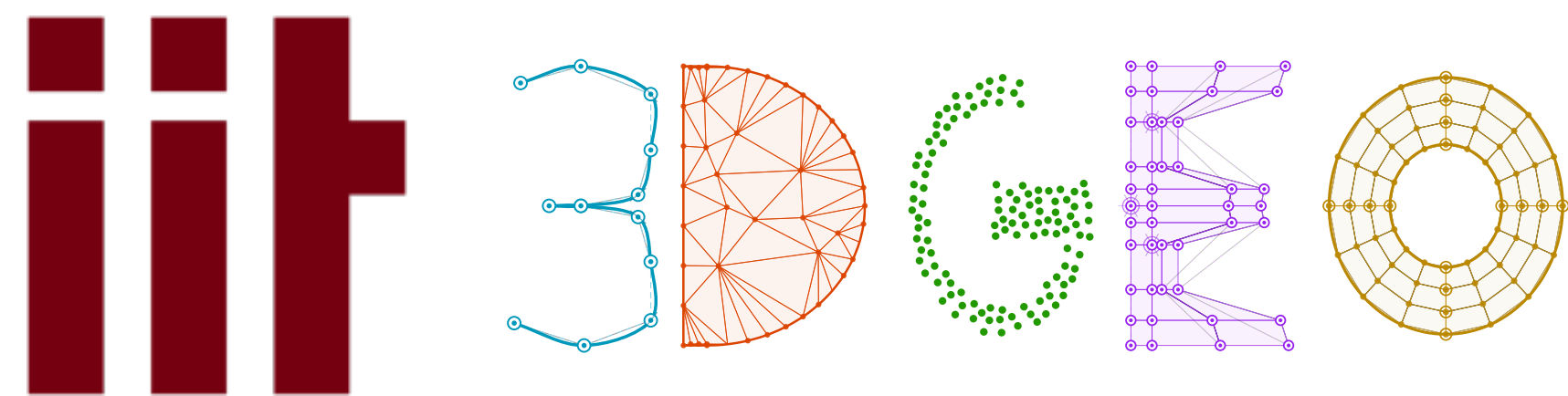
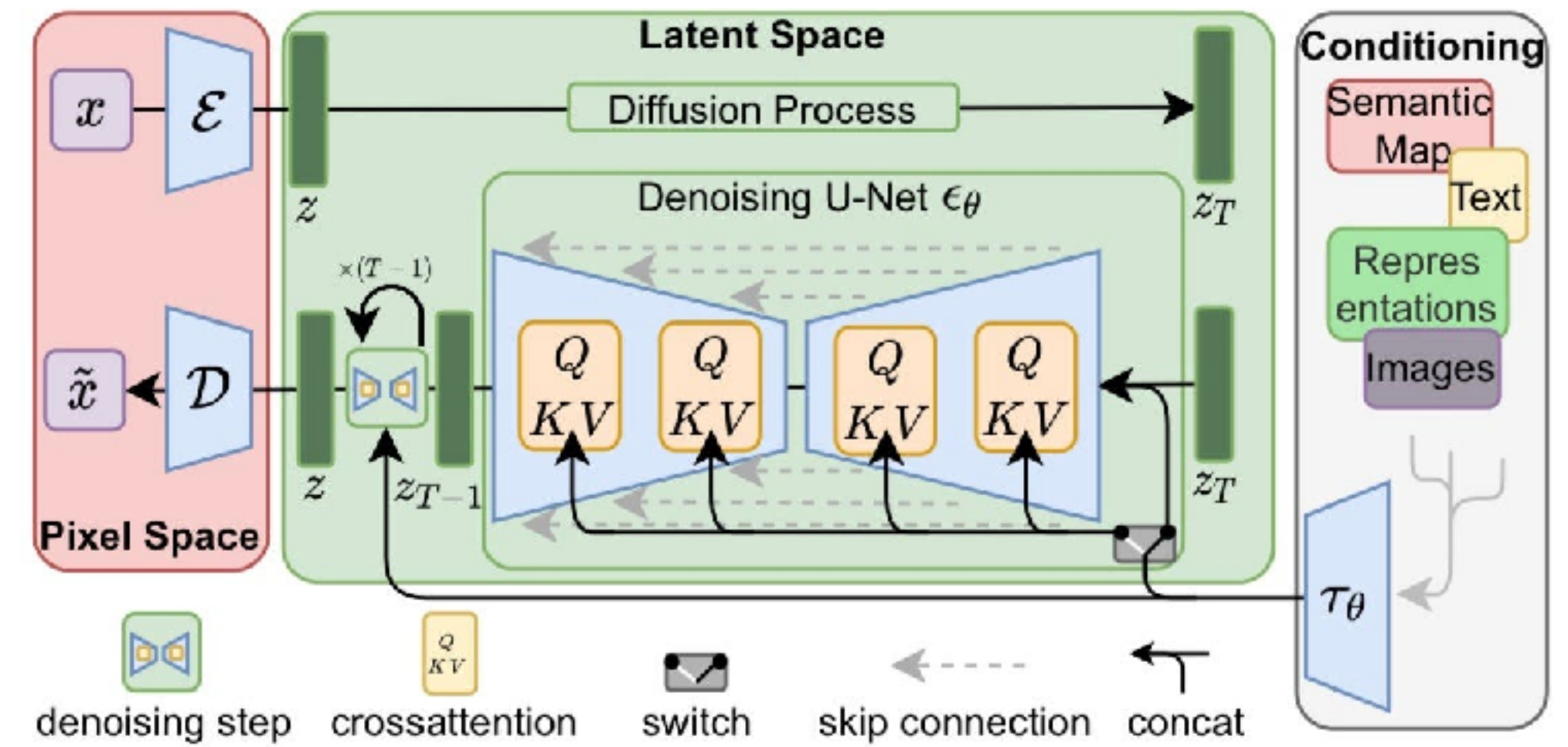


10. Előadás: Diffúziós Generálás a Gyakorlatban

Generatív AI és Inverz Módszerek a Képszintézisben
BME-VIK IIT, 2026



Dr. Vaitkus Márton

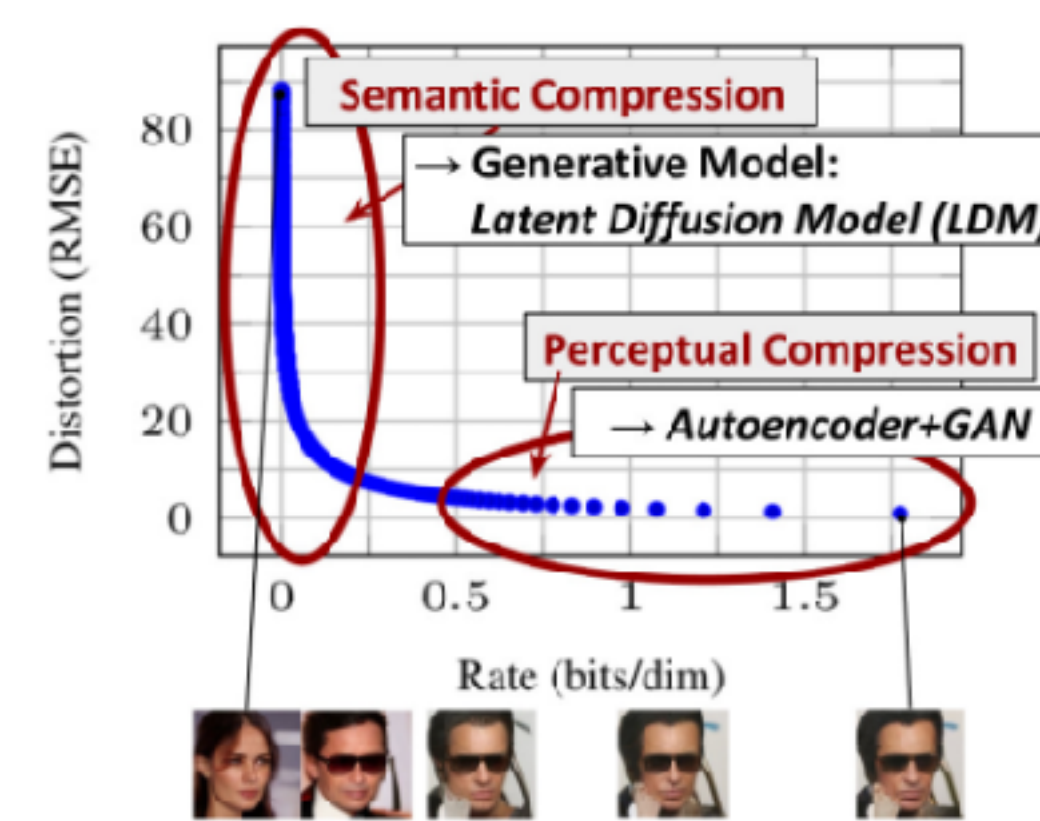
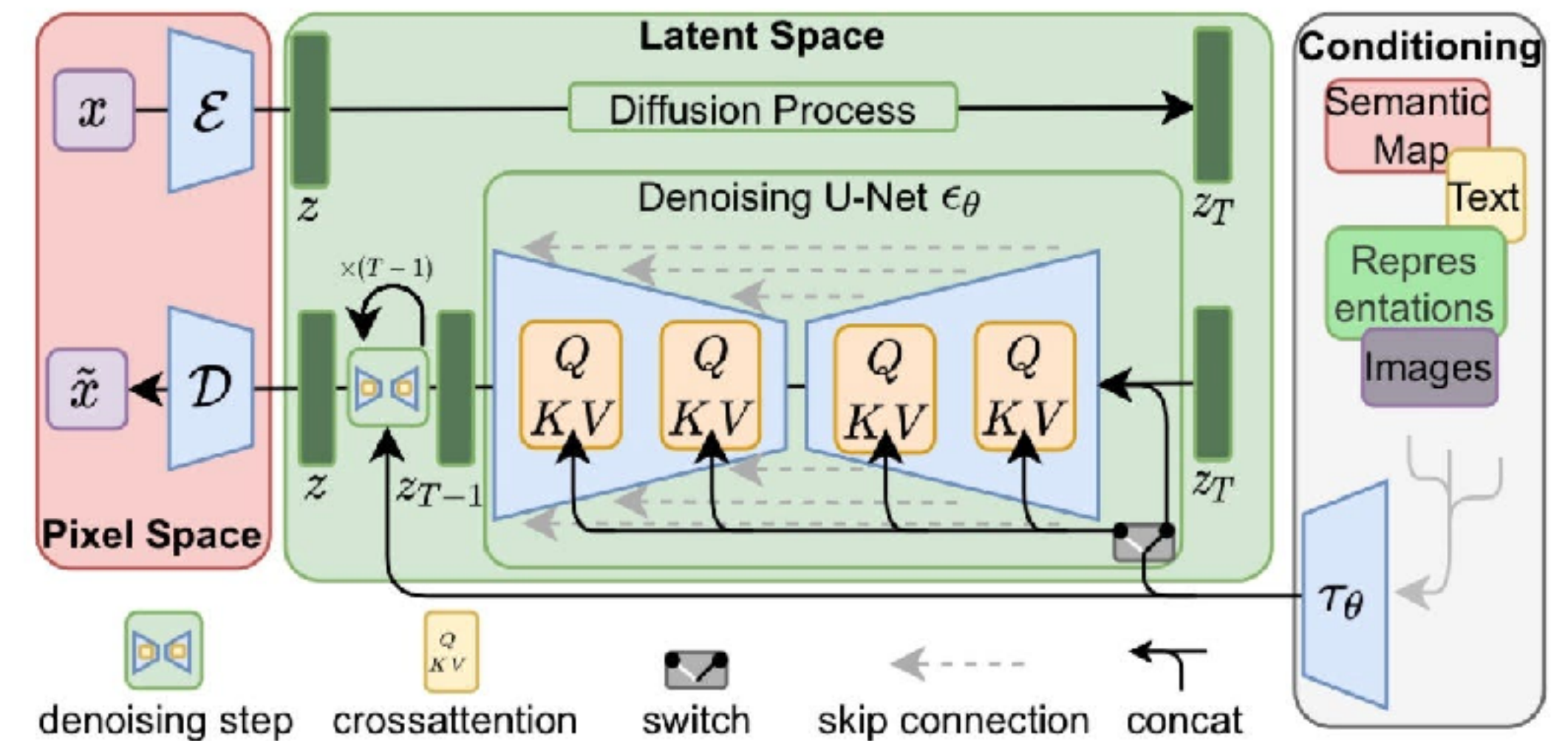


Látens Diffúzió

- Pixel térben a diffúziós modellek rosszul skálázódnak a generált felbontással...
 - Nem is biztos, hogy minden apró részletet diffúzióval akarunk rekonstruálni...
- Gyakori megoldás: diffúzió **látens térben**, majd dekódolás pixel térbe
- A Stable Diffusion nyomán tkp. az összes jelenlegi diffúziós képgenerátor látens térben dolgozik!

High-Resolution Image Synthesis with Latent Diffusion Models

Robin Rombach¹ * Andreas Blattmann¹ * Dominik Lorenz¹ Patrick Esser^ℜ Björn Ommer¹
¹Ludwig Maximilian University of Munich & IWR, Heidelberg University, Germany ^ℜRunway ML



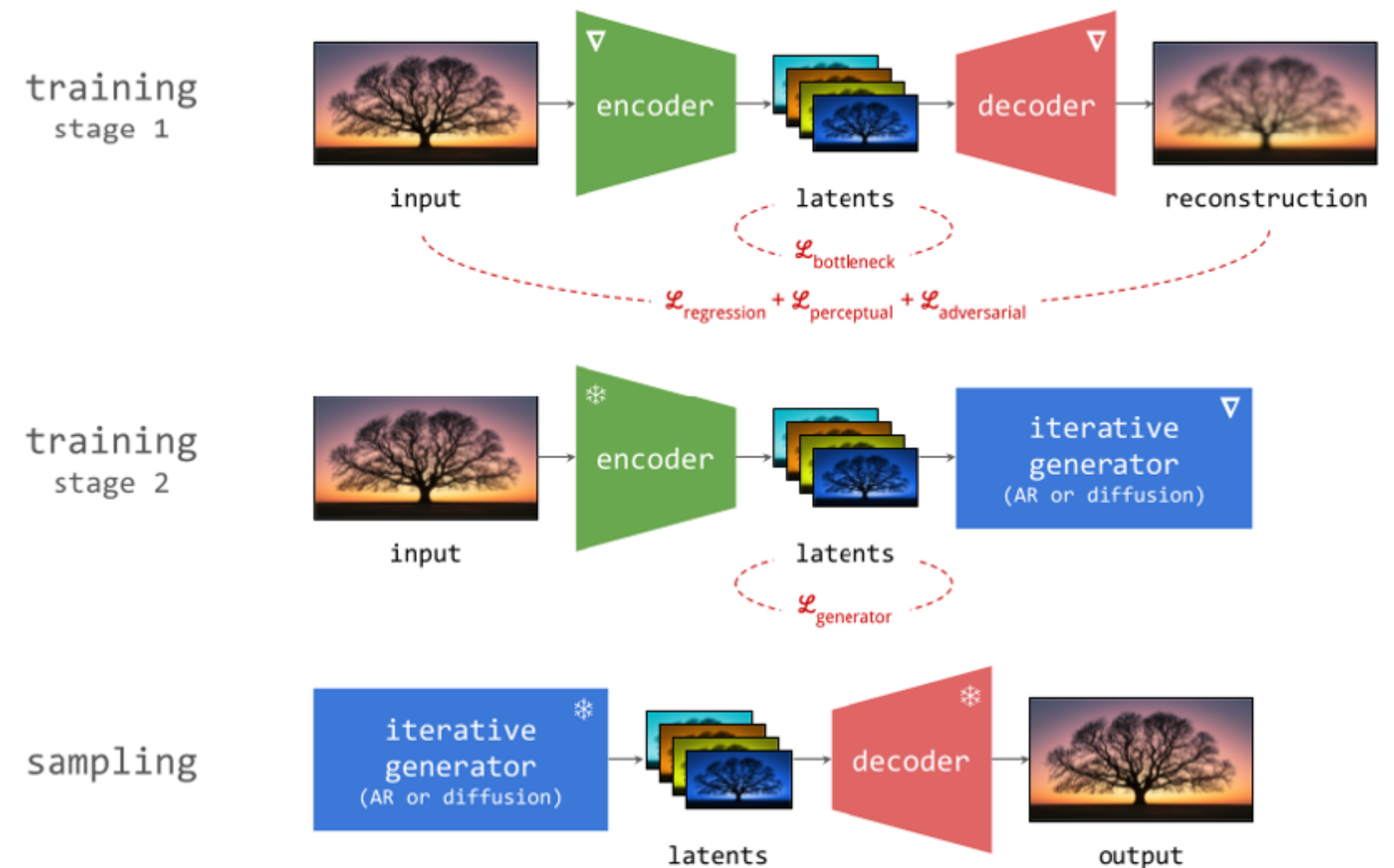
Látens Diffúzió

En-/dekóderek

- Enkóder/dekóder: tipikusan VAE (!), *külön* tanítva (VQ)GAN-jellegű loss-szal:

Rekonstrukciós hiba
+ percepciósi hiba (pl. LPIPS)
+ GAN diszkriminátor loss

- Elvileg “end-to-end” tanítás is lehetséges, de a két fázisú tanítás a bevett gyakorlat
- VAE helyett manapság gyakran használnak (multimodális) transformereket!



<https://sander.ai/2025/04/15/latents.html>

Esettanulmányok

Dall-E 2 [2022]

- “**unCLIP**”: Prompt alapján a szöveges CLIP látensbók diffúziós modell (“prior”) generál képi CLIP láténst
- A képi CLIP látensből diffúziós modell generálja a képet
- U-Net diffúziós architektúra, 3.5 milliárd paraméter

Hierarchical Text-Conditional Image Generation with CLIP Latents

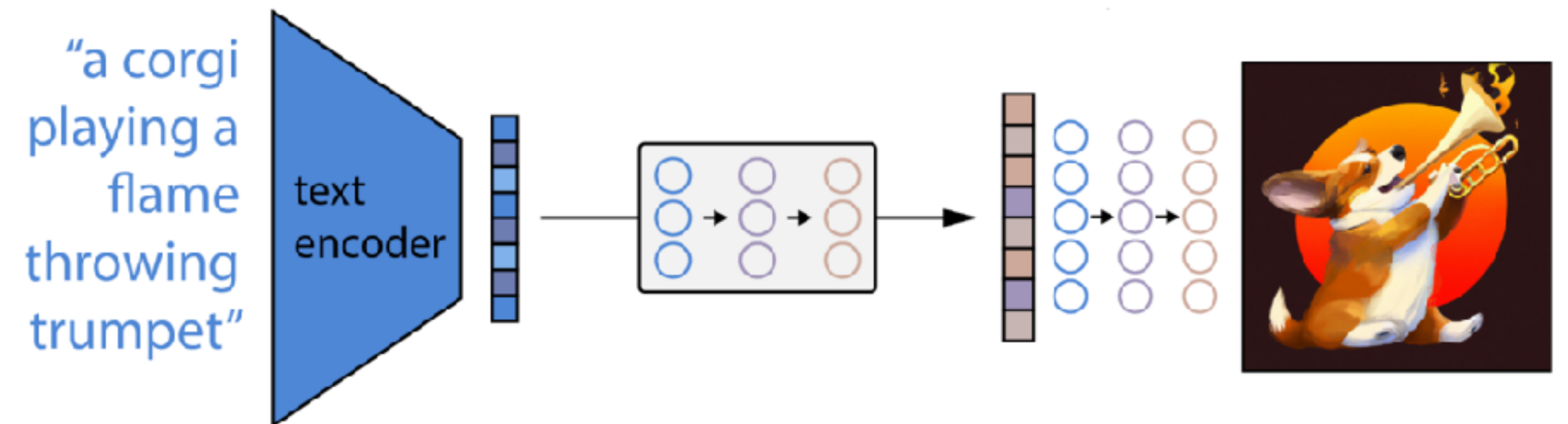
Aditya Ramesh*
OpenAI
aramesh@openai.com

Prafulla Dhariwal*
OpenAI
prafulla@openai.com

Alex Nichol*
OpenAI
alex@openai.com

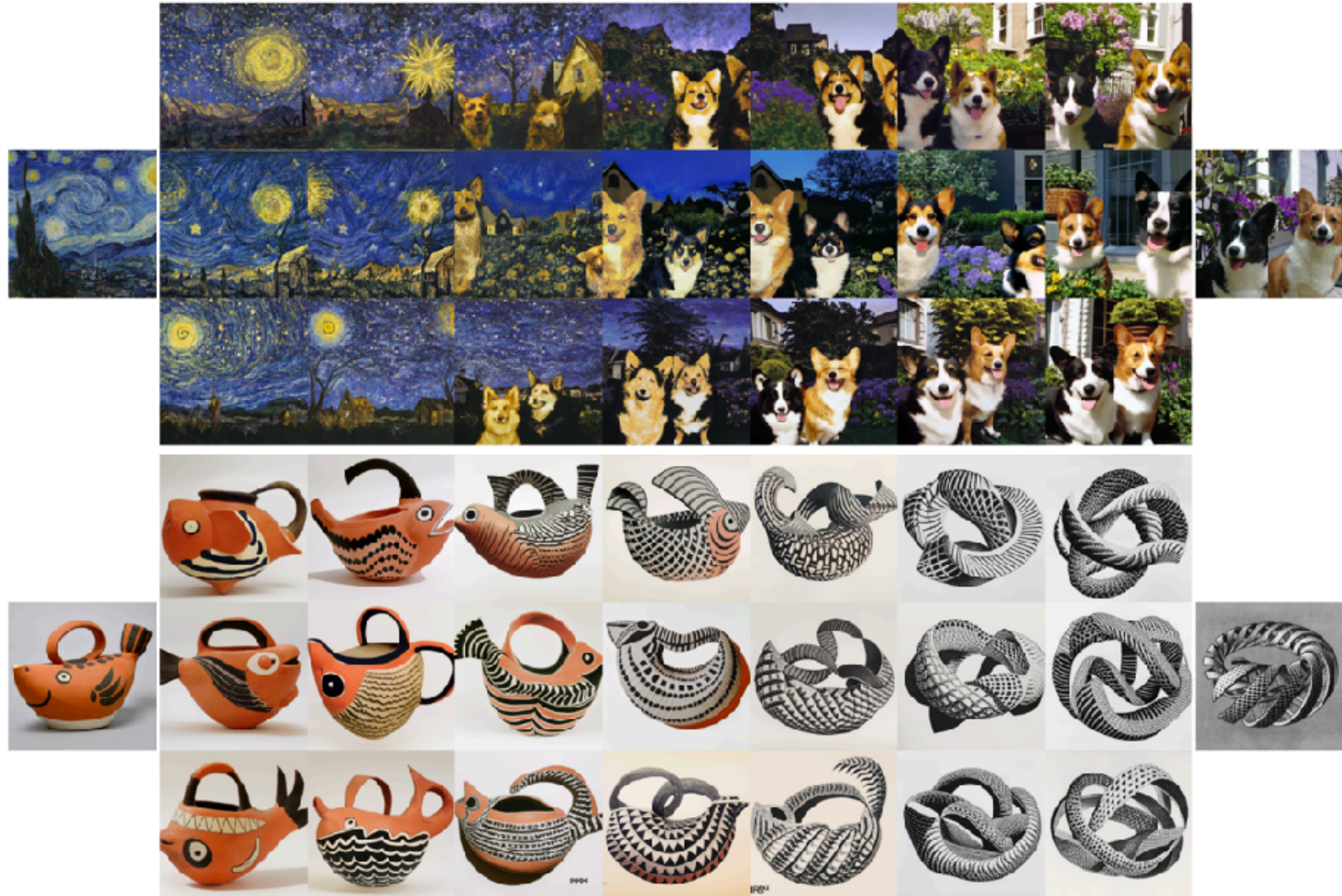
Casey Chu*
OpenAI
casey@openai.com

Mark Chen
OpenAI
mark@openai.com



Esettanulmányok

Dall-E 2 [2022]



CLIP látens interpoláció



vibrant portrait painting of Salvador Dalí with a robotic half face



a shiba inu wearing a beret and black turtleneck



a close up of a handpalm with leaves growing from it



an espresso machine that makes coffee from human souls, artstation



panda mad scientist mixing sparkling chemicals, artstation



a corgi's head depicted as an explosion of a nebula



a dolphin in an astronaut suit on saturn, artstation



a propaganda poster depicting a cat dressed as french emperor napoleon holding a piece of cheese



a teddy bear on a skateboard in times square

Esettanulmányok

Imagen [2022]

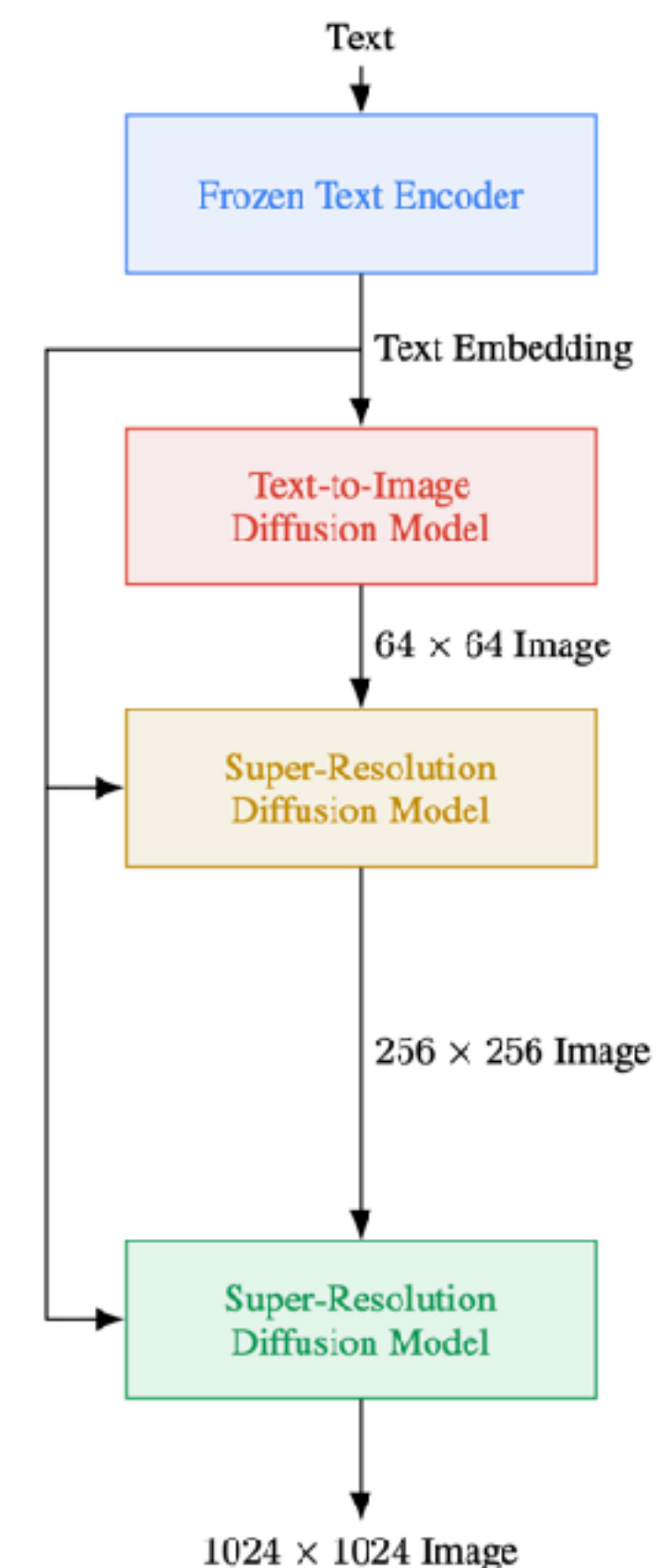
- [Google Brain, 2022]
- Szöveg enkóder: T5
- Diffúziós architektúra: U-Net
- Kaszkádosított generálás pixel-térben
 - Felskálázás speciális “efficient” U-Net architektúrájú diffúzióval
- 2B (diffúzió) + 1B (felskálázás) paraméter
- Tanítás: kb. 860 millió kép-szöveg páron

Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding

Chitwan Saharia^{*,} William Chan^{*,} Saurabh Saxena^{†,} Lala Li^{†,} Jay Whang^{†,}
Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan,
S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans,
Jonathan Ho^{†,} David J Fleet^{†,} Mohammad Norouzi^{*}

{sahariac, williamchan, mnorouzi}@google.com
{srbs, lala, jwhang, jonathanho, davidfleet}@google.com

Google Research, Brain Team
Toronto, Ontario, Canada



“A Golden Retriever dog wearing a blue
checked beret and red dotted turtleneck.”



Esettanulmányok

Imagen [2022]



“Küszöbözött” CFG



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



A photo of a Shiba Inu dog with a backpack riding a bike. It is wearing sunglasses and a beach hat.



A high contrast portrait of a very happy fuzzy panda dressed as a chef in a high end kitchen making dough. There is a painting of flowers on the wall behind him.



Teddy bears swimming at the Olympics 400m Butterfly event.



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



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



A brain riding a rocketship heading towards the moon.



A dragon fruit wearing karate belt in the snow.



A strawberry mug filled with white sesame seeds. The mug is floating in a dark chocolate sea.

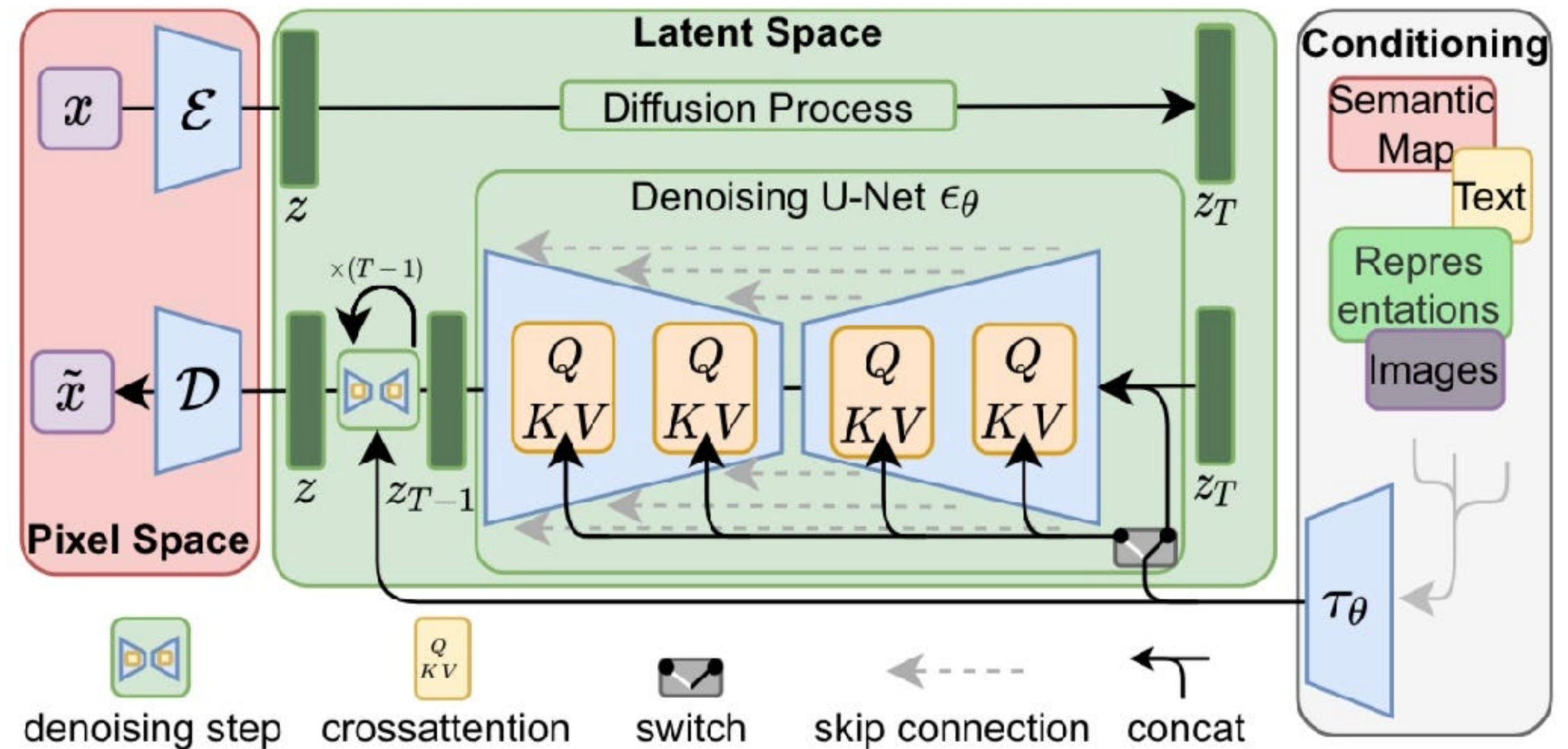
Esettanulmányok

Stable Diffusion 1.5 [2022]

High-Resolution Image Synthesis with Latent Diffusion Models

Robin Rombach^{1*} Andreas Blattmann^{1*} Dominik Lorenz¹ Patrick Esser^ℝ Björn Ommer¹
¹Ludwig Maximilian University of Munich & IWR, Heidelberg University, Germany ^ℝRunway ML
<https://github.com/CompVis/latent-diffusion>

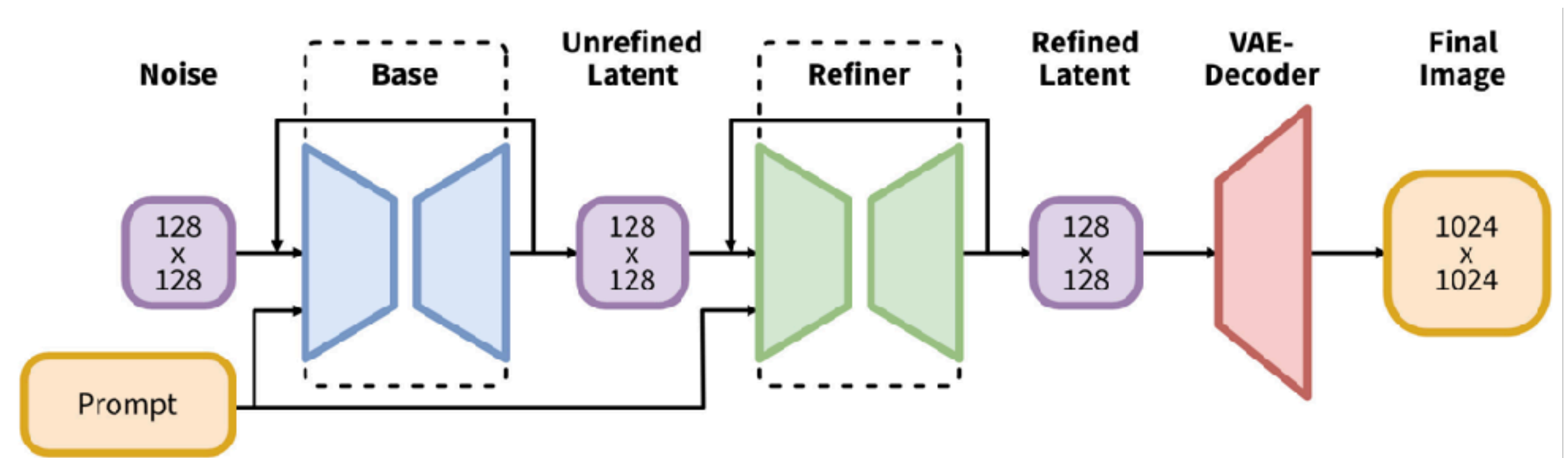
- Diffúzió látens térben
- VAE enkóder külön tanítva, GAN loss-szal
- Szöveg encoder: CLIP
- Diffúziós architektúra: U-Net — ResNet + attention rétegek
- 860M paraméter
- Tanító adat: LAION-5B
- Tanítás: 256 A100 GPU, 150k GPU-óra, becsült költség: kb. 600k USD



Esettanulmányok

SDXL [2023]

- Több lépéses generálás
- Nagyobb U-Net modell (2,6B)
- Nagyobb kontextus ablak
- Több szöveg-enkóder kombinációja



Model	<i>SDXL</i>	SD 1.4/1.5	SD 2.0/2.1
# of UNet params	2.6B	860M	865M
Transformer blocks	[0, 2, 10]	[1, 1, 1, 1]	[1, 1, 1, 1]
Channel mult.	[1, 2, 4]	[1, 2, 4, 4]	[1, 2, 4, 4]
Text encoder	CLIP ViT-L & OpenCLIP ViT-bigG	CLIP ViT-L	OpenCLIP ViT-H
Context dim.	2048	768	1024
Pooled text emb.	OpenCLIP ViT-bigG	N/A	N/A

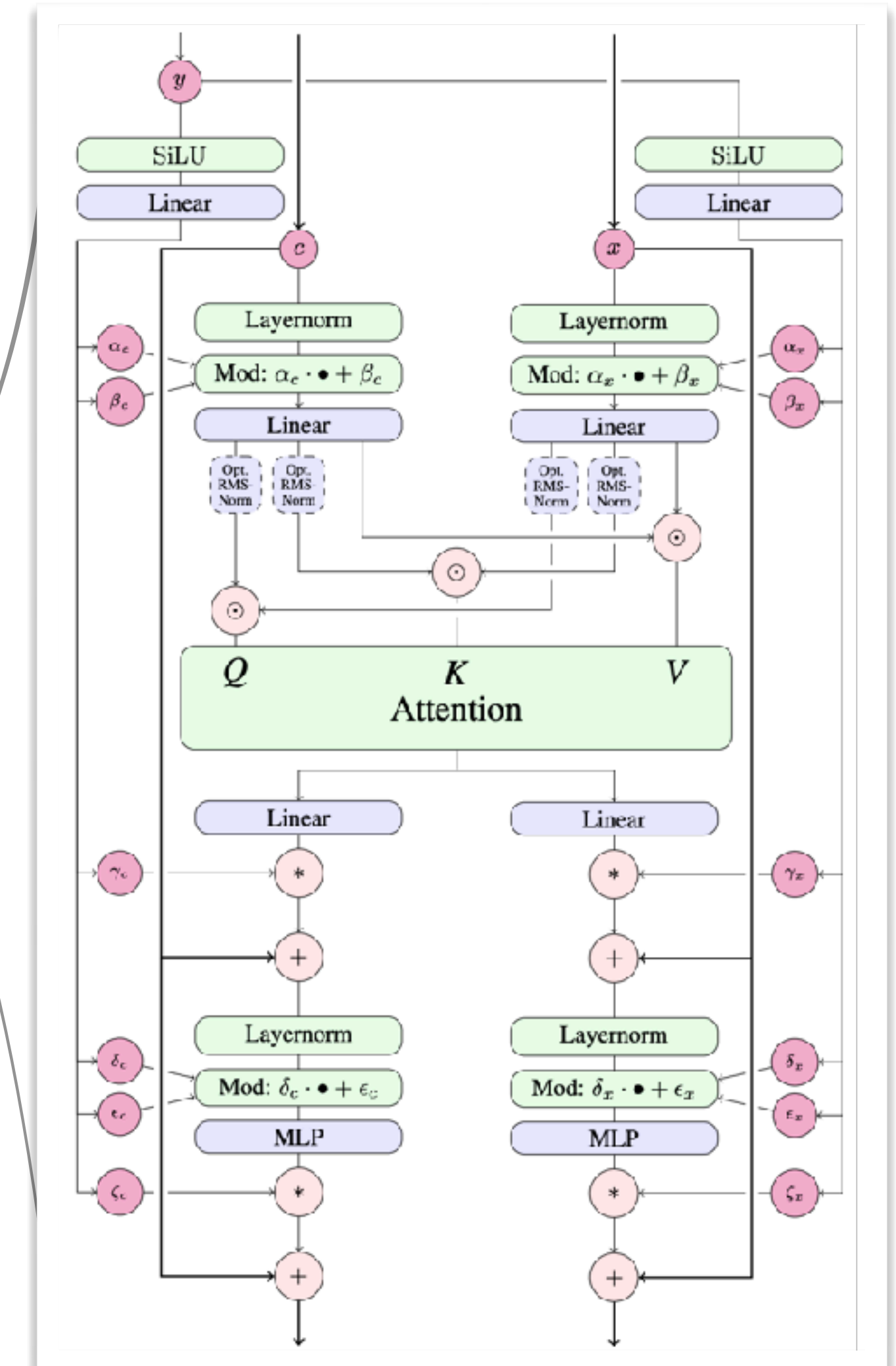
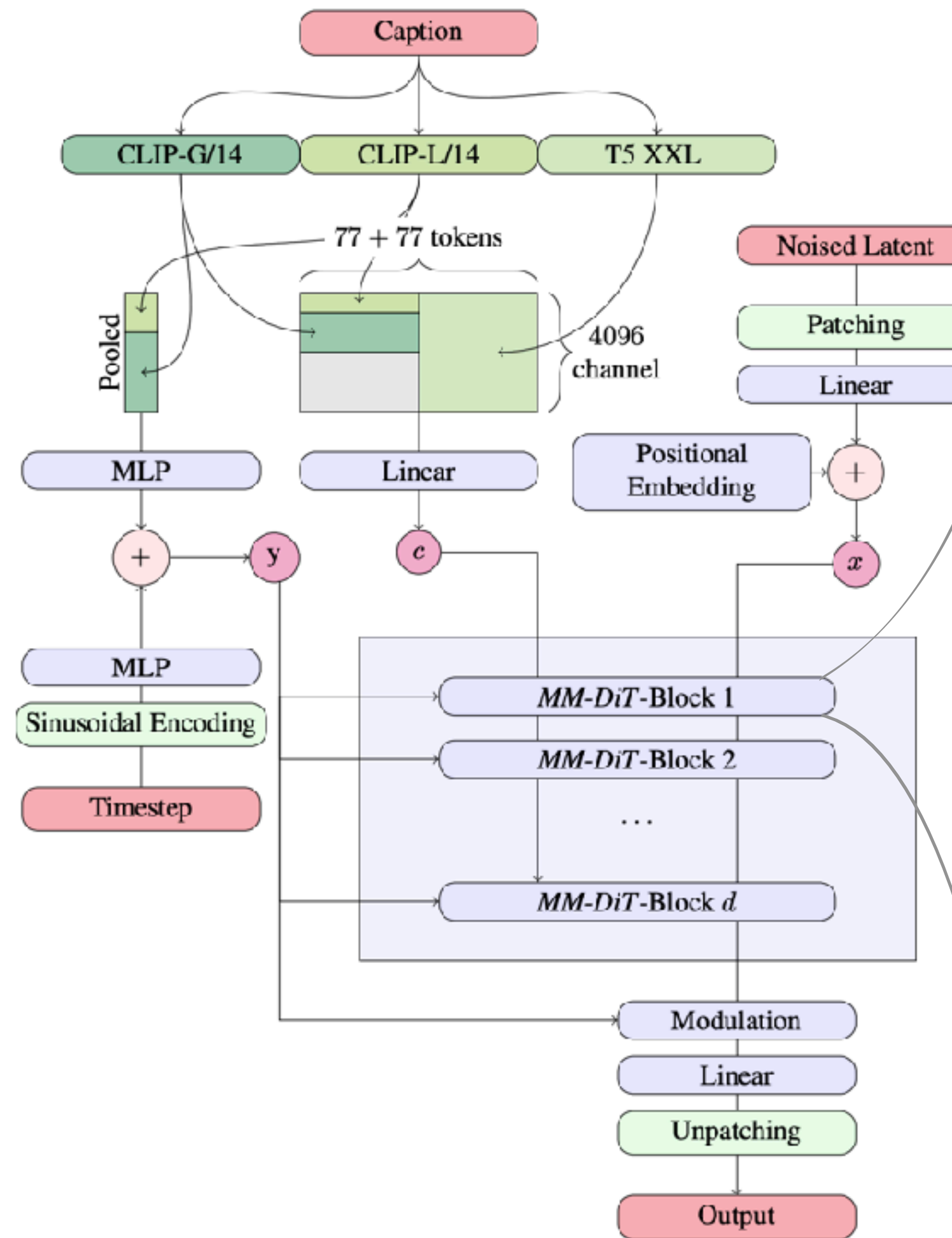
Esettanulmányok

Stable Diffusion 3 [2024]

- Diffúziós architektúra: multi-modális transzformer (MM-DiT)
- Közös attention szöveg és (látens) zajos képi tokenekre
- “Dual stream”: külön attention mátrixok a kétféle modalitásra
- Tanítás: flow matching (“rectified flow”) látens térben
- 3 szöveg enkóder kombinációja
- 8B paraméter

Scaling Rectified Flow Transformers for High-Resolution Image Synthesis

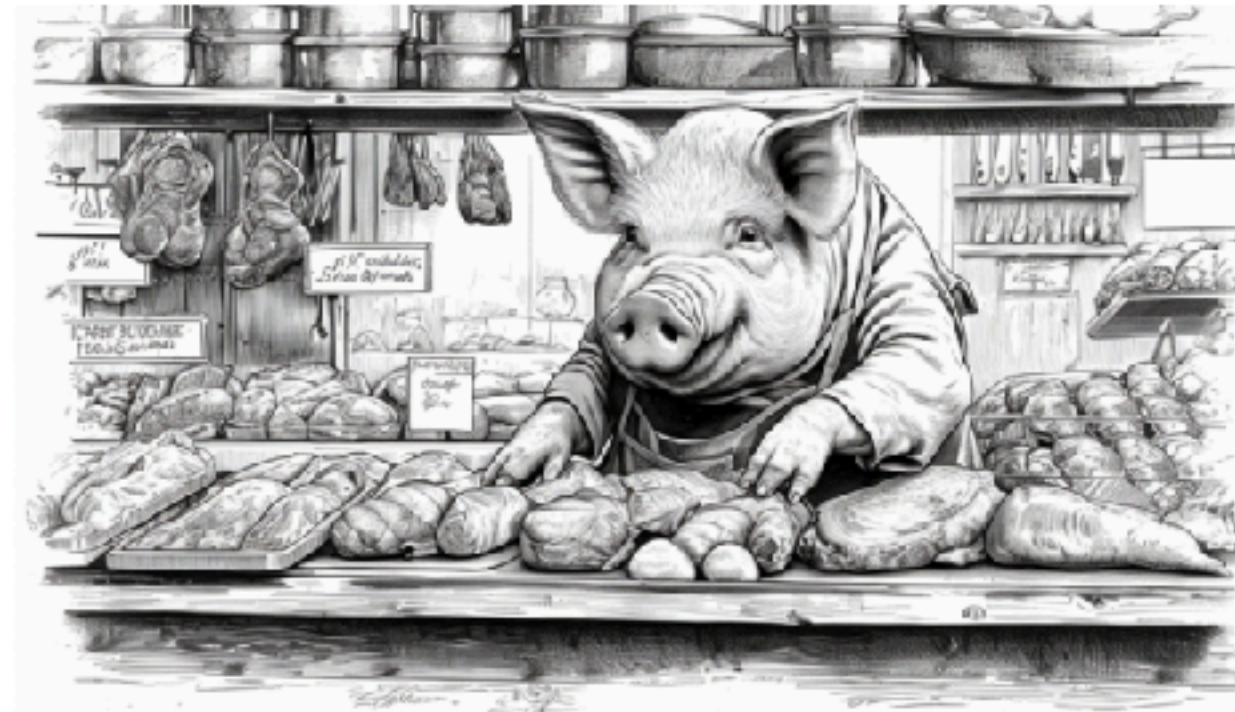
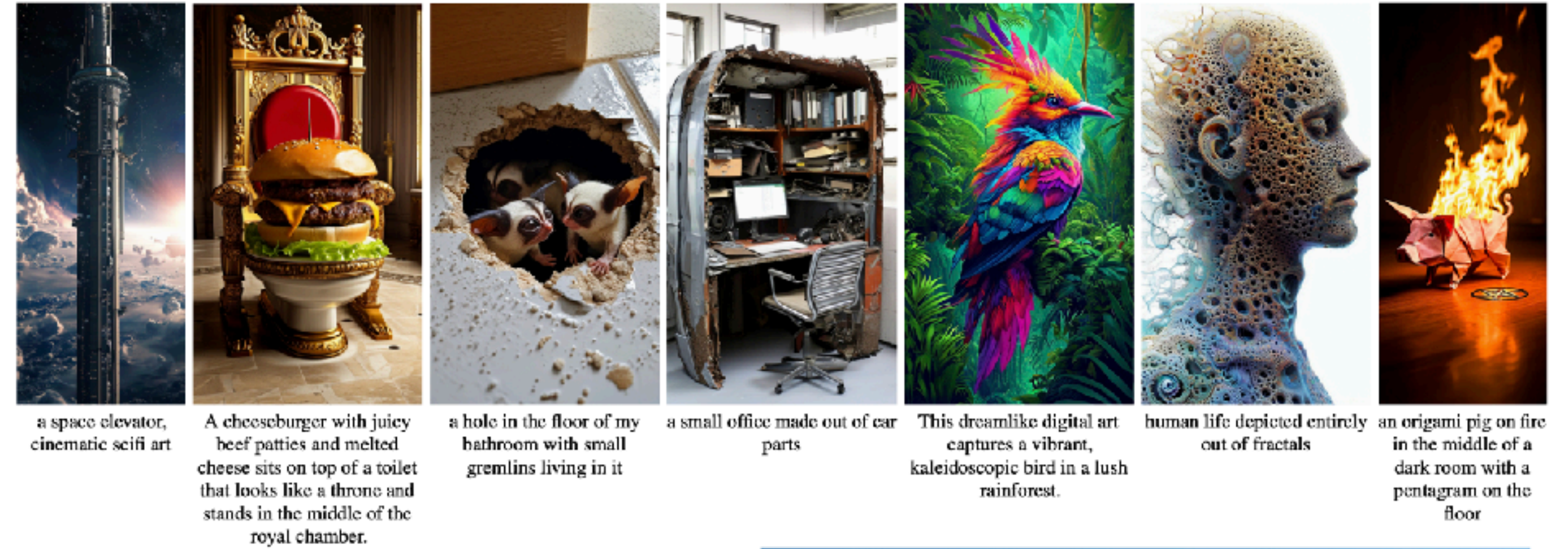
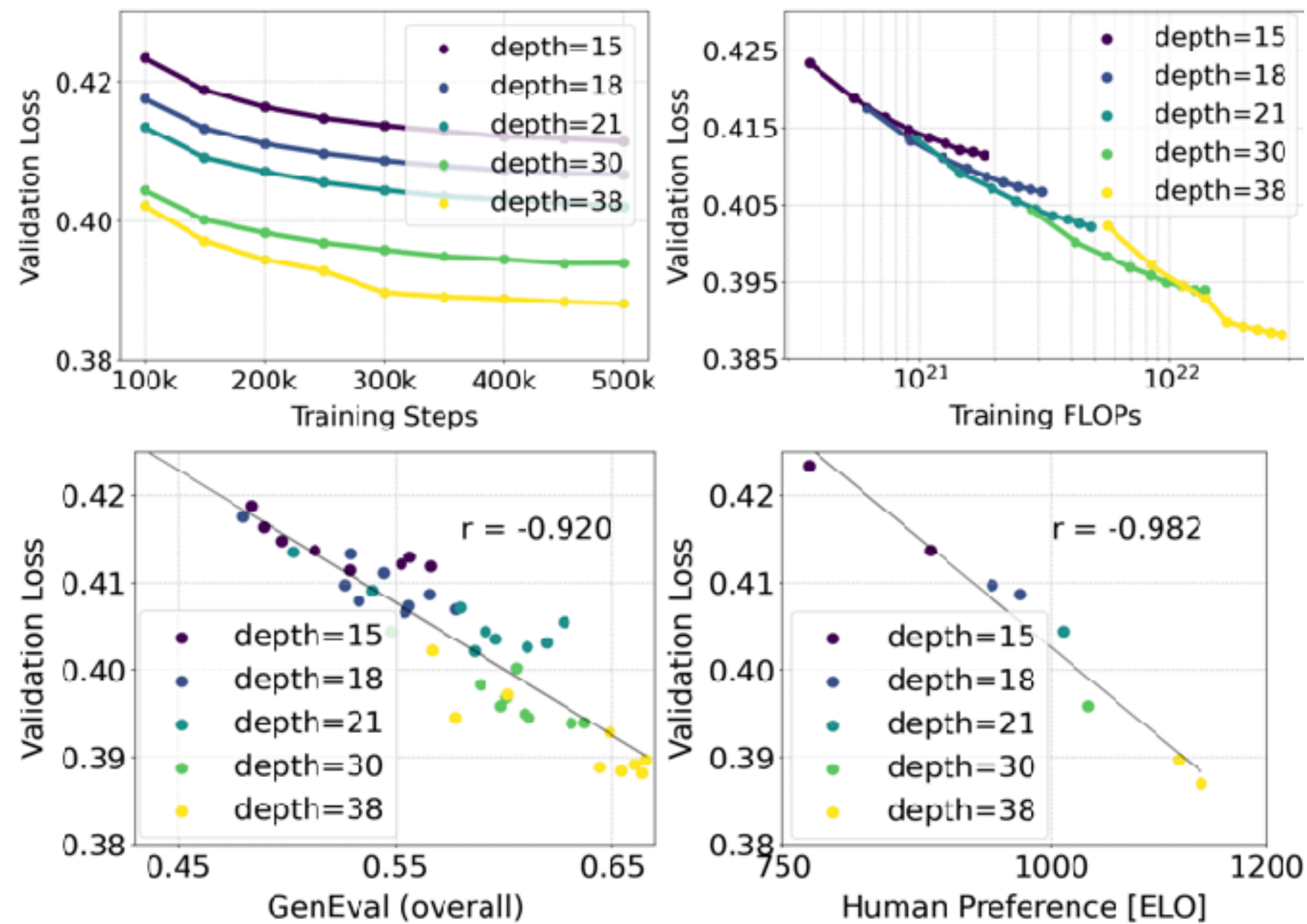
Patrick Esser* Sumith Kulal Andreas Blattmann Rahim Entezari Jonas Müller Harry Saini Yam Levi
 Dominik Lorenz Axel Sauer Frederic Boesel Dustin Podell Tim Dockhorn Zion English
 Kyle Lacey Alex Goodwin Yannik Marek Robin Rombach*
 Stability AI



Esettanulmányok

Stable Diffusion 3 [2024]

- Tanítás: 1 milliárd kép előtanításra + 30 millió finomhangolás + 3 millió RLHF
- A paraméterek számával jól skálázódik



Detailed pen and ink drawing of a happy pig butcher selling meat in its shop.



a massive alien space ship that is shaped like a pretzel.



A kangaroo holding a beer, wearing ski goggles and passionately singing silly songs.



An entire universe inside a bottle sitting on the shelf at walmart on sale.



A cheeseburger surfing the vibe wave at night



A swamp ogre with a pearl earring by Johannes Vermeer



A car made out of vegetables.



heat death of the universe, line art

Esettanulmányok

Stable Diffusion 3 [2024]



A crab made of cheese on a plate



Dystopia of thousand of workers picking cherries and feeding them into a machine that runs on steam and is as large as a skyscraper. Written on the side of the machine: "SD3 Paper"



translucent pig, inside is a smaller pig.



Film still of a long-legged cute big-eye anthropomorphic cheeseburger wearing sneakers relaxing on the couch in a sparsely decorated living room.



A whimsical and creative image depicting a hybrid creature that is a mix of a waffle and a hippopotamus. This imaginative creature features the distinctive, bulky body of a hippo, but with a texture and appearance resembling a golden-brown, crispy waffle. The creature might have elements like waffle squares across its skin and a syrup-like sheen. It's set in a surreal environment that playfully combines a natural water habitat of a hippo with elements of a breakfast table setting, possibly including oversized utensils or plates in the background. The image should evoke a sense of playful absurdity and culinary fantasy.

Esettanulmányok

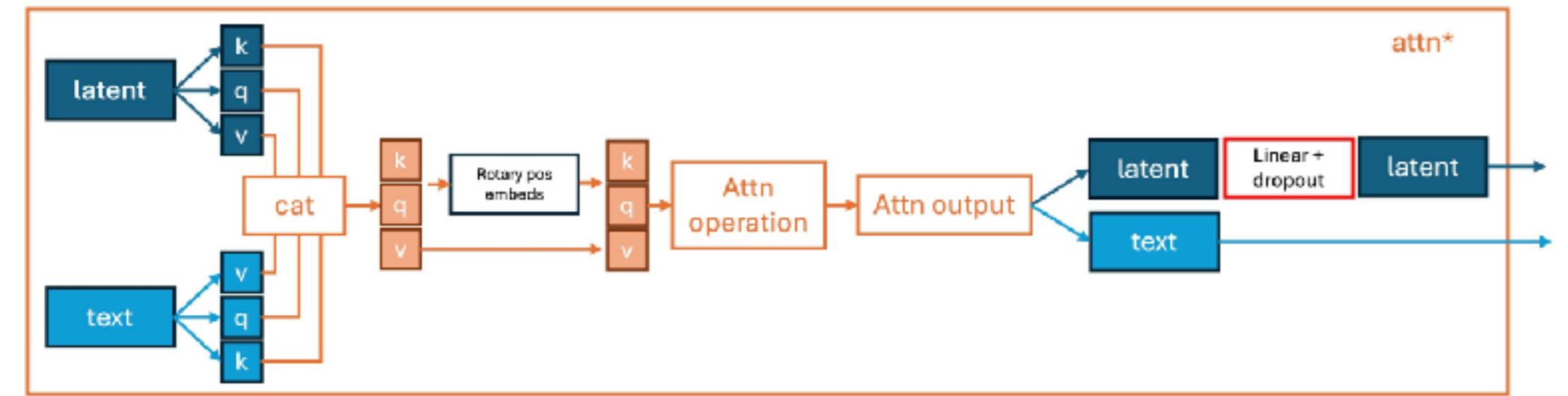
FLUX.1 [2024]

Demystifying Flux Architecture

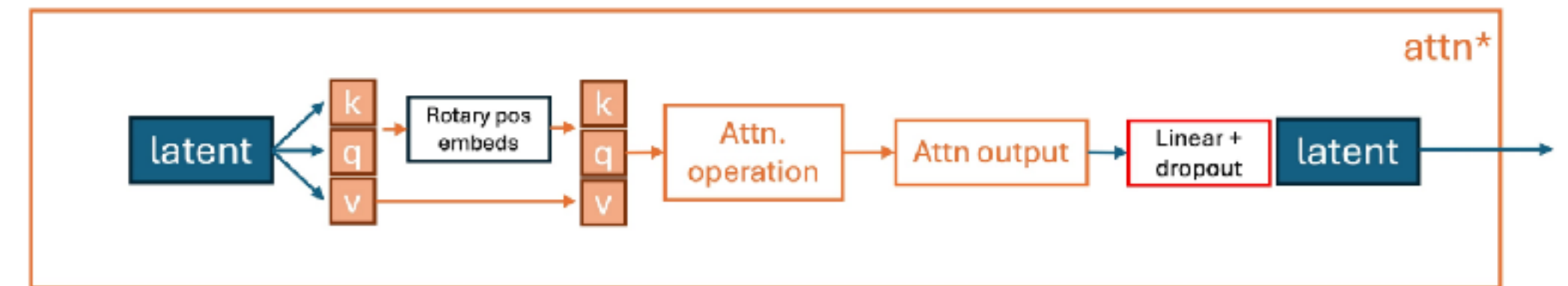
Or Greenberg^{1,2}

¹Hebrew University of Jerusalem, Israel

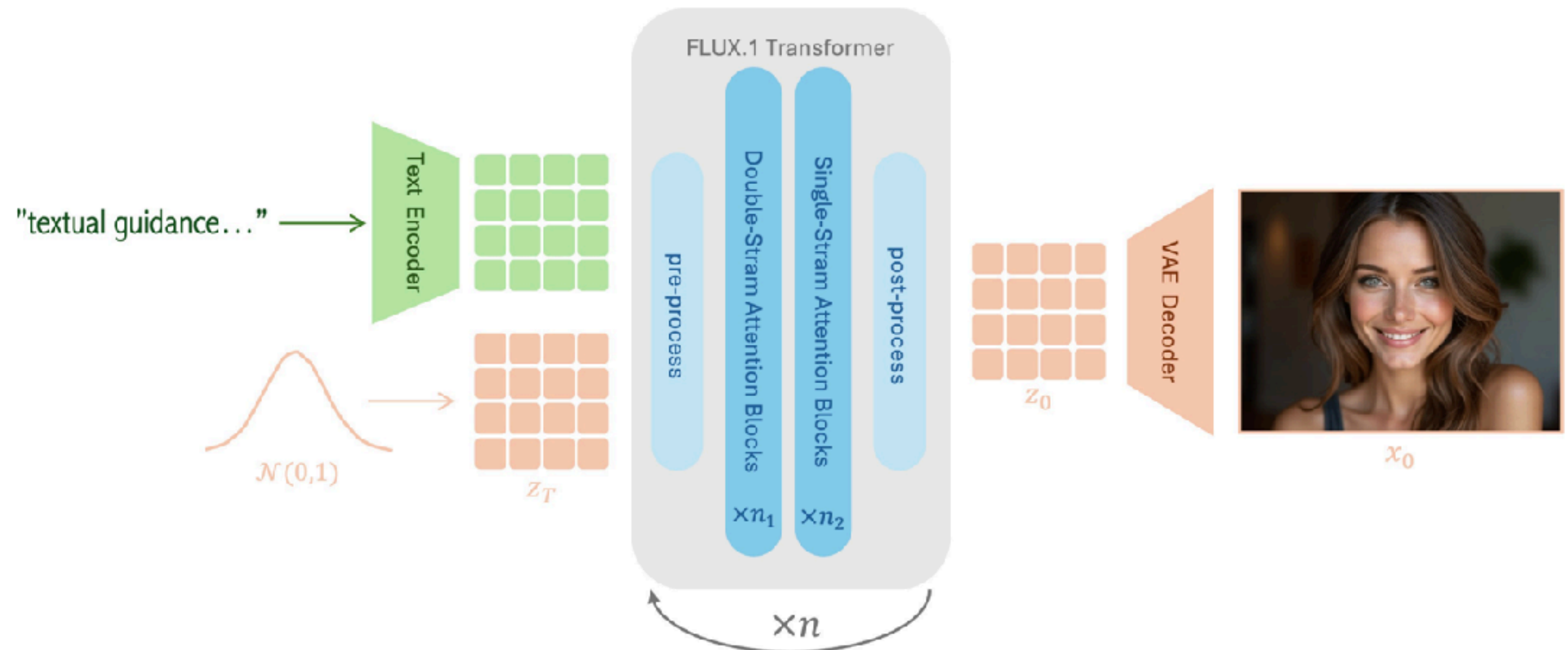
²General Motors R&D, SDVR



Dual-stream MMDIT (a la SD3)



Single-stream MMDIT



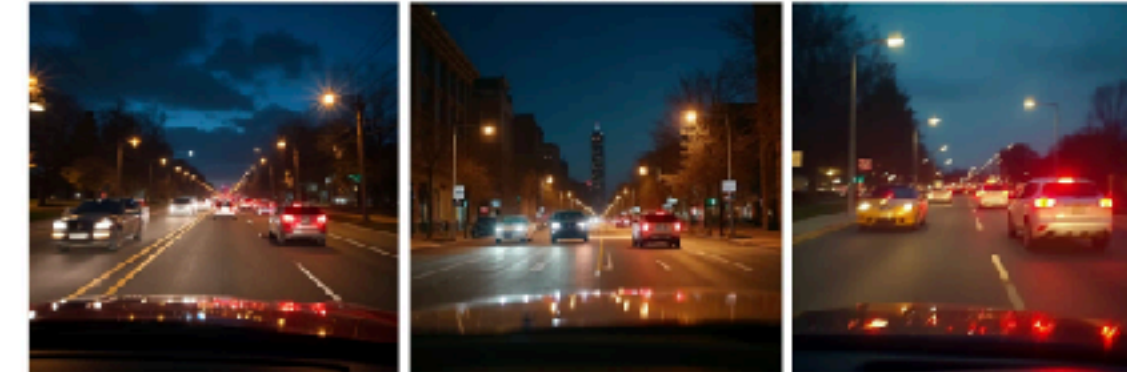
Esettanulmányok

FLUX.1 [2024]

Cars driving down the street on a **clear day**, taken from a vehicle's front dashcam. natural, realistic



Cars driving down the street on a **clear night**, taken from a vehicle's front dashcam. natural, realistic



Cars driving down the street on a **foggy day**, taken from a vehicle's front dashcam. natural, realistic



Stable Diffusion 2.1



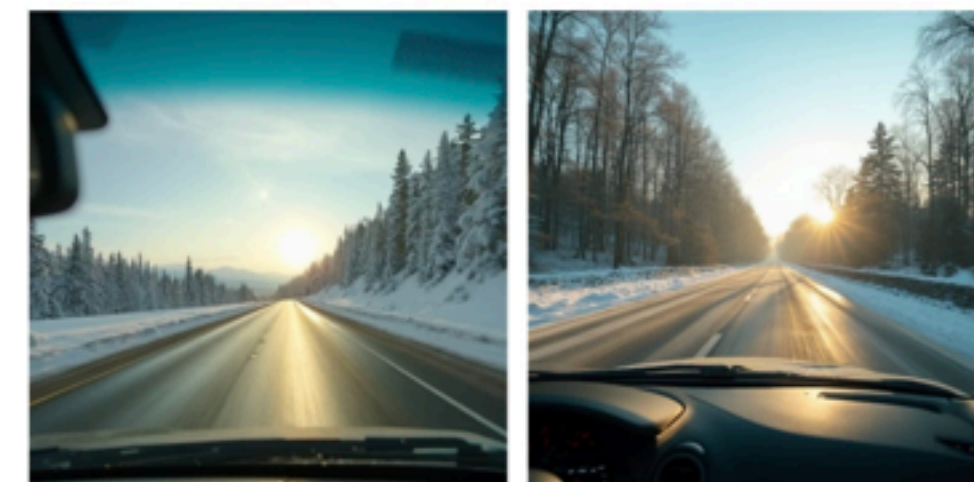
“... a slightly snowy day...”



“...a heavily snowy day...”



“...sunny day after a snowy period...”



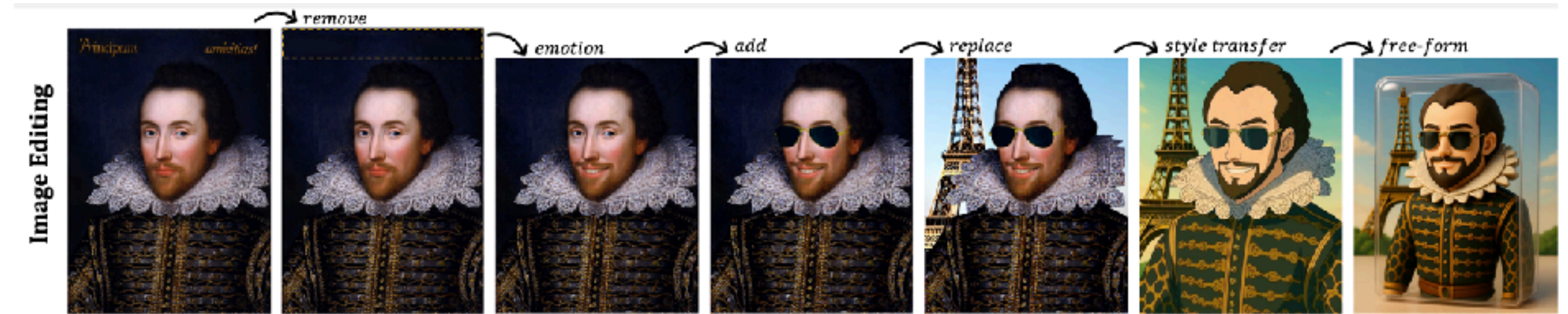
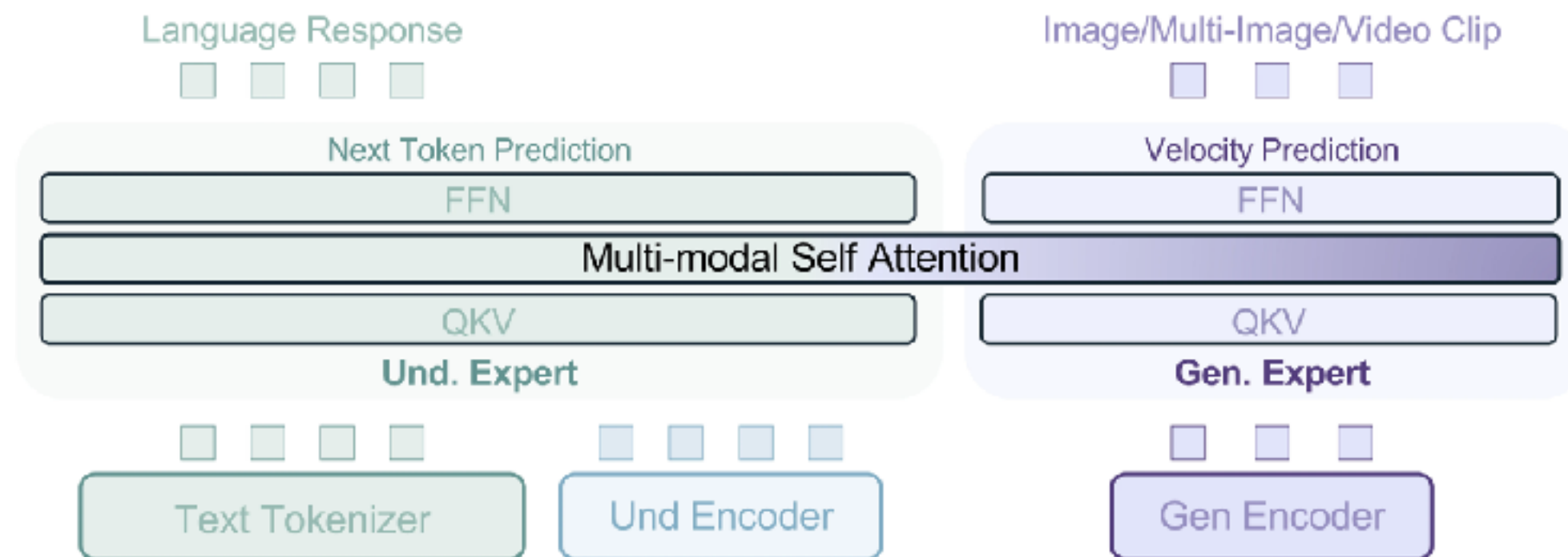
Esettanulmányok

BAGEL [2025]



Emerging Properties in Unified Multimodal Pretraining

Chaorui Deng^{*1}, Deyao Zhu^{*1}, Kunchang Li^{*2‡}, Chenhui Gou^{*3‡}, Feng Li^{*4‡}
 Zeyu Wang^{5‡}, Shu Zhong¹, Weihao Yu¹, Xiaonan Nie¹, Ziang Song¹, Guang Shi^{1§}
 Haoqi Fan^{*†}



Generation & Editing w/ Thinking

A car made of small cars.

<think> The model should generate an image where numerous small cars are combined to form the shape and structure of a larger car. The refined prompt is: A highly detailed and creative image showcasing a large car composed entirely of numerous small cars, meticulously arranged to replicate the shape and structure of a full-sized vehicle. The small cars are tightly interlocked, creating a seamless and realistic exterior with accurate proportions and details. The setting is neutral and clean, with soft, diffused lighting to emphasize the intricacy of the composition without any distractions... </think>



He is surrounded by the national flower of the country that invented sushi.

<think> The user wants to replace the statue's surroundings with a scene full of cherry blossoms, while keeping the main body of the statue unchanged. First, I have to make sure the details of the statue, such as muscle lines, facial expressions, and clothing folds, remain clear and complete. Next, cherry blossom elements need to be added around the statue to make it conform to the setting of 'the national flower of the country that invented sushi'. The cherry blossoms... </think>



Could you display the sculpture that takes after this design?

<think> The question_image shows a design of an octopus created by a child named Blake W. Torney, aged 10, titled "The Octopus." The description mentions that the sculpture was made in 2008 by Dante Marioni, an American glass artist born in 1964, using blown and hot-sculpted glass with applied bits. The sculpture is part... The question_text asks to display the sculpture that takes after this design. Based on the description, the final answer image should feature the actual glass sculpture created by Dante Marioni, which reflects the design of the child's drawing. The sculpture should have a green body with red tentacles, similar to the child's depiction... </think>

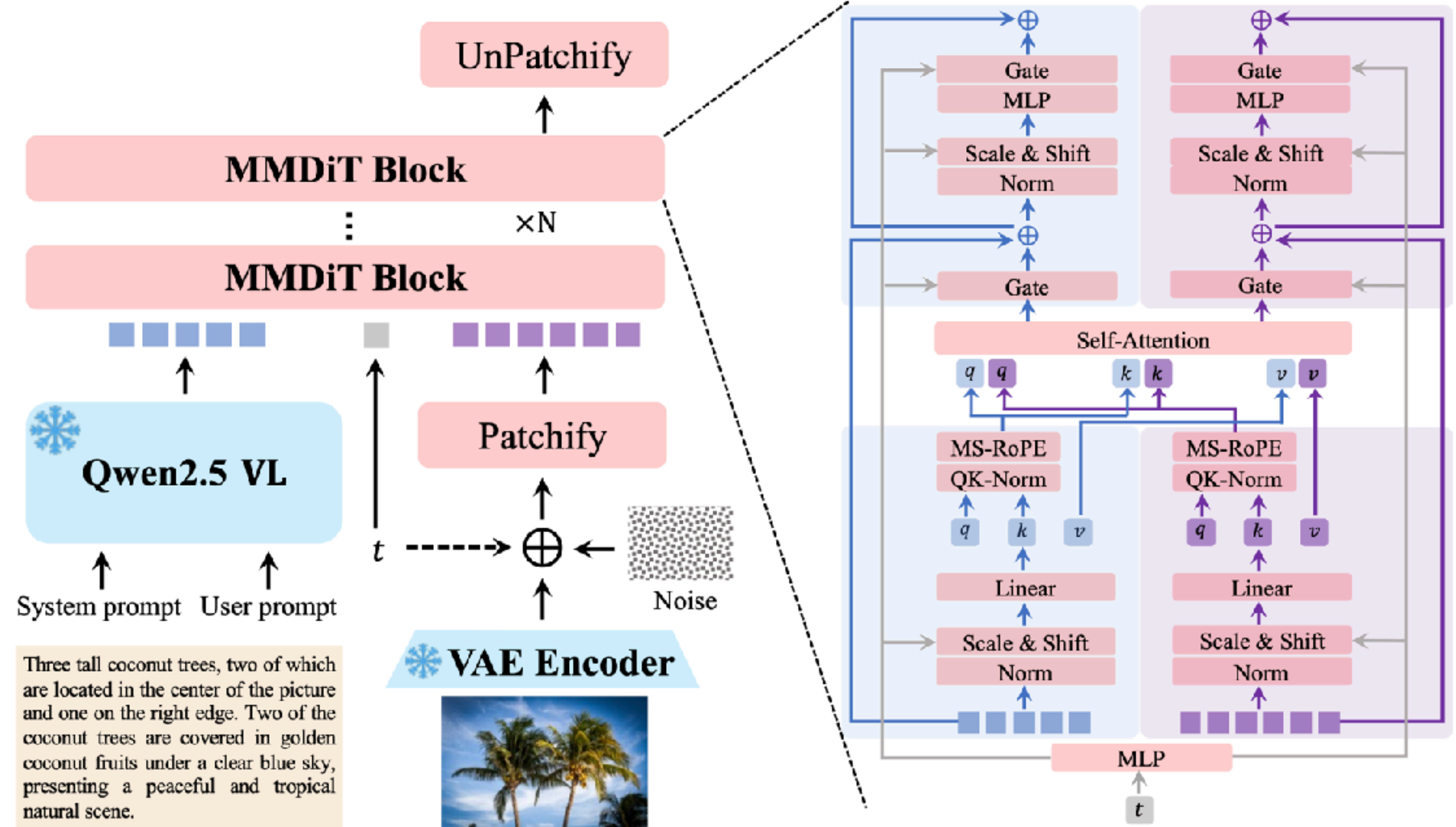


Esettanulmányok

Qwen-Image [2025]



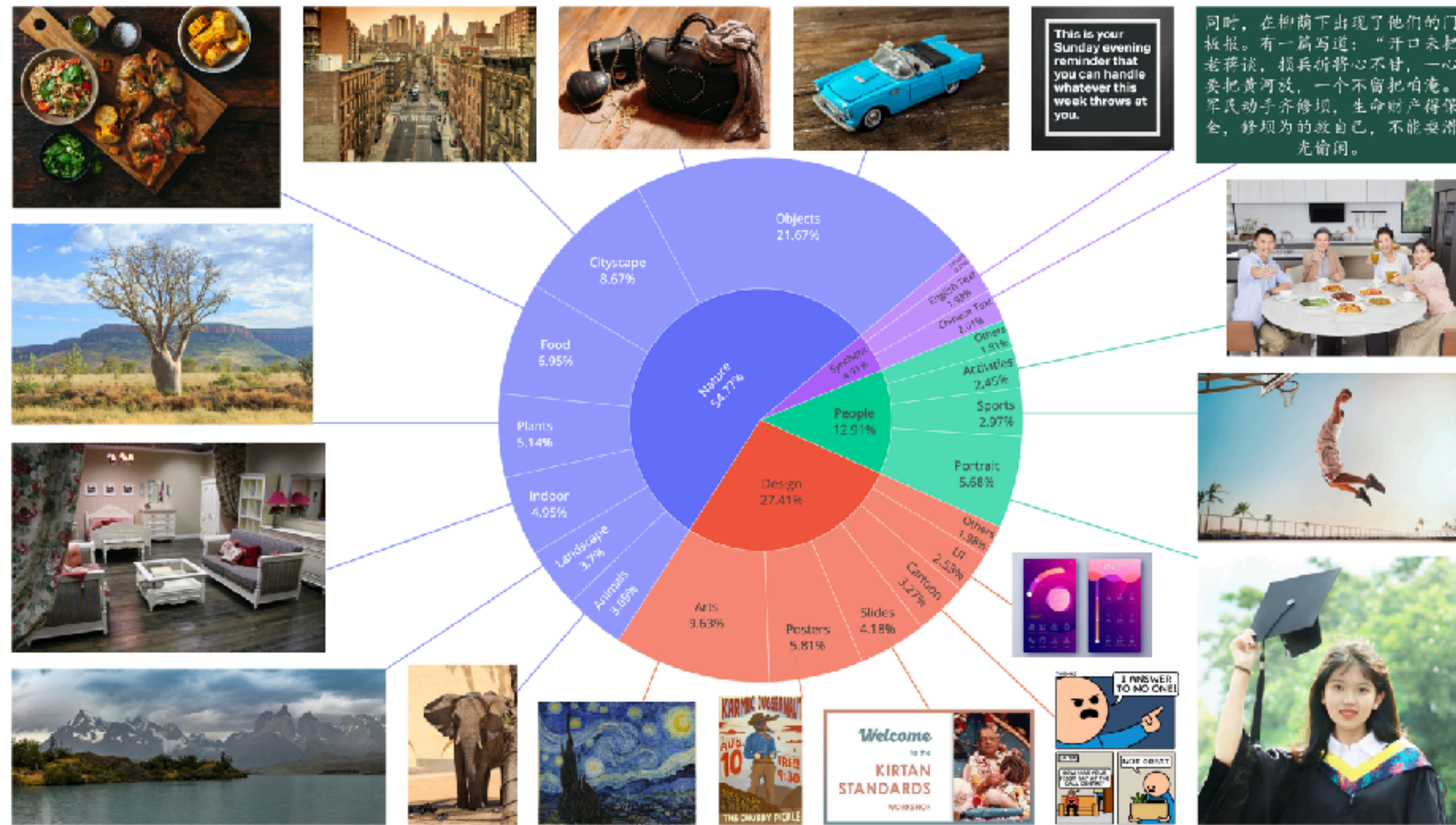
Configuration	VLM		VAE		MMDiT
	ViT	LLM	Enc	Dec	
# Layers	32	28	11	15	60
# Num Heads (Q / KV)	16 / 16	28 / 4	-	-	24 / 24
Head Size	80	128	-	-	128
Intermediate Size	3,456	18,944	-	-	12,288
Patch / Scale Factor	14	-	8x8	8x8	2
Channel Size	-	-	16	16	-
# Parameters	7B		54M	73M	20B



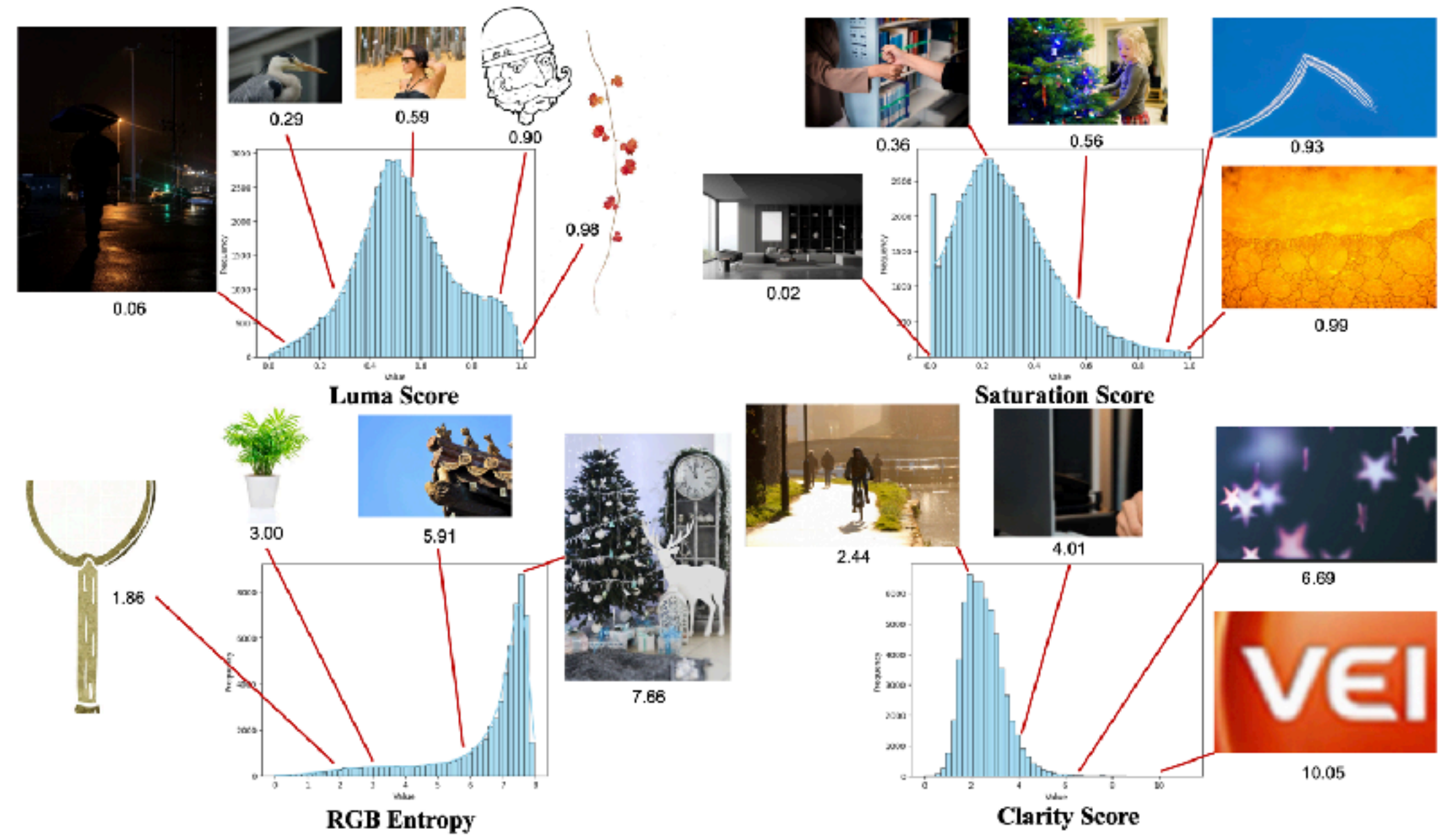
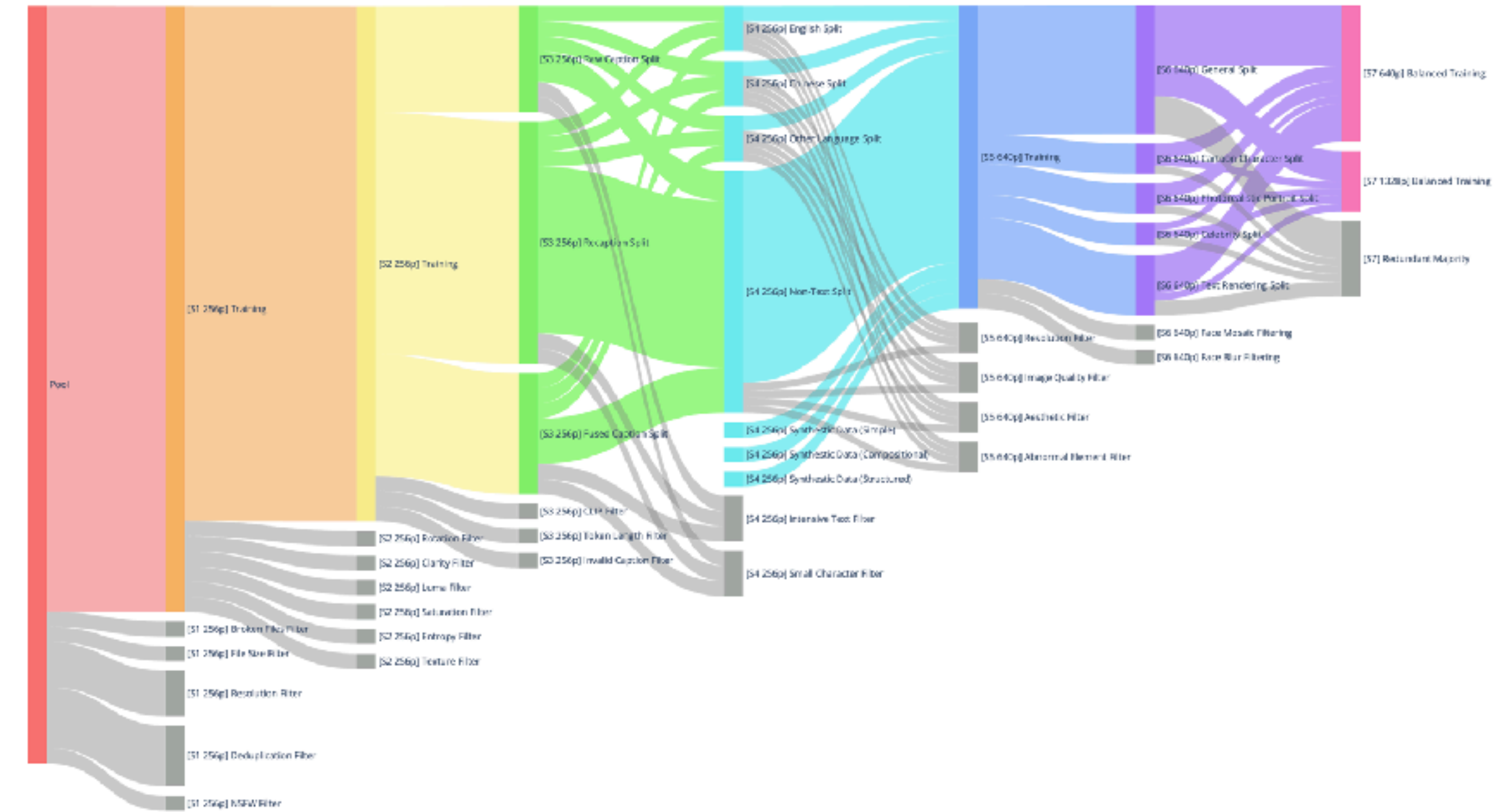
Three tall coconut trees, two of which are located in the center of the picture and one on the right edge. Two of the coconut trees are covered in golden coconut fruits under a clear blue sky, presenting a peaceful and tropical natural scene.

Esettanulmányok

Qwen-Image [2025]



Adathalmaz kuráció



Esettanulmányok

Qwen-Image [2025]

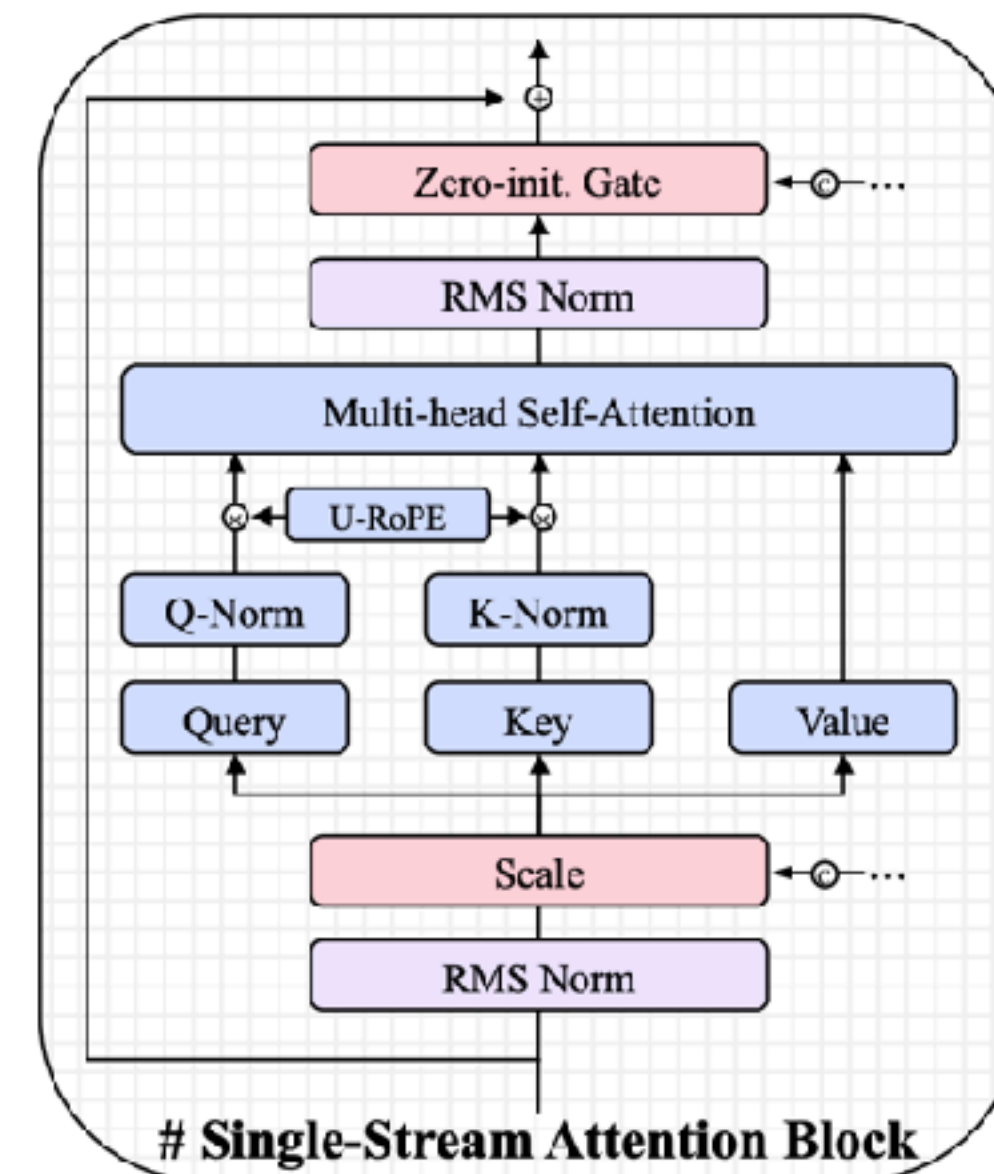
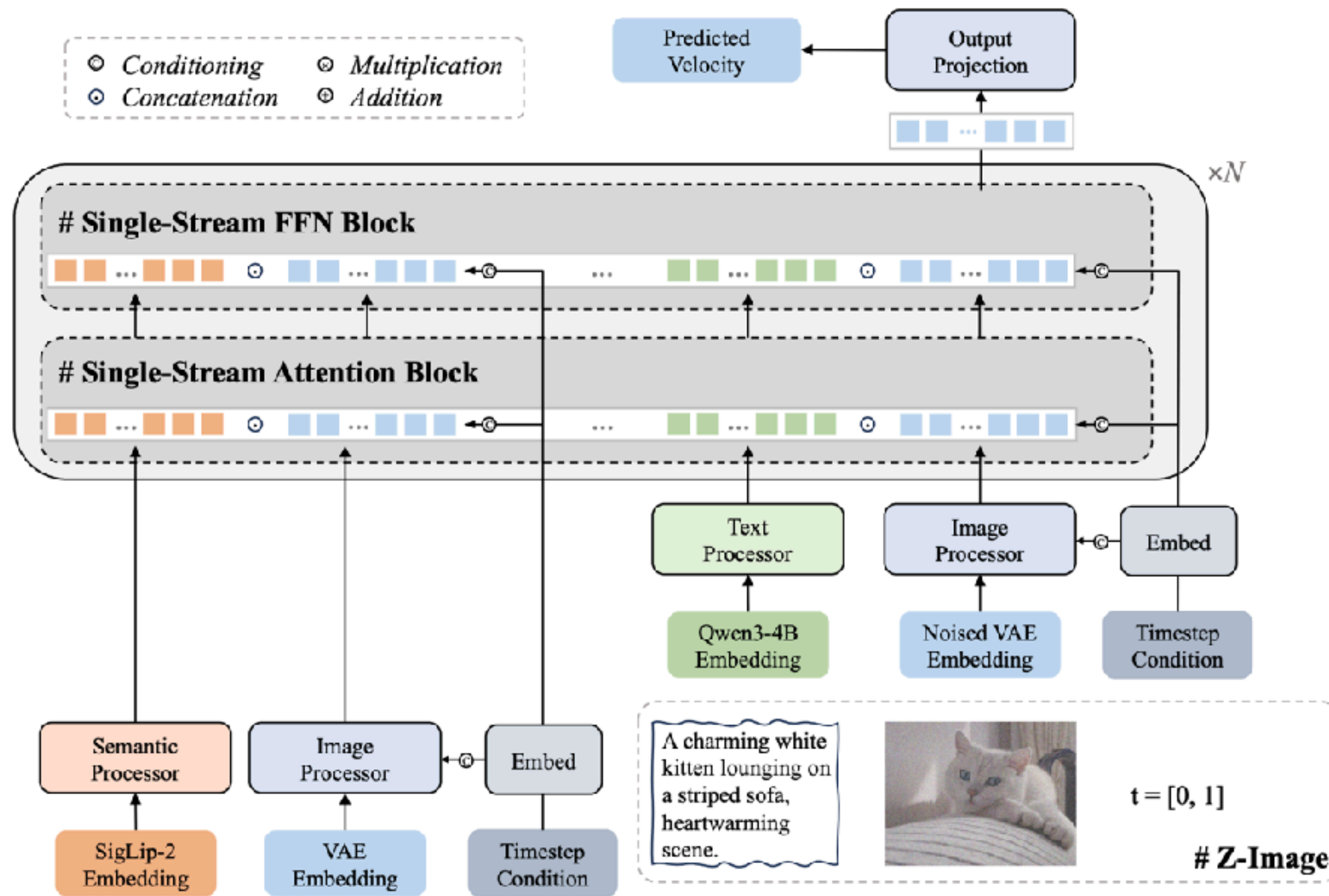


Esettanulmányok

Z-Image [2025]

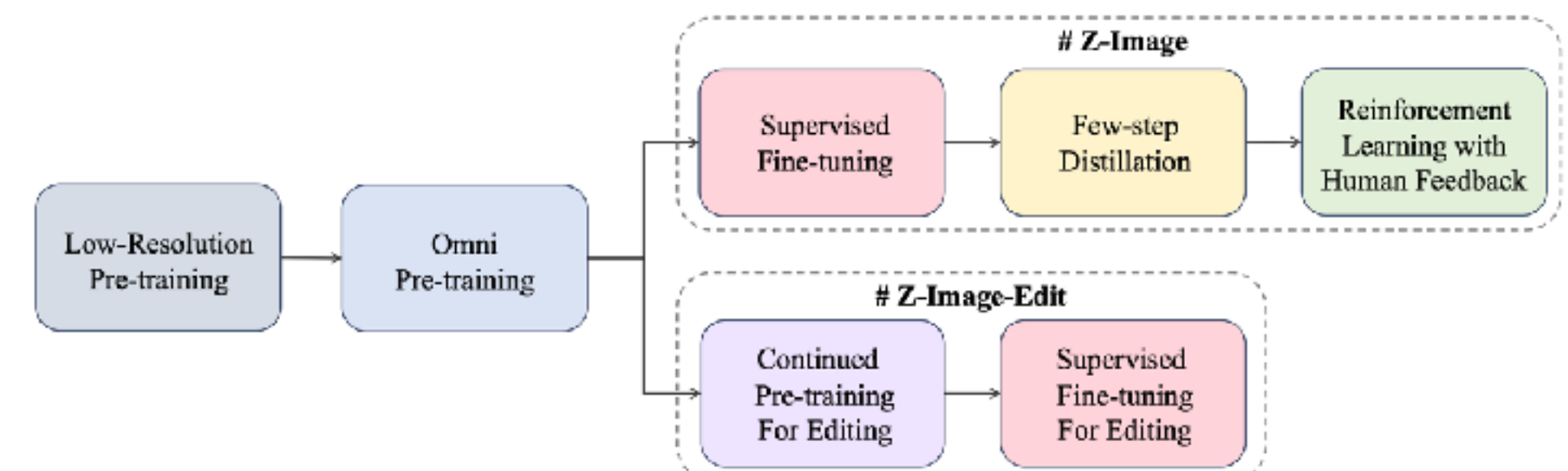
Z-Image: An Efficient Image Generation Foundation Model with Single-Stream Diffusion Transformer

Z-Image Team, Alibaba Group



Configuration	S3-DiT
Total Parameters	6.15B
Number of Layers	30
Hidden Dimension	3840
Number of Attention Heads	32
FFN Intermediate Dimension (d_v, d_h, d_w)	10240 (32, 48, 48)

- Image



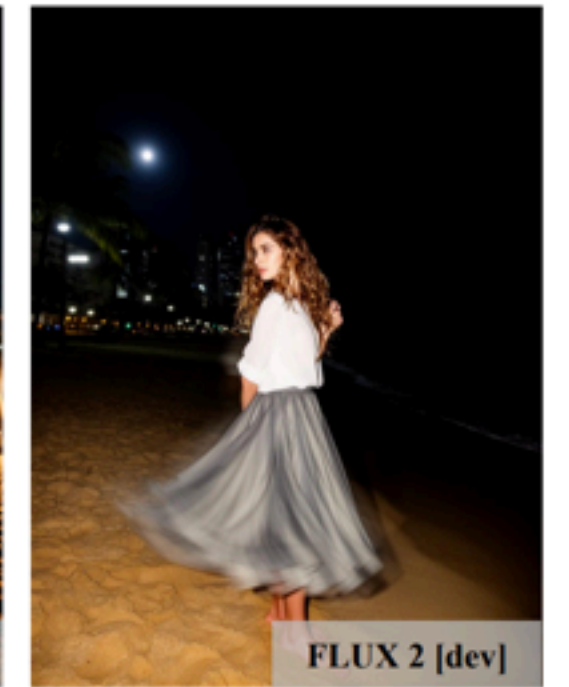
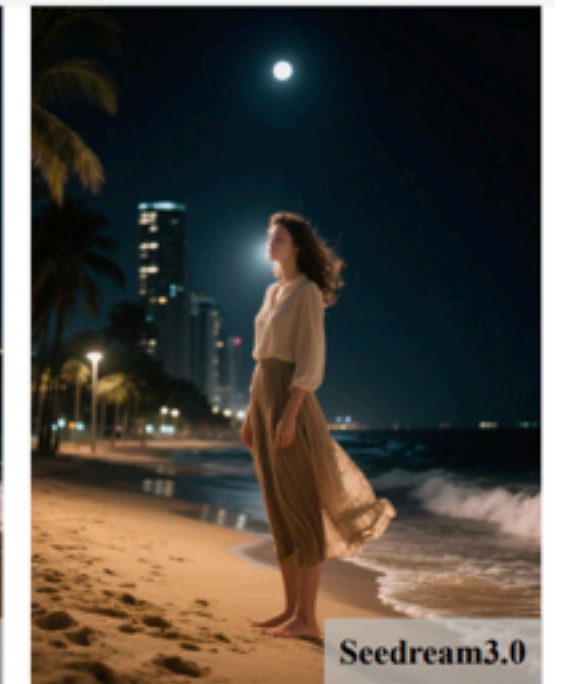
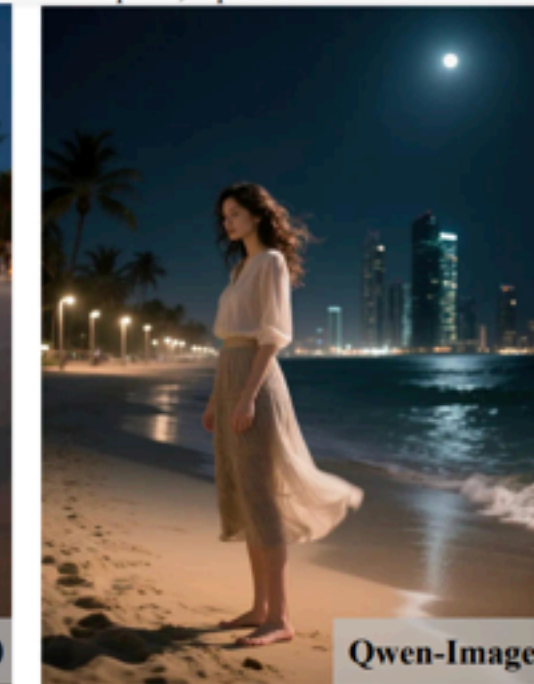
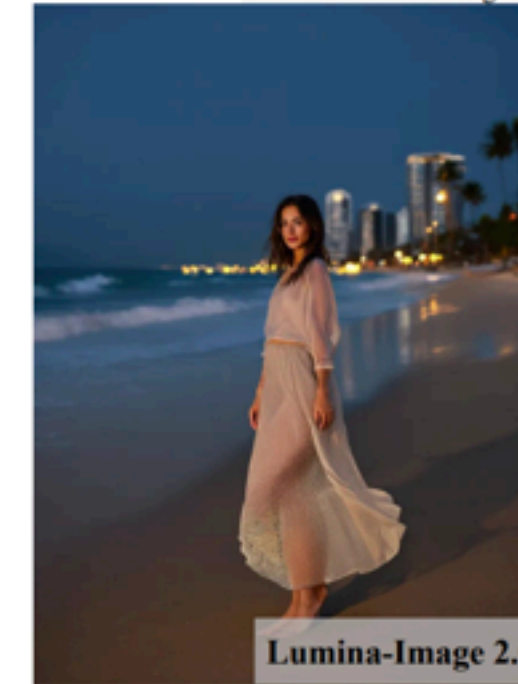
Esettanulmányok

Z-Image [2025]



Input prompt

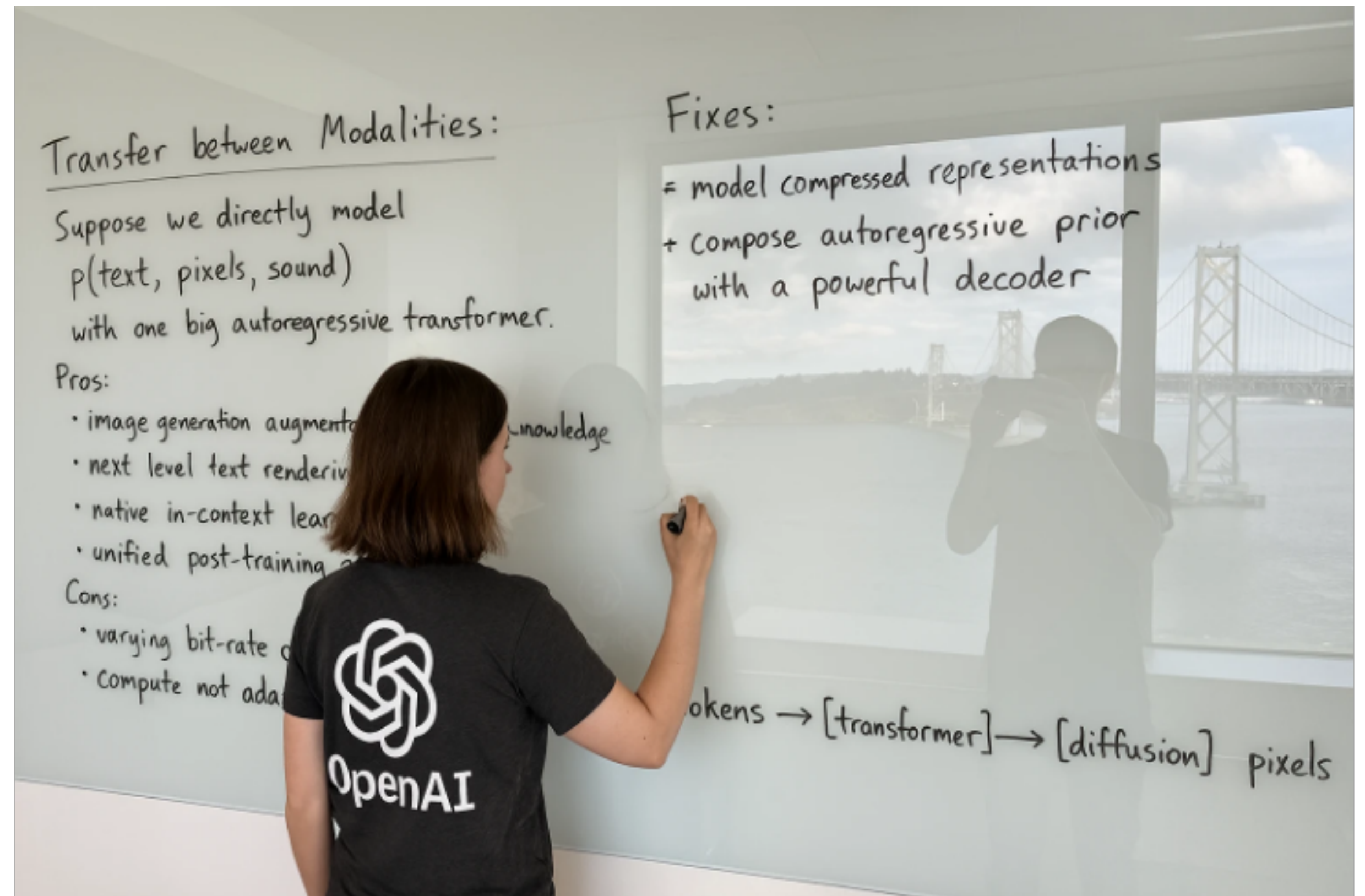
Under a sudden iPhone flash, she stands just where the surf and sand meet, the midi-skirt and fluid blouse catching the gentle seaside breeze, woven folds reflecting faint glimmers of distant city lights and moon's shy glow. Her bare feet nestle softly in the warm, textured sand, illuminated subtly by distant streetlamps that drift behind her, palms and skyscrapers whispering quiet stories to the night. Damp tendrils of wavy hair move gently, framing a serene, turned face softened by reflective air, while the expansive darkness spills across two-thirds of the frame in velvety repose. A delicate blur trails her skirt's hem, contrasting with the crisp focus of her poised stance, embodying timeless grace and effortless charm—flash-lit elegant beach snapshot, captured on iPhone.



Esettanulmányok

“Frontier” modellek [2024-25]

Elképesztő képességek...
de a cégek már *semmit* sem árulnak el róluk!



OpenAI GPT-4o



Google Gemini 2.5 (“Nano Banana”)

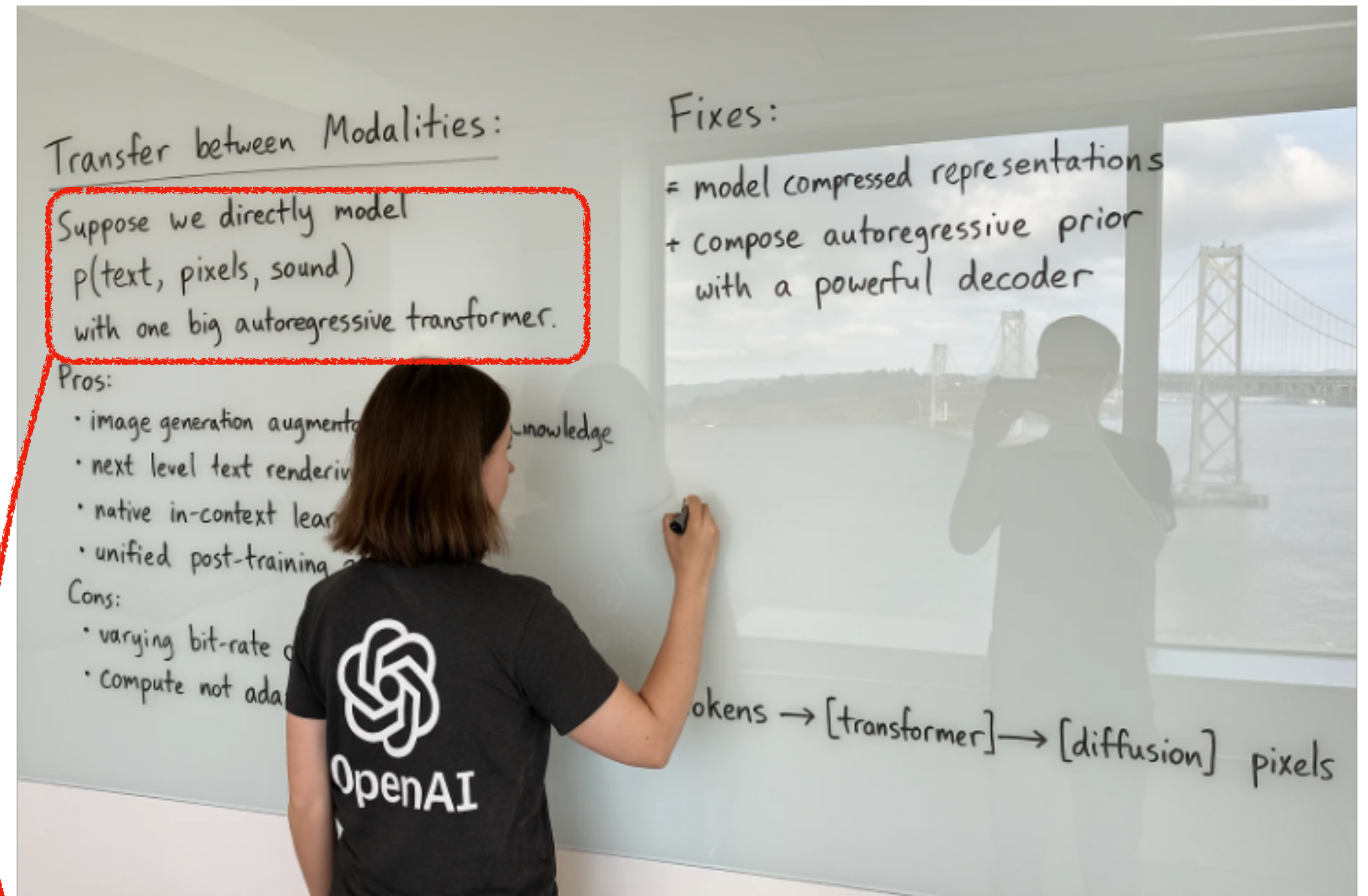
Esettanulmányok

“Frontier” modellek [2024-25]

Elképesztő képességek...
de a cégek már *semmit* sem árulnak el róluk!



Google Gemini 2.5 (“Nano Banana”)



OpenAI GPT-4o

Vagy mégis?

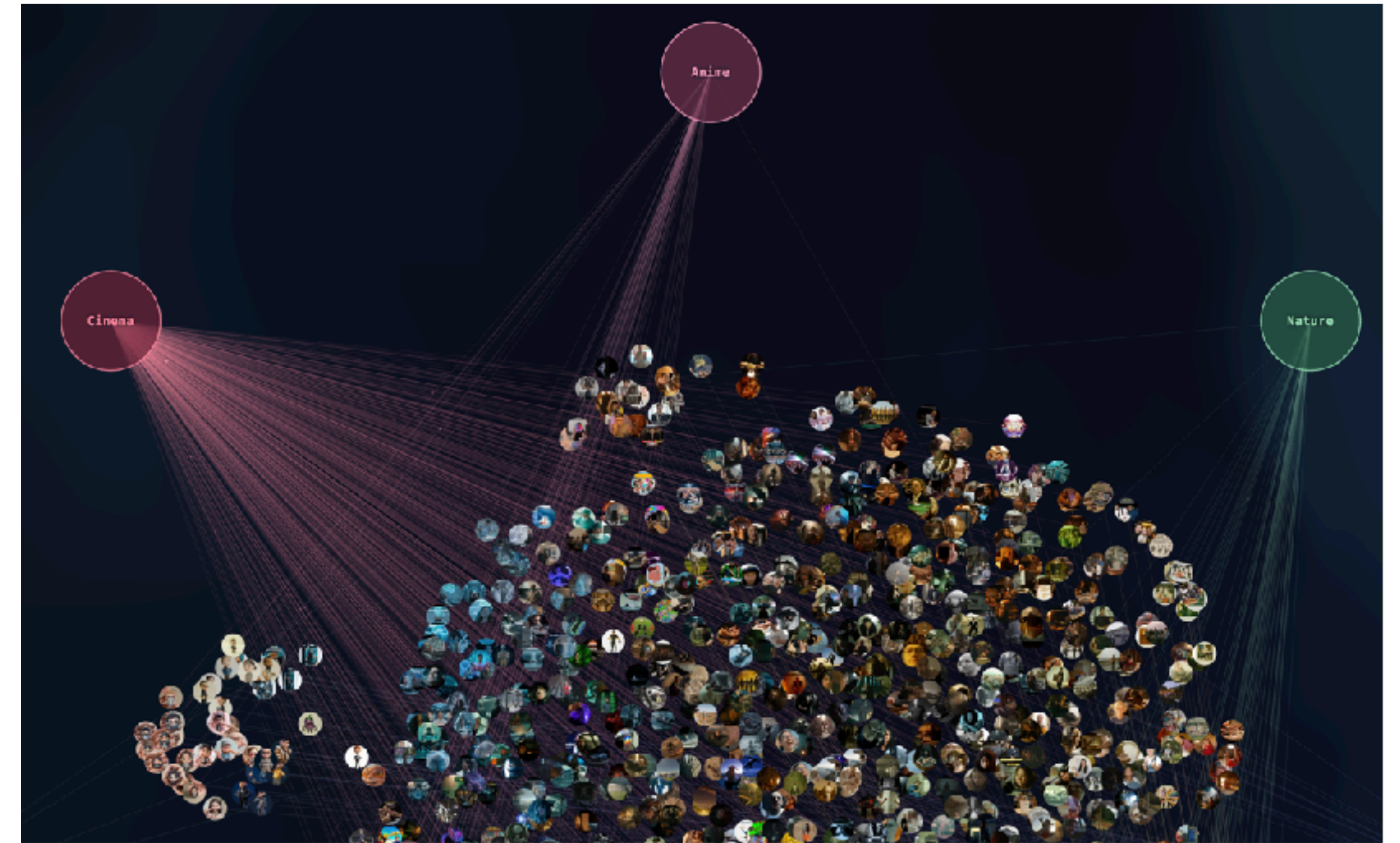
Esettanulmányok

Luma UNI-1 [2026]



<https://lumalabs.ai/uni-1>

Az autoregresszív transformerek diadala a képgenerálásban?



Látens tér vizualizáció

At its core, our model is a decoder-only autoregressive transformer. Text and images are represented in a single interleaved sequence, acting both as input and as output.

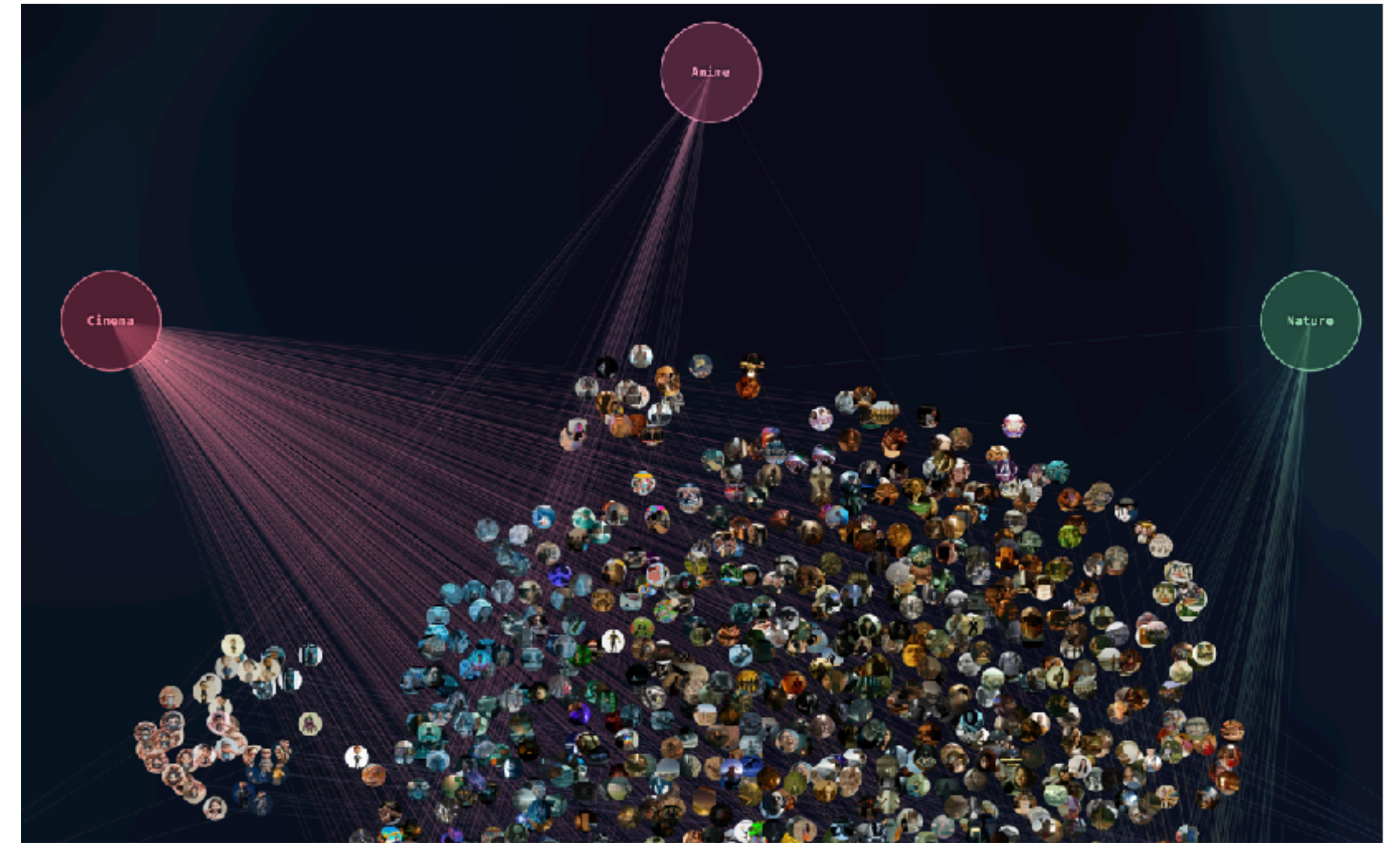
Esettanulmányok

Luma UNI-1 [2026]



<https://lumalabs.ai/uni-1>

Az autoregresszív transformerek diadala a képgenerálásban?



Látens tér vizualizáció

At its core, our model is a decoder-only autoregressive transformer. Text and images are represented in a single interleaved sequence, acting both as input and as output.

Diffúziós Generálás a Gyakorlatban

HuggingFace Diffusers

- HuggingFace Diffusers könyvtár: [LINK](#)
- Nyílt forráskódú implementációk diffúziós modellekre,
- Egységes API: DiffusionPipeline
 - Diffúziós modell, prompt enkóder, zaj ütemező, mintavetelező, stb. egybe csomagolva
- Sok-sok adapter és fine-tuning modul (LoRA, ControlNet, stb.)



```
import torch
from diffusers import DiffusionPipeline

pipeline = DiffusionPipeline.from_pretrained(
    "Qwen/Qwen-Image", torch_dtype=torch.bfloat16, device_map="cuda"
)

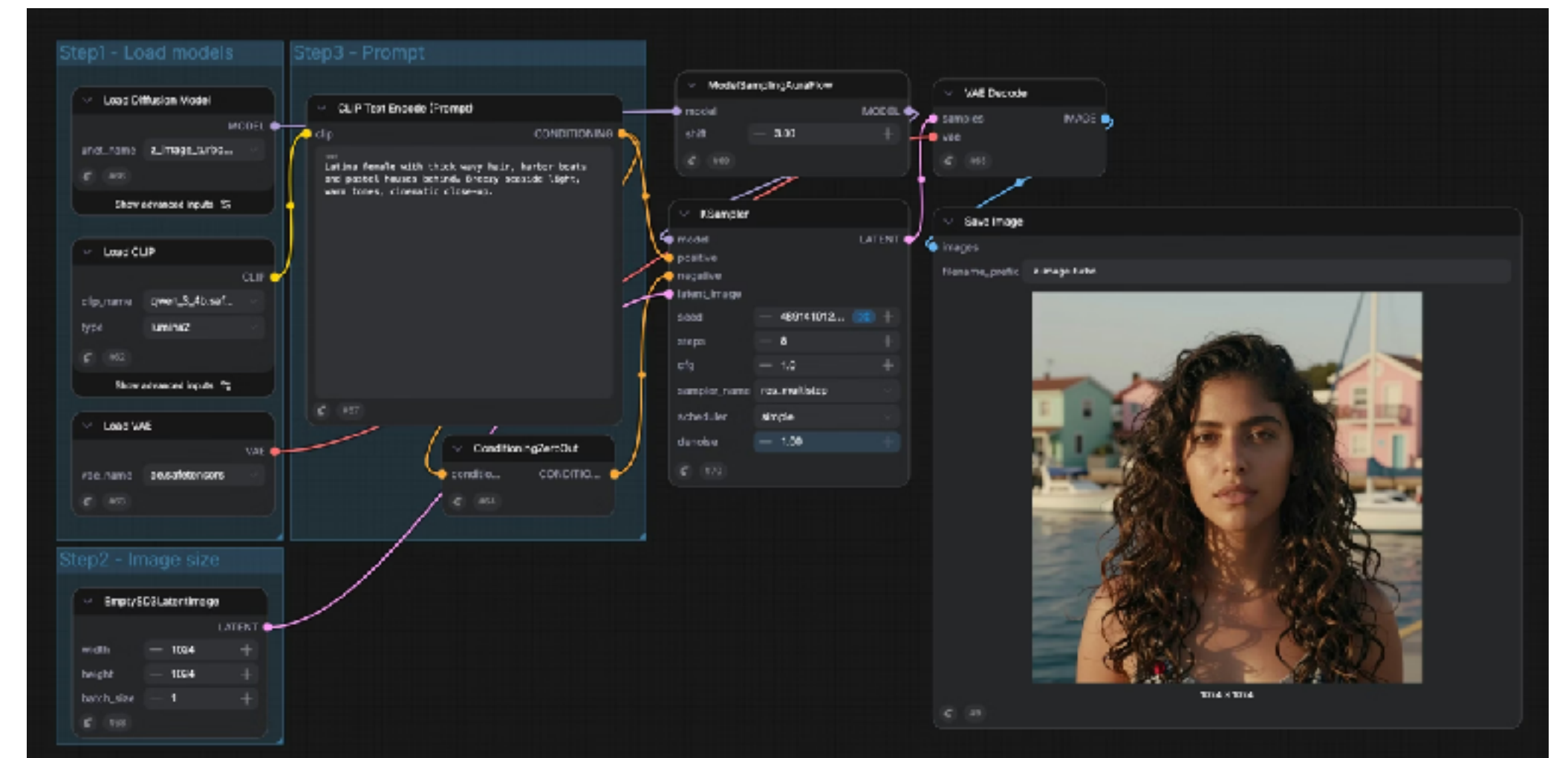
prompt = """
cinematic film still of a cat sipping a margarita in a pool in Palm Springs, California
highly detailed, high budget hollywood movie, cinemascope, moody, epic, gorgeous, film grain
"""

pipeline(prompt).images[0]
```

Diffúziós Generálás a Gyakorlatban

ComfyUI

- ComfyUI: grafikus programozási környezet generatív média előállításra
- Node-ok összekötése grafikus felületen keresztül
- Érett ökoszisztéma, sok száz custom node, template, stb.
- Rengeteg diffúziós modell, LLM, adapter, stb. támogatása
- Futtatás lokálisan, vagy felhőben



Következő előadás: Haladó Diffúziós Generálás

- Gyorsított generálás, modell disztilláció
- Képszerkesztés, diffúziós priorok
- Videó diffúzió, Világmodellek

