Towards Enhanced Image Inpainting: Mitigating Unwanted Object Insertion and Preserving Color Consistency

Supplementary Material

6. Brief Introduction of Backbone Models

We evaluate our proposed solution on two inpainting models: the Stable Diffusion v1.5 inpainting model (SD) [67] and the Control-Net fine-tuned FLUX inpainting model (FLUX) [2]. Both models are representative latent inpainting models that use a VAE [38] to compress images into a smaller latent space. In SD, a diffusion process [72] maps the latent space to random Gaussian noise, and a U-Net [68] learns the reverse denoising path. Text condition is introduced through cross-attention layers [77]. The inpainting version of SD extends the U-Net input by concatenating the masked image and mask with the noise along the channel dimension. Conversely, FLUX uses rectified flow [1, 49, 52] to map the latent space to noise and a vision transformer [63] for generation. Text condition is applied by concatenating text with image patches as transformer input, while a pooled text condition is injected into the normalization layers. Since the original FLUX [40] does not support inpainting, we use a Control-Net [96] fine-tuned version [2] that modifies FLUX's transformer output by conditioning on the masked image and mask. We demonstrate that our ASUKA effectively improves unwanted object mitigation and color consistency of these models.

7. Details about MISATO

The principle of constructing MISATO is to select the most representative and diverse examples. To this end, for first three datasets, we use CLIP visual model [64] to extract semantic visual features. Then we use BisectingKMeans [74] to cluster each dataset into 500 clusters, and select the cluster centers as the evaluation data. The selected data are center cropped and then resized to 512^2. For COCO, we focus on the background inpainting. To this end, for each data we identify the foreground with provided segmentation and remove it from the generated masks, yielding a dataset specified for purely background inpainting.

Combined together, MISATO contains 2000 examples from four inpainting domains, indoor, outdoor landscape, building, and background, as shown in Fig. 10. we adopt the masking strategy as in Sec. 3.1.1 but excluding the rectangle and complement rectangle masks. The masking ratio is set as $[0.2, 0.8]$.

8. Implementation Details

We use Places2 [101] to train ASUKA. For the MAE [31] used in ASUKA, we train on images of size 256^2, which is efficient and produce context-stable guidance for generative models to generate high-resolution images. We fine-tune the MAE with a batch size of 1024. We train the alignment module with AdamW [53] of learning rate 5e-2 with the standard diffusion objective. We set p as 100% and linearly decay it to 10% in the first 2K training steps and then freeze. For SD's decoder, we fine-tune from [103] for 50K steps with a batch size of 40 and learning rate of 8e-5 with cosine decay. For FLUX's decoder, we fine-tune from the original decoder with the same setup. We use ColorJitter for color augmentation, with brightness 0.15, contrast 0.2, saturation 0.1, and hue 0.03.

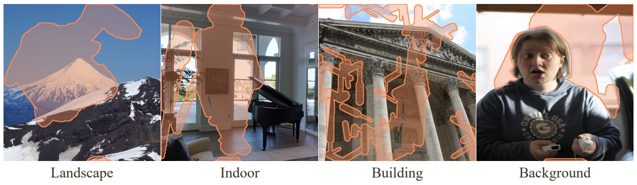

Figure 10. Different image domains in MISATO.

Table 5. Comparison of ASUKA with text-guided SD

Model	LPIPS↓	FID↓	U-IDS↑	P-IDS↑	C@m↑	G@e↓
SD (BLIP2)	0.163	12.536	0.370	0.225	0.880	70.846
ASUKA-SD	**0.150**	**11.495**	**0.423**	**0.312**	**0.958**	**47.753**

Table 6. Ablation of p

Model	LPIPS↓	FID↓	U-IDS↑	P-IDS↑	C@m↑	G@e↓
p=0	0.155	11.804	0.403	0.288	0.940	48.032
p=1	0.152	11.734	0.394	0.296	0.947	47.997
linear decay p	0.152	11.558	0.405	0.307	0.955	47.814
Ours	**0.150**	**11.495**	**0.423**	**0.312**	**0.958**	**47.753**

9. Further Analysis

Comparison with text-guided inpainting We compare ASUKA with text-guided SD model, as shown in Tab. 5. We run SD inpainting sing text captions generated by BLIP2 [43]. ASUKA performs better, since captions describe the entire image, while MAE focuses on reconstructing only the masked region, leading to more precise guidance.

Ablation of p We analyze how different values of p affect

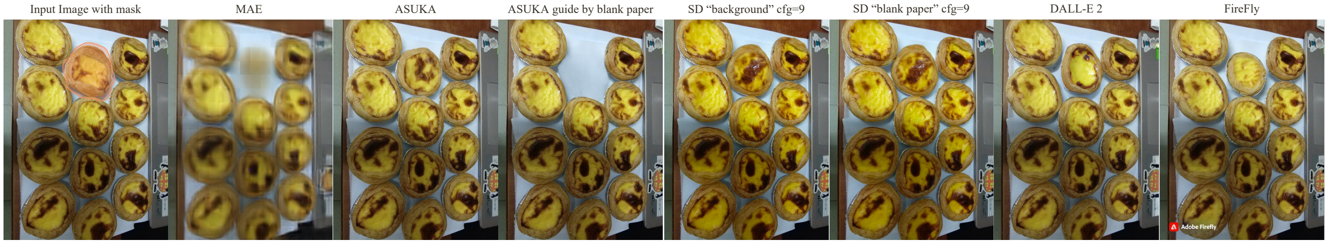

Figure 11. The curse of self-attention, causing the MAE falsely estimate the masked region and powerful text-guided diffusion models fail to generation content based on text prompts. ASUKA potential circumvents this issue by using a blank paper image as the input to the MAE to provide correct prior.

Table 7. Additional results on benchmark datasets

Dataset	Model	LPIPS↓	FID↓	U-IDS↑	P-IDS↑	C@m↑	G@e↓
CelebA-HQ	SD	0.132	11.968	0.282	0.101	0.939	42.870
	ASUKA-SD	**0.129**	**10.190**	**0.293**	**0.134**	**0.941**	**40.503**
FFHQ	SD	0.139	2.235	0.371	0.197	0.944	43.529
	ASUKA-SD	**0.131**	**2.060**	**0.386**	**0.205**	**0.955**	**30.848**

Table 8. Our Decoder in Text-Guided Inpainting.

Model	CLIPScore↑	LPIPS↓	FID↓	U-IDS↑	C@m↑	G@e↓
SD	0.297	0.180	30.255	0.312	0.930	57.136
ASUKA-SD	**0.298**	**0.175**	**29.350**	**0.350**	**0.931**	**38.123**

Table 9. Effect of each module.

MAE	LPIPS↓	FID↓	U-IDS↑	P-IDS↑	C@m↑	G@e↓
SD w/ MAE	0.157	12.093	0.397	0.236	0.953	62.845
SD w/ decoder	0.159	12.075	0.411	0.283	0.954	49.376
ASUKA-SD	**0.150**	**11.495**	**0.423**	**0.312**	**0.958**	**47.753**

Table 10. Comparison of ASUKA using pre-trained MAE v.s. fine-tuned MAE.

MAE	LPIPS↓	FID↓	U-IDS↑	P-IDS↑
pre-trained	0.151	11.513	0.354	**0.258**
fine-tuned	**0.150**	**11.460**	**0.368**	0.256

Table 11. User-study of top-1 ratio among all the inpainting results.

Model	UOM (%)	CC(%)
Co-Mod [99]	3.98	4.98
MAT [44]	7.40	3.20
LaMa [75]	8.18	8.28
MAE-FAR [10]	4.88	5.60
SD [67]	10.58	5.75
SD-text	7.70	15.83
SD-prompt	16.18	15.78
SD-Repaint [54]	1.60	0.55
ASUKA-SD	**39.43**	**40.05**

ASUKA in Tab. 6. The results show that our warm-up and freeze strategy outperforms other approaches.

Additional Results We further compare ASUKA with standard SD on two additional datasets: CelebA-HQ [36] and FFHQ [37]. As shown in Tab. 7, these results provide more evidence of ASUKA's effectiveness.

Our Decoder in Text-Guided Inpainting To test the generalizability of our decoder, we evaluate it on text-guided inpainting tasks. We compare our decoder with the original SD decoder using 1,000 randomly sampled images from “jackyhate/text-to-image-2M” [105]. The results in Tab. 8 confirm its effectiveness for general inpainting tasks.

Ablation on independent modules To understand the contribution of each module in ASUKA, we evaluate SD with the proposed modules added separately. The results, shown in Tab. 9, highlight the effectiveness of each module.

Ablation of MAE prior We compare our fine-tuned MAE with directly adopting the MAE trained in [10]. To this end, we train ASUKA with the MAE in [10] using the same training strategy and compare the results in Tab. 10. Results suggest the improvements of fine-tuning MAE, especially on FID and U-IDS. This improvement comes from the better adaptation on the real-world masks.

User-study To evaluate the user preference on inpainting algorithms, we conduct an user-study. Specifically, we randomly select 40 testing images. We ask the user to select the best inpainting results from the following perspectives respectively: i) Unwanted-object-mitigation (UOM): the generated region should be context-stable with surrounding unmasked region, with a preference of not generating new elements; ii) Color-consistency (CC) : the color consistency between masked and unmasked regions. We collect 100 valid anonymous questionnaire results, and report the average selection ratio among all the inpainting algorithms in Tab. 11. This result validate the efficacy of ASUKA on alignment with human preference.

Limitation: The ”curse” of self-attention The primary limitation of ASUKA stems from the inefficacy of the MAE prior, mainly due to issues within the self-attention module. Specifically, as shown in Fig. 11, the presence of multiple similar objects in an image may lead the MAE to incorrectly predict a similar object in the masked region, conflicting with the goal of object removal. Notably, this curse of self-attention significantly impacts diffusion-based generative models. It results in the inability to accurately follow ”blank paper” text prompts, even when employing a substantial classifier-free guidance scale of 9. This issue is not unique to SD but is also a common problem in other advanced text-guided diffusion models, such as OpenAI’s DALL-E 2 [65] and Adobe’s FileFly [18]. Nevertheless, ASUKA has the potential to circumvent this issue by modifying the MAE prior, for instance, by instead using a blank paper image as the input to MAE prior. A more comprehensive solution would involve extra control on self-attention layers in diffusion models, which we leave as future work.

Potential negative impact As an image editing tool, our proposed ASUKA will generate images based on user intentions for masking specific parts of the image, potentially resulting in unrealistic renderings and posing a risk of misuse.

10. More Qualitative Examples

Here we provide more qualitative examples on MISATO in Fig. 12, Fig. 13, Fig. 14, Fig. 15, and Fig. 16. We compare ASUKA with Co-Mod [99], MAT [44], LaMa [75], MAE-FAR [10], and SD [67]. SD performs unconditional generation. SD-text utilizes text prompt of “background”. SD-token utilizes trained prompt for inpainting task using the same training setting of ASUKA.

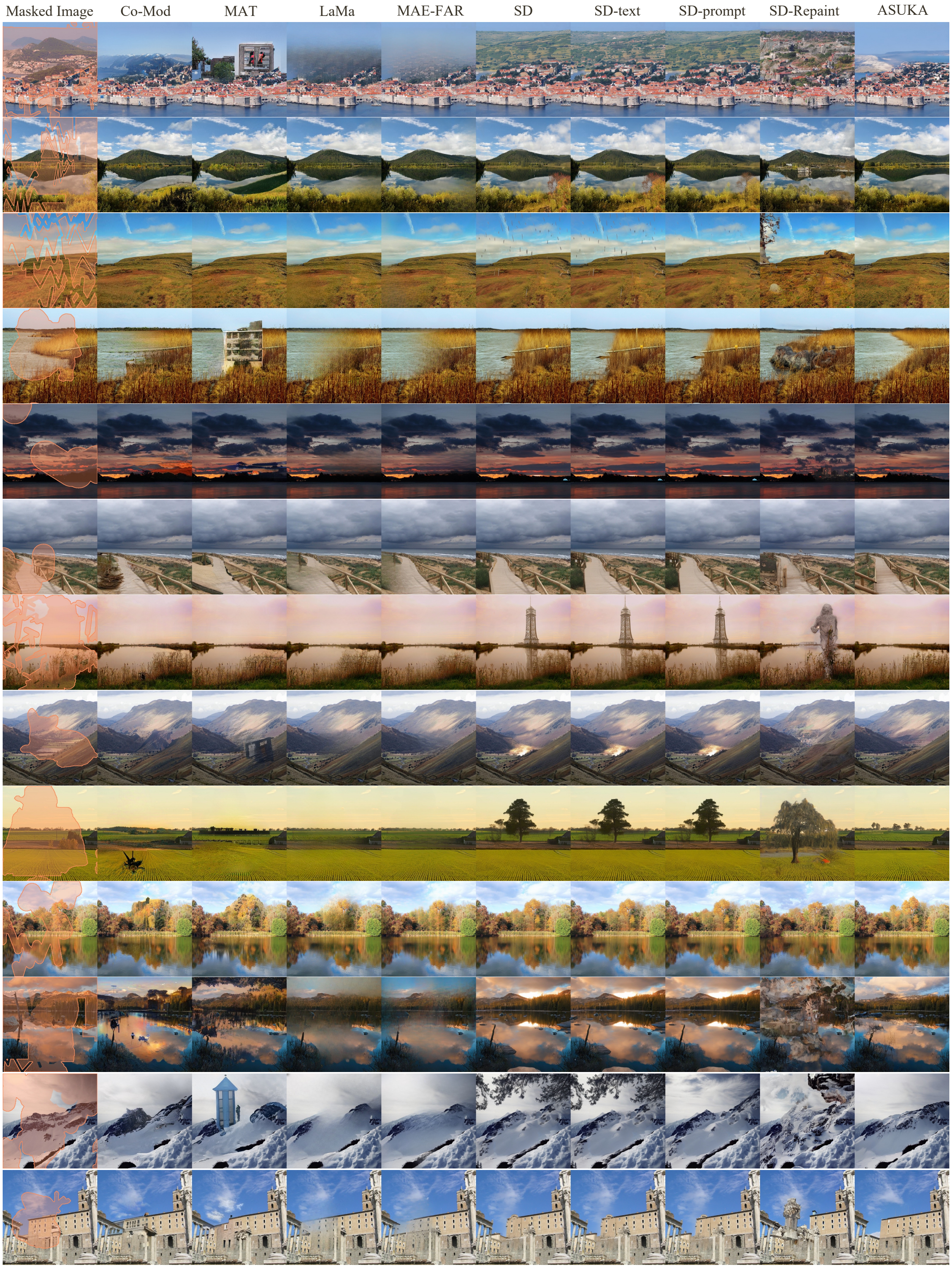

Figure 12. More qualitative comparison on MISATO.

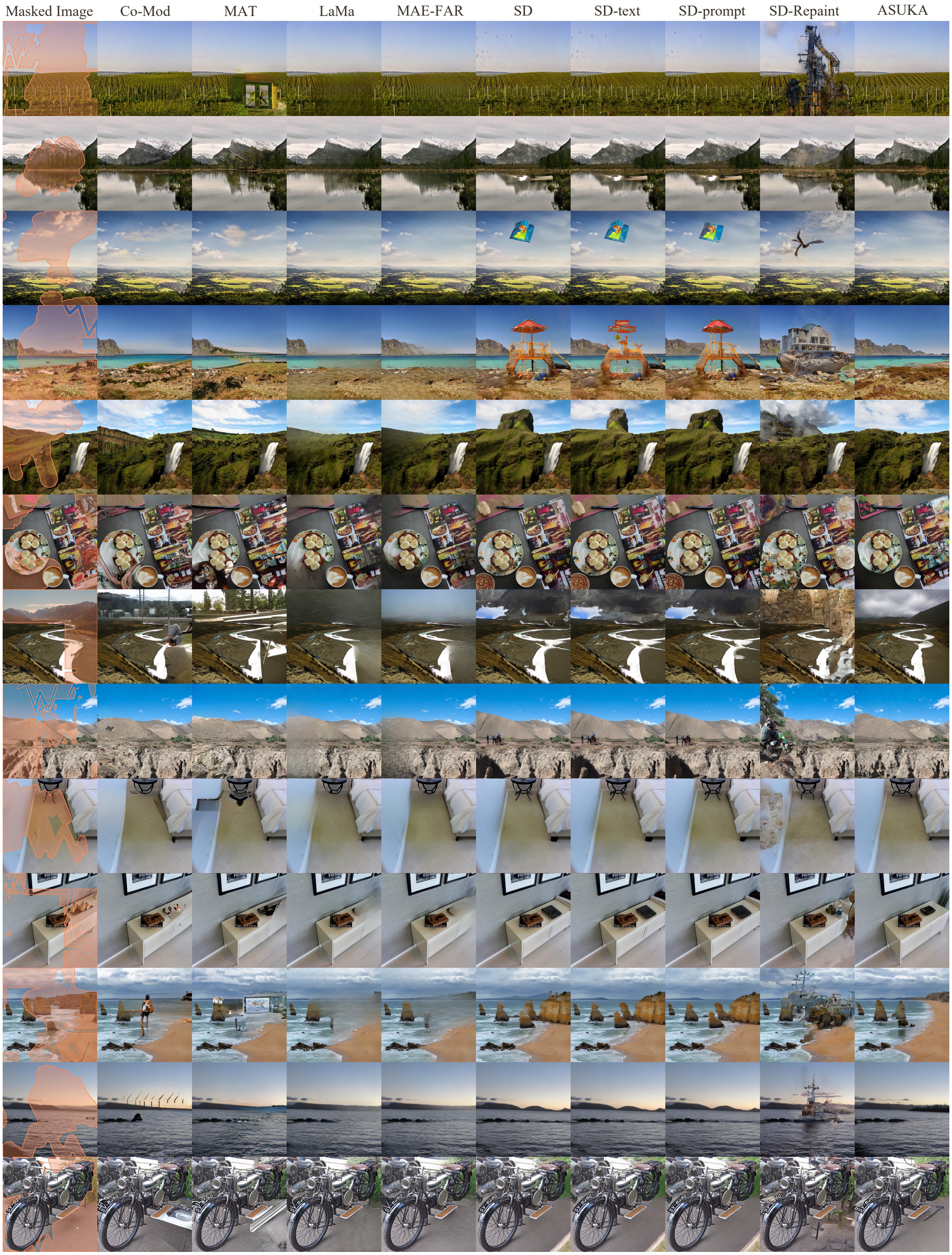

Figure 13. More qualitative comparison on MISATO.

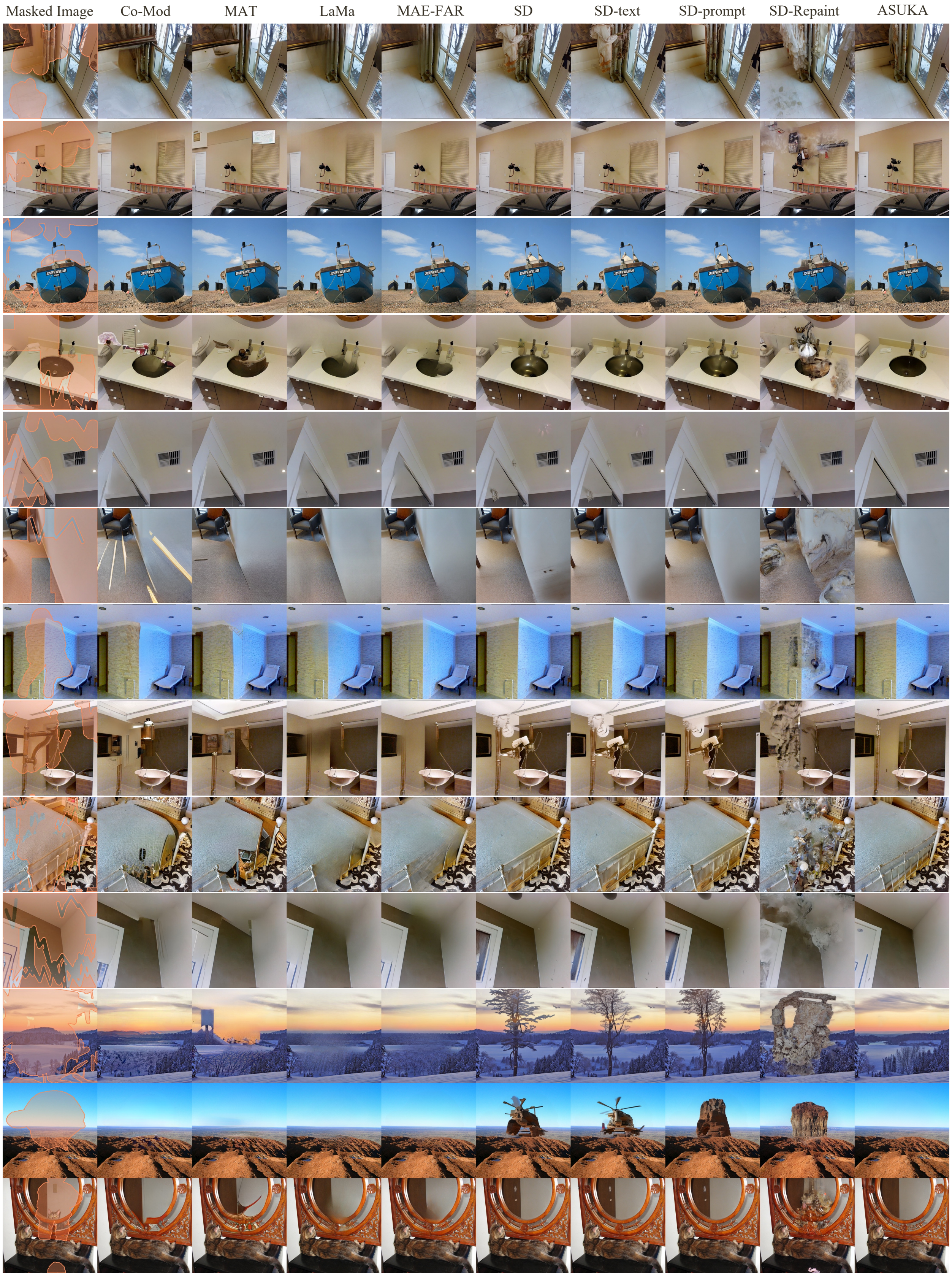

Figure 14. More qualitative comparison on MISATO.

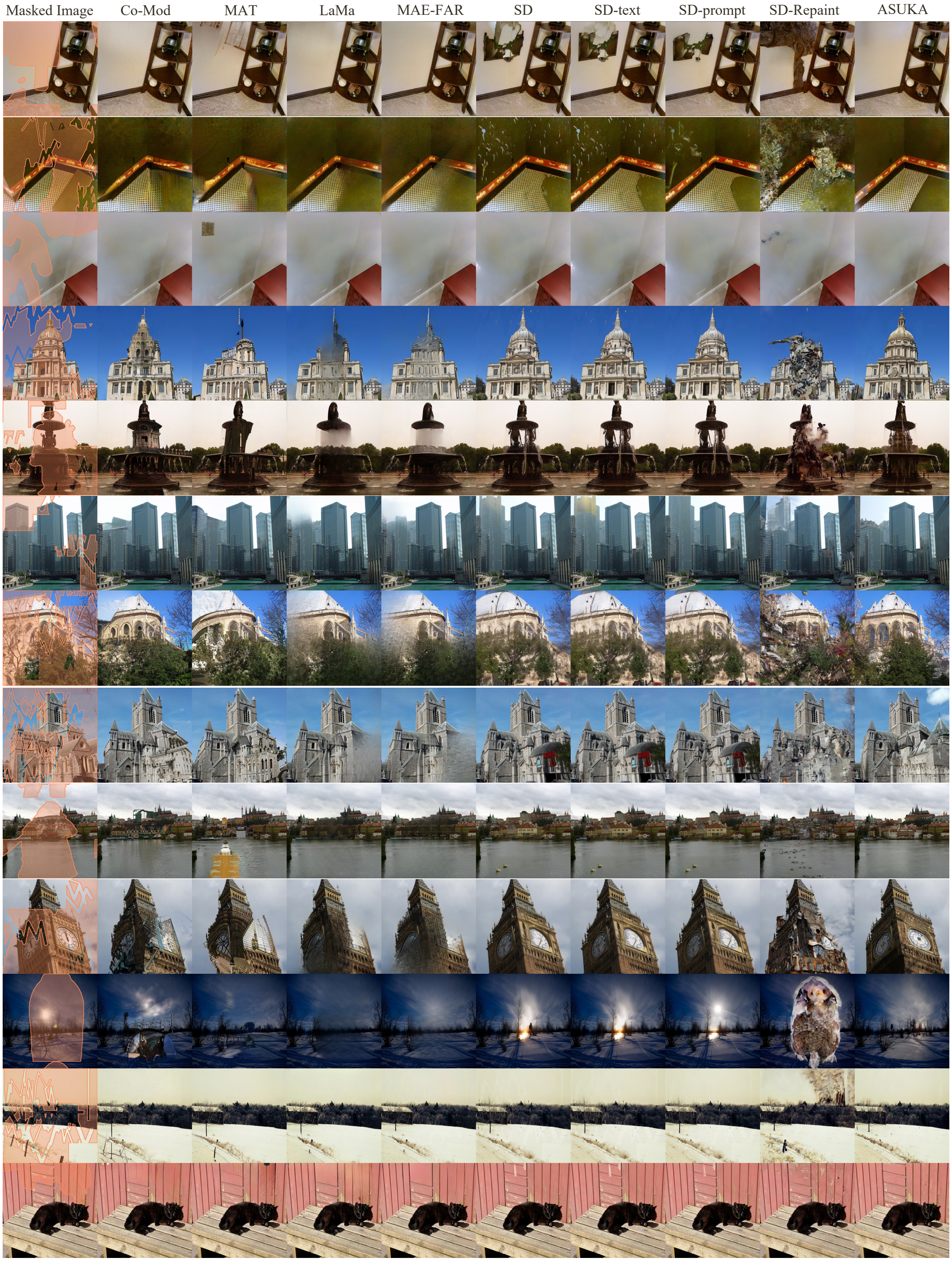

Figure 15. More qualitative comparison on MISATO.

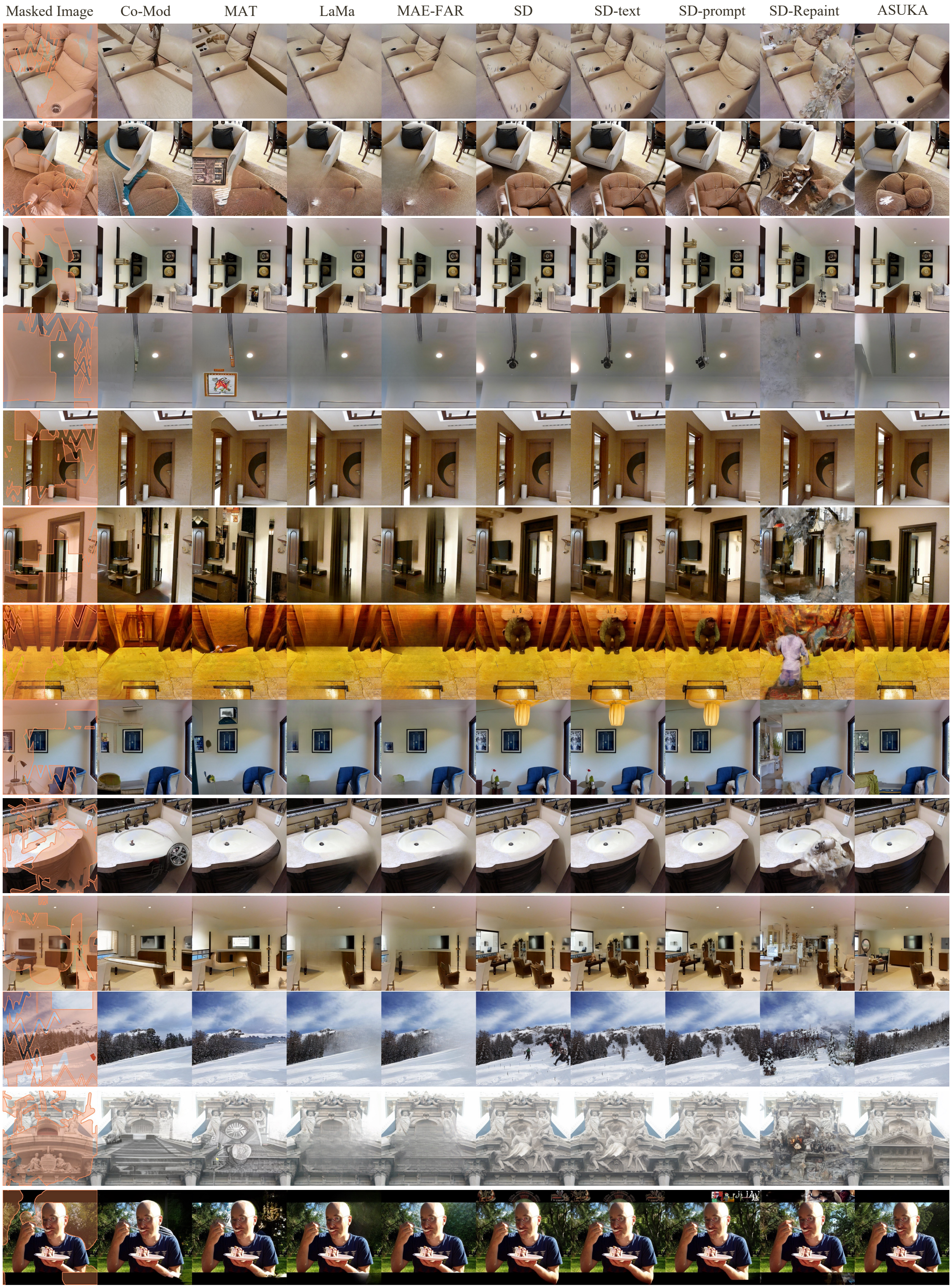

Figure 16. More qualitative comparison on MISATO.