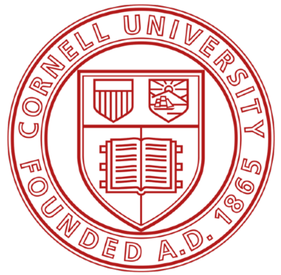
Lecture 15: Diffusion models - Part 2

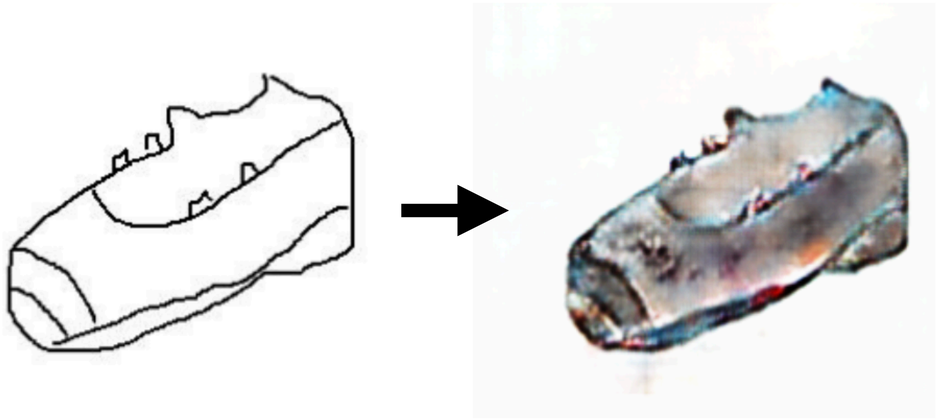
CS 5670: Introduction to Computer Vision



Announcements

- Take-home exam: Mon. 12pm - Weds. 12pm
  - Only expected to take a few hours.
  - No lectures, discussion, or office hours during exam time (private Ed Discussion questions only).
  - Theoretical questions and a bit of coding
  - See practice problems that we posted online.
- Guest lecture next Weds.: Xuanchen Lu
- Highly encourage you to do the midterm evaluation!

PS4: Generative models





“an oil“an oil painting of an old man”mpfire”

Image translation with GANs

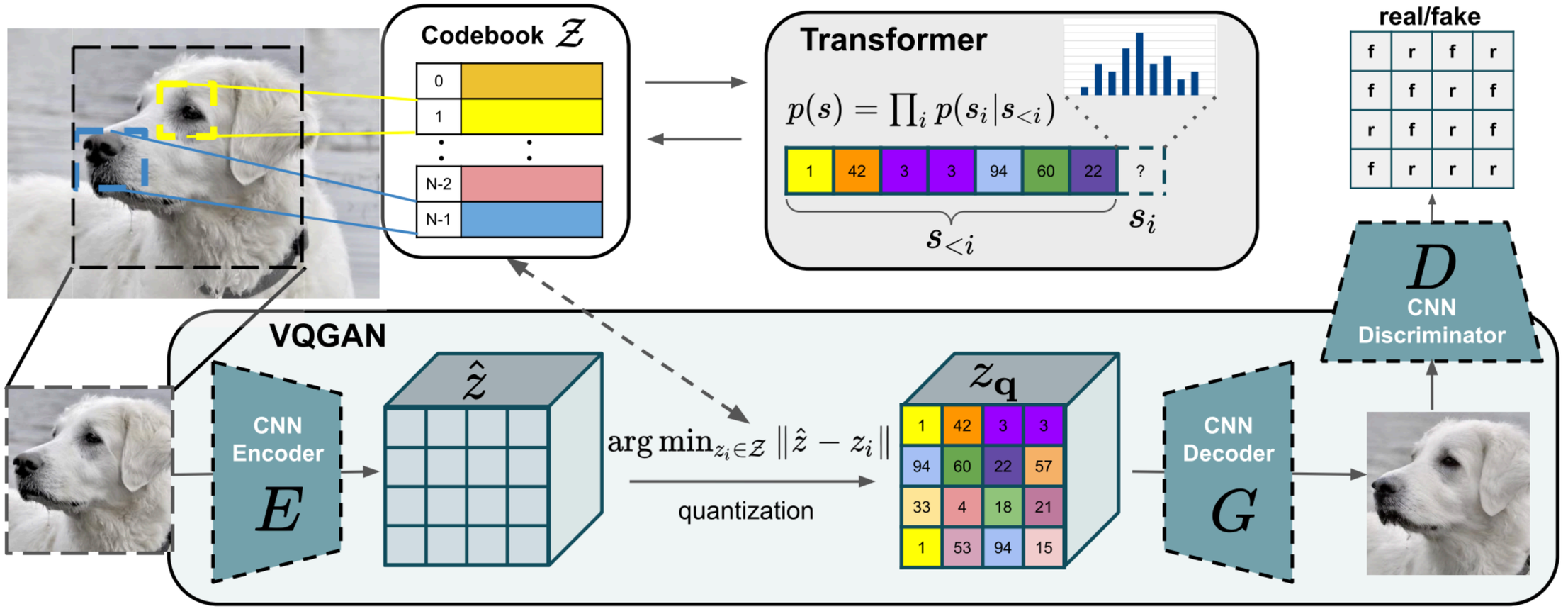
Image manipulation with diffusion

Which of the following is often a challenge of autoregressive image models?



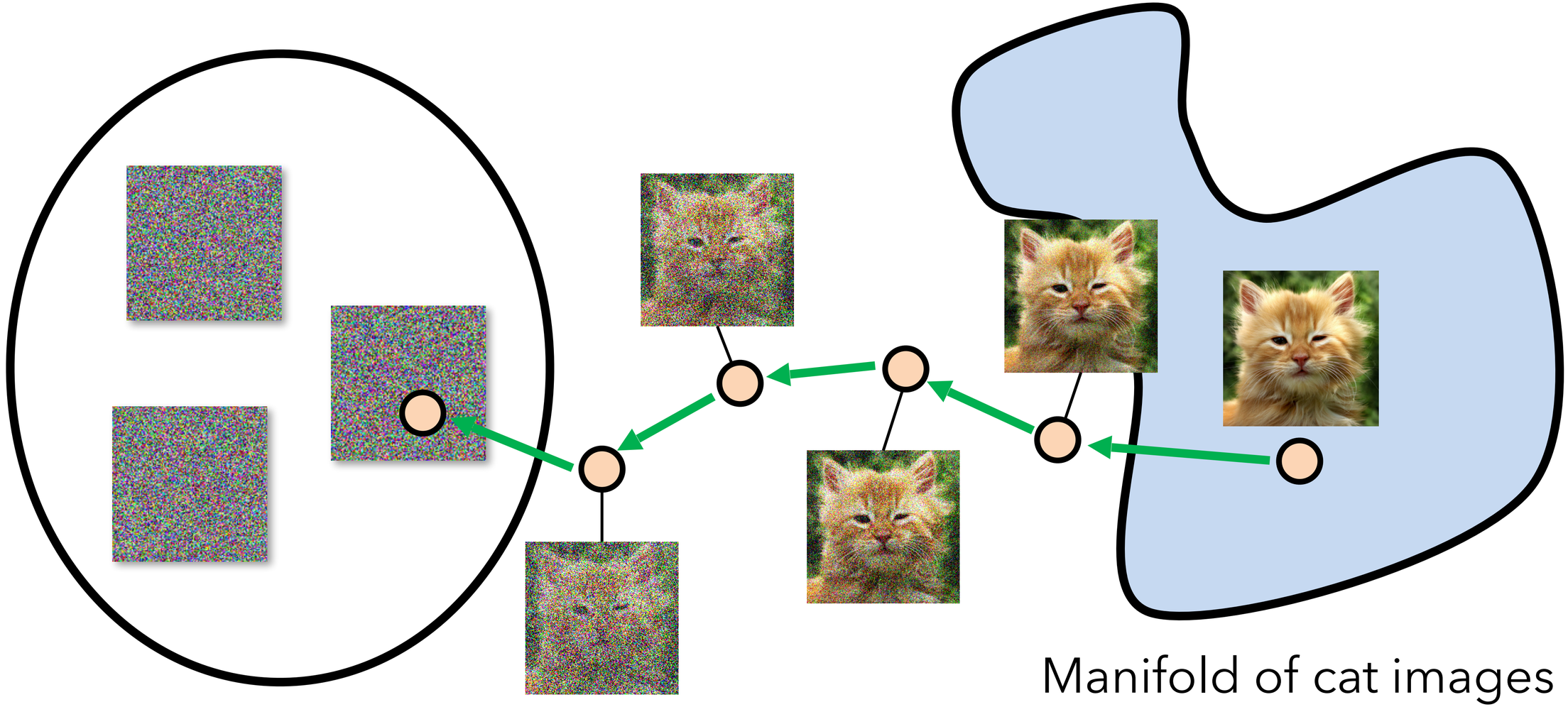
A. Mode collapse
B. Defining discrete tokens for image patches
C. Unstable training due to minimax game

Last class: autoregressive model in latent space



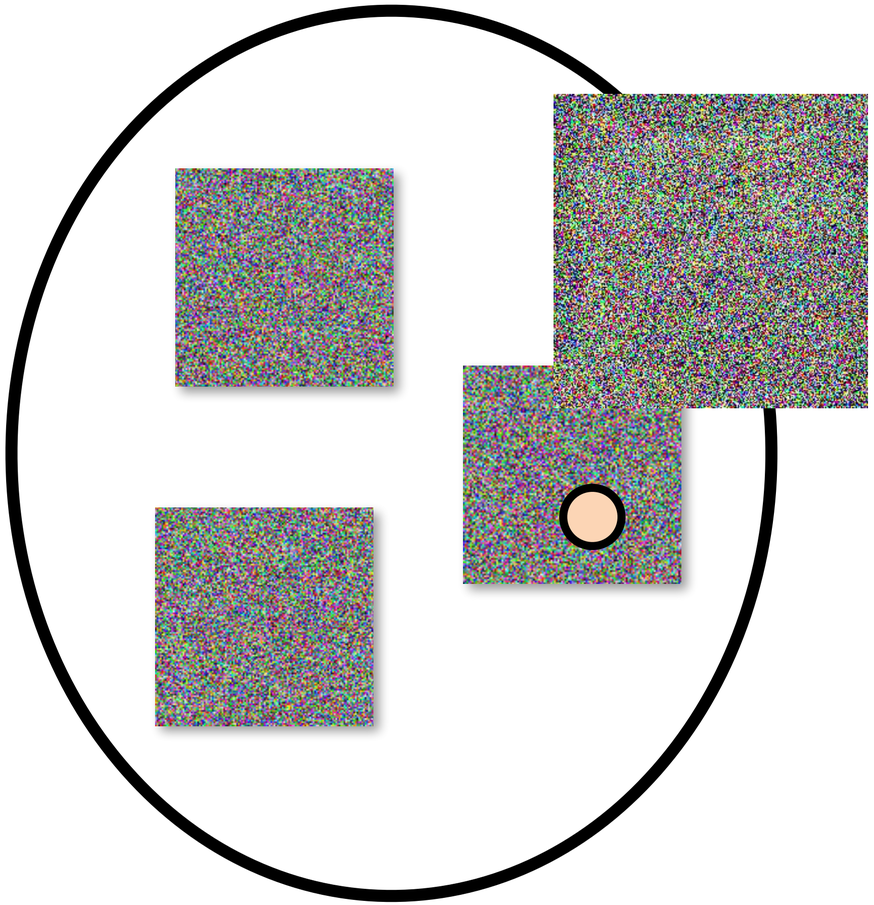
[Esser et al., "Taming Transformers for High-Resolution Image Synthesis", 2021]

Last class: diffusion models

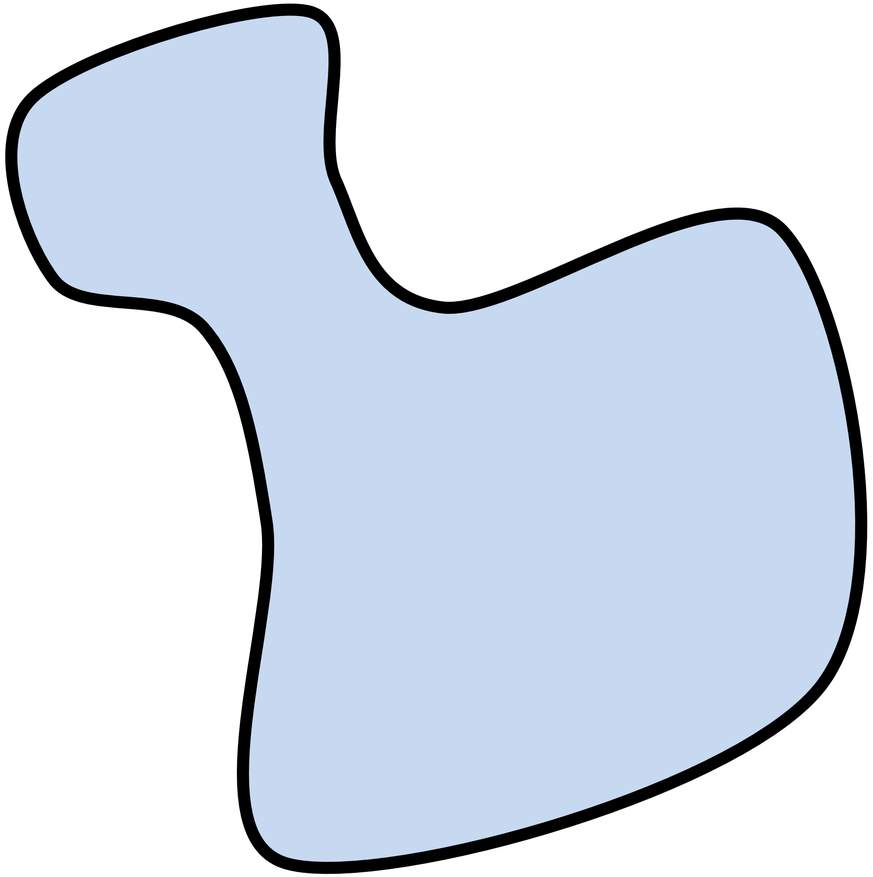


Slide source: N. Snavely & S. Seitz

Last class: diffusion models



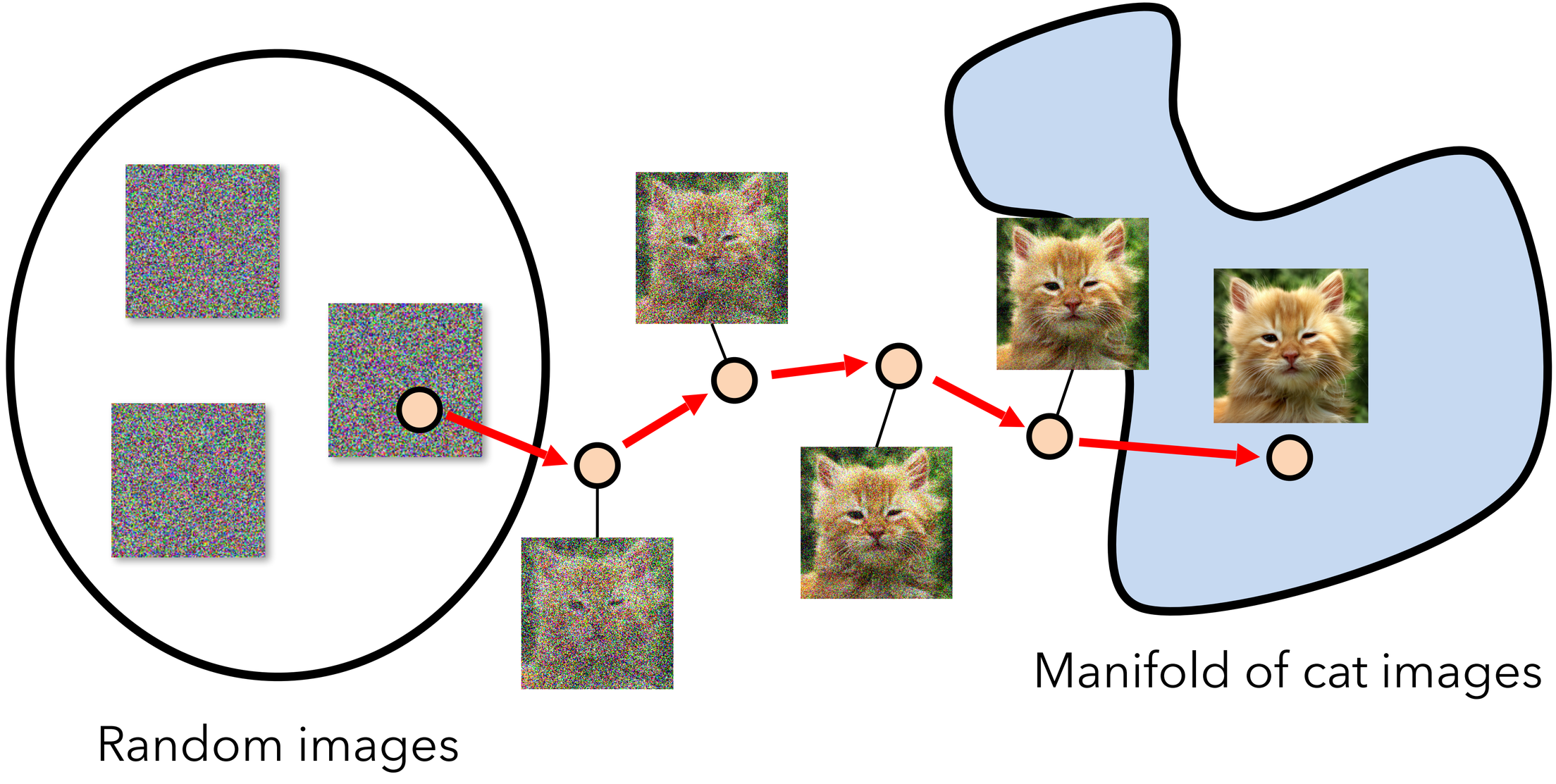
Random images



Manifold of cat images

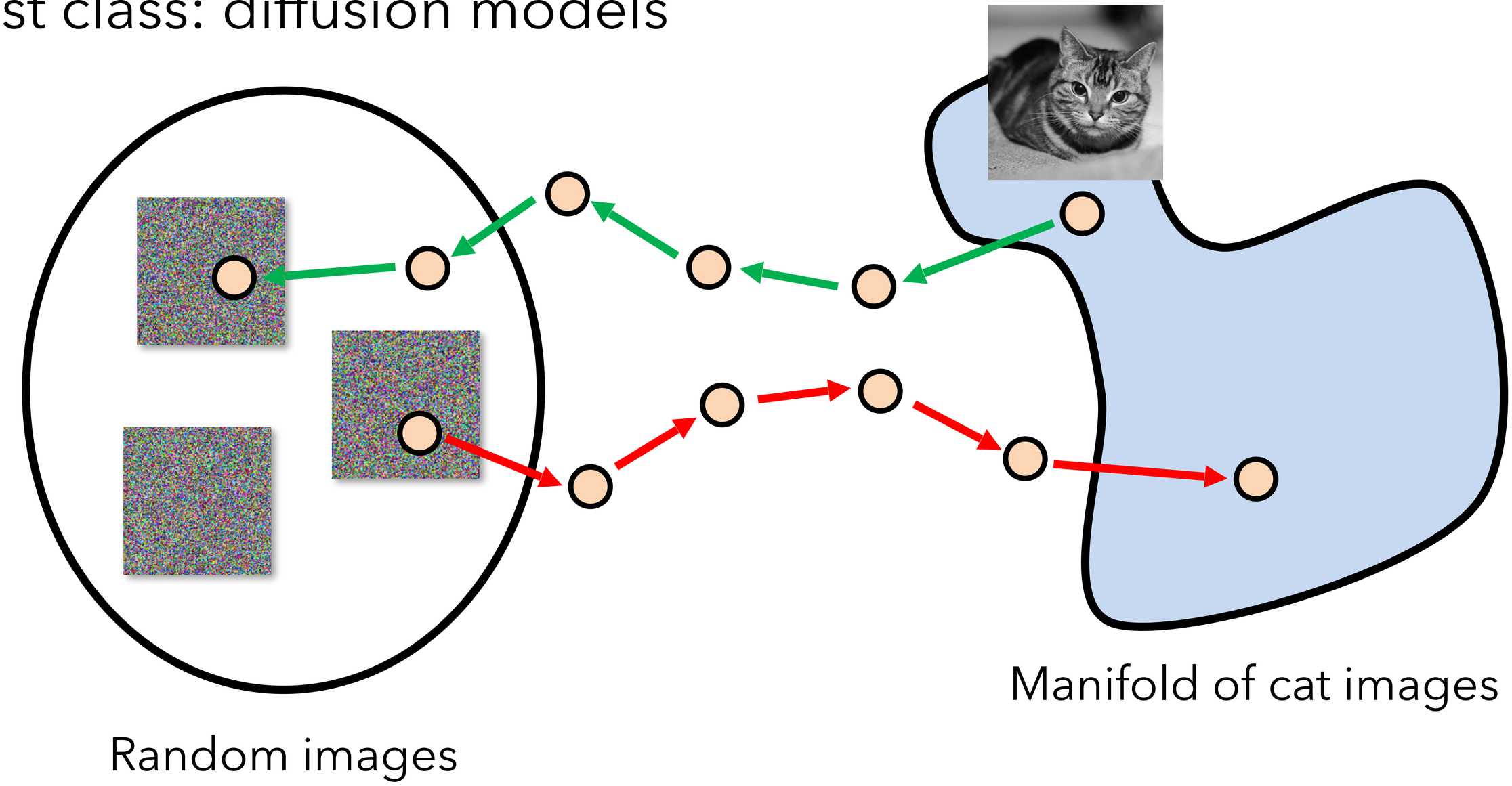
Slide source: N. Snavely & S. Seitz

Last class: diffusion models



Slide source: N. Snavely & S. Seitz

Last class: diffusion models

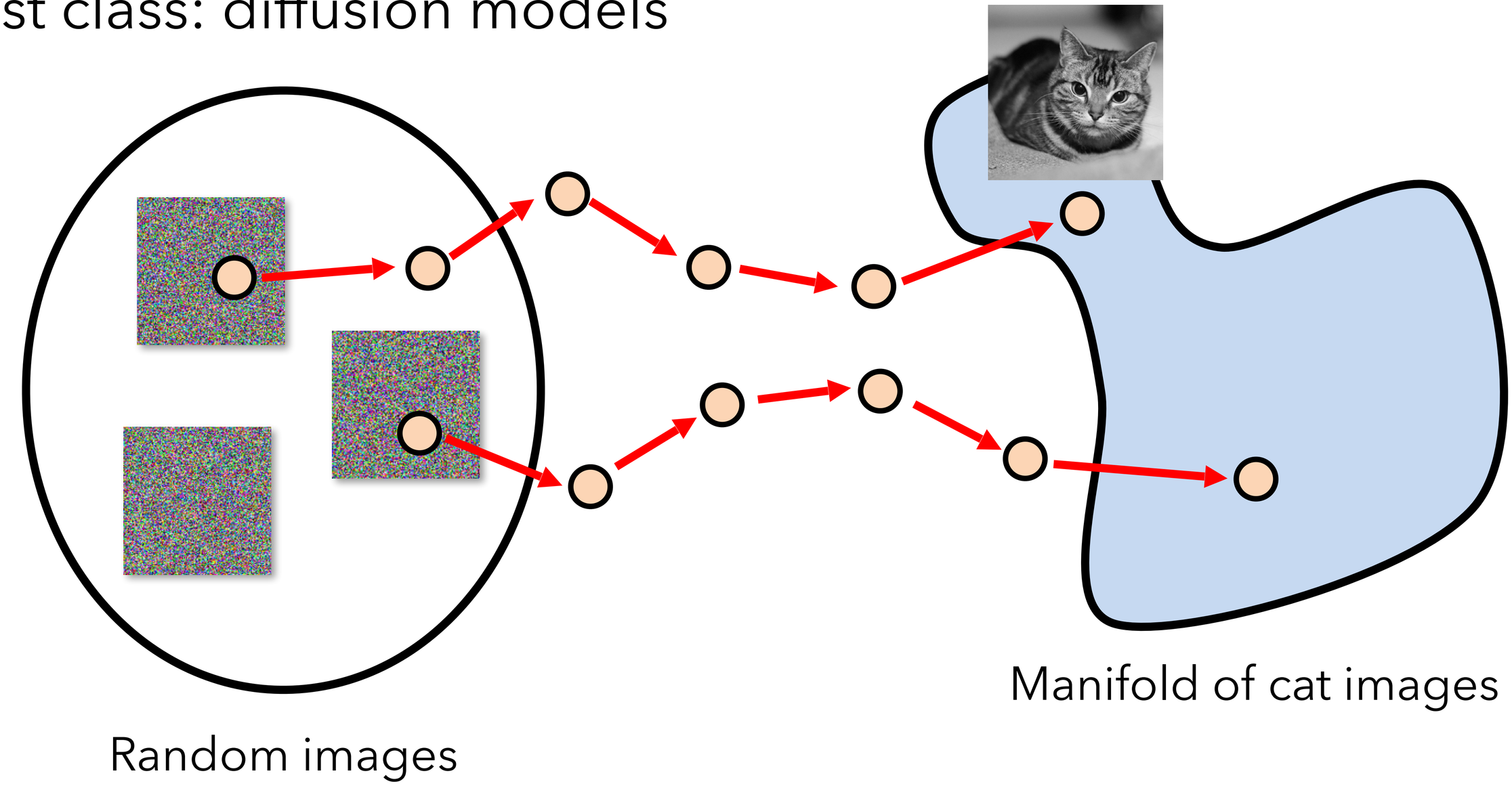


Random images

Manifold of cat images

Slide source: N. Snavely & S. Seitz

Last class: diffusion models

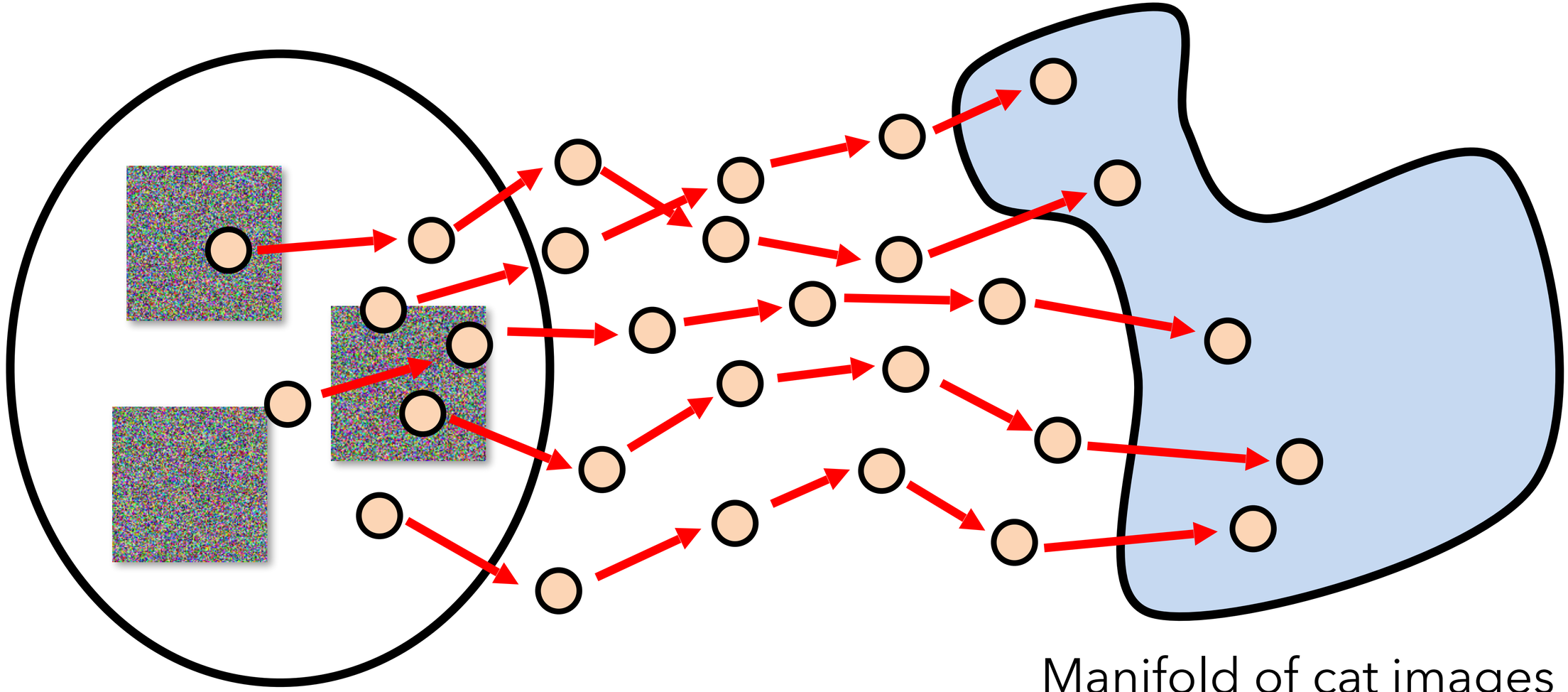


Random images

Manifold of cat images

Slide source: N. Snavely & S. Seitz

Last class: diffusion models

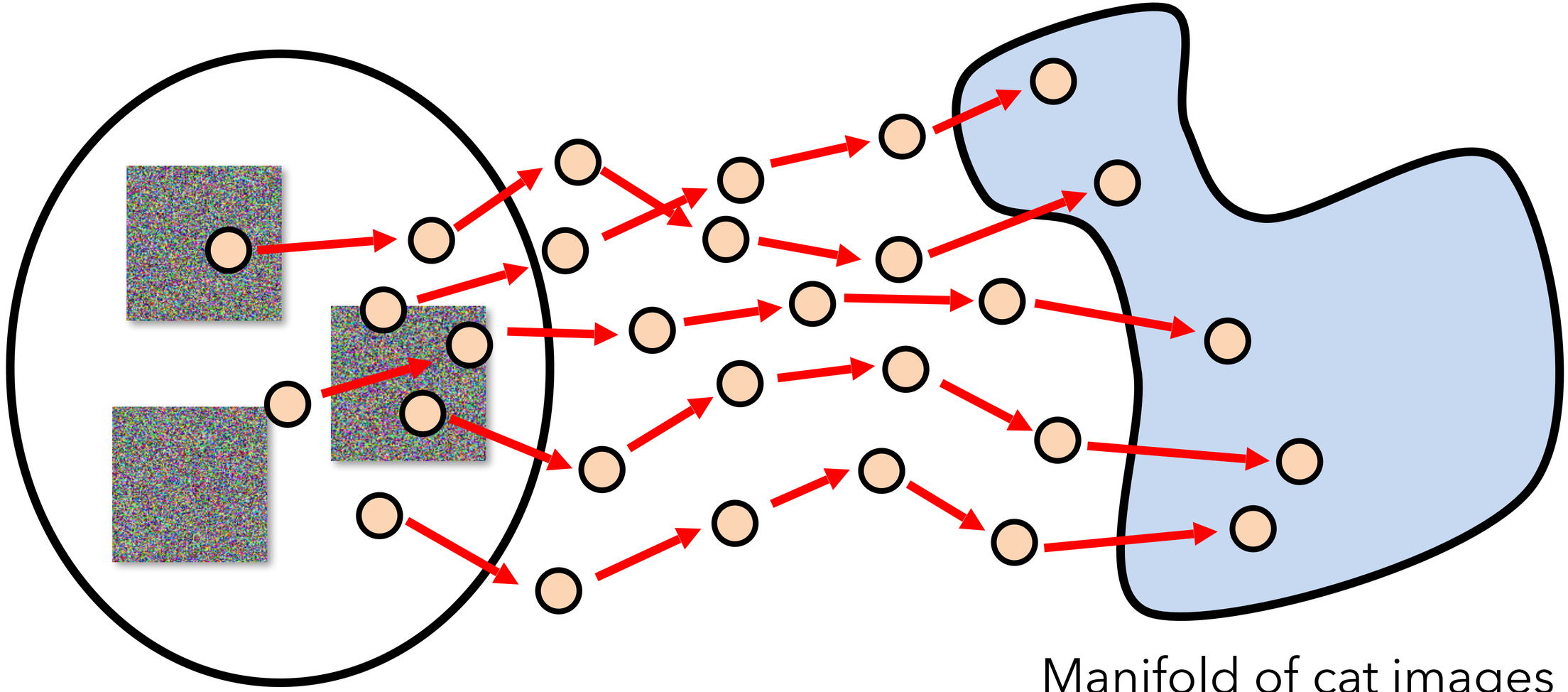


Random images

Manifold of cat images

Slide source: N. Snavely & S. Seitz

Last class: diffusion models



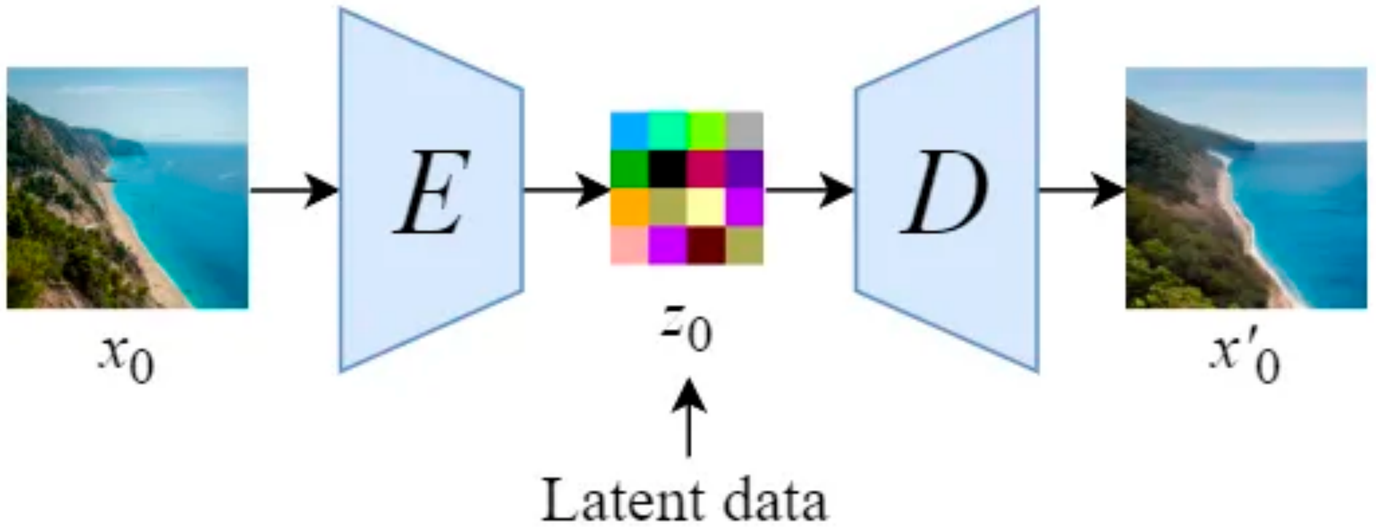
Random images

Manifold of cat images

Slide source: N. Snavely & S. Seitz

Latent diffusion models

- Use a separate *encoder* and *decoder* to convert images to and from a lower-dimensional latent space, run diffusion in latent space.
- Faster and focuses more on “perceptually important” details.

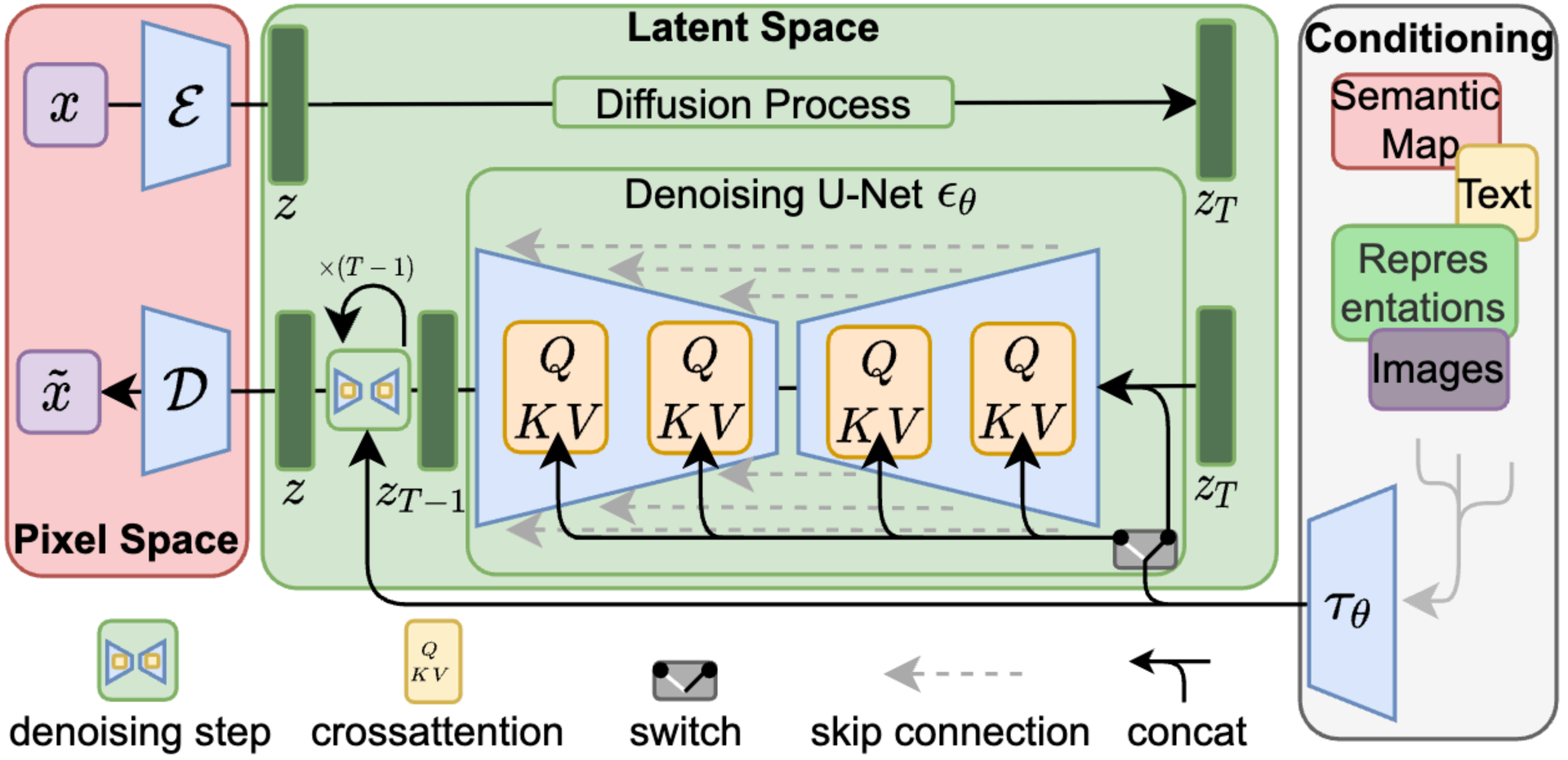


R. Rombach et al. High-Resolution Image Synthesis with Latent Diffusion Models. CVPR 2022

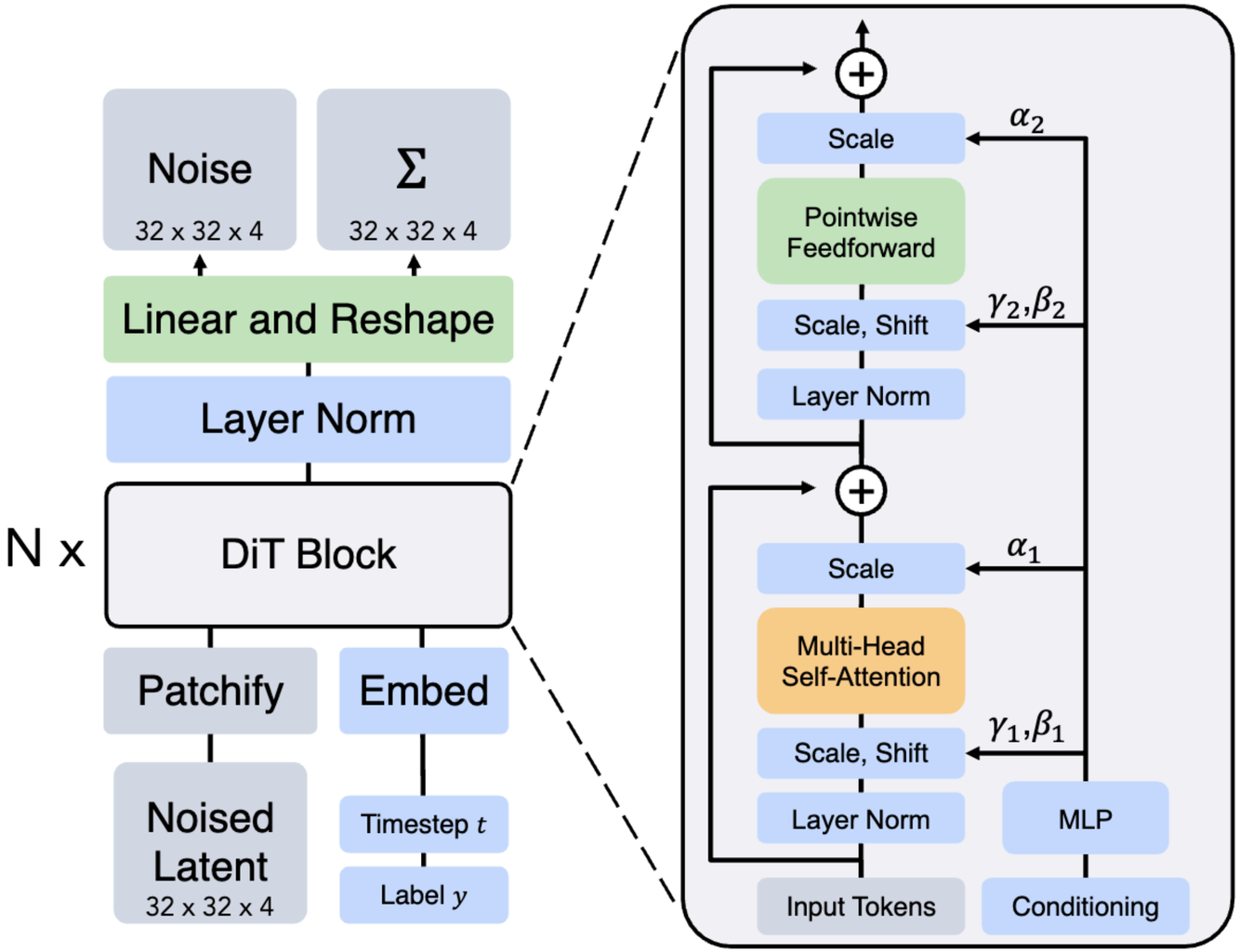
Figure source

Slide source: S. Lazebnik.

Architectures for latent diffusion: U-Net



Diffusion transformers (DiT)



Latent Diffusion Transformer

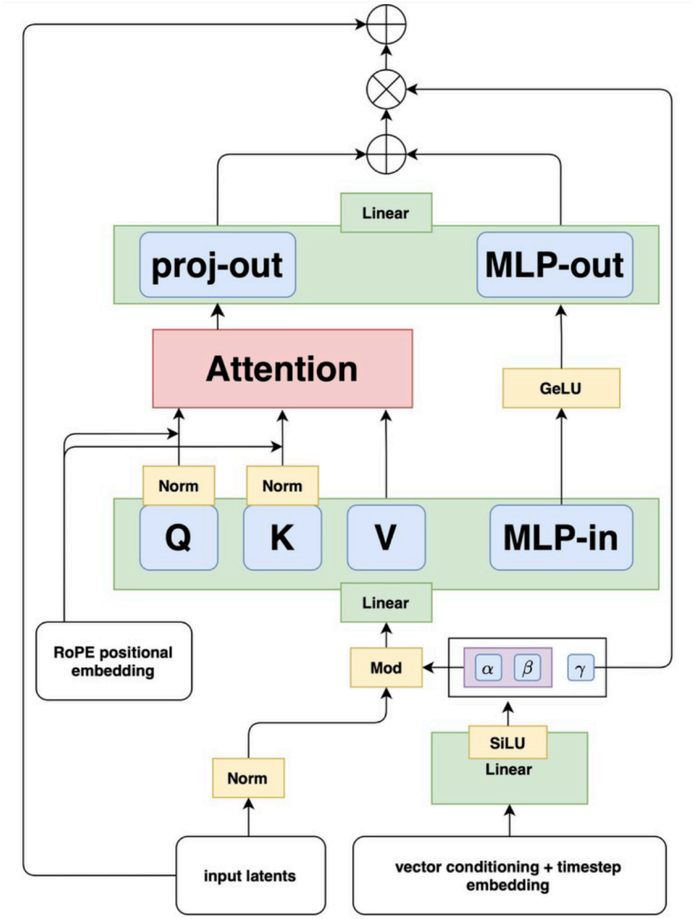
DiT Block with adaLN-Zero



Samples on ImageNet in 2023

[Peebles and Xie, 2023]

Diffusion transformers today (DiT)



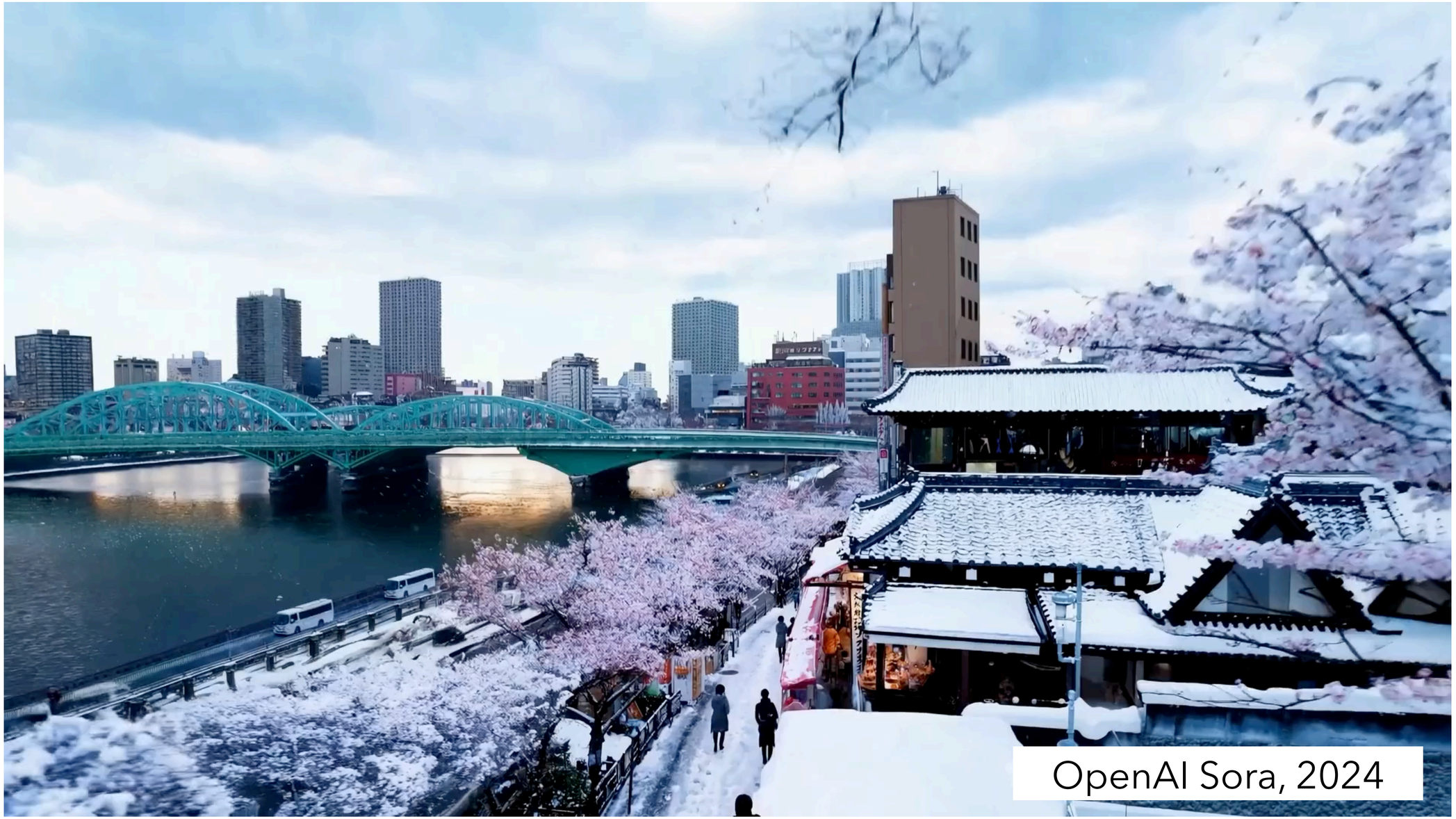
Scaled up DiT (12B parameters)



FLUX.1

Source: [FLUX.1 Kontext, 2024]

Video models, too



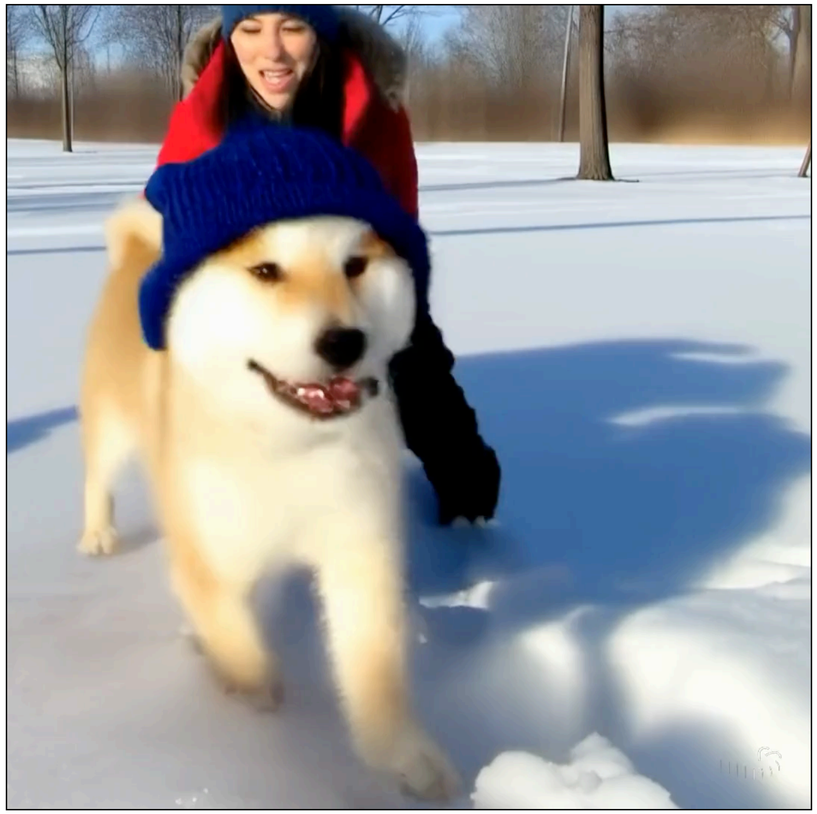
Video models, too



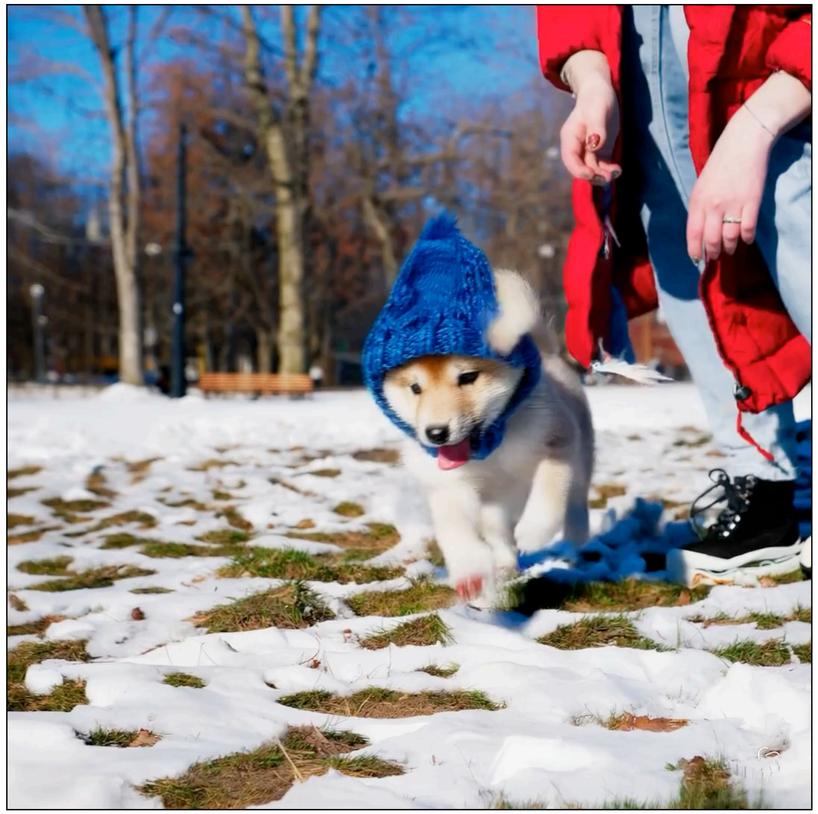
Generations as training progresses



1x training



4x training



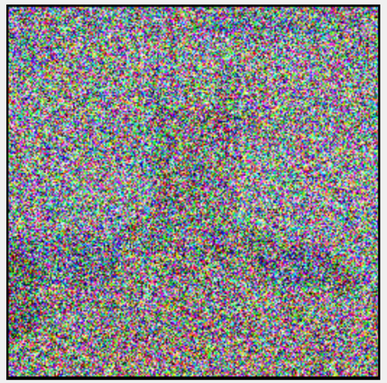
32x training

OpenAI Sora, 2024

Sketch to photo

Add noise to a sketch

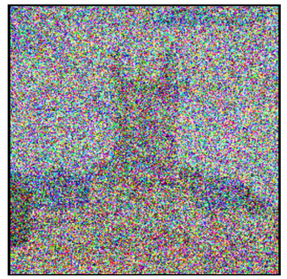




$\mathbf{x}_t$

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \text{ for } \epsilon \sim N(0,1)$$

Denoise from t to 0







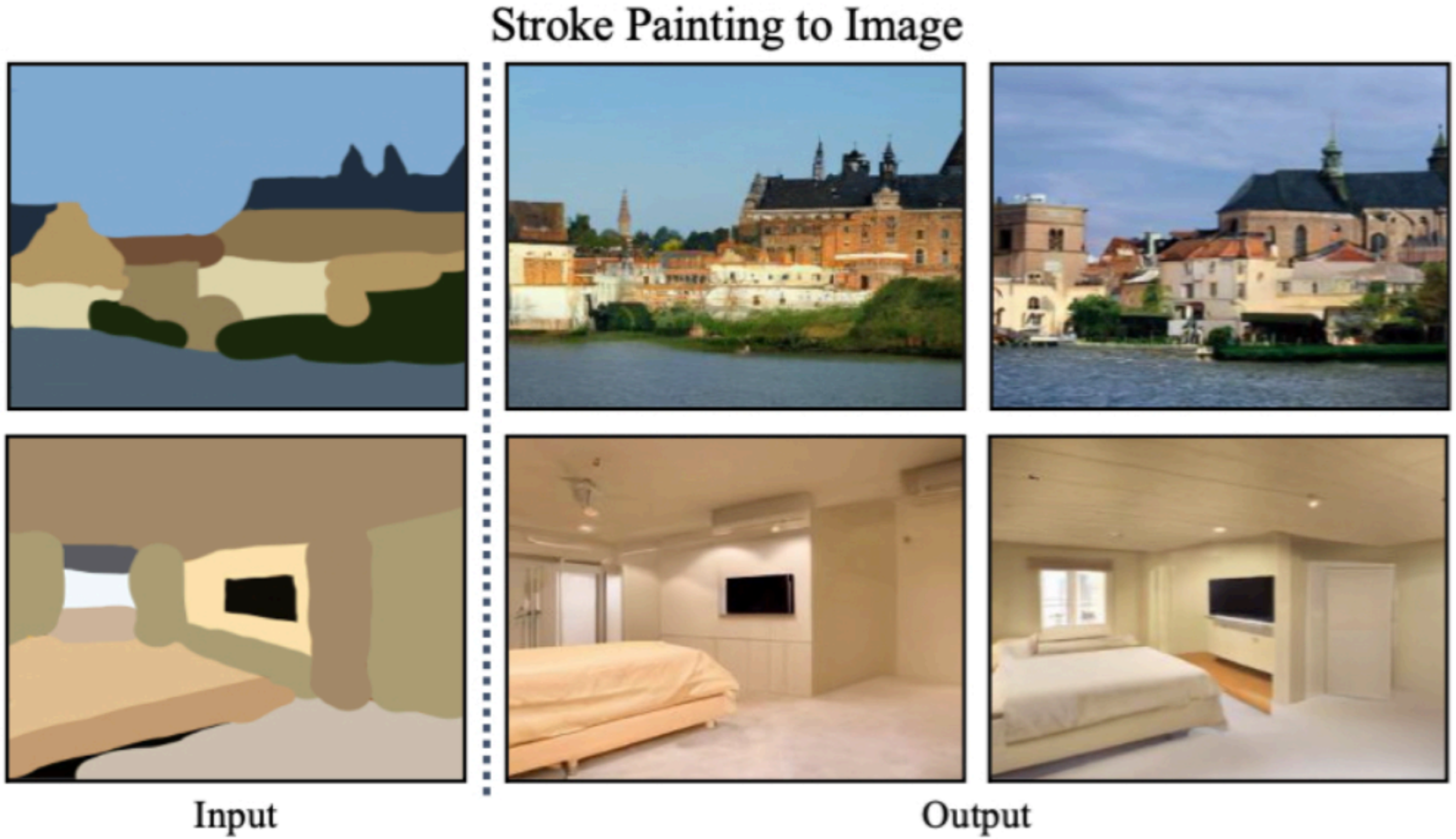


$\mathbf{x}_0$

Denoise using diffusion model trained on *real* images.

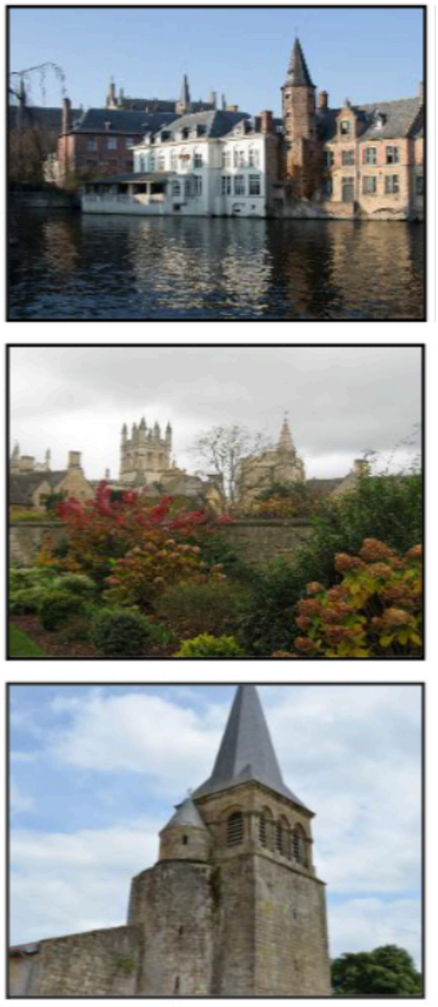
[Meng et al. "SDEdit", 2022]

Sketch to photo



Source: [Meng et al. “SDEdit”, 2022]

Image-to-image translation with SDEdit



Source: [Meng et al. “SDEdit”, 2022]

Video-to-video translation with SDEdit and Sora



original generated video



rewrite the video in a pixel art style

Video-to-video translation with SDEdit and Sora



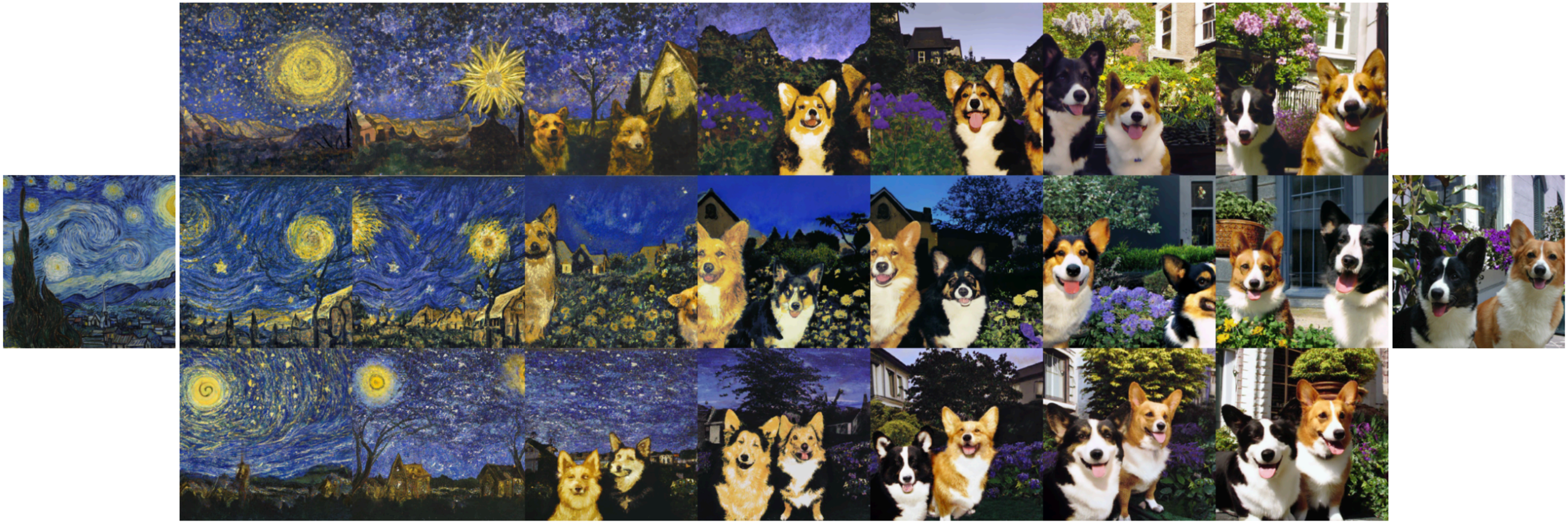
original generated video



change the video to a medieval theme

Image interpolation

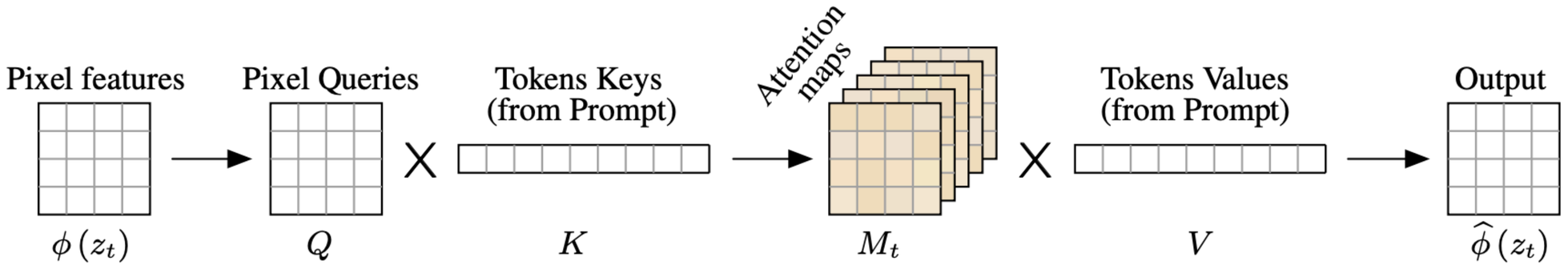
Interpolate between CLIP features, generating an image at each point.

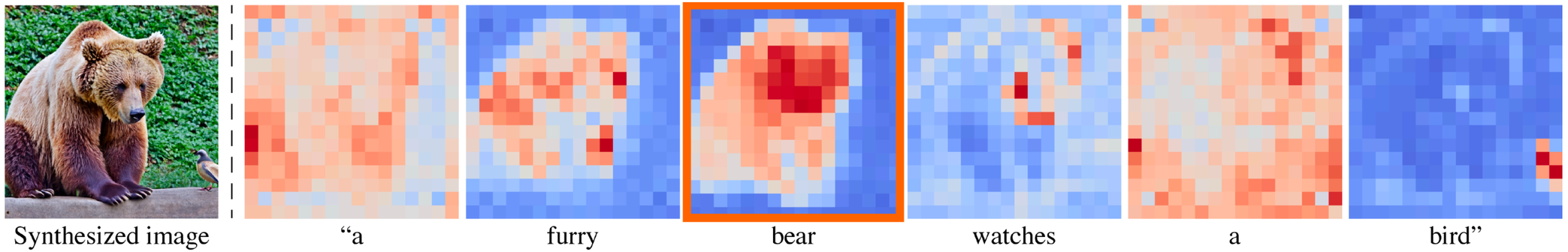


[Ramesh et al., “DALL-E 2”]

Network visualization

Empirical observation: cross-attention between text and image often conveys style, content, and structure.





Average cross–attention maps across all timestamps

Hertz et al. “Prompt-to-Prompt Image Editing with Cross-Attention Control”



Change text prompt, use same random seed.

Hertz et al. “Prompt-to-Prompt Image Editing with Cross-Attention Control”



Change text prompt, freeze random seed, and *freeze attention maps*.

Hertz et al. “Prompt-to-Prompt Image Editing with Cross-Attention Control”

Prompt-to-Prompt



Freeze the attention map for "apple" or "basket" during generation.

Hertz et al. "Prompt-to-Prompt Image Editing with Cross-Attention Control"

Prompt-to-Prompt

"Photo of a cat riding on a bicycle."



source image



cat → dog



cat → chicken



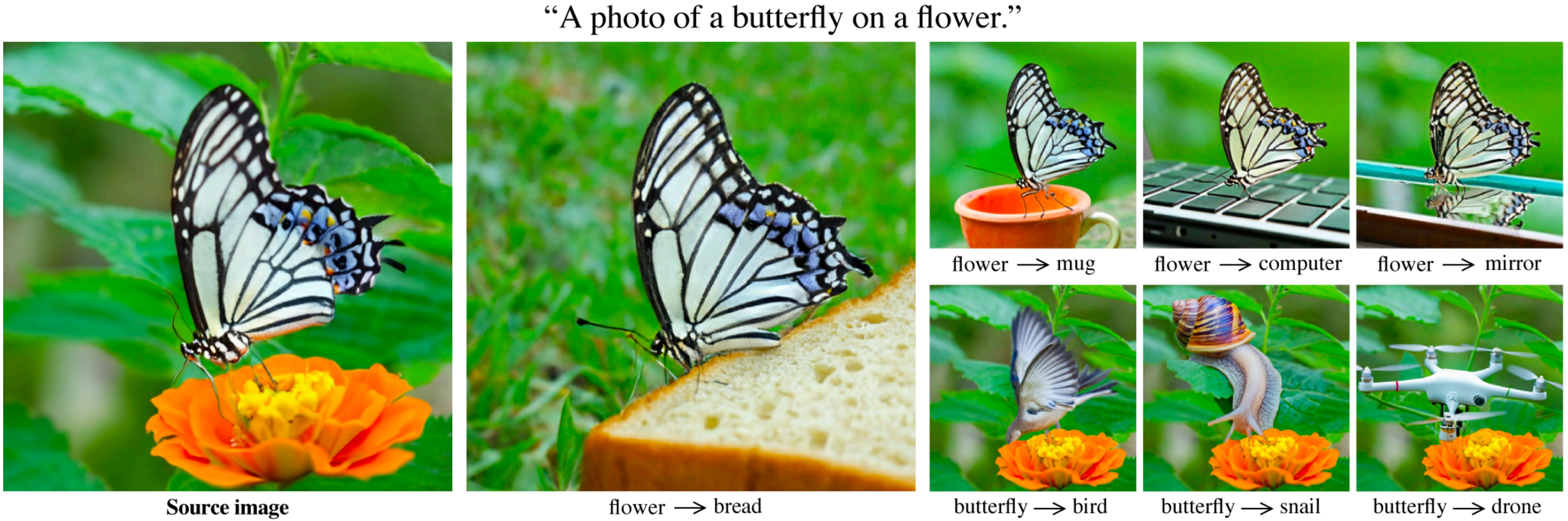
cat → squirrel



cat → elephant

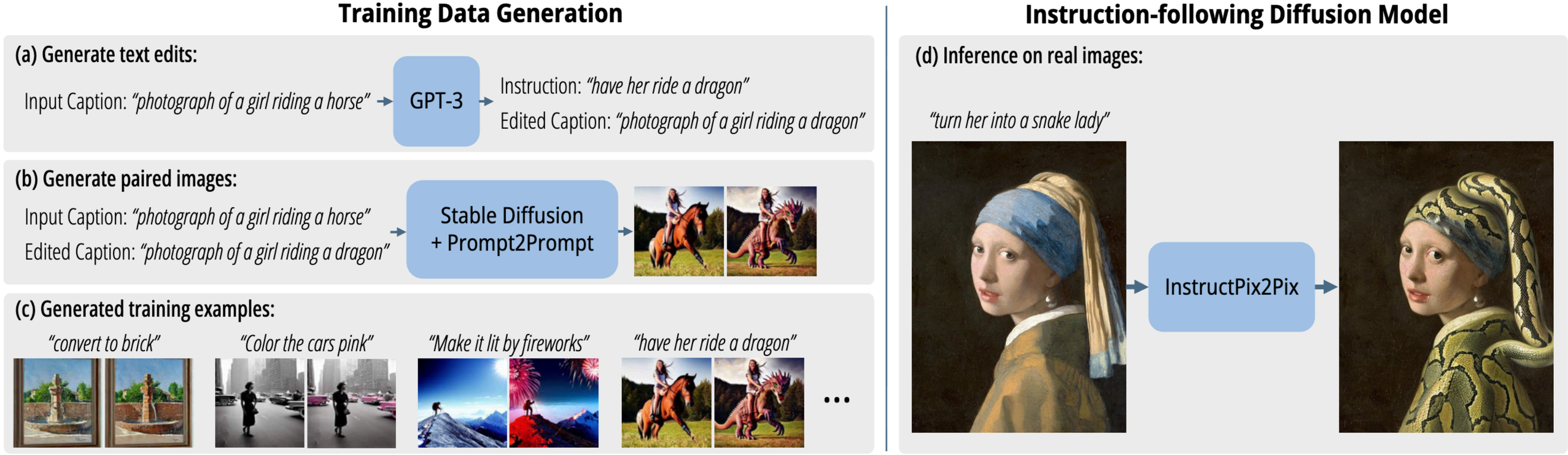
Hertz et al. "Prompt-to-Prompt Image Editing with Cross-Attention Control"

Prompt-to-Prompt

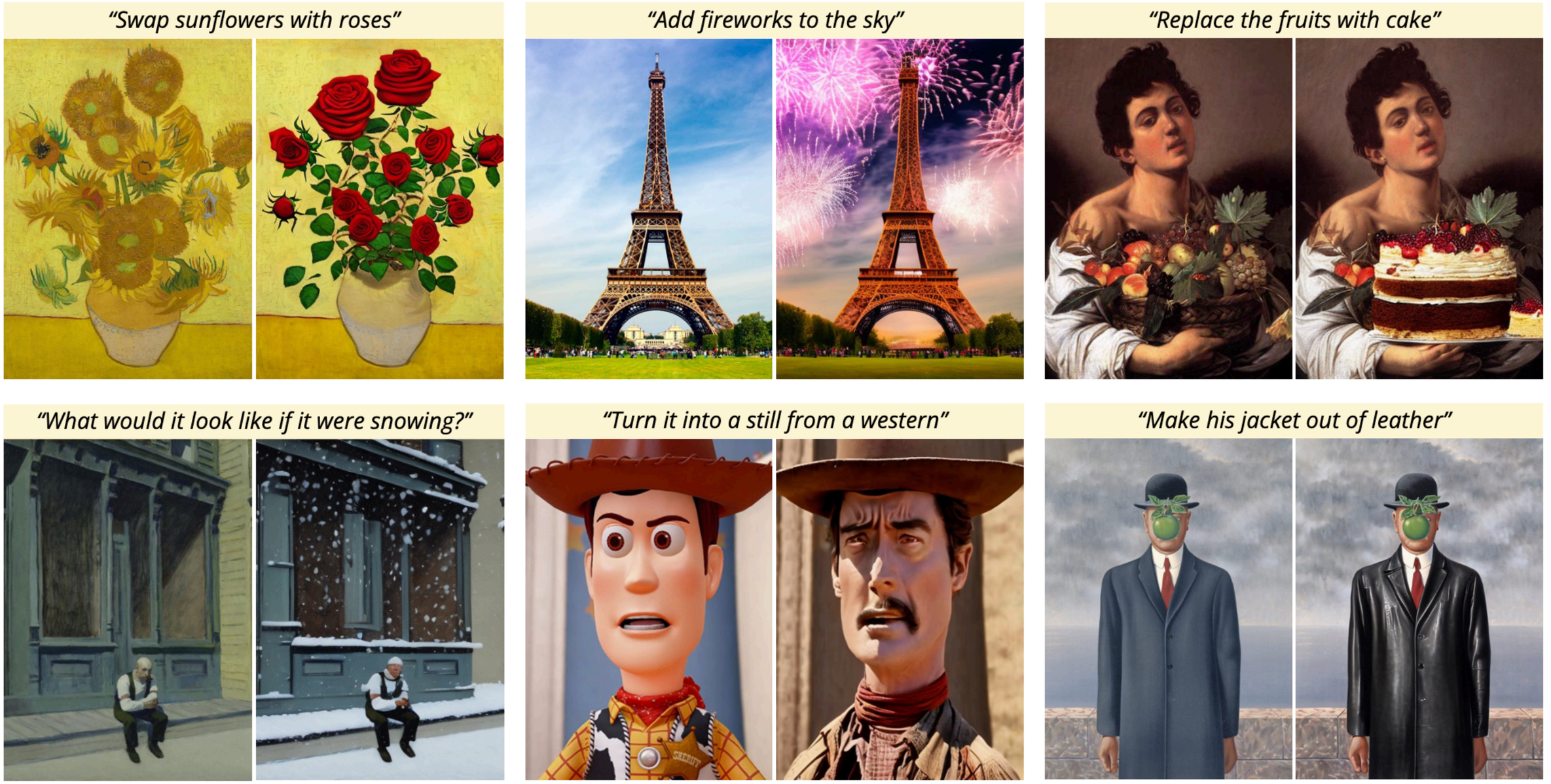


Hertz et al. "Prompt-to-Prompt Image Editing with Cross-Attention Control"

Learning to follow image editing instructions



[Brooks et al., "InstructPix2Pix", 2023]



[Brooks et al., "InstructPix2Pix", 2023]

Inpainting



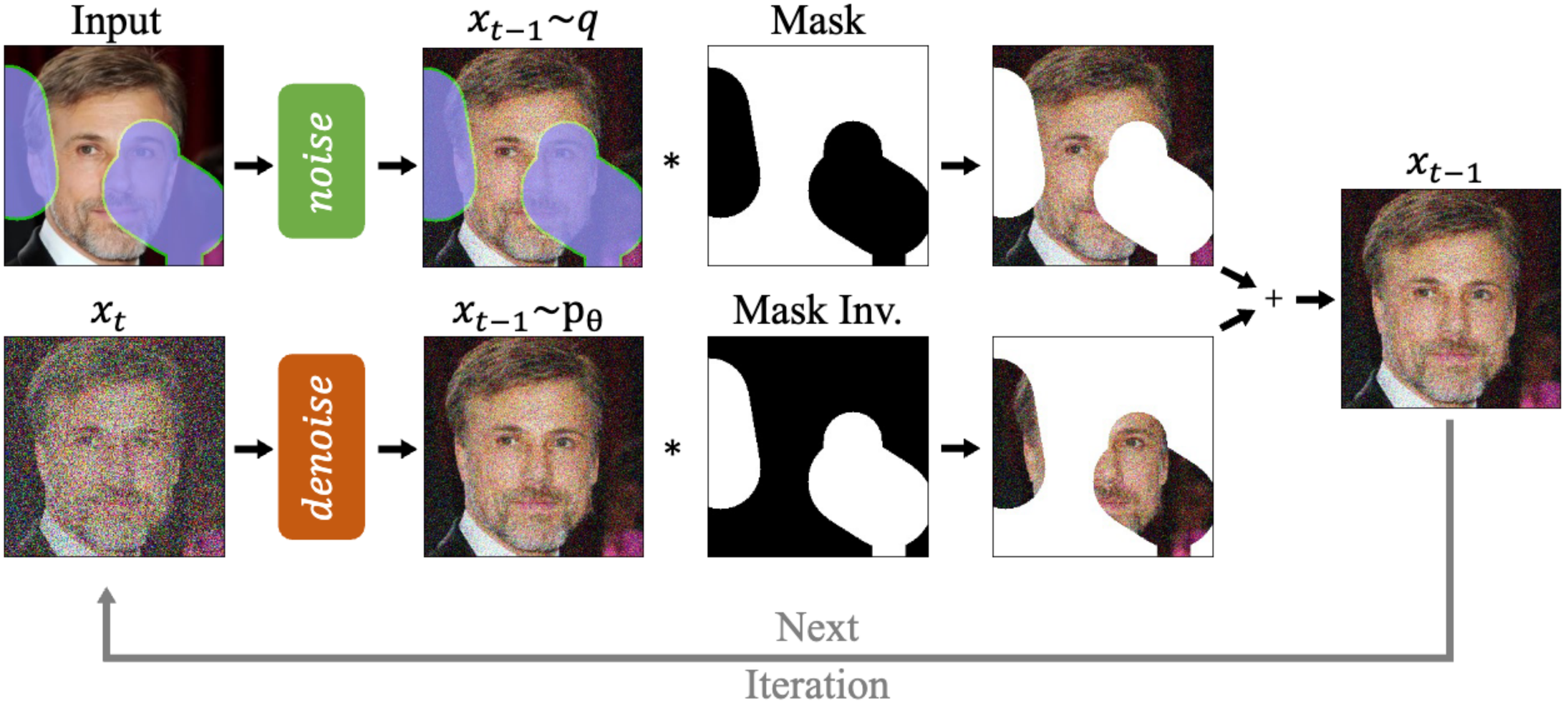
Input

Samples

Source: [Lugmayr et al., RePaint]

Inpainting

Constrain the pixels each iteration.



Source: [Lugmayr et al., RePaint]

Getting inpainting for free!

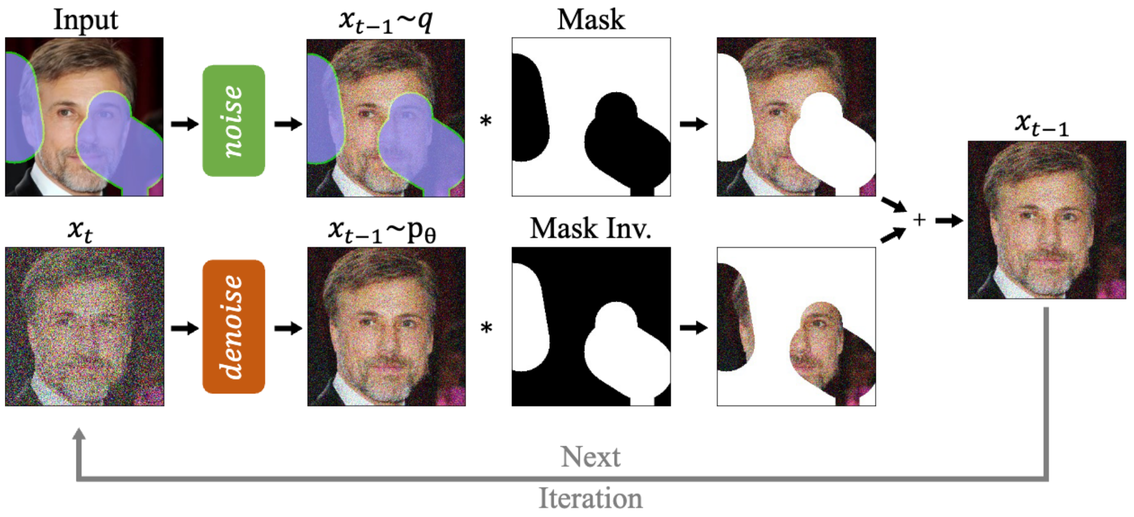
Pseudocode:

$$x_{t-1}^{\text{unknown}} \sim \mathcal{N}(\mu_\theta(x_t, t), \Sigma_\theta(x_t, t))$$

$$x_{t-1}^{\text{known}} \sim \mathcal{N}(\sqrt{\bar{\alpha}_t} x_0, (1 - \bar{\alpha}_t)\mathbf{I})$$

$$x_{t-1} = m \odot x_{t-1}^{\text{known}} + (1 - m) \odot x_{t-1}^{\text{unknown}}$$

Constrain pixels each iteration.

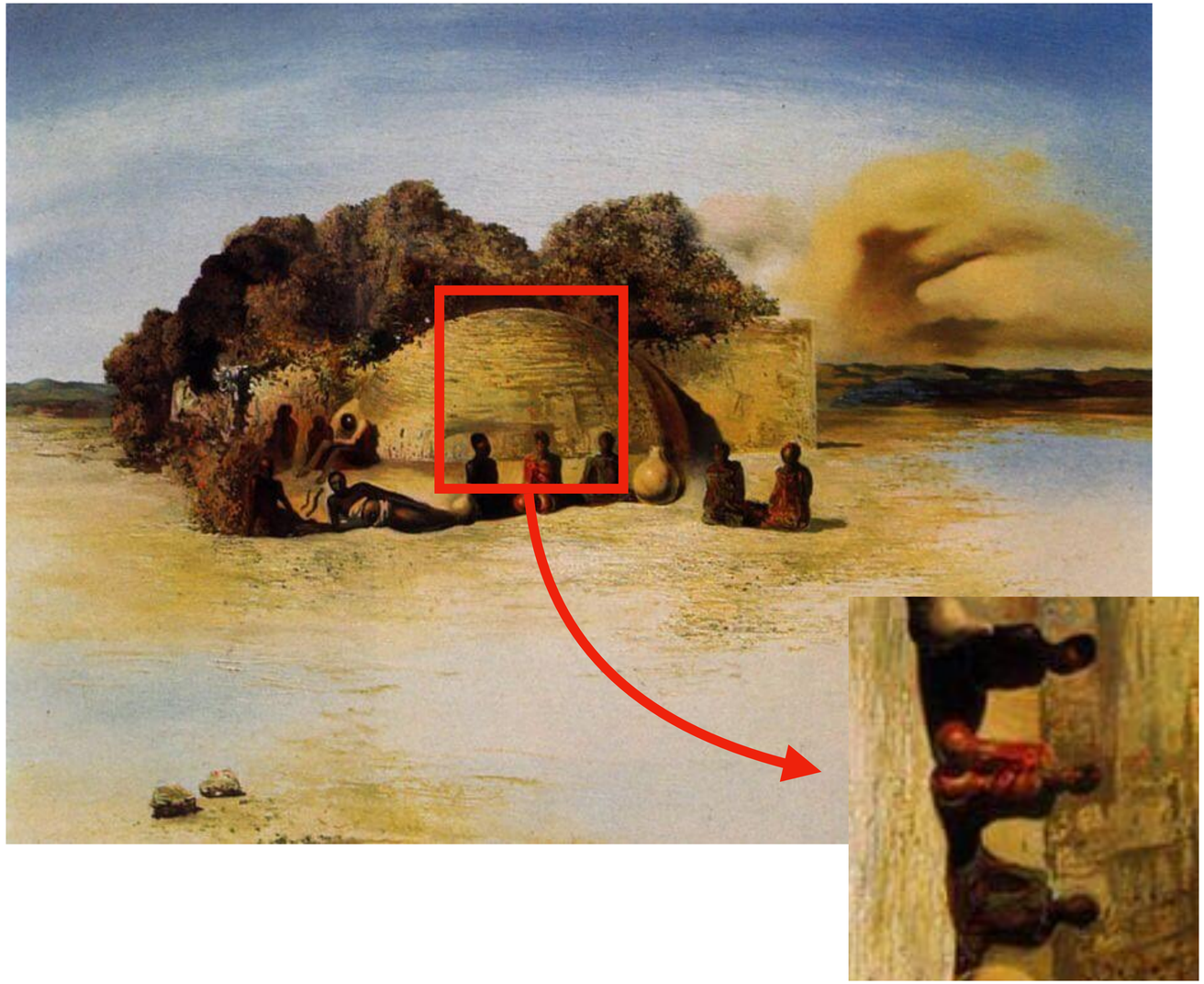


Source: [Lugmayr et al., RePaint]

How far can we take this?



Salvador Dalí,
Paranoiac Face. 1937.



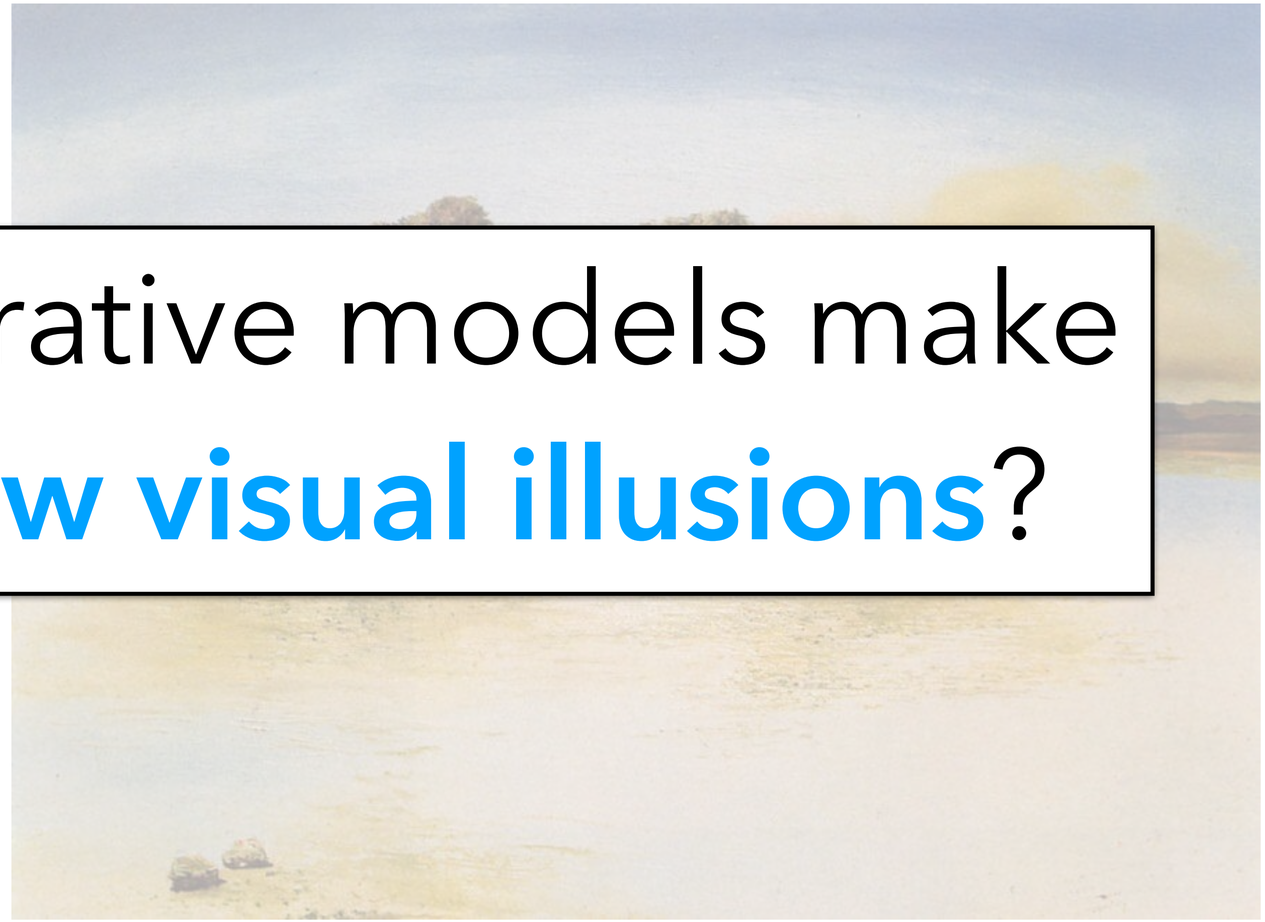


Daniel Geng



Aaron Park

[Visual Anagrams: Generating Multi-View Optical Illusions with Diffusion Models, CVPR 2024]



Can generative models make **multi-view visual illusions**?


Daniel Geng


Aaron Park

[Visual Anagrams: Generating Multi-View Optical Illusions with Diffusion Models, CVPR 2024]

Diffusion Models

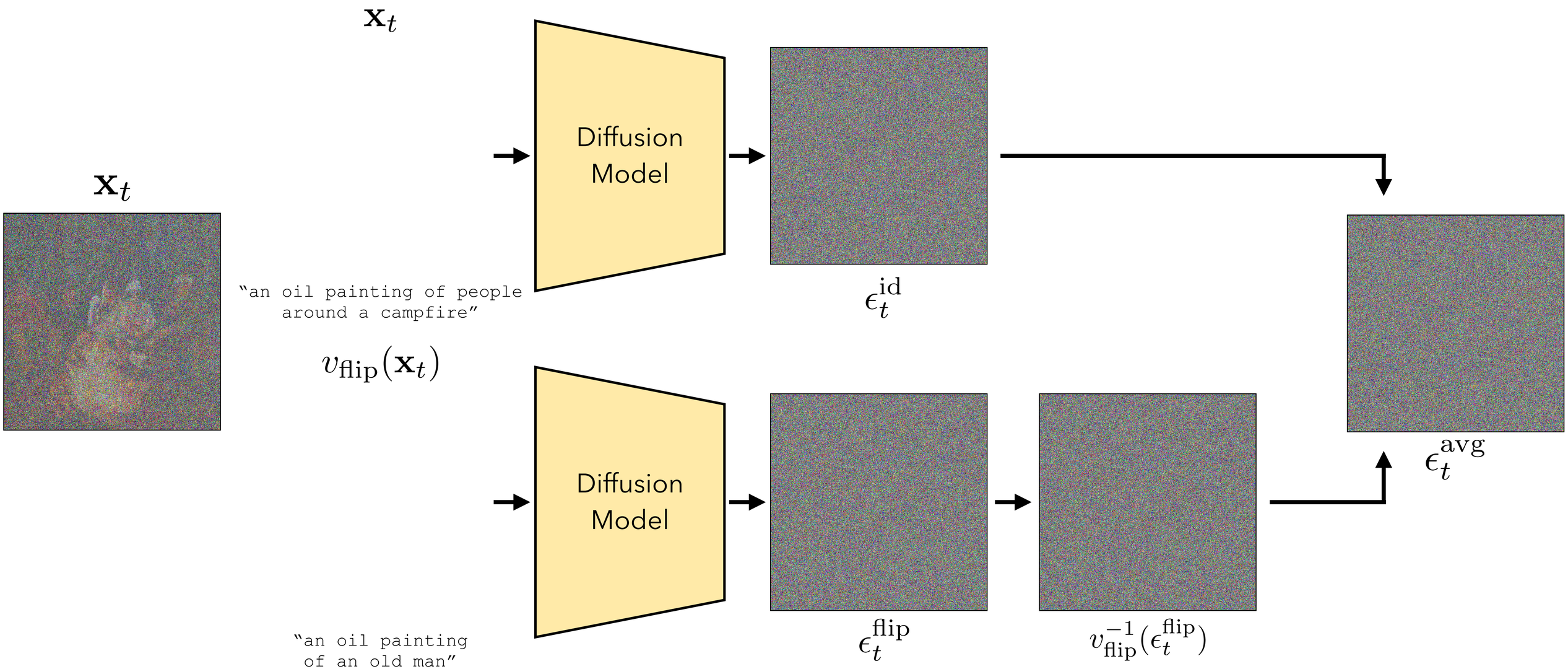
$\mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$



$\mathbf{x}_0$

"a photo of a puffin"

Generating an illusion



Generating an illusion

$\mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$



$\mathbf{x}_0$



an oil painting of people around a campfire

180°

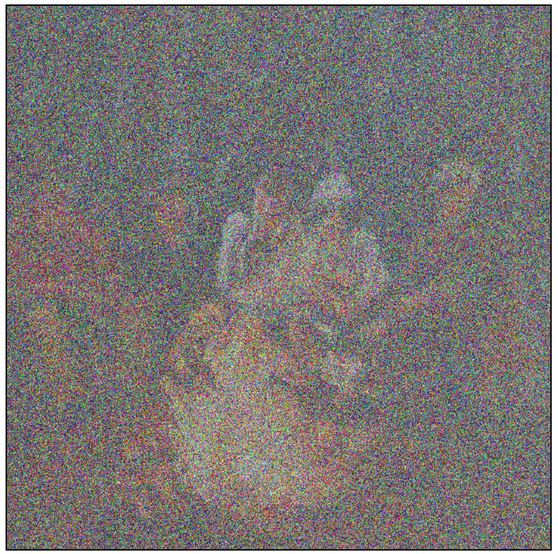


a pop art of
albert einstein

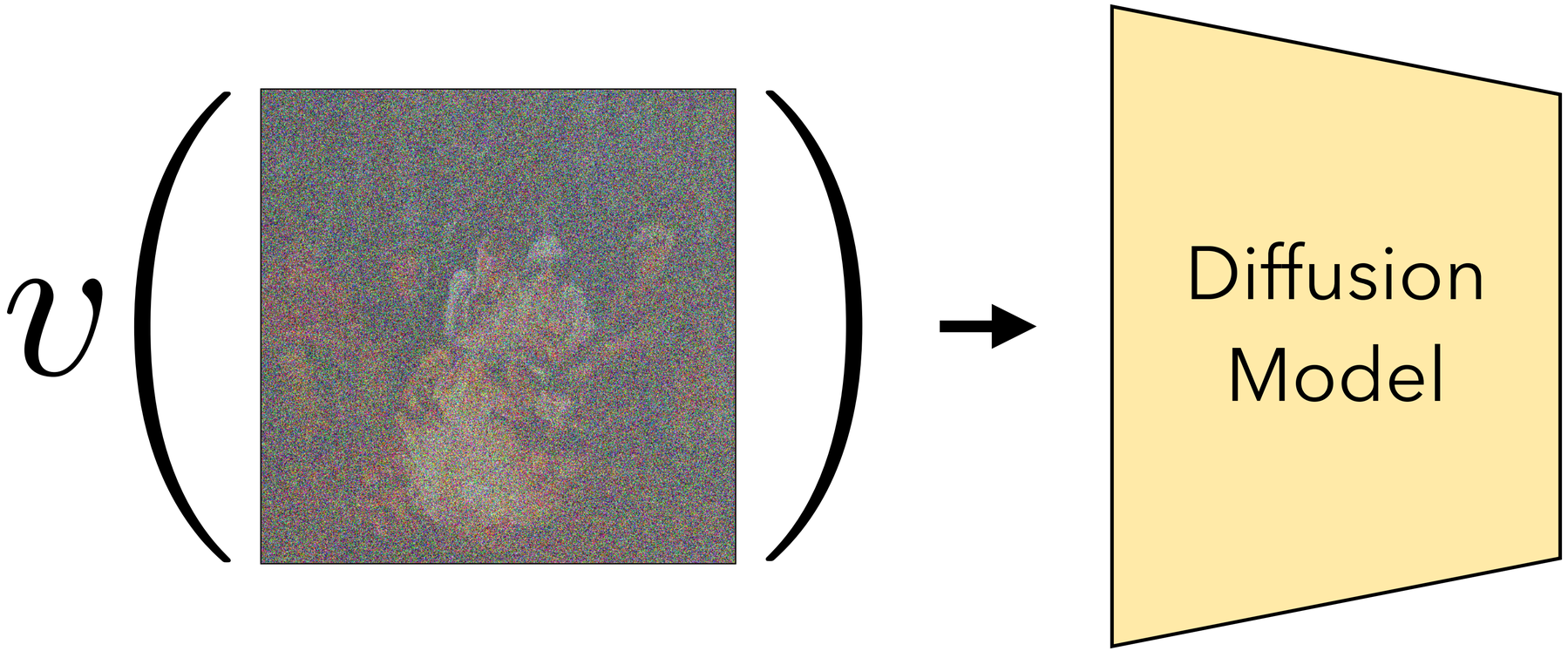
135°

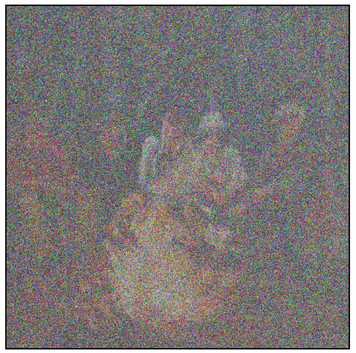


a pop art of
albert einstein



Diffusion Model

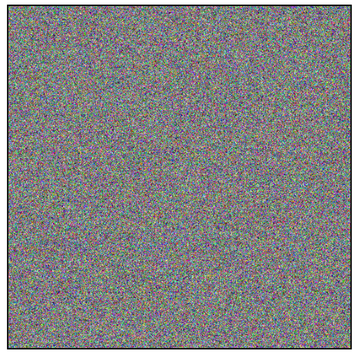

$v($
$)$
Diffusion
Model



$$= w_t^{\text{signal}}$$

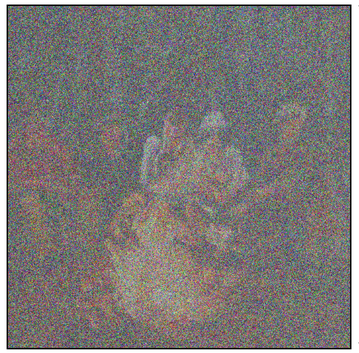
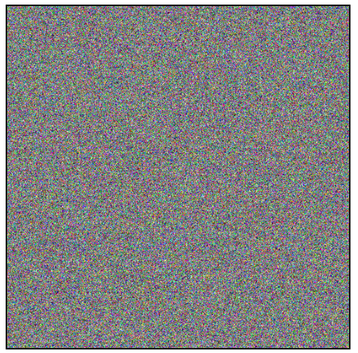


$$+ w_t^{\text{noise}}$$

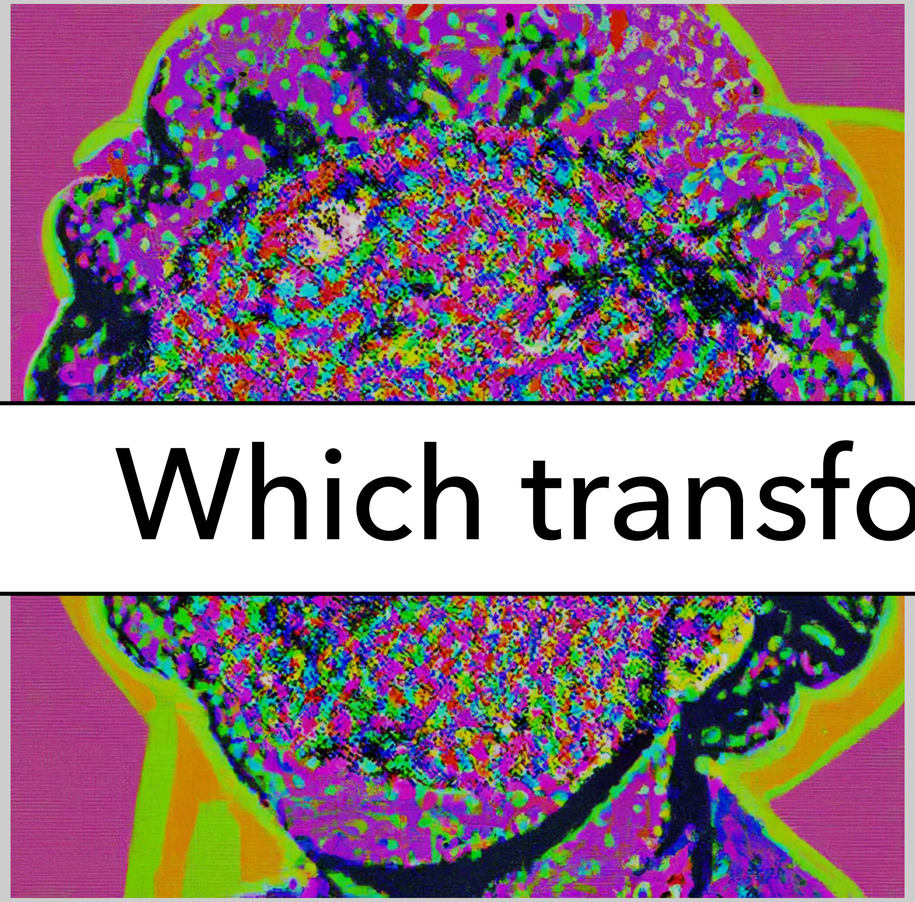


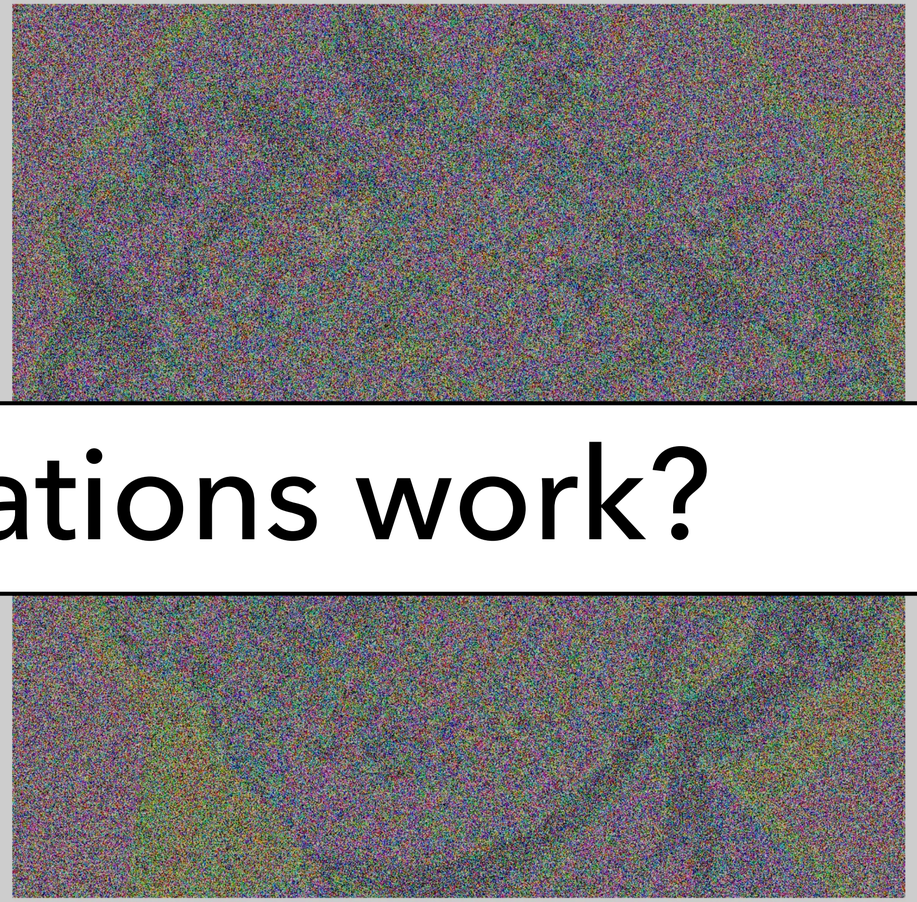
$$\sim \mathcal{N}(0, \mathbf{I})$$

1. Noisy image is weighted sum of signal and noise
2. Noise must be i.i.d. standard Gaussian

$$v\left(\text{  }\right) = v\left(w_t^{\text{signal}} \text{  } + w_t^{\text{noise}} \underbrace{\text{  }}_{\sim \mathcal{N}(0, \mathbf{I})}\right)$$

1. Noisy image is weighted sum of *signal* and *noise*
2. Noise must be i.i.d. standard Gaussian

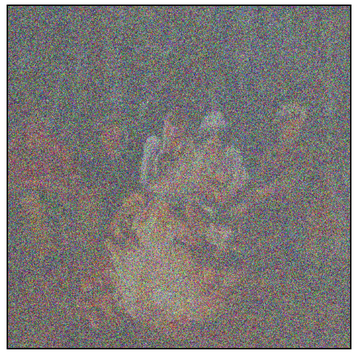
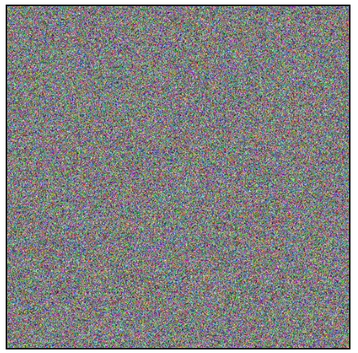




Which transformations work?

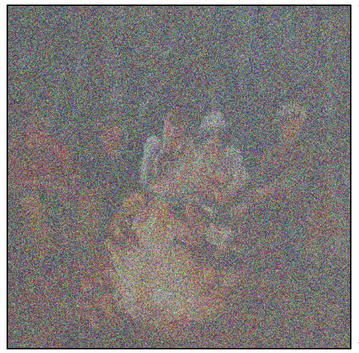
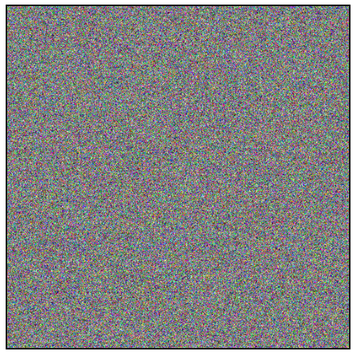
a pop art of
albert einstein

$\mathbf{x}_t$

$v\Big($  $\Big) = v\Big(w_t^{\text{signal}}$  $+ w_t^{\text{noise}}$  $\Big)$

$\sim \mathcal{N}(0, \mathbf{I})$

1. Noisy image is weighted sum of *signal* and *noise*
2. Noise must be i.i.d. standard Gaussian

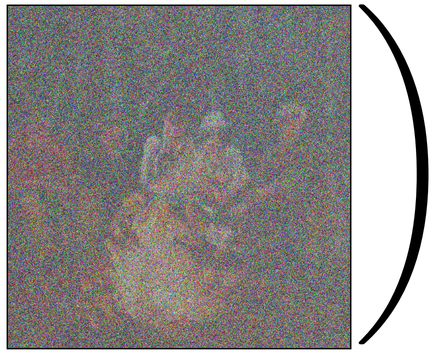
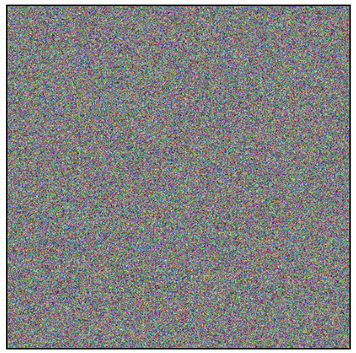
$v\Big($  $\Big) = \mathbf{A}\Big(w_t^{\text{signal}}$  $+ w_t^{\text{noise}}$  $\Big)$

$\sim \mathcal{N}(0, \mathbf{I})$

1. Noisy image is weighted sum of *signal* and *noise*
2. Noise must be i.i.d. standard Gaussian

Transformed Signal

Transformed Noise

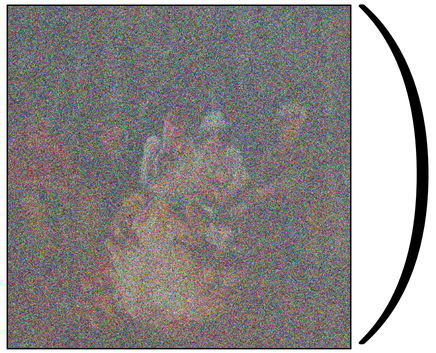
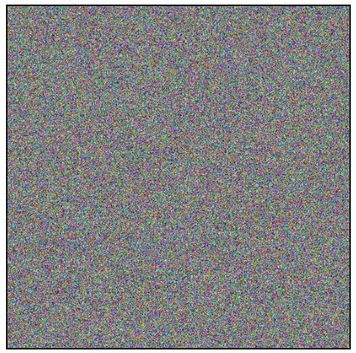
$$v\left(\text{  }\right) = w_t^{\text{signal}}\,\mathbf{A}\,\text{  } + w_t^{\text{noise}}\,\mathbf{A}\,\text{  }$$

$$\sim \mathcal{N}(0, \mathbf{I})$$

1. Noisy image is weighted sum of *signal* and *noise*
2. Noise must be i.i.d. standard Gaussian

Transformed Signal

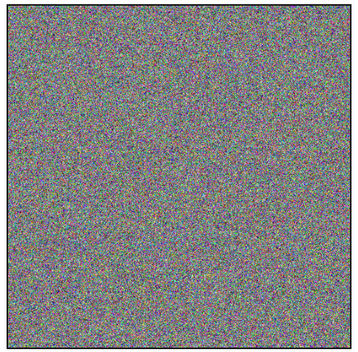
Transformed Noise

$$v\left(\text{  }\right) = w_t^{\text{signal}}\,\mathbf{A}\,\text{  } + w_t^{\text{noise}}\,\mathbf{A}\,\text{  }$$

$$\sim \mathcal{N}(0, \mathbf{I})$$

1. Noisy image is weighted sum of *signal* and *noise*
2. Noise must be i.i.d. standard Gaussian

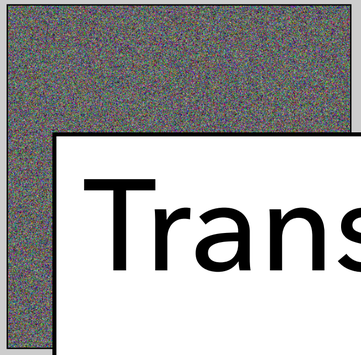
Transformed Noise

$$\mathbf{A}$$  $$\sim \mathcal{N}(0, \mathbf{I})$$

$$\sim \mathcal{N}(0, \mathbf{I})$$

1. Noisy image is weighted sum of *signal* and *noise*
2. Noise must be i.i.d. standard Gaussian

Transformed
Noise



A

nal

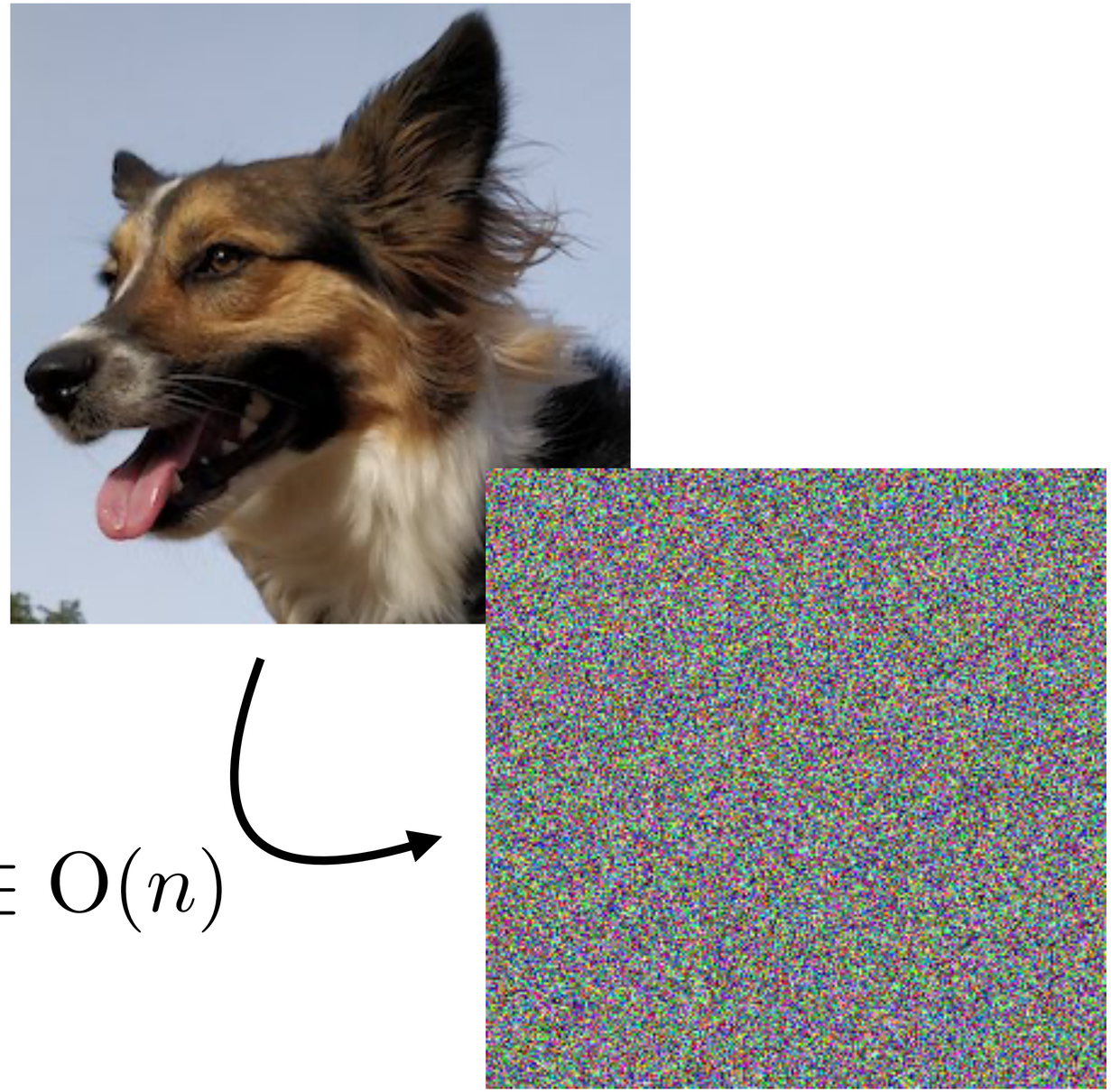
Transformations must be orthogonal

1. Noisy image is weighted sum of *signal* and *noise*
2. Noise must be i.i.d. standard Gaussian



Most orthogonal transformations are uninterpretable…

… but any permutation is orthogonal!



Visual Anagrams

An image that changes appearance
under a permutation of its pixels

(an ambigram!)



180° Rotations

the word "happy",
cursive writing

See Also: Boheng Zhao et al. “AmbiGen,” arXiv 2023

90° Rotations



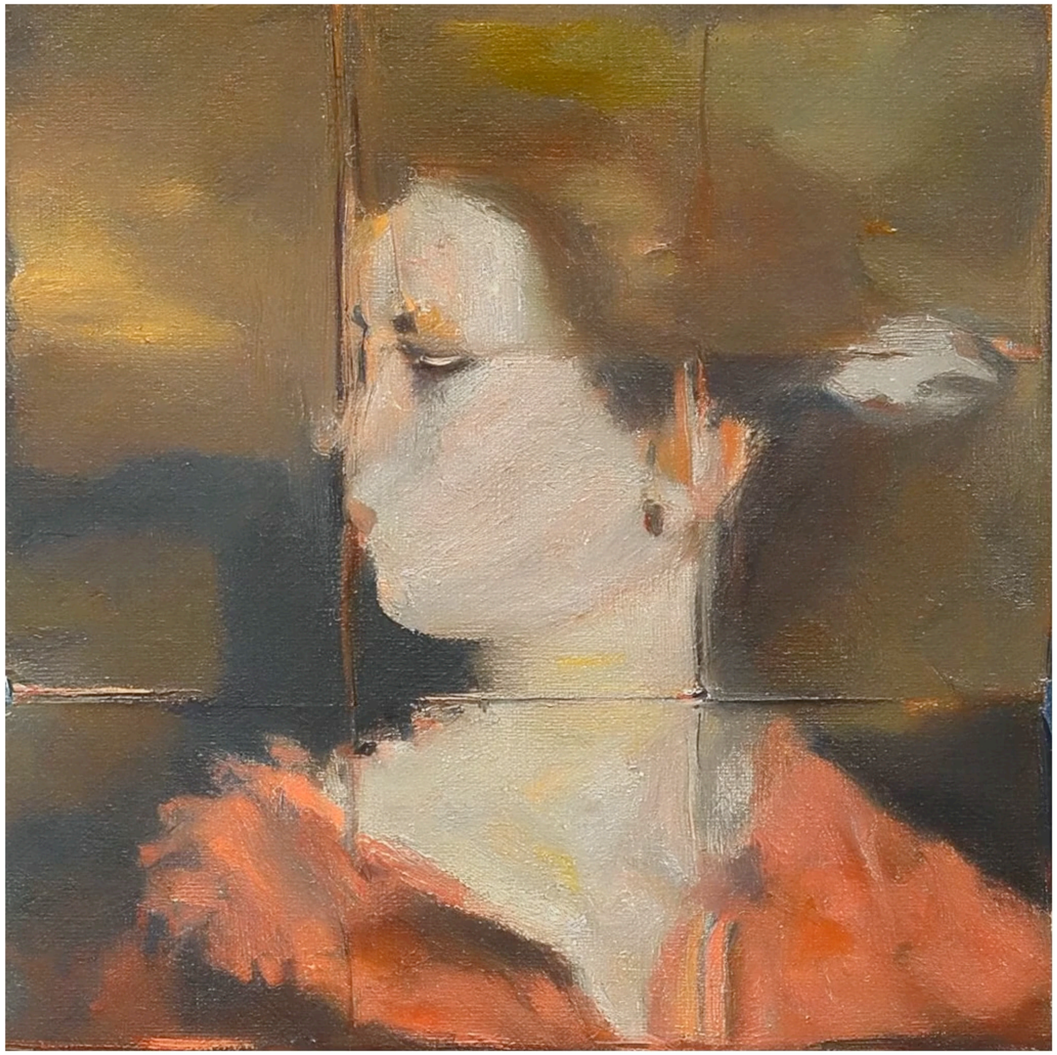
a lithograph of a table

“Inner Circle”



an oil painting
of an old man

“Square Hinge”



an oil painting
of a young lady

Thank you to Ryan Burgert for this idea! Inspired by an illusion from Steve Mould.

“Polymorphic” Jigsaw Puzzles



a watercolor of a kitten

“Polymorphic” Jigsaw Puzzles



a watercolor of a rabbit

Real puzzles!



Patch Permutations



a watercolor of a rabbit

Patch Permutations



a pencil sketch of a kangaroo



Hans Holbein, *The Ambassadors*. 1533

Skews

Skews



an oil painting
of a skull

an oil painting of
a tudor portrait

Inversions $v(\mathbf{x}) = -\mathbf{x}$

Inversions



a lithograph of a man

Inversions



a lithograph of a landscape



or



edge detector



gradient magnitude

The limits of edge-based vision?

Thanks to Ted Adelson for pointing this out.

3-view Illusions



an oil painting of houseplants

3-view Illusions



an oil painting of elvis

Analysis

Latent vs. Pixel Models

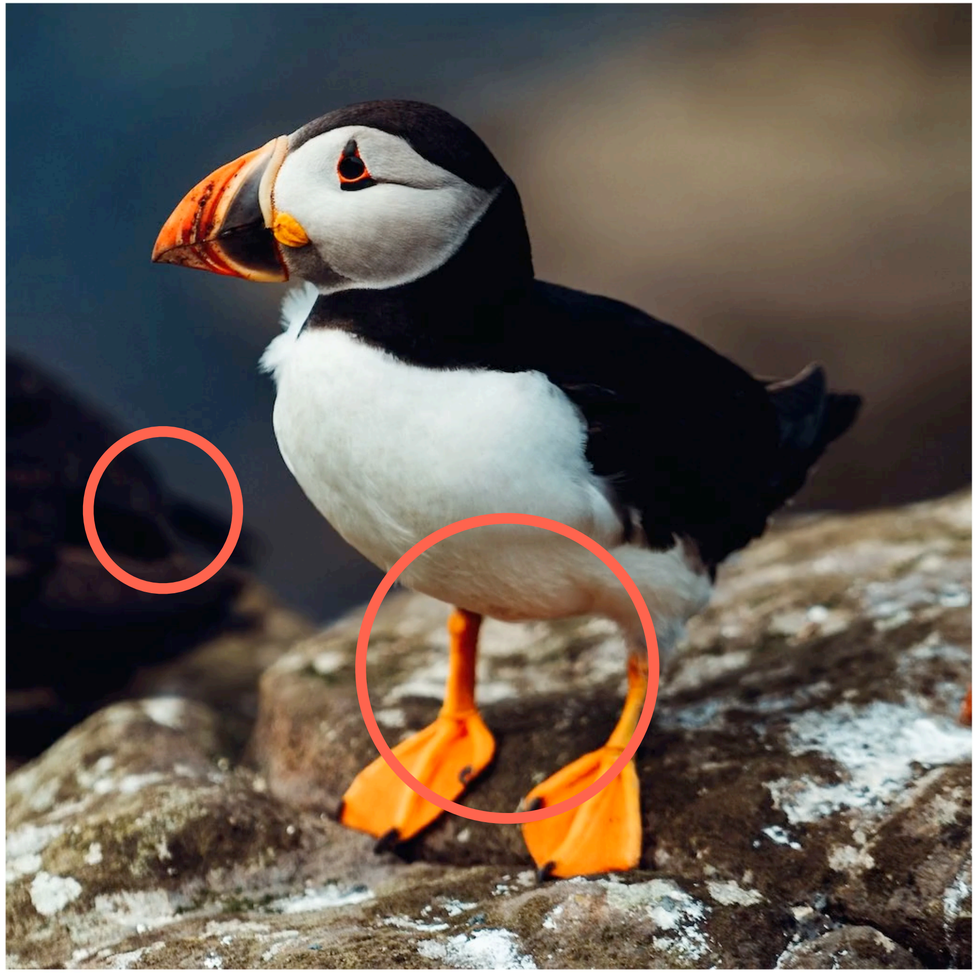
We use DeepFloyd, a
pixel diffusion model…

…because latent models
produce artifacts

Latent vs. Pixel Models

We use DeepFloyd, a pixel diffusion model…

…because latent models produce artifacts



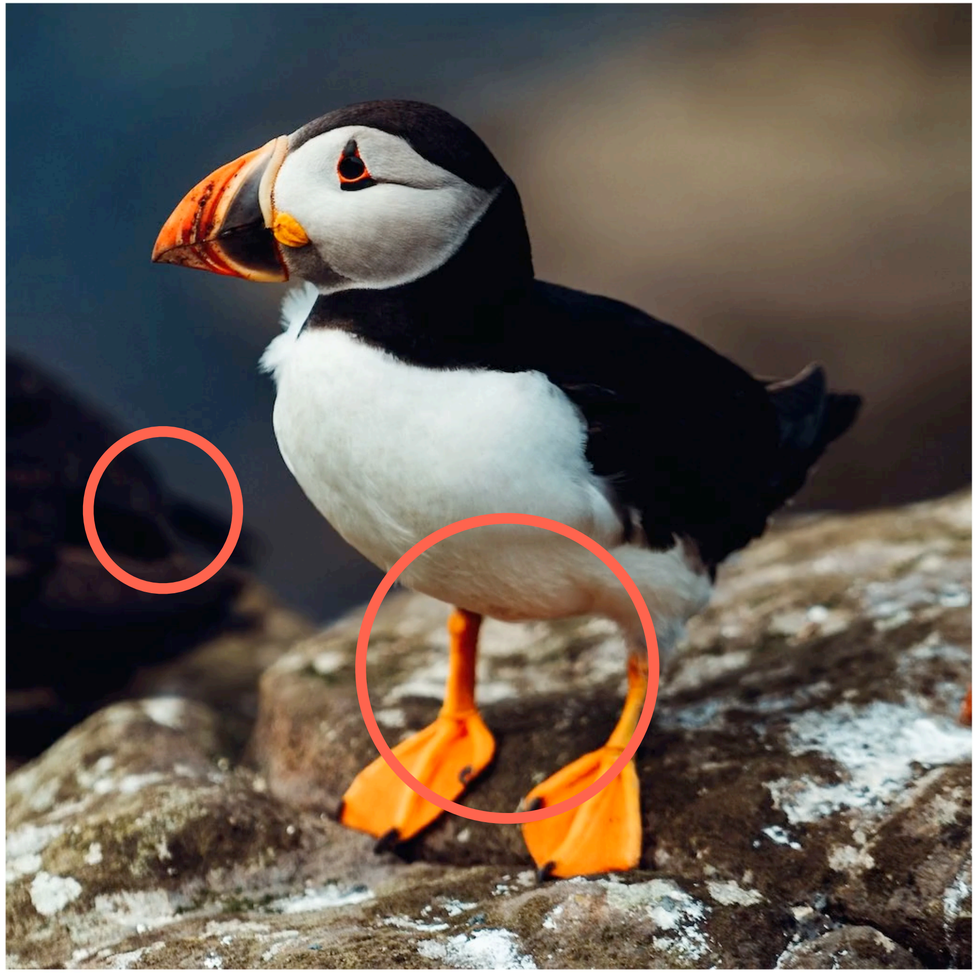
Latent vs. Pixel Models

Stable Diffusion v1.5



A cartoon drawing
of a bird

A cartoon drawing
of a frog



Related Work



Illusion Diffusion

Matthew Tancik

Colab notebook that denoises multiple views







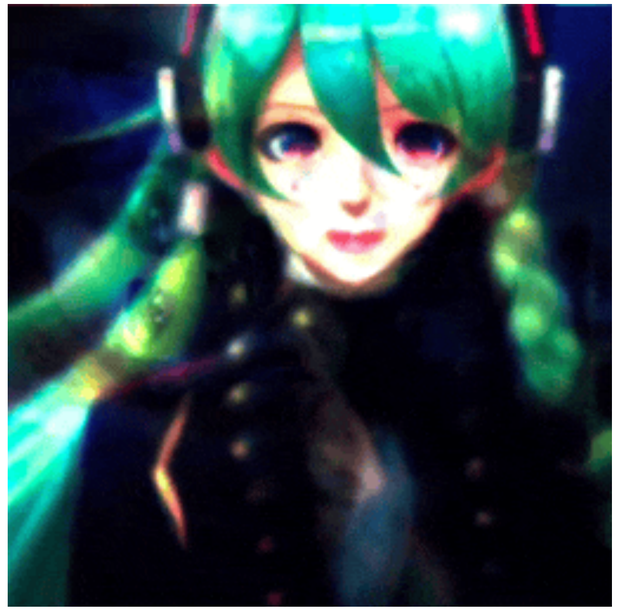

Diffusion Illusions

Ryan Burgert et al.

SIGGRAPH ’24, CVPR Outstanding Demo

diffusionillusions.com

Use SDS to optimize multiple views of an image



Which Prompts Work?

“an oil painting of people around a campfire”

“an oil painting of an old man”








“a photo of a goldfish”

“a photo of the queen of england”








Why does this work?

Recall connection to score function:

$$\epsilon_\theta(\mathbf{x}_t) \approx -\sigma_t \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t)$$

where $\mathbf{x}_t$ is the data noised at step t of the diffusion model with Gaussian noise σ_t.

Compositionality

By learning the score…

$$\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t)$$

…you get many other scores for free

$$\nabla_{\mathbf{x}_t} \log p_{\text{flipped},t}(\mathbf{x}_t)$$

$$\nabla_{\mathbf{x}_t} \log p_{\text{rotated},t}(\mathbf{x}_t)$$

$$\nabla_{\mathbf{x}_t} \log p_{\text{permuted},t}(\mathbf{x}_t)$$

$$\vdots$$

Compositionality

$$\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t) + \nabla_{\mathbf{x}_t} \log p_{\text{flipped},t}(\mathbf{x}_t)$$

$$= \nabla_{\mathbf{x}_t} \log \left(p_t(\mathbf{x}_t) p_{\text{flipped},t}(\mathbf{x}_t) \right)$$

$$\Longrightarrow$$

$$\mathbf{x} \sim p^{\text{prod}}(\mathbf{x}) \propto p(\mathbf{x}) p_{\text{flipped}}(\mathbf{x})$$

See also: Nan Liu, Shuang Li, Yilun Du, et al. "Compositional Visual Generation." ECCV 2022
Yilun Du, et al. "Reduce, Reuse, Recycle." ICML 2023

Limitations to this approach

Compositionality failures



Success rate depends on prompts



Orthogonal views

$$v\left(\boxed{}\right) = v\left(w_t^{\text{signal}} \boxed{} + w_t^{\text{noise}} \underbrace{\boxed{}}_{\sim \mathcal{N}(0, \mathbf{I})}\right)$$

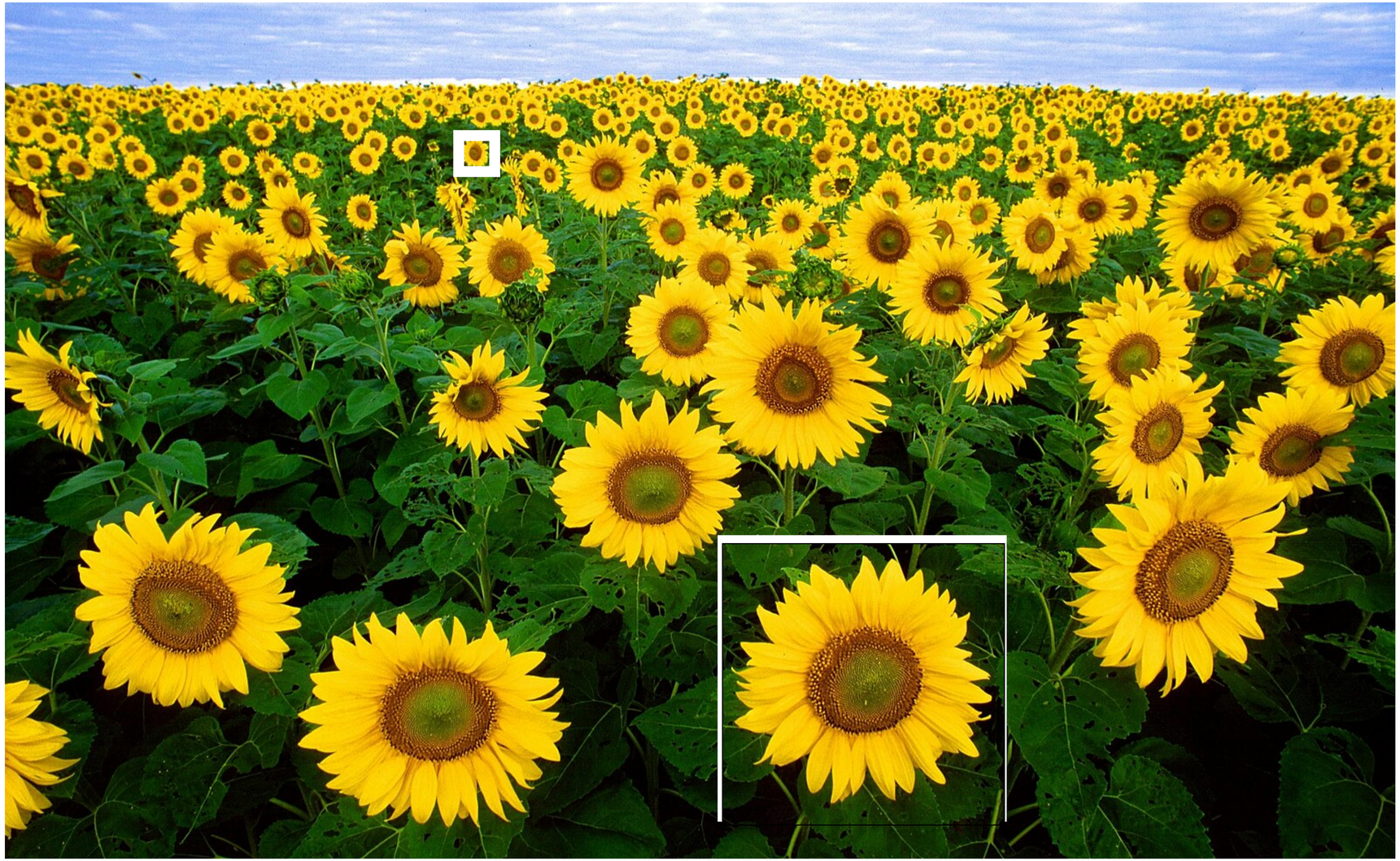


Charles Allan Gilbert
All Is Vanity. 1892.



See also [“Hybrid Images,” Oliva et al., 2006]

Frequency and distance



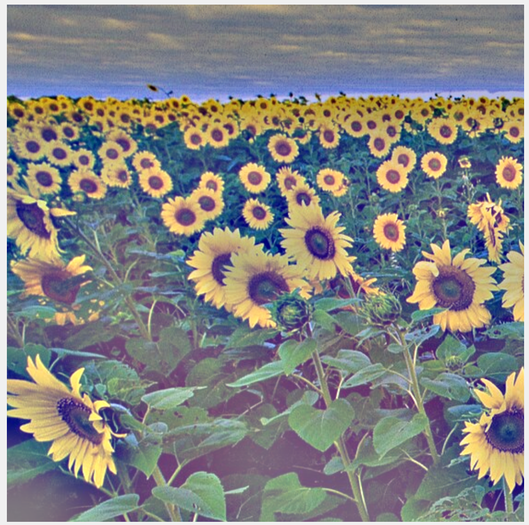
Frequency and distance

linear functions

Decomposition into linear factors



=



near

+



in between

+

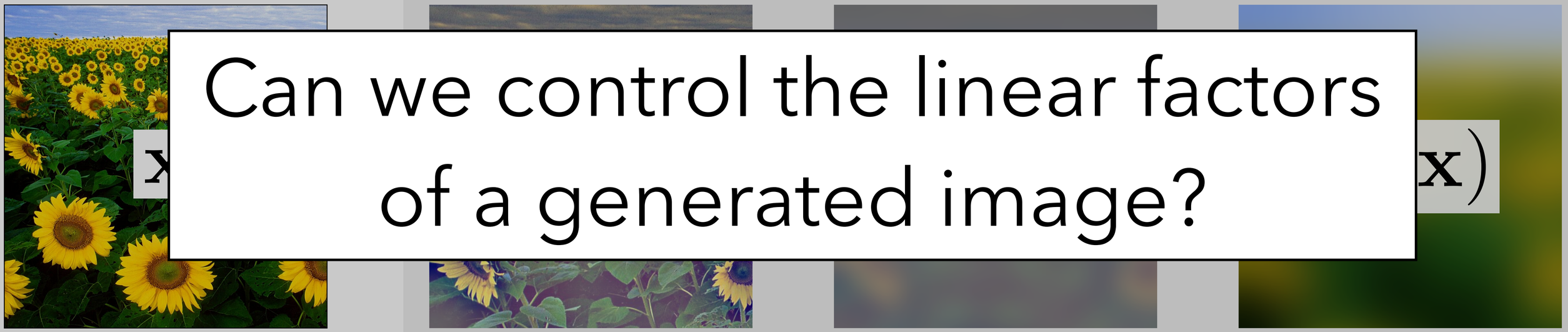


far

[Burt & Adelson, “Laplacian Pyramid”, 1983, Oliva et al., “Hybrid Images”, 2006]

Frequency and distance

Decomposition into linear factors



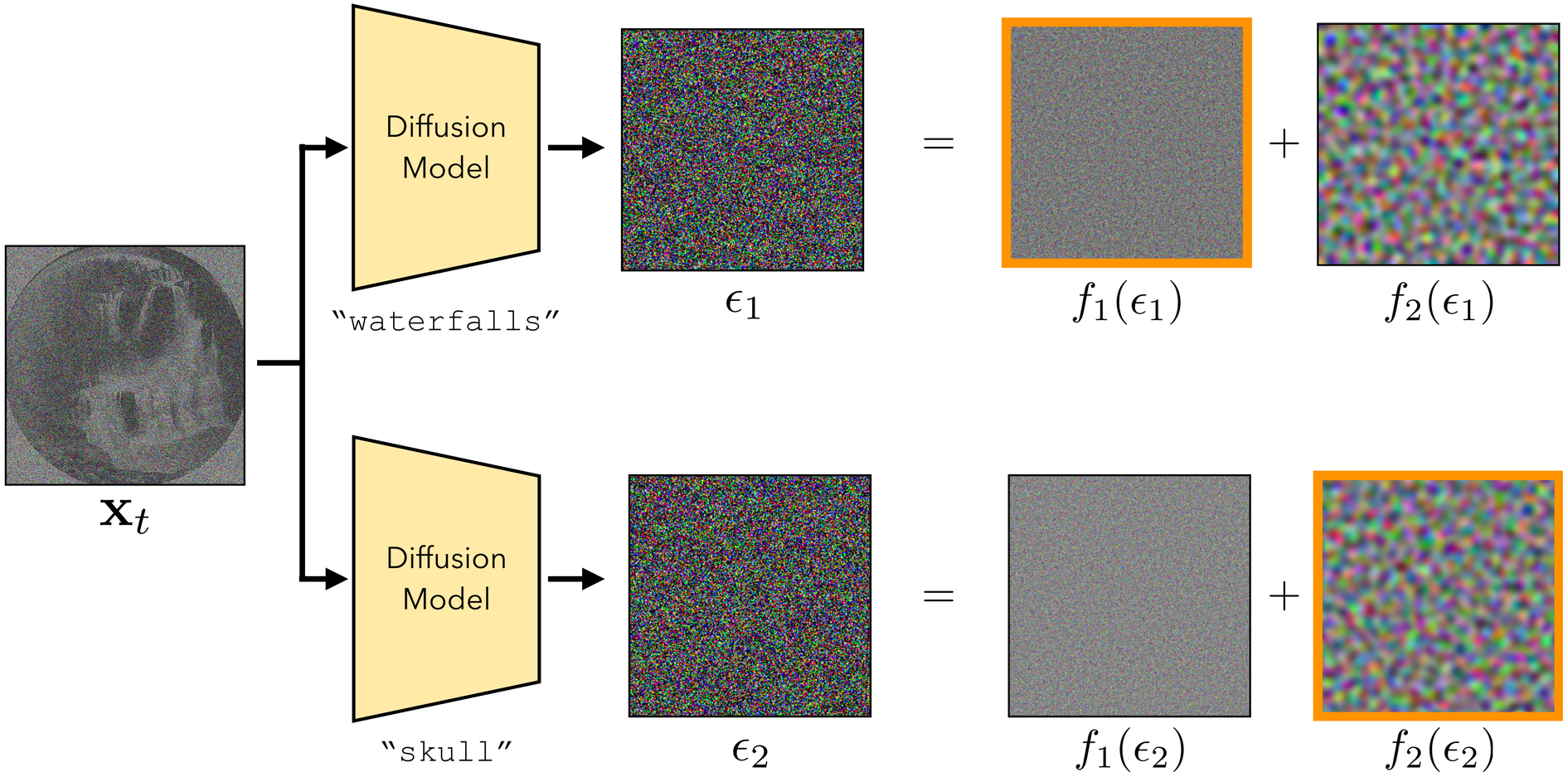
Can we control the linear factors of a generated image?


Daniel Geng*

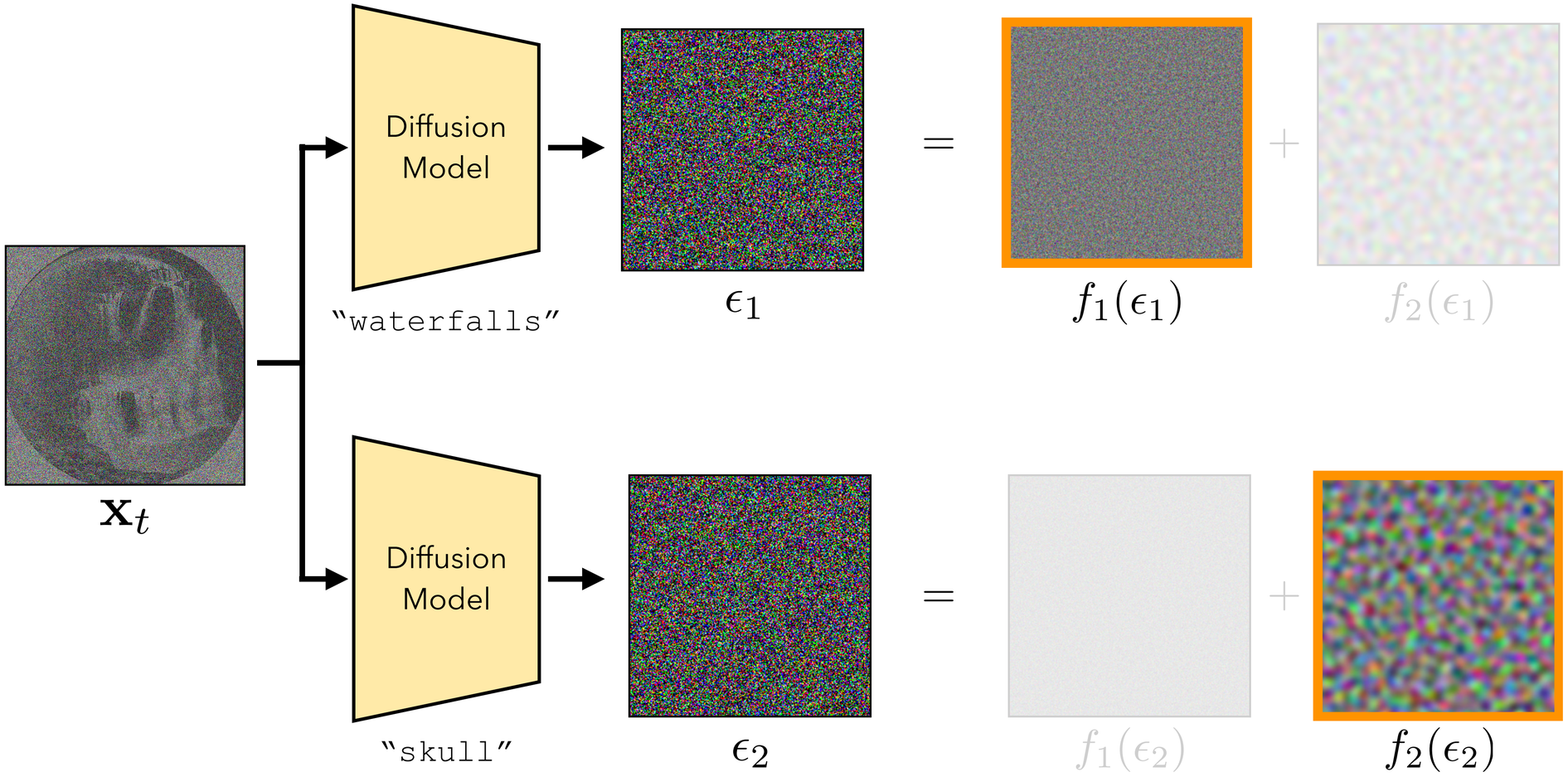

Aaron Park*

[Factorized Diffusion: Perceptual Illusions by Noise Decomposition, ECCV 2024]

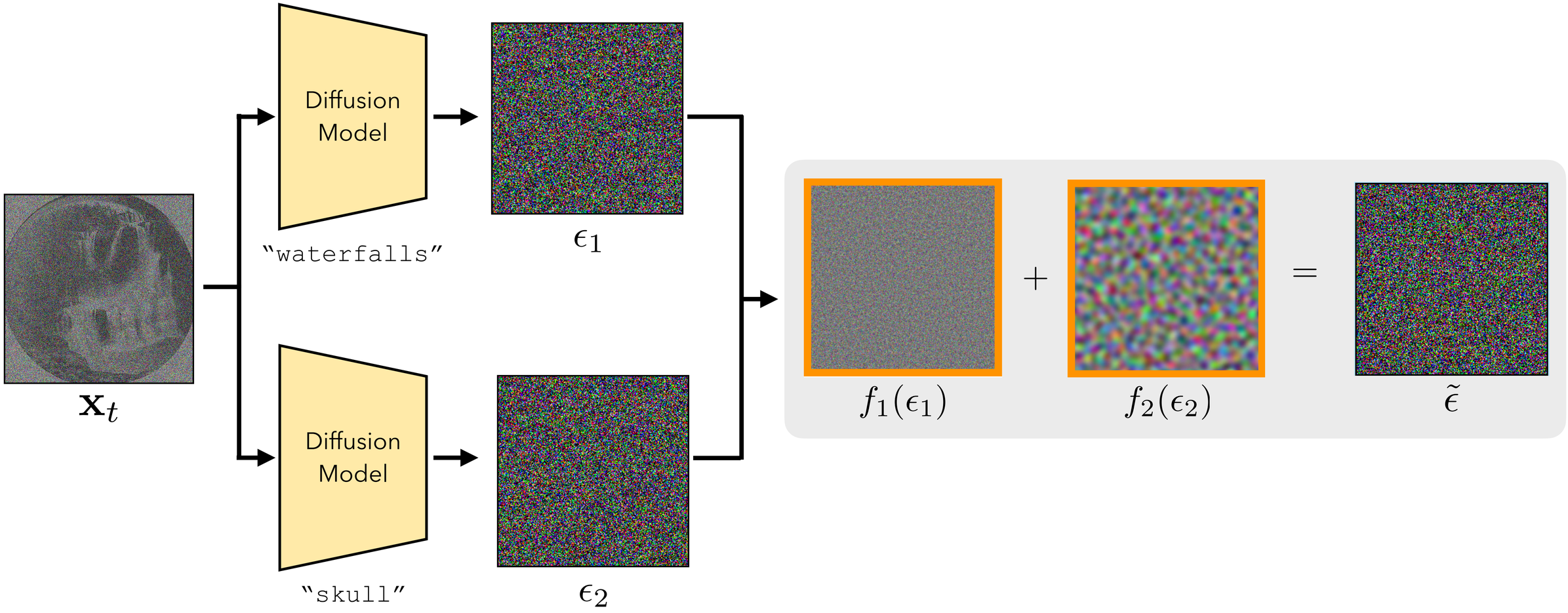
Factorized diffusion



Factorized diffusion



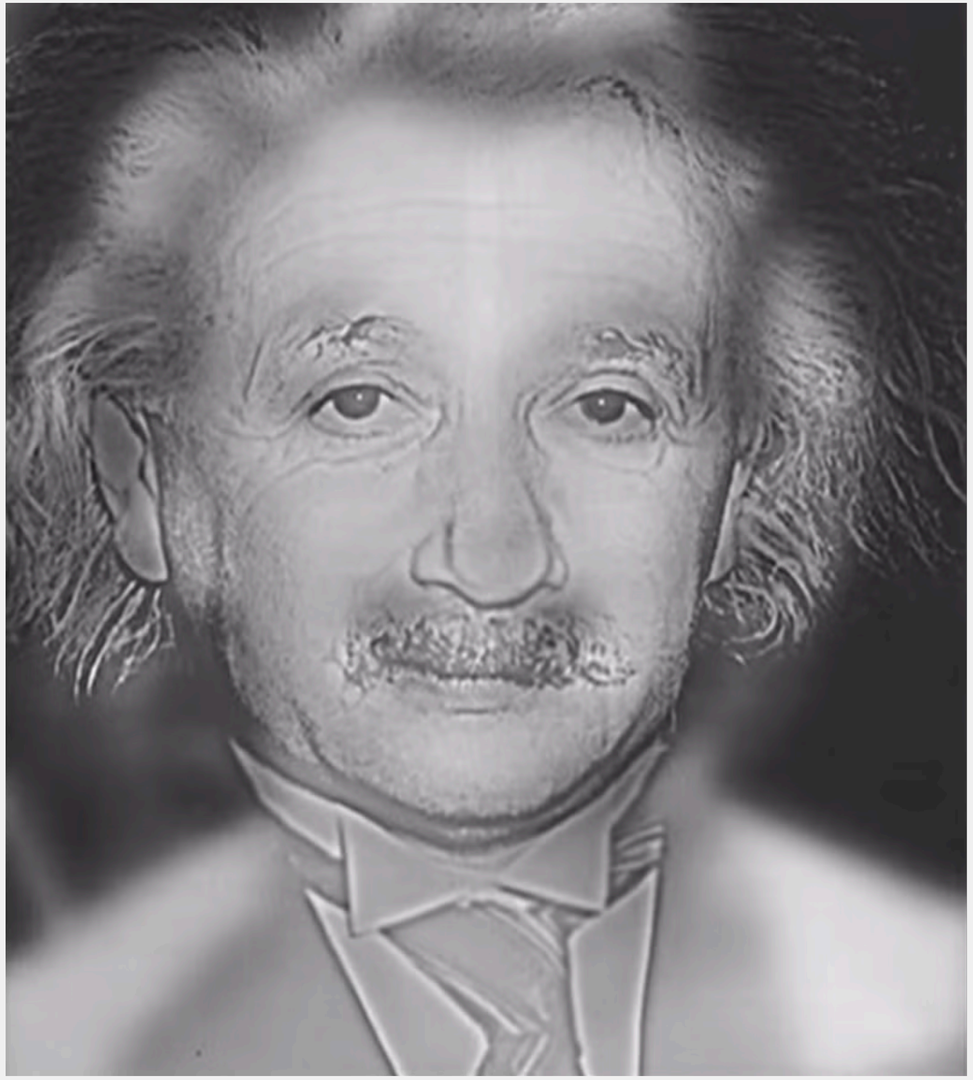
Factorized diffusion



Hybrid Decomposition

$$\mathbf{x} = \underbrace{\mathbf{x} - G_\sigma(\mathbf{x})}_{f_{\text{high}}(\mathbf{x})} + \underbrace{G_\sigma(\mathbf{x})}_{f_{\text{low}}(\mathbf{x})}$$

Related work

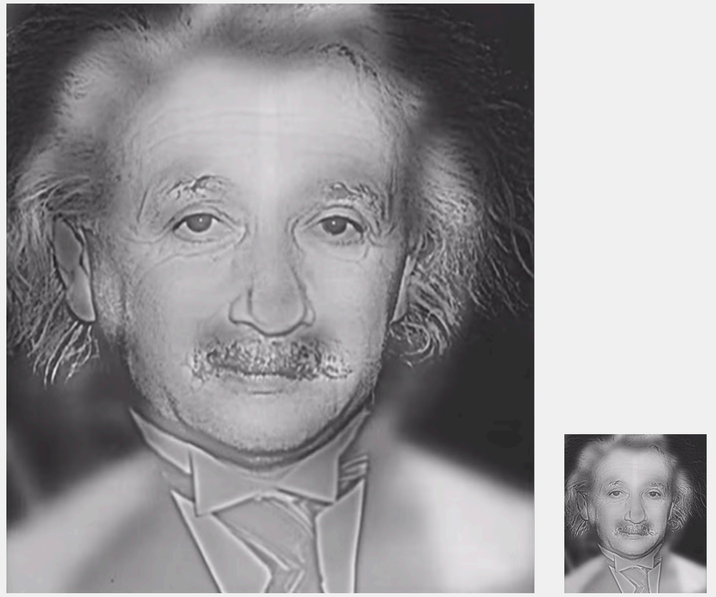


[Aude Oliva et al., “Hybrid Images,” SIGGRAPH 2006]

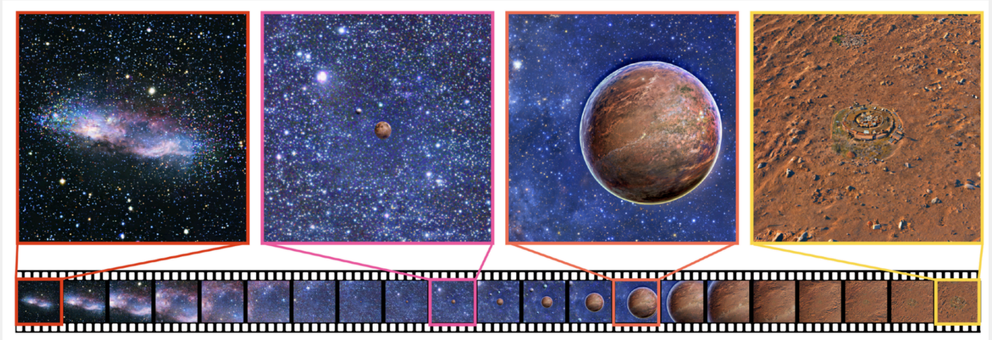
Hybrid Decomposition

$$\mathbf{x} = \underbrace{\mathbf{x} - G_\sigma(\mathbf{x})}_{f_{\text{high}}(\mathbf{x})} + \underbrace{G_\sigma(\mathbf{x})}_{f_{\text{low}}(\mathbf{x})}$$

Related work



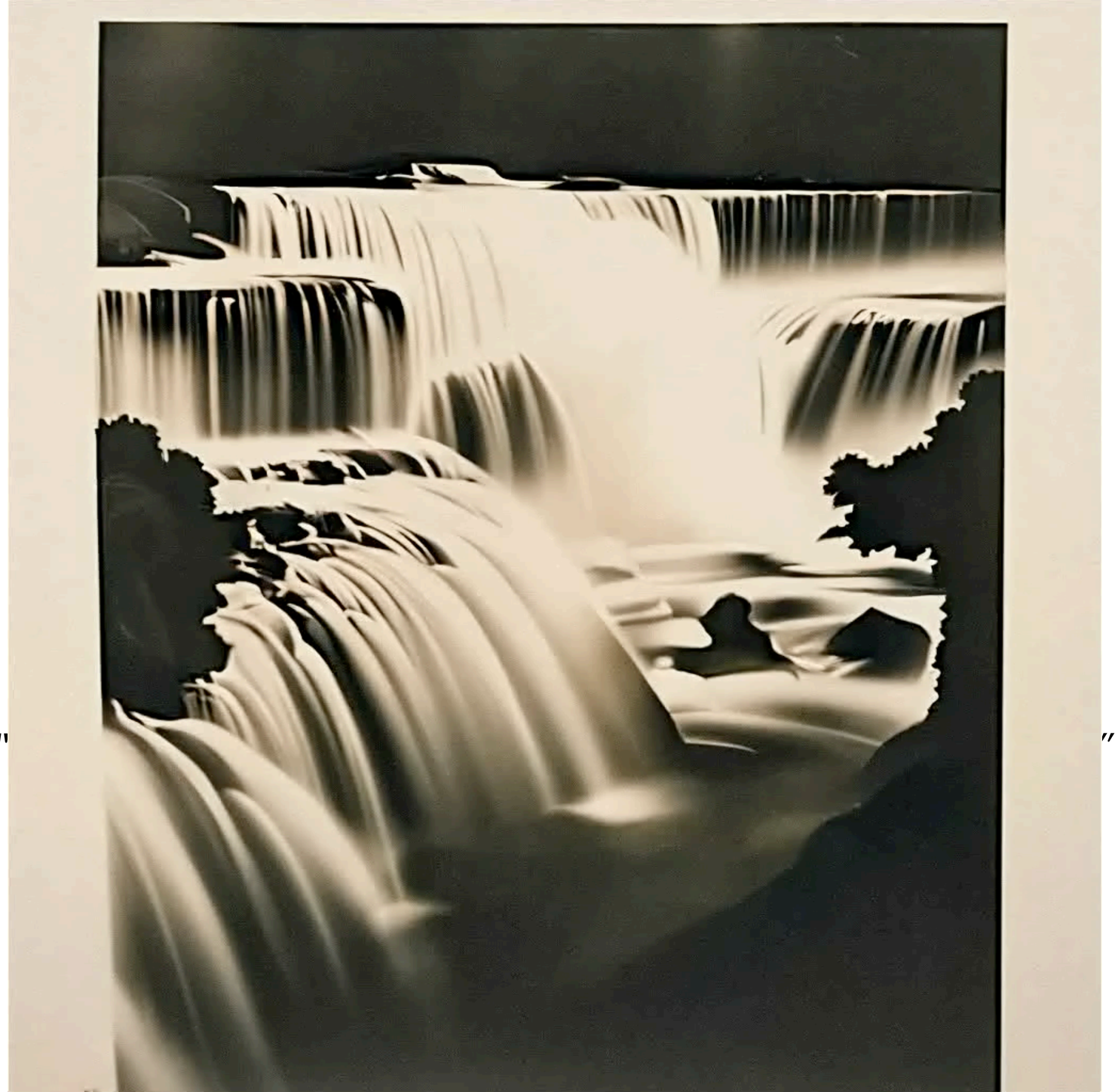
[Aude Oliva et al., “Hybrid Images,” SIGGRAPH 2006]



[Xiaojuan Wang et al., “Generative Powers of Ten”, CVPR 2024]









ph of a yin yang”

Other “natural” linear decompositions





Color Decomposition

$$f_{\text{gray}}(\mathbf{x}) = \frac{1}{3} \sum_{c \in \{R,G,B\}} \mathbf{x}_c$$

$$f_{\text{color}}(\mathbf{x}) = \mathbf{x} - f_{\text{gray}}(\mathbf{x})$$



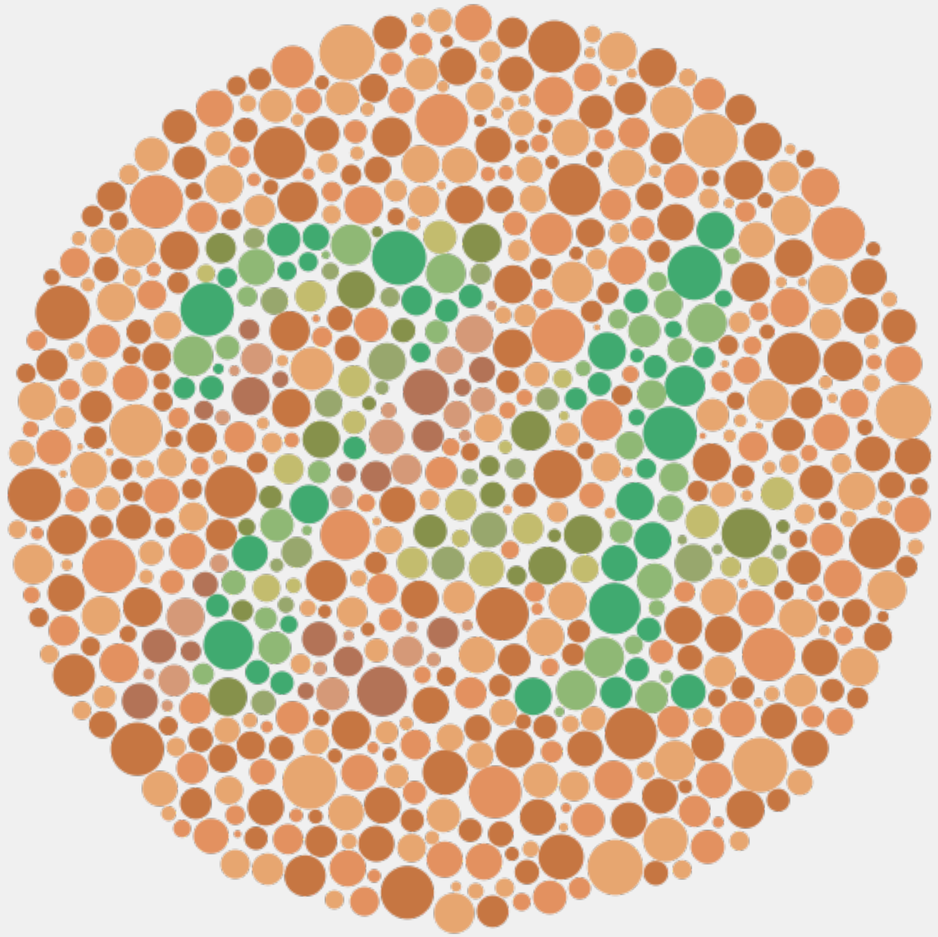


Color Decomposition

$$f_{\text{gray}}(\mathbf{x}) = \frac{1}{3} \sum_{c \in \{R,G,B\}} \mathbf{x}_c$$

$$f_{\text{color}}(\mathbf{x}) = \mathbf{x} - f_{\text{gray}}(\mathbf{x})$$

Related work



Color blindness test
[Ishihara 1917]



oil painting style, a flower arrangement

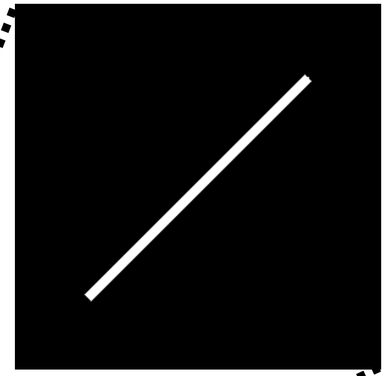


a lithograph of a landscape



oil painting style, a barn





Motion Blur Decomposition

$$f_{\text{motion}}(\mathbf{x}) = \mathbf{h} * \mathbf{x}$$

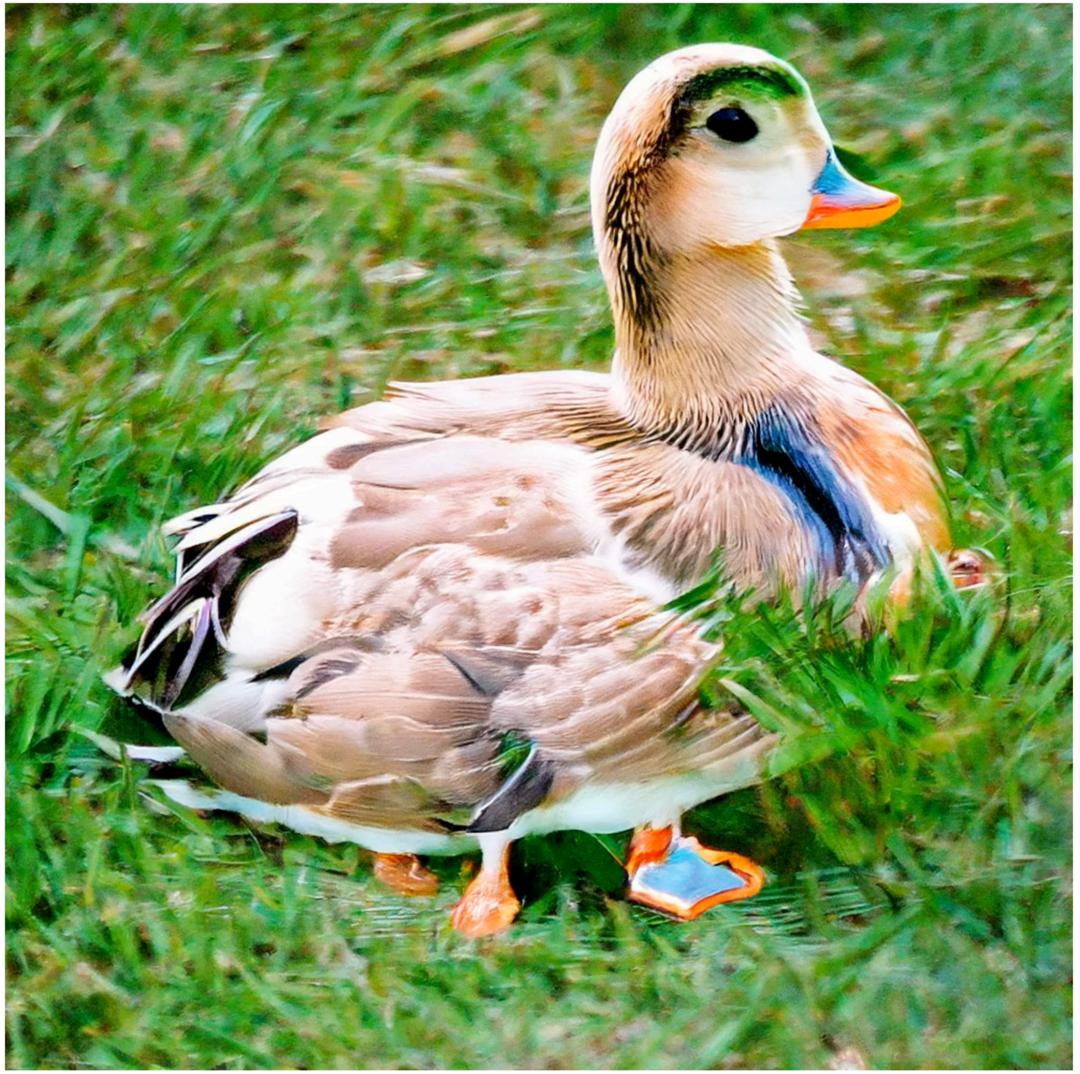
$$f_{\text{res}}(\mathbf{x}) = \mathbf{x} - f_{\text{motion}}(\mathbf{x})$$



Motion Blur Decomposition

$$f_{\text{motion}}(\mathbf{x}) = \mathbf{h} * \mathbf{x}$$

$$f_{\text{res}}(\mathbf{x}) = \mathbf{x} - f_{\text{motion}}(\mathbf{x})$$



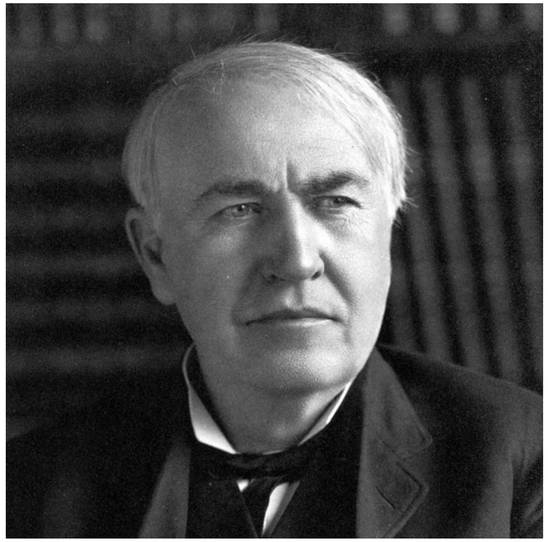
a photo of a duck

(Simulated motion)



an oil painting of a beehive

Freeze one component and generate the other



=



low pass

+



high pass

Freeze one component and generate the other

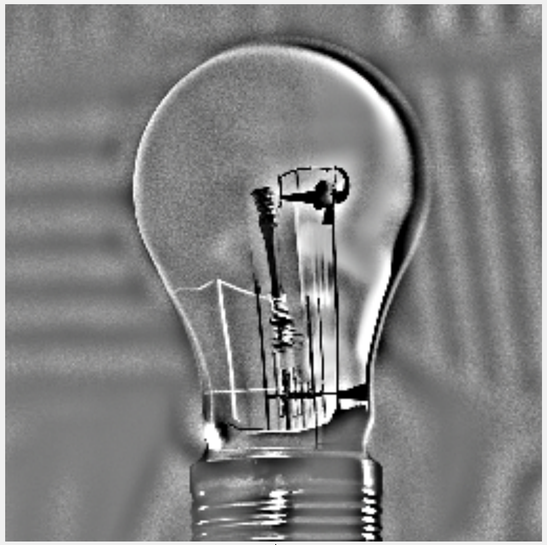


=



low pass

+

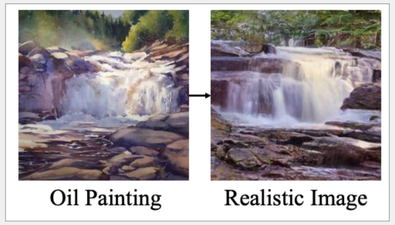


factorized diffusion

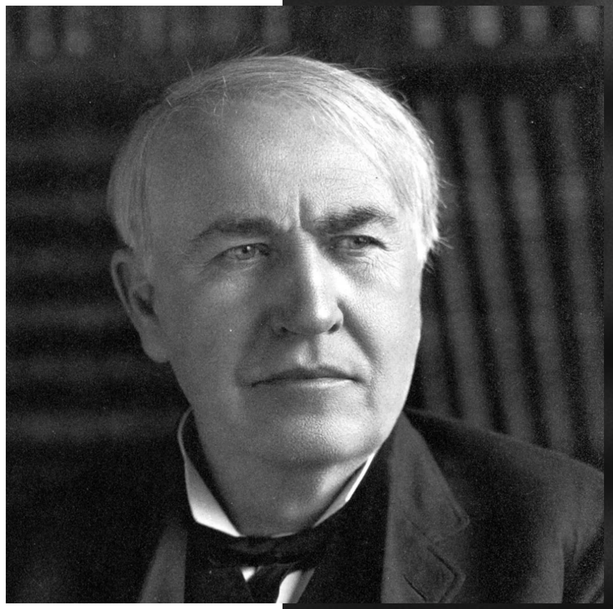
"a photo of a lightbulb"

Closely related to work on inverse problems:

[Choi et al., "ILVR", 2021], [Song et al., 2021], [Wang et al., 2022], [Kawar, "DDRM", 2022], [Lugmayr, et al., "RePaint", 2022], and many more!



[Choi et al., "ILVR", 2021]






ECCV


CVPR

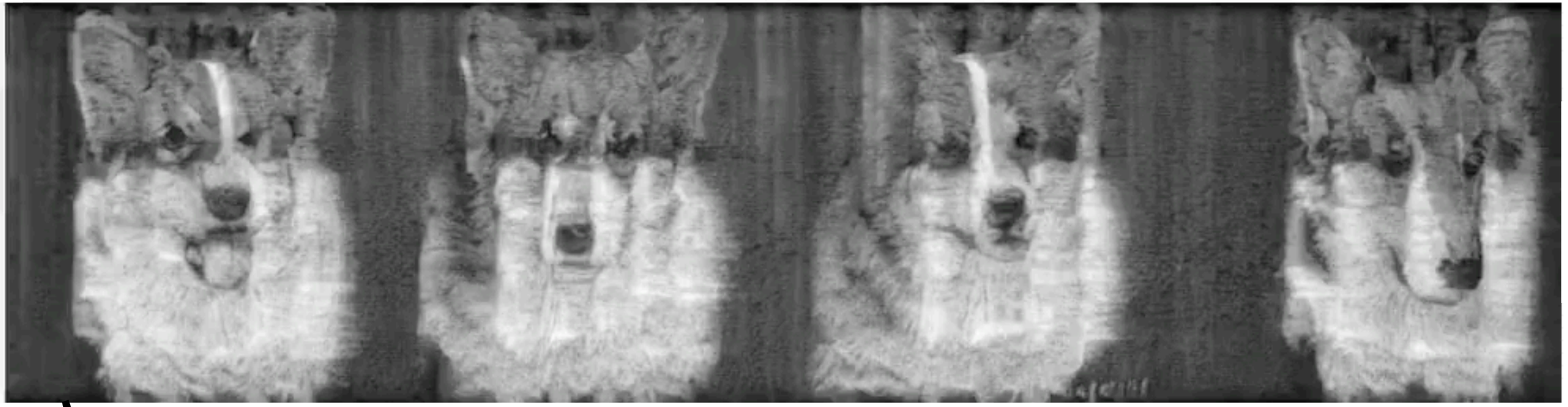


Thanks to Walter Scheirer and Luba Elliot for suggesting this!

Discussion: different models

- GANs
- Autoregressive models
- Diffusion models

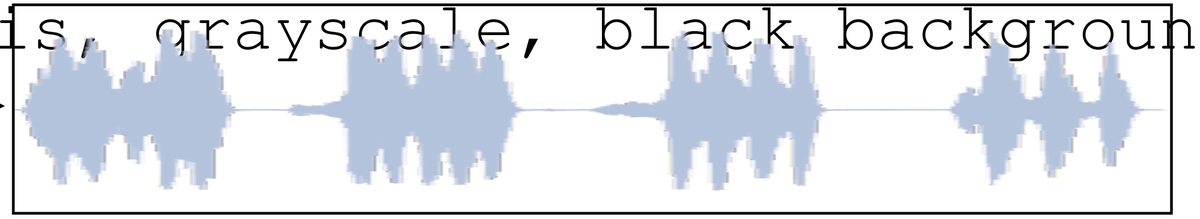
Next Monday: midterm



“a p orgis, grayscale, black background”

“dog

Spectrogram to audio





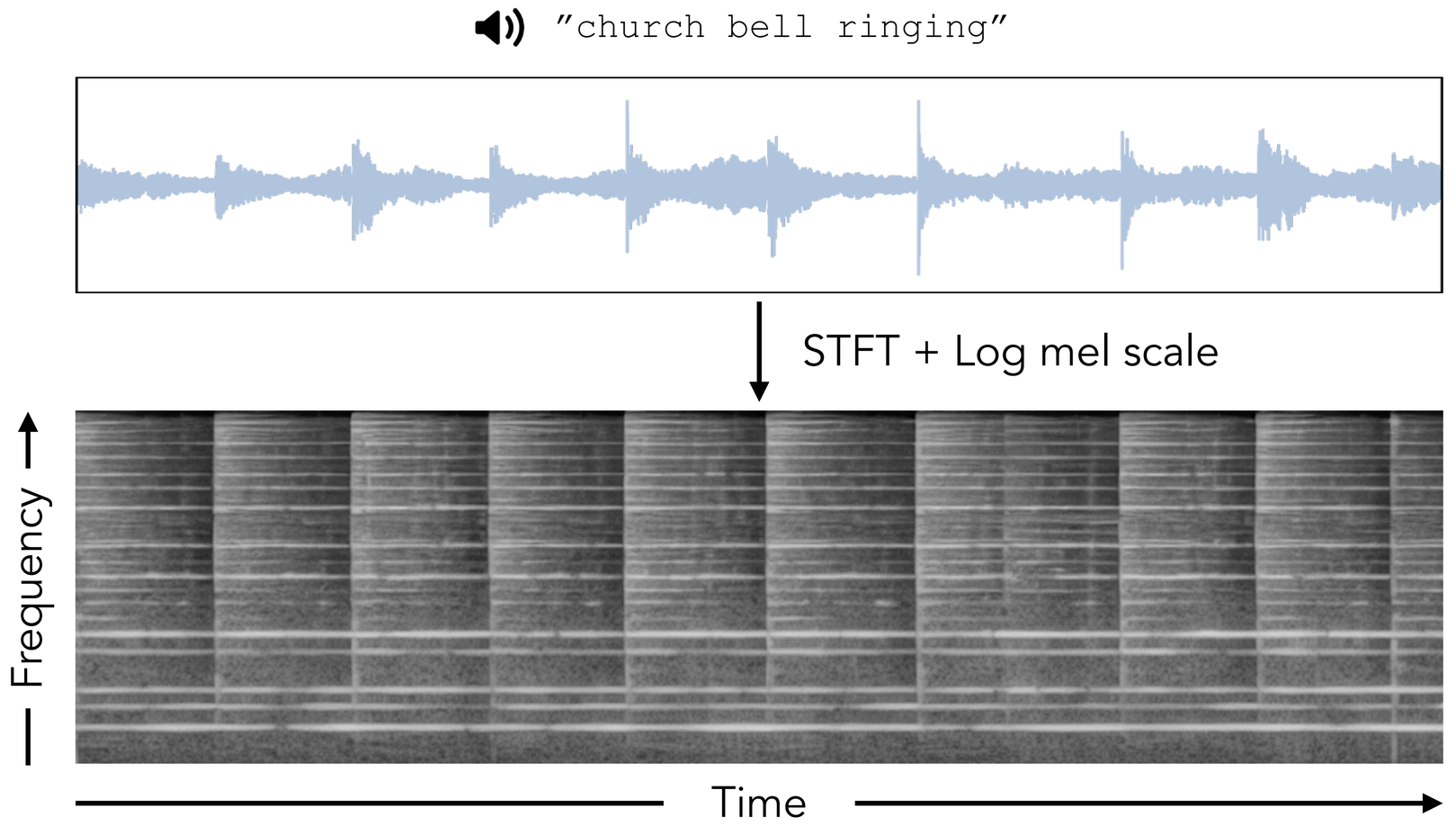
Ziyang Chen



Daniel Geng

[Chen, Geng, Owens. “Images that Sound: Composing Images and Sounds on a Single Canvas.” NeurIPS, 2024]

Audio spectrograms

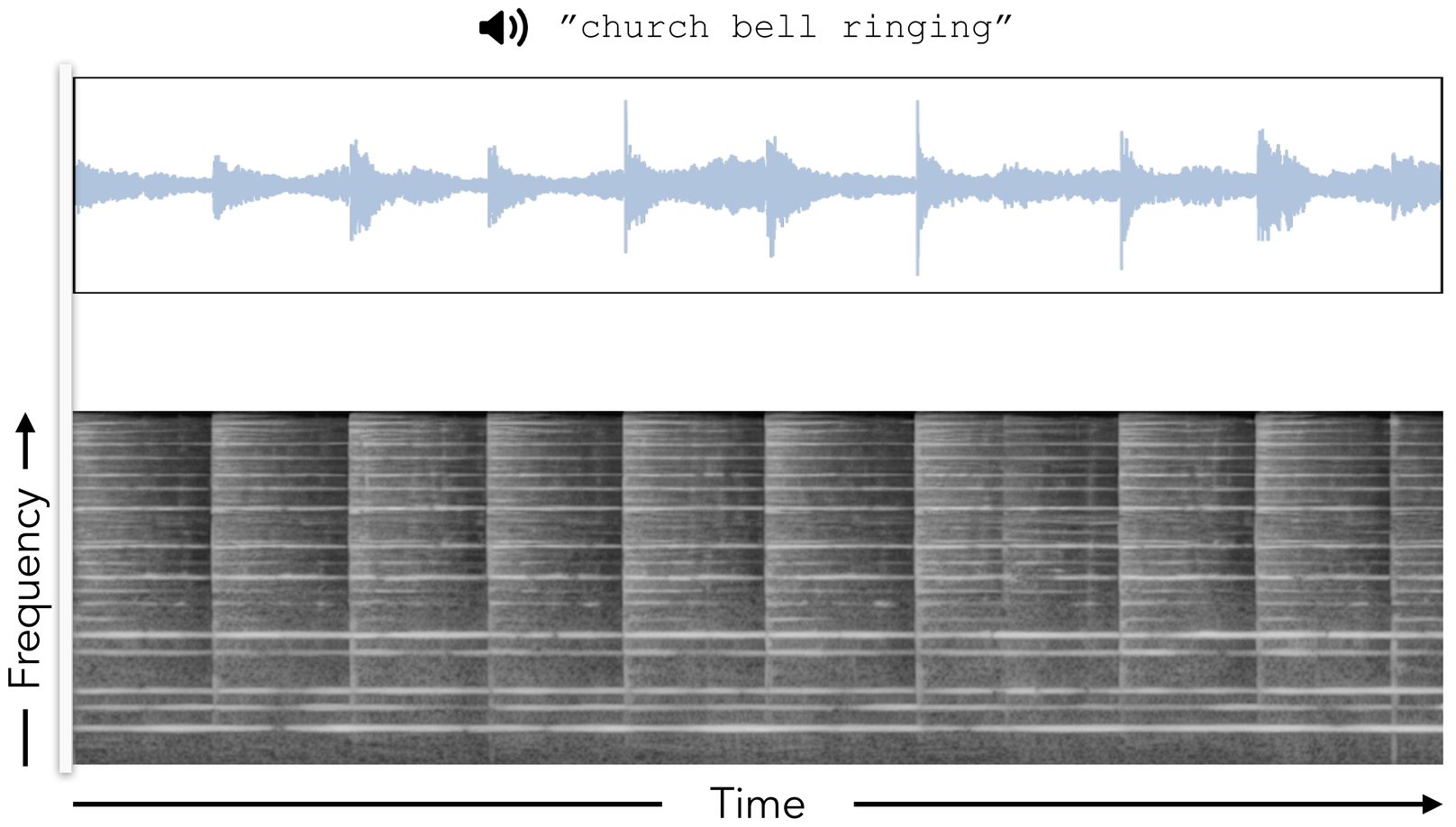


[Chen, Geng, Owens. "Images that Sound: Composing Images and Sounds on a Single Canvas." NeurIPS, 2024]


Ziyang Chen


Daniel Geng

Audio spectrograms



[Chen, Geng, Owens. "Images that Sound: Composing Images and Sounds on a Single Canvas." NeurIPS, 2024]


Ziyang Chen


Daniel Geng

“Seeing” spectrograms

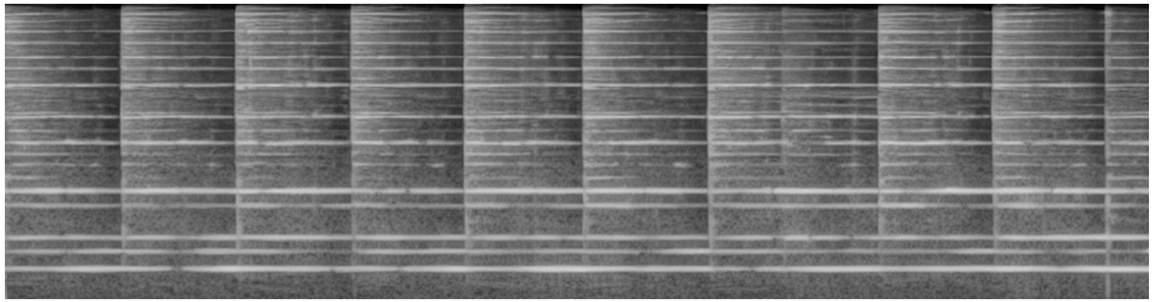
Images

”castles, grayscale”



Spectrograms

”church bell ringing”



“Seeing” spectrograms

Images

”castles, grayscale”



Spectrograms

”church bell ringing”

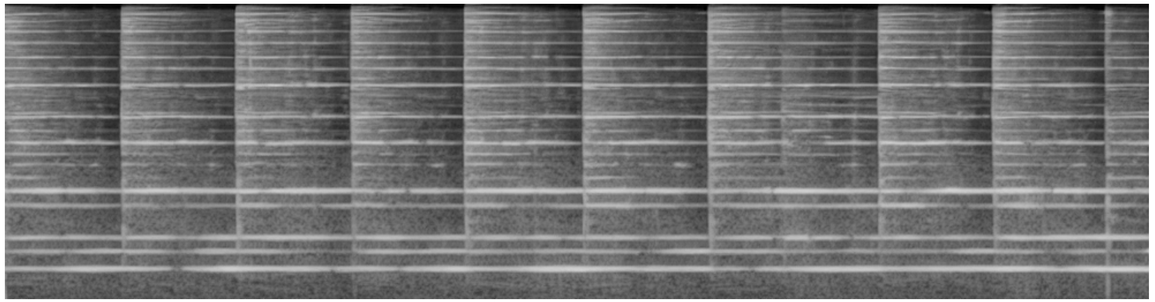
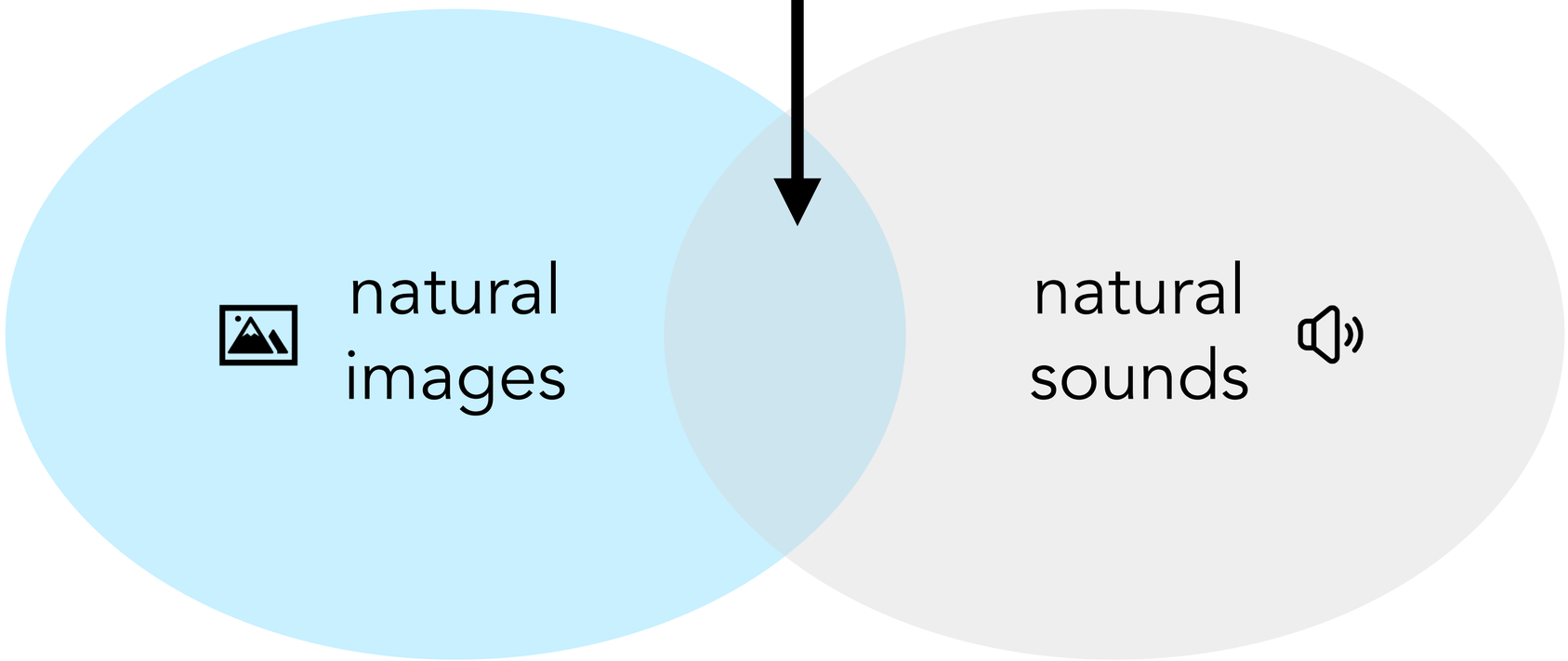


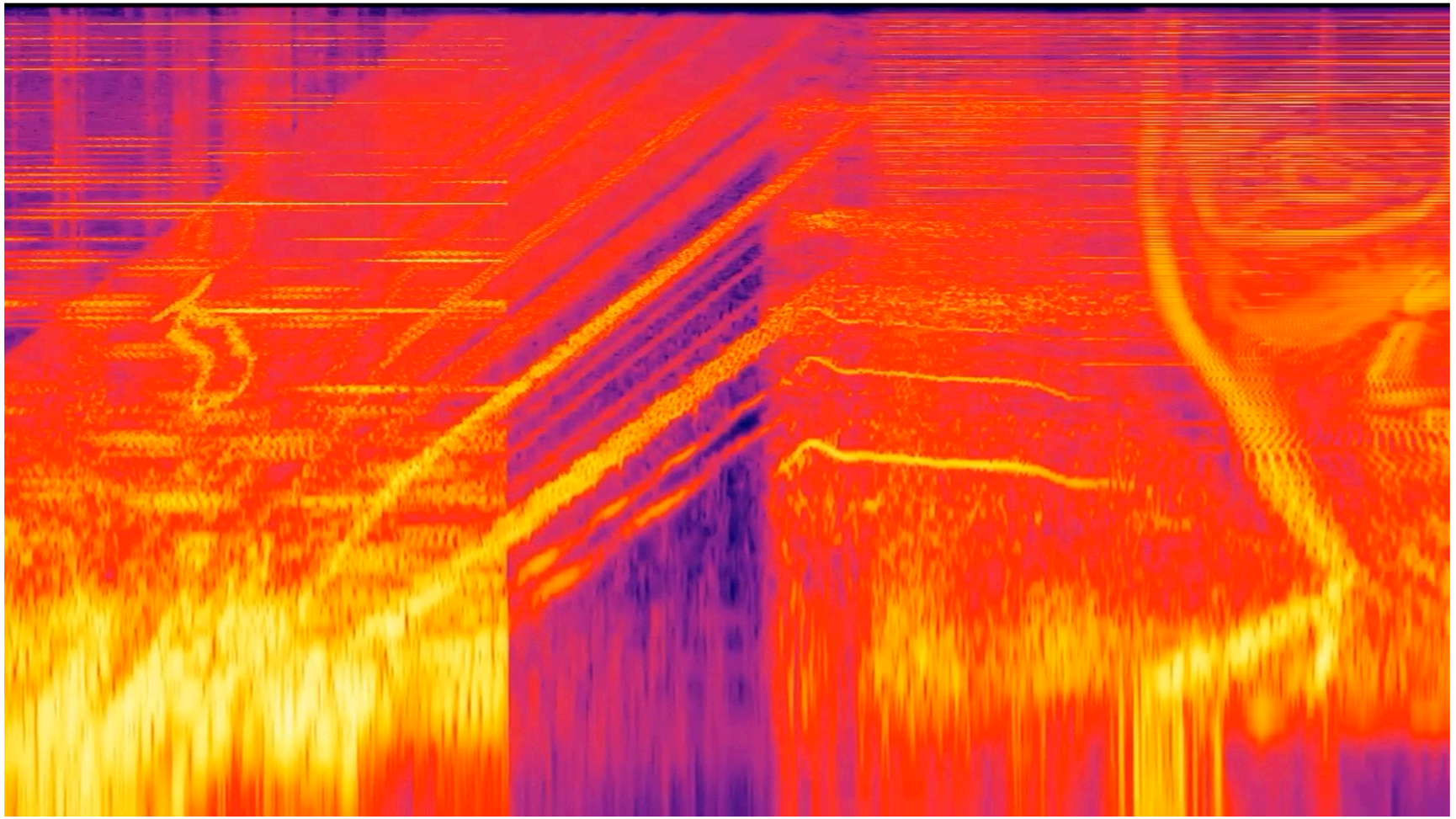
Image captioning
[Li et al., BLIP-2, 2023]

”a black and white photo of
a window with a curtain”

Can we find these?

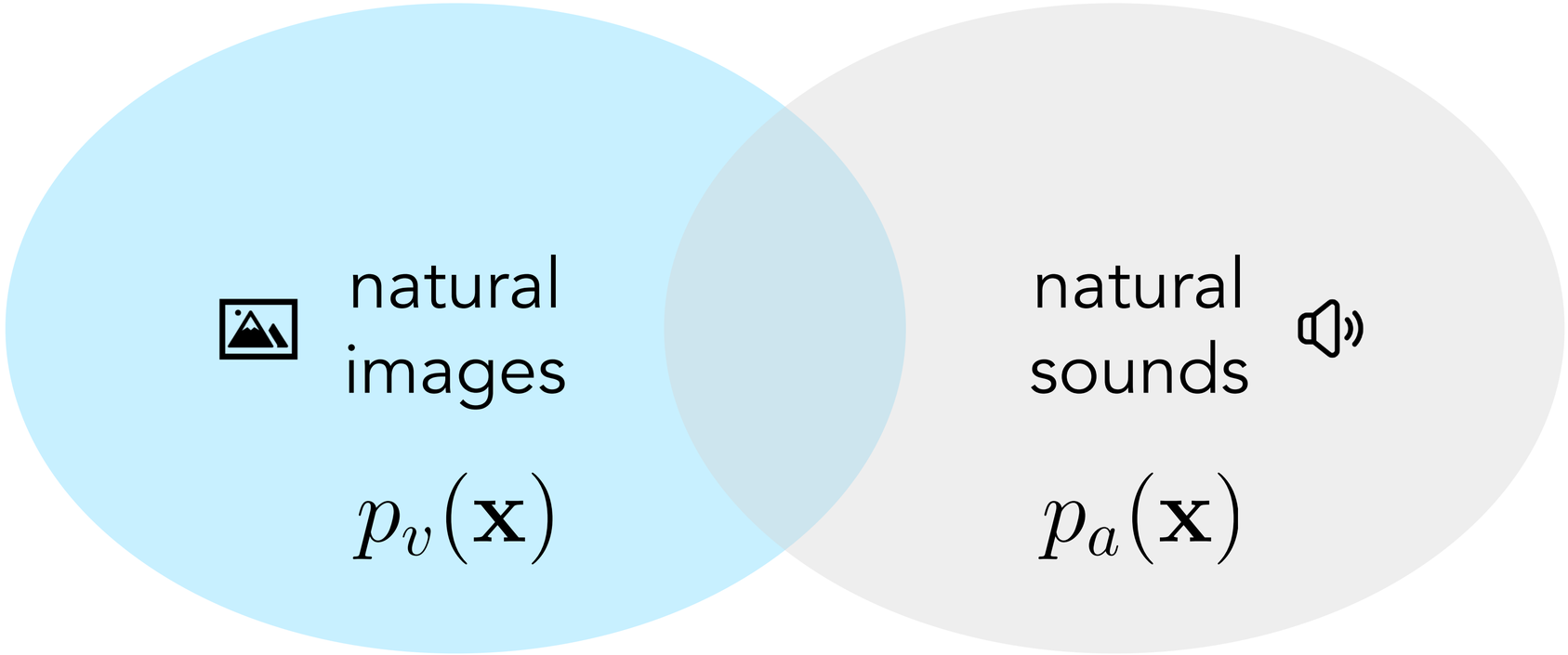


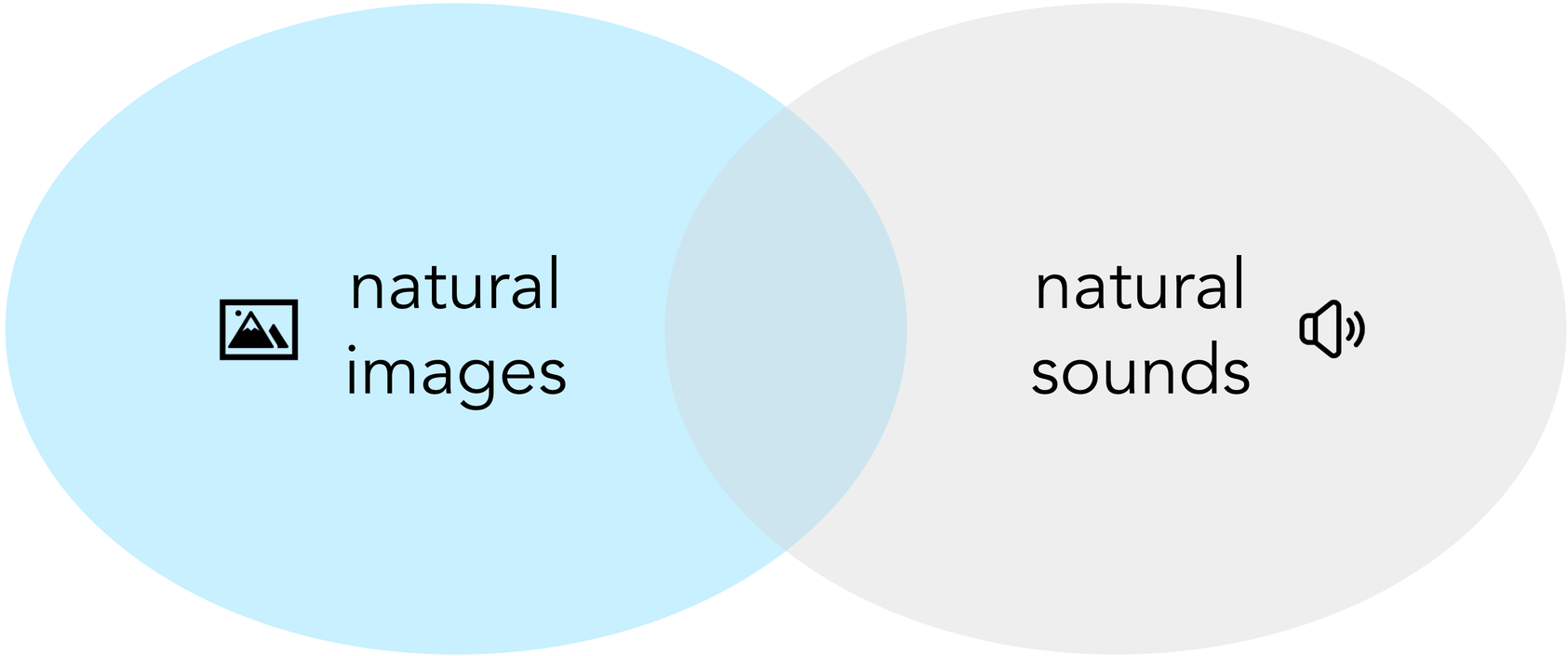
Related work: spectrogram art



Aphex Twin, *Formula, 2001*

See also: [Nine Inch Nails, “My Violent Heart”, 2007], [Venetian Snares, “Songs about my Cats”, 2006]


natural
images
$p_v(\mathbf{x})$
natural
sounds
$p_a(\mathbf{x})$

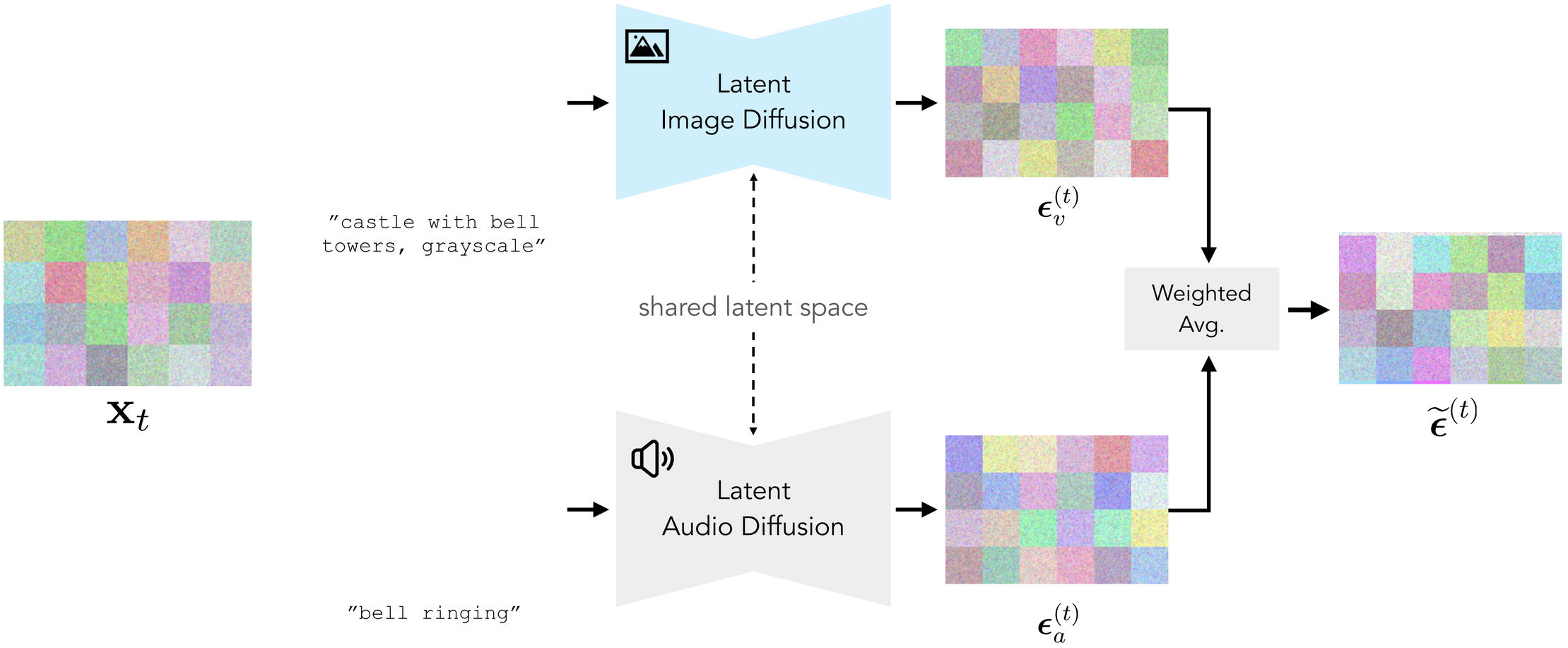


Sample from product of experts:

$$p_{av}(\mathbf{x}) \propto p_v(\mathbf{x}) p_a(\mathbf{x})$$

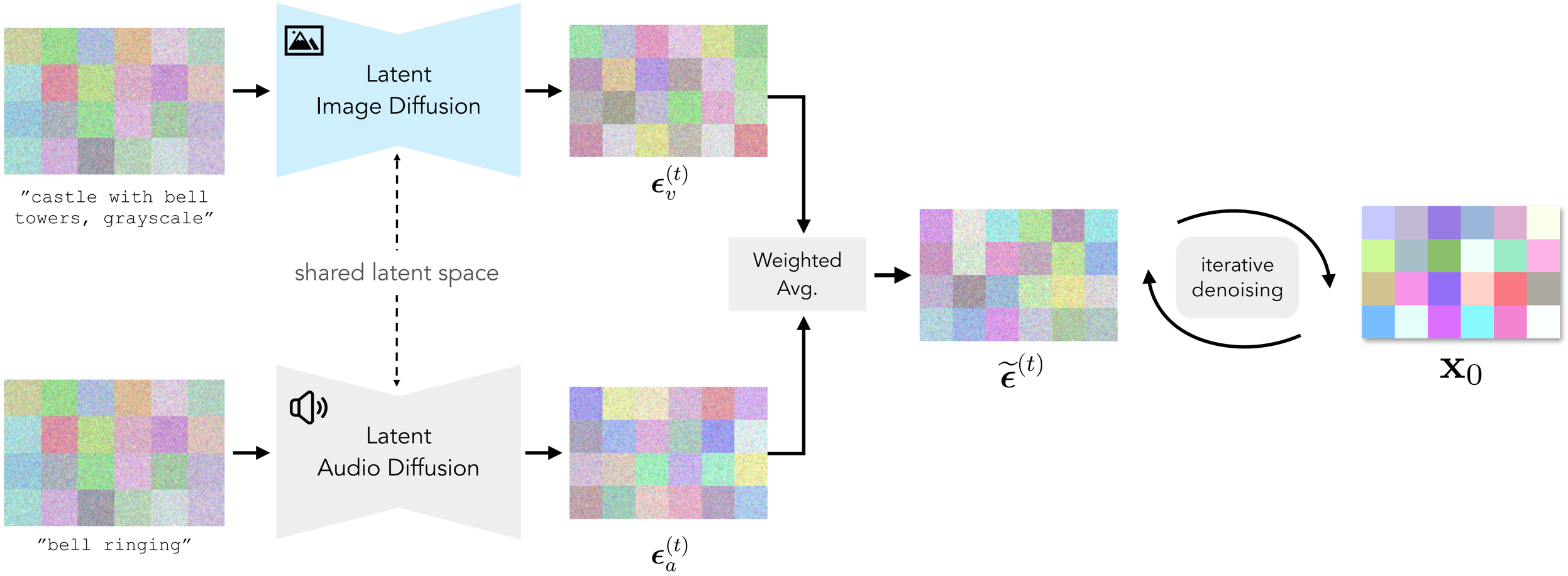
Related work on compositional generation: e.g., [Liu, Li, Du, et al. “Compositional Visual Generation.” ECCV 2022]

Composing sight with sound

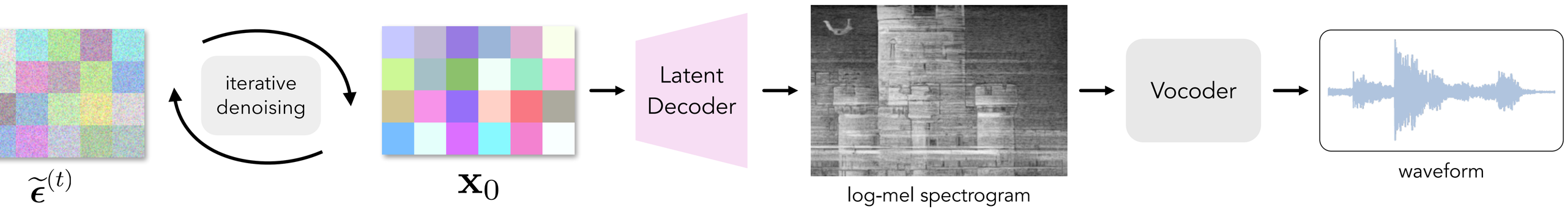


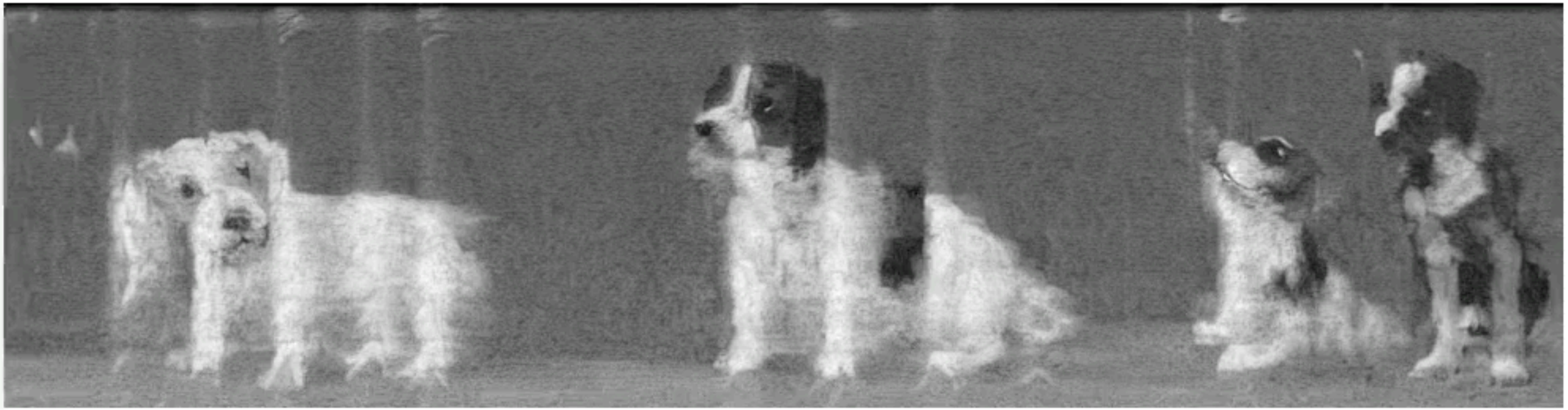
We use Auffusion [Xue et al., 2024], see also [Forsgren et al., "Riffusion", 2022]

Composing sight with sound



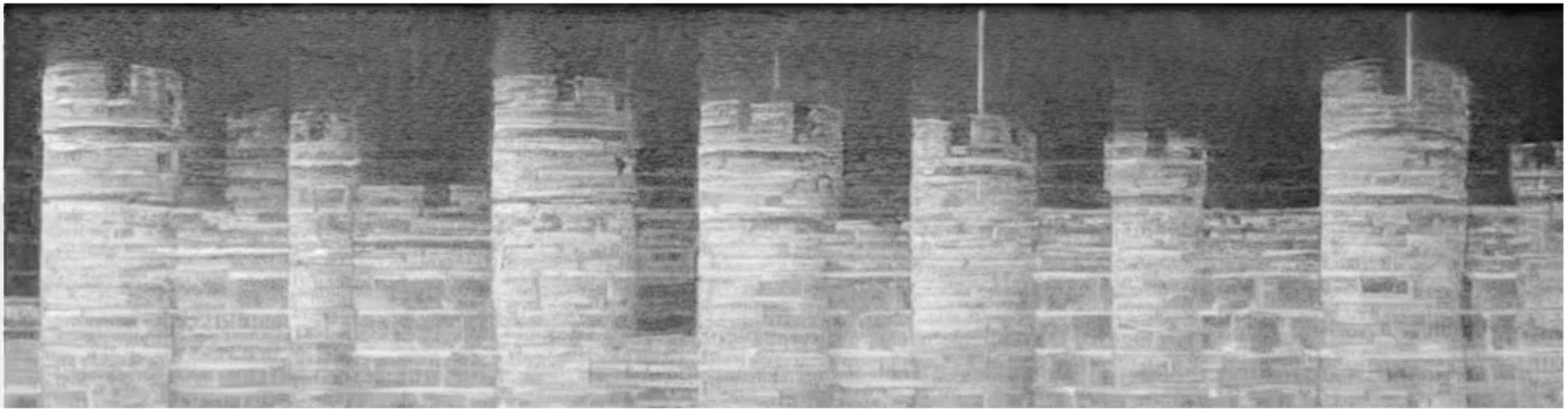
Composing sight with sound





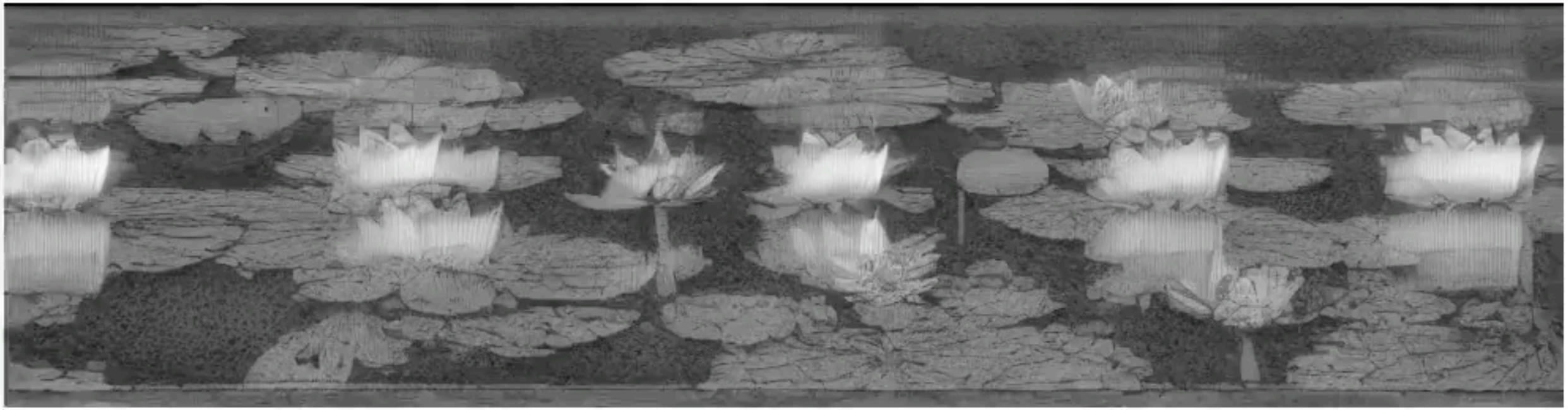
“a painting of cute dogs, grayscale”

“dog barking”



“a painting of castle towers, grayscale”

“bell ringing”



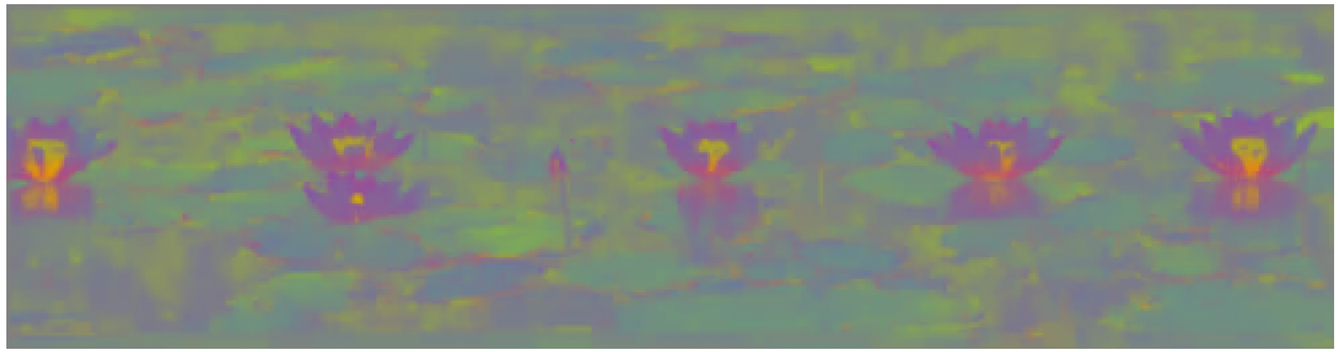
“frog croaking”

“a pond full of water lilies, grayscale, lithograph style”

Color images?



Factorized diffusion



“a colorful photo of a water-lily pond”

[Geng*, Park*, Owens, Factorized Diffusion: Perceptual Illusions by Noise Decomposition, ECCV 2024]