Cover Page..

Abstract:

[remember stating research gap here]

Table of Contents here

1.Introduction:

1.1 Overview

Artificial Intelligence Generated Content (AIGC) has become a prominent topic in the realm of AI technology, with applications spanning from textual content to images and audio. Defined by Cao et al. (2023a) as content synthesized automatically by AI tools based on human instructions within a short period, AIGC leverages advancements in generative AI models across various media. Notable examples include the text generation oriented multi-modal Chat-GPT (OpenAI, 2022), the image generation model DALL-E 3 (Betker et al., 2023), and the audio generator Udio v1.5 (Udio, 2024).

Among these, image generation particularly stands out as a focal area of interest, both academically and industrially. Given the significant expenditure in marketing, with over $1 trillion USD spent globally in 2023 on advertising (Statistica, 2024), there is a compelling business case for exploring the synthesis of images for commercial applications like advertisements, marketing campaign posters, or photography briefings for freelance photographers.

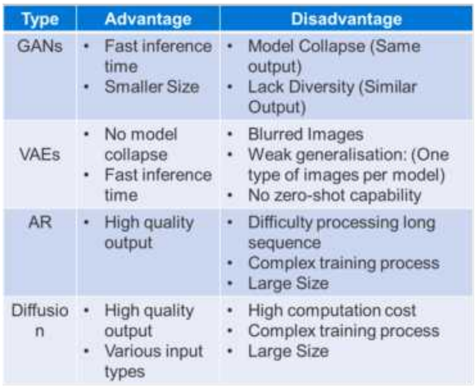
Despite the burgeoning use of AI in these contexts, there remains a scarcity of research directly comparing different diffusion-based image generation models for marketing purposes, nor is there a comprehensive framework guiding the selection of appropriate models based on specific marketing needs. This report addresses these gaps by examining whether generative AI models can produce high-quality images suited for marketing and proposes a framework to help users select the most suitable image generation models for various marketing requirements.

1.2 What is Image generation?

Image generation is a vital task in Computer Vision (CV) that leverages diverse inputs such as text, images, and audio to create new images. This area has seen significant advancements, particularly in text-to-image (T2I) generation and image-to-image (I2I) translation. T2I generation transforms textual descriptions into high-quality images, while I2I translation modifies input images based on textual descriptions, enhancing the field's scope and applicability (Elasri et al., 2022; Isola et al., 2018; Bie et al., 2023).

[Draw two diagrams that illustrate T2I and I2I process here]

The proliferation of sophisticated models in the T2I and I2I sectors highlights the rapid development within this domain. Mainstram types of image generation models are among Generative Neural Networks (GANs), Autoregressive Models (AR), Variational Autoencoders (VAEs) and Diffusion model. Details of these different model types are attached in the Appendix.



Type	Advantage	Disadvantage
GANs	• Fast inference time • Smaller Size	• Model Collapse (Same output) • Lack Diversity (Similar Output)
VAEs	• No model collapse • Fast inference time	• Blurred Images • Weak generalisation: (One type of images per model) • No zero-shot capability
AR	• High quality output	• Difficulty processing long sequence • Complex training process • Large Size
Diffusio n	• High quality output • Various input types	• High computation cost • Complex training process • Large Size

Crucially, the progress in image generation is closely tied to advancements in deep learning, especially in Natural Language Processing (NLP) and Computer Vision (CV) technologies, which enable the production of high-quality images (Bie et al., 2023).

Nowadays, an increasing number of cutting-edge image generation models are developed in T2I and I2I fields. However, Image generation cannot flourish without the development of deep learning. The advancement of deep learning, particularly the Natural Language Processing (NLP) and the Computer Visions (CV), has rendered the feasibility of high-quality image generation (Bie et al., 2023).

The development of deep learning can be traced back to 1950s in which Rosenblatt (1958) firstly proposed the idea of Perceptron and Widrow (1960) proposed the neural network with single layer called ADALINE (Adaptive Linear Neuron) to simulate the behaviour of human brains and solve linearly separable classification problems. Then in 1970s, Amari (1967) deployed the stochastic gradient descent method to a feedforward network to handle non-linearly separable classification problems. Later, the modern back-propagation method which relies on the chain-rule in gradient to update errors of each neuron has been firstly introduced by Linnainmaa (1976). This method was further generalised by Werbos (1982) and was intensively experimented by Rumelhart, Hinton and Williams (1986). Then inspired by the cognitive science, other crucial techniques such as Long-Short-Term Memmory (LSTM) (Hochreiter and Schmidhuber, 1997), dropout method (Hinton et al., 2013) that mitigates overfitting and the gradient based optimiser ADAM (Kingma and Ba, 2014) also emerged in the development of deep learning.

With these advancements of deep learning, many deep learning architectures such as Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Autoencoders, Generative Neural Network (GAN) have been developed in solving computer vision tasks and language processing problems. Particularly, the image generation has benefited by these advanced architectures. For instance, the CNN-based methods such as U-Net (Weng and Zhu, 2015) has been applied to deffusion-based image generation models in the iterative image denoising process and the PixelCNN (van den Oord et al., 2016a), an image generation model which is designed to predict future pixels based on given pixel-level information on an iterative manner, has been utilised to generate images. Similarly, van den

Oord et al. (2016b) also proposed the RNN-based image generation model PixelRNN to synthesise images by predicting next pixel’s colour conditioned on all previous information.

In recent years, with the new architecture Transformer and attention mechanism being introduced by Vaswani et al. (2017), transformer-based large-scale models such as T5 (Raffel et al., 2019), BERT (Devlin et al., 2019), CLIP (Radford et al., 2021), GPT (Brown et al., 2020) and Llama (Touvron et al., 2023) have been integrated into modern image generation models. For example, the T2I model Imagen (Saharia et al., 2022) deploys T5 (Raffel et al., 2019) as its text encoder while CLIP (Radford et al., 2021) is incorporated into another T2I model DALL-E (Ramesh et al., 2021). These innovative applications have enabled the image generation task to leverage the power of large-scale models and hence rendering the image generation into a new era.

1.3 About the Company

Company A is a UK-based large multinational enterprise in finance sector. To embrace the AI transformation and deploy potential benefits brought by Artificial Intelligence, company A has already established a Data Science team to help the group understand new AI techniques, analyse the feasibility of adapting these AI techniques and explore the potential benefits that may be brought to the firm.

Recently, the advancement of image generation models has rendered the company to explore the possibility of deploying them thus adding benefits to the firm, particularly the marketing team that utilises enormous images for marketing campaigns. However, there are some tangible challenges for Company A to utilise these models. Firstly, available models have not yet been systematically compared based on the requirement from the marketing team and no sufficient guidance has been given to users for model selection, which raise challenges for internal users to choose suitable models with appropriate image generation capabilities and associated costs. Secondly, there is no framework that guides users to generate and edit images for marketing use. This means even if appropriate models have been provided, users may not be able to synthesise right images without additional help. Finally, there is not yet known by the company that whether image generation models can synthesise high quality images for marketing purposes due to the lack of marketing-oriented experimentation. If they can generate desired images, it is not yet known what their capabilities are and how far we are from generating images that are immediately available to use by the marketing team. Therefore, company A sponsored this commercial project to answer their questions and concerns.

1.4 Objective:

The objectives of this project will be concentrated on the followings:

1. Exploring whether high quality images for marketing purposes can be produced by image generation models.

2. Designing an image generation framework for users from marketing team to apply appropriate models to synthesise and edit images based on various needs.

1.5 Preliminaries:

1.5.1 Environment & Platforms:

- **Python 3.8.19:** The Python 3.8.19 has been used through the whole project.
- **Azure Machine Learning (AML):** The AML is the main platform of deploying Stable Siffusion (SD) models for this project. Variants of SD models such as SDXL base, SD v1.4, SD v1.5 and SDXL refiner have been deployed on AML.
- **Azure OpenAI Studio:** DALL-E 3 is employed via Azure OpenAI Studio for this project to generate images.
- **DreamStudio:** Initial testing of SD models, testing the effectiveness of keywords and the deployment of SDXL inpainting model are conducted via DreamStudio
- **Fireworks AI:** Testing SD 3 Large and SD3 Medium are in Fireworks AI.

1.5.2 Tools:

- Image Masker: a tool that generates mask images for inpainting tasks.
- Packages:
    - torch, torchvision, transformers, pillow, cv2, skimage, numpy, pandas, datasets, matplotlib, base64, os, json, io are used for processing, converting and restoring image data.
    - piq and torchmetrics are used for evaluating generated images quality.
    - azure, ipykernel are used for deploying models from AML on python notebooks.

1.5.3 Key Definitions:

- **Prompt**: A prompt is the textual input given to models to guide the image generation process.
- **Photorealism**: Used to describe the synthetic images look like real photos taken by photographers.
- **Negative Prompt**: The textual instruction we want the model to avoid in the generated content. Typically, only Stable Diffusion 2 and later versions support this.
- **Mask Image:** Mask image is the black-and-white image that guides the inpainting models to edit the specific part of the given image. The white area in the mask image is where we wish to edit on the given image while the black area is where we wish to remain untouched.
- **SOTA**: The abbreviation of State-Of-The-Art.
- **Zero-shot**: A method where the AI model can solve problems that they have not seen before during the training stage.
- **Few-shot**: A method where the AI model can solve problems by giving them a few numbers of examples.
- **fine-grained**: Used to describe the model can understand and conduct instructions from the prompt at word-level.

- **Black-box model**: The model that is not open sourced and whose detailed structures and parameters are not publicly available.

1.6 Structures of Report

This report consists of 8 sections with 1 appendix attached.

Section one is the introduction of the project which describes the background and objective.

Section Two is the literature review in terms of image generation models, image evaluation metrics and AI in marketing.

Section Three is the methodology for systematically comparing image generation models and the framework setup for image generation task based on marketing purposes.

Section Four is the experimentations of methodology.

Section Five is the results and analysis for image generation models.

Section Six is the framework of image generation.

Section Seven is the application of the framework

Section Eight will give a conclusion and future research for this project.

Appendix will list detailed results from the experimentation.

2.Related Work:

[History of Text-to-image generation models graph here]

2.2 Text-to-Image (T2I) Generation

2.2.1 T2I with GANs

Generative Adversarial Networks (GANs) have significantly impacted text-to-image (T2I) tasks. Following the foundational work by Goodfellow et al. (2014), numerous GAN-based T2I models have emerged, categorized into multi-stage GANs, single-stage GANs, and scaled-up GANs with large pre-trained models or datasets.

In 2017, Zhang et al. introduced StackGAN, a T2I model generating low-resolution images in the first stage and refining them to high-resolution (256x256 pixels) images in the second stage, conditioned on the original text and initial sketches. This approach produced photo-realistic images with an FID score of 74.05 on the COCO dataset. However, the stacked architecture risked model collapse if the initial stage failed. To address this, Zhang et al. (2018) developed StackGAN++, incorporating a tree-like multi-stage GAN architecture and color-consistency regularization, enhancing image quality and stability.

Xu et al. (2018) further advanced the field with AttnGAN, a 230-million-parameter model using an attentional generative network and a deep attentional multimodal similarity model (DAMSM). This model significantly improved Inception Scores (IS) on COCO and CUB datasets, achieving an FID score of 35.49 on COCO, a 56% reduction compared to StackGAN++.

Despite their success, multi-stage GANs face limitations, such as entanglement among generators, reduced supervision capabilities, and computational costs associated with attention mechanisms (Tao et al., 2022). To overcome these, Tao et al. (2022) proposed DF-GAN, a 19-million-parameter single-stage model that efficiently generates high-resolution images without the complexities of multi-stage architectures. While simpler and smaller, DF-GAN's reliance on sentence-level text information limits its fine-grained image synthesis capabilities.

More recently, research has shifted towards GANs with contrastive learning to enhance text-image alignment. Zhang et al. (2022) introduced XMC-GAN, which leverages contrastive learning to maximize mutual information between text and image, achieving a remarkable FID score of 9.33 on the COCO dataset. Inspired by this, Tao et al. (2023) developed GALIP, powered by the CLIP model (Radford et al., 2021), achieving results comparable to modern autoregressive (AR) and diffusion models with significantly faster inference times and smaller model sizes.

Kang et al. (2023) explored the feasibility of training GANs with extremely large datasets, introducing the 1-billion-parameter GigaGAN trained on the LAION2B-en dataset, enabling high-quality image generation at a fast pace despite increased training times.

While GANs have unique advantages, their drawbacks, such as limited diversity, challenging training, and potential model collapse, have driven researchers to explore alternative generative models (Dhariwal & Nichol, 2021).

2.2.2 T21 with VAEs.

VAEs can also be employed for T2I tasks. In 2017, van den Oord, Vinyals and Kavukcuoglu (2017) proposed Vector Quantised – Variational Autoencoders (VQ-VAE) consists of discrete latent representations and an autoregressive prior. VQ-VAE can handle a variety of tasks such as image generation, speech recognition and video generation and can avoid the posterior in the VQ-VAE being collapsed (van den Oord, Vinyals and Kavukcuoglu, 2017), although there are no modern evaluation metrics contained in the paper to verify VQ-VAE’s capabilities. Inspired by this architecture, Razavi, van den Oord and Vinyals (2019) proposed VQ-VAE 2 to generate images with high quality and improved diversity. The newly proposed VQ-VAE 2 improves the autoregressive priors by scaling it into multi-scale hierarchical structures hence achieving higher image qualities that are comparable to previous GAN-based models while avoiding the risk of model collapse and the lack of diversity (Razavi, van den Oord and Vinyals, 2019).

VAEs seems not a popular choice for text-to-image generation due to its inherent limitations such as information loss and blurriness of generated images. However, VAE architectures have been widely used by numerous T2I models as encoders to reduce data dimensionalities, hence improving computational efficiencies. One example is the Stable Diffusion, a latent diffusion model that applies VAE's encoder-decoder structure for transferring image into latent representation and decoding it into generated image (Rombach et al., 2022). Another example can be CogView, an AR model proposed by Ding et al. (2021) which uses VQ-VAE to encode image tokens.

2.2.3 T2I with Autoregressive Models:

Autoregressive (AR) models, originally designed for text-based tasks, can be adapted for text-to-image (T2I) generation with certain modifications. Bie et al. (2023) describe how AR models, integrated with Transformers, generate images by sequentially predicting image tokens, conditioned on text tokens produced by a text encoder. This process divides the AR model into three components: a text encoder that converts text into tokens, a prior that uses these text tokens to generate image tokens, and a decoder that transforms the image tokens back into a visual image.

Significant advancements in AR models for T2I tasks have been noted recently. In 2021, Ramesh et al. introduced DALL-E, a pioneering T2I AR model utilizing a Transformers architecture, which processes language and image tokens as a unified data stream. DALL-E, equipped with 12 billion parameters, achieved a zero-shot FID score of 17.9 on the COCO dataset. The model also underwent specialised human evaluations, where over 90% of the participants found the images generated by DALL-E to be more photorealistic and accurately matched to captions compared to other models. DALL-E's training involves two stages for efficiency without compromising image quality: initially, a discrete variational autoencoder (dVAE) compresses and converts images into 32x32 image tokens. Subsequently, these tokens are merged with text descriptions encoded by a BPE encoder, and the combined stream is processed by the Transformer. This allows DALL-E to generate images from text prompts alone, providing zero-shot capabilities that suggest potential for generating previously unseen images. However, this zero-shot approach might be less effective for specialized datasets like CUB, indicating a need for model fine-tuning for improved performance. Simultaneously, Ding et al. (2021) developed CogView, a 4-billion-parameter AR model tailored for T2I tasks, employing VQ-VAE for image token generation. CogView achieved state-of-the-art (SOTA) FID scores on a blurred version of the COCO dataset, outperforming DALL-E and other models like DF-GAN and AttnGAN, while retaining zero-shot capabilities. However, images from CogView might exhibit some blurriness due to its VQ-VAE architecture.

Further enhancing the AR model's ability to produce photorealistic images, Yu et al. (2022) introduced the Pathways Autoregressive Text-to-Image (Parti) model. Parti, with 20 billion parameters, uses ViT-VQGAN as its image encoder, following a training methodology similar to DALL-E's. This adaptation has led Parti to achieve both SOTA zero-shot FID and fine-tuned FID scores on the COCO dataset, making it comparable to diffusion-based models like Imagen.

Additionally, the Parti model incorporates the holistic benchmark PartiPrompts, which evaluates its performance on complex tasks, illustrating its robustness in handling challenging prompts (Yu et al., 2022). However, the model does encounter issues such as color blending, omissions, and incorrect object counts as prompts increase in length and complexity.

Beyond the developments in models like DALL-E and Parti, there is also significant interest in enhancing the priors of autoregressive (AR) models. Radford et al. (2021) introduced the Contrastive Language-Image Pre-training (CLIP) model, designed to identify correct image-text pairings from batches of data, thus aiding in both image generation and captioning tasks. CLIP maintains zero-shot capabilities, marking a significant advancement in zero-shot classifiers for computer vision that are adaptable and applicable to various contexts. However, CLIP's training on unfiltered, uncurated data collected online raises concerns about the propagation of social biases and unethical content.

Addressing the limitations of CLIP, Pan et al. (2022) proposed Knowledge-CLIP, which integrates semantic information into the CLIP framework, thereby enhancing its semantic alignment between visual and language representations and improving its reasoning across different scenarios. To date, no open-source text-to-image generation models have utilized Knowledge-CLIP to enhance image quality, representing a potential area for future research to explore how semantic depth can refine image generation processes.

To date, we have highlighted the remarkable zero-shot capabilities of autoregressive (AR) models. However, several significant limitations also merit discussion. Firstly, many AR models exhibit slow image generation speeds, as they process images token-by-token, which can be inefficient (Ding et al., 2021). Moreover, the use of longer text prompts not only increases inference time but also raises the likelihood of errors in models like Parti (Yu et al., 2022). Another concern is the substantial size of AR models compared to their GAN-based and diffusion-based counterparts, which necessitates higher hardware capabilities for implementation. For instance, DALL-E contains 12 billion learnable parameters, whereas the diffusion model DALL-E 2, from the same developer, has only 6.5 billion parameters, showcasing the inherently large scale of AR models. Furthermore, the data sources for training these models also pose issues. The datasets, primarily sourced from the internet, consist of images and text that are neither filtered nor curated. This unmoderated data can lead to the generation of unethical or inappropriate (NSFW) content, presenting significant risks and ethical concerns in deploying these models without stringent oversight (Yu et al., 2022).

2.2.4 T2I with Diffusion Models

Diffusion models have become increasingly prominent in the field of text-to-image (T2I) generation, with popular systems like Stable Diffusion, DALL-E 3, and Midjourney exemplifying this trend. These models utilize advanced techniques discussed in Section 2.14, including U-Net architectures, Denoising Diffusion Implicit Models (DDIM), guidance methods, and ControlNet. Early iterations, such as the Guided Language to Image Diffusion for Generation and Editing (GLIDE) system introduced by Nichol et al. (2022), leveraged classifier-free guidance to produce photorealistic images in a zero-shot manner, surpassing AR-based models like DALL-E in human evaluations.

Following this, Saharia et al. (2022) developed the Imagen model, which combines the strengths of the Large Language Model T5 with a diffusion model to generate exceptionally realistic images. Imagen achieved a state-of-the-art FID score of 7.27 on the MS COCO dataset and performed better than DALL-E 2 in DrawBench human evaluations.

Recent developments have diversified into various branches of diffusion models, including those with a Prior-Decoder architecture. For instance, DALL-E 2, introduced by Ramesh et al. (2022), incorporates a diffusion-based prior for CLIP image embeddings, coupled with a decoder that generates images conditioned on these embeddings. This setup has allowed DALL-E 2 to achieve high image quality with greater diversity. Another example is the Corgi model proposed by Zhou et al. (2023), which integrates CLIP knowledge and trains with a dataset where less than 2% of images have associated text descriptions, achieving comparable FID scores to leading models.

The Mixture of Expert (MOE) Diffusion approach represents another significant branch. The ERNIE-VILG 2.0 model by Feng et al. (2023) combines fine-grained visual and textual information, utilizing various denoising experts at different stages. This model not only generates high-fidelity images but also ensures better alignment between text and image. Similarly, the eDiff-I model introduced by Balaji et al. (2023) uses expert denoisers and multiple encoders, like CLIP and T5, to handle textual and visual information effectively across different stages, enabling the creation of images with various artistic styles.

Additionally, Retrieval-Augmented (RA) Diffusion models have shown promising developments. Blattmann et al. (2022) applied a retrieval strategy to the Latent Diffusion Model (LDM), significantly enhancing image quality and diversity. Following this concept, Chen et al. (2022) introduced the Re-Imagen model, which improves the visual appearance accuracy of components in synthetic images by augmenting it with summarized semantic and detailed visual knowledge.

Finally, to afford more precise control over the semantic information and shapes of different parts in synthetic images, Avrahami et al. (2022) proposed the SpaText method. This approach utilizes segmentation maps and text prompts as inputs for diffusion models, facilitating the management of complex image generation tasks under varied conditions.

Similar to GAN-based image generation techniques, diffusion models have also integrated features from other generative methods to enhance their capabilities. Yin et al. (2024)

introduced the Distribution Matching Distillation (DMD) method, which incorporates GAN architectures to transform multi-stage diffusion models into a single-step generator. This adaptation significantly speeds up the image generation process while maintaining high-quality outputs. Additionally, Gu et al. (2022) developed the VQ-Diffusion model, which combines Vector Quantised-Variational Autoencoders (VQ-VAEs) and a conditional Discrete Denoising Diffusion Probabilistic Model (DDPM), effectively reducing biases and error accumulation throughout the generation process.

The realm of diffusion models is broad, including techniques like Diffusion with Prior-Decoder architecture, Mixture-of-Expert Diffusion, Retrieval-Augmented Diffusion, and hybrids with other generative model frameworks. However, one of the most significant areas of research involves Latent Space Diffusion, a method widely adopted by the Stable Diffusion model family. In 2022, Rombach et al. (2022a) unveiled the novel Latent Diffusion Model (LDM), which generates images in a compressed latent space using a cross-attention mechanism. This approach dramatically reduces computational costs during model training without sacrificing image quality, paving the way for models like Stable Diffusion versions 1 and 2. Building on this foundational model, Deci.ai (2023) introduced the Deci Diffusion 1.0 model, which utilizes an optimized U-Net architecture with fewer parameters, achieving approximately three times faster inference times compared to the SD 1.5 models while maintaining similar quality levels. This innovation has also reduced image generation costs by 66%. Concurrently, Betker et al. (2023) released DALL-E 3, which enhances prompt comprehension and image quality over its predecessors. Simultaneously, Podell et al. (2023) developed the SDXL model, a high-capacity Stable Diffusion variant with a larger U-Net and more attention blocks, increasing computational demands but outperforming previous models in human evaluations and achieving results on par with other leading systems like Midjourney v5.1.

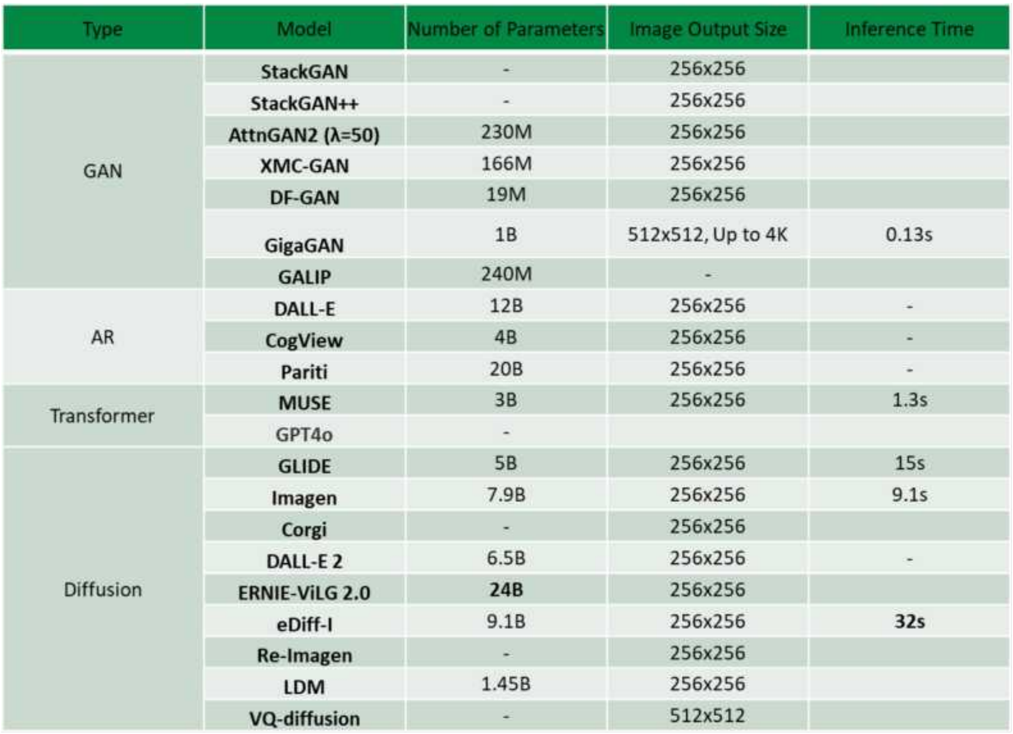
Further advancements came from Esser et al. (2024), who proposed a novel Rectified Flow (RF) architecture that utilizes separate transformers for text and image modalities, allowing bidirectional information flows and significantly improving both the quality and efficiency of generating high-resolution images. This technology underpins the latest in the Stable Diffusion lineup, the SD3 Large and SD3 Medium models.

Additionally, Sauer et al. (2024) introduced the Latent Adversarial Diffusion Distillation (LADD), which simplifies the training process and enhances overall performance. By combining LADD with RF, the team created the Stable Diffusion 3 Turbo (SD3 Turbo), which is capable of producing ultra-high-quality images rapidly, showcasing the evolving potential and multifaceted development within the field of diffusion models.

2.2.5 T2I with Other Models:

Apart from types of models mentioned above, there are also T2I models with different architectures. In 2023, Chang et al. (2023) proposed Muse, a transformer-based T2I model that achieves SOTA performances and is more efficient than diffusion and autoregressive models such as DALL-E 2 and Imagen because it employs discrete tokens and requires less sampling iterations. Additionally, Lai et al. (2023) leverages both large language models (LLMs) and diffusion models to make communications between users and models more effective hence improving the image quality. Lai et al. (2023) firstly used a router to analyse the response of the LLM and then they applied an adapter to change the image embedding or descriptions for the subsequent T2I models. With this setting, Lai et al. (2023) proposed interactive Text-to-Image framework, which may be a useful method to improve image quality.

[T2I model summary table here]



Type	Model	Number of Parameters	Image Output Size	Inference Time
GAN	**StackGAN**	-	256x256	
	StackGAN++	-	256x256	
	AttnGAN2 (λ=50)	230M	256x256	
	XMC-GAN	166M	256x256	
	DF-GAN	19M	256x256	
	GigaGAN	1B	512x512, Up to 4K	0.13s
	GALIP	240M	-	
AR	**DALL-E**	12B	256x256	-
	CogView	4B	256x256	-
	Pariti	20B	256x256	-
Transformer	**MUSE**	3B	256x256	1.3s
	GPT4o	-		
Diffusion	**GLIDE**	5B	256x256	15s
	Imagen	7.9B	256x256	9.1s
	Corgi	-	256x256	
	DALL-E 2	6.5B	256x256	-
	ERNIE-ViLG 2.0	**24B**	256x256	
	eDiff-I	9.1B	256x256	**32s**
	Re-Imagen	-	256x256	
	LDM	1.45B	256x256	
	VQ-diffusion	-	512x512	

2.3 Image-to-Image (I2I) Generation

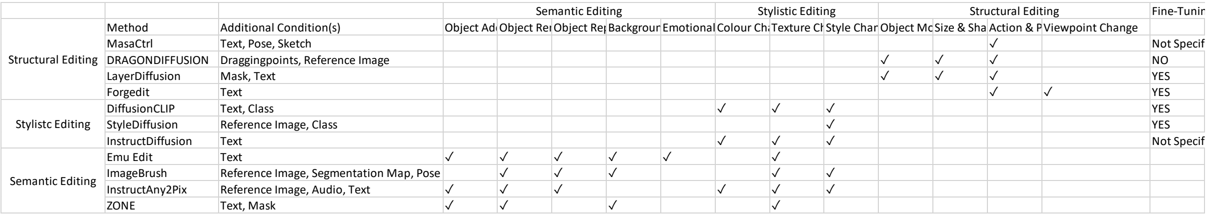
2.3.1 I2I Background

Pang et al. (2022) outlined how Image-to-Image (I2I) models address various computer vision tasks including image-inpainting, super-resolution, style transfer, and image extension. Initially, GAN-based models were prominent in these areas, with Isola et al. (2018) describing GANs as versatile solutions for I2I translation. Notable GAN-based models include CycleGAN (Zhu et al., 2020), BicycleGAN (Zhu et al., 2018), and those identified by Gonzalez-Garcia, Weijer, and Bengio (2018) for semantic synthesis, along with models by Pathak et al. (2016), Zhu et al. (2018), and Song et al. (2018) for inpainting. Additionally, models by Hertzmann et al. (2001) and Ren, Romano, and Elad (2016) excel in style transfer and super-resolution respectively.

Recent developments in I2I models have shifted towards Diffusion-based models, yet GANs, Transformers, and VAEs remain integral, particularly in enhancing face, artifact, and eye restoration within diffusion models. For example, the transformer-based CodeFormer (Zhou et al., 2022) corrects artifacts and blurriness in images produced by the Stable Diffusion model.

Furthermore, Text-to-Image (T2I) models that combine image encoders and decoders could theoretically undertake I2I tasks. As discussed in Section 2.2, models like GLIDE (Nichol et al., 2022) and AttnGAN (Xu et al., 2018) demonstrate capabilities in inpainting and image refinement, respectively. Large-scale models such as Stable Diffusion (Rombach et al., 2022a), SDXL (Podell et al., 2023), and eDiff-1 (Balaji et al., 2023) also support style transfer and image refinement. Moreover, SDXL can efficiently perform inpainting tasks when integrated with SDEdit (Meng et al., 2021).

Despite the proficiency of large-scale text-to-image diffusion models in producing high-quality images from detailed textual descriptions, they often struggle with precise editing of generated or real images (Mou et al., 2023). Consequently, this dissertation will focus on the latest diffusion-based Image-to-Image (I2I) models that retain image editing capabilities. Huang et al. (2024) examined over 100 diffusion-based I2I models, categorizing their editing functionalities into semantic, stylistic, and structural editing. Within semantic editing, further subdivisions include tasks such as adding or removing objects. The subsequent sections will organize I2I models according to their editing capabilities, adhering to the framework established by Huang et al. (2024).



			Semantic Editing					Stylistic Editing			Structural Editing				Fine-Tunin
	Method	Additional Condition(s)	Object Ad	Object Re	Object Re	Backgrour	Emotional	Colour Ch	Texture C	Style Char	Object Mc	Size & Sha	Action & P	Viewpoint Change	
Structural Editing	MasaCtrl	Text, Pose, Sketch											✓		Not Specif
	DRAGONDIFFUSION	Draggingpoints, Reference Image									✓	✓	✓		NO
	LayerDiffusion	Mask, Text									✓	✓	✓		YES
	Forgedit	Text											✓	✓	YES
Stylistc Editing	DiffusionCLIP	Text, Class						✓	✓	✓					YES
	StyleDiffusion	Reference Image, Class								✓					YES
	InstructDiffusion	Text						✓	✓	✓					Not Specif
Semantic Editing	Emu Edit	Text	✓	✓	✓	✓	✓		✓						
	ImageBrush	Reference Image, Segmentation Map, Pose		✓	✓	✓			✓	✓					
	InstructAny2Pix	Reference Image, Audio, Text	✓	✓	✓			✓	✓	✓					
	ZONE	Text, Mask	✓	✓		✓			✓						

2.3.2 Semantic Editing

In their semantic editing framework, Huang et al. (2023) identified key subtasks including object addition, removal, replacement, emotional changes, and color modifications. Several diffusion models are adept at these tasks. For instance, Li Singh and Grover (2023) introduced InstructAny2Pix, a versatile multi-modal instruction-following system capable of performing object addition, removal, and replacement using prompts, audio, and images. Subsequently, Yang et al. (2024) developed ImageBrush, which relies solely on visual instructions such as segmentation maps, poses, and reference images. Yang et al. (2024) suggests this approach better captures human intentions, enhancing its applicability in real-world scenarios. Concurrently, Li et al. (2024) unveiled Zero-Shot Instruction-Guided Local Editing, a method allowing precise, localized edits without requiring detailed masks or prompts, effective for tasks like background changes and object removal.

Additionally, Sheynin et al. (2024) proposed Emu Edit, a multi-task image editing model that operates solely on text input and can execute a broad spectrum of edits as directed by natural language. Unique to Emu Edit is its integration of learned task embeddings, which enhances its few-shot learning capabilities. To our knowledge, Emu Edit is the sole model capable of addressing all five subtasks delineated by Huang et al. (2024).

2.3.3 Stylistic Editing

When it comes to stylistic editing, Huang et al. (2024) divided it into colour change, texture change and style change. Kim, kwon & Ye (2022) proposed DiffusionCLIP method that leverages the power of CLIP models and pre-trained diffusion models. By fine-tuning the pre-trained model and CLIP losses, DiffusionCLIP has shown excellent performances in zero-shot capabilities. In 2023, Wang, Zhao & Xing (2023) introduced a novel framework StyleDiffusion for style transfer subtask that disentangles original images and reference images using diffusion models without relying on the traditional assumptions such as Gram matrices and GANs. In 2024, Geng et al. (2024) also proposed InstructDiffusion interface which is designed for various vision tasks by leveraging diffusion models. With a textual input, all subtasks of stylistic editing such as colour change, texture change and style change can be achieved by using InstructDiffusion.

Apart from these models, other models mentioned in section 2.3.2 can also handle stylistic editing tasks. For example, Emu Edit (Sheynin et al., 2024) can change the texture of object while InstructAny2Pix (Li, Singh & Grover, 2023) can cope with all subtasks in stylistic editing.

2.3.4 Structural Editing

There are various I2I models that can handle structural editing tasks. Huang et al. (2024) has divided structural editing task into smaller subtasks such as moving object to elsewhere in the image, changing the shape or size of object, changing the object pose or actions and changing the viewpoint of the image. In 2023, Cao et al. (2023b) developed MasaCtrl, a tuning-free method that can realise consistent image generation and editing blended and stretched images simultaneously by using text and object sketch images, hence allowing the change of actions and poses of objects. Similarly, DragonDiffusion proposed by Mou et al. (2023) and the Layer Diffusion introduced by Li et al. (2023) can not only change the action and poses of images, but they also further preserve capabilities in moving the position and changing the size and shapes of image objects compare to MasaCtrl (Cao et al., 2023b), although DragonDiffusion does not accept textual input and LayerDiffusion may require additional mask images. When it comes to changing the viewpoint of images, Forgedit proposed by Zhang, Xiao & Huang (2023) seems to be the only diffusion model from our best knowledge. To achieve this capability, Forgedit adapted a vison-language joint optimisation framework for faster image reconstruction, a vector projection mechanism for controlling identity similarity and editing strength (Zhang, Xiao & Huang, 2023).

In stylistic editing, Huang et al. (2024) classify the tasks into color change, texture change, and style change. Kim, Kwon, and Ye (2022) developed the DiffusionCLIP method, which combines CLIP models with pre-trained diffusion models. By fine-tuning the pre-trained model alongside CLIP losses, DiffusionCLIP has demonstrated strong zero-shot capabilities. In 2023, Wang, Zhao, and Xing (2023) introduced StyleDiffusion, a novel framework for the style transfer subtask that uses diffusion models to disentangle original from reference images, moving away from traditional methods like Gram matrices and GANs.

In 2024, Geng et al. (2024) introduced InstructDiffusion, an interface designed for various vision tasks using diffusion models. This model allows for completing all subtasks of stylistic editing with textual inputs, including colour, texture, and style changes. Additionally, other models discussed in Section 2.3.2 also contribute to stylistic editing tasks. For instance, Emu Edit (Sheynin et al., 2024) can modify the texture of objects, while InstructAny2Pix (Li, Singh, & Grover, 2023) is capable of addressing all stylistic editing subtasks.

2.4 Evaluation Metrics & Datasets

The quality of generated images can be mainly measured by quantitative evaluation metrics and human evaluations while many prevalent image generation models are trained via large scale image datasets. In this section, we will explore popular evaluation metrics used by image generation and will discuss common datasets used for image generation model training.

2.4.1 Quantitative Evaluation Metrics

The quantitative evaluation metrics can be roughly divided into 3 categories: the distribution-based metrics, no-reference based metrics and full-reference based metrics. Each type of metrics is used based on different needs.

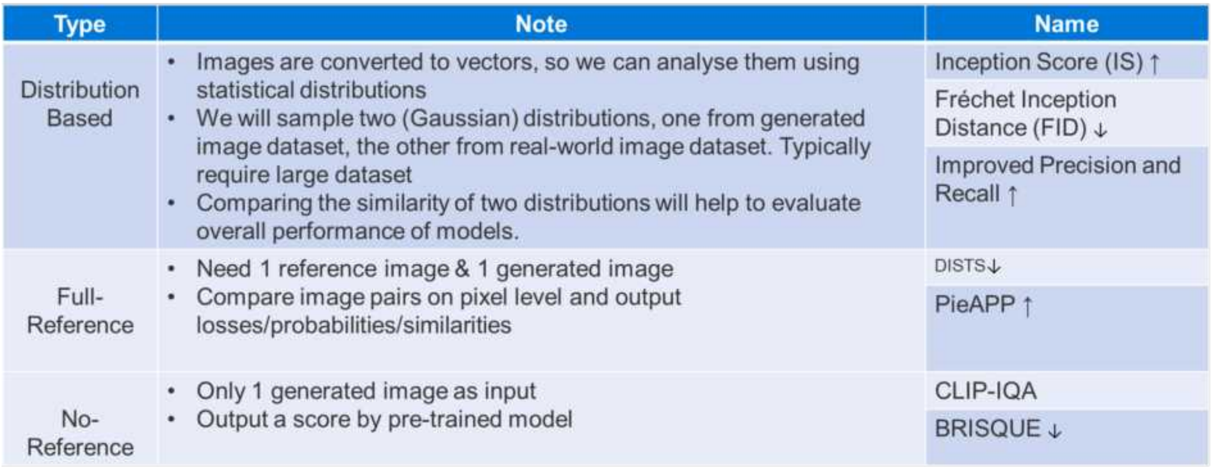
2.4.1.1 Distribution Based:

Distribution-based metrics are key in evaluating image generative models by comparing datasets of real-world and synthesized images transformed into high-dimensional vectors. These vectors, often assumed to follow multivariate Gaussian distributions, allow for the assessment of model quality through statistical analysis. In 2016, Salimans et al. introduced the Inception Score (IS) to measure the quality and diversity of images, positing that a higher IS indicates superior model performance (Salimans et al., 2016). However, IS overlooks the comparison with real-world images, a gap addressed by Heusel et al. in 2017 with the Fréchet Inception Distance (FID), which compares the distributions of actual and synthetic images, suggesting that a lower FID signifies better model performance (Heusel et al., 2017).

Today, FID is a prevalent metric in image generation evaluations, often used to benchmark new models. Despite its widespread use, FID has limitations, including the assumption that real-world images adhere to normality, which may not always be true due to biases and privacy concerns in training datasets (Jayasumana et al., 2024). Furthermore, Otani et al. (2023) and Jayasumana et al. (2024) have criticized FID for its lack of consistency with human judgments of image quality.

Addressing these concerns, Kynkäänniemi et al. (2019) introduced the Improved Precision and Recall (PR) metric, which assesses the overlap of image distributions to calculate precision and recall rates, thus evaluating the diversity and quality of synthesized images. According to Kynkäänniemi et al. (2019), this method offers a more reliable assessment than FID, where high-quality but low-diversity images result in high precision but low recall. This shift underscores the evolving landscape of evaluation metrics in image generation, emphasizing the need for measures that align more closely with human visual assessment.

[Draw a diagram]



Type	Note	Name
Distribution Based	• Images are converted to vectors, so we can analyse them using statistical distributions • We will sample two (Gaussian) distributions, one from generated image dataset, the other from real-world image dataset. Typically require large dataset • Comparing the similarity of two distributions will help to evaluate overall performance of models.	Inception Score (IS) ↑ Fréchet Inception Distance (FID) ↓ Improved Precision and Recall ↑
Full-Reference	• Need 1 reference image & 1 generated image • Compare image pairs on pixel level and output losses/probabilities/similarities	DISTS↓ PieAPP ↑
No-Reference	• Only 1 generated image as input • Output a score by pre-trained model	CLIP-IQA BRISQUE ↓

2.4.1.2 Full-reference Based:

Another type of quantitative metric is full-reference based evaluation metrics. The idea of this type of metric is that we compare the generated image with a ground truth image at pixel level hence deriving a score that compares the similarities of the image pairs. In 2018, Prashnani et al. (2018) proposed the PieAPP metric that measures the perceptual error of generated image when comparing to the reference image. In 2020, Ding et al. (2020) also introduced the DISTS metric that compares reference image and generated images.

Prashanani et al. (2018) claims that image quality judged by PieAPP is considerably correlated to the human opnions. Therefore, using these metrics may help users to perceive the quality of generated images.

[draw a diagram]

2.4.1.3 No-reference Based:

No-reference based metrics only requires images to be assessed as the input. In 2012, Mittal, Moorthy & Bovik (2012) proposed a no-reference based metric BRISQUE that evaluates image qualities on a spatial domain using the normalised luminance coefficients to quantify losses of “naturalness” in the image. By assessing the BRISQUE score, one can decide if the generated image is of high quality. By leveraging the power of CLIP (Radford et al., 2021), Hessel et al. (2021) firstly proposed a CLIPScore to evaluate prompt and image alignment, and then Wang, Chan & Loy (2023) further proposed the CLIP-IQA metric that can evaluate both physical qualities such as sharpness and abstract qualities such as happiness and aesthetics of images. By utilising the inherent capability of captioning images for CLIP model, CLIP-IQA chooses antonym prompt pairs to evaluate image qualities in different perspectives (Wang, Chan & Loy, 2023). To achieve the evaluation, CLIP-IQA will firstly convert the image into a textual description using pre-trained CLIP, then passing it to the model to compare this textual description with antonym prompt pairs to assess which prompt our textual description is more similar to by giving a similarity score from 0 to 1 (Wang, Chan & Loy, 2023). In this way, the CLIP-IQA score of each perspective could be used for assessing the quality of synthetic images.

[Draw a diagram for CLIP-IQA]

2.4.2 Human Evaluations

Human evaluation is another mainstream choice for assessing synthetic image quality. Due to its accessibility and low-level technical requirement, many image generation models have adapted human evaluations as part of their image quality evaluation process. For instance, models mentioned in section 2.2 such as DALL-E (Ramesh et al., 2021), Imagen (Saharia et al., 2022) and Parti (Yu et a., 2022) have used human evaluations such as ranking the output images from 1 to 7 and giving preferences over pairwise image comparison to assess their models and compare them with other models. However, the process of human evaluation varies in different models, meaning that they may not be comparable unless a complete comparison of intended models being tested by the human evaluation metrics. To cope with this issue, Otani et al. (2023) proposed a standardised and well-defined human evaluation protocol to facilitate the framework being verifiable and reproducible. Similarly, Petsiuk et al. (2022) proposed a human evaluation benchmark for text-to-image models over different applications to assess model capabilities.

2.4.3 Datasets for Image Generations

While some models utilise their own image data for training, many models have been using publicly available image datasets to train their own models.

- **MS COCO:**
  MS COCO (Lin et al., 2015) is the large-scale dataset containing images and their associated English image captions. It has been widely used for testing T2I model general performances with FID score when comparing with different models. For instance, Feng et al. (2023) test AR models such as DALL-E and Parti and Diffusion models such as GLIDE and Imagen using FID score on COCO dataset.
- **LAION:**
  The LAION dataset is the open-sourced dataset consisting of huge number of image and text pairs (LAION.ai, 2024). In 2022 Schuhmann et al. (2022) proposed LAION-5B, a dataset which contains more than 5.85 billion image-text pairs that are multilingual and filtered by pre-trained CLIP model. This dataset has significant importance to

Stable Diffusion models as many SD models such as SD2.1 and SDXL are trained by LAION-5B.

- **PartiPrompt:**
  Yu et al. (2022) also proposed PartiPrompt dataset when they introduced the Parti model. The PartiPrompt can be used for testing I2I model performances because they are designed to test different image editing capabilities.
- **DiffusionDB:**
  DiffusionDB is the dataset containing 2 – 14 million pairs of English prompts and Images generated by stable diffusion models (Wang et al., 2023). This dataset can potentially be used for testing T2I models as well.

2.5 AI Application in Marketing:

Dencheva (2023a) predicts that the AI market share in marketing will surge to $107.5 billion by 2028, highlighting the sector's growth. Dencheva (2023b) further supports this with a survey indicating that over 70% of U.S. marketing practitioners now employ AI tools, such as chatbots, within their operations. This uptake has spurred research into the use of generative AI models for marketing purposes, evidenced by a variety of applications across different platforms. For instance, Duolingo has integrated GPT-4 into a tool called Max, which aids in language learning through interactive role-play and explanatory feedback (Kshetri, 2023). Similarly, Goosehead Insurance has utilized Jasper.ai to generate marketing content for blogs (Kshetri et al., 2024), while Mattel has employed DALL-E 2 to design a model car toy (Mehta, 2023). Additionally, Jones Road Beauty and PrimeCare have adopted Meta’s AI sandbox and Midjourney, respectively, to produce diverse and artistic advertising content swiftly (Adams, 2023; Kshetri et al., 2024).

Exploring further, Zhang et al. (2024) have investigated the potential of diffusion models to create personalized marketing content, acknowledging a trade-off between textual alignment and image fidelity. This study underscores the feasibility of using diffusion models for marketing visual content, thus enhancing confidence in their application.

In a systematic comparison, Hartmann, Exner, and Domdey (2024) evaluated various diffusion models—including DALL-E 3, Midjourney v6, Firefly 2, Realistic Vision, and SDXL Turbo—across aspects such as image quality, realism, and aesthetics through human evaluations. Their findings reveal that while all models generally surpass human-made images in aesthetics, only Realistic Vision consistently achieves high realism. In terms of marketing efficacy, DALL-E 3 was shown to significantly improve click-through rates in A/B testing, positioning it as the most effective model for generating marketing content (Hartmann, Exner & Domdey, 2024).

However, the study by Hartmann, Exner, and Domdey (2024) also presents limitations. It overlooks the role of Image-to-Image (I2I) models, which are crucial for tasks such as editing original images multiple times to enhance advertisement quality. Additionally, the reliance solely on human evaluations rather than quantitative metrics could introduce bias and affect the robustness of the findings. This suggests a need for incorporating both I2I models and more objective evaluation metrics in future research to better assess the quality and impact of generated marketing images.

2.6 Literature Review Conclusion:

The comprehensive literature review conducted explores a wide range of image generation models, including Text-to-Image (T2I) and Image-to-Image (I2I) models, evaluation metrics, datasets, and applications within the marketing sector. The review highlights various architectures in T2I models such as GAN-based, autoregressive-based, and diffusion-based models, alongside I2I models that support semantic, stylistic, and structural editing.

[Visualisation of literature review]

Our analysis, reinforced by Table xx and Table xxx, indicates that diffusion-based models generally surpass other types in performance, albeit with a higher parameter count. This finding is corroborated by the Hugging Face leaderboard, which aligns with the conclusions drawn from our literature review. Given their superior performance in generating high-quality, marketing-oriented images, diffusion-based models are the preferred choice for this project. Users aiming to produce effective marketing visuals are therefore recommended to opt for diffusion-based models, which we will focus on in the subsequent phases of this project.

[https://huggingface.co/spaces/ArtificialAnalysis/Text-to-Image-Leaderboard]
[https://dreamstudio.com/start/]

However, there are tangible limitations for current T2I and I2I models as well. Since many diffusion models are trained on public data that do not filter out sensitive contents, there might be ethical issues and deepfakes when generating images. Also, when we empirically testing SDXL model on Dream Studio, we noticed that output images synthesised may contain artifacts, distorted face and illegible words hence requiring subsequent prompt engineering for image generation tasks.

When it comes to evaluation metrics, we have explored distribution-based metrics, no-reference-based metrics, full-reference-based metrics and human evaluations. Due to the nature of this project, we may initially choose no-reference-based metrics and human evaluations for the following reasons. Firstly, the distribution-based metrics such as FID and IS may not be suitable for this project. This is because these metrics primarily assess the general performances of models by generating millions of images but not for the single image. Since only limited number of images will be generated in this project, we may not use distribution-based metrics at this stage. Secondly, reference-based metrics may not be meaningful for this project either. This type of metrics is primarily used for comparing images manipulated by some compression algorithms to original images and therefore may not be meaningful for this project where the marketing team pays more attention to the quality of final output. Thirdly, although human evaluations are becoming more standardised than before, the biases, preference differences and knowledge differences among examiners may undermine the effectiveness of human evaluations.

As for generative AI application in marketing, we have displayed a couple of real-world applications by using AI models for marketing purposes. While these applications give us confidence in using image generation models to generate high quality marketing content, they also help us to form our research gap.

To start with, previous applications and research have limited comparisons among different diffusion based T2I models under image principles from marketing team using both quantitative metrics and human evaluations. Secondly, limited attentions have been paid to I2I models when generating images for marketing purposes in previous research. Occasionally, marketing teams may reuse their imagery assets with some editing procedures when they wish to provide customised advertisement based on specific requirement, meaning that using I2I models to edit these image assets may add tangible benefits for companies. Third, previous research has not combined T2I and I2I models together when generating marketing visual contents, which may hinder the full performances of diffusion models. Finally, there is no unified framework that guides users choosing models and generating images based on specific marketing needs.

These limitations have formed our research gap, and in this report, we will focus on answering if diffusion models can produce high-quality images for marketing purposes. This report will also provide a unified and reusable framework that can help users decide which model to use when they have specific requirement.

3. Method:

3.1 Background:

Company A has provided enormous resources to support the author during the project. In this project, Company A has provided their stock image database, photography principles, image briefings from marketing team and image generation models on Azure Machine Learning Studio and Azure OpenAI Studio. To leverage the power of SOTA models, we also obtain the access to Stable Diffusion 3 Large (SD3L), Stable Diffusion 3 Medium (SD3M), SDXL inpainting and GPT4o from the internet.

Based on resources provided by Company A and publicly available resources, we currently have access to the following models in the following table (Table 3):

Type	Name	Access From
T2I	Stable Diffusion v1.4 (SD1.4)	Azure Machine Learning Studio
	Stable Diffusion v1.5 (SD1.5)	
	Stable Diffusion v2.1 (SD2)	
	SDXL v1.0 base (SDXL)	
	DALL-E 3	Azure OpenAI Studio
	Stable Diffusion 3 Large (SD3L)	fireworks.ai []
	Stable Diffusion 3 Medium (SD3M)	fireworks.ai []
I2I	Stable Diffusion 1.5 inpainting	Azure Machine Learning
	Stable Diffusion 2.1 inpainting	
	SDXL-refiner	
	SD3L-refiner	fireworks.ai []
Multimodal	GPT4o	OpenAI

[redesign table]

[https://fireworks.ai/models/stability/sd3]

[https://beta.dreamstudio.ai/generate]

[https://portal.azure.com/#home]

[https://oai.azure.com/resource/overview?tid=5567eafd-e777-42a5-91bb-9440fd43b893]

[https://www.midjourney.com/home]

[https://github.com/AUTOMATIC1111]

Therefore, in this project, we will mainly focus on models available in company A. We will use images from company A's database due to availabilities. Other black-box models or Graphic User Interface (GUI) such as Midjourney and Automatic1111 are out of project scope and should be left as future research.

Furthermore, the general capabilities for large scale image generation models such as SDXL and SD2.1 have already been widely tested. However, there are insufficient comparisons of marketing-oriented capabilities for these models. Therefore, we wish to test T2I and I2I model capabilities that the marketing team emphasised on most.

3.1.1 Understand Needs of Company A

Since this is a commercial project, it is essential to evaluate models based on the sponsor's needs. Thus, we compare image generation models using criteria deemed important by the marketing team.

To achieve this, we analysed image content from company websites and other marketing sources, as well as gathered information directly from marketing teams. Additionally, we investigated the photography breifings when the marketing team commissions freelancers to create visual contents. Combining our findings, we summarised photo principles that the company considered important and extracted relevant keywords in Table 4.

[Photo Principles here]

Requirement	Description	Keywords
Principles	Legal & General photography adheres to a set of principles which embody our brand personality and ensure a strong sense of consistency. Our imagery is always colour and shot with natural lighting. They often feature people (real in appearance - rather than too obviously models), and they capture their subjects personality and activities in an authentic way. Our photography is always colour. This will ensure that our images have a collective warmth and realism. And since our branding is synonymous with colour, our images can include colourful highlights, such as an item of clothing or a given prop. (This should always feel like a natural element of any given image though, not a forced art direction or style straight jacket.)	consistency, colour (No black & white), authentic, natrual (not forced/staged)
Natural lighting	Our photography should appear as natural as possible. And so no obvious use of flash or studio lighting. Instead, it should be shot in daylight - not the harsh midday sun of unflattering shadows - but more the warm gentle light of morning and late afternoon, even some occasional grey skies and evenings are ok too.	Natrual looking, no obvious flash/studio lighting. Gentle light in the morning or late afternoon, grey skies or evnings OK sometimes
Real	Our subjects should always appear 'real' life. Candid, believable and unstaged. A natural smile for example, feels contagious in comparison to anything overtly forced, which moves no one.	Authentic, not staged, not forced
People	The people we choose to feature are like our customers themselves: diverse and from all walks of life (race, gender, sexuality, disability, age). Whilst they are often UK centric, we also represent our other markets, sometimes very specifically (this can often include the regional context of landscapes, houses, objects and pastimes).When talking about real people and real stories, don't shy away from eye contact. While a candid photo might feel more realistic, there is nothing unnatural about a genuine smile to the camera.And finally, when using photographs of Legal & General employees - please try to avoid the use of lanyards and passes within the image for security reasons.	diversity, UK centric, other regional context, eye contact, genuine smile, candid
Personality	Legal & General photography should exude a personality that's bright and real. Active and energetic. This reflects how many of our customers live their lives today. They're often volunteering or getting involved with challenges and charities, from running marathons or undertaking sailing trips, to going back to university and evening classes, to getting involved in community events.	bright and real, acitve, energetic, get involved in community events such as charity event, challenges, university, evening classes
Contextual imagery	In addition to people images, we use photography that adds context and layering to our understanding of our customers, their lives and what they choose to surround themselves with. These can be details or crops of activities, still life scenes or even eye-catching graphic details. They may show close-ups of people interacting with technology (such as a tablet), or a specific hobby. Still lifes may reveal customers lives, without the need to actually show them at all. From everyday accessories such as reading glasses, a favourite coat or hat, to 'lived in' household details, such as a piece of furniture with a half read newspaper or a vase of freshly cut garden flowers.	objects from daily life: daily accessories, household details; crops of activities: eye-catching graphic details, hobbies.
Depth of field	A narrower depth of field can really help focus on a chosen subject within any given frame. It makes for graphically simpler images that are free from unnecessary clutter. This is achieved by choosing to use a large aperture setting on good quality, fast lenses whilst shooting. The resulting images are easier to integrate and use (over busy, complex images) in combination with multiple other elements (such as headlines, supporting copy, overprinted panels and branding).	narrower depth of field
Middle ground	Using a narrow depth of field where both the foreground and the background are noticeably out of focus, can make for particularly graphic images. Images that can deliver much more atmosphere.	narrow depth of field where both the foreground and the background are noticeably out of focus
Angles and viewpoints	Unexpected or interesting angles and viewpoints, can create more dynamic, memorable images. Looking at the world differently, can help gain attention. This can be introduced in many ways, from seeing the world from a child's eye level, to an aerial shot of a cityscape.	Unexpected or interesting angles and viewpoints, seeing the world from a child's eye level, aerial shot of a cityscape
Avoid over cluttered environments	Photographing people in neutral, relatively uncluttered environments can help us focus on the subject in hand. By uncluttered environments, we don't mean completely empty spaces. With neutral environments, a subjects clothing takes on a more important role. It can either offer great contrast (if it's dark), or real stand out (by being colourful for example). The same can be achieved through using simpler flat colour backgrounds. Limited props and incidental details can offer context and insights into the subjects life and interests.	people in neutral, relatively uncluttered environments
Moments	Capturing meaningful moments that exude personality, requires a naturalism in our photography. And so rather than overly formal set ups, we should be capturing relaxed, real moments.	we should be capturing relaxed, real moments.
Graphic	Not all our photography is of people of course. We have areas of business that require different ways to illustrate them. Striking, graphic images of details or associated patterns are a good way to show complicated things, simply.	Striking, graphic images of details or associated patterns are a good way to show complicated things, simply
Abstract	Sometimes we need to talk about subjects that are tricky to convey in a single image. This is when we can use more abstract imagery. However, it is important to not fall back on clichéd storytelling devices. There are many creative ways to express ideas such as "growth" or "balance". Whatever the image might be, it should always depict real life, and not be too stylised, staged or fake.	not fall back on clichéd storytelling devices.depict real life, and not be too stylised, staged or fake.
Screens and interfaces	When showcasing an app, interface or website, we primarily try to show it being used in real life situations first. When more detail is needed, such as showcasing specific features or more of the user experience journey, a paired back minimal approach is used. This shows only the app itself, and no devices. We never show cutout monitors or devices shot in a studio, or in a fake white space.	show it being used in real life situations first. When more detail is needed, shows only the app itself, and no devices.
Photography don'ts	•Do not show overly staged or cheesy situations – it should always feel real •Do not use photography with awkward angles. •Do not use photography with obvious models. •Do not use black and white photography. •Do not use photography that is too dark. •Do not use photography that lacks colour, or appears washed out. •Do not go overboard with colour control – colour use should always appear natural •Do not rely on clichés when using abstract imagery.	overly staged, cheesy, awkward angles, obvious models (obvious gesture etc), blakc&white photo, too dark lighting, lack colours, overboard with colour control, cliched abstract imagery.

From Table 4, it has raised our attention that company A consider aspects such as authenticity, naturalness and positive atmosphere more when assessing qualities of images produced by photographers. Thus, our evaluation will focus more on photo principles to make the final output images more aligned to the real photographs.

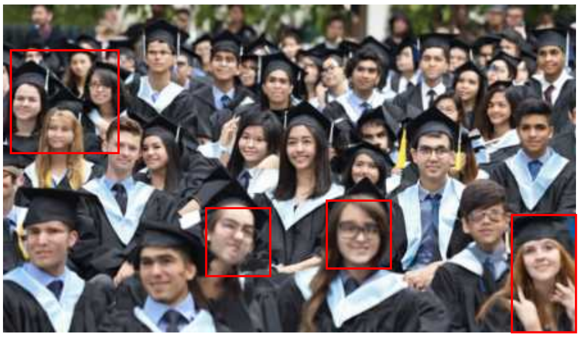
Moreover, we also categorise the marketing images to assess model performance across various use cases. From our investigation, we realised that natural landscapes, humans, cities, and illustrations are targeted images the company A utilised most frequently. Therefore, we will focus on producing a series of images regarding these categories when assessing the model performances.

3.1.2 Prompt Engineering Techniques

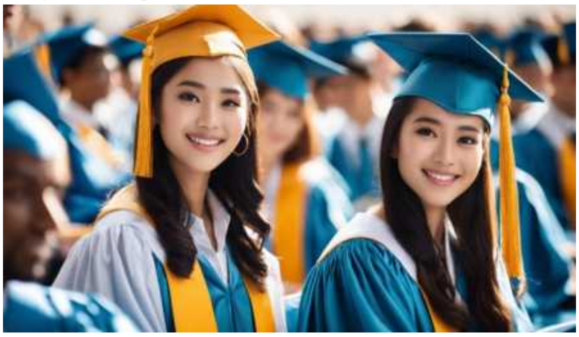
Since T2I models and I2I models leverage language encoders with natural language settings, it naturally gives us motivations to conduct prompt engineering for the textual input before utilising generative models. Recent advancement in AI has brought many powerful prompt engineering techniques for generative models. For instance, Wei et al. (2022) proposed the chain-of-thought (CoT) technique which enables generative models to achieve complex reasoning capabilities by dividing the task into finite intermediate steps. Also, Lewis et al. (2021) proposed a general-purpose approach for Retrieval-Augmented Generation (RAG) using fine-tuning techniques to solve complex tasks with the help of external sources. At the same time, Zhou et al. (2022a) also proposed Automatic Prompt Engineer (APE), a prompt engineering technique that can generate and select prompt instructions automatically.

Moreover, our initial testing also strengthened our motivations to improve overall performances using prompt engineering techniques. The following Figures xxx are generated via SDXL using different prompts. The left-hand-side images use the plain descriptions only while the right images use prompts with relevant keywords. Additionally, since SDXL can pass negative prompts into the model, we also use keywords as negative prompt to control contents we wish to avoid in our initial testing. For images synthesised by plain texts, we observed that faces on students and shapes of turbines within red boxes are blurred and distorted. Therefore, we may claim that adding keywords can substantially improve the image quality without changing the model architecture and we have realised the importance of properly designing the prompt before image generation.

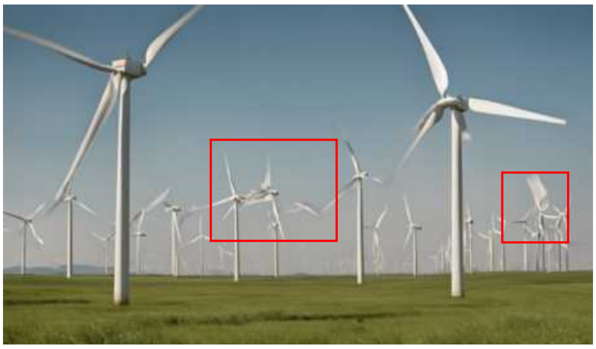
Prompt: *students are attending their graduation ceremony*



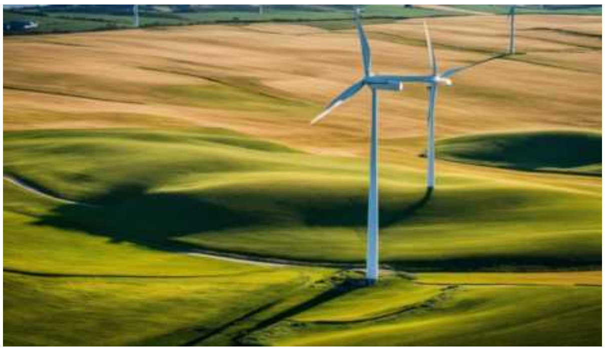
Prompt: *... ultra quality, sharp focus, clear, students are attending their graduation ceremony, detailed face…*
Negative Prompt: *blurriness, distorted face,…*



Prompt: large wind farm in a grassland



Prompt: *...land art, ultra quality, sharp focus, large wind farm in a grassland,, iStock, highly detailed...*
Negative Prompt: *3D, painting, illustration, disfigured...*



Although we have a number of techniques, we will adding keywords and main prompt as our prompt engineering technique for this project due to time constraints.

3.1.3 Keywords

Since we have discovered that adding keywords to the main prompt can considerably improve the overall image quality, we wish to explore if we could find some keywords that are useful in the image generation models and categorise them based on photography principles for marketing so that future users can design their prompt by finding and adding keywords from that summary table. To achieve this, we firstly proposed potentially useful keywords and then pass one keyword per time into SDXL to see if the model understands the single keyword. By filtering out ineffective ones, we illustrate some useful keywords and negative prompts in the Table 5. The full table is included in the appendix. We will mainly use these keywords to design our prompts when testing each candidate models.

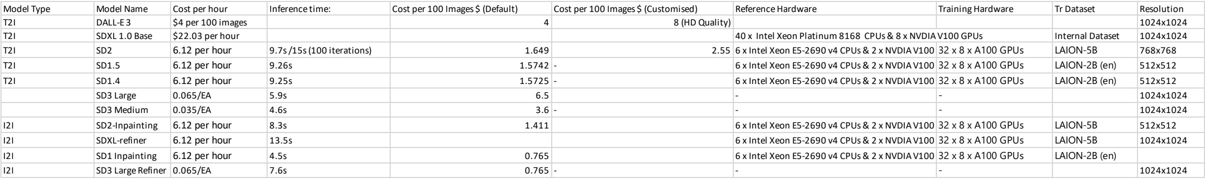
3.1.4. Evaluation Metrics:

Based on quantitative metrics discussed in the literature review, we found CLIP-IQA can examine both tangible quality of photographs such as resolution and sharpness and abstract quality such as aesthetics and happiness, which is more powerful than BRISQUE that can only predict image overall quality. Therefore, we will use CLIP-IQA as our quantitative metrics. For this project, we will choose “quality” and “natural” as the aspects being assessed. Here the “natural” aspect examines whether the image is real or generated by AI and “quality” aspect assesses the overall quality. However, before utilising CLIP-IQA, the initial testing will be conducted to check the robustness of this quantitative metrics.

As for human evaluation, inspired by Chatbot Arena, we are using pairwise comparisons over generated images. We will randomly select image pairs to ask respondents to choose a preferred one each time under the assumption that if they were members of marketing team. The results from respondents will be used to evaluate model performances and image quality.

3.1.4 Methodology Setup

Based on lit review, empirical test based on marketing requirement and resource availability, we tend to apply SD1.4, SD1.5, SD2, SDXL, DALL-E 3, SDXL Refiner, SD1.5 Inpainting, SD2.1 Inpainting and SDXL inpainting models for experimentation.



Model Type	Model Name	Cost per hour	Inference time:	Cost per 100 Images $ (Default)	Cost per 100 Images $ (Customised)	Reference Hardware	Training Hardware	Tr Dataset	Resolution
T2I	DALL-E 3	$4 per 100 images		4	8 (HD Quality)				1024x1024
T2I	SDXL 1.0 Base	$22.03 per hour				40 x Intel Xeon Platinum 8168 CPUs & 8 x NVDIA V100 GPUs		Internal Dataset	1024x1024
T2I	SD2	6.12 per hour	9.7s /15s (100 iterations)	1.649	2.55	6 x Intel Xeon E5-2690 v4 CPUs & 2 x NVDIA V100	32 x 8 x A100 GPUs	LAION-5B	768x768
T2I	SD1.5	6.12 per hour	9.26s	1.5742	-	6 x Intel Xeon E5-2690 v4 CPUs & 2 x NVDIA V100	32 x 8 x A100 GPUs	LAION-2B (en)	512x512
T2I	SD1.4	6.12 per hour	9.25s	1.5725	-	6 x Intel Xeon E5-2690 v4 CPUs & 2 x NVDIA V100	32 x 8 x A100 GPUs	LAION-2B (en)	512x512
	SD3 Large	0.065/EA	5.9s	6.5		-	-		1024x1024
	SD3 Medium	0.035/EA	4.6s	3.6	-	-	-		1024x1024
I2I	SD2-Inpainting	6.12 per hour	8.3s	1.411		6 x Intel Xeon E5-2690 v4 CPUs & 2 x NVDIA V100	32 x 8 x A100 GPUs	LAION-5B	512x512
I2I	SDXL-refiner	6.12 per hour	13.5s			6 x Intel Xeon E5-2690 v4 CPUs & 2 x NVDIA V100	32 x 8 x A100 GPUs	LAION-5B	1024x1024
I2I	SD1 Inpainting	6.12 per hour	4.5s	0.765		6 x Intel Xeon E5-2690 v4 CPUs & 2 x NVDIA V100	32 x 8 x A100 GPUs	LAION-2B (en)	
I2I	SD3 Large Refiner	0.065/EA	7.6s	0.765	-	-	-		1024x1024

The methodology will consist of 3 parts. In the first part, we will test T2I models, I2I models and ensemble models by generating a bunch of images and evaluating image using human evaluations. For the second part, we will propose a framework to guide users generating images for marketing purposes. As for the third part, we will apply the proposed framework to generate a series of images based on customer segmentation and assess whether these images have satisfied the photography principles from marketing team.

3.2 Methodology for testing Prompts:

To test what prompts will be the most suitable for our project, we will test different types of prompts in our report before passing them to different diffusion models.

In this project, there will be 6 types of prompts to be tested on SD2 and DALL-E 3. We will firstly choose the brief description of desired image output as our main prompt. Then we will add keywords to the main prompt at different positions hence generating other 3 types of prompts: “keywords in front”, “keywords at the end” and “main prompt in the middle”. Further, we will expand the main prompt with keywords to a long prompt for testing model performances. After testing these five types of prompts, we will have an idea on what type of them is generally better. Based on this type of prompts, we will further change the number of denoising steps to see if they will affect image qualities. As for SD2 model, we will further add negative prompt to the chosen prompt to see if the image quality will be further improved.

3.3 Methodology for testing T2I and I2I models:

3.3.1 Testing T2I Models:

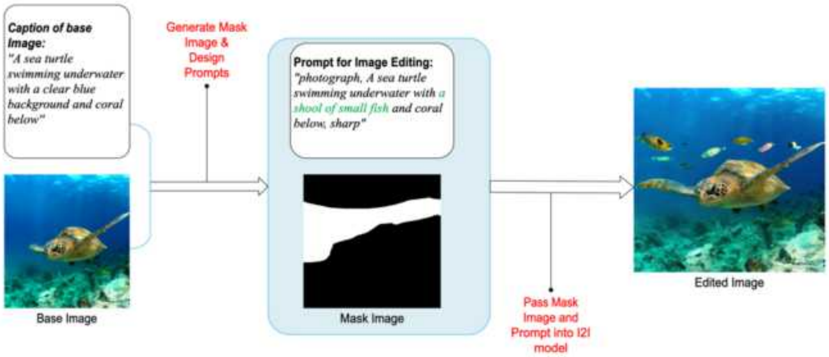
A series of prompts in terms of landscape, humans, city, illustrations, counting and words will be designed before testing T2I models. The idea of testing T2I models is that each time we will fix one prompt and pass it into different models. Once we used all prompts, we will evaluate these images using evaluation metrics. Based on these evaluations, a summary table that describes model performances will be given to help future users to choose suitable image generation models

3.3.2 Testing I2I Models:

Testing I2I models could be editing real stock images for reusage and editing synthetic images for restoring artifacts. In this report, we will test available I2I models for both scenarios.

3.3.2.1 Testing Inpainting Models

Using inpainting models generally require a prompt that instructs the model how to edit and a mask image that specifies the area that needs to edit. The following Flowchart demonstrates the process of utilising inpainting model for image editing.


Caption of base Image:
"A sea turtle swimming underwater with a clear blue background and coral below"
Generate Mask Image & Design Prompts
Prompt for Image Editing:
"photograph, A sea turtle swimming underwater with a shool of small fish and coral below, sharp"
Base Image
Mask Image
Pass Mask Image and Prompt into I2I model
Edited Image

To test inpainting models, we will firstly select multiple images from both Company A’s stock images and synthetic images produced by T2I models. Then we will draw corresponding mask images and design prompt for inpainting tasks before passing selected images into inpainting models.

There are a number of methods to produce mask images for inpainting tasks. Kirillov et al. (2023) proposed Segment Anything (SA) Model that can automatically generate mask images in a precise and accurate manner. Users now can access to SA via roboflow[] with Google colab. Alternatively, users can also generate mask images manually via Image Masker []. Since we have not tested the robustness of SA model empirically and due to limited time, we use Image Masker to generate mask images instead.

[https://colab.research.google.com/github/roboflow-ai/notebooks/blob/main/notebooks/how-to-segment-images-with-sam-2.ipynb]

[https://imagemasker.github.io]

Captions will be given to stock images as their prompts while synthetic images will reuse original prompts. To design prompt for inpainting tasks, we will slightly change part of original prompt while maintaining most of them to check if I2I models preserve certain capabilities. After generation task being done, an evaluation will also be given.

3.3.2.2 Testing Refiners

Unlike inpainting models which can edit part of images while remain other parts untouched, the I2I refiners will slightly change the whole image in editing while no mask image is required. Deploying Refiners generally requires a base image and a prompt which is edited on the original image caption. To test refiners, we will choose original stock images and then design prompts to see if these stock images can be edited by I2I refiners. Evaluations will also be given after editing all required images.

[diagram here]

3.3.3 Testing T2I and I2I Ensemble Models

Photographers sometimes edit their images multiple times to improve overall quality. Inspired by this real-world image production process, we wish to test if generating images using T2I models and editing them using I2I models will enhance the quality of synthetic images. In this section, we will test the quality of both stock images and T2I images by passing them to I2I inpainting models, I2I refiners and both inpainting and refiner models.

3.3.3.1 T2I with I2I Refiner:

In this section, we will test if the overall image quality will be improved by passing images generated from T2I models to SDXL Refiner and SD3L Refiner. Then we will use pairwise human evaluation to assess overall quality of assessed images.

[diagram here]

3.3.3.2 T2I with both I2I Inpainting Models & Refiners:

Given the presence of minor artifacts in some T2I synthetic images, we propose using inpainting models to modify these imperfections before refining images. Thus, images edited in section 3.3.2.1 will be processed through refiners to evaluate any improvement in overall quality.

[flowchart here]

3.3 Methodology for Framework:

Combining all results from section 3.1 and 3.2, we will form our framework that can guide future users generating images. The framework begins by determining whether the task is image generation or image editing. Based on the outcomes from section 3.3, we will select appropriate models and evaluation metrics to synthesize images. After generating the images, we will repeatedly assess and correct any artifacts until no significant flaws remain. Finally, images will be refined and presented to the marketing team for approval. If they are not satisfied, the process will be repeated until the desired quality is achieved.

3.4 Application of Framework:

To test our proposed framework, we will use customer segmentation in travel agency advertisements as an example. First, we will create prompts that depict people of various age groups enjoying the beach. Using these prompts, we will generate images with T2I models. Next, we will edit and refine the images to address any flaws or artifacts. Finally, we will evaluate the image quality using specific metrics to ensure they meet the marketing team's standards.

4. Experimentation

4.1 Experimentation on Keywords & Prompts Style:

4.1.1 Testing Keywords:

Based on the marketing team's general photo principles, we have identified potentially useful keywords and will test them using the SDXL model. We are not testing these keywords on DALL-E 3 because our literature review indicates that DALL-E 3 has a robust language encoder for Text-to-Image generation. Therefore, we assume DALL-E 3 can interpret the keywords effectively and focus our testing on Stable Diffusion models.

To evaluate the model's response to these keywords, we will use each keyword as a single prompt and analyse the output images. This will help us identify useful keywords for image generation in SDXL.

Table 6 presents a selection of keywords categorised by photography principles and found to be effective in SDXL. The full table is available in the Appendix. These keywords will be used to design prompts for our subsequent experiments.

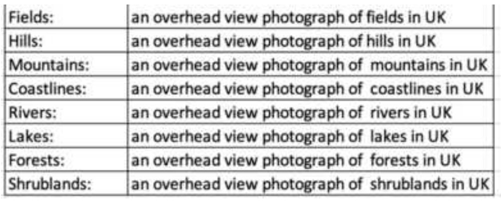
[keywords table]

Since Stable Diffusion models share similar language encoders, we assume these keywords will also be effective in other versions of the Stable Diffusion model family. However, we acknowledge that our assumptions regarding DALL-E 3 and other SD models may be incorrect. Due to time constraints, more comprehensive keyword experimentation is reserved for future research.

4.1.2 Testing Prompts:

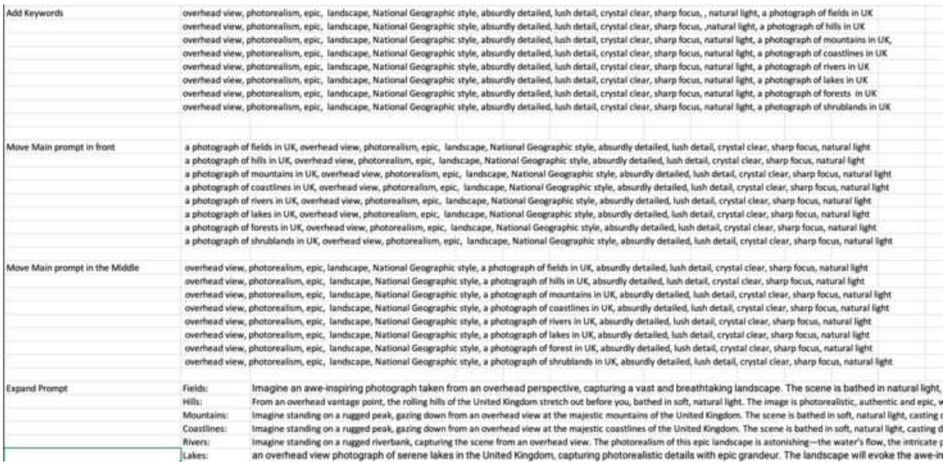
In this section, we use DALL-E 3 on Azure OpenAI Studio in a standard setting and SD2 from Azure Machine Learning.

The photography briefing used requires freelancers to produce a set of photographs depicting UK landscapes. Bearing this in mind, we have generated 8 prompts in the following as the main prompts:

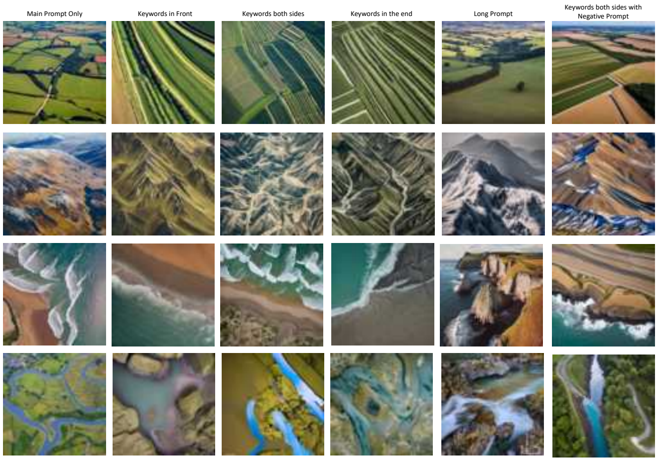


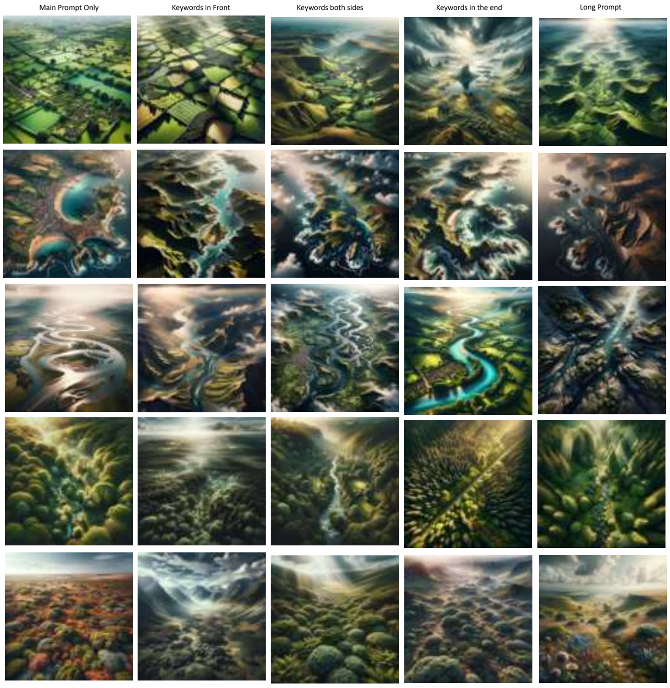
Fields:	an overhead view photograph of fields in UK
Hills:	an overhead view photograph of hills in UK
Mountains:	an overhead view photograph of mountains in UK
Coastlines:	an overhead view photograph of coastlines in UK
Rivers:	an overhead view photograph of rivers in UK
Lakes:	an overhead view photograph of lakes in UK
Forests:	an overhead view photograph of forests in UK
Shrublands:	an overhead view photograph of shrublands in UK

Then we add keywords at different positions, add negative prompts and expand the prompts to get other 6 types of prompts to be tested.



Add Keywords	overhead view, photorealism, epic, landscape, National Geographic style, absurdly detailed, lush detail, crystal clear, sharp focus, , natural light, a photograph of fields in UK	
	overhead view, photorealism, epic, landscape, National Geographic style, absurdly detailed, lush detail, crystal clear, sharp focus, ,natural light, a photograph of hills in UK	
	overhead view, photorealism, epic, landscape, National Geographic style, absurdly detailed, lush detail, crystal clear, sharp focus, natural light, a photograph of mountains in UK,	
	overhead view, photorealism, epic, landscape, National Geographic style, absurdly detailed, lush detail, crystal clear, sharp focus, natural light, a photograph of coastlines in UK	
	overhead view, photorealism, epic, landscape, National Geographic style, absurdly detailed, lush detail, crystal clear, sharp focus, natural light, a photograph of rivers in UK	
	overhead view, photorealism, epic, landscape, National Geographic style, absurdly detailed, lush detail, crystal clear, sharp focus, natural light, a photograph of lakes in UK	
	overhead view, photorealism, epic, landscape, National Geographic style, absurdly detailed, lush detail, crystal clear, sharp focus, natural light, a photograph of forests in UK	
	overhead view, photorealism, epic, landscape, National Geographic style, absurdly detailed, lush detail, crystal clear, sharp focus, natural light, a photograph of shrublands in UK	
Move Main prompt in front	a photograph of fields in UK, overhead view, photorealism, epic, landscape, National Geographic style, absurdly detailed, lush detail, crystal clear, sharp focus, natural light	
	a photograph of hills in UK, overhead view, photorealism, epic, landscape, National Geographic style, absurdly detailed, lush detail, crystal clear, sharp focus, natural light	
	a photograph of mountains in UK, overhead view, photorealism, epic, landscape, National Geographic style, absurdly detailed, lush detail, crystal clear, sharp focus, natural light	
	a photograph of coastlines in UK, overhead view, photorealism, epic, landscape, National Geographic style, absurdly detailed, lush detail, crystal clear, sharp focus, natural light	
	a photograph of rivers in UK, overhead view, photorealism, epic, landscape, National Geographic style, absurdly detailed, lush detail, crystal clear, sharp focus, natural light	
	a photograph of lakes in UK, overhead view, photorealism, epic, landscape, National Geographic style, absurdly detailed, lush detail, crystal clear, sharp focus, natural light	
	a photograph of forests in UK, overhead view, photorealism, epic, landscape, National Geographic style, absurdly detailed, lush detail, crystal clear, sharp focus, natural light	
	a photograph of shrublands in UK, overhead view, photorealism, epic, landscape, National Geographic style, absurdly detailed, lush detail, crystal clear, sharp focus, natural light	
Move Main prompt in the Middle	overhead view, photorealism, epic, landscape, National Geographic style, a photograph of fields in UK, absurdly detailed, lush detail, crystal clear, sharp focus, natural light	
	overhead view, photorealism, epic, landscape, National Geographic style, a photograph of hills in UK, absurdly detailed, lush detail, crystal clear, sharp focus, natural light	
	overhead view, photorealism, epic, landscape, National Geographic style, a photograph of mountains in UK, absurdly detailed, lush detail, crystal clear, sharp focus, natural light	
	overhead view, photorealism, epic, landscape, National Geographic style, a photograph of coastlines in UK, absurdly detailed, lush detail, crystal clear, sharp focus, natural light	
	overhead view, photorealism, epic, landscape, National Geographic style, a photograph of rivers in UK, absurdly detailed, lush detail, crystal clear, sharp focus, natural light	
	overhead view, photorealism, epic, landscape, National Geographic style, a photograph of lakes in UK, absurdly detailed, lush detail, crystal clear, sharp focus, natural light	
	overhead view, photorealism, epic, landscape, National Geographic style, a photograph of forest in UK, absurdly detailed, lush detail, crystal clear, sharp focus, natural light	
	overhead view, photorealism, epic, landscape, National Geographic style, a photograph of shrublands in UK, absurdly detailed, lush detail, crystal clear, sharp focus, natural light	
Expand Prompt	Fields:	Imagine an awe-inspiring photograph taken from an overhead perspective, capturing a vast and breathtaking landscape. The scene is bathed in natural light,
	Hills:	From an overhead vantage point, the rolling hills of the United Kingdom stretch out before you, bathed in soft, natural light. The image is photorealistic, authentic and epic,
	Mountains:	Imagine standing on a rugged peak, gazing down from an overhead view at the majestic mountains of the United Kingdom. The scene is bathed in soft, natural light, casting
	Coastlines:	Imagine standing on a rugged peak, gazing down from an overhead view at the majestic coastlines of the United Kingdom. The scene is bathed in soft, natural light, casting
	Rivers:	Imagine standing on a rugged riverbank, capturing the scene from an overhead view. The photorealism of this epic landscape is astonishing—the water's flow, the intricate
	Lakes:	an overhead view photograph of serene lakes in the United Kingdom, capturing photorealistic details with epic grandeur. The landscape will evoke the awe-in

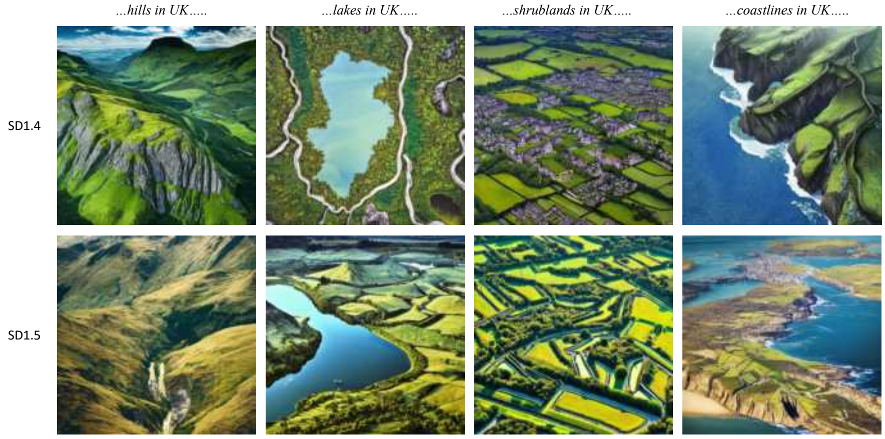




In SD2, we found using long prompt or using keywords with negative prompt are better than other types of prompts, while in DALL-E 3, we also found using long prompt can marginally generate better images. When it comes to other types of prompts being tested, no significant differences are perceived by the author. Moreover, since designing long prompts are time consuming and demanding, we will design our prompt by adding keywords both sides and will add negative prompts for SD model family where possible for subsequent experimentations.

4.1.3Testing SD1.4 & SD1.5:

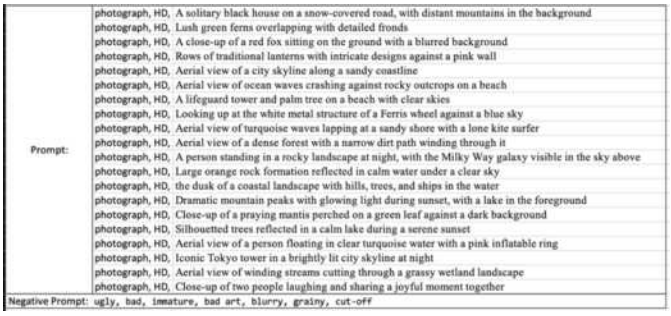
When testing landscape generation empirically, we found SD1.4 and SD 1.5 have unsatisfied performances. This is because they could only generate images with low resolutions (512x) and low authenticity. In the Figure 2 below, we further investigate that both SD1.4 and SD1.5 fail to synthesise shrublands images. Therefore, we will filter out SD1.4 and SD1.5 at this stage and they will no longer be used in the subsequent model comparison.



4.1.4 Testing CLIP-IQA:

To test CLIP-IQA, we firstly select 20 real photographs from Unsplash that are downloaded at least 6000 times. We use download times as a basis for image selection because this is a plausible indicator for assessing whether images are of good quality. Then, we give each image a caption and further craft each caption to a properly designed prompt in the following table xxx:

[https://unsplash.com]



Prompt:	photograph, HD, A solitary black house on a snow-covered road, with distant mountains in the background
	photograph, HD, Lush green ferns overlapping with detailed fronds
	photograph, HD, A close-up of a red fox sitting on the ground with a blurred background
	photograph, HD, Rows of traditional lanterns with intricate designs against a pink wall
	photograph, HD, Aerial view of a city skyline along a sandy coastline
	photograph, HD, Aerial view of ocean waves crashing against rocky outcrops on a beach
	photograph, HD, A lifeguard tower and palm tree on a beach with clear skies
	photograph, HD, Looking up at the white metal structure of a Ferris wheel against a blue sky
	photograph, HD, Aerial view of turquoise waves lapping at a sandy shore with a lone kite surfer
	photograph, HD, Aerial view of a dense forest with a narrow dirt path winding through it
	photograph, HD, A person standing in a rocky landscape at night, with the Milky Way galaxy visible in the sky above
	photograph, HD, Large orange rock formation reflected in calm water under a clear sky
	photograph, HD, the dusk of a coastal landscape with hills, trees, and ships in the water
	photograph, HD, Dramatic mountain peaks with glowing light during sunset, with a lake in the foreground
	photograph, HD, Close-up of a praying mantis perched on a green leaf against a dark background
	photograph, HD, Silhouetted trees reflected in a calm lake during a serene sunset
	photograph, HD, Aerial view of a person floating in clear turquoise water with a pink inflatable ring
	photograph, HD, Iconic Tokyo tower in a brightly lit city skyline at night
	photograph, HD, Aerial view of winding streams cutting through a grassy wetland landscape
	photograph, HD, Close-up of two people laughing and sharing a joyful moment together
Negative Prompt: ugly, bad, immature, bad art, blurry, grainy, cut-off	

Then we input these prompts and their corresponding original images into a relatively older model called SDXL Refiner to generate another 20 images that are similar to the original ones. Images generated in this section is attached in the Appendix xxx. Later, we evaluate original images and generated images using overall quality and naturalness aspects in CLIP-IQA to see if the results of metrics are consistent.

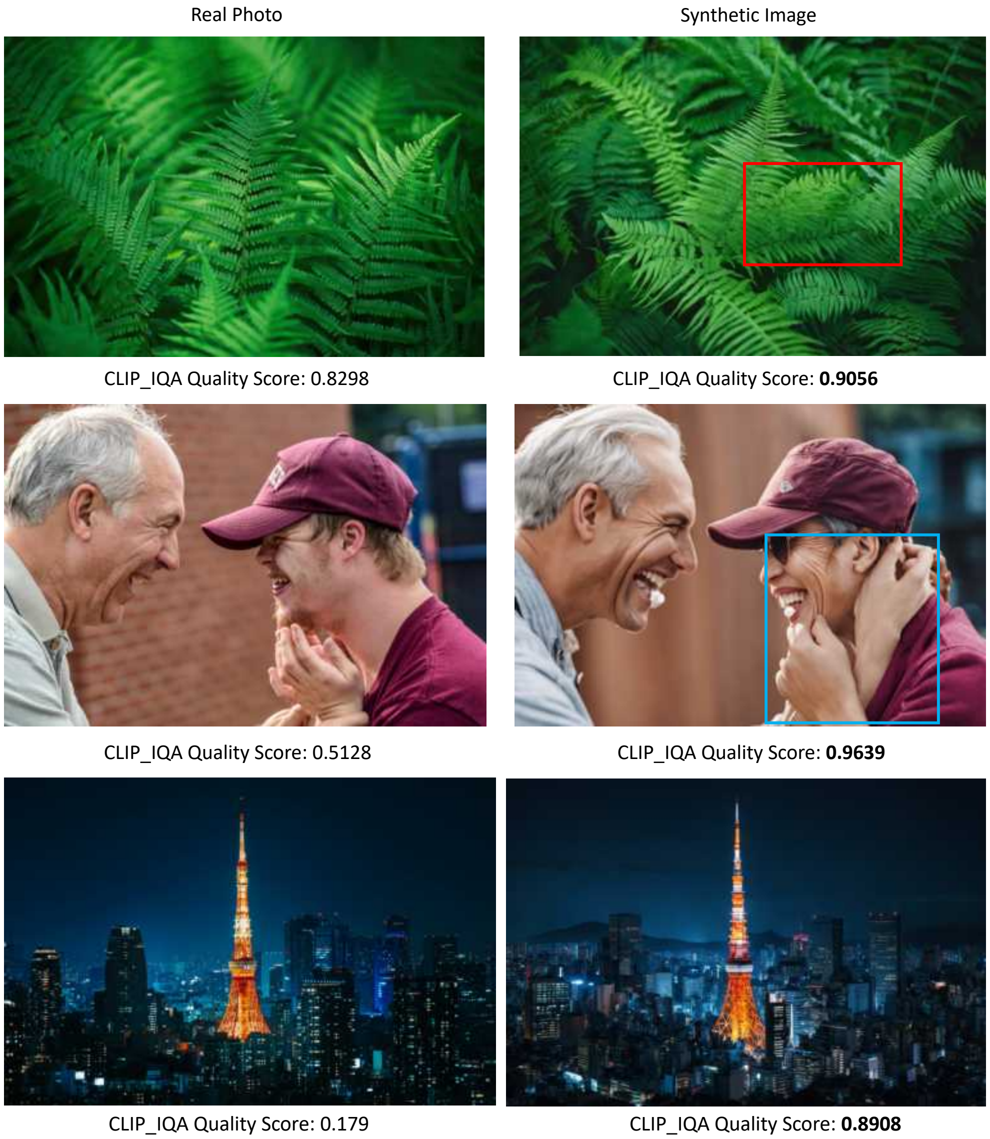
The idea of the experiment design is that if the CLIP-IQA is useful, it should assign photos a higher authenticity score because they are indeed photos produced by real photographers. Moreover, if generated images are given a higher quality score, it should free from flaws compared to real photos.

However, our experiment suggests that CLIP-IQA is not stable in assessing quality and authenticity of images. In the table xxx below, 40% of generated images that produced by older version of SD model have higher photo-realism than real photos, which contradicts to our assumptions.

	Overall Quality		Authenticity	
	Real Photo	Synthetic Image	Real Photo	Synthetic Image
Pair 1	0.6248	**0.8185**	0.9208	**0.937**
Pair 2	0.8298	**0.9056**	**0.941**	0.8503
Pair 3	0.669	**0.8757**	0.8993	**0.9444**
Pair 4	**0.7732**	0.739	**0.5921**	0.2025
Pair 5	0.6965	**0.8416**	**0.8479**	0.8416
Pair 6	0.9085	**0.972**	0.8388	**0.8639**
Pair 7	0.8233	**0.9369**	**0.9519**	0.946
Pair 8	0.5628	**0.8255**	**0.6436**	0.6327
Pair 9	0.7495	**0.9573**	0.8611	**0.8695**
Pair 10	0.6815	**0.9564**	**0.9774**	0.8403
Pair 11	0.535	**0.9734**	0.6478	**0.6652**
Pair 12	0.6657	**0.9081**	0.8388	**0.8597**
Pair 13	0.4994	**0.8713**	**0.9043**	0.684
Pair 14	0.7498	**0.7896**	**0.9237**	0.8691
Pair 15	0.4727	**0.9373**	0.926	**0.9461**
Pair 16	0.6327	**0.8671**	**0.8443**	0.8305
Pair 17	0.8555	**0.9276**	0.8679	**0.8955**
Pair 18	0.179	**0.8908**	**0.6337**	0.5134
Pair 19	0.8404	**0.9906**	**0.9582**	0.9223
Pair 20	0.5128	**0.9639**	**0.9056**	0.842
Count	1	19	12	8

In addition, table xxx indicates that 95% of generated images have higher quality score than highly downloaded stock photos. However, we noticed many inconsistencies through our manual image investigations.

The following figure xxxx demonstrates 3 image pairs containing real photos and synthetic images. In each pair, the generated image has higher quality score as well as tangible flaws. For the first row of images, the generated image on the right has higher CLIP-IQA quality score, but there is perceivable artifact of fern leaves in the red box. Similarly, for the second image pair in the second row, we observed unclear hands being generated in the blue box, although the synthetic image has much higher quality score compared to original photo. When it comes to third image pair, the original photo on the left only achieved 0.179 in quality score, which is far less than the generated image whose details of Tokyo Tower relatively vague compared to the real photo on the left. Therefore, we may conclude CLIP-IQA is not stable in terms of overall quality neither.



Real Photo

Synthetic Image

CLIP_IQA Quality Score: 0.8298

CLIP_IQA Quality Score: **0.9056**

CLIP_IQA Quality Score: 0.5128

CLIP_IQA Quality Score: **0.9639**

CLIP_IQA Quality Score: 0.179

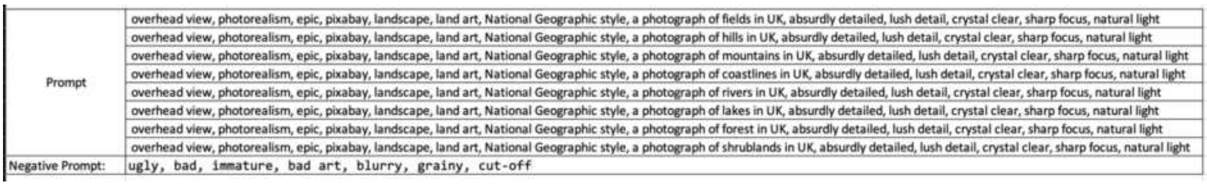
CLIP_IQA Quality Score: **0.8908**

Hence, we will only use human evaluation in our subsequent experimentations.

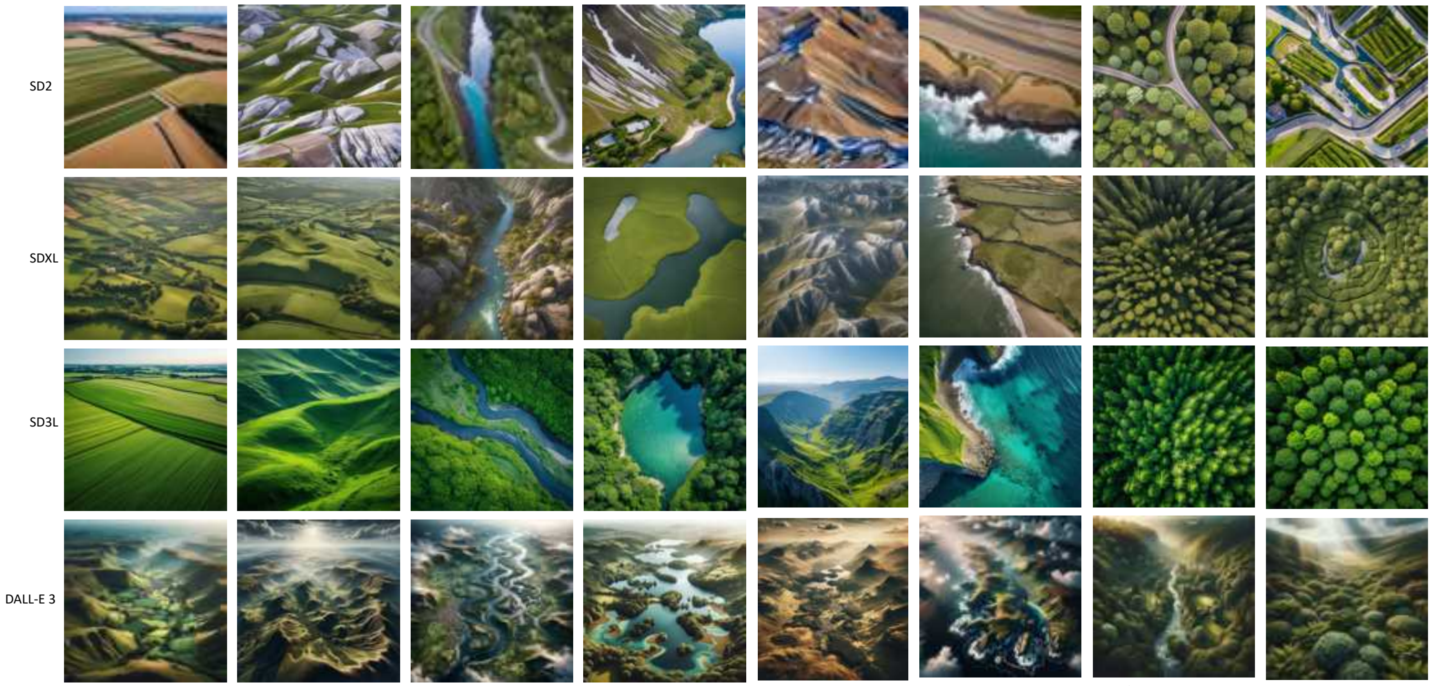
4.2 Experimentation for T2Is

4.2.1 Experimentation on Landscapes:

We firstly design 8 main prompts pertaining to landscapes such as fields, coastlines and rivers of United Kingdom. Then we craft these prompts by adding keywords both sides and display them in the following table xxxx:



Prompt	overhead view, photorealism, epic, pixabay, landscape, land art, National Geographic style, a photograph of fields in UK, absurdly detailed, lush detail, crystal clear, sharp focus, natural light
	overhead view, photorealism, epic, pixabay, landscape, land art, National Geographic style, a photograph of hills in UK, absurdly detailed, lush detail, crystal clear, sharp focus, natural light
	overhead view, photorealism, epic, pixabay, landscape, land art, National Geographic style, a photograph of mountains in UK, absurdly detailed, lush detail, crystal clear, sharp focus, natural light
	overhead view, photorealism, epic, pixabay, landscape, land art, National Geographic style, a photograph of coastlines in UK, absurdly detailed, lush detail, crystal clear, sharp focus, natural light
	overhead view, photorealism, epic, pixabay, landscape, land art, National Geographic style, a photograph of rivers in UK, absurdly detailed, lush detail, crystal clear, sharp focus, natural light
	overhead view, photorealism, epic, pixabay, landscape, land art, National Geographic style, a photograph of lakes in UK, absurdly detailed, lush detail, crystal clear, sharp focus, natural light
	overhead view, photorealism, epic, pixabay, landscape, land art, National Geographic style, a photograph of forest in UK, absurdly detailed, lush detail, crystal clear, sharp focus, natural light
	overhead view, photorealism, epic, pixabay, landscape, land art, National Geographic style, a photograph of shrublands in UK, absurdly detailed, lush detail, crystal clear, sharp focus, natural light
Negative Prompt:	`ugly, bad, immature, bad art, blurry, grainy, cut-off`



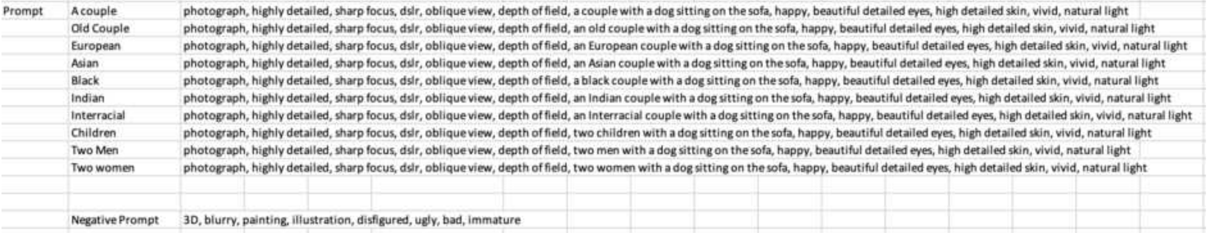
...fields in UK... ...hills in UK... ...rivers in UK... ...lakes in UK... ...mountains in UK... ...coastlines in UK... ...forests in UK... ...shrublands in UK...

According to the author's assessment, images produced by DALL-E 3 appear less authentic and more illustrative, yet they exhibit high aesthetic quality. Conversely, images from SD2 contain artifacts and the model often fails to accurately render shrublands, with outputs not consistently aligning with the textual prompts.

Images from SDXL, while generally appealing, do not achieve a high degree of photorealism. Meanwhile, images generated by SD3L show an improvement in quality, although this model struggles with accurately depicting shrublands in photographs as well.

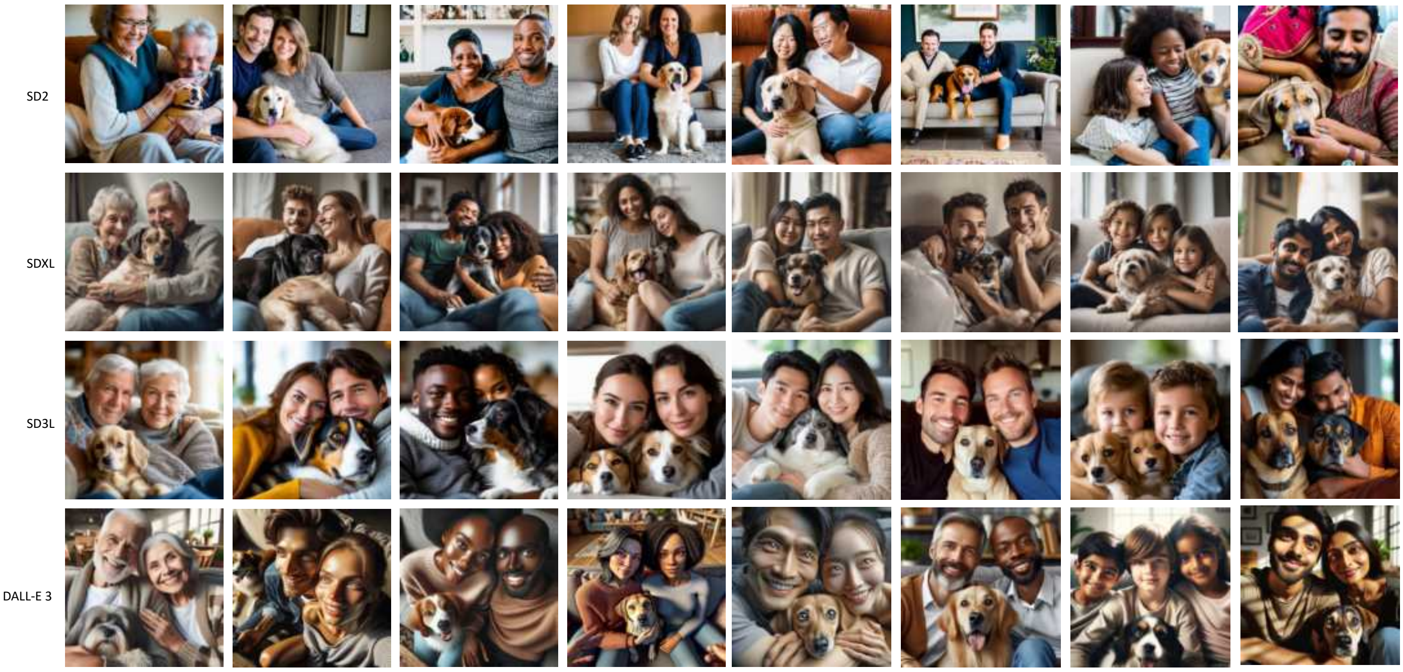
4.2.2 Experimentation on Generating Faces:

To evaluate the capability of available Text-to-Image (T2I) models in generating photographs featuring people, we devised 10 main prompts. These prompts specifically describe a couple, distinguished by ethnicity and age, accompanied by a dog seated on a sofa. Additionally, we incorporated keywords on either side of the main prompts, as detailed in the subsequent table:



Prompt	A couple	photograph, highly detailed, sharp focus, dslr, oblique view, depth of field, a couple with a dog sitting on the sofa, happy, beautiful detailed eyes, high detailed skin, vivid, natural light
	Old Couple	photograph, highly detailed, sharp focus, dslr, oblique view, depth of field, an old couple with a dog sitting on the sofa, happy, beautiful detailed eyes, high detailed skin, vivid, natural light
	European	photograph, highly detailed, sharp focus, dslr, oblique view, depth of field, an European couple with a dog sitting on the sofa, happy, beautiful detailed eyes, high detailed skin, vivid, natural light
	Asian	photograph, highly detailed, sharp focus, dslr, oblique view, depth of field, an Asian couple with a dog sitting on the sofa, happy, beautiful detailed eyes, high detailed skin, vivid, natural light
	Black	photograph, highly detailed, sharp focus, dslr, oblique view, depth of field, a black couple with a dog sitting on the sofa, happy, beautiful detailed eyes, high detailed skin, vivid, natural light
	Indian	photograph, highly detailed, sharp focus, dslr, oblique view, depth of field, an Indian couple with a dog sitting on the sofa, happy, beautiful detailed eyes, high detailed skin, vivid, natural light
	Interracial	photograph, highly detailed, sharp focus, dslr, oblique view, depth of field, an Interracial couple with a dog sitting on the sofa, happy, beautiful detailed eyes, high detailed skin, vivid, natural light
	Children	photograph, highly detailed, sharp focus, dslr, oblique view, depth of field, two children with a dog sitting on the sofa, happy, beautiful detailed eyes, high detailed skin, vivid, natural light
	Two Men	photograph, highly detailed, sharp focus, dslr, oblique view, depth of field, two men with a dog sitting on the sofa, happy, beautiful detailed eyes, high detailed skin, vivid, natural light
	Two women	photograph, highly detailed, sharp focus, dslr, oblique view, depth of field, two women with a dog sitting on the sofa, happy, beautiful detailed eyes, high detailed skin, vivid, natural light
	Negative Prompt	3D, blurry, painting, illustration, disfigured, ugly, bad, immature

We submitted the designed prompts to four different T2I models—SD2, SDXL, SD3L, and DALL-E 3—to produce images depicting people as specified. The selected results of this



...an old couple with a dog sitting on the sofa...

...a European couple with a dog sitting on the sofa...

...a black couple with a dog sitting on the sofa...

...two women with a dog sitting on the sofa...

...an Asain couple with a dog sitting on the sofa...

...two men with a dog sitting on the sofa...

...two children with a dog sitting on the sofa...

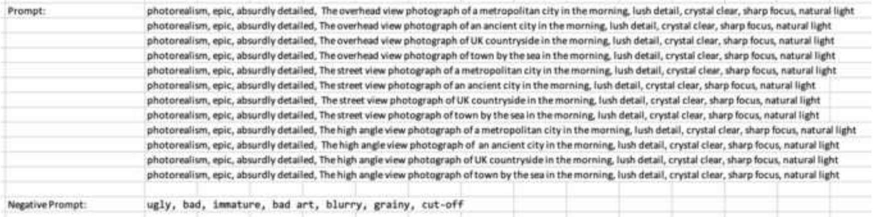
...an Indian with a dog sitting on the sofa...

Our evaluation indicates that SDXL encounters difficulties in rendering fingers accurately, while images produced by SD2 lack authenticity and contains distinct artifacts. DALL-E 3 tends to generate outputs that resemble illustrations, which also compromises their authenticity. Although SD3L produces comparatively better results, it is important to note that it operates as an internet demo via fireworks.ai, which might not fully represent its capabilities. In addition, SD2 fails to generate photograph of Indian couple, while both DALL-E 3 and SDXL did not generate images of children in correct numbers, indicating they may suffer from generating correct number of objects.

Furthermore, there is a noticeable bias within these T2I models, where they often default to representations of conventionally attractive individuals as the norm. This could potentially fail to meet the diversity and inclusion standards set by Company A, as the models may not adequately represent a broad spectrum of appearances and attributes.

4.2.3 Experimentation on City:

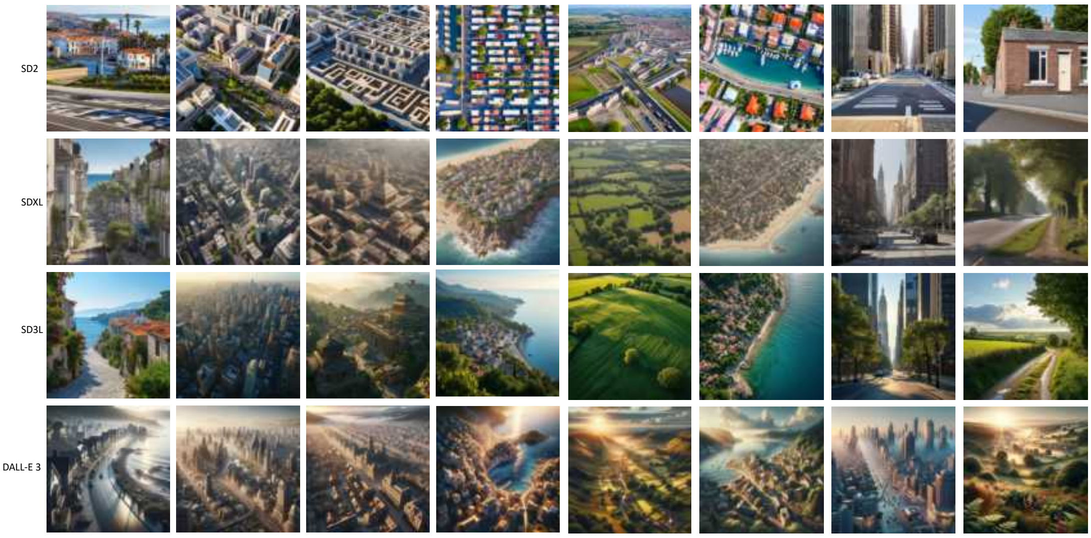
To assess the proficiency of our T2I models in generating cities, we tasked them with producing images of a metropolitan city, an ancient city, the UK countryside, and a seaside town, each from overhead, street, and high-angle views. This led to the creation of 12 specific prompts, the results of which have been collated into 48 images, with a selection showcased in Table XX.



Prompt:	photorealism, epic, absurdly detailed, The overhead view photograph of a metropolitan city in the morning, lush detail, crystal clear, sharp focus, natural light
	photorealism, epic, absurdly detailed, The overhead view photograph of an ancient city in the morning, lush detail, crystal clear, sharp focus, natural light
	photorealism, epic, absurdly detailed, The overhead view photograph of UK countryside in the morning, lush detail, crystal clear, sharp focus, natural light
	photorealism, epic, absurdly detailed, The overhead view photograph of town by the sea in the morning, lush detail, crystal clear, sharp focus, natural light
	photorealism, epic, absurdly detailed, The street view photograph of a metropolitan city in the morning, lush detail, crystal clear, sharp focus, natural light
	photorealism, epic, absurdly detailed, The street view photograph of an ancient city in the morning, lush detail, crystal clear, sharp focus, natural light
	photorealism, epic, absurdly detailed, The street view photograph of UK countryside in the morning, lush detail, crystal clear, sharp focus, natural light
	photorealism, epic, absurdly detailed, The street view photograph of town by the sea in the morning, lush detail, crystal clear, sharp focus, natural light
	photorealism, epic, absurdly detailed, The high angle view photograph of a metropolitan city in the morning, lush detail, crystal clear, sharp focus, natural light
	photorealism, epic, absurdly detailed, The high angle view photograph of an ancient city in the morning, lush detail, crystal clear, sharp focus, natural light
	photorealism, epic, absurdly detailed, The high angle view photograph of UK countryside in the morning, lush detail, crystal clear, sharp focus, natural light
	photorealism, epic, absurdly detailed, The high angle view photograph of town by the sea in the morning, lush detail, crystal clear, sharp focus, natural light
Negative Prompt:	ugly, bad, immature, bad art, blurry, grainy, cut-off

Our preliminary analysis revealed that images synthesized by SD2 consistently display noticeable artifacts. The images from SDXL, while clear, tend to lack photorealism, often resembling 3D renderings or screenshots from commercial mapping services. On the other hand, images from DALL-E 3, though highly aesthetic, appear illustrative and thus less authentic.

SD3L emerged as the most effective model for depicting urban scenes due to its superior photorealism, high resolution, and distinct aesthetic quality. Nonetheless, it is important to acknowledge that SD3L also exhibits some limitations, particularly in rendering high-angle views of metropolitan and ancient cities, where artifacts are still evident.



... The street view photograph of town by the sea in the morning...

... The high angle view photograph of a metropolitan city in the morning...

...The high angle view photograph of an ancient city in the morning...

...The high angle view photograph of town by the sea in the morning...

...The overhead view photograph of UK countryside in the morning...

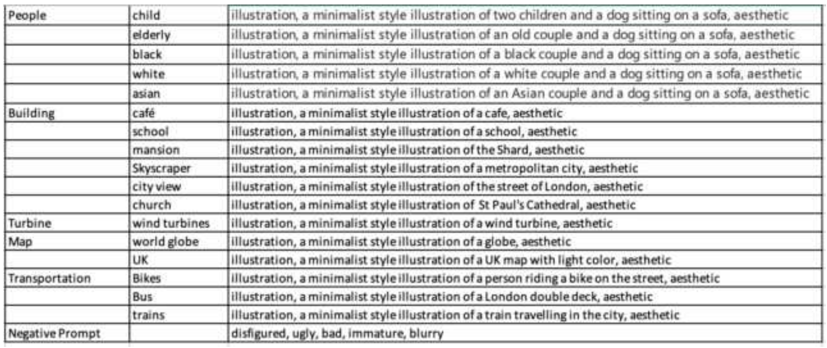
...The overhead view photograph of town by the sea in the morning...

...The street view photograph of a metropolitan city in the morning...

...The street view photograph of an ancient city in the morning...

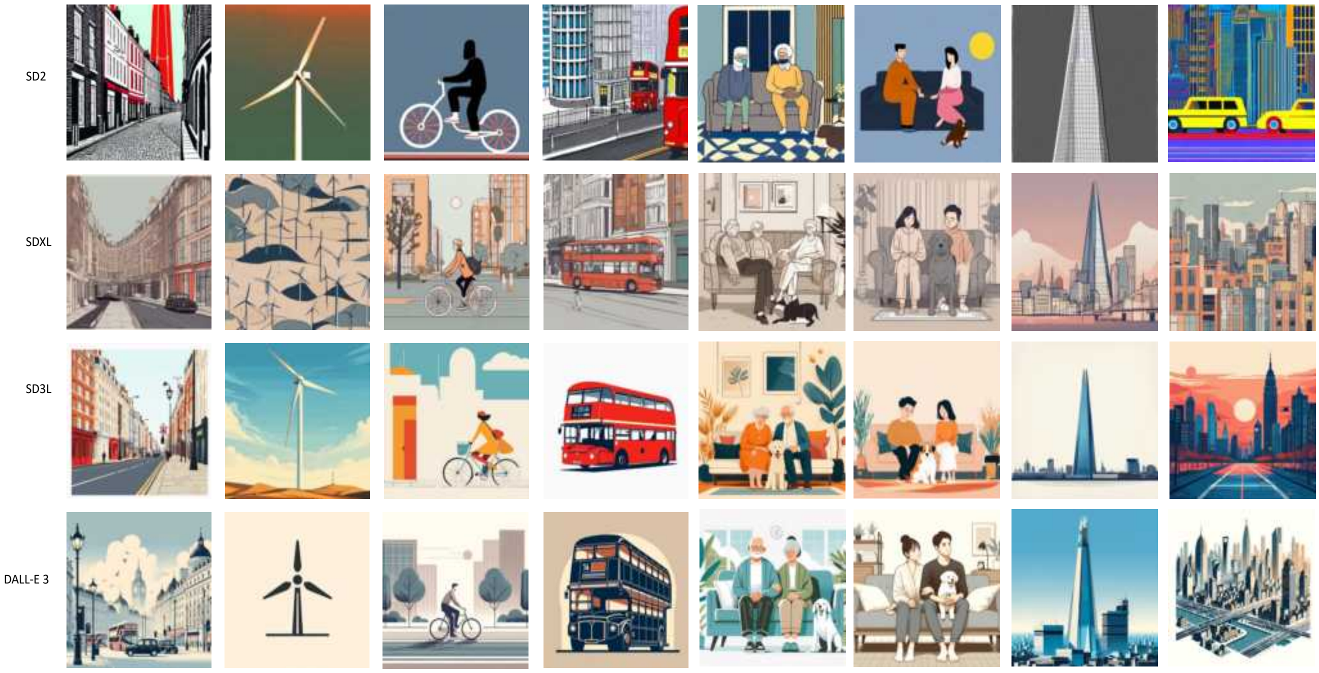
4.2.4 Experimentation on Illustration:

Similar to previous sections, we crafted 17 prompts of illustrations ranging from people, buildings and infrastructure, etc.



People	child	illustration, a minimalist style illustration of two children and a dog sitting on a sofa, aesthetic
	elderly	illustration, a minimalist style illustration of an old couple and a dog sitting on a sofa, aesthetic
	black	illustration, a minimalist style illustration of a black couple and a dog sitting on a sofa, aesthetic
	white	illustration, a minimalist style illustration of a white couple and a dog sitting on a sofa, aesthetic
	asian	illustration, a minimalist style illustration of an Asian couple and a dog sitting on a sofa, aesthetic
Building	café	illustration, a minimalist style illustration of a cafe, aesthetic
	school	illustration, a minimalist style illustration of a school, aesthetic
	mansion	illustration, a minimalist style illustration of the Shard, aesthetic
	Skyscraper	illustration, a minimalist style illustration of a metropolitan city, aesthetic
	city view	illustration, a minimalist style illustration of the street of London, aesthetic
	church	illustration, a minimalist style illustration of St Paul's Cathedral, aesthetic
Turbine	wind turbines	illustration, a minimalist style illustration of a wind turbine, aesthetic
Map	world globe	illustration, a minimalist style illustration of a globe, aesthetic
	UK	illustration, a minimalist style illustration of a UK map with light color, aesthetic
Transportation	Bikes	illustration, a minimalist style illustration of a person riding a bike on the street, aesthetic
	Bus	illustration, a minimalist style illustration of a London double deck, aesthetic
	trains	illustration, a minimalist style illustration of a train travelling in the city, aesthetic
Negative Prompt		disfigured, ugly, bad, immature, blurry

Inputting all 17 prompts into our T2I models, we generate 68 images of illustrations with some examples listed in the below figure xxx.



...illustration of the street of London...

... illustration of a wind turbine...

...illustration of a person riding a bike on the street...

... illustration of a London double deck...

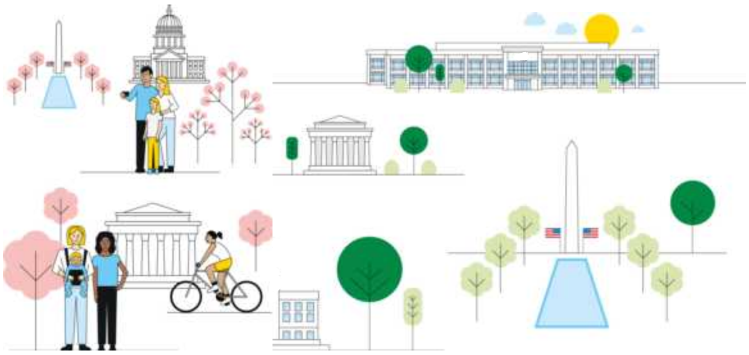
...illustration of an old couple and a dog sitting on a sofa...

...illustration of an Asian couple and a dog sitting on a sofa...

...illustration of the Shard...

...illustration of a metropolitan city...

From our investigation, it has raised our attention that the style of illustrations is difficult to control, which may undermine the usefulness of T2I models in generating illustrations for Company A. Company A usually use illustrations of consistent minimalist style with consistent colour palette in their online blog contents. The figure xxxx is the example of style of illustrations Company A prefers:



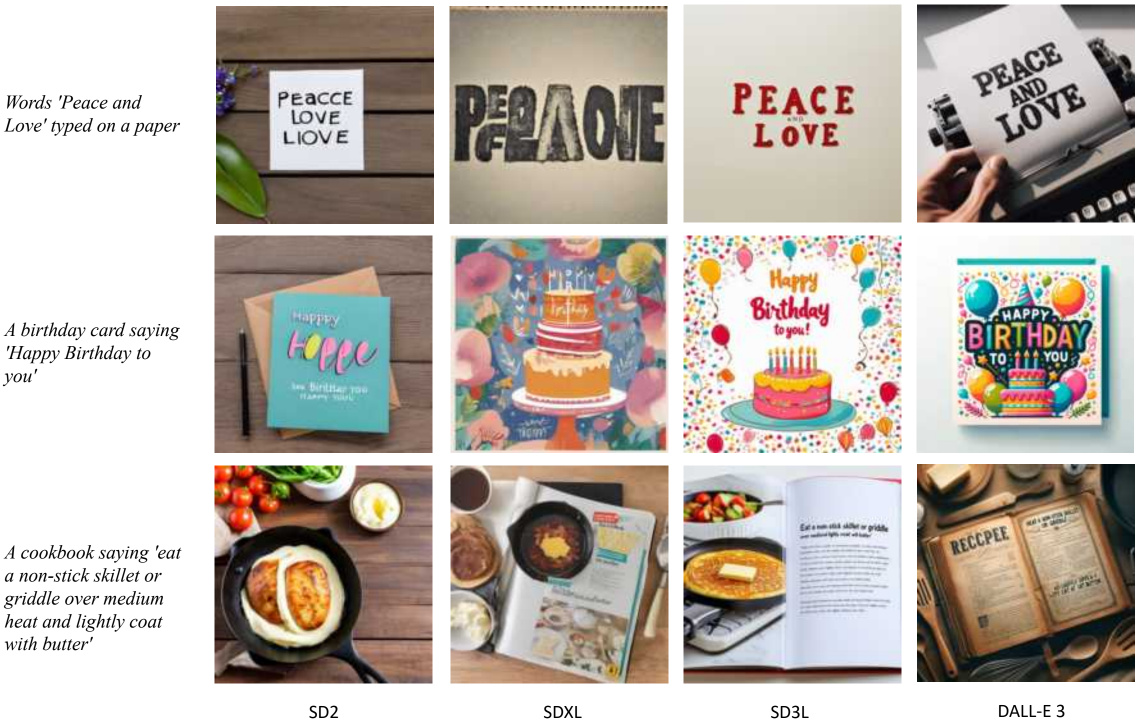
We could observe Company A prefers illustrations with minimalist style, white background, simple colours and simple lines to depict the whole image. Although we include minimalist style in our prompt, yet the style of generated images is neither consistent within each model, nor consistent with preference of Company A. Therefore, this issue has been one major problem for current T2I models.

Apart from style, there are also limitations for illustrations being generated. Overall, DALL-E 3 outperforms other models when generating illustrations because images from DALL-E 3 capture the most details compare to other models while producing minimum artifacts in each image. For images generated by SD3L, they are also of good quality, albeit minor flaws being produced. As for SDXL, illustrations synthesised may not as good as DALL-E 3 and SD3L because there is perceivable blurriness in illustrations such as turbines, London double deck and the old couple, making illustrations from SDXL look more like the sketch from artists. As for SD2, illustrations from this model may not be suitable for commercial purposes as they are similar to non-serious artworks such as cartoon arts.

4.2.5 Experimentation on Words:

One of findings from our initial model testing is that we found T2I models may struggle with generating legible words on their images. To test this capability, we designed 3 prompts which contain different lengths of words and passed them into available T2I models. The Figure xxx demonstrates the images with words in short, medium, and long lengths.

We found that only SD3L and DALL-E 3 can produce legible words in their images while SDXL and SD2 struggle the words generation. By comparing SD3L and DALL-E 3, we further noticed that only SD3L synthesised the correct textual content for words with long lengths, while DALL-E 3 add some incorrect alphabets in the text.

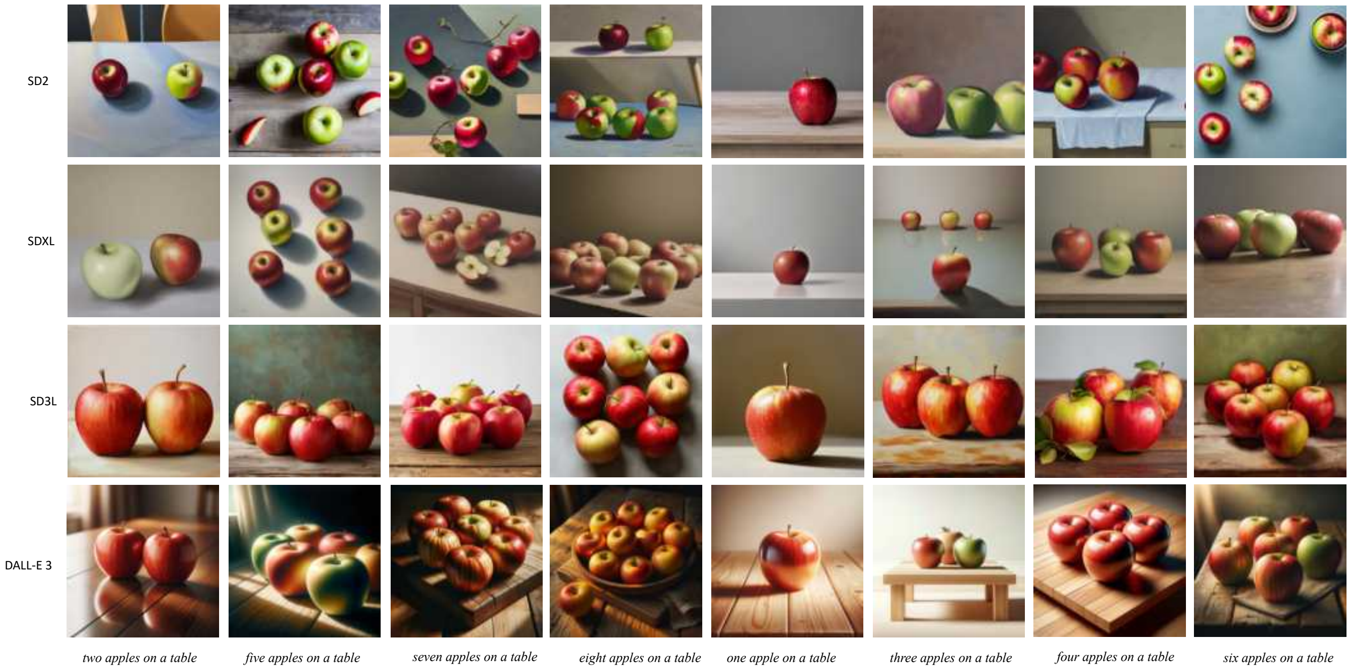

Words 'Peace and Love' typed on a paper
A birthday card saying 'Happy Birthday to you'
A cookbook saying 'eat a non-stick skillet or griddle over medium heat and lightly coat with butter'
SD2
SDXL
SD3L
DALL-E 3

Therefore, we may claim both SD3L and DALL-E 3 can generate images containing legible words, while SD3L is marginally better when it generates images with long text contents.

4.2.6 Experimentation on Counting:

We have seen that some T2I models may not generate the correct number of objects on the images. To further investigate this, we require our models to generate apple(s) from 1 to 8 using a series of prompts "…*apples on a table*" for experimentation.

The following Figure xxx is the selection of images regarding number of apples from different T2I models.

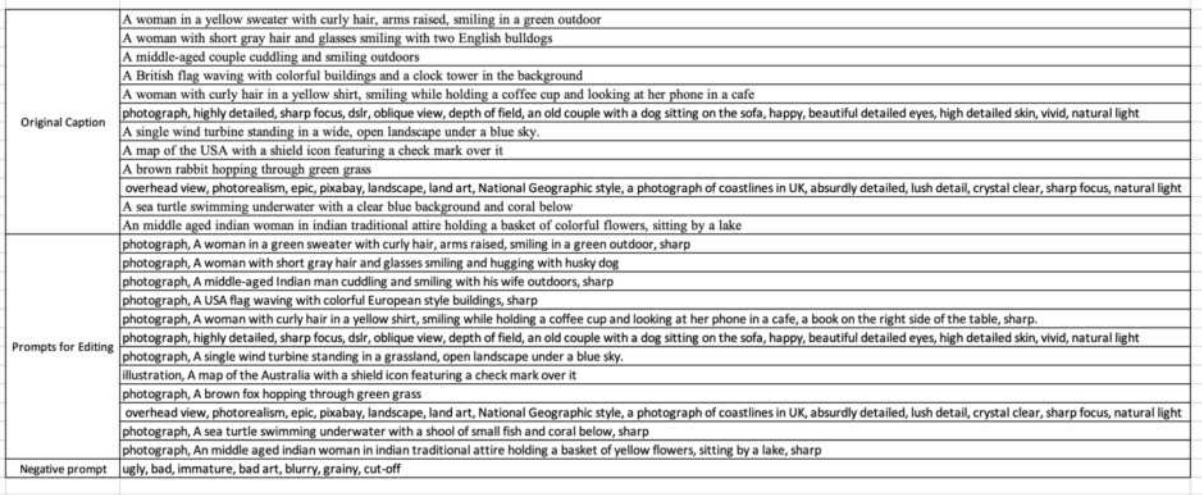


From our observations, we observed that all four T2I models are struggling with generating correct number of apples. While these T2I models can accurately generate images with one or two apples, some models are making mistakes from generating 3 apples or more. For SD2, it made mistakes in generating 5 apples as there are additional slices being included. For SDXL, they only correctly generate images with 1, 2 and 4 apples. As for SD3L, they made mistakes in generating 5, 6 and 7 apples. As for DALL-E 3, they made mistakes in generating 5, 7, and 8 apples as well. Based on these results, we may claim all four T2I models we are comparing are suffering from generating correct number of objects on the image.

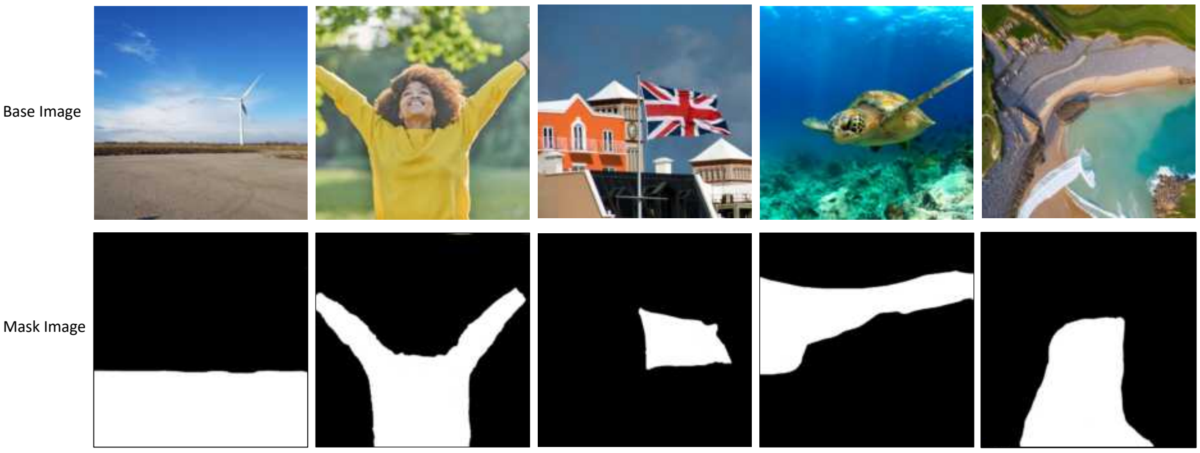
4.3 Experimentation on I2Is

4.3.1 Experimentation for Inpainting Models

To test the performances of inpainting models, we firstly select more than 10 images from Company A's stock image dataset and flawed images generated previously as the base image. Then we give each image a caption or use the prompt that synthesized that image. Later, we specify what we wish to change by generating customised prompts and mask images. Here we only add or partially change the original caption to obtain out prompts. This is to ensure that the edited part is consistent with the original image.



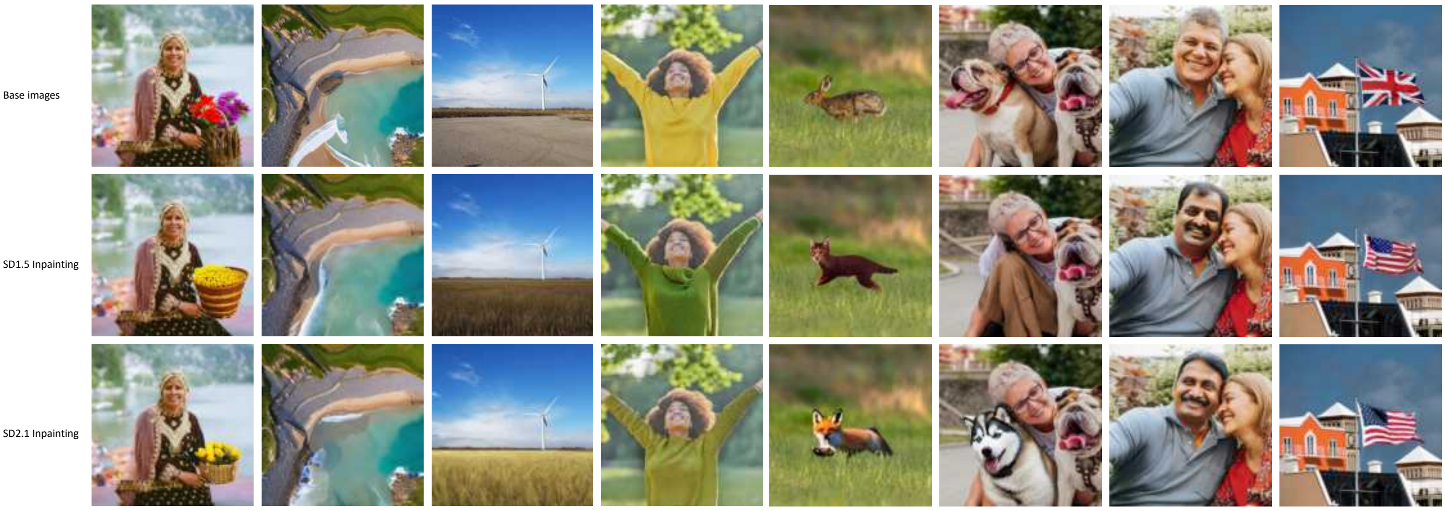
Original Caption	A woman in a yellow sweater with curly hair, arms raised, smiling in a green outdoor
	A woman with short gray hair and glasses smiling with two English bulldogs
	A middle-aged couple cuddling and smiling outdoors
	A British flag waving with colorful buildings and a clock tower in the background
	A woman with curly hair in a yellow shirt, smiling while holding a coffee cup and looking at her phone in a cafe
	photograph, highly detailed, sharp focus, dslr, oblique view, depth of field, an old couple with a dog sitting on the sofa, happy, beautiful detailed eyes, high detailed skin, vivid, natural light
	A single wind turbine standing in a wide, open landscape under a blue sky.
	A map of the USA with a shield icon featuring a check mark over it
	A brown rabbit hopping through green grass
	overhead view, photorealism, epic, pixabay, landscape, land art, National Geographic style, a photograph of coastlines in UK, absurdly detailed, lush detail, crystal clear, sharp focus, natural light
	A sea turtle swimming underwater with a clear blue background and coral below
	An middle aged indian woman in indian traditional attire holding a basket of colorful flowers, sitting by a lake
Prompts for Editing	photograph, A woman in a green sweater with curly hair, arms raised, smiling in a green outdoor, sharp
	photograph, A woman with short gray hair and glasses smiling and hugging with husky dog
	photograph, A middle-aged Indian man cuddling and smiling with his wife outdoors, sharp
	photograph, A USA flag waving with colorful European style buildings, sharp
	photograph, A woman with curly hair in a yellow shirt, smiling while holding a coffee cup and looking at her phone in a cafe, a book on the right side of the table, sharp.
	photograph, highly detailed, sharp focus, dslr, oblique view, depth of field, an old couple with a dog sitting on the sofa, happy, beautiful detailed eyes, high detailed skin, vivid, natural light
	photograph, A single wind turbine standing in a grassland, open landscape under a blue sky.
	illustration, A map of the Australia with a shield icon featuring a check mark over it
	photograph, A brown fox hopping through green grass
	overhead view, photorealism, epic, pixabay, landscape, land art, National Geographic style, a photograph of coastlines in UK, absurdly detailed, lush detail, crystal clear, sharp focus, natural light
	photograph, A sea turtle swimming underwater with a shool of small fish and coral below, sharp
	photograph, An middle aged indian woman in indian traditional attire holding a basket of yellow flowers, sitting by a lake, sharp
Negative prompt	ugly, bad, immature, bad art, blurry, grainy, cut-off



Base Image Caption	*A single wind turbine standing in a wide, open landscape under a blue sky.*	*A woman in a yellow sweater with curly hair, arms raised, smiling in a green outdoor*	*A British flag waving with colorful buildings and a clock tower in the background*	*A sea turtle swimming underwater with a clear blue background and coral below*	*...land art, National Geographic style, a photograph of coastlines in UK, absurdly...*

After completing preparations, we will pass all prompts, base images and mask images to the SD1.5 Inpainting and SD2.1 Inpainting, respectively.

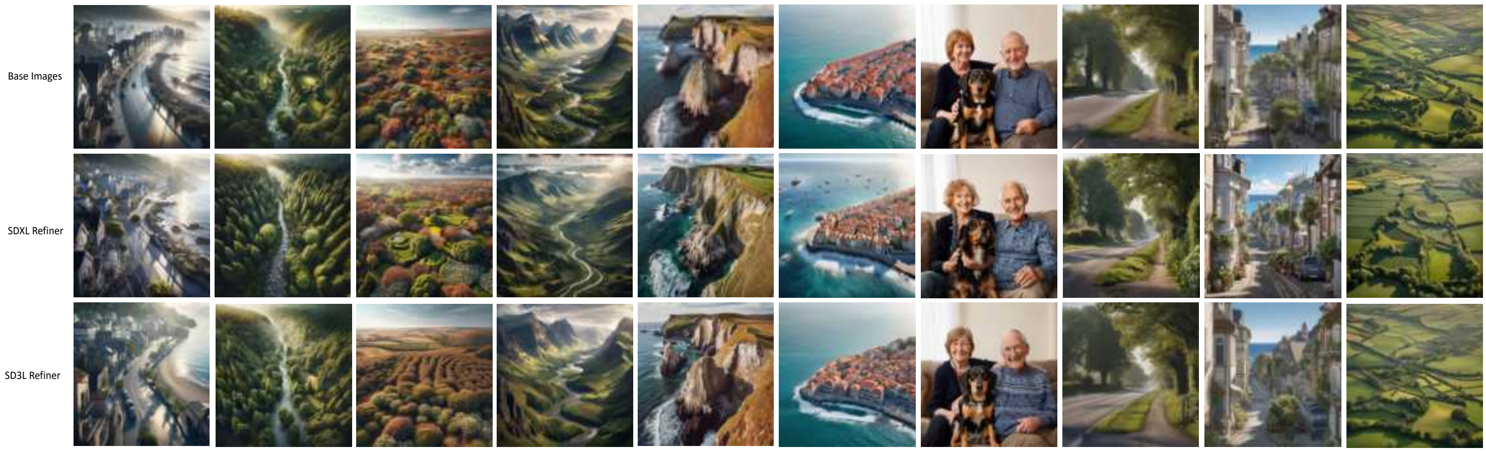
The following figure xxx illustrates editing base images via inpainting models using edited prompts and mask images:


Base images
SD1.5 Inpainting
SD2.1 Inpainting
photograph, An middle aged indian woman in indian traditional attire holding a basket of yellow flowers, sitting by a lake, sharp
overhead view, photorealism, epic, pixabay, landscape, land art, National Geographic style, a photograph of coastlines in UK, absurdly detailed, lush detail, crystal clear, sharp focus, natural light
photograph, A single wind turbine standing in a grassland, open landscape under a blue sky
photograph, A woman in a green sweater with curly hair, arms raised, smiling in a green outdoor, sharp
photograph, A brown cat/fox hopping through green grass
photograph, A woman with short gray hair and glasses smiling and hugging with husky dog
photograph, A middle-aged Indian man cuddling and smiling with his wife outdoors, sharp
photograph, A USA flag waving with colorful European style buildings, sharp

We observed that both SD2.1 Inpainting and SD1.5 Inpainting models may not be able to edit images perfectly. SD1.5 Inpainting model sometimes fails to edit images in line with textual descriptions while SD2.1 Inpainting model sometimes produces artifacts when editing images. However, SD2.1 Inpainting model seems slightly better than SD1.5 inpainting in terms of overall quality, which is consistent with the performances of their corresponding T2I models, namely, SD1 and SD2 T2I models.

4.3.2 Experimentation for Refiner Models

To test Refiner Models, we randomly select 10 images generated by DALL-E 3, SDXL, SD2 and SD3L in previous sections as our base images and their relevant prompts in the hope that image quality will be considerably enhanced. We then pass prompts and images to SDXL Refiner and SD3L Refiner to obtain new images that are similar to base images but with difference in details in the following table xxx:

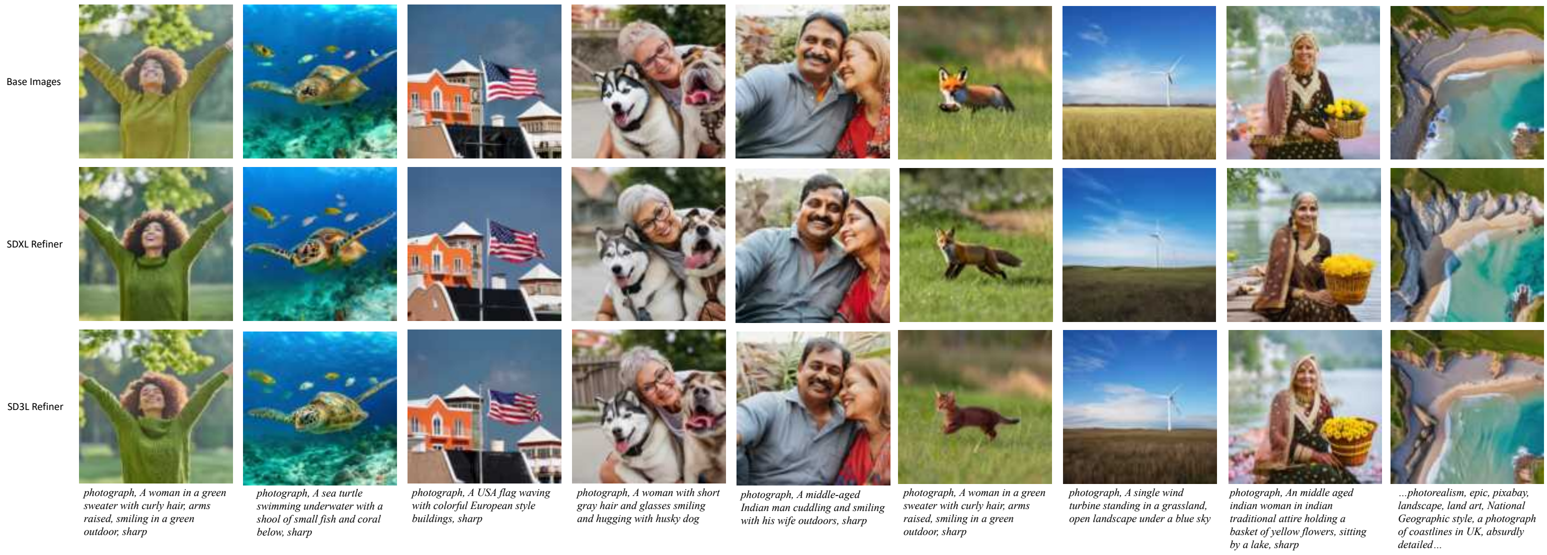

Base Images
SDXL Refiner
SD3L Refiner
...The street view photograph of town by the sea in the morning, lush detail...
...a photograph of forest in UK, absurdly detailed...
...a photograph of shrublands in UK...
...National Geographic style, a photograph of mountains in UK...
...absurdly detailed, The overhead view photograph of town by the sea in the morning...
...The street view photograph of town by the sea in the morning, lush detail...
... depth of field, an old couple with a dog sitting on the sofa, happy...
...absurdly detailed, The street view photograph of UK countryside in the morning, lush detail...
...absurdly detailed, The street view photograph of town by the sea in the morning, lush detail...
...National Geographic style, a photograph of fields in UK, absurdly detailed, lush detail...

From our experimentation, we observed that the authenticity of images from DALL-E 3 that look like illustrations has been improved when passing them into SDXL Refiner and SD3L Refiner. In addition, artifacts such as unclear hands of the man and blurriness of buildings by the sea have been mitigated as well, leading to higher authenticity and naturalness compared to base images. Hence, leveraging the results, we may consider passing images to refiners in our framework to improve overall image quality.

4.4 Experimentation for Ensemble Methods:

In section 4.3.1 and section 4.3.2, we only deploy one I2I model per time for image editing. Inspired by traditional machine learning ensemble method such as boosting method proposed by (Freund and Schapire, 1999), we also tested whether editing images by inpainting models and then adjusting them using refiners will improve the overall image editing quality compared to single model method.

In this section, we choose images generated by SD2.1 Inpainting model from section 4.3.1 and prompts used in section 4.3.2 as base images and prompts for editing. Then we input both the base image and the prompt into SDXL Refiner and SD3L Refiner respectively to assess the quality of new images. The table xxx below demonstrates the results of images modified by SDXL Refiner and SD3L Refiner:


Base Images
SDXL Refiner
SD3L Refiner
photograph, A woman in a green sweater with curly hair, arms raised, smiling in a green outdoor, sharp
photograph, A sea turtle swimming underwater with a shool of small fish and coral below, sharp
photograph, A USA flag waving with colorful European style buildings, sharp
photograph, A woman with short gray hair and glasses smiling and hugging with husky dog
photograph, A middle-aged Indian man cuddling and smiling with his wife outdoors, sharp
photograph, A woman in a green sweater with curly hair, arms raised, smiling in a green outdoor, sharp
photograph, A single wind turbine standing in a grassland, open landscape under a blue sky
photograph, An middle aged indian woman in indian traditional attire holding a basket of yellow flowers, sitting by a lake, sharp
…photorealism, epic, pixabay, landscape, land art, National Geographic style, a photograph of coastlines in UK, absurdly detailed…

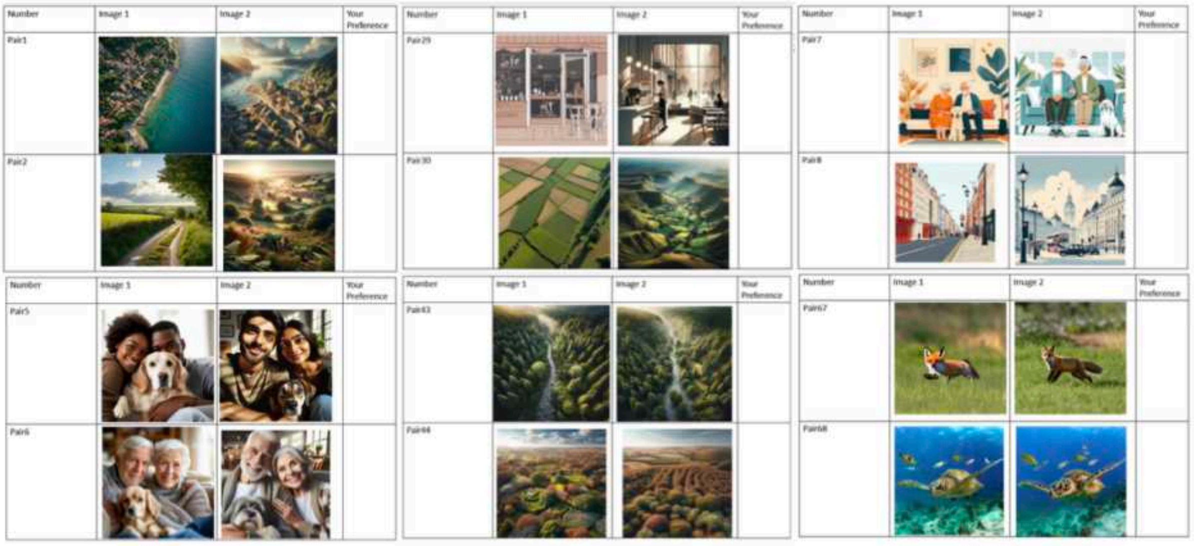
From author's observation, passing images edited by inpainting models into refiners could substantially improve the overall image quality because the refined images looked more natural and their artifacts at the edge of mask images are eliminated by refiners. Therefore, passing generated images to refiners before submitting to marketing team seems a plausible strategy.

4.5 Experimentation on Human Evaluation:

Since the quantitative metrics are not stable at the moment, we are using human evaluation to replace them. In this section, we have designed a survey which contains 69 image pairs synthesized by different models ranging from photographs like landscapes and cities to a variety of illustrations. The following table xxx descries the summary of our human evaluation used in this project:

Models Compared	Scope
SD3L vs DALL-E 3	city, human, illustrations, landscape
SDXL vs SD2	city, human, illustrations, landscape
SDXL vs DALL-E 3	human, illustration, landscape
SD1.5 Inpainting vs SD2.1 Inpainting	city, human, animal, landscape, object
SDXL Refiner vs SD3L Refiner	city, landscape, human
Inpainting to SD3L Refiner vs Inpainting to SDXL refiner	human, object, animal, landscape
SD2.1 Inpainting vs SD2.1 Inpainting to SDXL Refiner	human, object, landscape, animal

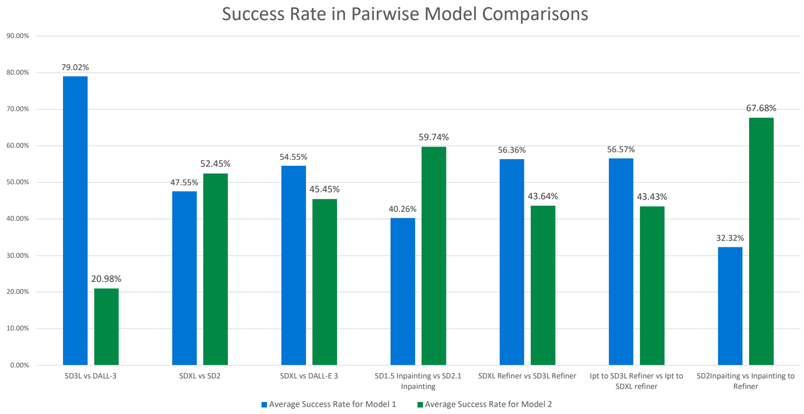
Each time, we pick up two images from different models to form the image pair, and we ask participants to specify which image they would prefer if they were members of marketing team who wish to choose one of those images for marketing purposes. The following table xxx illustrates part of pairwise human evaluation survey and results. The full survey and full results are included in the Appendix.



Within the survey, we pairwise compared T2I models such as DALL-E 3, SD3L and SDXL, I2I models such as SD2.1 Inpainting and SD3L Refiner to evaluate their overall performances by calculating the success rate of the model in each pair of images. Success rate for model refers to the percentage of time participants choose the image from that model in each image pair.

Model 1 vs Model 2	Pair Name	Preference	Preference	Preference	Preference	Preference	Preference	Preference	Preference	Preference	Preference	Preference	Success rate for Model 1	Success Rate for Model 2
SD3L vs DALL-3	Pair1	1	1	1	1	2	1	1	1	1	1	1	90.91%	9.09%
	Pair2	1	1	1	1	1	1	1	1	1	1	1	100.00%	0.00%
	Pair3	1	1	1	1	2	1	1	1	1	1	1	90.91%	9.09%
	Pair4	1	1	1	1	1	1	1	1	2	1	1	90.91%	9.09%
	Pair5	1	1	1	1	1	1	1	1	1	1	1	100.00%	0.00%
	Pair6	2	1	1	1	1	1	1	1	1	1	1	90.91%	9.09%
	Pair7	1	1	2	1	1	2	1	2	2	2	2	45.45%	54.55%
	Pair8	2	2	1	1	2	2	2	2	1	2	2	27.27%	72.73%
	Pair9	1	1	1	1	1	1	2	1	1	1	2	81.82%	18.18%
	Pair10	2	2	2	1	2	2	2	1	2	1	1	36.36%	63.64%
	Pair11	1	1	1	1	1	1	2	1	1	1	1	90.91%	9.09%
	Pair12	2	1	1	1	2	1	1	1	1	1	1	81.82%	18.18%
	Pair13	1	1	1	1	1	1	1	1	1	1	1	100.00%	0.00%
SDXL vs SD2	Pair14	2	2	2	2	2	2	2	2	2	2	2	0.00%	100.00%
	Pair15	1	1	2	1	2	2	1	1	1	2	2	54.55%	45.45%
	Pair16	2	2	2	2	2	2	2	1	2	1	2	18.18%	81.82%
	Pair17	1	1	1	2	1	2	2	2	2	2	2	36.36%	63.64%
	Pair18	1	1	1	1	2	2	2	2	2	2	2	36.36%	63.64%
	Pair19	1	1	1	1	2	2	1	1	1	2	1	72.73%	27.27%
	Pair20	1	2	2	2	2	2	1	1	1	1	2	45.45%	54.55%
	Pair21	1	1	1	1	1	1	1	1	1	1	1	100.00%	0.00%
	Pair22	1	1	1	2	1	1	1	1	1	1	2	81.82%	18.18%
	Pair23	1	1	1	2	1	1	1	1	1	1	1	90.91%	9.09%
	Pair24	1	1	2	2	2	2	2	1	2	2	2	27.27%	72.73%
	Pair25	2	2	2	1	1	2	2	1	1	2	2	36.36%	63.64%
	Pair26	2	1	2	2	2	2	1	2	2	2	2	18.18%	81.82%

We later analysed these responses with some visualisations, and we will obtain some insights from within. For instance, the following Figure xxx shows the average success rate of models during human evaluation. More in-depth analysis will be conducted in Section 5.



5. Results and Analysis:

5.1 Results from Keywords, Prompts and Evaluation Metrics

5.1.1 Keywords:

The table xxx below is the summary of effective keywords categorised by photography principles from company A's marketing team. Users may consult this table when designing their prompts. Full table is included in the Appendix.

[table here]

5.1.2 Prompts:

Based on the results from our experimentation, it is suggested that users may add effective keywords on both sides of main textual descriptions when designing prompts. Long prompts that contain more details can also be considered should users have sufficient time for prompt crafting.

5.1.3 Evaluation Metrics

As highlighted in section 4.1.4, the instability of the CLIP-IQA system necessitated the adoption of human evaluation as an alternative method, due to the inability to address this limitation within the current project scope. Despite its limitations, CLIP-IQA could still be of value in scenarios where large volumes of images are generated from the same prompt. In our project, we generated images one at a time using individual prompts. However, in practical applications, users may produce hundreds of images using a single prompt, selecting the best among them, which can be a laborious and time-intensive process.

CLIP-IQA could serve as a preliminary screening tool to eliminate images with lower scores, thereby reducing the number of options users need to manually evaluate. This approach would facilitate the selection process, allowing users to focus on choosing from a smaller, more refined set of high-quality images.

5.2. T2I Results

From section 4.2, we have collected relevant information of T2I models and briefly discussed general results of T2I models. The table xxx below has summarised the basic information of T2I with their associated cost and inference time during experimentation.

Model Name	Cost per hour	Inference time:	Cost per 100 Images $	Reference Hardware	Training Hardware	Tr Dataset	Resolution
DALL-E 3	-	16.16s	4 (Standard) / 8 (HD Images)	-	-	Internal Dataset	1024x1024
SDXL 1.0 Base	$22.03 per hour	21.79s	13.33	40 x Intel Xeon Platinum 8168 CPUs & 8 x NVDIA V100 GPUs		Internal Dataset	1024x1024
SD2	6.12 per hour	9.7s	1.65	6 x Intel Xeon E5-2690 v4 CPUs & 2 x NVDIA V100 GPUs	32 x 8 x A100 GPUs	LAION-5B	768x768
SD1.5	6.12 per hour	9.26s	1.57	6 x Intel Xeon E5-2690 v4 CPUs & 2 x NVDIA V100 GPUs	32 x 8 x A100 GPUs	LAION-2B (en)	512x512
SD1.4	6.12 per hour	9.25s	1.57	6 x Intel Xeon E5-2690 v4 CPUs & 2 x NVDIA V100 GPUs	32 x 8 x A100 GPUs	LAION-2B (en)	512x512
SD3L	0.065/EA	5.9s	6.50	-	-	Internal Dataset	1024x1024

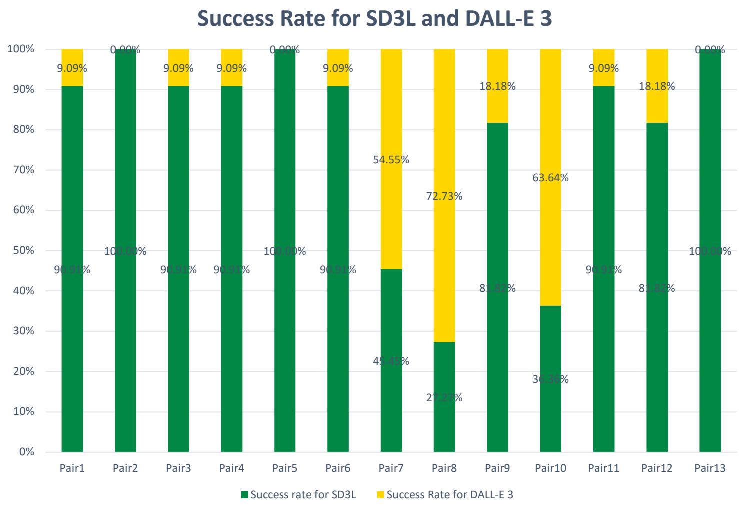
We noticed that SDXL is the most expensive model and requires the most inference time. On average, SDXL requires more than 21 seconds to generate an image and requires 13.33 USD generating 100 images. The DALL-E 3 is the second most expensive model and requires 16.16 seconds on average in generating one image. For SD1 and SD2 models, they preserve similar inference time and cost per generating 100 images. SD3L on the other hand is the model that can produce images at fastest pace albeit higher cost per generating 100 images compared to SD1 and SD2.

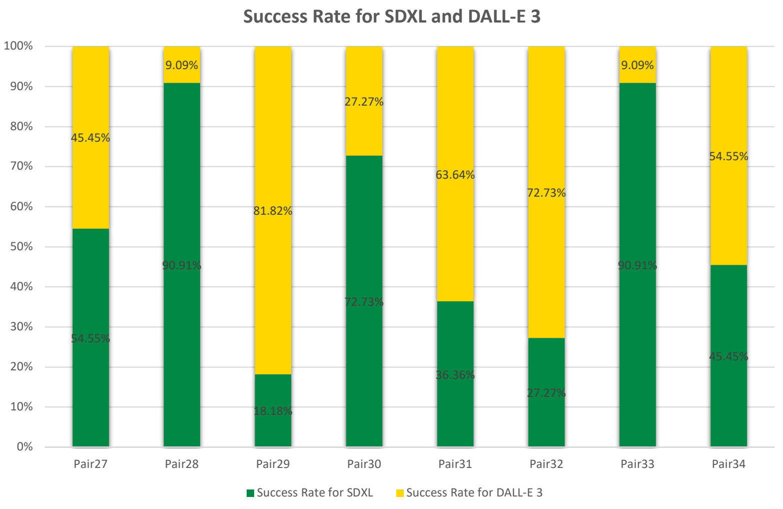
In our initial observation, we found that DALL-E 3 can generate high quality illustrations and could produce images with legible words, but it is struggling to synthesise photographs. Meanwhile, both SD3L and SDXL can potentially generate images pertaining to landscapes, human faces, and cities, etc and SD3L can further generate legible words in the image. However, they may occasionally fail to synthesise images aligned with prompts, generating perceivable artifacts and objects with incorrect numbers on the images.

To evaluate T2I models objectively, we have further conducted human evaluations. In the first 3 columns on Figure xxx in Section 4.5, we observed that SD3L was chosen more than 79% of time while SDXL was chosen for more than 54.55% time when comparing to DALL-E 3. Meanwhile, SD2 was selected more than 52% times compared to SDXL. These results indicate SD model families may outperform DALL-E 3 in overall performances. In addition, SDXL and SD2 preserve similar success rates in the pairwise comparison. Although the SD2 achieved marginally higher rate of 52.45% compared to SDXL, we cannot conclude that SD2 is better than SDXL due to potential statistical insignificance. This is because the size of human evaluation is small, if the number of response increase, the rates between these models may be altered. However, by considering the fact that SD2 can only synthesize images with lower resolutions, we may prefer SDXL than SD2 when generating images for marketing purposes.

Furthermore, by investigating results from human evaluation in detail, we noticed that DALL-E 3 performed significantly better than SD model families in terms of illustrations.

The Figures xxx blow illustrate the success rate of SD models and DALL-E 3 at pairwise level.





For illustration image pairs such as Pair8, Pair10, Pair29, Pair31 and Pair32, we observed that DALL-E 3 significantly outperforms both SD3L and DALL-E3. Therefore, we may suggest using DALL-E 3 to generate illustrations for marketing purposes.

5.2.1 T2I Summary

To summarise, based on the cost, inference times, image resolution and model performances, we recommended T2I models in the following Table xxx:

Task	T2I Model Rank	Model Name	Cost per 100 Images $	Average Inference Time	Resolution
Photogragh	1	SD3L	6.50	5.9s	1024x
	2	SDXL	13.33	21.79s	1024x
	3	SD2	1.65	9.7s	768x
Illustration	1	DALL-E 3	4~8	16.16s	1024x

It is recommended that deploying SD3L for generating photorealistic images and utilising DALL-E 3 for illustrations synthesis as they have higher performances, relatively lower costs, and inference time.

However, there are limitations that may undermine our conclusions to T2I models. Firstly, the size of experimentation is small due to the time constraint of the project, meaning that true performances of each model may not be fully exerted in our comparisons. Secondly, the limited number of participants, varied knowledge base among respondents and limited number of image pairs may introduce biases to the human evaluation results, making unfair judgement to T2I models. Thirdly, the SD3L model being utilised is from firework.ai whose parameters such as the number of denoising iterations are fixed, indicating that the model may not demonstrate its full potentials compared to models deployed on Azure Machine Learning. Therefore, further experimentation and more involved human evaluations are recommended to explore model's full performances.

5.3. I2I Results

In Section 4.3 and Section 4.4, we have explored the performances of I2I models ranging from inpainting models, refiners and their ensemble methods. The table xxx below illustrates information of I2I models used during experimentation.

Model Name	Cost per hour	Inference time:	Cost per 100 Images $	Inference Hardware	Training Hardware	Tr Dataset	Resolution
SD2.1 Inpainting	6.12 per hour	8.3s	1.41	6 x Intel Xeon E5-2690 v4 CPUs & 2 x NVDIA V100 GPUs	32 x 8 x A100 GPUs	LAION-5B	512x512
SDXL-refiner	6.12 per hour	13.5s	2.30	6 x Intel Xeon E5-2690 v4 CPUs & 2 x NVDIA V100 GPUs	32 x 8 x A100 GPUs	LAION-5B	1024x1024
SD1.5 Inpainting	6.12 per hour	4.5s	0.77	6 x Intel Xeon E5-2690 v4 CPUs & 2 x NVDIA V100 GPUs	32 x 8 x A100 GPUs	LAION-2B (en)	512x512
SD3 Large Refiner	0.065/EA	7.6s	6.50	-	-	Internal Dataset	1024x1024

We could see that SD3L Refiner is the most expensive refiner model and SD2.1 Inpainting is the most expensive inpainting model. In terms of inference time, SDXL Refiner requires most time for editing, which is 13.5s on average. For other models, the time varies from 4.5s to 8.3s. However, the time for producing mask image is not considered, meaning the actual inference time for inpainting models may be underestimated.

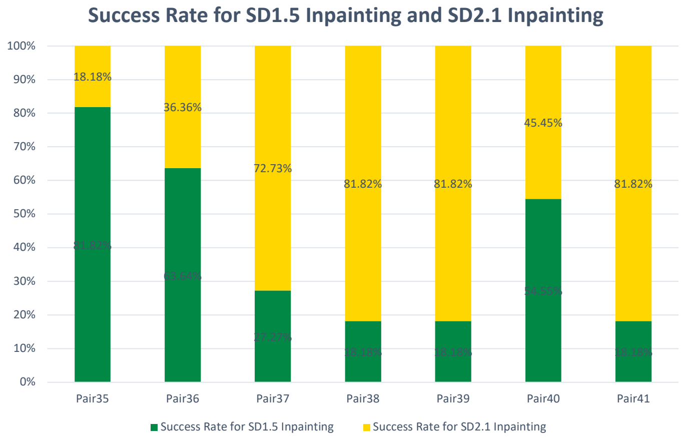
From our initial investigation, it is observed that SD2.1 inpainting model may outperform SD1.5 in terms of quality and text-image alignment, although both models may introduce artifacts in the edge of edited areas. For refiners, both SDXL and SD3L could edit images with under text instructions. As for ensemble methods, it seems passing edited images to refiners could improve overall image quality as the artifacts on edges are mitigated.

To justify our assumptions, we analysed results from human evaluations on Figure xxx in section 4.5 and we found that respondents prefer images edited by SD2.1 Inpainting than SD1.5 Inpainting more as SD2.1 Inpainting achieved the success rate of approximately 60%. Meanwhile, when passing T2I images and Inpainting Images to refiners, we noticed that SDXL Refiner achieved 56.36% success rate in refining T2I images but SD3L achieved 56.57% when refining I2I Inpainting images, indicating the different preference of refiners in different scenarios. However, since the difference of success rate between SD3L Refiner and SDXL Refiner is small, it is not yet clear if it is due to the limited size of human evaluation participants that leads to varied results and hence being statistically insignificant. Therefore, more comprehensive experimentation should be conducted to discover the full capabilities of these two refiners.

5.3.1 Results for Inpainting Models:

Although we have concluded SD2.1 Inpainting is recommended based on its average success rate, a detailed analysis of these two models is conducted further.

The Figure xxx below demonstrates the success rate between SD1.5 inpainting and SD2.1 Inpainting at pairwise level. We investigated Pair35, Pair36 and Pair40 whose success rates for SD1.5 Inpainting are higher and found the quality of images edited by SD1.5 Inpainting model is not satisfactory due to artifacts and unnaturalness on images. Since the human evaluation forced participants choose one preferred image per time, it may not be able to identify if both images in the pair are of unsatisfied quality. Therefore, users should be aware that inpainting models may not successfully edit images based on user instruction.

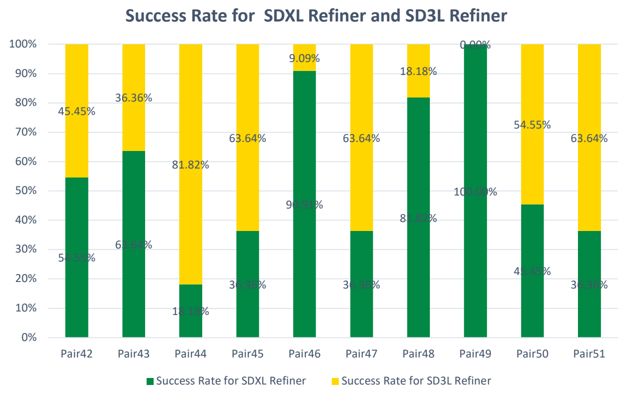


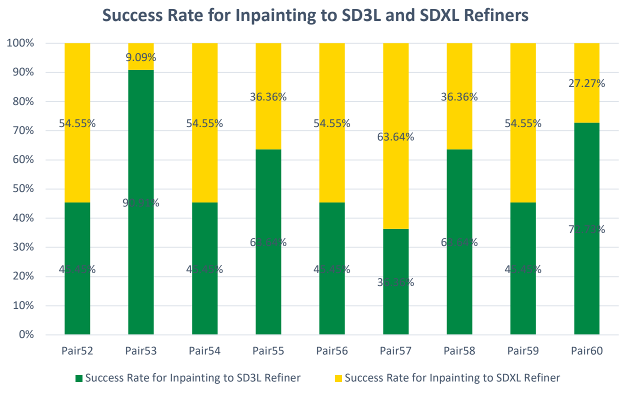
5.3.2 Results for Refiners:

From the Table xxx in section 4.5, we are aware that SDXL achieved marginally higher success rate of 56.36% in terms of refining T2I images while SD3L achieved slightly higher success rate of 56.57% in terms of refining inpainting images. However, by averaging success rates we noticed that SDXL achieved 49.90% success rate while SD3L achieved 50.1%, which are roughly equal.

By further exploring their pairwise results in the Figures xxx below, it seems both refiners preserve the comparable capabilities in refining images. But considering the fact that the cost of SD3L Refiner is 2.8 times more than SDXL Refiner, we may recommend SDXL Refiner as the first choice for marketing image generation.

However, due to the size and potential bias in our human evaluation and due to the nature that SD3L is an online demo whose parameters are fixed, our results may not accurate. Further experimentations are required to explore full performances of refiners.





5.3.3. Ensemble Method results:

Based on the average success rates of 67.68% for ensemble method in Table xxx from Section 4.5, we observed that passing inpainting images to refiners would significantly improve the overall quality of edited images. Hence, we will adopt this strategy of passing images to the refiner in our framework for image generation.

5.3.4 I2I Summary

Based on all results analysed in section 5.3, we have summarised advantages and disadvantages of each type of I2I models in the table xxx below:

I2I Model Type	Advantages	Disadvantages
Inpainting Models	Non-edited areas remain untouched Enables stock images to be reusable	• Requires Mask Image • May fail to edit in line with prompt. • Style may be inconsistent
Refiners	More consistent style No mask image required More controllable output	• May not significantly change the overall structure of images. • Similar to input image
Ensemble Methods	Improved overall image quality Reduced artifacts generated by Inpainting Models Could change the style of images	• More time consuming • More cost incurred

Considering these aspects and results from human evaluation, we may recommend SD2.1 Inpainting for inpainting tasks and SDXL Refiner for refining tasks. However, as mentioned before, users are supposed to be aware that inpainting models may sometimes fail and actual performances between SD3L Refiner and SDXL Refiner are not exactly understood.

Model Type	I2I Model Rank	Model Name	Cost per 100 Images $	Average Inference Time
Inpainting	1	SD2.1 Inpainting	1.41	8.3s
	2	SD1.5 Inpainting	0.77	4.5s
Refiner	1	SDXL Refiner	2.30	13.5s
	2	SD3L Refiner	6.50	7.6s

5.4 Analysis of Results:

[what we have done so far]

Using T2I models available may not be able to generate high quality images because of xxxx

Limitations summary table here:

[Authenticity, Diversity, artifacts, unreliable metrics, hard to control models]

How far are we from solving these limitations?

Artifacts: easiest one to slove: can use tools such as Codeformers, Automatic1111, ComfyUI etc

Diversity: need unbiased models, or fine tuning. We can fine tune the model by feeding them with images we would like to genenrate. But this also arises another problem, may involve ethical issues about data. Available data don’t mean we can use them immediately.

Authenticity: New models. In July 2024, new model SD3 Medium published. It is aimed at improving authenticity. So maybe SD3M can be a solution to improve authenticity.

Metrics: No reliable metrics so far. Most challenging part.

[table here]

Can’t use immediately but we can use them to guide photographers taking photos. Traditional photo briefings contain textual descriptions only. We can use these descriptions to generate images and attach them into briefings.

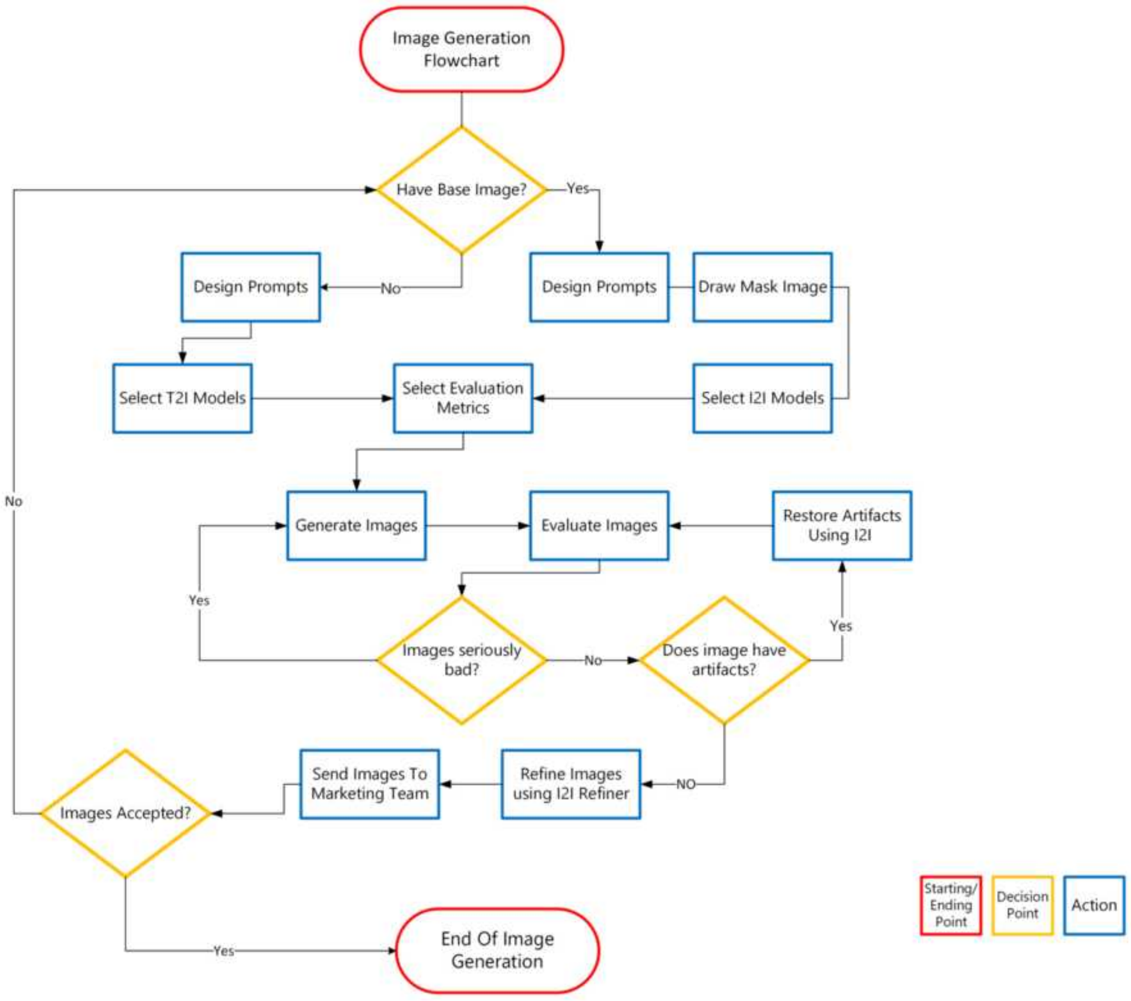
We only generate one image per time. Which may be biased.

Human evaluation may be biased as well.

If we want to solve current limitations, what else should we do? Reliable metrics, unbiased models, tools fixing artifacts, more controllable generation process (hard to control output, we want model better to understand what we think about)

[table of how far we are].

6. Framework:



7. Application of Framework:

8.Conclusion and Future Research

[where we are, how far we are from objective]

We tested GPT4o I2I capabilities to check if they can edit images. But from our initial testing, it seems we cannot. Maybe it is because we are not using the effective prompts to let GPT4o show its full capabilities. This is left as future research.

#Future research:

Use Gen images and Human made images for A/B testing to assess whether Gen Image ads have better or comparable click rates.

Test GPT4o in designing prompt

Test SDM3 Medium

Test black-box models such as Midjourney

Test GUIs like Automatic1111 (Style Aligned and ControlNet Reference to generate consistent style images), ConfyUI.

Use Codeformer (zhou et al., 2022b) restore face.

Fine-tuning (Liao et al 2024 enhance face quality)

Croud source human eval

Reference List: (Will use IEEE style in Formal Dissertation by LaTex)

[1]
Y. Zhou *et al.*, “LARGE LANGUAGE MODELS ARE HUMAN-LEVEL PROMPT ENGINEERS”, Accessed: Aug. 06, 2024. [Online]. Available: https://github.com/keirp/automatic_prompt_engineer.

[2]
P. Lewis *et al.*, “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” *Advances in Neural Information Processing Systems*, vol. 2020-December, May 2020, Accessed: Aug. 06, 2024. [Online]. Available: https://arxiv.org/abs/2005.11401v4

[3]
J. Wei *et al.*, “Chain-of-Thought Prompting Elicits Reasoning in Large Language Models,” *Advances in Neural Information Processing Systems*, vol. 35, Jan. 2022, Accessed: Aug. 06, 2024. [Online]. Available: https://arxiv.org/abs/2201.11903v6

[4]
Z. J. Wang, E. Montoya, D. Munechika, H. Yang, B. Hoover, and D. H. Chau, “DiffusionDB: A Large-scale Prompt Gallery Dataset for Text-to-Image Generative Models,” *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, vol. 1, pp. 893–911, Oct. 2022, doi: 10.18653/v1/2023.acl-long.51.

[5]
“LAION.” Accessed: Aug. 05, 2024. [Online]. Available: https://laion.ai/

[6]
C. Schuhmann *et al.*, “LAION-5B: An open large-scale dataset for training next generation image-text models,” *Advances in Neural Information Processing Systems*, vol. 35, Oct. 2022, Accessed: Aug. 05, 2024. [Online]. Available: https://arxiv.org/abs/2210.08402v1

[7]
T.-Y. Lin *et al.*, “Microsoft COCO: Common Objects in Context”.

[8]
“COCO - Common Objects in Context.” Accessed: Aug. 05, 2024. [Online]. Available: https://cocodataset.org/#home

[9]
“Advertising - Worldwide | Statista Market Forecast.” Accessed: Aug. 05, 2024. [Online]. Available: https://www.statista.com/outlook/amo/advertising/worldwide#ad-spending

[10]
Y. Ren, Y. Romano, and M. Elad, “Example-Based Image Synthesis via Randomized Patch-Matching,” *IEEE Transactions on Image Processing*, vol. 27, no. 1, pp. 220–235, Sep. 2016, doi: 10.1109/TIP.2017.2750419.

[11]
A. Gonzalez-Garcia, J. van de Weijer, and Y. Bengio, “Image-to-image translation for cross-domain disentanglement,” *Advances in Neural Information Processing Systems*, vol. 2018-December, pp. 1287–1298, May 2018, Accessed: Aug. 04, 2024. [Online]. Available: https://arxiv.org/abs/1805.09730v3

[12]
A. Hertzmann, C. E. Jacobs, N. Oliver, B. Curless, and D. H. Salesin, “Image Analogies,” 2001, Accessed: Aug. 04, 2024. [Online]. Available: http://grail.cs.washington.edu/projects/image-analogies/

[13]
Y. Song *et al.*, “Contextual-based Image Inpainting: Infer, Match, and Translate,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11206 LNCS, pp. 3–18, Nov. 2017, doi: 10.1007/978-3-030-01216-8_1.
[14]
D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, “Context Encoders: Feature Learning by Inpainting,” *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-December, pp. 2536–2544, Apr. 2016, doi: 10.1109/CVPR.2016.278.
[15]
H. Kazemi, S. Soleymani, F. Taherkhani, and N. M. Nasrabadi, “Unsupervised Image-to-Image Translation Using Domain-Specific Variational Information Bound”.
[16]
J. Y. Zhu *et al.*, “Toward Multimodal Image-to-Image Translation,” *Advances in Neural Information Processing Systems*, vol. 2017-December, pp. 466–477, Nov. 2017, Accessed: Aug. 04, 2024. [Online]. Available: https://arxiv.org/abs/1711.11586v4
[17]
J. Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks,” *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2017-October, pp. 2242–2251, Mar. 2017, doi: 10.1109/ICCV.2017.244.
[18]
“Introducing ChatGPT | OpenAI.” Accessed: Aug. 04, 2024. [Online]. Available: https://openai.com/index/chatgpt/
[19]
H. Touvron *et al.*, “LLaMA: Open and Efficient Foundation Language Models,” Feb. 2023, Accessed: Aug. 04, 2024. [Online]. Available: https://arxiv.org/abs/2302.13971v1
[20]
T. B. Brown *et al.*, “Language Models are Few-Shot Learners,” *Advances in Neural Information Processing Systems*, vol. 2020-December, May 2020, Accessed: Aug. 04, 2024. [Online]. Available: https://arxiv.org/abs/2005.14165v4
[21]
J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, vol. 1, pp. 4171–4186, Oct. 2018, Accessed: Aug. 04, 2024. [Online]. Available: https://arxiv.org/abs/1810.04805v2
[22]
C. Raffel *et al.*, “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer,” *Journal of Machine Learning Research*, vol. 21, pp. 1–67, Oct. 2019, Accessed: Aug. 04, 2024. [Online]. Available: https://arxiv.org/abs/1910.10683v4
[23]
C. Zhang, C. Zhang, M. Zhang, and I. S. Kweon, “Text-to-image Diffusion Models in Generative AI: A Survey,” Mar. 2023, Accessed: Aug. 04, 2024. [Online]. Available: https://arxiv.org/abs/2303.07909v2
[24]
A. van den Oord, N. Kalchbrenner, and K. Kavukcuoglu, “Pixel Recurrent Neural Networks,” Jan. 2016, Accessed: Aug. 04, 2024. [Online]. Available: https://arxiv.org/abs/1601.06759v3

[25]
A. van den Oord, N. Kalchbrenner, O. Vinyals, L. Espeholt, A. Graves, and K. Kavukcuoglu, “Conditional Image Generation with PixelCNN Decoders,” *Advances in Neural Information Processing Systems*, pp. 4797–4805, Jun. 2016, Accessed: Aug. 04, 2024. [Online]. Available: https://arxiv.org/abs/1606.05328v2
[26]
D. P. Kingma and J. L. Ba, “Adam: A Method for Stochastic Optimization,” *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, Dec. 2014, Accessed: Aug. 04, 2024. [Online]. Available: https://arxiv.org/abs/1412.6980v9
[27]
G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, “Improving neural networks by preventing co-adaptation of feature detectors,” Jul. 2012, Accessed: Aug. 04, 2024. [Online]. Available: https://arxiv.org/abs/1207.0580v1
[28]
A. Vaswani *et al.*, “Attention Is All You Need,” *Advances in Neural Information Processing Systems*, vol. 2017-December, pp. 5999–6009, Jun. 2017, Accessed: Aug. 04, 2024. [Online]. Available: https://arxiv.org/abs/1706.03762v7
[29]
P. J. Werbos, “Applications of advances in nonlinear sensitivity analysis,” *System Modeling and Optimization*, pp. 762–770, Oct. 1982, doi: 10.1007/BFB0006203.
[30]
S. Linnainmaa, “Taylor expansion of the accumulated rounding error,” *BIT*, vol. 16, no. 2, pp. 146–160, 1976, doi: 10.1007/BF01931367/METRICS.
[31]
S. Amari, “A Theory of Adaptive Pattern Classifiers,” *IEEE Transactions on Electronic Computers*, vol. EC-16, no. 3, pp. 299–307, 1967, doi: 10.1109/PGEC.1967.264666.
[32]
Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2323, 1998, doi: 10.1109/5.726791.
[33]
D. E. Rumerlhart, G. Hinton, and R. J. Williams, “Learning Representations by Back-propagating Errors,” *Nature*, vol. 323, Oct. 1986, Accessed: Aug. 03, 2024. [Online]. Available: https://www.iro.umontreal.ca/~vincentp/ift3395/lectures/backprop_old.pdf
[34]
S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/NECO.1997.9.8.1735.
[35]
B. Widrow, “An Adaptive ‘ADALINE’ Neuron Using Chemical ‘MEMISTORS,’” Stanford, 1960. Accessed: Jul. 31, 2024. [Online]. Available: https://www-isl.stanford.edu/~widrow/papers/t1960anadaptive.pdf
[36]
F. Rosenblatt, “THE PERCEPTRON: A PROBABILISTIC MODEL FOR INFORMATION STORAGE AND ORGANIZATION IN THE BRAIN 1,” *Psychological Review*, vol. 65, no. 6, pp. 19–27.
[37]
M. Elasri, O. Elharrouss, S. Al-Maadeed, and H. Tairi, “Image Generation: A Review,” *Neural Processing Letters*, vol. 54, no. 5, pp. 4609–4646, Oct. 2022, doi: 10.1007/S11063-022-10777-X/FIGURES/5.

[38]
“Introducing ChatGPT | OpenAI.” Accessed: Jul. 28, 2024. [Online]. Available: https://openai.com/index/chatgpt/
[39]
“Udio - Introducing v1.5.” Accessed: Jul. 28, 2024. [Online]. Available: https://www.udio.com/blog/introducing-v1-5
[40]
Y. Cao *et al.*, “A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT; A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT,” *J. ACM*, vol. 37, no. 111, 2018, Accessed: Jul. 26, 2024. [Online]. Available: https://doi.org/XXXXXXX.XXXXXXX
[41]
T. Kynkäänniemi, T. Karras, S. Laine, J. Lehtinen, and T. Aila, “Improved Precision and Recall Metric for Assessing Generative Models,” *Advances in Neural Information Processing Systems*, vol. 32, Apr. 2019, Accessed: Jul. 18, 2024. [Online]. Available: https://arxiv.org/abs/1904.06991v3
[42]
M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, “GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium,” *Advances in Neural Information Processing Systems*, vol. 2017-December, pp. 6627–6638, Jun. 2017, doi: 10.18034/ajase.v8i1.9.
[43]
T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, “Improved Techniques for Training GANs,” *Advances in Neural Information Processing Systems*, pp. 2234–2242, Jun. 2016, Accessed: Jul. 18, 2024. [Online]. Available: https://arxiv.org/abs/1606.03498v1
[44]
J. Wang, K. C. K. Chan, and C. C. Loy, “Exploring CLIP for Assessing the Look and Feel of Images”, Accessed: Jul. 18, 2024. [Online]. Available: https://github.com/IceClear/
[45]
A. Mittal, A. K. Moorthy, and A. C. Bovik, “No-reference image quality assessment in the spatial domain,” *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, 2012, doi: 10.1109/TIP.2012.2214050.
[46]
K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, “Image Quality Assessment: Unifying Structure and Texture Similarity,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 5, pp. 2567–2581, Apr. 2020, doi: 10.1109/TPAMI.2020.3045810.
[47]
E. Prashnani, H. Cai, Y. Mostofi, and P. Sen, “PieAPP: Perceptual Image-Error Assessment through Pairwise Preference,” *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1808–1817, Jun. 2018, doi: 10.1109/CVPR.2018.00194.
[48]
V. Petsiuk *et al.*, “Human Evaluation of Text-to-Image Models on a Multi-Task Benchmark,” Nov. 2022, Accessed: Jul. 18, 2024. [Online]. Available: https://arxiv.org/abs/2211.12112v1
[49]
S. Jayasumana *et al.*, “Rethinking FID: Towards a Better Evaluation Metric for Image Generation,” Nov. 2023, Accessed: Jul. 18, 2024. [Online]. Available: https://arxiv.org/abs/2401.09603v2
[50]

M. Otani *et al.*, “Toward Verifiable and Reproducible Human Evaluation for Text-to-Image Generation,” *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2023-June, pp. 14277–14286, Apr. 2023, doi: 10.1109/CVPR52729.2023.01372.

[51]

S. Li *et al.*, “ZONE: Zero-Shot Instruction-Guided Local Editing,” Dec. 2023, Accessed: Jul. 18, 2024. [Online]. Available: https://arxiv.org/abs/2312.16794v2

[52]

S. Li, H. Singh, and A. Grover, “InstructAny2Pix: Flexible Visual Editing via Multimodal Instruction Following”.

[53]

Y. Sun *et al.*, “ImageBrush: Learning Visual In-Context Instructions for Exemplar-Based Image Manipulation,” 2023.

[54]

S. Sheynin *et al.*, “Emu Edit: Precise Image Editing via Recognition and Generation Tasks”, Accessed: Jul. 18, 2024. [Online]. Available: https://emu-edit.metademolab.com/

[55]

Z. Geng *et al.*, “InstructDiffusion: A Generalist Modeling Interface for Vision Tasks”, Accessed: Jul. 18, 2024. [Online]. Available: https://gengzigang.github.io/instructdiffusion.github.io/

[56]

Z. Wang, L. Zhao, and W. Xing, “StyleDiffusion: Controllable Disentangled Style Transfer via Diffusion Models”.

[57]

G. Kim, T. Kwon, and J. C. Ye, “DiffusionCLIP: Text-Guided Diffusion Models for Robust Image Manipulation”, Accessed: Jul. 18, 2024. [Online]. Available: https://github.com/gwang-kim/DiffusionCLIP.git

[58]

S. Zhang, S. Xiao, and W. Huang, “Forgedit: Text-guided Image Editing via Learning and Forgetting”.

[59]

P. Li, Q. Huang, Y. Ding, and Z. Li, “LayerDiffusion: Layered Controlled Image Editing with Diffusion Models”.

[60]

C. Mou, X. Wang, J. Song, Y. Shan, and J. Zhang, “DRAGONDIFFUSION: ENABLING DRAG-STYLE MA-NIPULATION ON DIFFUSION MODELS”, Accessed: Jul. 18, 2024. [Online]. Available: https://mc-e.github.io/project/DragonDiffusion/

[61]

M. Cao, X. Wang, Z. Qi, Y. Shan, X. Qie, and Y. Zheng, “MasaCtrl: Tuning-Free Mutual Self-Attention Control for Consistent Image Synthesis and Editing”, Accessed: Jul. 18, 2024. [Online]. Available: https://github.com/TencentARC/MasaCtrl

[62]

P. Isola, J.-Y. Zhu, T. Zhou, A. A. Efros, and B. A. Research, “Image-to-Image Translation with Conditional Adversarial Networks”, Accessed: Jul. 18, 2024. [Online]. Available: https://github.com/phillipi/pix2pix.

[63]

Y. Pang, J. Lin, T. Qin, and Z. Chen, “Image-to-Image Translation: Methods and Applications,” *IEEE Transactions on Multimedia*, vol. 24, pp. 3859–3881, 2022, doi: 10.1109/TMM.2021.3109419.

[64]

Y. Huang *et al.*, “Diffusion Model-Based Image Editing: A Survey,” Feb. 2024, Accessed: Jul. 18, 2024. [Online]. Available: https://arxiv.org/abs/2402.17525v2

[65]

J. Hartmann, Y. Exner, and S. Domdey, “The power of generative marketing: Can generative AI create superhuman visual marketing content?,” *SSRN Electronic Journal*, Apr. 2024, doi: 10.2139/SSRN.4597899.

[66]

“Popularity of generative AI in marketing U.S. 2023 | Statista.” Accessed: Jul. 18, 2024. [Online]. Available: https://www.statista.com/statistics/1388390/generative-ai-usage-marketing/

[67]

“Global AI in marketing revenue 2028 | Statista.” Accessed: Jul. 18, 2024. [Online]. Available: https://www.statista.com/statistics/1293758/ai-marketing-revenue-worldwide/#

[68]

N. Kshetri, “The Economics of Generative Artificial Intelligence in the Academic Industry,” *Computer*, vol. 56, no. 8, pp. 77–83, Aug. 2023, doi: 10.1109/MC.2023.3278089.

[69]

“Meta broadens access to generative AI tools, seeking to save marketers time | Marketing Dive.” Accessed: Jul. 18, 2024. [Online]. Available: https://www.marketingdive.com/news/Meta-launches-AI-sandbox-tool-social-media-advertising/650117/

[70]

“Meta announces generative AI features for advertisers | TechCrunch.” Accessed: Jul. 18, 2024. [Online]. Available: https://techcrunch.com/2023/05/11/meta-announces-generative-ai-features-for-advertisers/?guccounter=1&amp;guce_referrer=aHR0cHM6Ly93d3cuZ29vZ2xlLmNvbS8&amp;guce_referrer_sig=AQAAAK5w074vOuB6tZfQwG2Z6Bi2vLDwLr7j6oJG_0NN7qw0xTJG9yXxYuea-WYb1EbEvhlfPbW-WteDxGRK6cFu3gaJHIE3WsPtMmW-ocArZBKYl6-4KoDIHre9JUTFYLcDDMwqQNOjA-5KAf5cA-5G5XaDF0ubxY3Wpb3kCCUluxm0

[71]

N. Kshetri, Y. K. Dwivedi, T. H. Davenport, and N. Panteli, “Generative artificial intelligence in marketing: Applications, opportunities, challenges, and research agenda,” *International Journal of Information Management*, vol. 75, p. 102716, Apr. 2024, doi: 10.1016/J.IJINFOMGT.2023.102716.

[72]

X. Zhang *et al.*, “A Survey on Personalized Content Synthesis with Diffusion Models”.

[73]

Z. J. Wang, E. Montoya, D. Munechika, H. Yang, B. Hoover, and D. H. Chau, “DiffusionDB: A Large-scale Prompt Gallery Dataset for Text-to-Image Generative Models,” *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, vol. 1, pp. 893–911, Oct. 2022, doi: 10.18653/v1/2023.acl-long.51.

[74]

A. Blattmann, R. Rombach, K. Oktay, J. Müller, and B. Ommer, “Retrieval-Augmented Diffusion Models,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 15309–15324, Dec. 2022, Accessed: Jul. 18, 2024. [Online]. Available: https://github.com/CompVis/retrieval-augmented-diffusion-models

[75]

A. Sauer, F. Boesel, T. Dockhorn, A. Blattmann, P. Esser, and R. Rombach, “Fast High-Resolution Image Synthesis with Latent Adversarial Diffusion Distillation,” 2024.

[76]

D. Podell *et al.*, “SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis,” Jul. 2023, Accessed: Jul. 18, 2024. [Online]. Available: https://arxiv.org/abs/2307.01952v1

[77]
C. Meng *et al.*, “SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations,” *ICLR 2022 - 10th International Conference on Learning Representations*, Aug. 2021, Accessed: Jul. 18, 2024. [Online]. Available: https://arxiv.org/abs/2108.01073v2

[78]
T. Yin *et al.*, “One-step Diffusion with Distribution Matching Distillation,” Nov. 2023, Accessed: Jul. 18, 2024. [Online]. Available: https://arxiv.org/abs/2311.18828v3

[79]
O. Avrahami *et al.*, “SpaText: Spatio-Textual Representation for Controllable Image Generation,” *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2023-June, pp. 18370–18380, Nov. 2022, doi: 10.1109/CVPR52729.2023.01762.

[80]
S. Gu *et al.*, “Vector Quantized Diffusion Model for Text-to-Image Synthesis,” *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2022-June, pp. 10686–10696, Nov. 2021, doi: 10.1109/CVPR52688.2022.01043.

[81]
P. Esser *et al.*, “Scaling Rectified Flow Transformers for High-Resolution Image Synthesis,” Mar. 2024, Accessed: Jul. 18, 2024. [Online]. Available: https://arxiv.org/abs/2403.03206v1

[82]
R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-Resolution Image Synthesis with Latent Diffusion Models,” *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2022-June, pp. 10674–10685, Dec. 2021, doi: 10.1109/CVPR52688.2022.01042.

[83]
W. Chen, H. Hu, C. Saharia, and W. W. Cohen, “Re-Imagen: Retrieval-Augmented Text-to-Image Generator,” Sep. 2022, Accessed: Jul. 18, 2024. [Online]. Available: https://arxiv.org/abs/2209.14491v3

[84]
Y. Balaji *et al.*, “eDiff-I: Text-to-Image Diffusion Models with an Ensemble of Expert Denoisers,” Nov. 2022, Accessed: Jul. 18, 2024. [Online]. Available: https://arxiv.org/abs/2211.01324v5

[85]
Z. Feng *et al.*, “ERNIE-ViLG 2.0: Improving Text-to-Image Diffusion Model with Knowledge-Enhanced Mixture-of-Denoising-Experts,” *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2023-June, pp. 10135–10145, Oct. 2022, doi: 10.1109/CVPR52729.2023.00977.

[86]
Y. Zhou, B. Liu, Y. Zhu, X. Yang, C. Chen, and J. Xu, “Shifted Diffusion for Text-to-image Generation,” pp. 10157–10166, Nov. 2022, doi: 10.1109/cvpr52729.2023.00979.

[87]
C. Saharia *et al.*, “Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding,” *Advances in Neural Information Processing Systems*, vol. 35, May 2022, Accessed: Jul. 18, 2024. [Online]. Available: https://arxiv.org/abs/2205.11487v1

[88]
A. Nichol *et al.*, “GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models,” *Proceedings of Machine Learning Research*, vol. 162, pp.

16784–16804, Dec. 2021, Accessed: Jul. 18, 2024. [Online]. Available: https://arxiv.org/abs/2112.10741v3

[89]

J. Ho and T. Salimans, “Classifier-Free Diffusion Guidance,” Jul. 2022, Accessed: Jul. 18, 2024. [Online]. Available: https://arxiv.org/abs/2207.12598v1

[90]

W. Weng and X. Zhu, “U-Net: Convolutional Networks for Biomedical Image Segmentation,” *IEEE Access*, vol. 9, pp. 16591–16603, May 2015, doi: 10.1109/ACCESS.2021.3053408.

[91]

L. Yang *et al.*, “Diffusion Models: A Comprehensive Survey of Methods and Applications,” *ACM Computing Surveys*, vol. 56, no. 4, p. 54, Sep. 2022, doi: 10.1145/3626235.

[92]

J. Song, C. Meng, and S. Ermon, “Denoising Diffusion Implicit Models,” *ICLR 2021 - 9th International Conference on Learning Representations*, Oct. 2020, Accessed: Jul. 18, 2024. [Online]. Available: https://arxiv.org/abs/2010.02502v4

[93]

Z. Lai, X. Zhu, J. Dai, Y. Qiao, and W. Wang, “Mini-DALLE3: Interactive Text to Image by Prompting Large Language Models,” Oct. 2023, Accessed: Jul. 18, 2024. [Online]. Available: https://arxiv.org/abs/2310.07653v2

[94]

H. Chang *et al.*, “Muse: Text-To-Image Generation via Masked Generative Transformers,” *Proceedings of Machine Learning Research*, vol. 202, pp. 4055–4075, Jan. 2023, Accessed: Jul. 18, 2024. [Online]. Available: https://arxiv.org/abs/2301.00704v1

[95]

A. Razavi, A. van den Oord, and O. Vinyals, “Generating Diverse High-Fidelity Images with VQ-VAE-2,” *Advances in Neural Information Processing Systems*, vol. 32, Jun. 2019, Accessed: Jul. 18, 2024. [Online]. Available: https://arxiv.org/abs/1906.00446v1

[96]

A. van den Oord, O. Vinyals, and K. Kavukcuoglu, “Neural Discrete Representation Learning,” *Advances in Neural Information Processing Systems*, vol. 2017-December, pp. 6307–6316, Nov. 2017, Accessed: Jul. 18, 2024. [Online]. Available: https://arxiv.org/abs/1711.00937v2

[97]

J. Yu *et al.*, “Scaling Autoregressive Models for Content-Rich Text-to-Image Generation,” Jun. 2022, Accessed: Jul. 18, 2024. [Online]. Available: https://arxiv.org/abs/2206.10789v1

[98]

M. Ding *et al.*, “CogView: Mastering Text-to-Image Generation via Transformers,” *Advances in Neural Information Processing Systems*, vol. 24, pp. 19822–19835, May 2021, Accessed: Jul. 18, 2024. [Online]. Available: https://arxiv.org/abs/2105.13290v3

[99]

A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen OpenAI, “Hierarchical Text-Conditional Image Generation with CLIP Latents,” Apr. 2022, Accessed: Jul. 18, 2024. [Online]. Available: https://arxiv.org/abs/2204.06125v1

[100]

A. Ramesh *et al.*, “Zero-Shot Text-to-Image Generation,” *Proceedings of Machine Learning Research*, vol. 139, pp. 8821–8831, Feb. 2021, Accessed: Jul. 18, 2024. [Online]. Available: https://arxiv.org/abs/2102.12092v2

[101]

P. Sun *et al.*, “Autoregressive Model Beats Diffusion: Llama for Scalable Image Generation,” Jun. 2024, Accessed: Jul. 18, 2024. [Online]. Available: https://arxiv.org/abs/2406.06525v1

[102]

M. Tao, B. K. Bao, H. Tang, and C. Xu, “GALIP: Generative Adversarial CLIPs for Text-to-Image Synthesis,” *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2023-June, pp. 14214–14223, Jan. 2023, doi: 10.1109/CVPR52729.2023.01366.

[103]

M. Kang *et al.*, “Scaling up GANs for Text-to-Image Synthesis,” pp. 10124–10134, Mar. 2023, doi: 10.1109/cvpr52729.2023.00976.

[104]

X. Pan, T. Ye, D. Han, S. Song, and G. Huang, “Contrastive Language-Image Pre-Training with Knowledge Graphs,” *Advances in Neural Information Processing Systems*, vol. 35, Oct. 2022, Accessed: Jul. 18, 2024. [Online]. Available: https://arxiv.org/abs/2210.08901v1

[105]

M. Patel, C. Kim, S. Cheng, C. Baral, and Y. Yang, “ECLIPSE: A Resource-Efficient Text-to-Image Prior for Image Generations,” Dec. 2023, Accessed: Jul. 18, 2024. [Online]. Available: http://arxiv.org/abs/2312.04655

[106]

A. Radford *et al.*, “Learning Transferable Visual Models From Natural Language Supervision,” *Proceedings of Machine Learning Research*, vol. 139, pp. 8748–8763, Feb. 2021, Accessed: Jul. 18, 2024. [Online]. Available: https://arxiv.org/abs/2103.00020v1

[107]

J. Betker *et al.*, “Improving Image Generation with Better Captions”.

[108]

Bagrov, Keren, and Tymchenko, “Deci Diffusion 1.0.” Accessed: Jul. 17, 2024. [Online]. Available: https://deci.ai/blog/decidiffusion-1-0-3x-faster-than-stable-diffusion-same-quality/

[109]

M. Tao, H. Tang, F. Wu, X. Jing, B. K. Bao, and C. Xu, “DF-GAN: A Simple and Effective Baseline for Text-to-Image Synthesis,” *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2022-June, pp. 16494–16504, Aug. 2020, doi: 10.1109/CVPR52688.2022.01602.

[110]

L. Zhang, A. Rao, and M. Agrawala, “Adding Conditional Control to Text-to-Image Diffusion Models,” *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3813–3824, Feb. 2023, doi: 10.1109/ICCV51070.2023.00355.

[111]

M. Tahmid, M. S. Alam, N. Rao, and K. M. A. Ashrafi, “Image-to-Image Translation with Conditional Adversarial Networks,” *Proceedings of 2023 IEEE 9th International Women in Engineering (WIE) Conference on Electrical and Computer Engineering, WIECON-ECE 2023*, pp. 468–472, Nov. 2016, doi: 10.1109/WIECON-ECE60392.2023.10456447.

[112]

S. Zhou, K. C. K. Chan, C. Li, and C. C. Loy, “Towards Robust Blind Face Restoration with Codebook Lookup Transformer,” *Advances in Neural Information Processing Systems*, vol. 35, Jun. 2022, Accessed: Jul. 12, 2024. [Online]. Available: https://arxiv.org/abs/2206.11253v2

[113]

“Hello GPT-4o | OpenAI.” Accessed: Jul. 12, 2024. [Online]. Available: https://openai.com/index/hello-gpt-4o/

[114]

A. Radford *et al.*, “Learning Transferable Visual Models From Natural Language Supervision,” *Proceedings of Machine Learning Research*, vol. 139, pp. 8748–8763, Feb. 2021, Accessed: Jul. 12, 2024. [Online]. Available: https://arxiv.org/abs/2103.00020v1

[115]

H. Zhang, J. Y. Koh, J. Baldridge, H. Lee, and Y. Yang, “Cross-Modal Contrastive Learning for Text-to-Image Generation,” *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 833–842, Jan. 2021, doi: 10.1109/CVPR46437.2021.00089.

[116]

T. Xu *et al.*, “AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks,” *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1316–1324, Nov. 2017, doi: 10.1109/CVPR.2018.00143.

[117]

H. Zhang *et al.*, “StackGAN++: Realistic Image Synthesis with Stacked Generative Adversarial Networks,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 8, pp. 1947–1962, Oct. 2017, doi: 10.1109/TPAMI.2018.2856256.

[118]

H. Zhang *et al.*, “StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks,” vol. 2017-Octob, pp. 5908–5916, Dec. 2016, Accessed: Jul. 12, 2024. [Online]. Available: https://arxiv.org/abs/1612.03242v2

[119]

P. Dhariwal and A. Nichol, “Diffusion Models Beat GANs on Image Synthesis,” *Advances in Neural Information Processing Systems*, vol. 11, pp. 8780–8794, May 2021, Accessed: Jul. 10, 2024. [Online]. Available: https://arxiv.org/abs/2105.05233v4

[120]

F. Bie *et al.*, “RenAIssance: A Survey into AI Text-to-Image Generation in the Era of Large Model,” Sep. 2023, Accessed: Jul. 10, 2024. [Online]. Available: https://arxiv.org/abs/2309.00810v1

[121]

D. P. Kingma and M. Welling, “An Introduction to Variational Autoencoders,” *Foundations and Trends in Machine Learning*, vol. 12, no. 4, pp. 307–392, Jun. 2019, doi: 10.1561/2200000056.

[122]

D. P. Kingma and M. Welling, “Auto-Encoding Variational Bayes,” *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*, Dec. 2013, doi: 10.61603/ceas.v2i1.33.

[123]

I. J. Goodfellow *et al.*, “Generative Adversarial Networks,” *Science Robotics*, vol. 3, no. January, pp. 2672–2680, Jun. 2014, Accessed: Jul. 10, 2024. [Online]. Available: https://arxiv.org/abs/1406.2661v1

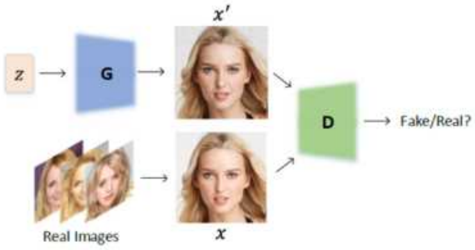
Appendix:

2.1 Image Generation Models:

Synthesising image is largely dependent on modern generative models. Apart from CNN and RNN based generative models such as PixelCNN and PixelRNN, mainstream generative models can range from Generative Adversarial Networks (GANs), Variational Auto-encoders (VAEs), Autoregressive Models (ARs) and Diffusion models.

2.1.1 Generative Adversarial Networks (GANs):

Generative Adversarial network (GAN) is an architecture that trains two models, the Generator (G) and the Discriminator (D), in a simultaneous manner (Goodfellow et al., 2014).

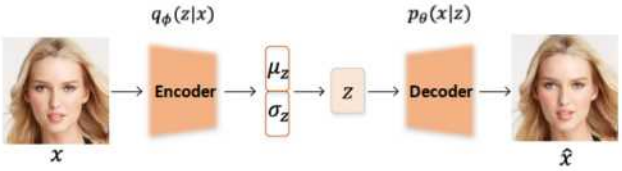


For image generation task, G is responsible for generating the image by capturing the distribution of the training image while D is required to determine if the sample image is generated or from the original training set (Goodfellow et al., 2014). In other words, GAN architecture is a two-player zero-sum game in which two players G and D manage to optimise their payoffs during training process (Goodfellow et al., 2014)

GANs preserve advantages and some inherent drawbacks for image generation. Goodfellow et al. (2014) stated that, firstly, the Markov Chains are not required for GANs in data sampling, thus having fair random draws for sampling; Secondly, the parameters of Generator (G) are not required for GANs, hence having lighter model size. Because of the GAN settings, the possibility of GAN collapse because of sampling data from noises and the implicit distribution of Generators are inherent disadvantages that are difficult to avoid (Goodfellow et al., 2014).

2.1.2 Variational Autoencoders (VAEs):

Variational Autoencoders (VAEs) consists of Bayesian distributions with reparameterization tricks and the encoder-decoder structure which can reconstruct input image data to a new generated version (Kingma & Welling, 2014). The encoder maps each data point x to the latent space z that follows a Gaussian distribution with mean and variance being generated by encoder (Kingma & Welling, 2014). The probability decoder, on the other hand, decodes the information from latent space z to a reconstructed image x' using Bayes rules (Kingma & Welling, 2014).



This setting enables VAEs to reduce dimensionality of input data and enables decoder of VAEs to generate images in a complex but reasonably fast manner, albeit sampling noises being introduced at training stage (Kingma & Welling, 2019). Although Bie et al. (2023) claims that the VAEs usually lose information when data are projected to the latent space with lower dimensions thus generating blurred images, yet the latent space structure has allowed VAEs to be integrated into image generation models as a tool for useful feature selection and computational cost reductions.

2.1.3 Autoregressive (AR) Models:

Zhang et al. (2021) clarified that models that regress a value of a variable on its previous values can be considered as Autoregressive (AR) models. For sequential data, estimating its conditional distributions of the variable at timestep t given previous information before t via calculating the conditional expectations will enable us to predict the value of the variable at next timestep (Zhang et al., 2021). In this regard, accompanied by Transformers, images may be generated by passing a sequence of predicted image tokens through an image decoder (Sun et al., 2024). When generating images, pixels of the image are firstly quantised and reshaped into a sequence of image tokens for model training, then the image tokens for evaluation are predicted by the trained AR model sequentially, finally the image will be generated by converting the predicted image tokens from some image tokens decoder (Sun et al., 2024).

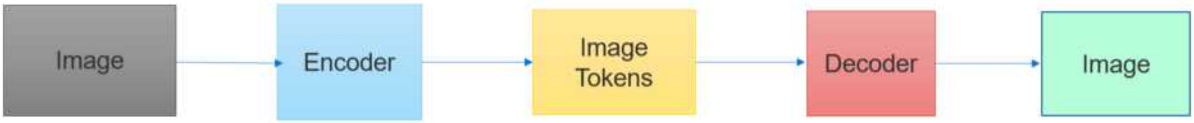
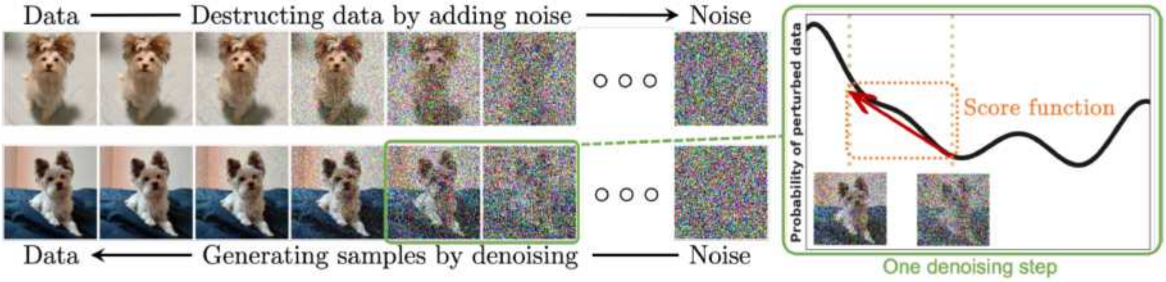


Image → Encoder → Image Tokens → Decoder → Image

2.1.4 Diffusion Models:

Nowadays, diffusion model has been dominating the visual generation in computer vision area (Sun et al., 2024). Yang et al. (2023) has summarised that the diffusion models are a group of probabilistic generative models which inject noises to destroy the input image gradually and then learn to reverse the process by gradually denoising the destroyed image thus generating a new image.


Data
Destructing data by adding noise
Noise
Data
Generating samples by denoising
Noise
Probability of perturbed data
Score function
One denoising step

According to Yang et al. (2023), diffusion models consist of three fundamental formulations: Denoising Diffusion Probabilistic Models (DDPMs), Score-based Generative Models (SGMs) and Stochastic Differential Equations (SDEs). The DDPM is composed of two Markov chains in opposite directions where the forward chain aims perturbing data into the noise and the reverse chain manages to reverse the noise to original data back (Yang et al., 2023). Apart from DDPMs, a Score-based Generative Model is also required for perturbing the image data with a series of Gaussian noises and for estimating the score function of noisy data distributions which will be used to guide the direction of denoising the perturbed image at each intermediate step (Yang et al., 2023). Moreover, when there are infinite timesteps or noise levels for diffusion models, DDPMs and SGMs can be generalised by a stochastic differential equation (SDE) whose roles are perturbing image data and denoising them by the guidance of the solutions to the SDE (Yang et al., 2023). These formulations are sufficient to establish a diffusion model, but they may lead to longer inference time and less stable models. To address these issues, a variety of methods have been further developed to facilitate model performances. To accelerate the denoising process thus rendering faster image generation, the U-net proposed by Ronneberger, Fischer & Brox (2015) and the Denoising Diffusion Implicit Models (DDIM) defined by Song, Meng & Ermon (2022) have been integrated into many diffusion models. U-net is a convolutional neural network used to deploy annotated image data more efficiently (Ronneberger, Fischer & Brox, 2015), while DDIM, on the other hand, is a Markov-chain-free and an alternative method for DDPM but with faster inference speed (Song, Meng & Ermon, 2022). When it comes to image generation stability, the concept of guidance for diffusion models which is based on the gradient of score functions have emerged to facilitate image generation. Dhariwal & Nichol (2021) firstly proposed classifier-guidance that can balance both the diversity and the image quality during denoising process. Then Ho & Salimans (2022) further introduced Classifier-Free Guidance which preserves the same capabilities with Classifier-Guidance but with less computational cost. With the help of guidance techniques, users can generate and modify high fidelity images with significantly reduced manipulations (Meng et al., 2022). Furthermore, to make the diffusion models more flexible and easier to deploy, Zhang, Rao & Agrawala (2023) proposed the ControlNet architecture which allows multiple inputs such as text prompts, human pose sketch, segmentation maps and images to be passed through diffusion models thus offering more flexible and controllable image generation capabilities.