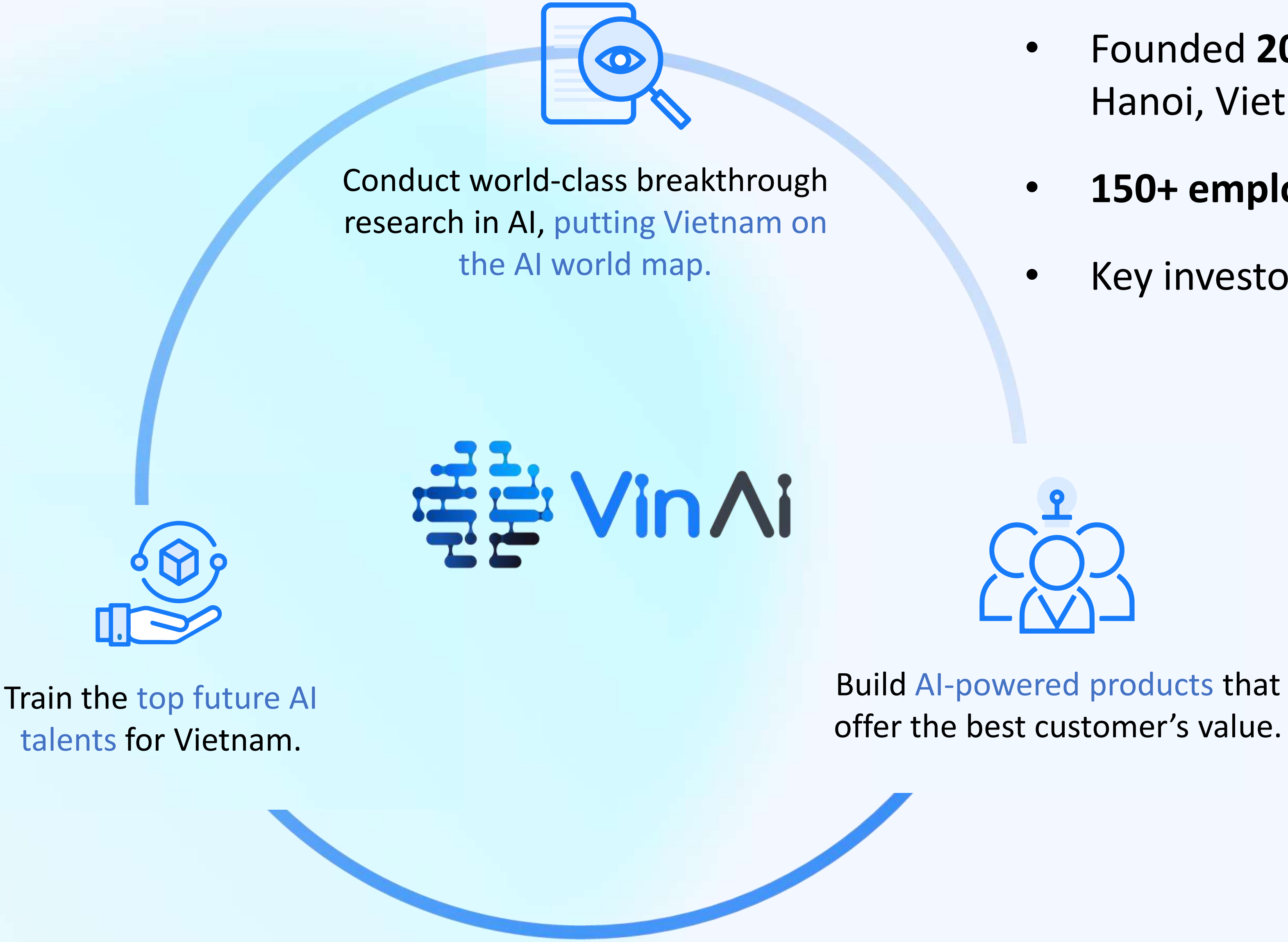
Recent Advances in Text-to-X Generation

SPEAKERS:

Anh Tran – Research Scientists at VinAI

Our Story

Three Missions



Conduct world-class breakthrough research in AI, putting Vietnam on the AI world map.

Train the top future AI talents for Vietnam.

Build AI-powered products that offer the best customer’s value.

- Founded **2019**, headquartered in Hanoi, Vietnam & USA, Australia.
- **150+ employees total**
- Key investor is **Vingroup**

Putting Vietnam on the Global AI Map

2022

104 top tier papers

2021

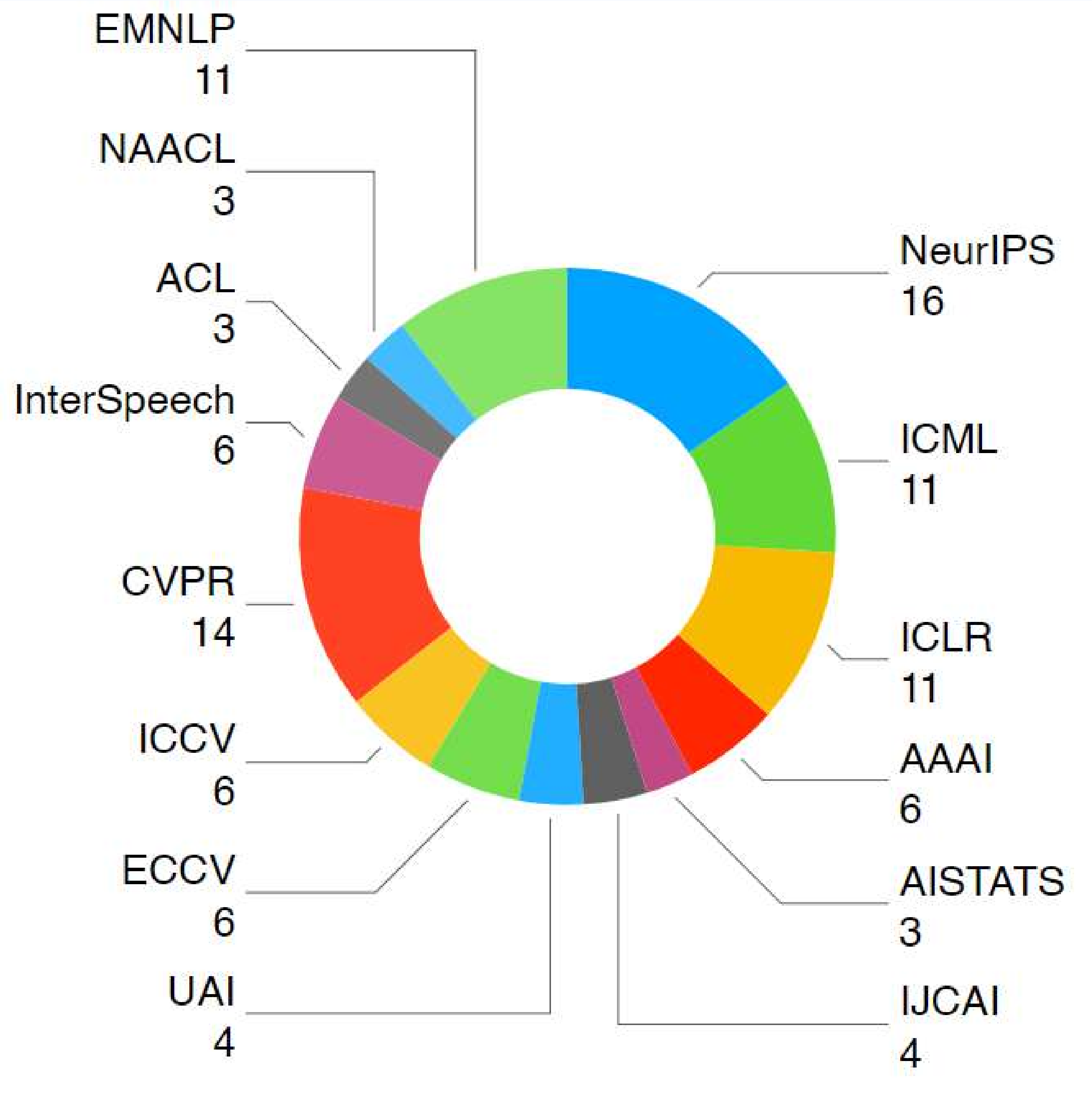
68 top tier papers

2020

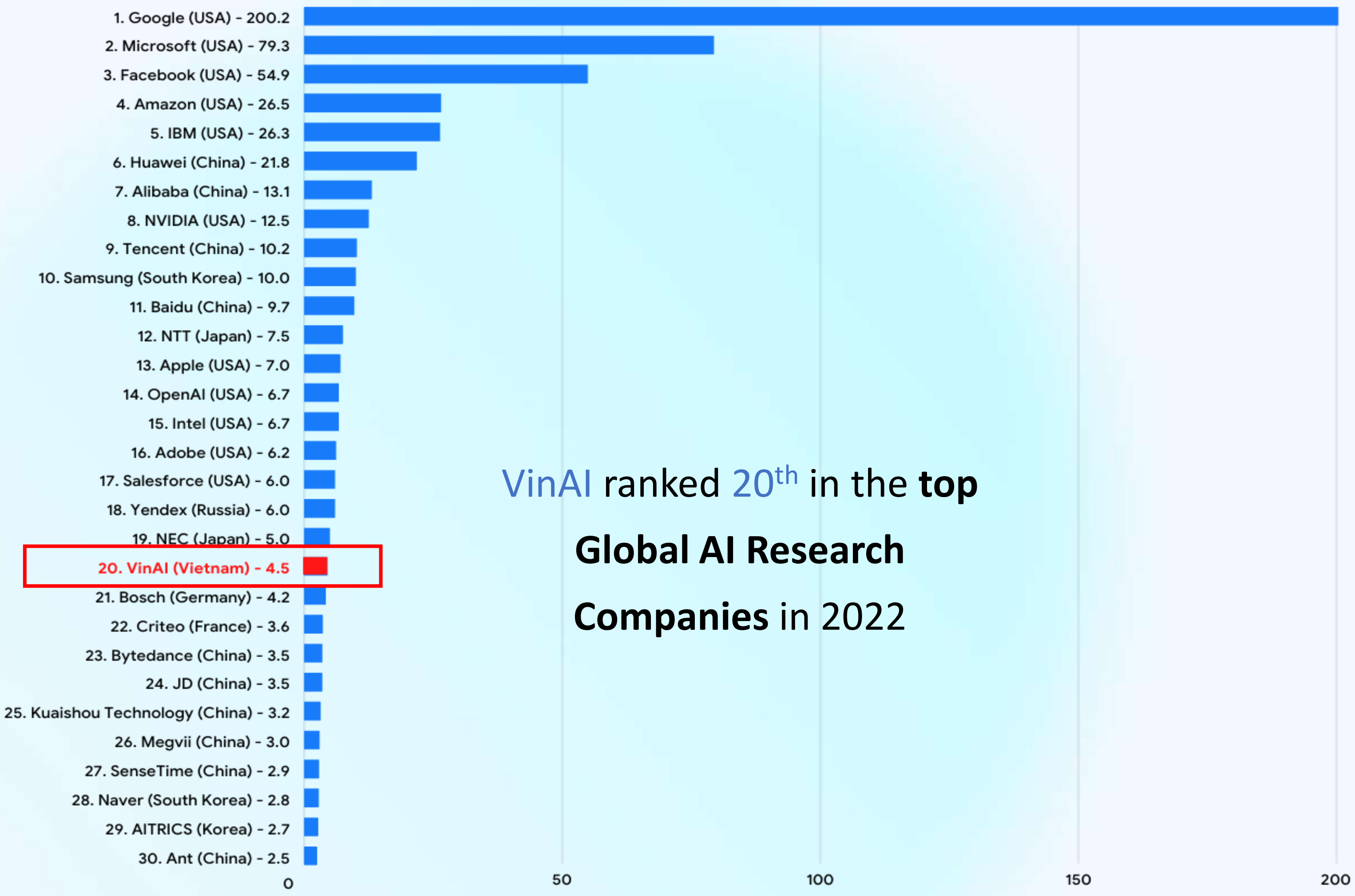
23 top tier papers

2019

6 top tier papers



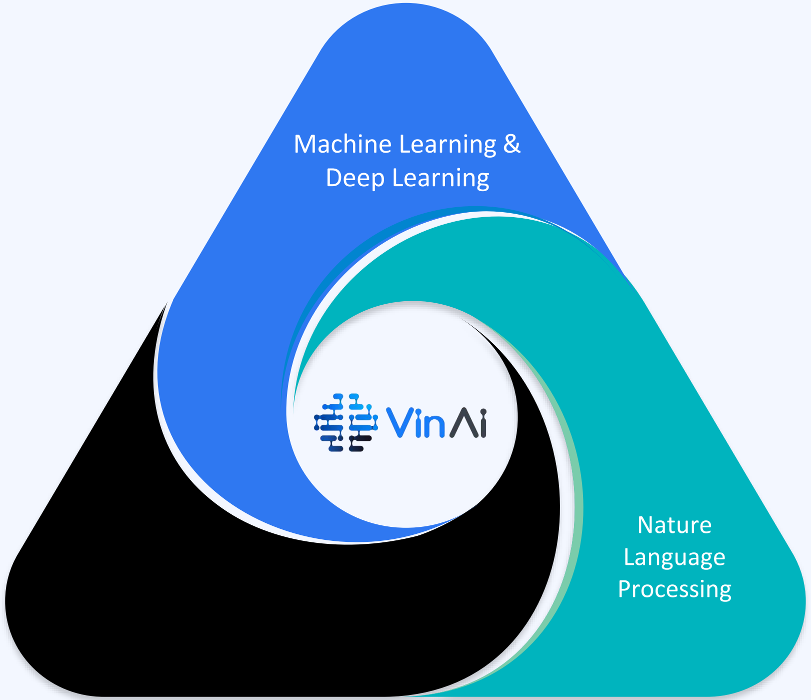
Our World Class AI Research



VinAI ranked 20th in the **top Global AI Research Companies** in 2022

Source: Thundermark Capital https://thundermark.medium.com/ai-research-rankings-2022-sputnik-moment-for-china-64b693386a4

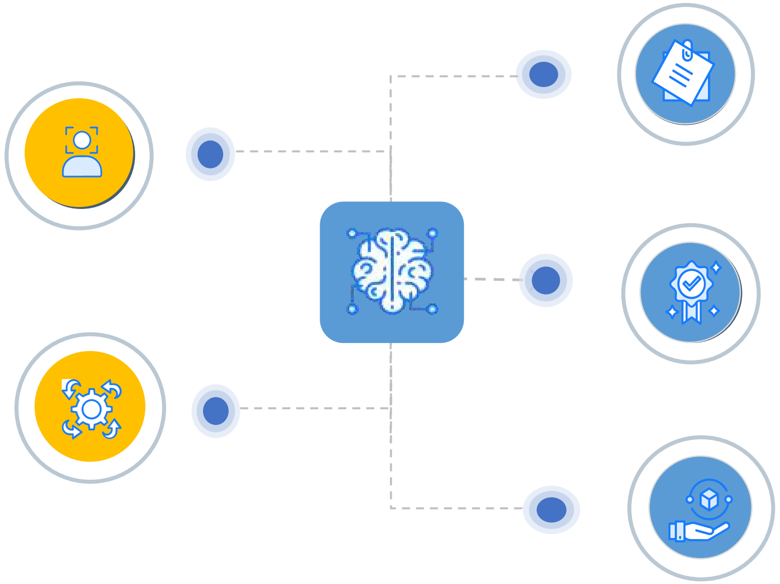
3 Research Pillars



Home grown talents in VN

Talent - AI Residency Program

The First and Most Prestigious AI Residency Program in Vietnam



80 Residents

World-class research training & Mentorship

43 accepted Top-tiers papers

66 Ph.D. Scholarships (from the world's TOP 20 CS universities)

45 Filed Patents

Our Current Global Alumni Network

University of Maryland

Carnegie Mellon University, Pittsburgh Pennsylvania 1900

Georgia Institute of Technology

Johns Hopkins University

Texas, The University of Texas at Austin

UCLA, University of California, Los Angeles

University of California, Irvine

University of Massachusetts Amherst

Wisconsin, University of Wisconsin-Madison

Trinity College Dublin, The University of Dublin

Cornell University

The University of Arizona

The University of Edinburgh

University of Rochester

NUS, National University of Singapore

UC San Diego

George Mason University

University of Stuttgart, Germany

Oregon State University

جامعة محمد بن زايد للذكاء الاصطناعي
Mohamed bin Zayed University of Artificial Intelligence

Monash University

The University of Iowa

Speaker

Anh Tran

- Research Scientist at VinAI Research
- Research field: Computer Vision
- Former Amazon
- PhD degree from University of Southern California (USA)
- VEF 2012
- 1st prize at Vietnam Talent 2010

Website : https://sites.google.com/site/anhttranusc/
Google Scholar: https://scholar.google.com/citations?user=FYZ5ODQAAAAJ

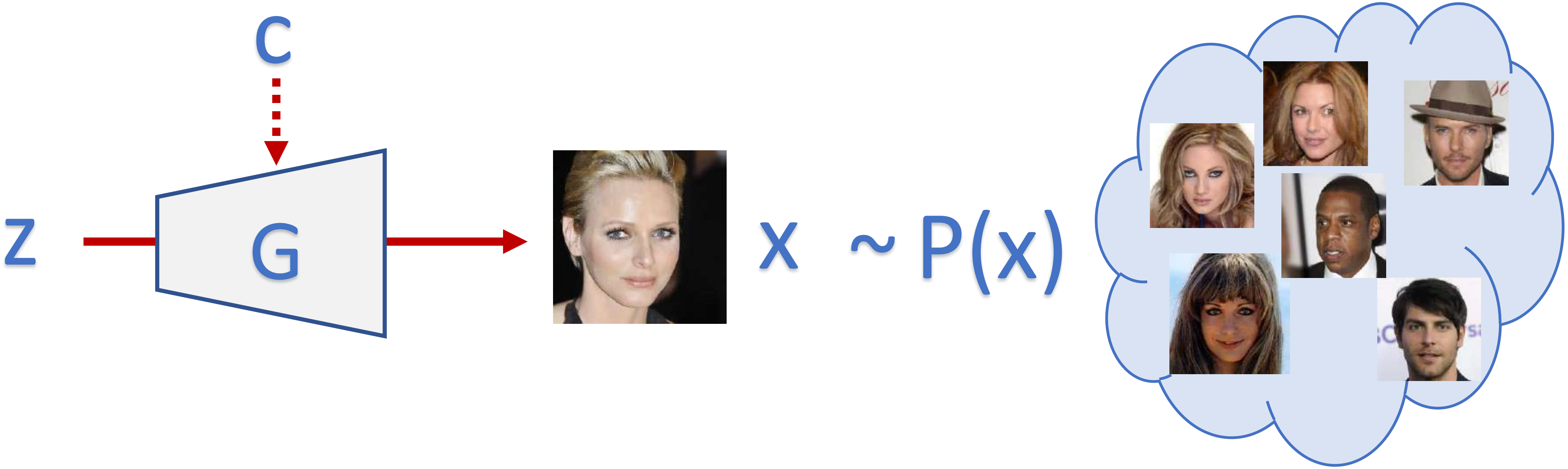
Outline

1. Diffusion Models

2. Text-to-image models

3. Other text-to-X models

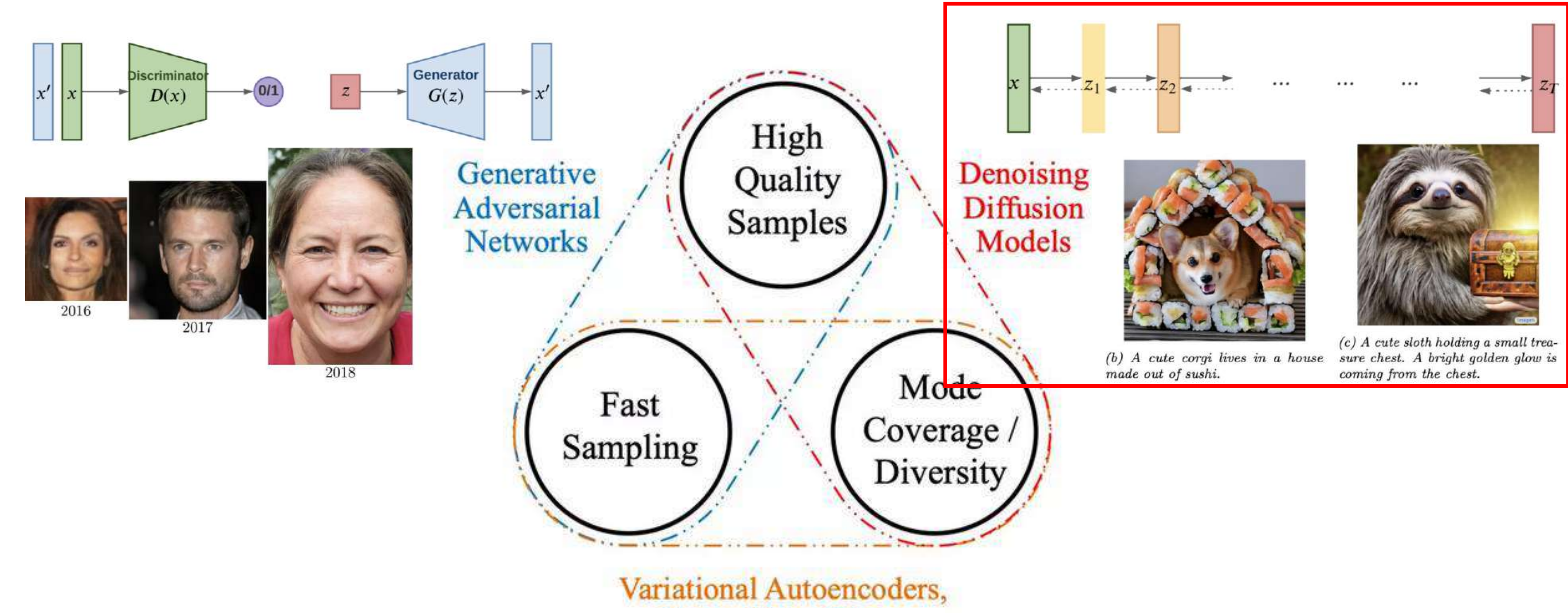
Image Generation



Map an input $z \sim N(0, I)$ to an output image x in the desired domain

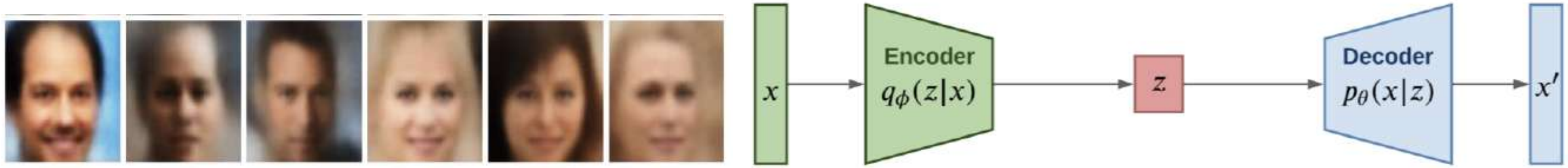
- **Unconditional generation:** $x = G(z)$
- **Conditional generation**: additional target attributes **c:** $x = G(z, c)$

Typical Approaches



(b) A cute corgi lives in a house made out of sushi.

(c) A cute sloth holding a small treasure chest. A bright golden glow is coming from the chest.



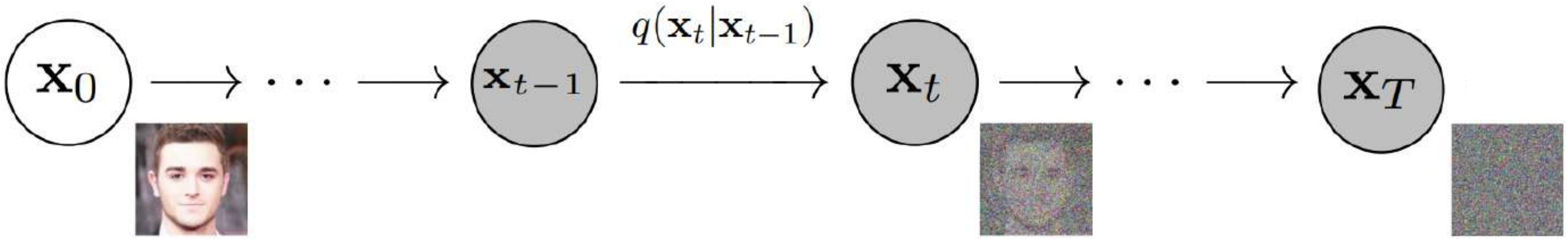
Others:

- Autoregressive models (ARM)
- Energy-based models (EBM)
- Normalizing flow models

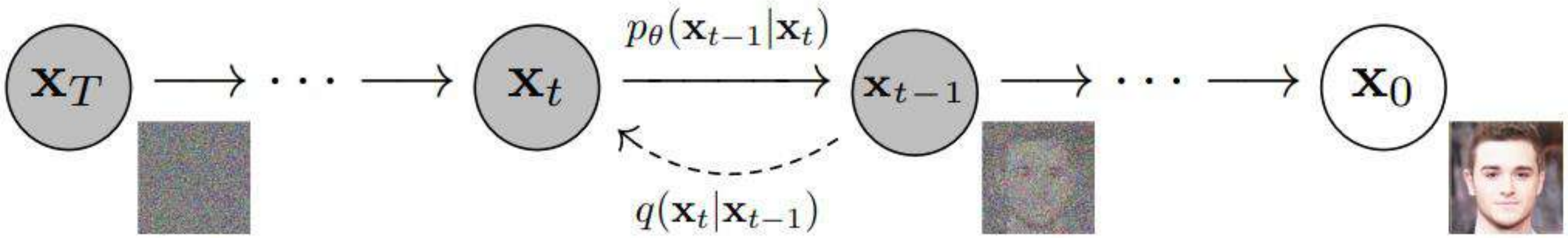
1. Diffusion Models

High-level Ideas

- Noise adding process (forwards diffusion)



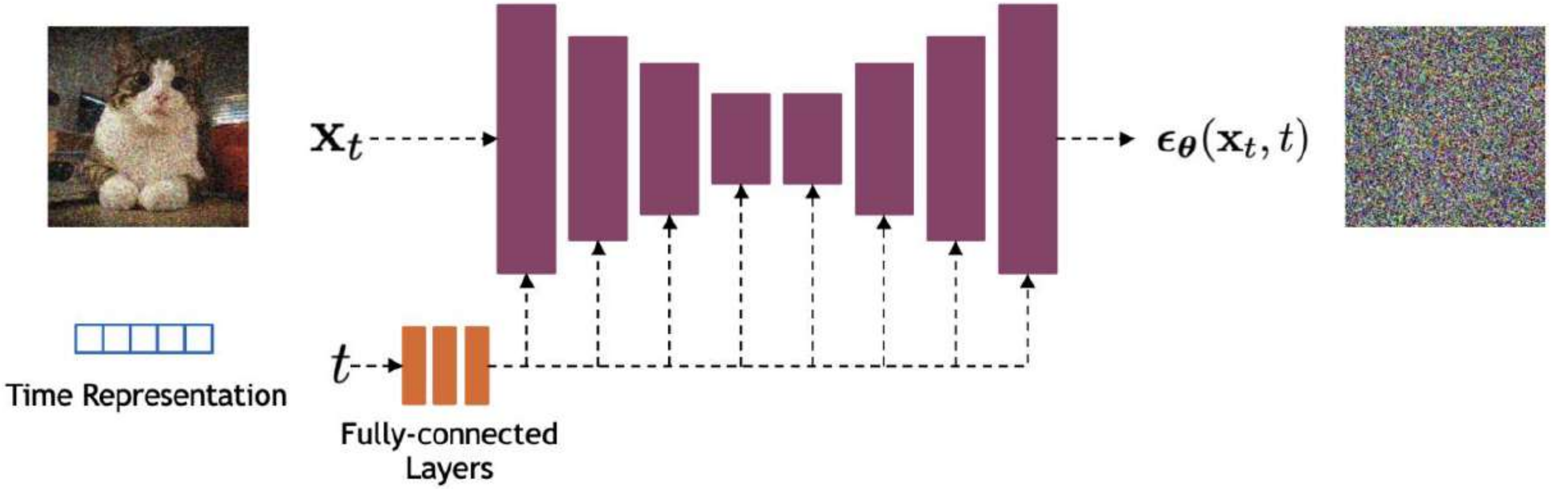
- Denoising process (reverse diffusion)
  - Can generate image from random noise!!!


$\mathbf{x}_T$
$\mathbf{x}_t$
$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$
$q(\mathbf{x}_t|\mathbf{x}_{t-1})$
$\mathbf{x}_{t-1}$
$\mathbf{x}_0$

J. Ho, A. Jain, and P. Abbeel. "Denoising diffusion probabilistic models". *In NeurIPS 2020, 33*, 6840-6851.

Details

- Noise adding steps for training: $x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon$
  - In training, we want to predict the noise $\epsilon \sim N(0, I)$ given x_t using a UNet:

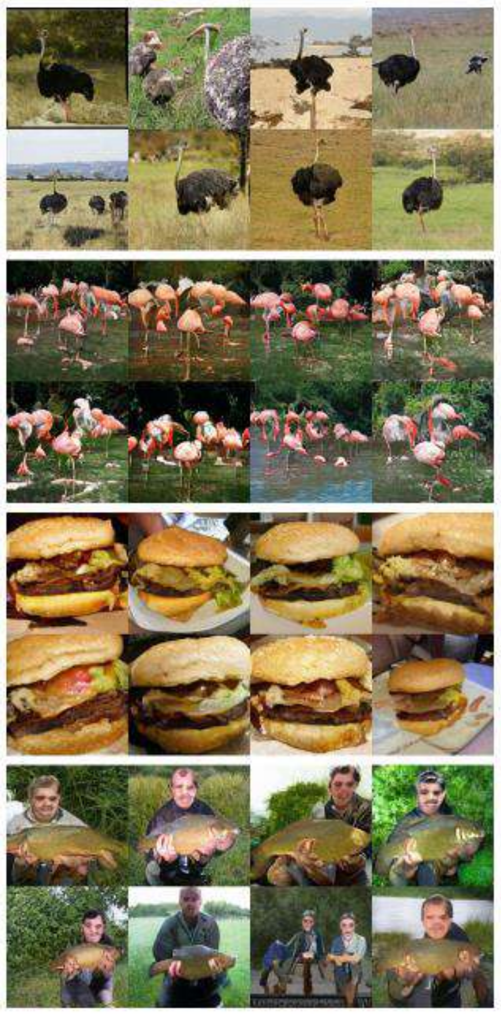


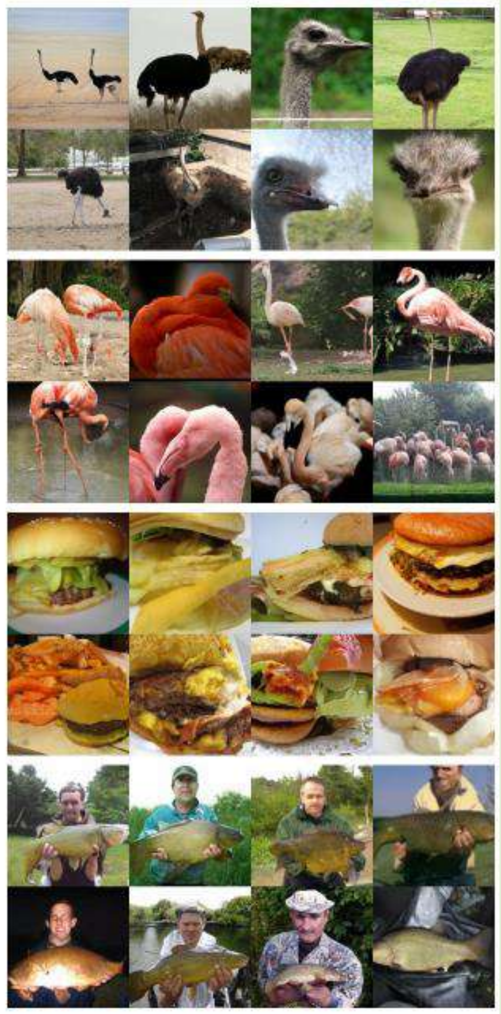
- Denoising strategy (sampling): $x_{t-1} = \frac{1}{\sqrt{\bar{\alpha}_t}} \left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta(x_t, t) \right) + \sigma_t \epsilon_t$

K. Kreis, R. Gao, and A. Vahdat. Denoising diffusion-based generative modeling: foundations and applications. CVPR Tutorial. 2022.

Image Quality: Diffusion models beat GANs

BigGAN

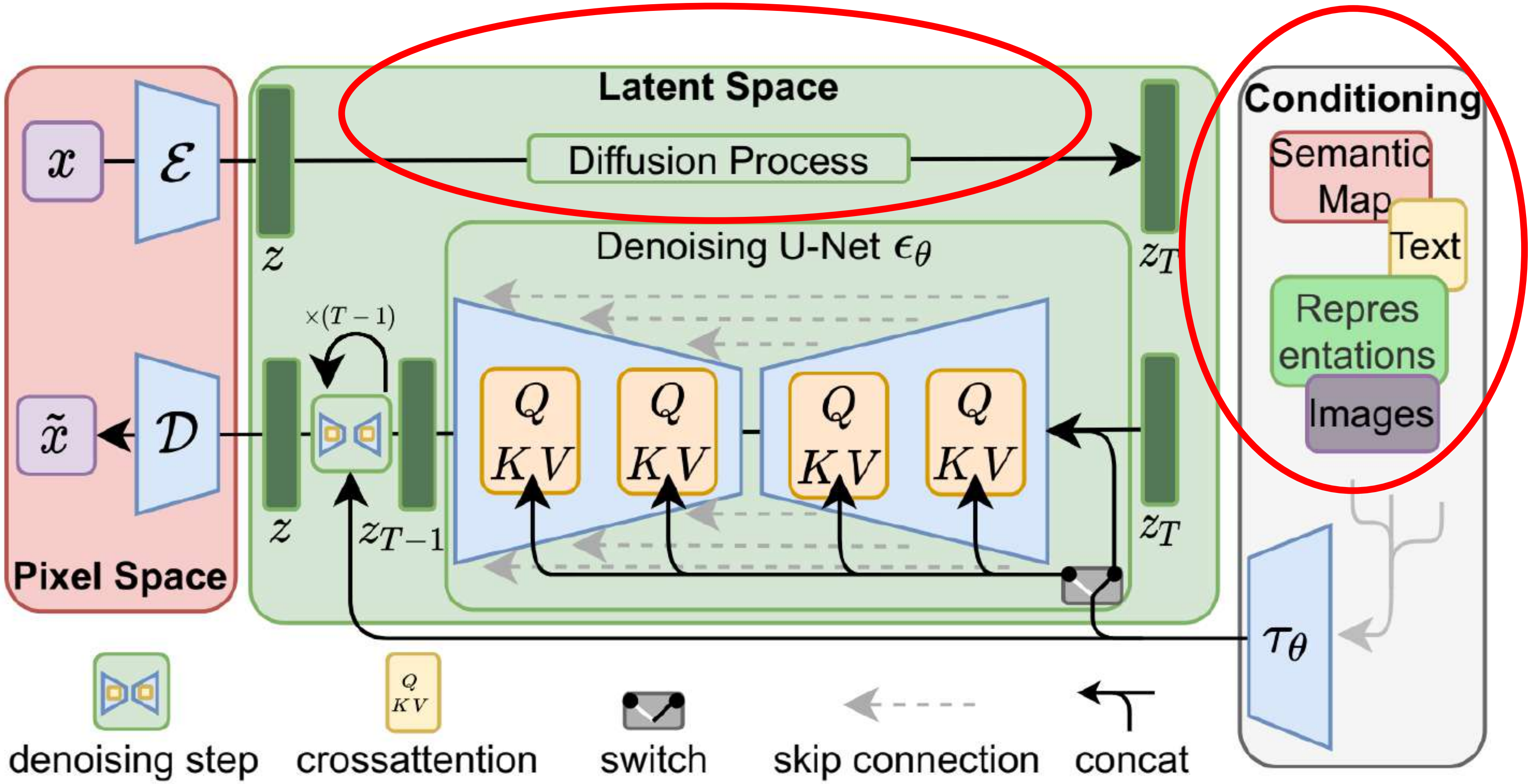




ADM
(diff. model)

P. Dhariwal and A. Nichol. Diffusion models beat GANs on image synthesis. In *NeurIPS* 2021.

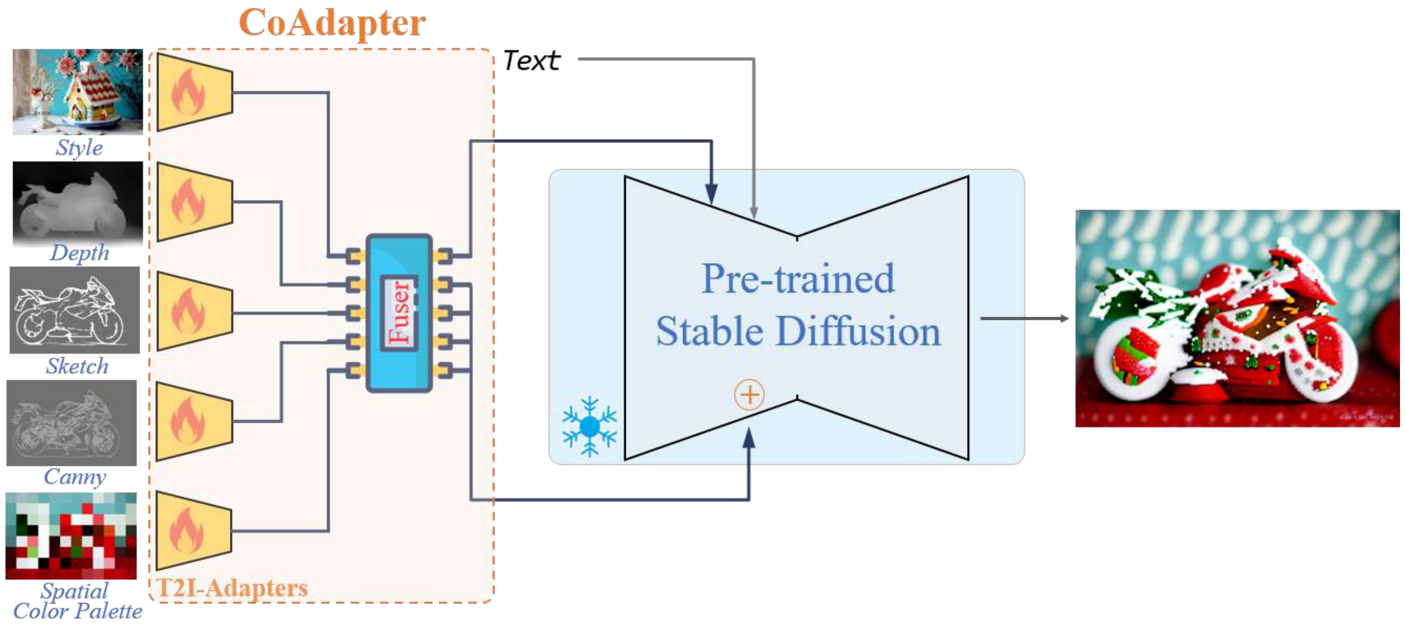
Latent Difussion Models (LDMs)



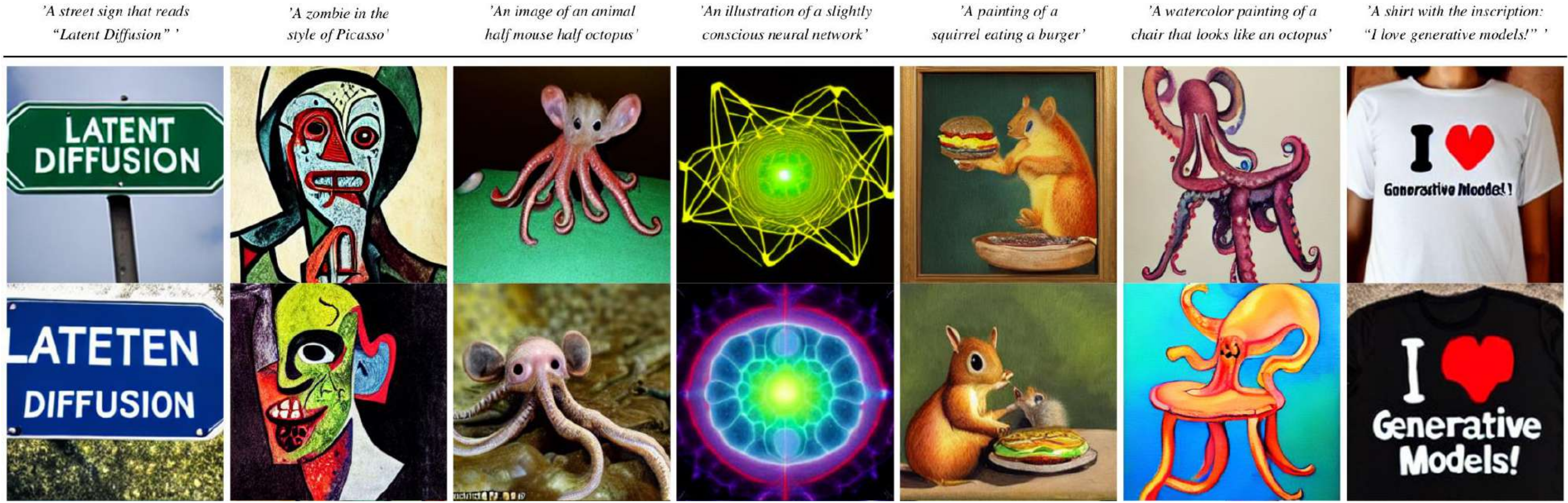
R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. “High-resolution image synthesis with latent diffusion models”. In *CVPR* 2022 (pp. 10684-10695).

Diversity: Much Better Than GANs

- Text: most common
- Sketches, bounding-box layouts, scene graph, semantic segmentation map, depth map, style image, human poses
  ➔ More control over the image generation!
- Representation from other modalities: audio, fMRI signal...



Text Conditioning: Text-to-image


'A street sign that reads "Latent Diffusion" '
'A zombie in the style of Picasso'
'An image of an animal half mouse half octopus'
'An illustration of a slightly conscious neural network'
'A painting of a squirrel eating a burger'
'A watercolor painting of a chair that looks like an octopus'
'A shirt with the inscription: "I love generative models!" '
LATENT DIFFUSION
LATETEN DIFFUSION
I Generative Models!
I Generative Models!

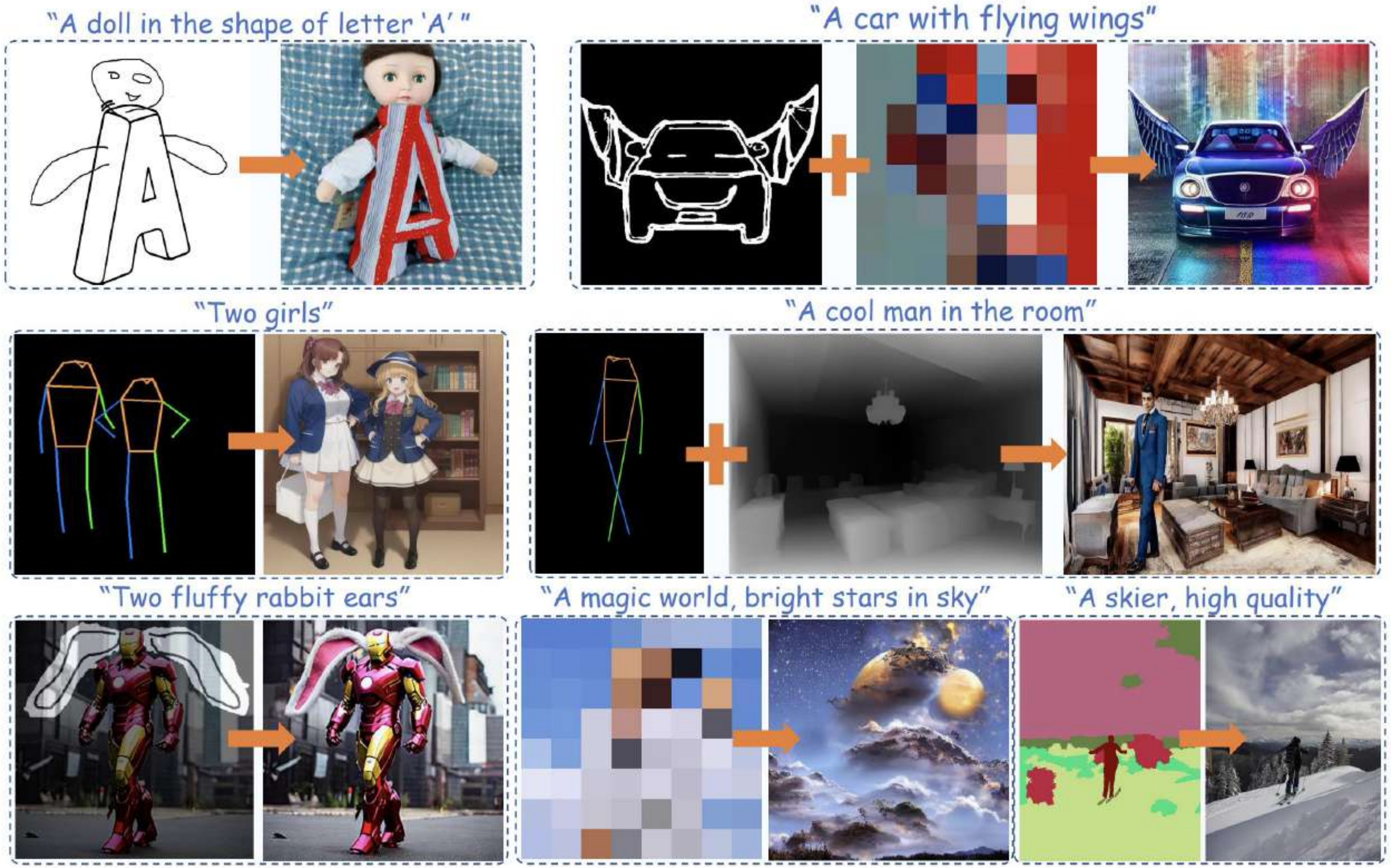
R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. "High-resolution image synthesis with latent diffusion models". In *CVPR* 2022 (pp. 10684-10695).

Image Conditioning

Image/Semantic maps conditional:
Image translation, Inpainting, Super-resolution/Restoration…

R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. “High-resolution image synthesis with latent diffusion models”. In *CVPR* 2022 (pp. 10684-10695).

Various Types of Conditioning



"A doll in the shape of letter 'A' "

"A car with flying wings"

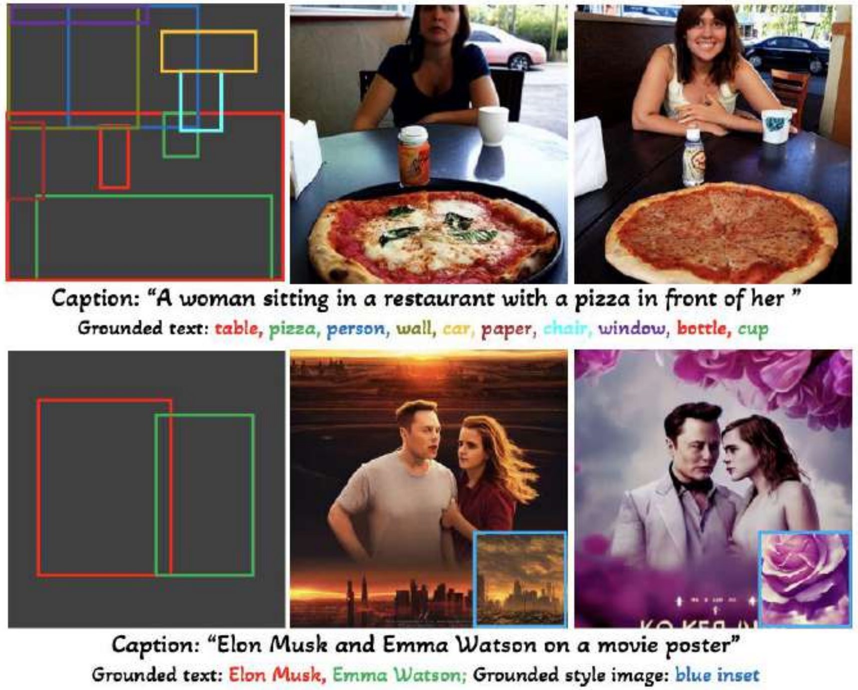
"Two girls"

"A cool man in the room"

"Two fluffy rabbit ears"

"A magic world, bright stars in sky"

"A skier, high quality"



Caption: "A woman sitting in a restaurant with a pizza in front of her "

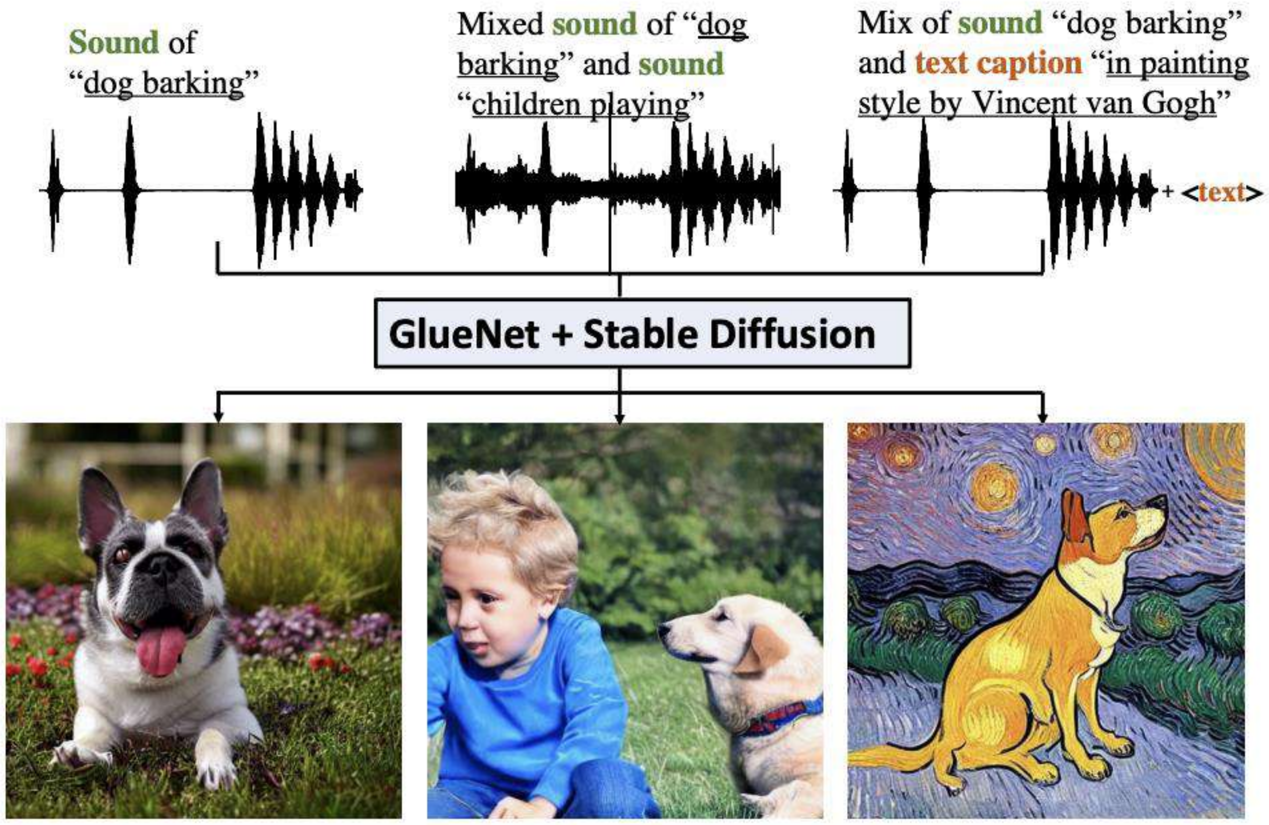
Grounded text: table, pizza, person, wall, car, paper, chair, window, bottle, cup

Caption: "Elon Musk and Emma Watson on a movie poster"

Grounded text: Elon Musk, Emma Watson; Grounded style image: blue inset

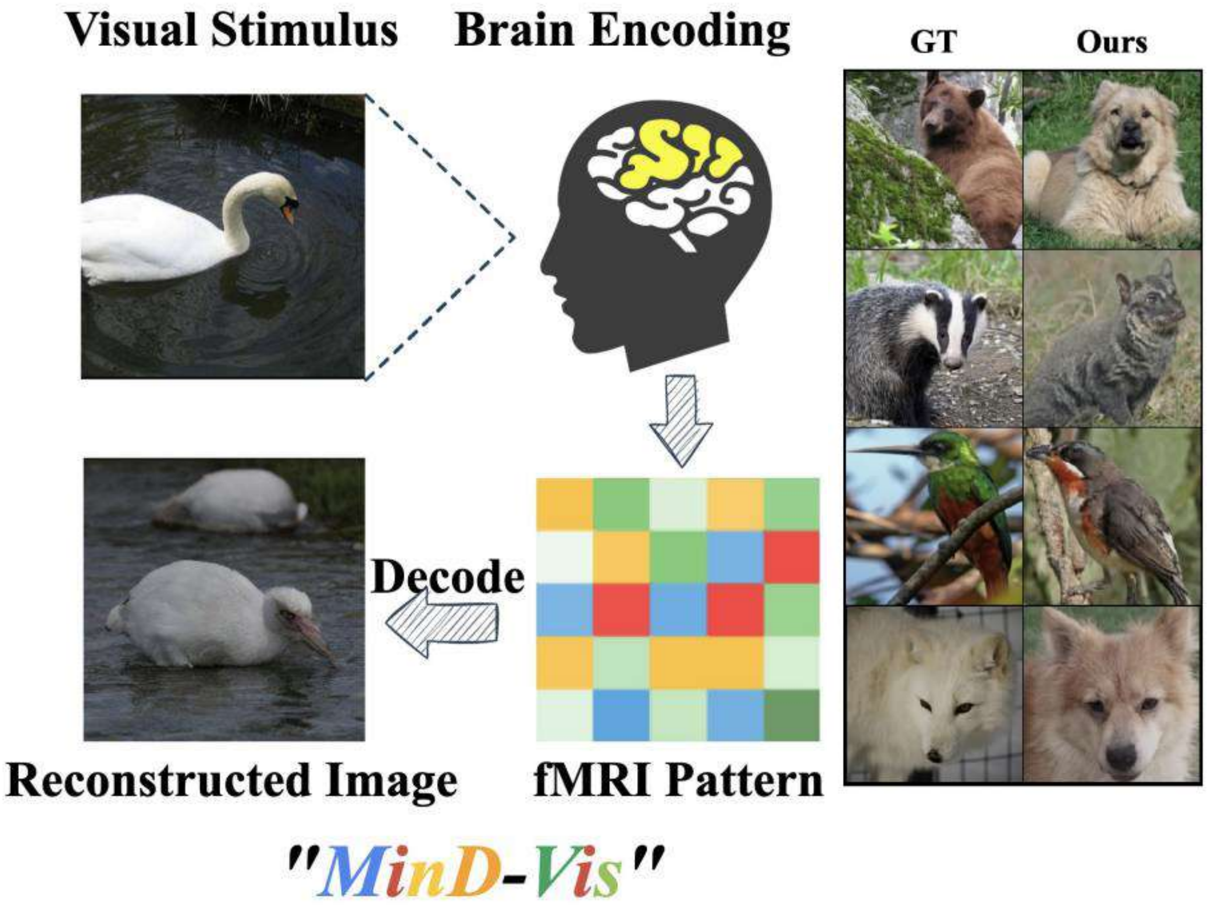
Mou, Chong, et al. "T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models." *arXiv* 2023.

Conditioning from Other Modalities


Sound of "dog barking"
Mixed sound of "dog barking" and sound "children playing"
Mix of sound "dog barking" and text caption "in painting style by Vincent van Gogh"
+ <text>
GlueNet + Stable Diffusion

(a) Single Sound (b) Mixed Sound (c) Sound-Text Mix

Qin, Can, et al. "GlueGen: Plug and Play Multi-modal Encoders for X-to-image Generation." *arXiv* 2023.



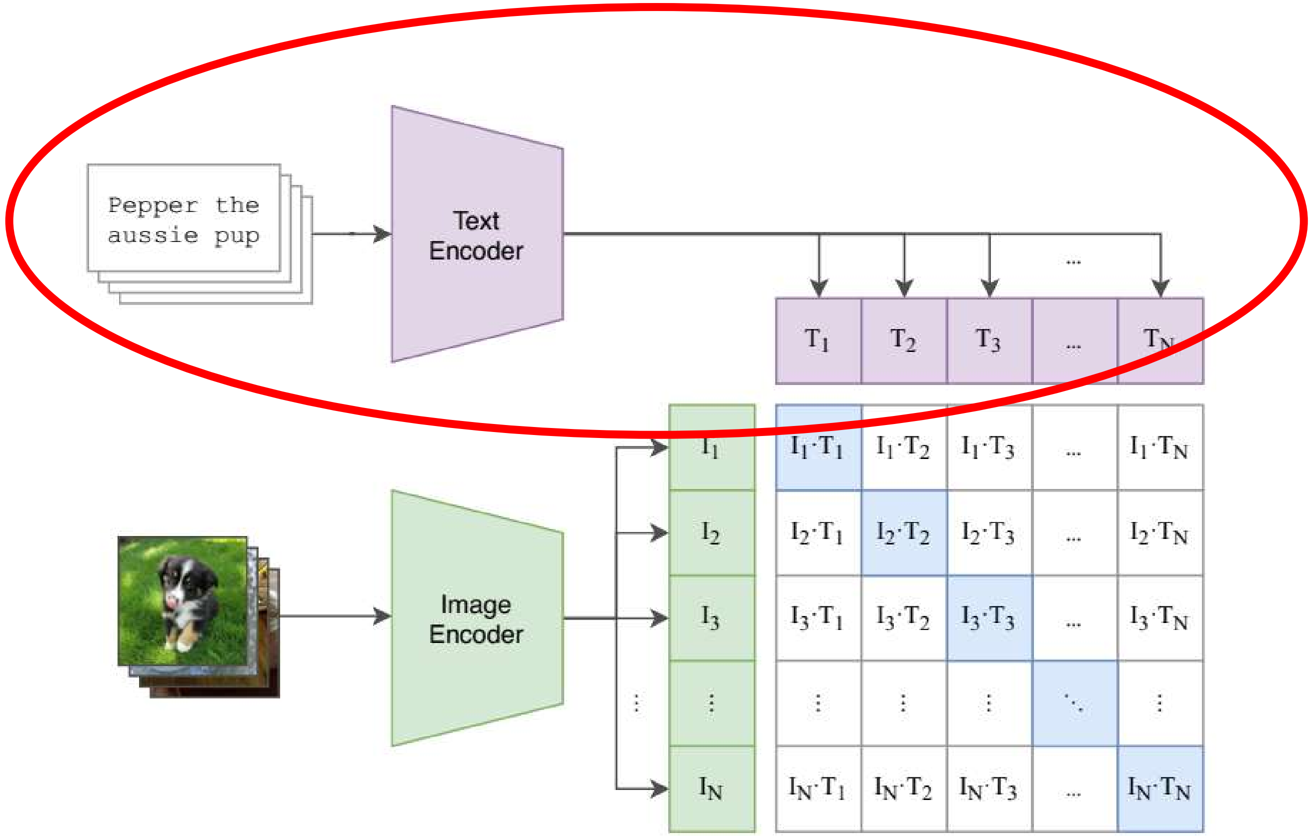
Chen, Zijiao, et al. "Seeing beyond the brain: Conditional diffusion model with sparse masked modeling for vision decoding." *arXiv 2022*

2. State-of-the-art Text-to-image Models

Preliminary: CLIP

Contrastive Language-Image Pretraining (CLIP)

- ➢ Pretrained **aligned** encoders using contrastive loss
  - ✓ Image
  - ✓ Text
- ➢ Only employ the text encoder in text-to-image models

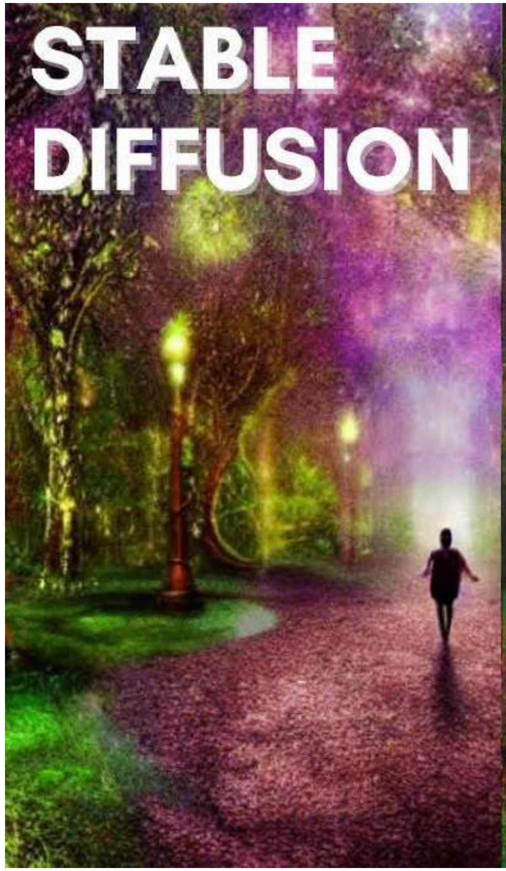


	T_1	T_2	T_3	...	T_N
I_1	$I_1 \cdot T_1$	$I_1 \cdot T_2$	$I_1 \cdot T_3$	...	$I_1 \cdot T_N$
I_2	$I_2 \cdot T_1$	$I_2 \cdot T_2$	$I_2 \cdot T_3$	...	$I_2 \cdot T_N$
I_3	$I_3 \cdot T_1$	$I_3 \cdot T_2$	$I_3 \cdot T_3$	...	$I_3 \cdot T_N$
⋮	⋮	⋮	⋮	⋱	⋮
I_N	$I_N \cdot T_1$	$I_N \cdot T_2$	$I_N \cdot T_3$	...	$I_N \cdot T_N$

A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, and G. Krueger. "Learning transferable visual models from natural language supervision". In *ICML* 2021.

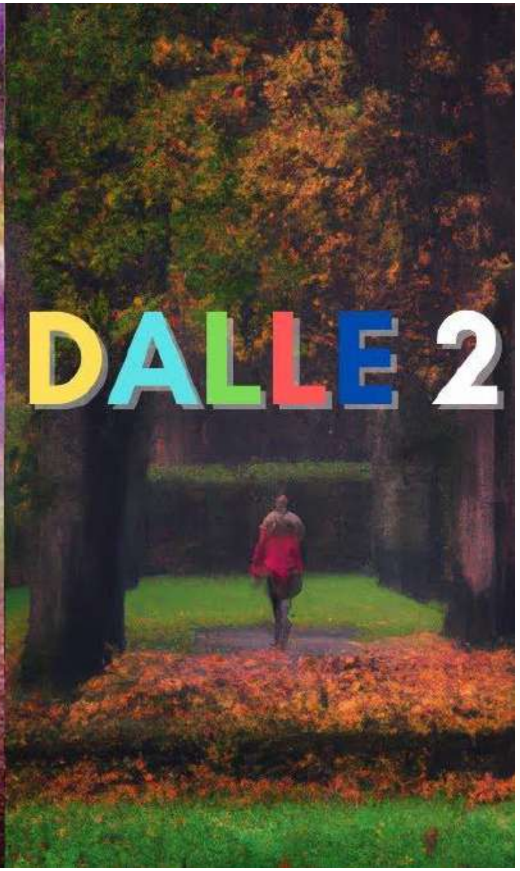
Text-to-image Models

Stability AI



Open-source

OpenAI



Query via website

Adobe

Query via website

Midjourney Lab

Query via Discord

Google/NVIDIA

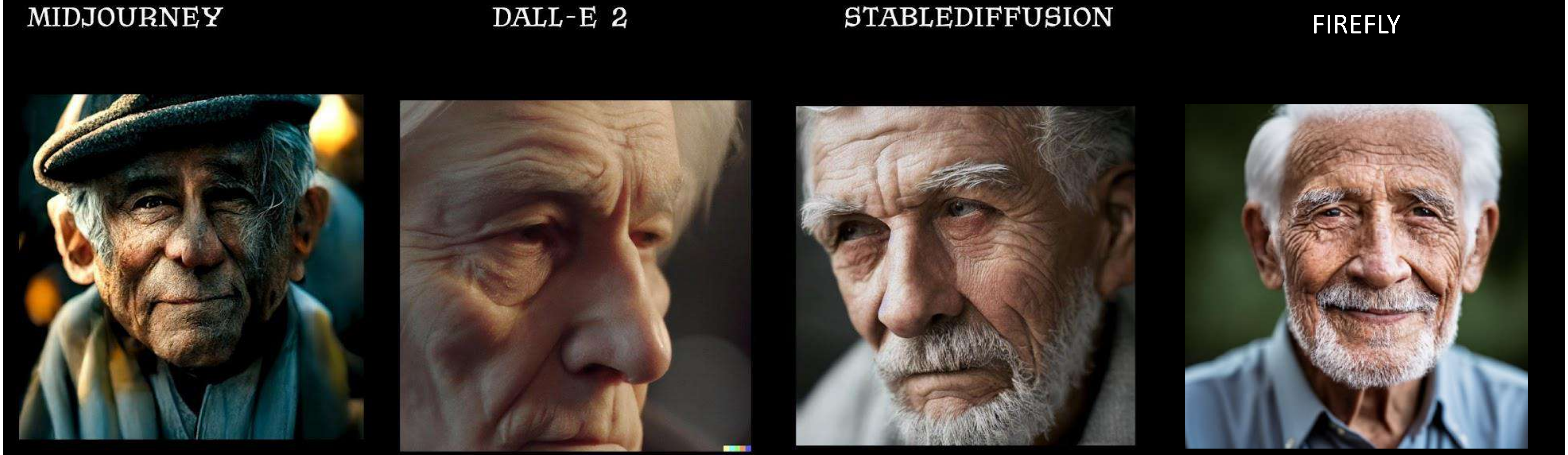
Imagen

?

eDiff-I

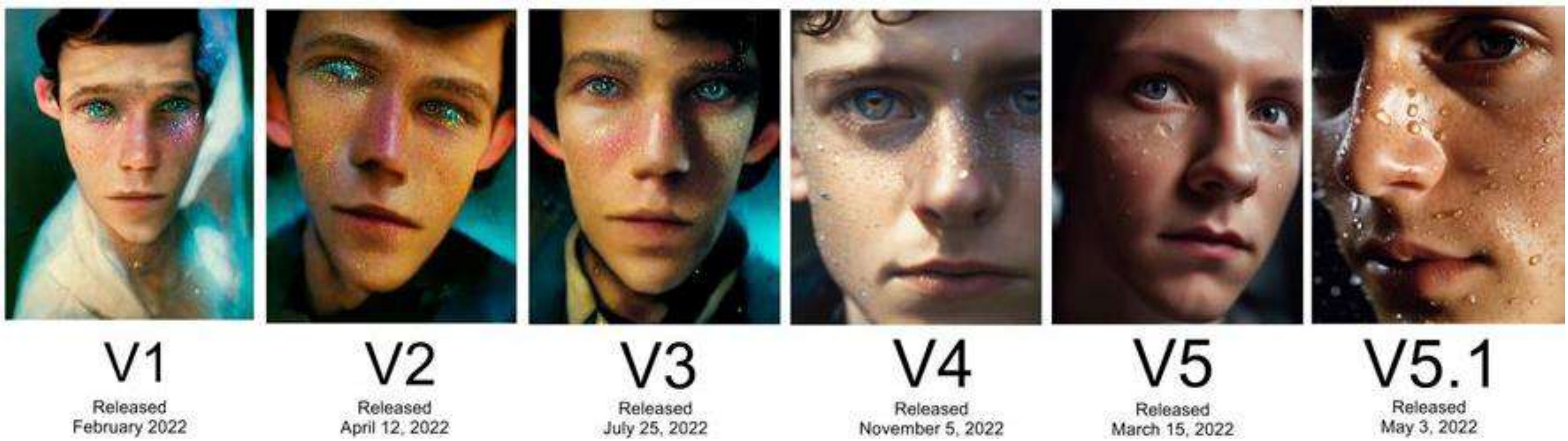
Closed to public

Text-to-image Models



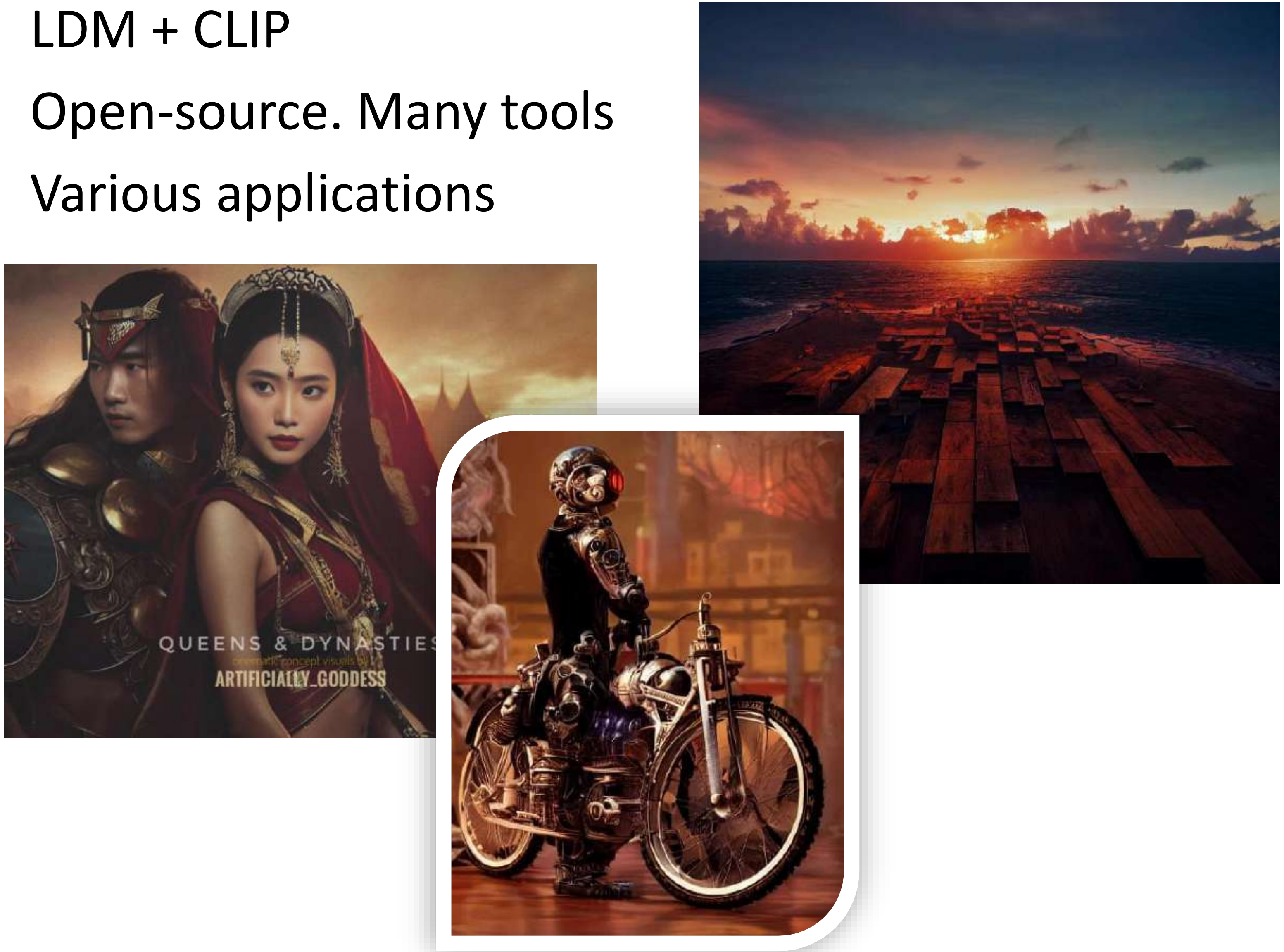
film still, portrait of an old man, wrinkles, dignified look, grey silver hair, peculiar nose, wise, eternal wisdom and beauty, incredible lighting and camera work, depth of field, bokeh, screenshot from a hollywood movie

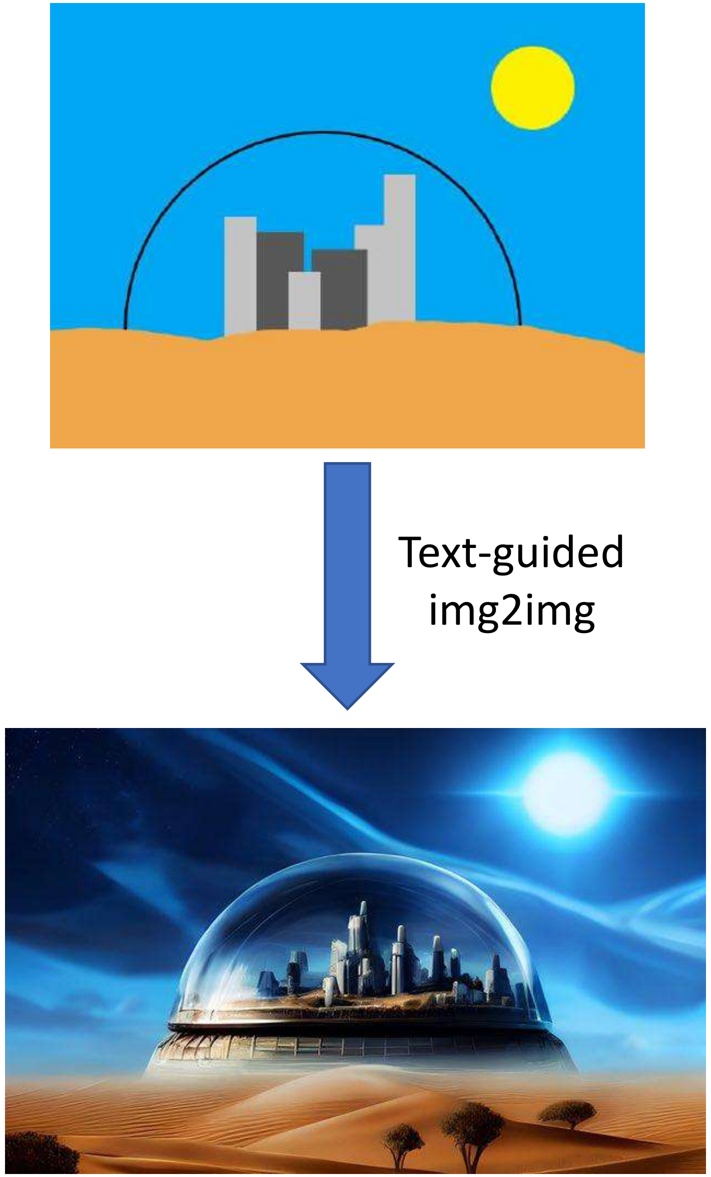
MidJourney versions



Stable Diffusion

- LDM + CLIP
- Open-source. Many tools
- Various applications





Text-to-image Applications

AI Art

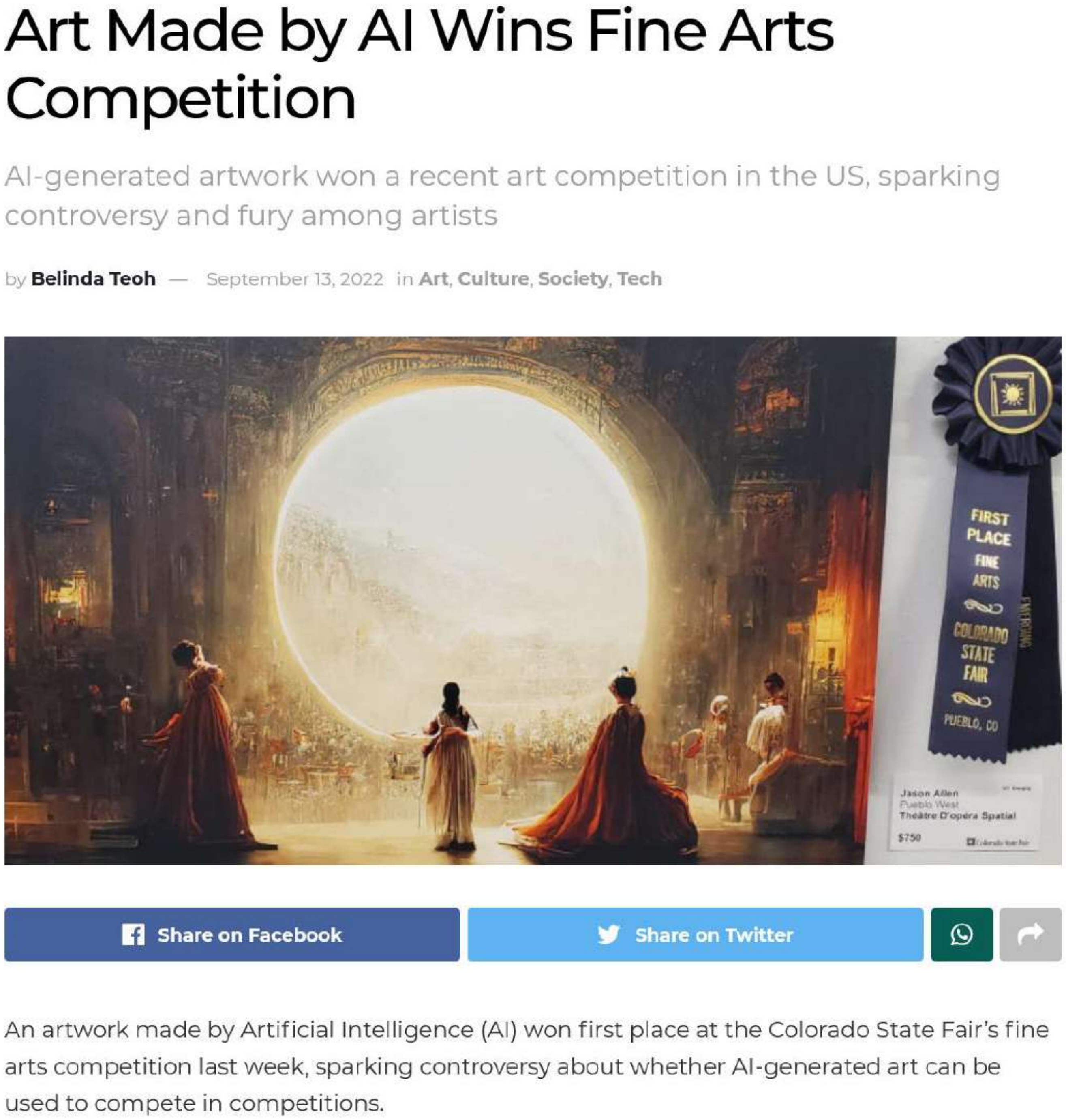
AI generation

Post-processing

(Manual, image restoration tools)

- ✓ Reduce time from days to hours/minutes
- ✓ Everyone can be an artist

AI Art



Art Made by AI Wins Fine Arts Competition

AI-generated artwork won a recent art competition in the US, sparking controversy and fury among artists

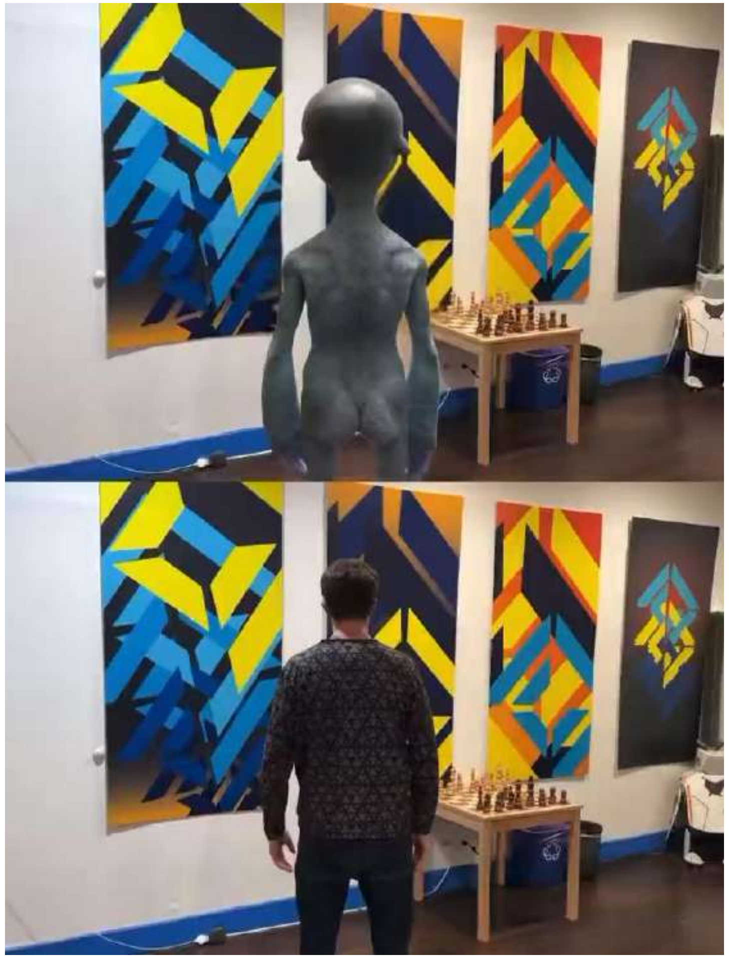
by **Belinda Teoh** — September 13, 2022 in **Art**, **Culture**, **Society**, **Tech**

An artwork made by Artificial Intelligence (AI) won first place at the Colorado State Fair's fine arts competition last week, sparking controversy about whether AI-generated art can be used to compete in competitions.

Path for future VFX

Path for future VFX

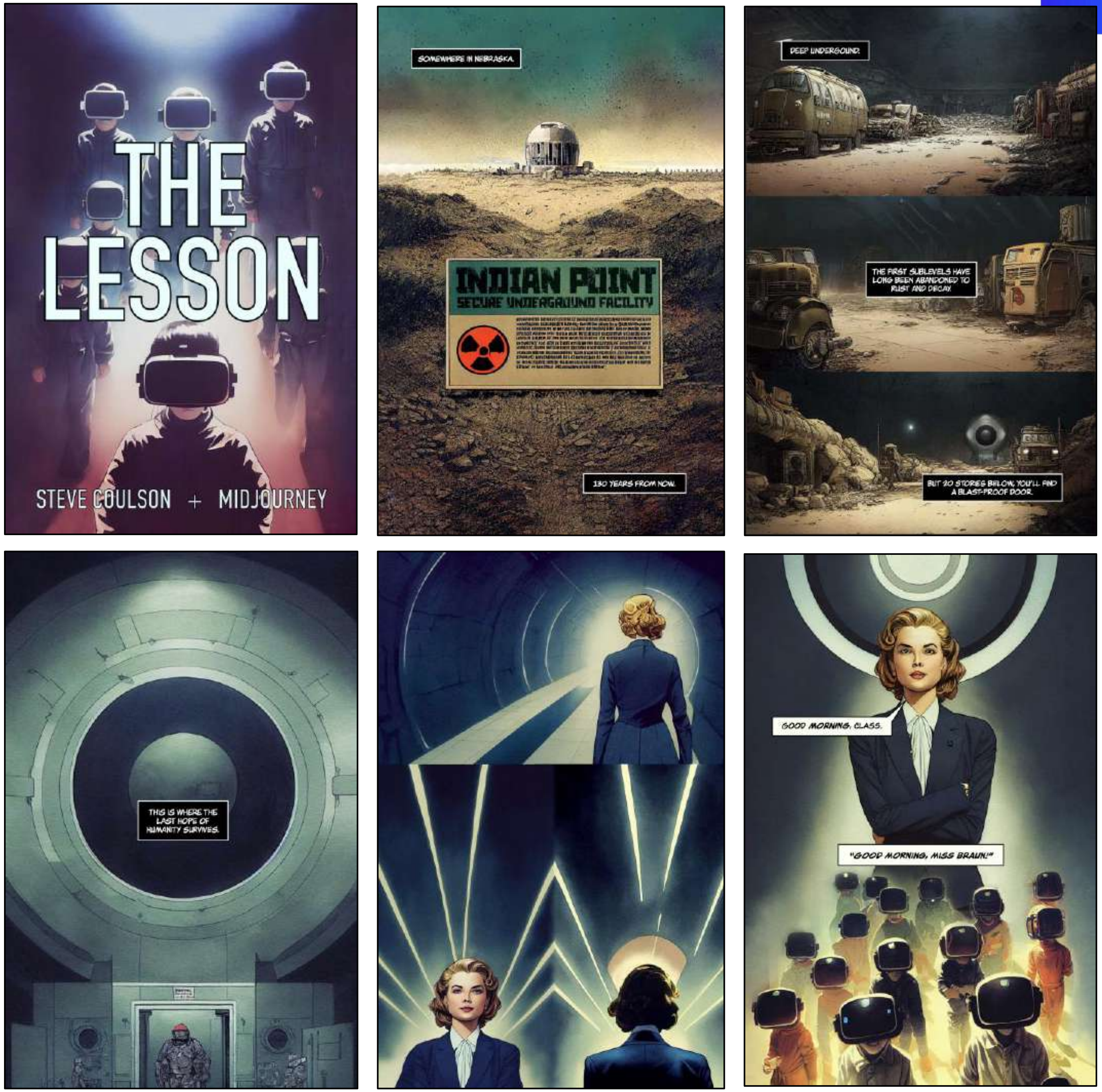
https://wonderdynamics.com/



https://twitter.com/SirWrender/status/1643319553789947905

https://twitter.com/eLPenry/status/1643931490290483201

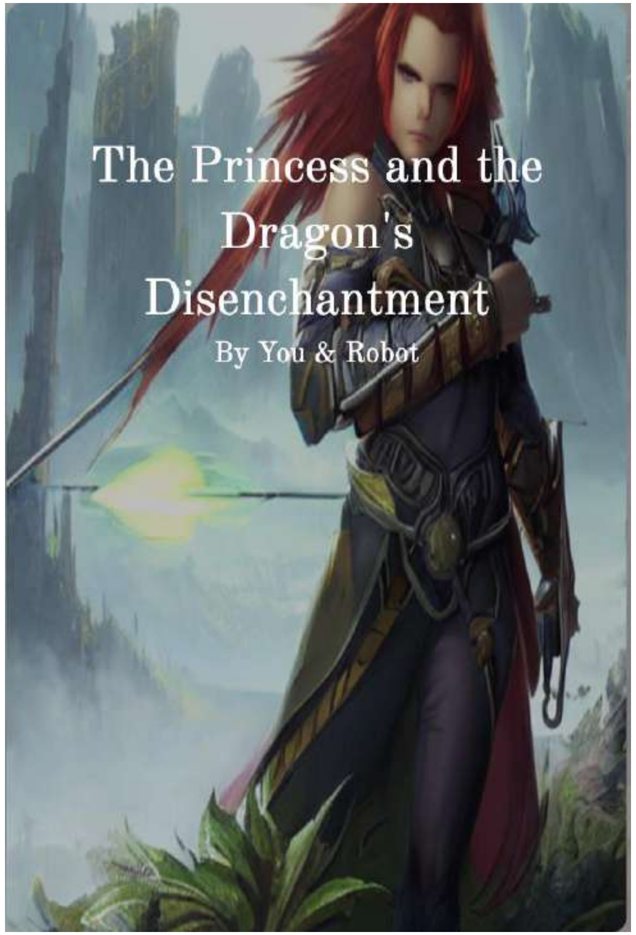
Story Synthesis


THE LESSON
STEVE COULSON + MIDJOURNEY
SOMEWHERE IN NEBRASKA.
INDIAN POINT
SECURE UNDERGROUND FACILITY
130 YEARS FROM NOW.
DEEP UNDERGROUND.
THE FIRST SUBLEVELS HAVE LONG BEEN ABANDONED TO RUST AND DECAY.
BUT 20 STORIES BELOW, YOU'LL FIND A BLAST-PROOF DOOR.
THIS IS WHERE THE LAST HOPE OF HUMANITY SURVIVES.
GOOD MORNING, CLASS.
"GOOD MORNING, MISS BRAUN!"

[Quick Start Guide - Midjourney Documentation (gitbook.io)]

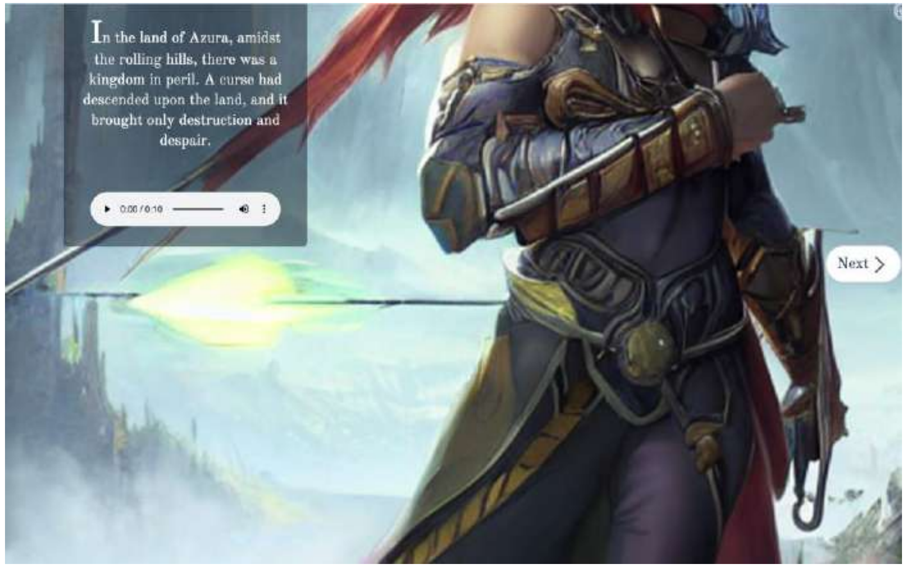
[CAMPFIRE / Comics (campfirenyc.com)]

Story Synthesis

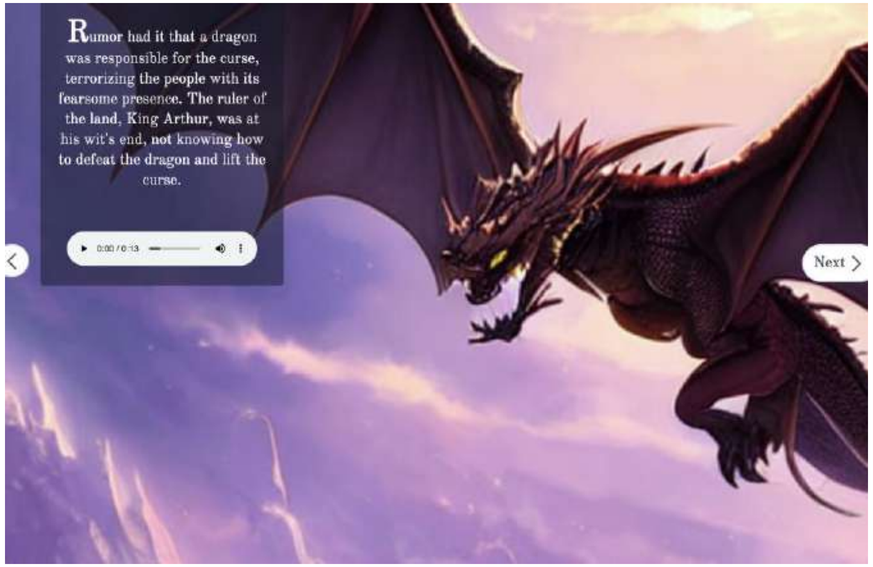


The Princess and the Dragon's Disenchantment

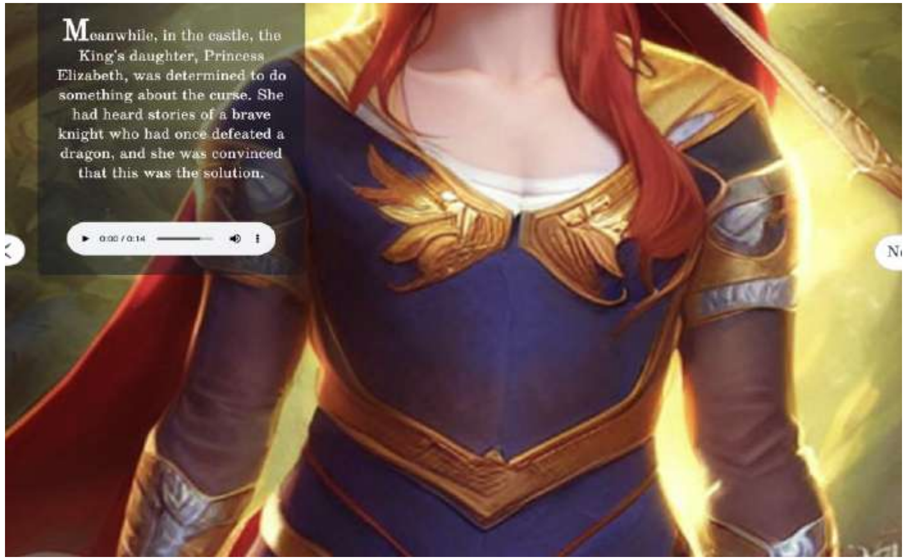
By You & Robot



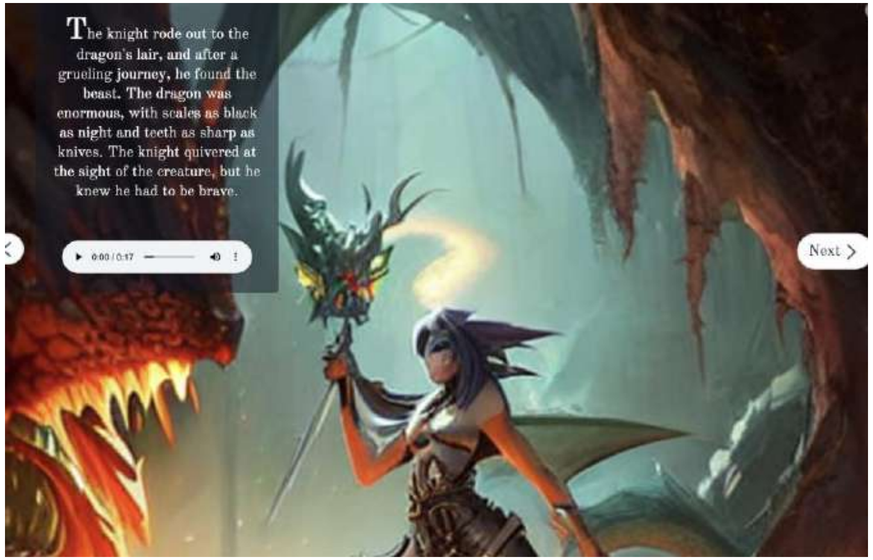
In the land of Azura, amidst the rolling hills, there was a kingdom in peril. A curse had descended upon the land, and it brought only destruction and despair.



Rumor had it that a dragon was responsible for the curse, terrorizing the people with its fearsome presence. The ruler of the land, King Arthur, was at his wit's end, not knowing how to defeat the dragon and lift the curse.



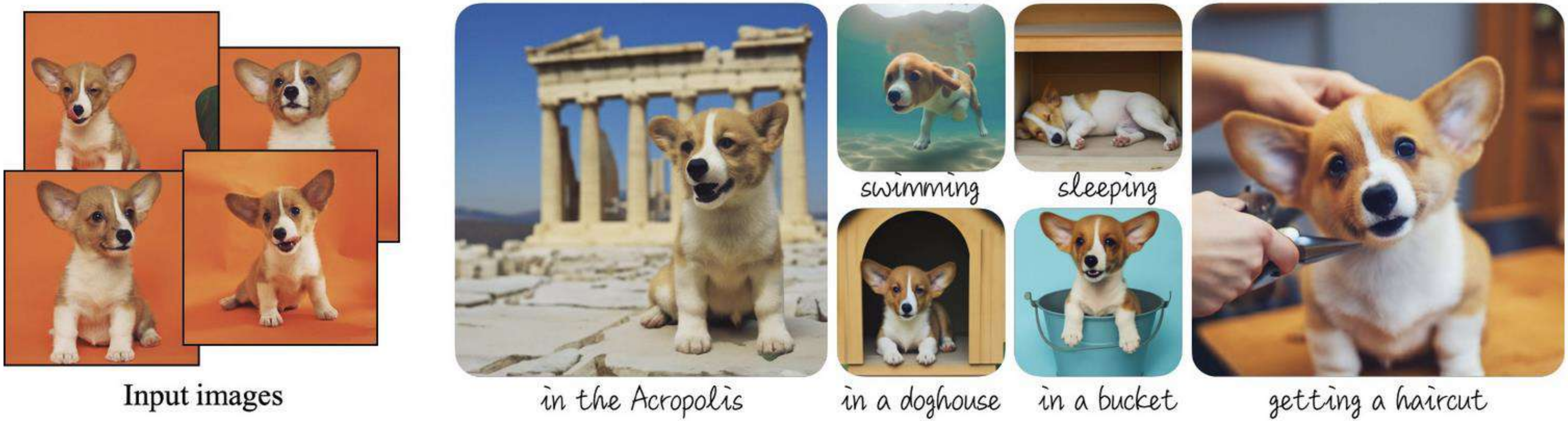
Meanwhile, in the castle, the King's daughter, Princess Elizabeth, was determined to do something about the curse. She had heard stories of a brave knight who had once defeated a dragon, and she was convinced that this was the solution.



The knight rode out to the dragon's lair, and after a grueling journey, he found the beast. The dragon was enormous, with scales as black as night and teeth as sharp as knives. The knight quivered at the sight of the creature, but he knew he had to be brave.

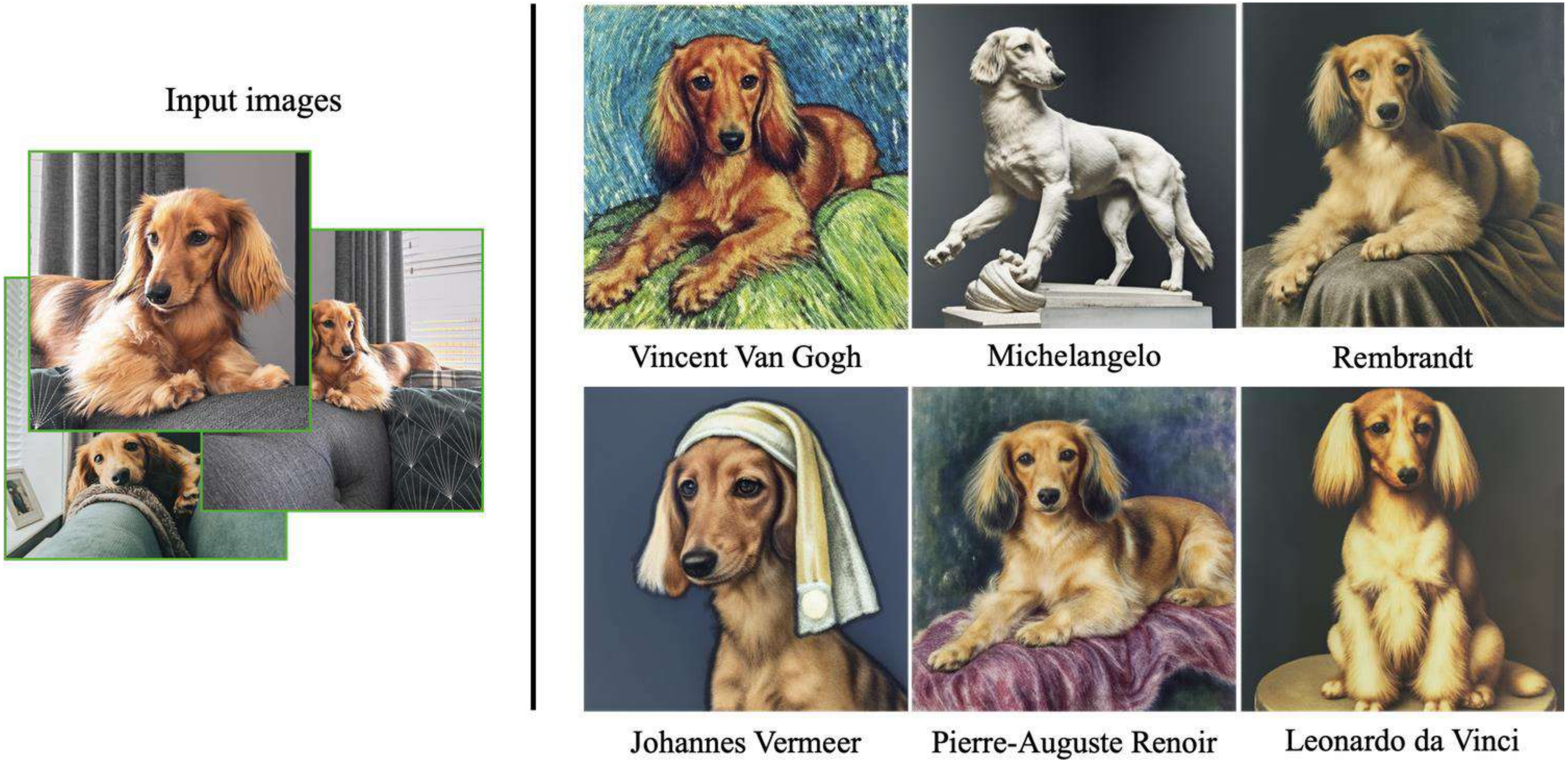
https://onceuponabot.com/story

Personalization



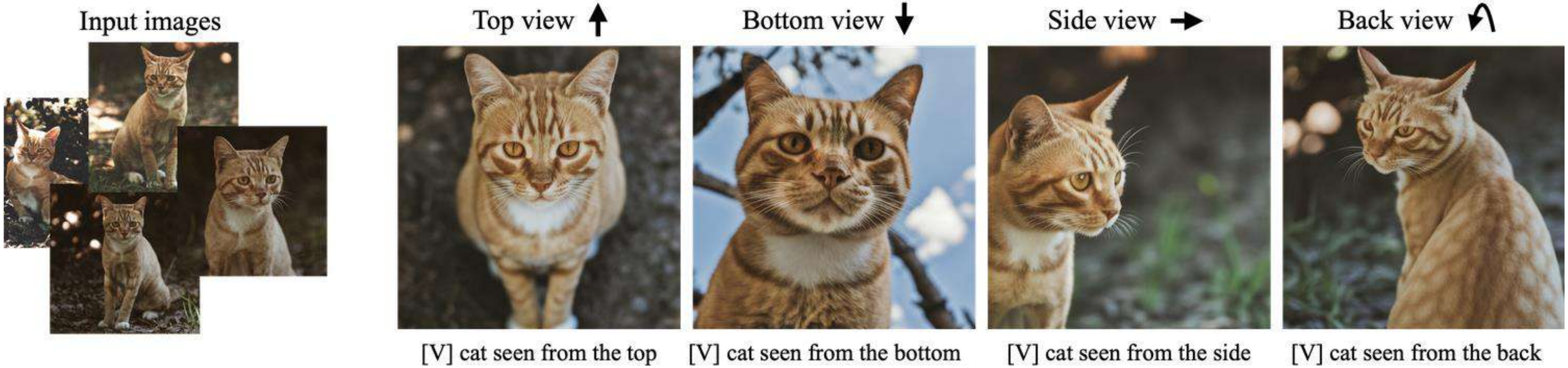
N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein and K. Aberman. “Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation”. *arXiv preprint arXiv:2208.12242*.

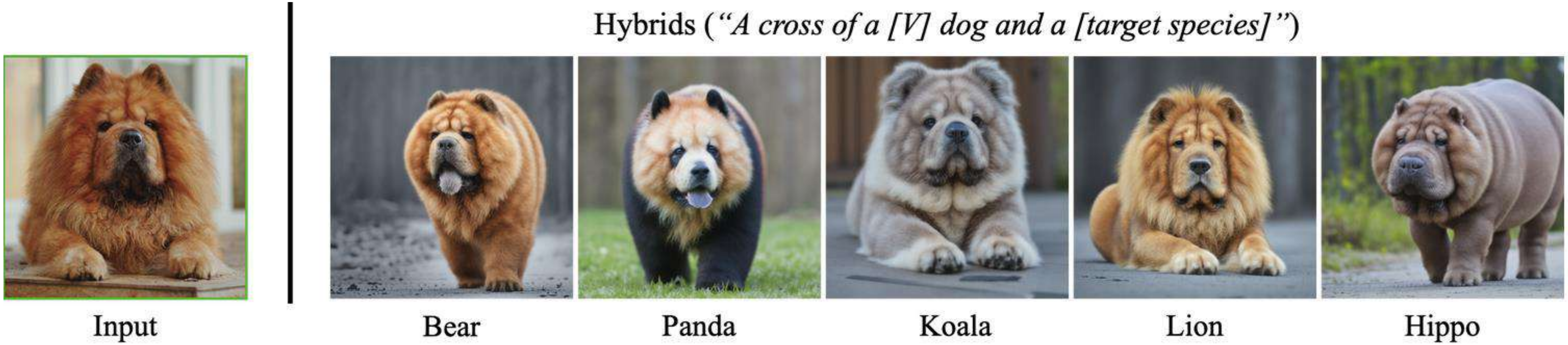
Personalization



N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein and K. Aberman. “Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation”. *arXiv preprint arXiv:2208.12242*.

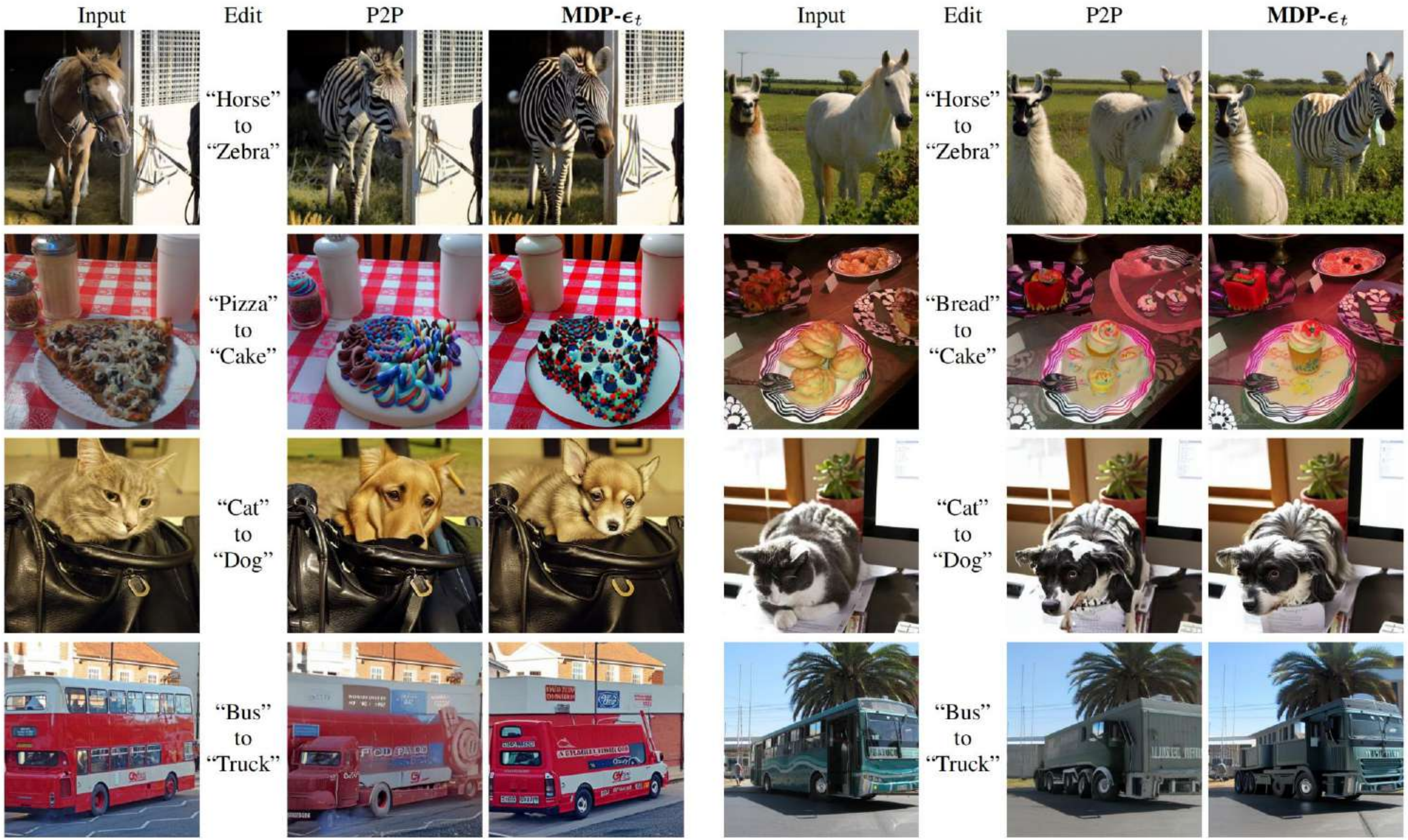
Personalization


Input images
Top view
Bottom view
Side view
Back view
[V] cat seen from the top
[V] cat seen from the bottom
[V] cat seen from the side
[V] cat seen from the back



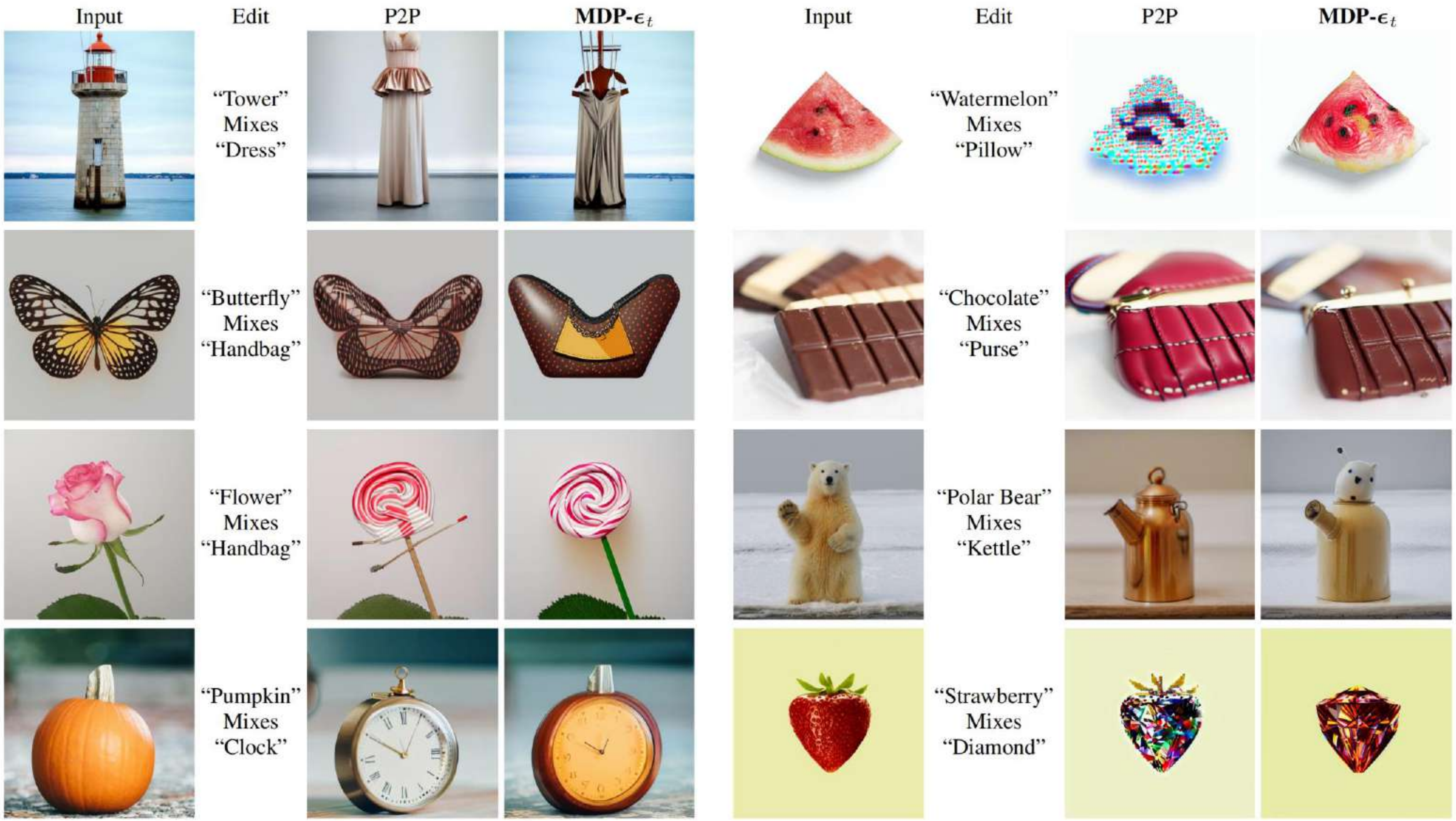
N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein and K. Aberman. "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation". *arXiv preprint arXiv:2208.12242*.

Text-guided Image Editing



Wang, Qian, et al. "MDP: A Generalized Framework for Text-Guided Image Editing by Manipulating the Diffusion Path." *arXiv* 2023

Text-guided Image Editing


Input
Edit
P2P
MDP-ϵ_t
Input
Edit
P2P
MDP-ϵ_t
“Tower” Mixes “Dress”
“Watermelon” Mixes “Pillow”
“Butterfly” Mixes “Handbag”
“Chocolate” Mixes “Purse”
“Flower” Mixes “Handbag”
“Polar Bear” Mixes “Kettle”
“Pumpkin” Mixes “Clock”
“Strawberry” Mixes “Diamond”

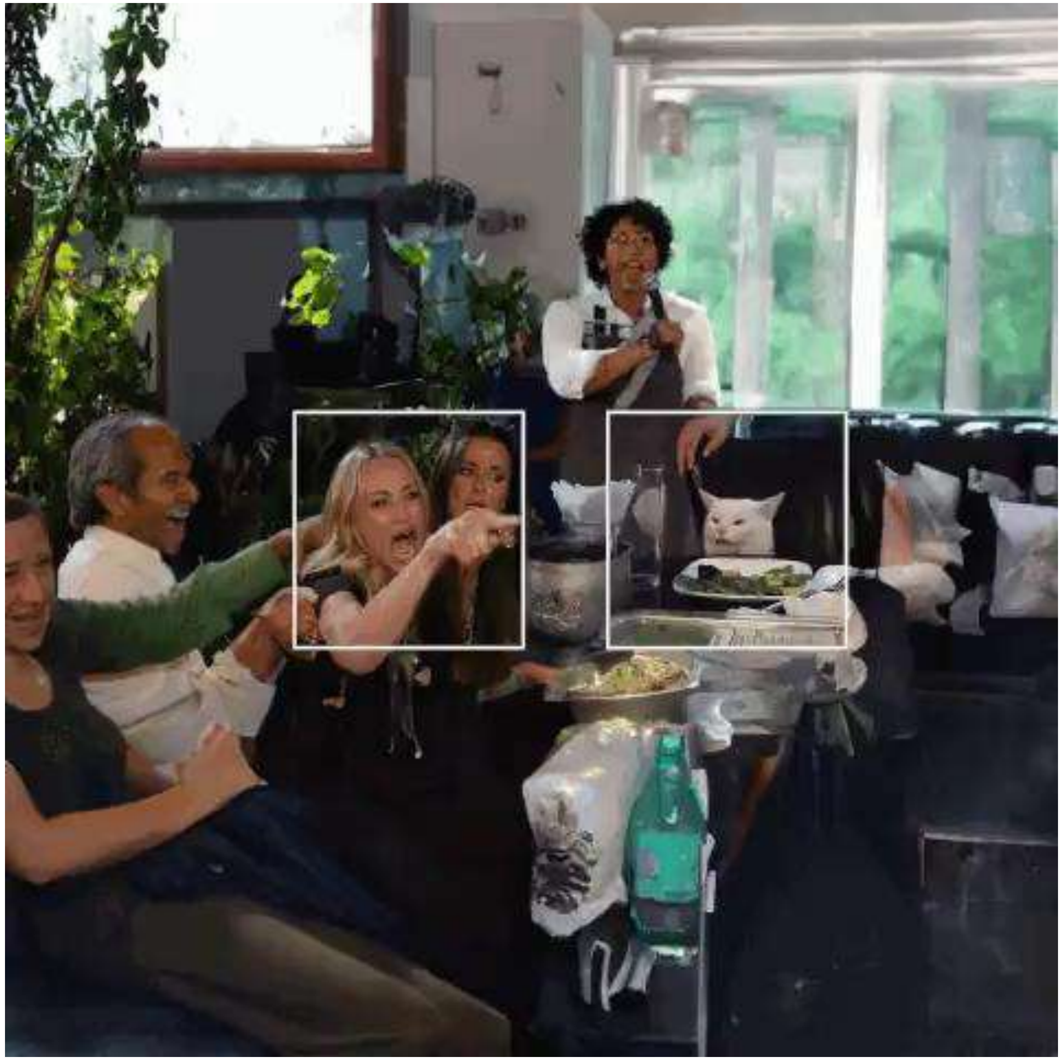
Wang, Qian, et al. "MDP: A Generalized Framework for Text-Guided Image Editing by Manipulating the Diffusion Path." *arXiv* 2023

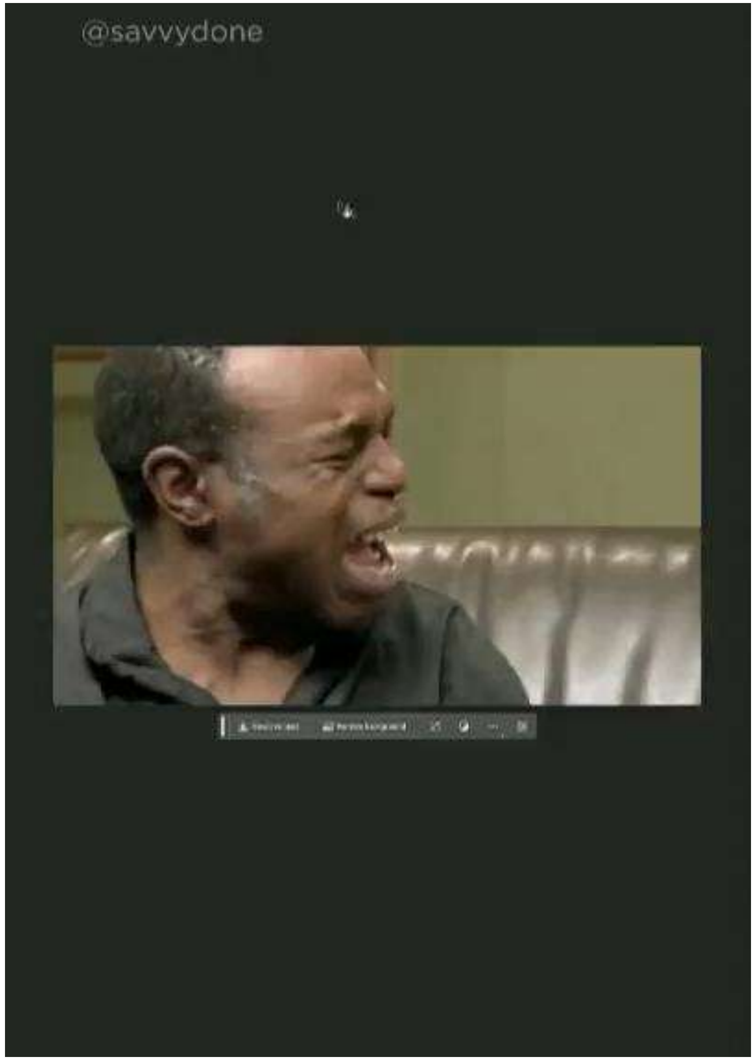
Outpainting

Generative Fill (Photoshop Beta)

Outpainting

Generative Fill (Photoshop Beta)



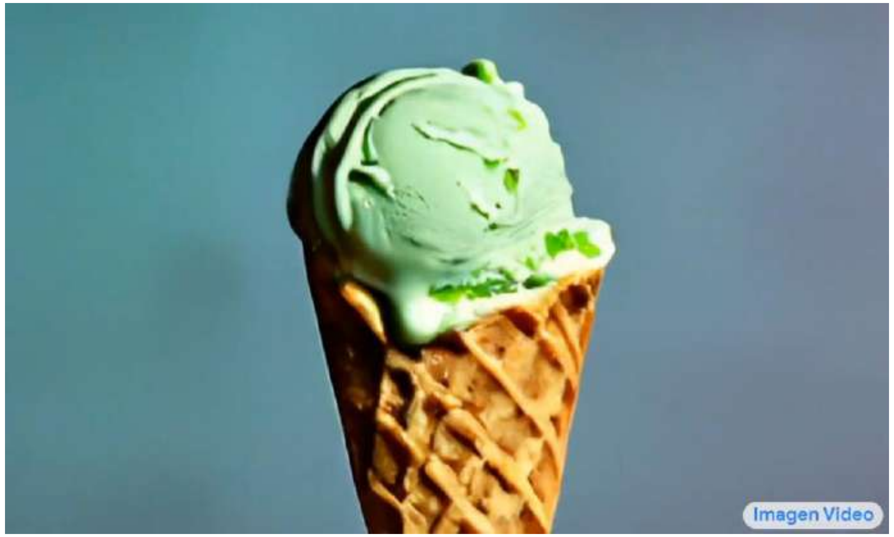


Outpainting

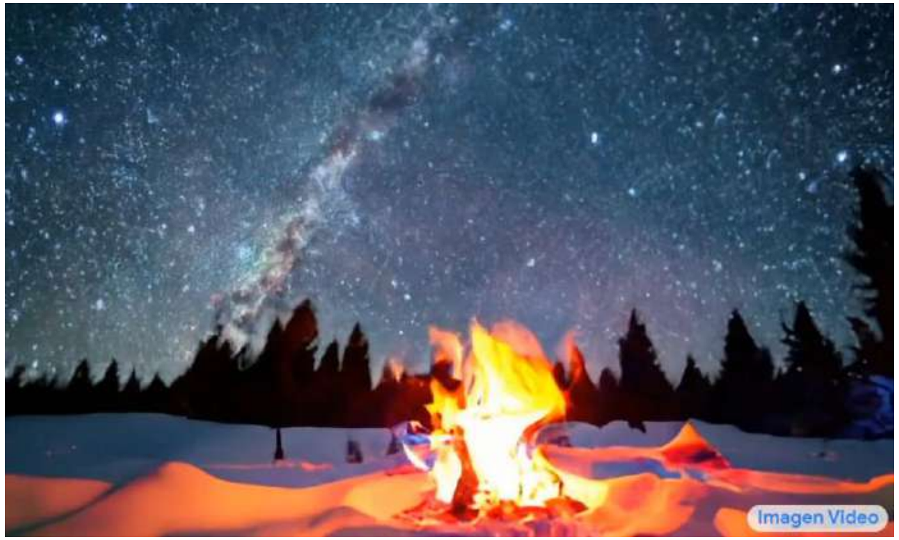
Zoom-out (Midjourney 5.2)

3. State-of-the-art Text-to-X Models

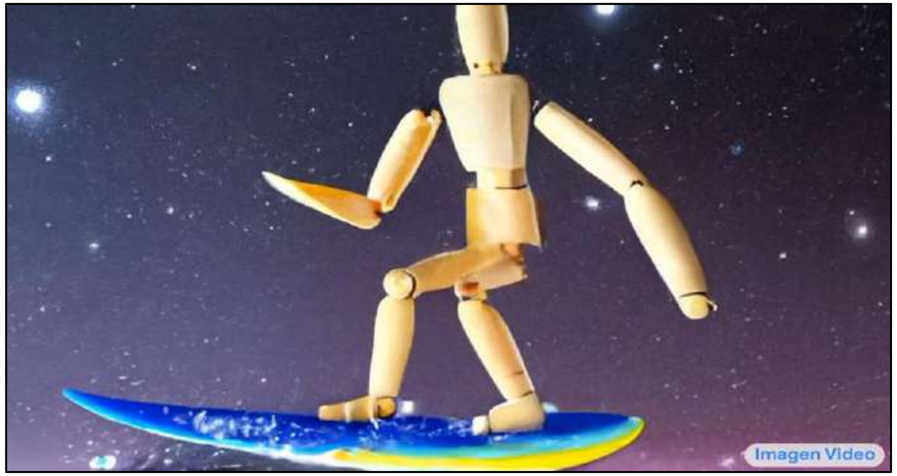
Text-to-Video



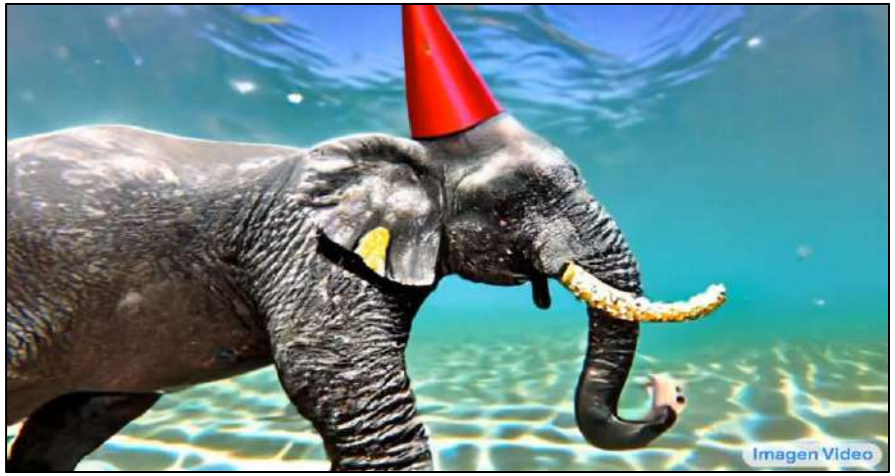
Melting ice cream dripping down the cone.



Campfire at night in a snowy forest with starry sky in the background.



Wooden figurine surfing on a surfboard in space

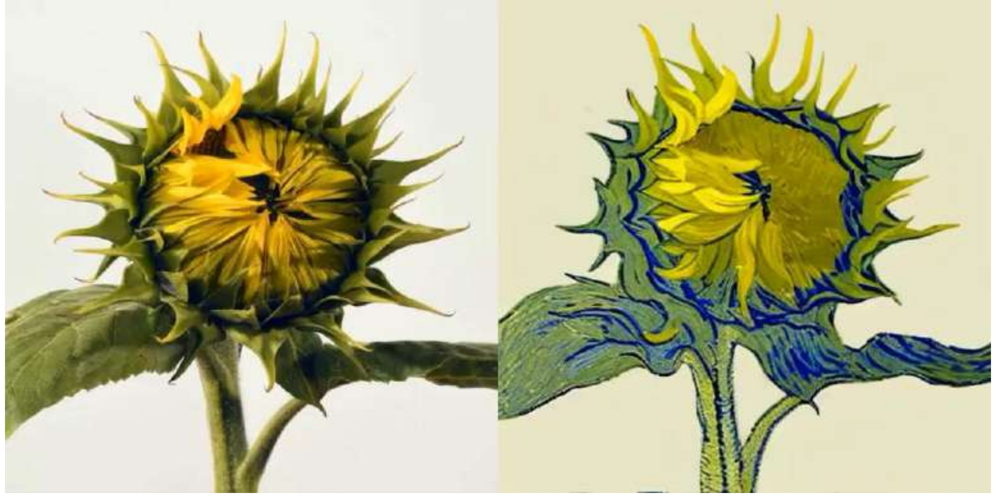


A happy elephant wearing a birthday hat walking under the sea

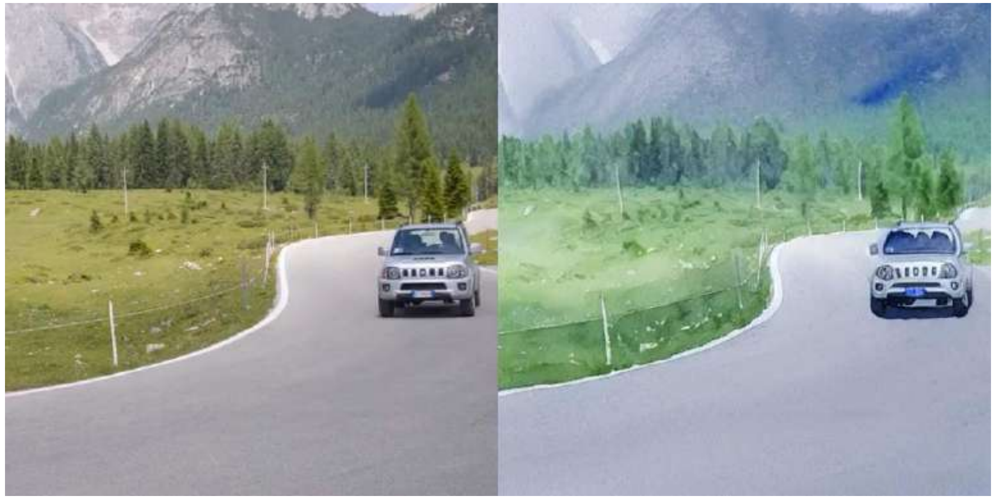
https://makeavideo.studio https://imagen.research.google/video/ https://phenaki.video/

Text-guided Video Editing

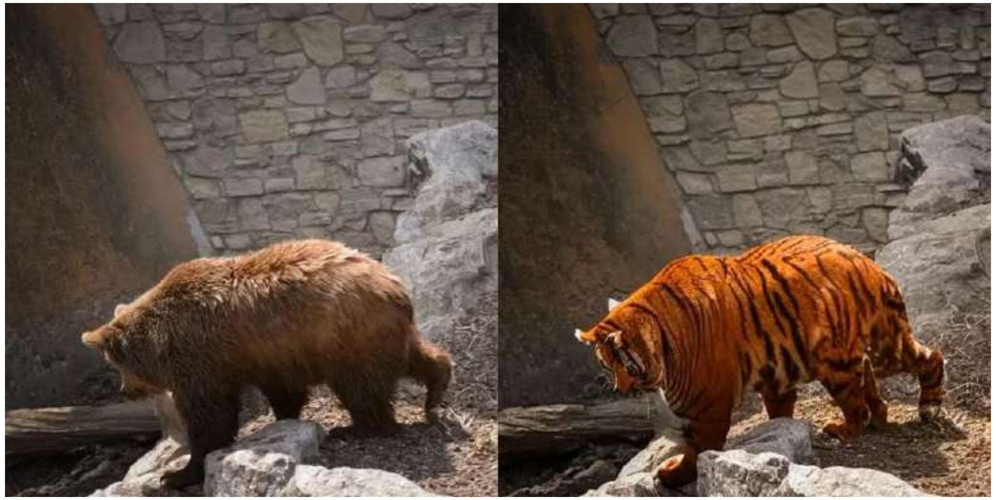
Fate/Zero (fate-zero-edit.github.io)



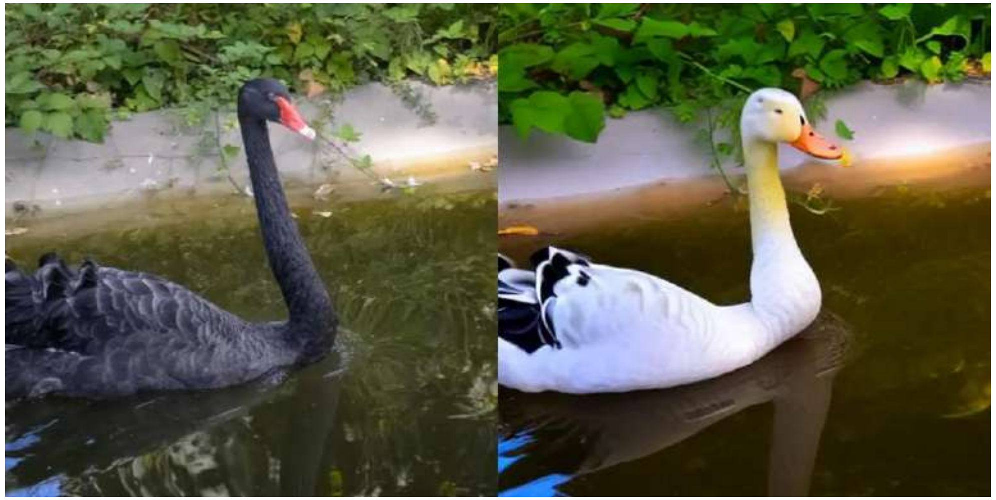
+ Van Gogh Style Painting



+ Watercolor Painting



Bear → A Red Tiger



Swan → White Duck*

Text-to-3D

https://dreamfusion3d.github.io/gallery.html

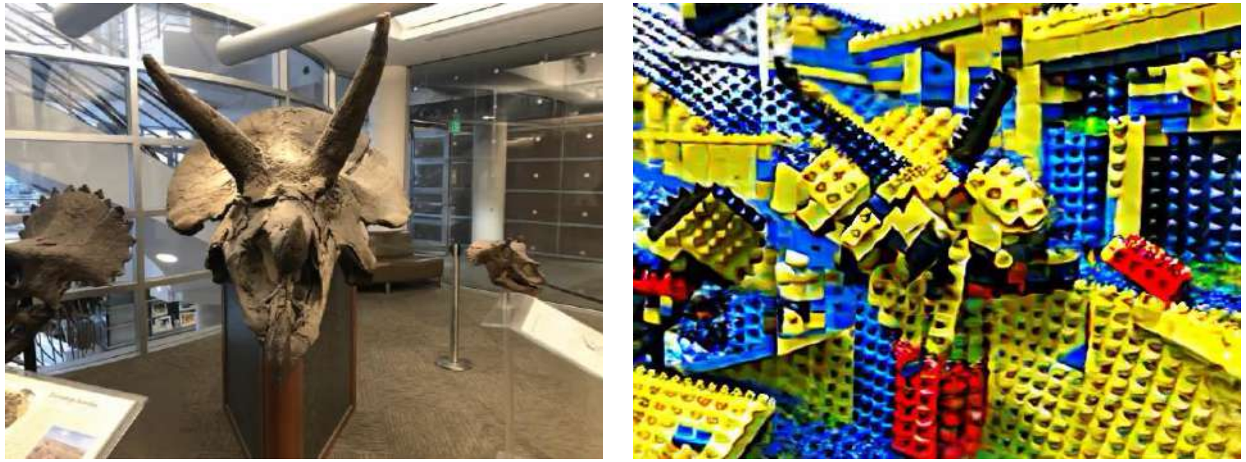
A car made out of sushi.

A peacock on a surfboard.

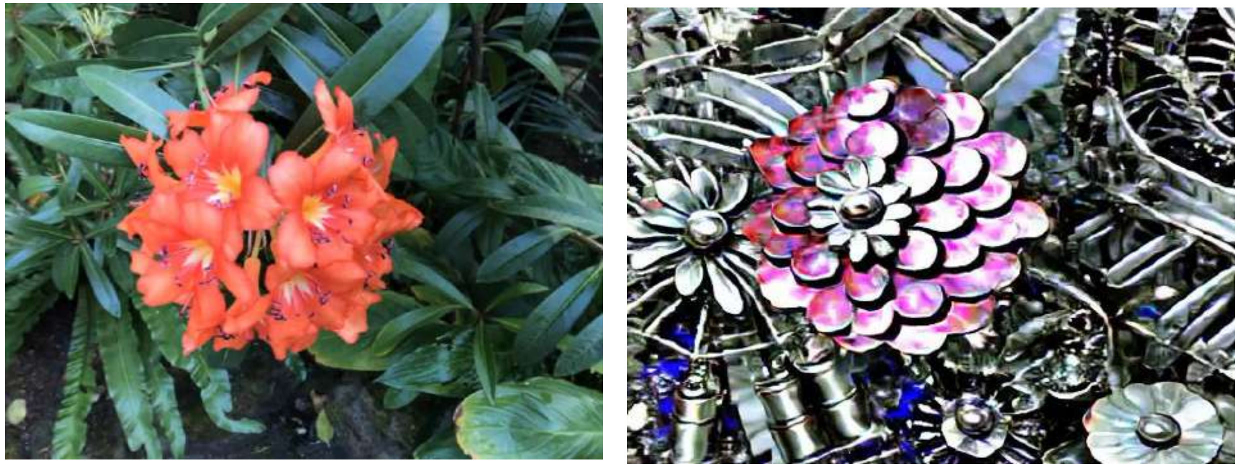
Magic3D: High-Resolution

Text-guided 3D Editing

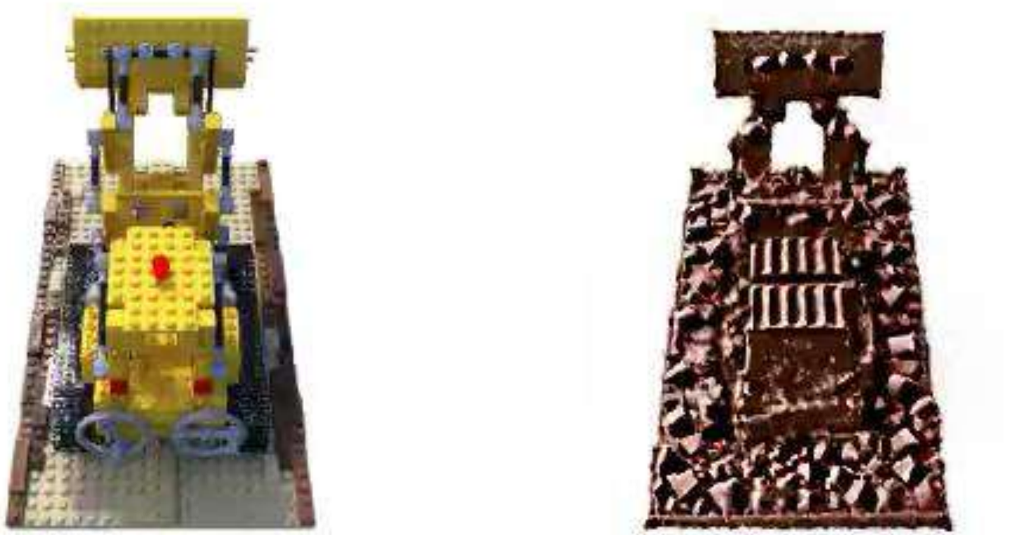
Instruct 3D-to-3D: Text Instruction Guided 3D-to-3D conversion (sony.github.io)



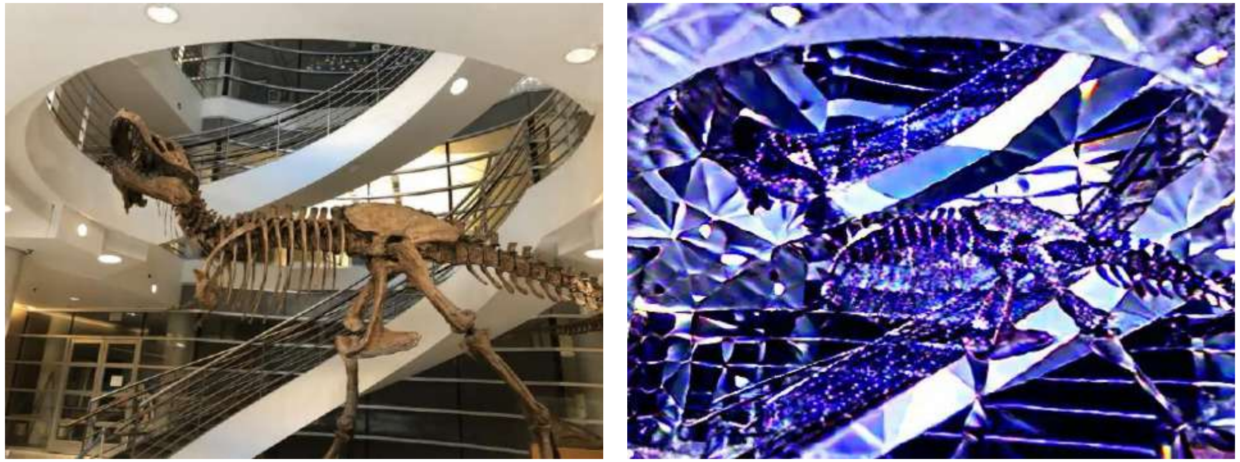
make it Lego blocks



convert it to a mechanical flower made of silver metal

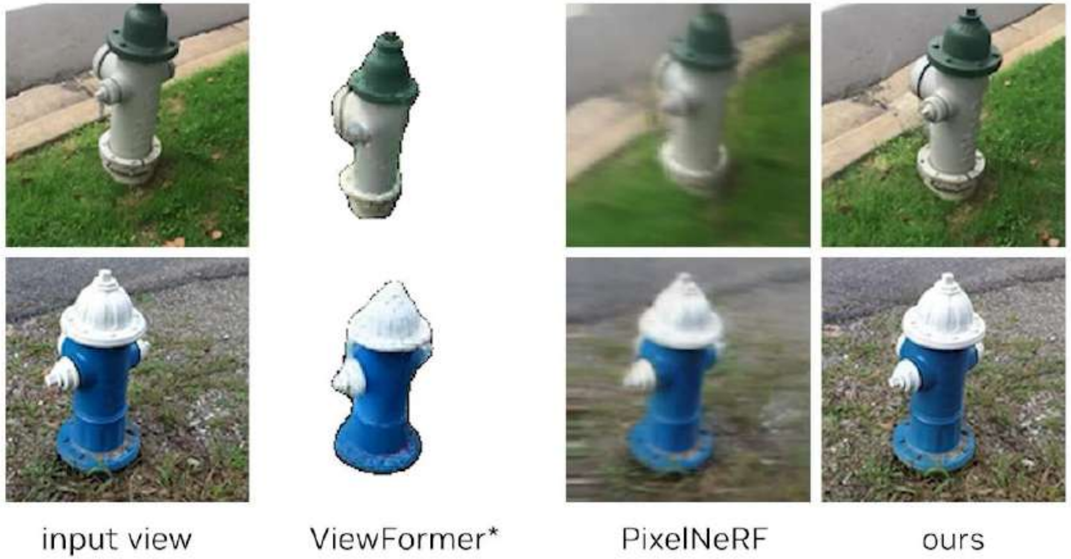


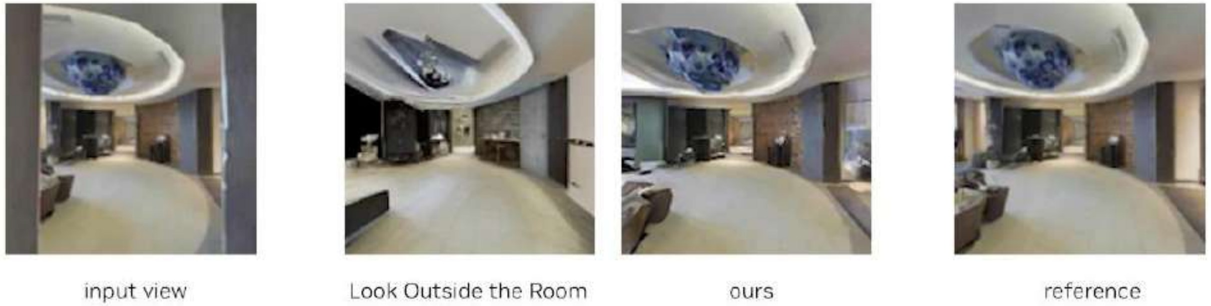
make it chocolate



What if it was made of diamonds?

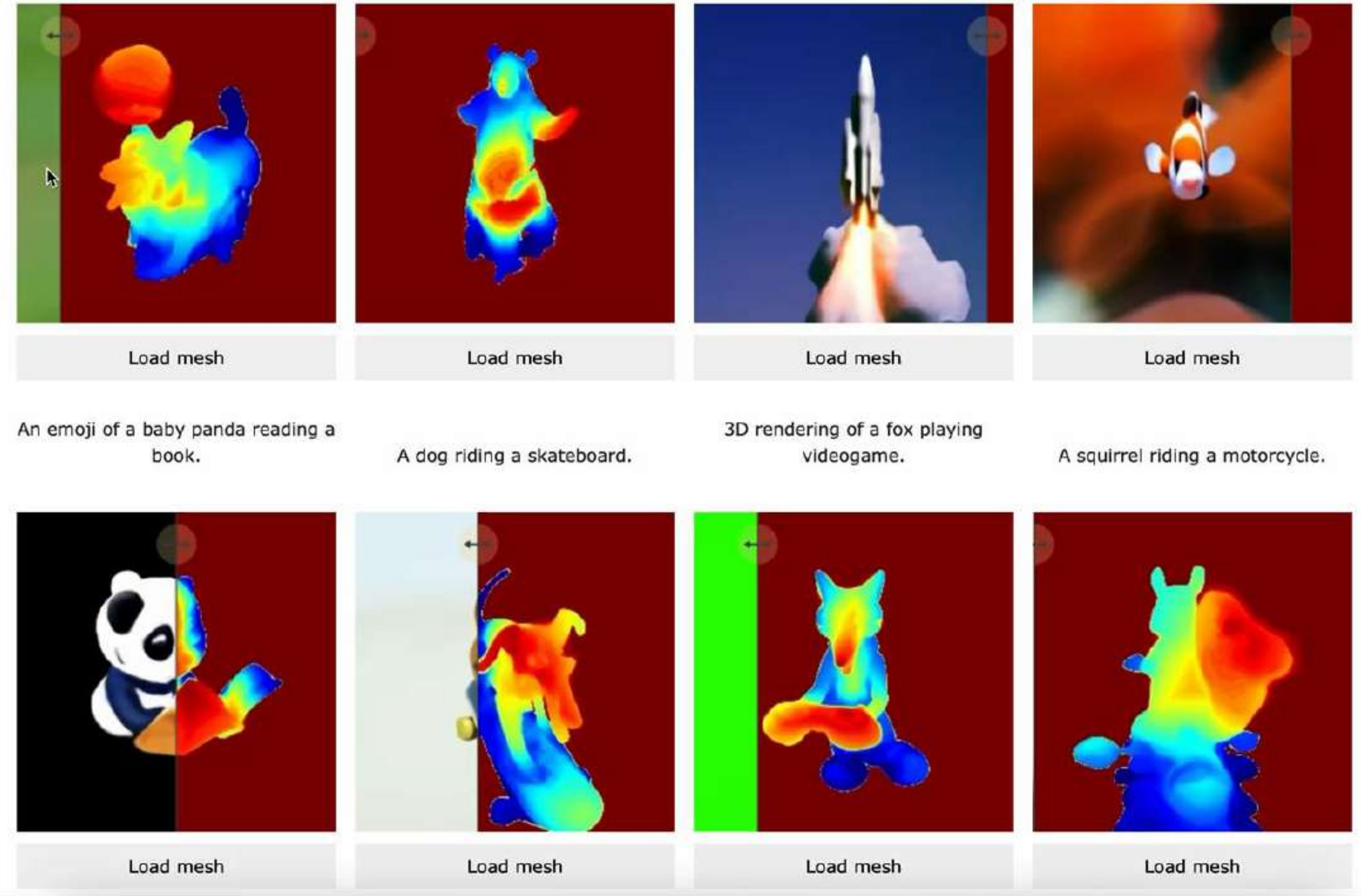
3D Models from a Single 2D Image





GeNVS (nvlabs.github.io)

Text-to-4D



An emoji of a baby panda reading a book.

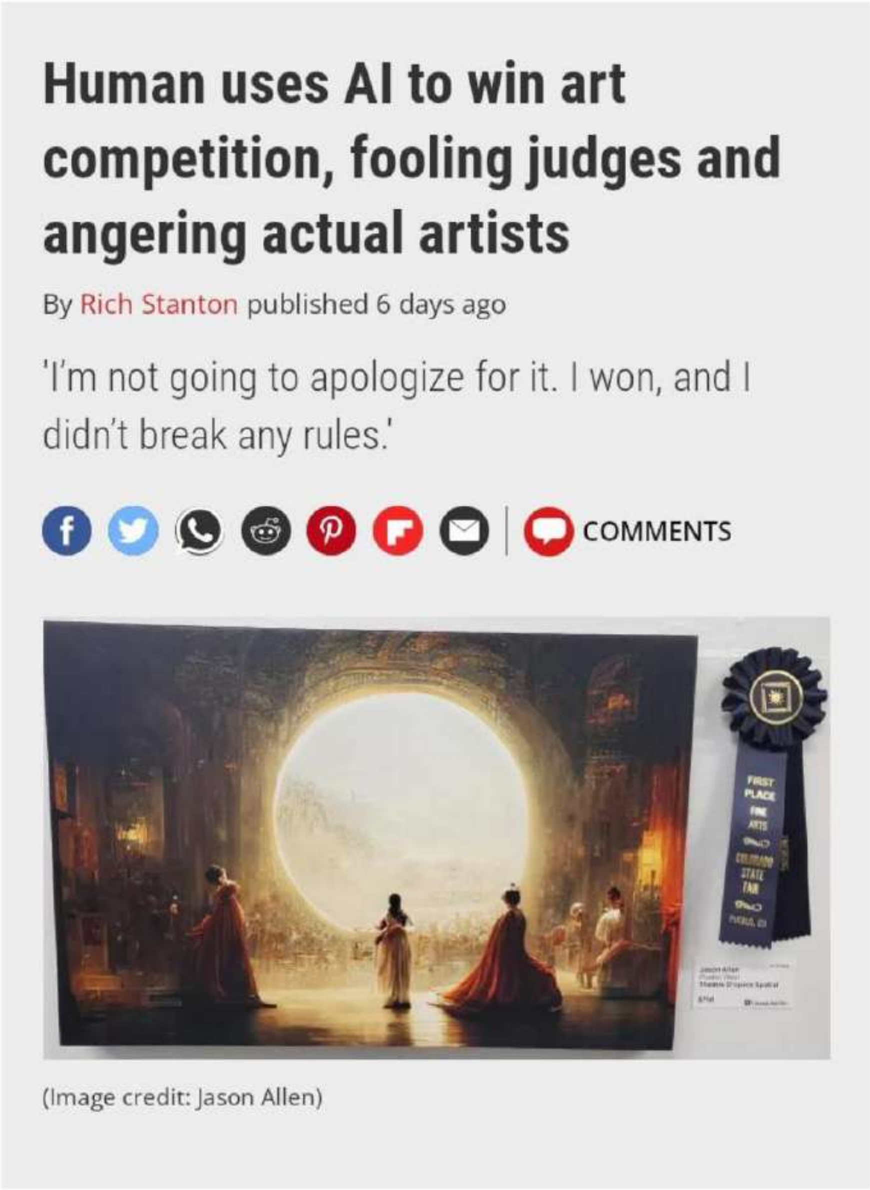
A dog riding a skateboard.

3D rendering of a fox playing videogame.

A squirrel riding a motorcycle.

[Text-To-4D Dynamic Scene Generation (make-a-video3d.github.io)](make-a-video3d.github.io)

Computer Vision is picking up steam!!!



Human uses AI to win art competition, fooling judges and angering actual artists

By Rich Stanton published 6 days ago

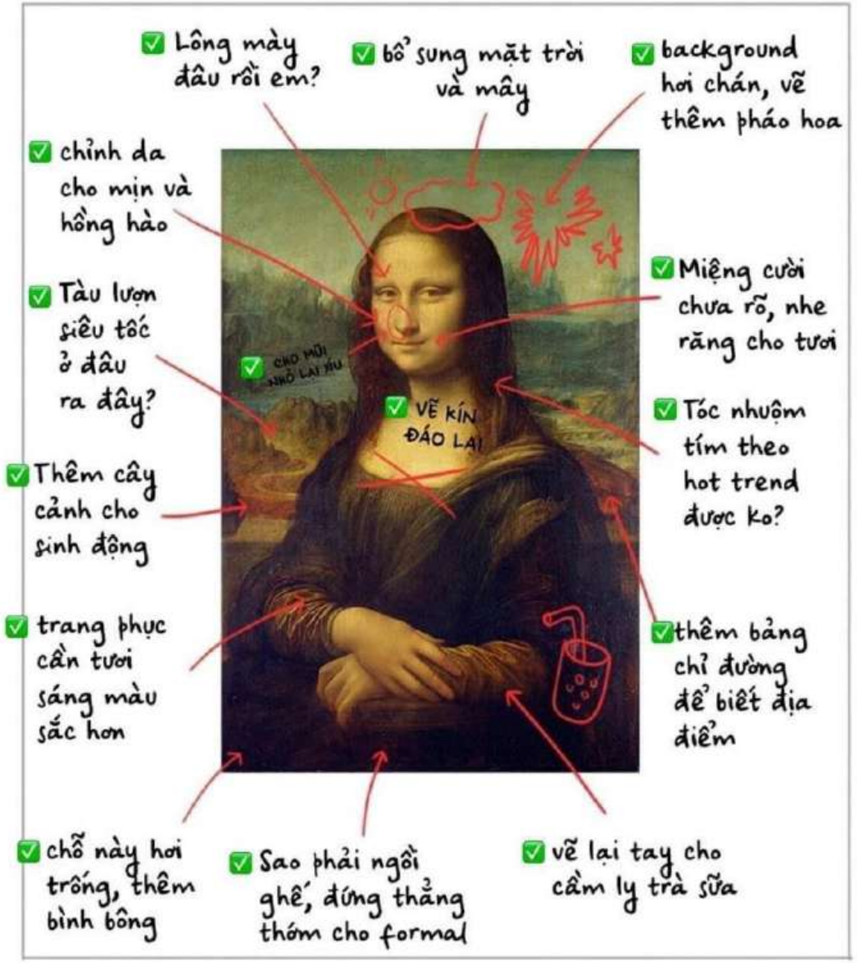
'I'm not going to apologize for it. I won, and I didn't break any rules.'

(Image credit: Jason Allen)

Hi Leonardo,

Please see attached document for feedbacks...

Many thanks.


☑ Lông mày đâu rồi em?
☑ bổ sung mặt trời và mây
☑ background hơi chán, vẽ thêm pháo hoa
☑ chỉnh da cho mịn và hồng hào
☑ Miệng cười chưa rõ, nhe răng cho tươi
☑ Tàu lượn siêu tốc ở đâu ra đây?
☑ CHO MŨI NHỎ LẠI XÍU
☑ VẼ KÍN ĐÁO LẠI
☑ Tóc nhuộm tím theo hot trend được ko?
☑ Thêm cây cảnh cho sinh động
☑ trang phục cần tươi sáng màu sắc hơn
☑ thêm bảng chỉ đường để biết địa điểm
☑ chỗ này hơi trống, thêm bình bông
☑ Sao phải ngồi ghế, đứng thẳng thớm cho formal
☑ vẽ lại tay cho cầm ly trà sữa

Thanks for listening