CreatiLayout: Siamese Multimodal Diffusion Transformer for Creative Layout-to-Image Generation

Supplementary Material

A. More Details on SiamLayout-FLUX

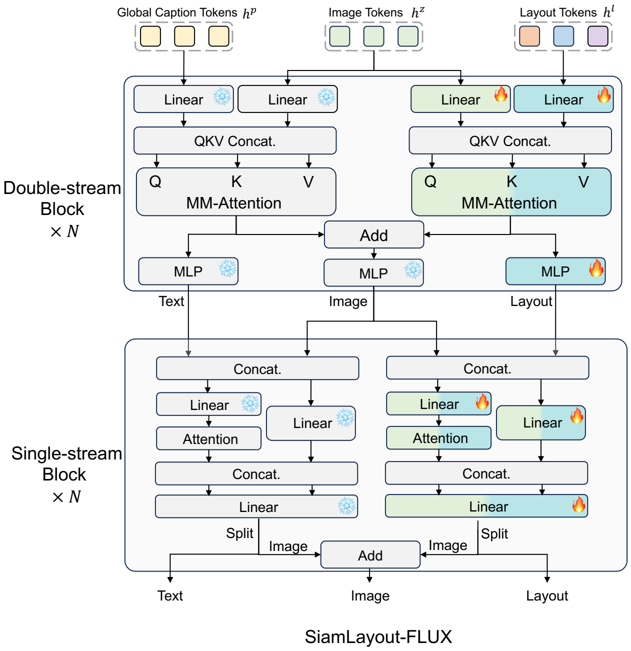

Figure 10. An overview of SiamLayout-FLUX.

FLUX is another outstanding text-to-image generative model based on the Multimodal Diffusion Transformer (MM-DiT) architecture, demonstrating remarkable performance alongside SD3. The core components of FLUX include the double-stream block and the single-stream block. To further verify the generality of our proposed SiamLayout approach, we integrate SiamLayout into FLUX and propose SiamLayout-FLUX, aiming to empower FLUX for layout-to-image generation.

Specifically, as illustrated in Fig. 10, for the Double Stream Block in FLUX, we adopt the same layout integration strategy as presented in Fig. 2. For the Single Stream Block, we maintain the core concept from SiamLayout, which emphasizes preserving the advantage of MM-Attention in facilitating effective multimodal interactions while reducing interference and competition among modalities. Therefore, we similarly introduce a Siamese branch dedicated to processing the interactions between images and layouts. Such a design enables the layout information to independently and concurrently guide the image content generation, analogous to the role played by the global text caption.

Both quantitative and qualitative experimental results demonstrate that our proposed SiamLayout-FLUX outperforms previous methods by a clear margin, validating the effectiveness of SiamLayout and its general applicability across different MM-DiTs.

B. LoRA Version of SiamLayout

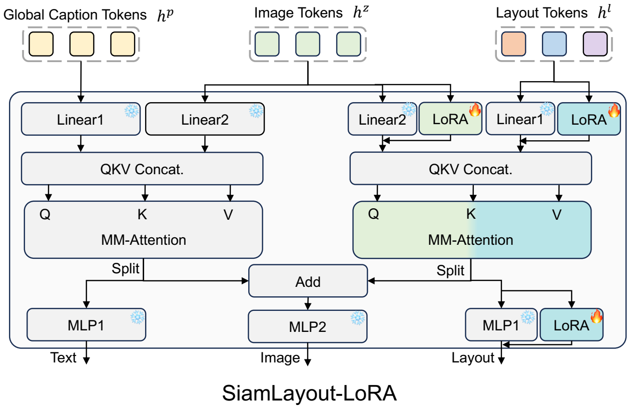

Figure 11. An overview of SiamLayout-LoRA.

We further propose a more lightweight variant for layout control, named SiamLayout-LoRA, which achieves comparable layout control accuracy with fewer additional parameters. As illustrated in Fig. 11, SiamLayout-LoRA integrates Low-Rank Adaptation (LoRA) modules into the Linear1, Linear2, and MLP1 layers, efficiently handling image-layout interactions based on the frozen pre-trained weights. This LoRA-based variant provides alternative trade-offs between extra parameter overhead and layout control accuracy. In our experiments, we set the LoRA rank to 256.

C. Discussion on Extra Parameters and Computation Introduced by Layout Control

	Extra Params ↓	Extra MACs ↓	Spatial ↑	Color ↑	Texture ↑	Shape ↑
GLIGEN	24.20%	30.80%	77.53	49.41	55.29	52.72
InstanceDiff	42.67%	155.83%	87.99	69.16	72.78	71.08
HiCo	42.03%	263.70%	87.04	69.19	72.36	71.1
Layout Adapter	16.90%	21.70%	88.43	71.67	73.56	72.61
M^3-Attention	48.50%	**15.70%**	79.15	60.19	62.96	61.29
SiamLayout-SD3	62.40%	38.70%	**92.67**	**74.45**	**77.21**	**75.93**
SiamLayout-SD3-LoRA	**15.90%**	39.90%	89.71	71.89	74.14	72.63

Table 7. Comparison of extra parameters and computation. **Bold**, underline represent the best and second best methods, respectively.

As shown in Tab. 7, we compare different approaches with respect to extra parameters and computational overhead (i.e., MACs required for one denoising step given a layout containing 10 entities). To ensure a fair comparison, we report the relative increase in parameters and computation introduced by integrating layout control compared to the base model.

Our proposed approach outperforms others at comparable additional parameter counts while requiring significantly lower extra computational overhead compared to InstanceDiff and HiCo. This is due to InstanceDiff and HiCo computing image-layout attention separately for each entity before integrating them, causing their computational overhead to increase linearly as the number of entities grows. In addition, our proposed LoRA variant achieves comparable layout control accuracy with substantially fewer extra parameters.

D. More Details on Datasets and Benchmarks

D.1. LayoutSAM Dataset and Benchmark

Layout annotation pipeline. We design a mechanism to automatically annotate the layout for any given image.

I) Image Filtering: We employ the LAION-Aesthetics predictor [34] to assign aesthetic scores to images and filter out those with low scores. For SAM [31], we analyze the aesthetic scores shown in Fig. 12 and curate a high visual quality subset consisting of images in the top 50% of aesthetic scores.

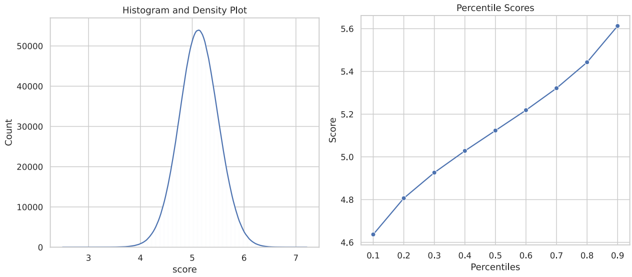

Figure 12. Distribution of aesthetic scores of SAM.

II) Global Caption Annotation: We generate the detailed descriptions for the query image using the Qwen-VL-Chat-Dense-Captioner [17], which is a vision-language model fine-tuned on generative and human-annotated data using LoRA. This model supports accurate and detailed image descriptions. The average length of the captions is 95.41 tokens, measured by the CLIP tokenizer.

III) Entity Extraction: Existing state-of-the-art open-set grounding models [44, 56] perform better at detecting entities using a list of short phrases than directly using dense captions. Thus, we utilize the large language model Llama3.1-8b-it [46] to extract the main entities from dense captions via in-context learning. The brief descriptions include simple attribute descriptions with an average length of 2.08 tokens.

IV) Entity Spatial Annotation: We use Grounding DINO [44] to annotate bounding boxes of entities. To clean noisy data, we design the following filtering rules. We first filter out bounding boxes that occupy less than 2% of the total image area, then only retain images with 3 to 10 bounding boxes. The average number of entities per image is 3.96.

V) Region Caption Recaptioning: We use the vision language model MiniCPM-V-2.6 [49] to generate fine-grained descriptions with complex attributes for each entity based on its visual content and brief description. The generated detailed descriptions generally cover attributes such as color, shape, texture, and some text details, with an average length of 15.07 tokens.

Finally, we contribute the large-scale layout dataset LayoutSAM, which includes 2.7 million image-text pairs and 10.7 million entities. Each entity is annotated with a bounding box and a detailed description. Fig. 13 shows some examples from LayoutSAM.

LayoutSAM-Eval Benchmark. The LayoutSAM-Eval benchmark, constructed from LayoutSAM, serves as a comprehensive tool for evaluating layout-to-image generation quality. It consists of 5,000 layout data points. We evaluate layout-to-image generation quality using LayoutSAM-Eval from two aspects: Spatial and Attribute, both of which are evaluated via the vision language model MiniCPM-V-2.6 [49] in a visual question-answering manner.

– *Spatial Accuracy.* To measure spatial adherence, for each bounding box, we ask the VLM whether the given entity exists within the bounding box, with the answer being either “Yes” or “No.” Finally, we divide the number of entities with a “Yes” answer by the total number of entities to obtain the spatial score.
– *Attribute Accuracy.* To measure attribute adherence, we ask the VLM whether the entity within the bounding box matches the attributes in the detailed description. For attributes like color, shape, and texture, each attribute is evaluated independently through visual question answering, and the score is obtained in the same manner as the spatial score.

D.2. Layout Planning Dataset and Benchmark

Layout Planning Dataset. To train LayoutDesigner, we construct a layout planning dataset derived from LayoutSAM. It consists of a total of 180,000 data points, covering the following three tasks, each with 60,000 data points:

– *Caption-to-layout generation.* We randomly select data from LayoutSAM to construct pairs of global captions and ground truth layouts of entities. Each entity includes a bounding box and a description. This portion of the data is used to train the generation of layouts based on global captions.

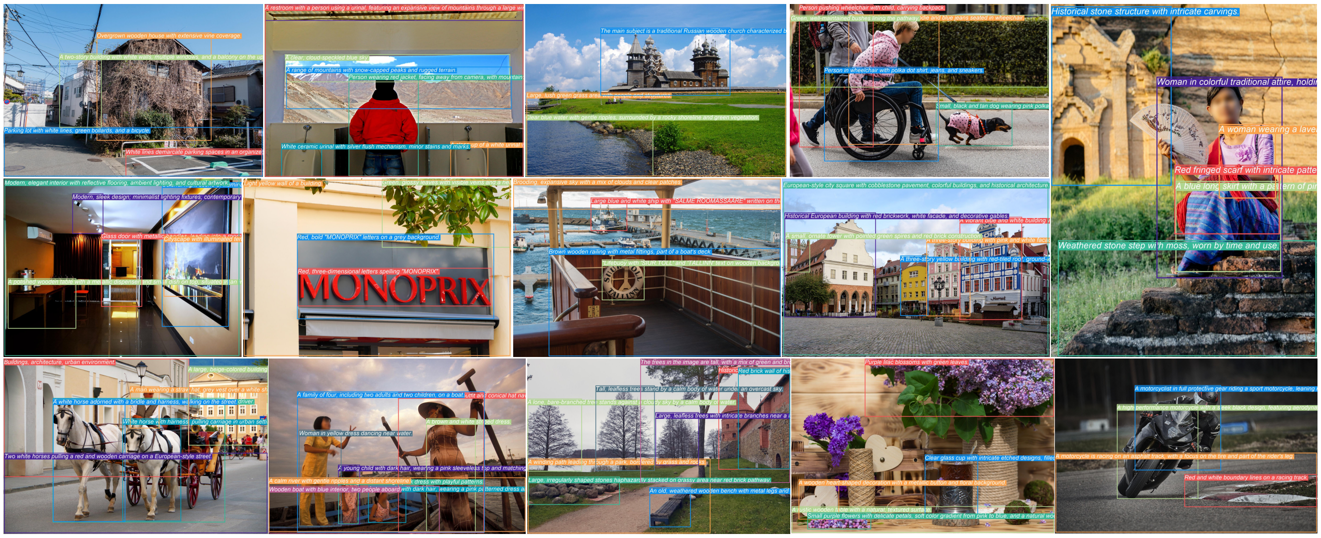
Historical stone structure with intricate carvings.
Woman in colorful traditional attire, holdi
Red fringed scarf with intricate patte
Weathered stone step with moss, worn by time and use.
MONOPRIX

Figure 13. Examples from the LayoutSAM dataset.

- *Center point-to-layout generation.* For each entity, we calculate the center point from its bounding box to construct pairs of center points and ground truth layouts of entities. This portion of the data is used to train the generation of layouts based on the center points of entities.
- *Suboptimal layout-to-layout optimization.* For each entity, we create suboptimal layouts by performing operations such as deletion, duplication, movement, and resizing with a certain probability. This portion of the data consists of pairs of suboptimal layouts and GT layouts and is used to train the optimization of layouts from suboptimal to better.

LayoutDesigner is a unified layout planning model that supports all three tasks simultaneously through joint training of a large language model on these three types of data.

Layout Planning Benchmark. We construct 1,000 data points for each layout planning task using the same method from LayoutSAM-Eval and conduct experiments to evaluate layout planning capabilities under a 3-shot in-context learning setting. First, we evaluate the formatting accuracy of the generated layouts, including the coordinates of the top-left corner being smaller than those of the bottom-right corner and the bounding box not exceeding the image boundaries. To assess the harmony and aesthetics of the layouts, we did not use metrics like AP to measure the adherence of the generated bounding boxes to the ground truth. This is because layout generation is an open-ended problem, and there can be multiple optimal solutions for the input. Even if a solution does not resemble the GT, it can still be an excellent layout. Therefore, we generate images based on the designed layouts and reflect the quality of the layouts through the quality of the images. Additionally, we evaluate the layout planning capabilities through qualitative results.

E. More Analysis on Modal Competition

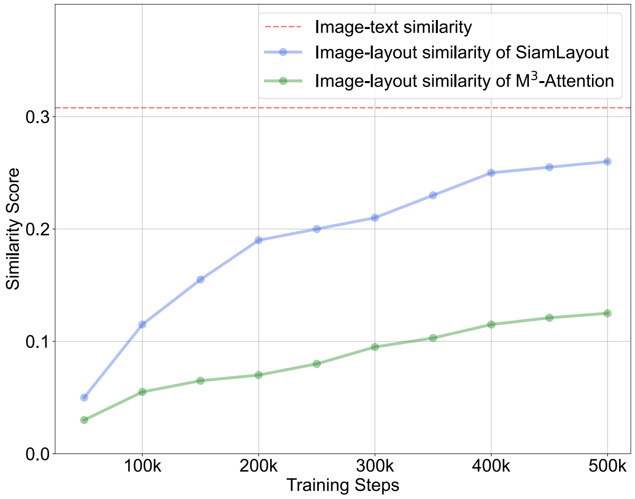

Figure 14. The trend of image-layout similarity in the attention maps of different network variants.

In Fig. 3, we analyze the issue of modality competition in M^3-Attention: the similarity between the layout and the image is much lower than that between the global caption and the image. Our proposed SiamLayout alleviates this issue by decoupling it into two siamese MM-Attention branches: image-text and image-layout. We investigate the image-text and image-layout similarity scores in the attention map to further illustrate the modality competition, as shown in Fig. 14. The image-layout similarity score is determined by applying softmax to the attention map of the M^3-Attention or image-layout MM-Attention, identifying the cross-region of the image and layout, and then taking the average of the top 1% scores in this cross-region. The image-text similarity score is calculated in the same way. Experimental results show that, as the training steps increase, the image-layout similarity score in M^3-Attention remains consistently much lower than the image-text similarity score, resulting in the layout having a weaker influence on the image compared to the global caption. In contrast, due to the independent guiding role of the layout in the image-layout branch of SiamLayout, the image-layout similarity gradually increases to a value close to the image-text similarity as the network training progresses, thereby allowing the layout to play a more significant guiding role in image generation.

F. More Qualitative Results

We present more qualitative results of SiamLayout-SD3 in Fig. 15. Experimental results show that our proposed method empowers MM-DiT for layout-to-image generation, achieving visually appealing and precisely controllable generation, as demonstrated by the high adherence to complex attributes such as color, texture, and shape.

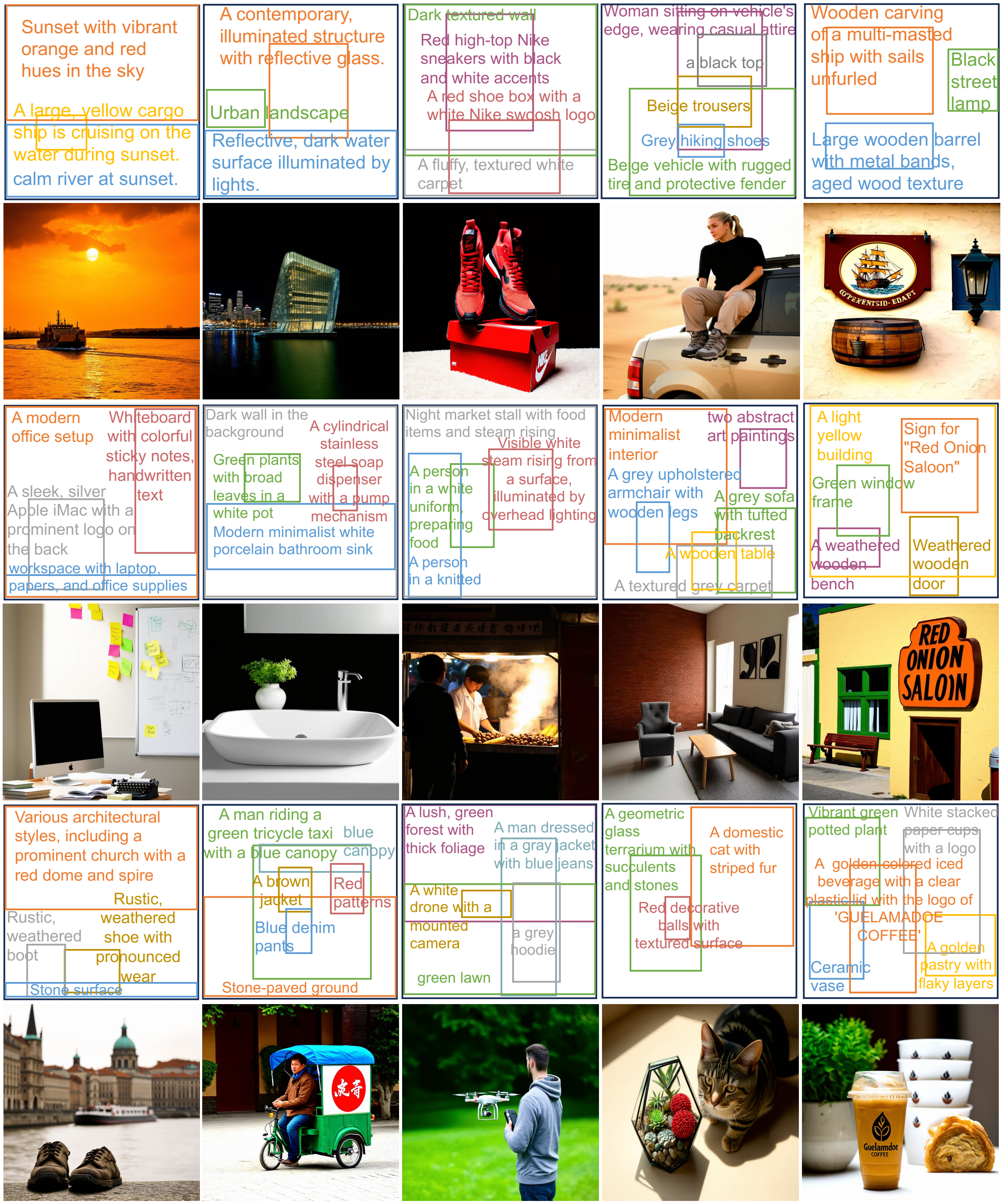
Sunset with vibrant orange and red hues in the sky
A large, yellow cargo ship is cruising on the water during sunset.
calm river at sunset.
A contemporary, illuminated structure with reflective glass.
Urban landscape
Reflective, dark water surface illuminated by lights.
Dark textured wall
Red high-top Nike sneakers with black and white accents
A red shoe box with a white Nike swoosh logo
A fluffy, textured white carpet
Woman sitting on vehicle's edge, wearing casual attire
a black top
Beige trousers
Grey hiking shoes
Beige vehicle with rugged tire and protective fender
Wooden carving of a multi-masted ship with sails unfurled
Black street lamp
Large wooden barrel with metal bands, aged wood texture
A modern office setup
Whiteboard with colorful sticky notes, handwritten text
A sleek, silver Apple iMac with a prominent logo on the back
workspace with laptop, papers, and office supplies
Dark wall in the background
A cylindrical stainless steel soap dispenser with a pump mechanism
Green plants with broad leaves in a white pot
Modern minimalist white porcelain bathroom sink
Night market stall with food items and steam rising
Visible white steam rising from a surface, illuminated by overhead lighting
A person in a white uniform, preparing food
A person in a knitted
Modern minimalist interior
two abstract art paintings
A grey upholstered armchair with wooden legs
A grey sofa with tufted backrest
A wooden table
A textured grey carpet
A light yellow building
Sign for "Red Onion Saloon"
Green window frame
A weathered wooden bench
Weathered wooden door
Various architectural styles, including a prominent church with a red dome and spire
Rustic, weathered boot
Rustic, weathered shoe with pronounced wear
Stone surface
A man riding a green tricycle taxi with a blue canopy
blue canopy
A brown jacket
Red patterns
Blue denim pants
Stone-paved ground
A lush, green forest with thick foliage
A man dressed in a gray jacket with blue jeans
A white drone with a mounted camera
a grey hoodie
green lawn
A geometric glass terrarium with succulents and stones
A domestic cat with striped fur
Red decorative balls with textured surface
Vibrant green potted plant
White stacked paper cups with a logo
A golden-colored iced beverage with a clear plastic lid with the logo of 'GUELAMADOE COFFEE'
Ceramic vase
A golden pastry with flaky layers
OTERENTSID·EDAFI
RED ONION SALOIN
Guelamdoe COFFEE

Figure 15. More qualitative results on layout-to-image generation.