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

Yikai Wang[1,2]*, Chenjie Cao[1,3,4]*, Junqiu Yu[1]*, Ke Fan[1], Xiangyang Xue[1], Yanwei Fu[1]
[1]Fudan University [2]Nanyang Technological University [3]Alibaba DAMO Academy [4]Hupan Lab
yi-kai.wang@outlook.com, yanweifu@fudan.edu.cn
Project page (include code, model, and dataset): https://yikai-wang.github.io/asuka

## Abstract

*Recent advances in image inpainting increasingly use generative models to handle large irregular masks. However, these models can create unrealistic inpainted images due to two main issues: (1)* ***Unwanted object insertion****: Even with unmasked areas as context, generative models may still generate arbitrary objects in the masked region that don't align with the rest of the image. (2)* ***Color inconsistency****: Inpainted regions often have color shifts that causes a smeared appearance, reducing image quality. Retraining the generative model could help solve these issues, but it's costly since state-of-the-art latent-based diffusion and rectified flow models require a three-stage training process: training a VAE, training a generative U-Net or transformer, and fine-tuning for inpainting. Instead, this paper proposes a post-processing approach, dubbed as ASUKA (Aligned Stable inpainting with UnKnown Areas prior), to improve inpainting models. To address unwanted object insertion, we leverage a Masked Auto-Encoder (MAE) for reconstruction-based priors. This mitigates object hallucination while maintaining the model's generation capabilities. To address color inconsistency, we propose a specialized VAE decoder that treats latent-to-image decoding as a local harmonization task, significantly reducing color shifts for color-consistent inpainting. We validate ASUKA on SD 1.5 and FLUX inpainting variants with Places2 and MISATO, our proposed diverse collection of datasets. Results show that ASUKA mitigates object hallucination and improves color consistency over standard diffusion and rectified flow models and other inpainting methods.*

## 1. Introduction

Image inpainting [6] fills masked areas of images while maintaining consistency with the unmasked regions. Traditional inpainting algorithms [6, 21, 30, 41, 69] often result in blurred synthesis when reconstructing masked regions [62]. The Generative Adversarial Networks (GANs) based models could fill complex mask structures, achieving impressive inpainting results [9, 24, 27, 33, 46, 59, 75, 78, 99]. However, they still struggle with general challenging inpainting cases, particularly in filling large holes. Recently, more powerful generative model like Stable Diffusion [67] and FLUX [40], with their large model capacity and extensive training dataset, act as versatile tools for image inpainting. These models mainly follow the latent generation pipeline, first encode the image into a small latent space, then train the inpainting model in this latent space.

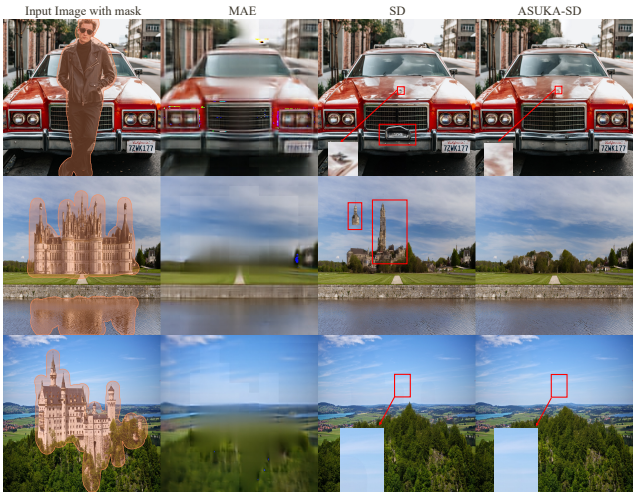


Figure 1. Image inpainting on $1024^2$ images. ASUKA solves two issues existed in current diffusion and rectified flow inpainting models: (1) Unwanted object insertion, where randomly elements that are not aligned with the unmasked region are generated; (2) Color-inconsistency: the color shift of the generated masked region, causing smear-like traces. ASUKA proposes a post-training procedure for these models, significantly mitigates object hallucination and improves color consistency of inpainted results.

However, these latent-based generative inpainting models still suffer from some issues, causing the inpainted im-

First three authors contribute equally. Most parts of this work was done when Yikai was at Fudan. Yanwei Fu is the corresponding author.

age lacking fidelity. In particular:

(1) The *unwanted object insertion* problem, where the model generates random, unreasonable elements to fill masked regions, as depicted in first to second rows in Fig. 1. This issue comes from the random masking strategy used to train generative models. This strategy introduces training cases where foreground objects are completely masked but the models are forced to fill masked regions with foreground objects. Consequently, these models will hallucinate unreasonable objects devoid of contextual information. Adjusting prompts may reduce this risk, but the best prompt is image-dependent, making it infeasible for practical applications.

(2) Furthermore, the inpainted results of latent inpainting models suffer from *color inconsistency* problem. This problem, less explored in academia but critical for real-world applications, results in color discrepancies between inpainted and unmasked regions, including mismatches of brightness, saturation, luminance and hue, and exhibits smear-like traces in the image, as shown in the second to third rows in Fig. 1. Essentially, this color-inconsistency comes from the misalignment between the pixel distributions of filled results and original images due to imperfect latent generative model and VAE, as illustrated in Fig. 4. Notably, this issue is not a big problem for generation, given that the whole image is generated, and the color shift is consistent across pixels. However, this issue is important for inpainting tasks, as we have ground-truth pixels for unmasked regions. When we replace the unmasked regions of the generated image with the ground-truth pixels, the color inconsistency largely influence the fidelity of the image. This issue may be solved by training a better VAE and explicitly enforce the color consistency. However, training or fine-tuning the VAE encoder introduces the subsequent fine-tuning of the latent generative models to match the new latent space, which is costly. In this paper, we propose to freeze the VAE encoder and the latent generative models, while fine-tuning the VAE decoder to improve color consistency. Specifically, we reformulate the decoding from latent to image as a local harmonization task, explicitly reduce the color inconsistency.

Formally, to mitigate object hallucination and enhance the color-consistency of image inpainting models, we present the Aligned Stable inpainting with UnKnown Areas prior (ASUKA) framework. ASUKA enhances the latent inpainting models with regression-based reconstruction and distribution-aligned generation. This results in improved image inpainting models that avoid generating unreasonable elements in the masked region and reduces mask-unmask color inconsistency. The stable diffusion models [67] and the rectified flow models [40] adopt a VAE to compress image into latent and perform inpainting in the latent space. We manipulate their generation and decoding processes to reduce object hallucination and improve color consistency.

We propose using the Masked Auto-Encoder (MAE)[31] as a prior to guide and stabilize the generation process. As shown in Fig. 1, MAE yields stable yet blurred results, while generative models may produce implausible content despite their impressive generation capacity. By aligning MAE prior with latent generative models, we reduce object hallucination without damaging performance.

We redesign the VAE decoder to address color inconsistencies between masked and unmasked regions by acting as a local harmonization model conditioned on unmasked image pixels. *Our decoder can be used as a plug-and-play module to improve general inpainting models, such as text-guided inpainting*.

These steps collectively enable ASUKA to achieve less object hallucination and more color-consistent inpainting results. We adopt ASUKA on two typical inpainting models, Stable Diffusion v1.5 [67] and FLUX [40], to validate the generalization ability of ASUKA on different generation architectures. To evaluate the effectiveness of inpainting algorithms across various scenarios and mask shapes, in addition to the benchmark dataset Places 2 [101], we further utilize an evaluation dataset named MISATO, which selects representative testing images from Matterport3D [13], Flickr-Landscape [47], MegaDepth [45], and COCO 2014 [48]. This dataset covers four distinct domains—landscape, indoor, building, and background—making it diverse to serve as a benchmark for evaluation. Experiments on MISATO and Places 2 with large irregular masks validate the efficacy of ASUKA.

**Contributions** ASUKA enhances image inpainting with color-consistency and mitigate object hallucination while leveraging the generation capacity of the frozen inpainting model. It achieves this through two main components: *(1) Context-Stable Alignment*: ASUKA aligns the stable MAE prior with generative models to provide a context-stable estimation of masked regions, replacing the text-condition with MAE prior. *(2) Color-Consistent Alignment*: ASUKA re-formulates the decoding from latent to image as a local harmonization task, trains an inpainting-specialized decoder to align masked and unmasked regions during decoding and thus mitigates color inconsistencies.

## 2. Related Works

**Image inpainting** is the task of filling missing image regions with consistent pixels. Traditional methods using patch matching [5, 22, 95] or differential equations [6, 7, 12] focus on low-level features and often struggle with large gaps. GAN [27]-based inpainting [10, 44, 62, 91, 99] introduces adaptive convolutions [50, 91, 93], attention [39, 89, 90, 92], and frequency-based learning for high-resolution results [17, 75, 87]. Methods like Co-Mod [99] address the challenging ill-posed inpainting issue [44, 100] and improve realism but may produce unstable outputs or unwanted artifacts due to random latent variables. Techniques

with higher reconstruction penalties [10, 59, 75] offer more stability but can appear blurry on larger missing areas. Recent diffusion models [3, 67, 70] and rectified flow models [25, 40] achieve impressive results yet share GANs' limitation of learning distributions over exact pixel alignment, which leads to unwanted object insertion.

**Adapting latent generative models** Latent diffusion models (LDMs) [67] and rectified flow models [25, 40] are popular due to their ability to encode image semantics at lower resolutions by combining a VAE to learn a latent space and a generative model within this space. Various methods have been developed to introduce new conditions to these models, such as image-inversion for text-guided image translation [56], textual-inversion for personalization [26], LoRA fine-tuning [34], and controlnet [96] to add diverse conditions. Our goal is to mitgate object hallucination while preserving generation quality, so we avoid fine-tuning the generative backbone. For inpainting, we remove the text condition and instead guide the generation using a Masked Auto-Encoder [31] prior for masked regions.

**Information loss in latent inpainting models** Although claimed only eliminates imperceptible details, the VAE used by diffusion and rectified flow models causes distortion in the reconstruction of images. In addition, the gap between generated latent and real latent also causes the color inconsistency. See Fig. 4 for illustrative examples. OpenAI [61] proposes a larger decoder to improve the decoding quality of SD's latent. Luo *et. al.* [55] propose a frequency-augmented decoder to address the super-resolution case. Zhu *et. al.* [103] propose to preserve unmasked regions during decoding. In this paper, we ensure the low-frequency color color-consistency in the decoding process.

**Masked Image-Modeling [4]** (MIM) is an active research area in self-supervised learning. Typical MIM methods [4, 14, 31, 86] split images into visible and masked patches, learning to estimate masked patches from visible patches. Training targets for visible patches encompass pixel values [31], HOG features [82], and high-level semantic features [83]. While the primary objective of MIM is representation learning, its potential effectiveness in image generation is also noteworthy. Cao *et. al.* [10] adopts MAE features and attention scores to assist the convolutional inpainting model better in handling long-distance dependencies. In contrast, this paper uses MAE prior to enhance the context-stability of diffusion and rectified flow models.

**Image harmonization** aims to blend a foreground object with a background image while keeping the final result realistic and visually consistent [76]. This task is often treated as an image translation problem [19, 20, 28, 29, 51, 57, 60, 66, 79, 102]. Similarly, our work addresses color inconsistency issues in latent generative models. However, unlike standard image harmonization, where inconsistencies arise from combining images from different sources and thus different real image distributions, color inconsistencies in latent generative models stem from imperfections in the VAE and the generative model itself.

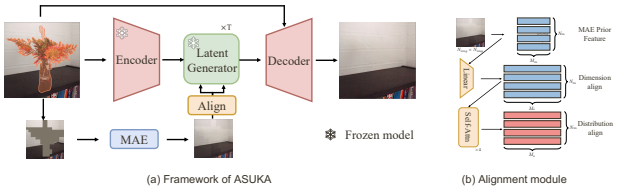


Figure 2. ASUKA tackles the unwanted object insertion issue by adopting the MAE to provide a stable prior for frozen latent generative models to maintain the generation capacity while mitigating object hallucination. For the color-inconsistency issue, ASUKA utilizes an inpainting-specialized decoder to achieve mask-unmask color consistency when decoding latent.

**Object insertion and removal** are two opposite tasks in image inpainting. Object insertion focuses on adding foreground objects to the image using various methods, such as shape-guided masks [85, 94], text prompts [8, 16, 80, 85], learnable prompts [16, 81, 104], extra network modules [15, 35], or reference images of objects [71], etc. Some studies also explore completing partial objects using reference images [11] or learnable prompts [81]. Object removal, on the other hand, aims to erase unwanted objects from an image. Common approaches include attention reweighting [42] and learnable prompts [81, 104]. These techniques can help create new datasets [23]. On the other hand, creating new datasets can also benefit these tasks [84]. While most research focuses on designing better inpainting models, our work takes a different approach. We analyze a fundamental problem with latent generative models: they often introduce unwanted objects in the inpainting area. We also propose solutions to address this issue.

## 3. Methodology

**Problem setup** Inpainting takes as inputs a masked image to complete with a mask to indicate the missing region. The target of inpainting is to fill the missing region based on the information of unmasked regions to generate high-fidelity images. In this paper, we focus on the standard inpainting task without utilizing other conditions. We focus on the general issues of inpainting models, (1) **unwanted object insertion**: unstable and uncontrollable hallucinations, yielding random elements generated in the masked region; (2) **color-inconsistency**: mask-unmask color inconsistency issue, yielding smear-like traces in the masked region.

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]. We provide a brief introduction of these models in the appendix. We will demonstrate that our ASUKA effectively improves unwanted object mitigation

and color consistency of these models.

**Overview** The framework of the proposed Aligned Stable inpainting with Unknown Areas prior (ASUKA) is illustrated in Fig. 2(a). ASUKA adopts the pre-trained latent inpainting models. Our target is to mitigate object hallucination and provide more color-consistent inpainting results while fully exploiting the generation capacity of frozen models. ASUKA includes **(1)** a *context-stable alignment* to align stable Masked Auto-Encoder (MAE) prior for masked region with generative models and **(2)** a *color-consistent alignment* to align ground-truth unmasked region with generated masked region during decoding. To this end, we freeze the latent generative models, while replacing the text-condition part with our proposed MAE prior to mitigate object hallucination. To align the MAE prior to generative models, we introduce an alignment module, trained via the training objective of generative models. Additionally, to align masked and unmasked regions during decoding and resolve the information loss issue from VAE decoder and generative model which causes mask-unmask color inconsistency, we train an inpainting-specialized decoder to decode the latent back to the image space for seamless color-consistency. Combined together, ASUKA achieves less object hallucination and more color-consistent inpainting.

## 3.1. Mitigate Object Hallucination via Stable Prior

### 3.1.1. Masked Auto-Encoder Prior

**Context-stable prior** While recent generative models rely on random noise to provide more diverse generation results, it leads to the generation of random objects unexpectedly. Some inpainting models also utilize the reconstruction loss to reconstruct the masked region, but they also incorporate other types of losses like perceptual-loss [75] which implicitly reduces the stability. In contrast, MAE is known to provide a context-stable estimation of masked regions based purely on the unmasked regions. In this paper, we utilize MAE to produce the stable prior such that *the improvement of inpainted result can be explicitly attributed to the improvement of mitigating object hallucination*.

**MAE as context-stable prior** As MAE is trained on the L2 reconstruction loss, we can regard the estimation of MAE as a mean estimation, which can be utilized to provide a context-stable prior for generative models to not generate new concepts. However, MAE itself results in average and blur generations and cannot reconstruct detailed textures of the masked region, and works poorly if we use MAE prior as the initial values for the inpainting models to inpaint in image-to-image style, as in Fig. 3. To this end, we adopt the MAE to provide prior to stabilizing diffusion models.

**Train MAE** The original MAE is trained to estimate random masks uniformly distributed in the image, while inpainting task usually contains large continuous masks. Inspired by Cao *et. al.* [10], we fine-tune the MAE to inpainting masks. To adapt MAE for more practical inpainting scenarios, we construct a systematic masking strategy. The mask basis contains: object-shape mask, irregular mask, and regular mask. We collect object-shape masks from COCO [48] object segments. We use irregular masks from previous studies, including Co-Mod mask [99] and LaMa [75] mask. The regular masks contains rectangle and complement rectangle mask. To ensure generalization and coverage, for each mask we generate from mask basis with the probability of 50% object-shape, 40% irregular, and 10% regular. For object-shape mask basis, we combine it with irregular mask with the chance of 50%. This construction of mask style estimates the masks occurs in inpainting tasks, especially for the object removal and user-specified irregular masks. We control the mask ratio in the range of $[0.1, 0.75]$ to follow the training scenario of MAE. For masks smaller than the ratio of 75%, we enlarge the mask ratio to 75% with randomly selected mask regions. This benefits ASUKA to tackle the large hole inpainting task.


Figure 3. Use MAE prior for image-to-image translation (start from 80% noise rate) via SD achieves poor inpainting results.

### 3.1.2. Align MAE Prior with Generator

**Replace text-condition with MAE prior** Generative inpainting models are not trained on MAE priors. As we do not assume a text condition for inpainting task, we propose to replace the text-condition of generative models with our proposed MAE prior condition. However, as we do not fine-tune the generative models, they cannot directly align well with the MAE prior. Hence, we introduce the alignment module to align MAE with generative models in both dimension and distribution perspective, as shown in Fig. 2(b).

**Dimension alignment** Particularly, the MAE prior $F_{\text{MAE}}$ is of size $N_m \times M_m$, where $N_m$ is the sequence length and $M_m$ is the feature dimension. To align it with the diffusion or flow condition of size $N_s \times M_s$, we adopt a linear layer to map the feature dimension from $M_m$ to $M_s$ and set $N_s = N_m$ to preserve the local MAE prior.

**Distribution alignment** After aligning the dimension, we use self-attention blocks to learn to better guiding generative models, leading to the condition $C_{\text{MAE}}$. We train our alignment module using the standard generation objective with the same masking strategy used to train the MAE, keeping other modules frozen.

**Handle misalignment** When training the alignment module with the set (input image, MAE prior, inpaint result), misalignment may arise. For example, if an object is completely masked, the MAE will predict the masked area with

background, whereas the generative models are trained to recreate the object. This difference can lead the alignment module to mistakenly disregard the MAE prior. To address this, we improve the generative models' adherence to the MAE prior by substituting the MAE predicted prior with the MAE reconstructed prior at a probability of $p$. The MAE reconstructed prior involves using MAE to recreate the image without masking any area, ensuring MAE has access to all information needed for reconstruction. This approach helps train the alignment module to better guidance.

## 3.2. Enhancing Color-Consistency in Decoding

### 3.2.1. Color-Inconsistency

**Color-inconsistency is a general problem** The color-inconsistency between masked and unmasked regions is a general problem in generative inpainting models. This inconsistency comes when the generative masked region suffers from a color shift compared with the unmasked region. As in Fig. 4, the color shift happens in all kinds of scenarios, including indoor and outdoor scenes, random or continuous masks, and may cause darker or lighter color shift. This shift comes from the imperfect VAE and latent generator.

**Information loss of VAE** Popular latent diffusion and rectified flow models perform all the generative processes in the latent space and subsequently decodes these latent codes back to image space using VAE. Despite the decoder being trained to reconstruct the image, it encounters challenges associated with information loss. Particularly in tasks like inpainting, we have ground-truth values for the unmasked region. Though Rombach *et. al.* [67] claimed that the diffusion model should prioritize the informative semantic compression, while the VAE is used to tackle perceptual compression with high-frequency details, we argue that *low-frequency semantic loss in VAE could not be neglected*, as verified in Fig. 6 (b). The VAE will not only noticeably degrades high-frequency details but also shifts in colors. This shift can be verified by repeated reconstruction with VAE, as shown in Fig. 6 (a) where larger shift is observed during repeated reconstruction. As human is sensitive to low-frequency information changes in the image, even subtle color shifts can induce significant inconsistencies. This issue is more severe in irregular or large mask cases.

**Gap between real and generated latents** Apart from the information loss of VAE in reconstruction, there is another gap between the generated and real latents. This gap also causes color inconsistency even if we alleviate the VAE reconstruction loss, as in Fig. 5. We need to solve both the loss of VAE and the latent generator for better color-consistency.

### 3.2.2. Mask-Unmask Align during Decoding

We propose to solve the color-inconsistency and ensure the mask-unmask alignment during VAE decoding.

**Unmask-region conditioned decoder** The basic solution

Figure 4. The color shift exists in all kinds of scenarios in inpainted images, including indoor and outdoor scenes, random or continuous masks, and may cause darker or lighter color shift.


Figure 5. Inpainting w/ v.s. w/o latent augmentation. The latent augmentation handles the gap between generated and real latent.

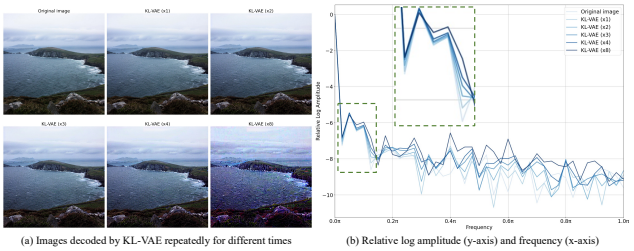


Figure 6. (a) The color of the reconstruted image is shifted, where larger shift is observed during repeated reconstruction. (b) VAE suffers from non-ignorable shifts in low-frequency fields.

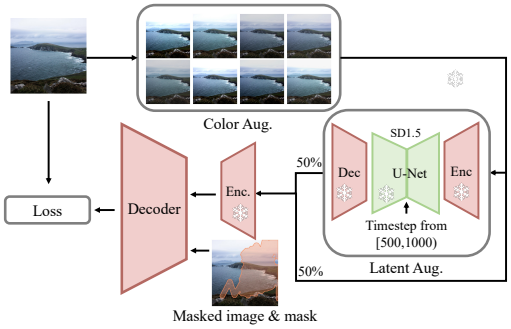


Figure 7. Decoder trained by local harmonization task, enhancing mask-unmask consistency by reconstructing original image guided by the unmasked region from augments in color and latent spaces.

is to incorporate the ground-truth unmasked region in the decoding, then we could have access to the unbiased color information. Zhu *et. al.* [103] adopts decoder with additional inputs of masked images. However, it still fails to handle the incompatible color and texture between the original images and compressed ones in challenging scenes as

(a) Masked image (b) Vanilla decoder (c) Conditional decoder (d) ASUKA decoder

Figure 8. SD1.5 inpainting results decoded by (b) vanilla decoder of SD [67], (c) conditional decoder [103], (d) our decoder. Our decoder largely alleviate the mask-unmask color inconsistency.

verified in Fig. 8 (c). The gap between degraded and original images makes it challenging to explicitly address this issue.

**Mask-unmask color-consistent decoder** To train the decoder to ensure color-consistency between generated latent and unmasked pixels, we re-formulate the decoding as a local harmonization task. Our decoder involves additional inputs of masked images in the pixel-wise color space and the 0-1 mask. To properly train the decoder, we propose the color and latent augmentation as shown in Fig. 7 to estimate and enlarge the color-inconsistency. We follow the standard VAE training pipeline, but replacing the inputs with augmented ones. Particularly, we use the original image as the reconstruction target and use color and latent augmentation to corrupt input image, simulating the information loss of VAE and domain gap between generated and real latent, respectively. This forces the decoder to reconstruct the clean image based on the ground-truth unmasked region.

**Color augmentation** We use color augmentation to capture the VAE loss as in Fig. 8 (b). Empirically, further conditioned on unmasked image alleviate but not solve the color inconsistency issue, as shown in Fig. 8 (c). Hence, we need to explicitly train the decoder to ensure color consistency. To this end, we augment all training images in brightness, contrast, saturation, and hue, and requires the decoder to reconstruct original image conditioned on the unaugmented unmasked image. This encourages the decoder to faithfully follow the unmasked regions.

**Latent augmentation** To simulate the gap between generated and real latent, we incorporate the artifacts generated from the generative models to train the decoder. However, denoising to real images iteratively is notably time-consuming, even with DDIM [73]. To balance the efficiency and efficacy, we design a one-step estimation. As our target is to capture the generation gap, we use the clean latent $\boldsymbol{z}_0$ and all-zero mask $\mathbf{M}$ as conditions. This tells the generator all the needed information to generate the clean latent, ensuring the generated latent preserves content and only shift from the generation gap. We follow the standard pipeline to estimate $\boldsymbol{z}_0$ with modified conditions as:

$$\hat{\boldsymbol{z}}_0 = \frac{1}{a}(\boldsymbol{z}_t - b\varepsilon_\theta\left([\boldsymbol{z}_t; \boldsymbol{z}_0; \mathbf{M}], t\right)), \tag{1}$$

where the timestep $t$ is randomly sampled from $[500, 1000)$; $a$ indicates the prescribed variance schedule, $a^2 + b^2 = 1$ in diffusion models while $a + b = 1$ in rectified flow models; $\varepsilon_\theta(\cdot)$ is the frozen generator take as inputs noised $\boldsymbol{z}_t$, unmasked $\boldsymbol{z}_0$, and all-zero masking $\mathbf{M}$. The large step denoising is chosen to increase the distribution gap, as empirically the generator could produce reliable results in small $t$ given the unmasked latent condition $\boldsymbol{z}_0$. Then we decode $\hat{\boldsymbol{z}}_0$ to image as the latent augmented inputs. This makes latent augmentation an off-line strategy. We apply latent augmentation to 50% training images. The fine-tuned decoder showcases superior consistency as compared in Fig. 8.

## 4. Experiments

**Evaluation datasets** We follow previous works to evaluate on the standard benchmark Places 2 [101] validation set of 36,500 images. In addition, to validate across different domains and mask styles, we construct a evaluation dataset, dubbed as MISATO, from Matterport3D [13], Flickr-Landscape [47], MegaDepth [45], and COCO 2014 [48] to handle indoor, outdoor, building and background inpainting, respectively. We select 500 representative examples of size $512^2$ and $1024^2$ from each dataset, forming a total of 2,000 testing examples. See details in the appendix.

**General evaluation metrics** We use the Learned Perceptual Image Patch Similarity (LPIPS) [97] to calculate the patch-level image distances, Fréchet Inception Distance (FID) [32] to compare the distribution distance between generated images and real images, and Paired/Unpaired Inception Discriminative Score (P-IDS/U-IDS) [99] to measure the human-inspired linear separability.

**Evaluate object hallucination and color-consistency** We introduce two new metrics to assess the object hallucination and color-consistency of inpainted images. (1) *CLIP@mask* (C@m): We use CLIP to get visual features from both the ground-truth and the inpainted masked region, then calculate their cosine similarity. Following the standard CLIP score, we multiply the result by 100 and clip negative values, yielding a range from 0 to 100. (2) *Gradient@edge* (G@e): We calculate the average pixel gradient difference along the edges of the masked region with respect to the ground-truth image to assess color smoothness. A smaller gradient difference means more similar color transitions to the ground-truth image and, therefore, less color shift.

**Competitors** We primarily use the SD v1.5 inpainting model [67] to analyze and compare ASUKA with competitors, while validating ASUKA's generalization ability with

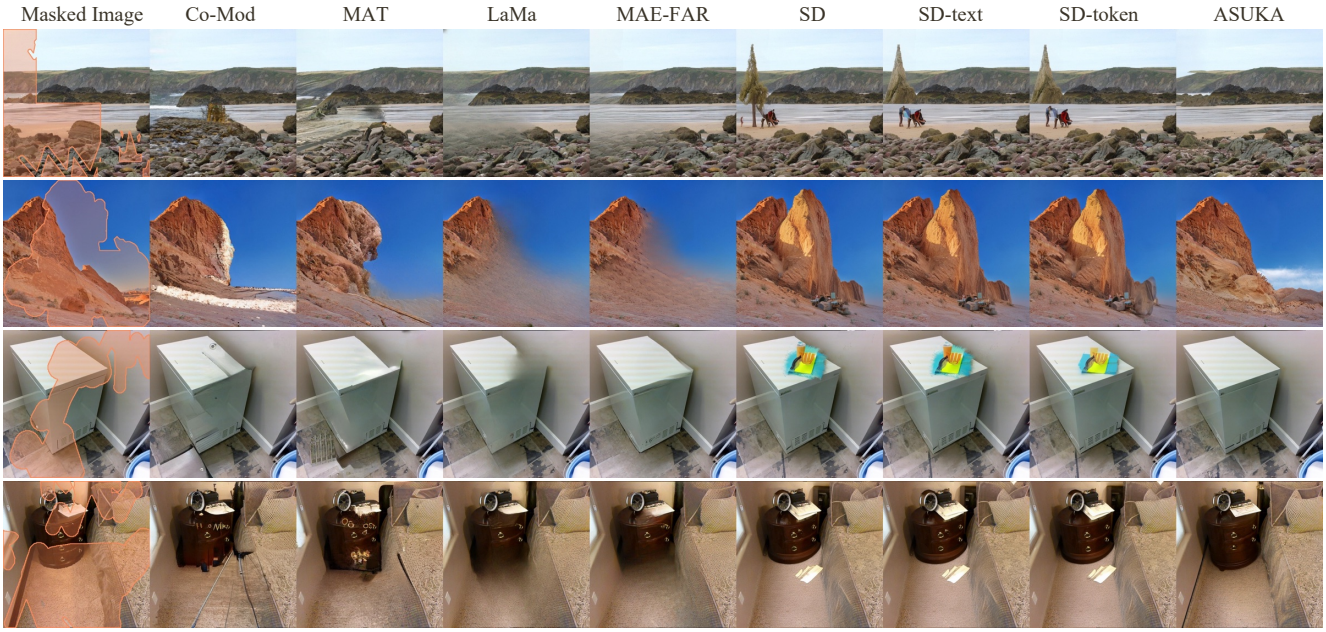


Figure 9. Inpainting results for $512^2$ images. GANs generate blurred results; SD variants hallucinate unreasonable objects and suffer from color shift. ASUKA achieves unwanted-object-mitigated and color-consistent inpainting. More results are in the appendix.

Table 1. Quantitative comparison on MISATO and Places 2. Top-3 results are colored.

| Dataset | MISATO | | | | | | Places 2 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | LPIPS↓ | FID↓ | U-IDS↑ | P-IDS↑ | C@m↑ | G@e↓ | LPIPS↓ | FID↓ | U-IDS↑ | P-IDS↑ | C@m↑ | G@e↓ |
| Co-Mod [99] | 0.179 | 17.421 | 0.243 | 0.109 | 0.924 | 52.106 | 0.267 | 5.794 | 0.274 | 0.096 | 0.951 | 166.914 |
| MAT [44] | 0.176 | 17.261 | 0.255 | 0.122 | 0.925 | 48.722 | 0.202 | 3.765 | 0.348 | 0.195 | 0.955 | 163.442 |
| LaMa [75] | 0.155 | 15.436 | 0.260 | 0.135 | 0.928 | 46.270 | 0.202 | 6.693 | 0.247 | 0.050 | 0.953 | 153.653 |
| MAE-FAR [10] | 0.142 | 13.283 | 0.282 | 0.153 | 0.940 | 43.613 | 0.174 | 3.559 | 0.307 | 0.105 | 0.958 | 149.843 |
| SD-Repaint [54] | 0.227 | 27.861 | 0.016 | 0.007 | 0.915 | 80.410 | 0.251 | 12.466 | 0.217 | 0.045 | 0.947 | 176.421 |
| SD [67] | 0.168 | 12.812 | 0.345 | 0.211 | 0.951 | 63.844 | 0.193 | 1.514 | 0.375 | 0.207 | 0.959 | 160.705 |
| SD-text | 0.164 | 12.603 | 0.337 | 0.207 | 0.952 | 63.776 | 0.191 | 1.506 | 0.373 | 0.202 | 0.959 | 160.418 |
| SD-token [81] | 0.160 | 12.517 | 0.331 | 0.204 | 0.955 | 61.700 | 0.189 | 1.477 | 0.390 | 0.234 | 0.960 | 158.924 |
| SD-IP [88] | 0.157 | 12.204 | 0.398 | 0.242 | 0.956 | 62.704 | 0.186 | 1.539 | 0.389 | 0.173 | 0.953 | 148.571 |
| SD-T2I [58] | 0.166 | 13.806 | 0.365 | 0.222 | 0.949 | 63.866 | 0.195 | 1.720 | 0.384 | 0.160 | 0.951 | 148.549 |
| SD-CAEv2 [98] | 0.157 | 29.179 | 0.193 | 0.045 | 0.901 | 69.890 | 0.192 | 6.887 | 0.287 | 0.065 | 0.921 | 151.863 |
| SD-LaMa [75] | 0.157 | 12.159 | 0.390 | 0.256 | 0.956 | 62.726 | 0.188 | 1.522 | 0.389 | 0.168 | 0.953 | 148.461 |
| ASUKA-SD | 0.150 | 11.495 | 0.423 | 0.312 | 0.958 | 47.753 | 0.183 | 1.230 | 0.413 | 0.287 | 0.961 | 147.733 |

FLUX. We consider three SD v1.5 inpainting variants: SD: uses a null-prompt for unconditional generation; SD-text: uses "background" as a prompt since no captions are used in inpainting; SD-token [81]: uses learnable tokens trained with ASUKA's pipeline. To test other ways of incorporating the MAE condition, we implement the following: SD-IP, uses IP-Adapter [88]; SD-T2I, uses T2I-Adapter [58]; SD-CAEv2, uses a CLIP-style alignment module CAEv2 [98]; We also test SD-LaMa, which inputs LaMa [75] inpainting results instead of MAE. We also compare with leading inpainting algorithms Co-Mod [99], MAT [44], LaMa [75], MAE-FAR [10], and SD-Repaint [54]. We provide implementation details in the appendix.

### 4.1. Comparison on Benchmarks

**Quantitative comparison** Results on SD are reported in Tab. 1. **(1)**: Although ASUKA-SD is based on a fixed SD model, it consistently outperforms SD across all evaluation metrics, achieving state-of-the-art results in FID, U-IDS, and P-IDS. Notably, U-IDS and P-IDS are closely aligned with human preferences [99] and have a potential maximum score of 0.5, highlighting ASUKA's strong performance. **(2)**: Compared to other adapters that align the MAE prior with SD, ASUKA-SD shows consistently supe-

Table 2. FLUX and ASUKA-FLUX on MISATO@512. Results on 1K and qualtitative results are in the appendix.

| Decoder | LPIPS↓ | FID↓ | U-IDS↑ | P-IDS↑ | C@m↑ | G@e↓ |
|---|---|---|---|---|---|---|
| FLUX | 0.254 | 12.839 | 0.351 | 0.223 | 0.951 | 65.928 |
| ASUKA-FLUX | **0.206** | **11.372** | **0.428** | **0.327** | **0.962** | **48.635** |

rior performance across all metrics. This demonstrates the effectiveness of our straightforward alignment module. **(3)**: While the LaMa condition improves inpainting quality, as shown by FID and IDS scores, it is less effective than the MAE condition. When using the MAE condition as a prior, improvements can be attributed to better mitigation of object hallucination. **(4)**: ASUKA-SD consistently performs better than all competitors on CLIP@mask, showcasing the strength of its improved mitigation of object hallucination. **(5)**: Pixel-based GAN inpainting models perform better in the Gradient@edge metric, suggesting that color shifts may originate from the compressed latent space. ASUKA-SD, however, still shows significant improvements over all SD variants, highlighting its enhanced color consistency. **(6)**: The second-to-best LPIPS scores are partially due to using a frozen SD, where ASUKA achieves consistent improvements but remains constrained by the frozen U-Net. These results confirm that ASUKA-SD improves color consistency and mitigation of object hallucination in inpainting, even when using frozen latent inpainting models. This advantage is evident both in the in-distribution dataset Places2 and the out-of-distribution dataset MISATO.

**Qualitative comparison** examples are shown in Fig. 9. **(1)** The state-of-the-art inpainting algorithms usually suffer from unnatural generation, for example the unnatural boundaries in the third and fourth rows, and failed inpainting of tower in the third-to-last row. LaMa and MAE-FAR sometimes lead to blurred inpainting results, especially in the scenario of large continuous masks. **(2)** The SD variants usually suffer from the unwanted object insertion issue and hallucinate unreasonable objects, in almost all the illustrated images. **(3)** In contrast, ASUKA enjoys unwanted-object-mitigated and color-consistent inpainting.

### 4.2. Further Analysis of ASUKA

In this part, we conduct more experiments to analysis ASUKA. More analysis can be found in the appendix.

**Extension to FLUX** To demonstrate ASUKA's versatility, we trained it on FLUX (see Tab. 2). ASUKA-FLUX consistently outperforms the original FLUX. Results on Places 2 and qualitative comparisons are in the appendix.

**Ablation of decoder** For the decoder, we compare ASUKA-SD with (1) *VAE*: the decoder used in SD; (2) + *cond.*: the decoder conditioned on unmasked image [103]; (3) + *color*: only trained with color augmentation ; Results are in Tab. 3, showing the superiority of our decoder.

Table 3. Comparison of different decoders for SD.

| Decoder | LPIPS↓ | FID↓ | U-IDS↑ | P-IDS↑ | C@m↑ | G@e↓ |
|---|---|---|---|---|---|---|
| VAE | 0.156 | 11.949 | 0.387 | 0.253 | 0.953 | 63.142 |
| + cond. | 0.151 | 11.634 | 0.410 | 0.272 | 0.955 | 48.588 |
| + color | 0.152 | 11.603 | 0.407 | 0.273 | 0.954 | 49.538 |
| Ours | **0.150** | **11.495** | **0.423** | **0.312** | **0.958** | **47.753** |

Table 4. Ablation of different alignment modules.

| Align | LPIPS↓ | FID↓ | U-IDS↑ | P-IDS↑ | C@m↑ | G@e↓ |
|---|---|---|---|---|---|---|
| linear | 0.155 | 11.934 | 0.400 | 0.263 | 0.953 | 48.983 |
| attn | 0.152 | 11.613 | 0.403 | 0.268 | 0.954 | 48.785 |
| cross x4 | 0.152 | 11.762 | 0.405 | 0.256 | 0.953 | 48.279 |
| Ours | **0.150** | **11.495** | **0.423** | **0.312** | **0.958** | **47.753** |

**Ablation of alignment module** We validate the efficacy of our alignment module step by step: (1) *linear*: Use linear layer to align feature dimension only; (2) *attn*: Based on *linear*, further use a single self-attention block to align the distribution; (3) *cross x4*: we instead use learnable query and 4 cross-attention layers to learn the MAE prior. ASUKA-SD adopts 4 self-attention blocks. Results are shown in Tab. 4. The self-attention block shows improved results compared with only align dimension and cross-attention block. Using 4 self-attention blocks improves the capacity.

## 5. Conclusion

In this paper, we proposed Aligned Stable inpainting with Unknown Areas prior (ASUKA) to achieve unwanted-object-mitigated and color-consistent inpainting via frozen latent inpainting models. To avoid unwanted object insertion, we adopt a reconstruction-based masked auto-encoder (MAE) as the context-stable prior for masked region purely from unmasked region. Then we align the context-stable prior to frozen generative models with the proposed alignment module. To achieve color-consistency, we resolve the mask-unmask color inconsistency in the latent decoding process. We train an unmask-region conditioned VAE decoder to perform local harmonization during the decoding process. To validate the efficacy of inpainting algorithms in different image domains and mask types, we introduce an evaluation dataset, named as MISATO, from existing datasets. We propose two new metrics to explicitly evaluate the object hallucination and color-consistency of inpainted images. ASUKA enjoys unwanted-object-mitigated and color-consistent inpainting results and superior than leading inpainting models.

**Acknowledgments** The authors would like to thank Huawei Ascend Cloud Ecological Development Project for the support of Ascend 910 processors.

## References


[1] Michael S Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *The Eleventh International Conference on Learning Representations*, 2023. 1

[2] AlimamaCreative. Flux-controlnet-inpainting, 2024. 3, 1

[3] Vladimir Arkhipkin, Andrei Filatov, Viacheslav Vasilev, Anastasia Maltseva, Said Azizov, Igor Pavlov, Julia Agafonova, Andrey Kuznetsov, and Denis Dimitrov. Kandinsky 3.0 technical report, 2023. 3

[4] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. In *International Conference on Learning Representations*, 2022. 3

[5] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. Patchmatch: A randomized correspondence algorithm for structural image editing. *ACM Trans. Graph.*, 28(3):24, 2009. 2

[6] Marcelo Bertalmio, Guillermo Sapiro, Vincent Caselles, and Coloma Ballester. Image inpainting. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 417–424, 2000. 1, 2

[7] Marcelo Bertalmio, Luminita Vese, Guillermo Sapiro, and Stanley Osher. Simultaneous structure and texture image inpainting. *IEEE transactions on image processing*, 12(8): 882–889, 2003. 2

[8] Alper Canberk, Maksym Bondarenko, Ege Ozguroglu, Ruoshi Liu, and Carl Vondrick. Erasedraw: Learning to insert objects by erasing them from images. In *European Conference on Computer Vision*, pages 144–160. Springer, 2024. 3

[9] Chenjie Cao and Yanwei Fu. Learning a sketch tensor space for image inpainting of man-made scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14509–14518, 2021. 1

[10] Chenjie Cao, Qiaole Dong, and Yanwei Fu. Learning prior feature and attention enhanced image inpainting. In *European Conference on Computer Vision*, pages 306–322. Springer, 2022. 2, 3, 4, 7

[11] Chenjie Cao, Yunuo Cai, Qiaole Dong, Yikai Wang, and Yanwei Fu. Leftrefill: Filling right canvas based on left reference through generalized text-to-image diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7705–7715, 2024. 3

[12] Tony F Chan and Jianhong Shen. Nontexture inpainting by curvature-driven diffusions. *Journal of visual communication and image representation*, 12(4):436–449, 2001. 2

[13] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3D: Learning from RGB-D data in indoor environments. *International Conference on 3D Vision (3DV)*, 2017. 2, 6

[14] Xiaokang Chen, Mingyu Ding, Xiaodi Wang, Ying Xin, Shentong Mo, Yunhao Wang, Shumin Han, Ping Luo, Gang Zeng, and Jingdong Wang. Context autoencoder for self-supervised representation learning. *International Journal of Computer Vision*, pages 1–16, 2023. 3

[15] Yifu Chen, Jingwen Chen, Yingwei Pan, Yehao Li, Ting Yao, Zhineng Chen, and Tao Mei. Improving text-guided object inpainting with semantic pre-inpainting. In *European Conference on Computer Vision*, pages 110–126. Springer, 2024. 3

[16] Mang Tik Chiu, Yuqian Zhou, Lingzhi Zhang, Zhe Lin, Connelly Barnes, Sohrab Amirghodsi, Eli Shechtman, and Humphrey Shi. Brush2prompt: Contextual prompt generator for object inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12636–12645, 2024. 3

[17] Tianyi Chu, Jiafu Chen, Jiakai Sun, Shuobin Lian, Zhizhong Wang, Zhiwen Zuo, Lei Zhao, Wei Xing, and Dongming Lu. Rethinking fast fourier convolution in image inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23195–23205, 2023. 2

[18] Adobe Creative Cloud. Adobe firefly, 2023. 3

[19] Wenyan Cong, Jianfu Zhang, Li Niu, Liu Liu, Zhixin Ling, Weiyuan Li, and Liqing Zhang. Dovenet: Deep image harmonization via domain verification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8394–8403, 2020. 3

[20] Wenyan Cong, Xinhao Tao, Li Niu, Jing Liang, Xuesong Gao, Qihao Sun, and Liqing Zhang. High-resolution image harmonization via collaborative dual transformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18470–18479, 2022. 3

[21] Antonio Criminisi, Patrick Pérez, and Kentaro Toyama. Object removal by exemplar-based inpainting. *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, 2:II–II, 2003. 1

[22] Antonio Criminisi, Patrick Pérez, and Kentaro Toyama. Region filling and object removal by exemplar-based image inpainting. *IEEE Transactions on image processing*, 13(9): 1200–1212, 2004. 2

[23] Pau de Jorge, Riccardo Volpi, Puneet K Dokania, Philip HS Torr, and Grégory Rogez. Placing objects in context via inpainting for out-of-distribution segmentation. In *European Conference on Computer Vision*, pages 456–473. Springer, 2024. 3

[24] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12873–12883, 2021. 1

[25] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024. 3

[26] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *The Eleventh International Conference on Learning Representations*, 2022. 3

[27] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 1, 2

[28] Zonghui Guo, Haiyong Zheng, Yufeng Jiang, Zhaorui Gu, and Bing Zheng. Intrinsic image harmonization. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 16367–16376, 2021. 3

[29] Zonghui Guo, Zhaorui Gu, Bing Zheng, Junyu Dong, and Haiyong Zheng. Transformer for image harmonization and beyond. *IEEE transactions on pattern analysis and machine intelligence*, 45(11):12960–12977, 2022. 3

[30] James Hays and Alexei A Efros. Scene completion using millions of photographs. *ACM Transactions on Graphics (ToG)*, 26(3):4–es, 2007. 1

[31] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 2, 3, 1

[32] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 6

[33] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1

[34] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2021. 3

[35] Xuan Ju, Xian Liu, Xintao Wang, Yuxuan Bian, Ying Shan, and Qiang Xu. Brushnet: A plug-and-play image inpainting model with decomposed dual-branch diffusion. In *European Conference on Computer Vision*, pages 150–168. Springer, 2024. 3

[36] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation, 2018. 2

[37] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 2

[38] Diederik P Kingma. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014. 1

[39] Keunsoo Ko and Chang-Su Kim. Continuously masked transformer for image inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13169–13178, 2023. 2

[40] Black Forest Labs. Flux.1, 2024. 1, 2, 3

[41] Anat Levin, Assaf Zomet, and Yair Weiss. Learning how to inpaint from global image statistics. *Proceedings Ninth IEEE International Conference on Computer Vision*, pages 305–312 vol.1, 2003. 1

[42] Fan Li, Zixiao Zhang, Yi Huang, Jianzhuang Liu, Renjing Pei, Bin Shao, and Songcen Xu. Magiceraser: Erasing any objects via semantics-aware control. In *European Conference on Computer Vision*, pages 215–231. Springer, 2024. 3

[43] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 1

[44] Wenbo Li, Zhe Lin, Kun Zhou, Lu Qi, Yi Wang, and Jiaya Jia. Mat: Mask-aware transformer for large hole image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2, 7, 3

[45] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 6

[46] Liang Liao, Jing Xiao, Zheng Wang, Chia-Wen Lin, and Shin'ichi Satoh. Guidance and evaluation: Semantic-aware image inpainting for mixed scenes. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16*, pages 683–700. Springer, 2020. 1

[47] Chieh Hubert Lin, Yen-Chi Cheng, Hsin-Ying Lee, Sergey Tulyakov, and Ming-Hsuan Yang. InfinityGAN: Towards infinite-pixel image synthesis. In *International Conference on Learning Representations*, 2022. 2, 6

[48] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 2, 4, 6

[49] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. 1

[50] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 85–100, 2018. 2

[51] Sheng Liu, Cong Phuoc Huynh, Cong Chen, Maxim Arap, and Raffay Hamid. Lemart: Label-efficient masked region transform for image harmonization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18290–18299, 2023. 3

[52] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023. 1

[53] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018. 1

[54] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vi-*

*sion and Pattern Recognition*, pages 11461–11471, 2022. 7, 2

[55] Feng Luo, Jinxi Xiang, Jun Zhang, Xiao Han, and Wei Yang. Image super-resolution via latent diffusion: A sampling-space mixture of experts and frequency-augmented decoder approach. *arXiv preprint arXiv:2310.12004*, 2023. 3

[56] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2021. 3

[57] Quanling Meng, Liu Qinglin, Zonglin Li, Xiangyuan Lan, Shengping Zhang, and Liqiang Nie. High-resolution image harmonization with adaptive-interval color transformation. *Advances in Neural Information Processing Systems*, 37: 13769–13793, 2024. 3

[58] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4296–4304, 2024. 7

[59] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Qureshi, and Mehran Ebrahimi. Edgeconnect: Structure guided image inpainting using edge prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019. 1, 3

[60] Li Niu, Junyan Cao, Wenyan Cong, and Liqing Zhang. Deep image harmonization with learnable augmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7482–7491, 2023. 3

[61] OpenAI. Openai's consistency decoder, 2023. 3

[62] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016. 1, 2

[63] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 1

[64] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1

[65] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint*, 2022. 3

[66] Mengwei Ren, Wei Xiong, Jae Shin Yoon, Zhixin Shu, Jianming Zhang, HyunJoon Jung, Guido Gerig, and He Zhang. Relightful harmonization: Lighting-aware portrait background replacement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6452–6462, 2024. 3

[67] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 1, 2, 3, 5, 6, 7

[68] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 1

[69] Stefan Roth and Michael J. Black. Fields of experts: a framework for learning image priors. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 2:860–867 vol. 2, 2005. 1

[70] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022. 3

[71] Nirat Saini, Navaneeth Bodla, Ashish Shrivastava, Avinash Ravichandran, Xiao Zhang, Abhinav Shrivastava, and Bharat Singh. Invi: Object insertion in videos using off-the-shelf diffusion models. *arXiv preprint arXiv:2407.10958*, 2024. 3

[72] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 1

[73] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*, 2021. 6

[74] Michael Steinbach, George Karypis, and Vipin Kumar. A comparison of document clustering techniques. *Department of Computer Science and Engineering, University of Minnesota*, 2000. 1

[75] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2149–2159, 2022. 1, 2, 3, 4, 7

[76] Yi-Hsuan Tsai, Xiaohui Shen, Zhe Lin, Kalyan Sunkavalli, Xin Lu, and Ming-Hsuan Yang. Deep image harmonization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3789–3797, 2017. 3

[77] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 1

[78] Ziyu Wan, Jingbo Zhang, Dongdong Chen, and Jing Liao. High-fidelity pluralistic image completion with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4692–4701, 2021. 1

[79] Ke Wang, Michaël Gharbi, He Zhang, Zhihao Xia, and Eli Shechtman. Semi-supervised parametric real-world image harmonization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5927–5936, 2023. 3

[80] Su Wang, Chitwan Saharia, Ceslee Montgomery, Jordi Pont-Tuset, Shai Noy, Stefano Pellegrini, Yasumasa Onoe, Sarah Laszlo, David J Fleet, Radu Soricut, et al. Imagen editor and editbench: Advancing and evaluating text-guided image inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18359–18369, 2023. 3

[81] Yikai Wang, Chenjie Cao, Ke Fan, Qiaole Dong, Yifan Li, Xiangyang Xue, and Yanwei Fu. Repositioning the subject within image. *Transactions on Machine Learning Research*, 2024. 3, 7

[82] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14668–14678, 2022. 3

[83] Longhui Wei, Lingxi Xie, Wengang Zhou, Houqiang Li, and Qi Tian. Mvp: Multimodality-guided visual pre-training. In *European Conference on Computer Vision*, pages 337–353. Springer, 2022. 3

[84] Daniel Winter, Matan Cohen, Shlomi Fruchter, Yael Pritch, Alex Rav-Acha, and Yedid Hoshen. Objectdrop: Bootstrapping counterfactuals for photorealistic object removal and insertion. In *European Conference on Computer Vision*, pages 112–129. Springer, 2024. 3

[85] Shaoan Xie, Zhifei Zhang, Zhe Lin, Tobias Hinz, and Kun Zhang. Smartbrush: Text and shape guided object inpainting with diffusion model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22428–22437, 2023. 3

[86] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9653–9663, 2022. 3

[87] Xingqian Xu, Shant Navasardyan, Vahram Tadevosyan, Andranik Sargsyan, Yadong Mu, and Humphrey Shi. Image completion with heterogeneously filtered spectral hints. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4591–4601, 2023. 2

[88] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint*, 2023. 7

[89] Zili Yi, Qiang Tang, Shekoofeh Azizi, Daesik Jang, and Zhan Xu. Contextual residual aggregation for ultra high-resolution image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7508–7517, 2020. 2

[90] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5505–5514, 2018. 2

[91] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4471–4480, 2019. 2

[92] Yu Zeng, Zhe Lin, Jimei Yang, Jianming Zhang, Eli Shechtman, and Huchuan Lu. High-resolution image inpainting with iterative confidence feedback and guided upsampling. In *European Conference on Computer Vision*, pages 1–17. Springer, 2020. 2

[93] Yanhong Zeng, Jianlong Fu, Hongyang Chao, and Baining Guo. Aggregated contextual transformations for high-resolution image inpainting. *IEEE Transactions on Visualization and Computer Graphics*, 2022. 2

[94] Yu Zeng, Zhe Lin, and Vishal M Patel. Shape-guided object inpainting. *arXiv preprint arXiv:2204.07845*, 2022. 3

[95] Dengyong Zhang, Zaoshan Liang, Gaobo Yang, Qingguo Li, Leida Li, and Xingming Sun. A robust forgery detection algorithm for object removal by exemplar-based image inpainting. *Multimedia Tools and Applications*, 77:11823–11842, 2018. 2

[96] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023. 3, 1

[97] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 6

[98] Xinyu Zhang, Jiahui Chen, Junkun Yuan, Qiang Chen, Jian Wang, Xiaodi Wang, Shumin Han, Xiaokang Chen, Jimin Pi, Kun Yao, et al. Cae v2: Context autoencoder with clip latent alignment. *Transactions on Machine Learning Research*, 2023. 7

[99] Shengyu Zhao, Jonathan Cui, Yilun Sheng, Yue Dong, Xiao Liang, I Eric, Chao Chang, and Yan Xu. Large scale image completion via co-modulated generative adversarial networks. In *International Conference on Learning Representations*, 2020. 1, 2, 4, 6, 7, 3

[100] Haitian Zheng, Zhe Lin, Jingwan Lu, Scott Cohen, Eli Shechtman, Connelly Barnes, Jianming Zhang, Ning Xu, Sohrab Amirghodsi, and Jiebo Luo. Image inpainting with cascaded modulation gan and object-aware training. In *European Conference on Computer Vision*, pages 277–296. Springer, 2022. 2

[101] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017. 2, 6, 1

[102] Jun-Yan Zhu, Philipp Krahenbuhl, Eli Shechtman, and Alexei A Efros. Learning a discriminative model for the perception of realism in composite images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3943–3951, 2015. 3

[103] Zixin Zhu, Xuelu Feng, Dongdong Chen, Jianmin Bao, Le Wang, Yinpeng Chen, Lu Yuan, and Gang Hua. Designing a better asymmetric vqgan for stablediffusion. *arXiv preprint arXiv:2306.04632*, 2023. 3, 5, 6, 8, 1

[104] Junhao Zhuang, Yanhong Zeng, Wenran Liu, Chun Yuan, and Kai Chen. A task is worth one word: Learning with task prompts for high-quality versatile image inpainting. In *European Conference on Computer Vision*, pages 195–211. Springer, 2024. 3

[105] zk. text-to-image-2m (revision e64fca4), 2024. 2

# 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

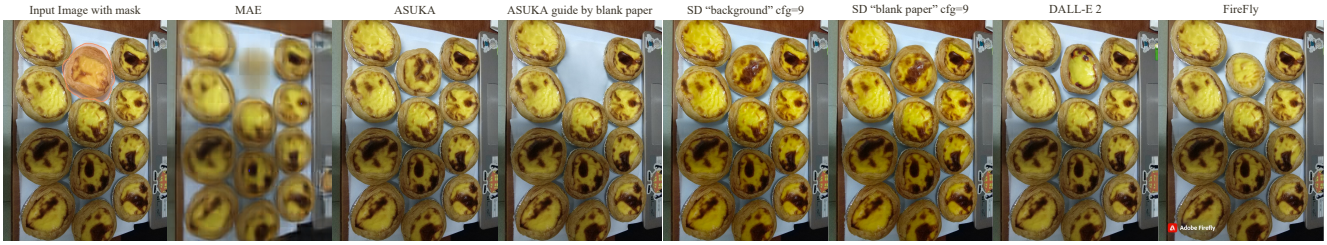


Figure 11. The curse of self-attention, causing the MAE falsely estimate the masked region and powerful text-guided diffusion models fail to generation content based on text prompts. ASUKA potential circumvents this issue by using a blank paper image as the input to the MAE to provide correct prior.

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** |

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.

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** |

**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 av-

erage 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.

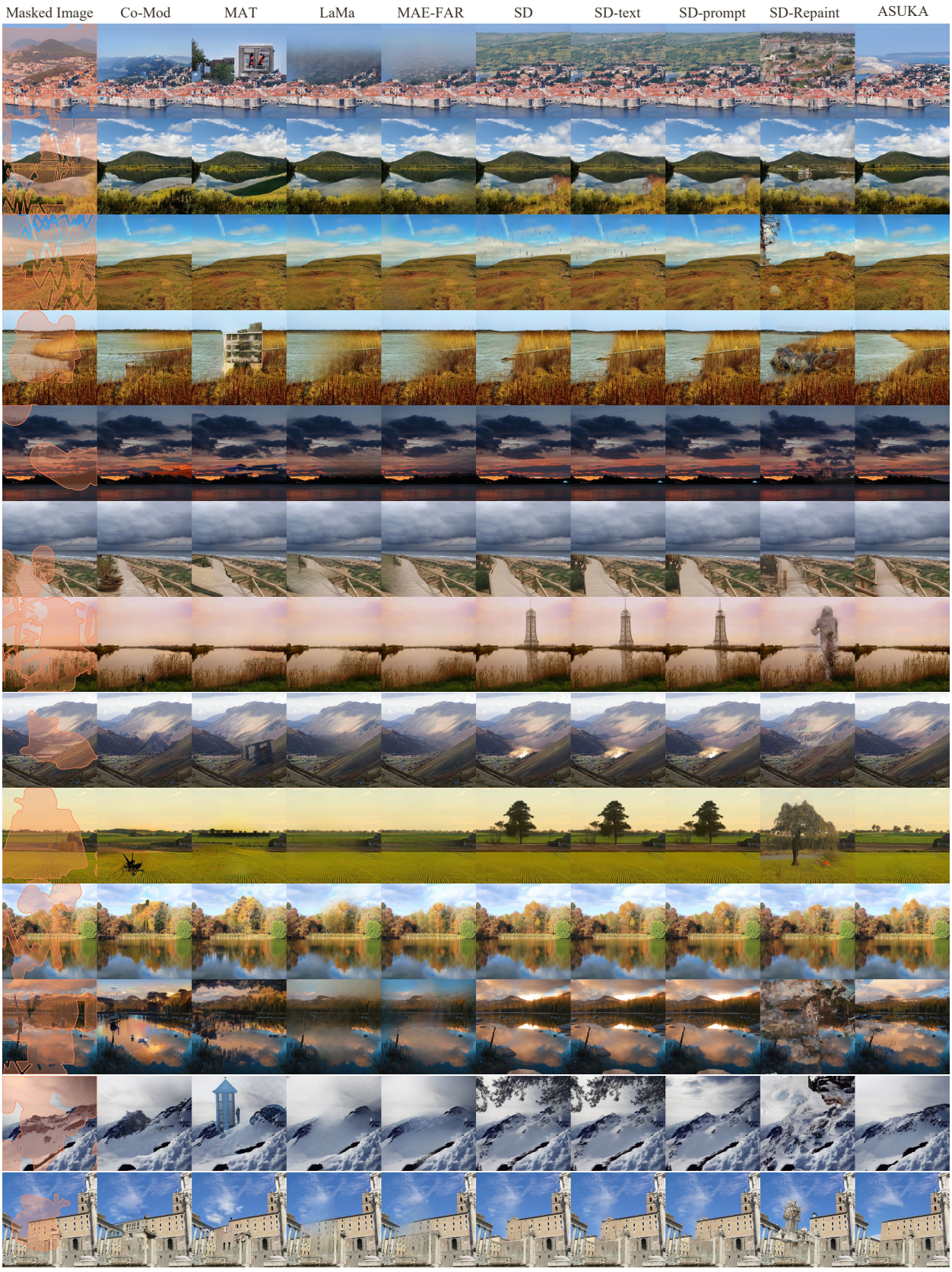


Figure 12. More qualitative comparison on MISATO.

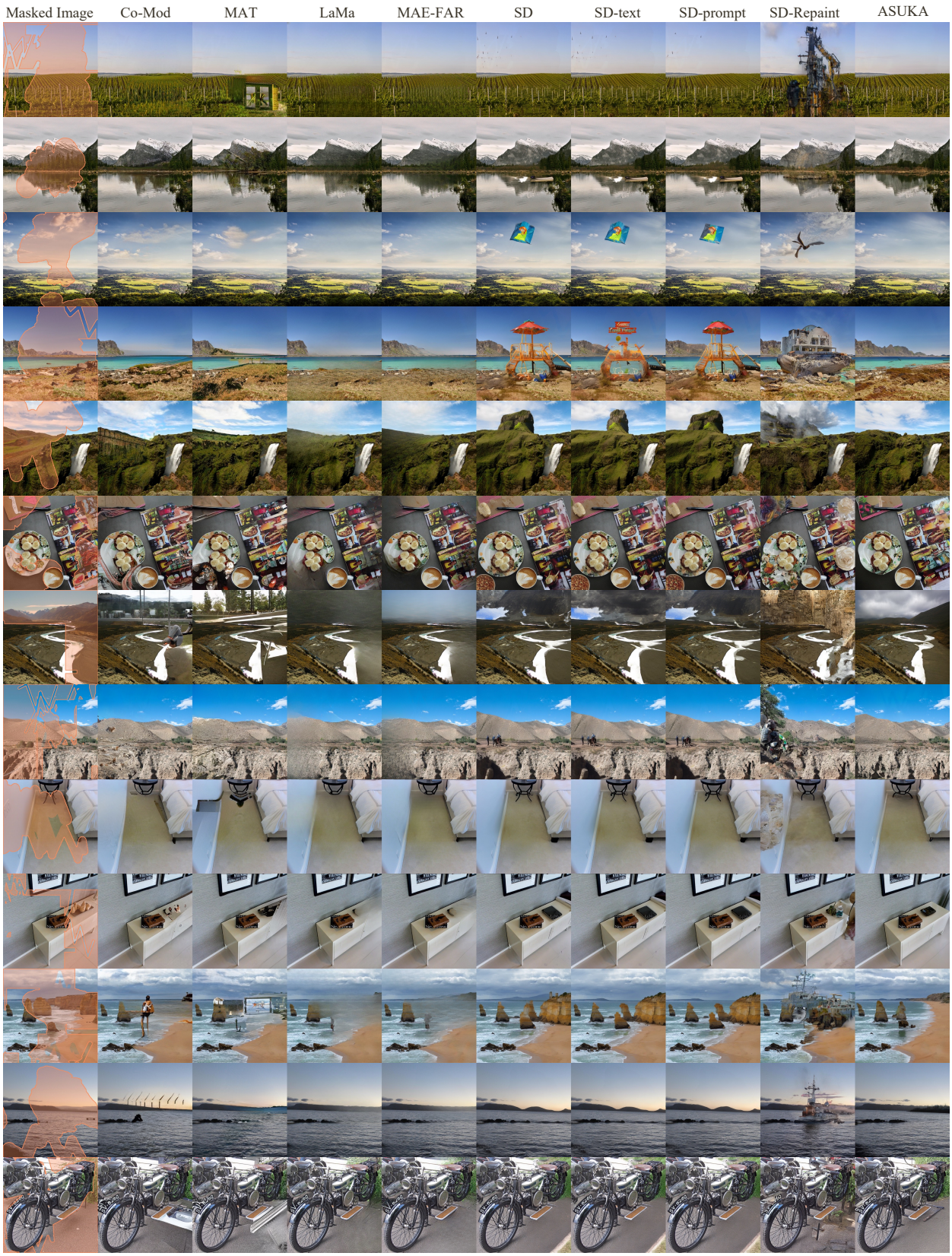


Figure 13. More qualitative comparison on MISATO.

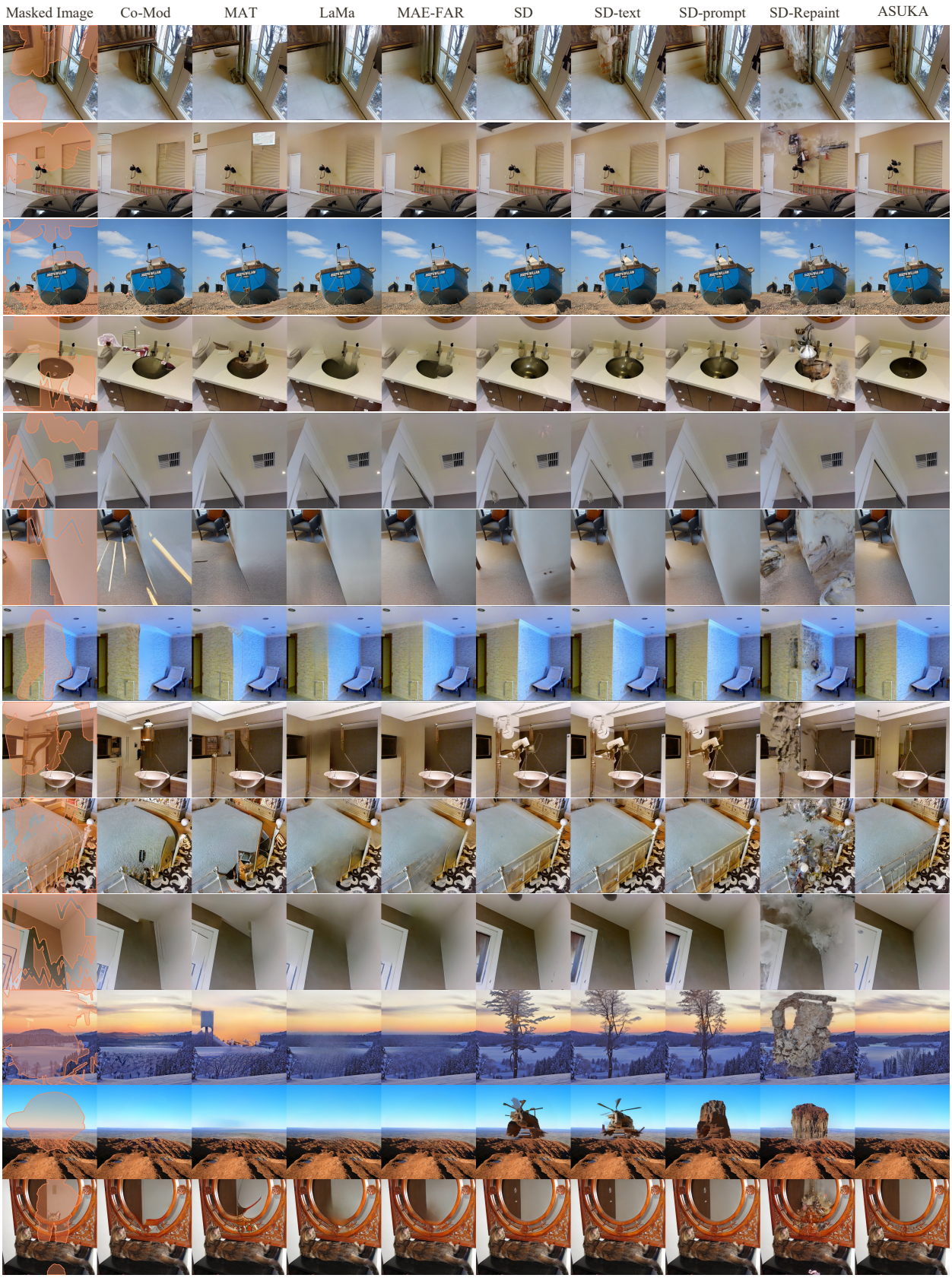


Figure 14. More qualitative comparison on MISATO.

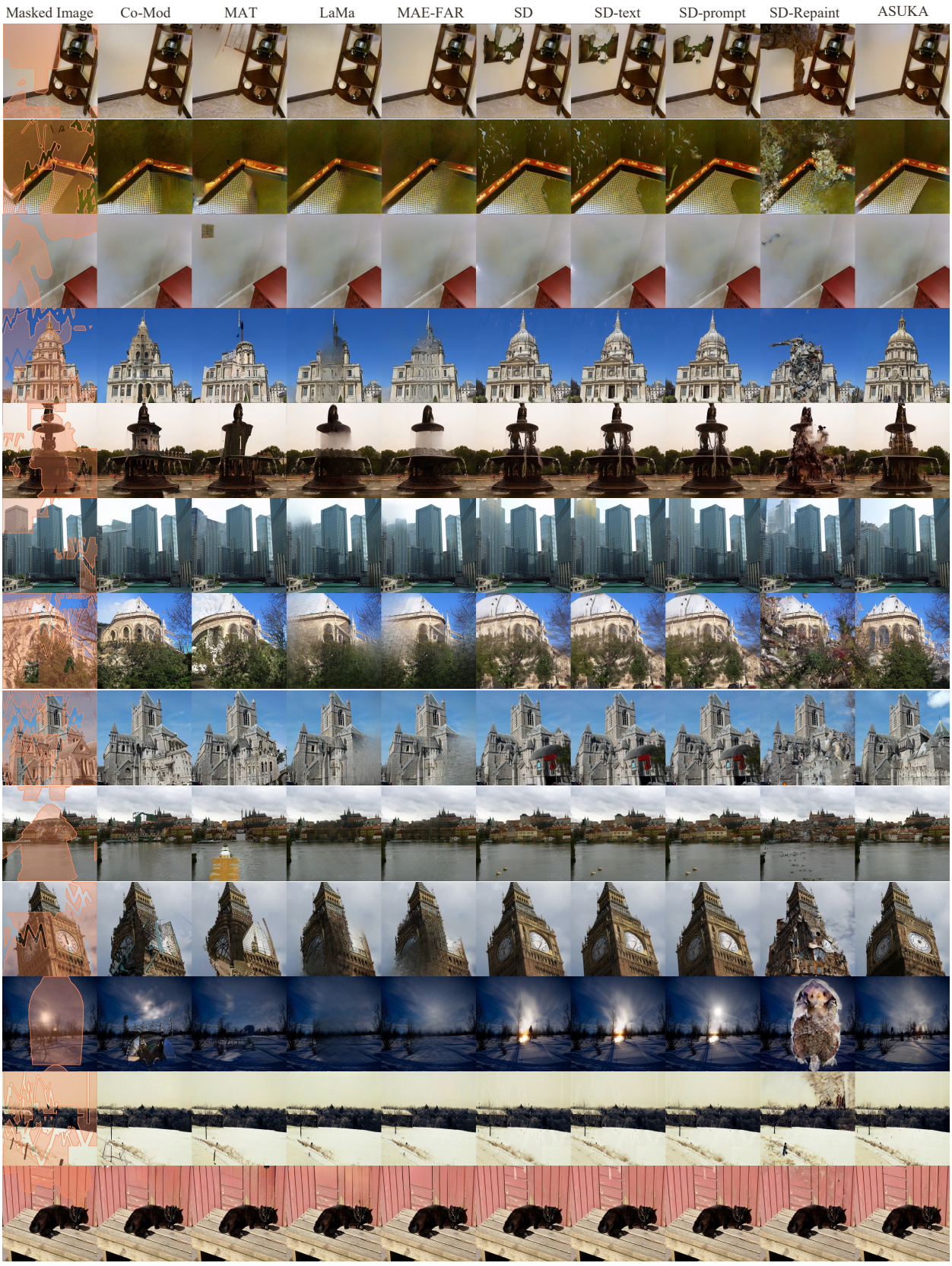


Figure 15. More qualitative comparison on MISATO.

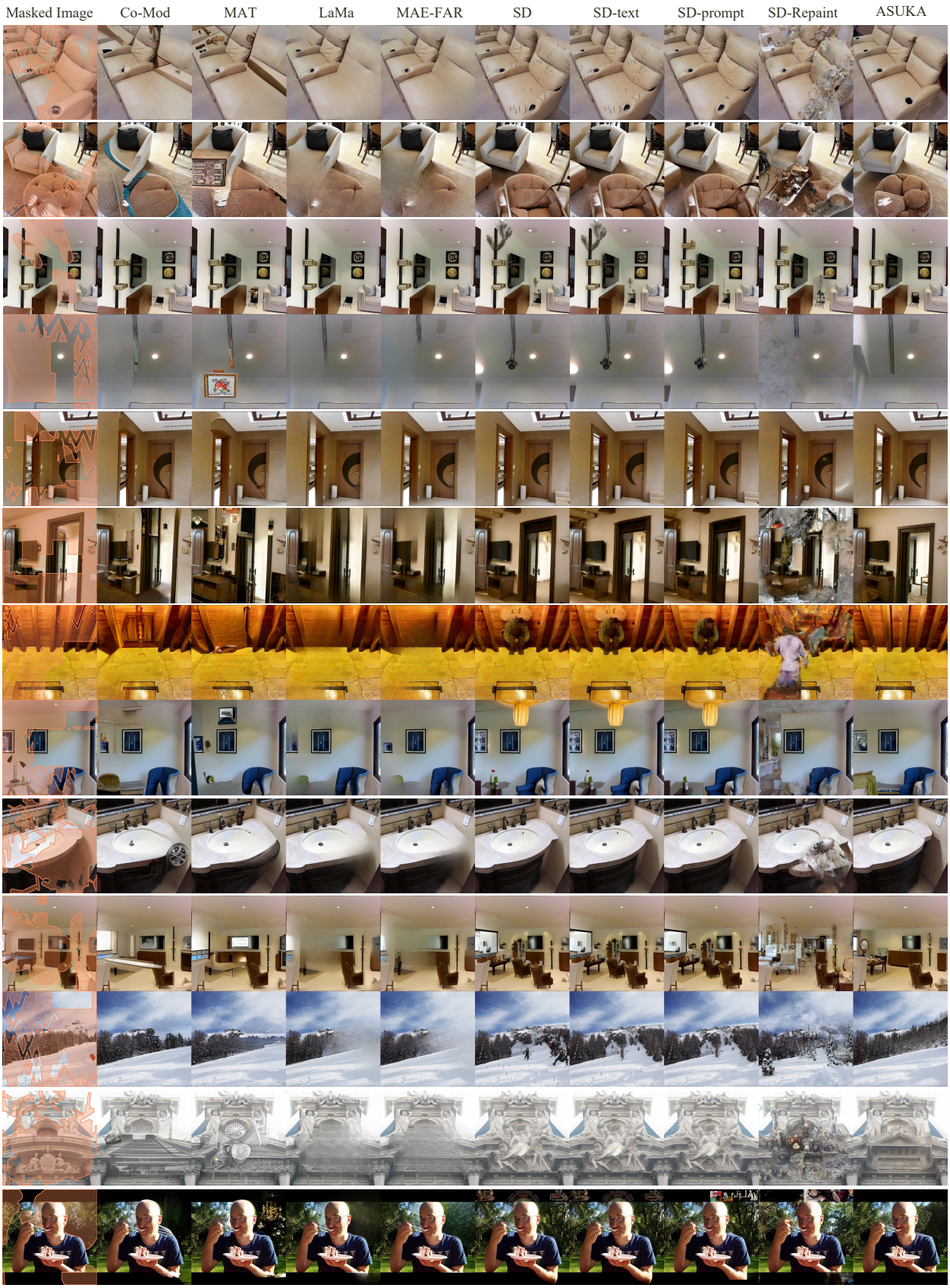


Figure 16. More qualitative comparison on MISATO.