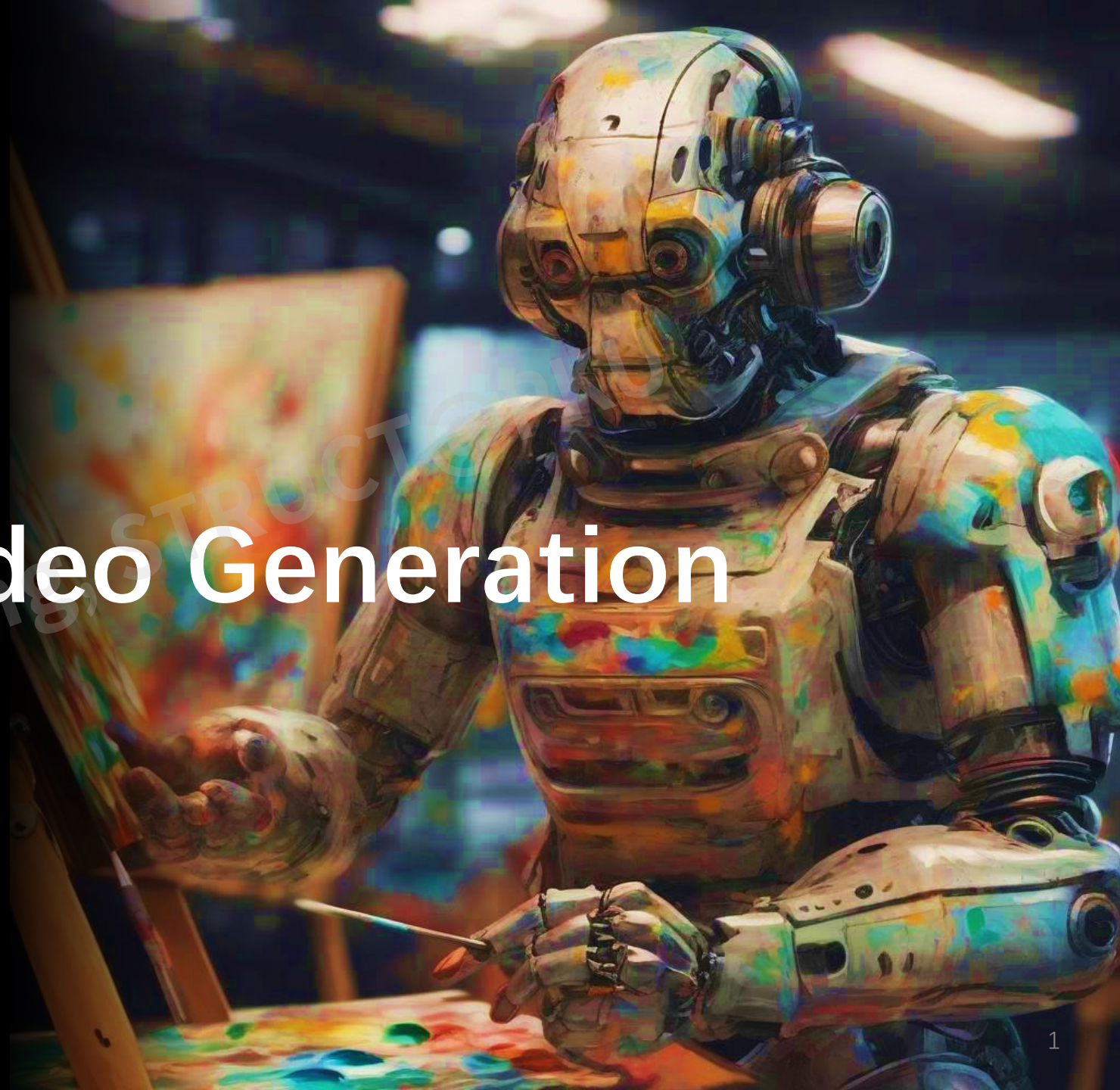
ISCAS 2024

AIGC for Image and Video Generation

Presenter: Shuai Yang

北京大学
PEKING UNIVERSITY



AI for image generation

Development of generative model

- Higher resolution
- Richer content
- More controllable


VAE
(2013)


GAN
(2014)


DCGAN
(2016)


ProGAN
(2018)


StyleGAN2
(2020)


StableDiffusion
(2022)

Traditional Era

GAN Era

Diffusion Era

AI for image generation

As the data scale increases, the generation quality has been greatly improved.

GAN Era

Diffusion Era

GAN (2016)

TFD
112k 48×48

DCGAN (2016)

Celeb-A
200K 64×64

BigGAN (2019)

ImageNet
1.2M 256×256

StyleGAN (2019)

FFHQ
70K 1024×1024

StableDiffusion (2022)

LAION-5B
5.85B 120~1080

AI for image generation

As the data scale increases, the generation quality has been greatly improved.

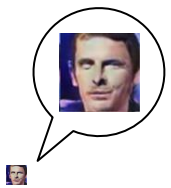
GAN Era

Diffusion Era

GAN (2016)

TFD
112k 48×48

DCGAN (2016)



Celeb-A
200K 64×64

BigGAN (2019)

ImageNet
1.2M 256×256

StyleGAN (2019)

FFHQ
70K 1024×1024

StableDiffusion (2022)

LAION-5B
5.85B 120~1080

AI for image generation

As the data scale increases, the generation quality has been greatly improved.

GAN Era

Diffusion Era

GAN
(2016)

TFD
112k 48×48

DCGAN
(2016)

Celeb-A
200K 64×64

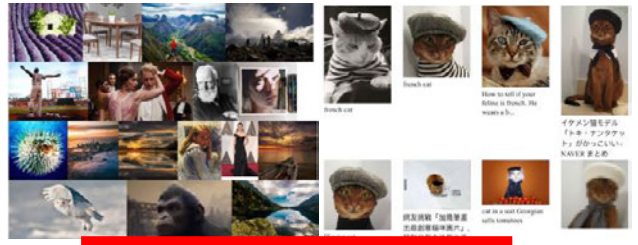
BigGAN
(2019)

ImageNet
1.2M 256×256

StyleGAN
(2019)

FFHQ
70K 1024×1024

StableDiffusion
(2022)



LAION-5B
5.85B 120~1080

AI for image generation

The enhancement of model representation capabilities benefits from the development of computing power and model scale.

GAN Era → **Diffusion Era**

DCGAN
(~70MB)

StyleGAN
(~300MB)

BigGAN
(~200MB)

StableDiffusion
(~4GB)

Single theme per model

Human face, animal face, indoor, vehicle

1000 themes per model

Arbitrary themes per model

AI for image generation

Opportunities in Diffusion Era

- Any theme: draw what I want
- **Diverse style support**

Diversity of Theme

greater expressive power than ever

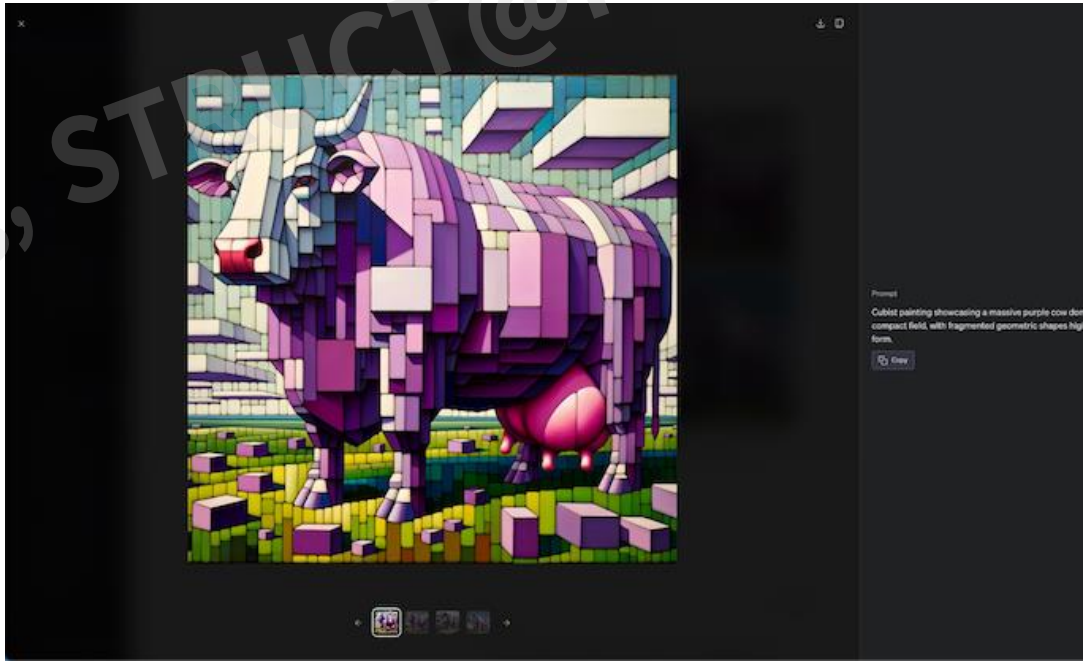
AI for image generation

Opportunities in Diffusion Era

- Any theme: draw what I want
- **Diverse style support**

- What you say is what you get
- **Interactive generation**

Interactive and Controllable Generation
combined with LLMs to achieve interactive and convenient generation



AI for image generation

Opportunities in Diffusion Era

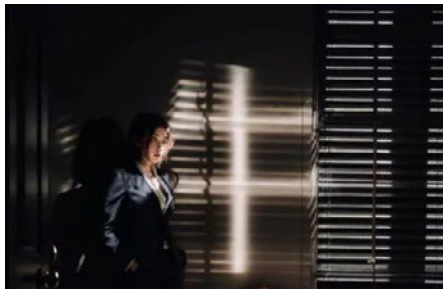
- Any theme: draw what I want
- **Diverse style support**

- What you say is what you get
- **Interactive generation**

- Creative generation and fusion
- **Powerful generation ability**

Challenge "Only human can do creative work"
creative integration and emergence driven by large-scale multi-modal data

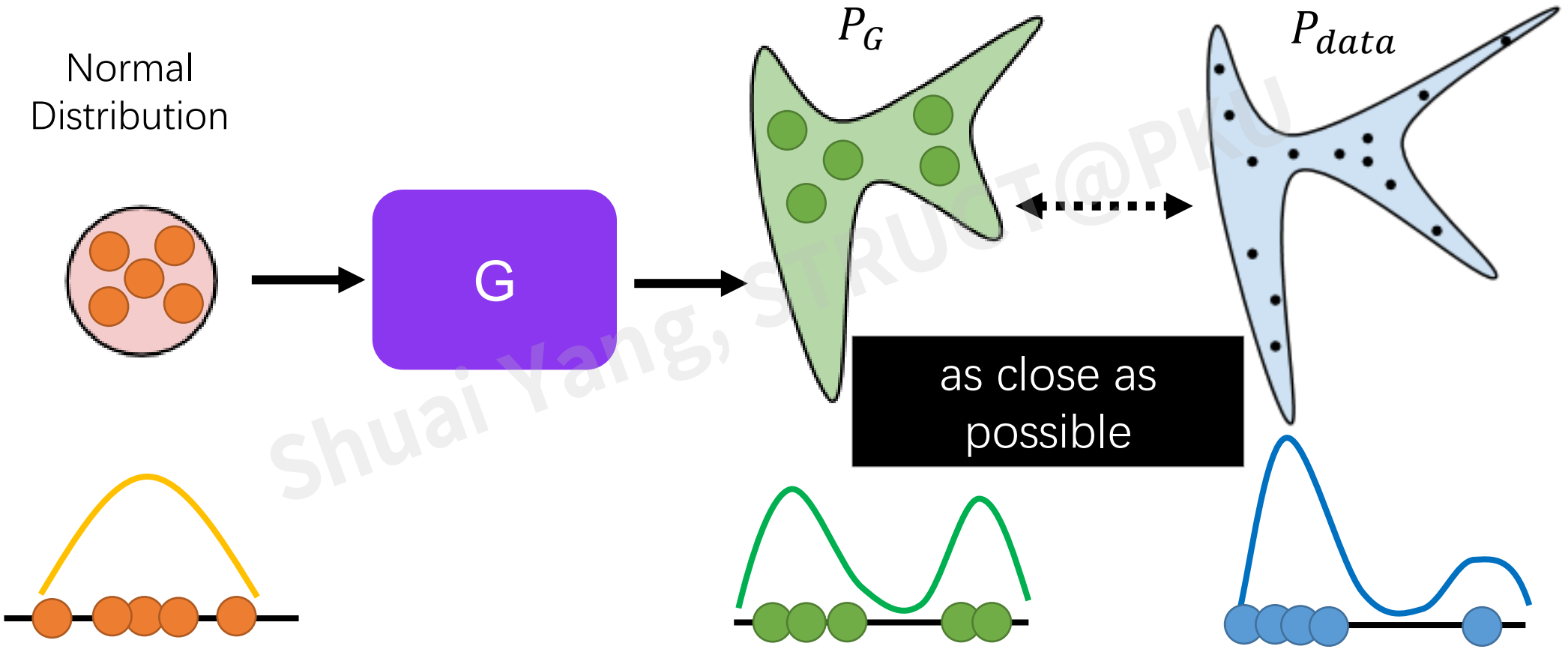



Diffusion

Preliminaries: Diffusion Models

Generative Models



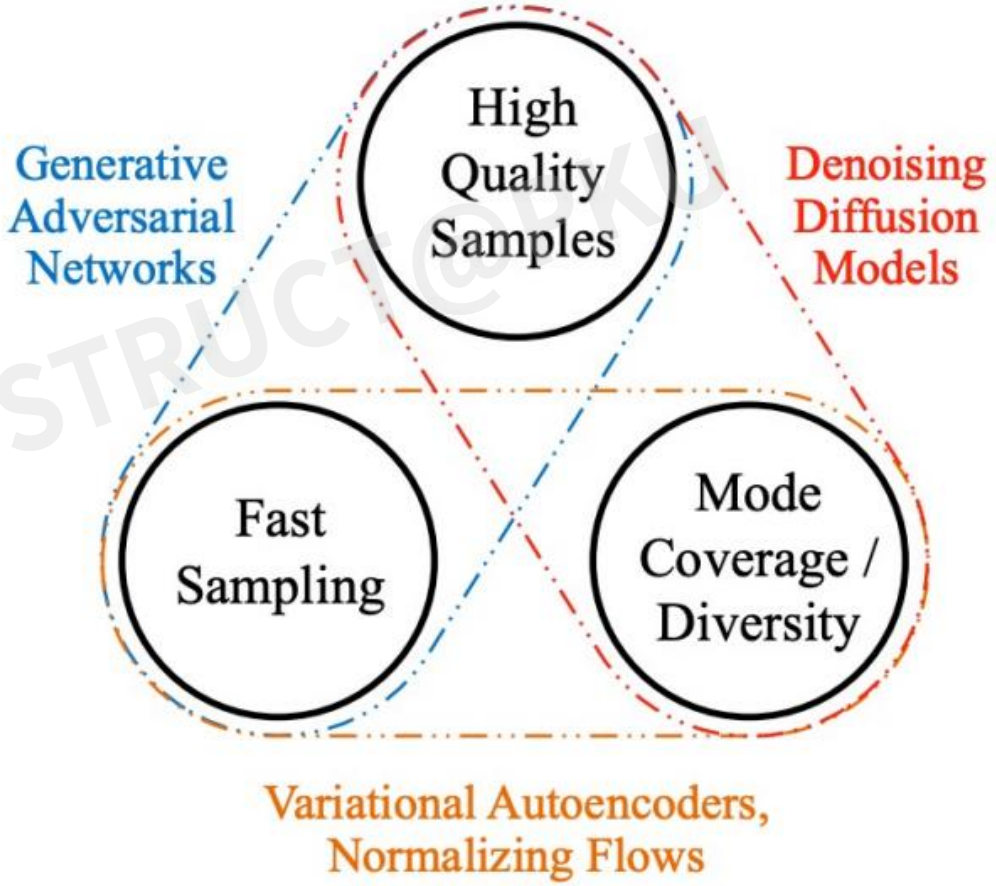
Machine Learning, https://speech.ee.ntu.edu.tw/~hylee/ml/2023-spring.php

Preliminaries: Diffusion Models

Generative Models

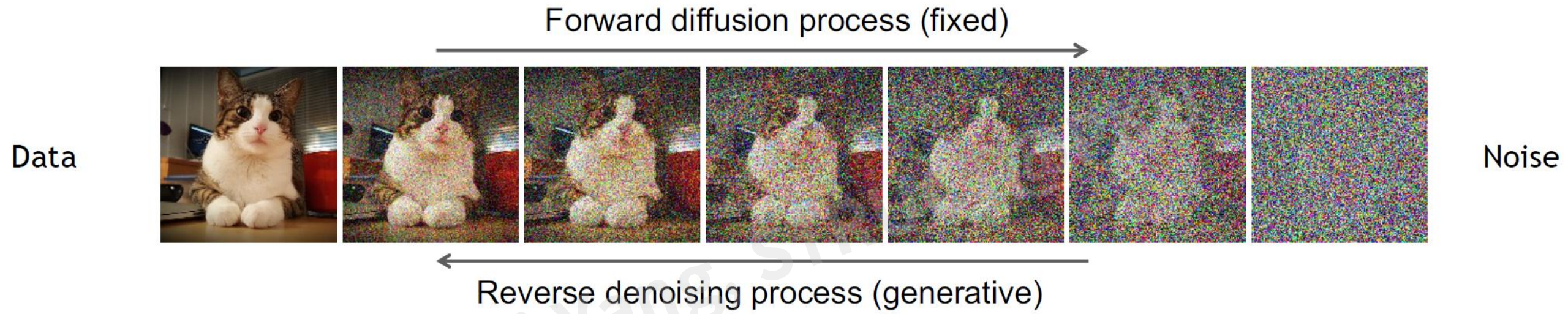
Previous generative models:

- VAEs: Lack of image quality
- GANs: Limited diversity



Denoising Diffusion-based Generative Modeling: Foundations and Applications? CVPR Tutorial, 2022

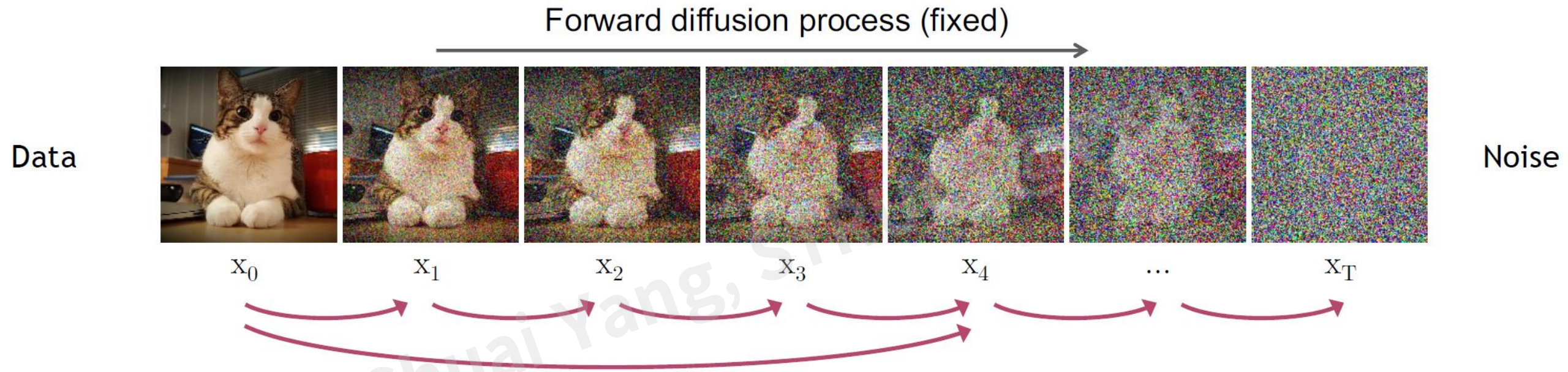
Preliminaries: Diffusion Models



- Adding noise: Images “diffuse” in noise, like ink “diffuses” in water.
- Build a bridge between Gaussian noise and image manifold.

Denoising Diffusion-based Generative Modeling: Foundations and Applications? CVPR Tutorial, 2022

Preliminaries: Diffusion Models

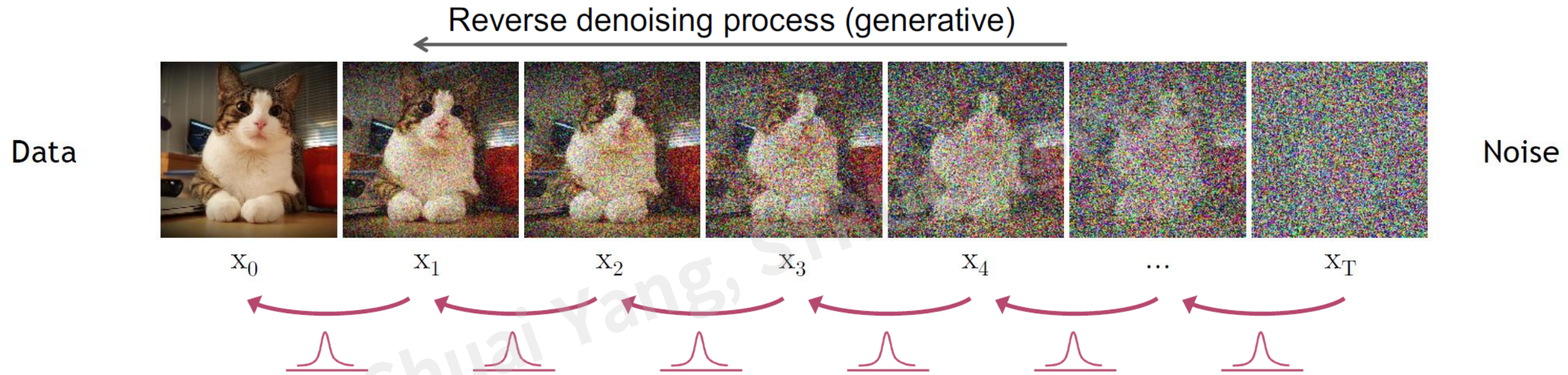


- Forward process: from image to noise

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I}))$$

Denoising Diffusion-based Generative Modeling: Foundations and Applications? CVPR Tutorial, 2022

Preliminaries: Diffusion Models

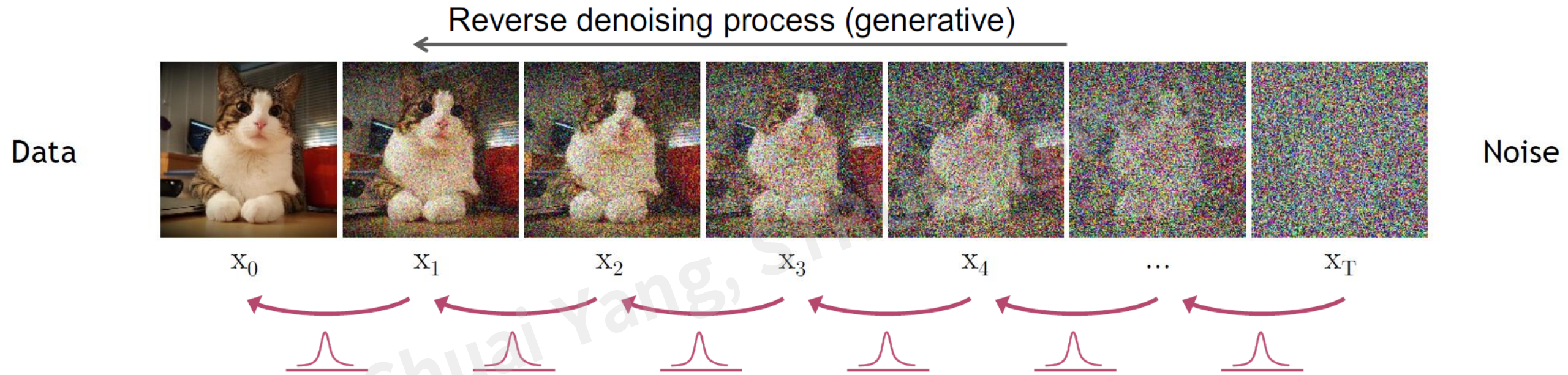


- Reverse process: gradually generate images from noise.

$$p(\mathbf{x}_T) = \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I}) \quad p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \sigma_t^2\mathbf{I})$$

Denoising Diffusion-based Generative Modeling: Foundations and Applications? CVPR Tutorial, 2022

Preliminaries: Diffusion Models

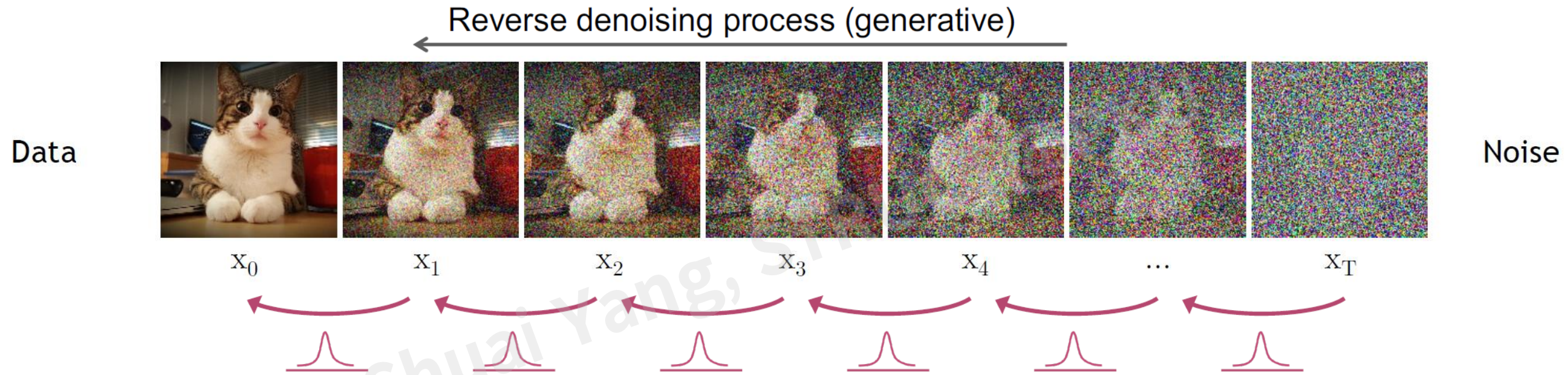


- Reverse process: given the original image $\mathbf{x_0}$, we have

$$q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_{t-1}; \tilde{\mu}_t(\mathbf{x}_t, \mathbf{x}_0), \tilde{\beta}_t \mathbf{I}),$$

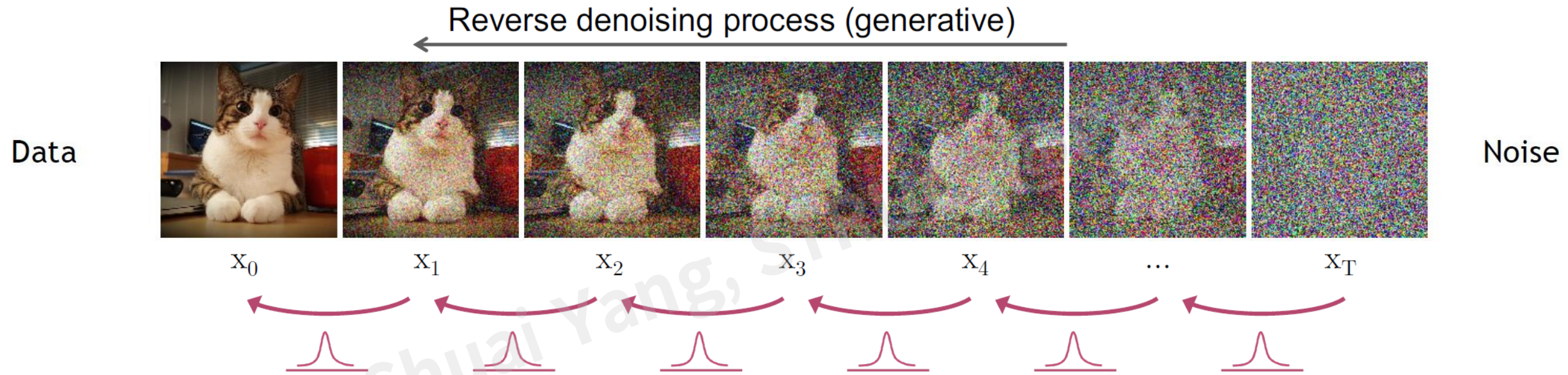
$$\text{where } \tilde{\mu}_t(\mathbf{x}_t, \mathbf{x}_0) := \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\mathbf{x}_0 + \frac{\sqrt{1-\beta_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t \text{ and } \tilde{\beta}_t := \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$$

Preliminaries: Diffusion Models



- Reverse process: replace the $\mathbf{x_0}$ as the predicted one:

Preliminaries: Diffusion Models

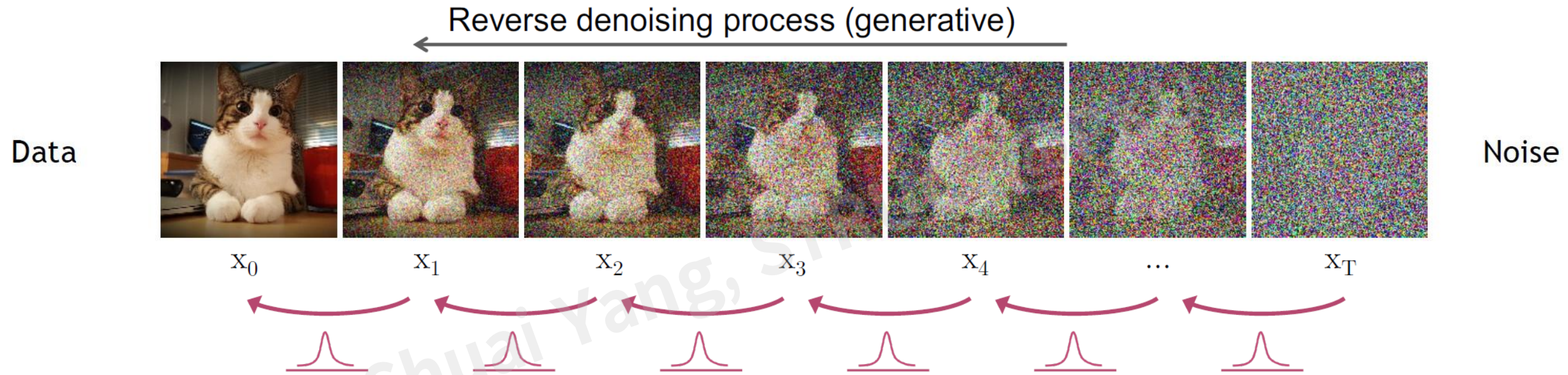


- Reverse process: predicting the noise in $\mathbf{x}_t$

$$\boldsymbol{\mu}_\theta(\mathbf{x}_t, t) = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\right)$$

Denoising Diffusion-based Generative Modeling: Foundations and Applications? CVPR Tutorial, 2022

Preliminaries: Diffusion Models



- Reverse process: training loss

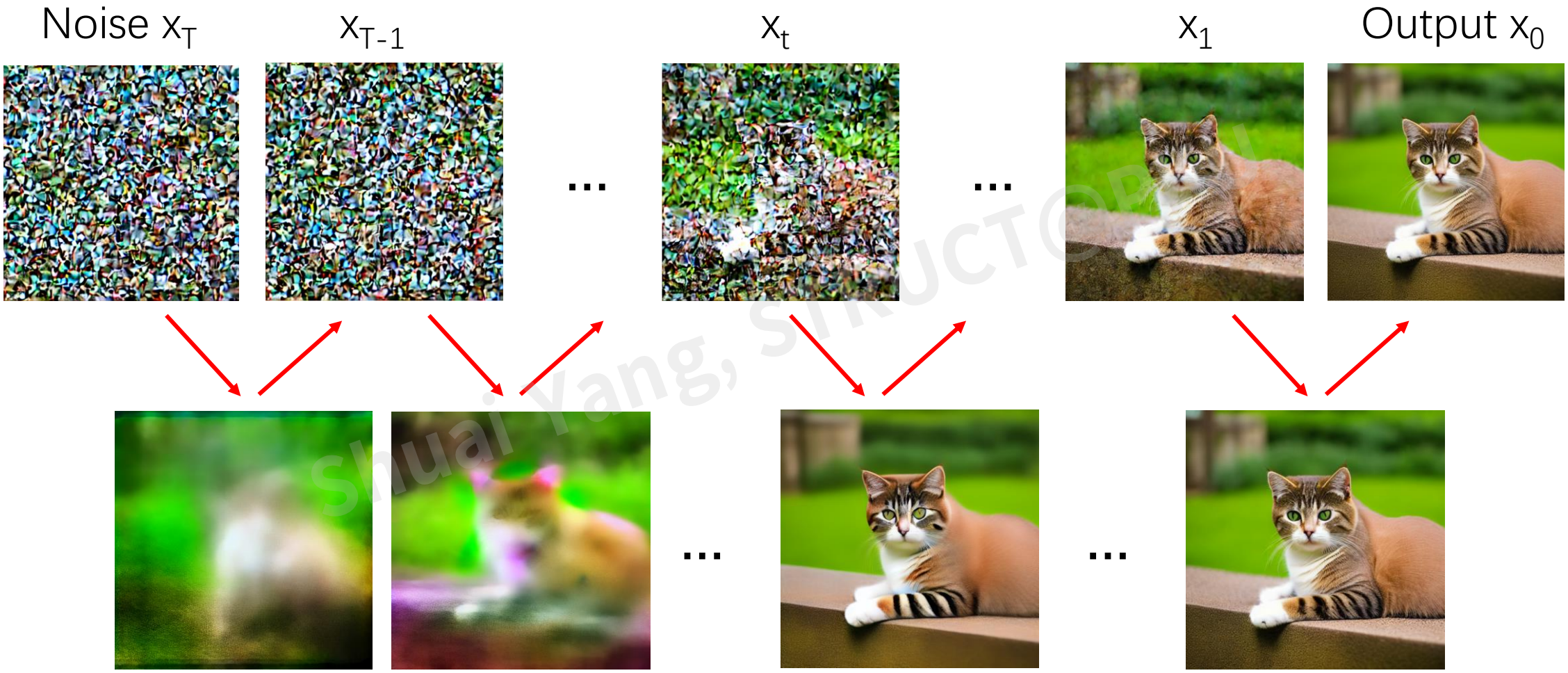
$$\mathbb{E}\left[-\log p_\theta(\mathbf{x}_0)\right] =: L$$

$$L_{\text{simple}} = \mathbb{E}_{\mathbf{x}_0 \sim q(\mathbf{x}_0), \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), t \sim \mathcal{U}(1,T)} \left[||\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}\, \mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon, t)||^2 \right]$$

Denoising Diffusion-based Generative Modeling: Foundations and Applications? CVPR Tutorial, 2022

Preliminaries: Diffusion Models

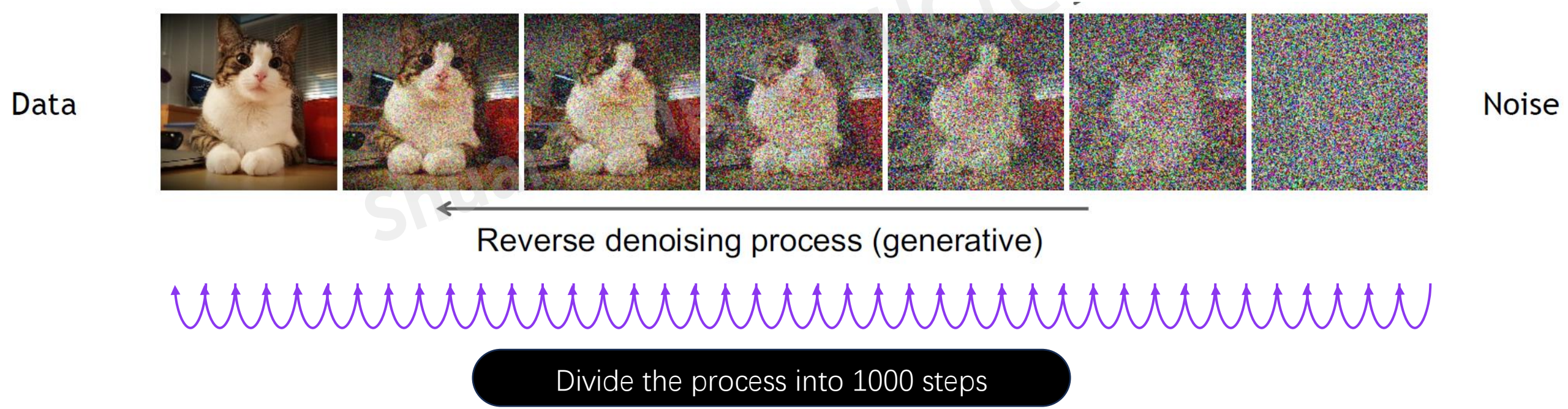
Diffusion Models



Preliminaries: Diffusion Models

Disadvantage: Sampling speed

- The generation process is gradual

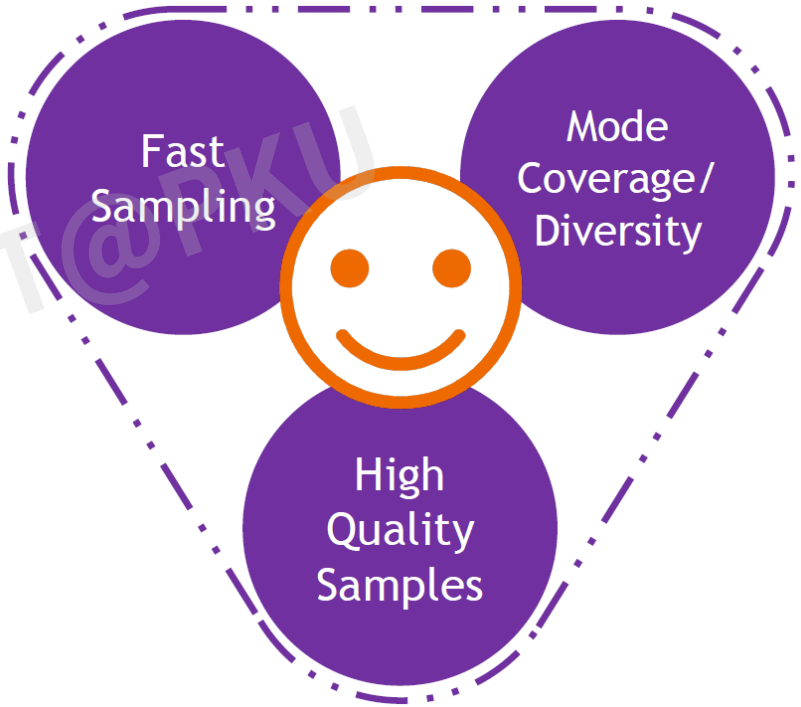


Denoising Diffusion-based Generative Modeling: Foundations and Applications? CVPR Tutorial, 2022

Preliminaries: Diffusion Models

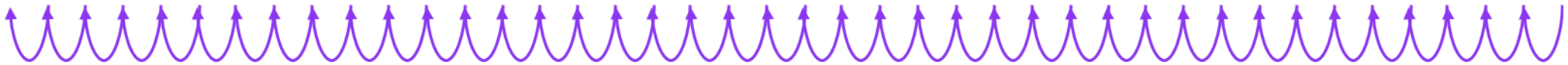
Disadvantage: Sampling speed

- The generation process is gradual



Accelerating:

- Determined sampling process
- Larger denoising step



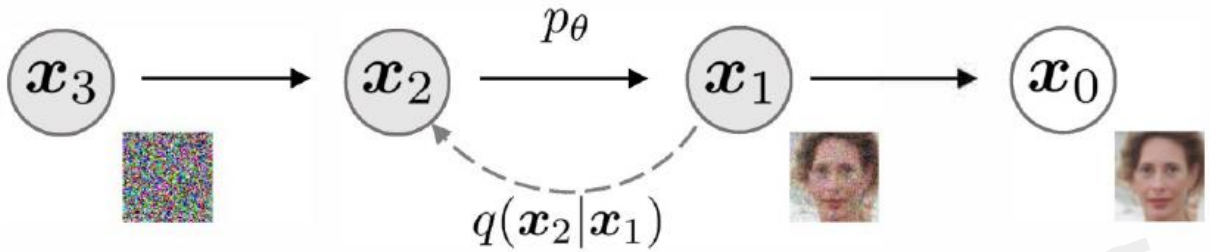
Divide the process into 50 or fewer steps

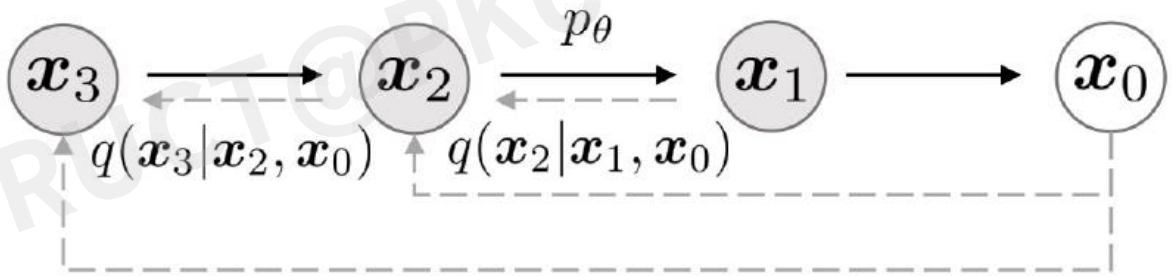
Denoising Diffusion-based Generative Modeling: Foundations and Applications? CVPR Tutorial, 2022

Preliminaries: Diffusion Models

DDIM

Non-Markovian process





$$q_\sigma(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) = \mathcal{N}\left(\sqrt{\alpha_{t-1}}\boldsymbol{x}_0 + \sqrt{1-\alpha_{t-1}-\sigma_t^2}\cdot\frac{\boldsymbol{x}_t - \sqrt{\alpha_t}\boldsymbol{x}_0}{\sqrt{1-\alpha_t}}, \sigma_t^2\boldsymbol{I}\right)$$

Determined when σ_t set to 0

Denoising Diffusion Implicit Models, ICLR 2021

Preliminaries: Diffusion Models

DDIM inversion

Determined sampling process means –
The start noise can also obtained by the image.

$$x_{t+1} - x_t = \sqrt{\bar{\alpha}_{t+1}} \left[\left(\sqrt{1/\bar{\alpha}_t} - \sqrt{1/\bar{\alpha}_{t+1}} \right) x_t + \left(\sqrt{1/\bar{\alpha}_{t+1} - 1} - \sqrt{1/\bar{\alpha}_t - 1} \right) \epsilon_\theta(x_t) \right]$$

The noise can be utilized in many ways (e.g., interpolating).

Diffusion Models Beat GANs on Image Synthesis, NeurIPS 2021.

Guidance of Diffusion

Problem: How to guide the diffusion models?

- A simple way: feed the noise-prediction network with the condition

$$\epsilon_\theta(x_t, t) \longrightarrow \epsilon_\theta(x_t, y, t)$$

Guidance of Diffusion

Problem: How to guide the sampling process?

- Classifier guidance: employing an external classifier:

$$\mu \longrightarrow \mu + s\Sigma \nabla_{x_t} \log \boxed{p_\phi(y|x_t)}$$

External classifier

- The classifier “pulls” the noisy image to the direction of the condition.

Diffusion Models Beat GANs on Image Synthesis, NeurIPS 2021.

Guidance of Diffusion

Problem: How to guide the sampling process?

- Classifier guidance: employing an external classifier:

$$\mu \longrightarrow \mu + s\Sigma \nabla_{x_t} \log \boxed{p_\phi(y|x_t)}$$

Hard to train!

Diffusion Models Beat GANs on Image Synthesis, NeurIPS 2021.

Guidance of Diffusion

Problem: How to guide the sampling process?

- Classifier-free guidance: implicit classifier

$$\tilde{\boldsymbol{\epsilon}}_\theta(\mathbf{z}_\lambda, \mathbf{c}) = (1+w)\boldsymbol{\epsilon}_\theta(\mathbf{z}_\lambda, \mathbf{c}) - w\boldsymbol{\epsilon}_\theta(\mathbf{z}_\lambda)$$

- Drop the condition randomly
- Apply the distance of conditioned and unconditioned

$$\boldsymbol{\epsilon}_\theta(\mathbf{z}_\lambda, \varnothing)$$

Classifier-Free Diffusion Guidance, NIPSW 2021.

Preliminaries: Diffusion Models

Guidance of Diffusion

$$\tilde{\epsilon}_\theta(\mathbf{z}_\lambda, \mathbf{c}) = (1+w)\epsilon_\theta(\mathbf{z}_\lambda, \mathbf{c}) - w\epsilon_\theta(\mathbf{z}_\lambda)$$

a cute cat in the garden

W=-1

W=0

W=1

W=3

W=6.5

W=9

Classifier-Free Diffusion Guidance, NIPSW 2021.

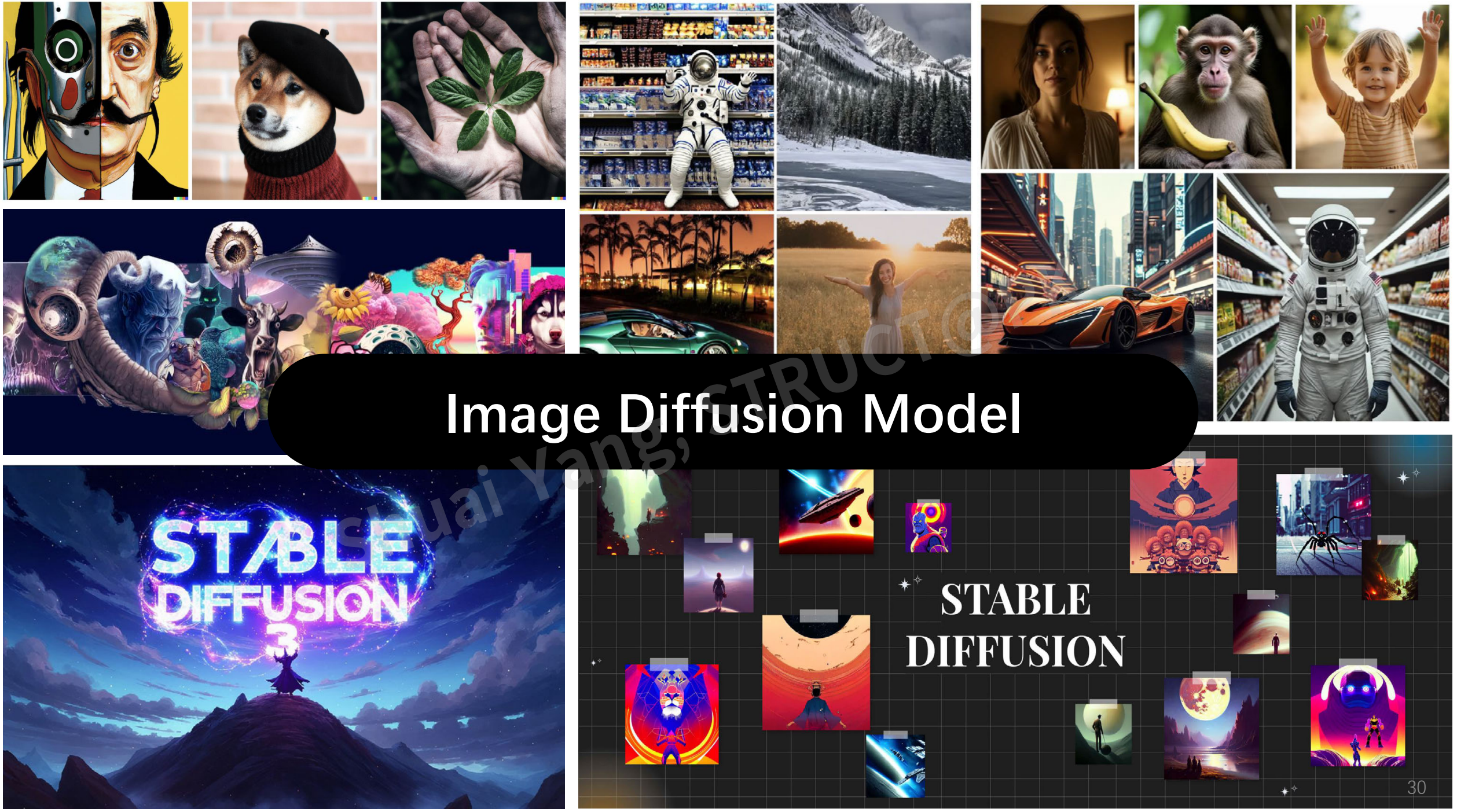


Image diffusion model

The World's Smartest Artificial Intelligence Just Made Its First Magazine Cover

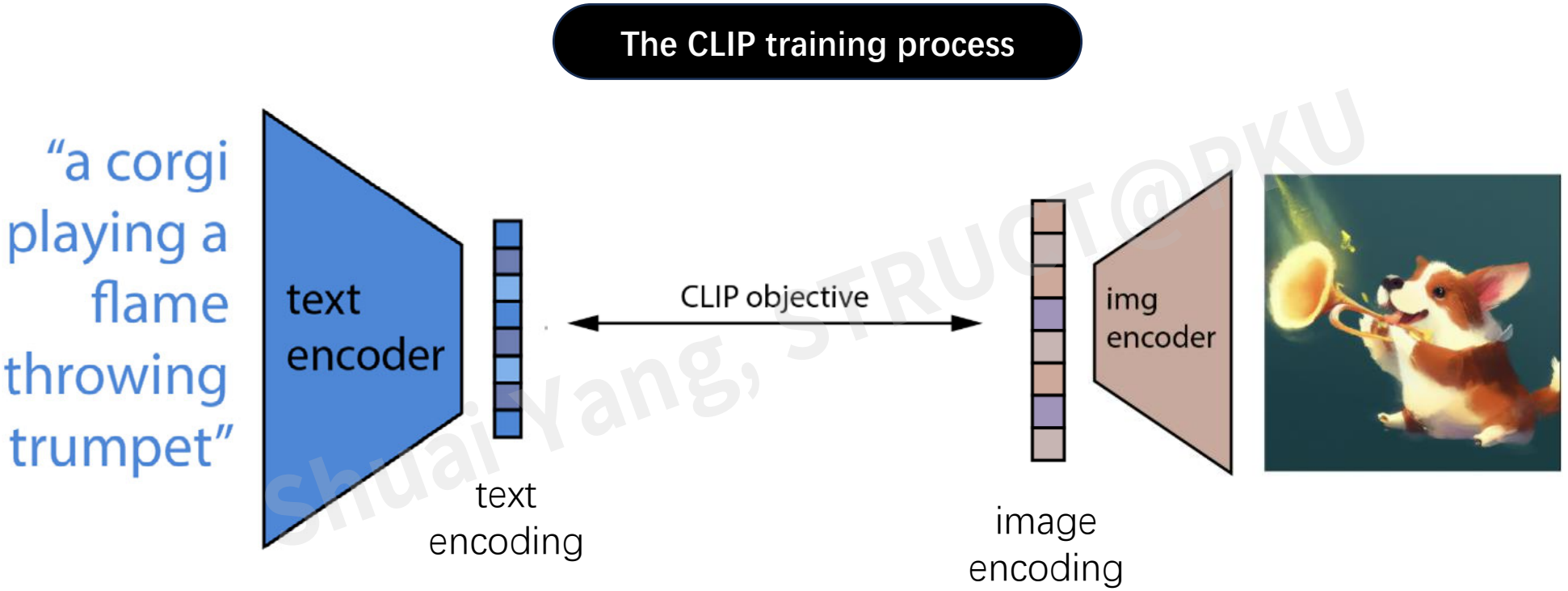
wide-angle shot from below of a female astronaut with an athletic feminine body walking with swagger toward camera on Mars in an infinite universe, synthwave digital art

digital artist — Karen X. Cheng

DALL·E 2

Image diffusion model

DALL·E 2

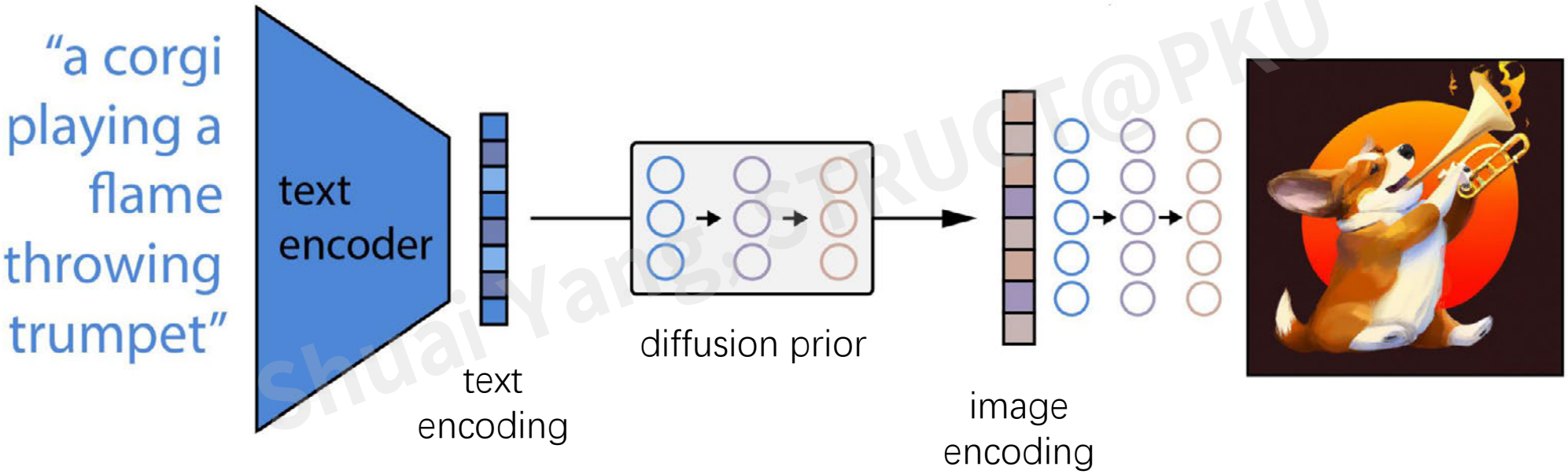


Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv'22

Image diffusion model

DALL·E 2

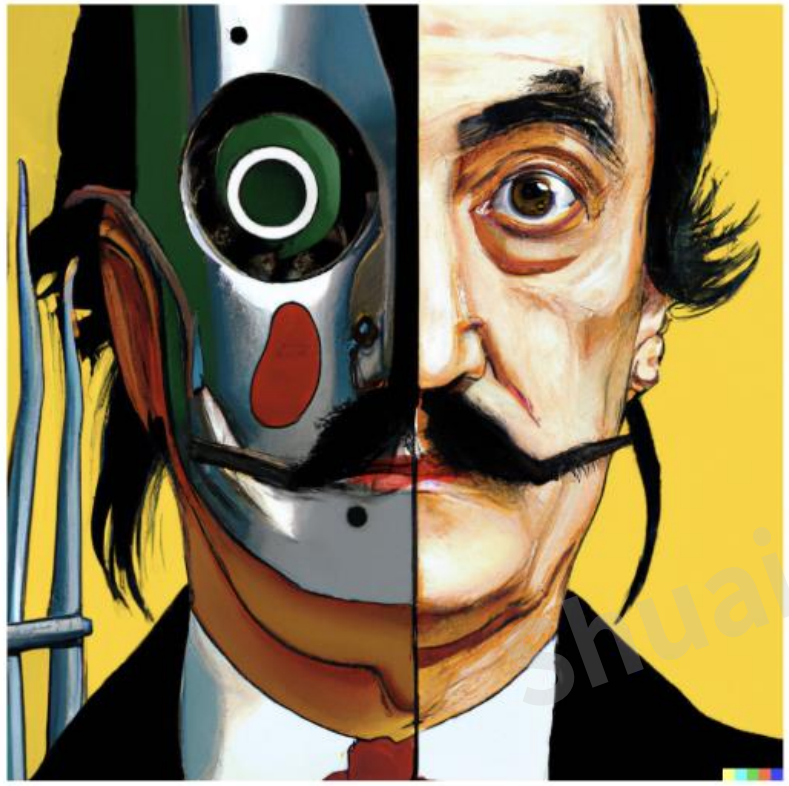
The text-to-image generation process



Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv'22

Image diffusion model

DALL·E 2



vibrant portrait painting of Salvador Dalí with a robotic half face

a shiba inu wearing a beret and black turtleneck

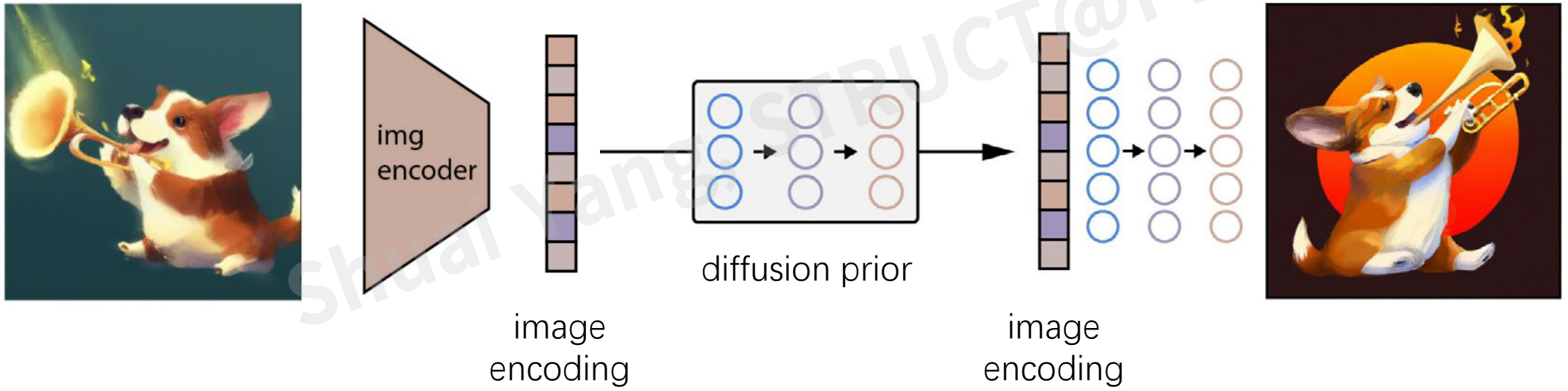
a close up of a handpalm with leaves growing from it

Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv'22

Image diffusion model

DALL·E 2

Encoding with CLIP and then decoding with a diffusion model



Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv'22

Image diffusion model

DALL·E 2

Encoding with CLIP and then decoding with a diffusion model

Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv'22

Image diffusion model

DALL·E 2

Interpolating CLIP image embeddings and then decoding with a diffusion model

Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv'22

Image diffusion model

DALL·E 2

Interpolating CLIP image embeddings and then decoding with a diffusion model

Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv'22

Image diffusion model

Latent Diffusion Models

To train diffusion models on **limited computational resources** while retaining quality and flexibility

- Not directly in the pixel space
- But using **the potential space**

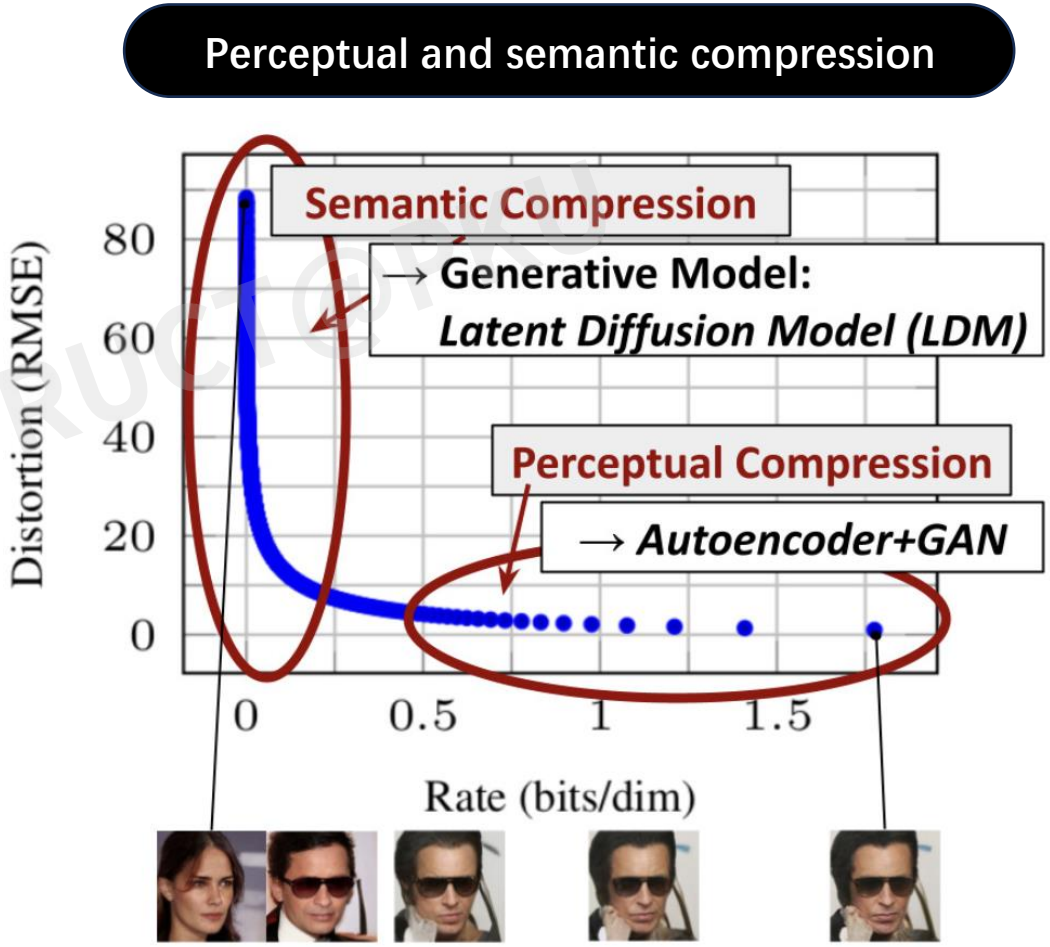
High-Resolution Image Synthesis with Latent Diffusion Models. CVPR'22

Image diffusion model

Latent Diffusion Models

To train diffusion models on **limited computational resources** while retaining quality and flexibility

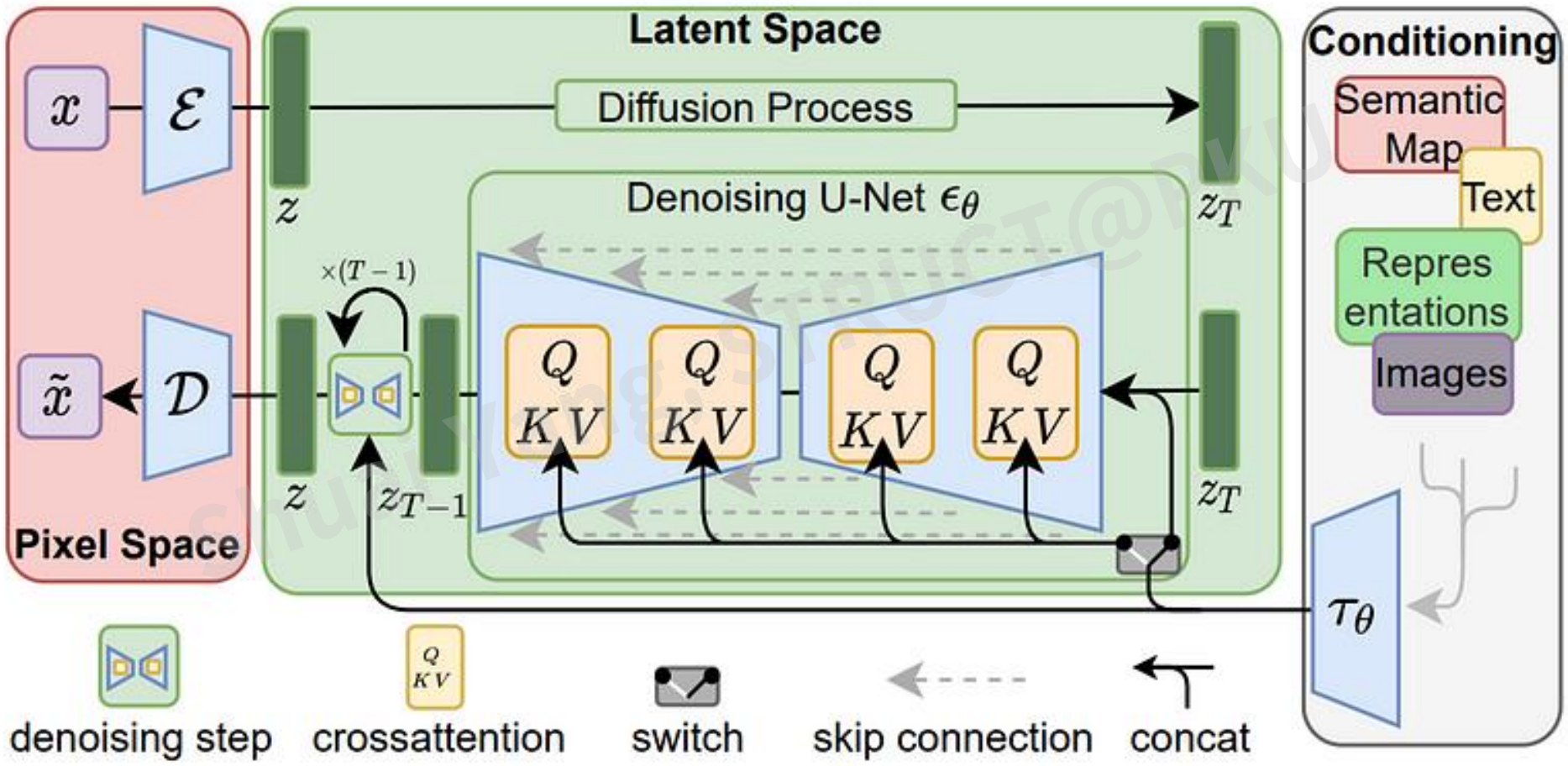
- Not directly in the pixel space
- But using **the potential space**



High-Resolution Image Synthesis with Latent Diffusion Models. CVPR'22

Image diffusion model

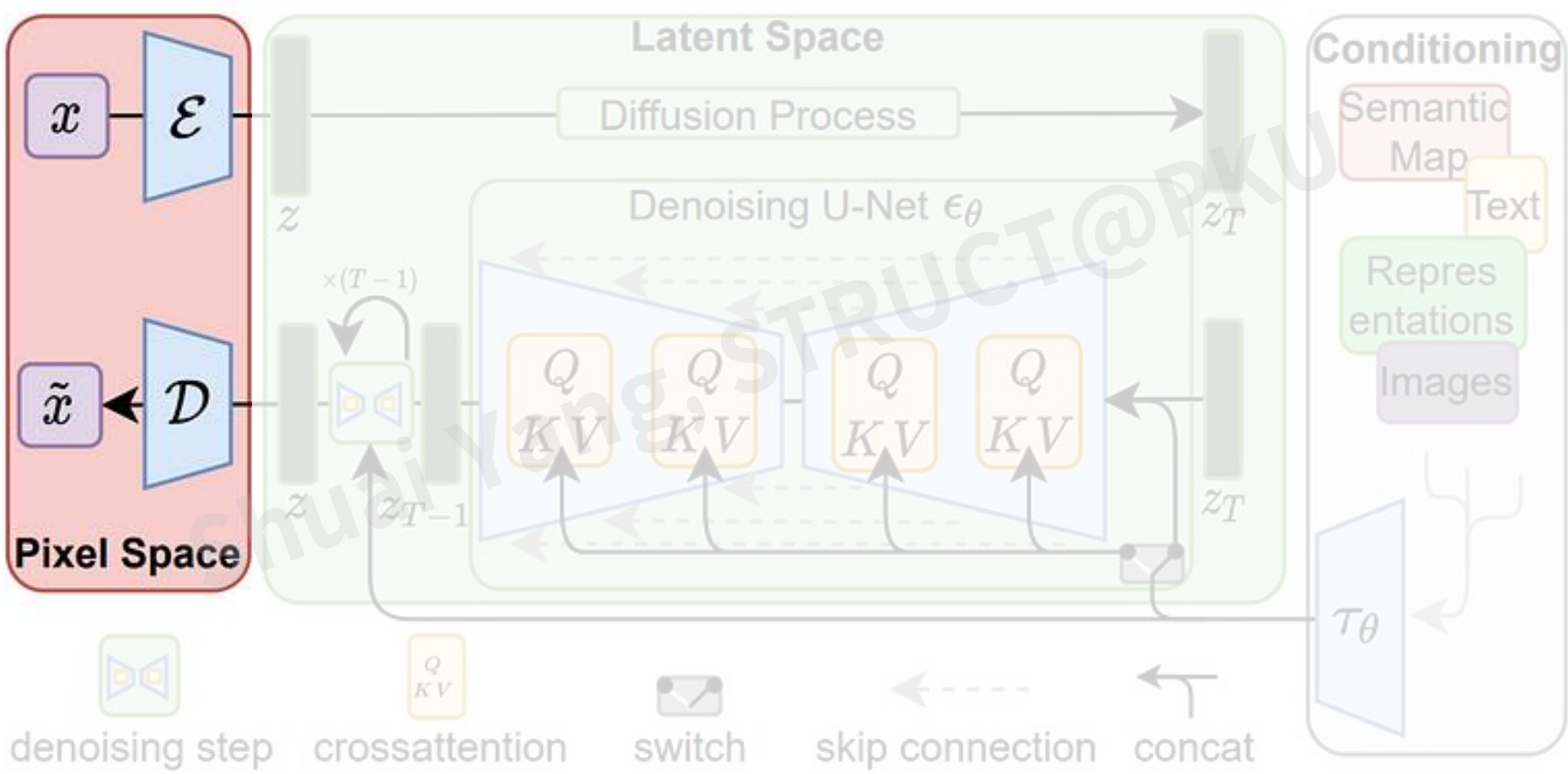
Latent Diffusion Models



High-Resolution Image Synthesis with Latent Diffusion Models. CVPR'22

Image diffusion model

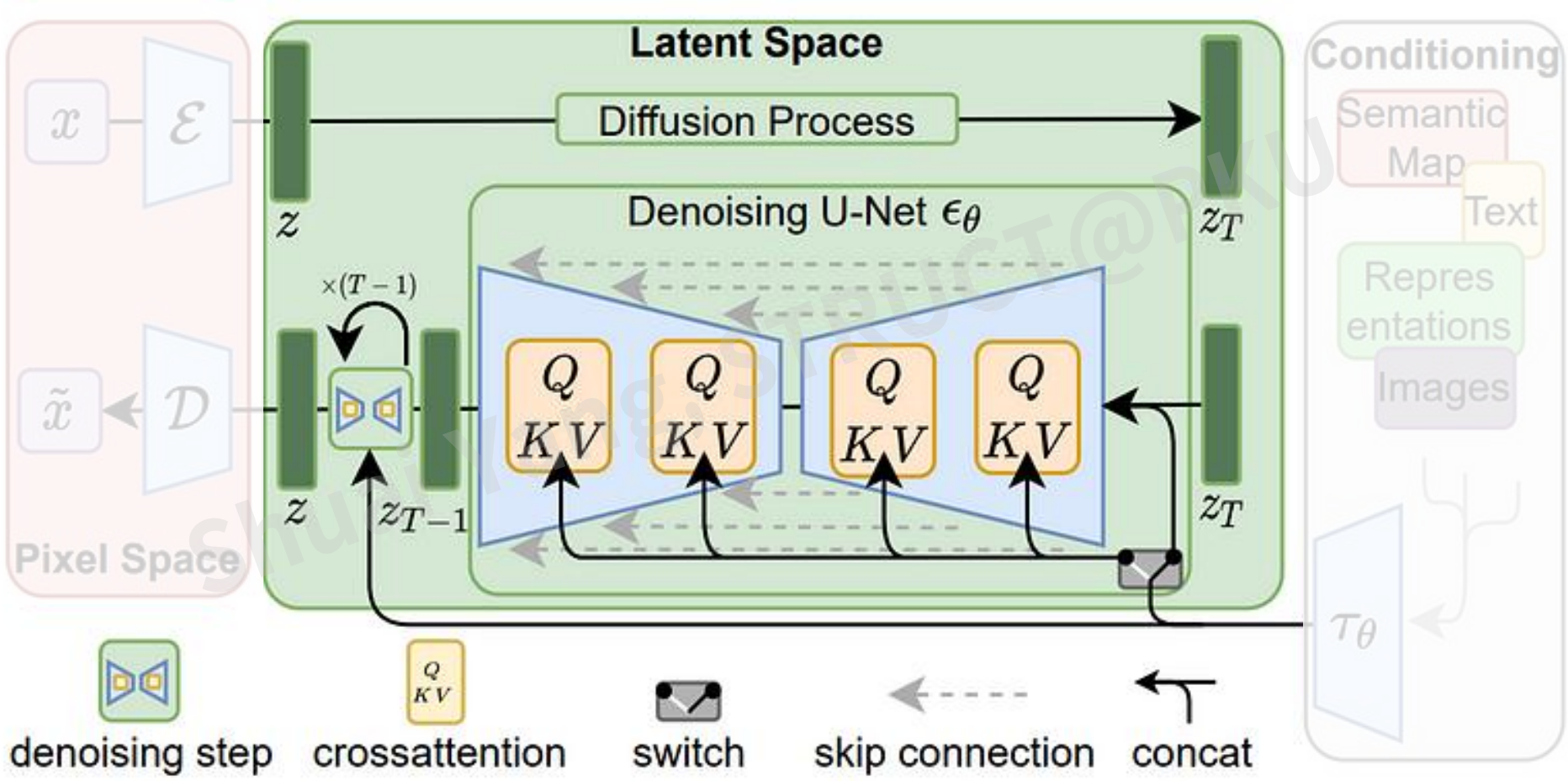
Latent Diffusion Models



High-Resolution Image Synthesis with Latent Diffusion Models. CVPR'22

Image diffusion model

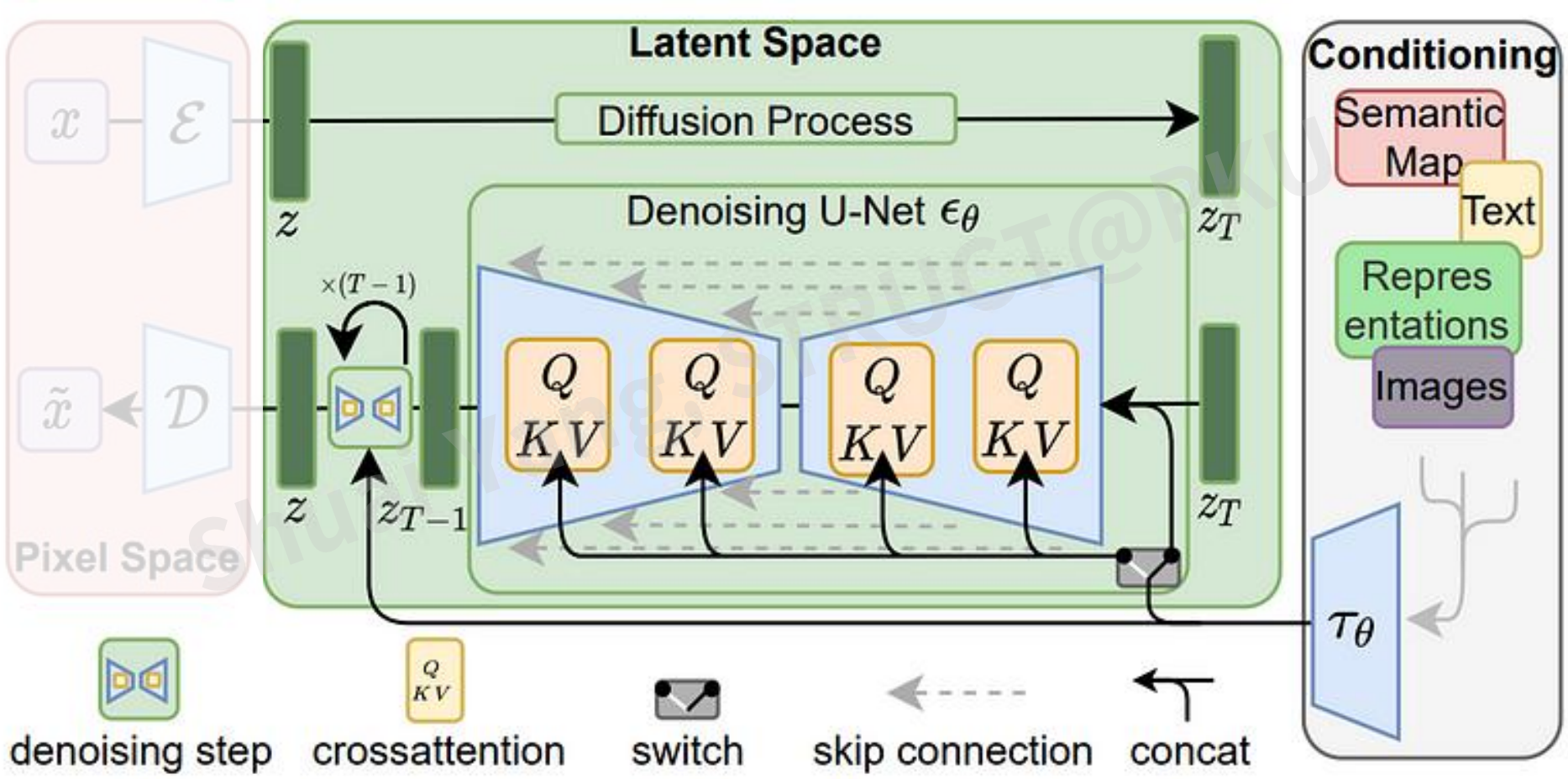
Latent Diffusion Models



High-Resolution Image Synthesis with Latent Diffusion Models. CVPR'22

Image diffusion model

Latent Diffusion Models



High-Resolution Image Synthesis with Latent Diffusion Models. CVPR'22

Image diffusion model

Stable Diffusion series

Image diffusion model

Stable Diffusion series

June 2022 — August 2022 — October 2022 — November 2022 — December 2022

SD 1.1
SD 1.2
SD 1.3

SD 1.4

SD 1.5

SD 2.0

SD 2.1

SD 1.x Models: resolution of 512x512 pixels and use a ViT-L/14 CLIP model for text conditioning

Image diffusion model

Stable Diffusion series

June 2022 — August 2022 — October 2022 — November 2022 — December 2022

SD 1.1
SD 1.2
SD 1.3

SD 1.4

SD 1.5

SD 2.0

SD 2.1

SD 2.x Models: resolution of 768x768 pixels and use a different text encoder: ViT-H/14 OpenCLIP, allowing the prompts to be more expressive.

Image diffusion model

Stable Diffusion series

December 2022 — SD 2.1

July 2023 — SD XL 1.0

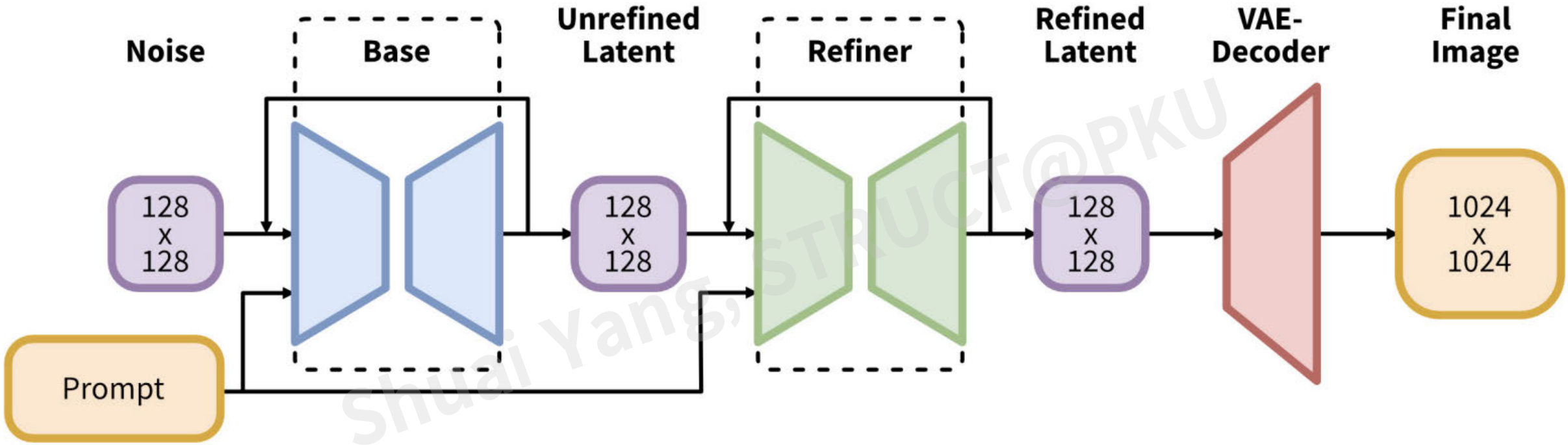
November 2023 — SD XL Turbo

February 2024 — SD Cascade

March 2024 — SD 3.0

Image diffusion model

Stable Diffusion XL



Number of parameters compared with older SD models:

- SD XL 2.6B v.s. SD 1.4/1.5 860M v.s. SD 2.0/2.1 865M

SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. ICLR'24

Image diffusion model

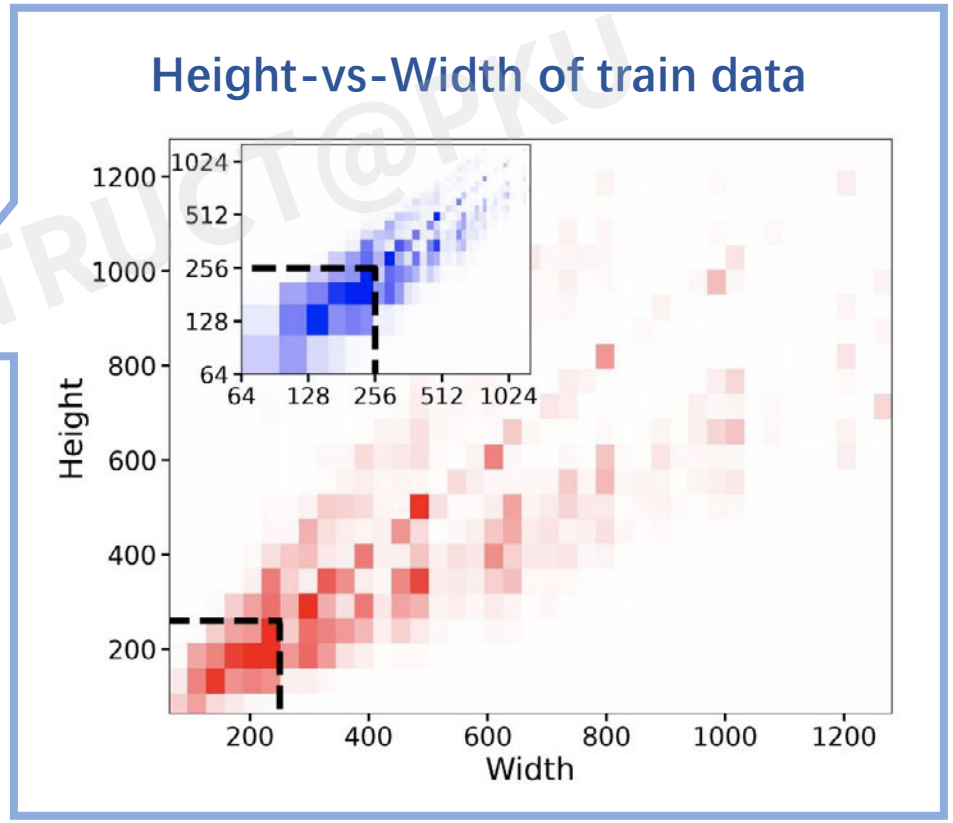
Stable Diffusion XL

The shortcoming of the LDM: training a model requires a minimal image size

Common solutions:

- Discard all training images below a certain minimal resolution.
  → Discard a lot of training data

- Upscale images that are too small.
  → Upscaling artifacts

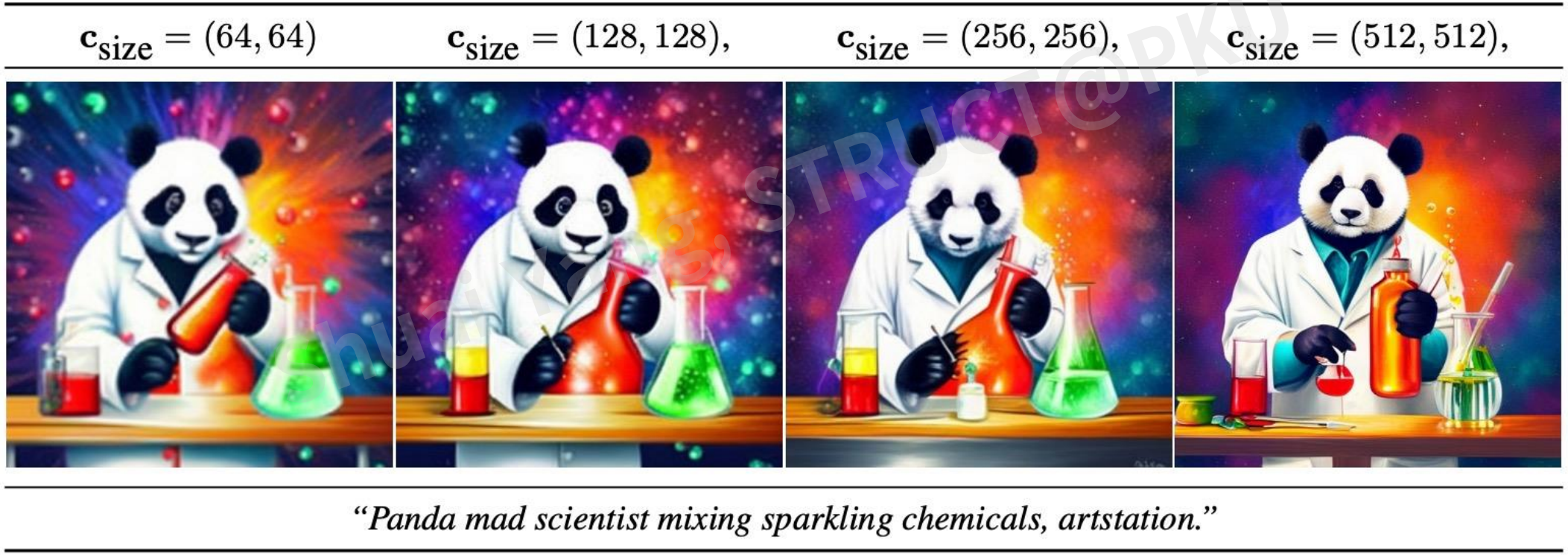


SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. ICLR'24

Image diffusion model

Stable Diffusion XL

Micro-Conditioning: 1. Size conditioning

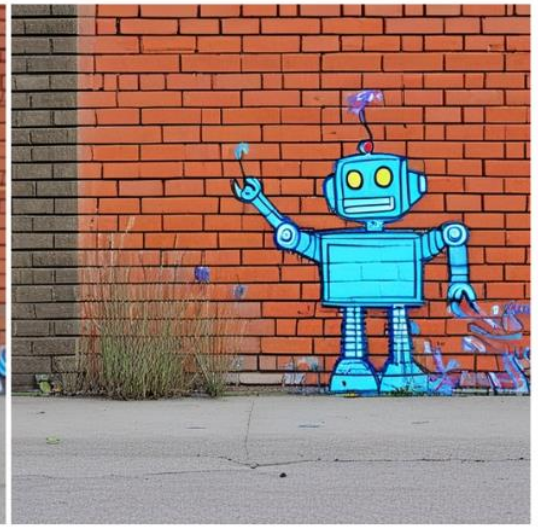


SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. ICLR'24

Image diffusion model

Stable Diffusion XL

Micro-Conditioning: 1. Size conditioning

$\mathbf{c}_{\text{size}} = (64, 64)$	$\mathbf{c}_{\text{size}} = (128, 128)$,	$\mathbf{c}_{\text{size}} = (256, 236)$,	$\mathbf{c}_{\text{size}} = (512, 512)$,
			

'A robot painted as graffiti on a brick wall. a sidewalk is in front of the wall, and grass is growing out of cracks in the concrete.'

SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. ICLR'24

Image diffusion model

Stable Diffusion XL

Failure cases of previous SD models

	"A propaganda poster depicting a cat dressed as french emperor napoleon holding a piece of cheese."	*"a close-up of a fire spitting dragon, cinematic shot."*
***SD* 1-5**		
***SD* 2-1**		

SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. ICLR'24

Image diffusion model

Stable Diffusion XL

Failure cases of previous SD models

	"A propaganda poster depicting a cat dressed as french emperor napoleon holding a piece of cheese."	*"a close-up of a fire spitting dragon, cinematic shot."*
***SD* 1-5**		
***SD* 2-1**		

SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. ICLR'24

Image diffusion model

Stable Diffusion XL

Micro-Conditioning: 2. Cropping parameters $(c_{\text{top}}, c_{\text{left}})$

SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. ICLR'24

Image diffusion model

Stable Diffusion XL

Micro-Conditioning: 2. Cropping parameters $(c_{\text{top}}, c_{\text{left}})$

$\mathbf{c}_{\text{crop}} = (0, 0)$	$\mathbf{c}_{\text{crop}} = (0, 256)$,	$\mathbf{c}_{\text{crop}} = (256, 0)$,	$\mathbf{c}_{\text{crop}} = (512, 512)$,
			

'An astronaut riding a pig, highly realistic dslr photo, cinematic shot.'

SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. ICLR'24

Image diffusion model

Stable Diffusion XL

Multi-Aspect Training

- Partition the data into buckets of different aspect ratios
- Keep the pixel count close to 1024^2 pixels
- Bucket size as a conditioning

Height	Width	Aspect Ratio
512	2048	0.25
512	1984	0.26
512	1920	0.27
512	1856	0.28
576	1792	0.32
576	1728	0.33
576	1664	0.35
640	1600	0.4
640	1536	0.42
704	1472	0.48
704	1408	0.5
704	1344	0.52
768	1344	0.57
768	1280	0.6
832	1216	0.68
832	1152	0.72
896	1152	0.78
896	1088	0.82
960	1088	0.88
960	1024	0.94

Height	Width	Aspect Ratio
1024	1024	1.0
1024	960	1.07
1088	960	1.13
1088	896	1.21
1152	896	1.29
1152	832	1.38
1216	832	1.46
1280	768	1.67
1344	768	1.75
1408	704	2.0
1472	704	2.09
1536	640	2.4
1600	640	2.5
1664	576	2.89
1728	576	3.0
1792	576	3.11
1856	512	3.62
1920	512	3.75
1984	512	3.88
2048	512	4.0

SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. ICLR'24

Image diffusion model

Stable Diffusion XL

SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. ICLR'24

Image diffusion model

Stable Diffusion XL

SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. ICLR'24

Image diffusion model

Stable Diffusion XL

SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. ICLR'24

Image diffusion model

Stable Diffusion series

December 2022 — SD 2.1

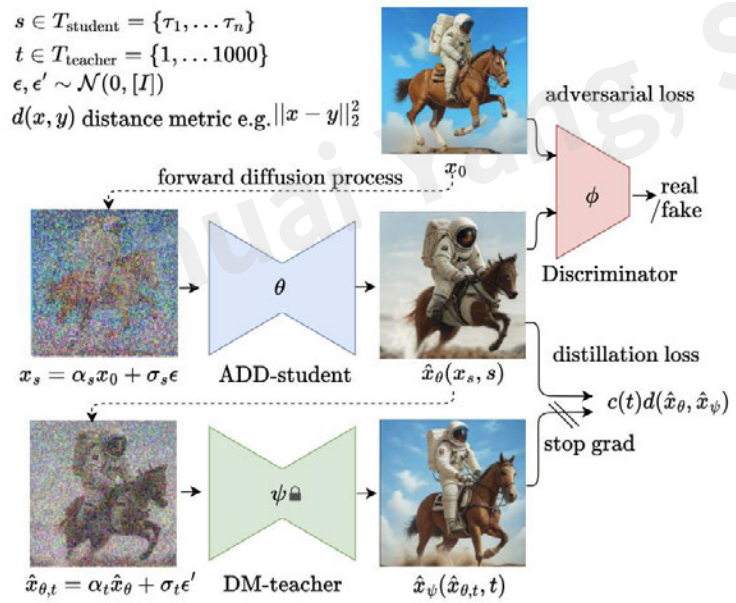
July 2023 — SD XL 1.0

November 2023 — SD XL Turbo

February 2024 — SD Cascade

March 2024 — SD 3.0

Adversarial Diffusion Distillation



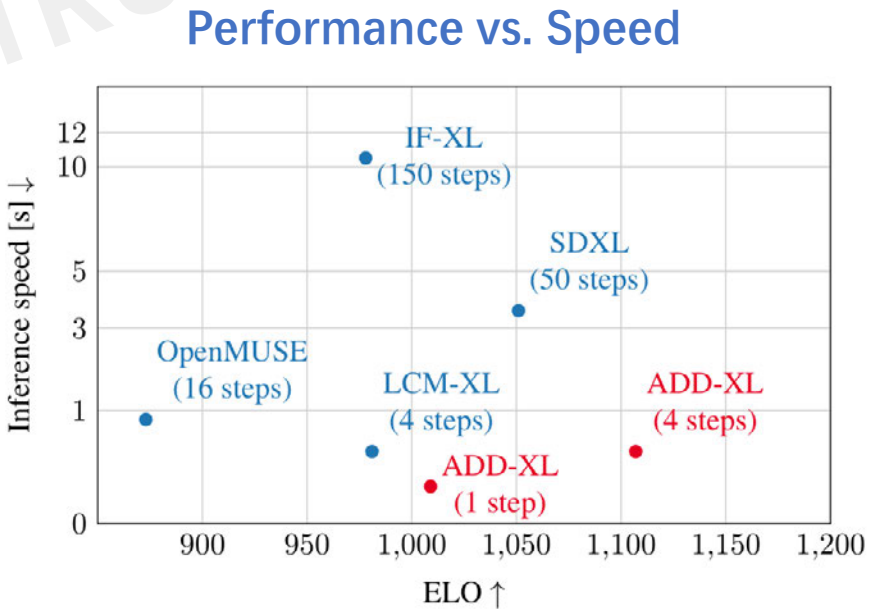


Image diffusion model

Stable Diffusion series

December 2022 — **SD 2.1**

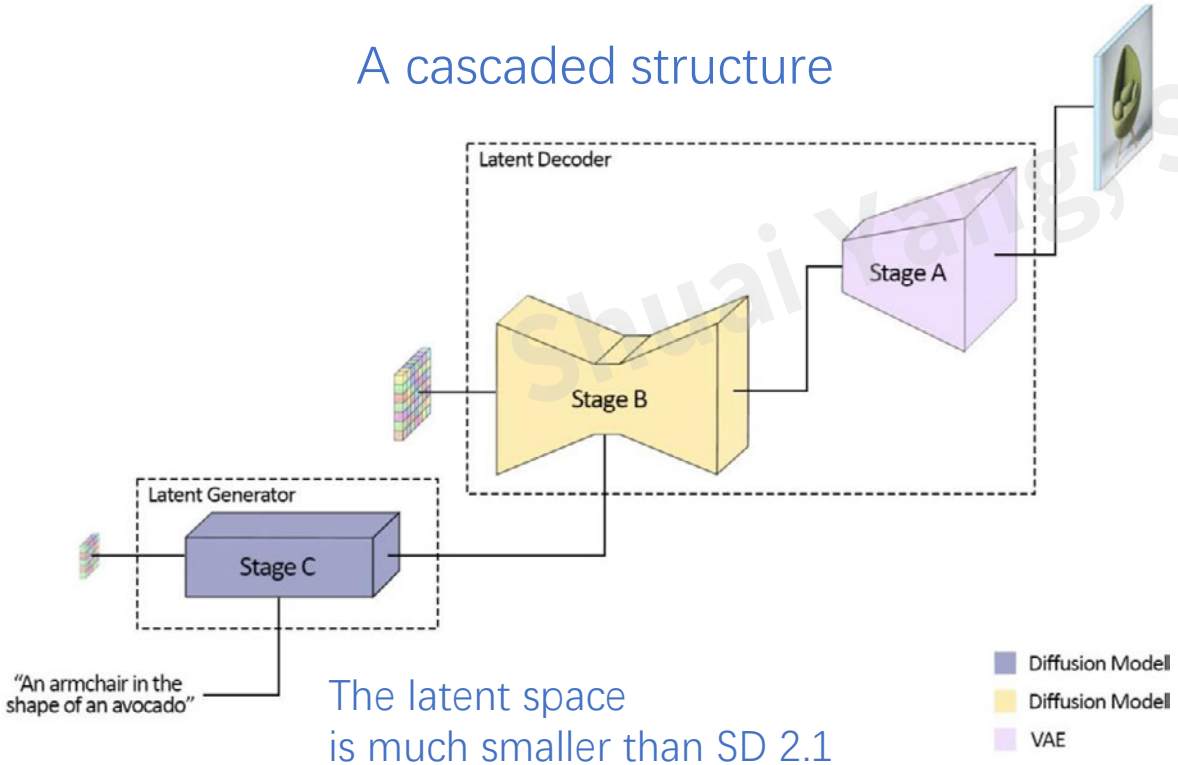
July 2023 — **SD XL 1.0**

November 2023 — **SD XL Turbo**

February 2024 — **SD Cascade**

March 2024 — **SD 3.0**

A cascaded structure



The latent space
is much smaller than SD 2.1

Image diffusion model

Stable Diffusion series

December 2022 — SD 2.1

July 2023 — SD XL 1.0

November 2023 — SD XL Turbo

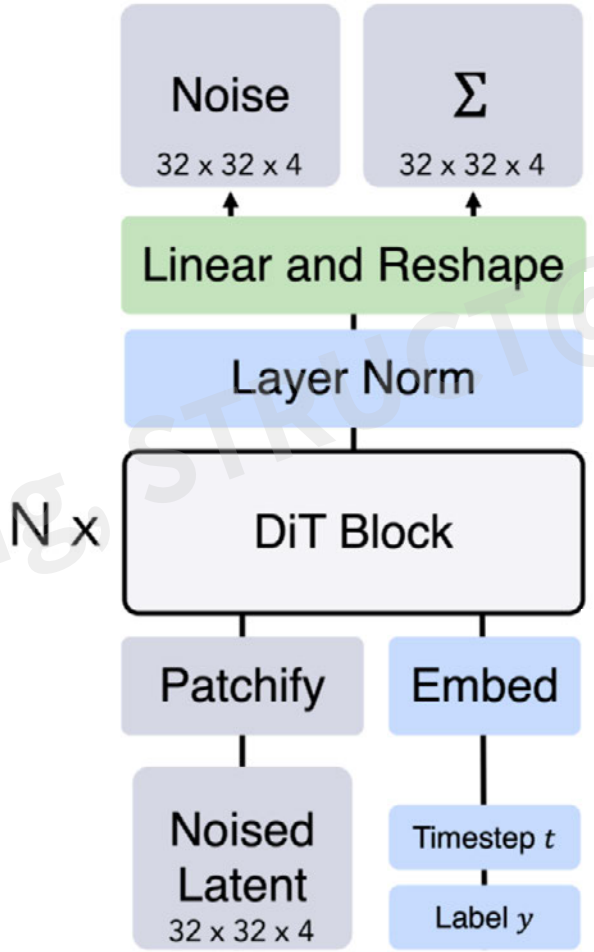
February 2024 — SD Cascade

March 2024 — SD 3.0

Image diffusion model

Diffusion Transformers

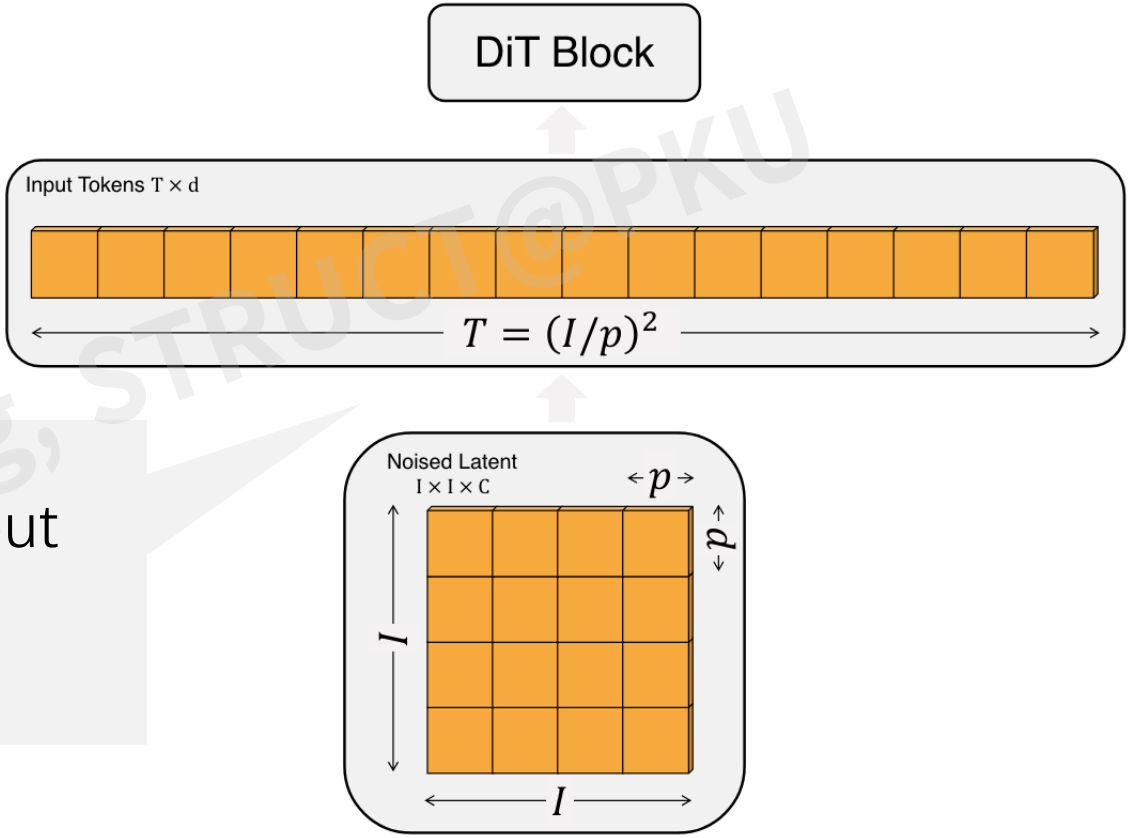
Latent Diffusion + Transformer



Scalable Diffusion Models with Transformers. ICCV'23

Image diffusion model

Diffusion Transformers

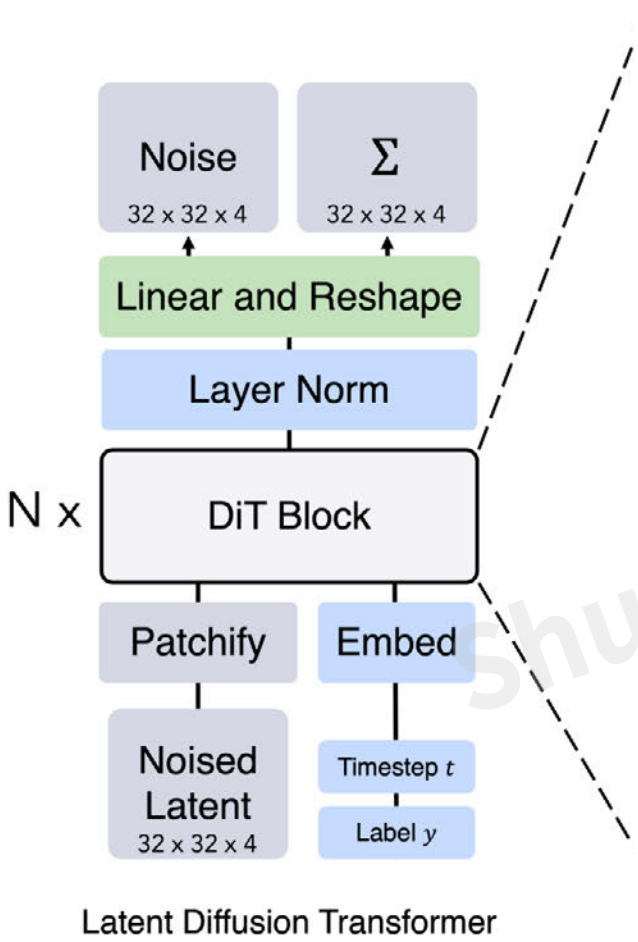


Patchify: convert the spatial input into a sequence of T tokens

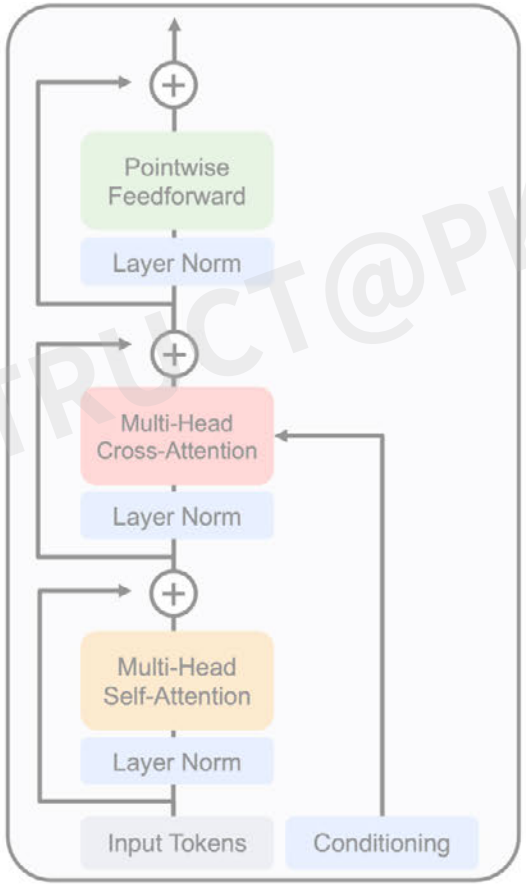
Scalable Diffusion Models with Transformers. ICCV'23

Image diffusion model

Diffusion Transformers



Latent Diffusion Transformer

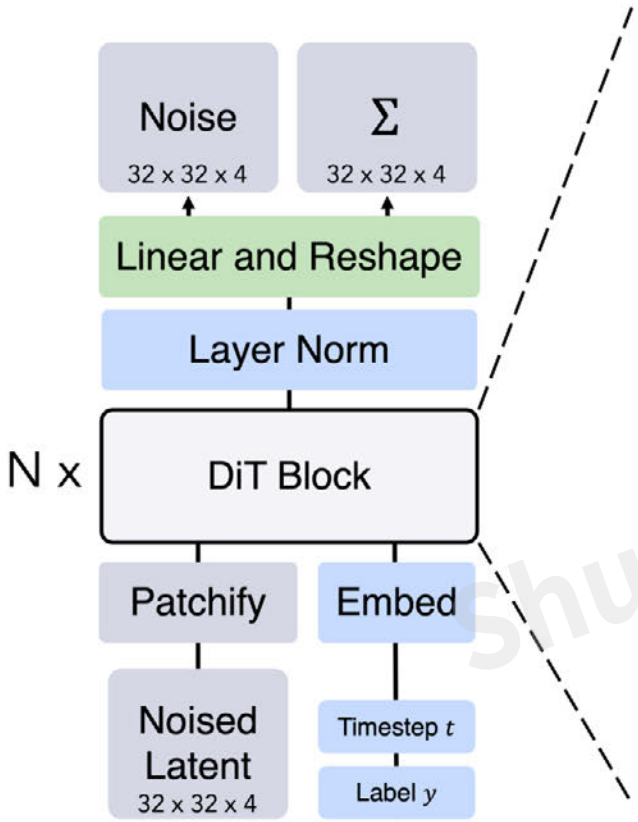


DiT Block with Cross-Attention

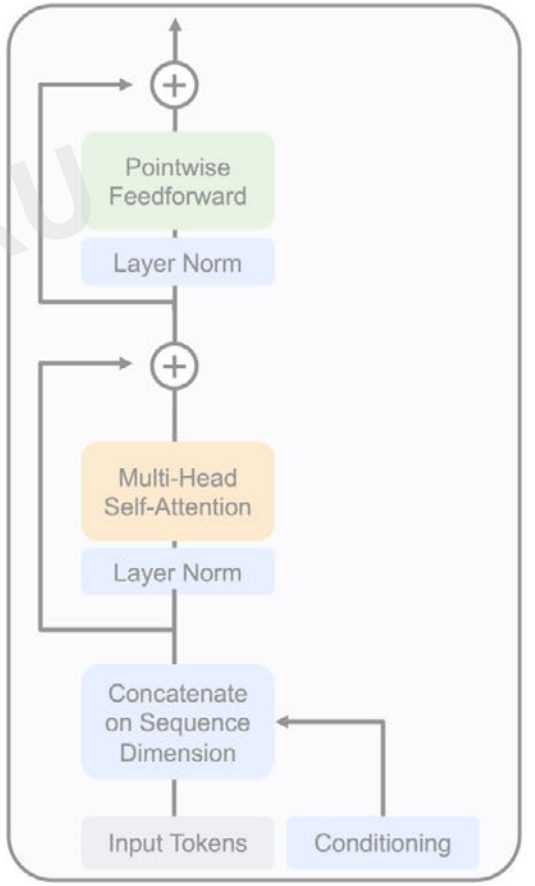
Scalable Diffusion Models with Transformers. ICCV'23

Image diffusion model

Diffusion Transformers



Latent Diffusion Transformer

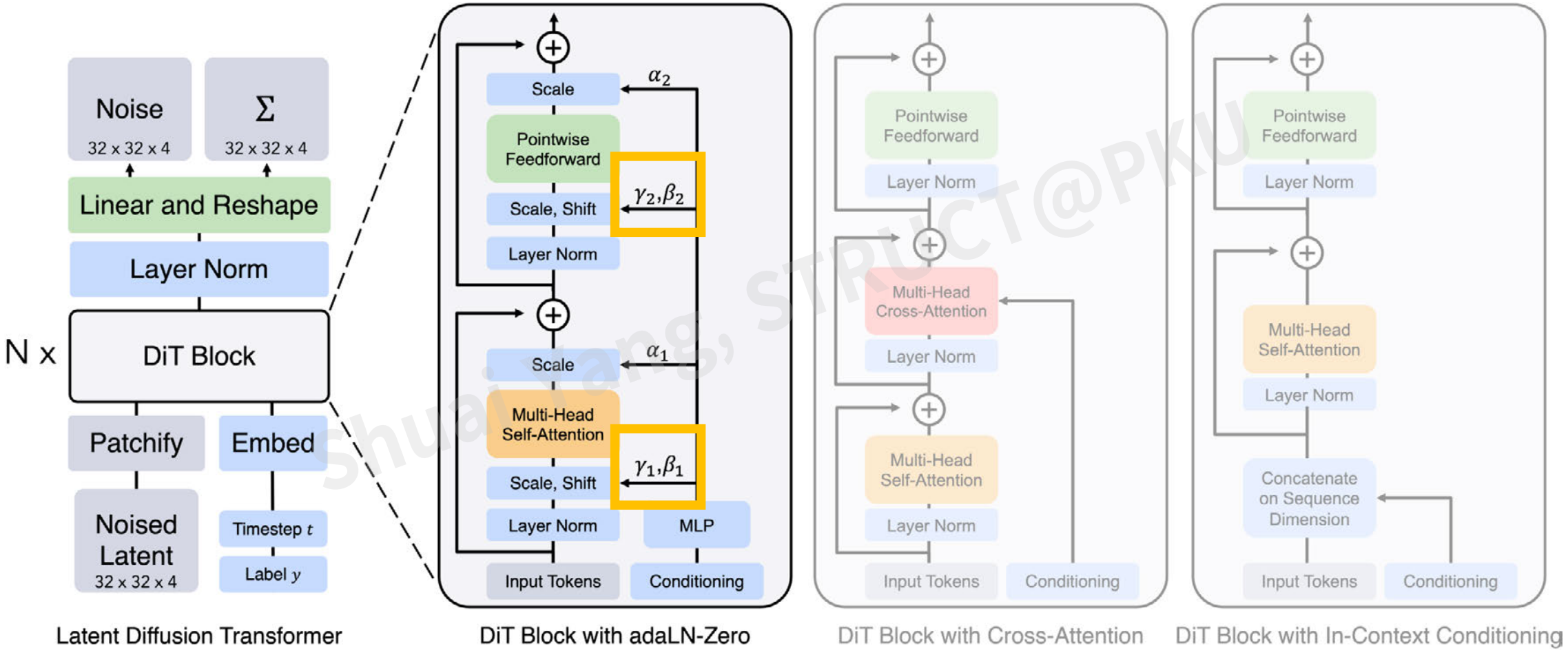


DiT Block with In-Context Conditioning

Scalable Diffusion Models with Transformers. ICCV'23

Image diffusion model

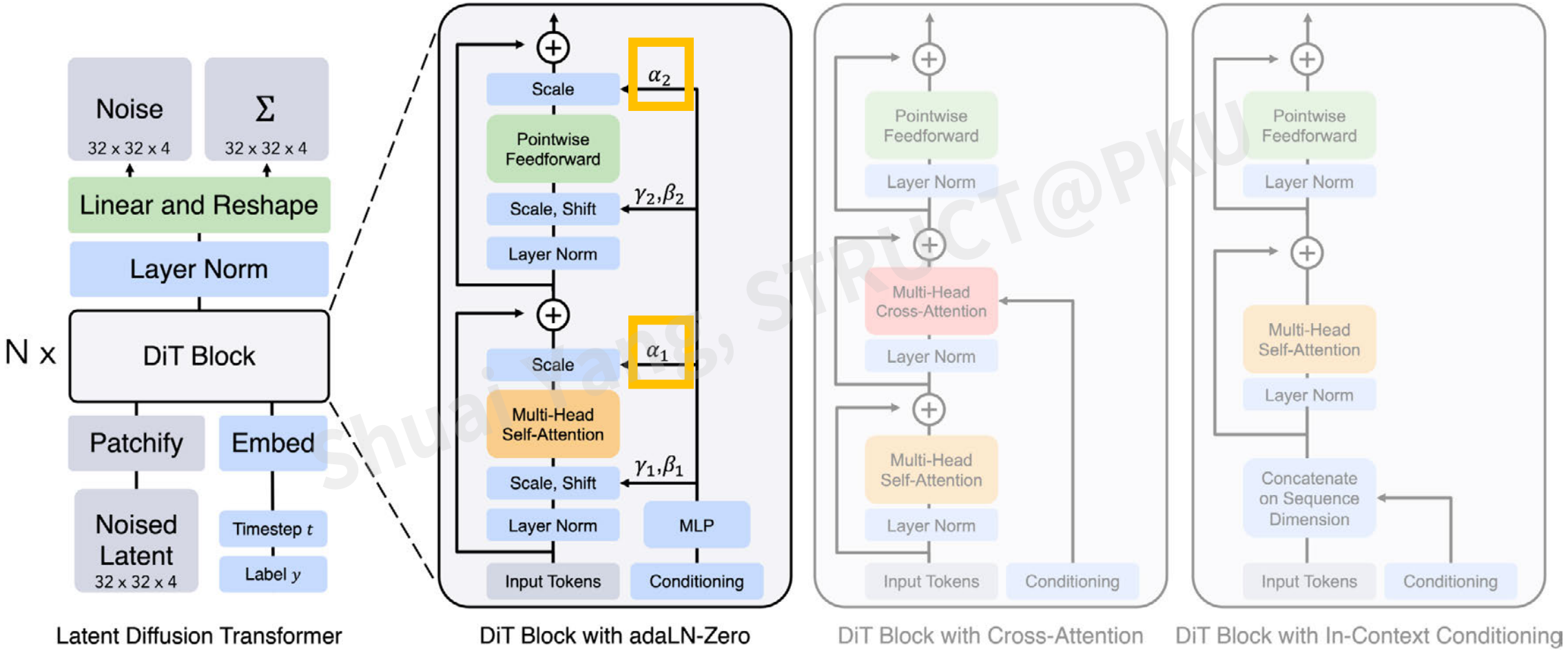
Diffusion Transformers



Scalable Diffusion Models with Transformers. ICCV'23

Image diffusion model

Diffusion Transformers

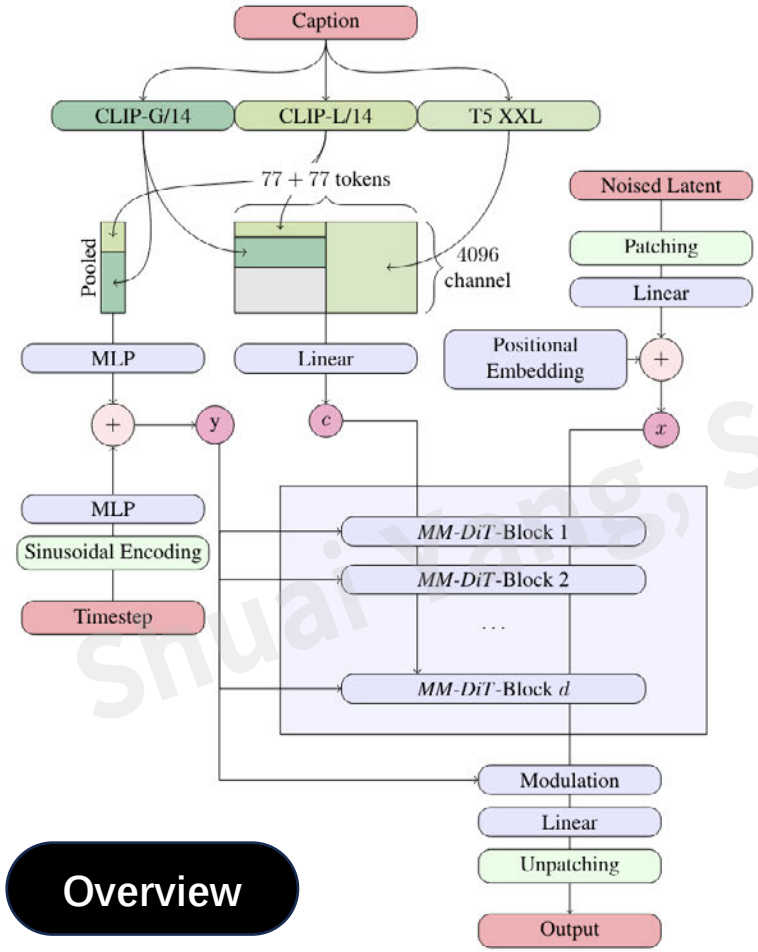


Latent Diffusion Transformer | DiT Block with adaLN-Zero | DiT Block with Cross-Attention | DiT Block with In-Context Conditioning

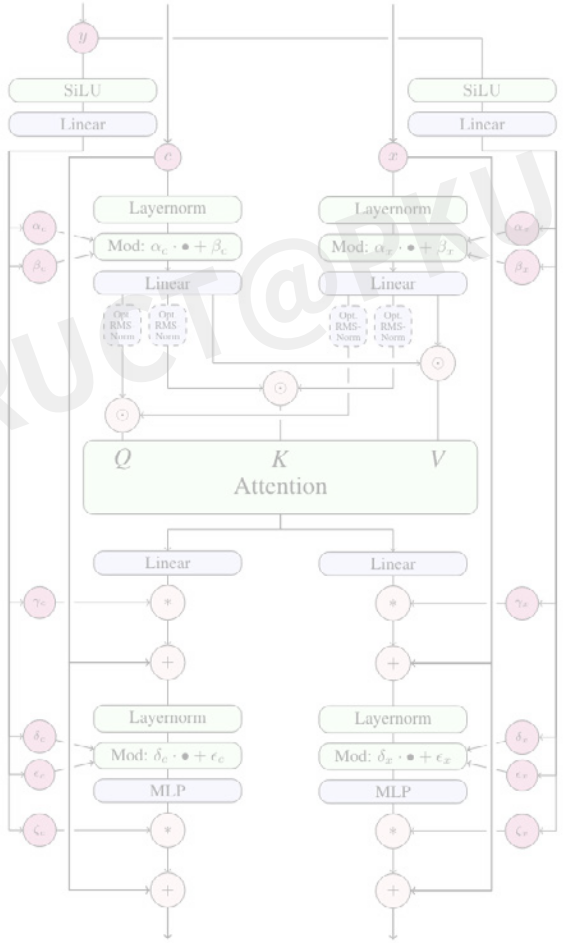
Scalable Diffusion Models with Transformers. ICCV'23

Image diffusion model

Stable Diffusion 3



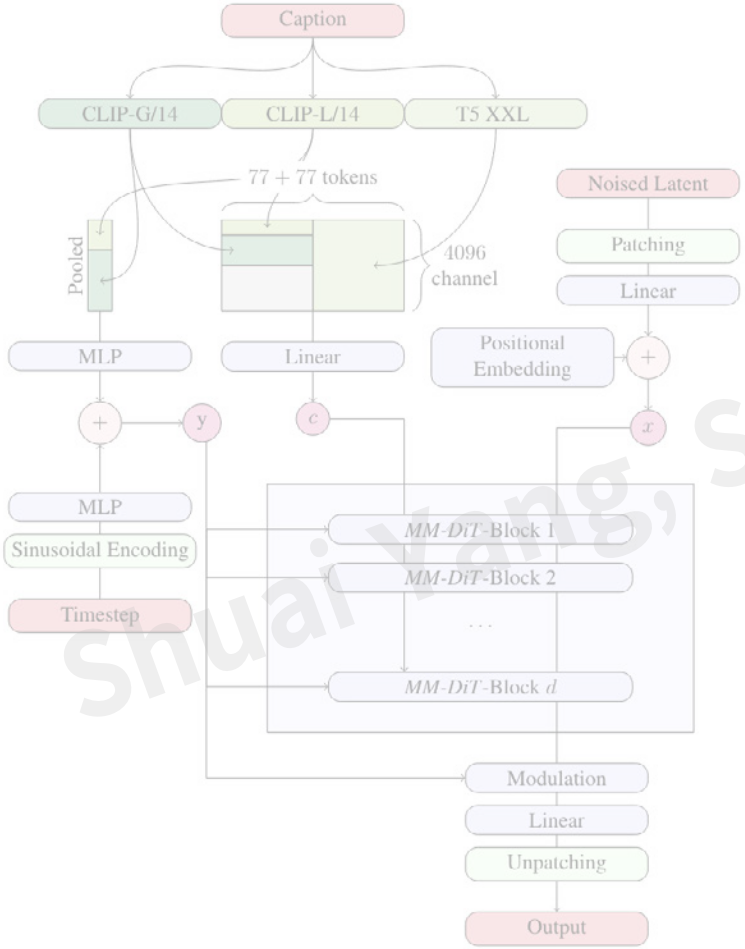
Overview

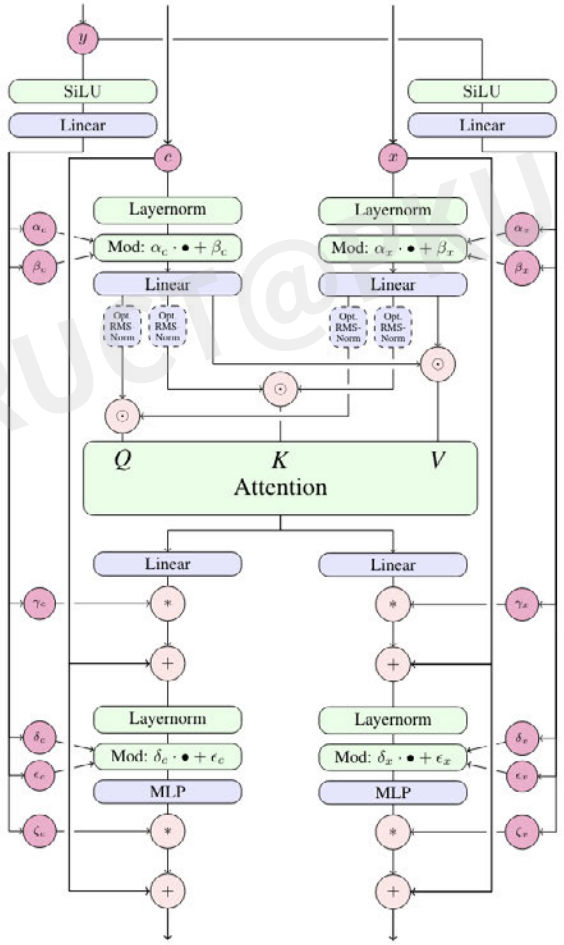


Scaling Rectified Flow Transformers for High-Resolution Image Synthesis. arXiv'24

Image diffusion model

Stable Diffusion 3





Two separate sets of weights for image and text

Scaling Rectified Flow Transformers for High-Resolution Image Synthesis. arXiv'24

Image diffusion model

Stable Diffusion 3

an old rusted robot wearing pants and a jacket riding skis in a supermarket.

smiling cartoon dog sits at a table, coffee mug on hand, as a room goes up in flames. “This is fine,” the dog assures himself.

Scaling Rectified Flow Transformers for High-Resolution Image Synthesis. arXiv’24

Image diffusion model

Stable Diffusion 3

Scaling Rectified Flow Transformers for High-Resolution Image Synthesis. arXiv'24

Image diffusion model

DALL·E 3

Compliance of visual content to text prompts

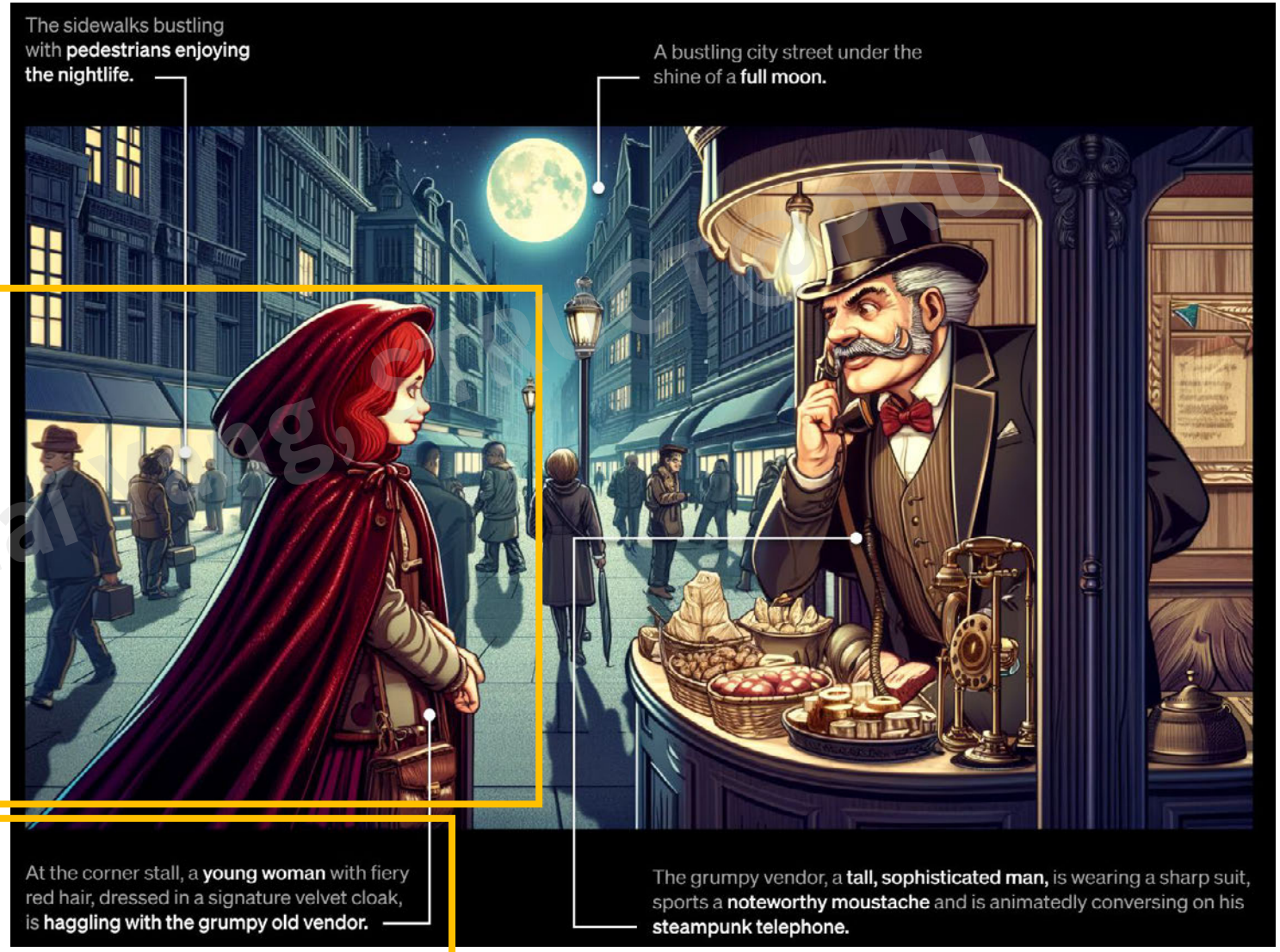


Image diffusion model

DALL·E 3

Problem of data scraped from the internet

Alt Text: now at victorian plumbing.co.uk

- **Problem 1: Incomplete descriptions.**
- **Problem 2: Irrelevant descriptions.**

Image diffusion model

DALL·E 3

Dataset Recaptioning

Short Synthetic Captions:

a white modern bathtub sits on a wooden floor

Descriptive Synthetic Captions:

this luxurious bathroom features a modern freestanding bathtub in a crisp white finish. the tub sits against a wooden accent wall with glass-like panels, creating a serene and relaxing ambiance. three pendant light fixtures hang above the tub, adding a touch of sophistication. a large window with a wooden panel provides natural light, while a potted plant adds a touch of greenery. the freestanding bathtub stands out as a statement piece in this contemporary bathroom.

Image diffusion model

DALL·E 3

DALL·E 3 + ChatGPT = Interactive and more user-friendly

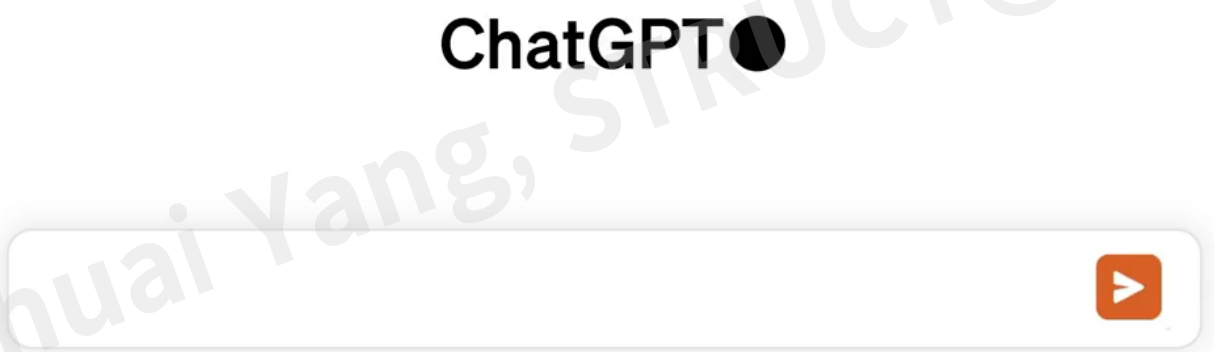


Image diffusion model

Midjourney

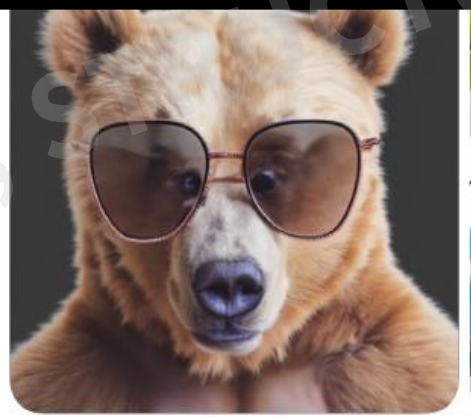
MIDJOURNEYAI

Image Supporting Model

swimming

sleeping

getting a haircut



worn by a bear

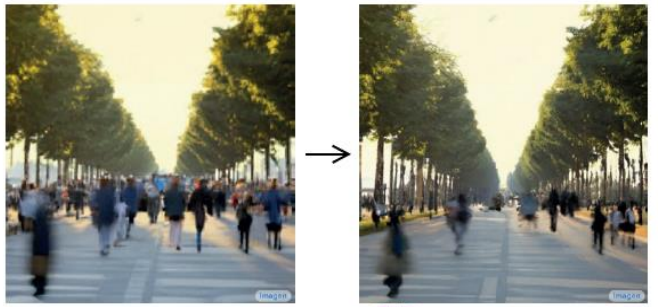
in the jungle

on red fabric

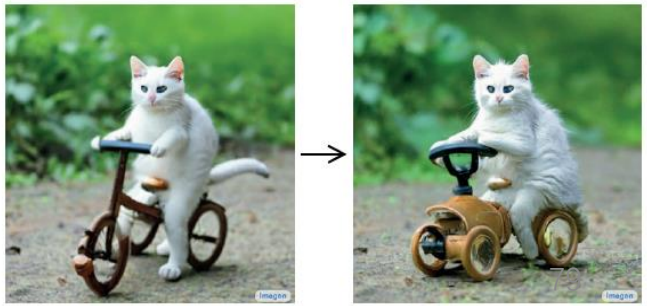
at Mt. Fuji

on top of snow

with Eiffel Tower



"The boulevards are crowded today."



"Photo of a cat riding on a bicycle."

Image supporting model

Approaches

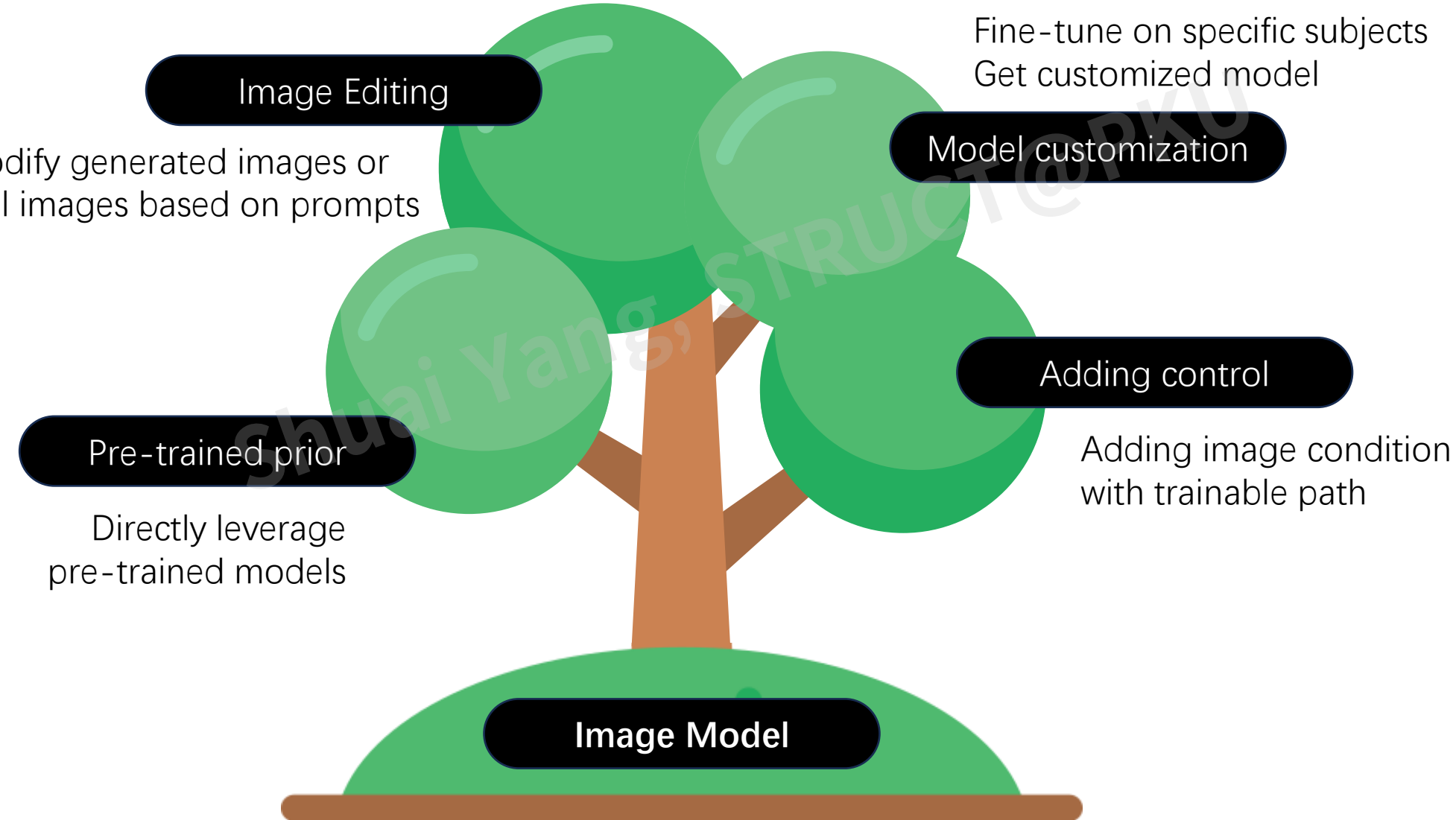
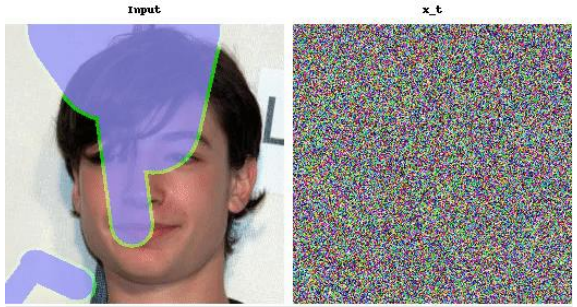


Image supporting model



Pre-trained prior

- No additional training
- Directly leveraging generative prior to specific task
- Limited customizability

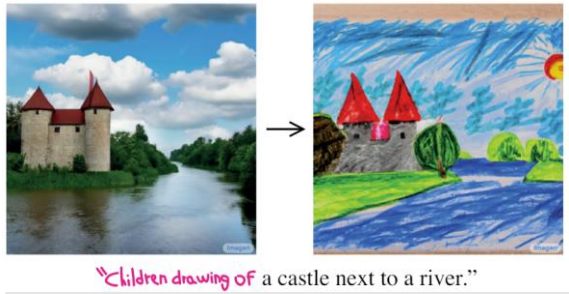


Image editing

- No additional training
- Use cross-attention map between text and image feature
- Cannot introduce new modalities

Model customization

- Subject-specific generation
- Highly customized
- Needs finetune on each case

Adding control

- Diverse modalities as guidance
- Highly controllable
- Needs training on customized dataset

Pre-trained models as prior

1

RePaint

CVPR 22

2

SDEdit

CVPR 22

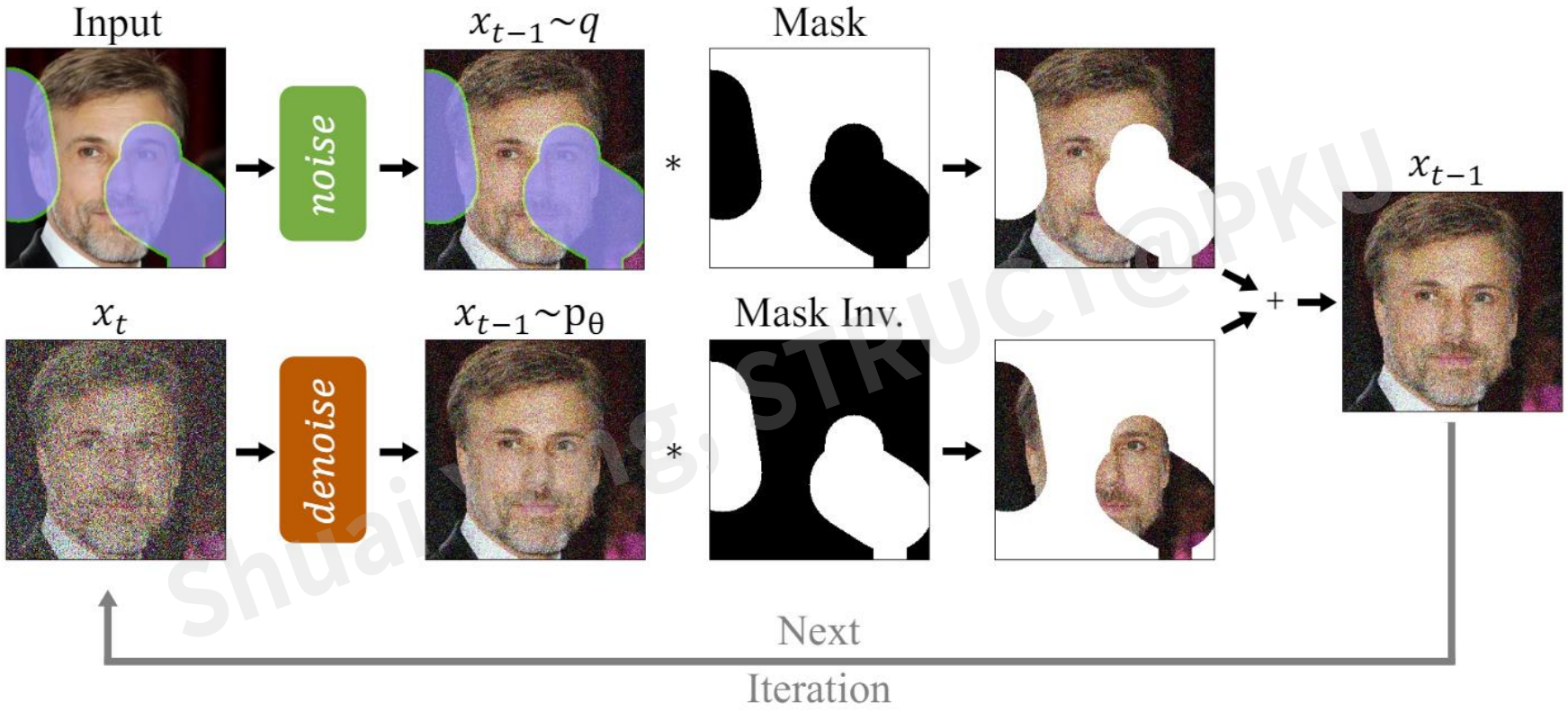
Pre-trained models as prior

1 RePaint

RePaint: Inpainting Using Denoising Diffusion Probabilistic Models. CVPR'22

Pre-trained models as prior

1 RePaint

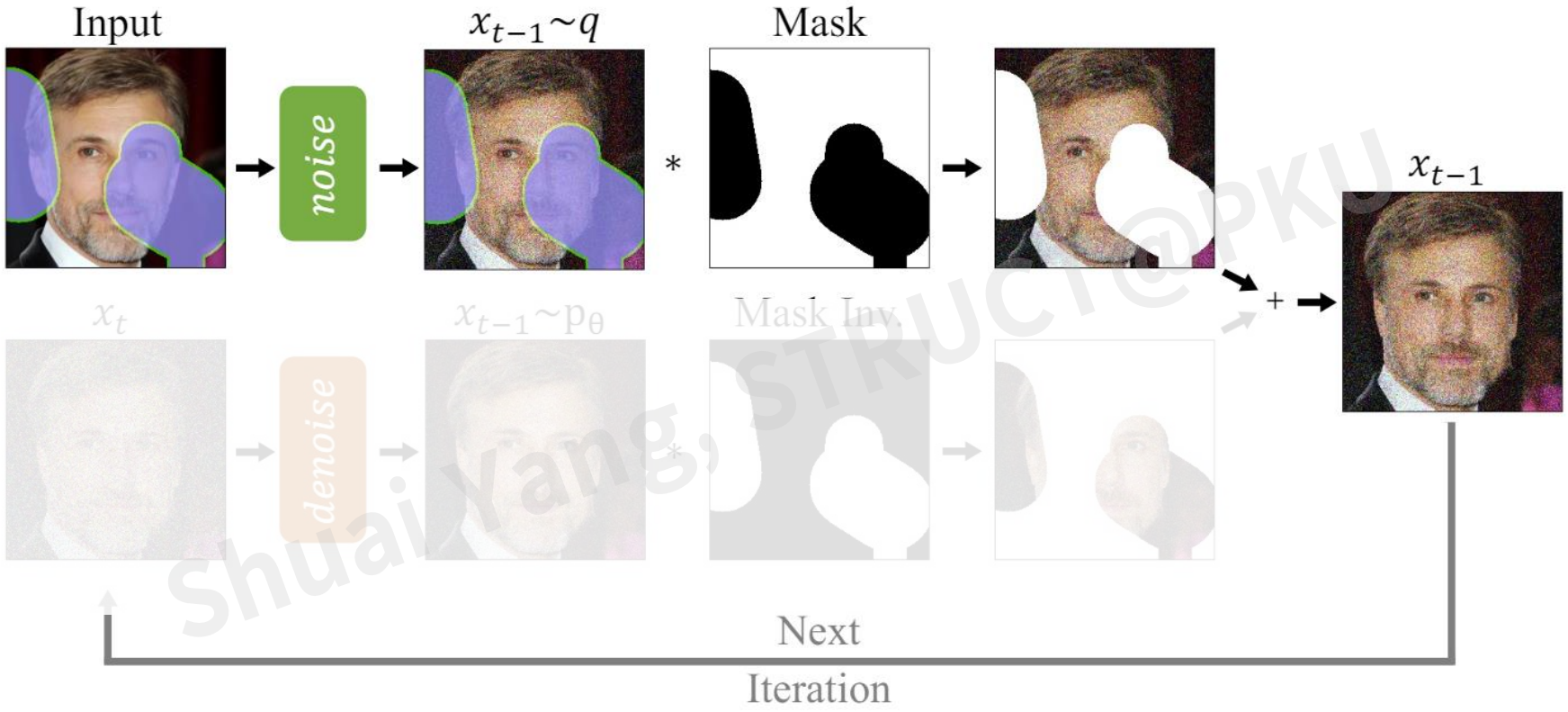


Iteratively generate missing region based on visible parts

RePaint: Inpainting Using Denoising Diffusion Probabilistic Models. CVPR'22

Pre-trained models as prior

1 RePaint

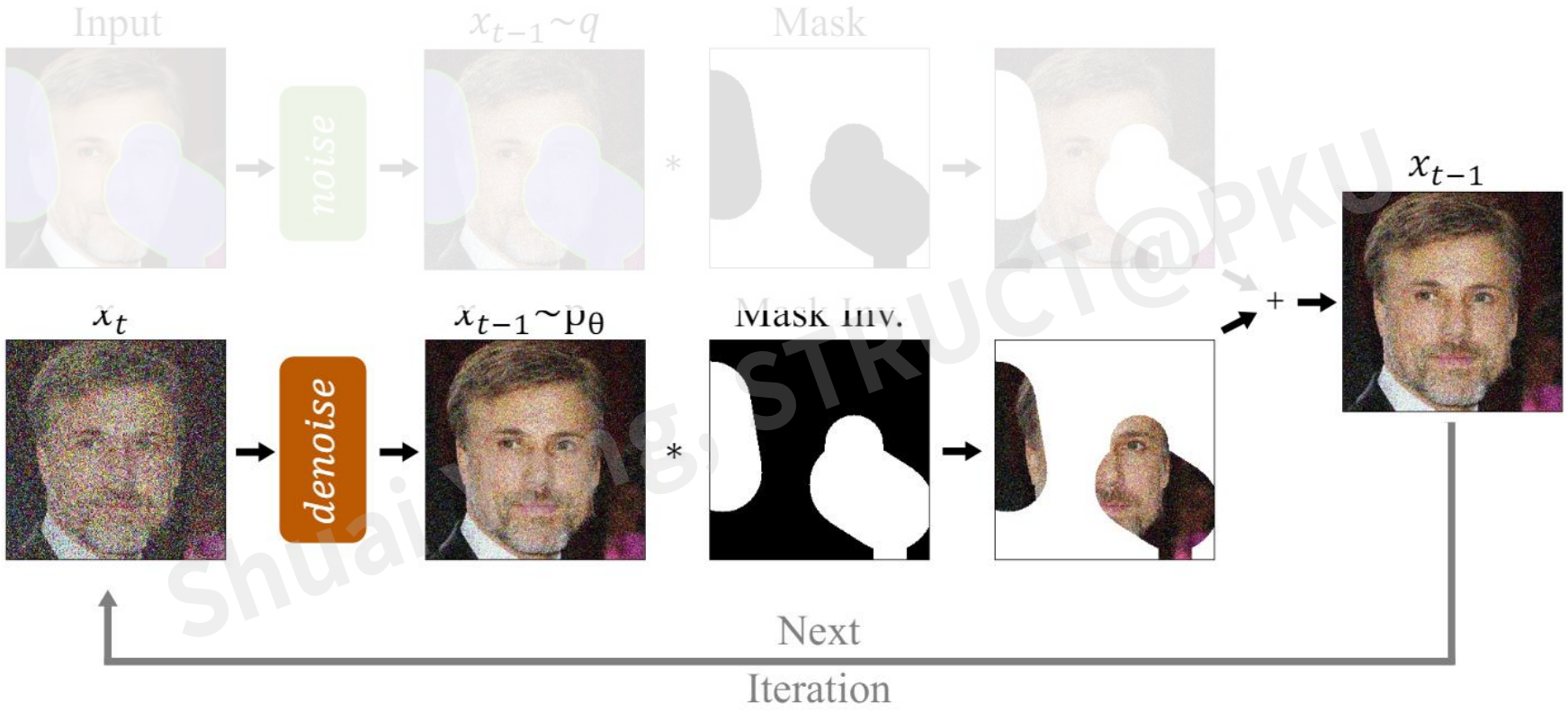


Iteratively generate missing region based on visible parts

RePaint: Inpainting Using Denoising Diffusion Probabilistic Models. CVPR'22

Pre-trained models as prior

1 RePaint

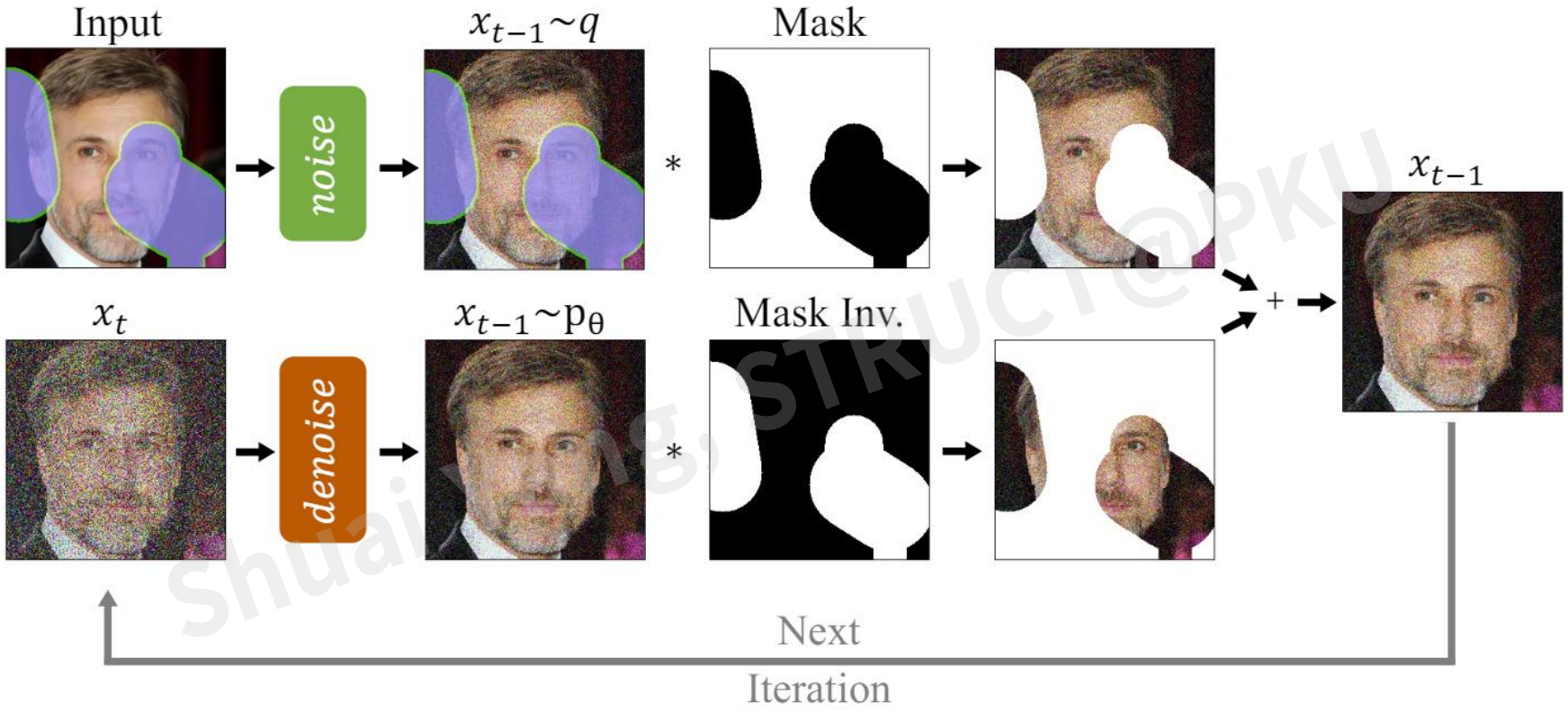


Iteratively generate missing region based on visible parts

RePaint: Inpainting Using Denoising Diffusion Probabilistic Models. CVPR'22

Pre-trained models as prior

1 RePaint



Iteratively generate missing region based on visible parts

RePaint: Inpainting Using Denoising Diffusion Probabilistic Models. CVPR'22

Pre-trained models as prior

1 RePaint

Multiple tasks as inpainting

RePaint: Inpainting Using Denoising Diffusion Probabilistic Models. CVPR'22

Pre-trained models as prior

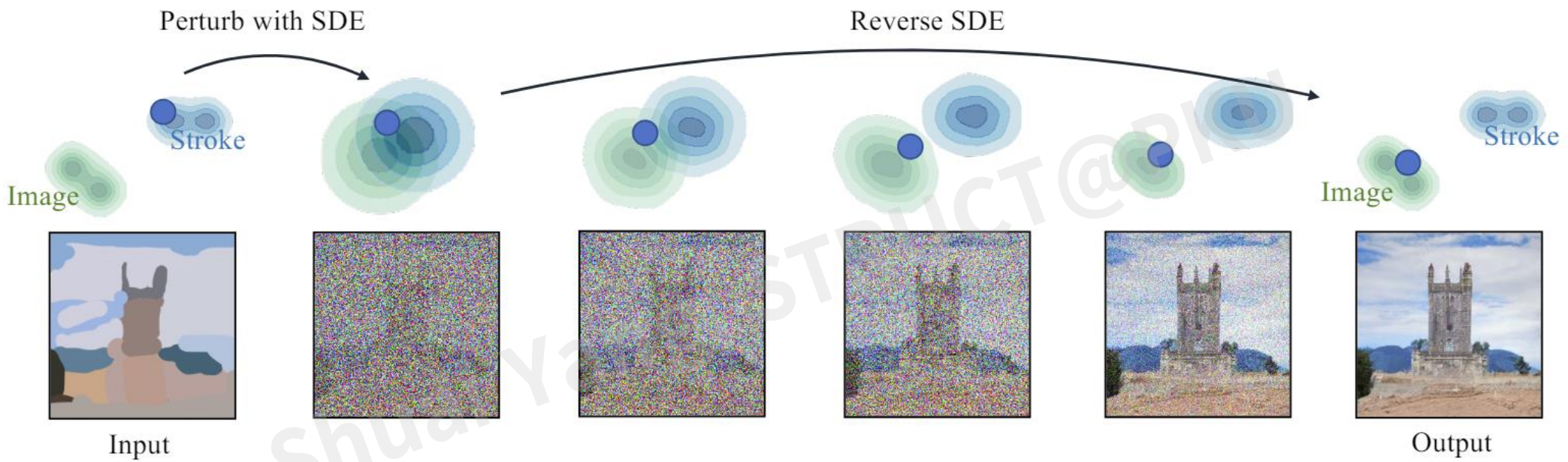
2 SDEdit

Cross-domain transformation

SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations. ICLR'22

Pre-trained models as prior

2 SDEdit

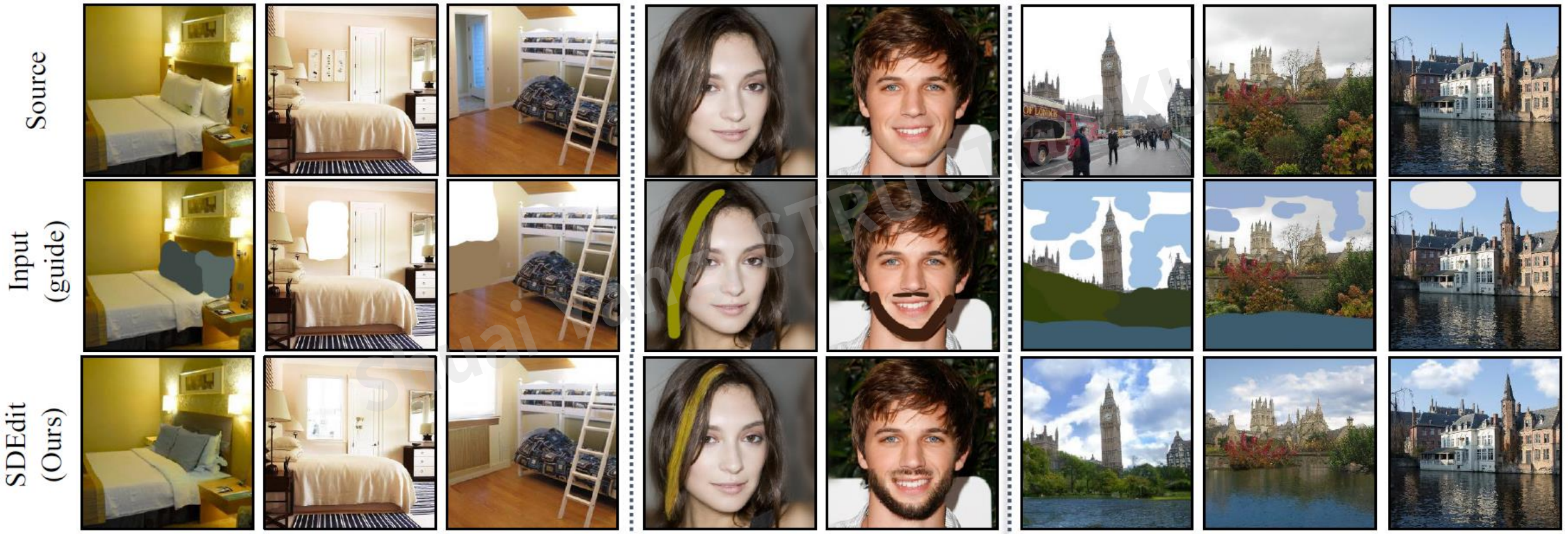


Source domain —add noise→ Noisy domain —remove noise→ Target domain

SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations. ICLR'22

Pre-trained models as prior

2 SDEdit



SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations. ICLR'22

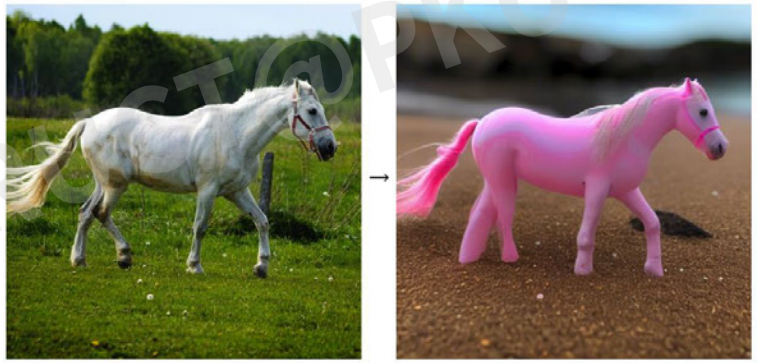
Image editing

1

Prompt2Prompt

ICLR 23

2



Plug-and-Play

CVPR 23

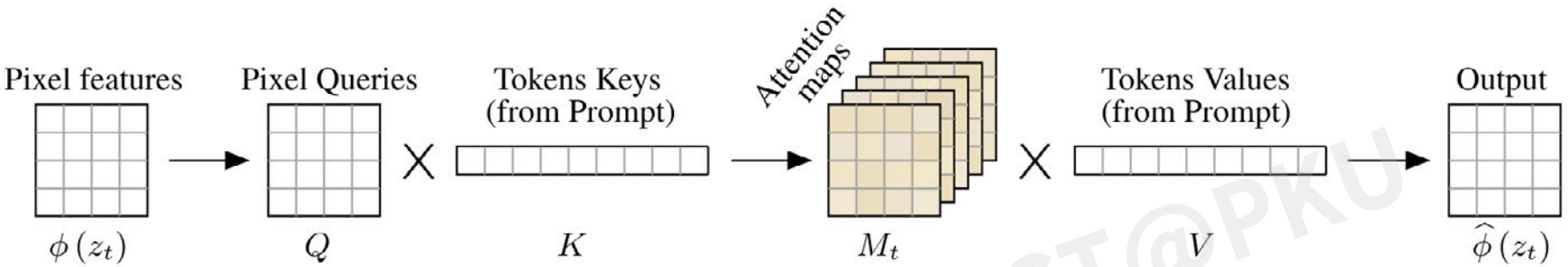
Image editing

1 Prompt-to-Prompt

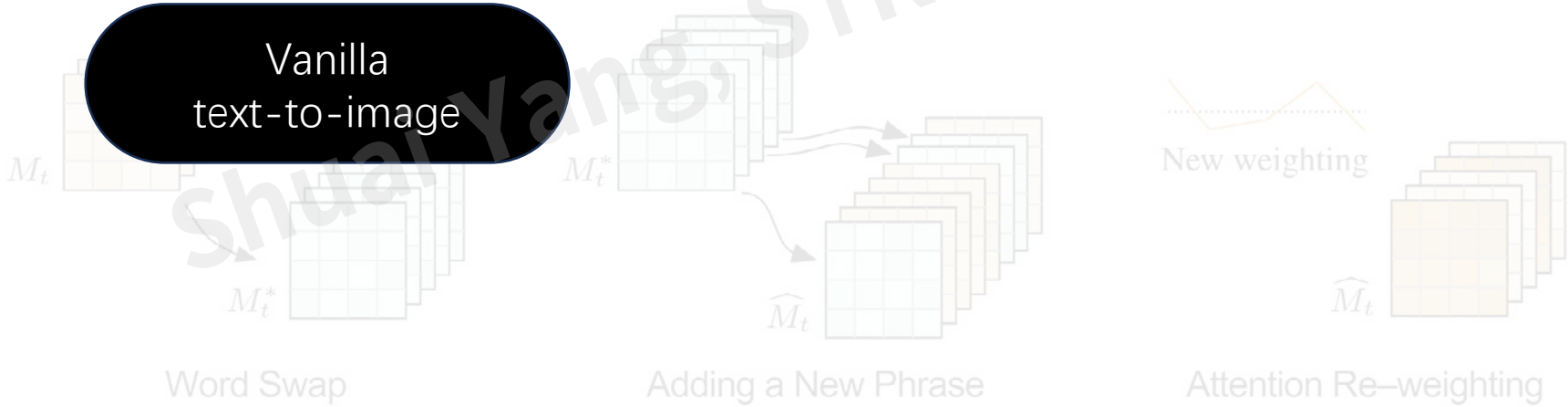
Prompt-to-Prompt Image Editing with Cross-Attention Control. ICLR'23

Image editing

1 Prompt-to-Prompt



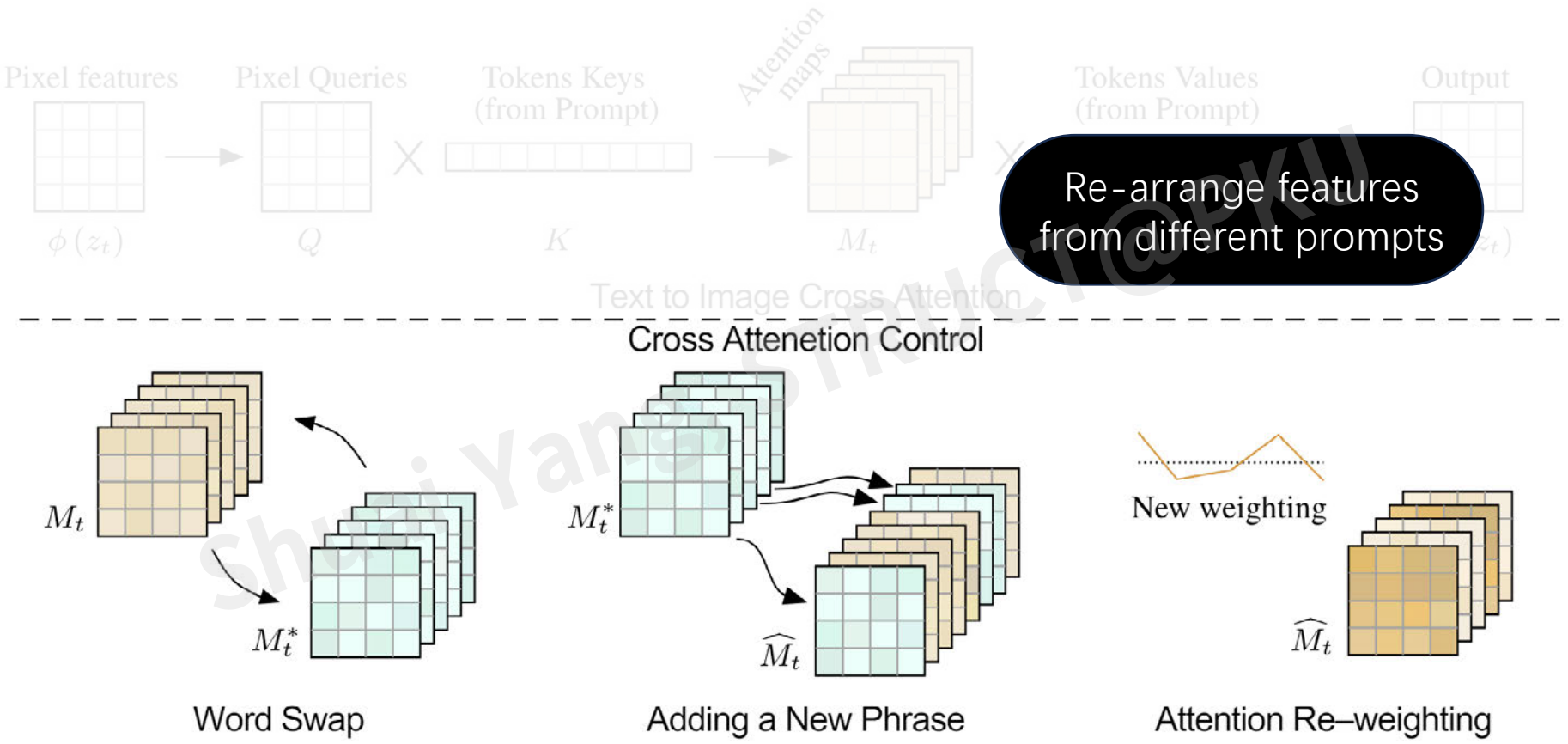
Vanilla text-to-image



Prompt-to-Prompt Image Editing with Cross-Attention Control. ICLR'23

Image editing

1 Prompt-to-Prompt



Re-arrange features from different prompts

Prompt-to-Prompt Image Editing with Cross-Attention Control. ICLR'23

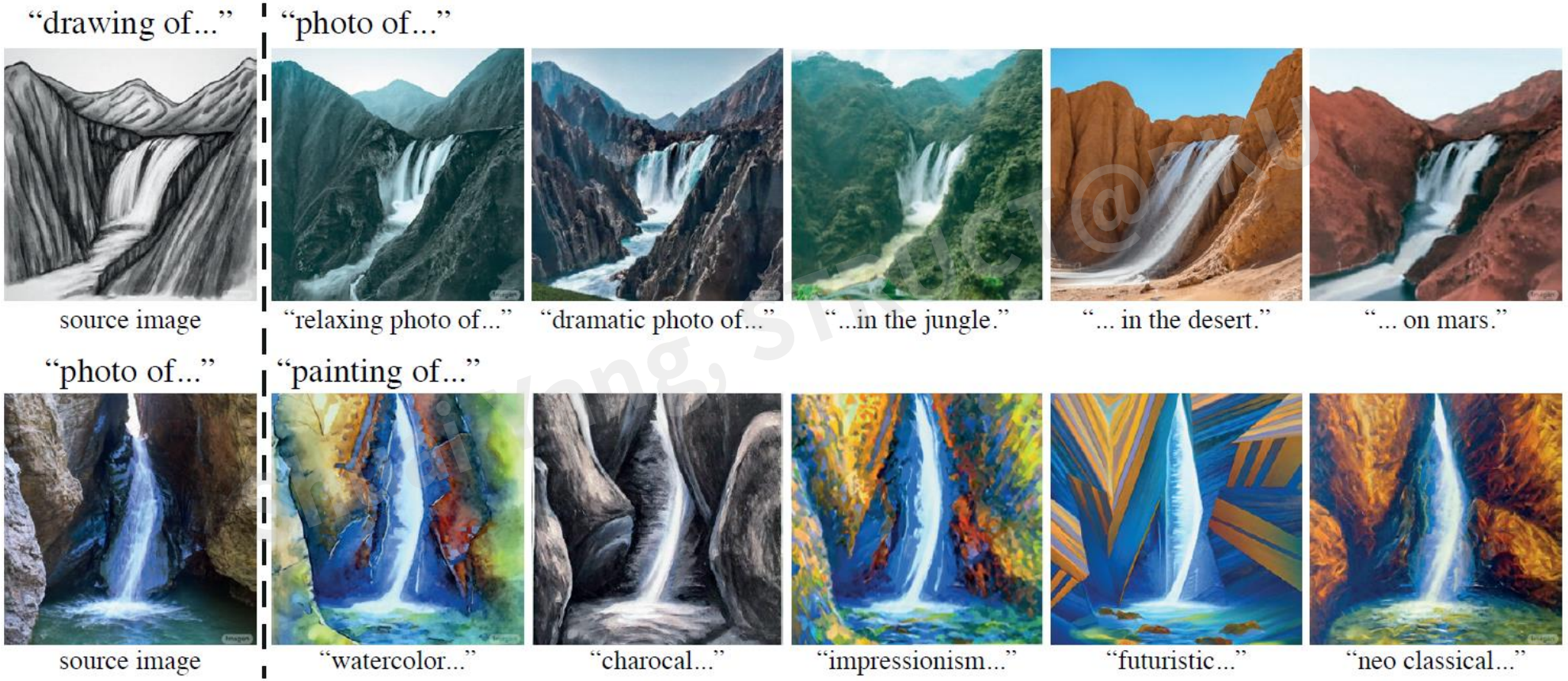
Image editing

1 Prompt-to-Prompt

Prompt-to-Prompt Image Editing with Cross-Attention Control. ICLR'23

Image editing

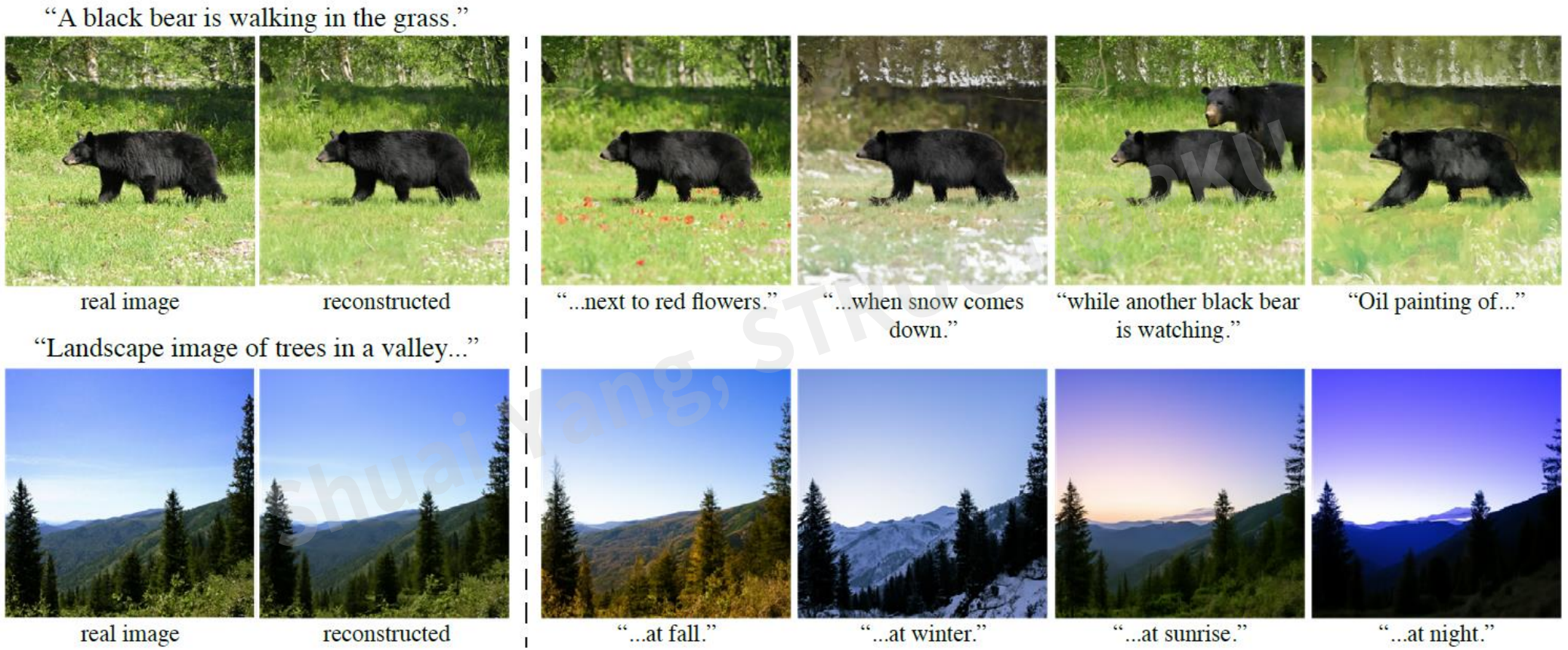
1 Prompt-to-Prompt



Prompt-to-Prompt Image Editing with Cross-Attention Control. ICLR'23

Image editing

1 Prompt-to-Prompt

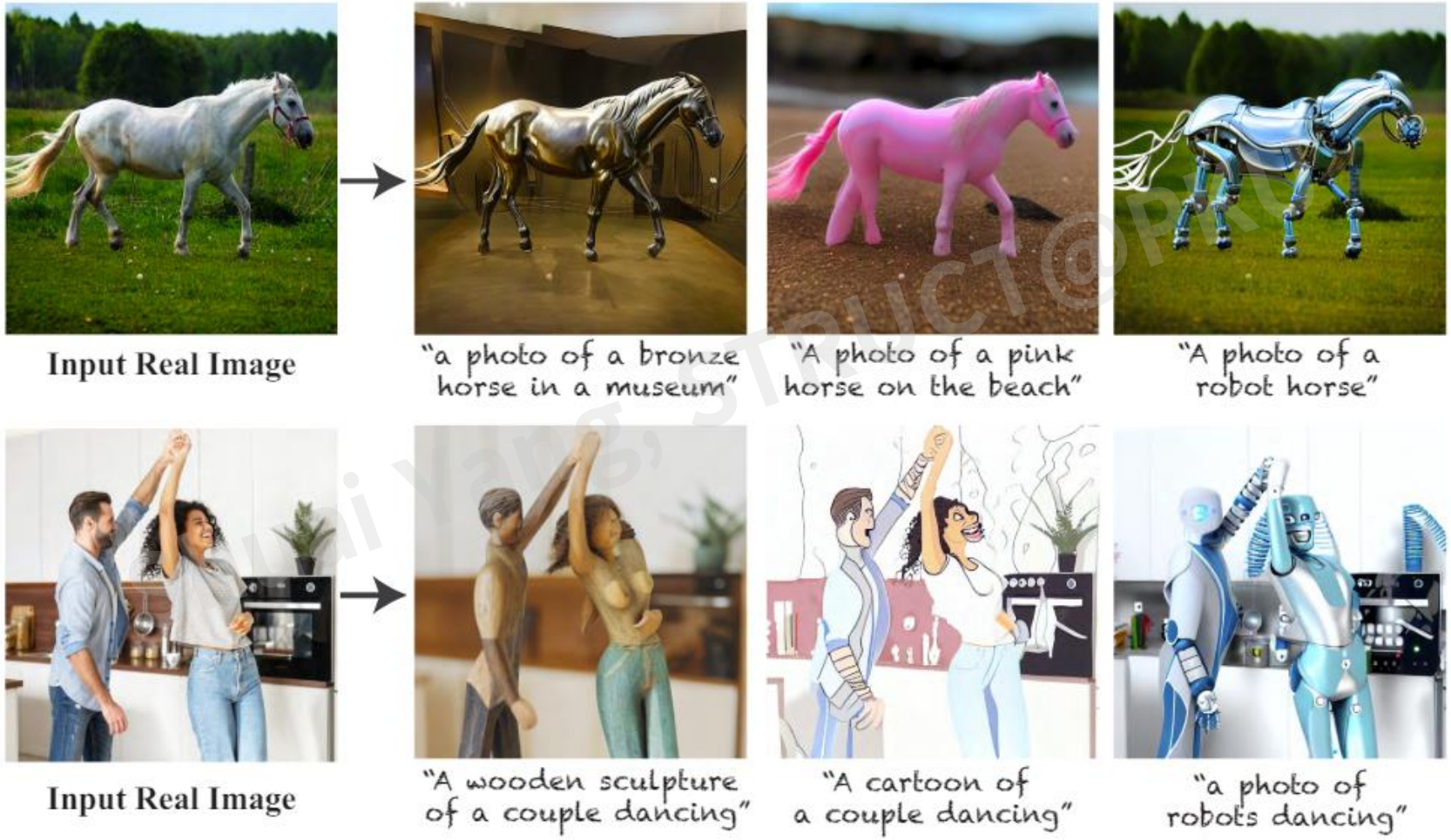


Real-world image editing

Prompt-to-Prompt Image Editing with Cross-Attention Control. ICLR'23

Image editing

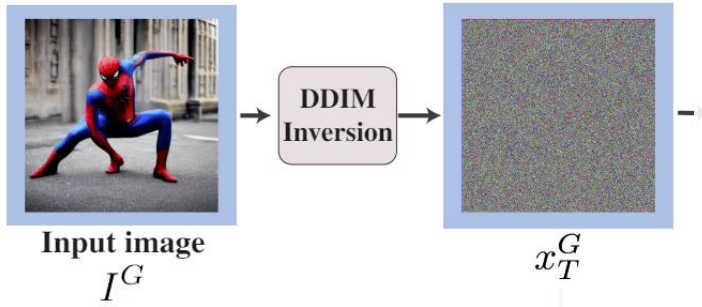
2 Plug-and-Play



Plug-and-Play Diffusion Features for Text-Driven Image-to-Image Translation. CVPR'23

Image editing

2 Plug-and-Play

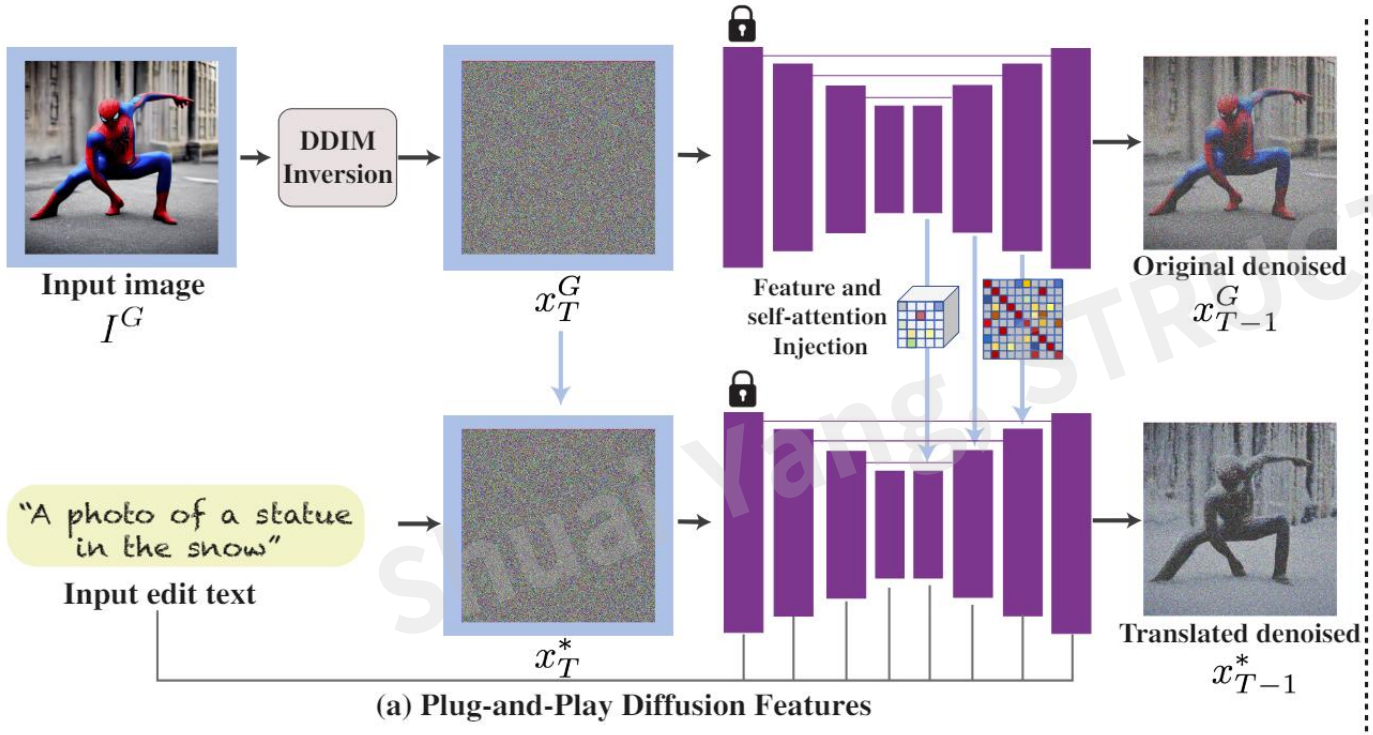


Diffusion inversion via DDIM

Plug-and-Play Diffusion Features for Text-Driven Image-to-Image Translation. CVPR'23

Image editing

2 Plug-and-Play



(a) Plug-and-Play Diffusion Features

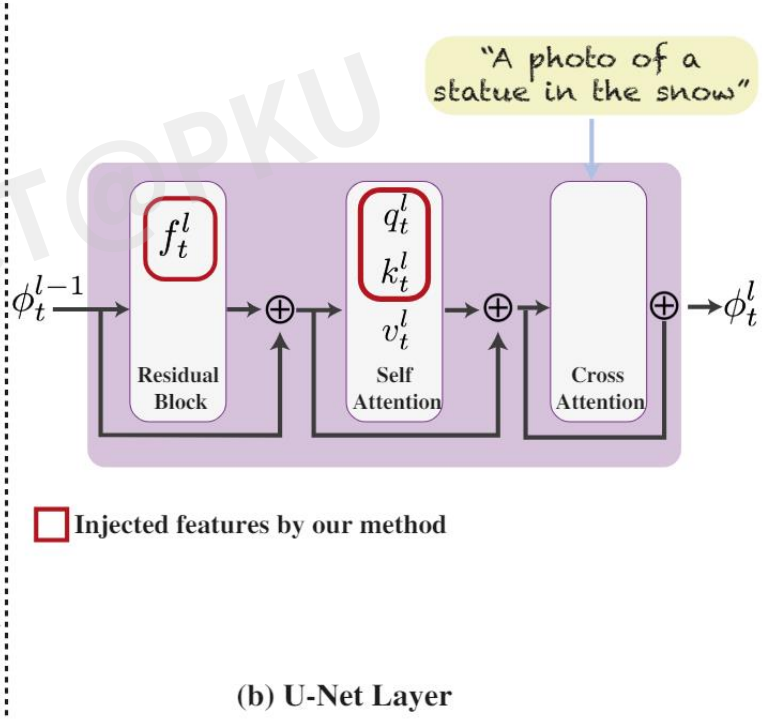
Inject previous features to new generation process

Plug-and-Play Diffusion Features for Text-Driven Image-to-Image Translation. CVPR'23

Image editing

2 Plug-and-Play

Feature injection in detail

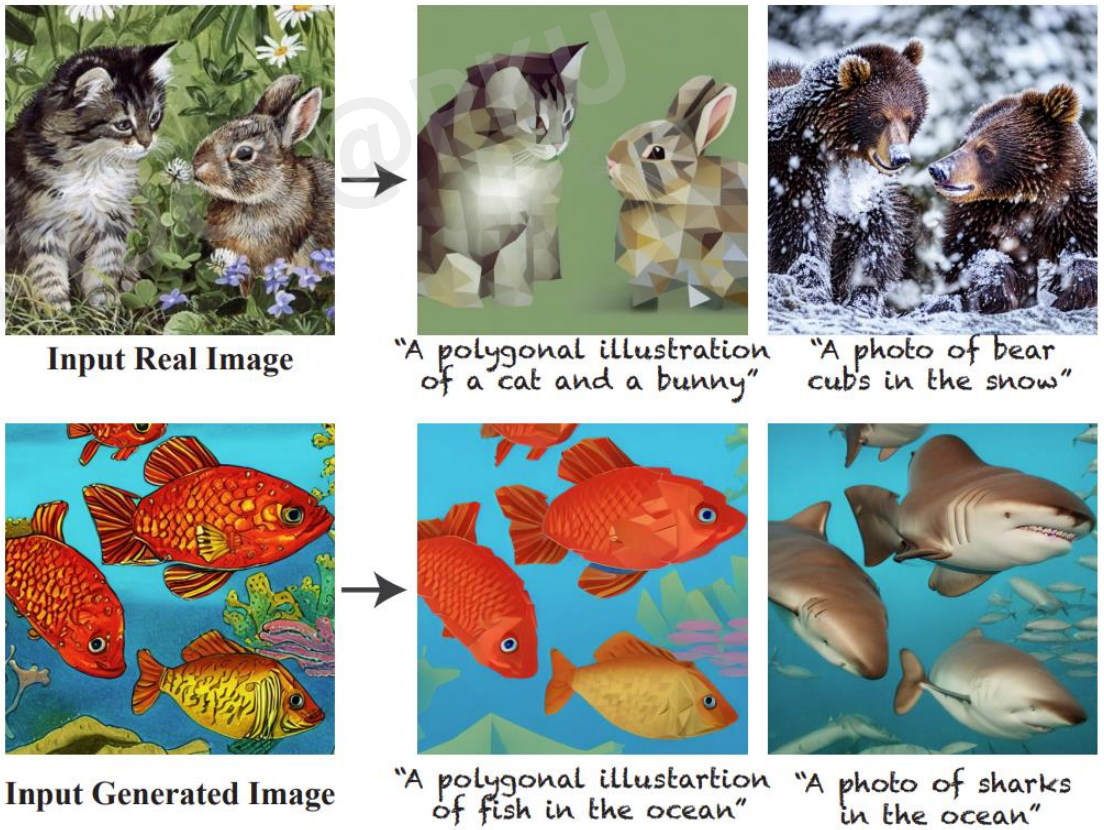


(b) U-Net Layer

Plug-and-Play Diffusion Features for Text-Driven Image-to-Image Translation. CVPR'23

Image editing

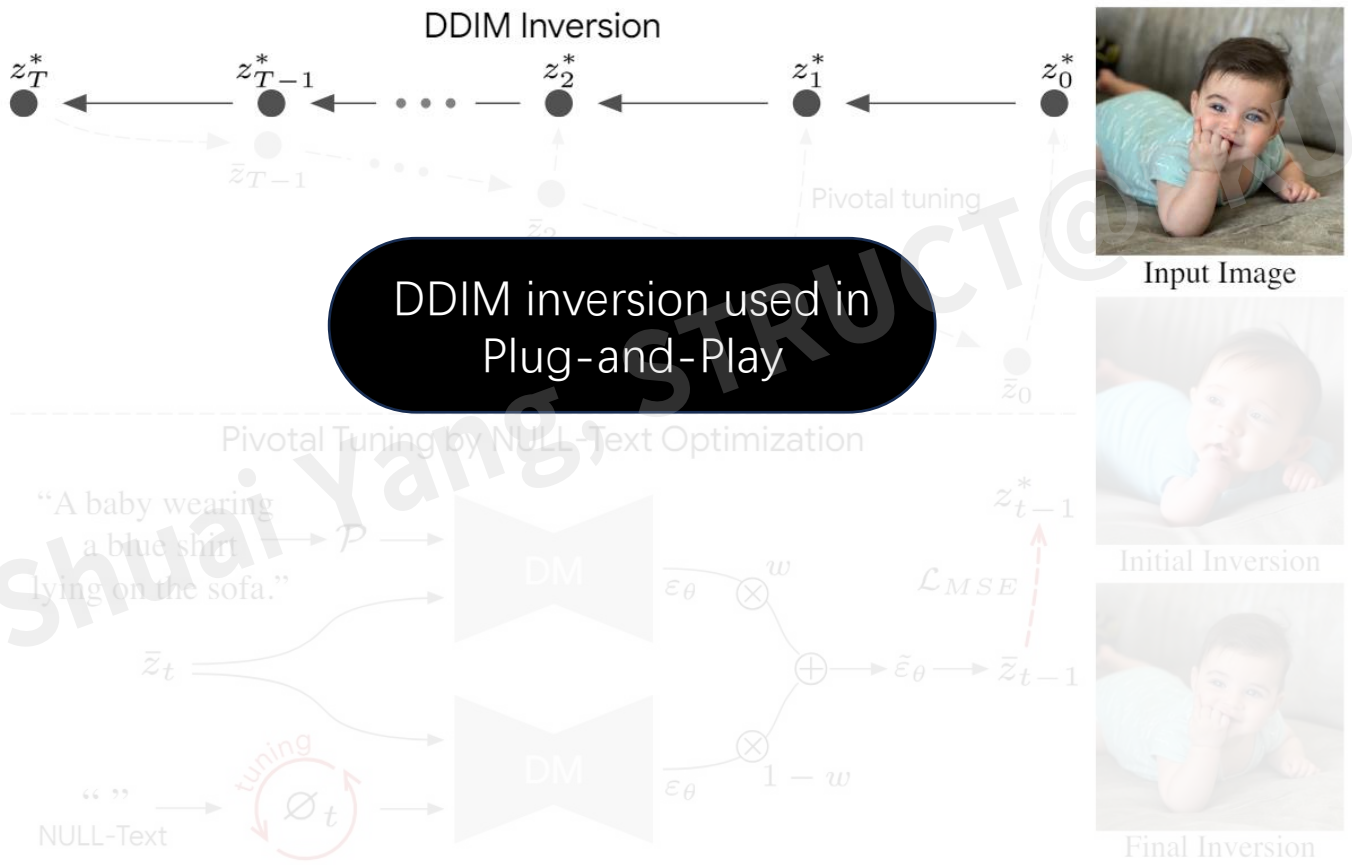
2 Plug-and-Play



Plug-and-Play Diffusion Features for Text-Driven Image-to-Image Translation. CVPR'23

Better inversion for editing

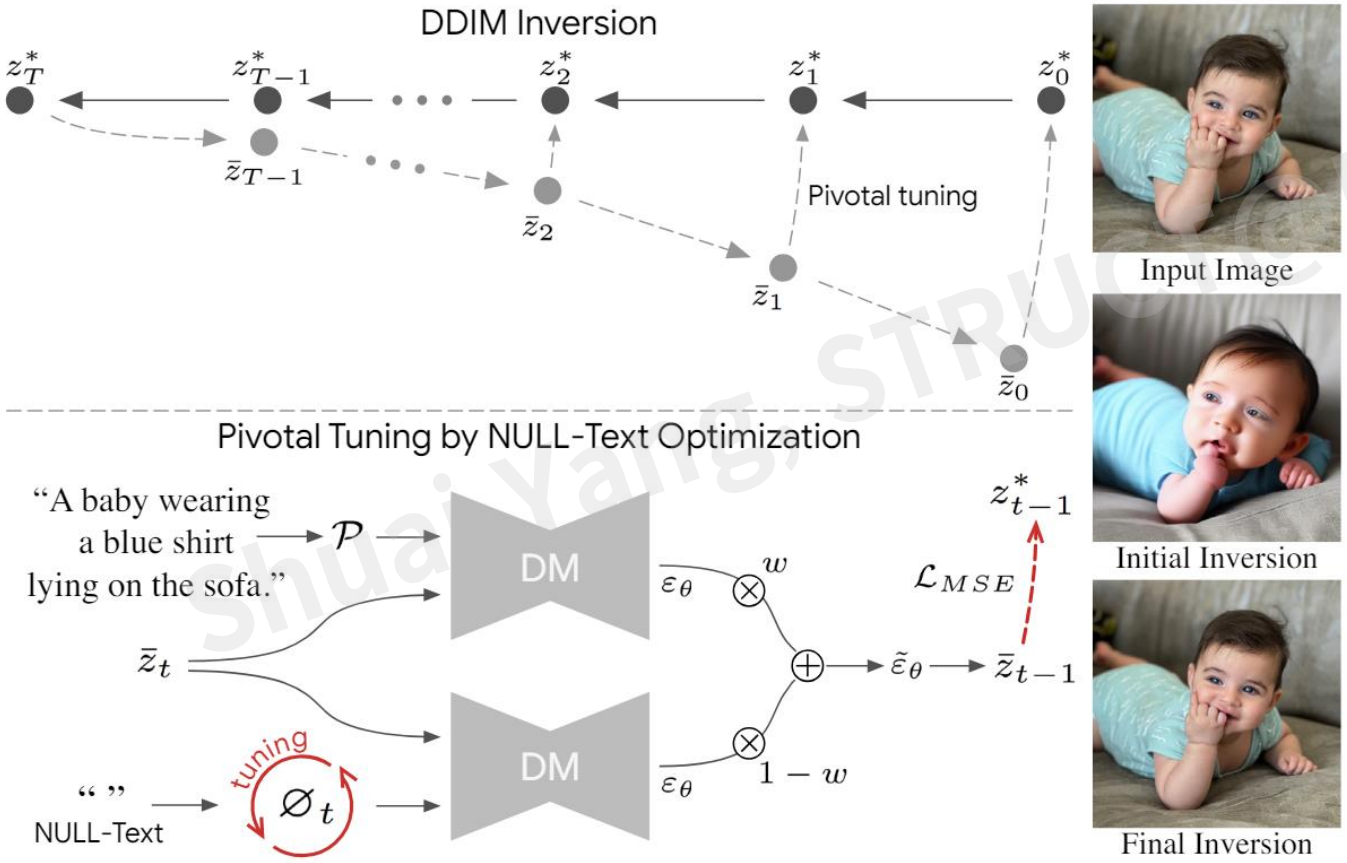
Null-text inversion



Null-Text Inversion for Editing Real Images Using Guided Diffusion Models. CVPR'23

Better inversion for editing

Null-text inversion



Finetuning "Null-Text" for better inversion

Null-Text Inversion for Editing Real Images Using Guided Diffusion Models. CVPR'23

Better inversion for editing

Null-text inversion

Input caption: "A baby wearing a blue shirt lying on the sofa."

Null-Text Inversion + Prompt to Prompt

Null-Text Inversion for Editing Real Images Using Guided Diffusion Models. CVPR'23

Better inversion for editing

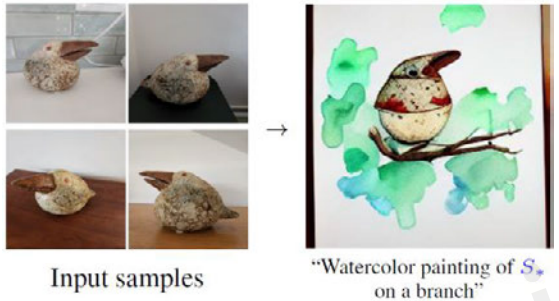
Null-text inversion

Null-Text Inversion + Prompt to Prompt

Null-Text Inversion for Editing Real Images Using Guided Diffusion Models. CVPR'23

Model customization

1



Textual Inversion

ICLR 23

2

DreamBooth

CVPR 23

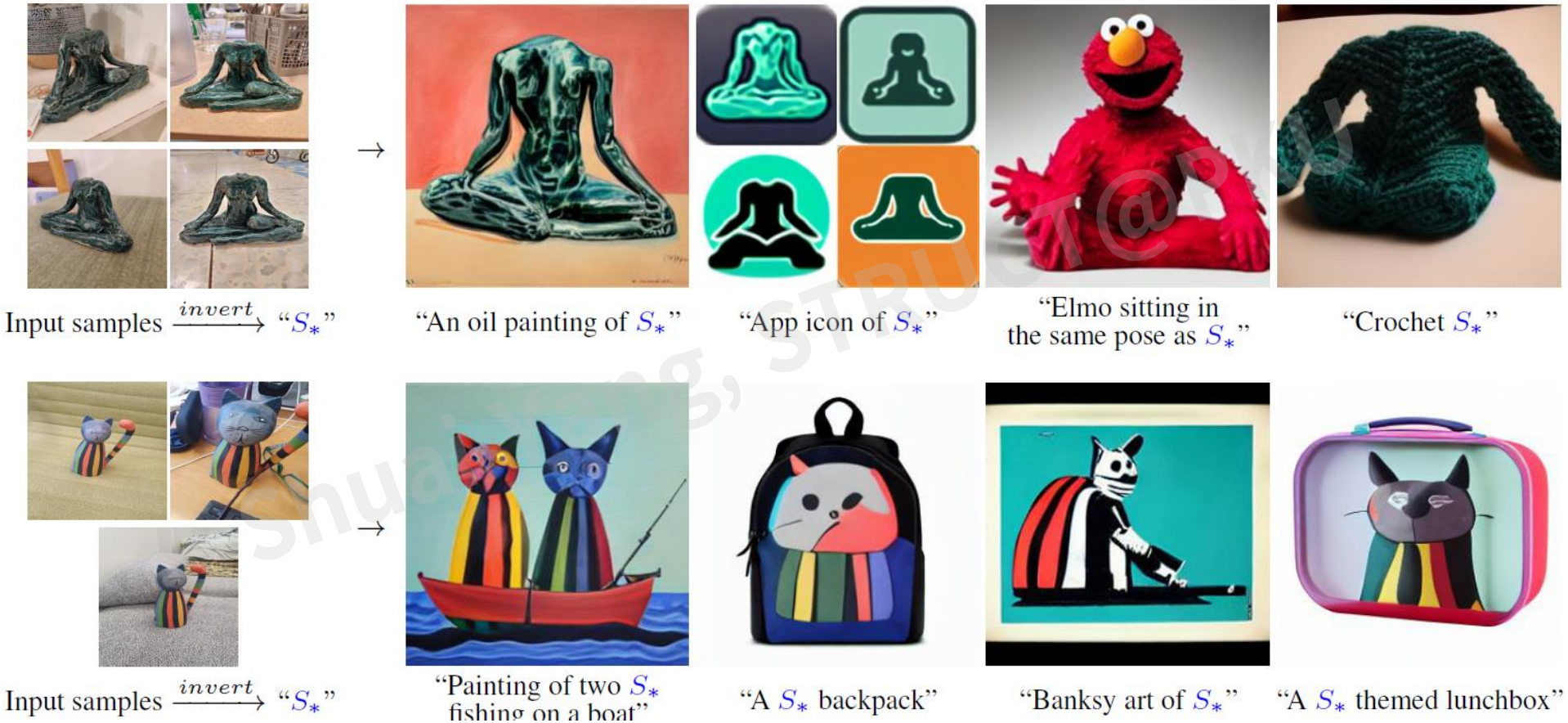
3

LoRA for image

GitHub Repo

Model customization

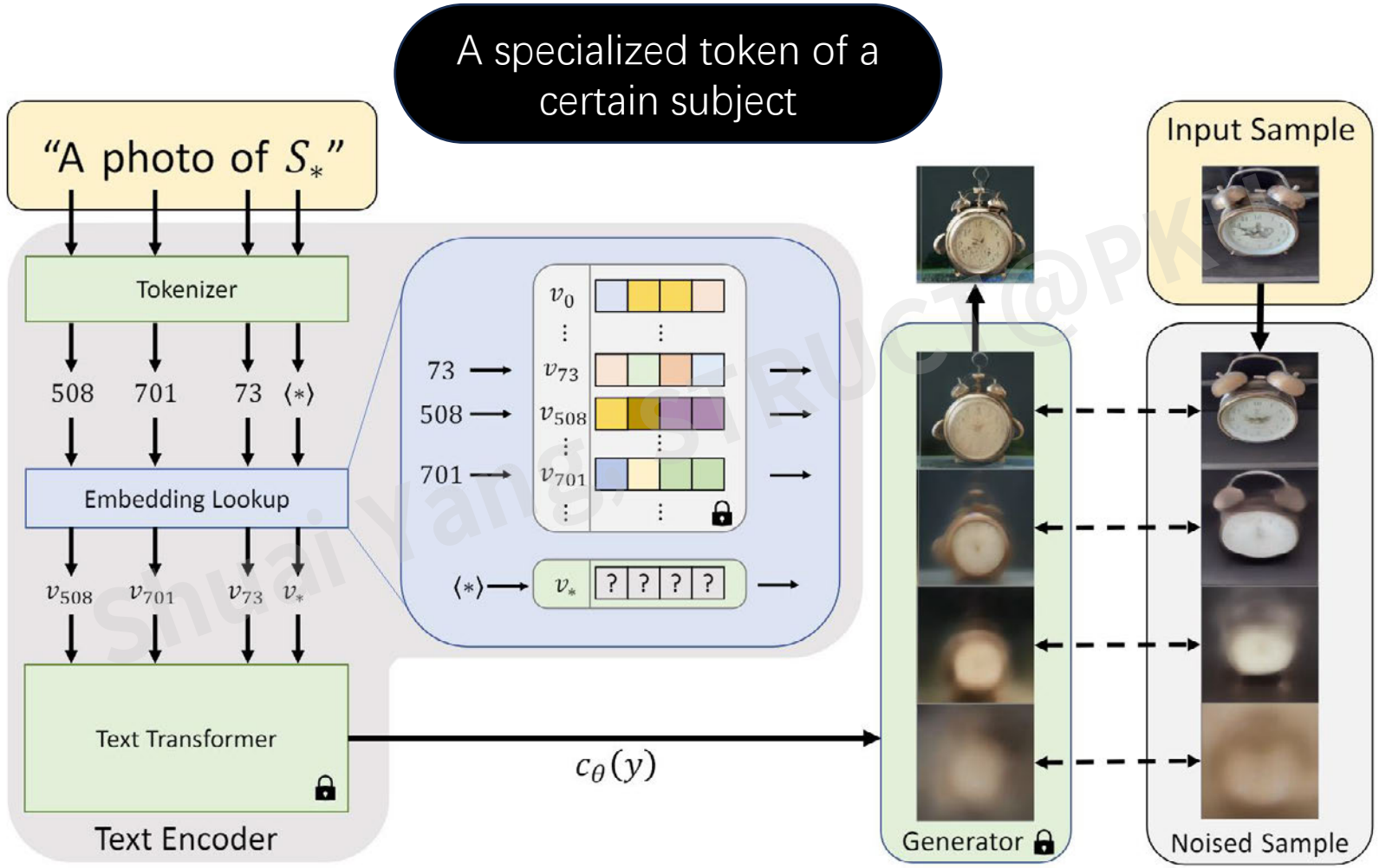
1 Textual Inversion



An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. ICLR'23

Model customization

1 Textual Inversion



An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. ICLR'23

Model customization

1 Textual Inversion

An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. ICLR'23

Model customization

1 Textual Inversion

An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. ICLR'23

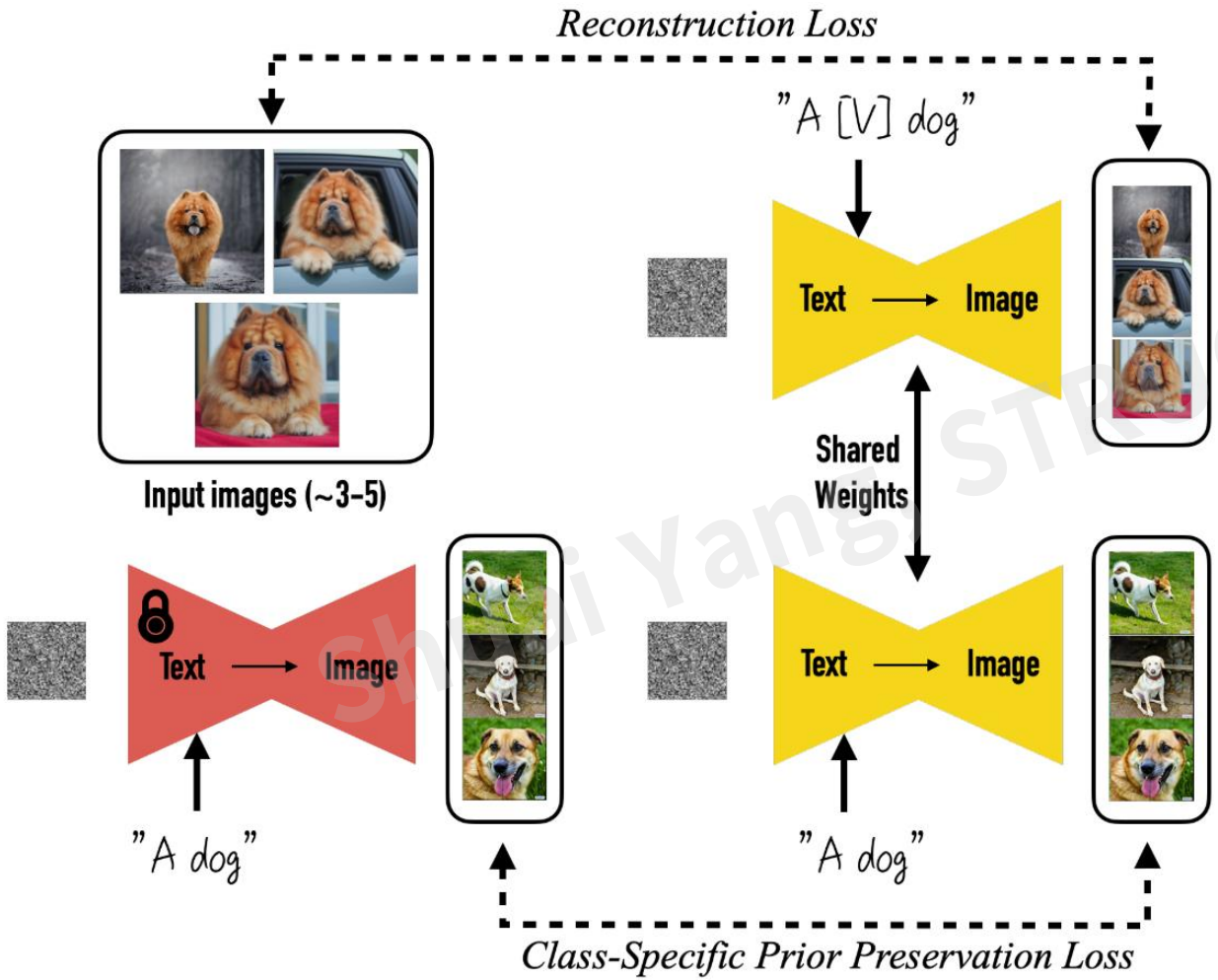
Model customization

2 DreamBooth

DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. CVPR'23

Model customization

2 DreamBooth



Specialized token [V]

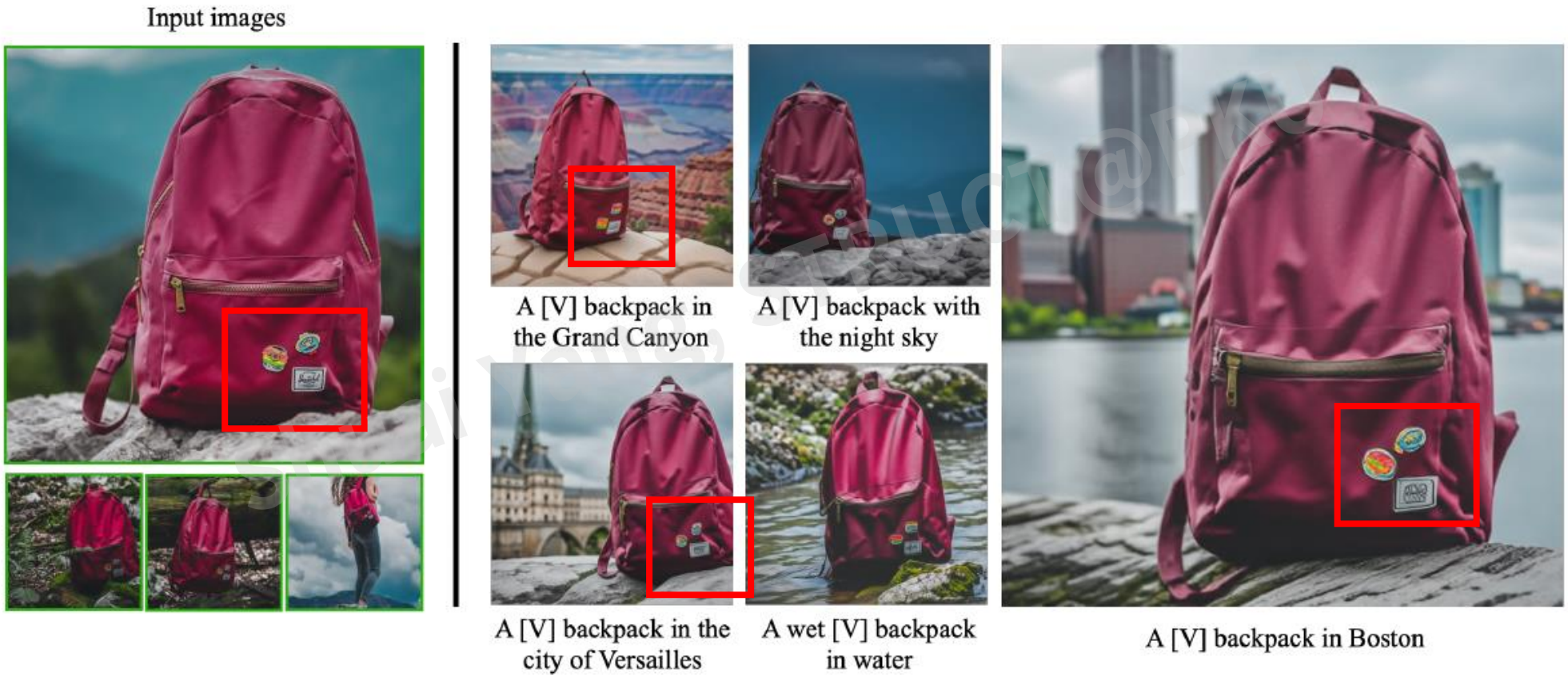
+

Finetune entire model

DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. CVPR'23

Model customization

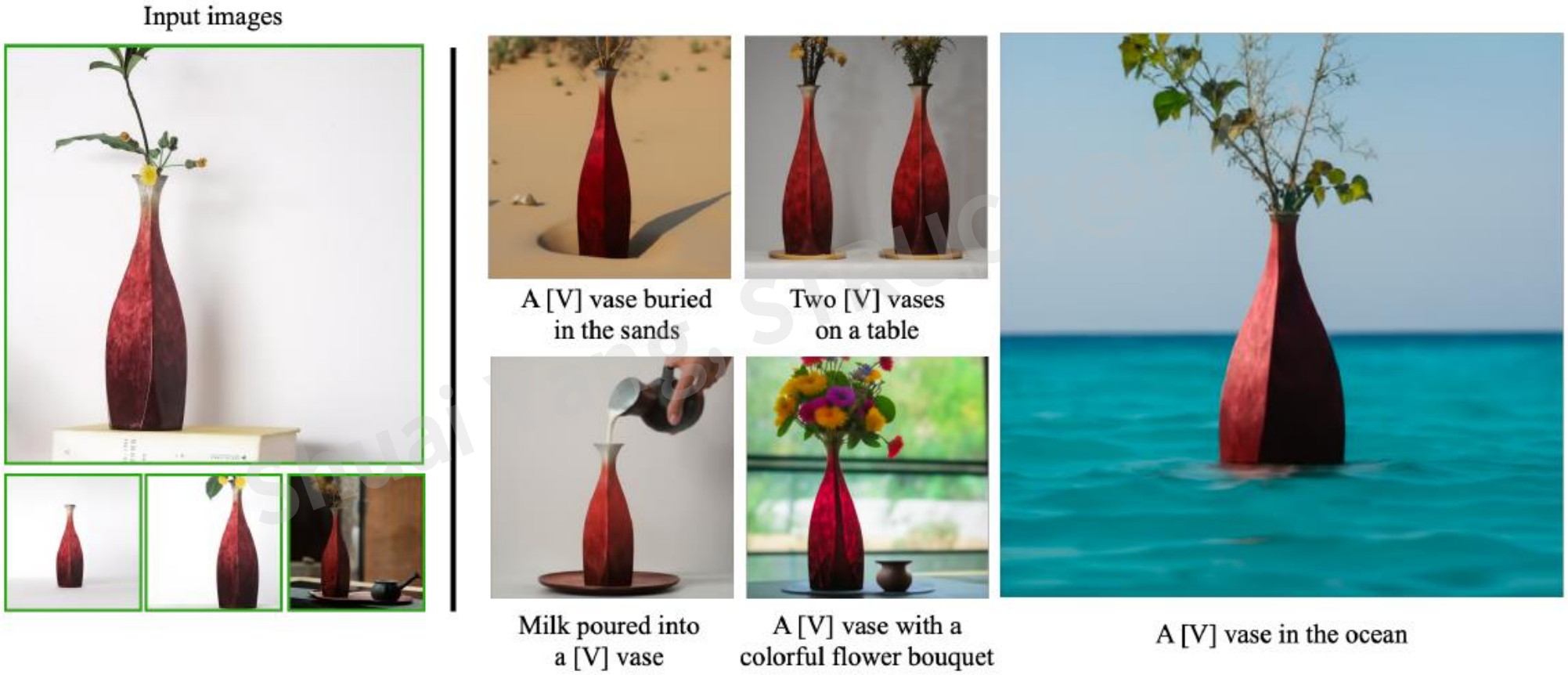
2 DreamBooth



DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. CVPR'23

Model customization

2 DreamBooth



DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. CVPR'23

Model customization

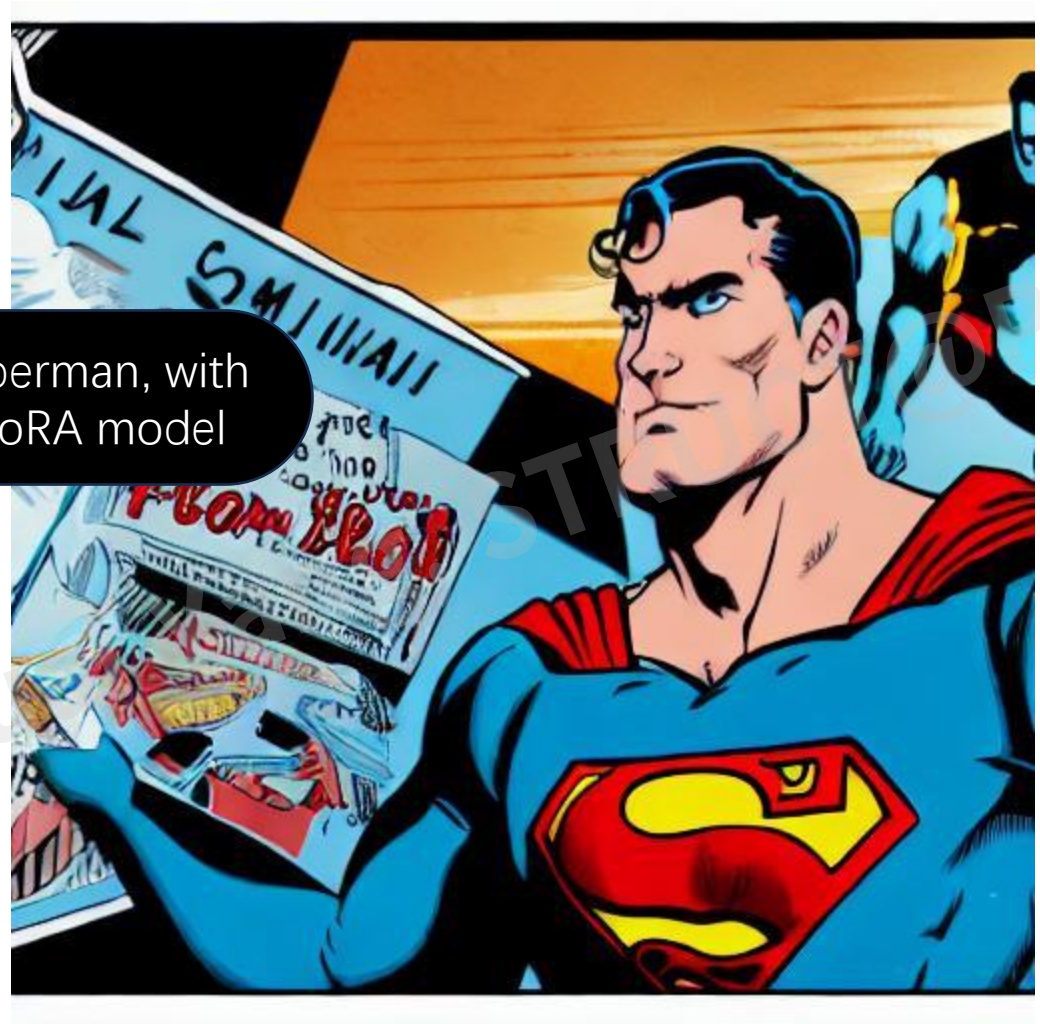
2 DreamBooth

DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. CVPR'23

Model customization

3 LoRA

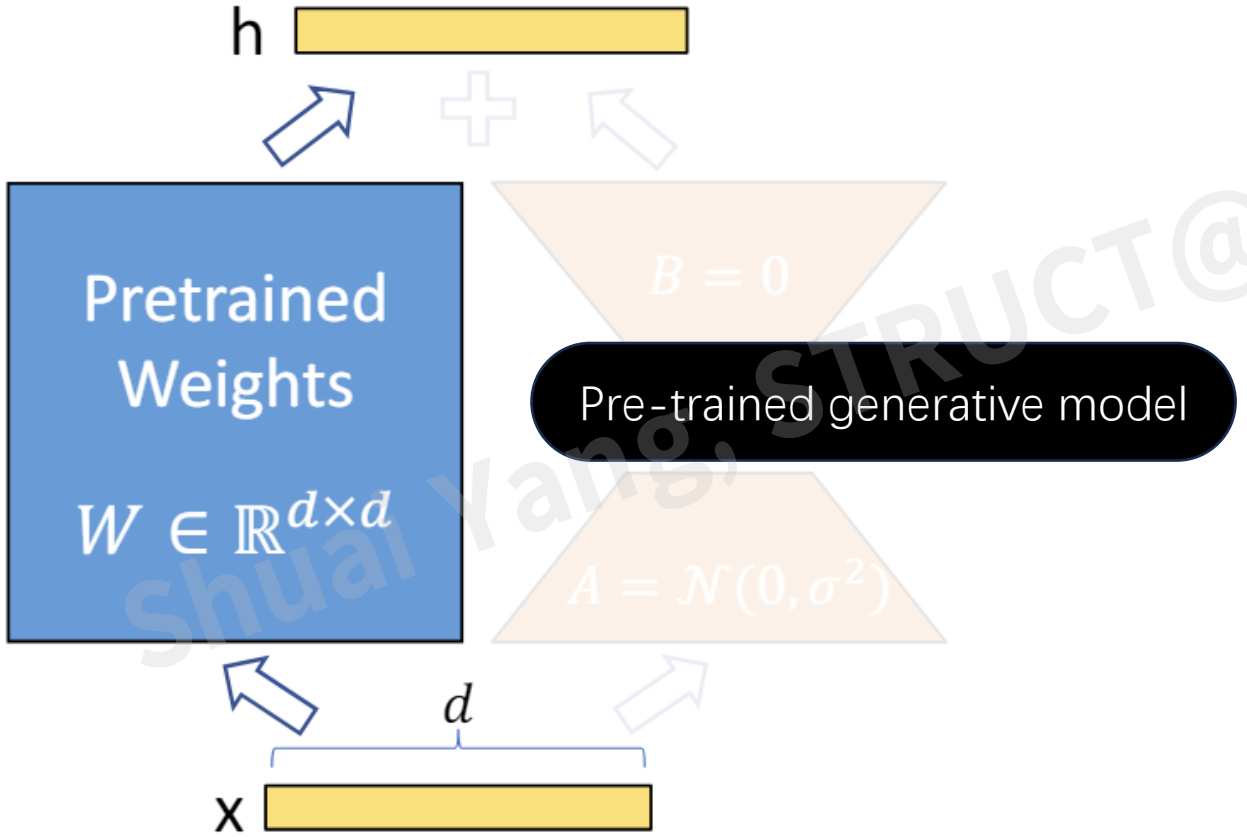
An image of superman, with pop-art style LoRA model



Low-rank Adaptation (LoRA) for Fast Text-to-Image Diffusion Fine-tuning. GitHub repo
LoRA: Low-Rank Adaptation of Large Language Models. ICLR'22

Model customization

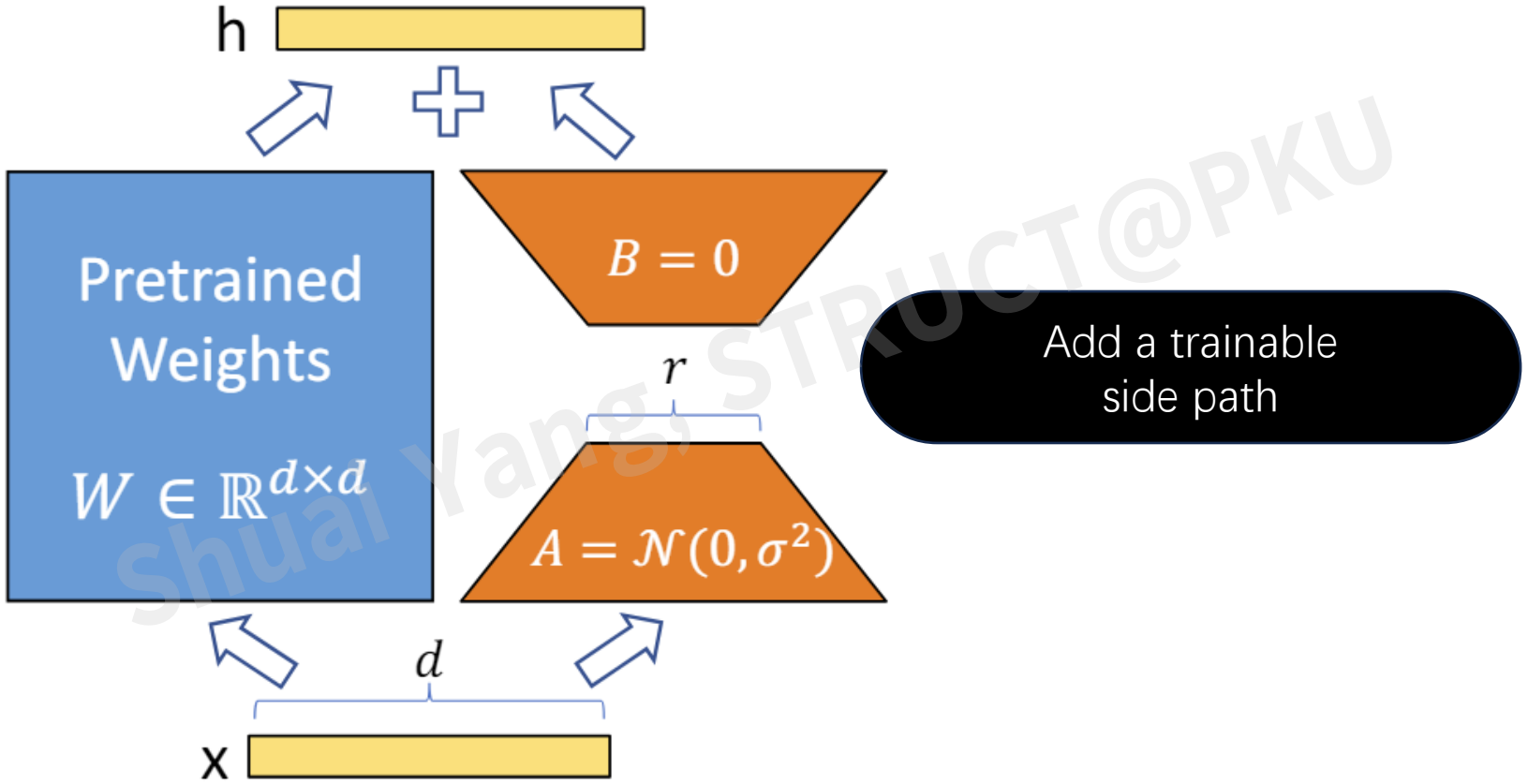
3 LoRA



Low-rank Adaptation (LoRA) for Fast Text-to-Image Diffusion Fine-tuning. GitHub repo
LoRA: Low-Rank Adaptation of Large Language Models. ICLR'22

Model customization

3 LoRA



Low-rank Adaptation (LoRA) for Fast Text-to-Image Diffusion Fine-tuning. GitHub repo
LoRA: Low-Rank Adaptation of Large Language Models. ICLR'22

Model customization

3 LoRA

An image of baby lion, with pop-art style LoRA model

Low-rank Adaptation (LoRA) for Fast Text-to-Image Diffusion Fine-tuning. GitHub repo
LoRA: Low-Rank Adaptation of Large Language Models. ICLR'22

Model customization

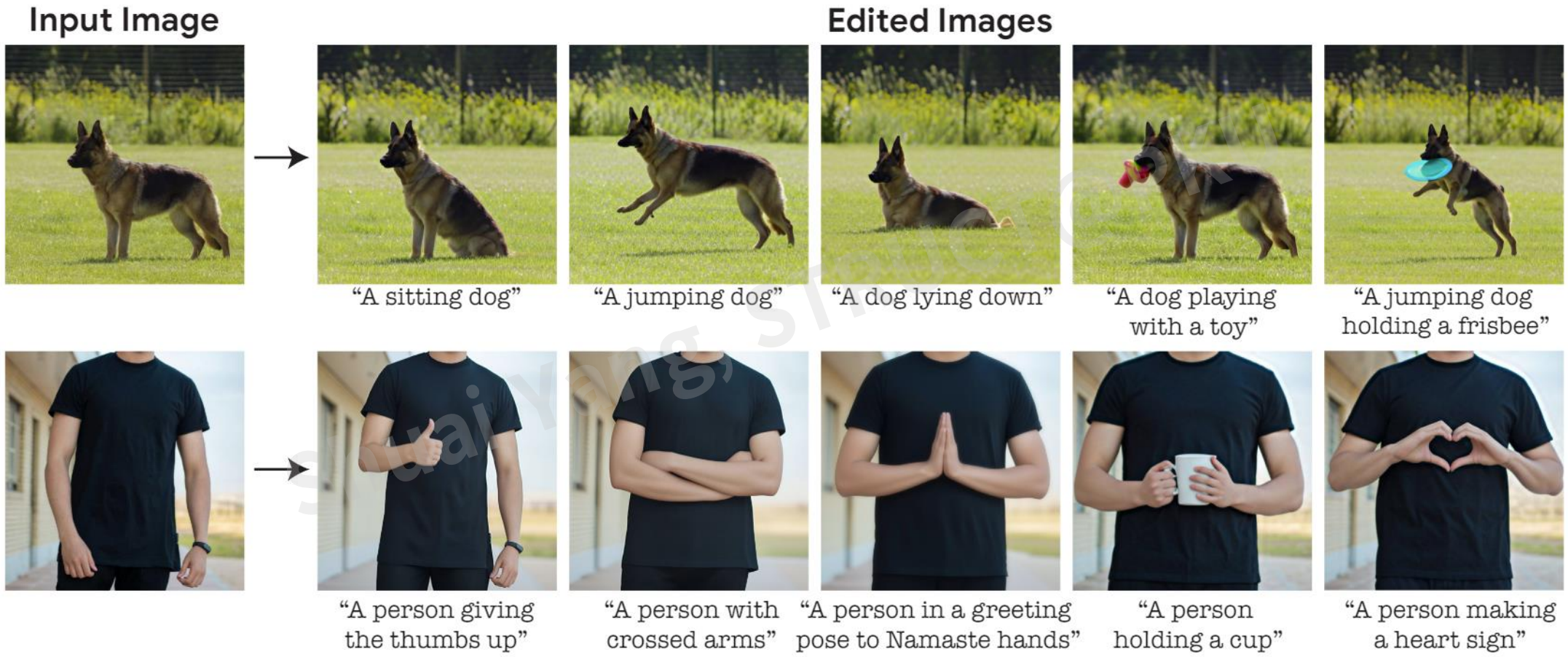
3 LoRA

Pokémon, with
Pokémon style LoRA model

Low-rank Adaptation (LoRA) for Fast Text-to-Image Diffusion Fine-tuning. GitHub repo
LoRA: Low-Rank Adaptation of Large Language Models. ICLR'22

Model customization for editing

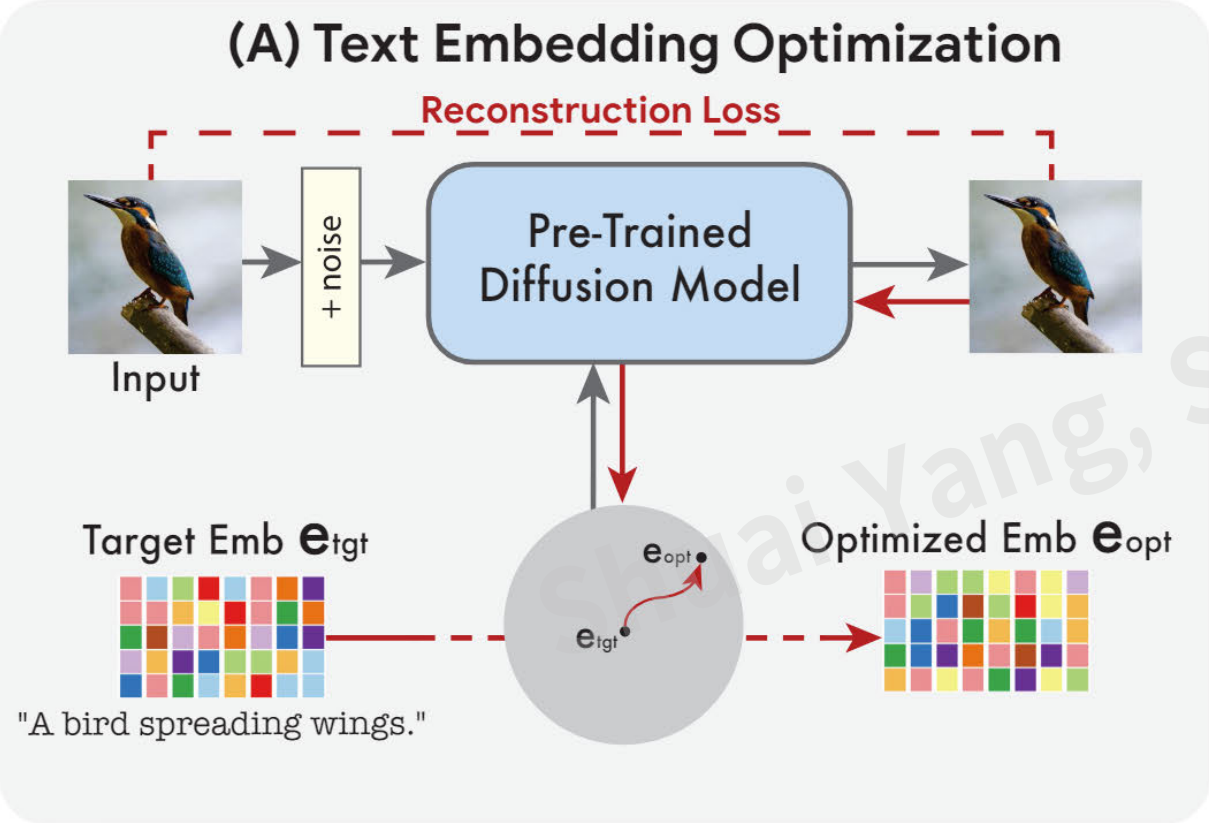
Imagic



Imagic: Text-Based Real Image Editing with Diffusion Models. CVPR'23

Model customization for editing

Imagic

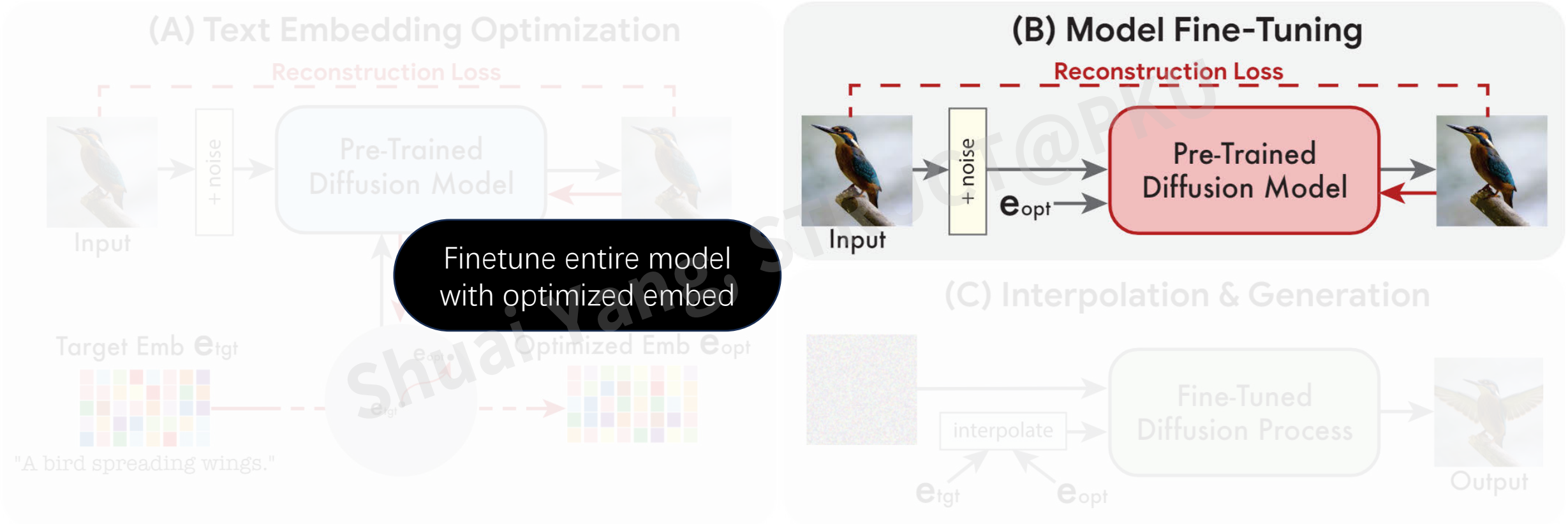


Optimize target prompt

Imagic: Text-Based Real Image Editing with Diffusion Models. CVPR'23

Model customization for editing

Imagic

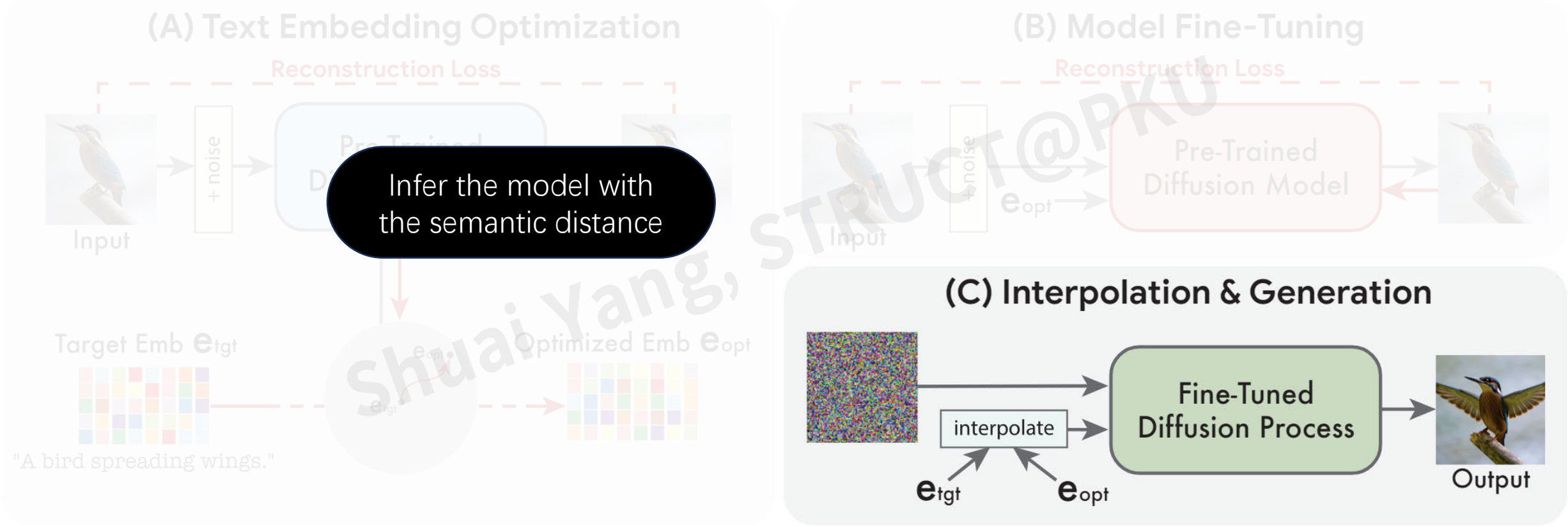


Finetune entire model with optimized embed

Imagic: Text-Based Real Image Editing with Diffusion Models. CVPR'23

Model customization for editing

Imagic



Infer the model with the semantic distance

Imagic: Text-Based Real Image Editing with Diffusion Models. CVPR'23

Model customization for editing

Imagic

Imagic: Text-Based Real Image Editing with Diffusion Models. CVPR'23

Controlling pre-trained models

1

ControlNet

ICCV 23

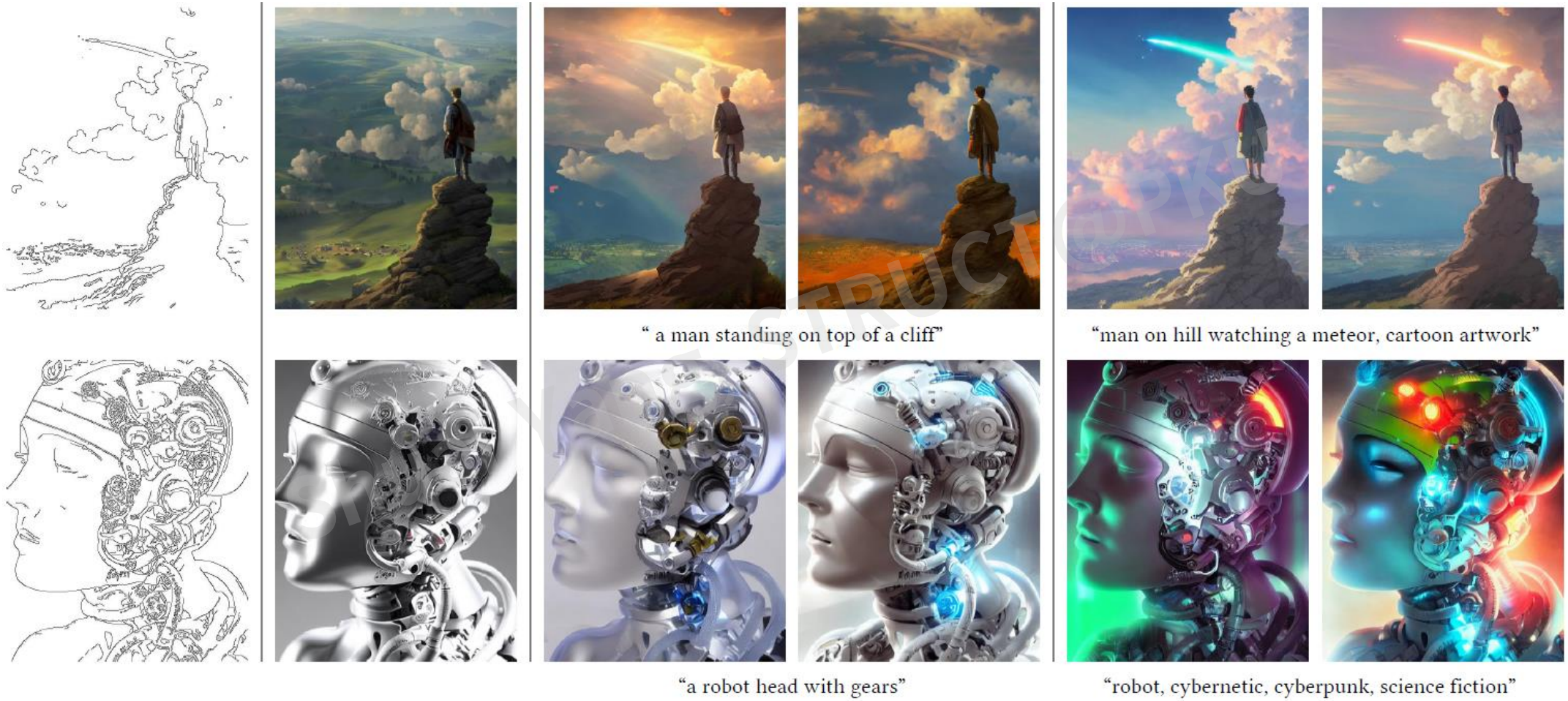
2

T2I Adapter

AAAI 24

Controlling pre-trained models

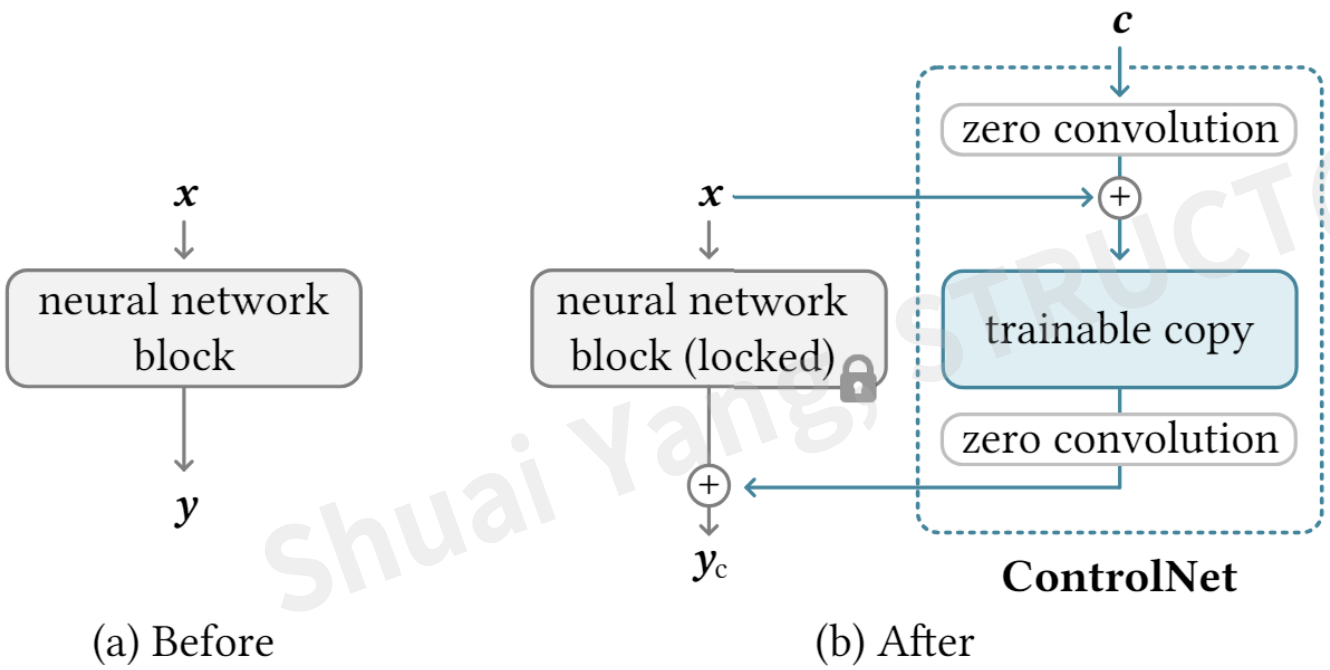
1 ControlNet



Adding Conditional Control to Text-to-Image Diffusion Models. ICCV'23

Controlling pre-trained models

1 ControlNet

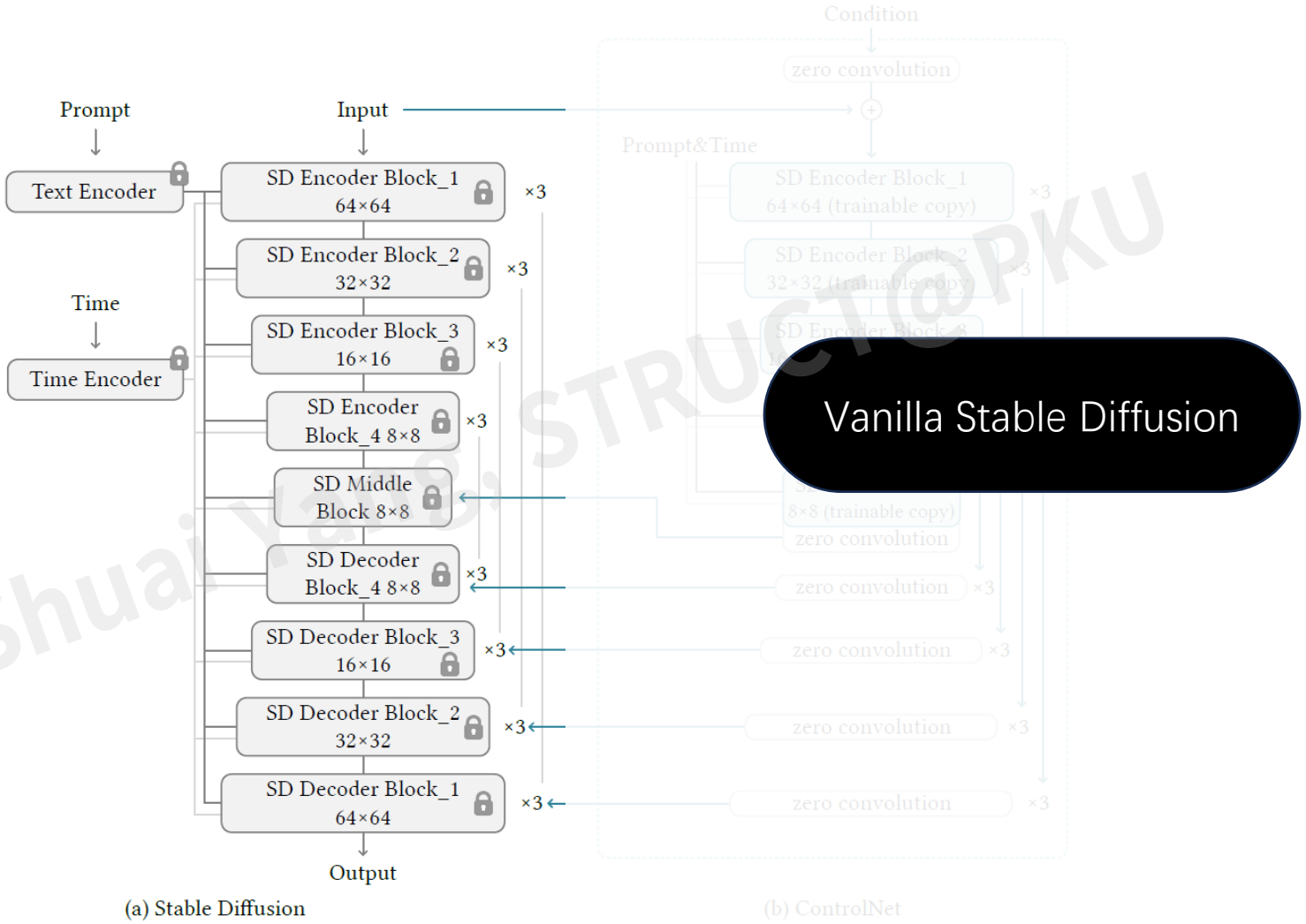


Employing zero-conv to inject guidance

Adding Conditional Control to Text-to-Image Diffusion Models. ICCV'23

Controlling pre-trained models

1 ControlNet

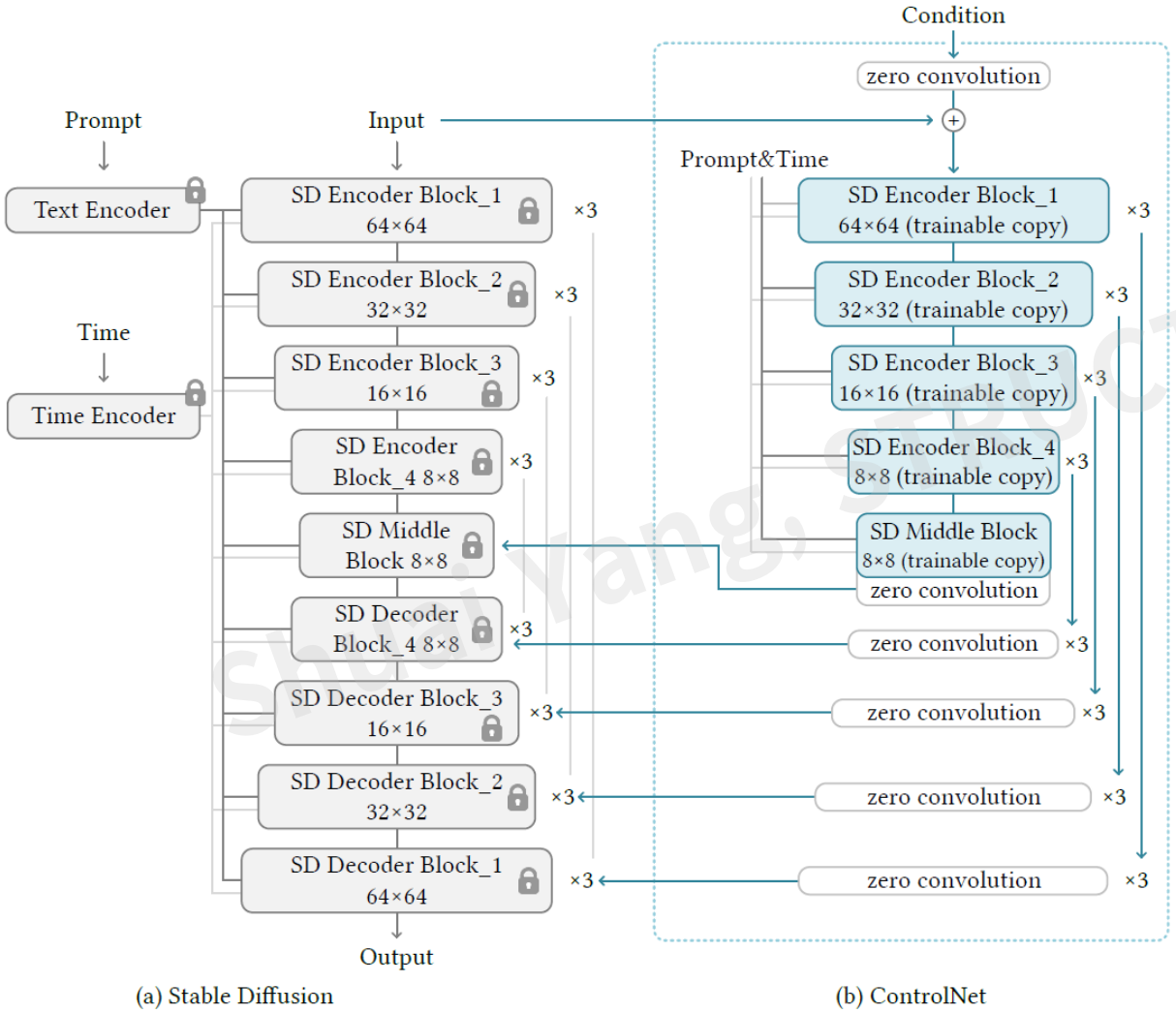


Vanilla Stable Diffusion

Adding Conditional Control to Text-to-Image Diffusion Models. ICCV'23

Controlling pre-trained models

1 ControlNet



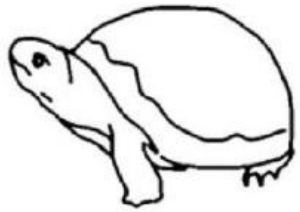
Implement ControlNet on SD

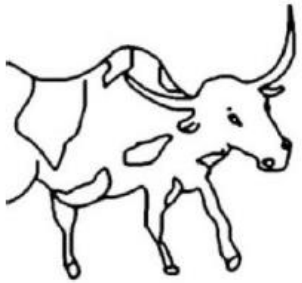
Adding Conditional Control to Text-to-Image Diffusion Models. ICCV'23

Controlling pre-trained models

1 ControlNet

Input (User Scribble)

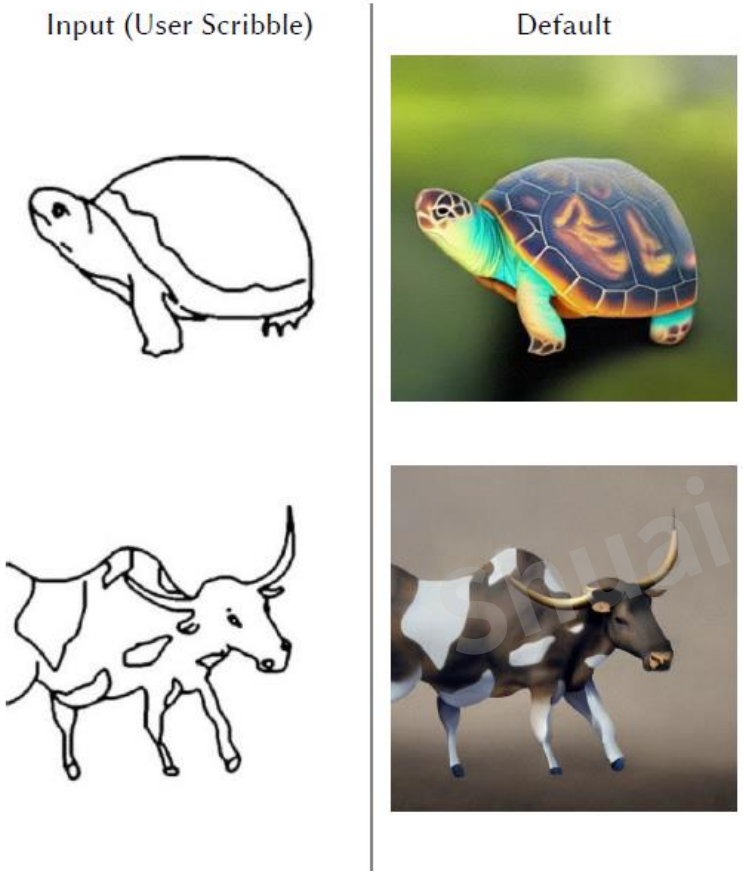




Adding Conditional Control to Text-to-Image Diffusion Models. ICCV'23

Controlling pre-trained models

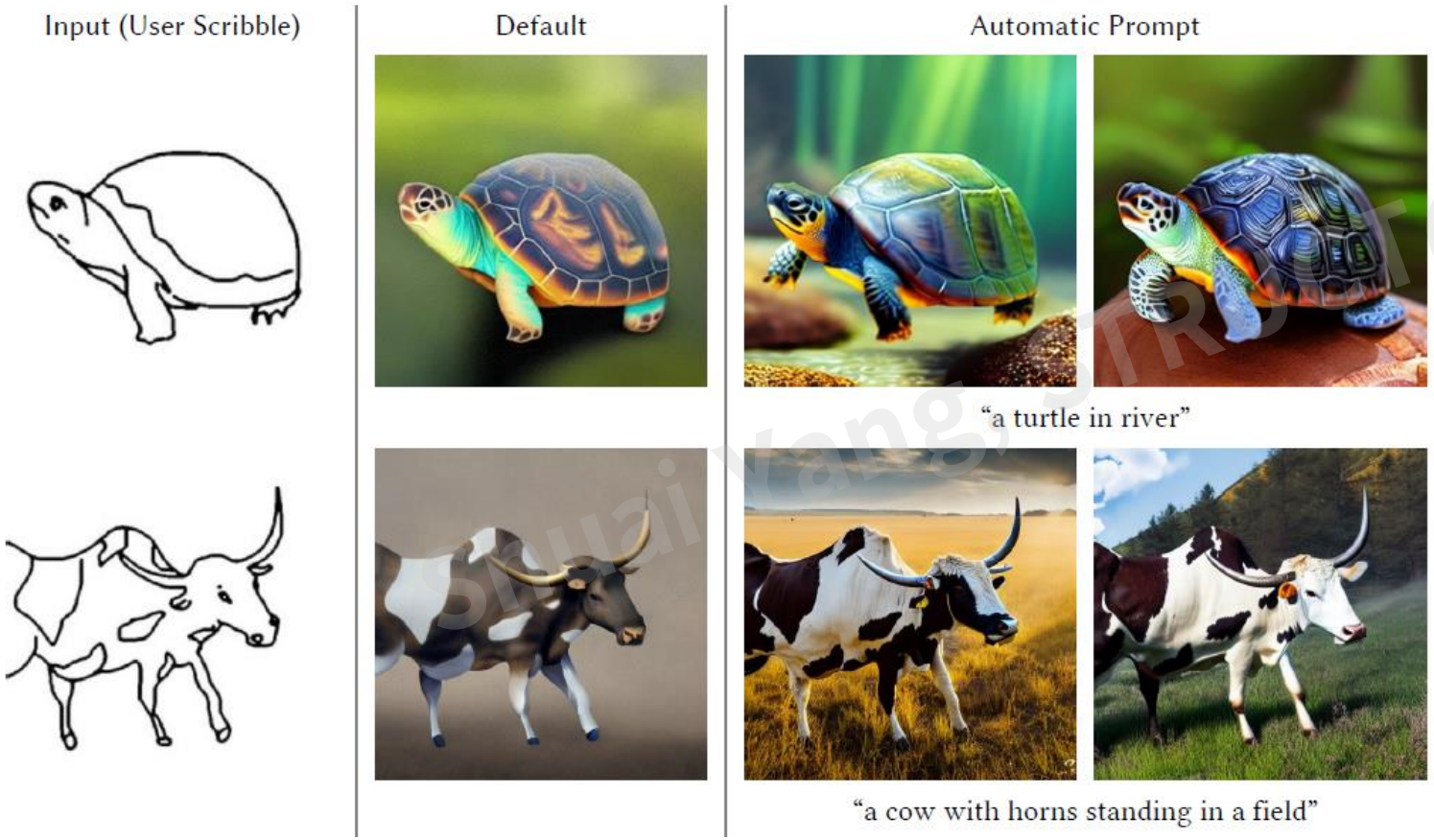
1 ControlNet



Adding Conditional Control to Text-to-Image Diffusion Models. ICCV'23

Controlling pre-trained models

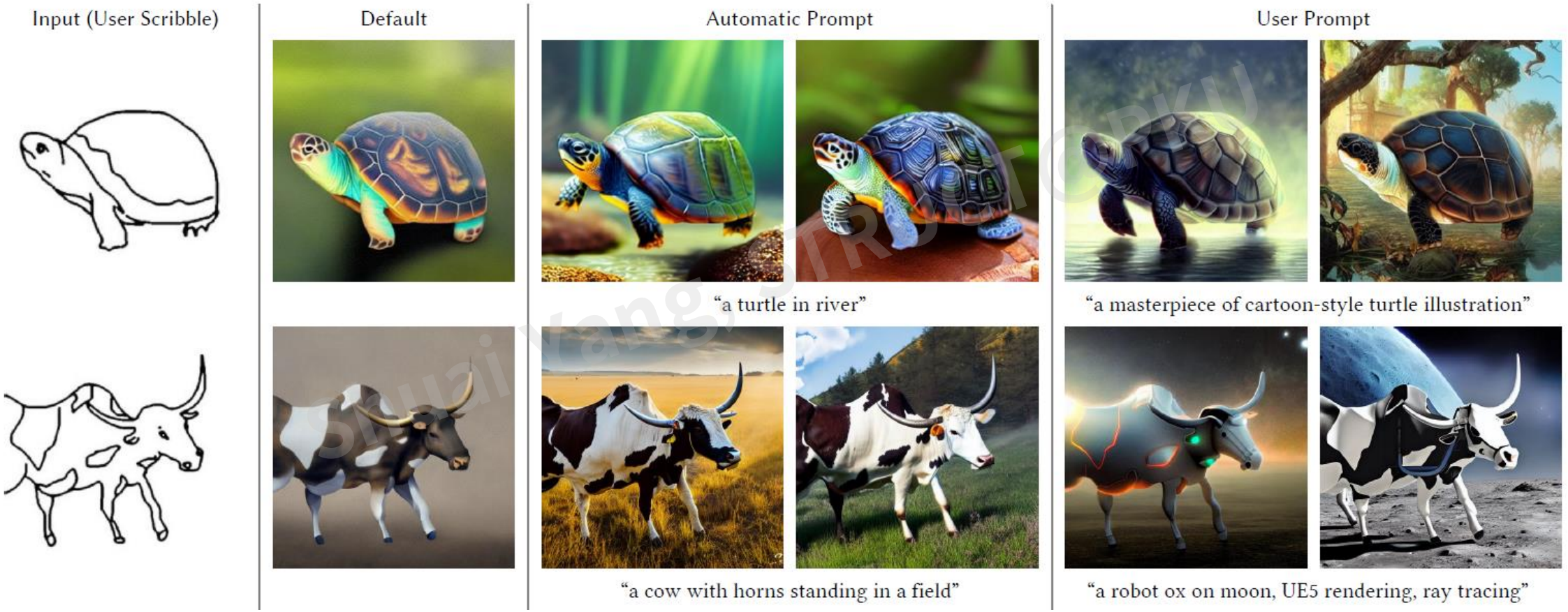
1 ControlNet



Adding Conditional Control to Text-to-Image Diffusion Models. ICCV'23

Controlling pre-trained models

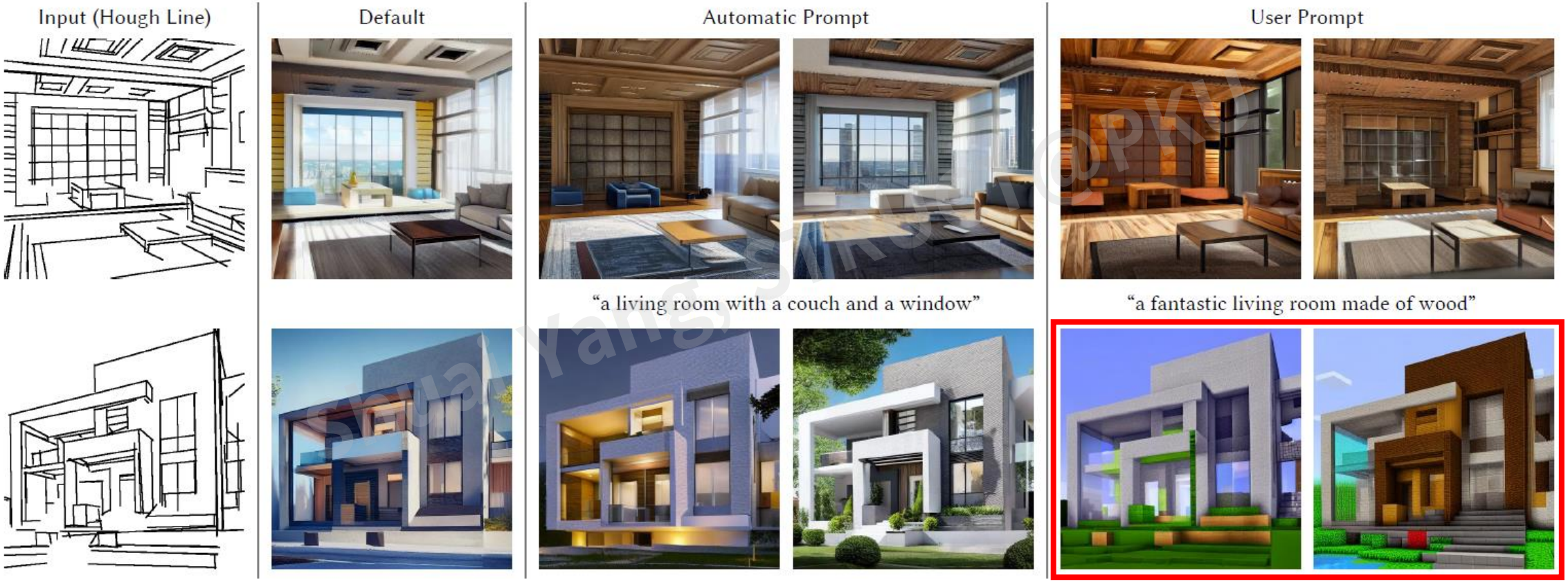
1 ControlNet



Adding Conditional Control to Text-to-Image Diffusion Models. ICCV'23

Controlling pre-trained models

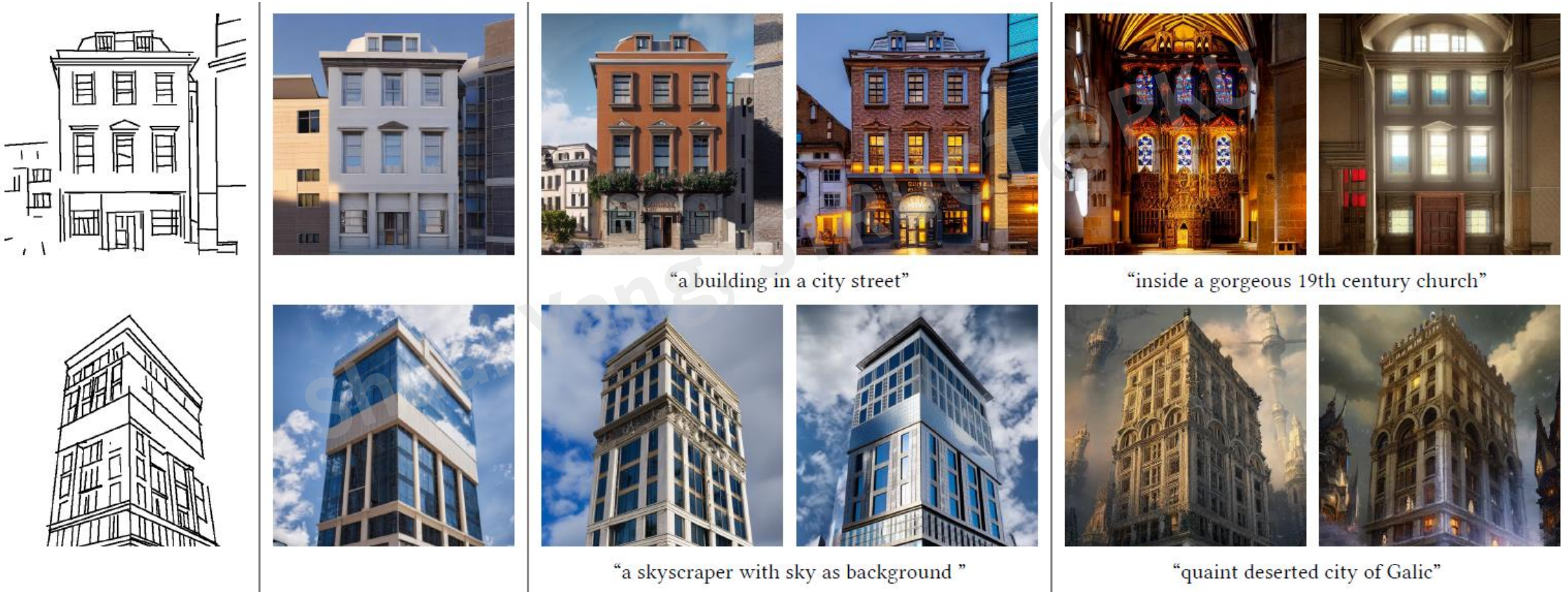
1 ControlNet



Adding Conditional Control to Text-to-Image Diffusion Models. ICCV'23

Controlling pre-trained models

1 ControlNet



Adding Conditional Control to Text-to-Image Diffusion Models. ICCV'23

Controlling pre-trained models

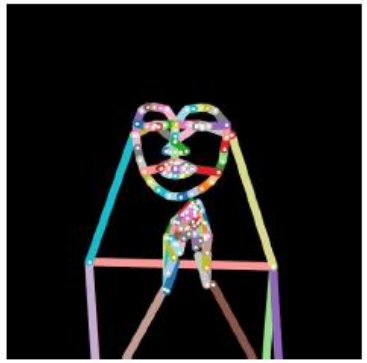
1 ControlNet

Adding Conditional Control to Text-to-Image Diffusion Models. ICCV'23

Controlling pre-trained models

1 ControlNet

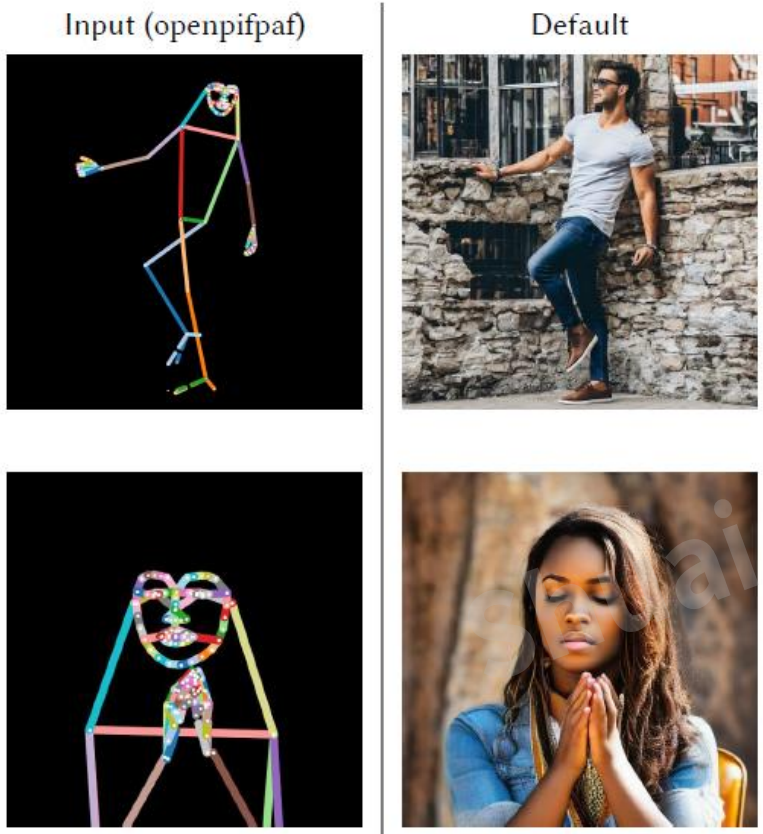
Input (openpifpaf)



Adding Conditional Control to Text-to-Image Diffusion Models. ICCV'23

Controlling pre-trained models

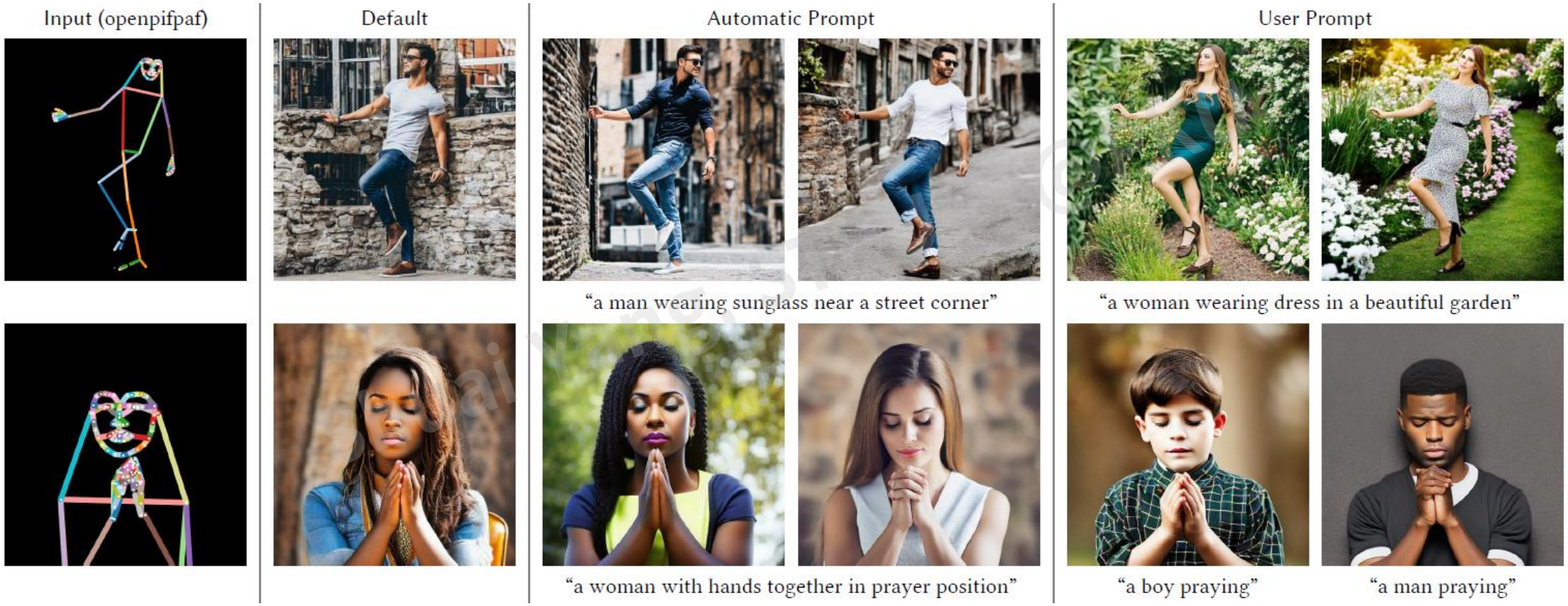
1 ControlNet



Adding Conditional Control to Text-to-Image Diffusion Models. ICCV'23

Controlling pre-trained models

1 ControlNet



Adding Conditional Control to Text-to-Image Diffusion Models. ICCV'23

Controlling pre-trained models

1 ControlNet

Adding Conditional Control to Text-to-Image Diffusion Models. ICCV'23

Controlling pre-trained models

1 ControlNet

Images and Midas Depth

Adding Conditional Control to Text-to-Image Diffusion Models. ICCV'23

Controlling pre-trained models

1 ControlNet

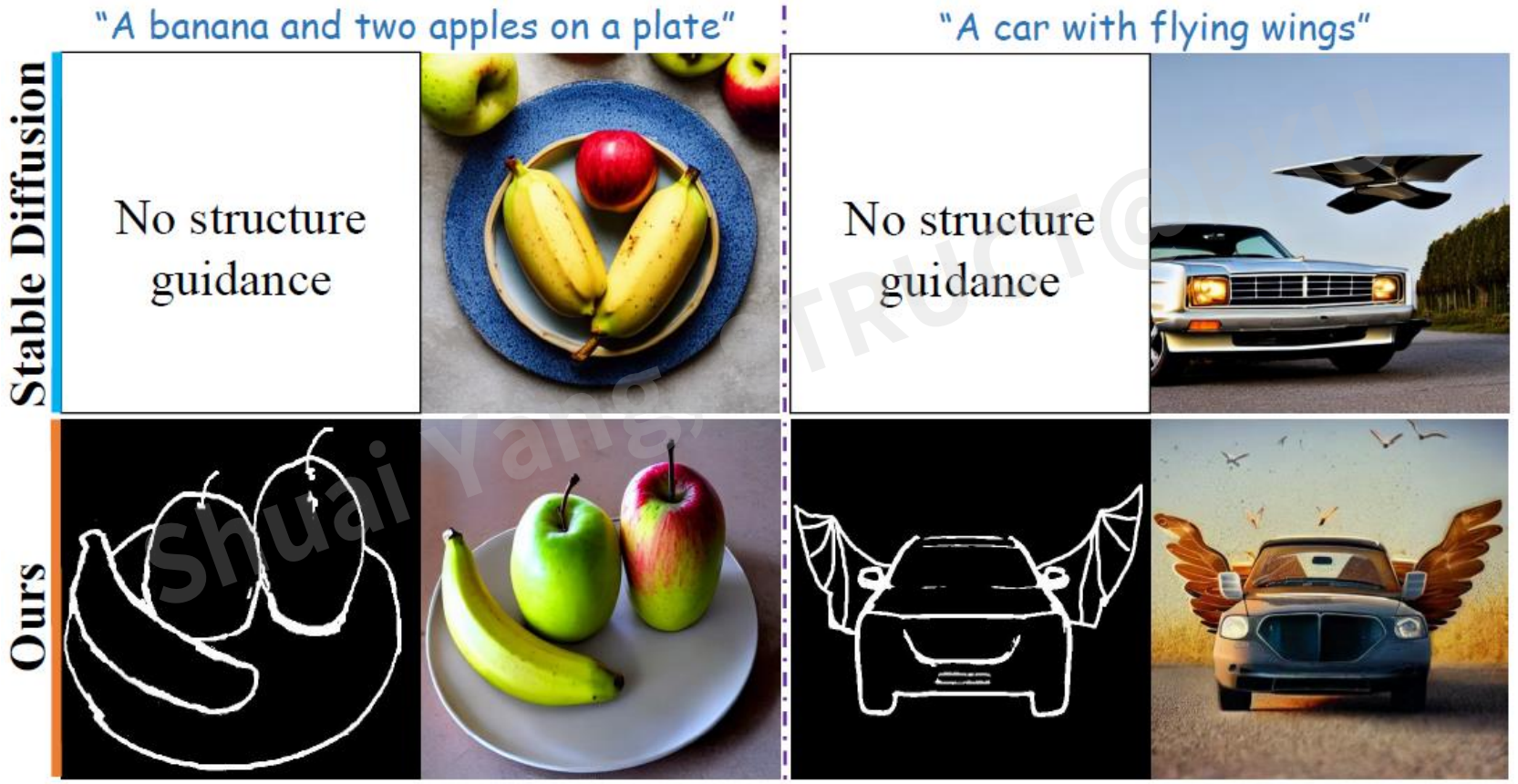
Cartoon line drawing

"1girl, masterpiece, best quality, ultra-detailed, illustration"

Adding Conditional Control to Text-to-Image Diffusion Models. ICCV'23

Controlling pre-trained models

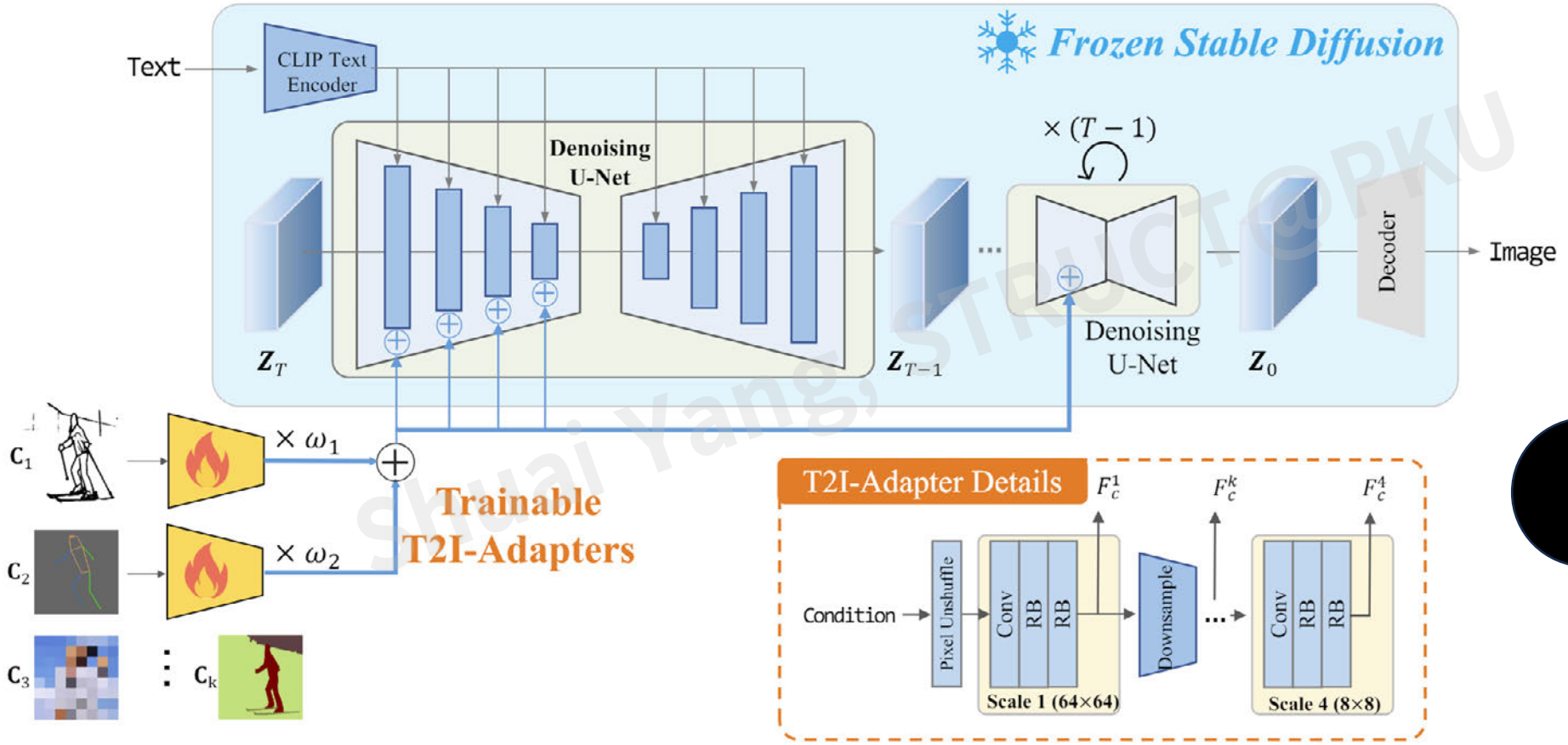
2 T2I Adapter



T2I-Adapter: Learning Adapters to Dig Out More Controllable Ability for Text-to-Image Diffusion Models. AAAI'24

Controlling pre-trained models

2 T2I Adapter



Similar design with ControlNet

T2I-Adapter: Learning Adapters to Dig Out More Controllable Ability for Text-to-Image Diffusion Models. AAAI'24

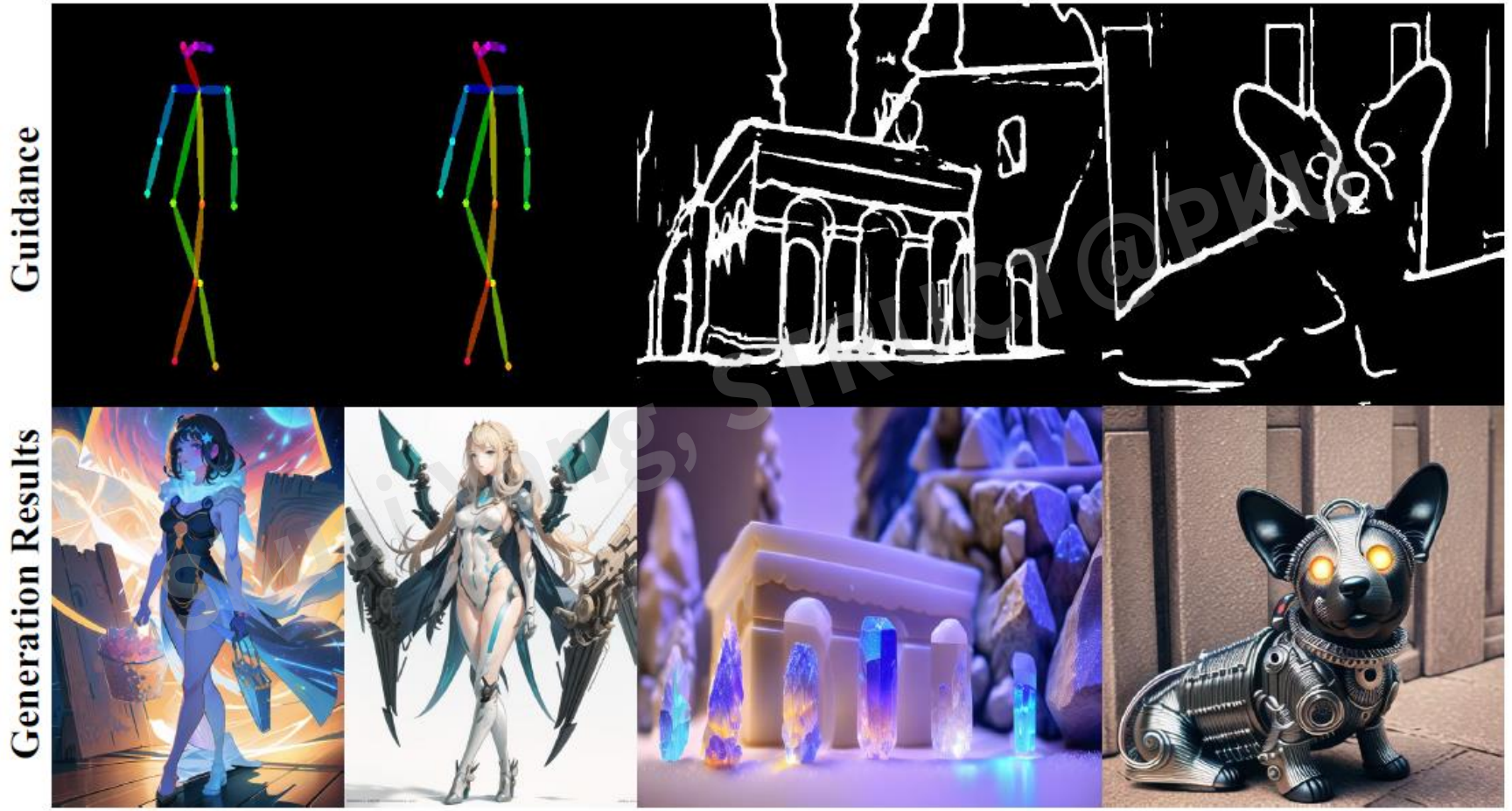
Controlling pre-trained models

2 T2I Adapter

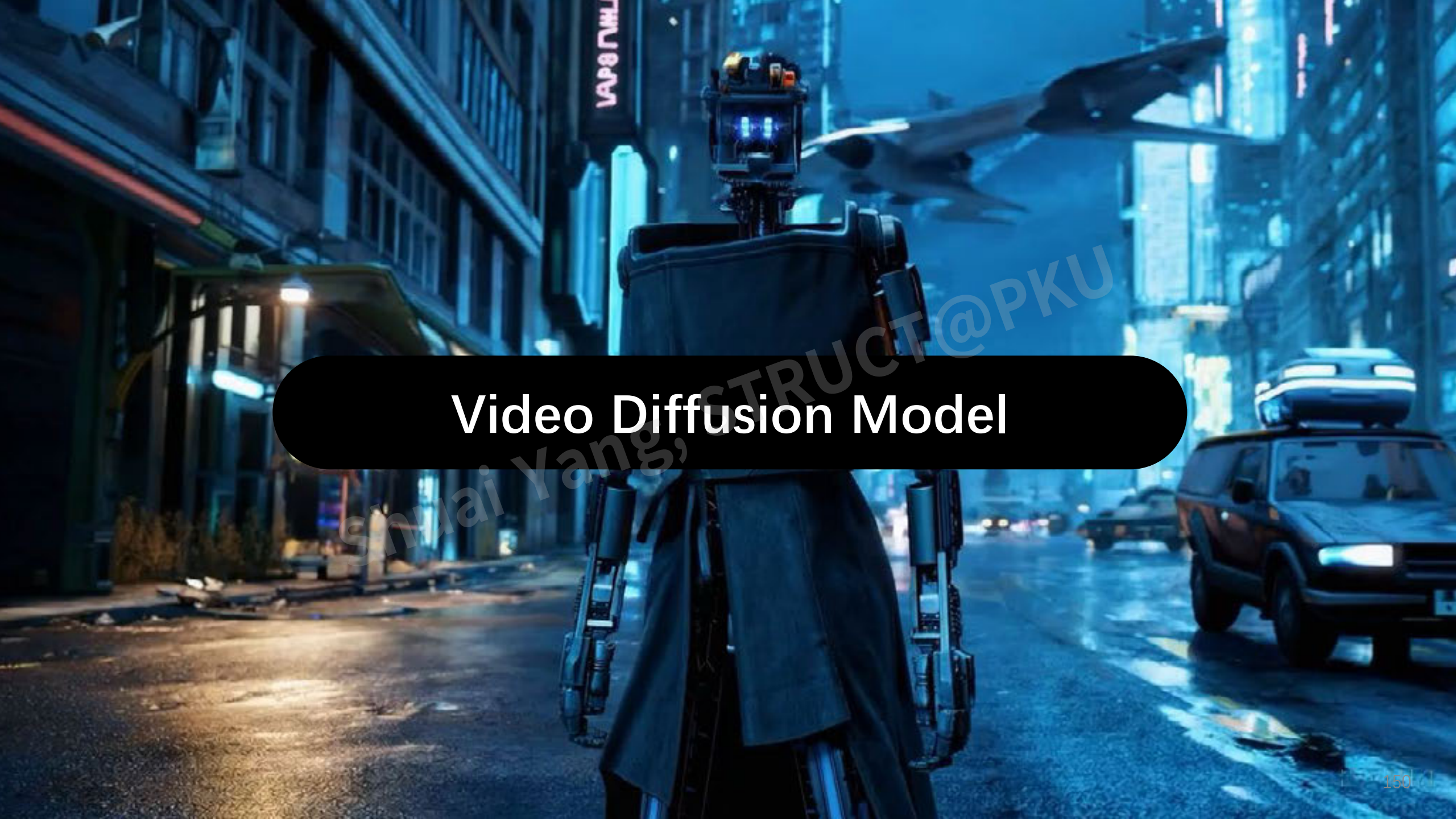
T2I-Adapter: Learning Adapters to Dig Out More Controllable Ability for Text-to-Image Diffusion Models. AAAI'24

Controlling pre-trained models

2 T2I Adapter

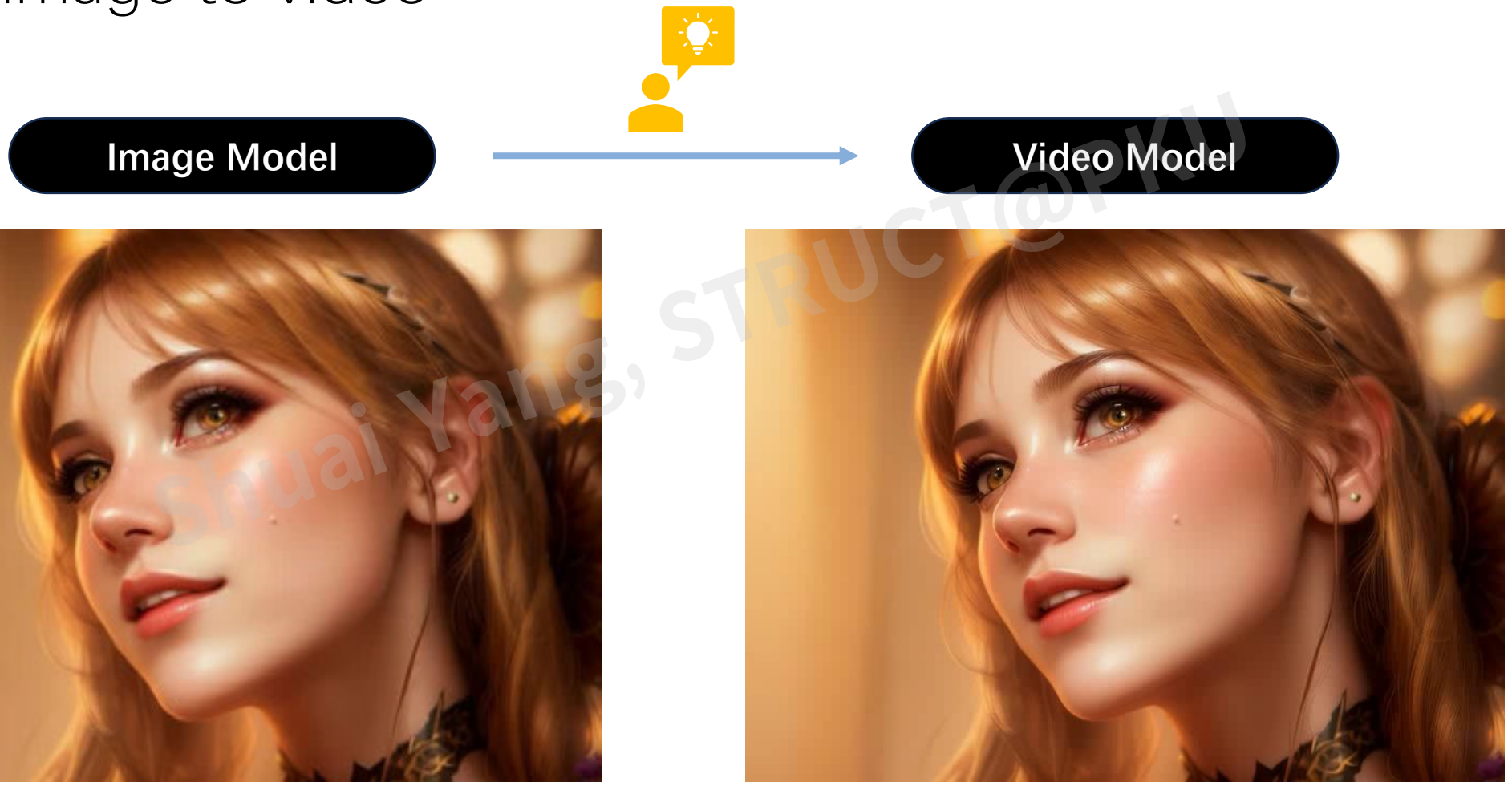


T2I-Adapter: Learning Adapters to Dig Out More Controllable Ability for Text-to-Image Diffusion Models. AAAI'24


Video Diffusion Model
Shidai Yang, STRUCT@PKU

Video Diffusion Model

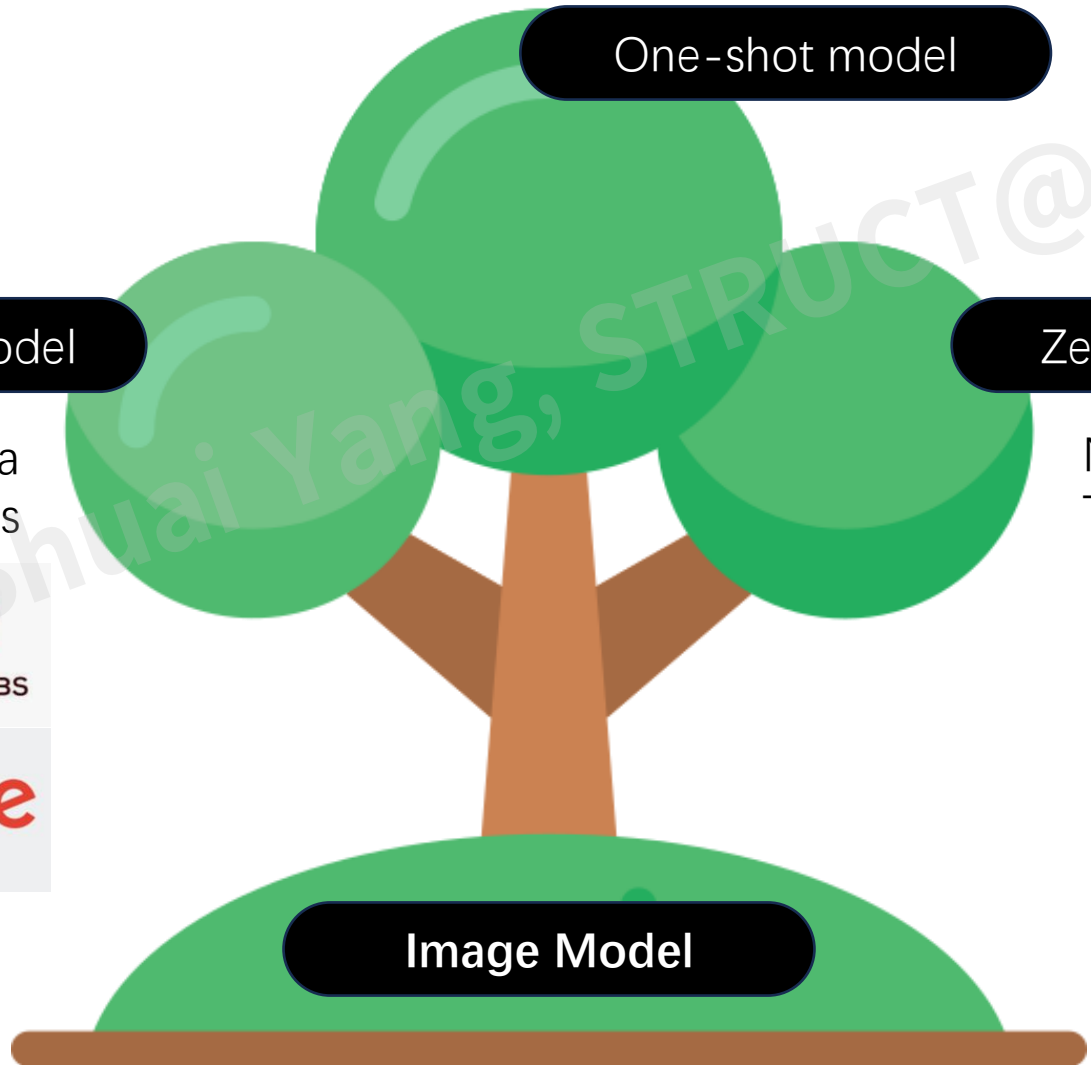
From image to video



A beautiful woman in CG style

Video Diffusion Model

From image to video



Fine-tune on one video
Test on that one video

Train on large data
Test on arbitrary videos

No training
Test on arbitrary videos

Sora
PIKA LABS
GEN-2 runway
Google Research

Video Diffusion Model

From image to video

Data-driven model

- Good performance
- Need large video data and high computational resources
- Good models are not open sourced

One-shot model

- Less training resource required
- Good compatibility
- Overfit to a single video
- Less efficient during testing

Zero-shot model

- No training resource required
- High compatible with image model
- Lack the knowledge of real-world motion

Video Diffusion Model

From image to video

Data-driven model

- Good performance
- Need large video data and high computational resources
- Good models are not open sourced

One-shot model

- Less training resource required
- Good compatibility
- Overfit to a single video
- Less efficient during testing

Zero-shot model

- No training resource required
- High compatible with image model
- Lack the knowledge of real-world motion

Video Diffusion Model

From image to video

Data-driven model

- Good performance
- Need large video data and high computational resources
- Good models are not open sourced

One-shot model

- Less training resource required
- Good compatibility
- Overfit to a single video
- Less efficient during testing

Zero-shot model

- No training resource required
- High compatible with image model
- Lack the knowledge of real-world motion

Video Diffusion Model

From image to video

Data-driven model

TASK

- Text-to-Video Generation
- Image-to-Video Generation
- Video Editing

One-shot model

TASK

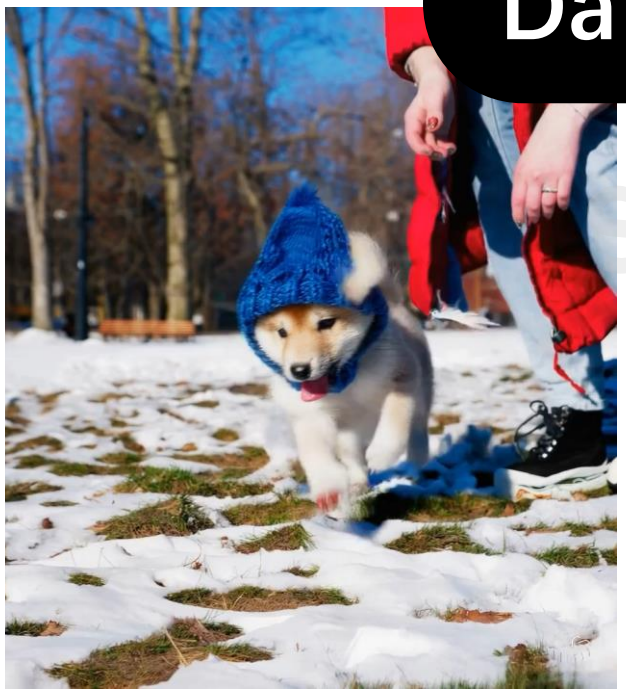
- Video Editing

Zero-shot model

TASK

- Text-to-Video Generation
- Video Editing

Data-Driven Video Diffusion Model



Data-driven model

Non-Diffusion Era → Diffusion Era

- GAN
- Auto-regressive transformers
  - **CogVideo**



CogVideo: Large-scale Pretraining for Text-to-Video Generation via Transformers. ICLR'23

Data-driven model

1 VDM

Non-Diffusion Era

- GAN
- Auto-regressive transformers

Diffusion Era

- Early exploration —— VDM
  - Architecture: factorized space-time U-Net
  - Training: joint images and videos

Video Diffusion Models. NeurIPS'22

Data-driven model

2 Make-A-Video

A dog wearing a Superhero outfit with red cape flying through the sky

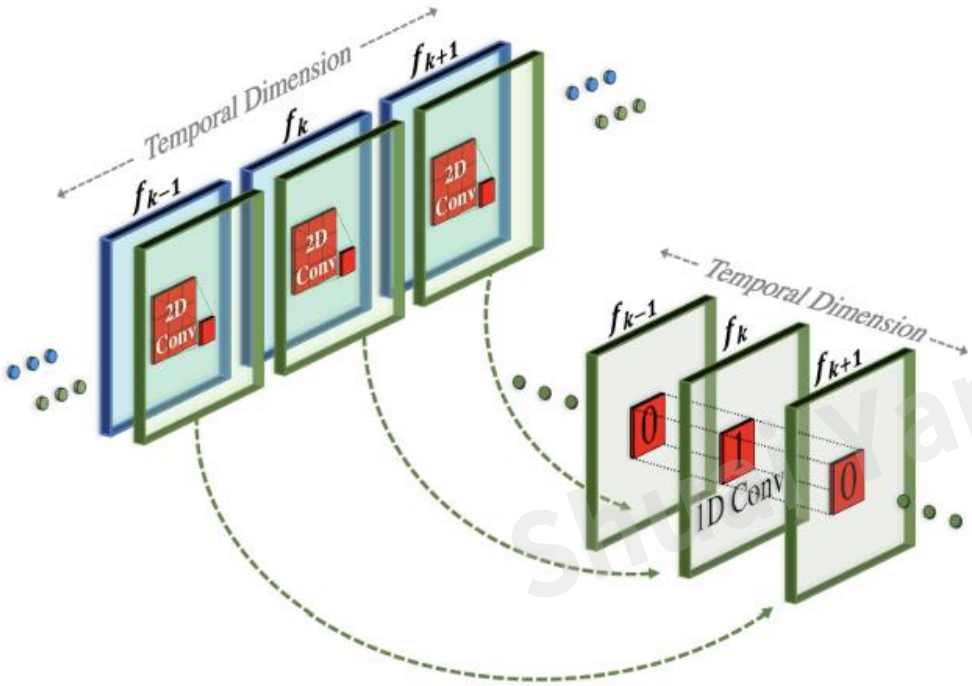
A teddy bear painting a portrait

Robot dancing in times square

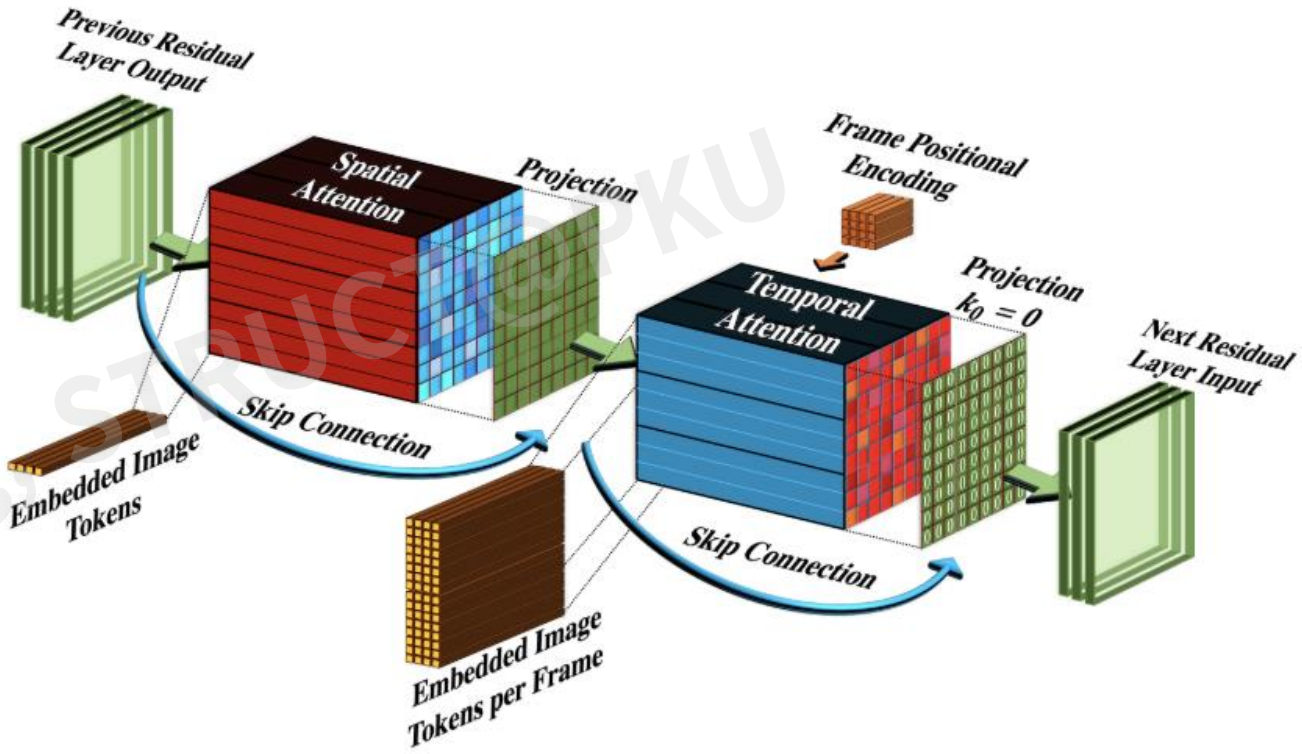
Make-A-Video: Text-to-Video Generation without Text-Video Data. ICLR'23

Data-driven model

2 Make-A-Video



pseudo-3D convolution

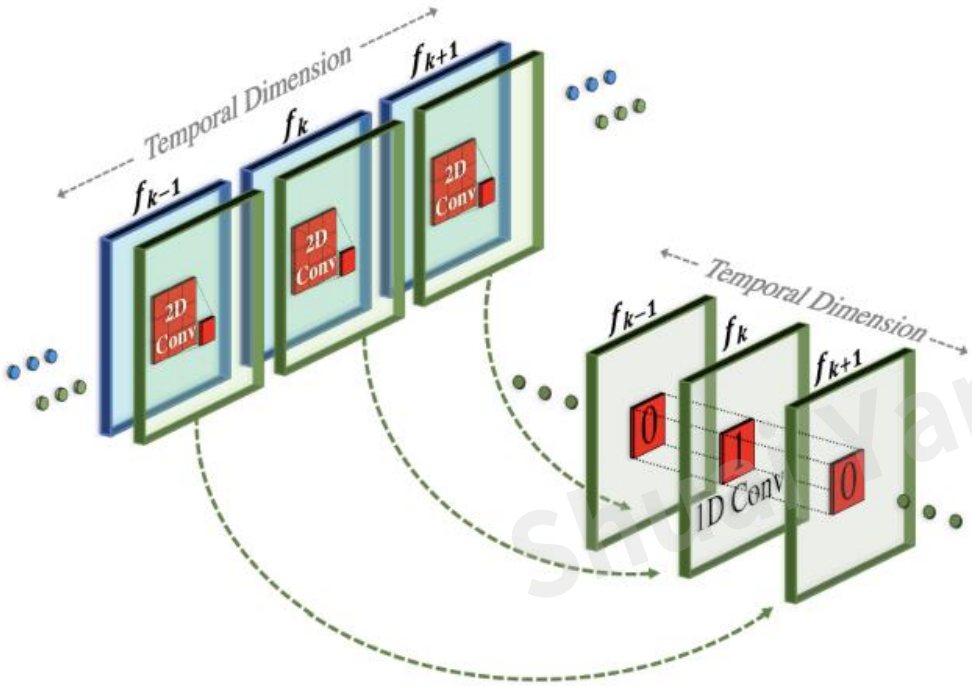


pseudo-3D attention

Make-A-Video: Text-to-Video Generation without Text-Video Data. ICLR'23

Data-driven model

2 Make-A-Video



pseudo-3D convolution

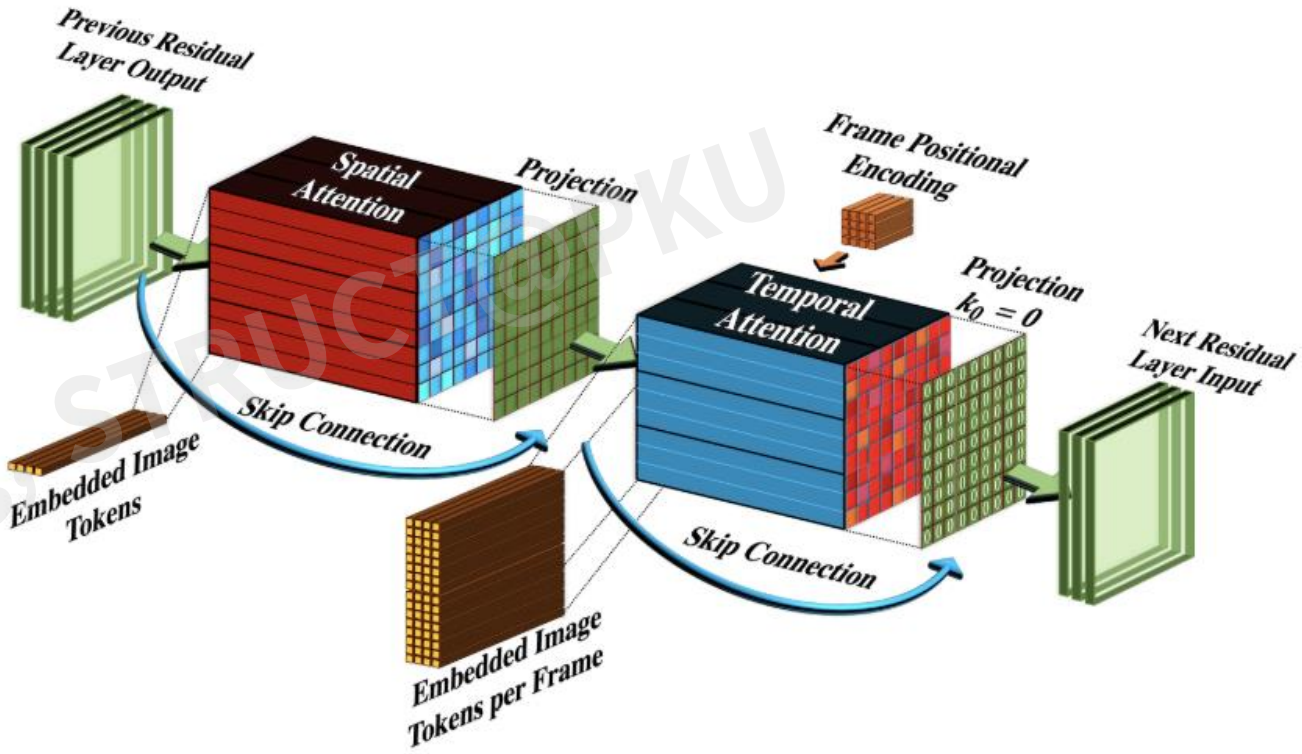
pseudo-3D attention

Make-A-Video: Text-to-Video Generation without Text-Video Data. ICLR'23

Data-driven model

2 Make-A-Video

pseudo-3D convolution

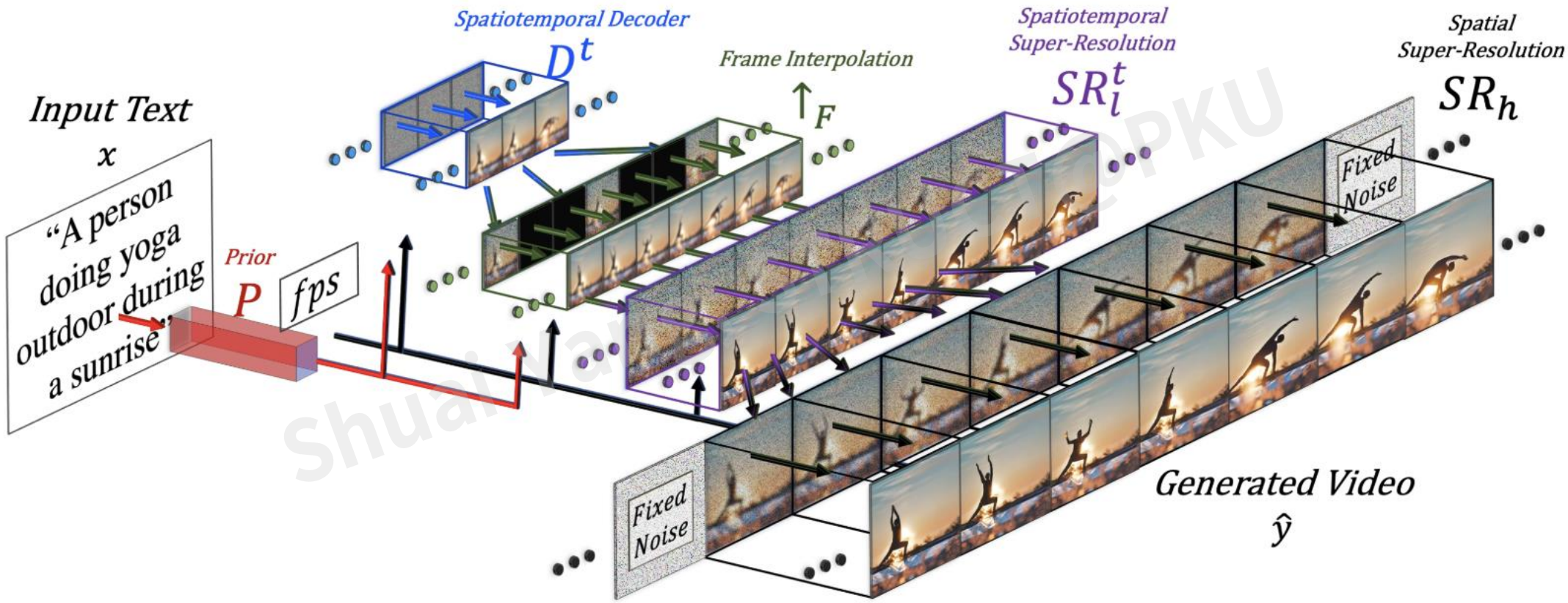


pseudo-3D attention

Make-A-Video: Text-to-Video Generation without Text-Video Data. ICLR'23

Data-driven model

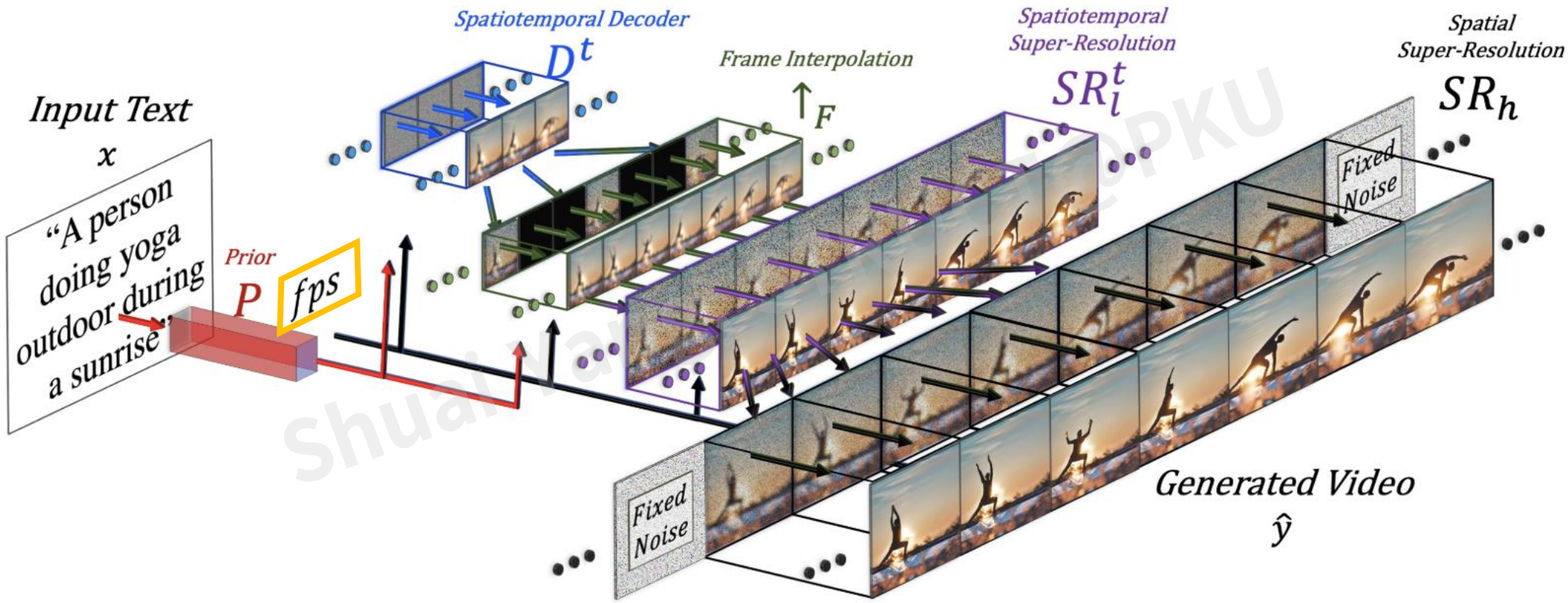
2 Make-A-Video



Make-A-Video: Text-to-Video Generation without Text-Video Data. ICLR'23

Data-driven model

2 Make-A-Video



Make-A-Video: Text-to-Video Generation without Text-Video Data. ICLR'23

Data-driven model

3 Imagen Video

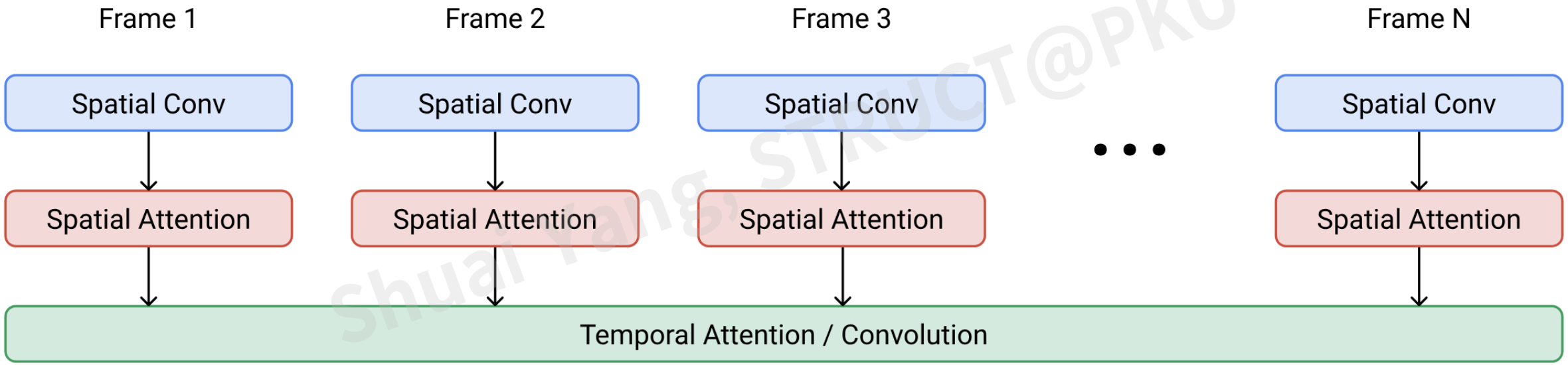
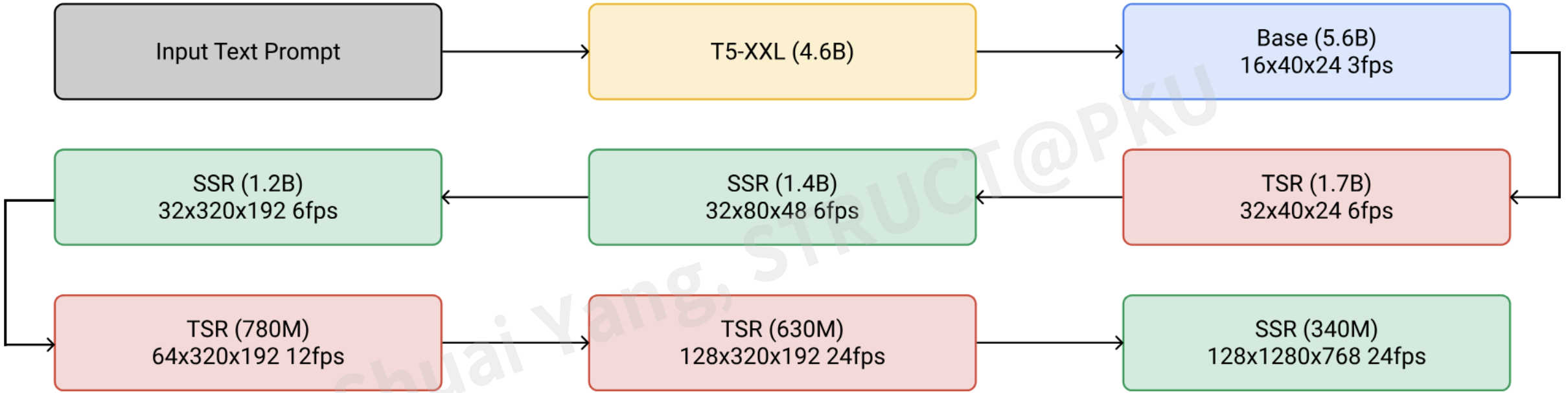


Imagen Video: High Definition Video Generation with Diffusion Models. arXiv'22

Data-driven model

3 Imagen Video



More than 11B parameters to learn

Imagen Video: High Definition Video Generation with Diffusion Models. arXiv'22

Data-driven model

3 Imagen Video

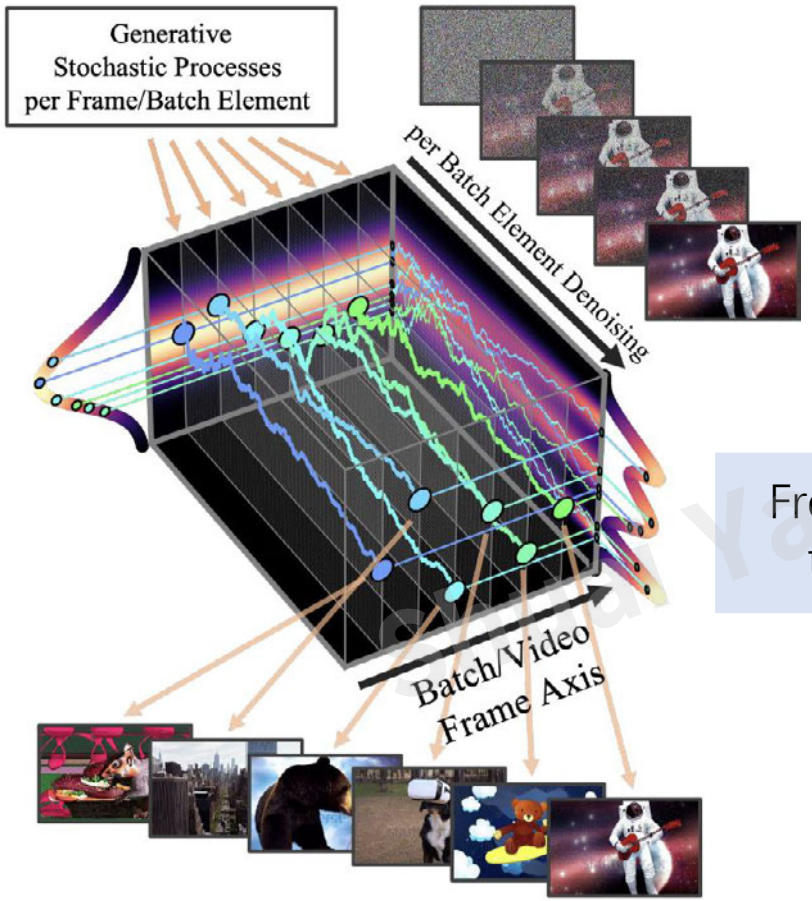
A happy elephant wearing a birthday hat walking under the sea

Wooden figurine surfing on a surfboard in space

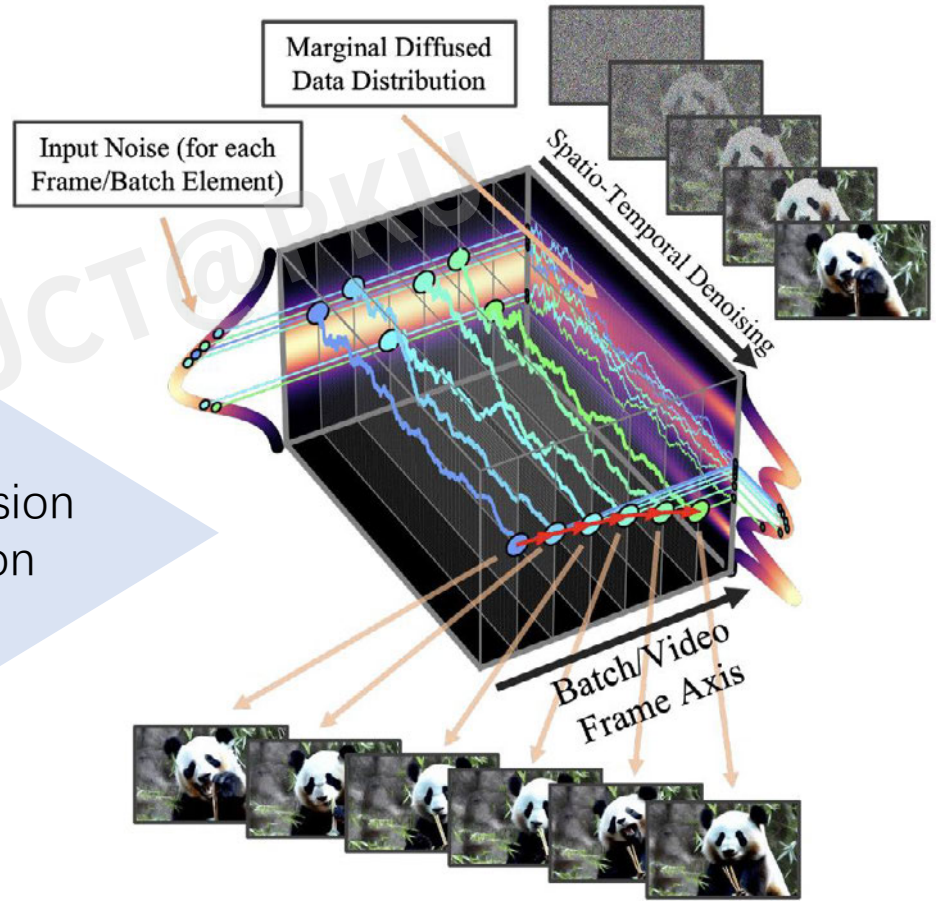
Imagen Video: High Definition Video Generation with Diffusion Models. arXiv'22

Data-driven model

4 Video LDM



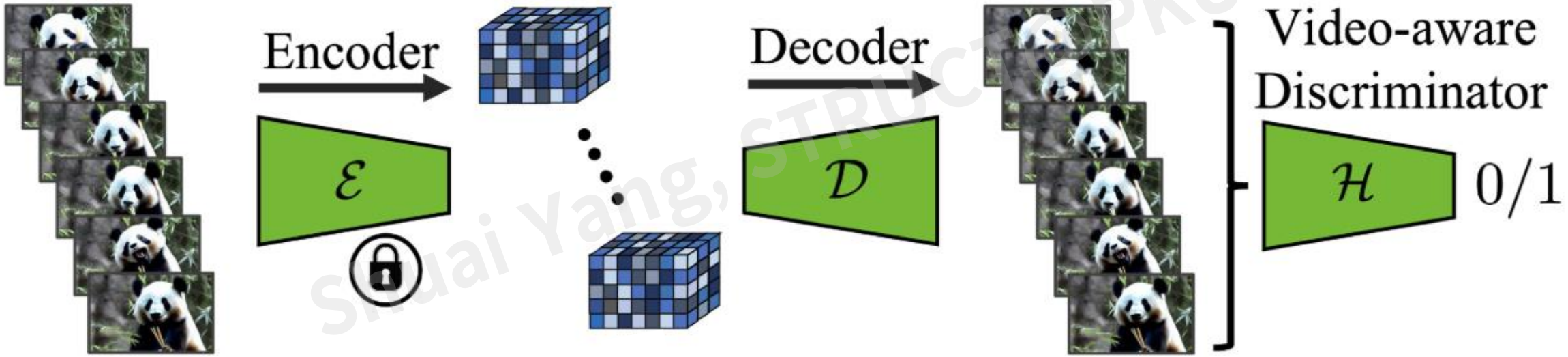
From *image* latent diffusion
to *video* latent diffusion



Align your Latents: High-Resolution Video Synthesis with Latent Diffusion Models. CVPR'23

Data-driven model

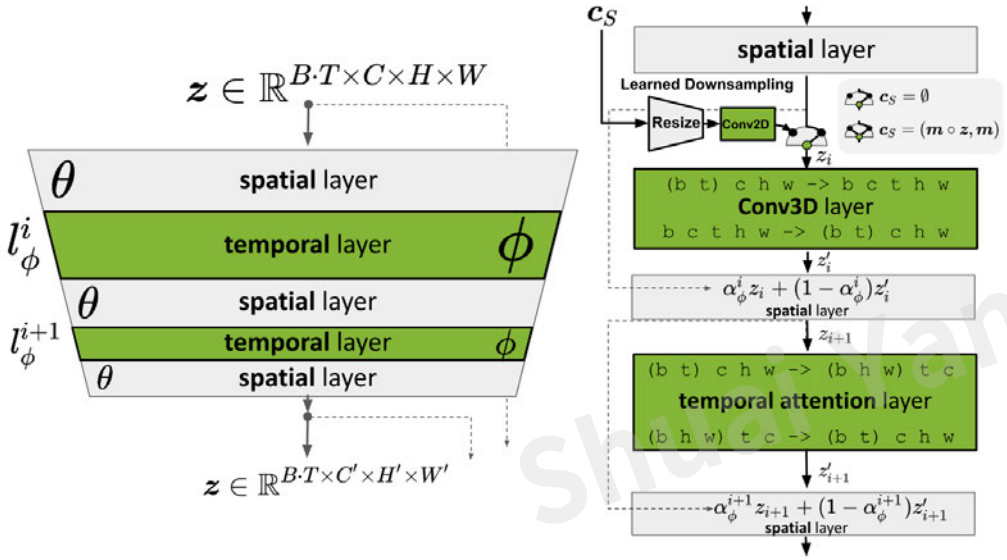
4 Video LDM

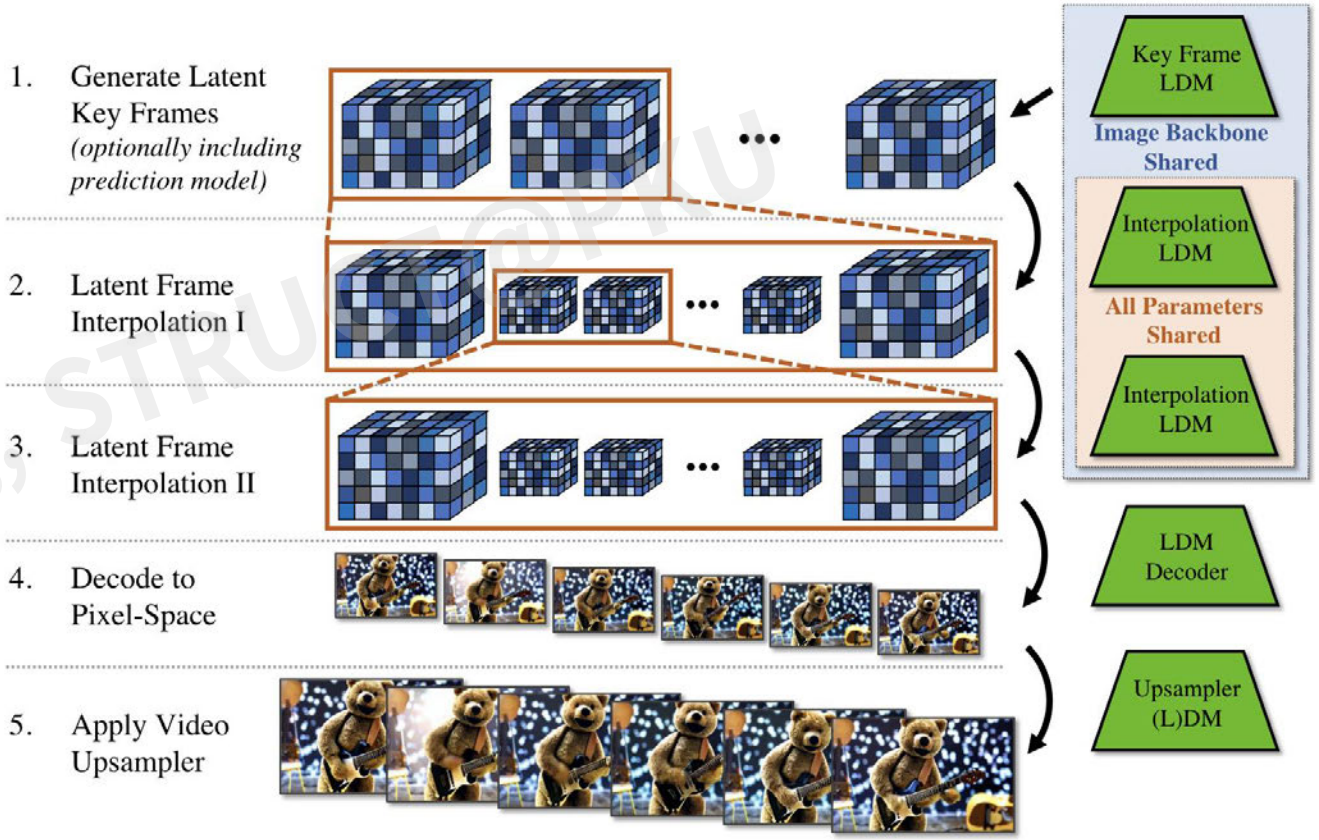


Align your Latents: High-Resolution Video Synthesis with Latent Diffusion Models. CVPR'23

Data-driven model

4 Video LDM





Align your Latents: High-Resolution Video Synthesis with Latent Diffusion Models. CVPR'23

Data-driven model

4 Video LDM

less than 3B parameters to learn

A storm trooper vacuuming the beach

A koala bear playing piano in the forest

Align your Latents: High-Resolution Video Synthesis with Latent Diffusion Models. CVPR'23

Data-driven model

4 Video LDM

Training images for DreamBooth.

Text prompt: "A sks frog playing a guitar in a band."

Align your Latents: High-Resolution Video Synthesis with Latent Diffusion Models. CVPR'23

Data-driven model

Problem: Lack of high-quality data

An example from the WebVid-10M dataset

Results of models trained on WebVid-10M

ModelScope T2V **LVDM**

Man walking in deep snow under the branches

Data-driven model

Ideal text-to-video training data

- **Video Contents**

  Clean and dynamic

- **Text Annotation**

  Corresponding text description

Example:

A person is holding a long haired dachshund in their arms.

RAW internet videos

- **Video Contents**

  × Multiple scenes

  × Unsuitable visual contents

- **Text Annotation**

  × No corresponding text

Example:

Data-driven model

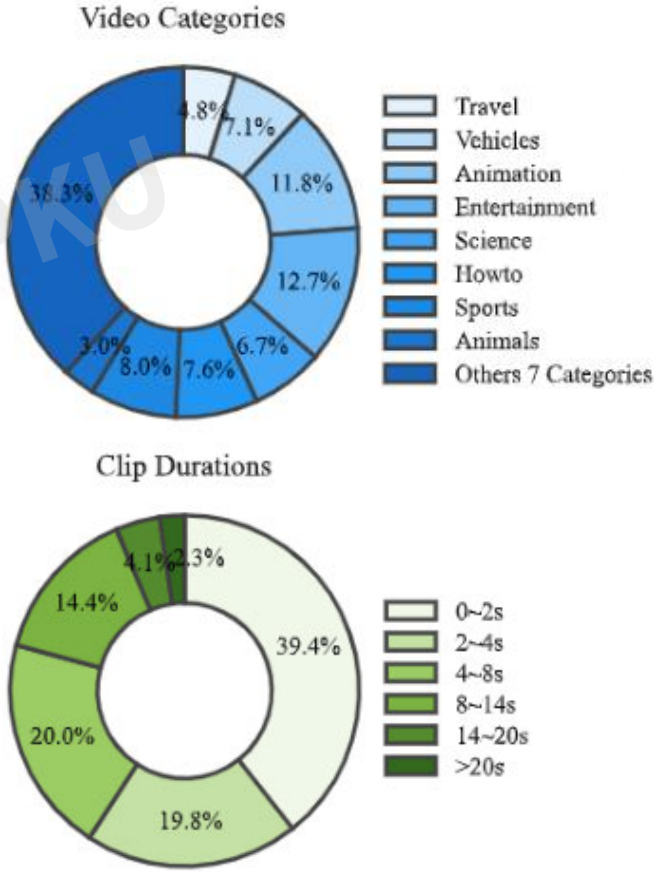
5 HD-VG-130M

Data Collection

- Open-domain videos collected from YouTube

Data Processing

- Detect and split scenes → 130M video clips
- Caption video clips with BLIP-2



Swap Attention in Spatiotemporal Diffusions for Text-to-Video Generation. arXiv'24

Data-driven model

5 HD-VG-130M

Further Data Processing

1. Text Detection

YouTube channel names and subtitles

Swap Attention in Spatiotemporal Diffusions for Text-to-Video Generation. arXiv'24

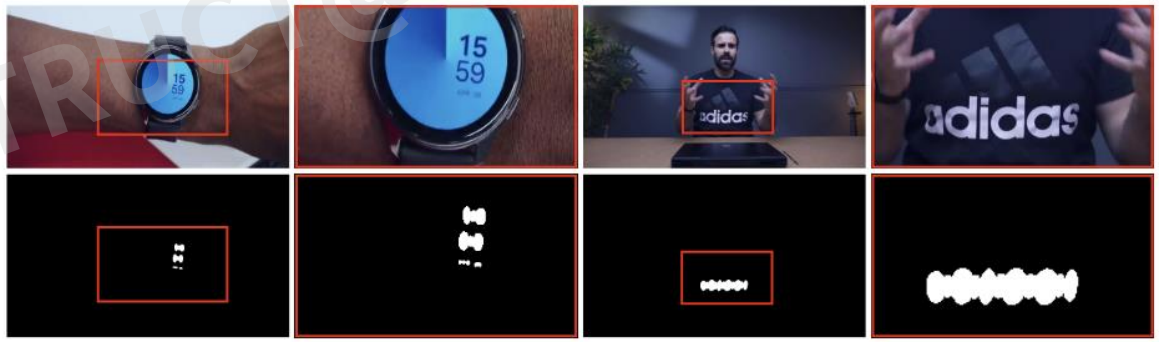
Data-driven model

5 HD-VG-130M

Further Data Processing

1. Text Detection

Videos filtered out due to unrelated text



Video not filtered out

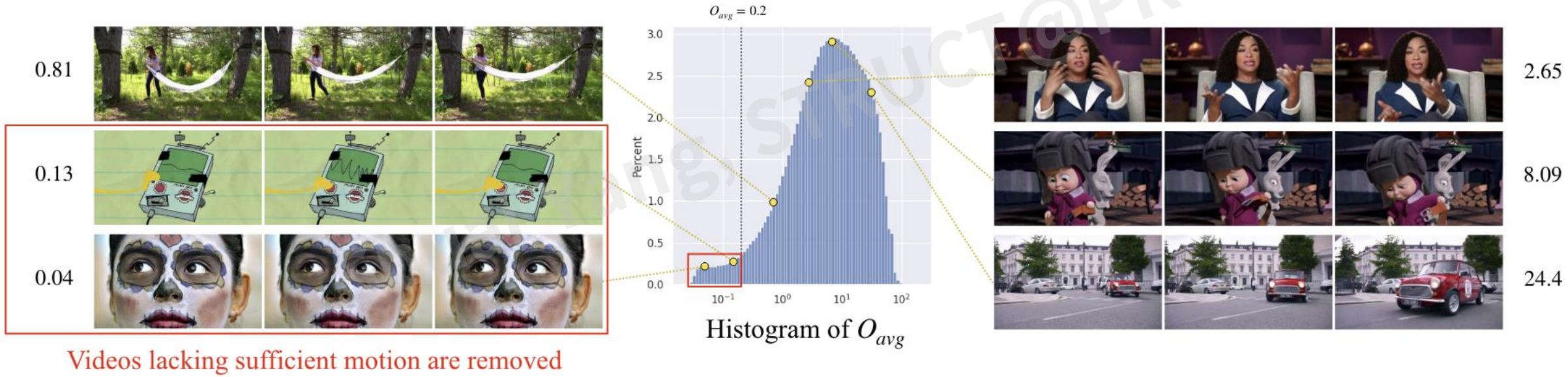
Swap Attention in Spatiotemporal Diffusions for Text-to-Video Generation. arXiv'24

Data-driven model

5 HD-VG-130M

Further Data Processing

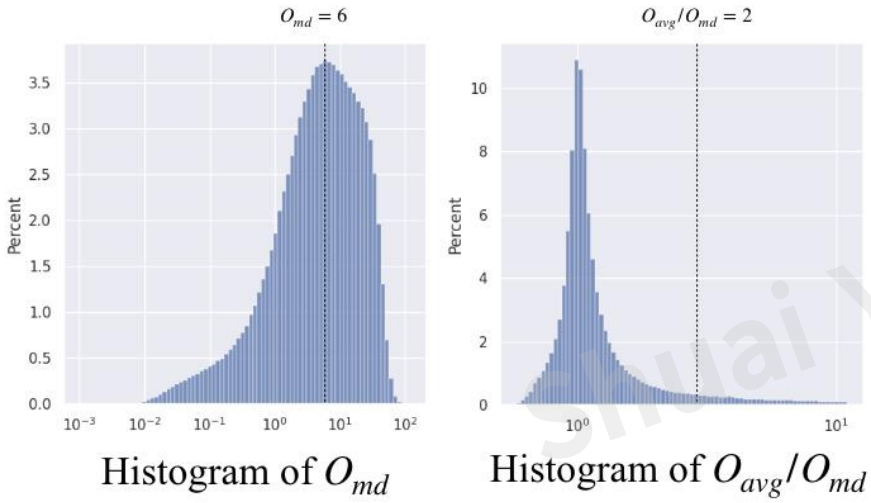
2. Motion Detection



Swap Attention in Spatiotemporal Diffusions for Text-to-Video Generation. arXiv'24

Further Data Processing

2. Motion Detection



Distribution of motion features

$$\begin{cases} O_{avg} = 4.24 \\ O_{md} = 1.38 \\ \dfrac{O_{avg}}{O_{md}} = 3.06 \end{cases}$$

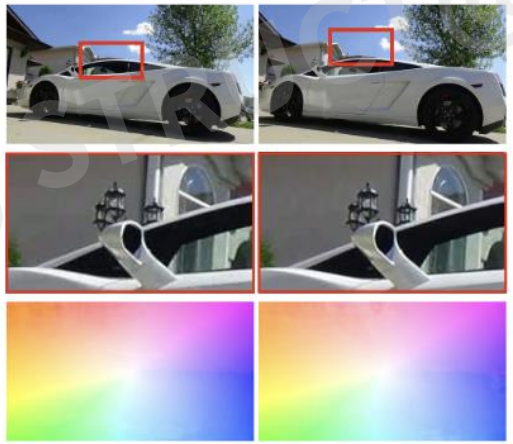


Image zooming transformation
(filtered out)

$$\begin{cases} O_{avg} = 40.9 \\ O_{md} = 9.93 \\ \dfrac{O_{avg}}{O_{md}} = 4.11 \end{cases}$$

Real-world zooming
(not filtered out)

Swap Attention in Spatiotemporal Diffusions for Text-to-Video Generation. arXiv'24

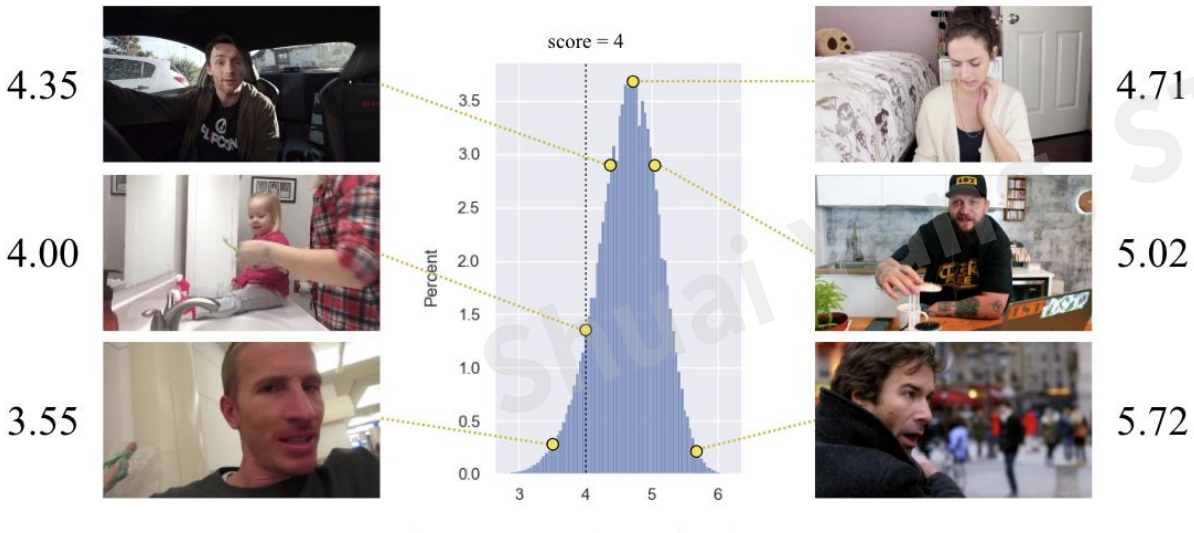
Further Data Processing

3. Aesthetics Evaluation

Swap Attention in Spatiotemporal Diffusions for Text-to-Video Generation. arXiv'24

Further Data Processing

3. Aesthetics Evaluation

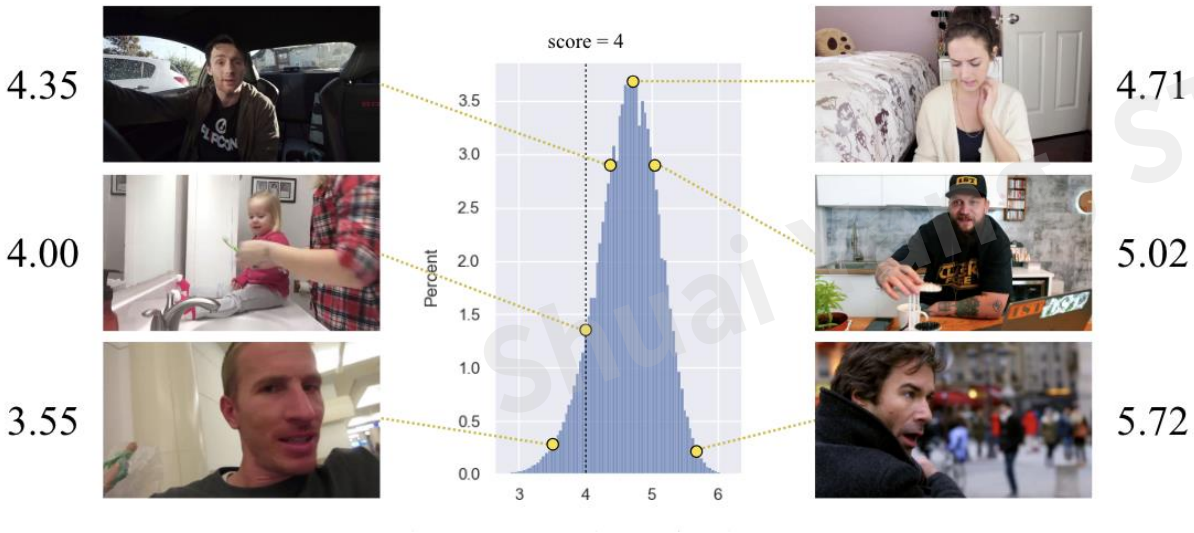


Histogram of aesthetic score

Swap Attention in Spatiotemporal Diffusions for Text-to-Video Generation. arXiv'24

Further Data Processing

3. Aesthetics Evaluation



Histogram of aesthetic score

Samples from HD-VG-130M with aesthetic scores exceeding 6

Swap Attention in Spatiotemporal Diffusions for Text-to-Video Generation. arXiv'24

Data-driven model

5 HD-VG-130M

Original Internet videos with subtitles as text labels

"He thought he was gonna get shows terrible communication on the teams part."

"Yeah, now everybody thought that we couldn't replace cat; yeah, because you're such animal lovers."

Examples from HD-VG-130M

"two men on a boat fishing on a lake"

"a group of rocks covered in seaweed"

Swap Attention in Spatiotemporal Diffusions for Text-to-Video Generation. arXiv'24

Data-driven model

6 Panda-70M

Data Processing

Panda-70M: Captioning 70M Videos with Multiple Cross-Modality Teachers. CVPR'24

Data Processing

- **Semantics-Aware Video Splitting**
  - Stage 1: shot boundary detection
  - Stage 2: merging clips back according to embeddings

Panda-70M: Captioning 70M Videos with Multiple Cross-Modality Teachers. CVPR'24

Data-driven model

6 Panda-70M

Data Processing

- **Semantics-Aware Video Splitting**
- **Captioning with Cross-Modality Teachers**
  - Select 8 video captioning models from 31 candidates based on user study

Panda-70M: Captioning 70M Videos with Multiple Cross-Modality Teachers. CVPR'24

Data-driven model

6 Panda-70M

Data Processing

- **Semantics-Aware Video Splitting**
- **Captioning with Cross-Modality Teachers**
  - Select 8 video captioning models from 31 candidates based on user study
- **Fine-grained Video-to-Text Retrieval**
  - Collect a 100K videos with human selection

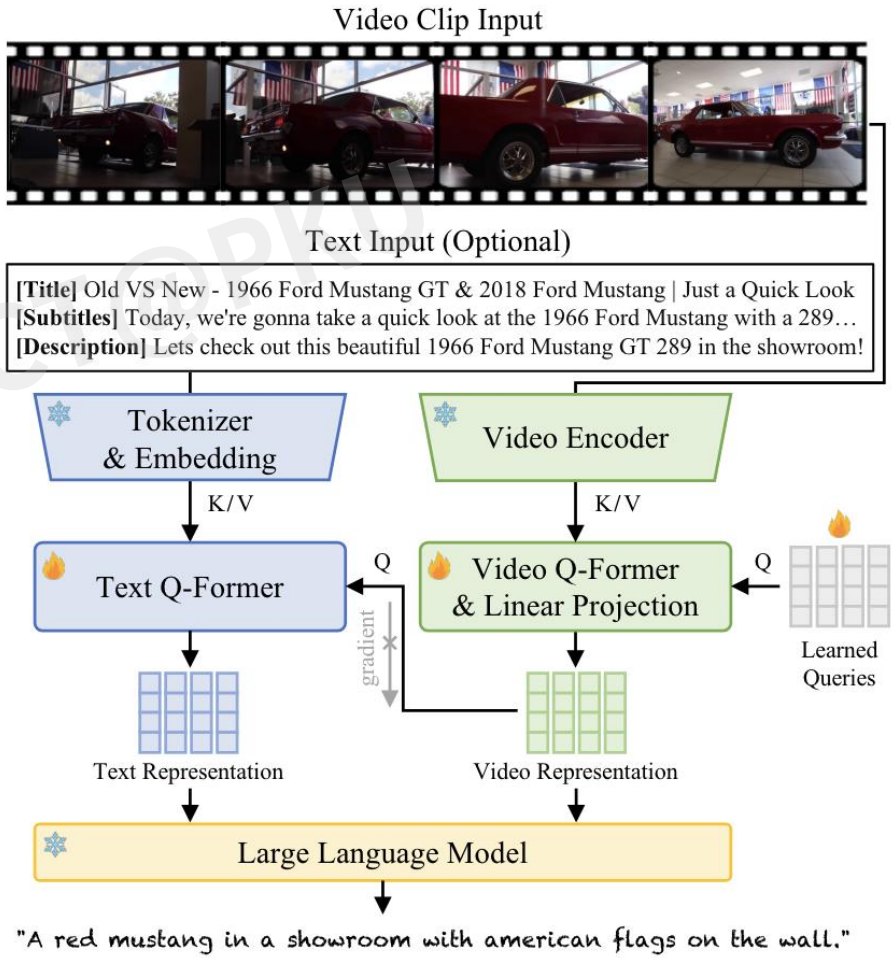
Panda-70M: Captioning 70M Videos with Multiple Cross-Modality Teachers. CVPR'24

Data-driven model

6 Panda-70M

Data Processing

- Semantics-Aware Video Splitting
- Captioning with Cross-Modality Teachers
- Fine-grained Video-to-Text Retrieval
- Multimodal Student Captioning Model



Panda-70M: Captioning 70M Videos with Multiple Cross-Modality Teachers. CVPR'24

Data-driven model

6 Panda-70M

A rhino and a lion are fighting in the dirt

A person is making a pie crust on a table

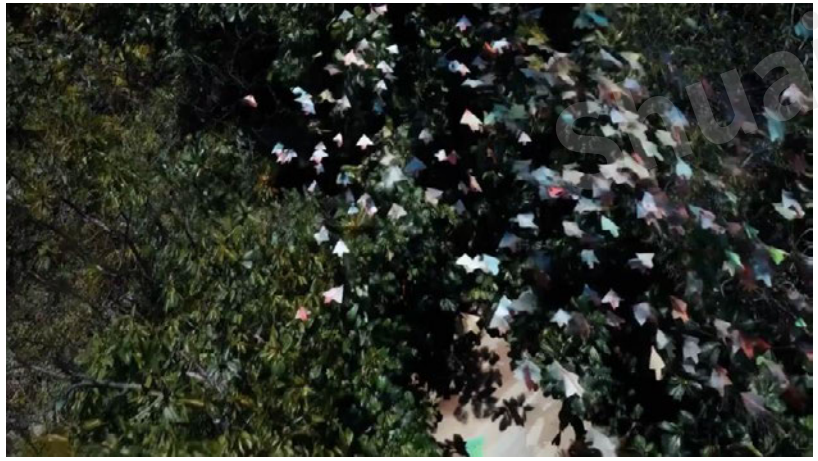
It is a rally car driving on a dirt road in the countryside, with people watching from the side of the road.

Panda-70M: Captioning 70M Videos with Multiple Cross-Modality Teachers. CVPR'24

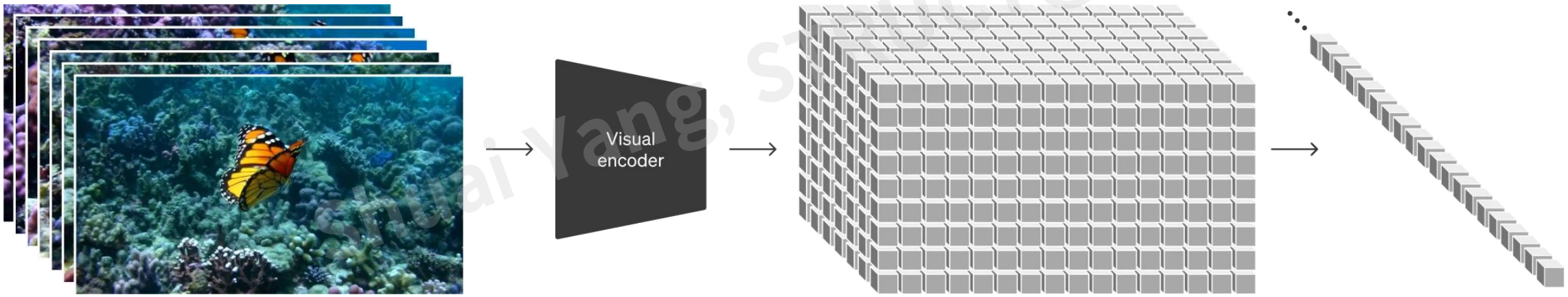
Data-driven model

7 Sora



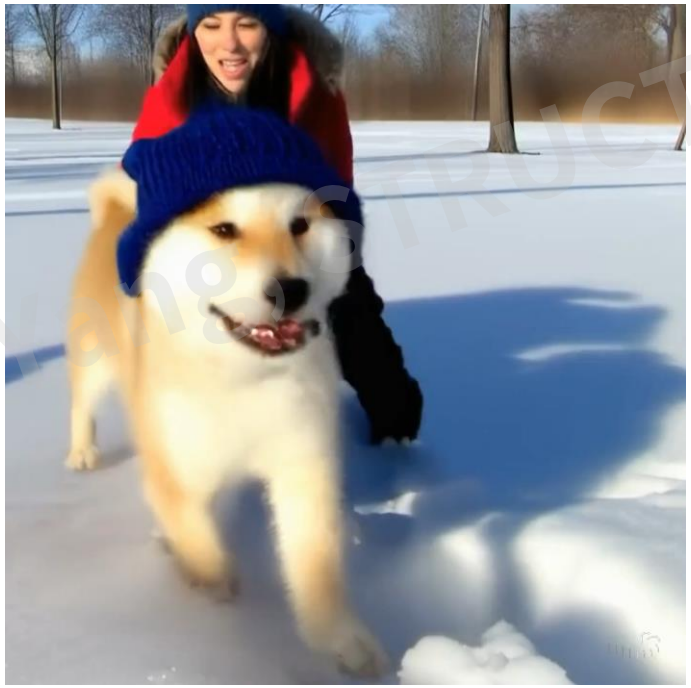
Turning visual data into patches



Diffusion Transformer (DiT)

Scaling transformers for video generation

Base compute



4x compute

32x compute

Sampling flexibility

Improved framing and composition

Trained on square crops

Trained at native aspect ratios

Re-captioning technique

a toy robot wearing **a green dress and a sun hat** taking a pleasant stroll in **Johannesburg, South Africa** during **a winter storm**.

Animating images

A Shiba Inu dog wearing a beret and black turtleneck.

Extending generated videos

Extended backward in time starting from the same video segment

Connecting videos

Data-driven model

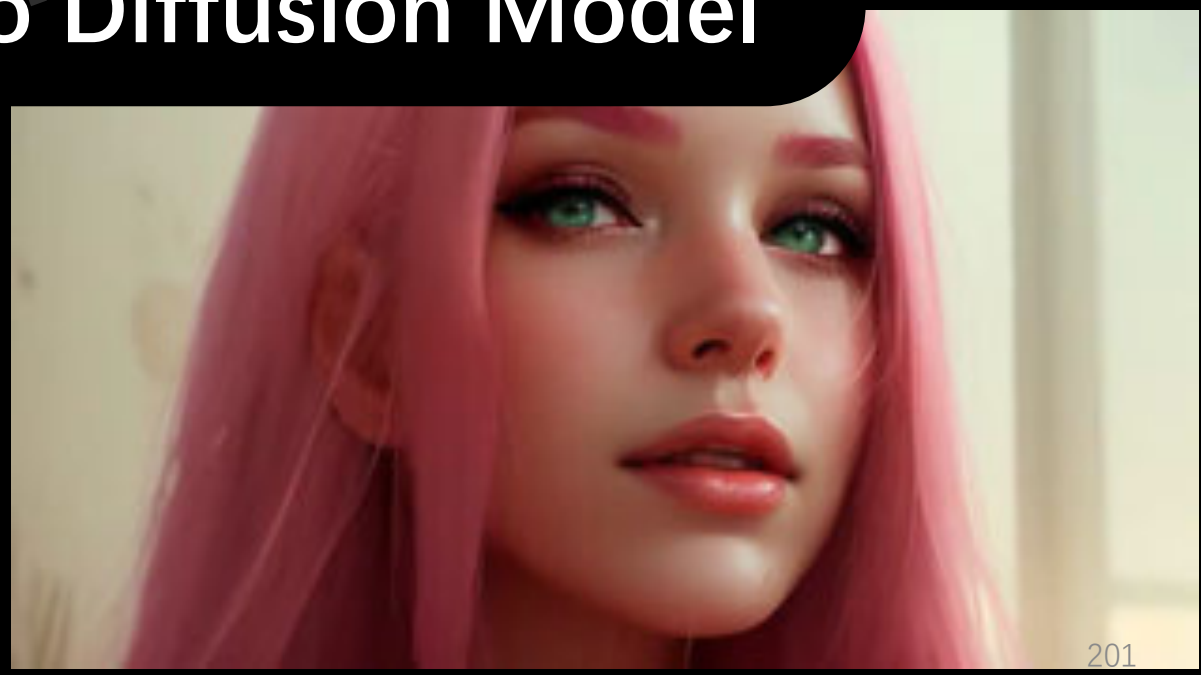
Huge amount of data

Huge computational cost



One/Zero-Shot Video Diffusion Model





One-shot model

1

Tune-A-Video

ICCV 23

2

Edit-A-Video

ACML 23

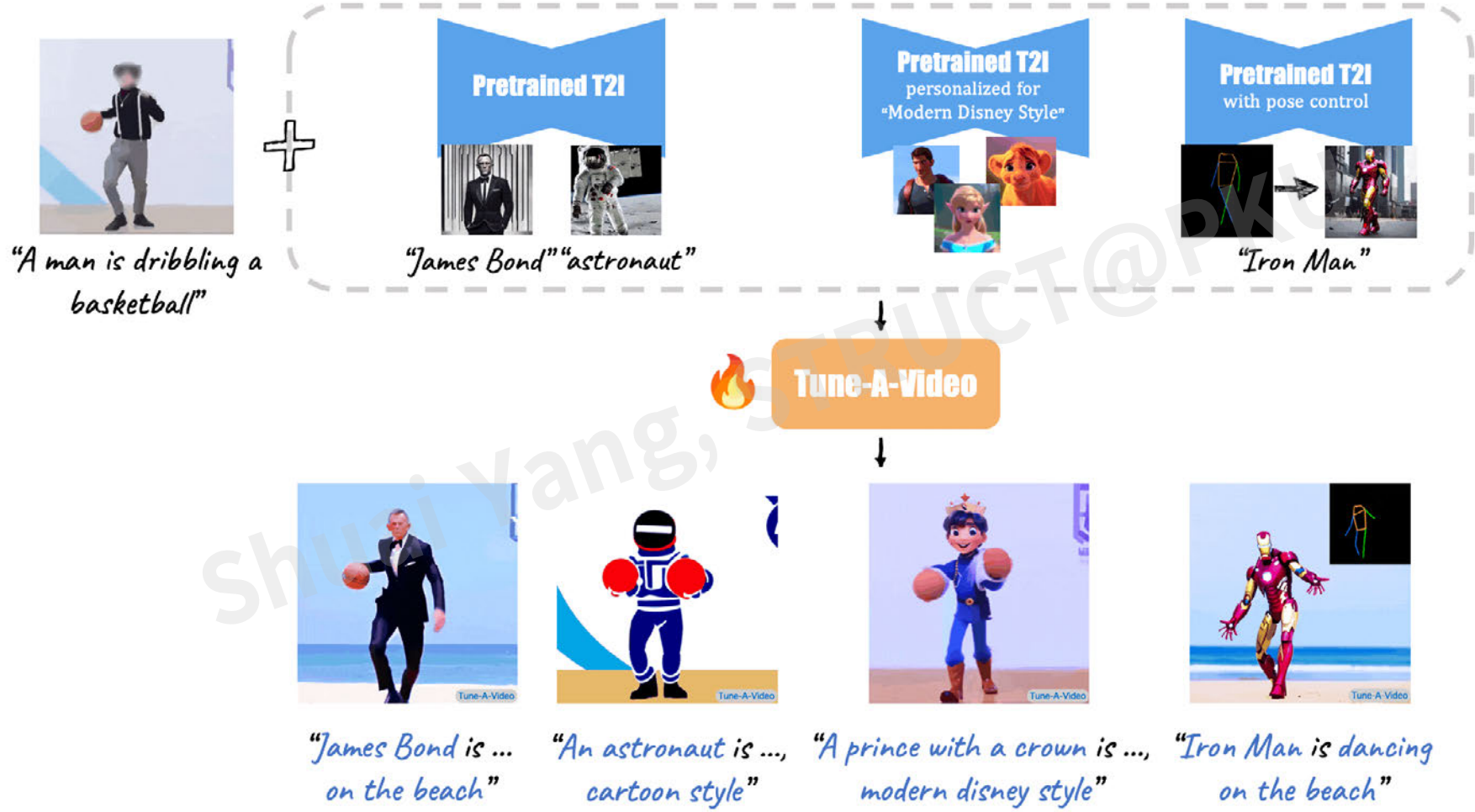
3

Video-P2P

CVPR 24

One-shot model

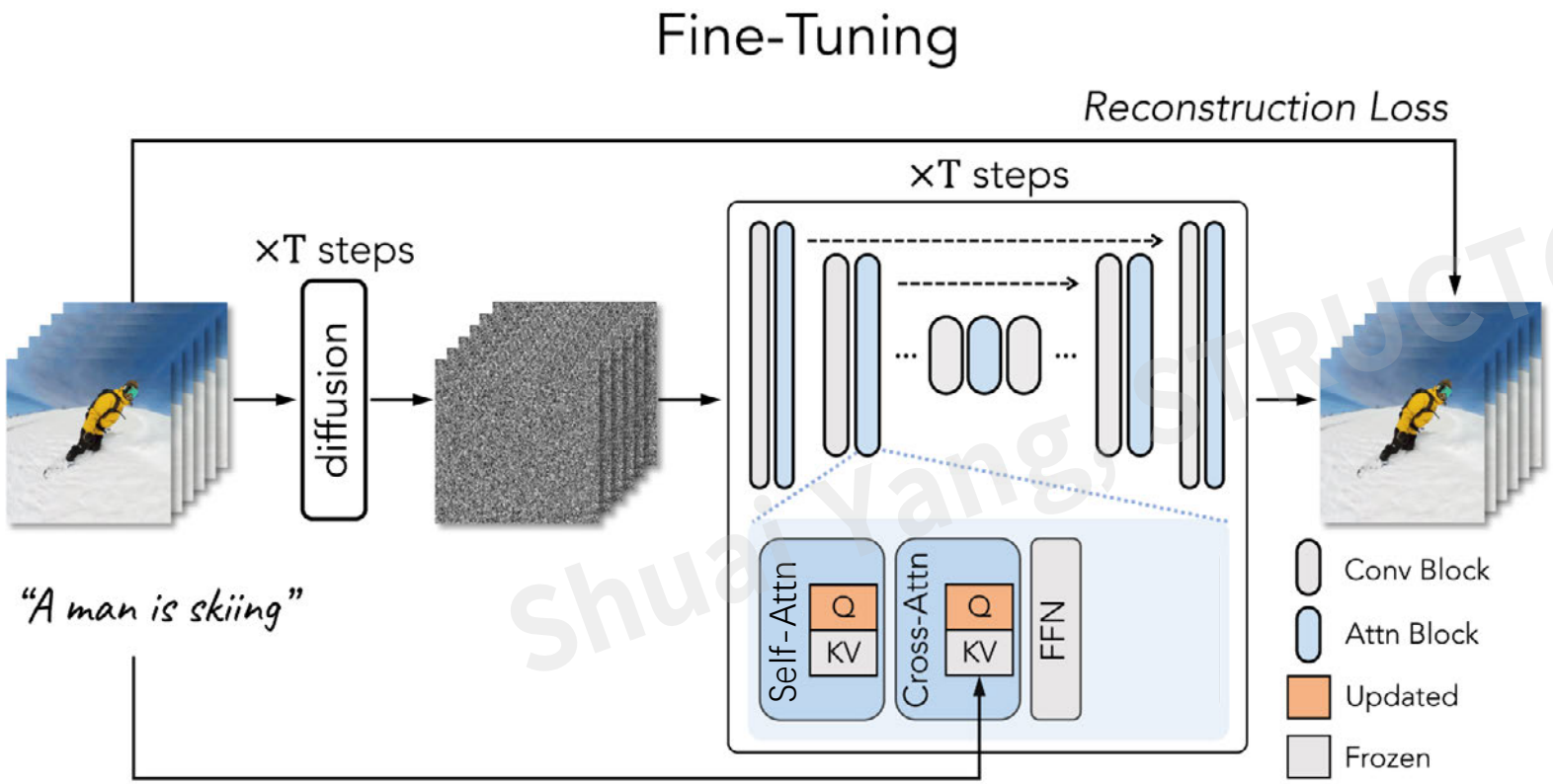
1 Tune-A-Video



Tune-A-Video: One-Shot Tuning of Image Diffusion Models for Text-to-Video Generation. ICCV'23

One-shot model

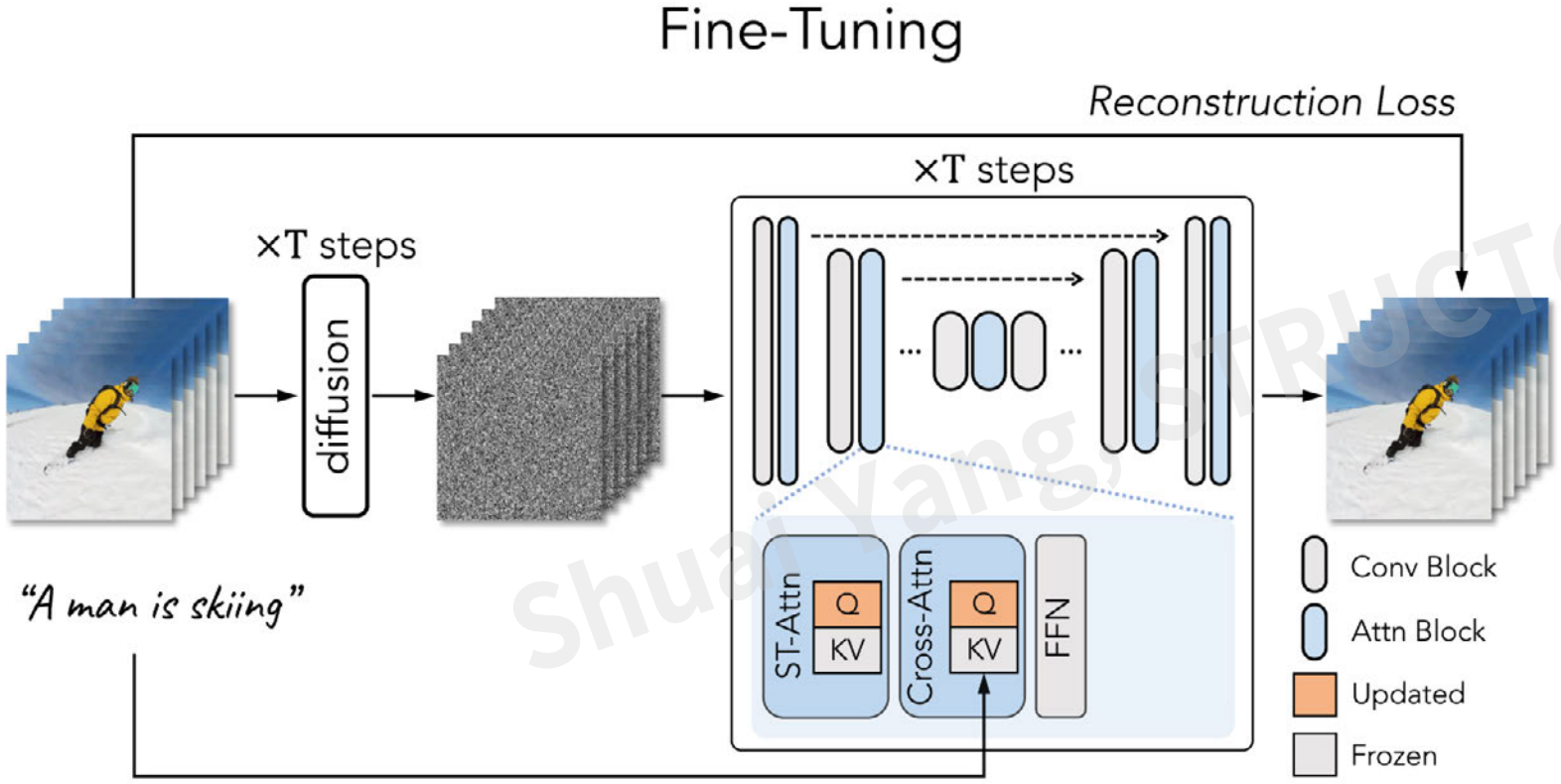
1 Tune-A-Video

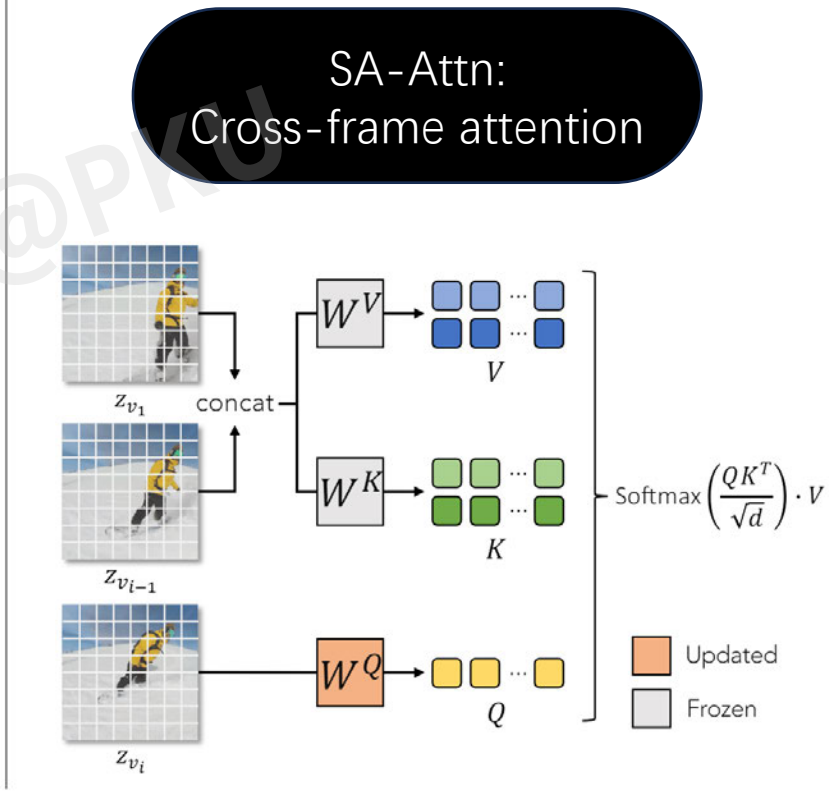


Tune-A-Video: One-Shot Tuning of Image Diffusion Models for Text-to-Video Generation. ICCV'23

One-shot model

1 Tune-A-Video

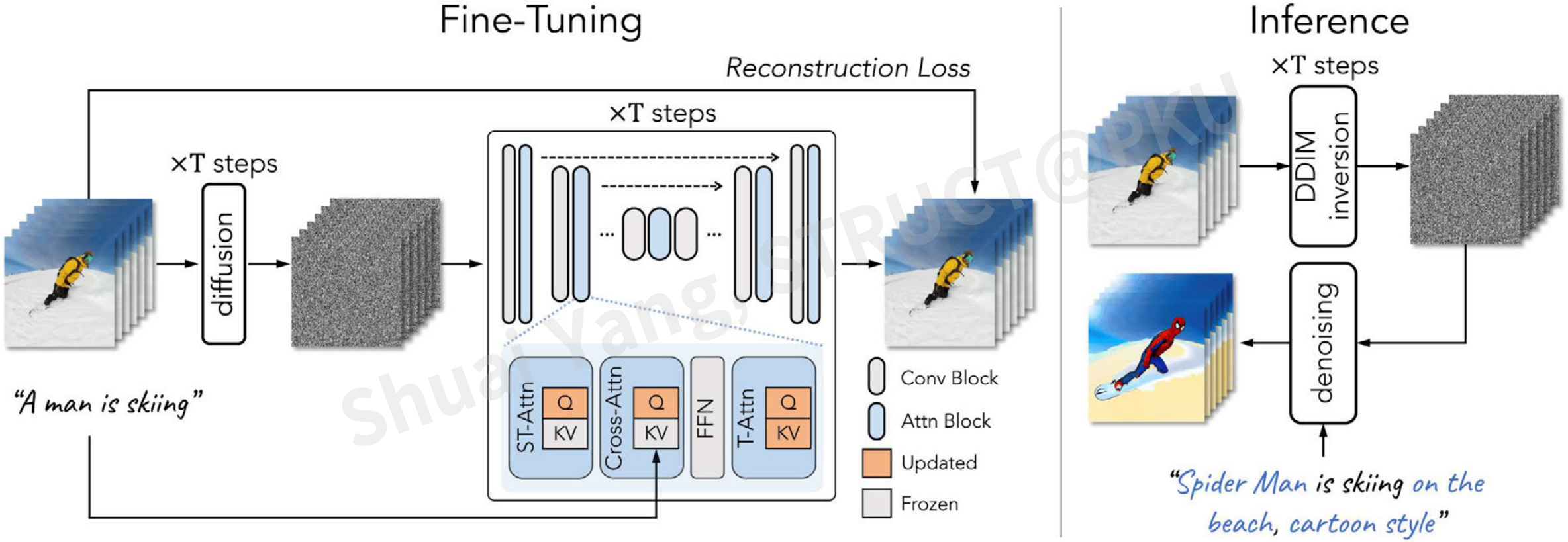




Tune-A-Video: One-Shot Tuning of Image Diffusion Models for Text-to-Video Generation. ICCV'23

One-shot model

1 Tune-A-Video



Tune-A-Video: One-Shot Tuning of Image Diffusion Models for Text-to-Video Generation. ICCV'23

One-shot model

1 Tune-A-Video

A jeep car is moving on the road

Change too much in the background

*A jeep car is moving on the road, **cartoon style***

*A jeep car is moving on the **snow***

Tune-A-Video: One-Shot Tuning of Image Diffusion Models for Text-to-Video Generation. ICCV'23

One-shot model

2 Edit-A-Video

Better background

a man is riding a motorcycle

~~*man*~~
gorilla

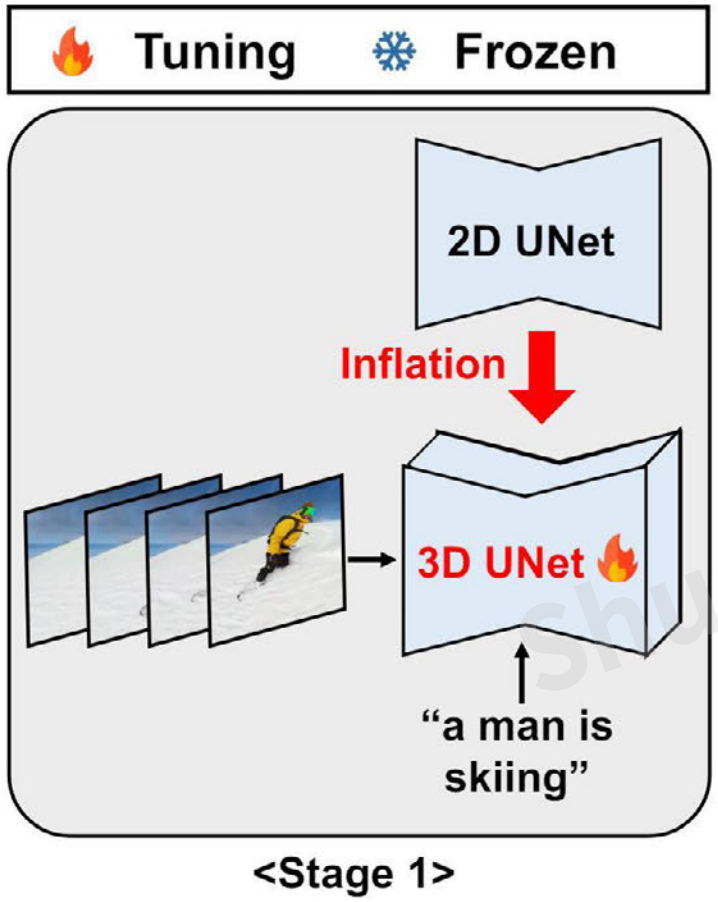
Tune-A-Video

Edit-A-Video

Edit-A-Video: Single Video Editing with Object-Aware Consistency. ACML'23

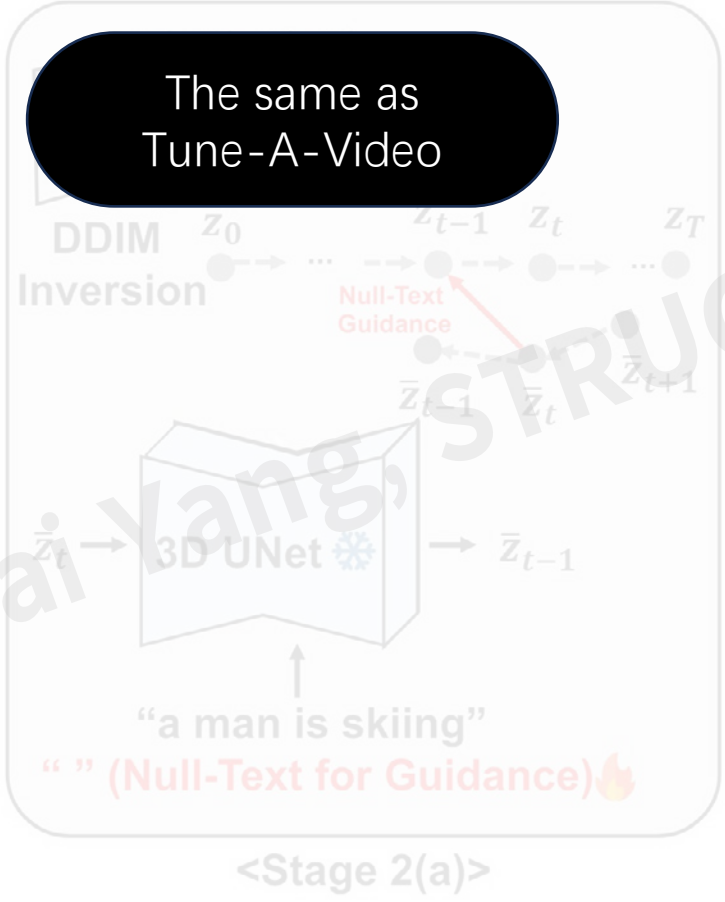
One-shot model

2 Edit-A-Video



<Stage 1>

The same as Tune-A-Video



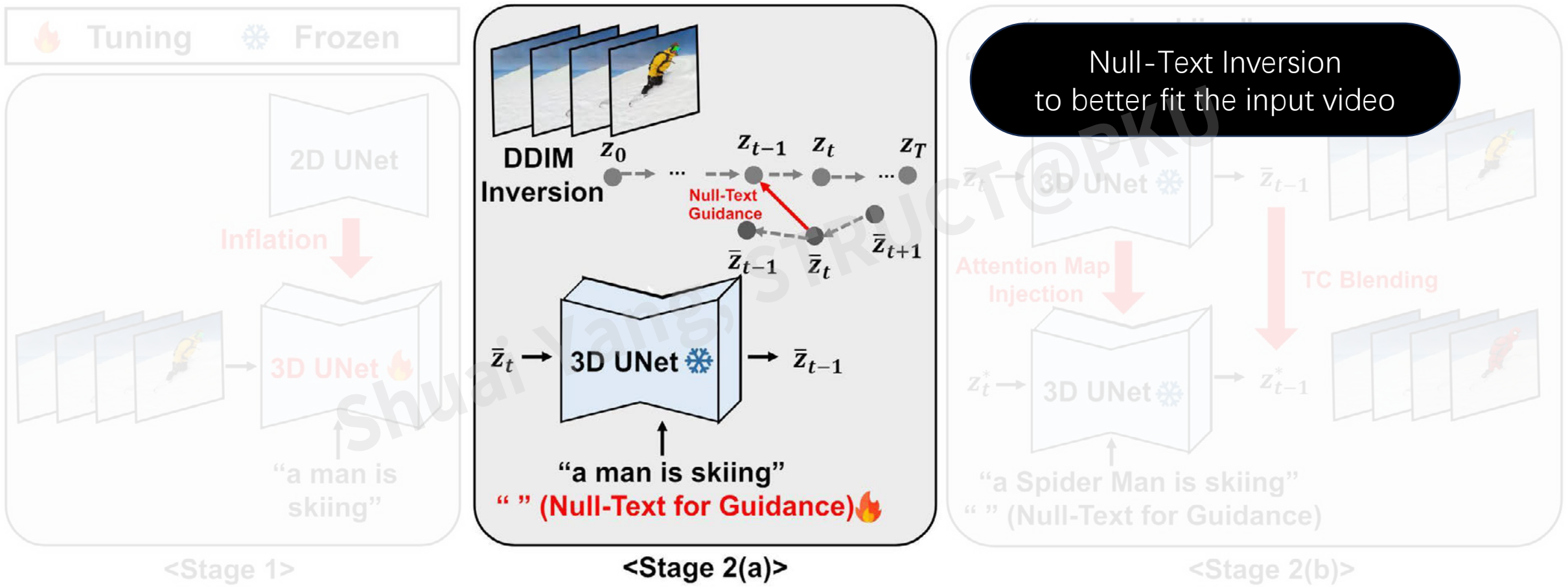
<Stage 2(a)>

<Stage 2(b)>

Edit-A-Video: Single Video Editing with Object-Aware Consistency. ACML'23

One-shot model

2 Edit-A-Video



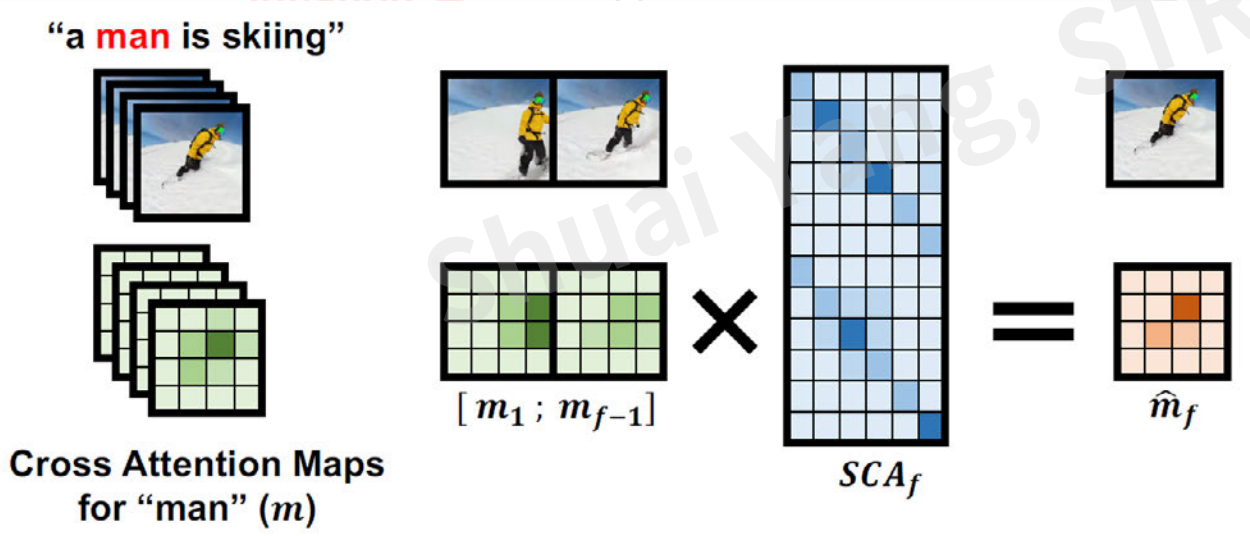
Edit-A-Video: Single Video Editing with Object-Aware Consistency. ACML'23

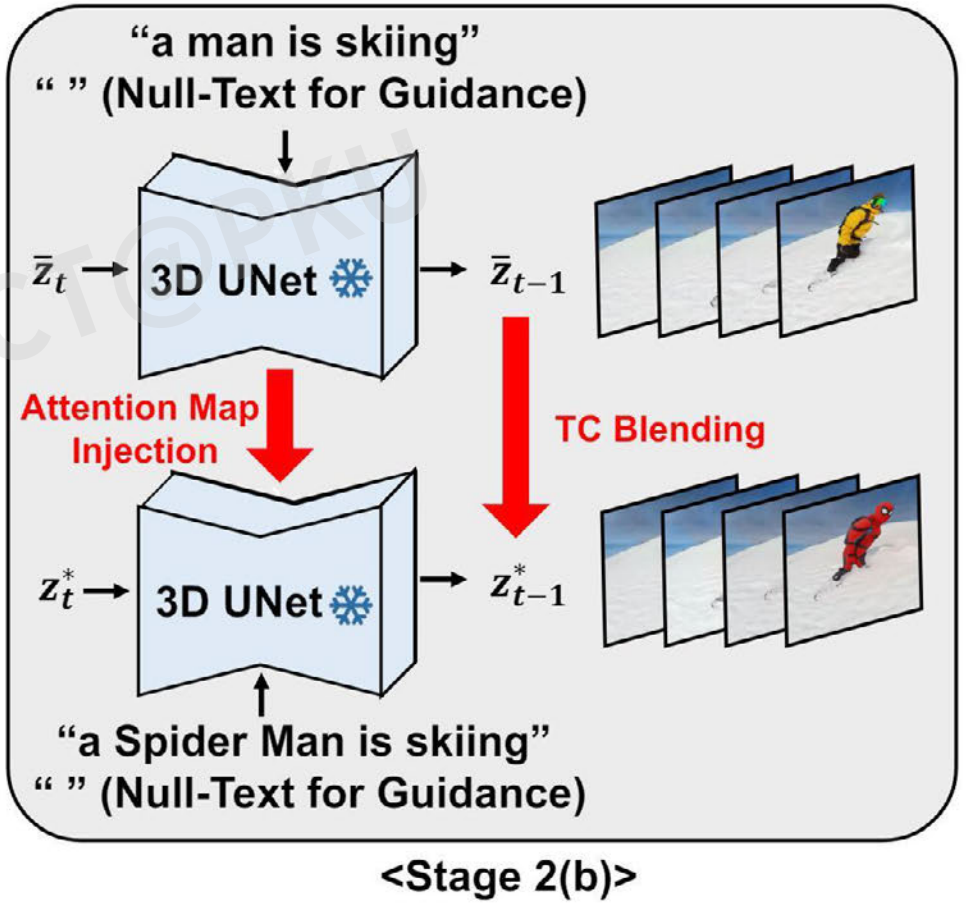
One-shot model

2 Edit-A-Video

Modified Prompt-to-Prompt

Background mask should be also temporal consistent



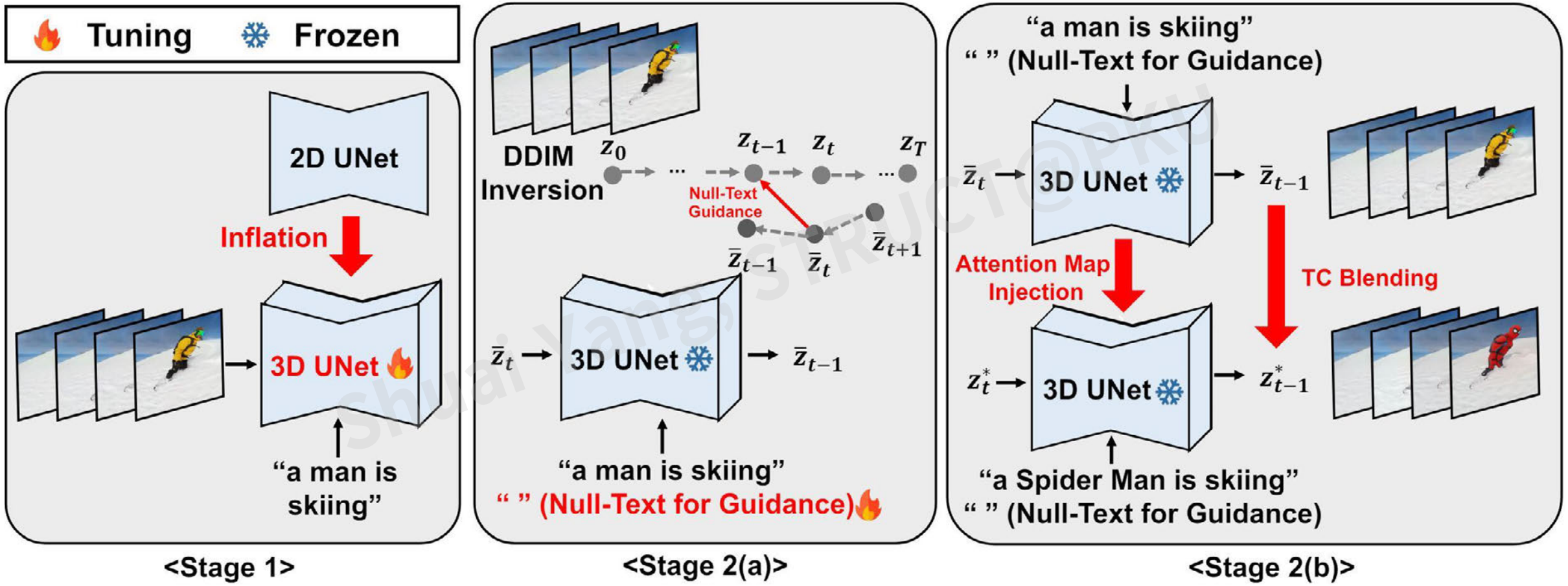


<Stage 2(b)>

Edit-A-Video: Single Video Editing with Object-Aware Consistency. ACML'23

One-shot model

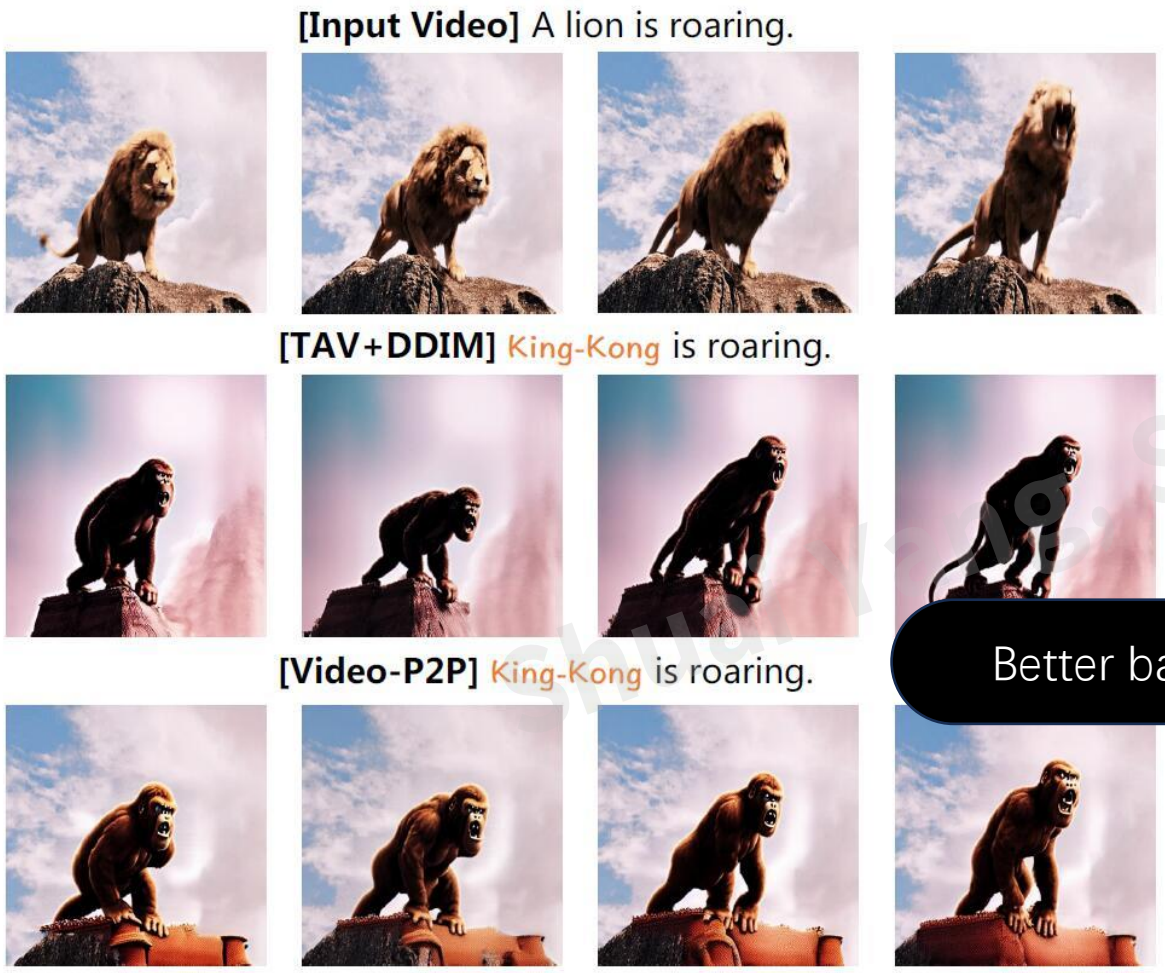
2 Edit-A-Video

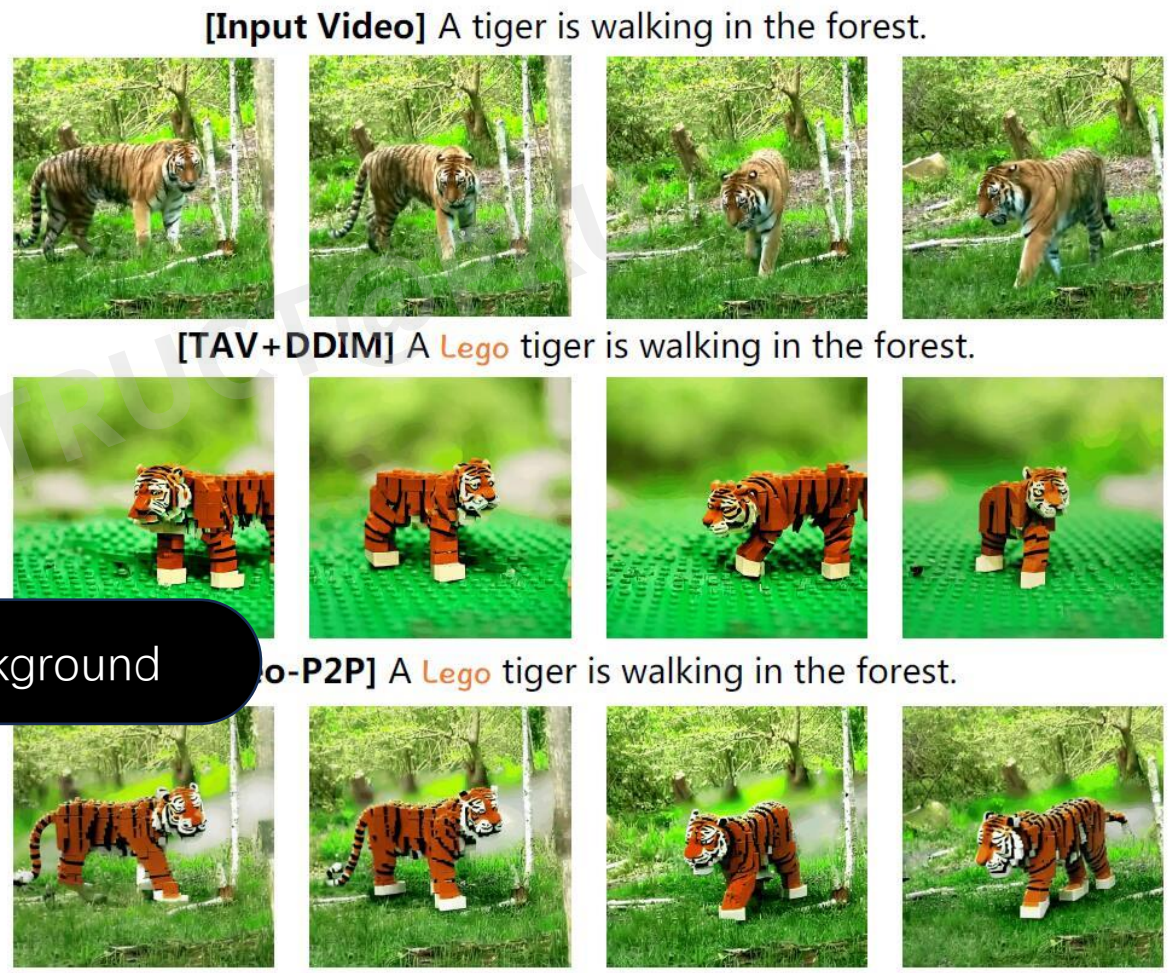


Edit-A-Video: Single Video Editing with Object-Aware Consistency. ACML'23

One-shot model

3 Video-P2P



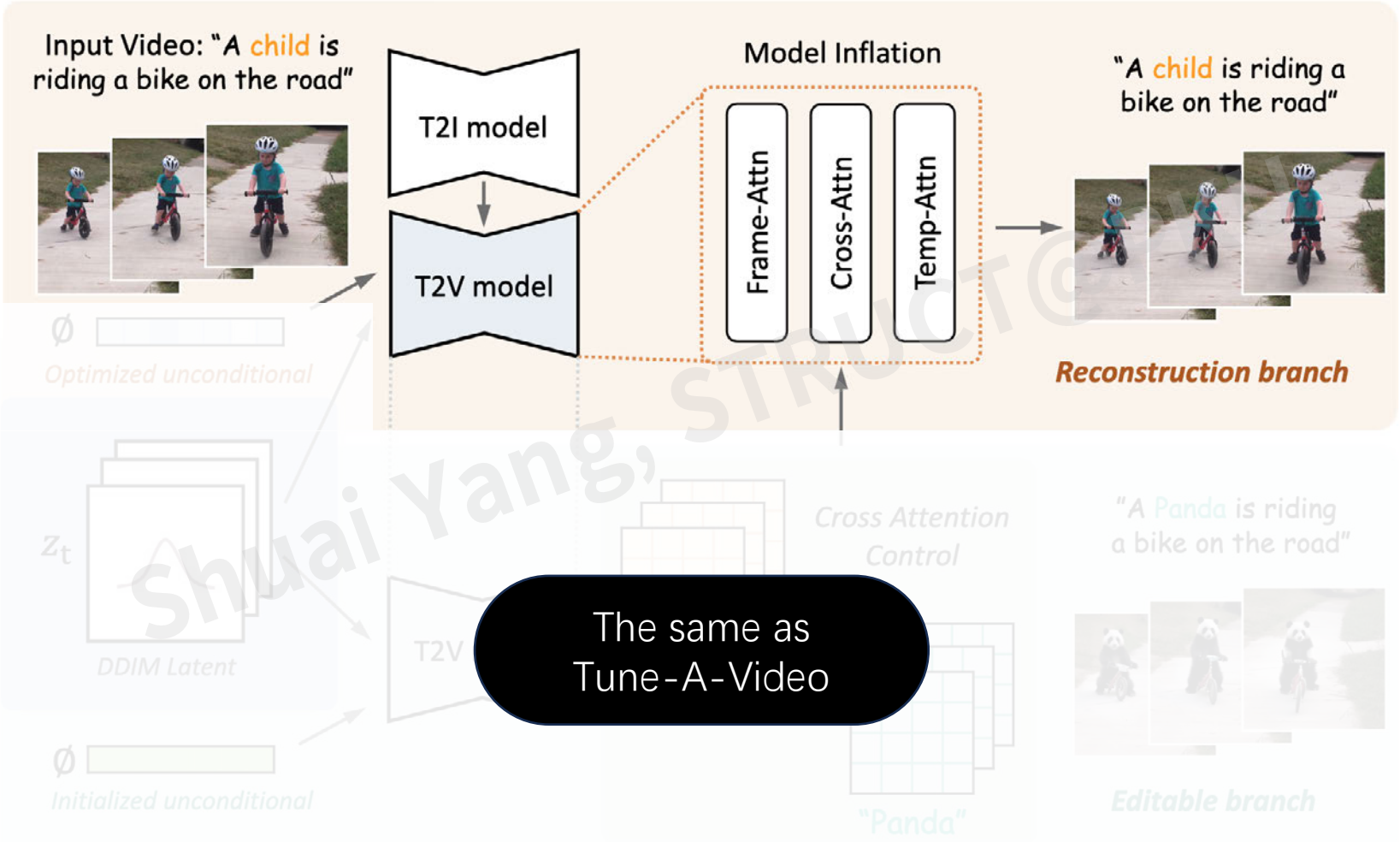


Better background

Video-P2P: Video Editing with Cross-attention Control. CVPR'24

One-shot model

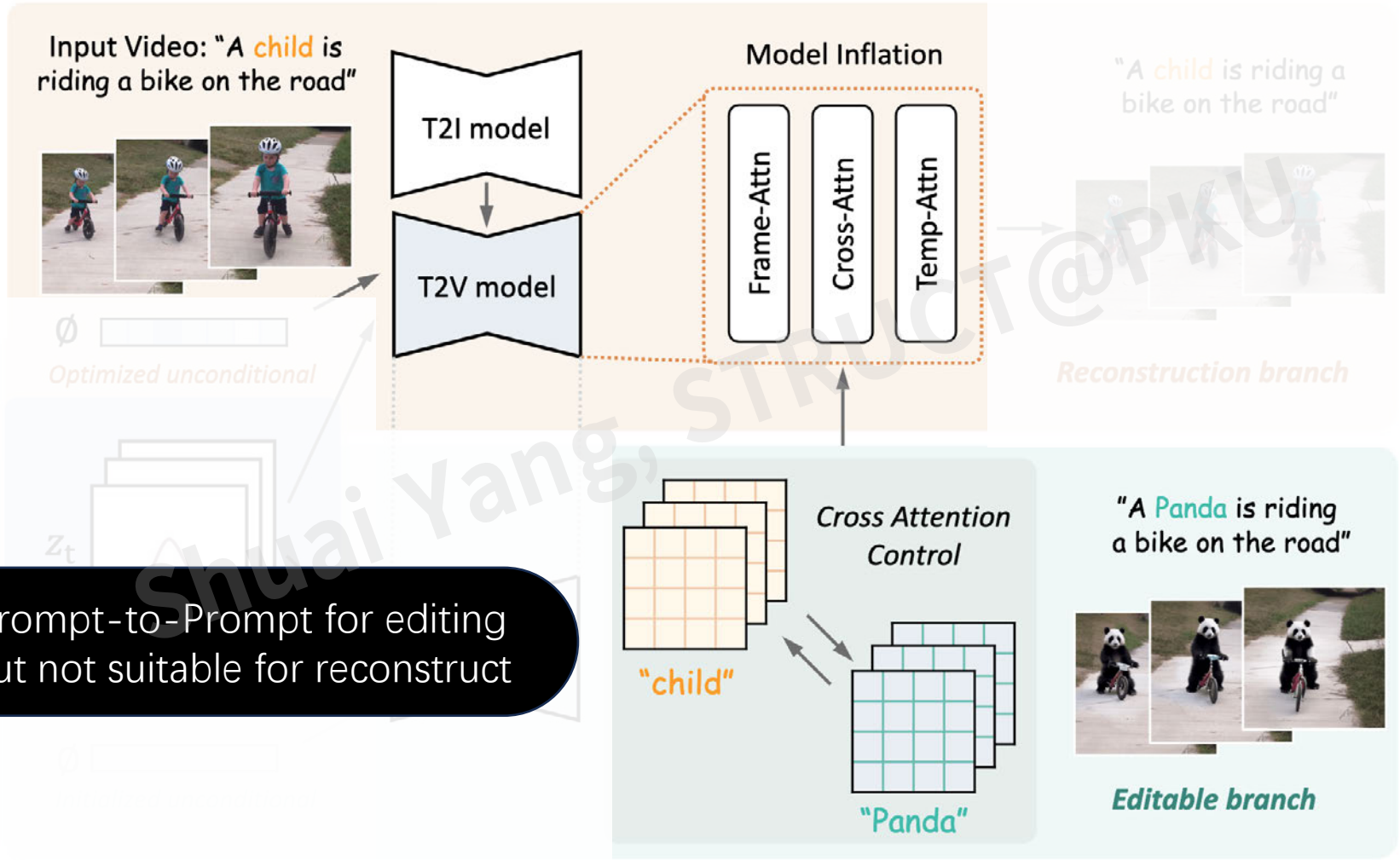
3 Video-P2P



Video-P2P: Video Editing with Cross-attention Control. CVPR'24

One-shot model

3 Video-P2P

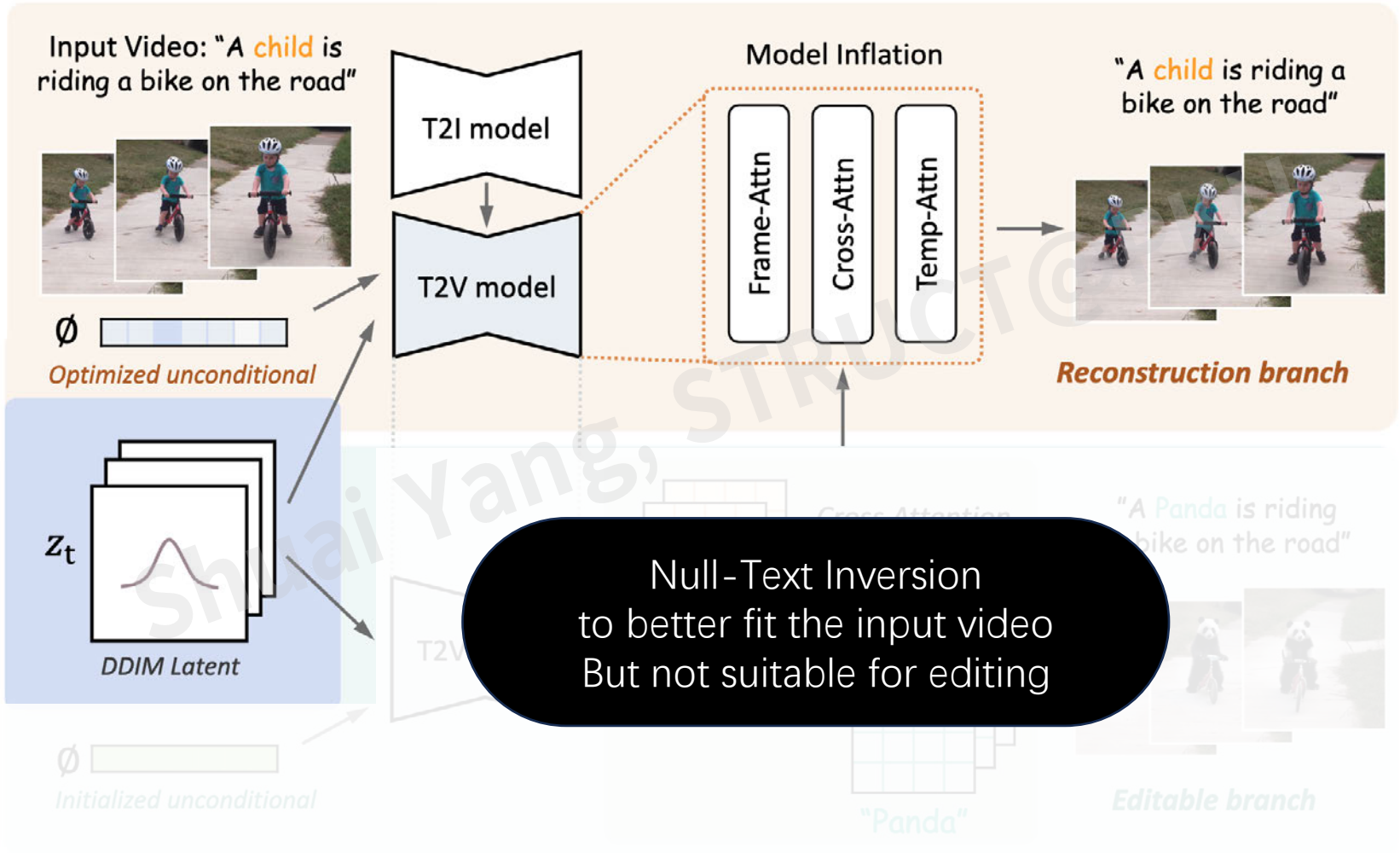


Prompt-to-Prompt for editing
But not suitable for reconstruct

Video-P2P: Video Editing with Cross-attention Control. CVPR'24

One-shot model

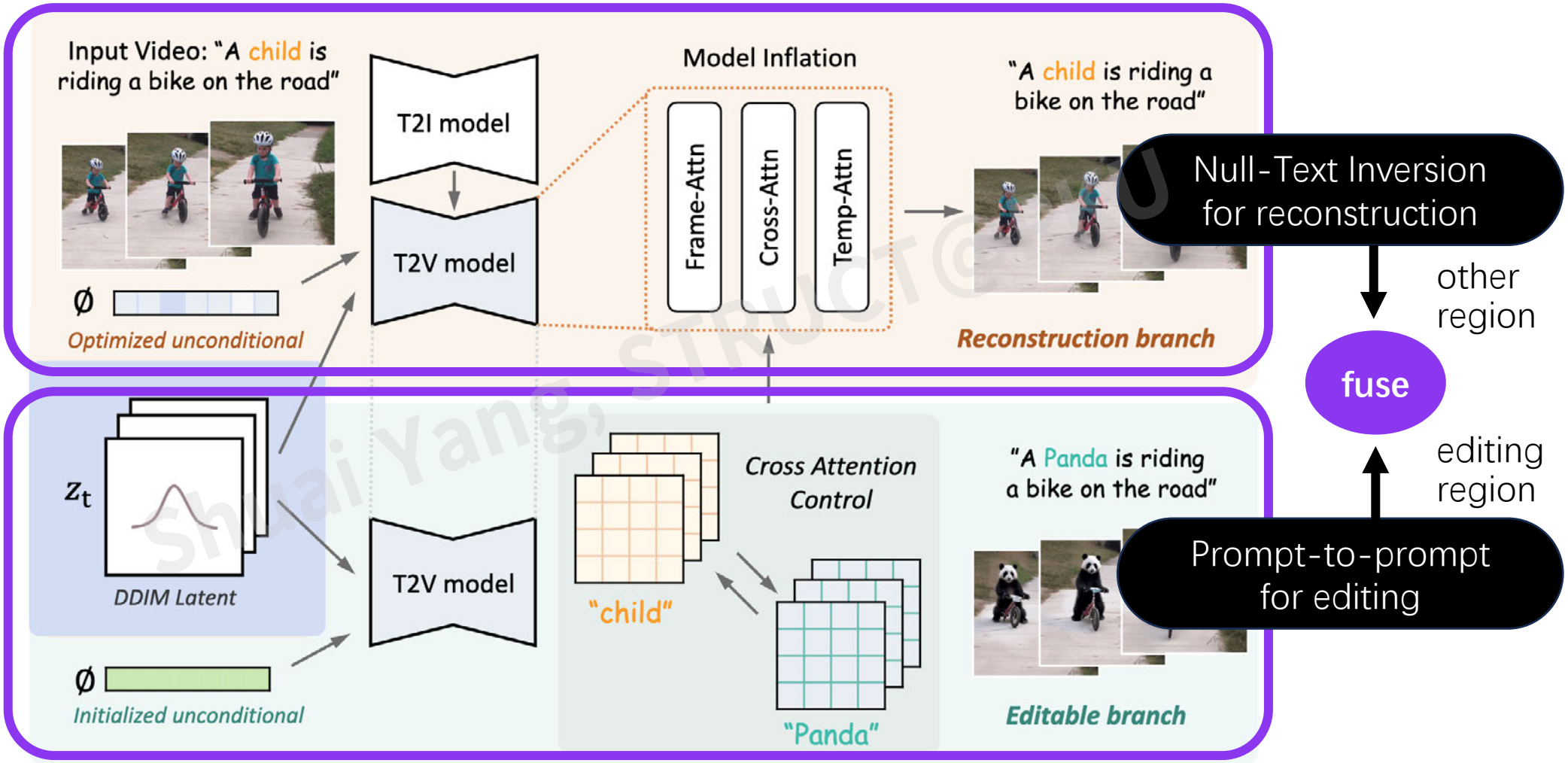
3 Video-P2P



Video-P2P: Video Editing with Cross-attention Control. CVPR'24

One-shot model

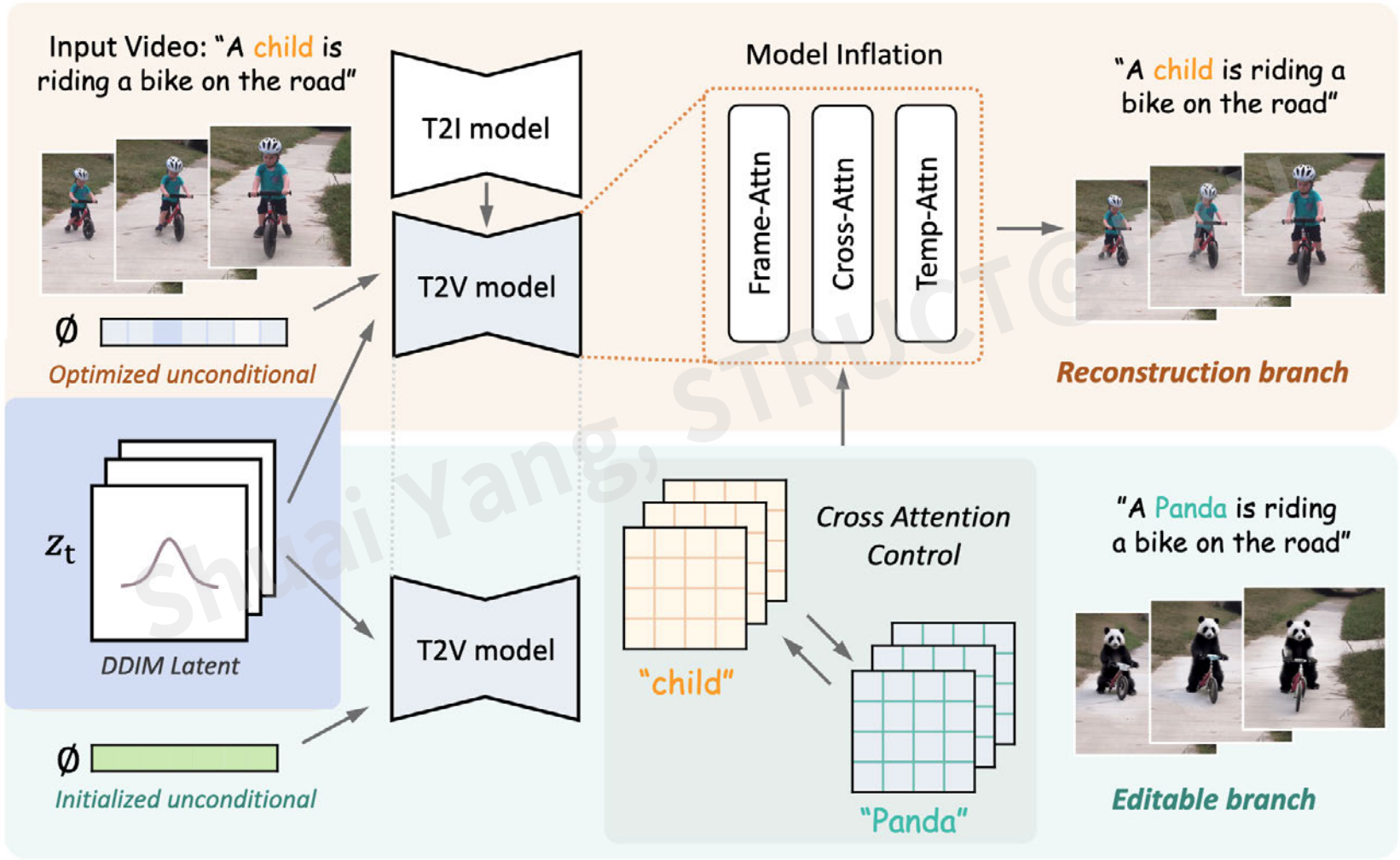
3 Video-P2P



Video-P2P: Video Editing with Cross-attention Control. CVPR'24

One-shot model

3 Video-P2P



Video-P2P: Video Editing with Cross-attention Control. CVPR'24

One-shot model

3 Video-P2P

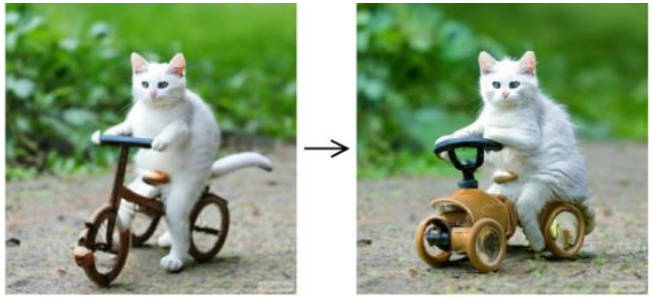
a **crochet** penguin is running on the ice

a jeep car is moving on the ~~road~~ **beach**

Video-P2P: Video Editing with Cross-attention Control. CVPR'24

Zero-shot model

Image Editing



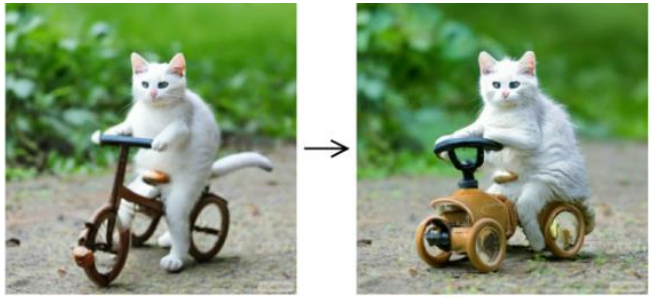
"Photo of a cat riding on a bicycle."

car

- Prompt-to-Prompt
- Null-Text Inversion
- Plug-and-Play

Zero-shot model

Image Editing



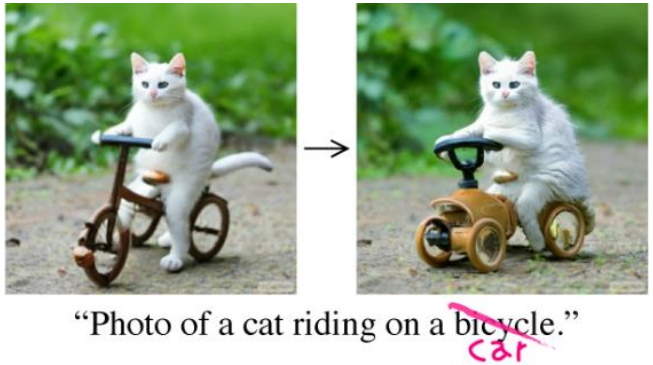
Prompt-to-Prompt
Null-Text Inversion
Plug-and-Play

Inversion-Based Model

ICCV'23 FateZero Pixel2Video
ICLR'24 TokenFlow FLATTEN
CVPR'24 VidToMe

Zero-shot model

Image Editing



Prompt-to-Prompt
Null-Text Inversion
Plug-and-Play

Inversion-Based Model

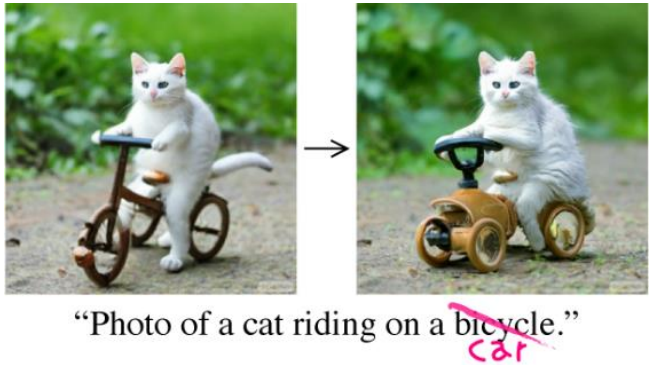
ICCV'23 FateZero Pixel2Video
ICLR'24 TokenFlow FLATTEN
CVPR'24 VidToMe

Conditional Generation

ControlNet
SDEdit

Zero-shot model

Image Editing



Prompt-to-Prompt
Null-Text Inversion
Plug-and-Play

Inversion-Based Model

ICCV'23 FateZero Pixel2Video
ICLR'24 TokenFlow FLATTEN
CVPR'24 VidToMe

Conditional Generation

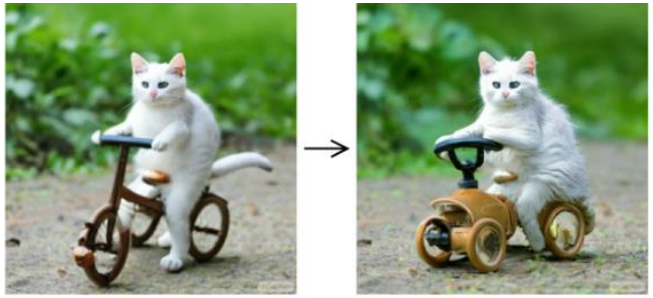
ControlNet
SDEdit

Inversion-Free Model

ICCV'23 Text2Video-Zero
SIGGRAPH Asia'23 Rerender-A-Video
ICLR'24 ControlVideo
CVPR'24 FRESCO Fairy

Zero-shot model

Image Editing



Prompt-to-Prompt
Null-Text Inversion
Plug-and-Play

Inversion-Based Model

ICCV'23 FateZero Pixel2Video
ICLR'24 TokenFlow FLATTEN
CVPR'24 VidToMe

Conditional Generation

ControlNet
SDEdit

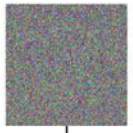
Inversion-Free Model

ICCV'23 Text2Video-Zero
SIGGRAPH Asia'23 Rerender-A-Video
ICLR'24 ControlVideo
CVPR'24 FRESCO Fairy

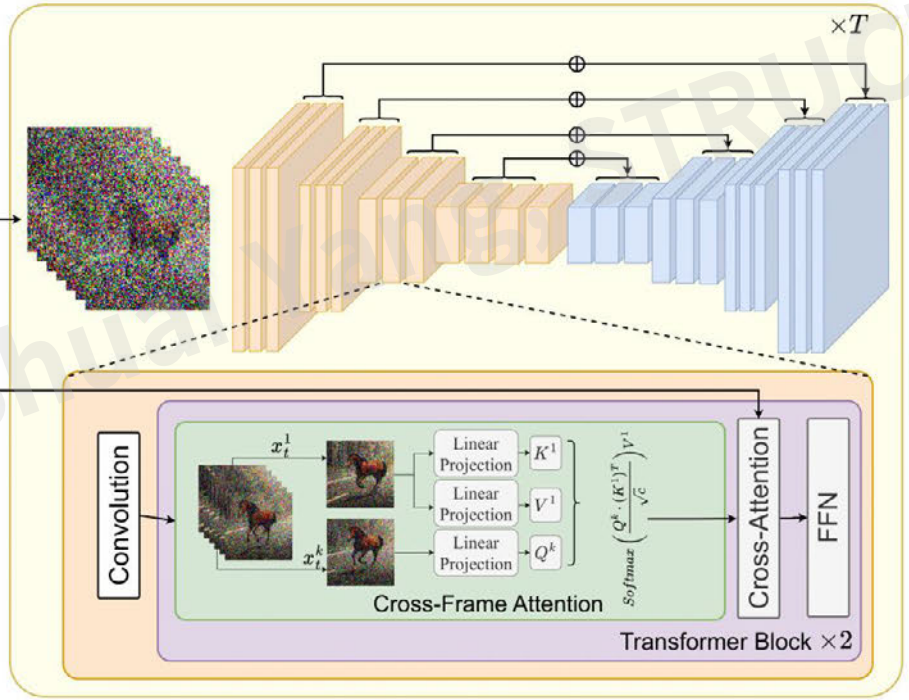
Inversion-free zero-shot model

1 Text2Video-Zero

cross-frame attention
Not enough due to random noises



$x_T^1 \sim \mathcal{N}(0, I)$



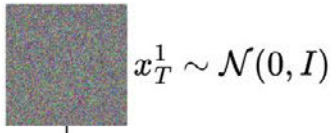
Text prompt: "A horse is galloping on the street"

Text2Video-Zero: Text-to-Image Diffusion Models are Zero-Shot Video Generators. ICCV'23

Inversion-free zero-shot model

1 Text2Video-Zero

Use the same noises with translation



$x_T^1 \sim \mathcal{N}(0, I)$

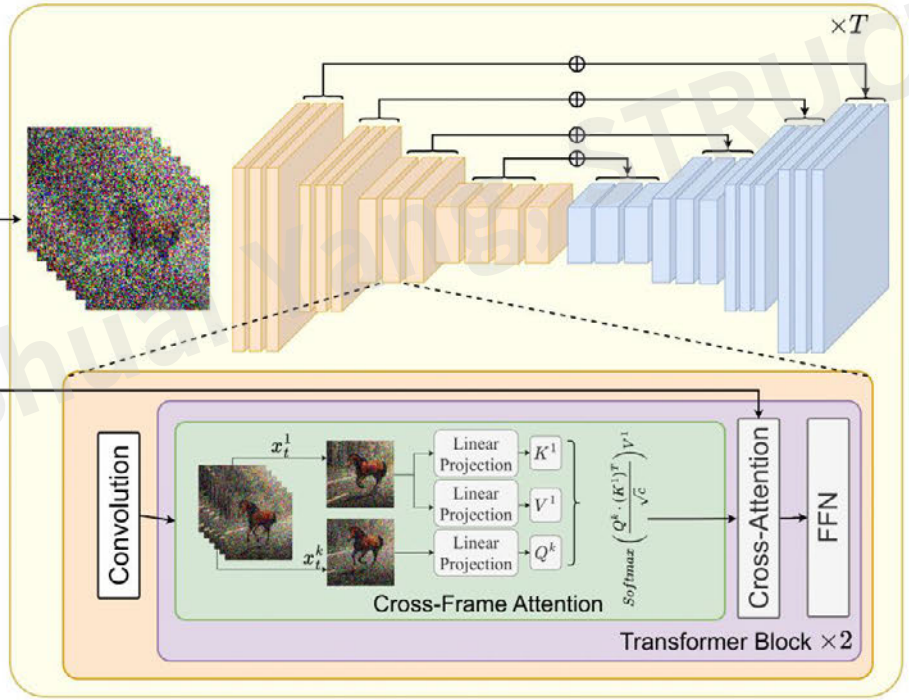
$x_{T'}^1 = \text{DDIM\_Backward}(x_T^1, \Delta t, SD)$

$\tilde{x}_{T'}^k = W_k(x_{T'}^1)$

$x_T^k = \text{DDPM\_Forward}(\tilde{x}_T^k, \Delta t)$

for $k = 2, 3, \ldots, m$

Text prompt: "A horse is galloping on the street"

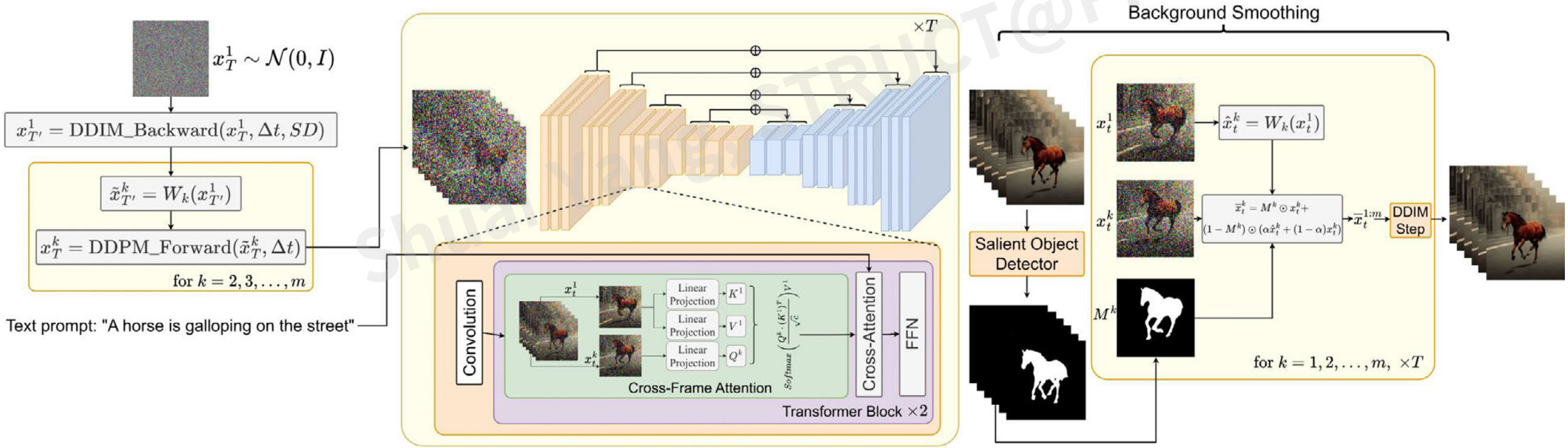


Text2Video-Zero: Text-to-Image Diffusion Models are Zero-Shot Video Generators. ICCV'23

Inversion-free zero-shot model

1 Text2Video-Zero

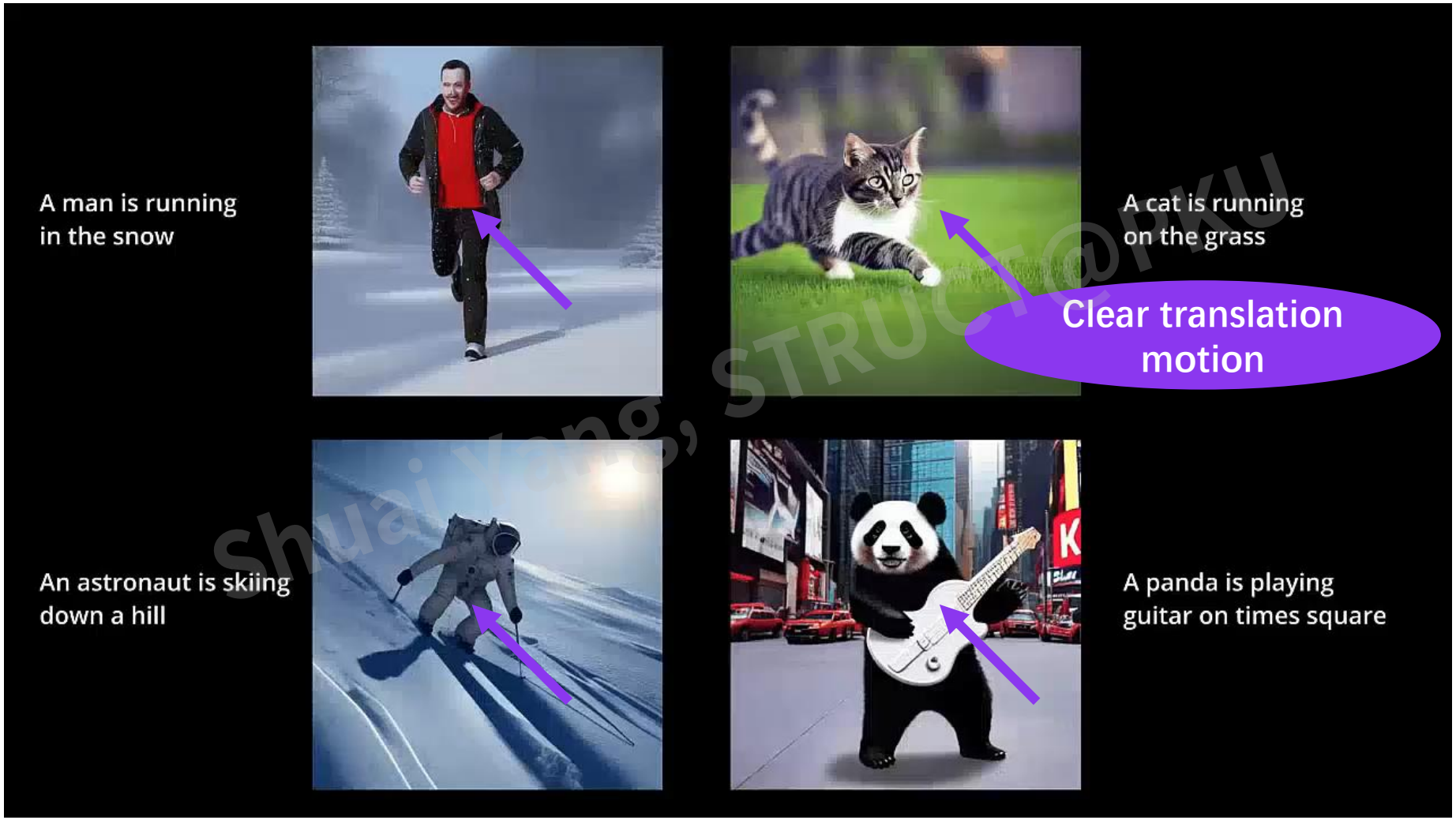
Background smooth
Keep the non-salient region the same



Text2Video-Zero: Text-to-Image Diffusion Models are Zero-Shot Video Generators. ICCV'23

Inversion-free zero-shot model

1 Text2Video-Zero



Text2Video-Zero: Text-to-Image Diffusion Models are Zero-Shot Video Generators. ICCV'23

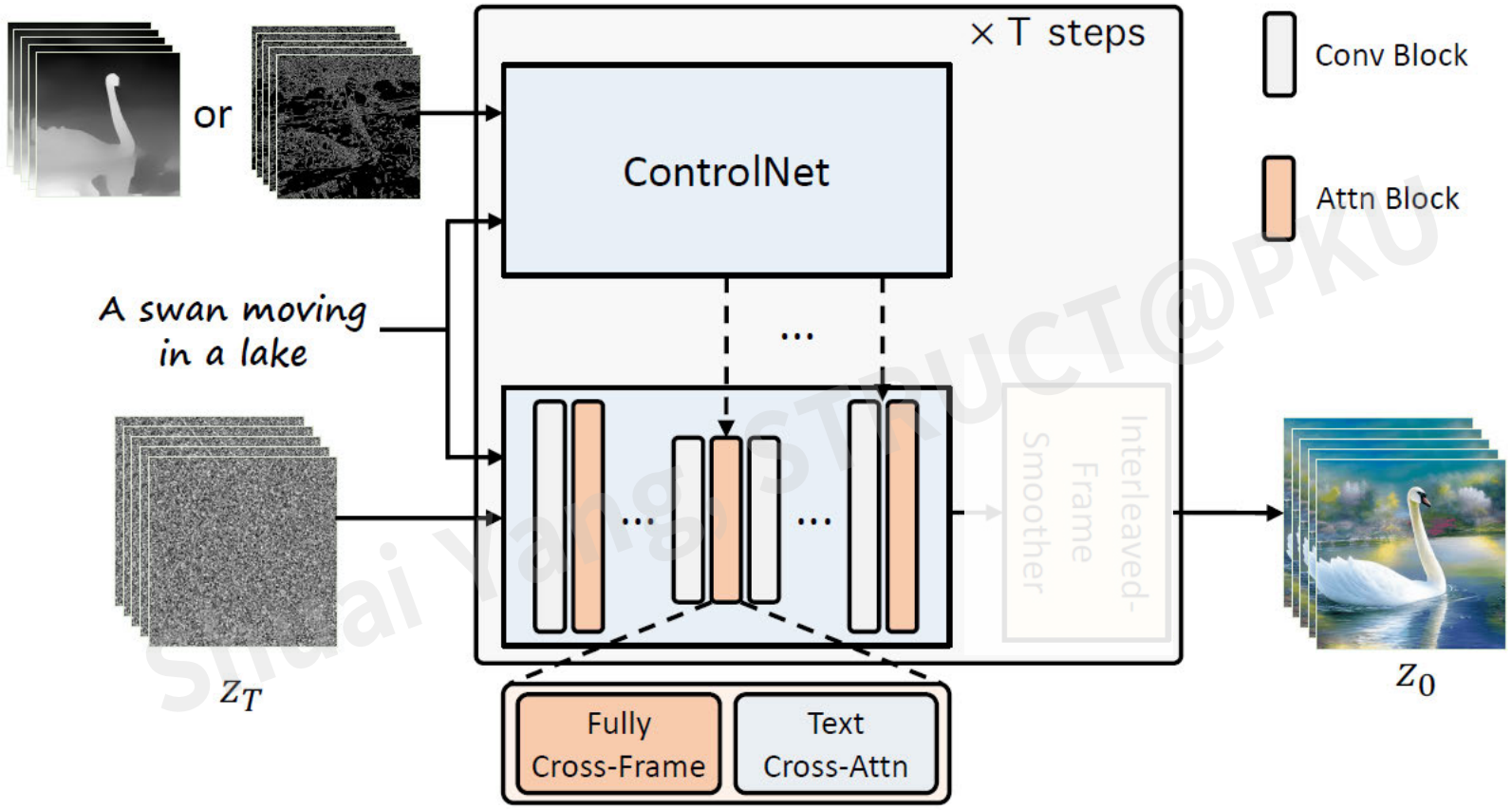
Inversion-free zero-shot model

1 Text2Video-Zero

Text2Video-Zero: Text-to-Image Diffusion Models are Zero-Shot Video Generators. ICCV'23

Inversion-free zero-shot model

2 ControlVideo

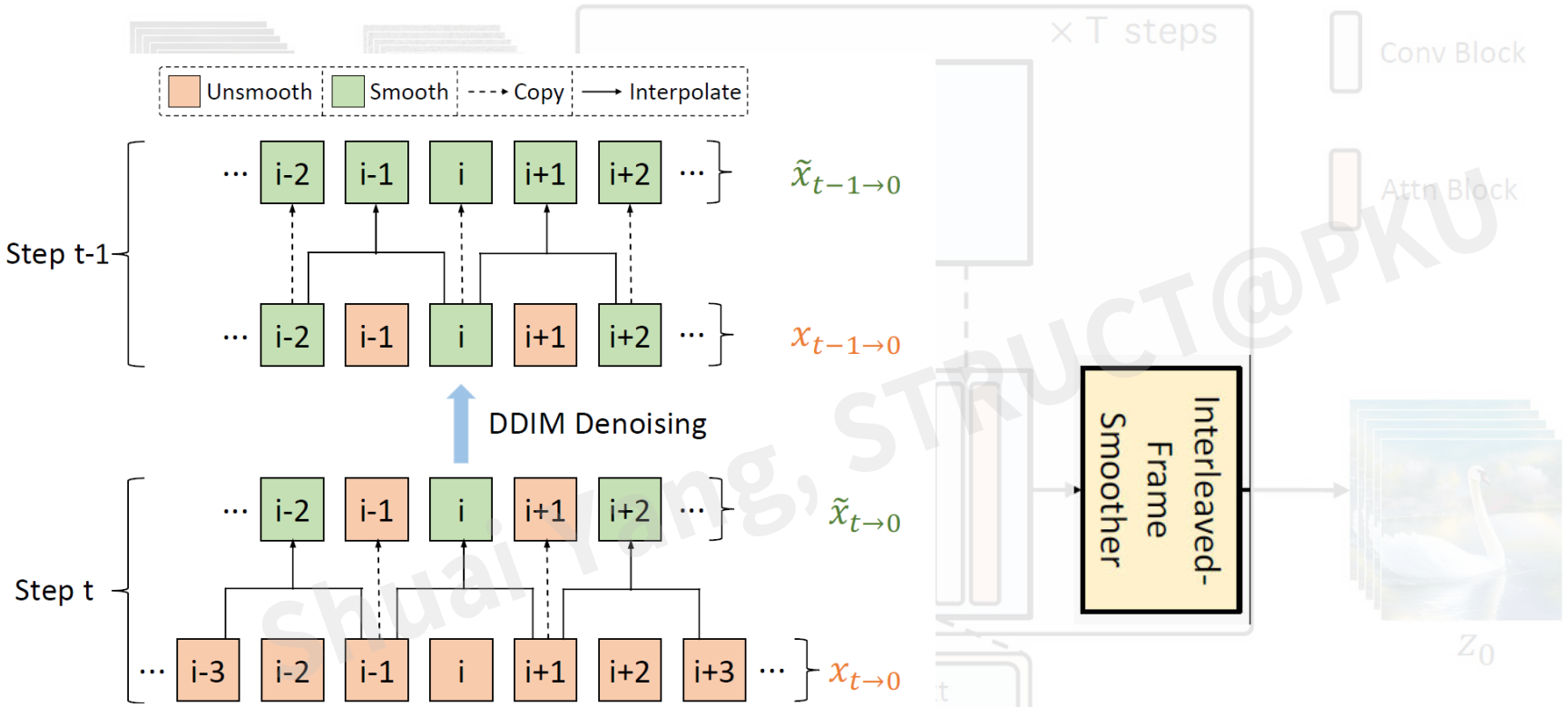


ControlNet with cross-frame attention

ControlVideo: Training-free Controllable Text-to-Video Generation. ICLR'24

Inversion-free zero-shot model

2 ControlVideo



Interleaved-frame smoother:
denoised frame interpolation

ControlVideo: Training-free Controllable Text-to-Video Generation. ICLR'24

Inversion-free zero-shot model

2 ControlVideo

still very flickering
Can we use stronger constraints, such as ··· optical flow?

Input

w/o smoother

full

ControlVideo: Training-free Controllable Text-to-Video Generation. ICLR'24

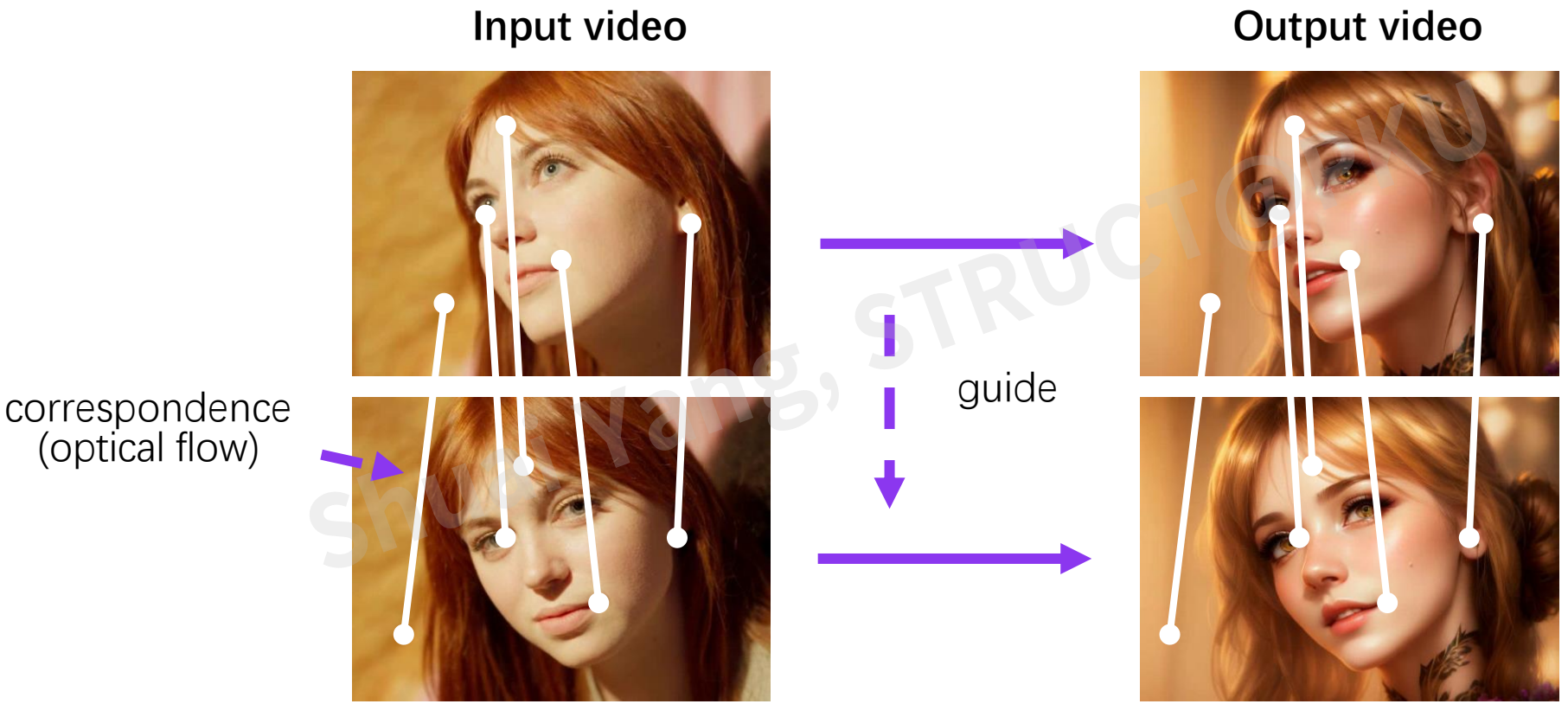
Inversion-free zero-shot model

3 Rerender-A-Video

Rerender A Video: Zero-Shot Text-Guided Video-to-Video Translation. SIGGRAPH Asia'23

Inversion-free zero-shot model

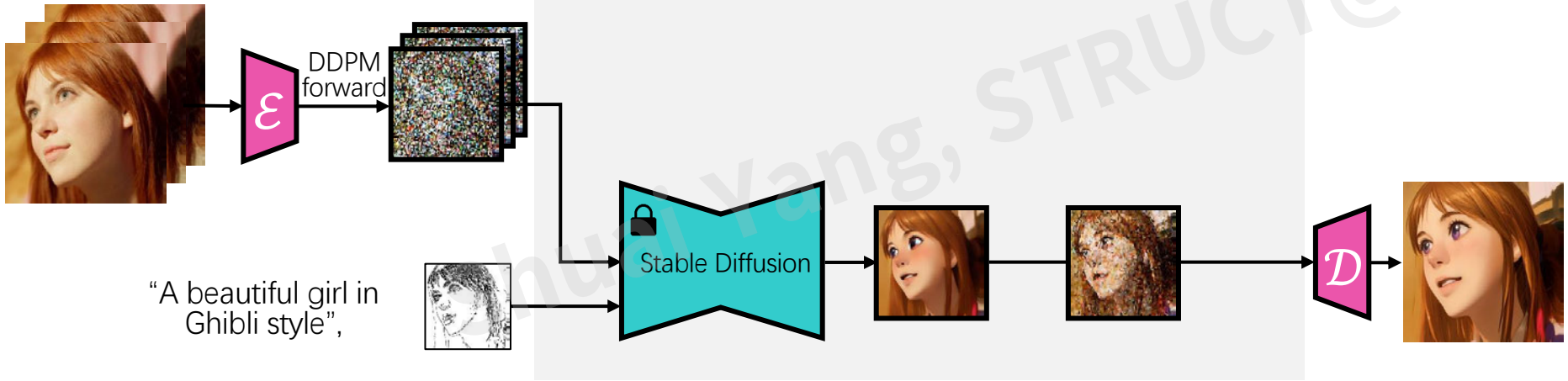
3 Rerender-A-Video



Rerender A Video: Zero-Shot Text-Guided Video-to-Video Translation. SIGGRAPH Asia'23

Inversion-free zero-shot model

3 Rerender-A-Video

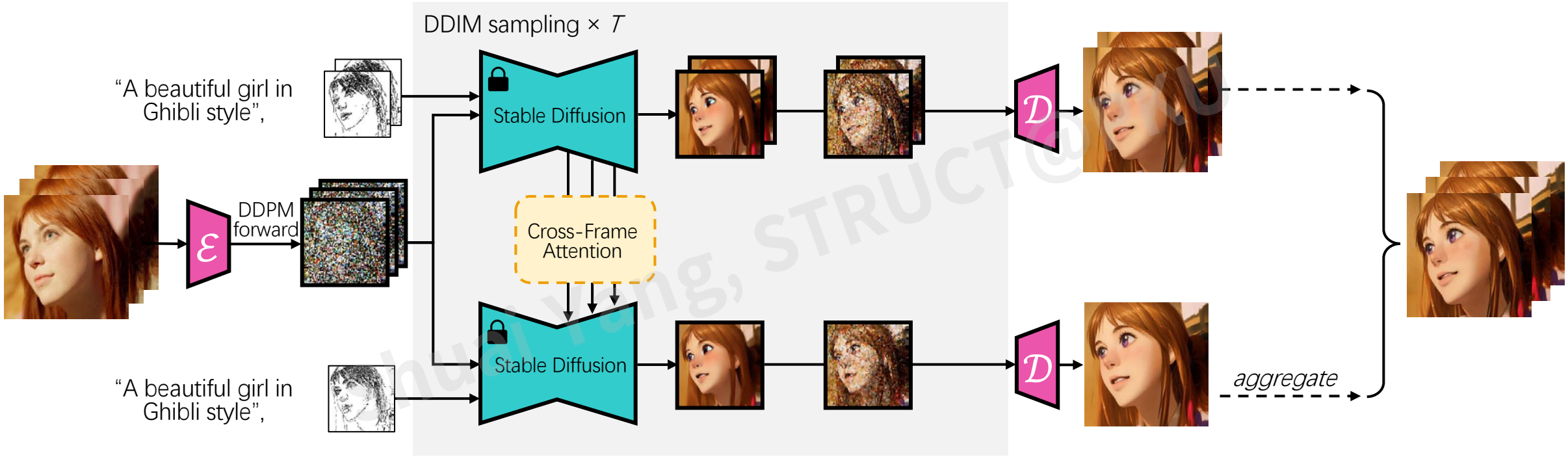


ControlNet+SDEdit image translation

Rerender A Video: Zero-Shot Text-Guided Video-to-Video Translation. SIGGRAPH Asia'23

Inversion-free zero-shot model

3 Rerender-A-Video

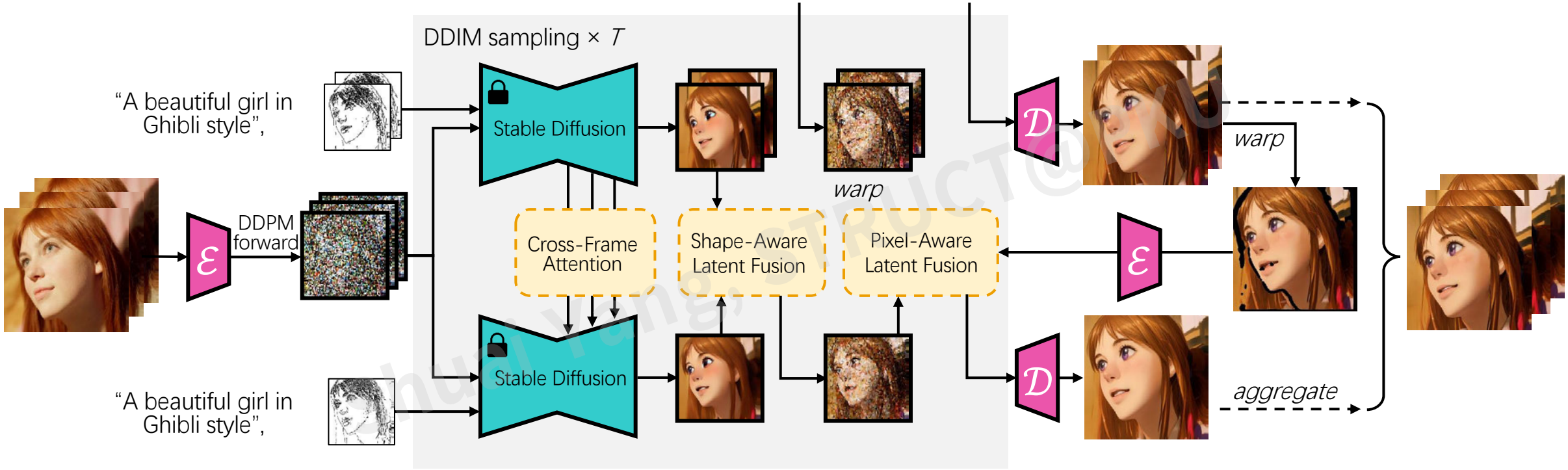


ControlNet+SDEdit image translation with cross-frame attention

Rerender A Video: Zero-Shot Text-Guided Video-to-Video Translation. SIGGRAPH Asia'23

Inversion-free zero-shot model

3 Rerender-A-Video

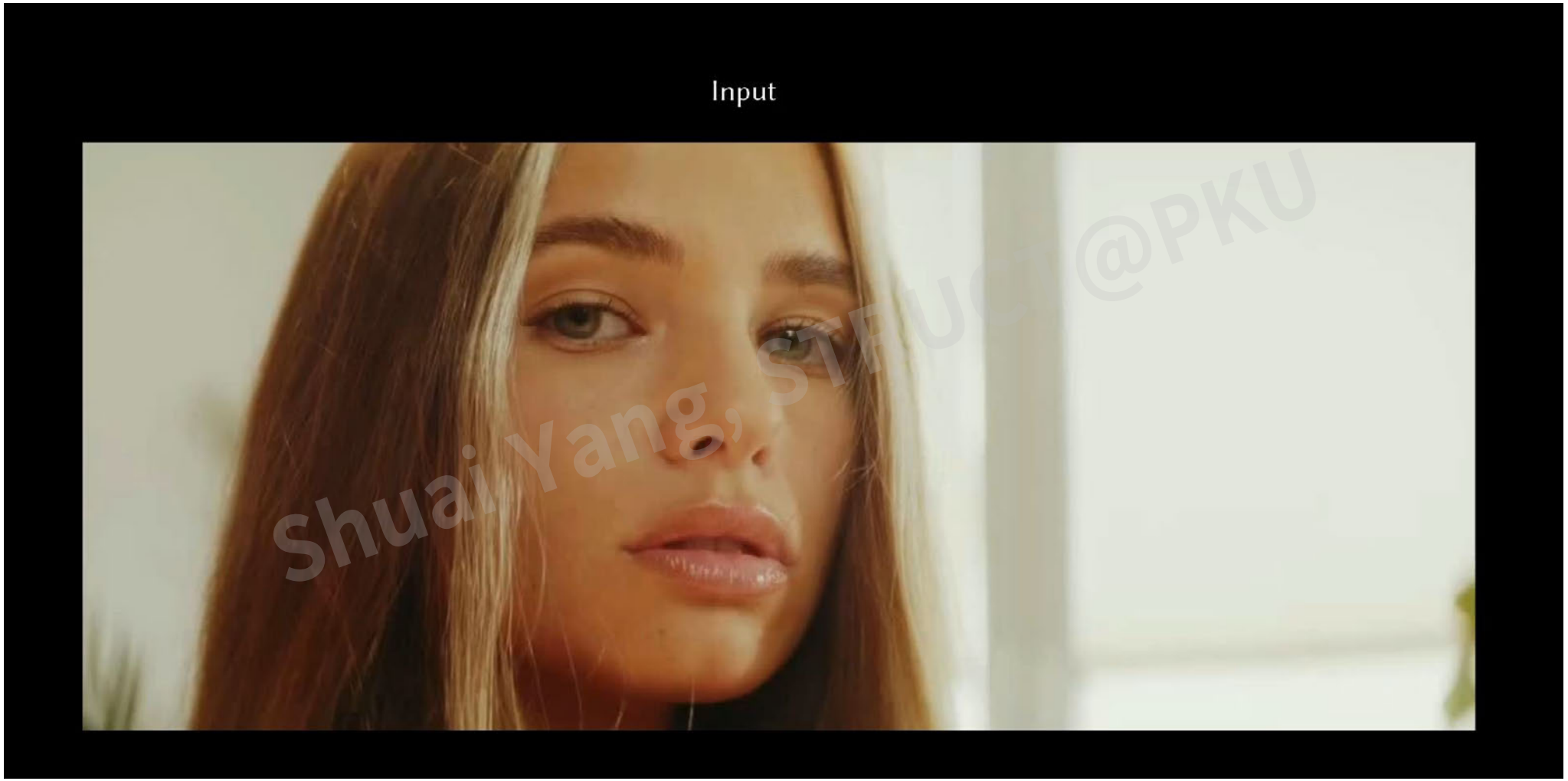


warp feature and image and fuse them

Rerender A Video: Zero-Shot Text-Guided Video-to-Video Translation. SIGGRAPH Asia'23

Inversion-free zero-shot model

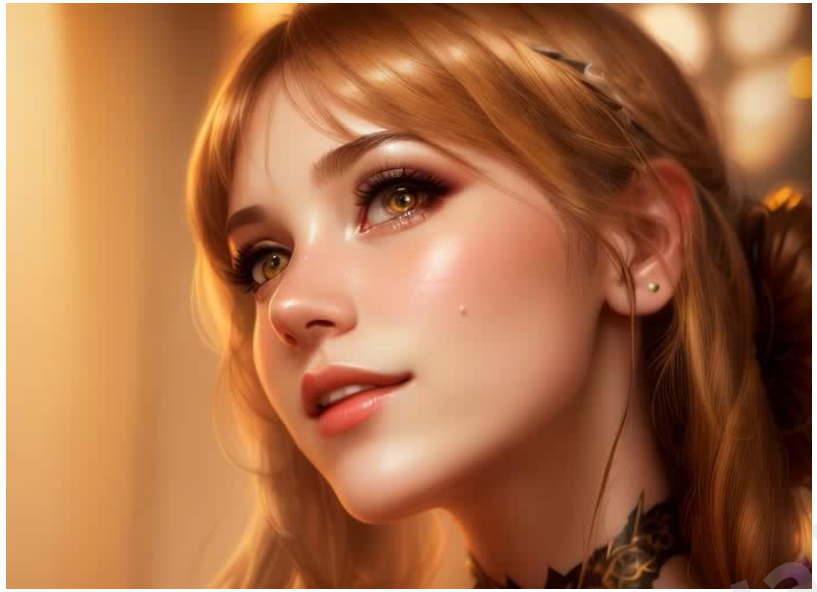
3 Rerender-A-Video



Rerender A Video: Zero-Shot Text-Guided Video-to-Video Translation. SIGGRAPH Asia'23

Inversion-free zero-shot model

3 Rerender-A-Video



Stable Diffusion

+ Cross-frame attn

+ Optical flow

Optical flow helps a lot
but when optical flow is hard to predict, what can we do?

Rerender A Video: Zero-Shot Text-Guided Video-to-Video Translation. SIGGRAPH Asia'23

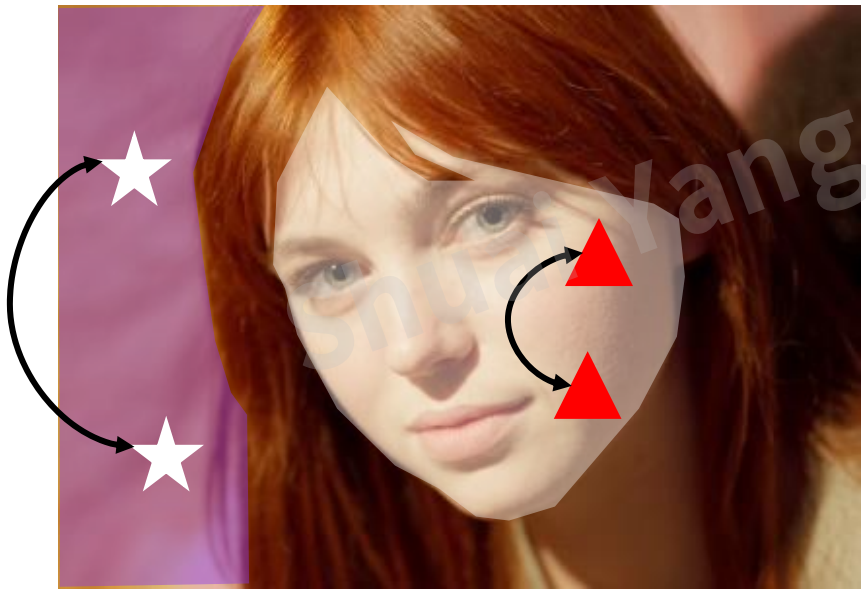
Inversion-free zero-shot model

4 FRESCO

While previous methods primarily focus on constraining inter-frame temporal correspondence,

preserving intra-frame spatial correspondence is equally crucial

semantically similar content is manipulated cohesively,
maintaining its similarity post-translation

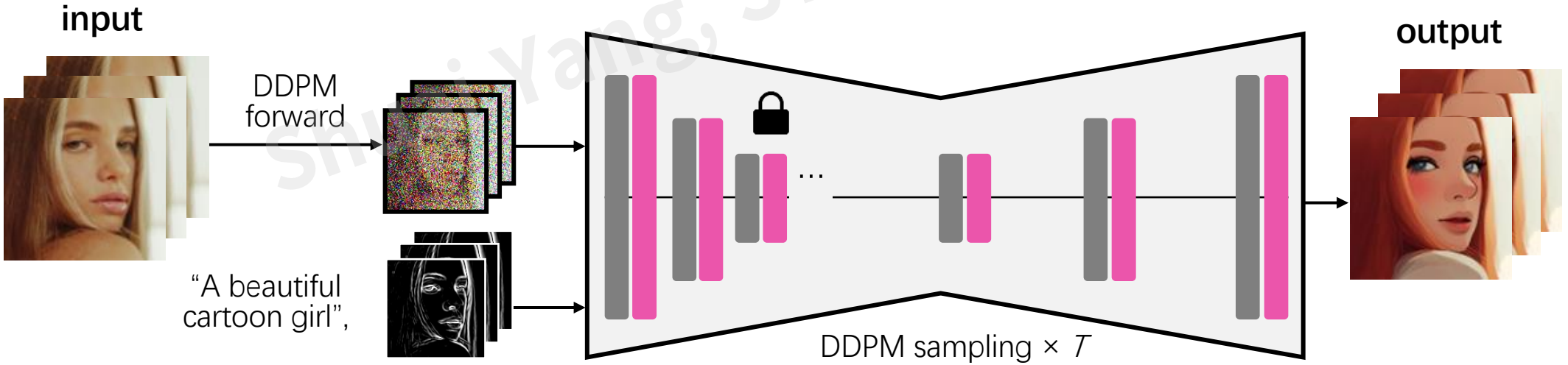


FRESCO: Spatial-Temporal Correspondence for Zero-Shot Video Translation. CVPR'24

Inversion-free zero-shot model

4 FRESCO

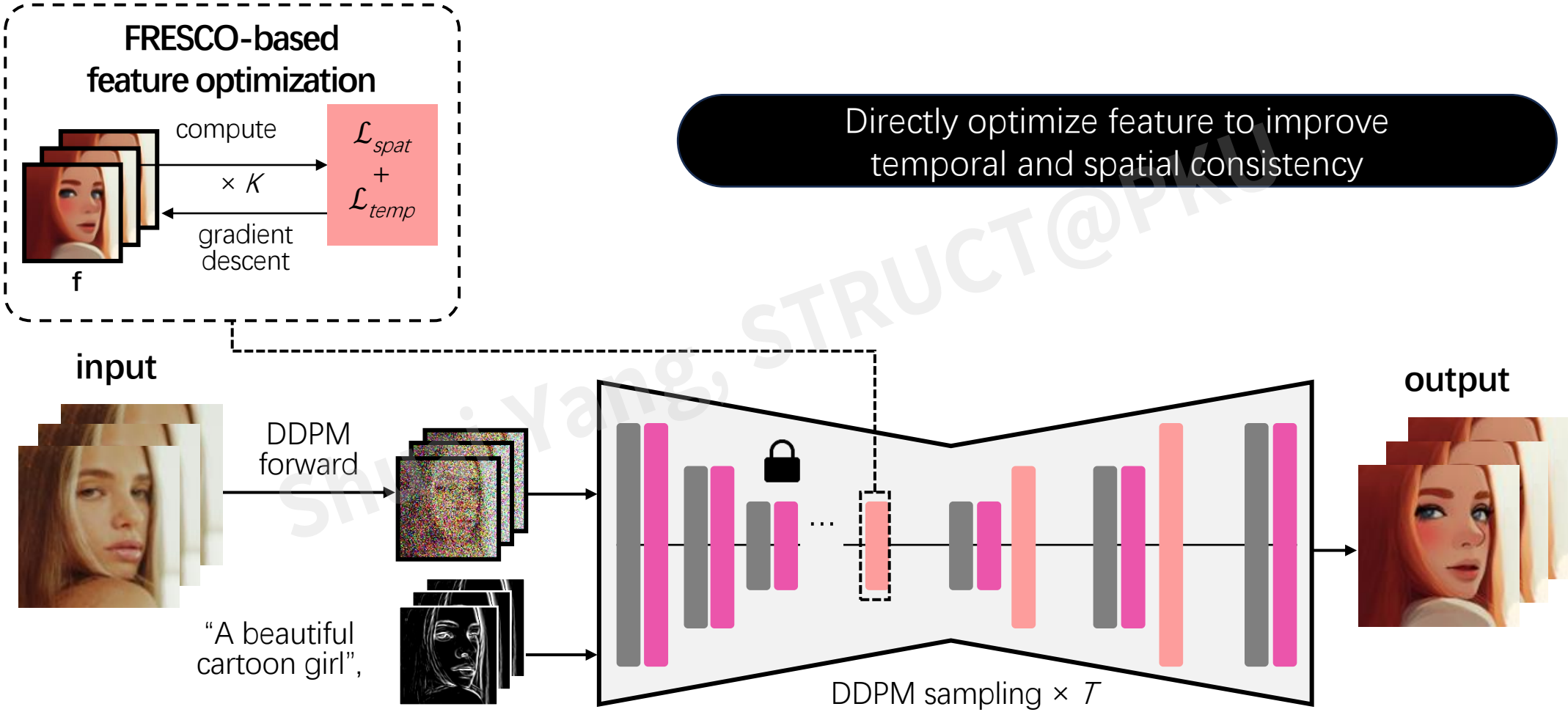
ControlNet+SDEdit image translation



FRESCO: Spatial-Temporal Correspondence for Zero-Shot Video Translation. CVPR'24

Inversion-free zero-shot model

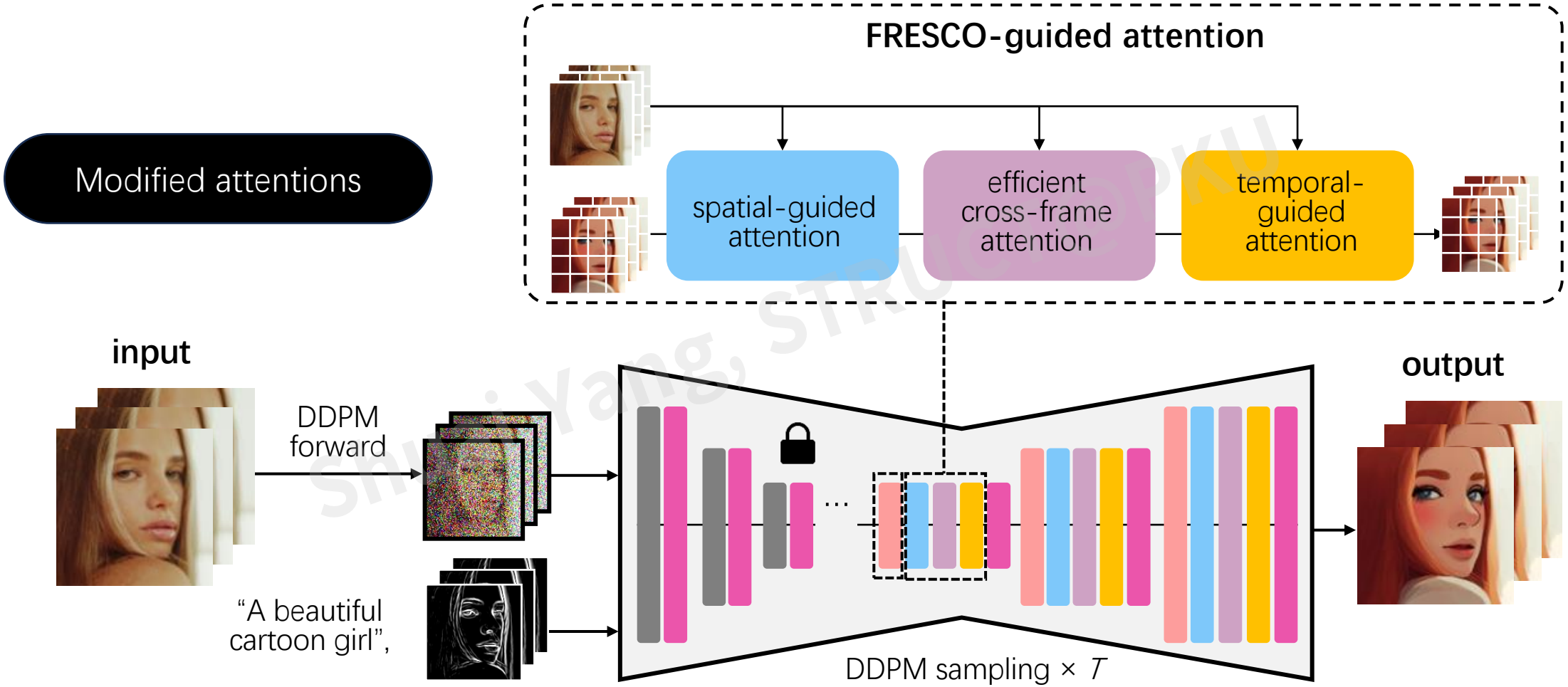
4 FRESCO



FRESCO: Spatial-Temporal Correspondence for Zero-Shot Video Translation. CVPR'24

Inversion-free zero-shot model

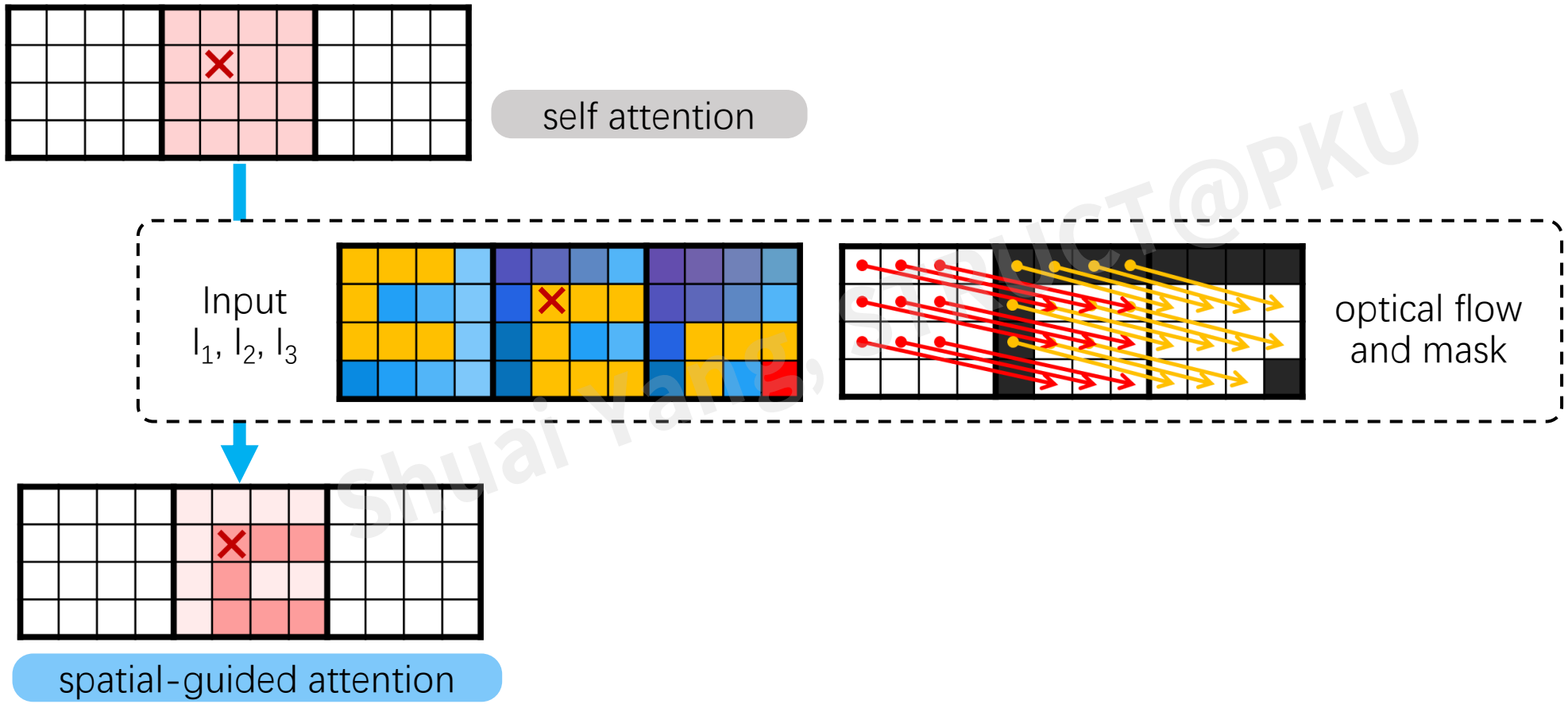
4 FRESCO



FRESCO: Spatial-Temporal Correspondence for Zero-Shot Video Translation. CVPR'24

Inversion-free zero-shot model

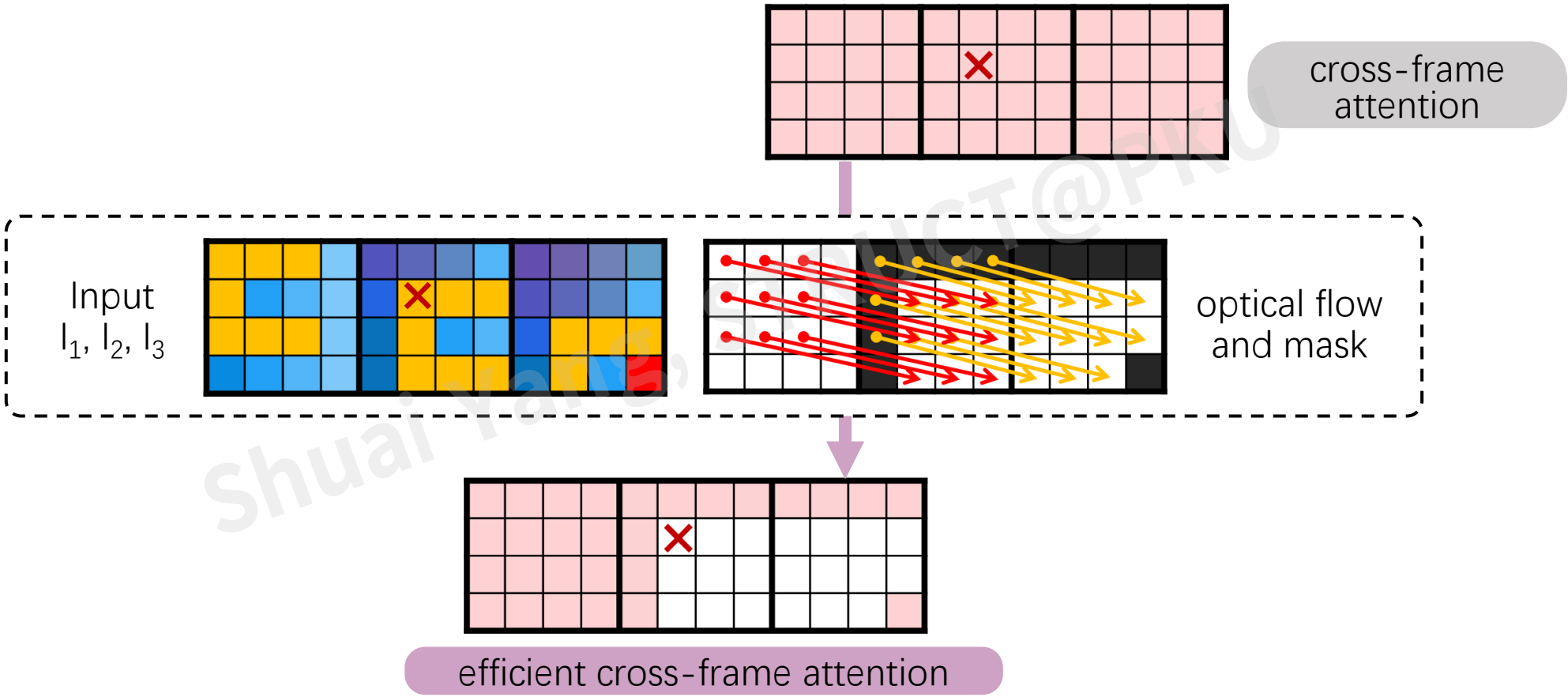
4 FRESCO



FRESCO: Spatial-Temporal Correspondence for Zero-Shot Video Translation. CVPR'24

Inversion-free zero-shot model

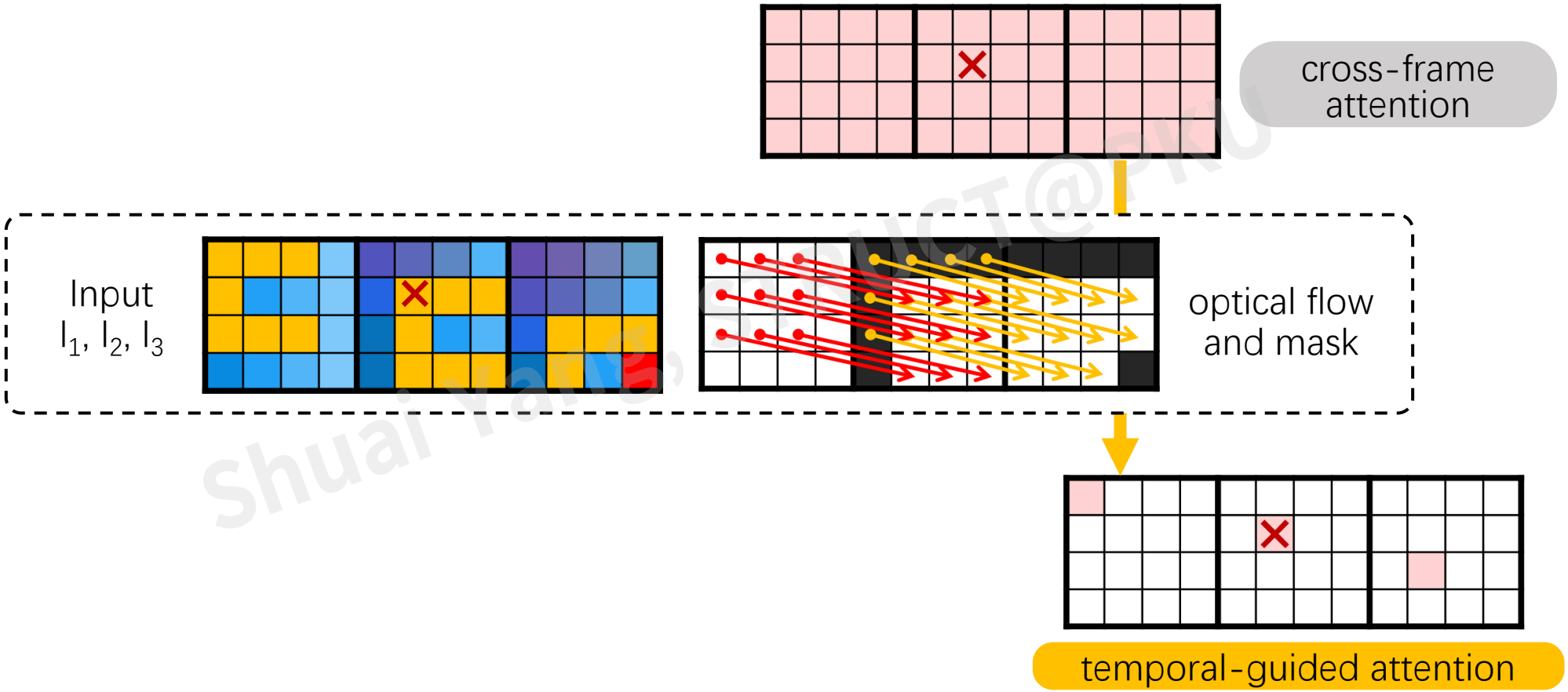
4 FRESCO



FRESCO: Spatial-Temporal Correspondence for Zero-Shot Video Translation. CVPR'24

Inversion-free zero-shot model

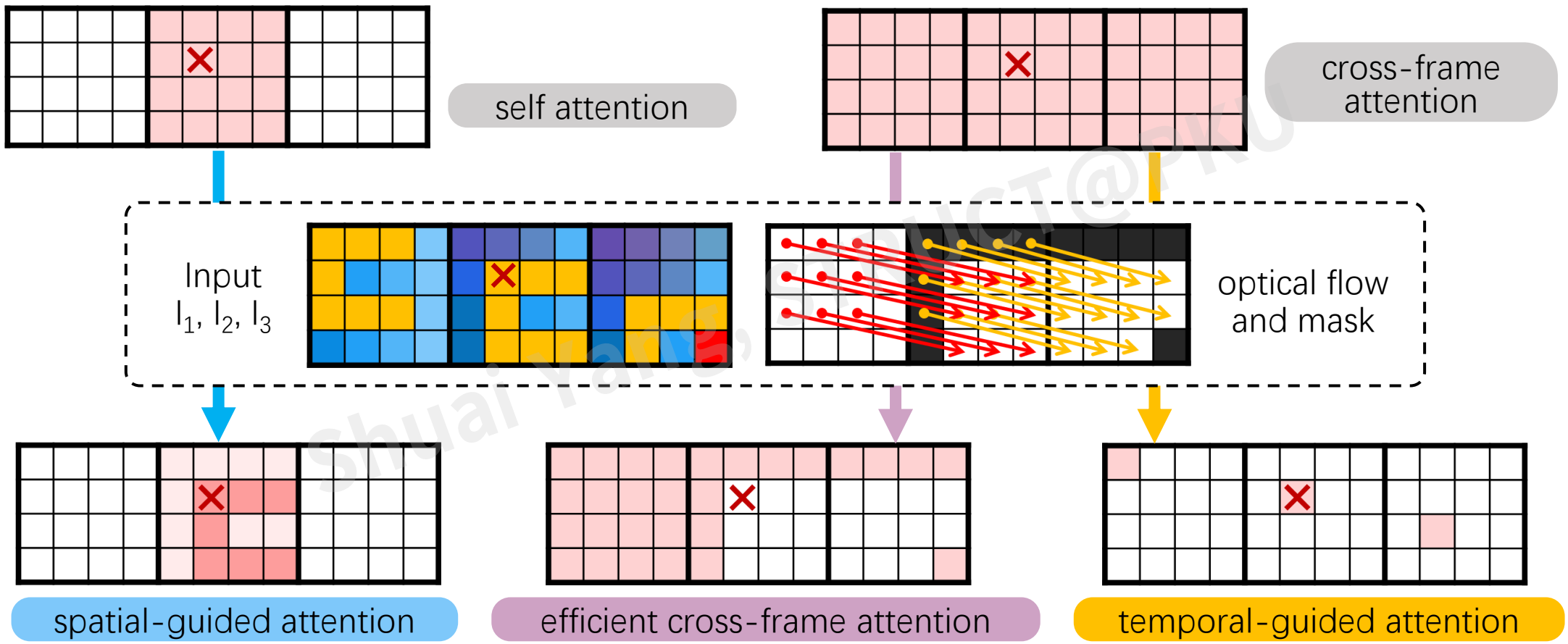
4 FRESCO



FRESCO: Spatial-Temporal Correspondence for Zero-Shot Video Translation. CVPR'24

Inversion-free zero-shot model

4 FRESCO



FRESCO: Spatial-Temporal Correspondence for Zero-Shot Video Translation. CVPR'24

Inversion-free zero-shot model

4 FRESCO

FRESCO: Spatial-Temporal Correspondence for Zero-Shot Video Translation. CVPR'24

Inversion-free zero-shot model

4 FRESCO

a white cat in pink background, cartoon style

input

Text2Video-Zero

ControlVideo

Rerender-A-Video

FRESCO

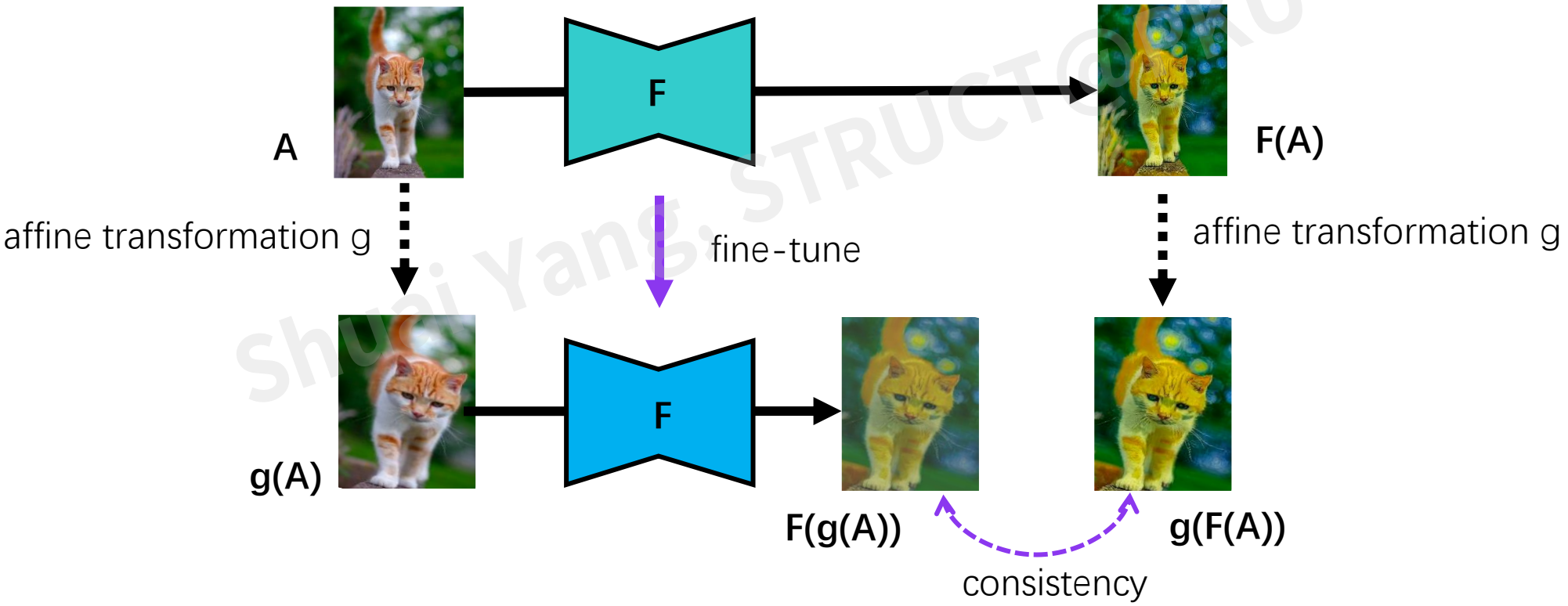
FRESCO: Spatial-Temporal Correspondence for Zero-Shot Video Translation. CVPR'24

Inversion-free zero-shot model

5 Fairy

While we are always fighting against the noises,

maybe we can allows some fine-tuning to make the model less sensitive to the noises



Fairy: Fast Parallellized Instruction-Guided Video-to-Video Synthesis. CVPR'24

Inversion-free zero-shot model

5 Fairy

~~cat~~ **fox** walking

input

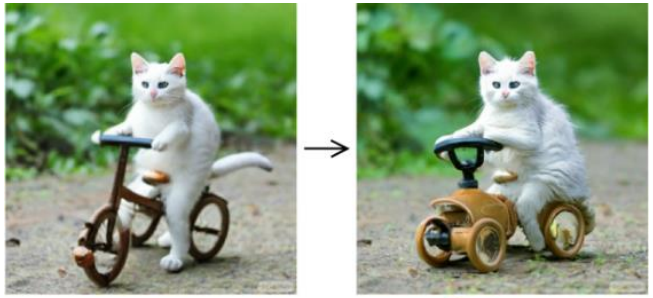
w/o fine-tune

full

Fairy: Fast Parallellized Instruction-Guided Video-to-Video Synthesis. CVPR'24

Zero-shot model

Image Editing



Prompt-to-Prompt
Null-Text Inversion
Plug-and-Play

Inversion-Based Model

ICCV'23 FateZero Pixel2Video
ICLR'24 TokenFlow FLATTEN
CVPR'24 VidToMe

Conditional Generation

ControlNet
SDEdit

Inversion-Free Model

ICCV'23 Text2Video-Zero
SIGGRAPH Asia'23 Rerender-A-Video
ICLR'24 ControlVideo
CVPR'24 FRESCO Fairy

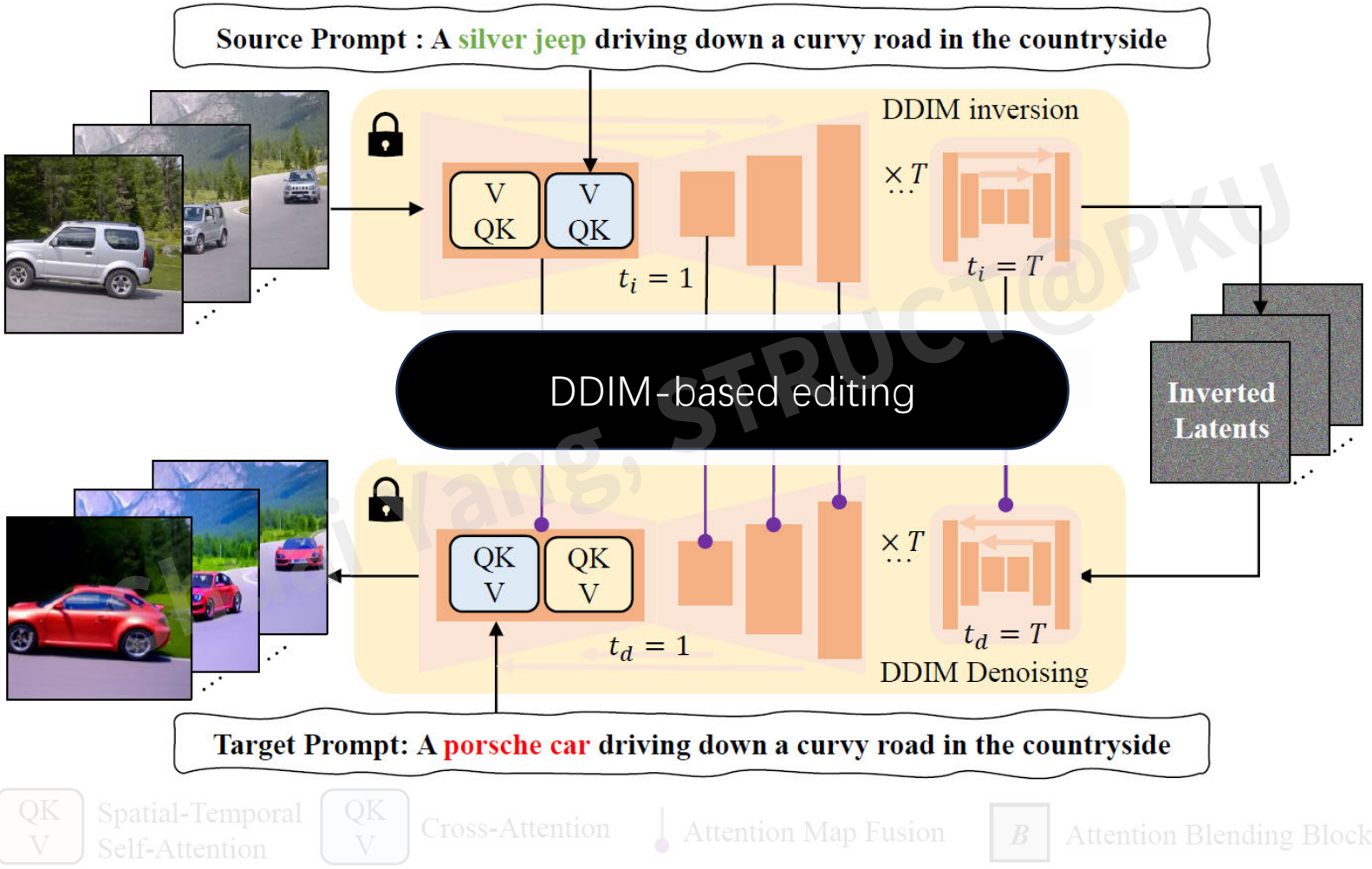
Inversion-based zero-shot model

1 FateZero

FateZero: Fusing Attentions for Zero-shot Text-based Video Editing. ICCV'23

Inversion-based zero-shot model

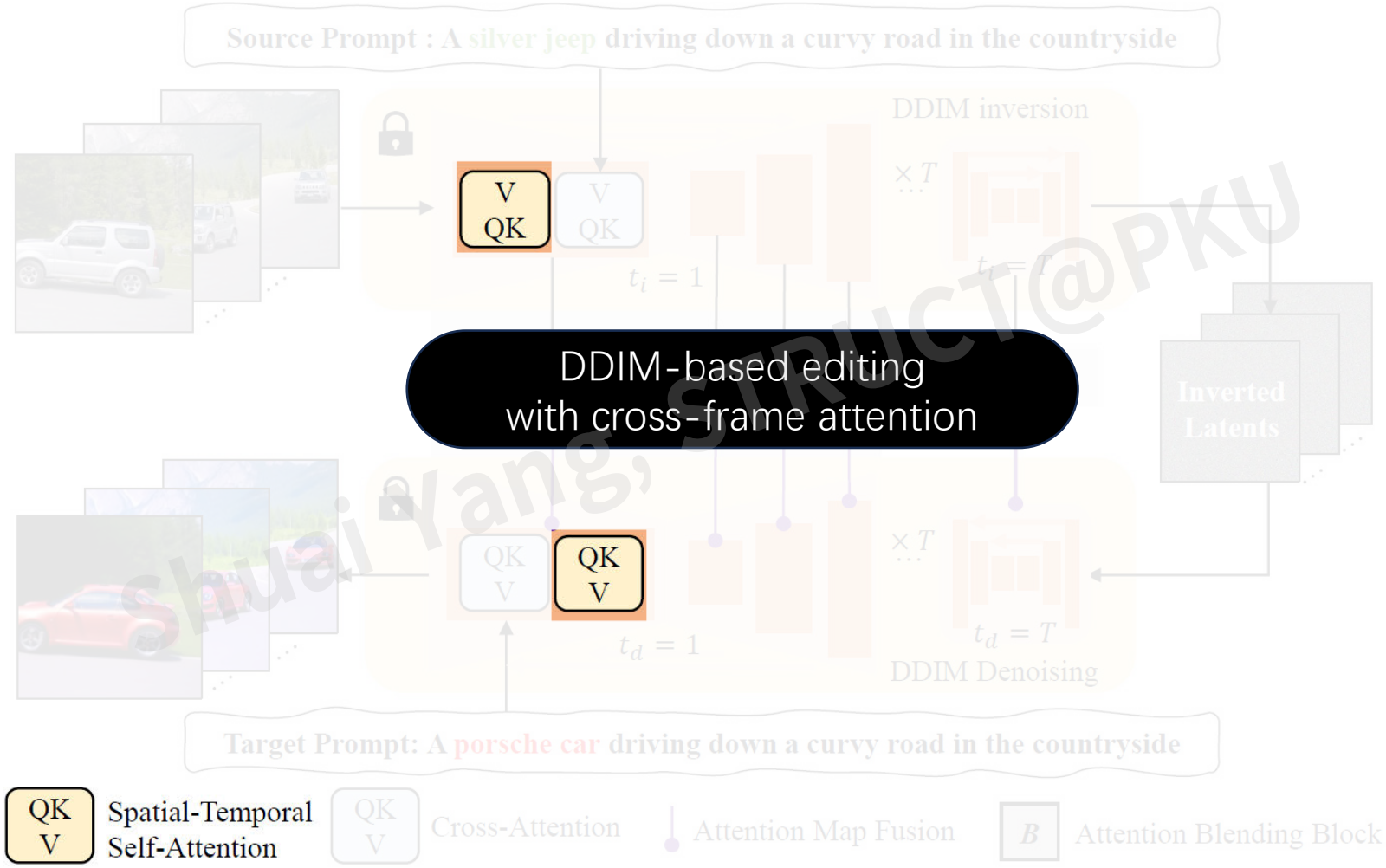
1 FateZero



FateZero: Fusing Attentions for Zero-shot Text-based Video Editing. ICCV'23

Inversion-based zero-shot model

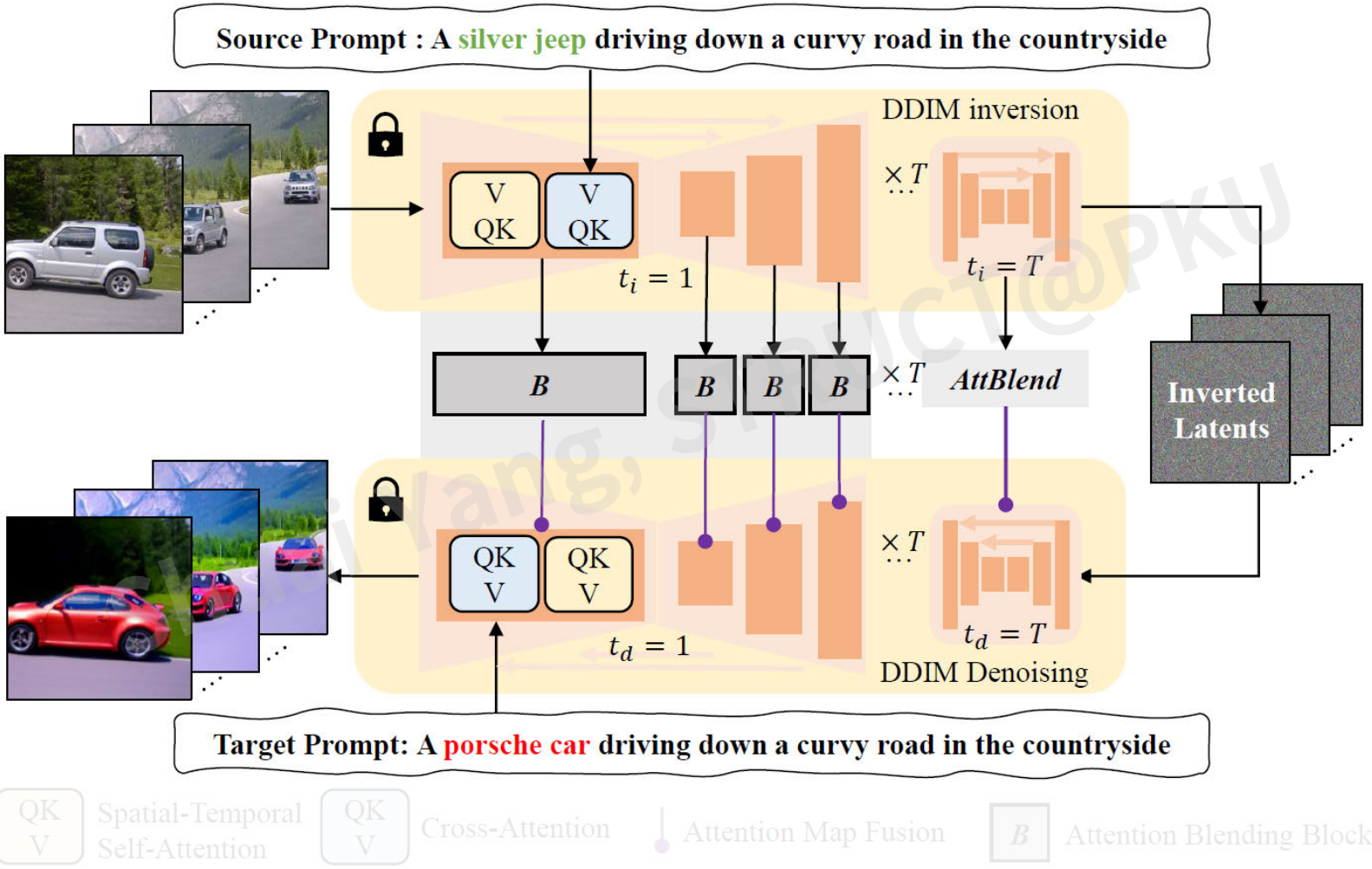
1 FateZero



FateZero: Fusing Attentions for Zero-shot Text-based Video Editing. ICCV'23

Inversion-based zero-shot model

1 FateZero



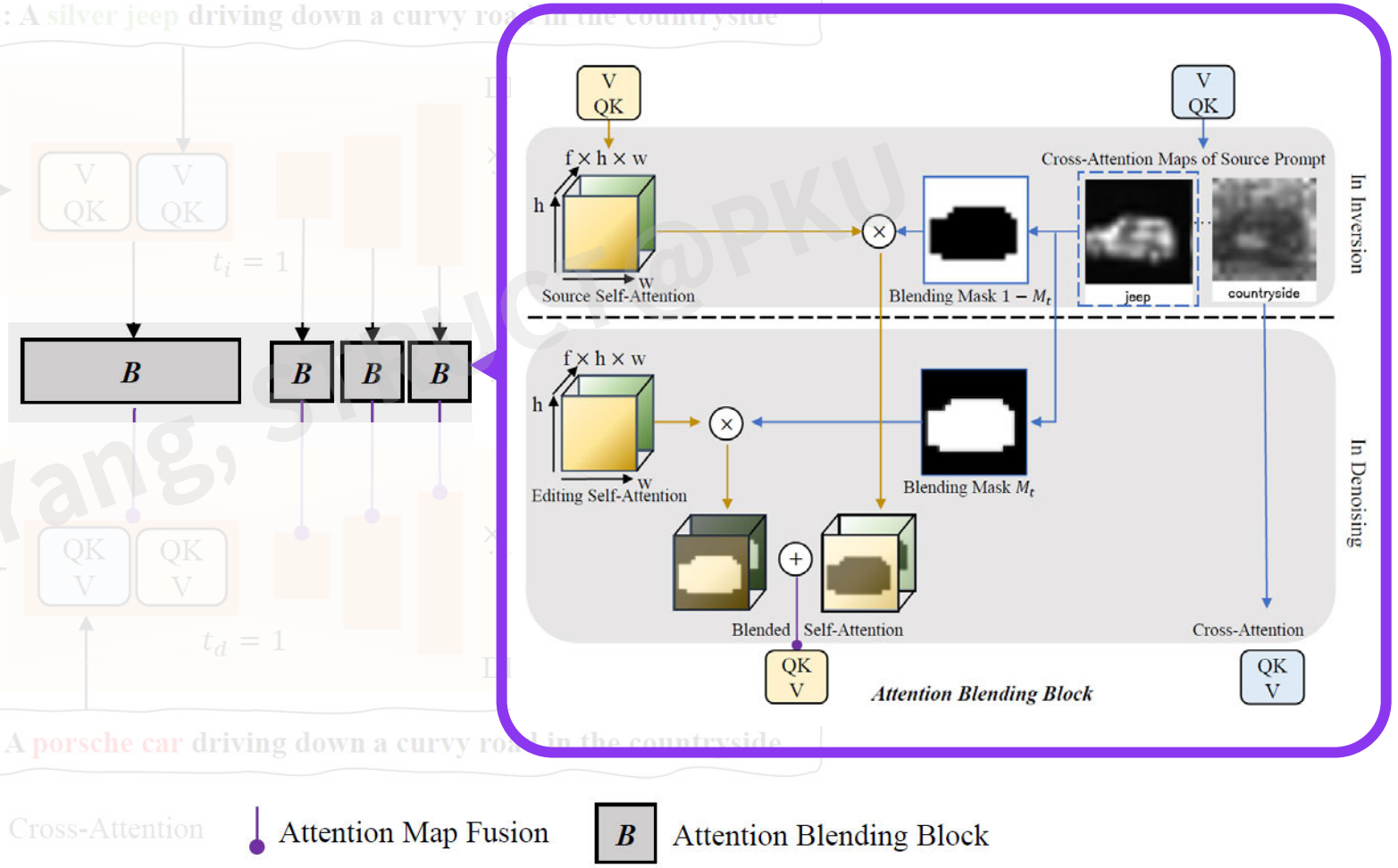
FateZero: Fusing Attentions for Zero-shot Text-based Video Editing. ICCV'23

Inversion-based zero-shot model

1 FateZero

Attention Blending

- Non-editing region uses the original self-attention map;
- Editing region uses the edited self-attention map



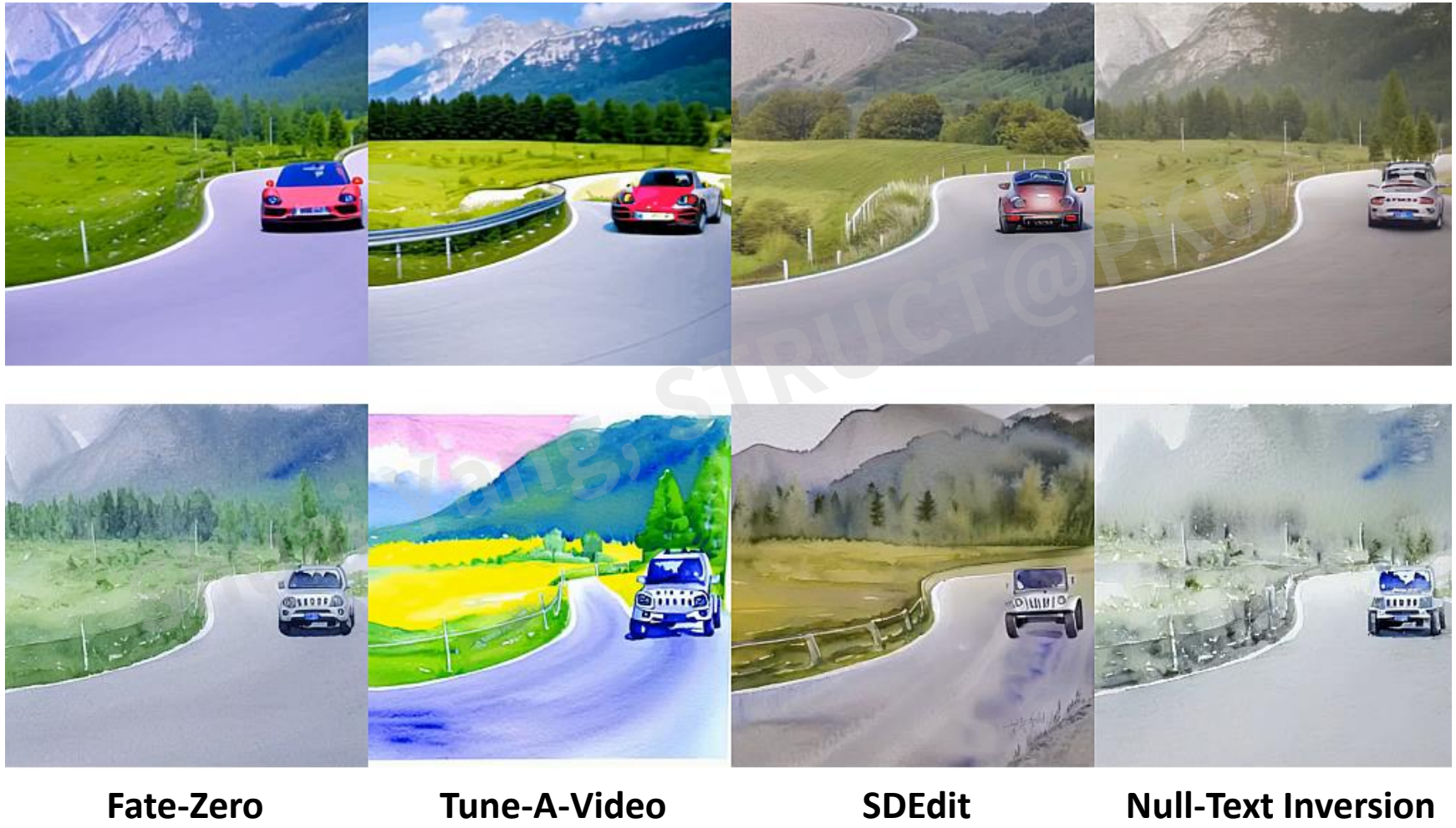
FateZero: Fusing Attentions for Zero-shot Text-based Video Editing. ICCV'23

Inversion-based zero-shot model

1 FateZero

~~*Car*~~
Posche Car

+
Watercolor Painting



Fate-Zero **Tune-A-Video** **SDEdit** **Null-Text Inversion**

FateZero: Fusing Attentions for Zero-shot Text-based Video Editing. ICCV'23

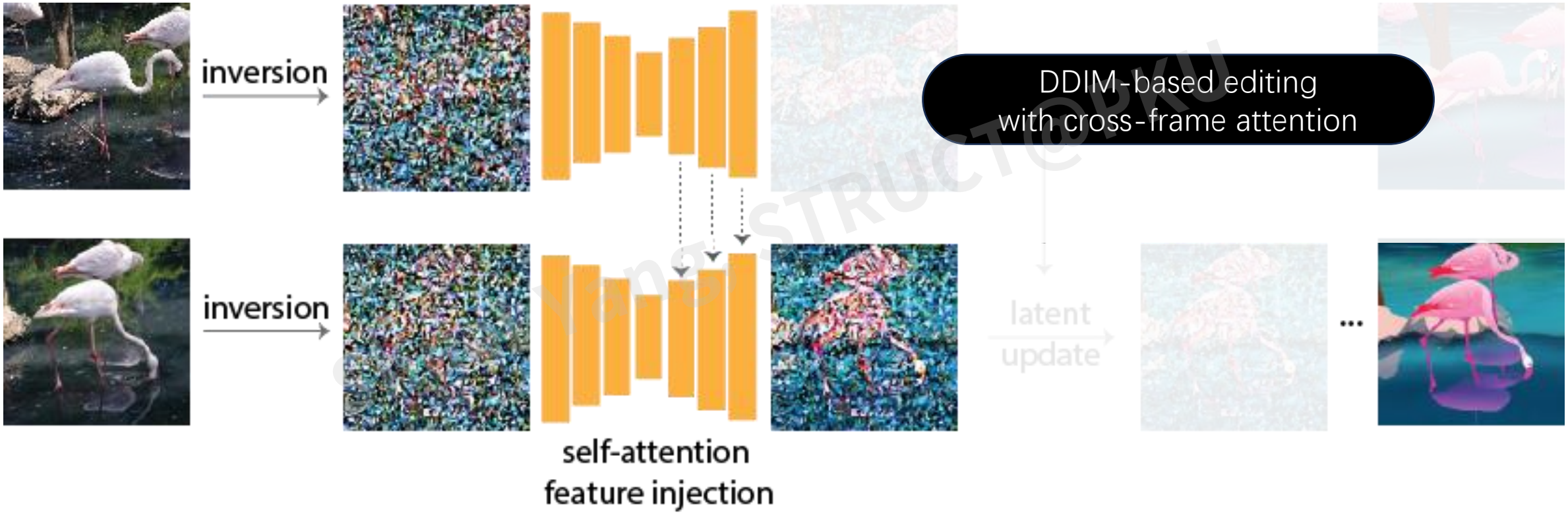
Inversion-based zero-shot model

2 Pix2Video

Pix2Video Video Editing using Image Diffusion. ICCV'23

Inversion-based zero-shot model

2 Pix2Video



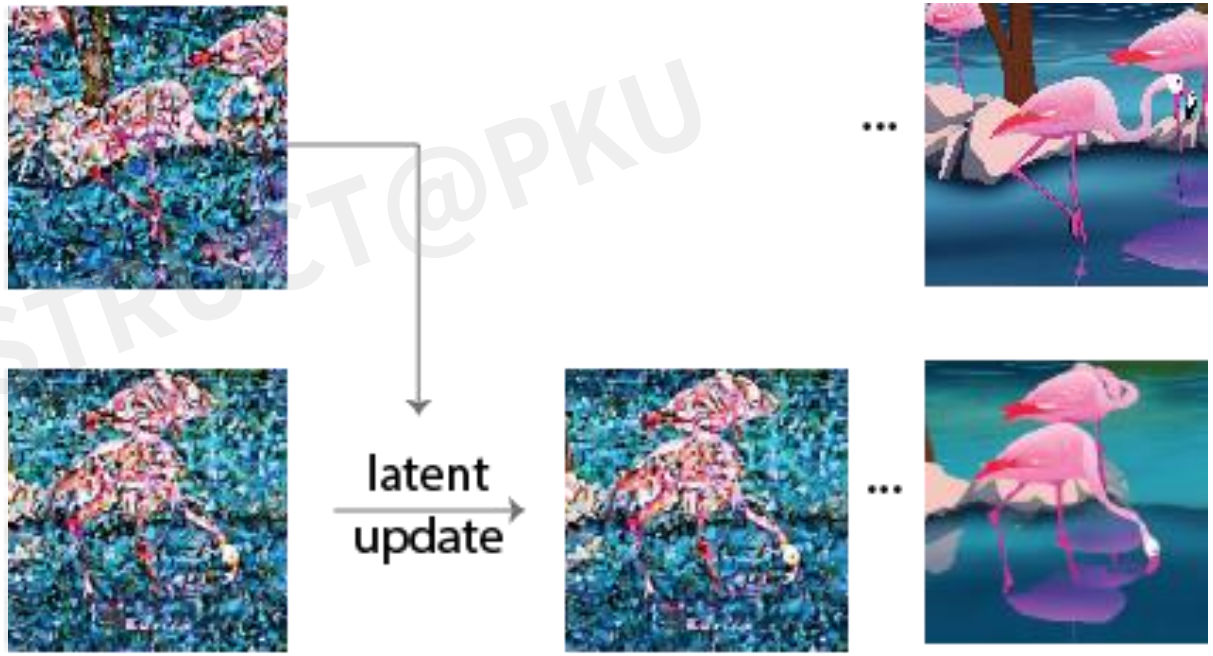
Pix2Video Video Editing using Image Diffusion. ICCV'23

Inversion-based zero-shot model

2 Pix2Video

Latent Update

- Optimize the denoised latent z_i such that
- z_i is similar to z_{i-1}

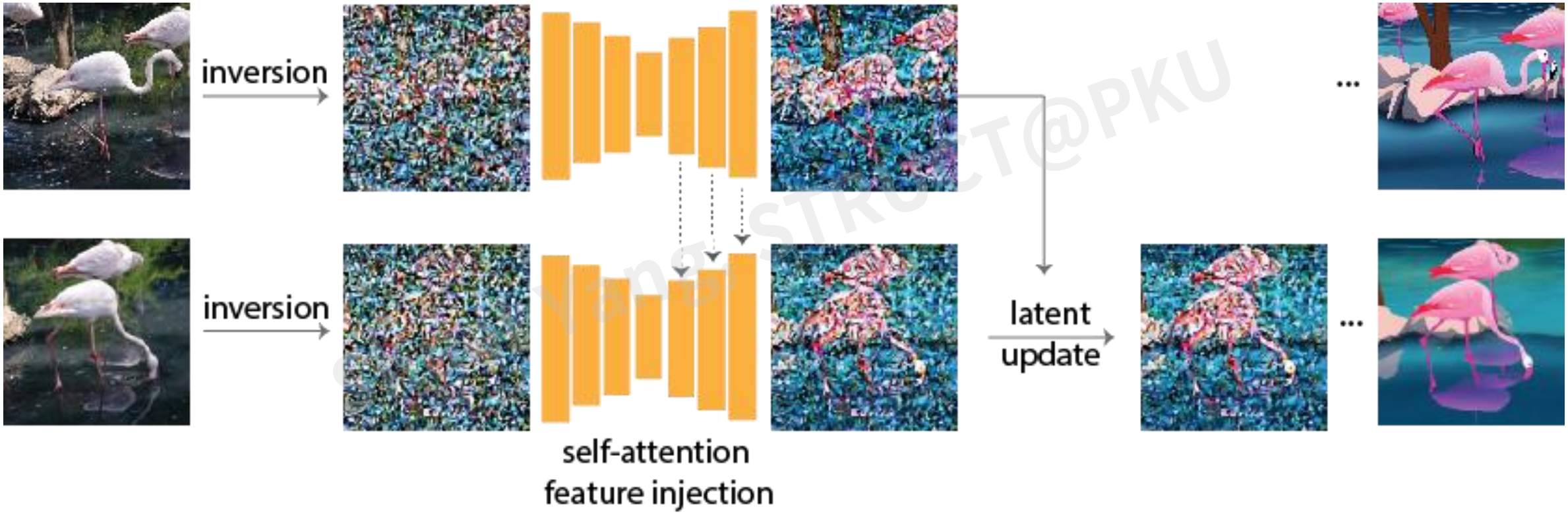


self-attention
feature injection

Pix2Video Video Editing using Image Diffusion. ICCV'23

Inversion-based zero-shot model

2 Pix2Video



Pix2Video Video Editing using Image Diffusion. ICCV'23

Inversion-based zero-shot model

2 Pix2Video

a group of pigs standing in the dirt, claymation

Input

Prompt-to-Prompt

Tune-A-Video

Pix2Video

Previous methods use inversion features globally via attention.
Can we use them more locally?

Pix2Video Video Editing using Image Diffusion. ICCV'23

Inversion-based zero-shot model

3 TokenFlow

Input

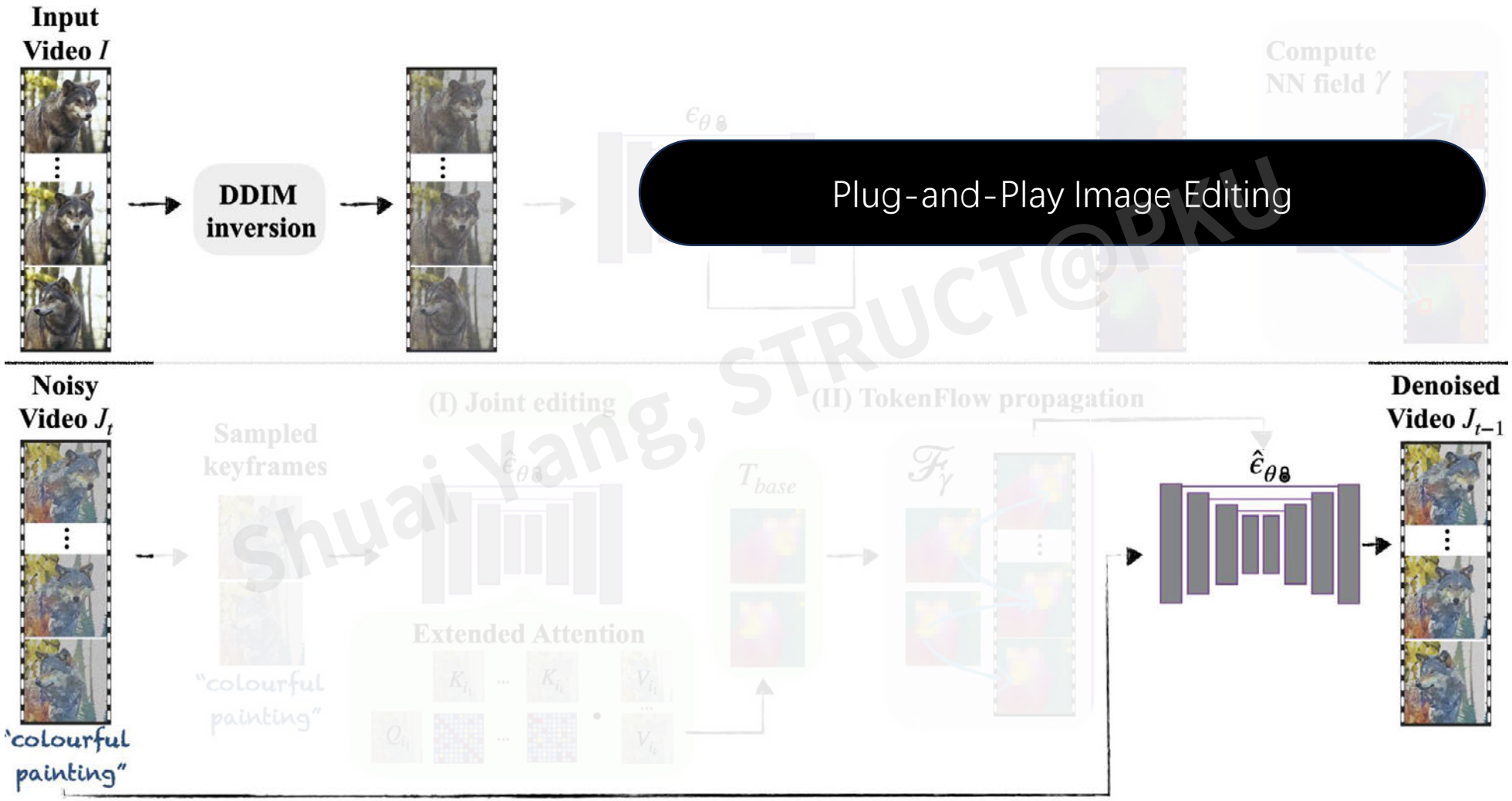
Visualized Diffusion feature

Diffusion features are semantically consistent.
Coherent features → coherent output!

TokenFlow: Consistent Diffusion Features for Consistent Video Editing. ICLR'24

Inversion-based zero-shot model

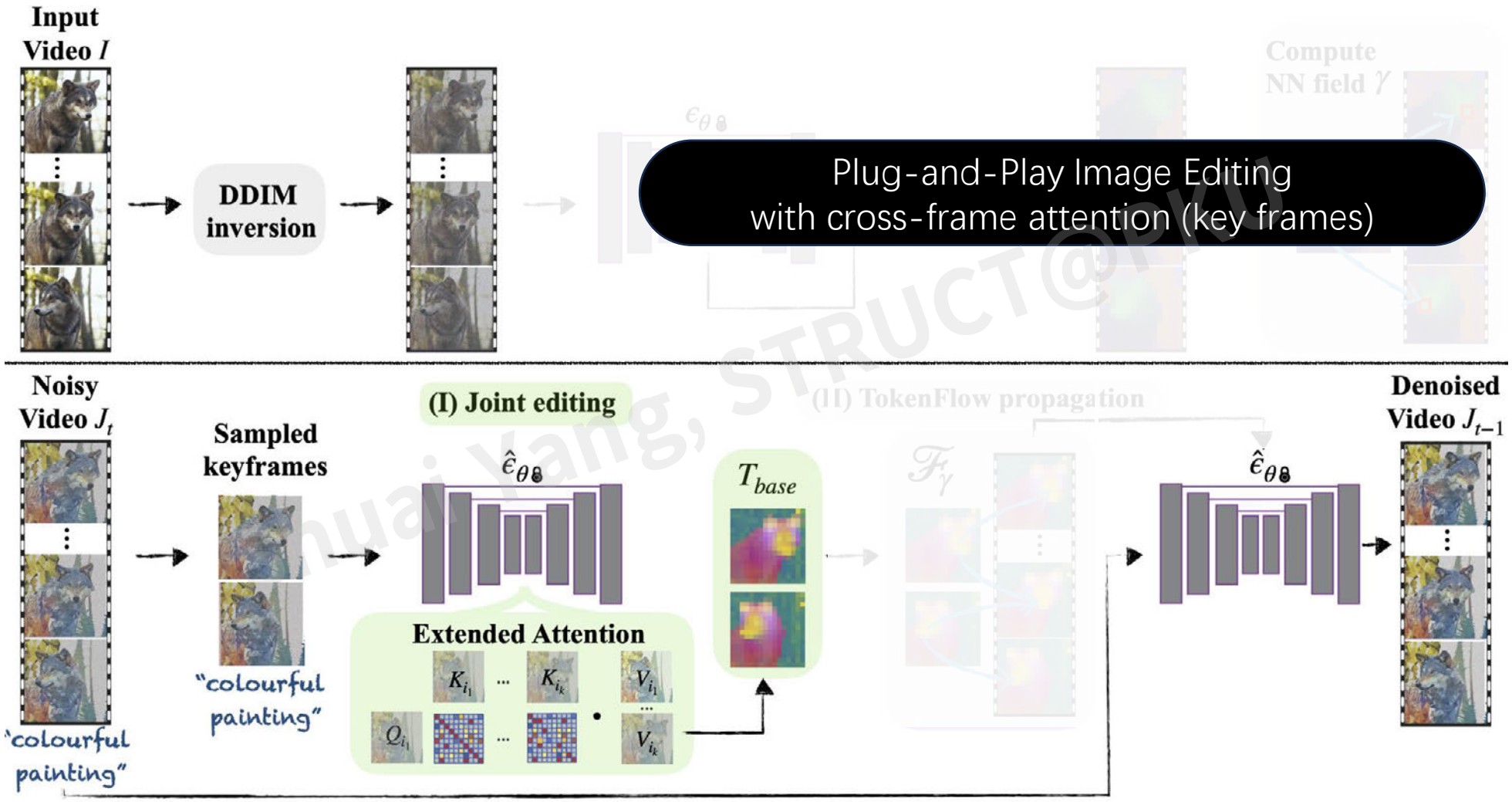
3 TokenFlow



TokenFlow: Consistent Diffusion Features for Consistent Video Editing. ICLR'24

Inversion-based zero-shot model

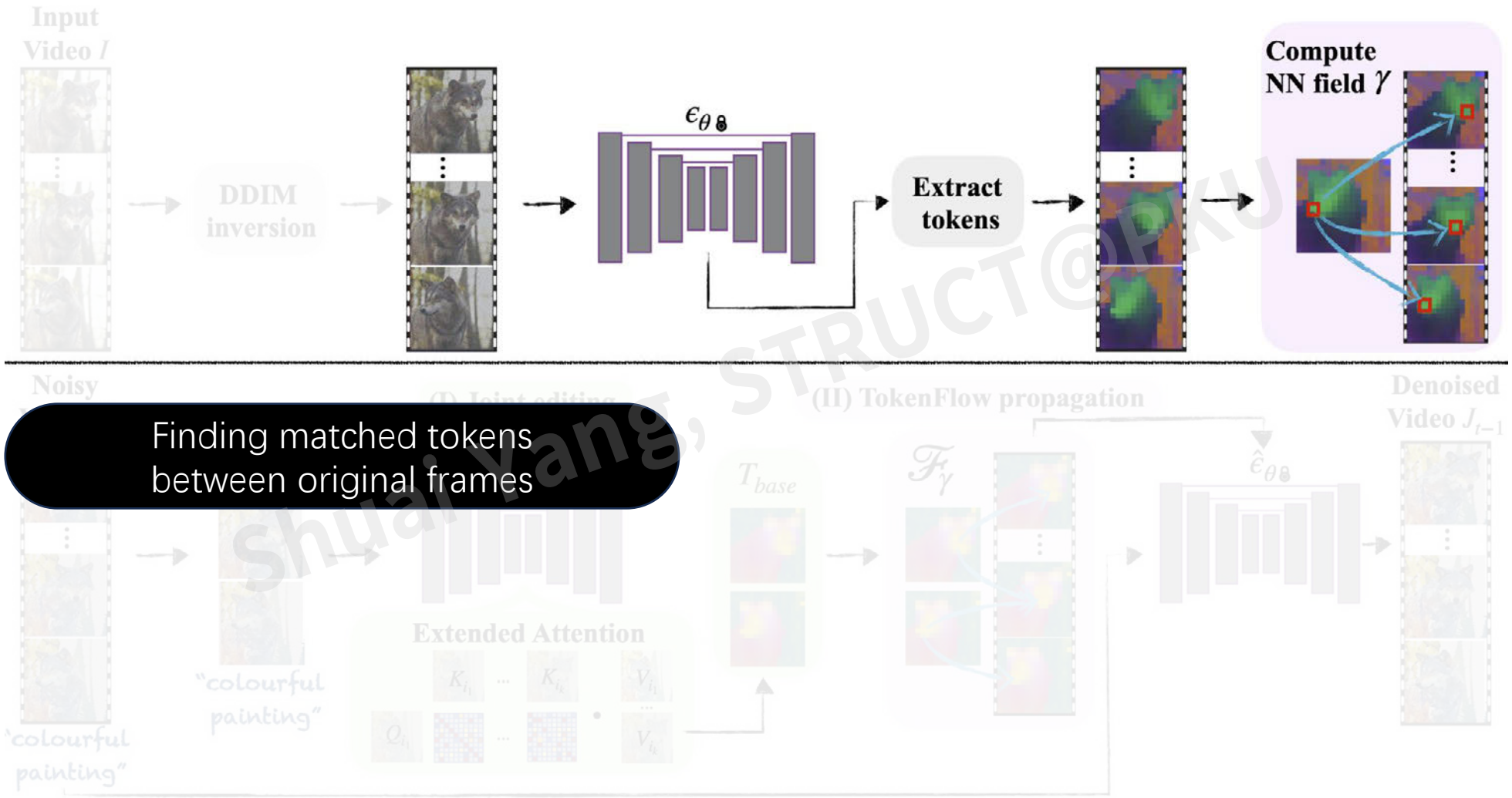
3 TokenFlow



TokenFlow: Consistent Diffusion Features for Consistent Video Editing. ICLR'24

Inversion-based zero-shot model

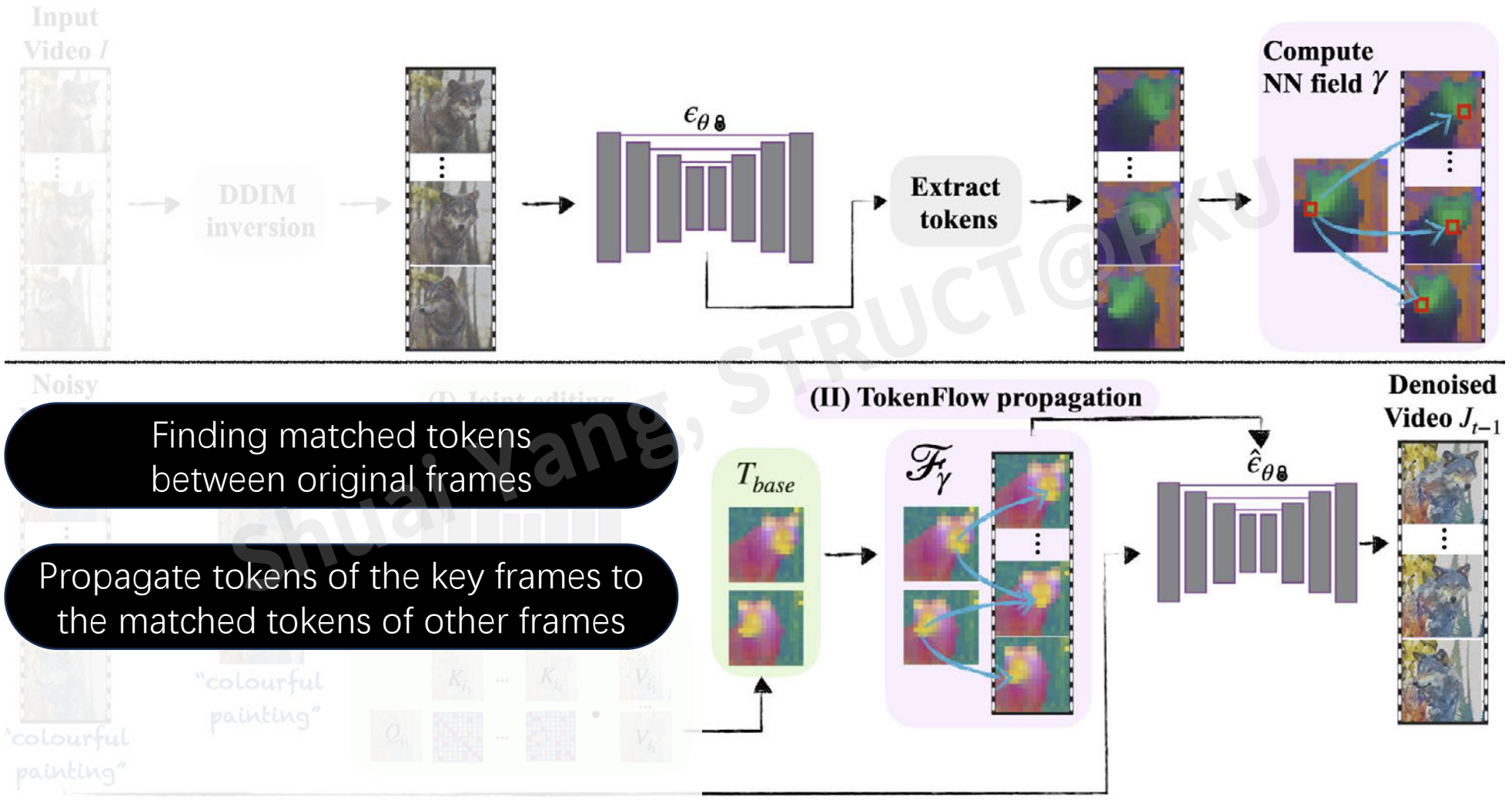
3 TokenFlow



TokenFlow: Consistent Diffusion Features for Consistent Video Editing. ICLR'24

Inversion-based zero-shot model

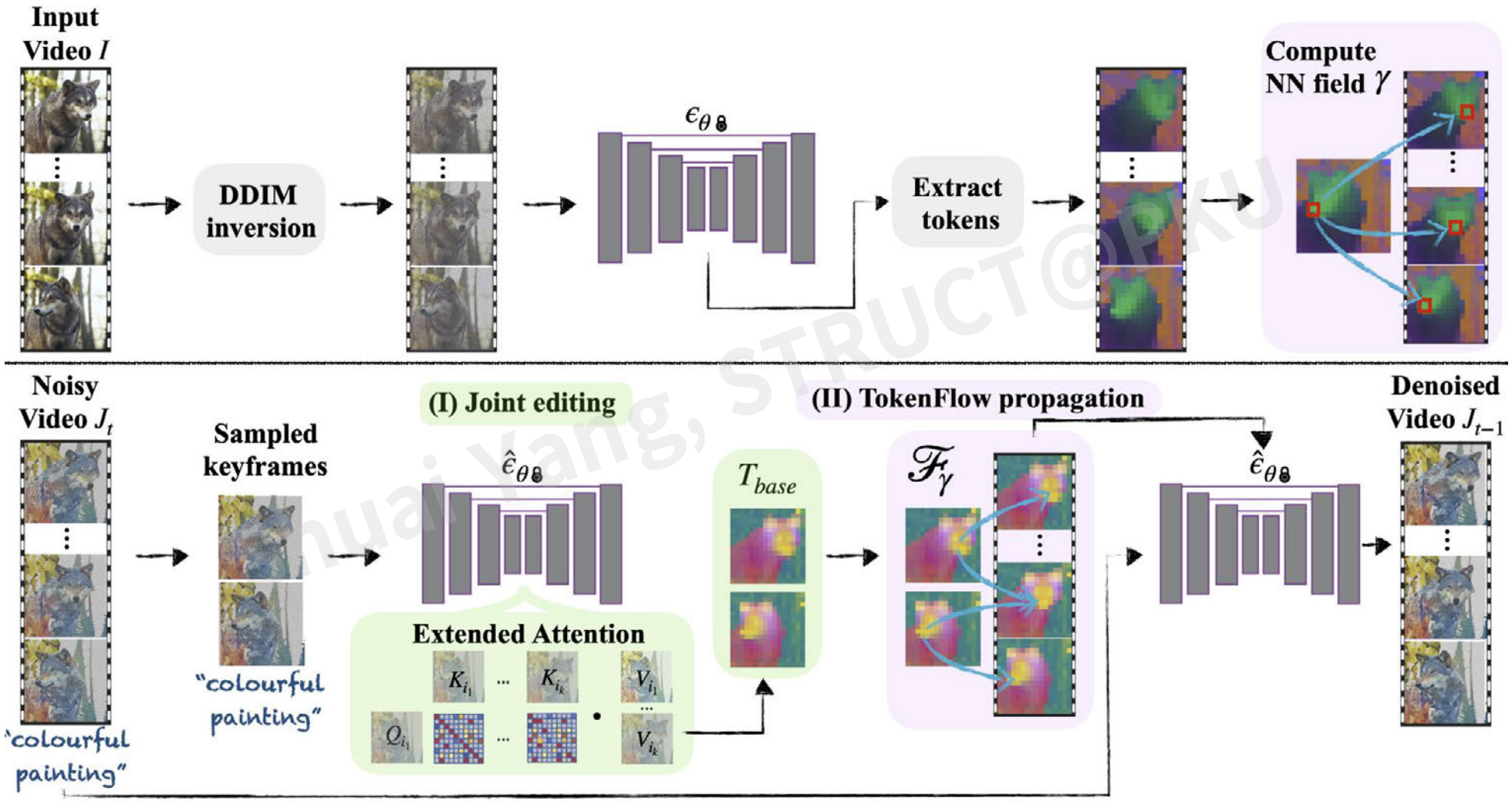
3 TokenFlow



TokenFlow: Consistent Diffusion Features for Consistent Video Editing. ICLR'24

Inversion-based zero-shot model

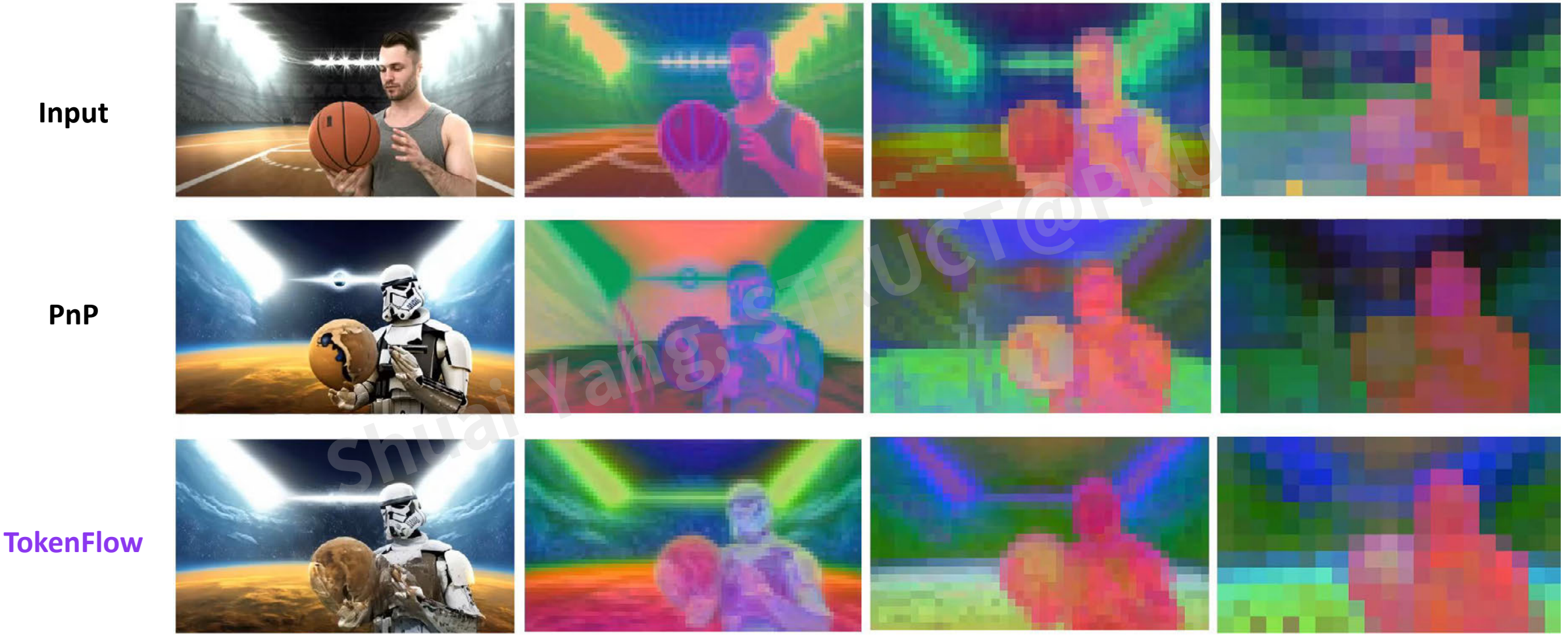
3 TokenFlow



TokenFlow: Consistent Diffusion Features for Consistent Video Editing. ICLR'24

Inversion-based zero-shot model

3 TokenFlow



Visualized features in different layers

TokenFlow: Consistent Diffusion Features for Consistent Video Editing. ICLR'24

Inversion-based zero-shot model

3 TokenFlow

a fluffy wolf doll

Input

Tune-A-Video

FateZero

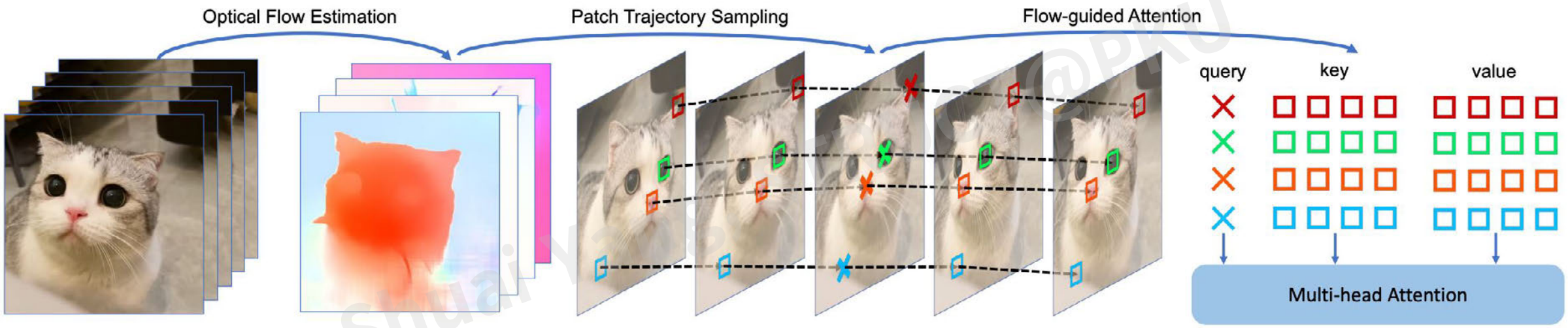
TokenFlow

Locally matched tokens can significantly improve the consistency.
What about finer-grained optical flow?

TokenFlow: Consistent Diffusion Features for Consistent Video Editing. ICLR'24

Inversion-based zero-shot model

4 FLATTEN

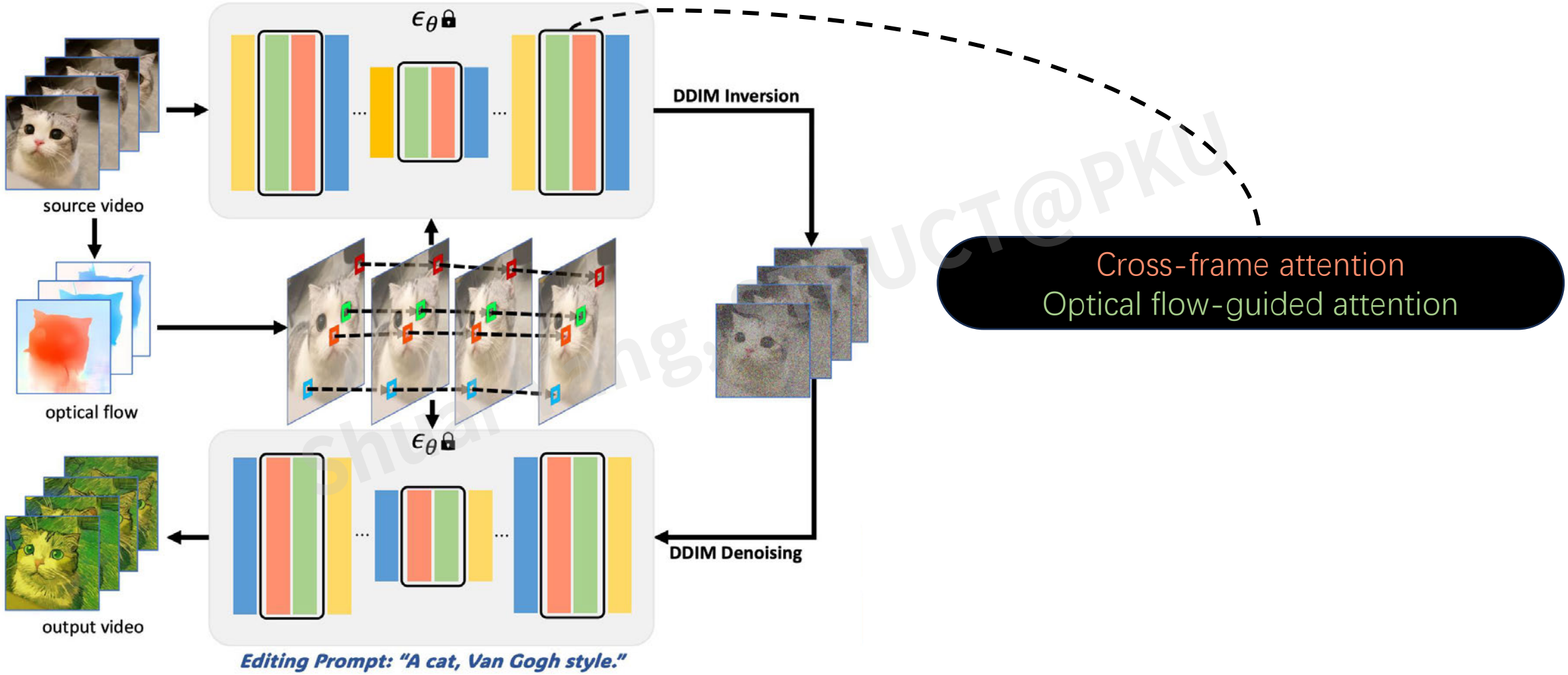


FLATTEN: optical FLow-guided ATTENtion

FLATTEN: Optical Flow-guided Attention for Consistent Text-to-Video Editing. ICLR'24

Inversion-based zero-shot model

4 FLATTEN



FLATTEN: Optical Flow-guided Attention for Consistent Text-to-Video Editing. ICLR'24

Inversion-based zero-shot model

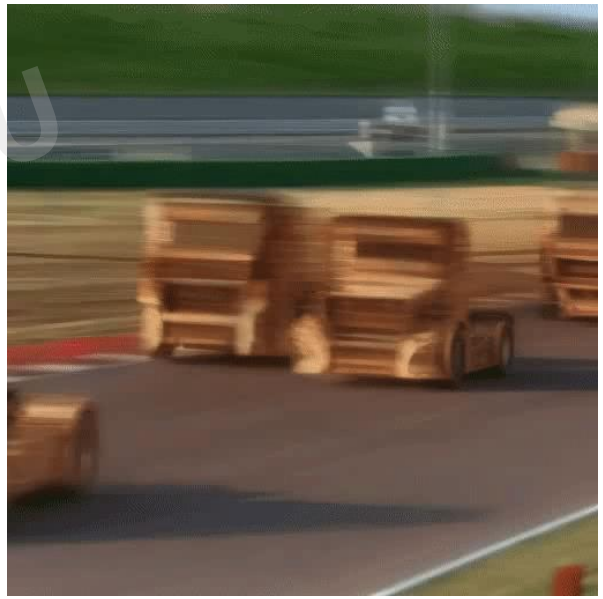
4 FLATTEN

***Wooden* trucks drive on a racetrack**

Input

FateZero

TokenFlow



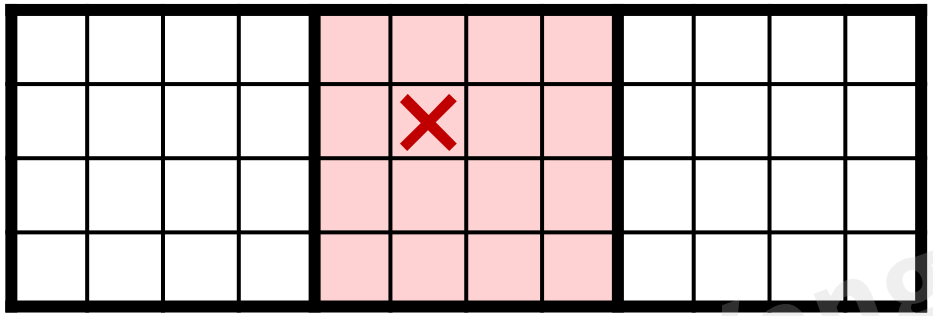
FLATTEN

Attention mask or feature blending may harm editability.
Only FLATTEN changes the appearance.

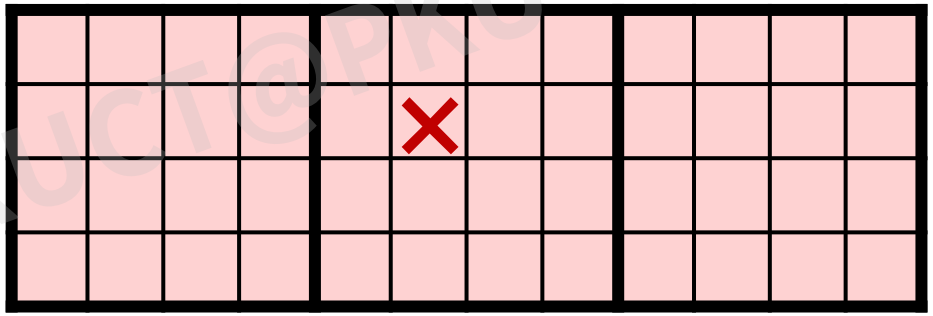
FLATTEN: Optical Flow-guided Attention for Consistent Text-to-Video Editing. ICLR'24

Inversion-based zero-shot model

5 VidToMe



Self attention



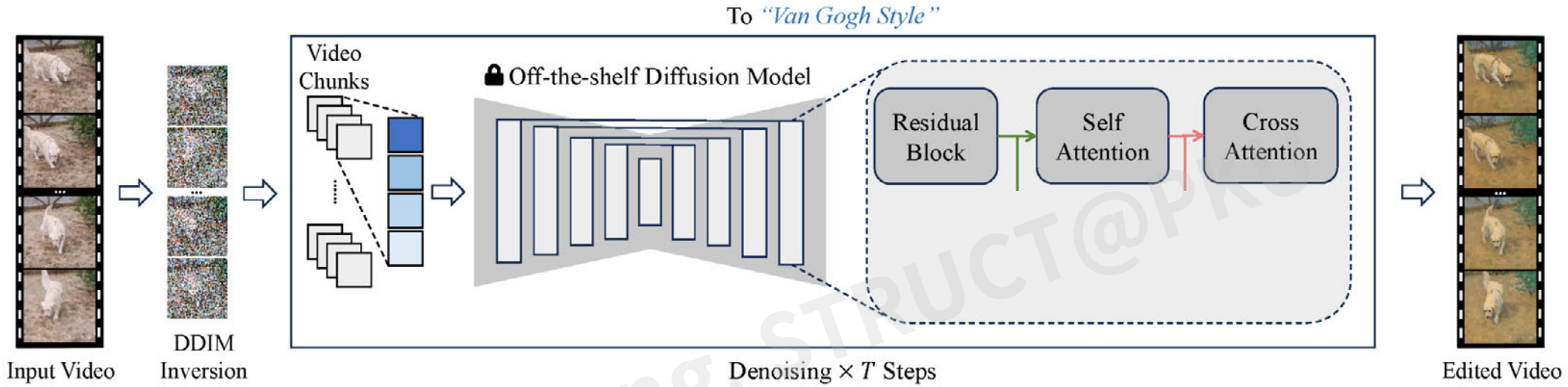
Cross-frame attention

With B frames, computation increase B*B times
Redundant tokens can be merged!

VidToMe: Video Token Merging for Zero-Shot Video Editing. CVPR'24

Inversion-based zero-shot model

5 VidToMe



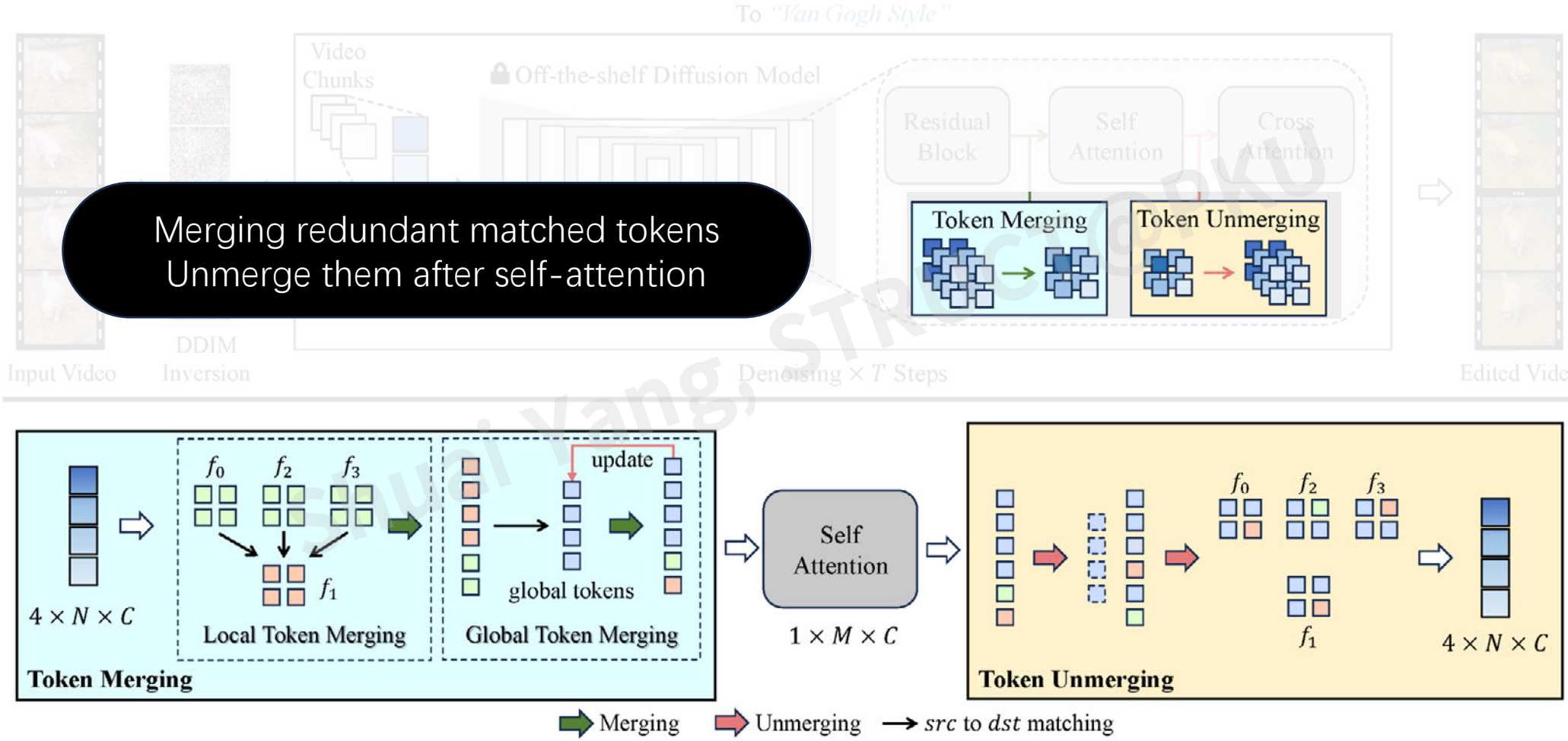
Plug-and-Play Image Editing

VidToMe: Video Token Merging for Zero-Shot Video Editing. CVPR'24

Inversion-based zero-shot model

5 VidToMe

Merging redundant matched tokens
Unmerge them after self-attention



VidToMe: Video Token Merging for Zero-Shot Video Editing. CVPR'24

Inversion-based zero-shot model

5 VidToMe

VidToMe: Video Token Merging for Zero-Shot Video Editing. CVPR'24

Zero-shot model

Image Editing

Inversion-Based Model

more stable, less editability

a white cat in pink background, cartoon style

+ inversion feature fusion

Conditional Generation

Inversion-Free Model

more editability, less stable

a white cat in pink background, cartoon style

+ depth condition

SUMMARY

- Big models and big data bring about key emergent capabilities
- Fundamental big models offer a lot of opportunities for creative work
- Even with limited resources, we can still use the fundamental big models for various downstream tasks
- Future: long and consistent video generation

→ make everyone a creator


Thank you for listening!
williamyang@pku.edu.cn
Shuai Yang, STRUCT@PKU

STRUCT Group

Intelligent image computing

Wangxuan Institute of Computer Technology, Peking University

Spatial and Temporal **Restoration**,
Understanding and **Compression** Team

PI: Jiaying Liu

- **Email:** liujiaying@pku.edu.cn
- **Website:** http://www.wict.pku.edu.cn/struct/