

SpotActor: Training-Free Layout-Controlled Consistent Image Generation

Jiahao Wang^{1,2}, Caixia Yan^{1,*}, Weizhan Zhang^{1,*}, Haonan Lin¹, Mengmeng Wang^{3,4},
Guang Dai⁴, Tieliang Gong¹, Hao Sun⁵, Jingdong Wang⁶

¹School of Computer Science and Technology, MOEKLINNS, Xi'an Jiaotong University

²State Key Laboratory of Communication Content Cognition

³College of Computer Science and Technology, Zhejiang University of Technology

⁴SGIT AI Lab, State Grid Corporation of China

⁵China Telecom Artificial Intelligence Technology Co.Ltd

⁶Baidu Inc

{uguisu,linhaonan}@stu.xjtu.edu.cn, {yancaixia,zhangwzh,gongtl}@xjtu.edu.cn, mengmewang@gmail.com,
gdai@gmail.com, sunh10@chinatelecom.cn, wangjingdong@outlook.com

Abstract

Text-to-image diffusion models significantly enhance the efficiency of artistic creation with high-fidelity image generation. However, in typical application scenarios like comic book production, they can neither place each subject into its expected spot nor maintain the consistent appearance of each subject across images. For these issues, we pioneer a novel task, *Layout-to-Consistent-Image* (L2CI) generation, which produces consistent and compositional images in accordance with the given layout conditions and text prompts. To accomplish this challenging task, we present a new formalization of *dual energy guidance* with optimization in a dual semantic-latent space and thus propose a training-free pipeline, **SpotActor**, which features a layout-conditioned optimizing stage and a consistent sampling stage. In the optimizing stage, we innovate a nuanced layout energy function to mimic the attention activations with a sigmoid-like objective. While in the sampling stage, we design *Regional Interconnection Self-Attention* (RISA) and *Semantic Fusion Cross-Attention* (SFCA) mechanisms that allow mutual interactions across images. To evaluate the performance, we present **ActorBench**, a specified benchmark with hundreds of reasonable prompt-box pairs stemming from object detection datasets. Comprehensive experiments are conducted to demonstrate the effectiveness of our method. The results prove that SpotActor fulfills the expectations of this task and showcases the potential for practical applications with superior layout alignment, subject consistency, prompt conformity and background diversity.

Project page —

<https://johnneywang.github.io/SpotActor-webpage/>

Introduction

Diffusion probabilistic models (Ho, Jain, and Abbeel 2020; Rombach et al. 2022; Sohl-Dickstein et al. 2015a) have achieved notable success in the realm of image generation. Within this domain, text-to-image (T2I) diffusion models

*Corresponding authors.

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

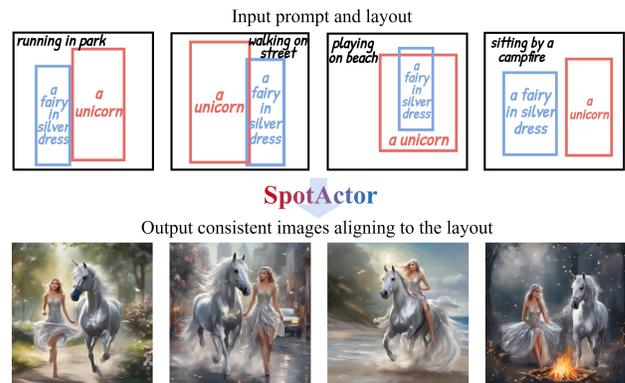


Figure 1: Given bounding boxes and text prompts of subjects, our method generates high-quality images where subjects align to the layout and share a consistent appearance.

(Ho and Salimans 2022; Podell et al. 2023) enable artists to generate high-quality images with descriptions of their desired subjects. Thus, their applicability extends to numerous practical contexts for their substantial contributions to artistic productivity. Despite the success, however, their performance in some application scenarios still exhibits aspects in need of further refinement. For instance, in real-world creation scenarios like comic book drawing, a natural process involves conceptualizing the appearance of a specific character, designing the visual layout of each scene, and then illustrating a series of images of the character. This process reflects two essential skills of professionals: the ability to preserve appearance consistency of the same character, and the capacity to render image content in alignment with the pre-defined layout—both of which are lacking in standard diffusion models.

Since the emergence of diffusion models, the challenges of subject consistency and layout controllability have continuously been two separate topics of ongoing research interest. For subject consistency, TheChosenOne (Avrahami et al. 2024) first introduces the task of *consistent subject generation*, which aims to generate consistent images of

the same subject solely driven by prompts, and differentiate it from other analogous tasks. It proposes a tuning-based approach to accomplish this task yet its iterative backbone tuning process results in high computational expenses and may degrade the image quality. Later, OneActor (Wang et al. 2024) achieves a $4\times$ faster tuning speed without sacrificing the image quality via intricate cluster-conditioned guidance. More recently, training-free methods (Tewel et al. 2024; Zhou et al. 2024b) are proposed to bypass the tuning process by enhancing the backbone with handicraft modules activated during the inference process. For the layout controllability, prevailing methods (Epstein et al. 2023; Mo et al. 2024; Chen, Laina, and Vedaldi 2024) manage to achieve the *layout-to-image generation* task via energy guidance in a training-free manner. To specify, they regulate the latent codes through the backward propagation driven by customized energy functions and thus align the visual elements to their expected positions. Nevertheless, in the context of the aforementioned creation scenario, no existing work embarks on addressing both challenges simultaneously. Besides, existing works in either task pay limited attention to the role of the semantic space in T2I models.

For these issues, we pioneer a novel task, *layout-to-consistent-image* (L2CI) generation. As shown in Fig. 1, given a series of expected bounding boxes, the corresponding subject descriptions (e.g. *fairy&unicorn*) and the plot descriptions, this task aims to generate a series of images where subjects share the consistent appearance as well as occur perfectly in the given boxes. To accomplish this challenging task, we propose **SpotActor**, the first L2CI generation pipeline in a training-free manner. We start from the insight that the semantic space and spatial latent space of diffusion models are inherently entangled and share certain properties (Li et al. 2024; Wang et al. 2024). Hence, we consider the semantic and latent space as a whole dual space and present a new formalization of *dual energy guidance*, which defines an update trajectory in the dual space. The formalization splits the pipeline into two stages: a layout-conditioned optimizing stage and a consistent sampling stage. In the optimizing stage, we design a sigmoid-like objective based on in-depth analysis of the network activations to regulate the attention distributions. The objective later drives a backward update of the latent codes and the semantic embeddings to search for an optimal alignment with the pre-defined boxes. Subsequently in the sampling stage, we enhance the ordinary backbone with *Regional Interconnection Self-Attention* (RISA) and *Semantic Fusion Cross-Attention* (SFCA) mechanisms in order to allow inter-image level spatial-spatial interactions and spatial-semantic interactions according to the layout conditions, respectively. To evaluate the performance in this task, we present **ActorBench**, the first L2CI benchmark including hundreds of prompt-box pairs and a set of evaluation metrics. We creatively utilize real-world object detection datasets to construct the prompt-box pairs that comply with objective principles. Comprehensive experiments verify our motivation and the effectiveness of our method. The balanced layout alignment, subject consistency, prompt conformity and background diversity confirm that SpotActor fulfills the expectations of this task.

To summarize, our main contributions are as follows:

- We pioneer the layout-to-consistent-image generation task that aims to maintain a consistent appearance of subjects as well as align them to the given layout.
- We consider the semantic and latent space as a whole dual space and formalize a novel dual energy guidance to jointly optimize in the semantic-latent space.
- We propose SpotActor pipeline to address the L2CI task, which features a backward update based on the nuanced layout energy and a forward sampling enhanced by two intricate attention mechanisms.
- We present the first L2CI generation benchmark, ActorBench, and conduct comprehensive experiments to evaluate the effectiveness of our method.

Related Work

Consistent Subject Generation. This task is first proposed by TheChosenOne (Avrahami et al. 2024) and focuses on generating images of the same subject based solely on descriptive prompts, which is different from the customization tasks (Song et al. 2023; Kwon et al. 2024; Zhou et al. 2024a). The pioneer work presents a repetitive process of generating, clustering and tuning to regulate the generation distribution into a cohesive cluster. Yet the laborious process takes 20 minutes to function and may harm the inner capacity of the backbone. Later, OneActor (Wang et al. 2024) proposes a cluster guidance paradigm, supersedes the backbone tuning with a projector optimization and reduces the required time to 5 minutes. More recently, training-free methods (Tewel et al. 2024; Zhou et al. 2024b) are introduced to eliminate the tuning process with new self-attention mechanisms that directly function during the inference. Nonetheless, the training-free methods hardly notice the spatial-semantic interaction of the diffusion process, and no attempts have been made to incorporate the layout control. Thus, we accomplish a novel layout-controlled consistent subject generation task by leveraging latent-semantic optimization.

Layout-to-Image Generation. As diffusion models prevail, numerous works manage to harness the diffusion backbone to generate images aligning with the given layout like boxes or blobs. Early tuning-based methods fine-tune the backbone with handicraft modules (Li et al. 2023; Nie et al. 2024) or specific token embeddings (Yang et al. 2023) to inject layout conditions, while training-free methods (Balaji et al. 2022; Rombach et al. 2022; Kim et al. 2023) achieve the goal by manipulating the attention procedure. Recently, a branch of training-free methods (Couairon et al. 2023; Xie et al. 2023; Epstein et al. 2023; Chen, Laina, and Vedaldi 2024) is gaining prominence for its elegant energy guidance. To specify, they design a backward propagation based on the layout loss, which optimizes the latent codes to align with the given layout. However, optimizing solely in the latent space restricts the search range as the other half, the semantic space, is continuously neglected. Hence, we formalize a new dual energy guidance approach to jointly optimize the latent codes and semantic embeddings, unleashing the full potential of the diffusion model.

Preliminaries

Before introducing our method, we first provide a brief review of diffusion models. From the score-based perspective (Song and Ermon 2019; Song et al. 2021), *diffusion models* (Sohl-Dickstein et al. 2015b; Ho, Jain, and Abbeel 2020; Rombach et al. 2022) essentially manage to estimate a score function of the latent distribution of real images, $\nabla_{\mathbf{z}_t} \log p(\mathbf{z}_t)$, where \mathbf{z}_t is contaminated data from a pre-terminated, time-dependent noise addition process. A denoising network θ is trained to estimate the score at each step: $\hat{\epsilon}_t = \epsilon_\theta(\mathbf{z}_t, t, \mathbf{c}) \approx -\sigma_t \nabla_{\mathbf{z}_t} \log p(\mathbf{z}_t)$, where σ_t are pre-defined constants, \mathbf{c} is the semantic embeddings of the given prompt. Then during generation, taking DDPM (Ho, Jain, and Abbeel 2020) as an example, high-quality images are sampled from random noise by iteratively predicting \mathbf{z}_{t-1} from \mathbf{z}_t based on the score in a step-by-step manner:

$$\mathbf{z}_{t-1} = \frac{1}{\sqrt{1-\beta_t}}(\mathbf{z}_t + \beta_t \nabla_{\mathbf{z}_t} \log p(\mathbf{z}_t)) + \sqrt{\beta_t} \epsilon, \quad (1)$$

where β_t is a set of pre-defined constants and $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. To incorporate more flexible control, *energy guidance* (Zhao et al. 2022) suggests that any energy function $e(\mathbf{z}_t, t, \mathbf{c})$ can be leveraged like a score to update the latent codes for different purposes:

$$\mathbf{z}_t \leftarrow \mathbf{z}_t - v \sigma_t \nabla_{\mathbf{z}_t} e(\mathbf{z}_t, t, \mathbf{c}), \quad (2)$$

where v is the energy guidance scale.

Inside the denoising network, θ , commonly implemented as a U-Net (Ronneberger, Fischer, and Brox 2015), self-attention layers project features of latent codes, \mathbf{h} , into queries, key and value through projection matrices \mathbf{W} : $\mathbf{Q}^{sa} = \mathbf{h} \cdot \mathbf{W}_{\mathbf{Q}}^{sa}$, $\mathbf{K}^{sa} = \mathbf{h} \cdot \mathbf{W}_{\mathbf{K}}^{sa}$, $\mathbf{V}^{sa} = \mathbf{h} \cdot \mathbf{W}_{\mathbf{V}}^{sa}$. While cross-attention layers project \mathbf{h} into queries and project semantic embeddings into keys and values: $\mathbf{Q}^{ca} = \mathbf{h} \cdot \mathbf{W}_{\mathbf{Q}}^{ca}$, $\mathbf{K}^{ca} = \mathbf{c} \cdot \mathbf{W}_{\mathbf{K}}^{ca}$, $\mathbf{V}^{ca} = \mathbf{c} \cdot \mathbf{W}_{\mathbf{V}}^{ca}$. The outputs, \mathbf{h}' , are then calculated via standard attention mechanism: $\mathbf{A} = \text{softmax}(\mathbf{Q} \cdot \mathbf{K}^\top / \sqrt{d_k})$, $\mathbf{h}' = \mathbf{A} \cdot \mathbf{V}$, where d_k is the feature dimension of $\mathbf{W}_{\mathbf{Q}}$ and $\mathbf{W}_{\mathbf{K}}$.

Method

Overview

In this task, users input a batch of N prompt embeddings, $\{\mathbf{c}_i\}_{i=1}^N$, where $N \geq 2$, with the token embeddings of the central subject, $\{\mathbf{c}_i^{sub}\}_{i=1}^N$, and bounding boxes, $\{\mathbf{b}_i = (h_i^{min}, w_i^{min}, h_i^{max}, w_i^{max})\}_{i=1}^N$. Our goal is to generate N consistent images of the subject with respect to the layout boxes. For this purpose, we propose a L2CI generation pipeline, SpotActor, as shown in Fig. 2. To elaborate, we first consider the latent and semantic space as a whole and formalize a new *dual energy guidance* approach which consists of two stages at each generation step. The optimizing stage manages to place the subject in the desired location, in which we update the latent codes and semantic embeddings with nuanced layout energy based on in-depth activation analysis. Subsequently, the sampling stage contributes to the consistent appearance of the subject, in which we enhance the U-Net sampling with *Regional Interconnection Self-Attention*

(RISA) and *Semantic Fusion Cross-Attention* (SFCA) mechanisms in Figs. 2(b) and 2(c). Note that though we illustrate with single subject generation for simplicity, our pipeline can be seamlessly extended to multiple subject generation.

Formalization of Dual Energy Guidance

The standard energy guidance consists of a backward update process p_b and a forward sampling process p_f to transform \mathbf{z}_t to \mathbf{z}_{t-1} at each step, which can be denoted as:

$$p(\mathbf{z}_{t-1} | \mathbf{z}_t) = p_f(\mathbf{z}_{t-1} | \mathbf{z}_t^*) \cdot p_b(\mathbf{z}_t^* | \mathbf{z}_t), \quad (3)$$

where \mathbf{z}_t^* is the optimized latent codes. In the optimizing stage, Eq. (2) with a control energy function is used to update the latent code until an optimal \mathbf{z}_t^* that minimizes the control energy is found. Subsequently, the forward sampling in Eq. (1) is performed to finish this step. However, the standard approach determines a sampling trajectory solely in the latent space and neglects the role of semantic condition. As proven in previous works (Li et al. 2024; Wang et al. 2024), the latent space and the semantic space are inherently entangled together and ought to be regarded as a whole. Thus, we propose to reform Eq. (3) into:

$$p(\mathbf{z}_{t-1}, \mathbf{c}_{t-1} | \mathbf{z}_t, \mathbf{c}_t) = p_f(\mathbf{z}_{t-1}, \mathbf{c}_{t-1} | \mathbf{z}_t^*, \mathbf{c}_t^*) \cdot p_b(\mathbf{z}_t^*, \mathbf{c}_t^* | \mathbf{z}_t, \mathbf{c}_t), \quad (4)$$

where p_b and p_f are layout-conditioned backward update and consistent forward sampling, which will be detailed in the following subsections.

Layout-Conditioned Backward Update

Composition Property of Cross-Attention. In the pursuit of designing an effective energy function for layout control without image quality degradation, we initiate our approach by analyzing the composition property of the cross-attention component. We conduct a standard generation process on SDXL (Podell et al. 2023) and collect the cross-attention maps \mathbf{A}^{ca} to explore the interaction between spatial pixels and semantic tokens. Fig. 3(a) presents the intra-token normalized maps (IntraM) from U-Net encoder layers, bottleneck layers, and decoder layers. It demonstrates that the activation of spatial pixels by semantic tokens corresponds to the composition of the final image, which echoes previous works (Hertz et al. 2023; Tumanyan et al. 2023). Beyond this established conclusion, we further observe that the correspondence intensifies with increasing network depth from encoder to decoder layers. Meanwhile, we normalize IntraM to the same scale to obtain the inter-token normalized maps (InterM) in Fig. 3(b) and thus reveal that different semantic tokens do not activate the spatial pixels equally, but exhibit different levels. These varying degrees of semantic-spatial interactions have a nuanced impact on the final image quality. We consider this discovery to be highly significant, yet it has been hardly utilized in existing works.

Nuanced Layout Energy Function. The analysis above inspires us to transform the layout conditions into precise target distributions to regulate the generated subjects. To this end, we proceed with a detailed distribution analysis concentrating on three nouns in Fig. 3(c). We can observe that

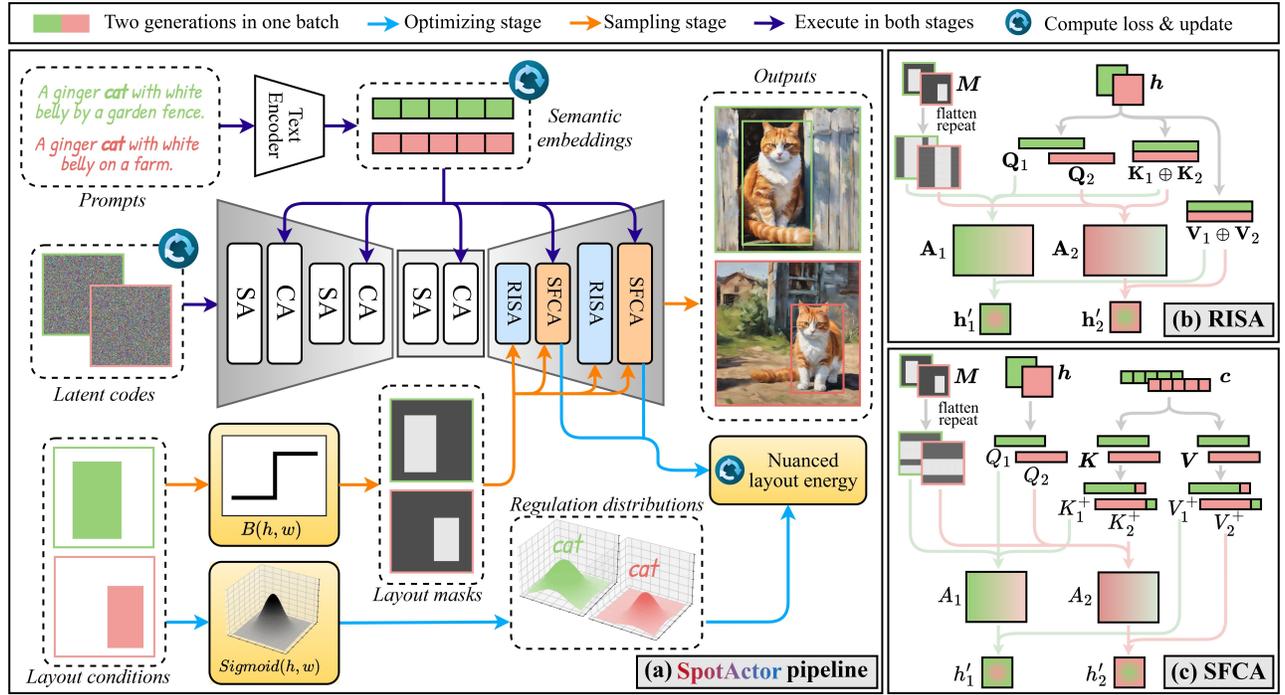


Figure 2: The overall architecture of SpotActor. (a) Our method consists of two stages at each sample step in a dual energy guidance manner. The optimizing stage adjusts the latent codes and semantic embeddings via the nuanced layout energy based on the sigmoid-like objective. Subsequently, the sampling stage is enhanced by two intricate attention mechanisms: (b) RISA and (c) SFCA.

the activation of each noun exhibits a peak shape aligning with its spatial location at a certain range level. For example, the activation of the *dog* token maintains high values in the center part of the *dog* spatial area. When approaching the edge part, it sharply declines to lower values. To mimic the intra-token peak distribution, we extend Sigmoid function to 2-dimensional:

$$\text{Sigmoid}(x, y) = \frac{1}{1 + e^{-s \cdot (1 - \frac{(x-\mu_1)^2}{\sigma_1^2} + \frac{(y-\mu_2)^2}{\sigma_2^2})}}, \quad (5)$$

where (μ_1, μ_2) marks the center; σ_1 and σ_2 establish the margin and s is a shape control factor. Yet for the inter-token distributions, directly regulating the range levels may degrade the image quality. For this issue, we employ spatial normalization to allow the model to allocate activation levels spontaneously. Therefore, the whole backward update can be articulated as below. Given a bounding box $\mathbf{b} = (h^{\min}, w^{\min}, h^{\max}, w^{\max})$ and a subject token embedding \mathbf{c}^{sub} , we first define the target distribution in Eq. (5):

$$\mu_1 = \frac{h^{\min} + h^{\max}}{2}, \mu_2 = \frac{w^{\min} + w^{\max}}{2}, \quad (6)$$

$$\sigma_1 = \frac{(h^{\max} - h^{\min})^2}{4}, \sigma_2 = \frac{(w^{\max} - w^{\min})^2}{4}. \quad (7)$$

We execute the forward sampling of U-Net to collect the cross-attention maps $\mathbf{A}^{ca} \in \mathbb{R}^{K \times S}$ of the subject token from the U-Net decoder, averaged across different layers. K is the

number of attention heads and $S = H \times W$ is the total number of flattened pixels, where H and W are the height and width, respectively. We then reshape it and perform min-max normalization along the spatial dimensions to obtain $\tilde{\mathbf{A}}^{ca} \in \mathbb{R}^{K \times H \times W}$. The energy function can then be calculated by:

$$e = \frac{1}{KHW} \sum_k \sum_h \sum_w \left(\tilde{\mathbf{A}}_{khw}^{ca} - \text{Sigmoid}\left(\frac{h}{H}, \frac{w}{W}\right) \right)^2, \quad (8)$$

where k, h, w are dimension indices. From the dual energy guidance perspective of Eq. (4), we simultaneously update the semantic embedding in the optimizing stage besides Eq. (2):

$$\mathbf{c}_t \leftarrow \mathbf{c}_t - w\sigma_t \nabla_{\mathbf{c}_t} e(\mathbf{z}_t, t, \mathbf{c}), \quad (9)$$

until a batch of optimal $(\mathbf{z}_t^*, \mathbf{c}_t^*)$ are found. Note that even though the semantic update doesn't necessarily need σ_t , we add it as a dynamic step-wise weight.

Consistent Forward Sampling

For the semantic space, we define the forward sampling as $\mathbf{c}_{t-1} = \mathbf{c}_t^*$. While for the latent forward sampling, we enhance the ordinary U-Net with two attention mechanisms to maintain consistent appearance of the subject.

Regional Interconnection Self-Attention. The self-attention mechanism in ordinary U-Net enables the spatial

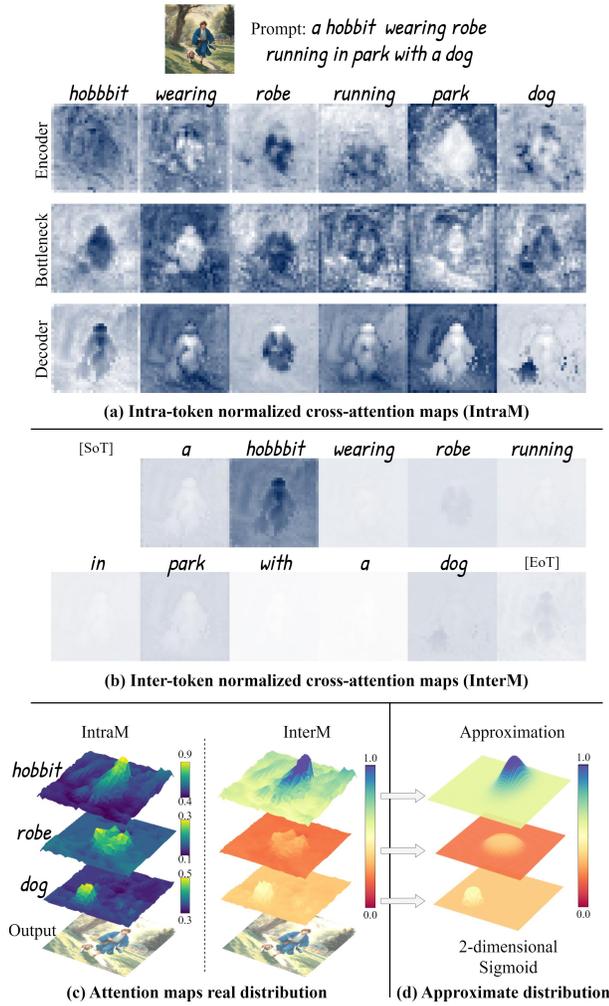


Figure 3: Illustration of the attention analysis. (a) IntraM is the attention map normalized within each token, while (b) InterM is normalized across all the tokens. We further visualize (c) 3D distributions of attention maps and propose (d) sigmoid-like approximate distributions.

pixels from one image to interact with each other, contributing to the final image with a consistent style and content. We desire to broaden the scope of this mechanism to an inter-image level with respect to the layout conditions, which gives rise to RISA. As illustrated in Fig. 2(b), given binary layout masks $\mathbf{M}_i \in \mathbb{R}^{H \times W}$ transformed from \mathbf{b}_i and features of latent codes \mathbf{h}_i , we first flatten the masks and expand to obtain $\mathbf{M}_i^{sa} \in \mathbb{R}^{K \times S \times S}$. Then we concatenate keys and values, respectively and utilize layout masks to precisely control the interconnection region:

$$\mathbf{K}^{sa+} = [\mathbf{K}_1^{sa} \oplus \mathbf{K}_2^{sa} \oplus \dots \oplus \mathbf{K}_N^{sa}], \quad (10)$$

$$\mathbf{V}^{sa+} = [\mathbf{V}_1^{sa} \oplus \mathbf{V}_2^{sa} \oplus \dots \oplus \mathbf{V}_N^{sa}], \quad (11)$$

$$\mathbf{M}_i^{sa+} = [\mathbf{M}_1^{sa} \dots \mathbf{M}_{i-1}^{sa} \oplus \mathbf{I} \oplus \mathbf{M}_{i+1}^{sa} \dots \oplus \mathbf{M}_N^{sa}], \quad (12)$$

$$\mathbf{h}_i^{sa} = \text{Softmax}(\mathbf{Q}_i^{sa} \cdot \mathbf{K}^{sa+T} / \sqrt{d_k} + \log \mathbf{M}_i^{sa+}) \cdot \mathbf{V}^{sa+}, \quad (13)$$

where \mathbf{I} is matrix of ones, \oplus indicates matrix concatenation and the superscript $+$ represents the enlarged matrix.

Semantic Fusion Cross-Attention. The role of semantic space, as we highlight throughout our work, has been continuously overlooked in consistent generation works (Tewel et al. 2024; Zhou et al. 2024b). Hence, we design SFCA to enable each image to interact with all the semantic conditions within the batch. As shown in Fig. 2(c), given \mathbf{M}_i , \mathbf{h}_i and the semantic embedding \mathbf{c}_i , $\mathbf{M}_i^{ca} \in \mathbb{R}^{K \times S \times 1}$ is obtained by flattening, transposing and expanding. We then locate the corresponding \mathbf{K}_i^{sub} and \mathbf{V}_i^{sub} of the subject token and cross-concatenate them for a fused interaction within the layout region:

$$\mathbf{K}_i^{ca+} = [\mathbf{K}_1^{sub} \oplus \dots \oplus \mathbf{K}_{i-1}^{sub} \oplus \mathbf{K}_i^{ca} \oplus \mathbf{K}_{i+1}^{sub} \oplus \dots \oplus \mathbf{K}_N^{sub}], \quad (14)$$

$$\mathbf{V}_i^{ca+} = [\mathbf{V}_1^{sub} \oplus \dots \oplus \mathbf{V}_{i-1}^{sub} \oplus \mathbf{V}_i^{ca} \oplus \mathbf{V}_{i+1}^{sub} \oplus \dots \oplus \mathbf{V}_N^{sub}], \quad (15)$$

$$\mathbf{M}_i^{ca+} = [\mathbf{M}_1^{ca} \dots \mathbf{M}_{i-1}^{ca} \oplus \mathbf{I} \oplus \mathbf{M}_{i+1}^{ca} \oplus \dots \oplus \mathbf{M}_N^{ca}], \quad (16)$$

$$\mathbf{h}_i^{ca} = \text{Softmax}(\mathbf{Q}_i^{ca} \cdot \mathbf{K}_i^{ca+T} / \sqrt{d_k} + \log \mathbf{M}_i^{ca+}) \cdot \mathbf{V}_i^{ca+}. \quad (17)$$

Experiment

Actor-Bench. To provide a fair and objective measurement of this novel task, we present ActorBench, the first layout-to-consistent-image generation benchmark. It includes 100 single-subject sets and 100 double-subject sets. Every set consists of four prompt-box pairs of the same central subject(s). In the endeavor for prompts and boxes that comply with objective principles, we direct our focus on COCO2017 (Lin et al. 2014), a detection dataset of real-world photos. We utilize the train set annotations to obtain naturally associated subject-box pairs. We first perform data cleaning to remove boxes that are excessively small or positioned too close to the edge. Subjects that are inherently uniform in appearance (e.g. *apple*) or unlikely to be personalized (e.g. *plane*) are also removed. Then we collect four boxes of the same subject as a single-subject set and four double-boxes of the same double-subjects that exist in the same image as a double-subject set. All the subjects are divided into three main types: *human*, *animal*, and *object* and we instruct ChatGPT (OpenAI 2023) to randomly convert some subjects to other more creative subjects of the same type (e.g. *dog*→*dragon*) and retain the corresponding boxes. Finally, we instruct ChatGPT to generate four formatted prompts for every set of subject(s): *[appearance]+[action]+[background]+[style]*. Note that *appearance* and *style* remain the same within one set and *action* is only applicable for *human* and *animal*.

Metrics. For evaluation, we perform subject-driven segmentation using Grounded-SAM (Ren et al. 2024) to obtain the subject boxes and separate the foregrounds (fg) and backgrounds (bg) of generated images. DINO (Oquab et al. 2023), CLIP (Radford et al. 2021) and LPIPS (Zhang et al. 2018) are utilized to extract visual embeddings. We introduce four dimensions of metrics: (1) layout alignment: we report the mean IoU (mIoU) between the detected boxes and the given boxes; (2) subject consistency: we calculate the cosine similarity among the visual embeddings of foregrounds to obtain DINO-fg and CLIP-fg. LPIPS-fg is also computed;

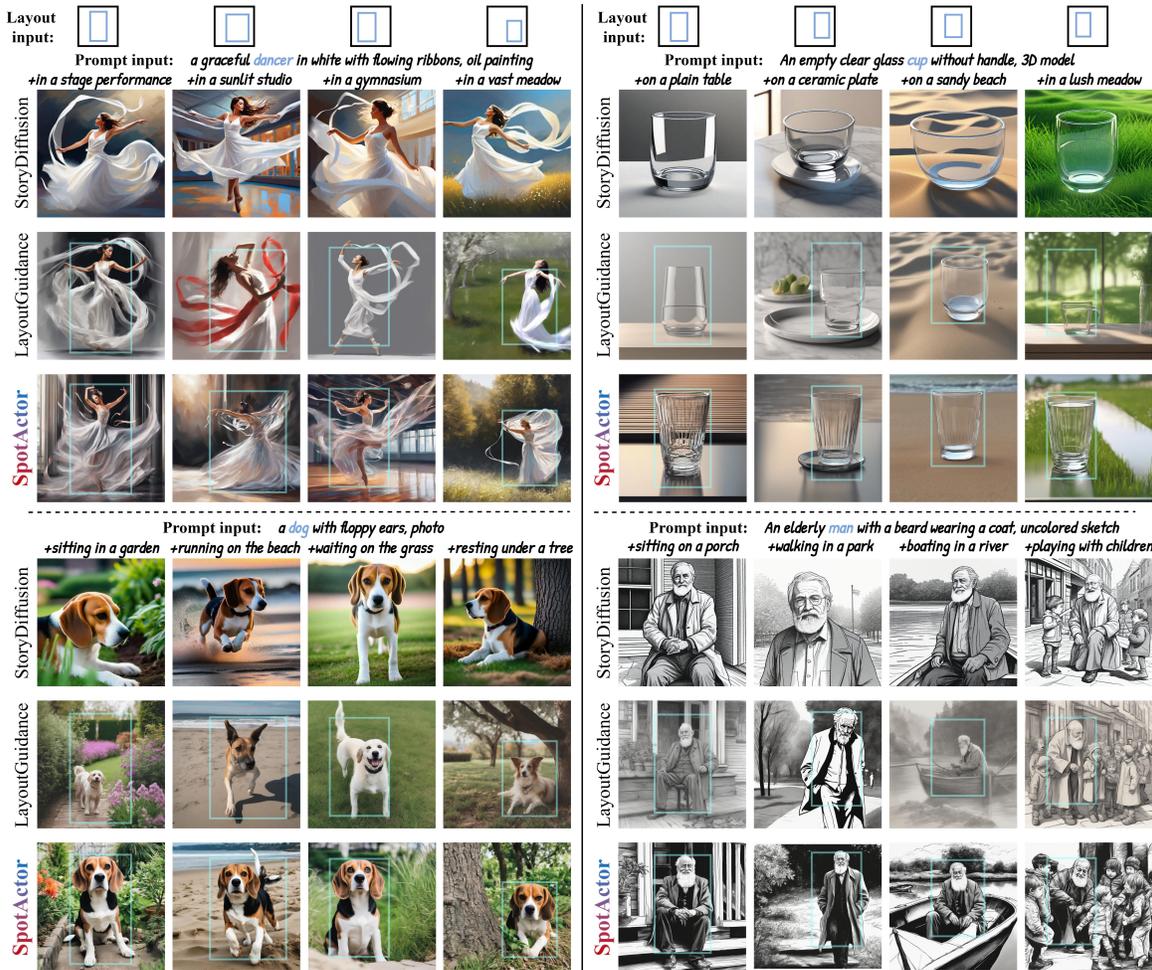


Figure 4: The qualitative comparison between baselines and our SpotActor. Our method shows superior layout controllability compared to LayoutGuidance and exhibits better subject consistency compared to StoryDiffusion. The central subjects are marked in blue and the given boxes are outlined in blue lines.

(3) prompt conformity: we report the CLIP-T-Score (Hessel et al. 2021) among the whole images; (4) background diversity: we calculate the scores of backgrounds to obtain DINO-bg, CLIP-bg and LPIPS-bg.

Baselines. To comprehensively evaluate the performance of SpotActor, we establish two training-free state-of-the-art models as baselines: consistent subject generation pipeline, StoryDiffusion (Zhou et al. 2024b) and layout control pipeline, LayoutGuidance (Chen, Laina, and Vedaldi 2024), both implemented on SDXL (Podell et al. 2023).

Qualitative Evaluation. We illustrate the single subject generation results of baselines and our method in Fig. 4. As shown, StoryDiffusion exhibits competent consistency of subject appearance among images yet fails to be controlled by layout. With energy guidance, LayoutGuidance is able to place the subjects according to the given bounding boxes, but its inadequate approximation of the activation distribution leads to imperfectly aligned subjects (e.g. *cup*, *dog*). By contrast, on the one hand, our SpotActor demonstrates superior

layout controllability. Benefiting from the nuanced and smooth sigmoid-based energy function, the subjects seamlessly adhere to the given edges and fill the whole boxes. On the other hand, our pipeline maintains subject consistency decently with the proposed intricate attention mechanisms. As illustrated in Fig. 5, our SpotActor naturally facilitates the multiple objects generation and continues to perform well in this scenario. The results prove that our method effectively fulfills the expectations of the L2CI generation.

Quantitative Evaluation. In Tab. 1, we display the results of the quantitative metrics between our method and baselines, which consist of four evaluation dimensions. For layout alignment, our method achieves a remarkable 67.1% on mIoU and surpasses the LayoutGuidance by a large margin, which demonstrates our superior layout controllability. Meanwhile, in the dimension of subject consistency, our method scores the best at 78.6% and 79.7% on DINO-fg and CLIP-fg, showing the excellent capacity to maintain a consistent appearance of subjects. For prompt con-

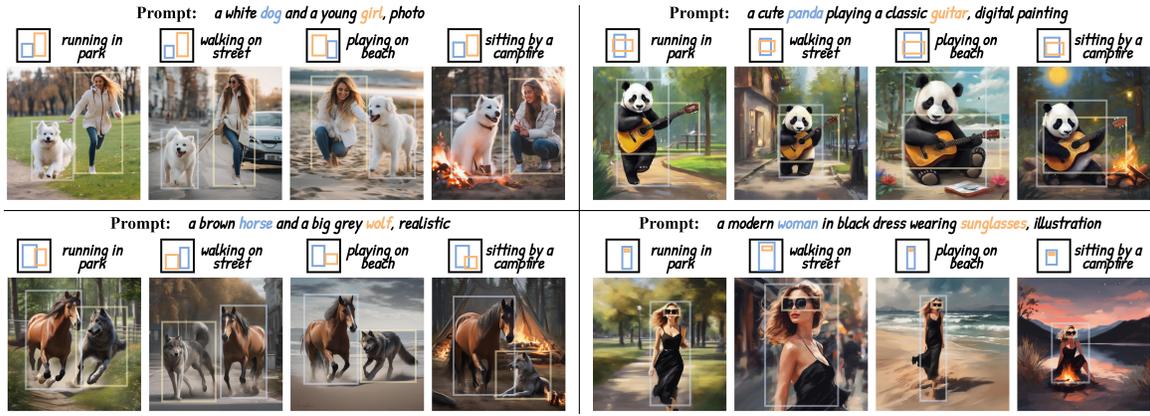


Figure 5: Illustration of double subject generation by SpotActor. Our method maintains excellent performance when handling multiple subjects. Different central subjects are marked in different colors.

Method	Layout alignment	Subject consistency			Prompt conformity	Background diversity		
	mIoU(↑)	DINO-fg(↑)	CLIP-fg(↑)	LPIPS-fg(↓)	CLIP-T(↑)	DINO-bg(↓)	CLIP-bg(↓)	LPIPS-bg(↑)
SDXL	32.1	53.6	54.9	43.6	65.4	30.9	38.9	60.8
StoryDiffusion	29.5	<u>75.2</u>	<u>78.1</u>	33.1	63.5	38.6	46.8	57.3
LayoutGuidance	<u>53.7</u>	53.7	55.3	40.8	54.3	29.6	35.7	<u>60.1</u>
Ours (full)	67.1	78.6	79.7	<u>34.9</u>	<u>63.6</u>	37.8	49.5	56.8

Table 1: The quantitative results of baselines and our SpotActor. All the values are represented in percentage form. The best and second-best results are denoted in bold and underlined.

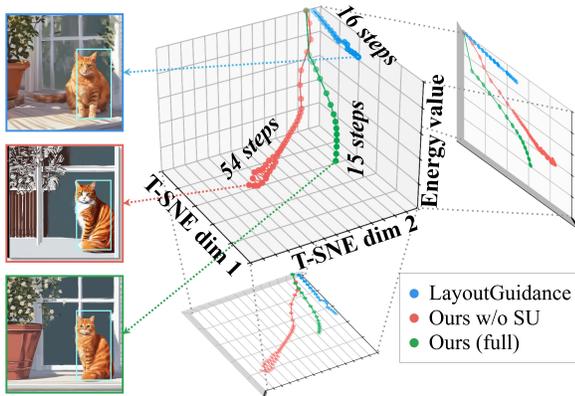


Figure 6: The optimization trajectories of different guidance strategies. The XY-plane represents the dual space after T-SNE and the Z-axis corresponds to the normalized energy.

formity, our method is second only to the original model with a narrow 1.8% margin, which proves that our method preserves the great prompt controllability of the original model. Furthermore, due to the inherent bias, consistent generation pipelines inevitably sacrifice background diversity compared to SDXL, and our method exhibits competitive performance with StoryDiffusion. To summarize, our model displays strong and balanced performance in the four-dimensional quantitative evaluation.

Ablation Study To evaluate the validity of dual energy guidance, which is the core of our work, we collect the latent codes, semantic embeddings and energy values in each update iteration of LayoutGuidance, our model excluding semantic update (ours w/o SU) and our full model. We perform T-SNE to the combined latent codes and embeddings to represent the dual space with two dimensions. Thus, we visualize the optimization trajectories of the dual space with the energy value in Fig. 6. Each sequence of energy values is normalized respectively. As illustrated, with dual energy guidance, our method rapidly converges to the optimum with improved layout alignment and no image quality degradation.

Conclusion

This paper pioneers a novel training-free pipeline, SpotActor, for layout-to-consistent-image generation task. Considering the latent and semantic as a cohesive unit, we propose a new formalization of dual energy guidance including two stages. To perfectly align the subject to the given layout in the optimizing stage, we design a nuanced layout energy based on in-depth analysis. Later in the sampling stage, we enhance the backbone with intricate attention mechanisms to strength the latent-semantic interactions, contributing to the consistent appearance of the generated subject. We further present a specialized benchmark, ActorBench, for evaluation. Comprehensive experiments highlight the effectiveness of our method with superior layout alignment, subject consistency as well as generation efficiency.

Acknowledgements

This work was supported by National Science and Technology Major Project under Grant No. 2022ZD0117103, the Key Research and Development Project in Shaanxi Province No. 2022GXLH-01-03, National Natural Science Foundation of China under Grant No. 62302384, No. 62192781, No. 62172326, No. 62137002 and No. 62403429, China Postdoctoral Science Foundation under Grant No. 2023M742790, Research Project Funded by the State Key Laboratory of Communication Content Cognition under Grant No. A202403, and by the Project of China Knowledge Centre for Engineering Science and Technology.

References

- Avrahami, O.; Hayes, T.; Gafni, O.; Gupta, S.; Taigman, Y.; Parikh, D.; Lischinski, D.; Fried, O.; and Yin, X. 2023. Spa-Text: Spatio-Textual Representation for Controllable Image Generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Avrahami, O.; Hertz, A.; Vinker, Y.; Arar, M.; Fruchter, S.; Fried, O.; Cohen-Or, D.; and Lischinski, D. 2024. The Chosen One: Consistent Characters in Text-to-Image Diffusion Models. In *ACM Special Interest Group for Computer GRAPHICS*.
- Balaji, Y.; Nah, S.; Huang, X.; Vahdat, A.; Song, J.; Zhang, Q.; Kreis, K.; Aittala, M.; Aila, T.; Laine, S.; et al. 2022. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*.
- Chen, M.; Laina, I.; and Vedaldi, A. 2024. Training-Free Layout Control with Cross-Attention Guidance. In *IEEE/CVF Winter Conference on Applications of Computer Vision*.
- Couairon, G.; Careil, M.; Cord, M.; Lathuilière, S.; and Verbeek, J. 2023. Zero-shot spatial layout conditioning for text-to-image diffusion models. In *IEEE/CVF International Conference on Computer Vision*.
- Epstein, D.; Jabri, A.; Poole, B.; Efros, A. A.; and Holynski, A. 2023. Diffusion Self-Guidance for Controllable Image Generation. In *Advances in Neural Information Processing Systems*.
- Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2023. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. In *International Conference on Learning Representations*.
- Hertz, A.; Mokady, R.; Tenenbaum, J.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2023. Prompt-to-Prompt Image Editing with Cross-Attention Control. In *International Conference on Learning Representations*.
- Hessel, J.; Holtzman, A.; Forbes, M.; Bras, R. L.; and Choi, Y. 2021. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In Moens, M.; Huang, X.; Specia, L.; and Yih, S. W., eds., *Conference on Empirical Methods in Natural Language Processing*.
- Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems*.
- Ho, J.; and Salimans, T. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.
- Kim, Y.; Lee, J.; Kim, J.-H.; Ha, J.-W.; and Zhu, J.-Y. 2023. Dense text-to-image generation with attention modulation. In *IEEE/CVF International Conference on Computer Vision*.
- Kwon, G.; Jenni, S.; Li, D.; Lee, J.; Ye, J. C.; and Heilbron, F. C. 2024. Concept Weaver: Enabling Multi-Concept Fusion in Text-to-Image Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Li, Z.; Cao, M.; Wang, X.; Qi, Z.; Cheng, M.-M.; and Shan, Y. 2024. Photomaker: Customizing realistic human photos via stacked id embedding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Lin, T.; Maire, M.; Belongie, S. J.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision*.
- Maharana, A.; Hannan, D.; and Bansal, M. 2022. StoryDALL-E: Adapting Pretrained Text-to-Image Transformers for Story Continuation. In *European Conference on Computer Vision*.
- 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Nie, W.; Liu, S.; Mardani, M.; Liu, C.; Eckart, B.; and Vahdat, A. 2024. Compositional Text-to-Image Generation with Dense Blob Representations. In *International Conference on Machine Learning*.
- OpenAI. 2023. ChatGPT. <https://chat.openai.com/>.
- Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. 2023. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.
- Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.
- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In Meila, M.; and Zhang, T., eds., *International Conference on Machine Learning*.
- Rahman, T.; Lee, H.; Ren, J.; Tulyakov, S.; Mahajan, S.; and Sigal, L. 2023. Make-A-Story: Visual Memory Conditioned Consistent Story Generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Ren, T.; Liu, S.; Zeng, A.; Lin, J.; Li, K.; Cao, H.; Chen, J.; Huang, X.; Chen, Y.; Yan, F.; et al. 2024. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*.

- Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention*.
- Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Sohl-Dickstein, J.; Weiss, E. A.; Maheswaranathan, N.; and Ganguli, S. 2015a. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. In Bach, F. R.; and Blei, D. M., eds., *International Conference on Machine Learning*.
- Sohl-Dickstein, J.; Weiss, E. A.; Maheswaranathan, N.; and Ganguli, S. 2015b. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. In *International Conference on Machine Learning*.
- Song, Y.; and Ermon, S. 2019. Generative Modeling by Estimating Gradients of the Data Distribution. In *Advances in Neural Information Processing Systems*.
- Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2021. Score-Based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations*.
- Song, Y.; Zhang, Z.; Lin, Z. L.; Cohen, S.; Price, B. L.; Zhang, J.; Kim, S. Y.; and Aliaga, D. G. 2023. ObjectStitch: Object Compositing with Diffusion Model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Tewel, Y.; Kaduri, O.; Gal, R.; Kasten, Y.; Wolf, L.; Chechik, G.; and Atzmon, Y. 2024. Training-free consistent text-to-image generation. *ACM Transactions on Graphics*.
- Tumanyan, N.; Geyer, M.; Bagon, S.; and Dekel, T. 2023. Plug-and-Play Diffusion Features for Text-Driven Image-to-Image Translation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Wang, J.; Yan, C.; Lin, H.; and Zhang, W. 2024. OneActor: Consistent Character Generation via Cluster-Conditioned Guidance. *arXiv preprint arXiv:2404.10267*.
- 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 *IEEE/CVF International Conference on Computer Vision*.
- Yang, Z.; Wang, J.; Gan, Z.; Li, L.; Lin, K.; Wu, C.; Duan, N.; Liu, Z.; Liu, C.; Zeng, M.; and Wang, L. 2023. ReCo: Region-Controlled Text-to-Image Generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *IEEE Conference on Computer Vision and Pattern Recognition*.
- Zhao, M.; Bao, F.; Li, C.; and Zhu, J. 2022. EGSDE: Unpaired Image-to-Image Translation via Energy-Guided Stochastic Differential Equations. In *Advances in Neural Information Processing Systems*.
- Zhou, D.; Li, Y.; Ma, F.; Zhang, X.; and Yang, Y. 2024a. MIGC: Multi-Instance Generation Controller for Text-to-Image Synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Zhou, Y.; Zhou, D.; Cheng, M.-M.; Feng, J.; and Hou, Q. 2024b. StoryDiffusion: Consistent Self-Attention for Long-Range Image and Video Generation. *arXiv preprint arXiv:2405.01434*.