Exploring Free Lunch in Diffusion U-Net

Ziqi Huang

MMLab@NTU | S-Lab, Nanyang Technological University

About Me

- Ziqi Huang 黄子琪
- *Ph.D. student at MMLab@NTU* *2022 Aug – Now*
  - advisor: Prof. Ziwei Liu
  - generative models, visual generation and manipulation
- *Undergraduate* *2018 Aug – 2022 May*
  - Nanyang Technological University (NTU)

S-LAB FOR ADVANCED INTELLIGENCE

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

FreeU: Free Lunch in Diffusion U-Net


Chenyang Si


Ziqi Huang

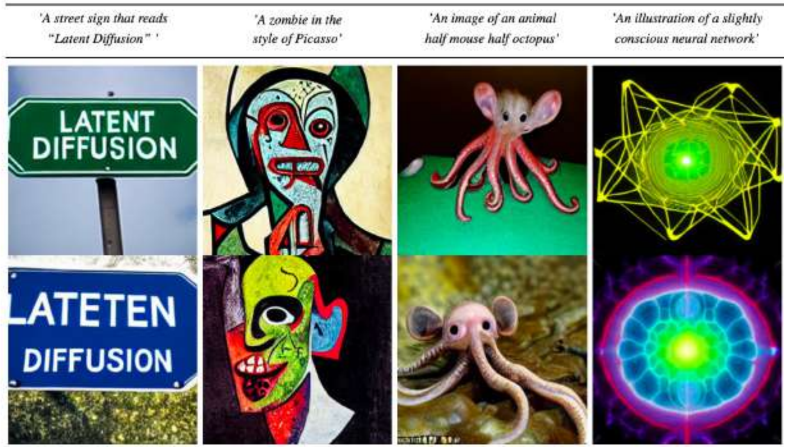

Yuming Jiang


Ziwei Liu

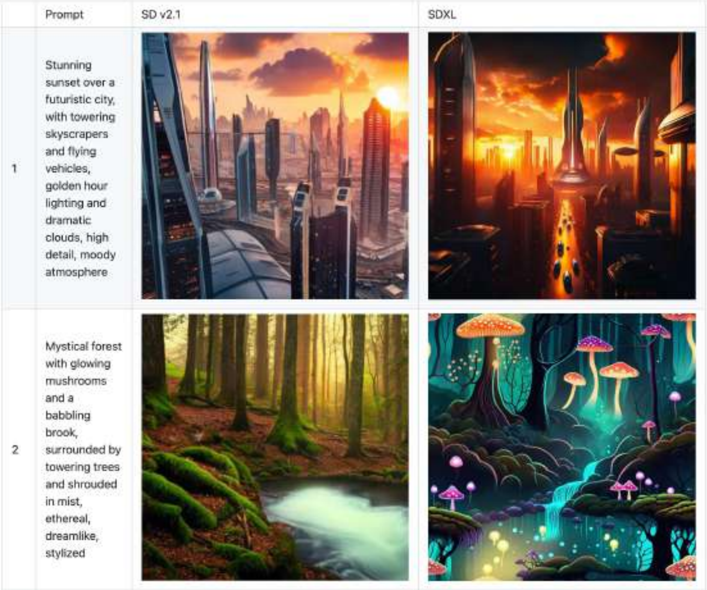
Pre-trained Diffusion Models

- image generation

ADM [1]



LDM [2]

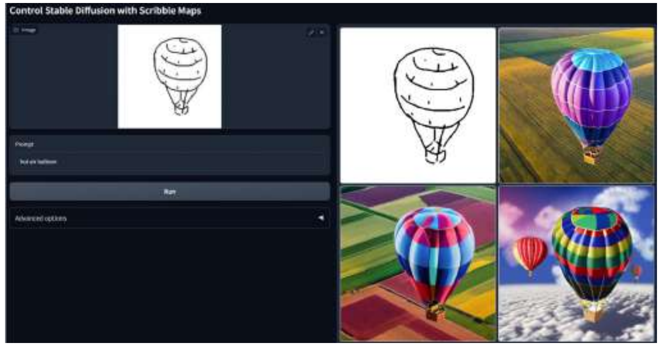


SDXL [3]

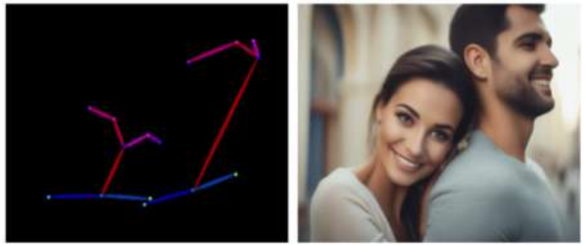
[1] *Dhariwal et al.* Diffusion Models Beat GANs on Image Synthesis
[2] *Rombach et al.* High-resolution image synthesis with latent diffusion models
[3] *Podell et al.* SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis

Pre-trained Diffusion Models

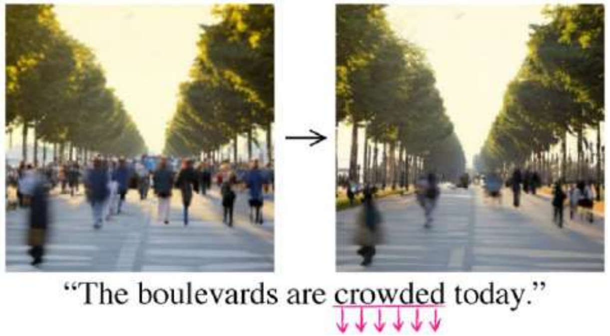
- controllable generation / editing / translation



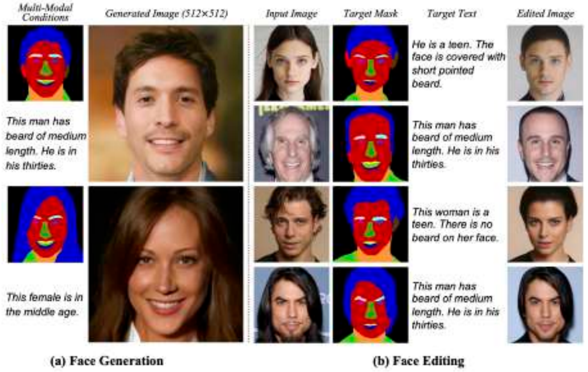
ControlNet [1]



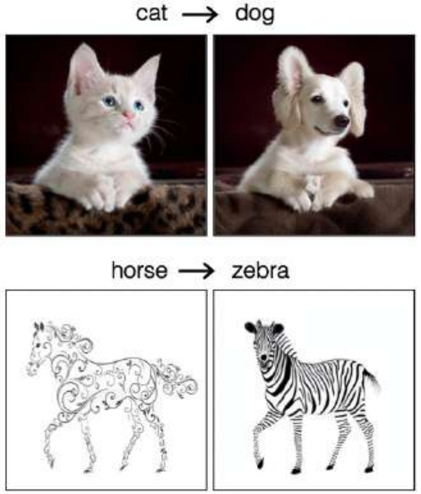
T2I-Adapter [2]



Prompt-to-Prompt [3]



Collaborative Diffusion [4]

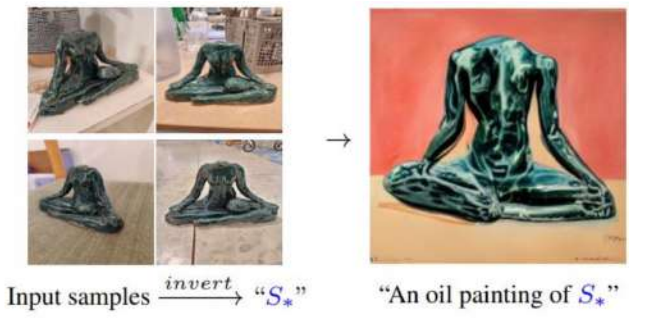


pix2pix-zero [5]

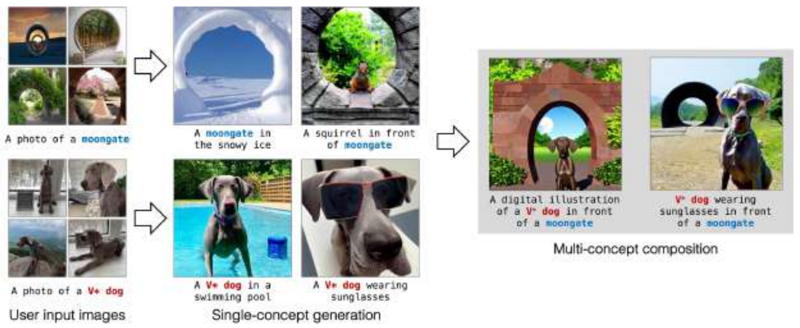
[1] *Zhang et al. Adding Conditional Control to Text-to-Image Diffusion Models*
[2] *Mou et al. T2I-Adapter: Learning Adapters to Dig out More Controllable Ability for Text-to-Image Diffusion Models*
[3] *Hertz et al. Prompt-to-Prompt Image Editing with Cross-Attention Control*
[4] *Huang et al. Collaborative Diffusion for Multi-Modal Face Generation and Editing*
[5] *Parmar et al.* Zero-shot Image-to-Image Translation

Pre-trained Diffusion Models

- add / remove concepts for a pre-trained diffusion model



Textual Inversion [1]

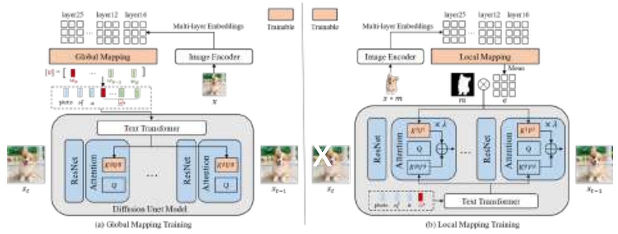


Custom Diffusion [3]

Ablating Concepts [5]

DreamBooth [2]

ReVersion [4]



ELITE [6]

[1] *Gal et al.* An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion
[2] *Ruiz et al.* DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation
[3] *Kumari et al.* Multi-Concept Customization of Text-to-Image Diffusion
[4] *Huang et al.* ReVersion : Diffusion-Based Relation Inversion from Images
[5] *Kumari et al.* Ablating Concepts in Text-to-Image Diffusion Models
[6] *Wei et al.* ELITE: Encoding Visual Concepts into Textual Embeddings for Customized Text-to-Image Generation

Pre-trained Diffusion Models

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

- video generation

VideoLDM [1]

VideoCrafter [2] […]

LaVie [3]

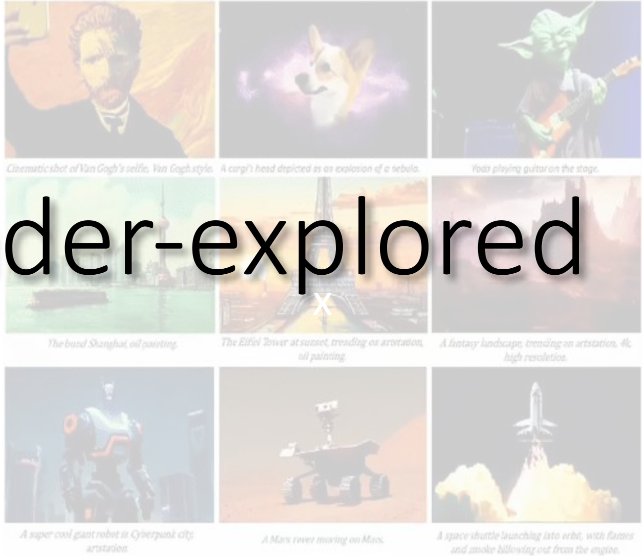
[1] *Blattmann et al.* Align your Latents: High-Resolution Video Synthesis with Latent Diffusion Models
[2] *He et el.* Latent Video Diffusion Models for High-Fidelity Long Video Generation (And more)
[3] *Wang et al.* LaVie: High-Quality Video Generation with Cascaded Latent Diffusion Models

Pre-trained Diffusion Models

- video generation

VideoLDM [1]

VideoCrafter [2] […]



LaVie [6]

Diffusion U-Net remains under-explored

[1] *Blattmann et al.* Align your Latents: High-Resolution Video Synthesis with Latent Diffusion Models
[2] *He et el.* Latent Video Diffusion Models for High-Fidelity Long Video Generation (And more)
[6] *Wang et al.* LaVie: High-Quality Video Generation with Cascaded Latent Diffusion Models

Motivation

- Downstream applications
  - directly utilizing pre-trained diffusion U-Nets
  - internal properties of diffusion U-Net features remain under-explored
- Train better foundation models
  - expensive (*e.g.,* SDXL)
  - besides scaling up (*e.g.,* data scale, model size), what else can we do?
- Why not exploit pre-trained diffusion models?
  - Let's take a closer look at *diffusion U-Net* and the *denoising process*

Diffusion Models

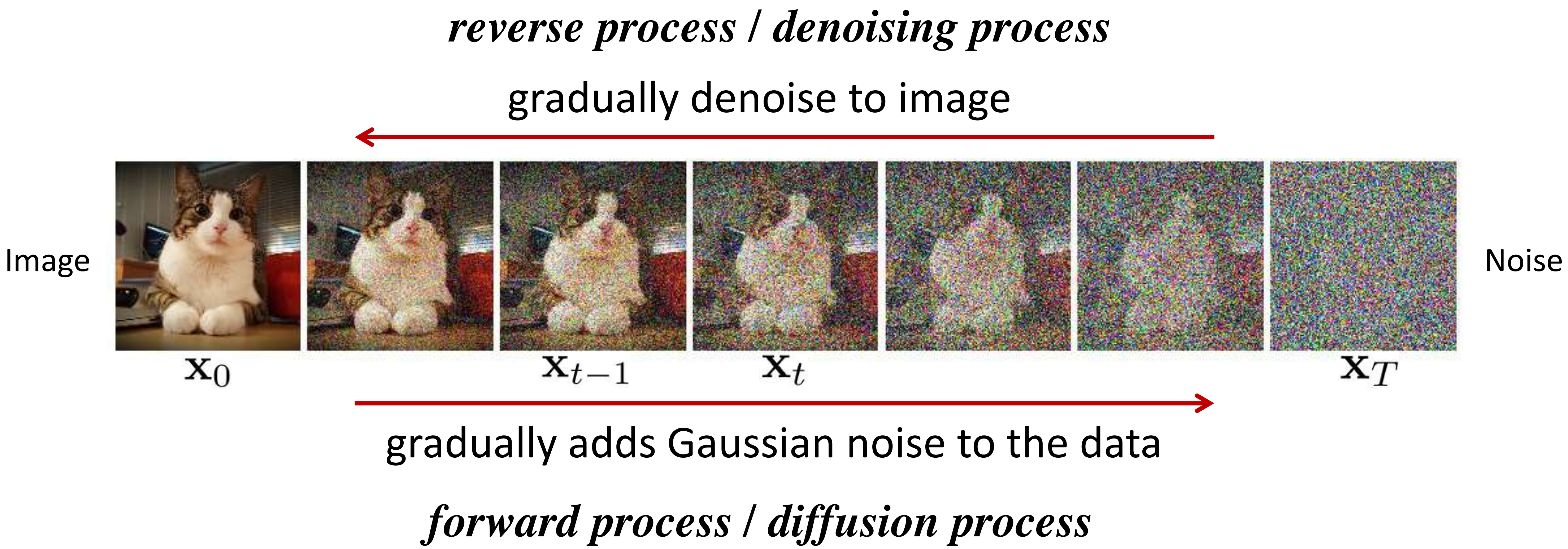
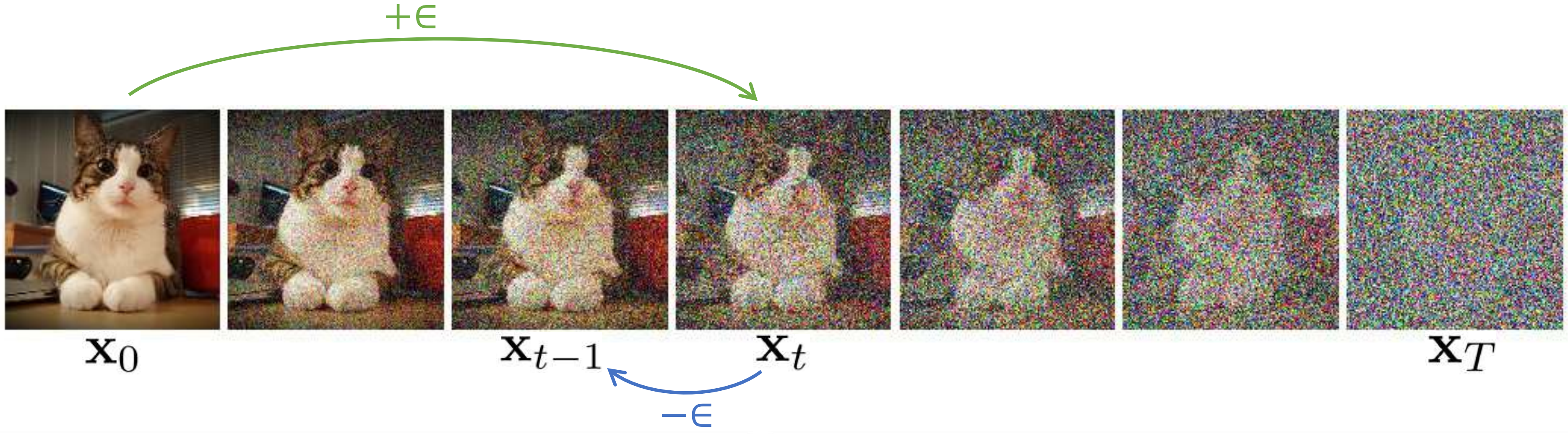


Image Credit: CVPR 2022 Tutorial: Denoising Diffusion-based Generative Modeling: Foundations and Applications

Training & Sampling



Algorithm 1 Training

1: **repeat**
2: $\mathbf{x}_0 \sim q(\mathbf{x}_0)$
3: $t \sim \mathrm{Uniform}(\{1, \ldots, T\})$
4: $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
5: Take gradient descent step on
$$\nabla_\theta \left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta\left(\boxed{\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}}, t\right) \right\|^2$$
6: **until** converged

Algorithm 2 Sampling

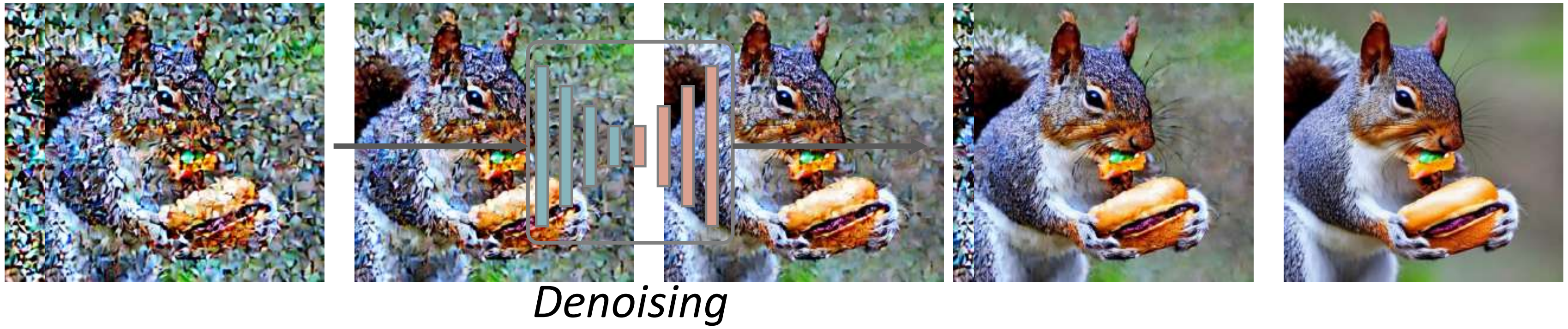
1: $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
2: **for** $t = T, \ldots, 1$ **do**
3: $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\mathbf{z} = \mathbf{0}$
4: $\mathbf{x}_{t-1} = \boxed{\frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\right)} + \sigma_t \mathbf{z}$
5: **end for**
6: **return** $\mathbf{x}_0$

Ho et al. Denoising Diffusion Probabilistic Models
Image Credit: CVPR 2022 Tutorial: Denoising Diffusion-based Generative Modeling: Foundations and Applications

Closer look at the denoising process

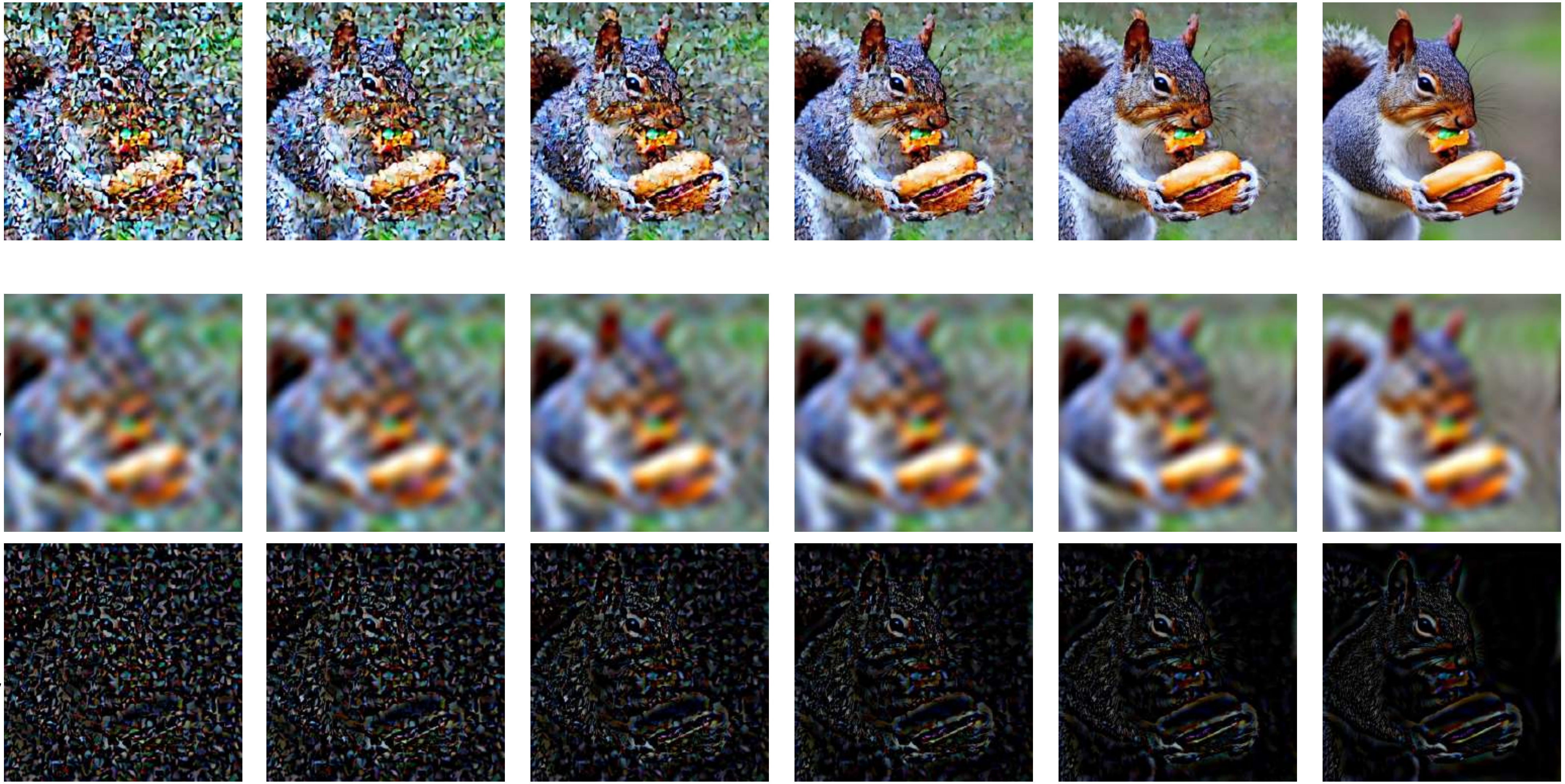
Denoising Process

Input: *A squirrel eating a burger*



Denoising Process

Input: *A squirrel eating a burger*

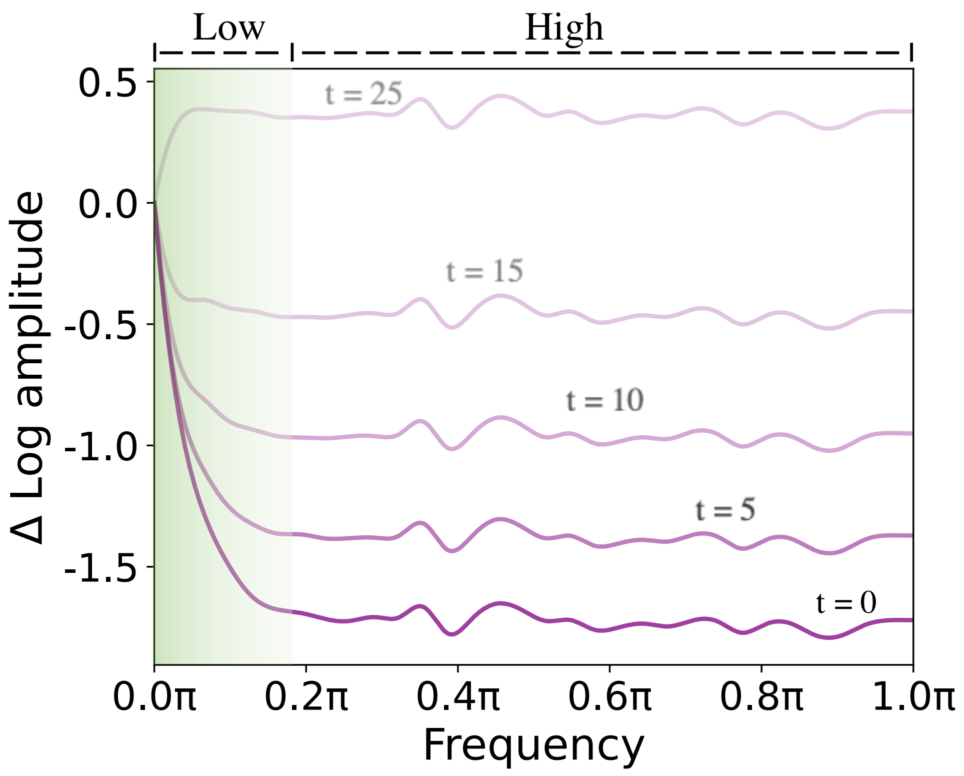


Low frequency

High frequency

Denoising Process

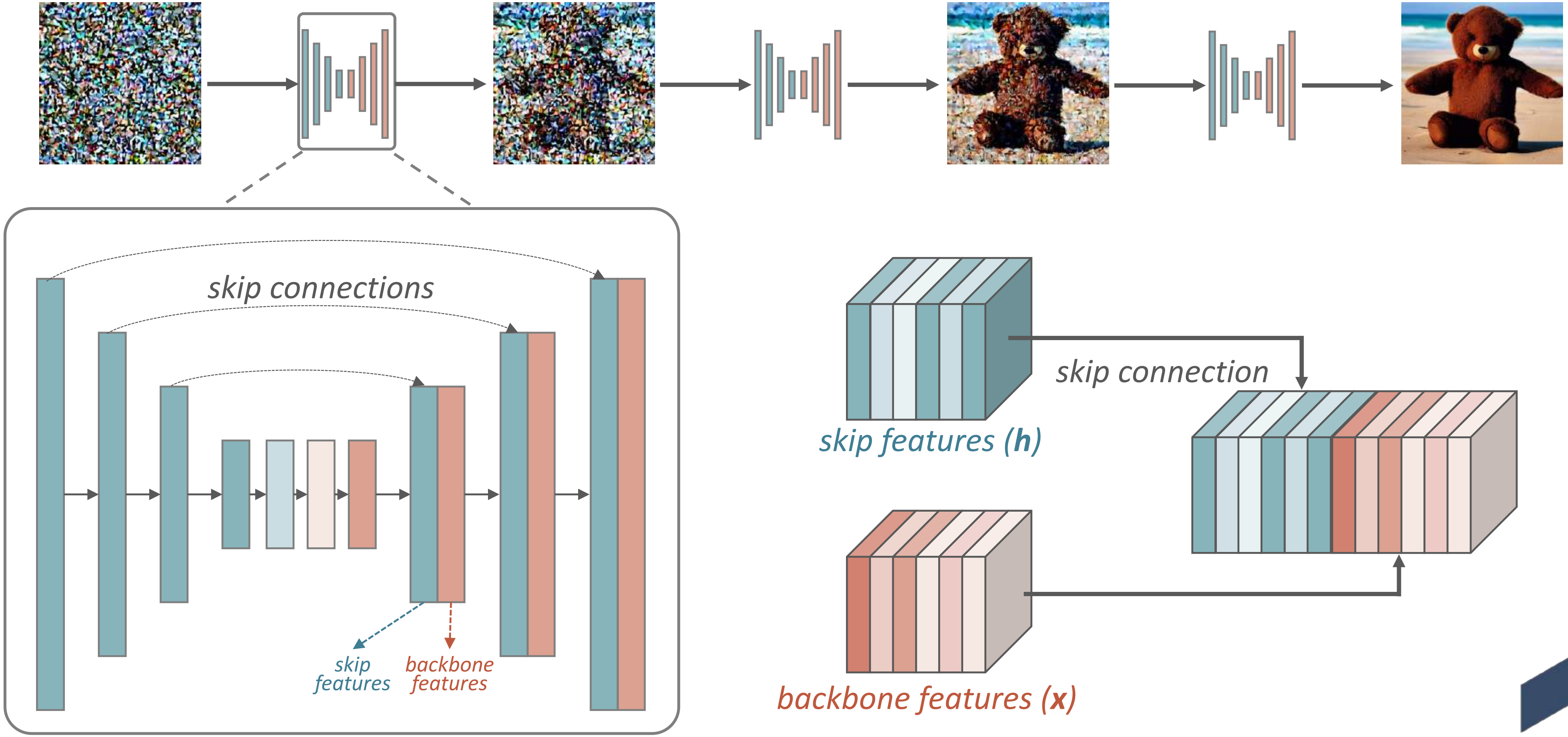
- The high-frequency components of $\boldsymbol{x}_t$ drops drastically during the denoising process



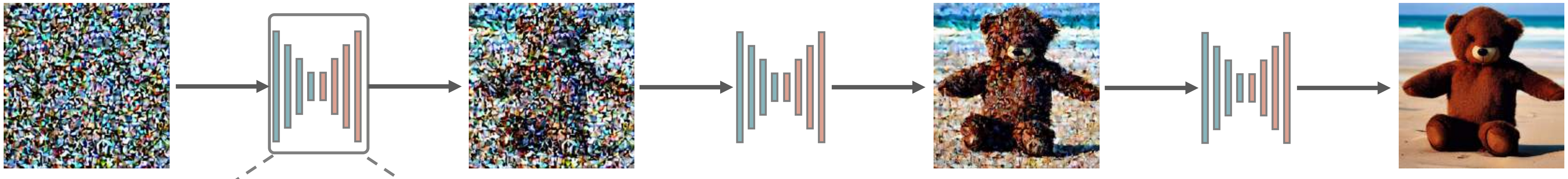
Relative log amplitudes of Fourier for diffusion intermediate steps

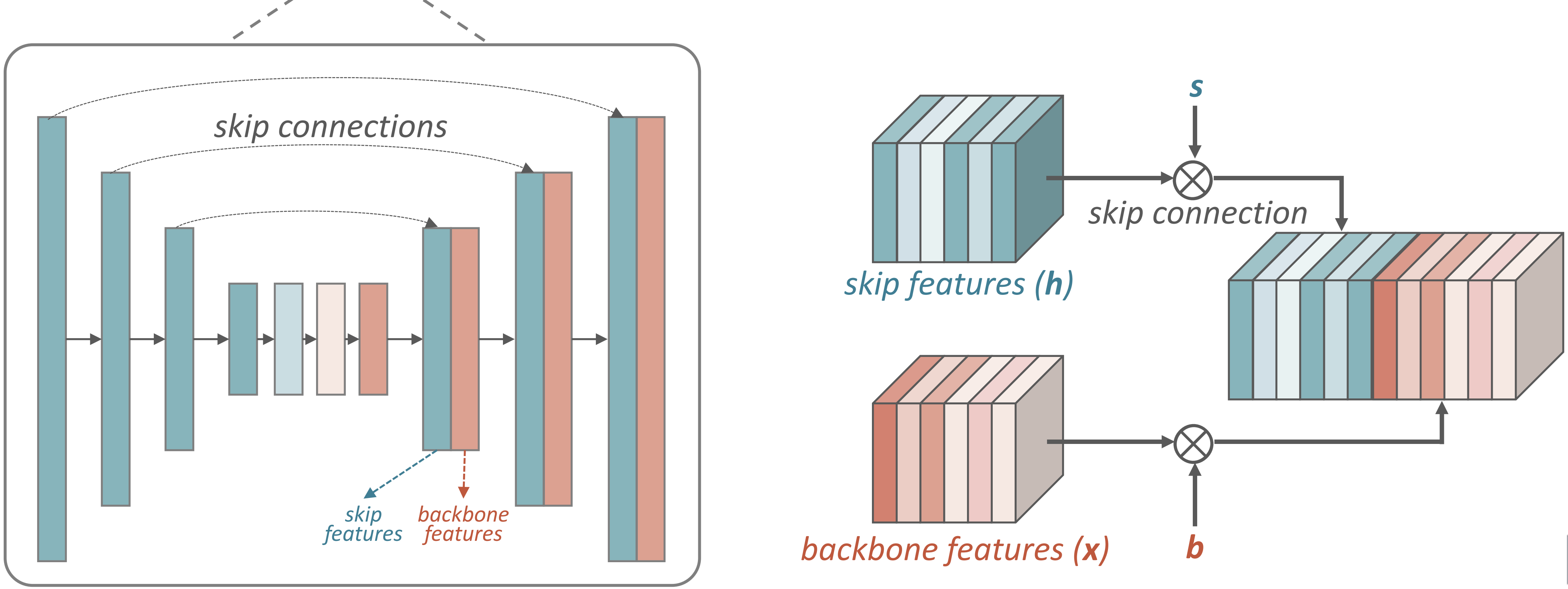
How does diffusion U-Net perform denoising?

Denoising Process: U-Net



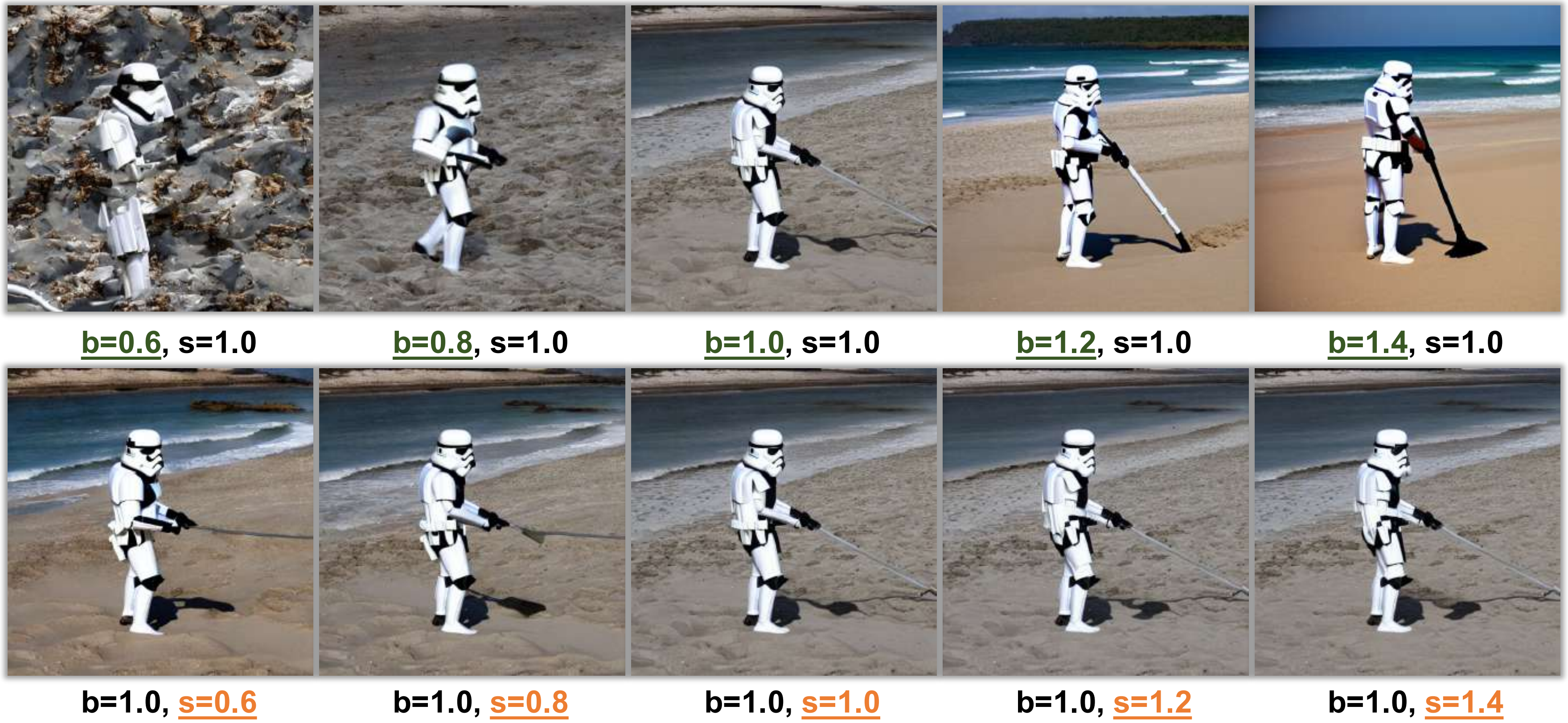
Denoising Process: U-Net





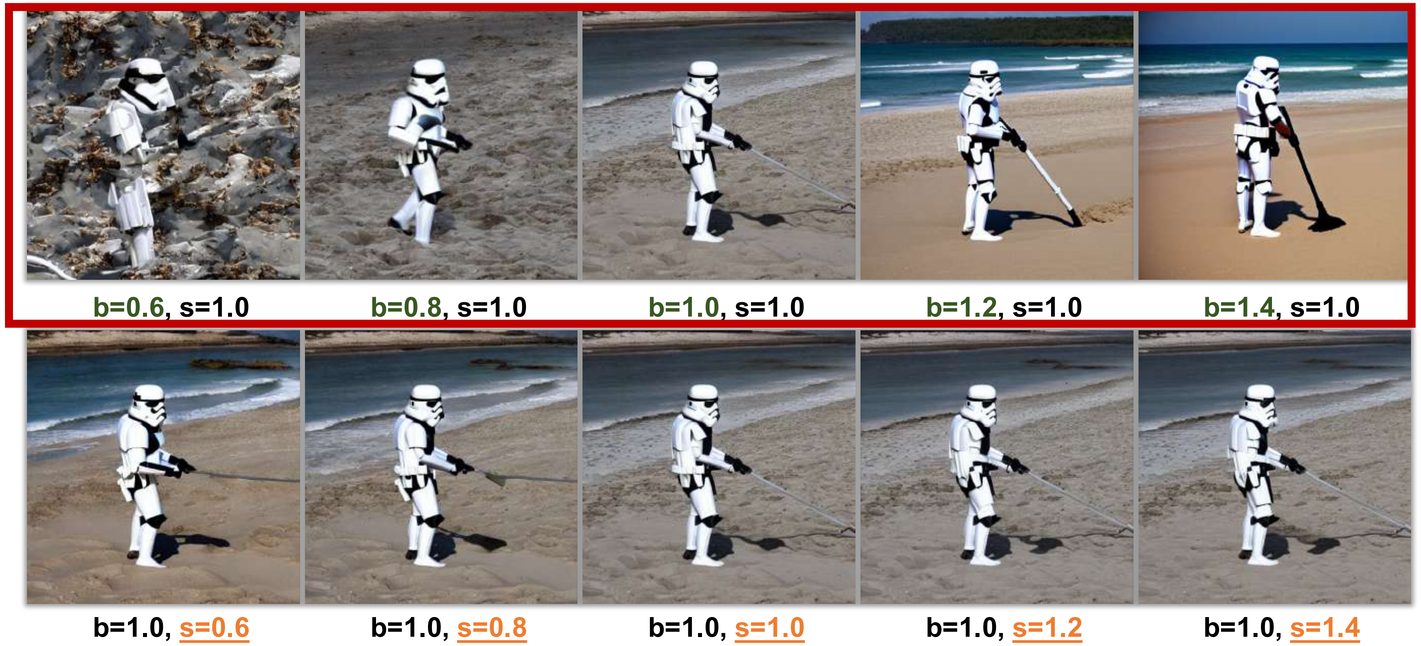
Role of Backbone and Skip Features

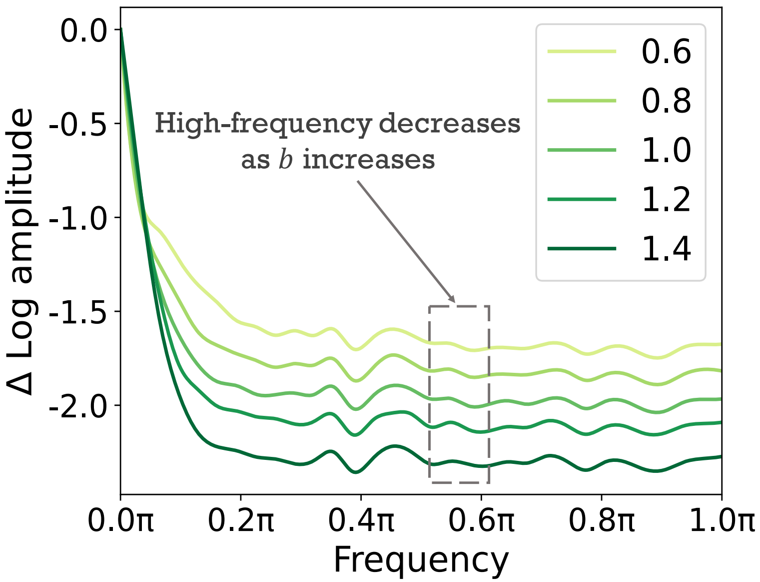
- **Backbone**: denoising
- **Skip**: limited impact during inference



How Diffusion U-Net Perform Denoising?

- **Backbone**: primarily contributes to denoising
  - Consistent with previous observation (next page)

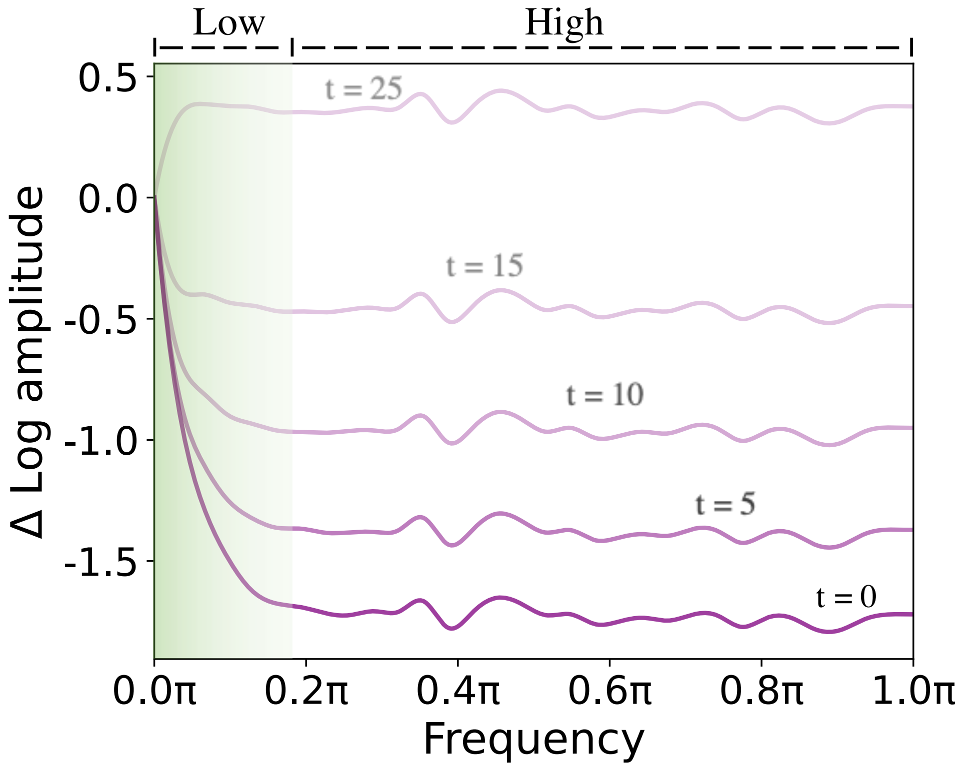




Fourier relative log amplitudes of variations of b

Denoising Process

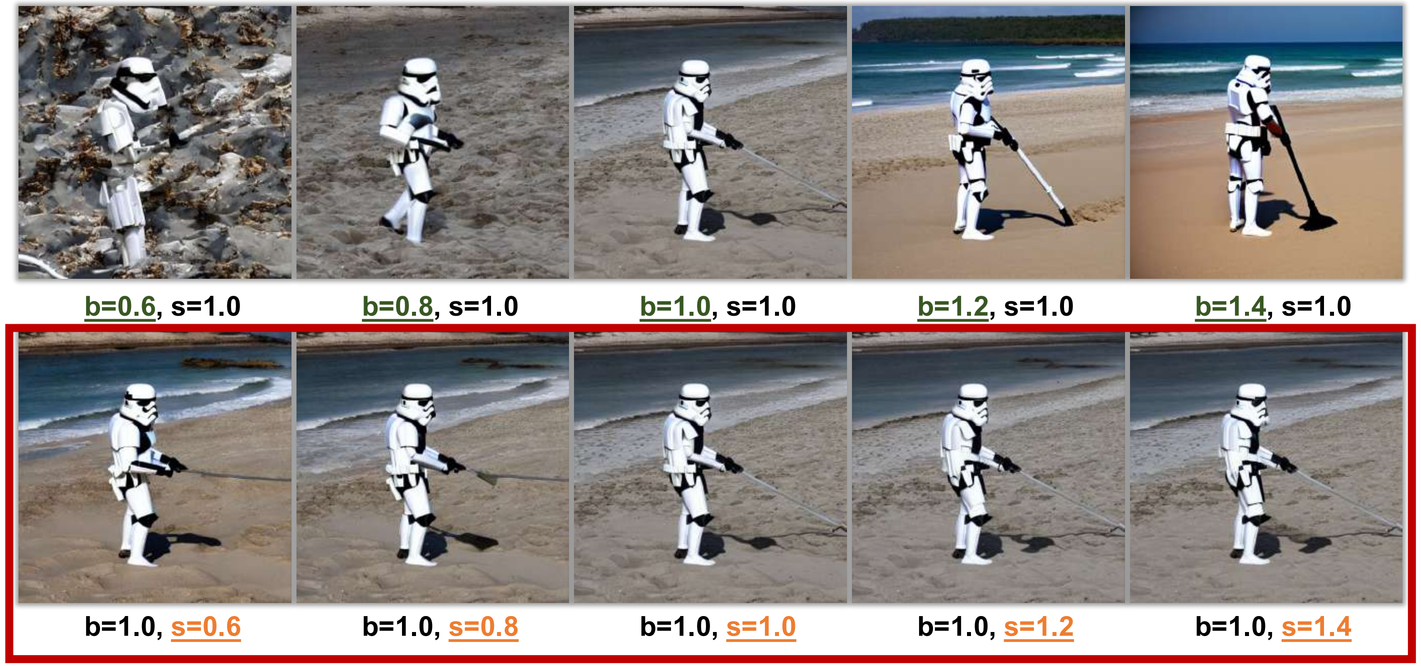
- The high-frequency components of $\boldsymbol{x}_t$ drops drastically during the denoising process

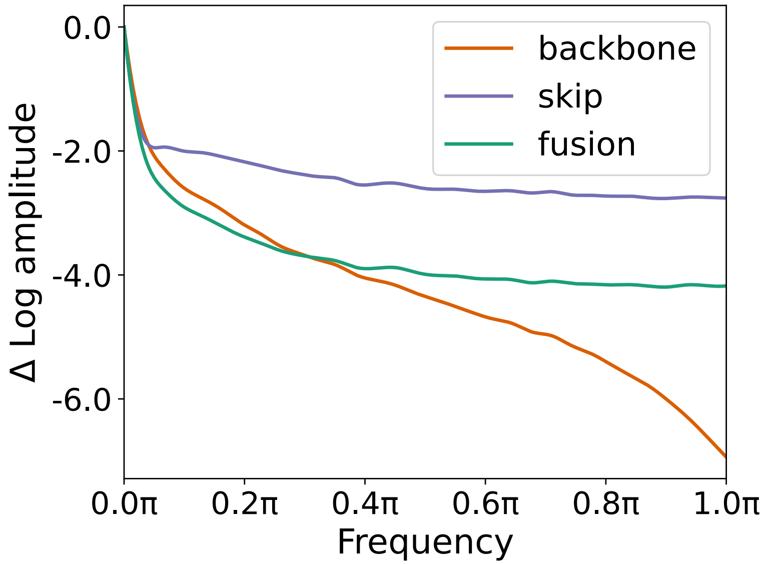


Relative log amplitudes of Fourier for diffusion intermediate steps

How Diffusion U-Net Perform Denoising?

- **Backbone**: primarily contributes to denoising
- **Skip**: introduce high-frequency features into the decoder module

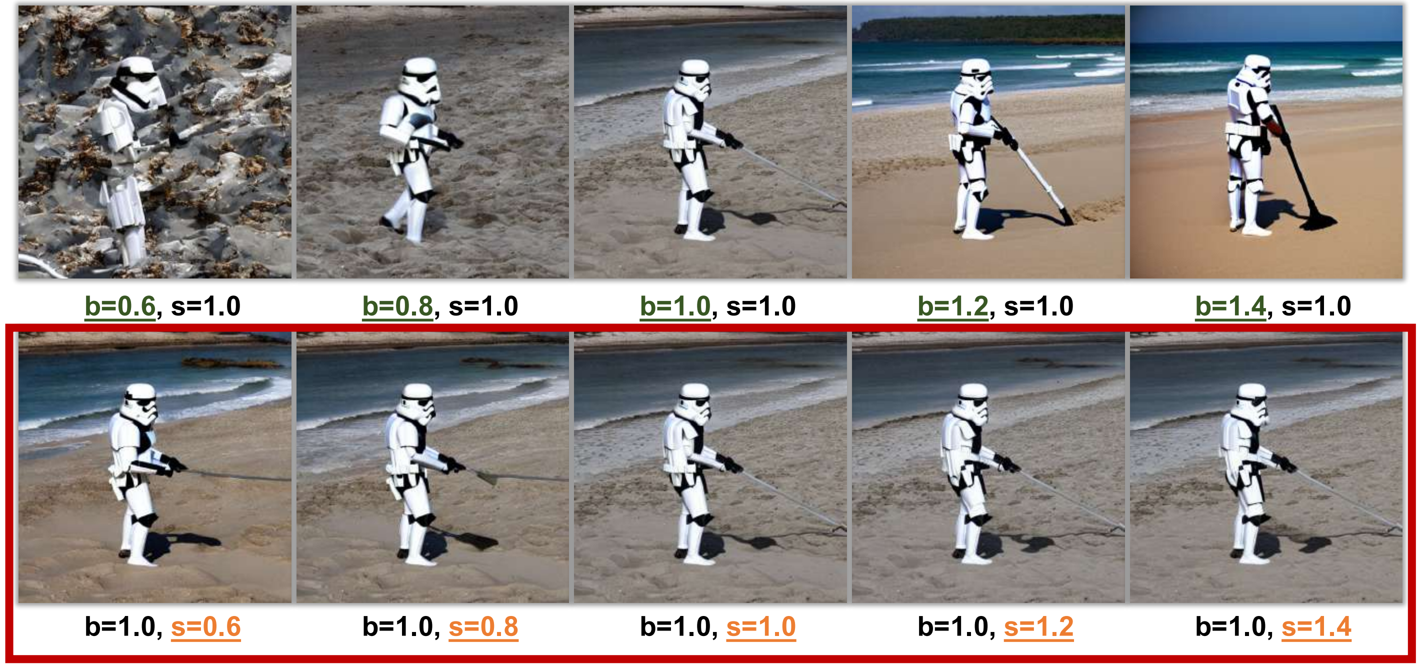


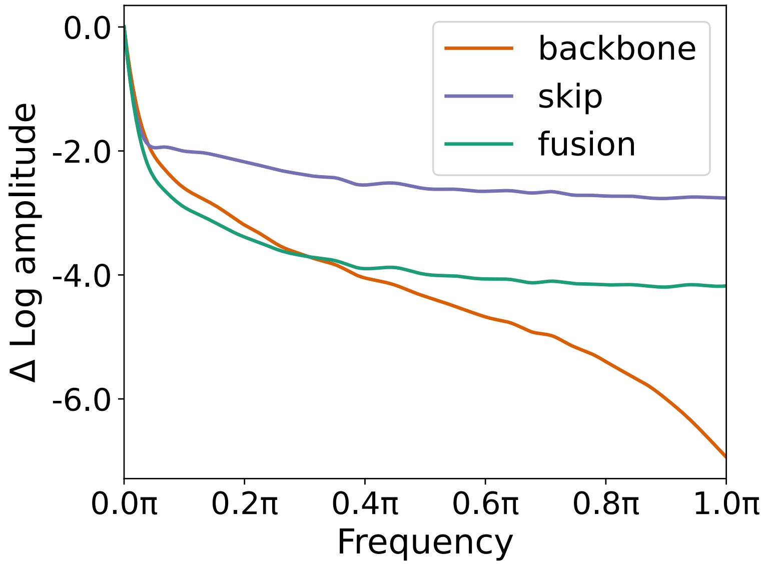


Fourier relative log amplitudes of backbone, skip, and their fused feature maps

How Diffusion U-Net Perform Denoising?

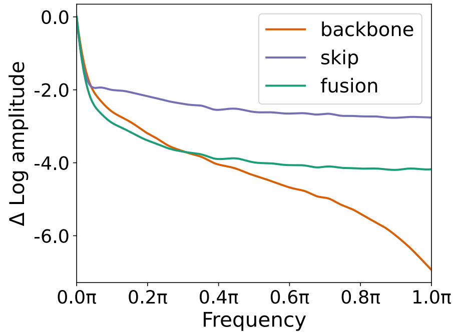
- Gap between training and sampling



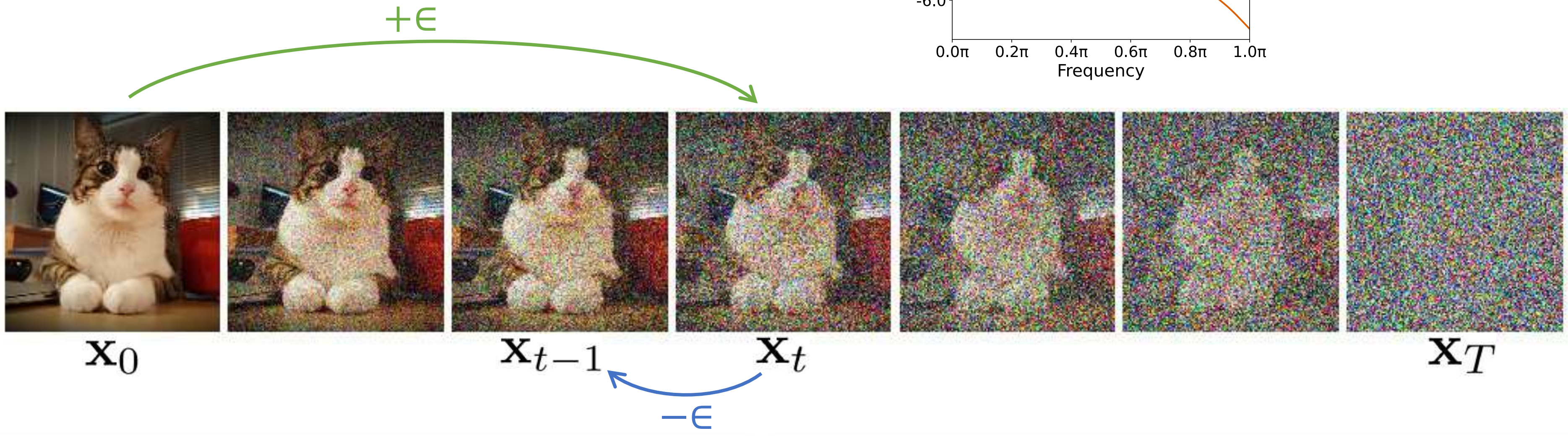


Fourier relative log amplitudes of backbone, skip, and their fused feature maps

Training & Sampling



Image



Noise

Algorithm 1 Training

1: **repeat**
2: $\quad \mathbf{x}_0 \sim q(\mathbf{x}_0)$
3: $\quad t \sim \mathrm{Uniform}(\{1, \ldots, T\})$
4: $\quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
5: $\quad$ Take gradient descent step on
$$\nabla_\theta \left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta\left(\boxed{\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}}, t\right) \right\|^2$$
6: **until** converged

Algorithm 2 Sampling

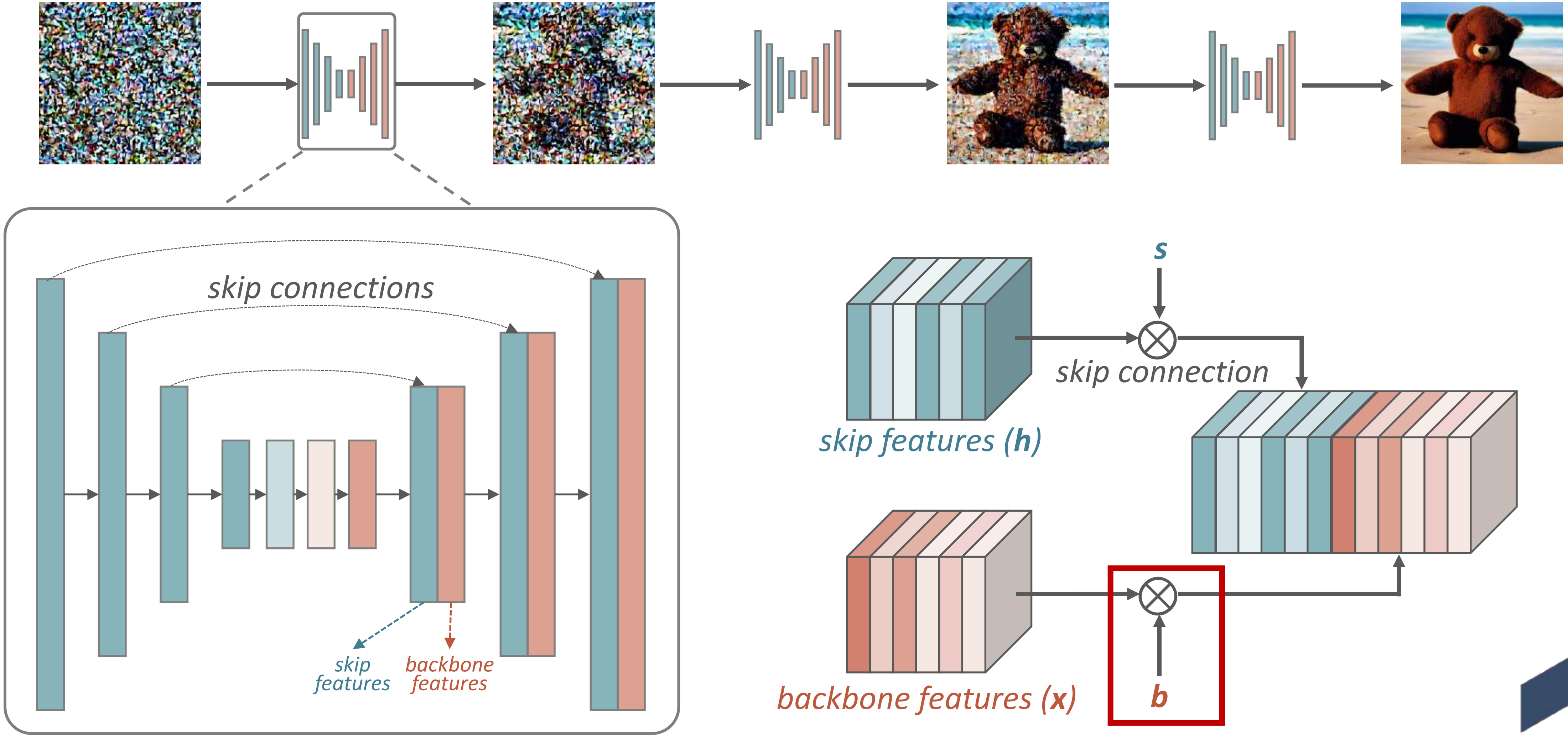
1: $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
2: **for** $t = T, \ldots, 1$ **do**
3: $\quad \mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\mathbf{z} = \mathbf{0}$
4: $\quad \mathbf{x}_{t-1} = \boxed{\frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\right)} + \sigma_t \mathbf{z}$
5: **end for**
6: **return** $\mathbf{x}_0$

Song et al. Denoising diffusion implicit models. (ICLR 2021)

Image Credit: CVPR 2022 Tutorial: Denoising Diffusion-based Generative Modeling: Foundations and Applications

FreeU Method

(1) enhance backbone features



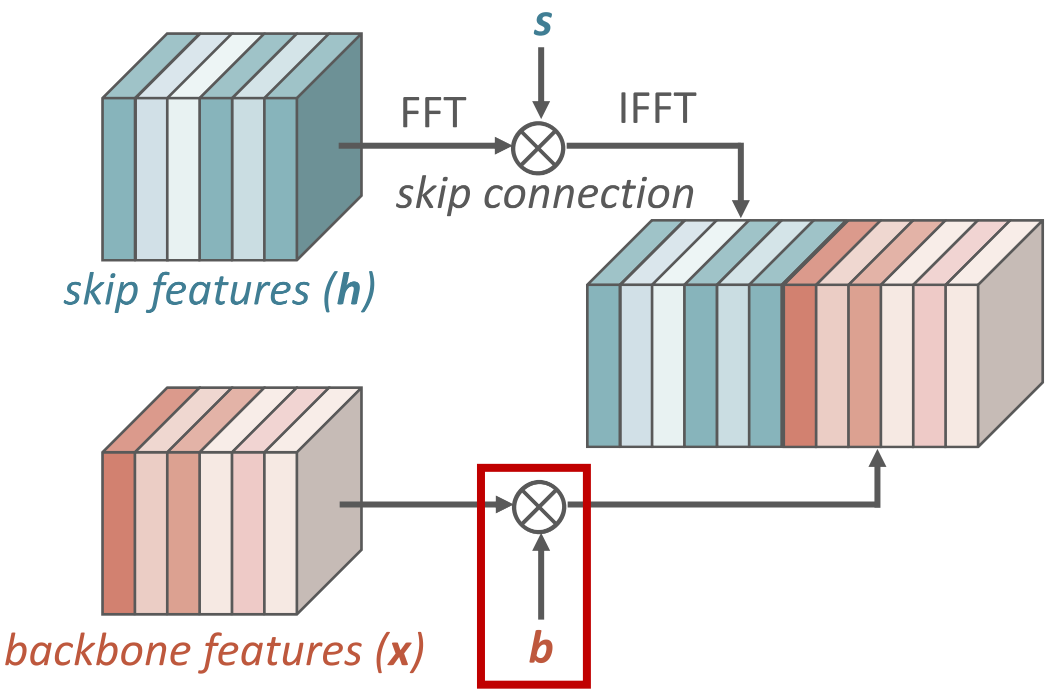
(a) UNet Architecture

(b) FreeU Operations

FreeU Method

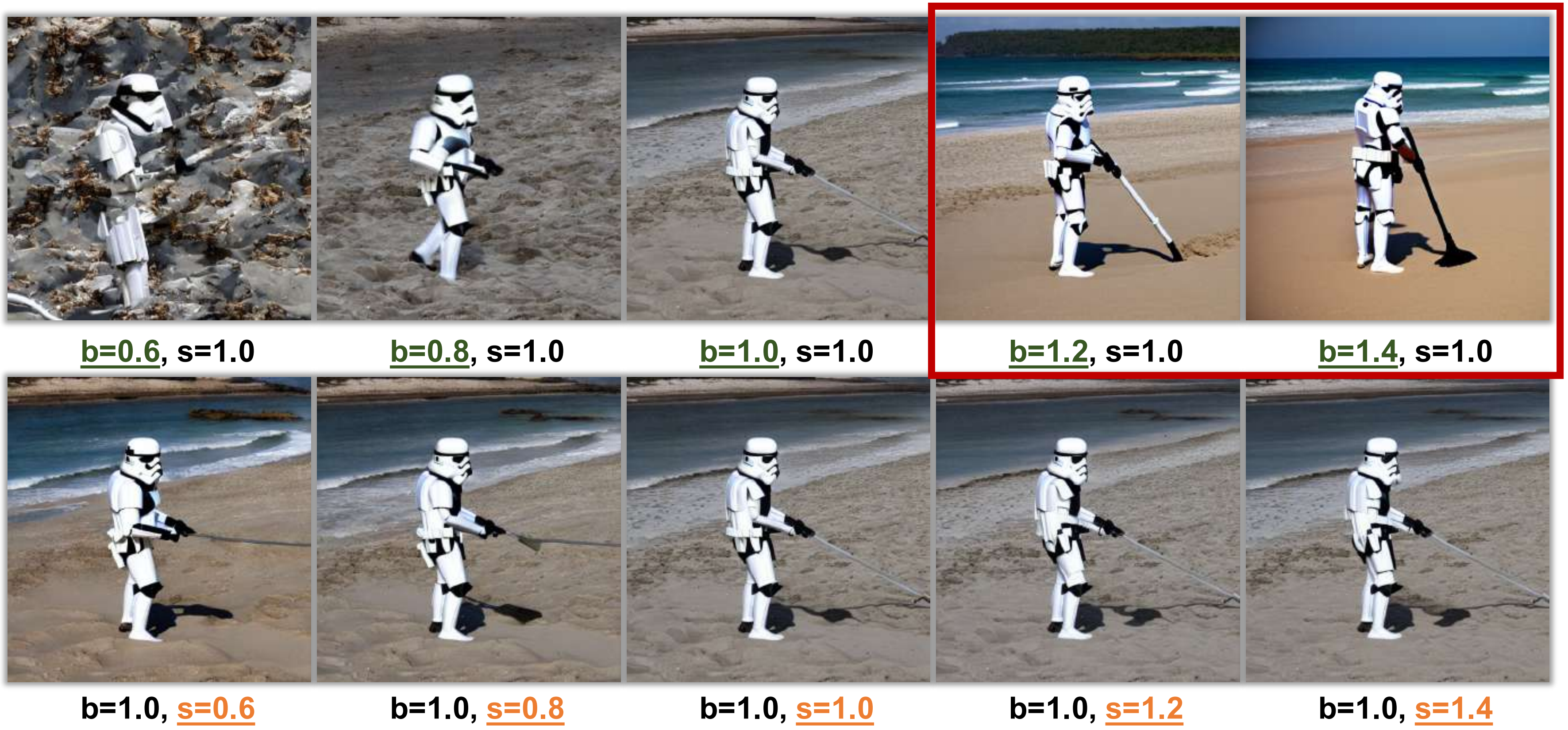
(1) enhance backbone features

Scale backbone features up
by a factor of b (*e.g.*, b=1.4)



Ablation: Backbone Scaling Factor

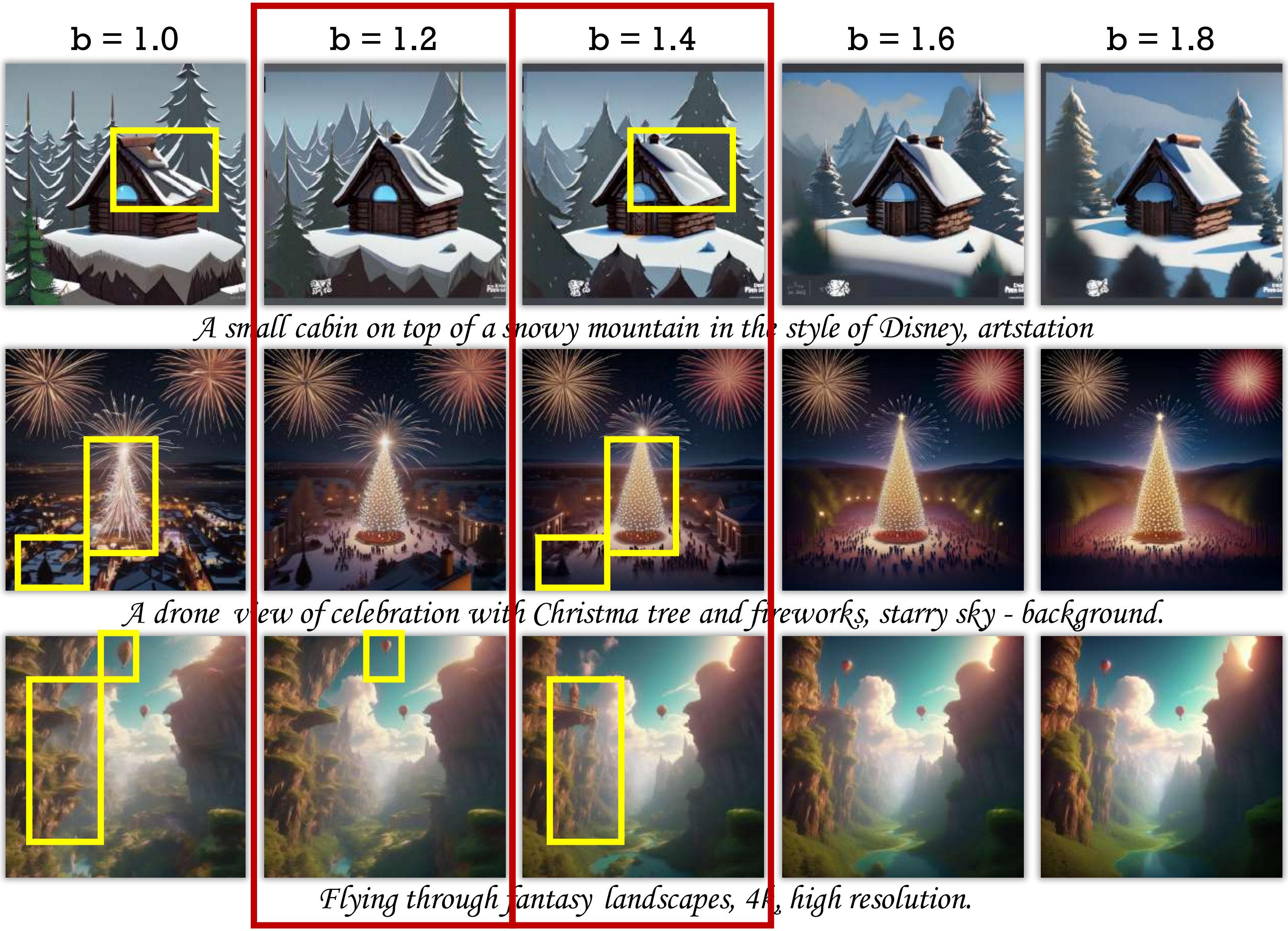
- Enhancing backbone features can improve image quality



Ablation: Backbone Scaling Factor

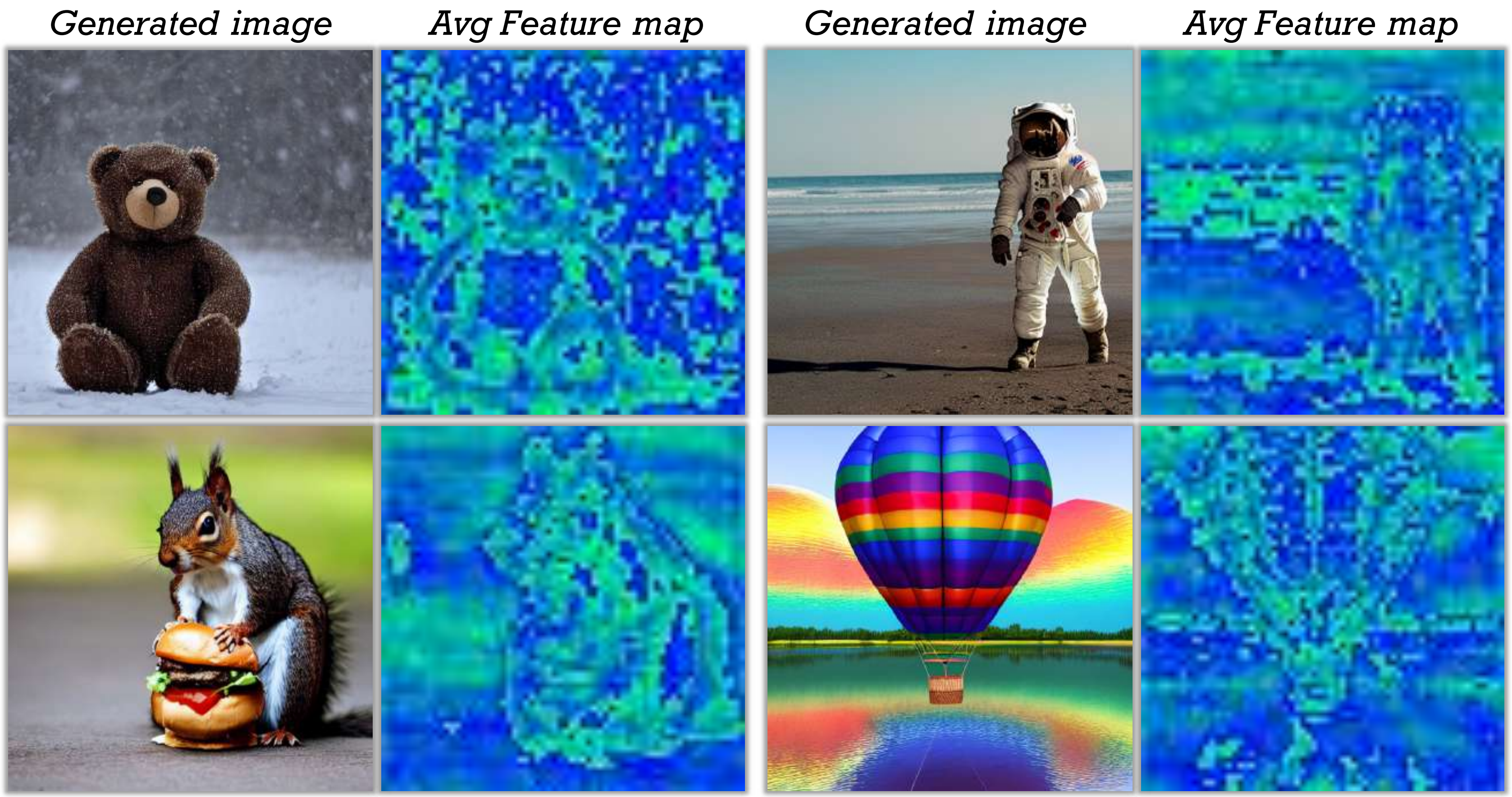
NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

S-LAB FOR ADVANCED INTELLIGENCE



Average **Backbone** Feature Maps

- Now: same backbone scaling everywhere.
- Is there a better way?



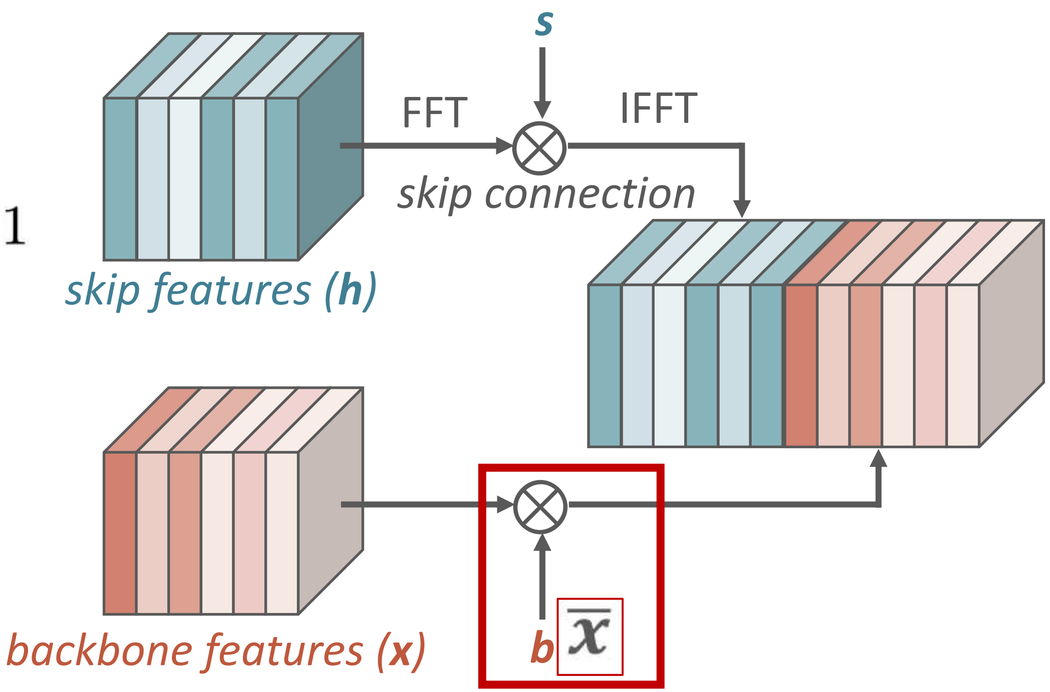
FreeU Method

(1) enhance backbone features

(2) content-aware backbone enhancement

$$\bar{\boldsymbol{x}}_l = \frac{1}{C}\sum_{i=1}^{C}\boldsymbol{x}_{l,i} \qquad \boldsymbol{\alpha}_l = (b_l - 1)\cdot\frac{\bar{\boldsymbol{x}}_l - Min(\bar{\boldsymbol{x}}_l)}{Max(\bar{\boldsymbol{x}}_l) - Min(\bar{\boldsymbol{x}}_l)} + 1$$

- spatially adaptive
- instance specific

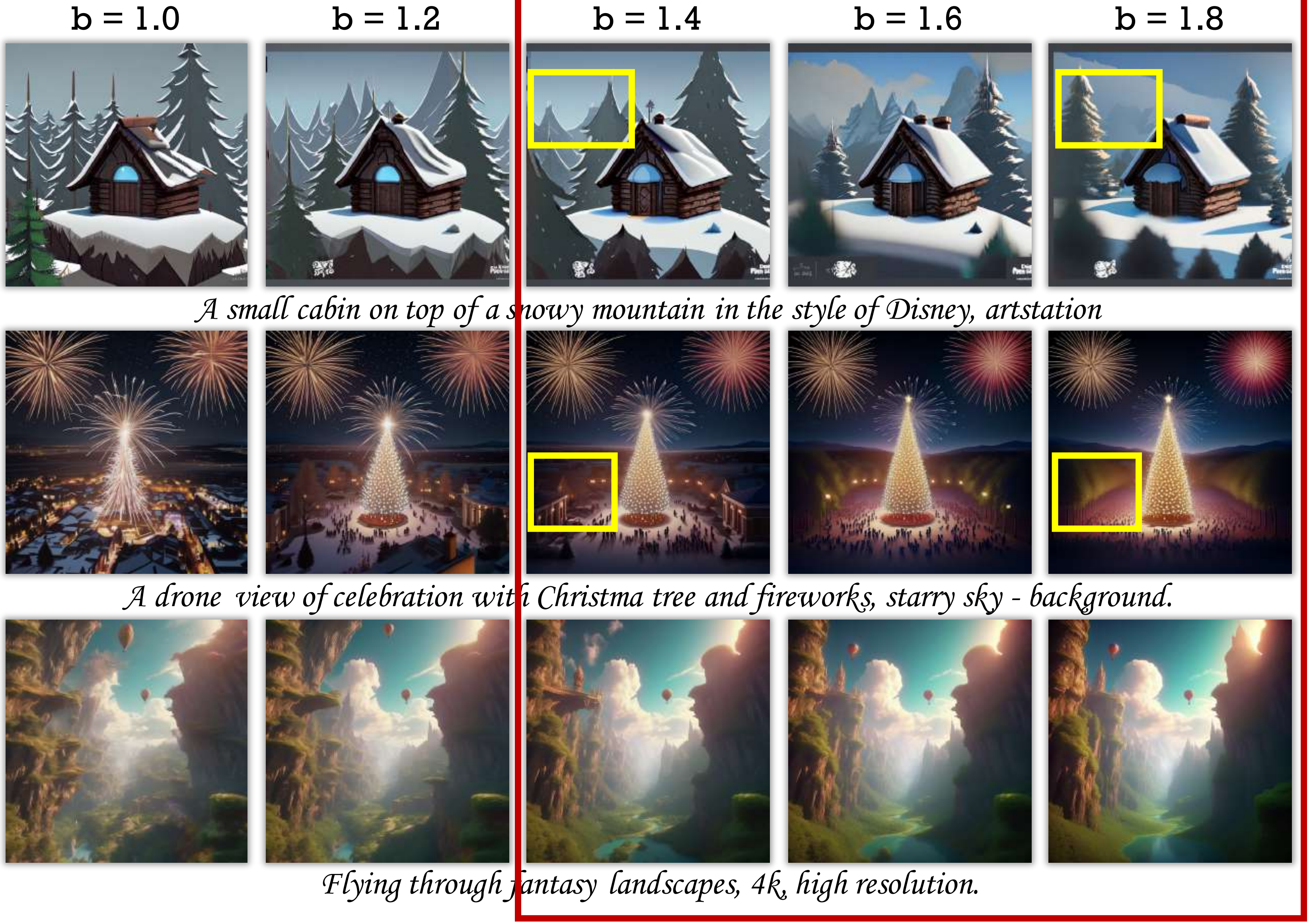


Content-Aware Backbone Scaling

Without FreeU | Constant Backbone Scaling | Content-Aware Backbone Scaling

(a) (b) (c)

Ablation: Backbone Scaling Factor



with increased backbone scaling, image can be oversmoothed

FreeU Method

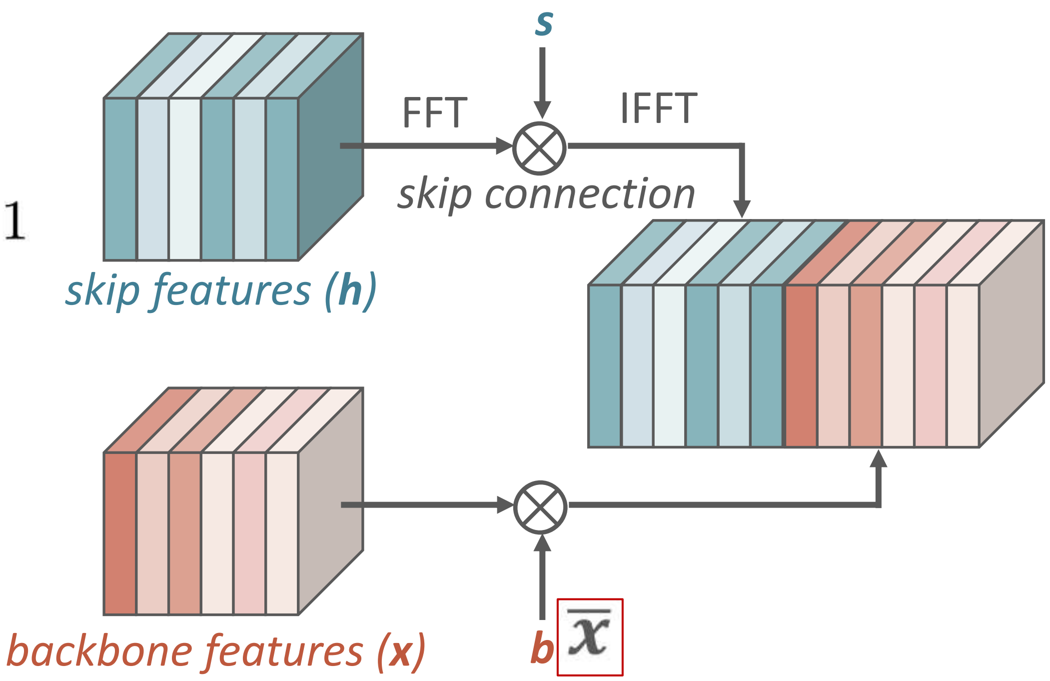
(1) enhance backbone features

(2) content-aware backbone enhancement

$$\bar{\boldsymbol{x}}_l = \frac{1}{C}\sum_{i=1}^{C} \boldsymbol{x}_{l,i} \qquad \boldsymbol{\alpha}_l = (b_l - 1) \cdot \frac{\bar{\boldsymbol{x}}_l - Min(\bar{\boldsymbol{x}}_l)}{Max(\bar{\boldsymbol{x}}_l) - Min(\bar{\boldsymbol{x}}_l)} + 1$$

(3) channel-selective backbone enhancement

$$\boldsymbol{x}^{'}_{l,i} = \begin{cases} \boldsymbol{x}_{l,i} \odot \boldsymbol{\alpha}_l, & \text{if } i < C/2 \\ \boldsymbol{x}_{l,i}, & \text{otherwise} \end{cases}$$



Channel Selection of **Backbone** Scaling

No Scaling | Scale All | Select First Half | Select Second Half | Uniform Selection

A drone view of celebration with Christma tree and fireworks, starry sky - background.

Flying through fantasy landscapes, 4k, high resolution.

A fat rabbit wearing a purple robe walking through a fantasy landscape.

FreeU Method

(1) enhance backbone features

(2) content-aware backbone enhancement

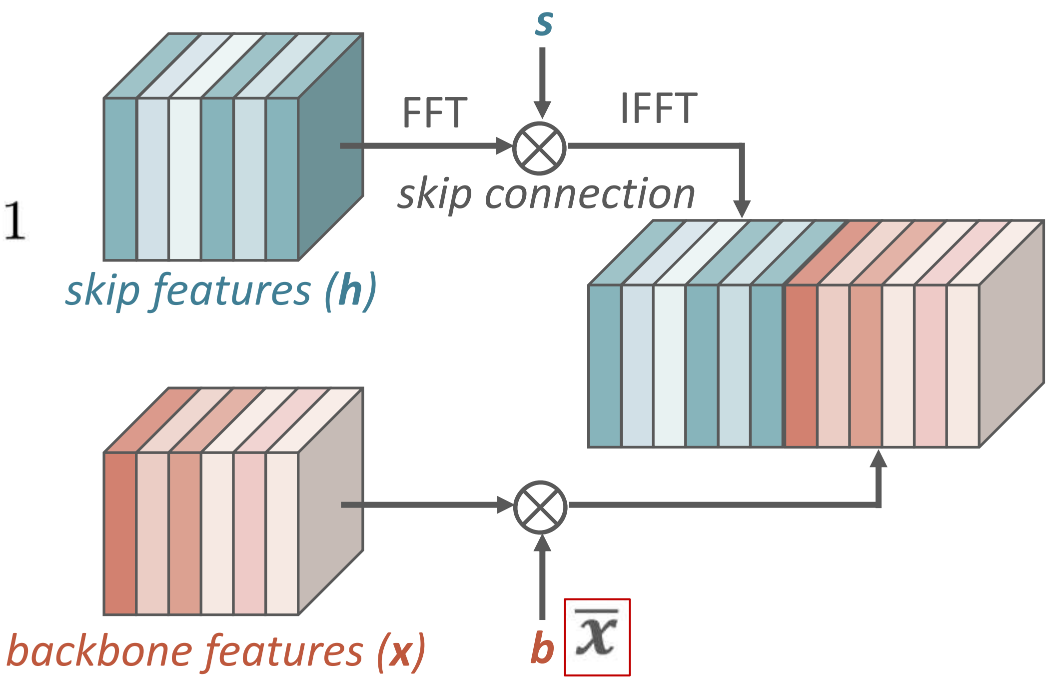
$$\bar{\boldsymbol{x}}_l = \frac{1}{C}\sum_{i=1}^{C} \boldsymbol{x}_{l,i} \qquad \boldsymbol{\alpha}_l = (b_l - 1) \cdot \frac{\bar{\boldsymbol{x}}_l - Min(\bar{\boldsymbol{x}}_l)}{Max(\bar{\boldsymbol{x}}_l) - Min(\bar{\boldsymbol{x}}_l)} + 1$$

(3) channel-selective backbone enhancement

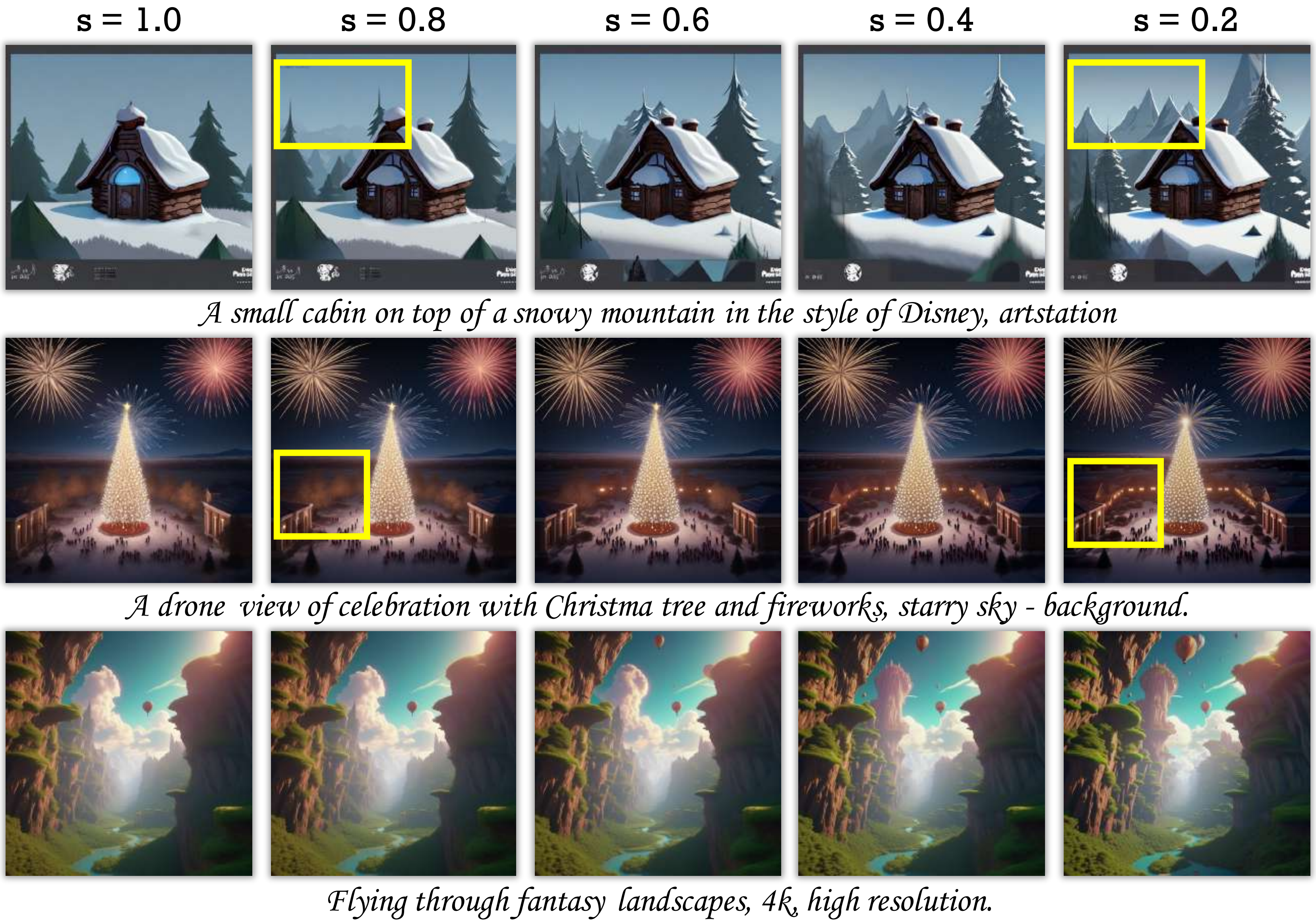
$$\boldsymbol{x}'_{l,i} = \begin{cases} \boldsymbol{x}_{l,i} \odot \boldsymbol{\alpha}_l, & \text{if } i < C/2 \\ \boldsymbol{x}_{l,i}, & \text{otherwise} \end{cases}$$

(4) suppress low-frequency in skip features

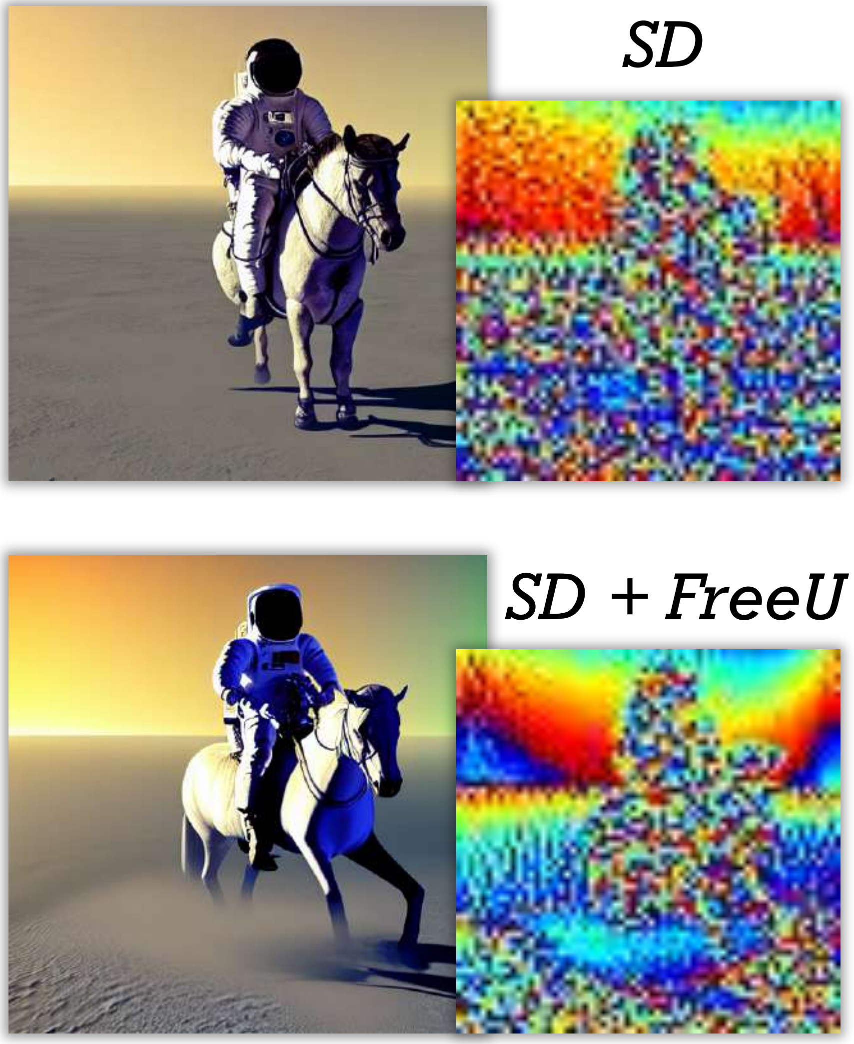
$$\boldsymbol{\beta}_{l,i}(r) = \begin{cases} s_l & \text{if } r < r_{\text{thresh}}, \\ 1 & \text{otherwise.} \end{cases} \qquad \begin{aligned} \mathcal{F}(\boldsymbol{h}_{l,i}) &= \text{FFT}(\boldsymbol{h}_{l,i}) \\ \mathcal{F}'(\boldsymbol{h}_{l,i}) &= \mathcal{F}(\boldsymbol{h}_{l,i}) \odot \boldsymbol{\beta}_{l,i} \\ \boldsymbol{h}'_{l,i} &= \text{IFFT}(\mathcal{F}'(\boldsymbol{h}_{l,i})) \end{aligned}$$

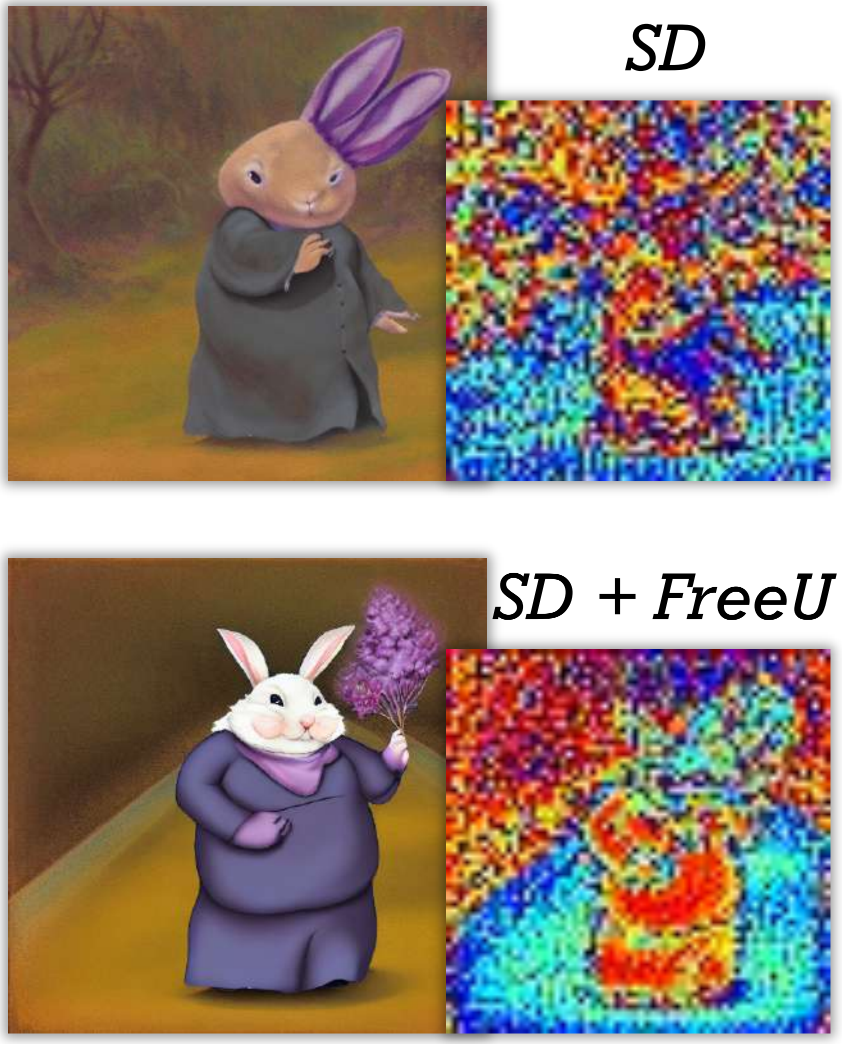


Ablation: **Skip** Scaling Factor

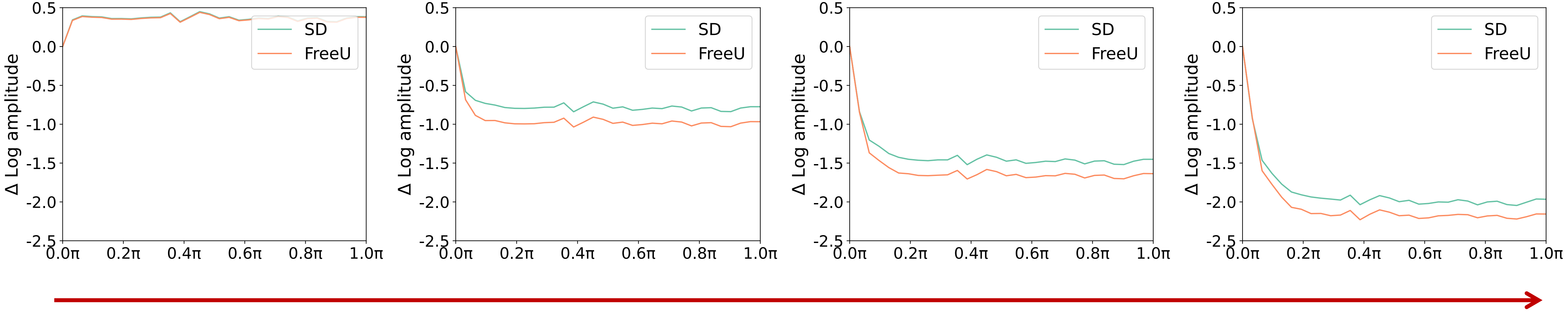


Feature Maps Visualization





FreeU's Impact to Frequency Domain



reverse process / denoising process

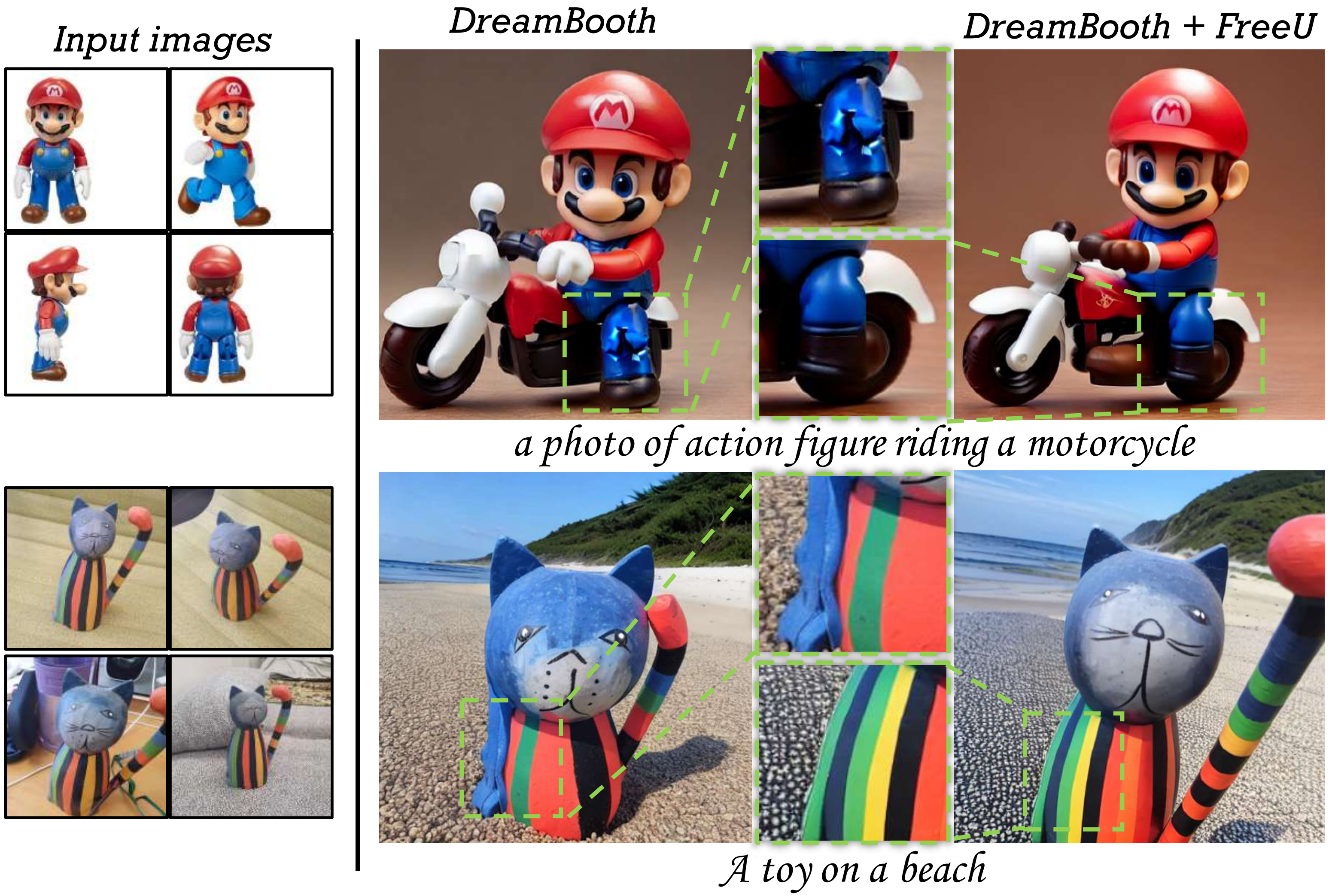
Gradually denoise to image

Visual Results: Text-to-Image

Visual Results: Text-to-Video

Visual Results: Text-to-Video

Visual Results: Personalized Text-to-Image



Ruiz et al. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

Visual Results: Personalized Text-to-Image

ReVersion | *ReVersion+FreeU* | *ReVersion* | *ReVersion+FreeU*

child <R> child
<R> = "sits back-to-back with"

dog <R> basket
<R> = "is contained inside of"

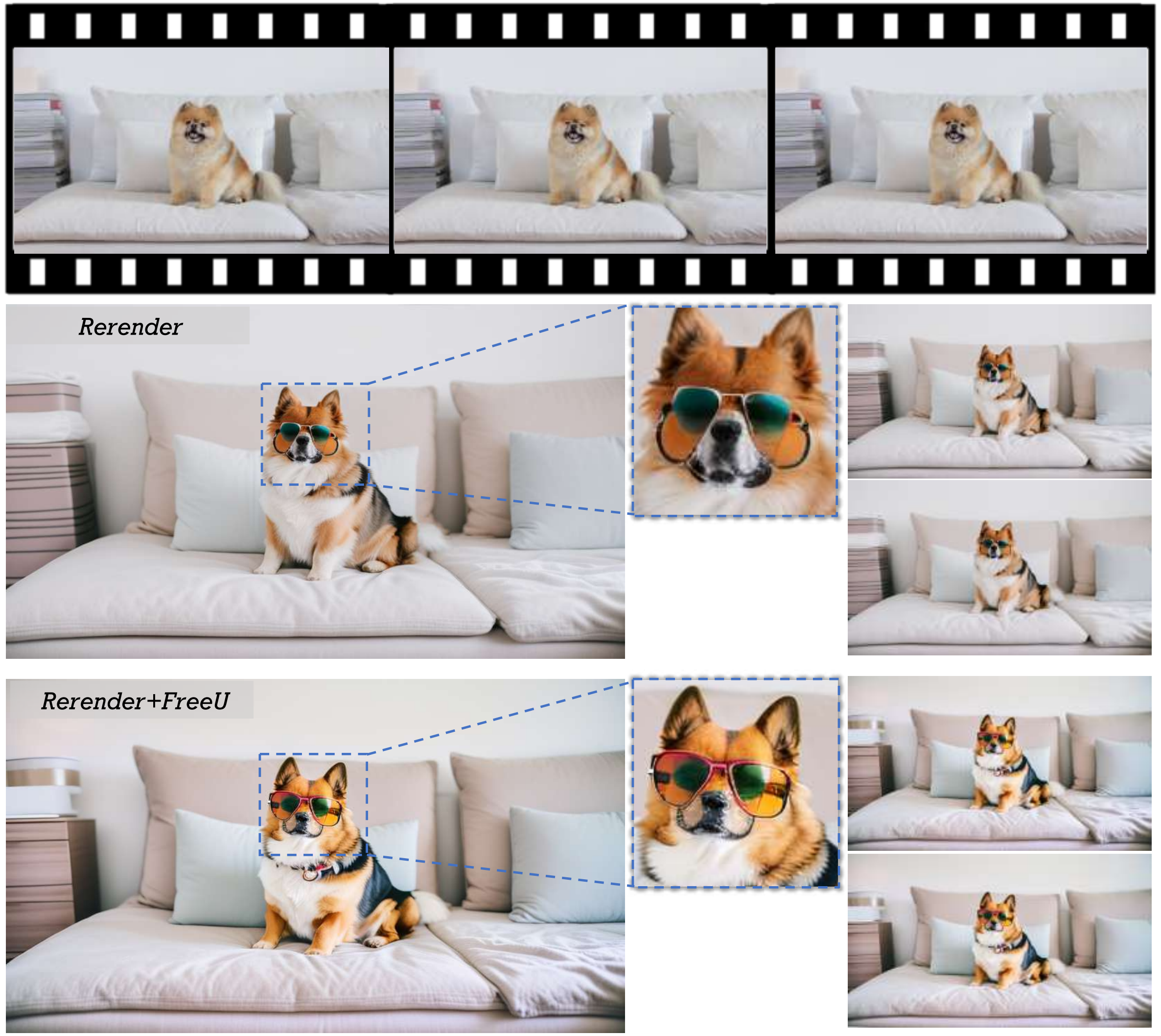
Spiderman <R> basket
<R> = "is contained inside of"

cat <R> motorbike
<R> = "ride on"

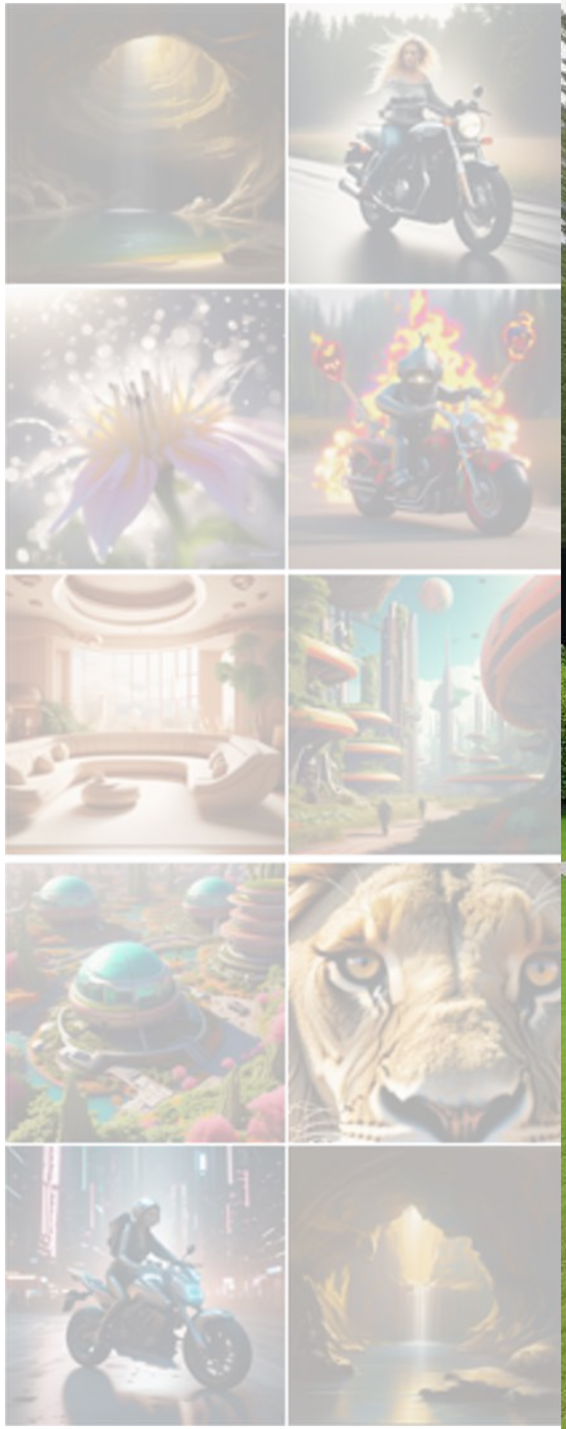
Huang et al. ReVersion : Diffusion-Based Relation Inversion from Images

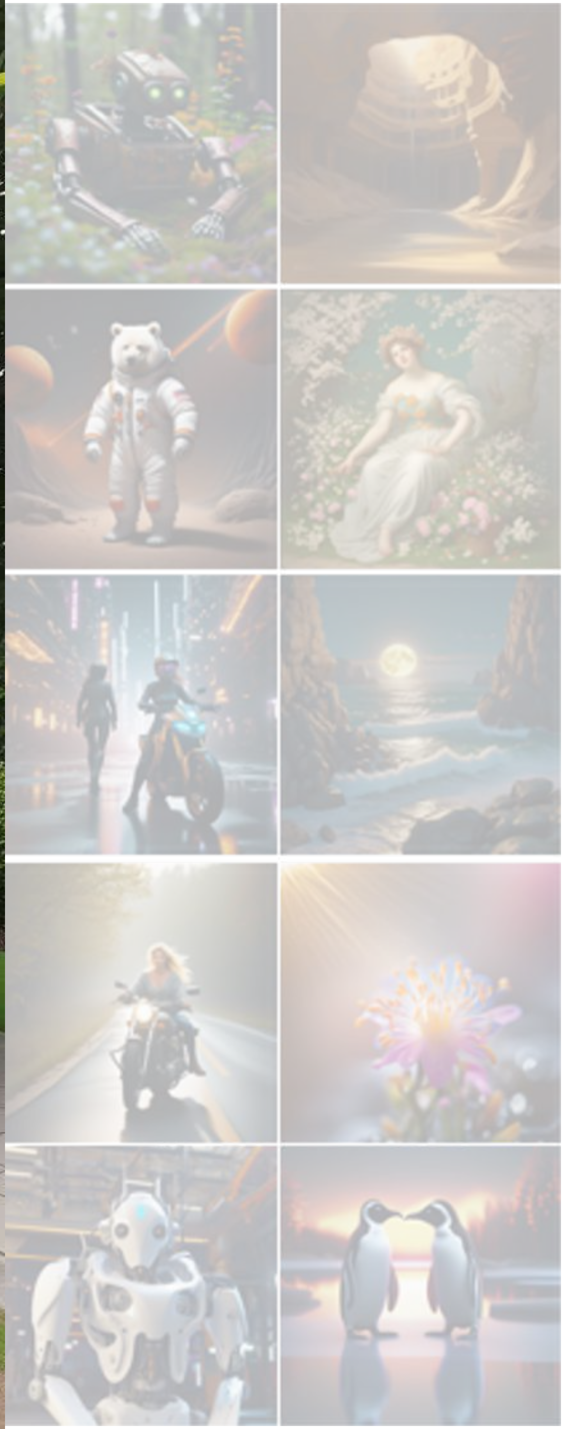
Visual Results: Video-to-Video

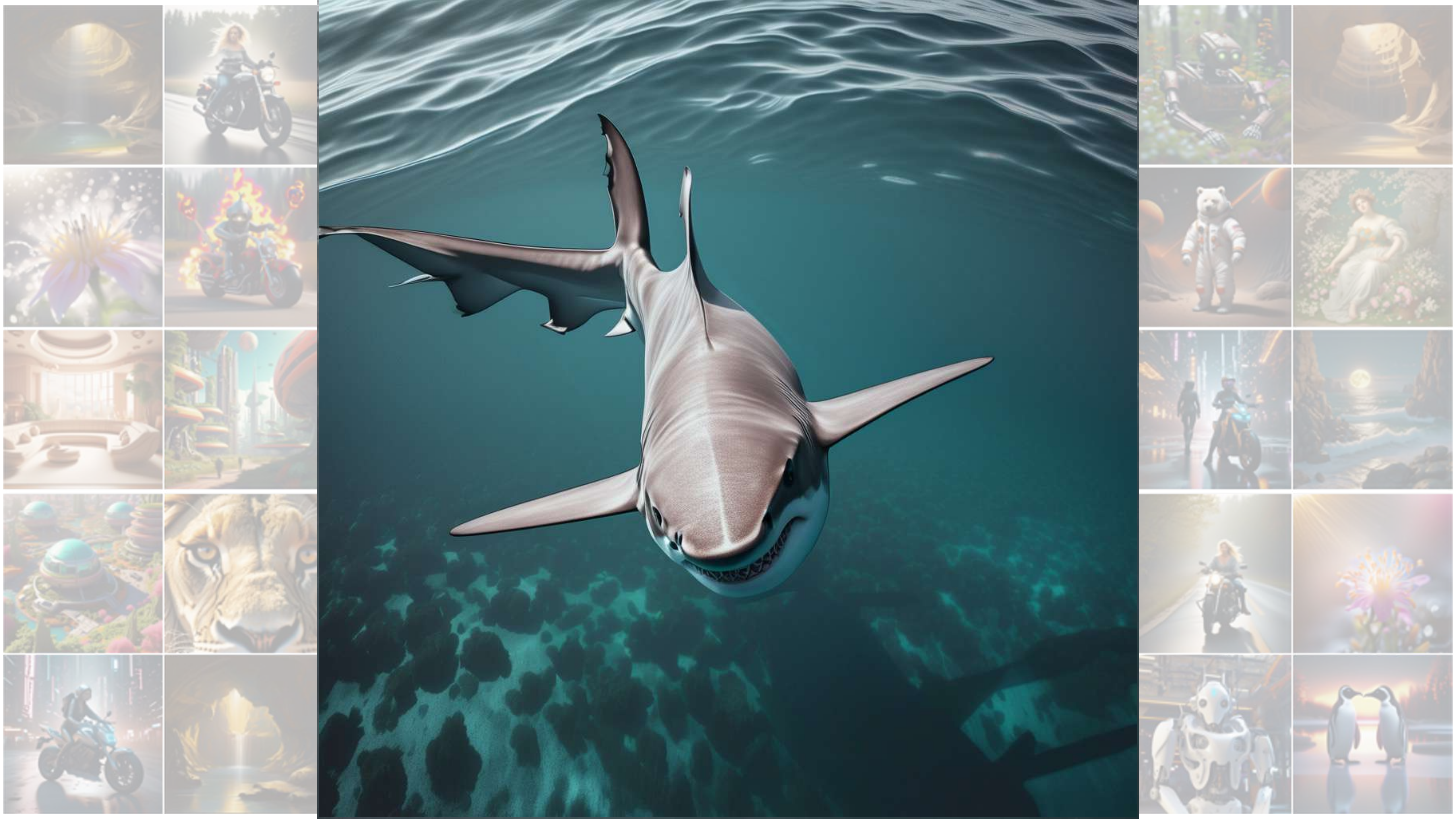
Visual Results: Video-to-Video

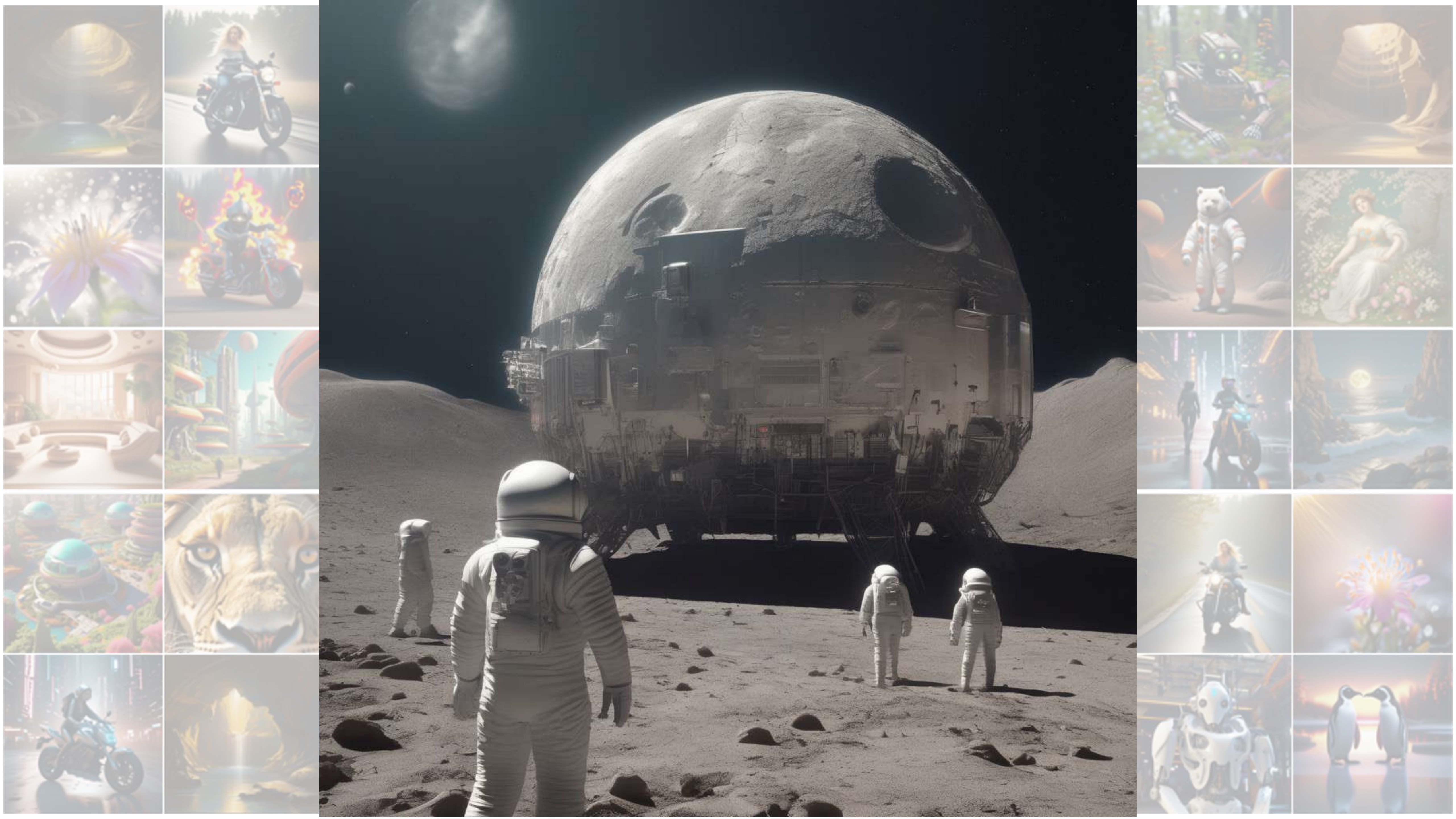


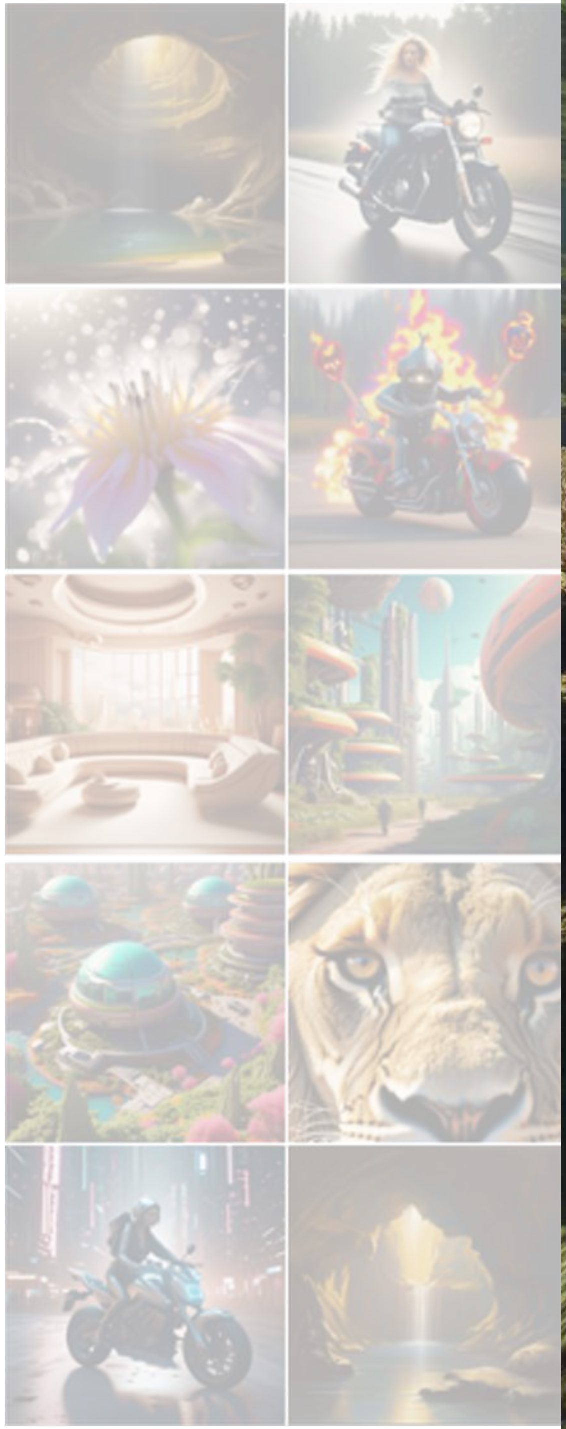
A dog wearing sunglasses

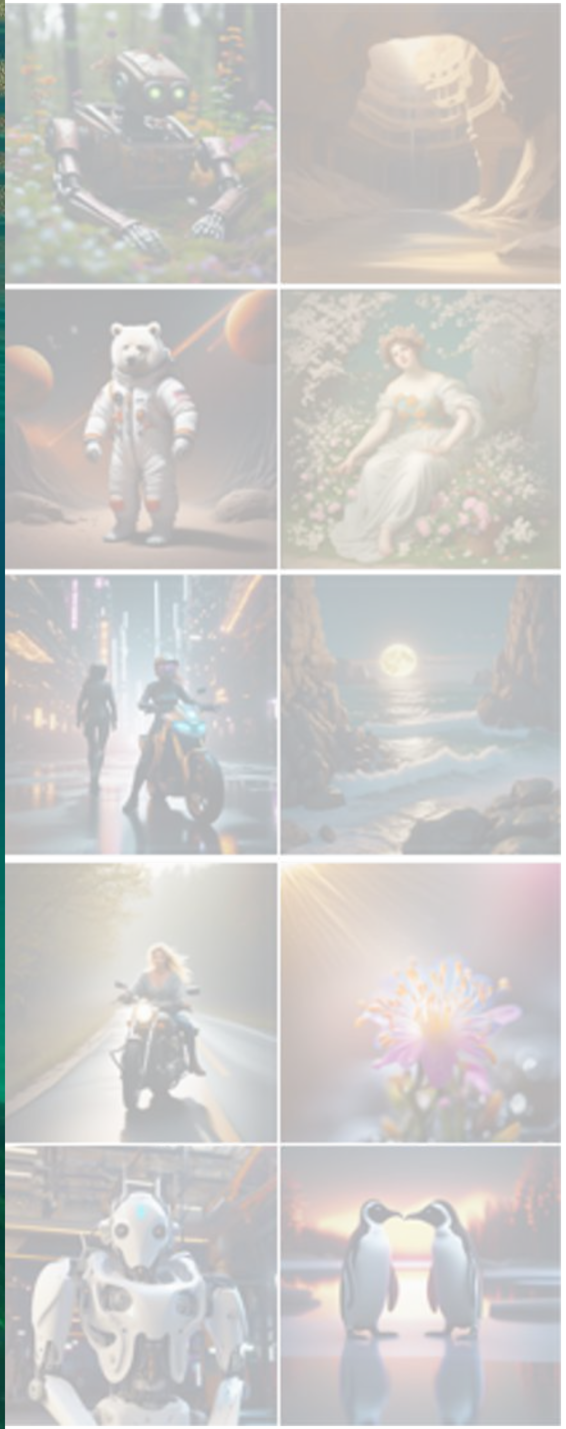


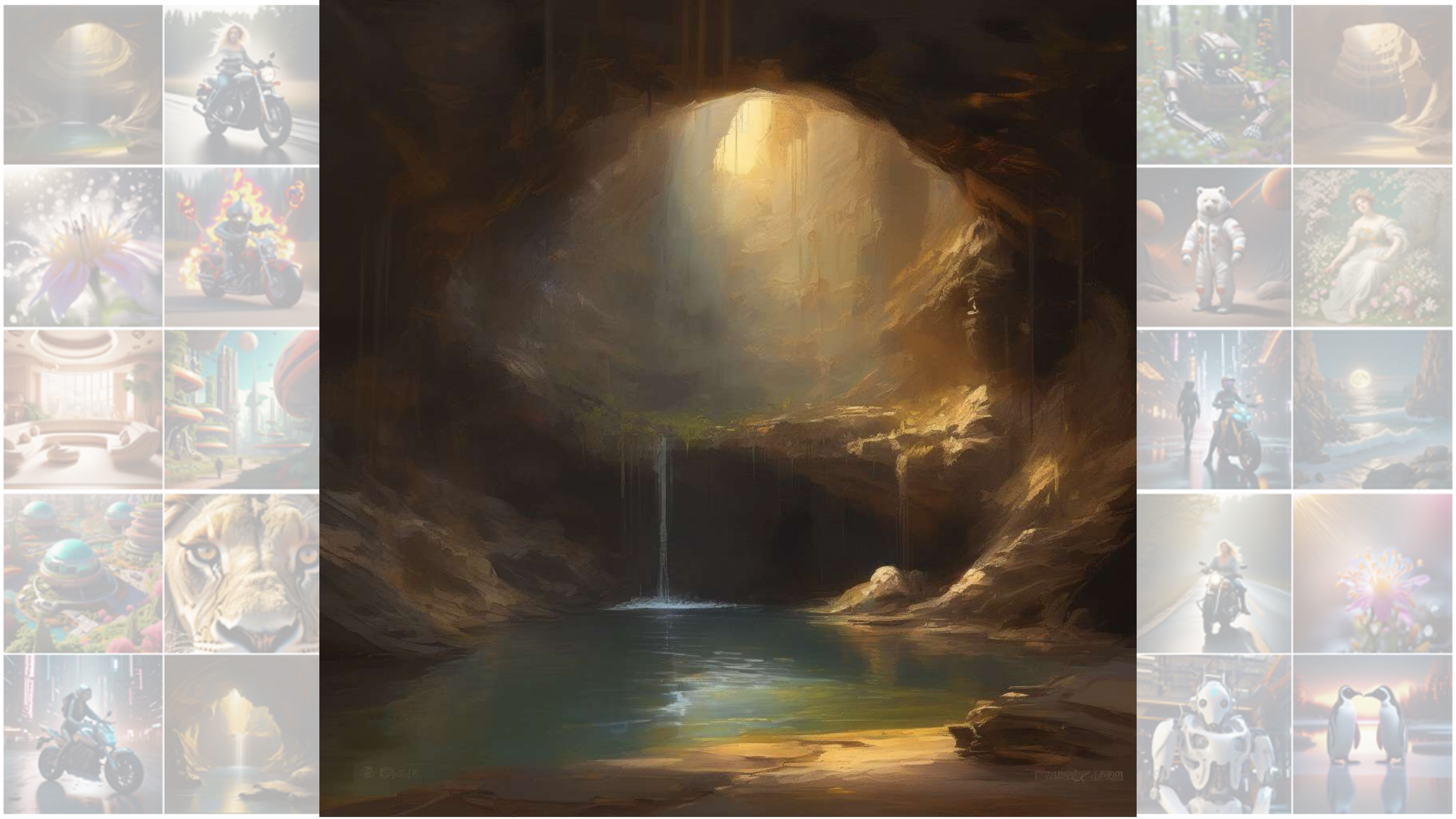




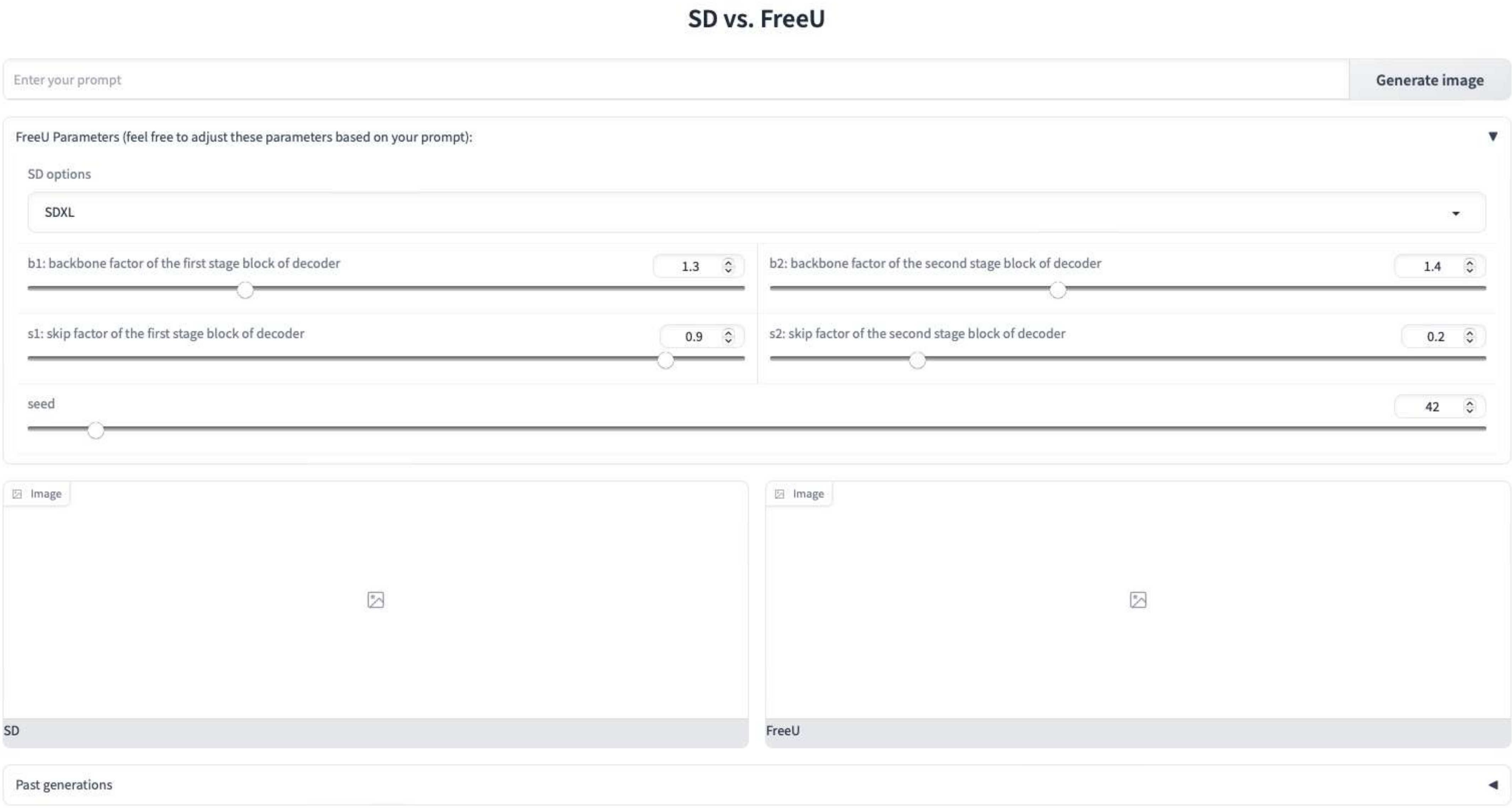




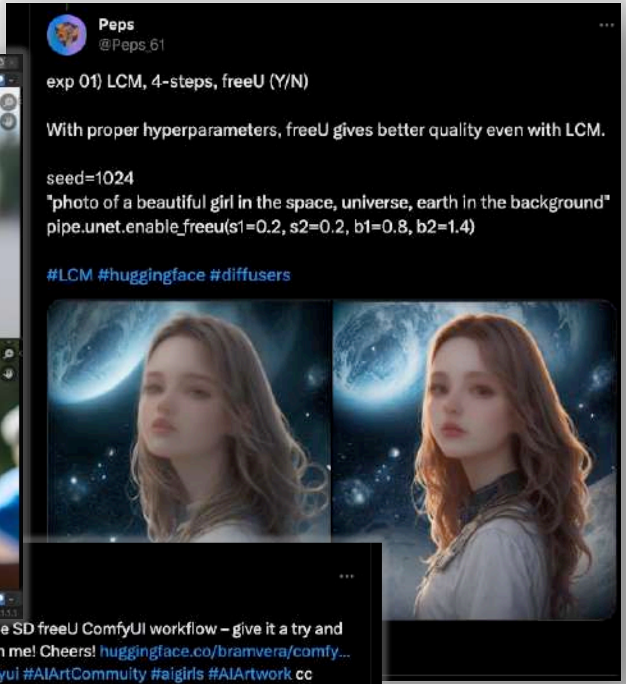


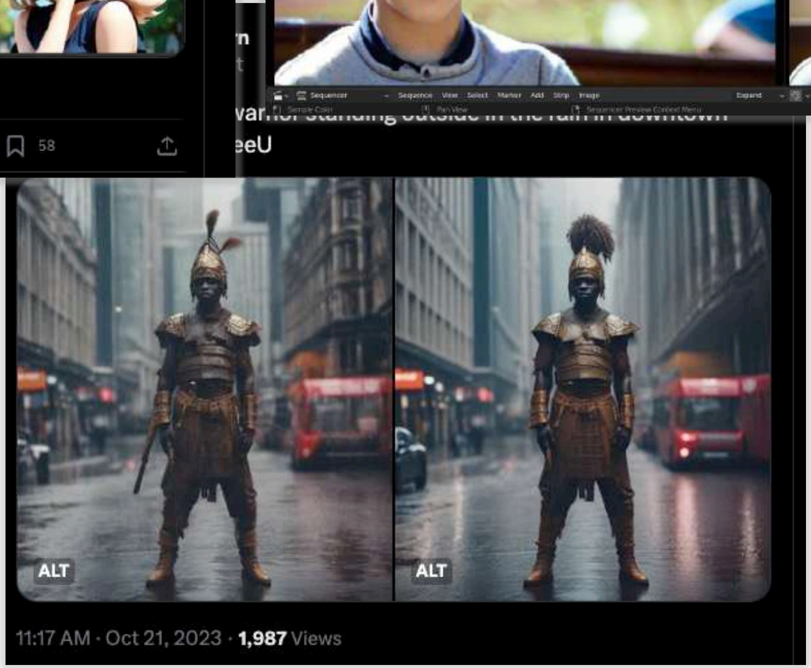


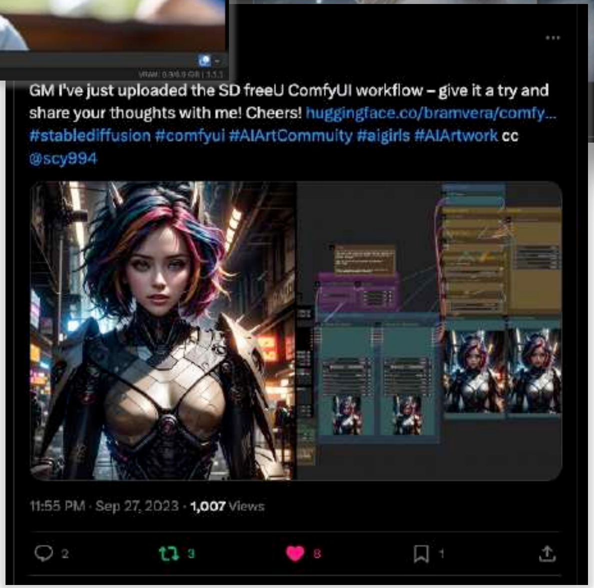
FreeU Demo


SD vs. FreeU
Enter your prompt
Generate image
FreeU Parameters (feel free to adjust these parameters based on your prompt):
SD options
SDXL
b1: backbone factor of the first stage block of decoder
1.3
b2: backbone factor of the second stage block of decoder
1.4
s1: skip factor of the first stage block of decoder
0.9
s2: skip factor of the second stage block of decoder
0.2
seed
42
Image
Image
SD
FreeU
Past generations

Community Contributions

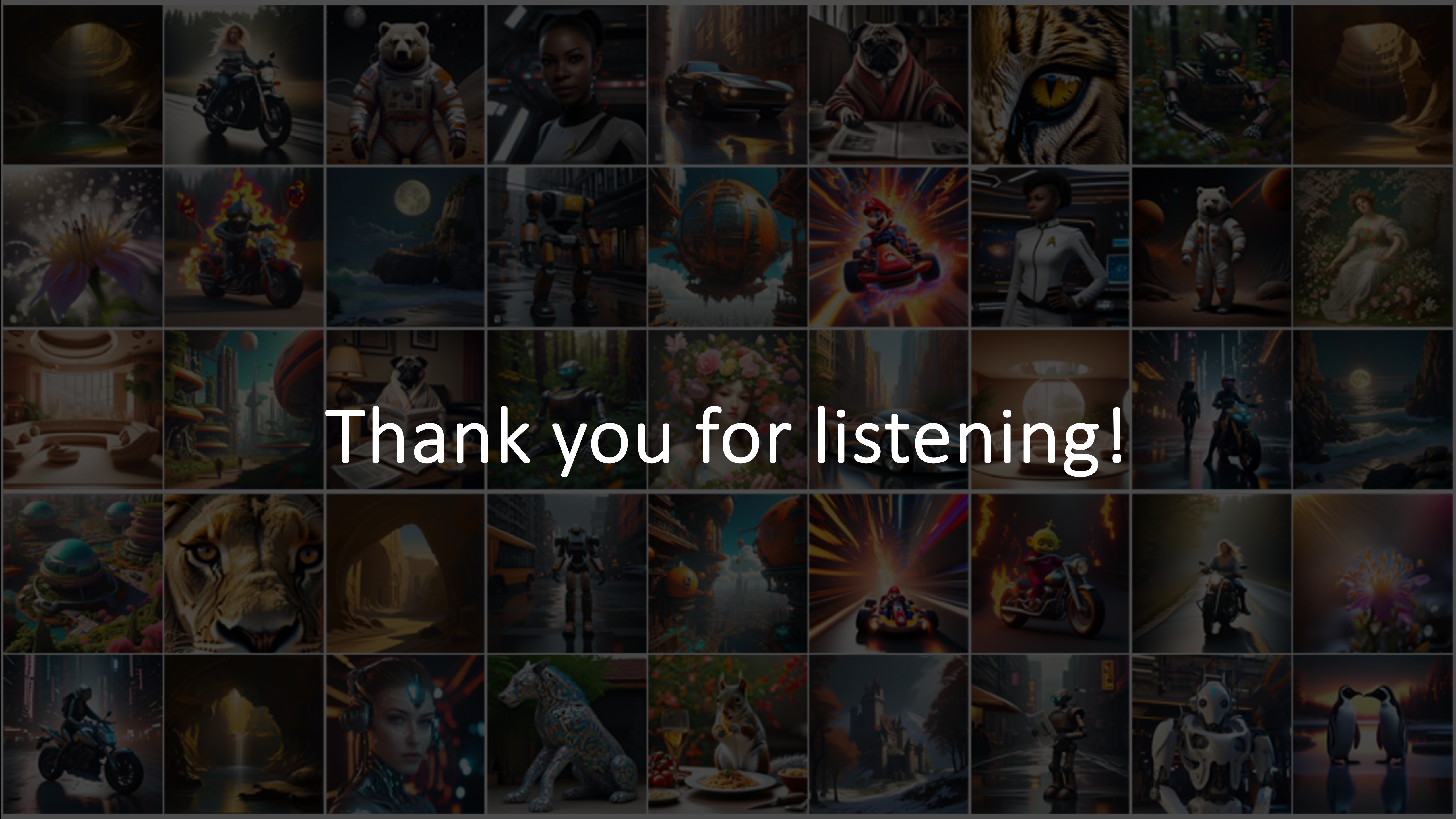






Future Works

- Different FreeU strategy across inference time
  - Backbone features: early stage
  - Skip features: later stage
- Further explanation on FreeU
  - Gap between training and inference
  - Insights for training strategies
- Automatic parameter search for FreeU
- FreeU for more modalities (*e.g.*, audio, video, 3D)


Thank you for listening!