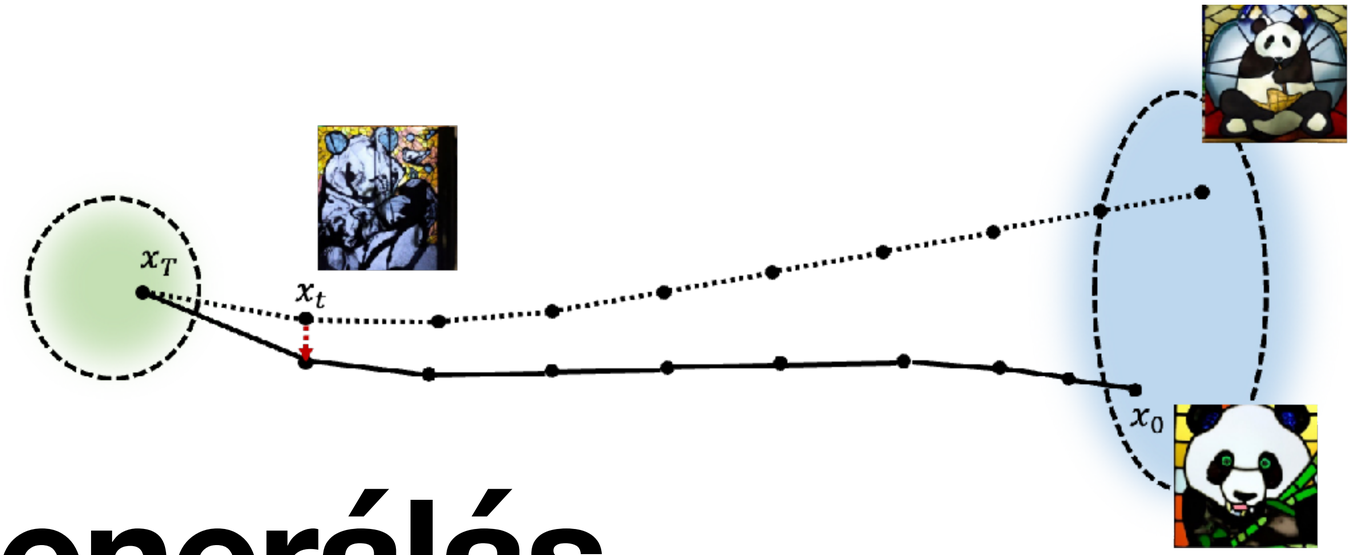


9. Előadás: Vezérelt Képgenerálás

Generatív AI és Inverz Módszerek a Képszintézisben

BME-VIK IIT, 2026

Dr. Vaitkus Márton

Vezérelt Generálás

Diffúziós modell (*<u>feltétel nélkül</u>*):

Becsült zaj

$$\epsilon_\theta(x_T, T)$$

Zajos kép

Zajszint

Vezérelt Generálás

Diffúziós modell (*feltétel nélkül*):

Becsült zaj → $\epsilon_\theta(x_T, T)$

Zajos kép: x_T

Zajszint: T

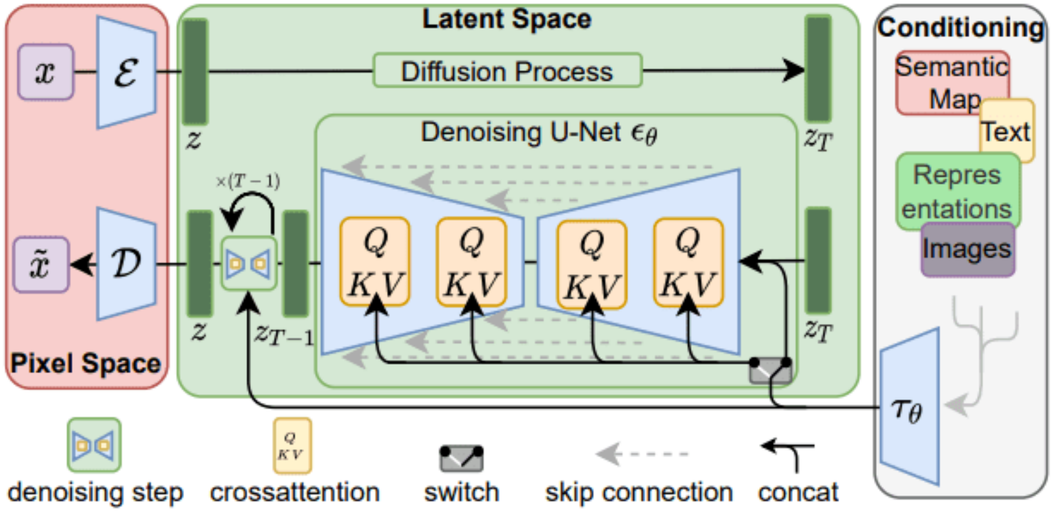
Diffúziós modell (*szöveges feltétellel*):

$$\epsilon_\theta(x_T, T, c)$$

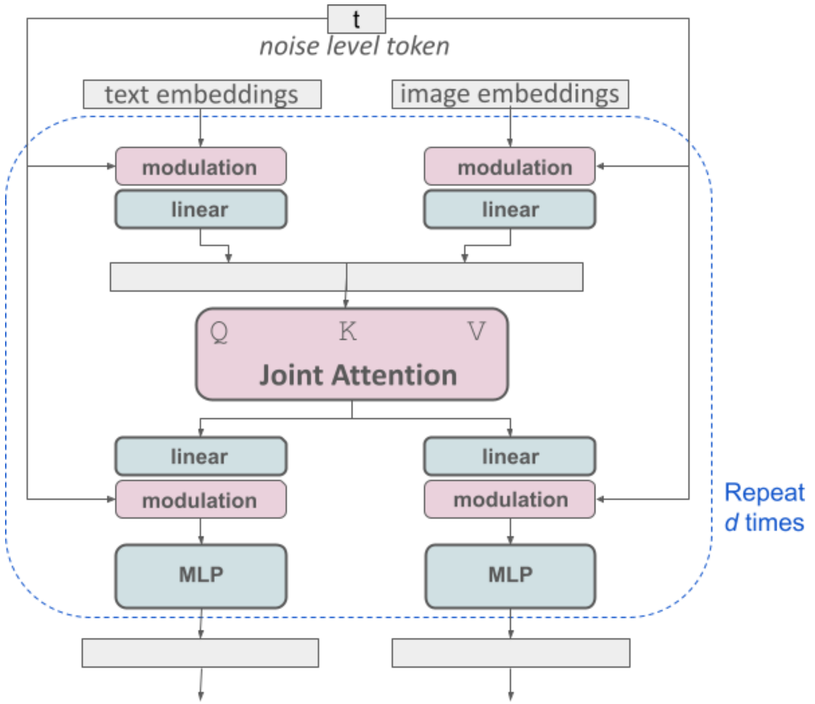
c = "A stained glass window of a panda eating bamboo."

Vezérelt Generálás

Neurális implementáció



Stable Diffusion 1.0
(U-Net + cross-attention)



Stable Diffusion 3.5
(Multimodális Transformer)

Vezérelt Generálás

Instrukciók követése

“A stained glass window of a panda eating bamboo.”

Vezérelt Generálás

Instrukciók követése

“A stained glass window of a panda eating bamboo.”

Vezérelt Generálás

Instrukciók követése

“A stained glass window of a panda eating bamboo.”

Tudunk jobbat?

Vezérelt Generálás

Instrukciók követése

“A stained glass window of a panda eating bamboo.”

Tudunk jobbat?

Vezérelt Generálás

Instrukciók követése

“A stained glass window of a panda eating bamboo.”

Tudunk jobbat?

Igen!

“Guidance”

Vezérelt Generálás

Instrukciók követése

“A stained glass window of a panda eating bamboo.”

Tudunk jobbat?

Igen!

“Guidance”

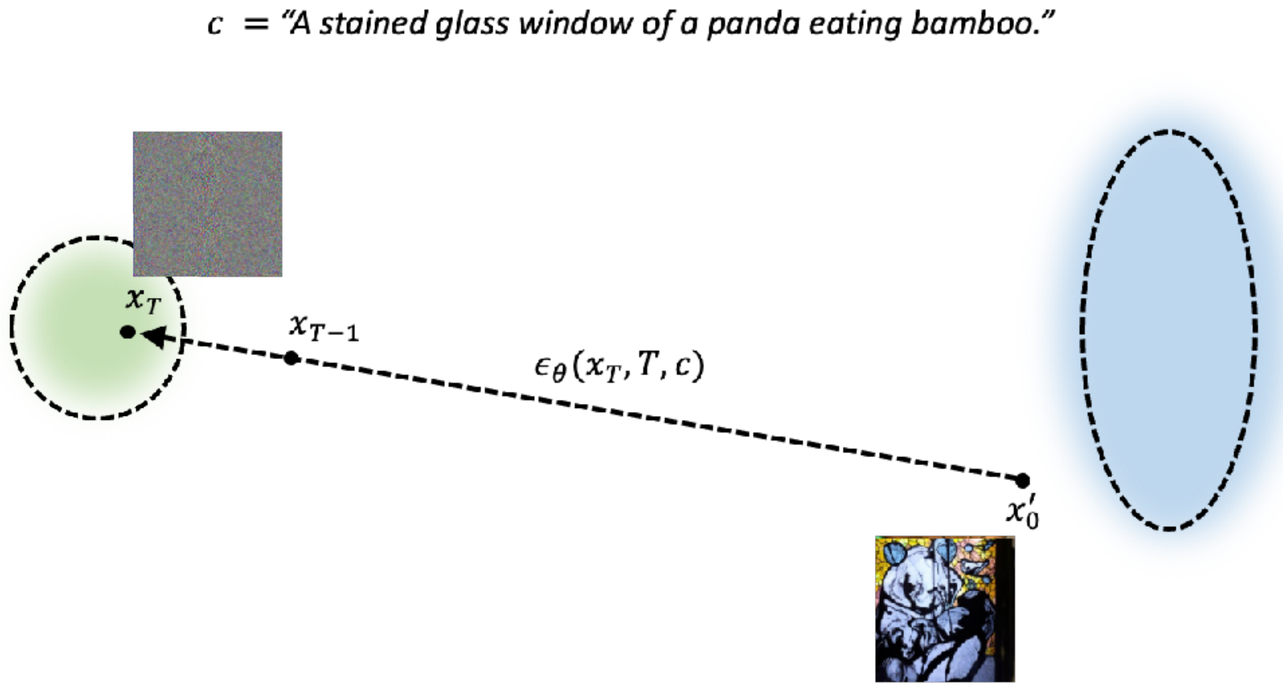
Kritikus eleme minden képgenerátornak!!!

Vezérelt Generálás

Classifier Guidance

Vezérelt Generálás

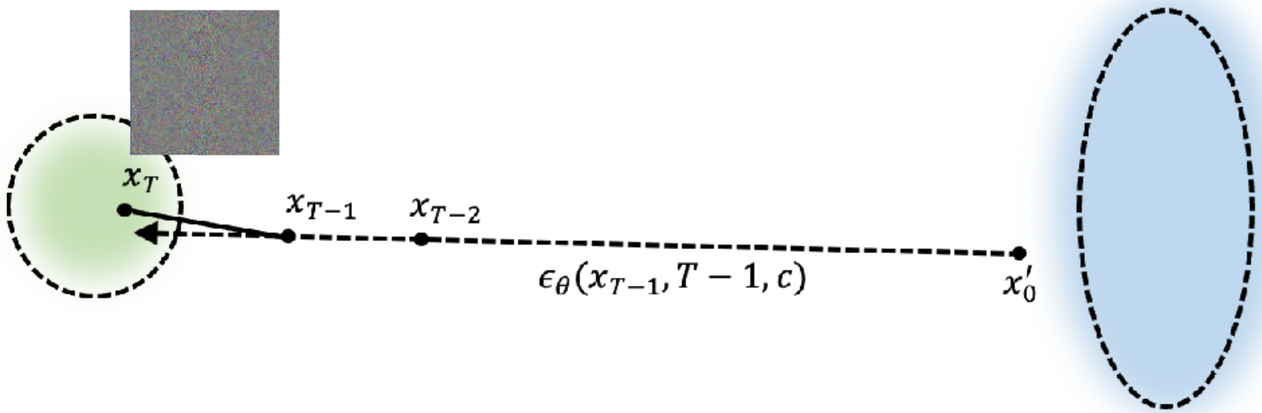
Classifier Guidance



Vezérelt Generálás

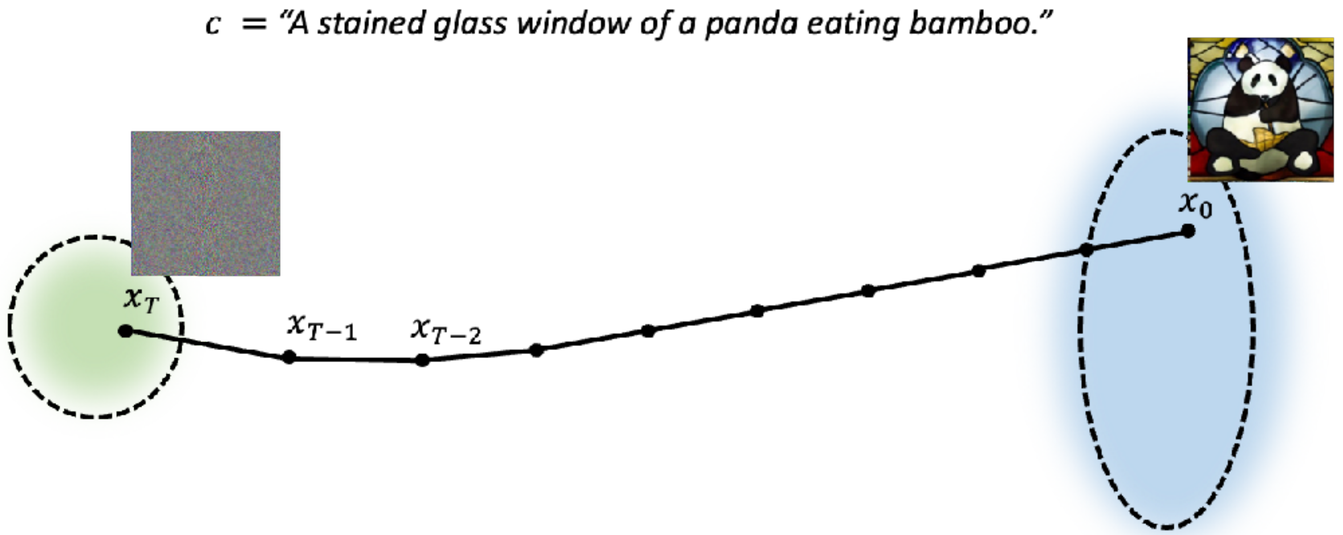
Classifier Guidance

c = *"A stained glass window of a panda eating bamboo."*



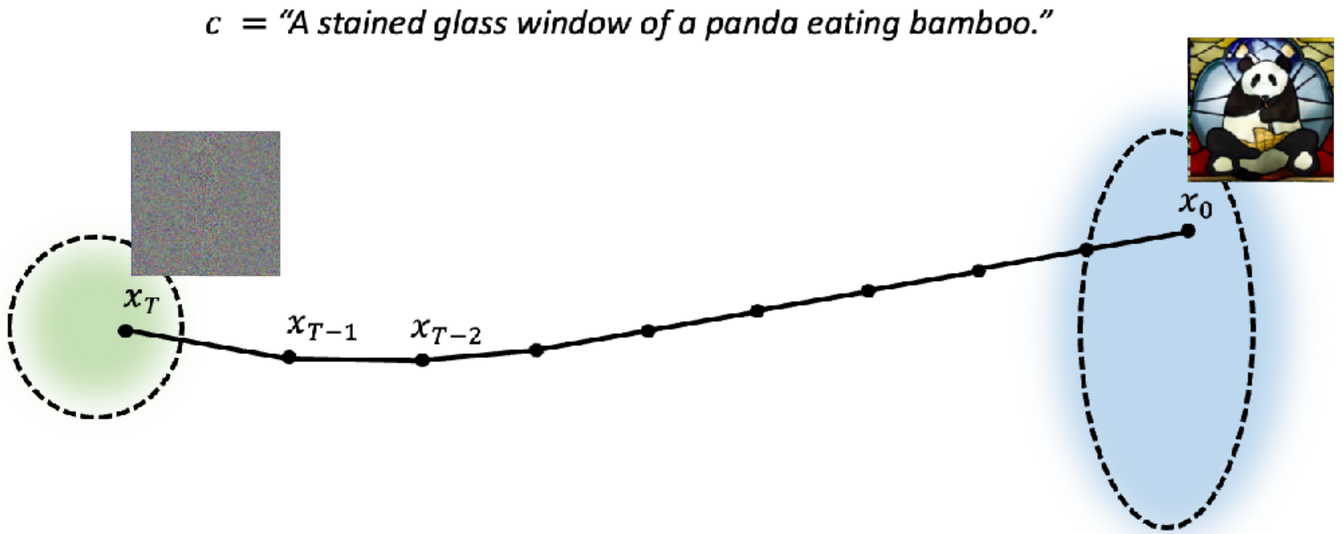
Vezérelt Generálás

Classifier Guidance



Vezérelt Generálás

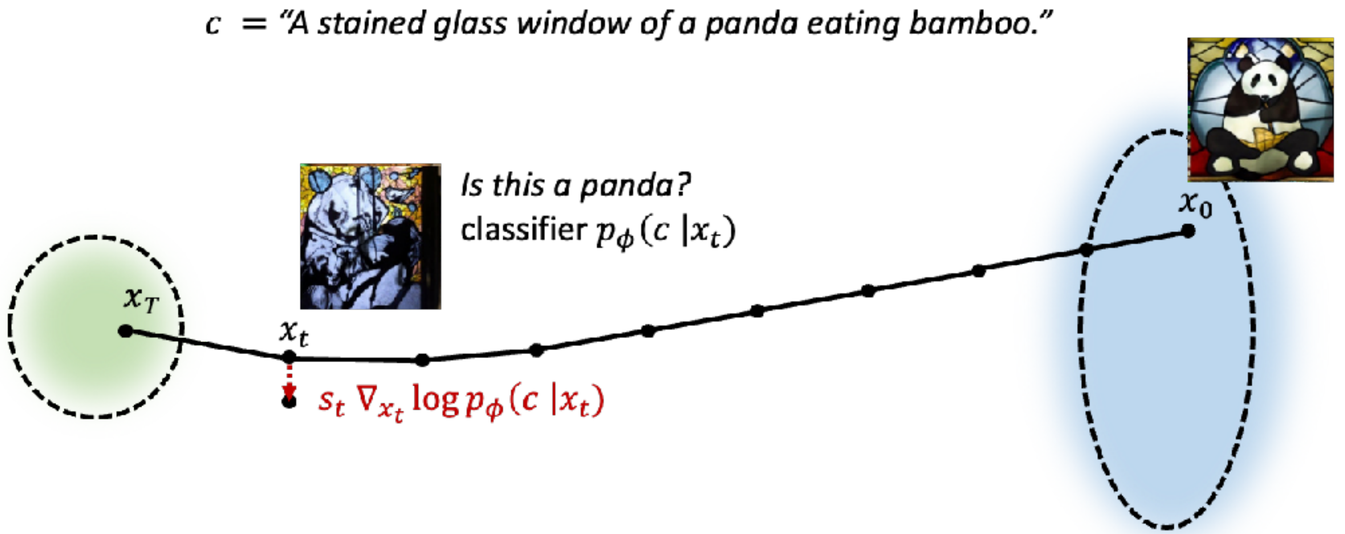
Classifier Guidance



Classifier Guidance — minősítsük a zajos kép prompthoz való igazodását — $p_\phi(c \mid x_t)$

Vezérelt Generálás

Classifier Guidance

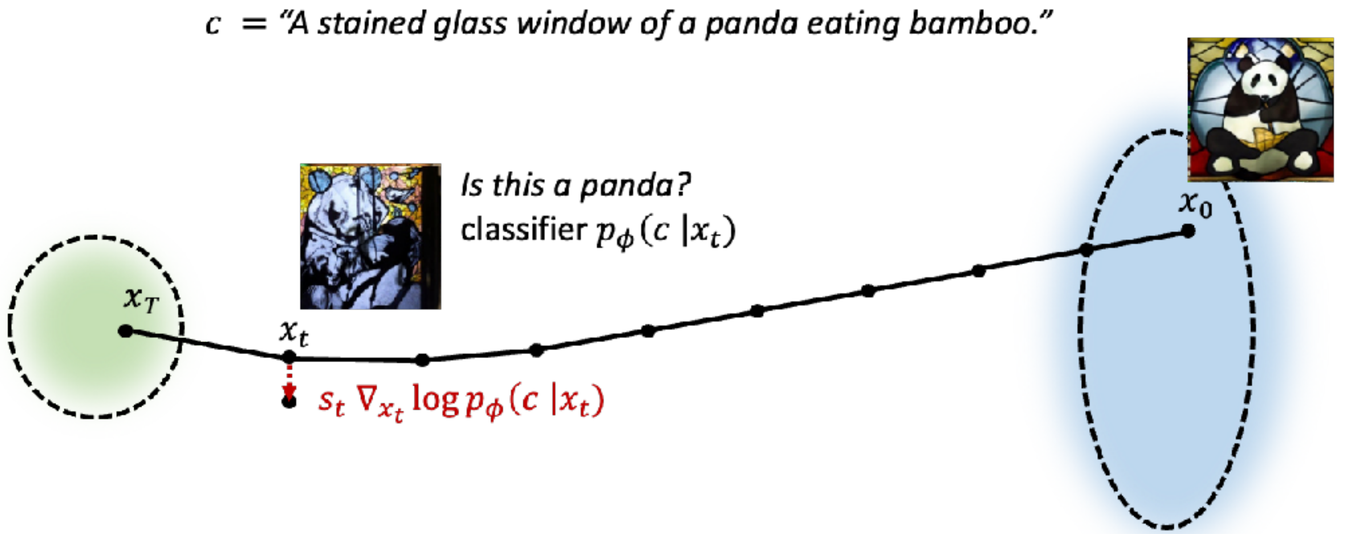


Classifier Guidance — minősítsük a zajos kép prompthoz való igazodását — $p_\phi(c \mid x_t)$

Bayes tételből: $\nabla_x \log p(x_t \mid c) = \nabla_x \log p(c \mid x_t) + \nabla_x \log p(x_t) - \cancel{\nabla_x \log p(c)}$

Vezérelt Generálás

Classifier Guidance



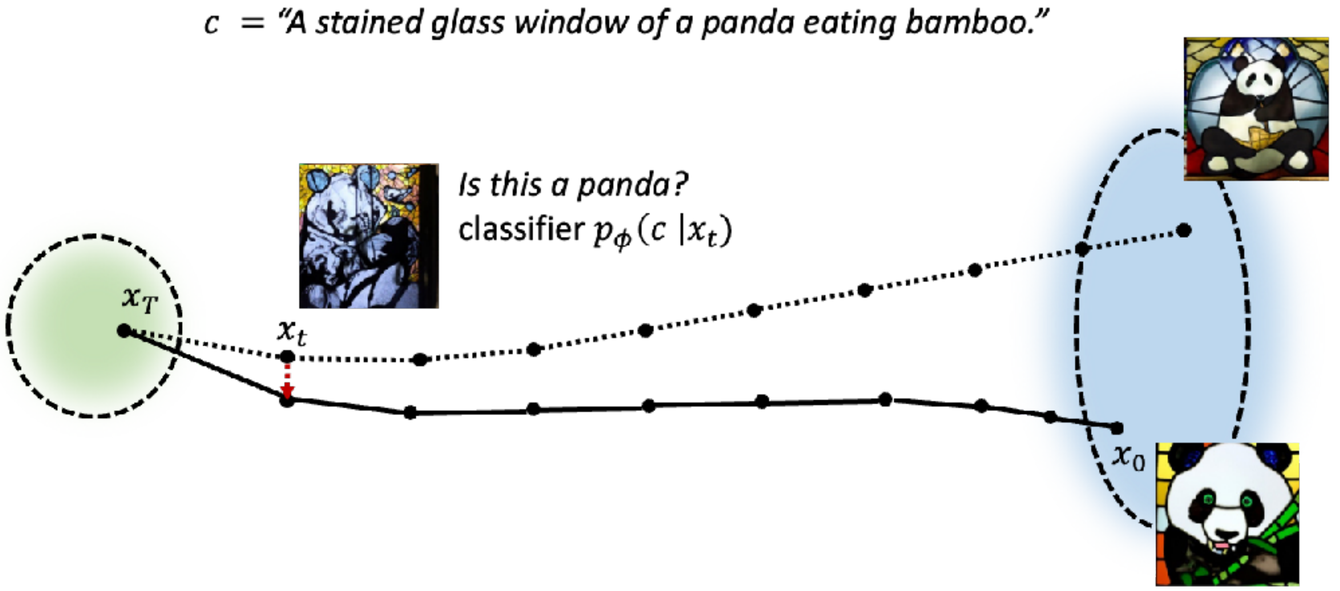
Classifier Guidance — minősítsük a zajos kép prompthoz való igazodását — $p_\phi(c \mid x_t)$

Bayes tételből: $\nabla_x \log p(x_t \mid c) = \nabla_x \log p(c \mid x_t) + \nabla_x \log p(x_t) - \cancel{\nabla_x \log p(c)}$

Skálázzuk fel a klasszifikáció erejét: $\nabla_x \log p(x_t) + s_t \cdot \nabla_x \log p(c \mid x_t)$

Vezérelt Generálás

Classifier Guidance



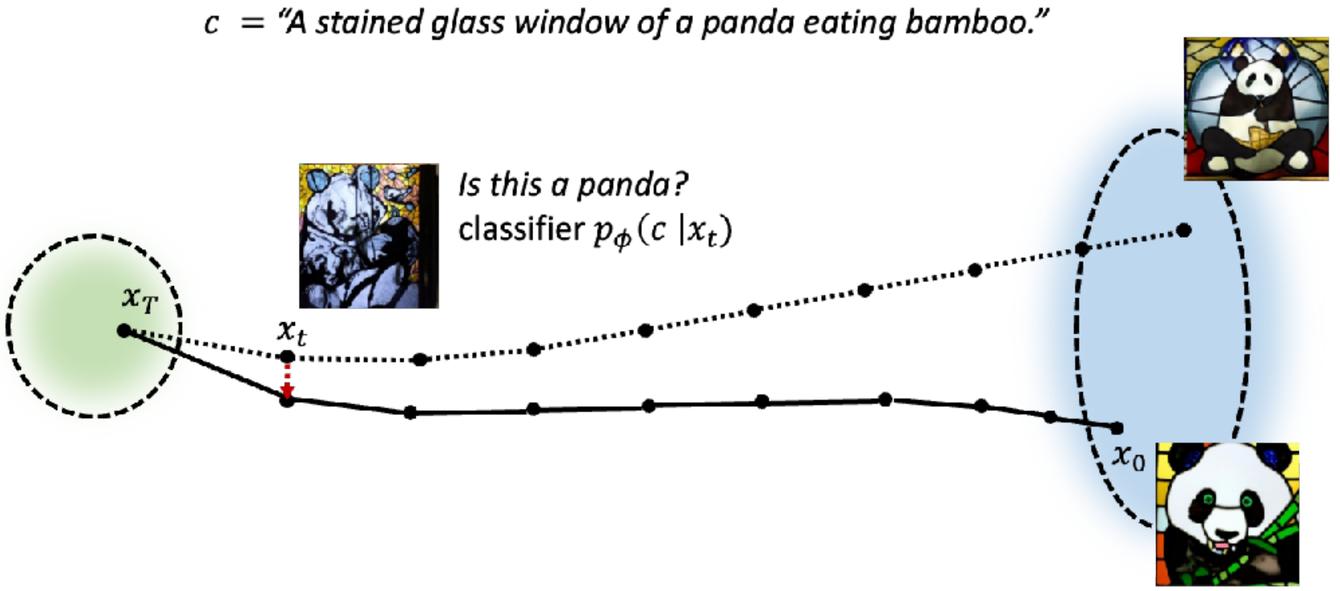
Classifier Guidance — minősítsük a zajos kép prompthoz való igazodását — $p_\phi(c \mid x_t)$

Bayes tételből: $\nabla_x \log p(x_t \mid c) = \nabla_x \log p(c \mid x_t) + \nabla_x \log p(x_t) - \cancel{\nabla_x \log p(c)}$

Skálázzuk fel a klasszifikáció erejét: $\nabla_x \log p(x_t) + s_t \cdot \nabla_x \log p(c \mid x_t)$

Vezérelt Generálás

Classifier Guidance



Classifier Guidance — minősítsük a zajos kép prompthoz való igazodását — $p_\phi(c \mid x_t)$

Bayes tételből: $\nabla_x \log p(x_t \mid c) = \nabla_x \log p(c \mid x_t) + \nabla_x \log p(x_t) - \cancel{\nabla_x \log p(c)}$

Skálázzuk fel a klasszifikáció erejét: $\nabla_x \log p(x_t) + s_t \cdot \nabla_x \log p(c \mid x_t)$

Ehhez külön klasszifikátort kell tanítani, méghozzá *zajos* képeken!

Vezérelt Generálás

Classifier-Free Guidance (CFG)

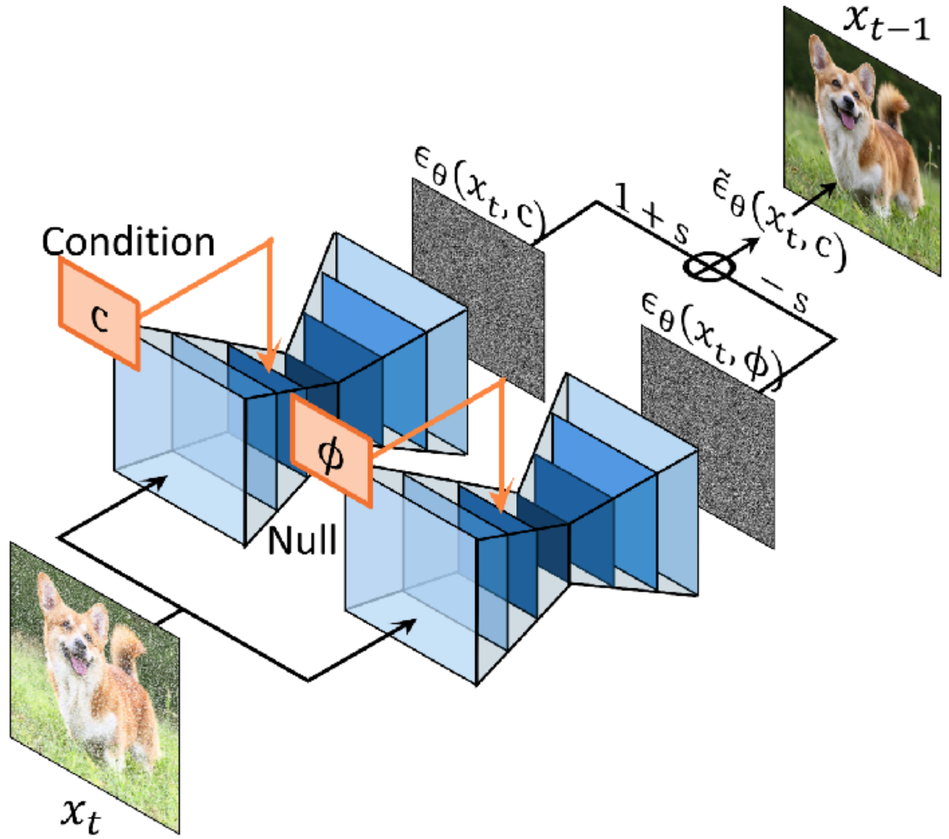
CLASSIFIER-FREE DIFFUSION GUIDANCE

Jonathan Ho & Tim Salimans
Google Research, Brain team

- Szabaduljunk meg a klasszifikációtól (Bayes újból):

$$\nabla_x \log p(x_t) + s_t \cdot \nabla_x \log p(y \mid x_t)$$
$$= \nabla_x \log(p(x_t)) + s_t \cdot (\nabla_x \log p(x_t \mid y) + \cancel{\nabla_x \log p(y)} - \nabla_x \log p(x_t))$$
$$= (1 - s_t) \underbrace{\nabla_x \log p(x_t)}_{\text{Denoizer feltétel nélkül}} + s_t \cdot \underbrace{\nabla_x \log p(x_t \mid y)}_{\text{Denoizer feltétellel}}$$

- **Classifier-Free Guidance (CFG)** — a denoizert 2x értékeljük ki: feltétellel és anélkül (“null” feltétel), majd kombináljuk!
- Címkézett képekkel való tanítás során valamilyen valószínűséggel elhagyjuk a feltételt (“null”-al helyettesítjük), hogy a háló feltétel nélkül is megtanuljon generálni!

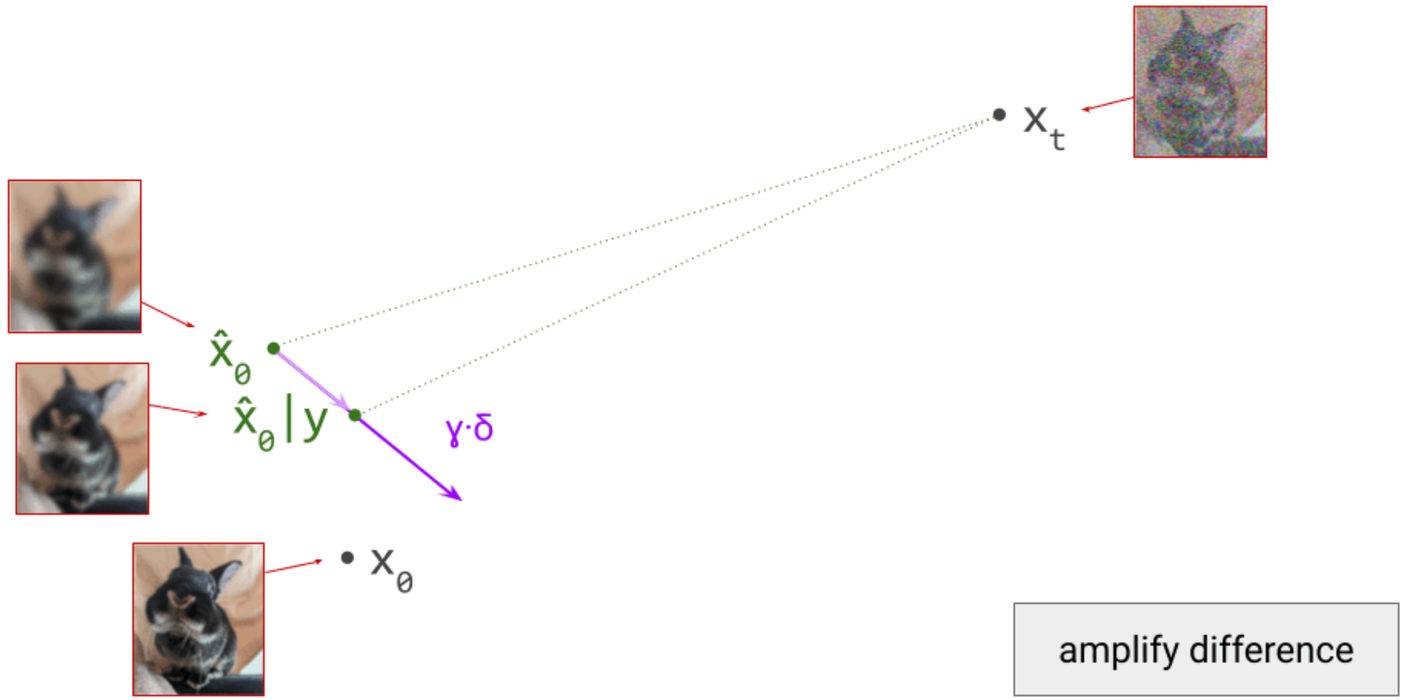


Vezérelt Generálás

Classifier-Free Guidance (CFG)

CLASSIFIER-FREE DIFFUSION GUIDANCE

Jonathan Ho & Tim Salimans
Google Research, Brain team



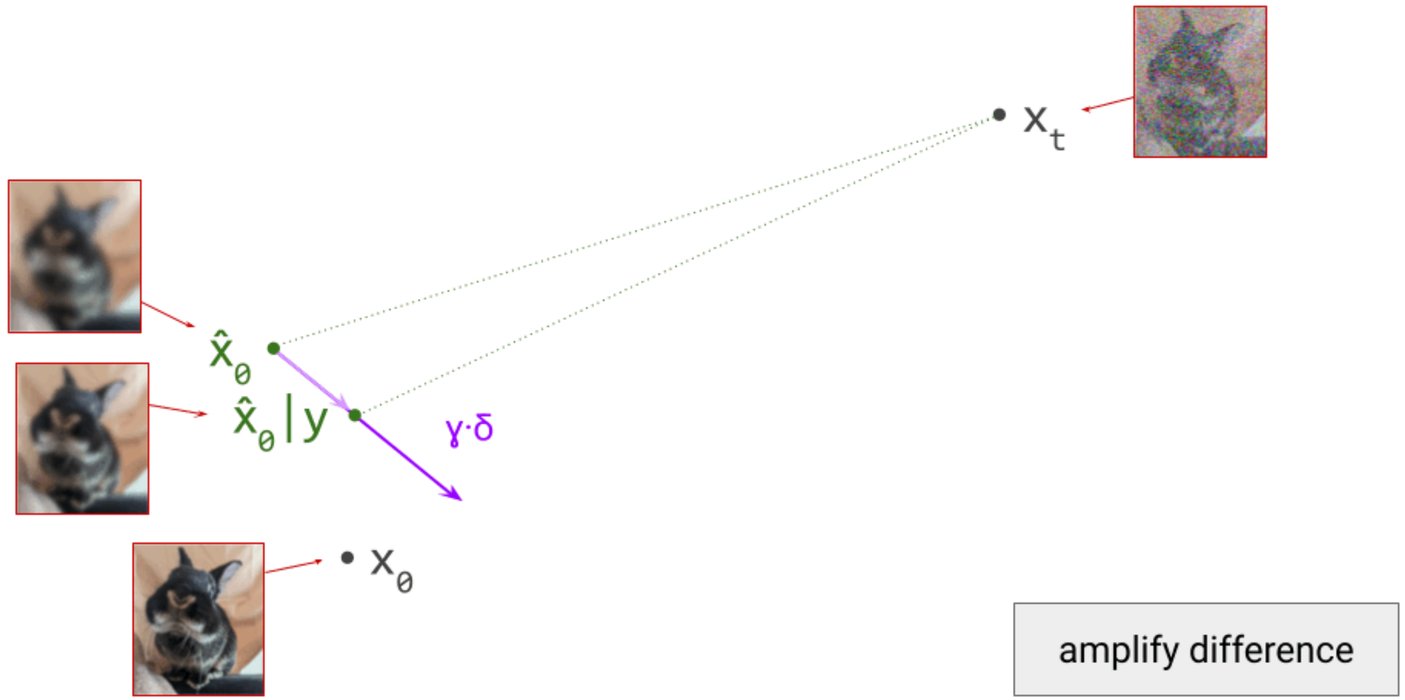
DDPM + CFG — Forrás: LINK

Vezérelt Generálás

Classifier-Free Guidance (CFG)

CLASSIFIER-FREE DIFFUSION GUIDANCE

Jonathan Ho & Tim Salimans
Google Research, Brain team



DDPM + CFG — Forrás: LINK

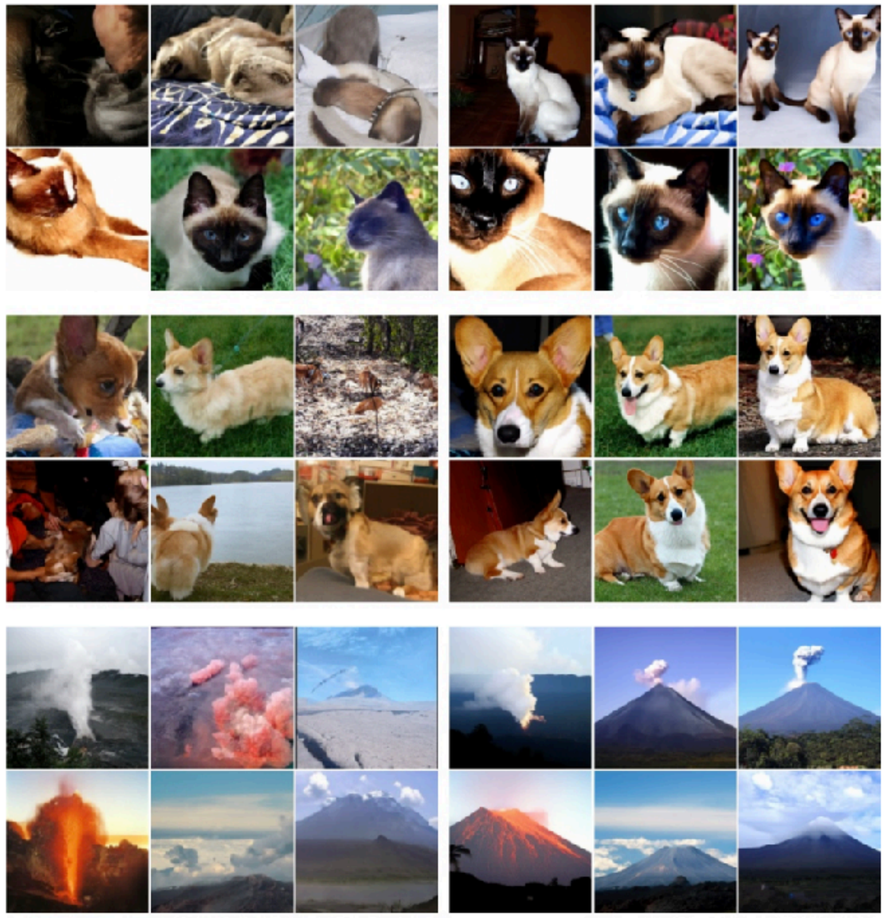
Vezérelt Generálás

Classifier-Free Guidance (CFG)

- Tapasztalati tény: erősebb CFG skálafaktorral ($s_t > 1$) jobb eredményeket kapunk!!!
- Gyakorlatban akár $s_t = 5 - 8$!!!
- Minden “rendes” képgenerátor erős CFG-t használ!

CLASSIFIER-FREE DIFFUSION GUIDANCE

Jonathan Ho & Tim Salimans
Google Research, Brain team



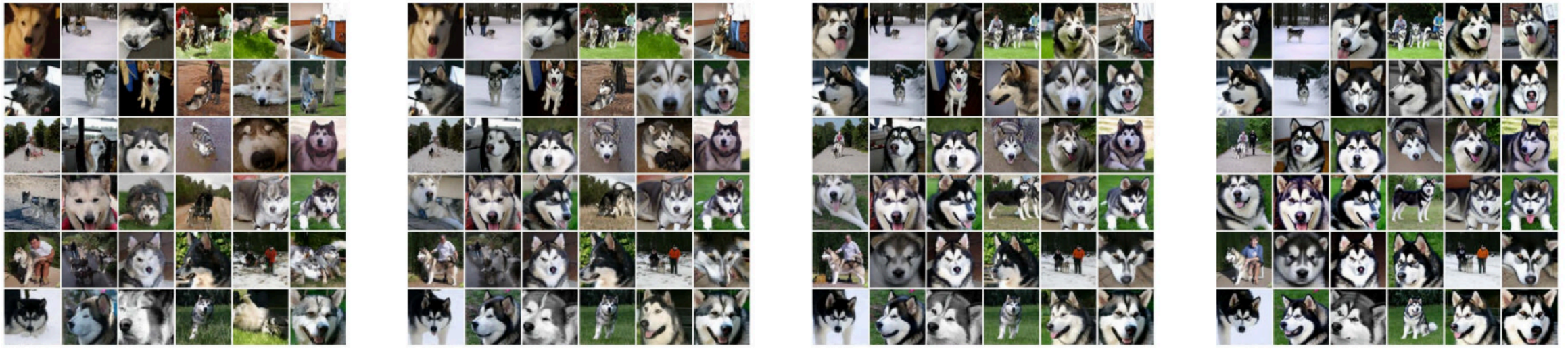
$s_t = 0$ $s_t = 3$

Vezérelt Generálás

Classifier-Free Guidance (CFG)

CLASSIFIER-FREE DIFFUSION GUIDANCE

Jonathan Ho & Tim Salimans
Google Research, Brain team



(a) $w = 0$ (b) $w = 1$ (c) $w = 2$ (d) $w = 4$

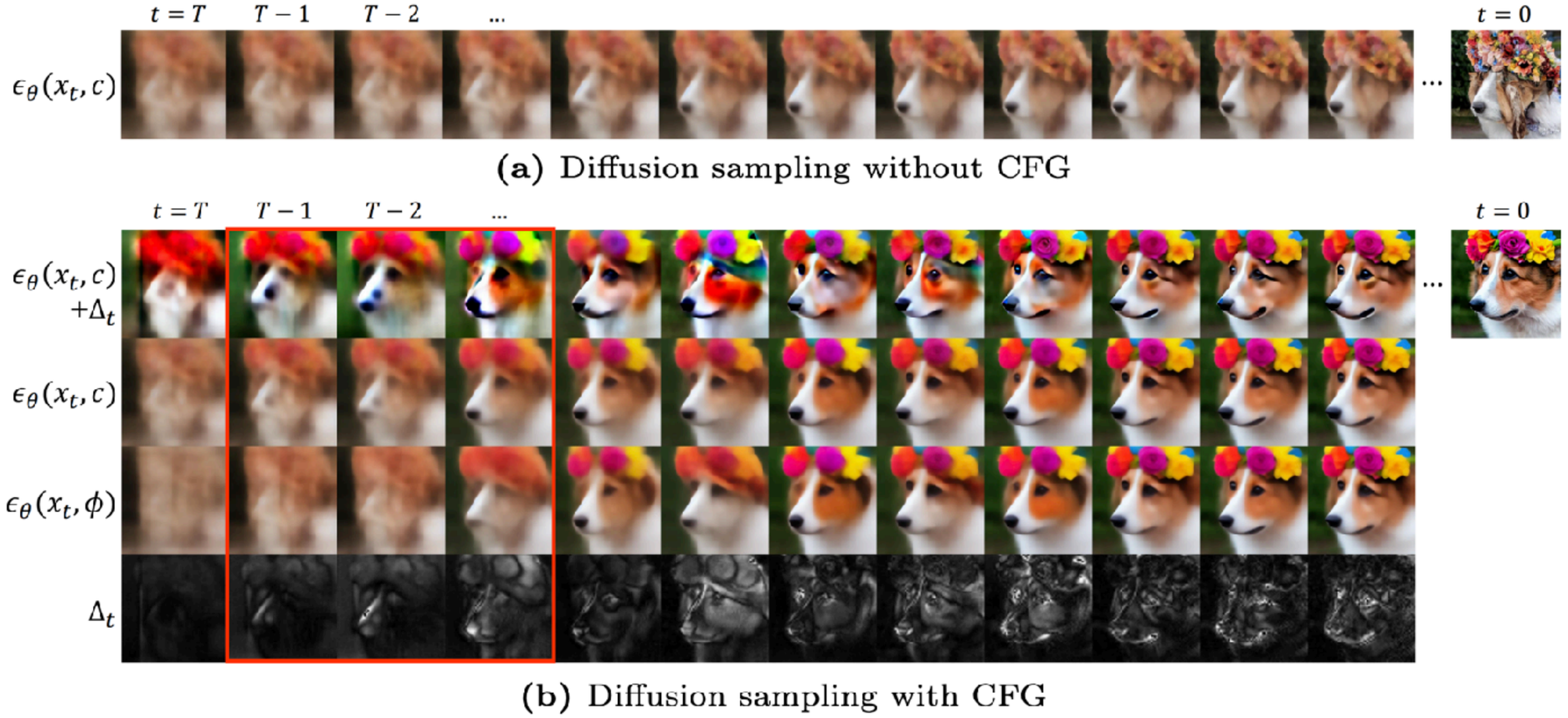
CFG skálázás hatása: jobb igazodás a prompthoz, de kisebb variancia és “szaturált” színek!

Vezérelt Generálás

Classifier-Free Guidance (CFG)

CLASSIFIER-FREE DIFFUSION GUIDANCE

Jonathan Ho & Tim Salimans
Google Research, Brain team



(a) Diffusion sampling without CFG

(b) Diffusion sampling with CFG

A CFG hamarabb a promptnak megfelelő irányba “húzza” a generálást!

Vezérelt Generálás

Perturbed-Attention Guidance (PAG)

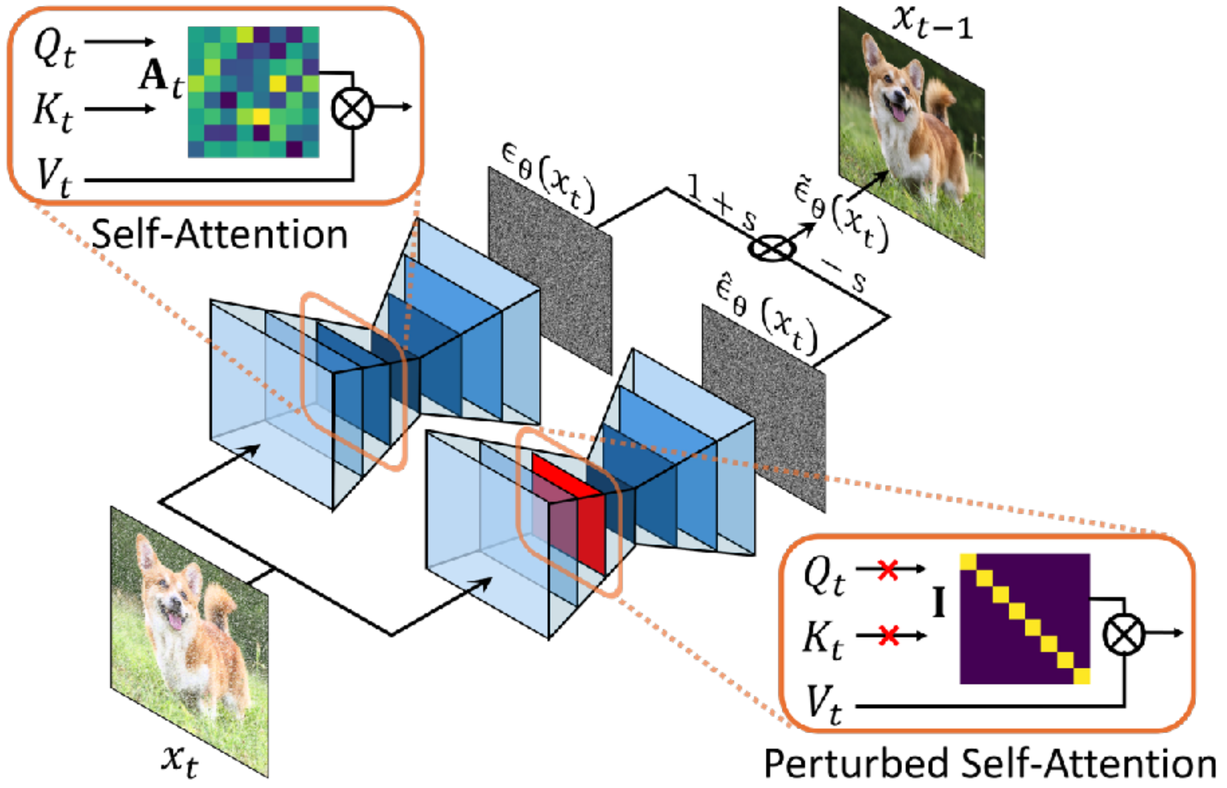
Self-Rectifying Diffusion Sampling with Perturbed-Attention Guidance

Donghoon Ahn*[1], Hyoungwon Cho*[1], Jaewon Min[1], Wooseok Jang[1], Jungwoo Kim[1], SeonHwa Kim[1], Hyun Hee Park[2], Kyong Hwan Jin†[1], and Seungryong Kim†[1]

[1] Korea University
[2] Samsung Electronics

https://ku-cvlab.github.io/Perturbed-Attention-Guidance

- Megfigyelés: a generálás korai (zajosabb) szakaszában néha “rossz” struktúrájú képek felé indulunk, amit nehéz korrigálni…
  - A struktúrát a self-attention (query-key szorzat) mátrixok határozzák meg!
- **Perturbed-Attention Guidance (PAG):** CFG-hez hasonló, csak feltétel nélküli kiértékelés helyett a self-attention mátrixokat “perturbáljuk" (egységmátrixra cseréljük) és kombináljuk a normál attention eredményével!



Vezérelt Generálás

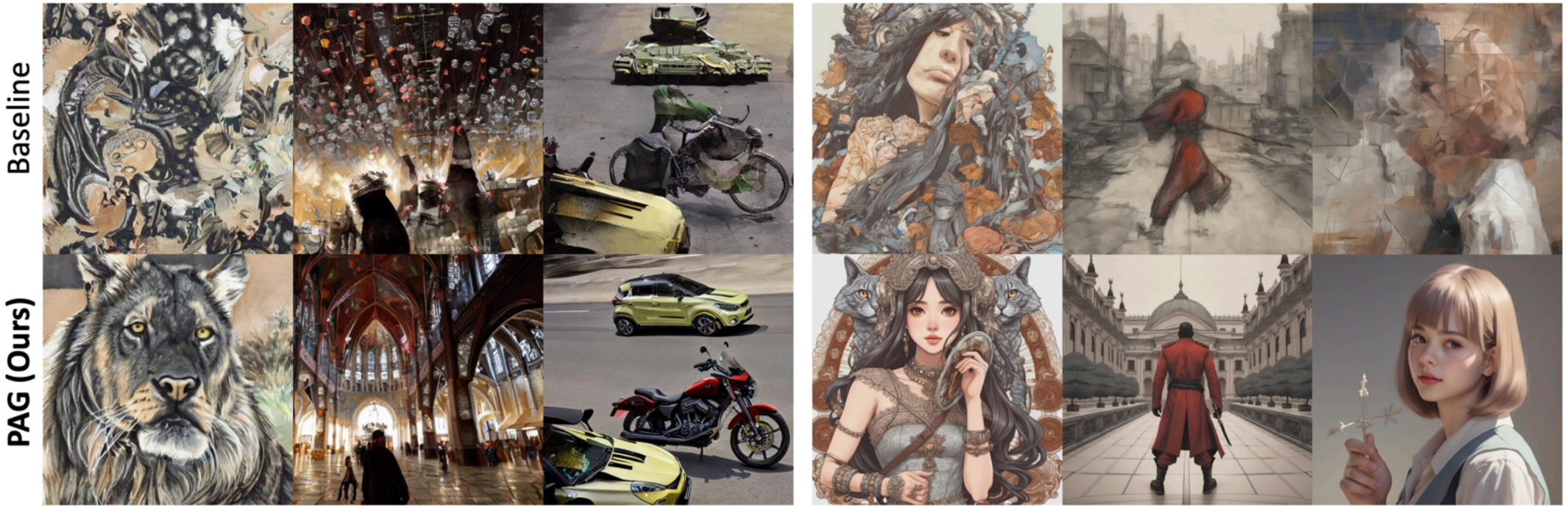
Perturbed-Attention Guidance (PAG)

Self-Rectifying Diffusion Sampling with Perturbed-Attention Guidance

Donghoon Ahn*[1], Hyoungwon Cho*[1], Jaewon Min[1], Wooseok Jang[1], Jungwoo Kim[1], SeonHwa Kim[1], Hyun Hee Park[2], Kyong Hwan Jin†[1], and Seungryong Kim†[1]

[1] Korea University
[2] Samsung Electronics

https://ku-cvlab.github.io/Perturbed-Attention-Guidance



Stable Diffusion 1.5

SDXL

A PAG a feltétel nélkül generált képeken is sokat javít!

Vezérelt Generálás

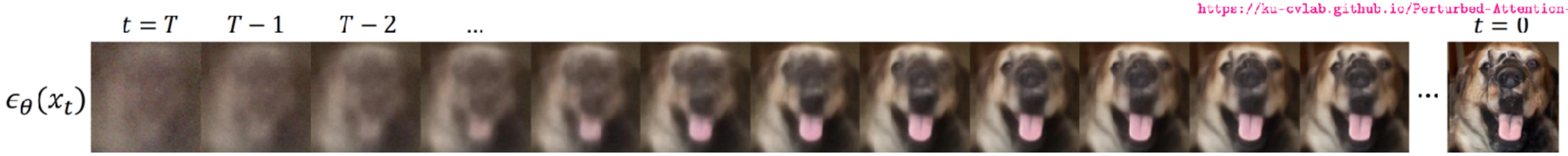
Perturbed-Attention Guidance (PAG)

Self-Rectifying Diffusion Sampling
with Perturbed-Attention Guidance

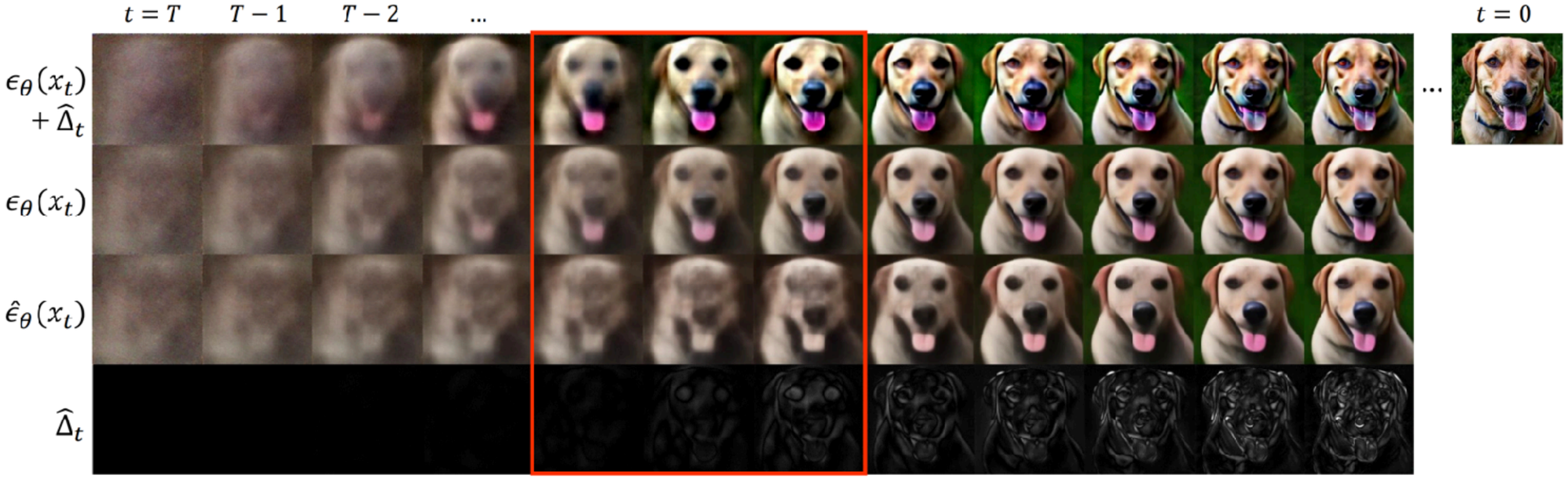
Donghoon Ahn*[1], Hyoungwon Cho*[1], Jaewon Min[1], Wooseok Jang[1],
Jungwoo Kim[1], SeonHwa Kim[1], Hyun Hee Park[2],
Kyong Hwan Jin†[1], and Seungryong Kim†[1]

[1] Korea University
[2] Samsung Electronics

https://ku-cvlab.github.io/Perturbed-Attention-Guidance



(a) Sampling process without PAG



(b) Sampling process with PAG

Vezérelt Generálás

Promptolás képpel

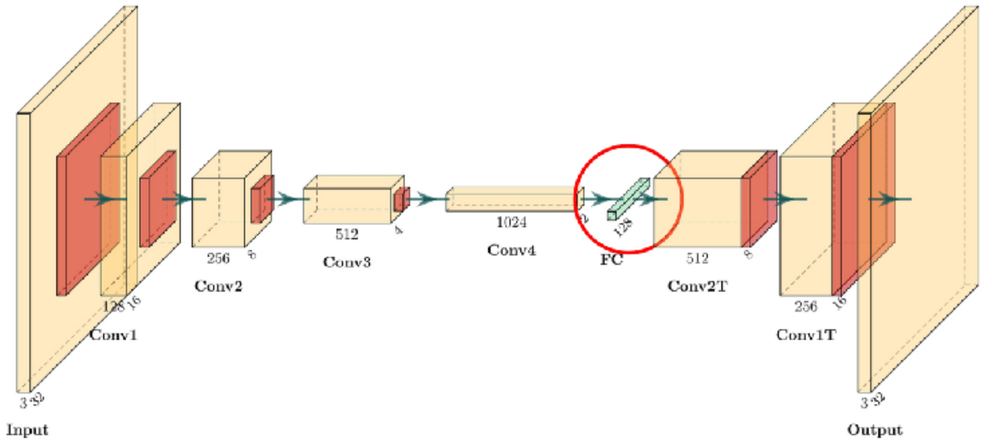
Képi promptok

Generált képek

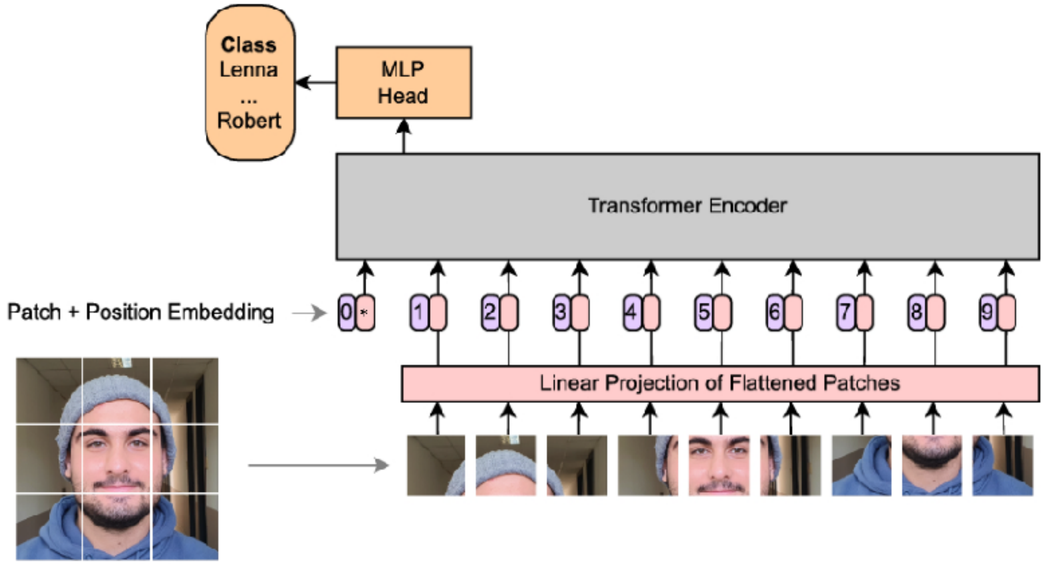
Hogyan kódoljunk be képi feltételeket?

Képek Kódolása

VAE, ViT



Konvolúciós (V)AE



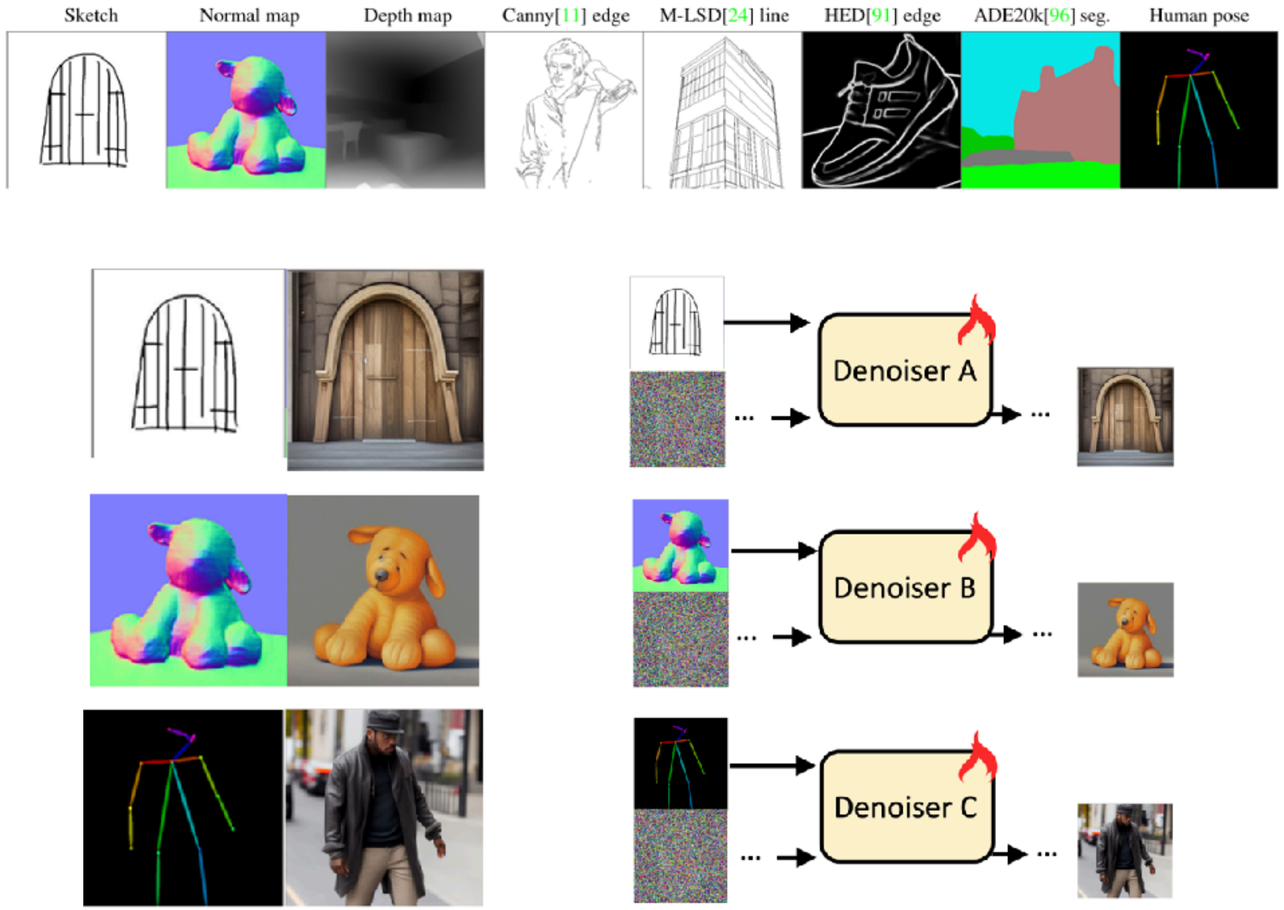
Vision Transformer (ViT)

(pl. CLIP / MAE / DINO)

Finomhangolás

Képi kondíciók figyelembe vétele

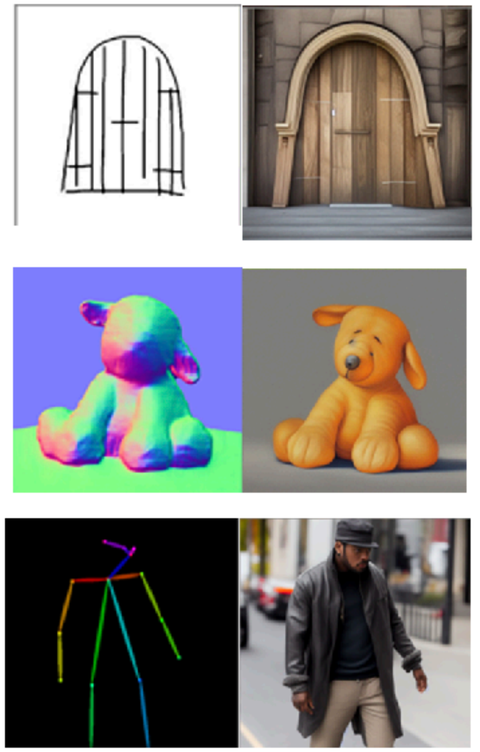
- Hogyan tanítsunk meg egy diffúziós modellt korábban nem látott kondíciókra — pl. képekre?
- Adjuk hozzá az új kondíciót a korábbiakhoz, aztán **finomhangoljuk** példák alapján?
  - Rengeteg paramétert kéne hangolni (pl. SD 3.5: 8B!)
  - Kevés lesz az adat (az eredeti hálót akár több *milliárd* képen tanították…)
  - Minden újabb kondíció komoly idő és compute ráfordítást igényelne
  - Összezavarhatjuk vele az eredeti hálót (katasztrofális felejtés)

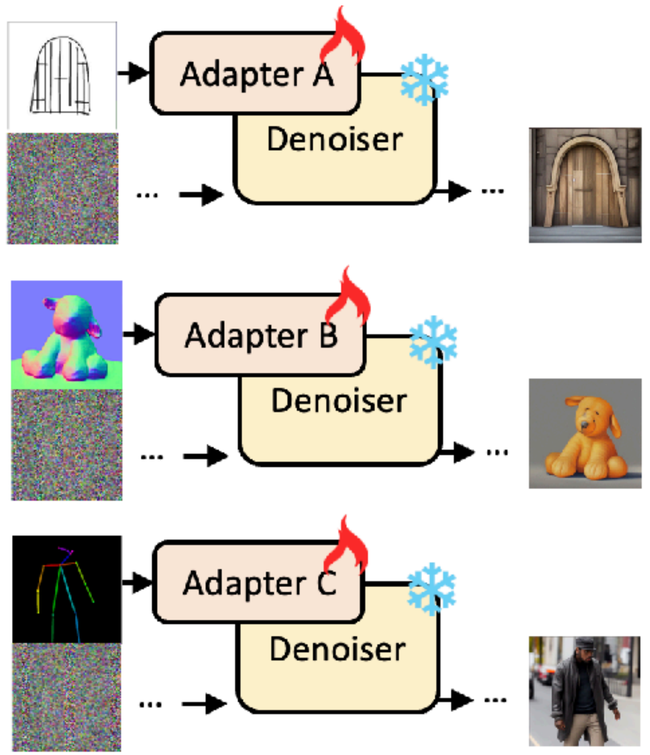


Finomhangolás

Adapterek

- Ötlet: az eredeti nagy diffúziós hálót “fagyasszuk be”, csak egy (jóval kisebb) **adapter** háló paraméteit tanítsunk!
- Népszerű adapterek:
  - LoRA
  - ControlNet
  - IP-Adapter

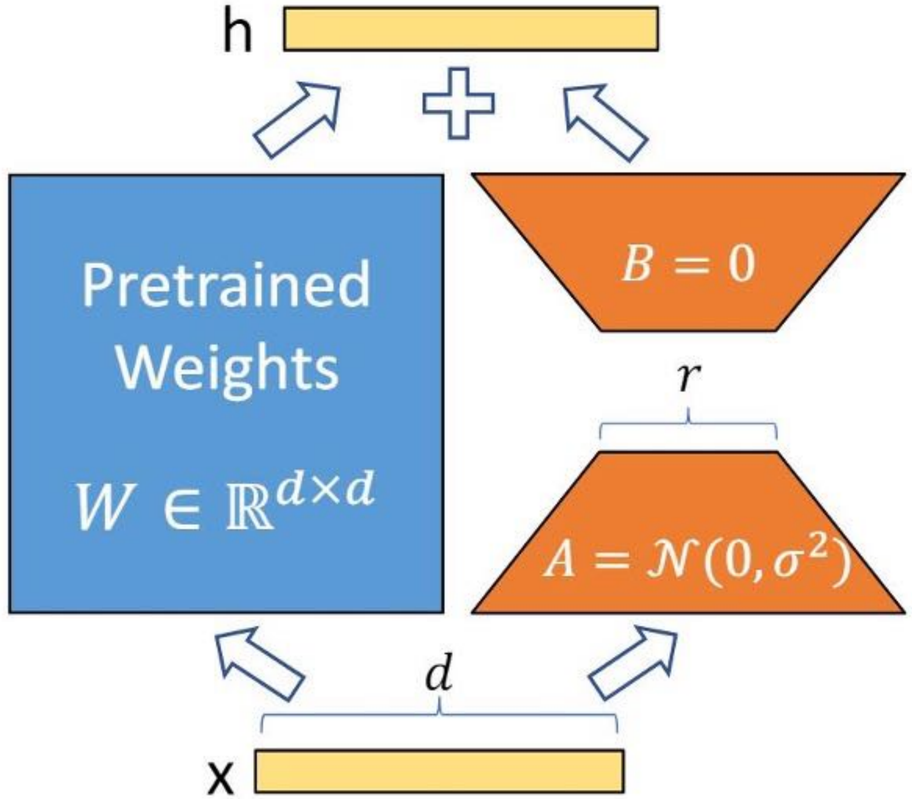




Finomhangolás

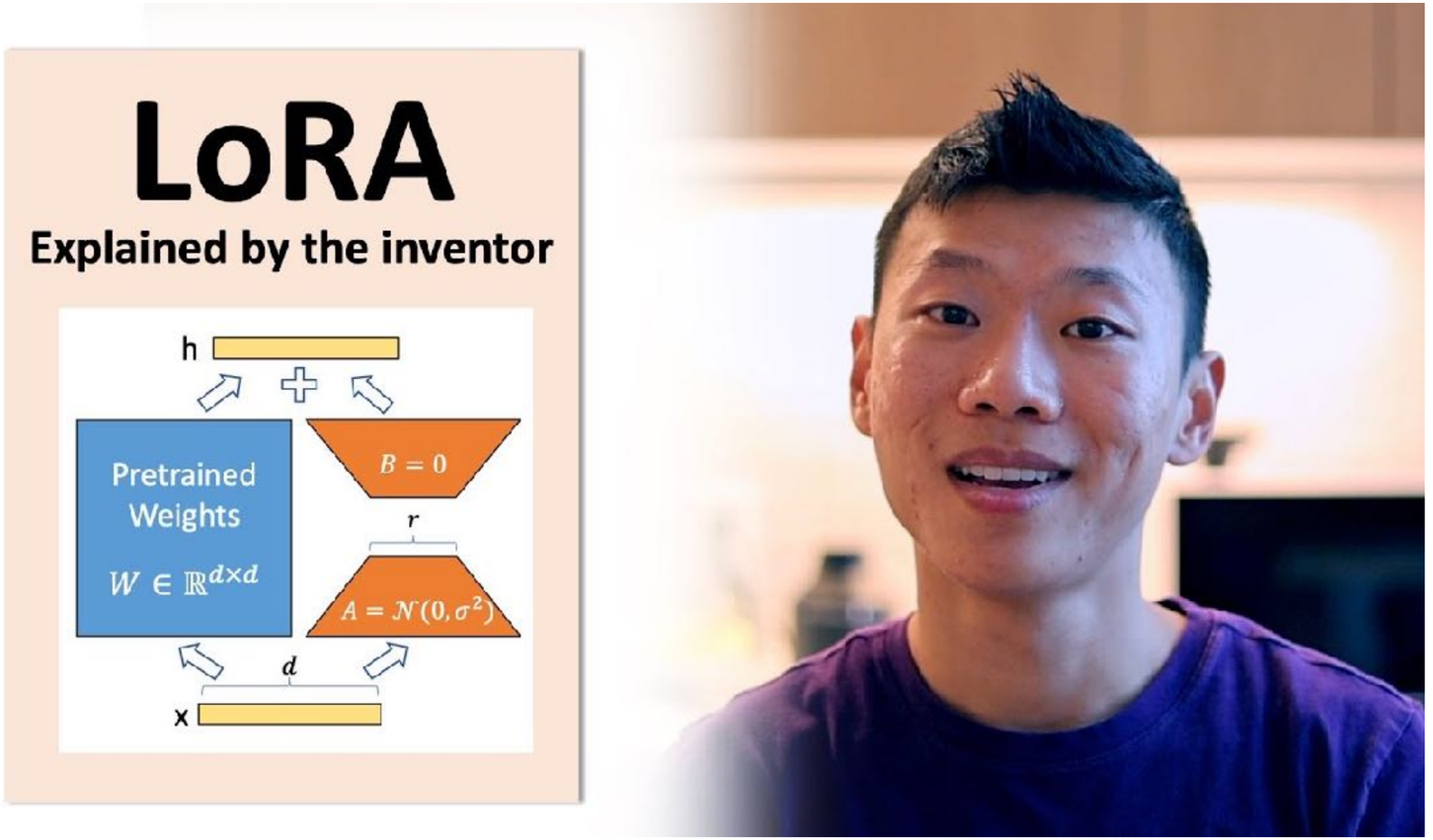
Adapterek — LoRA

- Megfigyelés: finomhangolás során a hálóparaméterek (pl. QKV mátrixok) alacsony dimenziós altérben változnak!
- Paraméterezzük a módosításokat alacsony rangú mátrixként (külső szorzatként) — **Low-Rank Adaptation (LoRA)**
- A LoRA mátrixok kis méretűek, gyorsan taníthatóak és több LoRA akár *kombinálható is*!
- Nagyon népszerű módszer — főleg LLM-ek és képgenerátorok attention rétegeinek hangolására
  - Nagy modellek finomhangolása (pl. A többi adapter implementációja is) általában LoRA-val történik!



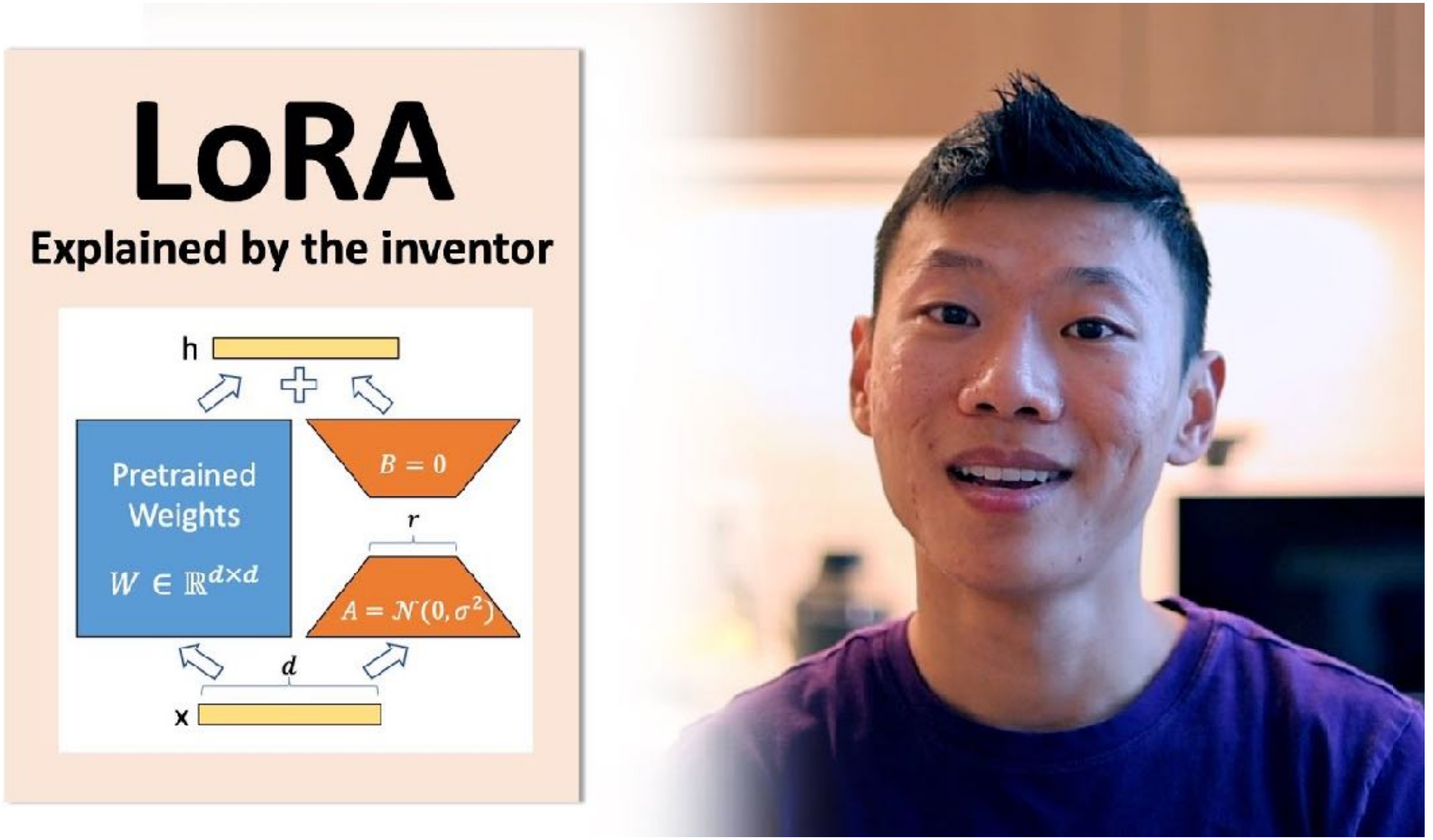
Finomhangolás

Adapterek — LoRA — Videóajánló



Finomhangolás

Adapterek — LoRA — Videóajánló

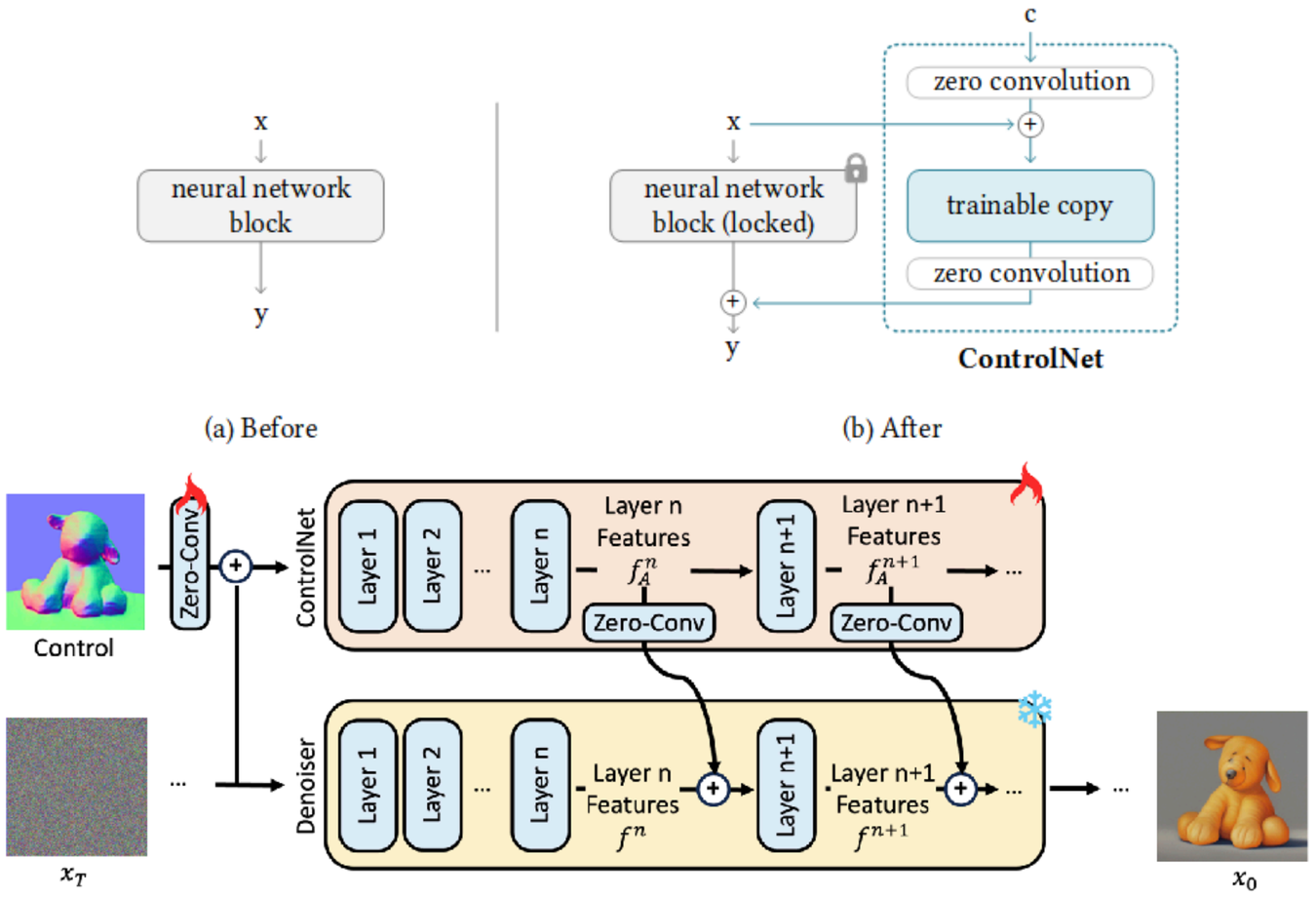


Finomhangolás

Adapterek — ControlNet

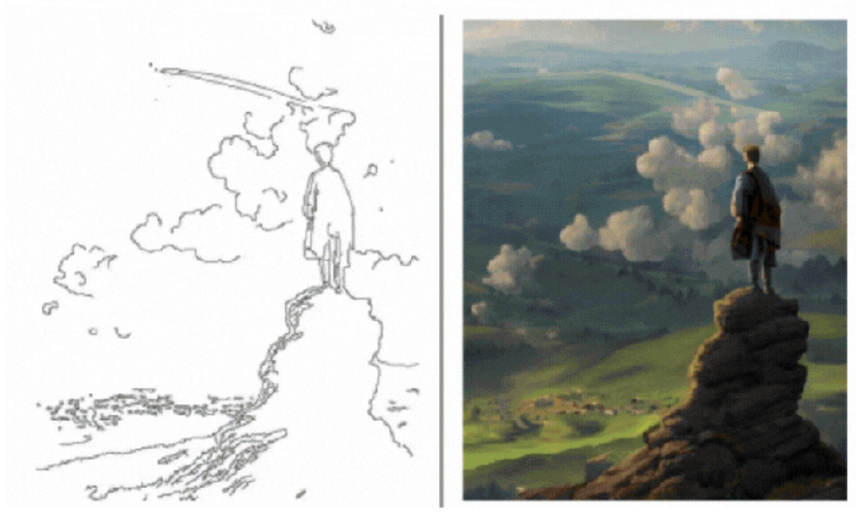
- Finomhangoljuk a háló másolatát, és adjuk hozzá az eredetihez — **ControlNet**

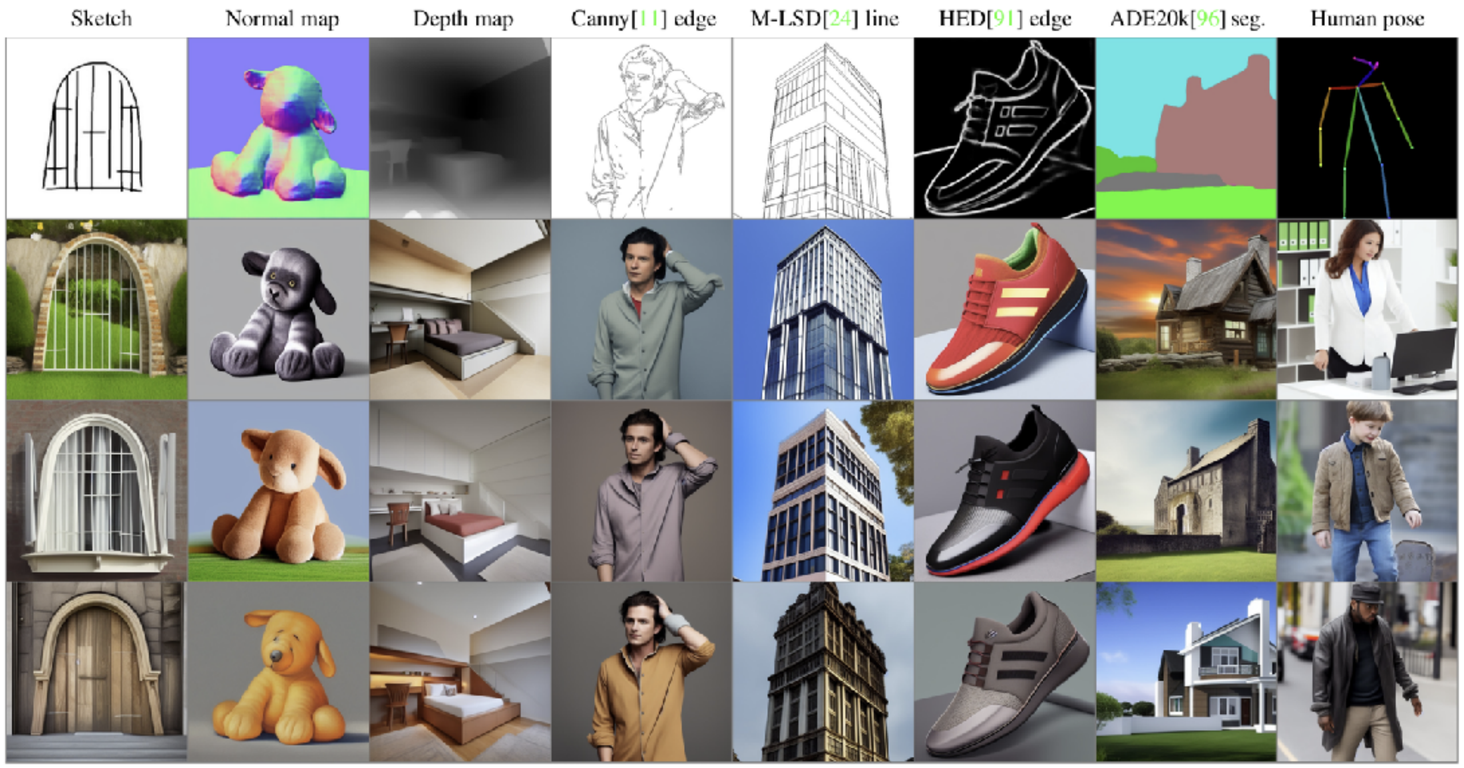
- Kezdetben a másolatnak ne legyen hatása — nullára inicializált konvolúciós (Zero-Conv) rétegekkel adjuk össze!



Finomhangolás

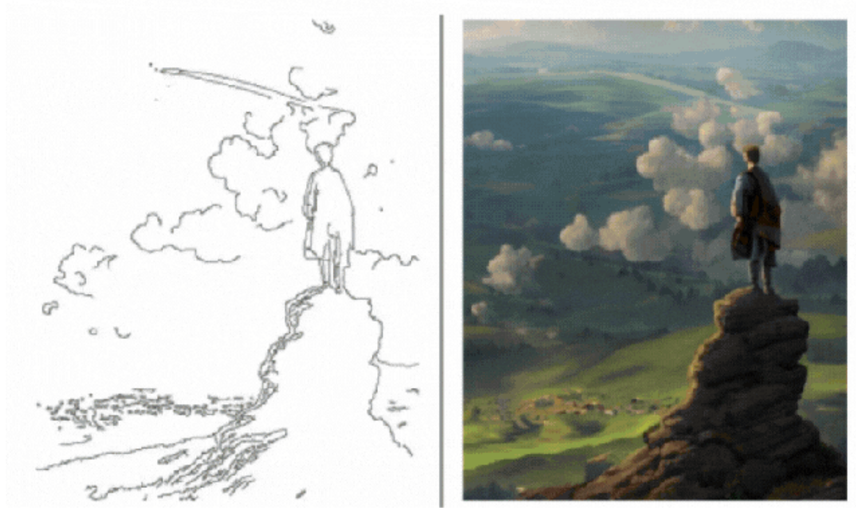
Adapterek — ControlNet

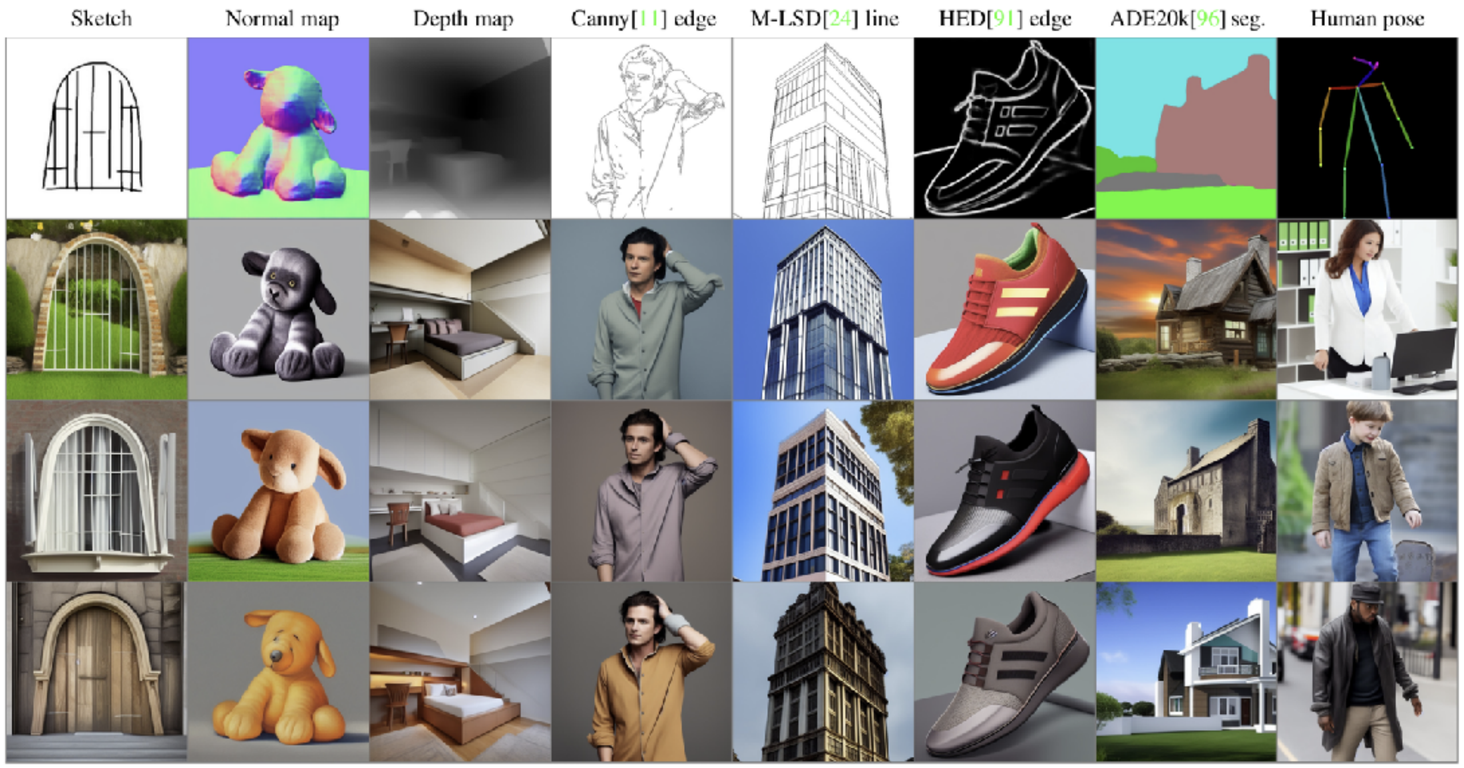




Finomhangolás

Adapterek — ControlNet





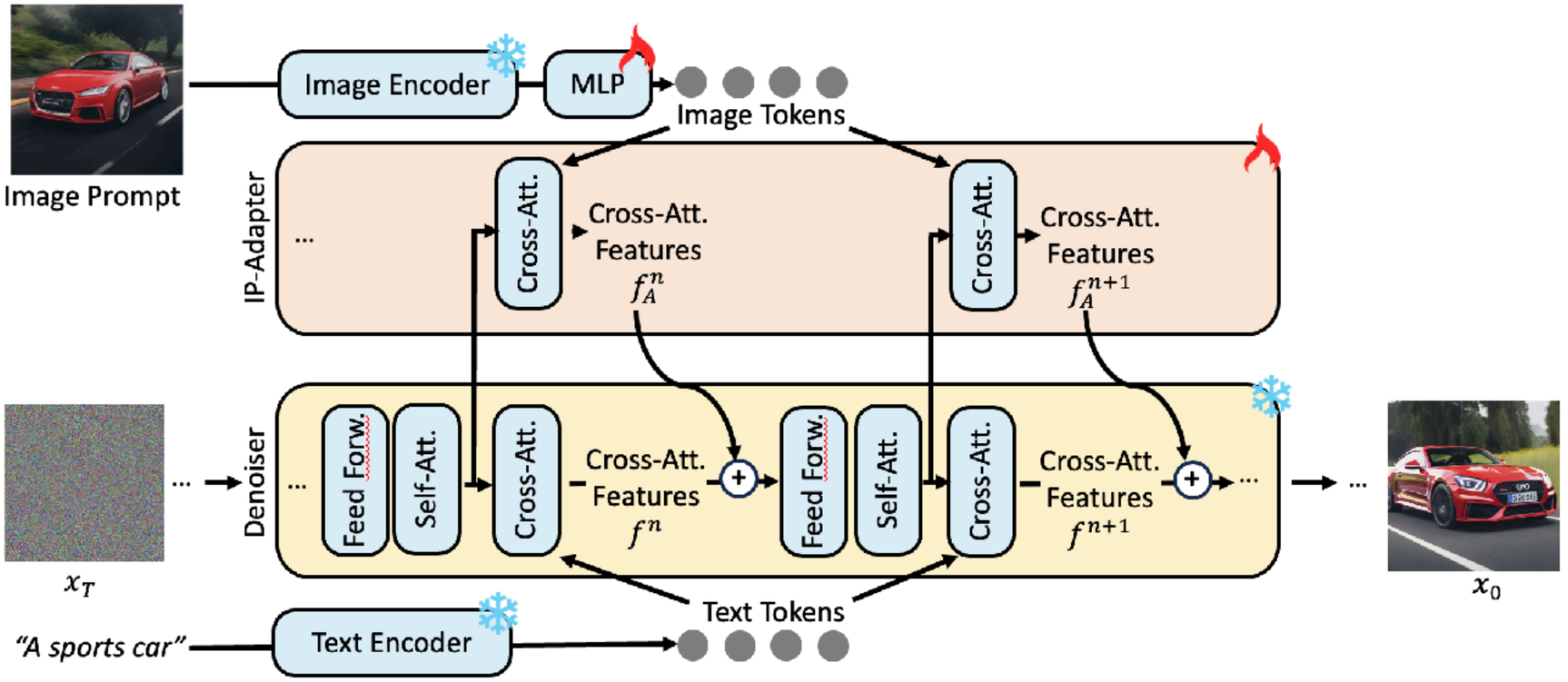
Finomhangolás

Adapterek — ControlNet

ControlNet (szöveges prompt nélkül) + PAG

Finomhangolás

Adapterek — IP-Adapter



Finomhangolás

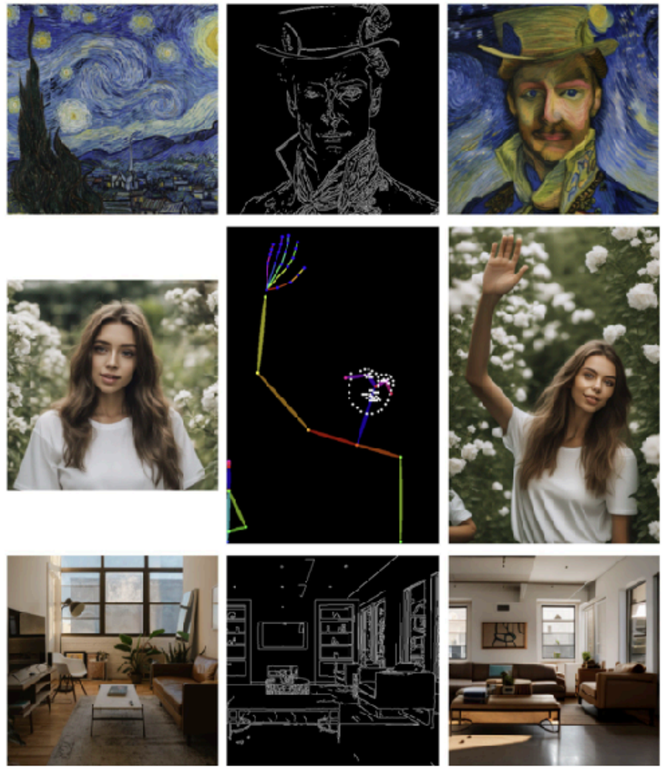
Adapterek — IP-Adapter

IP-Adapter Only

input generated

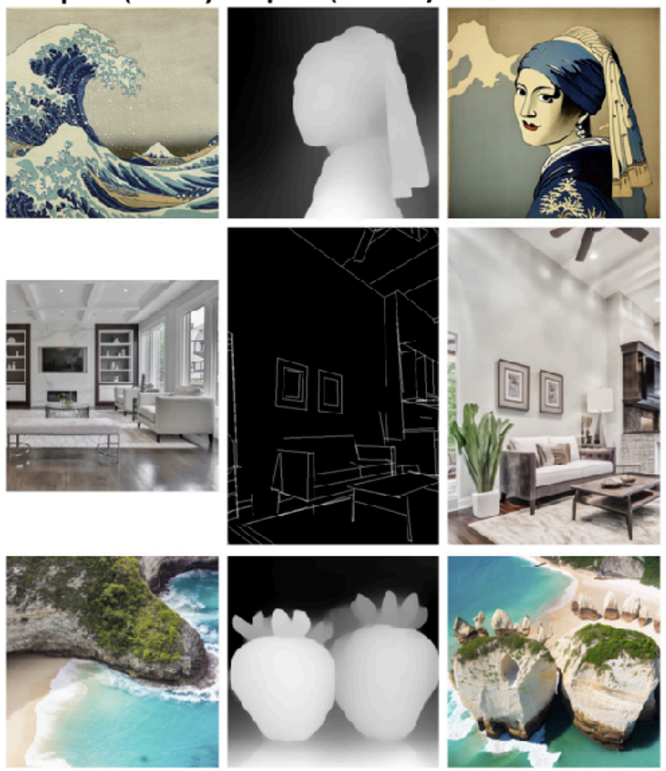
IP-Adapter + ControlNet

Input (IP-A.) Input (C.Net) generated



IP-Adapter + ControlNet

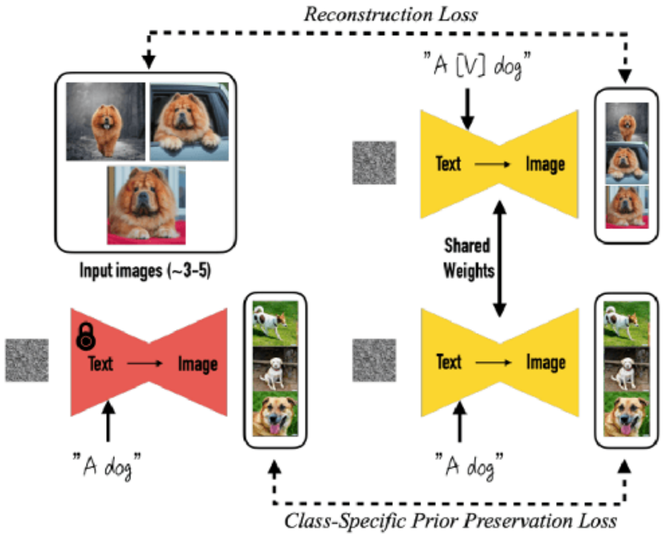
Input (IP-A.) Input (C.Net) generated



Finomhangolás

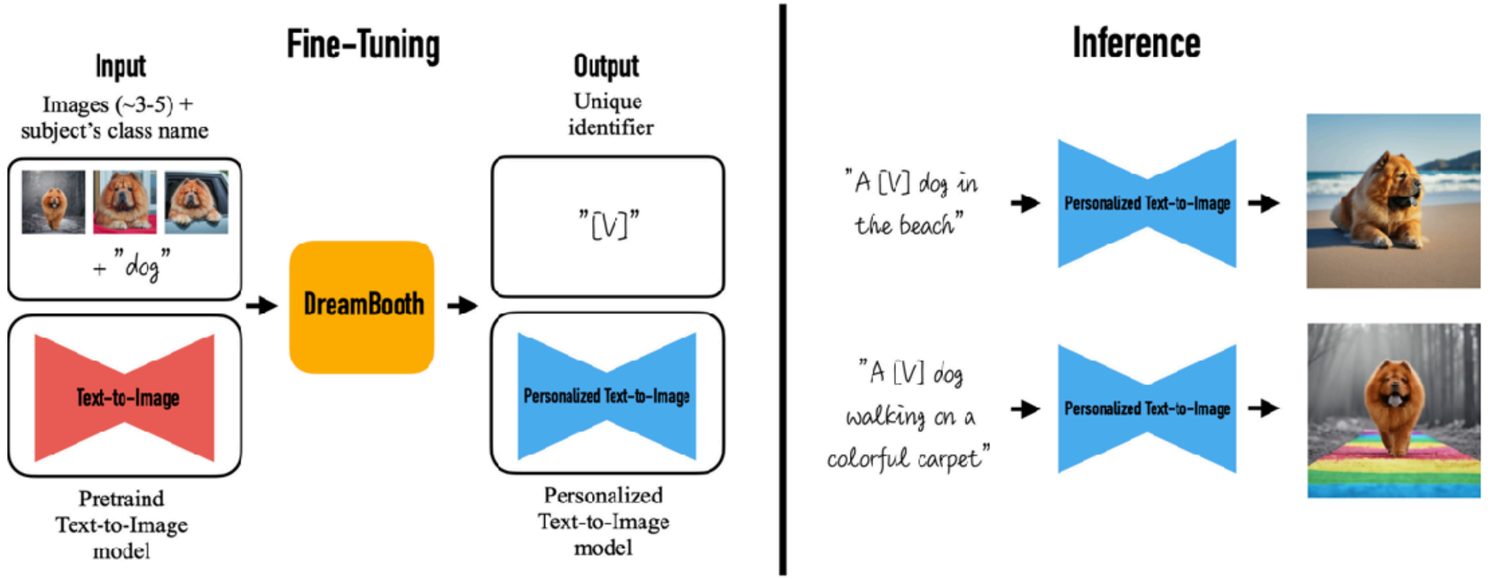
DreamBooth

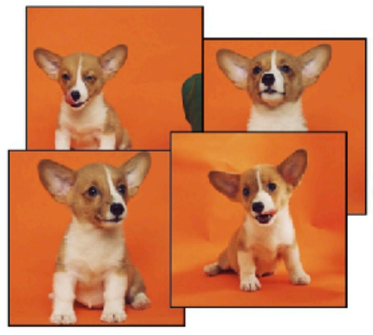
- **DreamBooth**: finomhangolás kevés (3-5) példakép alapján!
- Eredmény: finomhangolt modell + specializált szöveges token



DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation

Nataniel Ruiz*,[1,2] Yuanzhen Li[1] Varun Jampani[1]
Yael Pritch[1] Michael Rubinstein[1] Kfir Aberman[1]
[1] Google Research [2] Boston University



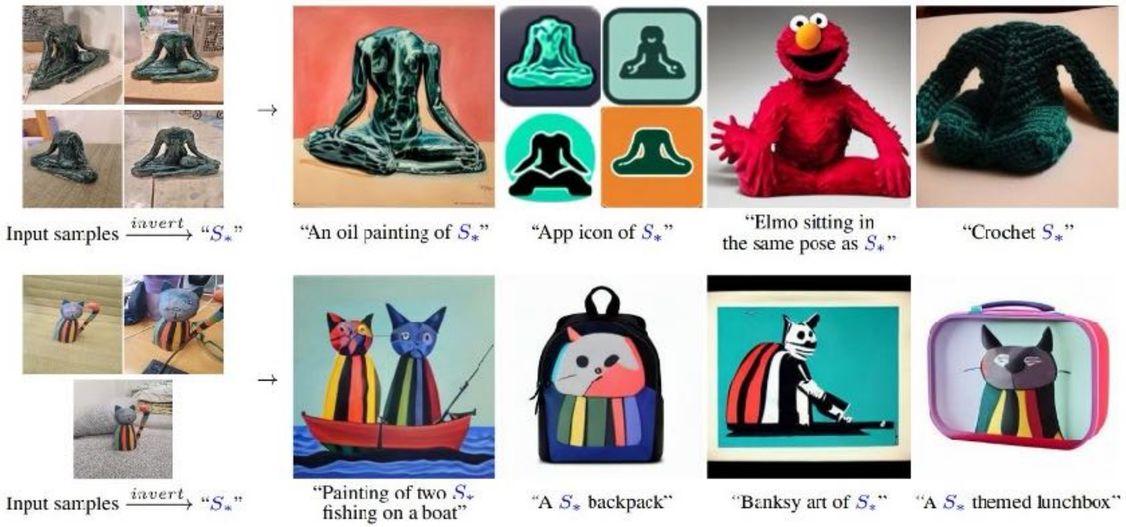


Input images

Finomhangolás

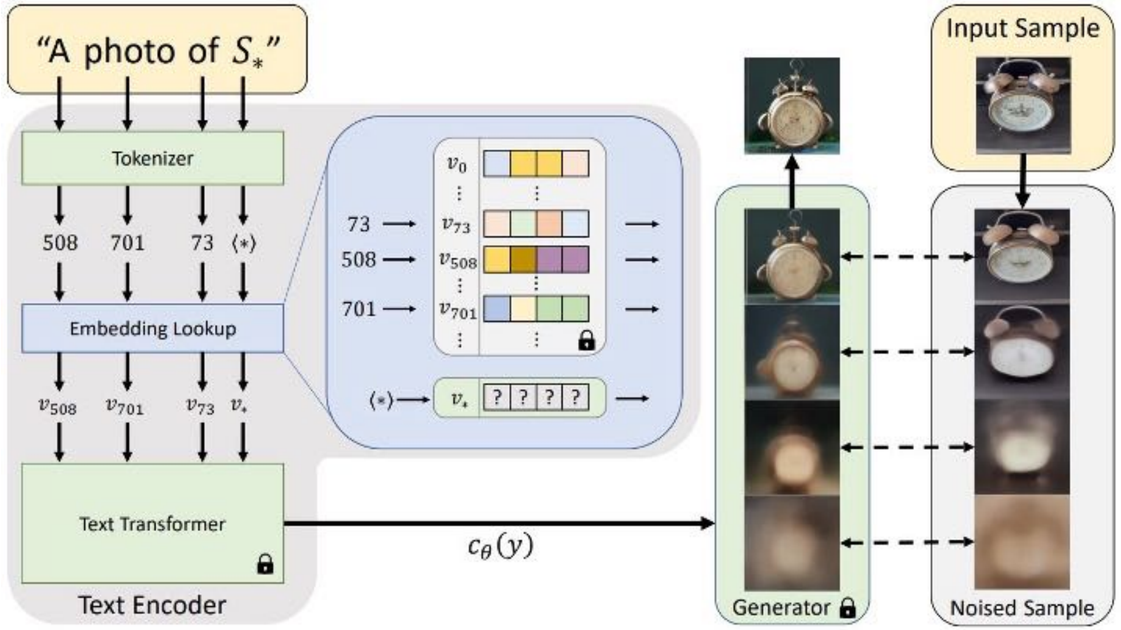
Szöveges inverzió

- **Szöveges Inverzió** — a példaképek optimalizálunk egy egyedi szöveges tokent!
- A képgenerátort nem finomhangoljuk, *csak a tokent*!



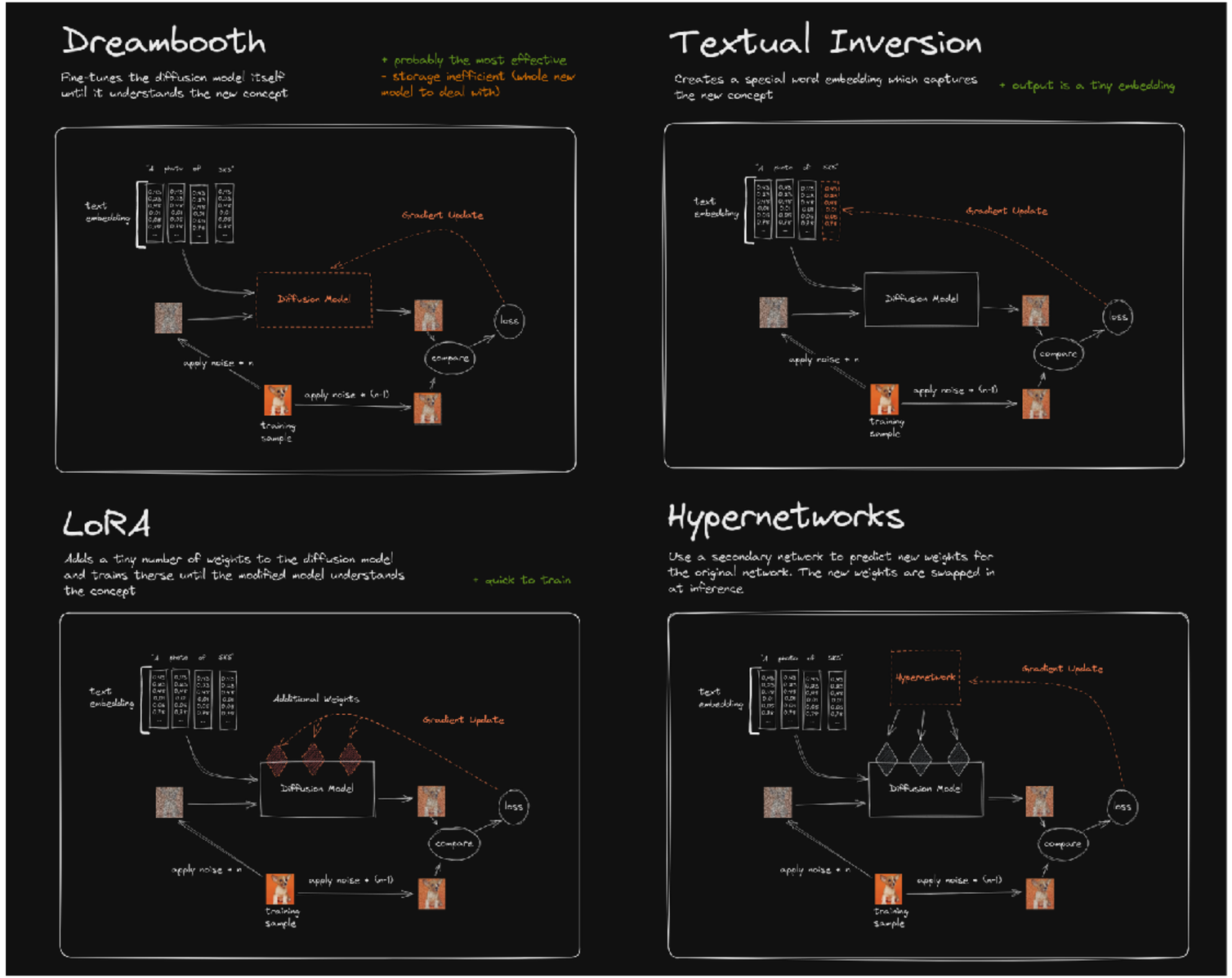
An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion

Rinon Gal[1,2], Yuval Alaluf[1], Yuval Atzmon[2], Or Patashnik[1], Amit H. Bermano[1], Gal Chechik[2], Daniel Cohen-Or[1],
[1]Tel Aviv University, [2]NVIDIA



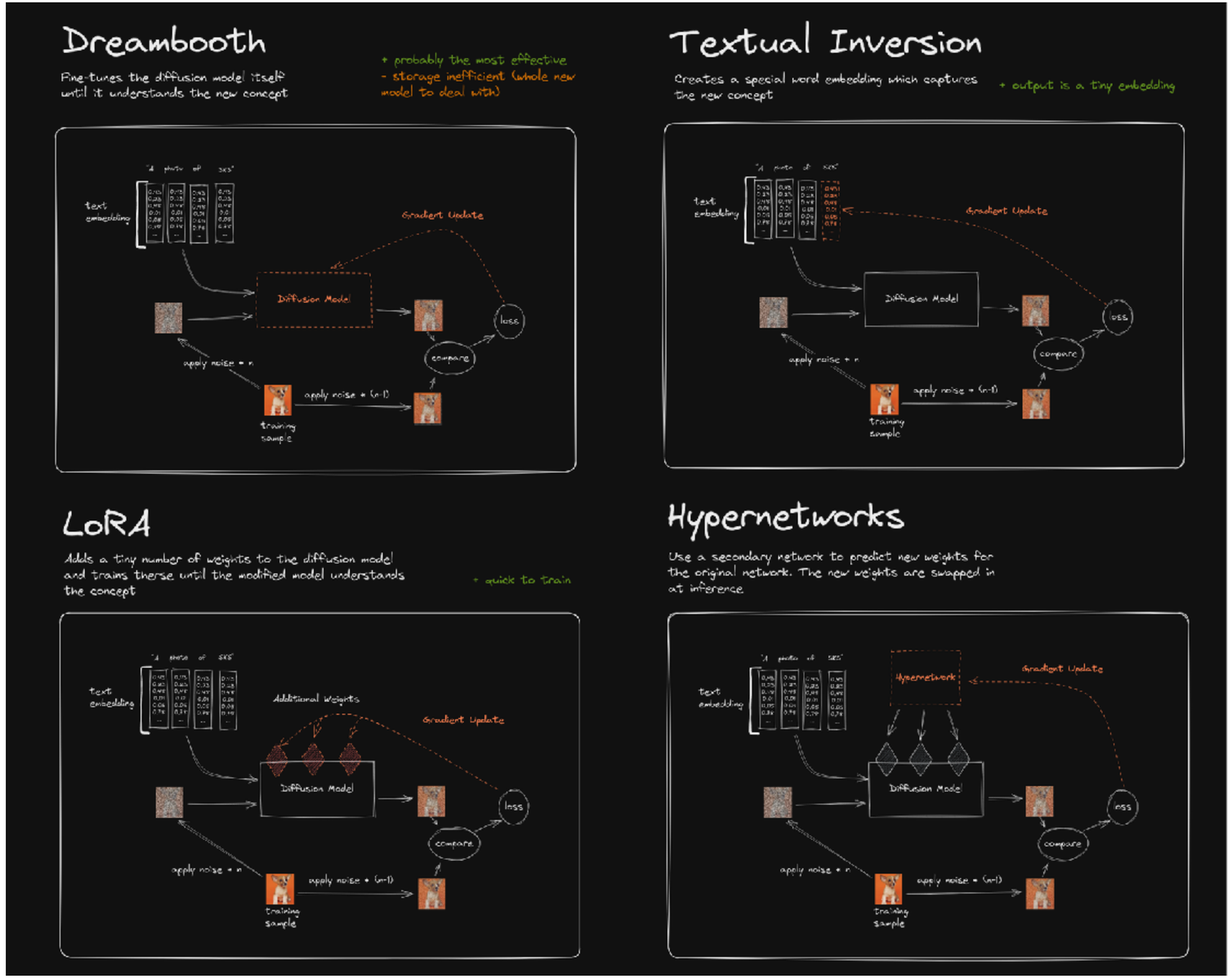
Finomhangolás

Áttekintés



Finomhangolás

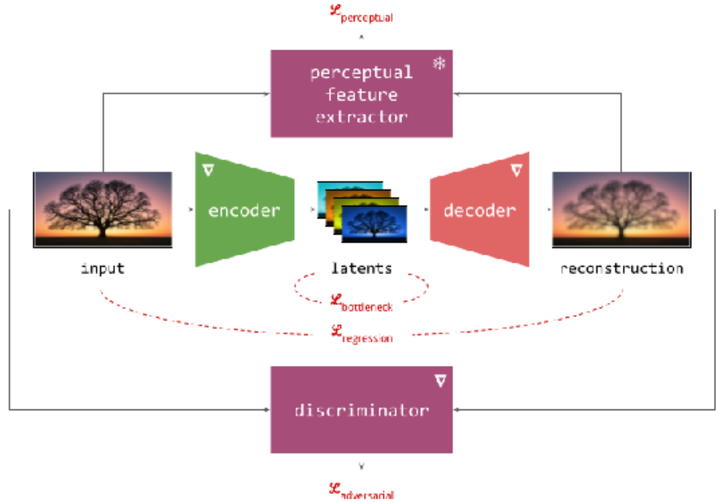
Áttekintés

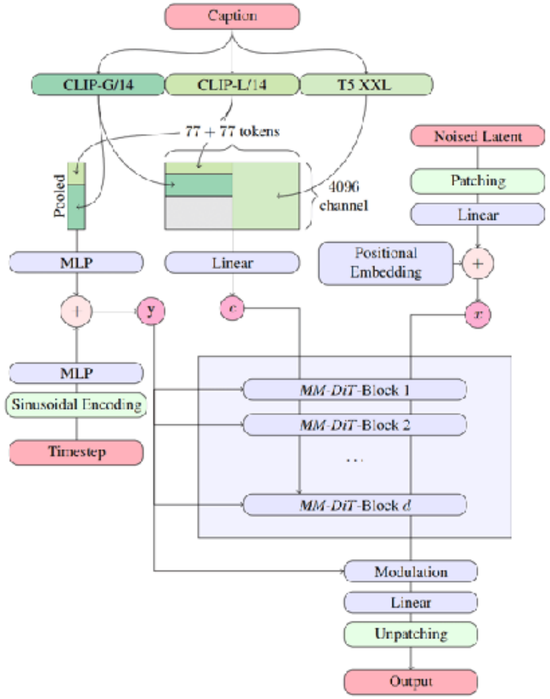


Következő előadás:

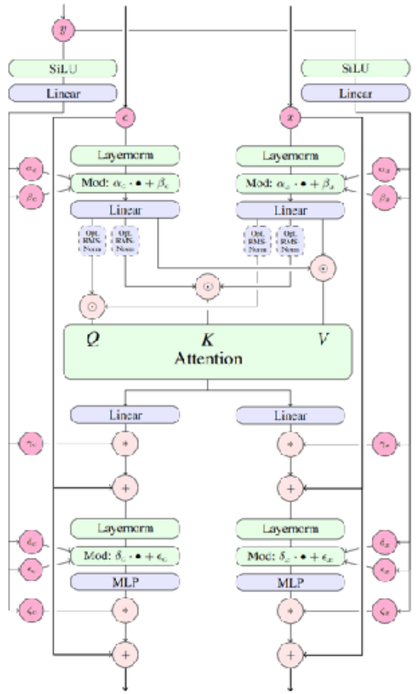
Diffúziós Generálás a Gyakorlatban

- Látens diffúzió
- Esettanulmányok:
  - Dall-E, Imagen
  - Stable Diffusion 1-3
  - Stb.





(a) Overview of all components.



(b) One *MM-DiT* block