Table 1. Effect of the Attention Entropy Scale Factor λ on Image Quality and Positional Information. We report two metrics for different scale factors: 1) Position Loss: the loss value derived from our position information quantitative analysis technique, measuring the amount of positional information of the feature map; and 2) ImageReward to assess image quality. The metrics are averaged over 100 randomly sampled prompts from the Laion-5B dataset. User Study: for each prompt, five participants independently chose the image with the highest visual quality among the six experiments. Each prompt was treated as an independent voting instance, and we calculated the selection probability of each experiment across all votes. We then calculated the selection probability for each experiment across all prompts. All experiments are conducted on *SD-2.1* with 1024×1024 resolution.

λ	Position Loss (Total) ↓	Position Loss (central 512^2) ↓	ImageReward ↑	User Study ↑
$\sqrt{\frac{1}{d}}$ (Training)	0.0225	0.0438	-0.0453	1.2%
$\sqrt{\frac{\log_T N}{d}}$ (AttenEtro Align [3])	0.0225	0.0425	-0.3116	18%
$1.05 * \sqrt{\frac{\log_T N}{d}}$	**0.0224**	0.0420	-0.2676	20.6%
$1.10 * \sqrt{\frac{\log_T N}{d}}$	0.0225	0.0415	-0.2454	**39.2%**
$1.15 * \sqrt{\frac{\log_T N}{d}}$	0.0225	**0.0408**	-0.2466	13.0%
$1.2 * \sqrt{\frac{\log_T N}{d}}$	0.0225	0.0418	**-0.2312**	8.0%

Table 2. Quantitative comparison across methods under SD1.5, SD2.1, and SD-XL. Because of the time constraint, we are unable to report the performance of demofusion in SD1.5 and SD 2.1. All experiments are at a resolution of 1024×1024 in SD-1.5 and SD2.1, 2048×2048 in SD-XL. User Study: five participants independently rated images on a scale from 1 (lowest) to 5 (highest) for their image visual quality based on 40 randomly selected prompts per method. The table below reports the average user scores.

Method	**FID ↓**	**KID ↓**	**IS ↑**	**IR ↑**	**HPS ↑**	**CR ↓**	**User Study↑**
SD-1.5							
DI	100.21	0.0022	11.38	-0.1708	0.1942	30.38	1.2
ScaleCrafter	98.40	0.0047	10.09	0.4402	0.1987	29.57	3.1
FouriScale	**83.64**	**0.0015**	11.47	0.5858	**0.1997**	29.28	4.2
PBC (Ours)	87.50	0.0022	**11.76**	**0.5902**	0.1967	**29.23**	**4.3**
SD-2.1							
DI	95.63	0.0069	**11.15**	0.0420	0.1956	30.67	1.1
ScaleCrafter	106.04	0.0076	9.82	0.5731	0.2014	30.14	3.4
FouriScale	89.06	0.0054	10.81	**0.7579**	**0.2023**	29.54	4.4
PBC (Ours)	**86.32**	**0.0026**	11.01	0.7457	0.2018	**29.43**	**4.6**
SD-XL							
DI	97.02	0.0092	9.25	0.645	20.76	30.30	1.1
ScaleCrafter	83.59	0.0067	9 .53	1.078	21.15	29.50	3.4
DemoFusion	**76.40**	0.0088	**11.07**	**1.090**	**21.33**	29.47	2.9
FouriScale	90.28	0.0103	9.38	1.023	21.13	29.61	4.0
PBC (Ours)	91.98	**0.0067**	9.95	1.036	21.20	**29.42**	**4.0**

Table 3. User study comparison on 4096×4096 resolution by SD-XL. Five participants independently rated images on a scale from 1 (lowest) to 5 (highest) for their image visual quality based on 20 randomly selected prompts per method.

	DI	ScaleCrafter	DemoFusion	FouriScale	PBC (Ours)
User Study ↑	1.1	<u>3.8</u>	3.3	3.6	**4.1**

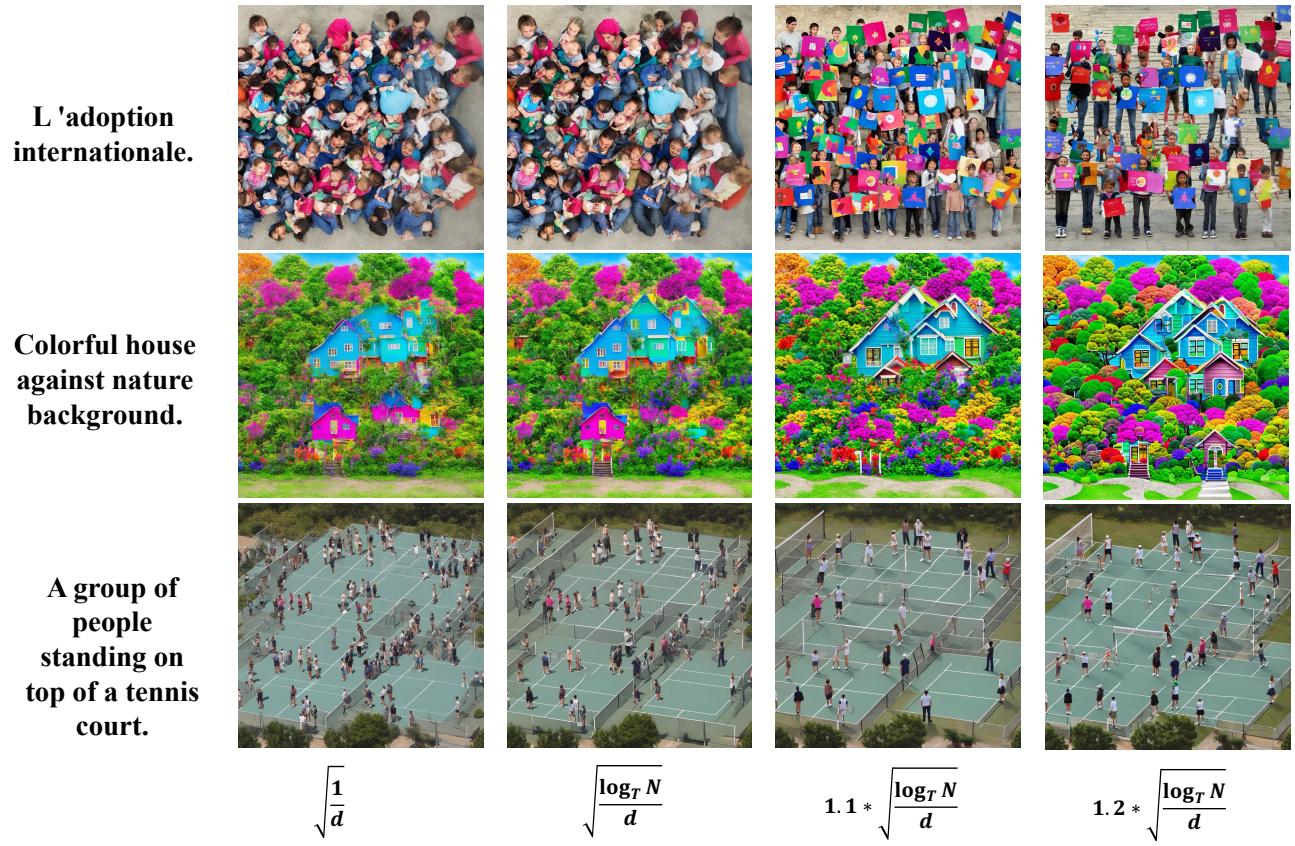


Figure 1. Qulitative Comparison between different Scale Factor λ. Reducing the attention entropy beyond the entropy alignment point $\lambda = \sqrt{\frac{\log_T N}{d}}$ leads to image quality enhancement.

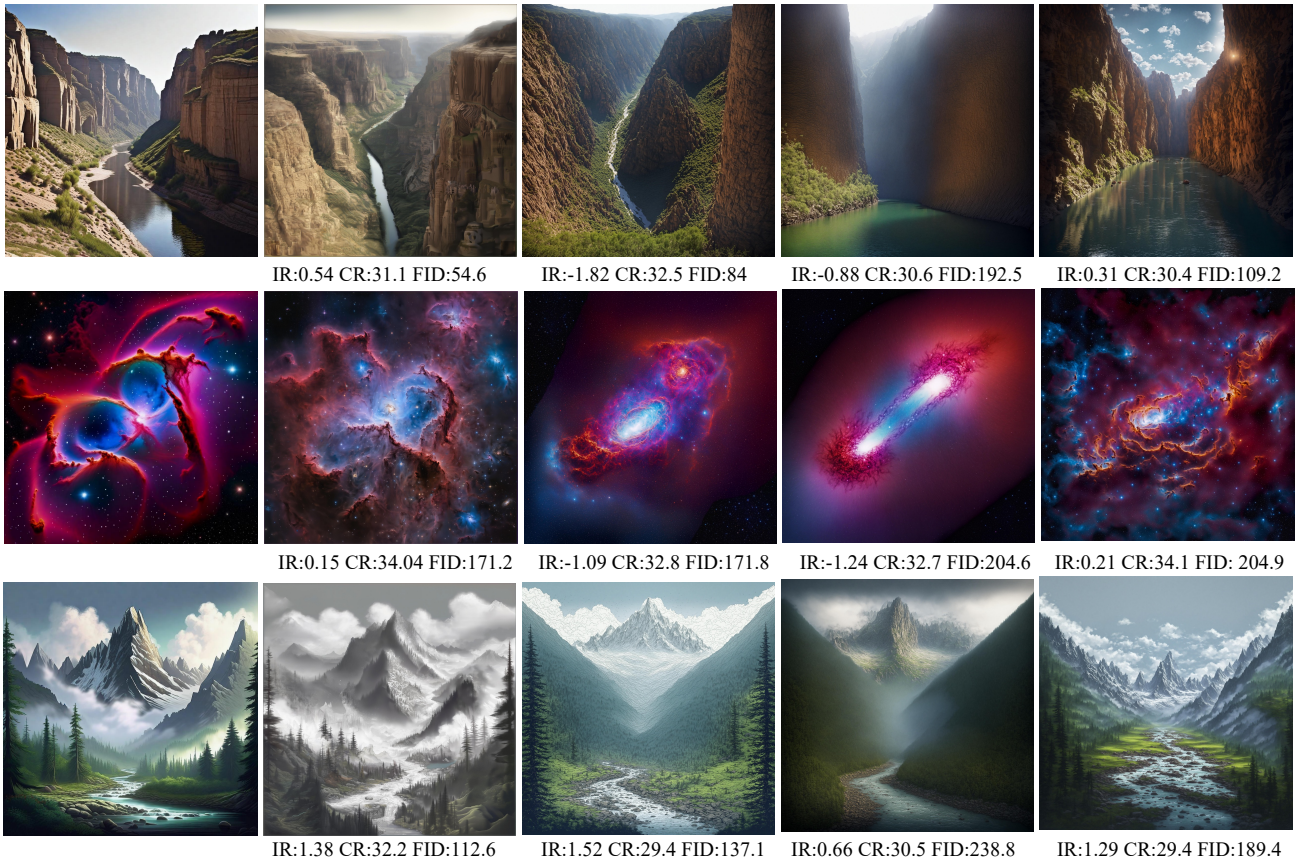


Figure 2. Comparison at 4096×4096 resolution by SD-XLF. We report three metrics for each image: ImageReward (IR) ↑, Content Richness (CR) ↓, and FID ↓. Our generated images demonstrate higher visual quality but yield worse FID scores, as the additional content they generate results in greater deviation from the training-resolution images.

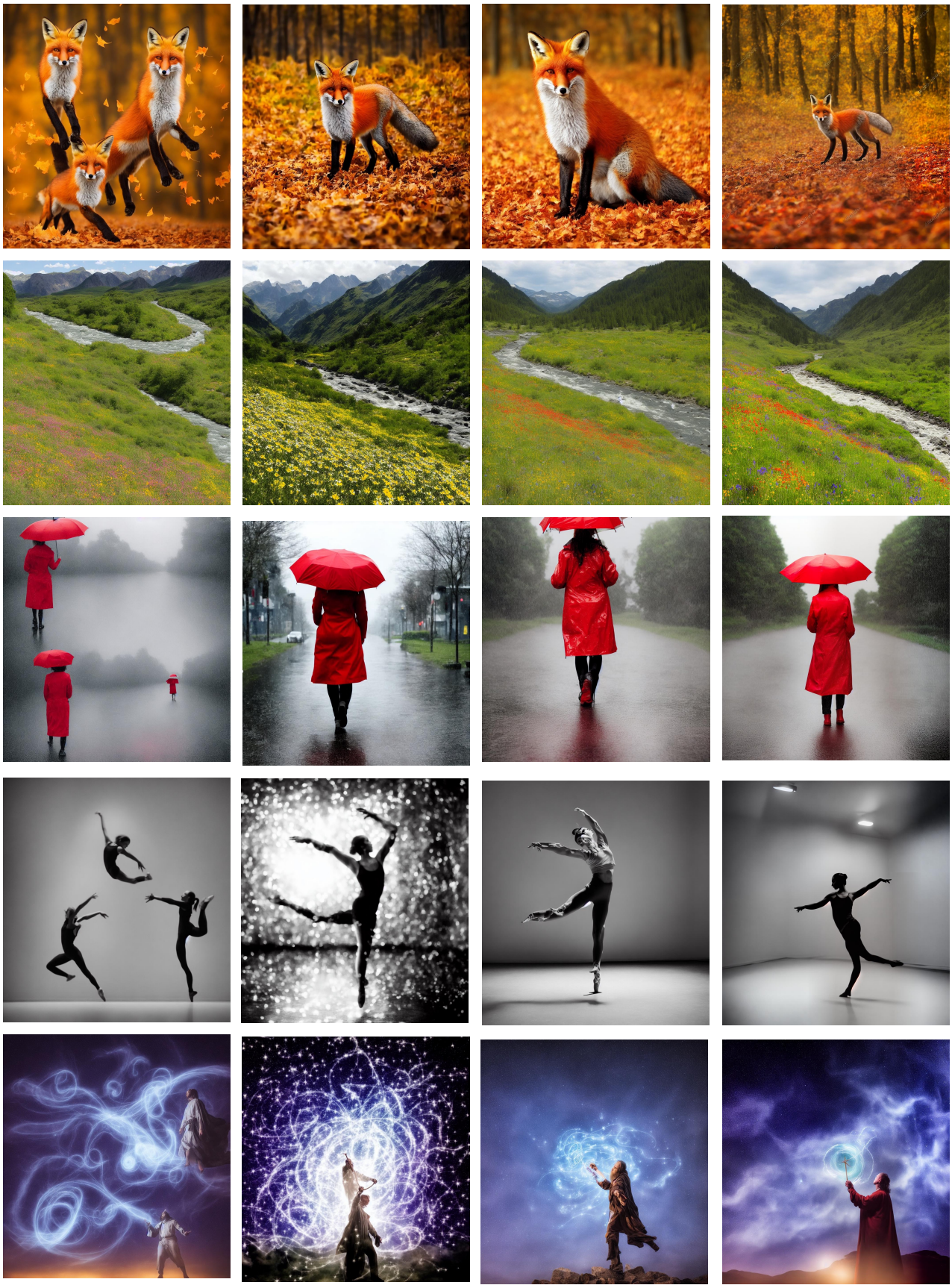


Figure 3. Qualitative experiments of SD-1.5. All experiments are at a resolution of 1024×1024.

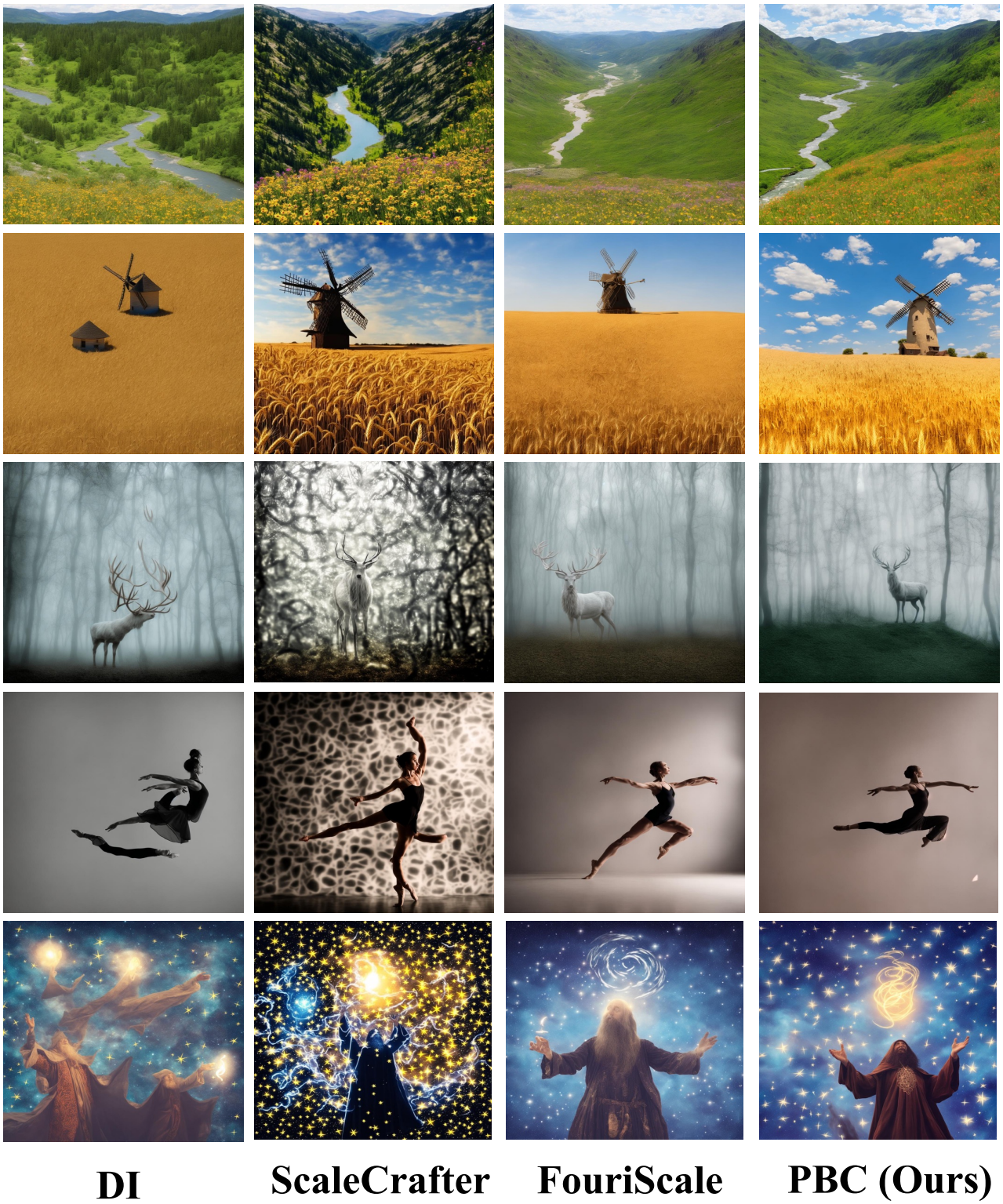
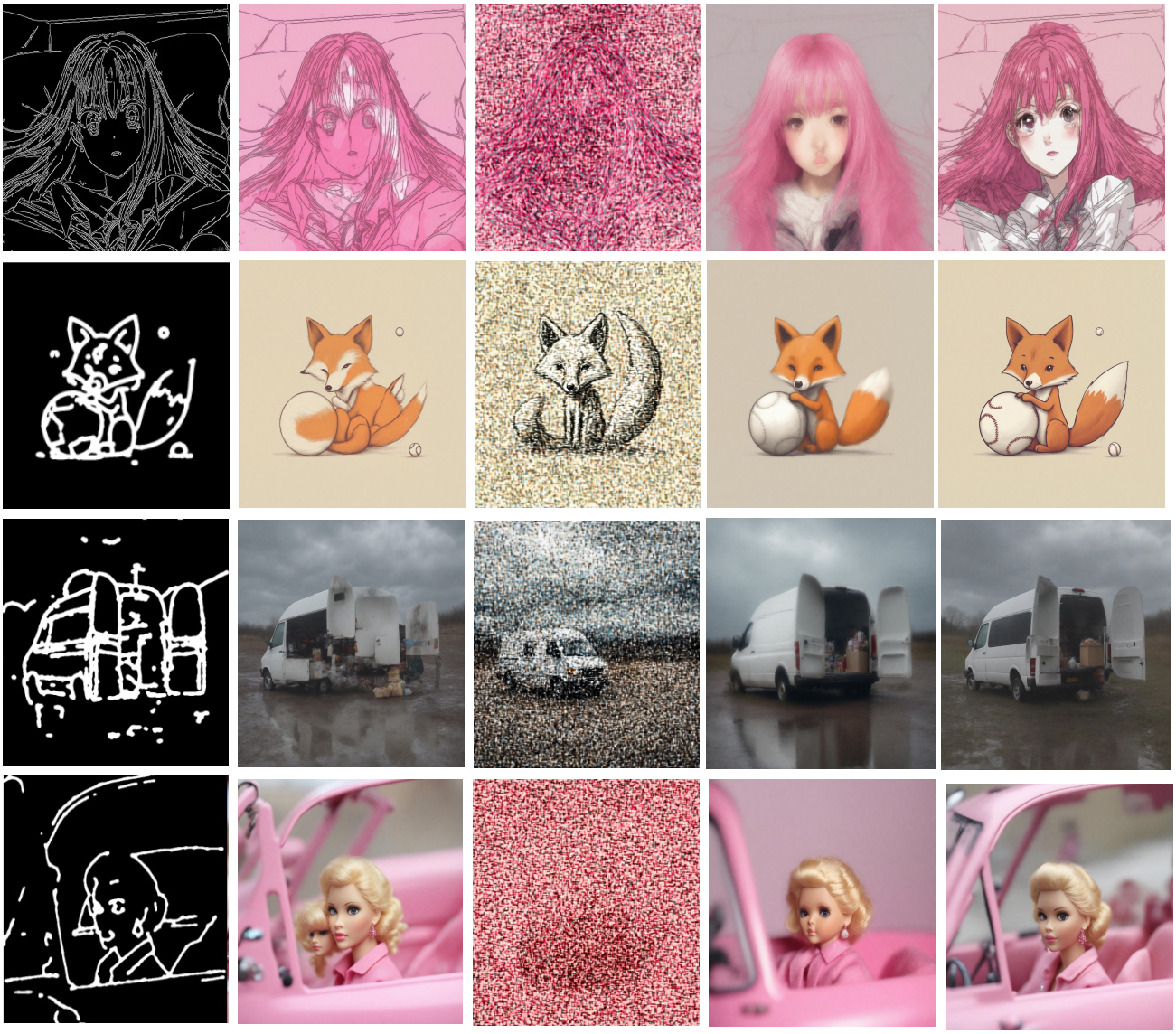


Figure 4. Qualitative experiments of SD-2.1. All experiments are at a resolution of 1024×1024.



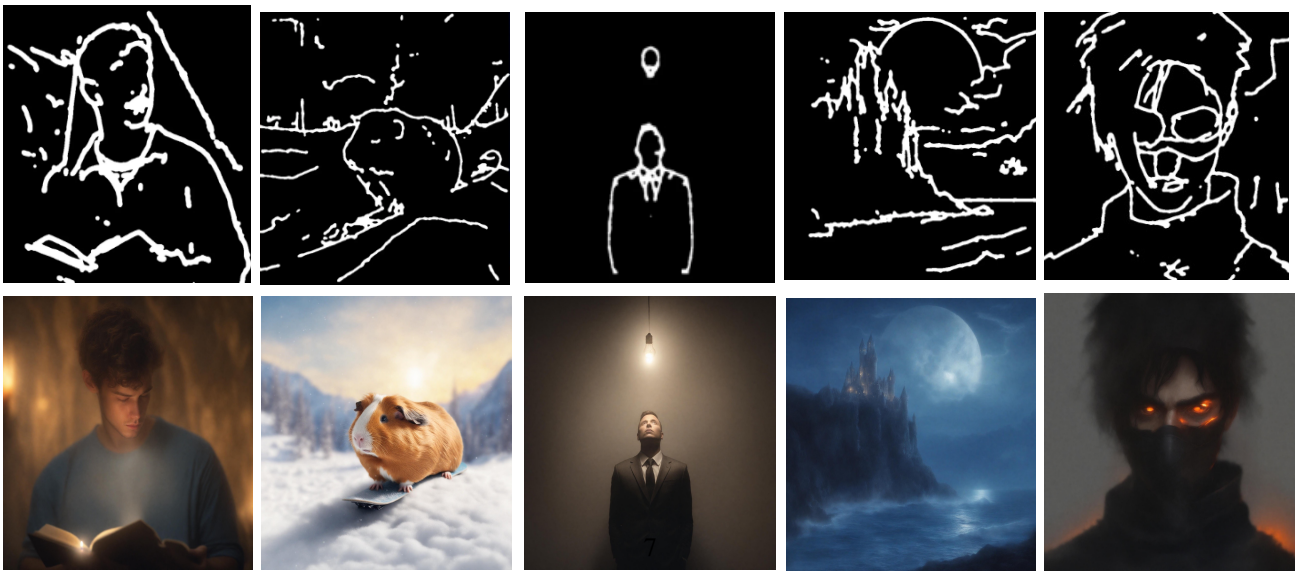


Figure 5. ControlNet integration experiment. All experiments are conducted by SD-XL at a resolution of 2048×2048.

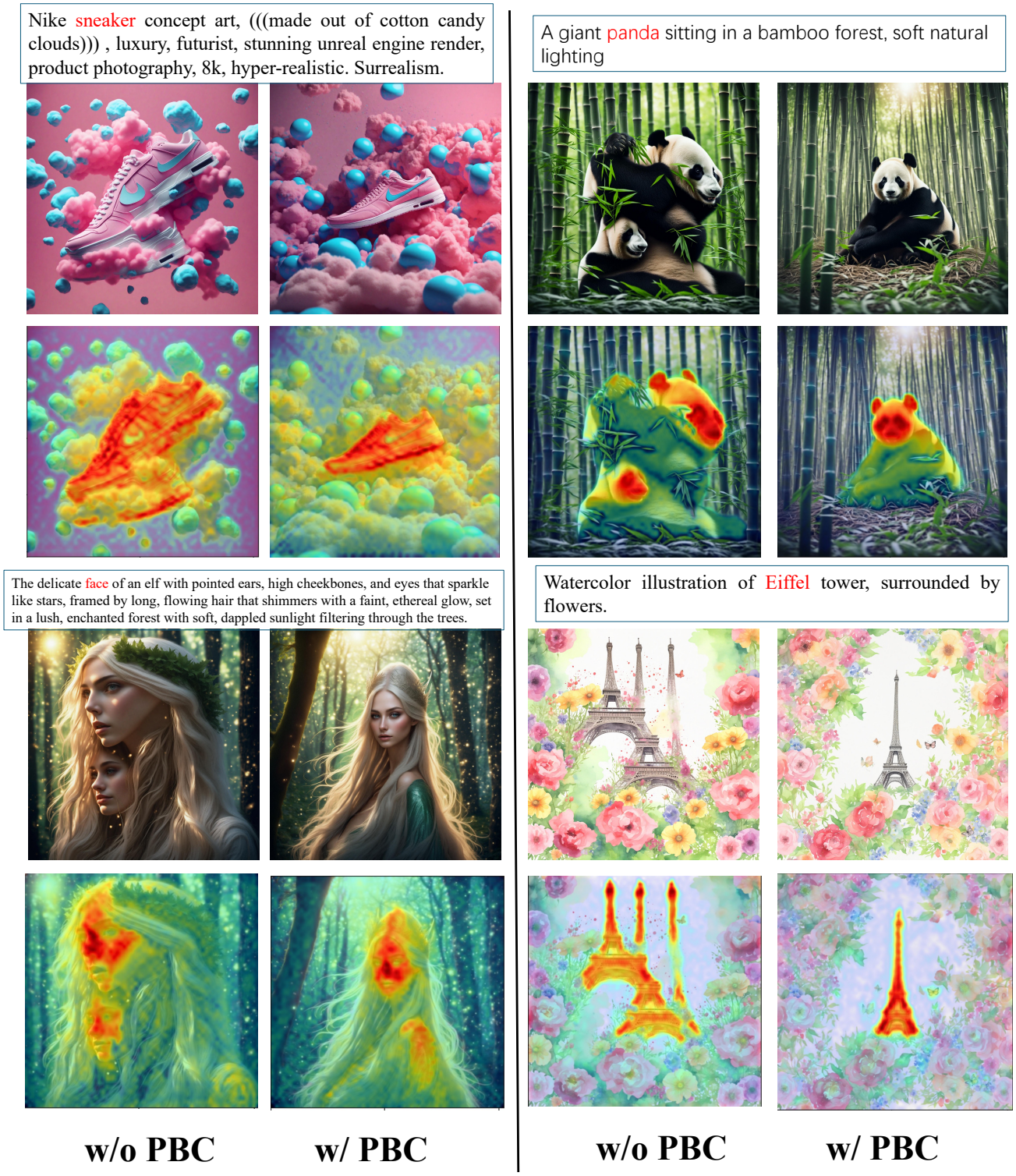


Figure 6. Feature Map visualization. We visualize the key text token's attention map to the whole feature map, which demonstrated the effectiveness of our method.

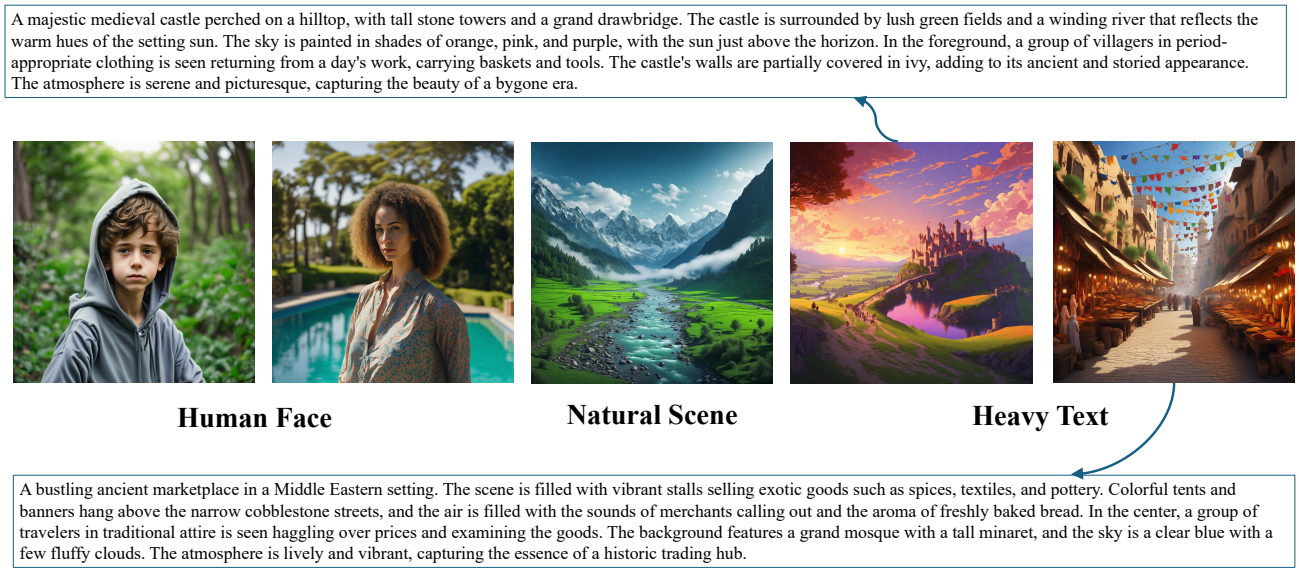


Figure 7. Qualitative performance across different categories. All experiments are conducted by SD-XL at a resolution of 2048×2048.