VLP for High-Resolution Visual Synthesis

Chenfei Wu

MSRA NCL Group

Overview

- **GAN**
  - **from global latents** *(LocoGAN, SinGAN, Infinite-GAN, MS-PIE, TileGAN)* — *Repetitive*
  - **from coordinate latents** *(ALIS, InfiniteGAN)* — *Non-repetitive But rough transition*
- **NAR** **mask by mask** *(MaskGIT)* — *Smooth transition between **masks***
- **AR** **token by token** *(Taming Transformers)* — *Smooth transition between **tokens** Rely on sketches*
- **Diffusion** **Vague to Clear** *(Latent Diffusion)* — *Fixed Size*
- **SR** **from low-resolution images** *(CogView, DALLE-2, Imagen)* — *Fixed Size*
- **AR over AR** *NUWA-Infinity* — *Infinite Size*

GAN-based Models

[1] from global latents *(LocoGAN, SinGAN, Infinite-GAN, MS-PIE, TileGAN)*

Repetitive

$$z \sim N(0, I) \longrightarrow w$$

LocoGAN generates different patches with the same global latents, which limits the diversity and leads to repetitive results.

GAN-based Models

[2] from coordinate latents *(ALIS, InfiniteGAN)*

Non-repetitive
But rough transition

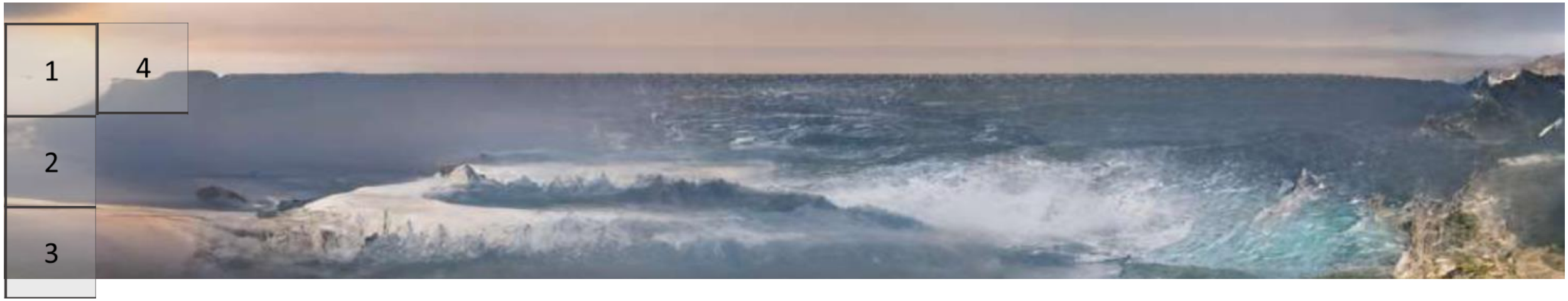
$$z \sim N(0, I) \longrightarrow w_1, w_2, w_3, \ldots$$

To address the repetition issue, ALIS generates different patches with different coordinate latents. However, this sometimes brings rough transition between patches when the coordinate latents are too different.

AR-based Models

(Taming Transformers)

Smooth transition between tokens
But slow inference



Auto-Regressive Models are naturally beneficial for generating smooth transition results. Taming Transformer views images as discrete tokens and generate a high-resolution image token by token, which considers the smooth transition between nearby tokens. As a result, it fails to model long-range consistency and has a slow inference speed.

NAR-based Models

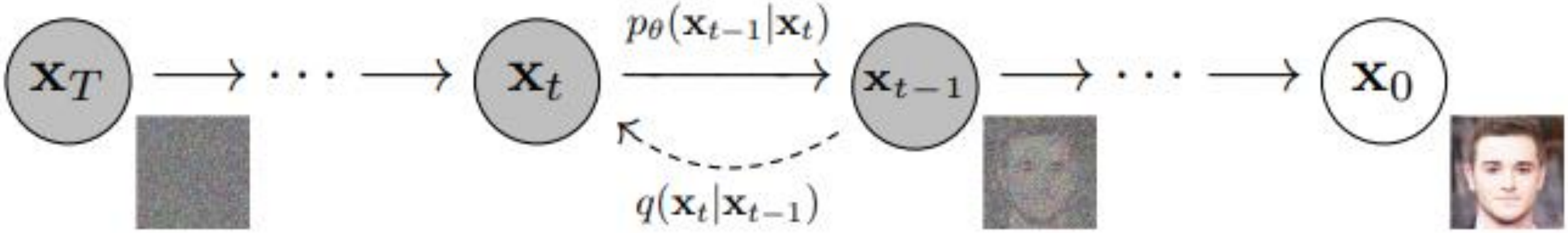
Fast
*Smooth transition between **masks***

(MaskGiT)

1	4	[M]
2	5	[M]
3	6	[M]

MaskGIT generates masked tokens based on visible tokens inside a sliding window. The inference speed is improved by generating masked tokens in parallel in a fixed steps. Although it considers the smooth transition between the generated masked tokens, it still fails to model long-range consistency.

Diffusion



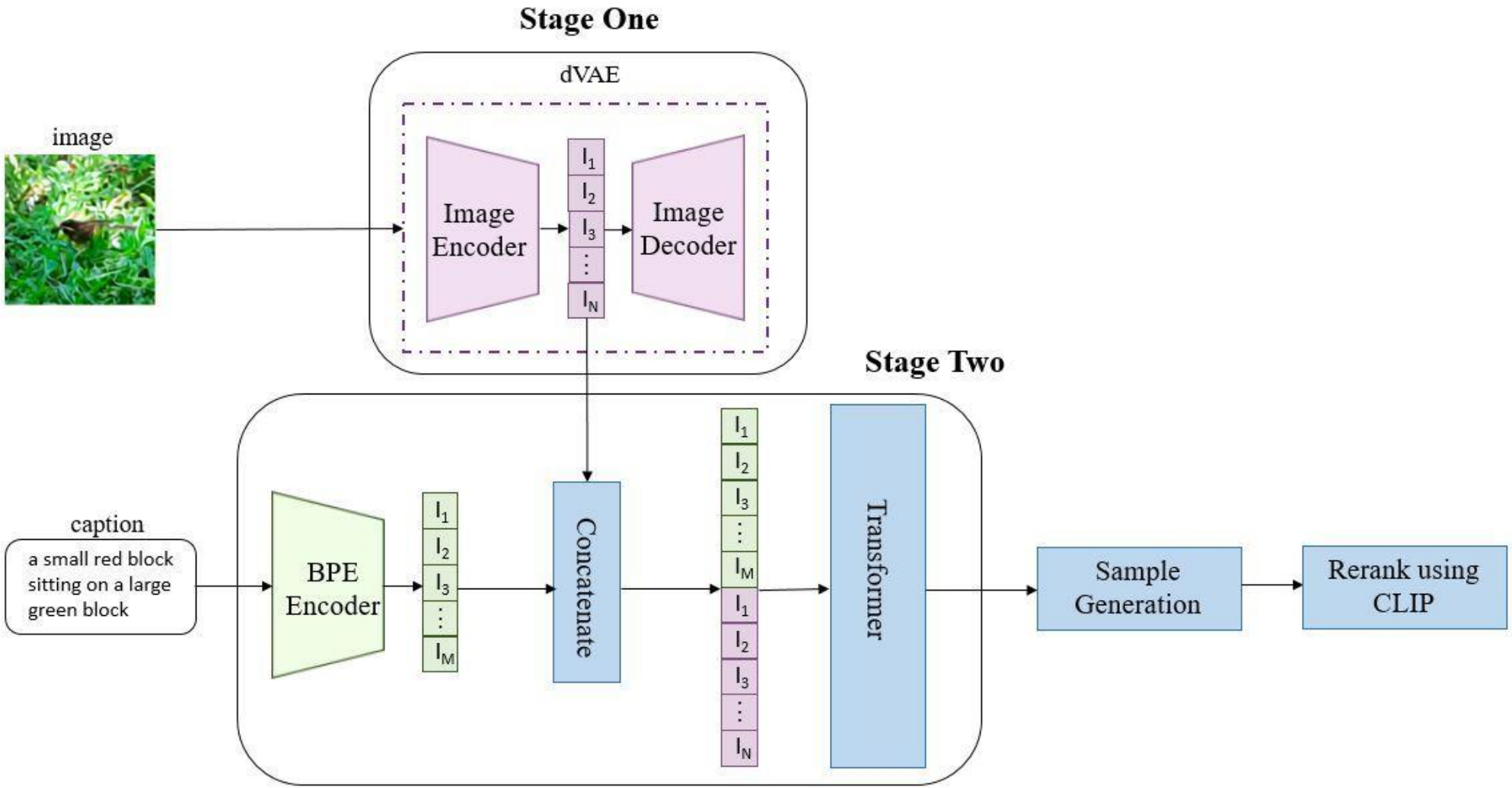
Algorithm 1 Training

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

Algorithm 2 Sampling

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

DALL-E


Stage One
dVAE
image
Image Encoder
I_1 I_2 I_3 ⋮ I_N
Image Decoder
Stage Two
caption
a small red block sitting on a large green block
BPE Encoder
I_1 I_2 I_3 ⋮ I_M
Concatenate
I_1 I_2 I_3 ⋮ I_M I_1 I_2 I_3 ⋮ I_N
Transformer
Sample Generation
Rerank using CLIP

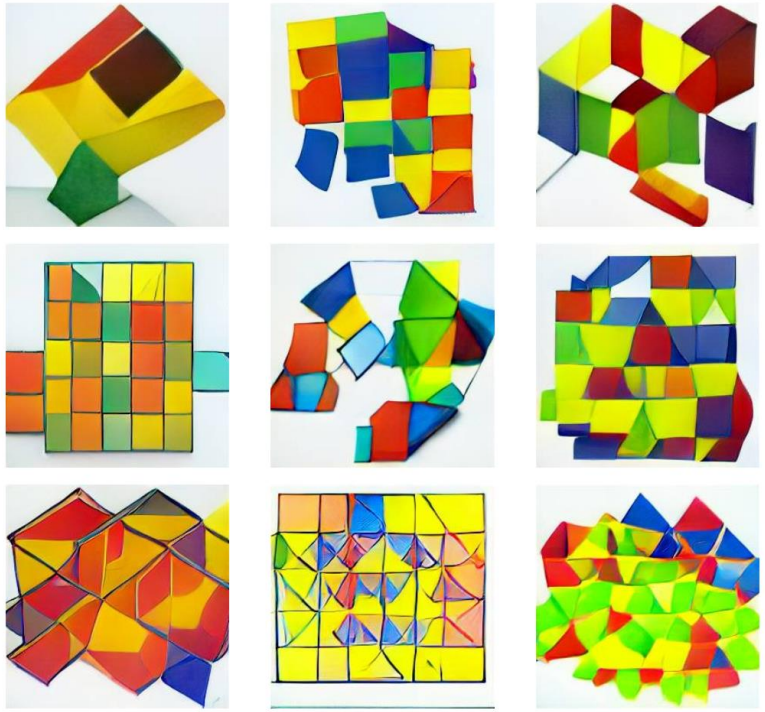
DALL-E

Generate images from text

What do you want to see?

gn multiple three-dimensional geometric squares with four main colors of red, yellow, blue and green

Design multiple three-dimensional geometric squares with four main colors of red, yellow, blue and green



These results have been obtained using model `wzoooa1c:v0` from an ongoing training run.

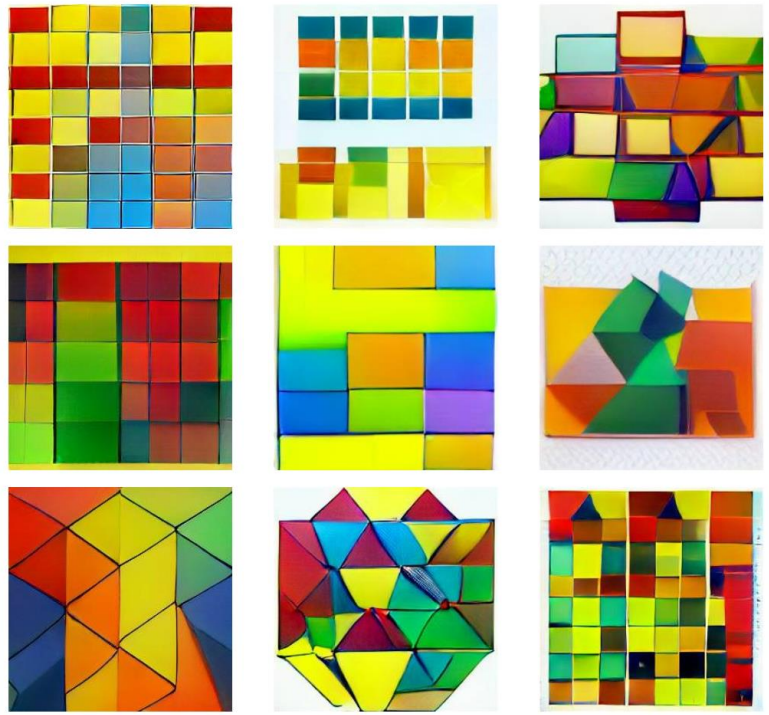
Again!

Generate images from text

What do you want to see?

olors of red, yellow, blue and green, to have a sense of composition and a sense of depth in the space.

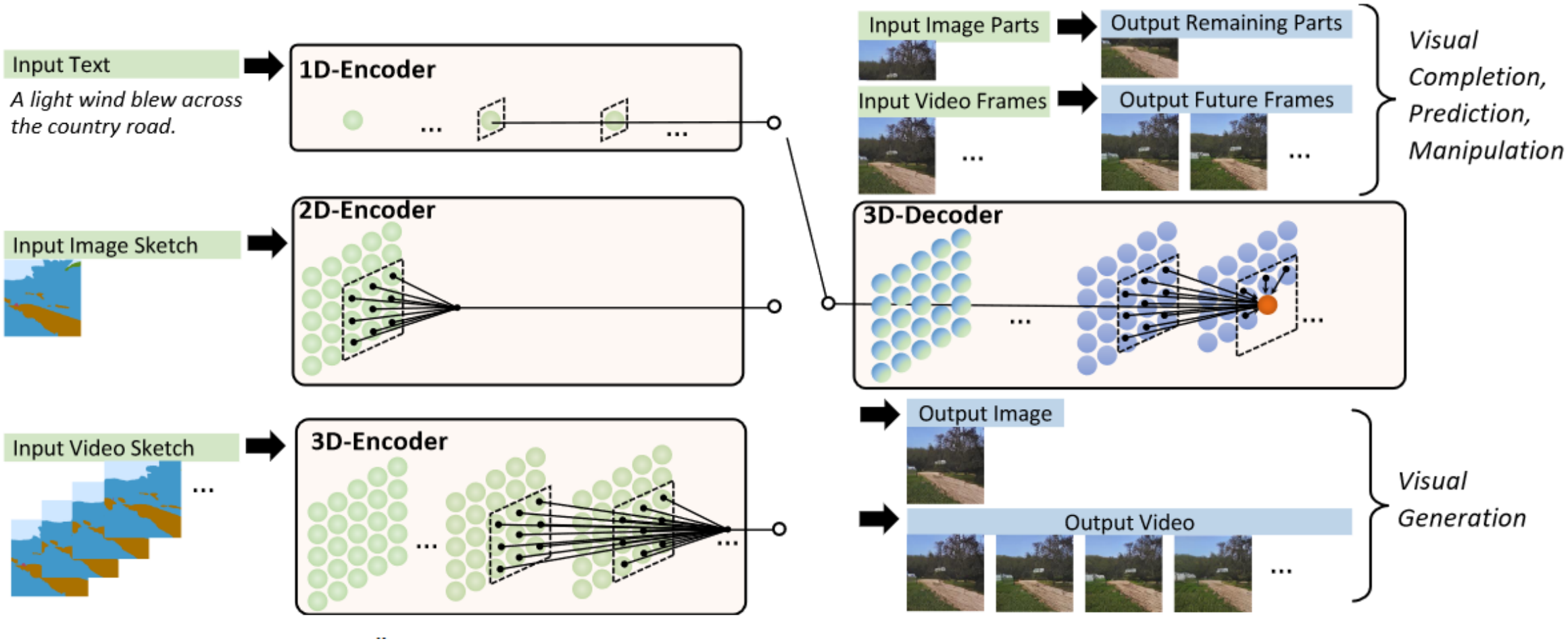
Design multiple three-dimensional geometric squares with four main colors of red, yellow, blue and green, to have a sense of composition and a sense of depth in the space.



These results have been obtained using model `wzoooa1c:v0` from an ongoing training run.

Again!

NUWA



[1] A unified multimodal pretrained model that can generate both images and videos
[2] A 3D encoder-decoder framework to improve quality and efficiency.

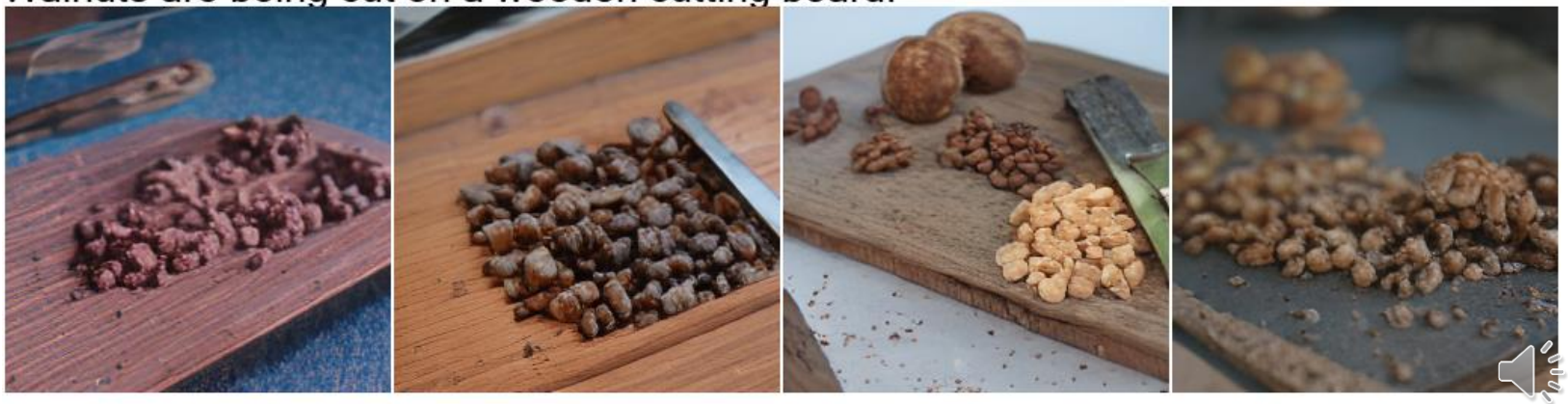
Task 1:
Text-To-Image (T2I)

NÜWA achieves a significant high FID score of **12.9**, outperforms **27.1** of DALL-E from Open AI.

A wooden house sitting in a field.

A young girl eating a very tasty looking slice of pizza.

Walnuts are being cut on a wooden cutting board.



Task 2:
Sketch-To-Image (S2I)

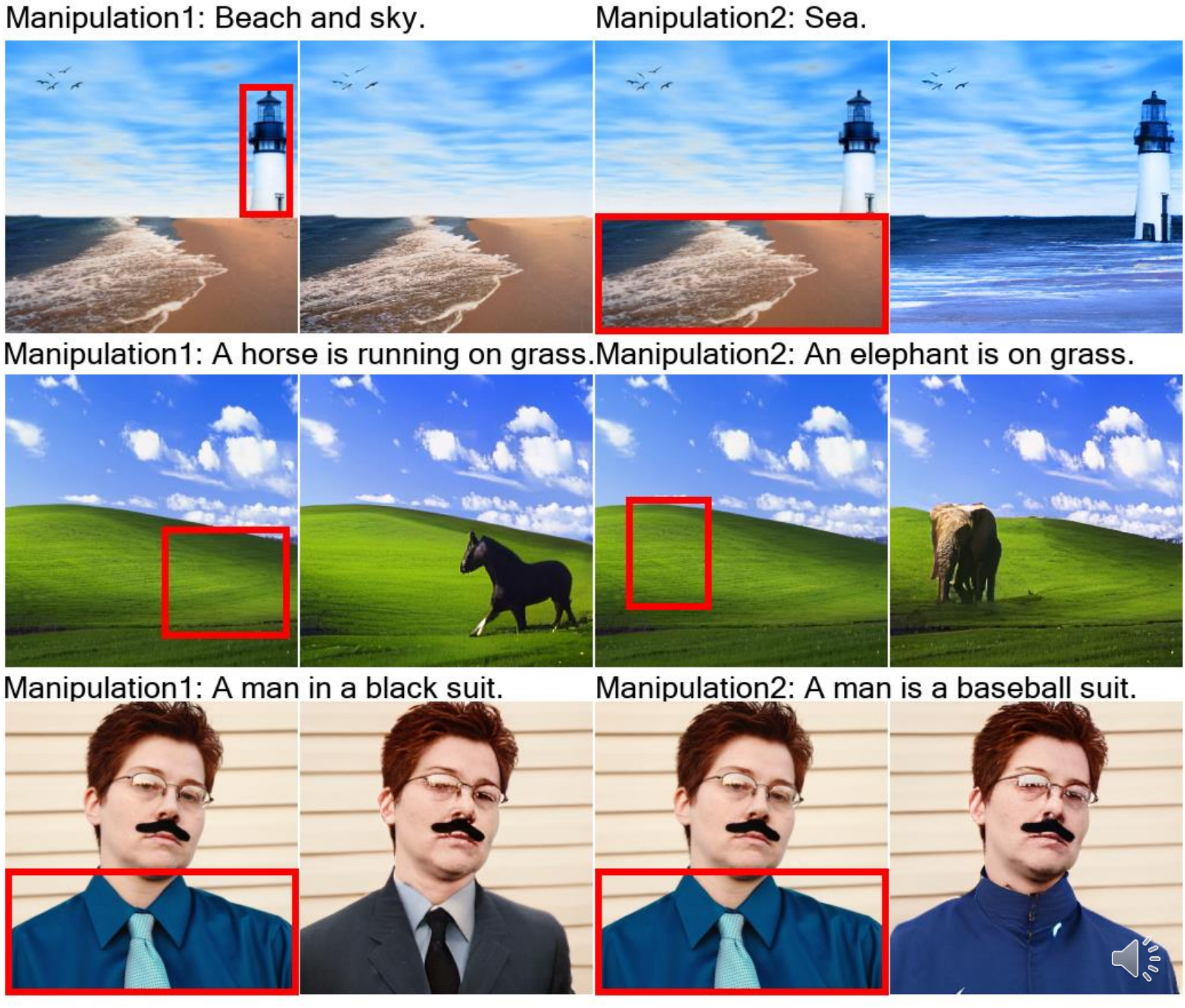
NÜWA generates a diverse realistic results, even the reflection of the bus window is clearly visible.

Task 3:
Image Completion (I2I)

NÜWA shows strong imagination and can even complete an image with only **5%** tokens.

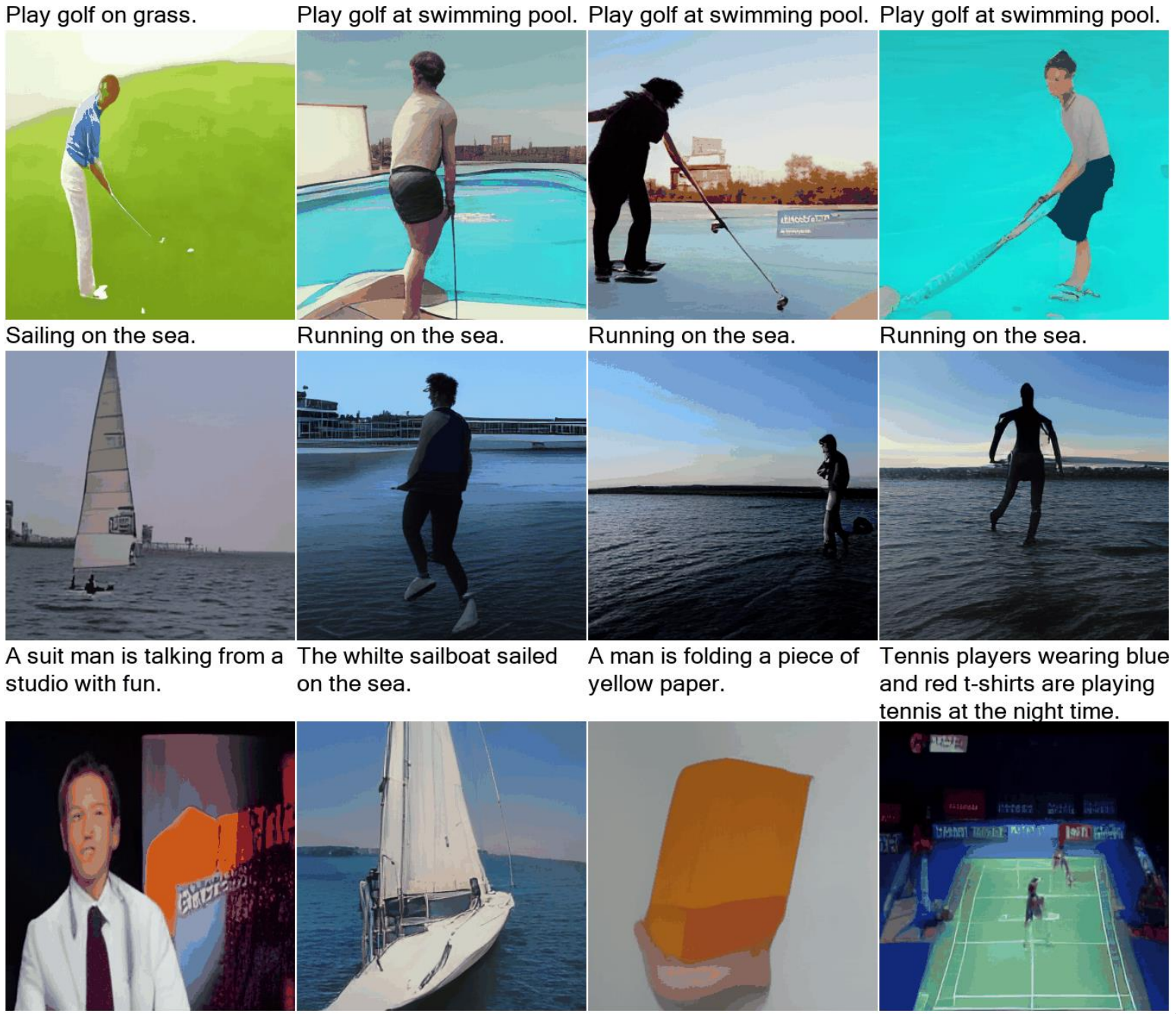
Task 4:
Image Manipulation (I2I)

NÜWA can
add, remove or change
any parts of an image to whatever you want.



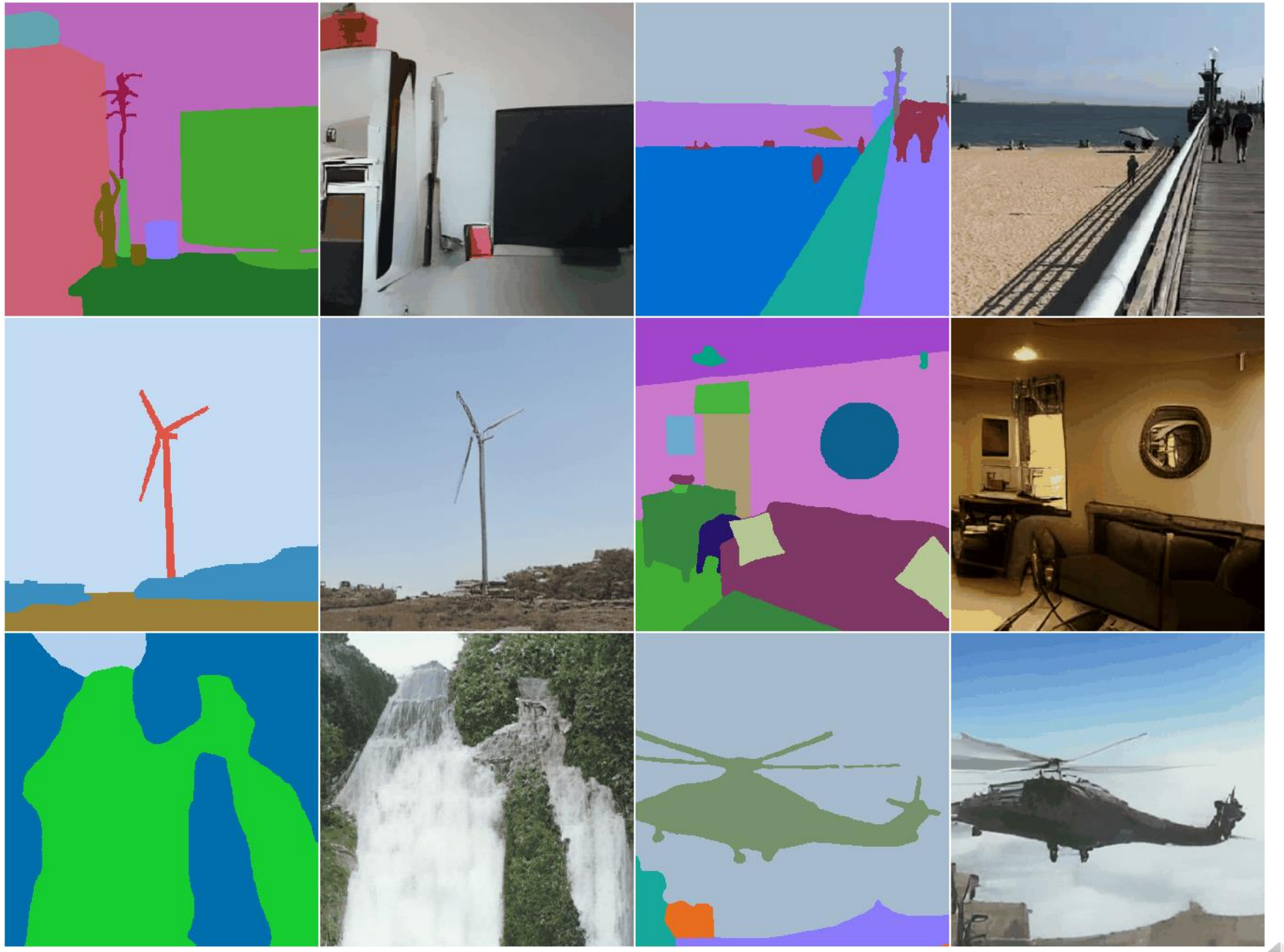
Task 5:
Text to Video(T2V)

NÜWA can generate unseen videos like “Running on the sea.”



Task 6:
Sketch to Video (S2V)

NÜWA generates temporally consistent open-domain videos.



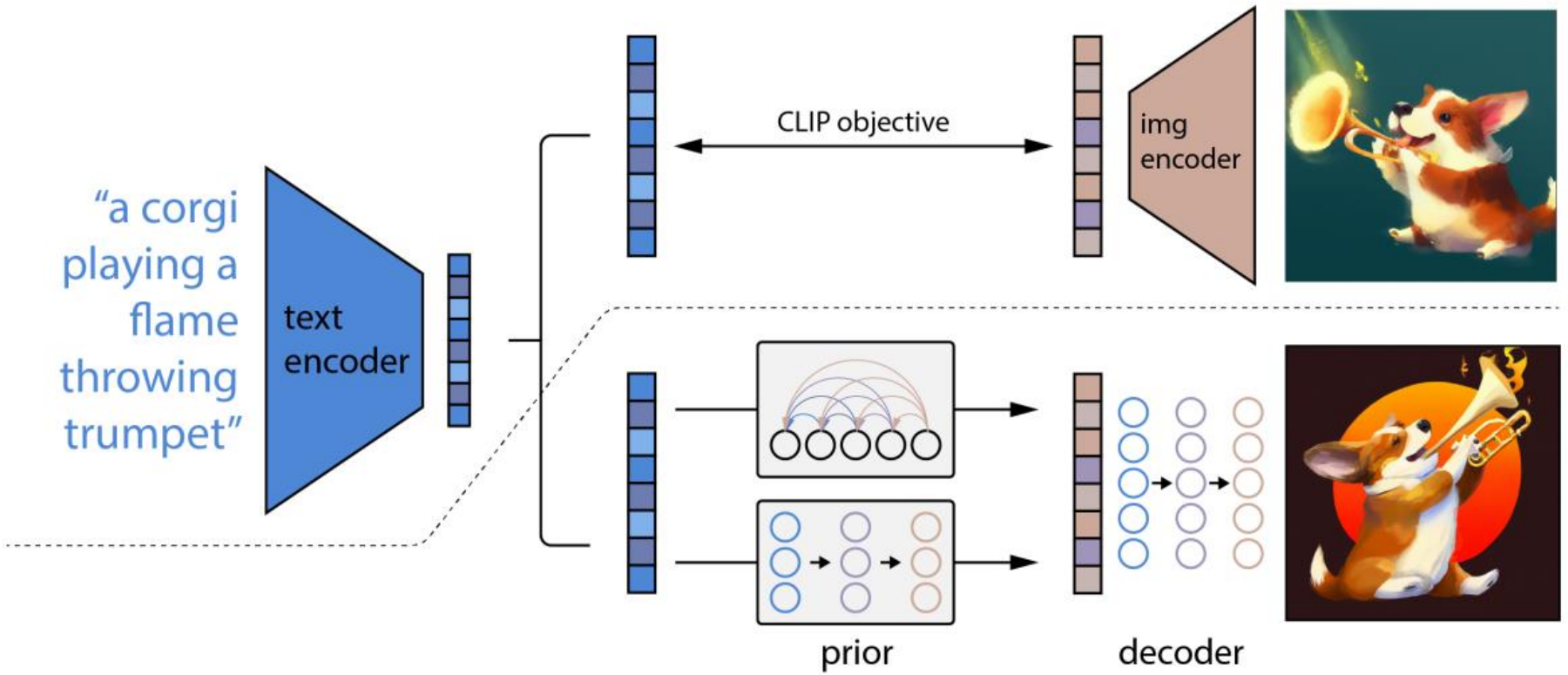
Task 7:
Video Prediction (V2V)

NÜWA predicts the future of an image.

Task 8:
Video Manipulation (TV2V)

NÜWA can even manipulate video futures with the control of language.

DALL-E2



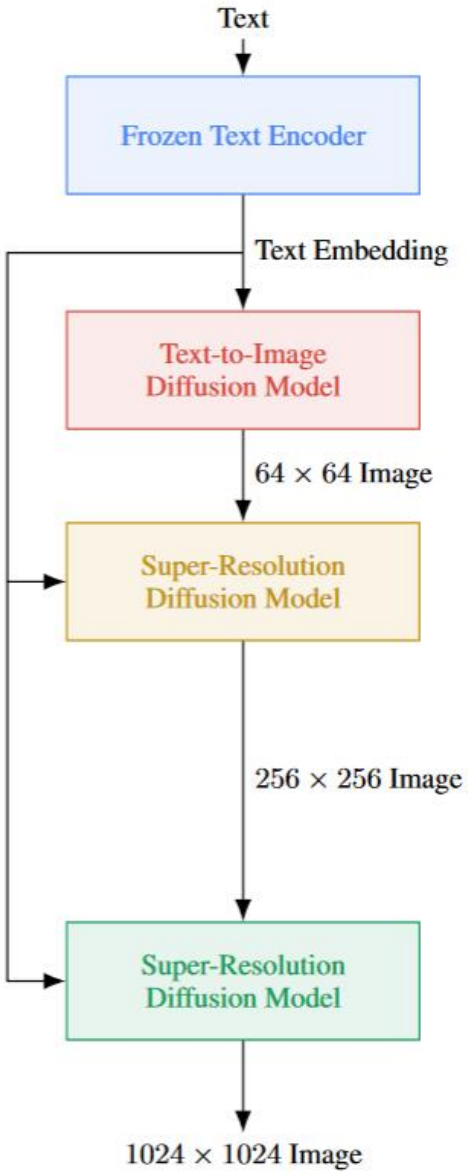
1. A prior that generates a CLIP image embedding given a text.

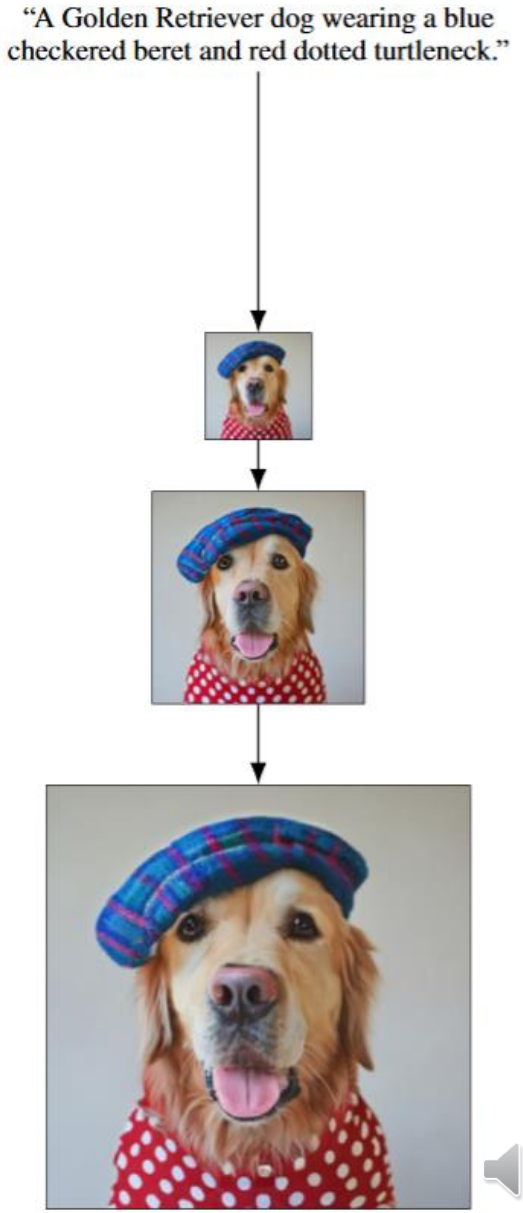
2. A decoder that generates an image conditioned on the image embedding.

DALL-E2

Imagen

Sprouts in the shape of text ‘Imagen’ coming out of a fairytale book.





Imagen

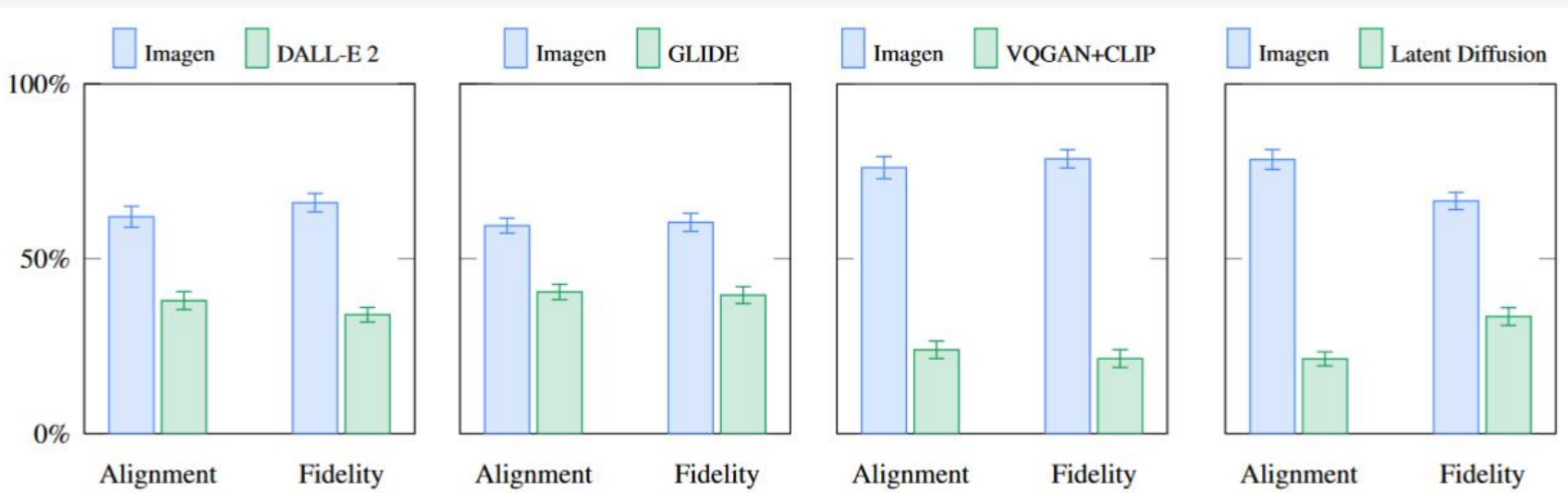
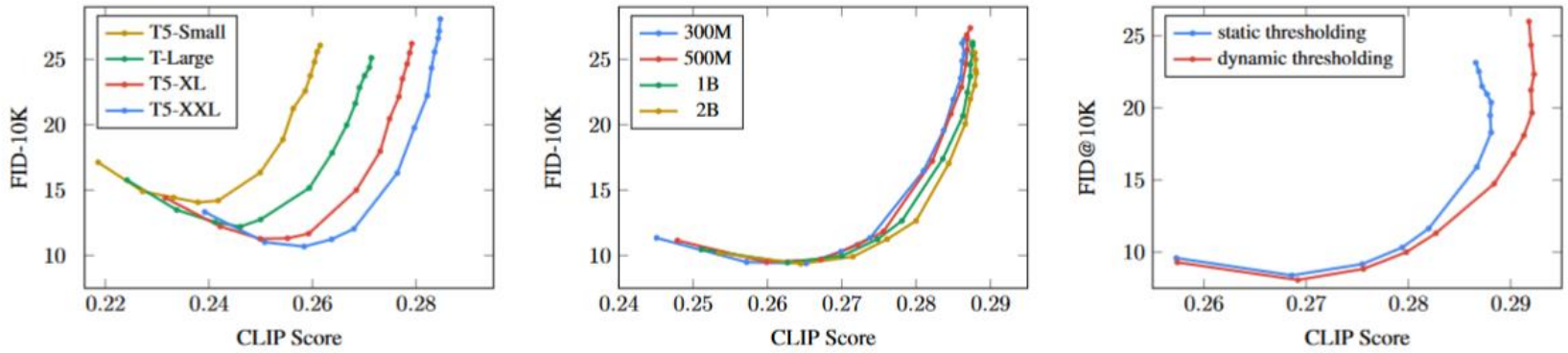


Figure 3: Comparison between Imagen and DALL-E 2 [54], GLIDE [41], VQ-GAN+CLIP [12] and Latent Diffusion [57] on DrawBench: User preference rates (with 95% confidence intervals) for image-text alignment and image fidelity.



(a) Impact of encoder size. (b) Impact of U-Net size. (c) Impact of thresholding.

Figure 4: Summary of some of the critical findings of Imagen with pareto curves sweeping over different guidance values. See Appendix D for more details.

Super-Resolution based Models

DALL-E

(256x256 images)

Jan 5, 2021

DALL-E 2

(1024x1024 images)

April 7, 2022

Imagen

(1024x1024 images)

May 17, 2022

Nov 5, 2021

NUWA

(256x256 **images and videos**)

June 2022

NUWA-Infinity

(**infinite-resolution images and videos**)

Why Infinite Resolution important?

64x64

256x256

1024x1024

Higher resolution implies not only more details, but also wider views.

Traditional Two-Stage Solution

(CogView, DALLE-2, Imagen)

Stage 1:

Coconut trees beside the sea.

Rendering

Stage 2:

Super-Resolution

Fix-sized, limited Resolution.
Supports images only.

Our End-to-End Solution

(NUWA-Infinity)

Coconut trees beside the sea.

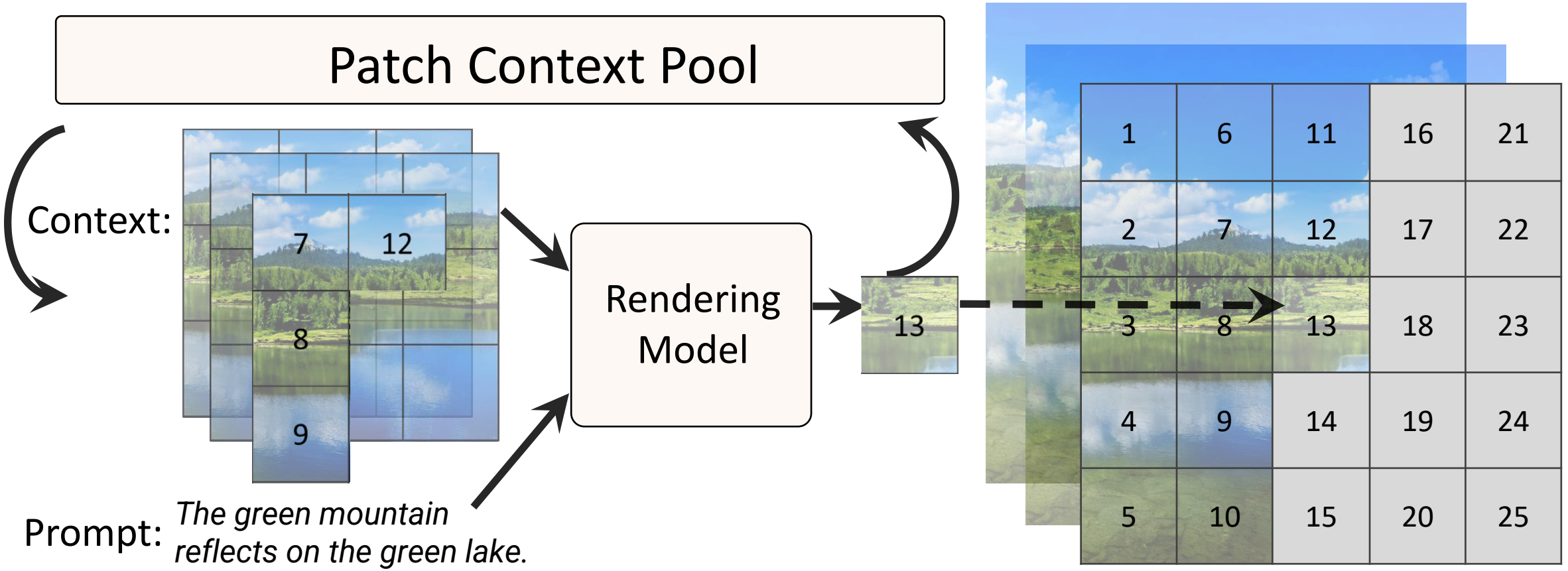
Context-aware Rendering

Patch Context Pool

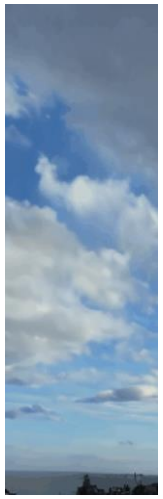
Patch Feedback

Arbitrary-sized, Infinite Resolution.
Supports both images and videos.

Our Solution: NUWA-Infinity



NUWA-Infinity Supported Tasks & Datasets



None

Unconditional Image Generation (UIG)

Image-Extension (IE)

Landscape (90K)
Design (2K)

Image-to-Video (I2V)

Landscape-V (120K)
FlintStones (20K)

White clouds floated in the sky.

Text-to-Image (T2I)

Text-to-Video (T2V)

PART 1: UIG Task

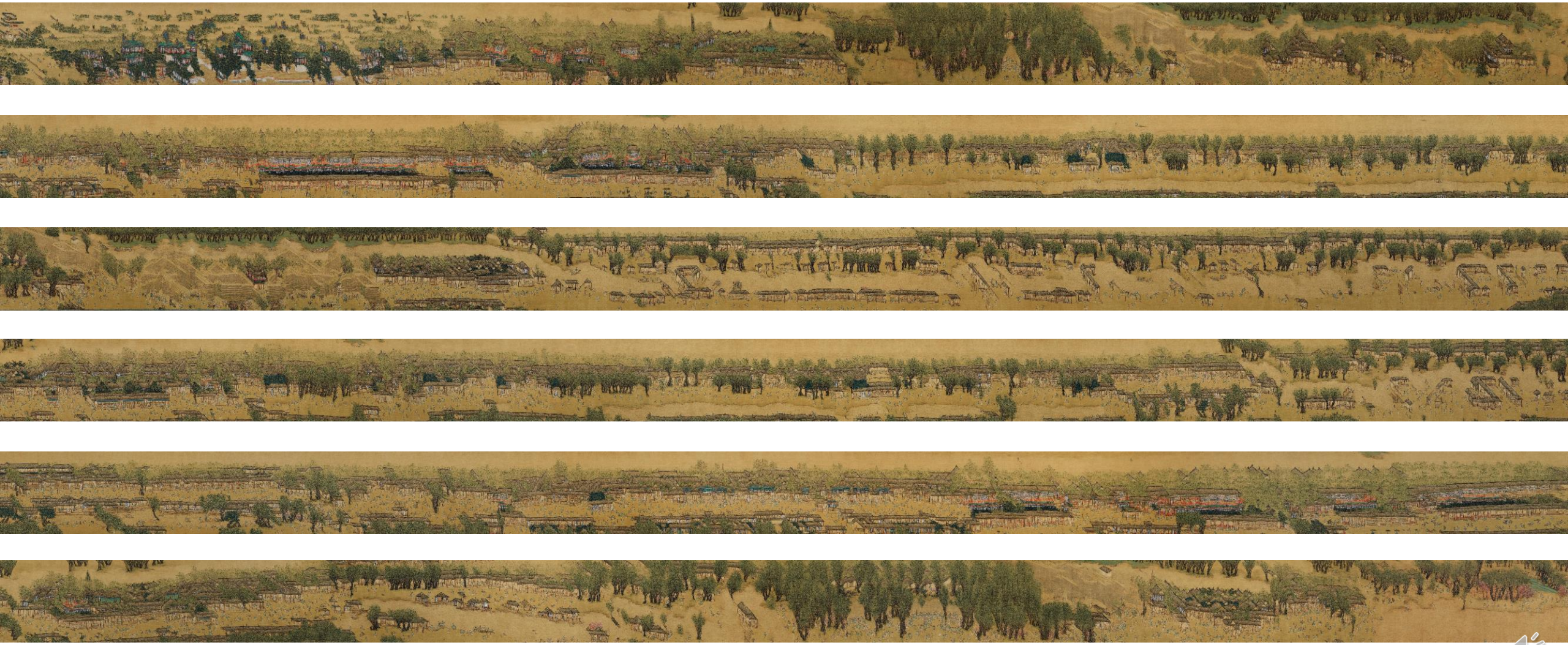
None

Unconditional Image Generation (UIG)

[1] Unconditional Image Generation Task on Landscape dataset

[2] Unconditional Image Generation Task on Design dataset

[3] Unconditional Image Generation Task on “Riverside Scene at Qingming Festival” (*One-shot Leaning from Scratch*)



2048 × 38912

[3] Unconditional Image Generation task on “Riverside Scene at Qingming Festival” (2048x6114)

PART 2: IE Task

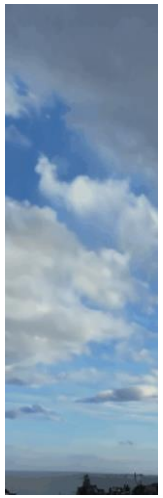
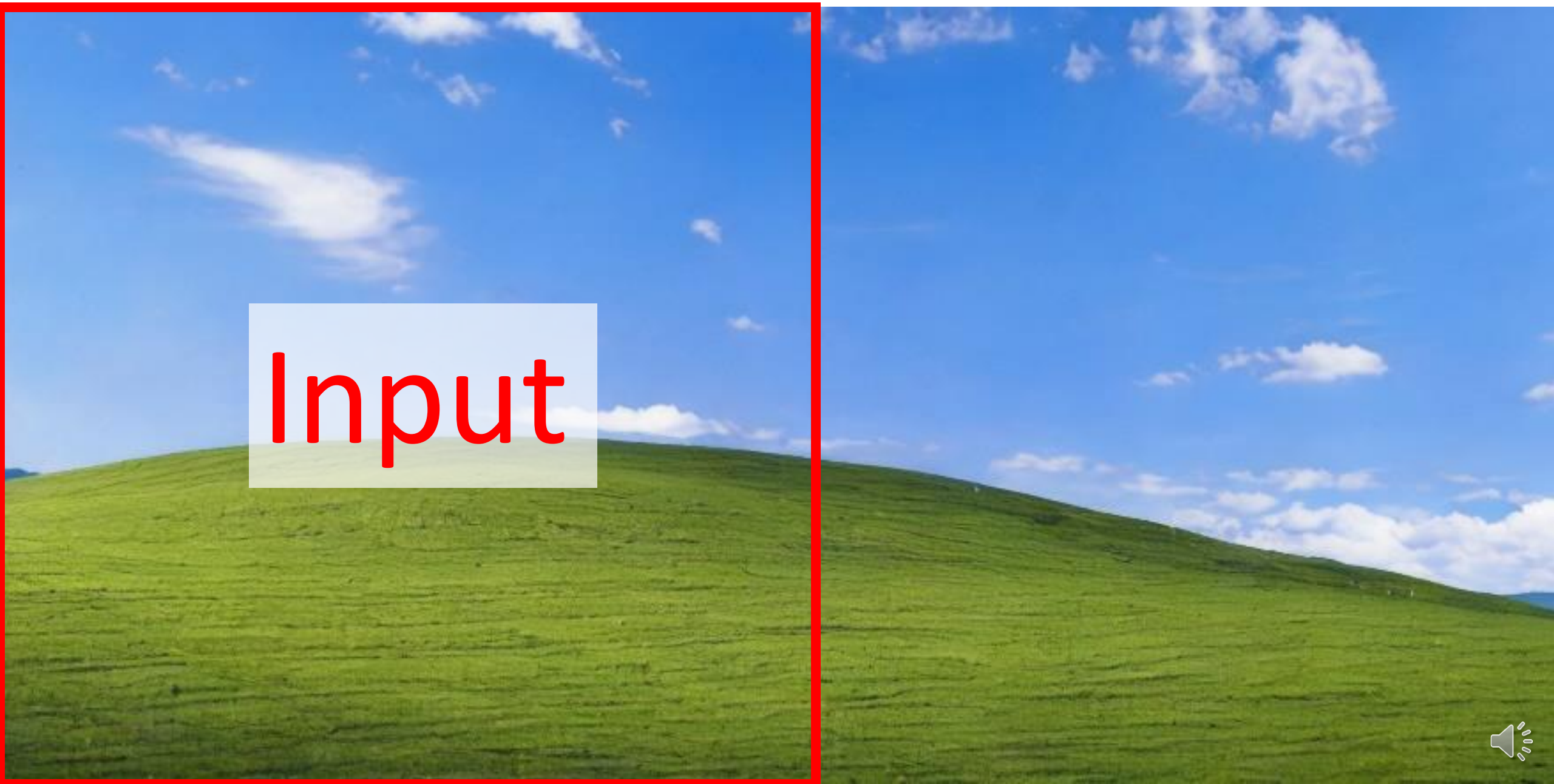


Image-Extension (IE)

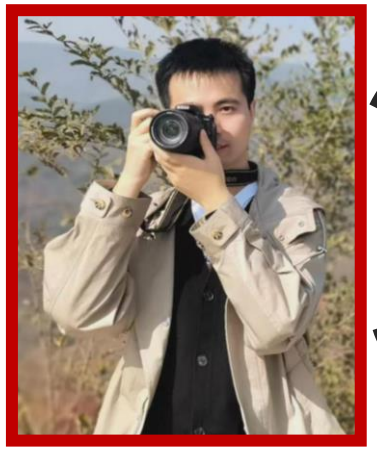
[1] Image-Extension Task on Landscape dataset

[1] Image-Extension Task on Windows Wallpaper (Zero-shot)



[1] Image-Extension Task on “The Starry Night” （Zero-Shot）

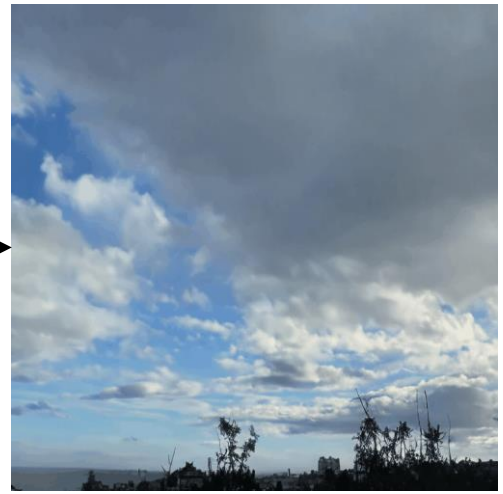
[1] Image-Extension Task on “The Gleaners” (Zero-Shot)



PART 3: T2I Task

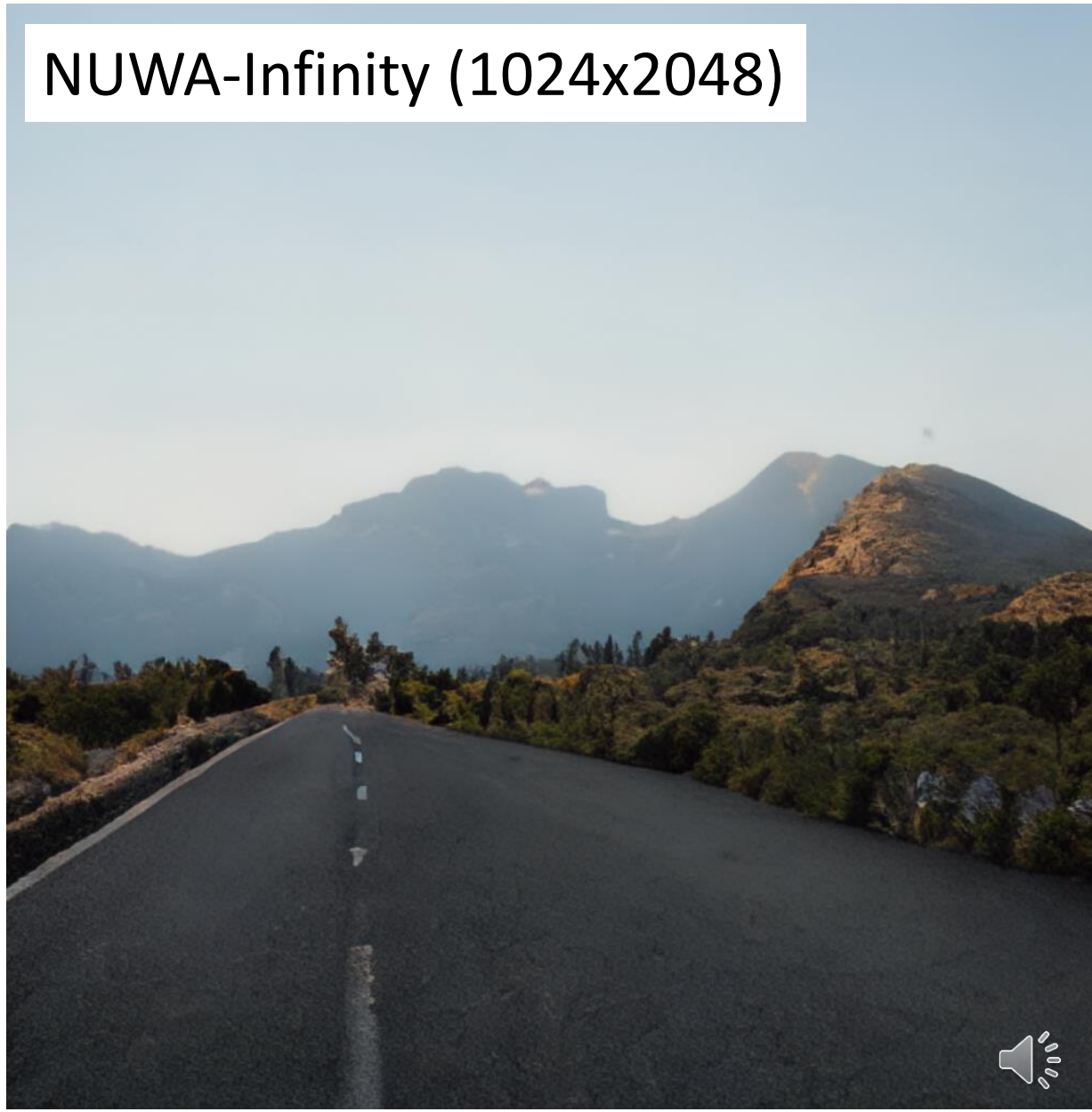
White clouds floated in the sky.

Text-to-Image (T2I)



[1] Text-to-Image Task on Landscape dataset

Input Text: a road that is going down a hill.



[1] Text-to-Image Task on Landscape dataset

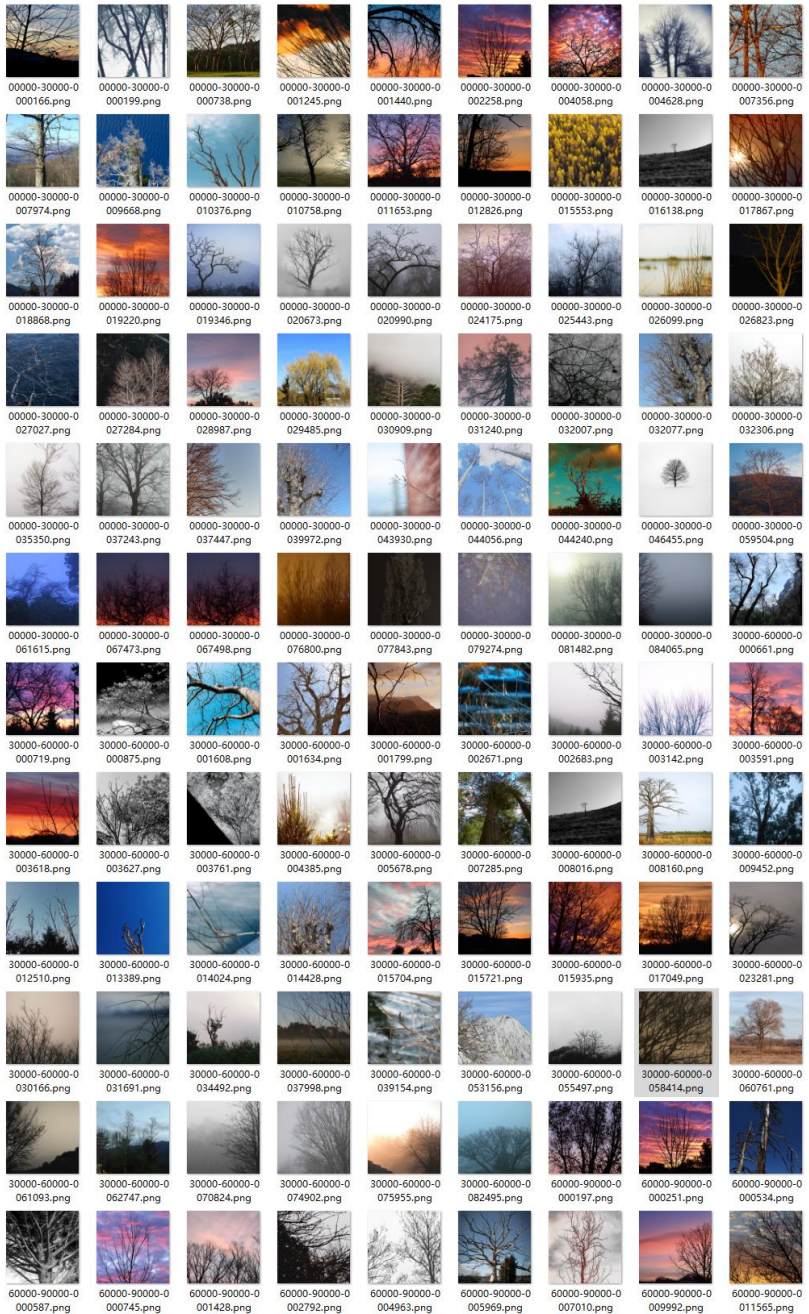
a path in a forest with tall trees.

a path through a forest with **fog** and trees.

a **snowy** forest with trees covered in snow.

a **snowy** forest with trees covered in snow.

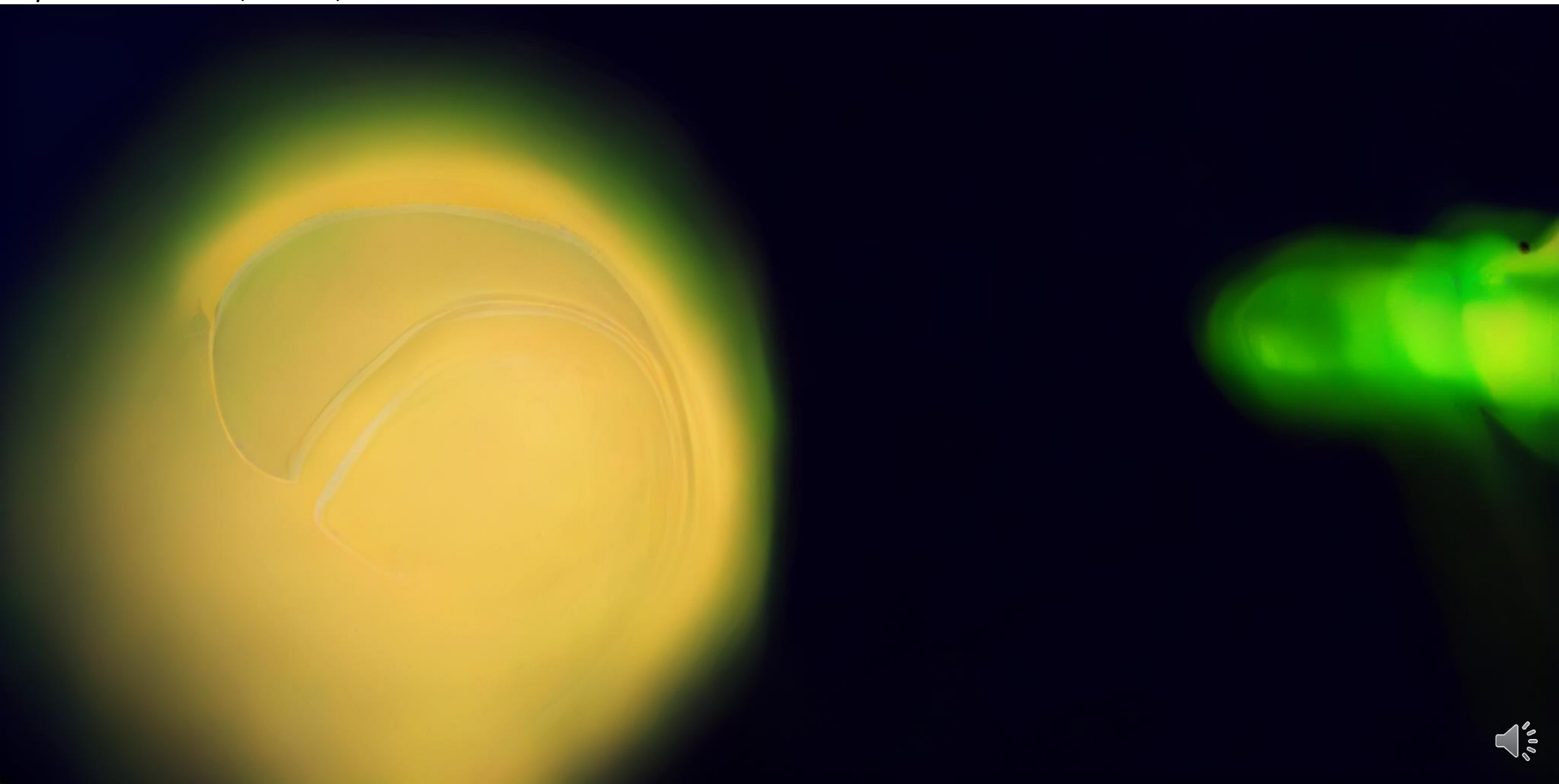
[1] Text-to-Image Task on Landscape dataset



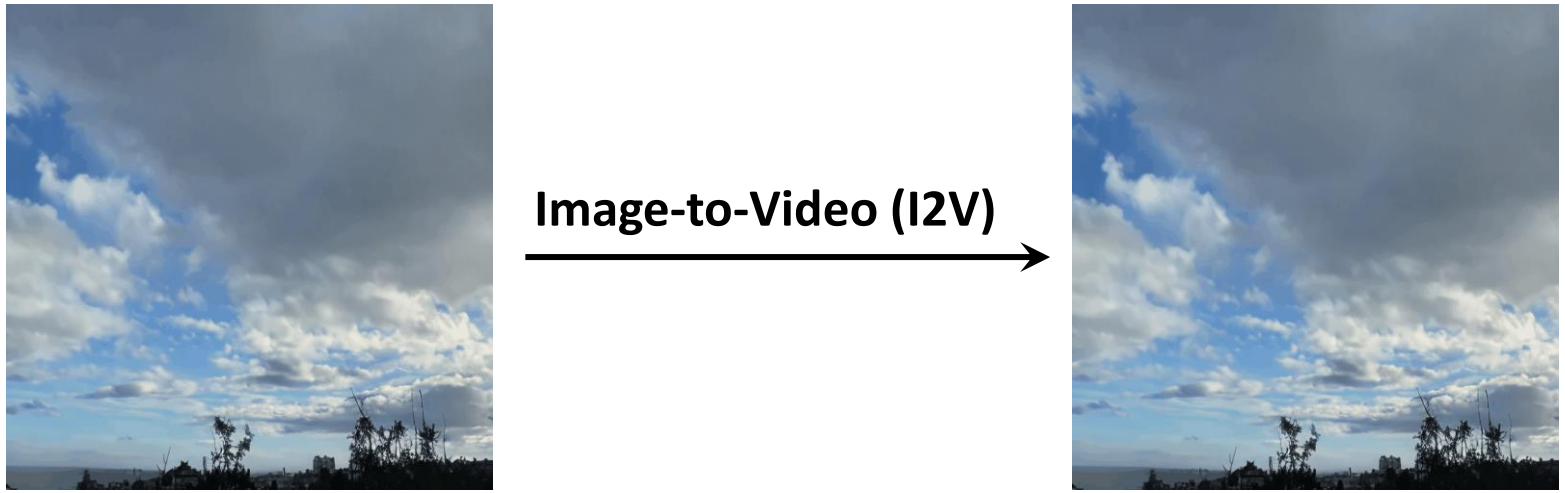
a tree with no leaves on it under a cloudy sky.

[2] Text-to-Image Task on Design dataset

Input Color: Black, Green, Yellow.



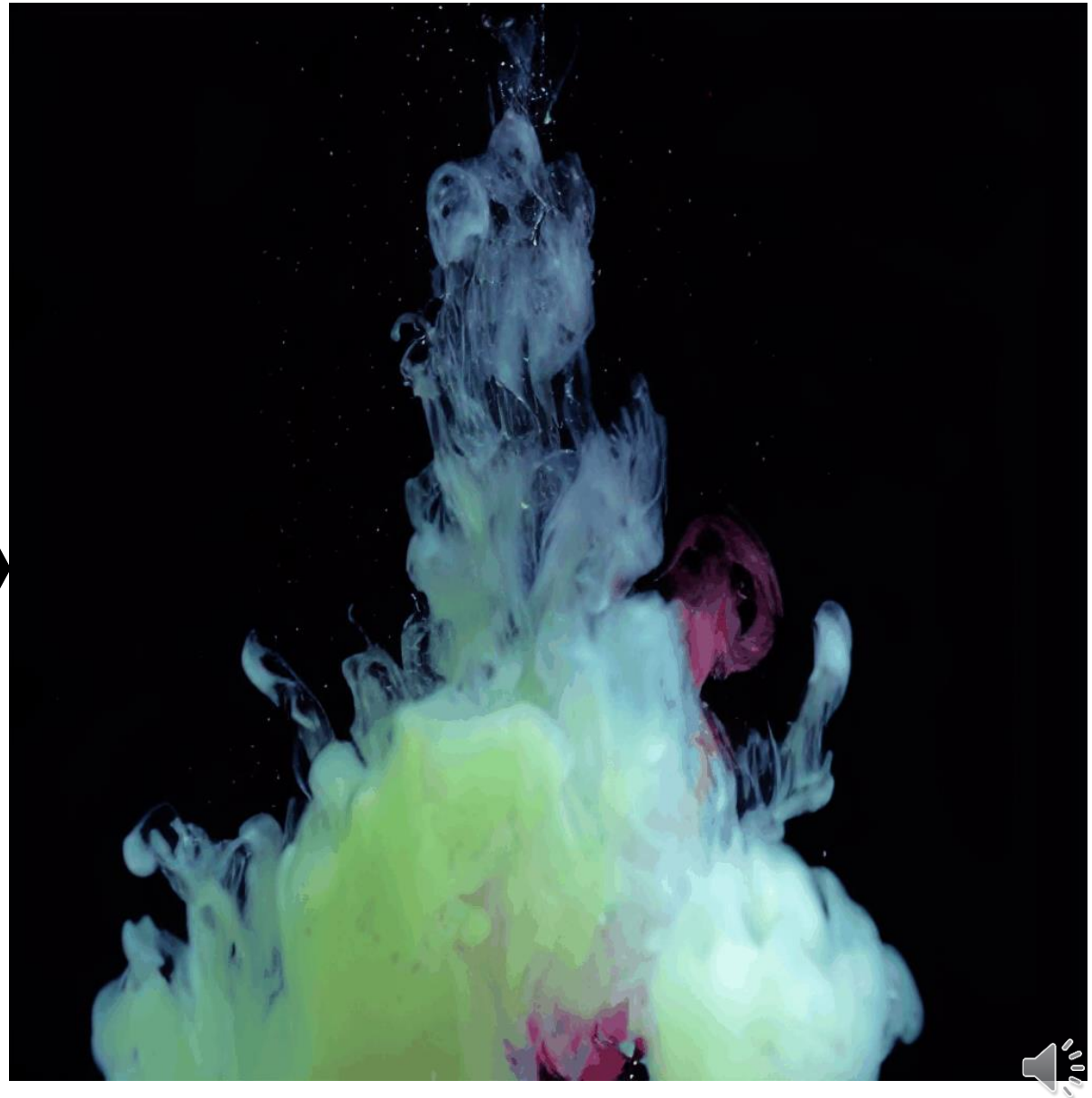
PART 4: I2V Task



[1] Image-to-Video Task on Pexels dataset

[1] Image-to-Video Task on Pexels dataset

[2] Image-to-Video Task on Design dataset



PART 5: T2V Task

White clouds floated in the sky.

Text-to-Video (T2V) →

Cartoon Generation on Flintstone dataset

Input Story:

1) Fred wants to find treasure.

2) He drives a car. He passes some trees.

3) He stops at a cave. He goes inside.

4) Fred finds the treasure.

5) He is very happy.

Output Cartoon:

Fred wants to find treasure.

He drives a car. He passes some trees.

He stops at a cave. He goes inside.

Fred finds the treasure.

He is very happy.

From Idea to Design

IDEA

Ideal design
(Object + Style + Color)

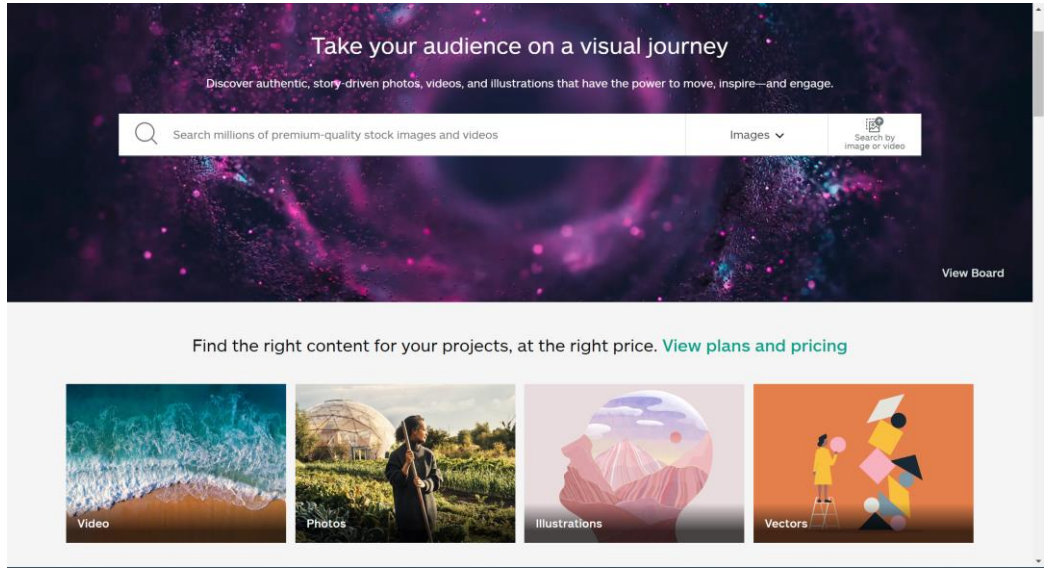
Input | AI | Output

Case 1

“

I want to design a concept map in Office PowerPoint.. There is a **blue 3D squares** in the background. The overall style should have a sense of three-dimensional composition, and the space should reflect the **sense of depth**.

”



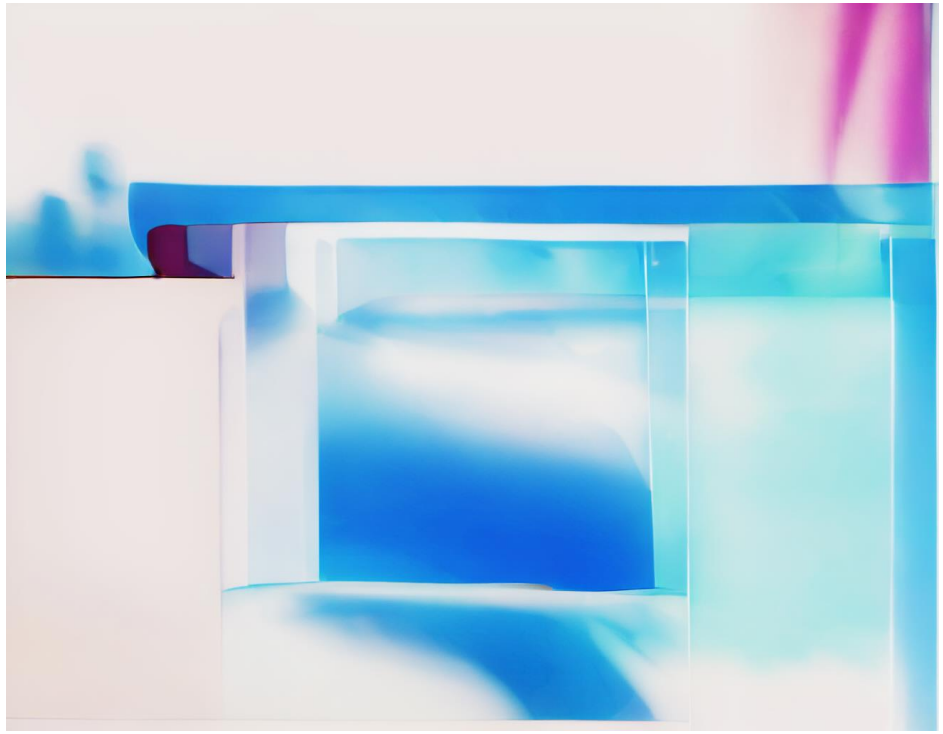
Current design: Search or Sketch or Software

Case 1

blue

3D squares

sense of depth

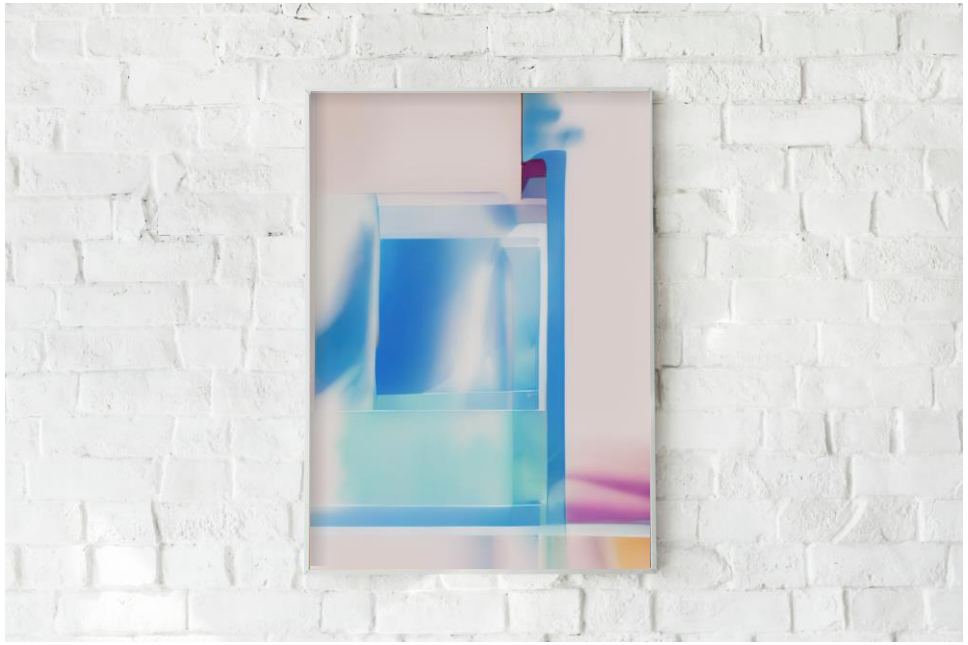


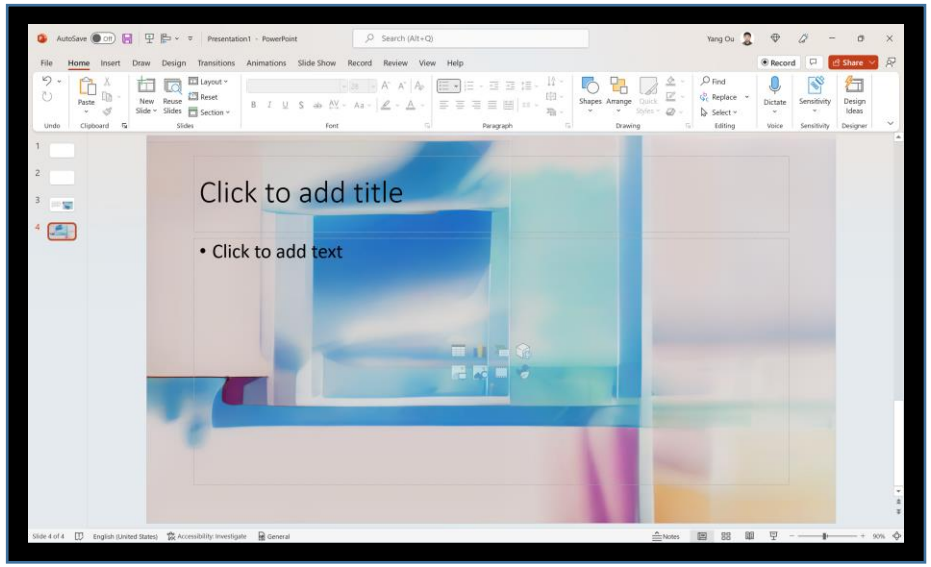
Text → Image Generation → Expand

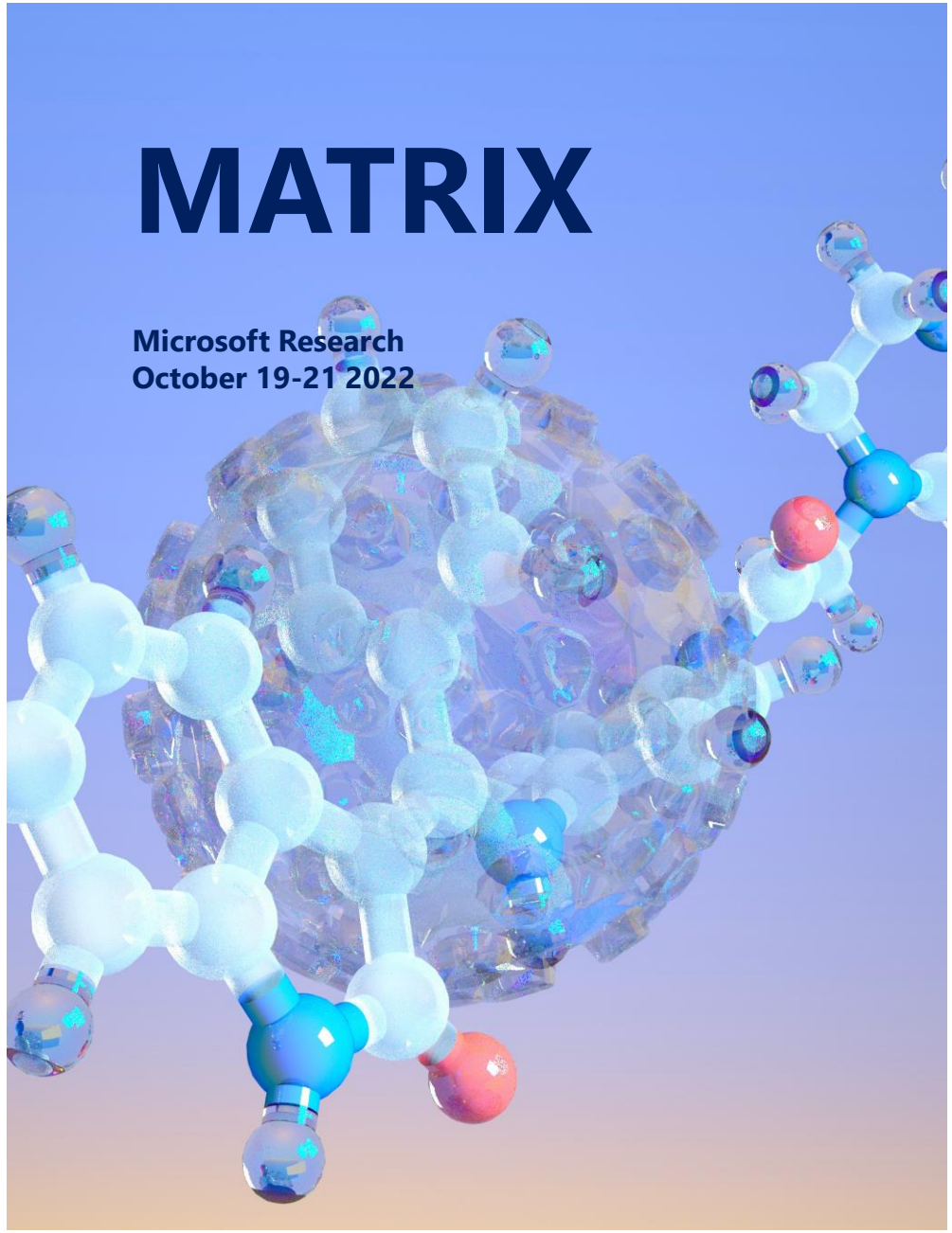
Stage1: Text-to-Image

Stage2: Image Extension

Case 1

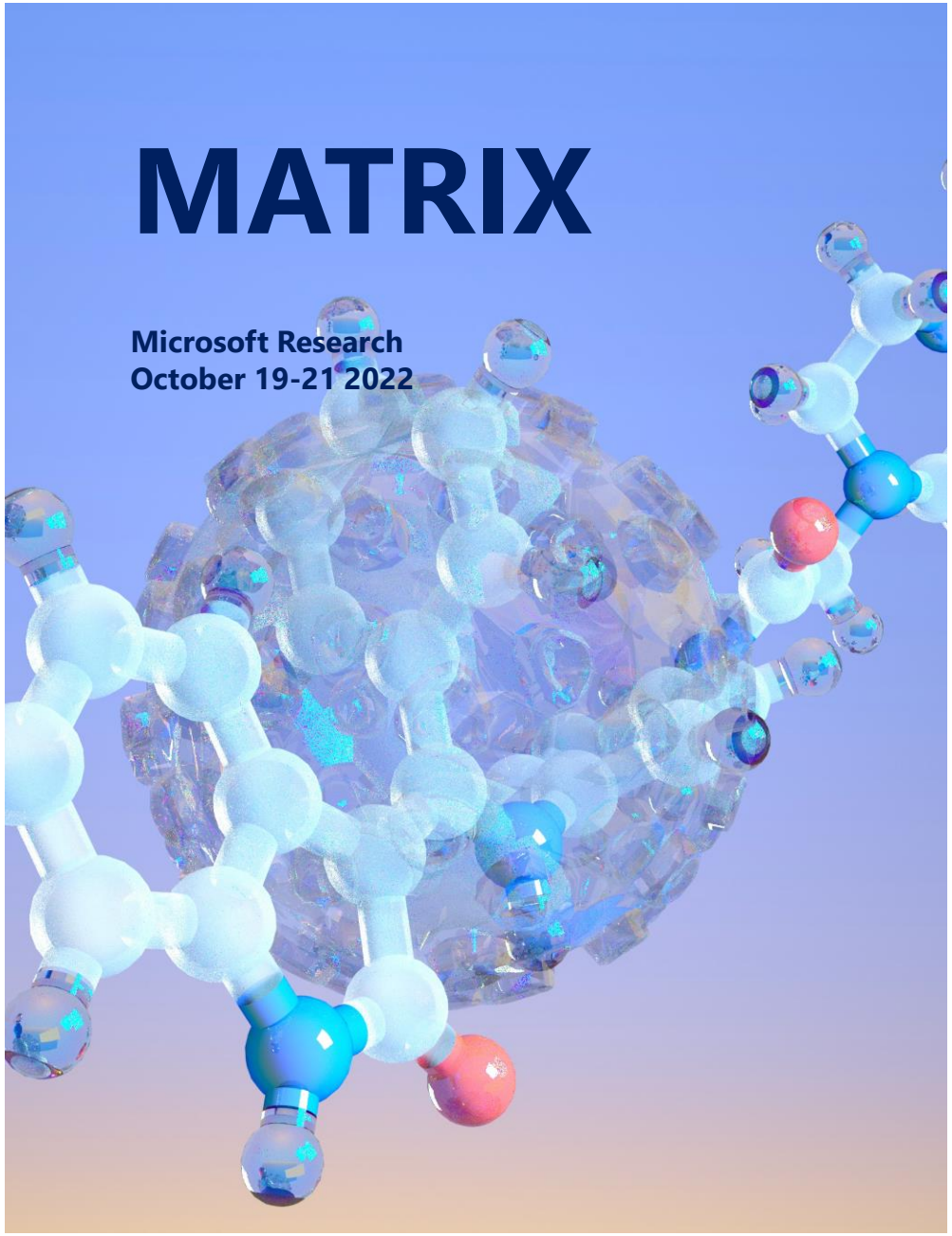


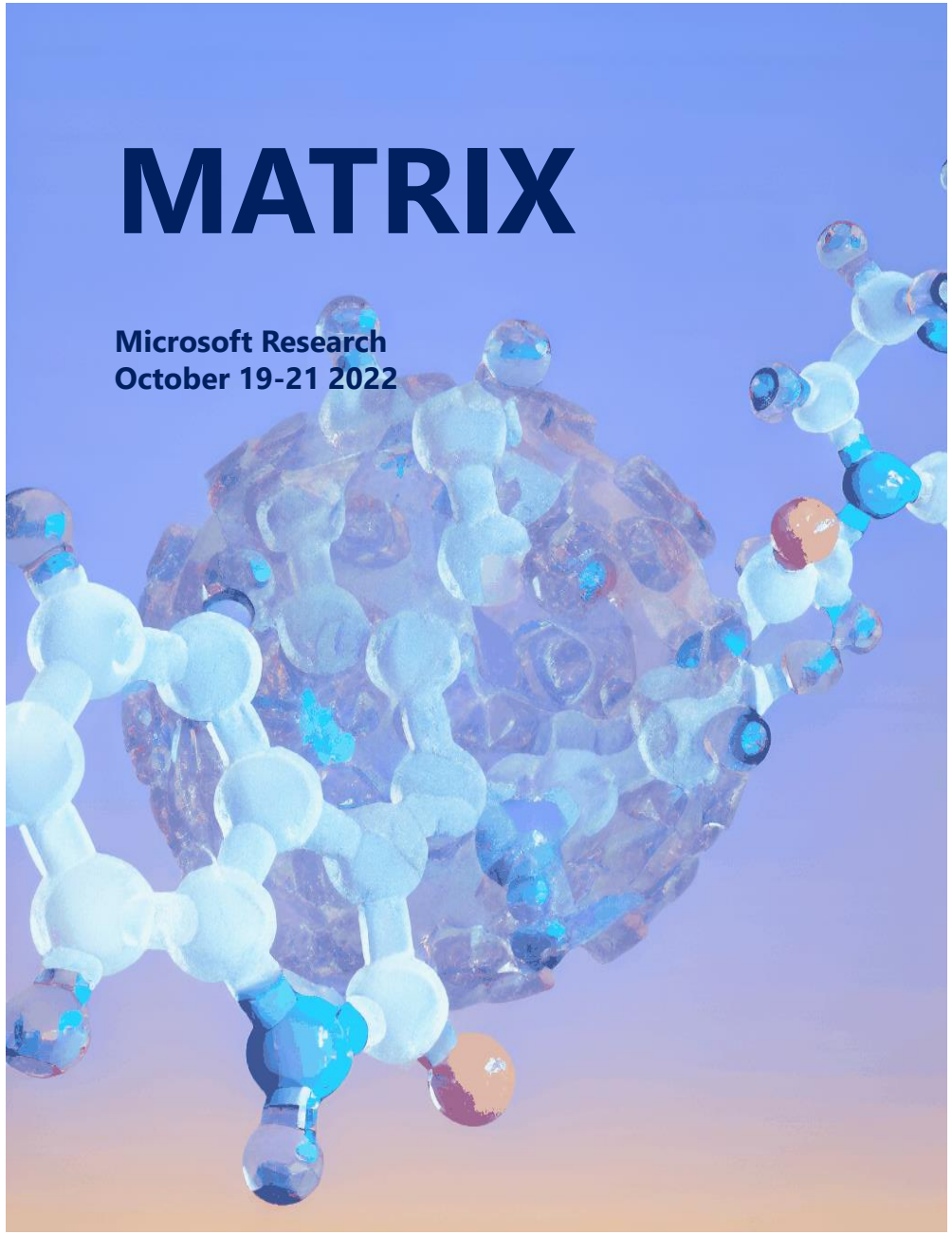




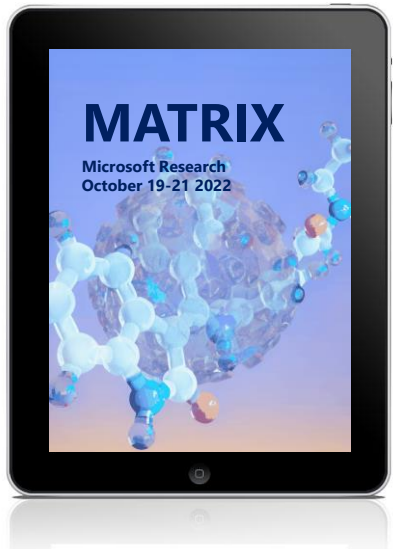
Case 2

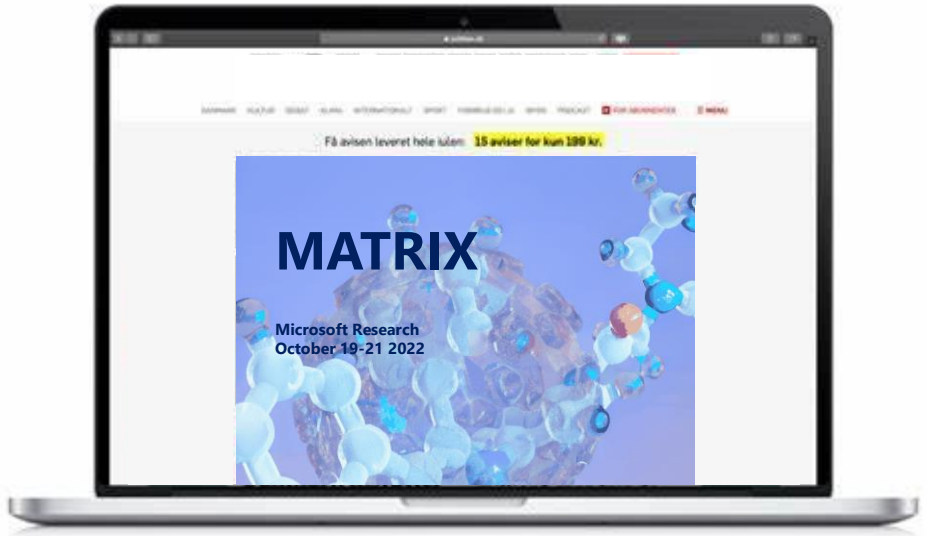
“I designed a cover for the magazine of MSRA Matrix, I want to animate it.”


MATRIX
Microsoft Research
October 19-21 2022


MATRIX
Microsoft Research
October 19-21 2022


MATRIX
Microsoft Research
October 19-21 2022


MATRIX
Microsoft Research
October 19-21 2022

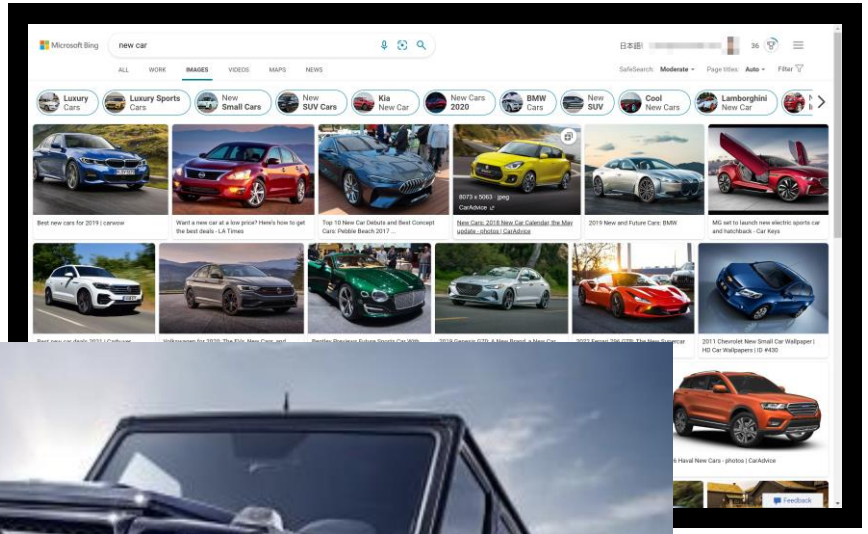

Få avisen leveret hele ulen: 15 aviser for kun 199 kr.
MATRIX
Microsoft Research
October 19-21 2022

Case 3

“

I want to put the car ads on Bing, but need to make it adaptive on web pages.

”



Case 3

Case 4

“

I am in a Microsoft Teams Meeting.
I want to animate my Teams Background wallpaper.

”

Case 4

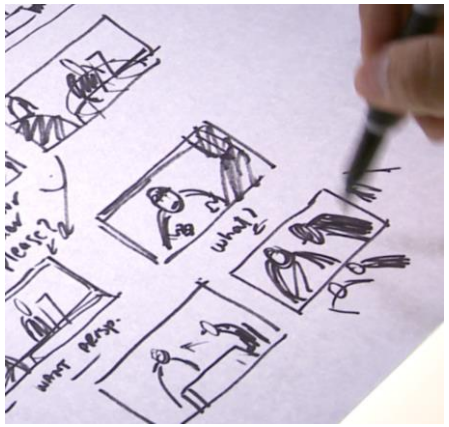

Weekly planning
00:22:06
Leave


Weekly planning
00:22:06
Leave

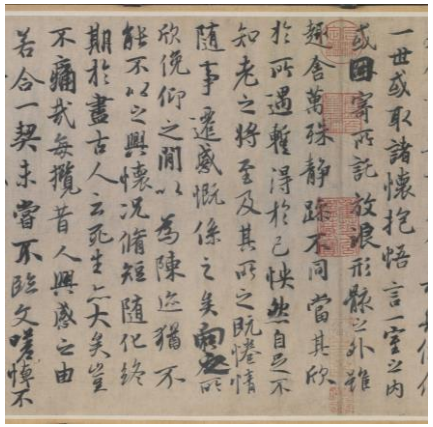
Other Scenarios in future

Restoration

Gaming



Story board



Calligraphy

Disability