A. Details on Evaluation Data

Our benchmark comprises a total of 31 tasks, with each task containing between 50 and 500 evaluation cases. We provide visualizations of the conditions and exemplary generation results for each task in Fig. 7. Specifically, an evaluation case should comprise *(Introduction, Target Caption, Source Image, Source Mask, Reference Images)* to facilitate the generation and evaluation process. A detailed illustration of a complete evaluation case is presented in Tab. 1. Most of the existing image generation models support only one or a few of the 31 evaluation tasks. We provide a detailed summary of the tasks supported and unsupported by the 10 evaluated models in Tab. 2.

B. Details on Evaluation Dimensions

B.1. Aesthetic Quality

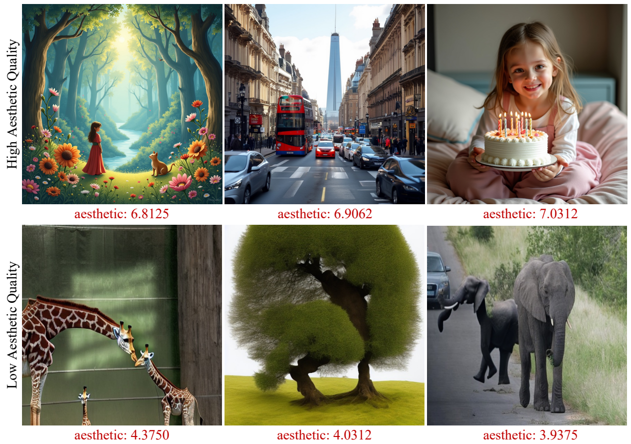


Figure 1. **Visualization of Aesthetic Quality.** Images that receive high aesthetic scores exhibit artistic appeal, whereas those with low aesthetic scores tend to appear unattractive.

Aesthetic Quality evaluates the principles of photographic composition, considering color harmony, subject arrangement, and the overall artistic impression of the image. We utilize a SigLip-based image aesthetic quality predictor to assess the aesthetic score of the generated image. The model produces a rating on a scale from 0 to 10, which we linearly normalize to a range of [0, 1] by dividing the raw score by 10.

$$S_{\text{AES}} = \frac{f_{\text{AES}}(\mathbf{I})}{10} \tag{1}$$

B.2. Imaging Quality

Imaging quality primarily examines the low-level characteristics of the generated image, such as edge sharpness, distortion, overexposure, noise, and blur. We employ the MUSIQ image quality predictor trained on the Koniq dataset, as implemented in IQA-Pytorch [10]. For consistency and fairness in comparison, we resize the height of all generated images to 1024 pixels before inputting them into the model to assess imaging quality. This approach inherently favors high-resolution images as they typically exhibit superior imaging quality compared to low-resolution images. The model produces a score on a scale from 0 to 100, which we linearly normalize to a range of [0, 1] by dividing the raw score by 100.

Table 1. **Detail of a complete evaluation case.**

<ItemID>:	b9de809c702c8cf23428ec175af3b0b9
<TaskLevel1>:	Reference Editing
<TaskLevel2>:	Subject Reference Editing
<Task>:	Subject-guided Inpainting
<SourceImageType>:	Real Image
<RegionBased>:	True
<SourceImage>:	images/reference_editing/subject_reference_editing/subject_guided_inpainting/b9de809c702c8cf23428ec175af3b0b9_src.png
<SourceMask>:	images/reference_editing/subject_reference_editing/subject_guided_inpainting/b9de809c702c8cf23428ec175af3b0b9_mask.png
<ReferenceImages>:	[”images/reference_editing/subject_reference_editing/subject_guided_inpainting/b9de809c702c8cf23428ec175af3b0b9_ref1.png”]
<Instruction>:	Take `<REF_1>` as a reference to repaint the masked part of `<SOURCE>`.
<SourceCaption>:	Eye-level view of a street scene featuring a fire hydrant in the foreground.
<TargetCaption>:	A small, brightly colored toy car sits on a weathered asphalt surface, positioned slightly off-center in the foreground. The car is predominantly red and yellow, with green accents.

$$S_{\text{IMG}} = \frac{f_{\text{MUSIQ}}(\mathbf{I})}{100} \tag{2}$$

B.3. Prompt Following

The prompt-following score evaluates the degree to which the generated image aligns with the provided textual instructions or descriptions. For image creation tasks and controllable generation

Table 2. Task-model correspondence.

Evaluation Tasks				OmniGen [54]	ACE [18]	FLUX [25]	OminiControl [47]	InstructPix2Pix [5]	MagicBrush [57]	UltraEdit [60]	FLUX-Control [48]	IP-Adapter [56]	ACE++ [29]
Creating	No-Ref		(1) Text-to-Image Creating	✓	✓	✓	×	×	×	×	×	×	×
	Ref		(2) Face Reference Creating	✓	✓	×	×	×	×	×	×	✓	✓
			(3) Style Reference Creating	✓	✓	×	×	×	×	×	×	✓	×
			(4) Subject Reference Creating	✓	✓	×	✓	×	×	×	×	✓	✓
Editing	No-Ref	Global	(5) Color Editing	✓	✓	×	×	✓	✓	✓	×	×	×
			(6) Motion Editing	✓	✓	×	×	✓	✓	✓	×	×	×
			(7) Face Editing	✓	✓	×	×	✓	✓	✓	×	×	×
			(8) Texture Editing	✓	✓	×	×	✓	✓	✓	×	×	×
			(9) Style Editing	✓	✓	×	×	✓	✓	✓	×	×	×
			(10) Scene Editing	✓	✓	×	×	✓	✓	✓	×	×	×
			(11) Subject Addition	✓	✓	×	×	✓	✓	✓	×	×	×
			(12) Subject Removal	✓	✓	×	×	✓	✓	✓	×	×	×
			(13) Subject Change	✓	✓	×	×	✓	✓	✓	×	×	×
			(14) Text Render	✓	✓	×	×	✓	✓	✓	×	×	×
			(15) Text Removal	✓	✓	×	×	✓	✓	✓	×	×	×
			(16) Composite Editing	✓	✓	×	×	✓	✓	✓	×	×	×
		Local	(17) Inpainting	✓	✓	×	×	×	×	✓	×	×	✓
			(18) Outpainting	✓	✓	×	×	×	×	✓	×	×	✓
			(19) Local Subject Addition	✓	✓	×	×	×	×	✓	×	×	✓
			(20) Local Subject Removal	✓	✓	×	×	×	×	✓	×	×	✓
			(21) Local Text Removal	✓	✓	×	×	×	×	✓	×	×	✓
			(22) Local Text Render	✓	✓	×	×	×	×	✓	×	×	✓
	Ref	Controllable	(23) Pose-guided Generation	✓	✓	×	×	×	×	×	✓	×	×
			(24) Edge-guided Generation	✓	✓	×	✓	×	×	×	✓	×	×
			(25) Depth-guided Generation	✓	✓	×	✓	×	×	×	✓	×	×
			(26) Image Colorization	✓	✓	×	✓	×	×	×	✓	×	×
			(27) Image Deblurring	✓	✓	×	✓	×	×	×	✓	×	×
			(28) Style Transfer	✓	✓	×	×	×	×	×	×	×	×
		Subject	(29) Subject-guided Inpainting	✓	✓	×	×	×	×	×	×	×	✓
			(30) Virtual Try On	✓	✓	×	×	×	×	×	×	×	✓
			(31) Face Swap	✓	✓	×	×	×	×	×	×	×	✓

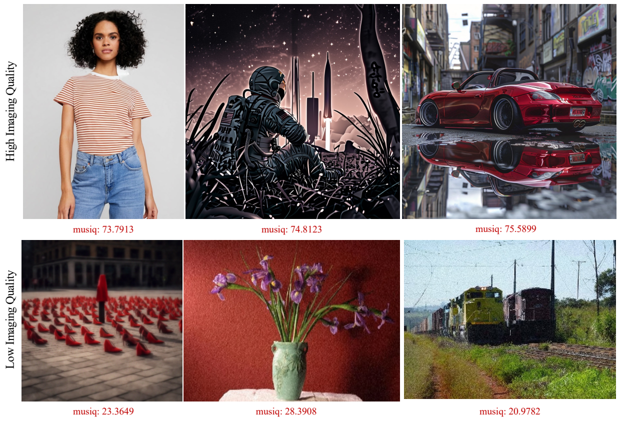
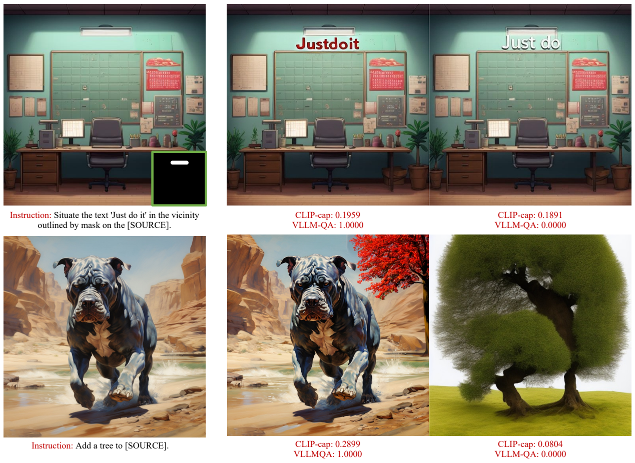


Figure 2. **Visualization of Imaging Quality.** Images that achieve high imaging quality scores are typically clear and possess sharp edges, whereas those with low scores tend to appear blurry and noisy.



Instruction: Situate the text 'Just do it' in the vicinity outlined by mask on the [SOURCE].

CLIP-cap: 0.1959
VLLM-QA: 1.0000

CLIP-cap: 0.1891
VLLM-QA: 0.0000

Instruction: Add a tree to [SOURCE].

CLIP-cap: 0.2899
VLLMQA: 1.0000

CLIP-cap: 0.0804
VLLM-QA: 0.0000

Figure 3. **Visualization of Prompt Following.** Both the CLIP-cap and VLLM-QA metrics effectively capture the successful execution of instructions.

tasks, we compute the CLIP [36] similarity between the target caption and the generated image directly. The prompt-following score is then obtained by normalizing the CLIP similarity, specifically by dividing it by 0.5.

$$S_{\text{PF}} = \frac{\langle d_{\text{prompt}} \cdot d_{\mathbf{I}} \rangle}{0.5} \tag{3}$$

Notably, for the Image Colorization and Image Deblurring tasks, CLIP similarity alone is insufficient to accurately assess prompt-following capability. For the Image Colorization task, the colorfulness score must also be considered an essential metric, leading us to adapt the prompt-following score accordingly:

$$S_{\text{PF}}^{\text{colorsize}} = \frac{\langle d_{\text{prompt}} \cdot d_{\mathbf{I}} \rangle}{0.5} + s_{\text{color}} \tag{4}$$

In the case of the Image Deblurring task, the Imaging score serves as the prompt-following metric, as the primary objective is to enhance image quality.

$$S_{\text{PF}}^{\text{deblur}} = S_{\text{IMG}} \tag{5}$$

For image editing tasks, relying solely on CLIP similarity is insufficient to determine whether instructions have been correctly executed. To address this, we introduce a novel VLLM-based metric called VLLM-QA to assess the success of instruction alignment. We employ the QWEN2-VL-72B [51] model as our QA tool, prompting it with all relevant input components, including the instruction, source image, reference images, source mask, and the generated image. The model is tasked with evaluating whether the instruction has been accurately implemented; it returns a score of 1 for success and 0 otherwise. We calculate the VLLM-QA score by averaging the results across all cases within a task. Subsequently, the prompt-following score is determined as follows:

$$S_{\text{PF}} = \frac{\frac{\langle d_{\text{prompt}} \cdot d_{\mathbf{I}} \rangle}{0.5} + f_{\text{QWEN}}(\cdot)}{2} \tag{6}$$

B.4. Source Consistency

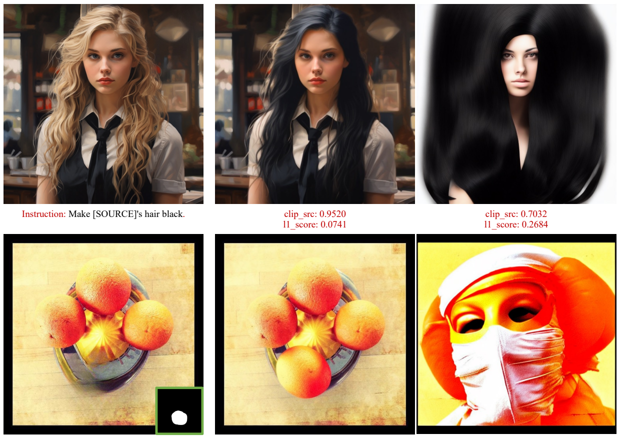


Figure 4. **Visualization of Source Consistency.** Images that exhibit strong pixel alignment with the source image attain higher CLIP-src scores and lower L1 scores. These outcomes underscore the effectiveness of our evaluation of Source Consistency.

For image editing tasks, it is crucial to maintain the pixels that are unrelated to the editing instructions unchanged. To evaluate the models' ability to preserve pixel alignment, we compute both the CLIP similarity and the mean L1 distance between the generated image and the source image. The Source Consistency score is then calculated as follows:

$$S_{\text{SRC}} = \frac{\langle d_{\mathbf{I}_{\text{src}}} \cdot d_{\mathbf{I}} \rangle + 1 - L1(\mathbf{I}_{\text{src}}, \mathbf{I})}{2} \tag{7}$$

B.5. Reference Consistency

Reference consistency evaluates the semantic alignment between the reference image and the generated image across specific aspects, such as face, style, and subject. To achieve this, we utilize different encoders to extract embeddings from both the reference image and the generated image. We then assess the reference consistency in these three dimensions by calculating the feature similarity between the extracted embeddings:

$$S_{\text{REF}} = \langle d_{\mathbf{I}_{\text{ref}}} \cdot d_{\mathbf{I}} \rangle \tag{8}$$

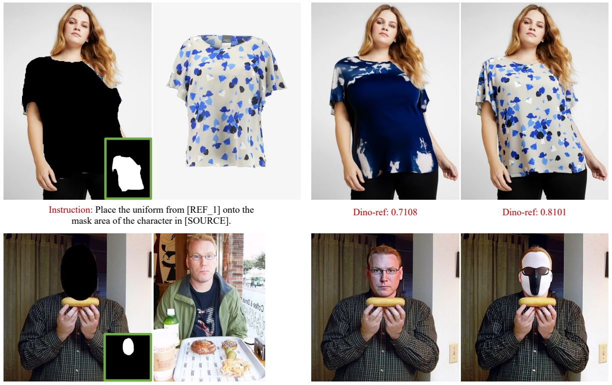

Instruction: Place the uniform from [REF_1] onto the mask area of the character in [SOURCE].
Dino-ref: 0.7108
Dino-ref: 0.8101
Instruction: Change the facial features in [SOURCE] mask area to match those in [REF_1].
Face-ref: 0.5662
Face-ref: 0.1458

Figure 5. **Visualization of Reference Consistency.** Images that maintain identity preservation with the reference image achieve higher CLIP-ref scores, highlighting the effectiveness of our Reference Consistency evaluation.

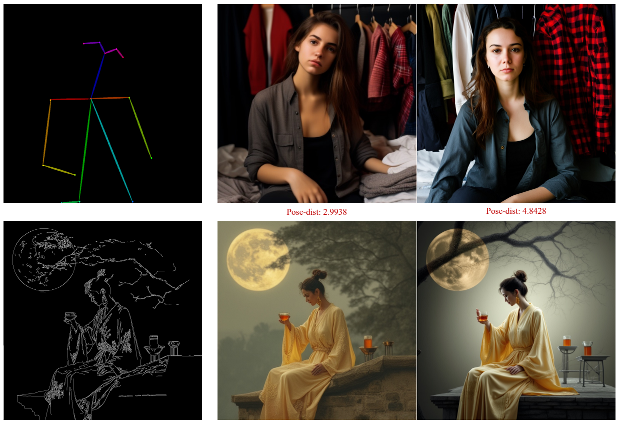


Figure 6. **Visualization of Controllability.** The Pose-dist and Canny-dist metrics effectively indicate controllability, with lower values generally signifying greater controllability.

B.6. Controllability

Controllability evaluates the alignment of low-level features in the generated image with the input condition image. For tasks such as Pose, Depth, Edge-guided Generation, and Image Colorization, we extract the relevant low-level feature map from the generated image and calculate the mean L1 score between this feature map and the input condition image. The controllability score is then determined as follows:

$$S_{\text{CTRL}} = 1 - (f_{\text{enc}}(\mathbf{I}) - \mathbf{I}_{\text{src}}) \tag{9}$$

While for Image Deblurring task, we employ the SSIM score as the controllability score:

$$S_{\text{CTRL}}^{\text{deblur}} = \text{SSIM}(\mathbf{I}, \mathbf{I}_{\text{src}}) \tag{10}$$

C. Details on Model Performance per Task

In this section, we present the detailed evaluation results for each metric across all tasks and models. The results for No-ref Image Creating are shown in Tab. 3. The results for Ref Image Creating are provided in Tab. 6. For No-ref Image Editing, the results are detailed in Tab. 4, Tab. 7, and Tab. 8. The results for Ref Image Editing are reported in Tab. 5.

Table 3. Metrics on No-ref Image Creating Task (Task 1).

Models	Aesthetic Score↑	Imaging Score↑	CLIP-cap↑
ACE	5.485	53.403	0.283
OmniGen	6.107	72.615	**0.285**
FLUX	**6.175**	**73.480**	**0.285**

Table 4. Metrics on Controllable Generation Tasks (Tasks 23-27).

Models	**Task 23: Pose-guided Generation**			
	Aesthetic Score↑	Imaging Score↑	CLIP-cap↑	L1-src↓
ACE	**5.568**	50.253	**0.299**	**0.009**
OmniGen	5.365	**61.463**	0.298	0.015
FLUX-Control	5.538	56.010	0.298	0.015

Models	**Task 24: Edge-guided Generation**			
	Aesthetic Score↑	Imaging Score↑	CLIP-cap↑	L1-src↓
ACE	5.319	49.506	0.298	0.091
OmniGen	4.897	**66.168**	0.293	0.102
FLUX-Control	5.493	54.225	0.296	0.104
OminiControl	**5.507**	51.301	**0.299**	**0.087**

Models	**Task 25: Depth-guided Generation**			
	Aesthetic Score↑	Imaging Score↑	CLIP-cap↑	L1-src↓
ACE	5.505	51.948	0.291	**0.095**
OmniGen	4.809	**60.266**	0.266	0.131
FLUX-Control	**5.844**	59.578	0.295	0.123
OminiControl	5.762	57.305	**0.296**	0.098

Models	**Task 26: Image Colorization**				
	Aesthetic Score↑	Imaging Score↑	CLIP-cap↑	Color Score↑	L1-src↓
ACE	5.325	50.484	0.295	**0.278**	0.059
OmniGen	5.275	**61.076**	0.289	0.189	0.185
FLUX-Control	**5.371**	51.891	**0.302**	0.210	0.067
OminiControl	5.272	50.995	0.301	0.161	**0.029**

Models	**Task 27: Image Deblurring**		
	Aesthetic Score↑	Imaging Score↑	SSIM↑
ACE	**5.556**	**50.229**	0.582
OmniGen	5.133	48.144	0.350
FLUX-Control	5.342	45.063	0.540
OminiControl	4.249	30.327	**0.650**

Table 5. Metrics on Ref Image Editing Tasks (Tasks 28-31).

Models	**Task 28: Style Transfer**						
	Aesthetic Score↑	Imaging Score↑	CLIP -cap↑	VLLM -QA↑	Style -ref↑	CLIP -src↑	L1-src↓
ACE	**5.346**	53.030	0.189	**0.323**	0.234	**0.762**	**0.186**
OmniGen	5.045	**62.995**	**0.193**	0.290	**0.359**	0.680	0.277

Models	**Task 29: Subject-guided Inpainting**						
	Aesthetic Score↑	Imaging Score↑	CLIP -cap↑	VLLM -QA↑	DINO -ref↑	CLIP -src↑	L1-src↓
ACE	4.812	52.544	**0.197**	0.171	0.562	**0.766**	**0.015**
OmniGen	4.459	59.995	0.186	0.093	0.555	0.642	0.149
ACE++	**4.835**	**63.419**	0.186	**0.257**	**0.563**	0.753	0.040

Models	**Task 31: Face Swap**						
	Aesthetic Score↑	Imaging Score↑	CLIP -cap↑	VLLM -QA↑	Face -ref↑	CLIP -src↑	L1-src↓
ACE	4.983	56.985	**0.232**	0.400	0.250	**0.763**	**0.018**
OmniGen	4.309	64.021	0.217	**0.484**	**0.477**	0.661	0.112
ACE++	**5.034**	**64.963**	0.231	0.442	0.378	0.760	0.054

Models	**Task 30: Virtual Try On**						
	Aesthetic Score↑	Imaging Score↑	CLIP -cap↑	VLLM -QA↑	DINO -ref↑	CLIP -src↑	L1-src↓
ACE	**4.837**	64.723	0.231	0.629	0.751	**0.889**	**0.006**
OmniGen	4.696	73.313	0.235	0.722	0.744	0.847	0.058
ACE++	4.577	**73.525**	**0.243**	**0.804**	**0.763**	0.882	0.029

Table 6. Metrics on Ref Image Creating Tasks (Tasks 2-4).

Models	**Task 2: Face Reference Creating**				**Task 3: Style Reference Creating**				**Task 4: Subject Reference Creating**			
	Aesthetic Score↑	Imaging Score↑	CLIP-cap↑	Face-ref↑	Aesthetic Score↑	Imaging Score↑	CLIP-cap↑	Style-ref↑	Aesthetic Score↑	Imaging Score↑	CLIP-cap↑	DINO-ref↑
ACE	5.352	54.953	0.265	0.329	5.312	58.960	0.116	**0.802**	5.228	55.748	0.249	**0.878**
OmniGen	**5.790**	**72.667**	**0.270**	0.573	**5.785**	**70.827**	**0.215**	0.432	**5.821**	71.355	**0.266**	0.753
IP-Adapter	5.055	64.239	0.254	**0.633**	5.773	69.629	0.144	0.749	5.726	70.329	0.242	0.841
ACE++	5.508	67.900	0.261	0.506	-	-	-	-	5.198	62.751	0.238	0.852
OminiControl	-	-	-	-	-	-	-	-	5.651	**72.273**	0.264	0.783

Table 7. Metrics on Global Editing Tasks (Tasks 5-16).

Models	**Task 5: Color Editing**						**Task 6: Motion Editing**						**Task 7: Face Editing**					
	Aesthetic Score↑	Imaging Score↑	CLIP-cap↑	VLLM-QA↑	CLIP-src↑	L1-src↓	Aesthetic Score↑	Imaging Score↑	CLIP-cap↑	VLLM-QA↑	CLIP-src↑	L1-src↓	Aesthetic Score↑	Imaging Score↑	CLIP-cap↑	VLLM-QA↑	CLIP-src↑	L1-src↓
ACE	**5.244**	55.219	**0.285**	**0.896**	**0.919**	**0.080**	**5.146**	57.679	**0.278**	**0.354**	**0.946**	**0.033**	4.798	56.851	**0.268**	**0.796**	**0.899**	**0.046**
OmniGen	4.918	**63.562**	0.277	0.789	0.880	0.119	4.927	**61.038**	0.262	0.329	0.870	0.106	4.735	**63.584**	0.247	0.636	0.818	0.095
InstructPix2Pix	4.990	53.124	0.267	0.452	0.828	0.217	4.796	57.453	0.211	0.081	0.719	0.134	**4.920**	57.941	0.192	0.364	0.669	0.151
MagicBrush	4.826	51.677	0.267	0.604	0.854	0.094	4.620	53.121	0.254	0.267	0.826	0.081	4.636	55.833	0.258	0.660	0.836	0.054
UltraEdit	5.136	52.398	0.274	0.485	0.864	0.098	4.970	55.514	0.266	0.199	0.871	0.059	4.774	57.159	0.247	0.655	0.786	0.057

Models	**Task 8: Texture Editing**						**Task 9: Style Editing**						**Task 10: Scene Editing**					
	Aesthetic Score↑	Imaging Score↑	CLIP-cap↑	VLLM-QA↑	CLIP-src↑	L1-src↓	Aesthetic Score↑	Imaging Score↑	CLIP-cap↑	VLLM-QA↑	CLIP-src↑	L1-src↓	Aesthetic Score↑	Imaging Score↑	CLIP-cap↑	VLLM-QA↑	CLIP-src↑	L1-src↓
ACE	**5.408**	57.106	**0.276**	0.605	**0.918**	**0.060**	**4.967**	51.081	**0.258**	0.470	**0.781**	0.158	5.076	47.345	**0.253**	0.392	**0.902**	**0.075**
OmniGen	5.151	**64.069**	0.257	0.558	0.819	0.156	4.935	**60.567**	0.250	**0.478**	0.763	0.183	**5.109**	**55.674**	0.246	0.414	0.806	0.169
InstructPix2Pix	4.847	59.220	0.240	0.422	0.703	0.193	4.630	48.674	0.228	0.416	0.627	0.218	5.048	45.324	0.224	0.381	0.657	0.219
MagicBrush	4.720	52.909	0.245	0.463	0.796	0.122	4.227	46.647	0.184	0.140	0.600	0.249	4.592	44.262	0.239	**0.464**	0.725	0.189
UltraEdit	5.148	54.875	0.270	**0.714**	0.821	0.093	4.697	49.067	0.246	0.414	0.726	**0.093**	5.023	44.961	0.255	0.453	0.764	0.098

Models	**Task 11: Subject Addition**						**Task 12: Subject Removal**						**Task 13: Subject Change**					
	Aesthetic Score↑	Imaging Score↑	CLIP-cap↑	VLLM-QA↑	CLIP-src↑	L1-src↓	Aesthetic Score↑	Imaging Score↑	CLIP-cap↑	VLLM-QA↑	CLIP-src↑	L1-src↓	Aesthetic Score↑	Imaging Score↑	CLIP-cap↑	VLLM-QA↑	CLIP-src↑	L1-src↓
ACE	4.920	50.514	**0.274**	**0.619**	**0.888**	**0.045**	4.877	45.559	**0.253**	**0.834**	0.855	**0.053**	**5.018**	52.386	**0.274**	0.500	**0.881**	**0.070**
OmniGen	**4.987**	**58.151**	0.266	0.611	0.877	0.077	4.884	**54.001**	0.231	0.611	0.830	0.107	4.997	**59.282**	0.262	0.460	0.812	0.115
InstructPix2Pix	4.884	52.320	0.205	0.234	0.703	0.144	4.827	48.625	0.170	0.119	0.711	0.141	4.746	53.884	0.229	0.360	0.691	0.179
MagicBrush	4.656	46.127	0.272	0.594	0.866	0.061	4.672	45.197	0.231	0.322	0.864	0.069	4.291	48.950	0.257	0.500	0.756	0.123
UltraEdit	4.932	47.651	0.259	0.537	0.830	0.064	**4.974**	47.308	0.223	0.256	**0.873**	0.056	4.868	51.984	0.269	**0.540**	0.788	0.082

Models	**Task 14: Text Render**						**Task 15: Text Removal**						**Task 16: Composite Editing**					
	Aesthetic Score↑	Imaging Score↑	CLIP-cap↑	VLLM-QA↑	CLIP-src↑	L1-src↓	Aesthetic Score↑	Imaging Score↑	CLIP-cap↑	VLLM-QA↑	CLIP-src↑	L1-src↓	Aesthetic Score↑	Imaging Score↑	CLIP-cap↑	VLLM-QA↑	CLIP-src↑	L1-src↓
ACE	3.981	51.104	**0.263**	0.517	0.800	**0.052**	**4.842**	49.714	**0.270**	**0.754**	**0.883**	**0.037**	**5.475**	49.984	0.270	0.420	**0.797**	0.194
OmniGen	4.351	**57.420**	**0.263**	**0.596**	0.815	0.075	4.500	**57.211**	0.223	0.330	0.767	0.125	5.259	**62.885**	0.272	**0.567**	0.753	0.229
InstructPix2Pix	**4.712**	51.201	0.213	0.010	0.718	0.187	4.400	44.069	0.194	0.147	0.655	0.163	4.827	50.006	0.258	0.280	0.698	**0.237**
MagicBrush	4.458	45.903	0.261	0.099	**0.845**	0.088	4.359	44.484	0.260	0.529	0.838	0.063	4.665	47.646	0.245	0.070	0.732	0.185
UltraEdit	4.465	46.965	0.262	0.187	0.813	0.059	4.640	47.908	0.255	0.246	0.861	0.044	5.180	48.372	**0.274**	0.395	0.731	0.147

Table 8. Metrics on Local Editing Tasks (Tasks 17-22).

Models	**Task 17: Inpainting**						**Task 18: Outpainting**						**Task 19: Local Subject Addition**					
	Aesthetic Score↑	Imaging Score↑	CLIP-cap↑	VLLM-QA↑	CLIP-src↑	L1-src↓	Aesthetic Score↑	Imaging Score↑	CLIP-cap↑	VLLM-QA↑	CLIP-src↑	L1-src↓	Aesthetic Score↑	Imaging Score↑	CLIP-cap↑	VLLM-QA↑	CLIP-src↑	L1-src↓
ACE	4.878	51.793	0.269	0.833	0.785	0.024	5.514	50.403	0.287	0.376	0.891	0.017	4.965	51.704	0.272	0.555	0.897	0.029
OmniGen	4.545	59.264	0.238	0.524	0.734	0.108	5.442	**65.758**	0.265	0.326	0.802	0.114	4.584	58.911	0.249	0.479	0.814	0.066
ACE++	**5.064**	**61.661**	**0.272**	**0.910**	0.776	**0.016**	**5.644**	64.156	**0.289**	**0.531**	0.908	**0.010**	**5.014**	**62.083**	0.268	**0.785**	0.894	**0.018**
UltraEdit	3.817	46.284	0.250	0.180	**0.952**	0.019	4.498	43.968	0.274	0.220	**0.945**	0.018	4.881	47.855	**0.275**	0.555	**0.909**	0.021

Models	**Task 20: Local Subject Removal**						**Task 21: Local Text Render**						**Task 22: Local Text Removal**					
	Aesthetic Score↑	Imaging Score↑	CLIP-cap↑	VLLM-QA↑	CLIP-src↑	L1-src↓	Aesthetic Score↑	Imaging Score↑	CLIP-cap↑	VLLM-QA↑	CLIP-src↑	L1-src↓	Aesthetic Score↑	Imaging Score↑	CLIP-cap↑	VLLM-QA↑	CLIP-src↑	L1-src↓
ACE	4.996	47.011	**0.258**	**0.757**	0.852	0.024	4.275	43.159	0.276	**0.791**	0.860	0.016	**4.896**	49.766	**0.273**	**0.801**	0.888	0.033
OmniGen	4.792	54.320	0.238	0.658	0.787	0.061	4.015	42.527	0.261	0.380	0.815	0.066	4.487	56.398	0.246	0.674	0.793	0.097
ACE++	**5.061**	**61.614**	0.229	0.312	**0.901**	**0.017**	4.231	**43.276**	**0.277**	0.834	0.899	**0.012**	4.694	**59.636**	0.260	0.704	0.905	**0.017**
UltraEdit	4.858	48.748	0.226	0.287	0.888	0.018	**4.506**	38.887	**0.277**	0.098	**0.946**	0.014	4.665	47.294	0.264	0.714	**0.910**	0.023

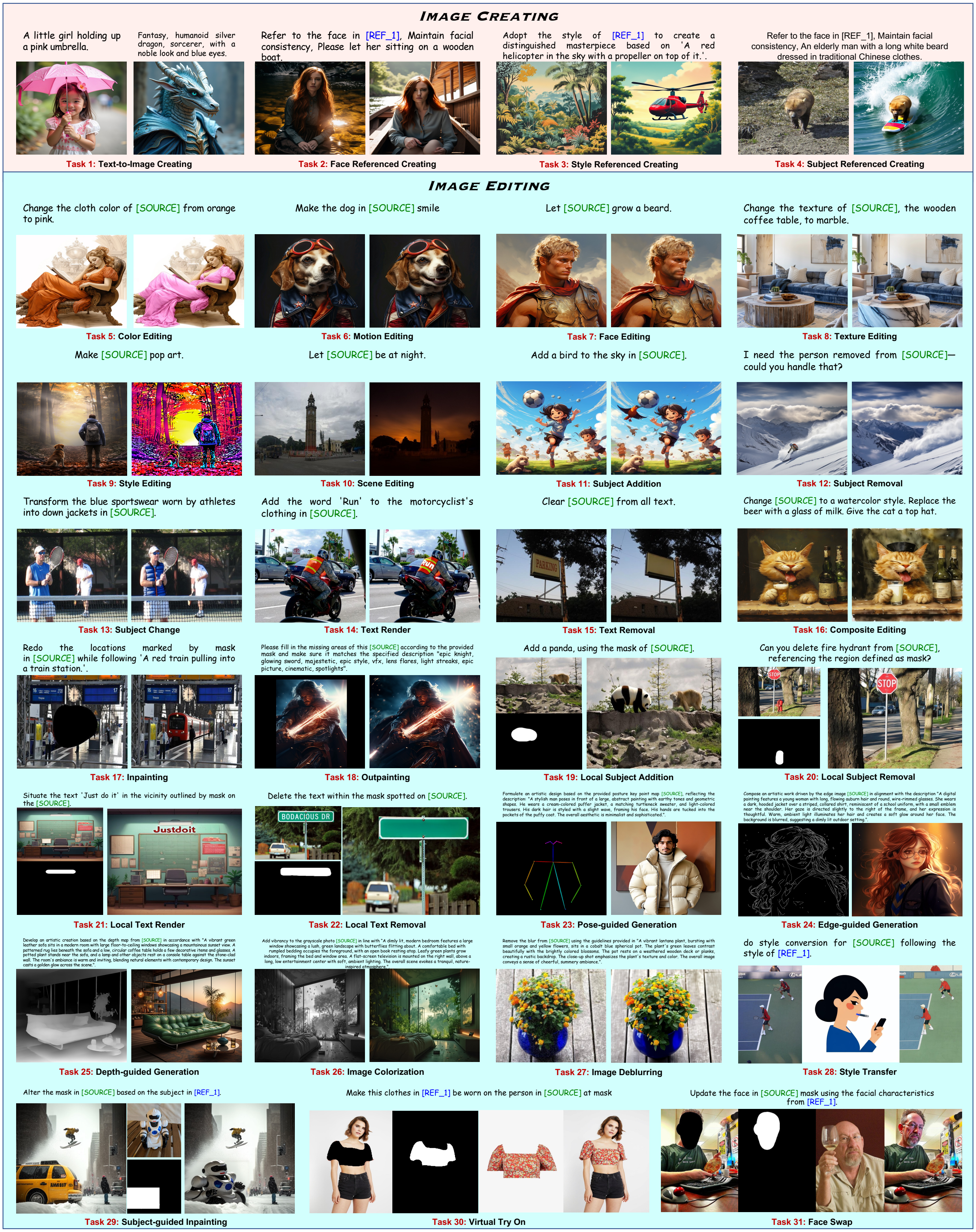


IMAGE CREATING
A little girl holding up a pink umbrella.
Fantasy, humanoid silver dragon, sorcerer, with a noble look and blue eyes.
Task 1: Text-to-Image Creating
Refer to the face in [REF_1], Maintain facial consistency, Please let her sitting on a wooden boat.
Task 2: Face Referenced Creating
Adopt the style of [REF_1] to create a distinguished masterpiece based on 'A red helicopter in the sky with a propeller on top of it.'.
Task 3: Style Referenced Creating
Refer to the face in [REF_1], Maintain facial consistency, An elderly man with a long white beard dressed in traditional Chinese clothes.
Task 4: Subject Referenced Creating
IMAGE EDITING
Change the cloth color of [SOURCE] from orange to pink.
Task 5: Color Editing
Make the dog in [SOURCE] smile
Task 6: Motion Editing
Let [SOURCE] grow a beard.
Task 7: Face Editing
Change the texture of [SOURCE], the wooden coffee table, to marble.
Task 8: Texture Editing
Make [SOURCE] pop art.
Task 9: Style Editing
Let [SOURCE] be at night.
Task 10: Scene Editing
Add a bird to the sky in [SOURCE].
Task 11: Subject Addition
I need the person removed from [SOURCE]—could you handle that?
Task 12: Subject Removal
Transform the blue sportswear worn by athletes into down jackets in [SOURCE].
Task 13: Subject Change
Add the word 'Run' to the motorcyclist's clothing in [SOURCE].
Task 14: Text Render
Clear [SOURCE] from all text.
Task 15: Text Removal
Change [SOURCE] to a watercolor style. Replace the beer with a glass of milk. Give the cat a top hat.
Task 16: Composite Editing
Redo the locations marked by mask in [SOURCE] while following 'A red train pulling into a train station.'.
Task 17: Inpainting
Please fill in the missing areas of this [SOURCE] according to the provided mask and make sure it matches the specified description "epic knight, glowing sword, majestetic, epic style, vfx, lens flares, light streaks, epic picture, cinematic, spotlights".
Task 18: Outpainting
Add a panda, using the mask of [SOURCE].
Task 19: Local Subject Addition
Can you delete fire hydrant from [SOURCE], referencing the region defined as mask?
Task 20: Local Subject Removal
Situate the text 'Just do it' in the vicinity outlined by mask on the [SOURCE].
Justdoit
Task 21: Local Text Render
Delete the text within the mask spotted on [SOURCE].
BODACIOUS DR
Task 22: Local Text Removal
Formulate an artistic design based on the provided posture key point map [SOURCE], reflecting the description: "A stylish man poses in front of a large, abstract painting with earthy tones and geometric shapes. He wears a cream-colored puffer jacket, a matching turtleneck sweater, and light-colored trousers. His dark hair is styled with a slight wave, framing his face. His hands are tucked into the pockets of the puffy coat. The overall aesthetic is minimalist and sophisticated.".
Task 23: Pose-guided Generation
Compose an artistic work driven by the edge image [SOURCE] in alignment with the description "A digital painting features a young woman with long, flowing auburn hair and round, wire-rimmed glasses. She wears a dark, hooded jacket over a striped, collared shirt, reminiscent of a school uniform, with a small emblem near the shoulder. Her gaze is directed slightly to the right of the frame, and her expression is thoughtful. Warm, ambient light illuminates her hair and creates a soft glow around her face. The background is blurred, suggesting a dimly lit outdoor setting.".
Task 24: Edge-guided Generation
Develop an artistic creation based on the depth map from [SOURCE] in accordance with "A vibrant green leather sofa sits in a modern room with large floor-to-ceiling windows showcasing a mountainous sunset view. A patterned rug lies beneath the sofa and a low, circular coffee table holds a few decorative items and glasses. A potted plant stands near the sofa, and a lamp and other objects rest on a console table against the stone-clad wall. The room's ambiance is warm and inviting, blending natural elements with contemporary design. The sunset casts a golden glow across the scene.".
Task 25: Depth-guided Generation
Add vibrancy to the grayscale photo [SOURCE] in line with "A dimly lit, modern bedroom features a large window showcasing a lush, green landscape with butterflies flitting about. A comfortable bed with rumpled bedding occupies the foreground, with an open book resting atop. Leafy green plants grow indoors, framing the bed and window area. A flat-screen television is mounted on the right wall, above a long, low entertainment center with soft, ambient lighting. The overall scene evokes a tranquil, nature-inspired atmosphere.".
Task 26: Image Colorization
Remove the blur from [SOURCE] using the guidelines provided in "A vibrant lantana plant, bursting with small orange and yellow flowers, sits in a cobalt blue spherical pot. The plant's green leaves contrast beautifully with the brightly colored blossoms. The pot rests on a weathered wooden deck or planks, creating a rustic backdrop. The close-up shot emphasizes the plant's texture and color. The overall image conveys a sense of cheerful, summery ambiance.".
Task 27: Image Deblurring
do style conversion for [SOURCE] following the style of [REF_1].
Task 28: Style Transfer
Alter the mask in [SOURCE] based on the subject in [REF_1].
Task 29: Subject-guided Inpainting
Make this clothes in [REF_1] be worn on the person in [SOURCE] at mask
Task 30: Virtual Try On
Update the face in [SOURCE] mask using the facial characteristics from [REF_1].
Task 31: Face Swap

Figure 7. Examples of 31 fine-grained evaluation tasks in our ICE-Bench.