**NANYANG TECHNOLOGICAL UNIVERSITY**

**AUTOMATED IMAGE GENERATION**

Hill Seah Wen Qi

College of Computing and Data Science 2024

**NANYANG TECHNOLOGICAL UNIVERSITY**

**AUTOMATED IMAGE GENERATION  
(CCDS24-0163)**

Submitted in Partial Fulfillment of the Requirements for the Degree of Bachelor of Computing (Hons) in Data Science and Artificial Intelligence of the Nanyang Technological University

by

Hill Seah Wen Qi

College of Computing and Data Science 2024

22nd March 2025

# Abstract

Multi-conditional image generation aims to synthesize images that satisfy diverse conditions, such as textual descriptions, segmentation masks, and landmark constraints. Current training-free approaches, which rely on off-the-shelf and open-source pre-trained networks to provide guidance, perform well for single conditions but fail to capture the complex interdependencies among multiple conditions. The purpose of this project is to develop a novel framework that overcomes these limitations by effectively modeling the interactions between conditions. To achieve this, the project analyzes existing methods and introduces an innovative design: a time-independent approximated energy guidance function enhanced with interaction modeling. This function captures non-linear dependencies and guides an iterative denoising process to progressively refine the generated images. Experimental results indicate that our approach outperforms existing techniques, producing images that are both coherent and condition-consistent. In conclusion, the framework not only resolves key challenges in multi-conditional image synthesis but also provides a basis for future research, with recommendations to further explore adaptive energy functions for even broader applicability.

# Acknowledgements

This project is standing on the shoulders of giants and would not have been possible without the foundational knowledge, resources, and inspiration provided by the authors and researchers whose works have been referenced in this project. Their pioneering contributions have laid the groundwork for advancements like mine. Hence, I would like to extend my deepest gratitude to them for their help in the completion of this project.

Additionally, I am extremely grateful to my project supervisor Professor Lu Shijian, and my student mentor Xu Muyu for their guidance and assistance in this endeavor, making this journey a memorable and fulfilling experience.

Table of Contents

[Abstract 3](#_Toc193233809)

[Acknowledgements 3](#_Toc193233810)

[1. Introduction 6](#_Toc193233811)

[1.1 Problem Statement 6](#_Toc193233812)

[1.2 Objective 6](#_Toc193233813)

[1.3 Scope 7](#_Toc193233814)

[2. Related Work 8](#_Toc193233815)

[2.1 Conditional Score Based Diffusion Models 8](#_Toc193233816)

[2.2 Energy Diffusion Guidance 8](#_Toc193233817)

[2.3 Time-Independent Distance Functions Approximating Time-Dependent Energy Guidance 9](#_Toc193233818)

[3. Proposed Methodology 10](#_Toc193233819)

[3.1 Multi-Conditional Training-Free Image Generation Modelling 10](#_Toc193233820)

[3.1.1 Multi-Conditional Energy Guidance Model – No Interaction Modelling 10](#_Toc193233821)

[3.1.2 Improved Multi-Conditional Energy Guidance Model – Interaction Modelling 10](#_Toc193233822)

[3.2 Interaction Models 11](#_Toc193233823)

[3.2.1 Simple Similarity Measures 11](#_Toc193233824)

[3.2.2 Polynomial Functions 11](#_Toc193233825)

[3.2.3 Sigmoid Functions 12](#_Toc193233826)

[3.2.4 Gaussian Kernels 12](#_Toc193233827)

[3.3 Configuration 13](#_Toc193233828)

[3.3.1 Pre-trained Models 13](#_Toc193233829)

[3.3.2.1 Multi-Condition Combinations 14](#_Toc193233830)

[3.3.2.2 Experimental Setup 15](#_Toc193233831)

[3.3.2.3 Hyperparameter Optimization 15](#_Toc193233832)

[4. Results 16](#_Toc193233833)

[4.1 Hyperparameters 16](#_Toc193233834)

[4.2 Quantitative Results 17](#_Toc193233835)

[4.2.1 Comparison between Interaction Models 17](#_Toc193233836)

[4.2.2 Baseline Comparison 18](#_Toc193233837)

[4.3 Qualitative Results 19](#_Toc193233838)

[4.3.1 Interaction Modelling Comparison 19](#_Toc193233839)

[4.3.2 Baseline Comparison 21](#_Toc193233840)

[5. Evaluation 22](#_Toc193233841)

[5.1 Models 22](#_Toc193233842)

[5.2 Experimental Limitations 24](#_Toc193233843)

[5.3 Future Work 24](#_Toc193233844)

[5.3.1 Anisotropic Gaussian Kernels 24](#_Toc193233845)

[5.3.2 Dynamic and Adaptive Weight Adjustment 25](#_Toc193233846)

[5.3.3 Extension to other generative tasks 26](#_Toc193233847)

[6. Conclusion 26](#_Toc193233848)

[7. Project Schedule 27](#_Toc193233849)

[8. Appendices 32](#_Toc193233850)

[8.1 Appendix A - Code Snippets 32](#_Toc193233851)

[8.1.1 Approximated clean image from an intermediate noisy image 32](#_Toc193233852)

[8.1.2 Multi-Conditional Energy Guidance with Interaction Modelling 33](#_Toc193233853)

[8.1.3 Condition-Specific Classes 33](#_Toc193233854)

[8.1.4 Final Sampling Formula 36](#_Toc193233855)

[8.1.5 Interaction Models 36](#_Toc193233856)

[8.2 Appendix B – Old Project Schedule 38](#_Toc193233857)

[9. References 40](#_Toc193233858)

[10. TODOS (Post-Interim Report) 42](#_Toc193233859)

Table of Tables

[Table 1. Pre-trained diffusion models used. 13](#_Toc193233901)

[Table 2. Pre-trained condition-specific models used extract distance information between given condition and image. 14](#_Toc193233902)

[Table 3. Conditional groups. 15](#_Toc193233903)

[Table 4. Search space and reasonings for each hyperparameter 16](#_Toc193233904)

[Table 5. Configuration of hyperparameter values. 17](#_Toc193233905)

[Table 6. Results of the interaction models. 18](#_Toc193233906)

[Table 7. Comparison Between Approximated Energy Function Guidance with the Best Interaction Model vs. Baseline (TediGAN) 19](#_Toc193233907)

[Table 8. Evaluation the pros and cons of each proposed interaction-modelling methods in the context of multi-conditional image generation tasks 24](#_Toc193233908)

[Table 9. Finalized Project Schedule 32](#_Toc193233909)

[Table 10. Old Project Schedule, 31 August 2024 41](#_Toc193233910)

Table of Figures

[Figure 1. Illustration of Sequential Multi-Conditional Image Generation 20](#_Toc193233867)

[Figure 2. Segmentation Map + Text Prompt Multi-Conditional Image Generation Result 20](#_Toc193233868)

[Figure 3. Sketch + Segmentation Map Multi-Conditional Image Generation Result 21](#_Toc193233869)

[Figure 4. Face ID + Landmark + Text Prompt Multi-Conditional Image Generation Result 21](#_Toc193233870)

[Figure 5. Sketch + Text Prompt Baseline Comparison 22](#_Toc193233871)

[Figure 6. Gantt Chart of Finalized Project Schedule 32](#_Toc193233872)

[Figure 7. Approximated clean image from intermediate noisy image 33](#_Toc193233873)

[Figure 8. Algorithm for multi-conditional energy guidance with interaction modelling 34](#_Toc193233874)

[Figure 9. Text condition class 34](#_Toc193233875)

[Figure 10. Facial landmark condition class 35](#_Toc193233876)

[Figure 11. Facial segmentation map condition class 35](#_Toc193233877)

[Figure 12. Facial ID condition class 36](#_Toc193233878)

[Figure 13. Sketch condition class 36](#_Toc193233879)

[Figure 14. Final sampling formula 36](#_Toc193233880)

[Figure 15. Euclidean distance interaction model 37](#_Toc193233881)

[Figure 16. Cosine similarity interaction model 37](#_Toc193233882)

[Figure 17. Pearson correlation interaction model 38](#_Toc193233883)

[Figure 18. Polynomial function interaction model 38](#_Toc193233884)

[Figure 19. Sigmoid function interaction model 38](#_Toc193233885)

[Figure 20. Gaussian kernel interaction model 39](#_Toc193233886)

# Introduction

## Problem Statement

Multi-conditional image generation has emerged as a critical area in computer vision, where the goal is to generate images that simultaneously satisfy multiple user-defined constraints.

Training-free approaches are rapidly gaining popularity due to advantages over training-required methods, as they eliminate the need for extensive datasets and computationally expensive training processes, making them faster to deploy, more cost-effective, and easier to adapt to new tasks or domains.

For single-condition image generation, training-free approaches often leverage off-the-shelf and open-source pre-trained networks to estimate the distance between an intermediate image and the condition, guiding the iterative denoising processes (Chung et al, 2023). However, extending these techniques to multi-conditional image generation presents challenges. Unlike single-condition tasks, where adhering to a single constraint is sufficient, multi-conditional generation requires optimizing multiple constraints simultaneously, which often leads to conflicts, trade-offs, or failures (e.g., aligning text descriptions while maintaining geometric consistency). Without explicit training, these methods struggle to balance constraints, resolve conflicts, and adapt to complex, real-world multimodal scenarios.

Some existing training-free methods attempts to address this limitation, for example through ControlNet and Adapter Modules (Yang et al, 2024), or with cross-attention mechanisms, to help balance constraints like depth, edges, and text alignment. However, these methods often rely on heuristic tuning and static constraint weighting, leading to conflicts, coherence loss, and limited adaptability.

Hence, this necessitates the development of new strategies to manage multi-condition interdependencies, dynamically resolve conflicts, and adjust constraints among conditions to ensure robust, high-quality image outputs that satisfy all specified requirements.

## Objective

This project aims to overcome the limitations of training-free multi-conditional image generation by integrating an approximated energy-based guidance function with effective interaction modeling within the diffusion model denoising process. This approach enables the dynamic handling of multiple, potentially conflicting conditions, ensuring coherent, high-quality image synthesis that satisfies all specified constraints. Ultimately, the goal is to enhance adaptability and stability in complex, real-world multimodal scenarios without requiring retraining.

We will evaluate our different proposed models to identify the best-performing approach based on qualitative results, such as visual fidelity and consistency in constraint adherence, as well as quantitative metrics like Fréchet Inception Distance (FID) and constraint-specific distances. This will allow us to comprehensively assess the balance between image quality and the successful integration of multiple conditions.

## Scope

Given resource limitations in GPU capacity, which lead to extended image generation times for larger diffusion models with general image generation capabilities (such as StableDiffusion or ControlNet), we opted to conduct facial image generation using a smaller pre-trained unconditional human face diffusion model.

The conditioning inputs for our multi-conditional image generation experiments were selected to provide a diverse and complementary set of information:

1. Face ID – Encodes identity-specific facial features in a numerical representation, ensuring consistency in facial appearance.
2. Sketch – Provides a structural outline of the subject, capturing overall shape and contours.
3. Landmark – Defines spatial key points (e.g., eyes, nose, mouth) to enforce geometric accuracy in facial features.
4. Segmentation Map – Specifies region-based attributes, guiding the model in differentiating facial parts and background elements.
5. Text – Offers high-level semantic descriptions, providing flexible and interpretable guidance for image generation.

However, the results and contributions of this project remain translatable to general image generation, as the methods and frameworks developed are model-agnostic and condition-agnostic, and can be applied to larger models. Additionally, the principles of interaction-aware image generation explored in this project are not limited to facial images but can be generalized to broader contexts, enabling scalability to more complex image generation tasks.

# Related Work

## Conditional Score Based Diffusion Models

For unconditional score-based diffusion models (SBDM) operating on score theory, its goal is to learn and estimate a time-dependent score function that guides the denoising phase of a noisy image to at time step during the iterative sampling process. The sampling formula is denoted as follows:

where is a hyperparameter and represents random Gaussian noise. (Song et al, 2021)

For conditional diffusion, a corrective gradient is added to the sampling formula Eq. (1) to guide to a hyperplane in the data space that aligns with the condition . (Song et al, 2021).

Training-required methods often retain the time-dependent nature of the corrective gradient , learning it through approaches like classifier training. In contrast, training-free methods aim to approximate the corrective gradient using time-independent functions.

## Energy Diffusion Guidance

One alternative method to model the corrective gradient , would be to use an energy function as follows:  
where represents a normalizing constant, represents the positive temperature constant and represents an energy function measuring the similarity between a given condition and a noisy image (LeCun et al, 2006).

The energy function value decreases as the similarity between and increases, reaching the value zero when and are perfectly similar. This, the corrective gradient can be remodeled to energy guidance as (LeCun et al, 2006).

The final sampling formula that incorporates energy guidance into the sampling formula Eq. (1) is as follows:  
where represents the learning rate of the energy guidance term.

## Time-Independent Distance Functions Approximating Time-Dependent Energy Guidance

To obtain the energy guidance function, most training-required methods revolve around training classifiers to calculate a time-dependent distance measuring function to approximate it, where represents the trained parameters of the classifier. (Dhariwal & Nichol, 2021). This is particularly problematic as it is extremely difficult to find an existing pre-trained model for the noisy image to ensure training-free. To circumvent this, we can estimate the time-dependent energy guidance with time-independent distance functions through a series of approximations.

Unlike time-dependent networks, time-independent functions for measuring distances in clean data, , are widely accessible (Yu et al, 2023). Open-source pre-trained models, such as those for classification, text encoding, segmentation, and face identification, are commonly available and highly effective for working with clean images.

First, we can approximate the time-dependent distance function with time-independent distance function where represents the pre-trained parameters, as follows:  
and this is reasonable because if the noisy image is close to the condition , then the corresponding clean image should also be close to (Yu et al, 2023).

Next, we need to approximate a clean image corresponding to an intermediate noisy image for each time step as follows:  
where and is the pre-trained score estimator (Chung et al, 2022).

Finally, we can combine the results to approximate the time-dependent energy guidance function with a time-independent distance function provided by the condition-specific open-source model. For a singular condition *,* we get (Yu et al, 2023).

# Proposed Methodology

## Multi-Conditional Training-Free Image Generation Modelling

### **Multi-Conditional Energy Guidance Model – No Interaction Modelling**

A primitive approach in modelling a multi-conditional energy guidance formula would be to extend the singular-conditional energy guidance formula as a weighted sum of the different distance functions for each of the respective conditions as follows:

where represents the distance between condition, and the approximated clean image at time step , computed by a pre-trained network that is specific to condition , and represents the weighting factor of .

For example, if represents a text condition, then then pre-trained network

could be a CLIP embedding model, and could be a Euclidean distance value between the CLIP embedding of and .

However, the primary limitation of this primitive model lies in its reliance on the naive assumption that all conditions are mutually independent and non-conflicting. Consequently, it struggles to generate high-quality images under diverse conditions, particularly when those conditions exhibit complex, non-linear interdependencies.

### **Improved** **Multi-Conditional Energy Guidance Model – Interaction Modelling**

To overcome the primary limitation and improve the multi-conditional energy guidance model, we can account for the interactions between all combinations of conditions as follows:

where represents a function that models the interactions between conditions , and the approximated clean image in their respective spaces, and represents a weighting factor of the interactions between .

## Interaction Models

To maintain the training-free nature of image generation in this project, we deliberately avoid interaction modelling methods that require training, such as attention mechanisms, graph-based models, bilinear models, latent factor models, or any other training-dependent neural networks. These methods demand additional training time and data, which would undermine the core advantages of the training-free conditional image generation framework.

### **Simple Similarity Measures**

A straight forward approach to interaction modelling is by simply computing the similarities between the different conditions with a chosen similarity metric. Here, we propose trying 3 different similarity metrics as follows:

1. Euclidean distance,
2. Cosine similarity,
3. Pearson correlation,

In general, these simple similarity measures are quick, interpretable, and computationally efficient interaction modelling in tasks where the relationships between features are simple or linear. However, for image conditions that require complex, non-linear, or higher-order interactions, these methods tend to still produce unsatisfiable results.

### **Polynomial Functions**

Polynomial functions can typically be used to model interactions by expanding features into a higher-dimensional space. They are particularly useful in capturing for complex and non-linear relationships (Shawe-Taylor & Cristianini, 2004).

Using polynomial functions to model the interaction terms between each possibly dependent condition and , we get the model as follows:

where represents the degree of the polynomial and represents a constant that controls the flexibility of the polynomial.

### **Sigmoid Functions**

Sigmoid functions are another popular method for modelling interactions between inputs, particularly for capturing non-linear relationships (Lin & Lin, 2003). They are inspired by the activation functions used in neural networks and can model complex dependencies between conditions. It is especially useful in situations where the interactions between conditions exhibit saturating behavior.

Using sigmoid functions to model the interaction terms between each possibly dependent condition and , we get the model as follows:

where represents the scaling factor that controls the sensitivity to the input distance and represents a bias term.

### **Gaussian Kernels**

The gaussian kernel is a radial basis function (RBF) that is widely used in machine learning to measure similarity between inputs in a smooth and interpretable manner, defined as

The Gaussian kernel computes a similarity score between two inputs and , where σ is a hyperparameter standard deviation term that controls the sensitivity to differences in input. Its smooth exponential decay enables effective modelling of nuanced relationships in high-dimensional feature spaces (Rasmussen & Williams, 2006).

Using Gaussian kernels to model the interaction terms between each possibly dependent condition and , we get

where and are Gaussian Kernels capturing the individual effects and is the Gaussian Covariance Factor capturing the dependency between conditions and .

The final multi-conditional energy function model with interaction-aware gaussian kernels is as follows:

## Configuration

### **Pre-trained Models**

#### **Base Unconditional Diffusion Model**

|  |  |
| --- | --- |
| **Model** | **Purpose** |
| Unconditional Human Face Diffusion Model  (Meng et al, 2022) | Base unconditional face generation model.  To test the proposed training-free multi-conditional energy guidance models. |

Table 1. Pre-trained diffusion models used.

#### **Condition Extracting Model**

Recall the proposed multi-conditional energy guidance model in Eq. 7. To ensure training free, we use pre-trained models specific to the condition to calculate .

Hence, for each unique condition type , a specific model is used to extract the given conditional information and the approximated clean image as and respectively and is a pre-determined distance function between and

Concretely,

The specific pre-trained model for each unique condition type is shown below in Table 2.

|  |  |  |
| --- | --- | --- |
| **Model** | **Purpose** |  |
| Open-source Face Segmentation Network (Yu et al, 2018) | Supports the Segmentation map condition.  Generates a facial Segmentation map of the image and the conditional image. | Euclidean distance |
| Open-source Landmark Extractor Network  (Chen, 2021) | Supports the landmark condition.  Generates a facial landmark of the image and the conditional image. | Euclidean distance |
| Open-source Face Identification Network  (Deng et al, 2019) | Supports the facial ID condition.  Generates a Segmentation map of the image and the conditional image. | Euclidean distance |
| Sketch  (Xiang et al, 2022) | Supports the sketch condition.  Generates a sketch of the image and the conditional image. | Euclidean distance |
| CLIP image encoder  (Radford et al, 2021) | Supports textual condition.  Encode the image and text condition into the same CLIP feature space. | Euclidean distance |

Table 2. Pre-trained condition-specific models used extract distance information between given condition and image.

* + 1. **Experiments**

### **Multi-Condition Combinations**

For multi-conditional image generation, we classify conditions into similar groups based on the type of information they provide. Conditions within the same group convey similar features or representations of the image, so testing them together would be redundant or meaningless.

|  |  |  |
| --- | --- | --- |
| **Group** | **Conditions** | **Description** |
| Structural Representation | Face ID,  Sketch | Defines the shape and structure of the subject |
| Spatial Feature | Landmark, Segmentation Map | Describes locations of key points and regions |
| Semantic Description | Text | Provides high-level conceptual information |

Table 3. Conditional groups.

Thus, for testing, we would avoid combining conditions from the same group but evaluate all possible combinations across different groups to ensure diverse and meaningful testing. For example, testing multi-condition combinations (Face ID, Landmark, Text), (Sketch, Landmark, Text), (Face ID, Segmentation Map, Text), (Sketch, Segmentation Map, Text).

### **Experimental Setup**

For each of the Multi-Conditional Energy Guidance Models proposed in Sections 3.1 and 3.2, we generate a set of 100 images for each multi-condition combination outlined in Section 3.3.2.1. To evaluate the quality and accuracy of the generated images, we measure key metrics, including condition-specific distances and Fréchet Inception Distance (FID). This setup enables a comprehensive comparison of different conditioning strategies and their impact on generation performance.

### **Hyperparameter Optimization**

Due to **GPU limitations, hyperparameter tuning** is conducted using a **grid search** over a predefined set of values. This approach ensures a systematic exploration of key parameters while maintaining computational feasibility. The search space for each hyperparameter and the reasonings for them are defined as follows:

|  |  |  |
| --- | --- | --- |
| **Hyperparameter** | **Search Space** | **Reasoning** |
|  | (By ratio, )  1, 10, 100, 1000 | These values span different orders of magnitude to balance guidance strength. |
|  | By ratio, )  1, 10, 100, 1000 | Keeping a similar range ensures meaningful weight adjustments for interaction terms. |
|  | 1, 2, 3, 4, 5 | The polynomial degree should remain low to prevent excessive complexity and overfitting. |
|  | 0.5, 1, 5, 10, 15 | This controls polynomial-modeled interaction terms; a range from small to moderate values helps test expressiveness. |
|  | 0.05, 0.1, 0.5, 1, 2 | The scaling factor for sigmoid-modeled interactions should remain small to maintain stability in optimization. |
|  | 0.5, 1, 5, 10, 15 | Similar to , this adjusts the strength of sigmoid-based interactions. |
|  | 0.5, 0.8, 1 | A reasonable range for standard deviations in Gaussian kernels, balancing smoothness vs. sharpness in interaction modeling. |

Table 4. Search space and reasonings for each hyperparameter

# Results

## Hyperparameters

After evaluating visual and quantitative results through grid search, the hyperparameter configuration in Table 5 was found to produce the most coherent and acceptable images within our resource constraints. While this configuration may not be globally optimal, it serves as a reasonable choice for our experiments. The subsequent results in Sections 4.2 and 4.3 are based on these hyperparameters.

|  |  |  |
| --- | --- | --- |
| **Hyperparameter** | **Value** | **Description** |
|  | = 1000 : 1 : 1000 : 1000 : 10 | Weighting factor for |
|  |  | Weighting factor for |
|  | 3 | Polynomial degree for polynomial-modelled interaction terms |
|  | 1 | Constant for polynomial-modelled interaction terms |
|  | 1 | Scaling factor for sigmoid-modelled interaction terms |
|  | 1 | Constant for sigmoid-modelled interaction terms |
|  | 0.5 | Standard deviation for gaussian-kernel-modelled interaction terms |

Table 5. Configuration of hyperparameter values.

## Quantitative Results

To compare the optimized multi-conditional energy guidance models—with and without interaction modeling—as well as to determine the best interaction modeling approach, we evaluate two key metrics:

1. Average Condition-Specific Distance – Measures how well each generated image adheres to its specific conditioning inputs.
2. Average FID Score – Measures the overall quality and realism of the generated images.

We compute these metrics across all multi-conditional combinations for a set of 100 generated images. This ensures a fair comparison of the models in terms of both fidelity to conditions and visual quality.

### **Comparison between Interaction Models**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Text** | | **Facial ID** | | **Segmentation Map** | | **Landmark** | | **Sketch** | |
| **Interaction Models** | **FID** | **Distance** | **FID** | **Distance** | **FID** | **Distance** | **FID** | **Distance** | **FID** | **Distance** |
| None (Primitive Model) | 112.721 | 21.812 | 158.014 | 84.122 | 129.301 | 2331.756 | 139.192 | 17.342 | 119.087 | 328.919 |
| Euclidean Distance | 162.891 | 34.921 | 187.928 | 132.271 | 146.027 | 2737.189 | 135.928 | 18.911 | 129.139 | 367.716 |
| Cosine Similarity | 130.475 | 24.572 | 144.371 | 73.490 | 118.163 | 2147.203 | 128.305 | 21.380 | 115.978 | 317.192 |
| Pearson Correlation | 148.907 | 29.230 | 148.039 | 83.211 | 131.394 | 2721.823 | 141.283 | 16.101 | 131.371 | 395.908 |
| Polynomial | 86.192 | 17.988 | 102.102 | 71.519 | 91.027 | 2174.201 | 95.318 | 15.283 | 91.378 | 292.830 |
| Sigmoid | 102.391 | 21.695 | 117.920 | 68.355 | 101.273 | 2271.521 | 127.311 | 16.309 | 122.202 | 322.337 |
| Gaussian Kernel | 74.189 | 13.353 | 86.346 | 58.745 | 77.523 | 1824.910 | 81.681 | 14.892 | 89.209 | 248.293 |

Table 6. Results of the interaction models.

As shown in Table 6, multi-conditional energy guidance models with appropriate interaction modelling significantly outperform primitive models without interaction modelling. This is evident from the lower average FID and distance values across the conditions.

Conversely, the table also highlights that inappropriate or insufficient interaction modelling can lead to worse results. For example, models using Euclidean distance, cosine similarity, or Pearson correlation coefficient for interaction modelling generally exhibit higher average FID and distance values compared to models without any interaction modelling.

Notably, Table 6 demonstrates that methods capable of capturing complex and non-linear interactions, such as Polynomial, Sigmoid, and Gaussian kernels, consistently outperform simpler methods that only capture linear relationships. Among these, Gaussian kernels stand out as the most effective, yielding the lowest average FID and distance scores across all conditions.

### **Baseline Comparison**

To further validate our results, we compare our proposed approximated energy function guidance, using the best interaction model (Gaussian kernel), with the baseline model TediGAN. TediGAN is a multi-modal image generation framework that enables text-guided synthesis through StyleGAN inversion and visual-linguistic similarity learning (Xia et al, 2021). It also supports image synthesis from diverse inputs, including segmentation maps, landmarks, and sketches, which aligns with our scope (Xia et al, 2021).

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Text** | | **Facial ID** | | **Segmentation Map** | | **Landmark** | | **Sketch** | |
| **Interaction Models** | **FID** | **Distance** | **FID** | **Distance** | **FID** | **Distance** | **FID** | **Distance** | **FID** | **Distance** |
| TediGAN (Xia et al, 2021) | 71.901 | 12.844 | 82.839 | 59.012 | 81.239 | 2073.721 | 86.959 | 19.371 | 98.201 | 287.839 |
| Gaussian Kernel | 74.189 | 13.353 | 86.346 | 58.745 | 77.523 | 1824.910 | 81.681 | 14.892 | 89.209 | 248.293 |

Table 7. Comparison Between Approximated Energy Function Guidance with the Best Interaction Model vs. Baseline (TediGAN)

From Table 7, our proposed approximated time-independent energy guidance with Gaussian kernel interaction modeling outperforms TediGAN in the segmentation map, landmark, and sketch conditions, achieving lower FID and conditional distances. However, it falls short in the text and facial ID conditions, where TediGAN attains lower FID and conditional distances. Despite this, our framework demonstrates competitive performance against state-of-the-art image generation models, highlighting its potential for integration into existing diffusion models to enhance their effectiveness.

## Qualitative Results

To demonstrate the effectiveness of our proposed model, we present qualitative results from the best-performing variant, which utilizes a Gaussian kernel for multi-conditional image generation. Specifically, we compare images generated with and without interaction modeling, followed by a comparison between our best-performing model and the baseline, TediGAN.

### **Interaction Modelling Comparison**

Visually, the images produced with interaction modeling exhibit improved visual fidelity and greater consistency in adhering to multiple constraints, demonstrating the benefits of capturing condition interdependencies.

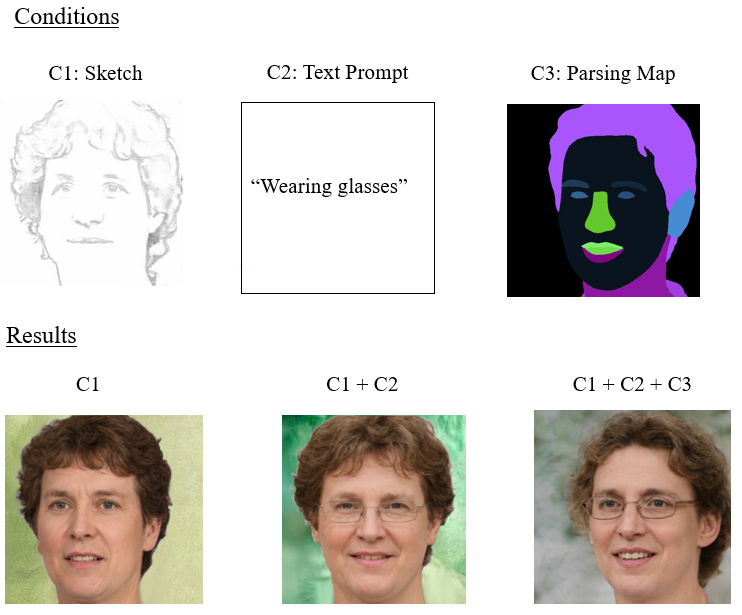


Figure 1. Illustration of Sequential Multi-Conditional Image Generation

*Illustration of Sequential Multi-Conditional Image Generation -*  
*Sketch + Text Prompt + Segmentation Map . This sequence demonstrates how the model integrates additional constraints while maintaining consistency in the output.*



Figure 2. Segmentation Map + Text Prompt Multi-Conditional Image Generation Result

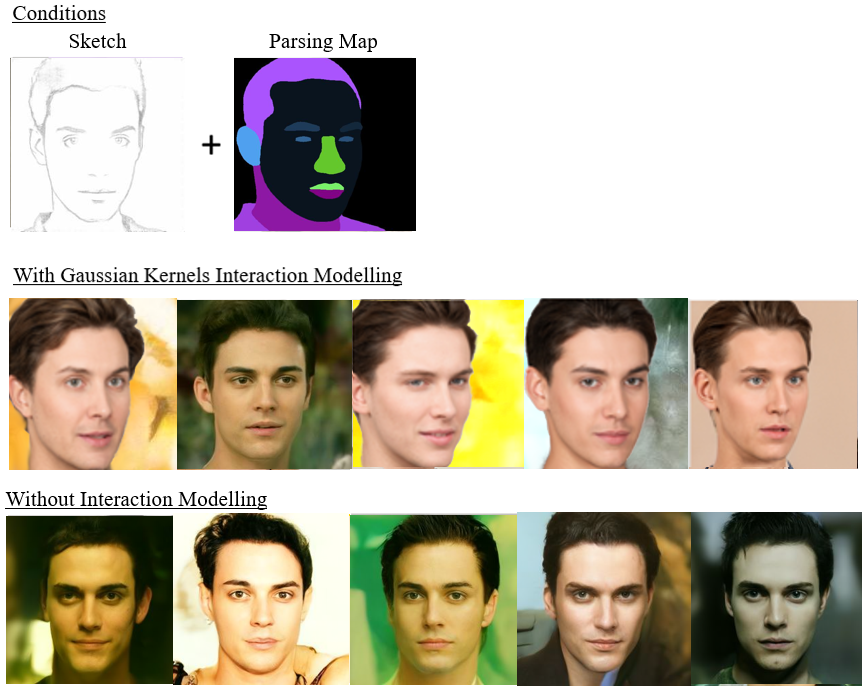


Figure 3. Sketch + Segmentation Map Multi-Conditional Image Generation Result

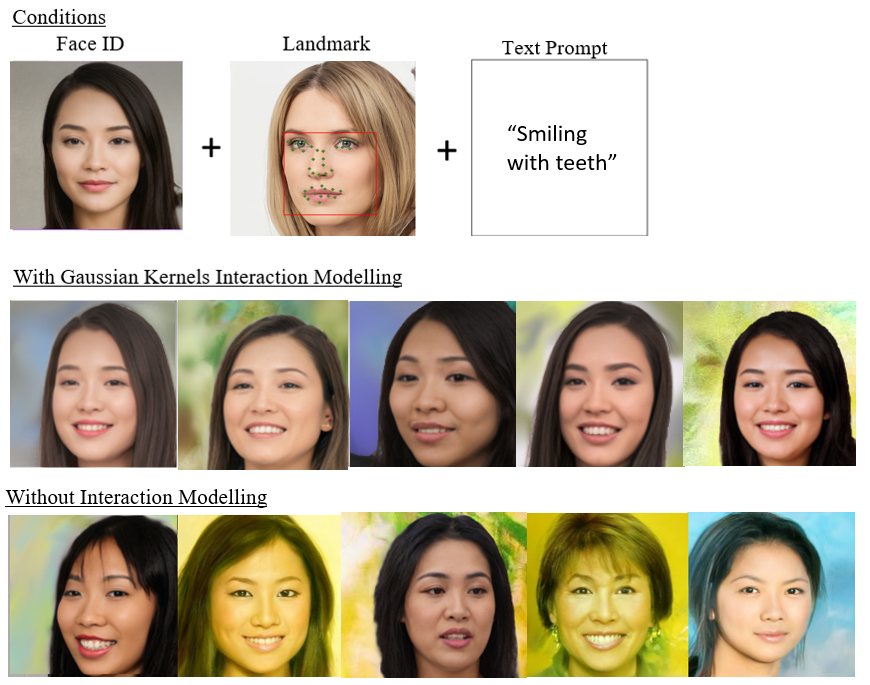


Figure 4. Face ID + Landmark + Text Prompt Multi-Conditional Image Generation Result

### **Baseline Comparison**

A comparison between images generated by our best-performing model and the baseline, TediGAN, reveals visually comparable results, with our model demonstrating slightly improved coherence and fidelity.

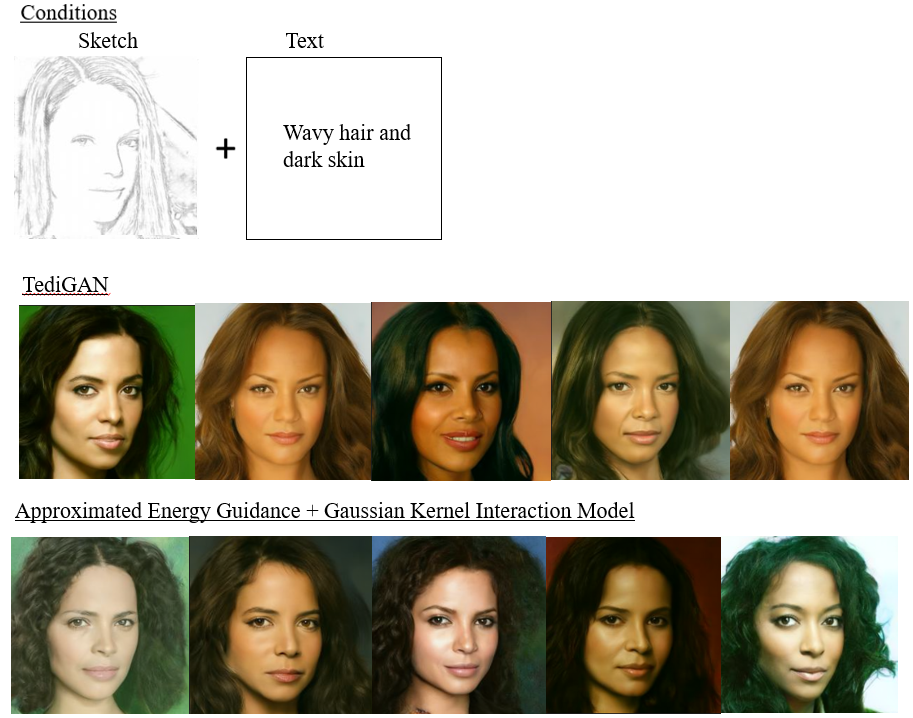


Figure 5. Sketch + Text Prompt Baseline Comparison

# Evaluation

## Models

Based on the observed quantitative results in Section 4.2.1, we observe that modeling interactions between potentially dependent conditions is crucial for achieving significantly better outcomes in multi-conditional image generation. However, the choice of interaction modeling approach is equally critical, as inappropriate modeling can lead to worse results than omitting interaction modeling altogether.

In this section, we attempt to evaluate the suitability of various approaches to interaction modeling between diverse conditions in the context of multi-conditional image generation, analyzing the potential reasons behind the observed results

|  |  |  |
| --- | --- | --- |
| **Model** | **Pros** | **Cons** |
| Simple Similarity Measures  (Euclidean Distance,    Cosine Similarity,  Pearson Correlation) | Simple and computationally efficient. | Fails to capture non-linear dependencies.  Poor performance in high-dimensional spaces. |
| Polynomial Functions | Captures non-linear interactions.  Flexible with adjustable degrees. | Sensitive to parameter tuning.  Risk of overfitting with higher-degree polynomials.  May exhibit abrupt changes for higher-degree terms, making optimization more challenging.  Also ensures localized influence, but may over-penalize small deviations due to the lack of squared distance in their formulation. |
| Sigmoid Functions | Captures non-linear interactions.  Effective for sigmoidal or threshold-like dependencies.  Useful in situations where the interactions between conditions exhibit saturating behavior. | Sensitive to parameter tuning.  Limited expressiveness for highly complex interactions.  Not symmetric and may introduce biases based on the direction of interactions. |
| Gaussian Kernels | Excellent for capturing complex, non-linear relationships without requiring explicit feature engineering (Hofmann et al, 2008).  Provides a smooth and differentiable measure of similarity, which is crucial for stable optimization and gradual alignment of conditions (Rasmussen & Williams, 2006).  Robust and adaptable across diverse data distributions.  Radial symmetry ensures that similarity depends only on the distance between conditions, not their orientation or scale, making it ideal for pairwise interactions. (Hofmann et al, 2008).  The exponential decay ensures that only closely related conditions strongly influence the output, minimizing the impact of irrelevant or conflicting conditions (Schölkopf & Smola, 2001). This focuses on the model’s attention on relevant, closely aligned conditions, preventing the overemphasis on distant or irrelevant conditions can lead to artifacts.  Flexible as the bandwidth parameter σ allows fine-grained control over how much dissimilarity is tolerated before the kernel value drops significantly (Rasmussen & Williams, 2006). Adjusting σ enables the model to adapt to the specific nature of each condition. | Computationally more expensive than simpler methods.  Requires careful bandwidth parameter selection. |

Table 8. Evaluation the pros and cons of each proposed interaction-modelling methods in the context of multi-conditional image generation tasks

## Experimental Limitations

The image generation process was highly time-intensive due to GPU resource constraints, particularly when exploring multiple hyperparameter combinations. As illustrated in Eq. (7), the number of possible configurations for interaction model-specific hyperparameters, along with the weighting factors for each condition, was extremely large. As a result, we acknowledge that our proposed grid search optimization method and the defined search space in Section 3.3.2.3 may not have been the most efficient approach for hyperparameter tuning. Consequently, the configurations presented in Section 4.1 may not represent the most optimal settings.

## Future Work

Aligned with the project's focus on enhancing training-free, multi-conditional image generation, the following suggestions for future work could further improve or address the limitations of the proposed framework.

### **Anisotropic Gaussian Kernels**

Given that the gaussian kernel shows the most promise, we aim to further enhance its performance. Currently, the standard gaussian kernel proposed in Eq. (13) can be written as:

where is assumed to be fixed across all dimension , reducing the hyperparameter search space to address GPU limitations and minimize computational complexity. However, with the fixed the standard gaussian kernel assumes equal influence in all directions, but some conditions (e.g., shape vs. color) influence different aspects of the image.

An improvement to this would be to introduce anisotropic kernels, which allow varying scaling across different feature directions with separate values for each dimension as follows:

where is predefined diagonal covariance matrix with varying values based on the perpetual importance of the different dimensions. This flexibility enables the kernel to capture the anisotropic nature of images, where different features (dimensions) may have different importance or units, requiring different levels of smoothness or variance along each axis (Berry & Sauer, 2014). Consequently, this modification can potentially add improvements to shape, spatial, and color constraints by ensuring conditions affect only relevant aspects of image features.

### **Dynamic and Adaptive Weight Adjustment**

Given that the weighting factors account for much of the hyperparameter search space, a potential improvement is to implement dynamic, adaptive weight adjustment to reduce reliance on exhaustive optimization. One approach involves computing the gradient norms of each condition’s energy term and applying a softmax function to obtain normalized weights that reflect real-time energy contributions. For interaction terms, the similarity between gradients can indicate whether conditions are synergistic or conflicting, allowing for corresponding adjustments in their weights. This process enables automatic and efficient scaling of both condition weights and interactions, thereby mitigating the need for labor-intensive hyperparameter tuning in high-dimensional spaces.

Our initial attempts to incorporate dynamic weight adjustment revealed several critical challenges. In early iterations, the gradient signals were extremely noisy, which compromised the stability of the softmax-based normalization of energy contributions. This instability hindered our ability to reliably adjust weights in real time. Moreover, the significant variability in the scales across different conditions made it difficult to establish robust thresholds for both individual conditions and their interaction terms. As a result, the dynamic adjustments did not consistently reflect the true energy dynamics, limiting the overall effectiveness of our approach. These findings underscore the inherent complexity of automating weight scaling in such systems and highlight the need for more refined strategies to mitigate noise and scale disparities.

### **Extension to other generative tasks**

Another promising avenue for future work is to extend interaction modeling in energy guided functions to other training-free multi-conditional generative tasks, such as video synthesis and 3D reconstruction.

In video synthesis, the guided energy function can be adapted to capture temporal interactions between consecutive frames. By modeling these interactions, the system can ensure smooth transitions, maintain motion coherence, and manage dynamic content effectively. This approach enables the generator to produce sequences where each frame is consistent with its neighbors, preserving the continuity and flow necessary for realistic video content.

Similarly, for 3D reconstruction, interaction modeling can be employed to capture spatial dependencies across multiple viewpoints. Here, the energy function integrates cues from different perspectives to enforce geometric and photometric consistency. This allows for the generation of volumetric representations that accurately reflect the underlying 3D structure of the scene without the need for additional training, paving the way for more flexible and efficient 3D content creation.

# Conclusion

In conclusion, this work successfully addresses the limitations of training-free multi-conditional image generation in handling multi-conditional dependencies. We propose integrating interaction modeling, particularly Gaussian kernels, with approximated time-independent energy guidance functions to enhance generation quality. The proposed approach not only improves the qualitative aspects of generated images but also delivers significant quantitative gains, demonstrating its effectiveness in overcoming conventional challenges in image synthesis.

The validity of these results is supported by notable improvements in key evaluation metrics, confirming that interaction-based energy modeling effectively refines generated outputs, especially when compared to the baseline, TediGAN, a state-of-the-art image generation network. However, certain limitations remain, such as potential trade-offs in computational efficiency and adaptability to highly complex constraints. Additionally, GPU limitations may lead to suboptimal hyperparameter configurations and hinder the testing of larger, more generalized generative models. Further research is needed to evaluate the method's generalizability across diverse datasets and real-world applications.

Future work could focus on optimizing computational efficiency, extending applicability to more complex generative models, and dynamically integrating additional constraints and weighting parameters. Exploring alternative kernel methods may further enhance adaptive scaling across different feature directions. This study lays a solid foundation for advancements in training-free multi-conditional image generation, opening avenues for more robust generative applications such as video synthesis and reconstruction.

# Project Schedule

This section provides an overview of the entirety of theproject schedule, detailing how the final-year project was systematically divided into phases, with each phase further broken down into specific objectives. The project was carefully planned across the designated timeline to ensure efficient progress from inception to completion.

In the original project schedule submitted on 31st August 2024, Table 9 in Appendix B, the research focus was to evaluate various image generation neural networks (GANs, VAEs, and Diffusion models) and reimplement the best-performing variation. However, at that stage, the project direction was still exploratory, and the final research topic had yet to be solidified.

Following discussions with project mentors, the initial scope was deemed too simplistic and lacking in originality to make a meaningful contribution to the field of image generation. After the research and literature review phase, the focus of this automated image generation project shifted to “Enhancing Training-Free Multi-Conditional Image Generation through Approximated Energy-Based Guidance and Interaction Modeling in Diffusion Model Denoising”. Thus, significant revisions were made to refine the research focus and enhance its impact.

The finalized project schedule, presented in Table 8 below, is the result of a comprehensive revision after the project's completion. It now accurately reflects the actual timeline, segmented into distinct periods with corresponding objectives, ensuring a detailed and precise alignment with the project lifecycle. This revised schedule provides a clear, systematic view of the research and development process, capturing the true progression of the project.

|  |  |
| --- | --- |
| **Time Period** | **Objective** |
| August  2024 | **Research and Literature Review Phase**  Research and Literature Review on:   1. Neural Networks for Image Generation    1. GANs    2. Variational Autoencoders    3. Diffusion Models       1. Unconditional Score-based diffusion models       2. Conditional Score-based diffusion models 2. Performance Metrics for image generation    1. FID, IS, KID, PPL 3. Learning Deep Learning Frameworks    1. TensorFlow    2. PyTorch |
| September 2024 | **Refinement of Research Scope Phase**   1. Identify potential research topics    1. Benchmarking different image generation neural networks (Initial).    2. Enhancing training-free image generation **(Chosen)**. 2. Research training-free image generation approaches    1. Classifier Guidance    2. Classifier-Free Guidance    3. Approximated Energy-Based Guidance    4. Latent Editing    5. Diffusion Inversion 3. Assessment of training-free image generation approaches to work on    1. Classifier Guidance       1. Requires a separate pre-trained classifier, which may not generalize well across conditions.    2. Classifier-Free Guidance       1. Less effective when multiple constraints need to be enforced simultaneously.    3. Approximated Energy-Based Guidance (Chosen)    4. Latent Editing       1. Requires additional optimization steps at inference, making it slow.    5. Diffusion Inversion       1. Can be computationally expensive and may introduce artifacts. 4. Research on approximated energy guidance.    1. Time independent approximations for energy guided functions with pre-trained networks. 5. Research on the pain points of training-free image generation.    1. To increase the accuracy of the Approximated Multi-Conditional Energy Guidance. 6. Finalize research topic    1. “Enhancing Training-Free Multi-Conditional Image Generation through Approximated Energy-Based Guidance and Interaction Modeling in Diffusion Model Denoising.” 7. Set up the project repository and define an initial experimental plan. |
| October 2024 | **Preliminary Experiments and Search Phase**   1. Experimenting with pre-trained networks for the primary image generation model.    1. Due to the limitations of my personal GPU, the denoising process for larger models like StableDiffusion and ControlNet either took too long to complete or resulted in memory exhaustion.    2. Chose to use a smaller, pre-trained face generating network 2. Evaluate different pre-trained networks for condition-specific guidance (condition-specific extracting models).    1. Dependent on the type of conditions to be experimented and the scope of images able to be generated by the primary image generation model.    2. Chosen conditions: Face Identity, Landmark, Segmentation Map, Sketch, Text. 3. Experiment with approximated time-independent single-conditional energy-based guidance. |
| November 2024 | **Multi-Conditional Energy Function Approximation & Initial Model Integration Phase**   1. Develop and test methods for approximating time-independent approximated multi-conditional energy-based guidance.    1. Initial model - extension of single-conditional energy-based guidance via weighted sum. 2. Integration of approximated multi-conditional energy-based guidance into denoising function. 3. Conduct initial qualitative evaluations to check the impact of individual condition constraints. 4. Begin logging results and debugging performance issues    1. To find the optimal ratio of weighting factors of condition-specific distance guidance functions for different condition combinations. |
| December 2024 | **Interaction Modeling for Multi-Condition Dependencies Phase**   1. Research and implement different training-free interaction modeling techniques to capture dependencies between multiple conditions.    1. Weighted Aggregation    2. Euclidean Distance    3. Pearson Correlation    4. Cosine Similarity    5. Polynomial Kernels    6. Sigmoid Kernels    7. Gaussian Kernels 2. Integration of interaction modelling into multi-conditional energy-based guidance function. 3. Experiment with how different conditions influence each other and refine the weighting factors of the interaction terms. 4. Ablation studies will be conducted to compare different interaction models. 5. Perform initial qualitative evaluations to determine the reasonable search space for the hyperparameters of each interaction modeling technique, before conducting hyperparameter optimization through grid search. |
| January 2025 | **Model Optimization & Performance Evaluation Phase**   1. Optimize all combinations of hyperparameters with grid search. 2. Conduct extensive evaluations using quantitative metrics (FID, constraint-specific distances). 3. Compare results with existing training-free image generation baselines. 4. Compare the different interaction modelling techniques to find the best. 5. Identify failure cases and areas for improvement. 6. Identify potential extension areas for the project. 7. Start drafting key findings and insights. 8. Interim Report Writing. 9. Submission of Interim Report – 27th January 2025 |
| February 2025 | **Documentation & Final Experiments Phase**   1. Code refactoring (Submission preparation). 2. Final Deployment. 3. Conduct final experiments to confirm findings. 4. Prepare figures, graphs, and tables for results visualization. 5. Final Report Writing    1. Document methodology, experimental design, results, and key conclusions.    2. Include all required deliverables mentioned in the report guide. |
| March 2025 | **Final Report Submission**   1. Thorough checks and ensure all formats are correct 2. Submission of Final Report – 24th March 2025 |
| April  2025 | **Amended Final Report**   1. Final amendments to the final report for submission. 2. Submission of Amended Final Report – 18th April 2025 |
| May 2025 | **Oral Presentation**   1. Creation of slides for oral presentation. 2. Creation of script for oral presentation. 3. Rehearsals for oral presentation 4. Oral Presentation – 9th May 2025 |

Table 9. Finalized Project Schedule

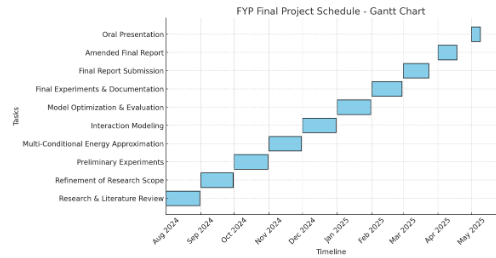


Figure 6. Gantt Chart of Finalized Project Schedule

# Appendices

## 8.1 Appendix A – Code Snippets

This appendix provides key code snippets that were instrumental in the implementation of the project. The following sections contain excerpts from core scripts of the proposed framework, including the time independent approximated time energy function and the various interaction models.

### **8.1.1 Approximated clean image from an intermediate noisy image**

This code snippet depicts the approximation formula to get a clean image corresponding to an intermediate noisy image for each time step as in Equation (5). This is crucial for the calculation of distance between the given condition and the intermediate image as the pre-trained condition networks only work well on cleaned images.

A screen shot of a computer program

AI-generated content may be incorrect.

Figure 7. Approximated clean image from intermediate noisy image

### **8.1.2 Multi-Conditional Energy Guidance with Interaction Modelling**

This code snippet depicts the algorithm for calculating the multi-conditional energy guidance as in Equation (7).

A screen shot of a computer code

AI-generated content may be incorrect.

Figure 8. Algorithm for multi-conditional energy guidance with interaction modelling

### **8.1.3 Condition-Specific Classes**

The following code snippets (Figures 9–13) illustrate the implementation of condition-specific classes: Text, Facial Landmark, Facial Segmentation Map, Face ID, and Sketch. Each class is initialized with its respective pre-trained model and a specified instance of the condition as an image. Additionally, these classes include functions for interaction models and conditional distance calculations. Approximated clean images can be passed into these functions to compute the required information.

A computer screen shot of a program code

AI-generated content may be incorrect.

Figure 9. Text condition class

A screen shot of a computer program

AI-generated content may be incorrect.

Figure 10. Facial landmark condition class

A screen shot of a computer code

AI-generated content may be incorrect.

Figure 11. Facial segmentation map condition class

A screen shot of a computer code

AI-generated content may be incorrect.

Figure 12. Facial ID condition class

A computer screen with text and images

AI-generated content may be incorrect.

Figure 13. Sketch condition class

### **8.1.4 Final Sampling Formula**

This code snippet depicts the final sampling formula with energy guidance as in Equation (3).

A screen shot of a computer code

AI-generated content may be incorrect.

Figure 14. Final sampling formula

### **8.1.5 Interaction Models**

The following code snippets (Figures 15–20) illustrate the implementation of various interaction models and similarity functions: Euclidean distance, cosine similarity, Pearson correlation coefficient, polynomial kernel, and Gaussian kernel. These figures specifically depict the interaction model class functions within the facial segmentation map class. While similar functions exist for other conditional classes—such as text, landmarks, face ID, and sketches—the calculations vary slightly due to differences in the output characteristics of each pre-trained model. However, the underlying logic remains consistent across all classes.

A computer screen shot of text

AI-generated content may be incorrect.

Figure 15. Euclidean distance interaction model

A computer screen shot of a program code

AI-generated content may be incorrect.

Figure 16. Cosine similarity interaction model

A screen shot of a computer code

AI-generated content may be incorrect.

Figure 17. Pearson correlation interaction model

A screen shot of a computer program

AI-generated content may be incorrect.

Figure 18. Polynomial function interaction model

A computer screen with text and images

AI-generated content may be incorrect.

Figure 19. Sigmoid function interaction model

A computer screen shot of text

AI-generated content may be incorrect.

Figure 20. Gaussian kernel interaction model

## 8.2 Appendix B – Old Project Schedule

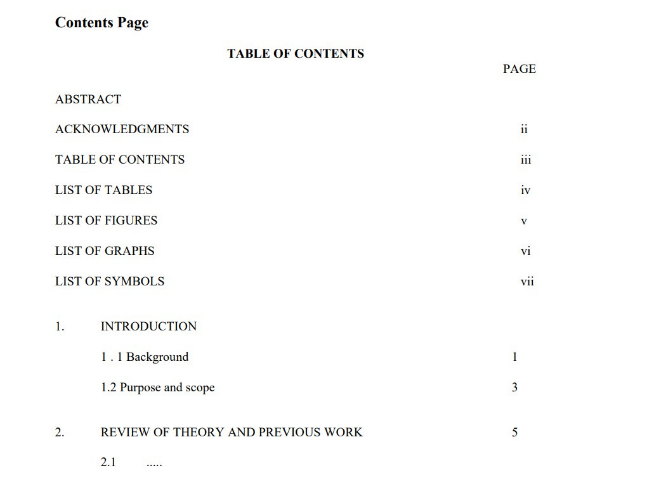
|  |  |
| --- | --- |
| **Time Period** | **Objectives** |
| August  2024 | **Background Phase**   1. Literature Review    1. Deep Learning Models (GANs, VAEs, Diffusion, ...etc.)    2. Performance Metrics (FID, IS, KID, PPL, ...etc.) 2. Learning Deep Learning Frameworks    1. TensorFlow    2. PyTorch 3. Understanding Image Processing Libraries    1. OpenCV    2. Pillow |
| September 2024 | **Background Phase**   1. Literature Review    1. Deep Learning Models (GANs, VAEs, Diffusion, ...etc.)    2. Performance Metrics (FID, IS, KID, PPL, ...etc.) 2. Learning Deep Learning Frameworks    1. TensorFlow    2. PyTorch 3. Understanding Image Processing Libraries    1. OpenCV    2. Pillow |
| October 2024 | **Reimplementation Phase**   * Build automated image generation models for the learned deep learning models   **Benchmarking Phase**   * Assess every model built with a set of appropriate assessment metrics   **Analysis Phase**   * Performance and limitation analysis of models * Comparative analysis |
| November 2024 | **Reimplementation Phase**   * ­Build automated image generation models for the learned deep learning models   **Benchmarking Phase**   * Assess every model built with a set of appropriate assessment metrics   **Analysis Phase**   * Performance and limitation analysis of models * Comparative analysis |
| December 2024 | **Reimplementation Phase**   * ­Build automated image generation models for the learned deep learning models   **Benchmarking Phase**   * Assess every model built with a set of appropriate assessment metrics   **Analysis Phase**   * Performance and limitation analysis of models * Comparative analysis |
| January 2025 | **Innovation Phase**   * Literature Review of new scarcely-researched for (potential) implementation and analysis * (Potential) modifications to existing methods to address identified limitations   **Final Development Phase**   * Finalize best performing automated image generation model. * Create a functioning interface for the best performing automated image generation model * Deployment   Submission of Interim Report – 27th January 2025 |
| February 2025 | **Project Finalization Phase**   * Code refactoring (Submission preparation) * Final Deployment * Report Finalization * Work on Presentation |
| March 2025 | **Presentation Preparation and rehearsal**  Submission of Final Report – 24th March 2025 |
| April  2025 | **Presentation Preparation and rehearsal**  Submission of Amended Final Report – 18th April 2025 |
| May 2025 | **Oral Presentation**  Oral Presentation – 9, 12-14th May (To be Confirmed) |

Table 10. Old Project Schedule, 31 August 2024

# References

1. Berry, T., and Sauer, T., 'Local kernels and the geometric structure of data', *arXiv*, https://doi.org/10.48550/arXiv.1407.1426, 2014.
2. Chen, C., 'PyTorch Face Landmark: A fast and accurate facial landmark detector', *github*, <https://github.com/cunjian/pytorch_face_landmark>, 2021.
3. Chung, H., Sim, B., and Ye, J. C., 'Come-closer-diffuse-faster: Accelerating conditional diffusion models for inverse problems through stochastic contraction', *arXiv*, https://doi.org/10.48550/arXiv.2112.05146, 2022.
4. Deng, J., Guo, J., Xue, N., and Zafeiriou, S., 'Arcface: Additive angular margin loss for deep face recognition', *arXiv*, https://doi.org/10.48550/arXiv.1801.07698, 2019.
5. Dhariwal, P., and Nichol, A., 'Diffusion models beat gans on image synthesis', *arXiv*, https://doi.org/10.48550/arXiv.2105.05233, 2021.
6. Hofmann, T., Schölkopf, B., and Smola, A. J., 'Kernel methods in machine learning', *arXiv*, https://doi.org/10.48550/arXiv.math/0701907, 2008.
7. LeCun, Y., Chopra, S., Hadsell, R., Ranzato, M., and Huang, F., 'A tutorial on energy-based learning', *MIT Press*, http://yann.lecun.com, 2006.
8. Lin, H.-T., and Lin, C.-J., 'A study on sigmoid kernels for SVM and the training of non-PSD kernels by SMO-type methods', 2003
9. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.-Y., and Ermon, S., 'SDEdit: Guided image synthesis and editing with stochastic differential equations', *arXiv*, https://doi.org/10.48550/arXiv.2108.01073, 2022.
10. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al., 'Learning transferable visual models from natural language supervision', *arXiv*, https://doi.org/10.48550/arXiv.2103.00020, 2021.
11. Rasmussen, C. E., and Williams, C. K. I., 'Gaussian processes for machine learning', *MIT Press*, https://doi.org/10.7551/mitpress/3206.001.0001, 2006.
12. Schölkopf, B., & Smola, A. J., ‘Learning with kernels: Support Vector Machines, regularization, optimization, and beyond.’, *MIT Press*, <https://doi.org/10.7551/mitpress/4175.001.0001>, 2001
13. Shawe-Taylor, J., and Cristianini, N., 'Kernel methods for pattern analysis', *Cambridge University Press*, <https://doi.org/10.1017/CBO9780511809682>, 2004.
14. Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B., 'Score-based generative modelling through stochastic differential equations', *arXiv*, https://doi.org/10.48550/arXiv.2011.13456, 2021.
15. Xia, W., Yang, Y., Xue, J.-H., and Wu, B., 'TediGAN: Text-guided diverse face image generation and manipulation', *arXiv*, https://doi.org/10.48550/arXiv.2012.03308, 2021.
16. Xiang, X., Liu, D., Yang, X., Zhu, Y., Shen, X., and Allebach, J. P, 'Adversarial open domain adaptation for sketch-to-photo synthesis', *arXiv*, https://doi.org/10.48550/arXiv.2104.05703, 2022.
17. Yang, Y., Wang, W., Peng, L., Song, C., Chen, Y., Li, H., Yang, X., Lu, Q., Cai, D., Wu, B., and Liu, W., 'LoRA-Composer: Leveraging low-rank adaptation for multi-concept customization in training-free diffusion models', *arXiv*, https://doi.org/10.48550/arXiv.2403.11627, 2024.
18. Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., and Sang, N., 'Bisenet: Bilateral segmentation network for real-time semantic segmentation', *arXiv*, https://doi.org/10.48550/arXiv.1808.00897, 2018.
19. Yu, J., Wang, Y., Zhao, C., Ghanem, B., and Zhang, J., 'FreeDoM: Training-free energy-guided conditional diffusion model', *arXiv*, https://doi.org/10.48550/arXiv.2303.09833, 2023.

# TODOS (Post-Interim Report)

1. **~~experiment section, describing the experiment section, give different results on different hyperparameter configuration~~**
   1. **~~FILL Experimental Setup~~**
      1. **~~100 images generated across different combinations of multi c~~**
      2. **~~Talk about what kind of c? eg what text~~**
   2. **~~FILL Hyperparameter Optimization~~**
2. **~~multi control vs single control (qualitative)progressive results, showing multi c better than single c with each condition added starting from single c,~~** 
   * ~~eg: sc: face id; mutlic: faceid, segmentation map~~
3. ~~To test out potentially new approaches to model interaction terms~~
   1. ~~Gaussian Kernel extensions: Anisotropic Gaussian Kernels, Multi-scale kernels~~
4. ~~To test out potentially new approaches to model interaction terms~~
5. ~~Fill Conclusion~~
6. ~~Modify results and evaluation sections~~
7. ~~Tidy model eval section table 5.1~~
8. ~~Ensure all citations are proper and present~~
   1. ~~And the names are all correct (search it up)~~
   2. ~~Format~~
9. ~~Research more on section 5.3~~
10. ~~Tidy up intro and abstract~~
11. ~~Tidy up references (SHOULD FOLLOW GUIDE, NUMERING)~~
12. **~~ENSURE ALL THE ITEMS IN THE GUIDE FOR THE REPORT ARE THERE~~**
    1. **~~In body, add~~**
       1. **~~PROJECT SCHEDULE~~**
          1. ~~The project timeline should be in chart format(? Check guide)~~
       2. **~~DESIGN DOCUMENTATION~~**
    2. **~~Last done - conclusion~~**
13. ~~ADD CLEAR THEME TO THE REPORT, overarching title: {approximated time independent energy function guidance formula with interaction modelling} for training free multi conditional image generation~~
14. ~~Add old project schedule and explain the deviation, can include gant charts for both~~
15. ~~Fill Appendix~~
16. ~~ADD BASELINE – TEDIGAN (QUANTITATIVE RESULTS)~~
17. ~~ADD BASELINE – TEDIGAN (QUALITITATIVE RESULTS)~~
18. READ EVERYTHING, SPELL CHECK (EDIT WEIRD FLOWS + SPELL CHECK)
19. final structuring of the report
20. Standardize all formatting to APA (spacings, fontsize, in text citation, references, tables, figures, labelling fomatings…etc)
21. Tidy table of content (page #, spacings, bolding…etc)
22. Tidy table of tables + update (page #, updated figures, 1.5 spacing between lines)
23. Tidy table of figures + update(page #, updated figures, 1.5 spacing between lines)
24. Update table of content
25. Possibly more qualitative generation results
26. Code optimization