Local Conditional Controlling for Text-to-Image Diffusion Models

Yibo Zhao[1,2], Liang Peng[2], Yang Yang[1,2], Zekai Luo[1,2], Hengjia Li[1,2], Yao Chen[2,3]
Zheng Yang[2], Xiaofei He[1,2], Wei Zhao[4], Qinglin Lu[5], Wei Liu[5], Boxi Wu[3*]
[1]State Key Lab of CAD&CG, Zhejiang University
[2]Fabu Inc
[3]The School of Software Technology, Zhejiang University
[4]Xidian University
[5]Tencent Inc

Abstract

Diffusion models have exhibited impressive prowess in the text-to-image task. Recent methods add image-level structure controls, *e.g.*, edge and depth maps, to manipulate the generation process together with text prompts to obtain desired images. This controlling process is globally operated on the entire image, which limits the flexibility of control regions. In this paper, we explore a novel and practical task setting: **local control**. It focuses on controlling specific local region according to user-defined image conditions, while the remaining regions are only conditioned by the original text prompt. However, it is non-trivial to achieve it. The naive manner of directly adding local conditions may lead to the local control dominance problem, which forces the model to focus on the controlled region and neglect object generation in other regions. To mitigate this problem, we propose Regional Discriminate Loss to update the noised latents, aiming at enhanced object generation in non-control regions. Furthermore, the proposed Focused Token Response suppresses weaker attention scores which lack the strongest response to enhance object distinction and reduce duplication. Lastly, we adopt Feature Mask Constraint to reduce quality degradation in images caused by information differences across the local control region. All proposed strategies are operated at the inference stage. Extensive experiments demonstrate that our method can synthesize high-quality images aligned with the text prompt under local control conditions.

1 Introduction

Large-scale text-to-image diffusion models (Saharia et al. 2022; Rombach et al. 2022; Ramesh et al. 2021) have showed remarkable capabilities in generating images conditioned on specified text prompts. However, text descriptions frequently fall short in sufficiently conveying detailed control (Wang et al. 2022) over the generation images. Recent methods (Zhang, Rao, and Agrawala 2023; Mou et al. 2023) have introduced the incorporation of image-level spatial controls, such as edges, depth, and segmentation, into the text-to-image generation process. These methods operate across the entire image, ensuring that the final synthesized image retains the complete global structural information of the original. Such advancements have attracted considerable attention from both the academic and industrial sectors, as it facilitates a more controllable image synthesis process.

Unfortunately, the conventional approach of enforcing global control at the entire-image level lacks the fine-grained image manipulation capability that users often require. Considering the example in Figure 1, we might want to exclusively apply the cat canny condition, while allowing the generation in other regions fully leverages the capabilities of the text-to-image model. This naturally leads to a question: Can we confine our control to specific local region using image conditions? In this paper, we specifically address this practical scenario and introduce the concept of local control as a novel and pragmatic paradigm. It focuses on controlling the generation of specific regions of an image according to user-defined local control conditions, while the rest of the image is conditioned by the text prompt. As illustrated in Figure 1, only the canny condition of the cat is employed for controlling, while the remainder of the image is generated solely based on the provided text prompt. This novel manner empowers users with the ability to exert flexible control over the image generation process.

An intuitive resolution for local control lies in directly adding local image conditions. However, this naive method leads to a catastrophic problem. The prompt concept that is most related to the local control condition dominates the generation process, while other prompt concepts are ignored. Consequently, the generated image cannot align with the input prompt. We call this problem **local control dominance**. We show the cases in the Figure 2. Delving into the underlying mechanism, we find that the attention map for the dog shows high response within the local control regions, which impedes the generation of multiple objects, as seen here with the dog's generation being hindered. It indicates that local control conditions force the model to focus on the controlled region and overlook other regions.

To address this issue, we introduce an object regeneration strategy. Specifically, we propose the Regional Discriminate loss to update the latents. It identifies the ignored objects in the cross attention map, enhancing them on the latent by adaptively operating attention scores under local and non-local regions. This manner guides the model to generate the most related objects to the image condition in the local region, meanwhile producing other desired objects in non-

*Corresponding author.

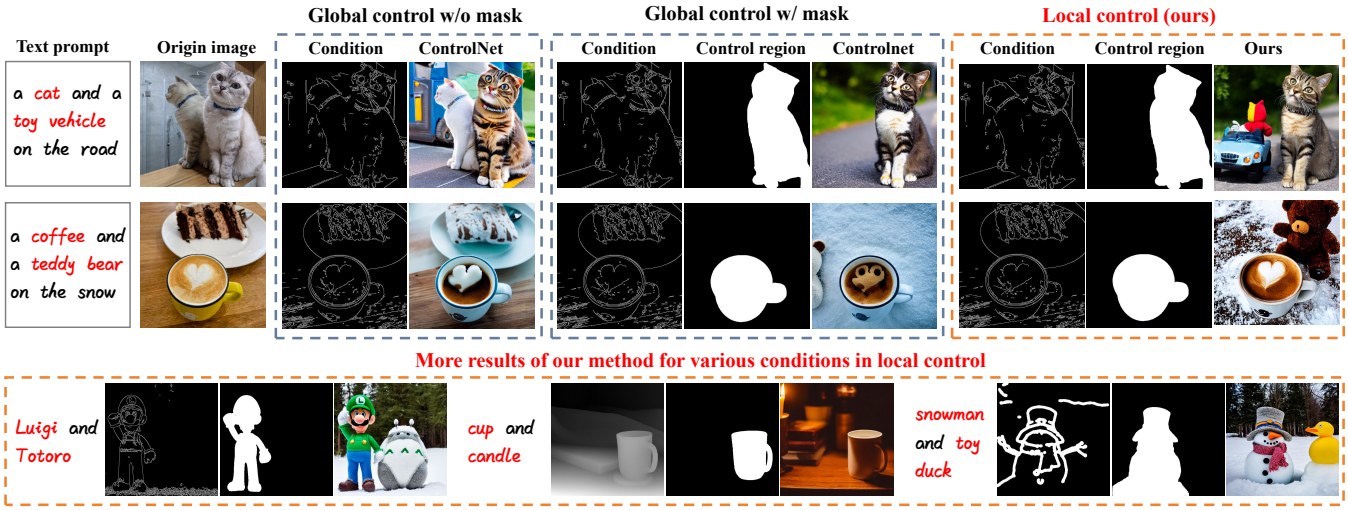


Figure 1: Previous global control mechanism mainly synthesizes images similar to structure conditions, but has difficulty generating results aligned with text prompts. Even adding control mask could only produce concepts closest to the local condition. Therefore, we explore local control, which leverages text prompts, image conditions, and user-defined regions for local control as inputs. Our proposed method successfully generates images that are faithful to both the prompts and local control conditions.

control regions for the text prompt. Additionally, we suppress attention scores for patches on the cross attention map which lack the strongest response in the token dimension. This technology alleviates token overlapping contribution to enhance object distinction and reduce duplication. Moreover, we use local region mask to constrain the feature outputs of ControlNet. This technology reduces the degradation of synthesized image quality caused by differences in control information between non-local and local control regions. All proposed strategies are operated at the inference stage. Overall, the proposed training-free approach achieves high-quality local control for both the text prompt and image conditions, without the need for any training or additional text-image paired datasets. Extensive experiments demonstrate the effectiveness of our method.

Our contributions can be summarized as follows:

- We introduce a new task setting: local control, allowing for fine-grained controlling with image conditions. It aims to control specific regions within the image as desired by the user, while the remaining regions are conditioned solely by the original text prompt.
- We pose the main challenge in local control, and propose a training-free method that includes three techniques to resolve it. The proposed method ensures the preservation of local control while maintaining the generative capability of the diffusion models.
- Extensive experiments demonstrate that our method can synthesize high-quality images, precisely aligning both local image conditions and text prompts.

2 Related Work

Text-to-Image Models. Early methods (Xu et al. 2018; Zhang et al. 2017) for text-to-image synthesis are limited in generating low-resolution images in specific domains. In recent years, extensive data volume LAION(Schuhmann et al. 2022) and abundant model capacity have led to significant progress in the field. The diffusion model (Dhariwal and Nichol 2021; Ho, Jain, and Abbeel 2020) has achieved amazing results in text-to-image generation tasks and has gradually become a representative method in this field. Such as Glide(Nichol et al. 2021), Imagen (Saharia et al. 2022), Stable Diffusion (Rombach et al. 2022), and DALL-E (Ramesh et al. 2021). They leverage the denoising process, gradual reduction of noise to high-quality images through iterative optimization. In addition, the cross-attention mechanism incorporates content from text prompts into the generated images through CLIP(Radford et al. 2021a) or large language models(Raffel et al. 2020).

Compositional Generation. Compositional Generation involves constructing or combining images semantically based on a set of given elements or attributes. (Yang et al. 2023; Avrahami, Lischinski, and Fried 2022; Lu, Liu, and Kong 2023) utilize diffusion models to tackle the problems of fusion inconsistency, enhancing the efficiency of image composition. Recent works focus on mutil-concept generation, Prompt2Prompt (Hertz et al. 2022) and Plug-and-Play (Tumanyan et al. 2023) explore the relationship between attention map and image representation. StructureDiffusion (Feng et al. 2022), Attend-and-Excite (Chefer et al. 2023) and Multi-Concept T2I-Zero (Tunanyan et al. 2023) refine the cross-attention map to strengthen all of the subjects to encourage the model to generate all concepts described in the text prompts. (Li et al. 2023a, 2024a,b) aims to tackle the problems of fusion inconsistency, enhancing the efficiency of image generation. Gligen (Li et al. 2023c) achieves grounded text2img generation with text prompt and bounding box condition inputs. Boxdiff (Xie et al. 2023) propose a training-free method to control objects and contexts. Unlike these methods, our approach aims to preserve the structural

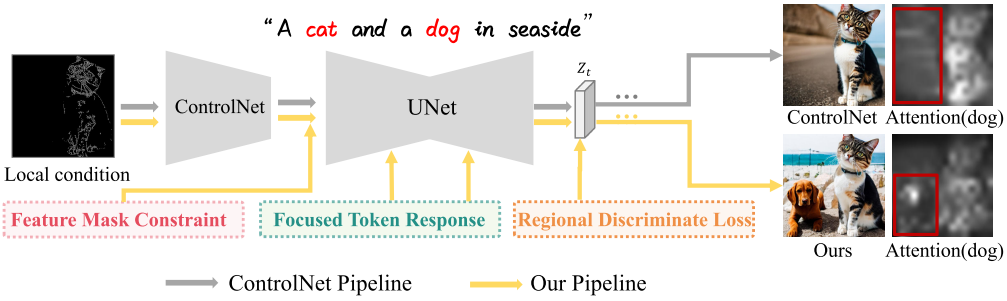


Figure 2: In the ControlNet pipeline, local conditions dominate the image generation. The attention map for "dog" exhibits a higher response in local areas while a lower response in non-local regions where the dog should be generated. Meanwhile, we showcase the result of our method applied to local control , where both "cat" and "dog" are successfully generated.

information of local regions during image generation, which facilitates finer-grained generation.

Controllable Generation. To facilitate controllable content generation, some works aim to provide reliable structural guidance. PITI (Wang et al. 2022) reduces the feature gap for structural guidance, while (Voynov, Aberman, and Cohen-Or 2023) uses similarity gradients between the target sketch and intermediate features. ControlNet (Zhang, Rao, and Agrawala 2023) learns condition-specific network to enable guide generation for the pre-trained diffusion model. T2i-Adapter(Mou et al. 2023) introduces a lightweight adapter to combine the internal knowledge in T2I models with control signals. Uni-ControlNet(Zhao et al. 2024) proposes a condition injection strategy to learn an adapter for various controls, focusing on combining multiple structural conditions. Recently, training-free approaches (Cao et al. 2023; Patashnik et al. 2023; Epstein et al. 2023; Mo et al. 2024) are proposed for controllable generation. Although these existing structure-controllable image generation methods have achieved good control at the image-level, they lack the ability to provide localized control for generating images that align with both prompts and control conditions.

3 Method

3.1 Preliminaries

Stable Diffusion. We apply our method on Stable Diffusion (Rombach et al. 2022), which efficiently operates on the latent space. Specifically, given an input image I, the encoder maps it to a latent z_0. Diffusion algorithms progressively add noise ϵ to the z_0 and produce a noisy latent z_t, where t represents the number of times noise is added. To achieve conditional generation, Stable Diffusion utilizes cross-attention between text embeddings $\mathbf{c}$ converted by CLIP (Radford et al. 2021b) and the UNet's intermediate features $\varphi(\mathbf{x}_t)$. The cross-attention map $\mathbf{A}$ can be defined as:

$$\mathbf{A} = \text{Softmax}(\mathbf{Q}\mathbf{K}^\top/\sqrt{d}), \tag{1}$$

where $\mathbf{Q} = \mathbf{W}_Q\varphi(\mathbf{x}_t)$ and $\mathbf{K} = \mathbf{W}_K\mathbf{c}$ are the query and key matrices, respectively. Finally, ϵ_θ can be trained by a MSE loss, which predicts the noise added to the latent $\boldsymbol{z}_t$:

$$\mathcal{L} = \mathbb{E}_{\boldsymbol{z}_0,\boldsymbol{t},\boldsymbol{c},\epsilon\sim\mathcal{N}(0,1)}\left[\|\epsilon - \epsilon_\theta(\boldsymbol{z}_t,\boldsymbol{t},\boldsymbol{c})\|_2^2\right] \tag{2}$$

ControlNet. Relying solely on text prompts often fails to generate images that align with the user's desires. ControlNet(Zhang, Rao, and Agrawala 2023) integrate global-level spatial controls which utilizes the encoding layers of the diffusion model. During training, an additional spatial control condition $\boldsymbol{c}_\text{f}$ is required, and the training loss is:

$$\mathcal{L}_c = \mathbb{E}_{\boldsymbol{z}_0,\boldsymbol{t},\boldsymbol{c},\boldsymbol{c}_\text{f},\epsilon\sim\mathcal{N}(0,1)}\left[\|\epsilon - \epsilon_\theta(\boldsymbol{z}_t,\boldsymbol{t},\boldsymbol{c},\boldsymbol{c}_\text{f})\|_2^2\right]. \tag{3}$$

In the context of local control, given a text prompt, the tokenizer produces a set of tokens $\mathcal{P} = \{\mathbf{p}_i\}$, and a set of corresponding spatial cross-attention maps $\mathcal{A}^t = \{\mathbf{A}_i^t\}$ in UNet can be accordingly obtained. The most straightforward method of enabling local control is to directly input $\boldsymbol{c}_\text{f}$ and $\mathcal{P}$ into the existing control methods. However, this approach leads to the issue of local control dominance (Figure 2).

3.2 Control Concept Matching

Our initial objective in local control is to identify the most suitable object for generation within the control region at timestep t. The resulting object token indices are denoted as $\text{C}^t_\text{control}$. In our method, the sum of attention scores within the local control region is employed as the criterion. At denoising steps $t > \beta T$, we identify the object with the highest sum attention score within the local control region as the $\text{C}^t_\text{control}$. β is a hyperparameter that acts on the total timesteps T. Given the impact of the denoising process's initial steps on the spatial structure of the synthesized image, we focus on stabilizing the selection of $\text{C}^t_\text{control}$. Thus, we employ the $Count_{max}(\cdot)$ function to determine the most frequently occurring $\text{C}^t_\text{control}$ in the early steps of the denoising process.

$$\text{C}^t_\text{control} = \begin{cases} \arg\max_i\Big(\sum(\mathbf{M}\cdot\mathbf{A}_i^t)\Big), & t > \beta T, \\ Count_{max}(\{\text{C}^{t^*}_\text{control} \mid t^* > \beta T\}), & t \le \beta T \end{cases} \tag{4}$$

where $\mathbf{M}$ denotes a binary spatial mask, which is generated from the local control region. It aims to selectively local control elements in the cross-attention maps. This approach ensures that the chosen object is optimally aligned with the user-defined local control conditions.

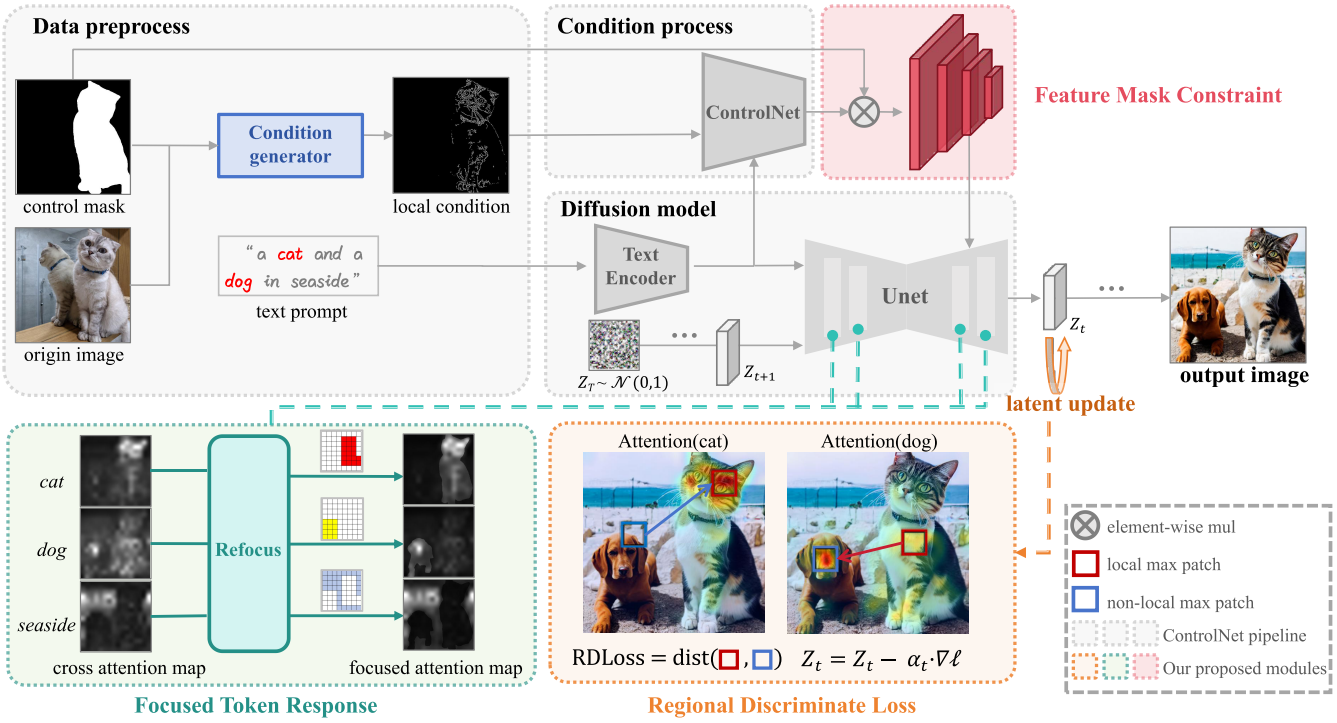


Figure 3: Overview of our method. Given the text condition and local control condition, a latent variable z_T is passed into the denoising network, *i.e.,* UNet. At each denoising step, we apply **Feature Mask Constraint** (see Section 3.5) to the ControlNet branch output features. Cross-attention maps generated from UNet are refined by **Focused Token Response** (see Section 3.4), which suppresses weaker interfering responses to enhance object distinction. Furthermore, we use **Regional Discriminate Loss** (see Section 3.3) to update the latent z_t, thereby identifying and regenerating ignored objects in the cross-attention map.

3.3 Regional Discriminate Loss

Due to the impact of local control dominance, most objects often respond to the local control regions in cross-attention maps. An effective strategy to counter this effect involves directing the attention responses of objects in non-control regions away from the local control region. To achieve this, we introduce the object regeneration strategy, which is specifically designed to enhance the difference in response between the inside and outside of the local control region.

Attend-and-excite(Chefer et al. 2023) indicates that the maximum value regions in the cross-attention map are key to defining the positioning of generated objects. Therefore, we choose the maximum value in the attention map to design a loss, named Regional Discriminate Loss (RDLoss). We adopt the max distance between the maximum values inside and outside the local region as the loss, so that each attention map can separate the objects in the local and non-local regions based on whether they are $\mathrm{C}^t_{\text{control}}$ adaptive.

$$\begin{aligned}\text{RDLoss}_{s_i} = \mathbb{I} \cdot (&\max(G(\mathbf{A}_{s_i}) \cdot \mathbf{M}) \\ &- \max(G(\mathbf{A}_{s_i}) \cdot (1 - \mathbf{M}))),\end{aligned} \tag{5}$$

we select a set of object token indices $\mathcal{S} = \{\mathbf{s}_i\}$ from the collection $\mathcal{P}$, where $\mathcal{P}$ represents the entire set of prompt tokens. $G(\cdot)$ represents a Gaussian smoothing operator. And the indicator function $\mathbb{I}$ is defined as follows:

$$\mathbb{I}(s_i = \mathrm{C}^t_{\text{control}}) = \begin{cases} -1, & \text{if } s_i = \mathrm{C}^t_{\text{control}}, \\ 1, & \text{otherwise.} \end{cases} \tag{6}$$

RDLoss uses the difference between the maximum attention scores inside and outside the local region to represent distance. In the attention maps for tokens not requiring local control conditions, we enhance the generation in non-local regions. Conversely, for tokens that require local control, we enhance generation in local regions. After defining the loss, we can update the latent in the following way:

$$z_t = z_t - \alpha_t \nabla l, \tag{7}$$

l is determined by $\max(RDLoss_{s_i})$ in eq 5 and is used to update the noised latent z_t at the current time step t, where α_t is a scalar defining the step size of the gradient update. l updates noised latents by increasing the attention scores that exhibit the maximum difference between patchs inside and outside the local control region. This approach directs the model to generate objects closely related to the control condition in the local region, while also creating objects in the non-control regions as given text prompt.

3.4 Focused Token Response

After updating the noised latent, while the image content closely aligns with the text prompt, non-local objects still

tend to be generated repeatedly, or near the local control region. This challenge is often linked to the overlapping contribution regions of corresponding token embeddings. Inspired by(Tunanyan et al. 2023), we utilize cross attention maps to mitigate weak response regions. Specifically, we reduce the attention scores for patchs in the attention map that do not demonstrate the strongest response in the token dimension. This technology encourages the reduction of responses from weaker tokens, thereby enforce the model to selectively respond to the most relevant tokens.

$$\mathbf{A}^t = O(\arg\max(\mathbf{A}^t), \gamma) \cdot \mathbf{A}^t, \tag{8}$$

$\arg\max$ identifies the indices of the highest responses across token dimension, while O represents a suppression operation that applies a scaling factor γ to the non-max attention scores. This technique minimizes the influence of irrelevant pixels in the cross-attention map, enhances specific token contributions, reduces overlap and duplication. Collaborating with RDLoss, it aids the model in generating across both local and non-local regions.

3.5 Feature Mask Constraint

During denoising, there is a notable difference in information between the inside and outside of the local control region. The local control carries only localized control information, leaving the rest of the regions with blank control information. As a consequence, the latent distribution resulting from the injection of local control conditions into the UNet differs from that obtained under global control conditions, potentially affecting the image quality.

To address this issue, we apply feature mask constraint to the output of ControlNet. This design directs the UNet's focus to the local control region, minimizing the disruptive effect of irrelevant control information in the non-control region. In practice, we extract the local feature maps from the output of ControlNet block. Subsequently, these local condition features are added in the UNet. Assume $\mathcal{F}(\cdot;\theta)$ represents a neural block within the UNet architecture, and x_t is the input feature map of the block during the t-th denoising step. Separately, $\mathcal{G}(\cdot;\cdot)$ represents a neural block in ControlNet, taking the spatial conditional input feature c_f and x_t. The fused feature output from each block is denoted as y_t:

$$y_t = \mathcal{F}(x_t;\theta) + \mathbf{M} \cdot \mathcal{G}(x_t; c_f). \tag{9}$$

After applying the feature mask, the ControlNet output is selectively integrated into the UNet. The local control condition is achieved by multiplying the feature mask $\mathbf{M}$ with the conditioned output from the ControlNet block $\mathcal{G}(x_t; c_f)$, then added to the original feature map $\mathcal{F}(x_t;\theta)$ from the UNet. This method improves image quality by preventing the UNet from focusing on irrelevant control information.

4 Experiments

4.1 Dataset and Evaluation

Dataset: We utilized the COCO(Lin et al. 2014) validation set with 80 object categories, selecting one random caption per image to create a dataset of 5k generated images. We also created the **Attend-Condition** dataset, inspired by (Chefer et al. 2023), featuring 11 object and 11 animal categories with text prompts like 'a [object] and a [animal]'. Images are collected online, with those having larger subject areas randomly selected for local control.

Figure 4: Visual results of our method under the local control setting. The proposed method can be employed for flexibly controlling the local regions with image conditions.

Evaluation Metrics: We assess image quality using the Fréchet Inception Distance (FID)(Seitzer 2020), and measure text-image similarity with the average CLIP(Radford et al. 2021c) score between text prompts and corresponding images. Using BLIP-2(Li et al. 2023b)'s descriptive capabilities, we generate captions for images produced under the local control paradigm. Subsequently, we calculate the CLIP text-text similarity between generated captions and original captions. To evaluate the alignment with the provided spatial condition, we caculate the LPIPS(Zhang et al. 2018) metrics between original and generated images in localized regions. To evaluate whether local object generation matches specified regions, we employ OneFormer (Jain et al. 2023) for local segmentation detection. Additionally, we use SAM(Kirillov et al. 2023) and DINO(Liu et al. 2023) to segment the Attend-Condition experimental results.

Baseline Settings We conducted a comparative analysis of our method against several baselines. We select two commonly used structure-controllable models, ControlNet(Zhang, Rao, and Agrawala 2023) and T2I-Adapter(Mou et al. 2023), as naive baselines by inputting the local control condition and text prompts into these models. Additionally, we design three competitive baselines specifically for the local control task, which do not require extra training and use global conditions as input. Specifically, 1)Noise-Mask: Inspired by (Nichol et al. 2021), we estimate the noise with adding the condition into ControlNet and the noise from vanilla T2I model, then combine them according to the mask. The combined noise is calculated as:

$$\epsilon_\theta = \epsilon_\theta(\boldsymbol{z}_t, \boldsymbol{t}, \boldsymbol{c}, \boldsymbol{c}_\mathrm{f}) \cdot \mathbf{M} + \epsilon_\theta(\boldsymbol{z}_t, \boldsymbol{t}, \boldsymbol{c}) \cdot (1 - \mathbf{M}). \tag{10}$$

2)Feature-Mask: We implement the addition of our proposed "Feature Mask Constraint" component into ControlNet across all denoising steps, which applies local control mask to ControlNet features. 3)Inpainting: We use ControlNet to generate an image with the spatial condition and text prompt, then use inpainting method(Lugmayr et al. 2022) to regenerate the parts without spatial condition.

Figure 5: Comparisons with several baselines under varying conditions. Our method, along with T2I-Adapter and ControlNet, uses local conditions as input. Conversely, Noise-Mask, Feature-Mask, and Inpainting use global conditions as inputs. Compared with these baselines, our method can synthesize high-quality images in local control, especially in terms of text alignment.

Dataset	metrics	ControlNet	T2I	Noise-Mask	Feature-Mask	Inpainting	Ours
COCO2017	FID ↓	27.87 ± 4.22	30.32 ± 3.75	22.72 ± 0.61	21.93 ± 0.64	25.72 ± 1.22	**21.86 ± 0.48**
	CLIP Score ↑	0.293 ± 0.010	0.295 ± 0.03	0.287 ± 0.003	0.290 ± 0.002	0.288 ± 0.003	**0.305 ± 0.02**
	CLIP T2T ↑	0.782 ± 0.019	0.780 ± 0.009	0.789 ± 0.006	0.779 ± 0.014	0.785 ± 0.006	**0.801 ± 0.006**
	IOU ↑	**0.55 ± 0.07**	0.44 ± 0.06	0.41 ± 0.02	0.37 ± 0.03	0.53 ± 0.07	0.51 ± 0.10
	$\text{LPIPS}_{\times 10^2}$ ↓	13.15 ± 1.19	14.00 ± 1.06	14.16 ± 1.12	15.01 ± 1.33	13.25 ± 1.51	**13.04 ± 1.36**
Attend-Condition	FID ↓	-	-	-	-	-	-
	CLIP Score ↑	0.261 ± 0.010	0.258 ± 0.004	0.279 ± 0.003	0.278 ± 0.003	0.273 ± 0.03	**0.285 ± 0.02**
	CLIP T2T ↑	0.732 ± 0.018	0.700 ± 0.015	0.782 ± 0.006	0.784 ± 0.019	0.753 ± 0.009	**0.804 ± 0.002**
	IOU ↑	**0.74 ± 0.04**	0.70 ± 0.02	0.66 ± 0.02	0.64 ± 0.04	0.72 ± 0.04	0.70 ± 0.05
	$\text{LPIPS}_{\times 10^2}$ ↓	12.67 ± 1.03	13.31 ± 1.15	13.25 ± 1.21	13.30 ± 1.27	12.97 ± 1.06	**12.57 ± 1.08**

Table 1: Quantitative comparisons to baselines. We do not provide FID on the Attend-Condition due to the small scale. Our method can maintain good alignment with the spatial condition while achieving the best image quality and text alignment.

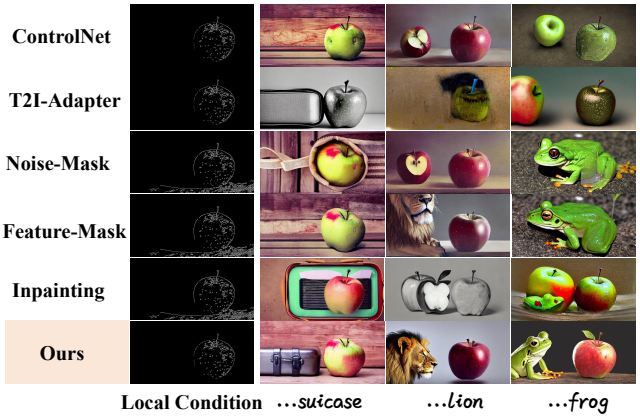


Figure 6: Comparisons with baselines on Attend-Condition.

4.2 Comparison with Baselines

Quantitative experiments. We evaluate the baselines and our method under various conditions. The result includes the following metrics: FID, CLIP-score, CLIP text-text similarity (CLIP T2T), IOU, and LPIPS, which are detailed in Table 1. Due to the small number of original images in the Attend-Condition, we do not calculate its FID, as such a small sample size leads to unreliable FID. In some cases, the small local control region leads to blank pixels, causing naive methods to generate artifacts and grayscale images, resulting in high FID scores. In evaluations on Attend-Condition dataset specifically designed for multi-object generation, our method significantly outperforms other baselines in both CLIP Score and CLIP T2T. Additionally, we evaluate the alignment ability of baselines and our method. Noise-Mask and Feature-Mask methods align poorly with the given local spatial conditions. ControlNet achieves the best spatial alignment performance. The inpainting method ensures spatial alignment capabilities similar to that of ControlNet. Moreover, our method also achieves similar spatial alignment capabilities as ControlNet in local control task. Compared to these baselines, our method maintains good alignment with the given spatial conditions, while achieving the best image quality and text alignment.

Qualitative Results. Figure 5 illustrates these differences, particularly in scenarios involving multiple objects in

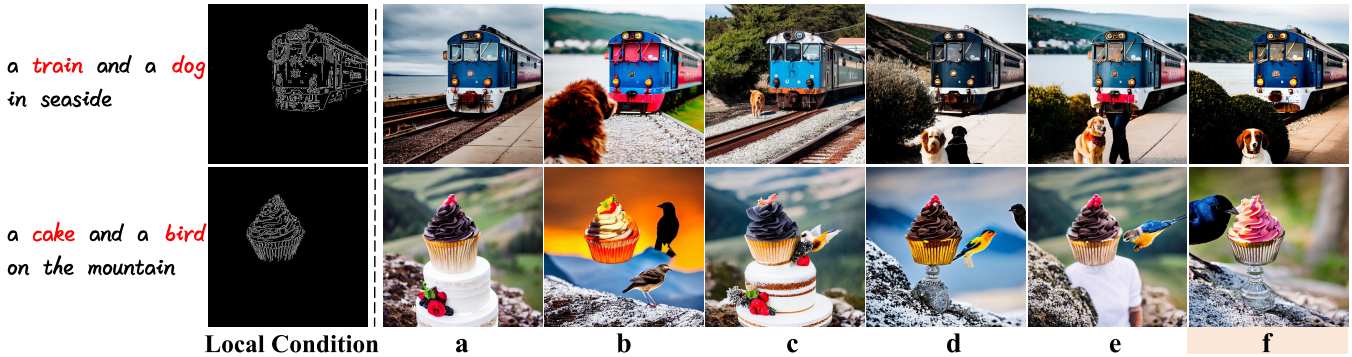


Figure 7: Ablation Studies: a) We use ControlNet as the vanilla baseline. b) RDLoss only. c) FMC only, referred to as the feature-mask baseline. d) Our method without the FTR module. e) Our method without the FMC module. f) Complete method.

prompts. In naive method, it can almost only generate objects in text prompt that best match the control region, while neglecting the generation of other objects. The other three baselines designed specifically for local control achieved good text-image alignment and image quality, but have significant drawbacks compared to our visualized results. Noise-Mask frequently generates untrustworthy pixels near the local region, and sometimes fails to generate mutilple objects. The results generated by the Feature-Mask approach align poorly with the given spatial conditions. The inpainting method sometimes fails to distinguish between the generated content within controlled and uncontrolled regions, leading to issues of repetition and omission in the generated images. Compared to these baselines, our method generates images that most closely match the text prompts.

4.3 Ablation Studies

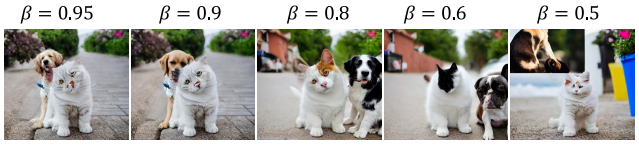
Main ablations. We utilize the canny condition on the COCO dataset to evaluate the effectiveness of our method's key components. Specifically, we focus on three components: 1) Regional Discriminate Loss(RDLoss), updates the noised latent to assist object generation in non-local regions; 2) Focused Token Response(FTR), designed to refine attention response; and 3) Feature Mask Constraint(FMC), a mechanism to constraint features of the ControlNet output. Only the FMC is included, referred to as the feature-mask baseline. Additionally, since the FTR depends on RDLoss, there is no ablation study for it alone. The images are generated using the same seed. The results are listed in Table 2. It reveals that RDLoss reduces the influence of local control dominance in non-local regions. FTR and FMC can assist RDLoss in improving the quality of generated images. It can be foreseen that the updates to the latent have somewhat diminished the capability for image alignment, yet the extent of this impact remains within an acceptable threshold. Figure 7 visualizes the results of the ablation experiments.

Impact of Control Concept Matching. Figure 8(a) shows the need to select concepts matching the local control region. The local content aligns with specified objects, but this naive method forces undesired objects into the region, hindering the model's alignment with the control condition. The right image shows the superiority of Control Concept Matching.

RDLoss	FTR	FMC	FID ↓	CLIP Score↑	CLIP T2T↑	IOU↑	LPIPS$_{\times 10^2}$↓
-	-	-	23.65	0.283	0.750	**0.62**	11.96
√	-	-	24.48	0.299	0.786	0.59	11.99
-	-	√	21.83	0.292	0.791	0.40	13.68
√	-	√	21.91	0.301	0.802	0.57	11.71
√	√	-	22.90	0.298	**0.806**	0.59	11.79
√	√	√	**21.38**	**0.303**	0.795	0.57	**11.68**

Table 2: Ablation studies of the proposed components under canny condition on the COCO dataset. Since the FTR depends on RDLoss, there is no ablation study for it alone.

(a) Control Concept Matching and manually specified



(b) The impact of different β on image quality

Figure 8: (a) Control Concept Matching adaptively matches the concept, (b) Different β values affect image quality.

Figure 8(b) shows the impact of different β values on image quality. Our focus is on stabilizing the selection of C^t_{control}. A β between 0.8 to 0.9 yields good results.

5 Conclusions

In this paper, we introduce local control as a new paradigm for image controllable generation, focusing on user-defined control of specific image regions. To mitigate local control dominance, we propose a training-free method that integrates smoothly with existing control models. This approach optimizes noised latents, enhancing object generation in non-local regions. We demonstrate a variety of results under the local control condition, successfully generating high-quality images faithful to the text prompts.

Acknowledgments

This research was supported by The National Nature Science Foundation of China (Grant Nos: 62402417, 62273302, 62036009, 61936006), in part by the Key R&D Program of Ningbo (Grant No. 2024Z115), in part by Yongjiang Talent Introduction Programme (Grant No: 2023A-197-G), in part by CCF-Tencent Open Fund (Grant No. JR2023TEG012).

References

Avrahami, O.; Lischinski, D.; and Fried, O. 2022. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18208–18218.

Cao, M.; Wang, X.; Qi, Z.; Shan, Y.; Qie, X.; and Zheng, Y. 2023. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22560–22570.

Chefer, H.; Alaluf, Y.; Vinker, Y.; Wolf, L.; and Cohen-Or, D. 2023. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4): 1–10.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34: 8780–8794.

Epstein, D.; Jabri, A.; Poole, B.; Efros, A.; and Holynski, A. 2023. Diffusion self-guidance for controllable image generation. *Advances in Neural Information Processing Systems*, 36: 16222–16239.

Feng, W.; He, X.; Fu, T.-J.; Jampani, V.; Akula, A.; Narayana, P.; Basu, S.; Wang, X. E.; and Wang, W. Y. 2022. Training-free structured diffusion guidance for compositional text-to-image synthesis. *arXiv preprint arXiv:2212.05032*.

Hertz, A.; Mokady, R.; Tenenbaum, J.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2022. Prompt-to-Prompt Image Editing with Cross Attention Control. *arXiv preprint arXiv:2208.01626*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Jain, J.; Li, J.; Chiu, M. T.; Hassani, A.; Orlov, N.; and Shi, H. 2023. Oneformer: One transformer to rule universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2989–2998.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4015–4026.

Li, H.; Liu, Y.; Lin, Y.; Zhang, Z.; Zhao, Y.; Zheng, T.; Yang, Z.; Jiang, Y.; Wu, B.; Cai, D.; et al. 2024a. UniHDA: Towards Universal Hybrid Domain Adaptation of Image Generators. *arXiv preprint arXiv:2401.12596*.

Li, H.; Liu, Y.; Xia, L.; Lin, Y.; Wang, W.; Zheng, T.; Yang, Z.; Zhong, X.; Ren, X.; and He, X. 2023a. Few-shot Hybrid Domain Adaptation of Image Generator. In *The Twelfth International Conference on Learning Representations*.

Li, H.; Liu, Y.; Zhao, Y.; Cheng, H.; Yang, Y.; Xia, L.; Luo, Z.; Qiu, Q.; Wu, B.; Zheng, T.; et al. 2024b. GCA-3D: Towards Generalized and Consistent Domain Adaptation of 3D Generators. *arXiv preprint arXiv:2412.15491*.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Li, Y.; Liu, H.; Wu, Q.; Mu, F.; Yang, J.; Gao, J.; Li, C.; and Lee, Y. J. 2023c. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22511–22521.

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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Li, C.; Yang, J.; Su, H.; Zhu, J.; et al. 2023. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*.

Lu, S.; Liu, Y.; and Kong, A. W.-K. 2023. TF-ICON: Diffusion-Based Training-Free Cross-Domain Image Composition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2294–2305.

Lugmayr, A.; Danelljan, M.; Romero, A.; Yu, F.; Timofte, R.; and Van Gool, L. 2022. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11461–11471.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7465–7475.

Mou, C.; Wang, X.; Xie, L.; Zhang, J.; Qi, Z.; Shan, Y.; and Qie, X. 2023. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*.

Nichol, A.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; and Chen, M. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*.

Patashnik, O.; Garibi, D.; Azuri, I.; Averbuch-Elor, H.; and Cohen-Or, D. 2023. Localizing object-level shape variations with text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 23051–23061.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.;

et al. 2021a. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

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

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

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1): 5485–5551.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, 8821–8831. PMLR.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

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. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35: 36479–36494.

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. *Advances in Neural Information Processing Systems*, 35: 25278–25294.

Seitzer, M. 2020. pytorch-fid: FID Score for PyTorch. https://github.com/mseitzer/pytorch-fid. Version 0.3.0.

Tumanyan, N.; Geyer, M.; Bagon, S.; and Dekel, T. 2023. Plug-and-Play Diffusion Features for Text-Driven Image-to-Image Translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1921–1930.

Tunanyan, H.; Xu, D.; Navasardyan, S.; Wang, Z.; and Shi, H. 2023. Multi-Concept T2I-Zero: Tweaking Only The Text Embeddings and Nothing Else. *arXiv preprint arXiv:2310.07419*.

Voynov, A.; Aberman, K.; and Cohen-Or, D. 2023. Sketch-guided text-to-image diffusion models. In *ACM SIGGRAPH 2023 Conference Proceedings*, 1–11.

Wang, T.; Zhang, T.; Zhang, B.; Ouyang, H.; Chen, D.; Chen, Q.; and Wen, F. 2022. Pretraining is all you need for image-to-image translation. *arXiv preprint arXiv:2205.12952*.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7452–7461.

Xu, T.; Zhang, P.; Huang, Q.; Zhang, H.; Gan, Z.; Huang, X.; and He, X. 2018. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1316–1324.

Yang, B.; Gu, S.; Zhang, B.; Zhang, T.; Chen, X.; Sun, X.; Chen, D.; and Wen, F. 2023. Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18381–18391.

Zhang, H.; Xu, T.; Li, H.; Zhang, S.; Wang, X.; Huang, X.; and Metaxas, D. N. 2017. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 5907–5915.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3836–3847.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zhao, S.; Chen, D.; Chen, Y.-C.; Bao, J.; Hao, S.; Yuan, L.; and Wong, K.-Y. K. 2024. Uni-controlnet: All-in-one control to text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36.