Suppress Content Shift: Better Diffusion Features via Off-the-Shelf Generation Techniques

Benyuan Meng, Qianqian Xu*, Zitai Wang,

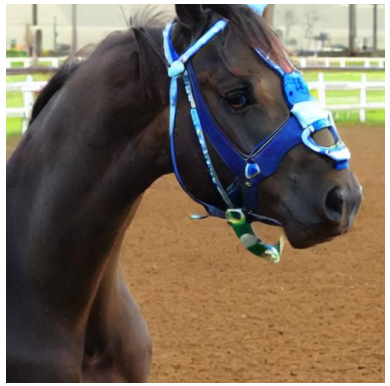
Zhiyong Yang, Xiaochun Cao, Qingming Huang*

Benyuan Meng

2024.11

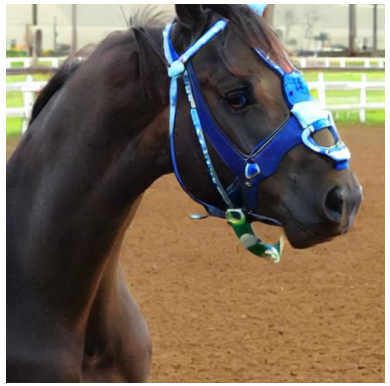
Background

Basic Generation




Background

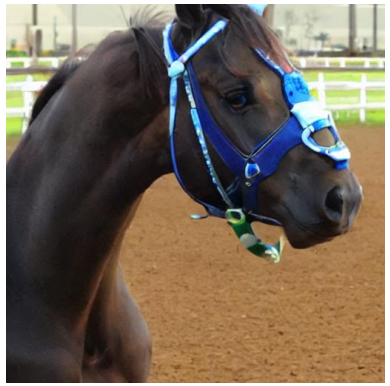
Basic Generation




Advanced Generation:
Variation, Inpainting, Personalization…

Background

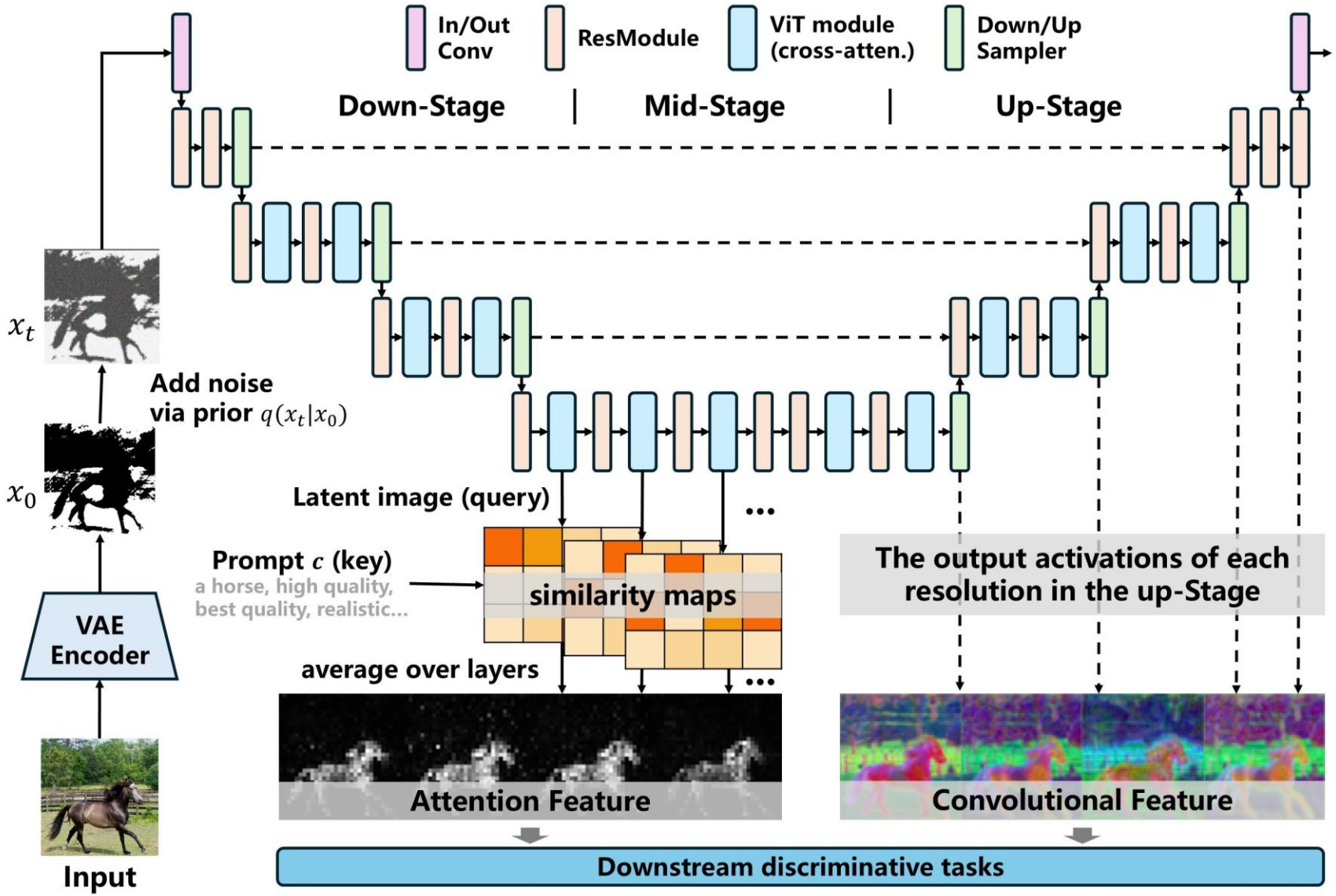
Basic Generation




Advanced Generation:
Variation, Inpainting, Personalization…

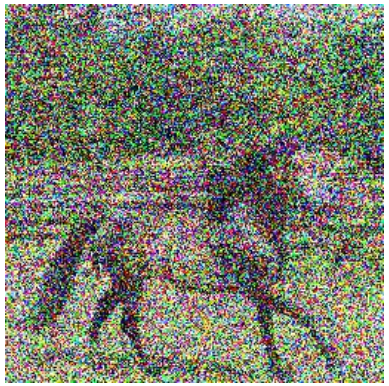
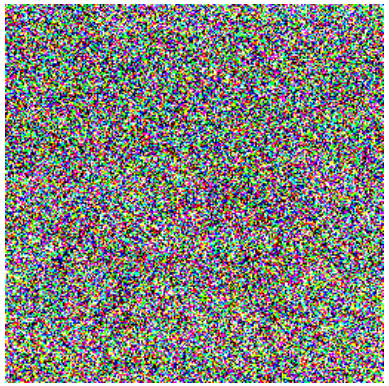
Generation → $p(x, y)$ → $p(y|x)$ → Discrimination

Background


In/Out Conv
ResModule
ViT module (cross-atten.)
Down/Up Sampler
Down-Stage
Mid-Stage
Up-Stage
x_t
Add noise via prior $q(x_t|x_0)$
x_0
VAE Encoder
Input
Latent image (query)
Prompt c (key)
a horse, high quality, best quality, realistic...
similarity maps
average over layers
Attention Feature
The output activations of each resolution in the up-Stage
Convolutional Feature
Downstream discriminative tasks

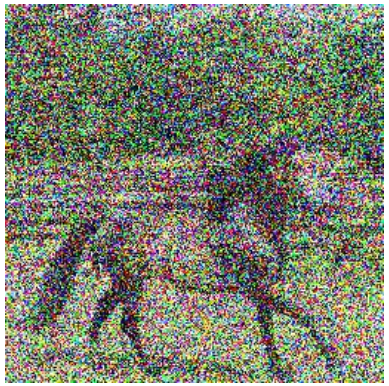
Key Observation

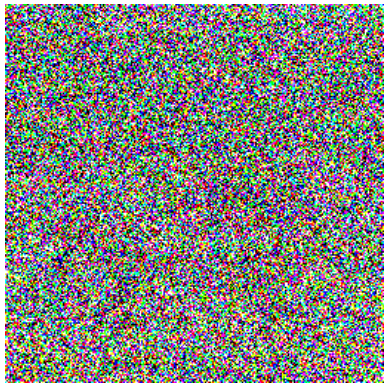


Key Observation

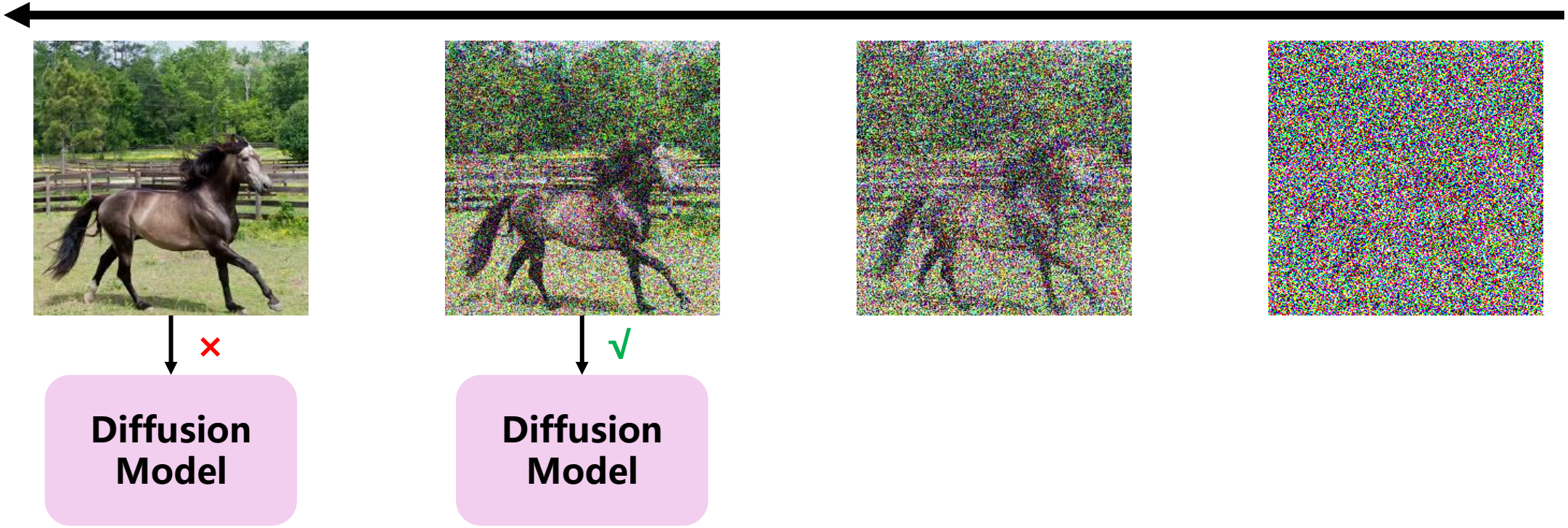
Key Observation





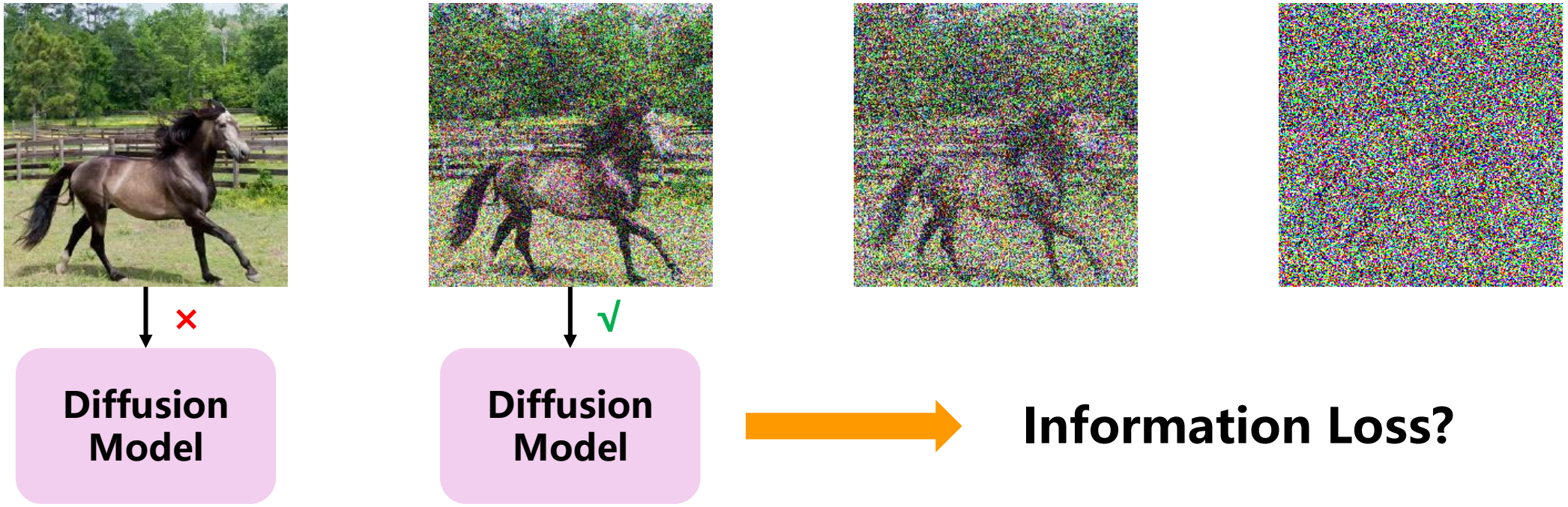
Key Observation

Key Observation



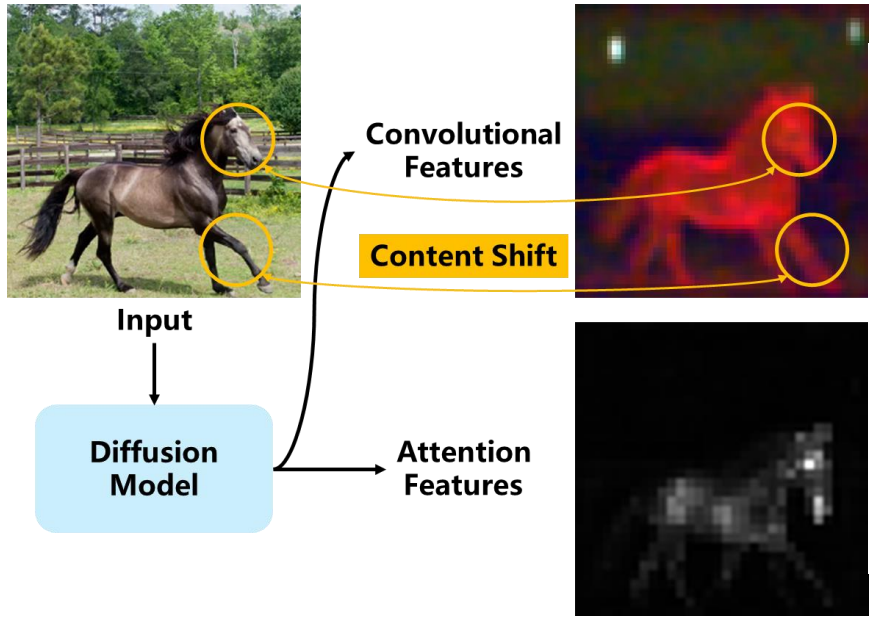
- **Adding noises to input images is widely considered necessary to avoid singularity of clean inputs.**

Key Observation



- **Adding noises to input images is widely considered necessary to avoid singularity of clean inputs.**

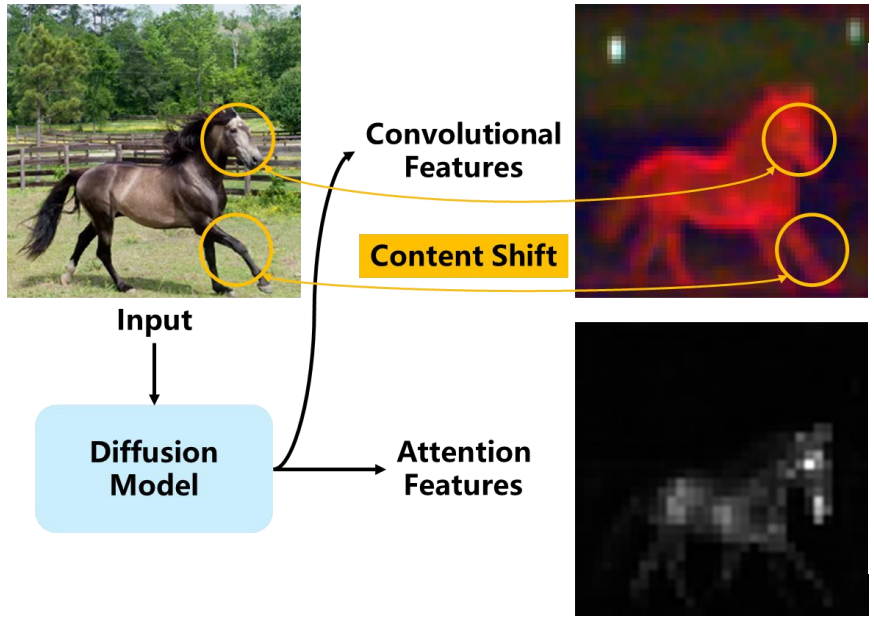
Key Observation



Dataset	Low Quality	Neutral	High Quality
Horse-21	58.90	59.03	59.33
CIFAR10	91.67	91.44	91.21

- Qualitative: perceivable content difference.
- Quantitative: non-negligible performance degradation.

Key Observation



Dataset	Low Quality	Neutral	High Quality
Horse-21	58.90	59.03	59.33
CIFAR10	91.67	91.44	91.21

- Qualitative: perceivable content difference.
- Quantitative: non-negligible performance degradation.
- Content Shift

Key Observation

Less Noises
+ Less Content Shift
- More Singularity

More Noises
+ Less Singularity
- More Content Shift

Key Observation

Less Noises
+ Less Content Shift
- More Singularity

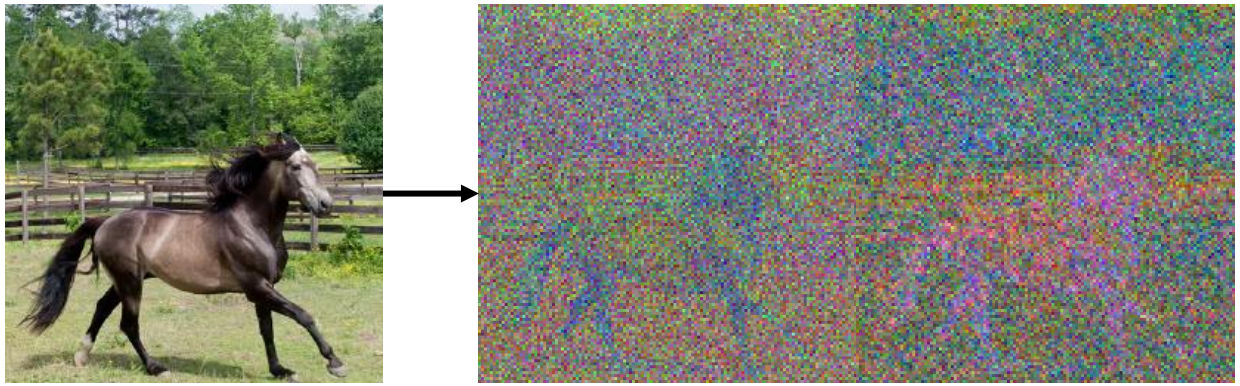
More Noises
+ Less Singularity
- More Content Shift

Additional Control Method

Better Feature

- Not seeking balance.
- But a new way to suppress content shift while maintaining noise strength.

Understanding Content Shift

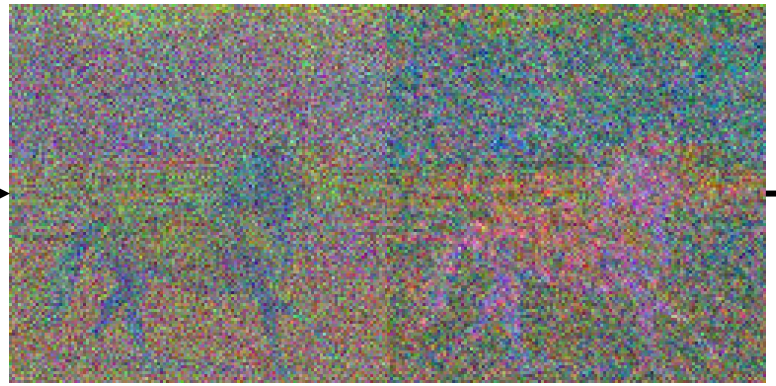


Original Image

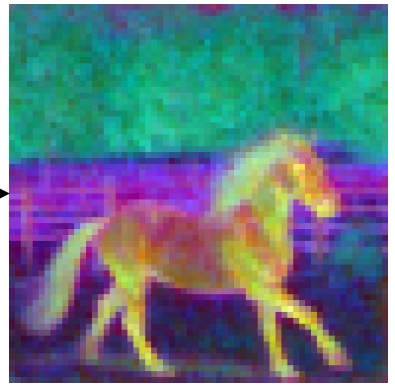
Noisy Activations
in Early Sections

Understanding Content Shift

Original Image

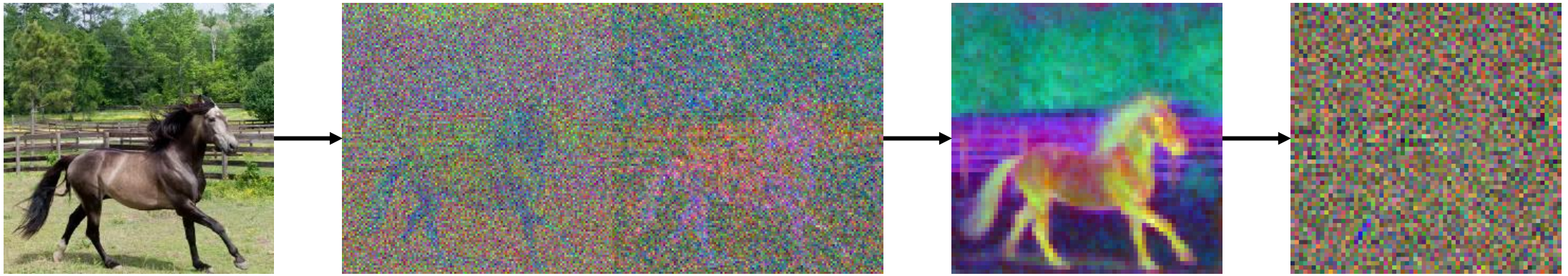


Noisy Activations
in Early Sections



Reconstructed
Clean Representations

Understanding Content Shift



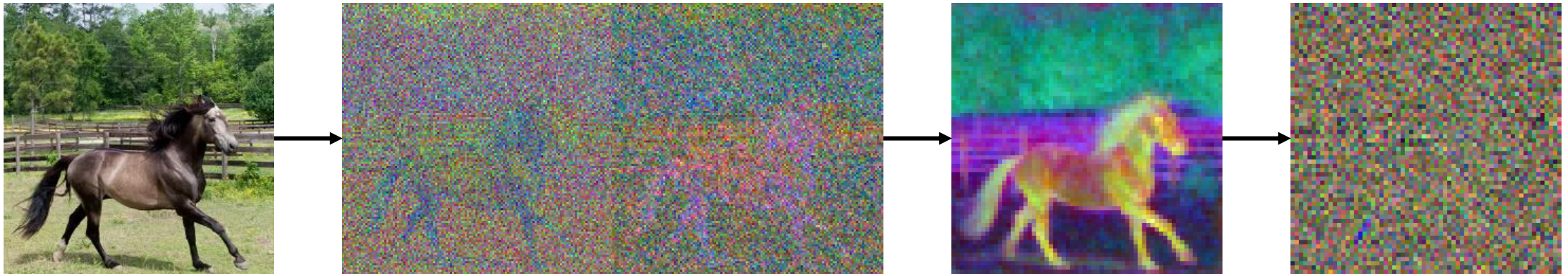
Original Image

Noisy Activations
in Early Sections

Reconstructed
Clean Representations

Output Noises

Understanding Content Shift



Original Image

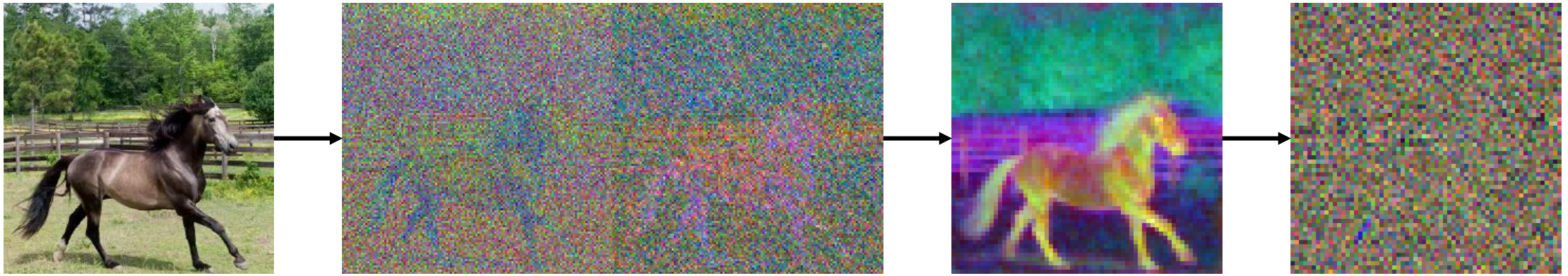
Noisy Activations in Early Sections

Reconstructed Clean Representations

Output Noises

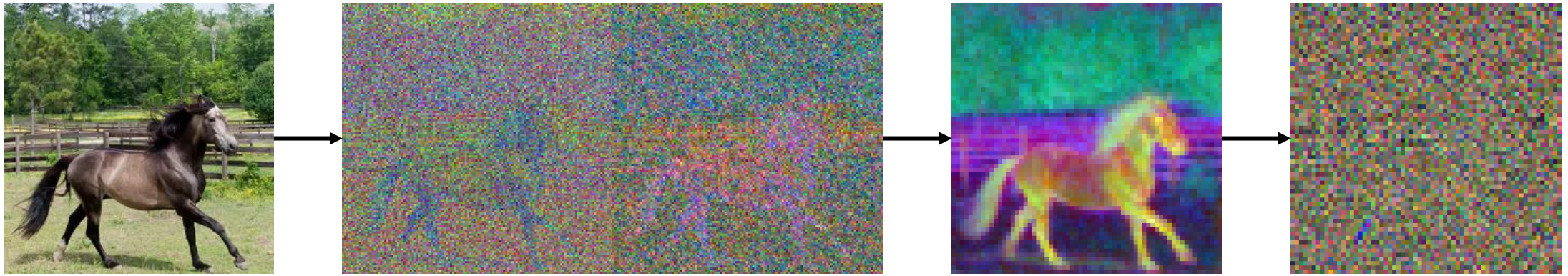
- **Diffusion models have learnt to reconstruct clean representations from noisy inputs.**

Understanding Content Shift



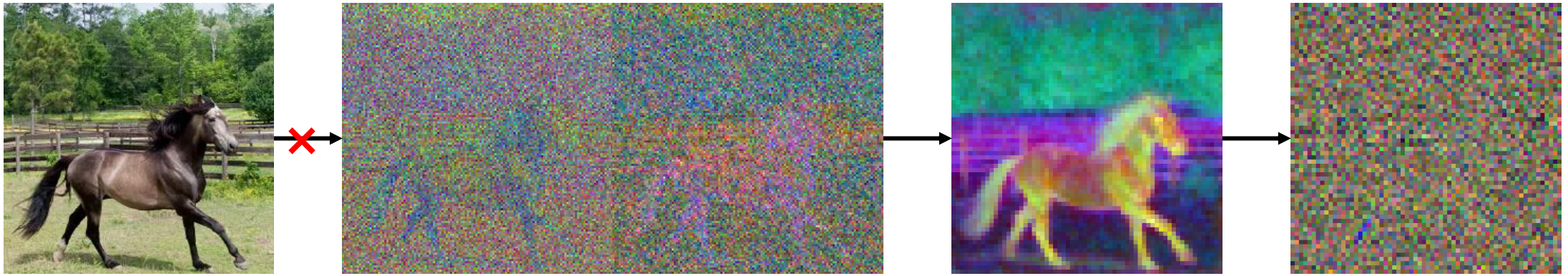
- **Diffusion models have learnt to reconstruct clean representations from noisy inputs.**
- **Reconstructed details are created based on context, not guaranteed to stay true to the original image.**

Suppressing Content Shift



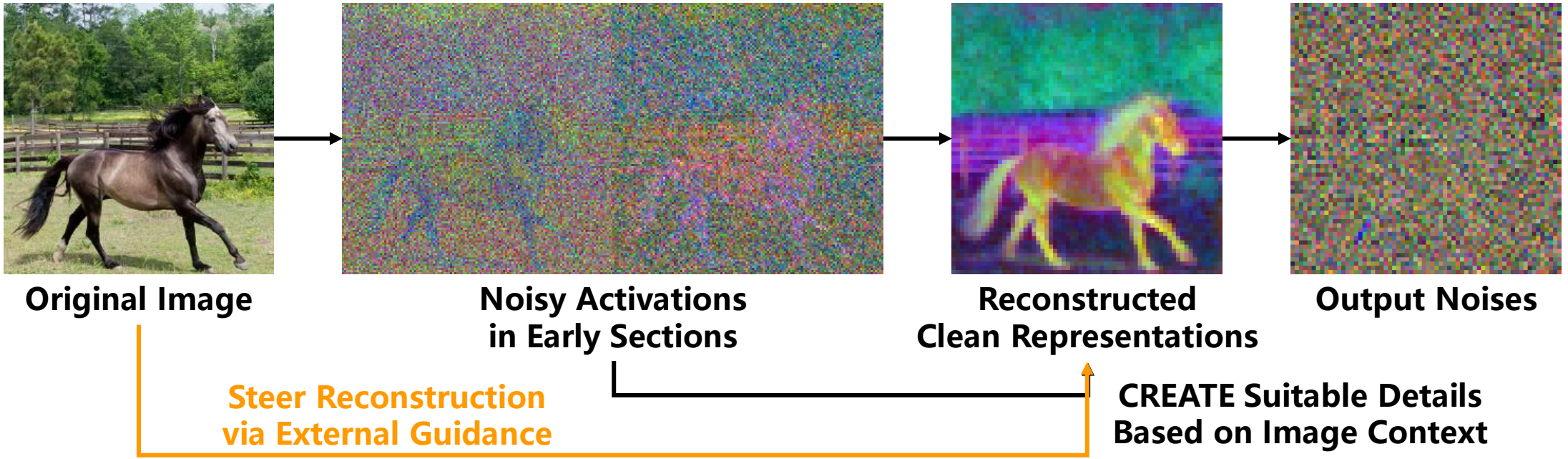
- **Steer reconstruction?**

Suppressing Content Shift



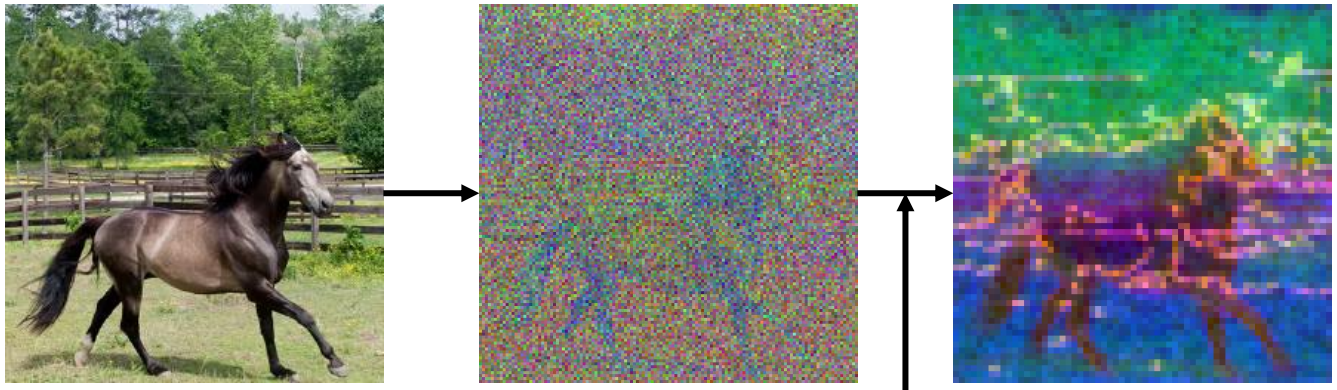
- **Steer reconstruction?**

Suppressing Content Shift



- **Steer reconstruction?**
- **Bypass the normal route to directly inject guidance information.**

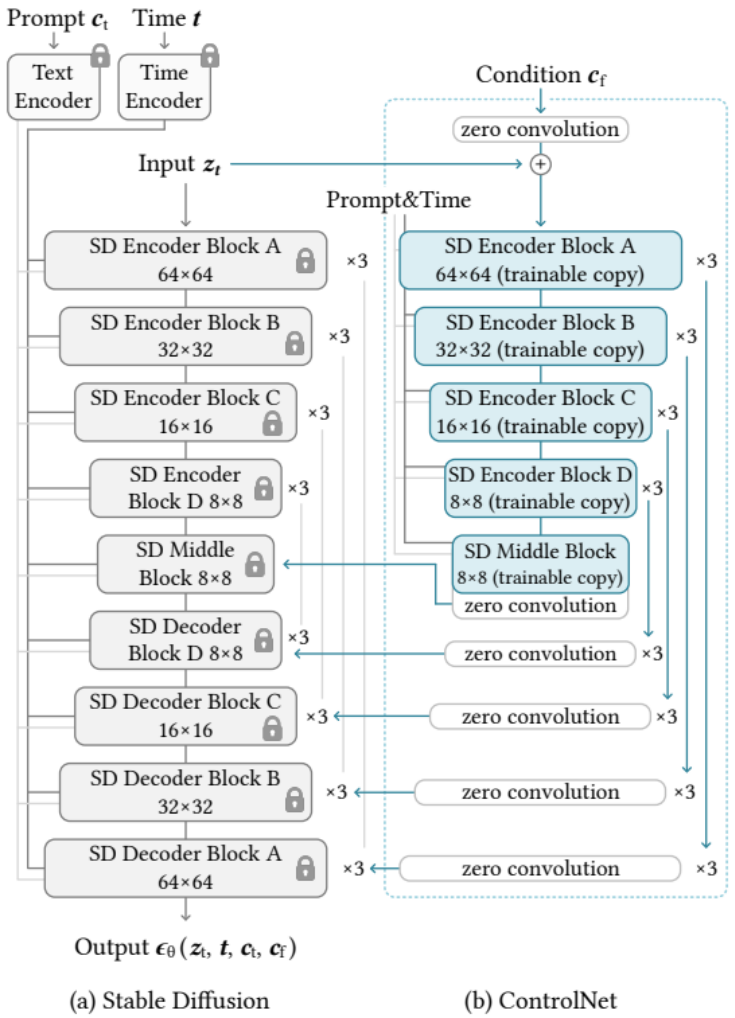
Suppressing Content Shift



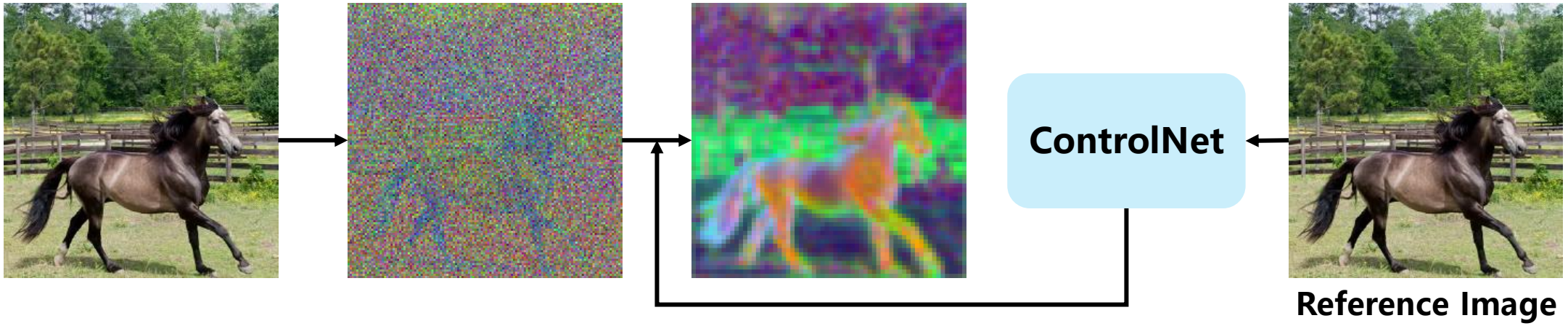
Reference Image

ControlNet

Taken from Zhang, Lvmin, Anyi Rao, and Maneesh Agrawala. "Adding conditional control to text-to-image diffusion models." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023.

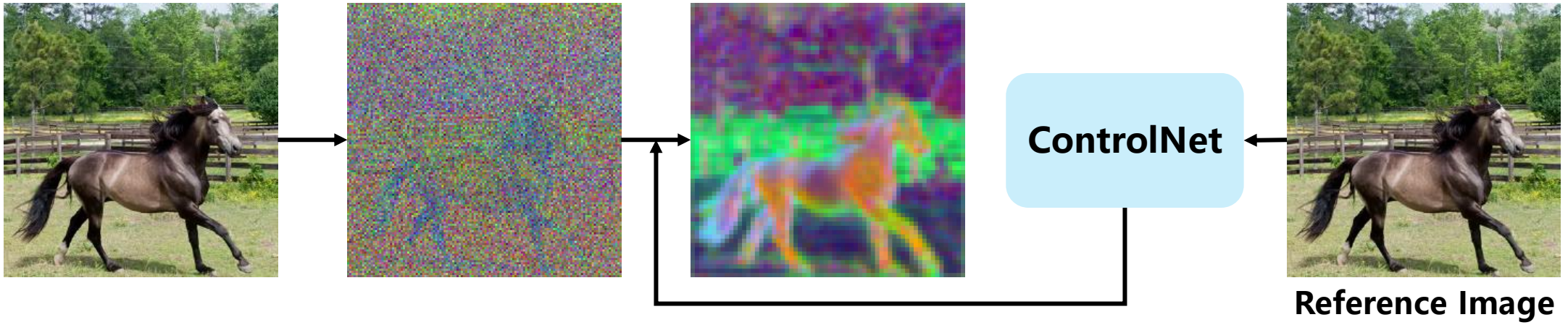


Suppressing Content Shift



- **Utilize generation control techniques to steer reconstruction to stay close to the original image.**
- **Named GATE**
  - **(GenerAtion Techniques Enhanced diffusion feature)**

Suppressing Content Shift



- **Utilize generation control techniques to steer reconstruction to stay close to the original image.**
  - **Fine-Grained Prompts**
  - **ControlNet**
  - **LoRA**
- **Feature amalgamation to harness enhanced diversity.**

Results

Table 1: The results of semantic correspondence (left, PCK@0.1) and label-scarce semantic segmentation (right, mIoU↑). **Red** for the best result and **blue** for the runner-up.

Method		PCK_{img} ↑	PCK_{bbox} ↑
Non-DF	DINO	51.68	41.04
	DHPF	55.28	42.63
DF	DIFT	-	52.90
	DHF	72.56	64.61
Baseline	nn	61.15	51.66
	conv	**73.96**	**65.74**
GATE	nn	64.47	55.72
	conv	**76.60**	**69.10**

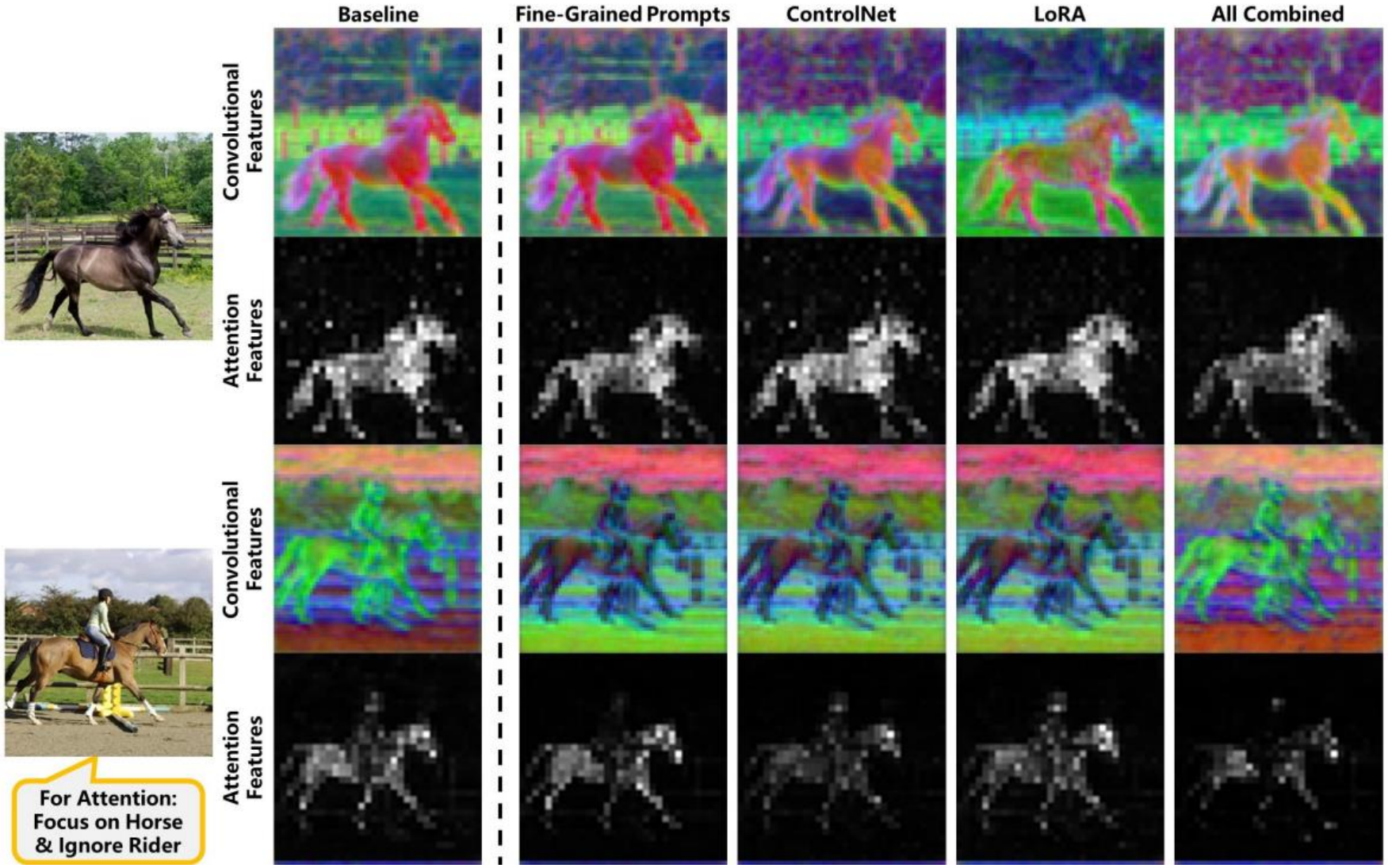
Method	Bedroom-28	Horse-21
ALAE	20.0 ± 1.0	–
GAN Inversion	13.9 ± 0.6	17.7 ± 0.4
GAN Encoder	22.4 ± 1.6	26.7 ± 0.7
SwAV	41.0 ± 2.3	51.7 ± 0.5
SwAVw2	42.4 ± 1.7	54.0 ± 0.9
MAE	45.0 ± 2.0	63.4 ± 1.4
DatasetGAN	31.3 ± 2.7	45.4 ± 1.4
DatasetDDPM	47.9 ± 2.9	60.8 ± 1.0
DDPM	**49.4 ± 1.9**	**65.0 ± 0.8**
GATE	**53.1 ± 2.7**	**67.2 ± 1.1**

Results

Table 2: Results on the two standard semantic segmentation datasets, ADE20K and CityScapes. **Red** for the best result and **blue** for the runner-up.

Category	Method	ADE20K			CityScapes		
		mIoU↑	aAcc↑	mAcc↑	mIoU↑	aAcc↑	mAcc↑
SOTA	MaskCLIP	23.70	-	-	-	-	-
	ODISE	29.90	-	-	-	-	-
	VPD	**37.63**	**79.16**	**50.08**	**55.06**	**90.14**	**68.96**
Ours	GATE	**40.51**	**79.68**	**54.90**	**64.20**	**92.83**	**76.98**

Results


Baseline
Fine-Grained Prompts
ControlNet
LoRA
All Combined
Convolutional Features
Attention Features
Convolutional Features
Attention Features
For Attention:
Focus on Horse
& Ignore Rider

NEURAL INFORMATION PROCESSING SYSTEMS

Thanks for your listening!