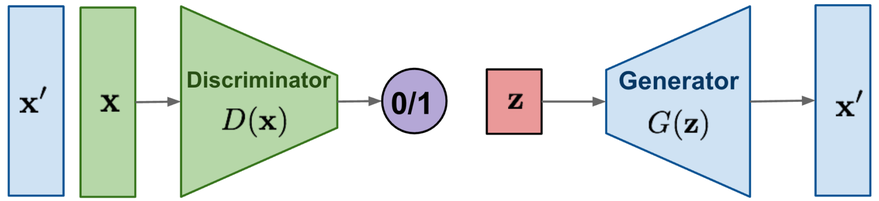
Structured Generative Vision:

From Context-Stable Inpainting to Next Visual Granularity Generation

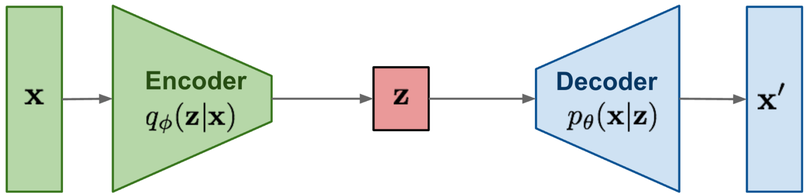
Yikai Wang

Generative Models

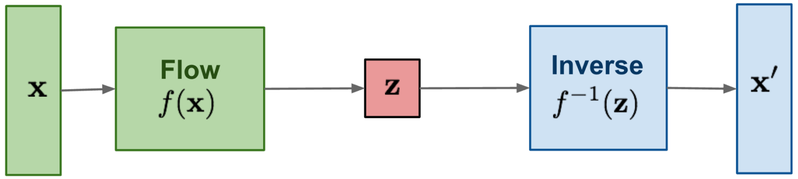
GAN: Adversarial training



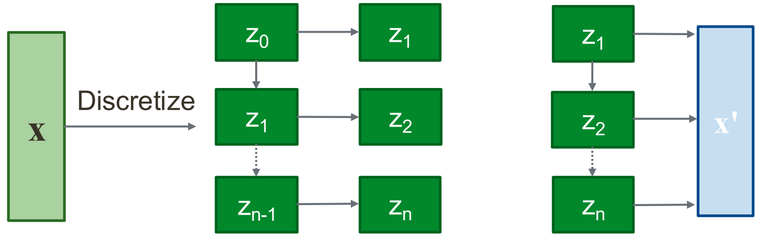
VAE: maximize variational lower bound



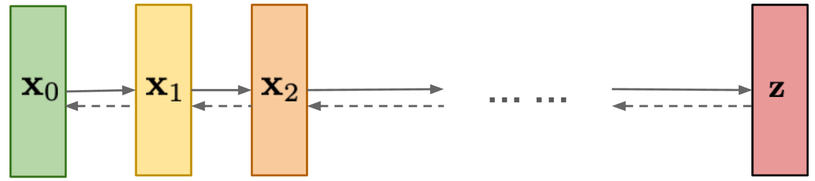
Flow-based models: Invertible transform of distributions



Transformers: Discretization and Autoregressive generation



Diffusion models: Gradually add Gaussian noise and then reverse



Motivation: Model the inherent relationship within images or between images and 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/

Image Inpainting: Task Definition

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

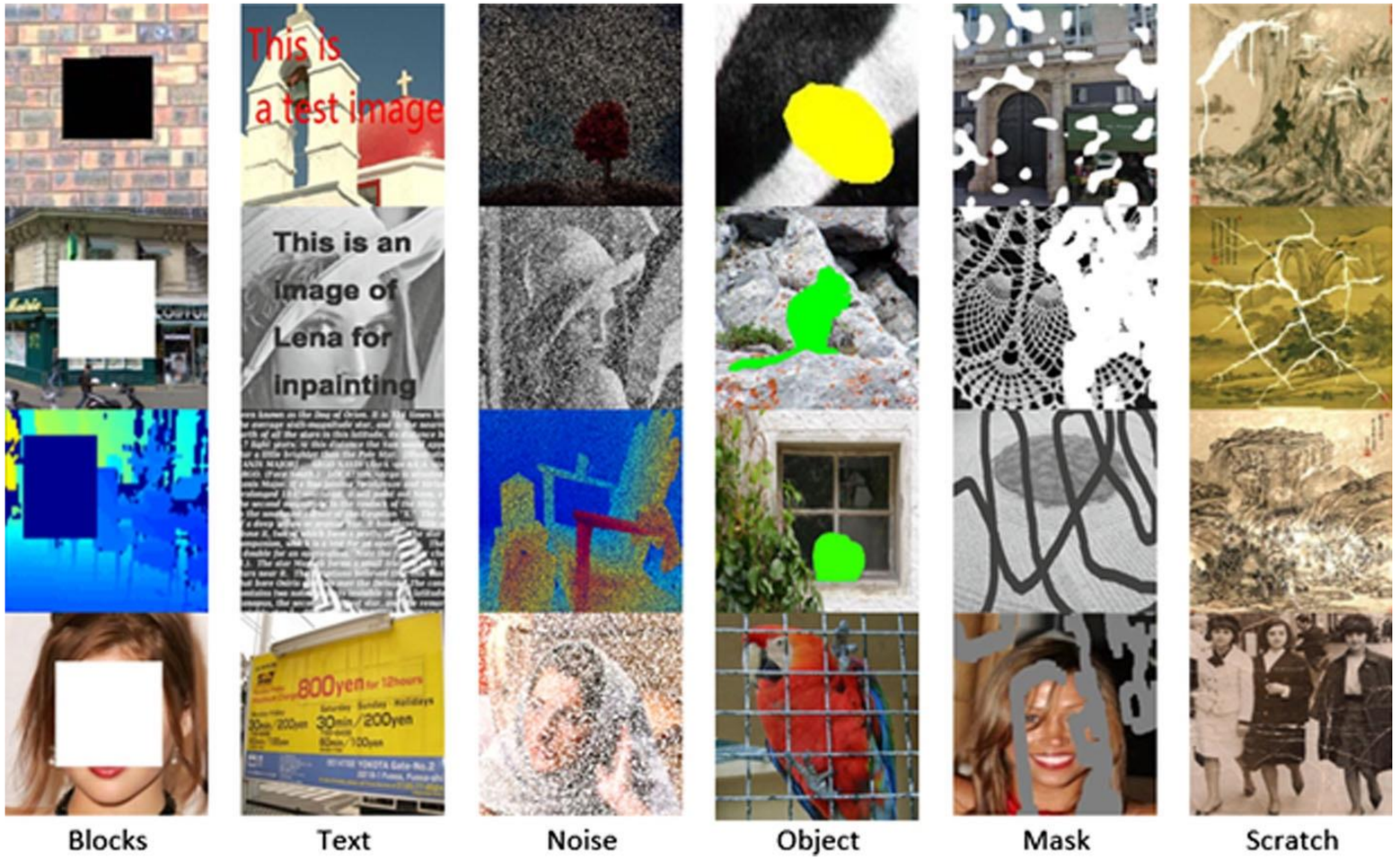


Image depicted from Elharrouss Omal, et al. " Image inpainting: A review. " Neural Processing Letters, 2020.

Conditional Image Inpainting

Input:
Masked Image
& Mask

Generative Model

Output:
Inpainted Image

Conditions

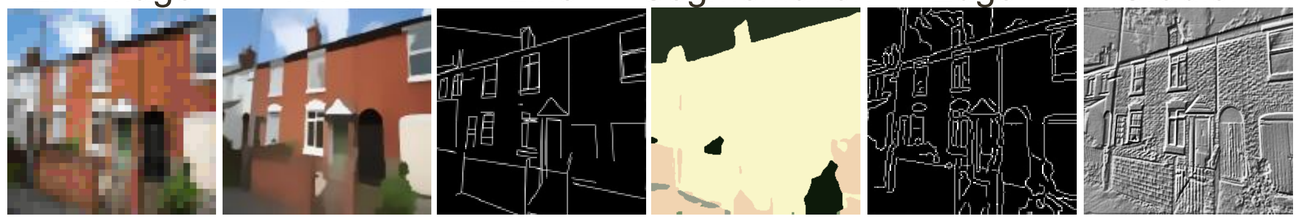
Text Description

A red brick house with a green door

Class

Building

LR-Image RTV Line Segmentation Edge Gradient



High-level (Semantic) Guidance

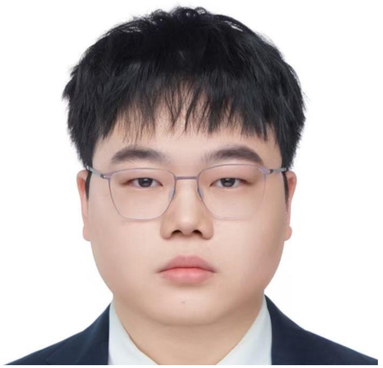
Low-level (Structure) Guidance

Highlight

Towards Enhanced Image Inpainting:

Mitigating Unwanted Object Insertion and Preserving Color Consistency






Yikai Wang[12*] Chenjie Cao[134*] 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

Context-Stability V.S. Variety

Reconstruction

unmasked region → masked region

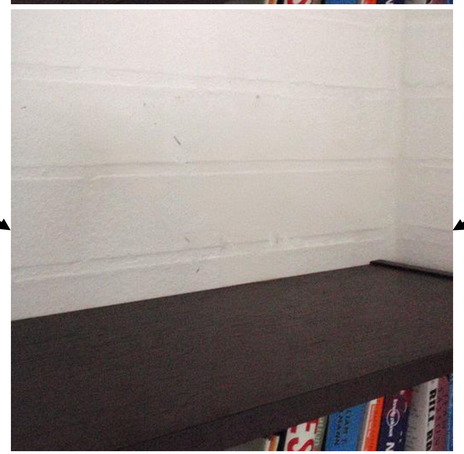
Masked Auto-Encoder

Pros:

Context-stable, no hallucination

Cons:

Averaged and blurred results, low-fidelity.

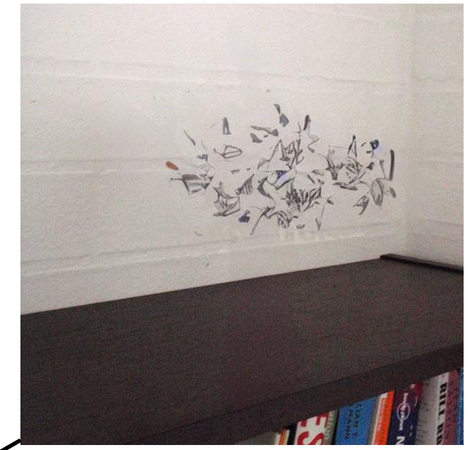


Generation

unmasked region

noise → image

Stable Diffusion Inpainting Model



Pros:

High-fidelity, high-variety

Cons:

Usually generate random elements.

How to align MAE with SD?

In an image-to-image translation manner?

Input Image with mask

MAE

Add noise and then denoise

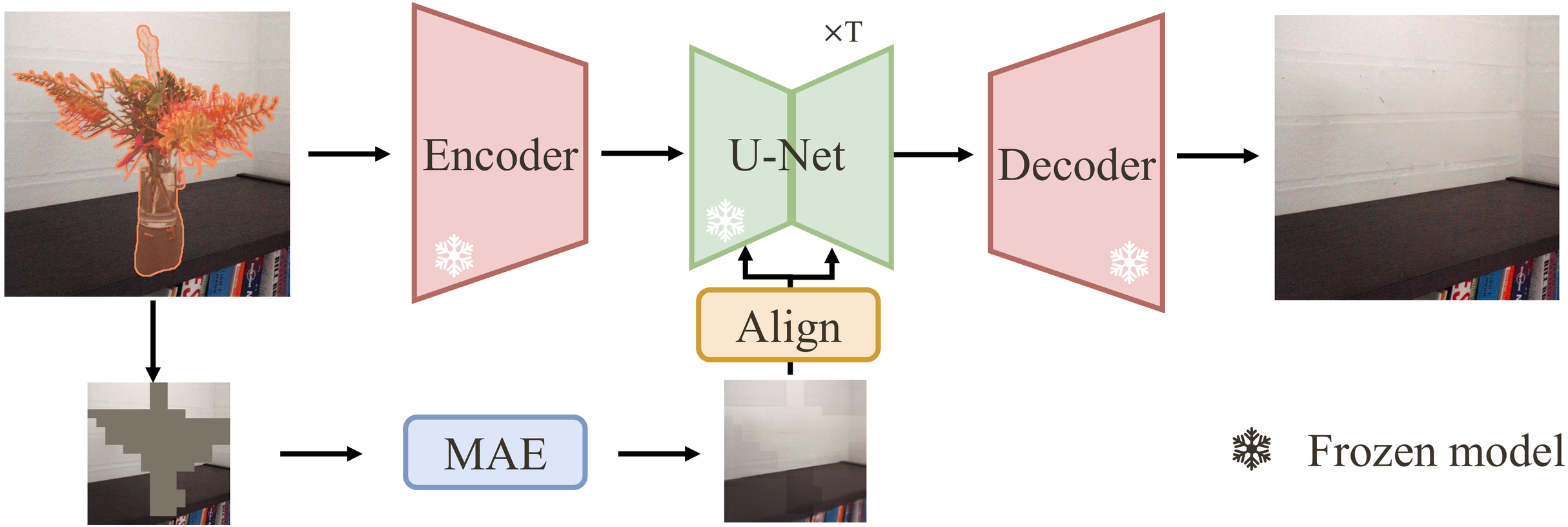
SD with MAE initial latent

Blurring initial latent leads to blurring generation result.

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

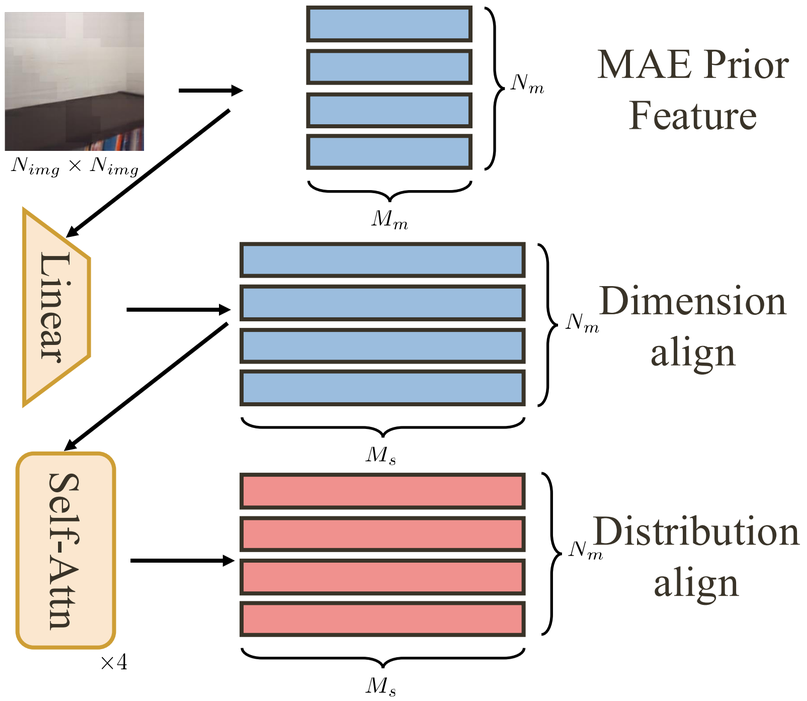
ASUKA

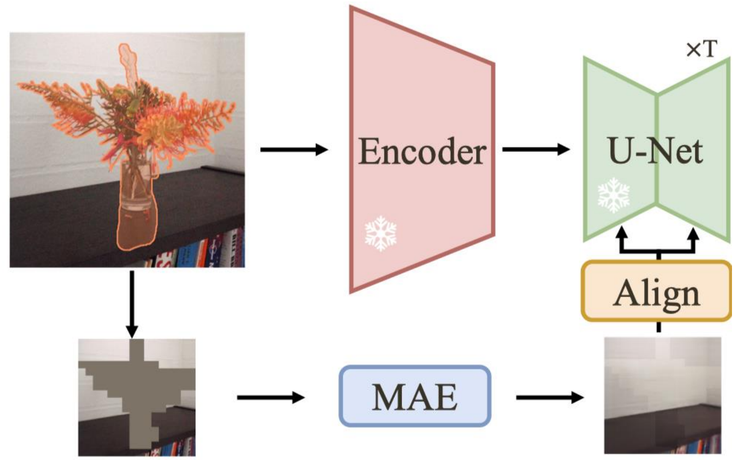
Stable Diffusion Inpainting Model with MAE Condition



Context-Stable Inpainting: Technique Details

Align Architecture:





(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**

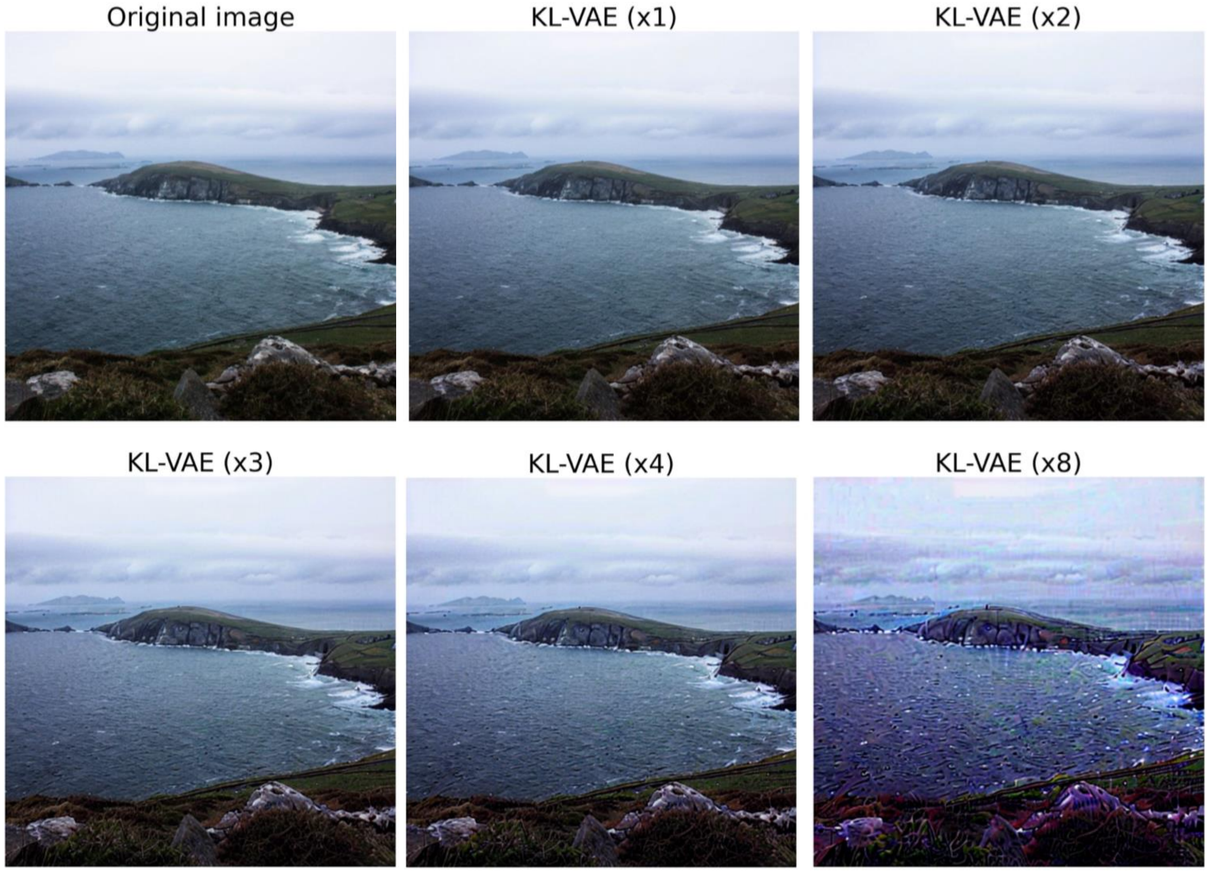
Remark:
The input to SD is 256x768, instead of 77x768 (text feature sequence) to preserve local guidance.

Separate Training:

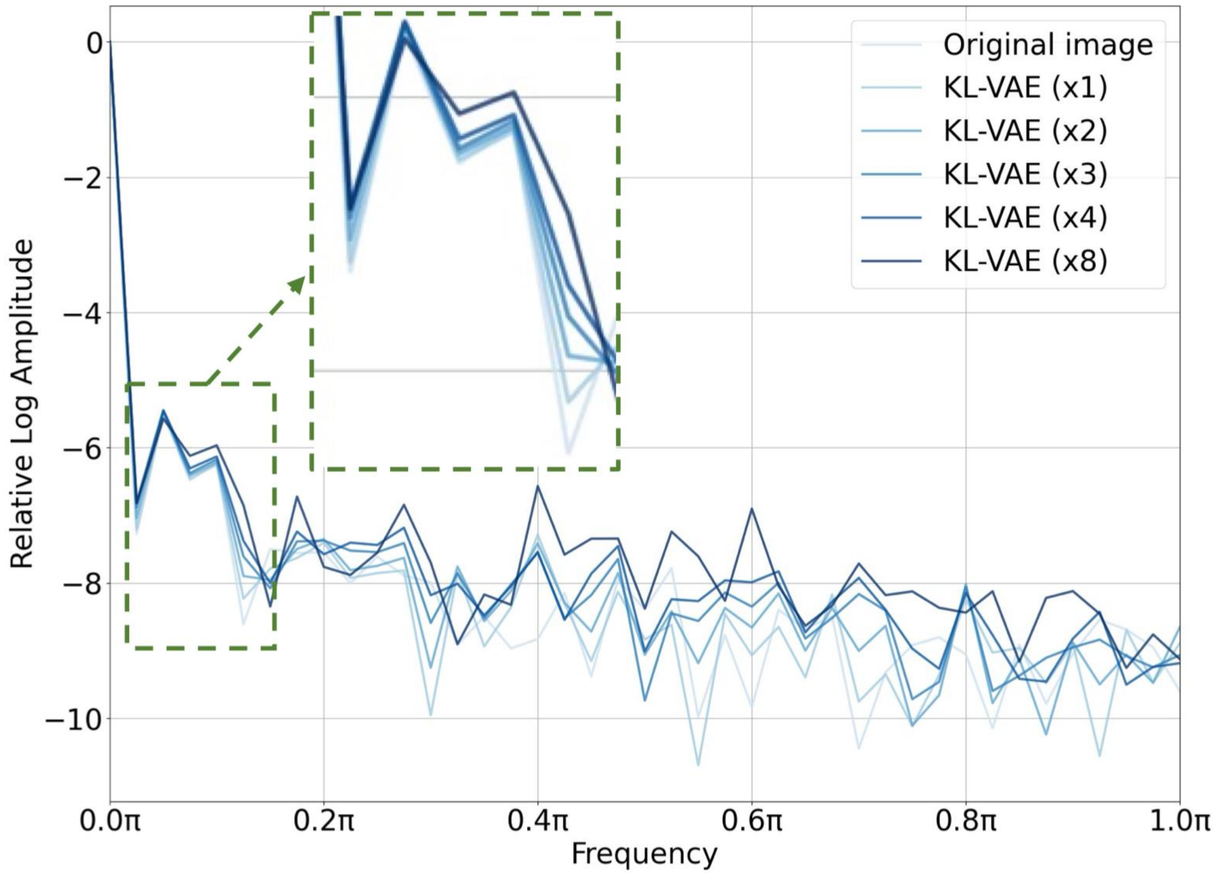
- MAE is fine-tuned to handle continuous masks.
- Alignment module is trained with standard diffusion objective.

Visual-Inconsistency Issue (of SD)

Information Loss of VAE used in SD



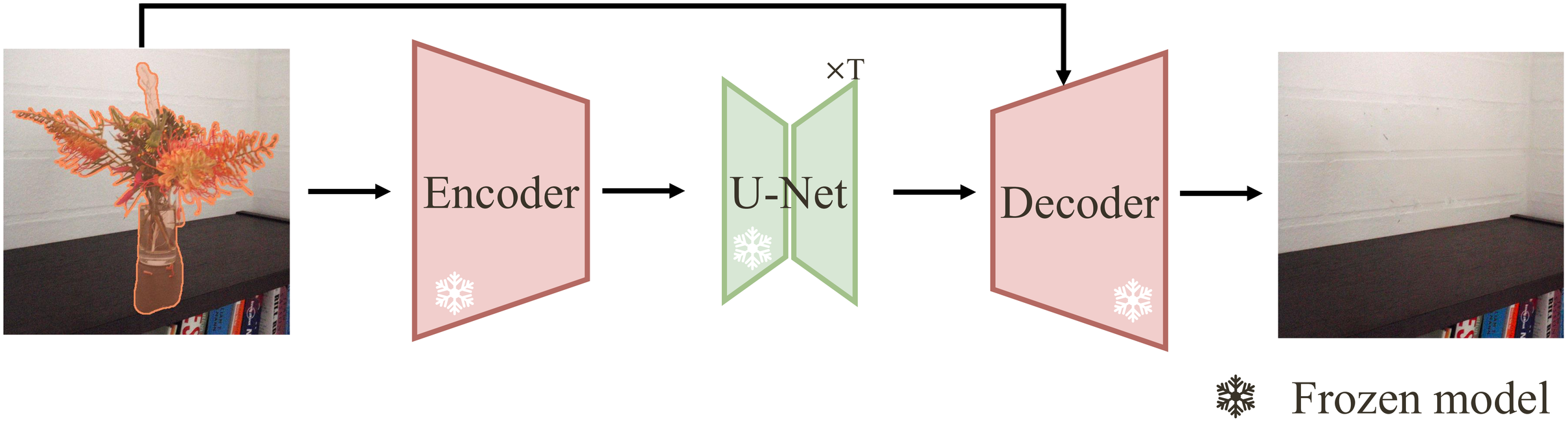
(a) Images decoded by KL-VAE repeatedly for different times



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

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.

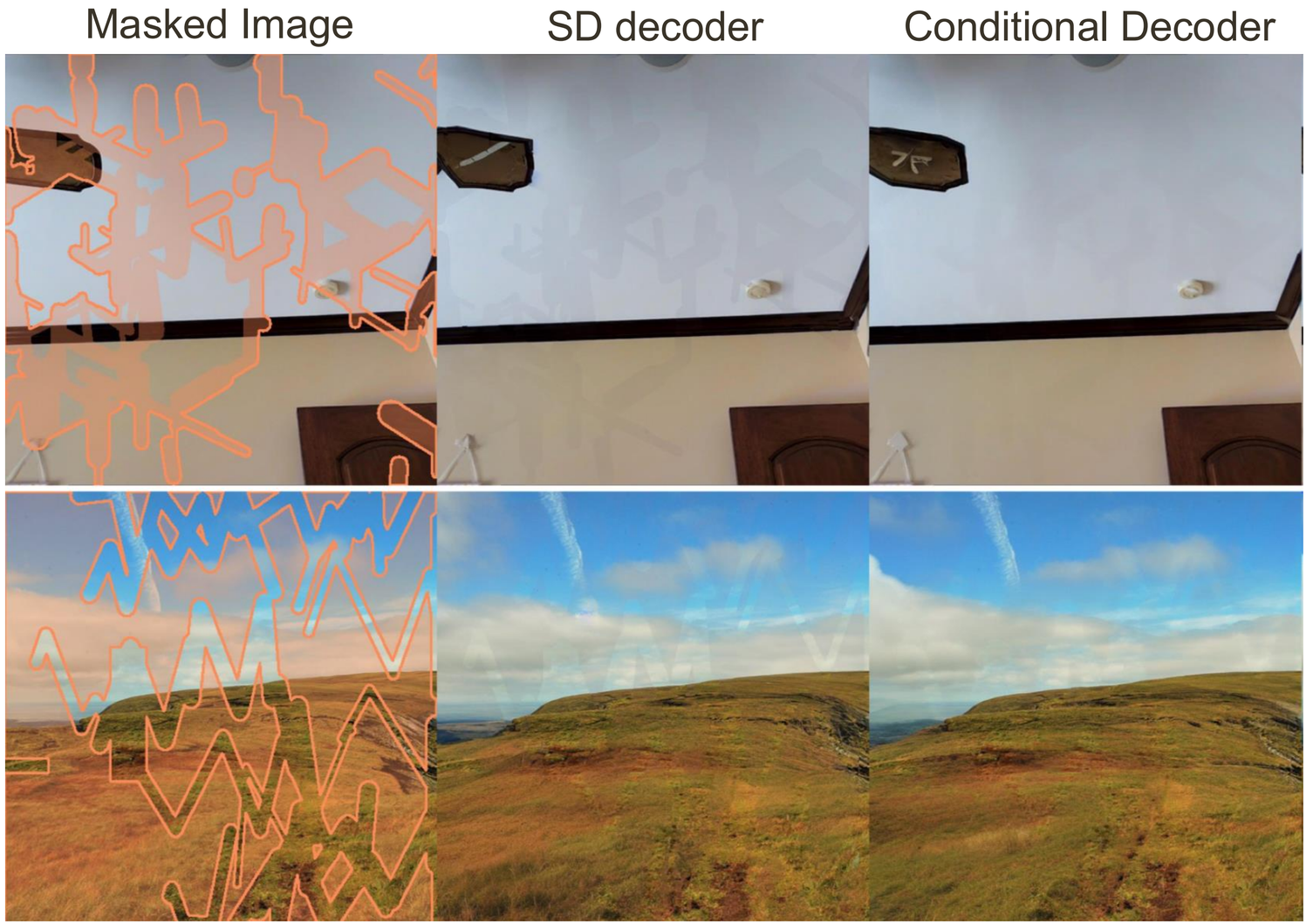


Better Decoder?

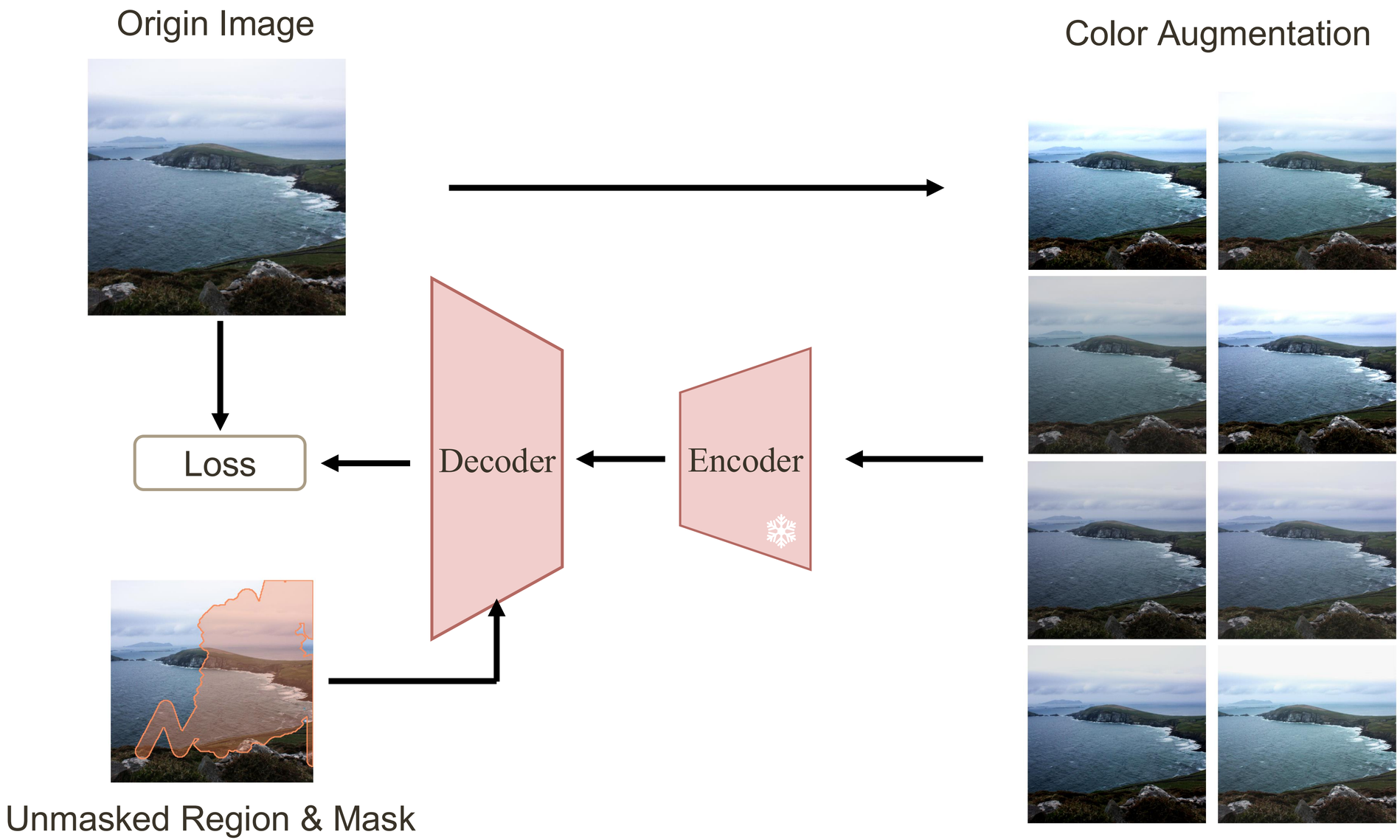
Yes, but not good enough.

The condition of unmasked region works.

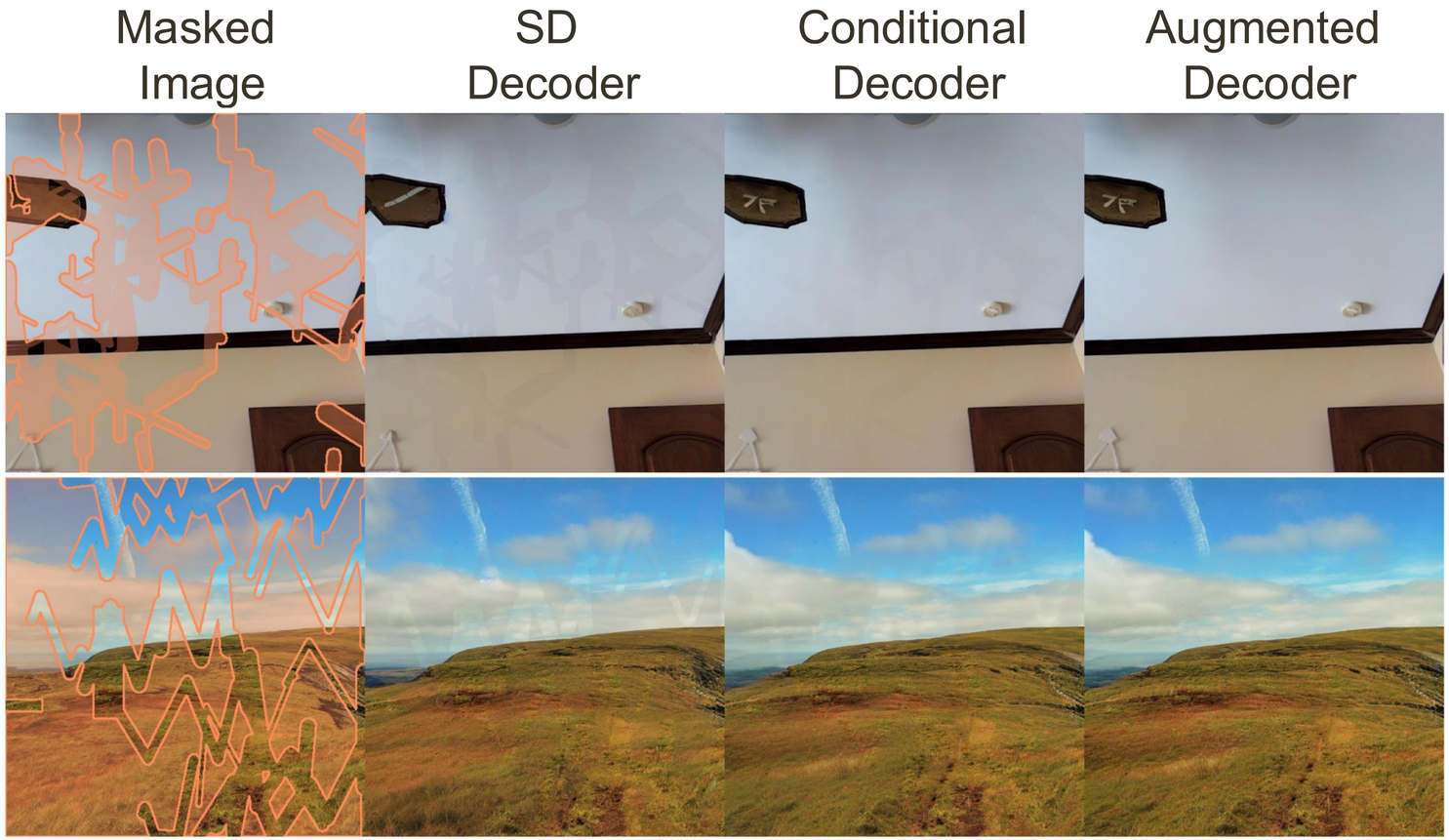
We need to train the decoder to reduce color shift explicitly.



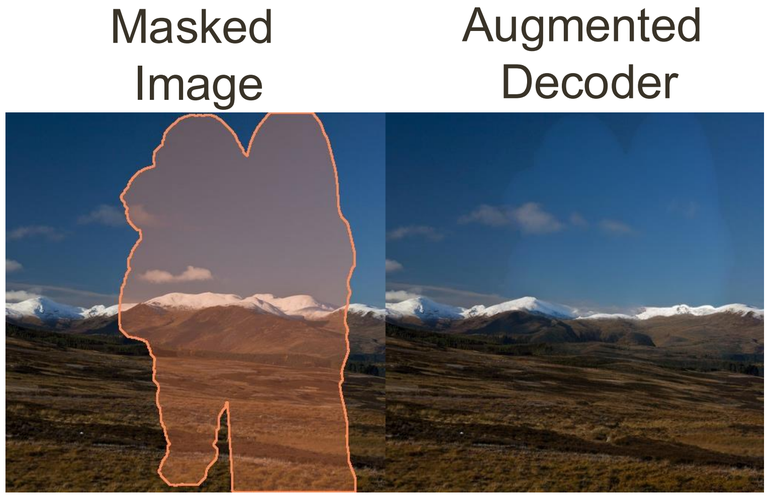
Harmonizing while Decoding: Color Augmentation



Better Decoder Now?

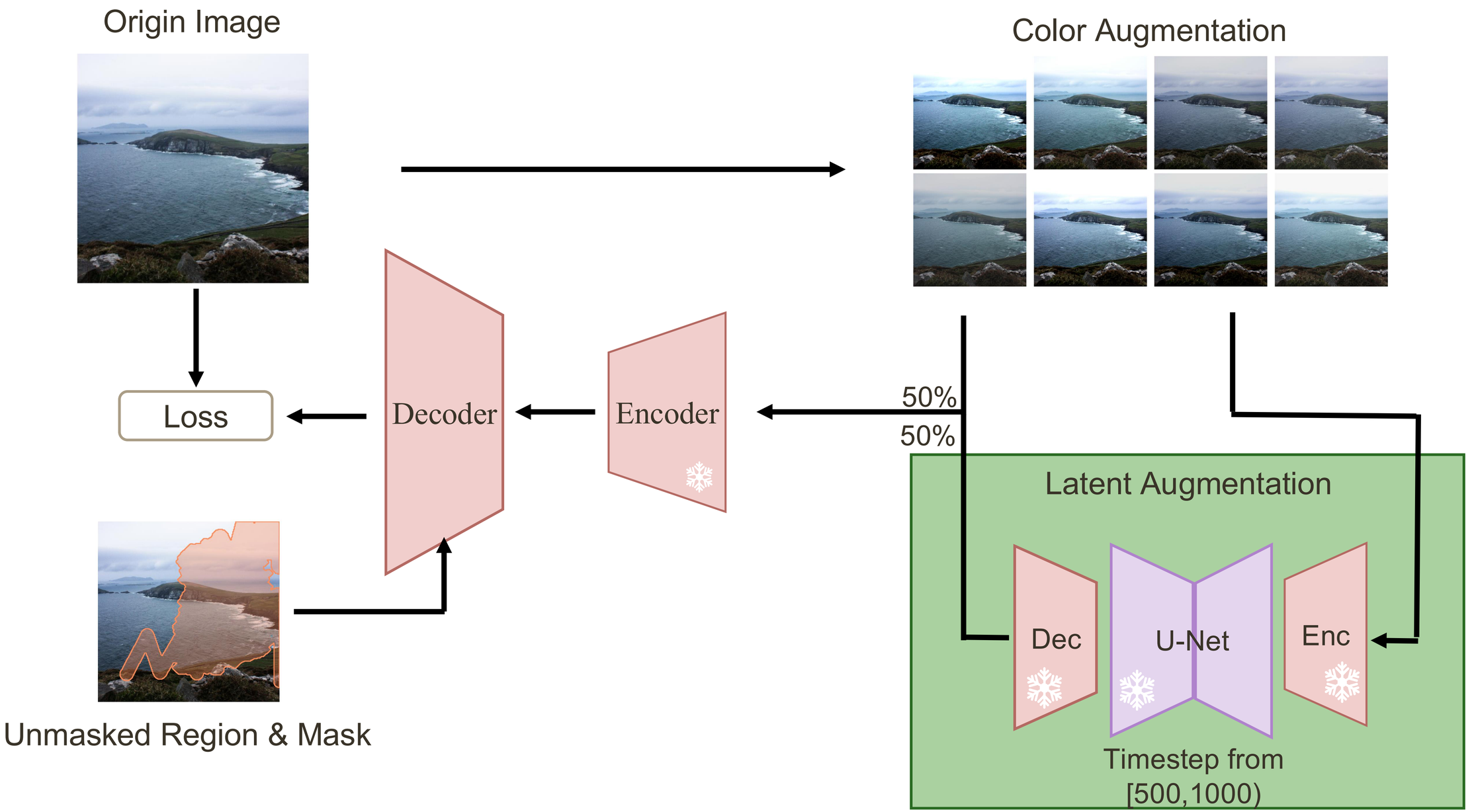


Yes, but not in all cases.



Information loss also exists in the U-Net.

Harmonizing while Decoding: Latent Augmentation



Better Decoder Now?

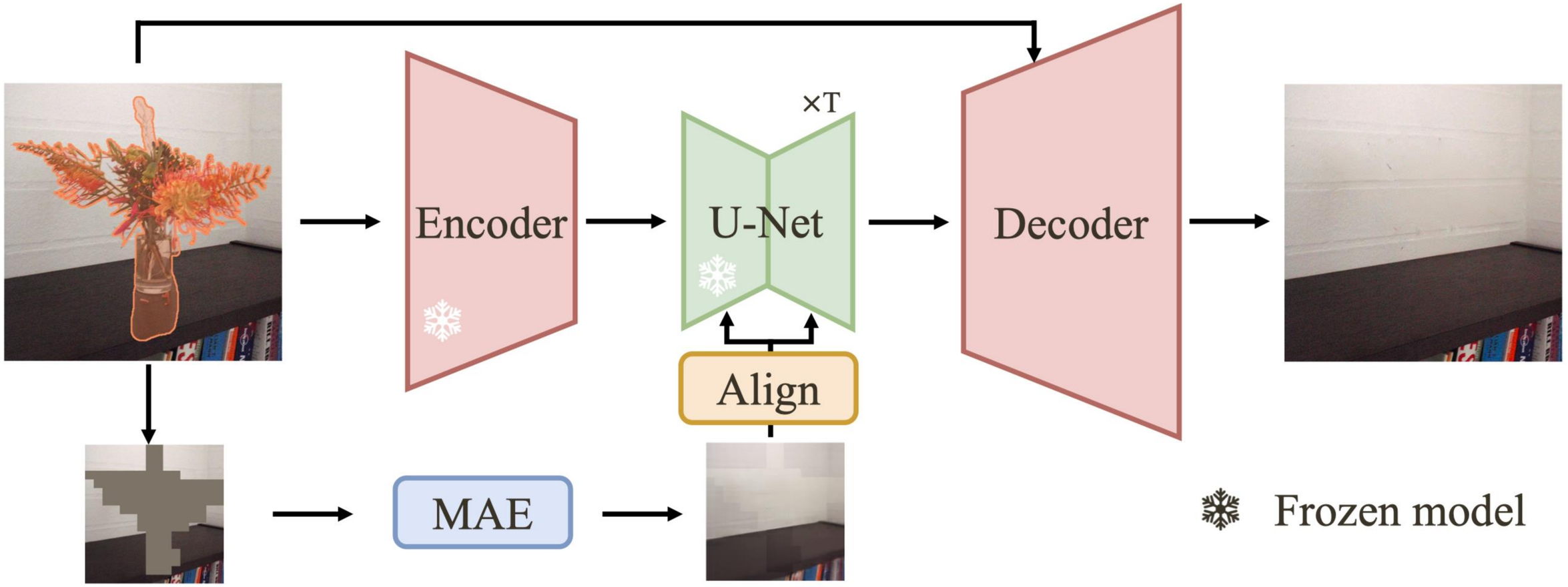
Masked Image | w/o latent aug | w/ latent aug

Yes!

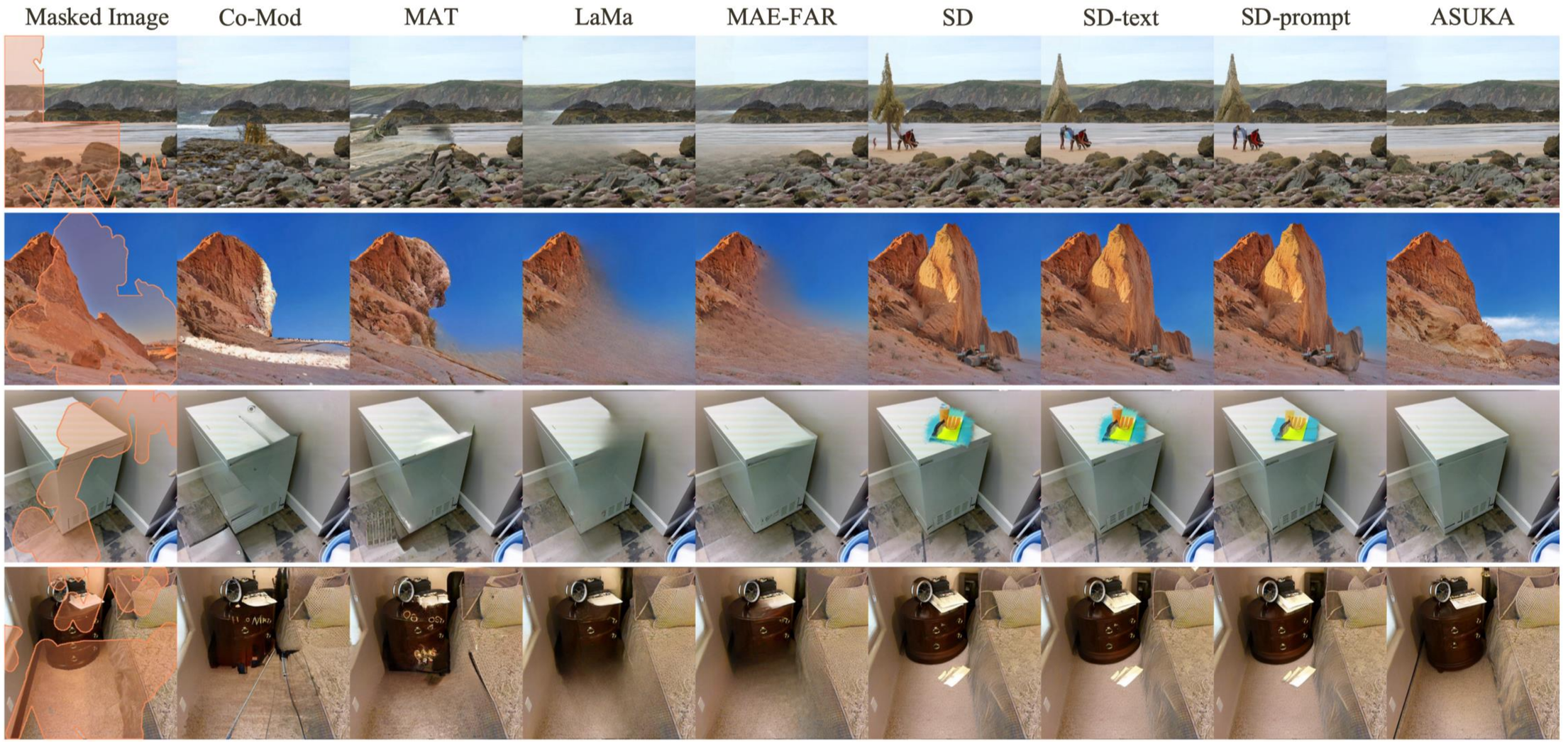
(d) Comparison of different decoders for SD. VAE [39] is the original decoder used by SD; + cond. [62] is the decoder conditioned on unmasked image; + color uses the color augmentation strategy to perform local harmonization task; Ours further combines latent augmentation strategy to handle the gap between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

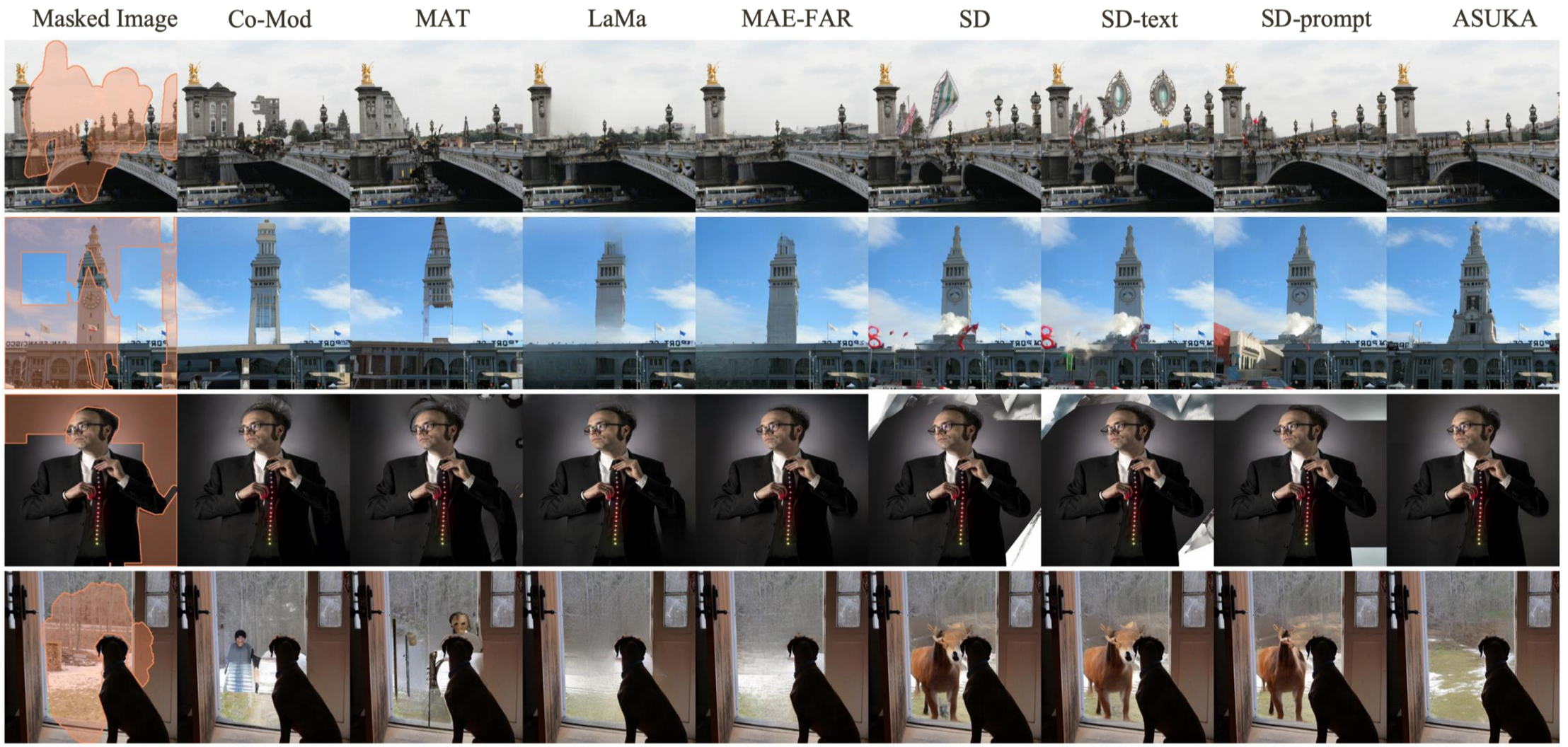
Aligned Stable Inpainting with UnKnown Areas Prior



Comparison


Masked Image
Co-Mod
MAT
LaMa
MAE-FAR
SD
SD-text
SD-prompt
ASUKA

Comparison (cont.)



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**

Summary

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

ICLR 2026

Next Visual Granularity Generation

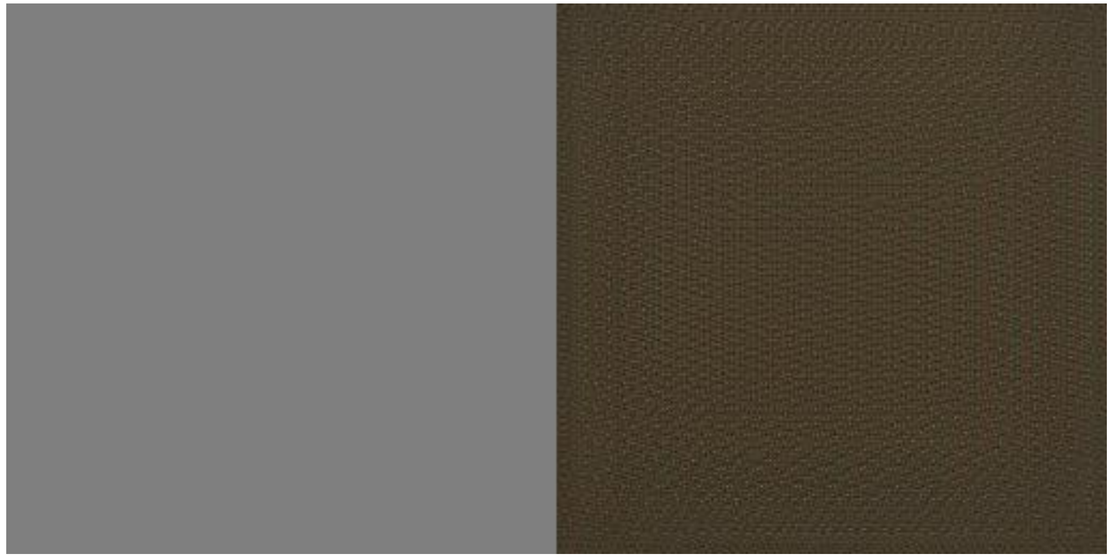


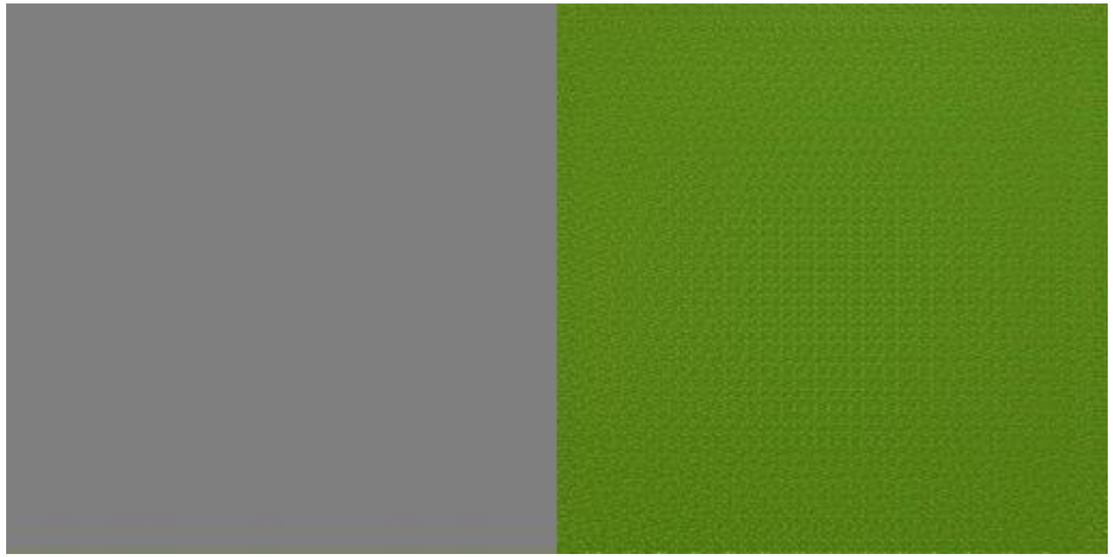


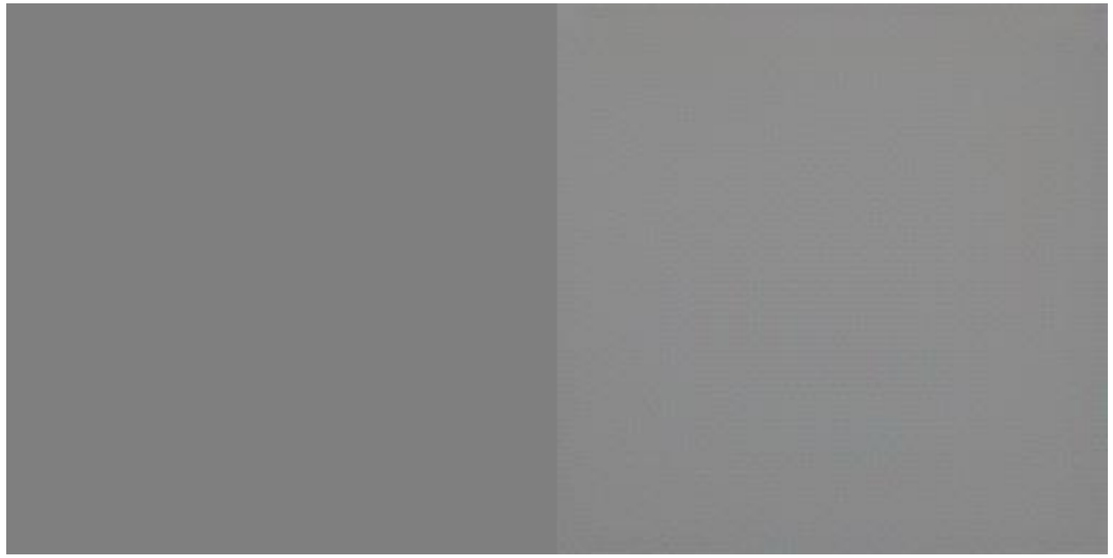

Yikai Wang[1] Zhouxia Wang[1] Zhonghua Wu[2] Qingyi Tao[2] Kang Liao[1] Chen Change Loy[1]

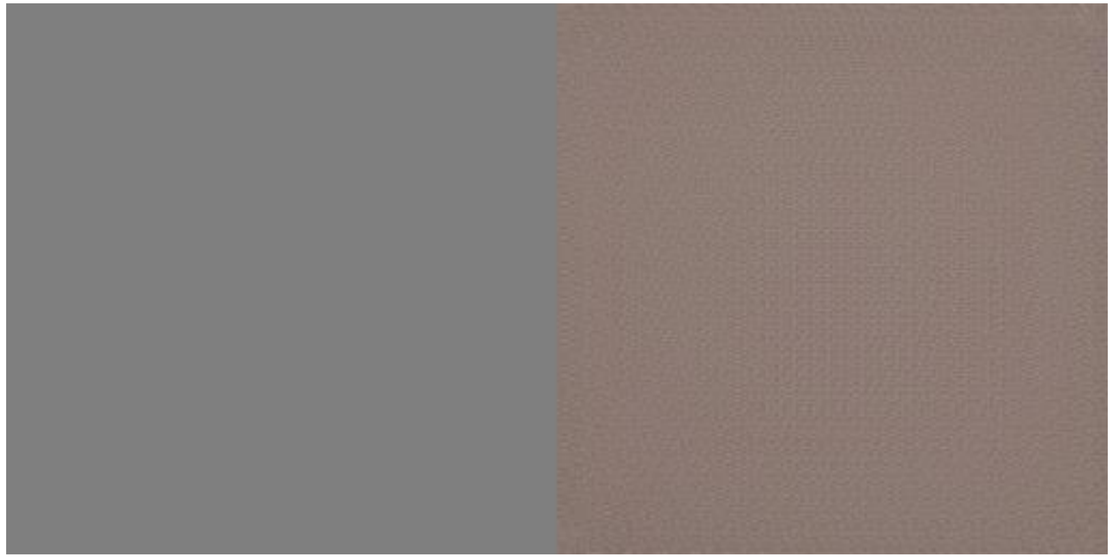
[1]: S-Lab, Nanyang Technological University; [2]: SenseTime Research

Illustration of Generation Process





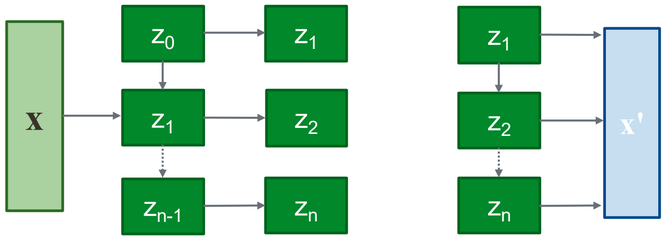




The Gap in Current Frameworks

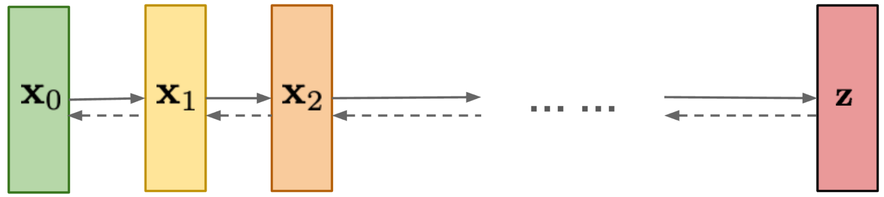
Standard Approaches:

Views image as a 1D visual sentence



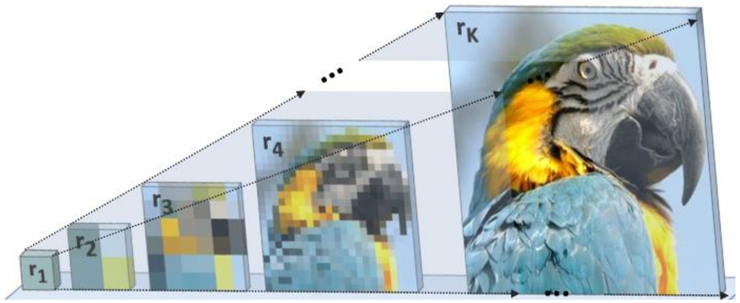
(Masked) Auto-regressive models

Views image as some high-dim distribution



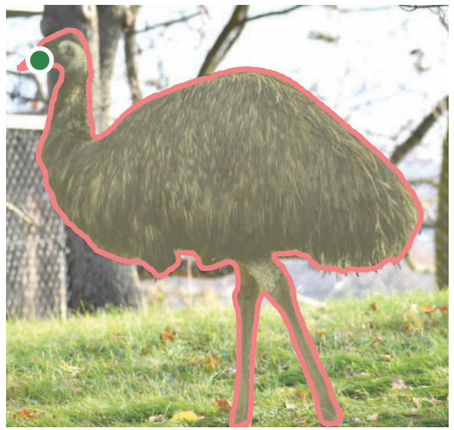
Diffusion and Flow models

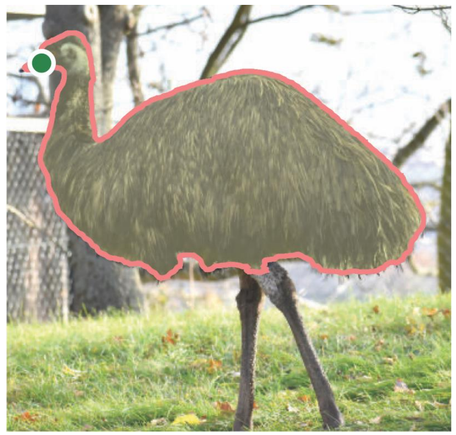
Views image as a resolution-based visual pyramid



Visual Auto-regressive models

Human Perception:





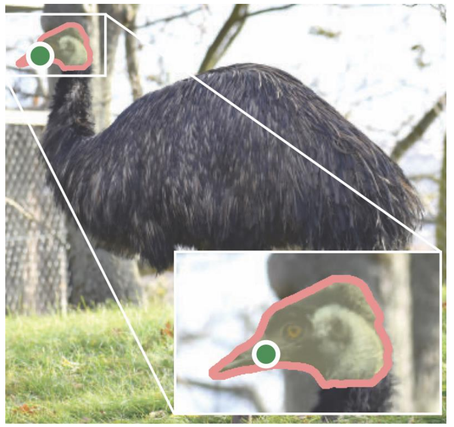
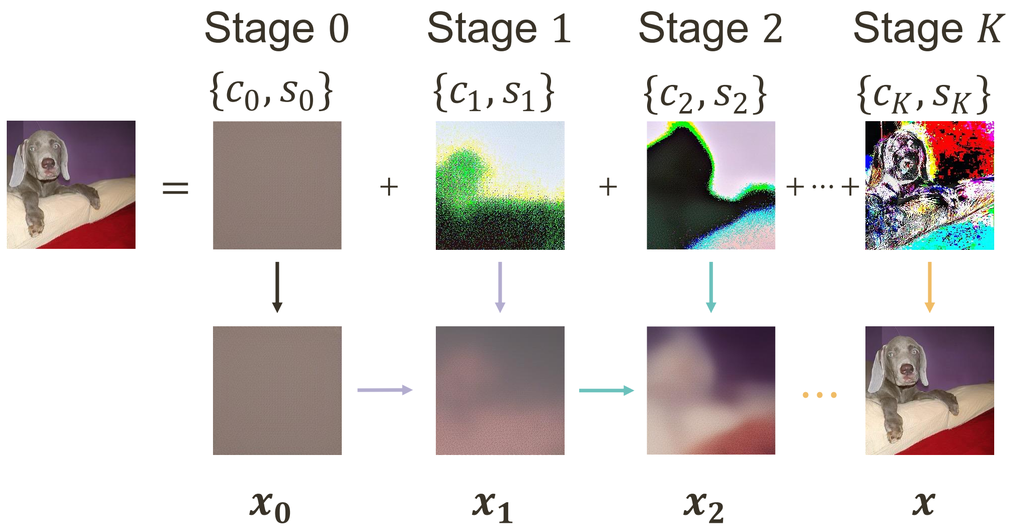
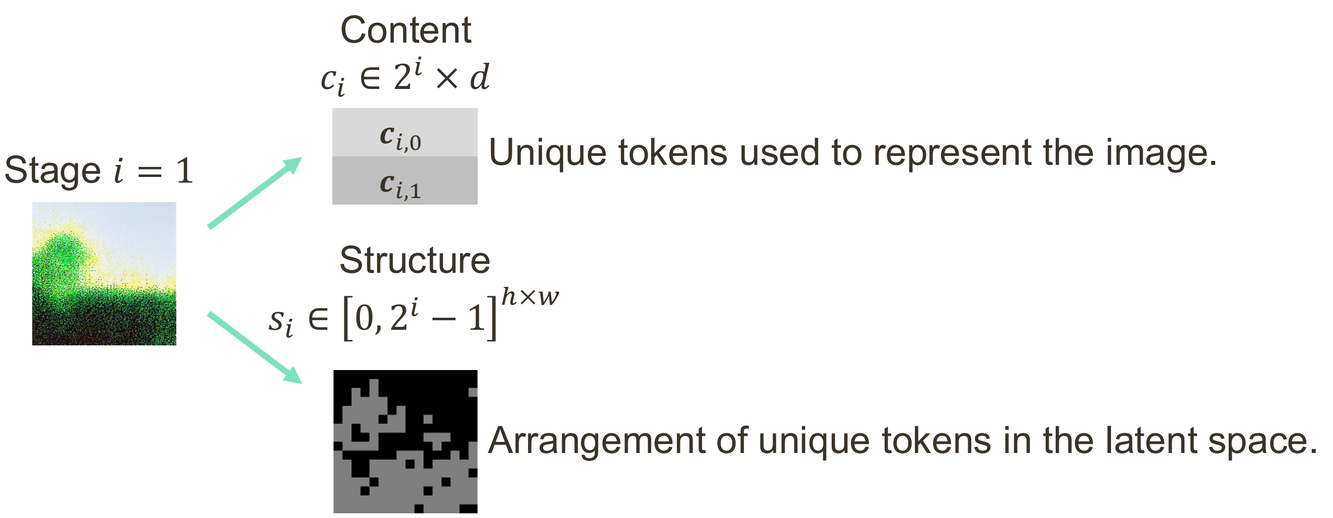
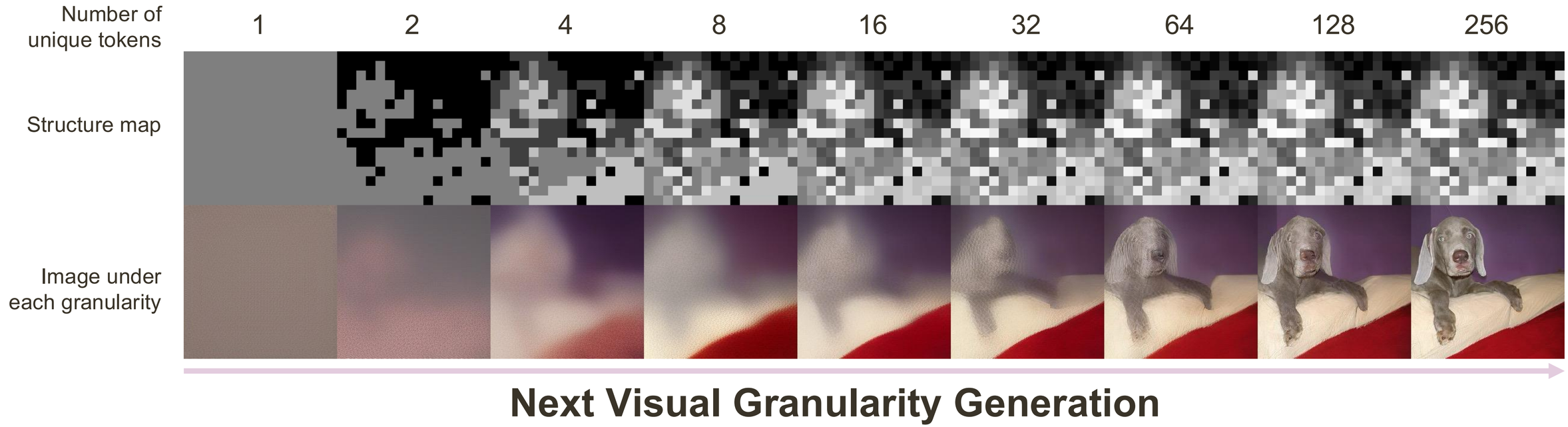


Image depicted from Alexander Kirillov, et al. Segment Anything. ICCV, 2023.

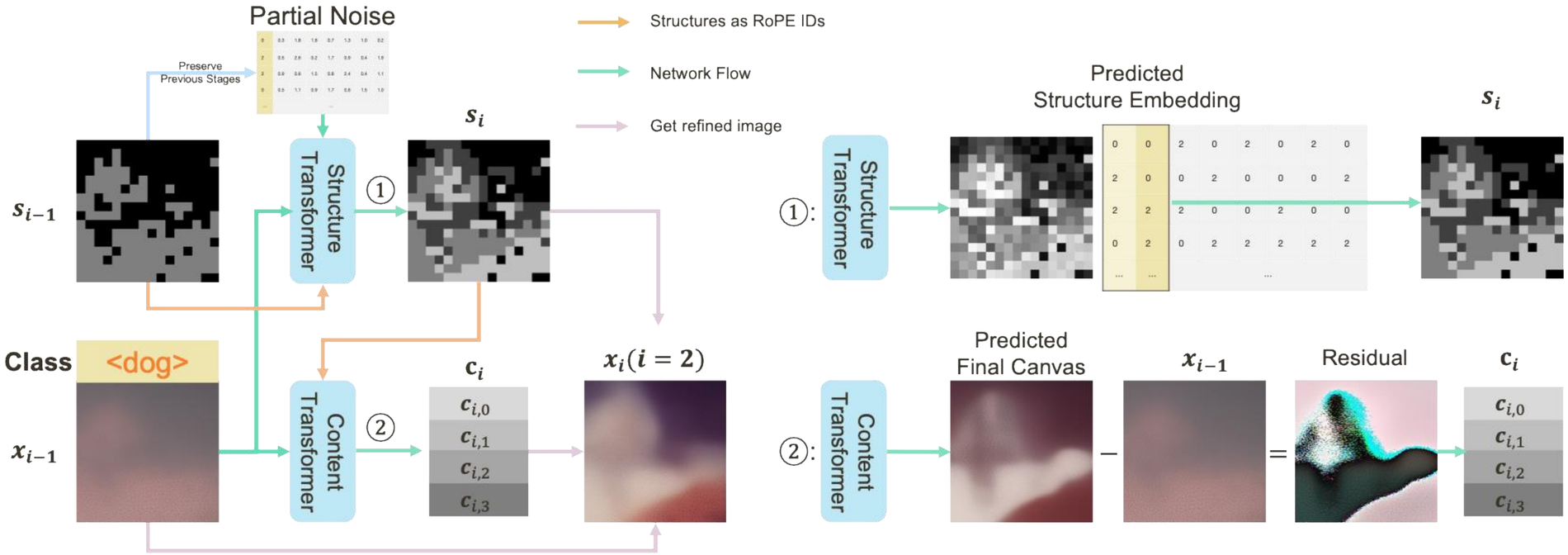
The Visual Granularity Sequence



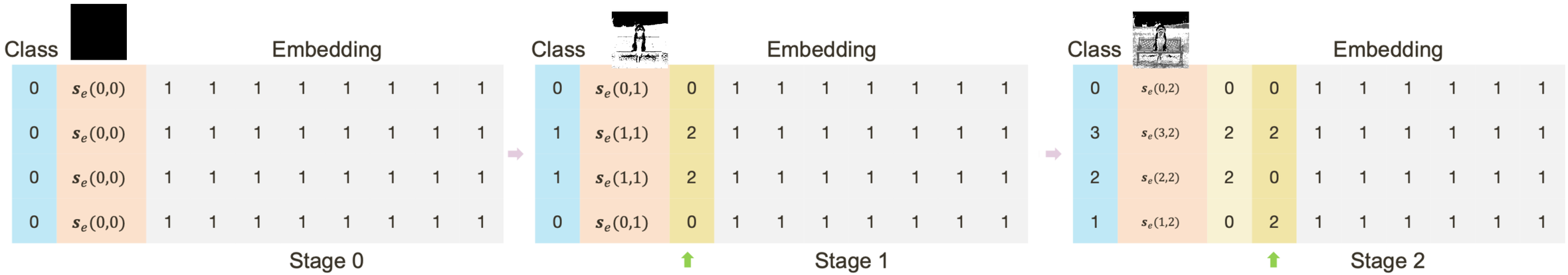




Canvas Refine Generation Framework



Structure Embedding:



Experimental Results: Reconstruction

Table 1: Reconstruction performance on ImageNet validation dataset. Results of VAR are reproduced, while other competitors are from IBQ. LPIPS are calculated via VGG (Simonyan & Zisserman, 2015).

Tokenizer	#Tokens	Ratio	Codebook	rFID(↓)	LPIPS(↓)	Usage (↑)
VQ-GAN (Esser et al., 2021)	16 × 16	16	1,024	7.94	-	44%
SG-VQGAN (Rombach et al., 2022)	16 × 16	16	16,384	5.15	-	-
VQGAN-LC (Zhu et al., 2024)	16 × 16	16	100, 000	2.62	0.2212	99%
MaskGIT (Chang et al., 2022)	16 × 16	16	1,024	2.28	-	-
LlamaGen (Sun et al., 2024)	16 × 16	16	16,384	2.19	0.2281	97%
Open-MAGVIT2 (Luo et al., 2024)	16 × 16	16	262,144	1.17	0.2038	100%
IBQ (Shi et al., 2024)	16 × 16	16	262,144	1.00	0.2030	84%
VAR (Tian et al., 2024)	680	16	4,096	1.06	0.1863	100%
NVG	511*	16	4,096	**0.74**	0.1875	100%

* We define #Tokens as the number of unique tokens. If consider the number of total tokens, our model uses 256×9 tokens. In comparison, VAR uses 256×10 tokens, showing that our tokenizer achieves better quantization with fewer tokens.

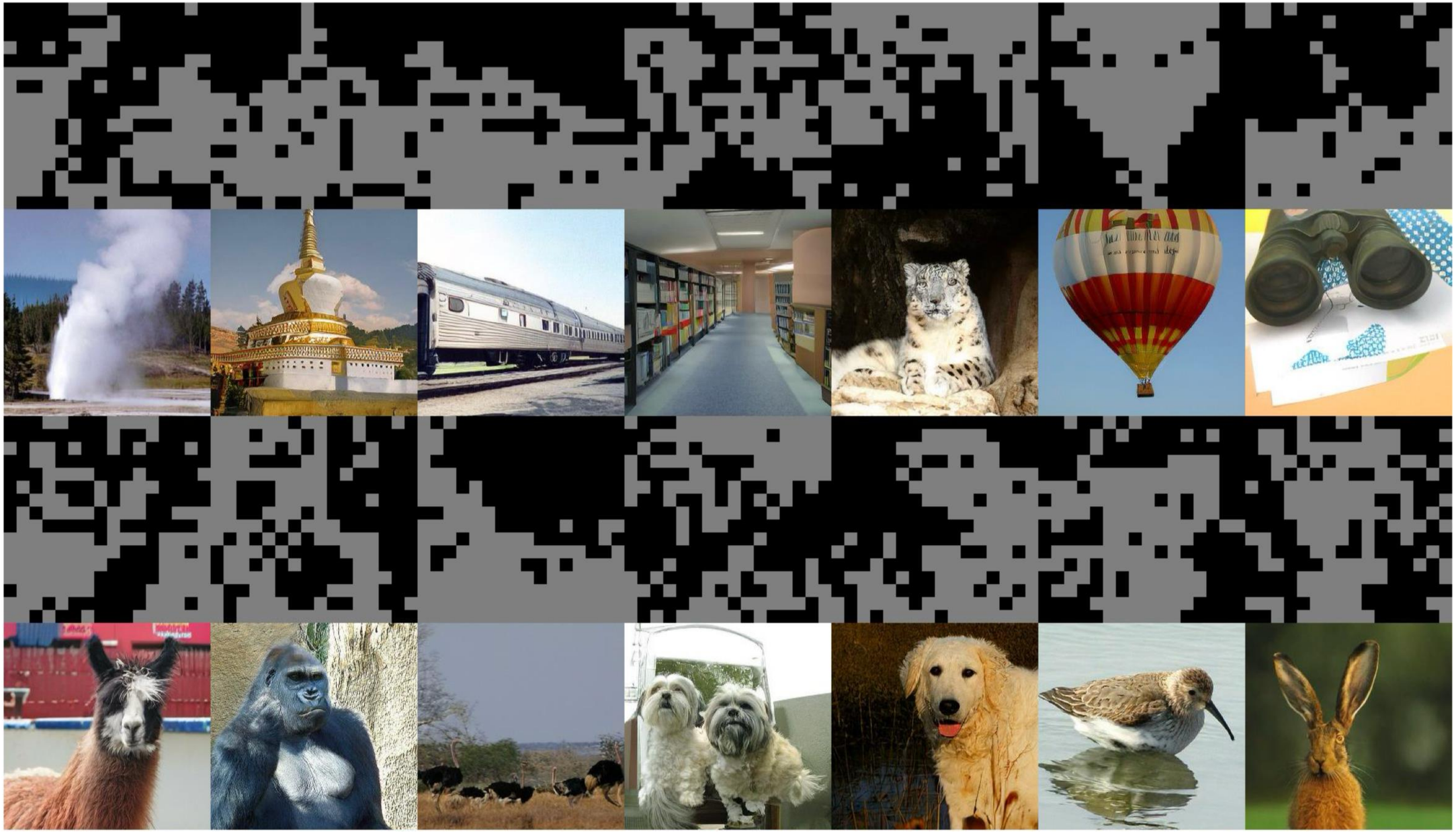
In the first stage, VAR's codebook utilization rate is **25.39%**, while ours is **68.55%**, indicating a more balanced codebook.

Experimental Results: Generation

Table 2: Generation performance on class-conditional ImageNet 256 × 256.

Type	Model	FID(↓)	IS(↑)	Pre(↑)	Rec(↑)	#Para	#Train†	#Step
Diff	LDM-4-G (Rombach et al., 2022)	3.60	247.7	–	–	400M	178K	250
	DiT-XL/2 (Peebles & Xie, 2023)	2.27	278.2	0.83	0.57	675M	7M	250
	SiT-X (Ma et al., 2024)	2.06	270.3	0.82	0.59	675M	7M	250
AR	MaskGIT (Chang et al., 2022)	6.18	182.1	0.80	0.51	227M	300e	8
	LlamaGen-L (Sun et al., 2024)	3.07	256.1	0.83	0.52	343M	300e	576
	LlamaGen-XXL (Sun et al., 2024)	2.34	253.9	0.80	0.59	1.4B	300e	576
	EAR-H (Shao et al., 2025)	1.97	289.6	0.81	0.59	937M	800e	64
	CausalFusion-XL (Deng et al., 2024)	1.77	282.3	0.82	0.61	676M	800e	250
VAR	VAR-*d*16 (Tian et al., 2024)	3.30	274.4	0.84	0.51	310M	200e	10
	VAR-*d*20 (Tian et al., 2024)	2.57	302.6	0.83	0.56	600M	250e	10
	VAR-*d*24 (Tian et al., 2024)	2.09	312.9	0.82	0.59	1.0B	350e	10
Ours*	NVG-*d*16 (255M+64M)	3.03	279.2	0.82	0.54	320M	200e	9
	NVG-*d*20 (497M+125M)	2.44	310.4	0.80	0.60	622M	250e	9
	NVG-*d*24 (856M+215M)	2.06	317.0	0.79	0.61	1.1B	350e	9

Generated Images align well with the binary structure map

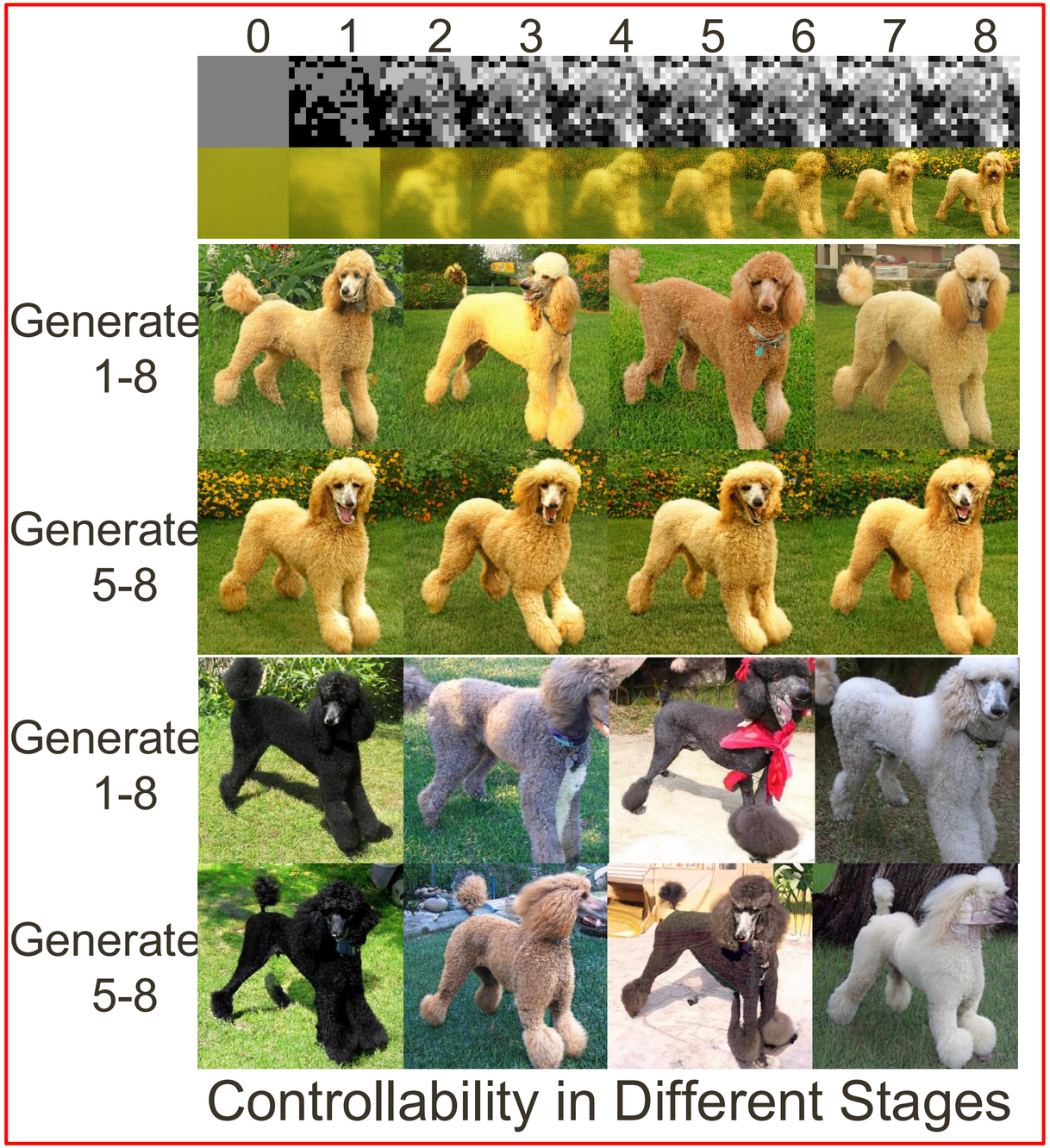


More generation results

Explicit Control Capabilities

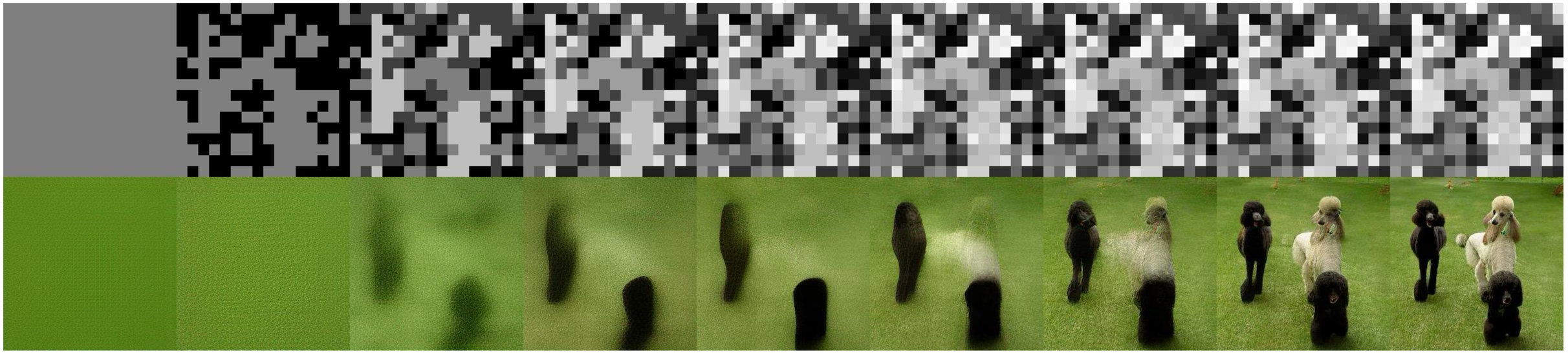
Structure Guided Generation

Structure Transfer



Controllability in Different Stages

Conclusion and Future Work



Conclusion: NVG is a novel generation framework that mimics human perception and offers explicit, built-in structure control.

Future Directions:

- *Region-Aware Generation:* Fine-grained control over specific domain annotations.
- *Physical-Aware Video Generation:* Tracking structured image regions over time for coherent physics.
- *Hierarchical Spatial Reasoning:* Global-to-local reasoning chains for unobserved patches.

Project Page:

https://yikai-wang.github.io/nvg/

Code & Model:

https://github.com/Yikai-Wang/nvg