Not All Diffusion Model Activations Have Been Evaluated as Discriminative Features

Benyuan Meng, Qianqian Xu*, Zitai Wang,

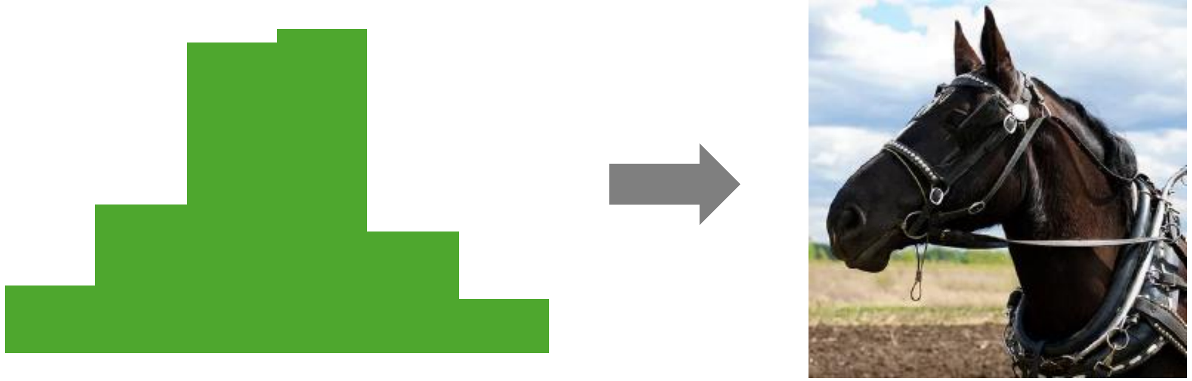
Xiaochun Cao, Qingming Huang*

Benyuan Meng

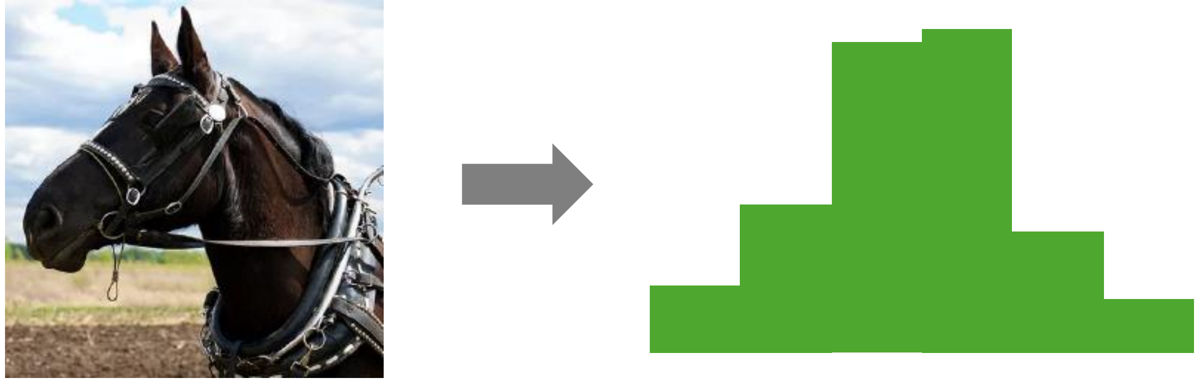
2024.11

Background

- **Generation** $p(x, y)$

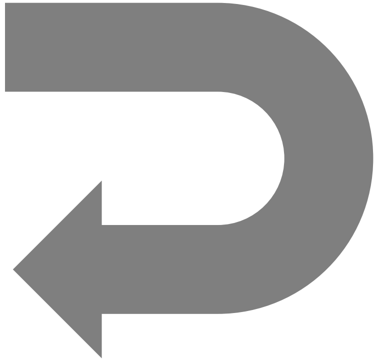


- **Discrimination** $p(y|x)$



Background

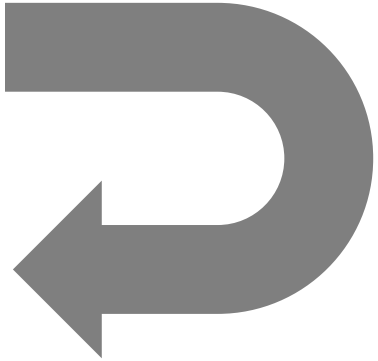
- **Generation** $p(x, y)$
- **Discrimination** $p(y|x)$



Generation Models for Discrimination

Background

- **Generation** $p(x,y)$
- **Discrimination** $p(y|x)$



Generation Models for Discrimination

GAN Liu, Xuanqing, and Cho-Jui Hsieh. "Rob-gan: Generator, discriminator, and adversarial attacker." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019.

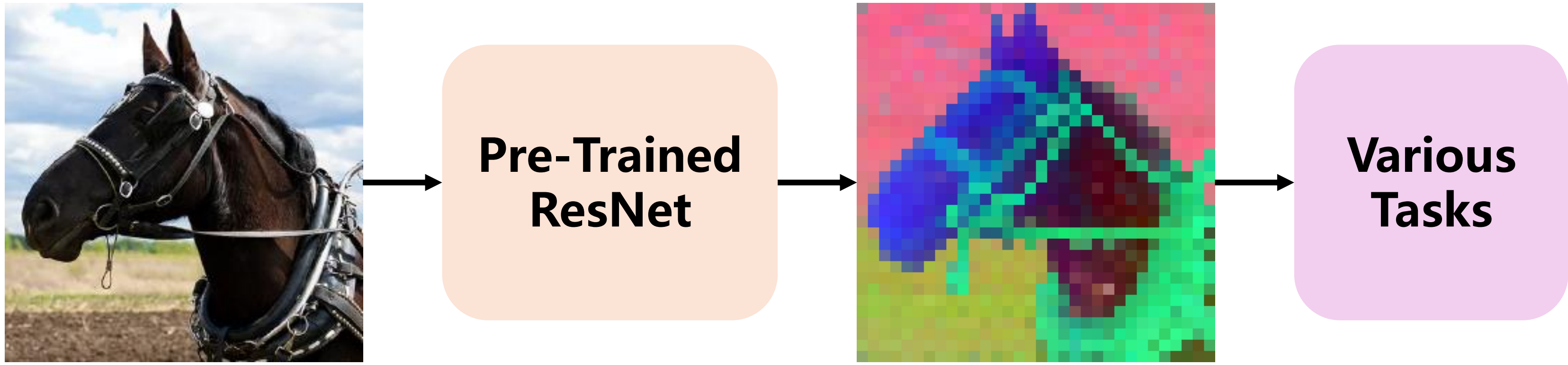
VAE Wu, Aming, and Cheng Deng. "Discriminating known from unknown objects via structure-enhanced recurrent variational autoencoder." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.

Background

- **Generation** $p(x, y)$
- **Discrimination** $p(y|x)$

Generation Models for Discrimination

Diffusion Models
Diffusion Feature

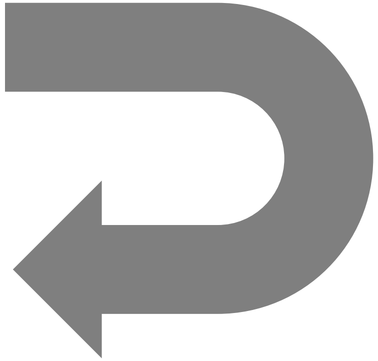


Pre-Trained ResNet

Various Tasks

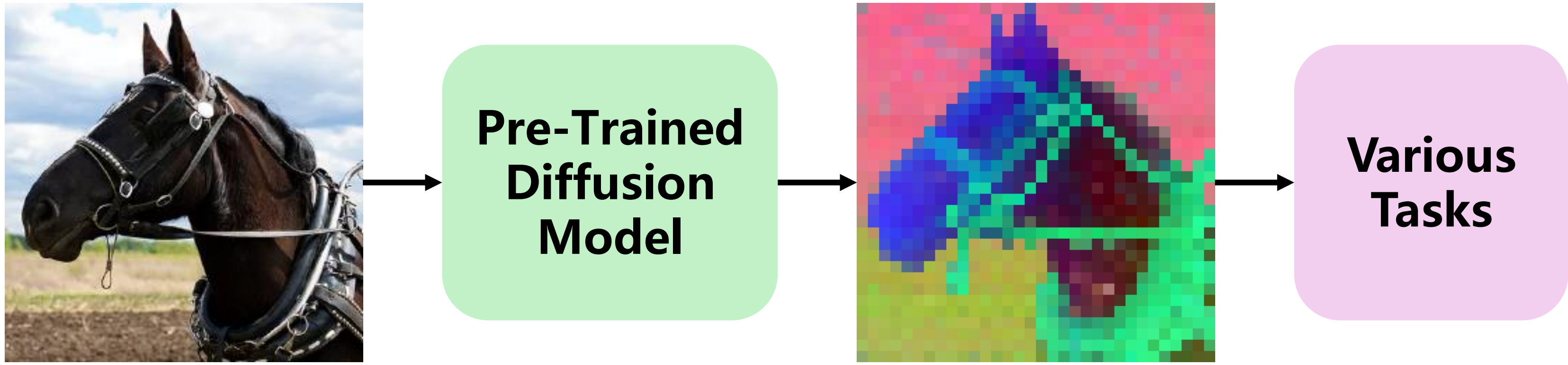
Background

- **Generation** $p(x, y)$
- **Discrimination** $p(y|x)$



Generation Models for Discrimination

Diffusion Models
Diffusion Feature



Key Observation

- **Popular diffusion models for diffusion feature study:**

Stable Diffusion v1.4

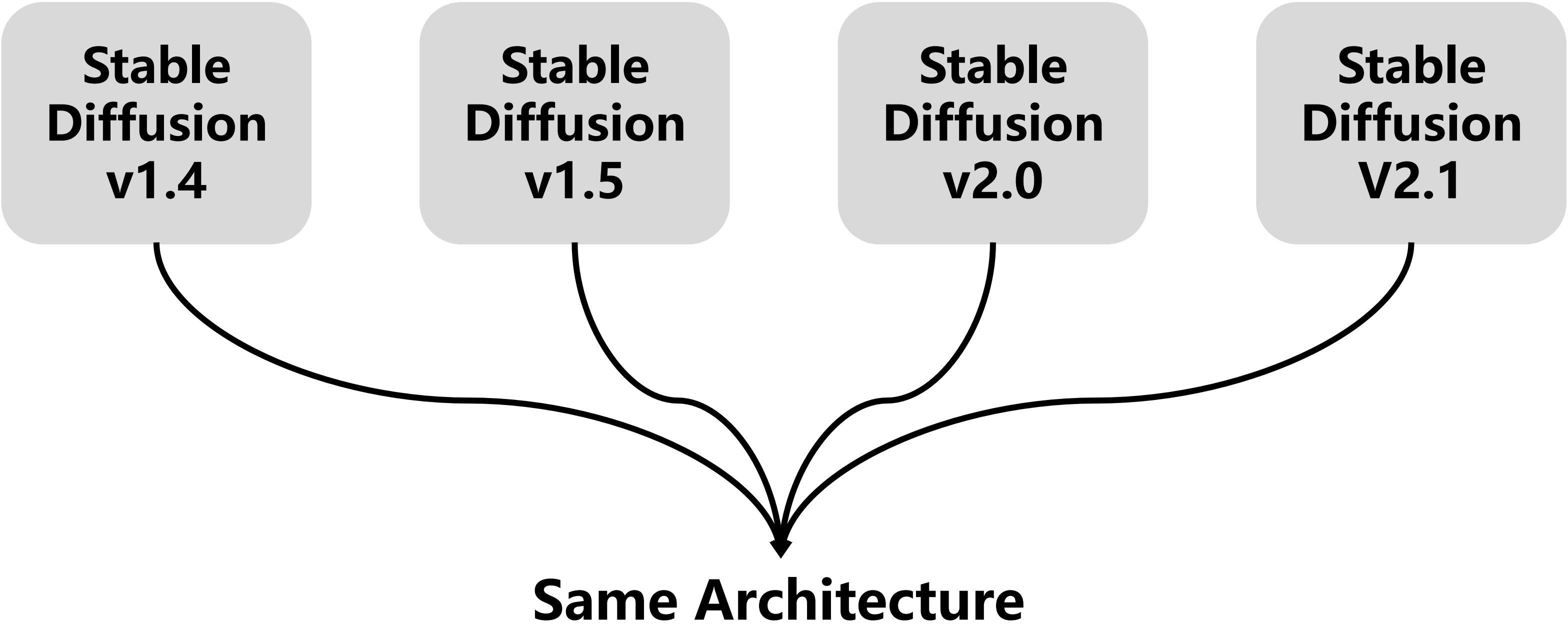
Stable Diffusion v1.5

Stable Diffusion v2.0

Stable Diffusion V2.1

Key Observation

- **Popular diffusion models for diffusion feature study:**



Stable Diffusion v1.4

Stable Diffusion v1.5

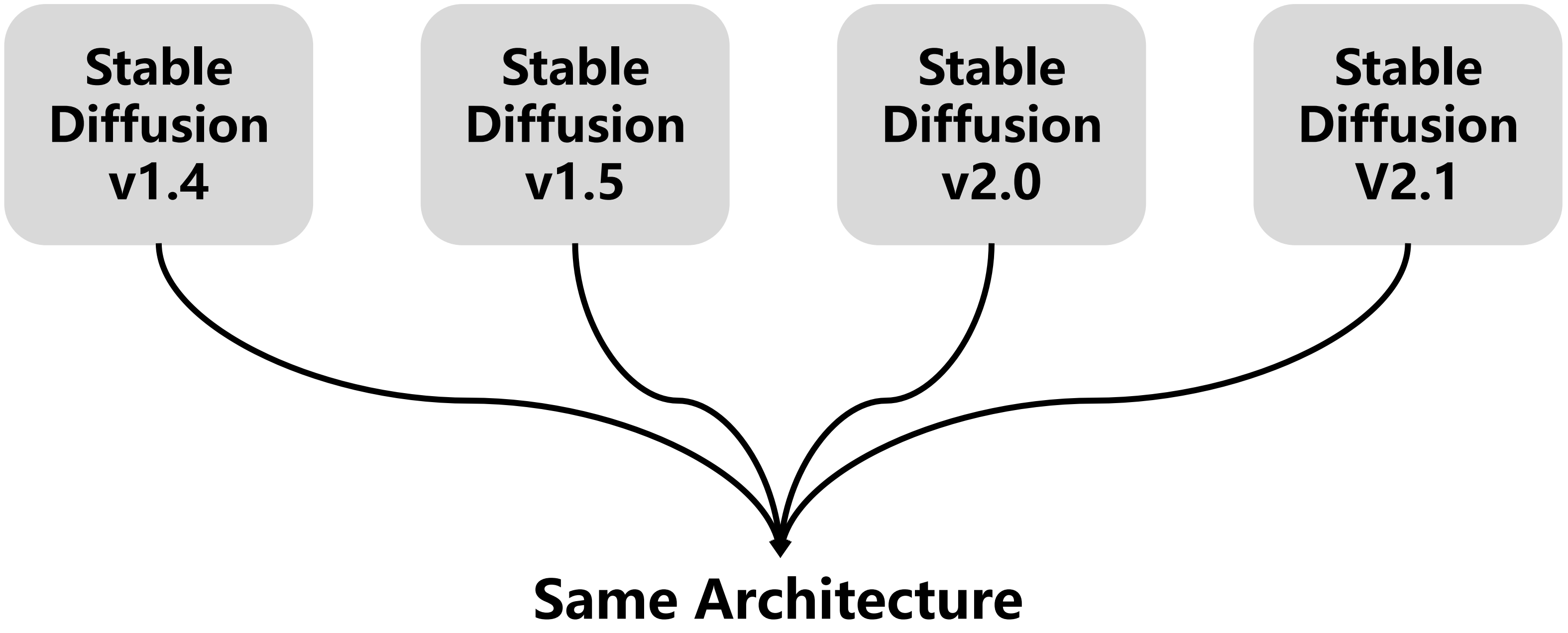
Stable Diffusion v2.0

Stable Diffusion V2.1

Same Architecture

Key Observation

- **Popular diffusion models for diffusion feature study:**



Stable Diffusion v1.4

Stable Diffusion v1.5

Stable Diffusion v2.0

Stable Diffusion V2.1

Same Architecture

Newer, Stronger
Different Architectures

Stable Diffusion XL

Playground v2

PixArt

…

Key Observation

Stable Diffusion v1.5

Unsatisfying Generation

66.7 PCK@0.1↑ Baseline Discrimination

Stable Diffusion XL

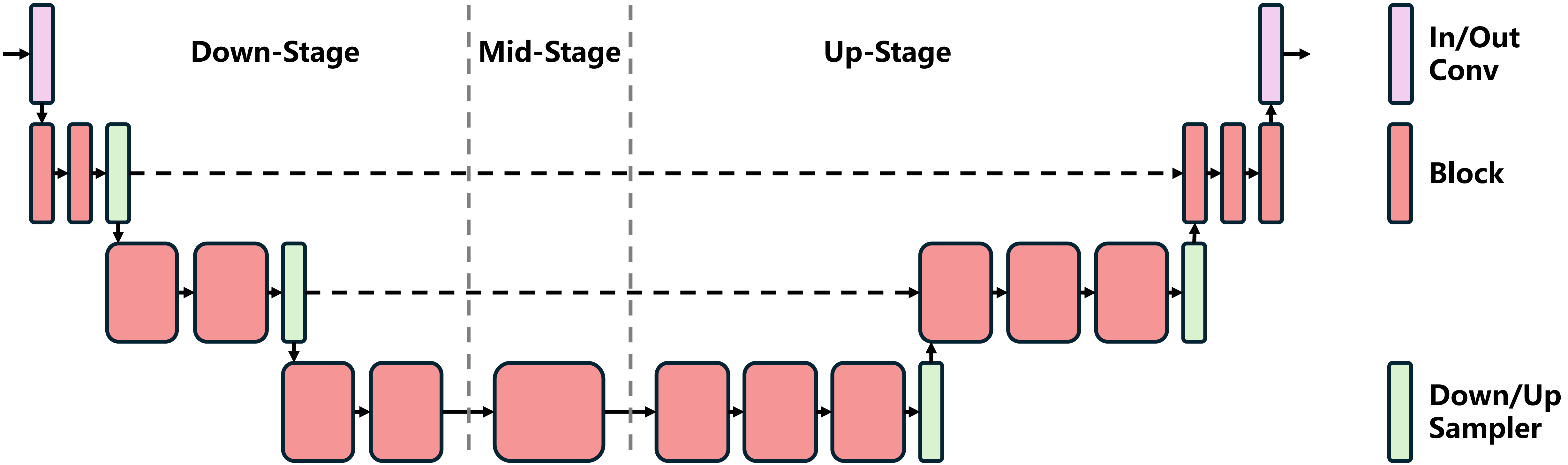
Impressive Generation

59.2 PCK@0.1↑ Worse Discrimination

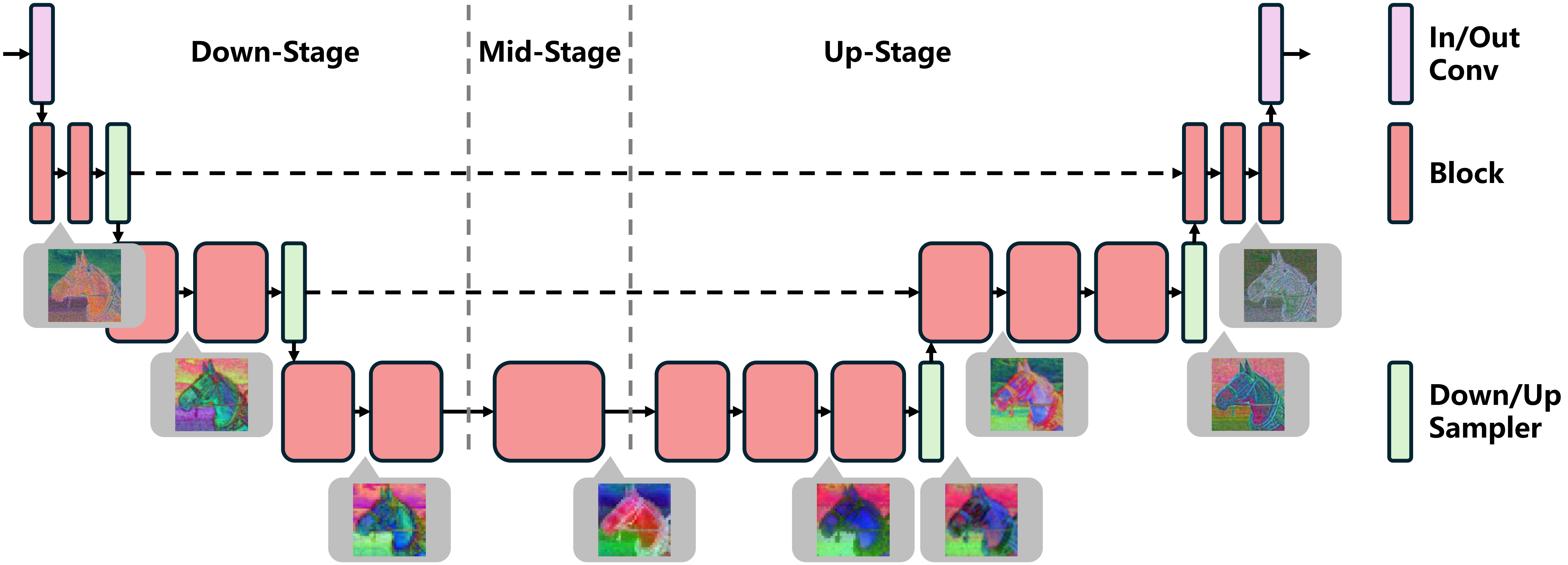
- **Stronger SDXL gives weaker features.**
- **Current feature extraction methods cannot fully unleash the advancements of diffusion models.**

Generation samples taken from Podell, Dustin, et al. "Sdxl: Improving latent diffusion models for high-resolution image synthesis." *arXiv preprint arXiv:2307.01952* (2023).

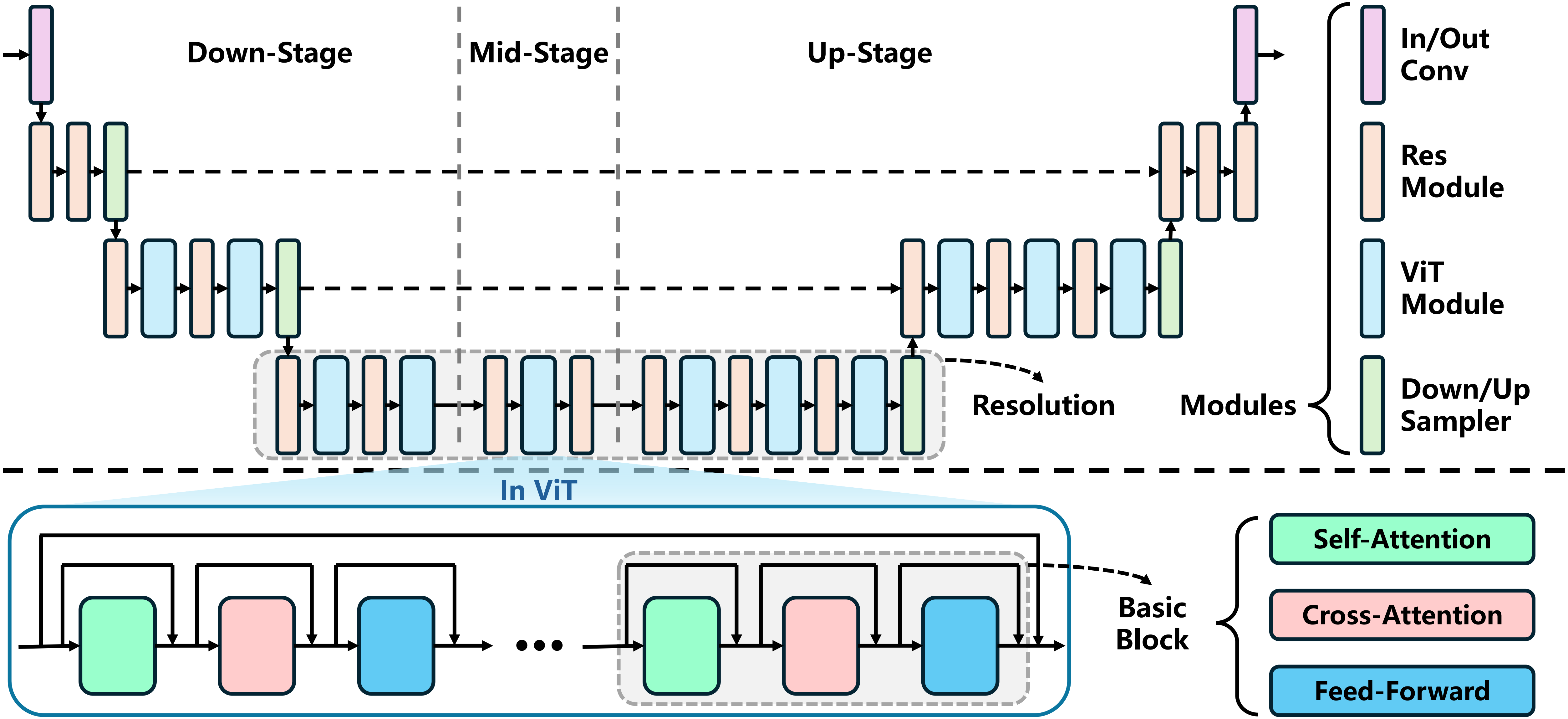
Key Observation



Key Observation



Key Observation



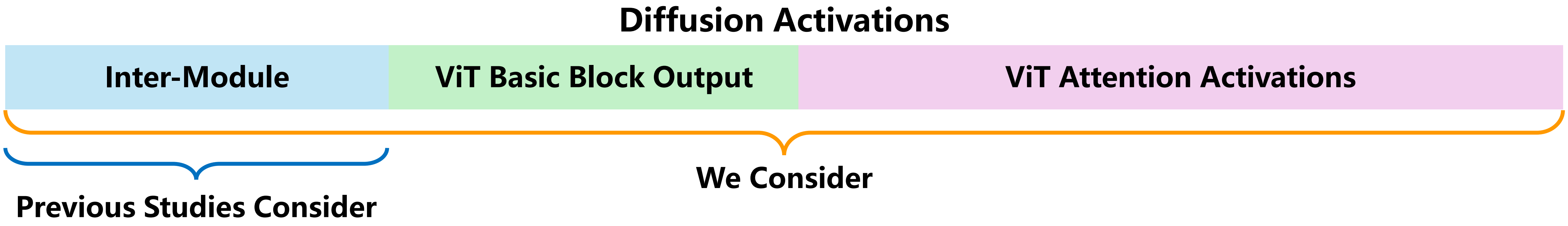
Method

Diffusion Activations

Inter-Module

Previous Studies Consider

Method



Diffusion Activations

Inter-Module

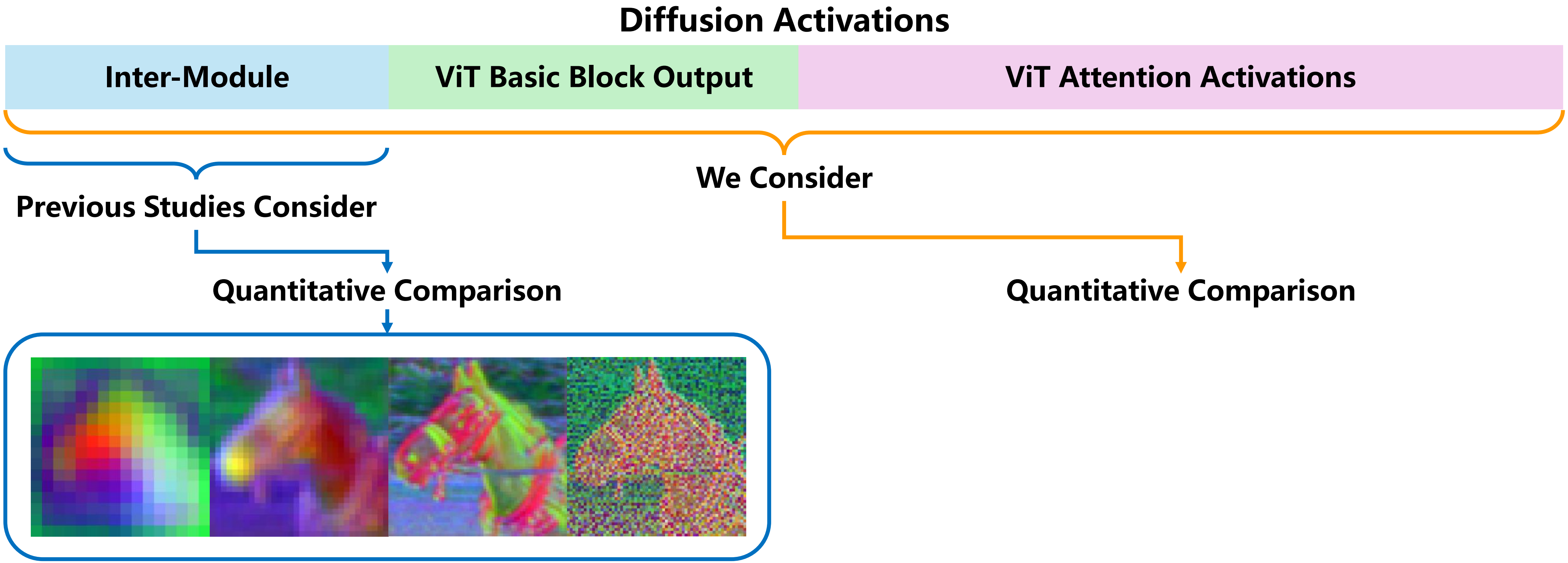
ViT Basic Block Output

ViT Attention Activations

We Consider

Previous Studies Consider

Method



Diffusion Activations

Inter-Module

ViT Basic Block Output

ViT Attention Activations

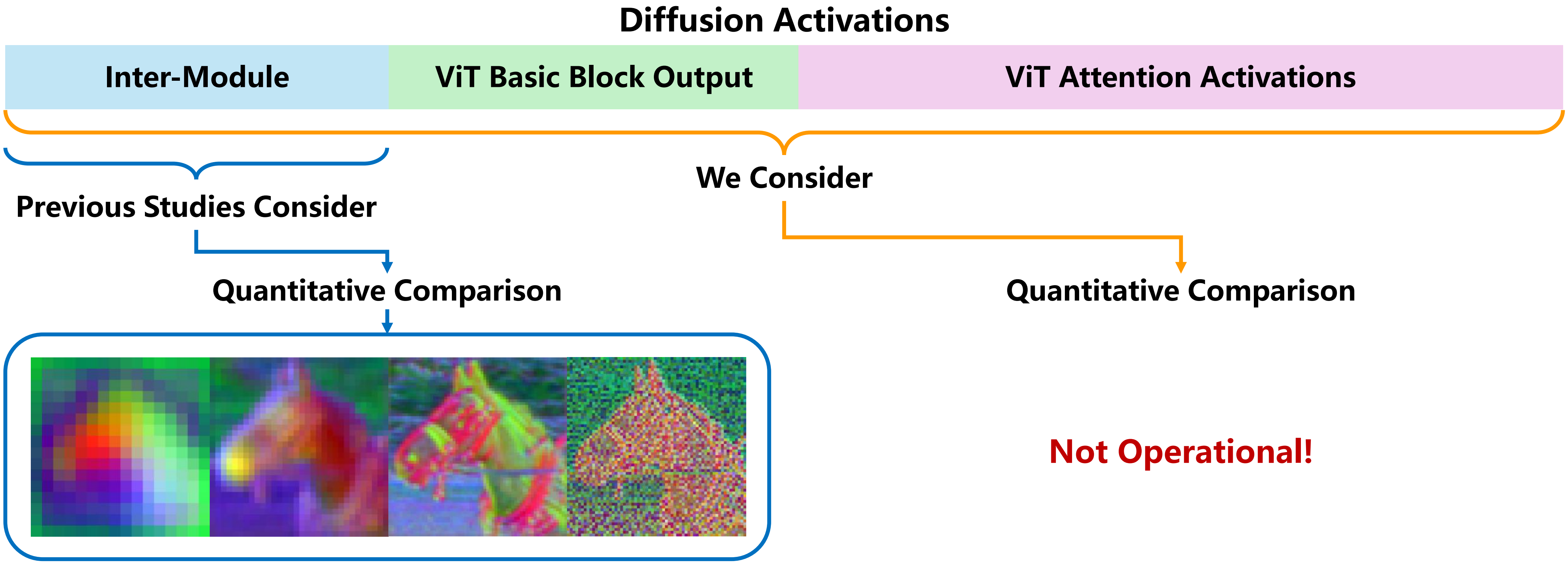
Previous Studies Consider

We Consider

Quantitative Comparison

Quantitative Comparison

Method



Diffusion Activations

Inter-Module

ViT Basic Block Output

ViT Attention Activations

Previous Studies Consider

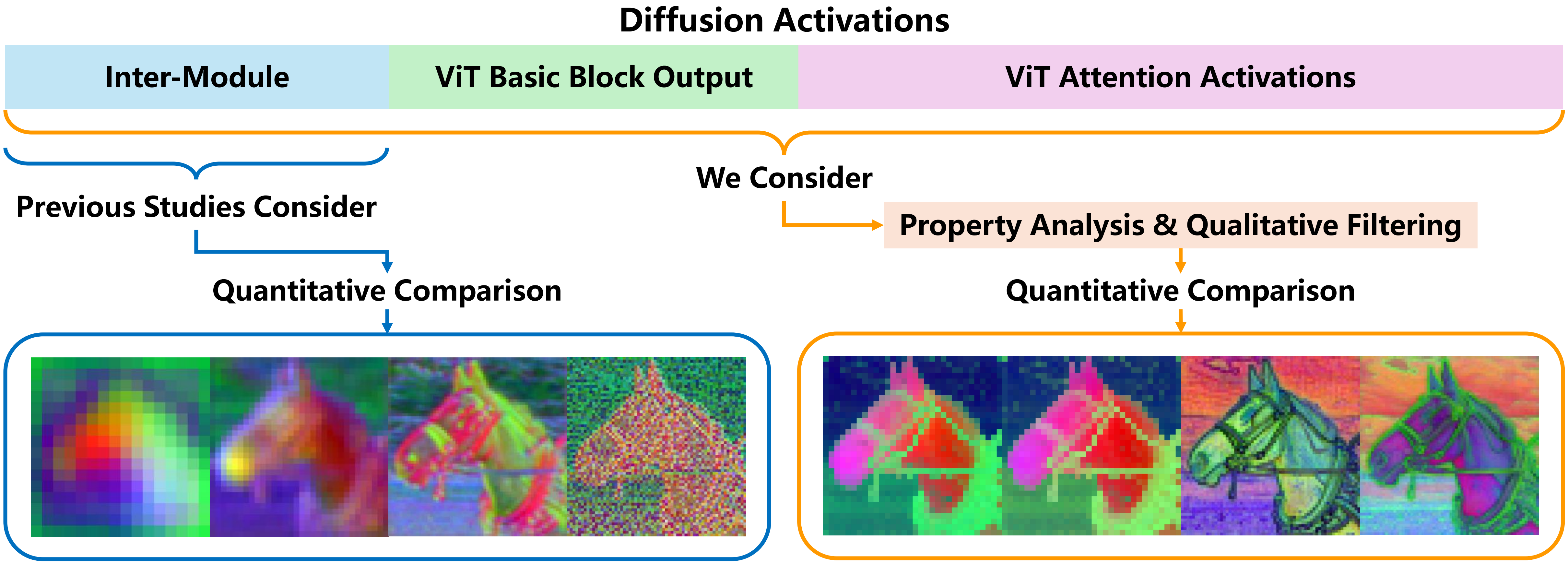
We Consider

Quantitative Comparison

Quantitative Comparison

Not Operational!

Method



Diffusion Activations

Inter-Module

ViT Basic Block Output

ViT Attention Activations

Previous Studies Consider

Quantitative Comparison

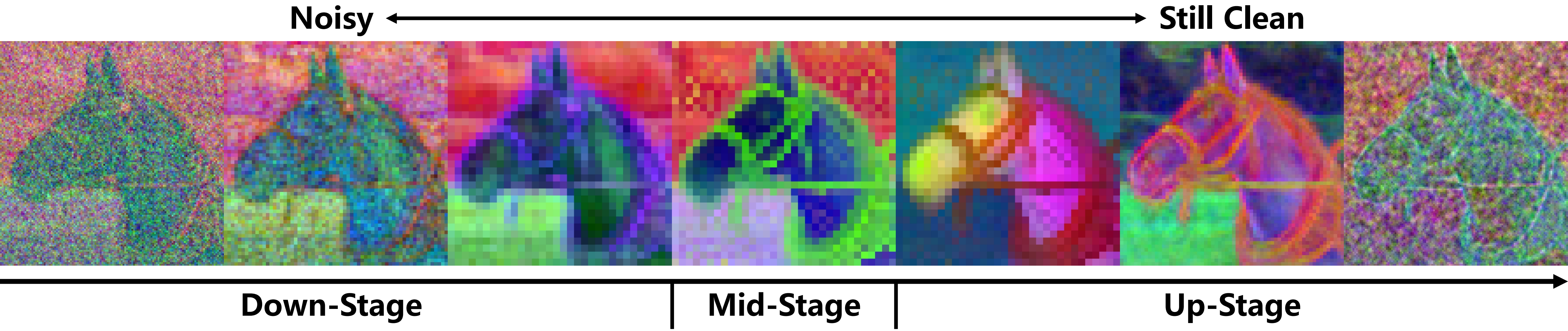
We Consider

Property Analysis & Qualitative Filtering

Quantitative Comparison

Property

- **Diffusion noises**



Noisy

Still Clean

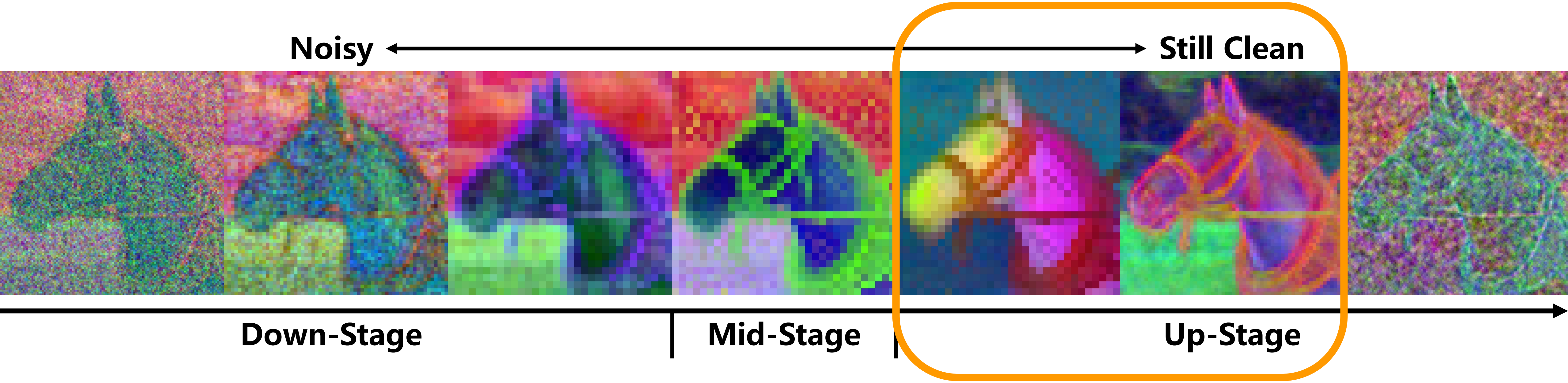
Down-Stage

Mid-Stage

Up-Stage

Property

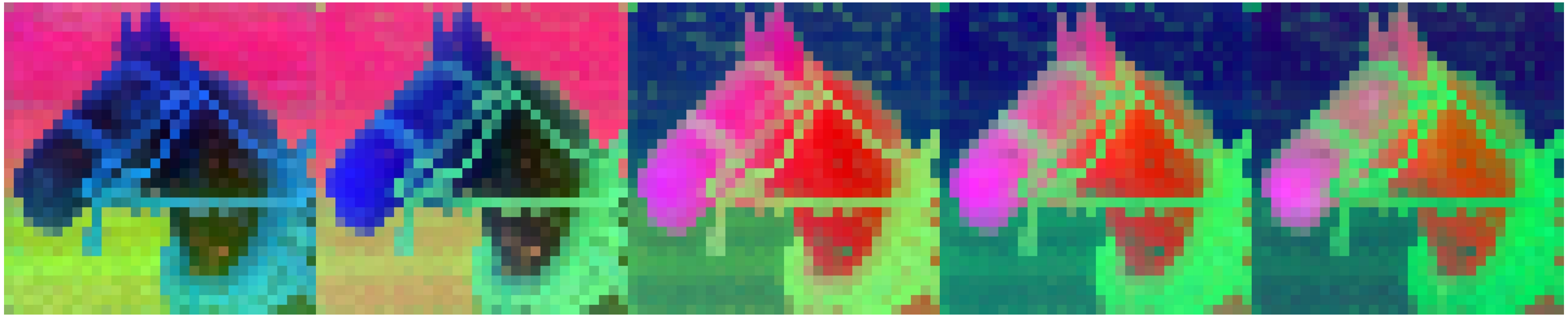
- **Diffusion noises**



- **The first half of the upsampling stage can provide high-quality features.**

Property

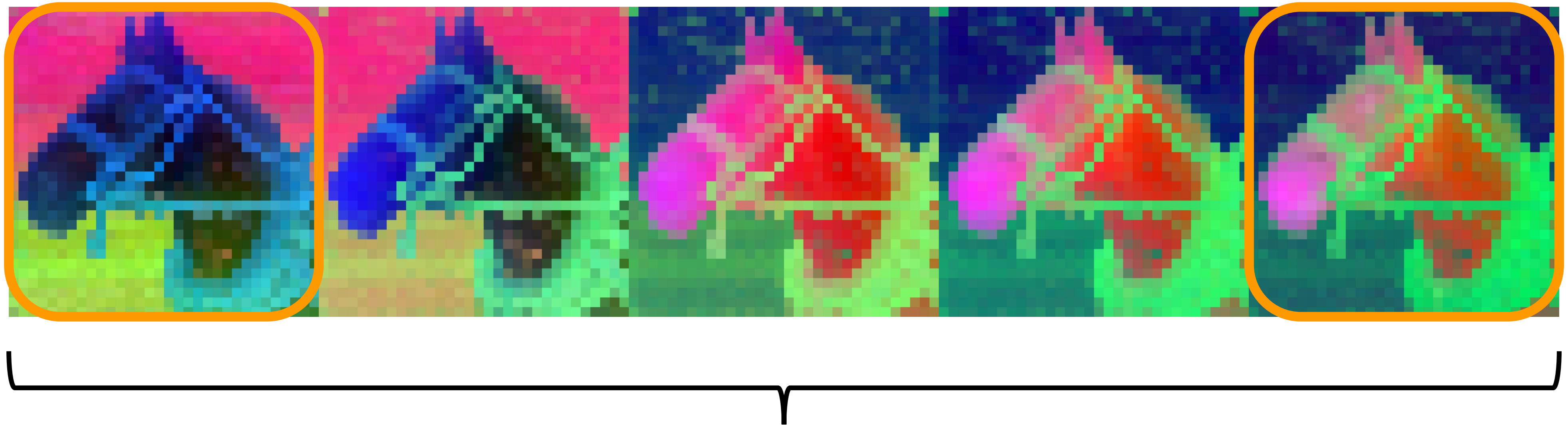
- **In-resolution granularity changes**



All Extracted from Resolution #0 in Up-Stage

Property

- **In-resolution granularity changes**

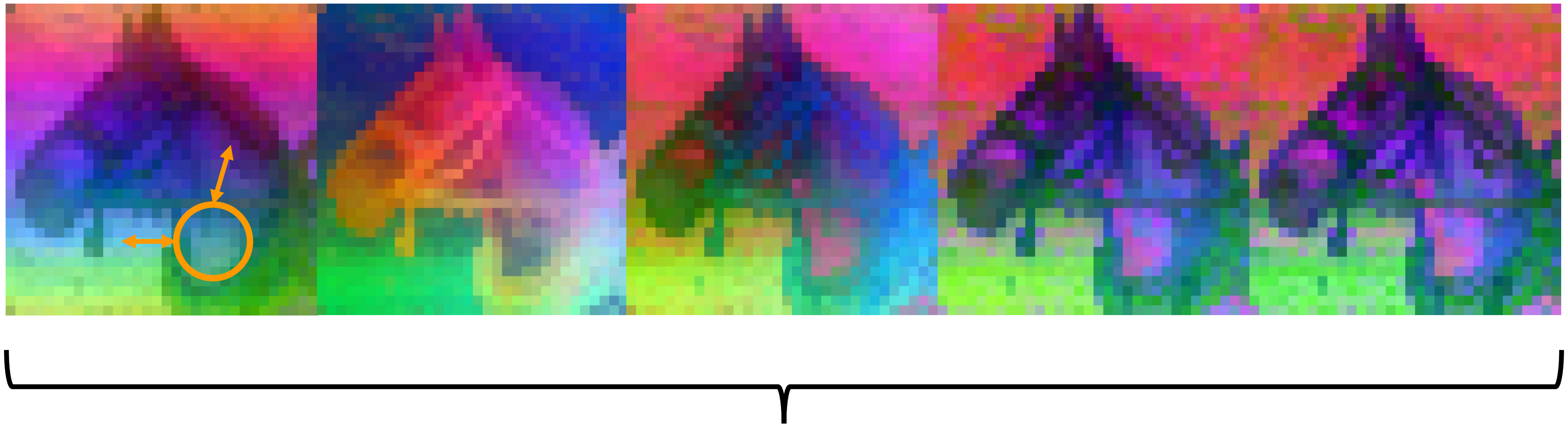


All Extracted from Resolution #0 in Up-Stage

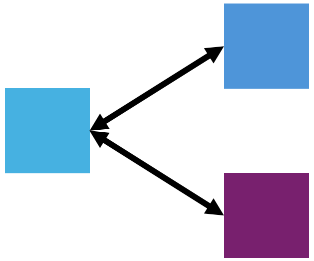
- **It makes sense to select more than one feature from the same resolution level.**

Property

- **Locality without positional embeddings**



Self-Attention Key Activations



Locality: A pixel is …

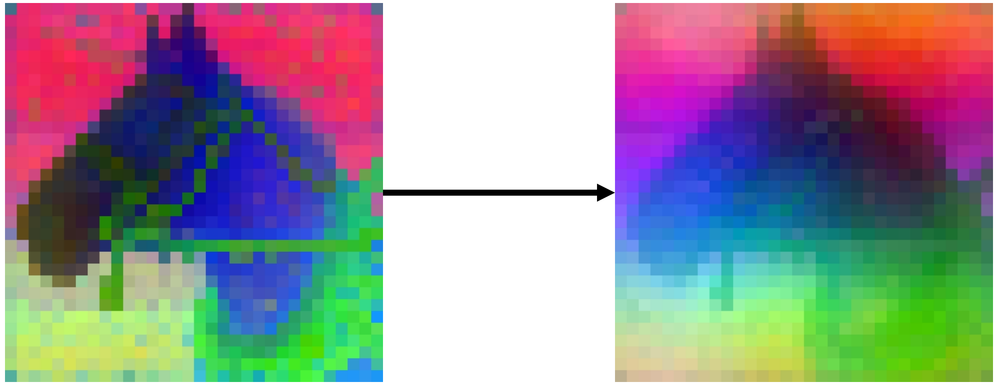
… more similar to nearby pixels

… less similar to pixels with similar semantics

Property

- **Locality without positional embeddings**

Most Time:

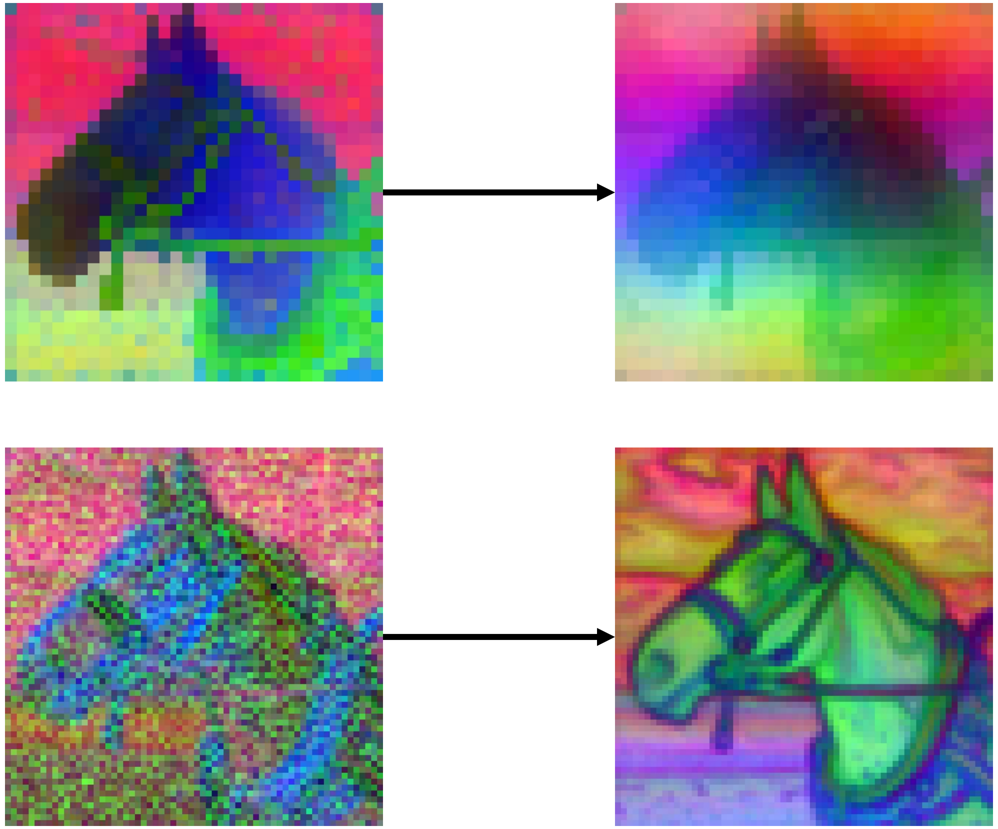


😟

- **Locality causes worse feature quality most of the time.**

Property

- **Locality without positional embeddings**



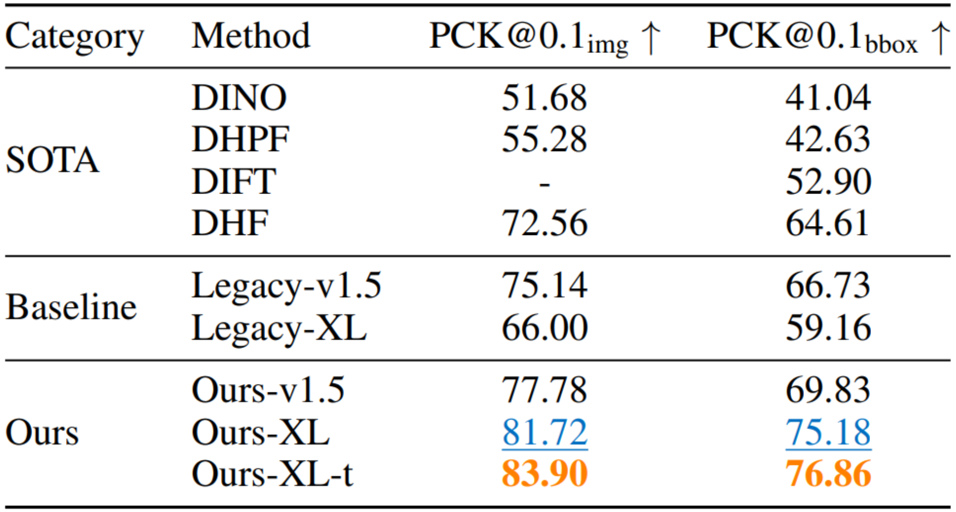
Most Time: 😟

Strong Noise Regions: 😊

- **But locality is also helpful to suppress strong diffusion noises.**

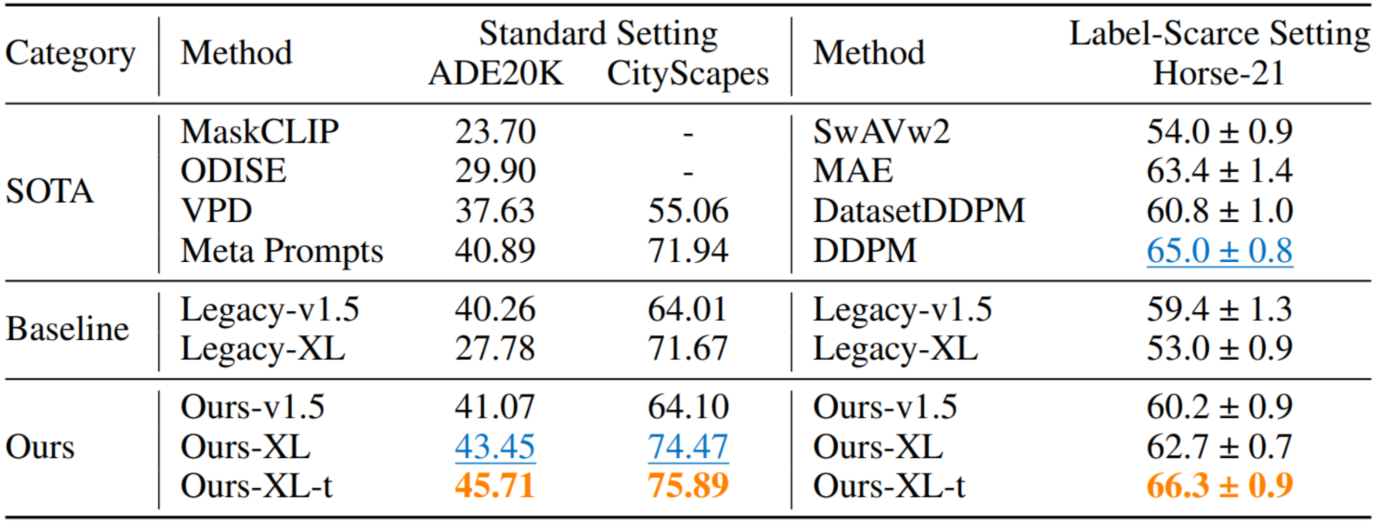
Results

- The best results marked as **bold** and runner-up marked as underlined.



Category	Method	PCK@0.1$_{\text{img}}$ ↑	PCK@0.1$_{\text{bbox}}$ ↑
SOTA	DINO	51.68	41.04
	DHPF	55.28	42.63
	DIFT	-	52.90
	DHF	72.56	64.61
Baseline	Legacy-v1.5	75.14	66.73
	Legacy-XL	66.00	59.16
Ours	Ours-v1.5	77.78	69.83
	Ours-XL	81.72	75.18
	Ours-XL-t	**83.90**	**76.86**

Semantic Correspondence Task

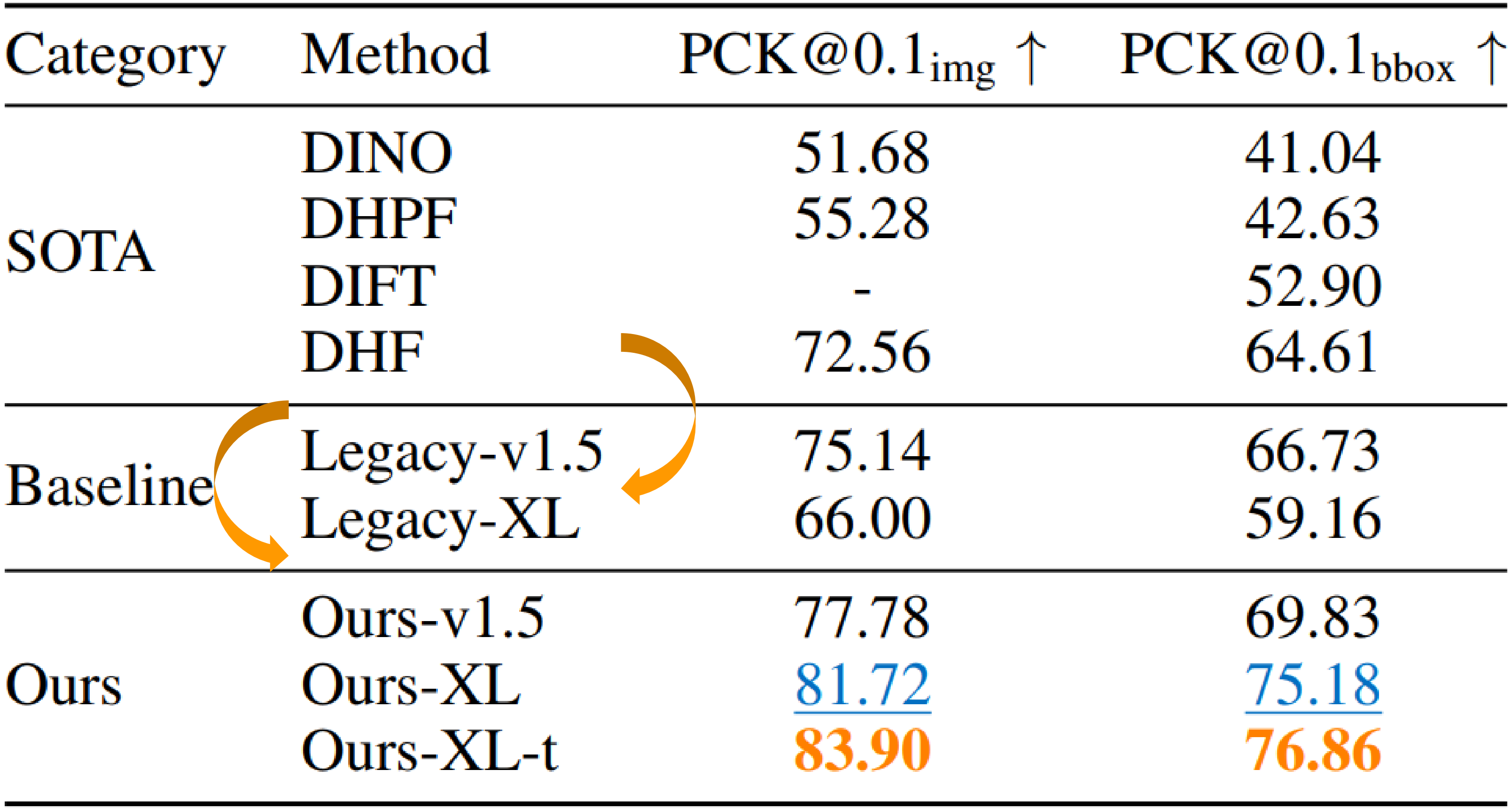


Category	Method	Standard Setting ADE20K	 CityScapes	Method	Label-Scarce Setting Horse-21
SOTA	MaskCLIP	23.70	-	SwAVw2	54.0 ± 0.9
	ODISE	29.90	-	MAE	63.4 ± 1.4
	VPD	37.63	55.06	DatasetDDPM	60.8 ± 1.0
	Meta Prompts	40.89	71.94	DDPM	65.0 ± 0.8
Baseline	Legacy-v1.5	40.26	64.01	Legacy-v1.5	59.4 ± 1.3
	Legacy-XL	27.78	71.67	Legacy-XL	53.0 ± 0.9
Ours	Ours-v1.5	41.07	64.10	Ours-v1.5	60.2 ± 0.9
	Ours-XL	43.45	74.47	Ours-XL	62.7 ± 0.7
	Ours-XL-t	**45.71**	**75.89**	Ours-XL-t	**66.3 ± 0.9**

Semantic Segmentation Task

Results

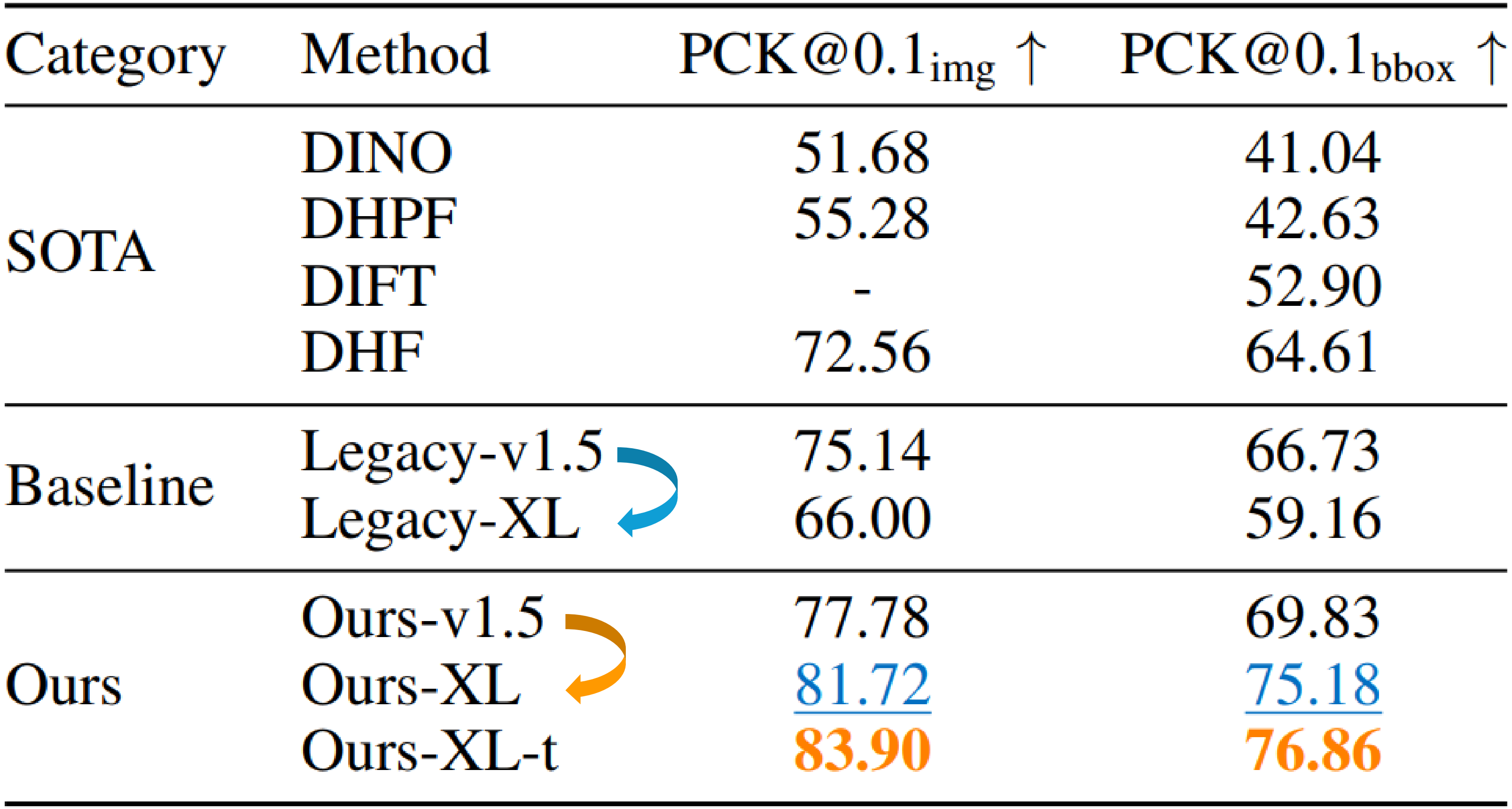
- **Better performance with the same model.**



Category	Method	PCK@0.1$_{\text{img}}$ ↑	PCK@0.1$_{\text{bbox}}$ ↑
SOTA	DINO	51.68	41.04
	DHPF	55.28	42.63
	DIFT	-	52.90
	DHF	72.56	64.61
Baseline	Legacy-v1.5	75.14	66.73
	Legacy-XL	66.00	59.16
Ours	Ours-v1.5	77.78	69.83
	Ours-XL	<u>81.72</u>	<u>75.18</u>
	Ours-XL-t	**83.90**	**76.86**

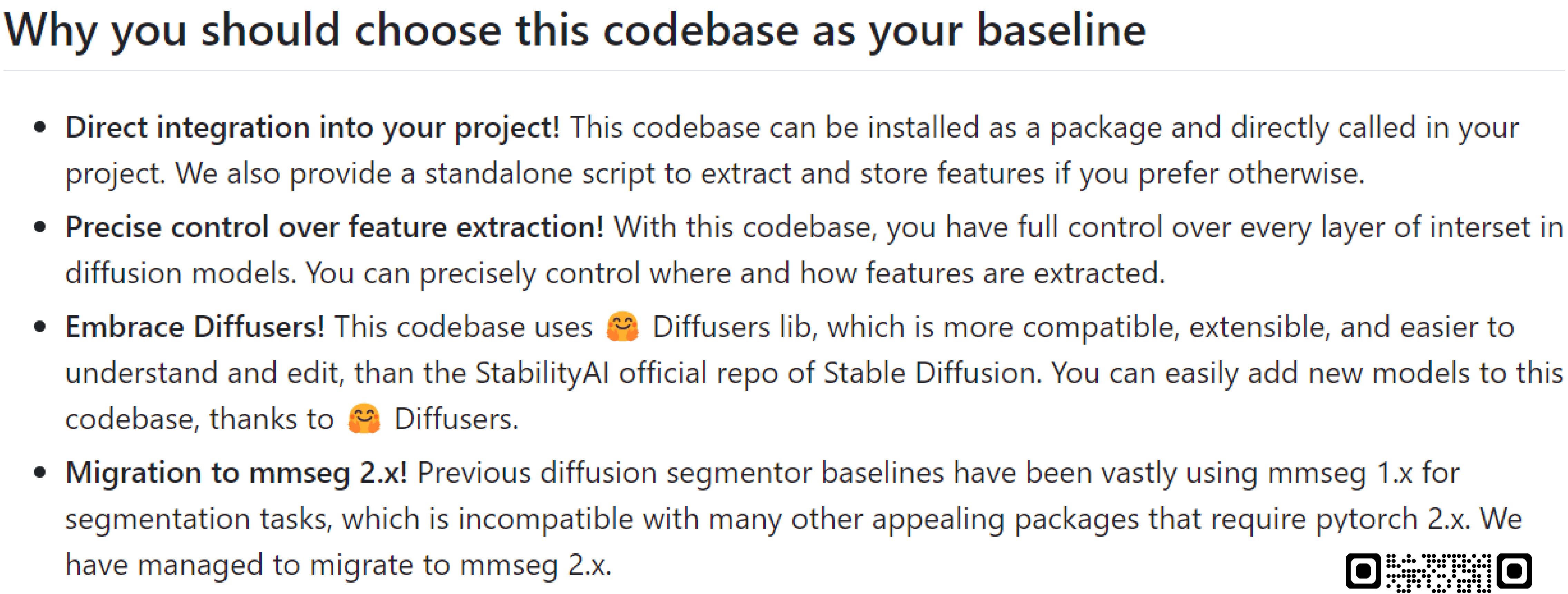
Results

- **Better performance from SDXL than SDv1.5.**



Category	Method	PCK@0.1$_{\text{img}}$ ↑	PCK@0.1$_{\text{bbox}}$ ↑
SOTA	DINO	51.68	41.04
	DHPF	55.28	42.63
	DIFT	-	52.90
	DHF	72.56	64.61
Baseline	Legacy-v1.5	75.14	66.73
	Legacy-XL	66.00	59.16
Ours	Ours-v1.5	77.78	69.83
	Ours-XL	81.72	75.18
	Ours-XL-t	**83.90**	**76.86**

Introducing Our Code Base



Why you should choose this codebase as your baseline

- **Direct integration into your project!** This codebase can be installed as a package and directly called in your project. We also provide a standalone script to extract and store features if you prefer otherwise.
- **Precise control over feature extraction!** With this codebase, you have full control over every layer of interset in diffusion models. You can precisely control where and how features are extracted.
- **Embrace Diffusers!** This codebase uses 🤗 Diffusers lib, which is more compatible, extensible, and easier to understand and edit, than the StabilityAI official repo of Stable Diffusion. You can easily add new models to this codebase, thanks to 🤗 Diffusers.
- **Migration to mmseg 2.x!** Previous diffusion segmentor baselines have been vastly using mmseg 1.x for segmentation tasks, which is incompatible with many other appealing packages that require pytorch 2.x. We have managed to migrate to mmseg 2.x.

GitHub page at: https://github.com/Darkbblue/generic-diffusion-feature

Thanks for your listening!