FreLay: Frequency-Aware Energy Function for Training-Free Layout-to-Image Generation

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

[1]University of Chinese Academy of Sciences
[2]National University of Singapore
libonan@ucas.ac.cn huyinhan16@mails.ucas.ac.cn songhua.liu@u.nus.edu {zeyuxiao,xinchao}@nus.edu.sg

Abstract

Layout-to-Image generation has significantly advanced content creation by enabling the rendering of visual text under predefined spatial layouts. Current approaches achieve training-free layout guidance by constructing attention-based energy functions to derive correction gradients. In this paper, we demonstrate that vanilla energy functions suffer from two limitations, resulting in imprecise layout control and visually unrealistic artifacts. First, the normalizing factor of the Boltzmann distribution defined by the energy functions is non-negligible when calculating correction gradients, yet current energy functions cannot compute this factor exactly. Furthermore, while attention varies over time during the denoising process, existing approaches employ a fixed formulation. To address these challenges, we introduce ***FreLay***, a novel training-free approach equipped with a frequency-aware energy function. Our method first reformulates the energy function to handle the normalization factor, enabling accurate computation of correction gradients. Simultaneously, leveraging the prior knowledge that low-frequency information deteriorates slower during noise addition, we design a time-specific energy function for each timestep from a frequency-domain perspective. Experimental results demonstrate that FreLay consistently outperforms existing state-of-the-art training-free methods by a large margin both qualitatively and quantitatively across multiple datasets.

Introduction

Recent advancements in Text-to-Image (T2I) generation (Rombach et al. 2022) have yielded significant progress, facilitating breakthroughs across diverse domains, *e.g.* style translation (Li et al. 2025b), image restoration (Qiu et al. 2023), image editing (Zhang et al. 2023) and Text-to-3D creation (Li et al. 2025c). However, these models still lack explicit structural reasoning capabilities. When handling complex scenes with multiple objects and precise spatial layouts, they often struggle to offer reliable controllability, which limits their applicability to structured visual content creation. This drawback has spurred growing interest in Layout-to-Image (L2I) approaches, which aim to synthesize images that faithfully reflect textual descriptions while adhering to spatial constraints.

A straightforward approach (Kim et al. 2023a; Zhou et al. 2024a; Wu et al. 2024) to obtaining L2I models is to finetune powerful T2I models with spatial conditioning. However, such methods entail substantial training costs and require large-scale, resource-intensive paired data. To enable more efficient adaptation of T2I models for L2I tasks, a widely adopted strategy (Chen, Laina, and Vedaldi 2024; Mo et al. 2024; Phung, Ge, and Huang 2024; Xiao et al. 2024; Li et al. 2025a) is to incorporate spatial conditions during the denoising stage, thereby avoiding the need for expensive training. Among these methods, backward guidance (Chen, Laina, and Vedaldi 2024) has emerged as the dominant paradigm by extending score-based diffusion models (*e.g.*, Stable Diffusion (Rombach et al. 2022)) into layout-guided conditional score functions and leveraging correction gradients to steer the generation process. Specifically, it defines an energy function that encourages the cross-attention layer to allocate more attention weights to user-specified regions (*e.g.*, bounding box), and injects constraint signals into the sampling trajectory via gradient estimation. Although these works offer high efficiency, they remain inadequate for real-world applications where precise spatial control and high visual fidelity are essential.

To bridge this gap, we initiate our investigation with a comprehensive analysis of existing attention energy functions and theoretically demonstrate that, despite their empirical effectiveness, they inherently suffer from limitations in spatial controllability and visual fidelity (see Fig.2). First, the normalization factor in the Boltzmann distribution defined by the energy functions is indeed a function of the current image. Consequently, accurate computation of correction gradients necessitates computing and differentiating this factor. However, due to its formulation as an intractable integral, the normalization factor cannot be analytically handled using existing energy functions. Second, they impose identical spatial constraints on attention maps across all timesteps. Note that even if the final synthesized image aligns perfectly with the given layout, the attention maps evolve from coarse to fine during the denoising process. Therefore, dynamically modifying spatial constraints throughout the generation process is crucial.

In this paper, we propose a novel training-free approach,

*These authors contributed equally.
†Corresponding author.

Figure 1: Given user-specified bounding boxes and textual prompts, our ***FreLay*** framework generates controllable and realistic images using a pre-trained Text-to-Image diffusion model, such as SDXL (Podell et al. 2024), without requiring any fine-tuning on paired data.

FreLay, which addresses the above limitations by redesigning a time-specific energy function that enables analytical treatment of the normalization factor. To handle the normalization factor, we reformulate the energy function as the squared Euclidean distance, reducing its Boltzmann distribution to a Gaussian distribution. Across timesteps, attention exhibits less variation in low-frequency components than in high-frequency ones. So low-frequency components contain more reliable information at every timestep, which means modeling low frequencies needs smaller variance compared to high frequencies. We therefore construct our time-specific energy function by adaptively modulating the variance of Gaussian distributions across timesteps. Specifically, we transform the distance into the frequency domain, assigning smaller variance to low-frequency components than high-frequency ones, with this variance gap progressively narrowing during the denoising process. Through this frequency energy function, FreLay provides temporally varying constraints and an analytical correction gradient, which enables precise spatial control over the placement and arrangement of subjects during synthesis.

Our contributions are fourfold: (i) We analyze inherent flaws in existing attention energy functions and propose FreLay, a training-free layout-guided generation framework.

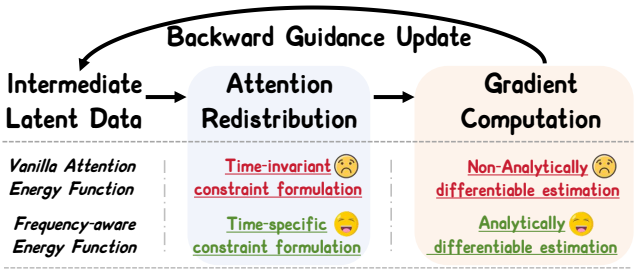

Figure 2: Energy function plays a central role in the backward guidance update by influencing both attention redistribution and correction gradient computation.

(ii) We design an Anchor Attention Energy Function that resolves normalization issues and enables precise gradient correction for spatial control. (iii) We introduce a Timestep Adaptivity Extension from the frequency domain to align spatial constraints with progressive denoising. (iv) Extensive experiments show that FreLay surpasses state-of-the-art training-free methods in controllability and visual quality.

Related Work

Text-to-Image Generation

Diffusion models (Ho, Jain, and Abbeel 2020; Nichol and Dhariwal 2021; Karras et al. 2022; Xu et al. 2023b) have recently surpassed generative adversarial networks (Goodfellow et al. 2020) in computer vision (Xiao et al. 2021, 2023; Xiao, Li, and Jia 2025; Xiao and Xiong 2025), driving significant progress in Text-to-Image generation (Rombach et al. 2022; Esser et al. 2024). Leveraging large-scale image-text datasets (Schuhmann et al. 2022), these models produce diverse, high-quality images guided by text and have been extended to other tasks. Despite these advances, fine-grained spatial control, such as layout alignment, remains challenging, limiting real-world applicability.

Layout-to-Image Generation

Layout-to-Image (L2I) methods (Xue et al. 2023; Zheng et al. 2023; Xie et al. 2023; Jia et al. 2024; Liu et al. 2024; Liu, Huang, and Xu 2024) aim to generate images that satisfy both textual prompts and explicit spatial constraints like bounding boxes or sketches. A common solution is fine-tuning large Text-to-Image models on paired layout–image datasets (Li et al. 2023; Zhou et al. 2024b,a; Wu et al. 2024), but this is costly and impractical. To avoid such bottlenecks, recent work (Bar-Tal et al. 2023; Kim et al. 2023b; Singh, Gould, and Zheng 2023; Couairon et al. 2023; Mo et al. 2024; Phung, Ge, and Huang 2024; Xiao et al. 2024; Li et al. 2025a) adopts training-free strategies by embedding layout information into the denoising process through forward or backward guidance. While efficient, these approaches often produce misaligned content and visual artifacts. In this paper, we provide a theoretical analysis of backward guidance and introduce an enhanced energy function to improve spatial accuracy and visual fidelity.

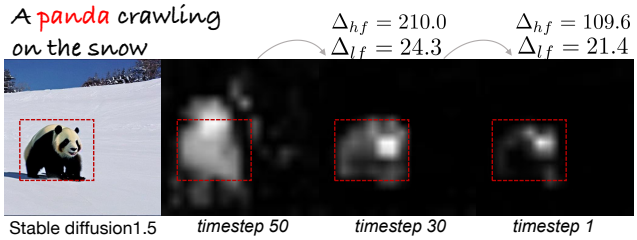

Figure 3: Visualization of cross-layer attention maps in the original Text-to-Image model illustrates that the attention maps progressively refine during the denoising process to better align with the bounding boxes, rather than strictly matching them from the beginning. Furthermore, we compute the variation magnitudes for high- and low-frequency components, denoted as Δ_{hf} and Δ_{lf}[1], respectively. The results clearly indicate that low-frequency information remains relatively stable, whereas high-frequency information exhibits significant fluctuations.

Preliminaries

Text-to-Image Generation

We build upon the Latent Diffusion Model (LDM) (Singh, Gould, and Zheng 2023) as our Text-to-Image backbone, which defines a generative process that progressively transforms Gaussian noise into an image conditioned on a text prompt. Concretely, LDM first maps an input image x_0 into a lower-dimensional latent $z_0 = \mathbf{E}(x_0)$ via a frozen, pre-trained encoder $\mathbf{E}(\cdot)$. The latent diffusion process then corrupts z_0 with noise at each timestep t, producing z_t. A corresponding frozen text encoder $\Phi(\cdot)$ converts the prompt p into token embeddings $y = \Phi(p)$. The denoiser network $\epsilon_\theta(\cdot)$, typically realized as a U-Net with convolutional, self-attention, and cross-attention layers, is trained to predict and remove the injected noise by minimizing

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

Finally, the clean latent $\hat{z}_0$ is decoded back into image space via a frozen decoder $\mathbf{D}(\cdot)$, yielding the reconstruction $\hat{x}_0 = \mathbf{D}(\hat{z}_0)$ that matches the given prompt.

Backward Guidance for Layout-to-Image

To precisely position the subject at the specified location, backward guidance aims to sample images from the conditional distribution $p(z|y, b, i)$, where b_i is the bounding box corresponding to the text token y_i of the target subject. Specifically, backward guidance introduces an optimizable attention energy function (Chen, Laina, and Vedaldi 2024) $\mathcal{E}_{\text{aef}}$ to redistribute the cross-attention map a from the feature z, thereby encouraging the i^{th} token's attention to concentrate within region b_i:

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

[1] Δ_f represents the mean distance of all frequencies in some set. $\Delta_f(P, Q) = \frac{1}{|S|}\sum_{k\in S} \|P_k - Q_k\|$

where $a_{ui}^{(\gamma)}$ measures the attention weight between spatial location u in the cross-attention layer γ and token y_i, and m_i is the binary mask derived from b_i, with pixels inside the box set to 1 and pixels outside set to 0. For clarity, we omit the token index i and layer index γ in the following. A simple way to continuously bias the attention map during generation is to compute the gradient of the energy function and backpropagate it to update the initial latent feature $\bar{z} = z$ via:

$$z \leftarrow \bar{z} - \eta \nabla_z \mathcal{E}_{\text{aef}}(a, m). \tag{3}$$

To ensure that updates remain within the data manifold while satisfying layout constraints, a recent method (Li et al. 2025a) introduces Langevin dynamics to achieve superior visual fidelity. Specifically, it defines a Gibbs distribution $p(m|z_t) \propto \exp(-\nu \mathcal{E}_{aef}(z_t, m))$, based on the energy function $\mathcal{E}_{aef}$. By applying Bayes' theorem, the conditional distribution is expressed as $p(z_t|m) \propto p(z_t)p(m|z_t)$, which characterizes the relationship between latent features and spatial constraints. This allows the $\mathcal{E}_{aef}$ to be reformulated as a conditional score function, $\nabla_{z_t} \log(p(z_t|m))$, which can be decomposed into two components:

$$\nabla_{z_t} \log(p(z_t|m)) = \nabla_{z_t} \log(p(z_t)) + \nabla_{z_t} \log(p(m|z_t)), \tag{4}$$

where the first term, $\nabla_{z_t} \log(p(z_t))$, is approximated by a pre-trained unconditional score estimator, and the second term, $\nabla_{z_t} \log(p(m|z_t))$, serves as a correction gradient that plays a critical role in injecting conditional information. Finally, updated latent feature can be obtained by :

$$z_t \leftarrow \bar{z}_t + \xi \nabla_{\bar{z}_t} \log(p(\bar{z}_t|m)) + \sqrt{2\xi}\epsilon, \tag{5}$$

where $\epsilon \sim \mathcal{N}(0, I)$ and ξ are the hyperparameters.

Proposed Method

In this section, we introduce FreLay, a novel training-free method for Layout-to-Image generation. First, we revisit existing attention energy functions and articulate two principal shortcomings. Next, the core Frequency-Aware Energy Function, grounded in rigorous theoretical analysis, is introduced to effectively mitigate these issues and substantially enhance generation quality.

Motivation

We commence with a brief revisit of the existing backward guidance update mechanism, demonstrating its inherent limitations, particularly the deficiencies in the current attention energy functions and their implications for the overall generation process.

Revisiting of Correction Gradient Computation. To enable the attention energy function to effectively guide the generative process, accurately computing the correction gradient is essential. Following the Boltzmann distribution, the conditional probability $p(m|z_t)$ can be rigorously formulated as:

$$p(m|z_t) = \frac{1}{C(z_t)} \exp(-\nu \mathcal{E}_{\text{aef}}(z_t, m)), \tag{6}$$

where $C(z_t)$ is the normalizing factor, given by:

$$C(z_t) = \int \exp(-\nu \mathcal{E}_{\text{aef}}(z_t, m))\, dm. \tag{7}$$

Then, the corresponding correction gradient can be expressed as follows:

$$\nabla_{z_t} \log p(m|z_t) = -\nu \nabla_{z_t} \mathcal{E}_{\text{aef}}(z_t, m) - \nabla_{z_t} \log C(z_t). \tag{8}$$

Note that current methods typically compute correction gradients with only the first term of the Equ.(8) while neglecting the second term, either due to distinct modeling approaches or simply by mistake, which leads to sub-optimal results. However, for a general attention energy function, computing $C(z_t)$ is intractable. Therefore, it is crucial to design an energy function that enables exact computation of the second term of Equ.(8), thereby ensuring accurate correction gradient computation and facilitating effective spatial control.

Revisiting of Attention Redistribution. Building on the formulation of the vanilla attention energy function $\mathcal{E}_{\text{aef}}$, we observe that existing modeling strategies typically enforce a fixed-size masked bounding box m across all denoising steps t. However, this design disregards a fundamental property of the generative process, which is the progressive refinement of attention maps from coarse to fine. In the early stages, attention distributions are inherently diffuse, capturing only approximate object layouts, whereas later steps demand more localized and precise spatial focus. As illustrated in Fig.3, the cross-attention map initially allocates an excessively large region to the object, which gradually contracts to match its actual size in the final image. Applying a static spatial constraint throughout this dynamic process creates a mismatch between the evolving attention patterns and the fixed spatial prior, leading to inaccurate object placement and degraded generation quality. To address this limitation, it is essential to develop strategies that adaptively synchronize spatial constraints with the natural evolution of attention during generation.

Frequency-Aware Energy Function

Building on the above motivation, we propose a time-specific energy function that allows for analytical computation of the correction gradient. We first construct an anchor energy function on clean data z_0, formulated as a squared Euclidean distance. Consequently, the distribution $p(m|z_0)$ adopts a Gaussian form, resulting in the normalization factor no longer being dependent on z_0. Subsequently, based on the differences of attention in the frequency domain, we generalize the formulation and derive time-specific energy functions from a spectral perspective for modeling $p(m|z_t)$.

Anchor Attention Energy Function. We start by modeling the canonical conditional distribution $p(m|z_0)$, which represents the likelihood of a particular mask m given a clean latent representation z_0. Reminding that our goal is to adjust z_0 such that its probability under this distribution is maximized. This distribution relies on the attention map obtained from z_0. Let a_0 denote the unnormalized attention and p its normalized counterpart. That is said we only need model $p(m|a_0)$ to get $p(m|z_0)$. Intuitively, a higher attention mass inside the mask region should correspond to a higher probability, which can be captured by the term $p \cdot m$. However, if the mask were allowed to cover the entire image, this term would always be maximal, leading to a trivial solution. To

address this, a regularization term is introduced based on the mask size, resulting in the following energy function $\mathcal{E}_{\text{aaef}}$:

$$\mathcal{E}_{\text{aaef}} = -p \cdot m + \lambda m \cdot m = \lambda \|m - \frac{1}{2\lambda} p\|^2 - \frac{1}{4\lambda} \|p\|^2, \quad (9)$$

where a smaller energy indicates a higher probability. Since p is given when modeling this distribution and can be treated as a constant, we can add $\frac{1}{4\lambda}\|p\|^2$ uniformly and then the energy function takes the form of a l_2 distance,

$$\mathcal{E}_{\text{aaef}} = \lambda \|m - sp\|^2, \quad (10)$$

with λ and s as hyperparameters. The corresponding conditional distribution is then:

$$p(m|a_0) \propto \exp(-\nu \|m - sp\|^2), \quad (11)$$

which effectively resembles the Gaussian distribution in the mask space. Here, we absorb λ into the distribution parameter ν to avoid introducing excessive hyperparameters. Although computing the normalization factor $C = \int \exp(-\nu \|m - sp\|^2)\, dm$ is still intractable under the binary mask constraint, this is not a practical issue in our setting because the mask is fixed during optimization. Consequently, we relax the mask domain to $\mathbb{R}^n$, making it a rigorous Gaussian distribution, resulting in that the normalization factor is independent of a_0 and irrelevant for the correction gradient computation. The conditional score function is:

$$\nabla_{a_0} \log p(m|a_0) = -\nu \nabla_{a_0} (-2sm \cdot p + s^2 \|p\|^2), \quad (12)$$

which suggests that the update should both increase attention mass inside the mask $m \cdot p$ and penalize excessive global attention spread $\|p\|^2$, implicitly acting as a regularizer.

Timestep Adaptivity Extension. To extend the formulation from $p(m|z_0)$ (*i.e.*, $p(m|a_0)$) to $p(m|z_t)$, a natural strategy would be to first estimate a clean latent $\hat{z}_0$ from z_t and then apply the previously defined distribution $p(m|z_0)$. Existing approaches using a fixed energy function are equivalent to adopting the simplest approximation by setting $\hat{z}_0 = z_t$, ignoring the degradation introduced by noise. However, a closer examination of the attention maps a_t, where a_t obtained from z_t, reveals an important observation: their low-frequency components remain much more consistent than their high-frequency counterparts (see Fig.3). Empirically, predicting the low-frequency part of a_0 from a_t yields lower variance and higher accuracy, whereas high-frequency prediction exhibits significantly larger uncertainty. This phenomenon can be explained by the diffusion process, which corrupts high-frequency information much faster than low-frequency structures (Falck et al. 2025). Building on this insight, instead of explicitly modeling $p(a_0|a_t)$, we directly adapt the energy function in Equ.(10) to define $p(m|z_t)$ from a spectral perspective. Let $M = \mathcal{F}(m)$ and $P = \mathcal{F}(p)$ denote the Fourier transforms of the mask and attention map, respectively. By Parseval's theorem, we have $\|m - sp\|^2 = \|M - sP\|^2$ and frequency energy function $\mathcal{E}_{\text{faef}}$ can be formulated as:

$$\mathcal{E}_{\text{faef}} = \|M - sP\|_A^2 = (M - sP)^H A (M - sP). \quad (13)$$

where A is a diagonal matrix that controls the weight of each frequency component's contribution to the overall distance. Subsequently, we formulate the conditional distribution as:

$$\begin{aligned} p(m|a_t) &\propto \exp(-(M - sP)^H \Gamma (M - sP)) \\ &= \exp(-\sum_j \nu_j \|M_j - sP_j\|^2). \end{aligned} \quad (14)$$

Different from Equ.(11), to capture the varying reliability of different frequencies, the scalar ν is now becoming a diagonal matrix Γ, where each diagonal entry ν_j controls the variation for j specific frequency. Intuitively, high-frequency components, which are less reliable, are assigned smaller ν, while low-frequency components receive larger ν. Hence, the corresponding conditional score function becomes:

$$\begin{aligned} \nabla_{a_t} \log p(m|a_t) &= -\nabla_{a_t} (M - sP)^H \Gamma (M - sP) \\ &= -\nabla_{a_t} (-2s \sum_j \nu_j \overline{M_j} P_j + s^2 \sum_j \nu_j \overline{P_j} P_j). \end{aligned} \quad (15)$$

Actually, this formulation introduces two key effects. The first term promotes the weighted correlation. Given that smaller weights are assigned to high frequency components, the layout guidance emphasizes alignment in low-frequency regions where attention remains more stable. The second term serves as a frequency-aware regularizer: it penalizes low-frequency energy more heavily than high-frequency energy, preventing excessive suppression of high-frequency details. Note that Equ.(15) emphasizes alignment in the low-frequency domain, effectively aligning with a low-pass filtered mask, thereby altering the original mask accordingly. In practice, we designate the central $n \times n$ region of the Fourier spectrum as the low-frequency band and treat all remaining components as high-frequency. We assign $\nu_{\text{low}} > 1$ to the low-frequency part and $\nu_{\text{high}} < 1$ to the high-frequency part. Additionally, we linearly decrease (increase) ν to 1 throughout the denoising steps for low-frequency (high-frequency) components to ensure that the mask region remains consistent with the evolving attention.

Experiments

In this section, we first describe the experimental setup, followed by both qualitative and quantitative evaluations comparing our method with state-of-the-art Layout-to-Image approaches. We further conduct ablation studies to validate the effectiveness of the proposed components.

Experimental Setup

Evaluation Benchmarks. Following prior work (Li et al. 2025a), we evaluate FreLay on two widely used benchmarks: COCO2014 (Lin et al. 2014) and Flickr30K (Plummer et al. 2015). For quantitative performance assessment, we adopt YOLOv7 (Wang, Bochkovskiy, and Liao 2023) for object detection and report Average Precision (Li et al. 2021), which reflects the accuracy of object localization and generation. To measure semantic alignment between images and text, we employ CLIP score (Radford et al. 2021). In addition, we use FID (Kynkäänniemi et al. 2023), PickScore (Kirstain et al. 2023), and ImageReward (Xu et al. 2023a) to assess image quality. Following common practice,

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

we set the text template to "A photo of [prompt]" to produce more realistic results.

Implementation Details. To ensure a fair comparison, we adopt Stable Diffusion 1.5 (Rombach et al. 2022), pre-trained on LAION-5B (Schuhmann et al. 2022), as our base Text-to-Image model. Generation is performed using the DDIM sampler with 50 steps and a guidance scale of 7.5 on 1 NVIDIA A100. Since spatial layout construction primarily occurs during the early stages of denoising, we apply the proposed layout constraint only within the first 10 steps, specifically on the middle and initial upsampling layers. We set n as 4, with the hyperparameter ν_{low} configured to 3 and ν_{high} configured to 0.01. These values consistently produce stable results across diverse cases, demonstrating the robustness and generalizability of FreLay. While these default settings work well in practice, further performance gains could be achieved through case-specific tuning.

Comparison with SOTA Methods

Quantitative Comparison. As shown in Tab.1, we report quantitative comparisons across two benchmarks under identical settings, adopting WinWinLay's configurations for fairness. FreLay achieves the highest performance on nearly all metrics, surpassing prior training-free methods (AttRe, R&B, CSG, WinWinLay). On COCO2014, it attains an AP of 21.91, exceeding the second-best by 2.17 points and achieving the best FID, indicating superior spatial alignment and visual fidelity. On Flickr30K, FreLay maintains a consistent lead with a 1.39-point AP gain over WinWinLay, along with the highest CLIP, PickScore, and ImageReward scores, reflecting strong semantic and perceptual quality. A large-scale user study with 150 participants and 7,500 responses further confirms these advantages—32.9% preferred FreLay for controllability and 26.6% for visual quality, both exceeding all competitors.

Model	COCO2014					Flicker30K					User Study	
	AP↑	CLIP-s↑	FID↓	PickScore↑	ImageReward↑	AP↑	CLIP-s↑	FID↓	PickScore↑	ImageReward↑	Controllability↑	Quality↑
AttRe (Phung, Ge, and Huang 2024)	15.51	0.296	27.51	21.23	0.7109	15.26	0.277	27.72	20.64	0.7095	13.5	17.2
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	11.3	14.2
CSG (Liu, Huang, and Xu 2024)	17.58	0.299	27.64	21.22	0.7027	15.11	0.282	27.90	20.51	0.7049	19.8	16.5
WinWinLay (Li et al. 2025a)	19.74	0.327	26.85	21.41	0.7218	17.28	0.309	27.04	20.85	0.7202	22.5	25.5
Ours	**21.91**	**0.339**	**26.39**	**21.52**	**0.7244**	**18.67**	**0.311**	**26.80**	**20.99**	**0.7278**	**32.9**	**26.6**

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

Model	COCO2014		
	AP↑	CLIP-s↑	FID↓
Att.Eng.Fun.	14.72	0.328	27.04
Anc.Att.Eng.Fun.	19.92	0.331	26.55
Anc.Att.Eng.Fun. + Time.Ada.Ext.	**21.91**	**0.339**	**26.39**

Table 2: Quantitative ablation on proposed Anchor Attention Energy Function and Timestep Adaptivity Extension on COCO2014.

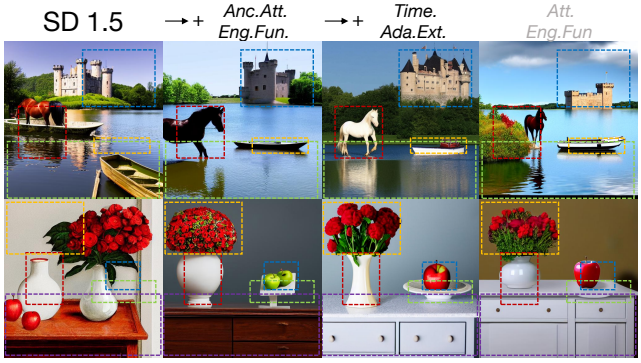

Figure 5: Qualitative ablation on proposed Anchor Attention Energy Function and Timestep Adaptivity Extension. Zoom in for more details.

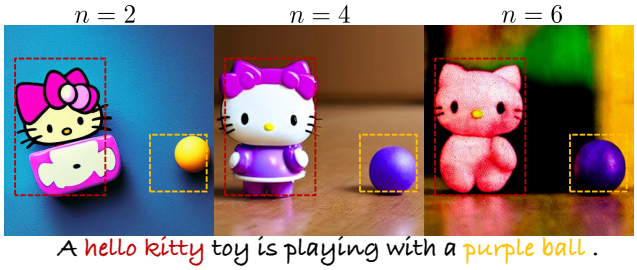

Figure 6: Ablation on Timestep Adaptivity Extension.

Qualitative Comparison. To illustrate generation quality and spatial controllability, we present qualitative comparisons in Fig.4 using diverse text prompts and layouts. For fairness, each method produces 10 samples under the same random seed, and the top-AP result is visualized. Two representative examples are shown per case. FreLay precisely places target objects within their bounding boxes, while other methods often suffer from misalignment, missing elements, or semantic confusion—especially in multi-object scenes where semantic content is misplaced. Moreover, FreLay generates diverse yet coherent results under identical prompts and constraints, demonstrating strong robustness and adaptability for practical L2I generation.

Ablation Study

Effect of Proposed Strategies. Fig.5 and Tab.2 show the progressive gains from our components. Incorporating the Anchor Attention Energy Function markedly improves spatial alignment, with most objects appearing near their target positions, though minor inconsistencies persist for small objects (*e.g.*, apples). Adding the Timestep Adaptivity Extension further refines layout adherence and preserves object integrity—for instance, the horse and boat are correctly positioned in the lake, and the red-flower vase fills its region naturally. In contrast, the vanilla Attention Energy Function often yields misplaced or overlapping objects. These results confirm that both anchor-based energy and timestep adaptation are vital for precise spatial control.

Hyperparameter of Timestep Adaptivity Extension. To evaluate the role of separating high- and low-frequency regions, we conduct an ablation on the parameter n. As shown in Fig.6, larger n increases the weight of low-frequency components, enhancing spatial alignment but yielding coarser details. Conversely, smaller n reduces low-frequency influence, improving detail sharpness but causing spatial misalignment and structural artifacts. Empirically, $n = 4$ provides the best overall balance and is used as the default. The tuning of ν_{low} and ν_{high} exhibits a similar trend, and after fixing n, we select their optimal values via grid search. These parameters can be adjusted for specific cases to further refine results.

Conclusion

In this paper, we present FreLay, a novel training-free Layout-to-Image generation method that overcomes key limitations of existing attention energy functions. Through theoretical analysis, we expose fundamental flaws in the vanilla formulation and introduce a Frequency-Aware Energy Function that precisely computes correction gradients and adaptively enforces spatial constraints during generation. Extensive experiments demonstrate that FreLay consistently outperforms state-of-the-art training-free baselines across multiple benchmarks.

Acknowledgments

This work is supported by Beijing Natural Science Foundation (1254050), National Natural Science Foundation of China (U23B2012, 12431012, 12471308), Fundamental Research Funds for the Central Universities, and Ministry of Education, Singapore, under its Academic Research Fund Tier 2 (Award Number: MOE-T2EP20122-0006).

References

Bar-Tal, O.; Yariv, L.; Lipman, Y.; and Dekel, T. 2023. MultiDiffusion: Fusing Diffusion Paths for Controlled Image Generation. In *ICML*, 1737–1752.

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

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

Esser, P.; Kulal, S.; Blattmann, A.; Entezari, R.; Müller, J.; Saini, H.; Levi, Y.; Lorenz, D.; Sauer, A.; Boesel, F.; et al. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*.

Falck, F.; Pandeva, T.; Zahirnia, K.; Lawrence, R.; Turner, R.; Meeds, E.; Zazo, J.; and Karmalkar, S. 2025. A Fourier Space Perspective on Diffusion Models. *arXiv preprint arXiv:2505.11278*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144.

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

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

Karras, T.; Aittala, M.; Aila, T.; and Laine, S. 2022. Elucidating the design space of diffusion-based generative models. In *NeurIPS*, 26565–26577.

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

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

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

Kynkäänniemi, T.; Karras, T.; Aittala, M.; Aila, T.; and Lehtinen, J. 2023. The Role of ImageNet Classes in Fréchet Inception Distance. In *ICLR*.

Li, B.; Hu, Y.; Liu, S.; and Wang, X. 2025a. Control and Realism: Best of Both Worlds in Layout-to-Image without Training. In *ICML*.

Li, B.; Zhang, Z.; Nie, X.; Han, C.; Hu, Y.; Qiu, X.; and Guo, T. 2025b. Styo: Stylize your face in only one-shot. In *AAAI*, 4625–4633.

Li, B.; Zhang, Z.; Yang, X.; and Wang, X. 2025c. CoSER: Towards Consistent Dense Multiview Text-to-Image Generator for 3D Creation. In *CVPR*, 2880–2890.

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

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

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

Liu, J.; Huang, T.; and Xu, C. 2024. Training-free Composite Scene Generation for Layout-to-Image Synthesis. In *ECCV*, 37–53.

Liu, S.; Ma, A.; Wu, X.; Leng, D.; Yin, Y.; et al. 2024. HiCo: Hierarchical Controllable Diffusion Model for Layout-to-image Generation. In *NeurIPS*.

Mo, S.; Mu, F.; Lin, K. H.; Liu, Y.; Guan, B.; Li, Y.; and Zhou, B. 2024. Freecontrol: Training-free spatial control of any text-to-image diffusion model with any condition. In *CVPR*, 7465–7475.

Nichol, A. Q.; and Dhariwal, P. 2021. Improved denoising diffusion probabilistic models. In *ICML*, 8162–8171.

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

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

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2024. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. In *ICLR*.

Qiu, X.; Han, C.; Zhang, Z.; Li, B.; Guo, T.; and Nie, X. 2023. Diffbfr: Bootstrapping diffusion model for blind face restoration. In *ACM MM*, 7785–7795.

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

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

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

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

Wang, C.-Y.; Bochkovskiy, A.; and Liao, H.-Y. M. 2023. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *CVPR*, 7464–7475.

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

Xiao, J.; Lv, H.; Li, L.; Wang, S.; and Huang, Q. 2024. R&B: Region and Boundary Aware Zero-shot Grounded Text-to-image Generation. In *ICLR2024*.

Xiao, Z.; Fu, X.; Huang, J.; Cheng, Z.; and Xiong, Z. 2021. Space-time distillation for video super-resolution. In *CVPR*, 2113–2122.

Xiao, Z.; Li, Z.; and Jia, W. 2025. Occlusion-Embedded Hybrid Transformer for Light Field Super-Resolution. In *AAAI*, 8700–8708.

Xiao, Z.; Liu, Y.; Gao, R.; and Xiong, Z. 2023. Cutmib: Boosting light field super-resolution via multi-view image blending. In *CVPR*, 1672–1682.

Xiao, Z.; and Xiong, Z. 2025. Incorporating degradation estimation in light field spatial super-resolution. *Computer Vision and Image Understanding*, 252: 104295.

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

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

Xu, X.; Wang, Z.; Zhang, G.; Wang, K.; and Shi, H. 2023b. Versatile diffusion: Text, images and variations all in one diffusion model. In *ICCV*, 7754–7765.

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

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

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

Zhou, D.; Li, Y.; Ma, F.; Yang, Z.; and Yang, Y. 2024a. Migc++: Advanced multi-instance generation controller for image synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

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