Devil is in the Detail: Towards Injecting Fine Details of Image Prompt in Image Generation via Conflict-free Guidance and Stratified Attention

Supplementary Material

A. Overview

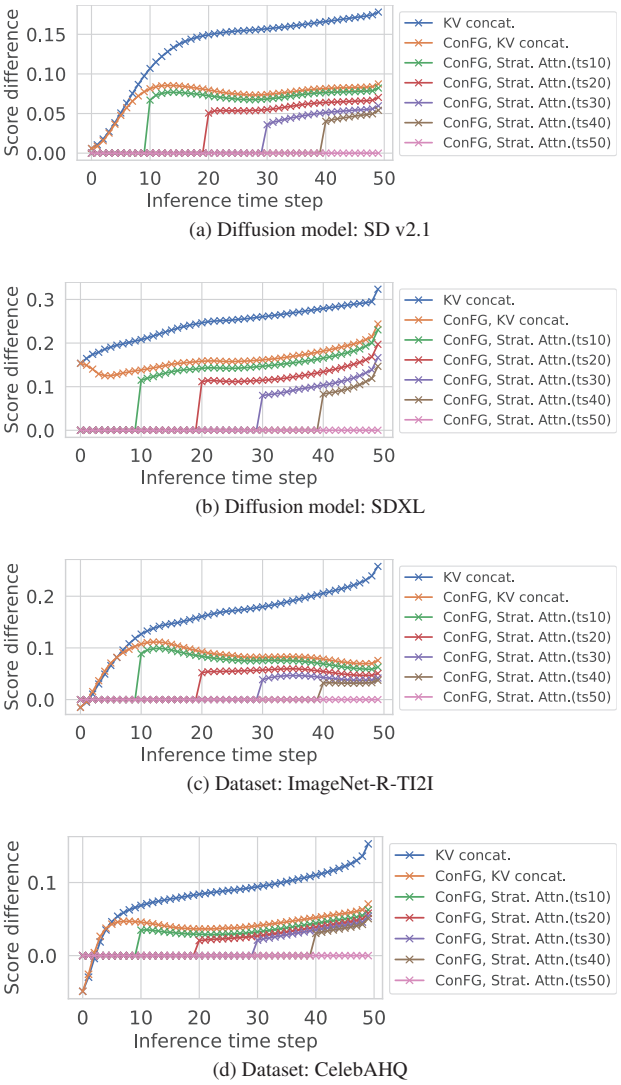
In this supplementary, we first provide the implementation details of the proposed method (Sec. B). Next, expanding the analysis in Sec. 5.2, we further analyze the impact of the proposed method on alignment with the image prompt across three aspects: self-attention layers, diffusion models, and datasets (Sec. C). Subsequently, we conduct ablation studies (Sec. D) and analyze the impact of the negative image prompt and weighted attention in the proposed conflict-free guidance (Sec. E) and Stratified Attention (Sec. F), respectively, to demonstrate the method's effectiveness in faithfully reflecting the given image prompt and its details in the generated images. In addition, we explain the role of text and image prompts in our model (Sec. G) and compare the performance in image prompt alignment with identity-preserving methods (Sec. H). Finally, we highlight the superiority of the proposed method over existing approaches as well as its potential for cross-prompt image generation through additional qualitative results (Sec. I).

B. Implementation Details

We conduct all experiments with 50 inference steps for each diffusion model. For the classifier-free guidance scale, we set the value to 7.5 for Stable Diffusion (SD) [13] and 5.0 for Stable Diffusion XL [11]. Additionally, the self-attention modification is applied exclusively to the decoder blocks, and Stratified Attention performs a weighted sum of the attention for the generated image and the image prompt, assigning weights of $\lambda_G = 0.5$ and $\lambda_R = 0.5$, respectively. We also employ Stratified Attention up to 10 time steps, except for structure-guided image generation using ControlNet, where it is applied up to 25 time steps. The positive text prompt is appended with ", best quality, extremely detailed." at the end of the sentence, while the negative text prompt is set to "monochrome, bad anatomy, bad hands, cropped, worst quality.". To ensure a fair comparison and eliminate the influence of text prompts, we use the same text prompt across all models, including the baselines. Following RIVAL's setting, we use null prompts for DDIM inversion in cross-prompt image generation, while employing positive text prompts in all other experiments.

C. Effects on the Alignment of Image Prompt

In Sec. 5.2, we demonstrate the effectiveness of two key components, conflict-free guidance and Stratified Attention, of the proposed method in increasing the attention score of the image prompt by visualizing changes in attention scores within the self-attention layer of Stable Diffusion, using 100 randomly sampled images from the COCO validation set (Fig. 11). However, this analysis is limited to results from the first of nine self-attention layers in the decoder in Stable



(a) Diffusion model: SD v2.1

(b) Diffusion model: SDXL

(c) Dataset: ImageNet-R-TI2I

(d) Dataset: CelebAHQ

Figure A. Comparison of the impact of conflict-free guidance (ConFG) and Stratified Attention (Strat. Attn.) on the attention scores of the image prompt across *diffusion models* (a and b) and *datasets* (c and d). A smaller difference on the y-axis indicates stronger attention to the image prompt, resulting in a more faithful alignment in the generated images.

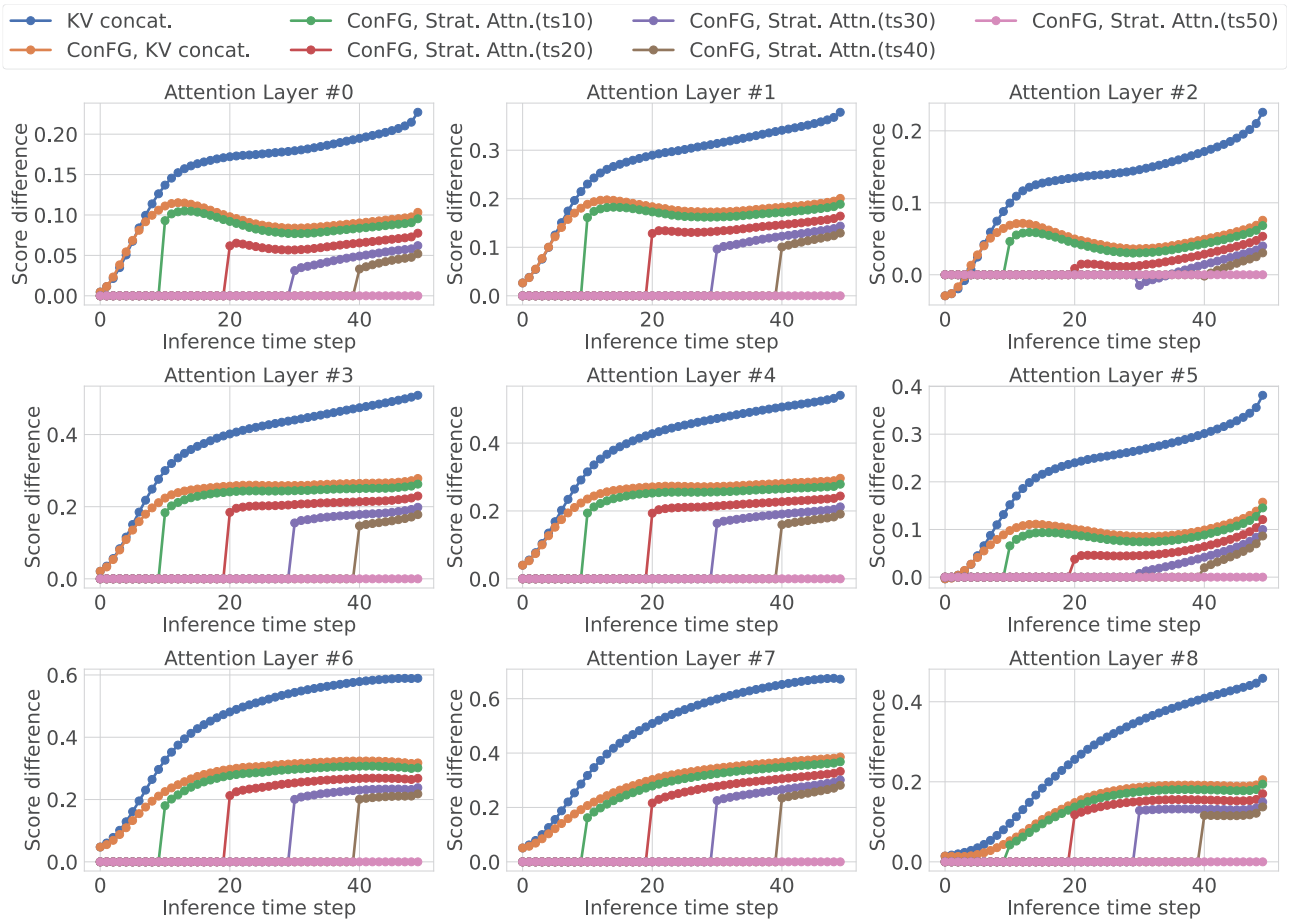


Figure B. Comparison of the impact of conflict-free guidance (ConFG) and Stratified Attention (Strat. Attn.) on the attention scores of the image prompt across *self-attention layers*. A smaller difference on the y-axis indicates stronger attention to the image prompt, resulting in a more faithful alignment in the generated images.

Diffusion version 1.5 and relies on a single dataset. In this section, we broaden the analysis by reporting results from additional layers, diffusion models, and datasets to confirm the consistent performance of the proposed method across these three factors.

First, we present the results for all layers of SD version 1.5, including the first layer, in Fig. B. While the attention scores of the image prompt show slight variations across layers, they consistently improve as we apply conflict-free guidance (ConFG) or use longer time steps in Stratified Attention (Strat. Attn.). Furthermore, we observe that these trends remain consistent across all layers in the subsequent experiments, even though we visualize the results from the first layer.

Next, we validate that the two proposed components consistently enhance the attention score of the image prompt across different diffusion models by conducting the same experiment with SD version 2.1 (Fig. Aa) and SDXL (Fig. Ab). The results indicate that, while the attention score values differ slightly from those of SD version 1.5, both models exhibit the same overall trend. Specifically, a minor difference arises at the initial time step: SD models generally produce nearly identical attention scores for the image prompt and the generated features, whereas SDXL yields slightly higher attention scores for the generated features. Despite this variation, the proposed method consistently improves the attention score of the image prompt across all models. These findings confirm that the proposed method reliably increases the attention score of the image prompt, regardless of the diffusion model. Additionally, as shown in Tab. 3 and Fig. 12, this improvement strengthens alignment with the image prompt, further demonstrating the robustness of the method.

Finally, we confirm that the two key components consistently enhance the attention score of the image prompt across datasets by conducting the same experiments on

	CLIP-I ↑	CLIP-T ↑
KV concatenation	0.770	**0.316**
ConFG + KV concatenation	0.882	0.311
ConFG + Stratified Attention (ts10)	0.886	0.311
ConFG + Stratified Attention (ts20)	0.893	0.310
ConFG + Stratified Attention (ts30)	0.899	0.308
ConFG + Stratified Attention (ts40)	0.901	0.307
ConFG + Stratified Attention (ts50)	**0.902**	0.306

Table A. Ablation studies for key components: Conflict-free guidance (ConFG) and Stratified Attention. Applying ConFG or increasing the time steps of Stratified Attention enhances the faithful reflection of the image prompt in the generated image.

λ_P (Image prompt)	λ_G (Generated fetures)	CLIP-I ↑	CLIP-T ↑
1.00	0.00	**0.917**	0.301
0.67	0.33	0.908	0.303
0.50	0.50	0.902	0.306
0.33	0.67	0.884	0.309
0.00	1.00	0.724	**0.313**

Table B. Quantitative impact of the weights applied to the attention of the image prompt and generated features in Stratified Attention on the alignment with the image and text prompts. λ_P and λ_G represent the weights applied to the image prompt and generated features, respectively. Since we use the diffusion model in a training-free manner, the sum of the two weights is set to 1.

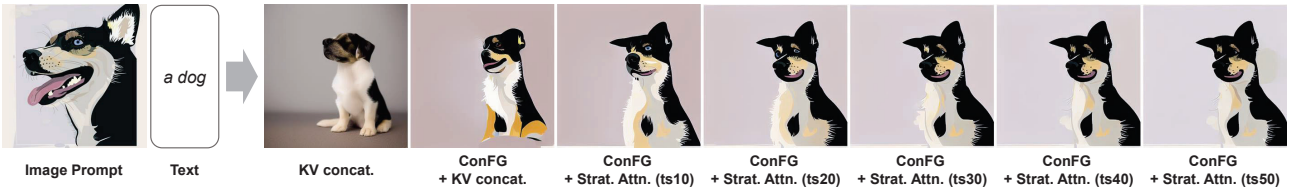


Figure C. Qualitative results of ablation studies. When Conflict-Free Guidance (ConFG) is not applied, the image prompt serves as both positive and negative guidance, creating conflicts that hinder the generated image from reflecting the given prompt. However, with ConFG, the generated image aligns effectively with the provided prompt, and Stratified Attention (Strat. Attn.) further improves this alignment.

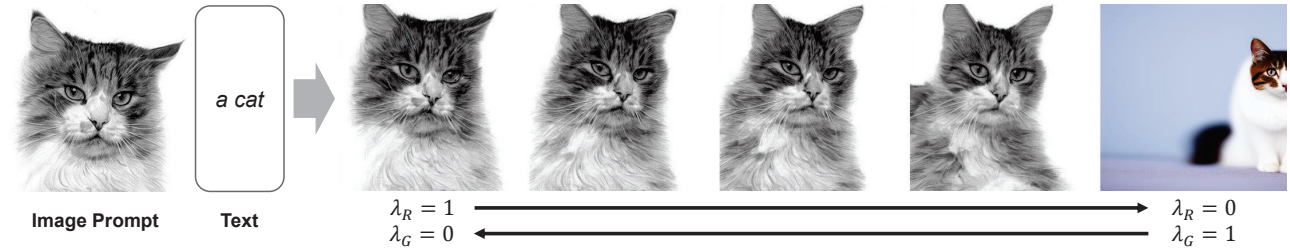


Figure D. Visual results showing the impact of varying the weights applied to the attention of the image prompt and generated features in Stratified Attention on the generated images. When the weight for the image prompt, λ_P, is set to 0, the image prompt is not incorporated at all. However, the separated attention score computation of Stratified Attention ensures that even a small weight for the image prompt allows it to be faithfully reflected. Additionally, as λ_P increases, the generated images progressively resemble the structure of the image prompt.

the ImageNet-R-TI2I dataset [16] (Fig. Ac) and the CelebAHQ dataset [7] (Fig. Ad). Consistent with previous results, applying conflict-free guidance or increasing the time steps with Stratified Attention reliably improves the attention score of the image prompt, regardless of the dataset. These findings align with the results presented in Tab. 1.

From these analyses, we conclude that the proposed method consistently enhances the attention score of the image prompt, ensuring its faithful incorporation into the generated images across diverse diffusion models and datasets. This conclusion is well-supported by the results presented in Sec. 5.

D. Ablation Study

This section presents ablation studies on the two key components of the proposed method: conflict-free guidance (ConFG) and Stratified Attention (Strat. Attn.). To assess the effectiveness of these components in faithfully incorporating the image prompt into the generated images, we generate 20 images per image prompt from the ImageNet-R-TI2I dataset. Since the proposed method employs an Attention Fusion strategy [22] that combines KV concatenation with Stratified Attention, its performance is compared to KV concatenation (concat.) as the baseline.

First, we quantitatively demonstrate the advantages of

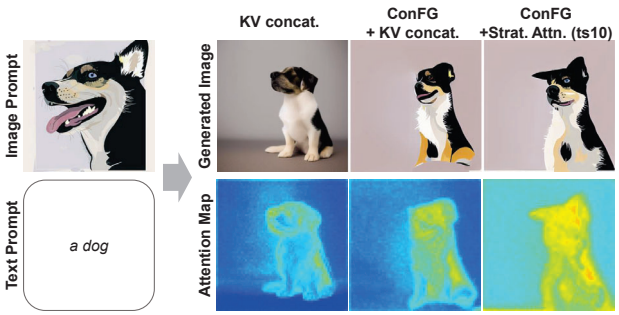


Figure E. Visualization of the attention map to show the effects of conflict-free guidance (ConFG) and Stratified Attention (Strat. Attn.) on image prompt alignment. The color scale of the attention maps ranges from red, indicating high alignment with the image prompt, to blue, indicating low alignment.

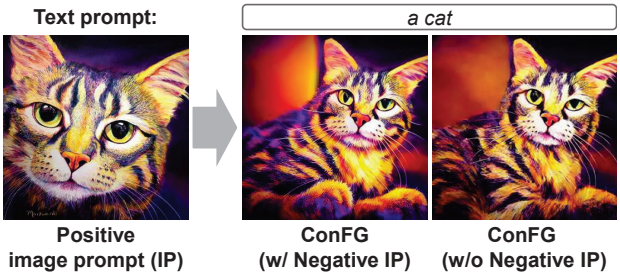


Figure F. Qualitative results showing the impact of using and not using negative image prompts. Using negative image prompts introduces unintended effects, such as shifts in color tone, in the generated images.

the two proposed components in reflecting the image prompt. As shown in Tab. A, the quantitative results reveal that the conflict-free guidance slightly decreases the alignment with the text prompt but significantly enhances the incorporation of the image prompt by removing its role as negative guidance. Furthermore, this improvement becomes more evident as the time steps utilizing Stratified Attention increase. However, since applying Stratified Attention across all time steps reduce the alignment with text, we recommend using Stratified Attention for time steps between 10 and 30.

Also, the qualitative analysis further supports these findings (Fig. C). Without conflict-free guidance, the image prompt simultaneously acts as both positive and negative guidance, hindering its incorporation into the generated images. In contrast, conflict-free guidance resolves these conflicting roles, allowing the generated images to faithfully reflect the image prompt, including its color tones and textures. Additionally, increasing the time steps with Stratified Attention further improves the ability of the generated images to faithfully reflect the image prompt, including its color and content.

For further analysis, in Fig. E, we visualize the self-attention maps of the results from KV concatenation, conflict-free guidance, and Stratified Attention presented in Fig. C. According to these results, conflict-free guidance generally enhances the alignment of the image prompt with the subject, while Stratified Attention improves alignment across the entire generated image, including the subject.

From these results, we affirm the crucial role of conflict-free guidance and Stratified Attention in achieving faithful integration of the image prompt into the generated images, thereby demonstrating the effectiveness of the proposed method.

E. Impact of Negative Image Prompt on Conflict-free Guidance

Conflict-free guidance can be implemented in two ways: one where no image is supplied to the negative branch, and the other where an image dissimilar to the provided image prompt is used for the negative branch. To assess the efficacy of our proposed approach, we conduct experiments comparing image prompt alignment under both conditions. In these experiments, a negative image, generated from a negative text prompt, was used (the first image in Figure Fig. 5-a). The results yielded a CLIP-I score of 0.854 and a CLIP-T score of 0.319. When compared to the results obtained using identical image prompts for both branches (the first row in Table Tab. A), it is evident that directing the image prompt solely to the positive branch leads to improved alignment. However, the CLIP-I score remains lower than that of conflict-free guidance, where no image prompt is provided to the negative branch (second row). We attribute this discrepancy to the negative image prompt containing not only undesirable features (*e.g.*, poor anatomical structure) but also other characteristics (*e.g.*, color), which introduce unintended effects (*e.g.*, shifts in color tone) in the generated images (Fig. F).

F. Effects of Weighted Attention in Stratified Attention

In this section, we analyze the influence of the two hyperparameters, λ_P and λ_G, in Stratified Attention on achieving alignment with the image prompt (Eq. (3)). As mentioned in Sec. 4.2, Stratified Attention tackles the problem of attention bias, where queries disproportionately focus on keys and values derived from the same generated features in KV concatenation, thereby enhancing the integration of the image prompt. Specifically, it separately computes attention scores for the image prompt and the generated features, subsequently combining them through a weighted sum using λ_P and λ_G. Therefore, these parameters individually deter-

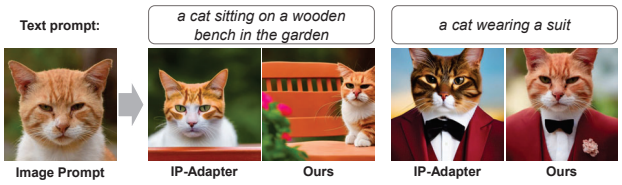


Figure G. Generated images from realistic scenarios

mine the extent to which the image prompt and the generated features are integrated into the final generated images. In these experiments, to clearly observe the effects of the weights, Stratified Attention is applied to all time steps.

To evaluate this, we quantitatively and visually assess the incorporation of the image prompt into the generated images by varying the values of the two hyperparameters. As shown in Tab. B, the quantitative analysis reveals that increasing λ_P, the weight assigned to the image prompt, while decreasing λ_G, the weight assigned to the generated features, enhances the faithful integration of the image prompt into the generated images. Conversely, reversing this adjustment diminishes the influence of the image prompt. The visual results in Fig. D corroborate this pattern, demonstrating that setting λ_P to 0 entirely excludes the image prompt, while progressively increasing λ_P incorporates not only the color tones and textures but also the structure of the image prompt. Based on these findings, we recommend setting λ_P within the range of 0.3 to 0.7.

G. Role of Text and Image Prompts

Since we use the text-to-image diffusion model in a training-free manner, we input text prompts alongside the image prompt to maintain the model's mechanism. In this setup, we typically assume the text and image prompts are aligned and aim to faithfully reflect the details of the image prompt, following the experimental setup of previous studies. However, when the image and text prompts are misaligned (Fig. 9b), we can enhance the influence of the text prompt by adjusting hyperparameters, similar to the baselines. As an additional example of misalignment between image and text prompts, we generate images using two text prompts in conjunction with a cat image prompt, and compare the results with those generated by IP-Adapter [19]. This experiment serves to demonstrate the superiority of our method, even in realistic scenarios. As shown in Fig. G, our approach produces images that more accurately capture the details of the cat in the image prompt, in contrast to the results from IP-Adapter.

H. Comparison with Identity-preserving Method

We conduct a comparative analysis with identity-preserving approaches [17], which generate images while maintaining the identity of a given face, as both models aim to capture the details of the provided image. We use InstantID [17] as the baseline for identity-preserving methods and evaluate both models using 58 images from the CelebA-HQ dataset [7], where InstantID successfully detects faces. The comparison is based on image prompt alignment, and the results demonstrate that our model outperforms InstantID, achieving a CLIP-I score 0.124 higher. This discrepancy underscores the superiority of our method, which not only preserves the identity of the image prompt but also captures additional details, such as texture, while InstantID focuses solely on identity preservation.

I. Additional Qualitative Results

In addition to the qualitative results provided in Sec. 5, this section presents additional visual results across three tasks to demonstrate the superiority of our method: cross-prompt image generation with various text prompts, cross-prompt image generation with structural conditions, and image variation.

First, we present the results of our method in cross-prompt generation tasks with various text prompts. In Sec. 5.1, we compare its performance to baseline models by generating images from a single image-text pair. Extending this, we pair a single image prompt with multiple text prompts, generating images for each pair to show that our method consistently integrates both prompts. As shown in Fig. H, our method reliably reflects the color tones and textures of the image prompt while adapting to diverse text prompts. These results demonstrate that the proposed method enhances controllability in image generation and holds promise for applications such as style transfer.

Next, we demonstrate the performance of the proposed method in cross-prompt image generation with a given structural condition, combining the two tasks presented in Sec. 5.1. To incorporate the structural condition as input, we employ ControlNet [20], following the approach used in the main paper. As illustrated in Fig. I, the proposed method faithfully integrates the structural condition while generating text-prompted images that consistently reflect the style of the image prompt, including its color tone and texture, as observed in the earlier cross-prompt image generation results.

Finally, along with the results on the CelebAHQ dataset shown in Fig. 9a, we provide additional visual results generated from the COCO validation dataset and ImageNet-R-TI2I to demonstrate that the proposed method consistently outperforms baseline models by faithfully reflecting the details of the given image prompt. Starting with the COCO validation dataset (Fig. Ja), the proposed method accurately reproduces the details of the image prompt, such as the red dot pattern on a black tie, whereas the baselines simplify it to a plain red tie. Additionally, our method preserves

intricate features like mural patterns, bedspread designs, and elongated light decorations in the generated images, while these elements are missing in images produced by the baselines. This trend is also observed in the results on the ImageNet-R-TI2I dataset (Fig. Jb). The proposed method faithfully reflects textures, such as the pencil strokes in a sketch and the textures and colors of a painting. In contrast, IP-Adapter and SSR-Encoder exhibit color shifts, and RIVAL tends to smooth out the sketch textures or make paintings appear overly photorealistic.

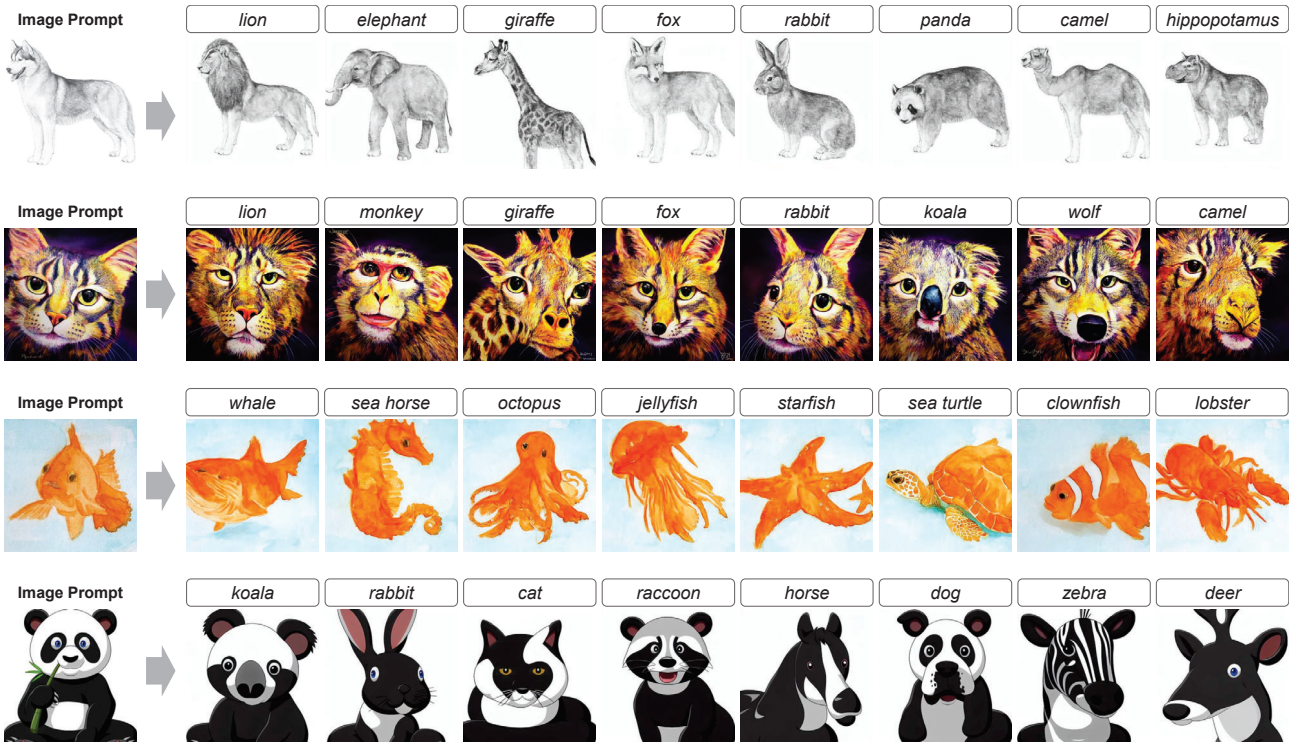


Figure H. Qualitative results of our method on cross-prompt image generation tasks. The text within the rounded rectangle serves as a text prompt. Our method generates images from text prompts that consistently capture the color tones and textures of the image prompts.

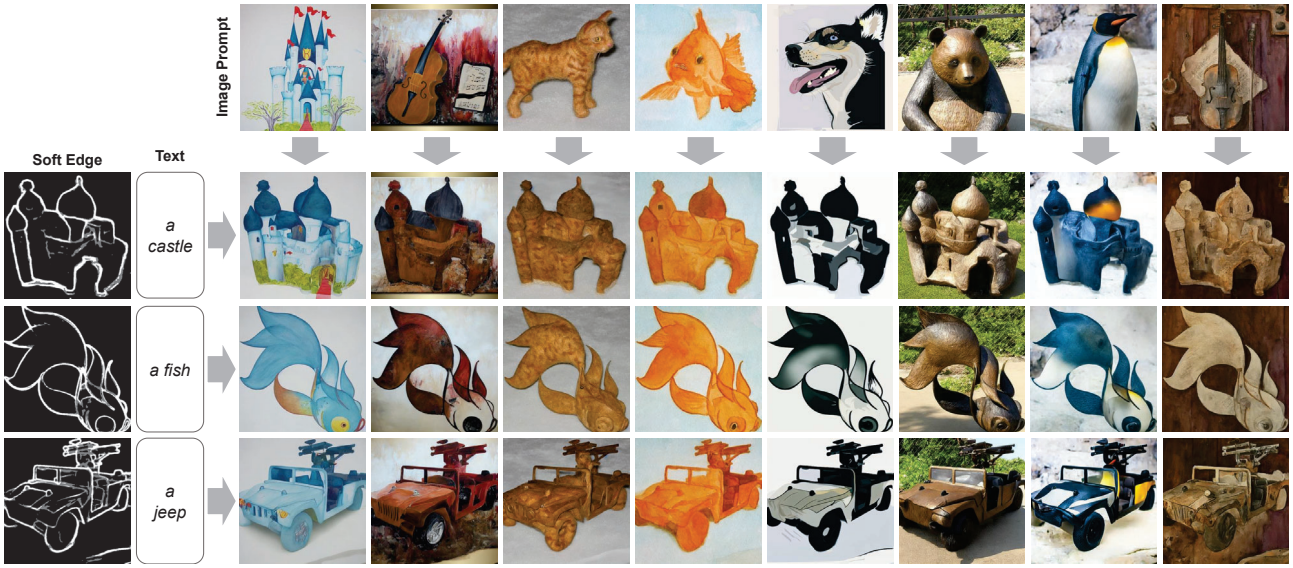
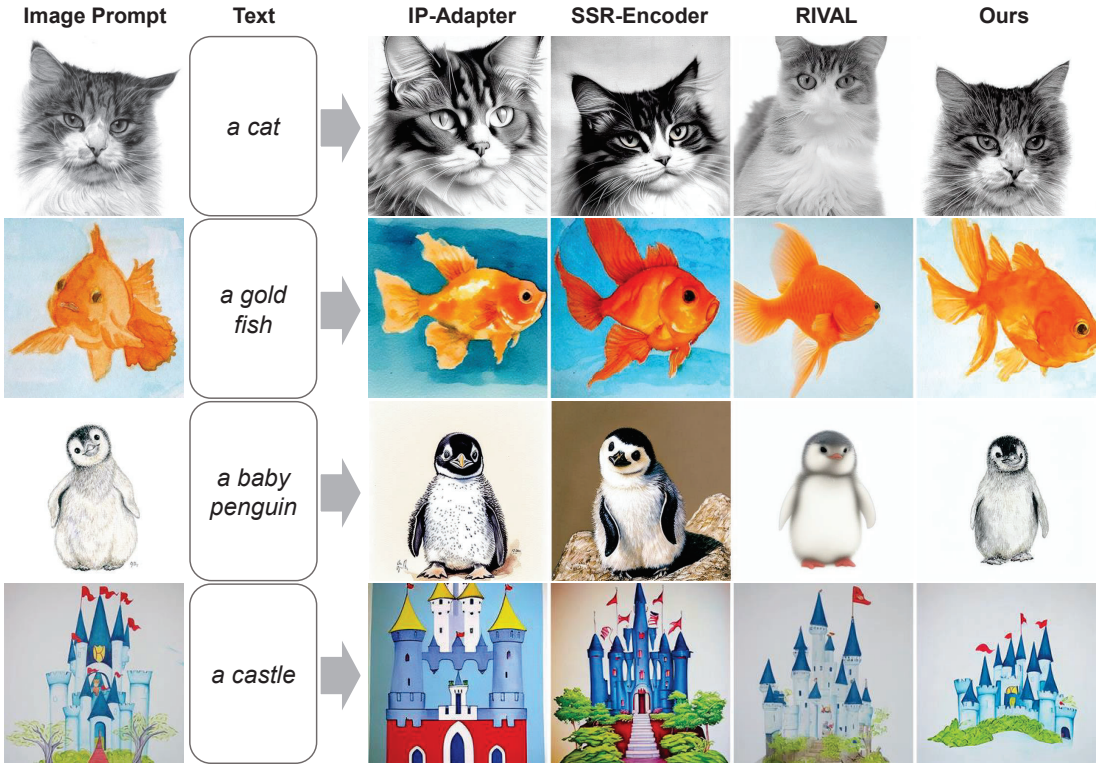


Figure I. Qualitative results of our method on the combination of cross-prompt image generation and structure-guided image generation tasks. Our method generates images from text prompts that consistently reflect the color tones and textures of the image prompts while capturing the structure of images with soft edges.

(a) COCO validation. The baselines struggle to reflect the intricate details of the given image prompt, such as the necktie's pattern, the mural, the bed blanket's design, or the light decorations, whereas our method faithfully incorporates these elements.



(b) ImageNet-R-I2I. The baselines fail to capture the color tone and texture of the given image prompt, whereas our method faithfully incorporates them.

Figure J. Qualitative results of image variation tasks, with corresponding quantitative results shown in Tab. 1.