

Table 1. Effect of the Attention Entropy Scale Factor  $\lambda$  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.

$\lambda$	Position Loss (Total) $\downarrow$	Position Loss (central 512 $^2$ ) $\downarrow$	ImageReward $\uparrow$	User Study $\uparrow$
$\sqrt{\frac{1}{d}}$ (Training)	0.0225	0.0438	-0.0453	1.2%
$\sqrt{\frac{\log_T N}{d}}$ (AttenEntro Align [3])	0.0225	0.0425	-0.3116	18%
$1.05 * \sqrt{\frac{\log_T N}{d}}$	<b>0.0224</b>	0.0420	-0.2676	20.6%
$1.10 * \sqrt{\frac{\log_T N}{d}}$	0.0225	0.0415	-0.2454	<b>39.2%</b>
$1.15 * \sqrt{\frac{\log_T N}{d}}$	0.0225	<b>0.0408</b>	-0.2466	13.0%
$1.2 * \sqrt{\frac{\log_T N}{d}}$	0.0225	0.0418	<b>-0.2312</b>	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 $\downarrow$	KID $\downarrow$	IS $\uparrow$	IR $\uparrow$	HPS $\uparrow$	CR $\downarrow$	User Study $\uparrow$
<b>SD-1.5</b>							
DI	100.21	0.0022	11.38	-0.1708	0.1942	30.38	1.2
ScaleCrafter	98.40	0.0047	10.09	0.4402	<u>0.1987</u>	29.57	3.1
Fouriscale	<b>83.64</b>	<b>0.0015</b>	<u>11.47</u>	<u>0.5858</u>	<b>0.1997</b>	<u>29.28</u>	4.2
PBC (Ours)	<u>87.50</u>	<u>0.0022</u>	<b>11.76</b>	<b>0.5902</b>	0.1967	<b>29.23</b>	<b>4.3</b>
<b>SD-2.1</b>							
DI	95.63	0.0069	<b>11.15</b>	0.0420	0.1956	30.67	1.1
ScaleCrafter	106.04	0.0076	9.82	0.5731	0.2014	30.14	3.4
Fouriscale	<u>89.06</u>	<u>0.0054</u>	10.81	<b>0.7579</b>	<b>0.2023</b>	<u>29.54</u>	4.4
PBC (Ours)	<b>86.32</b>	<b>0.0026</b>	<u>11.01</u>	<u>0.7457</u>	<u>0.2018</u>	<b>29.43</b>	<b>4.6</b>
<b>SD-XL</b>							
DI	97.02	0.0092	9.25	0.645	20.76	30.30	1.1
ScaleCrafter	<u>83.59</u>	<u>0.0067</u>	<u>9.53</u>	<u>1.078</u>	21.15	29.50	3.4
DemoFusion	<b>76.40</b>	0.0088	<b>11.07</b>	<b>1.090</b>	<b>21.33</b>	<u>29.47</u>	2.9
Fouriscale	90.28	0.0103	9.38	1.023	21.13	29.61	4.0
PBC (Ours)	91.98	<b>0.0067</b>	<u>9.95</u>	1.036	<u>21.20</u>	<b>29.42</b>	<b>4.0</b>

110  
111  
112  
113  
114115 *Table 3.* User study comparison on 4096x4096 resolution by SD-XL. Five participants independently rated images on a scale from 1  
116 (lowest) to 5 (highest) for their image visual quality based on 20 randomly selected prompts per method.117  
118  
119

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

120  
121  
122  
123  
124  
125  
126  
127  
128129  
130  
131**L 'adoption  
internationale.**132  
133  
134  
135  
136  
137  
138**Colorful house  
against nature  
background.**139  
140  
141  
142**A group of  
people  
standing on  
top of a tennis  
court.**143  
144  
145  
146

$$\sqrt{\frac{1}{d}}$$

$$\sqrt{\frac{\log_T N}{d}}$$

$$1.1 * \sqrt{\frac{\log_T N}{d}}$$

$$1.2 * \sqrt{\frac{\log_T N}{d}}$$

153  
154  
155156 *Figure 1.* Qualitative Comparison between different Scale Factor  $\lambda$ . Reducing the attention entropy beyond the entropy alignment point  
157  $\lambda = \sqrt{\frac{\log_T N}{d}}$  leads to image quality enhancement.  
158159  
160  
161  
162  
163  
164

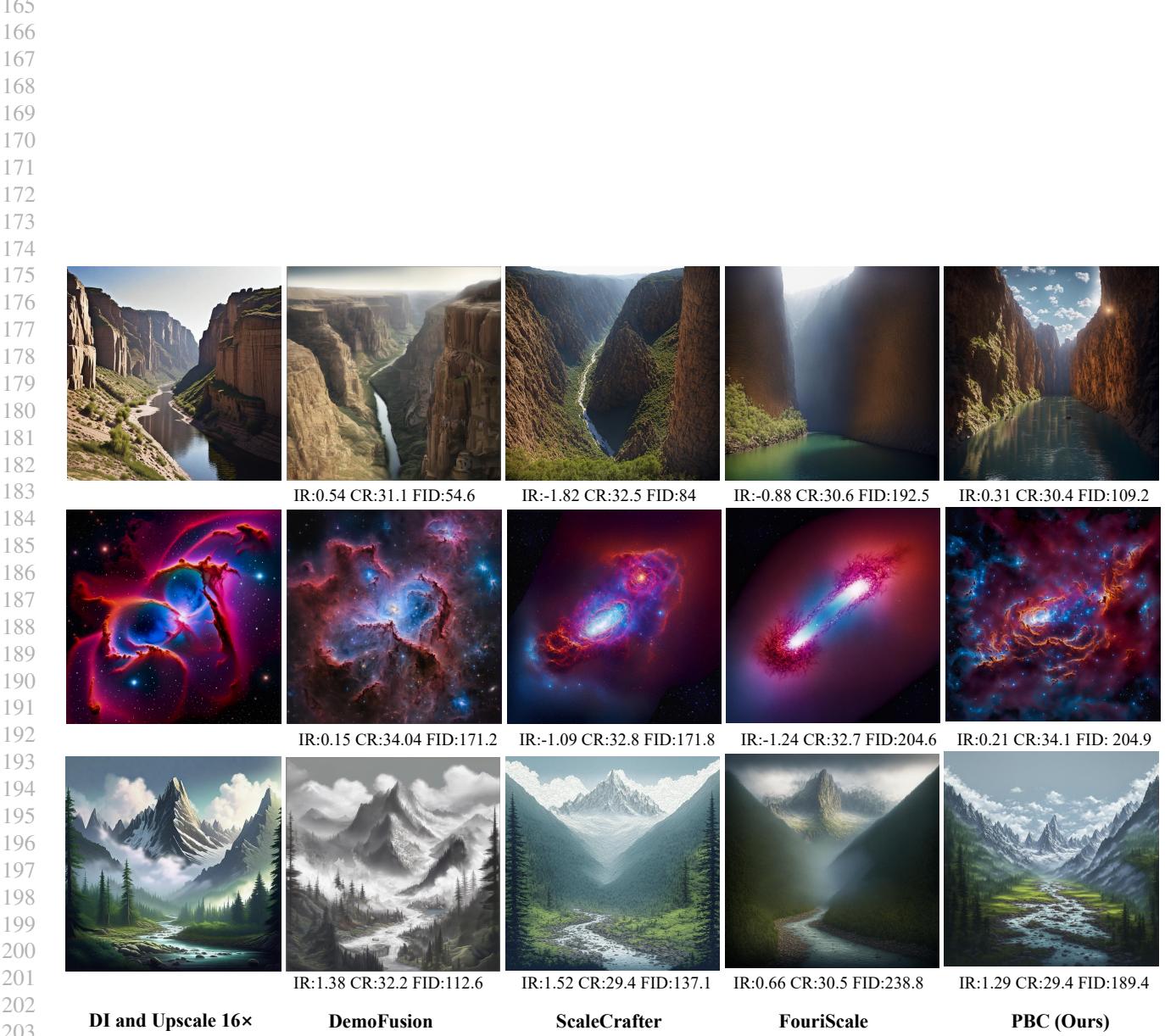
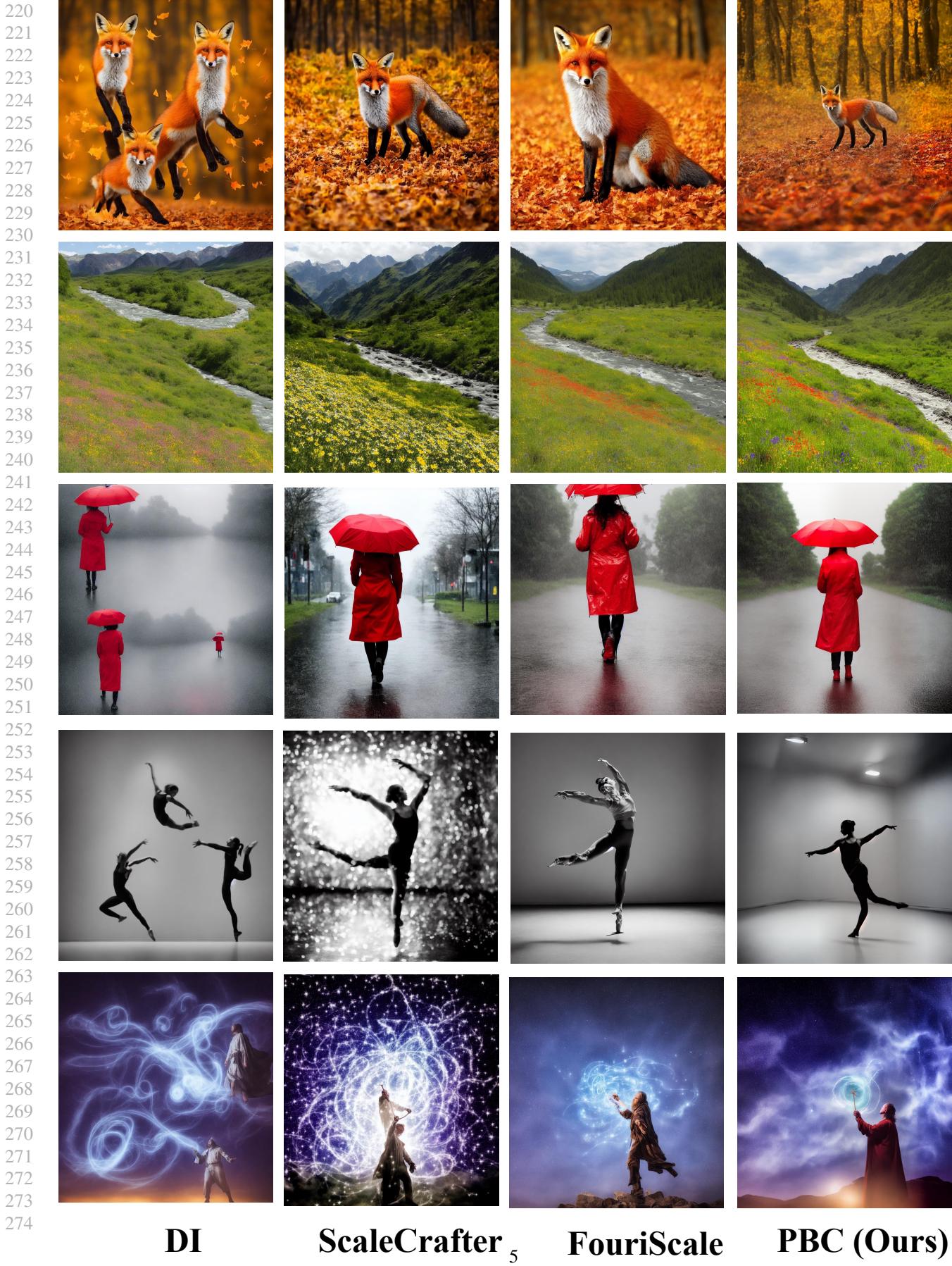


Figure 2. Comparison at 4096x4096 resolution by SD-XLF. We report three metrics for each image: ImageReward (IR)  $\uparrow$ , Content Richness (CR)  $\downarrow$ , and FID  $\downarrow$ . Our generated images demonstrate higher visual quality but yield worse FID scores, as the additional content they generate results in greater deviation from the training-resolution images.



**DI**

**ScaleCrafter**

**FouriScale**

**PBC (Ours)**

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

275  
276  
277  
278  
279  
280  
281  
282  
283  
284  
285  
286  
287  
288  
289  
290  
291  
292  
293  
294  
295  
296  
297  
298  
299  
300  
301  
302  
303  
304  
305  
306  
307  
308  
309  
310  
311  
312  
313  
314  
315  
316  
317  
318  
319  
320  
321



**DI**

**ScaleCrafter**

**FouriScale**

**PBC (Ours)**

322  
323  
324  
325  
326  
327  
328  
329

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

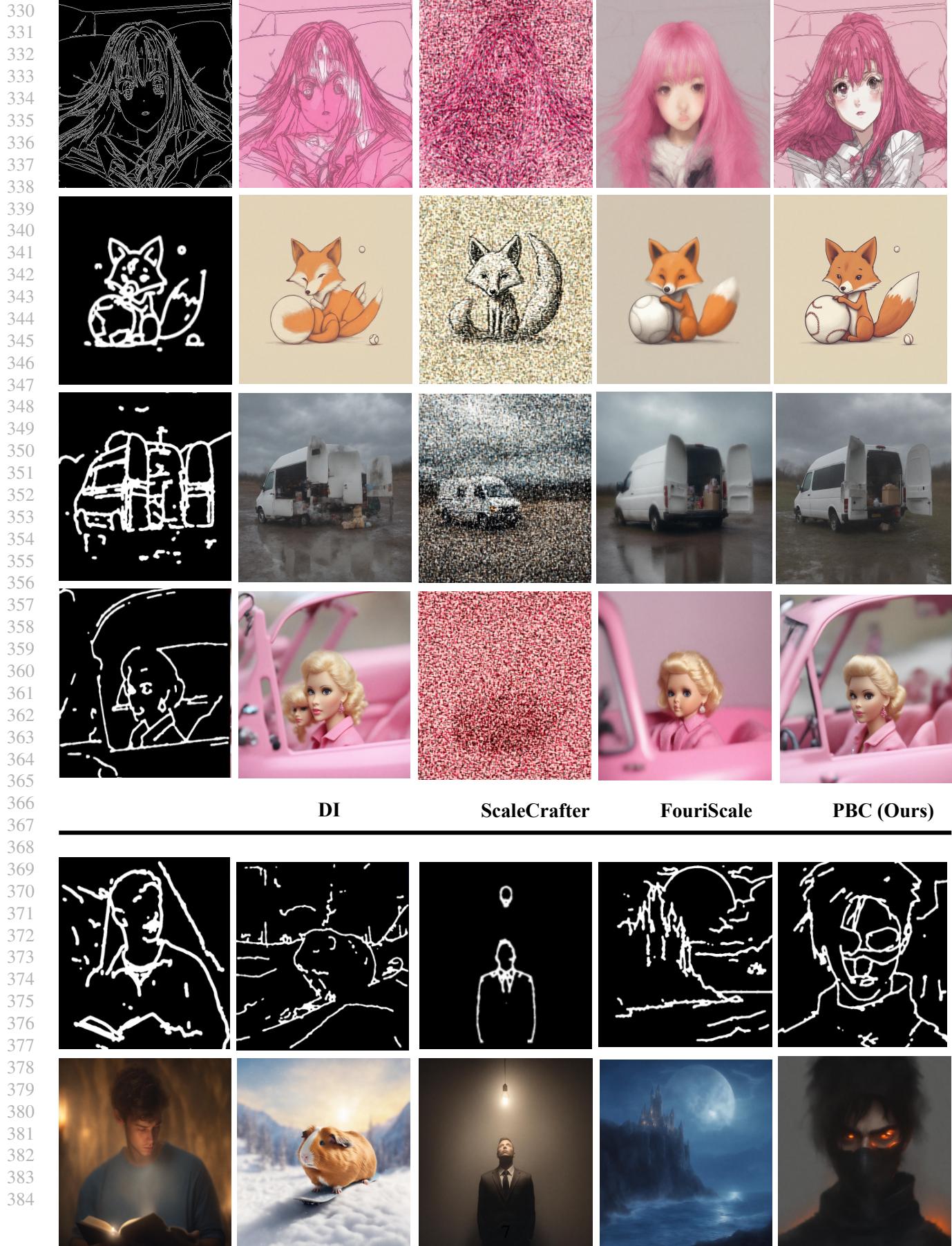
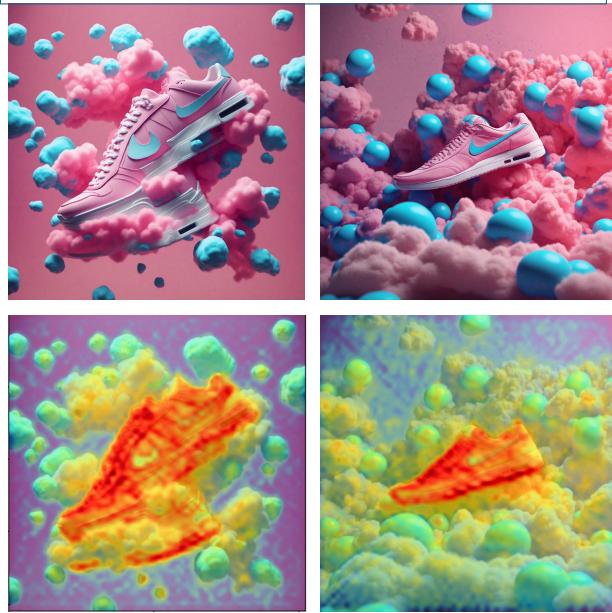


Figure 5. ControlNet integration experiment. All experiments are conducted by SD-XL at a resolution of 2048×2048.

385  
386  
387  
388  
389  
390  
391  
392  
393  
394  
395  
396  
397  
398  
399  
400  
401  
402  
403  
404  
405  
406  
407  
408

Nike **sneaker** concept art, (((made out of cotton candy clouds))) , luxury, futurist, stunning unreal engine render, product photography, 8k, hyper-realistic. Surrealism.



A giant **panda** sitting in a bamboo forest, soft natural lighting



409  
410  
411  
412  
413  
414  
415  
416  
417  
418  
419  
420  
421  
422  
423  
424  
425  
426  
427  
428  
429  
430  
431  
432  
433  
434

The delicate **face** of an elf with pointed ears, high cheekbones, and eyes that sparkle like stars, framed by long, flowing hair that shimmers with a faint, ethereal glow, set in a lush, enchanted forest with soft, dappled sunlight filtering through the trees.



**w/o PBC**

**w/ PBC**

Watercolor illustration of **Eiffel** tower, surrounded by flowers.



**w/o PBC**

**w/ PBC**

435 *Figure 6.* Feature Map visualization. We visualize the key text token’s attention map to the whole feature map, which demonstrated the  
436 effectiveness of our method.  
437  
438  
439

440  
441  
442  
443  
444  
445  
446  
447  
448  
449  
450  
451  
452  
453  
454  
455  
456

457 A majestic medieval castle perched on a hilltop, with tall stone towers and a grand drawbridge. The castle is surrounded by lush green fields and a winding river that reflects the warm hues of the setting sun. The sky is painted in shades of orange, pink, and purple, with the sun just above the horizon. In the foreground, a group of villagers in period-appropriate clothing is seen returning from a day's work, carrying baskets and tools. The castle's walls are partially covered in ivy, adding to its ancient and storied appearance.  
458 The atmosphere is serene and picturesque, capturing the beauty of a bygone era.



460  
461  
462  
463  
464  
465  
466  
467  
468  
469  
470  
471  
472  
473  
474  
475  
476  
477  
478  
479  
480  
481  
482  
483  
484  
485  
486  
487  
488  
489  
490  
491  
492  
493  
494

**Human Face**

**Natural Scene**

**Heavy Text**

471 A bustling ancient marketplace in a Middle Eastern setting. The scene is filled with vibrant stalls selling exotic goods such as spices, textiles, and pottery. Colorful tents and  
472 banners hang above the narrow cobblestone streets, and the air is filled with the sounds of merchants calling out and the aroma of freshly baked bread. In the center, a group of  
473 travelers in traditional attire is seen haggling over prices and examining the goods. The background features a grand mosque with a tall minaret, and the sky is a clear blue with a  
474 few fluffy clouds. The atmosphere is lively and vibrant, capturing the essence of a historic trading hub.

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