FreeScale: Unleashing the Resolution of Diffusion Models via Tuning-Free Scale Fusion (ICCV 2025)

Haonan Qiu[1] Shiwei Zhang*[2] Yujie Wei[3] Ruihang Chu[2] Hangjie Yuan[2]
Xiang Wang[2] Yingya Zhang[2] Ziwei Liu*[1]

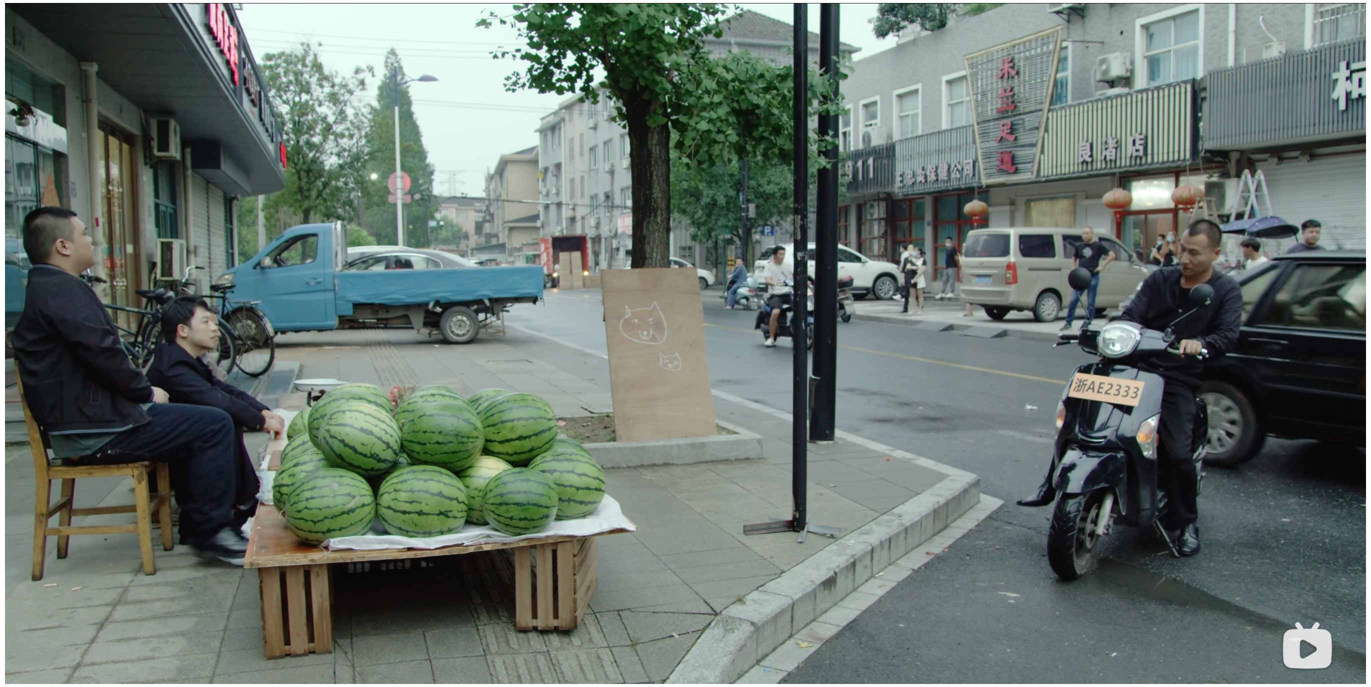
64× Resolution
(8192×8192)
16× Resolution
(4096×4096)
(2048×4096)
(2048×4096)
16× Resolution
(4096×4096)
(4096×2048)
(4096×2048)

[1]Nanyang Technological University [2]Alibaba Group [3]Fudan University

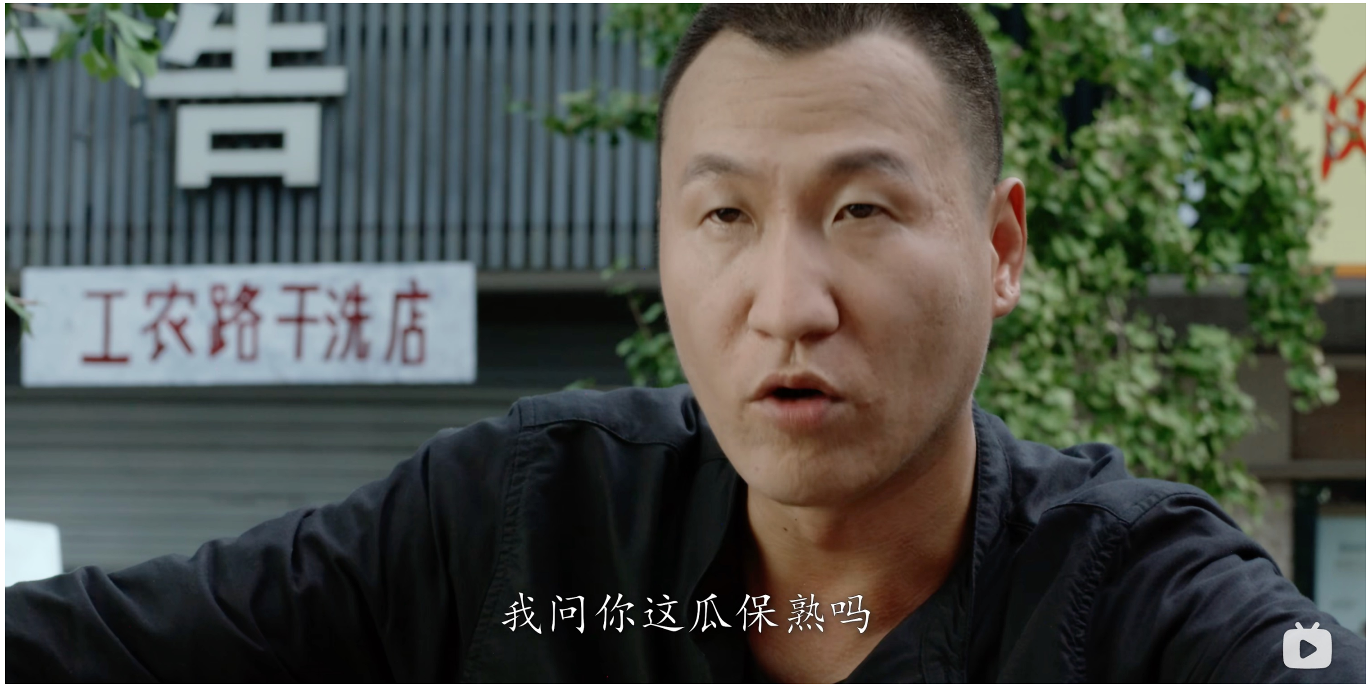
Motivation

Exploring potential higher-resolution visual generation of pre-trained diffusion models.

Face
Long Shot

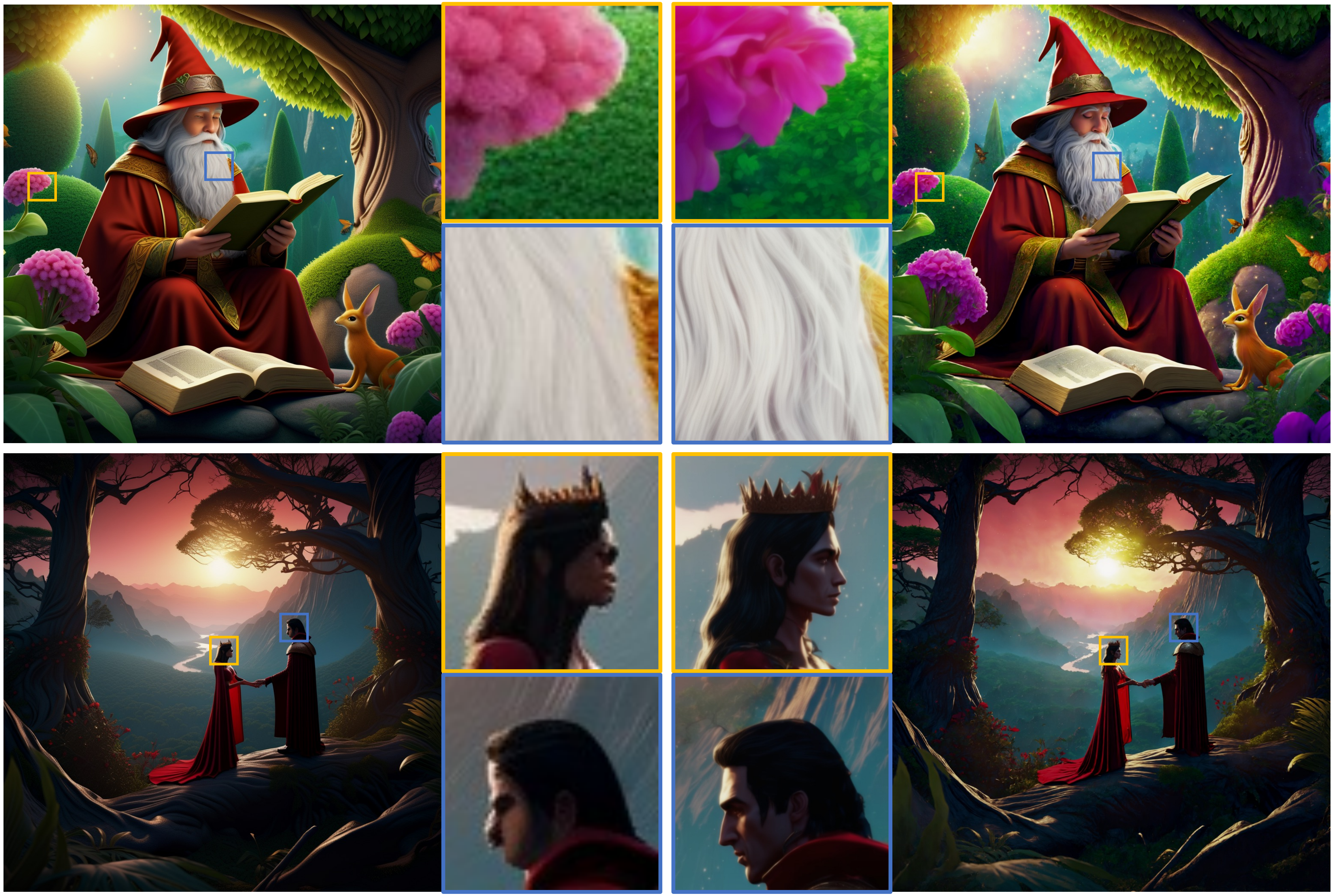


Face
Close-Up



Illustrative Diagram

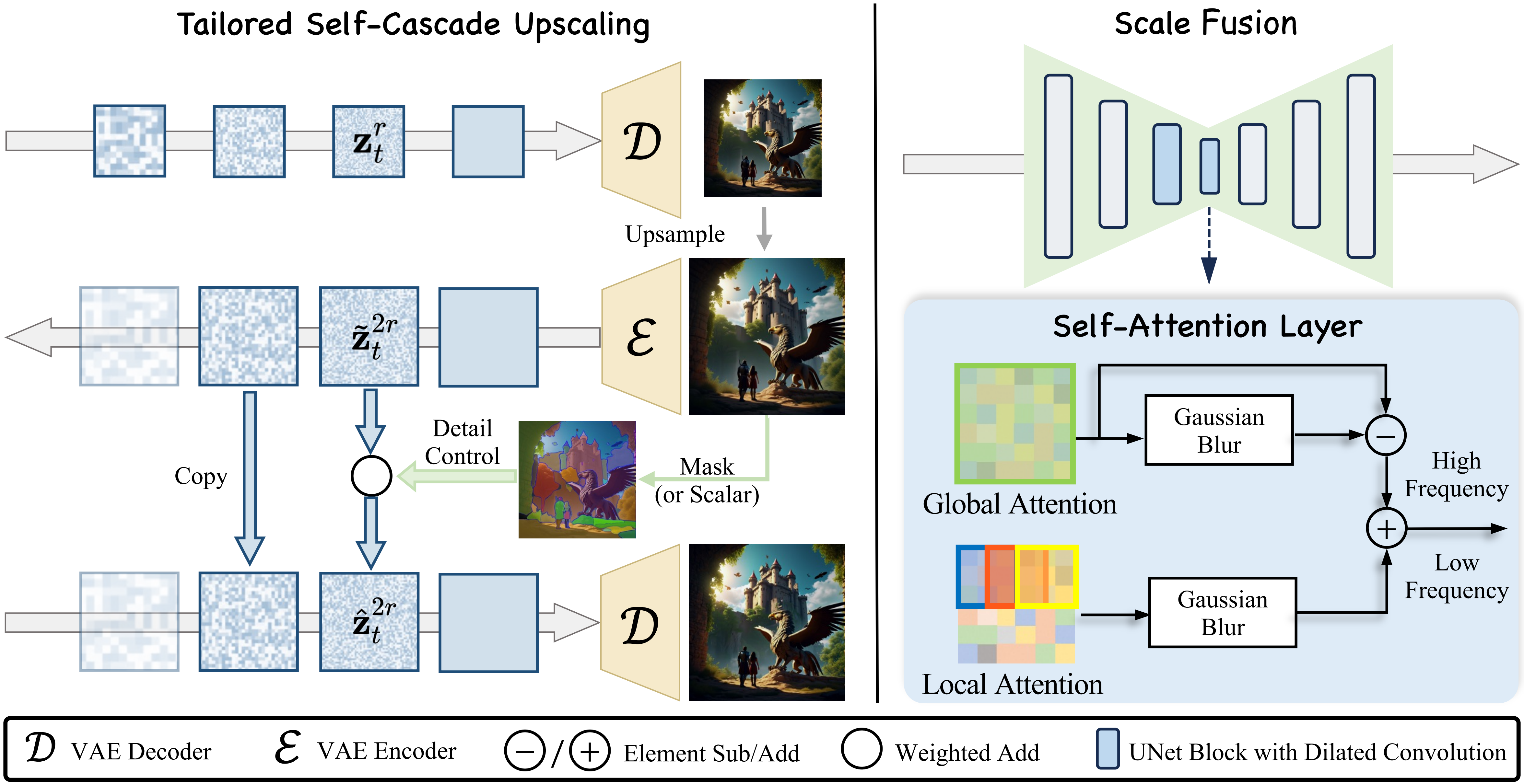
Close-up priors can aid detail recovery in long shots



Resolution: 1024 × 1024

Resolution: 8192 × 8192

Method Overview



Tailored Self-Cascade Upscaling

$\mathbf{z}_t^r$

$\mathcal{D}$

Upsample

$\tilde{\mathbf{z}}_t^{2r}$

$\mathcal{E}$

Copy

Detail Control

Mask (or Scalar)

$\hat{\mathbf{z}}_t^{2r}$

$\mathcal{D}$

Scale Fusion

Self-Attention Layer

Global Attention

Gaussian Blur

High Frequency

Low Frequency

Local Attention

Gaussian Blur

$\mathcal{D}$ VAE Decoder

$\mathcal{E}$ VAE Encoder

⊖/⊕ Element Sub/Add

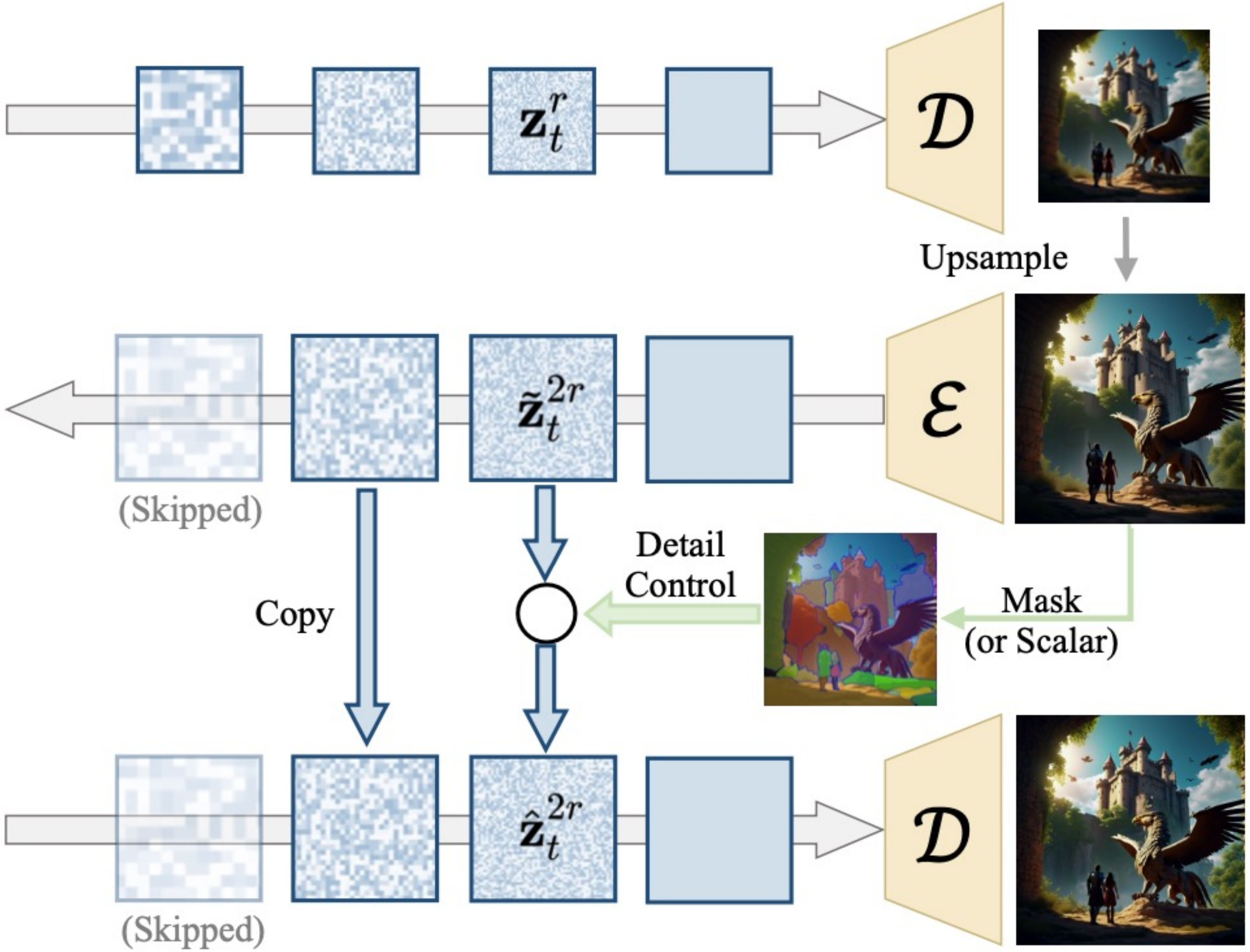
○ Weighted Add

▯ UNet Block with Dilated Convolution

Tailored Self-Cascade Upscaling

$$\hat{\mathbf{z}}_t^r = c \times \tilde{\mathbf{z}}_t^r + (1 - c) \times \mathbf{z}_t^r, \tag{4}$$

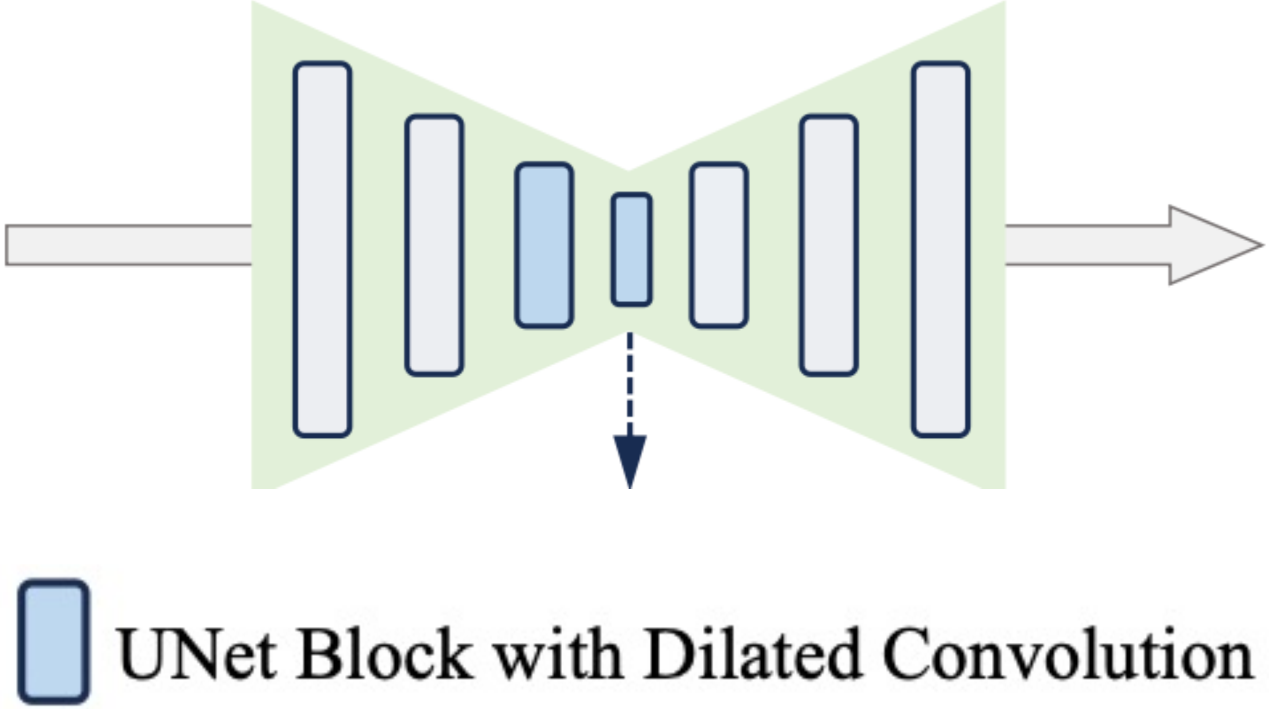
where $c = \left(\left(1 + \cos\left(\frac{T-t}{T} \times \pi\right)\right)/2\right)^{\alpha}$ is a scaled cosine decay factor with a scaling factor α.



Restrained Dilated Convolution

$$f_{\boldsymbol{k}}^{d}(\mathbf{h}) = \mathbf{h} \circledast \Phi_d(\boldsymbol{k}), \left(\mathbf{h} \circledast \Phi_d(\boldsymbol{k})\right)(o) = \sum_{s+d\cdot t=p} \mathbf{h}(p)\cdot \boldsymbol{k}(q), \tag{5}$$

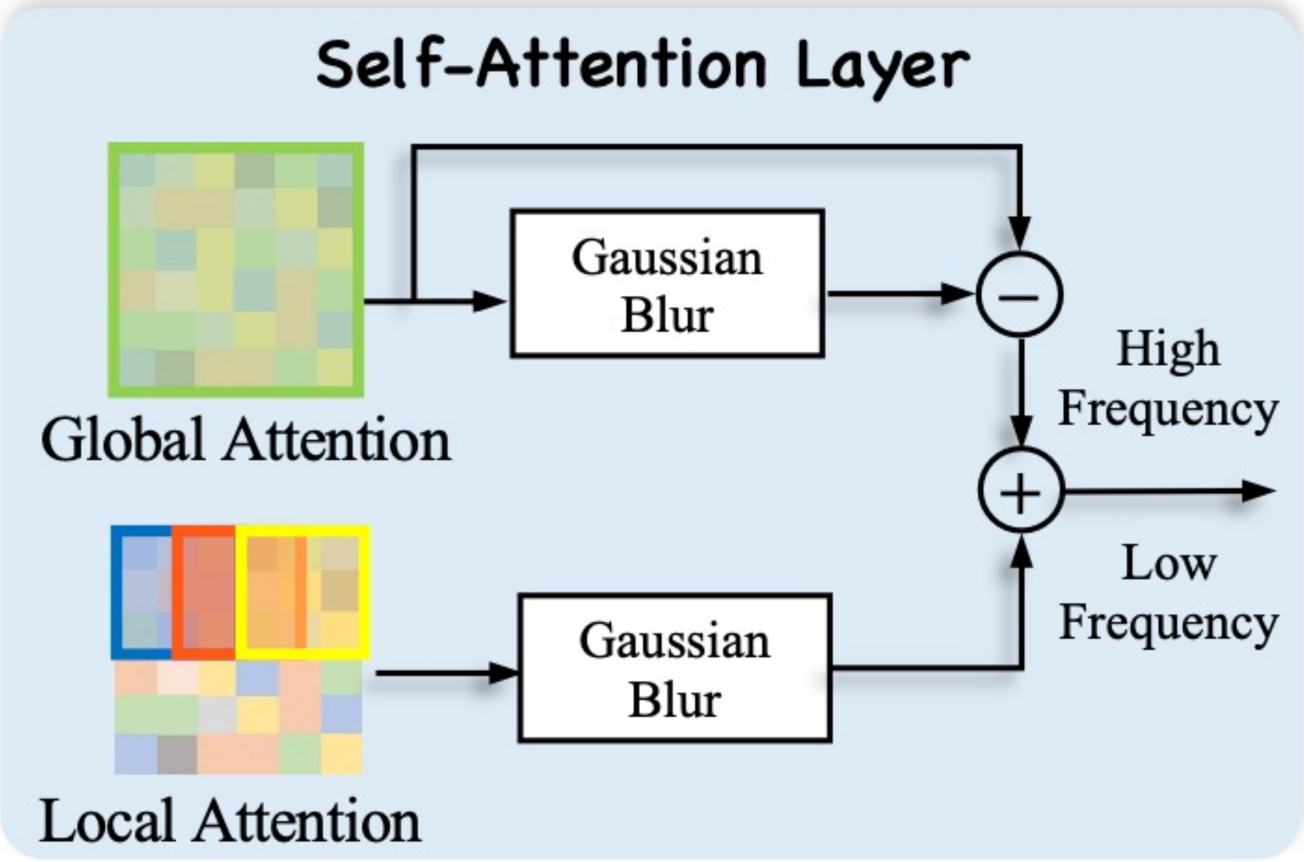
where o, p, and q are spatial locations used to index the feature or kernel. $\circledast$ denotes convolution operation.



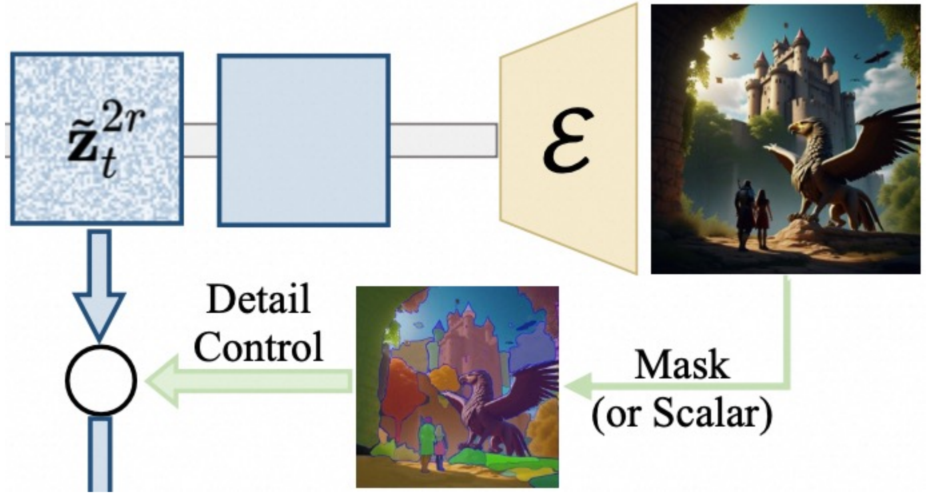
Scale Fusion

$$\mathbf{h}_{\text{out}}^{\text{fusion}} = \underbrace{\mathbf{h}_{\text{out}}^{\text{global}} - G\left(\mathbf{h}_{\text{out}}^{\text{global}}\right)}_{\text{high frequency}} + \underbrace{G\left(\mathbf{h}_{\text{out}}^{\text{local}}\right)}_{\text{low frequency}}, \qquad (7)$$

where G is a low-pass filter implemented as a Gaussian blur, and $\mathbf{h}_{\text{out}}^{\text{global}} - G\left(\mathbf{h}_{\text{out}}^{\text{global}}\right)$ acts as a high pass of $\mathbf{h}_{\text{out}}^{\text{fusion}}$.



Detail Level Control

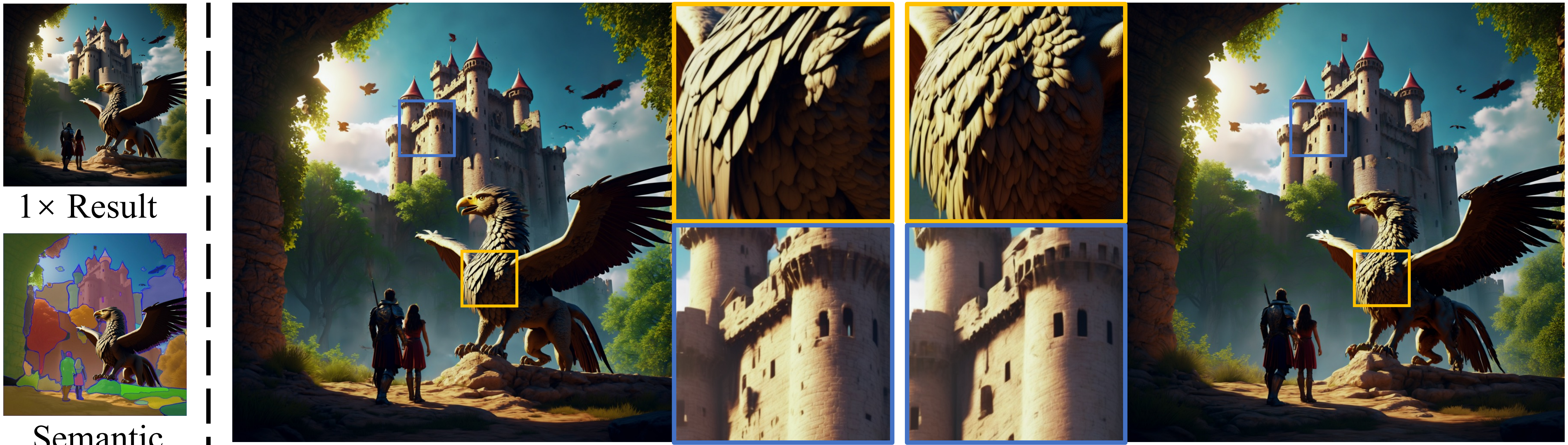


$\tilde{\mathbf{z}}_t^{2r}$

$\mathcal{E}$

Detail Control

Mask (or Scalar)



1× Result

Semantic Mask

Weighted with Scalar

Weighted with Mask

Quantitative Comparison

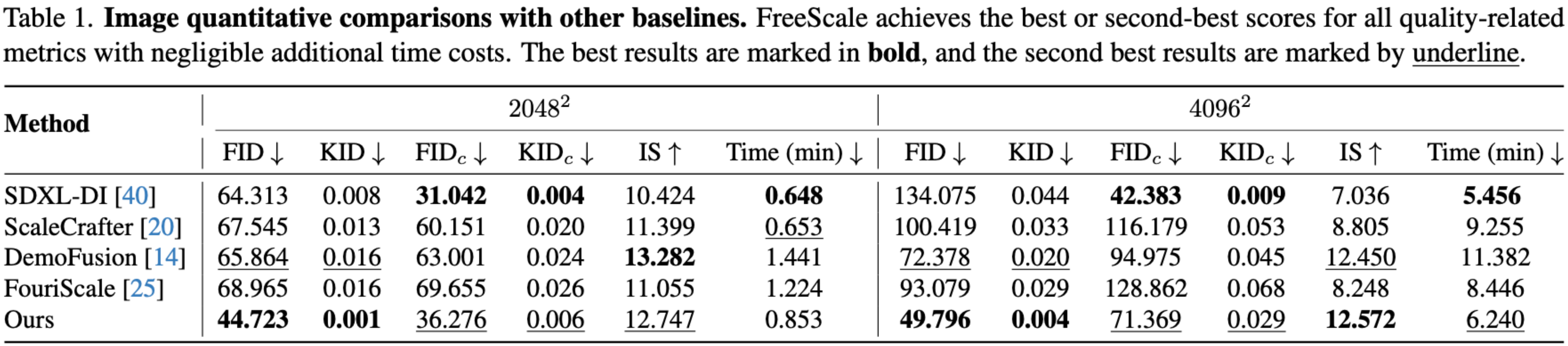


Table 1. **Image quantitative comparisons with other baselines.** FreeScale achieves the best or second-best scores for all quality-related metrics with negligible additional time costs. The best results are marked in **bold**, and the second best results are marked by underline.

Method	2048^2						4096^2					
	FID ↓	KID ↓	FID_c ↓	KID_c ↓	IS ↑	Time (min) ↓	FID ↓	KID ↓	FID_c ↓	KID_c ↓	IS ↑	Time (min) ↓
SDXL-DI [40]	64.313	0.008	**31.042**	**0.004**	10.424	**0.648**	134.075	0.044	**42.383**	**0.009**	7.036	**5.456**
ScaleCrafter [20]	67.545	0.013	60.151	0.020	11.399	0.653	100.419	0.033	116.179	0.053	8.805	9.255
DemoFusion [14]	65.864	0.016	63.001	0.024	**13.282**	1.441	72.378	0.020	94.975	0.045	12.450	11.382
FouriScale [25]	68.965	0.016	69.655	0.026	11.055	1.224	93.079	0.029	128.862	0.068	8.248	8.446
Ours	**44.723**	**0.001**	36.276	0.006	12.747	0.853	**49.796**	**0.004**	71.369	0.029	**12.572**	6.240

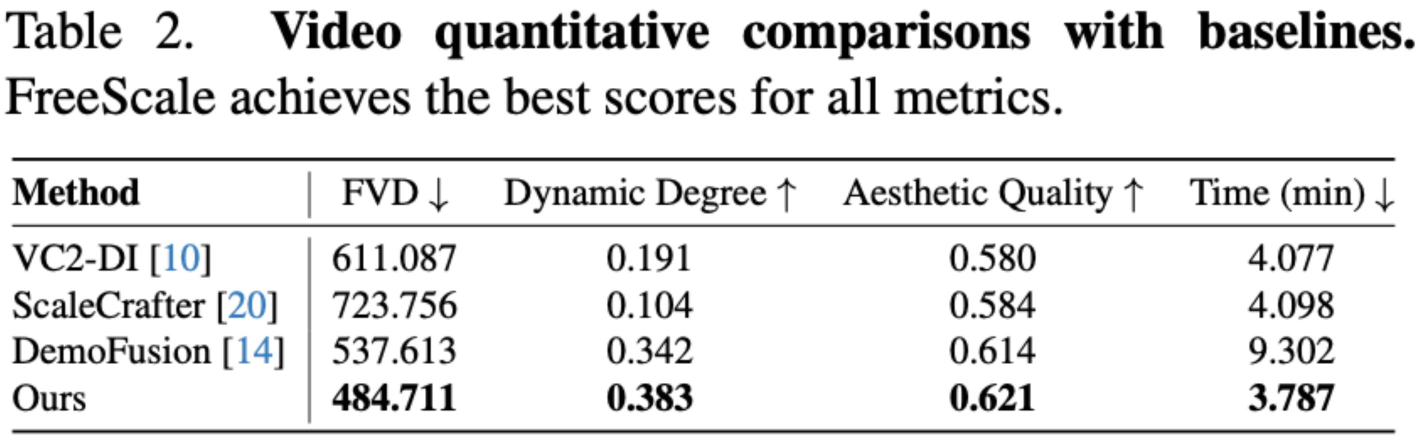


Table 2. **Video quantitative comparisons with baselines.** FreeScale achieves the best scores for all metrics.

Method	FVD ↓	Dynamic Degree ↑	Aesthetic Quality ↑	Time (min) ↓
VC2-DI [10]	611.087	0.191	0.580	4.077
ScaleCrafter [20]	723.756	0.104	0.584	4.098
DemoFusion [14]	537.613	0.342	0.614	9.302
Ours	**484.711**	**0.383**	**0.621**	**3.787**

Ablation Study

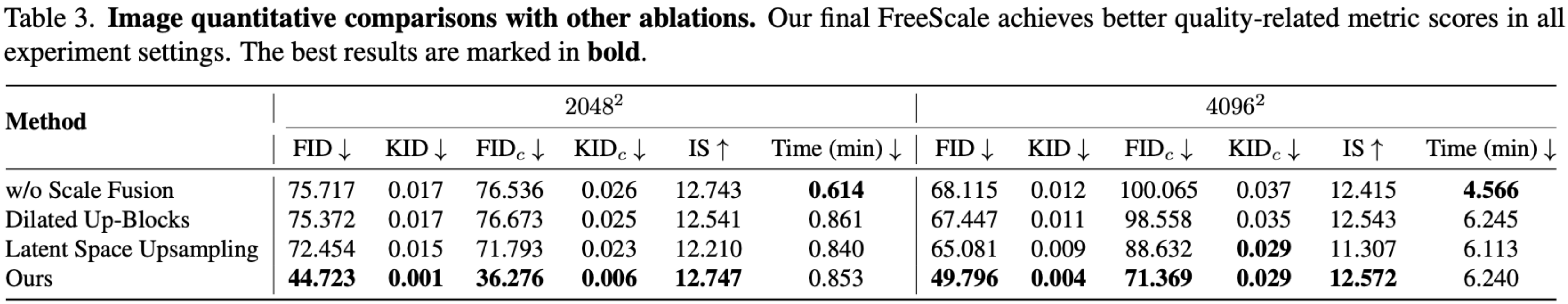


Table 3. **Image quantitative comparisons with other ablations.** Our final FreeScale achieves better quality-related metric scores in all experiment settings. The best results are marked in **bold**.

Method	2048^2						4096^2					
	FID ↓	KID ↓	FID_c ↓	KID_c ↓	IS ↑	Time (min) ↓	FID ↓	KID ↓	FID_c ↓	KID_c ↓	IS ↑	Time (min) ↓
w/o Scale Fusion	75.717	0.017	76.536	0.026	12.743	**0.614**	68.115	0.012	100.065	0.037	12.415	**4.566**
Dilated Up-Blocks	75.372	0.017	76.673	0.025	12.541	0.861	67.447	0.011	98.558	0.035	12.543	6.245
Latent Space Upsampling	72.454	0.015	71.793	0.023	12.210	0.840	65.081	0.009	88.632	**0.029**	11.307	6.113
Ours	**44.723**	**0.001**	**36.276**	**0.006**	**12.747**	0.853	**49.796**	**0.004**	**71.369**	**0.029**	**12.572**	6.240

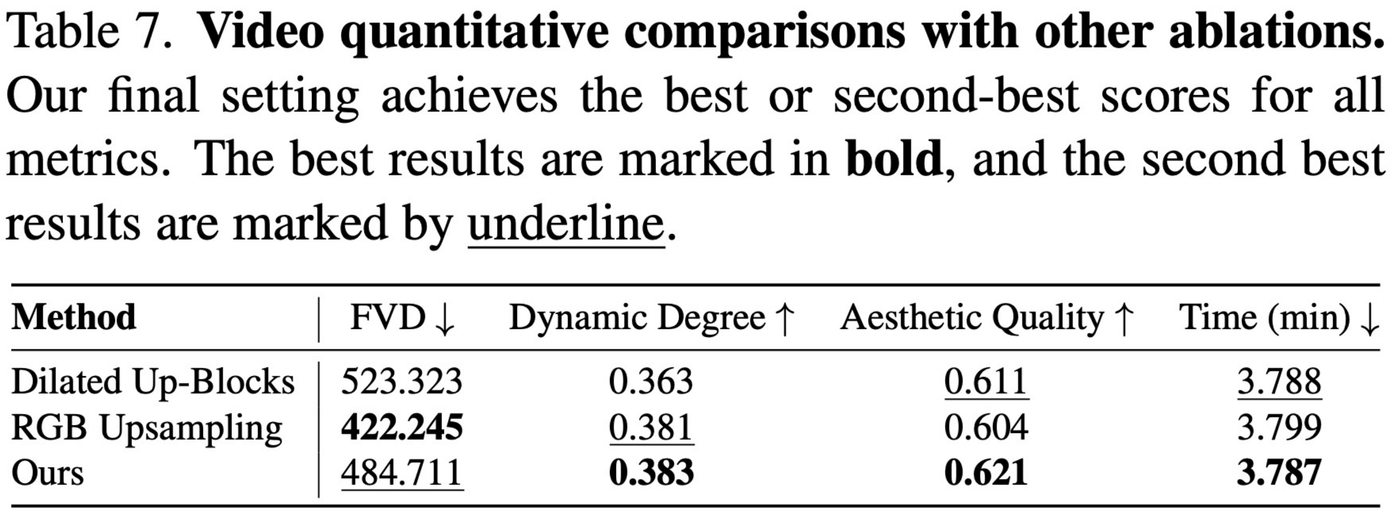


Table 7. **Video quantitative comparisons with other ablations.** Our final setting achieves the best or second-best scores for all metrics. The best results are marked in **bold**, and the second best results are marked by underline.

Method	FVD ↓	Dynamic Degree ↑	Aesthetic Quality ↑	Time (min) ↓
Dilated Up-Blocks	523.323	0.363	0.611	3.788
RGB Upsampling	**422.245**	0.381	0.604	3.799
Ours	484.711	**0.383**	**0.621**	**3.787**

User Study

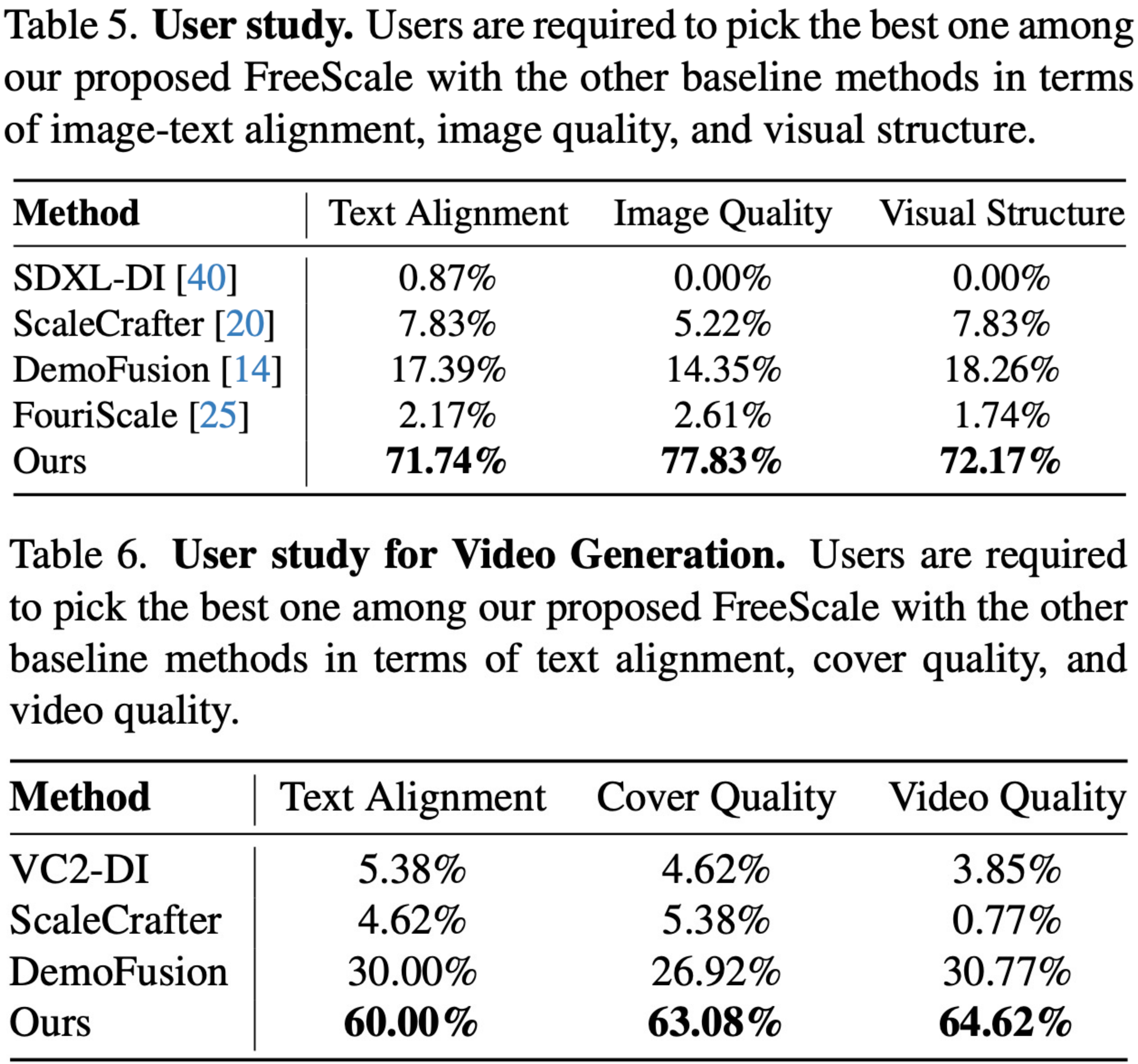


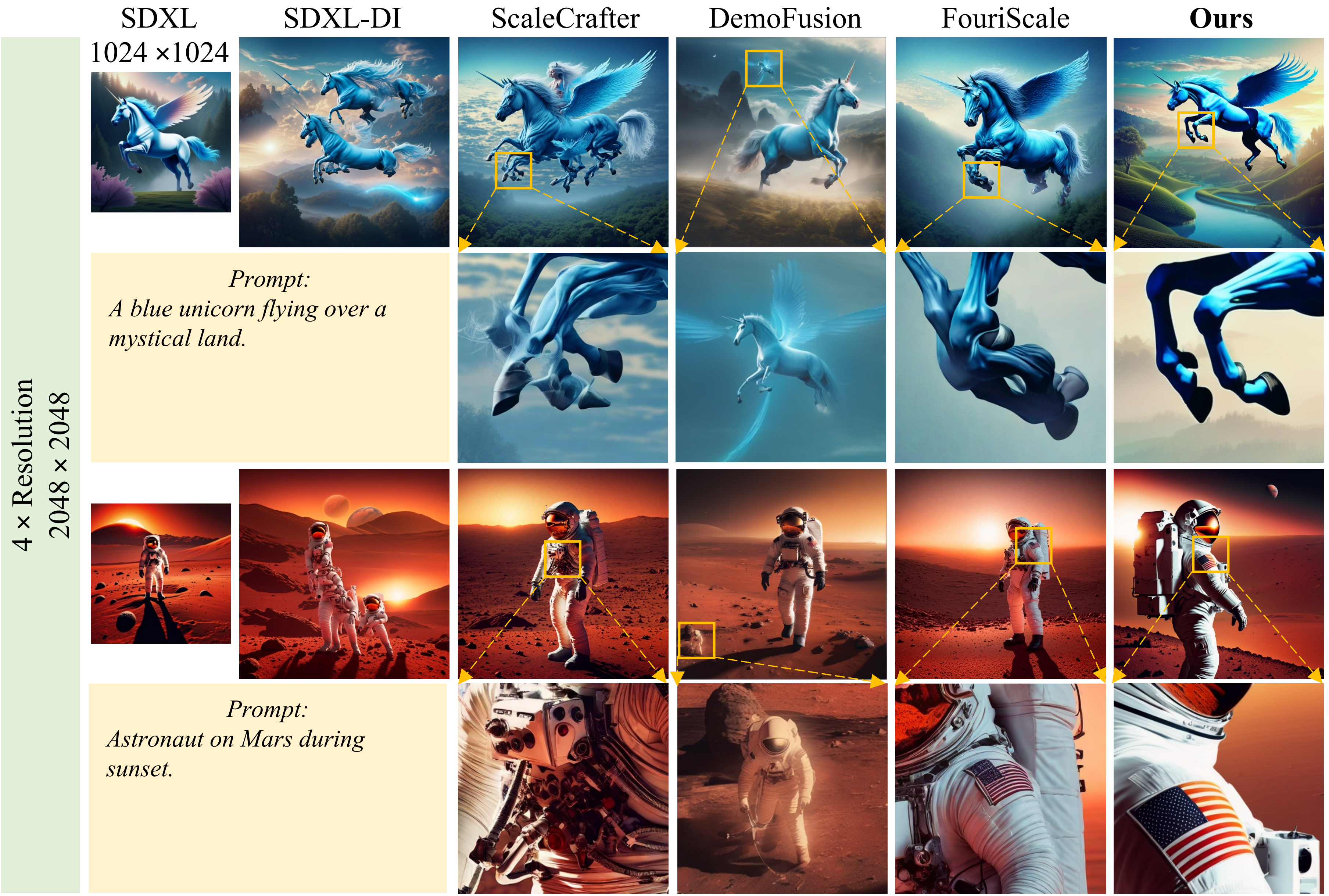
Table 5. **User study.** Users are required to pick the best one among our proposed FreeScale with the other baseline methods in terms of image-text alignment, image quality, and visual structure.

Method	Text Alignment	Image Quality	Visual Structure
SDXL-DI [40]	0.87%	0.00%	0.00%
ScaleCrafter [20]	7.83%	5.22%	7.83%
DemoFusion [14]	17.39%	14.35%	18.26%
FouriScale [25]	2.17%	2.61%	1.74%
Ours	**71.74%**	**77.83%**	**72.17%**

Table 6. **User study for Video Generation.** Users are required to pick the best one among our proposed FreeScale with the other baseline methods in terms of text alignment, cover quality, and video quality.

Method	Text Alignment	Cover Quality	Video Quality
VC2-DI	5.38%	4.62%	3.85%
ScaleCrafter	4.62%	5.38%	0.77%
DemoFusion	30.00%	26.92%	30.77%
Ours	**60.00%**	**63.08%**	**64.62%**

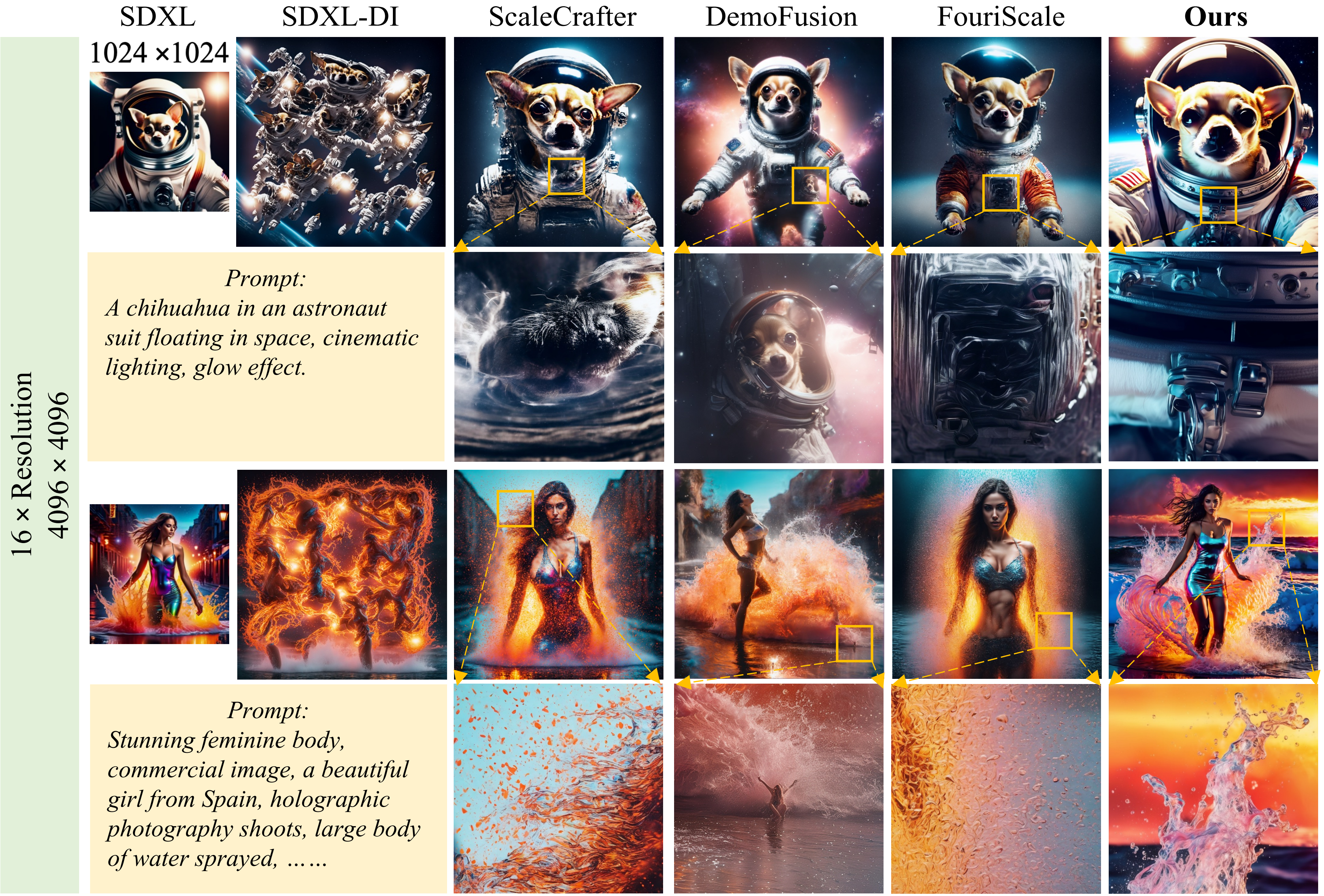
Qualitative Image Comparison



Prompt:
A blue unicorn flying over a mystical land.

Prompt:
Astronaut on Mars during sunset.

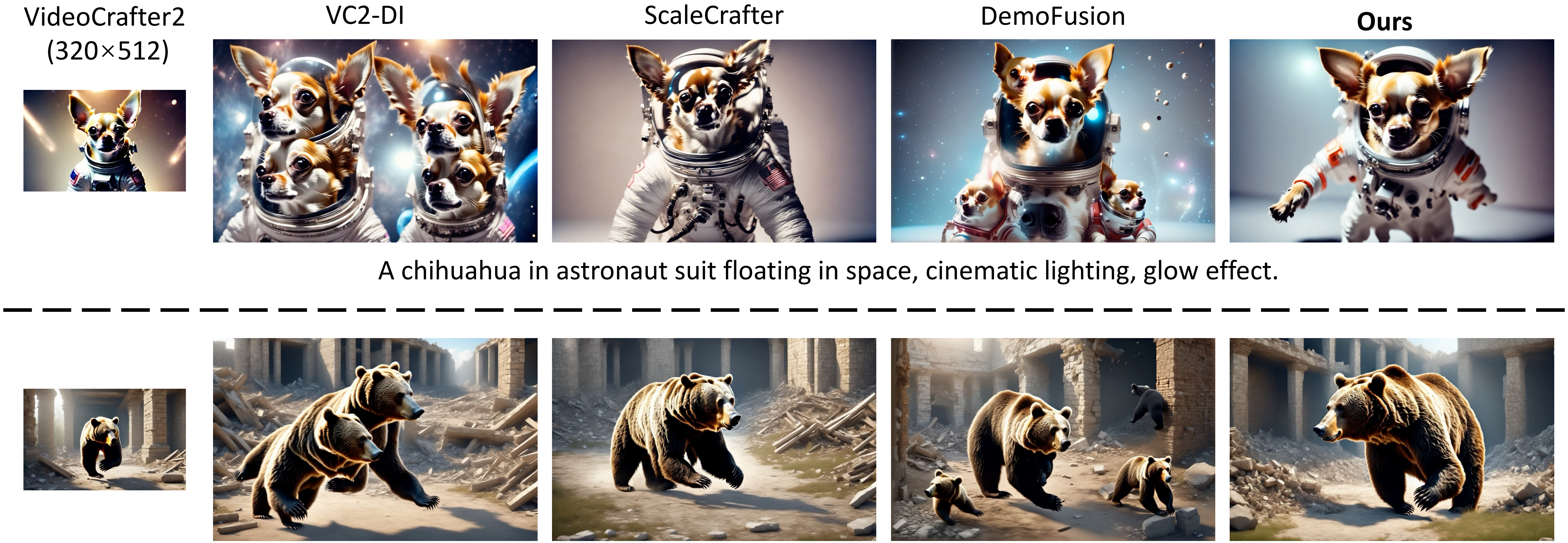
Qualitative Image Comparison



Prompt:
A chihuahua in an astronaut suit floating in space, cinematic lighting, glow effect.

Prompt:
Stunning feminine body, commercial image, a beautiful girl from Spain, holographic photography shoots, large body of water sprayed, ……

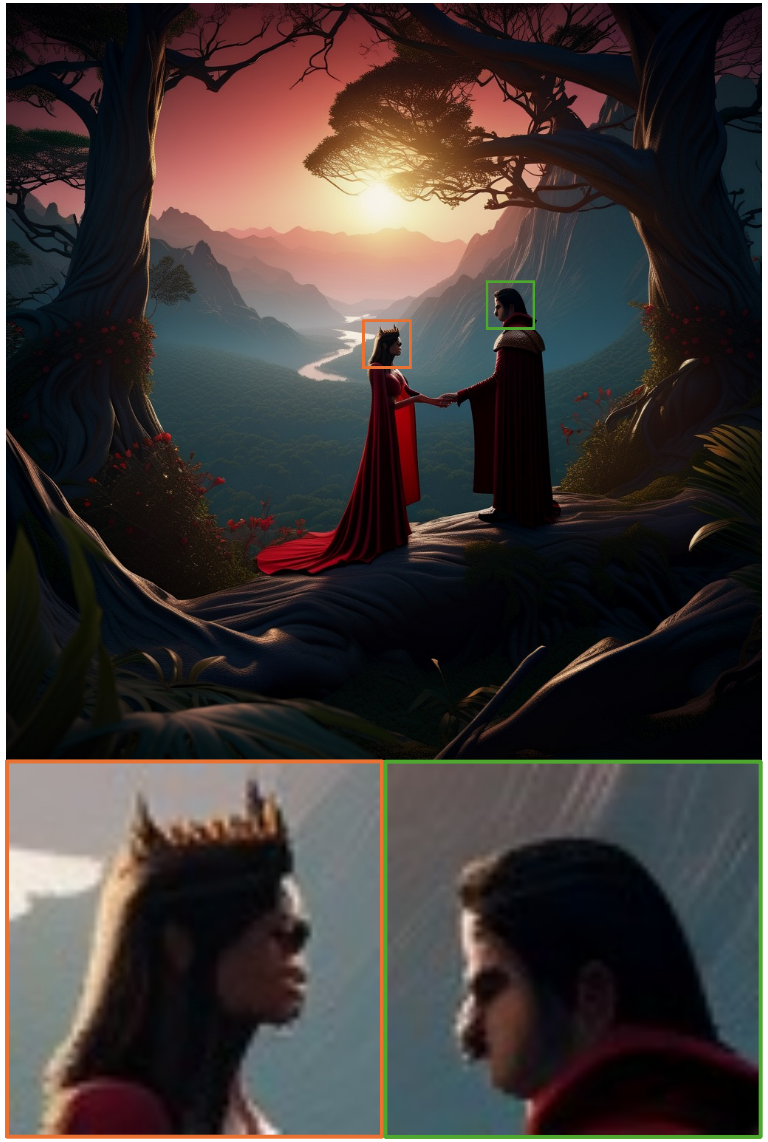
Qualitative Video Comparison



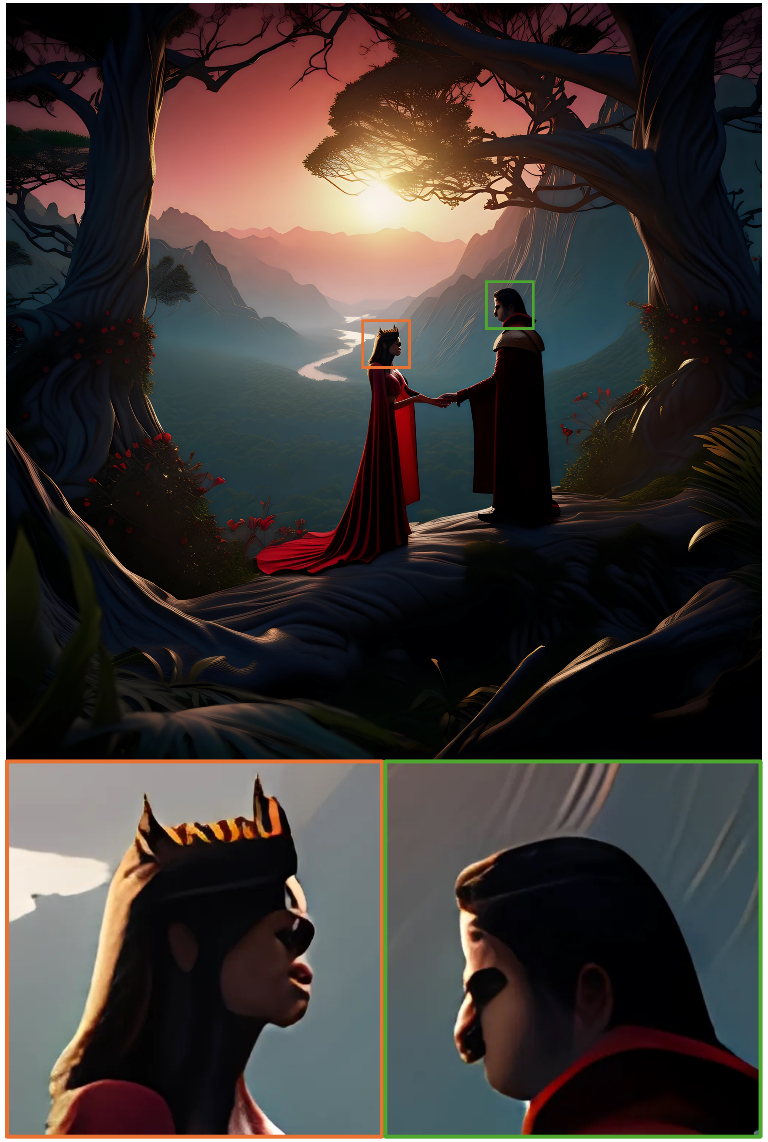
A chihuahua in astronaut suit floating in space, cinematic lighting, glow effect.

A bear running in the ruins, photorealistic, 4k, high definition.

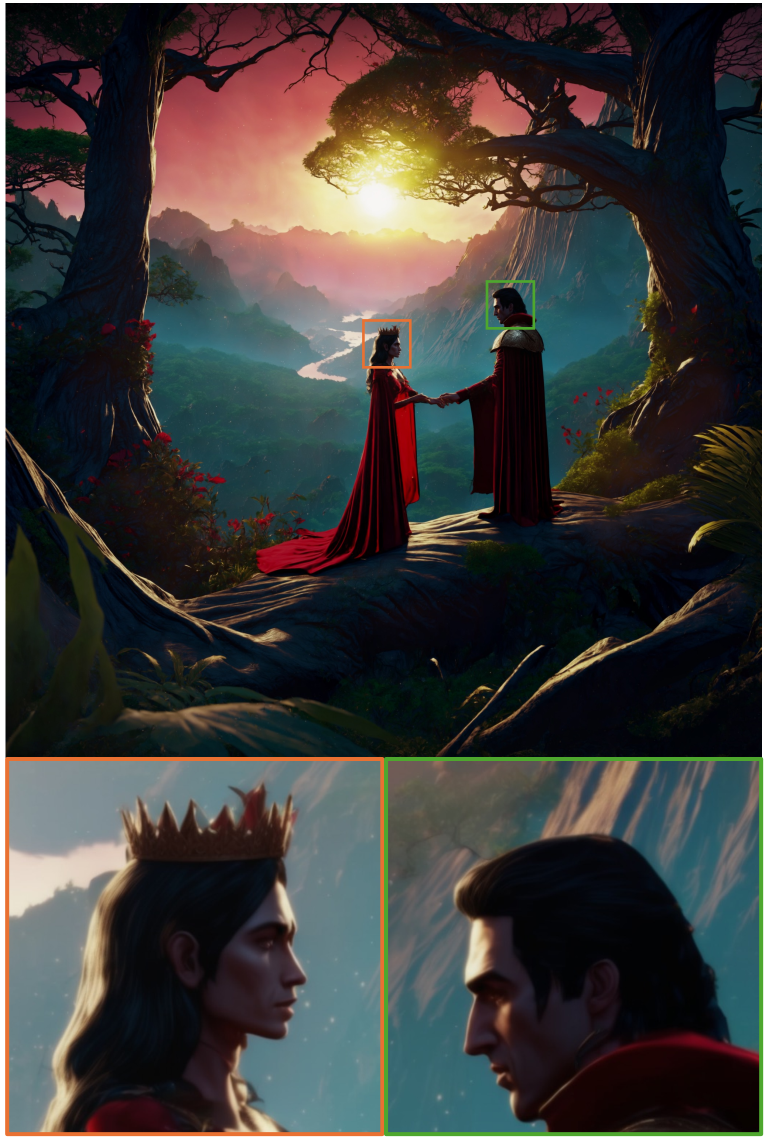
Difference with SR



SDXL (1024 × 1024)

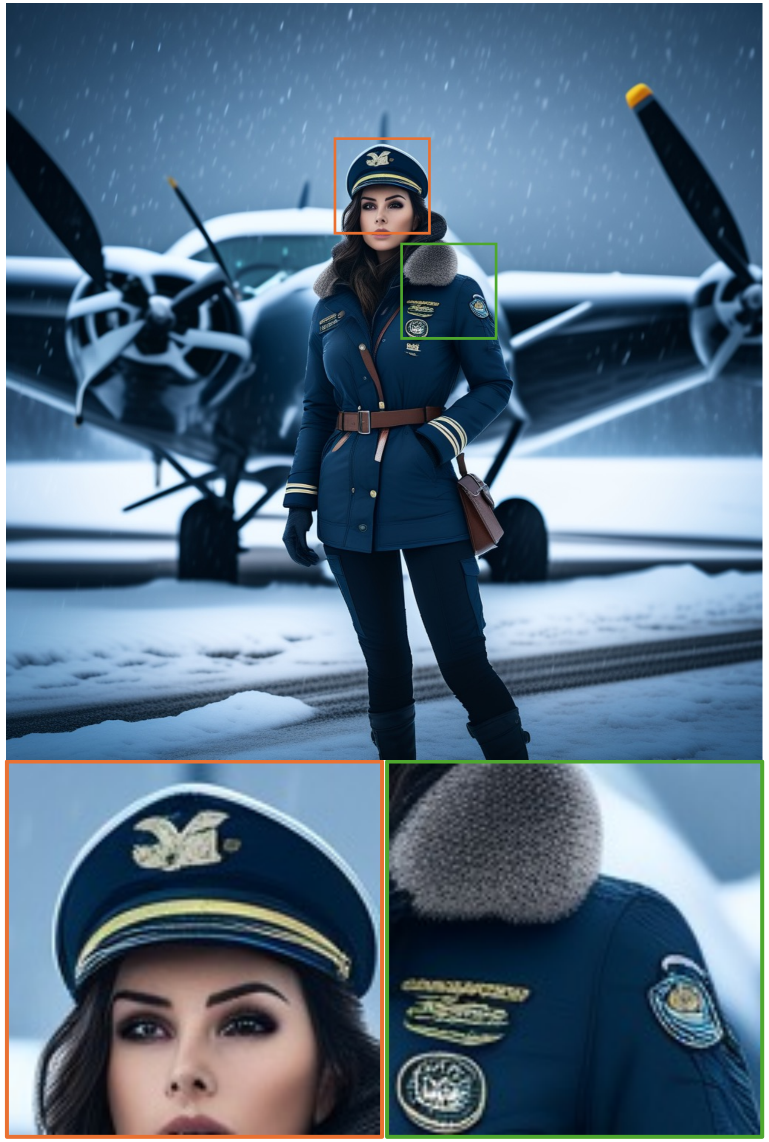


Real-ESRGAN (8192 × 8192)

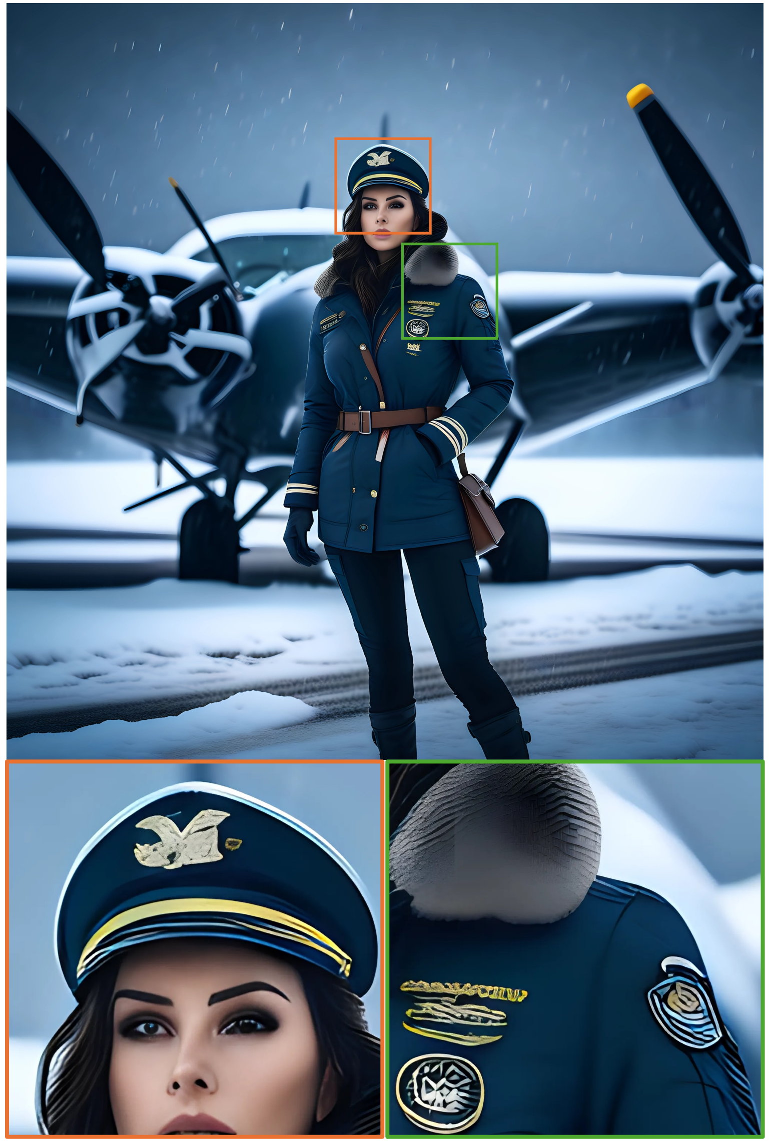


FreeScale (8192 × 8192)

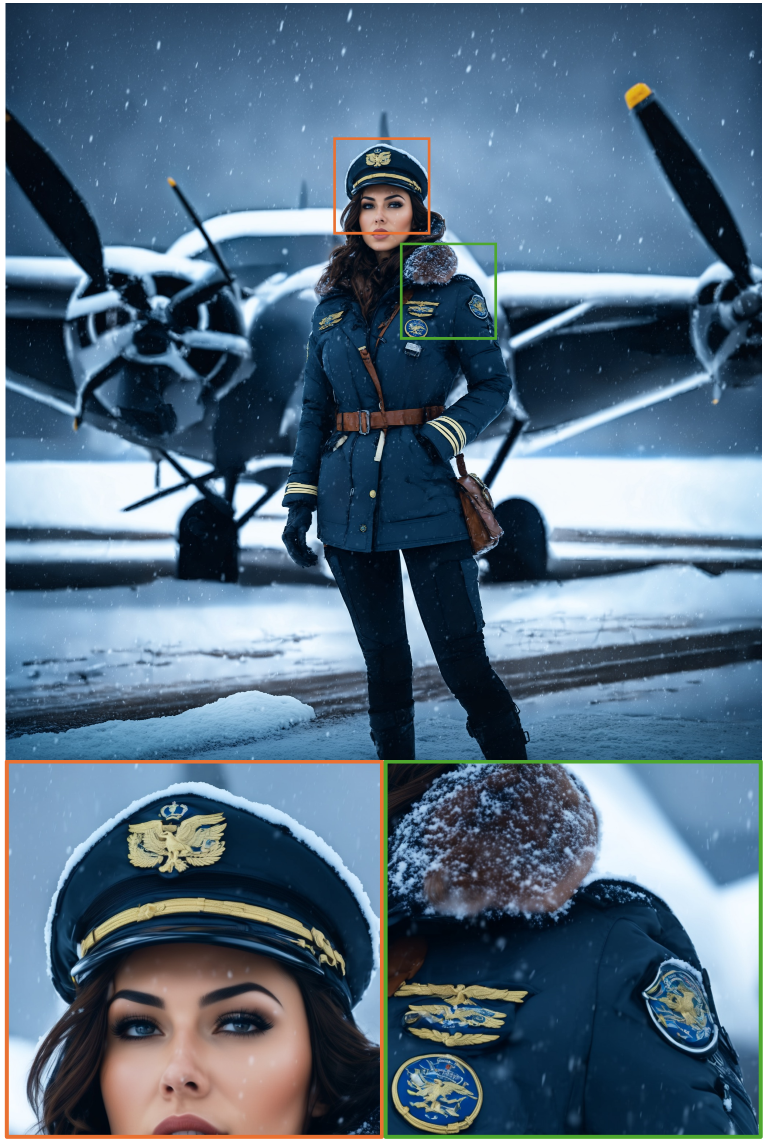
Difference with SR



SDXL (1024 × 1024)



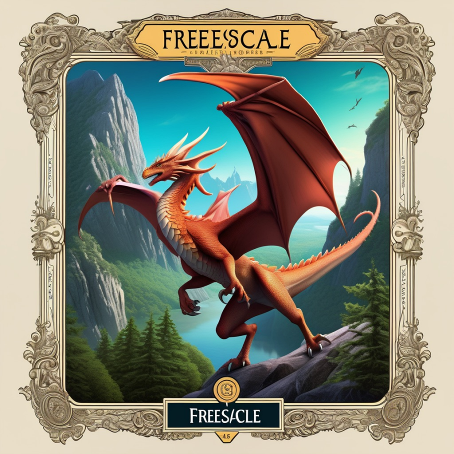
Real-ESRGAN (8192 × 8192)



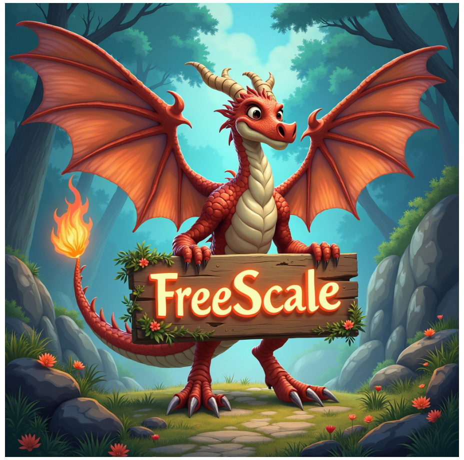
FreeScale (8192 × 8192)

Img-to-Img Generation

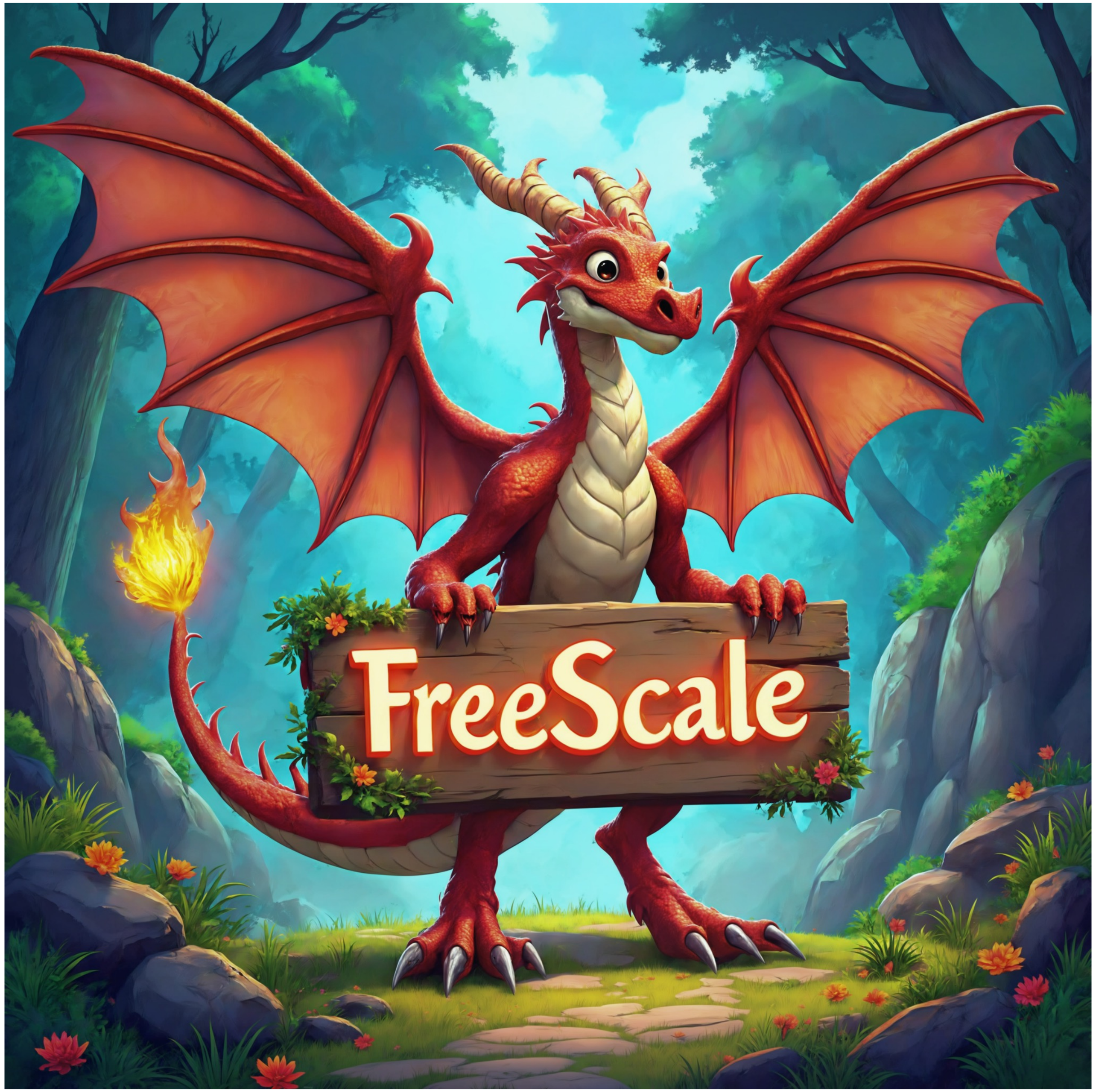
SDXL 1024x1024



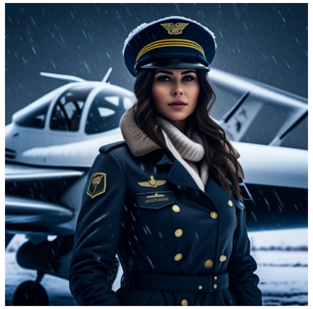
FLUX 1024x1024



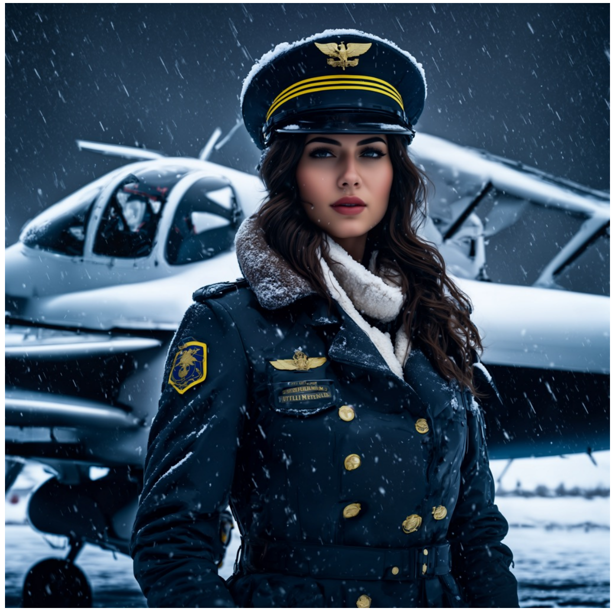
FLUX+FreeScale(SDXL) 8192x8192



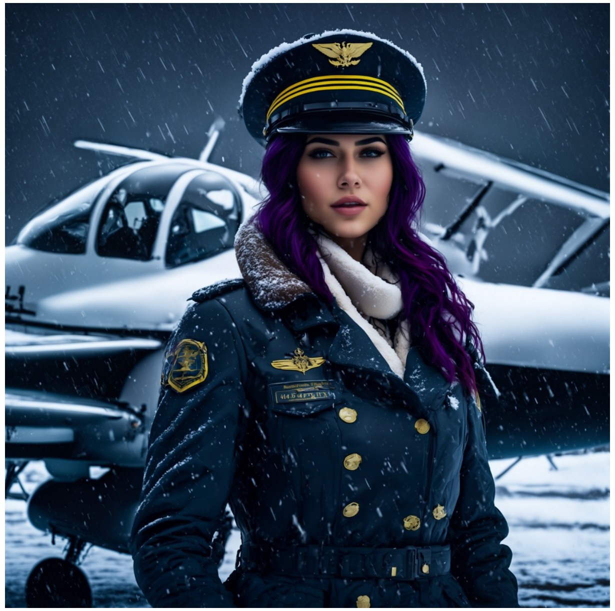
Local Semantic Editing



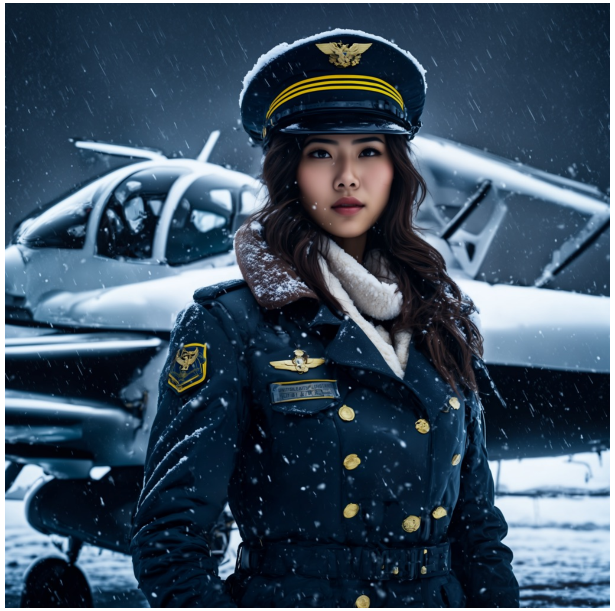
1× Result
(1024x1024)



No Editing
(4096x4096)

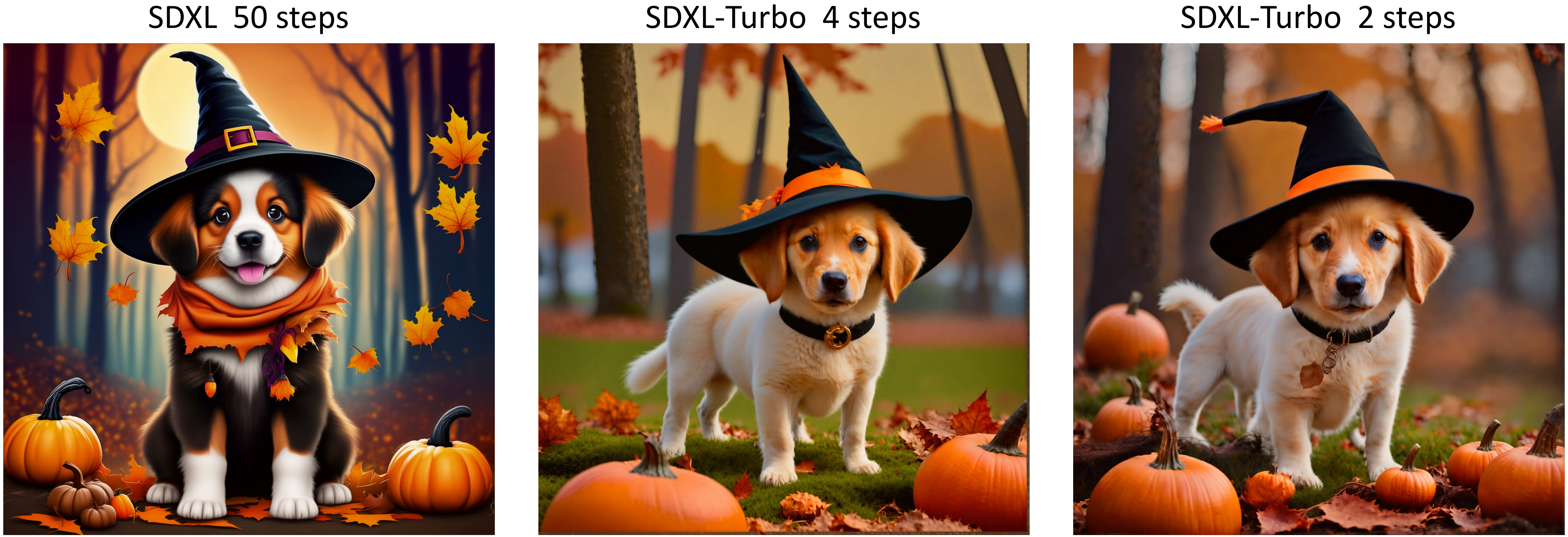


Hair Editing
(4096x4096)



Face Editing
(4096x4096)

FreeScale + SDXL-Turbo



SDXL 50 steps

SDXL-Turbo 4 steps

SDXL-Turbo 2 steps

“A cute and adorable fluffy puppy wearing a witch hat in a Halloween autumn evening forest, falling autumn leaves, brown acorns on the ground, Halloween pumpkins spiderwebs, bats, and a witch’s broom.”

More 8K Results

More 8K Results

Limitation and Future Work

1. Inference Cost:
   8k image = 55 GB and 1 hour on NVIDIA A800.

2. Knowledge Limitation:
   The endless higher-resolution result will have either the same level of detail or unnatural messy detail.

3. Generalization:
   Current version does not work for DiT-based models.
   (Solved in extended work, CineScale)

Thanks for Watching

Project Page

Code Repo