Ctrl-X: Controlling Structure and Appearance for Text-To-Image Generation Without Guidance

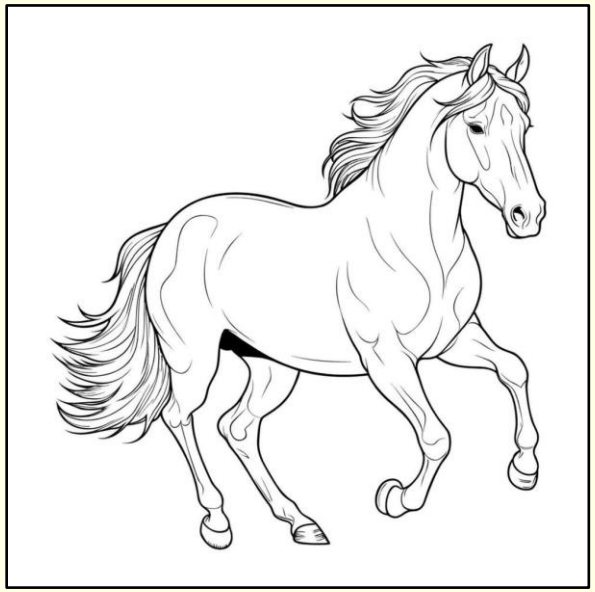
Kuan Heng Lin[1]*, Sicheng Mo[1]*, Ben Klingher[1], Fangzhou Mu[2], Bolei Zhou[1]
[1]University of California, Los Angeles [2]NVIDIA

*Indicates equal contribution

https://genforce.github.io/ctrl-x

Structure and appearance control

Structure



+

Appearance

Provided by the user

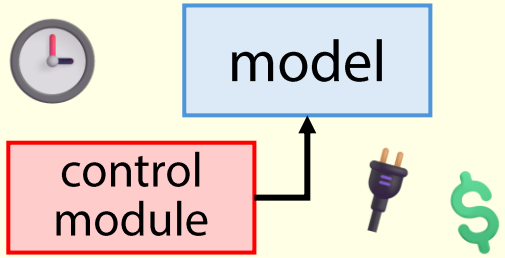
=

Output

Key challenges

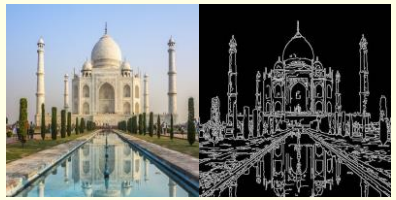
Training-based methods

Requires expensive training for *each* model architecture



Requires large amounts of paired data, difficult to gather for challenging conditions

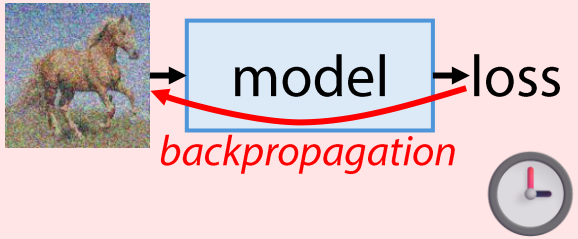
Canny
Easy to obtain



3D mesh
Difficult to obtain

Guidance-based methods

High inference time and GPU memory due to guidance

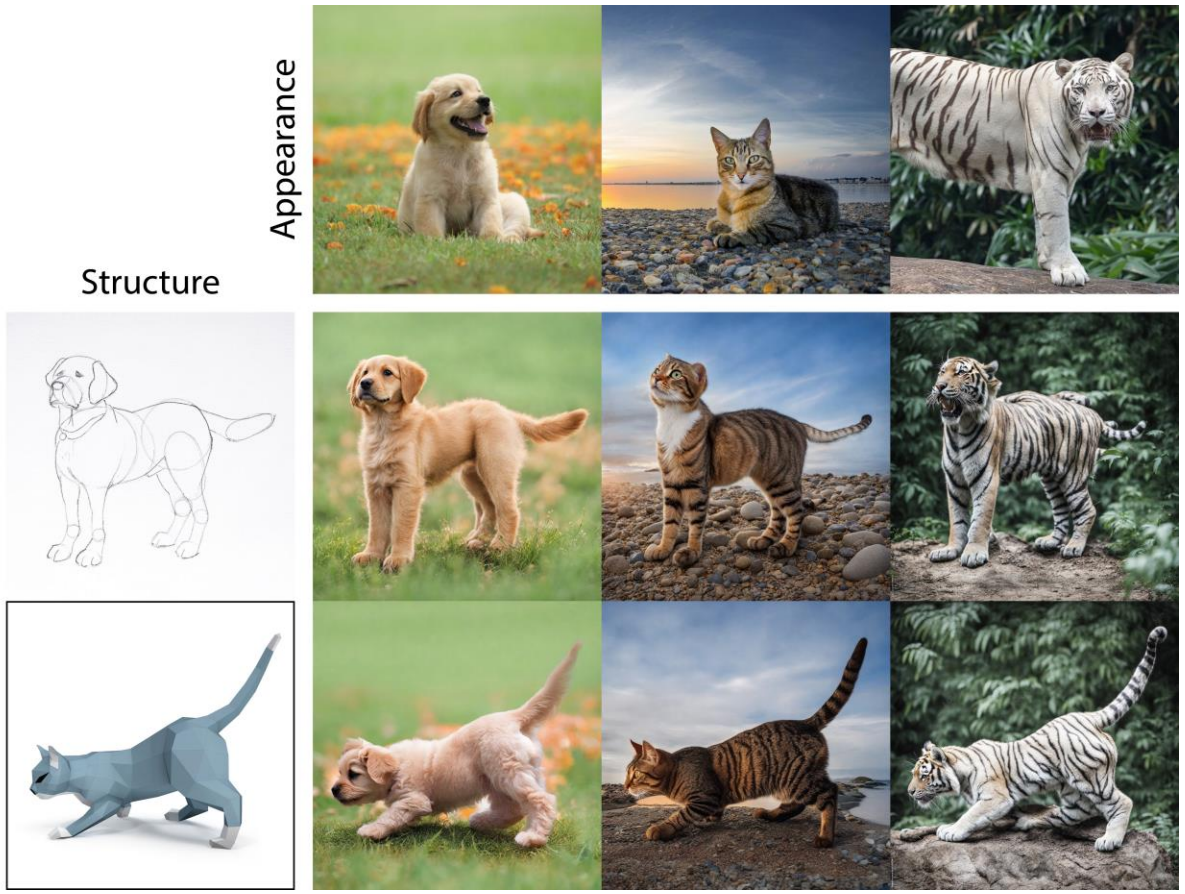


Sensitive to guidance weights, needing to be tuned per condition/loss type and per-image

Prone to artifacts from latents OOD, requiring more diffusion steps

Can we achieve controllable generation *without* **training** or **guidance**?

Our solution: Ctrl-X

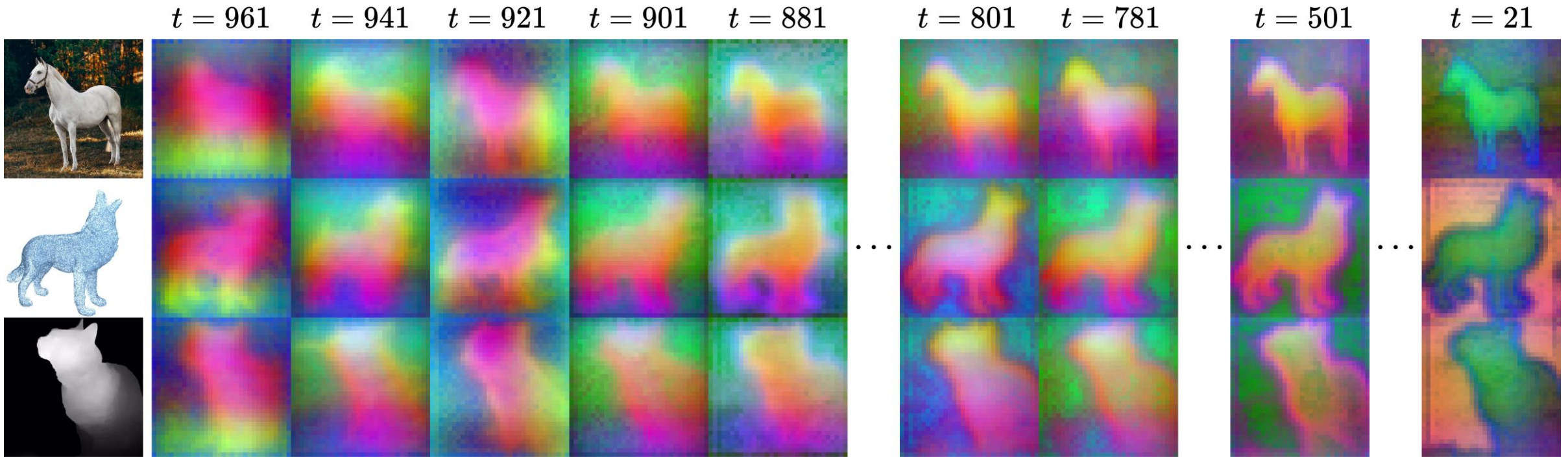


Training-free & guidance-free
Simple plug-and-play method

Multiple condition signals
Structure & appearance control

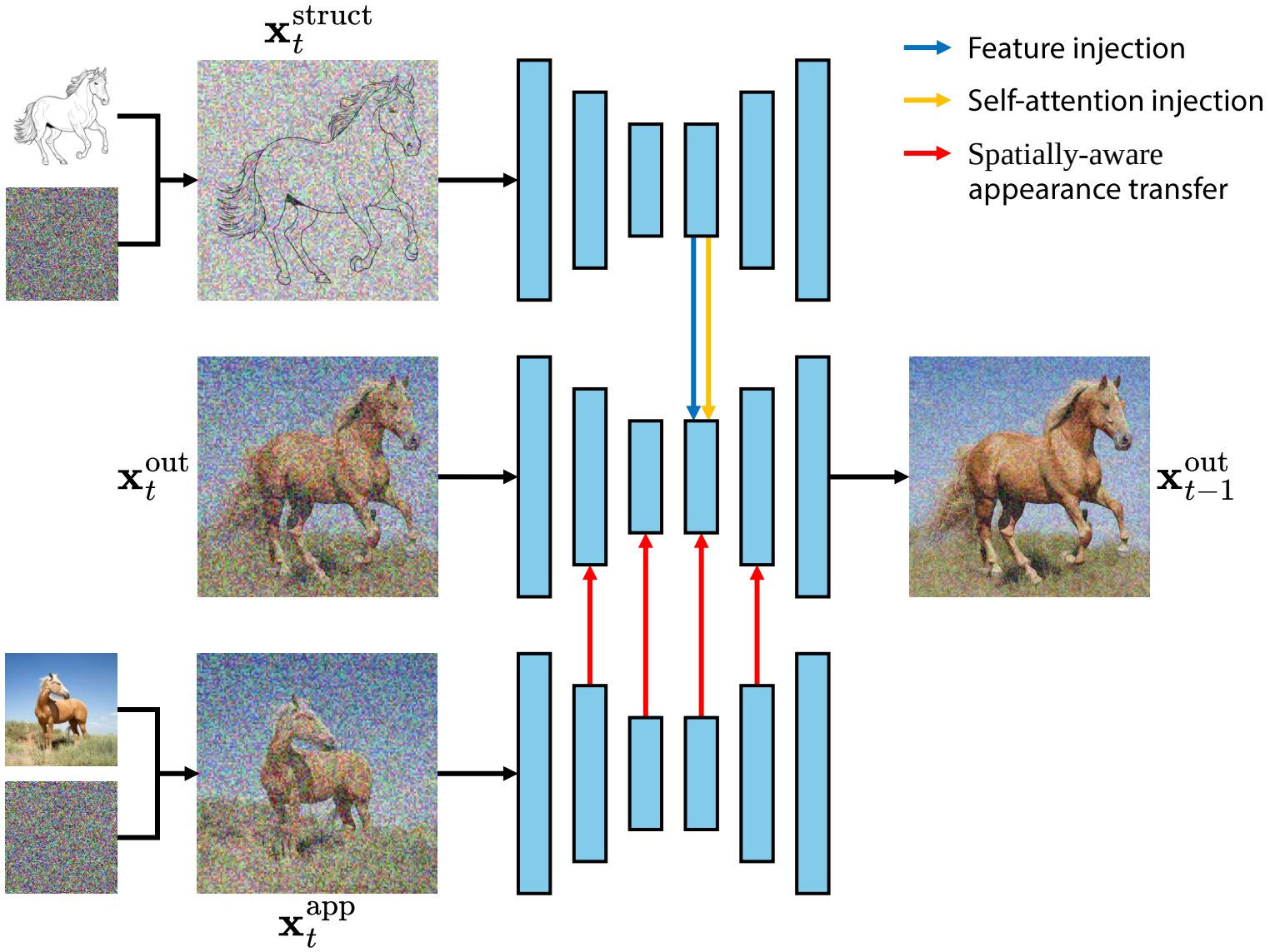
Lightweight and flexible
Any architecture and checkpoint

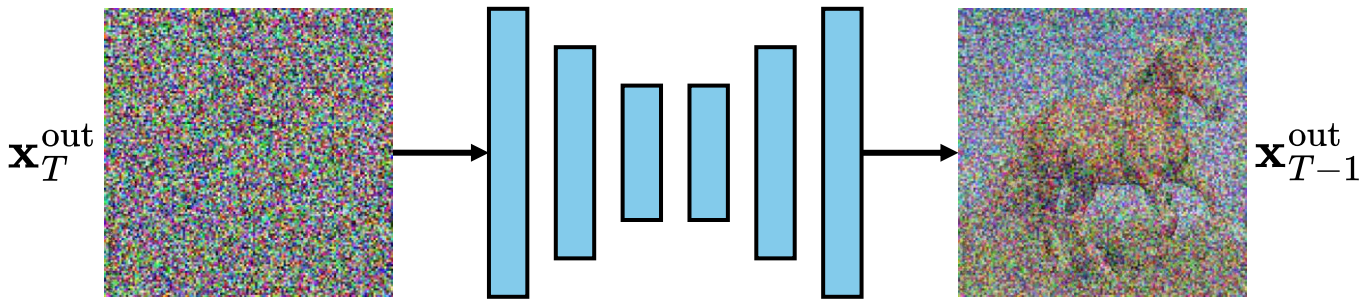
Key insight

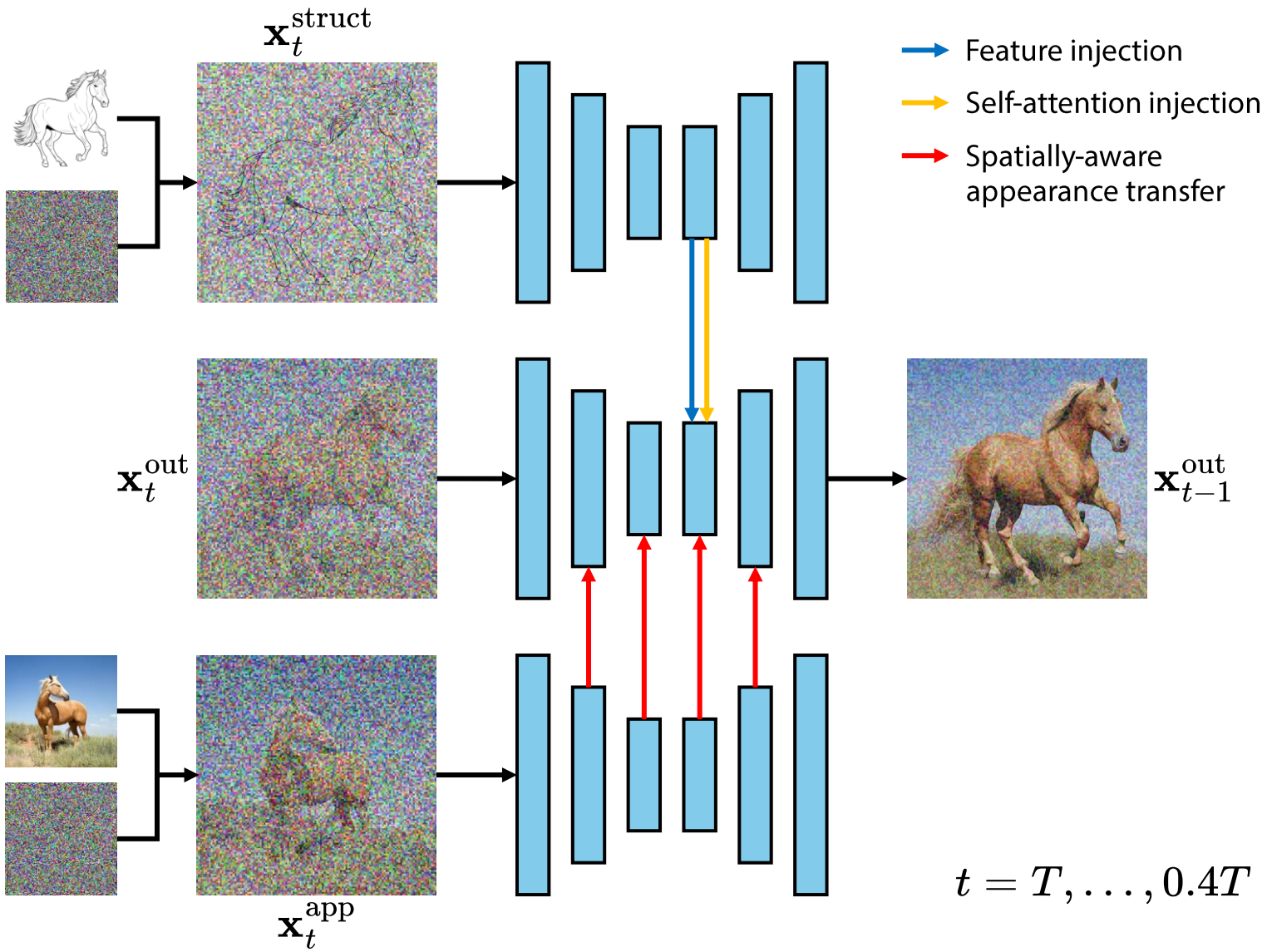


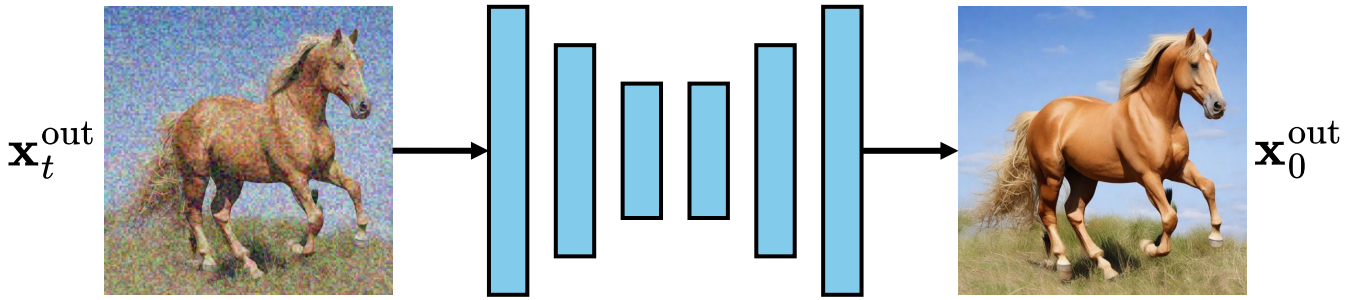
Forward diffusion features contain sufficient structure information at *very early* timesteps across modalities. Thus, we have semantic correspondence between two images via self-attention $\mathbf{QK}^\top$ without inversion.

Method overview

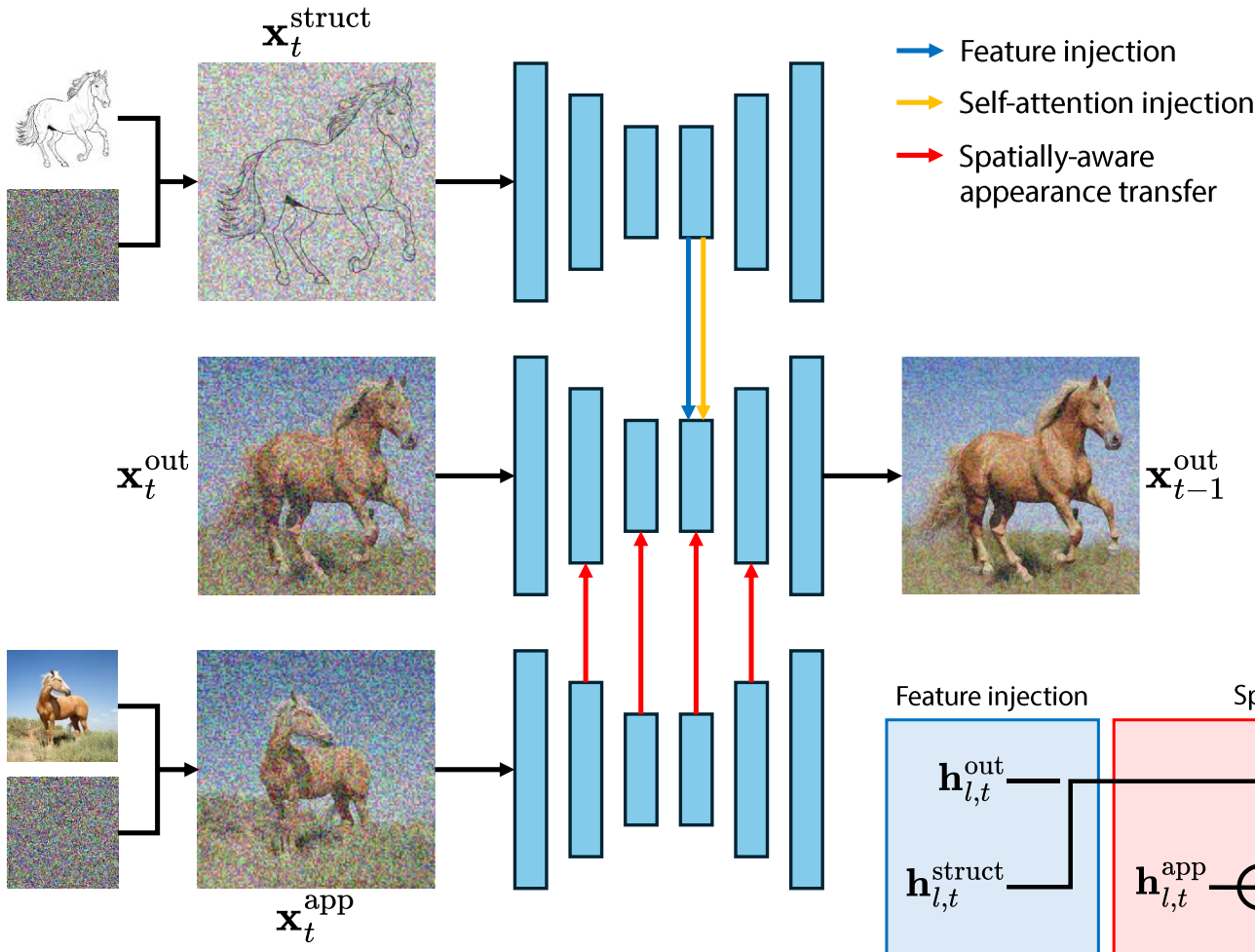


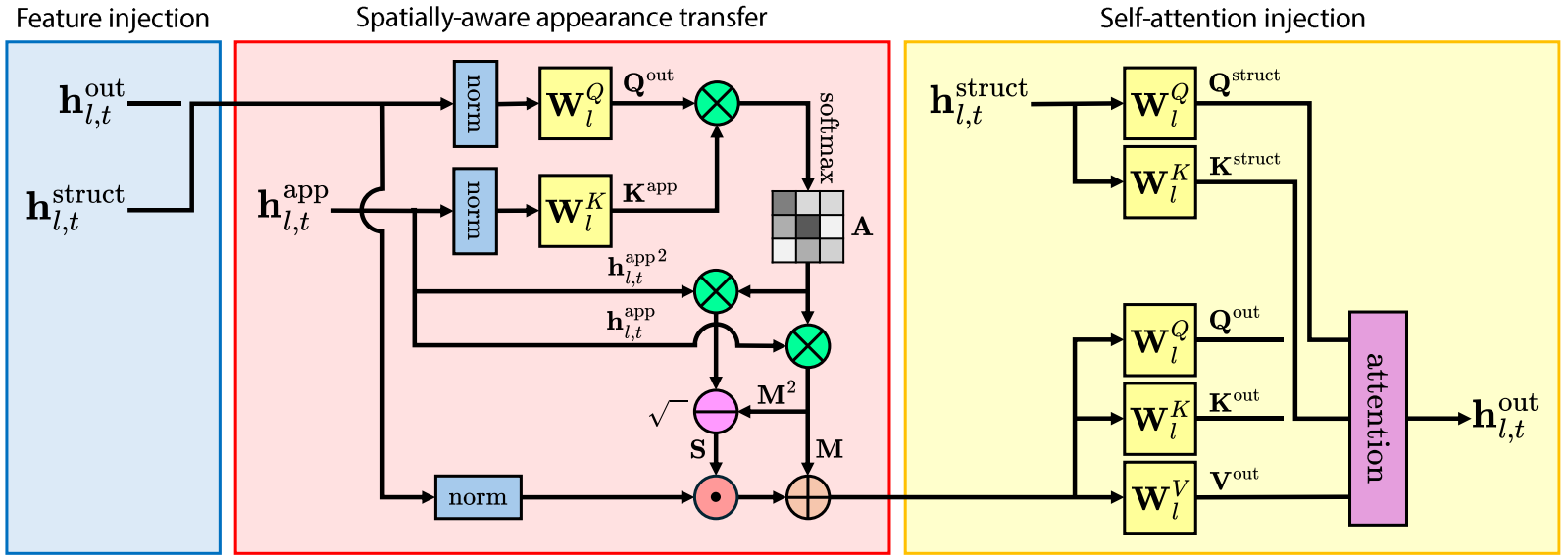

$\mathbf{x}_T^{\text{out}}$
$\mathbf{x}_{T-1}^{\text{out}}$


$\mathbf{x}_t^{\text{struct}}$
Feature injection
Self-attention injection
Spatially-aware appearance transfer
$\mathbf{x}_t^{\text{out}}$
$\mathbf{x}_{t-1}^{\text{out}}$
$\mathbf{x}_t^{\text{app}}$
$t = T, \ldots, 0.4T$

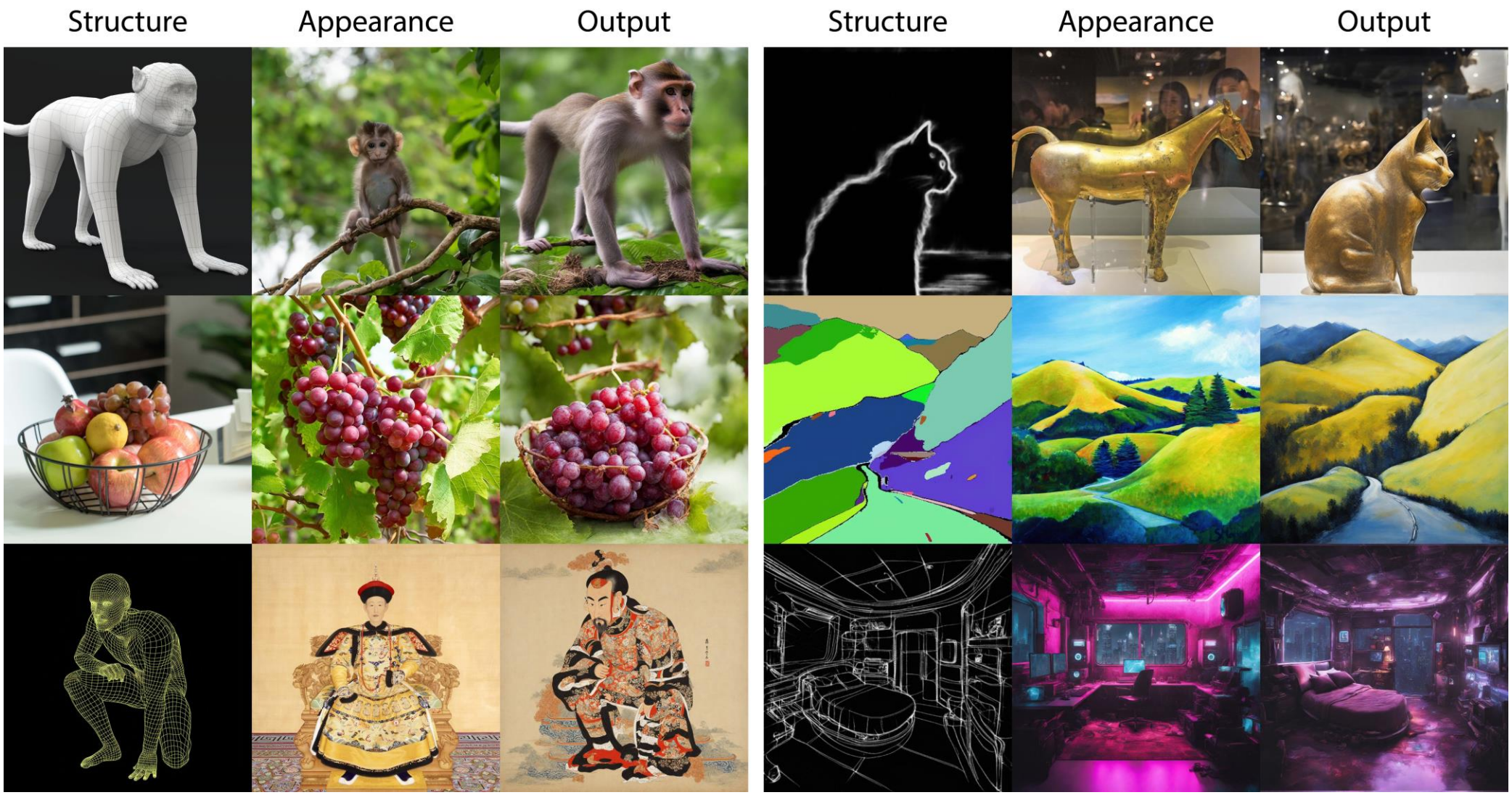


$t = 0.4T, \ldots, 0$


$\mathbf{x}_t^{\text{struct}}$
$\mathbf{x}_t^{\text{out}}$
$\mathbf{x}_t^{\text{app}}$
$\mathbf{x}_{t-1}^{\text{out}}$
Feature injection
Self-attention injection
Spatially-aware appearance transfer


Feature injection
Spatially-aware appearance transfer
Self-attention injection
$\mathbf{h}_{l,t}^{\text{out}}$
$\mathbf{h}_{l,t}^{\text{struct}}$
$\mathbf{h}_{l,t}^{\text{app}}$
norm
$\mathbf{W}_l^Q$
$\mathbf{Q}^{\text{out}}$
$\mathbf{W}_l^K$
$\mathbf{K}^{\text{app}}$
softmax
$\mathbf{A}$
$\mathbf{h}_{l,t}^{\text{app}\,2}$
$\mathbf{M}^2$
$\mathbf{S}$
$\mathbf{M}$
$\mathbf{K}^{\text{struct}}$
$\mathbf{Q}^{\text{struct}}$
$\mathbf{V}^{\text{out}}$
attention
$\mathbf{h}_{l,t}^{\text{out}}$

Structure and appearance control



Multi-subject controllable generation

Appearance

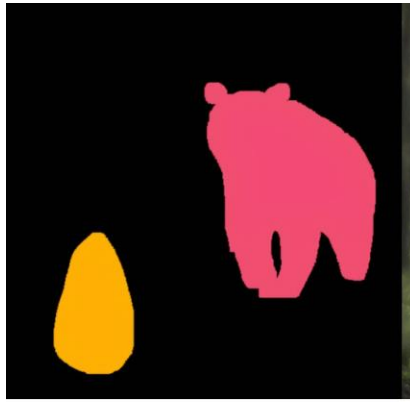
Structure

Multi-subject controllable generation

Appearance

Structure

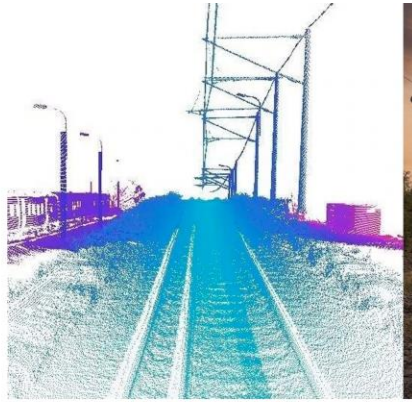
Prompt-driven conditional generation



Structure

a realistic photo of a bear and an avocado in a forest

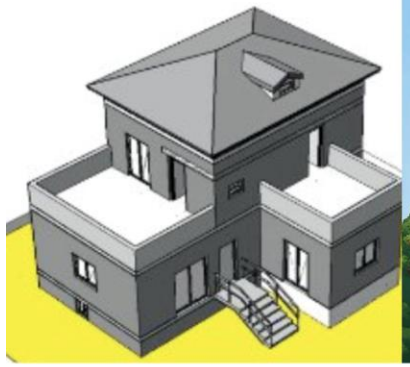
a painting of a tiger looking at a large white egg on a beach



Structure

a photo of a railway during sunset

a painting of a railway during the harsh winter



Structure

a video game pixel art of a mansion

a photo of a gingerbread house in space

Structure

a photo of a Karate man in a park

an embroidery of a man scuba diving in the ocean

Prompt-driven conditional generation

Structure

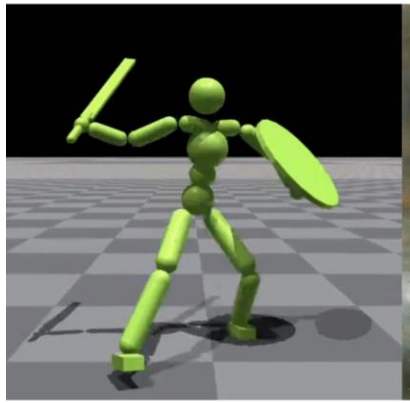
a cartoon of an evil goblin holding a piece of gold

a rough sketch of a kangaroo on top of a mountain

Structure

a photo of a city intersection at night, bird's eye view

a photo of a river during winter, bird's-eye view



Structure

an oil painting of a warrior holding a sword and shield in a river

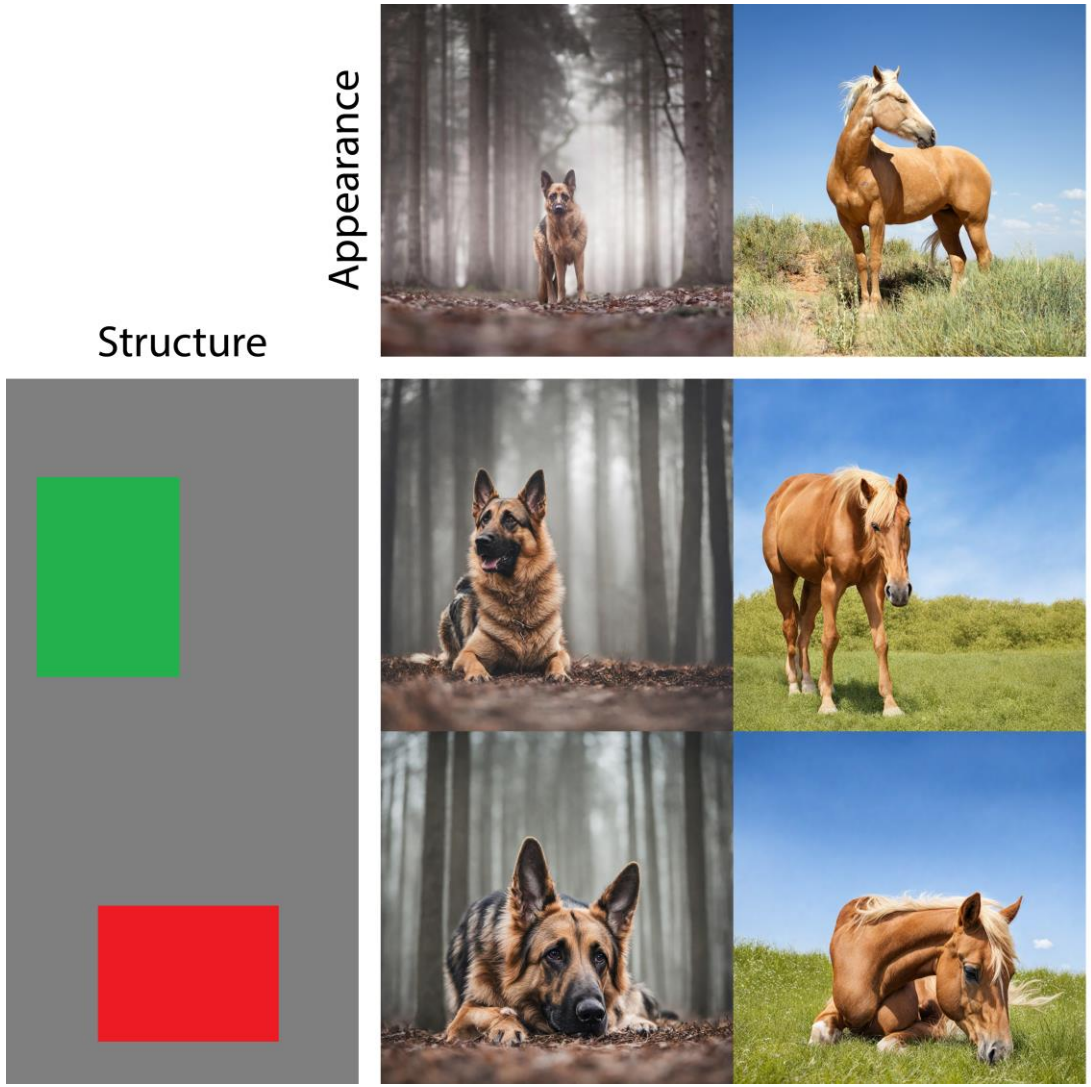
a photo of a robot in a Cyberpunk city holding weapons

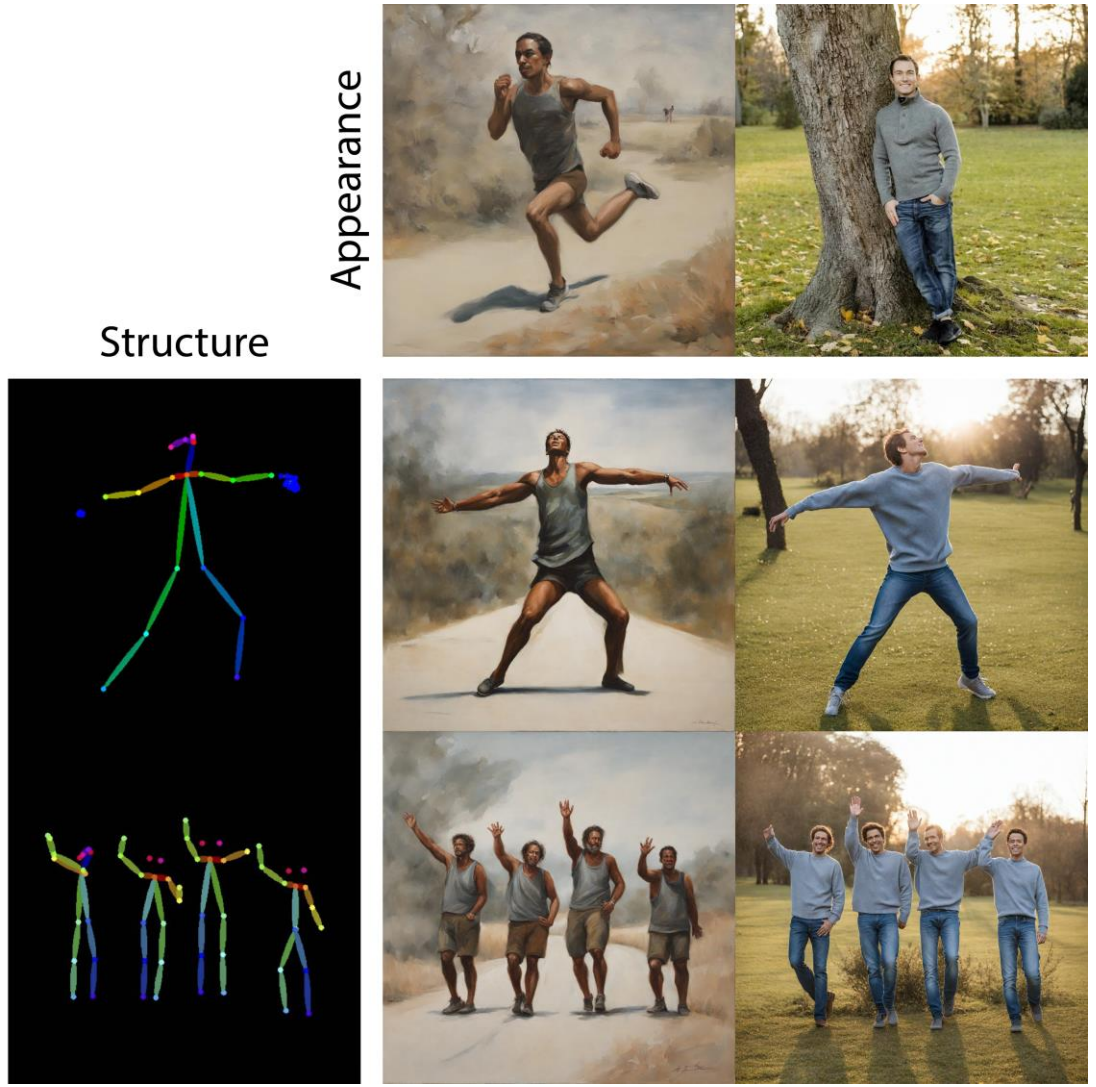
Structure

a photo of a mechanical wolf howling in a cave

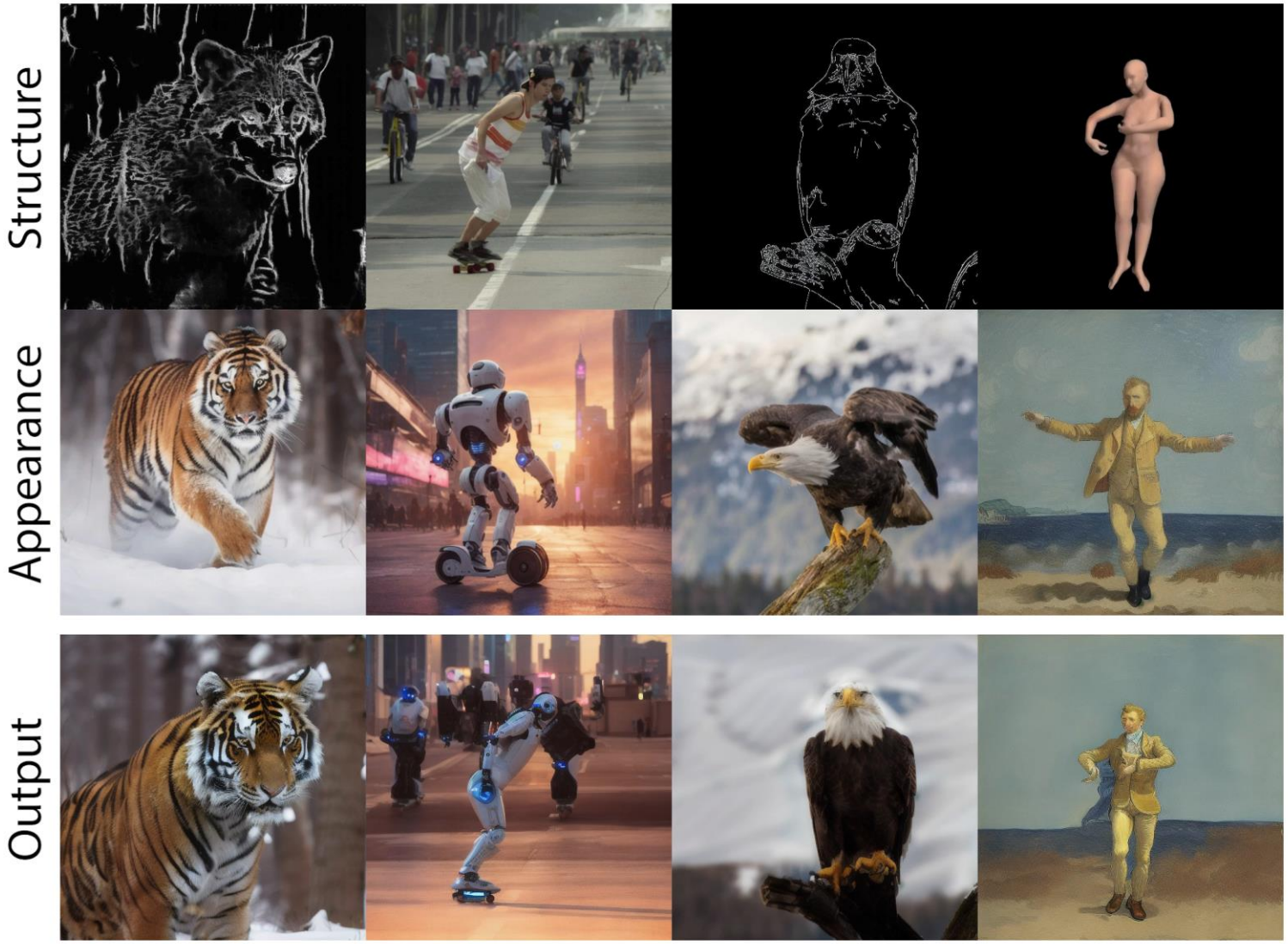
a cartoon of a wolf howling at the moon

Higher-level conditions





Extension to text-to-video generation



Inference efficiency

Method	Training	Base model	Inference time (s)	Peak GPU memory usage (GiB)
Splicing ViT Features	✓	Custom U-Net	1557.09	3.95
Uni-ControlNet	✓	SD v1.5	6.96	7.36
ControlNet + IP Adapter	✓	SDXL v1.0	6.21	18.09
T2I-Adapter + IP-Adapter	✓	SDXL v1.0	4.37	13.28
Cross-Image Attention	✗	SD v1.5	42.80	8.85
FreeControl	✗	SDXL v1.0	378.89	44.34
Ctrl-X (ours)	✗	SDXL v1.0	10.91	11.51

Qualitative evaluation

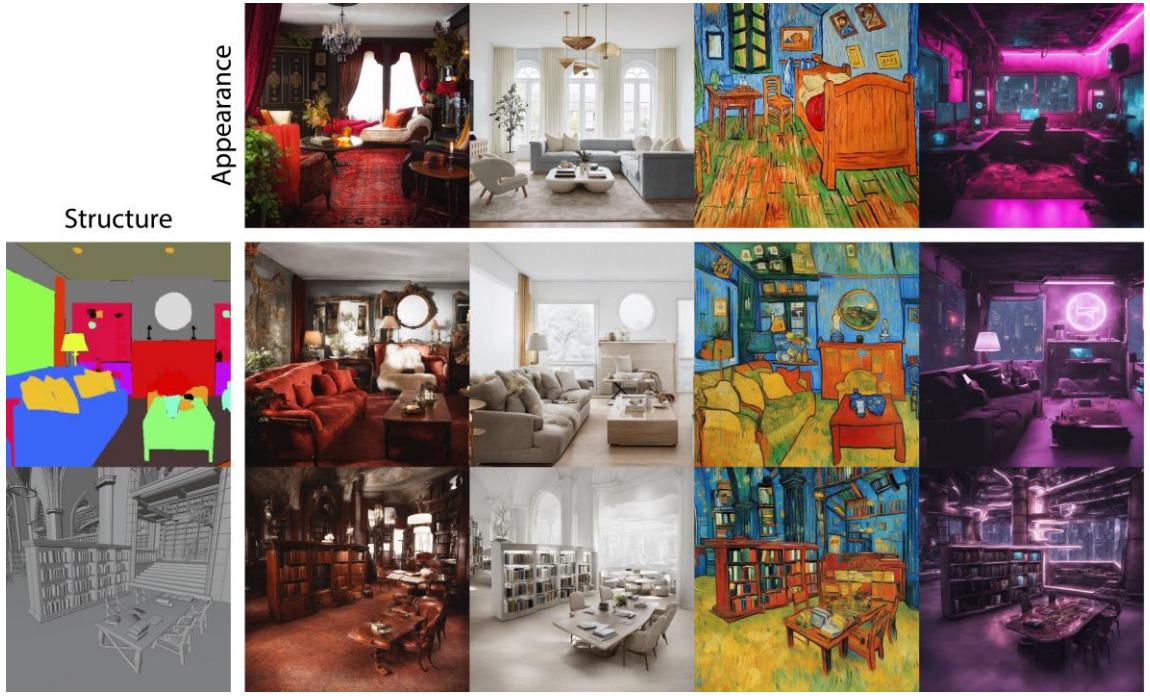
Method	Training	Natural image		ControlNet-supported		New condition	
		Self-sim ↓	DINO-I ↑	Self-sim ↓	DINO-I ↑	Self-sim ↓	DINO-I ↑
Splicing ViT Features	✓	0.030	0.907	0.043	0.864	0.037	0.866
Uni-ControlNet	✓	**0.045**	0.555	**0.096**	0.574	**0.073**	0.506
ControlNet + IP-Adapter	✓	0.068	*0.656*	0.136	*0.686*	0.139	*0.667*
T2I-Adapter + IP-Adapter	✓	*0.055*	0.603	0.118	0.586	0.109	0.566
Cross-Image Attention	✗	0.145	0.651	0.196	0.510	0.175	0.570
FreeControl	✗	0.058	0.572	*0.101*	0.585	*0.089*	0.567
Ctrl-X (ours)	✗	0.057	**0.686**	0.121	**0.698**	0.109	**0.676**

User study: Average user preference

Method	Training	Result quality ↑	Structure fidelity ↑	Appearance fidelity ↑	Overall fidelity ↑
Splicing ViT Features	✓	95%	87%	56%	78%
Uni-ControlNet	✓	86%	17%	96%	74%
ControlNet + IP-Adapter	✓	46%	61%	41%	50%
T2I-Adapter + IP-Adapter	✓	74%	53%	67%	58%
Cross-Image Attention	✗	95%	83%	83%	83%
FreeControl	✗	64%	48%	79%	74%
Ctrl-X (ours)	✗	-	-	-	-

Ctrl-X

*A **training-free** and **guidance-free** method for **structure** and **appearance** control of text-to-image generation*



Thank you :D

UCLA NVIDIA

Wednesday, December 11th
4:30–7:30 p.m. PST, Poster Session 2

Paper and code