Region-Adaptive Sampling for Diffusion Transformers

Supplementary Material

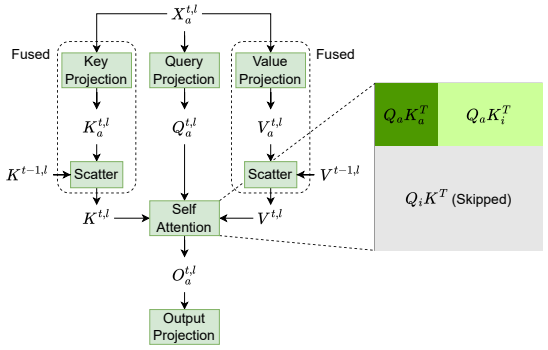


Figure 7. Illustration of the ***RAS*** self-attention module with *Attention Recovery* for enhanced generation quality. $X_a^{t,l}$, $Q_a^{t,l}$, $K_a^{t,l}$, $V_a^{t,l}$, and $O_a^{t,l}$ denote the input hidden states, query, key, value, and attention output of *active tokens* in layer l at timestep t, respectively. $K^{t,l}$ and $V^{t,l}$ represent the key and value caches, which are partially updated via a fused scatter operation integrated into the preceding projection layer using a PIT GeMM kernel. The keys and values of the non-focused regions ($K_i^{t,l}$ and $V_i^{t,l}$) are estimated using cached representations from the previous timestep ($K^{t-1,l}$ and $V^{t-1,l}$). This design preserves contextual consistency while minimizing redundant computation.

Method	FID ↓	sFID ↓	CLIP ↑	time/image (s)
RFlow	36.54	40.25	34.29	3.90
R-75%	37.24	40.23	34.18	3.05
R-50%	38.96	41.17	34.12	2.40
R-25%	40.82	41.41	34.00	1.81
R-12.5%	42.13	40.25	33.96	1.59

Table 7. Comparison with detailed prompts on ParaImage-3000[52]. R-X% represents RAS with X% sample ratio.

6. Evaluation of detailed prompts, objects, positions, and counts.

To evaluate the effect of *RAS* in scenarios when using extremely detailed prompts, and when the user requires exact numbers or positions of the objects, we test *RAS* on the ParaImage-3000 [52] and GenEval[15] dataset, which evaluates the model's ability to generate single, two, multiple objects, colors, and positions with a fixed set of prompts and gives an overall score. As is shown in Table 6 and 7, RAS has little effect on the overall score and provides Pareto improvement in multiple fields.

6.1. More Visualization of *RAS*

This section presents *RAS* accelerating Lumina-Next-T2I and Stable Diffusion 3 with a 50% sampling ratio. As illustrated in Figure 9, the main object receives more sampling steps compared to the background, demonstrating the significance of our region-adaptive sampling strategy. This approach ensures that the primary subject in the generated image consistently undergoes more sampling, while relatively smooth regions receive fewer sampling steps. For instance, in the example shown in Figure 9 with the prompt "hare in snow," the weeds in the snow are sampled more frequently, while the smooth snow receives fewer sampling steps.

In Figure 10, we visualize the standard deviation of the noise across dimensions, as well as the decoded images derived from the noise. This stems from our observation that the noise's standard deviation is consistently smaller in the main subject areas. A preliminary hypothesis is that this occurs because the main subject contains more information. When mixed with a certain proportion of noise at each diffusion step, the foreground tends to retain more deterministic information compared to the background. This allows the model to predict more consistent denoising directions. We acknowledge that further study is needed to fully understand this phenomenon.

Method	Step	Sing. Obj. ↑	Two Obj. ↑	Count ↑	Color ↑	Pos. ↑	SpeedUp ↑	Overall Score ↑
RFlow	30	0.92	0.44	0.40	0.70	0.08	1	0.45
RAS-25%	30	**0.92**	**0.44**	0.39	0.69	0.07	1.25	0.44
RAS-50%	30	**0.92**	0.41	**0.40**	0.68	0.08	1.56	0.44
RFlow	15	0.91	0.39	0.37	0.67	0.07	2.01	0.42
RAS-75%	30	**0.91**	0.37	**0.37**	**0.67**	**0.07**	2.25	**0.42**
RAS-87.5%	30	0.89	0.33	0.35	**0.67**	0.05	2.70	0.40

Table 6. GenEval of RAS and RFlow on Lumina. GenEval evaluates the method's ability to follow instructions, including single object, two objects, object counting, colors, and positions, and gives an overall score. RAS has little effect on the overall score while providing high speedup.

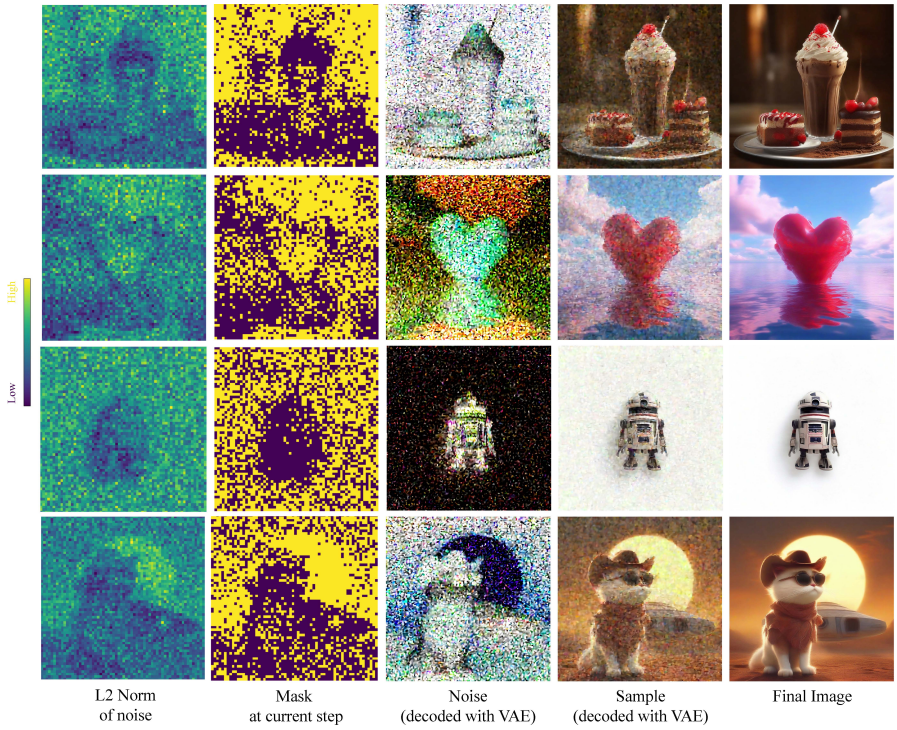


Figure 8. *RAS* using norm as the metric, accelerating Lumina-Next-T2I with 50% sample ratio and 30 total steps. The noise, masks and samples are from the 20th step.

The primary contribution of this work is to highlight that employing different sampling steps for different regions can significantly enhance the efficiency of diffusion model sampling. The method for selecting these regions is not limited to the aforementioned approach based on the noise standard deviation across dimensions. For example, we also experimented with using the $l-2$ norm of the noise output by the network as a criterion for selection. By targeting regions with larger noise norms, which indicate areas the network deems requiring more refinement, we observed a preference for more complex regions in the frequency domain as in Figure 8. This approach also achieves high-quality imaging results, as shown in Table 8. It can be seen that the methods using the l_2 norm and standard deviation (std) yield relatively similar results, and both significantly outperform random selection, particularly when the cache ratio is higher.

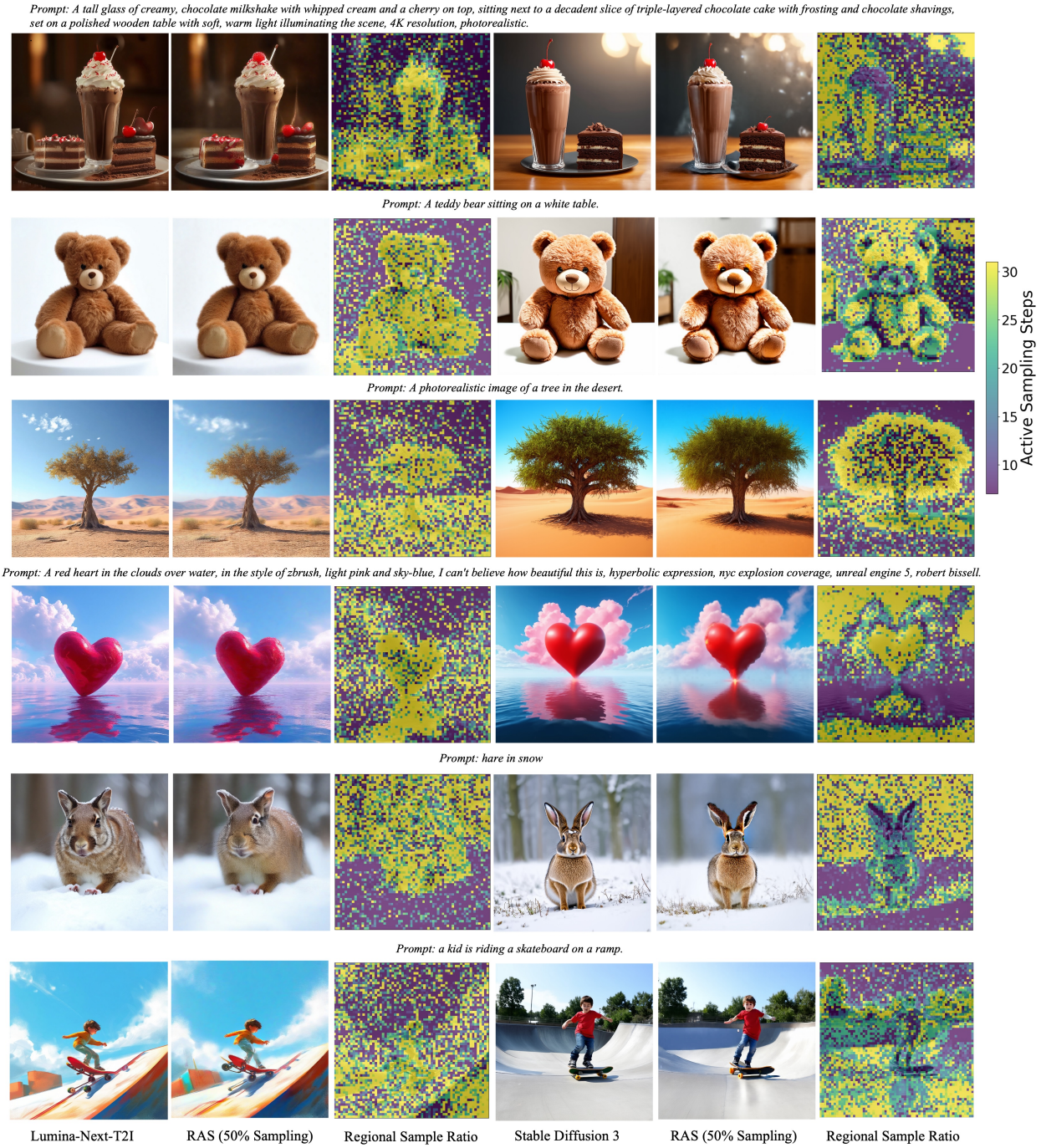


Figure 9. *RAS* VS default sampling and the active sampling step for each latent token.

Method	Sample Steps	Sampling Ratio	Image/s ↑	FID ↓	sFID ↓	CLIP score ↑
RFlow	7	100.0%	1.01	27.23	17.76	30.87
RAS-Std	7	25.0%	1.45	31.99	21.7	30.64
RAS-Norm	7	25.0%	1.45	31.65	21.24	30.59
Random	7	25.0%	1.45	33.26	22.10	30.67

Table 8. Experiments on using L2 Norm as the metric for *RAS* on Stable Diffusion 3. The sample ratio of the first 4 steps is 100% to guarantee generation qualities.

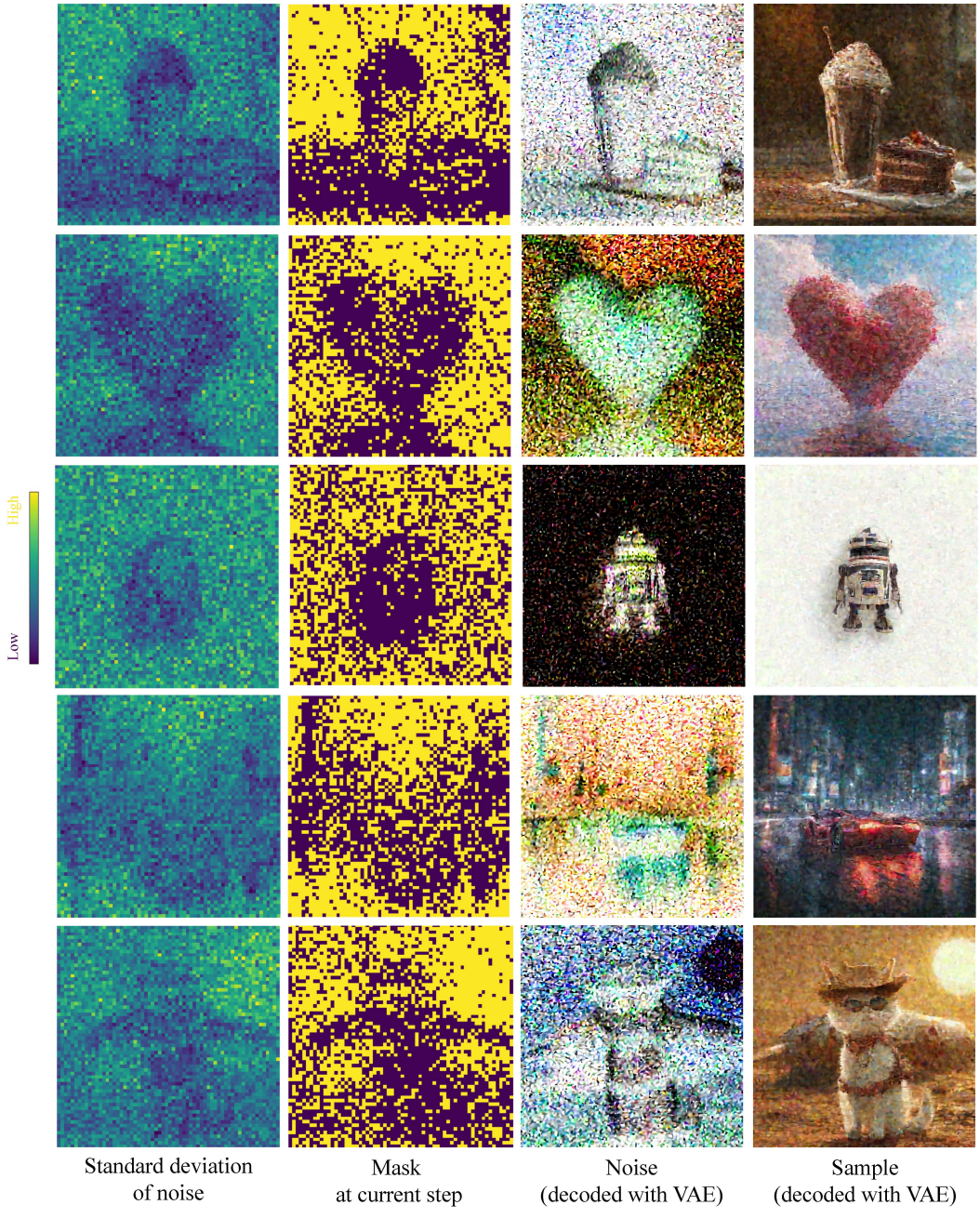


Figure 10. The 20th sampling step (out of 30) of Lumina-Next-T2I using *RAS*.

7. Full Experiment Results of *RAS*

Method	Sample Steps	Sampling Ratio	Image/s ↑	FID ↓	sFID ↓	CLIP score ↑
RFlow	30	100.0%	0.11	22.46	16.59	30.47
RAS	30	75.0%	0.14	23.31	17.73	30.49
RFlow	23	100.0%	0.15	23.10	17.91	30.42
RAS	30	50.0%	0.18	24.10	18.83	30.51
RFlow	15	100.0%	0.23	24.88	21.02	30.25
RAS	30	25.0%	0.26	27.44	20.95	30.45
RAS	15	75.0%	0.27	26.82	23.33	30.26
RAS	30	12.5%	0.31	33.64	23.44	30.36
RAS	15	50.0%	0.33	28.48	25.17	30.29
RFlow	10	100.0%	0.34	31.35	27.84	29.74
RAS	10	75.0%	0.40	34.19	30.57	29.79
RAS	15	25.0%	0.43	33.28	27.41	30.24
RAS	15	12.5%	0.48	39.75	28.88	30.14
RAS	10	50.0%	0.48	36.18	32.36	29.86
RFlow	7	100.0%	0.49	48.19	38.60	28.65
RAS	7	75.0%	0.54	50.45	40.19	28.78
RAS	10	25.0%	0.59	42.96	33.51	29.91
RAS	7	50.0%	0.61	51.78	40.51	28.82
RAS	6	75.0%	0.62	66.12	46.58	27.80
RAS	10	12.5%	0.65	47.34	32.70	29.75
RAS	6	50.0%	0.67	66.54	46.71	27.83
RAS	7	25.0%	0.70	53.93	39.80	28.85
RAS	7	12.5%	0.74	54.62	40.23	28.83
RAS	6	25.0%	0.74	67.16	46.46	27.85
RAS	5	75.0%	0.75	99.01	56.26	26.02
RAS	6	12.5%	0.78	67.88	45.89	27.83
RFlow	5	100.0%	0.69	96.53	59.26	26.03
RAS	5	50.0%	0.83	99.81	56.57	26.01
RAS	5	25.0%	0.95	101.50	56.40	25.93
RAS	5	12.5%	1.00	102.90	55.25	25.84
RFlow	3	100.0%	1.15	256.90	94.80	19.67

Table 9. Full experiment results of *RAS* and rectified flow on Lumina-Next-T2I and COCO Val2014 1024×1024.

Method	Sample Steps	Sampling Ratio	Image/s ↑	FID ↓	sFID ↓	CLIP score ↑
RFlow	28	100%	0.26	25.8	15.32	31.4
RAS	28	75.0%	0.33	24.43	15.94	31.39
RAS	28	50.0%	0.42	24.86	16.88	31.36
RFlow	14	100%	0.51	24.49	14.78	31.34
RAS	28	25.0%	0.55	25.16	17.11	31.29
RFlow	12	100%	0.59	24.36	14.89	31.3
RAS	14	75.0%	0.62	23.61	15.92	31.35
RAS	28	12.5%	0.63	25.72	17.3	31.22
RFlow	10	100%	0.71	24.17	15.39	31.22
RAS	14	50.0%	0.74	24.6	17.24	31.32
RAS	14	25.0%	0.91	25.88	17.97	31.24
RAS	10	75.0%	0.91	24.39	16.29	31.12
RAS	14	12.5%	0.98	26.48	18.14	31.18
RAS	10	50.0%	1.0	27.1	17.5	30.93
RFlow	7	100%	1.01	27.23	17.76	30.87
RAS	7	75.0%	1.16	27.57	18.76	30.81
RAS	10	25.0%	1.2	30.97	18.36	30.67
RAS	10	12.5%	1.3	35.81	18.41	30.13
RAS	7	50.0%	1.3	30.04	20.34	30.73
RAS	6	75.0%	1.3	31.23	19.98	30.48
RAS	6	50.0%	1.41	32.21	20.86	30.43
RFlow	5	100%	1.43	39.7	22.34	29.84
RAS	7	25.0%	1.45	31.99	21.7	30.64
RAS	7	12.5%	1.48	32.86	22.1	30.55
RAS	6	25.0%	1.52	33.24	21.51	30.36
RAS	6	12.5%	1.57	33.81	21.62	30.33
RAS	5	75.0%	1.59	44.02	23.14	29.53
RAS	5	50.0%	1.75	48.65	24.51	29.29
RFlow	4	100%	1.79	61.92	27.42	28.45
RAS	5	25.0%	1.94	51.92	25.67	29.06
RAS	5	12.5%	1.99	53.24	26.04	28.94
RFlow	3	100%	2.38	121.61	36.92	25.32

Table 10. Full experiment results of *RAS* and rectified flow on Stable Diffusion 3 and COCO Val2014 1024×1024.

8. Questionnaire for Human Evaluation

This section contains the questionnaire we used for the human evaluation we mentioned in Section 4.

Text-to-Image Quality Preference Survey

We are conducting an evaluation of two image generation methods. You will be presented with 14 pairs of images, each created by one of the two methods, with the order of the images shuffled for objectivity. Please select your preference for the shown images. Thank you for your participation and cooperation.

Q1. A massive alien spaceship that is shaped like a pretzel.

A.

B.

☐ A is obviously better than B.
☐ A is slightly better than B.
☐ They are of similar qualities.
☐ B is slightly better than A.
☐ B is obviously better than A.

Q2. Upper body of a young woman in a Victorian-era outfit with brass goggles and leather straps. Background shows an industrial revolution cityscape with smoky skies and tall, metal structures.

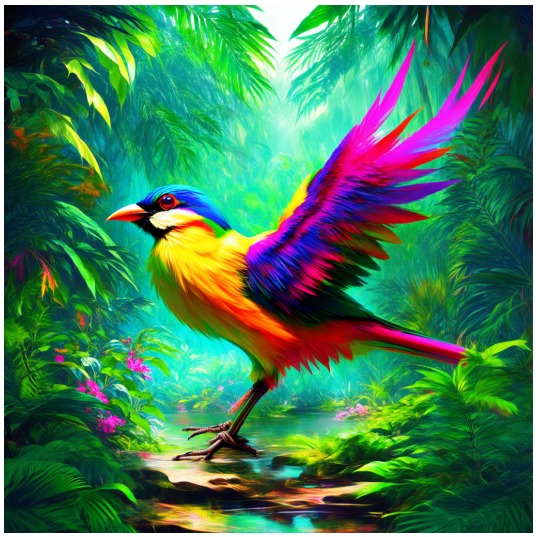
A.

B.

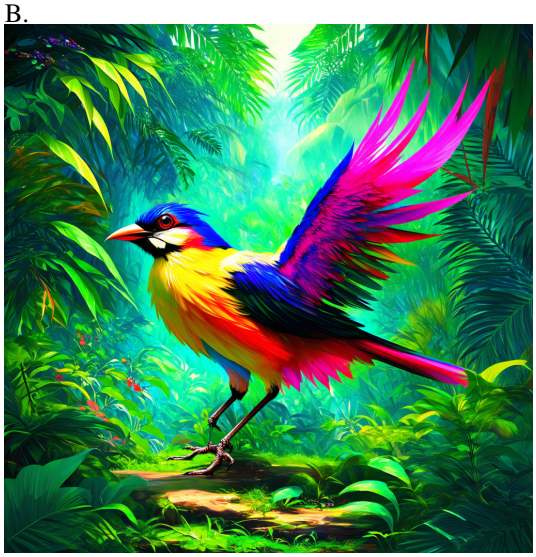
☐ A is obviously better than B.
☐ A is slightly better than B.
☐ They are of similar qualities.
☐ B is slightly better than A.
☐ B is obviously better than A.

Q3. This dreamlike digital art captures a vibrant, kaleidoscopic bird in a lush rainforest.

A.



B.



☐ A is obviously better than B.
☐ A is slightly better than B.
☐ They are of similar qualities.
☐ B is slightly better than A.
☐ B is obviously better than A.

A.

B.

☐ A is obviously better than B.
☐ A is slightly better than B.
☐ They are of similar qualities.
☐ B is slightly better than A.
☐ B is obviously better than A.

Q4. A cat wearing a cowboy hat and sunglasses and standing in front of a rusty old white spaceship at sunrise. Pixar cute. Detailed anime illustration.

Q5. A kangaroo holding a beer, wearing ski goggles and passionately singing silly songs.

A.

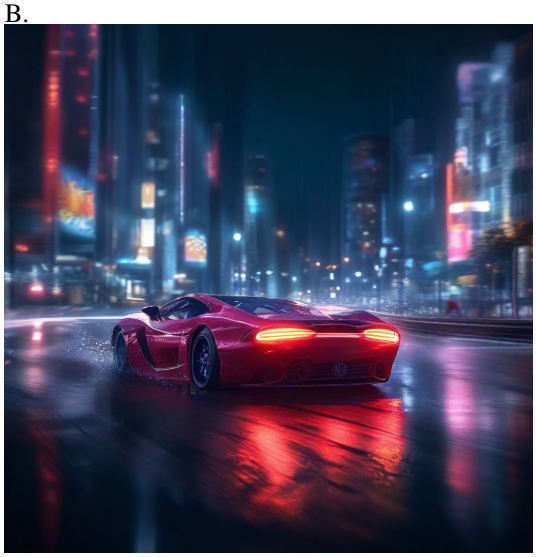
B.

☐ A is obviously better than B.
☐ A is slightly better than B.
☐ They are of similar qualities.
☐ B is slightly better than A.
☐ B is obviously better than A.

Q6. A photorealistic image of a Pagani Huayra driving through a city at night with glowing city lights in the background.

A.

B.



☐ A is obviously better than B.
☐ A is slightly better than B.
☐ They are of similar qualities.
☐ B is slightly better than A.
☐ B is obviously better than A.

Q7. A cheeseburger with juicy beef patties and melted cheese sits on top of a toilet that looks like a throne and stands in the middle of the royal chamber.

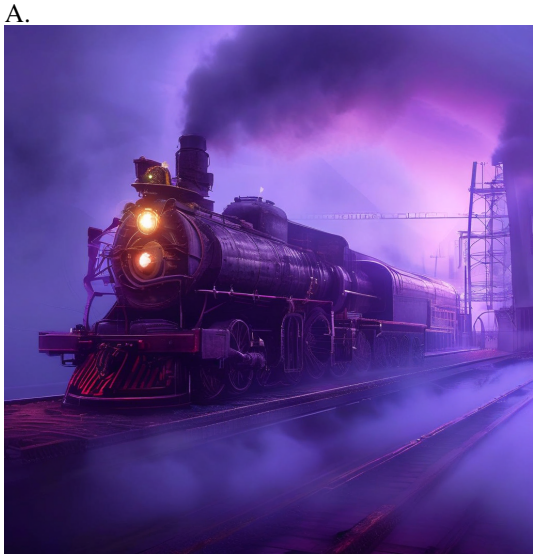
A.

B.

☐ A is obviously better than B.
☐ A is slightly better than B.
☐ They are of similar qualities.
☐ B is slightly better than A.
☐ B is obviously better than A.

Q8. A detailed photorealistic image of a steampunk locomotive on a platform with sharp lines, surrounded by light purple fog.

A.



B.

☐ A is obviously better than B.
☐ A is slightly better than B.
☐ They are of similar qualities.
☐ B is slightly better than A.
☐ B is obviously better than A.

Q9. An entire universe inside a bottle sitting on the shelf at Walmart on sale.

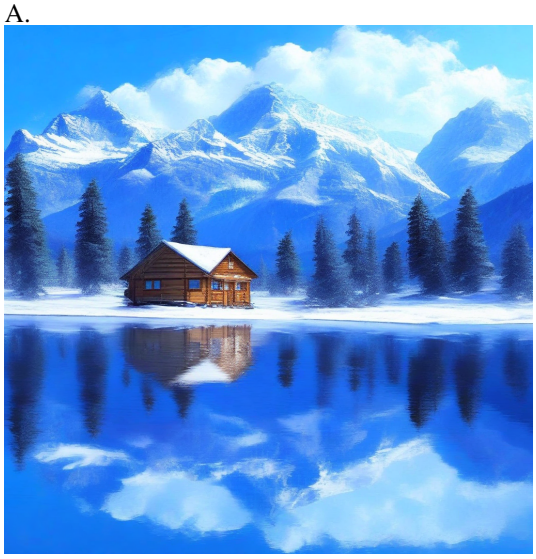
A.

B.

☐ A is obviously better than B.
☐ A is slightly better than B.
☐ They are of similar qualities.
☐ B is slightly better than A.
☐ B is obviously better than A.

Q10. Snow-covered mountains reflected in a crystal-clear alpine lake, with a small wooden cabin nestled among tall pine trees.

A.

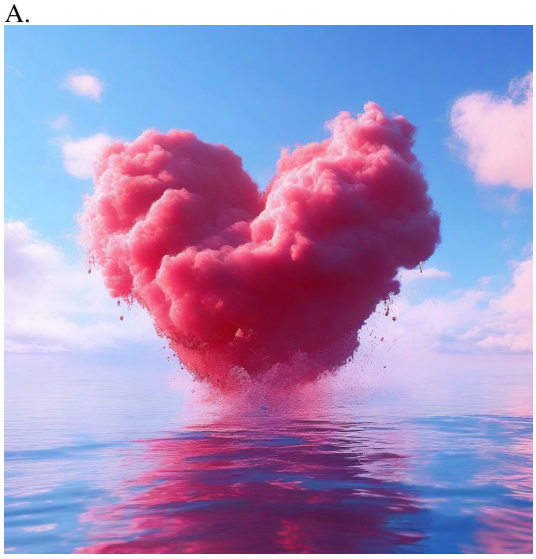


B.

☐ A is obviously better than B.
☐ A is slightly better than B.
☐ They are of similar qualities.
☐ B is slightly better than A.
☐ B is obviously better than A.

Q11. A red heart in the clouds over water, in the style of zbrush, light pink and sky-blue, I can't believe how beautiful this is, hyperbolic expression, nyc explosion coverage, unreal engine 5, robert bissell.

A.

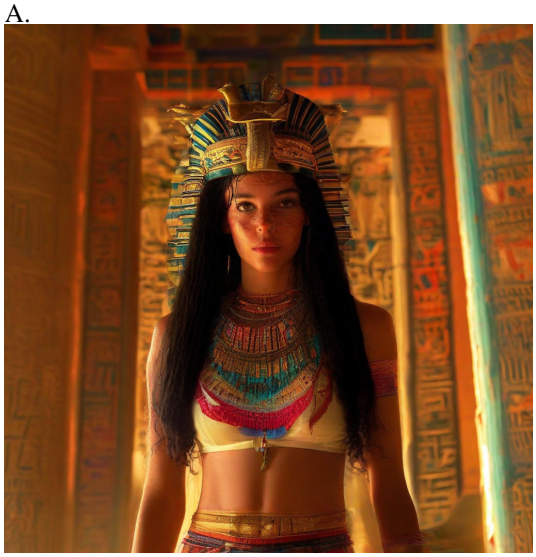


B.

☐ A is obviously better than B.
☐ A is slightly better than B.
☐ They are of similar qualities.
☐ B is slightly better than A.
☐ B is obviously better than A.

Q12. Upper body of a young woman adorned in elaborate ancient Egyptian clothing, with a headdress featuring golden ornaments and colorful gemstones.The background shows the inside of a grand temple with hieroglyphics on the walls.

A.

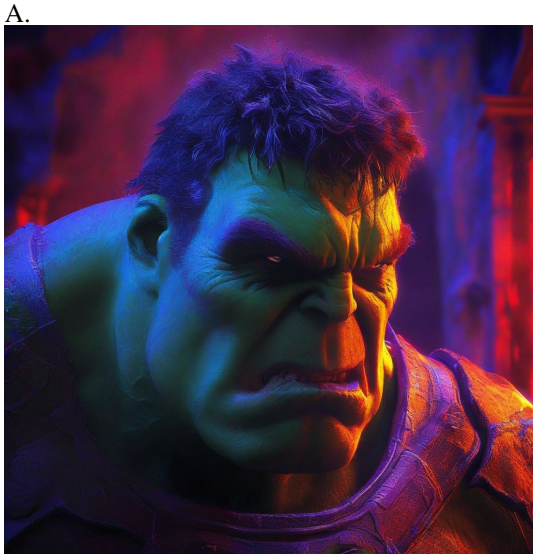


B.

☐ A is obviously better than B.
☐ A is slightly better than B.
☐ They are of similar qualities.
☐ B is slightly better than A.
☐ B is obviously better than A.

Q13. The Hulk is in a colorful gothic background, with highly detailed dramatic lighting and a photo realistic style, rendered in 8K resolution.

A.



B.

☐ A is obviously better than B.
☐ A is slightly better than B.
☐ They are of similar qualities.
☐ B is slightly better than A.
☐ B is obviously better than A.

Q14. A car made out of vegetables.

A.

B.

☐ A is obviously better than B.
☐ A is slightly better than B.
☐ They are of similar qualities.
☐ B is slightly better than A.
☐ B is obviously better than A.