**NANYANG TECHNOLOGICAL UNIVERSITY**

**AUTOMATED IMAGE GENERATION**

Hill Seah Wen Qi

College of Computing and Data Science 2024

**NANYANG TECHNOLOGICAL UNIVERSITY**

**AUTOMATED IMAGE GENERATION**

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

# Abstract

Multi-conditional image generation aims to synthesize images that satisfy diverse conditions, such as textual descriptions, segmentation masks, and landmark constraints. While training-free approaches to image generation often excel in single-condition scenarios by leveraging off-the-shelf, open-source pre-trained networks to estimate the distance between an intermediate image and the condition, they struggle with multi-conditional tasks. This limitation arises from the inability of training-free methods to effectively handle interactions and dependencies between multiple conditions. To address this challenge, this project investigates strategies for effective multi-conditional image generation. It introduces a novel framework that integrates a naive weighted sums of distance functions with product of Gaussian kernels interaction models that capture complex, non-linear interdependencies between conditions, to approximate the energy function and guide the iterative denoising process. Experimental results demonstrate that the proposed method outperforms existing techniques in generating coherent, condition-consistent images across a variety of conditions, showcasing its effectiveness in addressing the challenges of multi-conditional image generation.

# 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 gratefully to my project supervisor Professor Lu Shijian and my student mentor Xu Muyu for their guidance and assistance to this endeavor, making this journey a memorable and fulfilling experience.

Contents

[Abstract 3](#_Toc188825395)

[Acknowledgements 3](#_Toc188825396)

[1. Introduction 5](#_Toc188825397)

[1.1 Problem Statement 5](#_Toc188825398)

[1.2 Objective 5](#_Toc188825399)

[1.3 Scope 5](#_Toc188825400)

[2. Related Work 6](#_Toc188825401)

[2.1 Conditional Score Based Diffusion Models 6](#_Toc188825402)

[2.2 Energy Diffusion Guidance 6](#_Toc188825403)

[2.2 Time-Independent Distance Functions Approximating Time-Dependent Energy Guidance 7](#_Toc188825404)

[3. Methodology 8](#_Toc188825405)

[3.1 Multi-Conditional Training-Free Image Generation Modelling 8](#_Toc188825406)

[3.1.1 Primitive Multi-Conditional Energy Guidance Model – No Interaction Modelling 8](#_Toc188825407)

[3.1.2 Improved Multi-Conditional Energy Guidance Model – Interaction Modelling 8](#_Toc188825408)

[3.2 Interaction Modelling 9](#_Toc188825409)

[3.2.1 Simple Similarity Measures 9](#_Toc188825410)

[3.2.2 Polynomial Functions 9](#_Toc188825411)

[3.2.3 Sigmoid Functions 9](#_Toc188825412)

[3.2.4 Gaussian Kernels 10](#_Toc188825413)

[3.3 Configuration 11](#_Toc188825414)

[3.3.1 Pre-trained Models 11](#_Toc188825415)

[3.3.2 Hyperparameters 12](#_Toc188825416)

[4. Results 13](#_Toc188825417)

[4.1 Quantitative Results 13](#_Toc188825418)

[4.2 Qualitative Results 14](#_Toc188825419)

[Bibliography 14](#_Toc188825420)

# Introduction

## 1.1 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. Training-free techniques 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, 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 training-free techniques to multi-conditional image generation introduces significant challenges. Unlike single-condition tasks, where alignment to a single constraint is sufficient, multi-conditional generation requires simultaneous optimization across diverse constraints, which may conflict or compete with one another. For instance, aligning an image to textual descriptions while maintaining geometrical consistencies and segmentation adherences often leads to trade-offs or failures in satisfying one or more conditions. Furthermore, the absence of explicit training processes makes it difficult for training-free methods to effectively balance these constraints or resolve potential conflicts during image generation. The inefficacies in interaction modelling between multiple conditions leads to several issues, including the treatment of conflicting conditions in isolation, a loss of coherence in generated outputs, and limited adaptability to complex, real-world multimodal scenarios

Hence, this necessitates the development of new strategies to manage interaction effects among conditions and to ensure robust, high-quality image outputs that satisfy all specified requirements.

## 1.2 Objective

The objective of this project is to address the limitations of training-free multi-conditional image generation by developing effective interaction modelling techniques that can handle multiple, potentially conflicting conditions. The goal is to ensure robust, high-quality image outputs that simultaneously satisfy all specified requirements, enabling better adaptability to complex, real-world multimodal scenarios.

## 1.3 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, along with the conditions — textual description, facial landmarks, facial ID, facial parsing map and facial sketch.

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 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.

## 2.2 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 both conditionality and energy guidance is as follows:  
where represents the learning rate of the energy guidance term.

## 2.2 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. 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).

# Methodology

## Multi-Conditional Training-Free Image Generation Modelling

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

A primitive and straightforward approach in modelling a multi-conditional energy guidance formula would be to simply extend the singular-conditional energy guidance formula as a weighted sum of the different distance functions 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 weighing 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 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 weighing factor of the interactions between .

## Interaction Modelling

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. 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).

The reason why gaussian kernels are particularly well-suited for modelling interactions between image conditions in multi-conditional tasks because they satisfy key requirements for such applications:

1. Smooth Similarity Modelling — Gaussian kernels provide a smooth and differentiable measure of similarity, which is crucial for stable optimization and gradual alignment of conditions (Rasmussen & Williams, 2006).
2. Non-Linear Dependency Capture — Gaussian kernels naturally handle non-linear relationships between conditions without requiring explicit feature engineering (Hofmann et al, 2008).
3. Localized Influence — The exponential decay ensures that only closely related conditions strongly influence the output, minimizing the impact of irrelevant or conflicting conditions (Schölkopf & Smola, 2002).
4. Flexibility via Bandwidth Parameter (σ) — The adjustable σ allows fine-tuning of sensitivity to condition alignment, enabling the kernel to adapt to diverse condition types and scales (Rasmussen & Williams, 2006).
5. Radial Symmetry and Interpretability — Gaussian kernels treat condition pairs symmetrically and provide interpretable similarity scores, helping maintain fairness and transparency in interaction modelling (Hofmann et al, 2008).

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

where .

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 as:

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, .

|  |  |  |
| --- | --- | --- |
| **Model** | **Purpose** |  |
| Open-source Face Parsing Network (Yu et al, 2018) | Supports the parsing map condition.  Generates a facial parsing 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 parsing 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*

### Hyperparameters

After evaluating visual and quantitative results from experimentations through grid search, the following configuration of hyperparameter values were found to be as follows:

|  |  |  |
| --- | --- | --- |
| **Hyperparameter** | **Value** | **Description** |
|  | = 1000 : 1 : 1000 : 1000 : 10 | Weighing factor for |
|  |  | Weighing 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 3: Configuration of hyperparameter values*

# Results

## Quantitative Results

To compare the multi-conditional energy guidance models, with and without interaction modelling, and the best interaction modelling approach, we compare the average condition-specific distance and the average FID score across 100 generated images.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Text** | | **Facial ID** | | **Parsing Map** | |
| **Interaction Modelling** | **FID** | **Distance** | **FID** | **Distance** | **FID** | **Distance** |
| None (Primitive Model) | 112.721 | 21.812 | 158.014 | 84.122 | 129.301 | 2331.756 |
| Euclidean Distance | 162.891 | 34.921 | 187.928 | 132.271 | 146.027 | 2737.189 |
| Cosine Similarity | 130.475 | 24.572 | 144.371 | 73.490 | 118.163 | 2147.203 |
| Pearson Correlation | 148.907 | 29.230 | 148.039 | 83.211 | 131.394 | 2721.823 |
| Polynomial | 72.391 | 15.988 | 64.920 | 46.519 | 89.027 | 2271.201 |
| Sigmoid | 65.192 | 16.695 | 79.102 | 52.355 | 71.273 | 2174.521 |
| Gaussian Kernel | 55.189 | 12.353 | 65.346 | 38.745 | 57.523 | 1724.910 |

*Table 4: Results of the different proposed multi-conditional energy guidance models*

As shown in Table 4, 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 exhibit higher average FID and distance values compared to models without any interaction modelling.

Notably, Table X 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.

## Qualitative Results

Conditions

Parsing Map Text Prompt



With Gaussian Kernels Interaction Modelling



Without Interaction Modelling



# Bibliography

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. (2021) Learning

transferable visual models from natural language supervision.

Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. (2018) Bisenet: Bilateral segmentation network for real-time semantic segmentation.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. (2022) SDEdit: Guided image synthesis and editing with stochastic differential equations.

Cunjian Chen. (2021) PyTorch Face Landmark: A fast and accurate facial landmark detector

Hofmann, T., Schölkopf, B., & Smola, A. J. (2008). Kernel Methods in Machine Learning. The Annals of Statistics.

Hyungjin Chung, Byeongsu Sim, and Jong Chul Ye. (2022) Come-closer-diffuse-faster: Accelerating conditional diffusion models for inverse problems through stochastic contraction.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. (2019) Arcface: Additive angular margin loss for deep face recognition.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. (2021) In Advances in Neural Information Processing Systems.

Rasmussen, C. E., & Williams, C. K. I. (2006). Gaussian Processes for Machine Learning. MIT Press.

Schölkopf, B., & Smola, A. J. (2002). Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.

Shawe-Taylor, J., & Cristianini, N. (2004). Kernel Methods for Pattern Analysis. Cambridge University Press.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. (2021) Score-based generative modelling through stochastic differential equations.

Xiaoyu Xiang, Ding Liu, Xiao Yang, Yiheng Zhu, Xiaohui Shen, and Jan P Allebach. (2022) Adversarial open domain adaptation for sketch-to-photo synthesis.

Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and Fujie Huang. (2006) A tutorial on energy-based learning.

Yu, J., Wang, Y., Zhao, C., Ghanem, B., Zhang, J. (2023) FreeDoM: Training-Free Energy-Guided Conditional Diffusion Model.

**PS**

< CRUX: training-free image generation often works well for single conditions but struggle with multi conditions due to the inability for training-free approaches to handle interactions between multi conditions well >

< Training-free over training required methods>

<Training-free cannot handle multi conditions as well (FIND OUT MORE) + citations>

<Insert how training-free approaches struggle with mutli conditions + reasons + citations>

TODO:

* Seriously read the report + research done previously
* Identify the crux of the report
* See how to organize body \*see research portion headers (decide the main flow + idea you want to present)
* TO ADD PICS AND RESULTS
* Eval portion, quant metrics
* Conclusion, everything absically

**Table of Contents:**

1. **Introduction**
   1. **Problem Statement (Focus on generally, single cond heavy focused)**
   2. **Objective (want to create multi cond + effective multi cond)**
2. **Body**
   1. **Preliminaries (all the est formulas)**
      1. **Talk abt training free over training**
   2. **Related Work**
   3. **Models (Naïve Multi Cond denoising formula, m1, m2, m3, …)**
   4. **Evaluation**
      1. < in the end, so far, gaussain kernals are the best approach>
3. **Conclusion**
   1. **Limitations**
   2. **Recommendations**
4. **References**
   1. **Biblo**
   2. **Appendix**
   3. **Glossary**