Understanding and Enforcing Precise Control in Generative Models via Graph-Based Attention

by

Achint Soni

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Computer Science

Waterloo, Ontario, Canada, 2025

© Achint Soni 2025

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

The following publications form the basis on my thesis. I was the primary author on the first publication [120] and second author on the second publication [10].

Chapters 6 and 7 of this thesis include content from my primary author work which is currently under review at the time of thesis submission: Achint Soni, Meet Soni and Sirisha Rambhatla. "LOCATEdit : Graph Laplacian Optimized Cross Attention for Localized Text-Guided Image Editing" [120].

In this work, I proposed the main method, developed the codebase, conducted the analysis, and wrote the paper. Meet Soni contributed by running a part of experiments for the proposed method and assisted in formatting the manuscript. Sirisha Rambhatla participated in the discussions and provided feedback throughout the project and helped with manuscript writing.

Chapters 4 and 5 of this thesis include content from the work which has been published at ICML 2024: Pratik Bhowal, Achint Soni and Sirisha Rambhatla. "Why do Variational Autoencoders Really Promote Disentanglement?" [10]

In this work, I contributed by verifying the correctness of the proposed method, proofread all the mathematical theorems and propositions, and helped refine the manuscript. Pratik Bhowal came up with the proposed method and ran the experiments. Sirisha Rambhatla offered feedback, reviewed drafts, helped with verifying the theorems and contributed to manuscript writing.

Abstract

Generative models have significantly advanced in recent years, enabling unprecedented capabilities for data generation, manipulation, and editing. However, their practical applicability depends heavily on their ability to disentangle the underlying factors of variation, allowing precise and controllable modifications. This thesis explores disentanglement from two complementary perspectives: latent-space disentanglement in Variational Autoencoders (VAEs) and spatial disentanglement in diffusion-based text-guided image editing.

In the first part of the thesis, we investigate the mechanisms behind disentanglement in VAEs. By proposing a local non-linear approximation of the VAE decoder, we provide a rigorous theoretical analysis that reveals orthogonality of the decoder's Jacobian as a fundamental condition for disentanglement. To support this finding, we introduce a quantitative measure termed the Orthogonality Deviation Score (OD-Score) and empirically demonstrate across multiple benchmark datasets (dSprites, 3D Faces, 3D Shapes, and MPI3D) that increased orthogonality directly corresponds to improved disentanglement as measured by established metrics such as Mutual Information Gap (MIG) and MIG-Sup.

In the second part, we address the challenge of spatial disentanglement in text-guided image editing using diffusion models. Traditional diffusion-based methods rely primarily on cross-attention maps derived from textual prompts to determine regions for editing, often resulting in unintended alterations and compromised spatial coherence. To overcome this, we introduce LOCATEdit, a novel approach that refines attention maps using a graph-based regularization framework. LOCATEdit constructs a Cross and Self-Attention (CASA) graph, leveraging patch relationships derived from self-attention to promote spatial consistency and to constrain edits precisely within designated areas. Extensive evaluations on the PIE-Bench dataset illustrate that LOCATEdit achieves superior performance in localized editing tasks, substantially outperforming existing baselines in both semantic alignment and background preservation.

Together, these contributions offer a unified understanding of disentanglement in generative modeling, bridging theoretical insights from latent-space analysis with practical advancements in spatially coherent, text-guided image editing. Ultimately, this thesis provides a principled foundation for developing interpretable, reliable, and highly controllable generative systems.

Acknowledgements

I owe many thanks to my advisor Sirisha Rambhatla, who always had my best interest at heart and provided me with invaluable guidance through every possible aspect of a successful Masters study. Thank you for being always available to chat and for your time and patience during our countless hours of meetings. I am also very grateful to Charles Clarke, Prof. Stephan Mann and Yuntian Deng for serving as committee members for my thesis.

I would like to thank my grandparents, Bharti Soni and Dinesh Soni, who have always loved me and prayed for me throughout my life

I would also like to thank my younger sister, Harshvi Soni. She has always been my biggest support and true backbone.

Finally, I would like to thank my parents, Sanket Soni and Megha Soni. Both have sacrificed so much throughout their lives to support my studies. Without them, I would not be able to achieve anything. They were always there for me during tough times, believed in me unconditionally, and guided me when I went wrong. This master is dedicated to them.

Table of Contents

Author's Declaration ii

Statement of Contributions iii

Abstract iv

Acknowledgements v

List of Figures ix

List of Tables x

1 Introduction 1

1.1 Disentanglement in Generative Modeling 1

1.2 Latent-Space Disentanglement with VAEs 2

1.2.1 Defining Disentangled Representations 2

1.2.2 The VAE Framework and ELBO 2

1.2.3 Mechanisms of Disentanglement 2

1.3 Spatial Disentanglement in Text-Guided Editing 3

1.4 Challenges in Text-Guided Image Editing 3

1.5 Graph-Based Approach for Precise Editing 4

1.6 Integration into the Editing Pipeline 5

1.7 Unifying Latent and Spatial Disentanglement . . . 6
1.8 Thesis Organization . . . 7

2 Related Work **8**
2.1 Disentangled Representation Learning . . . 8
2.2 VAE-Based Architectures for Disentanglement . . . 9
2.3 Mechanistic Analyses of VAE Disentanglement . . . 9
2.4 Text-guided Image Editing . . . 9
2.5 Training Free Image Editing . . . 10
2.6 Graph Laplacian . . . 10

3 Background **12**
3.1 Variational Autoencoders . . . 12
3.2 The Evidence Lower Bound (ELBO) . . . 13
3.3 From Variational Autoencoders to Diffusion Models . . . 13
3.4 Diffusion Models . . . 14
3.5 Attention Mechanism . . . 14

4 From Local Linearity to Introducing Non-Linearity in VAEs **16**
4.1 Problem Setup and the β-VAE Objective . . . 16
4.2 The Polarized Regime and Local Decoder Approximation . . . 17
4.3 Decomposing the Reconstruction Loss . . . 17
4.4 Orthogonality from Stochastic Minimization . . . 18
4.5 From Orthogonality to Disentanglement . . . 19

5 Experiments on VAE Disentanglement **20**
5.1 Datasets . . . 20
5.2 Evaluation Metrics . . . 21
5.3 Models and Implementation . . . 21

5.4 Results . . . 22
5.4.1 Effect of Local Nonlinearity . . . 22
5.4.2 Approximation Method Comparison . . . 22
5.4.3 Orthogonality vs. Disentanglement . . . 23

6 LOCATEdit **25**
6.1 Dual-Branch Editing Paradigm . . . 25
6.2 Selective Embedding Interpolation . . . 27
6.3 Formulating CASA Graph . . . 28
6.4 Graph Laplacian Regularization . . . 29

7 Experiments on text guided image editing **32**
7.1 Dataset and Evaluation Metrics . . . 32
7.2 Comparison with Existing Methods . . . 33
7.3 Ablation Study . . . 33

8 Conclusion and Limitations **37**
8.1 Conclusion . . . 37
8.2 Limitations . . . 37
8.3 Future Directions . . . 38

References **40**
.1 Broader Impact . . . 58
.2 Proof of Lemma 1 . . . 58
.3 Proof of Theorem 1 . . . 59
.3.1 Optimization Problem . . . 59
.3.2 Existence and Uniqueness of the Solution . . . 60
.3.3 Additional Qualitative Results . . . 60

List of Figures

1.1 Illustration of disentangled representations 2
1.2 Example of over-editing caused due to imprecise masks. 6

5.1 Performance of different VAE architectures 24

6.1 An overview of the text-guided image editing pipeline 26
6.2 CASA (Cross and Self-Attention) Graph Construction Workflow. 27
6.3 Illustration of the convex objective 31

1 Refined masks after Graph Laplacian Regularization 61

List of Tables

5.1 Average approximation error: nonlinear vs. linear decoder. Lower is better. 22

5.2 Error of stochastic reconstruction approximation: our nonlinear f_D vs. prior linear Jacobian. 22

7.1 Qualitative comparision of LOCATEdit 35

7.2 Comparison of different methods based on structure, background preservation, and CLIP similarity metrics. 36

7.3 Comparison of different methods based on structure, background preservation, and CLIP similarity metrics. 36

1 Additional qualitative results on PIE-Bench 62

Chapter 1

Introduction

1.1 Disentanglement in Generative Modeling

In the last ten years, deep generative models have attained significant success in generating high-fidelity images, music, and various other data modalities. Variational Autoencoders (VAEs) [63, 106] proposed the concept of learning a continuous, probabilistic latent space; Generative Adversarial Networks (GANs) [38] attained photorealism through adversarial training; and, more recently, diffusion models [46, 119] have further enhanced sample quality. Nonetheless, despite these advancements, a significant obstacle to practical use persists: controllability. In practical applications—spanning medical imaging to digital art—users must be capable of directing the creative process in precise, predictable manners.

Central to controllability is **disentanglement**: the decomposition of a complex creative process into independent, semantically significant components. When a model's internal representation is disentangled, altering one factor (e.g., rotation, color, expression) does not unintentionally affect others. This attribute not only improves interpretability but also facilitates dependable downstream applications, including interactive editing [134, 40] and structured data synthesis.

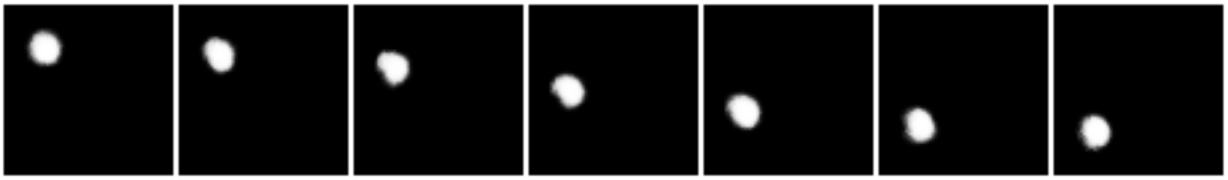


Figure 1.1: Varying a single latent variable, keeping all the other latent variables the same in the dSprites dataset changes only the vertical position of the object in the images while keeping the rest of its attributes constant, demonstrating disentanglement.

1.2 Latent-Space Disentanglement with VAEs

1.2.1 Defining Disentangled Representations

Disentangled Representation Learning (DRL) aims to identify latent codes $\mathbf{z} = (z_1, \ldots, z_d)$ such that each z_i represents an individual ground-truth source of variation within the data [44]. For example, in the dSprites dataset, altering a single coordinate may alone modify an object's vertical location (Fig. 1.1), while all other properties remain unchanged.

1.2.2 The VAE Framework and ELBO

VAEs cast representation learning as approximate inference under a probabilistic encoder–decoder model. Given data $\mathbf{x}$, the encoder $q_\phi(\mathbf{z} \mid \mathbf{x})$ and decoder $p_\theta(\mathbf{x} \mid \mathbf{z})$ optimize the Evidence Lower Bound (ELBO):

$$\mathcal{L}(\theta, \phi) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x} \mid \mathbf{z})] - \beta \, D_{\mathrm{KL}}\big(q_\phi(\mathbf{z} \mid \mathbf{x}) \,\|\, p(\mathbf{z})\big),$$

where $p(\mathbf{z}) = \mathcal{N}(0, I)$. By adjusting β, models like β-VAE [45], DIP-VAE [70], and β-TCVAE [20] encourage latent independence—at the cost of reconstruction fidelity.

1.2.3 Mechanisms of Disentanglement

Despite strong empirical findings, the precise mechanisms by which VAEs attain DRL remain contested. Initial analyses associated disentanglement with the diagonal posterior and local linearity of the decoder [107]. Subsequent research contended that the non-linear complexity of real-world data necessitates models that surpass local linear approximations [160]. In Chapter 4, we enhance this comprehension by (1) reevaluating the constraints

of local linearity and (2) presenting a local non-linear decoder component whose induced orthogonality further facilitates disentanglement.

1.3 Spatial Disentanglement in Text-Guided Editing

Diffusion models have emerged as an innovative method for image generation, rapidly gaining significance due to their capacity to create high-quality, realistic images. Their popularity has generated various practical applications—from digital art and advertising to medical imaging—where visual content is essential. However, although these models are proficient at generating images from inception, real applications often necessitate exact and controlled editing functionalities. Content creators, designers, and artists are increasingly seeking tools that enable them to alter images according to natural language descriptions. An intuitive editing interface would let users to define modifications, such as "render the sky sunset orange" or "add a smiling expression," rather than painstakingly adjusting individual pixels.

Text-guided image editing has therefore attracted significant attention in the research community [148, 137, 16]. By integrating the semantic understanding of language with the generative capabilities of diffusion models, these techniques offer an ideal combination of automation and creative control. The ability to direct image modifications via clear textual instructions could transform workflows across multiple creative fields, such as digital art, advertising [48, 33, 65, 125], and critical sectors like medical imaging [136, 91].

1.4 Challenges in Text-Guided Image Editing

Text-guided image editing aims to leverage the semantic power of language to modify images in a precise and user-friendly manner. In many methods, this is achieved by utilizing cross-attention mechanisms to associate text tokens with specific image regions, thereby guiding the editing process based on the provided prompt. However, these cross-attention maps are often imprecise, leading to modifications that spill over into unintended regions. Such inaccuracies can result in severe issues including object identity loss [66, 115, 138] and background drift [82, 148].

A key aspect of diffusion-based image editing involves the use of two complementary attention mechanisms. Cross-attention links textual tokens to targeted image regions, effectively determining where the edit should occur based on the semantic content of the prompt.

In contrast, self-attention captures the inherent spatial relationships and structural consistencies among different parts of the image, thus preserving underlying object boundaries and the overall layout. Although cross-attention is essential for localizing the edit, methods that rely exclusively on it can end up applying changes too broadly. Conversely, while self-attention helps maintain structure, naively merging it with cross-attention without careful handling may lead to information loss and does not fully address the issue of over-editing. This delicate balance between ensuring global consistency and achieving local specificity remains one of the central challenges in text-guided image editing.

Recent techniques have attempted to address these difficulties by integrating both cross- and self-attention cues. Several approaches adopt a dual-branch editing paradigm that leverages both types of attention to refine the editing masks. Additionally, some methods incorporate target image embeddings derived from source images and prompt pairs to provide auxiliary guidance. Despite these efforts, current solutions often fall short of achieving the desired precision. Inaccuracies in the computed attention maps can lead to modifications that are too diffuse, inadvertently affecting regions that should remain unchanged.

This persistent challenge of localizing edits accurately without compromising the overall image structure underscores the necessity for a more principled approach. In the following sections, we introduce a graph-based framework that explicitly enforces spatial consistency, thereby addressing these fundamental challenges in text-guided image editing.

1.5 Graph-Based Approach for Precise Editing

Our work addresses these challenges by introducing a novel graph-based framework for text-guided image editing. We observe that conventional attention maps can benefit from an explicit spatial regularization that enforces coherence and confines edits to intended regions. To achieve this, we propose modeling the attention maps as CASA graphs, where nodes represent image patches (derived from cross-attention maps) and edges are weighted according to self-attention relationships. This graph abstraction naturally integrates both local and global contexts, effectively linking fine-grained details with the broader semantic structure of the image.

The major innovation in our approach is to impose a Graph Laplacian regularizer on these CASA graphs. Graph Laplacian regularization has been widely used in image denoising, mesh editing, and segmentation tasks [95, 121, 133, 154] due to its ability to enforce smooth variations across interconnected regions. In our context, by applying this

regularizer to the attention maps, we promote spatial consistency: attention values are encouraged to vary gradually over regions that are closely connected in the graph, thereby preventing abrupt transitions and reducing the risk of over-editing.

For instance, consider a scenario in which a user instructs the model to "paint the car red." The raw cross-attention may highlight not only the car but also loosely include areas from the surrounding background. With our graph-based refinement, the self-attention-derived CASA graph delineates the boundaries of the car more accurately, and the Graph Laplacian regularizer smooths the attention values across patches that naturally belong together. As a result, the editing operation is confined to the car with minimal spillover to adjacent regions, thus preserving the overall structural integrity of the image.

This refined editing process is integrated directly into the diffusion sampling procedure. Importantly, the optimization for the graph regularizer admits a closed-form solution [137], which enables us to enforce spatial consistency efficiently without requiring iterative refinements or additional training.

1.6 Integration into the Editing Pipeline

To implement our approach, we adopt a dual-branch editing paradigm. In our framework, one branch reconstructs the original image while the other generates the edited version. Both branches share intermediate latent representations and originate from the same initial noise, ensuring structural consistency throughout the editing process. The CASA graphs are computed separately for the source and target branches from their respective cross- and self-attention maps. After applying the Graph Laplacian regularization, the optimized attention masks guide the subsequent denoising steps of the diffusion process. This integration ensures that the semantic changes prompted by the text are localized precisely, with the edited regions adhering closely to the structural cues derived from the self-attention graph.

Figure 1.2 exemplifies the consequence of imprecise masks, where over-editing results in unintended alterations of the background. ViMAEdit [137], using raw attention, generates an entangled and imprecise mask which results in spills in the mask generated. Our method explicitly mitigates this risk by confining modifications to the designated regions and preserving object boundaries.

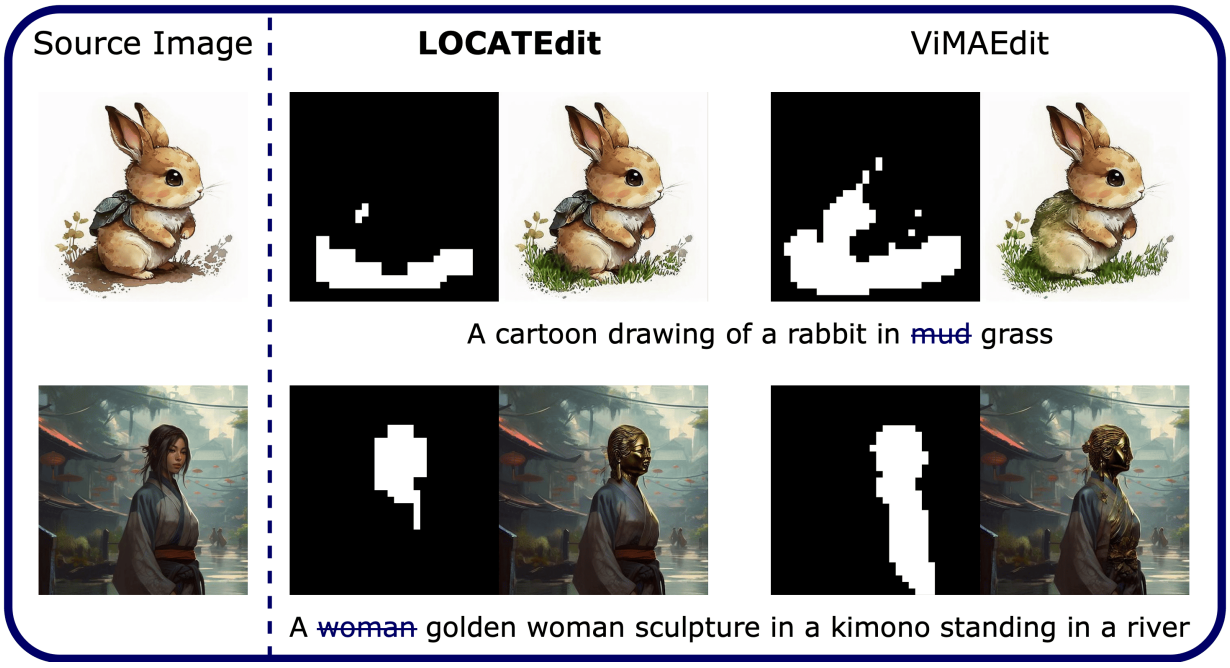


Figure 1.2: Example of over-editing caused due to imprecise masks.

1.7 Unifying Latent and Spatial Disentanglement

Both VAEs [10] and LOCATEdit [120] pursue the same fundamental goal—decomposing a complex generative process into independent factors—but at different levels of abstraction.

- **Latent disentanglement** isolates semantic attributes (e.g. pose, color) along orthogonal axes in the bottleneck layer of VAEs, i.e.,$\mathbf{z}$.
- **Spatial disentanglement** isolates pixel-patch correlations and generated precise masks, ensuring edits remain localized.

Insights into one domain can inform the other: orthogonality and smoothness priors that aid latent factorization also suggest graph-based constraints for attention maps. Together, they form a coherent framework for fine-grained, interpretable control of generative models.

1.8 Thesis Organization

- **Chapter 2 and 3** reviews related work in disentangled representation learning and text-guided editing.
- **Chapter 4** analyzes the mechanisms of DRL in VAEs, introducing local non-linearity and orthogonality.
- **Chapter 6** presents LOCATEdit, detailing the CASA graph and Graph Laplacian regularization.
- **Chapter 5 and 7** reports experiments evaluating both latent and spatial disentanglement against state-of-the-art baselines.
- **Chapter 8** concludes with limitations and future directions.

Chapter 2

Related Work

Recent progress in image editing has exploited a diverse array of conditioning signals—including text, reference images, and segmentation maps—to drive semantic, structural, and stylistic changes [49]. In our work, we focus particularly on text-guided image editing with an emphasis on retaining the original content and achieving effective foreground-background separation.

2.1 Disentangled Representation Learning

Disentangled representation learning (DRL) aims to factorize high-dimensional data into independent, human-interpretable factors. This pursuit is motivated by the need for interpretability, transparency, and causal reasoning in deep learning models [74, 8, 99]. A common definition of DRL requires that each latent variable capture exactly one underlying generative factor of the data [44], though some works highlight the role of inductive biases and dataset structure in enabling disentanglement [83, 84]. Well-disentangled representations facilitate downstream tasks—from interactive image editing and one-shot synthesis in computer vision [122, 135, 41] to fair decision making and causal inference [24, 99]—by allowing fine-grained control over individual attributes.

DRL methods have been applied across a broad range of domains. In adversarial learning, disentangled factors improve robustness [127, 88]; in graph neural networks, they help isolate and debias causal substructures [30]; in temporal modeling, they support independent dynamics [150]; and in fairness and interpretability, they prevent protected attributes from entangling with decision-relevant features [24, 118]. In computer vision,

disentanglement powers tasks such as zero-shot segmentation [41], cross-view localization [156], 3D retrieval [94], and medical image analysis [39].

2.2 VAE-Based Architectures for Disentanglement

Variational Autoencoders (VAEs) [63, 106] form the backbone of many DRL approaches. The β-VAE [45] introduces a weighting parameter β to up-weight the KL term in the ELBO, promoting latent independence. FactorVAE [62] and β-TCVAE [20] further decompose the KL divergence into terms that directly penalize total correlation among latent dimensions. Extensions such as guided-VAE [26] incorporate auxiliary labels to steer disentanglement, while sequential VAE variants (S3VAE) [159] and multi-view VAEs (Multi-VAE) [145] adapt the framework to temporal and multi-modal settings. Recent work on identifiability employs conditionally factorized priors [60, 90] and group-theoretic constraints [158], further strengthening the theoretical foundations of disentanglement.

2.3 Mechanistic Analyses of VAE Disentanglement

Given VAEs were not originally designed for DRL, several studies have sought to explain why and how they succeed. The information bottleneck perspective [15] argues that the KL term enforces minimal sufficient representations. Analyses of the variational family's geometry show that the ELBO's reconstruction term induces local orthogonality in the decoder Jacobian [69, 107]. Others have drawn parallels between locally linear VAEs and Principal Component Analysis [160], suggesting that small perturbations in latent space translate predictably in data space. However, these linearization arguments can fall short when modeling highly non-linear real-world variations. In our work, we extend these analyses by introducing a local non-linear decoder component and by rigorously linking decoder orthogonality to disentanglement both theoretically and empirically.

2.4 Text-guided Image Editing

Early work in text-guided image editing capitalized on the power of CLIP [103] to align visual content with textual information. For example, [61] refined diffusion models during reverse diffusion using a CLIP-based loss to alter image attributes, although these methods were mainly limited to broad, global modifications and often led to a decrease in

image quality. Later approaches, such as DiffuseIT [71] and StyleDiffusion [139], improved upon this by introducing additional losses that explicitly promote semantic or style disentanglement. However, these methods tend to be computationally heavy and are typically confined to specific style changes. More recently, frameworks like InstructPix2Pix [13] have been developed to preserve the source content by following text instructions. Yet, these methods often depend on carefully curated pairs of instruction and image examples along with supervised training. Other studies have explored manipulating text embeddings for better disentangled editing [141], but these adjustments have provided only incremental improvements over earlier approaches.

2.5 Training Free Image Editing

Recent developments in text-to-image synthesis [31, 92, 104, 108, 111] have paved the way for training-free editing techniques that take advantage of the detailed feature and attention maps available in pre-trained diffusion models. These methods use a dual-branch approach to extract rich signals from the diffusion process of the source image, allowing modifications to be made without further training. For instance, PRedITOR [105] generates a target CLIP embedding via a diffusion prior model, although it sometimes struggles with fine details and background consistency. Other methods, such as P2P [42] and PnP [132], focus on replacing or injecting attention maps—whether cross-attention for spatial consistency or self-attention for structural integrity—directly into decoder layers. Additional techniques, such as MasaCtrl [16] and attention regularization losses used in DPL [148] and refined in ViMAEdit [137], further enhance structural control. Despite these efforts, precisely localizing edits and maintaining consistent outputs continue to pose challenges.

2.6 Graph Laplacian

A recurring element in many of these methods is the use of graph Laplacian regularization. In the fields of optimization and semi-supervised learning, Laplacian regularization is known for promoting smooth changes along the graph, similar to how Conditional Random Fields (CRFs) enforce spatial and color consistency in segmentation tasks [73]. Unlike CRFs, Laplacian smoothing is fully differentiable, making it easy to integrate into modern neural network architectures. This technique has shown its effectiveness in applications such as image matting [77], where it helps maintain edge details while interpolating ambiguous regions, as well as in action localization [96], where it refines class activation maps

to produce more coherent results. In our work, such regularization plays a crucial role in ensuring spatial consistency during the editing process.

Chapter 3

Background

3.1 Variational Autoencoders

Variational Autoencoders (VAEs) are a class of deep generative models that have been widely adopted for learning interpretable latent representations of data. Introduced by [63] and [106], VAEs combine principles from variational inference with neural network-based encoding and decoding processes. Unlike traditional autoencoders, which typically learn deterministic mappings between an input and a latent representation, VAEs model the latent space probabilistically. This probabilistic treatment not only allows for robust data generation but also provides a natural framework for exploring the factors of variation inherent in the data.

A typical VAE comprises two main components: an encoder, $Enc_\phi : \mathbb{X} \to \mathbb{Z}$, and a decoder, $Dec_\theta : \mathbb{Z} \to \mathbb{X}$. Here, $\mathbb{X} \in \mathbb{R}^n$ represents the data space, and $\mathbb{Z} \in \mathbb{R}^d$ is the latent space. During training, each data sample $\boldsymbol{x}^{(i)}$ is encoded into a distribution $q_\phi(\boldsymbol{z}|\boldsymbol{x}^{(i)})$, which is typically modeled as a diagonal Gaussian:

$$q_\phi(\boldsymbol{z}|\boldsymbol{x}^{(i)}) = \mathcal{N}(\mu_\phi(\boldsymbol{x}^{(i)}), \text{diag}(\sigma^2_\phi(\boldsymbol{x}^{(i)}))),$$

where $\mu_\phi(\cdot)$ and $\sigma_\phi(\cdot)$ are outputs of the encoder network. A sample $\boldsymbol{z}^{(i)}$ drawn from this approximate posterior is then passed through the decoder to reconstruct the original image, leading to a reconstruction $\tilde{\boldsymbol{x}}^{(i)}$.

3.2 The Evidence Lower Bound (ELBO)

Training a VAE involves maximizing the marginal log-likelihood of the data

$$\sum_{i=1}^{N} \log(p(\boldsymbol{x}^{(i)})),$$

which is intractable in general. Instead, we maximize a tractable lower bound on the log-likelihood known as the Evidence Lower Bound (ELBO). The ELBO is defined as

$$L := \underbrace{\sum_{\boldsymbol{x}^{(i)} \in \mathbb{X}} \mathbb{E}_{\boldsymbol{z}^{(i)} \sim q_\phi(\boldsymbol{z}^{(i)}|\boldsymbol{x}^{(i)})}[\log p(\boldsymbol{x}^{(i)}|\boldsymbol{z}^{(i)})]}_{L_{MLE}} - \beta \underbrace{\sum_{\boldsymbol{x}^{(i)} \in \mathbb{X}} D_{\mathrm{KL}}\Big(q_\phi(\boldsymbol{z}^{(i)}|\boldsymbol{x}^{(i)}) \,\|\, p(\boldsymbol{z}^{(i)})\Big)}_{L_{KL}},$$

where $p(\boldsymbol{z})$ is generally chosen as a fixed isotropic Gaussian, $\mathcal{N}(\mathbf{0}, \mathbf{I})$, and β is a weighting factor that can be tuned to control the trade-off between reconstruction fidelity and the complexity of the latent distribution. The first term, L_{MLE}, serves as the reconstruction loss (commonly approximated by a squared error loss), while the second term, L_{KL}, enforces that the approximate posterior remains close to the prior.

3.3 From Variational Autoencoders to Diffusion Models

While Variational Autoencoders (VAEs) have long served as a powerful framework for learning interpretable and disentangled latent representations, recent advances in generative modeling have shifted considerable attention towards diffusion models. Unlike VAEs, which rely on a probabilistic encoder-decoder architecture and the evidence lower bound to capture the underlying factors of variation, diffusion models achieve high-fidelity image generation through a process of gradual noise injection and removal.

The key innovation behind diffusion models is their two-phase approach: a forward process that systematically corrupts a clean sample with Gaussian noise, and a backward process that reconstructs the image by learning to reverse this corruption. This framework enables not only the generation of photorealistic images but also, when combined with attention mechanisms, facilitates highly precise and localized image editing. In the following sections, we provide an in-depth overview of diffusion models and subsequently discuss the role of attention mechanisms in enhancing these models, thereby establishing a foundation for our work on precise, text-guided image editing.

3.4 Diffusion Models

Diffusion models [46, 119, 92] represent a class of generative methods that rely on two intertwined processes: forward and backward diffusion. In the forward process, one starts with a clean sample $\mathbf{z}_0$ (sampled from the data distribution) and iteratively adds Gaussian noise over timesteps $t = 1, 2, \ldots, T$. In this process, the formulation is given by

$$\mathbf{z}_t = \sqrt{\alpha_t}\,\mathbf{z}_0 \;+\; \sqrt{1-\alpha_t}\,\boldsymbol{\epsilon}_t, \quad t = 1, \ldots, T,$$

where $\boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ stands for the independent Gaussian noise added at each time step. The coefficients $\{\alpha_t\}_{t=1}^{T}$ control the noise variance, ensuring that after T steps, the resulting $\mathbf{z}_T$ closely follows a standard Gaussian distribution.

The backward diffusion process reverses this noising trajectory. It incrementally denoises the corrupted sample starting from $\mathbf{z}_T$, reconstructing intermediate states $\mathbf{z}_{T-1}$, $\mathbf{z}_{T-2}$, and so on, until arriving at a final clean approximation $\mathbf{z}_0$. This is achieved by sampling $\mathbf{z}_{t-1}$ from a learned conditional distribution parameterized by a denoising function. Formally, the reverse update can be expressed as

$$\mathbf{z}_{t-1} = \boldsymbol{\mu}_t\big(\mathbf{z}_t, \theta\big) + \sigma_t\,\tilde{\boldsymbol{\epsilon}}_t, \quad t = T, \ldots, 1,$$

where $\tilde{\boldsymbol{\epsilon}}_t$ is an added Gaussian noise, and $\boldsymbol{\mu}_t$ along with σ_t represent the mean and variance of the sampling distribution for $\mathbf{z}_{t-1}$, with θ encompassing the learnable parameters. For instance, the DDIM formulation [119] employs a deterministic version by fine-tuning the variance schedule, which not only accelerates sampling but also preserves high-quality outputs.

A cornerstone of modern diffusion frameworks is the noise prediction network $\boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t)$. Instead of directly predicting $\mathbf{z}_0$ or $\mathbf{z}_{t-1}$, the network estimates the noise present in $\mathbf{z}_t$. When well trained, this predictor effectively drives the reverse diffusion process, incrementally eliminating the Gaussian noise to yield a refined reconstruction.

3.5 Attention Mechanism

In practical implementations, the noise prediction component is commonly realized using a U-Net architecture, favored for its strong performance in pixel-wise prediction tasks. Each U-Net module is composed of: (i) a residual convolutional sub-block that enhances the

spatial features of intermediate outputs, (ii) a self-attention sub-block that captures long-distance dependencies among patches, and (iii) a cross-attention sub-block that integrates image features with corresponding textual information.

Within this framework, the feature maps are first transformed into three different embeddings: queries (Q), keys (K), and values (V). The attention mechanism is then computed using the formula

$$\text{Attention}(Q, K, V) = \text{Softmax}\Big(\frac{QK^\top}{\sqrt{d}}\Big)\, V,$$

where d denotes the dimensionality of the query and key vectors. In both self-attention and cross-attention contexts, the query Q is derived from the spatial features. For self-attention, both K and V originate from the same spatial features, while in cross-attention, they are computed from textual embeddings. These projection mappings are refined during training to optimize the network's performance.

Chapter 4

From Local Linearity to Introducing Non-Linearity in VAEs

This chapter is based on [10] which demonstrates why Variational Autoencoders (VAEs) – though not explicitly designed for disentangled representation learning—nevertheless tend to discover independent latent factors. Our key insight is that when a VAE operates under a polarized regime—where some latent dimensions are essentially ignored and others carry almost all the reconstruction information—a local nonlinear analysis reveals that minimizing the stochastic component of the reconstruction error against a fixed KL penalty naturally forces the decoder's Jacobian to have orthogonal columns. Orthogonality, in turn, aligns each latent dimension with a unique generative factor, thereby achieving disentanglement. We develop this argument in four steps: problem setup, the polarized regime and local decoder approximation, loss decomposition, and the orthogonality result.

4.1 Problem Setup and the β-VAE Objective

Let $\{\mathbf{x}^{(i)}\}_{i=1}^{N} \in \mathbb{R}^n$ denote our dataset of observations. A VAE encodes each $\mathbf{x}^{(i)}$ into a diagonal Gaussian posterior

$$q_\phi(\mathbf{z} \mid \mathbf{x}^{(i)}) \;=\; \mathcal{N}\big(\mu_\phi(\mathbf{x}^{(i)}),\, \mathrm{diag}(\sigma_\phi^2(\mathbf{x}^{(i)}))\big),$$

from which we sample $\mathbf{z}^{(i)} = \mu_\phi(\mathbf{x}^{(i)}) + \boldsymbol{\epsilon}^{(i)}$, $\boldsymbol{\epsilon}^{(i)} \sim \mathcal{N}(0, \mathrm{diag}(\sigma_\phi^2))$. The decoder $p_\theta(\mathbf{x} \mid \mathbf{z})$ then reconstructs $\tilde{\mathbf{x}}^{(i)}$.

Training minimizes the β-VAE loss

$$\mathcal{L} = \sum_{i=1}^{N} \underbrace{\mathbb{E}_{q_\phi} \|\tilde{\mathbf{x}}^{(i)} - \mathbf{x}^{(i)}\|^2}_{L_{\text{rec}}^{(i)}} + \beta \underbrace{D_{\text{KL}}\big(q_\phi(\mathbf{z}^{(i)} \mid \mathbf{x}^{(i)}) \parallel p(\mathbf{z})\big)}_{L_{\text{KL}}^{(i)}}.$$

Here $p(\mathbf{z})$ is the standard normal prior. The hyperparameter β balances reconstruction fidelity against the compactness of the latent representation. Empirically, VAEs often enter a regime in which some latent dimensions have posterior variances near one (and means near zero), contributing little to reconstruction, while a small subset of dimensions carries the essential information. We refer to this as the polarized regime.

4.2 The Polarized Regime and Local Decoder Approximation

In the polarized regime, we partition the latent indices into active set $\mathbb{V}_a$ and passive set $\mathbb{V}_p$. Passive latents $j \in \mathbb{V}_p$ satisfy $\sigma_j^2 \approx 1$ and $\mu_j \approx 0$, and the decoder's output is effectively independent of these coordinates. Active latents $j \in \mathbb{V}_a$ have $\sigma_j^2 \ll 1$ and nontrivial means μ_j, driving the reconstruction.

To analyze how these active latents influence disentanglement, we perform a local Taylor approximation of the decoder around each posterior mean $\mu^{(i)}$. Denote by $M_D^{(i)} \in \mathbb{R}^{n \times d}$ the Jacobian matrix of Dec_θ at $\mathbf{z} = \mu^{(i)}$. Then for small perturbations $\boldsymbol{\epsilon}$, we write

$$\text{Dec}_\theta\big(\mu^{(i)} + \boldsymbol{\epsilon}\big) \;\approx\; \text{Dec}_\theta\big(\mu^{(i)}\big) \;+\; f_D\big(M_D^{(i)}\,\boldsymbol{\epsilon}\big),$$

where f_D is a smooth nonlinear function capturing higher-order effects (see Appendix A for derivation). In this local model, the decoder splits into a linear map $M_D^{(i)}$ followed by a nonlinearity f_D.

4.3 Decomposing the Reconstruction Loss

Under the local approximation, the expected reconstruction error decomposes neatly into two parts:

$$L_{\text{rec}}^{(i)} = \mathbb{E}_{\boldsymbol{\epsilon}} \big\| \text{Dec}_\theta(\mu^{(i)} + \boldsymbol{\epsilon}) - \mathbf{x}^{(i)} \big\|^2 = L_{\text{rec}}^{\mu^{(i)}} \;+\; L_{\text{rec}}^{\text{stoch}^{(i)}},$$

where

$$L_{\text{rec}}^{\mu^{(i)}} = \left\|\text{Dec}_\theta(\mu^{(i)}) - \mathbf{x}^{(i)}\right\|^2, \quad L_{\text{rec}}^{\text{stoch}^{(i)}} = \mathbb{E}_{\boldsymbol{\epsilon}}\left\|f_D\big(M_D^{(i)}\boldsymbol{\epsilon}\big)\right\|^2.$$

The deterministic term $L_{\text{rec}}^{\mu^{(i)}}$ depends only on how well the decoder approximates $\mathbf{x}^{(i)}$ at the posterior mean, while the stochastic term measures the variance induced by random sampling. Crucially, $L_{\text{rec}}^{\text{stoch}^{(i)}}$ depends on the geometry of $M_D^{(i)}$: if the columns of $M_D^{(i)}$ are large or highly correlated, noise amplifies the reconstruction error.

4.4 Orthogonality from Stochastic Minimization

We now consider minimizing the total VAE loss under the constraint that the sum of KL terms remains fixed at C_{KL}. Intuitively, once the deterministic error L_{rec}^{μ} is optimized and the KL budget is exhausted, the only remaining degree of freedom is how to arrange the decoder Jacobians $M_D^{(i)}$ to minimize the stochastic penalty.

Theorem 1 (Orthogonality of Local Decoder Jacobians)**.** *Suppose $\sum_i L_{\text{KL}}^{(i)} = C_{\text{KL}}$ is held constant and the deterministic reconstructions $\{L_{\text{rec}}^{\mu^{(i)}}\}$ are fixed. Then the global minimizer of*

$$\sum_{i=1}^{N} \log L_{\text{rec}}^{\text{stoch}^{(i)}}$$

occurs precisely when, for every i, the columns of $M_D^{(i)}$ are orthogonal. Moreover, under this optimum each posterior variance satisfies

$$\sigma_j^{(i)2} \;\propto\; \frac{1}{\left\|c_j^{(i)}\right\|^2},$$

where $c_j^{(i)}$ denotes the jth column of $M_D^{(i)}$.

Proof sketch. The stochastic term $L_{\text{rec}}^{\text{stoch}}$ can be shown (via a second-order expansion of f_D) to depend on the sum of variances along each column direction of M_D. Under a fixed KL penalty, which itself constrains the magnitudes $\{\sigma_j^2\}$, the only way to minimize the induced noise is to ensure that different columns do not overlap—that is, are orthogonal.

4.5 From Orthogonality to Disentanglement

Orthogonal decoder axes have two important consequences for disentanglement:

1. **Coordinate Independence:** Varying a single latent coordinate z_j while holding others fixed moves the reconstruction only along the jth axis in feature space, so each latent controls a distinct direction of variation.

2. **Local Isometry under KL Divergence:** A smaller KL divergence causes encodings of nearby data points to cluster in latent space, and with orthogonal decoder axes, these clusters map back to data variations along a single generative factor.

Together, these imply that each active latent dimension aligns with one interpretable factor in the data, yielding the desired disentangled representation.

With this orthogonality in place, the VAE's latent space naturally factorizes into independent axes—each governing a single data attribute. In the next chapter, we translate this insight into the spatial domain, constructing a graph-based regularizer that disentangles "where" to edit from "what" to preserve in diffusion-based image editing.

Chapter 5

Experiments on VAE Disentanglement

In this chapter, we empirically validate the theoretical insights developed in Chapter 4. Our goals are threefold: (1) to demonstrate that a local nonlinear approximation of the decoder is more faithful than a purely linear one, (2) to introduce and compute an Orthogonality Deviation Score (OD-Score) that quantifies how closely each local decoder Jacobian approaches true orthogonality, and (3) to show that higher orthogonality—as measured by OD-Score—correlates with stronger disentanglement, as evaluated by the MIG and MIG-Sup metrics.

5.1 Datasets

We conduct experiments on four standard disentanglement benchmarks:

- **dSprites**, synthetic 2D shapes with five generative factors.
- **3D Faces** [98], rendered faces varying in pose and lighting.
- **3D Shapes** [14], synthetic 3D primitives with six factors.
- **MPI3D Complex** [35], real-world images of objects on a turntable.

5.2 Evaluation Metrics

Disentanglement: We report the Mutual Information Gap (MIG) [19] and its supervised variant MIG-Sup [79], where higher scores indicate more disentangled representations.

Approximation Error: To compare decoder approximations, we define for each sample $\mathbf{x}^{(i)}$ and noise $\boldsymbol{\epsilon}$:

$$\delta = \left\|\mathrm{Dec}_\theta(\mu^{(i)} + \boldsymbol{\epsilon}) - \mathrm{Dec}_\theta(\mu^{(i)}) - f_D\big(M_D\boldsymbol{\epsilon}\big)\right\|^2,$$

$$\delta_{\text{linear}} = \left\|\mathrm{Dec}_\theta(\mu^{(i)} + \boldsymbol{\epsilon}) - \mathrm{Dec}_\theta(\mu^{(i)}) - M_D\boldsymbol{\epsilon}\right\|^2.$$

We report the dataset-average of δ and δ_{linear}.

Orthogonality Deviation Score (OD-Score): For each local Jacobian M_D, let $M_D = U\Sigma V^\top$. We compute the normalized distance of U, Σ, V to the closest orthogonal factors (see Appendix C):

$$\mathrm{d}_U = \frac{\|U - U_{\text{orth}}\|_F^2}{\max_\phi \|U - U_{\text{orth},\phi}\|_F^2}, \quad \mathrm{d}_\Sigma,\ \mathrm{d}_V \text{ similarly},$$

and define

$$\mathrm{OD{-}Score}(M_D) = \frac{\mathrm{d}_U + \mathrm{d}_\Sigma + \mathrm{d}_V}{3}.$$

Lower OD-Score indicates closer to true orthogonality.

5.3 Models and Implementation

We evaluate three VAE variants: VAE, β-VAE $\big(\beta = 4\big)$, and β-TCVAE $\big(\beta = 6\big)$. All models share the same convolutional encoder/decoder architecture (see Appendix D for exact layer details), are trained for 200 epochs with Adam (learning rate 10^{-3}), and use minibatches of size 64.

5.4 Results

5.4.1 Effect of Local Nonlinearity

Table 5.1 compares the average δ and δ_{linear} across models and datasets. In every setting, the one-layer nonlinearity f_D yields a substantially smaller error than the purely linear approximation, confirming the necessity of modeling decoder nonlinearity.

Table 5.1: Average approximation error: nonlinear vs. linear decoder. Lower is better.

Model	dSprites	3D Faces	3D Shapes	MPI3D
		δ_{linear}		
VAE	0.6680	0.6039	0.7344	0.8042
β-VAE	0.5243	0.5799	0.7215	0.8001
β-TCVAE	0.4866	0.5690	0.7257	0.7836
		δ		
VAE	0.6643	0.5887	0.7316	0.8014
β-VAE	0.4907	0.5296	0.7010	0.7908
β-TCVAE	0.4532	0.5498	0.6879	0.7818

5.4.2 Approximation Method Comparison

Table 5.2 contrasts our nonlinear modeling with the purely linear Taylor expansion used in prior analyses [107]. Again, the nonlinear approach yields lower error consistently.

Table 5.2: Error of stochastic reconstruction approximation: our nonlinear f_D vs. prior linear Jacobian.

Method	dSprites	3D Faces	3D Shapes	MPI3D
Linearized (prior)	0.4866	0.5799	0.7344	0.8042
Ours (nonlinear)	0.4532	0.5296	0.7010	0.7908

5.4.3 Orthogonality vs. Disentanglement

Figure 5.1 plots OD-Score against MIG and MIG-Sup for each model and dataset. Panels (i)-(l) demonstrate the MIG versus OD-Score($\boldsymbol{M}_D$) scores for all the datasets for each of the three VAE-based architectures. Further, panels (m)-(p) demonstrate the MIG-Sup versus OD-Score($\boldsymbol{M}_D$) scores. We note that as the MIG and the MIG-Sup scores increase, the orthogonality as measured by OD-Score($\boldsymbol{M}_D$) decreases (lower is better). This establishes that orthogonality promotes disentanglement. We observe a strong negative correlation: as OD-Score decreases (i.e., columns become more orthogonal), both MIG and MIG-Sup increase, empirically linking decoder orthogonality to disentanglement.

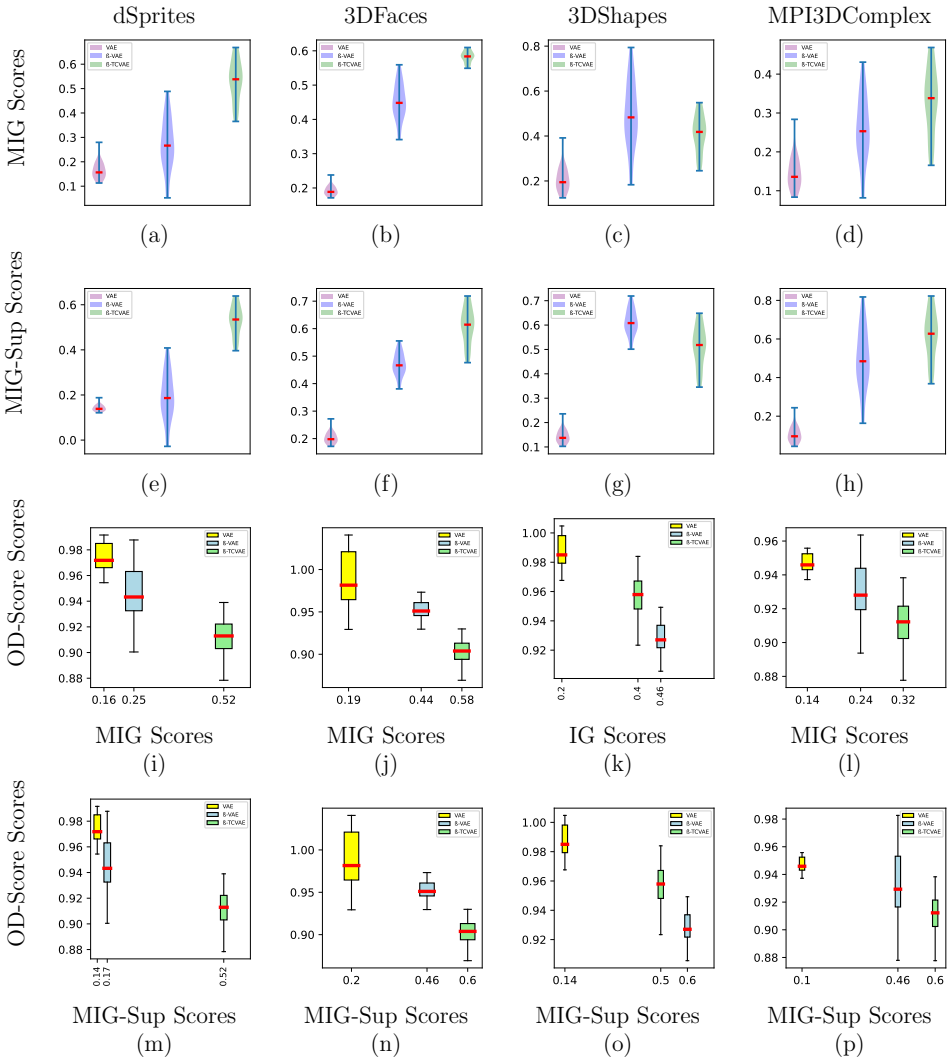


Figure 5.1: **Performance of different VAE architectures** Panels (a)-(d) illustrate the MIG scores (a higher score promotes disentanglement) of different VAE architectures for dSprites, 3DFaces, 3DShapes and MPI3DComplex, respectively. Panels (e)-(h) illustrate the MIG-Sup scores (a higher score promotes disentanglement). Panels (i)-(l) illustrate positive correlation between orthogonality, measured by OD-Score($\boldsymbol{M}_D$) (a lower score promotes orthogonality) and MIG scores for specified models and datasets. Finally, Panels (m)-(p) illustrate the OD-Score($\boldsymbol{M}_D$) vs the MIG-Sup score.

Chapter 6

LOCATEdit

In this chapter, we introduce LOCATEdit, a method designed for precise, localized text-guided image editing through the refinement of cross-attention maps. Our method combines two complementary modules to achieve this goal. First, we employ the CASA graph, which enforces spatial coherence and effectively limits edits strictly to the intended regions of the image. Second, following insights from previous research [137], we incorporate an enhanced denoising process that leverages image embeddings further refined by a selective pruning operation on text embedding offsets. This pruning selectively removes minor semantic variations, significantly reducing unintended modifications and preventing edits in areas outside the target regions. By integrating these two components, LOCATEdit successfully preserves the structural integrity of the original image while accurately applying the specified edits.

6.1 Dual-Branch Editing Paradigm

Our pipeline, as shown in Figure 6.1, uses a dual-branch architecture consisting of a source branch that reconstructs the original image and a target branch that generates the edited version. To ensure structural consistency, both branches originate from the same initial noise $\mathbf{z}_T$ and share intermediate latent representations. Importantly, cross-attention maps from the source branch are integrated into the target branch [42] to preserve spatial structure. Formally, given that Q^{src} and K^{src} represent the query and key embeddings from the source branch, and V^{tgt} represents the value embeddings from the target branch, the target cross-attention is computed as:

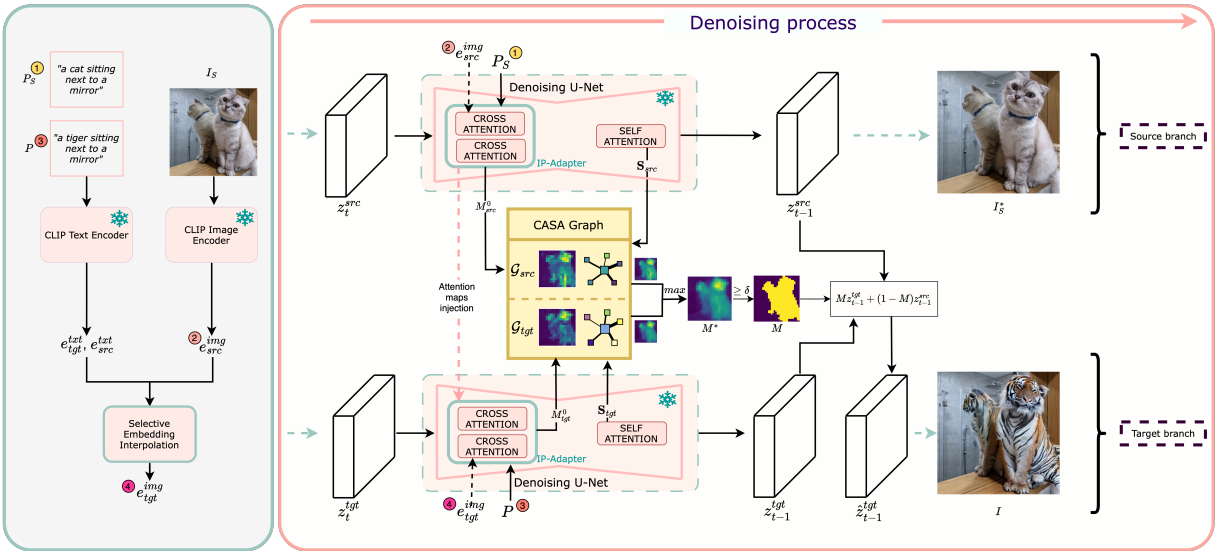


Figure 6.1: **An overview of the text-guided image editing pipeline.** LOCATEdit refines the cross-attention maps via graph Laplacian regularization to ensure spatial consistency, incorporates an IP-Adapter for extra guidance, and performs selective pruning on text embeddings to minimize noise, thus preserving essential structural details in the edited image.

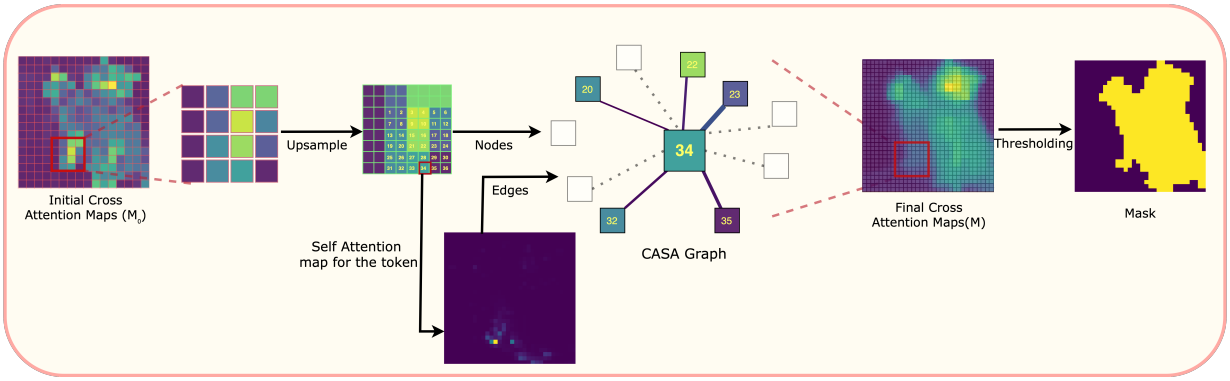


Figure 6.2: **CASA (Cross and Self-Attention) Graph Construction Workflow.** The initial cross-attention maps are upsampled to form a patch-level adjacency graph, after which Laplacian regularization enforces spatial consistency. Thresholding the refined maps yields the final, robust attention masks.

$$\text{Attention}\big(Q^{\text{src}}, K^{\text{src}}, V^{\text{tgt}}\big) = \text{Softmax}\Big(\frac{Q^{\text{src}}(K^{\text{src}})^{\top}}{\sqrt{d}}\Big) V^{\text{tgt}}.$$

6.2 Selective Embedding Interpolation

Following previous work [137], we employ an IP-Adapter [151] to provide explicit guidance for target image generation. After extracting the source image embedding e^{img}_{src} and the CLIP-based text embeddings e^{txt}_{src} and e^{txt}_{tgt} corresponding to the source and target prompts respectively, the conventional target image embedding is computed as

$$e^{img}_{\text{tgt}} = e^{img}_{\text{src}} + \Big(e^{txt}_{\text{src}} - e^{txt}_{\text{tgt}}\Big). \tag{6.1}$$

This embedding is then processed by the IP-Adapter, which projects it into a latent feature space that is integrated into the diffusion model's cross-attention mechanism. Specifically, given the query Q (derived from the noisy latent), the IP-Adapter produces additional key and value features K^{IP} and V^{IP} from the projected target embedding. These are then combined with the original key K and value V features to form the final cross-attention:

$$Z = \text{Attention}(Q, K, V) + \lambda \text{Attention}(Q, K^{\text{IP}}, V^{\text{IP}})$$

thereby incorporating semantic guidance into the diffusion process without requiring any additional training.

A limitation of directly using the difference $e^{txt}_{\text{src}} - e^{txt}_{\text{tgt}}$ in Equation (6.1) is that low-magnitude components, inherent in the entangled nature of CLIP text embeddings [86], can lead to unintended edits. To mitigate this, we introduce a selective pruning operator $\mathcal{H}$ that thresholds the text difference, retaining only the dominant semantic offsets. Formally, we replace Equation (6.1) with

$$e^{img}_{\text{tgt}} = e^{img}_{\text{src}} + \mathcal{H}\Big(e^{txt}_{\text{src}} - e^{txt}_{\text{tgt}}\Big), \tag{6.2}$$

where $\mathcal{H} : \mathbb{R}^d \to \mathbb{R}^d$ is defined elementwise as

$$\big[\mathcal{H}(\mathbf{y})\big]_i = \begin{cases} y_i, & \text{if } |y_i| \geq \tau, \\ 0, & \text{otherwise.} \end{cases} \tag{6.3}$$

Here, d is the embedding dimension and τ is determined via a percentile threshold on the absolute values of the difference. This selective pruning ensures that only significant semantic shifts contribute to the target image embedding, thereby reducing the risk of global edits and preserving the structural consistency of non-target regions. The pruned embedding is then processed through the IP-Adapter as described above, ensuring that the final embedding captures the intended modifications.

6.3 Formulating CASA Graph

While the IP-Adapter provides explicit semantic guidance, the cross-attention maps extracted during denoising may still contain spills that lead to unintended edits. To address this, we refine these attention maps by modeling them as a CASA graph, where each node represents an image patch and the edges capture patch-to-patch relationships obtained from self-attention, as illustrated in Figure 6.2. The graph Laplacian regularization enforces a smoothness constraint across the CASA graph, penalizing abrupt differences in attention between strongly connected patches. This smoothing suppresses isolated high responses that can cause over-editing, ensuring that only spatially coherent regions receive significant modifications. By harmonizing the attention values over connected patches, LOCATEdit robustly confines edits to the intended regions and preserves the overall spatial consistency.

Formally, within each U-Net block, each prompt word is associated with a cross-attention map; however, only the maps corresponding to the blend word(s) are necessary. Following prior work [18, 148, 42], we compute the average of the cross-attention maps obtained from multiple U-Net blocks to derive the initial maps. We obtain initial attention maps for both the source and target branches, denoted as $M_0^{\text{src}} \in \mathbb{R}^{r\times r}$ and $M_0^{\text{tgt}} \in \mathbb{R}^{r\times r}$. These masks are upsampled to a higher resolution of $R \times R$ (where $R = \gamma r$ and $\gamma > 1$) to capture fine spatial details and are subsequently flattened to yield the initial saliency maps $\mathbf{m}_0^{\text{src}}$, $\mathbf{m}_0^{\text{tgt}} \in \mathbb{R}^{R^2}$.

To prioritize high-confidence regions, we compute a weight for each patch by applying the sigmoid function $\sigma(\cdot)$ to the corresponding element of $\mathbf{m}_0^{\text{src}}$ and then squaring the output. This emphasizes higher values while suppressing lower ones, enhancing the reliability of high-confidence regions. These weights are compiled into a diagonal confidence matrix with a scaling factor α:

$$\mathbf{\Lambda}^{\text{src}} = \text{diag}\Big(\sigma\Big(\alpha \mathbf{m}_0[1]\Big)^2, \ldots, \sigma\Big(\alpha \mathbf{m}_0[R^2]\Big)^2\Big). \tag{6.4}$$

A similar formulation is used for $\mathbf{\Lambda}^{\text{tgt}}$.

Next, we extract self-attention maps $\mathbf{S}^{\text{src}} \in \mathbb{R}^{R^2\times R^2}$ and $\mathbf{S}^{\text{tgt}} \in \mathbb{R}^{R^2\times R^2}$ for the source and target branches, respectively. To treat mutual relationships uniformly and to ensure convexity in the optimization process, we symmetrize both maps as follows:

$$\mathbf{S}_{\text{sym}} = \frac{1}{2}\Big(\mathbf{S} + \mathbf{S}^\top\Big). \tag{6.5}$$

For each branch, we then construct a CASA graph $\mathcal{G} = (V, E)$, where each node $v_i \in V$ corresponds to a patch in the flattened saliency map $\mathbf{m}_0$. The edge weight between nodes v_i and v_j is defined by the symmetrized self-attention map $\mathbf{S}_{\text{sym}}$. This graph structure, with nodes representing initial saliency values and edges encoding inter-patch relationships, forms the foundation of the CASA graph.

6.4 Graph Laplacian Regularization

After initializing the CASA graphs $\mathcal{G}_{\text{src}}$ and $\mathcal{G}_{\text{tgt}}$ for both branches, we optimize the node values using graph Laplacian regularization.

Formally, the graph Laplacian is defined as:

$$\mathbf{L} = \mathbf{D} - \mathbf{S}_{\text{sym}},$$

where $\mathbf{D}$ is the degree matrix of $\mathbf{S}_{\text{sym}}$, computed by:

$$\mathbf{D}(i,i) = \sum_{j=1}^{R^2} \mathbf{S}_{\text{sym}}(i,j), \quad \mathbf{D}(i,j) = 0 \quad \text{for } i \neq j.$$

Lemma 1. *The graph Laplacian $\mathbf{L} \in \mathbb{R}^{R^2 \times R^2}$ is positive semidefinite.*

A detailed proof is provided in Appendix .2.

We refine the initial saliency maps $\mathbf{m}_0^{\text{src}}$ and $\mathbf{m}_0^{\text{tgt}}$ by solving the following convex optimization problem:

Theorem 2. *Let $\mathbf{m}_0 \in \mathbb{R}^{R^2}$ denote the initial saliency map, and let $\mathbf{\Lambda}$ and $\mathbf{L}$ be defined as above. The optimal saliency map $\mathbf{m}^* \in \mathbb{R}^{R^2}$ is the unique minimizer of the following objective:*

$$J(\mathbf{m}) = (\mathbf{m} - \mathbf{m}_0)^\top \mathbf{\Lambda} (\mathbf{m} - \mathbf{m}_0) + \lambda\, \mathbf{m}^\top \mathbf{L}\, \mathbf{m},$$

with the closed-form solution:

$$\mathbf{m}^* = (\mathbf{\Lambda} + \lambda\, \mathbf{L})^{-1}\, \mathbf{\Lambda}\, \mathbf{m}_0.$$

The proof is included in Appendix .3 and an illustration of the convexity can be viewed in Figure 6.3.

The refined saliency maps $\mathbf{m}^{*\text{src}}$ and $\mathbf{m}^{*\text{tgt}}$ are then reshaped back to $M^{*\text{src}}$ and $M^{*\text{tgt}}$, respectively. We compute their element-wise maximum to obtain the final map:

$$M^* = \max\{M^*_{\text{src}},\, M^*_{\text{tgt}}\}.$$

Thresholding M^* with a parameter δ yields the final spatial mask M. This optimized CASA graph results in a smooth, spatially coherent mask that retains high-confidence regions while preventing over-editing in less reliable areas.

To further ensure that non-editable regions and background areas remain unchanged, we apply the optimized mask at each denoising step by replacing the target branch's latent representation as follows:

$$\hat{\mathbf{z}}^{\text{tgt}}_{t-1} = M \odot \mathbf{z}^{\text{tgt}}_{t-1} \;+\; (1 - M) \odot \mathbf{z}^{\text{src}}_{t-1}, \quad t = T, \ldots, 1.$$

Here, $\odot$ represents the Hadamard (element-wise) product. This step ensures consistency in the background and unedited regions throughout the denoising process.

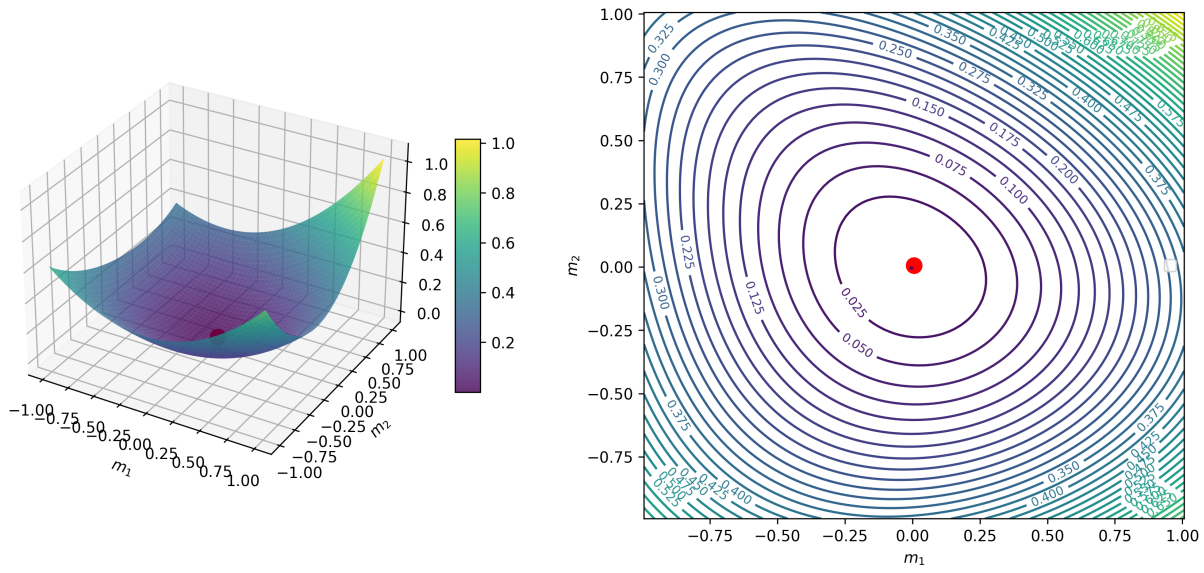


Figure 6.3: Illustration of the convex objective $J(\mathbf{m})$ in a 2D slice of the higher-dimensional space. The single global minimum, marked in red, highlights the function's convex nature.

Chapter 7

Experiments on text guided image editing

7.1 Dataset and Evaluation Metrics

Following recent work [57, 146, 137], we evaluate our approach using the PIE-Bench dataset [57], which is currently the only established benchmark specifically designed for prompt-based image editing. PIE-Bench contains 700 images divided across ten different editing tasks. Each image in the dataset is paired with a source prompt, a target prompt, a set of blend words (i.e., words that highlight the intended change), and an editing mask. While only the source prompt, target prompt, and blend words are required to perform the editing operation, the editing mask is used to evaluate how well the method preserves the image background.

To conduct a thorough evaluation, we adopt the strategy described in [57], focusing on three key criteria:

1. **Structure consistency**: Measured by computing the difference between DINO self-similarity maps [131], which capture the preservation of overall structural layout.
2. **Background preservation**: Quantified using several metrics including PSNR, LPIPS [155], MSE, and SSIM [140], all of which assess how closely the non-edited regions match the original image.

3. **Target prompt–image alignment**: Assessed through CLIP similarity [43], which evaluates how well the generated image aligns with the semantics of the target prompt.

7.2 Comparison with Existing Methods

LOCATEdit consistently delivers superior spatial consistency and semantic alignment compared to state-of-the-art text-guided image editing methods, as shown in Table 7.3. Unlike P2P-Zero [97] and PnP-based techniques [131, 57], which often result in global modifications and exhibit spatial inconsistencies, LOCATEdit restricts edits to the intended regions, thereby preserving the structural details of the source image. While mask-guided methods such as ViMAEdit [137] improve spatial localization, they can still introduce unintended artifacts in areas that are meant to remain unchanged as visible in Figure 1.2. By contrast, our use of graph Laplacian regularization refines cross-attention maps through smooth and coherent patch-to-patch interactions, effectively addressing these limitations.

In addition, methods like Edit-Friendly Inversion [50] and InfEdit with Virtual Inversion [146] demonstrate stronger semantic alignment with the target prompt but often fail to disentangle the editable content from the preserved background. LOCATEdit overcomes this challenge by robustly separating these regions, ensuring that edits are both precise and spatially confined, as illustrated in the qualitative results in Figure 7.1. For example, in the cat and the tiger example, the tiger in our image better matches the pose than the other methods, and doesn't have the artifacts of InfEdit, MasaCtrl, or LEDITS++. Another example is about closing the cat's eyes. None of the methods other than LOCATEdit is able to close the cat's eyes. Similar improvements can be seen in other examples.

Overall, the experimental results highlight that LOCATEdit not only improves the quality of the edited content but also maintains the integrity of the source image. By leveraging CASA graph-based attention refinement, our method outperforms existing approaches across multiple evaluation metrics, reinforcing the importance of spatially consistent and disentangled editing in practical text-guided image editing tasks.

7.3 Ablation Study

To validate the effectiveness of my model, we present results from three ablation variants. In the first variant, we remove the diagonal weighting matrix and instead use a uniform

L^2 penalty as the first term in Equation 1. The second variant omits the symmetrization of the self-attention matrix $\mathbf{S}$, which is a critical step for ensuring that the resulting graph Laplacian is positive semidefinite. In the third variant, we eliminate α-based confidence control by fixing $\alpha = 1$ in Equation 6.4, removing the adaptive emphasis on high-confidence regions.

Table 7.2 demonstrates that each of these components meaningfully contributes to the performance of the full model, particularly in terms of structure consistency and background preservation. Interestingly, we observe that when all components are combined, the performance on structure and background similarity metrics can slightly degrade compared to some ablated versions. However, this configuration achieves the highest CLIP similarity, indicating a stronger semantic alignment with the target prompt. It is also worth noting that even the slightly lower scores in structure and background metrics still surpass those reported by all baseline methods in Table 7.3.

While tuning the α parameter, we found that higher values lead to much stronger CLIP similarity but at a significant cost to the other metrics. This is expected, as large α values introduce a sharper thresholding effect, clearly separating trusted and adjusted regions. Although this enhances semantic alignment, it often results in abrupt transitions and reduced visual smoothness in non-edited areas.

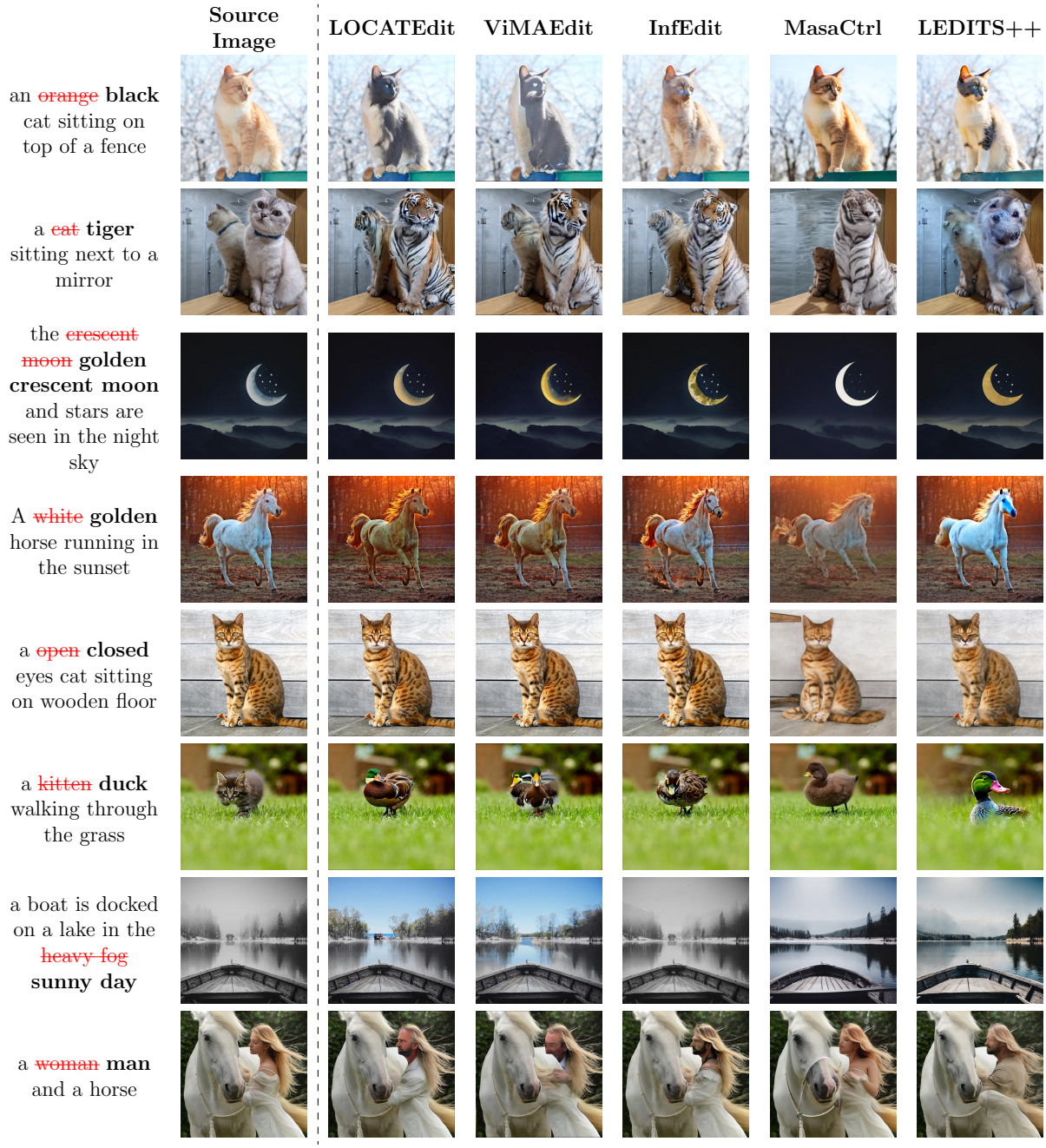


Table 7.1: Qualitative comparisons with competing text-guided editing methods. LOCATEdit yields more localized edits while preserving overall structure, outperforming baselines in both fidelity and consistency.

Method	Structure	Background Preservation				CLIP Similarity	
	Distance $\downarrow \times 10^3$	PSNR $\uparrow$	LPIPS $\downarrow \times 10^3$	MSE $\downarrow \times 10^4$	SSIM $\uparrow \times 10^2$	Whole $\uparrow$	Edited $\uparrow$
LOCATEdit	13.19	29.20	41.60	26.90	86.53	**25.96**	**23.02**
w/o diagonal weighting matrix	8.68	**29.59**	**38.16**	**23.17**	**86.83**	25.33	22.34
w/o symmetric self-attention	**8.59**	29.42	38.75	24.09	85.66	25.29	22.26
w/o α-based control	8.86	29.26	38.96	24.53	86.60	25.31	22.28
with high α	20.37	24.27	87.37	60.82	82.22	26.58	23.22

Table 7.2: Comparison of different methods based on structure, background preservation, and CLIP similarity metrics.

Method		Editing	Structure	Background Preservation				CLIP Similarity	
Inverse	Sampling (steps)		Distance $\downarrow \times 10^3$	PSNR $\uparrow$	LPIPS $\downarrow \times 10^3$	MSE $\downarrow \times 10^4$	SSIM $\uparrow \times 10^2$	Whole $\uparrow$	Edited $\uparrow$
VI	DDCM(12)	InfEdit	13.78	28.51	47.58	32.09	85.66	25.03	22.22
VI	DDIM(50)	ViMAEdit	12.65	28.27	44.67	30.29	85.65	25.91	22.96
PnP-I	DDIM(50)	P2P-Zero	51.13	21.23	143.87	135.00	77.23	23.36	21.03
		MasaCtrl	24.47	22.78	87.38	79.91	81.36	24.42	21.38
		PnP	24.29	22.64	106.06	80.45	79.68	25.41	22.62
		P2P	<u>11.64</u>	27.19	54.44	33.15	84.71	25.03	22.13
		ViMAEdit	11.90	28.75	43.07	28.85	85.95	25.43	22.40
		LOCATEdit (Ours)	13.19	**29.20**	<u>41.60</u>	<u>26.90</u>	**86.53**	<u>25.96</u>	**23.02**
EF	DPM-Solver++(20)	LEDITS++	23.15	24.67	80.79	118.56	81.55	25.01	22.09
		P2P	14.52	27.05	50.72	37.48	84.97	25.36	22.43
		ViMAEdit	14.16	28.12	45.62	33.56	85.61	25.51	22.56
		LOCATEdit (Ours)	**8.71**	<u>29.16</u>	**39.31**	**24.01**	<u>86.52</u>	**26.07**	22.43

Table 7.3: Comparison of different methods based on structure, background preservation, and CLIP similarity metrics.

Chapter 8

Conclusion and Limitations

8.1 Conclusion

In this thesis, we have introduced LOCATEdit, a text-guided image editing framework that combines the strengths of pre-trained diffusion models with a novel graph-based attention refinement. By modeling cross-attention maps as nodes in a CASA graph and using self-attention affinities as edge weights, we impose a Graph Laplacian regularizer to enforce spatial coherence. This produces editing masks that remain tightly localized to the intended regions, preventing unwanted spill-over and preserving background fidelity. Extensive evaluations on PIE-Bench and diverse baselines demonstrate that LOCATEdit achieves state-of-the-art performance in both semantic alignment and structure preservation, while remaining training-free and integrable into existing diffusion pipelines.

8.2 Limitations

Despite its strengths, LOCATEdit has several noteworthy limitations:

1. **Dependence on Attention Map Quality.** Both the initial cross-attention mask and the self-attention affinities must be reasonably accurate. If cross-attention mis-aligns the prompt word with unintended regions, or if self-attention produces spurious patch relations (e.g. in highly textured or abstract scenes), the CASA graph may reinforce incorrect connections, and the Laplacian smoothing cannot fully recover the true edit region.

2. **Hyperparameter Sensitivity.** Constructing and regularizing the CASA graph relies on several tunable parameters—upsampling factor γ, confidence weight α, Laplacian weight λ, and threshold δ. Although we identify default values that work well on PIE-Bench, adapting to new domains or significantly different prompts often requires manual retuning, which can hinder out-of-the-box usability.

3. **Limited Support for Complex Multi-Object Edits.** LOCATEdit is optimized for single-object or single-attribute changes. When a prompt requests multiple, interacting edits (e.g. "make both the cat and the dog wear hats"), the dual-branch paradigm and single mask per blend word may fail to disentangle overlapping edit regions, leading to incomplete or merged modifications.

4. **Lack of Interactive Mask Refinement.** Once the CASA graph produces its mask, there is no built-in mechanism for user correction within the pipeline. If the computed mask partially includes unwanted regions or omits small target areas, the user must resort to external image-editing tools, breaking the seamless text-to-image workflow.

8.3 Future Directions

Addressing these limitations suggests several promising avenues for further research:

- Robust Attention Priors: Incorporate stronger semantic or geometric priors—such as user-provided bounding boxes or learned object detectors—to correct initial attention errors.

- Automated Hyperparameter Adaptation: Develop meta-learning or reinforcement-learning strategies that tune γ, α, λ, and δ on the fly based on prompt characteristics and feedback signals.

- Multi-Mask and Multi-Prompt Extensions: Extend the CASA graph formulation to handle multiple blend words or hierarchical masks, enabling complex multi-object and scene-level edits.

- Interactive Mask Editing Interface: Integrate a lightweight UI component that allows users to click, brush, or verbally specify corrections to the mask before the final denoising stages.

By pursuing these directions, we aim to make LOCATEdit more robust, flexible, and user-friendly, further bridging the gap between human creative intent and diffusion-based image generation.

References

[1] Alessandro Achille and Stefano Soatto. Emergence of invariance and disentanglement in deep representations. The Journal of Machine Learning Research, 19(1):1947–1980, 2018.

[2] Tameem Adel, Zoubin Ghahramani, and Adrian Weller. Discovering interpretable representations for both deep generative and discriminative models. In International Conference on Machine Learning, pages 50–59. PMLR, 2018.

[3] Aishwarya Agarwal, Srikrishna Karanam, KJ Joseph, Apoorv Saxena, Koustava Goswami, and Balaji Vasan Srinivasan. A-star: Test-time attention segregation and retention for text-to-image synthesis. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 2283–2293, 2023.

[4] Francis R Bach and Michael I Jordan. Kernel independent component analysis. Journal of machine learning research, 3(Jul):1–48, 2002.

[5] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, et al. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. arXiv preprint arXiv:2211.01324, 2022.

[6] Jonathan T Barron and Jitendra Malik. Shape, albedo, and illumination from a single image of an unknown object. In 2012 IEEE Conference on Computer Vision and Pattern Recognition, pages 334–341. IEEE, 2012.

[7] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. Neural computation, 15(6):1373–1396, 2003.

[8] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. IEEE transactions on pattern analysis and machine intelligence, 35(8):1798–1828, 2013.

[9] Yoshua Bengio, Yann LeCun, et al. Scaling learning algorithms towards ai. Large-scale kernel machines, 34(5):1–41, 2007.

[10] Pratik Bhowal, Achint Soni, and Sirisha Rambhatla. Why do variational autoencoders really promote disentanglement? In Forty-first International Conference on Machine Learning, 2024.

[11] Piotr Bojanowski, Armand Joulin, David Lopez-Paz, and Arthur Szlam. Optimizing the latent space of generative networks. arXiv preprint arXiv:1707.05776, 2017.

[12] Manuel Brack, Felix Friedrich, Katharia Kornmeier, Linoy Tsaban, Patrick Schramowski, Kristian Kersting, and Apolinário Passos. Ledits++: Limitless image editing using text-to-image models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8861–8870, 2024.

[13] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 18392–18402, June 2023.

[14] Chris Burgess and Hyunjik Kim. 3D shapes dataset. https://github.com/deepmind/3dshapes-dataset/, 2018.

[15] Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-vae. arXiv preprint arXiv:1804.03599, 2018.

[16] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 22560–22570, 2023.

[17] J-F Cardoso. Source separation using higher order moments. In International Conference on Acoustics, Speech, and Signal Processing,, pages 2109–2112. IEEE, 1989.

[18] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. ACM Transactions on Graphics (TOG), 42(4):1–10, 2023.

[19] Ricky TQ Chen, Xuechen Li, Roger Grosse, and David Duvenaud. Isolating sources of disentanglement in vaes. In Proceedings of the 32nd International Conference on Neural Information Processing Systems, volume 2615, page 2625, 2018.

[20] Ricky TQ Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. Advances in neural information processing systems, 31, 2018.

[21] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. Advances in neural information processing systems, 29, 2016.

[22] Pierre Comon. Independent component analysis, a new concept? Signal processing, 36(3):287–314, 1994.

[23] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. arXiv preprint arXiv:2210.11427, 2022.

[24] Elliot Creager, David Madras, Jörn-Henrik Jacobsen, Marissa Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. Flexibly fair representation learning by disentanglement. In International conference on machine learning, pages 1436–1445. PMLR, 2019.

[25] Rémi Cuingnet, Marie Chupin, Habib Benali, and Olivier Colliot. Spatial and anatomical regularization of svm for brain image analysis. Advances in Neural Information Processing Systems, 23, 2010.

[26] Zheng Ding, Yifan Xu, Weijian Xu, Gaurav Parmar, Yang Yang, Max Welling, and Zhuowen Tu. Guided variational autoencoder for disentanglement learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 7920–7929, 2020.

[27] Sunny Duan, Loic Matthey, Andre Saraiva, Nicholas Watters, Christopher P Burgess, Alexander Lerchner, and Irina Higgins. Unsupervised model selection for variational disentangled representation learning. arXiv preprint arXiv:1905.12614, 2019.

[28] Cian Eastwood and Christopher KI Williams. A framework for the quantitative evaluation of disentangled representations. In International Conference on Learning Representations, 2018.

[29] Jesse Engel, Matthew Hoffman, and Adam Roberts. Latent constraints: Learning to generate conditionally from unconditional generative models. arXiv preprint arXiv:1711.05772, 2017.

[30] Shaohua Fan, Xiao Wang, Yanhu Mo, Chuan Shi, and Jian Tang. Debiasing graph neural networks via learning disentangled causal substructure. Advances in Neural Information Processing Systems, 35:24934–24946, 2022.

[31] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618, 2022.

[32] Lijian Gao, Qirong Mao, Ming Dong, Yu Jing, and Ratna Chinnam. On learning disentangled representation for acoustic event detection. In Proceedings of the 27th ACM International Conference on Multimedia, pages 2006–2014, 2019.

[33] Yifan Gao, Jinpeng Lin, Min Zhou, Chuanbin Liu, Hongtao Xie, Tiezheng Ge, and Yuning Jiang. Textpainter: Multimodal text image generation with visual-harmony and text-comprehension for poster design. In Proceedings of the 31st ACM International Conference on Multimedia, pages 7236–7246, 2023.

[34] Partha Ghosh, Mehdi SM Sajjadi, Antonio Vergari, Michael Black, and Bernhard Schölkopf. From variational to deterministic autoencoders. arXiv preprint arXiv:1903.12436, 2019.

[35] Muhammad Waleed Gondal, Manuel Wuthrich, Djordje Miladinovic, Francesco Locatello, Martin Breidt, Valentin Volchkov, Joel Akpo, Olivier Bachem, Bernhard Schölkopf, and Stefan Bauer. On the transfer of inductive bias from simulation to the real world: a new disentanglement dataset. Advances in Neural Information Processing Systems, 32, 2019.

[36] Abel Gonzalez-Garcia, Joost Van De Weijer, and Yoshua Bengio. Image-to-image translation for cross-domain disentanglement. Advances in neural information processing systems, 31, 2018.

[37] Ian Goodfellow, Honglak Lee, Quoc Le, Andrew Saxe, and Andrew Ng. Measuring invariances in deep networks. Advances in neural information processing systems, 22, 2009.

[38] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. Advances in neural information processing systems, 27, 2014.

[39] Luyi Han, Yuanyuan Lyu, Cheng Peng, and S Kevin Zhou. GAN-based disentanglement learning for chest x-ray rib suppression. Medical Image Analysis, 77:102369, 2022.

[40] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. Advances in neural information processing systems, 33:9841–9850, 2020.

[41] Shuting He, Henghui Ding, and Wei Jiang. Primitive generation and semantic-related alignment for universal zero-shot segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 11238–11247, 2023.

[42] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv:2208.01626, 2022.

[43] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. arXiv preprint arXiv:2104.08718, 2021.

[44] Irina Higgins, David Amos, David Pfau, Sebastien Racaniere, Loic Matthey, Danilo Rezende, and Alexander Lerchner. Towards a definition of disentangled representations. arXiv preprint arXiv:1812.02230, 2018.

[45] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In International conference on learning representations, 2016.

[46] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. Advances in neural information processing systems, 33:6840–6851, 2020.

[47] Qiyang Hu, Attila Szabó, Tiziano Portenier, Paolo Favaro, and Matthias Zwicker. Disentangling factors of variation by mixing them. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3399–3407, 2018.

[48] Nisha Huang, Weiming Dong, Yuxin Zhang, Fan Tang, Ronghui Li, Chongyang Ma, Xiu Li, and Changsheng Xu. CreativeSynth: Creative blending and synthesis of visual arts based on multimodal diffusion. arXiv preprint arXiv:2401.14066, 2024.

[49] Yi Huang, Jiancheng Huang, Yifan Liu, Mingfu Yan, Jiaxi Lv, Jianzhuang Liu, Wei Xiong, He Zhang, Shifeng Chen, and Liangliang Cao. Diffusion model-based image editing: A survey. arXiv preprint arXiv:2402.17525, 2024.

[50] Inbar Huberman-Spiegelglas, Vladimir Kulikov, and Tomer Michaeli. An edit friendly ddpm noise space: Inversion and manipulations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 12469–12478, 2024.

[51] Aapo Hyvarinen and Hiroshi Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ica. Advances in neural information processing systems, 29, 2016.

[52] Aapo Hyvärinen and Petteri Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. Neural networks, 12(3):429–439, 1999.

[53] Aapo Hyvarinen, Hiroaki Sasaki, and Richard Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. In The 22nd International Conference on Artificial Intelligence and Statistics, pages 859–868. PMLR, 2019.

[54] Yeonwoo Jeong and Hyun Oh Song. Learning discrete and continuous factors of data via alternating disentanglement. In International Conference on Machine Learning, pages 3091–3099. PMLR, 2019.

[55] Ananya Harsh Jha, Saket Anand, Maneesh Singh, and VS Rao Veeravasarapu. Disentangling factors of variation with cycle-consistent variational auto-encoders. In Proceedings of the European Conference on Computer Vision (ECCV), pages 805–820, 2018.

[56] Mengxi Jia, Xinhua Cheng, Shijian Lu, and Jian Zhang. Learning disentangled representation implicitly via transformer for occluded person re-identification. IEEE Transactions on Multimedia, 25:1294–1305, 2022.

[57] Xuan Ju, Ailing Zeng, Yuxuan Bian, Shaoteng Liu, and Qiang Xu. Pnp inversion: Boosting diffusion-based editing with 3 lines of code. In The Twelfth International Conference on Learning Representations, 2024.

[58] Christian Jutten and Jeanny Herault. Blind separation of sources, part i: An adaptive algorithm based on neuromimetic architecture. Signal processing, 24(1):1–10, 1991.

[59] Christian Jutten, Juha Karhunen, et al. Advances in nonlinear blind source separation. In Proc. of the 4th Int. Symp. on Independent Component Analysis and Blind Signal Separation (ICA2003), volume 2003, pages 245–256, 2003.

[60] Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In International Conference on Artificial Intelligence and Statistics, pages 2207–2217. PMLR, 2020.

[61] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 2426–2435, June 2022.

[62] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In International Conference on Machine Learning, pages 2649–2658. PMLR, 2018.

[63] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114, 2013.

[64] David Klindt, Lukas Schott, Yash Sharma, Ivan Ustyuzhaninov, Wieland Brendel, Matthias Bethge, and Dylan Paiton. Towards nonlinear disentanglement in natural data with temporal sparse coding. arXiv preprint arXiv:2007.10930, 2020.

[65] Hyung-Kwon Ko, Gwanmo Park, Hyeon Jeon, Jaemin Jo, Juho Kim, and Jinwook Seo. Large-scale text-to-image generation models for visual artists' creative works. In Proceedings of the 28th international conference on intelligent user interfaces, pages 919–933, 2023.

[66] Eunseo Koh, Sangeek Hyun, MinKyu Lee, Jiwoo Chung, and Jae-Pil Heo. Structure-preserving text-based editing for few-step diffusion models.

[67] Mark A Kramer. Nonlinear principal component analysis using autoassociative neural networks. AIChE journal, 37(2):233–243, 1991.

[68] Tejas D Kulkarni, William F Whitney, Pushmeet Kohli, and Josh Tenenbaum. Deep convolutional inverse graphics network. Advances in neural information processing systems, 28, 2015.

[69] Abhishek Kumar and Ben Poole. On implicit regularization in β-vaes. International Conference on Machine Learning, 2020.

[70] Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. Variational inference of disentangled latent concepts from unlabeled observations. arXiv preprint arXiv:1711.00848, 2017.

[71] Gihyun Kwon and Jong Chul Ye. Diffusion-based image translation using disentangled style and content representation. arXiv preprint arXiv:2209.15264, 2022.

[72] Sébastien Lachapelle, Tristan Deleu, Divyat Mahajan, Ioannis Mitliagkas, Yoshua Bengio, Simon Lacoste-Julien, and Quentin Bertrand. Synergies between disentanglement and sparsity: Generalization and identifiability in multi-task learning. In International Conference on Machine Learning, pages 18171–18206. PMLR, 2023.

[73] John Lafferty, Andrew McCallum, Fernando Pereira, et al. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Icml, volume 1, page 3. Williamstown, MA, 2001.

[74] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. Behavioral and brain sciences, 40, 2017.

[75] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. nature, 521(7553):436–444, 2015.

[76] Karel Lenc and Andrea Vedaldi. Understanding image representations by measuring their equivariance and equivalence. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 991–999, 2015.

[77] Anat Levin, Dani Lischinski, and Yair Weiss. A closed-form solution to natural image matting. IEEE transactions on pattern analysis and machine intelligence, 30(2):228–242, 2007.

[78] Huafeng Li, Kaixiong Xu, Jinxing Li, and Zhengtao Yu. Dual-stream reciprocal disentanglement learning for domain adaptation person re-identification. Knowledge-Based Systems, 251:109315, 2022.

[79] Zhiyuan Li, Jaideep Vitthal Murkute, Prashnna Kumar Gyawali, and Linwei Wang. Progressive learning and disentanglement of hierarchical representations. arXiv preprint arXiv:2002.10549, 2020.

[80] Yiyi Liao, Katja Schwarz, Lars Mescheder, and Andreas Geiger. Towards unsupervised learning of generative models for 3d controllable image synthesis. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 5871–5880, 2020.

[81] Jungbin Lim, Jihwan Kim, Yonghyeon Lee, Cheongjae Jang, and Frank C Park. Graph geometry-preserving autoencoders. In Forty-first International Conference on Machine Learning, 2024.

[82] Bingyan Liu, Chengyu Wang, Tingfeng Cao, Kui Jia, and Jun Huang. Towards understanding cross and self-attention in stable diffusion for text-guided image editing. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 7817–7826, 2024.

[83] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In international conference on machine learning, pages 4114–4124. PMLR, 2019.

[84] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. A sober look at the unsupervised learning of disentangled representations and their evaluation. arXiv preprint arXiv:2010.14766, 2020.

[85] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial autoencoders. arXiv preprint arXiv:1511.05644, 2015.

[86] Joanna Materzyńska, Antonio Torralba, and David Bau. Disentangling visual and written concepts in clip. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 16410–16419, 2022.

[87] Emile Mathieu, Tom Rainforth, Siddharth Narayanaswamy, and Yee Whye Teh. Disentangling disentanglement. arXiv preprint arXiv:1812.02833, 2018.

[88] Michael F Mathieu, Junbo Jake Zhao, Junbo Zhao, Aditya Ramesh, Pablo Sprechmann, and Yann LeCun. Disentangling factors of variation in deep representation using adversarial training. Advances in neural information processing systems, 29, 2016.

[89] Lars Mescheder, Sebastian Nowozin, and Andreas Geiger. Adversarial variational bayes: Unifying variational autoencoders and generative adversarial networks. In International conference on machine learning, pages 2391–2400. PMLR, 2017.

[90] Graziano Mita, Maurizio Filippone, and Pietro Michiardi. An identifiable double vae for disentangled representations. In International Conference on Machine Learning, pages 7769–7779. PMLR, 2021.

[91] Abdullah al Nomaan Nafi, Md Alamgir Hossain, Rakib Hossain Rifat, Md Mahabub Uz Zaman, Md Manjurul Ahsan, and Shivakumar Raman. Diffusion-based approaches in medical image generation and analysis. arXiv preprint arXiv:2412.16860, 2024.

[92] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741, 2021.

[93] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In International conference on machine learning, pages 8162–8171. PMLR, 2021.

[94] Jie Nie, Ting Zhang, Tianbao Li, Shusong Yu, Xuanya Li, and Zhiqiang Wei. Image-based 3D model retrieval via disentangled feature learning and enhanced semantic alignment. Information Processing & Management, 60(2):103159, 2023.

[95] Jiahao Pang and Gene Cheung. Graph laplacian regularization for image denoising: Analysis in the continuous domain. IEEE Transactions on Image Processing, 26(4):1770–1785, 2017.

[96] Jungin Park, Jiyoung Lee, Sangryul Jeon, Seungryong Kim, and Kwanghoon Sohn. Graph regularization network with semantic affinity for weakly-supervised temporal action localization. In 2019 IEEE International conference on image processing (ICIP), pages 3701–3705. IEEE, 2019.

[97] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In ACM SIGGRAPH 2023 Conference Proceedings, pages 1–11, 2023.

[98] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3D face model for pose and illumination invariant face recognition. In 2009 sixth IEEE international conference on advanced video and signal based surveillance, pages 296–301. Ieee, 2009.

[99] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Elements of causal inference: foundations and learning algorithms. The MIT Press, 2017.

[100] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952, 2023.

[101] Yifang Qin, Yifan Wang, Fang Sun, Wei Ju, Xuyang Hou, Zhe Wang, Jia Cheng, Jun Lei, and Ming Zhang. DisenPOI: Disentangling sequential and geographical influence for point-of-interest recommendation. In Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining, pages 508–516, 2023.

[102] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In International conference on machine learning, pages 8748–8763. PMLR, 2021.

[103] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, Proceedings of the 38th International Conference on Machine Learning, volume 139 of Proceedings of Machine Learning Research, pages 8748–8763. PMLR, 18–24 Jul 2021.

[104] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125, 1(2):3, 2022.

[105] Hareesh Ravi, Sachin Kelkar, Midhun Harikumar, and Ajinkya Kale. Preditor: Text guided image editing with diffusion prior. arXiv preprint arXiv:2302.07979, 2023.

[106] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In International conference on machine learning, pages 1278–1286. PMLR, 2014.

[107] Michal Rolinek, Dominik Zietlow, and Georg Martius. Variational autoencoders pursue pca directions (by accident). In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 12406–12415, 2019.

[108] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 10684–10695, 2022.

[109] Delian Ruan, Rongyun Mo, Yan Yan, Si Chen, Jing-Hao Xue, and Hanzi Wang. Adaptive deep disturbance-disentangled learning for facial expression recognition. International Journal of Computer Vision, 130(2):455–477, 2022.

[110] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 22500–22510, 2023.

[111] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. Advances in neural information processing systems, 35:36479–36494, 2022.

[112] Jürgen Schmidhuber. Learning factorial codes by predictability minimization. Neural computation, 4(6):863–879, 1992.

[113] Bernhard Schölkopf. Causality for machine learning. In Probabilistic and Causal Inference: The Works of Judea Pearl, pages 765–804. 2022.

[114] Matthias Scholz and Ricardo Vigário. Nonlinear PCA: a new hierarchical approach. In Esann, pages 439–444, 2002.

[115] Dazhong Shen, Guanglu Song, Zeyue Xue, Fu-Yun Wang, and Yu Liu. Rethinking the spatial inconsistency in classifier-free diffusion guidance. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9370–9379, 2024.

[116] Xinwei Shen, Furui Liu, Hanze Dong, Qing Lian, Zhitang Chen, and Tong Zhang. Weakly supervised disentangled generative causal representation learning. The Journal of Machine Learning Research, 23(1):10994–11048, 2022.

[117] Zhixin Shu, Ersin Yumer, Sunil Hadap, Kalyan Sunkavalli, Eli Shechtman, and Dimitris Samaras. Neural face editing with intrinsic image disentangling. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 5541–5550, 2017.

[118] Jiaming Song, Pratyusha Kalluri, Aditya Grover, Shengjia Zhao, and Stefano Ermon. Learning controllable fair representations. In The 22nd International Conference on Artificial Intelligence and Statistics, pages 2164–2173. PMLR, 2019.

[119] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502, 2020.

[120] Achint Soni, Meet Soni, and Sirisha Rambhatla. LOCATEdit: Graph laplacian optimized cross attention for localized text-guided image editing. arXiv preprint arXiv:2503.21541, 2025.

[121] Olga Sorkine, Daniel Cohen-Or, Yaron Lipman, Marc Alexa, Christian Rössl, and H-P Seidel. Laplacian surface editing. In Proceedings of the 2004 Eurographics/ACM SIGGRAPH symposium on Geometry processing, pages 175–184, 2004.

[122] Wolfgang Stammer, Marius Memmel, Patrick Schramowski, and Kristian Kersting. Interactive disentanglement: Learning concepts by interacting with their prototype representations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10317–10328, 2022.

[123] David Stap, Maurits Bleeker, Sarah Ibrahimi, and Maartje Ter Hoeve. Conditional image generation and manipulation for user-specified content. arXiv preprint arXiv:2005.04909, 2020.

[124] Zihan Su, Junhao Zhuang, and Chun Yuan. TextureDiffusion: Target prompt disentangled editing for various texture transfer. arXiv preprint arXiv:2409.09610, 2024.

[125] Yanan Sun, Yanchen Liu, Yinhao Tang, Wenjie Pei, and Kai Chen. Anycontrol: create your artwork with versatile control on text-to-image generation. In European Conference on Computer Vision, pages 92–109. Springer, 2024.

[126] Raphael Suter, Djordje Miladinovic, Bernhard Schölkopf, and Stefan Bauer. Robustly disentangled causal mechanisms: Validating deep representations for interventional robustness. In International Conference on Machine Learning, pages 6056–6065. PMLR, 2019.

[127] Attila Szabó, Qiyang Hu, Tiziano Portenier, Matthias Zwicker, and Paolo Favaro. Challenges in disentangling independent factors of variation. arXiv preprint arXiv:1711.02245, 2017.

[128] Luan Tran, Xi Yin, and Xiaoming Liu. Disentangled representation learning gan for pose-invariant face recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1415–1424, 2017.

[129] Frederik Träuble, Elliot Creager, Niki Kilbertus, Francesco Locatello, Andrea Dittadi, Anirudh Goyal, Bernhard Schölkopf, and Stefan Bauer. On disentangled representations learned from correlated data. In International conference on machine learning, pages 10401–10412. PMLR, 2021.

[130] Michael Tschannen, Olivier Bachem, and Mario Lucic. Recent advances in autoencoder-based representation learning. arXiv preprint arXiv:1812.05069, 2018.

[131] Narek Tumanyan, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Splicing vit features for semantic appearance transfer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10748–10757, 2022.

[132] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 1921–1930, 2023.

[133] Shinji Uchinoura and Takio Kurita. Graph laplacian regularization based on the differences of neighboring pixels for conditional convolutions for instance segmentation. In 2022 26th International Conference on Pattern Recognition (ICPR), pages 3611–3617. IEEE, 2022.

[134] Andrey Voynov and Artem Babenko. Unsupervised discovery of interpretable directions in the gan latent space. In International conference on machine learning, pages 9786–9796. PMLR, 2020.

[135] Duomin Wang, Yu Deng, Zixin Yin, Heung-Yeung Shum, and Baoyuan Wang. Progressive disentangled representation learning for fine-grained controllable talking head synthesis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 17979–17989, 2023.

[136] Haoshen Wang, Zhentao Liu, Kaicong Sun, Xiaodong Wang, Dinggang Shen, and Zhiming Cui. 3D MedDiffusion: A 3D Medical Diffusion Model for Controllable and High-quality Medical Image Generation. arXiv preprint arXiv:2412.13059, 2024.

[137] Kejie Wang, Xuemeng Song, Meng Liu, Jin Yuan, and Weili Guan. Vision-guided and mask-enhanced adaptive denoising for prompt-based image editing. arXiv preprint arXiv:2410.10496, 2024.

[138] Qian Wang, Biao Zhang, Michael Birsak, and Peter Wonka. Mdp: A generalized framework for text-guided image editing by manipulating the diffusion path. arXiv preprint arXiv:2303.16765, 2023.

[139] Zhizhong Wang, Lei Zhao, and Wei Xing. Stylediffusion: Controllable disentangled style transfer via diffusion models. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 7677–7689, October 2023.

[140] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing, 13(4):600–612, 2004.

[141] Qiucheng Wu, Yujian Liu, Handong Zhao, Ajinkya Kale, Trung Bui, Tong Yu, Zhe Lin, Yang Zhang, and Shiyu Chang. Uncovering the disentanglement capability in text-to-image diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 1900–1910, 2023.

[142] Shishi Xiao, Suizi Huang, Yue Lin, Yilin Ye, and Wei Zeng. Let the chart spark: Embedding semantic context into chart with text-to-image generative model. IEEE Transactions on Visualization and Computer Graphics, 2023.

[143] Jinheng Xie, Yuexiang Li, Yawen Huang, Haozhe Liu, Wentian Zhang, Yefeng Zheng, and Mike Zheng Shou. Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 7452–7461, 2023.

[144] Qingsong Xie, Yuexiang Li, Nanjun He, Munan Ning, Kai Ma, Guoxing Wang, Yong Lian, and Yefeng Zheng. Unsupervised domain adaptation for medical image segmentation by disentanglement learning and self-training. IEEE Transactions on Medical Imaging, 2022.

[145] Jie Xu, Yazhou Ren, Huayi Tang, Xiaorong Pu, Xiaofeng Zhu, Ming Zeng, and Lifang He. Multi-VAE: Learning disentangled view-common and view-peculiar visual representations for multi-view clustering. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 9234–9243, 2021.

[146] Sihan Xu, Yidong Huang, Jiayi Pan, Ziqiao Ma, and Joyce Chai. Inversion-free image editing with language-guided diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9452–9461, 2024.

[147] Zeyue Xue, Guanglu Song, Qiushan Guo, Boxiao Liu, Zhuofan Zong, Yu Liu, and Ping Luo. Raphael: Text-to-image generation via large mixture of diffusion paths. Advances in Neural Information Processing Systems, 36:41693–41706, 2023.

[148] Fei Yang, Shiqi Yang, Muhammad Atif Butt, Joost van de Weijer, et al. Dynamic prompt learning: Addressing cross-attention leakage for text-based image editing. Advances in Neural Information Processing Systems, 36:26291–26303, 2023.

[149] Tao Yang, Yuwang Wang, Yan Lv, and Nanning Zh. DisDiff: Unsupervised disentanglement of diffusion probabilistic models. arXiv preprint arXiv:2301.13721, 2023.

[150] Weiran Yao, Guangyi Chen, and Kun Zhang. Temporally disentangled representation learning. Advances in Neural Information Processing Systems, 35:26492–26503, 2022.

[151] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721, 2023.

[152] Ilker Yildirim, Mario Belledonne, Winrich Freiwald, and Josh Tenenbaum. Efficient inverse graphics in biological face processing. Science advances, 6(10):eaax5979, 2020.

[153] Jin Zeng, Gene Cheung, Michael Ng, Jiahao Pang, and Cheng Yang. 3d point cloud denoising using graph laplacian regularization of a low dimensional manifold model. IEEE Transactions on Image Processing, 29:3474–3489, 2019.

[154] Jin Zeng, Jiahao Pang, Wenxiu Sun, and Gene Cheung. Deep graph laplacian regularization for robust denoising of real images. In Proceedings of the ieee/cvf conference on computer vision and pattern recognition workshops, pages 0–0, 2019.

[155] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 586–595, 2018.

[156] Xiaohan Zhang, Xingyu Li, Waqas Sultani, Yi Zhou, and Safwan Wshah. Cross-view geo-localization via learning disentangled geometric layout correspondence. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 37, pages 3480–3488, 2023.

[157] Zijian Zhang, Zhou Zhao, and Zhijie Lin. Unsupervised representation learning from pre-trained diffusion probabilistic models. Advances in Neural Information Processing Systems, 35:22117–22130, 2022.

[158] Xinqi Zhu, Chang Xu, and Dacheng Tao. Commutative lie group vae for disentanglement learning. In International Conference on Machine Learning, pages 12924–12934. PMLR, 2021.

[159] Yizhe Zhu, Martin Renqiang Min, Asim Kadav, and Hans Peter Graf. S3vae: Self-supervised sequential vae for representation disentanglement and data generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6538–6547, 2020.

[160] Dominik Zietlow, Michal Rolinek, and Georg Martius. Demystifying inductive biases for (beta-) vae based architectures. In International Conference on Machine Learning, pages 12945–12954. PMLR, 2021.

APPENDICES

.1 Broader Impact

Our work advances the precision of text-guided image editing by ensuring that modifications are both spatially consistent and semantically faithful. This improvement has the potential to benefit a wide range of applications—from enhancing creative workflows in digital art and advertising to supporting critical tasks in medical imaging and scientific visualization—by reducing the need for extensive manual post-processing. At the same time, the increased reliability of automated editing tools underscores the importance of establishing robust ethical guidelines for their use, particularly in contexts where the authenticity of visual information is paramount. By delivering a method that better preserves the structural integrity of the source images, our approach paves the way for more trustworthy and accessible image editing solutions that can democratize creative technologies and support various high-stakes applications.

.2 Proof of Lemma 1

Proof. To prove that $\mathbf{L}$ is PSD, we must show that for any $\mathbf{x} \in \mathbb{R}^n$, the quadratic form $\mathbf{x}^\top \mathbf{L}\mathbf{x}$ is nonnegative:

$$\mathbf{x}^\top \mathbf{L}\mathbf{x} = \mathbf{x}^\top (\mathbf{D} - \mathbf{S}_{\text{sym}})\mathbf{x}.$$

Expanding this expression, we have:

$$\mathbf{x}^\top \mathbf{L}\mathbf{x} = \mathbf{x}^\top \mathbf{D}\mathbf{x} - \mathbf{x}^\top \mathbf{S}_{\text{sym}}\mathbf{x}.$$

The degree matrix $\mathbf{D}$ is diagonal, with entries $\mathbf{D}(i,i) = \sum_{j=1}^{n} \mathbf{S}_{\text{sym}}(i,j)$. Therefore:

$$\mathbf{x}^\top \mathbf{D}\mathbf{x} = \sum_{i=1}^{n} \mathbf{D}(i,i) x_i^2 = \sum_{i=1}^{n} \left(\sum_{j=1}^{n} \mathbf{S}_{\text{sym}}(i,j) \right) x_i^2.$$

The second term, $\mathbf{x}^\top \mathbf{S}_{\text{sym}}\mathbf{x}$, is given by:

$$\mathbf{x}^\top \mathbf{S}_{\text{sym}}\mathbf{x} = \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbf{S}_{\text{sym}}(i,j) x_i x_j.$$

Substituting these into the quadratic form, we get:

$$\mathbf{x}^\top \mathbf{L}\mathbf{x} = \sum_{i=1}^{n} \left(\sum_{j=1}^{n} \mathbf{S}_{\text{sym}}(i,j) x_i^2 \right) - \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbf{S}_{\text{sym}}(i,j) x_i x_j.$$

Reorganizing terms:

$$\mathbf{x}^\top \mathbf{L}\mathbf{x} = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{S}_{\text{sym}}(i,j)\left(x_i^2 + x_j^2 - 2x_i x_j\right).$$

This simplifies to:

$$\mathbf{x}^\top \mathbf{L}\mathbf{x} = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{S}_{\text{sym}}(i,j)(x_i - x_j)^2.$$

Since $\mathbf{S}_{\text{sym}}(i,j) \geq 0$ (by definition of the symmetrized self-attention matrix) and $(x_i - x_j)^2 \geq 0$, every term in the summation is nonnegative. Therefore:

$$\mathbf{x}^\top \mathbf{L}\mathbf{x} \geq 0 \quad \forall \mathbf{x} \in \mathbb{R}^n.$$

Thus, $\mathbf{L}$ is positive semidefinite. □

.3 Proof of Theorem 1

.3.1 Optimization Problem

We consider the following optimization problem:

$$\min_{x \in \mathbb{R}^{R^2}} J(x), \tag{1}$$

where the objective function is defined as

$$J(x) = (x - x^{(0)})^\top \mathbf{\Lambda}(x - x^{(0)}) + \lambda\, x^\top L\, x.$$

Here, the fidelity term $(x - x^{(0)})^\top \mathbf{\Lambda}(x - x^{(0)})$ penalizes deviations from the initial mask $x^{(0)}$ with stronger penalties in regions of higher confidence (as encoded by the diagonal weight matrix $\mathbf{\Lambda}$). The smoothness term $\lambda\, x^\top L\, x$ promotes a spatially coherent solution by enforcing that the mask varies smoothly across similar patches, as determined by the self-attention structure. The hyperparameter $\lambda > 0$ balances the trade-off between fidelity and smoothness.

.3.2 Existence and Uniqueness of the Solution

To obtain the refined mask, we solve the minimization problem in Equation (1). The first term is strictly convex since $\mathbf{\Lambda}$ is positive definite, and the second term is convex because L is positive semidefinite. Thus, the overall objective $J(x)$ is strictly convex and has a unique minimizer.

Taking the gradient with respect to x yields:

$$\nabla J(x) = 2\,\mathbf{\Lambda}(x - x^{(0)}) + 2\lambda\, L\, x. \tag{2}$$

Setting $\nabla J(x) = 0$ gives:

$$\mathbf{\Lambda}(x - x^{(0)}) + \lambda\, L\, x = 0. \tag{3}$$

Rearranging, we obtain:

$$(\mathbf{\Lambda} + \lambda\, L)\, x = \mathbf{\Lambda}\, x^{(0)}. \tag{4}$$

Since $\mathbf{\Lambda} + \lambda\, L$ is positive definite, it is invertible, and the unique solution is

$$x^* = (\mathbf{\Lambda} + \lambda\, L)^{-1}\, \mathbf{\Lambda}\, x^{(0)}.$$

The positive semidefiniteness of L ensures the convexity of the regularization term, thereby guaranteeing the existence and uniqueness of the solution.

.3.3 Additional Qualitative Results

This section presents qualitative results for refined masks achieved through graph Laplacian regularization and compares the editing outcomes with existing image editing methods. They can be view in Figure 1 and Table 1.

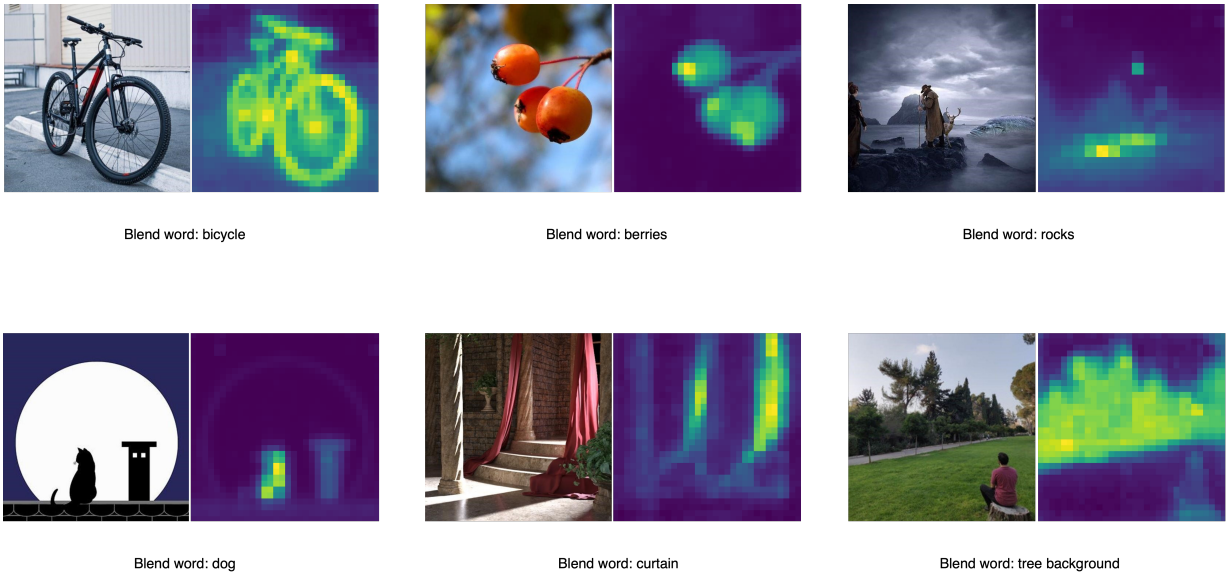


Figure 1: Refined masks after Graph Laplacian Regularization

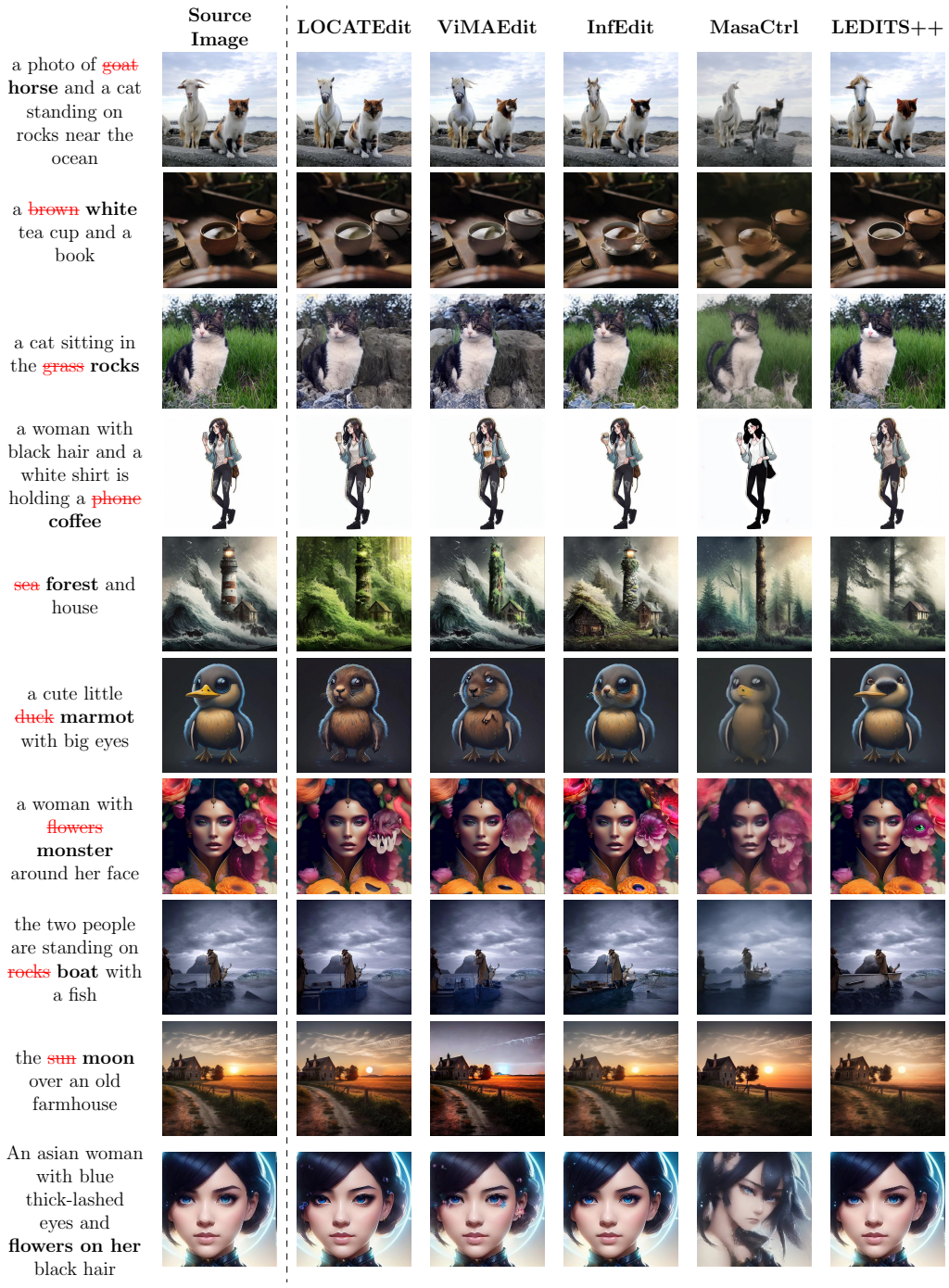


Table 1: Additional qualitative results on PIE-Bench