**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

15th 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.

Contents

[Abstract 3](#_Toc188913619)

[Acknowledgements 3](#_Toc188913620)

[1. Introduction 5](#_Toc188913621)

[1.1 Problem Statement 5](#_Toc188913622)

[1.2 Objective 5](#_Toc188913623)

[1.3 Scope 5](#_Toc188913624)

[2. Related Work 6](#_Toc188913625)

[2.1 Conditional Score Based Diffusion Models 6](#_Toc188913626)

[2.2 Energy Diffusion Guidance 6](#_Toc188913627)

[2.3 Time-Independent Distance Functions Approximating Time-Dependent Energy Guidance 6](#_Toc188913628)

[3. Methodology 7](#_Toc188913629)

[3.1 Multi-Conditional Training-Free Image Generation Modelling 7](#_Toc188913630)

[3.1.1 Primitive Multi-Conditional Energy Guidance Model – No Interaction Modelling 7](#_Toc188913631)

[3.1.2 Improved Multi-Conditional Energy Guidance Model – Interaction Modelling 8](#_Toc188913632)

[3.2 Interaction Modelling 8](#_Toc188913633)

[3.2.1 Simple Similarity Measures 8](#_Toc188913634)

[3.2.2 Polynomial Functions 8](#_Toc188913635)

[3.2.3 Sigmoid Functions 8](#_Toc188913636)

[3.2.4 Gaussian Kernels 9](#_Toc188913637)

[3.3 Configuration 10](#_Toc188913638)

[3.3.1 Pre-trained Models 10](#_Toc188913639)

[3.3.2 Hyperparameters 11](#_Toc188913640)

[4. Results 11](#_Toc188913641)

[4.1 Quantitative Results 11](#_Toc188913642)

[4.2 Qualitative Results 12](#_Toc188913643)

[5. Evaluation 13](#_Toc188913644)

[5.1 Models 13](#_Toc188913645)

[5.2 Limitations 14](#_Toc188913646)

[5.3 Future Work 15](#_Toc188913647)

[6. Conclusion 15](#_Toc188913648)

[7. Appendices 15](#_Toc188913649)

[8. References 15](#_Toc188913650)

[9. TODOS (Post-Interim Report) 16](#_Toc188913651)

# 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, 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.

## 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.

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.

## 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.

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.

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

## 2.3 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 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 (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,

|  |  |  |
| --- | --- | --- |
| **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*

* + 1. **Experiments**

### **3.3.2.1 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).

### **3.3.2.2 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.

### **3.3.2.3 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

## 4.1 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 | 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.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Text** | | **Facial ID** | | **Parsing Map** | | **Landmark** | | **Sketch** | |
| **Interaction Modelling** | **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 different proposed multi-conditional energy guidance 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.

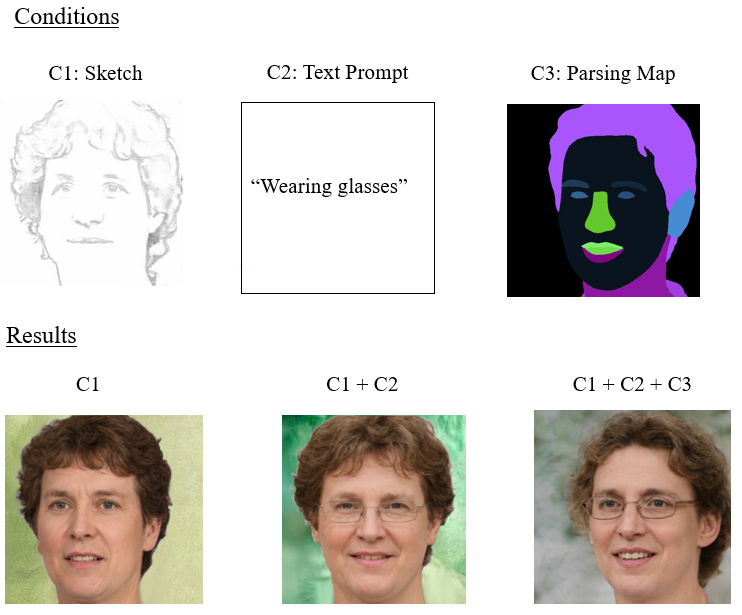
Conversely, the table also highlights that inappropriate or insufficient interaction modelling can lead to worse results. For example, models using Euclidean distance, cosine similarity, or Pearson correlation coefficient for interaction modelling generally exhibit higher average FID and distance values 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.

## Qualitative Results

To illustrate the effectiveness of our proposed model, we present some of the qualitative results from the best-performing model using a Gaussian kernel for multi-conditional image generation. Specifically, we compare images generated with and without interaction modeling.

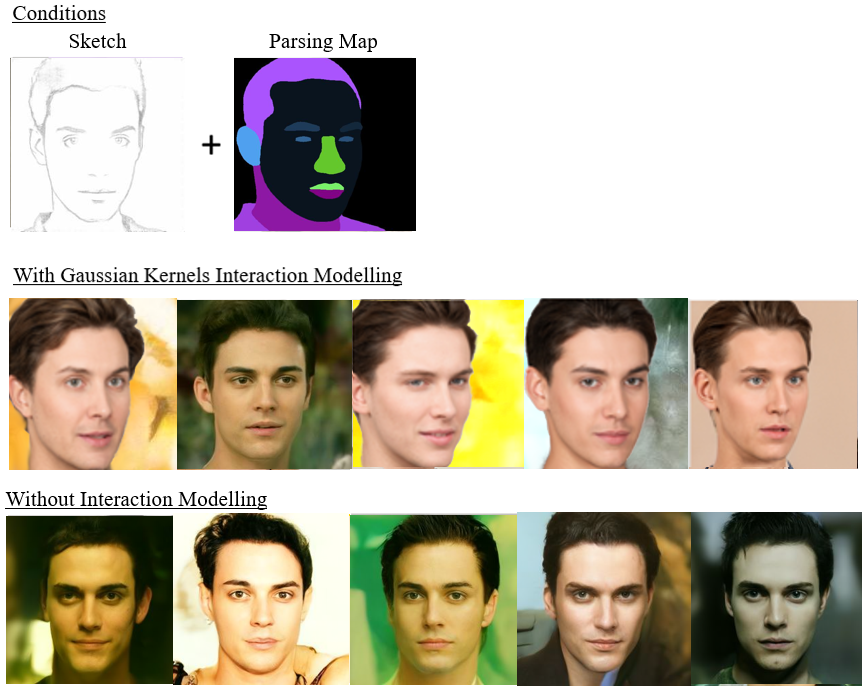
Visually, the images produced with interaction modeling exhibit improved visual fidelity and greater consistency in adhering to multiple constraints, demonstrating the benefits of capturing condition interdependencies.



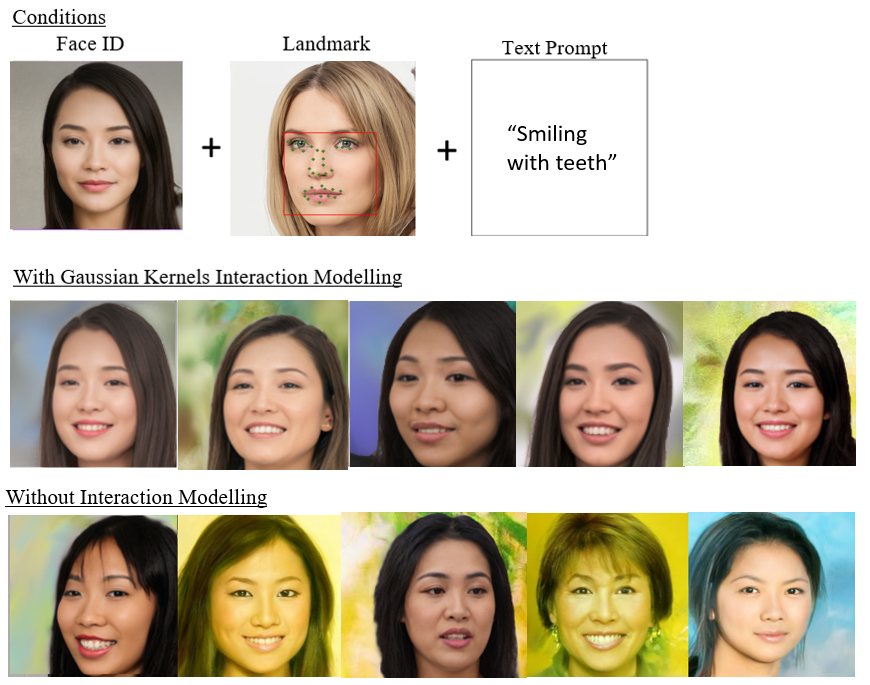
*Figure 1: Illustration of Sequential Multi-Conditional Image Generation -   
Sketch + Text Prompt + Parsing Map.* *This sequence demonstrates how the model integrates additional constraints while maintaining consistency in the output.*



*Figure 2: Parsing Map + Text Prompt Multi-Conditional Image Generation Result*



*Figure 3: Sketch + Parsing Map Multi-Conditional Image Generation Result*



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

# Evaluation

## 5.1 Models

Based on the observed quantitative results in Section 4.2, 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, 2002). 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 7: 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 several inherent limitations of training‐free multi‐conditional image generation. We propose integrating interaction modeling, particularly gaussian kernels, with approximated time‐independent energy function for this purpose. The proposed approach has demonstrated that it has not only enhances the qualitative aspects of generated images but also delivers significant quantitative improvements. These results underscore the potential of interaction-based energy modeling to overcome conventional barriers in image synthesis, laying a solid foundation for further exploration and optimization in the field.

# Appendices

* To insert code snippets (Currently Optimizing Code)

# References

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.

Berry, T., & Sauer, T. (2014). Local kernels and the geometric structure of data.

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.

Lin, H.-T., & Lin, C.-J. (2003). A Study on Sigmoid Kernels for SVM and the Training ofnon-PSD Kernels by SMO-type Methods

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.

Yang, Y., Wang, W., Peng, L., Song, C., Chen, Y., Li, H., Yang, X., Lu, Q., Cai, D., Wu, B., & Liu, W. (2024). LoRA-Composer: Leveraging low-rank adaptation for multi-concept customization in training-free diffusion models.

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.

# 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. **potential additional baseline**
3. **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
4. ~~To test out potentially new approaches to model interaction terms~~
   1. ~~Gaussian Kernel extensions: Anisotropic Gaussian Kernels, Multi-scale kernels~~
5. ~~To test out potentially new approaches to model interaction terms~~
6. Code optimization
7. ~~Fill Conclusion~~
8. Fill Appendix
9. ~~Modify results and evaluation sections~~
10. Possibly more qualitative generation results
11. Standardize all formatting to APA (spacings, fontsize, in text citation, references, tables, figures, labelling fomatings…etc)
12. Tidy table of content (spacings, bolding…etc)
13. Update table of content
14. ~~Tidy model eval section table 5.1~~
15. Ensure all citations are proper and present
    1. And the names are all correct (search it up)
    2. Format
16. ~~Research more on section 5.3~~
17. ~~Tidy up intro and abstract~~
18. Tidy up references
19. **ENSURE ALL THE ITEMS IN THE GUIDE FOR THE REPORT ARE THERE**
    1. **Last done - abstract**
20. Some special acknowledgement to freedom, esp for code?
21. bgn for qr