TOWARDS SPATIALLY GROUNDED DIFFUSION MODELS FOR LAYOUT-CONTROLLABLE IMAGE GENERATION

by

Susim Mukul Roy

May 11 2026

A dissertation submitted to the
Faculty of the Graduate School of
the University at Buffalo, State University of New York
in partial fulfilment of the requirements for the
degree of

Master of Science

Department of Computer Science and Engineering

Acknowledgments

I would like to express my sincere gratitude to the Computer Science Department of University at Buffalo and Dr. Nalini Ratha for providing me with the invaluable opportunity to pursue my research interests in the field of Computer Vision, with a particular focus on Image Generation. Their guidance, encouragement, and support have been instrumental in shaping my academic and research journey. I am also deeply thankful to my lab mates - Bharat Chandra Yalavarthi, Arjun Ramesh Kaushik, and Sai Bharadwaj Sirigadi - for their constant support, thoughtful discussions, and timely assistance. Their collaboration and encouragement greatly enriched both my research experience and personal growth throughout this endeavor.

Table of Contents

List of Tables

List of Figures

Abstract

Diffusion models have demonstrated remarkable success in image generation across a variety of settings, particularly in controlled settings. Recent work has shown that one can generate subjects in desired locations using either guiding the attention signal to attend to specific tokens or framing optimization objectives that penalize subjects or attributes being generated outside the bounding boxes. However, the role of memorization, which occurs during training, in the inference process for conditional generation has been largely unexplored. In this work, we take the first step in understanding how memorized subject characteristics incur a heavy toll during the reverse process. Based on the observations, we build a coherent pipeline that automatically detects which regions to suppress as they drift towards a particular attribute, followed by showing that we can mitigate it promptly, thereby giving the diffusion model enough timesteps to appropriately fill it. We demonstrate the validity of our ideas through experiments on the specific task of layout-to-image generation and show that we can achieve competitive performance on publicly available datasets without requiring computationally expensive fine-tuning. Our results demonstrate consistently superior quantitative performance across multiple baselines, encompassing both multi-subject and overlapping-subject configurations, while maintaining photorealistic fidelity free of visually incoherent or perceptually implausible artifacts.

Chapter 1

Introduction

Recent advances in generative AI have significantly pushed the boundaries of visually coherent and structurally grounded image synthesis. Early approaches, including GANs[52, 57], Variational Autoencoders[56, 50], and autoregressive models[64, 63, 36, 30], laid the groundwork for high-fidelity generation, with diffusion models[48, 1, 49, 10] subsequently establishing a new state of the art. These advances have been further accelerated by controllable image synthesis methods, particularly Text-to-Image (T2I) generation[65, 67, 60], which leverage large-scale pretrained models on image-text paired data to enable semantically rich synthesis. Notable models such as Stable Diffusion[47], GLIGEN[35], and DALL-E[45] are among the popular ones to have demonstrated impressive generation quality under text-only conditions, yet remain constrained in their ability to support fine-grained, multimodal control signals beyond language.

Contemporary research has elaborately demonstrated their inability to process textual information while simultaneously being vigilant to maintain the spatial consistency [5, 21] of the objects being generated. Text processing is usually done through pretrained large language models[29] as a text encoder[44], as it needs to harness their deep level of language understanding. Nevertheless, it still falls short since relying solely on text leads to a semantic gap between images and text, which hinders the spatial understanding of the

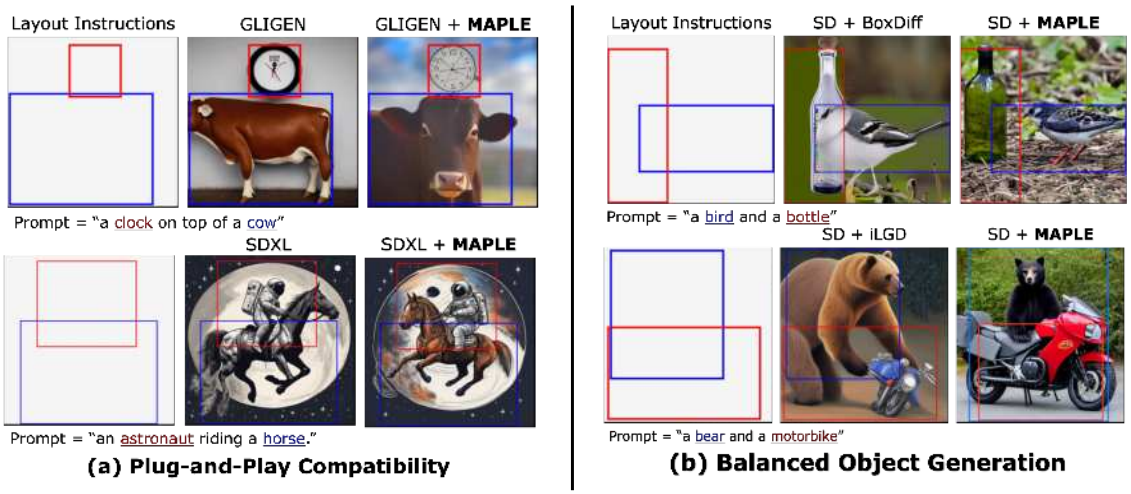


Figure 1.1: Qualitative illustration of MAPLE's two key properties. (a) **Plug-and-play compatibility**: MAPLE integrates seamlessly into existing models such as GLIGEN, correcting object placement without any additional training. (b) **Balanced object generation**: In the presence of competing objects, MAPLE mitigates inter-object dominance arising from memorized spatial priors. Given the text prompts, both BoxDiff and iLGD exhibits object dominance where the bird suppresses the bottle or the bear suprreses the motorbike, whereas MAPLE produces a spatially balanced and semantically faithful generation.

generation task.

To address these limitations, a growing body of work has explored Layout-to-Image Synthesis(LIS)[66, 14, 20, 61], wherein spatial bounding box layouts are introduced alongside text prompts to provide explicit localization guidance for object placement. This additional conditioning can be incorporated into existing frameworks via two principal paradigms: supervised fine-tuning[33, 55, 13], which augments pretrained models with additional control modules trained on paired image-text data, and training-free approaches[69, 22, 16], which instead manipulate cross-attention maps to steer the generation process without any parameter updates. While the former has demonstrated strong performance, it incurs substantial computational overhead and relies heavily on paired layout-text datasets such as COCO[37], which are both scarce and costly to curate. The latter circumvents these constraints by extracting and modulating cross-attention maps across the layers of the UNet, offering a more accessible and flexible alternative. A complementary line of work focuses on optimizing intermediate latent representations by quantifying the alignment between

text token attention and target spatial regions. Beyond latent optimization, a parallel direction targets the initial Gaussian noise[53, 23], operating on the premise that the seed distribution itself plays a critical role in determining object localization in the final generated output.

In light of the above advancements, our objective is to dynamically identify the degree to which individual objects are either dominated or neglected during the generation process, and to rectify this imbalance in a spatially informed manner. We hypothesize that memorized spatial priors accumulated during large-scale pretraining cause T2I models to systematically over-attend or under-attend to specific objects with respect to their prescribed bounding box layout, manifesting as object amplification or neglect in the final output. To validate this, we conduct a thorough analysis of the cross-attention and self-attention maps within the diffusion model, demonstrating that these maps serve as reliable indicators of inter-object dominance and semantic misalignment. Guided by these observations, we derive a principled optimization objective, $\boldsymbol{L_{MAPLE}}$, that quantifies the degree of attention imbalance across objects with respect to the provided layout.

Upon identifying the dominated and neglected objects, we employ a combination of attention reweighting, Mask Guidance(**MG**), and Latent Optimization($\boldsymbol{L_{SAE}}$ and $\boldsymbol{L_{GAP}}$) to correct this imbalance at inference time. Specifically, mask-based guidance spatially constrains the influence of each object to its designated region and prevents semantic leakage, while attention reweighting redistributes semantic focus across objects in proportion to their prescribed layout. These corrections are consolidated into a modified latent representation at inference, requiring no additional training or parameter updates. Our method, ***MAPLE***, as shown in Figure 1.1, is training-free and plug-and-play, integrating seamlessly into existing pretrained diffusion models and enabling accurate, controllable layout-conditioned image synthesis. We also conduct comprehensive experiments, comparing our method with various approaches in the training-free layout-to-image synthesis literature. Our results demonstrate state-of-the-art performance, showcasing significant

improvements both quantitatively and qualitatively over prior methods. In summary, our contribution can be summarized as follows:

- We introduce **MAPLE**, a memorization-aware, training-free framework for layout-conditioned image synthesis that operates on text prompts and bounding box specifications as spatial constraints.
- We propose two complementary mechanisms: *fine-grained loss guidance* via $\boldsymbol{L_{MAPLE}}$, which dynamically identifies and rectifies inter-object dominance and neglect during the generation process, and *coarse attention steering* via **MG**, which complements this by capturing individual and overall semantic structure and redistributing focus toward underperforming objects accordingly as shown in Figure 2.1.
- We conduct extensive experiments on [62] for quantitative superiority on well-known metrics. Furthermore, we qualitatively compare against prior art on different levels of prompts and corresponding layouts.

Chapter 2

Literature Review

2.1 Diffusion Models: An Introduction

We briefly introduce the Denoising Diffusion Probabilistic Model (DDPM) [27], which operates via a forward noising process and a learned reverse denoising process. Let $\{x_t\}_{t=0}^{T}$ denote a discrete Markov chain, where $t \in \{0, 1, \ldots, T\}$ is the timestep index and T is the total diffusion length. The **forward process** gradually corrupts the data by adding Gaussian noise according to:

$$q(x_t \mid x_{t-1}) \;=\; \mathcal{N}\Big(x_t;\; \sqrt{1-\beta_t}\, x_{t-1},\; \beta_t \mathbf{I}\Big)\,, \tag{2.1}$$

where β_t is the noise schedule (*e.g.*, linear or cosine). Defining $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$, a closed-form expression for sampling x_t directly from x_0 is:

$$q(x_t \mid x_0) \;=\; \mathcal{N}\big(x_t;\; \sqrt{\bar{\alpha}_t}\, x_0,\; (1-\bar{\alpha}_t)\mathbf{I}\big)\,. \tag{2.2}$$

The **reverse process** learns to denoise by predicting x_{t-1} from x_t:

$$x_{t-1} \;=\; \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\, \epsilon_\theta(x_t, t) \right) + \sigma_t\, \mathbf{z}, \qquad \mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{2.3}$$

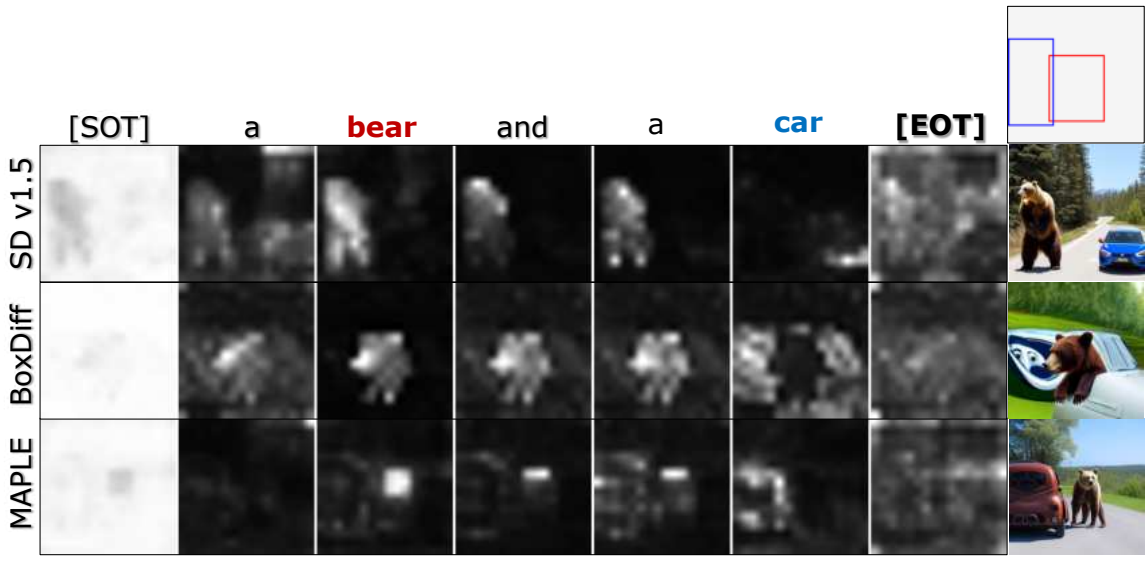


Figure 2.1: The top and bottom row shows the cross-attention maps of the text tokens on the different spatial regions in the h-space or layer 14 of Stable Diffusion at t=19, which is the final step for our Mask Guidance.

where $\sigma_t = \sqrt{\beta_t(1-\bar{\alpha}_{t-1})/(1-\bar{\alpha}_t)}$ is the posterior standard deviation and ϵ_θ is a neural network parameterized by θ. The model is trained by minimizing the simplified denoising objective derived from the variational lower bound:

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{t,\, x_0,\, \epsilon}\Big[\big\|\epsilon - \epsilon_\theta\big(\sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1-\bar{\alpha}_t}\,\epsilon,\ t\big)\big\|_2^2\Big]. \tag{2.4}$$

2.2 Diffusion Models and Controllability

The landscape of T2I synthesis is currently defined by the success of Diffusion Models (DMs), which generate high-fidelity samples by learning to reverse a gradual Gaussian denoising process[51, 18]. A pivotal advancement in this domain was the introduction of Latent Diffusion Models (LDM) [47], which utilize a pre-trained autoencoder to perform the diffusion process in a low-dimensional latent space z after which the latent is decoded back to the required resolution, significantly reducing computational complexity

while maintaining visual quality. These models typically optimize a denoising objective:

$$\mathcal{L}_{LDM} = \mathbb{E}_{\mathcal{E}(x),y,\epsilon,t}\left[\|\epsilon - \epsilon_\theta(z_t, t, \tau_\theta(y))\|^2\right] \tag{2.5}$$

where $\tau_\theta(y)$ represents conditioning embeddings from encoders like CLIP[44]. While scaling efforts such as SDXL[43] and the adoption of Diffusion Transformers (DiT) [40, 11] have further enhanced photorealism, precise spatial control remains a significant challenge. Architectures like ControlNet [68] address this by incorporating auxiliary spatial priors such as segmentation masks[15] or depth masks[54] into frozen T2I backbones to provide structural guidance. However, despite these advancements, existing models frequently struggle with accurately binding attributes to specific objects and fulfilling complex positional descriptions from text prompts alone [6]. This highlights a critical gap in the inherent spatial intelligence and semantic grounding of current diffusion frameworks, necessitating new approaches that can bridge the divide between high-fidelity generation and precise compositional control.

2.3 Memorization in Diffusion Models

There has been a substantial body of recent work on the memorization phenomenon in diffusion models[8, 41, 19, 26, 7], identifying one of its primary failure modes as the inability of the reverse process to escape learned attraction basins[31]. A closely related line of inquiry into concept dominance[32] demonstrates that the visual characteristics of individual objects systematically suppress those of co-occurring objects during multi-object generation. Complementing this, [24] proposes adjusting the initial noise distribution such that the sampling trajectory exits the attraction basin at an earlier stage, while [2] establishes that specific noise patches in the initial sample are causally responsible for determining object generation locations. Collectively, these works provide compelling evidence for the rigidity of the reverse diffusion process and its resistance to user-prescribed spatial

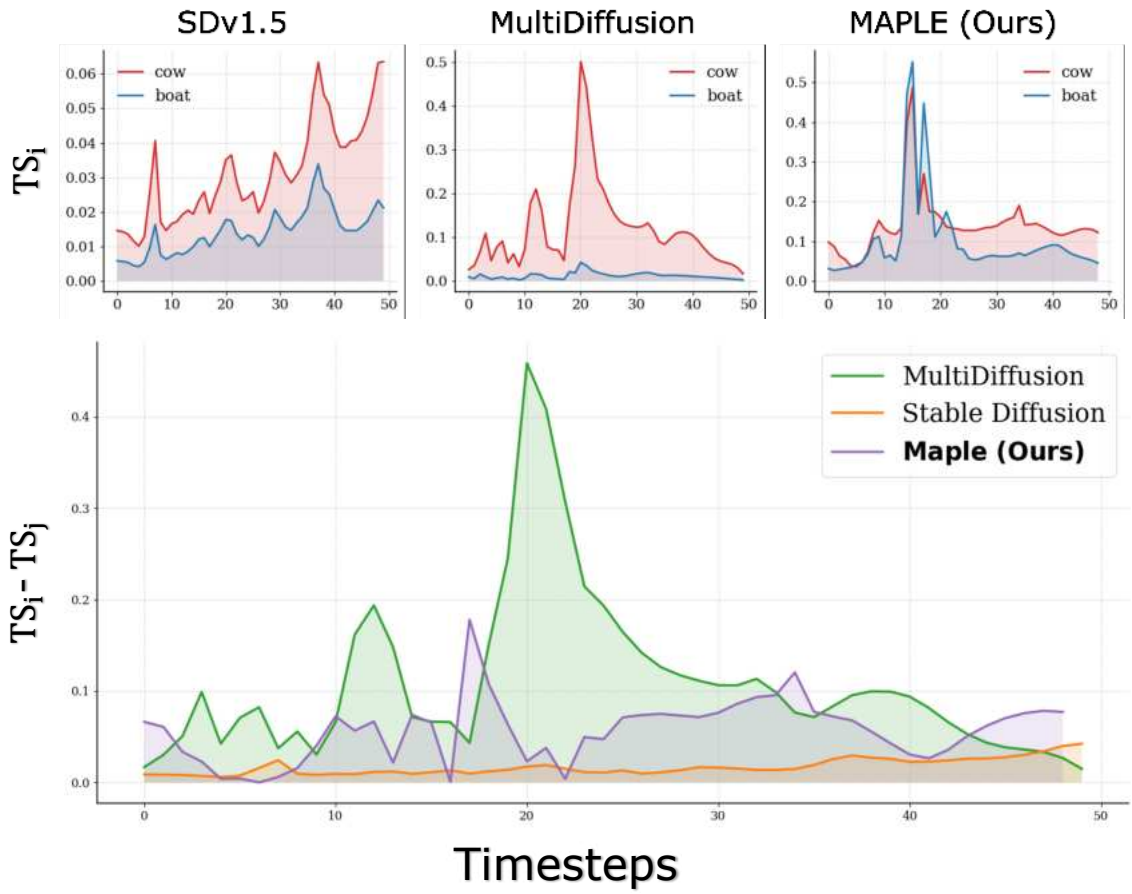


Figure 2.2: The top three plots demonstrate the gradient norm of two subjects, cow and boat, in the prompt "a cow and a boat", over all the timesteps. The bottom plot shows the absolute difference of the gradient norm between the two subjects.

constraints. However, despite this growing body of evidence, only a limited number of works, notably [23], which also has not open-sourced its integration for LIS, have sought to leverage these insights toward addressing the practical challenge of LIS.

2.4 Layout-to-Image Synthesis

LIS aims to generate images that faithfully adhere to both textual prompts and spatial layout instructions, such as bounding boxes or semantic masks. A prevalent approach to achieving this involves extending pre-trained diffusion models, like Stable Diffusion, by fine-tuning them on paired layout-image datasets[38]. For instance, SceneComposer[66] trains

a layout-to-image model using paired images and segmentation maps, while GLIGEN[35] fine-tunes a gated self-attention layer to incorporate bounding box information directly into the model. Other strategies integrate specialized adapters or additional components to achieve layout control [70, 59]. While these fully-supervised, training-based pipelines demonstrate noteworthy generative results, they inherently suffer from significant drawbacks. Primarily, they grapple with the labor-intensive and time-consuming curation of paired datasets, alongside high computational resource consumption and prolonged inference times.

To circumvent these heavy training requirements, a growing body of work focuses on training-free layout guidance by directly modifying the generation process [42, 62, 9, 39]. A foundational insight driving these methods was highlighted by works like Prompt-to-prompt[25], which showed that there is a strong intrinsic correlation between the spatial layout of generated content and the model's cross-attention maps. Leveraging this, approaches such as [62] manipulate or optimize the cross-attention layers during sampling to align the generated visuals with the provided spatial layouts. Other techniques like MultiDiffusion [3] perform regional denoising processes and fuse the predicted scores. Despite their efficiency, strictly optimizing multiple values within attention maps can sometimes degrade overall image quality or fail to reliably meet strict positional requirements.

Chapter 3

Method

3.1 Preliminaries

Cross-Attention Layers. To ensure cross-modal alignment between the user-given text prompts, which is treated as the condition, and the image to be generated, the text is first tokenized, followed by encoding using CLIP[44]. Specifically, given the condition **y**, we first obtain a set of text tokens as an embedding **e** = $\tau_\theta(y) \in \mathbb{R}^{N \times d_e}$, where d_e is the embedding dimension and N is the number of tokens. Since the text can be of variable size, it is usually padded to a fixed length p. Thereafter, it is fused in the attention layers of the UNet via the cross-attention layers via the key, $K \in \mathbb{R}^{p \times d_k}$, and value, $K \in \mathbb{R}^{p \times d_v}$ matrices while the query matrix, $Q \in \mathbb{R}^{rs \times d_q}$ comes from the latent at the current timestep z_t. Through this, we get the cross-attention map for a single head of the lth layer as:

$$A^{ca}_{l,p} = \text{softmax}(\frac{QK^T}{\sqrt{d}}) \in \mathbb{R}^{rs^2 \times p}$$

Here, rs is the resolution of the attention map at the lth layer. $A^{ca}_{l,p}$ is therefore a set of cross-attention maps for the prompt y, defined as $A^{ca}_l = \{A^{ca}_{l,0}, \ldots, A^{ca}_{l,p}\}$ where the $A^{ca}_{l,i}$ denotes the cross-attention map for the ith text token.

Self-Attention Layers. During the denoising process, objects should learn to only attend

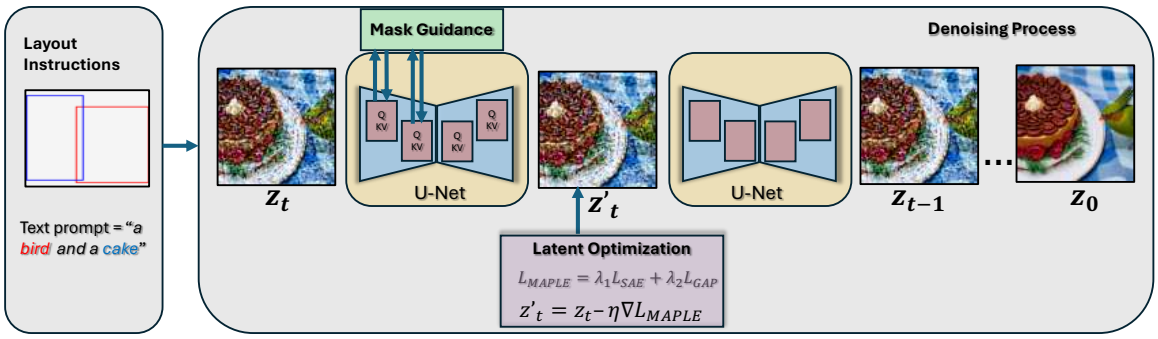


Figure 3.1: Overview of the MAPLE Framework. At each denoising timestep t, the pipeline operates in two stages. First, cross and self attention maps are extracted from the U-Net and passed to the Mask Guidance module, which generates subject-specific spatial masks that are fed back into the attention layers while simultaneously computing the layout losses L_{SAE} and L_{GAP}. Second, Latent Optimization combines these losses into a unified objective L_{MAPLE}, and updates the current latent z_t via a gradient step to produce the refined latent z_t, which seeds the next denoising step.

to themselves to prevent semantic leakage. We define the corresponding self-attention map at the lth layer as:

$$A^{sa}_{l,rs^2} = \text{softmax}(\frac{QK^T}{\sqrt{d}}) \in \mathbb{R}^{rs^2 \times rs^2}$$

Here, both the query $\mathbf{Q} \in \mathbb{R}^{rs^2 \times d_q}$ and key $\mathbf{K} \in \mathbb{R}^{rs^2 \times d_k}$ is derived from the latent z_t.

Bounding Boxes. Additionally, if we assume there are K_1 subjects in the prompt, then we are given a set of bounding boxes for subjects in the prompt y as $\mathrm{B} = \{B_0, B_1, \ldots, B_{K_2}\}$. Here, we note two changes from prior work. First, K_1 may or may not be equal to K_2. Secondly, we only focus on the attention layers at the downsampling and middle blocks, as the former is known to control the style and appearance, and the latter controls the semantics and shape.

3.2 Text Token Impact Scoring

To effectively generate objects in their respective locations, the model should implicitly and adaptively learn to give importance to the objects that are failing to either produce in the assigned location, with the right semantics, or with the correct size, shape, etc. Previous

studies[32, 46] have shown that certain objects tend to overshadow the other object during generation from the very early timesteps, as a result of which the dominated object does not recover. This effect can be seen in Figure 2.1 where the $bear$ is excessively being attended to, even by the filter tokens. This is particularly harmful for our cause as it may be the reasons to many problems found in the previous works, such as object disappearance, semantic leakage, misplacement of the objects, unrealism, etc. To analyze this, we use the following to measure the significance score for each token at position $i \in [0, \ldots, N-1]$:

$$TS_i = \|\nabla_{e_i} L(z_t, e)\|_2$$

$$\text{where } L(z_t, e) = \|\epsilon_\theta(x_t, e) - \epsilon_\theta(x_t, e_\emptyset)\|_2$$

A careful analysis of [16], as observed in Figure 2.2, shows that these methods inadvertently minimize $L(z_t, e)$ for the subject, which dominates the generated image. This can provide an additional explanation as to the reason behind these methods working. In our work, we use this as an indicator for the object to focus on during the optimization process. We hypothesize that one does not need to focus on all the objects being generated, as one of them can more easily be generated while the model finds it hard to generate the other object. Therefore, as long as we can dynamically change our object of focus, we should more efficiently create the desired image.

3.3 Mask Guidance

Following the works of [16, 39, 12], we guide both the self and cross attention layers towards the desired layout structure through masks. However, different from both, we develop our own mask based on the observations from [9, 34] that the attention guidance needs to be in harmony with the diffused state of the noise maps in the early steps and the coarse-grained structure of the cross-attention maps. Let us define the mask $\mathbf{M}$ which is added to the lth layer attention block. $\mathbf{M}$ contains ones in the desired objects locations,

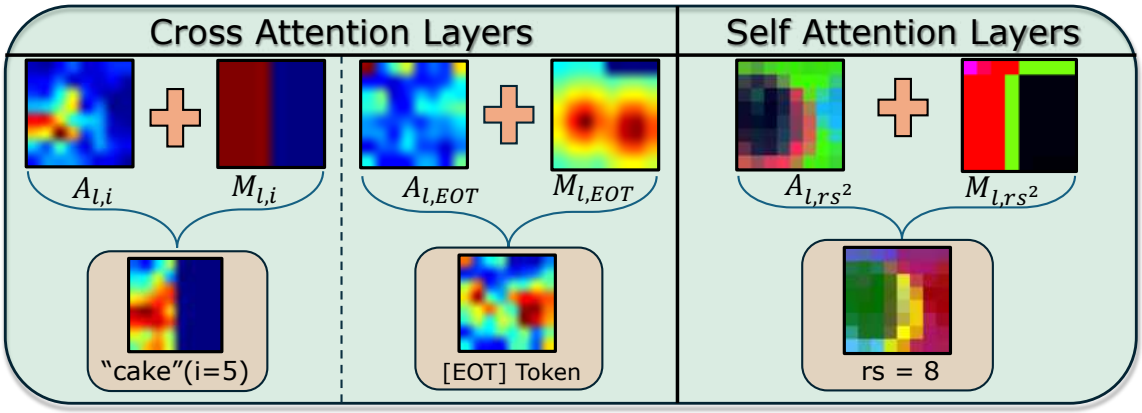


Figure 3.2: Mask Guidance.(Left) Denotes the visualization of the modulation to the cross-attention layers for a specific subject token and the $[EOT]$ token at $l = 13$ for the last timestep after which MG is stopped. (Right) Visualizes the update on self-attention maps, where we see that the cake is only attending to its regions, while the $bird$ region is dark.

i.e., $M[B_i, i] = 1$ while $M[\sim B_i, i] = 0$. We also set all the pixel-locations for the tokens other than the K_1 subject tokens to $-\infty$ in $\mathbf{M}$. We also note the importance of the $[EOT]$ as its role is to capture the overall semantics of the image at each timestep. It has also been used previously in works like [69] for noise optimization, while [12] empirically found it to affect the attention maps too aggressively. We argue that it is an effect of the naive linear interpolation, and when altered carefully, as in our formulation, we can obtain significant control over the image. Formally, let $\boldsymbol{S}$ denote the scaling applied over the N attention tokens. We begin by defining the initial scaling factor as follows:

$$S_1 = \alpha \cdot \log(1 + \sigma_t) \cdot \max_j(QK^T) \tag{3.1}$$

This scaling factor $\boldsymbol{S_1}$ serves to direct the model's attention toward the maximum object signal within each designated region. However, assigning a uniform signal to all spatial locations within a given bounding box is insufficient, as it disregards the inherent shape and fine-grained visual characteristics of the subject. To address this, we introduce a spatially-aware decay function:

$$S_2 = S_1 \cdot \beta e^{-\frac{(x-x_0)}{D}} \tag{3.2}$$

This decay serves a dual purpose. First, it progressively attenuates the attention signal as spatial locations deviate from the centroid x_0 of the bounding box B_i, ensuring that the highest semantic focus is concentrated at the centroid. Second, it enforces a smooth transition at the boundaries of the bounding boxes, mitigating abrupt discontinuities in the attention distribution. Here, D denotes the maximum distance between the centroid x_0 and the farthest spatial location within the respective bounding box B_i.

Given that the `[EOT]` token captures aggregated semantic information across the entire prompt, we apply $\boldsymbol{S_2}$ to scale the spatial locations of each subject's bounding box within the `[EOT]` token of the mask M, while suppressing background token contributions to zero. This formulation additionally addresses the challenge of overlapping bounding box regions, wherein mixed attention signals across subjects can induce undesired semantic contamination. Under our approach, provided that the degree of overlap does not substantially subsume the region of either subject, the decay function facilitates a gradual transfer of the attention signal across the overlap boundary, enabling more precise and controllable subject separation. The final modified attention map is consequently defined as:

$$\begin{aligned} S &= [S_1, S_2] \\ A'_{l,p} &= \text{softmax}\left(\frac{QK^T + S \cdot M}{\sqrt{d}}\right) \end{aligned} \tag{3.3}$$

We perform these updates for the first 20 timesteps only.

3.4 Latent Optimization

The main purpose of this step is to backpropagate through the diffusion model and steer the latent signal towards the layouts. To formalize, we want to minimize the log-likelihood

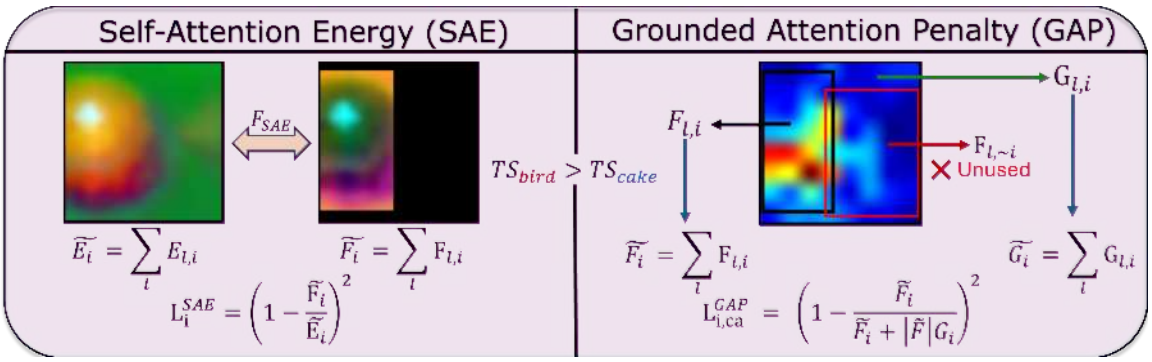


Figure 3.3: Latent Optimization.(Left) We regularize the self-attention maps by constraining each subject's attention to remain within its designated layout region, preventing cross-region attention leakage between subjects. (Right) We regularize the cross-attention maps by enforcing token-level spatial exclusivity, restricting each text token's attention distribution to its assigned layout region while penalizing activations that spread into background areas.

$\log p(z_t|y)$ and from Bayes Theorem, we obtain $\log p(z_t|y) \propto \log p(z_t) \log p(y|z_t)$ which brings us to minimizing $\log p(y|z_t) := L(z_t|y)$.

Grounded Attention Penalty. For each subject s_i, we accumulate attention mass inside its target box and in the global background, then form an IoU-style ratio. Let $\mathcal{B} = \{B_1, \ldots, B_{K_2}\}$ be the set of subject boxes and define the background region as $\Omega_{bg} = \Omega \setminus \bigcup_{j=1}^{K_2} B_j$, where Ω is the spatial grid at the current attention resolution. For cross-attention, let $\mathcal{T}_i$ be the token indices of subject s_i (in our implementation, we also include the EOS token). For a selected layer l, with attention tensor $A_l^{ca} \in \mathbb{R}^{H \times |\Omega| \times P}$ (heads $\times$ pixels $\times$ tokens), we define:

$$F_{l,i}^{ca} = \sum_{h=1}^{H} \sum_{x \in B_i} \sum_{t \in \mathcal{T}_i} A_l^{ca}[h, x, t], \qquad G_{l,i}^{ca} = \sum_{h=1}^{H} \sum_{x \in \Omega_{bg}} \sum_{t \in \mathcal{T}_i} A_l^{ca}[h, x, t]. \tag{3.4}$$

For self-attention, using $A_l^{sa} \in \mathbb{R}^{H \times |\Omega| \times |\Omega|}$, we use spatial indices of the same subject region as keys/columns:

$$F_{l,i}^{sa} = \sum_{h=1}^{H} \sum_{x \in B_i} \sum_{y \in B_i} A_l^{sa}[h, x, y], \qquad G_{l,i}^{sa} = \sum_{h=1}^{H} \sum_{x \in \Omega_{bg}} \sum_{y \in B_i} A_l^{sa}[h, x, y]. \tag{3.5}$$

Averaging across the optimizing layers gives:

$$\tilde{F}_i^\star = \frac{1}{|\mathcal{L}_\star|} \sum_{l \in \mathcal{L}_\star} F_{l,i}^\star, \qquad \tilde{G}_i^\star = \frac{1}{|\mathcal{L}_\star|} \sum_{l \in \mathcal{L}_\star} G_{l,i}^\star, \tag{3.6}$$

where $\star \in \{ca, sa\}$. Thereby, we define the IoU-style score as:

$$L_i^{GAP} = \left(1 - \frac{\tilde{F}_i^{ca}}{\tilde{F}_i^{ca} + |\mathcal{B}|\,\tilde{G}_i^{ca}}\right)^2 + \left(1 - \frac{\tilde{F}_i^{sa}}{\tilde{F}_i^{sa} + |\mathcal{B}|\,\tilde{G}_i^{sa}}\right)^2, \tag{3.7}$$

This represents the per-subject loss for token-aware localization.

Self-Attention Energy. We use a self-attention energy regularizer that measures how much of a subject's self-attention mass lies inside its assigned box. This is to restrict it from attending outside the box, even if there is a corresponding higher attention value. For subject s_i at layer l, we define total self-attention energy and in-box self-attention energy as:

$$E_{l,i}^{sa} = \sum_{h=1}^{H} \sum_{x \in \Omega} \sum_{y \in B_i} A_l^{sa}[h, x, y], \qquad F_{l,i}^{sa} = \sum_{h=1}^{H} \sum_{x \in B_i} \sum_{y \in B_i} A_l^{sa}[h, x, y]. \tag{3.8}$$

Averaging over the selected self-attention layers gives

$$\tilde{E}_i^{sa} = \frac{1}{|\mathcal{L}_{sa}|} \sum_{l \in \mathcal{L}_{sa}} E_{l,i}^{sa}, \qquad \tilde{F}_i^{sa} = \frac{1}{|\mathcal{L}_{sa}|} \sum_{l \in \mathcal{L}_{sa}} F_{l,i}^{sa}. \tag{3.9}$$

Following our implementation, the self-attention energy loss is

$$L_i^{SAE} = \left(1 - \frac{\tilde{F}_i^{sa}}{\tilde{E}_i^{sa} + \epsilon}\right)^2, \tag{3.10}$$

where ϵ is a small constant for numerical stability.

As described in Section 3.2, the L_{MAPLE} is calculated for only the subjects that are being dominated by the other subjects, and the latents are being updated only with them. This is

determined by the following formulation:

$$\mathcal{I}_{sub} = \bigcup_{j=1}^{n} \mathcal{T}_j, \qquad \hat{m} = \arg\max_{i \in \mathcal{I}_{sub}} \left\| \nabla_{e_i} L(z_t, e) \right\|_2, \tag{3.11}$$

where $\mathcal{T}_j$ is the token-index set of subject s_j. Let $\pi(i)$ map token index i to its subject index, and define the dominant subject as $k^\star = \pi(\hat{m})$. We then optimize all non-dominant subjects:

$$\mathcal{S}_{opt} = \{1, \ldots, K_1\} \setminus \{k^\star\}, \tag{3.12}$$

with aggregated losses

$$L_{GAP} = \frac{1}{|\mathcal{S}_{opt}|} \sum_{j \in \mathcal{S}_{opt}} L_j^{GAP}, \qquad L_{SAE} = \frac{1}{|\mathcal{S}_{opt}|} \sum_{j \in \mathcal{S}_{opt}} L_j^{SAE}. \tag{3.13}$$

Our final optimization objective is as follows:

$$L(z_t|y) := L_{MAPLE} = \lambda_1 L_{GAP} + \lambda_2 L_{SAE} \tag{3.14}$$

where the λ_1, λ_2 hyperparameter controls the strength of each loss function. With this selection, we update the latent at the current timestep as:

$$z_t' = z_t - \eta \nabla_{z_t} L_{MAPLE} \tag{3.15}$$

This is done for all the timesteps. With our methodology, we can maintain high attention values over the underdeveloped regions while adhering to the layout. This process leads to the synthesis of desired objects in the user-provided locations.

Chapter 4

Experiments

4.1 Experimental Setup

Dataset, Benchmarks and Hyperparameters. To evaluate our method, we follow [62] and use a dataset consisting of 200 unique prompts spanning 27 object categories, where each prompt is paired with 15 to 16 bounding box configurations representing different spatial arrangements, yielding 2821 prompt and bounding box pairs in total. Each prompt follows either the structure "a {} . . . ," or "a {} and a {} . . . ." Additionally, following [39], we kept the same settings for the benchmark models as them. We compare our proposed method, **MAPLE**, against BoxDiff[62], [12], iLGD[39], MultiDiffusion[3], and standard Stable Diffusion v1.5. We utilize the DDIM scheduler with 50 timesteps, setting the beta values to start at 0.00085 and end at 0.012. We utilize the classifier-free guidance[28] for MAPLE and all baselines with a fixed guidance scale of 7.5. Finally, we use the same negative prompt as [69]. We set $\lambda_1 = 5$, $\lambda_2 = 3$, $\beta = 1.2$, $\alpha = 0.75$ and τ follows the cosine scheduling with the start step and end step being 1 and 50, respectively.

Evaluation Metrics. To assess performance across multiple dimensions, we employ a diverse set of evaluation metrics. For text-to-image alignment, we use the T2I-Sim metric, which measures cosine similarity between text and image features in CLIP space to

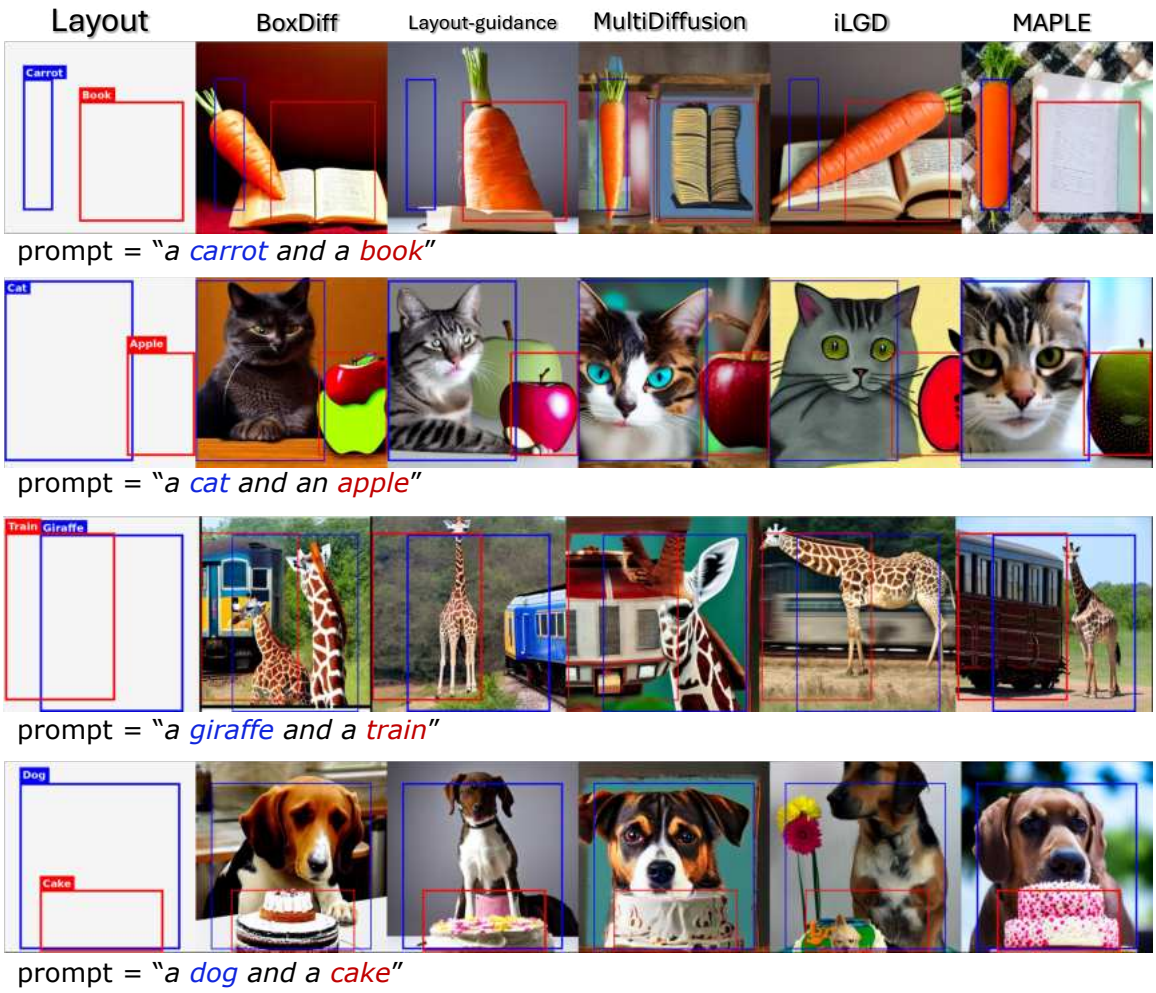


Figure 4.1: Visual Comparisons with Previous Methods. We compare our method against training-free layout-to-image baselines, with layout instructions indicated by solid bounding boxes. Our approach consistently outperforms prior methods in spatial layout adherence and visual quality.

evaluate how faithfully the generated images reflect the semantics of the input prompt. For image quality assessment, we use CLIP-IQA[58], implemented via TorchMetrics[17], which compares the CLIP features of a pair of semantically opposite descriptors y1, y2 against those of the generated image, with the final score indicating how well y1 over y2 describes the image. Specifically, we evaluate overall image quality using high quality, low quality, blurriness using clear, blurry, and naturalness using natural, synthetic. Finally, to measure spatial faithfulness to the prescribed bounding boxes, we use YOLOv4[4] to predict bounding boxes over the generated images and compare them against the ground truth, reporting average precision at IOU = 0.5.

In our implementation, the denoising scheduler is configured for 50 timesteps, with the noise schedule parameters set to $\beta_{start} = 0.00085$ and $\beta_{end} = 0.012$. The Mask Guidance (MG) module is applied exclusively during the first 20 timesteps, whereas the Latent Optimization pipeline is applied across all denoising timesteps. For the ablation studies, the loss weighting parameters λ_1 and λ_2 are adjusted to compensate for the absence of individual components, in order to maintain adequate guidance strength through the remaining loss terms. Specifically, when MG is omitted entirely, λ_1 and λ_2 are set to 7 and 5, respectively. When only L_{SAE} is applied in conjunction with MG, λ_1 is increased to 9, and when only L_{GAP} is applied with MG, λ_2 is increased to 7. These adjustments ensure that the optimization objective retains sufficient supervisory signal in the absence of the complementary loss term.

All experiments are conducted over five random seeds; 42, 160, 77, 123, and 10; selected without bias. For quantitative comparison against state-of-the-art methods, we adopt the results as reported by iLGD, ensuring a fair and consistent evaluation protocol.

Table 4.1: Quantitative comparison with state-of-the-art methods. ↑ denotes that higher is better.

Method	T2I-Sim (↑)	CLIP-IQA (↑)			AP@0.5 (↑)
		Quality	Natural	Clear	
SD	0.303	0.928	0.705	0.736	–
BoxDiff	0.305	0.922	0.613	0.6945	0.192
Layout-guidance	0.301	0.936	0.640	0.814	0.118
MultiDiffusion	0.295	0.920	0.547	0.792	**0.411**
iLGD	0.309	0.961	0.654	0.817	0.202
MAPLE (Ours)	**0.311**	**0.976**	**0.744**	**0.978**	0.267

4.2 Quantitative Evaluation

In Table 4.1, we evaluate methods across image quality (CLIP-IQA), layout precision (AP@IoU=0.5), and semantic alignment (T2I-Sim). MAPLE achieves the best performance on T2I-Sim and all three CLIP-IQA metrics - Quality, Natural, and Clear - by a notable margin over all competing methods. Although MultiDiffusion leads on AP@IoU=0.5, it records the lowest T2I-Sim score and weak CLIP-IQA metrics, revealing that it sacrifices an unacceptable degree of image quality and semantic fidelity in exchange for layout control. BoxDiff achieves a modest AP@IoU=0.5 of 0.192 but falls considerably behind MAPLE on all quality metrics. Layout-guidance shows competitive CLIP-IQA scores relative to other baselines, yet its AP@IoU=0.5 of 0.118 is the lowest among methods that report it, limiting its practical utility. iLGD presents a more balanced profile, outperforming SD, BoxDiff, Layout-guidance, and MultiDiffusion on most metrics, but is nonetheless surpassed by MAPLE across all dimensions except AP@IoU=0.5, where MAPLE still achieves a strong score. These results demonstrate that MAPLE strikes the most favorable *balance* between perceptual image quality, semantic alignment, and layout control, establishing it as a meaningful advancement over all considered baselines.

4.3 Qualitative Evaluation

To stress-test fine-grained spatial controllability, bounding boxes are deliberately configured at varying degrees of overlap and spatial complexity, with qualitative comparisons presented in Figure 4.1. Among the baselines, BoxDiff, Layout-Guidance, and iLGD exhibit a consistent failure mode: upon encountering subject dominance, these methods attempt to force the suppressed object into the residual spatial regions, inevitably compromising the structural integrity and visual quality of the generated image. This behavior highlights a fundamental limitation in their spatial optimization strategies, which lack the granularity needed to enforce strict, subject-specific layout constraints under complex spatial configurations. MultiDiffusion, while offering a more balanced distribution of subject occupancy, adopts an aggressive optimization strategy that over-regularizes the denoising process, resulting in generations that, although spatially balanced, appear unnatural and lack perceptual realism. In contrast, our method directly addresses subject dominance through the dynamic handling of subjects. This principled design enables our method to consistently generate subjects within their assigned layout regions while preserving visual fidelity and natural image appearance across all evaluated configurations.

Figure 4.2 provides a qualitative comparison for the core problem addressed in this work. The top row illustrates the output of iLGD, which exhibits the most severe form of subject dominance: the horse disproportionately occupies the image, marginalizing the banana to a negligible spatial presence. Furthermore, unrelated semantic attributes of the dominating subject are superimposed onto the dominated subject, resulting in semantic attribute leakage that fundamentally corrupts the visual identity of the co-generated object. The middle row illustrates the output of MultiDiffusion, which, while alleviating semantic leakage to a degree, still fails to enforce spatially faithful generation, such as the $horse's$ head excessively occupies its designated bounding box region, extending beyond its prescribed spatial boundary and encroaching on the compositional balance of the scene. These two

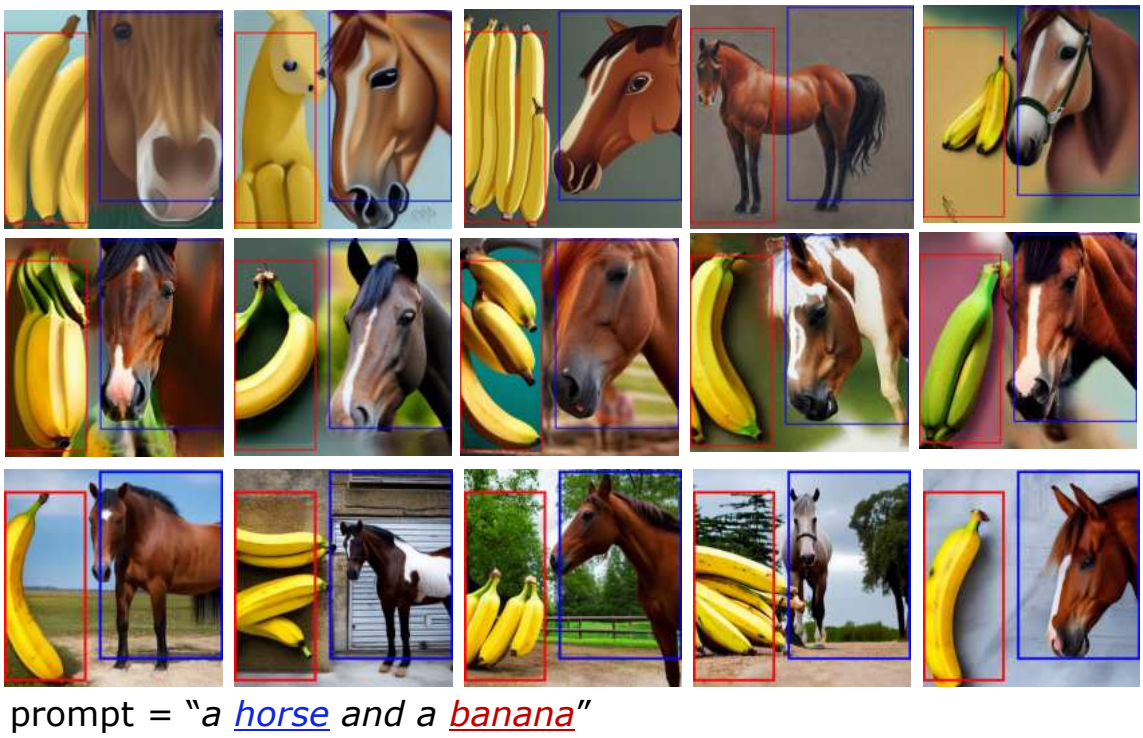


Figure 4.2: The columns represent the images created from the same seed. (Top and Middle) iLGD and MultiDiffusion fails to enforce region constraints, with "$horse$" dominating the image.(Bottom) MAPLE produces balanced spatial allocation between both objects.

failure modes collectively reveal the critical consideration in LIS: the pre-learned semantic prior of each subject must be carefully balanced during the generation process, as disproportionate focus on a dominant subject not only disrupts spatial layout adherence but also induces undesired semantic contamination across co-generated subjects. The bottom row demonstrates that our proposed method, *MAPLE*, successfully mitigates both failure modes, producing spatially balanced subject occupancy wherein each subject faithfully retains its distinct visual features and semantic attributes in the absence of cross-subject interference.

Figure 4.3: Visual Ablations. Per-component qualitative analysis of MAPLE, with layout instructions shown as solid bounding boxes and prompt as "a {} and a {}"

4.4 Ablation Studies

SD + MG. We take a closer look at the first column. Here, we notice that the model can generate both of the objects with quite good precision. However, there is unwanted occlusion, missing objects, or feature mixing, which has led to degradation in the quality around the overlap region, i.e., one of the objects has failed to generate fully. We also notice a bit of dominance of one subject with respect to the other.

SD + L_{MAPLE}. Here, we again observe that the subjects have generated almost equally well. However, the image looks unrealistic, specifically the image background, and in general, the image has contrast on the higher end. Also, one of the subjects seems to cover only a part of it's destined location while the other overflows into the unwanted layout region.

SD + MG + L_{SAE}. Although the generated images exhibit visual realism, subjects frequently extend beyond their designated layout regions, indicating that L_{SAE} alone lacks the cross-attention localization necessary to enforce strict spatial confinement. This is expected as we explicitly control them to remain confined even if it's an overlap region.

SD + MG + L_{GAP}. While spatial layout adherence is notably improved, cross-subject semantic leakage persists in the form of undesired mixing of fine-grained visual attributes, suggesting that l_{GAP} alone cannot fully resolve inter-subject feature contamination without self-attention regularization.

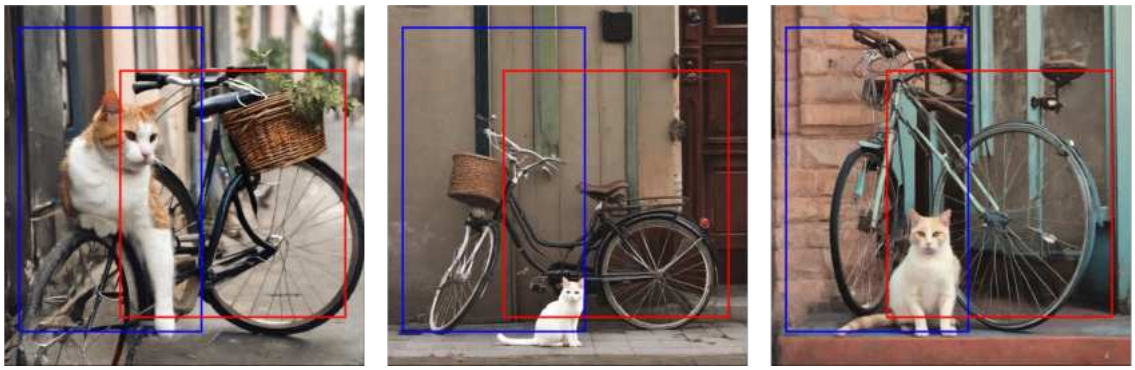
MAPLE effectively synthesizes the complementary strengths of each component into a unified framework. Mask Guidance steers subject-specific attention toward designated spatial regions. L_{SAE} builds upon this by enforcing self-attention regularity, ensuring balanced spatial occupancy, and preventing any single subject from disproportionately dominating the generated image. L_{GAP}, in turn, suppresses cross-region semantic leakage by constraining each token's cross-attention distribution to its assigned layout region, preserving the semantic integrity of each subject. The synergistic interaction of these three components directly addresses the individual shortcomings observed in each ablated variant, spatial misalignment, subject dominance, and semantic leakage.

4.5 Additional Visual Results

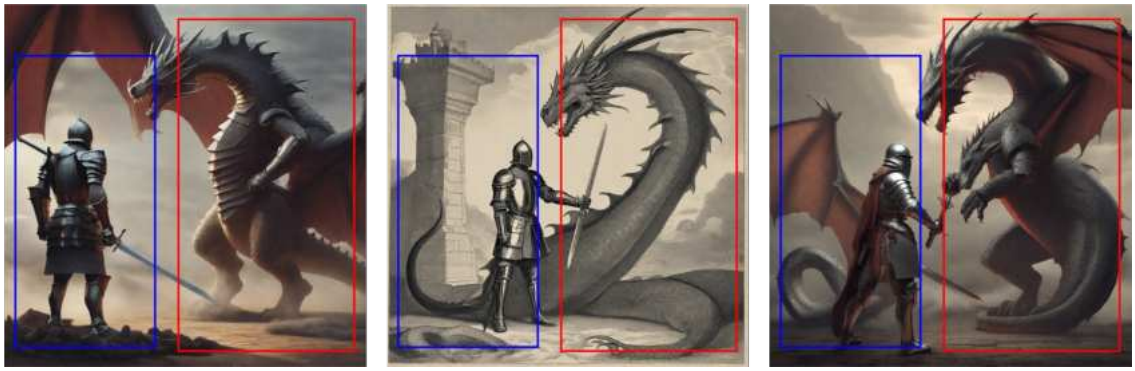
4.5.1 Plug-and-Play More Examples

Figure 4.4 presents additional qualitative results demonstrating the integration of MAPLE with Stable Diffusion XL 1.0. Despite the presence of subjects that inherently occupy larger spatial extents, our method successfully confines each subject within its prescribed bounding box region, thereby preserving adequate spatial allocation for co-occurring subjects that would otherwise be dominated.

Furthermore, MAPLE demonstrates robustness to compositional complexity, handling prompts involving two primary subjects while simultaneously accommodating fine-grained attribute specifications such as object orientation without compromising spatial fidelity or semantic coherence of the generated output.



prompt = "*a cat and a bicycle*"



prompt = "*a knight facing a dragon*"

prompt = "*a woman reading under a large tree*"

Figure 4.4: Stable-Diffusion XL was used to generate the images. The rows represent images generated from the same prompt but different seeds.

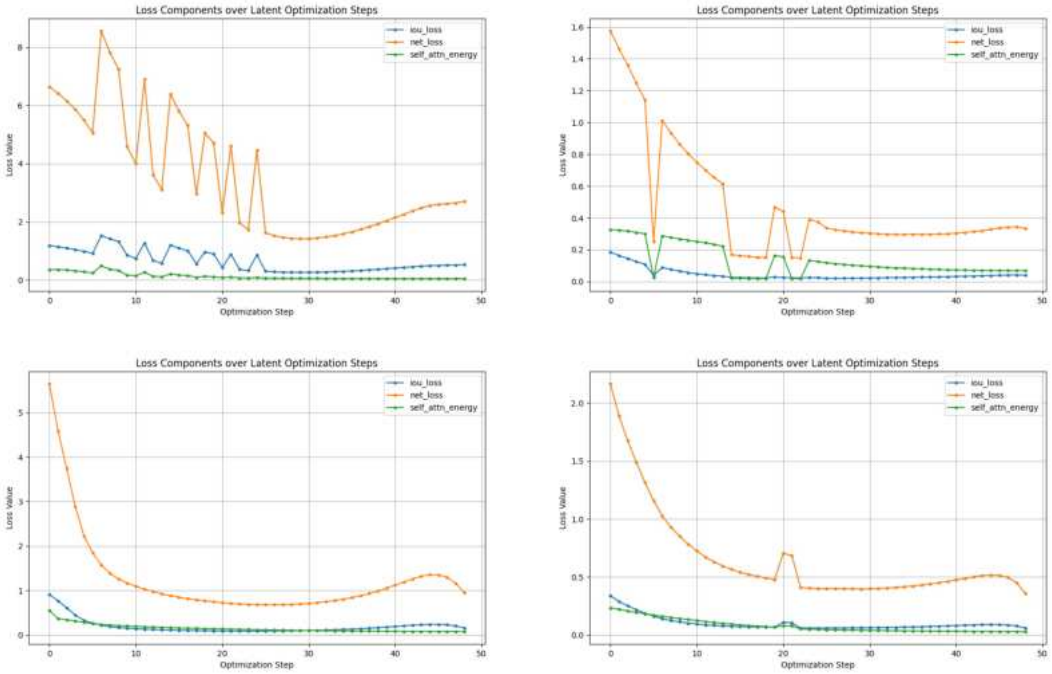


Figure 4.5: Optimization trajectories of the proposed MAPLE objective and its constituent loss terms across representative generation examples. The curves report L_{GAP}, denoted as `iou_loss`; L_{SAE}, denoted as `self_attn_energy`; and the overall objective L_{MAPLE}, denoted as `net_loss`. Peaks and troughs reflect adaptive shifts in guidance toward the currently dominant subject during denoising.

4.5.2 Optimization Objective Descent

Figure 4.5 illustrates the optimization trajectory of the proposed objective L_{MAPLE}, together with its constituent losses L_{SAE} and L_{GAP}. Across the four examples, the loss curves demonstrate how the proposed dynamic guidance mechanism adaptively regulates subject dominance during generation. Local peaks and troughs in the objective indicate transitions in the dominant subject receiving guidance. Specifically, an increase in the loss suggests that another subject has become relatively overemphasized and therefore requires stronger spatial or semantic regularization. Conversely, a decrease indicates that the currently guided subject has been sufficiently constrained, allowing the optimization process to shift attention toward the subject exhibiting greater dominance. When a single subject remains dominant throughout the denoising process, the losses follow a comparatively

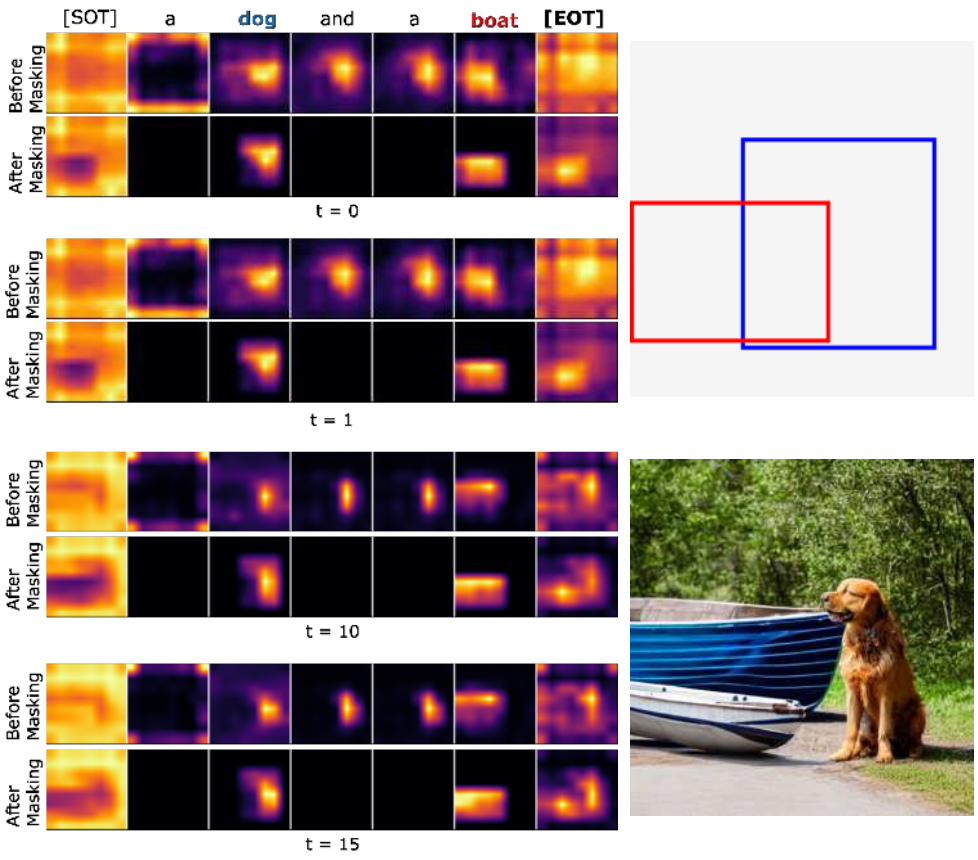


Figure 4.6: Comparative study of the cross-attention maps for the 13th layer of the UNet in Stable Diffusion at different timesteps.

smooth descent, reflecting stable and consistent guidance without frequent subject-level switching.

4.5.3 Cross-Attention Maps

Figure 4.6 illustrates the evolution of cross-attention maps across successive timesteps, up to the point at which our **MG** module is applied. At the initial timestep, the attention distribution is notably dispersed, exhibiting considerable overlap with spatially undesired regions beyond the prescribed bounding boxes. As the denoising process progresses, a discernible trend emerges wherein the attention progressively converges toward the regions delineated by the masking operation, with the spatial extent and morphology of each subject's attention map increasingly resembling the shape and scale of the corresponding ob-

ject.

Of particular note is the behavior of the [EOT] token throughout this process. At the initial timestep, the [EOT] token exhibits a strong semantic bias toward the boat, failing to adequately encode the presence of the co-occurring subject. However, as the timesteps advance, the [EOT] token progressively captures the aggregated semantics of both subjects, ultimately producing an attention response that correctly encompasses both the boat and the dog while appropriately suppressing background regions.

Chapter 5

Conclusion

This work introduces MAPLE, a training-free layout-to-image framework built on two core components: Mask Guidance, which steers subject-specific attention toward designated spatial regions, and two novel optimization constraints combined into L_{MAPLE}, which enforce balanced spatial occupancy and suppress semantic leakage. Together, these components directly tackle subject dominance, yielding $\sim \mathbf{9}\%$ average improvement in CLIP-IQA metrics along with other metrics in Layout to Image Synthesis over existing methods, while integrating seamlessly into existing diffusion models without additional training. As future work, we intend to apply this dynamic token relevance to other T2I tasks as it is a general method relevant to Diffusion Models.

Reference

[1] Yogesh Balaji et al. “ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers”. In: *arXiv preprint arXiv:2211.01324* (2022).

[2] Yuanhao Ban et al. “The crystal ball hypothesis in diffusion models: Anticipating object positions from initial noise”. In: *arXiv preprint arXiv:2406.01970* (2024).

[3] Omer Bar-Tal et al. “MultiDiffusion: Fusing Diffusion Paths for Controlled Image Generation”. In: *arXiv preprint arXiv:2302.08113* (2023).

[4] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. “Yolov4: Optimal speed and accuracy of object detection”. In: *arXiv preprint arXiv:2004.10934* (2020).

[5] Agneet Chatterjee et al. “Getting it right: Improving spatial consistency in text-to-image models”. In: *European Conference on Computer Vision*. Springer. 2024, pp. 204–222.

[6] Hila Chefer et al. “Attend-and-Excite: Attention-Based Semantic Guidance for Text-to-Image Diffusion Models”. In: *ACM Transactions on Graphics (TOG)* 42 (2023), pp. 1–10. URL: https://api.semanticscholar.org/CorpusID:256416326.

[7] Chen Chen, Daochang Liu, and Chang Xu. “Towards memorization-free diffusion models”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 8425–8434.

[8] Chen Chen et al. “Exploring local memorization in diffusion models via bright ending attention”. In: *arXiv preprint arXiv:2410.21665* (2024).

[9] Huancheng Chen et al. “Training-Free Layout-to-Image Generation with Marginal Attention Constraints”. In: *arXiv preprint arXiv:2411.10495* (2024).

[10] Jingye Chen et al. “Textdiffuser: Diffusion models as text painters”. In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 9353–9387.

[11] Junsong Chen et al. *PixArt-α: Fast Training of Diffusion Transformer for Photorealistic Text-to-Image Synthesis*. 2023. arXiv: 2310.00426 [cs.CV].

[12] Minghao Chen, Iro Laina, and Andrea Vedaldi. "Training-Free Layout Control with Cross-Attention Guidance". In: *arXiv preprint arXiv:2304.03373* (2023).

[13] Bo Cheng et al. "Hico: Hierarchical controllable diffusion model for layout-to-image generation". In: *Advances in neural information processing systems* 37 (2024), pp. 128886–128910.

[14] Jiaxin Cheng et al. "Layoutdiffuse: Adapting foundational diffusion models for layout-to-image generation". In: *arXiv preprint arXiv:2302.08908* (2023).

[15] Guillaume Couairon et al. "Zero-shot spatial layout conditioning for text-to-image diffusion models". In: *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2023, pp. 2174–2183. DOI: 10.1109/ICCV51070.2023.00207.

[16] Omer Dahary et al. "Be yourself: Bounded attention for multi-subject text-to-image generation". In: *European Conference on Computer Vision*. Springer. 2024, pp. 432–448.

[17] Nicki Skafte Detlefsen et al. "TorchMetrics - Measuring Reproducibility in PyTorch". In: *Journal of Open Source Software* 7.70 (2022), p. 4101. DOI: 10.21105/joss.04101. URL: https://doi.org/10.21105/joss.04101.

[18] Prafulla Dhariwal and Alexander Nichol. "Diffusion models beat gans on image synthesis". In: *Advances in neural information processing systems* 34 (2021), pp. 8780–8794.

[19] Raman Dutt et al. " MemControl: Mitigating Memorization in Diffusion Models via Automated Parameter Selection ". In: *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. Los Alamitos, CA, USA: IEEE Computer Society, Mar. 2025, pp. 4491–4501. DOI: 10.1109/WACV61041.2025.00441. URL: https://doi.ieeecomputersociety.org/10.1109/WACV61041.2025.00441.

[20] Oran Gafni et al. "Make-a-scene: Scene-based text-to-image generation with human priors". In: *European conference on computer vision*. Springer. 2022, pp. 89–106.

[21] Tejas Gokhale et al. "Benchmarking spatial relationships in text-to-image generation". In: *arXiv preprint arXiv:2212.10015* (2022).

[22] Biao Gong et al. "Check locate rectify: A training-free layout calibration system for text-to-image generation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 6624–6634.

[23] Xiefan Guo et al. “Initno: Boosting text-to-image diffusion models via initial noise optimization”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 9380–9389.

[24] Hyeonggeun Han et al. “Adjusting Initial Noise to Mitigate Memorization in Text-to-Image Diffusion Models”. In: *arXiv preprint arXiv:2510.08625* (2025).

[25] Amir Hertz et al. “Prompt-to-prompt image editing with cross attention control”. In: (2022).

[26] Dominik Hintersdorf et al. “Finding NeMo: Localizing Neurons Responsible For Memorization in Diffusion Models”. In: *Advances in Neural Information Processing Systems*. Ed. by A. Globerson et al. Vol. 37. Curran Associates, Inc., 2024, pp. 88236–88278. DOI: 10.52202/079017-2800.

[27] Jonathan Ho, Ajay Jain, and Pieter Abbeel. “Denoising diffusion probabilistic models”. In: *Advances in neural information processing systems* 33 (2020), pp. 6840–6851.

[28] Jonathan Ho and Tim Salimans. *Classifier-Free Diffusion Guidance*. 2022. arXiv: 2207.12598 [cs.LG]. URL: https://arxiv.org/abs/2207.12598.

[29] Xiwei Hu et al. “Ella: Equip diffusion models with llm for enhanced semantic alignment”. In: *arXiv preprint arXiv:2403.05135* (2024).

[30] Mengqi Huang et al. “Not all image regions matter: Masked vector quantization for autoregressive image generation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 2002–2011.

[31] Anubhav Jain et al. “Classifier-free guidance inside the attraction basin may cause memorization”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2025, pp. 12871–12879.

[32] Hayeon Jeong and Jong-Seok Lee. “Dominating vs. Dominated: Generative Collapse in Diffusion Models”. In: *arXiv preprint arXiv:2512.20666* (2025).

[33] Chengyou Jia et al. “SSMG: spatial-semantic map guided diffusion model for free-form layout-to-image generation”. In: *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*. AAAI’24/IAAI’24/EAAI’24. AAAI Press, 2024. ISBN: 978-1-57735-887-9. DOI: 10.1609/aaai.v38i3.28024. URL: https://doi.org/10.1609/aaai.v38i3.28024.

[34] Bonan Li et al. “FreLay: Frequency-aware Energy Function for Training-free Layout-to-Image Generation”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 40 (Mar. 2026), pp. 5992–6000. DOI: 10.1609/aaai.v40i8.37522.

[35] Yuheng Li et al. “GLIGEN: Open-Set Grounded Text-to-Image Generation”. In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 22511–22521. DOI: 10.1109/CVPR52729.2023.02156.

[36] Zongming Li et al. “Controlar: Controllable image generation with autoregressive models”. In: *arXiv preprint arXiv:2410.02705* (2024).

[37] Tsung-Yi Lin et al. “Microsoft coco: Common objects in context”. In: *European conference on computer vision*. Springer. 2014, pp. 740–755.

[38] Alex Nichol et al. “Glide: Towards photorealistic image generation and editing with text-guided diffusion models”. In: *arXiv preprint arXiv:2112.10741* (2021).

[39] Zakaria Patel and Kirill Serkh. “Enhancing image layout control with loss-guided diffusion models”. In: *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE. 2025, pp. 3916–3924.

[40] William S. Peebles and Saining Xie. “Scalable Diffusion Models with Transformers”. In: *2023 IEEE/CVF International Conference on Computer Vision (ICCV)* (2022), pp. 4172–4182. URL: https://api.semanticscholar.org/CorpusID:254854389.

[41] Bao Pham et al. “Memorization to generalization: Emergence of diffusion models from associative memory”. In: *arXiv preprint arXiv:2505.21777* (2025).

[42] Quynh Phung, Songwei Ge, and Jia-Bin Huang. “Grounded Text-to-Image Synthesis with Attention Refocusing”. In: *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024, pp. 7932–7942. DOI: 10.1109/CVPR52733.2024.00758.

[43] Dustin Podell et al. “SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis”. In: *ArXiv* abs/2307.01952 (2023).

[44] Alec Radford et al. *Learning Transferable Visual Models From Natural Language Supervision*. 2021. arXiv: 2103.00020 [cs.CV]. URL: https://arxiv.org/abs/2103.00020.

[45] Aditya Ramesh et al. “Zero-shot text-to-image generation”. In: *International conference on machine learning*. Pmlr. 2021, pp. 8821–8831.

[46] Jie Ren et al. “Unveiling and Mitigating Memorization in Text-to-Image Diffusion Models Through Cross Attention”. In: *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LXXVII*. Milan, Italy: Springer-Verlag, 2024, pp. 340–356. ISBN: 978-3-031-72979-9. DOI: 10.1007/978-3-031-72980-5_20. URL: https://doi.org/10.1007/978-3-031-72980-5_20.

[47] Robin Rombach et al. “High-Resolution Image Synthesis with Latent Diffusion Models”. In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), pp. 10674–10685.

[48] Nataniel Ruiz et al. “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*. 2023, pp. 22500–22510.

[49] Chitwan Saharia et al. “Photorealistic text-to-image diffusion models with deep language understanding”. In: *Advances in neural information processing systems* 35 (2022), pp. 36479–36494.

[50] Huajie Shao et al. “Controlvae: Controllable variational autoencoder”. In: *International conference on machine learning*. PMLR. 2020, pp. 8655–8664.

[51] Yang Song et al. “Score-Based Generative Modeling through Stochastic Differential Equations”. In: *International Conference on Learning Representations*. 2021. URL: https://openreview.net/forum?id=PxTIG12RRHS.

[52] Wei Sun and Tianfu Wu. “Learning layout and style reconfigurable gans for controllable image synthesis”. In: *IEEE transactions on pattern analysis and machine intelligence* 44.9 (2021), pp. 5070–5087.

[53] Wenqiang Sun et al. “Spatial-aware latent initialization for controllable image generation”. In: *arXiv preprint arXiv:2401.16157* (2024).

[54] Yanan Sun et al. *AnyControl: Create your artwork with versatile control on text-to-image generation*. 2024.

[55] Ashkan Taghipour et al. “Box It to Bind It: Unified Layout Control and Attribute Binding in Text-to-Image Diffusion Models”. In: *IEEE Transactions on Multimedia* 27 (2025), pp. 8393–8407. DOI: 10.1109/TMM.2025.3607759.

[56] Arash Vahdat and Jan Kautz. “NVAE: A deep hierarchical variational autoencoder”. In: *Advances in neural information processing systems* 33 (2020), pp. 19667–19679.

[57] Bo Wang et al. “Interactive Image Synthesis With Panoptic Layout Generation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2022, pp. 7783–7792.

[58] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. “Exploring clip for assessing the look and feel of images”. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 37. 2. 2023, pp. 2555–2563.

[59] Xudong Wang et al. *InstanceDiffusion: Instance-level Control for Image Generation*. 2024. arXiv: 2402.03290 [cs.CV].

[60] Qiucheng Wu et al. “Uncovering the disentanglement capability in text-to-image diffusion models”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2023, pp. 1900–1910.

[61] Qiang Xiang et al. “InstanceAssemble: Layout-Aware Image Generation via Instance Assembling Attention”. In: *arXiv preprint arXiv:2509.16691* (2025).

[62] Jinheng Xie et al. “BoxDiff: Text-to-Image Synthesis with Training-Free Box-Constrained Diffusion”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2023, pp. 7452–7461.

[63] Jing Xiong et al. “Autoregressive models in vision: A survey”. In: *arXiv preprint arXiv:2411.05902* (2024).

[64] Jiahui Yu et al. “Scaling autoregressive models for content-rich text-to-image generation”. In: *arXiv preprint arXiv:2206.10789* 2.3 (2022), p. 5.

[65] Yu Zeng et al. “Jedi: Joint-image diffusion models for finetuning-free personalized text-to-image generation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 6786–6795.

[66] Yu Zeng et al. “SceneComposer: Any-Level Semantic Image Synthesis”. In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), pp. 22468–22478. URL: https://api.semanticscholar.org/CorpusID:253734941.

[67] Gong Zhang et al. “Forget-me-not: Learning to forget in text-to-image diffusion models”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2024, pp. 1755–1764.

[68] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. “Adding Conditional Control to Text-to-Image Diffusion Models”. In: *2023 IEEE/CVF International Conference on Computer Vision (ICCV)* (2023), pp. 3813–3824.

[69] Peiang Zhao et al. “LoCo: Training-Free Layout-to-Image Synthesis with Localized Constraints”. In: *Proceedings of the 33rd ACM International Conference on Multimedia*. MM ’25. Dublin, Ireland: Association for Computing Machinery, 2025, pp. 9481–9490. ISBN: 9798400720352. DOI: 10.1145/3746027.3754905. URL: https://doi.org/10.1145/3746027.3754905.

[70] Dewei Zhou et al. “Migc: Multi-instance generation controller for text-to-image synthesis”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 6818–6828.