

000 **FORMATTING INSTRUCTIONS FOR ICLR 2026** 001 **CONFERENCE SUBMISSIONS** 002 **– *Supplementary Material* –**

006 **Anonymous authors**

007 Paper under double-blind review

010 **CONTENTS OF APPENDIX**

013 A Overview of Appendix	2
015 B Use of Large Language Models (LLMs)	2
017 C Limitations	2
019 D Detailed Analysis of Probing Results	3
020 D.1 Baseline Models and Implementation Details	3
021 D.2 Constructed Dataset for Probing Analysis	3
022 D.3 Amount and CLIP Score Analysis for All Models	4
023 D.4 Details of Evaluation for Probing Analysis	4
026 E Detailed Setup of Evaluation Results	6
027 E.1 Inference Details	6
028 E.2 Evaluation Benchmarks	6
029 E.3 Token-level Enhancement Mask Localization	6
030 E.4 Editing details	6
031 E.5 Acceleration Details	7
034 F More Quantitative Results	9
035 F.1 Analysis on the selection of enhanced blocks	9
036 F.2 More Ablation Studies on Scale Schemes	9
037 F.3 Original Data for Enhancing Scale and Block Ablation	10
038 F.4 More Results about Token-level Enhancement	10
039 F.5 Detailed Acceleration Results	11
042 G More Qualitative Results	12
043 G.1 More SD3.5 Results	12
044 G.2 More FLUX Results	13
045 G.3 More Qwen Image Results	14
046 G.4 More Editing Results	15
047 G.5 Failure Cases	16
048 G.6 All-block Showcases of Probing Analysis	17
051 H Human Evaluation Details	20

A OVERVIEW OF APPENDIX

The appendix provides supplementary information and results supporting the main text. It begins with a discussion on the use of large language models (LLMs) and Limitations. This is followed by a detailed analysis of probing results (Sec. D), including descriptions of baseline models and implementation details, datasets used for probing, quantitative metrics such as amount and CLIP score analysis, and evaluation procedures. Next, the appendix (Sec. E) presents the detailed setup of evaluation results, covering inference configurations, evaluation benchmarks, token-level enhancement mask localization, editing task details, and acceleration experiments. Additional quantitative results (Sec. F) are provided, including analyses on enhanced block selection, scaling schemes, ablation studies, and token-level enhancements. More qualitative results (Sec. G) show visual examples across different models and tasks, such as generation, editing, and model-specific outputs. The appendix also discusses failure cases and limitations observed during experiments. Last, the human evaluation protocol used to assess output quality and alignment (Sec. H). Together, these sections offer a comprehensive resource for reproducing experiments, understanding model behavior, and exploring additional findings beyond the main text.

B USE OF LARGE LANGUAGE MODELS (LLMs)

We made limited use of Large Language Models (LLMs) in preparing this work, specifically for the following purposes. (1). Writing assistance. The LLMs were used to polish grammar and phrasing in parts of the paper. But all substantive content, ideas, and claims were written by the authors. (2). Prompt generation for probing analysis and editing prompts. GPT-5¹ was used to generate an initial set of candidate prompts for probing analysis and editing prompts. The final prompts were manually screened, refined, and verified by the authors before use. All research contributions, analyses, and conclusions are solely the responsibility of the authors.

C LIMITATIONS

Although our proposed method delivers substantial performance gains across multiple tasks, certain limitations remain.

Dependence on the preliminary block-wise analysis. A core part of our enhancement pipeline relies on an initial, automatic analysis that identifies block-wise interactions inside the model. Luckily, our analysis is totally automatic and could be performed across various models with different numbers of blocks, varied model sizes, as well as different models (e.g., SD3.5, FLUX, Qwen Image).

Limited fidelity on very complex prompts. For extremely complex or highly detailed prompts, our method may fail to capture all fine-grained elements. This limitation mainly stems from the pretraining data distribution, where rare object combinations or subtle high-frequency details are underrepresented, sometimes resulting in missing elements or artifacts in the output. As could also be observed from the failure cases in Fig. A9.

Evaluation and generalization limits. Our evaluation on standard benchmarks may not fully reflect perceptual quality or robustness, and gains might not generalize to niche domains or prompts requiring world knowledge absent from pretraining.

¹<https://chatgpt.com>

D DETAILED ANALYSIS OF PROBING RESULTS

D.1 BASELINE MODELS AND IMPLEMENTATION DETAILS

We implement our probing analysis based on the widely adopted MMDiT-based text-to-image models, including:

- Stable Diffusion 3.5-Large² (Stability-AI, 2024): A latent diffusion model with approximately 8 billion parameters, based on the Multimodal Diffusion Transformer (MMDiT) architecture. It demonstrates strong performance in prompt adherence, typography, and supports a mature ecosystem of extensions.
- FLUX.1-Dev³ (Labs, 2024): A 12B-parameter rectified-flow transformer model that adopts advanced training techniques and a substantially larger dataset to enhance visual fidelity. It has attracted significant community attention for its improvements in prompt alignment, detail rendering, and efficient sampling.
- Qwen Image⁴ (Wu et al., 2025): A 20B-parameter MMDiT model developed within the Qwen series, designed for robust multimodal reasoning and high-quality image synthesis. It is particularly noted for its strong performance in complex text rendering (especially Chinese) and text-guided image editing.

We use the official checkpoints provided by the authors and the *diffusers* library (Team, 2025) for implementation. During inference, model weights are loaded in 16-bit precision. No acceleration techniques such as xformers or memory-efficient attention are used. The default parameters during inference are summarized in Tab. A1.

Table A1: Model information and default parameters during inference.

Models	SD3.5-large	FLUX.1-Dev	Qwen Image
MMDiT Blocks	[0,37]	[0,57]	[0,59]
Parameters	8B	12B	20B
Inference Steps	28	28	50
CFG Scale	7.0	3.5	4.0
Size	(1024,1024)	(1024,1024)	(1024,1024)

During probing analysis, we use identical hyperparameters for all models to ensure fair comparison, as summarized in Tab. A1. For each model with N blocks, we fix a random seed and generate one baseline image, N *disable* images, N *remove* images, and N *enhance* images—constituting one experimental group. For each of the three MMDiT-based models, we conduct five experimental groups using five different random seeds (0, 42, 329, 1234, 99514). Final results are reported as the average across these five groups.

D.2 CONSTRUCTED DATASET FOR PROBING ANALYSIS

We construct a challenging prompt dataset with GPT-5, comprising 333 diverse and complex prompts across three attributes: color (129 prompts), spatial relationship (104 prompts), and amount (100 prompts). For color attributes, we focus on objects with distinctive colors (e.g., “red apple”, “yellow banana”). For spatial relationships, we include prompts describing eight positional relations (e.g., “left”, “right”, “above”, “below”, “upper left”, “upper right”, “lower left”, “lower right”). For amount attributes, we cover a range of quantities from “three” to “nine”. We also ensure diversity in object categories, including human, animal, natural scenes, indoor scenes, food, clothing & accessories, vehicles, and so on.

The distribution of prompts across these attributes is illustrated in Fig. A1. We ensure that the prompts are diverse and challenging, covering various object categories, colors, spatial relations, and quantities.

²<https://huggingface.co/stabilityai/stable-diffusion-3.5-large>

³<https://huggingface.co/black-forest-labs/FLUX.1-dev>

⁴<https://huggingface.co/Qwen/Qwen-Image>

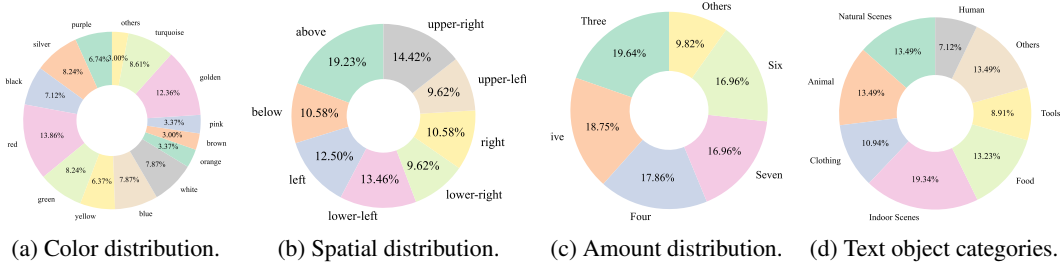


Figure A1: Statistics of our constructed datasets for probing analysis. Each subfigure presents the distribution of prompts for a specific attribute: color, spatial relation, amount, and object category.

D.3 AMOUNT AND CLIP SCORE ANALYSIS FOR ALL MODELS

Due to space limitations, the main paper only presents the block-wise analysis results for the color and spatial relationship attributes on three models under the removing, disabling, and enhancing strategies, as well as the corresponding changes in DINOv2 similarity relative to the baseline. Here, we provide additional results, including the block-wise analysis for the amount attribute on all three models, and the overall CLIP score trends under the removing, disabling, and enhancing strategies for all models. The detailed results are shown in Fig. A2. Notably, the amount attribute exhibits a different sensitivity to the number of blocks compared to color and spatial attributes, and enhancement does not lead to a clear improvement, which may be attributed to the inherent limitations of MMDiT models in understanding quantity. The overall CLIP score trends are consistent with those of DINOv2 similarity, further validating the effectiveness and stability of our method.

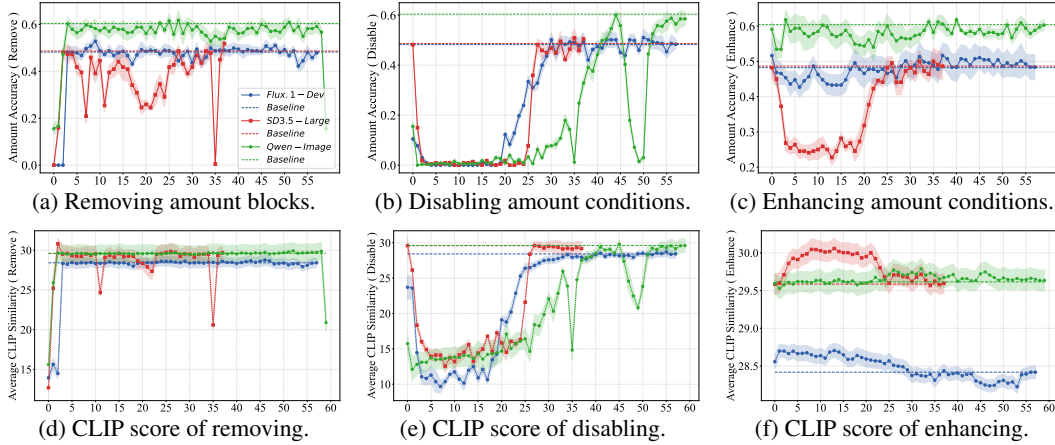


Figure A2: More detailed probing results on three MMDiT models: Stable Diffusion 3.5-Large, FLUX.1-Dev, and Qwen Image. Each subfigure presents the performance curves for a specific attribute (amount or overall CLIP score) under different probing strategies (removing, disabling, enhancing).

D.4 DETAILS OF EVALUATION FOR PROBING ANALYSIS

In probing analysis, we use the open-sourced Qwen2.5-VL 72B⁵ (Bai et al., 2025) model for color and spatial relationship evaluation. We design specific systematic prompts to guide the model in accurately assessing whether the generated images align with the intended attributes in the text prompts. The detailed prompts for color and spatial relationship evaluation are provided in following colored boxes.

⁵<https://github.com/QwenLM/Qwen2.5-VL>

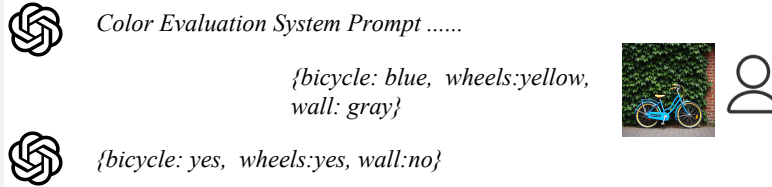
Color Evaluation Prompt

You are given an image, its caption, and a set of objects with their expected colors. Your task: 1).For each object: check if the color in the caption matches the actual color in the image. 2).If the color matches, return "Yes". If the color does not match or the object is not visible, return "No".

Rules:

- Output ONLY a single valid JSON object.
- The JSON keys must be exactly the provided object names.
- The values must be strictly "Yes" or "No".
- Do not generate any other words.
- Do not add explanations, extra text, or formatting outside the JSON.

Example:



Spatial Evaluation Prompt

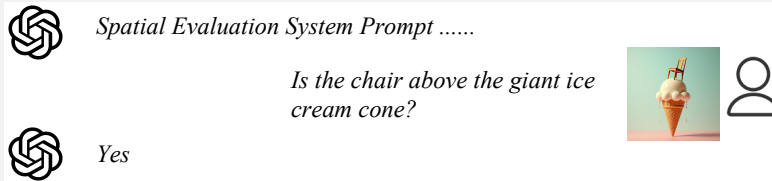
You are given an image, its caption, and a question about the spatial relationship between two objects in the image.

Your task: 1).Check whether the spatial relationship described in the question can be confirmed from the image.2).If the relationship is clearly visible and correct, return "Yes".3).If the relationship is not correct, cannot be seen, or the objects are unclear, return "No".

Rules:

- Output ONLY a single string. The value must be strictly "Yes" or "No".
- Do not generate any other words.
- Do not add explanations, extra text, or formatting outside the answer.

Example:



For amount evaluation, we use CountGD⁶ (Amini-Naieni et al., 2024), an open-world object counting model based on Grounding-DINO (Liu et al., 2024). The detection confidence threshold is set to 0.5 for higher precision. Counting accuracy is computed as the proportion of images where the predicted count exactly matches the ground-truth specified in the prompt.

Statistical Significance. We acknowledge that both the LLM-based evaluation and CountGD have inherent limitations. LLMs may misinterpret visual details or be affected by biases in their training data, while CountGD may produce inaccurate counts for small, overlapping, or occluded objects. To address these issues, we conduct multiple experimental runs with different random seeds and report averaged results, thereby reducing the impact of individual evaluation errors. To further ensure the validity and robustness of our conclusions, we additionally employ alternative evaluation methods that are independent of the primary approaches. This cross-validation helps to mitigate the influence of dataset bias and evaluation inaccuracies on our experimental findings.

⁶<https://github.com/niki-amini-naeni/CountGD>

E DETAILED SETUP OF EVALUATION RESULTS

E.1 INFERENCE DETAILS

During evaluation of text-to-image generation, editing, and acceleration, we use the same hyperparameters in Tab. A1. The enhancement strength λ is set to 1.5 by default. For the enhanced blocks, we select the block index in Tab. 1.

For all attributes except amount, we adopt a sentence-level approach, which has already demonstrated strong performance. For token-level enhancement, we construct the enhancement mask M by passing the target phrases (e.g., “two apples”, “a person”, “seven”).

E.2 EVALUATION BENCHMARKS

For text-to-image generation, we evaluate our method on the widely used T2I-CompBench++ (Huang et al., 2025) and GenEval (Ghosh et al., 2023) benchmarks. T2I-CompBench++ contains 8,000 compositional prompts spanning color, spatial, 3D spatial, shape, texture, non-spatial relations, numeracy, and complex attributes. It extends the original benchmark (Huang et al., 2023) and introduces more challenging tasks (e.g., 3D spatial and numerical compositionality). We use all prompts and generate one sample per prompt. GenEval consists of 553 structured prompts targeting single-object, two-object, counting, color, position, and color-attribute binding. Each prompt is paired with four generated samples, and performance is computed with an automatic evaluation pipeline based on object detection, counting, and attribute classification, providing interpretable error types (e.g., missing objects, incorrect color, or miscount).

We evaluate image quality and text-image alignment using LAION Aesthetic v2 (Schuhmann, 2022) and HPSv2 (Wu et al., 2023). LAION Aesthetic v2 measures visual appeal, while HPSv2 evaluates prompt-image alignment relative to human judgments. As these metrics capture different aspects, we report both and supplement them with task-specific evaluations and human studies to ensure a comprehensive assessment.

E.3 TOKEN-LEVEL ENHANCEMENT MASK LOCALIZATION

During inference, we record the multi-head self-attention of the concatenated features Z_{in} at each MMDiT block and denoising step. To obtain a stable token-region corresponding, we aggregate the attention maps across all heads and denoising steps. Eventually, we get attention maps of shape $[N, H \times W, T]$, where N is the number of MMDiT blocks, $H \times W$ is the spatial dimension of the image features, and T is the number of text tokens. For visualization, we average all the MMDiT blocks’ attention maps to get a single attention map of shape $[H \times W, T]$. We then normalize the attention maps along the spatial dimension and then resize them to the original image size. The visualization results are shown in Figure A3. We can see that ‘dog’ and ‘cat’ tokens have high attention values in the corresponding image regions, indicating that the token-level enhancement can effectively target specific areas in the image.

Based on the above analysis, we first tokenize the input prompt \mathcal{P} using the same tokenizer as the MMDiT model, obtaining a sequence of token IDs $\mathbf{P} = [p_1, p_2, \dots, p_N]$. Given a target phrase \mathcal{Q} (e.g., “brown”, “firetruck”), we tokenize it as $\mathbf{Q} = [q_1, q_2, \dots, q_M]$. We then search for all subsequences in \mathbf{P} that match \mathbf{Q} . The starting indices of these matches are collected in the set $\mathcal{I} = \{i \mid (p_i, p_{i+1}, \dots, p_{i+M-1}) = (q_1, q_2, \dots, q_M)\}$. The mask M is constructed with the following rule:

$$M_j = \begin{cases} 1, & \text{if } \mathcal{I} = \emptyset \text{ or } \exists (i, i+M-1) \in \mathcal{I} \text{ with } j \in [i, i+M-1], \\ 0, & \text{otherwise.} \end{cases}$$

Here, M_j indicates whether the j -th token in the prompt should be enhanced. If not matched, the token-level enhancement defaults to sentence-level enhancement by setting all entries of M to 1.

E.4 EDITING DETAILS

Stable Flow (Avrahami et al., 2025) is adopted as the baseline, which selects vital blocks based on the perceptual similarity. However, its block selection is not task-specific and may be inaccurate for

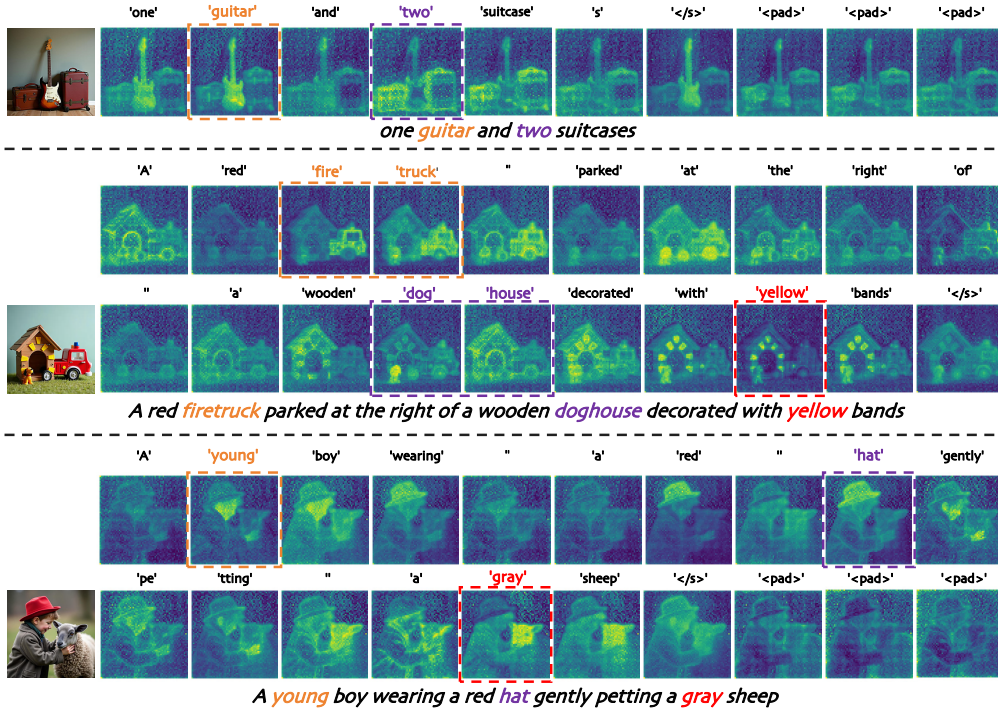


Figure A3: Visualization of token-level attention maps. Colored boxes indicate the target tokens in the prompt. The attention maps highlight the corresponding text tokens' indices in the image.

fine-grained editing, whereas our approach leverages probing analysis to identify blocks tailored to the editing task. Prior benchmarks lack coverage of quantity, attribute binding, and spatial relationship editing. To address this, we construct a new editing benchmark comprising 1,000 images and corresponding editing prompts. Each source prompt is paired with four target prompts, covering object addition (with varying colors), background changes, color and lighting adjustments, shape and direction modifications, positional changes, quantity variations, and object actions. Evaluation is performed for each image-prompt pair using CLIP-image similarity and CLIP score to assess image quality and prompt adherence. Human evaluation details are provided in Section H.

E.5 ACCELERATION DETAILS

Our method is conceptually related to TeaCache (Liu et al., 2025), which accelerates video diffusion by using timestep embeddings to estimate output differences and cache intermediate results selectively. In contrast, we skip blocks deemed irrelevant for the current editing task based on probing analysis. While TeaCache reduces redundancy across timesteps, our approach reduces computation across feature blocks, enabling acceleration without affecting editing quality.

For acceleration, we skip one-third of the CFG steps and remove three MMDDiT blocks. Applying our method to both the FLUX baseline and TeaCache demonstrates significant speedup while maintaining comparable image quality. In each experiment, we randomly select 400 prompts from T2I-CompBench++ and generate one sample per prompt. The CFG steps skipped are [5, 10, 15, 20, 25, 30, 35, 40, 45, 50], and the removed MMDDiT blocks are [30, 40, 50]. We repeat the experiments on both NVIDIA 4090 and H100 GPUs to verify the stability and robustness of our approach.

We present additional examples of accelerated generation in Fig. A4. The results in Tab.5 and Fig. A4 demonstrate that our method can seamlessly integrate with TeaCache, achieving significant speedup while maintaining high-quality generation.

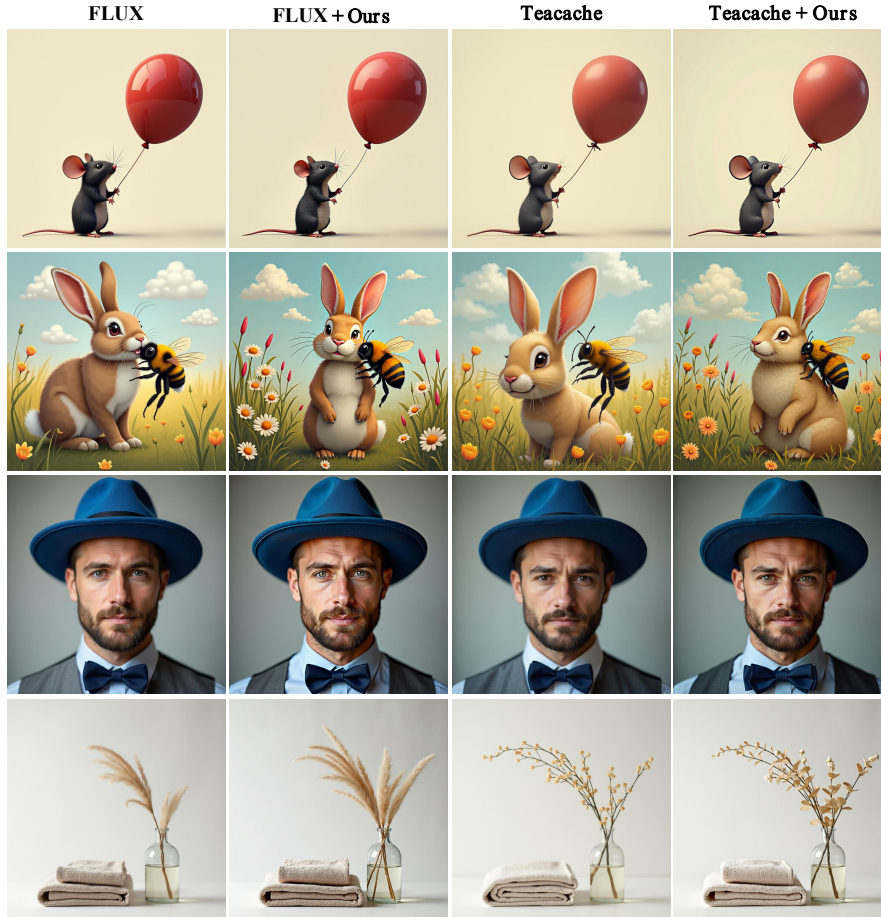


Figure A4: Examples of accelerated generation. Our method removes certain blocks, which may lead to different details in the generated image, yet the overall synthesis quality and textual semantics are consistent with the user prompts. The four prompts are (1) “a balloon on the right of a mouse”, (2) “a rabbit hidden by a bee”, (3) “a man in a blue and blur hat with a gray shirt and bowtie”, (4) “a fabric towel and a glass vase”.

F MORE QUANTITATIVE RESULTS

F.1 ANALYSIS ON THE SELECTION OF ENHANCED BLOCKS

In the main paper, Tab. 6 only shows the results of color, shape, and spatial of different enhanced blocks and different strengthening scales. Here we provide the full results of T2I-CompBench++ in Tab. A2.

For the weakening experiments, we keep the same number and position of the enhanced blocks in the Tab. 1. All the metrics drop compared to the baseline and the 0.7 scale, even worse than the 0.9 scale, indicating that these vital blocks are indeed important for compositional generation. We also try to randomly select five blocks and all blocks to strengthen. Enhancing random five blocks can also improve the performance, but not as good as our selected blocks. This further validates the effectiveness of our probing analysis. Enhancing all blocks leads to a significant performance drop, which may be due to the excessive enhancement that distorts the original feature distribution and harms the generation quality.

Table A2: Full results of small $\lambda(l)$ enhancement and block selection experiments. All the experiments are conducted with FLUX on T2I-CompBench++.

Methods	Attribute Binding			Object Relationship			Amount	Complex
	Color	Shape	Texture	2D Spatial	3D Spatial	Non-Spatial		
FLUX	0.7322	0.4908	0.6490	0.2935	0.3739	0.3044	0.5877	0.3597
0.7	0.3161	0.2653	0.2707	0.1100	0.1831	0.2890	0.3944	0.2630
0.9	0.6891	0.4395	0.5622	0.2611	0.3247	0.3029	0.5482	0.3419
Random 5 blocks	0.7624	0.5072	0.6492	0.3119	0.3797	0.3042	0.5842	0.3605
All blocks	0.2360	0.2736	0.2581	0.0495	0.1522	0.2928	0.2448	0.2233
Ours	0.7804	0.5482	0.6980	0.3280	0.3900	0.3054	0.5860	0.3691

F.2 MORE ABLATION STUDIES ON SCALE SCHEMES

We study the impact of different enhancement scale schemes $\lambda(l)$, including uniform ($U(1.2, 1.8)$, $U(1.8, 1.2)$), exponential ($Exp(1.6, 0.95)$), and fixed (1.5) scales. As shown in Tab. A3, all schemes improve over the baseline, confirming the robustness of our selected enhancing blocks. The fixed scale is notably simple and achieves balanced performance across all dimensions, making it practical for general use. In contrast, other schemes have their own strengths: linearly increasing scales favor spatial performance, while decaying scales (e.g., $U(1.8, 1.2)$) further boost fine-grained attributes like color and shape. These findings indicate that both the magnitude and distribution of enhancement strength across blocks are important for compositional generation.

Table A3: Comparison of different scale schemes across blocks on T2I-CompBench++.

Methods	Attribute Binding			Object Relationship			Amount	Complex
	Color	Shape	Texture	2D Spatial	3D Spatial	Non-Spatial		
FLUX	0.7322	0.4908	0.6490	0.2935	0.3739	0.3044	0.5877	0.3597
$U(1.2, 1.8)$	0.7795	0.5438	0.6827	0.3311	0.3865	0.3051	0.5845	0.3685
$U(1.8, 1.2)$	0.8035	0.5551	0.6940	0.3203	0.3893	0.3051	0.5654	0.3722
$Exp(1.6, 0.95)$	0.7783	0.5507	0.7049	0.3296	0.3947	0.3060	0.5763	0.3710
1.5 (Fixed)	0.7804	0.5482	0.6980	0.3280	0.3900	0.3054	0.5860	0.3691

F.3 ORIGINAL DATA FOR ENHANCING SCALE AND BLOCK ABLATION

In Fig.7, we present experiments conducted with a fixed enhancing scale ranging from 1.2 to 2.0, as well as experiments varying the number of enhancing blocks. In the main text, the vertical axes were normalized to emphasize consistent trends across different models and evaluation metrics. Here, we provide the original, unnormalized data for reference (see Tab. A4 and Tab. A5 for details). In addition, we include the corresponding results for SD3.5 under the same enhancing scale settings. It could be observed that the results consistently reflect identical conclusions with the main paper.

Table A4: Results of the enhancing scale experiments. Except the FLUX data presented in the main text, we also provide the corresponding results for SD3.5 under the same experimental settings.

Methods/ $\lambda(l)$		Baseline	1.2	1.3	1.4	1.5	1.6	1.8	2.0
FLUX	color	0.7322	0.7776	0.7768	0.7775	0.7804	0.7708	0.7643	0.7415
	shape	0.4908	0.5389	0.5489	0.5381	0.5482	0.5449	0.5226	0.5093
	spatial	0.6603	0.6827	0.7244	0.6987	0.7340	0.7147	0.7083	0.6795
SD3.5	color	0.7284	0.7992	0.8064	0.8072	0.8052	0.7874	0.7785	0.7608
	shape	0.5592	0.6432	0.6656	0.6642	0.6744	0.6659	0.6255	0.6048
	spatial	0.6418	0.6683	0.7596	0.7716	0.7885	0.7933	0.7740	0.7486

Table A5: Results of experiments varying the number of enhancing blocks on FLUX.

Methods	Baseline	1	3	5	7	9
color	0.7365	0.7533	0.7776	0.7863	0.7613	0.7073
shape	0.4720	0.5159	0.5181	0.5254	0.5201	0.4919
spatial	0.2985	0.3109	0.3152	0.3327	0.2841	0.2581

F.4 MORE RESULTS ABOUT TOKEN-LEVEL ENHANCEMENT

As shown in Table A6, token-level enhancement generally provides more precise guidance compared to sentence-level enhancement, leading to consistent but modest improvements across amount-related attributes. The gains, however, remain limited, which can be attributed to the intrinsic weakness of current diffusion models in numerical reasoning and counting.

Table A6: Token-level vs. sentence-level enhancement on amount-related attributes.

Methods	Amount(T2I-CompBench++)	Count(GenEval)
SD3.5	0.5969	0.6344
ours(sentence)	0.5929	0.6125
ours(token)	0.6088	0.6375
FLUX	0.5877	0.6375
ours(sentence)	0.5860	0.6000
ours(token)	0.6091	0.6438
Qwen Image	0.7406	0.8562
ours(sentence)	0.7359	0.8275
ours(token)	0.7616	0.8594

F.5 DETAILED ACCELERATION RESULTS

In the main paper Tab.5, we evaluate the results of our method for inference acceleration by removing less critical blocks. Here we provide more detailed results in Tab. A7, including the time cost on both NVIDIA 4090 and H100 GPUs, and the image quality metrics (HPSV2, LaionAesthetic V2, CLIP-Text) on different models. We test two baseline models: the original FLUX.1-Dev and the TeaCache-optimized version. For each baseline, we apply our method with different CFG step skipping and block removal strategies.

The CFG steps skipped are $seq(5, 50, 10)$, $seq(5, 55, 5)$, and $seq(6, 58, 3)$, and the removed MMDiT blocks are $[30, 40, 50]$, where $seq(a, b, c)$ denotes the arithmetic sequence starting from a to b with step c . We can see that skipping more CFG steps leads to faster inference without significantly affecting image quality. We finally choose to skip CFG $seq(6, 58, 3)$ steps and remove $[30, 40, 50]$ blocks as the default setting for a good trade-off between speed and quality.

Table A7: Details of acceleration

Method	Time(4090)↓	Time(H100)↓	HPSV2↑	LaionAes V2↑	CLIP-Text↑
FLUX	36.7889	13.0876	29.0533	6.1903	26.9986
skip CFG $seq(5, 50, 10)$	35.5859	12.6414	29.0395	6.2081	26.9761
skip CFG $seq(5, 55, 5)$	33.6433	11.9846	28.7259	6.1340	26.8841
skip CFG $seq(6, 58, 3)$	31.6931	11.3010	28.8408	6.1034	26.7460
Ours	33.3387	11.8734	28.9212	6.1874	26.7807
Teacache	26.6187	9.6125	28.8951	6.2067	26.8346
skip CFG $seq(5, 50, 10)$	26.1385	9.4414	28.9054	6.1983	26.8646
skip CFG $seq(5, 55, 5)$	25.2994	9.1565	28.8822	6.1973	26.8984
skip CFG $seq(6, 58, 3)$	24.5276	8.8804	28.8647	6.1801	26.7946
Ours	24.9992	9.0743	28.9481	6.2256	26.7722

G MORE QUALITATIVE RESULTS

G.1 MORE SD3.5 RESULTS

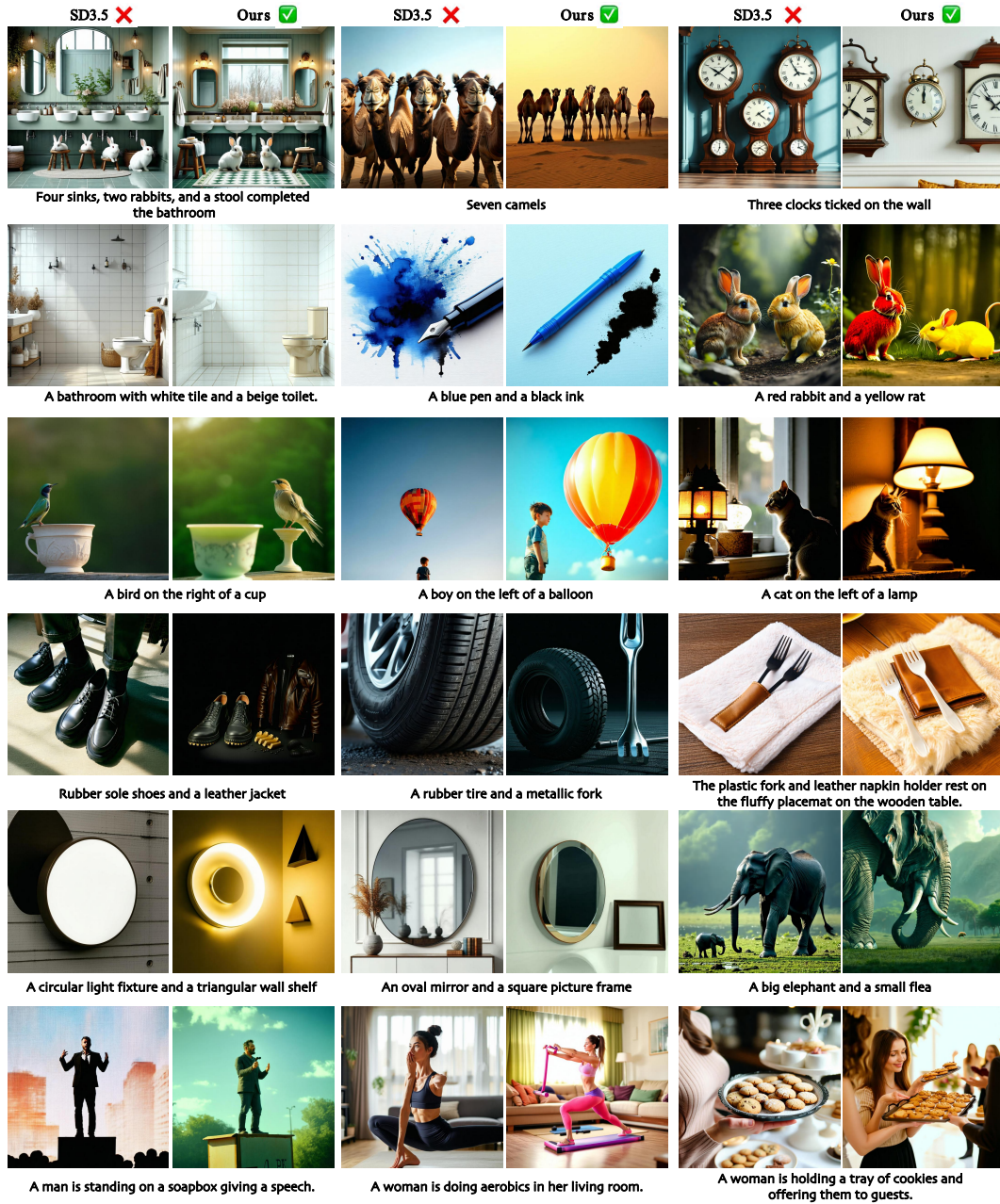


Figure A5: More qualitative results of SD3.5 and our method, covering aspects such as amount, color, spatial arrangement, texture, shape, and non-CLIP attributes. Our method consistently demonstrates better text alignment.

G.2 MORE FLUX RESULTS

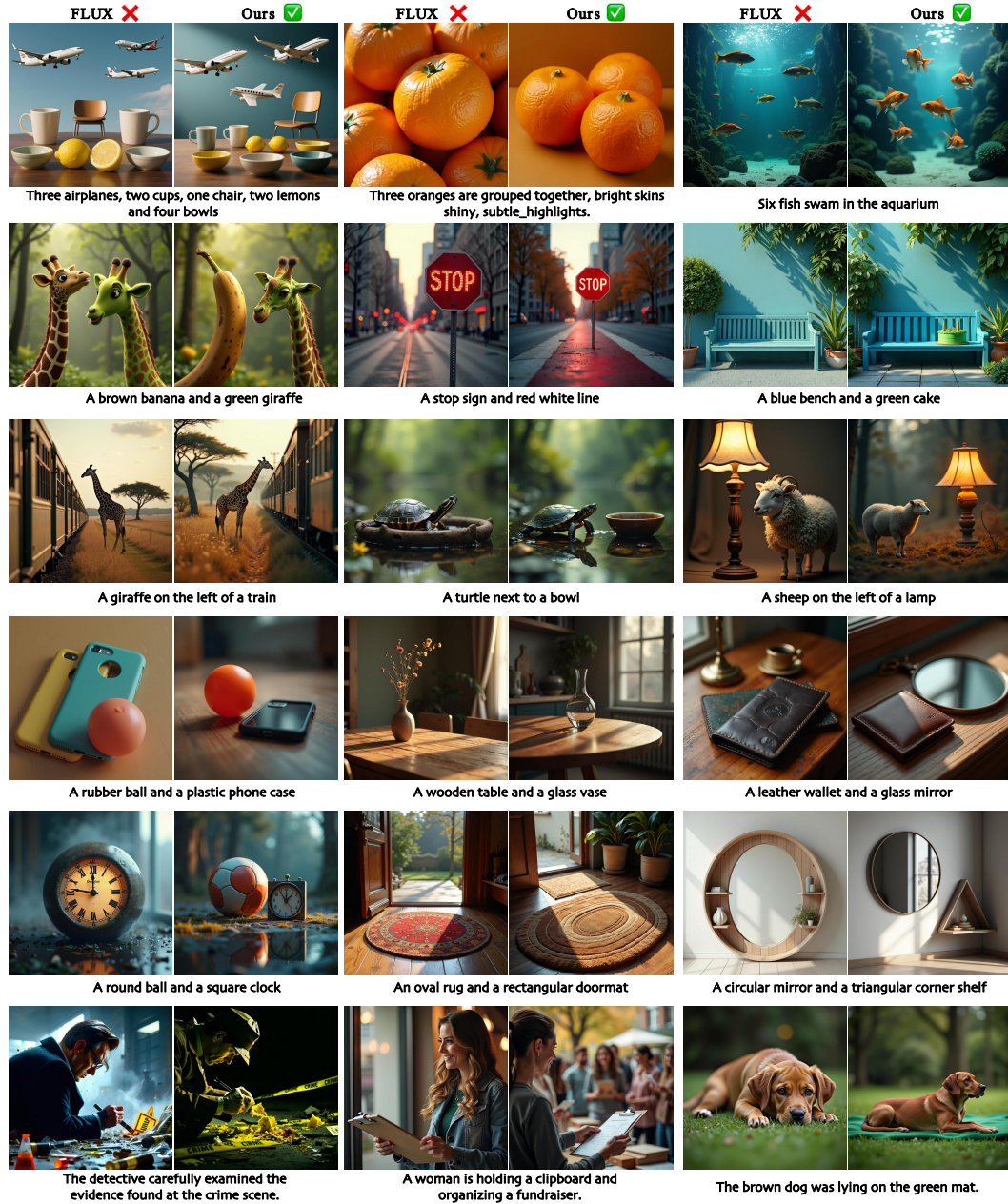


Figure A6: More qualitative results of FLUX and our method. Our approach achieves better text-image alignment and, in some cases, improved aesthetics over the baseline.

G.3 MORE QWEN IMAGE RESULTS

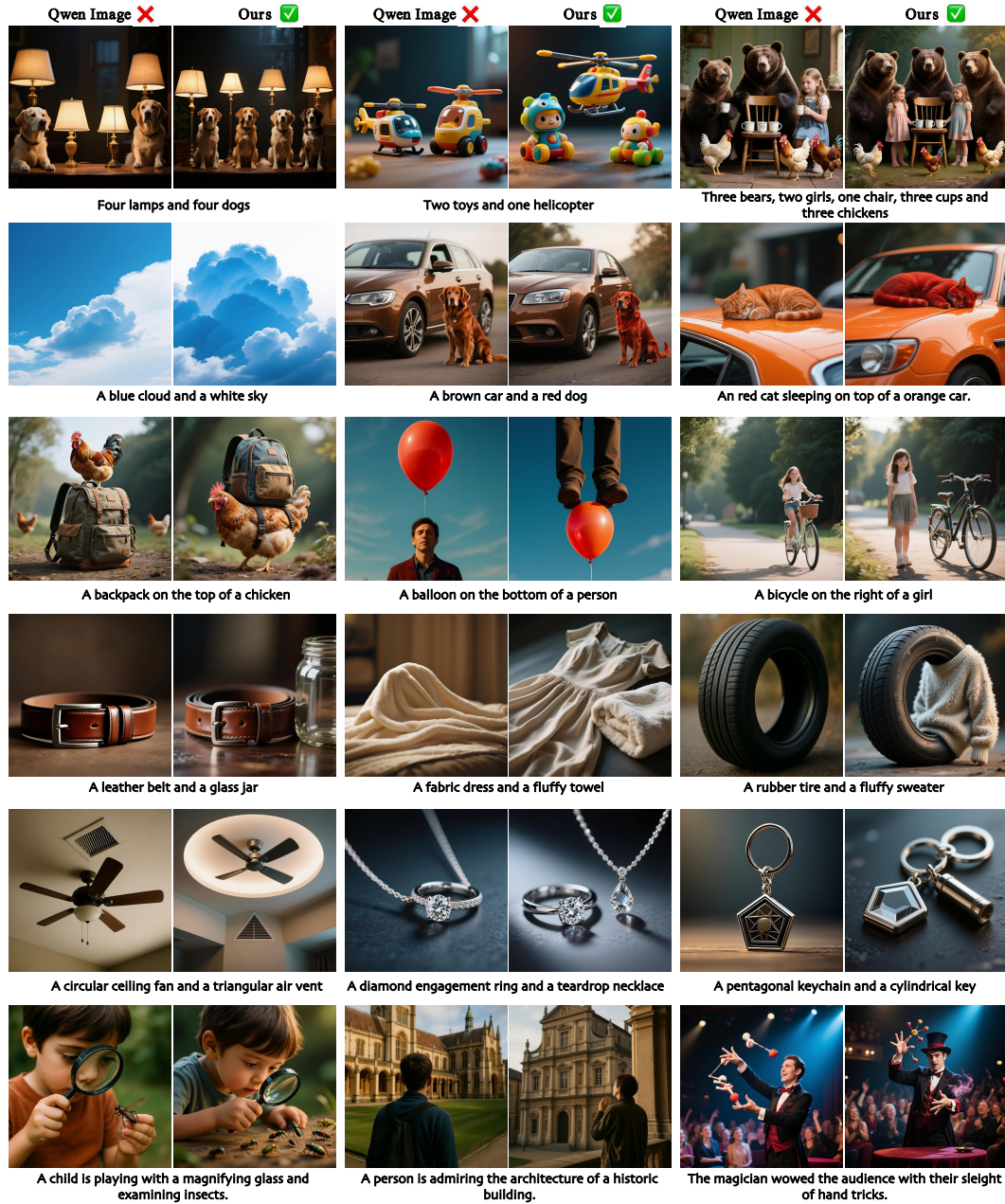


Figure A7: More qualitative results of Qwen Image and our method. Although the baseline fails on "A blue cloud and a white sky", our method succeeds.

G.4 MORE EDITING RESULTS

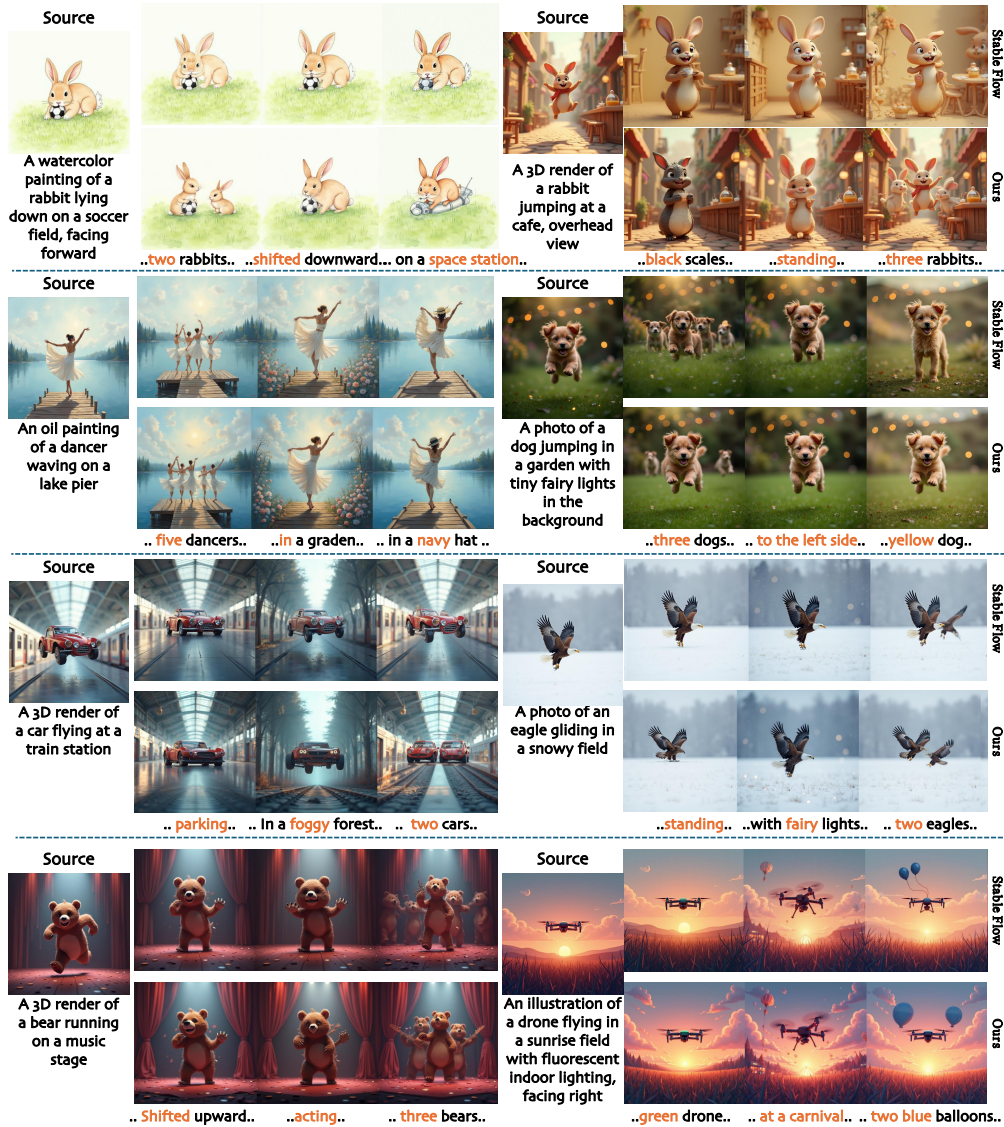


Figure A8: More editing examples on FLUX with our method and StableFlow. Our approach particularly surpasses StableFlow in quantity while maintaining high fidelity and strong text-image alignment.

G.5 FAILURE CASES

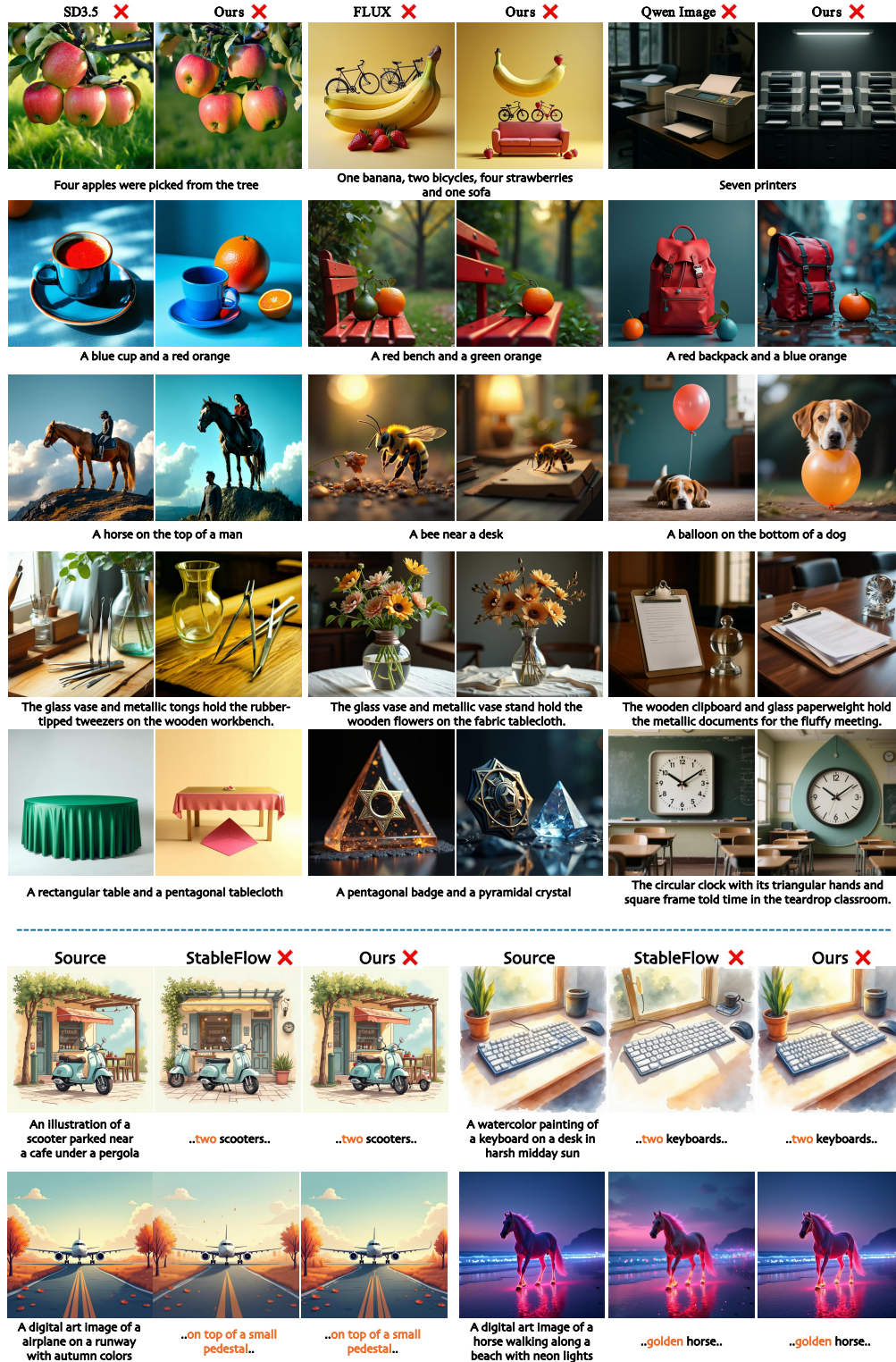
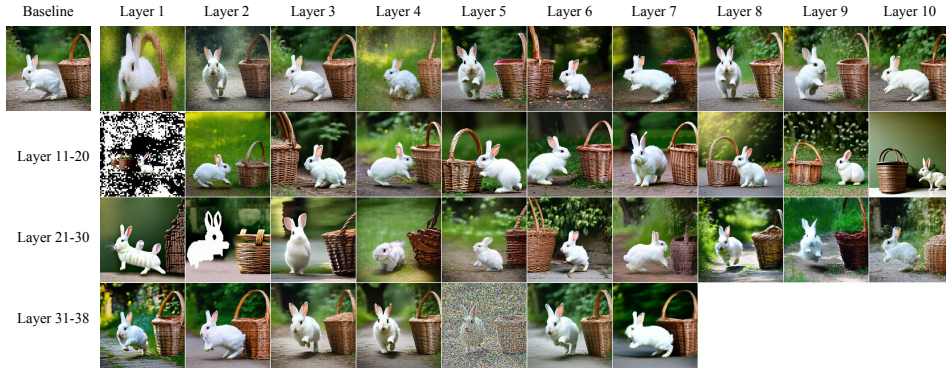
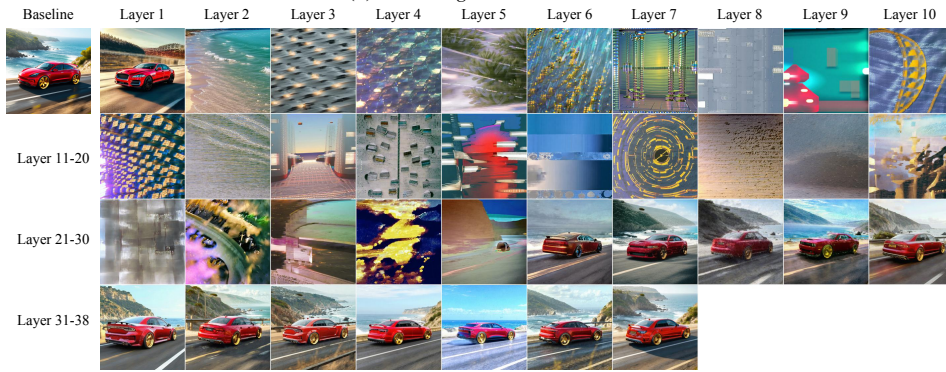


Figure A9: Failure cases of generation and editing. SD3.5 sometimes misidentifies attributes, confusing colors and objects. For rare real-world cases, our method may produce correct attributes but also hallucinations, e.g., a dog missing its body. In editing, hard cases mainly involve amount, reflecting the model's limited counting ability.

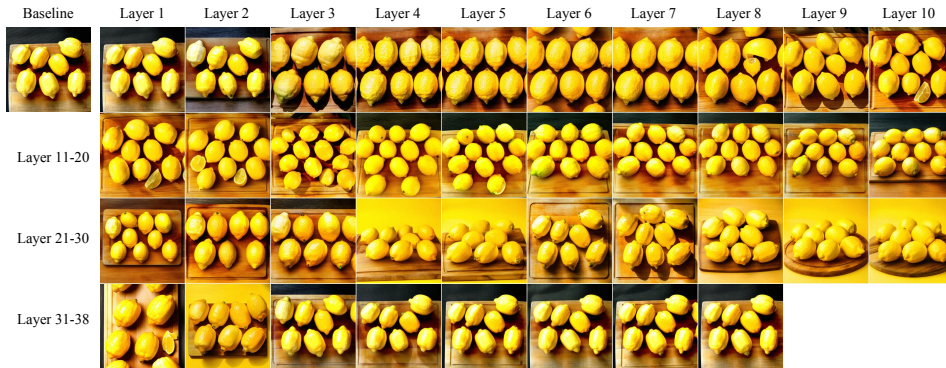
G.6 ALL-BLOCK SHOWCASES OF PROBING ANALYSIS



(a) removing all-block showcase.



(b) disabling all-block showcase.

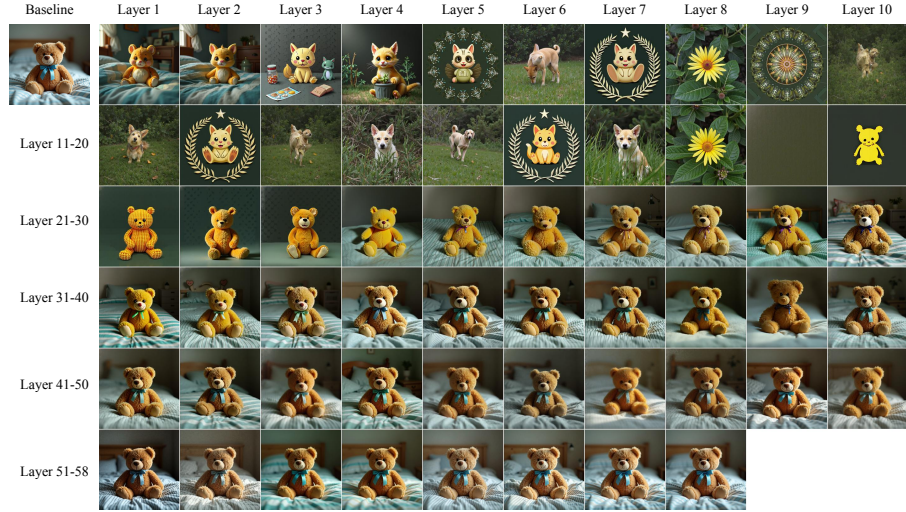


(c) enhancing all-block showcase.

Figure A10: All-block showcases of probing analysis on Stable Diffusion 3.5-large. From top to bottom, the prompts are: (a).“A white rabbit is hopping to the right of a brown basket.”,(b).“A red car with golden rims speeds along a coastal road.”, (c).“Seven lemons are arranged on a wooden cutting board, skins textured, color bright.”



(a) removing all-block showcase.

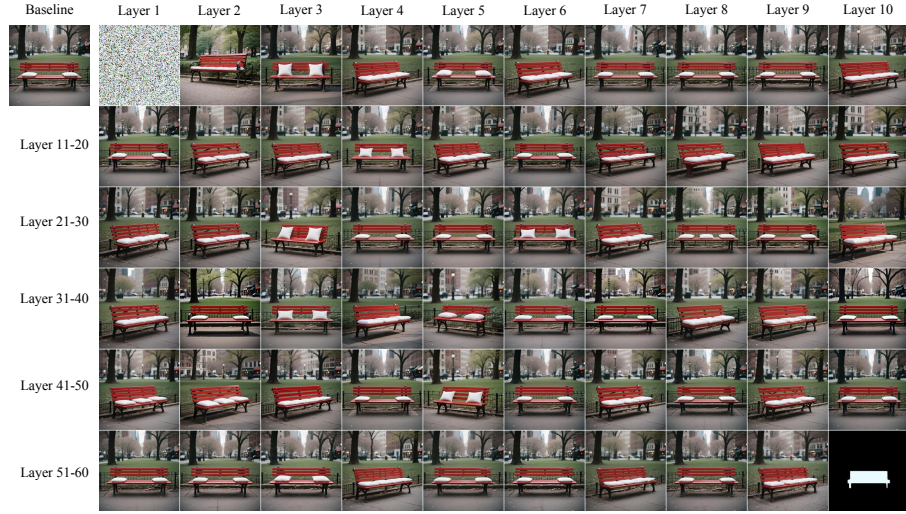


(b) disabling all-block showcase.

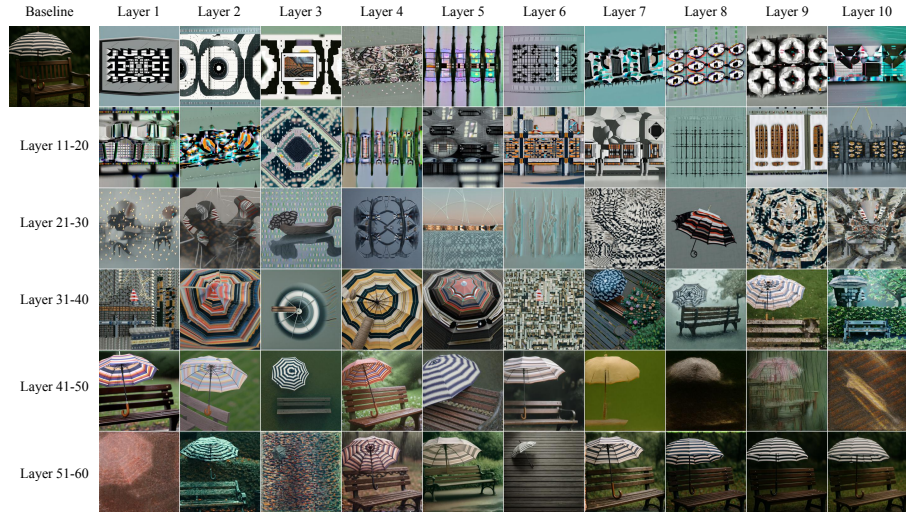


(c) enhancing all-block showcase.

Figure A11: All-block showcases of probing analysis on FLUX.1-Dev. From top to bottom, the prompts are: (a).“Eight green limes rest in a basket, shiny skins and small droplets of water visible.”,(b).“A brown teddy bear with a blue ribbon sits on a child’s bed with striped sheets.”, (c).“A small turtle moves to the right of a seashell.”



(a) removing all-block showcase.



(b) disabling all-block showcase.



(c) enhancing all-block showcase.

Figure A12: All-block showcases of probing analysis on Qwen Image. From top to bottom, the prompts are: (a).“A red bench with white cushions stands in a quiet city park.”,(b).“A striped umbrella is positioned at the top-left of a wooden bench.”, (c).“Five seashells lie in a row on sand, each detailed with ridges and soft reflections.”

H HUMAN EVALUATION DETAILS

We conduct human evaluations for both generation and editing tasks. For text-to-image generation, we sample 100 prompts from T2I-Compbench++ and generate results using FLUX.1-Dev and our method. Participants are provided with the prompt and the two generated images in random order, and answer the questions in Fig. A13a. The human preference score is defined as

$$\text{Human Preference} = \frac{1}{4}\text{Alignment} + \frac{1}{4}\text{Aesthetic} + \frac{1}{2}\text{Overall}.$$

For image editing, we randomly select 100 samples from the editing dataset and apply our method and the baseline Avrahami et al. (2025) with identical prompts. Three participants are recruited and, given the original image, the editing prompt, and two edited results in random order, they answer the questions in Fig. A13b. The final preference score is computed as

$$\text{Human Preference} = \frac{1}{4}\text{Alignment} + \frac{1}{4}\text{Preservation} + \frac{1}{2}\text{Overall}.$$

The corresponding quantitative results are summarized in Tab. A8 and Tab. A9.

Table A8: Human evaluation results for generation.

Methods	Alignment	Aesthetic	Overall	Human Preference
FLUX	118	134	109	39.17%
+ Ours	182	166	191	60.83%

Table A9: Human evaluation results for editing.

Methods	Alignment	Preservation	Overall	Human Preference
StableFlow	124	136	115	40.83%
Our Method	176	164	185	59.17%

Model Generation Evaluation

Prompt: A blue bench and a green cake



1. Which result better matches the text prompt?

☐ Result 1
☒ Result 2

2. Which result is more visually pleasing and reasonable?

☐ Result 1
☒ Result 2

3. Overall, which result is better?

☐ Result 1
☒ Result 2

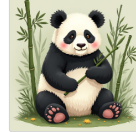
Submit

(a) Interface for generation evaluation.

Image Editing Evaluation

Source Prompt: An illustration of a panda sitting with bamboo nearby
Target Prompt: An illustration of three pandas sitting with bamboo nearby

Source Image



Editing Result 1



Editing Result 2



1. Which of the results is better in adhering to the text prompt?

☐ Result 1
☒ Result 2

2. Which of the results is better in preserving the information of the input image?

☒ Result 1
☐ Result 2

3. Which of the results is better overall?

☐ Result 1
☒ Result 2

Submit

(b) Interface for editing evaluation

Figure A13: Human evaluation interfaces for (a) generation and (b) editing. Participants are asked three questions corresponding to the given prompts and images.

REFERENCES

- Niki Amini-Naieni, Tengda Han, and Andrew Zisserman. Countgd: Multi-modal open-world counting. In *Advances in Neural Information Processing Systems*, volume 37, pp. 48810–48837, 2024.
- Omri Avrahami, Or Patashnik, Ohad Fried, Egor Nemchinov, Kfir Aberman, Dani Lischinski, and Daniel Cohen-Or. Stable flow: Vital layers for training-free image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7877–7888, 2025.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibor Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.
- Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. In *Advances in Neural Information Processing Systems*, volume 36, pp. 52132–52152, 2023.
- Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. In *Advances in Neural Information Processing Systems*, volume 36, pp. 78723–78747, 2023.
- Kaiyi Huang, Chengqi Duan, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench++: An enhanced and comprehensive benchmark for compositional text-to-image generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.
- Black Forest Labs. Flux. <https://github.com/black-forest-labs/flux>, 2024.
- Feng Liu, Shiwei Zhang, Xiaofeng Wang, Yujie Wei, Haonan Qiu, Yuzhong Zhao, Yingya Zhang, Qixiang Ye, and Fang Wan. Timestep embedding tells: It’s time to cache for video diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.
- Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *Proceedings of the European Conference on Computer Vision*, pp. 38–55. Springer, 2024.
- Chrisoph Schuhmann. Laion aesthetics. <https://github.com/LAION-AI/aesthetic-predictor.git>, 2022.
- Stability-AI. Stable diffusion 3.5. <https://github.com/black-forest-labs/flux>, 2024.
- Hugging Face Team. Diffusers. <https://github.com/huggingface/diffusers>, 2025.
- Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.
- Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.