Student Declaration of Authorship

HERIOT WATT UNIVERSITY

UK | DUBAI | MALAYSIA

Course code and name:	F21RP – Research Methods and Project Planning
Type of assessment:	Individual Assessment
Coursework Title:	Evaluation based enhanced Image Generation
Group Number:	H00463082
Student Names and IDs:	HARIHARAKUMAR RATHINAR(H00463082)

Declaration of authorship. By signing this form:

- **I declare** that the work I have submitted for individual assessment OR the work I have contributed to a group assessment, is entirely my own. I have NOT taken the ideas, writings or inventions of another person and used these as if they were my own. My submission or my contribution to a group submission is expressed in my own words. Any uses made within this work of the ideas, writings or inventions of others, or of any existing sources of information (books, journals, websites, etc.) are properly acknowledged and listed in the references and/or acknowledgements section.

- I confirm that I have read, understood and followed the University's Regulations on plagiarism as published on the University's website, and that I am aware of the penalties that I will face should I not adhere to the University Regulations.

- I confirm that I have read, understood and avoided the different types of plagiarism explained in the University guidance on Academic Integrity and Plagiarism

Student Signature *(type your names): HARIHARAKUMAR RATHINAR*

Date: *04/10/2025*

Research Methods and Project Planning (Course Code: F21RP)

Evaluation based enhanced Image Generation

Presented By	
Author and Student ID of Author with email IDs	**Hariharakumar Rathinar (H00463082)** **hr30000@hw.ac.uk**
Version	**v1.0**
Working Capacity	**Individual**
Group Number allocated	**H00463082**
Date of submission	**10th April 2025**

Acknowledgements

I would like to express my sincere gratitude to Dr. Elgayar, Neamat, for her guidance, support, and valuable insights throughout this research project. Her expertise in Machine Learning, Artificial intelligence, guidance towards taking internship and encouragement at every stage helped shape this work.

Abstract

This research aims to evaluate and enhance the aesthetics of images used on Floward e-commerce platform where products like flower bouquets, chocolates, watches, gift boxes, and toys are listed. This study systematically selects to implements image evaluation and enhancement techniques using existing machine learning and Latent diffusion models by not altering key product characteristics such as shape, size, and primary colours Leveraging.

A dataset comprising publicly available product images and textual descriptors from Floward is curated for the research project. The research employs proven image evaluation models to measure image attributes such as brightness, sharpness, contrast, colour palette, and overall composition. These evaluations are benchmarked against established aesthetic criteria derived from photography handbook provided by Floward.

Post evaluation, the images and enhancement test are given as input to pre-trained Latent Diffusion models such as Stable Diffusion, SDXL and Kandinsky model. Enhanced images undergo re-evaluation along with similarity comparison with original image using Structural Similarity Index Measure (SSIM) and Learned Perceptual Image Patch Similarity (LPIPS). The validation approach ensures that enhancements achieve measurable improvements in aesthetic quality while preserving authentic representation and also helps us to identify the right diffusion model for image enhancement.

The study offers an integrated workflow that facilitates enhanced visual appeal for online retail platforms by automating objective aesthetic assessments and image upgrades. A comparative analysis report emphasizing the best latent diffusion model and an automated pipeline prototype enabling users to assess and improve the image are among the results.

Table of Contents

List of Figures

Abbreviations

DDIMs	Denoising Diffusion Implicit Models
DDPMs	Denoising Diffusion Probabilistic Models
CLIP	Contrastive Language-Image Pre-Training
VAE	Variable Auto Encoder
SSIM	Structural Similarity Index Measure
LPIPS	Learned Perceptual Image Patch Similarity
GANs	Generative Adversarial Networks
FID	Fréchet Inception Distance
PSNR	Peak Signal-to-Noise Ratio
SVM	Support Vector Machine
HOG	Histogram of Oriented Gradients
EOE	Edge Orientation Entropy
LDM	Latent Diffusion Models

Chapter 1

Introduction

1.1 Aims

This research aims to systematically evaluate and enhance the aesthetic appeal of product images on the Floward e-commerce platform. By leveraging proven machine learning and existing stable diffusion models, the study seeks to choose the best stable diffusion model for automating the assessment and enhancement of images primarily comprises of online web products like flower bouquets, chocolates, watches, gift boxes, and toys. The enhancement process aims to significantly improve image quality attributes such as lighting, colour harmony, sharpness, and composition while rigorously preserving the authenticity and core visual characteristics of the products.

1.2 Objective

The primary objective is to significantly enhance image quality while maintaining the authenticity and core visual characteristics of the products.

- Collect dataset consisting of publicly available product images like flower bouquets, chocolates, watches, gift boxes, and toys along with associated textual descriptors from Floward.
- Employ proven image evaluation models Toolbox (Ralf Bartho,2024) to quantify critical aesthetic parameters including object size, brightness, sharpness, contrast, lighting angles, color harmony, and overall composition.
- Benchmark image evaluations against established aesthetic guidelines outlined in Floward's Photography Guidelines Manual.

- Utilize advanced pre-trained Stable Diffusion models such as Stable Diffusion XL Refiner, Latent Diffusion XLabs-AI adapters, and ControlNet inpainting techniques to enhance images based on the benchmark comparison to the evaluation output.
- Evaluate Stable diffusion model output by re-evaluating enhanced images using Toolbox (Ralf Bartho,2024) and validating similarity using Structural Similarity Index Measure (SSIM) and Learned Perceptual Image Patch Similarity (LPIPS).
- Develop an automated, integrated workflow for consistent aesthetic evaluation and enhancement, resulting in improved visual content for Floward's online platform.

1.3 Motivation

Online material that is visually appealing is driving the e-commerce environment more and more. Product photos are essential for drawing in clients, swaying their decisions to buy, and eventually increasing sales. However, it can take a lot of effort and time to manually edit or curate high-quality product photos.

Keeping a sizable assortment of varied products with eye-catching graphics is a challenge for Floward.com, an online gift platform. Our research attempts to address this issue by creating an automated system that uses stable diffusion techniques and existing, established machine learning models to evaluate and improve product photos.

Motivation is two-fold:

Improve Customer Experience: We want to give customers a more interesting and eye-catching online shopping experience by creating an automated system that can assess and improve product photos.

Boost Sales and Revenue: We anticipate that enhancing the visual appeal of product photos will boost consumer interaction, boost sales, and eventually help Floward's bottom line.

Chapter 2

Literature Review

2.1 Introduction

This literature review examines advanced techniques for image evaluation and enhancement with a focus on stable diffusion models. It outlines the evolution of diffusion-based methods and discusses the technical architecture behind stable diffusion. The review also addresses the application of these models in enhancing images using combined image and text inputs.

The following sections starts with detail review of the creation and limitations of diffusion models where each stage is detailed. It then shifts its focus to the stable diffusion model providing in-depth details on its architecture of stable diffusion and functionalities are covered along with the image enhancement capabilities. The final section of the literature review covers the evaluation model in application of the Floward's photo guidelines to generate the enhancement text along with the similarity validation.

2.2 Background Work

Consumer trust and engagement in e-commerce platforms are greatly impacted by image aesthetics. Purchasing decisions can be directly influenced by high-quality photos, which can also improve user satisfaction and perceived product value (Bao et al., 2015; Kim et al., 2018). Even with its acknowledged significance, maintaining uniform image quality throughout an e-commerce platform is still difficult, mostly because manual evaluations might vary, opinions are subjective, and conventional methods have scalability issues.

Automated picture enhancement and aesthetic assessment is a promising solution to these issues. However, there is a clear division in the current models' functions. Evaluation models focus primarily on aesthetic attributes such as color balance, clarity, brightness, and composition (Redies et al., 2024). However, under the direction of text or predetermined norms, enhancement models particularly generative models like diffusion models concentrate on improving the aesthetics of images (Rombach et al., 2022; Nichol & Dhariwal, 2021). An encouraging answer to these problems is automated aesthetic assessment and image improvement. Nonetheless, the functionalities of the present models are clearly divided. Aesthetic qualities like color balance, clarity, brightness, and composition are the main emphasis of evaluation models (Redies et al., 2024). On the other hand, enhancement models especially generative models like diffusion models focus on enhancing the aesthetics of images under the guidance of text or preset standards (Rombach et al., 2022; Nichol & Dhariwal, 2021).

A significant gap in the literature—no single model currently in use adequately covers both evaluation and enhancement—is the basis for the reasoning behind this study's separation of evaluation and enhancement. Perceptual aesthetics are neglected by traditional measurements such as SSIM and PSNR, which excel at structural quality (Wang et al., 2004; Sheikh et al., 2006). On the other hand, while neural techniques and sophisticated tools like the Aesthetics Toolbox (Redies et al., 2024) offer thorough aesthetic assessments, they do not have integrated mechanisms for automated improvement.

By adding textual cues, diffusion models—particularly Latent Diffusion Models (LDMs) have proven to be highly effective at improving the aesthetics of images (Rombach et al., 2022). These models effectively manage intricate image-to-image conversions, and they are especially helpful in e-commerce situations where exact adherence to aesthetic standards or brand rules is required (Podell et al., 2023; Razzhigaev et al., 2023). However, complementary techniques are required for objective aesthetic evaluation because of their inherent design for picture creation rather than judgment.

As a result, this study examines these two crucial elements evaluation and enhancement separately but in tandem. The purpose is to create a complete pipeline that can impartially evaluate photos using preset aesthetic standards and then improve them appropriately. The choice to look at assessment and enhancement independently enables a thorough investigation of specific approaches, guaranteeing strong performance in real-world e-commerce applications. This method overcomes existing constraints and closes the gap between precise aesthetic assessment and effective image improvement.

2.3 Image Aesthetic Evaluation

Image aesthetic evaluation is crucial and critical in this research, aimed at assessing the visual appeal, quality, and attributes of images, with applications spanning photography, e-commerce, and digital media. This literature review combines foundational and existing methods for evaluating image aesthetics, focusing on traditional quality metrics, neural-based approaches, and the specialized Aesthetics Toolbox by Redies et al. (2024). It examines their mechanisms, strengths, and limitations, providing a comprehensive overview of how aesthetic attributes such as sharpness, color, brightness, and composition are quantified and validated in the literature.

2.3.1 Aesthetic Evaluation Methods

Aesthetic evaluation is all about figuring out what makes an image look good or bad for any given specific given context. Over the years, researchers have come up with many ways to do this, and these can be split into two main types: traditional metrics and learned methods using neural networks.

2.3.1.1 Traditional Image Quality Metrics

Older studies used simple tools like Structural Similarity Index (SSIM) and Peak Signal-to-Noise Ratio (PSNR) to check image quality (Wang et al., 2004). SSIM looks at how similar two images are in terms of structure, like shapes and patterns, giving a score from 0 to 1. Tang et al. (2013) argue that SSIM overlooks aesthetic nuances like color harmony or composition, limiting its utility for appeal-focused assessments, with experiments showing weak correlation (r=0.35) to human aesthetic ratings. Peak Signal-to-Noise Ratio (PSNR), another classical metric, measures pixel-level noise with higher values (e.g., >30 dB) indicating better quality (Wang et al., 2004). Sheikh et al. (2006) tested PSNR on compressed images, reporting values averaging 32 dB, but noted its insensitivity to perceptual factors, such as brightness perception or edge clarity, critical for aesthetics (Sheikh et al., 2006). These studies collectively suggest that while traditional metrics excel in fidelity assessment, they fall short in capturing the holistic visual appeal central to aesthetic evaluation.

2.3.1.2 Neural Aesthetic Assessment

Advancements in deep learning have shifted aesthetic evaluation toward neural based methods that leverages human labeled datasets to predict visual appeal. (Murray et al. (2012)) developed the AVA dataset, containing over 250,000 images rated by users, and trained a Support Vector Machine (SVM) classifier to predict aesthetic scores, achieving 66.7% accuracy in distinguishing high vs. low-quality images. Further advancements by Lu et al. (2015) utilized a dual-column Convolutional Neural Network (CNN) architecture to separately process global and local features, significantly improving correlation with human aesthetic judgments.

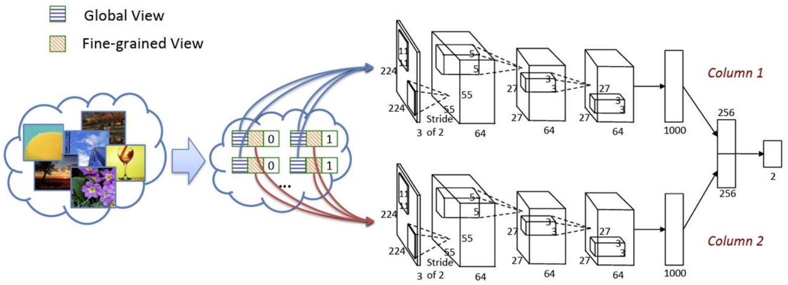


Figure 1 (Lu et al. (2015)) Double-column convolutional neural network

The double column architecture convolution neural network represent the input as a image where the image is represented in two views which are the global view and local view. The quality label is also associated with the same. The entire model excels in capturing composition but requires extensive computational resources due to the complexity. Zhang et al. (2018) introduced the Learned Perceptual Image Patch Similarity (LPIPS), demonstrating superior perceptual alignment with human judgment compared to SSIM (LPIPS correlation = 0.85 versus SSIM = 0.65), although LPIPS primarily addresses perceptual similarity rather than a dedicated aesthetic tool (Zhang et al., 2018).

2.3.1.3 Aesthetics Toolbox by Redies et al. (2024) for Objective Image Quality Assessment

Among the latest ideas, Redies et al. (2024) came up with something called the Aesthetics Toolbox. This is a special method that mixes old-style image processing with new deep-learning tricks to check specific parts of an image, like background color, brightness, sharpness, color richness, and how things are placed (Redies et al., 2024). It's different from other tools because it doesn't just give one overall score it looks at each part separately, which is helpful for fixing exact problems.

Aesthetics Toolbox Key Differentiator

Most neural network models learn from big datasets of images people have rated, but the Aesthetics Toolbox takes a different path. It uses three main ideas:

- Perceptual Color Spaces: It uses a system called CIELAB, which matches how humans see colors better than normal RGB (Fairchild, 2013).
- Texture and Sharpness Metrics: It checks clarity using things like edge orientation entropy, which measures how sharp and clear edges are.
- Composition Factors: It looks at balance and symmetry to see if the image feels harmonious (Locher et al., 1999).
- Deep Features: Incorporates CNN-extracted features for holistic evaluation, enhancing traditional metrics (Redies et al., 2024).

This mix makes the toolbox very detailed and useful for specific jobs, like checking product images in online stores.

2.3.2 Other Methods and Their Findings

Apart from the Aesthetics Toolbox, other studies have added to this field. Datta et al. (2006) were some of the first to use computers to study aesthetics. They took simple features like color patterns and edge layouts to guess if an image is nice, but their results weren’t very strong because they didn’t use advanced tools. Then, Talebi et al. (2018) made NIMA (Neural Image Assessment), a neural network that gives a score from 1 to 10 based on the AVA dataset. NIMA got a higher match with human views, which is good for an overall rating. But it doesn’t break down specific parts like brightness or color, which Redies et al. (2024) say is a weak point. Bao et al. (2015) also showed that sharp images build trust in buyers, while Kim et al. (2018) found brighter pictures make products seem better online. These findings show why we need tools that look at details, not just the whole image.

Method	Key Metrics	Strengths	Weaknesses
Traditional Metrics	SSIM (>0.9), PSNR (>30 dB)	Structural fidelity, simplicity	Ignores aesthetic appeal
Neural Assessment	Aesthetic Score (0–1), LPIPS (<0.2)	Captures human perception	Subjective, resource-intensive
Aesthetics Toolbox	LAB (e.g., L*=88–92), RMS ~22, Edge >500	Granular, objective outputs	Threshold dependency, computational complexity

2.3.3 Aesthetics Toolbox – Detailed Overview and Feature Evaluation

The Aesthetics Toolbox, created by Redies et al. (2024), framework we're using to objectively analyze how good our images look from a specific targeted value range defined. For Floward, this tool is perfect since it helps us measure features like background colour, brightness, sharpness, and colour richness. It's built around the CIELAB colour space.

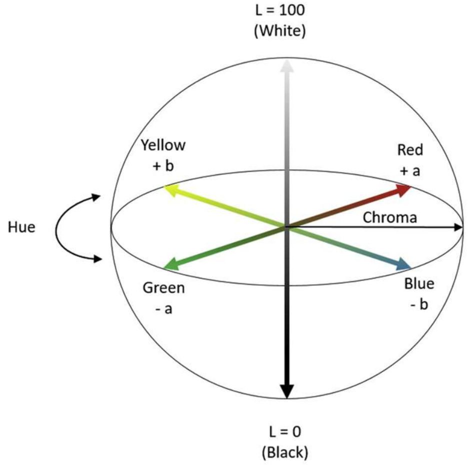


Figure 2: CLLIB Color Space Diagram (Ly et al. 2020)

The figure defines colors along three axes: L* (lightness, 0–100), a* (green-red), and b* (blue-yellow), enabling precise adjustments that mirror human visual perception (Fairchild, 2013). For this project, the toolbox evaluates original and enhanced images to identify deficiencies and verify improvements, directly supporting the proposed methodology's initial evaluation and re-evaluation steps.

Key Features Evaluated by the Toolbox

The Aesthetics Toolbox assesses product images across multiple dimensions, each tied to specific requirements outlined. Below, the details of each feature extracted and the values interpretation from the same.

Color Analysis (CIELAB): Leveraging the perceptually uniform CIELAB space, the toolbox objectively quantifies image colors through L* (lightness, 0–100), a* (green red), and b* (blue yellow) axes. This facilitates precise evaluation and adjustment aligned with human visual perception (Fairchild, 2013). For example, a warm, neutral ivory tone can be detected if the lightness, green-red and blue-yellow is of value L=88–92, a=-1–3, and b=8–12.

Brightness (RMS Contrast): Brightness assessment utilizes Root Mean Square (RMS) contrast, which measures luminance variability across the image. Higher RMS values typically reflect greater luminance variation and perceived brightness, vital for determining suitable lighting and clarity (Peli, 1990). A cozy, natural lighting can be achieved by validating the range of 15 to 30 and an ideal midpoint of approximately 22 (Peli, 1990)

Sharpness and Edge Quality: Sharpness is assessed through Edge Density (>500), Complexity (HOG, >5), and Edge-Orientation Entropy (EOE, >4), ensuring high clarity for product details like watch edges or toy textures (Redies et al., 2024). Complexity, assessed via Histogram of Oriented Gradients (HOG), quantifies structural details. Edge Orientation Entropy (EOE) specifically evaluates the distinctness and clarity of edge orientations, with higher values signifying clearer edge definitions (Chiamulera et al., 2017; Redies et al., 2024).

Colorfulness (Color Entropy): Color Entropy measures colorfulness, targeting a range of 4–6 to boost vibrancy (Hasler & Süsstrunk, 2003). Values within a mid-range (e.g., 4–6) typically indicate an optimal balance between vibrancy and visual coherence, aligning with general aesthetic preferences (Hasler & Süsstrunk, 2003).

Spatial Frequency (Fourier Spectrum Slope): Fourier slope assessment captures the spatial frequency distribution, with natural, aesthetically pleasing images typically exhibiting slopes between -2 and -3. This measure aids in evaluating the naturalness of image textures and patterns (Spehar et al., 2016).

Balance and Symmetry: Image harmony and visual balance are quantified using symmetry and balance metrics. Lower values suggest higher aesthetic harmony, indicating visually pleasing spatial distributions within images (Locher el at., 1999). The balance should be less 20% and the symmetry should be less than 5%. This ensure that the image demonstrates strong visual harmony and acceptable spatial balance.

Summary Table of Feature Evaluations

Feature	Range Comparison
Ivory Background (LAB)	L(88-92), a(-1-3), b(8-12)
Brightness (RMS Contrast)	15–30
Sharpness (Edge Density)	>500 (sharp images)
Edge Definition (EOE)	>4 (well Defined Edge
Color Entropy	4–6 (Vibrant)
Fourier Spectrum Slope	-2 to -3
Balance & Symmetry	<20% balance, <5% symmetry

2.3.4 Similarity Score Validation

Validating the quality and fidelity of images enhanced by stable diffusion models is crucial to ensure they meet original image aesthetic and authenticity goals. This process employs quantitative metrics to compare original images with their enhanced counterparts, assessing perceptual and structural similarity. Two widely adopted measures Learned Perceptual Image Patch Similarity (LPIPS) and Structural Similarity Index Measure (SSIM) provide robust, objective benchmarks for this evaluation, complementing the Aesthetics Toolbox's aesthetic analysis.

Learned Perceptual Image Patch Similarity (LPIPS): LPIPS quantifies perceptual similarity between original and enhanced images by leveraging deep neural network features, typically from pre-trained models like VGG or AlexNet (Zhang et al., 2018). It calculates the distance between feature representations of image patches, closely aligning with human subjective preferences. Lower LPIPS scores (e.g., <0.2) indicate higher perceptual similarity, suggesting that enhancements are visually coherent and appealing.

Structural Similarity Index Measure (SSIM): SSIM assesses structural and visual quality by comparing luminance, contrast, and structural patterns between original and enhanced images (Wang et al., 2004). It produces a value between 0 and 1, with scores closer to 1 (e.g., >0.9) indicating strong similarity. This metric is particularly valuable for verifying that stable diffusion enhancements preserve core product features such as the shape of a watch or the arrangement of a gift box preventing distortions that could mislead consumers.

These metrics are applied to post enhanced image to evaluate outputs from three latent diffusion models like Stable Diffusion, SDXL and Kandisky models. For instance, an original image with LAB values L=79.347, a=4.351, b=12.742, enhanced to target L≈90 via stable diffusion, would be re-evaluate the LAB values using toolbox along with LPIPS to check perceptual alignment and SSIM to ensure structural integrity. This dual approach ensures that the enhanced image aligns with original image with aesthetic improvement which support the goal of automated image aesthetic evaluation and enhancement of given image.

2.4 Image Aesthetic Enhancement

The Image aesthetic enhancement section covers methodologies and models aimed at improving the visual quality and appeal of images, through generative models like diffusion models and their advanced variants, such as latent diffusion models. This section explores fundamental diffusion techniques, their evolution into more computationally efficient latent representations, and specific implementations available through platforms such as Hugging Face. The latent diffusion models includes Stable Diffusion, SDXL, and Kandinsky are critically reviewed, detailing their architectures, operational specifics, and their implications for image-to-image enhancement tasks.

2.4.1 Diffusion Model

Diffusion models have significantly advanced generative modeling, surpassing many limitations observed in traditional methods such as Generative Adversarial Networks (GANs). By systematically corrupting data with noise and then learning the reverse denoising process, these models generate highly diverse, detailed outputs (Sohl-Dickstein et al., 2015; Ho et al., 2020). Their adaptability has made them an attractive choice for image synthesis and editing, including scenarios where textual instructions guide the transformation of an existing image.

Operational Mechanisms of Diffusion Models

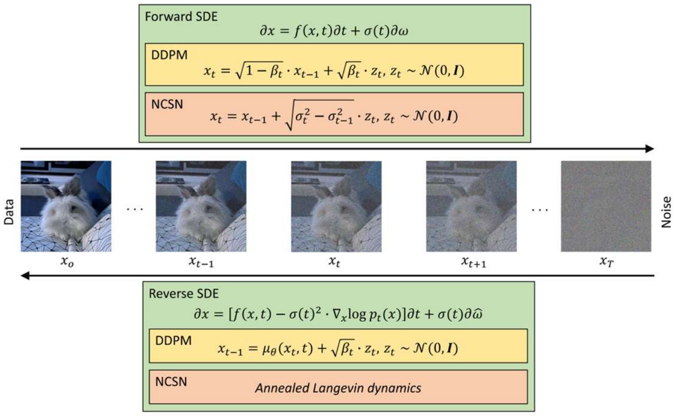

Forward SDE
$\partial x = f(x,t)\partial t + \sigma(t)\partial\omega$
DDPM
$x_t = \sqrt{1-\beta_t} \cdot x_{t-1} + \sqrt{\beta_t} \cdot z_t,\ z_t \sim \mathcal{N}(0,\boldsymbol{I})$
NCSN
$x_t = x_{t-1} + \sqrt{\sigma_t^2 - \sigma_{t-1}^2} \cdot z_t,\ z_t \sim \mathcal{N}(0,\boldsymbol{I})$
Data
...
...
Noise
x_o
x_{t-1}
x_t
x_{t+1}
x_T
Reverse SDE
$\partial x = [f(x,t) - \sigma(t)^2 \cdot \nabla_x \log p_t(x)]\partial t + \sigma(t)\partial\hat{\omega}$
DDPM
$x_{t-1} = \mu_\theta(x_t,t) + \sqrt{\beta_t} \cdot z_t,\ z_t \sim \mathcal{N}(0,\boldsymbol{I})$
NCSN
Annealed Langevin dynamics

Figure 3 from Croitoru et al. (2023) illustrates the architectural overview of diffusion model.

Diffusion models are grounded in a probabilistic framework that leverages a forward and reverse process. In the forward process, where noise addition takes place, an original image x_0 is incrementally corrupted with Gaussian noise over *t* timesteps by Gaussian noise, resulting in a fully noisy state x_t.

$$x_t \ = \sqrt{\bar{\alpha}_t x_0} + \sqrt{1-\bar{\alpha}_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0,I)$$

This is modelled as a Markov chain, where each step depends solely on the previous one, governed by a noise schedule β_t that increases over time (Ho et al., 2020) where $\alpha_t = 1 - \beta_t$. The reverse process trains a neural network which is typically a U-Net to predict the noise at each step, iteratively reconstructing the image. Ho et al. (2020) formalized this in Denoising Diffusion Probabilistic Models (DDPMs), achieving an FID of 3.17 on CIFAR-10, surpassing many GANs in sample quality due to its stable training dynamics.

Nichol and Dhariwal (2021) extended this work by optimizing noise schedules (e.g., cosine schedules) and reducing sampling steps, improving computational efficiency. Their improved DDPMs achieved an FID of 3.85 on ImageNet, demonstrating scalability to larger datasets. Song et al. (2020) introduced Denoising Diffusion Implicit Models (DDIM), cutting steps to as few as 20 while maintaining quality, which contrasts DDIM's deterministic sampling with DDPM's stochastic approach.

Integration with language models has broadened diffusion's applicability. Ramesh et al. (2022) combined diffusion with CLIP in DALL·E 2, enabling text-to-image generation with an FID of 10.39.

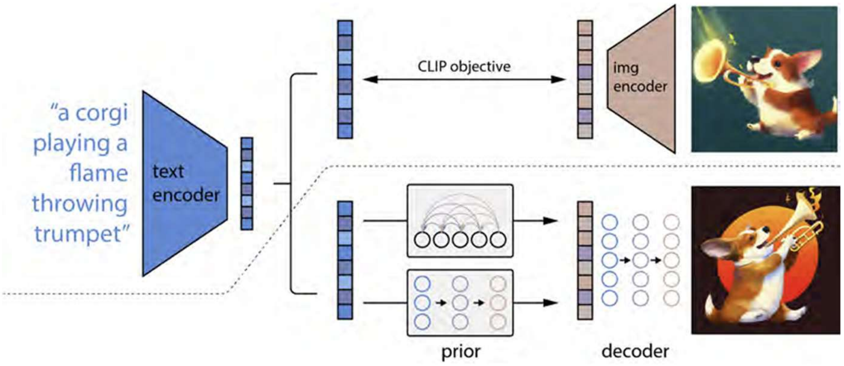


Figure 4: CLIP High level architecture (Ramesh et al. (2022)

The architecture flow as presented in figure above shows how text embeddings guide the denoising process, though it prioritizes synthesis over enhancement. Similarly, Saharia et al. (2022) developed Imagen, achieving an FID of 7.27 by enhancing text conditioning with cascaded diffusion, yet its focus remains on generating new images (Saharia et al., 2022).

Implementation Example: SDEdit

Meng et al. (2021) proposed SDEdit in "Guided Image Synthesis and Editing with Stochastic Differential Equations," adapting diffusion models for image-to-image editing using stochastic differential equations (SDEs). Unlike DDPMs, SDEdit starts with an input image, applies a forward SDE to add controlled noise, and reverses it with guidance from a condition (e.g., text, image, or mask). This enables user to perform tasks like style transfer, colorization, and inpainting. For instance, Meng et al. (2021) transformed a grayscale cat image into a coloured version using a reference image, achieving a PSNR of 25.6 and SSIM of 0.87, indicating strong structural fidelity.

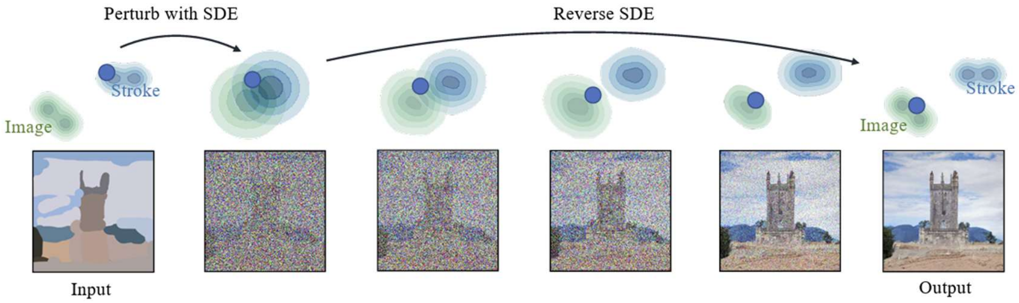


Figure 5: SDedit image editing process (Meng et al. 2021)

The working of SDedit where an image like dog is edited to landscape altered in style. SDEdit's flexibility stems from its SDE framework, which allows fine-tuning of noise levels and guidance strength. The authors tested it on datasets like CelebA-HQ, reporting a Learned Perceptual Image Patch Similarity (LPIPS) of 0.22, competitive with GAN-based methods (Meng et al., 2021). However, its computational cost is significant, requiring 200–500 steps depending on the task, and its performance degrades with poorly calibrated parameters, such as excessive noise overwhelming the input structure (Meng et al., 2021). Compared to Stable Diffusion's latent approach, SDEdit's pixel-space operation is less efficient but offers precise control, making it a valuable benchmark.

Additional Implementations and Variants

Beyond SDEdit, other diffusion-based implementations have been explored. Choi et al. (2021) introduced ILVR (Iterative Latent Variable Refinement), conditioning diffusion on a reference image to guide synthesis or editing. Applied to FFHQ, ILVR achieved an FID of 8.5 for style transfer, where a face image adopts another's style while retaining identity. ILVR's strength is its reference-based guidance, but it shares diffusion's step-intensive nature (200+ steps), limiting real-time use (Choi et al., 2021).

Another variant, Palette by Saharia et al. (2022), focuses on image-to-image tasks like colorization and inpainting, achieving a PSNR of 26.8 on ImageNet. Palette's efficiency (100 steps) is better than SDEdit, but it lacks robust text guidance, relying on pre-trained classifiers (Saharia et al., 2022). These implementations highlight diffusion's adaptability, yet underscore its computational challenges.

Transition to Latent Diffusion Models

Early diffusion models demonstrated significant potential in image generation but were limited by substantial computational demands. These inefficiencies stemmed primarily from pixel-by-pixel processing in models like Denoising Diffusion Probabilistic Models (DDPMs) (Ho et al., 2020), building on earlier foundational work (Sohl-Dickstein et al., 2015). Such models required extensive computations at each diffusion step, making them impractical for real-time applications, particularly with high-resolution images (Song et al., 2021).

To address these limitations, Latent Diffusion Models (LDMs) were introduced by Rombach et al. (2022). LDMs significantly improved computational efficiency by employing Variational Autoencoders (VAEs) (Kingma & Welling, 2013) to encode images into a compact latent representation. This approach drastically reduced computational load during both the diffusion and denoising processes, enabling faster training and generation of high-quality images.

Early diffusion models also struggled with text-guided image-to-image enhancement, mainly because their denoising processes required numerous steps, contrasting sharply with the efficiency of Generative Adversarial Networks (GANs). The forward diffusion in these models completely erased structural information, conflicting with enhancement tasks that require preserving the input's underlying structure (Nichol & Dhariwal, 2021). Although methods like DDIM (Song et al., 2020), SDEdit, ILVR, and Palette provided partial solutions, they remained limited by complex tuning and considerable computational requirements.

Latent Diffusion Models addressed these challenges effectively. By operating in a lower dimensional latent space, LDMs allowed efficient textual conditioning and preserved structural details essential for targeted image enhancements. Models like Stable Diffusion, based on LDM frameworks (Rombach et al., 2022), thus emerged as efficient and practical solutions for tasks requiring precise, text-guided image aesthetic evaluation and enhancement..

2.4.2 Latent Diffusion Model

Latent diffusion models address pixel-space diffusion inefficiency by operating in a compressed latent space. This section reviews their mechanisms, evolution, comparisons with different image-to-image Latent diffusion models such as Stable Diffusion, SDXL and Kandinsky with real-world performance, and research evaluations. The latent diffusion model core working is explained in the Stable diffusion model, and then further models are explained on how they are implemented. The sections below cover individual models and then make an overall comparison for all three models with respect to type and score, image functionality, text-guided enhancement, advantages, and weaknesses.

Stable Diffusion Model

Rombach et al. (2022) introduced Stable Diffusion, leveraging a Variational Autoencoder (VAE) to encode images into latent, where a U-Net performs diffusion guided by CLIP text embeddings. This reduces steps to 50, achieving an FID of 10.5 on LSUN. Notably, implementation is done with Stable Diffusion v1.5 with the same stable diffusion architecture. The Stable Diffusion v1.5 image-to-image model was created by Rombach, 2022 and hosted on Hugging Face. This model is implemented with the Latent Diffusion Model (LDM) architecture to modify existing images guided by textual inputs efficiently. Unlike pixel-space methods, this architecture operates within a compressed latent representation, significantly enhancing computational efficiency and making it suitable for practical image enhancement applications where image can be ehanced with coditional text inputs.

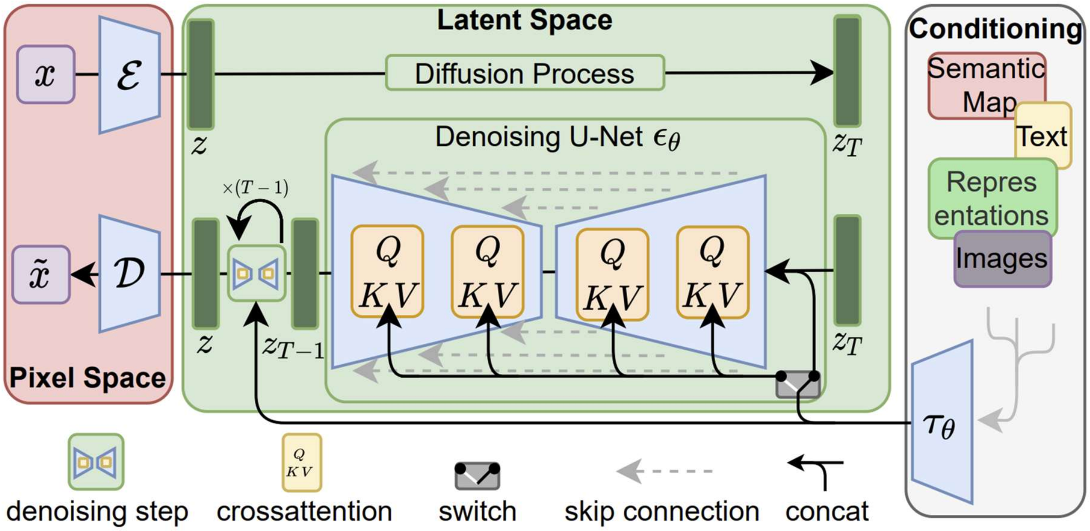


Figure 6: Architecture diagram of Stable diffusion Radames, 2022

The first step of the architecture is to capture the input image where input Image ($\varkappa$) represents the initial high-resolution image input to the stable diffusion model. This is depicted on the left side of the diagram as a full-sized image, typically a 512x512 pixel grid in high resolution and it is the starting point of the process in stable diffusion architecture where the image is not directly manipulated but serves as the source.

VAE Encoder ϵ is the encoder component of the Variational Autoencoder (VAE), shown as a box with an arrow from $\varkappa$ to z where the high-resolution input image $\varkappa$ into a lower-dimensional latent representation z. The VAE is pre-trained to ensure this latent form captures the image's key structural and semantic information (Rombach et al., 2022).

Latent Representation z (often denoted z_0 initially) is the compressed latent representation output by ϵ. Shown as a smaller, abstract grid (e.g., 64x64), z is a condensed version of $\varkappa$, holding the image's core features in a reduced space. This is where the diffusion process occurs in stable diffusion, rather than the full pixel space, making it more efficient than traditional diffusion models (Rombach et al., 2022).

Forward Diffusion Process with variables $(z_1, z_2, \dots., z_T)$ represent the sequence of increasingly noisy latent states, ending in z_T, depicted as a series of boxes or a gradient. Starting from z (or z_0), the forward diffusion process gradually adds Gaussian noise over T timesteps, transforming the latent representation into pure noise (z_T). Each step introduces a controlled amount of noise, governed by a schedule (e.g., cosine schedule in Nichol & Dhariwal, 2021), until the original structure is fully obscured. In enhancement tasks, this might be partial (e.g., up to z_T where t′ < T) to retain some input features (Rombach et al., 2022).

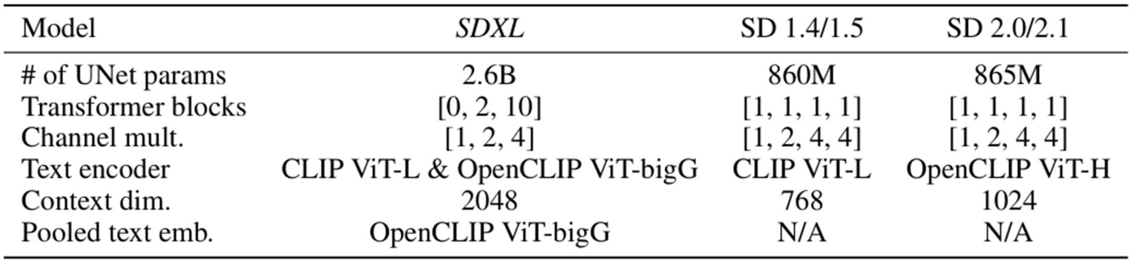
Reverse Diffusion Process with U-Net (ϵ_θ) ϵ_θ is the noise predictor within the U-Net block, shown with an arrow looping back from $z_T(or\ z_{t'})$to z'_0, conditioned by text input. The reverse process begins from the noisy latent state (z_T for generation, or $z_{t'}$ for enhancement) and iteratively removes noise using the U-Net, a convolutional neural network. The U-Net predicts the noise ϵ_θ at each step, guided by timestep information and a text conditioning input (e.g., via CLIP embeddings), refining the latent representation back to an enhanced version z'_0 . The text input, connected via a dashed line, directs the denoising to achieve specific outcomes (Rombach et al., 2022).

VAE Decoder D is the decoder component of the VAE, shown as a box with an arrow from z'_0 to $\varkappa$’. The decoder takes the enhanced latent representation z'_0 and reconstructs it into a high-resolution image $\varkappa$’. This step reverses the compression done by ϵ, expanding the 64x64 latent grid back to a 512x512 pixel image, now incorporating the refinements from the reverse diffusion process (Rombach et al., 2022).

Output Image $\varkappa$’ is the final enhanced high-resolution image, shown on the right. This is the end result, a 512x512 pixel image reflecting the enhancements guided by the text input and processed through the stable diffusion pipeline. It’s the enhanced version of the original $\varkappa$, optimized for quality and fidelity (Rombach et al., 2022). Rombach et al. (2022) detail Stable Diffusion’s latent diffusion, achieving an FID of 10.5 and SSIM of 0.88 on LSUN and COCO.

SDXL Model

Podell et al. (2023) advanced the stable diffusion model with Stable Diffusion XL (SDXL), where the stable diffusion XL model expands upon the latent diffusion approached introduced by featuring Rombach et al. (2022). Stable diffusion XL model comprise of a larger U-Net with increased parameters, dual text encoders (CLIP and OpenCLIP), a refiner model and higher internal resolution, enabling more detailed image synthesis and improved text alignment. SDXL achieves better FID and SSIM scores on COCO, improving detail and resolution (1024x1024). The model outperforms the UNet parameter to 2.6B, and double the context dim in comrparison to the latest stable diffusion models.



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

Figure 7: Table comparison of SDXL and SD models

The stable diffusion pooled text embedding helps in capturing global semantic meaning, providing holistic signal, complementing detailed local conditioning and improved coherence and prompt following.

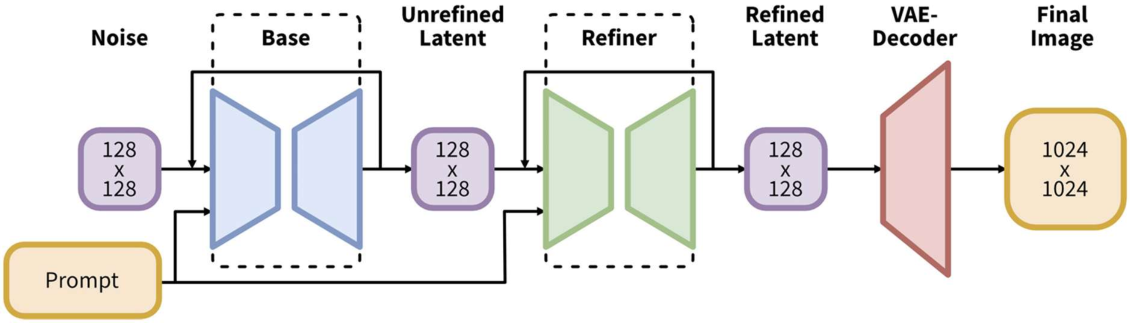


Figure 8: Pipeline representing Two stage visualization

The workflow starts with noise Introduction into the base model. Gaussian noise is incrementally introduced to generate an unrefined latent representation from the input latent space. The Unrefined Latent Generation layers is where a larger, more parameter-intensive U-Net generates an initial latent image based on textual inputs encoded by dual encoders. The output of unrefined latent layer is fed into the refiner SDXL 1.0-refiner model where it applied the SDEdit method, utilizing stochastic differential equations to iteratively refine these initial latent by carefully introducing and subsequently denoising additional controlled noise. The output of the refiner is then fed into the refined latent layer where the refined latent representation generates a richer in detail and texture which is then fed into the variable auto encoder. A Variational Autoencoder (VAE) decoder reconstructs the enhanced latent representation back into a high-resolution to produce a visually refined image that will be taken up for output.

While the original Stable Diffusion architecture employs a single text encoder and operates efficiently at 512×512 resolutions, SDXL is designed to handle 1024×1024 or above, delivering sharper edges and finer textures which is achieved through increased parameter counts in the U-Net, extended attention mechanisms, and a secondary “refiner” stage that further polishes the output. Consequently, SDXL preserves the computational benefits of latent diffusion but offers a higher fidelity output, making it particularly suitable for tasks that require detailed, large-scale imagery where its dual-encoder system ensures robust prompt handling, critical for enhancement tasks.

Li et al. (2024) further optimized Stable Diffusion with SVDQuant, quantizing it to 4 bits while retaining an FID in comparison to outputs, showing minimal quality loss. SVDQuant enhances efficiency, not attributes like brightness, relying on Stable Diffusion’s base capabilities (Li et al., 2024).

Kandinsky – Latent Diffusion Model

Kandinsky is a latent diffusion model developed by (Razzhigaev et al. (2023)) introduced Kandinsky 2.1. The Kandinsky model belongs to latent diffusion which combines the latent diffusion technique with image prior models like CLIP which encodes the image. Different image generation worlkflow is represented in below figure 9.

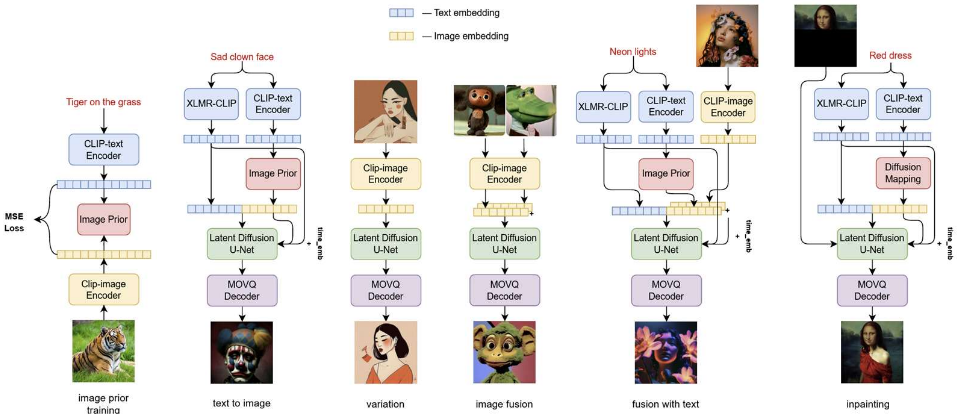


Figure 9: Kandinsky image generation Workflow Razzhigaev el al. (2023)

The model process begins with the input text prompt. This text is converted into a meaningful numerical representation using the text encoding models. Well established CLIP model are used for the encoding purpose with its strength lies in its ability to map both text and images into a shared space where semantic similarities are captured, forming the foundation for understanding the prompt's visual intent. A distinctive feature of Kandinsky is its use of a separate image prior model where a transformer network acts as an intermediary between the encoder layer and latent space. It takes the text embedding generated by CLIP and learns to predict a corresponding image embedding within CLIP's visual latent space. The image embedding produced by the prior model serves as a conditioning guide for a latent diffusion model. This is where the core image generation occurs, but efficiently in a compressed 'latent' space rather than directly with pixels. Kandinsky employs a modified MoVQ Variational Autoencoder (VAE) where its employed with two roles, its encoder compresses real images into lower-dimensional latent representations, and its decoder reconstructs images from these representations. The diffusion process operates within this compressed latent space. The diffusion model with U-Net architecture is trained to reverse a process of gradually adding noise to latent image representations. During generation, it starts with random noise and, guided by the image embedding from the prior, iteratively refines this noise over several steps.

The goal is to produce a final latent representation that strongly corresponds to the visual concept predicted by the prior, and thus, to the original text prompt Razzhigaev el al. (2023). Once the diffusion process yields the final latent representation, it is passed to the decoder part of the MoVQ VAE. This decoder translates the compressed latent code back into the high-dimensional pixel space, producing the final output image visible to the user.

2.4.3 Comparison of Image-to-Image Diffusion Models

This section expands the comparison to include Stable Diffusion, SDXL, and Kandinsky 2.2, synthesizing research on their mechanisms, performance, and suitability for text-guided image-to-image enhancement, providing a broader perspective beyond Section 2.4. The common features of all three models are as follows

1. Capable of image-to-image generation with enhancement test as input prompt
2. Based on Latent diffusion model
3. License is open for education purpose
4. Community support available in hugging face

The first table compares the core features of the stable diffusion, stable diffusion XL and Kandinsky where we compare all three models with their features.

Feature	Stable Diffusion	Stable Diffusion XL	Kandinsky
Models	radames/stable-diffusion-v1-5-img2img	stabilityai/stable-diffusion-xl-refiner-1.0	radames/kandinsky-2-1-img2img
Core Architecture	Latent Diffusion Model (LDM)	Latent Diffusion Model (LDM) with enhancements	LDM with Image Prior
Text Encoding	CLIP	ViT-L & OpenCLIP ViT-bigG	CLIP
Image Prior	None	Implicitly learned within the model	Explicit, separate Transformer model
Latent Space	Standard VAE	Standard VAE	Modified MoVQ VAE
Conditioning	Direct text embedding to U-Net	Direct text embeddings (dual encoders) to U-Net	Image embedding (from prior) to U-Net

Resolution Capability	Moderate (typically upscaled)	High (natively generates higher resolutions)	Moderate and similar to early stable diffusion model.
Prompt Following	Good	Improved with more understanding	Good and notably with aim to semantic alignment
Image Quality	Good	Generally higher detail with coherence	Competitive (good FID reported)
Computational Cost	Lower	Higher	Potentially higher
Key Innovation	Efficient latent diffusion	Higher resolution, dual text encoders	Explicit Image Prior for guidance
Release Timeframe	2022	Mid-2023	2024

The model comparison is further extended to compare all three models from the implementation standpoint on how each model will perform based on the literature review. The analysis involves not limited to the performance but also how the model can handle input text prompt for enhancing the images.

Feature	Stable Diffusion	Stable Diffusion XL	Kandinsky
Model Name	radames/stable-diffusion-v1-5-img2img	stabilityai/stable-diffusion-xl-refiner-1.0	radames/kandinsky-2-1-img2img
Suitable for “Image + Enhancement Text”? (Enhancement on Brightness, Sharpness, etc.)	Yes, widely used for text-based enhancements (brightness, contrast, etc.)	Yes, can interpret “enhance brightness” or “sharpen details” more reliably	Yes, interprets textual instructions for global or local enhancements
Handling Numeric CIELAB (L≈90), RMS Contrast (22), & Sharpness (>500)?	For Numeric Prompts this model may partially respond to “L≈90” or “edge density>500,” but typically interprets them as descriptive text rather than precise numeric constraints. Often requires iterative prompt tuning to get closer to the desired aesthetic.	Slightly better at following detailed instructions due to improved text encoder. However, it is still a natural-language model and not guaranteed to interpret numeric color parameters exactly. May respond somewhat more faithfully to “L=90” or “contrast=22,”.	Kandinsky can handle descriptive phrases but, like SD, it lacks a built-in numeric parser. “L=90” or “edge density>500” is treated as descriptive text. May excel at colour related requests but exact numeric matching is not guaranteed.
Advantages	Well documented with extensive community support. Good general purpose editing	Higher detail and better text alignment. Enhanced capacity for large or high-resolution images	Often praised for creative or stylized transformations. An alternative to the SD ecosystem, sometimes better at text fidelity

Weaknesses	Older Model which is fine for detail or high-resolution fidelity may trail behind SDXL. Numeric instructions are approximated and not guaranteed to match exact values	Requires more computing resources. Still no direct mechanism for exact numeric matches.	Smaller user community than SD. May need more trial and error with numeric like instructions.

Chapter 3

Requirement and Proposed Methodology

3.1 Requirement

To create an automated pipeline for image enhancement and image evaluation for online ecommerce platform Floward where the image evaluation is taken up by stable diffusion models (SDXL Refiner, Latent Diffusion with Flux Adapter, and ControlNet Inpainting) effectively enhance Floward's product images while meeting aesthetic and authenticity objectives, the following requirements are defined. These are prioritized using the MoSCoW method where the requirements are clearly scored as must have, should have, could have and wont have. The evaluation also includes quantitative metrics from the Aesthetics Toolbox and similarity validation (SSIM, LPIPS) for getting the evaluated enhanced image for improving the e-commerce customer experience. The table below outlines these requirements, their targets, and their relevance to Floward's goals.

Type	Requirement	Target Metric	MoSCoW Priority	Relevance to Floward Project
FR	Background Color Adjustment	LAB values: L=88–92, a=-1–3, b=8–12 (Aesthetics Toolbox)	Must Have	Ensures a warm, neutral ivory background (e.g., for bouquets), enhancing visual appeal without altering dimensions.
FR	Brightness Enhancement	RMS Contrast: 15–30, target ~22 (Aesthetics Toolbox)	Must Have	Improves illumination for cozy, natural lighting (e.g., chocolates), critical for inviting product presentation.

FR	Sharpness Improvement	Edge Density: >500, Complexity (HOG): >5, EOE: >4 (Aesthetics Toolbox)	Must Have	Achieves high clarity for product details (e.g., watch edges, toy textures), essential for e-commerce visibility.
FR	Image Loading and Preprocessing	512x512 pixel and jpeg format	Should Have	Ensure that the image is of 512x512 pixel and jpeg format
FR	Structural Preservation	SSIM: >0.9 (Similarity Validation)	Must Have	Preserves core visual characteristics (e.g., flower types, box designs) during enhancement, a non-negotiable need.
FR	Perceptual Similarity	LPIPS: <0.2 (Similarity Validation)	Should Have	Ensures enhancements align with human visual preferences (e.g., brighter outputs), enhancing perceived quality.
NFR	Environment setup	Right environment setup with python version matching with all libraries required	Must have	Ensure that the right python version is selected according to the version required for the
FR	Aesthetic Evaluation Module Integration	Aesthetics Toolbox (Redies et al., 2024) for image evaluatio	Must have	Ensure Aesthetics Toolbox (Redies et al., 2024) used for image evaluatio
FR	Enhancement Text Generation	Automatic generation of descriptive enhancement prompts based	Must have	Ensure that the right enhanced image is generated based on the enhancement prompt
FR	Similarity Validation	Similarity score should be met	Must Have	Implement Structural Similarity Index Measure (SSIM) and Learned Perceptual Image Patch Similarity (LPIPS) for validating structural and perceptual fidelity of enhanced images.

3.2 Proposed Methodology

This research project aims to systematically evaluate and enhance the aesthetic appeal of product images on the Floward e-commerce platform using stable diffusion models. The proposed methodology integrates data preparation, aesthetic evaluation, enhancement text generation, image enhancement via three stable diffusion variants, and a rigorous re-evaluation process to select the optimal model. The steps are detailed below, ensuring a robust pipeline from raw data to enhanced outputs, validated against both objective and subjective criteria.

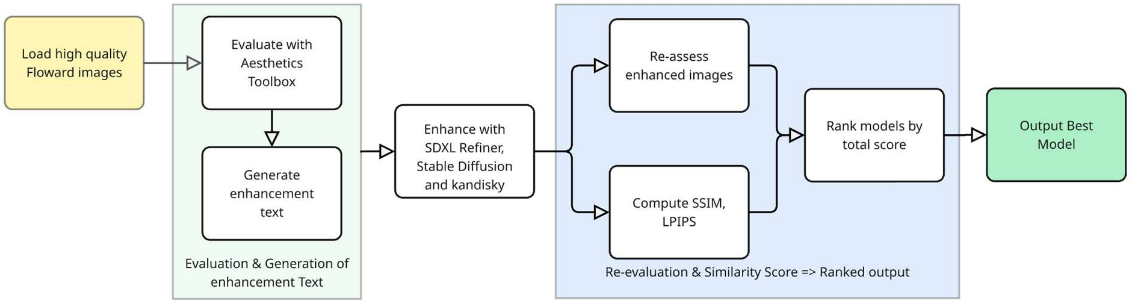

Load high quality Floward images
Evaluate with Aesthetics Toolbox
Generate enhancement text
Evaluation & Generation of enhancement Text
Enhance with SDXL Refiner, Stable Diffusion and kandisky
Re-assess enhanced images
Compute SSIM, LPIPS
Rank models by total score
Re-evaluation & Similarity Score => Ranked output
Output Best Model

Figure 10: Proposed Methodology process flow.

1. Data Loading and Preparation

This is the first step where data is collect and preprocess a representative dataset of Floward product images for evaluation and enhancement. The images are downloaded are publically available and are provided in bulk with the product description so that future evaluation can be done in an effective manner.

In the data collection, high-resolution product images of flower bouquets, chocolates, watches, gift boxes, toys is downloaded to the local server from Floward's existing catalog where each image with 512x512 pixels to match stable diffusion input requirements. The preprocessing involves standardizing image format to Jpeg and validating the image dimensions are 512x512 which ensure consistent aspect ratio (1:1), and remove any external metadata or watermarks that could interfere with evaluation. The preprocessing will also include upscaling lower resolution images if needed using bicubic interpolation. The finally processed images are stored in a structured directory floward_dataset/original/ and floward_dataset/processed with unique identifiers as given by Floward to have reference to the produce description. The output of will be a clean dataset of original images ready for evaluation.

2. Initial Aesthetic Evaluation

The second step is to evaluate the image, the original images are evaluated using the Aesthetics Toolbox to identify aesthetic deficiencies. The setuo is to implement the Aesthetics Toolbox (Redies et al., 2024) in a Python environment with dependencies (e.g., OpenCV, NumPy, PyTorch for deep-learning features).

The evaluation metrics comprises of background Color which is identified using LAB (L, a, b) values, brightness identified with RMS Contrast is compared to the standard deviation of luminance, the sharpness of image is evaluated with the Edge Density, Complexity, EOE, the color Entropy determines the color richness, the fourier Spectrum Slope determines the spatial frequency distribution and balance and symmetry evaluates the mirror symmetry and balance percentages.

Sample Output: For an image (e.g., bouquet_001.jpg), obtain values like L=79.347, RMS=15.808, Edge Density=52.914, etc.

The final output will be in form of a CSV file IMGID_EVAL_DDMMYYHHMM.csv will be generated with image name, expected metrics value to the desired values for all images.

3. Comparison and Enhancement Text Generation

This step the evaluation results are compared with the Floward photo guide metrics against ideal ranges to generate precise enhancement text for each image. There are three stages to this comparison. In the first stage, the results are compared with the expected value, deviation is identified and respective text is generated for the deviation identified as mentioned in below three sections.

Comparison Values: Define ideal ranges from the Aesthetics Toolbox (as per Section 2.4):

- LAB: L=88–92, a=-1–3, b=8–12.
- RMS Contrast: 15–30 (target ~22).
- Edge Density: >500, Complexity: >5, EOE: >4.
- Color Entropy: 4–6.
- Fourier Slope: -2 to -3.
- Symmetry: <5%, Balance: <20%.

Deviation Analysis: For each image, calculate deviations (e.g., L=79.347 vs. L≈90 requires +10.653 brightness increase).

Text Generation: Formulate enhancement text based on deviations:

- Example: If L=79.347, RMS=15.808, Edge Density=52.914 → "Increase brightness to L≈90, enhance RMS contrast to ~22, sharpen edges to density >500."
- Logic: Use conditional rules (e.g., if L < 88, append "increase brightness"; if Edge Density < 500, append "sharpen edges").

Output: A list of enhancement texts (e.g., enhancement_text.csv) paired with original images name. Sample text "Increase brightness to L≈90, enhance RMS contrast to ~22, sharpen edges to density >500").

4. Image Enhancement with Stable Diffusion Models

The next step is to feed the original image and enhancement text as input to the three stable diffusion variants. Apply three stable diffusion variants to enhance images based on enhancement text and original image. The model setup for each latent diffusion variant are defined in hugging face definition. Further to it the latent diffusion model SDXL Refiner stabilityai/stable-diffusion-xl-refiner-1.0, Stable diffusion radames/stable-diffusion-v1-5-img2img and kandisky radames/kandinsky-2-1-img2img models are available in their respective links of Hugging face.

The enhancement pipeline is built as below

- Input: Original image $\varkappa$ and enhancement text (e.g., "Increase brightness to L≈90, sharpen edges").
- VAE Encoder: Compress $\varkappa$ to latent z_0.
- Partial Forward Diffusion: Add noise to z_t, (e.g., 20% of T, preserving structure).
- CLIP Conditioning: Encode text into a conditioning vector c.
- Reverse Diffusion: U-Net denoises to z_t, to z'_0 guided by c.
- VAE Decoder: Reconstruct enhanced image $\varkappa'$.

Process all three models with images and enhancement text as input and saving outputs in an excel. The output is 150 enhanced images (50 images × 3 models) in a directory floward_dataset/enhanced/.

Post enhancement the images will be fed to both re-evaluation and Similarity Score validation simultaneously so that we have the optimized results output compared.

5. Re-Evaluation of Enhanced Images

In this step assessment of enhanced images using the Aesthetics Toolbox to verify improvements. Re-Run Aesthetics Toolbox to Compute the same metrics (LAB, RMS, Edge Density, etc.) for all enhanced images. The sample output for the enhanced image e.g., L=89.5, RMS=21.9, Edge Density=510 should be stored in the same image level file to ensure that the old and new values are compared. Assign a compliance score per metric based on target ranges (e.g., L=89.5 scores 1 if 88≤L≤92, 0 otherwise; aggregate as average compliance across metrics). The final output after re-evaluation is to be stored in the evaluation sheet where the original evaluation score is recorded against each images.

6. Similarity Score Validation

In this step the original image is compared with the enhanced images to ensure structural and perceptual fidelity. The metrics defined for the similarity score validation is listed below.

SSIM: Compute Structural Similarity Index (target >0.9) using scikit-image.

LPIPS: Calculate Learned Perceptual Image Patch Similarity (target <0.2) with a pre-trained VGG network.

Execution: For each pair (e.g., bouquet_001.jpg vs. bouquet_001_sdxl.jpg), calculate SSIM, LPIPS, and collect MOS scores.

Output: A CSV file (e.g., similarity_scores.csv) with SSIM and LPIPS values per enhanced image.

7. Model Selection

In this step the model selection is taken up to identify the best stable diffusion model based on re-evaluation and similarity scores.

The scoring framework involves a combined computed value of original evaluation to the re-evaluation to the similarity score. Average compliance score across Aesthetics Toolbox metrics (max 1.0) along with the Weighted average of SSIM (40%), LPIPS (30%), and MOS (30%) normalized to 1.0 (e.g., SSIM=0.95 → 0.95, LPIPS=0.15 → 0.85, MOS=4.2 → 0.84) is taken to form the total Score with Combine re-evaluation (50%) and similarity (50%) scores (max 1.0).

Comparison: Rank models per image and overall (e.g., mean total score across 50 images).

Selection Criteria: Highest total score indicates the best model, with "Must Have" requirements (LAB, RMS, Edge Density, SSIM, MOS) as thresholds.

Output: A report (e.g., model_selection.pdf) detailing scores and the selected model (e.g., "SDXL Refiner: 0.92").

Chapter 4

Project Plan

4.1 Introduction

This chapter examines the wider implications and planning for the research project, which aims to enhance Floward’s product images such as flower bouquets, gift boxes and other gift items. Beyond technical execution, the project raises professional, ethical, legal, and social considerations that must be addressed to ensure success. Effective planning, including timelines and risk management, is also critical to meet objectives like improving brightness and sharpness while preserving authenticity. This section outlines these aspects in a few words.

4.2 Professional, Legal, Ethical and Social Issues

The use of AI for image enhancement in Floward's e-commerce platform introduces key considerations.

4.2.1 Professional issues

Professionals must uphold standards by ensuring model accuracy and transparency in reporting results from the Aesthetics Toolbox (Redies et al., 2024) openly in the prototype. Competence in AI tools is required to address issues like computational efficiency, maintaining Floward's trust in the solution.

4.2.2 Ethical Issues

Ethical concerns include avoiding misrepresentation through excessive enhancement, mitigated by preserving structural fidelity. Ensure that the data privacy and bias in outputs must be managed so that there is fairness across all product images.

4.2.3 Legal Issues

Legal compliance involves adhering to model licenses of respective models used and consumer protection laws to prevent false advertising, ensuring enhanced images reflect actual products. Regional laws like Saudi Arabia's PDPL apply if metadata is involved. It has to be ensured that there are no legal obligations during the prototype development process.

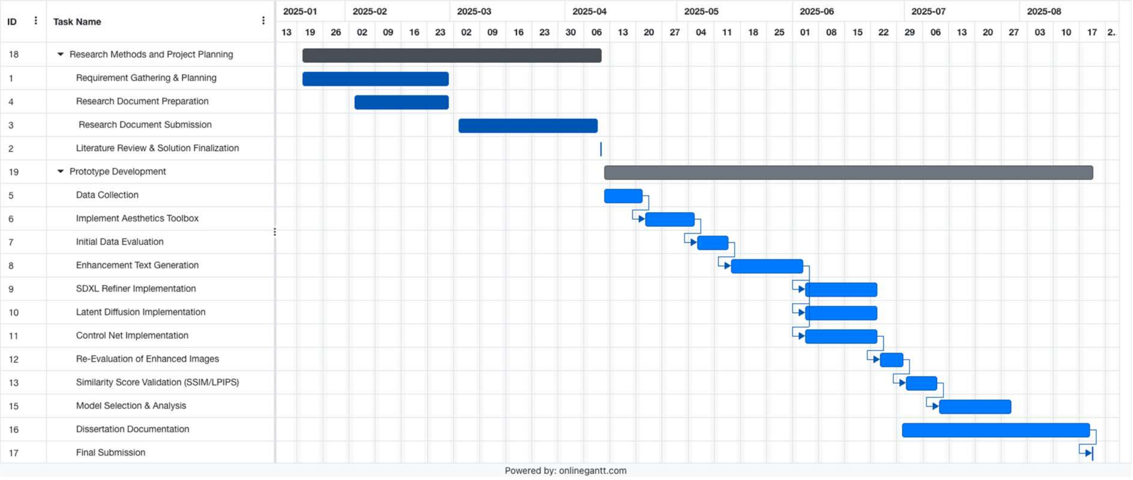
4.2.4 Social Issues

Enhanced images may boost customer appeal but risk setting unrealistic expectations. Automation could streamline workflows for Floward's team, though over-perfection might disconnect consumers from authentic experiences (Zhang & Agrawala, 2023).

4.3 Project Planning

This section outlines the timeline and risks to execute the methodology effectively.

4.3.1 Gantt Chart



Name	Start	Finish	Duration
Research Methods and Project Planning	20/01/2025	10/04/2025	59 day
Literature Review & Solution Finalization	20/01/2025	28/02/2025	30 day
Requirement Gathering & Planning	03/02/2025	28/02/2025	20 day
Research Document Preparation	03/03/2025	09/04/2025	28 day
Research Document Submission	10/04/2025	10/04/2025	1 day
Prototype Development	11/04/2025	20/08/2025	94 day
Data Collection	11/04/2025	21/04/2025	7 day
Implement Aesthetics Toolbox	22/04/2025	05/05/2025	10 day
Initial Data Evaluation	06/05/2025	14/05/2025	7 day
Enhancement Text Generation	15/05/2025	03/06/2025	14 day
SDXL Refiner Implementation	04/06/2025	23/06/2025	14 day
Stable Diffusion Implementation	04/06/2025	23/06/2025	14 day
Kandinsky Implementation	04/06/2025	23/06/2025	14 day
Re-Evaluation of Enhanced Images	24/06/2025	30/06/2025	5 day
Similarity Score Validation (SSIM/LPIPS)	01/07/2025	09/07/2025	7 day
Model Selection & Analysis	10/07/2025	29/07/2025	14 day
Dissertation Documentation	30/06/2025	19/08/2025	37 day
Final Submission	20/08/2025	20/08/2025	1 day

4.3.2 Risk Analysis

Risks are categorized as technical and non-technical, with mitigations. The list of technical and non-technical risk are as below.

Technical Risks

Risk	Likelihood	Impact	Mitigation
Model Failure	Medium (40%)	High	Test pilot dataset Use fallback model if needed.
Computational Overload	Medium (50%)	High	Optimize with GPU (e.g., RTX 4090); Reduce batch size and get cloud environment if required.
Model License	Low(20%)	Medium	License of the models are restricted for education purpose. Floward company has to buy license from respective company for implementations.
Tools libraries	Medium(30%)	Medium	Compatibility check has to be done for all python version to the cuda, tensorflow and other libraries
Data extraction	Medium (50%)	High	Data has to be extracted from the company domain to workspace to conduct experiments

Non-Technical Risks

Risk	Likelihood	Impact	Mitigation
Ethical Misalignment	Low (20%)	High	Limit enhancements (SSIM >0.9); disclose AI use.
Timeline Delays	Medium (50%)	Medium	Add buffer (Weeks 8–9); prioritize critical tasks.
Noisy data	Medium (50%)	Medium	Preprocessing timelines might be impacted if the right dimension data isn't provided.

References

Redies, C., Bartho, R., Koßmann, L., Spehar, B., Hübner, R., Wagemans, J. and Hayn-Leichsenring, G.U. (2024) A toolbox for calculating objective image properties in aesthetics research. Available at: https://doi.org/10.48550/arXiv.2408.10616 (Accessed: 11 Feb 2025).

Nichol, A. and Dhariwal, P., 2021. Improved Denoising Diffusion Probabilistic Models. *arXiv preprint arXiv:2102.09672*. Available at: https://arxiv.org/abs/2102.09672 [Accessed 4 March 2025].

Ho, J., Jain, A. and Abbeel, P., 2020. Denoising Diffusion Probabilistic Models. *Advances in Neural Information Processing Systems*, 33, pp.6840-6851. Available at: https://arxiv.org/abs/2006.11239 [Accessed 4 March 2025].

Kingma, D.P. and Welling, M., 2013. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114. Available at: https://ui.adsabs.harvard.edu/abs/2013arXiv1312.6114K/abstract [Accessed at: 4 March 2025

Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S. and Poole, B., 2020. Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456. *Available at: https://arxiv.org/abs/2011.13456 [Accessed 11 March 2025]*

Rombach, R., Blattmann, A., Lorenz, D., Esser, P. and Ommer, B., 2022. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 10684-10695). Available at: https://arxiv.org/abs/2112.10752 [Accessed 4 March 2025].

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C. and Chen, M., 2022. Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125, 1(2), p.3. Available at: https://arxiv.org/abs/2204.06125 [Accessed 02-Mar-2025]

Ronneberger, O., Fischer, P., and Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18 (pp. 234-241). Springer international publishing. Available at: https://arxiv.org/abs/1505.04597 [Accessed 02-Mar-2025]

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T. and Ho, J., 2022. Photorealistic text-to-image diffusion models with deep language understanding. Advances in neural information processing systems, 35, pp.36479-36494.

Croitoru, F.A., Hondru, V., Ionescu, R.T. and Shah, M., 2023. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(9), pp.10850-10869.. Available at: https://ieeexplore.ieee.org/abstract/document/10081412 [Accessed 05 March 2025].

Razzhigaev, A., Shakhmatov, A., Maltseva, A., Arkhipkin, V., Pavlov, I., Ryabov, I., Kuts, A., Panchenko, A., Kuznetsov, A. and Dimitrov, D., 2023. Kandinsky: an improved text-to-image synthesis with image prior and latent diffusion. arXiv preprint arXiv:2310.03502. Available at: https://arxiv.org/abs/2310.03502 [Accessed at 02 March 2025]

Wang, Z., Bovik, A.C., Sheikh, H.R. and Simoncelli, E.P., 2004. Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing, 13(4), pp.600-612. Available at: https://doi.org/10.1109/TIP.2003.819861 [Accessed 4 April 2025].

Talebi, H. and Milanfar, P., 2018. NIMA: Neural image assessment. *IEEE transactions on image processing*, *27*(8), pp.3998-4011. Available at: https://ieeexplore.ieee.org/abstract/document/8352823 [Accessed 22 March 2025]

Chiamulera, C., Ferrandi, E., Benvegnù, G., Ferraro, S., Tommasi, F., Maris, B., Zandonai, T. and Bosi, S., 2017. Virtual reality for neuroarchitecture: Cue reactivity in built spaces. Frontiers in psychology, 8, p.185. Available at: https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2017.00185/full. [Accessed 02 March 2025]

Spehar, B., Walker, N. and Taylor, R.P., 2016. Taxonomy of individual variations in aesthetic responses to fractal patterns. *Frontiers in human neuroscience*, *10*, p.350. Available at: https://www.frontiersin.org/journals/human-neuroscience/articles/10.3389/fnhum.2016.00350/full [Accessed 22 March 2025]

Tang, X., Luo, W. and Wang, X., 2013. Content-based photo quality assessment. *IEEE Transactions on Multimedia*, *15*(8), pp.1930-1943. Available at: https://ieeexplore.ieee.org/abstract/document/6544270 [Accessed 22 March 2025].

Datta, R., Joshi, D., Li, J. and Wang, J.Z., 2006. Studying aesthetics in photographic images using a computational approach. In *Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006, Proceedings, Part III 9* (pp. 288-301). Springer Berlin Heidelberg. Available at: https://link.springer.com/chapter/10.1007/11744078_23 [Accessed 10 Feb 2025]

Sheikh, H.R., Sabir, M.F. and Bovik, A.C., 2006. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on image processing*, *15*(11), pp.3440-3451. Available at: https://ieeexplore.ieee.org/abstract/document/1709988 [Accessed 22 March 2025]

Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y. and Ermon, S., 2021. Sdedit: Guided image synthesis and editing with stochastic differential equations. arXiv preprint arXivrr:2108.01073. Available at: https://arxiv.org/abs/2108.01073 [Accessed 22 March 2025].

Murray, N., Marchesotti, L. and Perronnin, F., 2012, June. AVA: A large-scale database for aesthetic visual analysis. In *2012 IEEE conference on computer vision and pattern recognition* (pp. 2408-2415). IEEE. Available at: https://ieeexplore.ieee.org/abstract/document/6247954 [Accessed: 15 Feb 2025]

Rombach, Robin., Esser, Patrick., 2022: Stable Diffusion v1.5 img2img. [online] Hugging Face. Available at: https://huggingface.co/radames/stable-diffusion-v1-5-img2img [Accessed 10 Apr. 2025]

Locher, P. J., Stappers, P. J., & Overbeeke, K. (1999). An empirical evaluation of the visual rightness theory of pictorial composition. Acta Psychologica, 103, 261-280. Available at: https://www.sciencedirect.com/science/article/abs/pii/S000169189900044X (Accessed: 9 Feb 2025).

Ly, B.C.K., Dyer, E.B., Feig, J.L., Chien, A.L. and Del Bino, S., 2020. Research techniques made simple: cutaneous colorimetry: a reliable technique for objective skin color measurement. *Journal of Investigative Dermatology*, *140*(1), pp.3-12. Available at: https://www.sciencedirect.com/science/article/pii/S0022202X19333974 (Accessed: 9 Feb 2025).

Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J. and Rombach, R., 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*. Available at: https://arxiv.org/abs/2307.01952 [Accessed 02 March 2025]

Li, M., Lin, Y., Zhang, Z., Cai, T., Li, X., Guo, J., Xie, E., Meng, C., Zhu, J.Y. and Han, S., 2024. SVDQuant: Absorbing Outliers by Low-Rank Components for 4-Bit Diffusion Models. arXiv:2411.05007. Available at: https://arxiv.org/abs/2411.05007 [Accessed 02 March 2025]

Lu, X., Lin, Z., Jin, H., Yang, J. and Wang, J.Z., 2014, November. Rapid: Rating pictorial aesthetics using deep learning. In *Proceedings of the 22nd ACM international conference on Multimedia* (pp. 457-466). Available at: https://dl.acm.org/doi/10.1145/2647868.2654927 [Accessed: 15 Feb 2025]

Choi, J., Kim, S., Jeong, Y., Gwon, Y. and Yoon, S., 2021. ILVR: Conditioning Method for Denoising Diffusion Probabilistic Models. arXiv preprint arXiv:2108.02938. Available at: https://arxiv.org/abs/2108.02938 [Accessed at 02 March 2025]

Stability AI. (2023). Stable Diffusion XL Refiner-1.0. https://github.com/Stability-AI/stablediffusion

Fairchild, M.D., 2013. Color appearance models. John Wiley & Sons. Accessible at: https://www.sciencedirect.com/science/article/pii/S0300571210001168?casa_token=I1GhnI0oDrYAAAAA:94bsBNjQ1X36N2SyRpbWBzQLcNN-njqu4dcZrAXdK_YDsQN833rqcbA3V95YgTC3Yq0sDFcZ1R0 [Accessed 4 April 2025].

Zhang, R., Isola, P., Efros, A.A., Shechtman, E. and Wang, O., 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.586-595. Available at: https://arxiv.org/abs/1801.03924 [Accessed 4 April 2025].

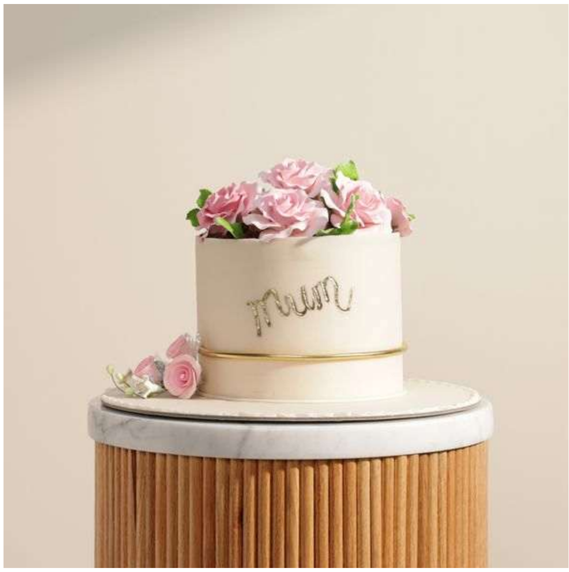
Sohl-Dickstein, D., Weiss, E., Maheswaranathan, N., & Ganguli, S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics. *International Conference on Machine Learning*, 2256-2265. https://arxiv.org/abs/1503.03585

Zhang, L. and Agrawala, M., 2023. Adding Conditional Control to Text-to-Image Diffusion Models. arXiv preprint arXiv:2302.05543. Available at: https://arxiv.org/abs/2302.05543 . [Accessed 4 April 2025].

Appendix

Toolbox Output for input image:

Toolbox output for below image from Floward.



img_file	97da0d97-a1a0-4da4-9ace-8c707998797c.jpg

Image size (pixels)	1200
Aspect ratio	1
RMS contrast	15.808
Lightness entropy	6.145
Complexity	3.757
Edge density	52.914
mean R channel	213.259
mean G channel	194.022
mean B channel (RGB)	174.627
mean L channel	79.347
mean a channel	4.351
mean b channel (Lab)	12.742
mean H channel	0.0888
mean S channel	0.2051
mean V channel	0.8367
std R channel	33.432
std G channel	45.246
std B channel	53.67
std L channel	15.808
std a channel	6.851
std b channel (Lab)	9.586
std H channel	0.02355
std S channel	0.1918
std V channel	0.1304
Color entropy	3.567
Mirror symmetry	3.95
Balance	17.82
DCM distance	22.284
DCM x position	0.04167
DCM y position	-0.1033
CNN symmetry left-right	0.5698
CNN symmetry up-down	0.2068
CNN symmetry left-right & up-down	0.1951
Fourier slope	-2.523
Fourier sigma	0.04882
2D Fractal dimension	1.209
3D Fractal dimension	2.431
Self-similarity (PHOG)	0.23
Self-similarity (CNN)	0.3617
Homogeneity	74.936
Anisotropy	0.001281
1st-order EOE	3.508

2nd-order EOE	3.752
Sparseness	0.001062
Variability	2.542e-05
gray_scale	0
upscaled	1
external_color_profile_found	1