

SoftREPA for better Text-to-Image Alignment

Supplementary Materials

A Information Theoretical Analysis of Diffusion Models

Several studies have explored diffusion models from an information-theoretic viewpoint. According to [18], the mutual information between random variables X and Y can be expressed as the expectation of pointwise mutual information over their joint distribution:

$$I(X, Y) = \mathbb{E}_{p(\mathbf{x}, \mathbf{y})}[i(\mathbf{x}, \mathbf{y})], \quad (19)$$

where the pointwise mutual information (PMI) is defined as the difference between the conditional and marginal log-likelihoods of a given sample \mathbf{x} :

$$i(\mathbf{x}, \mathbf{y}) = \log p(\mathbf{x}|\mathbf{y}) - \log p(\mathbf{x}). \quad (20)$$

This quantity measures how much the presence of condition \mathbf{y} affects the probability distribution near \mathbf{x} .

Song et al. [35] and Kong et al. [18] showed that, under the assumption of an optimal diffusion model trained on given data, the log-likelihood of an image \mathbf{x} can be formulated as:

$$-\log p(\mathbf{x}) = \frac{1}{2} \int_0^T \lambda(t) \mathbb{E}_{\epsilon} [\|\epsilon_{\theta}(\mathbf{x}_t, t) - \epsilon\|^2] dt + C \quad (21)$$

This result also extends to conditional diffusion models, where the likelihood of \mathbf{x} given condition \mathbf{y} (e.g., text embeddings) is given by:

$$-\log p(\mathbf{x}|\mathbf{y}) = \frac{1}{2} \int_0^T \lambda(t) \mathbb{E}_{\epsilon} [\|\epsilon_{\theta}(\mathbf{x}_t, t, \mathbf{y}) - \epsilon\|^2] dt + C \quad (22)$$

B Implementation Details

Architecture Details For Stable Diffusion 3, the main experiments use soft tokens of length 4, applied across 5 layers. In Stable Diffusion 1.5, soft tokens of length 4 are applied only to the Down block layers of the UNet’s conditional text embeddings. For Stable Diffusion XL, soft tokens of length 8 are used, applied to both the Down and Middle block layers of the UNet’s conditional text embeddings, without incorporating timestep dependency.

Training Details The hyperparameters used during training are listed in table 3. For Stable Diffusion 3, the soft tokens are optimized solely using the contrastive score matching loss (L_{SoftREPA}). In contrast, for Stable Diffusion 1.5 and Stable Diffusion XL, optimization combines the contrastive score matching loss with a small weighting of the denoising score matching loss. We observed that initializing the soft tokens with a random distribution led to performance degradation in Stable Diffusion 1.5. To address this, we initialized the soft tokens using unconditional text embeddings. The code is publicly available at <https://github.com/softrepa/SoftREPA>.

Models	lr	wd	batch size (positive, negative)	iterations	token init	optimizer	lr scheduler
SD1.5	1e-3	1e-4	32(4, 28)	26,000	\emptyset	AdamW	CosineAnnealingWarmRestarts
SDXL	1e-3	1e-4	16(1, 15)	30,000	$N(0, 0.02)$	AdamW	CosineAnnealingWarmRestarts
SD3	1e-3	1e-4	16(4, 12)	30,000	$N(0, 0.02)$	AdamW	CosineAnnealingWarmRestarts

Table 3: The implementation details for training.

Inference Details A detailed description of the image generation algorithm on MM-Dit is provided in algorithm 1. At each layer, a distinct set of soft tokens is prepended to the text features and used exclusively within that layer. These soft tokens are not carried over to subsequent layers; instead, new soft tokens are introduced or omitted as appropriate. Their primary role is to guide the text tokens toward better text-image alignment, particularly during the early stages of the model.

Algorithm 1 Image Generation with Soft Tokens in MM-DiT

Require: Gaussian noise $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$

Require: Text $\mathbf{Y} \sim p_{\text{data}}$

Require: Soft token $\mathbf{s} \sim \text{Embedding}(k, t)$

Require: Number of layers N , Threshold layer L

Require: Time steps $\{t_T, t_{T-1}, \dots, t_0\}$

```
1: Initialize  $\mathbf{H}_{\text{img}}^{(0,T)} \leftarrow \mathbf{z}$ 
2: Initialize  $\mathbf{H}_{\text{text}}^{(0,T)} \leftarrow \text{TextEncoder}(\mathbf{Y})$ 
3:  $n = |\mathbf{H}_{\text{text}}^{(0,T)}|$ 
4: for  $t$  in  $\{t_T, t_{T-1}, \dots, t_0\}$  do
5:   for  $l = 1$  to  $N$  do
6:     if  $k \leq L$  then
7:        $\mathbf{s}^{(k,t)} \leftarrow \text{Embedding}(k, t)$ 
8:        $\hat{\mathbf{H}}_{\text{text}}^{(k-1,t)} \leftarrow [\mathbf{s}^{(k,t)}; \mathbf{H}_{\text{text}}^{(k-1,t)}]$ 
9:     else
10:       $\hat{\mathbf{H}}_{\text{text}}^{(k-1,t)} \leftarrow \mathbf{H}_{\text{text}}^{(k-1,t)}$ 
11:    end if
12:     $\mathbf{H}_{\text{img}}^{(k,t)}, \hat{\mathbf{H}}_{\text{text}}^{(k,t)} \leftarrow \text{Layer}_l(\mathbf{H}_{\text{img}}^{(k-1,t)}, \hat{\mathbf{H}}_{\text{text}}^{(k-1,t)})$ 
13:     $\mathbf{H}_{\text{text}}^{(k,t)} \leftarrow \hat{\mathbf{H}}_{\text{text}}^{(k,t)}[-n :, :]$  ▷ Drop soft tokens
14:  end for
15: end for
16: return  $\hat{\mathbf{X}} = \text{Decoder}(\mathbf{H}_{\text{img}}^{(N,t_0)})$ 
```

736 **Memory Efficiency** For the memory efficiency metrics reported in table 1, values were measured
737 by averaging results over 50 runs with an A100 GPU. Image resolution was set to 512×512 for
738 SD1.5 and 1024×1024 for SDXL and SD3.

739 C Prompt for Editing dataset

740 To generate source/target text prompt from images, we leverage LLaVA [24]¹ and Llama [10]²
741 sequentially. Specifically, we extract source text description from 800 training images of DIV2K
742 dataset by giving each image and following text prompt to LLaVA.

743 "Describe the object and background in the image"

744 Then, we generate the target text prompt that has only a single different concept compared with the
745 source text prompt using Llama with the following instruction.

746 "You are an AI assistant for generating paired text prompts for real image editing
747 tasks. Your goal is to modify a given text description by replacing an object with
748 other while strictly following these rules:

- 749 • 1) Modify only one object (i.e., a single meaningful concept such as an object).
750 It could be small object.
- 751 • 2) The replacement must be significantly different from the original concept
752 but contextually appropriate. Avoid unrealistic substitutions (e.g., changing
753 "rabbit on grass" to "rocket on grass").
- 754 • 3) Ensure diversity in word choices across different modifications.
- 755 • 4) Preserve all other words exactly as they are. Do not change sentence
756 structure, introduce new elements, or modify additional details.
- 757 • 5) Do not provide any additional words—output only the modified text de-
758 scription.
- 759 • 6) Do not change or add colors. Specifically, when modifying a building,
760 change only the appearance, not the type of building (e.g., do not change
761 "building" to "church" or "lighthouse").

¹We use checkpoint from 4bit/llava-v1.5-13b-3GB

²We use Llama-3.1-8B-Instruct model.

- 7) Modify only one feature at a time. If changing an object (e.g., "starfish" to "sea turtle"), do not alter its color, shape, or other attributes.

Example:

Input: The image features a close-up of a brown dog with a blue nose. The dog is standing in a grassy field, and the background is blurred, creating a focus on the dog's face. The dog's ears are perked up, and its eyes are open, giving it a curious and attentive expression. The dog's fur is brown, and the grass in the background is green, creating a natural and vibrant scene.

Output: The image features a close-up of a brown fox with a blue nose. The fox is standing in a grassy field, and the background is blurred, creating a focus on the fox's face. The fox's ears are perked up, and its eyes are open, giving it a curious and attentive expression. The fox's fur is brown, and the grass in the background is green, creating a natural and vibrant scene."

D Discussion on counting metric

Adopting Soft Tokens Only A Few Layers As shown in section 4, our model demonstrates notable improvements in generating single and multiple objects, as well as in color and position-aware synthesis. However, it exhibits a relative decline in accurately generating the specified number of objects described in the text prompts. We hypothesize that this drop in the counting metric stems from the soft tokens excessively emphasizing textual cues, which can lead to overgeneration of object instances. To address this, we limited the use of soft tokens to the early layers of the model. As presented in table 4, applying soft tokens only to layers 1 ~ 2 preserves strong overall performance while mitigating the degradation in counting accuracy observed when soft tokens are applied across layers 1 ~ 5.

Incorporating Counting Loss during training To further enhance counting fidelity, we trained SoftREPA with a lightweight object-counting loss. Specifically, we employ a mean squared error (MSE) loss between the predicted object count, which is derived from denoised images, and ground-truth labels using the lightweight object detection module, YOLOv8 [39]. This variant, shown in the last row of table 4, shows improved counting accuracy without sacrificing overall generation quality, suggesting that it effectively reduces the tendency to replicate objects due to textual overemphasis.

Model	Layers	GenEval						
		Mean↑	Single↑	Two↑	Counting↑	Colors↑	Position↑	Color Attribution↑
SD3		0.68	<u>0.99</u>	0.86	<u>0.56</u>	0.85	0.27	0.55
SD3 (Ours)	1	0.70	<u>0.99</u>	0.91	0.51	0.89	<u>0.31</u>	<u>0.58</u>
SD3 (Ours)	1~2	<u>0.69</u>	1.00	0.89	0.50	0.88	0.34	0.56
SD3 (Ours)	1~5	0.70	1.00	0.95	0.29	0.92	0.34	0.68
SD3 (Ours) + count	1~5	<u>0.69</u>	<u>0.99</u>	0.88	0.59	0.86	0.25	0.55

Table 4: Additional quantitative comparison on the GenEval [9] benchmark. **Bold** indicates the best performance, and underline denotes the second best. "Layers" refers to the transformer layers where soft tokens are applied.

E Additional Results on T2I Generation

Additional qualitative results for SD1.5 and SDXL on the COCO-val5K dataset are presented in fig. 7 and fig. 8, respectively. Furthermore, qualitative examples on the GenEval benchmark are provided in fig. 9.

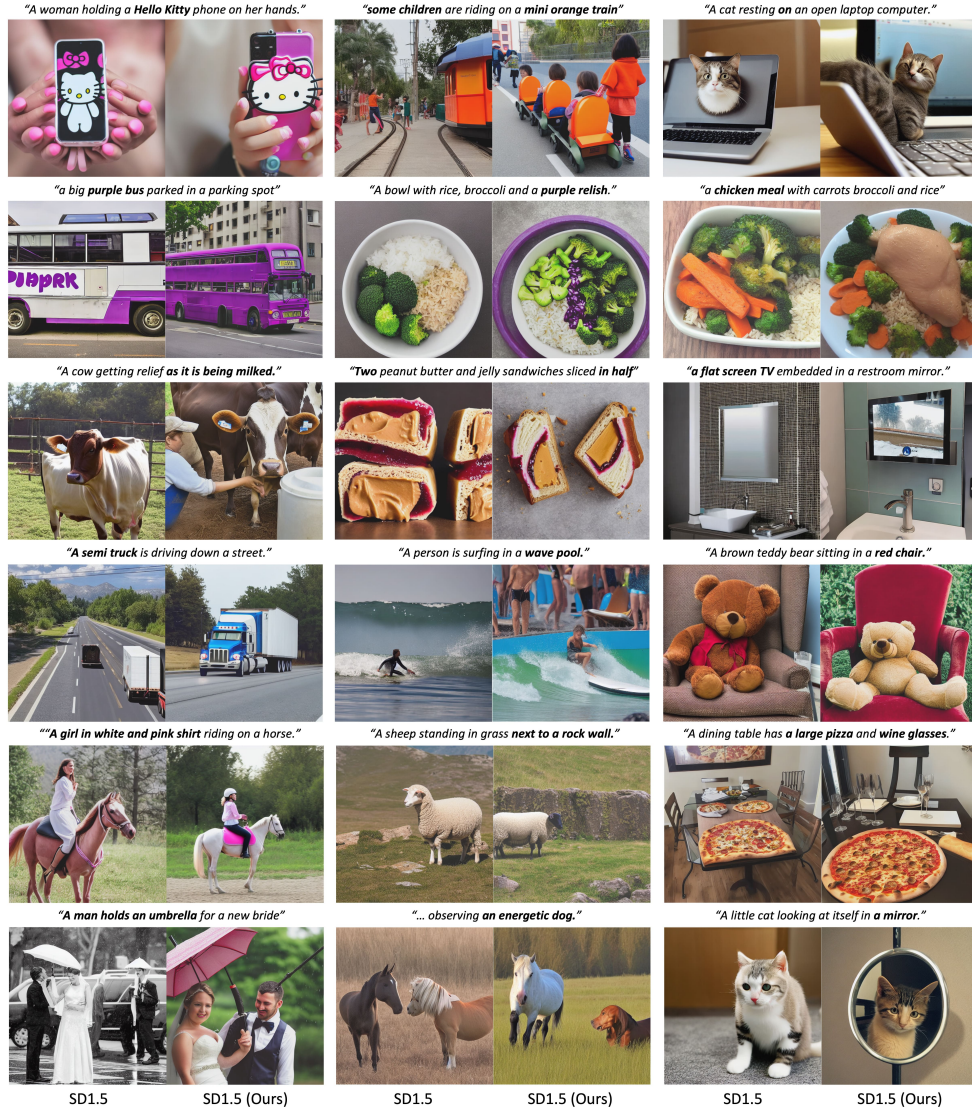


Figure 7: The qualitative results of text-to-image generation comparing SD1.5 and SD1.5 with proposed method. The given text is from COCO dataset.

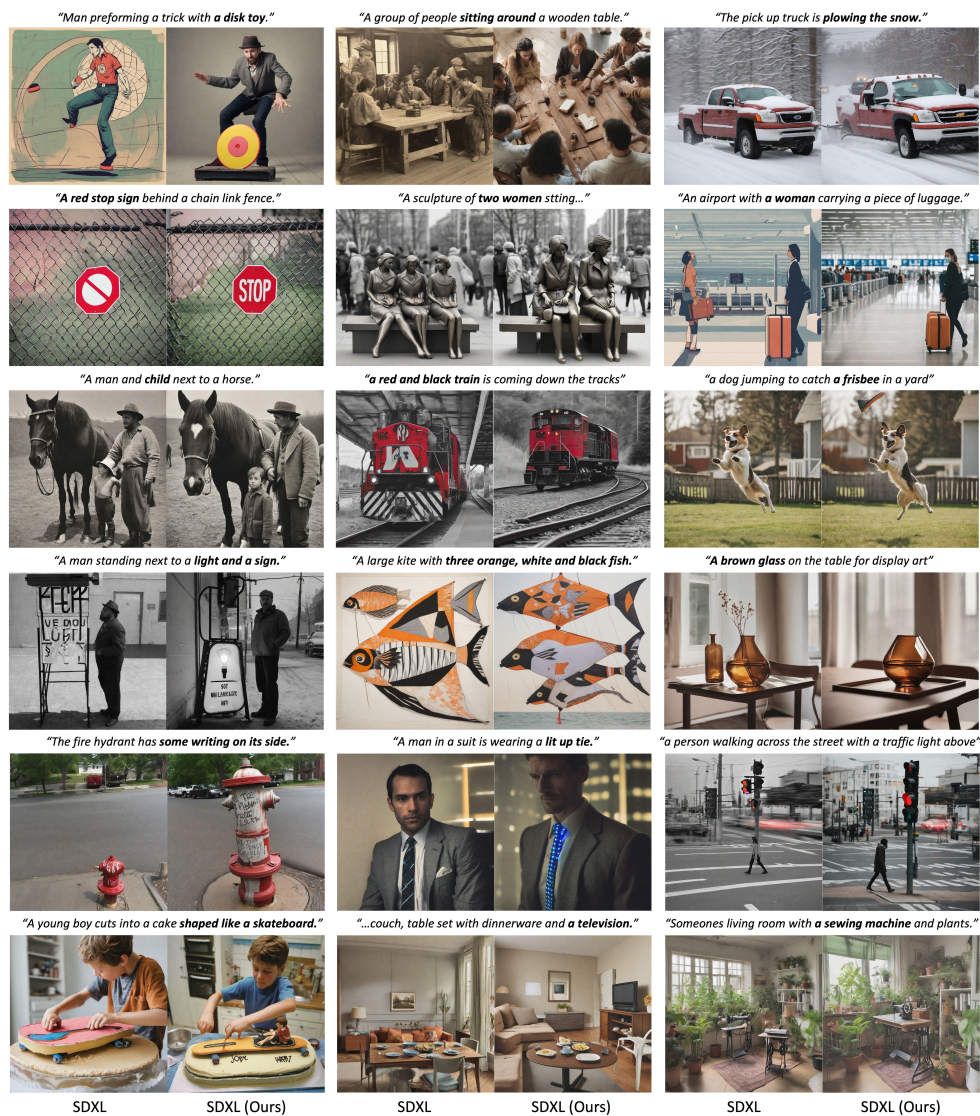


Figure 8: The qualitative results of text-to-image generation comparing SDXL and SDXL with proposed method. The given text is from COCO dataset.

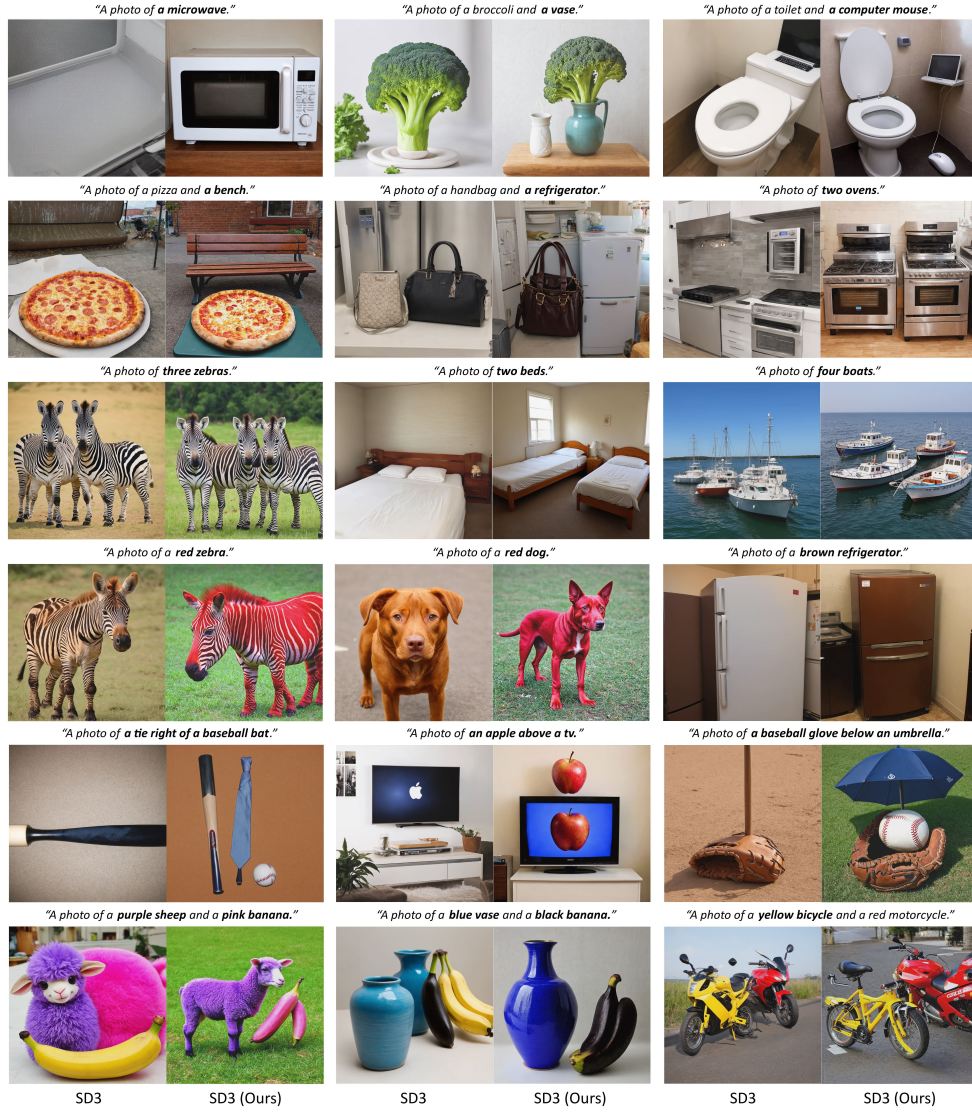


Figure 9: The qualitative results of text-to-image generation comparing SD3 and SD3 with proposed method. The given text is from GenEval dataset.

F Additional Results on Image Editing

Qualitative comparisons of SoftREPA across various editing methods, including PnP [37], MasaCtrl [4] with DDIM [34], Direct Inversion [14], and FlowEdit [19], on the PIEBench dataset are presented in fig. 11, corresponding to the quantitative results in table 2.

Quantitative results on the DIV2K and Cat2Dog datasets under varying target CFG scales are reported in table 5 and table 6, respectively. The corresponding qualitative results for different CFG scales are also shown in table 5. Additionally, qualitative examples across different numbers of editing steps are provided in fig. 12 and fig. 10.

Model	NMAX	target CFG	Human Preference		Text Alignment		Image Quality	
			ImageReward↑	PickScore↑	CLIP↑	HPS↑	FID↓	LPIPS↓
FlowEdit	33	13.5	0.380	21.749	0.261	0.255	52.038	0.154
FlowEdit	33	16.5	0.466	21.824	0.263	0.259	55.371	0.180
FlowEdit	33	19.5	0.510	21.869	0.265	0.261	58.069	0.200
FlowEdit +Ours	30	9	0.466	21.884	0.263	0.260	52.633	0.149
FlowEdit +Ours	30	11	0.564	21.985	0.268	0.265	56.900	0.178
FlowEdit +Ours	30	13	0.627	22.050	0.272	0.270	59.710	0.201

Table 5: Quantitative results of image editing regarding image quality and target text alignment of generated images with various target CFG scales on **DIV2K** dataset.

Model	NMAX	target CFG	Human Preference		Text Alignment		Image Quality	
			ImageReward↑	PickScore↑	CLIP↑	HPS↑	FID↓	LPIPS↓
FlowEdit	33	13.5	0.937	20.726	0.225	0.266	197.1	0.199
FlowEdit	33	16.5	0.908	20.641	0.229	0.272	201.05	0.217
FlowEdit	33	19.5	0.882	20.572	0.232	0.276	202.92	0.234
FlowEdit +Ours	30	7	1.111	21.143	0.226	0.269	198.71	0.173
FlowEdit +Ours	30	9	1.144	21.221	0.234	0.283	208.18	0.204
FlowEdit +Ours	30	11	1.144	21.224	0.239	0.290	214.72	0.230
FlowEdit +Ours	30	13	1.135	21.200	0.241	0.293	219.35	0.252

Table 6: Quantitative results of image editing regarding image quality and target text alignment of generated images with various target CFG scales on **Cat2Dog** dataset.

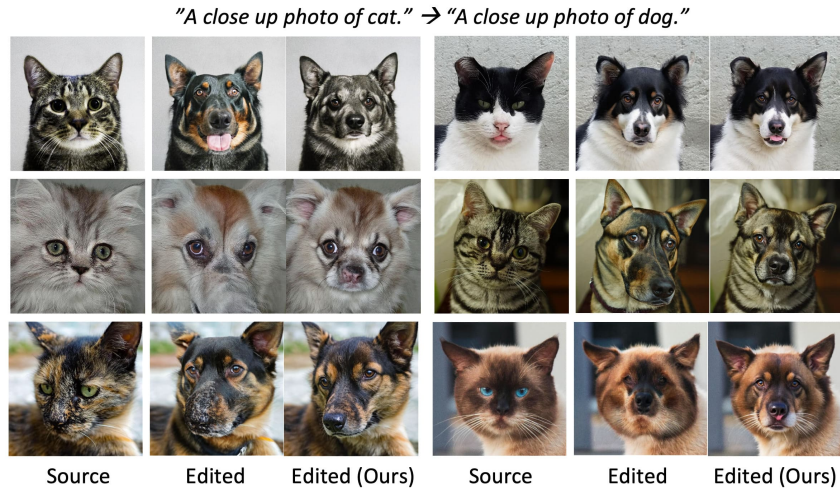


Figure 10: Additional qualitative results of text-guided image editing on Cat2Dog dataset.



Figure 11: The qualitative results of various editing methods including PnP, MasaCtrl, FlowEdit with DDIM and Direct inversion on PIEBench.

G Additional Results on Ablation Study

Quantitative results for varying the number of layers and soft tokens are shown in table 7, with corresponding qualitative examples in fig. 13 and fig. 14. As seen in fig. 13, text prompts are increasingly emphasized as more layers adopt soft tokens. The same input image is used in fig. 14 to illustrate the effect of different token counts.

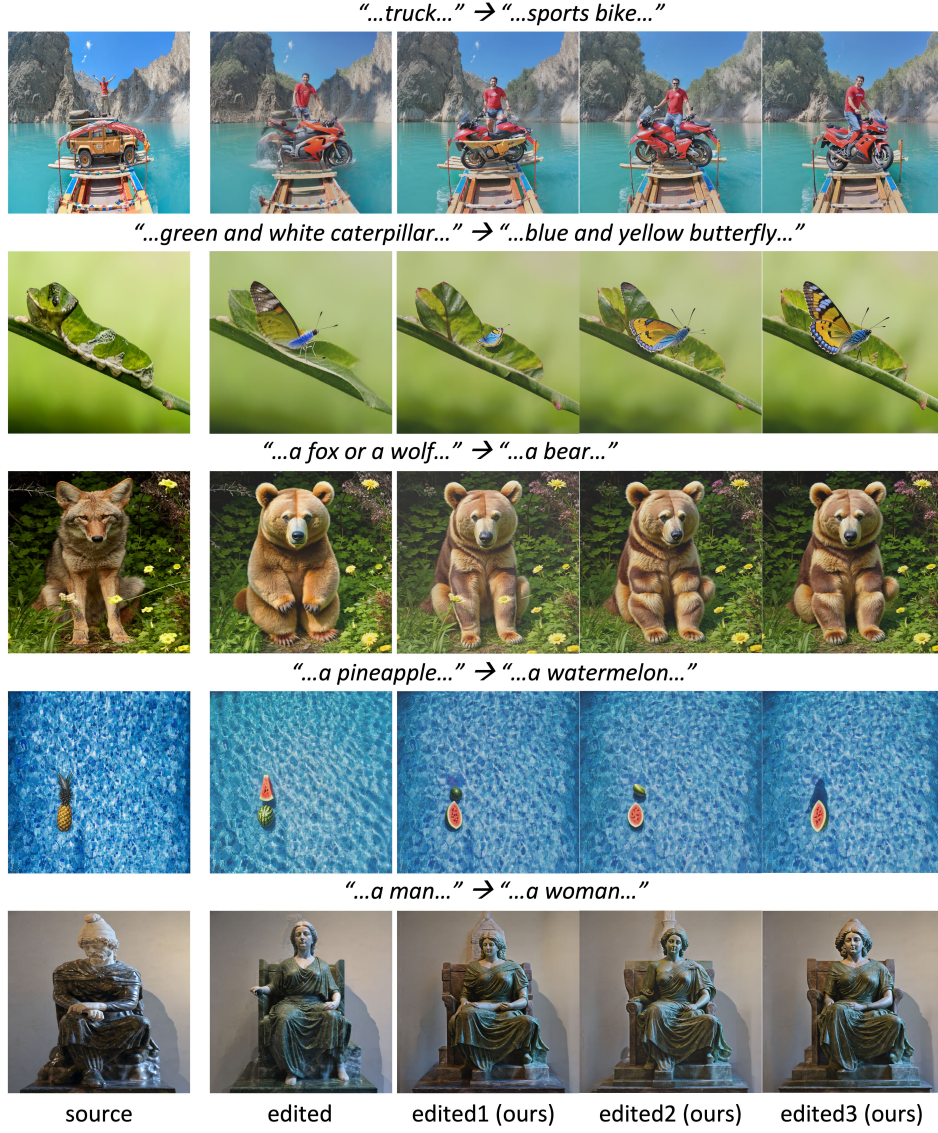


Figure 12: The additional qualitative results on editing using FlowEdit. The edited1-3 using soft tokens are vary on the number of editing steps, 27, 29, 30 out of 50 total timesteps, respectively.

# of tokens	# of layers	Human Preference		Text Alignment		Image Quality	
		ImageReward↑	PickScore↑	CLIP↑	HPS↑	FID↓	LPIS↓
8	2	0.993	22.556	0.263	0.286	72.253	0.428
8	4	1.009	22.464	0.266	0.285	72.308	0.424
8	6	0.984	22.403	0.267	0.286	73.840	0.427
8	7	0.944	22.197	0.269	0.280	74.546	0.425
1	5	1.054	22.492	0.272	0.283	58.430	0.426
4	5	1.063	22.493	0.269	0.287	60.766	0.429
8	5	1.056	22.516	0.268	0.288	62.644	0.431
16	5	0.801	22.095	0.265	0.278	60.743	0.431
32	5	0.675	21.876	0.257	0.275	61.623	0.426

Table 7: Quantitative results of image generation ablation study on the number of soft tokens and the number of layers. The evaluation is conducted on COCO val 1K dataset using SD3 backbone.



Figure 13: Qualitative results of the ablation study on the **number of layers** adopting soft tokens.



Figure 14: Qualitative results of the ablation study on the **number of tokens** adopting soft tokens.