Hybrid Layout Control for Diffusion Transformer: Fewer Annotations, Superior Aesthetics

Supplementary Material

High Aesthetics and Diverse styles Image Generation Prompt

Given the original caption, your task is to generate a high-quality, visually rich version of the description, infused with the given artistic style.
Please enhance the caption with the following requirements:
- **Subject**: Focus on specific, concrete nouns (e.g., ”a majestic lion”, ”a futuristic cityscape”).
- **Action/Context**: What is happening? (e.g., ”roaring in a vast savannah”, ”exploring a neon-lit alley”).
- **Environment**: Mention the surroundings clearly (e.g., ”amidst towering skyscrapers”, ”under a golden sunset sky”).
- **Lighting**: Define light conditions (e.g., ”soft diffused lighting”, ”dramatic shadows with high contrast”).
- **Composition**: Include details about framing and perspective (e.g., ”close-up portrait with bokeh effect”, ”wide-angle shot emphasizing the skyline”).
- **Color Palette**: Mention dominant colors (e.g., ”vibrant reds and yellows”, ”soft pastels”).
- **Modifiers**: Use descriptors like ”sharp focus,” ”realistic textures,” ”ultra-detailed,” and ”high resolution” for clarity and quality.
- **Negative Prompts**: Optionally include terms like ”no distortion”, ”clear object boundaries”, ”no blur”, ”sharp edges”, to help the detector recognize object contours easily.

Your task is to generate a cinematic, evocative caption that maximizes clarity, sharpness, and detail. The goal is to generate a highly defined image where object boundaries are easily detectable for further processing or analysis.
Ensure that the generated caption incorporates these elements while matching the artistic style,with a upper limit of 300 words. Your response should be a string that includes only the enhanced caption, do not output additional content.
Here is an example:
Origional caption: a monk in traditional red robes sitting on the steps of a temple. The monk is holding a string of prayer beads in his hand, seemingly in deep thought or prayer. The background of the photo is the entrance of the temple, with a decorative door and a colorful flag hanging on the wall. The sunlight shines on the monk and the steps, creating a tranquil and solemn atmosphere. Style: Line Drawing Style. Style description: A visual rendering created entirely with clean, continuous lines that emphasize minimalistic form and detailed contours, set against a monochromatic background for a refined, graphic look.
Enhanced caption: A meticulously rendered line drawing portrays a pensive monk in elaborately detailed, traditional red robes, seated gracefully on timeworn temple steps. He holds a string of prayer beads in mid-reflection, embodying deep, solemn meditation. The scene is captured with a close-up cinematic perspective that emphasizes every refined contour of his serene face and flowing garment. Behind him, the ancient temple entrance—with its ornate, intricately outlined door and a boldly defined flag—provides a stately backdrop rendered in monochromatic simplicity. Soft diffused light interacts with dramatic shadows, accentuating the crisp, continuous line work and creating a striking balance between clarity and minimalism. The composition is executed in ultra-detailed, high-resolution style with clear object boundaries, no distortion, and sharp edges, ensuring every element from the graceful posture of the monk to the precise architectural forms stands out with defined sharpness.
Now, output the enhanced caption based on given style and example:
Origional caption:
Style:
Style description:
Enhanced caption:

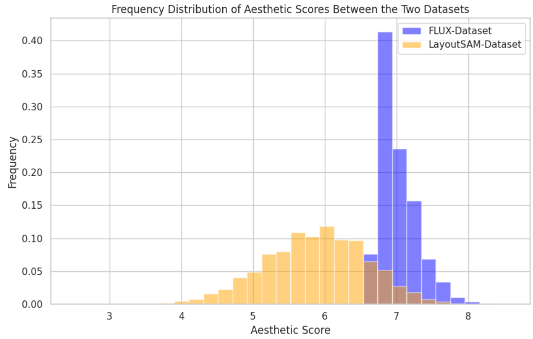


Figure 1. Quantitative comparison of aesthetic score distribution between our dataset and LayoutSAM dataset.

Instruction for High Aesthetic and Diverse Style Dataset Generation. We leverage high-aesthetic training data to address the issue of aesthetic degradation in image generation for the Layout-to-Image task. To further enhance performance, we explore re-optimization using different datasets. Specifically, for the Quality-Tuning data in this paper, we utilize captions from the LayoutSAM dataset. We then refine these captions using a state-of-the-art large language model, following carefully designed instructions to ensure improved quality and coherence.

Aesthetic Score Statistical Comparison between Our Dataset and the Original LayoutSAM Dataset. To demonstrate that our dataset surpasses the original in quality, we randomly sampled 200k data points from the LayoutSAM dataset and visually compared their aesthetic distributions with our dataset. The results clearly show that our dataset exhibits significantly higher quality than the LayoutSAM dataset. The visualization results are presented in Figure 1.

Implementaion Detaill of Scalable Anonymous Layout-Image Data Engine. For the Anonymous Object Detector in our paper, we choose the most advanced YOLO11x as our detector. We use about 500k data in the LayoutSAM dataset as the training set and train our anonymous detector for 40 epochs. We then applied the anonymous detector to our high-aesthetic, multi-style datasets to identify the primary objects in each image. The detector's parameters were configured accordingly, and following the approach of previous work [5, 7], we filtered out bounding boxes with an area smaller than 2% or larger than 80% of the entire image. The number of detected boxes per image was constrained to 3–10, yielding a total of approximately 200k samples. Next, we performed aesthetic sorting and selected the top 50k samples as the quality-tuning dataset. Finally, we manually refined the dataset by filtering out 1k images based on criteria such as removing irrelevant backgrounds and overlapping boxes.

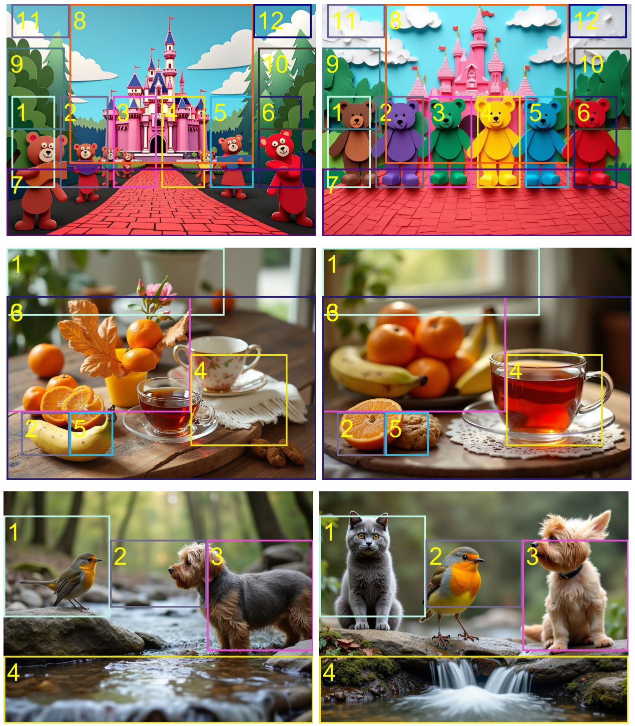


Figure 2. Qualitative comparison results between Regional-Prompting-FLUX (left column) and Ours (right column). Our approach achieves significantly bettera layout adherence while ensuring extremely high visual appealing.

Method	Layout Quality			Inference Time
	AP ↑	AP^{50} ↑	AR ↑	
GroundingDINO	44.5	53.7	61.2	2.37s
Our Anonymous Detector	63.3	74.9	72.0	13.7ms

Table 1. Comparison of Different Anonymous Detector.

Advantages of Our Anonymous Layout Detector. Our anonymous layout detector offers two notable advantages over GroundingDINO. First, it achieves significantly better detection performance, with 63.3/74.9/72.0 on relevant metrics compared to GroundingDINO's 44.5/53.7/61.2. Second, it provides a substantial speedup in inference time, requiring only 13.7ms per image versus GroundingDINO's 2.37 seconds. This efficiency is largely attributed to its YOLO11-based architecture, which enables both high-quality and real-time layout detection. The results are shown in Table 1.

More explanation about our Layout control mechanisms. Our layout control mechanism is rooted in two key innovations: First, the use of regional visual token supervision ensures that essential content is accurately generated within designated bounding boxes, tightly aligning the generation with spatial constraints. Second, by sharing position embeddings between regional and global tokens occupying the same spatial positions, our model enables effective propagation of layout-relevant information throughout the image, reinforcing layout fidelity. Beyond these algorithmic insights, we also introduce a novel downsampled regional diffusion transformer, which encodes layout implicitly through a set of learnable regional tokens—departing from the explicit layout token designs of prior works like GLIGEN or SiamLayout. Finally, our hybrid layout control scheme reduces the reliance on expensive semantic layout annotations while maintaining strong performance. We believe these architectural and training innovations represent meaningful progress toward more efficient and effective layout-aware generation.

Effectiveness in Complex Layout Scenarios and Conflicting Regional Prompts. To directly address the concern regarding our model's effectiveness in complex layout scenarios and conflicting regional prompts, we construct two dedicated evaluation subsets from the LayoutSAM-Eval benchmark. The first is LayoutSAM-Complex, containing 153 samples with ultra-dense layouts where each sample includes at least 7 bounding boxes and severe overlapping of multiple objects, designed to test layout-following performance under extreme object density. The second is LayoutSAM-Conflict, a 100-sample subset specifically curated to contain conflicting and overlapping regional prompts, challenging the model's ability to resolve semantic ambiguity in layout instructions. As shown in Table 3 and Table 2, our method significantly outperforms the previous state-of-the-art (SiamLayout-FLUX) across all layout and image quality metrics. These results demonstrate that our method effectively resolves the challenges posed by complex layouts and ambiguous regional prompts through robust layout-aware generation.

Method		**Layout Quality ↑**				**Image Quality**		
	Params	Spatial	Color	Texture	Shape	IR ↑	Pick ↑	CLIP ↑
SiamLayout-FLUX	20.5B	94.98	75.23	75.68	75.99	79.20	21.96	34.15
Ours w/ FLUX.1[dev]	12B	97.26	91.49	93.31	93.47	85.92	22.09	35.13

Table 2. Quantitative comparison on LayoutSAM-Eval conflict region prompts subset. Pick: PickScore.

Method		**Layout Quality ↑**				**Image Quality**		
	Params	Spatial	Color	Texture	Shape	IR ↑	Pick ↑	CLIP ↑
SiamLayout-FLUX	20.5B	95.60	71.55	74.74	74.05	78.75	22.04	33.60
Ours w/ FLUX.1[dev]	12B	96.90	89.22	91.12	90.86	86.38	22.16	34.61

Table 3. Quantitative comparison on LayoutSAM-Eval complex layout subset. Pick: PickScore.

Quality-Tuning Only Slightly Affects the Performance. We empirically find that the quality-tuning step, while improving image fidelity, may slightly harm the generalization ability of the model on unseen prompts or layouts. Specifically, when evaluating on the COCO2017 dataset in a zero-shot setting, our model trained with quality tuning achieves 12.7 / 30.0 / 19.1 on AP / AP50 / AR, whereas the model without quality tuning performs slightly better with 14.0 / 30.2 / 20.6. This suggests a minor trade-off between visual quality and layout generalization.

Effect of Choosing Different CFG Values. We need to clarify that it is non-trivial to fine-tune FLUX.1-[dev], as it is trained with guidance distillation for efficiency. We find that simply fine-tuning FLUX.1-[dev] with the default guidance value of 3.5 results in severe artifacts after fine-tuning for thousands of iterations. To this end, we set the guidance to 1.0 during the entire fine-tuning stage, following [4]. We also adopt the additional true CFG from ART [3] to enhance the visual appeal of the generated images. We conduct ablation experiments to study the effect of choosing different CFG values for the layout control model based on FLUX.1-[dev]. We clarify that we do not tune the CFG values for the experiments based on SD3 to ensure fair comparisons with the previous SiamLayout [6].

Visualization of Average Attention Scores of Our Layout Model. To demonstrate the planning capability of our model, we average the attention scores for all visual tokens within each bounding box as they attend to specific entities in the global prompt. As shown in Figure 4, these scores concentrate on the entities within the bounding boxes, indicating that the model effectively determines which object should be assigned to each region of the image.

CFG values.	Layout Quality				Image Quality			
	Spatial↑	Color↑	Texture↑	Shape↑	HPSv2 ↑	AEV2.5 ↑	L-AE ↑	IR ↑
1.5-1.5	94.37	87.69	89.57	89.05	0.280	<u>6.095</u>	5.597	0.819
2.0-2.0	94.94	89.15	90.19	89.72	0.290	6.000	5.698	0.874
2.5-2.5	94.57	87.17	89.25	88.68	0.293	5.777	5.742	0.887
3.5-3.5	95.04	86.96	89.41	88.63	<u>0.293</u>	5.408	<u>5.741</u>	<u>0.854</u>

Table 4. Effect of choosing different CFG values after quality tuning: the first value is used in the guidance distillation, while the second value is used in the additional true CFG settings.

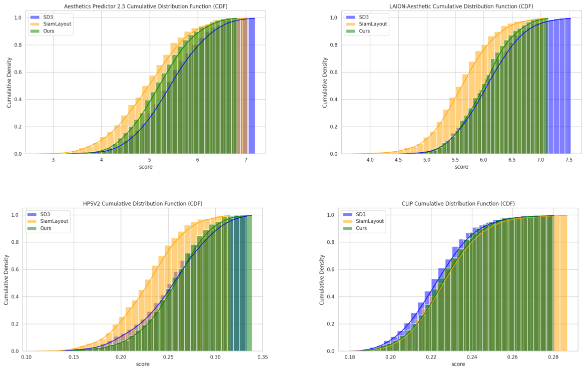


Figure 3. Cumulative distribution function of SD3, SiamLayout and our model on style test benchmark.

Details of Our User Study. To comprehensively assess the quality and visual effectiveness of images generated by our model, we conducted a user study with 15 participants from diverse professional backgrounds. Each participant was presented with 100 image pairs generated by our model and SiamLayout [6] across various artistic styles. They were asked to compare the images based on three key criteria:

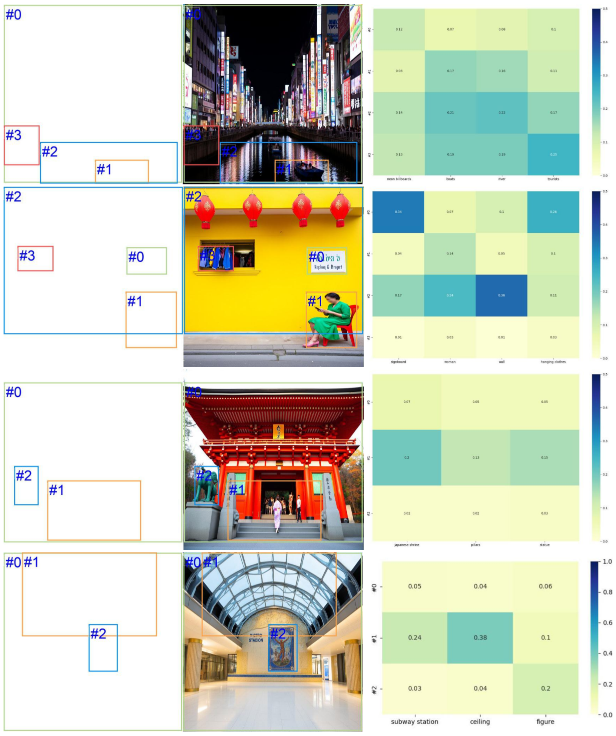


Figure 4. Visualization of the average attention scores for all visual tokens within the same bounding box (as query) attending to the entities in the global prompt (as key). These scores illustrate that our model assigns significant attention weights to the entities present within the region.

- **Visual Appeal:** Which method produced more aesthetically compelling and visually striking results.
- **Caption-Content Consistency:** The degree to which the generated image aligns with the text description, with a focus on stylistic coherence and content completeness.
- **Positional Accuracy of Key Objects:** Whether the main objects described in the captions were correctly placed within their designated bounding boxes.

Participants provided responses through a structured questionnaire. The results of the study indicate that our model significantly outperformed SiamLayout across all evaluated metrics, achieving higher user approval in terms of aesthetic quality, caption-content consistency, and the accurate placement of key objects.

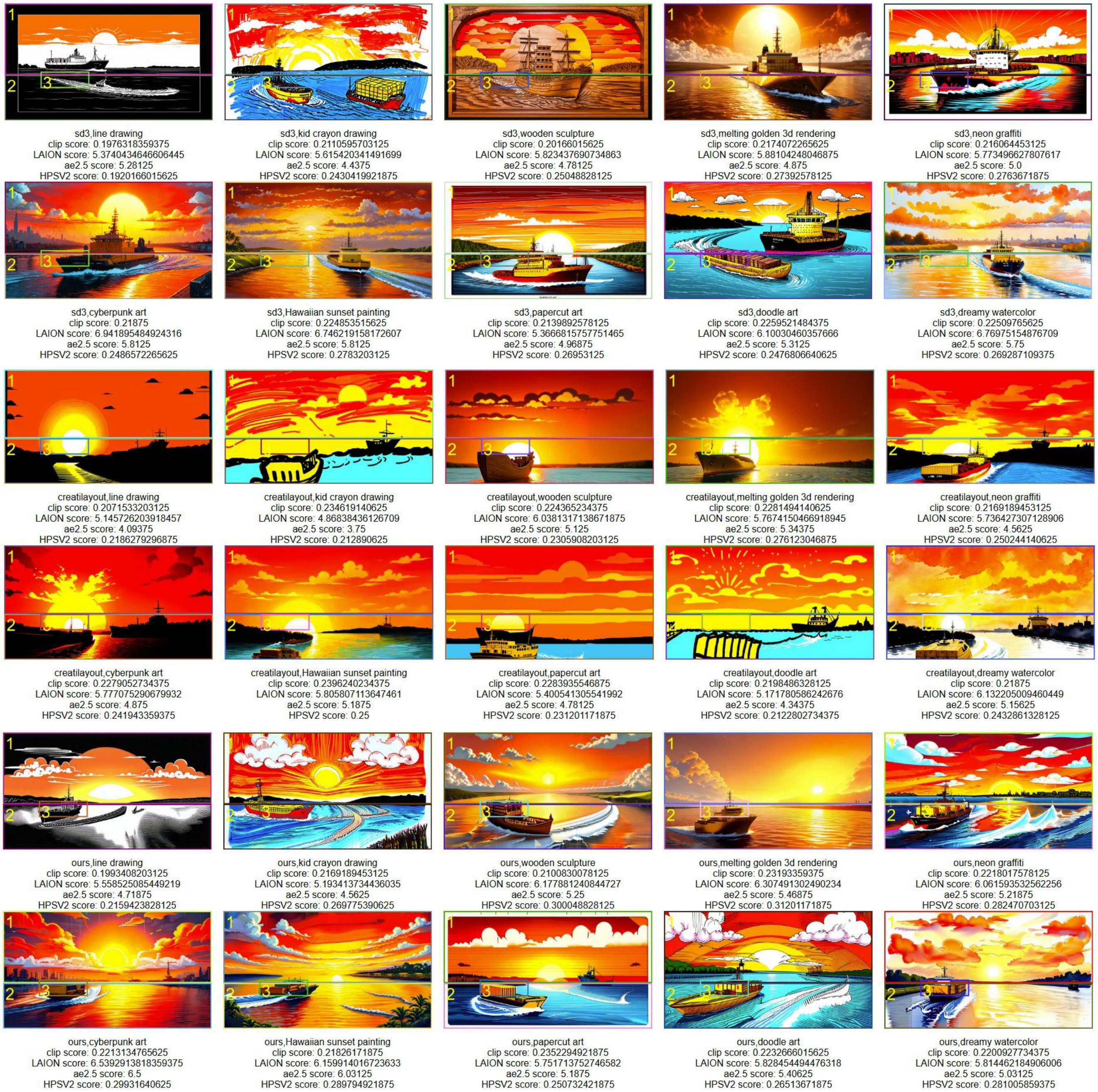
Construction and Use of Our Style Evaluation. In previous Layout-to-Image tasks, little attention has been given to addressing the decline in aesthetic quality in generated images. In this work, we systematically investigate how to preserve the original aesthetic capabilities of the T2I model while ensuring effective layout control. Our study focuses on two key aspects: aesthetic retention and prompt adherence, particularly from the perspective of style consistency. To evaluate the aesthetic retention ability of different models, we generate test prompts with distinct artistic styles. We then assess and compare these models using multiple metrics, including CLIP Score, HPS v2, Aesthetic Predictor 2.5, and the LAION-Aesthetics Predictor. The visual comparison is shown in Figure 3. It can be seen from the figure that our model is closer to the original SD3 model in multiple indicators. The qualitative comparison is shown in Figure 5. We also show the comparison results between our FLUX-based model and the original FLUX in Figure 6 for reference.

Comparison between Our Model and Regional-Prompting-FLUX. Regional-Prompting-FLUX [1] represents the most advanced training-free Layout-to-Image model. Since it does not modify the original Flux architecture, its aesthetic capabilities remain comparable to those of the base Flux model. In Figure 2, we present a visual comparison between our model and Regional-Prompting-FLUX. The results demonstrate that our model achieves precise control over complex objects and attributes while preserving high aesthetic quality. Overall, our approach delivers superior visual results compared to Regional-Prompting-FLUX.

Details of Inference Speed Computation in Table 5 of the Main Text. To benchmark inference speed under different downsampling settings, we measured the generation time for 1024×1024 images using 50 diffusion steps and 5 bounding boxes (each sized at 341×341), following SiamLayout to ensure a fair comparison. To mitigate randomness, we averaged the generation time over 20 runs after a warm-up of 5 runs. All experiments were conducted on an NVIDIA A100 80G GPU. Our method exhibits a clear trade-off between generation speed and spatial precision. Specifically, the runtime under downsample ratios of $1/1$, $1/2$, $1/4$, and $1/8$ are 103.01s, 72.13s, 61.78s, and 59.96s, respectively. In contrast, SiamLayout-FLUX reports a runtime of 68.96s. While our full-resolution model ($\times 1$) is slower due to fine-grained layout conditioning, it offers stronger spatial alignment. The lower ratio variants (e.g., $1/4$ or $1/8$) significantly accelerate inference while maintaining competitive layout adherence. This demonstrates the flexibility of our framework to balance quality and efficiency depending on deployment needs.

Captions Corresponding to the Figure in the Paper. Tables 5, 6, and 7 provide the detailed region caption descriptions corresponding to Figures 1, 5, and 8 in the main text of the paper.

Limitations. While our method demonstrates strong performance and layout control capabilities, we acknowledge two key limitations. First, the quality fine-tuning stage, though beneficial for improving visual fidelity, can introduce implicit style biases. This phenomenon—also observed in prior work such as Emu [2] may lead to overly saturated or stylized outputs. To mitigate this, we plan to expand the style diversity of our training dataset in future work. Second, although our method consistently improves results on the FLUX backbone, its performance on SD3 is slightly lower than expected. We attribute this to architectural differences between the two models, which may violate some of the assumptions made by our layout module. While the SD3 results are included for completeness and to encourage future research, our main claims and findings are centered around FLUX. We advise caution when adapting our module to other architectures like SD3, which may require non-trivial tuning or re-alignment to fully realize the benefits.



sd3,line drawing
clip score: 0.1976318359375
LAION score: 5.3740434646606445
ae2.5 score: 5.28125
HPSV2 score: 0.1920166015625

sd3,kid crayon drawing
clip score: 0.2110595703125
LAION score: 5.615420341491699
ae2.5 score: 4.4375
HPSV2 score: 0.2430419921875

sd3,wooden sculpture
clip score: 0.20166015625
LAION score: 5.823437690734863
ae2.5 score: 4.78125
HPSV2 score: 0.25048828125

sd3,melting golden 3d rendering
clip score: 0.2174072265625
LAION score: 5.88104248046875
ae2.5 score: 4.875
HPSV2 score: 0.27392578125

sd3,neon graffiti
clip score: 0.216064453125
LAION score: 5.773496627807617
ae2.5 score: 5.0
HPSV2 score: 0.2763671875

sd3,cyberpunk art
clip score: 0.21875
LAION score: 6.941895484924316
ae2.5 score: 5.8125
HPSV2 score: 0.2486572265625

sd3,Hawaiian sunset painting
clip score: 0.224853515625
LAION score: 6.746219158172607
ae2.5 score: 5.8125
HPSV2 score: 0.2783203125

sd3,papercut art
clip score: 0.2139892578125
LAION score: 5.3666815757751465
ae2.5 score: 4.96875
HPSV2 score: 0.26953125

sd3,doodle art
clip score: 0.2259521484375
LAION score: 6.10030460357666
ae2.5 score: 5.3125
HPSV2 score: 0.2476806640625

sd3,dreamy watercolor
clip score: 0.22509765625
LAION score: 6.76975154876709
ae2.5 score: 5.75
HPSV2 score: 0.269287109375

creatilayout,line drawing
clip score: 0.2071533203125
LAION score: 5.145726203918457
ae2.5 score: 4.09375
HPSV2 score: 0.2186279296875

creatilayout,kid crayon drawing
clip score: 0.234619140625
LAION score: 4.86838436126709
ae2.5 score: 3.75
HPSV2 score: 0.212890625

creatilayout,wooden sculpture
clip score: 0.224365234375
LAION score: 6.0381317138671875
ae2.5 score: 5.125
HPSV2 score: 0.2305908203125

creatilayout,melting golden 3d rendering
clip score: 0.2281494140625
LAION score: 5.7674150466918945
ae2.5 score: 5.34375
HPSV2 score: 0.276123046875

creatilayout,neon graffiti
clip score: 0.2169189453125
LAION score: 5.736427307128906
ae2.5 score: 4.5625
HPSV2 score: 0.250244140625

creatilayout,cyberpunk art
clip score: 0.2279052734375
LAION score: 5.777075290679932
ae2.5 score: 4.875
HPSV2 score: 0.241943359375

creatilayout,Hawaiian sunset painting
clip score: 0.2396240234375
LAION score: 5.805807113647461
ae2.5 score: 5.1875
HPSV2 score: 0.25

creatilayout,papercut art
clip score: 0.2283935546875
LAION score: 5.400541305541992
ae2.5 score: 4.78125
HPSV2 score: 0.231201171875

creatilayout,doodle art
clip score: 0.2198486328125
LAION score: 5.171780586242676
ae2.5 score: 4.34375
HPSV2 score: 0.2122802734375

creatilayout,dreamy watercolor
clip score: 0.21875
LAION score: 6.132205009460449
ae2.5 score: 5.15625
HPSV2 score: 0.2432861328125

ours,line drawing
clip score: 0.1993408203125
LAION score: 5.558525085449219
ae2.5 score: 4.71875
HPSV2 score: 0.2159423828125

ours,kid crayon drawing
clip score: 0.2169189453125
LAION score: 5.193413734436035
ae2.5 score: 4.5625
HPSV2 score: 0.269775390625

ours,wooden sculpture
clip score: 0.2100830078125
LAION score: 6.177881240844727
ae2.5 score: 5.25
HPSV2 score: 0.300048828125

ours,melting golden 3d rendering
clip score: 0.23193359375
LAION score: 6.307491302490234
ae2.5 score: 5.46875
HPSV2 score: 0.31201171875

ours,neon graffiti
clip score: 0.2218017578125
LAION score: 6.061593532562256
ae2.5 score: 5.21875
HPSV2 score: 0.282470703125

ours,cyberpunk art
clip score: 0.2213134765625
LAION score: 6.5392913818359375
ae2.5 score: 6.5
HPSV2 score: 0.29931640625

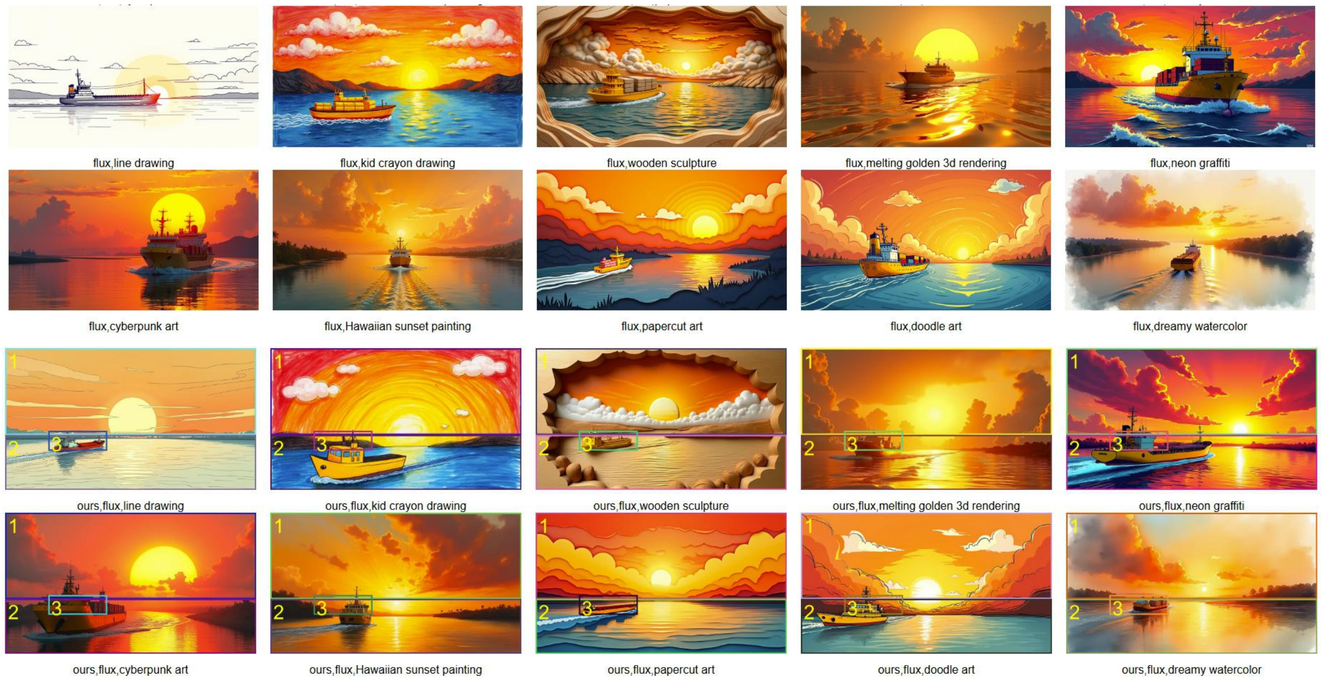
ours,Hawaiian sunset painting
clip score: 0.21826171875
LAION score: 6.159914016723633
ae2.5 score: 6.03125
HPSV2 score: 0.289794921875

ours,papercut art
clip score: 0.2352294921875
LAION score: 5.751713752746582
ae2.5 score: 5.1875
HPSV2 score: 0.250732421875

ours,doodle art
clip score: 0.2232666015625
LAION score: 5.828454494476318
ae2.5 score: 5.40625
HPSV2 score: 0.26513671875

ours,dreamy watercolor
clip score: 0.2200927734375
LAION score: 5.814462184906006
ae2.5 score: 5.03125
HPSV2 score: 0.281005859375

Figure 5. Qualitative comparison results of SD3, SiamLayout, and Ours-SD3 on Style-Benchmark.


flux,line drawing

flux,kid crayon drawing

flux,wooden sculpture

flux,melting golden 3d rendering

flux,neon graffiti

flux,cyberpunk art

flux,Hawaiian sunset painting

flux,papercut art

flux,doodle art

flux,dreamy watercolor

ours,flux,line drawing

ours,flux,kid crayon drawing

ours,flux,wooden sculpture

ours,flux,melting golden 3d rendering

ours,flux,neon graffiti

ours,flux,cyberpunk art

ours,flux,Hawaiian sunset painting

ours,flux,papercut art

ours,flux,doodle art

ours,flux,dreamy watercolor

Figure 6. Qualitative comparison results of FLUX, and Ours-FLUX on .

Image	Detailed Regional Prompts
Row 1, Col1	Region#1: Lush green rice plants, traditional harvesting activity. Region#2: A blurred image of a house with a red-tiled roof, surrounded by greenery. Region#3: A person wearing a conical hat, harvesting crops in a green field. Region#4: A person in a conical hat working in lush green rice paddies.
Row 1, Col2	Region#1: A gray British Shorthair standing on a rock in the woods. Region#2:: A yellow American robin standing on the rock. Region#3: A brown Maltipoo dog standing on the rock. Region#4: A close up of a small waterfall in the woods.
Row 1, Col3	Region#1: A male customer in casual attire, engaged on a phone call. Region#2: A man with a bag over his shoulder, wearing casual attire and standing in an outdoor setting. Region#3: City street with souvenir stand, pedestrians, buildings.
Row 1, Col4	Region#1: Small, black and tan dog wearing pink polka-dotted coat, blue harness with red trim, and a leash. Region#2: Person in pink polka dot hoodie and blue jeans seated in wheelchair. Region#3: Person pushing wheelchair with child, carrying backpack. Region#4: Person in wheelchair with polka dot shirt, jeans, and sneakers. Region#5: Green, well-maintained bushes lining the pathway.
Row 1, Col5	Region#1: Clear blue water with gentle ripples, surrounded by a rocky shoreline and green vegetation. Region#2: Large, lush green grass area with people and structures. Region#3: The main subject is a traditional Russian wooden church characterized by its intricate design, multiple onion domes, and the use of wood as the primary building material. This architectural style reflects historical Russian religious structures and cultural heritage.
Row 2, Col1	Region#1: A worker in blue uniform and cap, carrying a yellow hose. Region#2: A railway worker cleaning a train platform. Region#3: A red train is being cleaned by workers in blue uniforms.”
Row 2, Col2	Region#1: Person wearing red jacket, facing away from camera, with mountainous backdrop. Region#2: A clear, cloud-speckled blue sky. Region#3: Rusted corrugated metal with bird perched on beam. Region#4: A close-up of a white urinal with water marks and small debris on the surface, set against a plain background. Region#5: White ceramic urinal with silver flush mechanism, minor stains and marks. Region#6: A restroom with a person using a urinal, featuring an expansive view of mountains through a large window. Region#7: A range of mountains with snow-capped peaks and rugged terrain.
Row 2, Col3	Region#1: Large, ornate classical-style building with columns and statues. Region#2: People dining and socializing at an outdoor cafe. Region#3: People enjoying food and drinks at an outdoor cafe.
Row 2, Col4	Region#1: A contemporary, illuminated structure with reflective glass and prominent branding. Region#2: Reflective, dark water surface illuminated by lights. Region#3: Urban landscape with illuminated buildings and modern architecture.
Row 2, Col5	Region#1: A modern, illuminated suspension bridge at night with streetlights and fireworks in the background. Region#2: A calm river reflecting lights from nearby sources, creating a colorful and serene nighttime scene. Region#3: Cityscape at night with illuminated buildings and fireworks.

Table 5. Detailed regional prompts for the generated images shown in Figure 1.

Image	Detailed Regional Prompts
Col1	Region#1: A contemporary structure with a flat roof, large windows, and an open design. Region#2:: Spacious stone-paved area with concrete barriers and landscaped sections. Region#3: A large, rugged mountain range with varying vegetation and a clear sky above. Region#4: A low stone wall with decorative rocks and plants, bordered by a paved area.
Col2	Region#1: Calm sea with gentle ripples and a distant sailboat. Region#2: A panoramic view of a bustling urban skyline with various architectural styles, under a clear sky. Region#3: A cityscape with a mix of historic and modern buildings, under a clear sky.
Col3	Region#1: Clear blue water with gentle ripples, surrounded by a rocky shoreline and green vegetation. Region#2: Large, lush green grass area with people and structures. Region#3: The main subject is a traditional Russian wooden church characterized by its intricate design, multiple onion domes, and the use of wood as the primary building material. This architectural style reflects historical Russian religious structures and cultural heritage.
Col4	Region#1: A contemporary, illuminated structure with reflective glass and prominent branding. Region#2: Reflective, dark water surface illuminated by lights. Region#3: Urban landscape with illuminated buildings and modern architecture.
Col5	Region#1: A pedestrian in a blue coat and black boots walks through snowy conditions, carrying a bag. Region#2: City street illuminated with festive lights, decorated for winter holidays. Region#3: Illuminated blue star-shaped lights hanging above a city street.

Table 6. Detailed regional prompts for the generated images shown in Figure 5.

Image	Detailed Regional Prompts
Col1	Region#1: A blurred close-up of a firearm with a scope, held by an individual in partial view. Region#2:: A white armored figure, poised with a weapon in hand. Region#3: White stormtrooper helmet with black visor, blue mouthpiece, and detailed facial markings. Region#4: White armored figures with black detailing, one holding a weapon. Region#5: Imperial Stormtroopers, iconic white armor soldiers from the Star Wars franchise.
Col2	Region#1: A majestic, weathered bronze statue of a horse with ornate details and a rider atop. Region#2: A tall, pointed white structure with horizontal bands and a spherical finial at the peak against a blue sky. Region#3: Bronze equestrian statue atop ornate pedestal with lions, historical figures depicted. Region#4: The main subject in the input image is a white building, which appears to be an architectural structure with historical significance. It features intricate carvings and golden circular medallions, flanked by lion statues on either side. The building's design suggests it may serve as a monument or landmark, possibly used for ceremonial purposes or as part of a larger complex like a palace or museum.
Col3	Region#1: Large, ornate classical-style building with columns and statues. Region#2: People dining and socializing at an outdoor cafe. Region#3: People enjoying food and drinks at an outdoor cafe.
Col4	Region#1: A young tiger with distinctive stripes being petted. Region#2: Human hand with visible skin texture and white soap suds. Region#3: A blue collar worn by a sleeping cat. Region#4: The main subject is a wet tiger with visible fur patterns, being attended to by a person.
Col5	Region#1: Colorful buildings, outdoor market, cobblestone street. Region#2: Various clothing items displayed on racks in a store. Region#3: Leafless tree branches against a cloudy sky. Region#4: Clothes hanging on a clothesline against a yellow wall.
Col6	Region#1: Sunset with vibrant orange and red hues in the sky. Region#2: Large yellow and blue barge moving on calm river at sunset. Region#3: A large, yellow cargo ship is cruising on the water during sunset with a backdrop of distant hills.
Col7	Region#1: White lines demarcate parking spaces in an organized lot. Region#2: Overgrown wooden house with extensive vine coverage. Region#3: Parking lot with white lines, green bollards, and a bicycle. Region#4: A two-story building with white walls, multiple windows, and a balcony on the upper floor.
Col8	Region#1: Stone archway leading to quaint European street. Region#2: Vintage wall-mounted street lamp with a glass shade, black metal frame, and hanging mechanism. Region#3: A classic, ornate street lamp mounted on a building with green shutters.
Col9	Region#1: A serene river reflecting vibrant city lights at night. Region#2: Historic European castle complex, illuminated at night with vibrant lighting. Region#3: City illuminated with vibrant lights, showcasing historical architecture and reflections on water. Region#4: A deep blue sky serves as a backdrop for the illuminated castle.

Table 7. Detailed regional prompts for the generated images shown in Figure 8.

References

[1] Anthony Chen, Jianjin Xu, Wenzhao Zheng, Gaole Dai, Yida Wang, Renrui Zhang, Haofan Wang, and Shanghang Zhang. Training-free regional prompting for diffusion transformers. *arXiv preprint arXiv:2411.02395*, 2024. 4

[2] Xiaoliang Dai, Ji Hou, Chih-Yao Ma, Sam Tsai, Jialiang Wang, Rui Wang, Peizhao Zhang, Simon Vandenhende, Xiaofang Wang, Abhimanyu Dubey, et al. Emu: Enhancing image generation models using photogenic needles in a haystack. *arXiv preprint arXiv:2309.15807*, 2023. 4

[3] Yifan Pu, Yiming Zhao, Zhicong Tang, Ruihong Yin, Haoxing Ye, Yuhui Yuan, Dong Chen, Jianmin Bao, Sirui Zhang, Yanbin Wang, Lin Liang, Lijuan Wang, Ji Li, Xiu Li, Zhouhui Lian, Gao Huang, and Baining Guo. Art: Anonymous region transformer for variable multi-layer transparent image generation. In *CVPR*, 2025. 3

[4] John Shi. Why flux lora so hard to train and how to overcome it?, 2024. 3

[5] Binbin Yang, Yi Luo, Ziliang Chen, Guangrun Wang, Xiaodan Liang, and Liang Lin. Law-diffusion: Complex scene generation by diffusion with layouts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22669–22679, 2023. 1

[6] Hui Zhang, Dexiang Hong, Tingwei Gao, Yitong Wang, Jie Shao, Xinglong Wu, Zuxuan Wu, and Yu-Gang Jiang. Creatilayout: Siamese multimodal diffusion transformer for creative layout-to-image generation. *arXiv preprint arXiv:2412.03859*, 2024. 3

[7] Guangcong Zheng, Xianpan Zhou, Xuewei Li, Zhongang Qi, Ying Shan, and Xi Li. Layoutdiffusion: Controllable diffusion model for layout-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22490–22499, 2023. 1