# Image Inpainting and Editing with Various Prior Guidance

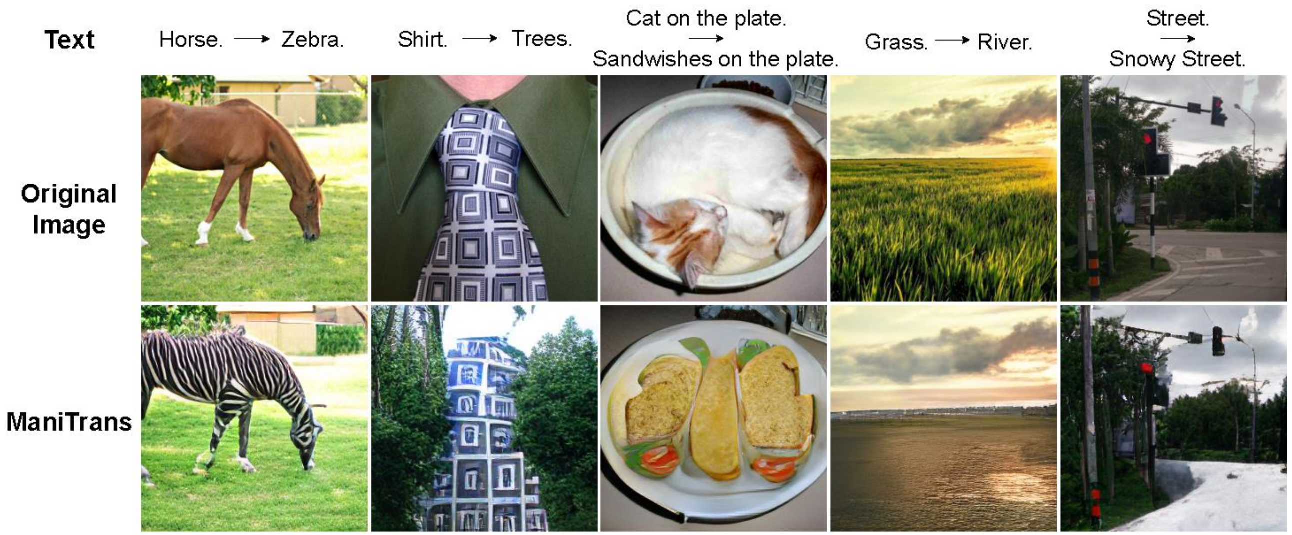
**Qiaole Dong**
**School of Data Science, Fudan University**
qldong18@fudan.edu.cn

# The Tasks

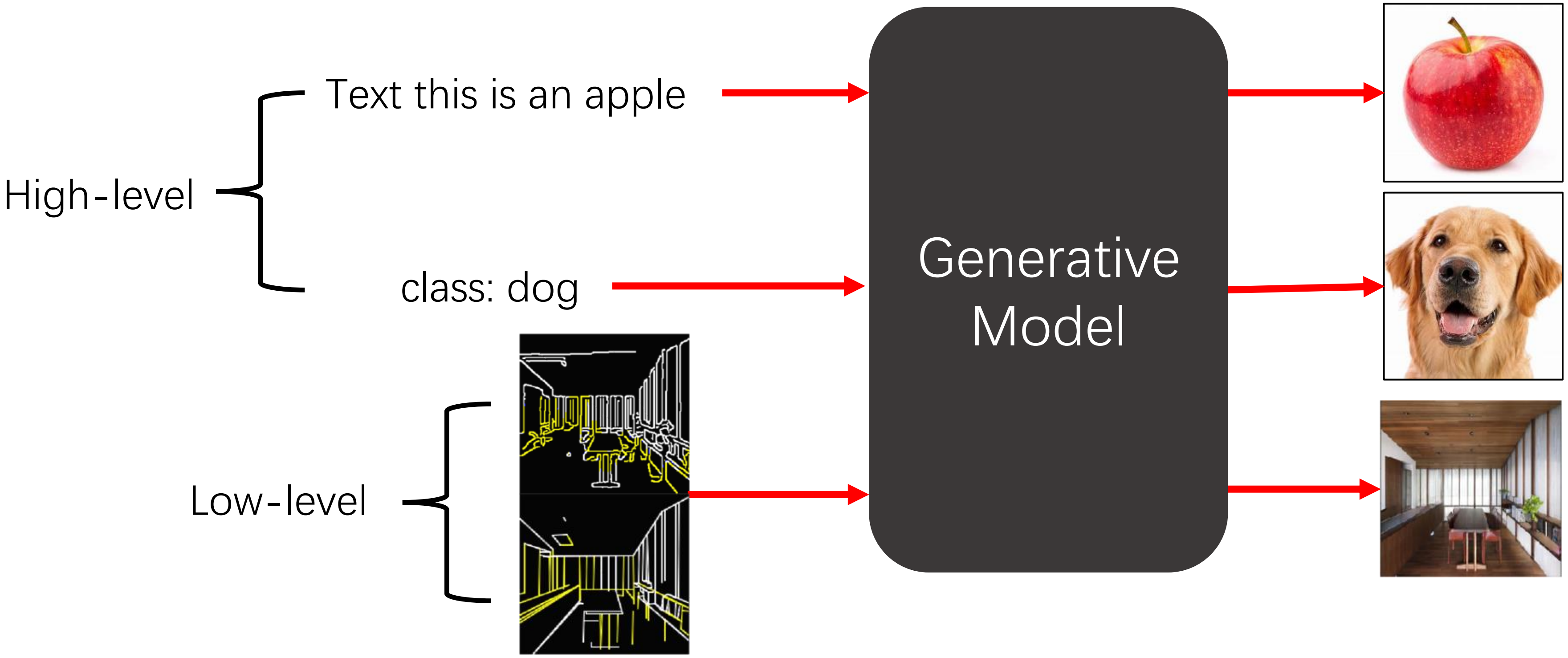
- Image Inpainting at High Resolution

Original (2K) LAMA Our work

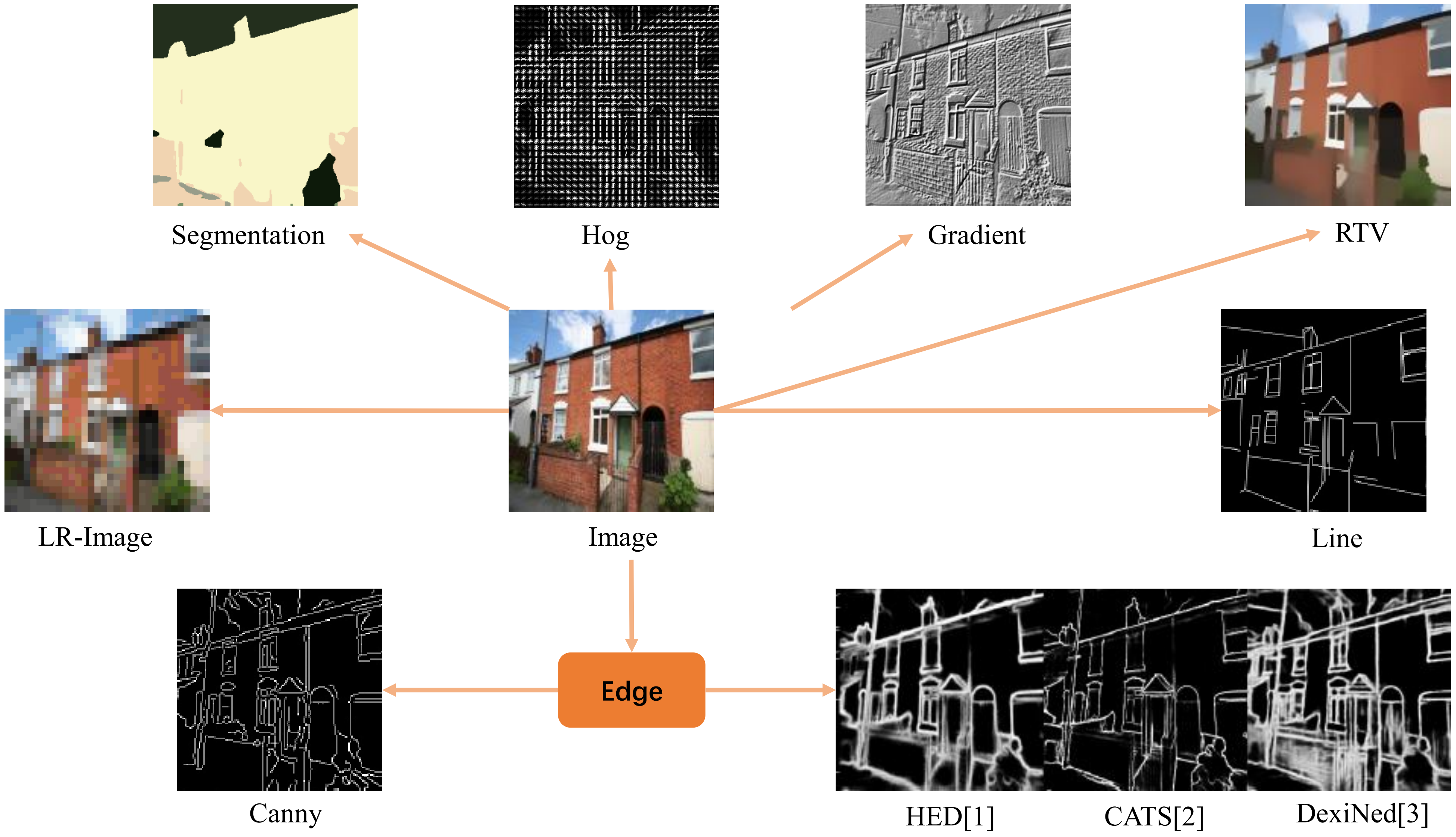
- Entity-level Image Editing



# Recap: What are the Priors?



# Various **Priors**





[1] S. Xie and Z. Tu, “Holistically-nested edge detection,” in Proceedings of the IEEE international conference on computer vision, 2015, pp.1395–1403.
[2] L. Huan, N. Xue, X. Zheng, W. He, J. Gong, and G.-S. Xia, “Unmixing convolutional features for crisp edge detection,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021.
[3] X. S. Poma, A. Sappa, P. Humanante, and A. Arbarinia, “Dense extreme inception network for edge detection,” arXiv preprint arXiv:2112.02250, 2021.

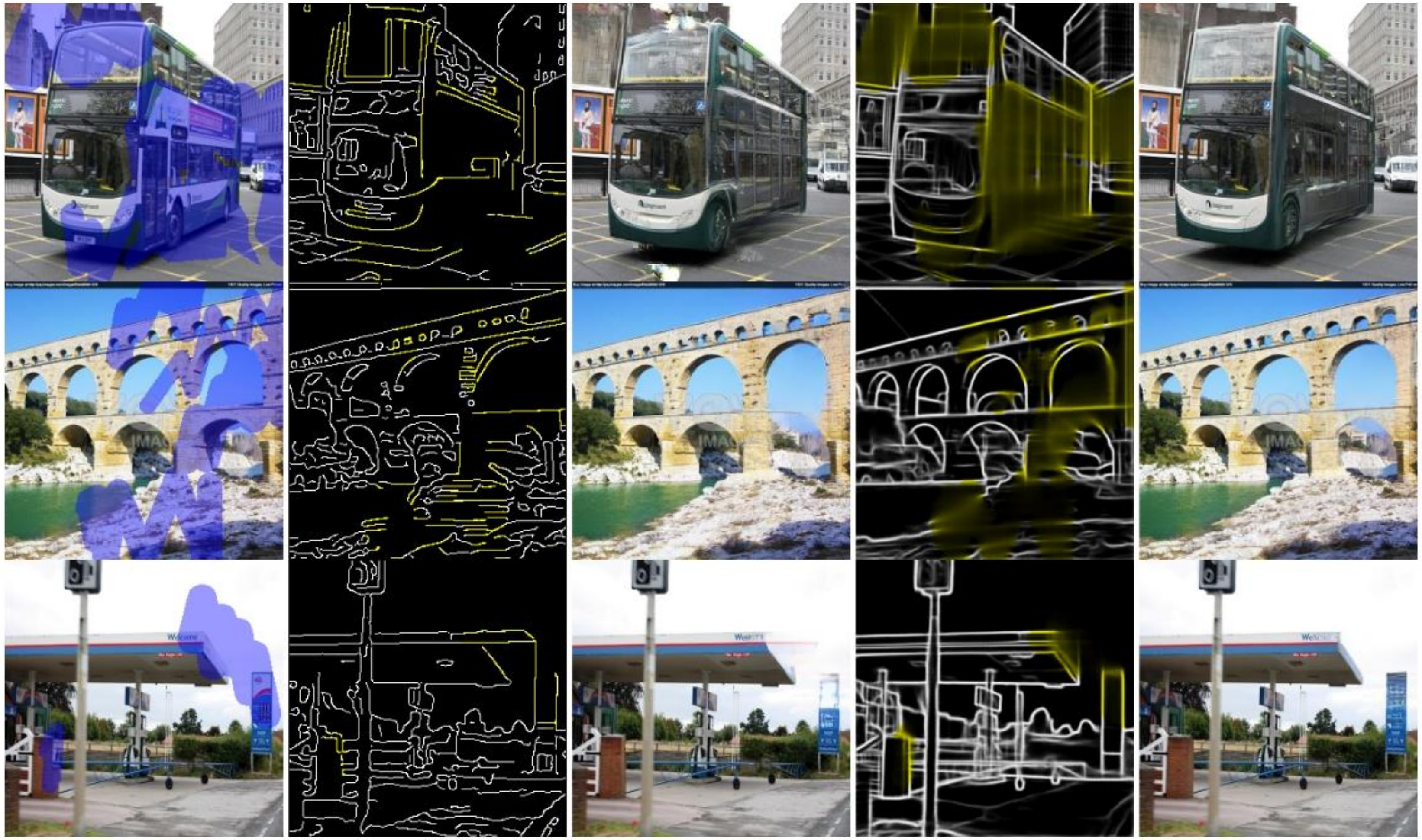
# ZITS++: Image Inpainting by Improving the Incremental Transformer on Structural Priors

Chenjie Cao*, Qiaole Dong*, Yanwei Fu†


(f) High-resolution inpainting results compared with LaMa (first) and our ZITS++ (second).

ZITS++, in submission

# ZITS++ compares different **Edges** for inpainting

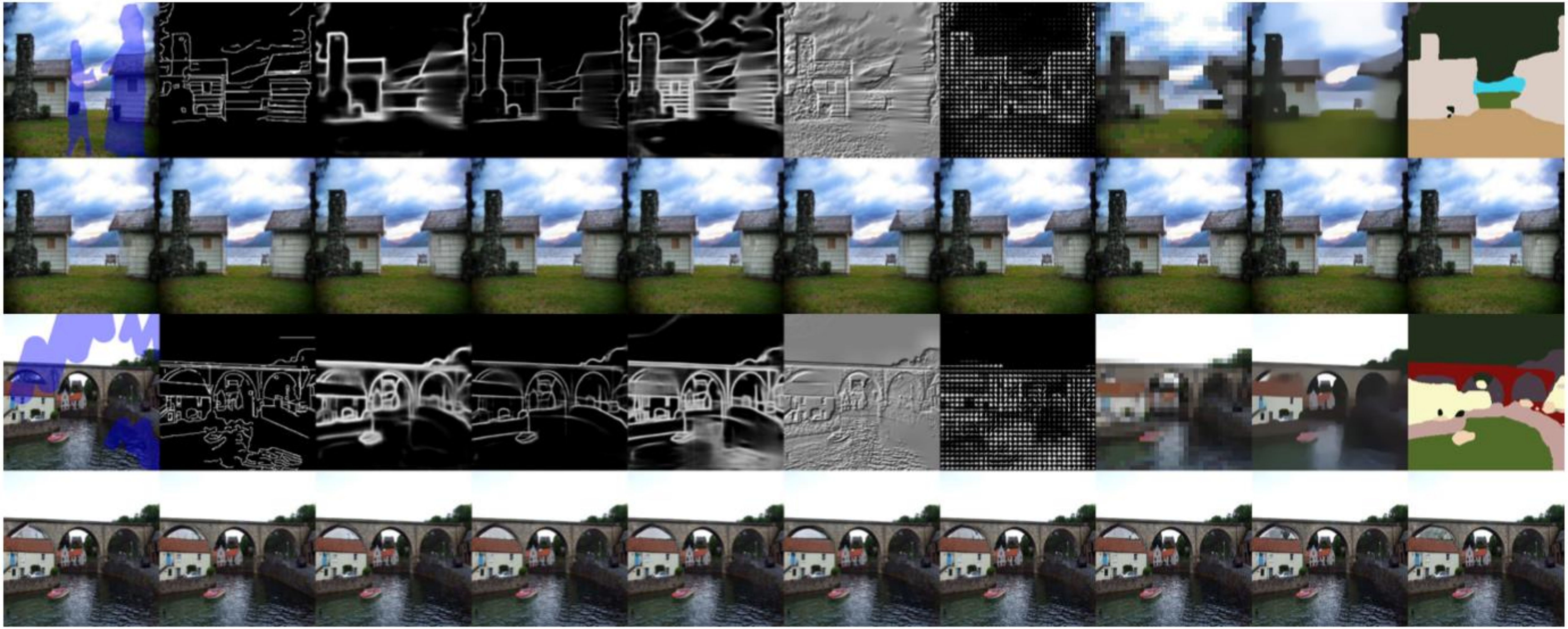


(a) Masked image (b) Canny from ZITS (c) ZITS results (d) CATS from ZITS++ (e) ZITS++ results

Using Learning based Edges (CATS [1]) instead of Canny edge.

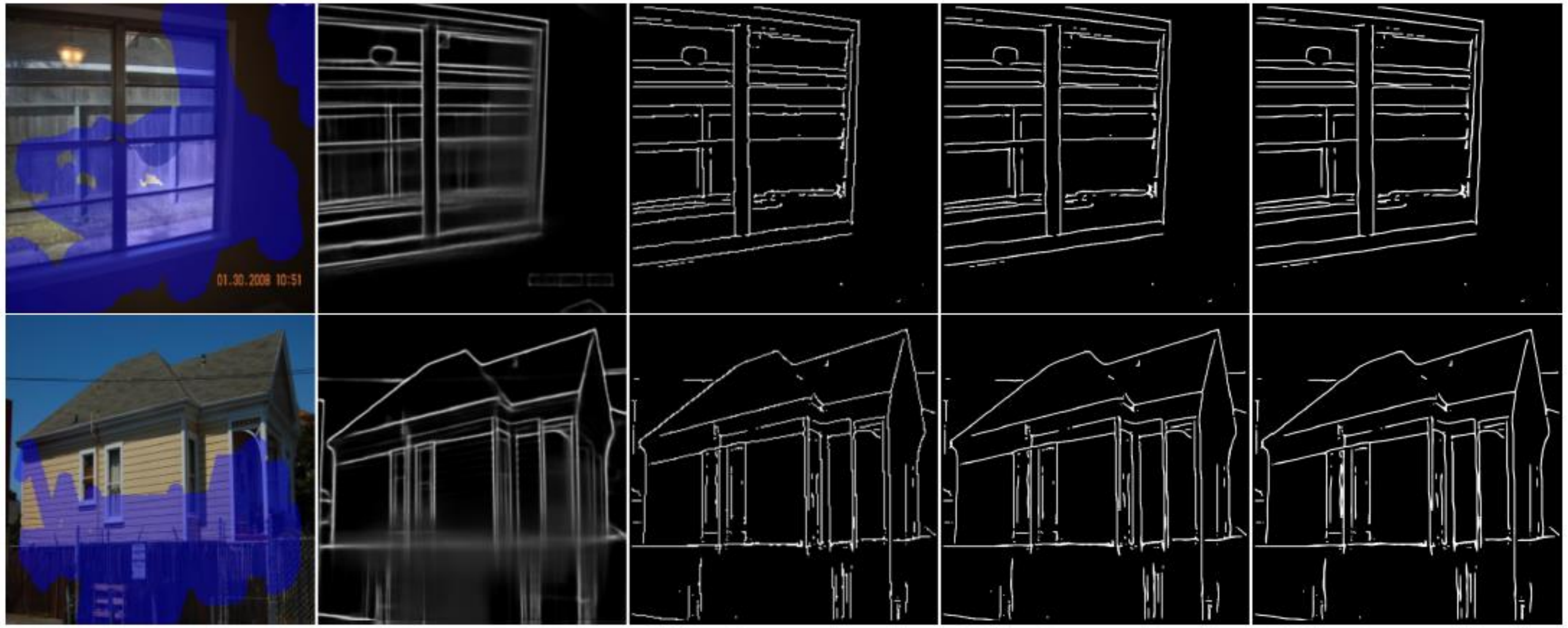
[1] L. Huan, N. Xue, X. Zheng, W. He, J. Gong, and G.-S. Xia, "Unmixing convolutional features for crisp edge detection," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021.

# ZITS++ compares different **priors** for inpainting



(a) Masked image /No prior (b) Canny+line (c) HED (d) CATS (e) DexiNed (f) Gradients (g) HOG (h) LR-RGB (i) RTV (j) Segmentation

# ZITS++ compares different **priors** for inpainting

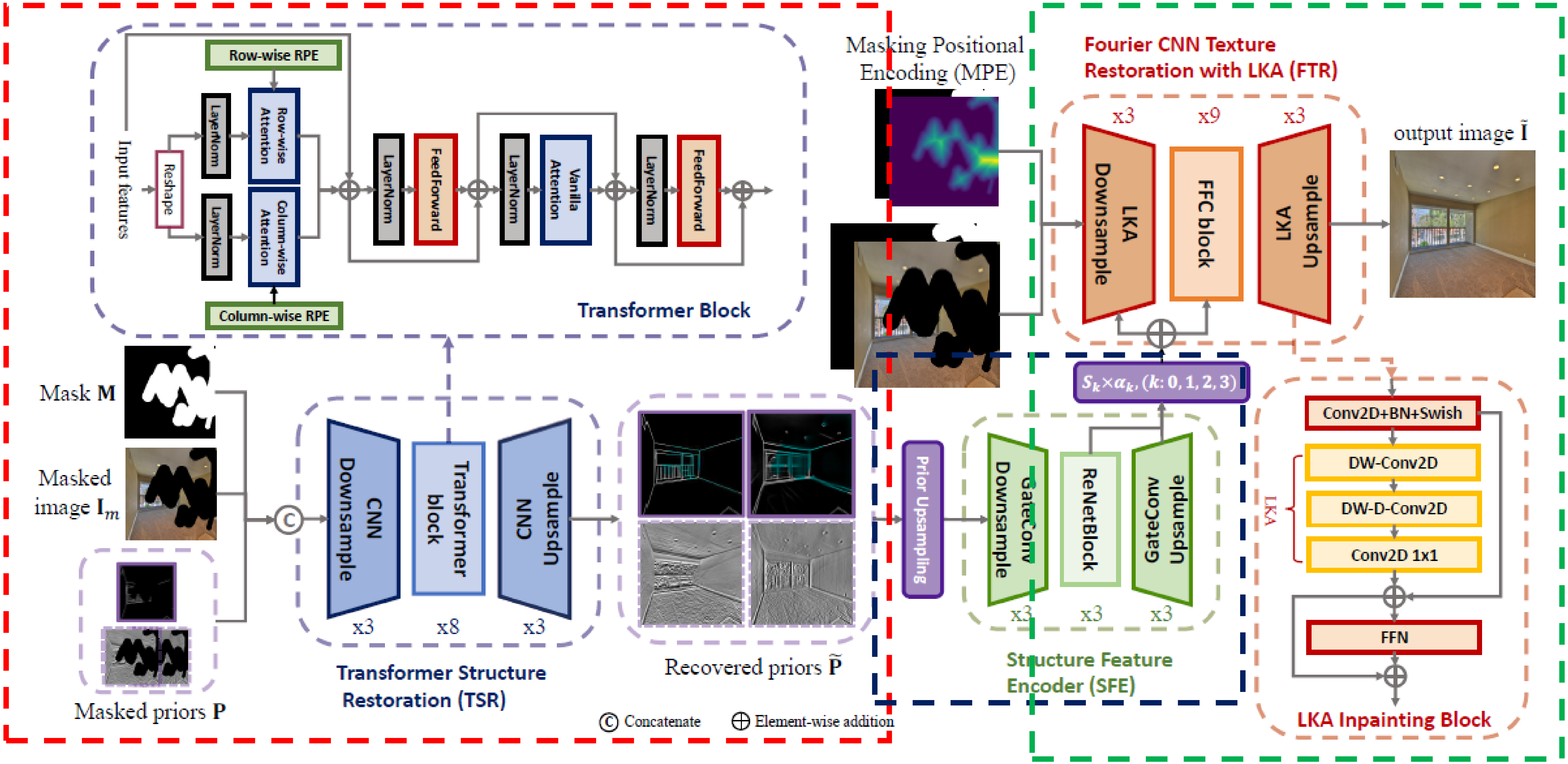


(a) Masked Image (b) TSR CATS (256) (c) After E-NMS (256) (d) 512x512 from SSU (e) 1024x1024 from SSU

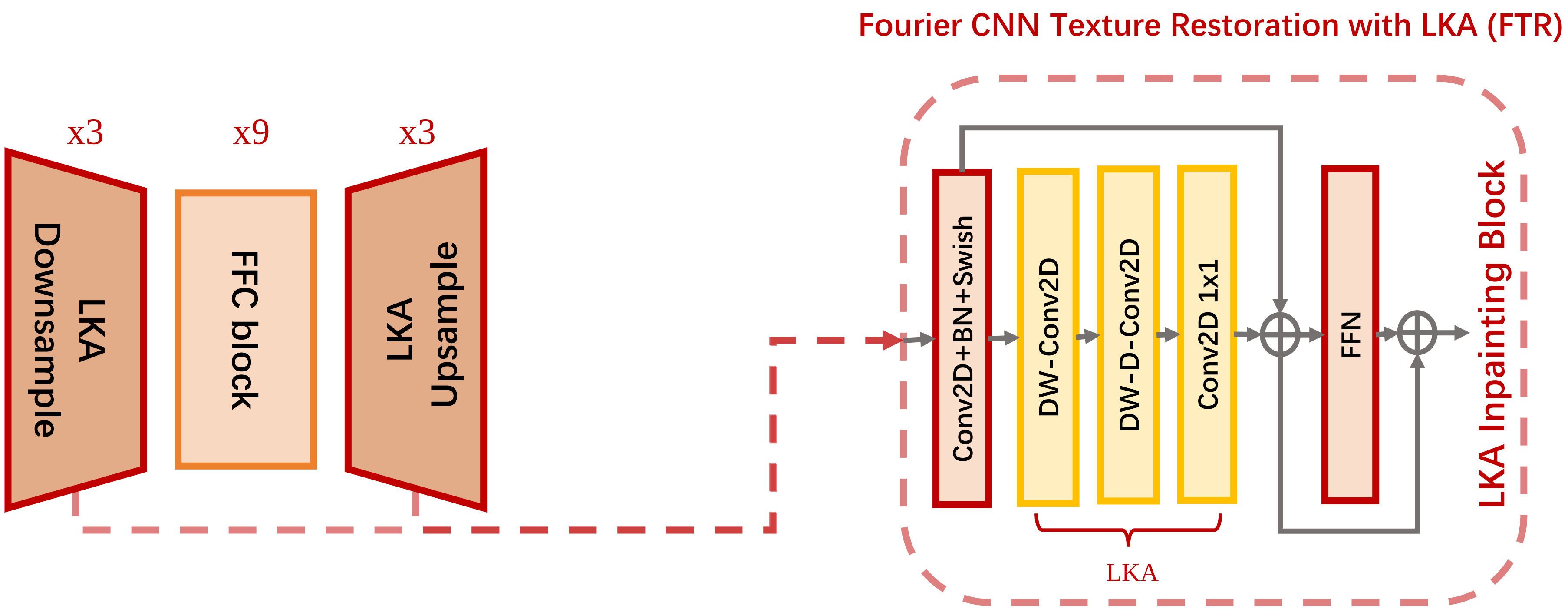
Masked HR image RTV inpainted result CATS inpainted result

Masked HR image Grad inpainted result CATS inpainted result

# ZITS++: Image Inpainting by Improving the Incremental Transformer on Structural Priors



# ZITS++: Further improve the FTR training with **large kernel attention** (**LKA** [1])

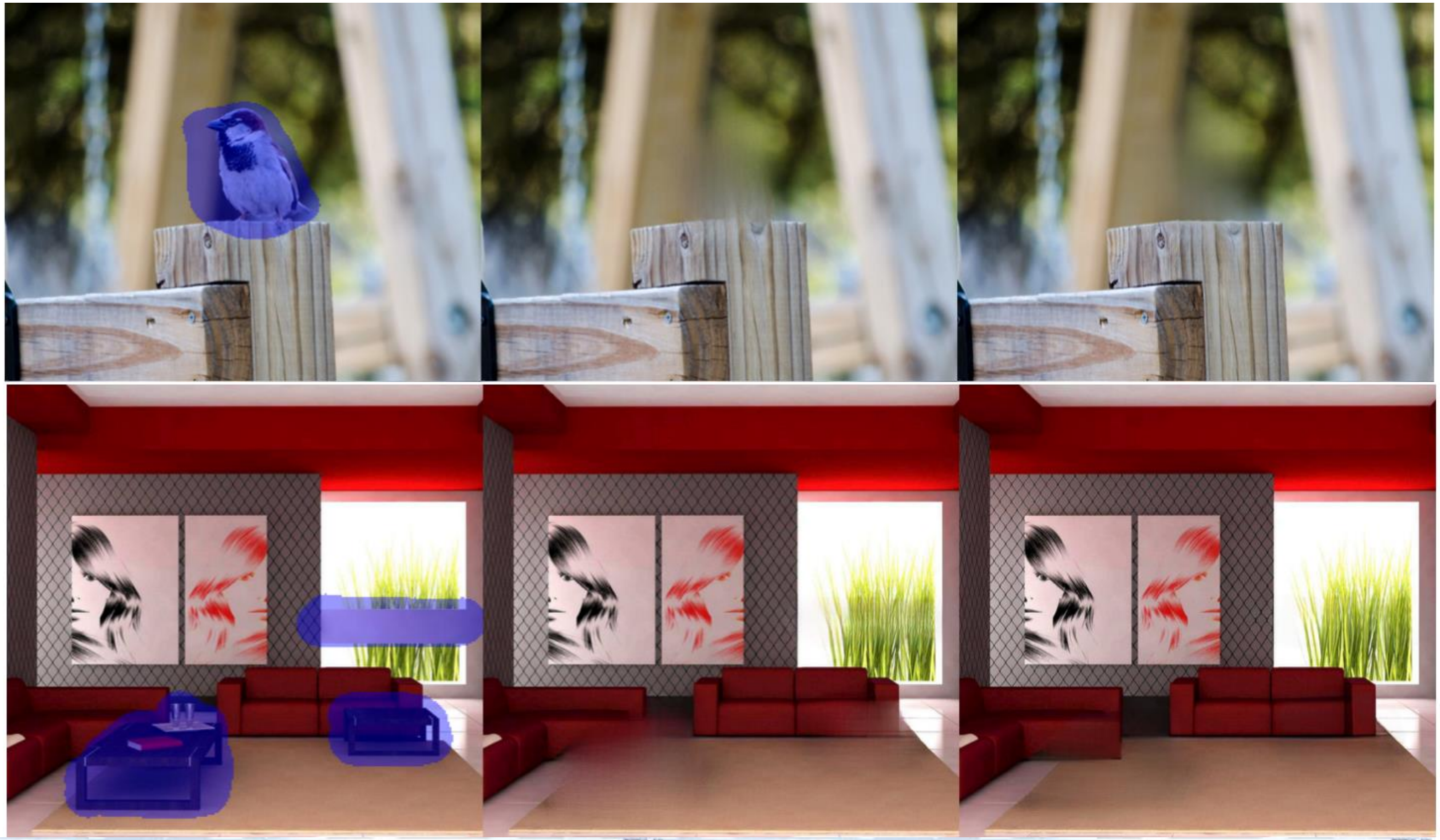




[1] M.-H. Guo, C.-Z. Lu, Z.-N. Liu, M.-M. Cheng, and S.-M. Hu, “Visual attention network,” arXiv preprint arXiv:2202.09741, 2022

**High-resolution (1K, 2K)** object removal results compared with LaMa

From left to right:
masked image, LaMa, ZITS++



# How about Data-driven Priors?

# Learning Prior Feature and Attention Enhanced Image Inpainting

Chenjie Cao*, Qiaole Dong*, and Yanwei Fu†

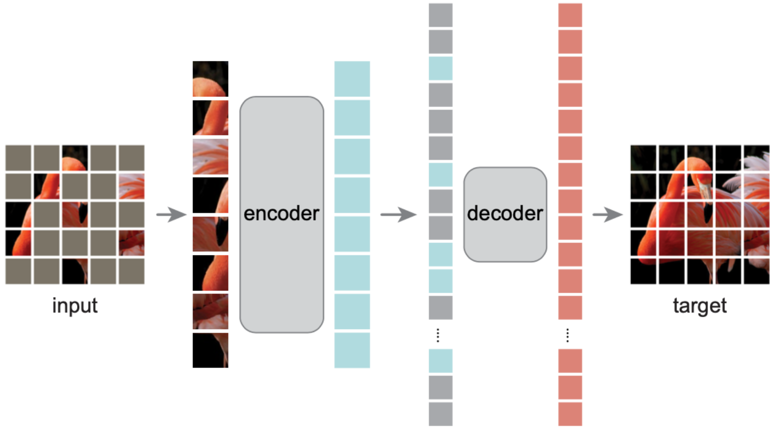
School of Data Science, Fudan University
{20110980001,qldong18,yanweifu}@fudan.edu.cn

**ECCV 2022**

Codes and pre-trained models are released in https://github.com/ewrfcas/MAE-FAR.

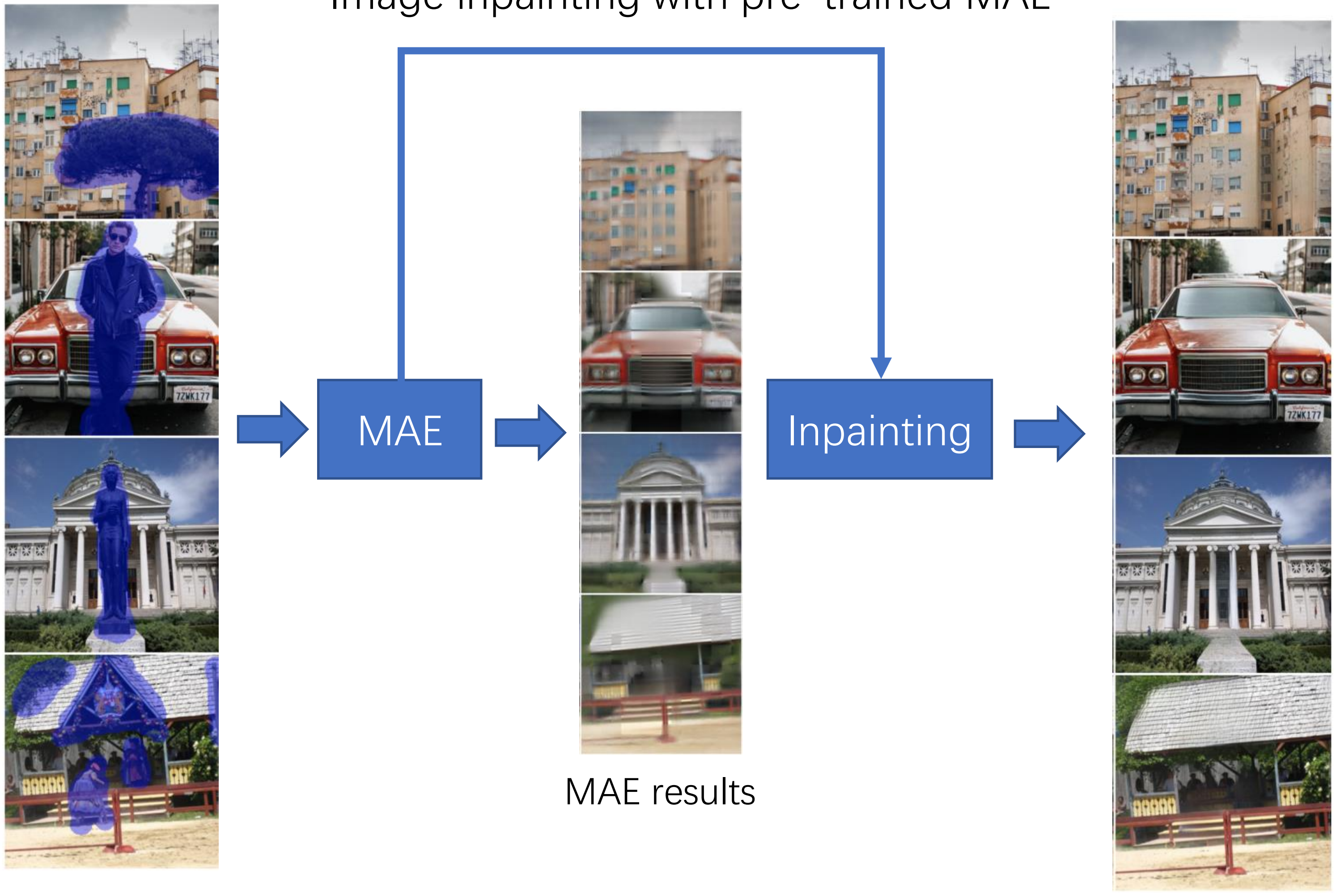

(a) Masked image (b) MAE (c) LaMa (d) Ours

# Data-driven Priors



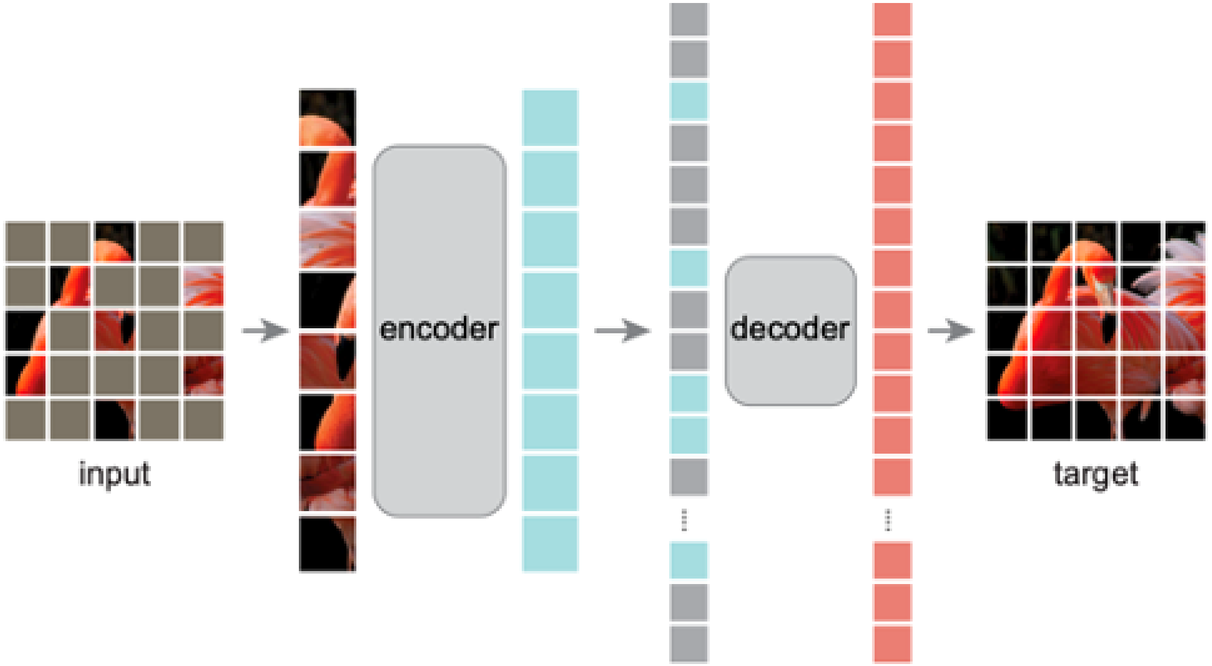

**Masked AutoEncoder (MAE)**[1]: A vision transformer that is pre-trained with 75% random masking prediction

Our model provides proper priors for Image inpainting with pre-trained MAE

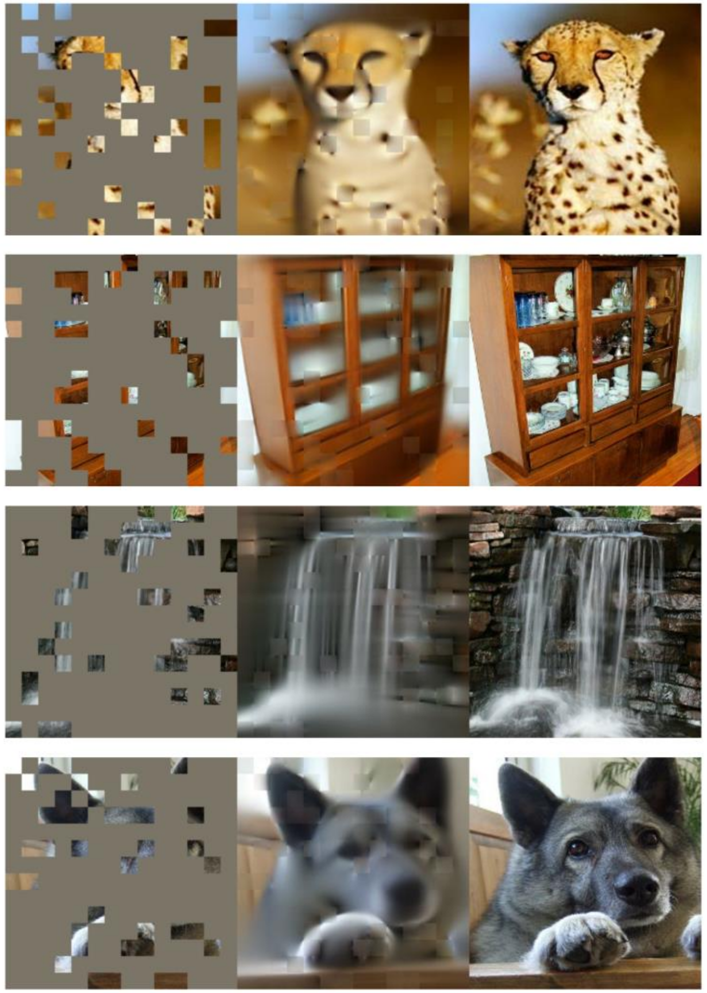



[1] He K, Chen X, Xie S, et al. Masked autoencoders are scalable vision learners[C]//CVPR2022: 16000-16009.

# MAE: Masked Autoencoders Are Scalable Vision Learners



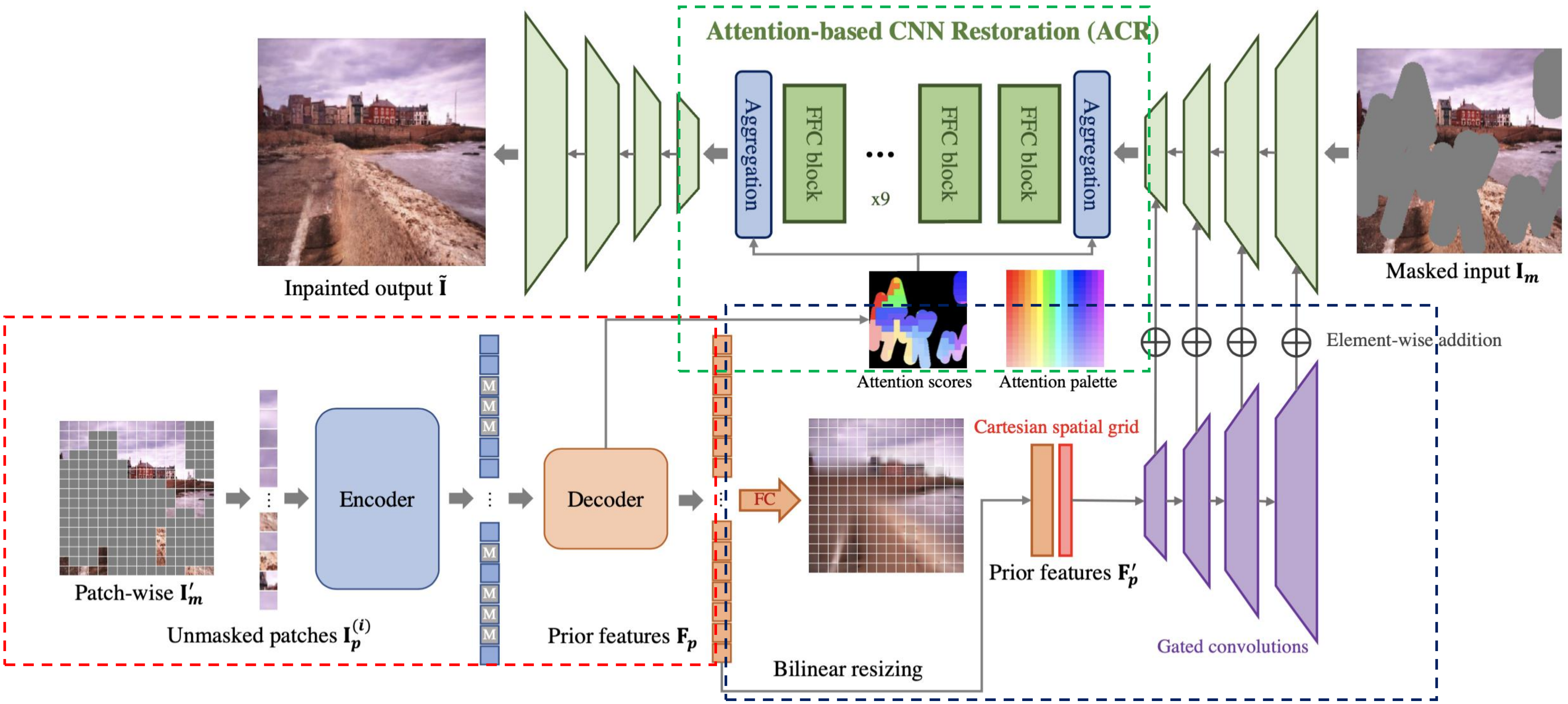

MAE structure



MAE Reconstruction

He, Kaiming, et al. "Masked autoencoders are scalable vision learners." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

# Method Overview



# Training Setting of MAE for Inpainting
## Masking Strategy

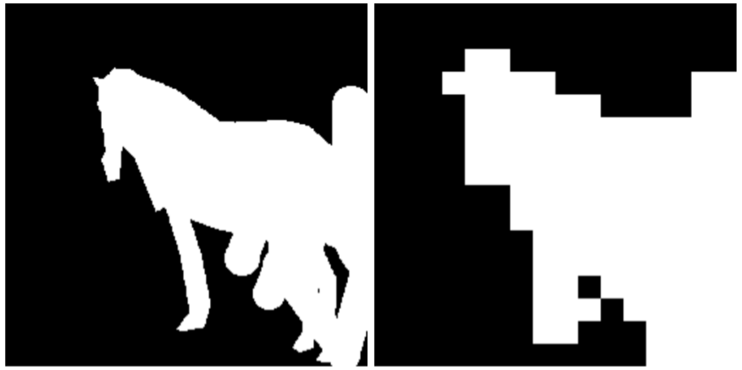
Noisy and random masks are easier[1]

(a) Ground truth image (b) Random masked image (75%) (c) Continues mask (d) Resized continuous mask (e) Hybird masked image (75%)

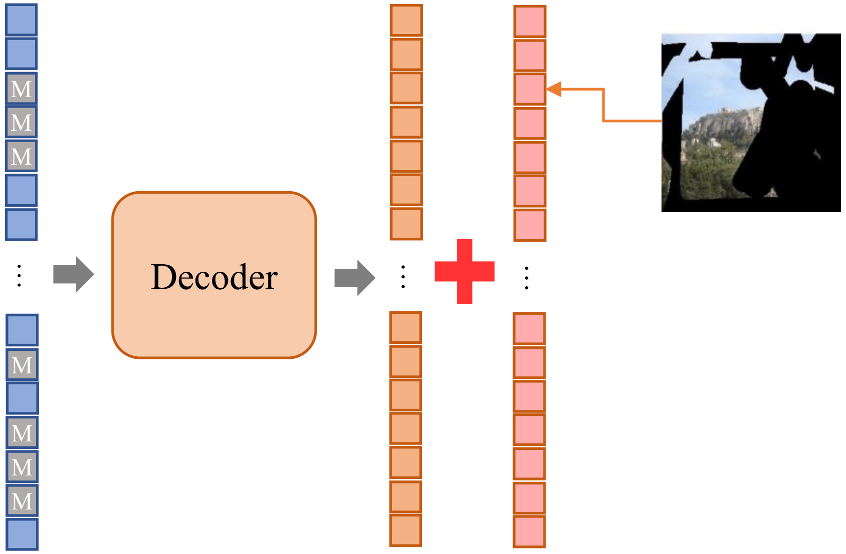
| MAE mask type | attention type | PSNR↑ | SSIM↑ | FID↓ | LPIPS↓ |
|---|---|---|---|---|---|
| mixed | no attention | 24.34 | 0.860 | 26.84 | 0.117 |
| mixed | trainable CA | 24.13 | 0.859 | 26.99 | 0.123 |
| random | prior attention | 24.39 | 0.861 | 26.25 | 0.117 |
| mixed | prior attention | **24.51** | **0.864** | **25.49** | **0.113** |

[1] Ntavelis, Evangelos, et al. "AIM 2020 challenge on image extreme inpainting." *European Conference on Computer Vision*. Springer, Cham, 2020.

# Training Setting of MAE for Inpainting
## Finetuning for Partially Masked Patches



(c) Continues mask (d) Resized continuous mask



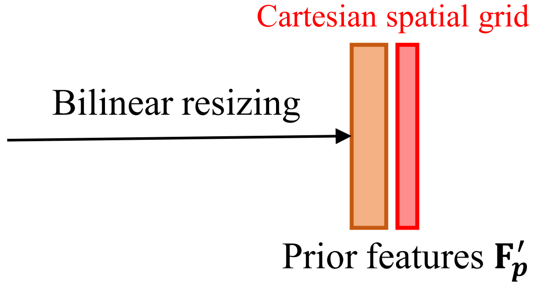

Partially masked embedding

(a) Origin input (b) Masked input (c) MAE w/o ft (d) MAE w/o ft+ACR (e) MAE with ft (d) MAE with ft+ACR

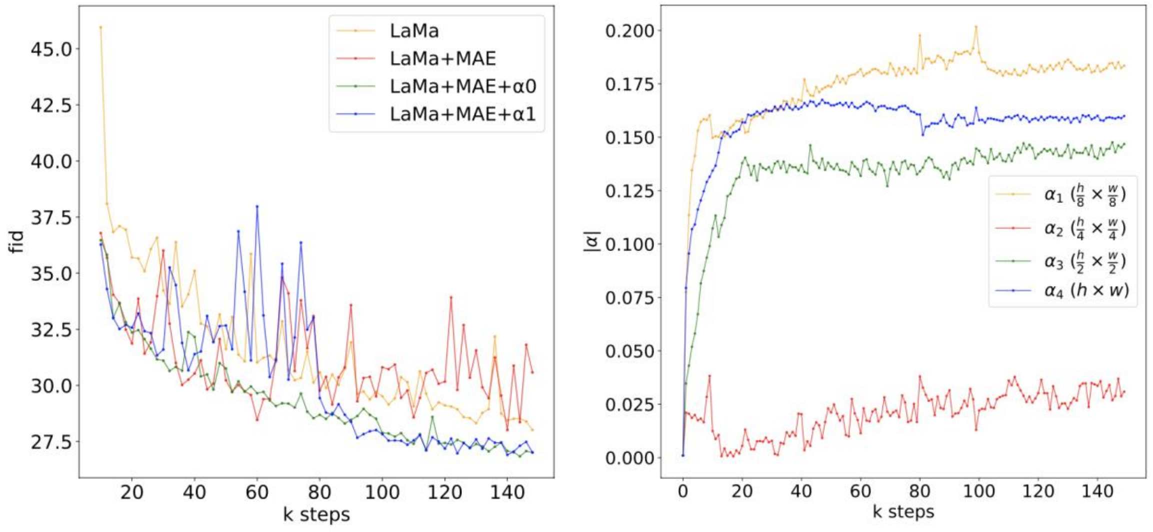
Overfitting

# Attention-based CNN Restoration (ACR)

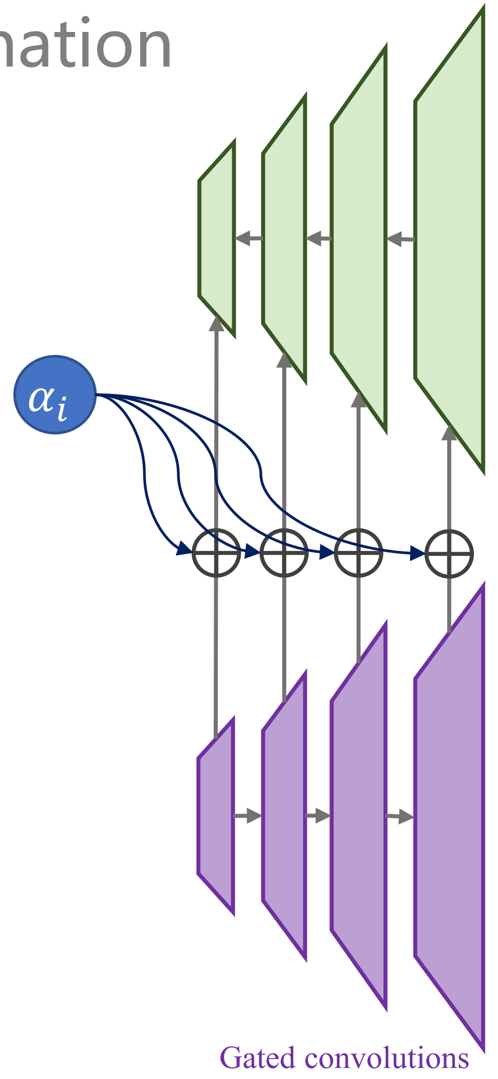
## Prior Features Upsampling and Prior Features Combination




$$\mathbf{F}'_p = \mathrm{Concat}(\mathrm{BilinearResize}(\mathbf{F}_p), \mathbf{C}) \in \mathbb{R}^{\frac{h}{8} \times \frac{w}{8} \times (d+2)},$$




(a) (b)



# Attention-based CNN Restoration (ACR)

## Prior Attentions from MAE vs. Contextual Attention

Masked input | MAE output | MAE attention | Contextual attention | Masked input | MAE output | MAE attention | Contextual attention

$$\cos_{u,m} = \left\langle \frac{\mathbf{F}_u}{||\mathbf{F}_u||}, \frac{\mathbf{F}_m}{||\mathbf{F}_m||} \right\rangle$$

$$\mathbf{R}_{u,m} = \mathrm{softmax}_u(\cos_{u,m}),$$

Contextual Attention

$$\mathbf{R}^{(l)}_{u,m} = \mathrm{softmax}\left(\frac{\mathbf{Q}^{(l)}{\mathbf{K}^{(l)}}^T}{\sqrt{d}} - inf \cdot \mathbf{M}\right),$$

$$\mathbf{R}_p = \frac{\sum_{l=1}^{L} \mathbf{R}^{(l)}_{u,m}}{L}, L = 8.$$

Prior Attention

# Qualitative results

(a) Masked image (b) EC (c) Co-Mod (d) LaMa (e) CTSDG (f) ICT (g) Ours

256x256 in Places2

(a) Masked input (b) HiFill (c) Co-Mod (d) LaMa (e) Ours (f) MAE

512x512 in Places2

# Qualitative results of faces and 1k images

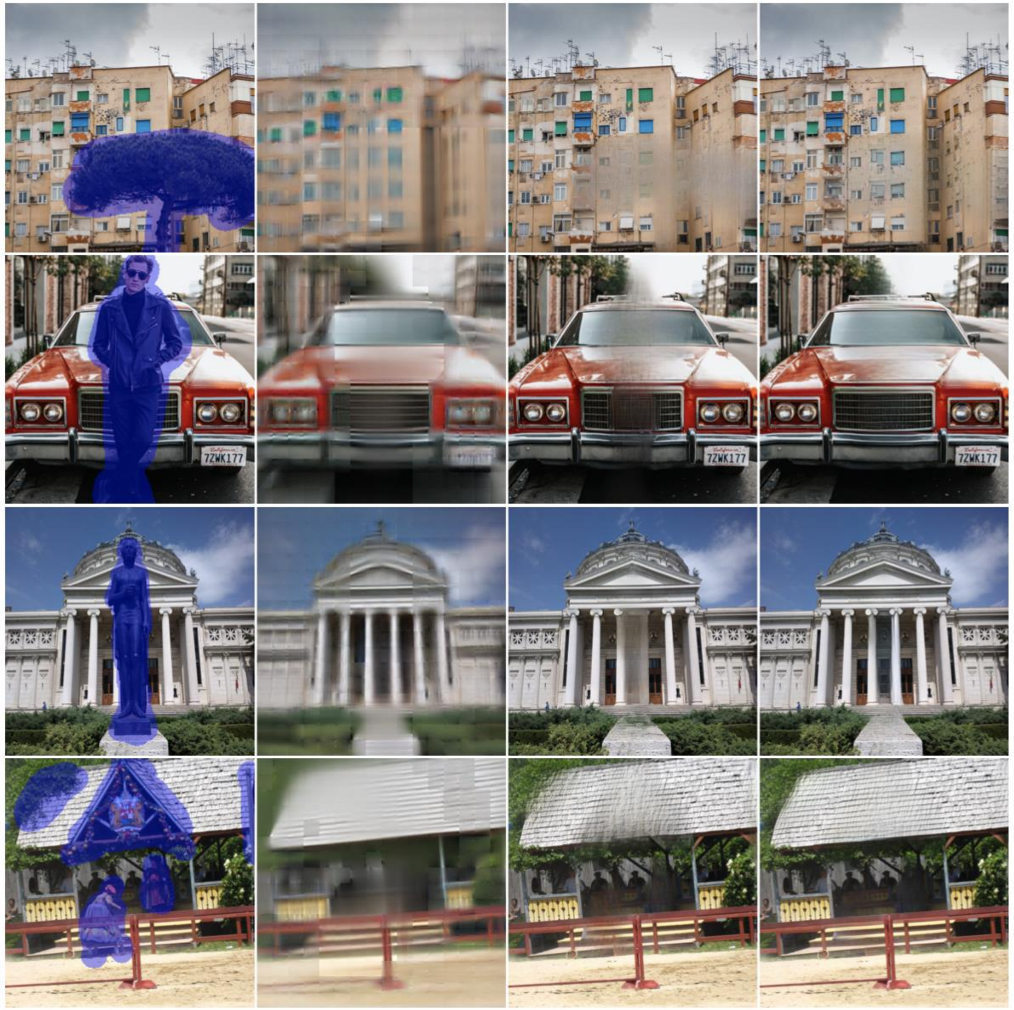
(A) 256x256 FFHQ

(a) Masked input (b) Co-Mod (c) LaMa (d) Ours

(B) 1024x1024 results

(a) Masked input (b) MAE output (c) LaMa (d) Ours

# More High-Resolution Results



(a) Masked input (b) MAE (c) LaMa (d) Ours

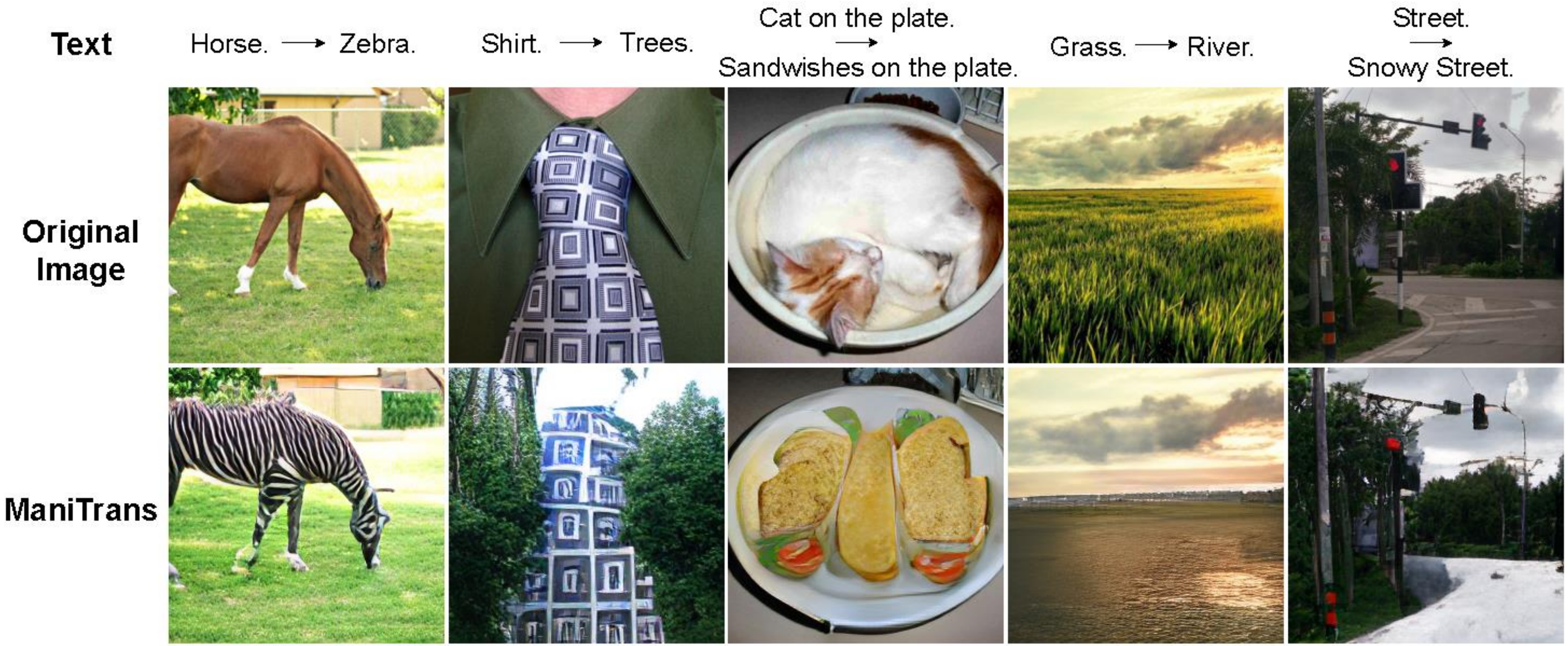
(a) Masked input (b) MAE (c) LaMa (d) Ours

Can we combine the priors with textual-conditions?

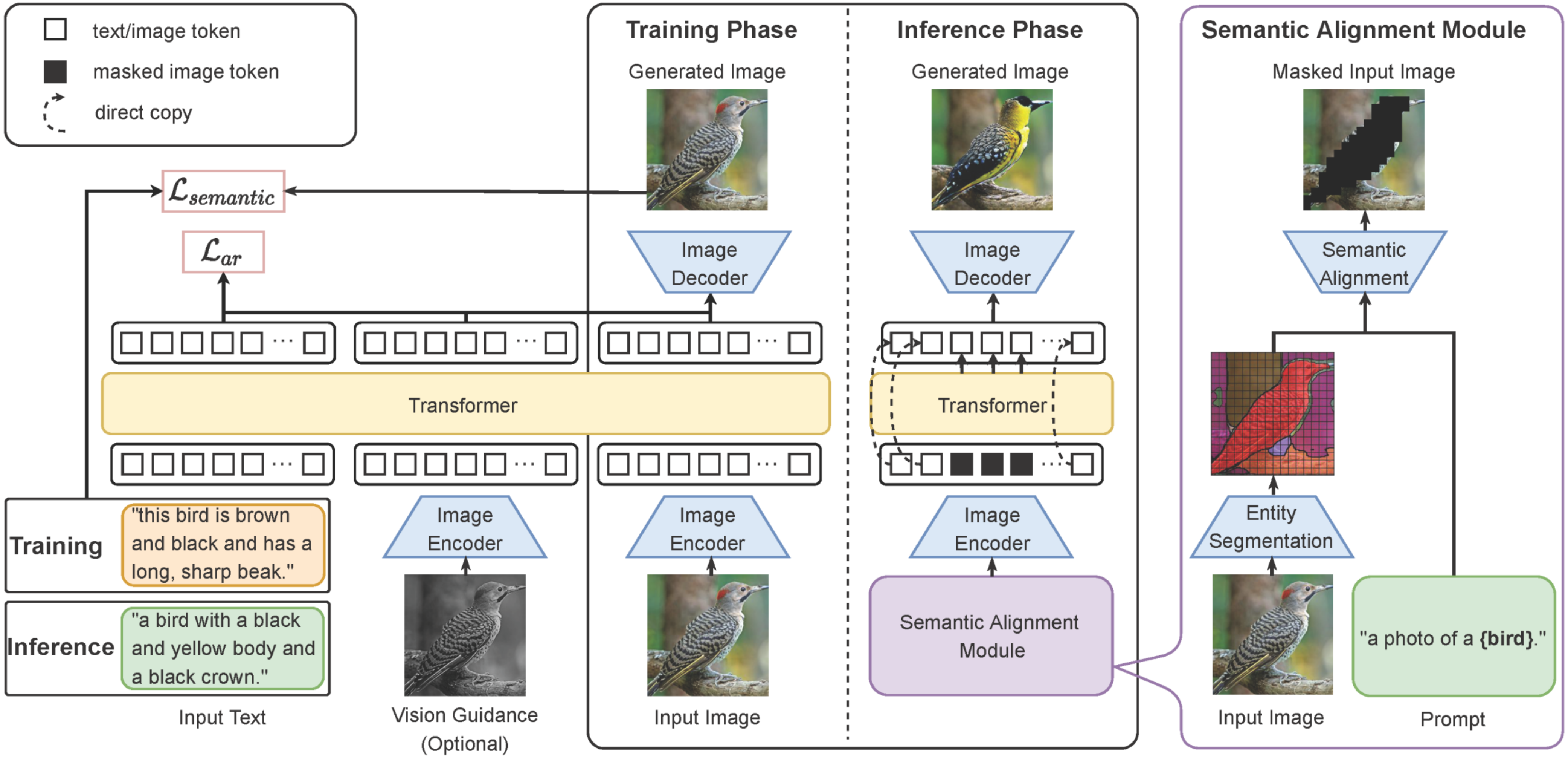
# ManiTrans: Entity-Level Text-Guided Image Manipulation via Token-wise Semantic Alignment and Generation

Jianan Wang[1] Guansong Lu[2] Hang Xu[2] Zhenguo Li[2] Chunjing Xu[2] Yanwei Fu[1]
[1]School of Data Science, Fudan University [2]Huawei Noah's Ark Lab
{jawang19, yanweifu}@fudan.edu.cn {luguansong, xu.hang, li.zhenguo, xuchunjing}@huawei.com

**CVPR 2022 (Oral)**



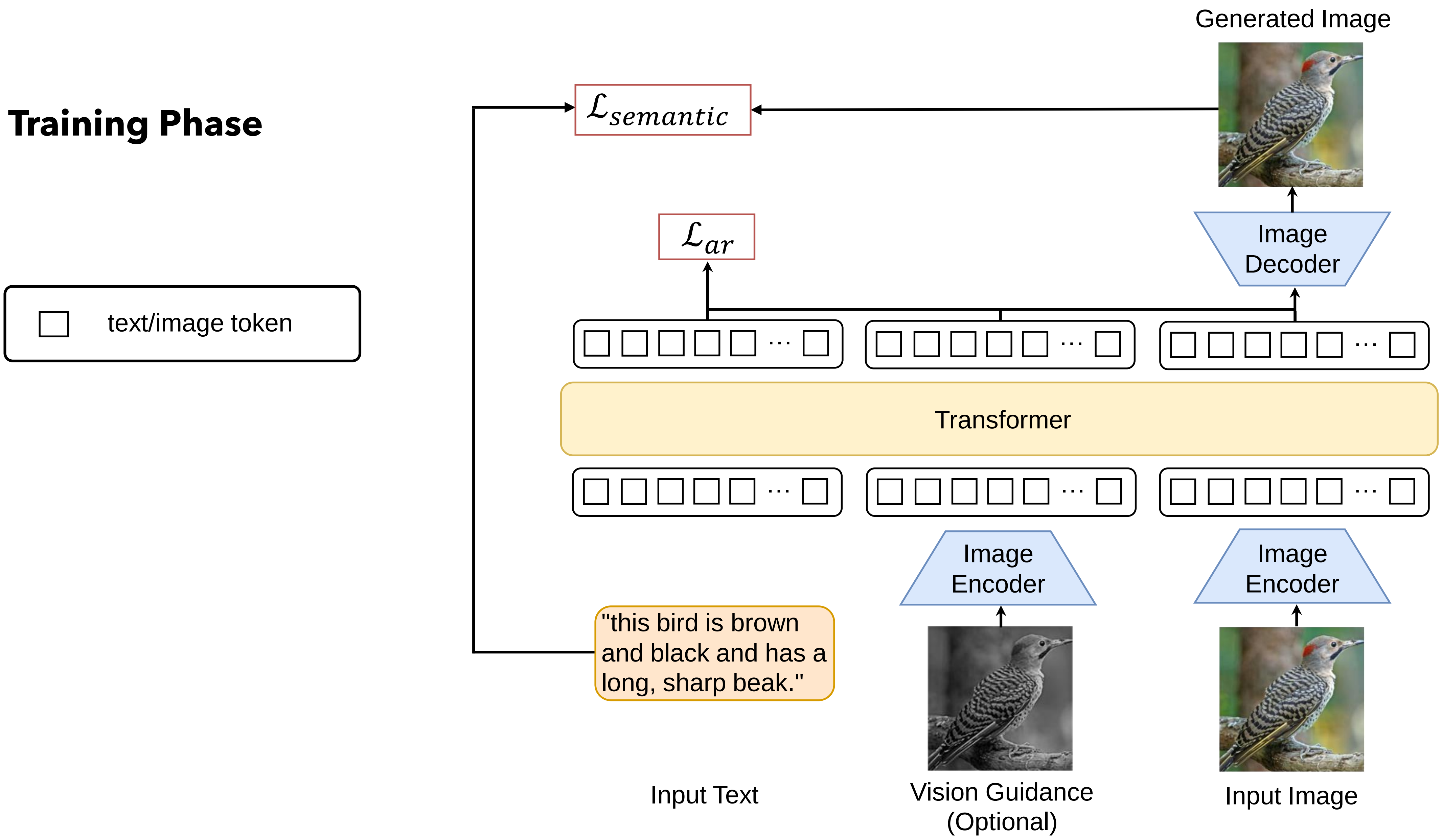
# ManiTrans



text/image token
masked image token
direct copy
$\mathcal{L}_{semantic}$
$\mathcal{L}_{ar}$
Training Phase
Generated Image
Image Decoder
Inference Phase
Generated Image
Image Decoder
Semantic Alignment Module
Masked Input Image
Semantic Alignment
Transformer
Transformer
Training
"this bird is brown and black and has a long, sharp beak."
Inference
"a bird with a black and yellow body and a black crown."
Input Text
Image Encoder
Vision Guidance (Optional)
Image Encoder
Input Image
Image Encoder
Semantic Alignment Module
Entity Segmentation
Input Image
"a photo of a **{bird}**."
Prompt

# ManiTrans

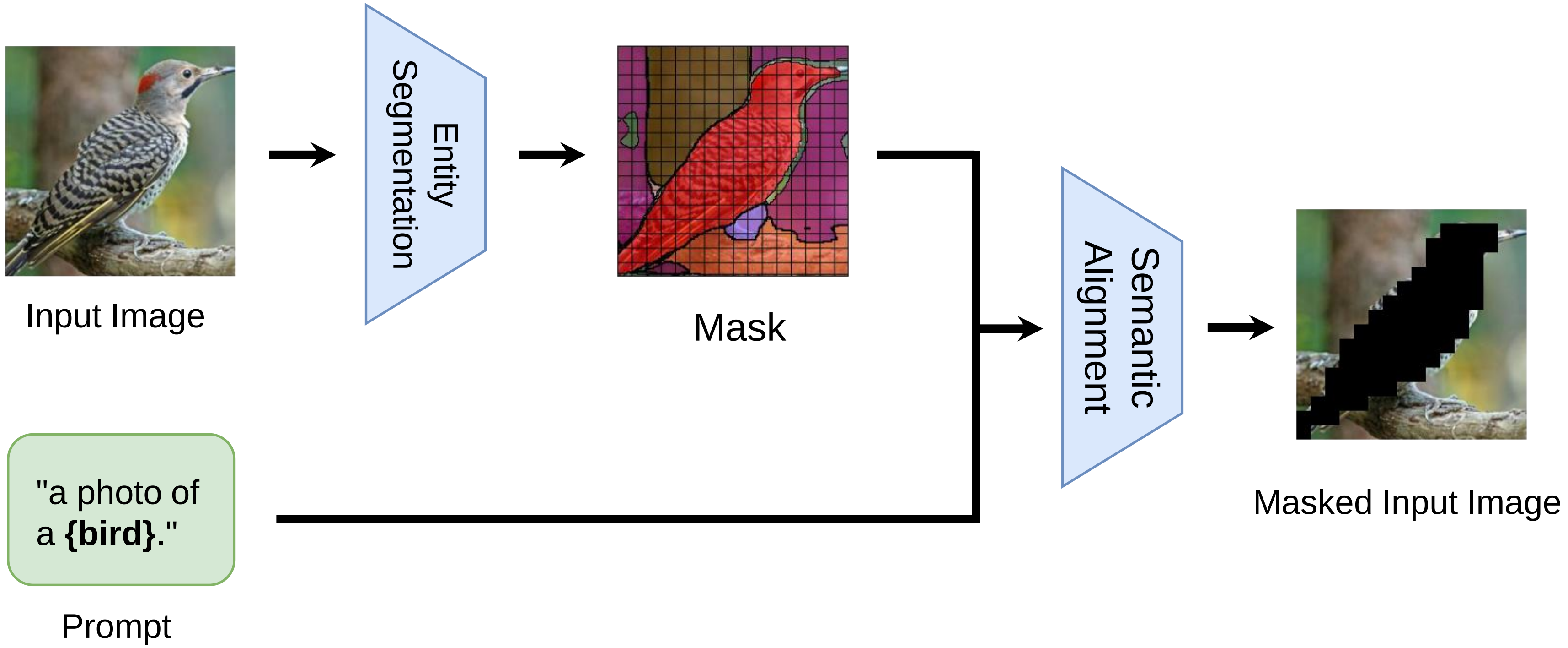
**Training Phase**



Mani

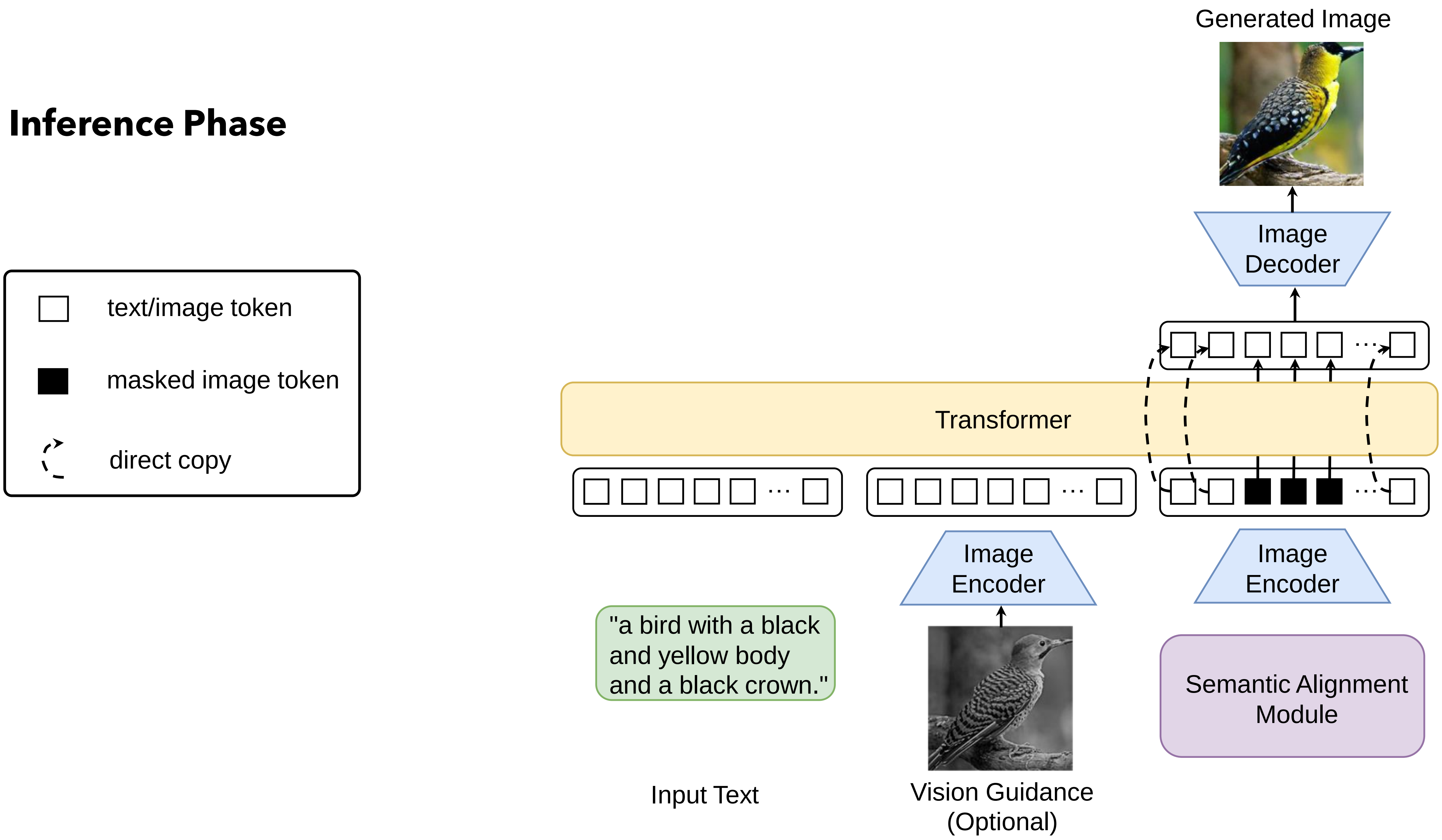
Semantic Alignment Module

**Semantic Alignment Module**

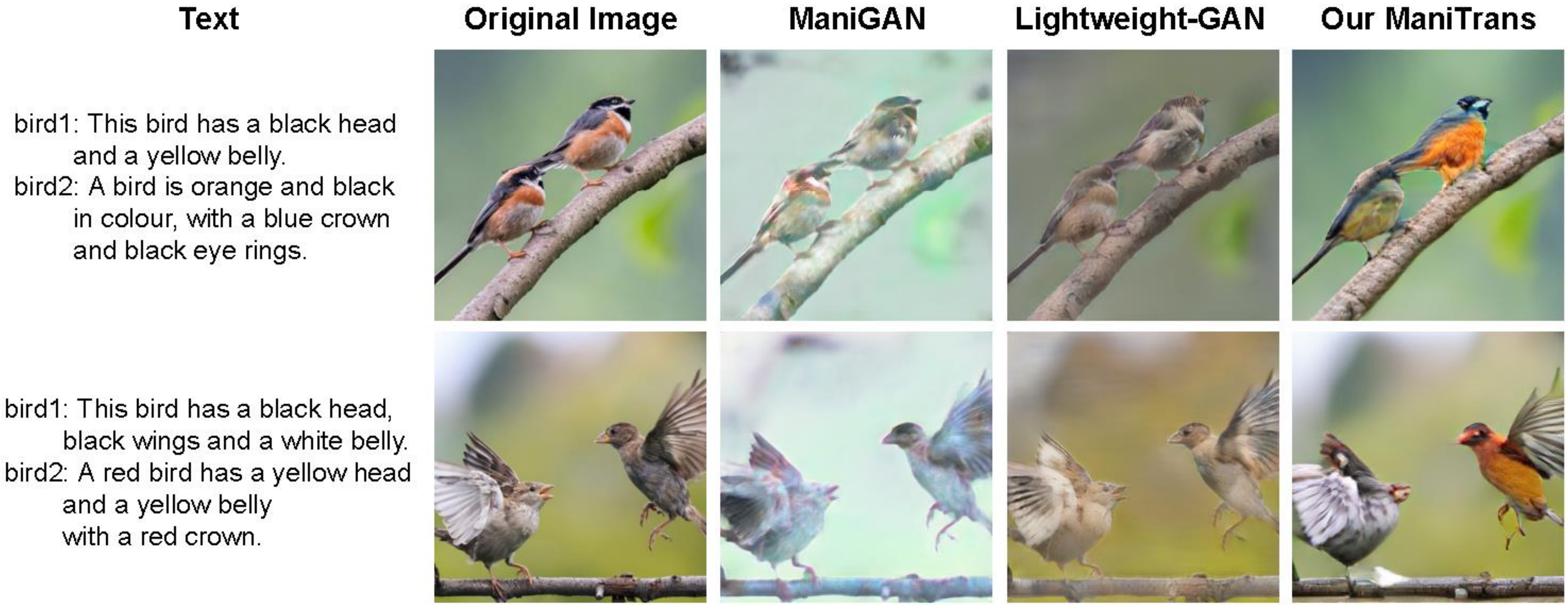


Mani

## Inference Phase



# Main Results



# Main Results

# Thanks!