

Latent region-wise image composition using Multimodal Large Language Model

by

Tan Chen

Thesis submitted to the University of Ottawa
in partial fulfillment of the requirements for the degree of
Master of Computer Science

School of Electrical Engineering and Computer Science
Faculty of Engineering
University of Ottawa

© Tan Chen, Ottawa, Canada, 2026

Declaration of Authorship

I hereby certify that this thesis is entirely my own original work except where otherwise indicated. I am aware of the University of Ottawa regulations concerning plagiarism, including those regarding consequent disciplinary actions. Any use of the works of any other author, in any form, is properly acknowledged at their point of use.

Abstract

Recent advances in text-to-image generative models, particularly latent diffusion models and more recent rectified-flow-based variants [12,41], have substantially improved the ability to synthesize images from textual prompts. Notably, models such as IP-Adapter [59] provide effective image-prompt conditioning and can capture coarse global appearance cues from reference images. However, these models typically condition generation at the image level, which may limit their ability to faithfully reproduce fine-grained regional details or spatially localized features from the reference. As a result, they may struggle to preserve specific object identities or detailed visual elements from individual reference images, especially when multiple diverse regions or entities are involved.

To address this limitation, we propose a novel training-free plugin that adopts a segmented binding and generation scheme. Specifically, we divide the input prompt and image space into semantically meaningful subregions and associate each region with its own reference image. A Multimodal Large Language Model (MLLM) is used as a planner to parse the prompt into region-wise descriptions, while each region is conditioned separately during the diffusion process. This enables more targeted control over specific elements in each region using dedicated reference images, thereby improving the alignment between the generated image and the intended local semantics.

In our experiments, the proposed approach achieves stronger visual and semantic consistency than baseline methods such as SDXL [35], EasyRef [63], Gen-4 by Runway [44], and DALL·E 2 [38], particularly in multi-entity and identity-preserving scenarios. Our method provides a flexible and interpretable framework for region-aware, reference-driven image synthesis.

Acknowledgements

I hereby certify that this thesis is entirely my own original work except where otherwise indicated. I am aware of the University's regulations concerning plagiarism, including those concerning consequent disciplinary actions. Any use of the works of any other author, in any form, is properly acknowledged at their point of use.

Table of Contents

List of Tables	x
List of Figures	xi
1 Introduction	1
1.1 Motivation	1
1.2 Problem Overview	3
1.2.1 Ambiguity in Multi-Reference Feature Fusion	3
1.2.2 Entanglement Between Style and Content	3
1.2.3 Spatial Allocation and Region-Wise Control Deficiency	4
1.2.4 Attention Conflicts and Overwriting Effects	4
1.3 Research Contributions	5
1.3.1 Region-Aware Multi-Reference Fusion Framework for Coherent Image Synthesis	6
1.3.2 Human-Aligned Benchmark for Multi-Reference Generation Evaluation via LLM-Based Scoring and User Studies	7

1.3.3	Additional Contributions and Broader Implications	8
1.4	Thesis structure	8
2	literature review	10
2.1	Basic models	10
2.1.1	Variational Autoencoders (VAEs)	10
2.1.2	Autoregressive Models	12
2.2	Diffusion models	14
2.2.1	Diffusion Probabilistic Models	14
2.2.2	Efficiency and Latent-space Improvements	17
2.2.3	Conditioning Mechanisms in Diffusion Models	18
2.3	Reference Image Generation	19
2.3.1	Single-reference Conditioning Methods	19
2.3.2	Regional and Spatially-aware Conditioning	22
2.4	Evaluation Metrics	24
2.4.1	Perceptual and Semantic Similarity Metrics	24
2.4.2	Aesthetic and Human-aligned Evaluation	27
2.5	Summary	28
3	Methodology	30
3.0.1	LLM-Guided Regional Prompt Decomposition	30

3.0.2	Region-Aware Complementary Generation via Latent Fusion	31
3.1	Preliminaries	33
3.2	Prompt Recaption via Compositional Language Understanding	36
3.2.1	Why Flat Prompts Fail	37
3.2.2	Prompt Recaption Task Definition	39
3.2.3	Design Benefits	40
3.3	Region Planning via Spatial-Aware Layout Reasoning	40
3.3.1	Task Definition	40
3.3.2	LLM-Based Layout Inference Pipeline	41
3.3.3	Serialization for Diffusion Control	41
3.3.4	Example: Layout Representation	42
3.3.5	Advantages Over Prior Methods	43
3.4	Region-wise Conditioning and Visual Injection	44
3.4.1	Stable Diffusion	44
3.4.2	Region-wise Conditioning Fusion	47
3.4.3	Enhancing Control with Visual Reference Injection	49
3.4.4	Benefits and Generalization	51
3.4.5	Illustrative Example	52
3.5	Summary	52

4	Experiment	55
4.1	Experiment Setup	55
4.1.1	Evaluation Dataset and Sample Selection	56
4.1.2	T2I-CompBench Evaluation	57
4.1.3	Rethinking Evaluation: Beyond Traditional Metrics	58
4.1.4	Qualitative Success Criteria for Region-Aware Multi-Reference Fusion	60
4.1.5	GPT-based Evaluation	62
4.2	Experiment Result of a human-aligned benchmark powered by advanced multi-modal GPT models	65
4.3	Experiment Result of real human evaluation	69
4.4	Ablation Study	74
4.4.1	Condition Scale	74
4.4.2	Ablation on Recaption and Regional Diffusion	75
5	Conclusion and future work	79
5.1	Conclusion	79
5.2	Future work	81
	References	84
	APPENDICES	93
	Appendix 1: Human Evaluation Protocol	93

.1	Appendices 2 for GPT evaluation	95
	Appendix 2: GPT-Based Evaluation Protocol	95

List of Tables

2.1	Representative Generative Models Across Architectures	24
2.2	Summary of Evaluation Metrics in Text-to-Image Generation	28
3.1	Example of LLM-based Regional Decomposition	31
4.1	Attribute Binding and Object Relationship Benchmark	58
4.2	Compositional Generation Benchmark	68
4.3	Human preference results comparing LCCR against baseline models. Each cell reports the number of participants who preferred LCCR versus the corresponding baseline, with the percentage of LCCR preference shown in parentheses. Each comparison contains 105 valid responses.	73
4.4	Ablation settings for analyzing the effects of prompt recaption and regional diffusion.	76

List of Figures

1.1	Comparison of generation results with and without regional planning. . . .	5
2.1	The architecture of Vector Quantised Variational AutoEncod	12
2.2	The architecture of Generative Pre-trained Transformer (GPT)	14
2.3	A overall architecture of Stable Diffusion model.	21
3.1	Failure cases of flat prompts.	38
3.2	Illustration of the LLM-based recaptioning process.	43
3.3	Comparison of Stable Diffusion	46
3.4	Illustration of regional attention-guided latent generation.	48
3.5	Integration of IP-Adapter with Regional Planning	50
4.1	Detailed comparison with other multi-reference generation models.	67
4.2	Ablation study	74
4.3	Qualitative ablation results	78

Chapter 1

Introduction

1.1 Motivation

Recent diffusion models have substantially advanced image synthesis in realism, diversity, and controllability, with latent diffusion and its successors (e.g., SDXL) becoming de facto backbones for high-fidelity generation [35, 41]. Building on these foundations, a rich ecosystem for style transfer and editing has emerged—ranging from style inversion and disentangled style control to instruction-driven editing—demonstrating impressive single-reference performance in preserving global aesthetics or subject identity [2, 14, 27, 55]. Adapter-based approaches further reduce deployment cost by injecting visual priors without full model fine-tuning, broadening applicability in practical pipelines [52, 59].

However, region-aware multi-reference image generation—where a model must jointly incorporate several visual references (e.g., identity, texture, palette, layout, material, or brand style), assign them to semantically appropriate regions, and fuse them into a single coherent output—remains underexplored and markedly fragile in current systems. In real

design and content workflows (e.g., product ideation, creative direction, marketing composites, multi-character scenes), creators frequently require compositional fusion: structure from one reference, color scheme from another, and material/texture from a third, while still adhering to a textual prompt. Empirically, most solutions struggle with (i) semantic alignment between references and prompt (prompt drift vs. reference fidelity), (ii) style/identity interference across references (one reference dominates or contaminates the others), (iii) entanglement of style and content that impedes targeted attribute transfer, and (iv) spatial incoherence when attributes should localize to specific regions. While fine-tuning methods such as DreamBooth personalize a model to a single subject with a handful of images [42], scaling this paradigm to multiple heterogeneous references is computationally expensive and often yields overfitting or mode locking. Adapter-based routes (e.g., IP-Adapter) offer efficient inference-time conditioning, yet balancing multi-source constraints against the prompt remains challenging, especially without curated data pairing the same subject/style with diverse textual descriptions [30, 59].

Recent advances in controllable or region-aware generation partially alleviate these issues by offering explicit handles for structure or spatial semantics [26, 52]. Nevertheless, most pipelines still assume single reference dominance or lack principled mechanisms to compose multiple references with fine-grained where-to-apply and how-much-to-apply control. This gap is acute for prompts that demand multi-entity, multi-style, or multi-attribute assembly—precisely the scenarios common in professional creation. Consequently, there is a pressing need for frameworks that (1) disentangle style and content cues across references, (2) allocate them to target regions or layers with calibrated influence, and (3) harmonize them with textual intent, yielding coherent, faithful, and controllable multi-reference synthesis [2, 26, 59]. Addressing this need would translate directly into improved creative workflows—reducing trial-and-error cycles, enhancing reproducibility, and enabling scal-

able, prompt-aligned fusion of heterogeneous visual sources.

1.2 Problem Overview

In this thesis, we focus on region-aware multi-reference text-to-image generation, where a model is given a global textual prompt and multiple reference images, and is expected to bind different reference-specific visual cues to semantically corresponding regions while preserving global coherence and prompt alignment. While models such as Stable Diffusion and SDXL [35, 41] have demonstrated strong realism and fidelity in text-to-image generation, they are primarily designed for text-conditioned generation rather than explicit multi-reference composition.

1.2.1 Ambiguity in Multi-Reference Feature Fusion

Many existing tuning-free approaches rely on simple operations such as feature averaging, linear interpolation, or straightforward concatenation of embeddings, which can dilute distinctive visual characteristics from each reference and lead to inconsistent results. [62, 63].

1.2.2 Entanglement Between Style and Content

A persistent challenge in reference-based diffusion models is the entanglement of style, content, and structural cues. Methods such as DreamBooth [42] are designed for subject-driven personalization rather than explicit disentanglement, which makes independent control and combination of different attributes difficult. Consequently, when multiple references from

distinct domains are provided, the model can exhibit semantic interference—identity attributes from one image may leak into regions governed by another, producing unrealistic blends or spatial inconsistency.

1.2.3 Spatial Allocation and Region-Wise Control Deficiency

Another major difficulty is controlling where and how much each reference influences generation. Standard diffusion models lack explicit spatial assignment mechanisms [8, 41], causing *attribute bleeding* in multi-entity compositions [6], while style transfer still struggles to preserve structure [2, 27].

1.2.4 Attention Conflicts and Overwriting Effects

Within the attention modules that mediate text and image conditions, multiple conditioning signals can compete during generation. This imbalance can lead to local artifacts, inconsistent textures, or the loss of subtle prompt semantics, especially in scenes requiring multi-entity alignment [6].

Summary of Challenges

In summary, the difficulty of multiple references arises from several interacting bottlenecks: limited modeling of relationships among reference embeddings [63], entanglement between style and content [15, 55], insufficient spatial selectivity in conditioning [6, 41], and unstable balancing of multiple conditions during attention-based generation [52, 59]. These challenges help explain why methods that work well for single-subject personalization or single-image conditioning [42, 56, 59] remain less reliable when multiple heterogeneous

references must be fused in a balanced and semantically aligned way. Addressing these deficiencies requires a framework that can (1) Dealing well with interactions among reference images [56,63], (2) appropriately allocate reference images to their respective regions, and (3) align them with textual intent during the denoising process [6,11]—motivating the approach proposed in this thesis.

1.3 Research Contributions



Figure 1.1: Comparison of generation results with and without regional planning. The upper row (with regional planning) demonstrates smoother scene composition and more natural fusion between the two reference images, while the lower row (without regional planning) shows weaker integration and less coherent blending across regions.

1.3.1 Region-Aware Multi-Reference Fusion Framework for Coherent Image Synthesis

Building upon recent advances in diffusion-based image generation, this research addresses the challenging problem of region-aware multi-reference fusion, where multiple reference images must be semantically assigned to appropriate regions and harmonized into a single prompt-aligned output. While existing frameworks such as Stable Diffusion, SDXL, and IP-Adapter [35, 41, 59] have demonstrated strong performance in text-to-image generation and image-prompt conditioning, they often struggle to balance diverse visual cues when conditioned on multiple references simultaneously [1, 63]. Existing conditioning pipelines also provide limited support for disentangling localized content, aligning reference cues with region-wise semantics, or modeling inter-reference interactions [6, 29, 63]. To address these limitations, this thesis contributes a region-aware multi-reference fusion framework that performs semantic-to-region allocation without requiring explicit user-provided masks or bounding boxes. [58]

1. **Semantic Confusion** by assigning explicit regional semantics through recaptioning;
2. **Style Contamination** by constraining reference influences to their designated regions;
3. **Spatial Inconsistency** by coordinating local generation control with global scene coherence.

Empirical experiments across diverse prompts and references show that the proposed framework achieves stronger reference preservation, reference-region consistency, and compositional coherence than baselines such as SDXL, EasyRef, and commercial image-generation systems evaluated in this thesis [35, 63]. The approach also improves controllability and interpretability in complex multi-entity scenes by making the correspondence between textual sub-prompts, reference cues, and regional generation more explicit.

1.3.2 Human-Aligned Benchmark for Multi-Reference Generation Evaluation via LLM-Based Scoring and User Studies

The second major contribution of this thesis is a comprehensive evaluation protocol for measuring the quality of multi-reference image synthesis using both model-based assessment and human judgment. Existing automated metrics, such as FID [17], LPIPS [61], and CLIP-based image-text alignment scores [16], quantify distributional similarity, perceptual similarity, or text-image compatibility, but they do not directly evaluate perceptual dimensions central to multi-reference generation, such as fusion harmony, aesthetic coherence, and cross-reference semantic consistency. Conversely, human evaluations are often more informative, but they are labor-intensive, subjective, and difficult to reproduce.

To bridge this gap, we design a two-stage hybrid evaluation pipeline that combines a multimodal large language model with structured human feedback. In the first stage, the model compares the generated image against two or more references along two perceptual axes: Concept Preservation and Prompt Following. By constraining the output to a discrete score range (0–4) and removing free-form explanations, the evaluation protocol improves consistency across samples. In the second stage, human participants evaluate the same image sets using the same criteria. Statistical analysis shows a strong correlation between model-generated and human-assigned scores, supporting the reliability of the proposed evaluation protocol.

This benchmark provides a standardized setting for comparing multi-reference generation models under aligned criteria. It enables systematic analysis of model behavior beyond single-reference personalization, especially in more challenging multi-condition scenarios. Furthermore, the benchmark exposes common weaknesses in current systems, such as unbalanced fusion and inconsistent local details, while also highlighting the strengths of our

proposed region-aware model in fusion quality and object consistency. By combining automated assessment with human judgment, this benchmark offers a practical evaluation framework for human-centered and interpretable analysis of multi-reference image generation.

1.3.3 Additional Contributions and Broader Implications

Beyond these two main innovations, this thesis introduces several auxiliary contributions that strengthen the overall framework and expand its applicability:

1. **Training-Free Deployment** The proposed architecture operates entirely without model fine-tuning, preserving the pre-trained weights of SDXL and IP-Adapter. This makes the system computationally lightweight, portable, and suitable for large-scale or interactive generation scenarios.
2. **Modular Regional Control** The region-aware design allows users to customize different spatial areas independently, supporting tasks such as composite portrait editing, product visualization, and cross-style scene composition.
3. **Explainability and Interpretability** The explicit link between textual sub-prompts, regional segmentation, and reference conditioning enables transparent reasoning about the generation process. This enhances model interpretability and aligns with emerging research on explainable generative AI.

1.4 Thesis structure

This section is a brief description of the structure of this thesis. This thesis has 5 chapters.

1. **Chapter 2** provides a comprehensive literature review of diffusion-based image generation and multi-reference compositional modeling. It discusses recent advances in latent diffusion architectures, adapter-based personalization, and region-aware control methods, highlighting the research gap that motivates this thesis.
2. **Chapter 3** presents the proposed methodology of our *Regional Planning System*. It details the overall framework design, including the LLM-guided recaption and spatial planning strategy, the IP-Adapter-based regional fusion mechanism, and the integration of multi-reference conditioning into the diffusion process.
3. **Chapter 4** describes the experimental setup, benchmark construction, and quantitative as well as qualitative evaluation results. It compares our model against state-of-the-art diffusion frameworks in terms of compositional fidelity, style preservation, and object consistency, supported by both LLM-based and human evaluation.
4. **Chapter 5** concludes the thesis by summarizing the key findings and contributions. It also discusses the current limitations and potential directions for future research, including improved multi-modal reasoning, dynamic regional attention, and real-time interactive generation.

Chapter 2

literature review

2.1 Basic models

2.1.1 Variational Autoencoders (VAEs)

Variational Autoencoders (VAEs) are a foundational class of likelihood-based generative models introduced by Kingma and Welling in 2014 [24]. They are built upon the architecture of autoencoders but reinterpret the encoding-decoding process through a probabilistic lens. Rather than mapping an input image to a single latent vector, VAEs learn a probability distribution over the latent space, typically assuming a multivariate Gaussian prior. The encoder network maps inputs to the parameters (mean and variance) of this latent distribution, while the decoder samples from this latent space to reconstruct the original data. This formulation allows VAEs to perform efficient sampling and inference within a coherent probabilistic framework.

The training objective of VAEs maximizes a variational lower bound on the data like-

likelihood, balancing two terms: a reconstruction loss that encourages accurate decoding, and a Kullback-Leibler (KL) divergence term that regularizes the latent space toward the prior distribution. This tradeoff ensures that the latent space remains smooth, continuous, and semantically meaningful—properties that make VAEs attractive for interpolation, representation learning, and conditional generation.

Over time, many extensions to the basic VAE framework have been proposed to address its blurry reconstructions and limited expressivity. For example, β -VAE introduces a hyperparameter to control the KL divergence term, enabling more disentangled latent representations [18], while VQ-VAE replaces the continuous latent space with discrete latent codes to better capture local image structure and details [53]. Other improvements include hierarchical VAEs, PixelVAEs, and VAEs with normalizing flows [40] that expand the flexibility of the latent distribution. Despite these innovations, VAEs typically underperform compared to GANs and diffusion models in terms of image sharpness and fine detail, due to their reliance on pixel-wise reconstruction losses and limitations in decoder expressiveness. However, they offer distinct advantages in terms of theoretical grounding, training stability, and explicit likelihood estimation, making them a useful component in hybrid generative systems and in scenarios where latent space interpretability is critical.

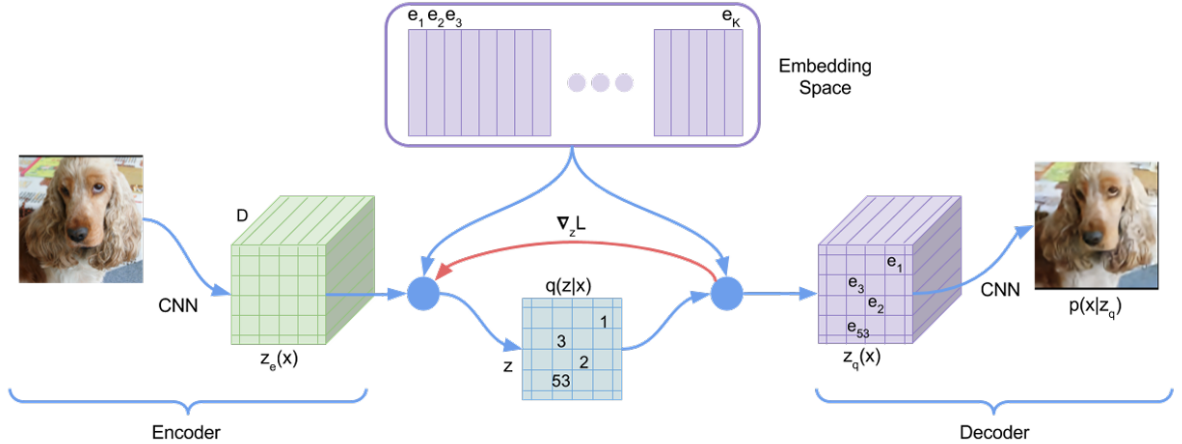


Figure 2.1: The architecture of Vector Quantised Variational AutoEncoder (VQ-VAE) consists of an encoder that maps input data to a latent space and a decoder that reconstructs the data from this latent representation.

In the context of modern generative modeling, VAEs have been used in combination with diffusion processes (e.g., Latent Diffusion Models [41]) to reduce the computational complexity of sampling and improve efficiency. These hybrids leverage VAEs to compress high-dimensional pixel space into compact latent spaces, where denoising diffusion processes can operate more efficiently. As such, while VAEs alone may no longer represent the state-of-the-art in image fidelity, they continue to play an important role in scalable and controllable generative pipelines.

2.1.2 Autoregressive Models

Autoregressive (AR) models represent another major class of likelihood-based generative models, characterized by sequential factorization of the data distribution. They decom-

pose the joint probability of an image or sequence $x = (x_1, x_2, \dots, x_T)$ into a product of conditional probabilities:

$$\log p(x) = \sum_{t=1}^T \log p(x_t \mid x_{<t}) \quad (2.1)$$

This formulation explicitly captures dependencies among elements by generating each token (or pixel) conditioned on all previous ones. Early image autoregressive models such as PixelRNN and PixelCNN demonstrated the effectiveness of this factorization for modeling fine-grained spatial dependencies in images. PixelCNN, in particular, introduced masked convolutions that preserve causality while making training more parallelizable than recurrent pixel-level models.

The introduction of Transformer-based architectures further expanded autoregressive modeling. Models such as GPT [37] and ImageGPT [7] replaced recurrent or convolutional components with self-attention mechanisms, enabling stronger long-range dependency modeling and improved scalability. ImageGPT treated images as sequences of discrete tokens, analogous to language modeling, and was trained with a next-token prediction objective based on cross-entropy loss. This formulation enabled the generation of diverse and coherent image samples and also supported strong unsupervised visual representation learning.

Autoregressive models provide tractable likelihoods, making them useful for density estimation, evaluation, and conditional generation. Their factorization also naturally supports sequential generation, which is beneficial in settings where output structure can be modeled token by token.

However, a major limitation of AR models is slow sampling speed, since tokens must typically be generated sequentially rather than in parallel. This bottleneck becomes in-

creasingly problematic for high-resolution image synthesis and real-time applications. In addition, incorporating complex conditioning signals can introduce extra architectural and computational overhead, especially in image generation settings that require spatial control or multimodal guidance.

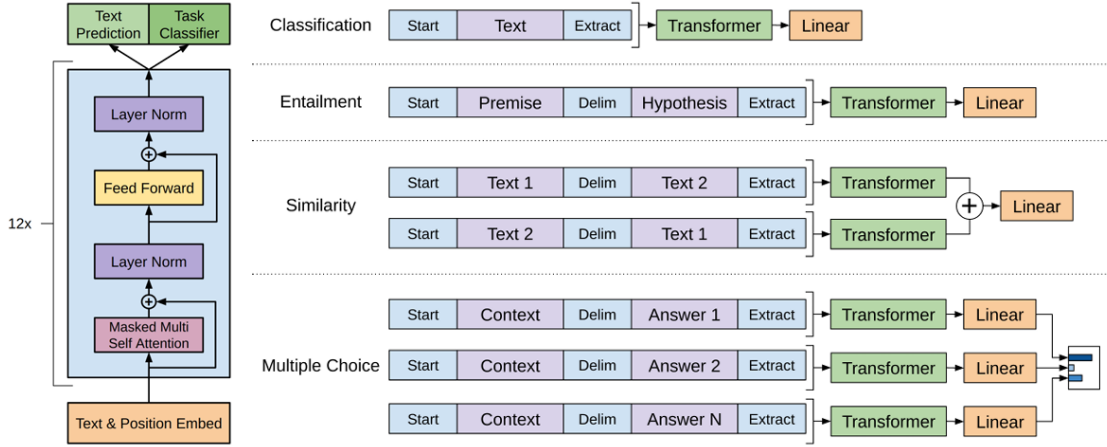


Figure 2.2: The architecture of Generative Pre-trained Transformer (GPT) consists of a stack of transformer decoder layers that utilize self-attention mechanisms to model the dependencies in sequential data. The model is pre-trained on large text corpora using a language modeling objective, enabling it to generate coherent and contextually relevant text. GPT can be fine-tuned for various downstream tasks, such as text generation, translation, and summarization. [37]

2.2 Diffusion models

2.2.1 Diffusion Probabilistic Models

Diffusion Probabilistic Models (DPMs), also known as denoising diffusion probabilistic models (DDPMs), have emerged as a powerful and stable class of generative models in

recent years. Initially introduced by Sohl-Dickstein et al. [48] and significantly improved by Ho et al. [19], diffusion models learn to synthesize data by reversing a gradual noising process. Specifically, they define a forward process that incrementally corrupts data by adding Gaussian noise over multiple time steps, and a learned reverse process that denoises samples step by step to recover data from noise.

Formally, the forward diffusion process constructs a Markov chain that transforms a data sample x_0 into a noisy version x_T by applying Gaussian perturbations at each step. The reverse process, parameterized by a neural network (often a U-Net), learns to model the reverse transition distribution $p_\theta(x_{t-1} \mid x_t)$ conditioned on the noisy input at step t . Training is typically formulated through a variational bound on the negative log-likelihood, which Ho et al. showed can be simplified into a noise-prediction objective closely related to a weighted mean squared error loss [19].

Unlike GANs, which rely on adversarial training and can suffer from instability or mode collapse, diffusion models are generally easier to optimize and have achieved strong sample quality in modern image generation. Large-scale systems such as Imagen [45] and Stable Diffusion [41] illustrate the practical success of diffusion-based text-to-image generation. Extensions such as Classifier-Free Guidance [20] further improve conditional generation by combining unconditional and conditional denoising predictions at inference time.

Despite their advantages, diffusion models are computationally expensive at inference time because sampling typically requires many denoising steps. To mitigate this limitation, a range of acceleration methods have been proposed, including DDIM [49], PNDM-style fast solvers [28], rectified-flow-based methods, and consistency models [50].

Diffusion Probabilistic Models (DPMs), particularly Denoising Diffusion Probabilistic Models (DDPMs), define a generative process by learning to reverse a predefined forward

noising process. The training and sampling procedures of DDPMs are described in Algorithm 2.1 and Algorithm 2.2, respectively. During training, the model learns to predict the Gaussian noise added to noisy inputs. Sampling is performed by iteratively denoising samples starting from Gaussian noise according to the learned reverse process.

Algorithm 2.1 DDPM Training

- 1: **repeat**
- 2: $\mathbf{x}_0 \sim q(\mathbf{x}_0)$
- 3: $t \sim \text{Uniform}(\{1, \dots, T\})$
- 4: $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$
- 5: Take gradient descent step on

$$\nabla_{\boldsymbol{\epsilon}} \left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_{\theta} \left(\sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \boldsymbol{\epsilon}, t \right) \right\|^2$$

- 6: **until** converged
-

Algorithm 2.2 DDPM Sampling

- 1: $\mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$
 - 2: **for** $t = T, \dots, 1$ **do**
 - 3: $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$ if $t > 1$, else $\mathbf{z} = 0$
 - 4: $\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \boldsymbol{\epsilon}_{\theta}(\mathbf{x}_t, t) \right) + \sigma_t \mathbf{z}$
 - 5: **end for**
 - 6: **return** $\mathbf{x}_0$
-

Overall, diffusion probabilistic models have become the de facto standard for high-fidelity image synthesis, particularly when control, stability, and compositional generalization are required—motivating their widespread adoption in both academic research and real-world applications.

2.2.2 Efficiency and Latent-space Improvements

As diffusion models have grown in popularity for high-fidelity image synthesis, their computational cost—particularly during inference—has become a major bottleneck. Traditional diffusion models often require many denoising steps to generate a single image, which limits their practicality in real-time or large-scale applications. To address this, researchers have proposed a variety of efficiency-oriented techniques and latent-space reformulations.

One major direction involves reducing the number of denoising steps. Denoising Diffusion Implicit Models (DDIM) [49] introduced a non-Markovian deterministic variant of DDPM sampling that enables faster, high-quality generation with fewer steps. More recently, Consistency Models [50] and other fast-sampling formulations have further reduced the number of required sampling steps while maintaining competitive image quality.

Another key advancement lies in latent-space modeling, where diffusion is applied not directly in pixel space, but in a compressed latent representation. This approach was popularized by Latent Diffusion Models (LDMs) [41], which use a pretrained autoencoder to map images into a lower-dimensional latent space before applying the diffusion process. By operating in this compressed space, LDMs substantially reduce memory usage and computational cost while preserving strong generation quality. This idea underpins many recent systems, including Stable Diffusion and SDXL.

Additional work has explored discrete latent representations and improved learned compression mechanisms to better balance efficiency and reconstruction quality. These developments have expanded the practical deployment of diffusion models in applications that require lower computational cost and faster inference.

Overall, advances in sampling efficiency and latent-space modeling have played an important role in making diffusion-based generation more practical and scalable.

2.2.3 Conditioning Mechanisms in Diffusion Models

Conditioning mechanisms are a central component in enabling diffusion models to generate targeted and controllable outputs based on user-provided inputs such as text, images, masks, or keypoints. In contrast to unconditional generation, where the model samples data from a learned prior, conditional diffusion models steer the denoising process using external signals. This capability underpins the success of modern text-to-image and image-to-image systems.

A widely adopted conditioning mechanism in modern diffusion models is cross-attention-based conditioning, where text or other conditional embeddings are injected into the denoising network to guide generation. Classifier-Free Guidance (CFG) [20] further improves conditional generation by combining conditional and unconditional denoising predictions at inference time, providing a simple and effective way to trade off image fidelity and prompt adherence.

Beyond global conditioning, recent works have explored spatially aware conditioning that modulates internal representations according to structured inputs. This includes ControlNet [60], which augments a pretrained diffusion backbone with trainable control branches for conditions such as depth maps, edge maps, or poses while preserving the capabilities of the original model. Similarly, Attend-and-Excite [6] guides cross-attention during inference to strengthen neglected subject tokens, improving semantic faithfulness in text-to-image generation.

For image-prompt control, methods such as IP-Adapter [59] inject image features into pretrained text-to-image diffusion models through decoupled cross-attention and lightweight adapter modules. Related work such as FastComposer [57] extends personalized conditioning to multi-subject generation by localizing subject-specific attention to the appropriate

regions of the image.

Text-conditioned systems such as Imagen [45] and SDXL leverage powerful text encoders for context-rich conditioning, while latent diffusion frameworks [41] further improve scalability by combining conditioning with compressed latent representations.

Overall, conditioning mechanisms in diffusion models have evolved from relatively simple global conditioning strategies to more structured control frameworks, supporting a wide range of tasks including semantic image synthesis, inpainting, personalization, and compositional generation.

2.3 Reference Image Generation

2.3.1 Single-reference Conditioning Methods

Single-reference conditioning is a core paradigm in personalized image generation with diffusion models. These methods use a single visual reference image to guide generation together with a user-provided text prompt. Depending on the method design, personalization is achieved either by fine-tuning a pre-trained diffusion model or by injecting image-derived features into the denoising process.

One of the most influential techniques in this domain is DreamBooth [42], which fine-tunes a text-to-image diffusion model using a few images of a target subject together with a rare textual identifier. This allows the model to generate the subject in diverse contexts while preserving its identity under new prompts. However, DreamBooth requires per-subject optimization and can suffer from overfitting or degraded text controllability without careful regularization.

To reduce the overhead of per-subject fine-tuning, plug-and-play approaches such as IP-Adapter [59] introduce image-prompt conditioning without modifying the core diffusion weights. IP-Adapter uses a visual encoder to extract image features and injects them into the denoising network through decoupled cross-attention, enabling efficient image-guided generation while remaining compatible with pretrained models such as Stable Diffusion and SDXL.

A growing line of work has focused on more efficient personalization. Custom Diffusion [25] reduces adaptation cost by updating only a subset of cross-attention parameters instead of fine-tuning the full model. InstantBooth [47] further removes test-time fine-tuning by learning a general image encoder and lightweight adapter layers for instant personalization. HyperDreamBooth [43] uses a hypernetwork to generate lightweight personalized weights from a single reference image, followed by fast subject-specific refinement.

A common limitation of many lightweight single-reference methods is that conditioning is applied mainly at the subject or image level, with limited spatial granularity. As a result, the reference signal often influences the output holistically, making it difficult to localize or disentangle attributes such as background, hairstyle, clothing, or facial expression. This limitation becomes more pronounced when the prompt requires semantic recomposition or more explicit control over different visual attributes.

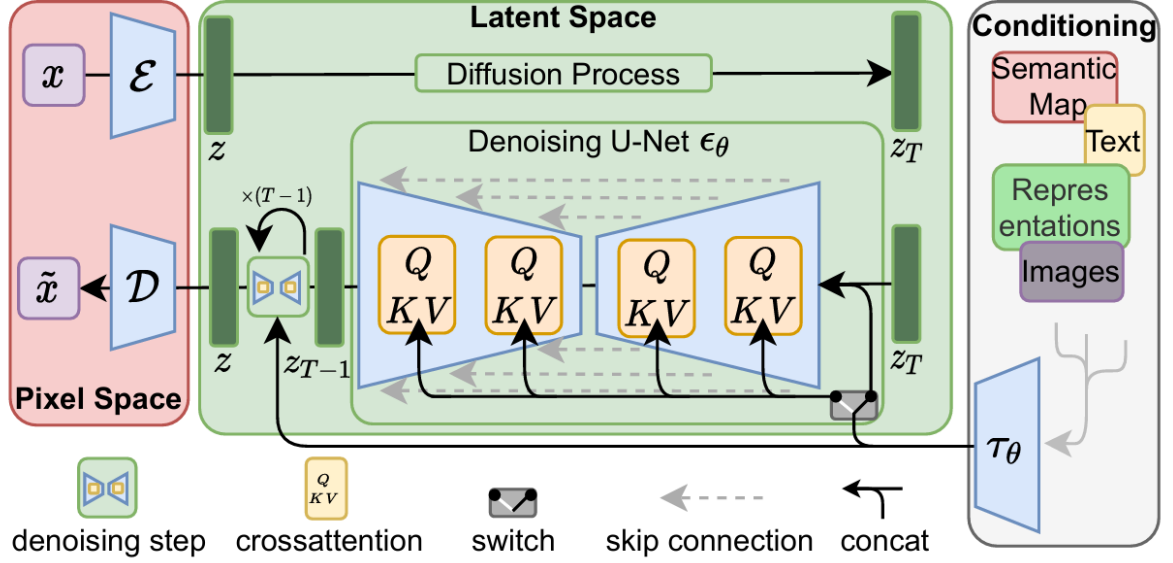


Figure 2.3: A overall architecture of Stable Diffusion model. The model consists of a U-Net based denoising network that operates in a latent space, a variational autoencoder (VAE) for encoding and decoding images, and a text encoder (such as CLIP) for processing textual prompts. During training, the model learns to reverse a diffusion process that gradually adds noise to images, enabling it to generate high-quality images from random noise when conditioned on text inputs. [41]

Despite these limitations, single-reference conditioning methods remain a foundational component in text-to-image personalization pipelines. They serve as the basis for further advancements in multi-subject and multi-style scenarios, where the demand for spatial disentanglement, semantic consistency, and controllable fusion becomes more pronounced. These shortcomings motivate the shift toward more modular, interpretable frameworks that can allocate reference information dynamically and at finer spatial resolutions—setting the stage for the next generation of compositional generative models.

2.3.2 Regional and Spatially-aware Conditioning

While multi-reference generation expands the expressive power of diffusion models, a key challenge is not only whether multiple references can be used, but also where each reference cue should influence the generated image. Many conditioning pipelines apply reference information mainly at the image or subject level, which makes it difficult to localize and modulate distinct conditioning signals in scenes involving multiple entities, attributes, or spatial relationships.

To address this limitation, recent methods have introduced region-aware or spatially guided conditioning frameworks that explicitly control where information should influence generation. For instance, StreamMultiDiffusion [26] proposes a real-time framework for region-based text-to-image generation using semantic masks, enabling interactive control over different spatial regions of the output. This type of design aligns more naturally with human spatial reasoning and supports applications such as scene layout design and multi-object composition.

Some approaches leverage semantic segmentation, bounding boxes, or layout priors as auxiliary inputs. For example, ControlNet [60] conditions generation on structured signals such as edge maps, segmentation masks, or human poses, enabling precise spatial control. Grounded customization methods such as GroundingBooth [58] further demonstrate the importance of associating personalized concepts with spatial grounding signals during text-to-image customization. These methods show that explicit localization cues can effectively improve spatial controllability and reduce semantic ambiguity.

Related work such as FastComposer [57] addresses spatial conflicts in multi-subject generation through localized attention supervision. Instead of relying purely on global subject conditioning, FastComposer supervises cross-attention maps so that subject-specific tokens

attend to the correct image regions during generation, which helps reduce identity blending and improves spatial correspondence across multiple subjects.

However, these methods mainly address *explicit spatial control*, where users or auxiliary systems provide localization signals such as masks, bounding boxes, poses, segmentation maps, or grounded layouts. In this setting, the model is directly informed about what content should appear in which spatial area. Therefore, the distinction of our work is not simply that it is training-free or region-aware, since both properties have already been explored in prior spatial-control methods. Instead, this thesis focuses on *implicit semantic-to-region allocation*: given a global prompt and multiple reference images, the model aims to infer which reference cues should be assigned to which semantic regions from prompt decomposition and reference semantics, without requiring precise user-provided masks or bounding boxes.

From this perspective, LCCR addresses a different but related problem. Rather than treating regional control as an externally specified layout constraint, our method uses LLM-guided prompt decomposition to identify semantic components, associates reference cues with corresponding regions, and performs region-aware fusion while maintaining global coherence. This formulation is particularly important for multi-reference generation, where the success criterion is not merely whether each concept appears in the image, but whether reference-specific attributes are bound to the correct regions, prevented from leaking into unrelated regions, and harmonized into a coherent prompt-aligned output.

Table 2.1: Representative Generative Models Across Architectures

Model	Type	Method	Architecture / Key Feature
StyleGAN2 [23]	GAN	Adversarial	Progressive layers
BigGAN [4]	GAN	Adversarial	ResNet backbone
StyleGAN3 [22]	GAN	Adversarial	Alias-free equivariant generator
VAE [24]	VAE	Variational	Encoder + Decoder with Gaussian prior
VQ-VAE-2 [39]	VAE	Discrete latent	Hierarchical latent quantization
ImageGPT [7]	Autoreg.	Transformer	GPT-style attention for image tokens
DDPM [19]	Diffusion	Denoising	UNet + time embeddings
DDIM [49]	Diffusion	Deterministic	Accelerated sampling
LDM (Stable Diff.) [41]	Diffusion	Latent Diffusion	VAE + CLIP + UNet
DiT [34]	Diffusion	Transformer-based	Vision Transformer backbone
ControlNet [60]	Diffusion	Conditional Diff.	Extra control branch on UNet
DreamBooth [42]	Diffusion	Fine-tuning	Personalized UNet
FastComposer [57]	Diffusion	Compositional	Per-reference attention fusion

2.4 Evaluation Metrics

2.4.1 Perceptual and Semantic Similarity Metrics

Traditional metrics such as FID and LPIPS primarily measure feature-distribution similarity or learned perceptual similarity. However, as generative models become more text- or reference-driven, these metrics do not directly evaluate whether the semantic content of

generated images aligns with the input prompts or reference identities. To address this limitation, several perceptual and semantic similarity metrics have been introduced to better reflect semantic alignment and task relevance.

CLIPScore [16] measures the alignment between an image and its conditioning text using CLIP embeddings. This metric captures semantic correspondence rather than low-level visual similarity and is often used as a proxy for text-image alignment in text-to-image evaluation.

$$\text{CLIPScore} = w \cdot \max(\cos(f_{\text{text}}(t), f_{\text{image}}(x)), 0) \quad (2.2)$$

where f_{text} and f_{image} are CLIP’s text and image encoders, respectively, t is the input text prompt, x is the generated image, and w is a scaling factor commonly set to 2.5 following Hessel et al. [16].

DINO Feature Similarity is computed using self-supervised vision transformers such as DINO or DINOv2 [5, 32]. It measures similarity between generated and reference images in a high-level feature space that captures semantic and structural information. Because DINO features are learned in a self-supervised manner, they are often used as a complementary alternative to perceptual metrics such as LPIPS.

$$\text{DINO Sim} = \cos(f_{\text{DINO}}(x), f_{\text{DINO}}(r)) \quad (2.3)$$

where x is the generated image, r is the reference image, and f_{DINO} is the DINO or DINOv2 feature extractor.

Identity Similarity [9] is commonly used in tasks requiring human face identity preservation, such as portrait personalization or face-oriented subject-driven generation. It com-

puts the cosine similarity between identity embeddings extracted by a face recognition model such as ArcFace from the generated and reference face images.

$$\text{IDSim} = \cos(f_{\text{id}}(x), f_{\text{id}}(r)) \quad (2.4)$$

where f_{id} is the ArcFace feature extractor, x is the generated face image, and r is the reference face image.

Directional CLIP Similarity [33] evaluates whether a generated image reflects a desired semantic transformation, such as a prompt change or style edit. It compares the direction of change in CLIP image embedding space with the corresponding direction of change in CLIP text embedding space. This metric is particularly useful in editing tasks where the goal is to measure semantic shifts rather than absolute similarity.

$$\text{Direction Score} = \cos(\Delta f_{\text{image}}, \Delta f_{\text{text}}) \quad (2.5)$$

where

$$\Delta f_{\text{image}} = f_{\text{image}}(x_2) - f_{\text{image}}(x_1), \quad (2.6)$$

and

$$\Delta f_{\text{text}} = f_{\text{text}}(t_2) - f_{\text{text}}(t_1). \quad (2.7)$$

Here, x_1 and x_2 denote the images before and after editing, while t_1 and t_2 denote the corresponding source and target text prompts.

These semantic-aware metrics provide a richer evaluation signal in scenarios involving prompt alignment, face identity preservation, or semantic transformation.

2.4.2 Aesthetic and Human-aligned Evaluation

While the above metrics quantify alignment and fidelity, they often overlook how well an image appeals to human perception in terms of beauty, composition, and realism. Aesthetic and human-aligned evaluation aims to address this aspect by modeling subjective visual preferences.

Learned Aesthetic Predictors, such as NIMA [51] and aesthetic predictors trained on the AVA dataset [31], are learned from large-scale datasets with human ratings. These models often employ deep CNN backbones to predict either a distribution of aesthetic ratings or an aggregated aesthetic score. Their outputs are intended to correlate with human judgments of visual quality, capturing attributes such as composition, lighting, and visual harmony.

Human Evaluation remains an important way to assess subjective aspects such as creativity, realism, or style faithfulness. Participants are typically asked to rank or score generated images based on predefined criteria, such as identity preservation, prompt fidelity, or visual appeal. Despite its cost and limited scalability, human evaluation provides valuable insights into model usability and perceptual quality, especially in tasks where quality is open-ended or nuanced.

LLM-based Evaluation has recently emerged as a scalable proxy for some forms of human evaluation. Multimodal large language models can assess prompt-image alignment through captioning, comparison, or reasoning-based scoring. For example, a generated image may first be described by the model, and the resulting description can then be compared with the original prompt. This provides a more contextual form of evaluation than purely perceptual similarity metrics, although its reliability still depends on the judge model and evaluation protocol.

CLIP-based Aesthetic Predictors, such as the LAION aesthetic predictor [46], combine CLIP embeddings with lightweight predictors trained to estimate human aesthetic preference. These methods provide a practical compromise between semantic sensitivity and aesthetic assessment, and are often used in style-sensitive or prompt-driven generation settings.

Table 2.2: Summary of Evaluation Metrics in Text-to-Image Generation

Metric	Source	Focus / Strengths
FID	Heusel et al. (2017) [17]	Distributional similarity using Inception features
LPIPS	Zhang et al. (2018) [61]	Perceptual similarity via deep features
SSIM	Wang et al. (2004)	Structural similarity (pixel-level)
PSNR	Traditional	Pixel-level fidelity
CLIPScore	Hessel et al. (2021) [16]	Semantic alignment with CLIP embeddings
Human Evaluation	–	Captures subjective quality and realism
Aesthetic Score	Internal models (e.g., LAION)	Measures human-like aesthetic quality

2.5 Summary

In this Chapter, we reviewed the evolution and current landscape of conditioning mechanisms in text-to-image diffusion models, with a particular focus on personalization, compositional generation, and fine-grained control. We categorized conditioning strategies into single-reference, multi-reference, and regional/spatially-aware methods, highlighting recent advances that enable more flexible and context-aware synthesis.

Single-reference methods, such as DreamBooth and its successors, demonstrate strong capabilities in subject-driven generation but often require finetuning or additional optimization steps. Recent lightweight approaches, like CustomDiffusion and HyperDreamBooth, offer promising directions for faster and more scalable personalization.

Multi-reference and compositional generation has emerged as a key area of innovation, enabling the synthesis of images that integrate multiple visual or semantic concepts. Works like FastComposer leverage localized attention or compositional guidance to improve identity fidelity, style consistency, and structural coherence across subjects.

Regional and spatially-aware conditioning methods, such as ControlNet and Stream-MultiDiffusion, provide users with interactive and fine-grained control over image layout and semantic regions. These approaches are essential for tasks that demand spatial alignment, inpainting, or structured control.

We also discussed evaluation metrics commonly used in generative modeling. While traditional statistical metrics like FID and LPIPS remain standard for benchmarking, they fall short in assessing semantic fidelity and user intent. Thus, human-aligned evaluation methods—including aesthetic scoring, user studies, and large language model (LLM)-based assessments—are gaining traction for more holistic and scalable quality evaluation.

Overall, the field is rapidly progressing toward more controllable, personalized, and semantically aligned generation. Future research will likely focus on developing more generalizable conditioning mechanisms, efficient adaptation methods, and unified evaluation frameworks that better reflect user intent and human judgment.

Chapter 3

Methodology

In this chapter, we propose a unified framework that enhances regional controllability and evaluation in text-to-image diffusion models by integrating large language models for semantic planning and multimodal large language models for human-aligned evaluation. The methodology is centered around two major modules, each addressing a distinct but complementary challenge in region-aware image synthesis: (1) fine-grained regional planning and multi-reference conditioning during generation, and (2) semantic, aesthetic, and human-aligned evaluation powered by multimodal reasoning.

3.0.1 LLM-Guided Regional Prompt Decomposition

To enable spatially controllable generation from a single prompt, we first decompose the global text prompt into multiple semantically meaningful sub-prompts using a large language model (LLM). This step allows for fine-grained control over different spatial regions, enabling heterogeneous multi-subject composition.

Given a user input such as:

‘Generate an image with two different styles of view. On the left is an ancient Japanese temple, and on the right is a modern nightclub.’

we employ an LLM prompted to convert the holistic sentence into regionally grounded instructions. This results in a structured table of sub-prompts and spatial boundaries:

Table 3.1: Example of LLM-based Regional Decomposition

Region	Sub-Prompt	Spatial Area
Left	Ancient Japanese temple, with traditional wooden architecture, curved rooftops, and a serene, spiritually tranquil atmosphere—evoking history and cultural depth.	[0.0, 0.5]
Right	A vibrant modern nightclub glowing with neon lights, filled with contemporary architecture, pulsating colors, and a lively, futuristic party ambiance.	[0.5, 1.0]

Each sub-prompt is paired with a corresponding reference image, forming conditioning pairs used in the generation process.

3.0.2 Region-Aware Complementary Generation via Latent Fusion

Following the semantic decomposition of the global prompt into multiple region-specific sub-prompts, the next stage involves synthesizing a visually coherent image that captures both the localized characteristics of each region and the overall compositional integrity. To achieve this, we introduce a generation strategy named **Latent Composition with Complementary Regions (LCCR)**. This approach enhances spatial controllability without

imposing rigid layout constraints, offering a more flexible alternative to conventional region-constrained methods.

Compared with many existing region-control methods, LCCR enables a more fluid and adaptive integration of sub-regional features. Each region is generated independently based on its assigned semantic sub-prompt, while the resulting features are fused in a way that encourages smooth transitions across boundaries. This helps reduce the visual fragmentation that can arise in layout-driven diffusion processes.

An important aspect of our design is the inclusion of a globally conditioned latent representation derived from the full prompt. This global context serves as a guiding signal, ensuring that the synthesized image not only respects the individual sub-prompts but also reflects the intended scene-level semantics.

By combining regional specificity with global coherence, LCCR facilitates the generation of complex multi-object or multi-style scenes without requiring additional training or fine-tuning of the underlying diffusion model. When regional branches are processed independently, the method can also support efficient generation as visual complexity increases.

Core Benefits of the LCCR Approach:

- *Flexible Spatial Composition:* Avoids rigid layouts by supporting soft integration across regions.
- *Scene Coherence:* Maintains consistency with the holistic prompt while preserving local details.
- *Zero Training Overhead:* Achieves controllable generation without modifying pre-trained diffusion models.

- *Parallelizability*: Supports efficient generation by allowing region-wise sampling to occur concurrently.

3.1 Preliminaries

This chapter introduces the core foundations and notations underpinning our proposed framework, which enhances regional controllability and evaluation in text-to-image diffusion models. In particular, we review essential concepts from diffusion-based generative models, large language models (LLMs), and multimodal conditioning strategies. A clear understanding of these components is necessary for interpreting the region-aware generation and evaluation techniques discussed in subsequent sections. Rather than presenting new methods, this section aims to establish the conceptual and mathematical groundwork upon which our framework is built.

As mentioned in the literature review, text-to-image diffusion models [19, 41, 49] generate high-quality images by learning to reverse a gradual noising process. This process starts from a clean sample x_0 and adds Gaussian noise iteratively over T timesteps, producing a noisy sample x_T . A neural network ϵ_θ is trained to predict the noise component at each timestep t , thereby learning to reverse the diffusion process and gradually reconstruct the clean image. The denoising trajectory is guided by a conditioning signal, most commonly a natural language prompt $\mathbf{c}$. Classifier-free guidance [20] is often used to strengthen the semantic alignment between the prompt and the generated image by interpolating between unconditional and conditional generations.

To improve computational efficiency, modern approaches frequently adopt the latent diffusion paradigm [41], where the generation process operates in a lower-dimensional latent space $\mathcal{Z}$ instead of pixel space. The input image is first encoded into this latent space

using a pretrained autoencoder, allowing the model to focus on semantically meaningful structures while reducing memory and compute requirements. After the denoising process, the latent is decoded back into image space, resulting in high-resolution, photorealistic outputs.

Then, understanding how conditioning affects generation is crucial for grasping the architectural design of our framework. In typical diffusion pipelines, the text prompt $\mathbf{c}$ is encoded (e.g., via a CLIP-based encoder [36]) into a sequence of tokens, which modulate the generation via cross-attention layers in the UNet backbone. This global prompt conditioning, while effective for single-scene images, struggles with compositional prompts that involve multiple entities or spatially localized content. For example, a flat prompt like “*a dragon on the left, a castle on the right*” often results in entangled or blended outputs, as the model lacks explicit mechanisms to localize distinct semantics.

In addition to textual conditioning, visual guidance from reference images plays a critical role in controllable generation—especially for capturing style, identity, or fine-grained details that text alone may not convey. Techniques such as IP-Adapter [59] enable the injection of visual features into the diffusion pipeline through decoupled cross-attention modules. This multimodal conditioning strategy allows the model to reflect both the semantic intent of the prompt and the stylistic or structural cues from example images.

Lastly, maintaining global coherence across spatially distinct regions requires balancing local guidance with scene-level consistency. In our framework, a global context latent derived from the original prompt acts as a semantic anchor during generation, while region-specific latents guide local structures. Understanding this global-local interaction is essential for interpreting the fusion-based techniques we present later in the methodology.

Notation Summary:

- x_0 : Clean image; x_t : Noisy latent at timestep t ; x_T : Fully noised latent after T steps.
- $\mathbf{z} \in \mathcal{Z}$: Latent representation of the image in the diffusion process; typically a tensor of shape (H, W, D) .
- $\epsilon_\theta(x_t, t, \mathbf{c})$: Denoising network that predicts noise at timestep t , conditioned on prompt $\mathbf{c}$.
- $\mathcal{Z}$: Compressed latent space where diffusion occurs.
- $\mathbf{c}$: Input text prompt; can be global or regional.
- $\mathbf{c}_{\text{global}}$: Global scene-level prompt used to preserve overall coherence.
- c_r : Sub-prompt associated with region r .
- I_r : Reference image providing appearance or style information for region r .
- $[a_r, b_r]$: Normalized horizontal coordinates defining the spatial extent of region r .
- $\mathbf{t}_j$: Text embedding of sub-prompt c_j .
- $\mathbf{v}_j$: Visual embedding extracted from reference image I_j .
- $\mathbf{e}_j$: Hybrid embedding combining text and visual features for region r_j .
- $\mathcal{T} = \{c_j\}$: Set of recaptioned sub-prompts derived from the global prompt.
- $\mathcal{R} = \{r_j\}$: Set of spatial regions, each mapped to a sub-prompt and spatial area.
- $\mathcal{L}$: Inferred layout specification representing region arrangement.
- **Ratio String**: Serialized string encoding relative widths per row (e.g., `0.5,0.5;1.0`).

- **Prompt String**: Flattened sequence of sub-prompts separated by delimiters.
- $\text{Attn}_{\text{cross}}$: Cross-attention module used to inject conditional embeddings into the denoising process.

This foundational knowledge prepares the reader to understand the architectural and algorithmic choices in our framework. In the next sections, we describe how these concepts are operationalized to enable structured prompt decomposition, reference-guided generation, and region-aware diffusion sampling.

3.2 Prompt Recaption via Compositional Language Understanding

Text-to-image diffusion models have achieved remarkable generative fidelity, but they still struggle with spatial compositionality and region-level controllability when processing prompts involving multiple semantic subjects. Most existing pipelines treat the input prompt as a holistic conditioning signal, lacking mechanisms to disentangle semantic entities or associate them with localized regions in the image [13, 30].

To address this, we first introduce a novel intermediate representation: Prompt To Region. This module leverages the semantic abstraction capabilities of Large Language Models (LLMs) to decompose a holistic prompt $\mathbf{c}$ into a set of localized, semantically disentangled, and stylistically enriched sub-prompts $\{c_r\}$. Each c_r corresponds to a spatial region r , and provides fine-grained guidance during the diffusion generation process.

3.2.1 Why Flat Prompts Fail

Despite their expressiveness, flat prompts—i.e., single-sentence descriptions treated as unified global embeddings—often lead to suboptimal generation quality in complex scenes due to the following issues:

- **Entity Overlap:** When multiple distinct entities are mentioned (e.g., “A tiger sitting beside a firefighter in a city street”), the model may generate entangled subject features or fail to preserve clear entity boundaries.
- **Style Leakage:** When different regions are intended to have distinct visual styles (e.g., “A realistic landscape with a cartoon-style dragon flying above”), the model may average the styles across the image. As a result, both subjects appear in a mismatched, intermediate visual domain rather than their intended individual styles.
- **Undirected Semantics:** Flat prompts lack explicit spatial grounding. For example, in “A desert on the left and an ocean on the right with a volcano in the middle” the model may swap the positions or blend the elements. This is because the text embedding does not provide an explicit mechanism to bind concepts to designated image regions.

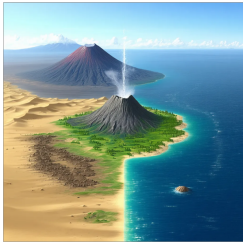
	Entity Overlap	Style Leakage	Undirected Semantics
Stable Diffusion 2.1			
Stable Diffusion 3.5			
Prompt	A tiger sitting beside a firefighter in a city street	A realistic landscape with a cartoon-style dragon flying above	A desert on the left and an ocean on the right with a volcano in the middle

Figure 3.1: Failure cases of flat prompts. Flat prompts such as **beside firefighter city street**, **cartoon-style dragon**, and ambiguous spatial descriptions lead to semantic entanglement, style leakage, or spatial inconsistency in generation.

These limitations arise because conventional diffusion pipelines process prompts holistically, embedding all semantics into a single latent vector. Without structural disentanglement, the model has no mechanism to reason about spatial layout, object boundaries, or style separation.

3.2.2 Prompt Recaption Task Definition

Given an input prompt $\mathcal{P}$, our goal is to extract a list of region-specific sub-prompts:

$$\mathcal{T} = \{c_r\}_{r=1}^N, \quad \text{where each } c_r \text{ is a semantically enriched textual description for region } r$$

This task is solved using a structured prompt protocol applied to an LLM, designed to enable step-wise decomposition and recaptioning:

1. **Entity Extraction:** Identify discrete visual entities and attributes.
2. **Disentangled Semantics:** Group visual descriptions by entity.
3. **Recaptioning:** For each entity, generate a vivid, stylistically coherent sub-prompt c_r .

Example Input: *‘Generate an image with two different style of view On the left is an ancient Japanese temple, and on the right is a modern nightclub ’*

Recaptioned Prompts:

- c0: “On the left, a timeless Japanese temple stands in tranquil harmony with nature. Its traditional wooden pillars and gracefully sloping rooftops are adorned with finely carved details, weathered by time.....”
- c1: “On the right, a vibrant modern nightclub explodes with color and motion. Neon signs blaze across the glass exterior, casting electric blues and pinks into the night.....”

3.2.3 Design Benefits

- Operates in natural language space—no masks or boxes required.
- Improves semantic disentanglement during generation.
- Compatible with any diffusion pipeline supporting per-region conditioning.

The recaptioned sub-prompts $\mathcal{T}$ are passed along with layout metadata (see Section 3.3) to guide region-aware diffusion sampling.

3.3 Region Planning via Spatial-Aware Layout Reasoning

While recaptioning decomposes a flat prompt into entity-specific sub-descriptions, it does not address where in the image each entity should appear. To enable spatial controllability, we propose a lightweight but effective **Region Planning** module, which infers 2D spatial layouts purely from natural language cues.

3.3.1 Task Definition

Given:

- A set of region-level sub-prompts $\mathcal{T} = \{c_r\}_{r=1}^N$, where each c_r describes an individual entity.

- Spatial cues embedded in the original prompt (e.g., “on the left”, “above”, “foreground”).

Output:

- A compositional layout $\mathcal{L} = \{\text{Row}_i\}_{i=1}^M$, where each row has a normalized height $h_i \in (0, 1]$.
- A region map $\mathcal{R} = \{r_r\}_{r=1}^N$, where each region $r_r = (\text{Row}_i, w_r, e_r)$ assigns:
 - Row_i : the row index the region belongs to,
 - w_r : normalized horizontal width of the region (e.g., 0.5 for half-width),
 - e_r : the visual entity described by sub-prompt c_r .

3.3.2 LLM-Based Layout Inference Pipeline

We instruct the LLM to perform structured reasoning over prompt semantics, including:

- **Row Splitting:** Infer vertical segmentation based on entity type or composition logic.
- **Region Splitting:** Assign horizontal regions to co-located or side-by-side entities.
- **Ratio Estimation:** Predict normalized width and height for each region.

3.3.3 Serialization for Diffusion Control

We serialize the layout as:

- **Ratio String:** e.g., 0.5,0.5;1.0 representing widths in each row.
- **Prompt String:** Recaptioned prompts joined by delimiter **BREAK**.

3.3.4 Example: Layout Representation

Prompt: “Generate an image with two different style of view On the left is an ancient Japanese temple, and on the right is a modern nightclub.”

- Row0: Left (temple) – $w = 0.5$; Right (nightclub) – $w = 0.5$
- Row1: Background (city) – $w = 1.0$

Serialized Layout: ‘0.5,0.5;1.0’

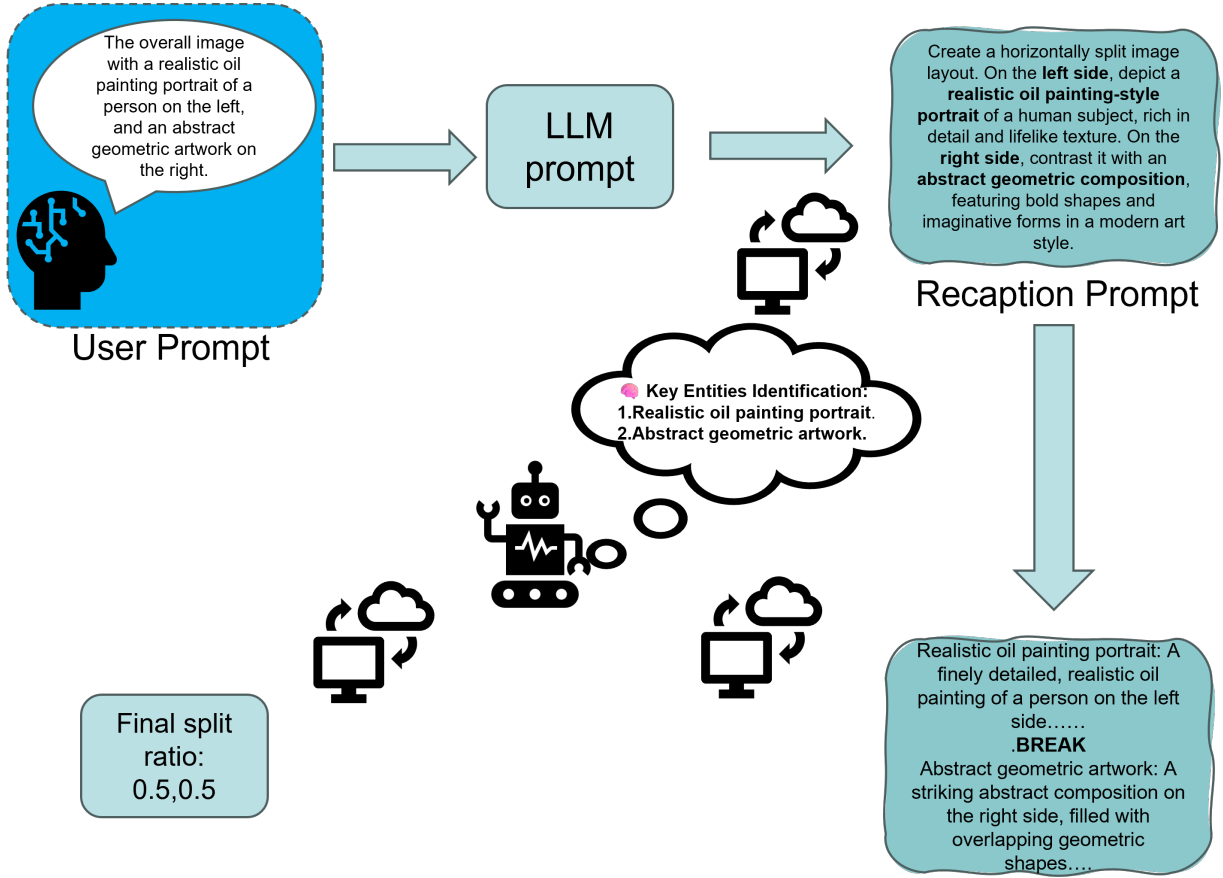


Figure 3.2: Illustration of the LLM-based recaptioning process. The system refines the original user prompt by identifying key entities and generating a structured recaption prompt that explicitly separates visual regions and stylistic elements. This enables clearer semantic guidance and improved image composition for multi-region generation.

3.3.5 Advantages Over Prior Methods

- **No Visual Supervision:** No bounding boxes or masks required.
- **Semantic Coherence:** Aligns spatial assignment with language-level meaning.
- **Generalization:** Works for abstract, stylized, and multi-subject prompts.

3.4 Region-wise Conditioning and Visual Injection

3.4.1 Stable Diffusion

Stable Diffusion XL (SDXL) [35] serves as the backbone of our proposed region-aware generation framework. As a state-of-the-art latent diffusion model, SDXL improves significantly over its predecessors in terms of generation fidelity, semantic alignment, and flexibility. It introduces several architectural innovations and training optimizations that make it particularly suitable for tasks requiring spatial compositionality and modular conditioning.

SDXL adopts a dual-stage architecture comprising a *base model* and an optional *refiner model*. The base model generates semantically coherent low-resolution images in a high-dimensional latent space, while the refiner enhances image fidelity and fine details in a subsequent stage. Our method primarily operates on the base model, leveraging its high-resolution latent representations for regional injection and control.

One of SDXL’s key advancements is its dual text-encoder design. It combines two pretrained text encoders, OpenCLIP and CLIP, to improve semantic understanding and stylistic expressivity. These text embeddings are fused and injected into the denoising UNet via cross-attention layers at each denoising timestep. This dual-path conditioning provides a rich latent prior that enables nuanced control over both content and style.

Furthermore, SDXL operates in a larger latent space $\mathcal{Z} \in \mathbb{R}^{128 \times 128 \times 4}$ for 1024×1024 image resolution, which offers enhanced spatial granularity—critical for regional-level editing and synthesis. The model is trained with diverse aspect ratios and scene compositions, giving it strong generalization across both natural and abstract scenes.

During inference, SDXL supports classifier-free guidance (CFG), where the final pre-

diction is interpolated between conditional and unconditional branches of the denoising network. This allows controllable trade-offs between prompt fidelity and image diversity without requiring retraining.

Importantly, SDXL is designed to support flexible conditioning pathways. The attention-based architecture allows seamless injection of auxiliary conditioning signals—such as regional embeddings, spatial masks, or visual features—making it compatible with plug-and-play modules like IP-Adapter and regional prompt planning. Our method builds on this extensibility to achieve fine-grained regional control by injecting multimodal conditioning tokens into designated spatial areas during the denoising process.

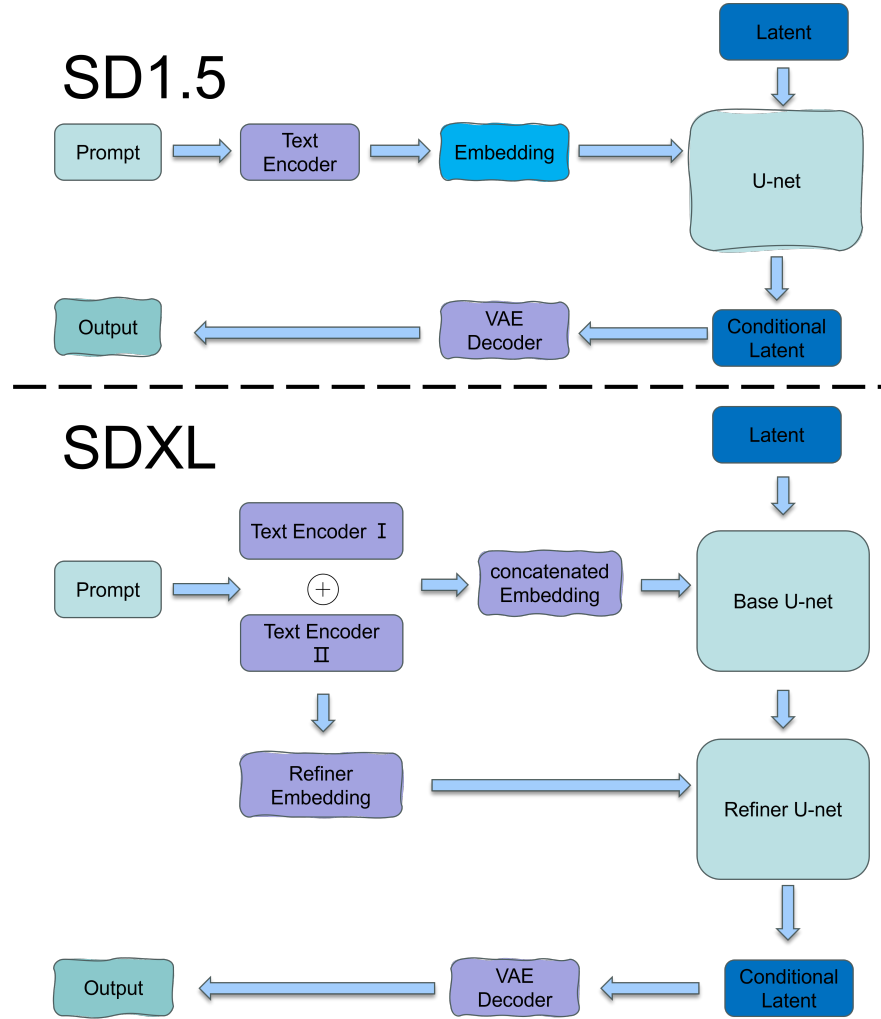


Figure 3.3: Comparison of Stable Diffusion (top) and Stable Diffusion XL (bottom) architectures. SD employs a single text encoder and UNet pipeline, while SDXL introduces dual text encoders, a two-stage generation process with Base and Refiner UNets, and enhanced latent conditioning for higher-resolution synthesis.

By taking advantage of these architectural properties, we are able to introduce a novel region-aware generation mechanism that remains fully training-free while supporting localized semantics, per-region style transfer, and compositional control.

Our approach operates in two complementary stages:

- **Text-based Regional Conditioning:** Every region is independently conditioned on a recaptioned sub-prompt c_r , ensuring semantic disentanglement.
- **Reference-based Visual Injection:** When visual references I_r are available, they are projected into the same embedding space and fused with text embeddings to enhance visual specificity.

Together, these mechanisms enable per-region control while preserving global coherence during image synthesis. The regional prompts $\{c_j\}$ are generated via an LLM-based recaptioning pipeline that decomposes a flat input into semantically and spatially disentangled sub-prompts.

3.4.2 Region-wise Conditioning Fusion

To enable spatially disentangled generation in multi-entity prompts, we associate different parts of the image with distinct semantic and stylistic signals with Latent Composition with Complementary Regions (LCCR). This mechanism allows each spatial region to be conditioned independently during the diffusion process while maintaining global coherence.

Given a set of regional prompts $\mathcal{T} = \{c_r\}_{r=1}^N$, optionally aligned reference images $\mathcal{I} = \{I_r\}_{r=1}^N$, and corresponding normalized horizontal spans $\mathcal{R} = \{[a_r, b_r]\}_{r=1}^N$, our goal is to synthesize a latent image $\mathbf{z}$ such that each region r reflects the semantics of c_r and the appearance of I_r , without introducing inter-region interference. To achieve this, we modify the cross-attention layers of the U-Net by injecting region-aware logic into their forward passes. At each denoising step t , the latent tensor $\mathbf{z}_t \in \mathbb{R}^{H \times W \times D}$ is partitioned into spatial sub-blocks.

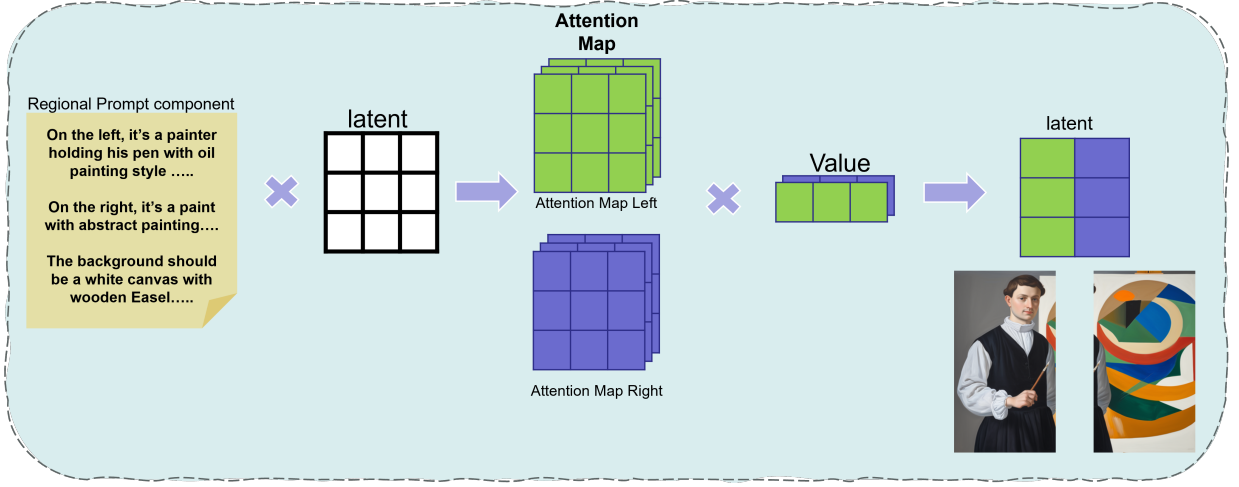


Figure 3.4: Illustration of regional attention-guided latent generation. The Regional Prompt component provides distinct textual descriptions for different spatial regions (e.g., left and right). Each region’s prompt produces a corresponding attention map that modulates the latent representation through cross-attention, resulting in spatially separated latent features. This enables coherent composition while preserving the unique visual semantics of each regional prompt.

Each region’s conditioning signal is a hybrid embedding $\mathbf{e}_j = \text{Concat}(\mathbf{t}_j, \mathbf{v}_j)$, where $\mathbf{t}_j$ is obtained via a CLIP-based text encoder and $\mathbf{v}_j$ comes from a visual encoder call IP-Adapter if reference images are provided. These embeddings are projected into key-value pairs (K_j, V_j) for cross-attention. During attention computation, each spatial location in the latent is routed to its associated region, and its query attends only to (K_j, V_j) . We compute attention outputs for each region independently and reconstruct the full latent tensor by spatially stitching together the region outputs.

To ensure continuity across region boundaries and prevent visual seams, we adopt a soft fusion mechanism with a global base context. When a base region is defined, its attention

output $\mathbf{z}_{\text{base}}$ is combined with regional output $\mathbf{z}_{\text{region}}$ via a weighted blend:

$$\mathbf{z}_{\text{final}} = \alpha \cdot \mathbf{z}_{\text{region}} + (1 - \alpha) \cdot \mathbf{z}_{\text{base}}, \quad \alpha \in [0, 1]$$

where α is a manually or adaptively specified coefficient indicating the strength of regional influence. This allows our model to preserve stylistic and semantic consistency in global context while maintaining fine-grained local control.

3.4.3 Enhancing Control with Visual Reference Injection

While regional prompts bring semantic disentanglement and spatial grounding to multi-subject scenes, textual descriptions alone often lack the specificity required to guide fine-grained visual attributes such as facial identity, artistic style, or object texture. To address this limitation, we enhance region-wise control by introducing visual reference injection—a mechanism that supplements textual prompts with appearance-grounded embeddings derived from user-provided reference images.

For each spatial region r_j , we assume a pair (c_j, I_j) , where c_j is a recaptioned textual description and I_j is an optional visual reference image. The text prompt is encoded into an embedding $\mathbf{t}_j$ via the frozen text encoder(s) of the backbone model, while the reference image is passed through an IP-Adapter [59] module to yield a visual embedding $\mathbf{v}_j$. Specifically, I_j is first processed by a frozen ViT-based encoder to extract visual tokens, which are then projected into the conditioning space using a learned adapter. These two modalities are fused via concatenation followed by a linear projection, yielding a hybrid conditioning token:

$$\mathbf{e}_j = \text{Concat}([\mathbf{t}_j \parallel \mathbf{v}_j])$$

This design allows each region to carry a rich, multimodal descriptor that combines both semantic intent and concrete visual cues. The resulting hybrid embeddings are injected into the U-Net cross-attention layers at region-specific spatial locations as described in the previous section. During inference, each spatial query is primarily influenced by its designated $\mathbf{e}_j$, ensuring that both the content and style of the region are guided by the provided reference image.

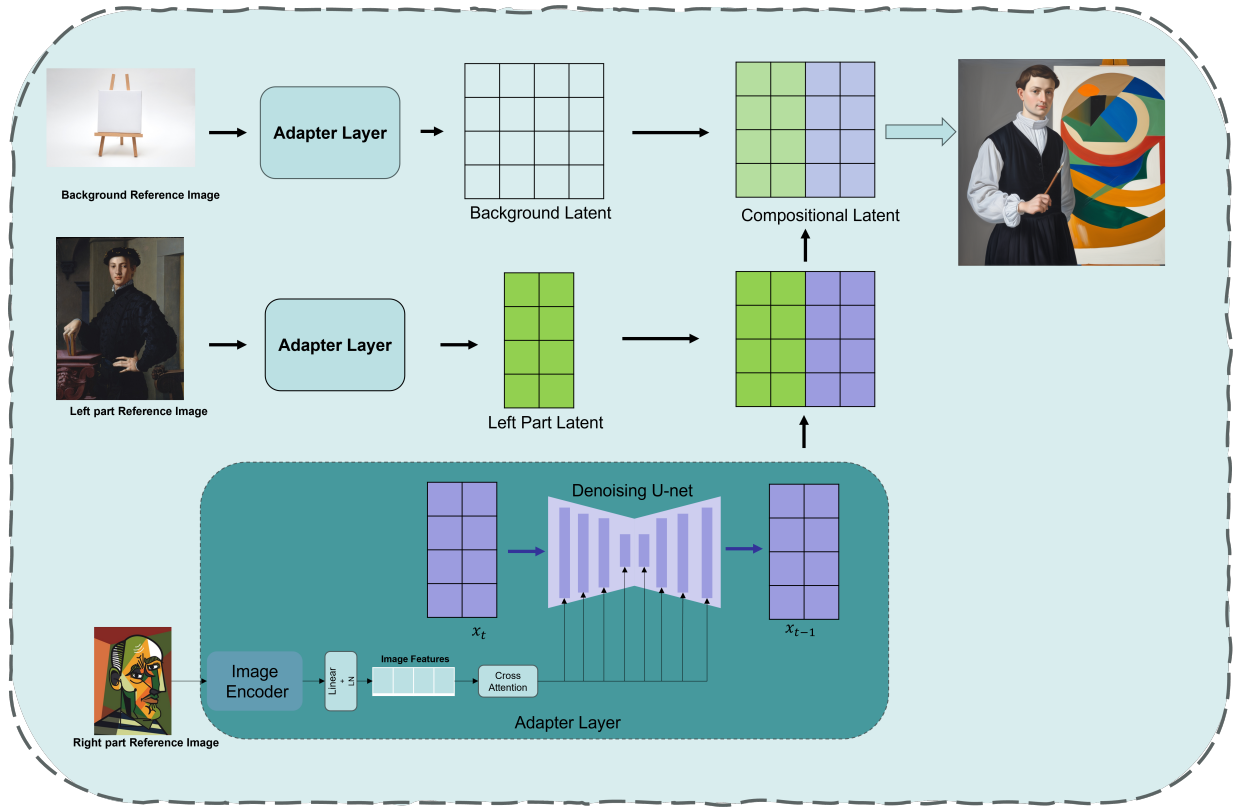


Figure 3.5: Integration of IP-Adapter with Regional Planning for compositional generation. Each regional reference image (background, left, right) is processed through individual adapter layers to produce corresponding latent features. The resulting compositional latent enables coherent blending of multiple reference styles—maintaining spatial structure while preserving each region’s visual identity.

To ensure compatibility with training-free pipelines, the visual reference injection is implemented in a plug-and-play fashion that does not require model fine-tuning. The injected embeddings flow seamlessly through the standard attention mechanism and can be toggled per region. Additionally, our system supports multiple reference images per batch and aligns them dynamically with the corresponding token spans and spatial regions through index-aware attention hooks.

This visual injection strategy helps resolve ambiguity in natural language prompts, reinforces stylistic consistency across generated regions, and allows users to guide the appearance of specific entities with high fidelity. It is particularly effective for tasks requiring identity preservation (e.g., character consistency across frames), complex aesthetics (e.g., watercolor vs cyberpunk), or fine detail control (e.g., clothing, hair, texture).

3.4.4 Benefits and Generalization

Our method offers several advantages:

- **Fine-Grained Control:** Each region independently controls identity, style, and position.
- **Style Disentanglement:** Prevents leakage across unrelated sub-scenes.
- **Modularity:** Each region can be swapped or updated without re-processing the entire prompt.
- **Compatibility:** Plug-and-play with any CLIP-guided latent diffusion model (e.g., SD 1.5, SDXL, ControlNet).

3.4.5 Illustrative Example

Input:

- Prompt: *“On the left, a cyberpunk warrior. On the right, a Renaissance noblewoman.”*
- Ref A: Illustration of a cyberpunk character.
- Ref B: Portrait of a Renaissance woman.

Effect: Without reference images, the left and right may share texture palettes or facial structure. With our injection, each side inherits not only semantic cues but also concrete visual attributes (e.g., armor reflectance, hair style, facial proportions).

Remark: Our region-aware fusion strategy can be viewed as a generalization of multi-subject personalization, enabling multi-reference blending with semantic modularity.

3.5 Summary

In this chapter, we presented a unified framework for enhancing regional controllability and evaluation in text-to-image diffusion models. By leveraging the semantic reasoning capabilities of Large Language Models (LLMs), our method enables flexible, compositional, and human-aligned generation of multi-region images without requiring model fine-tuning. The key contributions and components of our methodology are as follows:

- **LLM-Guided Regional Decomposition:** We introduced a prompt decomposition module that uses LLMs to convert a single holistic prompt into multiple semantically

enriched and spatially grounded sub-prompts. This enables precise regional control over image content and layout without relying on visual annotations.

- **Latent Composition with Complementary Regions (LCCR):** To ensure both regional fidelity and global coherence, we proposed a latent-space fusion strategy where independently conditioned region-wise latents are softly merged, guided by a global prompt embedding. This enables seamless composition across heterogeneous styles and subjects.
- **Prompt Recaptioning:** We addressed the limitations of flat prompts through an LLM-based recaptioning pipeline that disentangles entities and enriches descriptions. This improves semantic clarity and ensures that each region receives specific and stylistically coherent guidance.
- **Language-Driven Layout Reasoning:** Without requiring any spatial supervision, we designed a layout inference mechanism that extracts 2D compositional structure from natural language. The result is a serialized layout specification that controls spatial arrangement in the latent fusion stage.
- **Region-Wise Conditioning and Reference Injection:** Each region is further grounded through hybrid conditioning that combines text and visual reference embeddings. Leveraging IP-Adapter modules, this mechanism enables precise control over visual identity, texture, and style in a plug-and-play fashion.
- **LLM-Based Multimodal Evaluation:** Finally, we adopt LLMs as evaluators to assess the generated images across semantic alignment, aesthetic quality, and regional accuracy, enabling scalable and human-aligned feedback without ground truth annotations.

Overall, our methodology integrates LLMs throughout the generation pipeline—as planners, decomposers, layout reasoners, and evaluators—yielding a modular and extensible approach for complex, multi-region image synthesis. It achieves high spatial fidelity, rich semantic grounding, and visual coherence, all without requiring additional training or modification of the base diffusion model.

Chapter 4

Experiment

4.1 Experiment Setup

To evaluate the effectiveness of our approach, we compare against several baseline models, including EasyRef, Gen-4, DALL·E 2, IP-composer, MS-Diffusion, and Stable Diffusion. For models that support reference-based conditioning, generation is performed using the same reference images and textual prompt. For text-only baselines, evaluation is conducted using the textual prompt alone. To ensure fair and reproducible evaluation, we construct a standardized dataset of prompts and reference images, and use a consistent evaluation protocol across all models. Detailed descriptions of the metrics, GPT-based evaluation, and human ratings are provided in later sections.

4.1.1 Evaluation Dataset and Sample Selection

To avoid selection bias and ensure fair comparison across models, the evaluation prompts and reference image sets were determined before generating any model outputs. The evaluation set contains 50 prompt-reference cases in total, covering diverse multi-reference scenarios. Specifically, 30 cases are person-centric and focus on identity preservation, character consistency, clothing attributes, and region-level subject binding; 10 cases are object-centric and evaluate object appearance, texture, material, and compositional placement; and 10 cases are style-centric and focus on visual style transfer, color palette preservation, and cross-style fusion. This distribution is designed to cover the main use cases of region-aware multi-reference generation, including identity preservation, style transfer, texture fusion, object composition, and region-aware semantic binding.

Once the prompt-reference cases were finalized, the same inputs were used for all evaluated models. For models that support reference-based conditioning, both the textual prompt and the corresponding reference images were provided. For text-only baselines, only the textual prompt was used, while the same prompt remained unchanged.

Importantly, no prompt or reference case was added, removed, or replaced based on the visual quality of any generated result. Quantitative results were computed over the full evaluation set rather than over a manually selected subset. This ensures that the reported scores reflect overall model behavior instead of isolated favorable cases.

The qualitative examples shown in the main text were selected from the same fixed evaluation set for representativeness. They are intended to illustrate typical success cases, common failure modes, and visible differences between models under identical input conditions. These qualitative figures are not used as the sole basis for our conclusions; the main claims are supported by quantitative evaluation over the complete evaluation set and by

the human preference study. Additional examples are provided in the appendix to improve transparency and reproducibility.

4.1.2 T2I-CompBench Evaluation

For quantitative evaluation, we assess compositional text-image alignment on the T2I-CompBench benchmark [21], a comprehensive benchmark for compositional text-to-image generation that covers attribute binding, object relationships, and complex compositions. While this benchmark is not specifically designed for multi-reference generation, it provides a useful complementary evaluation of whether a model can correctly bind attributes, preserve object relations, and handle compositional prompts. These abilities are closely related to the region-aware planning and semantic binding objectives of our method.

As shown in Table 4.1, LCCR achieves the highest scores across all reported categories. In attribute binding, our method obtains the best performance on color, shape, and texture, with scores of 0.8073, 0.6544, and 0.8577, respectively. These results indicate that LCCR is effective at associating visual attributes with the correct semantic entities, which is essential for multi-reference generation where different references may provide different colors, textures, styles, or object-level details.

For object relationship reasoning, LCCR also achieves the highest scores in both spatial and non-spatial categories. In particular, the improvement in the spatial category suggests that the proposed planning-guided regional generation framework better preserves relative positioning and spatial correspondence between entities. This is consistent with the design goal of LCCR, which aims to allocate semantic and reference cues to appropriate regions while maintaining global coherence.

Notably, LCCR achieves the highest score on the Complex metric, reaching 0.4575.

Table 4.1: Attribute Binding and Object Relationship Benchmark

Model	Attribute Binding			Object Relationship		Complex↑
	Color↑	Shape↑	Texture↑	Spatial↑	Non-Spatial↑	
Stable Diffusion v1.4 [41]	0.3765	0.3576	0.4156	0.1246	0.3079	0.3080
Stable Diffusion v2 [41]	0.5065	0.4221	0.4922	0.1342	0.3096	0.3386
Composable Diffusion [29]	0.4063	0.3299	0.3645	0.0800	0.2980	0.2898
DALL-E 2 [3]	0.5750	0.5464	0.6374	0.1283	0.3043	0.3696
SDXL [35]	0.6369	0.5408	0.5637	0.2032	0.3110	0.4091
EasyRef [63]	0.4421	0.4139	0.4258	0.2246	0.3157	0.3351
IP-Composer [10]	0.6214	0.4864	0.7584	0.3037	0.3122	0.3331
MS-Diffusion [54]	0.7945	0.6360	0.8233	0.3267	0.3608	0.4505
LCCR (Ours)	0.8073	0.6544	0.8577	0.3481	0.3621	0.4575

This indicates stronger capability in handling multi-attribute and multi-object compositional prompts. MS-Diffusion [54] is the strongest baseline in this benchmark and performs competitively across most categories, especially in texture and spatial reasoning. Nevertheless, LCCR consistently outperforms it across all metrics, suggesting that the proposed region-aware fusion and planning mechanism provides more stable compositional control. Overall, these results support the effectiveness of LCCR not only in reference-guided generation, but also in broader compositional text-to-image synthesis.

4.1.3 Rethinking Evaluation: Beyond Traditional Metrics

Traditional evaluation metrics such as Fréchet Inception Distance (FID) [17] and Learned Perceptual Image Patch Similarity (LPIPS) [61] have long been standard tools for measuring image quality in generative models. These metrics, while useful for assessing distributional alignment or perceptual similarity, fall short in capturing the complex, high-level semantics required for evaluating modern reference-guided text-to-image generation tasks.

In particular, they are insufficient in scenarios where the primary challenge lies not in generating photorealistic images, but in achieving nuanced fusion and alignment between multiple sources of conditioning information, such as two or more reference images combined with a guiding textual prompt.

In our work, we choose not to report traditional metrics like FID or LPIPS for three primary reasons. First, these metrics do not directly evaluate the semantic integration of multiple conditioning inputs, which is the central challenge in multi-reference generation. Second, the core innovation in our approach lies not in architectural changes that would directly influence raw image quality, but rather in the conditioning and fusion mechanisms used to guide generation based on multiple references. In other words, we do not propose changes to the generative backbone that would justify expectations of improved FID scores. Instead, our contribution focuses on *how well* the model is able to semantically integrate multiple inputs—style, identity, structure—into a coherent and aesthetically pleasing image that aligns with user intent. This type of evaluation requires tools that can reason about *semantic fidelity*, not just pixel-level similarity or visual plausibility.

Third and most importantly, existing automated metrics are poorly aligned with human perception in the context of multi-reference generation. In multi-input scenarios, users often expect the generated image to preserve identifiable elements from all reference images, including specific textures, facial features, and stylistic details, while also achieving a visually harmonious composition. Traditional metrics provide no direct mechanism for evaluating *fusion quality*—a critical aspect of such generation tasks. For instance, LPIPS may score two images as similar due to superficial textural alignment, even when the underlying semantic fusion between reference styles is incoherent or visually jarring.

To address these shortcomings, we adopt an evaluation protocol that better aligns with human perceptual criteria while remaining reproducible and scalable. We design

a GPT-based evaluation system capable of reasoning about different visual attributes. Our system leverages the visual understanding capabilities of multimodal GPT models to provide structured, human-aligned judgments while reducing reliance on costly manual evaluation.

By rethinking how we evaluate generative models in multi-reference, style-fusion contexts, we aim to shift the focus of the field from low-level fidelity to higher-order semantic alignment. This shift is necessary as generation tasks become more personalized, instruction-driven, and semantically rich. Our evaluation protocol thus reflects not only the goals of our own approach, but also the broader trajectory of generative model research—toward more controllable, human-centered, and semantically grounded synthesis.

4.1.4 Qualitative Success Criteria for Region-Aware Multi-Reference Fusion

Although quantitative scores provide a useful summary of model performance, they do not fully capture the specific goals of region-aware multi-reference generation. In this thesis, a successful result is not defined merely by whether visual elements from all reference images appear in the generated output. Instead, the generated image should satisfy a joint set of qualitative criteria that reflect the core objective of LCCR: preserving reference-specific attributes, assigning them to appropriate semantic regions, fusing them coherently, preventing cross-reference interference, and maintaining alignment with the global prompt.

Specifically, we analyze generated images according to the following qualitative success criteria:

- **Reference Preservation:** The generated image should preserve key visual at-

tributes from each reference image, such as identity, style, texture, color, clothing, material, or object-specific details. A result is not considered successful simply because a similar object or person appears; the generated content should retain recognizable reference-specific characteristics.

- **Reference-Region Binding:** Each reference cue should influence the intended semantic or spatial region. For example, visual attributes from the left reference should not incorrectly appear in the right region, and identity or style cues should not be transferred to unrelated subjects or areas.
- **Fusion Quality:** The generated image should integrate multiple references into a visually coherent whole rather than appearing as a mechanical collage. Important aspects include smooth regional transitions, consistent lighting, compatible perspective, natural boundaries, and overall scene harmony.
- **Prompt Following:** The output should remain faithful to the textual prompt, including the specified subjects, actions, spatial relationships, scene context, and stylistic requirements. Strong reference preservation alone is insufficient if the generated image fails to follow the prompt.
- **Non-Interference:** Different reference images should not contaminate one another. Common failure cases include identity blending, style leakage, color transfer to unrelated regions, over-dominance of one reference, or averaging multiple references into an ambiguous intermediate appearance.

These criteria are used to guide our qualitative analysis and failure-case discussion. They are not treated as separate quantitative metrics in this thesis; rather, they clarify the visual and semantic properties that a successful region-aware multi-reference fusion

result should exhibit. In summary, a successful output should preserve reference-specific attributes, assign them to the correct regions, fuse them coherently, avoid cross-reference interference, and remain aligned with the prompt.

4.1.5 GPT-based Evaluation

To effectively assess region-aware multi-reference generation, we design a task-specific GPT-based evaluation protocol. The goal is not to claim novelty from using a multi-modal GPT model as an evaluator, but to define a structured evaluation setting tailored to multi-reference generation. In this setting, the evaluator must judge whether reference-specific visual attributes are preserved, whether the generated image follows the textual prompt, and whether multiple visual cues are fused into a coherent and visually plausible output. Traditional metrics fail to capture the complex semantic alignment required when conditioning on multiple visual sources; in contrast, multimodal GPT models provide stronger visual-semantic reasoning capabilities, making them suitable for scalable and human-aligned evaluation under a carefully defined rubric.

Evaluation Setup: In our evaluation protocol, each input consists of two reference images, the original textual prompt, and one generated image. The evaluator assesses the generated image solely based on the prompt, the reference images, and the generated result. It is not provided with internal knowledge of the generation process or model architecture. This design ensures that the evaluation focuses on observable visual and semantic quality rather than implementation-specific details.

The evaluation is organized into three groups of criteria, corresponding to the dimensions reported in Table 4.2. First, **Concept Preservation** measures how well the generated image preserves visual information from the reference images. It is reported us-

ing two sub-scores: **Subject Fidelity**, which focuses on identity, object structure, and subject-specific details, and **Style Consistency**, which focuses on visual style, texture, color palette, and appearance-level cues from the references. Second, **Prompt Following** measures how well the generated image satisfies the textual instruction. It is reported using two sub-scores: **Style Transfer**, which evaluates whether the requested stylistic transformation is correctly reflected, and **Imaginative Understanding**, which evaluates whether the model can follow more creative or compositional prompt requirements. Third, the **Overall Composition** score provides a holistic assessment of the final generation quality. This score considers reference preservation, prompt alignment, fusion naturalness, reference-region consistency, non-interference, and global visual coherence together.

Scoring Scale: For each reported dimension, the evaluator outputs an integer score from 0 to 4, corresponding to five levels of semantic and perceptual alignment:

- **Very Poor (0):** The generated image fails to satisfy the target dimension, with severe semantic mismatch, weak reference preservation, poor prompt alignment, or obvious visual inconsistency.
- **Poor (1):** The generated image shows only weak alignment with the target dimension and contains clear errors, missing attributes, unstable reference usage, or unnatural fusion.
- **Fair (2):** The generated image partially satisfies the target dimension, but some reference attributes, prompt requirements, or fusion details remain incomplete or inconsistent.
- **Good (3):** The generated image satisfies the target dimension reasonably well, with recognizable reference preservation, prompt alignment, or coherent fusion, but still

contains minor flaws.

- **Excellent (4):** The generated image strongly satisfies the target dimension, preserving reference-specific attributes, following the prompt accurately, and producing a coherent and visually natural result.

The GPT model is prompted with a fixed instruction template for each evaluation dimension and is instructed to return only the final numerical score. This design reduces variation from free-form textual explanations and simplifies score aggregation across samples.

Instruction Template: Below is the core structure of the prompt used for GPT-based evaluation:

You will be provided with one generated image, two reference images, and the original text prompt.

Your task is to evaluate the generated image according to the specified dimension.

Please consider whether the generated image preserves the relevant reference attributes, follows the text prompt, and forms a coherent multi-reference output.

Score Range: 0 (Very Poor) to 4 (Excellent).

Return only the final score.

Score: [Your Score]

Score Aggregation: For each model and each reported dimension, we compute the average GPT-assigned score over the full evaluation set. The two concept-preservation scores are reported as **Subject Fidelity** and **Style Consistency**, while the two prompt-following scores are reported as **Style Transfer** and **Imaginative Understanding**. The

Overall score is assigned as a separate holistic score rather than computed as a direct arithmetic average of the four sub-scores. This allows the overall score to capture multi-reference fusion quality, reference-region consistency, cross-reference non-interference, visual coherence, and prompt alignment jointly.

Reproducibility and Alignment: To ensure consistency, we apply the same evaluation template across all models and samples. The use of a fixed instruction template and discrete scoring scale makes the evaluation reproducible while still allowing the evaluator to reason about high-level visual semantics. Furthermore, our pilot experiments indicate that GPT-based scores are positively correlated with human judgment, supporting its use as a scalable and cost-effective approximation to human evaluation.

Overall, this GPT-based evaluation strategy is intended not as a generic image-quality metric, but as a task-specific evaluation protocol for multi-reference generation. The reported scores quantify concept preservation, prompt following, and overall compositional quality, while the holistic overall score further reflects fusion naturalness, reference-region consistency, cross-reference non-interference, and global-local coherence.

4.2 Experiment Result of a human-aligned benchmark powered by advanced multi-modal GPT models

We evaluate our model, **LCCR**, against several strong baselines using the proposed GPT-based benchmark. As shown in Table 4.2, LCCR achieves the highest overall compositional score among all evaluated methods. More importantly, it maintains strong and balanced performance across both concept preservation and prompt following, indicating its effectiveness in region-aware multi-reference generation where reference fidelity, textual alignment,

and coherent fusion must be jointly satisfied.

In terms of **concept preservation**, LCCR achieves the highest subject fidelity score of 2.788 and the highest style consistency score of 2.680. Although MS-Diffusion [54] achieves a very close subject fidelity score of 2.786, LCCR still obtains the best overall performance in the concept preservation category when both subject fidelity and style consistency are considered together. Compared with SDXL+IP-Adapter [35, 59], EasyRef [63], and IP-Composer [10], LCCR shows substantially stronger reference preservation, suggesting that the proposed region-aware fusion mechanism can better retain visual attributes from multiple reference images. In particular, IP-Composer obtains moderate concept preservation scores but remains clearly below LCCR in both subject fidelity and style consistency, indicating that concept-level composition alone may be insufficient for preserving reference-specific attributes in complex multi-reference settings.

For **Prompt Following**, LCCR achieves the highest imaginative understanding score of 3.826 and a strong style transfer score of 3.821. MS-Diffusion obtains a slightly higher style transfer score of 3.829, but LCCR remains highly competitive in this dimension while outperforming all baselines in imaginative understanding. Compared with IP-Composer, which obtains lower scores in both style transfer and imaginative understanding, LCCR demonstrates stronger ability to preserve prompt alignment while integrating multiple visual references. This result suggests that LCCR does not improve reference preservation at the expense of textual alignment. Instead, it maintains strong prompt-following ability while producing coherent and semantically aligned multi-reference outputs.

Overall, LCCR achieves the best comprehensive performance, with an overall compositional score of 3.279. This surpasses MS-Diffusion (3.238), DALL-E 2 (3.083), Gen-4 (2.903), SDXL+IP-Adapter (2.543), IP-Composer (2.518), and EasyRef (2.159). These results support the effectiveness of LCCR for region-aware multi-reference fusion. Rather

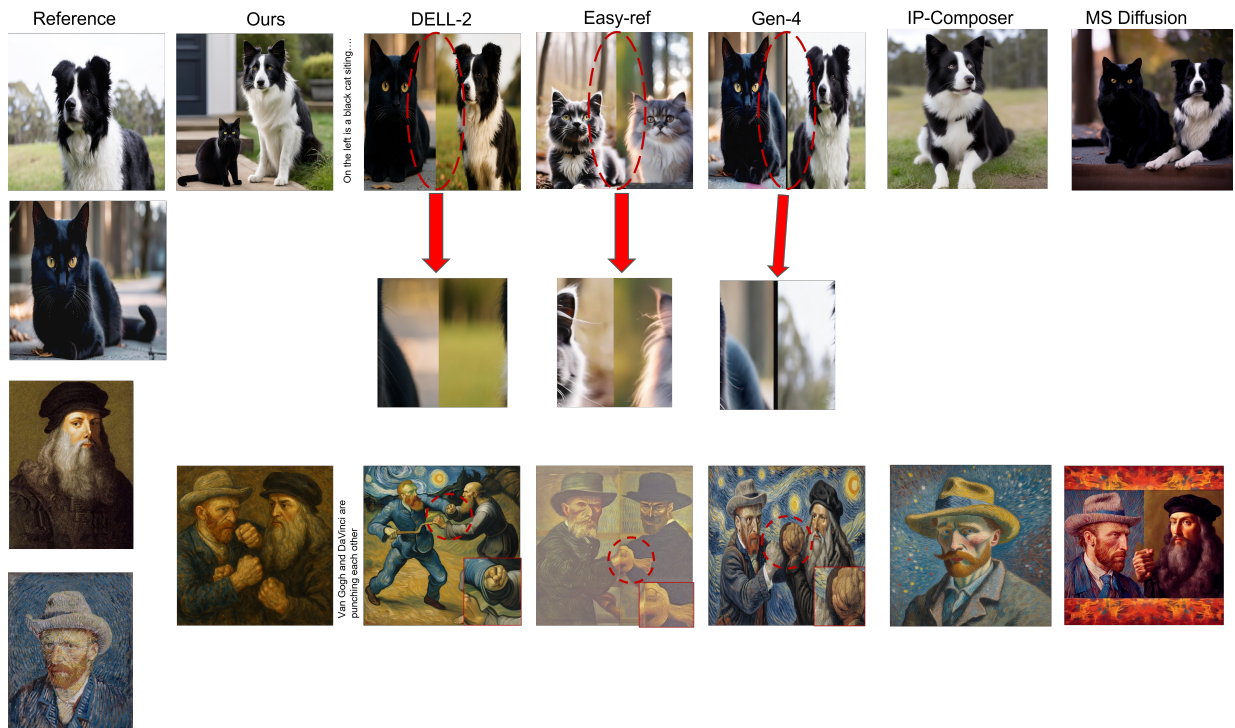


Figure 4.1: Detailed comparison with other multi-reference generation models. Our method preserves identity-specific details and achieves coherent fusion across multiple reference images, while competing methods exhibit artifacts such as feature distortion, inconsistent blending, semantic leakage, or weaker global-local consistency. Highlighted regions show common failure cases in baseline models, including mismatched textures, inaccurate facial attributes, and unstable cross-reference alignment.

Table 4.2: Compositional Generation Benchmark

Model	Concept Preservation		Prompt Following		Comp.
	Subj. (0-4)↑	Style (0-4)↑	Style Trans. (0-4)↑	Imaginative (0-4)↑	Overall (0-4)↑
Stable Diffusion v1.4 [41]	–	–	2.423	2.337	–
Stable Diffusion v2 [41]	–	–	2.875	2.809	–
DALL-E 2 [3]	2.765	2.657	3.298	3.610	3.083
SDXL + IP-Adapter [35, 59]	2.578	1.566	3.004	3.025	2.543
EasyRef [63]	0.800	0.860	3.543	3.436	2.159
Gen-4 (Runway)	2.693	1.361	3.764	3.794	2.903
IP-Composer [10]	2.024	2.127	2.911	3.011	2.518
MS-Diffusion [54]	2.786	2.651	3.829	3.689	3.238
LCCR (Ours)	2.788	2.680	3.821	3.826	3.279

than merely preserving isolated visual concepts, LCCR better balances reference preservation, prompt following, and global-local coherence. Qualitative analysis further supports this trend: baseline models often either overemphasize one reference, weaken reference-specific attributes, fail to maintain stable concept-region correspondence, or produce less coherent fusion across regions. In contrast, LCCR demonstrates more consistent spatially aware fusion and stronger alignment with both reference constraints and textual conditions.

4.3 Experiment Result of real human evaluation

Please answer the following two question based on the references images and text description

Reference image


man_caucasian_blue_shirt.jpg


woman_asian_red_sweater.jpg

Generated Image

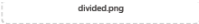
Generated Image A


gen4.png

Generated Image B


Gen-4 Turbo - Two different people chatting in a cozy cafe with warm lighting, the image is





textDescription

A cinematic split-screen shot of two different people chatting in a cozy cafe with warm ambient lighting. On the left side of the frame: a Caucasian man with short neatly styled brown hair, clean-shaven, wearing a crisp light blue button-up shirt, smiling naturally as he speaks. On the right side: an Asian woman with long black wave hair cascading over her shoulders, wearing a textured red knit sweater, smiling

Q 1: Which image is more similar to the reference image?

☐ Image A

☐ Image B

Q 2: Which image is more similar to the prompt?

☐ Image A

☐ Image B

Q 3: Overall which image do you think is better?

☐ Image A

☐ Image B

☐ I don't know

To further validate the effectiveness of our method and assess its alignment with human perception, we conducted a user study via Amazon Mechanical Turk (MTurk). This human

evaluation complements our GPT-based assessment and provides an additional human-centered benchmark for evaluating region-aware multi-reference generation. A total of 50 participants were recruited for the study, all of whom were registered MTurk workers. Each participant completed the questionnaire independently based on the provided textual prompts, reference images, and generated results.

Study Design: We adopted a blind pairwise comparison protocol. For each comparison, participants were presented with the same textual prompt, the same reference images, and two generated images produced by different models. One image was generated by LCCR, and the other was generated by a baseline model. The model names were hidden from participants, and the display order of the two generated images was randomized to reduce positional bias.

The human evaluation was conducted on a subset of the fixed evaluation set rather than on samples selected after inspecting generation results. Specifically, we used 35 prompt-reference sets for the human study. Each set contains one textual prompt, the corresponding reference images, one generated image from LCCR, and one generated image from a baseline model. For each baseline comparison, the same 35 prompt-reference sets were used. Each LCCR-vs-baseline comparison was independently evaluated by three different participants, resulting in 105 valid human judgments for each baseline model.

A total of 50 participants were recruited. To balance the workload, 25 participants completed 10 pairwise comparisons, and the remaining 25 participants completed 11 pairwise comparisons. The assigned comparisons were randomly distributed across participants while ensuring that each comparison set was evaluated by three independent participants. This design reduces individual annotator bias and ensures that the reported human preference results are not based on a single subjective judgment.

For each pairwise comparison, participants were asked three forced-choice questions:

1. **Prompt Following:** *Which image better follows the textual prompt?* This question evaluates whether the generated image correctly reflects the described subjects, spatial relationships, scene context, actions, and style requirements.
2. **Reference Preservation:** *Which image better preserves the visual attributes of the reference images?* This question evaluates whether the generated image retains reference-specific visual cues, such as identity, style, texture, color, clothing, material, object details, or other recognizable attributes.
3. **Overall Preference:** *Overall, which image better satisfies the prompt-reference generation task?* This question asks participants to make a holistic judgment by considering prompt following, reference preservation, fusion naturalness, visual coherence, and whether the final image appears as a unified and plausible generation result.

For each question, participants selected either Image A or Image B. We did not include a tie option, since the goal of the study is to obtain a clear pairwise preference between LCCR and each baseline under the same prompt-reference condition. The first two questions are used as diagnostic dimensions: Prompt Following measures textual alignment, while Reference Preservation measures fidelity to the visual references. The third question, Overall Preference, is treated as the primary human evaluation result because it reflects the combined perceptual quality of the generated image under the multi-reference generation task.

It is possible for a participant’s response to the Overall Preference question to differ from their responses to Prompt Following or Reference Preservation. Such cases are not treated as invalid responses. Instead, they reflect the fact that overall visual preference may depend on additional perceptual factors such as fusion quality, boundary naturalness,

artifact reduction, aesthetic coherence, or cross-reference interference. Therefore, the three questions are reported separately, with Overall Preference serving as the main result and the other two questions serving as explanatory dimensions.

Preference Result Reporting: For each baseline model and each question, we report the number of valid participant responses selecting LCCR and the number of valid participant responses selecting the corresponding baseline. Let b denote a baseline model and q denote one of the three questions. If $n_{\text{LCCR},b}^{(q)}$ is the number of valid responses selecting LCCR, and N_b is the total number of valid responses for that baseline comparison, then the number of responses selecting the baseline is $N_b - n_{\text{LCCR},b}^{(q)}$. The percentage shown in parentheses is computed as:

$$\text{Preference}_{\text{LCCR},b}^{(q)} = \frac{n_{\text{LCCR},b}^{(q)}}{N_b} \times 100\%.$$

In our setting, each baseline comparison contains 35 prompt-reference sets, and each set is evaluated by three independent participants. Thus, $N_b = 105$ for each baseline model and each question. In Table 4.3, each cell is reported in the format *LCCR chosen* / *Baseline chosen*, with the LCCR preference percentage shown in parentheses. The Overall Preference column is used as the main human evaluation result, while Prompt Following and Reference Preservation provide additional diagnostic analysis.

Table 4.3: Human preference results comparing LCCR against baseline models. Each cell reports the number of participants who preferred LCCR versus the corresponding baseline, with the percentage of LCCR preference shown in parentheses. Each comparison contains 105 valid responses.

Baseline Model	Prompt Following	Reference Preservation	Overall Preference
	<i>LCCR / Baseline</i>	<i>LCCR / Baseline</i>	<i>LCCR / Baseline</i>
EasyRef	70 / 35 (66.7%)	88 / 17 (83.8%)	82 / 23 (78.1%)
Gen-4	56 / 49 (53.3%)	67 / 38 (63.8%)	62 / 43 (59.0%)
DALL-E 2	58 / 47 (55.2%)	66 / 39 (62.9%)	60 / 45 (57.1%)
IP-Composer	64 / 41 (61.0%)	78 / 27 (74.3%)	75 / 30 (71.4%)
MS Diffusion	61 / 44 (58.1%)	73 / 32 (69.5%)	80 / 25 (76.2%)

4.4 Ablation Study

4.4.1 Condition Scale

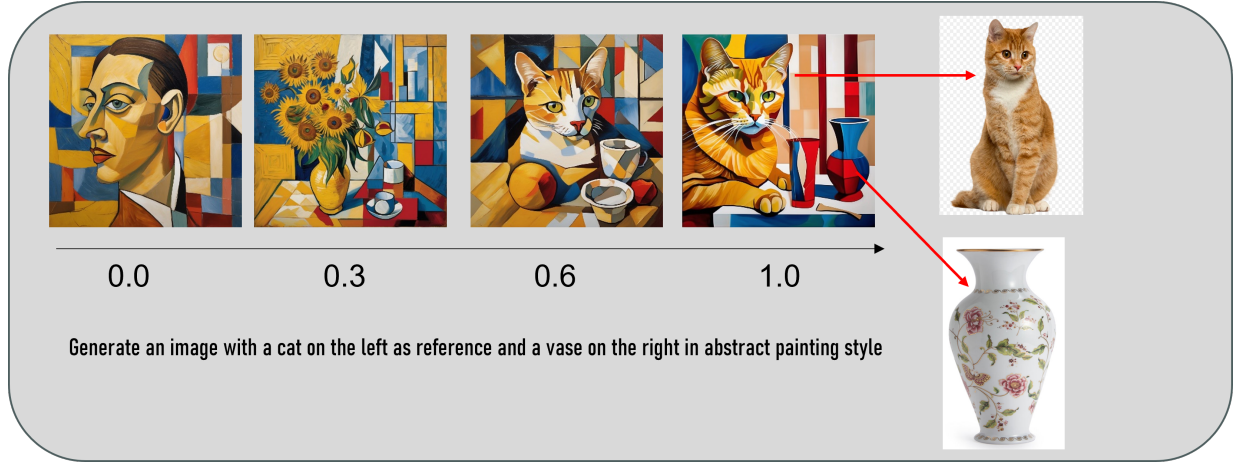


Figure 4.2: Ablation study on the effect of Condition Scale in regional image generation. As the condition scale increases from 0.0 to 1.0, the generated image progressively aligns more closely with the reference style and content. Lower scales (e.g., 0.0–0.3) yield weaker conditioning and looser resemblance, while higher scales (e.g., 0.6–1.0) produce stronger stylistic fidelity and more precise visual alignment with the conditioning reference.

We systematically evaluate the impact of our image injection component by analyzing the scaling factor of the IP-Adapter across different reference images. A critical hyperparameter within the IP-Adapter framework is the scaling factors, which serves as a weighting coefficient that modulates the relative contribution of the reference image embedding with respect to the text prompt embedding during the denoising process. Specifically, this scaling factor controls the strength of image-conditioned features injected into the cross-attention layers of the UNet, thereby providing a continuous mechanism to balance visual fidelity to the reference image and semantic alignment with the textual prompt.

To analyze its effect, we conduct a controlled ablation study by progressively increasing

the scaling factor while keeping all other generation parameters fixed. As illustrated in Figure 4.3, lower values of s lead to generations that primarily follow the textual description, exhibiting high semantic diversity but weaker adherence to the visual characteristics of the reference image. In contrast, as the scaling factor increases, the generated outputs gradually incorporate more visual cues from the reference image, resulting in improved preservation of stylistic attributes and object-level details such as color distribution, texture patterns, and structural layout.

Notably, when the scaling factor becomes sufficiently large, the generation increasingly prioritizes the reference image, often at the expense of textual flexibility. In this regime, the outputs tend to closely resemble the reference image in both style and object appearance, occasionally exhibiting reduced diversity or overfitting to specific visual features. These observations indicate that the scaling factor plays a crucial role in mediating the trade-off between controllability and generalization, enabling fine-grained adjustment between prompt-driven creativity and reference-driven fidelity.

Overall, this study demonstrates that the scaling factor acts as an effective and interpretable control knob for image-conditioned generation. By appropriately tuning s , practitioners can flexibly adapt the generation behavior to different application scenarios, such as style transfer, identity preservation, or prompt-guided creative synthesis.

4.4.2 Ablation on Recaption and Regional Diffusion

We conduct a 2×2 factorial ablation to analyze the effects of two core components in LCCR: the LLM-based recaption module and the regional diffusion module. The goal of this study is to examine how semantic prompt restructuring and region-aware generation affect the final synthesis quality. Rather than treating recaption as a standalone prompt-

engineering step, this ablation studies how recaption interacts with regional diffusion in a multi-reference generation setting.

Table 4.4 summarizes the four controlled variants. The base variant disables both modules and directly uses the original prompt under global conditioning. The recaption-only variant enables LLM-based prompt rewriting but keeps the generation process global. The regional-diffusion-only variant enables region-wise generation while removing LLM-based recaption. The full model enables both modules.

Table 4.4: Ablation settings for analyzing the effects of prompt recaption and regional diffusion.

Variant	Recaption	Regional Diffusion
Base Variant	No	No
Recaption Only	Yes	No
Regional Diffusion Only	No	Yes
Full LCCR	Yes	Yes

For a fair comparison, all variants use the same original prompt, the same reference images, the same random seed, and the same generation parameters. The only changed factors are whether the recaption module is enabled and whether the regional diffusion module is enabled. This controlled setting allows us to observe how each component affects semantic clarity, subject organization, and reference-to-region correspondence.

Figure 4.3 shows a representative qualitative comparison. When both recaption and regional diffusion are disabled, the model captures the coarse intent of the prompt, such as a coffee-shop conversation, but the result remains visually and semantically loose. The generated scene does not strongly preserve the intended reference-specific appearance, and the output tends to rely on generic global conditioning.

Enabling recaption alone improves the clarity of the textual guidance. In this set-

ting, the model better captures the intended two-person interaction and coffee-table scene. However, because the generation is still globally conditioned, the reference cues are not explicitly bound to separate regions or subjects. As a result, the generated image can still appear generic and may not fully preserve the desired identity, clothing, or scene-specific attributes.

Enabling regional diffusion without recaption improves local realism and provides stronger spatial separation between subjects. However, without structured semantic decomposition, the region-level guidance may still be ambiguous. In the example shown, this can lead to unstable subject organization, such as the appearance of an additional person, indicating that regional control alone does not fully resolve semantic assignment.

The full LCCR setting combines recaption and regional diffusion. In this example, it produces a more stable two-person composition, better preserves the intended subject attributes, and maintains a more coherent coffee-shop interaction. These observations suggest that the two modules play complementary roles: recaption clarifies what should be generated, while regional diffusion helps control where different semantic and reference cues should be applied.



No recaption + No Regional Diffusion



No recaption + Regional Diffusion



Recaption + No Regional Diffusion



Recaption + Regional Diffusion(Ours)

Figure 4.3: Qualitative ablation results of recaption and regional diffusion. The four settings are arranged as: top-left, *w/o Recaption + w/o Regional Diffusion*; top-right, *w/ Recaption + w/o Regional Diffusion*; bottom-left, *w/o Recaption + w/ Regional Diffusion*; and bottom-right, *w/ Recaption + w/ Regional Diffusion*. Without both modules, the model mainly captures the coarse scene intent but produces a loose and generic result. Recaption improves prompt clarity but does not fully ensure reference-to-subject binding. Regional diffusion improves spatial separation but may still suffer from ambiguous semantic assignment. The full setting produces a more coherent two-person composition with clearer subject correspondence and stronger global-local consistency.

Chapter 5

Conclusion and future work

5.1 Conclusion

In this work, we propose LCCR, a framework for multi-reference image generation that improves concept fusion and prompt alignment under a human-aligned evaluation setting. Our motivation stems from the limitations of existing methods in maintaining semantic fidelity when dealing with complex compositional inputs. While current approaches often focus on either subject preservation or stylization in isolation, LCCR is designed to handle both simultaneously through a layered cross-conditioned representation integration strategy.

We introduce a unified architecture that accommodates multiple reference images and textual prompts, enabling a fine-grained fusion of semantic attributes and visual styles. Our approach builds upon the foundation of diffusion models while incorporating an advanced alignment mechanism across visual layers, allowing for robust interpolation between content and style dimensions. Through extensive empirical validation, we show that LCCR achieves

strong performance across multiple dimensions of concept preservation, prompt following, and compositional complexity.

A key contribution of our study is the adoption of a multimodal GPT-based evaluation framework. Traditional pixel-level or embedding-based metrics fail to capture nuanced semantic relationships, especially when synthesizing content from multiple references. By leveraging advanced multimodal large language models, we conduct fine-grained, human-aligned assessments that better reflect real-world preferences and expectations. Our GPT-based evaluation provides a scalable and reproducible benchmarking protocol, and its usefulness is further supported by our complementary user study via Amazon Mechanical Turk (MTurk).

Beyond quantitative gains, qualitative results highlight LCCR’s ability to generate images that are both coherent and visually appealing, often exhibiting a seamless blend of reference attributes. In scenarios where competing methods tend to overfit one reference or exhibit unnatural fusion artifacts, our model demonstrates robustness and adaptability. These findings are supported by both human evaluations and the GPT-based scoring mechanism, offering a holistic understanding of performance.

Moreover, LCCR addresses important challenges in controllable multi-reference generation by offering a controllable and interpretable design. The architecture allows for explicit modulation of reference influence, making it adaptable to various real-world applications such as digital art creation, personalized avatar synthesis, and visual storytelling. Our design also lends itself to better transparency and alignment with user intent.

The experimental benchmark we propose further contributes to the field by providing a structured evaluation protocol under human-aligned criteria. The inclusion of subjective dimensions such as aesthetic coherence and imaginative integration pushes the boundary

beyond typical fidelity-focused metrics. By offering a comprehensive assessment suite, we lay the groundwork for future benchmarking efforts that better reflect human-centric values.

In conclusion, LCCR represents a meaningful step forward in the domain of multi-reference generative modeling. It combines architectural innovation, evaluative rigor, and practical relevance. Through both automated and human-centric assessments, we verify its effectiveness in our evaluation setting. We believe that LCCR opens up new directions for controllable, human-aligned generative systems and supports future advancements in this rapidly evolving field.

5.2 Future work

While LCCR achieves promising results in multi-reference generation and sets a new benchmark in human-aligned evaluation, several avenues for future exploration remain.

First, enhancing the interpretability and controllability of fusion mechanisms remains a promising direction. Although LCCR introduces explicit cross-conditioned layers, users still have limited control over how specific attributes from each reference influence the final output. Future work could explore disentangled latent representations or dynamic reference weighting mechanisms that allow fine-grained user control. Such advancements would be especially valuable in applications like creative design or educational visualization, where user-guided content manipulation is essential.

Second, extending the model’s generalization to a broader range of domains—such as medical imaging, scientific illustration, or cultural heritage reconstruction—would increase its applicability and social impact. Each of these domains introduces unique challenges

in terms of semantic alignment and visual fidelity, requiring the model to operate with high precision and interpretability. This may necessitate the integration of domain-specific knowledge into the generation process or the use of hybrid symbolic-neural approaches.

Another potential direction is real-time and interactive generation. Currently, the diffusion-based architecture of LCCR involves a computationally intensive sampling process. Future research could focus on optimizing inference speed via techniques such as latent-space acceleration, distillation, or consistency models. Enabling low-latency interaction would unlock use cases in immersive media, real-time content creation, and human-AI co-design.

Moreover, as multimodal models grow more sophisticated, the boundary between evaluator and generator may become more dynamic. One exciting prospect is the co-training or mutual alignment of generation and evaluation models. A generator that adapts its outputs based on feedback from a self-improving multimodal evaluator could lead to significantly more robust and aligned outputs. This may also enhance safety, as the system could better identify and avoid unintended artifacts or biases.

Additionally, future studies should further examine the ethical implications of personalized and compositional generation. While LCCR allows for greater fidelity and personalization, it also raises questions around consent, identity ownership, and potential misuse. Incorporating privacy-preserving mechanisms or auditable generation traces may help address these concerns. Establishing ethical guidelines and transparency protocols will be critical as such systems move closer to real-world deployment.

From an evaluation perspective, expanding beyond preference-based judgments to include measures of trustworthiness, long-term engagement, or emotional resonance could provide a richer understanding of generative model performance. Integrating affective

computing techniques or longitudinal user studies could support this broader evaluative scope.

Finally, exploring the integration of LCCR into broader multimodal agents represents a long-term vision. As vision-language models evolve toward general-purpose intelligence, the ability to perform nuanced, human-aligned generation from complex prompts and references will be a key component. Embedding LCCR as a generation module within larger embodied or dialog-based systems could enable collaborative and context-aware creativity at scale.

In summary, while LCCR advances the current frontier of controllable multi-reference generation, it also reveals new questions and challenges. By building on its foundations and addressing its limitations, future research can continue to push the envelope toward more general, transparent, and ethically aligned generative AI systems.

References

- [1] Aishwarya Agarwal, Srikrishna Karanam, Tripti Shukla, and Balaji Vasan Srinivasan. An image is worth multiple words: Multi-attribute inversion for constrained text-to-image synthesis. pages 6053–6062, 2025.
- [2] Namhyuk Ahn, Junsoo Lee, Chunggi Lee, Kunhee Kim, Daesik Kim, Seung-Hun Nam, and Kibeom Hong. Dreamstyler: Paint by style inversion with text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 674–681, 2024.
- [3] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science*. [https://cdn. openai. com/papers/dall-e-3. pdf](https://cdn.openai.com/papers/dall-e-3.pdf), 2(3):8, 2023.
- [4] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. 2018.
- [5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision trans-

- formers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021.
- [6] 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. volume 42, pages 1–10. ACM New York, NY, USA, 2023.
- [7] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *International conference on machine learning*, pages 1691–1703. PMLR, 2020.
- [8] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. Ilvr: Conditioning method for denoising diffusion probabilistic models. *arXiv preprint arXiv:2108.02938*, 2021.
- [9] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. pages 4690–4699, 2019.
- [10] Sara Dorfman, Dana Cohen-Bar, Rinon Gal, and Daniel Cohen-Or. Ip-composer: Semantic composition of visual concepts. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pages 1–11, 2025.
- [11] Lunhao Duan, Shanshan Zhao, Wenjun Yan, Yinglun Li, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, Mingming Gong, and Gui-Song Xia. Unic-adapter: Unified image-instruction adapter with multi-modal transformer for image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7963–7973, 2025.

- [12] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.
- [13] Yu Gao, Lixue Gong, Qiushan Guo, Xiaoxia Hou, Zhichao Lai, Fanshi Li, Liang Li, Xiaochen Lian, Chao Liao, Liyang Liu, et al. Seedream 3.0 technical report. *arXiv preprint arXiv:2504.11346*, 2025.
- [14] Zhen Han, Chaojie Mao, Zeyinzi Jiang, Yulin Pan, and Jingfeng Zhang. Stylebooth: Image style editing with multimodal instruction. pages 1947–1957, 2025.
- [15] Amir Hertz, Andrey Voynov, Shlomi Fruchter, and Daniel Cohen-Or. Style aligned image generation via shared attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4775–4785, 2024.
- [16] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clip-score: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021.
- [17] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. volume 30, 2017.
- [18] 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*, 2017.

- [19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. volume 33, pages 6840–6851, 2020.
- [20] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.
- [21] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747, 2023.
- [22] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. volume 34, pages 852–863, 2021.
- [23] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.
- [24] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- [25] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1931–1941, 2023.
- [26] Jaerin Lee, Daniel Sungho Jung, Kanggeon Lee, and Kyoung Mu Lee. Streammulti-diffusion: real-time interactive generation with region-based semantic control. *CoRR*, 2024.

- [27] Tianle Li, Max Ku, Cong Wei, and Wenhua Chen. Dreamedit: Subject-driven image editing. *arXiv preprint arXiv:2306.12624*, 2023.
- [28] Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo numerical methods for diffusion models on manifolds. *arXiv preprint arXiv:2202.09778*, 2022.
- [29] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. pages 423–439, 2022.
- [30] Junsheng Luan, Zhanjie Zhang, Wei Xing, and Lei Zhao. Personalized text-to-image generation with large language and vision assistant enhanced training. *Engineering Applications of Artificial Intelligence*, 161:112116, 2025.
- [31] Naila Murray, Luca Marchesotti, and Florent Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *2012 IEEE conference on computer vision and pattern recognition*, pages 2408–2415. IEEE, 2012.
- [32] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.
- [33] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2085–2094, 2021.
- [34] William Peebles and Saining Xie. Scalable diffusion models with transformers. pages 4195–4205, 2023.

- [35] 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.
- [36] 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.
- [37] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.
- [38] 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.
- [39] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. volume 32, 2019.
- [40] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR, 2015.
- [41] 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.
- [42] 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.
- [43] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Wei Wei, Tingbo Hou, Yael Pritch, Neal Wadhwa, Michael Rubinstein, and Kfir Aberman. Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models. pages 6527–6536, 2024.
 - [44] Runway Research. Introducing runway gen-4. <https://runwayml.com/research/introducing-runway-gen-4>, 2025. Accessed: 2025-11-02.
 - [45] 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. volume 35, pages 36479–36494, 2022.
 - [46] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.
 - [47] Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8543–8552, 2024.
 - [48] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. pages 2256–2265, 2015.
 - [49] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. 2020.

- [50] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. 2023.
- [51] Hossein Talebi and Peyman Milanfar. Nima: Neural image assessment. *IEEE transactions on image processing*, 27(8):3998–4011, 2018.
- [52] Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Omini-control: Minimal and universal control for diffusion transformer. pages 14940–14950, 2025.
- [53] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.
- [54] Xierui Wang, Siming Fu, Qihan Huang, Wanggui He, and Hao Jiang. Ms-diffusion: Multi-subject zero-shot image personalization with layout guidance. *arXiv preprint arXiv:2406.07209*, 2024.
- [55] 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*, pages 7677–7689, 2023.
- [56] Yuxiang Wei, Yiheng Zheng, Yabo Zhang, Ming Liu, Zhilong Ji, Lei Zhang, and Wangmeng Zuo. Personalized image generation with deep generative models: A decade survey. *arXiv preprint arXiv:2502.13081*, 2025.
- [57] Guangxuan Xiao, Tianwei Yin, William T Freeman, Frédo Durand, and Song Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *International Journal of Computer Vision*, 133(3):1175–1194, 2025.

- [58] Zhexiao Xiong, Wei Xiong, Jing Shi, He Zhang, Yizhi Song, and Nathan Jacobs. Groundingbooth: Grounding text-to-image customization. *arXiv preprint arXiv:2409.08520*, 2024.
- [59] Hu Ye, Jun Zhang, Sibio 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.
- [60] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023.
- [61] 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.
- [62] Xulu Zhang, Xiaoyong Wei, Wengyu Zhang, Jinlin Wu, Zhaoxiang Zhang, Zhen Lei, and Qing Li. A survey on personalized content synthesis with diffusion models. *arXiv preprint arXiv:2405.05538*, 2024.
- [63] Guanglu Song Hao Shao Dazhong Shen Yu Liu Hongsheng Li Zhuofan Zong, Dongzhi JiangBingqi Ma. Easyref: Omni-generalized group image reference for diffusion models via multimodal llm. *arXiv preprint arXiv:2412.09618*, 2024.

APPENDICES

A1.1 Purpose of the Human Evaluation

The purpose of the human evaluation is to assess the perceptual quality and semantic fidelity of images generated under multi-reference conditioning. While automated metrics provide scalable and reproducible measurements, they often fail to fully capture human perception of visual coherence, identity preservation, and compositional harmony. Therefore, a controlled user study was conducted to serve as a human-aligned benchmark for evaluating multi-reference image generation.

A1.2 Participants

A total of 50 participants were recruited through Amazon Mechanical Turk (MTurk). All participants were registered MTurk users with prior task approval experience, ensuring a baseline level of familiarity with evaluation tasks. Participants were geographically diverse and no personally identifiable information was collected during the study.

A1.3 Evaluation Interface and Procedure

Participants interacted with a web-based evaluation interface. For each evaluation instance, participants were presented with:

- Two generated image,
- The corresponding textual prompt,
- Two reference images associated with different semantic regions.

The order of methods was randomized across trials to reduce ordering bias. Participants evaluated each image independently and were not subject to time constraints.

A1.4 Evaluation Criteria

Each generated image was evaluated according to the following criteria:

1. **Concept Preservation:** The degree to which the generated image preserves key visual concepts, identities, and attributes from the reference images.
2. **Prompt Following:** The extent to which the generated image accurately reflects the semantic intent and spatial relationships described in the textual prompt.

A1.5 Data Aggregation and Analysis

For each method, scores were averaged across all participants and evaluation samples. Comparative analysis was performed using mean preference rates and standard deviation

statistics. The results demonstrate that the proposed region-aware framework consistently achieves higher human preference scores than baseline methods, indicating improved perceptual quality and semantic alignment.

.1 Appendices 2 for GPT evaluation

A2.1 Motivation

To complement human evaluation and address scalability limitations, a GPT-based evaluation protocol was introduced. This protocol leverages multimodal large language models as automated evaluators, aiming to approximate human judgment while ensuring consistency, reproducibility, and low evaluation cost.

A2.2 Evaluation Inputs

For each evaluation instance, the model was provided with:

- The generated image,
- The original textual prompt,
- The set of reference images used during generation.

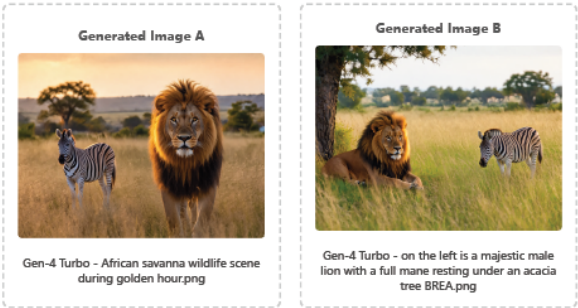
All inputs were presented in a structured and consistent format to ensure fair and unbiased comparison across different methods.

Please answer the following two question based on the references images and text description

Reference image



Generated Image



textDescription

A majestic African savanna wildlife scene during golden hour. A small herd of African elephants slowly crosses the golden grassland, kicking up soft red dust with each step. Acacia trees with flat crowns cast long dramatic shadows across the plains. Warm amber sunlight bathes the landscape from a low angle, illuminating the dust particles in the air. In the background, distant giraffes silhouette against a

Q 1: Which image is more similar to the reference image?

- ☐ Image A
- ☐ Image B

Q 2: Which image is more similar to the prompt?

- ☐ Image A
- ☐ Image B

Q 3: Overall which image do you think is better?

- ☐ Image A
- ☐ Image B
- ☐ I don't know

Please answer the following two question based on the references images and text description

Reference image



owl.png



eagal.png

Generated Image

Generated Image A



Gen-4 Turbo - Majestic birds perched on ancient tree branches.png

Generated Image B



Gen-4 Turbo - on the left is a bald eagle with white head and yellow beak spreading its wings on a h.png

textDescription

Majestic bald eagle with pure white head, sharp yellow beak, and intense pale eyes, wings fully spread in a powerful landing pose. Golden hour sunlight illuminates the individual feathers from behind, creating a dramatic rim light. Dark mountain forest background with shallow depth of field. Slow-motion wing flap, photorealistic,

Q 1: Which image is more similar to the reference image?

- ☐ Image A
- ☐ Image B

Q 2: Which image is more similar to the prompt?

- ☐ Image A
- ☐ Image B

Q 3: Overall which image do you think is better?

- ☐ Image A
- ☐ Image B
- ☐ I don't know

A2.3 Evaluation Prompt Design

The evaluation prompt instructed the model to assess the generated image according to the same two criteria used in the human evaluation:

1. Concept Preservation,
2. Prompt Following.
3. Prompt Following.

To improve reproducibility and reduce output variance, the model was explicitly constrained to output a single integer score between 0 and 4 for each criterion, without providing free-form reasoning or explanatory text.

A2.4 Scoring Protocol

Each image was evaluated independently by the model. Final scores for each method were obtained by averaging across all evaluation samples. The constrained output format ensures deterministic scoring behavior and minimizes subjective drift across evaluations.

A2.5 Correlation with Human Evaluation

Statistical analysis reveals a strong positive correlation between GPT-based evaluation scores and human-assigned scores. This alignment indicates that the proposed automated evaluation protocol serves as a reliable proxy for human judgment in multi-reference image generation tasks, particularly for assessing compositional fidelity and semantic consistency.

A2.6 Limitations

Despite its advantages, GPT-based evaluation remains limited by the perceptual and reasoning biases of current multimodal models. As such, this evaluation protocol is intended to complement rather than replace human evaluation, providing a scalable and standardized benchmark for large-scale experimental analysis.



OK

Please follow the following prompt to measure the quality of this generated image and its reference images





Prompt: Generate an image that contain two reference object, on the left there is an audio player and on the right, there is a laptop





Score: 2

Color & toning: The generated image uses warm, cinematic lighting and richer gradients, while the reference images rely on clean, neutral, product-style lighting. The tonal mood differs noticeably.

Texture & rendering: The generated image emphasizes atmospheric lighting and soft reflections, whereas the references show sharper, more uniform textures typical of studio product photography.

Style representation: The references are minimal, isolated, and commercial in presentation; the generated image is more lifestyle-oriented and expressive. While the objects are similar, the *artistic intent and presentation style* diverge.

Overall: There is a clear thematic connection (audio + tech products), but the stylistic treatment differs enough in lighting, mood, and presentation that the resemblance is **moderate rather than strong**, justifying a **score of 2**.

Please follow the following prompt to measure the quality of this generated image and its reference images



OK



Prompt: Generate an image that contain two reference object, on the left there is an audio player and on the right, there is a laptop



Score: 1

Style mismatch (major):The reference images are clean, isolated, studio product shots with controlled lighting, neutral backgrounds, and commercial catalog aesthetics. The generated image is a casual, cluttered lifestyle scene with desk context, random props, shallow depth-of-field, and informal composition.

Color & toning (inconsistent):References use **neutral whites and balanced tones** for accurate product presentation. The generated image introduces **uneven ambient lighting**, mixed color temperatures, and distracting highlights that break stylistic consistency.

Please follow the following prompt to measure the quality of this generated image and its reference images



OK



The image should be left and right, on the left red sports car with gleaming chrome color, on the right side there should be a boat



Score: 3

Composition & layout:The generated image correctly follows the required left–right structure: a red sports car on the left and a boat on the right. Spatial arrangement matches the reference intent well.

Style consistency:Both objects are rendered in a realistic, commercial showcase style, similar to exhibition or showroom photography. Lighting is controlled, surfaces are clean, and materials (paint, chrome, wood, sail fabric) are clearly defined.

Color & material handling:The red sports car’s glossy finish and the boat’s neutral, natural tones are visually coherent and balanced. While they are different objects, they share a consistent level of realism and polish.

Minor discrepancies:The background context and lighting environment are not perfectly unified between the car and the boat, which slightly weakens stylistic cohesion. This prevents a near-identical (4) score.

101



OK

Please follow the following prompt to measure the quality of this generated image and its reference images





A split-screen interior scene divided vertically into left and right. **On the left**, a cozy, rustic living room with warm wooden beams. On the right, a modern minimalist living space with large floor-to-ceiling black-framed windows





Score: 2

Visual confusion at the split boundary: Although the image is intended to be a clean left–right split, the transition between the rustic and modern spaces is not sufficiently clear. Elements (lighting, furniture scale, and spatial depth) visually bleed across the division, weakening semantic separation.

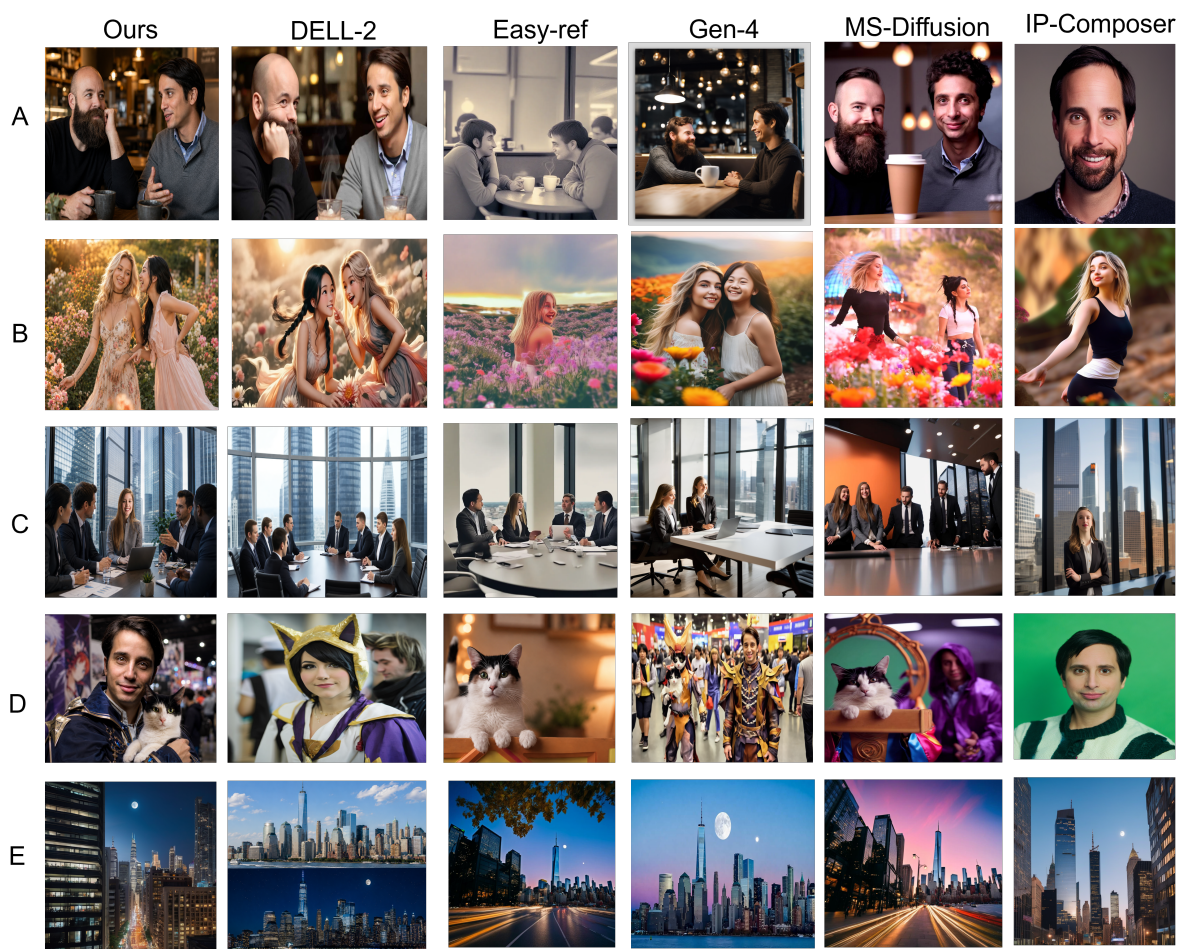
Style mixing reduces clarity: The left side’s warm, rustic cues (wood beams, fireplace glow) and the right side’s modern minimalist cues (black frames, clean surfaces) are present, but not isolated enough. This makes the scene feel like a single hybrid interior rather than two distinctly contrasted styles.

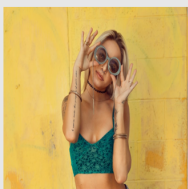
Lighting inconsistency: The lighting temperature and direction are too similar across both sides. A stronger contrast (warmer, dimmer left vs. cooler, brighter right) is expected for clear stylistic differentiation

.Semantic intent vs. execution gap: The concept is correct, but the execution lacks visual discipline. For a high score, viewers should instantly and unambiguously read “two different interior styles.” Here, that reading requires effort.

A2.7 More examples and meta data

102



	Reference image 1	Reference image 2	Prompt
A			Two men sit across from each other in a cozy café, engaged in a warm and relaxed conversation. On the left, a bald man with a thick full beard wears a black long-sleeve shirt, resting his cheek on one hand while holding a dark coffee mug with the other. He looks thoughtfully toward his companion with a slight smile. On the right, a dark-haired man wearing a gray sweater over a light blue collared shirt smiles as he speaks, gesturing naturally with one hand. The wooden table in front of them holds two dark ceramic mugs and a small pastry, while the softly blurred café background, warm hanging lights, and intimate indoor atmosphere create a friendly and comfortable mood.
B			Two men are sitting at a wooden table in a warm, softly lit café, having a relaxed conversation. The man on the left is bald with a thick beard, wearing a black long-sleeve shirt. He rests his head on one hand while holding a dark coffee mug, looking thoughtfully at the other man. The man on the right has short dark hair and wears a gray sweater over a light blue collared shirt. He smiles and gestures with one hand as he speaks. Two dark mugs, a small pastry, and a cozy blurred café background with warm hanging lights create an intimate and friendly atmosphere.
C			A group of business professionals sit around a conference table in a modern high-rise office, engaged in a strategic discussion. At the center, a smiling woman in a gray blazer sits behind an open laptop, appearing confident and attentive. Around her, several colleagues in dark suits listen and participate, with one man on the right gesturing as he speaks. The table is arranged with notebooks, printed charts, pens, glasses of water, and a small potted plant. Floor-to-ceiling windows reveal tall city skyscrapers outside, creating a polished corporate atmosphere of collaboration, professionalism, and modern business efficiency.
D			A man in an elaborate blue cosplay costume holds a black-and-white cat close to his chest at a lively anime convention. The man has dark hair and looks calmly toward the camera, wearing a detailed outfit with metallic shoulder armor, gold chains, and decorative fantasy-style accents. The cat faces forward with wide green eyes, its white fur and black markings clearly visible as it rests comfortably in the man's arms. The background is softly blurred, showing colorful anime posters, convention lights, and people moving through the crowded event space, creating a playful and energetic atmosphere.
E			A nighttime city skyline is seen from an elevated perspective between tall office buildings. A bright full moon hangs in the deep blue sky above illuminated skyscrapers, while rows of windows glow across the surrounding towers. In the center, a long avenue stretches into the distance, filled with warm streetlights and moving traffic that create a bright line through the city. The mix of dark building silhouettes, lit windows, and the moonlit sky gives the scene a calm yet vibrant urban atmosphere.