Control and Realism: Best of Both Worlds in Layout-to-Image without Training

Bonan Li [*1,2] **Yinhan Hu** [*1] **Songhua Liu** [2] **Xinchao Wang** [2]

Abstract

Layout-to-Image generation aims to create complex scenes with precise control over the placement and arrangement of subjects. Existing works have demonstrated that pre-trained Text-to-Image diffusion models can achieve this goal without training on any specific data; however, they often face challenges with imprecise localization and unrealistic artifacts. Focusing on these drawbacks, we propose a novel training-free method, ***WinWinLay***. At its core, WinWinLay presents two key strategies—Non-local Attention Energy Function and Adaptive Update—that collaboratively enhance control precision and realism. On one hand, we theoretically demonstrate that the commonly used attention energy function introduces inherent spatial distribution biases, hindering objects from being uniformly aligned with layout instructions. To overcome this issue, non-local attention prior is explored to redistribute attention scores, facilitating objects to better conform to the specified spatial conditions. On the other hand, we identify that the vanilla backpropagation update rule can cause deviations from the pre-trained domain, leading to out-of-distribution artifacts. We accordingly introduce a Langevin dynamics-based adaptive update scheme as a remedy that promotes in-domain updating while respecting layout constraints. Extensive experiments demonstrate that WinWinLay excels in controlling element placement and achieving photorealistic visual fidelity, outperforming the current state-of-the-art methods.

1. Introduction

Recent advances in Text-to-Image (T2I) generation (Rombach et al., 2022; Podell et al., 2023) have profoundly revolutionized the vision landscape, facilitating the synthesis of highly authentic assets from textual prompts, *e.g.*, text-driven Image-to-Image translation (Tumanyan et al., 2023; Parmar et al., 2023; Ruiz et al., 2023; Shi et al., 2024; Tosi et al., 2025) and video generation (Wu et al., 2023; Zhang et al., 2023; Jiang et al., 2024; Qing et al., 2024; Kwon et al., 2025). Nevertheless, designing comprehensive prompts to meticulously control every aspect of an image can be both labor-intensive and time-consuming, posing challenges for efficient generation workflows. As a remedy, Layout-to-Image (L2I) models (Xue et al., 2023; Zheng et al., 2023; Chen et al., 2024b; Liu et al., 2024a) have been developed, guiding the process in a desired direction by incorporating user-provided inputs, such as bounding boxes.

To acquire L2I models, an intuitive framework is to fine-tune powerful T2I models with with spatial conditioning. However, these approaches (Li et al., 2023; Wu et al., 2024) incurs expensive training cost and faces challenges when collecting resource-intensive paired data. To overcome the aforementioned limitations, existing methods (Couairon et al., 2023; Xie et al., 2023) have explored the incorporation of layout guidance during the sampling phase, establishing the training-free paradigm for L2I. Among them, backward guidance (Chen et al., 2024d), which combines attention redistribution and backpropagation update rules, has been demonstrated as a promising scheme (Liu et al., 2024a; Chen et al., 2024b). Specifically, attention redistribution encourages cross-attention activation at designated positions via the energy function, while the backpropagation rule directly updates feature maps using corresponding gradients to match the specified layout. However, while these works offer higher efficiency, they fall short of training-based approaches in terms of spatial fidelity and image quality.

To bridge this gap, we initiate our exploration with an in-depth analysis of existing mechanisms and demonstrate that, although effective, they are still prone to suboptimal results in terms of control capabilities and visual realism. Firstly, we theoretically verify that the ***attention energy function tends to favor patches with higher initial attention*** within the bounding box during the optimization process, lead-

[*]Equal contribution [1]University of Chinese Academy of Sciences, Beijing, China [2]National University of Singapore, Singapore. Correspondence to: Xinchao Wang <xinchao@nus.edu.sg>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

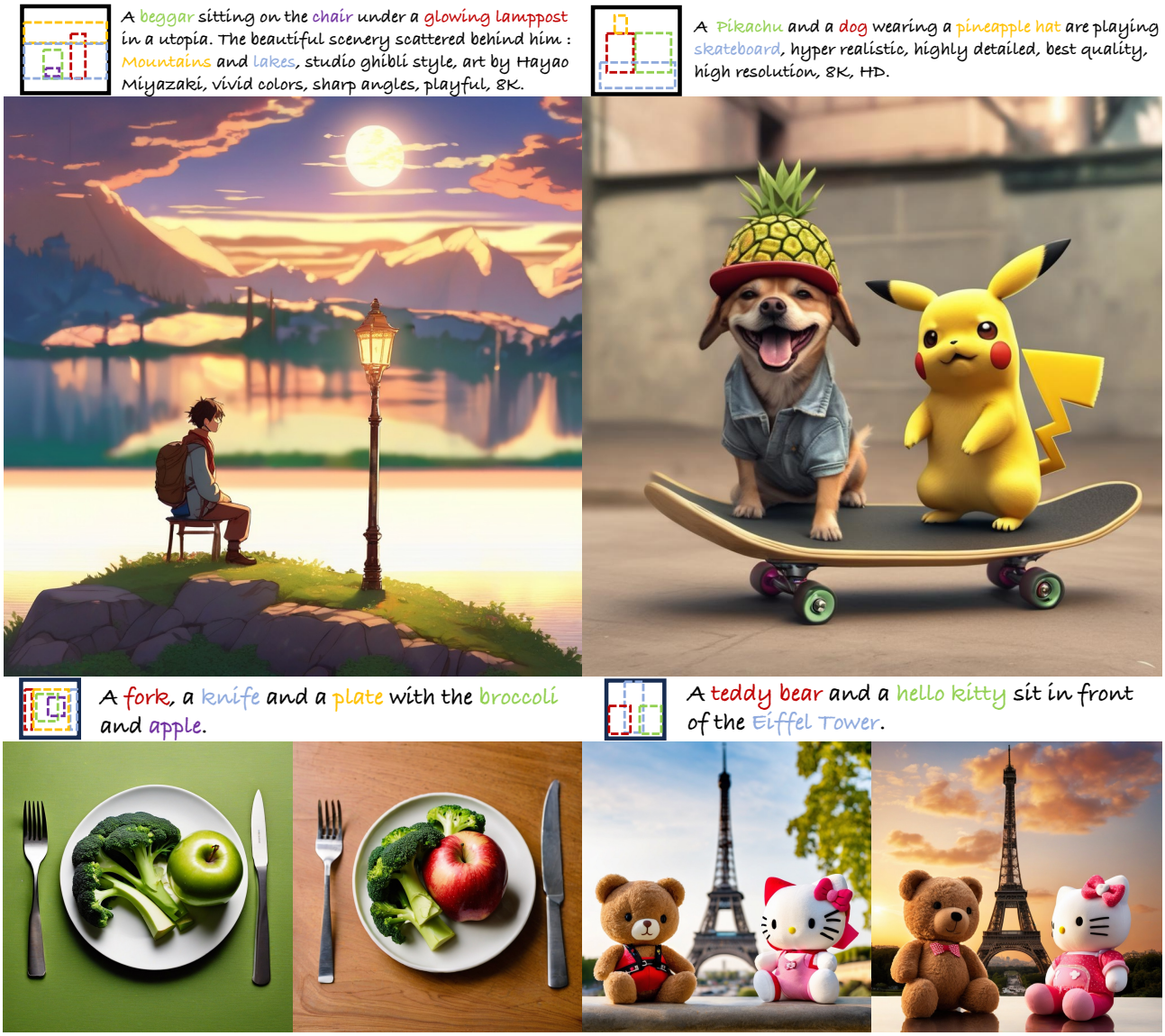

Figure 1: Given user-provided bounding boxes and prompts of subjects, our ***WinWinLay*** generates controllable and realistic images with pre-trained diffusion model, such as SDXL (Podell et al., 2023), without any finetuning with paired data.

ing to the final generated objects being confined to smaller regions rather than evenly distributed across the specified box. Secondly, the ***backpropagation update rule fails to account for the pre-trained distribution when biasing features***, resulting in a trade-off between control and image quality, where stronger control often comes at the cost of poorer visual appearance. To the best of our knowledge, this is the first study that theoretically analyze these two core components of backward guidance in Layout-to-Image task.

In this paper, we propose a novel model named WinWinLay, which generates high-quality training-free L2I results by explicitly accounting for above challenges. On one hand, to better align the given spatial guidance, we augment the attention energy function with the non-local attention prior to encourage attention to uniformly cover the specified region. Additionally, to avoid constraining irregularly shaped objects into a rigid box-like form (*e.g.*, coconut trees usually have broad leaves and slender trunks), we introduce a decaying schedule that gradually decreases the strength of prior along the denoising step that facilitates natural structure. On the other hand, focusing on the trade-off between controllability and realism, we design the update rule based on Langevin dynamics, which brings the best of two worlds by simultaneously incorporating the updating directions given by both layout controls and the original T2I model. Specifically, we introduce an adaptive weighting strategy to balance the two directions across different sampling steps, eliminating the need for cumbersome hyperparameter search.. Experiments demonstrate that the proposed approach mitigates

the above issues successfully and generates satisfactory L2I results (see Figure 1) in a training-free manner.

In summary, our contributions can be summarized as: (i) We provide the first theoretical analysis of previous backward guidance methods to the best of our knowledge. Inspired by the theoretical insights, we present an advanced approach, WinWinLay, for Layout-to-Image generation which exhibits significantly control and realistic quality. (ii) We propose a novel Non-local Attention Energy Function that guides the model to better adhere to spatial constraints while preserving the natural structure of objects. (iii) We explore a Langevin dynamics-based Adaptive Update scheme to eliminate the trade-off between layout instruction and realistic appearance while maintaining efficiency. (iv) We conduct extensive experiments to highlight the effectiveness of WinWinLay for both controllability and quality, thereby advancing the practical application of L2I generation.

2. Related Work

2.1. Text-to-Image Generation

Diffusion models (Ho et al., 2020) have recently disrupted the longstanding dominance of generative adversarial networks (GANs) (Goodfellow et al., 2014) in image synthesis (Dhariwal & Nichol, 2021; Song et al., 2021a; Ho et al., 2022), further accelerating advancements in Text-to-Image (T2I) generation (Saharia et al., 2022; Rombach et al., 2022; Podell et al., 2023; Peebles & Xie, 2023; Chen et al., 2024c). Benefitting from training on large-scale image-text datasets (Schuhmann et al., 2022b), they exhibited remarkable ability to generate diverse, creative images controlled by text prompts. Moreover, recent developments also unlock the potential of T2I to tackle challenging vision tasks, such as image editing (Brooks et al., 2023; Kawar et al., 2023; Ruiz et al., 2023; Mokady et al., 2023; Xu et al., 2024), style transfer (Sohn et al., 2024; Chen et al., 2024a; Ahn et al., 2024) and 3D generation (Chen et al., 2024e; Li et al., 2024; Wang et al., 2024b). Despite substantial progress, existing works still struggle to precisely control image details, such as layout, which significantly hampers their applicability in real-world scenarios. In this paper, we focus on controlling subject synthesis within the complex scene through user-specified layout condition in a training-free manner.

2.2. Layout-to-Image Generation

Layout-to-Image (L2I) (Xue et al., 2023; Zheng et al., 2023; Xie et al., 2023; Jia et al., 2024; Liu et al., 2024b;a) focus on generating images that simultaneously adhere to the textual prompt and corresponding layout instructions, *e.g.*, bounding boxes and scribble. To achieve this, existing works (Li et al., 2023; Yang et al., 2023; Wang et al., 2024a; Zhou et al., 2024) propose to finetune the powerful Text-to-Image models with paired data. However, collecting such extensive labeled images is not non-trivial and high-cost overheads also limit the usage of L2I in practice. To overcome the aforementioned challenges, recent studies (Bar-Tal et al., 2023; Kim et al., 2023; Singh et al., 2023; Couairon et al., 2023; Chen et al., 2024d;b; Liu et al., 2024a) explore training-free approaches that utilize forward or backward guidance mechanisms within the denoising process. Despite this success, the generated results often deviate from the predefined positions and exhibit severe unrealistic artifacts. Here, we theoretically analyze the backward guidance and propose improved strategies to alleviate these problems.

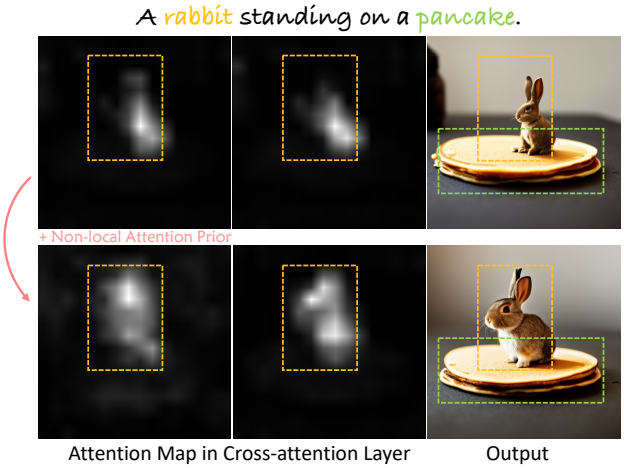

Figure 2: Visualization of cross-attention between text token "rabbit" and intermediate features of the denoiser. It can be observed that the attention energy function leads to attention focusing on a local region. Conversely, after incorporating non-local attention prior, the attention attempts to encompass the entire bounding box. To demonstrate the robustness of this prior, "pancake" is consistently equipped with non-local attention prior in this context.

3. Preliminaries

3.1. Latent Diffusion Model

Our work recruits the Latent Diffusion Model (LDM) (Rombach et al., 2022) as prior, which defines a generative process that gradually transforms a noise latent ϵ and a text prompt p to an image x. Specifically, LDM first encodes the image x_0 into in a low-dimensional latent space using a pre-trained encoder E, *i.e.*, $z_0 = E(x_0)$ and operates the diffusion process. Then, the representation z_0 are decoded back into the image x_0 using a pre-trained decoder D. Specifically, given a noised latent z_t at step t and text tokens $y = \phi(p)$ where ϕ is a frozen text encoder (Radford et al., 2021), the denoiser ϵ_θ is optimized to remove the noise ϵ added to the latent code z_0:

$$\min_{\theta} \mathbb{E}_{z_0,\epsilon,t}||\epsilon - \epsilon_\theta(z_t, t, y)||_2^2. \quad (1)$$

Here, learnable parameters θ is typically integrated into a U-Net (Ronneberger et al., 2015) architecture, which comprises convolutional layers, self-attention and cross-attention mechanisms.

3.2. Backward Guidance for Layout-to-Image

To precisely position the subject at the specified location, backward guidance (Chen et al., 2024d) aims to sample desired images from the distribution $p(z|y, b, i)$, where the bounding box b_i corresponding to the text token y_i of target subject. Specifically, backward guidance begins by designing an optimizable object, *e.g.*, attention energy function $\mathcal{E}_{aef}$, to redistribute attention map a in feature z, encouraging the attention values of the i^{th} token to focus within region b_i:

$$\mathcal{E}_{aef}(a^{(\gamma)}, m, i) = \left(1 - \frac{\sum_u m_{ui} \cdot a_{ui}^{(\gamma)}}{\sum_u a_{ui}^{(\gamma)}}\right)^2, \quad (2)$$

where $a_{ui}^{(\gamma)}$ quantifies the strength of the association between each location u in cross-attention layer γ and token y_i, m_i denotes the binary mask which is transformed from b_i with pixels inside the box region marked as 1 and those outside as 0. To bias attention maps, the gradient of the Equation (2) is computed via backpropagation to update the initial latent feature $\bar{z} = z$:

$$z \leftarrow \bar{z} - \eta \nabla_z \sum_\gamma \sum_i \mathcal{E}_{aef}(a^{(\gamma)}, m, i), \quad (3)$$

where η is a hyperparameter used to control the strength.

4. Method

In this section, we introduce WinWinLay, a training-free Layout-to-Image generation framework. First, we provide a detailed introduction to Non-local Attention Energy Function (Section 4.1), which is used to enhance layout constrain. Subsequently, we shift focus to explore Adaptive Update (Section 4.2) to eliminate the trade-off between control and quality.

4.1. Non-local Attention Energy Function

Attention energy function (Chen et al., 2024d) is a widely used loss term for guiding attention redistribution, but it often leads to objects occupying local region of the bounding box, hindering precise control. To address this, non-local attention prior is introduced to encourage attention to be smoothly distributed across the specified location.

Revisiting of Attention Energy Function. Based on the overview of attention energy function (Section 3.2), we can straightforwardly reformulate Equation (2) as following for intuitive analysis:

$$\mathcal{E}_{aef}(a, m) = (1 - \sum_u \tilde{a}_u \cdot m_u)^2, \quad (4)$$

where $\tilde{a}_u = a_u / \sum_u a_u$ denotes the normalization of a_u which denotes attention value of u^{th} patch in attention map a. Notably, for simplicity and clarity of presentation, the notation for subjects and cross-attention layer is omitted. Given $\max m_u = 1$, so we have $\max_{\tilde{a}} \sum_u \tilde{a}_u \cdot m_u = 1$ and then obtain following equation:

$$\mathcal{E}_{aef}(a, m) = (\max_{\tilde{a}} \sum_u \tilde{a}_u \cdot m_u - \sum_u \tilde{a}_u \cdot m_u)^2. \quad (5)$$

Here, it is evident that minimizing $\mathcal{E}_{aef}$ is equivalent to maximizing $\tilde{a} \cdot m$. Specifically, when $\tilde{a} \cdot m$ attains maximum value, the support set of $\tilde{a}$ is guaranteed to be contained within support set of m. Building on this insight, we first observe that the optimal solution to the attention energy function is not unique. Actually, when the support set of $\tilde{a}$ is entirely contained within m , $\tilde{a} \cdot m$ can be maximized. However, this non-uniqueness may lead to the support of $\tilde{a}$ concentrating in localized regions of m, thereby compromising effective control over the spatial layout. Meanwhile, we notice that $\nabla_{\tilde{a}} \mathcal{E}_{aef}(a, m) \parallel m$, causing all patches within the masked region to receive identical gradient magnitudes. This gives patches with larger initial values a significant advantage during the optimization process, thereby amplifying localized effects (see Figure 2). To support this perspective, we start by considering a simple yet universal optimization objective as follows:

$$\max_v f(v) = m \cdot \mathrm{softmax}(v), where\, v \in \mathbb{R}^n,\ m \in \{0, 1\}^n, \quad (6)$$

and we denote $q = \mathrm{softmax}(v)$ for simplicity.

Theorem 4.1. *Assume that during the optimization process at a certain step, there exist $m_j = m_k = 1$ and $q_j > q_k$. After a single gradient update with step size $\beta > 0$, the updated values q'_j, q'_k satisfy $q'_j > q'_k$ and $\frac{q'_j}{q'_k} > \frac{q_j}{q_k}$.*

Proof. First, Jacobian matrix of q with respect to v can be computed as follows:

$$J = diag(q) - qq^T. \quad (7)$$

According to the chain rule, we have the gradient of v by:

$$\nabla_v f(v) = J^T m = \begin{pmatrix} q_1(m_1 - q^T m) \\ \vdots \\ q_n(m_n - q^T m) \end{pmatrix}. \quad (8)$$

Then, v^j and v^k are updated as:

$$v'_j = v_j + \beta(m_j - q^T m)q_j = v_j + \beta' q_j, \quad (9)$$

$$v'_k = v_k + \beta(m_k - q^T m)q_k = v_k + \beta' q_k, \quad (10)$$

Naturally, combining the above equations gives:

$$\begin{aligned}\frac{q'_j}{q'_k} &= \exp(v'_j - v'_k) = \exp(v_j - v_k + \beta'(q_j - q_k)) \\ &= \exp(v_j - v_k) \cdot \exp(\beta'(q_j - q_k)) \\ &= \frac{q_j}{q_k} \cdot \exp(\beta'(q_j - q_k)) > \frac{q_j}{q_k}. \end{aligned} \quad (11)$$

□

Through the analysis of the above problem, it can be concluded that patches within the masked region with larger initial values tend to amplify their relative prominence during the optimization process, thereby suppressing the growth of other regions. This implies that the attention map redistributed by energy function exhibits an implicit bias, favoring regions with larger initial values. Consequently, it becomes challenging to evenly cover the entire box.

Non-local Attention Prior. In this regard, a simple and effective non-local attention prior is introduced to facilitate global attention responses. Different from intuitively conceivable uniform constrain, this prior promotes the placement of objects near the center of the bounding box while encouraging maximal coverage of the entire region. Concretely, given the bounding box b (width as W, height as H) and its center point $c = (c_x, c_y)$, the normalized distance from point $u = (u_x, u_y)$ within the masked region S to the center can be calculated as $d_u = \frac{(u_x - c_x)^2}{W} + \frac{(u_y - c_y)^2}{H}$. Accordingly, we construct the prior distribution $\tau_u \propto exp(-\lambda d_u)$, where $\lambda \geq 0$ is used to control the variance of the distribution. This design ensures that points farther from the center are assigned smaller probability values. Subsequently, local bias is alleviated by maximizing the KL divergence between the distribution of the attention within S and the prior τ:

$$\mathcal{R}_{nap} = \sum_{u \in S} \hat{a}_u \log \frac{\hat{a}_u}{\tau_u}, \quad (12)$$

where $\hat{a}_u = a_u / \sum_{u \in S} a_u$ and a denotes attention value.

Total Loss. Here, non-local attention energy function is formulated as the summation of $\mathcal{E}_{aef}$ and $\mathcal{R}_{nap}$ across all subject i and layer γ:

$$\begin{aligned}\mathcal{E}_{naef} = \sum_{\gamma} \sum_{i} [(1 - \frac{\sum_u m_{ui} \cdot a_{ui}^{(\gamma)}}{\sum_u a_{ui}^{(\gamma)}})^2 \\ + \rho \sum_{ui \in S_i} \hat{a}_{ui}^{(\gamma)} \log \frac{\hat{a}_{ui}^{(\gamma)}}{\tau_{ui}}]. \end{aligned} \quad (13)$$

To account for the irregular shape of objects in real-world scenarios, we introduce a hyperparameter ρ that decreases linearly with the denoising timesteps, enabling objects to adapt to naturally structure. Similar to existing work (Chen et al., 2024d), we only reallocate cross-attentions with corresponding tokens in the middle and first up layers.

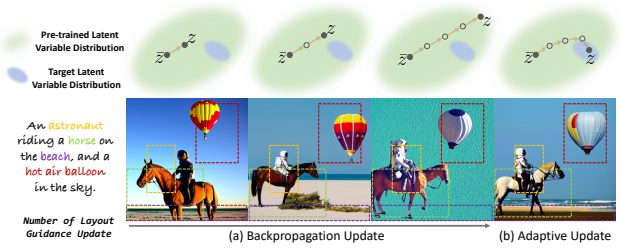

Figure 3: Given the initial feature $\bar{z}$ in a pre-trained distribution, our goal is to iteratively update it within layout constraints to achieve a target latent vector z. Here, (a) backpropagation update struggles to effectively maintain spatial control with a limited number of updates, while an increased number of updates often leads to deviations from the pre-trained distribution, leading to a trade-off between control and generation quality. In contrast, (b) adaptive update strategies simultaneously account for the pre-trained distribution and layout constraints, thereby consistently improving overall performance. Zoom in for more details.

4.2. Adaptive Update

Although backpropagation update is simple, it fails to balance layout constraints and image quality. Therefore, we propose Adaptive Update to consistently enhance the appearance of outputs based on Langevin dynamics and adaptive distribution construction.

Revisiting of Backpropagation Update. Given latent feature z_{t+1} as timestep $t+1$, conditional probability $p(z_t|z_{t+1})$ of pretrained diffusion model to obtain an initial estimation $\bar{z}_t$. Subsequently, the gradient update step refines $\bar{z}_t$ to minimize the non-local attention energy function $\mathcal{E}_{naef}$ (replace $\mathcal{E}_{aef}$ for consistent description) to obtained optimized latent variable $z_t = \arg\min \mathcal{E}_{naef}(\bar{z}_t, m)$ via Equation (3) where $a = attention\_map(\bar{z})$. However, update in this manner fails to adequately account for the constraints imposed by the latent variable distribution $p(z_t)$, which may lead to a trade-off between generation control and output quality. Specifically, when the gradient updates are insufficient, the optimization of $\mathcal{E}_{naef}$ remains suboptimal, resulting in poor layout control. Conversely, when the gradient updates are overly exhaustive, the resulting z_t may deviate significantly from the initial estimate $\bar{z}_t$, leading to a reduced likelihood $p(z_t)$. This misalignment adversely affects subsequent denoising steps, ultimately degrading the quality of the generated images. Moreover, the visualization results in Figure 3 under varying levels of control intensity further experimental substantiate this conclusion.

Langevin Dynamics Updating. To eliminate the stubborn trade-off, we propose that both the attention redistribution and $p(z_t)$ should be concurrently considered during the update process. For attention redistribution function, the corresponding Gibbs distribution can be constructed as

$p(m|z_t) \propto exp(-\nu\mathcal{E}_{naef}(z_t, m))$, where ν is a hyperparameter that controls the shape of the distribution. Given timestep t, our ultimate goal is to sample z_t from $p(z_t|m)$. Based on Bayes' Theorem, we have:

$$p(z_t|m) \propto p(z_t)p(m|z_t), \tag{14}$$

then the score function of $p(z_t|m)$ can be derived as:

$$\begin{aligned}\nabla_{z_t}\log(p(z_t|m)) &= \nabla_{z_t}\log(p(z_t)) + \nabla_{z_t}\log(p(m|z_t)) \\ &= \nabla_{z_t}\log(p(z_t)) - \nu\nabla_{z_t}\mathcal{E}_{naef}(z_t, m).\end{aligned} \tag{15}$$

Here, $\nabla_{z_t}\log(p(z_t))$ represents the unconditional score function which is approximated by pre-trained diffusion model. According to (Song & Ermon, 2019), we can use Langevin dynamics to sample from any distribution with a known score function. Specifically, given a step size $\xi > 0$ and an initial value $\bar{z}_t^{(0)}$, Langevin dynamics iteratively updates as follows:

$$\bar{z}_t^{(k+1)} = \bar{z}_t^{(k)} + \xi\nabla_{\bar{z}_t^{(k)}}\log(p(\bar{z}_t^{(k)}|m)) + \sqrt{2\xi}\epsilon_k, \tag{16}$$

where $\epsilon_k \sim \mathcal{N}(0, I)$ and $0 \leq k \leq O$. As $\xi \to 0$ and $O \to \infty$, the distribution of $\bar{z}_t^O$ will converge to $p(z_t|m)$. Note that for step size $\xi > 0$ and finite O, the sampling process can be corrected using the Metropolis-Hastings method to convert it into a strict MCMC sampling procedure. However, this correction step is typically omitted for convenience in practice. Here, similar to (Song et al., 2021b), we determine step size $\xi = 2(r||\epsilon_k||_2/||\nabla_{\bar{z}_t}\log(p(\bar{z}_t|m))||_2)^2$ where r is the signal-to-noise ratio.

Adaptive Distribution Construction. Although Langevin dynamics effectively mitigates the trade-off, the introduction of the additional hyperparameter $\nu > 0$ of distribution consequently reduces generation efficiency. From Equation (15), it follows that ν adjusts the weight of $\nabla_{z_t}\mathcal{E}_{naef}(z_t, m)$ in score function $\nabla_{z_t}\log(p(z_t|m))$. Intuitively, a larger ν results in a steeper distribution, where the optimization process focuses more on minimizing the non-local attention energy function, leading to faster convergence (smaller O) but requiring larger step sizes ξ to accelerate the optimization process of $\log(p(z_t))$, which can increase the error of Langevin dynamics and then degrade image quality. Conversely, a smaller ν produces a flatter distribution, prioritizing image quality preservation, which slows the optimization process (larger O) and requires smaller ξ but leads to more iterations, reducing sampling efficiency. Therefore, selecting the appropriate ν is critical to balance image quality and generation efficiency. Here, we propose treating Equation (16) as a multi-task optimization problem to explore the optimal ν, where one task minimizes the attention energy and the other maximizes the distribution probability. Inspired by Nash-MTL (Navon et al., 2022), we model the gradient combination of these two tasks as a bargaining game, solving for the Nash Bargaining Solution. Let $\{g_j \in \mathbb{R}^d | 1 \leq j \leq K\}$ represent the gradients of K tasks, the optimal gradient combination coefficients $\alpha \in \mathbb{R}_+^K$ satisfy $G^TG\alpha = \frac{1}{\alpha}$, where G is the matrix whose columns are the gradients g_j. Nash-MTL uses optimization to approximate the solution for α, and we find that when $K = 2$, this equation has a simple analytical solution:

Corollary 4.2. *Given $G = (g_1, g_2) \in \mathbb{R}^{d\times 2}$ and $\alpha = (\alpha_1, \alpha_2)^T \in \mathbb{R}_+^2$, if $G^TG\alpha = \frac{1}{\alpha}$, then we have $\frac{\alpha_1}{\alpha_2} = \frac{||g_2||}{||g_1||}$.*

Proof. According to $G^TG\alpha = \frac{1}{\alpha}$, we have:

$$\alpha_1^2||g_1||^2 + \alpha_1\alpha_2 g_1^T g_2 = 1, \tag{17}$$

$$\alpha_1\alpha_2 g_2^T g_1 + \alpha_2^2||g_2||^2 = 1. \tag{18}$$

By subtracting Equation (18) from Equation (17):

$$\alpha_1^2||g_1||^2 = \alpha_2^2||g_2||^2, \tag{19}$$

we can derive the conclusion as $\frac{\alpha_1}{\alpha_2} = \frac{||g_2||}{||g_1||}$. □

Based on the above proof, we propose apative update rule by formalize ν as an adaptive parameter for each iteration:

$$\nu = \frac{||\nabla_{z_t}\log(p(z_t))||}{||\nabla_{z_t}\mathcal{E}_{naef}(z_t, m))||}. \tag{20}$$

This design enables us to effectively mitigate the trade-off at negligible cost, making it more suitable for practical applications.

5. Experiments

In this section, we first provide the experimental setup and then conduct both qualitative and quantitative experiments to compare our method with previous SOTA Layout-to-Image methods. Additionally, we perform an ablation study to demonstrate the effectiveness of the proposed approaches.

5.1. Experimental setup

Evaluation Benchmarks. Akin to prior work (Chen et al., 2024d), we quantitatively evaluate our WinWinLay on COCO2014 (Lin et al., 2014) and Flickr30K (Plummer et al., 2015). For performance evaluation, we leverage YOLOv7 (Wang et al., 2023) for object detection, employing metrics such as AP (Li et al., 2021) to assess the effectiveness of our method in accurately locating and generating objects. Furthermore, the CLIP-s (Radford et al., 2021) is utilized to quantitatively evaluate image-text compatibility, thereby measuring the semantic accuracy of the synthesized images. Additionally, we also use the advantage metric FID (Kynkäänniemi et al., 2023), PickScore (Kirstain et al., 2023) and ImageReward (Xu et al., 2023) to evaluate image quality. Here, we set text template as "A photo of [prompt]" to acquire more realistic results.

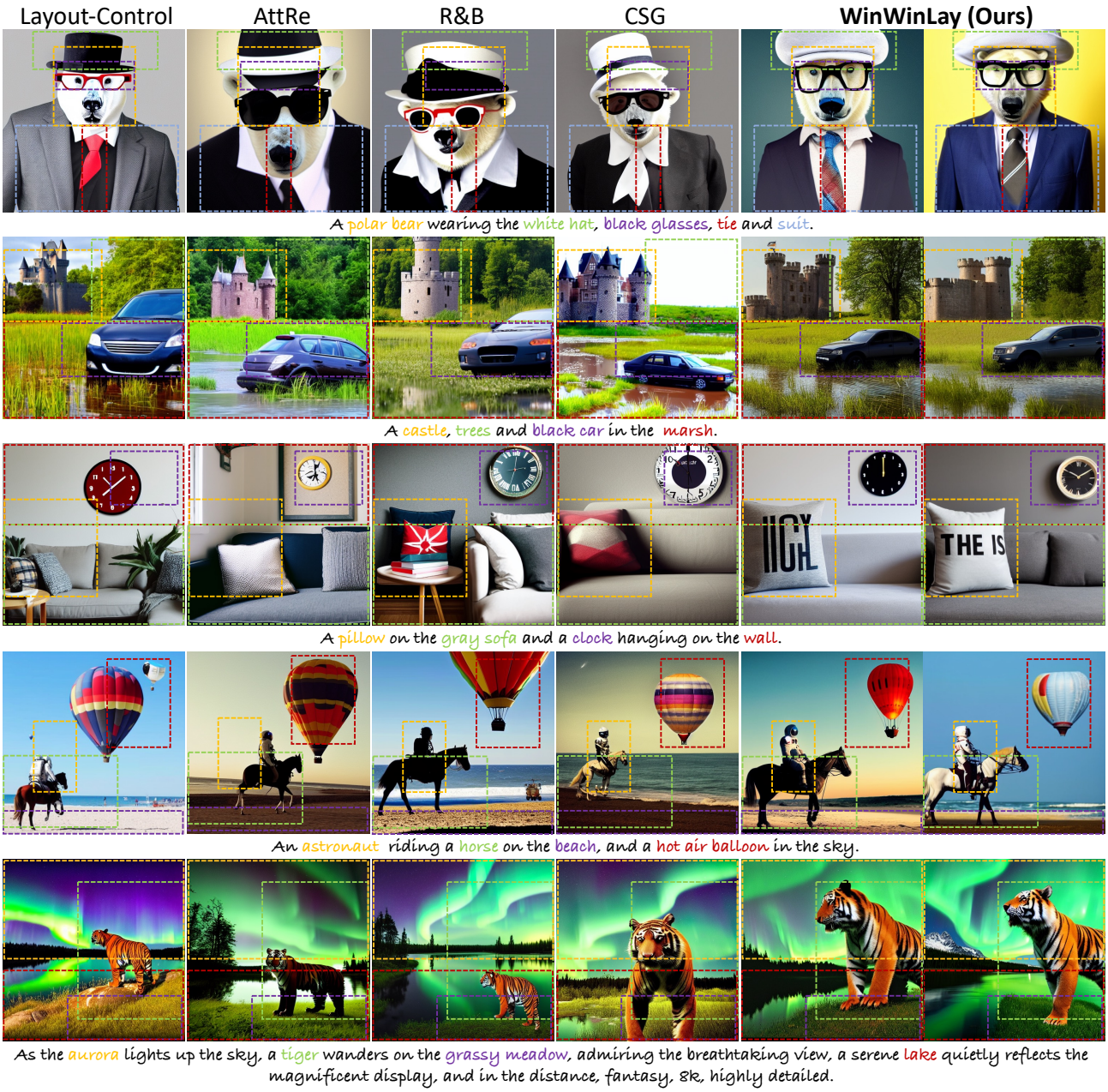

Figure 4: Qualitative comparison of our WinWinLay and state-of-the-art methods. Zoom in for more details.

Model	COCO2014					Flicker30K					User Study	
	AP↑	CLIP-s↑	FID↓	PickScore↑	ImageReward↑	AP↑	CLIP-s↑	FID↓	PickScore↑	ImageReward↑	Controllability↑	Quality↑
Layout-Control (Chen et al., 2024d)	8.42	0.310	29.79	21.09	0.7038	14.19	0.288	29.83	20.39	0.7016	12.1	7.5
AttRe (Phung et al., 2024)	15.51	0.296	27.51	21.23	0.7109	15.26	0.277	27.72	20.64	0.7095	18.7	22.3
R&B (Xiao et al., 2024)	17.63	0.306	28.22	21.16	0.7071	14.80	0.291	28.18	20.58	0.7114	20.6	19.4
CSG (Liu et al., 2024a)	17.58	0.299	27.64	21.22	0.7027	15.11	0.282	27.90	20.51	0.7049	20.9	21.0
Ours	**19.74**	**0.327**	**26.85**	**21.41**	**0.7218**	**17.28**	**0.309**	**27.04**	**20.85**	**0.7202**	**27.7**	**29.8**

Table 1: Quantitative comparison of our WinWinLay and state-of-the-art methods.

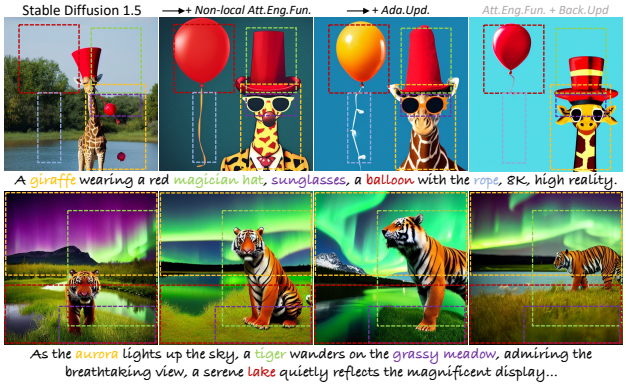

Figure 5: Qualitative ablation on proposed Non-local Attention Energy Function and Adaptive Update. Zoom in for more details.

Model	COCO2014		
	AP_{50}↑	AP↑	CLIP-s↑
Att.Eng.Fun. + Back.Upd.	25.8	8.4	0.310
Non-local Att.Eng.Fun. + Back.Upd.	44.1	17.4	0.318
Att.Eng.Fun. + **Ada.Upd.**	36.7	14.9	0.324
Non-local Att.Eng.Fun. + Ada.Upd.	**49.2**	**19.7**	**0.327**

Table 2: Quantitative ablation on proposed Non-local Attention Energy Function and Adaptive Update on COCO2014.

Implementation Details. We adopt the Stable Diffusion 1.5 (Rombach et al., 2022), pre-trained on the LAION-5B (Schuhmann et al., 2022a), as our base Text-to-Image model. During generation, we employ the DDIM sampler with 50 steps and set the scale guidance to 7.5 for generation. Since layout construction typically occurs during the early stages of denoising, we apply the layout constraint only within the initial 10 steps. The hyperparameters ρ of non-local attention prior is set to 5/0 for max/min, respectively. For adaptive update, we set steps O of Langevin dynamics is set as 4 and signal-to-noise ratio r as 0.06. We observe that these parameters generally work well in most cases, proving the generalizability of WinWinLay. We also point that better results may be obtained with a customized setting, *e.g.*, a larger ρ or more iterations for Langevin dynamics.

5.2. Comparison With SOTA Methods

We compare our approach with four representative state-of-the-arts of Layout-Control (Chen et al., 2024d), AttRe (Phung et al., 2024), R&B (Xiao et al., 2024) and CSG (Liu et al., 2024a) to show the advantage of WinWinLay. All methods are implemented by official codes.

Quantitative Comparison. As presented in Table 1, we first quantitatively evaluate generated images for our test dataset. Compared to methods Layout-Control and AttRe, CSG shows a significant improvement in object placement accuracy. However, we observed in our experiments that it is highly sensitive to gradient strength, where higher accuracy often leads to a severe decline in image quality, especially when generating a large number of objects. In contrast, our method consistently outperforms across multiple datasets and evaluation metrics, demonstrating a more robust improvement. Additionally, we resort to user studies to evaluate which method generates results that are most favored by humans. We conducted two user studies on the results in terms of the Controllability and Quality. In the first study, participants are asked to select the images that best align with the given layout. In the second study, participants are tasked with identifying the images that exhibit the most realistic appearance. To ensure clarity and reproducibility, we conducted the user study on Wenjuanxing, a platform similar to Amazon Mechanical Turk. 150 participants evaluated 50 image pairs, yielding 7500 responses per study. Images were shown side-by-side with layout prompts, and both question order and image positions were randomized to avoid bias. As shown in Table 1, over 27.7% of our results are selected as the best in both two metrics, which proves a significant advantage in generation.

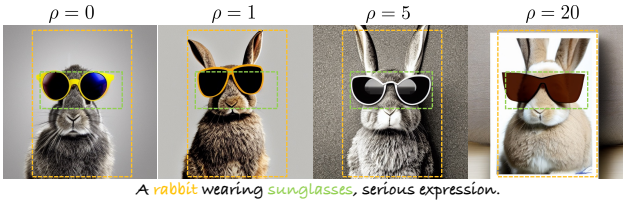

Figure 6: Ablation on hyperparameter of non-local attention prior.

Qualitative Comparison. To present a more detailed and visual comparison of our model, we carry out experiments on a smaller hand-crafted dataset with 3-5 objects. For fair comparison, we generate 10 images for each method under the same random seed and subsequently select the optimal image based on the AP_{50} for display. To demonstrate the effectiveness of WinWinLay, 2 images are presented for each case. From the results in Figure 4, we can draw the following conclusions: (i) Our method effectively places the target object within the given region while filling the entire bounding box without compromising the natural structure of the object, representing a significant improvement over existing methods. In contrast, other approaches often fail to generate images that faithfully adhere to the layout (*e.g.*, 1^{th} row), and parts of the object may collapse into localized regions of the bounding box (*e.g.*, 4^{th} row); (ii) Our method successfully eliminates the trade-off between control and quality, without compromising the generative capability of the underlying model despite the additional layout constraints. However, existing works focus on layout adherence while neglecting the realism of the generated objects (*e.g.*, 3^{th} row). Furthermore, multiple distinct results generated under the same prompt and spatial constraints demonstrate

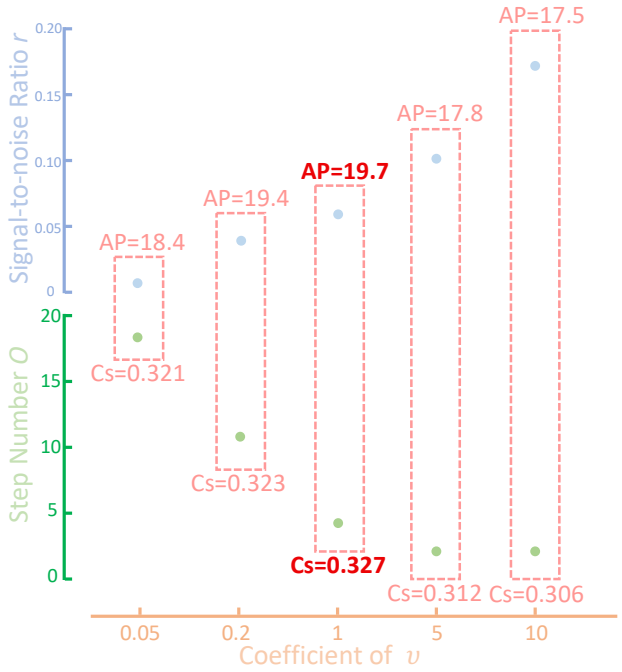

Figure 7: Ablation on hyperparameter of Adaptive Update on COCO2014. Cs denotes CLIP-s metric.

the robustness of WinWinLay, thereby further advancing the progress of Layout-to-Image in practical applications.

5.3. Ablation Study

Effect of Proposed Strategies. To further demonstrate the efficacy of the proposed method, we incrementally introduce Non-local Attention Energy Function and Adaptive Update to the baseline model and observe the resulting changes. As shown in Figure 5, Non-local Attention Energy Function significantly enhances the control over the layout, while also ensuring an accurate representation of all target objects. On the other hand, Adaptive Update not only strengthens the spatial placement accuracy but also improves the overall image quality (*e.g.*, "giraffe" appears more realistic). Furthermore, the quantitative results provided in Table 2 align with the visual observations, with Non-local Attention Energy Function achieving a substantial increase in AP and AP_{50}, while Adaptive Update further refines the spatial positioning and enhances image quality.

Hyperparameter of Non-local Attention Prior. Previous Attention Energy Function often suffers from the problem of attention collapse to local regions. Hence, non-local attention prior is introduced to constrain the attention to focus on the center of the bounding box and gradually expand to cover the entire box. Here, ρ serves as the hyperparameter that controls the strength of the non-local attention prior. As shown in Figure 6, with the gradual increase of ρ, the objects in the image progressively align with the edges of the bounding box, allowing for more precise layout control. However, when ρ becomes too large, it may lead to unnatural object placements within the bounding box, such as a "rabbit" appearing on a square canvas. In our experiments, we find that setting ρ to 5 generally yields optimal results, which is used across all experiments.

Hyperparameter of of Adaptive Update. Adaptive parameter ν plays a critical role in the effectiveness of the proposed Adapative Update and impact the overall performance of WinWinLay. In Section 4.2, we analyzed the impact of different ν on efficiency and performance, proposing an adaptive strategy to significantly reduce the complexity of hyperparameter tuning. To validate its effectiveness, we introduce coefficients of varying magnitudes to the adaptive parameter and conduct grid searches to determine the signal-to-noise ratio r of optimal step size and the number of update steps O for each coefficient. As shown in Figure 7, larger ν generally require larger step sizes and fewer update steps, which substantially degrade both accuracy and quality. Conversely, smaller ν has less pronounced effects on performance but significantly increase generation time.

6. Conclusion

The paper introduces WinWinLay, a novel training-free framework for Layout-to-Image generation, which achieves significant improvements in layout precision and visual fidelity. Addressing limitations in existing methods, WinWinLay incorporates two innovative components: Non-local Attention Energy Function, which ensures uniform attention distribution within specified layouts while preserving natural object structures, and Adaptive Update, which leverages Langevin dynamics to effectively balance layout control and image quality. Extensive experiments on standard benchmarks demonstrate that WinWinLay outperforms state-of-the-art approaches in both controllability and photorealism, making it a robust and efficient solution for L2I tasks.

Impact Statement

This project provides a training-free method for layout-controlled image synthesis, enhancing controllability while preserving generative strength of base models; however, like other generative techniques, it may be misused for disinformation, highlighting the need for future research to address ethical risks associated with layout-guided generation.

Acknowledgements

This paper is supported by National Natural Science Foundation of China under Grants (U23B2012, 12471308), Beijing Natural Science Foundation (1254050), Fundamental Research Funds for the Central Universities, and National Research Foundation, Singapore, under its Medium Sized Center for Advanced Robotics Technology Innovation.

References

Ahn, N., Lee, J., Lee, C., Kim, K., Kim, D., Nam, S.-H., and Hong, K. Dreamstyler: Paint by style inversion with text-to-image diffusion models. In *AAAI*, pp. 674–681, 2024.

Bar-Tal, O., Yariv, L., Lipman, Y., and Dekel, T. Multidiffusion: Fusing diffusion paths for controlled image generation. In *ICML*, pp. 1737–1752, 2023.

Brooks, T., Holynski, A., and Efros, A. A. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, pp. 18392–18402, 2023.

Chen, D.-Y., Tennent, H., and Hsu, C.-W. Artadapter: Text-to-image style transfer using multi-level style encoder and explicit adaptation. In *CVPR*, pp. 8619–8628, 2024a.

Chen, H., Li, J., Zhuang, W., Vikalo, H., and Lyu, L. Boundary attention constrained zero-shot layout-to-image generation. *arXiv preprint arXiv:2411.10495*, 2024b.

Chen, J., Jincheng, Y., Chongjian, G., Yao, L., Xie, E., Wang, Z., Kwok, J., Luo, P., Lu, H., and Li, Z. Pixart-alpha: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *ICLR*, 2024c.

Chen, M., Laina, I., and Vedaldi, A. Training-free layout control with cross-attention guidance. In *WACV*, pp. 5343–5353, 2024d.

Chen, Y., Pan, Y., Yang, H., Yao, T., and Mei, T. Vp3d: Unleashing 2d visual prompt for text-to-3d generation. In *CVPR*, pp. 4896–4905, 2024e.

Couairon, G., Careil, M., Cord, M., Lathuiliere, S., and Verbeek, J. Zero-shot spatial layout conditioning for text-to-image diffusion models. In *ICCV*, pp. 2174–2183, 2023.

Dhariwal, P. and Nichol, A. Diffusion models beat gans on image synthesis. In *NeurIPS*, pp. 8780–8794, 2021.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. C., and Bengio, Y. Generative adversarial nets. In *NeurIPS*, 2014.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. In *NeurIPS*, pp. 6840–6851, 2020.

Ho, J., Saharia, C., Chan, W., Fleet, D. J., Norouzi, M., and Salimans, T. Cascaded diffusion models for high fidelity image generation. *Journal of Machine Learning Research*, 23(47):1–33, 2022.

Jia, C., Luo, M., Dang, Z., Dai, G., Chang, X., Wang, M., and Wang, J. Ssmg: Spatial-semantic map guided diffusion model for free-form layout-to-image generation. In *AAAI*, pp. 2480–2488, 2024.

Jiang, Y., Wu, T., Yang, S., Si, C., Lin, D., Qiao, Y., Loy, C. C., and Liu, Z. Videobooth: Diffusion-based video generation with image prompts. In *CVPR*, pp. 6689–6700, 2024.

Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., Mosseri, I., and Irani, M. Imagic: Text-based real image editing with diffusion models. In *CVPR*, pp. 6007–6017, 2023.

Kim, Y., Lee, J., Kim, J.-H., Ha, J.-W., and Zhu, J.-Y. Dense text-to-image generation with attention modulation. In *ICCV*, pp. 7701–7711, 2023.

Kirstain, Y., Polyak, A., Singer, U., Matiana, S., Penna, J., and Levy, O. Pick-a-pic: An open dataset of user preferences for text-to-image generation. In *NeurIPS*, pp. 36652–36663, 2023.

Kwon, M., Oh, S. W., Zhou, Y., Liu, D., Lee, J.-Y., Cai, H., Liu, B., Liu, F., and Uh, Y. Harivo: Harnessing text-to-image models for video generation. In *ECCV*, pp. 19–36, 2025.

Kynkäänniemi, T., Karras, T., Aittala, M., Aila, T., and Lehtinen, J. The role of imagenet classes in fréchet inception distance. In *ICLR*, 2023.

Li, M., Zhou, P., Liu, J.-W., Keppo, J., Lin, M., Yan, S., and Xu, X. Instant3d: Instant text-to-3d generation. *International Journal of Computer Vision*, pp. 1–17, 2024.

Li, Y., Liu, H., Wu, Q., Mu, F., Yang, J., Gao, J., Li, C., and Lee, Y. J. Gligen: Open-set grounded text-to-image generation. In *CVPR*, pp. 22511–22521, 2023.

Li, Z., Wu, J., Koh, I., Tang, Y., and Sun, L. Image synthesis from layout with locality-aware mask adaption. In *ICCV*, pp. 13819–13828, 2021.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *ECCV*, pp. 740–755, 2014.

Liu, J., Huang, T., and Xu, C. Training-free composite scene generation for layout-to-image synthesis. In *ECCV*, pp. 37–53, 2024a.

Liu, S., Ma, A., Wu, X., Leng, D., Yin, Y., et al. Hico: Hierarchical controllable diffusion model for layout-to-image generation. In *NeurIPS*, 2024b.

Mokady, R., Hertz, A., Aberman, K., Pritch, Y., and Cohen-Or, D. Null-text inversion for editing real images using guided diffusion models. In *CVPR*, pp. 6038–6047, 2023.

Navon, A., Shamsian, A., Achituve, I., Maron, H., Kawaguchi, K., Chechik, G., and Fetaya, E. Multi-task learning as a bargaining game. *ArXiv*, abs/2202.01017, 2022.

Parmar, G., Kumar Singh, K., Zhang, R., Li, Y., Lu, J., and Zhu, J.-Y. Zero-shot image-to-image translation. In *SIGGRAPH*, pp. 1–11, 2023.

Peebles, W. and Xie, S. Scalable diffusion models with transformers. In *ICCV*, pp. 4195–4205, 2023.

Phung, Q., Ge, S., and Huang, J.-B. Grounded text-to-image synthesis with attention refocusing. In *CVPR*, pp. 7932–7942, 2024.

Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., and Lazebnik, S. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, pp. 2641–2649, 2015.

Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., and Rombach, R. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

Qing, Z., Zhang, S., Wang, J., Wang, X., Wei, Y., Zhang, Y., Gao, C., and Sang, N. Hierarchical spatio-temporal decoupling for text-to-video generation. In *CVPR*, pp. 6635–6645, 2024.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *ICML*, pp. 8748–8763, 2021.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *CVPR*, pp. 10684–10695, 2022.

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pp. 234–241, 2015.

Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., and Aberman, K. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, pp. 22500–22510, 2023.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, pp. 36479–36494, 2022.

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S., Crowson, K., Schmidt, L., Kaczmarczyk, R., and Jitsev, J. Laion-5b: An open large-scale dataset for training next generation image-text models. *ArXiv*, abs/2210.08402, 2022a.

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al. Laion-5b: An open large-scale dataset for training next generation image-text models. In *NeurIPS*, pp. 25278–25294, 2022b.

Shi, J., Xiong, W., Lin, Z., and Jung, H. J. Instantbooth: Personalized text-to-image generation without test-time finetuning. In *CVPR*, pp. 8543–8552, 2024.

Singh, J., Gould, S., and Zheng, L. High-fidelity guided image synthesis with latent diffusion models. In *CVPR*, pp. 5997–6006, 2023.

Sohn, K., Jiang, L., Barber, J., Lee, K., Ruiz, N., Krishnan, D., Chang, H., Li, Y., Essa, I., Rubinstein, M., et al. Styledrop: Text-to-image synthesis of any style. In *NeurIPS*, 2024.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. In *ICLR*, 2021a.

Song, Y. and Ermon, S. Generative modeling by estimating gradients of the data distribution. In *NeurIPS*, 2019.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021b.

Tosi, F., Ramirez, P. Z., and Poggi, M. Diffusion models for monocular depth estimation: Overcoming challenging conditions. In *ECCV*, pp. 236–257, 2025.

Tumanyan, N., Geyer, M., Bagon, S., and Dekel, T. Plug-and-play diffusion features for text-driven image-to-image translation. In *CVPR*, pp. 1921–1930, 2023.

Wang, C.-Y., Bochkovskiy, A., and Liao, H.-Y. M. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *CVPR*, pp. 7464–7475, 2023.

Wang, X., Darrell, T., Rambhatla, S. S., Girdhar, R., and Misra, I. Instancediffusion: Instance-level control for image generation. In *CVPR*, pp. 6232–6242, 2024a.

Wang, Z., Lu, C., Wang, Y., Bao, F., Li, C., Su, H., and Zhu, J. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. In *NeurIPS*, 2024b.

Wu, J. Z., Ge, Y., Wang, X., Lei, S. W., Gu, Y., Shi, Y., Hsu, W., Shan, Y., Qie, X., and Shou, M. Z. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *ICCV*, pp. 7623–7633, 2023.

Wu, Y., Zhou, X., Ma, B., Su, X., Ma, K., and Wang, X. Ifadapter: Instance feature control for grounded text-to-image generation. *arXiv preprint arXiv:2409.08240*, 2024.

Xiao, J., Lv, H., Li, L., Wang, S., and Huang, Q. R&b: Region and boundary aware zero-shot grounded text-to-image generation. In *ICLR*, 2024.

Xie, J., Li, Y., Huang, Y., Liu, H., Zhang, W., Zheng, Y., and Shou, M. Z. Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In *ICCV*, pp. 7452–7461, 2023.

Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J., and Dong, Y. Imagereward: Learning and evaluating human preferences for text-to-image generation. In *NeurIPS*, pp. 15903–15935, 2023.

Xu, S., Huang, Y., Pan, J., Ma, Z., and Chai, J. Inversion-free image editing with language-guided diffusion models. In *CVPR*, pp. 9452–9461, 2024.

Xue, H., Huang, Z., Sun, Q., Song, L., and Zhang, W. Freestyle layout-to-image synthesis. In *CVPR*, pp. 14256–14266, 2023.

Yang, Z., Wang, J., Gan, Z., Li, L., Lin, K., Wu, C., Duan, N., Liu, Z., Liu, C., Zeng, M., et al. Reco: Region-controlled text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14246–14255, 2023.

Zhang, Z., Li, B., Nie, X., Han, C., Guo, T., and Liu, L. Towards consistent video editing with text-to-image diffusion models. In *NeurIPS*, pp. 58508–58519, 2023.

Zheng, G., Zhou, X., Li, X., Qi, Z., Shan, Y., and Li, X. Layoutdiffusion: Controllable diffusion model for layout-to-image generation. In *CVPR*, pp. 22490–22499, 2023.

Zhou, D., Li, Y., Ma, F., Zhang, X., and Yang, Y. Migc: Multi-instance generation controller for text-to-image synthesis. In *CVPR*, pp. 6818–6828, 2024.