Cover Page..

Abstract:

[remember stating research gap here]

Table of Contents here

1. Introduction
2. Related Work
3. Method
4. Experimentation
5. Results and Analysis
6. Conclusion and Future work

1.Introduction:

1.1 Overview

The Artificial Intelligence Generated Content (AIGC) has been one of the most popular topics in the era of AI technology expansion. Cao et al. (2023a) has claimed the AIGC can be defined as the content synthesised automatically by human instructions with Generated AI tools within short period of time. In recent years, the form of AIGC has been spanned from various areas such as textual-based contents, images and audios with the advancement of corresponding generative AI models. 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) are representative examples of Generative models used for AIGC generation. Among these forms of AIGC, image generation has been an intriguing topic that has attracted enormous attentions from both academia and industries. In the industry, huge costs for marketing purposes incur every year for companies. In 2023, more than 1000 billion US dollars have been spent in the advertisement sector in the world (Statistica, 2024). Therefore, it has naturally rendered us to explore the possibilities of synthesising images under business contexts such as generating company advertisement and posters for marketing campaigns or generating sample images in the photography briefings for freelancer photographers using image generation models. In recent years, previous research rarely focused on comparing different diffusion-based image generation models when generating images for marketing purposes nor they provide a unified framework for generating these images required by marketing teams from scratch. Hence, combining the natural interests and the research gap, this report will aim at answering the question of whether Generative AI models can synthesise high quality images for marketing purposes and will provide a framework of choosing suitable image generation models for users under different marketing requirement.

1.2 What is Image generation?

Image generation is one of the most important tasks in Computer Vision (CV) which utilises various forms of input such as texts, images and audios to generate new images (Elasri et al, 2022). Among image generation subtasks, text-to-image (T2I) generation and image-to-image (I2I) translation has been attracting enormous attention in recent years. Text-to-Image (T2I) is a task that converts textual descriptions into images with high qualities (Bie et al., 2023), while the Image-to-Image (ITI) translation is a method that edits input images conditioned on textual input (Isola et al., 2018).

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

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.
* **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 will contain 6 main 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 results and analysis for image generation models.

Section Five is the application of the framework in generating a series of images for marketing campaign based on customer segmentations.

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

**Appendix** will list detailed results from the experimentation.

2.Related Work:

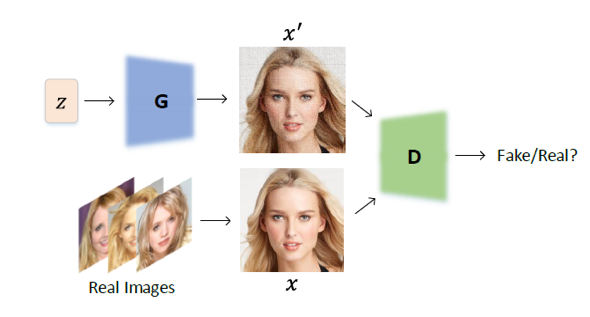
2.1 Image Generation Models:

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.

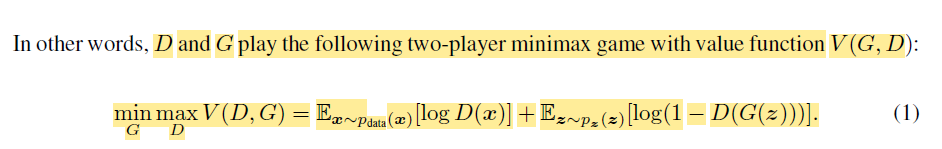
[Table 1: advantages and disadvantages of image generation models]

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 by the following formula (Goodfellow et al., 2014):

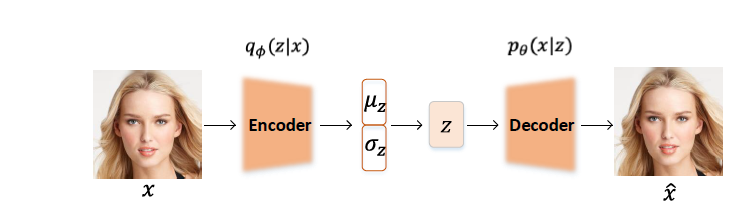


Where V(D,G) is the value function of the game.

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):

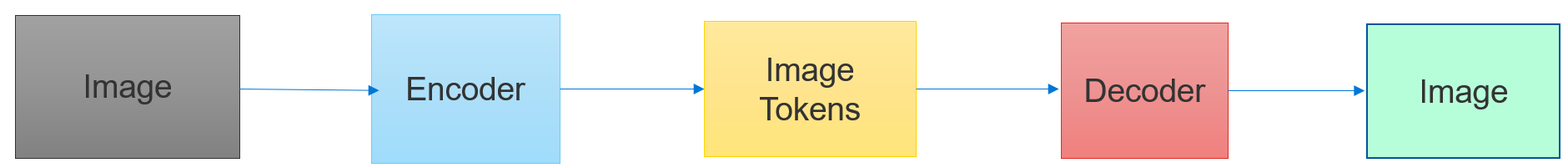
Variational Autoencoders (VAEs) consists of Bayesian distributions with reparameterization tricks and the encoder-decoder structure which can reconstruct input image data to a new generated version (Kingma & Welling, 2014). The encoder maps each data point x to the latent space z that follows a Gaussian 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.

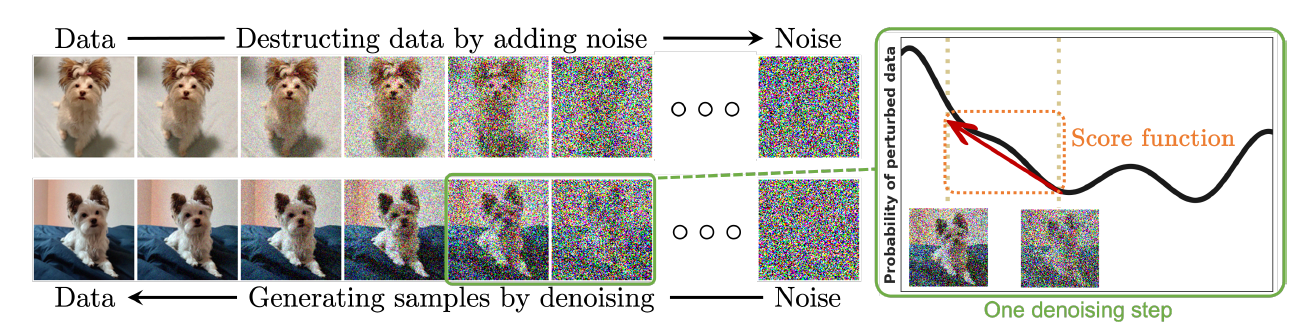
2.1.3 Autoregressive (AR) Models:

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).



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.



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.

[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 family have been playing a pivotal role and are continuously impacting T2I tasks. Based on theoretical foundations established by Goodfellow et al. (2014), numerous GAN-based T2I models emerge in recent years. Recent GAN-based models can mainly be categorised into multi-stage GANS, single-stage GANs and scaled-up GANs with large pre-trained model or large dataset. In 2017, Zhang et al. (2017) proposed StackGAN, a GAN-based T2I model that generates the low-resolution images with only rough sketch of basic shapes at first stage and then output the higher resolution image (256 x256 pixels) conditioned on original text and the sketch image generated previously. Zhang et al. (2017) observed their innovative settings can generate high-resolution image with photo-realistic attributes, corrects minor defects of images generated in previous stage, and achieves the FID scores of 74.05 in COCO dataset. However, the stacked architecture will be inevitably affected by the GAN-based model collapse. Because the StackGAN is conditioned on the low-resolution image generated in the first stage, the collapse in stage one will lead to the collapse of StcakGAN as a whole (Zhang et al., 2017). To mitigate the risk of collapse, Zhang et al. (2018) proposed another StcakGAN++ model which deployed tree-like multi-stage GAN architecture for T2I tasks. According to Zhang et al. (2018), StackGAN++ consists of StackGAN-v1, which is the same with StackGAN, and StackGAN-V2, a multi-stage GAN with multiple Generators and Discriminators that generate the image at varying scales. In addition, Zhang et al. (2018) also introduced colour-consistency regularisation at each stage in StackGAN-V2 to ensure that colours are consistent among different scales thus improving the overall image qualities. By leveraging the multi-stage setting and the attention mechanism, Xu et al. (2018) have established 230-million-parameter-model AttnGAN consisting of attentional generative network and a deep attentional multimodal similarity model (DAMSM). According to Xu et al. (2018), the attentional generative network enables AttnGAN to output images with fine-grained details through a multi-stage refinement process, while the DASMS facilitates the training process of the Generator by a text encoder with Long-Short Term Memory (LSTM) and a CNN based image encoder. This novel GAN-based model has achieved significant improvements in Inception Scores (IS) on COCO and CUB datasets, showing the capabilities in generating high quality images (Zhang et al., 2018). The performances of AttnGAN have been further verified by (Zhang et al., 2021), achieving the FID score of 35.49 on COCO dataset, which is 56% reduction compared to that of StackGAN++. Despite these cheerful capabilities preserved by previously discussed multi-stage GANs such as generating high-resolution fine-grained images, the multi-stage setting may affect model performances as well. Tao et al. (2022) has summarised three limitations of GAN models with stacked architectures: Firstly, the multi-stage GANs may induce entanglements among different Generators that could lead to the refined output image being a combination of simple details and blurred shapes; Secondly, GANs in stacked architectures such as StackGAN++ intend to deploy extra networks to fix faulty image parts generated in previous stages for semantic consistency between the text and the image, which may limit the supervision capabilities of these networks; Finally, the attention-based approach that previous models such as AttnGAN emphasised on may only be effective for a certain image scales because of the computational cost. To tackle these inherent limitations, Tao et al. (2022) proposed a 19-million-parameter model DF-GAN which is smaller in size but more efficient than precious models. DF-GAN is a single-stage T2I that generates high-resolution images without entanglements with other Generators (Tao et al., 2022). Although it is simpler and has significantly smaller size than other GAN-based models, yet it has achieved better results in COCO dataset compared with AttnGAN and XMC-GAN (Tao et al., 2022). Indeed, DF-GAN is simpler and smaller in size, but this is at the cost of only introducing sentence level text information, hindering the capabilities of image synthesis at fine-grained level (Tao et al., 2022). To balance the model complexity and its sensitivity to textual input, research in GANs with contrastive learning approach has been extensively conducted nowadays. Previous T2I models have been long emphasising on crafting the overall GAN structures such that improving overall performances, which has been proved feasible. However, Zhang et al. (2022) proposed XMC-GAN empowered by contrastive learning for better language-image alignment and they displayed how the single-stage XMC-GAN can outweigh other GAN-based models by the approach of maximising the mutual information between language and image. According to Zhang et al. (2022), the XMC-GAN utilises the losses from contrastive learning and has achieved strong coherence between generated image and text prompt while maintaining high-level photo-realistic output. Compared to AttnGAN and DF-GAN, XMC-GAN further improves its FID score to 9.33 on COCO dataset, achieving SOTA results at the time of model being released (Zhang et al., 2022). These advances are key steps for generating images by natural language descriptions hence guiding the future T2I research (Zhang et al., 2022). Inspired by XMC-GAN, Tao et al. (2023) proposed GALIP that leverages the power of a pretrained contrastive learning language image model (CLIP) proposed by Radford et al. (2021). Empowered by CLIP, the GALIP model has achieved comparable results with modern AR and diffusion models but with approximately 120 times faster inference times, 33 times less data required for training and 16 times smaller model size measured by parameters (Tao et al., 2023).

Finally, current research also discovers the feasibility of directly deploying extremely large dataset for training GAN-based models. In 2023, Kang et al. (2023) introduced 1-billion-parameter GigaGAN trained by LAION2B-en (Schuhmann et al., 2022) dataset containing 2 billion image and text descriptions. GigaGAN has enabled high-quality image generation in a rather fast pace, albeit increased model training times (Kang et al., 2023).

Despite unique advantages preserved by GANs, its drawbacks such as capturing less diversity than VQ-VAEs (van den Oord, Vinyals and Kavukcuoglu, 2017), challenging model training and possibility of model collapse have made scientists to research on other types of 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 models can also be deployed for text-to-image generations with some modifications, although they are predominately designed for text-related tasks. According to Bie et al., (2023), AR models with Transformers integrated generate images by predicting image tokens in a sequential manner conditioned on text-encoder-generated text tokens in a T2I task. This means our goal is to firstly use text prompt to predict image tokens sequentially, then pass the predicted image tokens into a decoder thus generating a new image. Therefore, we may divide the AR model into 3 parts: a text encoder that will transfer texts into text tokens, a prior that will deploy text tokens to generate image tokens, and a decoder that will convert image tokens into an image.

Recently, a number of AR models have been developed to cope with T2I tasks with innovations of AR models. In 2021, Ramesh et al. (2021) proposed an unprecedented T2I AR model with Transformers approach, DALL-E, whose objective is to train a transformer-based AR model that treats the language tokens and image tokens as a single series of tokens data. According to Ramesh et al. (2021), DALL-E contains 12 billion parameters and has achieved zero-shot FID score of 17.9 on COCO dataset. Moreover, specialised Human Evaluation metrics is also applied to DALL-E in which participants reckon images generated by DALL-E are more photo-realistic and match captions better more than 90% of time during evaluation (Ramesh et al., 2021). In order to improve training efficiency without hurting output image quality, Ramesh et al. (2021) have defined two training stages for DALL-E: At the first stage, a discrete variational autoencoder (dVAE) has been trained as image encoder to reduce the original image size and to convert the compressed image in to a 32 x 32 image tokens; At the second stage, the text descriptions found by CLIP that match the input image the most have been encoded by BPE encoder and has been concatenated with image tokens to a single stream, and finally the transformer has been trained by the single stream of data as the prior to process these tokens and decode them into generated images. At inference stage, the DALL-E can generate images with text prompt only and has even gained zero-shot capabilities for T2I tasks, indicating the DALL-E equip with the potential to generate images they did not see before, making it more competitive than models with domain knowledge (Ramesh et al., 2021). However, Ramesh et al. (2021) also claims that the zero-shot approach may not be rather effective for specialised dataset such as CUB, an image dataset for birds. Therefore, fine-tuning may be required for improvement in zero-shot performances (Ramesh et al., 2021). In the meantime, Ding et al. (2021) also proposed CogView, a 4-billion-parameter AR model specialised for T2I tasks. Unlike DALL-E, CogView have utilised VQ-VAE for generating image tokens, which helps CogView achieves SOTA FID on blurred COCO dataset and renders it to outperform DALL-E and previous GAN-based models such as DF-GAN and AttnGAN while maintaining zero-shot capabilities that DALL-E preserves (Ding et al., 2021). Despite the improved performances in terms of FID score, images generated by CogView may preserve blurriness because of the VQ-VAE architecture being used as image encoder (Ding et al., 2021). To enable the AR models to generate more photo-realistic images, Yu et al. (2022) has come up with Pathways Autoregressive Text-to-Image (Parti) model for high fidelity image generations. Parti is a 20-billion-parameter AR model which deploys ViT-VQGAN as image encoder to generate image tokens while following the training process of DALL-E (Yu et al., 2022). This slight change made by Yu et al. (2022) has enabled Parti to achieve SOTA zero-shot FID as well as SOTA fine-tuned FID on COCO dataset and has made Parti to be comparable to diffusion-based model Imagen. Additionally, the newly introduced holistic benchmark PartiPrompts shows the efficiency of Parti when dealing with challenging tasks (Yu et al., 2022). However, Yu et al. (2022) also have clarified that errors such as colour blending, omissions and wrong object counting are likely to occur for Parti as the prompt becomes longer and more complex.

Other than newly proposed models such as DALL-E or Parti, recent research also pays attention to the innovations in priors of AR models. In 2021, Radford et al. (2021) introduced the Contrastive Language-Image Pre-training (CLIP) model whose objective is to predict the correct image-text pairings when a batch of images and text prompts are given thus facilitating image generation and captioning tasks. According to Radford et al. (2021), CLIP surprisingly preserves the zero-shot capabilities and is one of the key steps to zero-shot classifiers of computer vision that are both feasible and adaptive. However, since CLIP is trained on online collected data that are neither filtered nor curated, social biases and unethical contents may be generated. In addition, Pan et al. (2022) argues that the CLIP is only trained with simple image-text pairs that ignore the semantic connections. Therefore, they further introduced Knowledge-CLIP which feeds sematic information into CLIP models thus making Knowledge-CLIP semantically align both vision and language representations and enhancing the capability of reasoning among different scenarios (Pan et al., 2022). From our best knowledge, no open-source text-to-image generation models have now utilised Knowledge-CLIP to improve image qualities. However, this will be left for future research as the semantic meanings can potentially refine generated images.

Up to now, we have shown the unprecedented zero-shot capability the AR models possess, yet there are tangible limitations for AR models to discuss as well. To start with, many AR models are slow in generating images because these models generate images token-by-token (Ding et al., 2021). In addition, the longer text prompt not only will increase inference time, but it is also more likely to makes error for AR models like Parti (Yu et al., 2022). Meanwhile, the number of parameters for AR models are significantly larger than GAN-based and Diffusion-based models, resulting higher hardware requirement to implement AR models. As shown in Table 1, DALL-E contains 12 billion learnable parameters while the diffusion model DALL-E 2 designed by the same company has 6.5 billion learnable parameters, indicating the nature of large size for AR models. Moreover, the dataset used for training is another issue that raises concerns. Since the dataset for AR models is mainly from the internet, images and text within the dataset are neither filtered nor curated, making the risk of generating unethical or not suitable for work (NSFW) contents inevitable (Yu et al., 2022).

2.2.4 T2I with Diffusion Models

One of the most prevalent types of Text-to-Image generation models is diffusion model. Popular image generation models such as Stable Diffusion, DALL-E 3 and Midjourney are all diffusion-based. In recent years, many diffusion-based text-to-image models have emerged by leveraging contemporary methods mentioned in Section 2.14 such as U-Net, DDIM, Guidance methods and ControlNet. In early stage, Nichol et al. (2022) introduced the Guided Language to Image Diffusion for Generation and Editing (GLIDE) system for T2I generation tasks. The GLIDE that deploys classifier-free guidance preserves the abilities in generating photo-realistic images in zero-shot fashions and outperforms the AR-based DALL-E in terms of human evaluations (Nichol et al., 2022). At the same time, Saharia et al. (2022) proposed Imagen model which can generate photo-realistic images of unprecedented level through the combination of the empowerment of the Large Language Model T5 and the diffusion model. According to Saharia et al. (2022), Imagen has achieved SOTA FID score of 7.27 in MS COCO dataset and has outperformed DALL-E 2 under DrawBench human evaluation. More recently, diffusion models have been developed in different branches. One branch worth mentioning is diffusion models with Prior-Decoder architecture. In 2022, Ramesh et al. (2022) introduced DALL-E 2, a diffusion model that incorporates a diffusion-based prior for CLIP image embeddings and a decoder for generating image conditioned on image embeddings from the prior. This innovative architecture has made DALLE-2 achieved comparable image quality to GLIDE but with higher diversity (Ramesh et al., 2022). Similarly, Zhou et al. (2023) also introduced Corgi, a diffusion model that encodes the knowledge of CLIP seamlessly and is trained by a dataset with less than 2% of images contain associated textual descriptions. The Prior-Decoder architecture setting has enabled Corgi to achieve comparable FID to SOTA models (Zhou et al., 2023). Mixture of Expert (MOE) Diffusion is another branch for T2I generation tasks. Feng et al. (2023) proposed ERNIE-VILG 2.0 model, a T2I diffusion model that combines fine-grained visual and text-based information and deploys several distinct denoising experts at different stages for denoising process. The MOE setting has enabled not only high-fidelity image generation, but it also has enabled higher text and image alignment (Feng et al., 2023). Similarly, Balaji et al. (2023) also introduced eDiff-I model which utilises a series of expert denoisers specialised at different stages to enable the model to rely on textual information more at early denoising stages and visual information at later stages. In addition, the eDiff-I utilises several encoders such as CLIP and T5 to achieve multiple conditional embeddings, facilitating to generate images with different artistic styles (Balaji et al., 2023). Another branch for diffusion model research is Retrieval-Augmented (RA) Diffusion models. Blattmann et al. (2022) introduced the retrieval strategy and they later applied this to the Latent Diffusion Model (LDM) proposed by Rombach et al. (2022). Blattmann et al. (2022) has demonstrated that the retrieval strategy can significantly improve the model performances, particularly image quality and diversity. With this idea, Chen et al. (2022) proposed Re-Imagen model which is augmented by the summarised semantic information and the detailed visual knowledge thus achieving improved visual appearance accuracy of different components in synthetic images. Diffusion models discussed so far are mainly conditioned on textual input. In order to gain more controls at fine-grained level for both the semantic information and shapes of different parts in synthetic images, Avrahami et al. (2022) proposed SpaText method which utilises the segmentation map and text prompts as input for diffusion models. This method has enabled the diffusion models to manage complex image generation tasks under different conditions (Avrahami et al., 2022).

Like GAN-based image generation methods, diffusion models can also leverage the advantages of other generative models. Yin et al. (2024) have proposed the Distribution Matching Distillation (DMD) method which leverages the GAN architecture to change the multi-stage diffusion models into a single-step generator hence accelerating image generation process while preserving the high quality. Gu et al. (2022) also proposed VQ-Diffusion model that leverages VQ-VAEs and a conditional DDPM thus reducing the unidirectional biases and accumulation of errors.

So far, we have discussed different types of diffusion models such as Diffusion with Prior-Decoder architecture, Mixture-of-Expert Diffusion, Retrieval-Augmented Diffusion, and Diffusion with other generative model architecture. However, one of the most important research branches in diffusion models may be the Latent Space Diffusion, a latent-space architecture that has been widely adapted by Stable Diffusion model family. In 2022, Rombach et al. (2022a) proposed a novel Latent Diffusion Model (LDM) that executes image generation process over a compressed latent space with cross-attention mechanism. This method has significantly reduced the computational cost for training diffusion models without compromising the quality of synthetic images (Rombach et al., 2022a). Because of this innovative architecture setting, models such as Stable Diffusion version 1 family (SD1) and version 2 family (SD2) have emerged in recent years. Based on this fundamental model setting, advanced variants of Stable Diffusion models also come into place. In 2023, Deci.ai (2023) proposed a Deci Diffusion 1.0 model that leverages the newly designed U-Net and uses a smaller number of parameters to achieve approximately 3 times faster inference time than SD 1.5 models while maintaining the comparable qualities. This change has enabled Deci Diffusion 1.0 to realise similar FID score to SD 1.5 model but allows the image generation cost to be reduced by 66% (Deci.ai, 2023). Meanwhile, Betker et al. (2023) also proposed DALL-E 3, a T2I model that understands prompt instructions better than previous versions and can generate images with higher quality given the same prompt. At the same time, Podell et al. (2023) proposed 2.6-billion-parameter SDXL, a new Stable Diffusion model that employs a larger U-Net and more attention blocks compared to SD1 and SD2 at the cost of significantly increased computational complexity for training (Podell et al., 2023). According to Podell et al. (2023), this model has consistently defeated all previous Stable Diffusion models based on human evaluation and has achieved comparable results compared to black-box image generation model such as Midjourney v5.1. More recently, Esser et al. (2024) proposed a new innovative approach to further improve both the quality and the efficiency of generating high-resolution images via Rectified Flow (RF), a new transformer-based architecture that separates the weights for text and image modalities thus enabling bidirectional information flows. The most advanced Stable Diffusion models such as Stable Diffusion 3 Large (SD3 Large) and Stable Diffusion 3 Medium (SD3 Medium) are largely based on this innovative approach. Furthermore, Sauer et al. (2024) also proposed the Latent Adversarial Diffusion Distillation (LADD) which can not only simplify model training process, but it also improves the overall performances. Combining LADD (Sauer et al., 2024) and RF (Esser et al., 2024), Sauer et al. (2024) introduced the Stable Diffusion 3 Turbo (SD3 Turbo) which can generate ultra-high-quality images in a fast manner.

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]

2.3 Image-to-Image (I2I) Generation

**2.3.1 I2I Background**

Pang et al. (2022) has summarised the I2I models can handle various computer vision tasks such as image-inpainting, super resolution, style transfer and image extension, etc. In the last decades, various I2I models are invented based on different architectures. At early stage, GAN-based I2I models have dominated various image editing subtasks. Isola et al. (2018) claims that GANs can act as general-purpose solutions for I2I translation tasks. (Gonzalez-Garcia, Weijer and Bengio, 2018), CycleGAN (Zhu et al., 2020), BicycleGAN (Zhu et al., 2018) and (Kazemi et al., 2018) are GAN-based models used for semantic synthesis. (Pathak et al., 2016), (Zhu et al., 2018) and (Song et al., 2018) are used for inpainting tasks. In addition, (Hertzmann et al., 2001) can transfer the style of images while (Ren, Romano and Elad, 2016) is used to image super resolution.

Nowadays, I2I models have emphasised more on Diffusion-based models. However, GANs, Transformers, and VAEs still play important roles in improving diffusion model performances. They are widely used for restoring faces, artifacts and eyes in diffusion models. A telling example is the transformer-based model CodeFormer (Zhou et al., 2022) which is used to fix artifacts and blurriness of faces generated by Stable Diffusion model.

Moreover, since T2I models integrate image encoder and image decoder into their architecture, some of T2I models can theoretically solve I2I tasks. Some models we have discussed in Section 2.2 also preserve image-to-image translation capabilities. For instance, GLIDE proposed by Nichol et al. (2022) can solve inpainting tasks while AttnGAN can refine synthesised images to improve image quality (Xu et al., 2018). Similarly, large-scale models such as the Stable Diffusion moedels (Rombach et al., 2022a) are able to conduct inpainting tasks, SDXL (Podell et al., 2023) could act as an image refiner and eDiff-1 (Balaji et al., 2023) can transfer the style of images. Furthermore, by leveraging SDEdit proposed by Meng et al. (2021), SDXL can handle inpainting tasks as well.

Despite the ability of existing large-scale text-to-image diffusion models being able to generate high quality images from detailed textual descriptions, they often lack the ability to precisely edit the generated or real images (Mou et al., 2023). Therefore, we will mainly explore the latest diffusion based I2I models that preserve certain image editing capabilities. Huang et al. (2024) has researched over 100 diffusion-based I2I models that can edit images in terms of semantic editing, stylistic editing and structural editing. In each part of image editing, Huang et al. (2024) further divided it into smaller subtasks such as adding object or removing object in the semantic editing. In the following sections, we will classify I2I models based on their capabilities by following the template of Huang et al. (2024)



**2.3.2 Semantic Editing**

In terms of semantic editing, Huang et al. (2023) concludes it can be further divided into object addition, removal, replacement, emotional change and object colour change. We found that many diffusion models can handle these tasks. In 2023, Li Singh & Grover (2023) proposed the flexible multi-modal instruction-following system InstructAny2Pix that can handle object addition, removal and replacement tasks and supports various instructions such as prompt, audios and images. Later, Yang et al. (2024) proposed ImageBrush that only uses visual instructions such as segmentation maps, poses and reference images. According to Yang et al. (2024), this method can better capture human intentions and facilitate real-world applications. At the same time, Li et al. (2024) also introduced Zero-Shot Instruction-Guided Local Editing, a method that enables precise and localized editing without the need for detailed mask images or prompts. This method can solve several subtasks such as change of image background and object removal (Li et al., 2024). Furthermore, Sheynin et al. (2024) also proposed the Emu Edit, a multi-task image editing model using text input only that is designed to perform a wide range of editing tasks using natural language instructions. Emu Edit innovatively introduced the learned task embeddings to its architecture hence improving its few shot capabilities. From our best knowledge, Emu Edit is the only model that can handle all 5 subtasks that Huang et al. (2024) defined.

**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).

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:**

The distribution-based metrics have been widely used when evaluating the general performances of an image generative model. The idea behind the distribution-based metric is that one can firstly derive two datasets by converting real world images and synthesised images into high-dimensional vectors. Vectors in each dataset can be fitted into some statistical distributions such as multi-variate Gaussian distribution under normality assumption. Then by comparing the similarities or distances of these two distributions, we can evaluate the overall quality of images generated by proposed model. In 2016, Salimans et al. (2016) proposed the Inception Score (IS) metric to evaluate the quality and diversity of images generated by image generation models. According to Salimans et al. (2026), the higher IS may indicate the higher performances of that image generation model. However, Heusel et al. (2017) noted drawbacks of IS that they did not consider real-world images into account when comparing the distributions. Therefore, Heusel et al. (2017) further proposed the Fréchet Inception Distance (FID) score that compares the distributions of ground-truth images and synthetic images. Heusel et al. (2017) claimed that the lower the FID indicates the better overall model performances. Nowadays, FID has been one of the most deployed quantitative metrics for image generation models. FID is often used when a new model being proposed, and many comparisons of different models are based on FID scores. However, there are also limitations to FID score. One of the most crucial limitations is that real-world images may not follow normality assumptions due to the bias and privacy issues for the dataset used for training (Jayasumana et al., 2024). In addition, Otani et al. (2023) and Jayasumana et al. (2024) also criticised that FID is inconsistent with human evaluations when assessing the quality of images. Inspired by traditional machine learning, Kynkäänniemi et al. (2019) proposed Improved Precision and Recall (PR) for image generation models, and they found PR is consistently better than FID score. By comparing the overlap of image distributions, one can derive precision rate and recall rate for the model hence indicating the diversity and quality of images synthesised by the model. According to (Kynkäänniemi et al., 2019). According to Kynkäänniemi et al. (2019), high quality but low diversity images being generated will lead to high precision and a low recall.

[Draw a diagram]

**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) has estimated the market share of AI in marketing is expected to increase to $107.5 billion by 2028. At the same time, Dencheva (2023b) also claims that a survey conducted by market practitioners in the US suggests more than 70% of respondents deploy AI tools such as chatbot as part of their company work. Driven by the popularities of AI in marketing sector, researchers have been exploring the feasibility of using generative AI models for marketing purposes. There are a bunch of real-world applications of using AI to generate marketing contents. To start with, the large foreign language learning platform, Duolingo, has utilised GPT-4 to develop an AI tool called Max which can practise spoken skills with users through role play and explain answers in the learning process (Kshetri, 2023). Then the Goosehead Insurance company has utilised the Jasper.ai platform to create blog articles for marketing campaign (Kshetri et al., 2024). Apart from textual contents, there are also examples using generative AI tools to create visual marketing contents. In 2023, Mehta (2023) states that the toy company Mattel has deployed the DALL-E 2 to create a model car toy. At the same time, Jones Road Beauty company employs Meta’s AI sandbox to generate more versions of advertisement in a faster pace (Adams, 2023), while PrimeCare company uses Midjourney to synthesise art contents for their blogs (Kshetri et al., 2024). Previous examples mainly use black-box models that are backboned by other base models. To assess whether we could generate high quality images for marketing purposes, Zhang et al. (2024) analysed a survey regarding personalised content generation using diffusion models. Zhang et al. (2024) claimed that diffusion models can potentially generate personalised content with the trade-off between textual alignment and image fidelity. This research has bridged the feasibilities of synthesising marketing visual contents and using diffusion models, strengthening our confidence for this project. With this idea, Hartmann, Exner & Domdey (2024) has systematically compared diffusion models such as DALL-E 3, Midjourney v6, Firefly 2, Realistic Vision and SDXL Turbo based on image quality, realism and aesthetics using human evaluations to answer whether high quality images can be generated by diffusion models and hence being used for marketing purposes. They managed to answer this question by firstly answering whether generated images can surpass human perception with a hypothesis testing, then answering whether T2I models can reach similar performances compared to images shot by humans with another hypothesis testing, and finally answering whether generated images can improve website click-through rate via A/B testing (Hartmann, Exner & Domdey, 2024). According to Hartmann, Exner & Domdey (2024), they found while all models surpass human-made images in terms of aesthetics, only Realistic Vision, a model backboned by SD1.5, can produce images with high realism and SDXL Turbo is inferior to real images in terms of realism. Combining all results, Hartmann, Exner & Domdey (2024) conclude that DALL-E 3 is overall the best model in terms of generating marketing contents and can distinctly increase more click-through rate than other models in the A/B testing. Although Hartmann, Exner & Domdey (2024) has compared these diffusion models under marketing context, there are some limitations as well. Firstly, their research did not take I2I models into account. When freelancers produce advertisement, they may edit the original image multiple times to achieve best quality. Therefore, I2I models should also be considered when assessing generated image quality. Secondly, this paper only uses human evaluations instead of quantitative metrics when examining the synthetic images, meaning that the bias may exist, and results may not be robust.

2.6 Literature Review Conclusion:

So far, we have conducted in depth literature review regarding various text-to-image generation models, image-to-image models, evaluation metrics and datasets for image generations, and the image generation applications in marketing sector.

[Visualisation of literature review]

We have discussed a variety of T2I models with different architectures such as GAN-based models, autoregressive-based models and diffusion-based models. We also explore Image-to-Image models, we have explored various GAN-based and diffusion-based models that preserve the capabilities in image semantic editing, stylistic editing and structural editing.

Based on Table 1 and Table 2, we noticed that the diffusion-based models are consistently outperforming other types of models at the cost of increased size of parameters. Additionally, the Hugging face leader board also verifies the conclusion from out literature review. Therefore, intended users may choose diffusion-based models when generating marketing-oriented images and diffusion models will be selected in 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 | SDXL-refiner | Azure Machine Learning |
| Stable Diffusion v2.1 inpainting |
| Stable Diffusion v1.5 inpainting |
| SDXL - inpainting | dreamstudio.ai [] |
| Multimodal | GPT4o | OpenAI |

[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]

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 images (Figure 1) 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. In our initial testing, we found that adding keywords can substantially improve the image quality without changing the model architecture and we realised the importance of properly designing the prompt before image generation.

[Student graduation here]

[Turbines here]

Albeit these powerful prompt engineering techniques, we will combine keywords and main prompt as our textual input for candidate models considering the limited time for this project.

**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, the Table 5 illustrates the useful keywords and negative prompt. We will mainly use these keywords to properly design our prompts when testing each candidate models.

**3.1.3. Evaluation Metrics:**

Since we have discussed the quantitative metrics in the literature review, we will use CLIP-IQA as our quantitative metrics. For this project, we will choose “natural”, “quality” and “relaxing” as the aspects being assessed. Here the “natural” aspect examines whether the image is real or generated by AI, “quality” aspect assesses the overall quality of images and “relaxing” tests if the image gives a relaxing feel for audience.

Moreover, inspired by Chatbot Arena, we are using pairwise comparisons over generated images. We will firstly randomly select 2 images per time to ask respondents to choose a preferred one and then record all the responses. 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.

The methodology will consist of 3 parts. In the first part, we will test chosen T2I models and I2I models by generating a bunch of images and evaluating image qualities using CLIP-IQA and human evaluation. 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 by both CLIP-IQA and human evaluation. 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**

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 via Image Masker and design prompt for inpainting tasks before passing selected images into inpainting models.

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.

[https://imagemasker.github.io]

**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. However, the advantage for refiner is that they do not require mask images. To test refiners, we will choose original stock images and design prompts to see if these stock images can be edited by I2I refiners. Evaluation will also be given after editing all required images.

**3.3.3 Testing T2I & I2I Model Combinations**

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 CLIP-IQA and pairwise human evaluation to assess overall quality of assessed images.

**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 correct these imperfections before refining the images. Thus, images edited in section 3.3.2.1 will be processed through refiners to evaluate any improvement in overall quality.

[summary table here]

**3.3 Methodology for Framework setup:**

Combining all results from section 3.1 and 3.2, we will form our framework in the following structure.

**A screenshot of a computer

Description automatically generated**

**3.4 Application: (Customer Segmentation)**

**Step 1: Generate Prompt:** One main prompt, several sub-prompts for customer segmentation (age group, ethnicity group)

**Step2: Generate images using 2 T2I models**

**Step3: Editing images using I2I models**

**Step4: Human Evaluation + Quant Evaluation , conclude whether Gen Image can do customer segmentation.**

4. Experimentation

4.1 Experimentation on Keywords & Prompts Style:

We firstly test keywords to see if they are useful.

Then we test prompts style (front, middle,…etc)

We conclude using main in middle is suitable for this project.

However, due to limited testing, may be wrong.

3.2.1 Testing Keywords:

Based on marketing team requirement, we summarised keywords and will test them in SD, DALLE…

How to test keywords? Just use the single keyword as prompt to generate image. Then look at images will let us know if models have such capability.

3.2.2Testing Prompts:

Test on DALLE, SD2, using landscape, (DALLE 1024x, vivid, HD) (SD2: 768x, 50 iter and 100 iter)

Testing keywords + main prompt

Testing different positions

Testing Long Prompt:

Testing with negative prompt,

Testing negative prompt with 100 iterations

4.2 Experimentation for T2Is

4.2.1 Testing SD1.4 & SD1.5:

When testing landscape, we found SD1.4 and SD 1.5 have poor performance. We will not use SD1.4 and SD1.5 for its low resolution and low authenticity. (Mention Azure environment I’m using, confidential data, challenges in VMs..security...)

4.2.1 Experimentation on Landscapes:

[selected picture here]

[evaluation results here]

4.2.2 Experimentation on Human:

4.2.3 Experimentation on City:

4.2.4 Experimentation on Illustration:

4.2.5 Experimentation on Words:

4.2.6 Experimentation on Counting:

4.2.7 Evaluations of I2I:

4.2.7.1 Quantitative Evaluation

4.2.7.2 Human Evaluation

4.2.8 Discussion:

From our experimentation, we noticed that the a,b,c,d…

[A summary table specifying pros and cons of T2I models]

Using T2I models available may not be able to generate high quality images. Therefore, we are considering if editing them using I2I would help.

4.3 Experimentation on I2Is

We screen out SD1.5 inpainting here.

4.3.1 Experimentation for SD2.1 Inpainting

4.3.2 Experimentation for SDXL refiner:

4.3.3 Experimentation for SD3L refiner:

4.4 Experimentation for combined Inpainting & Refiner:

We pass some inpainted images to refiner, and then track quant metrics improvement and human evaluation

Pass dalle3 image to refiner to see if authenticity improved

Pass images from T2I to refiner (sdxl and sd3) to see if quality improved

4.3.4 Discussion:

[Summary table of I2I]

We find a,b,c,d,……(CLIP-IQA not stable)

4.4 Framework

4.5 Application of Framework

5. Results and Analysis:

Results from CLIP-IQA: (can be used for screening)

Results from Human Evaluation

(Limitations summary table here)

Can’t use immediately but can do for ideation.

6.Conclusion and Future Research

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

(#Research Gap: Previous research does not have systematic comparisons among 8 variations of diffusion models under marketing context. They don’t use both quantitative performance metrics and human evaluations.) [Rephrase this]

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.

GPT4o

SDM3 Medium

Midjourney

Automatic1111 (Style Aligned and ControlNet Reference to generate consistent style images)

ConfyUI.

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

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

Appendix: