Advancing Image Inpainting: From Versatility to Consistency

Yikai Wang
Fudan University

Image Inpainting: Task Definition

Image inpainting is the process of completing or recovering the missing region in the image.

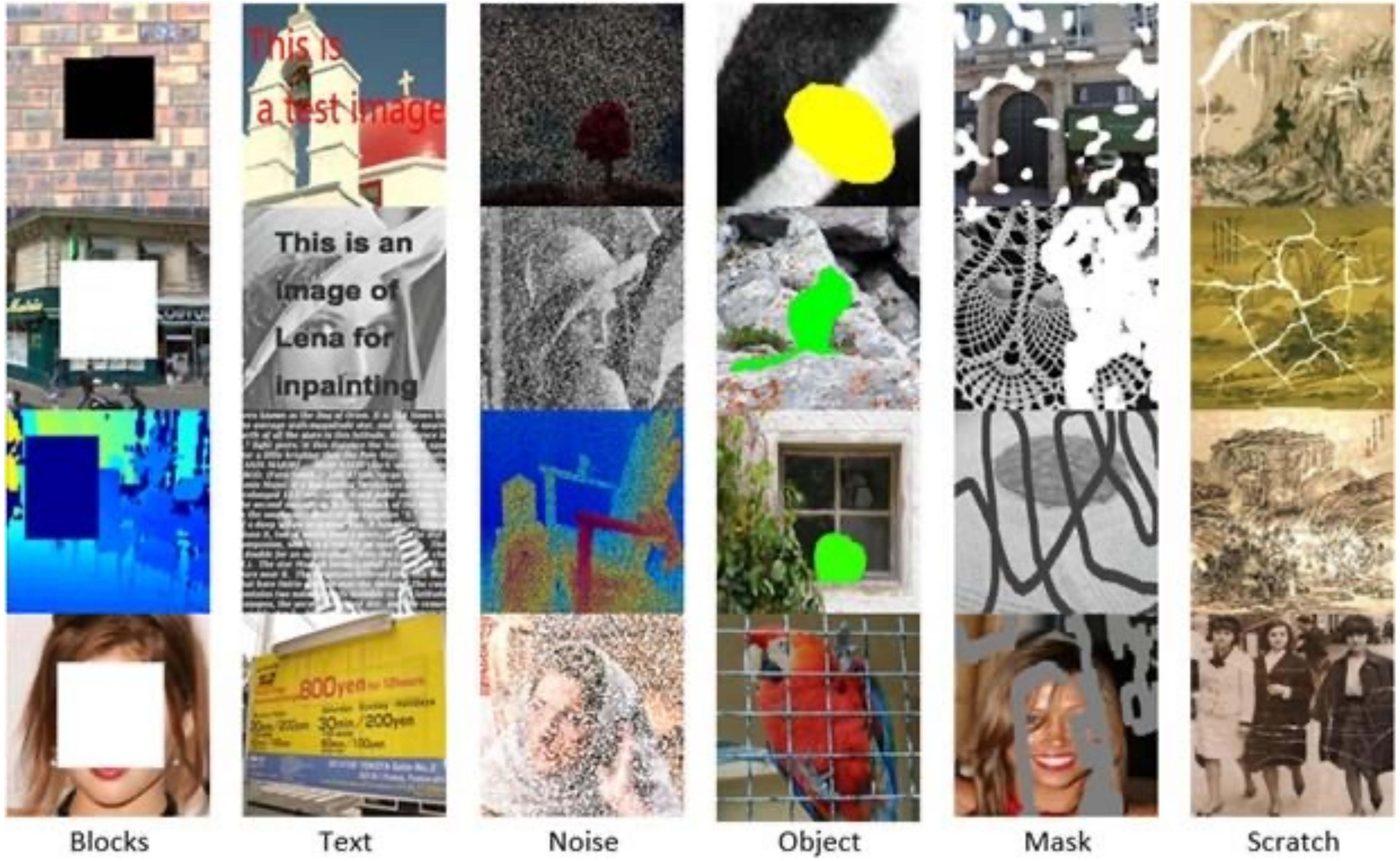


Image depicted from Elharrouss Omal, et al. " Image inpainting: A review. " Neural Processing Letters, 2020.

Conditional Image Inpainting

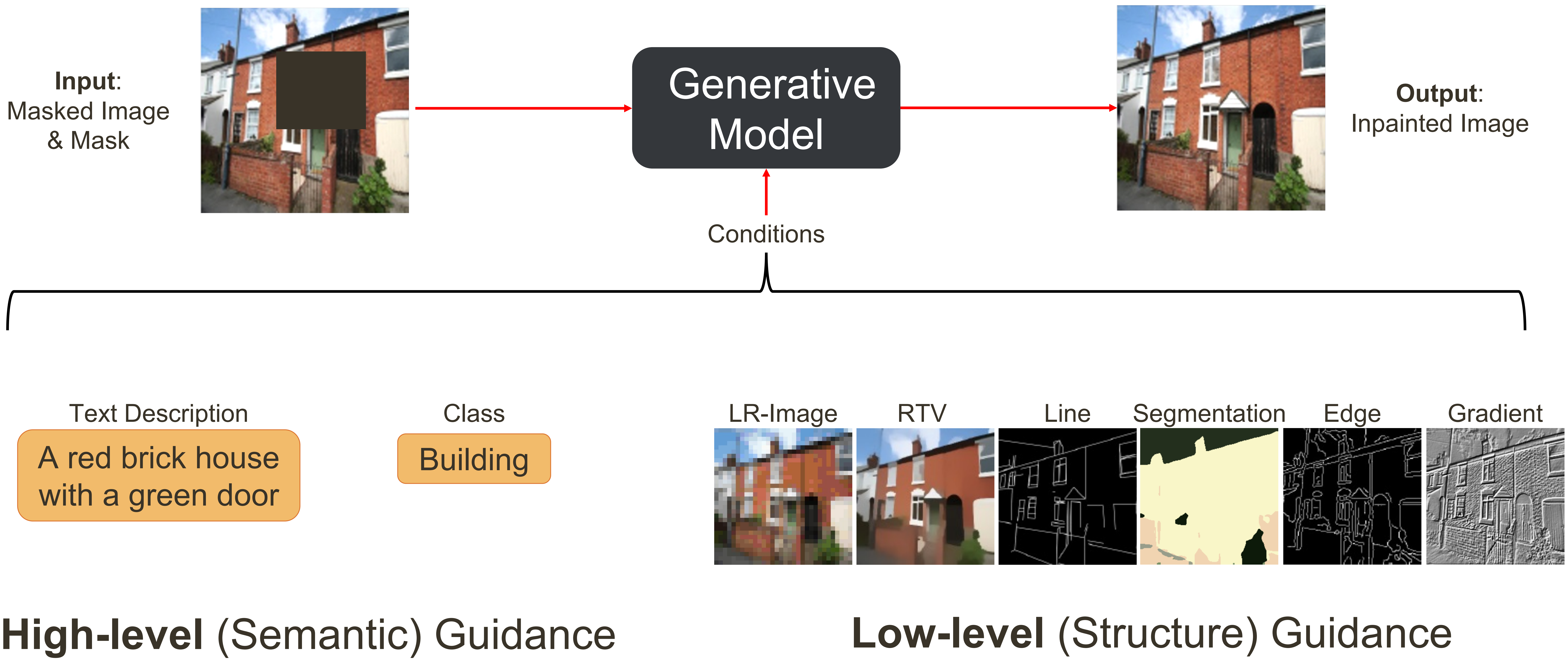
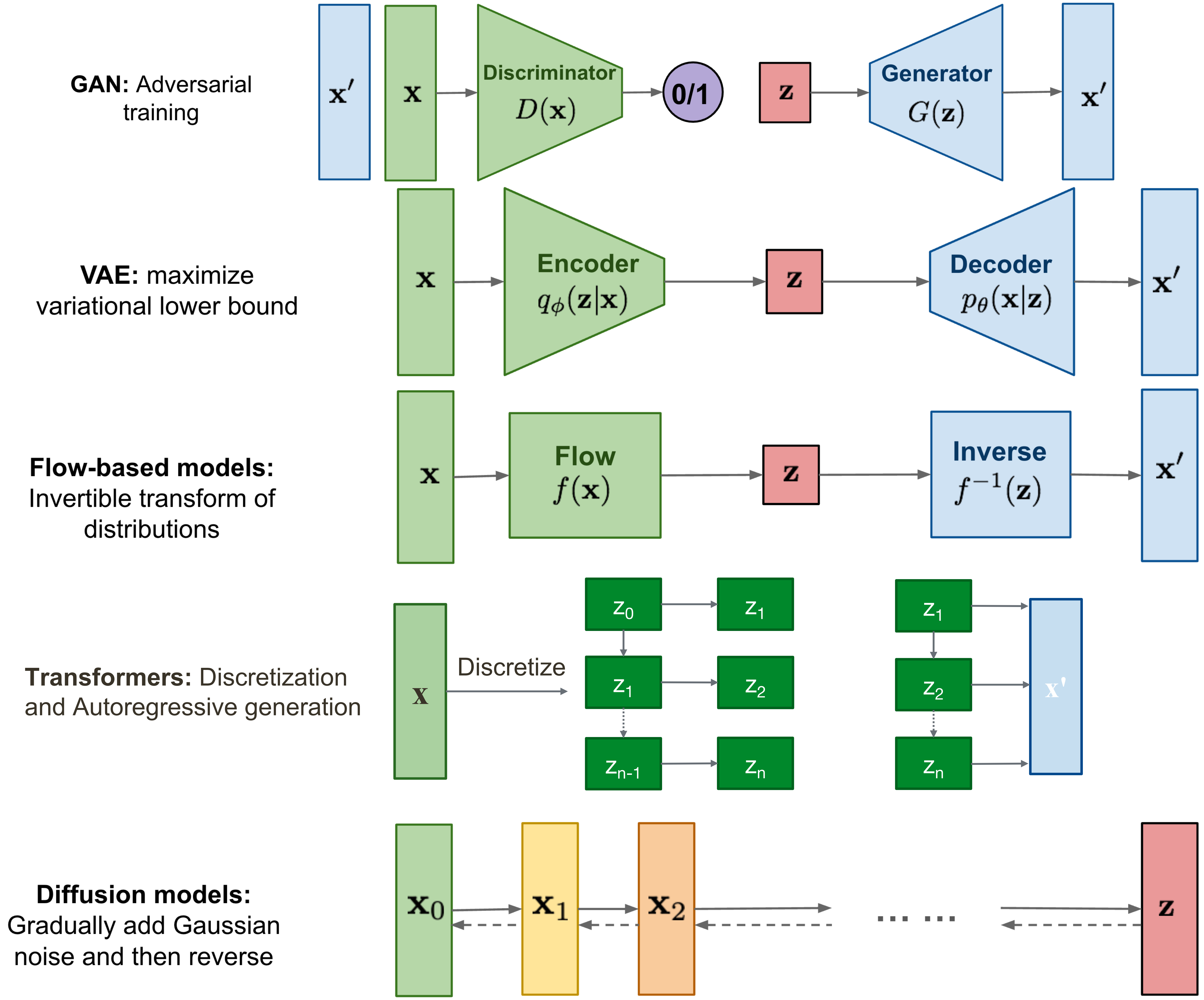


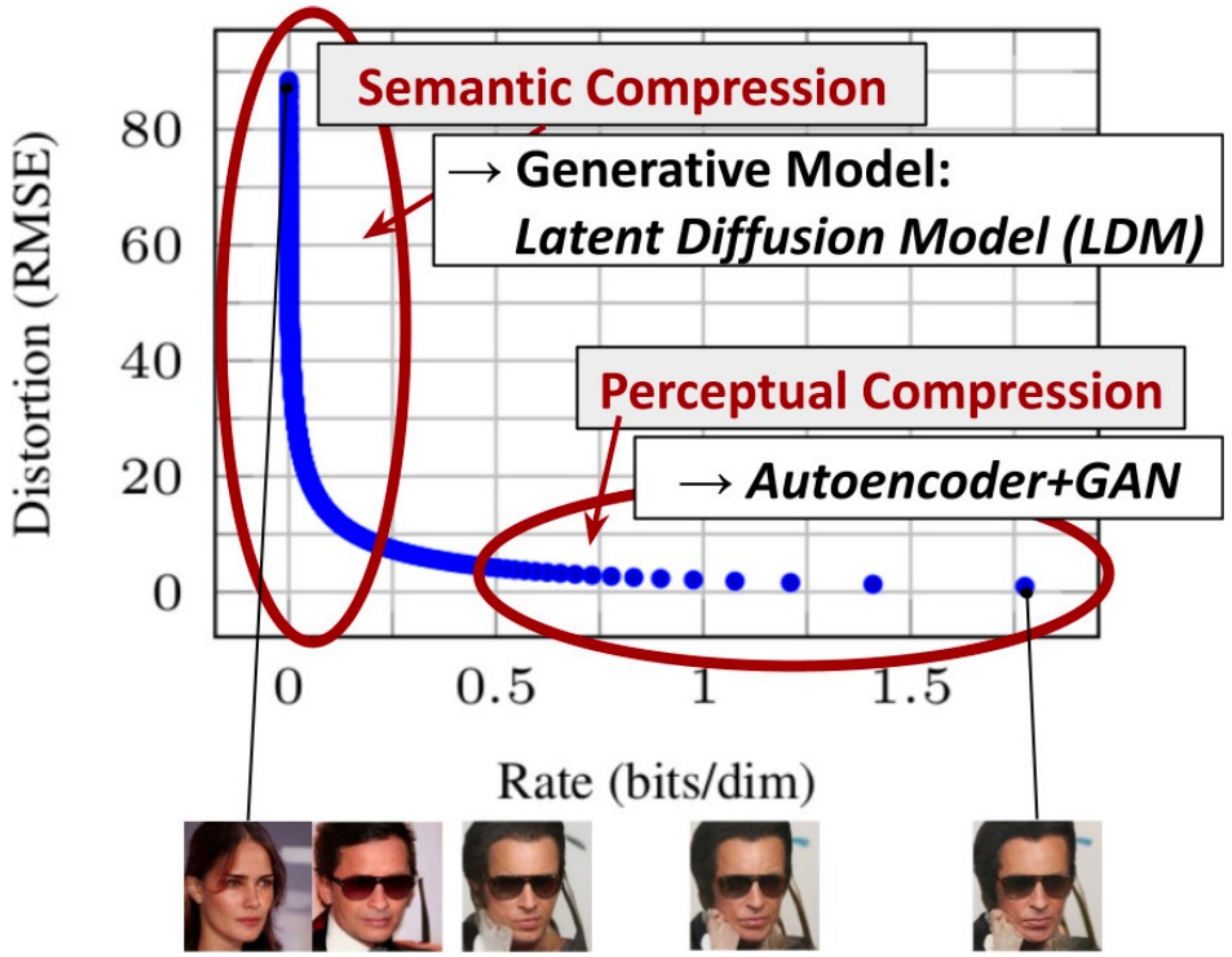
Image Inpainting: Generative Models

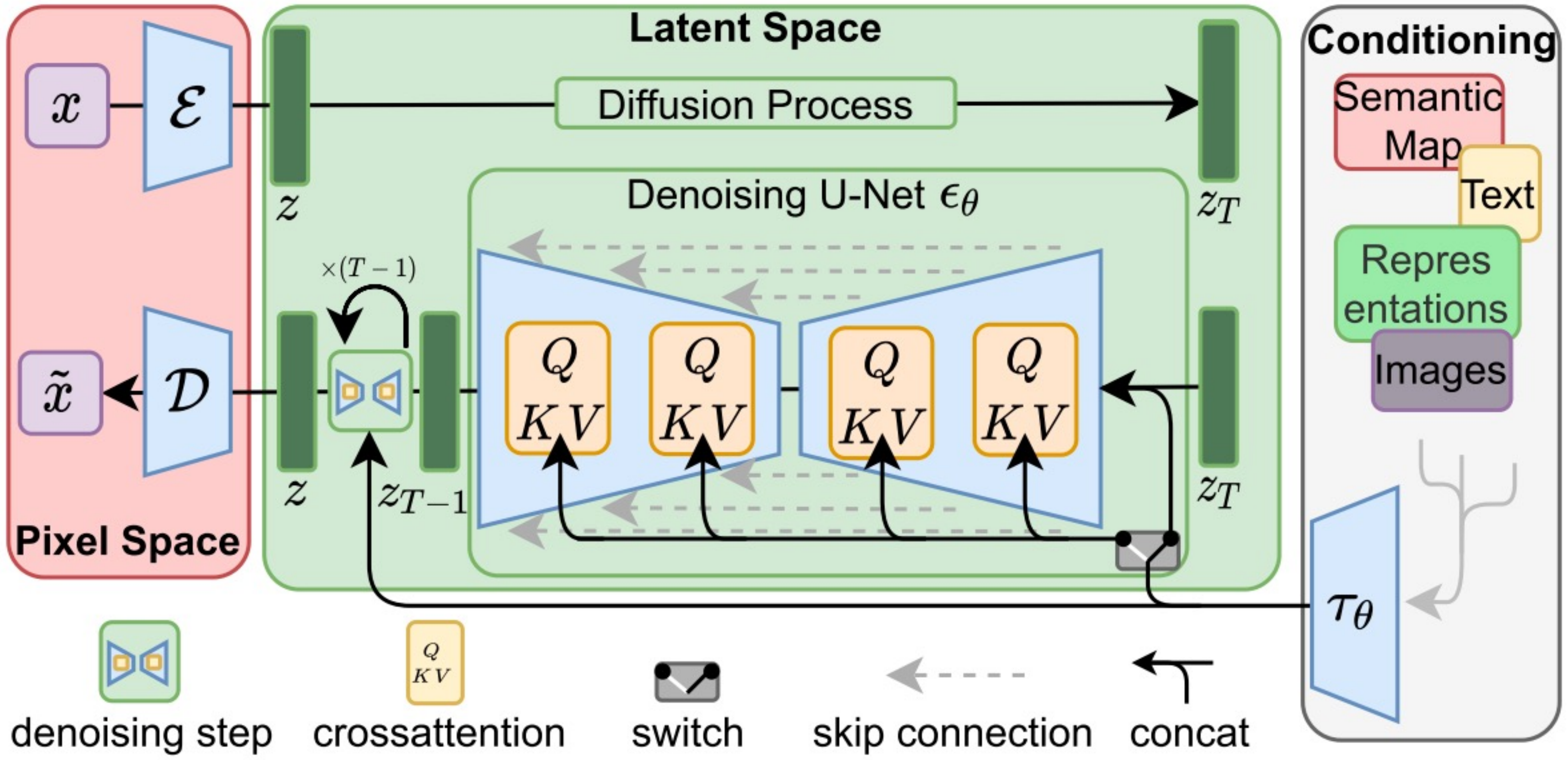
Main Idea: Model the inherence relationship within images or between images and some random distribution.



Part of the Image depicted from Weng, Lilian. (Jul 2021). What are diffusion models? Lil'Log. https://lilianweng.github.io/posts/2021-07-11-diffusion-models/

Stable Diffusion Inpainting Model



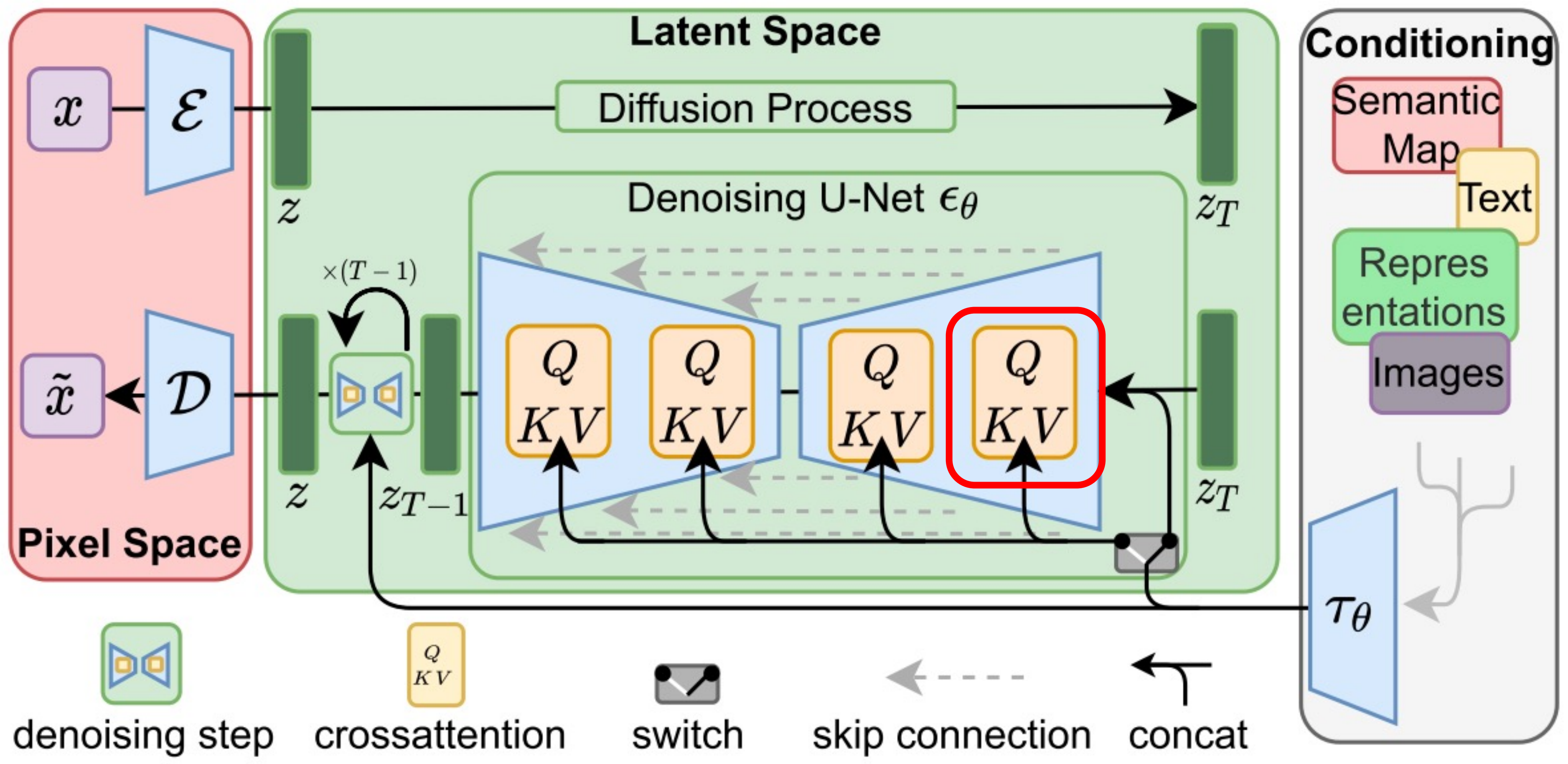


Perceptual Compression:
Down-sample the input sizes from the pixel-level x to latent z via VQ-VAE (discrete) or KL-VAE (continuous).

Latent Diffusion:
Perform diffusion process and inverse generation process in the latent space.

Robin Rombach, et al. "High-Resolution Image Synthesis with Latent Diffusion Models." *CVPR*. 2022.

Enhance frozen SD for non-text conditions



The cross-attention layers in text-to-image stable diffusion inpainting model is powerful enough to show emergency assumption:
can adapt to other non-text conditions without fine-tuning

Outline:

1. **Versatility**: Use a frozen SD to tackle all kinds of inpainting tasks.
2. **Consistency**: Improve the SD to more context-stable and visual-consistent inpainting.

Versatile Image Inpainting

for Subject Repositioning

1. Advancing Image Inpainting: Versatility
2. Advancing Image Inpainting: Consistency

Subject Repositioning

Segment

↓

Generate

↓

Blend

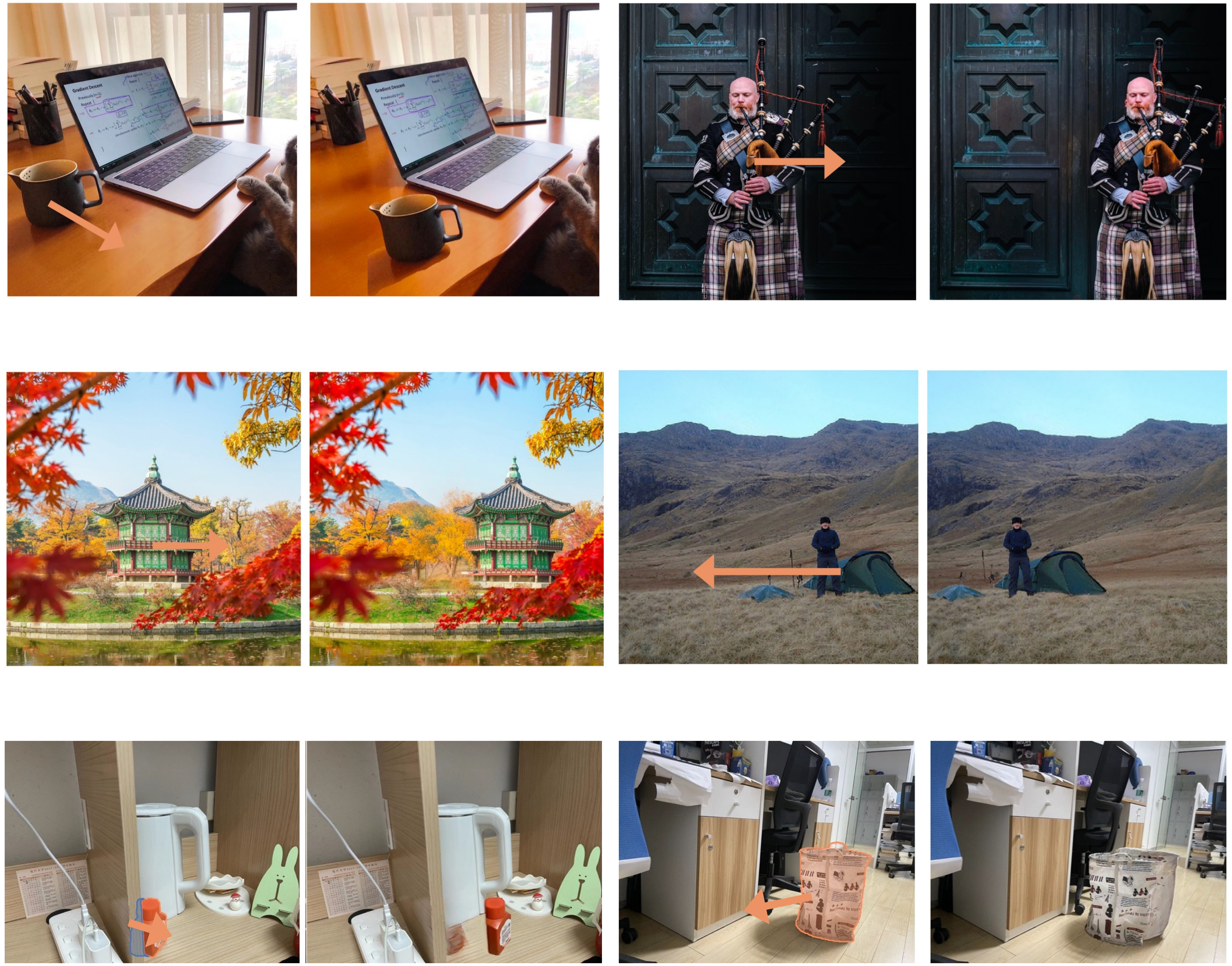
Yikai Wang et al. Repositioning the Subject within Image. Preprint, 2023.

Challenges in Subject Repositioning: Inconsistency

Appearance Inconsistency

Geometry Inconsistency

Semantic Inconsistency



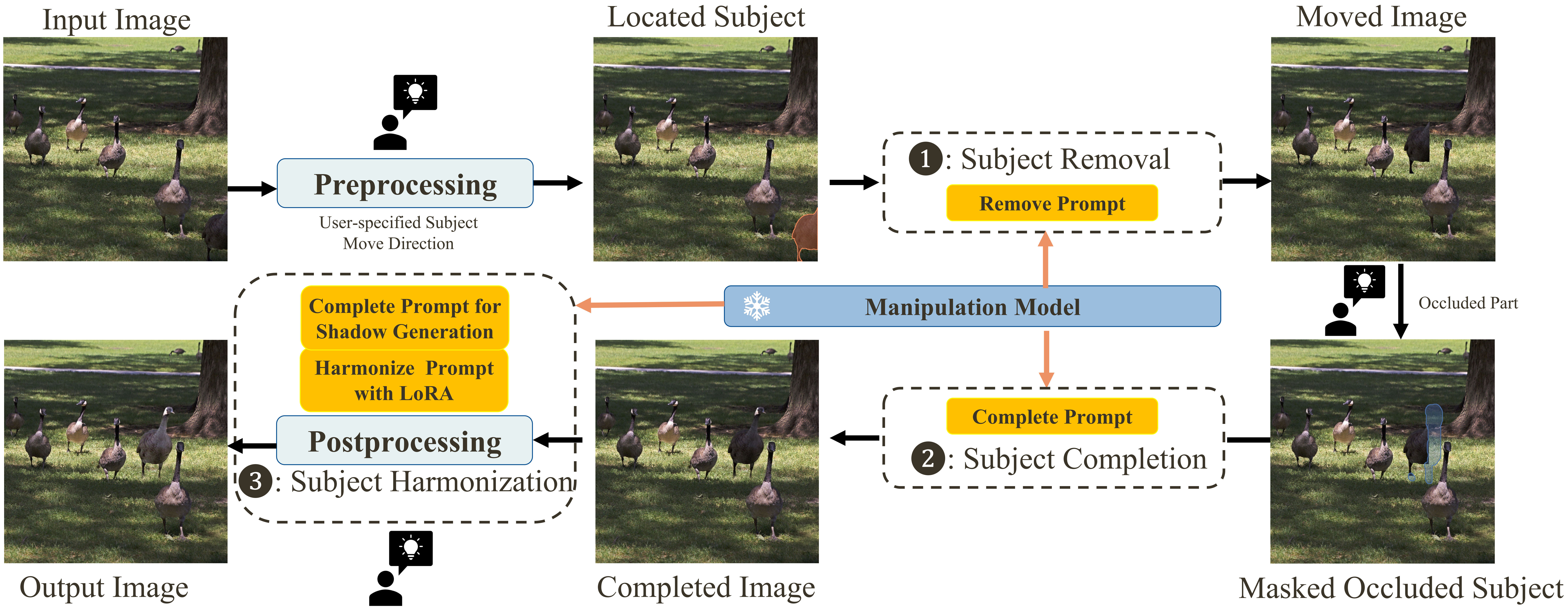
Shadows & Lightning

Occlusion & Perspective

Object & Background

Yikai Wang et al. Repositioning the Subject within Image. Preprint, 2023.

Deconstruct Subject Repositioning



Yikai Wang et al. Repositioning the Subject within Image. Preprint, 2023.

Generative Sub-Tasks in Subject Repositioning

Input Image SEELE

(a) Subject Removal

Input Image SEELE

(b) Subject Completion

Input Image SEELE

(c) Subject Harmonization

They are all image inpainting: take as inputs the masked image with mask, and take as output the inpainted image.

They require different generation capacity:

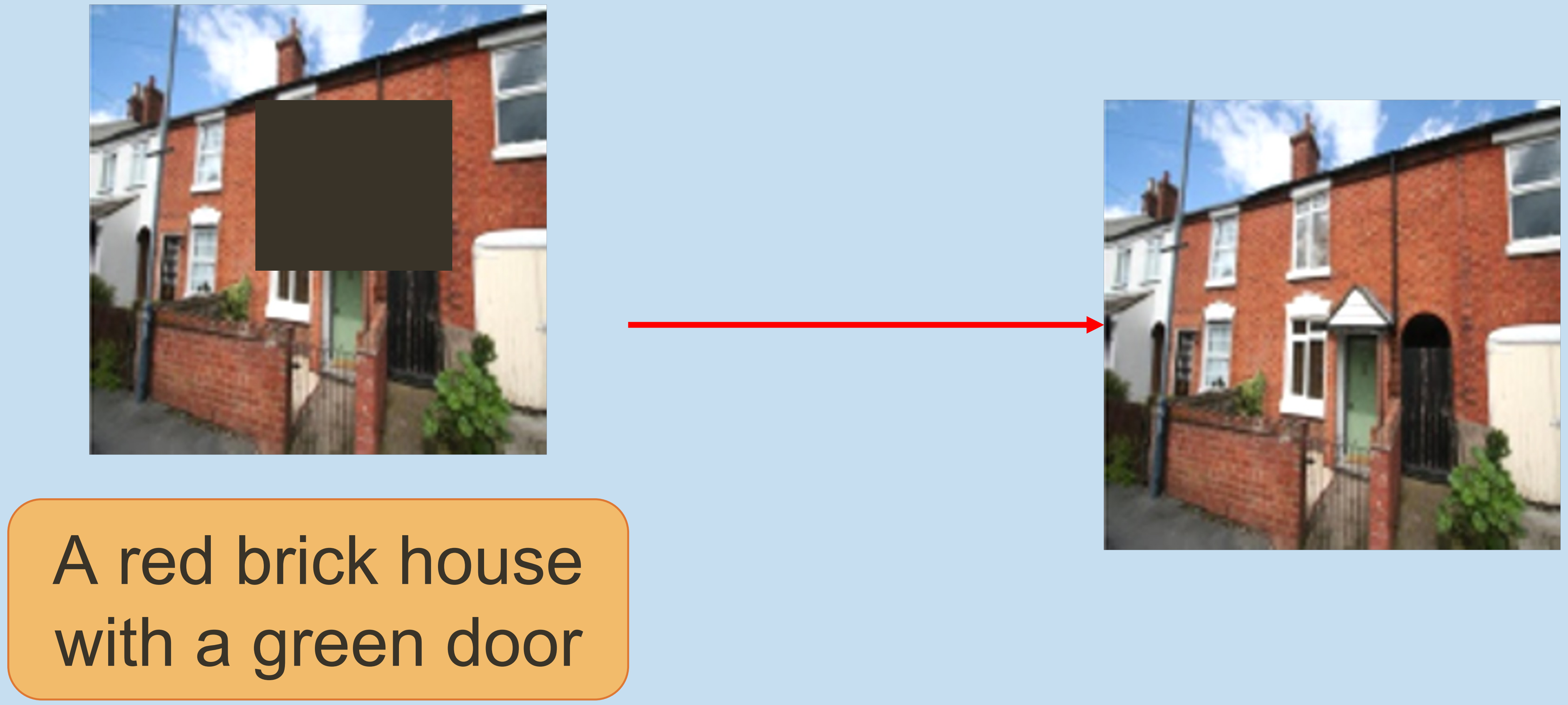
- **Subject removal** fills the void in original area without creating new subjects; semantic-less
- **Subject completion** completes the repositioned subject within masked region; semantic-rich
- **Subject harmonization** blends subject without inducing new elements. semantic-preserving

Can we tackle all these tasks within a single generative model?

Yikai Wang et al. Repositioning the Subject within Image. Preprint, 2023.

Task Inversion: Task-Level Instruction on SD

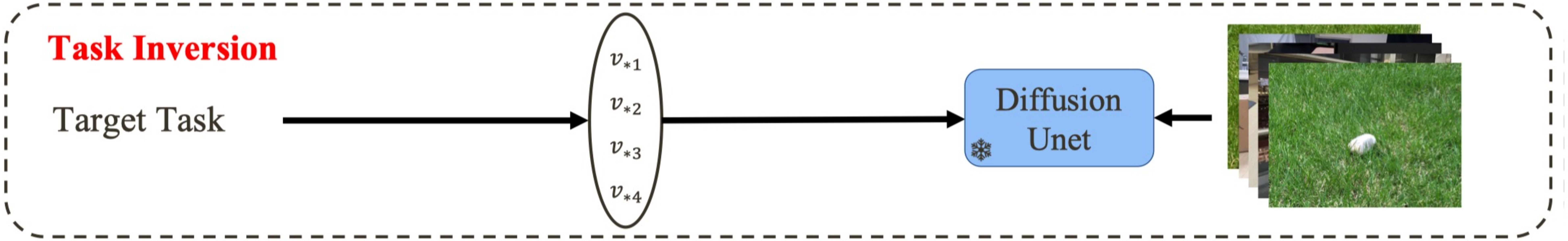
Text-to-image: Optimal but non-generalizable



Task-to-image: Generalizable but not trained

"Emergent" Assumption:
The cross-attention layer in stable diffusion inpainting model is powerful enough to enable non-text guidance.

Target: Train learnable prompts to approximate the behavior of image-dependent caption-style text guidance.



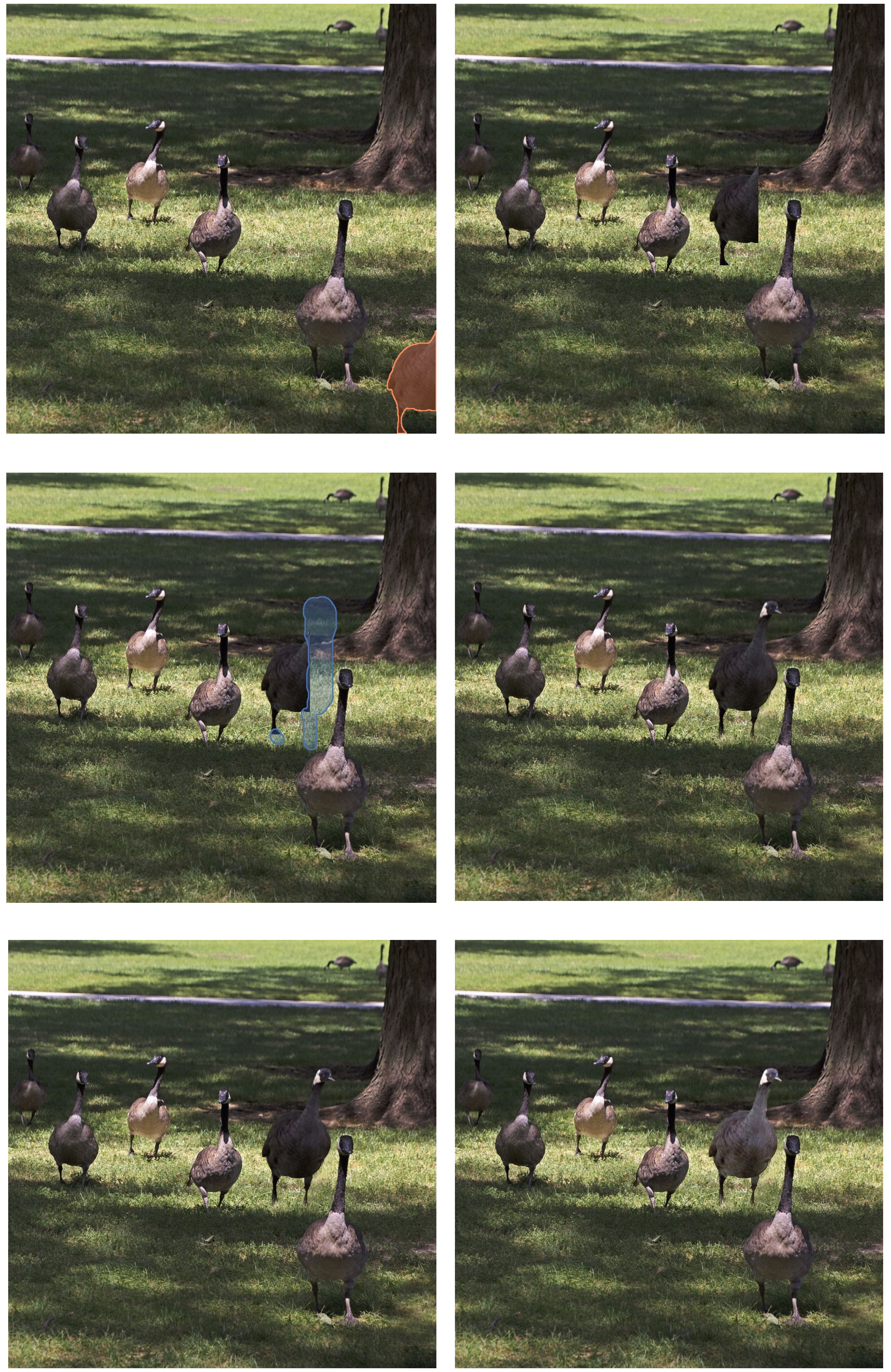
Yikai Wang et al. Repositioning the Subject within Image. Preprint, 2023.

Training Task Inversion: Training-Testing Consistency

Subject removal

Subject completion

Subject harmonization



move mask

↓

move subject

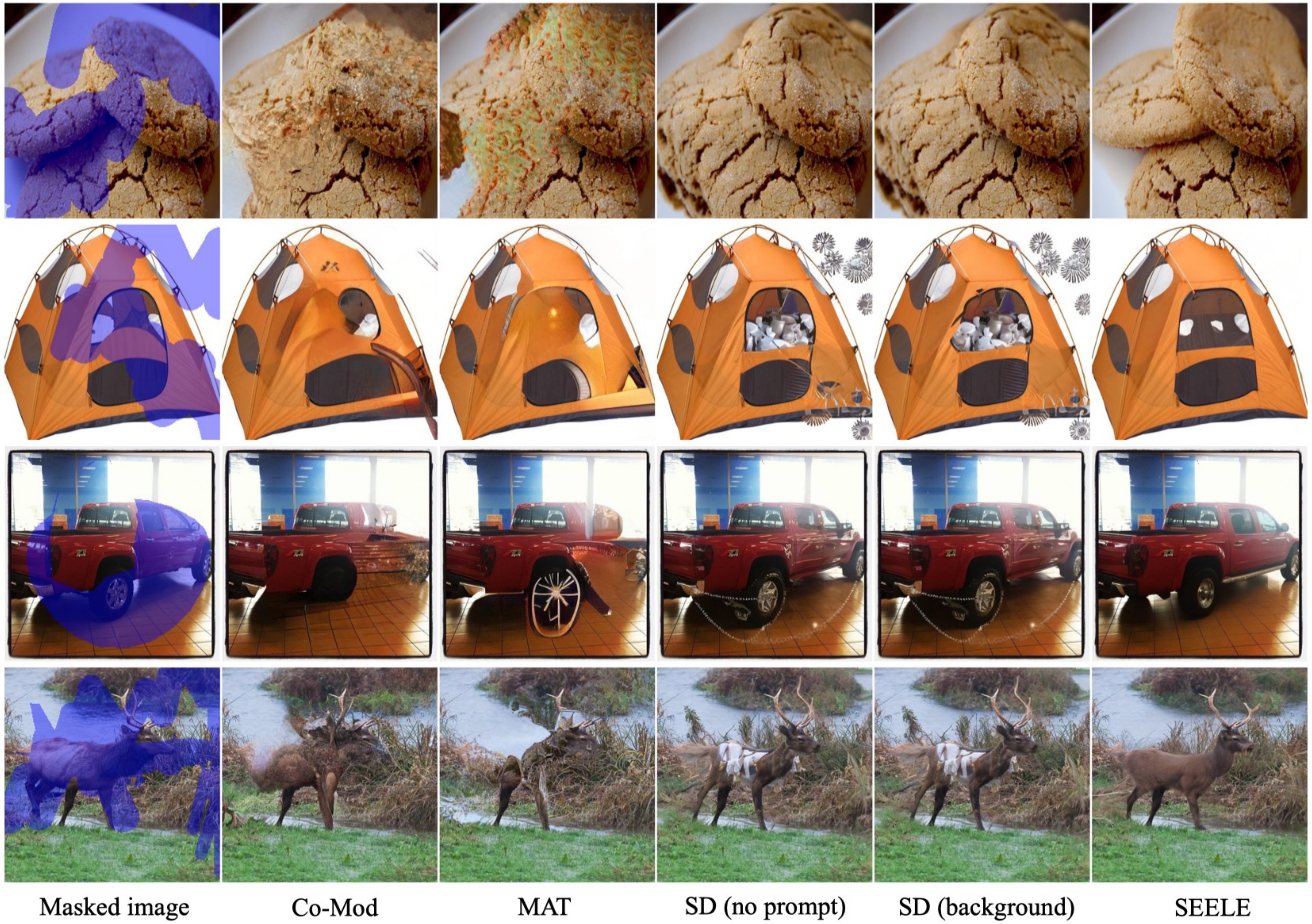
LoRA is used to perform subject harmonization

Yikai Wang et al. Repositioning the Subject within Image. Preprint, 2023.

Effectiveness of Task Inversion: Standard Inpainting

(a) Inpainting on Places2 [82].

Methods	PSNR↑	SSIM↑	FID↓	LPIPS↓
Co-Mod	21.09	0.84	30.04	0.17
MAT	20.68	0.84	32.44	0.16
SD("NA")	20.35	0.84	29.63	0.16
SD("bkg")	20.59	0.84	29.31	0.16
SEELE	**21.98**	**0.87**	**24.40**	**0.13**



Yikai Wang et al. Repositioning the Subject within Image. Preprint, 2023.

Effectiveness of Task Inversion: Standard Outpainting

(b) Outpainting on Flickr-Scenery [10].

Methods	SD("NA")	SD("bkg")	SEELE
PSNR↑	14.48	14.60	**16.00**
SSIM↑	0.69	0.70	**0.73**
FID↓	53.52	46.58	**29.06**
LPIPS↓	0.35	0.34	**0.31**

SD

SEELE

SD

SEELE

Yikai Wang et al. Repositioning the Subject within Image. Preprint, 2023.

Example of Subject Repositioning on 1k Images

Yikai Wang et al. Repositioning the Subject within Image. Preprint, 2023.

Example of Subject Repositioning

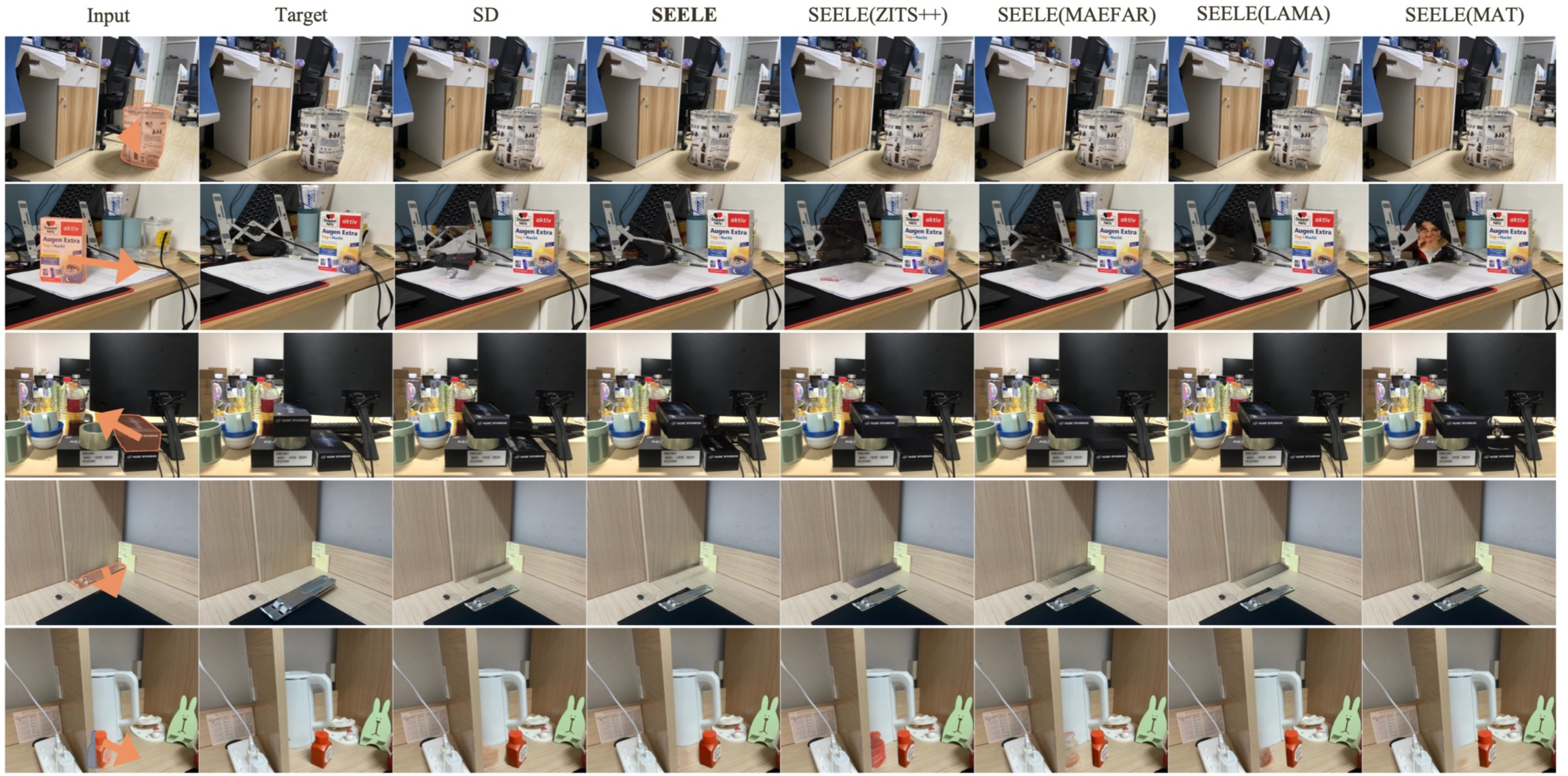


Table 1: Quantitative comparison and user-study on ReS. (○): SD; (*): SEELE; Quality: the fidelity of the results; Consist.: the consistency with surrounding area. SEELE consistently works better than SD variants.

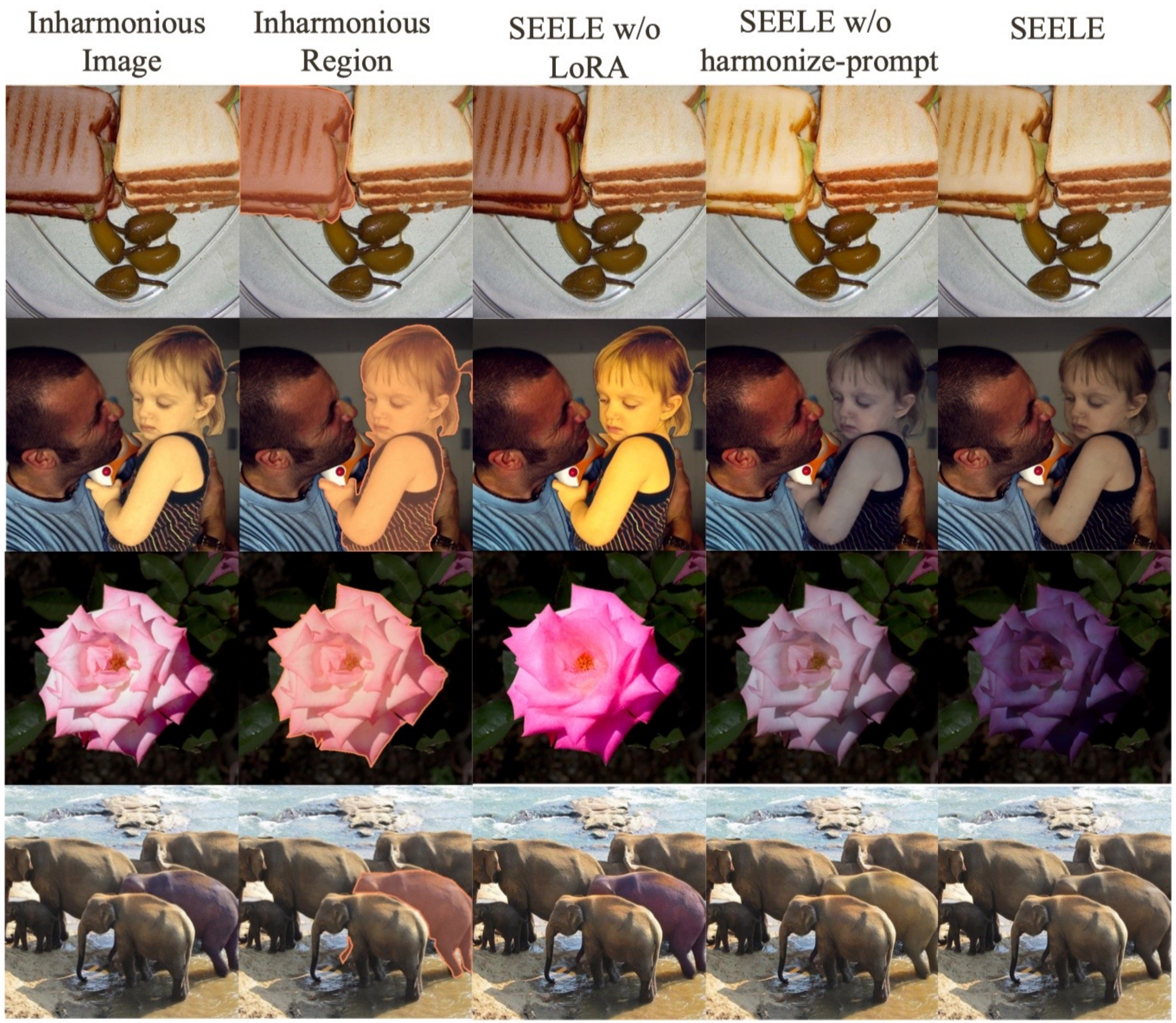
Model	$\circ_{no}$	$\circ_{simple}$	$\circ_{complex}$	$\circ_{lora}$	SEELE	$*_{ZITS++}$	$*_{MAE-FAR}$	$*_{LaMa}$	$*_{MAT}$
LPIPS(↓)	0.157	0.157	0.157	0.162	**0.156**	0.176	0.172	0.163	0.163
Quality(↑)	0.057	0.090	0.073	0.207	**0.290**	0.080	0.053	0.073	0.076
Consist.(↑)	0.054	0.057	0.050	0.036	**0.329**	0.089	0.114	0.168	0.104

Yikai Wang et al. Repositioning the Subject within Image. Preprint, 2023.

Different Prompt leads to Different Generation Direction

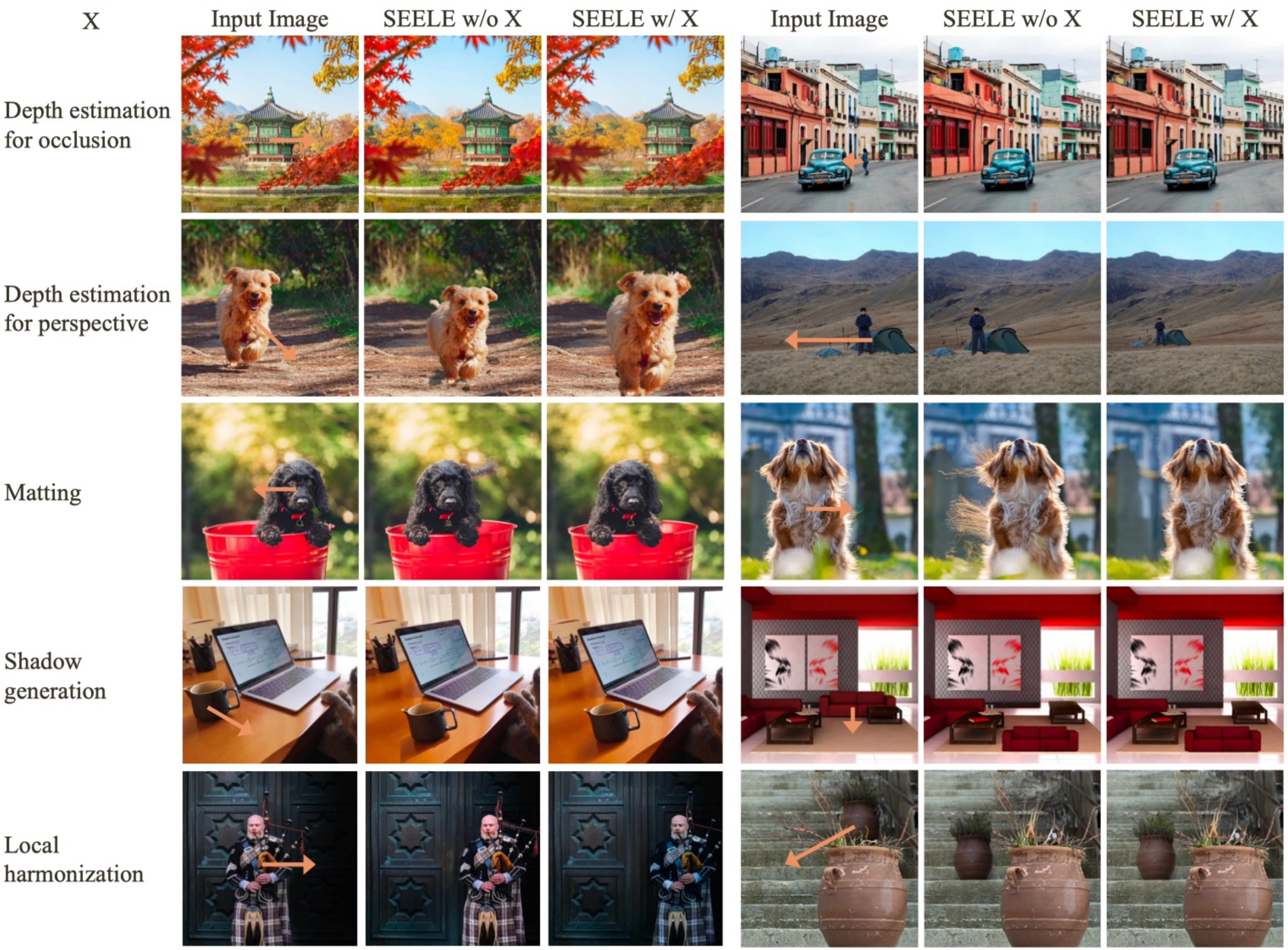
Yikai Wang et al. Repositioning the Subject within Image. Preprint, 2023.

Ablation of Local Harmonization Component



Yikai Wang et al. Repositioning the Subject within Image. Preprint, 2023.

Ablation of Other Modules



Yikai Wang et al. Repositioning the Subject within Image. Preprint, 2023.

Consistent Image Inpainting

context-stability and visual-consistency

1. Advancing Image Inpainting: Versatility
2. Advancing Image Inpainting: Consistency

Context-Stability V.S. Variety

Reconstruction

unmasked region → masked region

Masked Auto-Encoder

Generation

noise → image

↑

unmasked region

Stable Diffusion Inpainting Model

Pros:

Context-stable, no hallucination

Cons:

Averaged and blurred results, low-fidelity.

Pros:

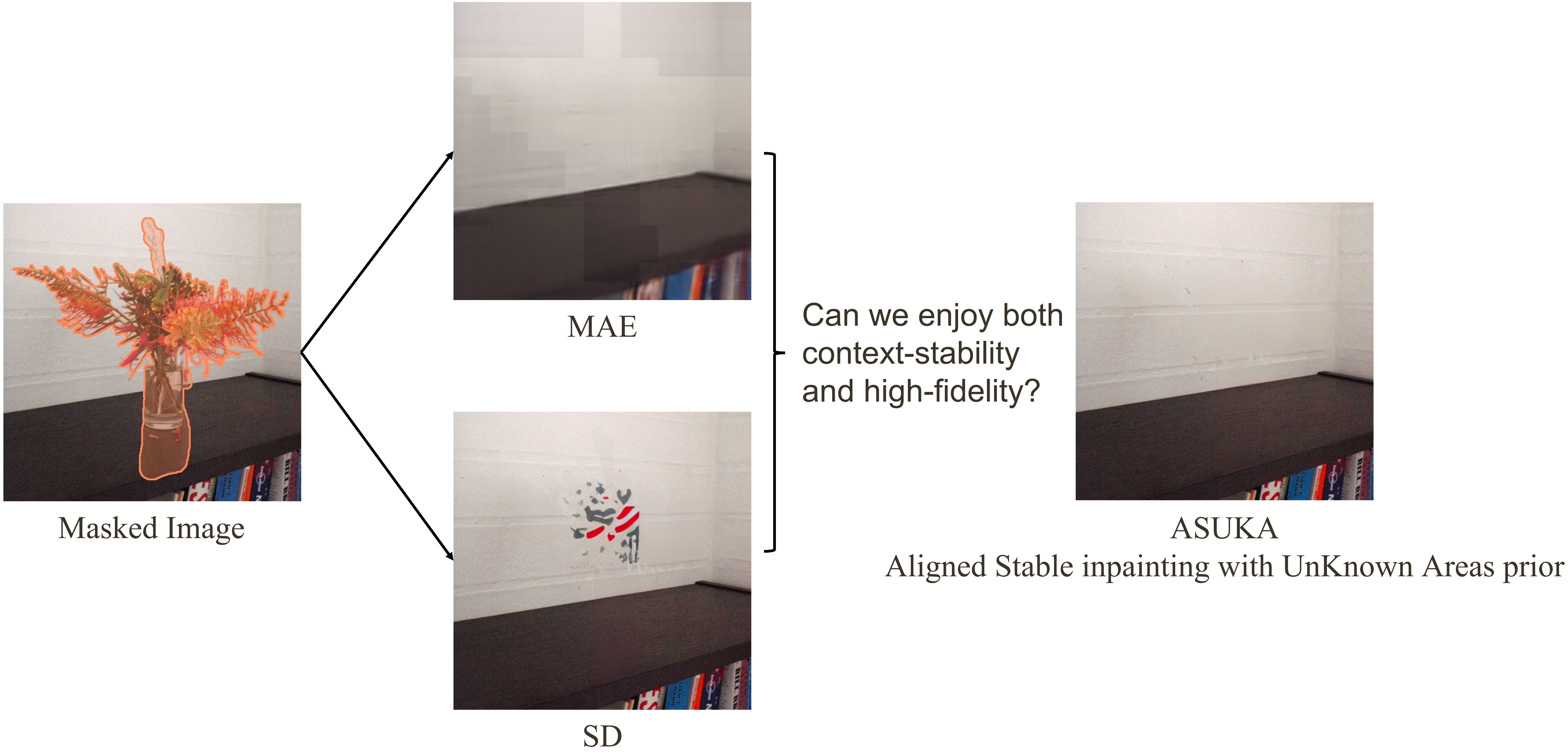
High-fidelity, high-variety

Cons:

Usually generate random elements.

Yikai Wang et al. Towards Context-Stable and Visual-Consistent Image Inpainting. Preprint, 2023.

Context-Stable Inpainting



Yikai Wang et al. Towards Context-Stable and Visual-Consistent Image Inpainting. Preprint, 2023.

How to align MAE with SD?

In an image-to-image translation manner?

Input Image with mask

MAE

Add noise and then denoise

SD with MAE initial latent

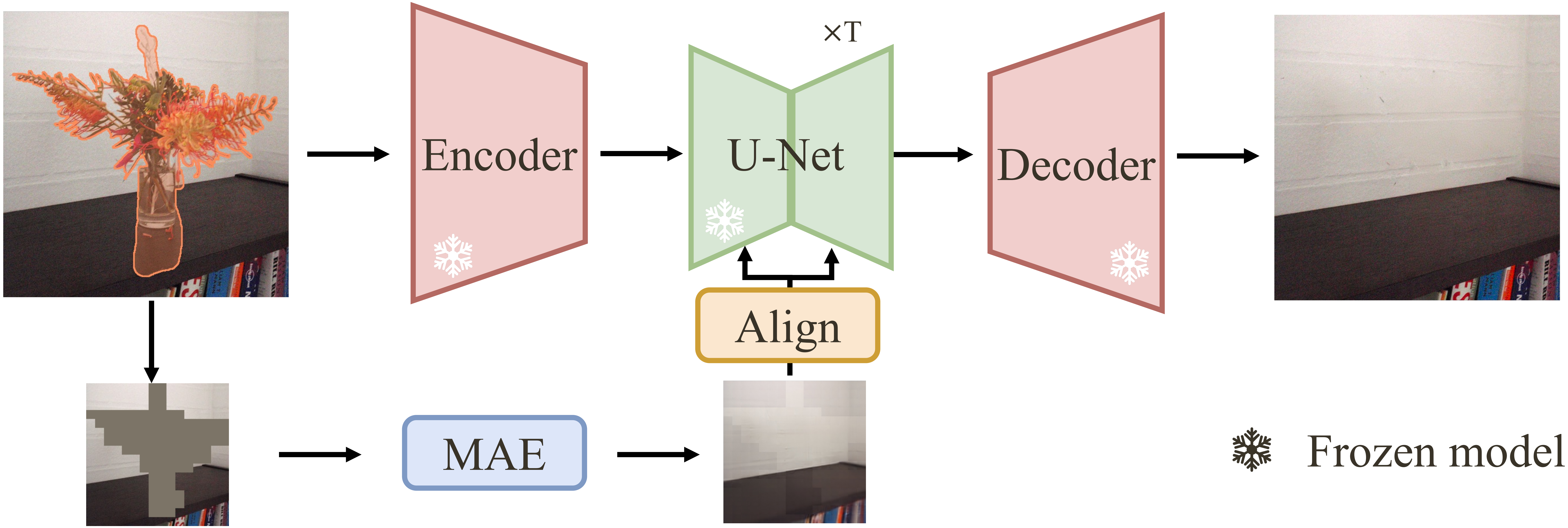
ASUKA

Blurring initial latent leads to blurring generation result.

Use MAE result as condition to selectively guide the generation of SD.

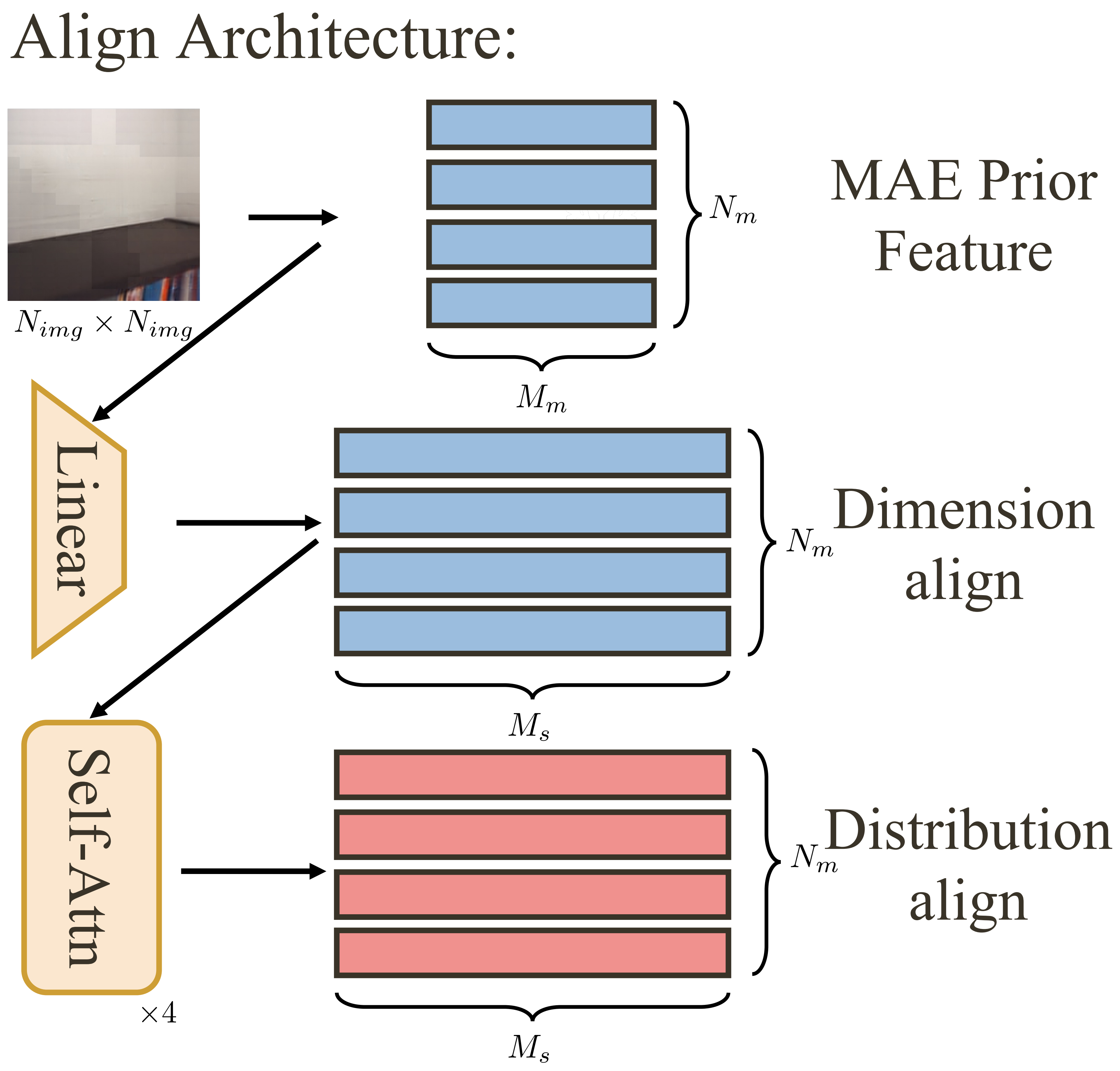
Yikai Wang et al. Towards Context-Stable and Visual-Consistent Image Inpainting. Preprint, 2023.

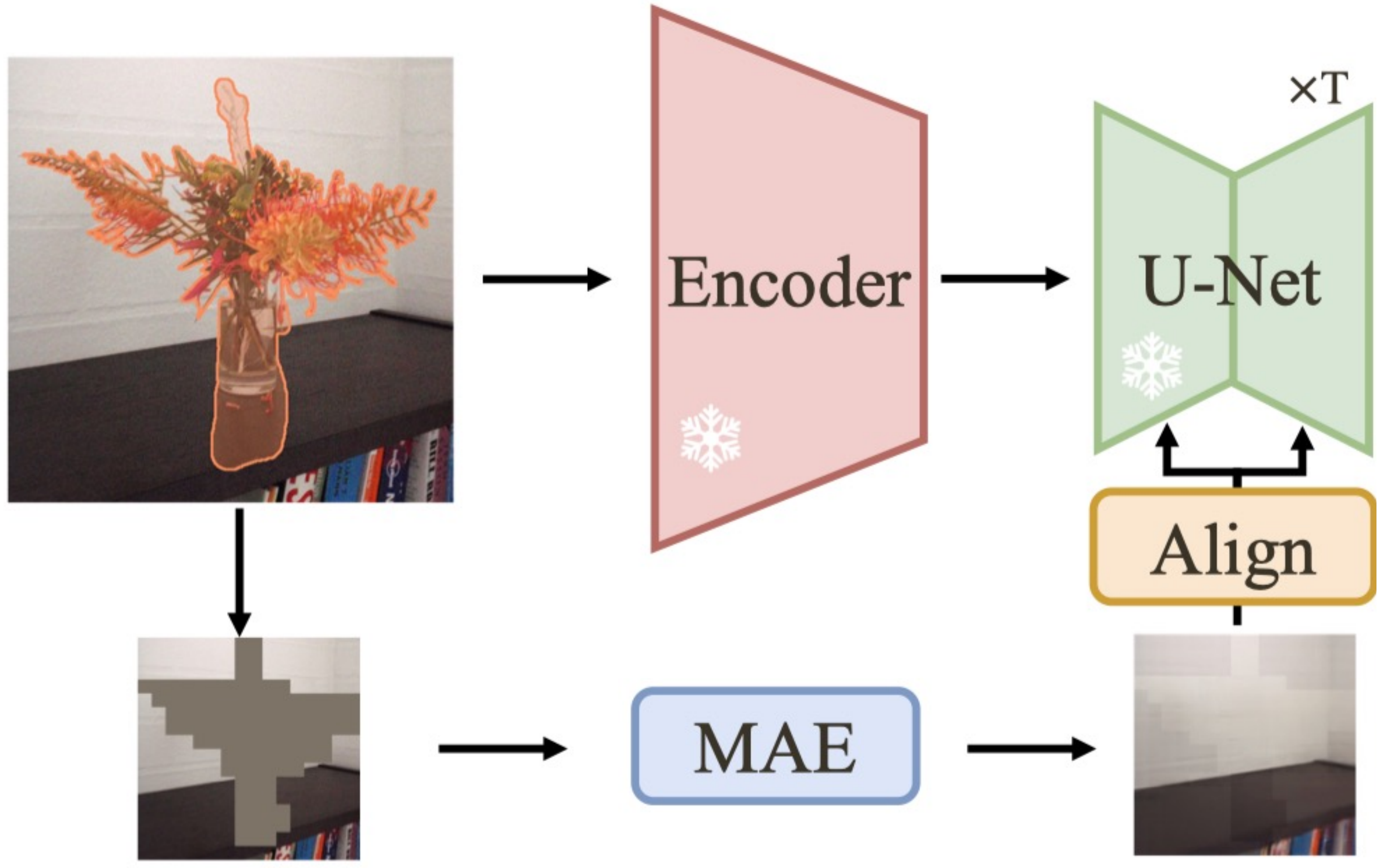
Stable Diffusion Inpainting Model with MAE Condition



Yikai Wang et al. Towards Context-Stable and Visual-Consistent Image Inpainting. Preprint, 2023.

Context-Stable Inpainting: Technique Details





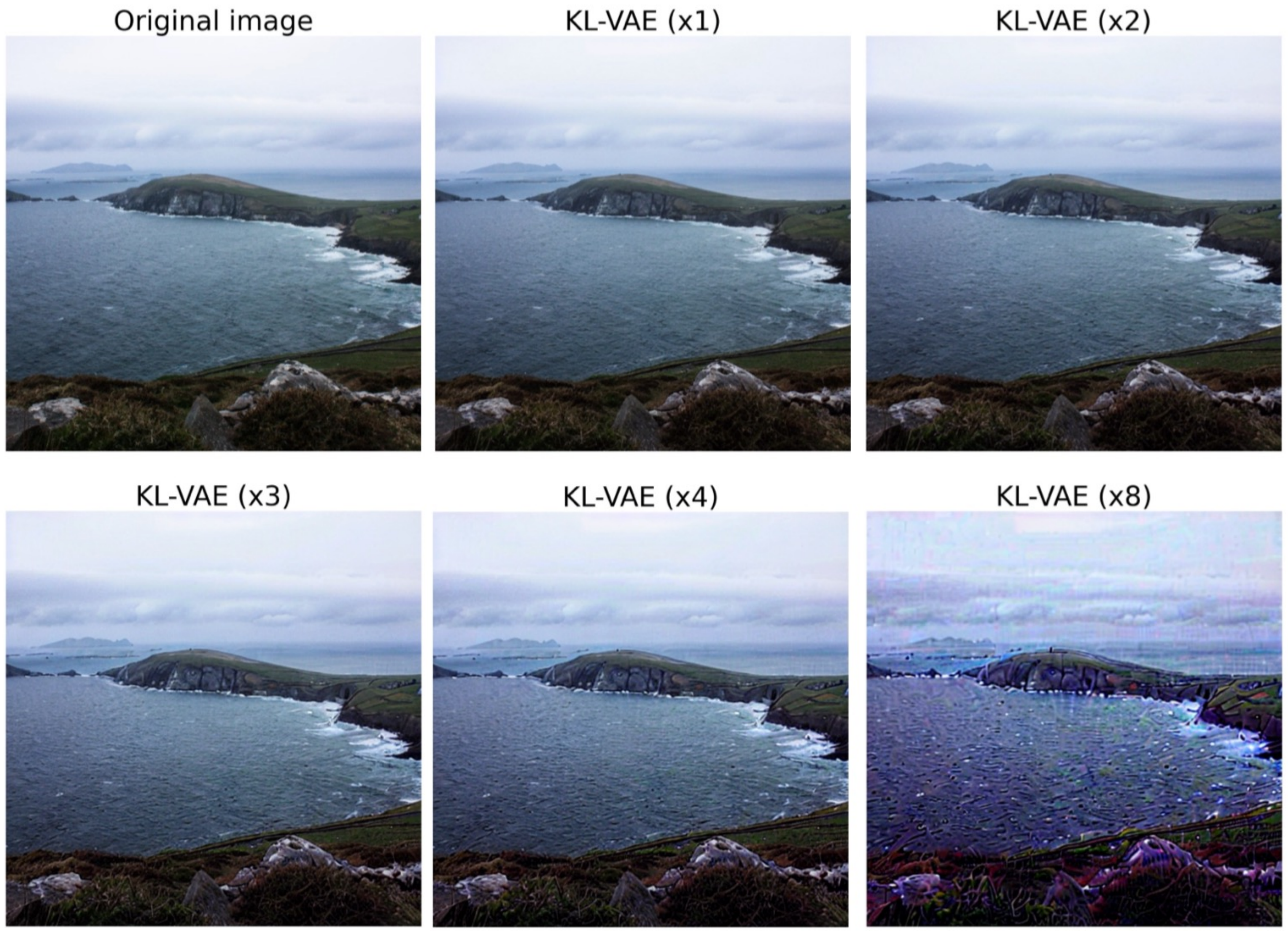
Remark:
The input to SD is 256x768, instead of 77x768 (text feature sequence) to preserve local guidance.

Separate Training:
- MAE is fine-tuned to handle continuous masks.
- Alignment module is trained with standard diffusion objective.

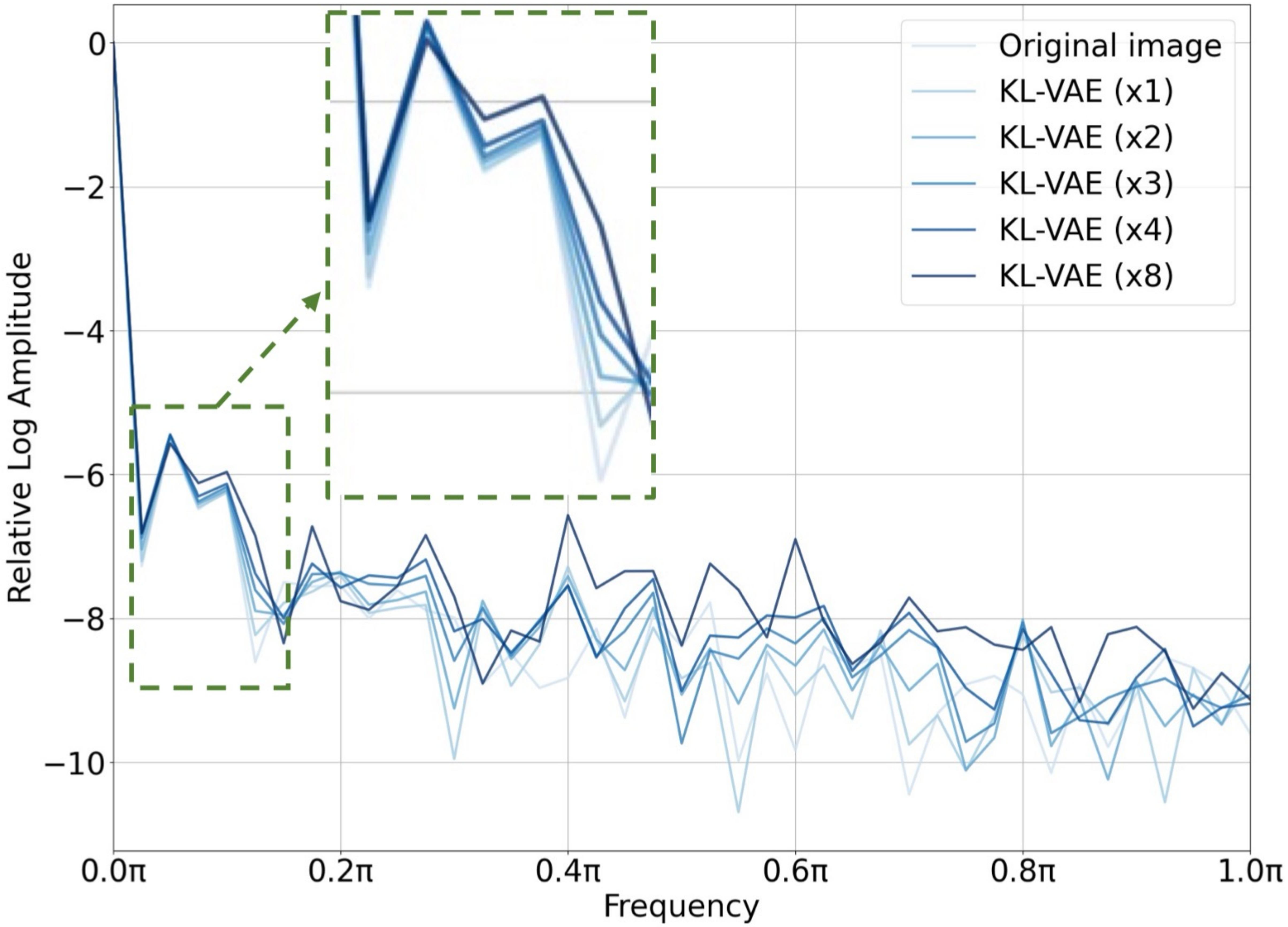
Visual-Inconsistency Issue (of SD)

Yikai Wang et al. Towards Context-Stable and Visual-Consistent Image Inpainting. Preprint, 2023.

Information Loss of VAE used in SD



(a) Images decoded by KL-VAE repeatedly for different times

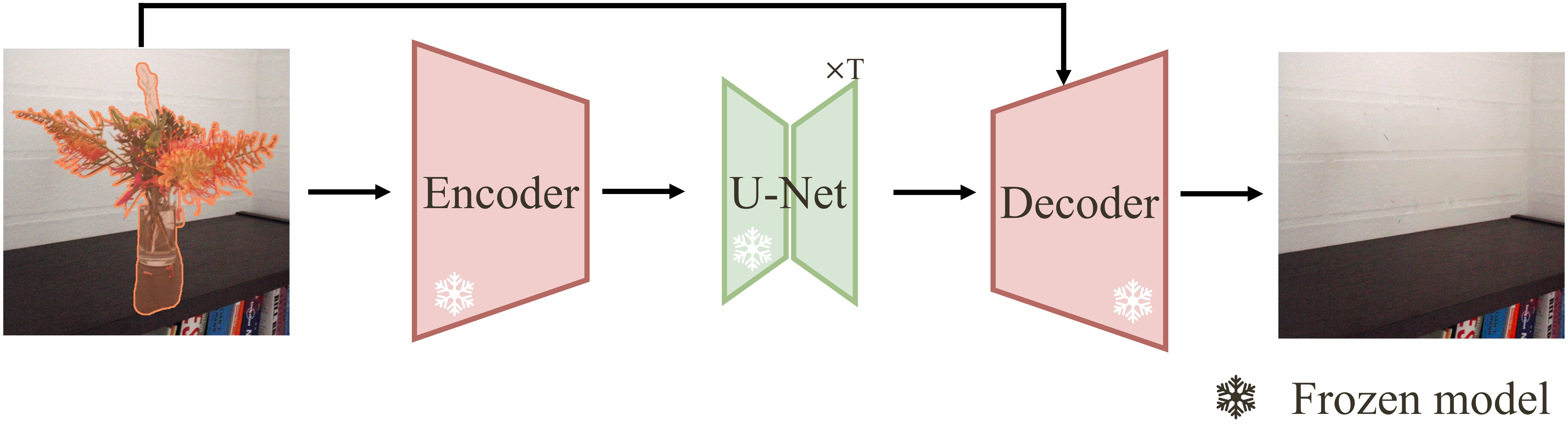


(b) Relative log amplitude (y-axis) and frequency (x-axis)

Yikai Wang et al. Towards Context-Stable and Visual-Consistent Image Inpainting. Preprint, 2023.

Alleviate the Information Loss

- Ideal way: Train a better VAE to solve the information loss.
  Re-train the VAE → Different latent space → Need to re-train the U-Net → Train another SD. Bad.
- Efficient way: Train a better VAE decoder to solve the information loss during decoding.
  Preserve the latent space → No need to re-train the U-Net. Good.
- How to train a better decoder?
  Utilize the ground-truth pixel value of unmasked region.



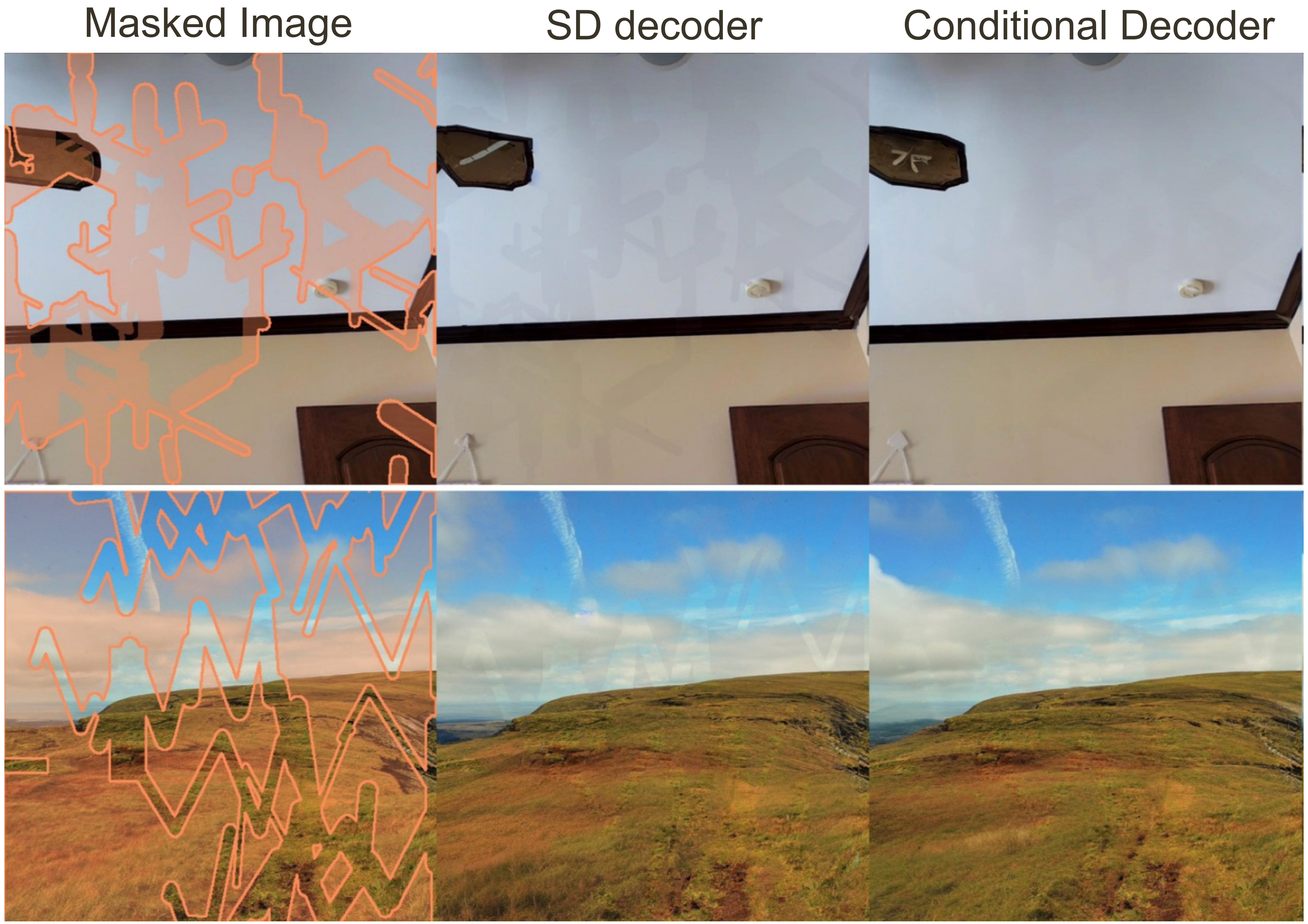
Yikai Wang et al. Towards Context-Stable and Visual-Consistent Image Inpainting. Preprint, 2023.

Better Decoder?

Yes, but not good enough.

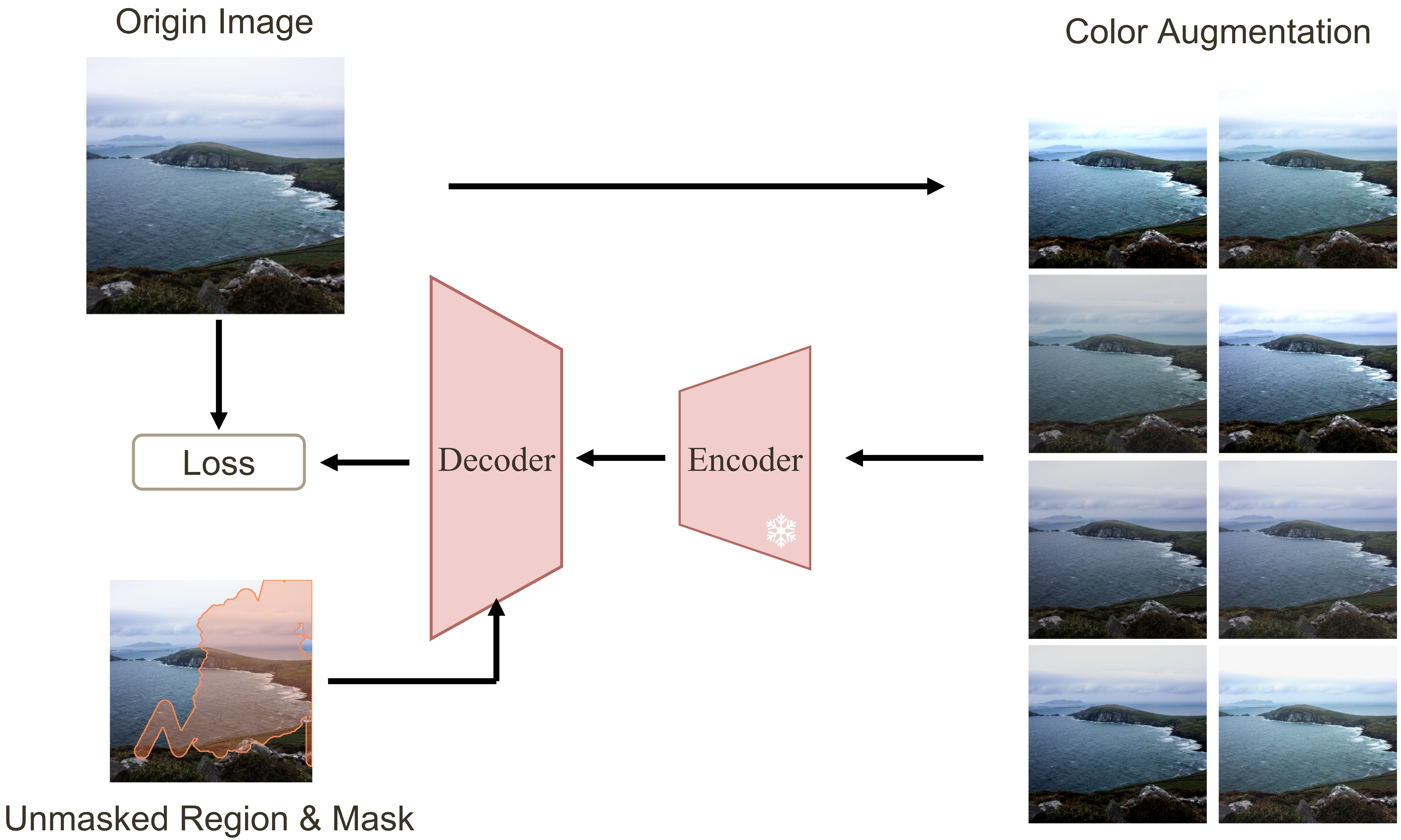
It already knows the knowledge to use, but it is not trained to use.

We need to train the decoder to reduce color shift.



Yikai Wang et al. Towards Context-Stable and Visual-Consistent Image Inpainting. Preprint, 2023.

Harmonizing while Decoding: Color Augmentation



Yikai Wang et al. Towards Context-Stable and Visual-Consistent Image Inpainting. Preprint, 2023.

Better Decoder Now?

Masked Image | SD Decoder | Conditional Decoder | Augmented Decoder

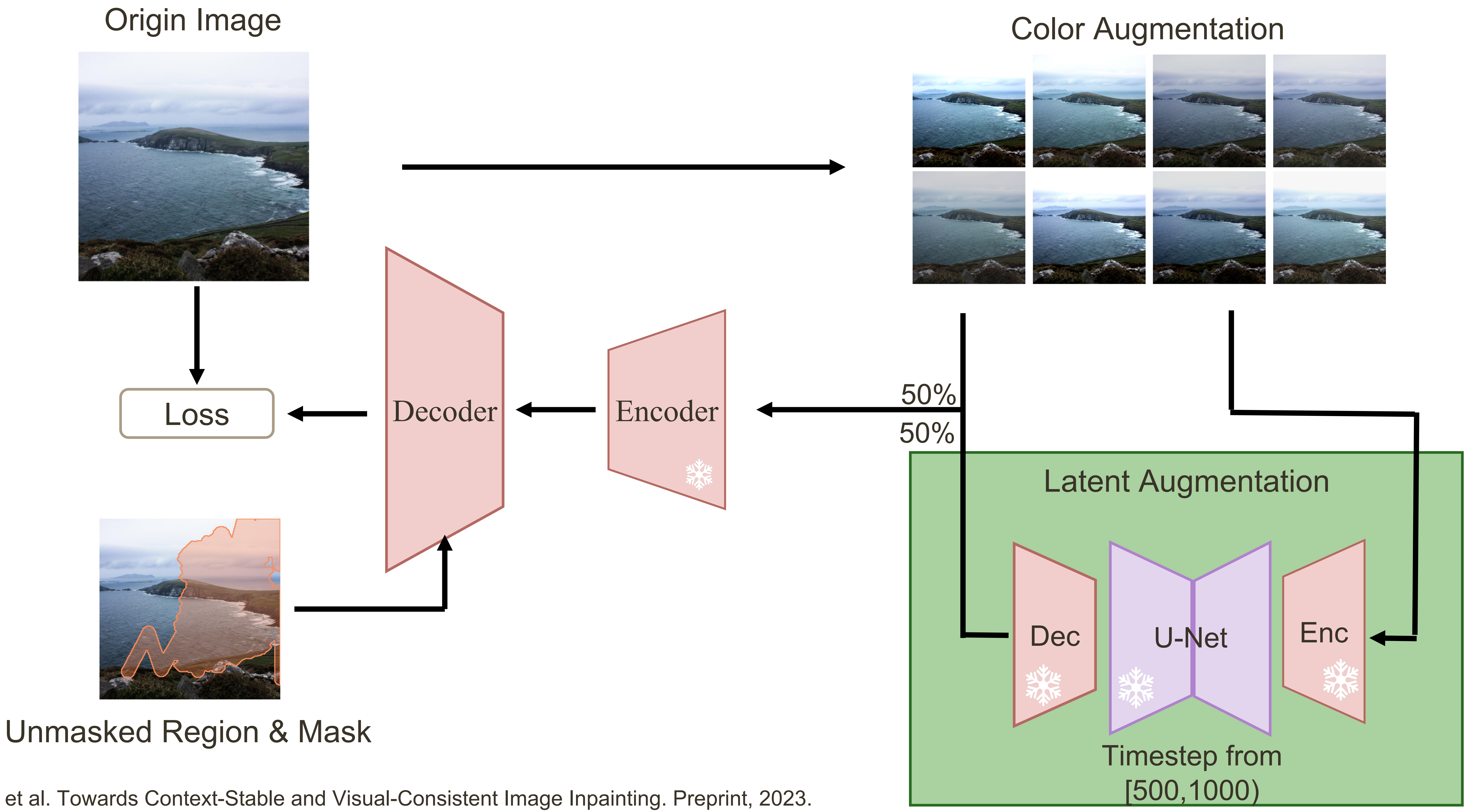
Yes, but not in all cases.

Masked Image | Augmented Decoder

Information loss also exists in the U-Net.

Yikai Wang et al. Towards Context-Stable and Visual-Consistent Image Inpainting. Preprint, 2023.

Harmonizing while Decoding: Latent Augmentation



Yikai Wang et al. Towards Context-Stable and Visual-Consistent Image Inpainting. Preprint, 2023.

Better Decoder Now?

Masked Image | w/o latent aug | w/ latent aug

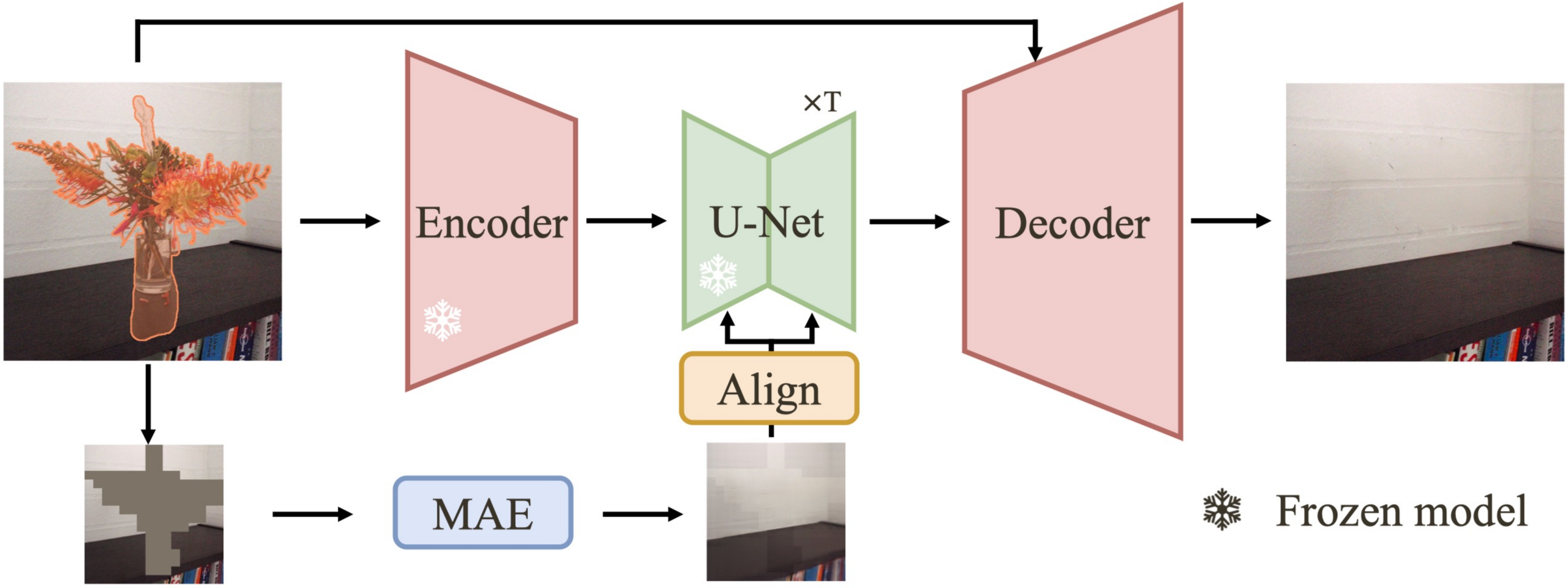
Yes!

(d) Comparison of different decoders for SD. VAE [39] is the original decoder used by SD; + cond. [62] is the decoder conditioned on unmasked image; + color uses the color augmentation strategy to perform local harmonization task; Ours further combines latent augmentation strategy to handle the gap between generated latent and real latent.

Decoder	LPIPS↓	FID↓	U-IDS↑	P-IDS↑
VAE	0.156	11.949	0.343	0.208
+ cond.	0.151	11.634	0.361	0.231
+ color	0.152	11.603	0.357	0.229
Ours	**0.150**	**11.460**	**0.368**	**0.256**

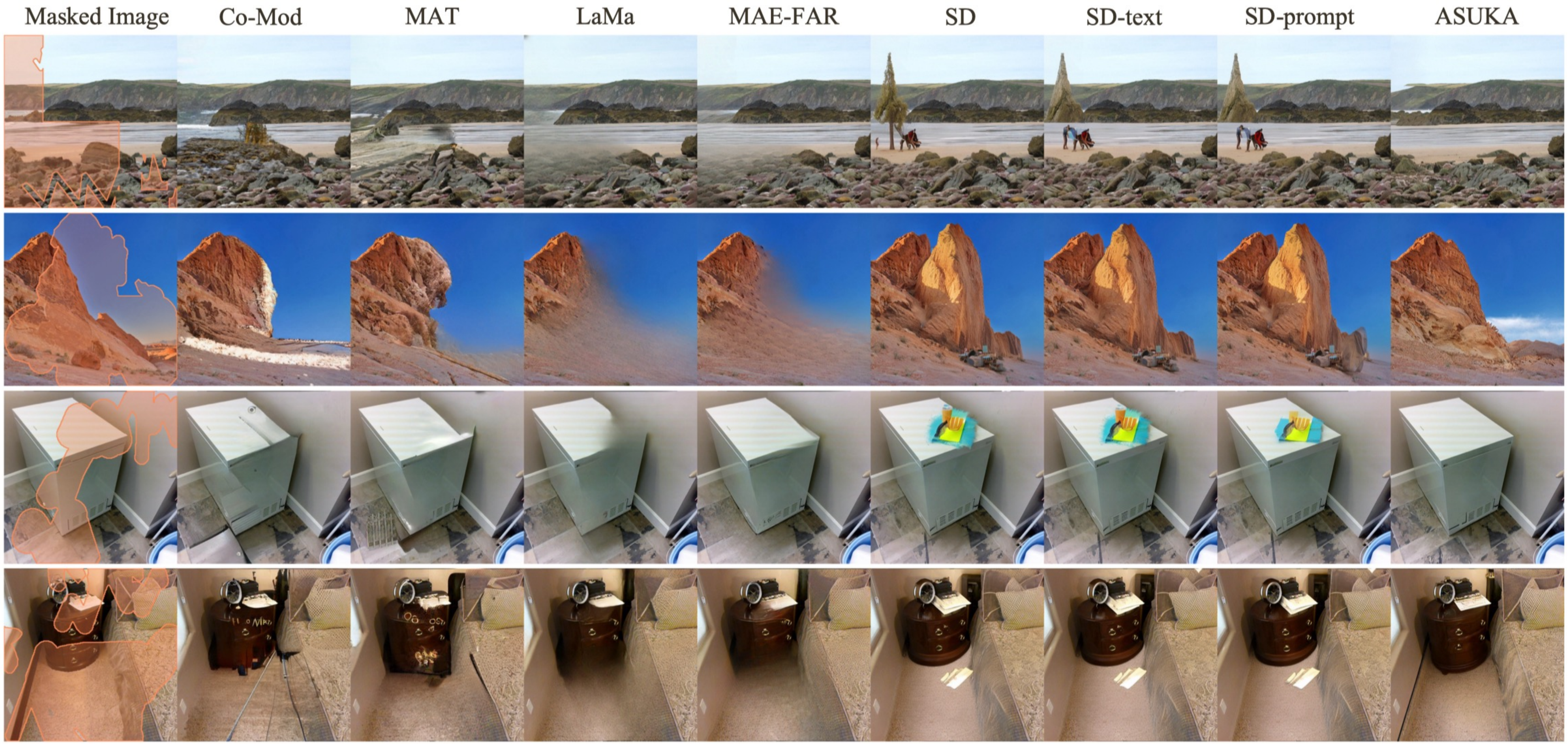
Yikai Wang et al. Towards Context-Stable and Visual-Consistent Image Inpainting. Preprint, 2023.

Aligned Stable Inpainting with UnKnown Areas Prior



Yikai Wang et al. Towards Context-Stable and Visual-Consistent Image Inpainting. Preprint, 2023.

Comparison



Yikai Wang et al. Towards Context-Stable and Visual-Consistent Image Inpainting. Preprint, 2023.

Comparison (cont.)

Yikai Wang et al. Towards Context-Stable and Visual-Consistent Image Inpainting. Preprint, 2023.

Quantitative Comparison

Table 1: Quantitative comparison on MISATO and Places 2. Co-Mod [59], MAT [25], LaMa [45], MAE-FAR [7] and SD-Repaint [32] are state-of-the-art inpainting methods. SD [39] performs unconditional generation. SD-text uses "background" text prompt to guide generation. SD-prompt uses learnable prompts trained specifically for inpainting, using the same training setting as ASUKA, performing prompt-guided generation. ASUKA and SD variants use the stable diffusion text-guided inpainting model v1.5.

Dataset	MISATO (2k images)				Places 2 (36.5k images)			
Method	LPIPS↓	FID↓	U-IDS↑	P-IDS↑	LPIPS↓	FID↓	U-IDS↑	P-IDS↑
Co-Mod [59]	0.179	17.421	0.243	0.109	0.267	5.794	0.274	0.096
MAT [25]	0.176	17.261	0.255	0.122	0.202	3.765	0.348	0.195
LaMa [45]	0.155	15.436	0.260	0.135	0.202	6.693	0.247	0.050
MAE-FAR [7]	**0.142**	13.283	0.282	0.153	**0.174**	3.559	0.307	0.105
SD [39]	0.168	12.812	0.345	0.211	0.193	1.514	0.375	0.207
SD-text	0.164	12.603	0.337	0.207	0.191	1.506	0.373	0.202
SD-prompt	0.160	12.517	0.331	0.204	0.189	1.477	0.390	0.234
SD-Repaint [32]	0.227	27.861	0.016	0.007	0.251	12.466	0.217	0.045
ASUKA	0.150	**11.460**	**0.368**	**0.256**	0.183	**1.230**	**0.413**	**0.287**

Yikai Wang et al. Towards Context-Stable and Visual-Consistent Image Inpainting. Preprint, 2023.

Ablation Studies

(a) Comparison of ASUKA using pre-trained (p.t.) MAE v.s. fine-tuned (f.t.) MAE.

MAE	LPIPS↓	FID↓	U-IDS↑	P-IDS↑
p.t.	0.151	11.513	0.354	**0.258**
f.t.	**0.150**	**11.460**	**0.368**	0.256

(b) Ablation of different alignment modules. *Linear* adopts linear layer; *attn* adopts a self-attention layer; *cross x4* adopts 4 cross-attention layers; ASUKA adopts 4 self-attention layers.

Align	LPIPS↓	FID↓	U-IDS↑	P-IDS↑
linear	0.155	11.934	0.361	0.227
attn	0.152	11.613	0.362	0.234
cross x4	0.152	11.762	**0.368**	0.238
ASUKA	**0.150**	**11.460**	**0.368**	**0.256**

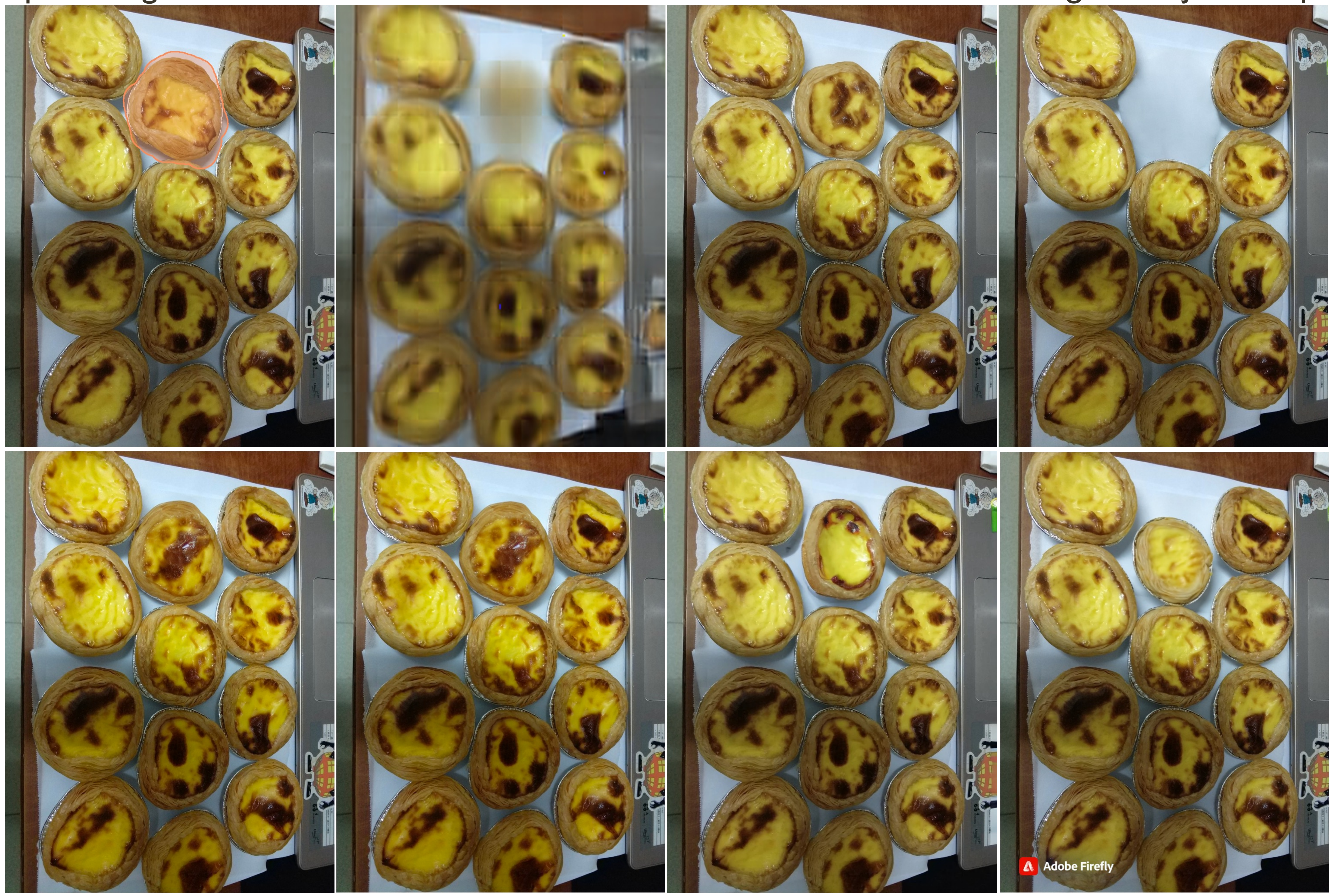
(c) User-study of top-1 ratio among all the inpainting results. Context-stability (C.S.) measures the coherence between masked region and unmasked surroundings, with a preference of not generating new elements; Visual-consistency (V.C.) measures the color consistency.

Model	C.S.(%)	V.C.(%)
Co-Mod [59]	3.98	4.98
MAT [25]	7.40	3.20
LaMa [45]	8.18	8.28
MAE-FAR [7]	4.88	5.60
SD [39]	10.58	5.75
SD-text	7.70	15.83
SD-prompt	16.18	15.78
SD-Repaint [32]	1.60	0.55
ASUKA	**39.43**	**40.05**

Yikai Wang et al. Towards Context-Stable and Visual-Consistent Image Inpainting. Preprint, 2023.

Limitation: The "curse" of Self-Attention

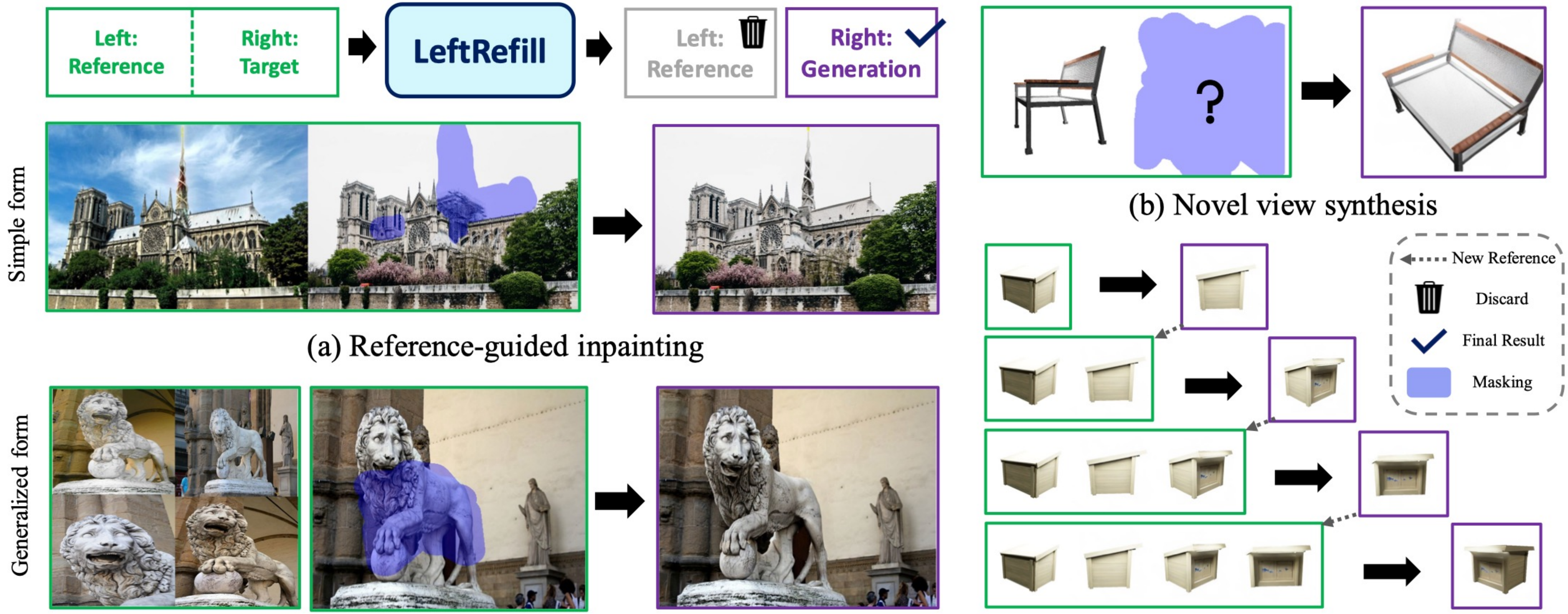
Input Image with mask | MAE | ASUKA | ASUKA guide by blank paper



SD “background” cfg=9 | SD “blank paper” cfg=9 | DALL-E 2 | FireFly

Yikai Wang et al. Towards Context-Stable and Visual-Consistent Image Inpainting. Preprint, 2023.

Use Cross-Attention Layers As Self-Attention Layers



(a) Reference-guided inpainting

(b) Novel view synthesis

(c) Inpainting one target view through multiple references

(d) Generating multiple targets from single view

Chenjie Cao, Yunuo Cai, Qiaole Dong, **Yikai Wang**, Yanwei Fu. LeftRefill: Filling Right Canvas based on Left Reference through Generalized Text-to-Image Diffusion Model. CVPR 2024.

Summary

- We delve into the advanced text-to-image stable diffusion inpainting model.
- We explore its "emergent property", which allows various non-textual guidance to be used as conditions for a wide range of image inpainting tasks, and combine these capabilities to address the challenging task of subject repositioning.
- We analyze two common issues found in popular generative image inpainting models, highlighting the importance of maintaining context stability and visual consistency throughout the inpainting process.
- Using the stable diffusion model as a case study, we illustrate how enhancing these consistencies can significantly improve its performance in general image inpainting tasks.

Yikai Wang et al. Repositioning the Subject within Image. Preprint, 2023.
Yikai Wang et al. Towards Context-Stable and Visual-Consistent Image Inpainting. Preprint, 2023.

THANKS