CVPR
JUNE 17-21, 2024
SEATTLE, WA

S-LAB FOR ADVANCED INTELLIGENCE

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

FreeU: Free Lunch in Diffusion U-Net

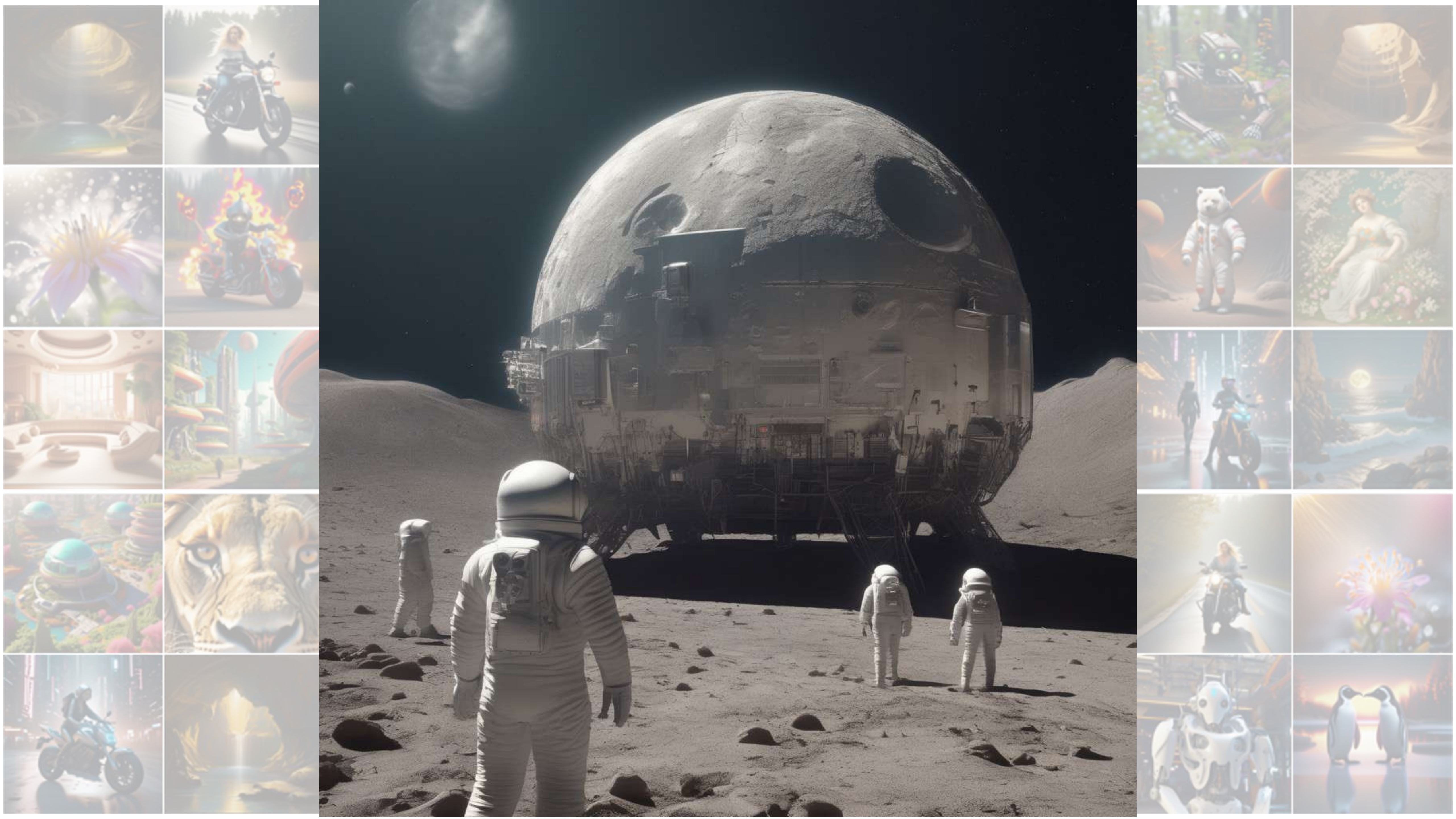

Chenyang Si


Ziqi Huang


Yuming Jiang


Ziwei Liu

MMLab@NTU | S-Lab, Nanyang Technological University

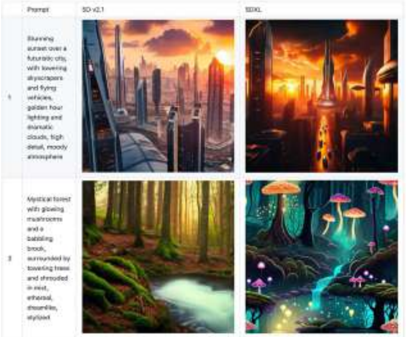


Pre-trained Diffusion Models

image generation

ADM

LDM

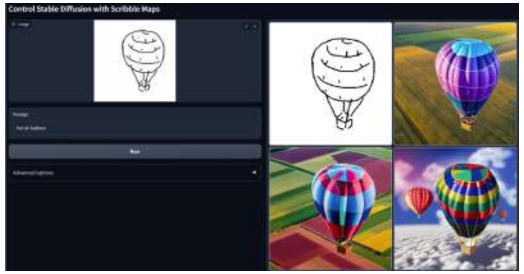


SDXL

video generation

VideoCrafter

controllable generation, customization, editing



ControlNet

Input images

in the Acropolis

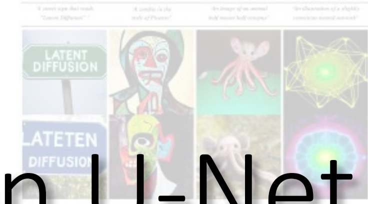
DreamBooth

InstructPix2Pix

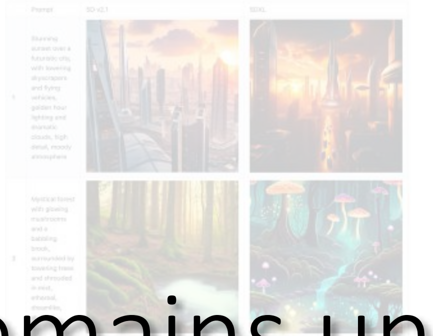
Pre-trained Diffusion U-Nets

image generation

ADM



LDM



SDXL

video generation

VideoCrafter

Diffusion U-Net remains under-explored

controllable generation, editing, customization

ControlNet

DreamBooth

InstructPix2Pix

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*

Recap: Diffusion Models

reverse process / denoising process

gradually denoise to image

Image

Noise

$\mathbf{x}_0$ $\mathbf{x}_{t-1}$ $\mathbf{x}_t$ $\mathbf{x}_T$

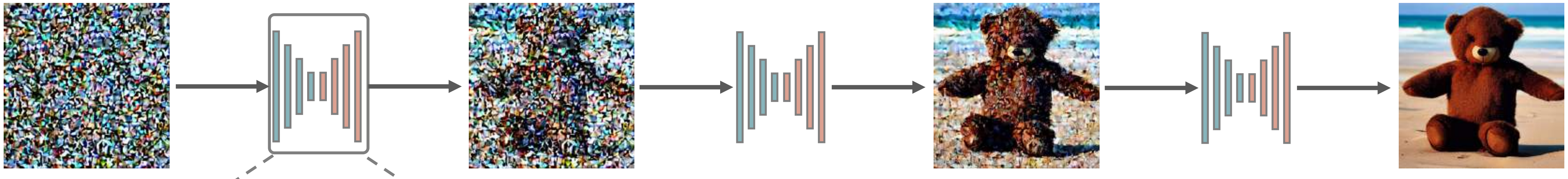
gradually adds Gaussian noise to the data

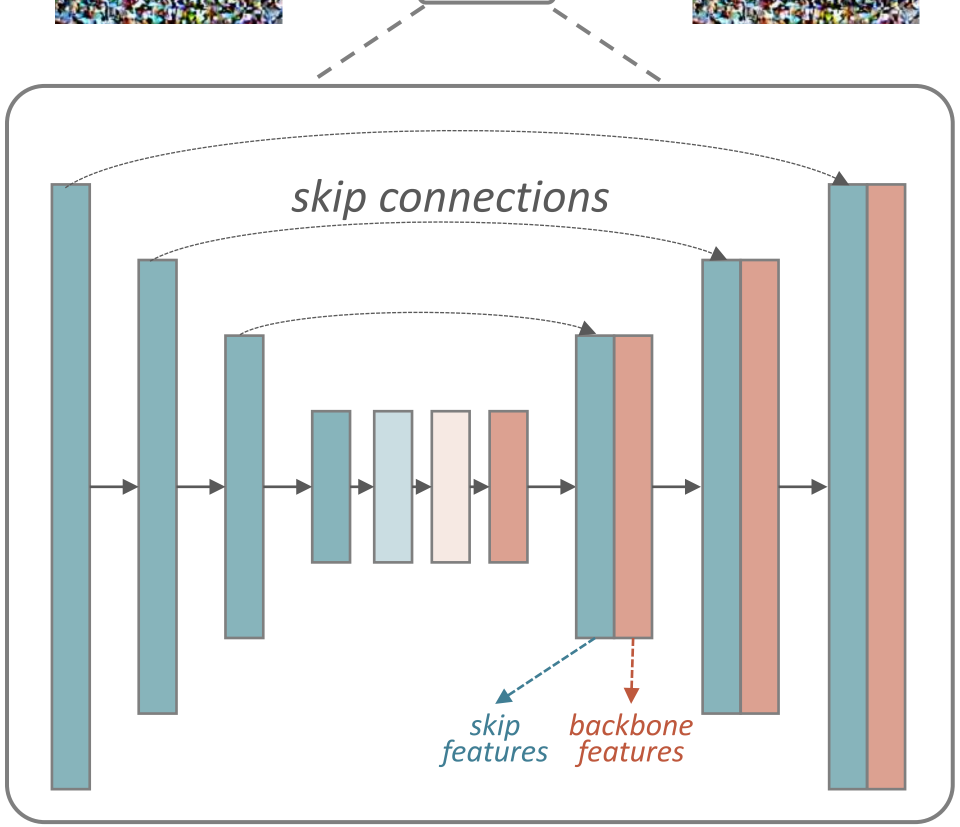
forward process / diffusion process

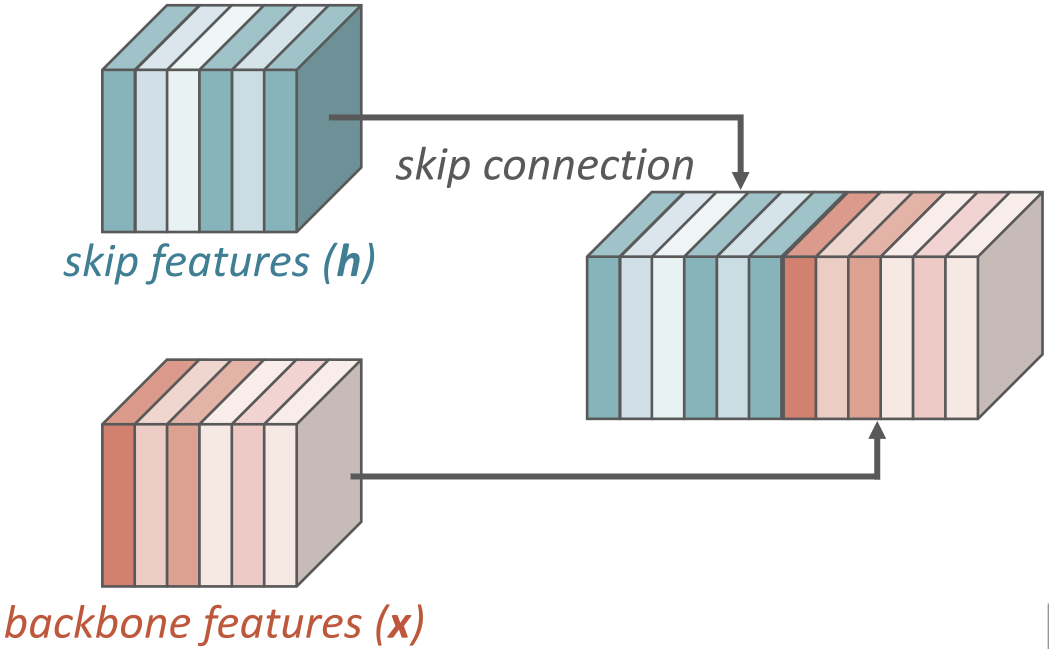
Image Credit: CVPR 2022 Tutorial: Denoising Diffusion-based Generative Modeling: Foundations and Applications

How does diffusion U-Net perform denoising?

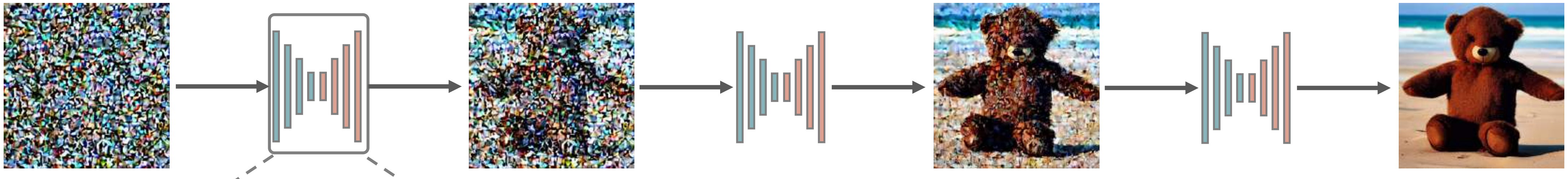
Denoising Process: U-Net

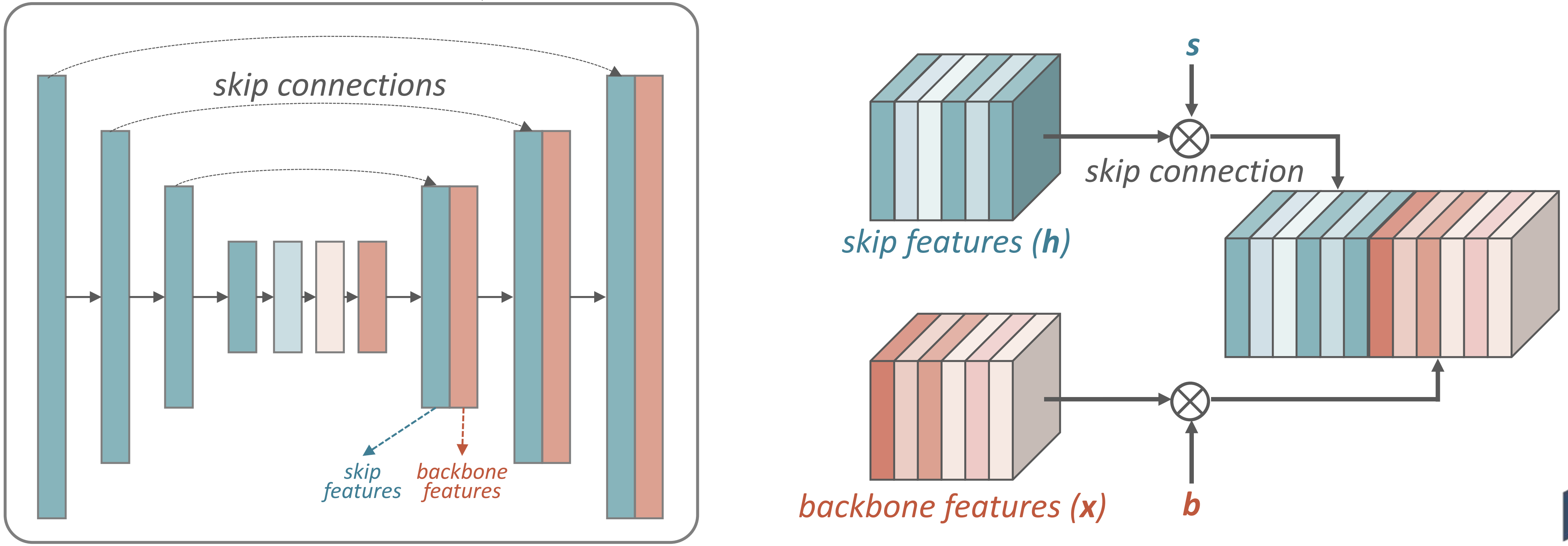






Denoising Process: U-Net





Role of Backbone and Skip Features

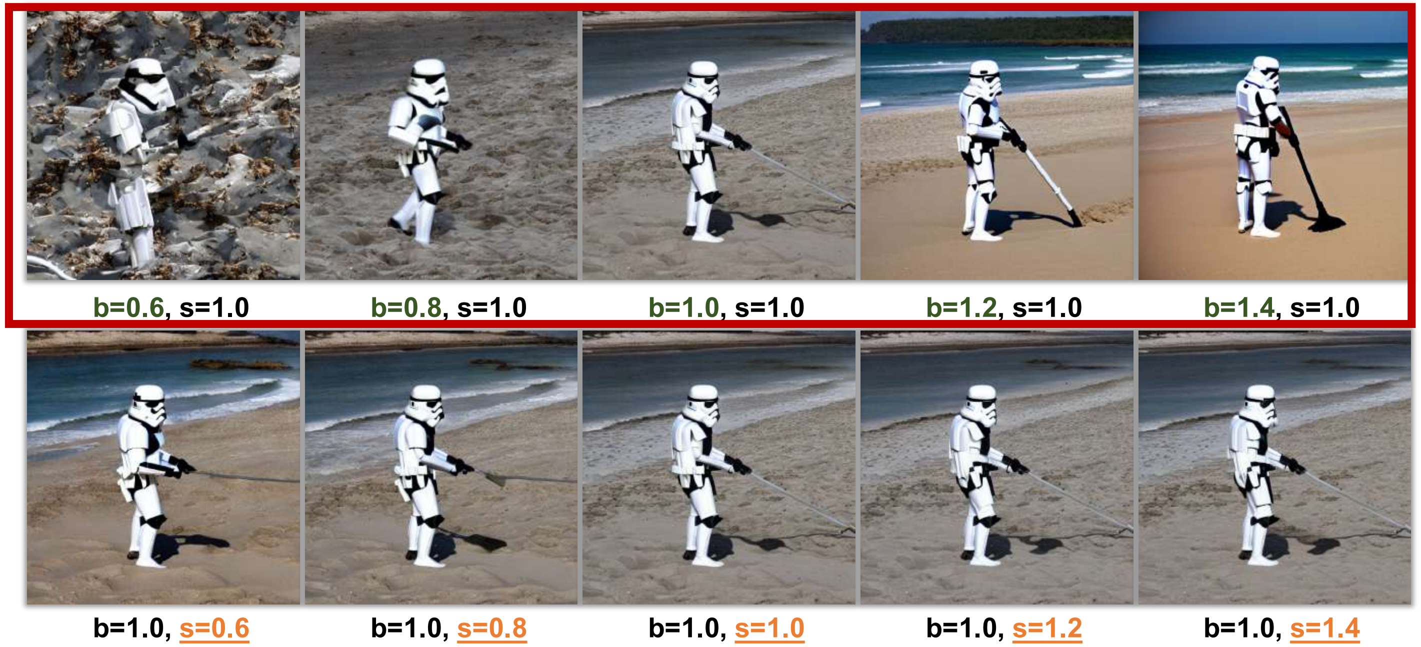
- **Backbone**: denoising
- **Skip**: limited impact during inference

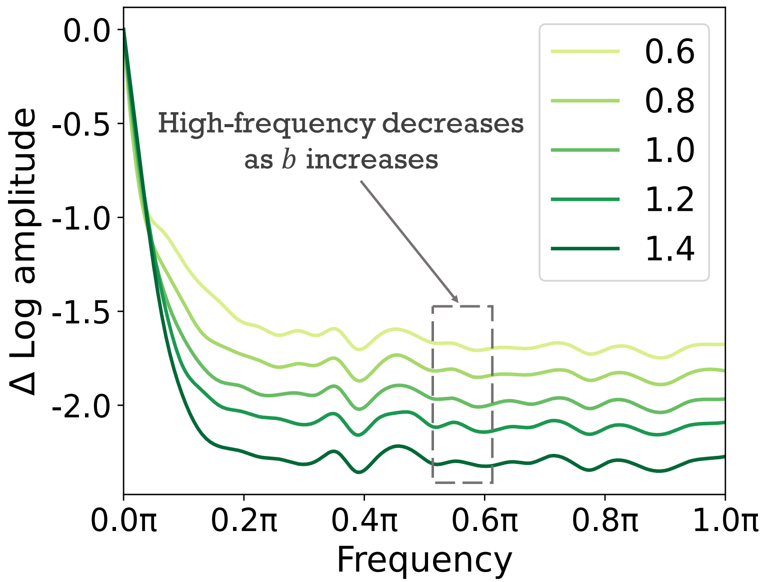
b=0.6, s=1.0 | b=0.8, s=1.0 | b=1.0, s=1.0 | b=1.2, s=1.0 | b=1.4, s=1.0

b=1.0, s=0.6 | b=1.0, s=0.8 | b=1.0, s=1.0 | b=1.0, s=1.2 | b=1.0, s=1.4

How Diffusion U-Net Perform Denoising?

- **Backbone features**: primarily contributes to denoising
  - Consistent with visualization on the next page

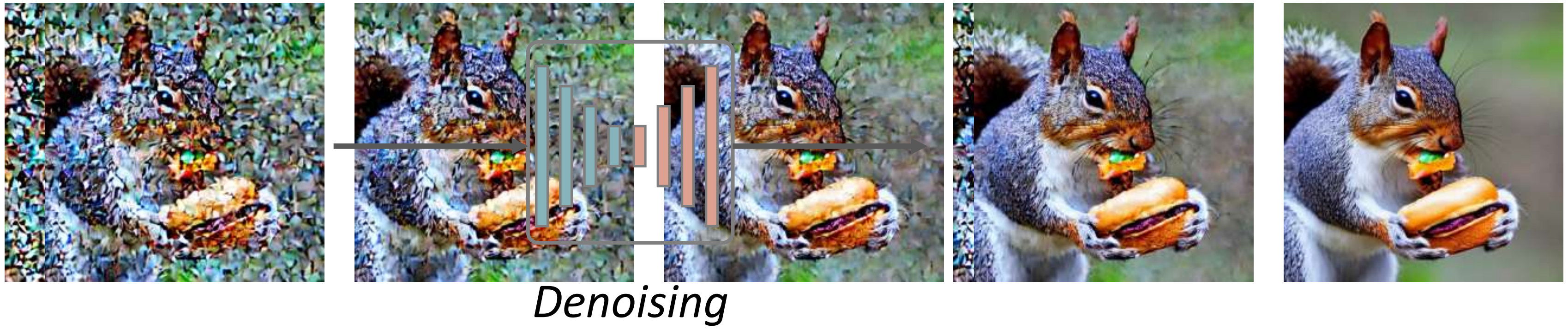




Fourier relative log amplitudes of variations of b

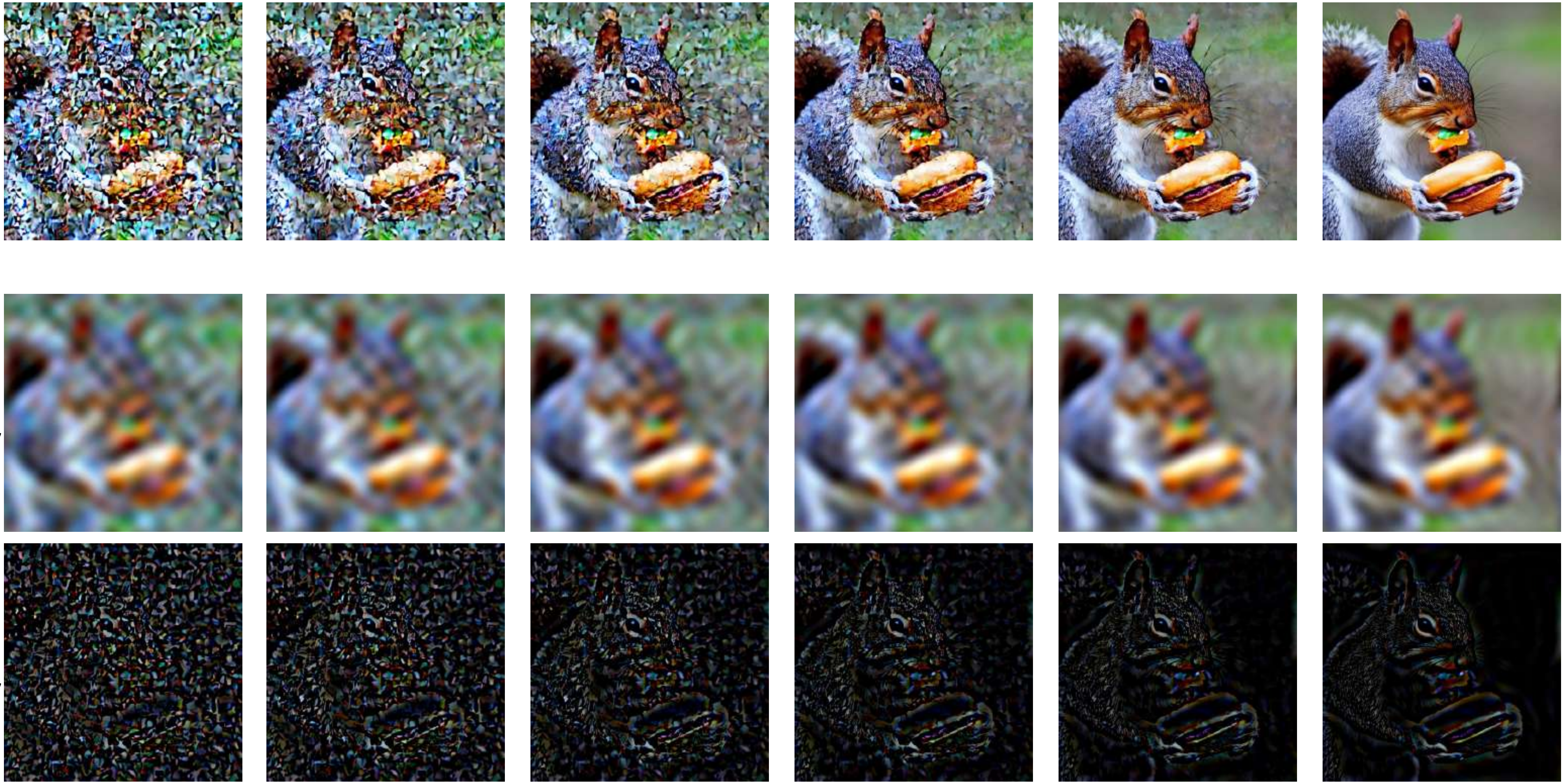
Denoising Process

Input: *A squirrel eating a burger*



Denoising Process

Input: *A squirrel eating a burger*

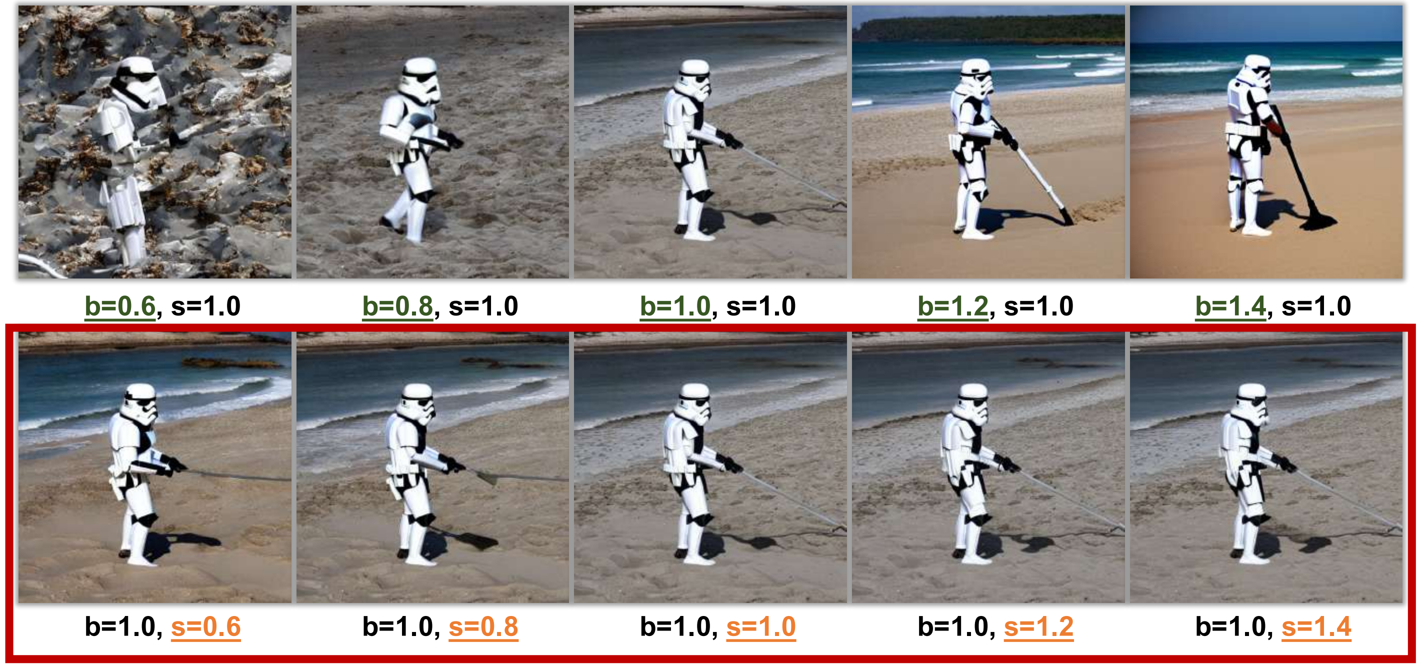


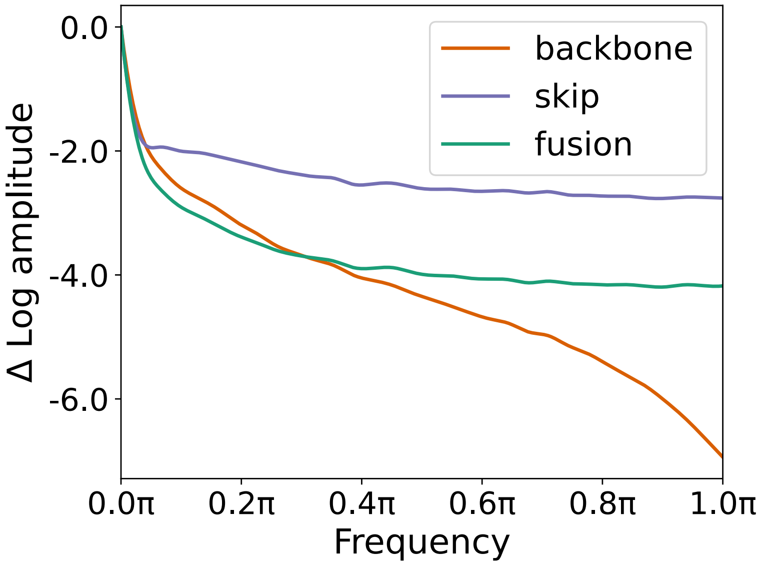
Low frequency

High frequency

How Diffusion U-Net Perform Denoising?

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

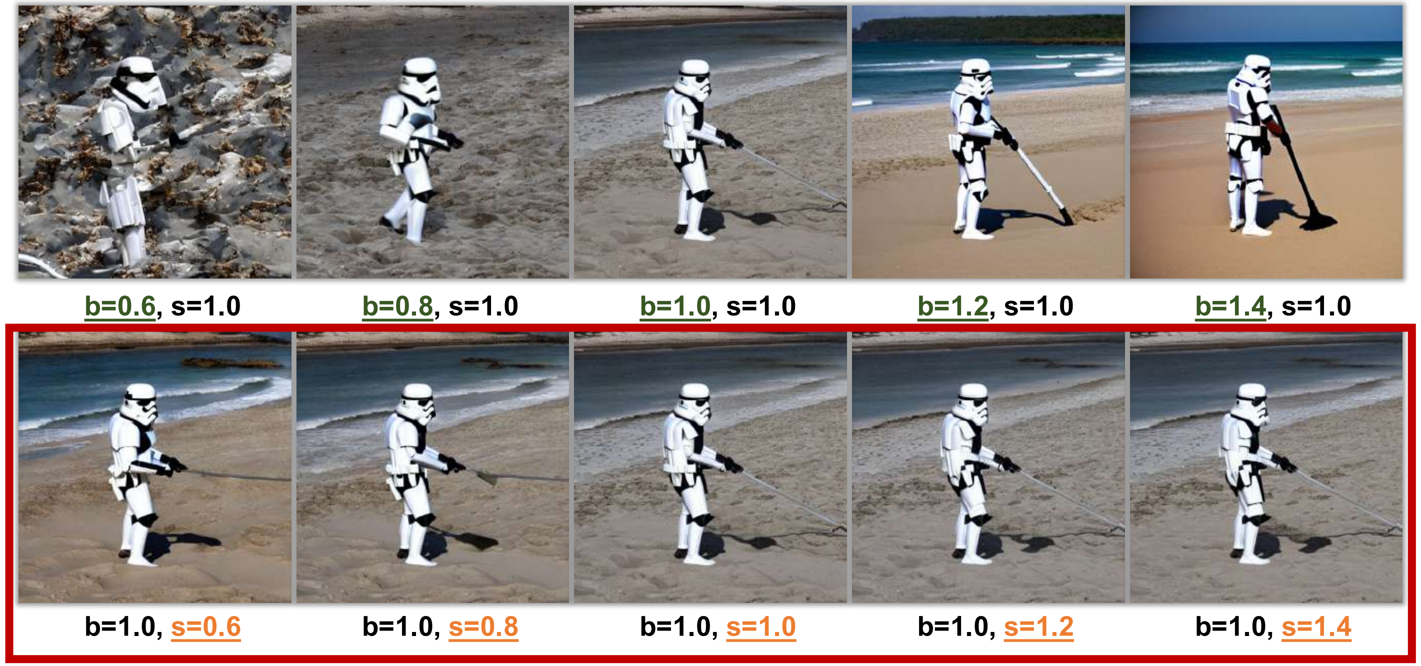


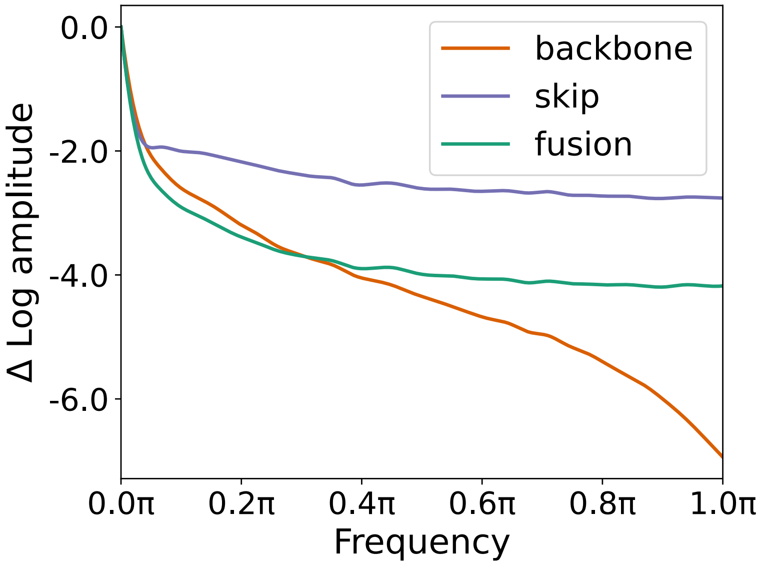


Fourier relative log amplitudes of backbone, skip, and their fused feature maps

How Diffusion U-Net Perform Denoising?

- Gap between training and sampling



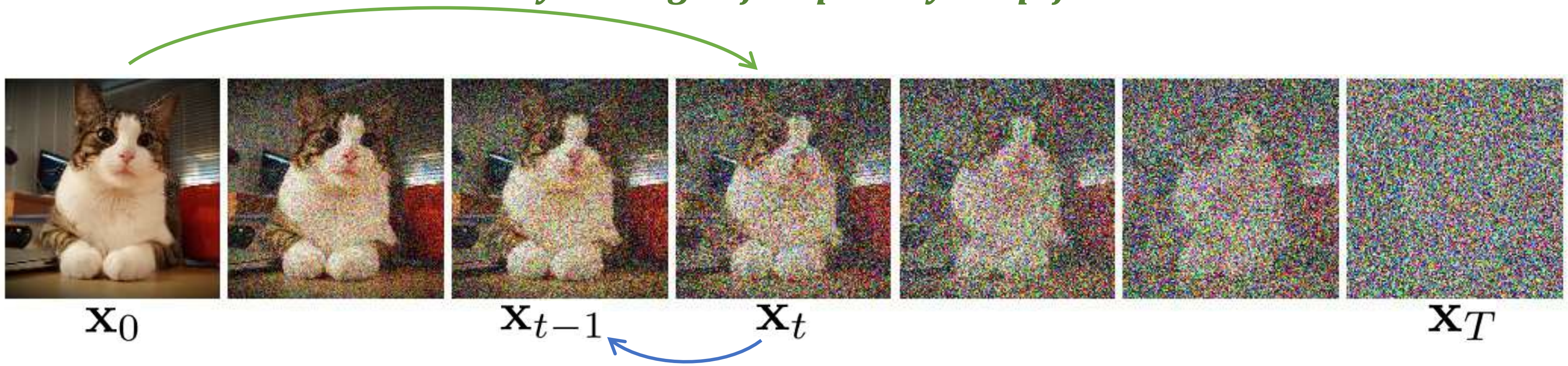


Fourier relative log amplitudes of backbone, skip, and their fused feature maps

Training & Sampling

$+\epsilon$ *rely on high-frequency skip features*

Image



Noise

$-\epsilon$

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

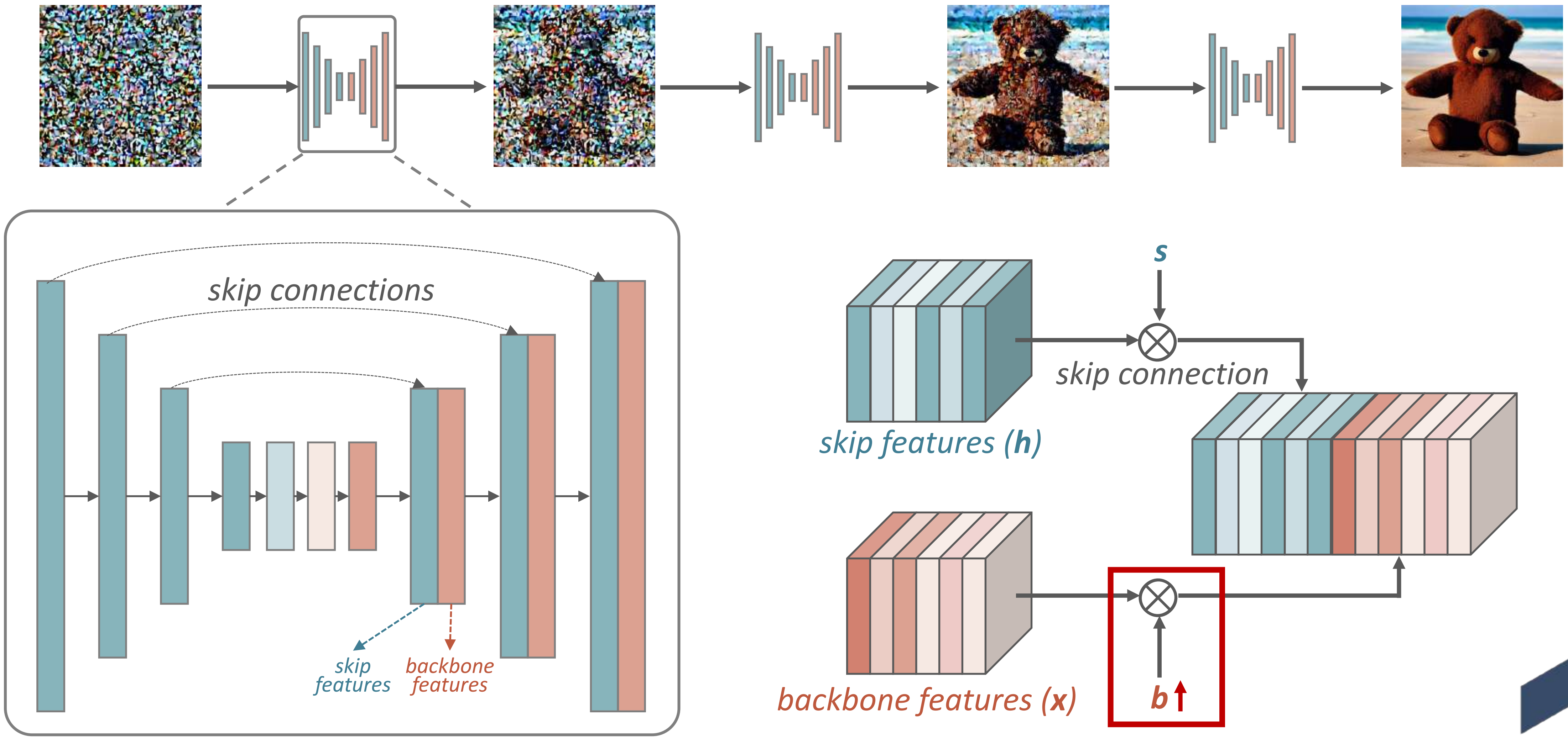
Algorithm 2 Sampling

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

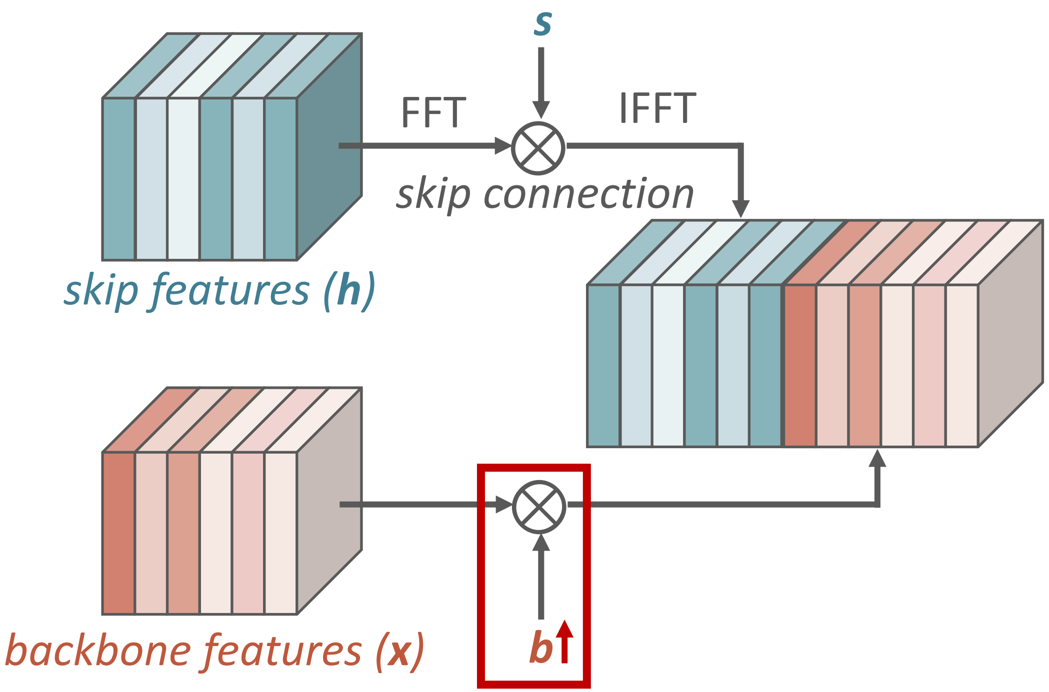


(a) UNet Architecture (b) FreeU Operations

FreeU Method

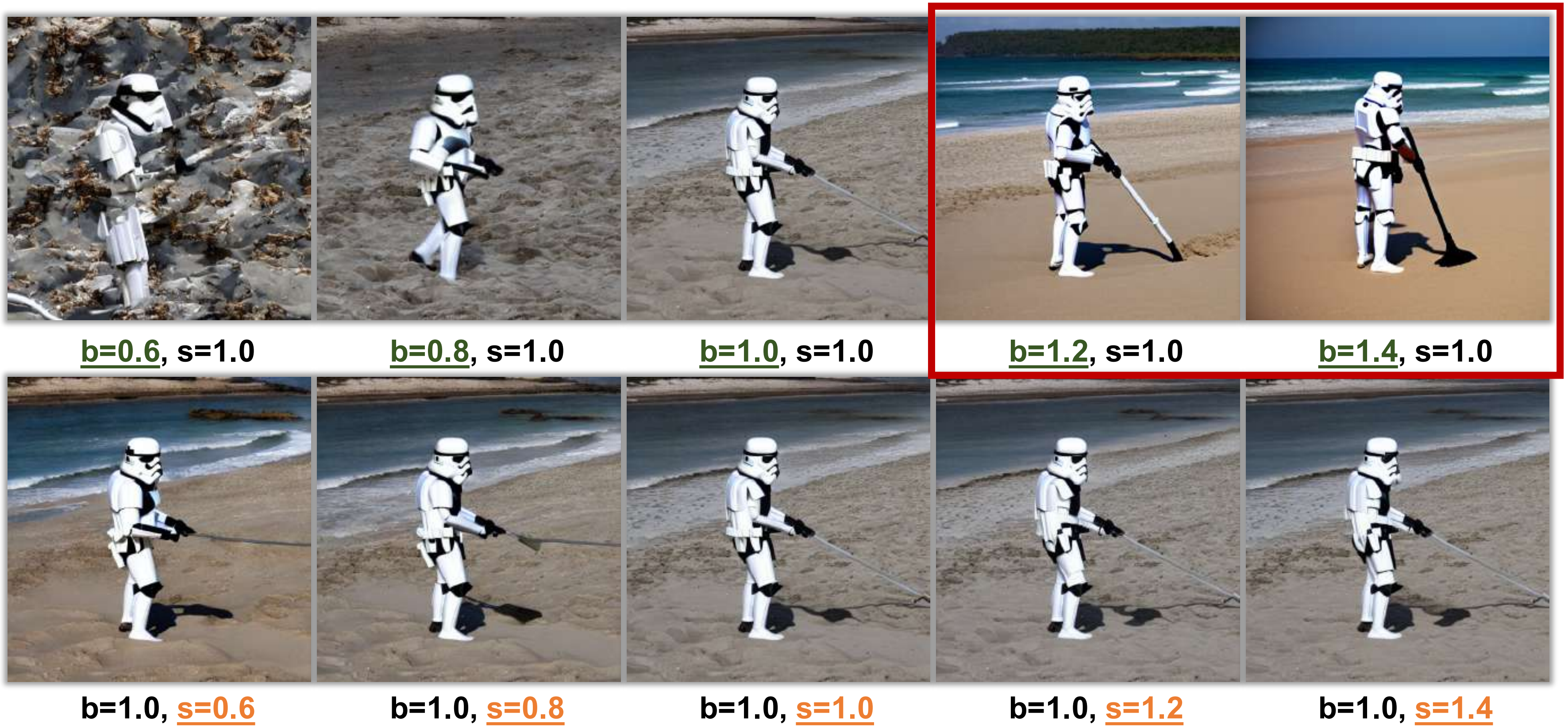
(1) enhance backbone features

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

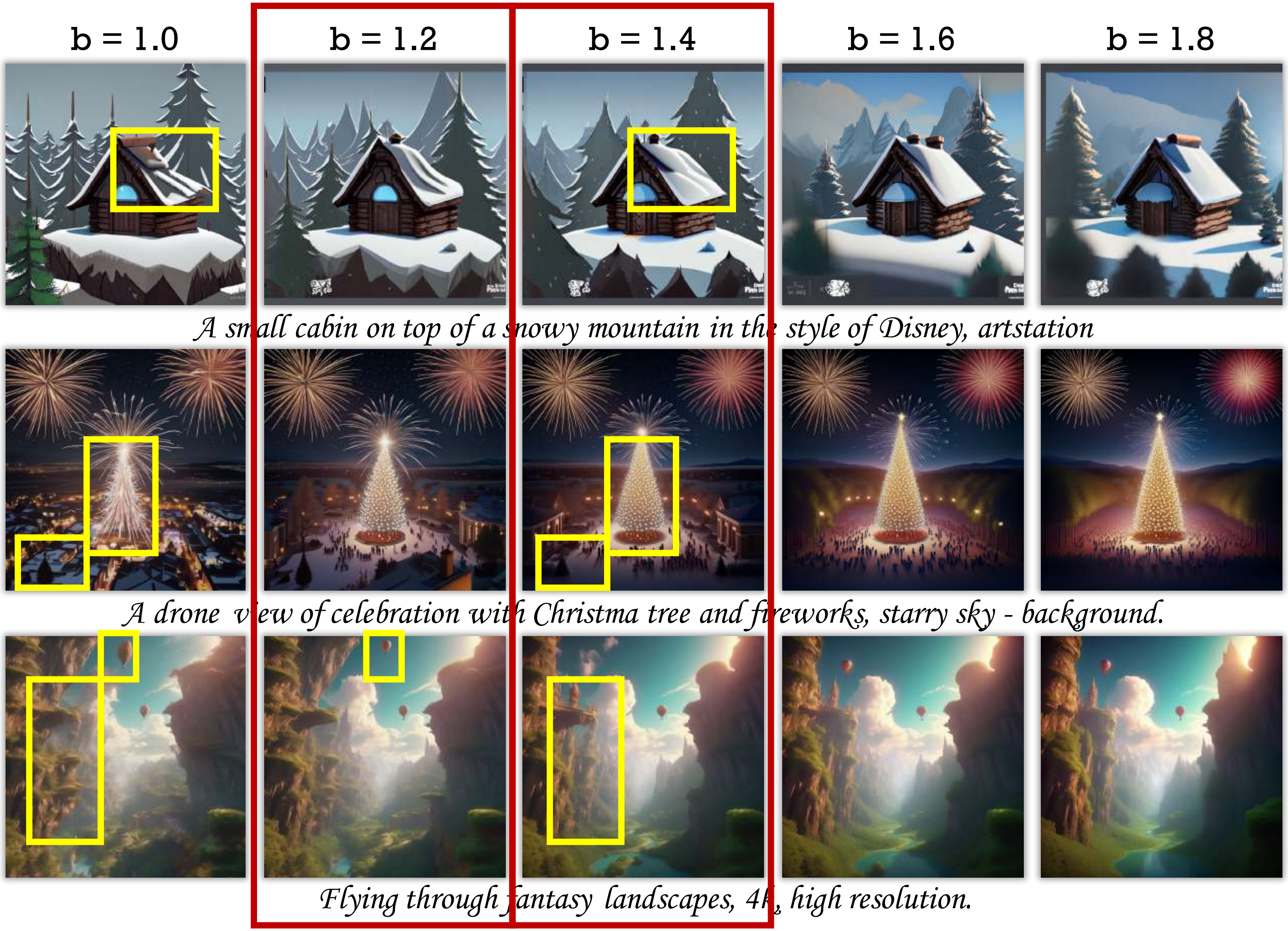


Ablation: Backbone Scaling Factor

- Enhancing backbone features can improve image quality

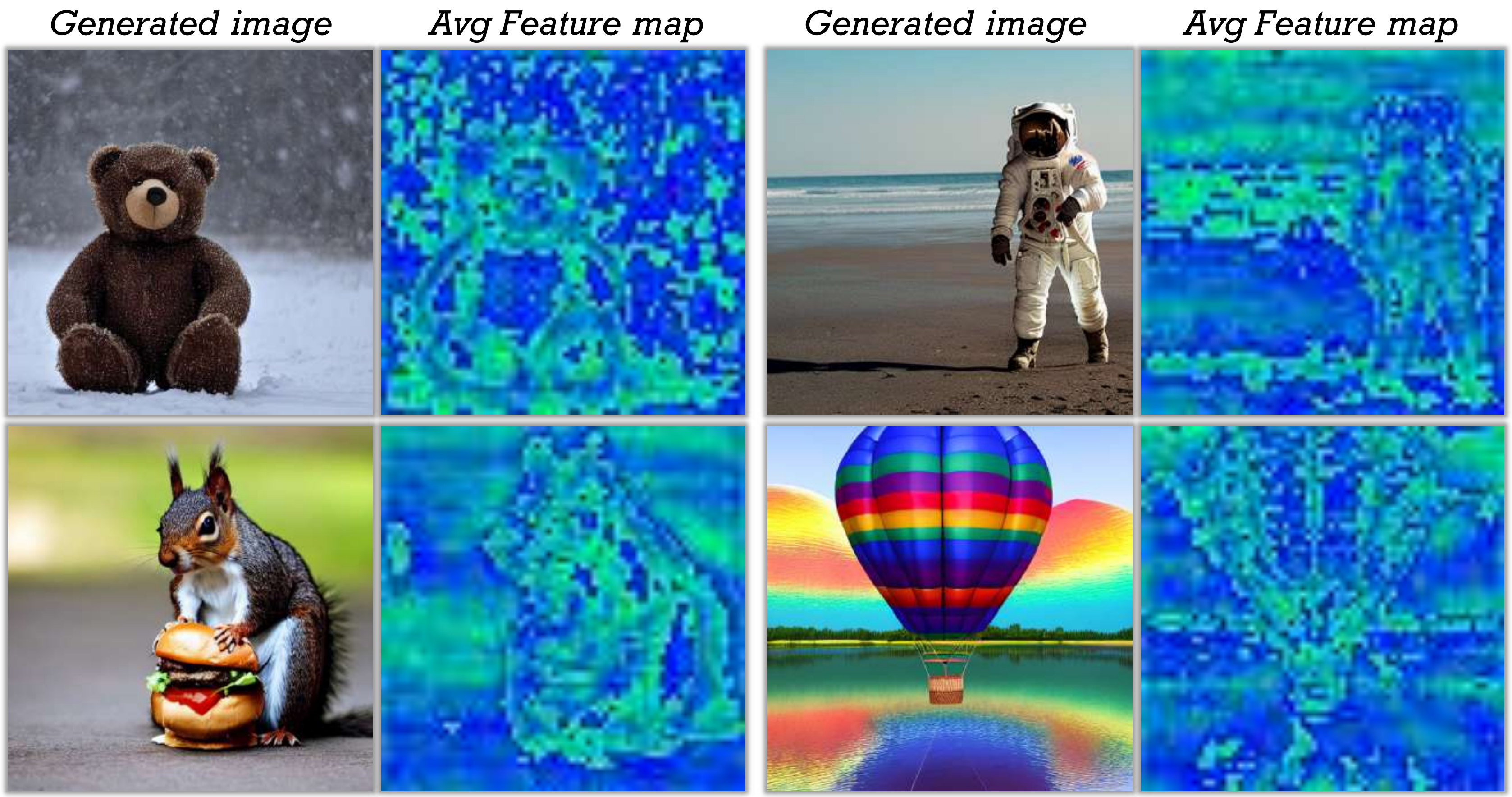


Ablation: Backbone Scaling Factor



Average Backbone Feature Maps

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



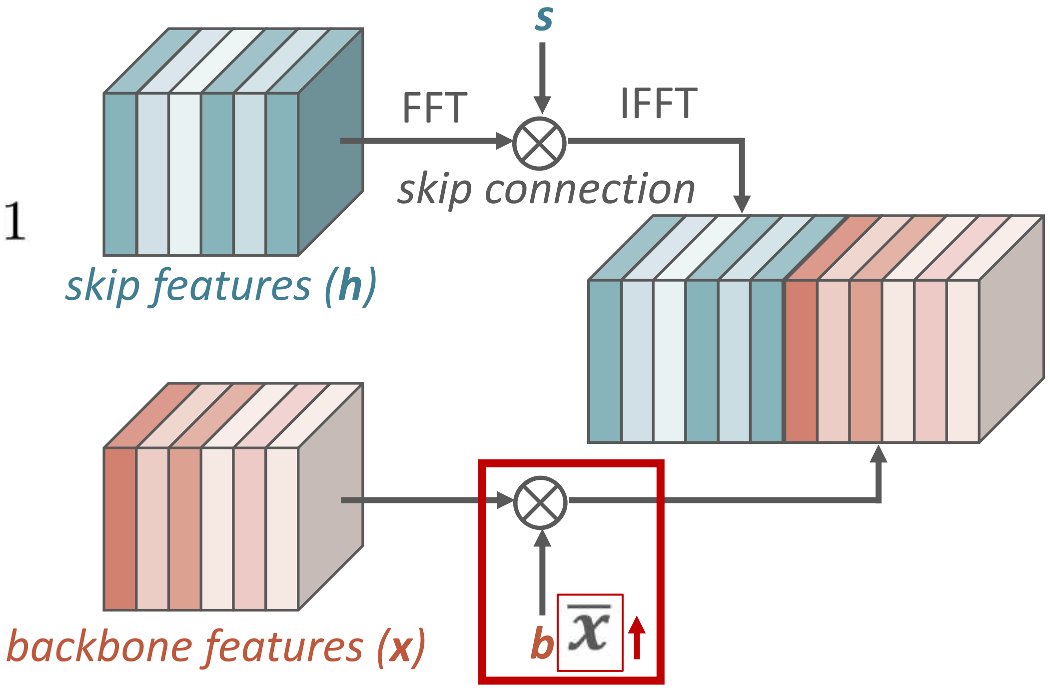
FreeU Method

(1) enhance backbone features

(2) content-aware backbone enhancement

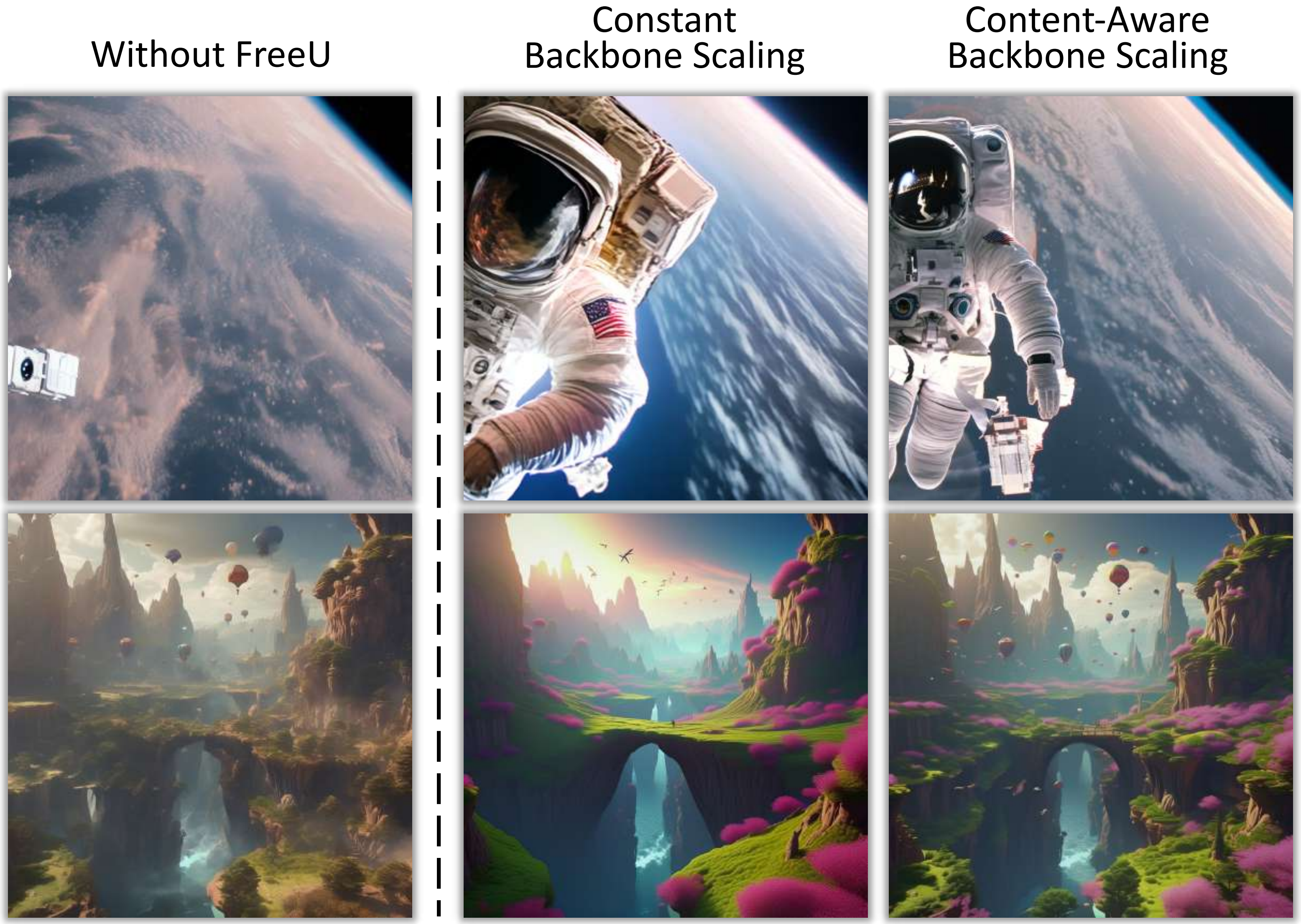
$$\bar{\boldsymbol{x}}_l = \frac{1}{C}\sum_{i=1}^{C} \boldsymbol{x}_{l,i} \qquad \boldsymbol{\alpha}_l = (b_l - 1)\cdot \frac{\bar{\boldsymbol{x}}_l - Min(\bar{\boldsymbol{x}}_l)}{Max(\bar{\boldsymbol{x}}_l) - Min(\bar{\boldsymbol{x}}_l)} + 1$$

- spatially adaptive
- instance specific



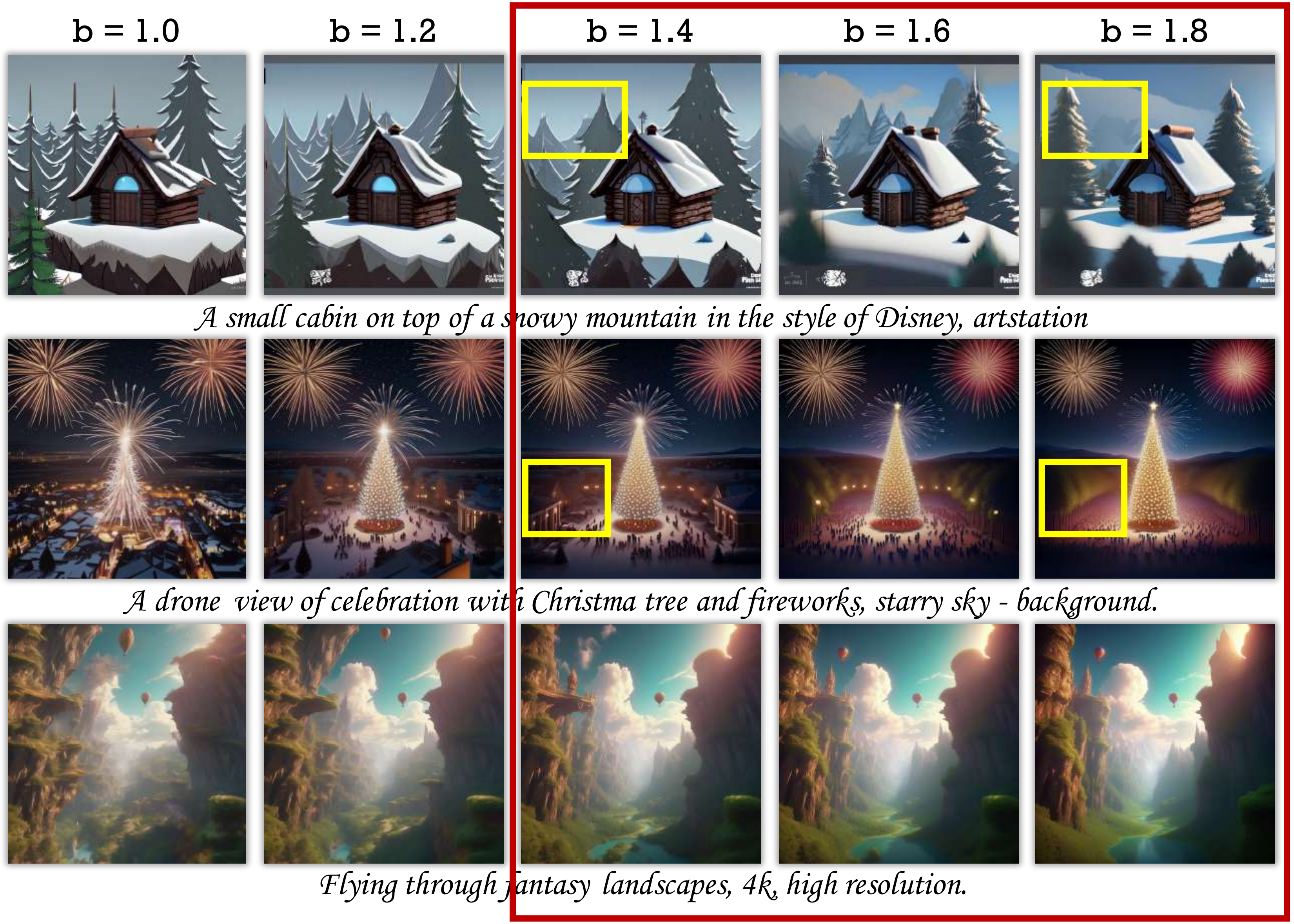
Content-Aware Backbone Scaling

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE
S-LAB FOR ADVANCED INTELLIGENCE



(a) (b) (c)

Ablation: Backbone Scaling Factor



with increased backbone scaling, image can be oversmoothed

FreeU Method

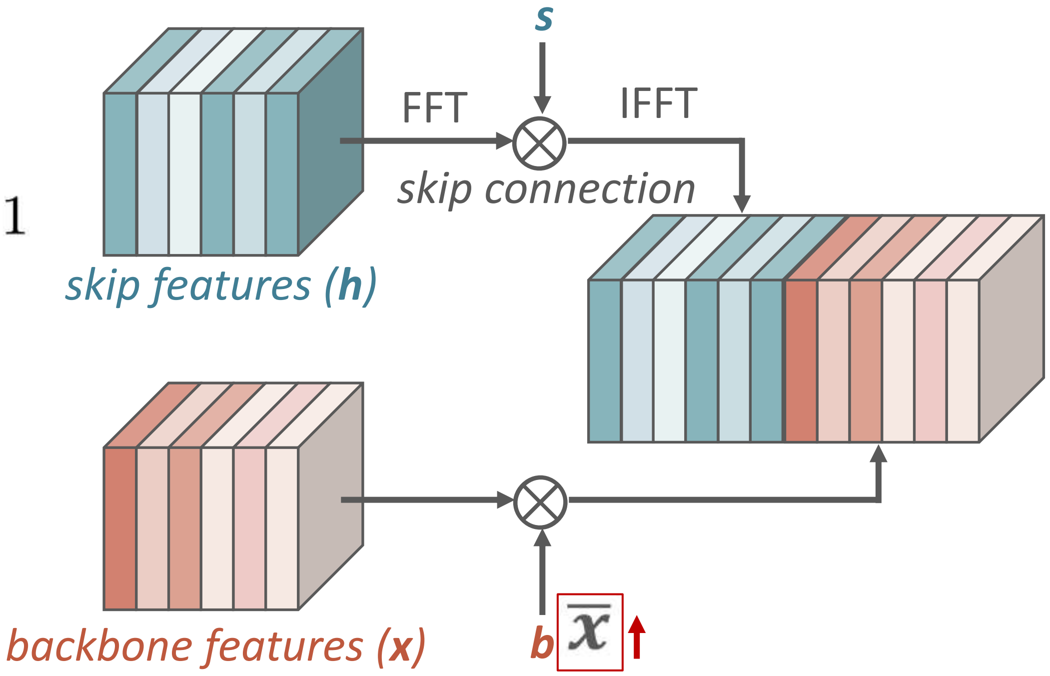
(1) enhance backbone features

(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 <u>**Backbone**</u> Scaling

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

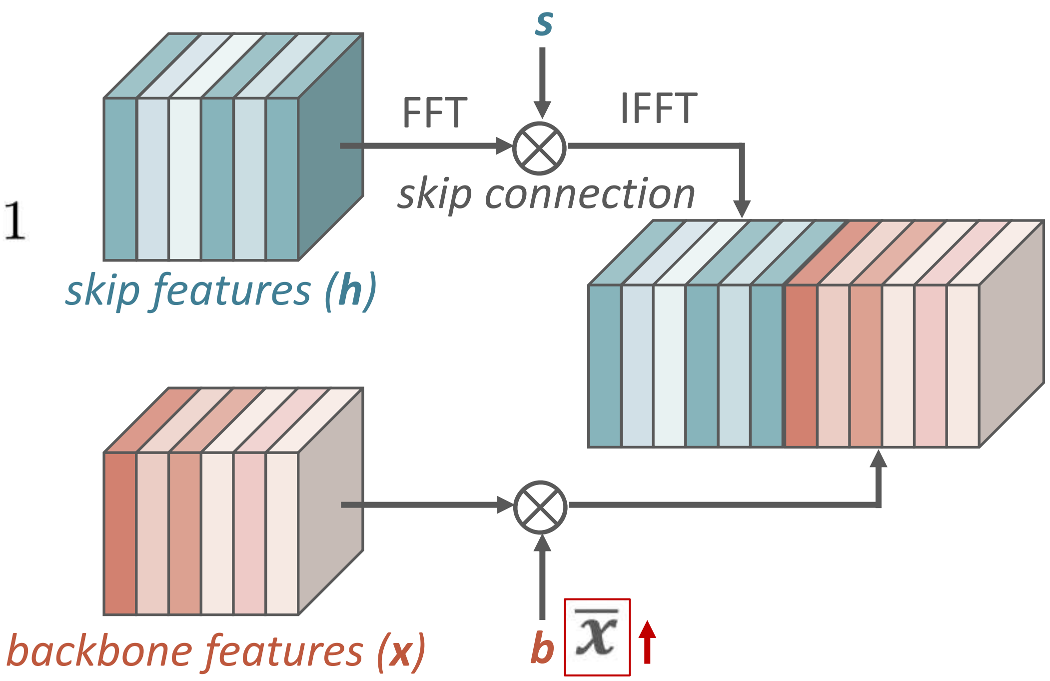
$$\bar{\boldsymbol{x}}_l = \frac{1}{C}\sum_{i=1}^{C} \boldsymbol{x}_{l,i} \qquad \boldsymbol{\alpha}_l = (b_l - 1)\cdot \frac{\bar{\boldsymbol{x}}_l - Min(\bar{\boldsymbol{x}}_l)}{Max(\bar{\boldsymbol{x}}_l) - Min(\bar{\boldsymbol{x}}_l)} + 1$$

(3) channel-selective backbone enhancement

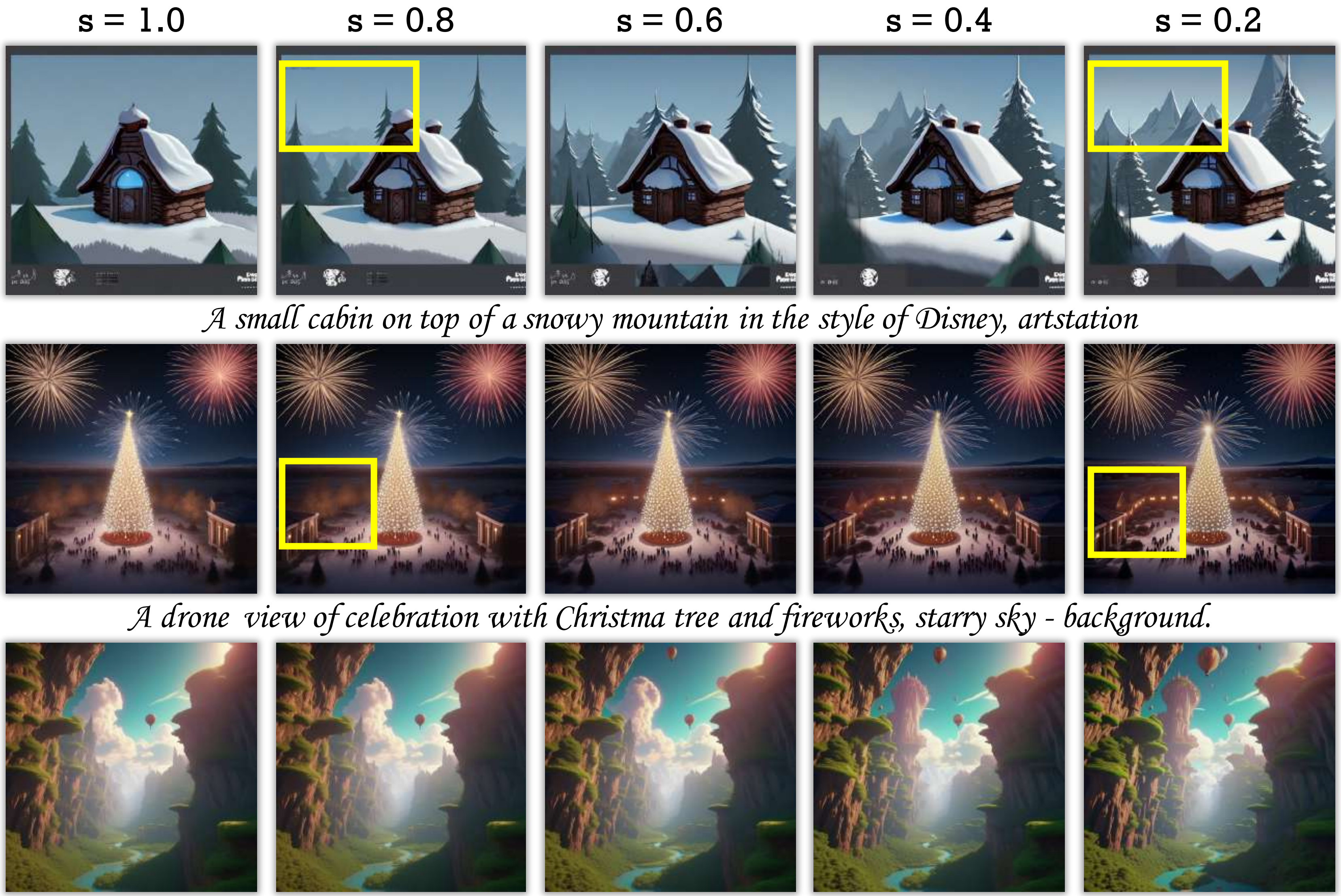
$$\boldsymbol{x}'_{l,i} = \begin{cases} \boldsymbol{x}_{l,i} \odot \boldsymbol{\alpha}_l, & \text{if } i < C/2 \\ \boldsymbol{x}_{l,i}, & \text{otherwise} \end{cases}$$

(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



A small cabin on top of a snowy mountain in the style of Disney, artstation

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

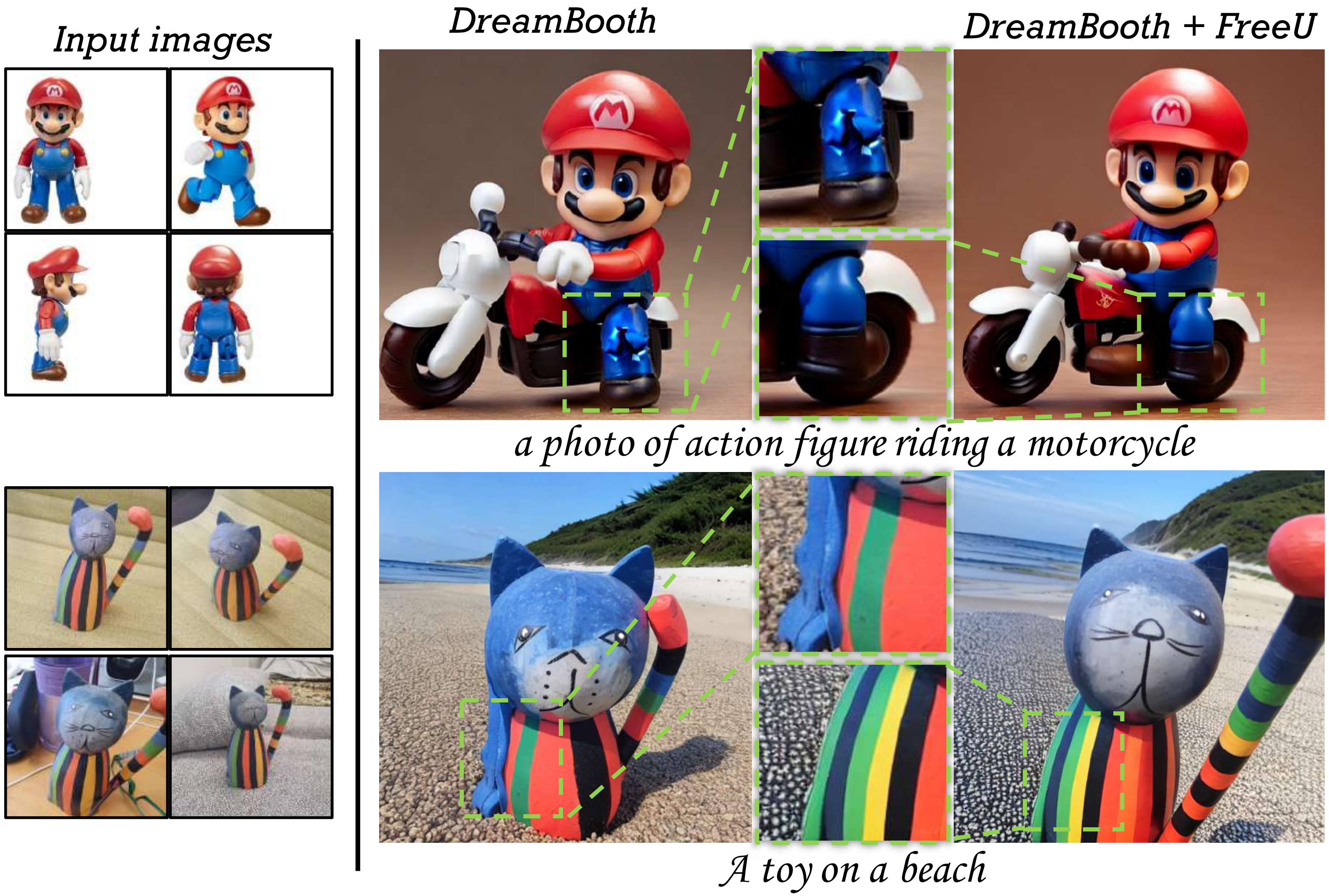
Flying through fantasy landscapes, 4k, high resolution.

Visual Results: Text-to-Image

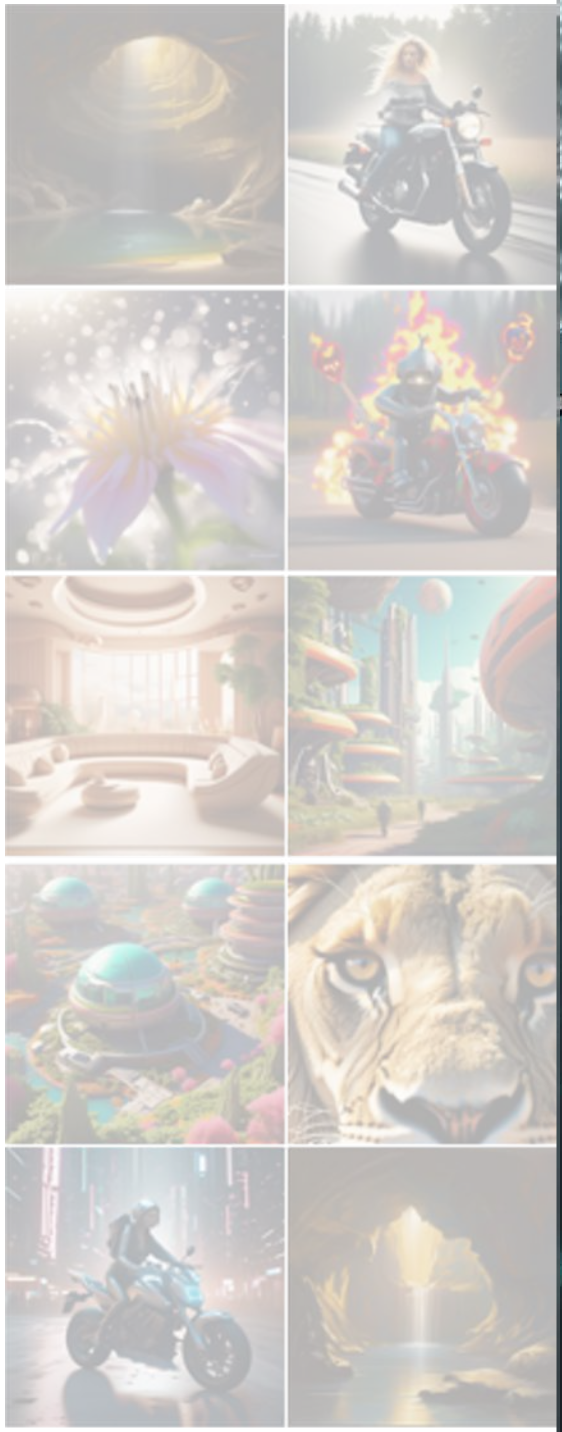
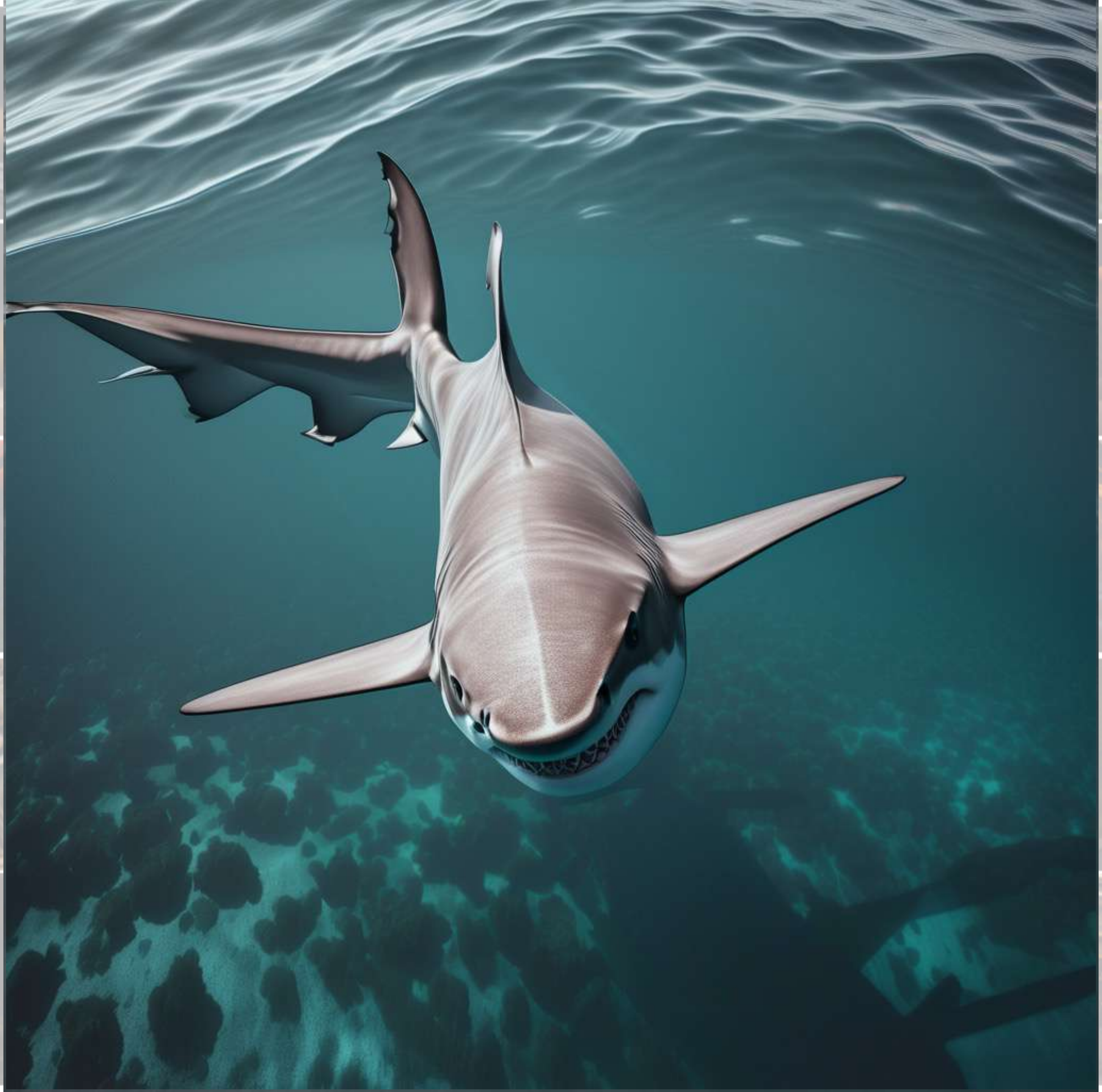
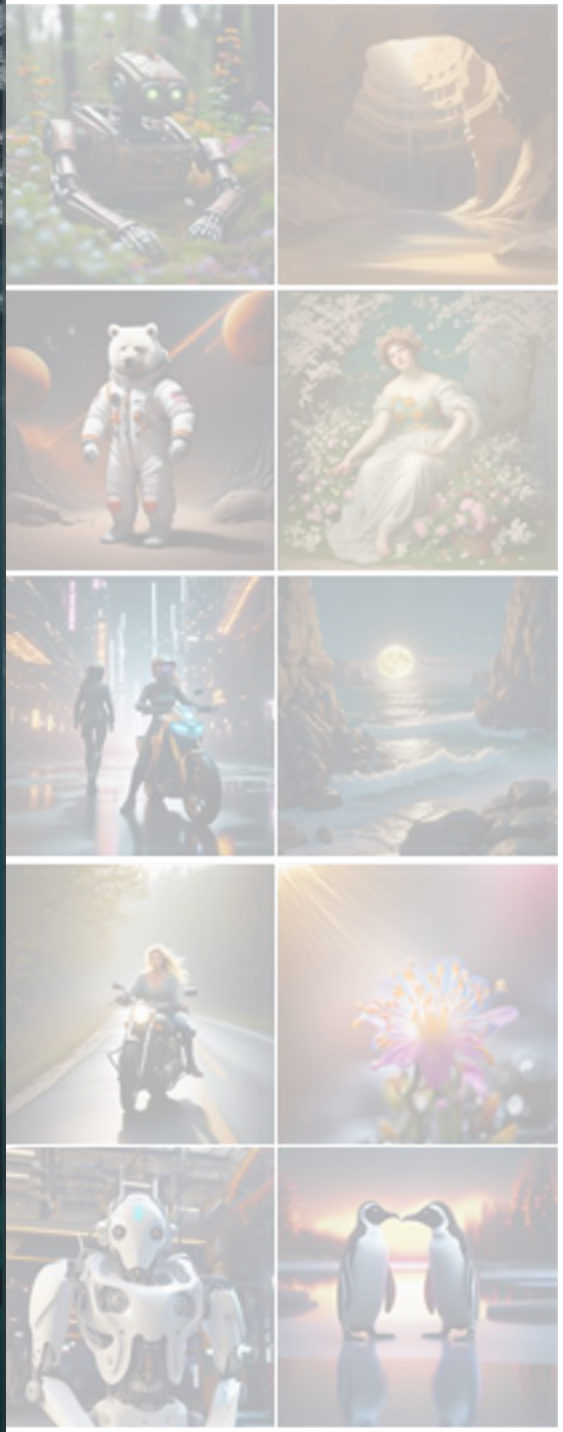
Visual Results: Text-to-Video

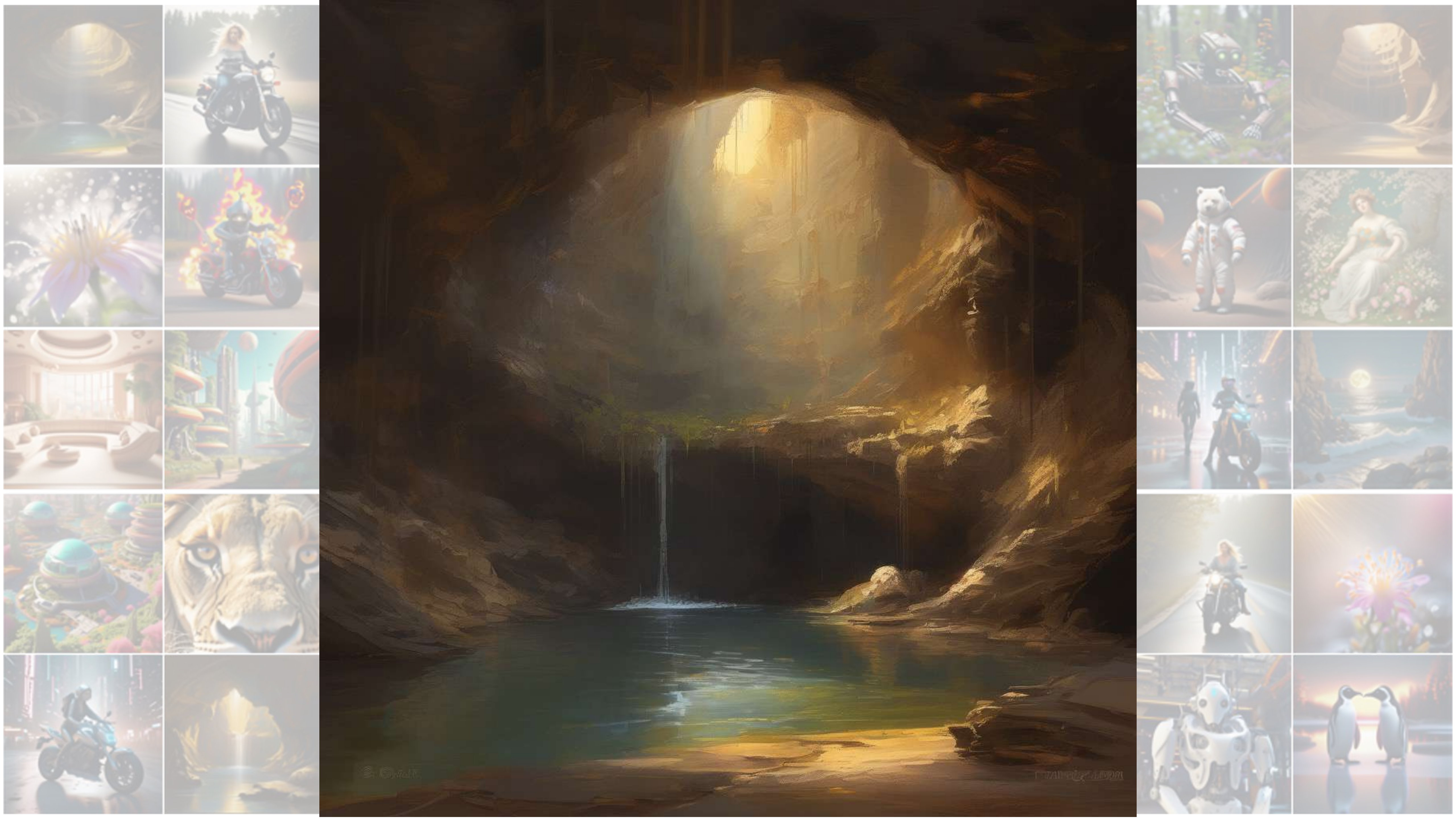
Visual Results: Text-to-Video

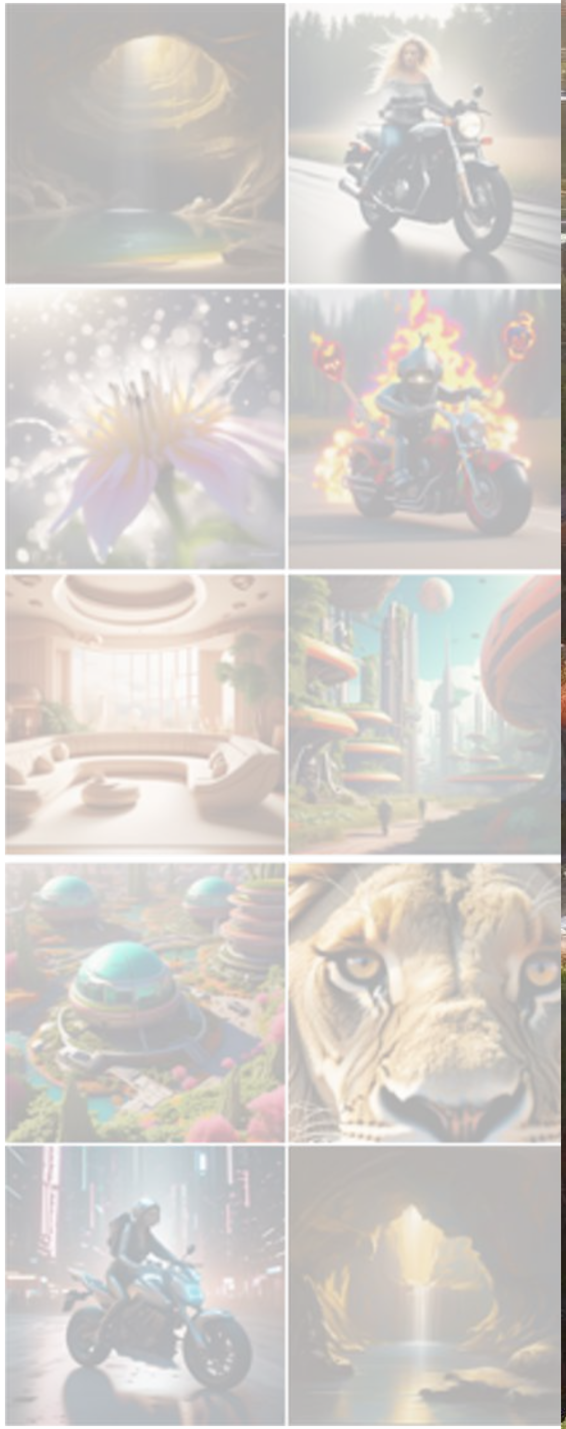
Visual Results: Personalized Text-to-Image

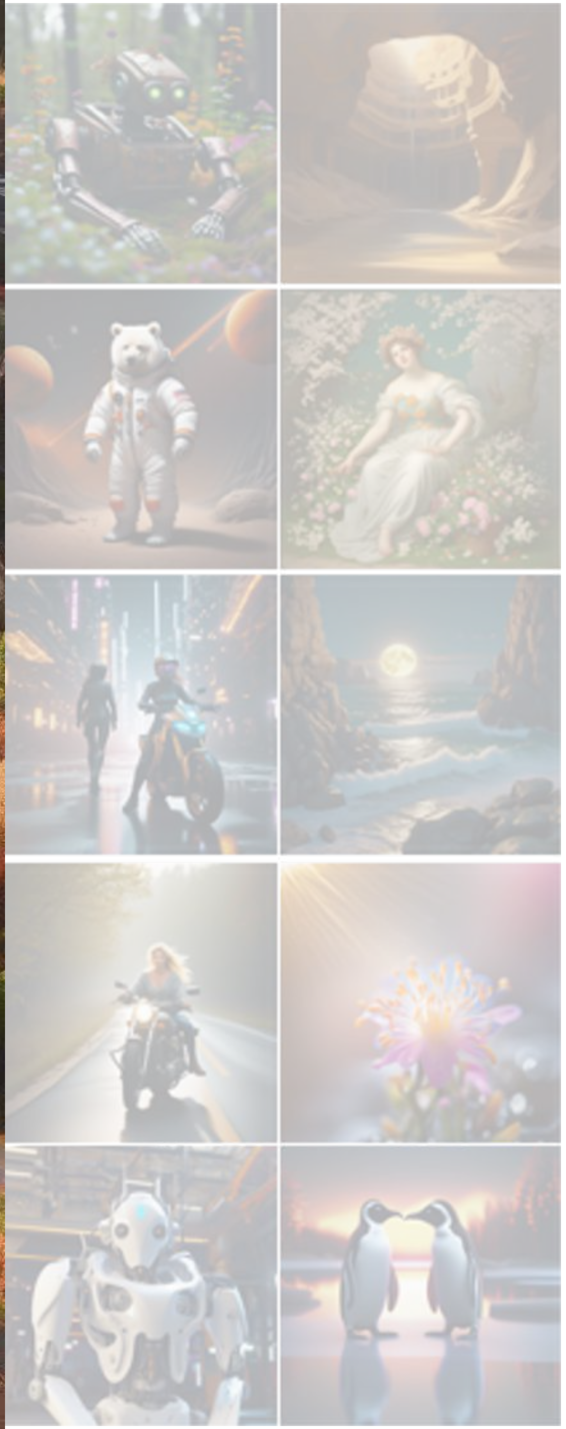


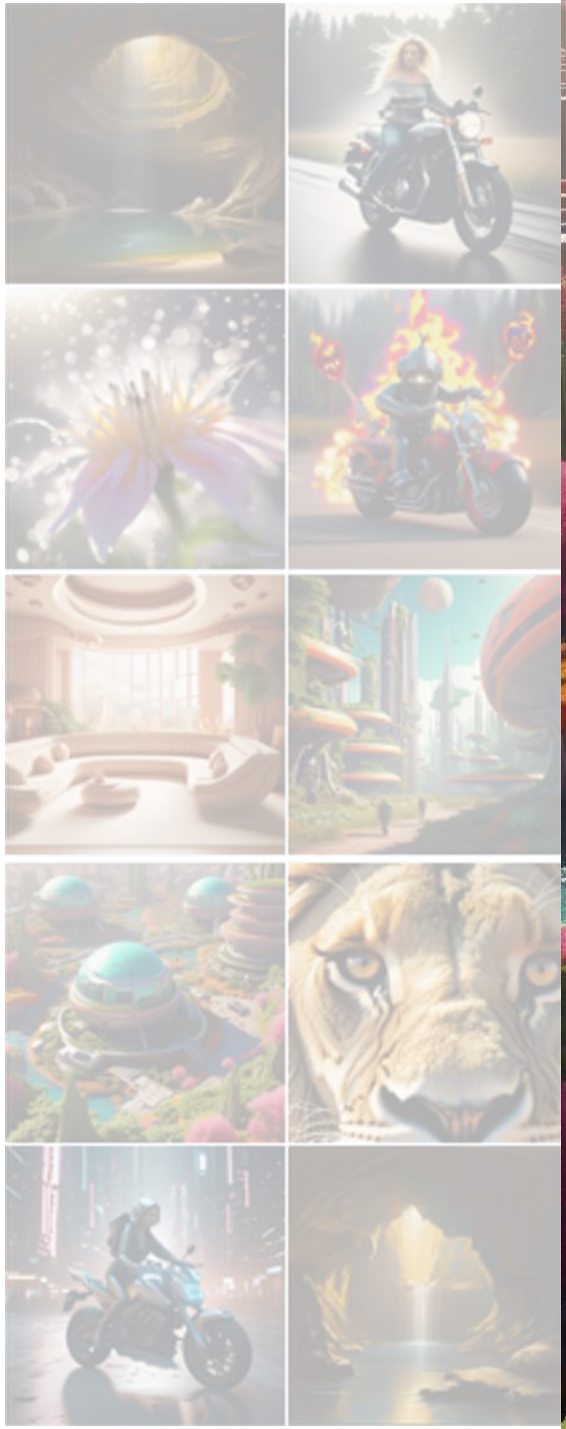
Ruiz et al. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation

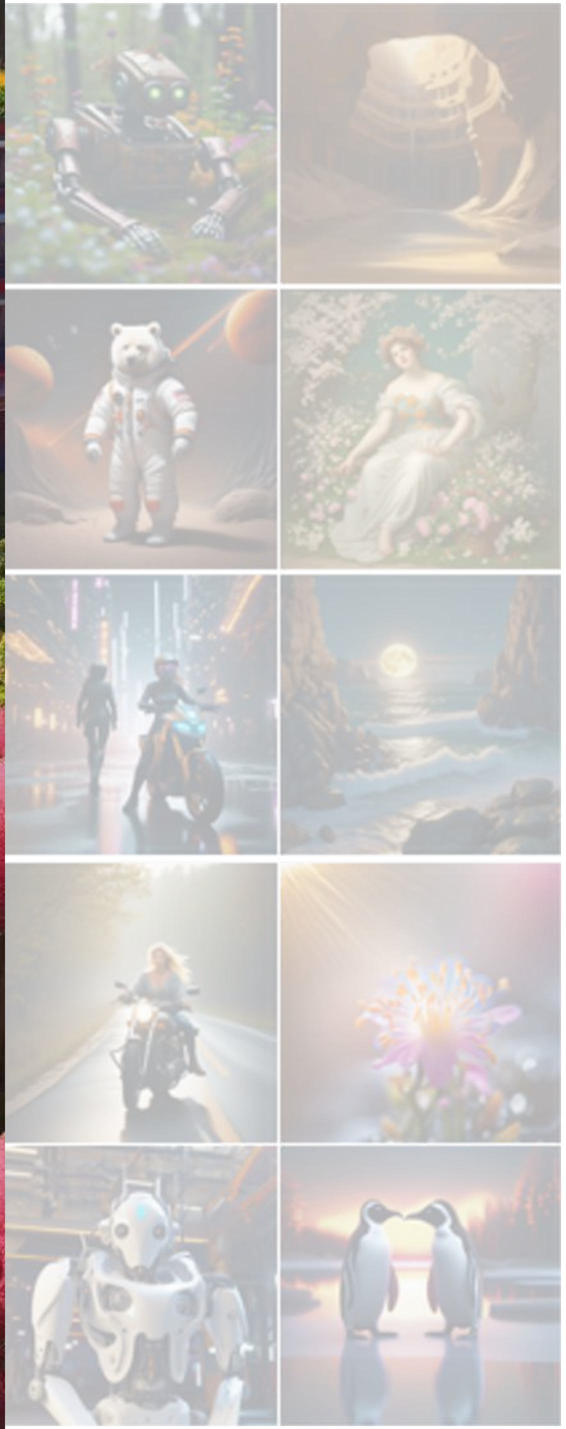






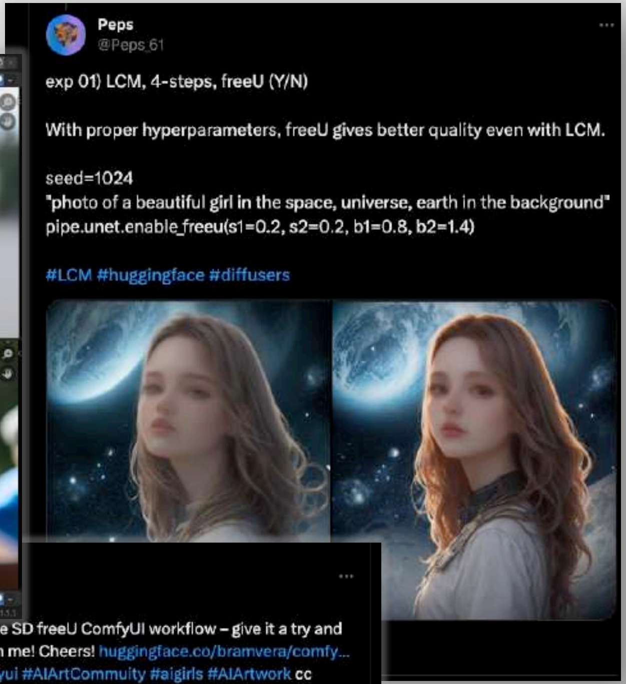


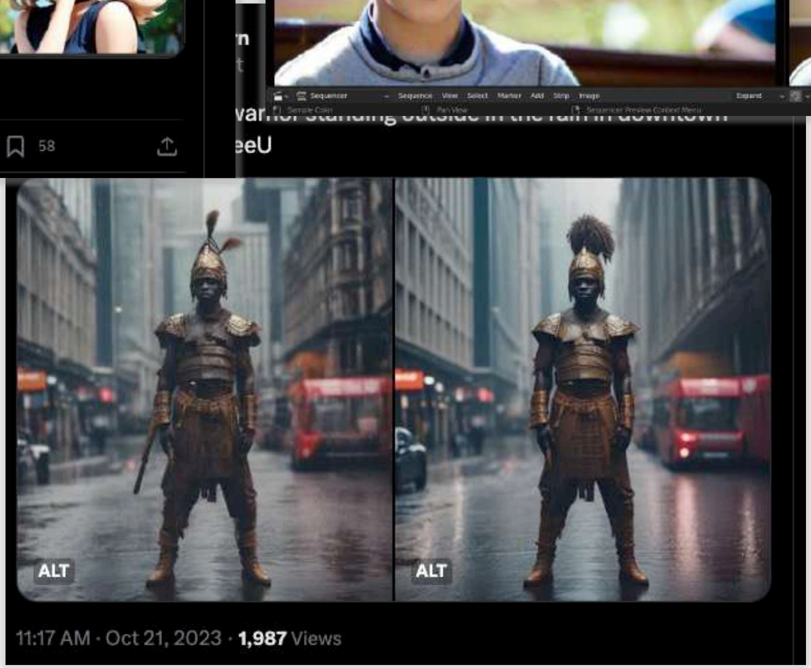


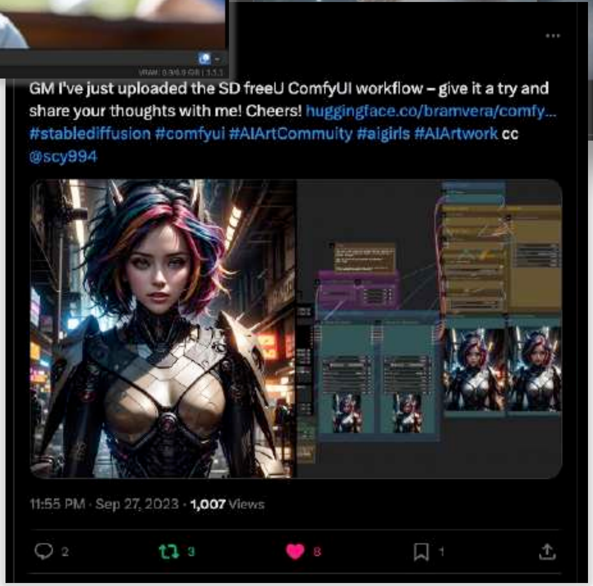




Community Contributions







Q&A

Poster Session

- Today 5 PM
- Arch 4A-E Poster #153
- Welcome any questions & discussions

CVPR JUNE 17-21, 2024 SEATTLE, WA

S-LAB FOR ADVANCED INTELLIGENCE

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

FreeU: Free Lunch in Diffusion U-Net

Chenyang Si Ziqi Huang Yuming Jiang Ziwei Liu

MMLab@NTU | S-Lab, Nanyang Technological University

Code

Project Page