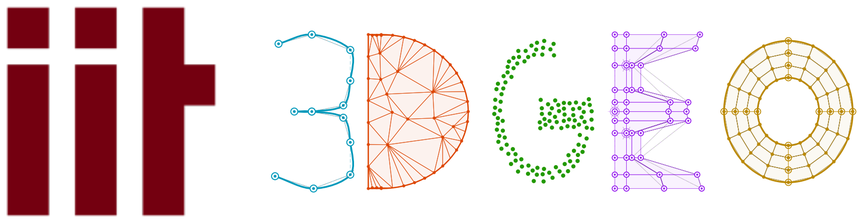
11. Előadás: Haladó Diffúziós Generálás

Generatív AI és Inverz Módszerek a Képszintézisben

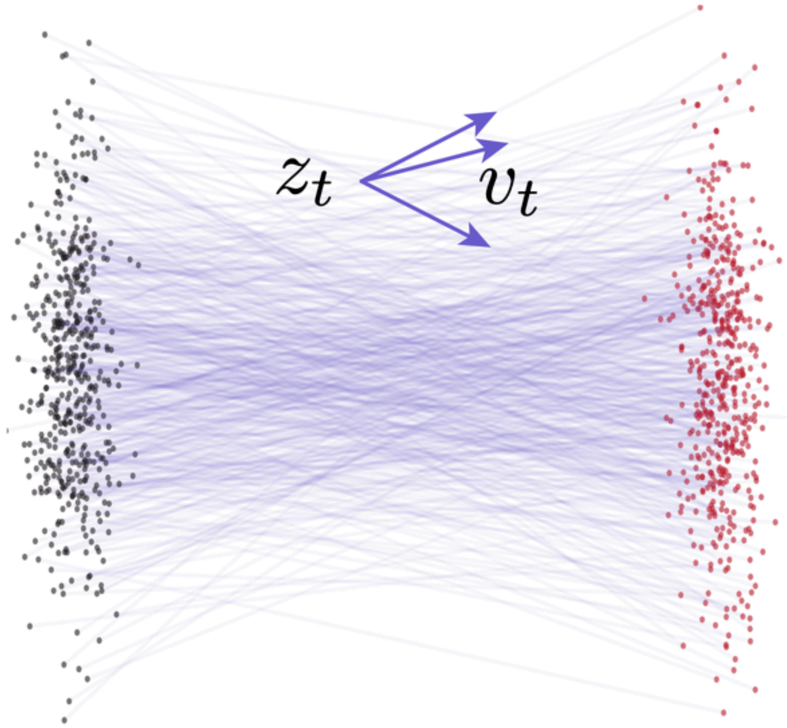
BME-VIK IIT, 2026



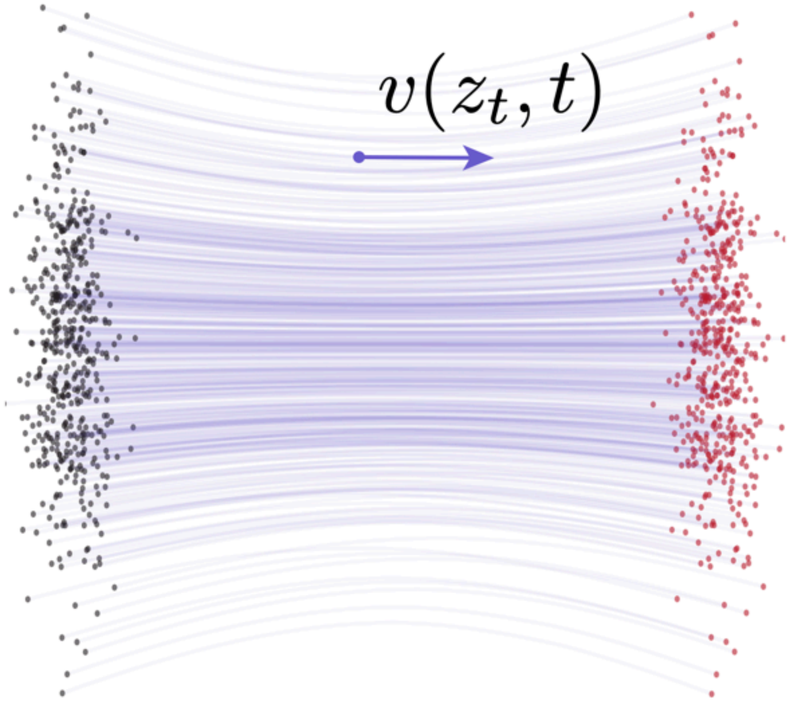
Dr. Vaitkus Márton

Emlékeztető

Folyamillesztés



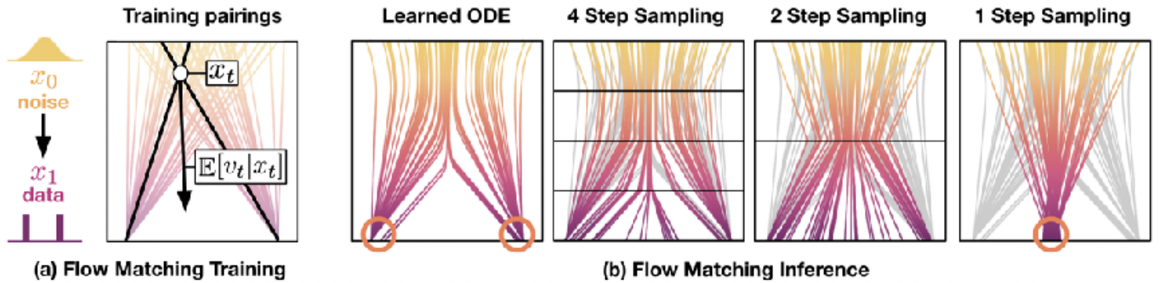
Feltételes folyam



Marginális folyam

Gyorsított Generálás

- A diffúziós és folyam-alapú generatív modellek sok iterációt igényelnek…
- A lépésszám csökkentésének egy ponton túl a minőség látja kárát…
- Két lehetőség a gyorsított generálásra:
  - **<u>Módosítjuk</u>** a modellt, hogy kevesebb lépésre is “jó” eredményeket generáljon (pl. Rektifikált folyamok)
  - **<u>Disztilláció</u>**: a lassú, de jó minőségű modell “tanárként" szolgál egy gyorsabb “diák" modell tanításához

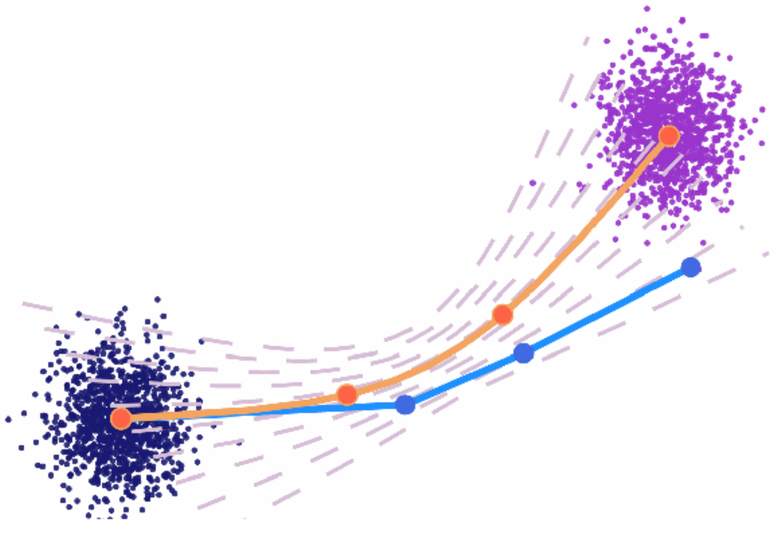


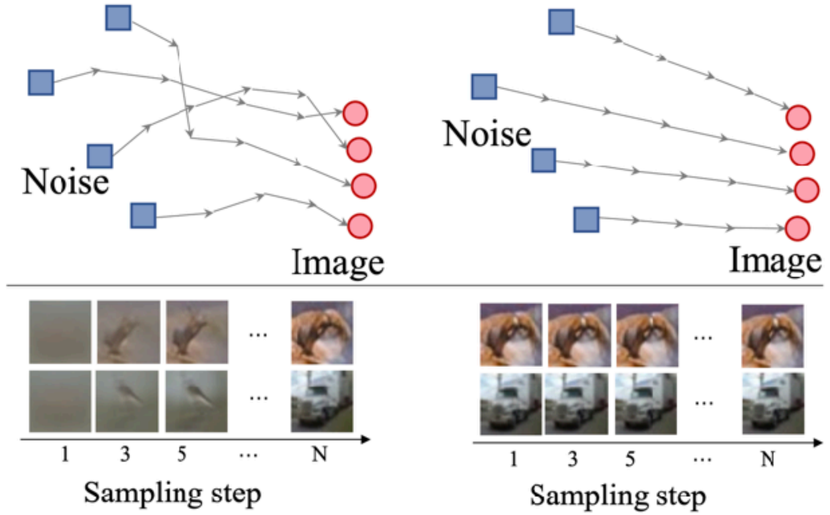
https://sander.ai/2024/02/28/paradox.html

Gyorsított Generálás

Rektifikált folyam (Rectified flow)

- Megfigyelés: “egyenesesebb” trajektóriák esetén kevesebb integrációs lépés is elég
- A folyamillesztéssel produkált trajektóriák általában elég egyenesek — alternatív elnevezés/formalizmus: **rektifikált folyam** (rectified flow)
  - Figyelem: a kisdimenziós példák valamennyire félrevezetőek!

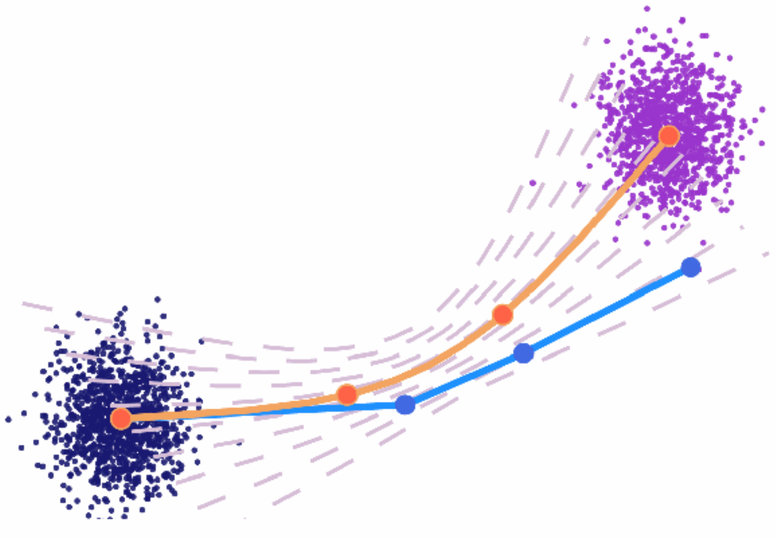


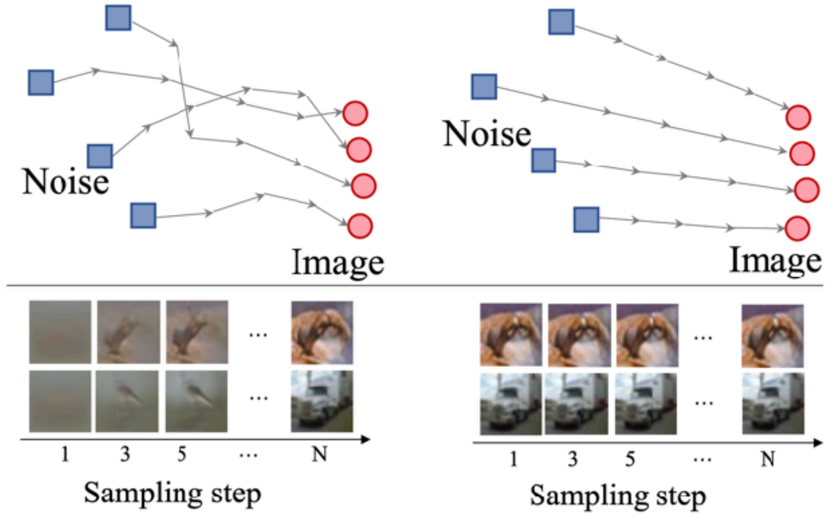


Gyorsított Generálás

Rektifikált folyam (Rectified flow)

- Megfigyelés: “egyenesesebb” trajektóriák esetén kevesebb integrációs lépés is elég
- A folyamillesztéssel produkált trajektóriák általában elég egyenesek — alternatív elnevezés/formalizmus: **rektifikált folyam** (rectified flow)
  - Figyelem: a kisdimenziós példák valamennyire félrevezetőek!



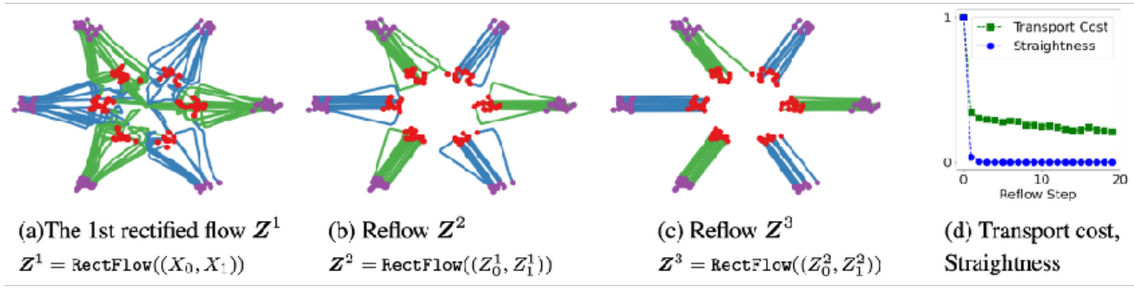


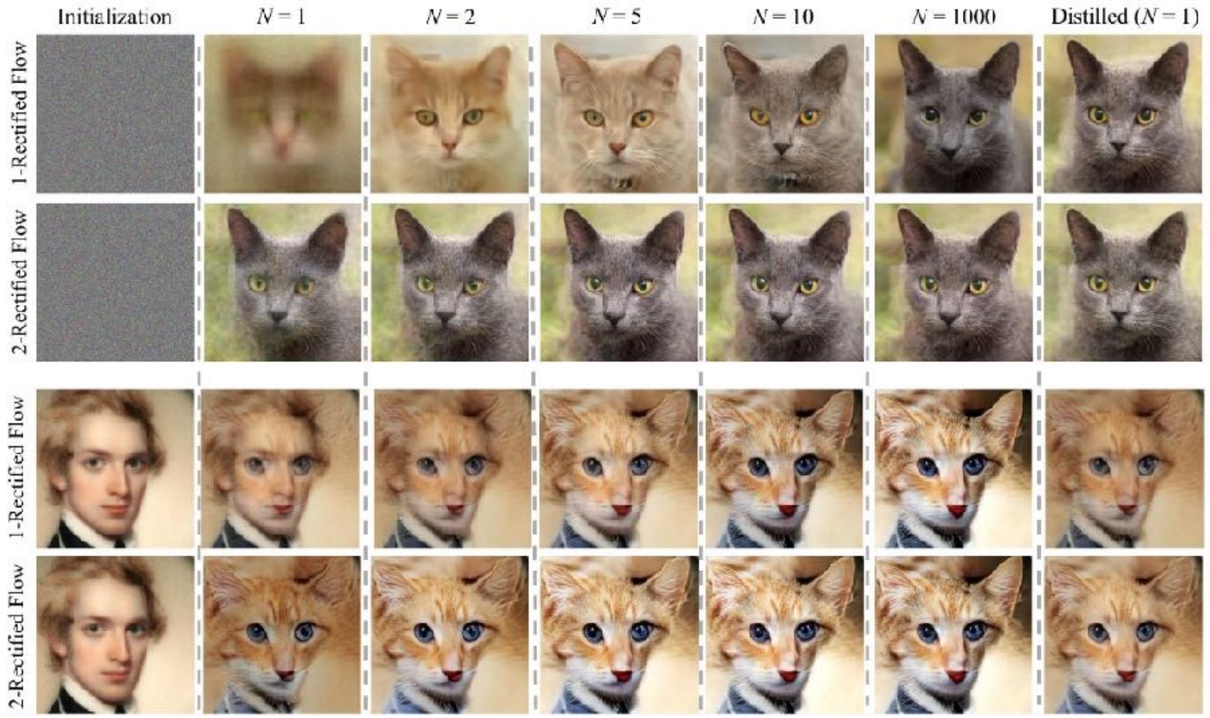
Gyorsított Generálás

Rektifikált folyam (Rectified flow)

- Reflow — iterált rektifikált folyam: definiáljuk át a zaj-adat összerendelést az aktuálisan illesztett folyam trajektóriái szerint és ez alapján ismételjük a folyamillesztést

- A folyam egyre egyenesebb trajektóriákat definiál!



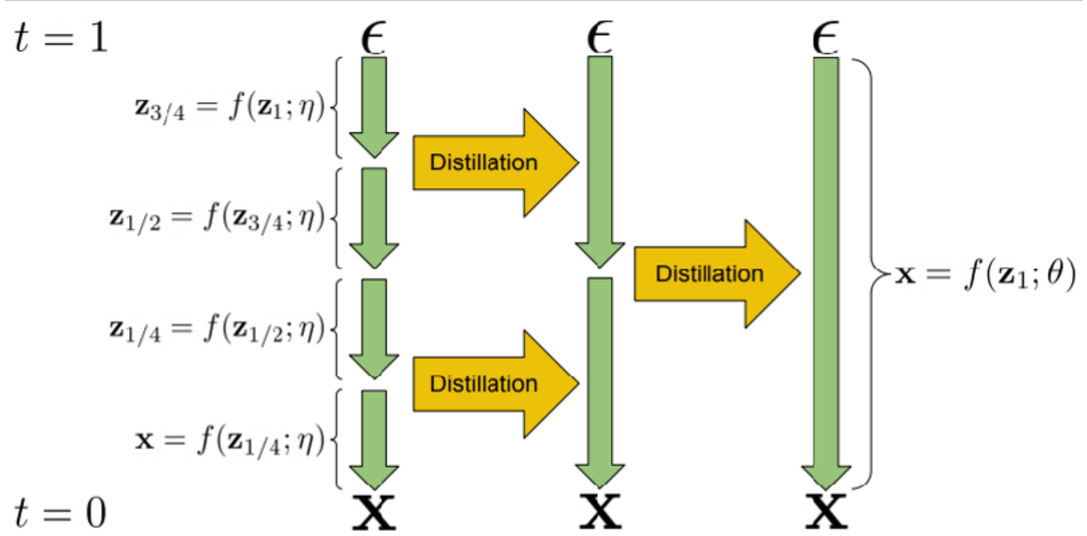


Gyorsított Generálás

Rektifikált folyam (Rectified flow)

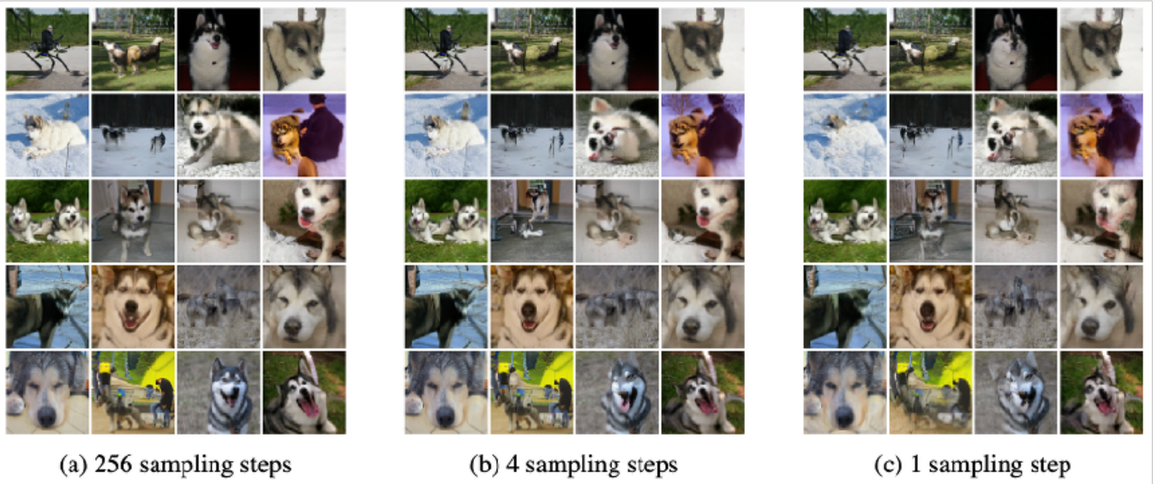
Gyorsított Generálás

Disztilláció — Progresszív tanítás



PROGRESSIVE DISTILLATION FOR FAST SAMPLING OF DIFFUSION MODELS

Tim Salimans & Jonathan Ho
Google Research, Brain team



(a) 256 sampling steps (b) 4 sampling steps (c) 1 sampling step

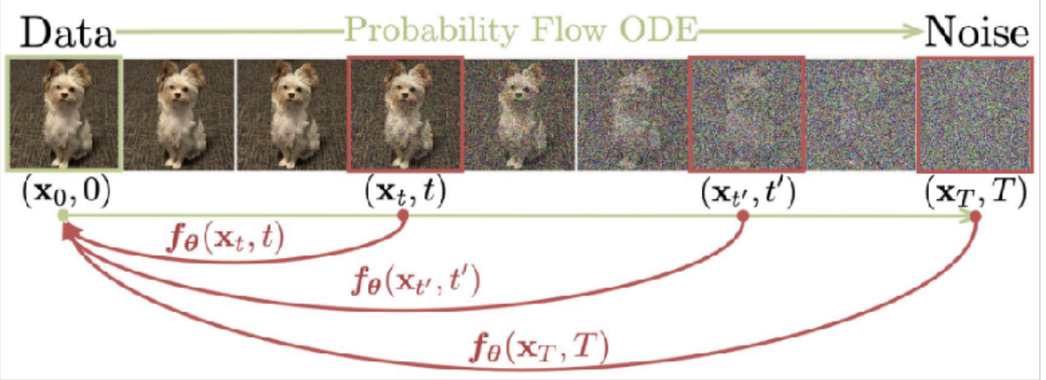
Gyorsított Generálás

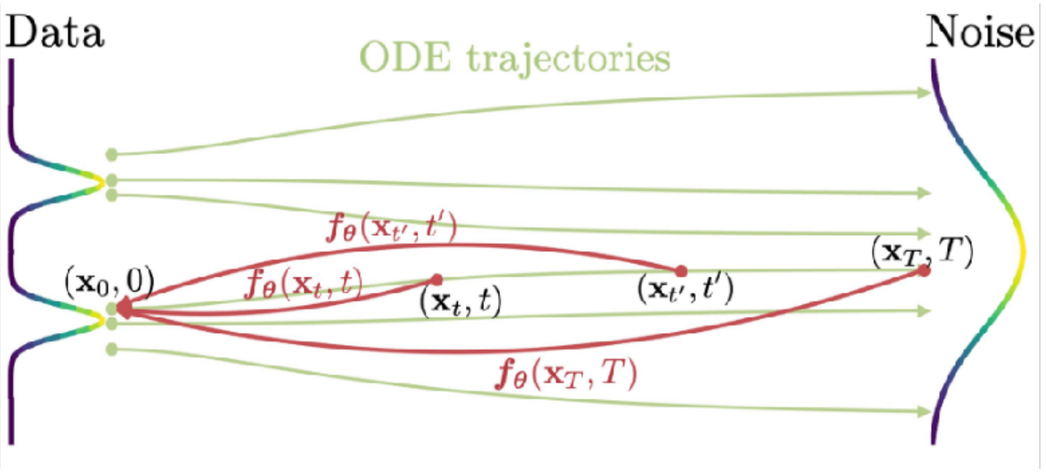
Disztilláció — Konzisztencia modellek

Consistency Models

Yang Song [1] Prafulla Dhariwal [1] Mark Chen [1] Ilya Sutskever [1]

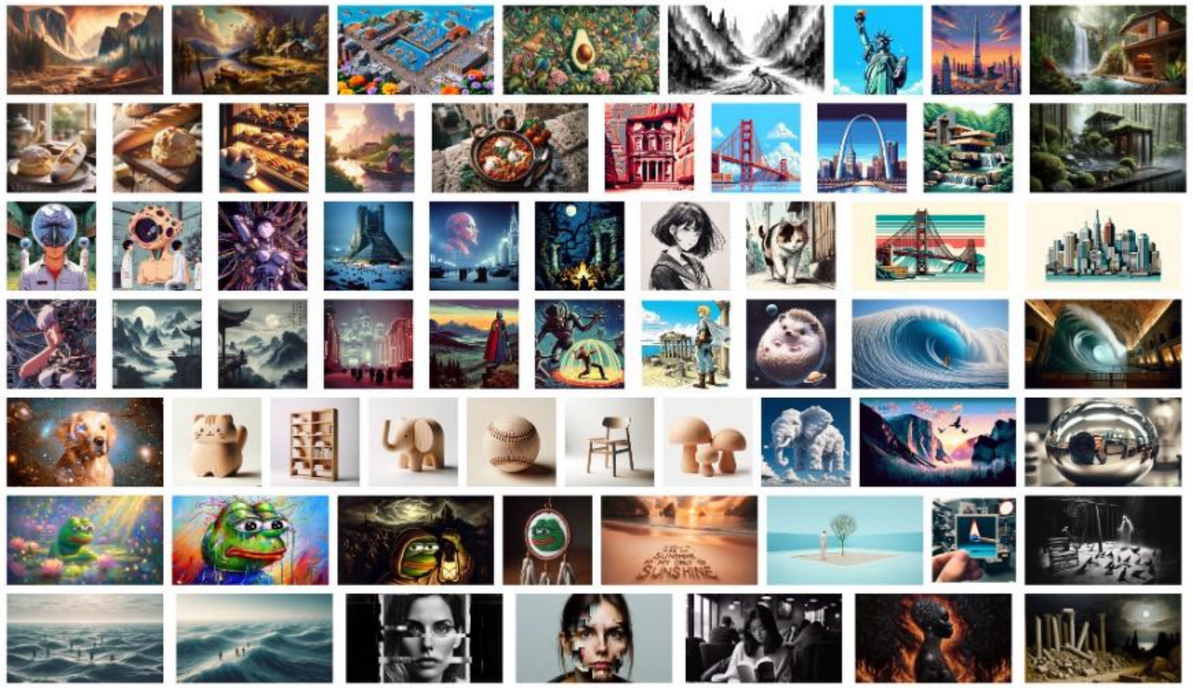
- Tanítsuk a hálót arra, hogy a trajektória végpontjára (a tiszta képre) ugorjon
- Konzisztencia tulajdonság: adott trajektória mentén minden pontból ugyanabba a végpontba ugrunk!
- Lehet disztillálni egy már betanított modellt, vagy akár eleve konzisztenciára tanítani
- Számos variánst inspirált





Gyorsított Generálás

Disztilláció — Konzisztencia modellek



OpenAI

DALL-E 3

SIMPLIFYING, STABILIZING & SCALING CONTINUOUS-TIME CONSISTENCY MODELS

Cheng Lu & Yang Song
OpenAI

Figure 2: Selected 2-step samples from a continuous-time consistency model trained on ImageNet 512×512.

Gyorsított Generálás

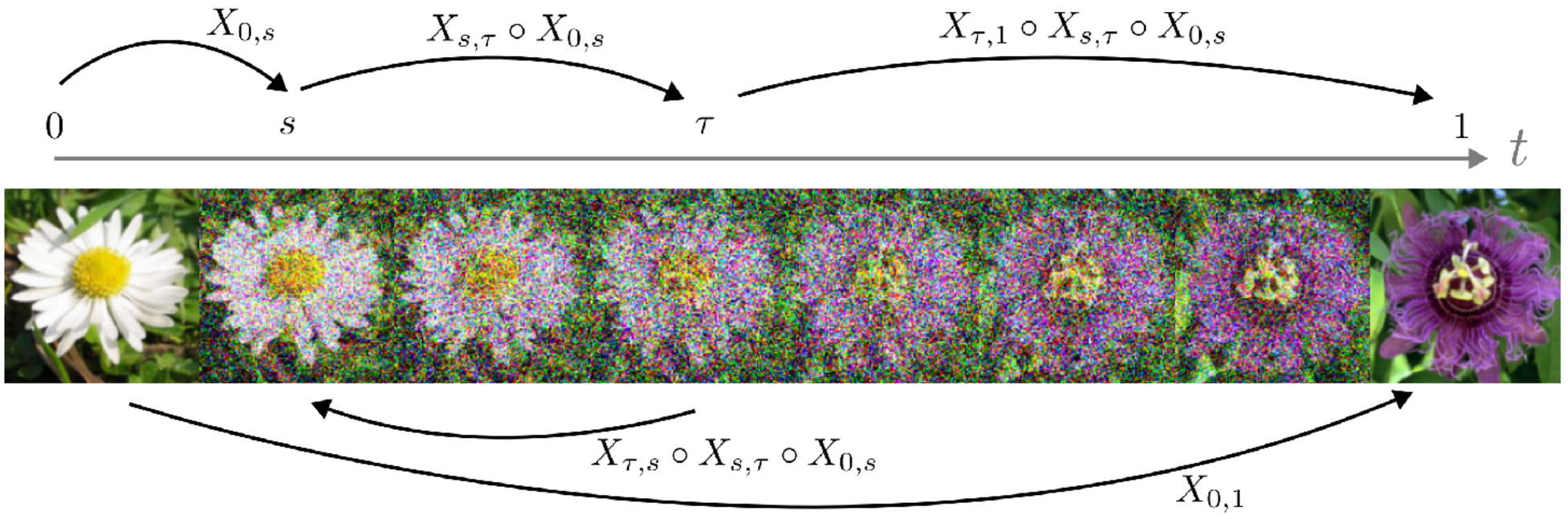
Disztilláció — Folyam leképezés

Flow map matching with stochastic interpolants: A mathematical framework for consistency models

Nicholas M. Boffi* boffi@cims.nyu.edu
Courant Institute of Mathematical Sciences

Michael S. Albergo* albergo@nyu.edu
Courant Institute of Mathematical Sciences

Eric Vanden-Eijnden eve2@cims.nyu.edu
Courant Institute of Mathematical Sciences



Folyam leképezés (Flow map) — $X_{s,\tau}$: a folyam (vektormező) “követése” az s paramétertől a τ paraméterig

Gyorsított Generálás

Disztilláció — Folyam leképezés

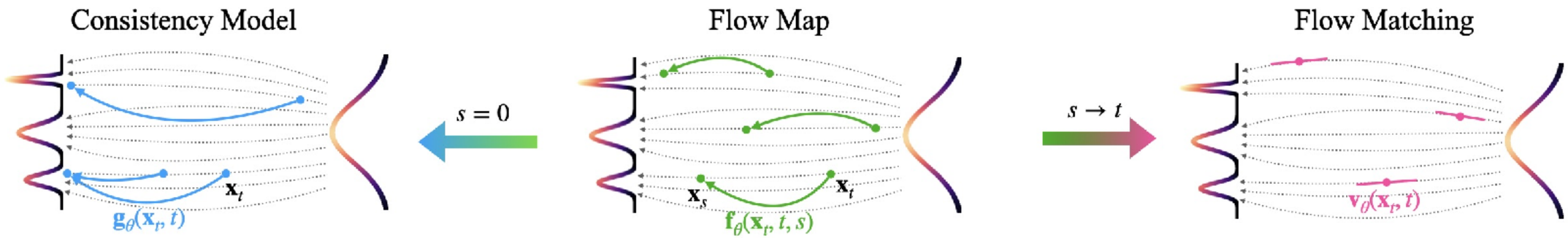
Align Your Flow:
Scaling Continuous-Time Flow Map Distillation

Amirmojtaba Sabour[1,2,3] Sanja Fidler[1,2,3] Karsten Kreis[1]

[1] NVIDIA [2] University of Toronto [3] Vector Institute

Project Page: https://research.nvidia.com/labs/toronto-ai/AlignYourFlow/

Method



Objective

stop gradient: nem kell magasabb deriváltakat számolni!

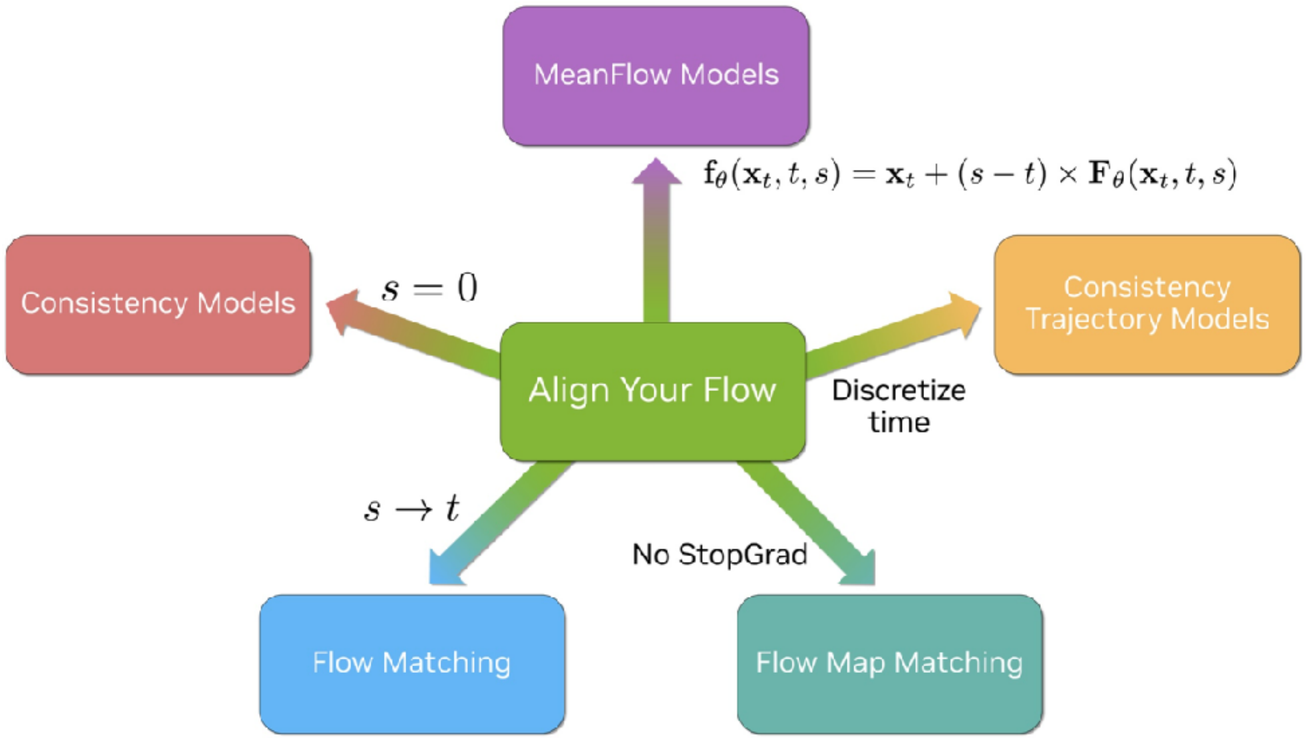
$$\nabla_\theta \mathbb{E}_{\mathbf{x}_t}\left[\mathbf{g}_\theta(\mathbf{x}_t,t)^\top \cdot \frac{\mathrm{d}\mathbf{g}_{\theta^-}(\mathbf{x}_t,t)}{\mathrm{d}t}\right] \xleftarrow{s=0} \nabla_\theta \mathbb{E}_{\mathbf{x}_t}\left[\mathrm{sign}(t-s)\cdot \mathbf{f}_\theta(\mathbf{x}_t,t,s)^\top \cdot \frac{\mathrm{d}\mathbf{f}_{\theta^-}(\mathbf{x}_t,t,s)}{\mathrm{d}t}\right] \xrightarrow{s\to t} \nabla_\theta \mathbb{E}_{\mathbf{x}_t}\left[\left|\left|\mathbf{v}_\theta(\mathbf{x}_t,t) - \frac{\mathrm{d}\mathbf{x}_t}{\mathrm{d}t}\right|\right|_2^2\right]$$

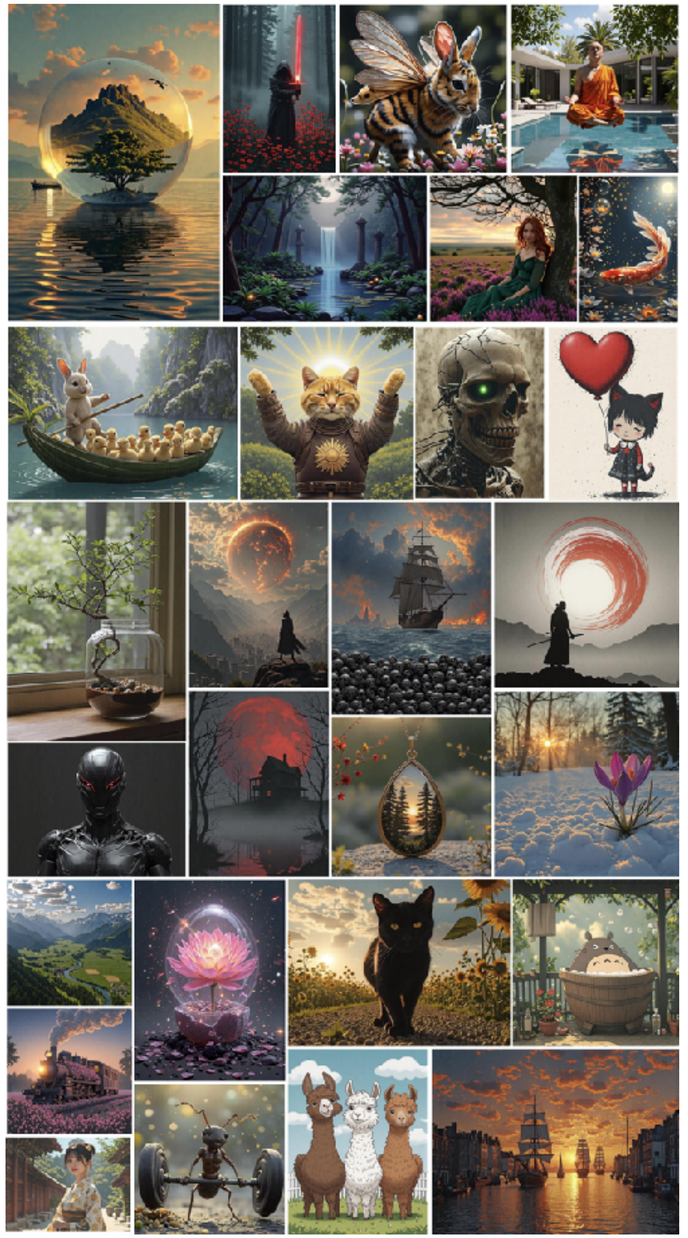
“A folyamleképezés eredménye minden trajektória mentén maradjon állandó”

A konzisztencia és a flow matching loss a folyamleképezés loss degenerált esete!

Gyorsított Generálás

Disztilláció — Folyam leképezés





Képszerkesztés

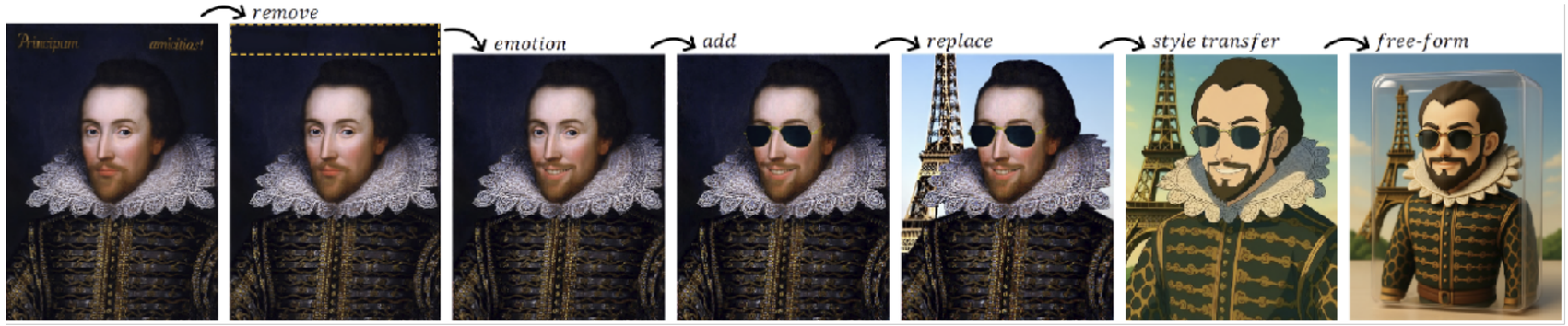
Képből kép — Képszerkesztés

Szöveggel

Más módon

Képszerkesztés

Képből kép — Összetett feladatok



Több lépéses generálás

Komplex képi + szöveges instrukciók

Képszerkesztés

Inverzió

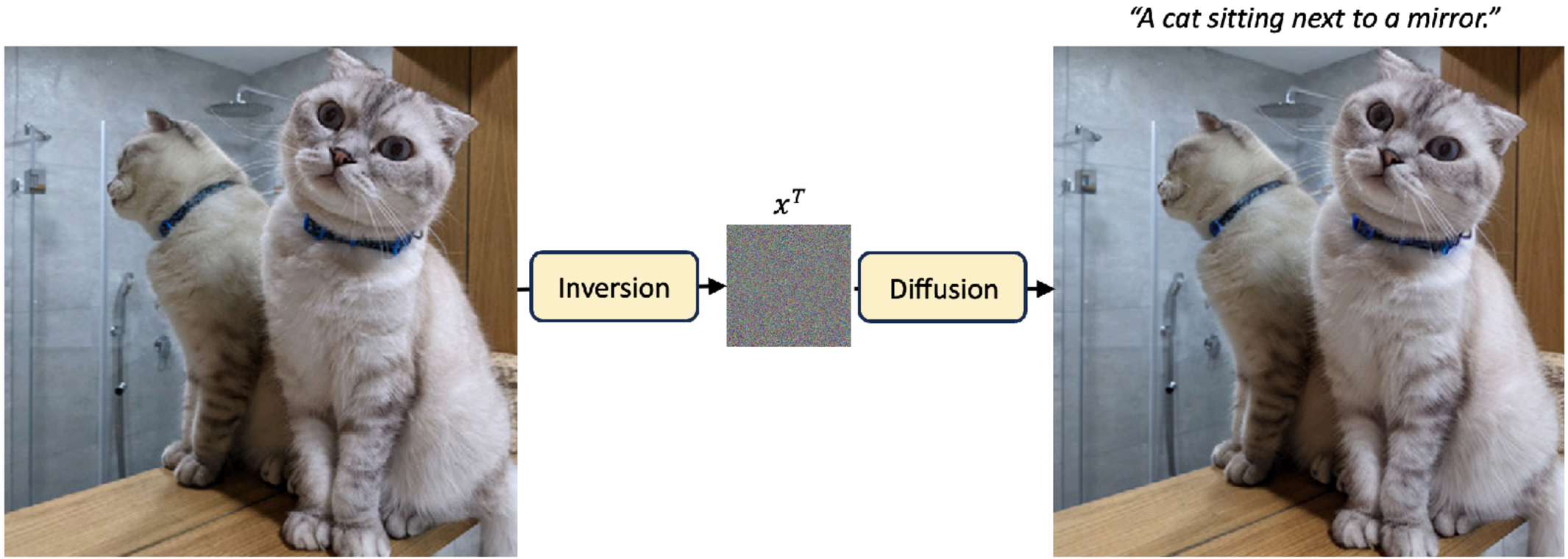
Képszerkesztés

Inverzió

“A cat sitting next to a mirror.”

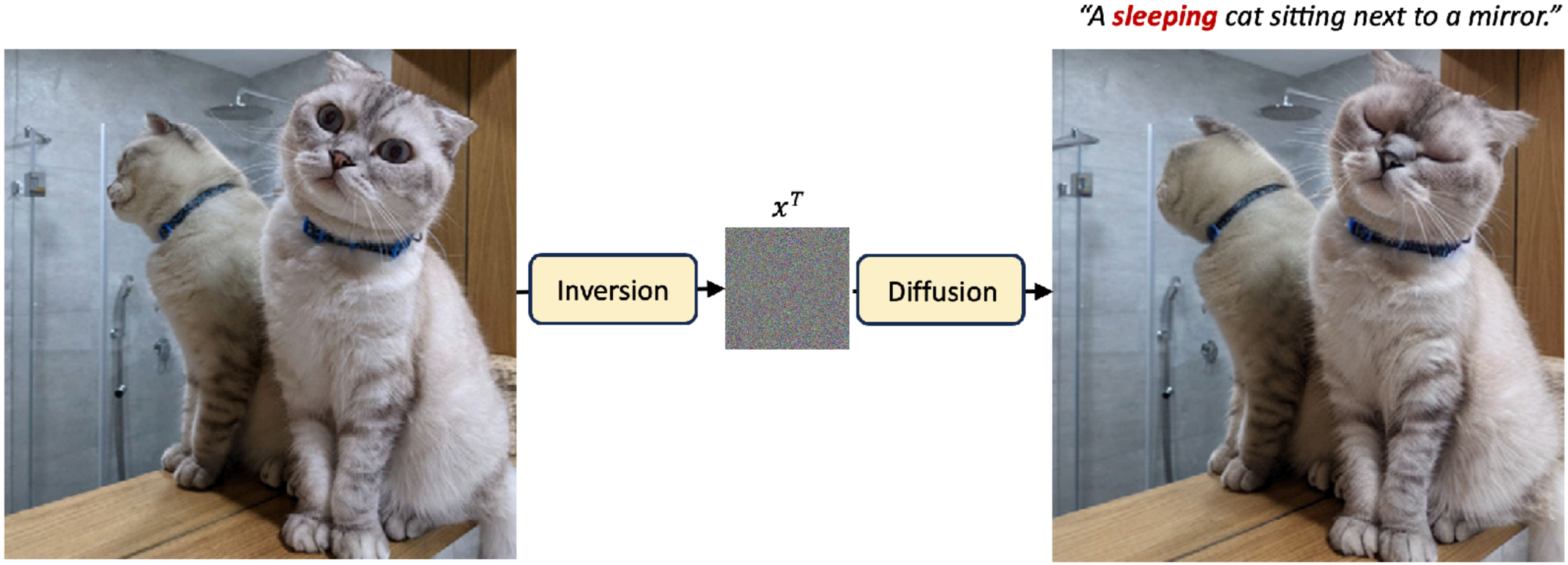
Képszerkesztés

Inverzió



Képszerkesztés

Inverzió



Képszerkesztés

Null-szöveg inverzió

Null-text Inversion for Editing Real Images using Guided Diffusion Models

Ron Mokady*† [1,2], Amir Hertz*† [1,2], Kfir Aberman[1], Yael Pritch[1], and Daniel Cohen-Or† [1,2]
[1]Google Research, [2]The Blavatnik School of Computer Science, Tel Aviv University

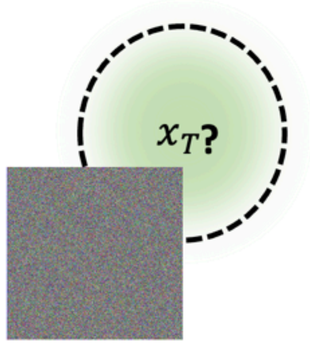
Képszerkesztés

Null-szöveg inverzió

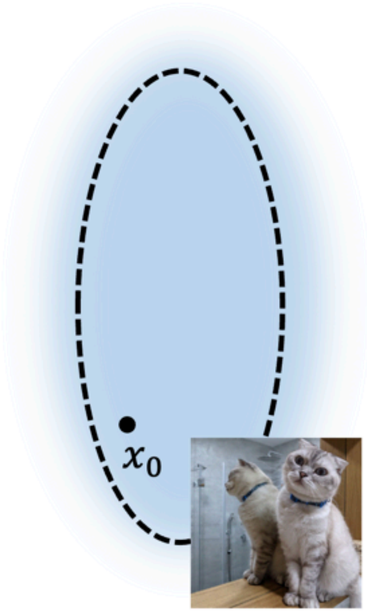
Null-text Inversion for Editing Real Images using Guided Diffusion Models

Ron Mokady*† [1,2], Amir Hertz*† [1,2], Kfir Aberman[1], Yael Pritch[1], and Daniel Cohen-Or† [1,2]
[1]Google Research, [2]The Blavatnik School of Computer Science, Tel Aviv University

"A cat sitting next to a mirror."



noise distribution
$p(x_T)$



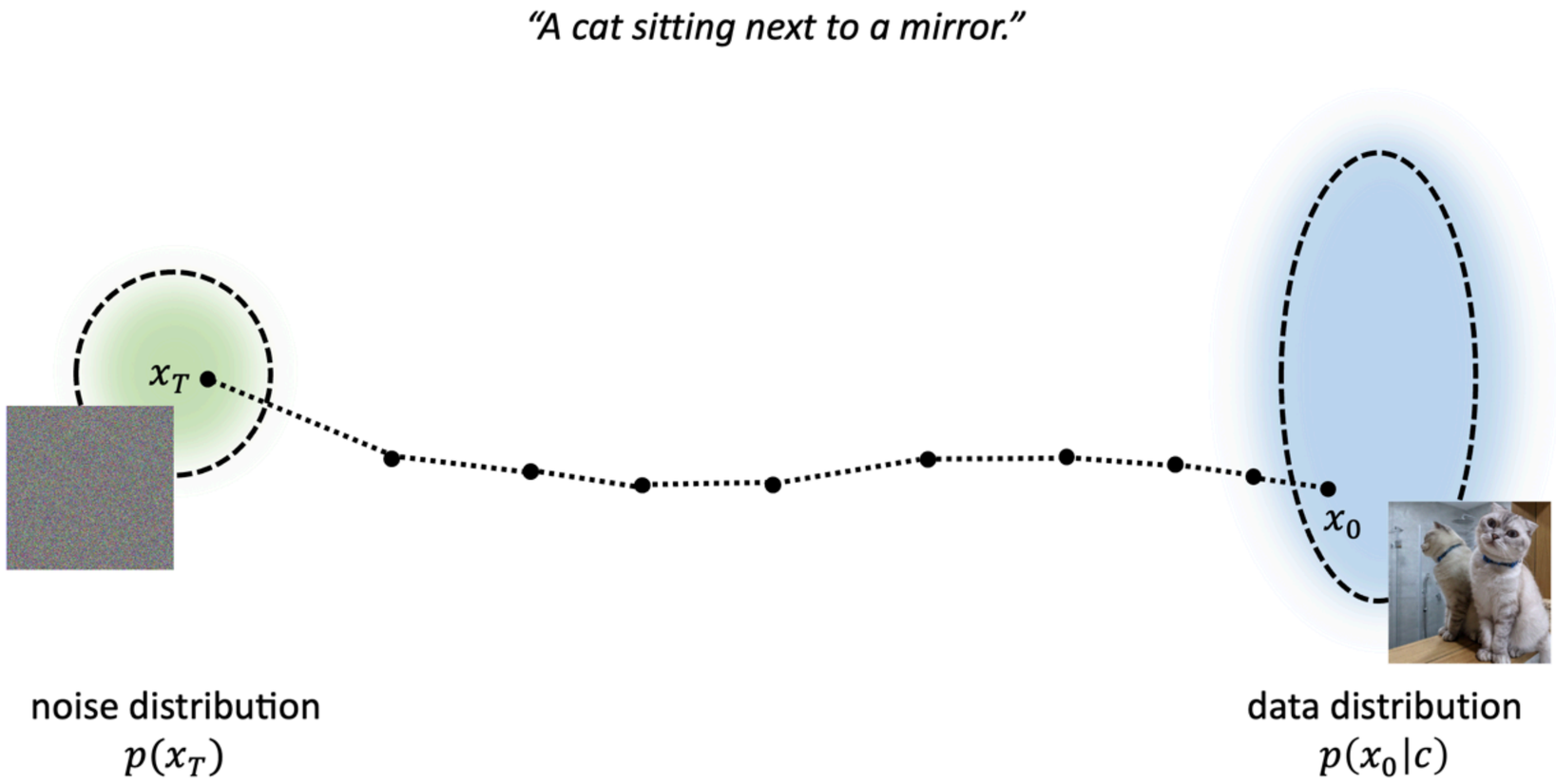
data distribution
$p(x_0|c)$

Képszerkesztés

Null-szöveg inverzió

Null-text Inversion for Editing Real Images using Guided Diffusion Models

Ron Mokady*† [1,2], Amir Hertz*† [1,2], Kfir Aberman[1], Yael Pritch[1], and Daniel Cohen-Or† [1,2]
[1]Google Research, [2]The Blavatnik School of Computer Science, Tel Aviv University

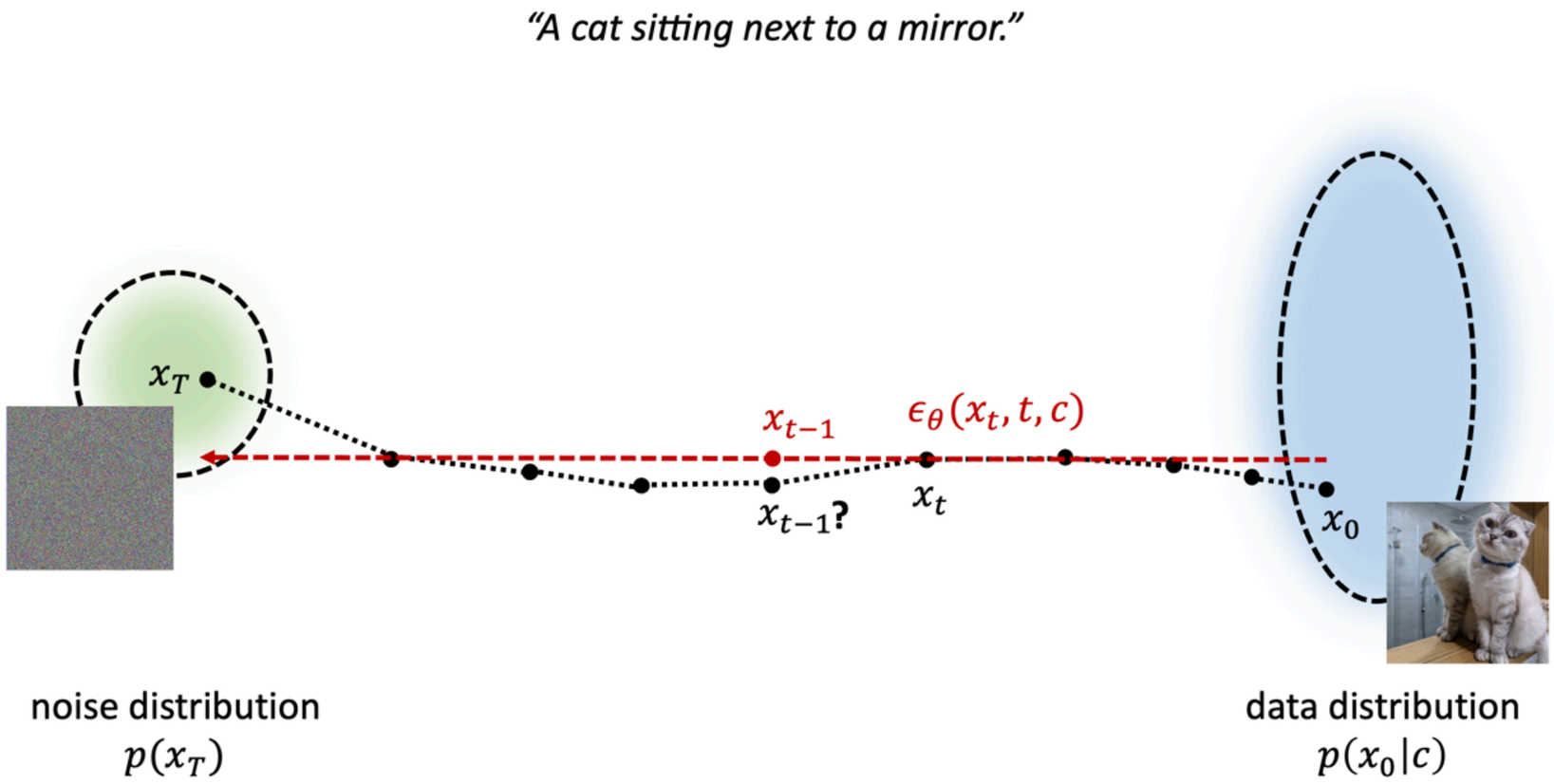


Képszerkesztés

Null-szöveg inverzió

Null-text Inversion for Editing Real Images using Guided Diffusion Models

Ron Mokady*† [1,2], Amir Hertz*† [1,2], Kfir Aberman[1], Yael Pritch[1], and Daniel Cohen-Or† [1,2]
[1]Google Research, [2]The Blavatnik School of Computer Science, Tel Aviv University

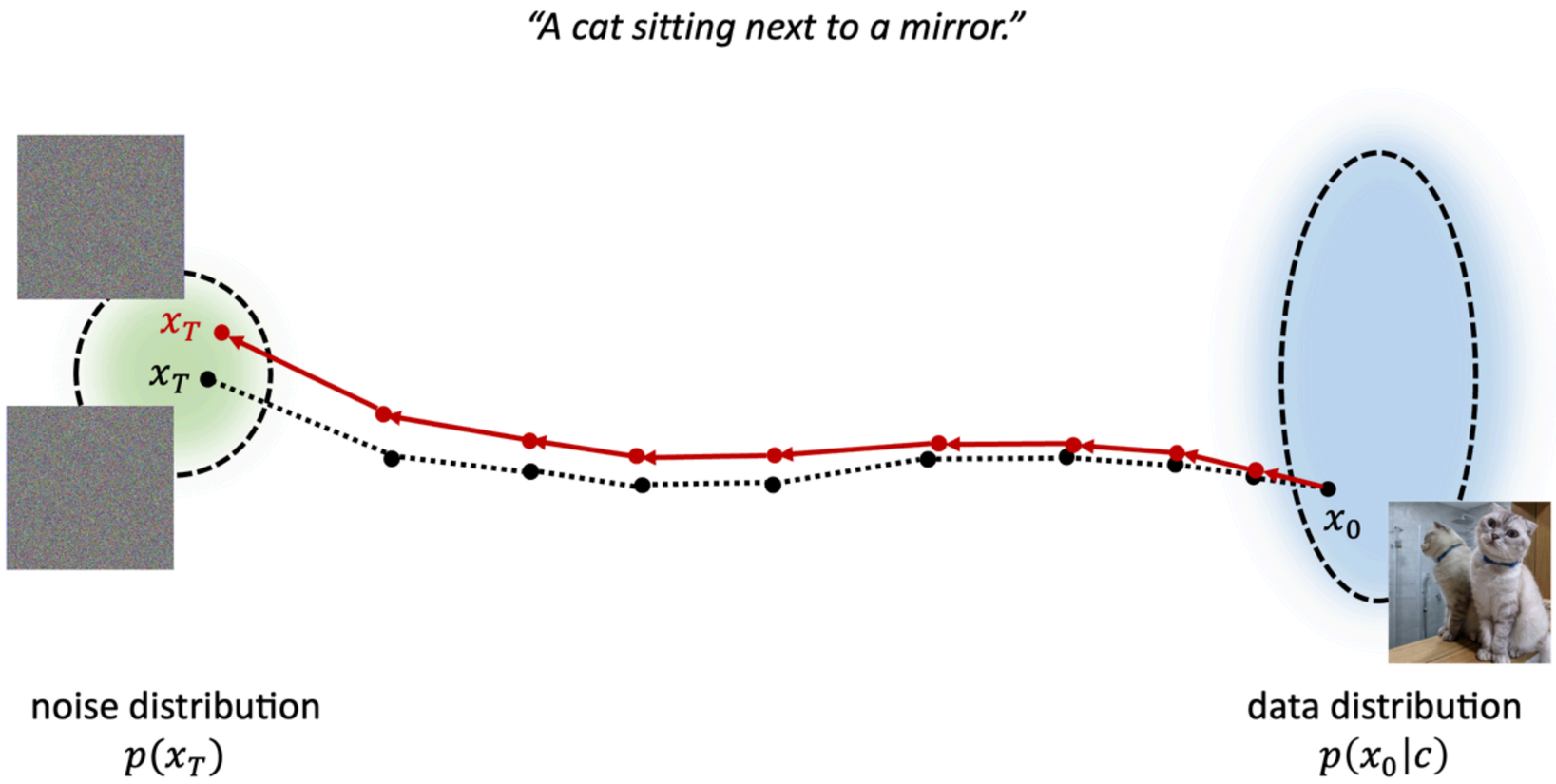


Képszerkesztés

Null-szöveg inverzió

Null-text Inversion for Editing Real Images using Guided Diffusion Models

Ron Mokady*† [1,2], Amir Hertz*† [1,2], Kfir Aberman[1], Yael Pritch[1], and Daniel Cohen-Or† [1,2]
[1]Google Research, [2]The Blavatnik School of Computer Science, Tel Aviv University

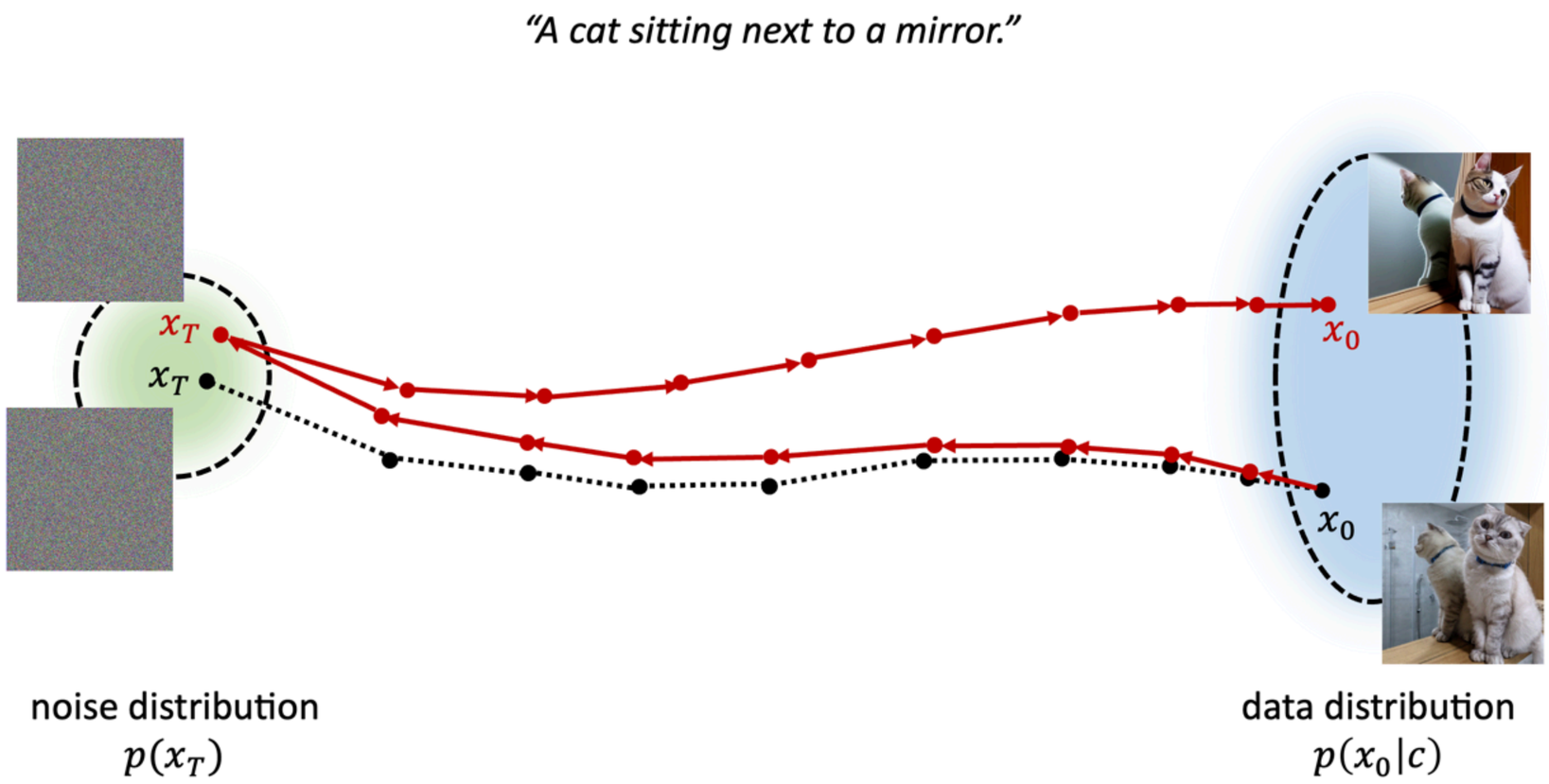


Képszerkesztés

Null-szöveg inverzió

Null-text Inversion for Editing Real Images using Guided Diffusion Models

Ron Mokady*† [1,2], Amir Hertz*† [1,2], Kfir Aberman[1], Yael Pritch[1], and Daniel Cohen-Or† [1,2]
[1]Google Research, [2]The Blavatnik School of Computer Science, Tel Aviv University

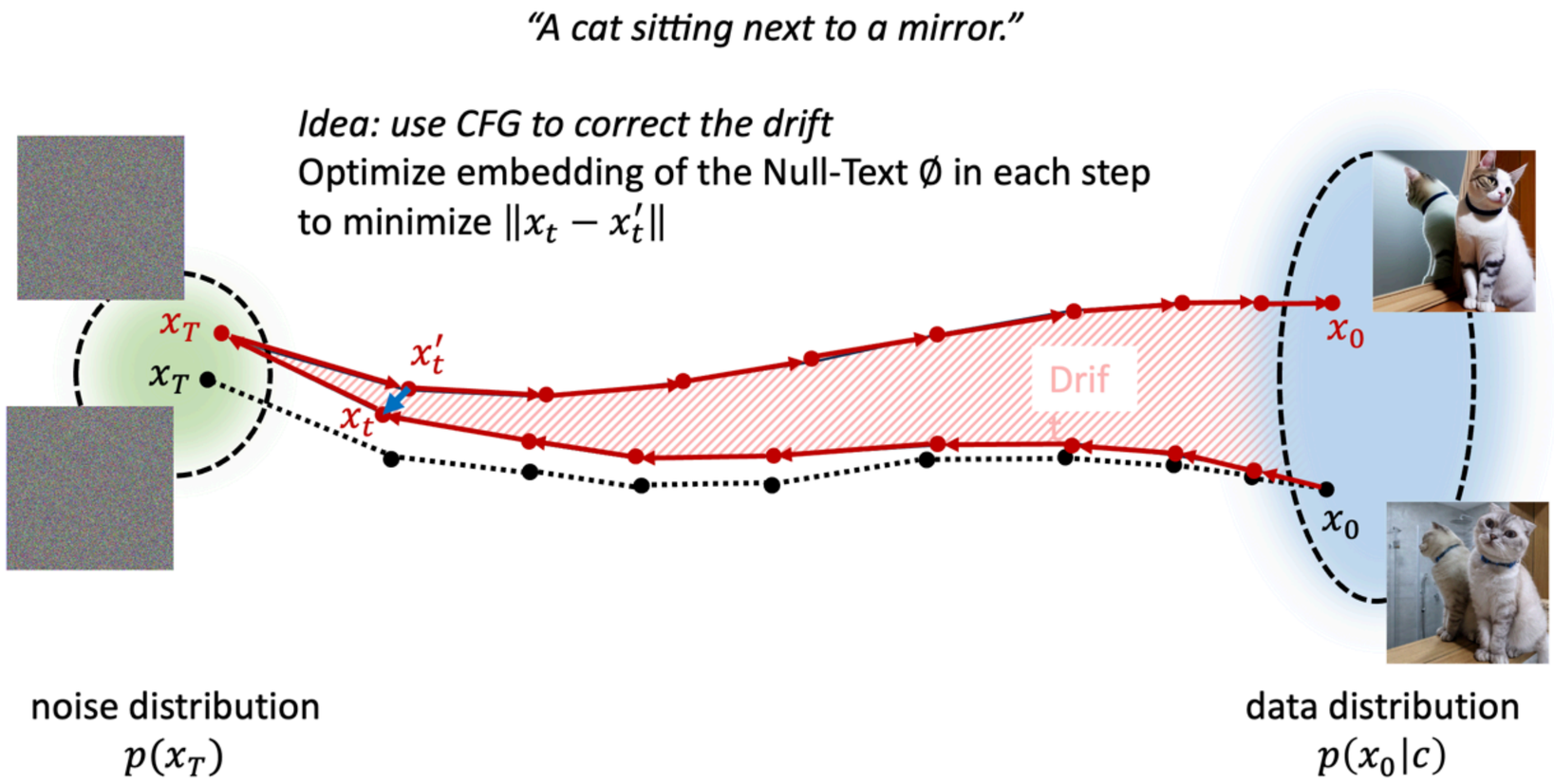


Képszerkesztés

Null-szöveg inverzió

Null-text Inversion for Editing Real Images using Guided Diffusion Models

Ron Mokady*† [1,2], Amir Hertz*† [1,2], Kfir Aberman[1], Yael Pritch[1], and Daniel Cohen-Or† [1,2]
[1]Google Research, [2]The Blavatnik School of Computer Science, Tel Aviv University

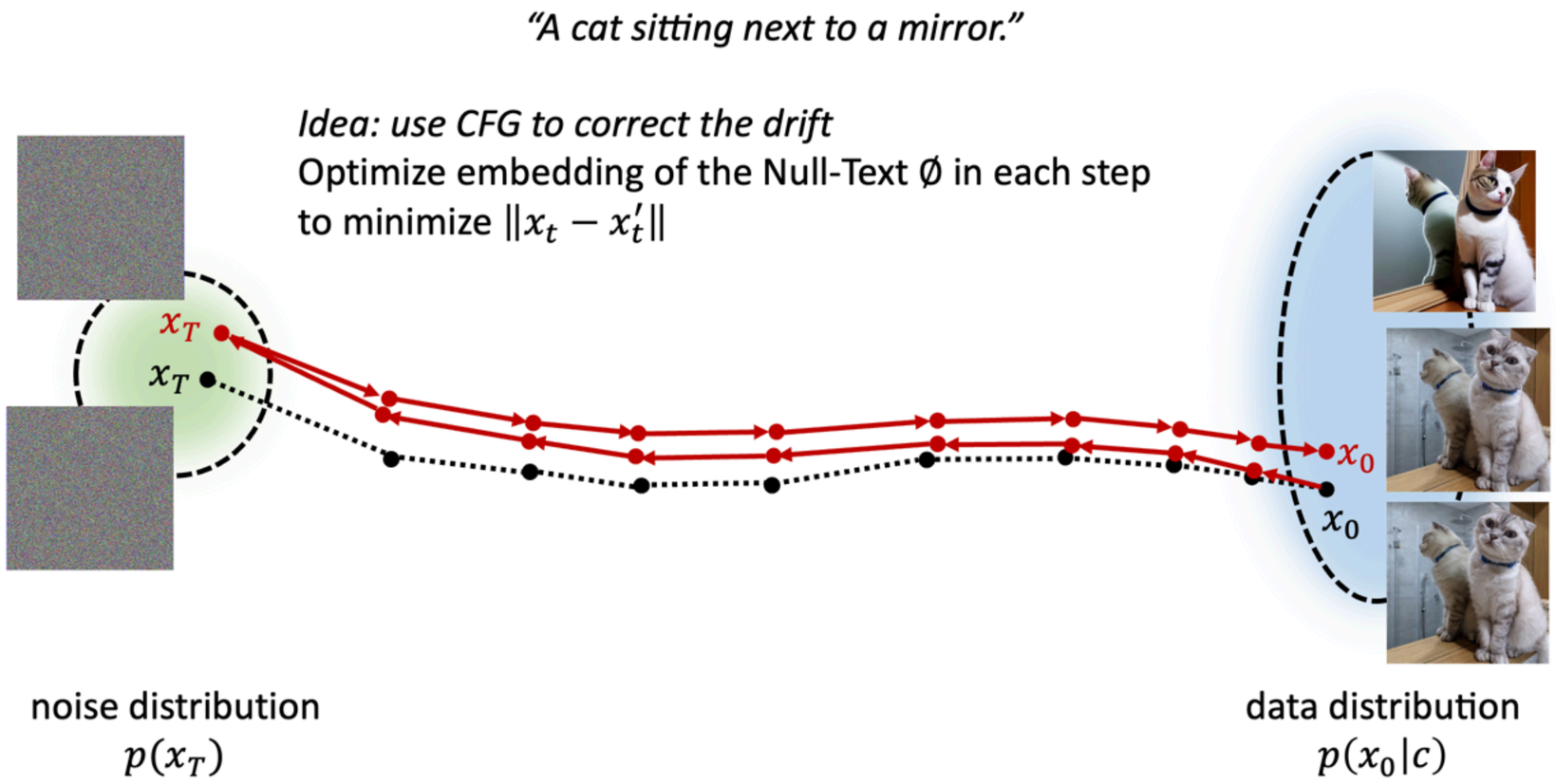


Képszerkesztés

Null-szöveg inverzió

Null-text Inversion for Editing Real Images using Guided Diffusion Models

Ron Mokady*† [1,2], Amir Hertz*† [1,2], Kfir Aberman[1], Yael Pritch[1], and Daniel Cohen-Or† [1,2]
[1]Google Research, [2]The Blavatnik School of Computer Science, Tel Aviv University



Képszerkesztés

Zajosítás — SDEdit

SDEDIT: GUIDED IMAGE SYNTHESIS AND EDITING WITH STOCHASTIC DIFFERENTIAL EQUATIONS

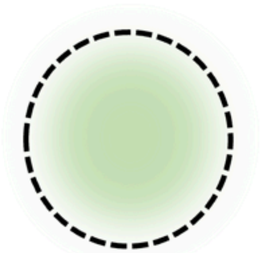
Chenlin Meng[1] **Yutong He**[1] **Yang Song**[1] **Jiaming Song**[1]
Jiajun Wu[1] **Jun-Yan Zhu**[2] **Stefano Ermon**[1]
[1]Stanford University [2]Carnegie Mellon University

Képszerkesztés

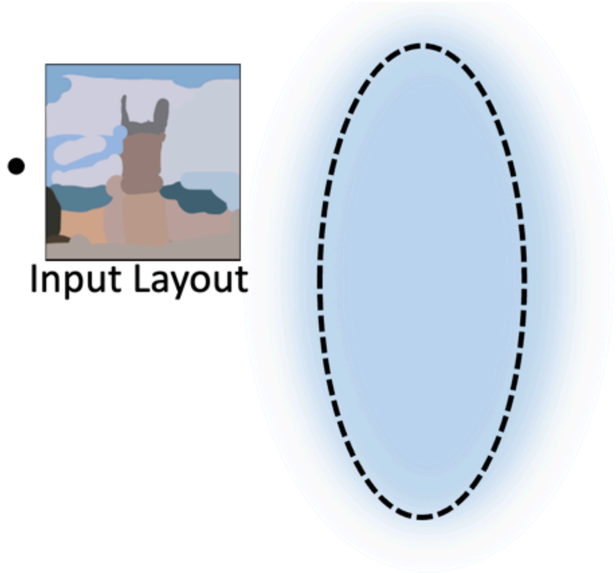
Zajosítás — SDEdit

SDEDIT: GUIDED IMAGE SYNTHESIS AND EDITING WITH STOCHASTIC DIFFERENTIAL EQUATIONS

Chenlin Meng[1] **Yutong He**[1] **Yang Song**[1] **Jiaming Song**[1]
Jiajun Wu[1] **Jun-Yan Zhu**[2] **Stefano Ermon**[1]
[1]Stanford University [2]Carnegie Mellon University



noise distribution
$p(x_T)$



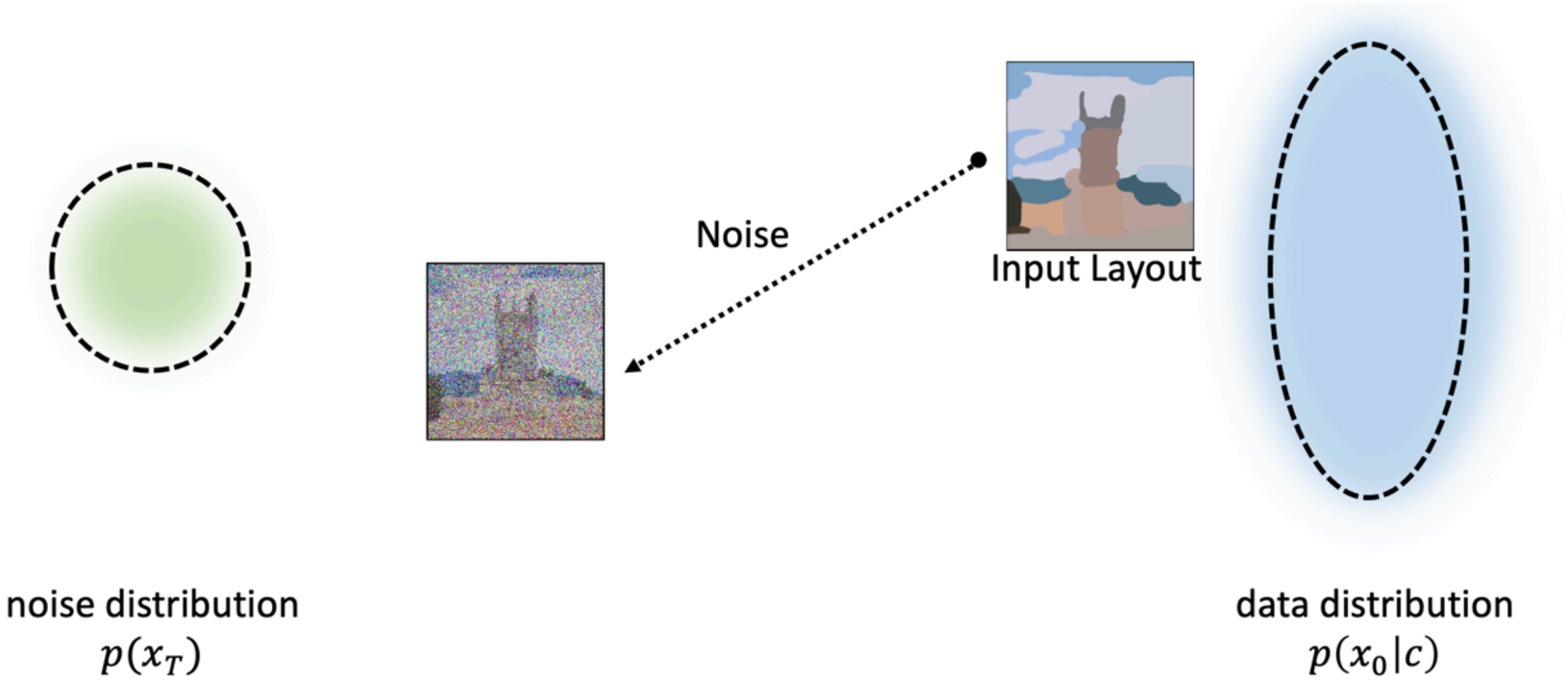
data distribution
$p(x_0|c)$

Képszerkesztés

Zajosítás — SDEdit

SDEDIT: GUIDED IMAGE SYNTHESIS AND EDITING WITH STOCHASTIC DIFFERENTIAL EQUATIONS

Chenlin Meng[1] **Yutong He**[1] **Yang Song**[1] **Jiaming Song**[1]
Jiajun Wu[1] **Jun-Yan Zhu**[2] **Stefano Ermon**[1]
[1]Stanford University [2]Carnegie Mellon University

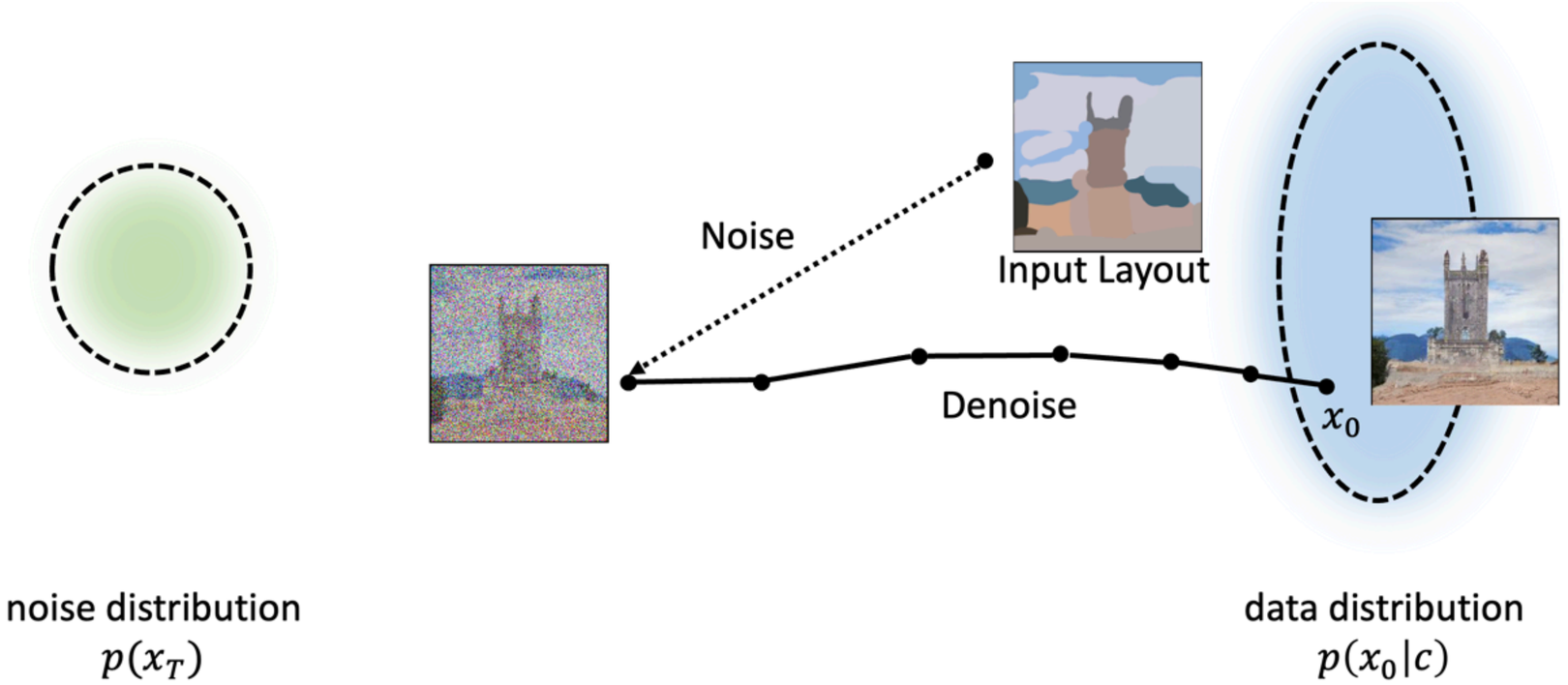


Képszerkesztés

Zajosítás — SDEdit

SDEDIT: GUIDED IMAGE SYNTHESIS AND EDITING WITH STOCHASTIC DIFFERENTIAL EQUATIONS

Chenlin Meng[1] **Yutong He**[1] **Yang Song**[1] **Jiaming Song**[1]
Jiajun Wu[1] **Jun-Yan Zhu**[2] **Stefano Ermon**[1]
[1]Stanford University [2]Carnegie Mellon University

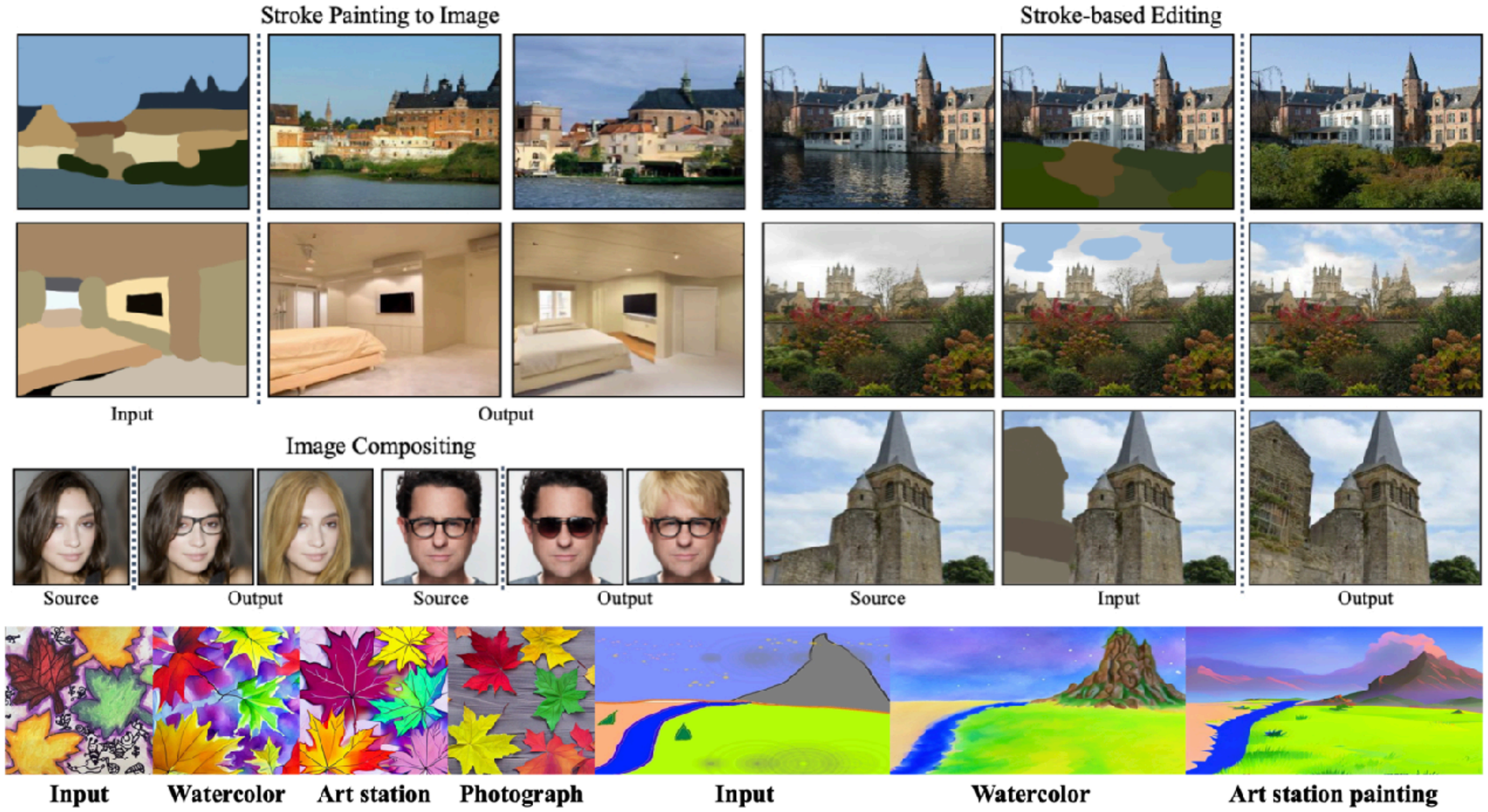


Képszerkesztés

Zajosítás — SDEdit

SDEDIT: GUIDED IMAGE SYNTHESIS AND EDITING WITH STOCHASTIC DIFFERENTIAL EQUATIONS

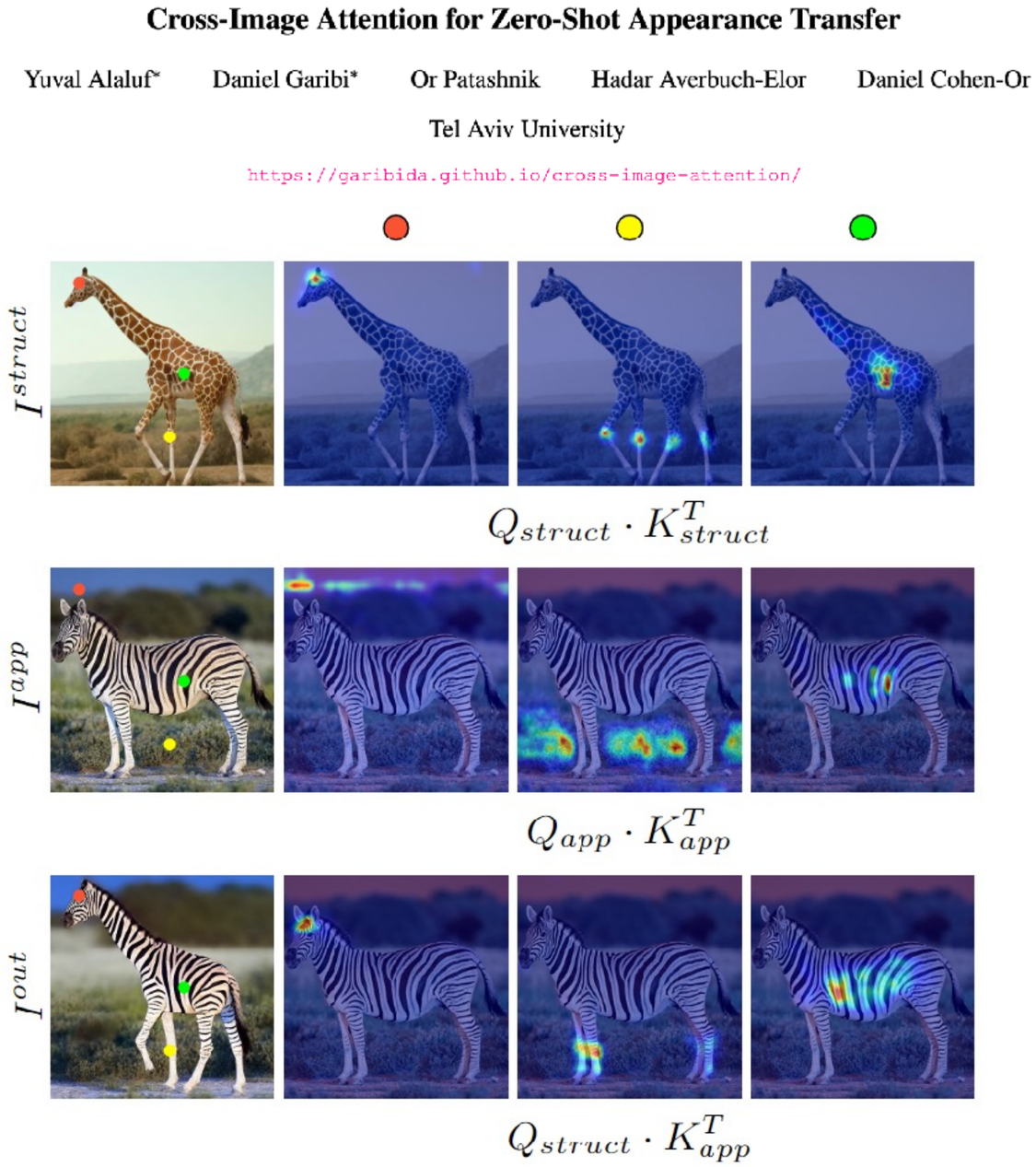
Chenlin Meng[1] **Yutong He**[1] **Yang Song**[1] **Jiaming Song**[1]
Jiajun Wu[1] **Jun-Yan Zhu**[2] **Stefano Ermon**[1]
[1]Stanford University [2]Carnegie Mellon University



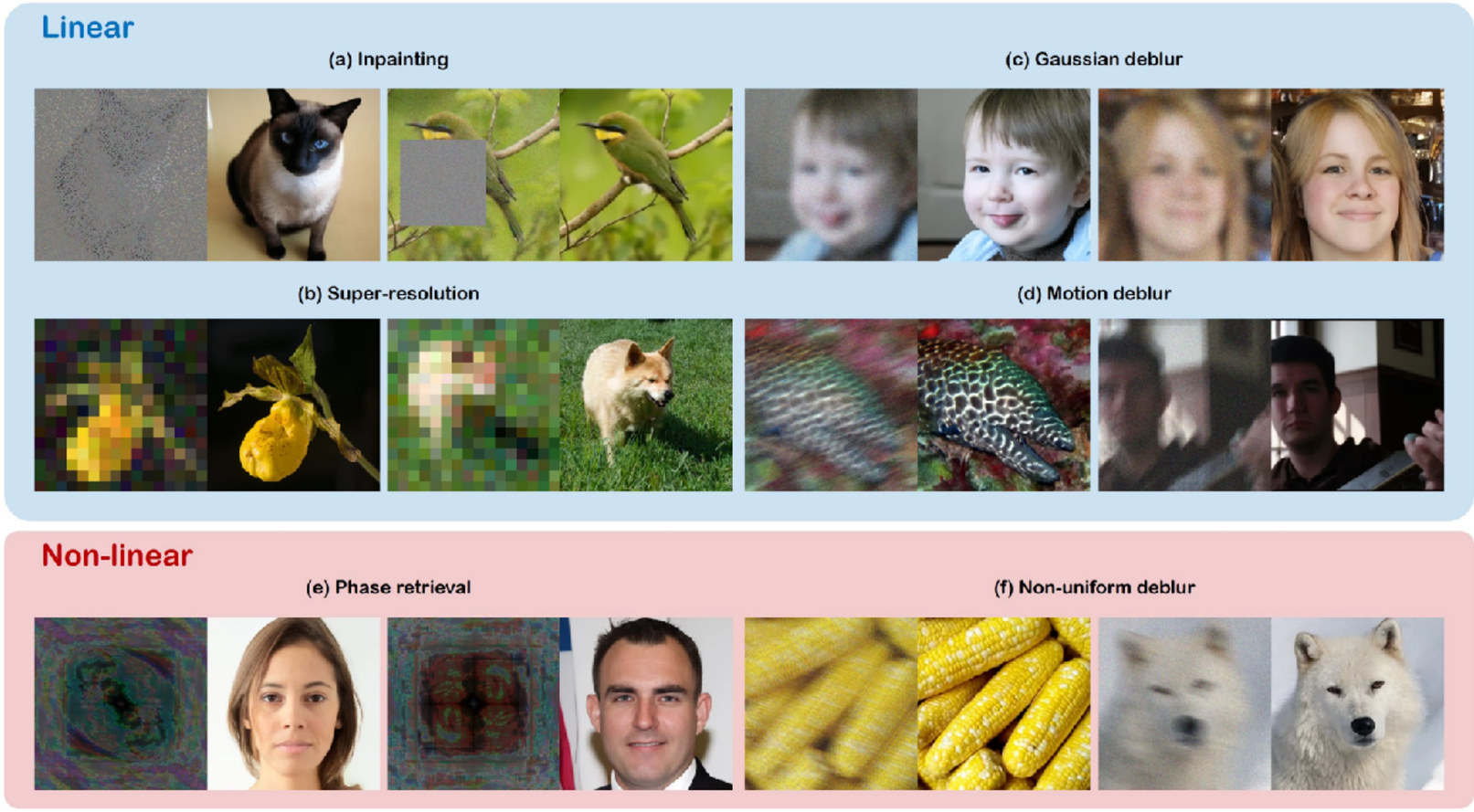
Képszerkesztés

Cross-Image Attention

- Cross-attention által az egyik kép struktúrája kombinálható egy másik kinézetével !

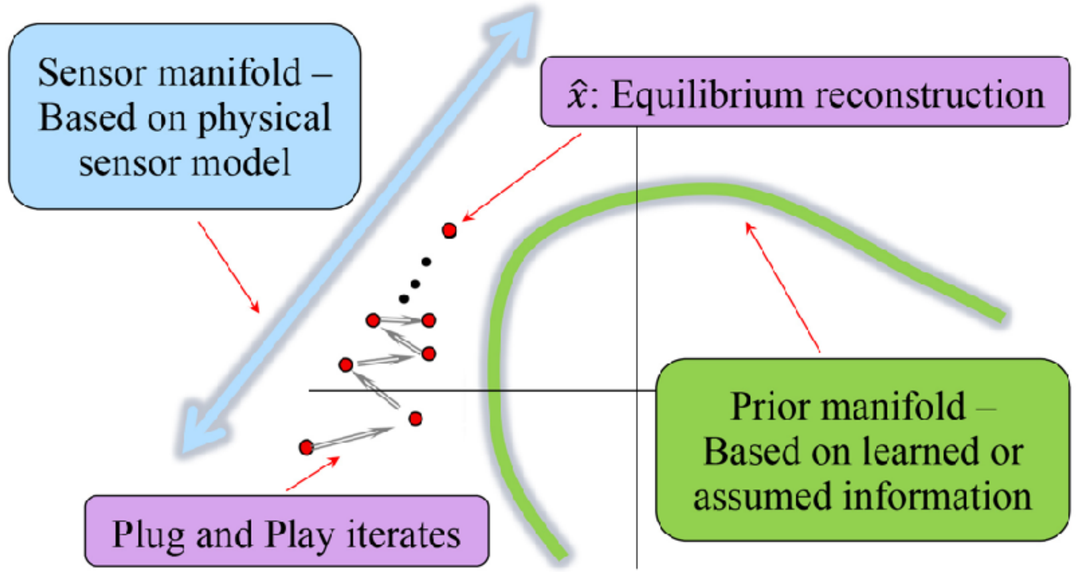


Inverz problémák*



Inverz problémák*

Diffúziós priorok



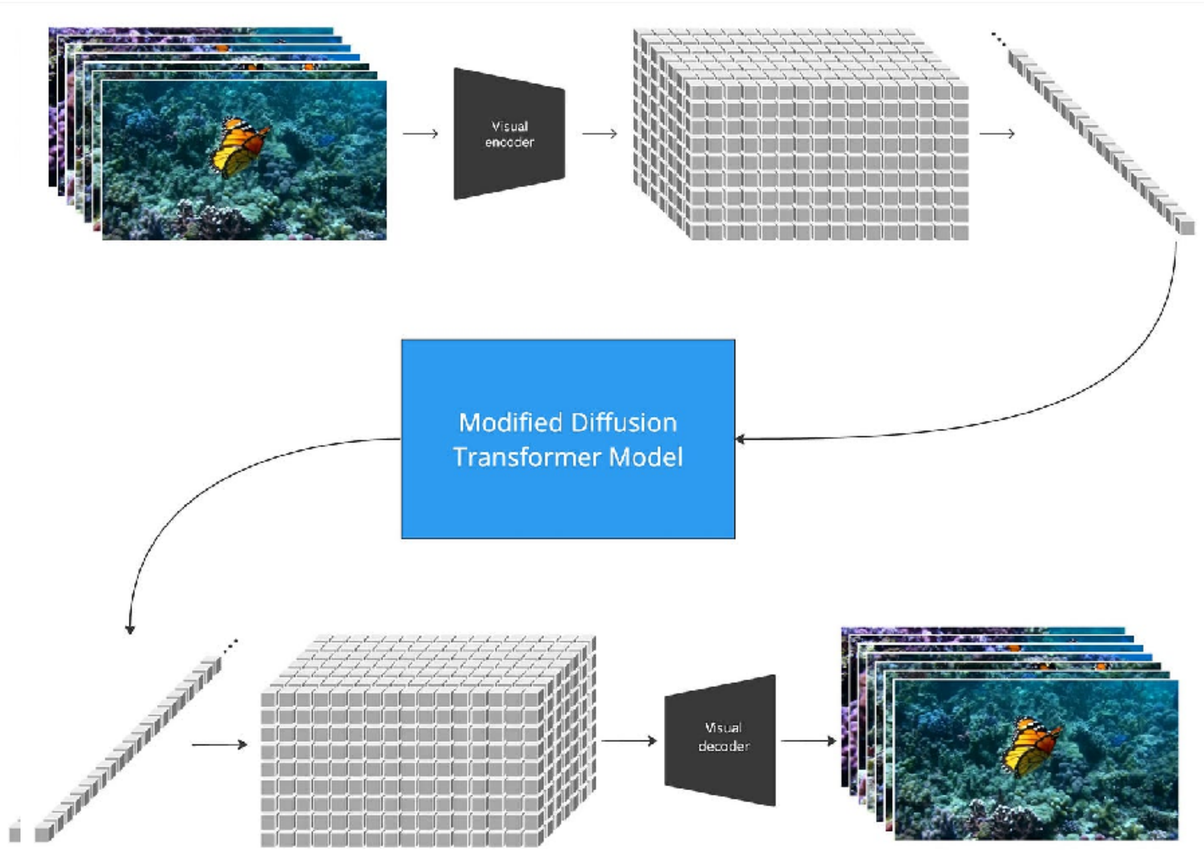
Videó Generálás

- Állókép helyett mozgóképet is generálhatunk diffúzióval / folyamillesztéssel
- 1080p felbontás @ 60 FPS: 7.5 milliárd pixel percenként!
- Nem elég több képet generálni, időbeli konzisztenciát biztosítani kell!
- A tanulásra használható adatok potenciális mennyisége óriási:
  - Csak a YouTube-ra naponta 720 000 órányi (80 évnyi) videót töltenek fel!
- Videó generátorok mint “világmodellek”: a “valódi” MI potenciális alapja?

Videó Generálás

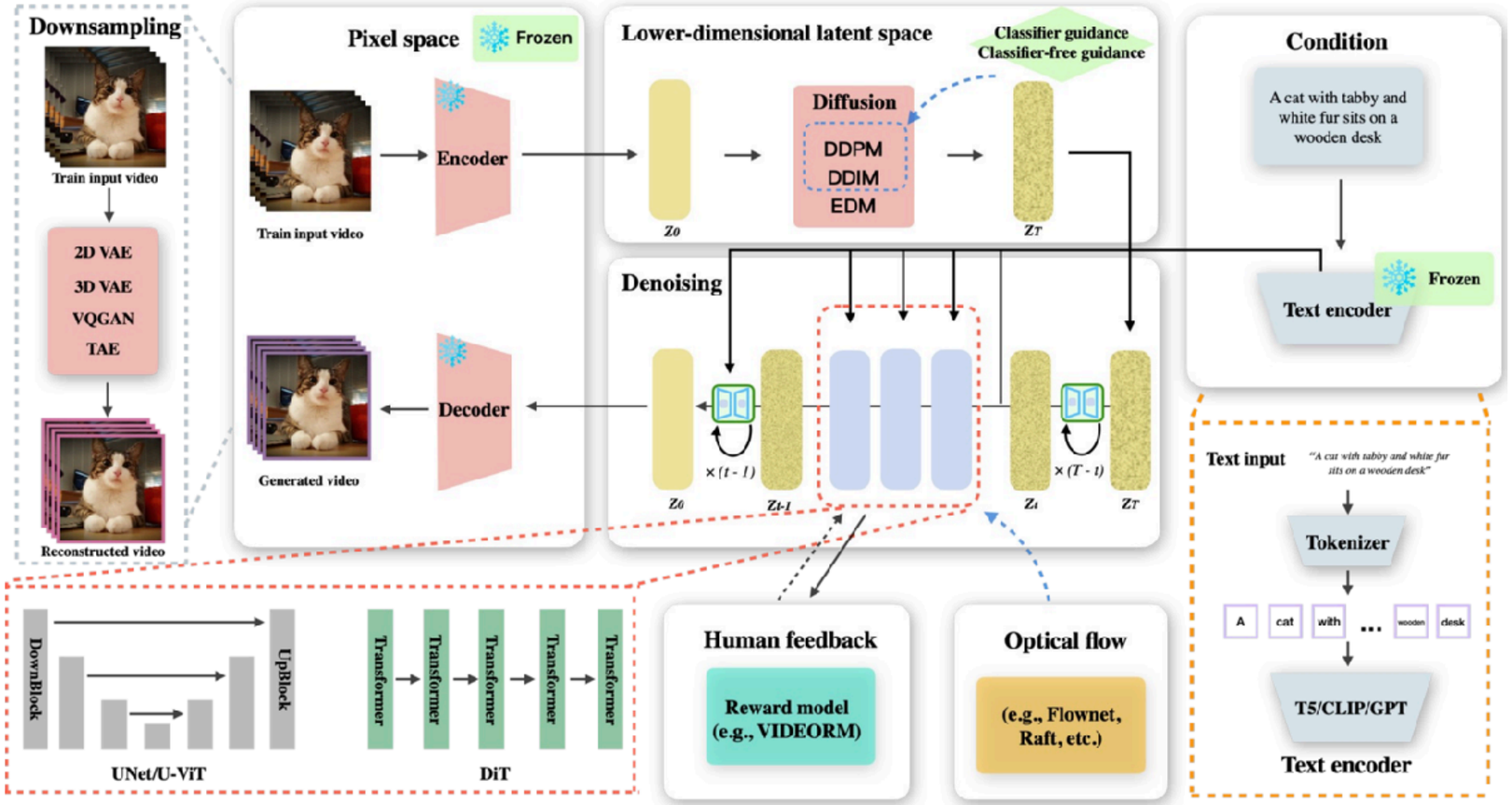
Látens generálás

- Videók diffúziós generálását szükségszerű látens térben végezni
- Általában külön tanított (variációs) autóenkódert használunk
  - Számos variáció, a legtöbb innováció ehhez kapcsolódik
- A generatív modell mostanában tipikusan “egyszerű” folyamillesztés, kisebb módosításokkal



Videó Generálás

Látens generálás

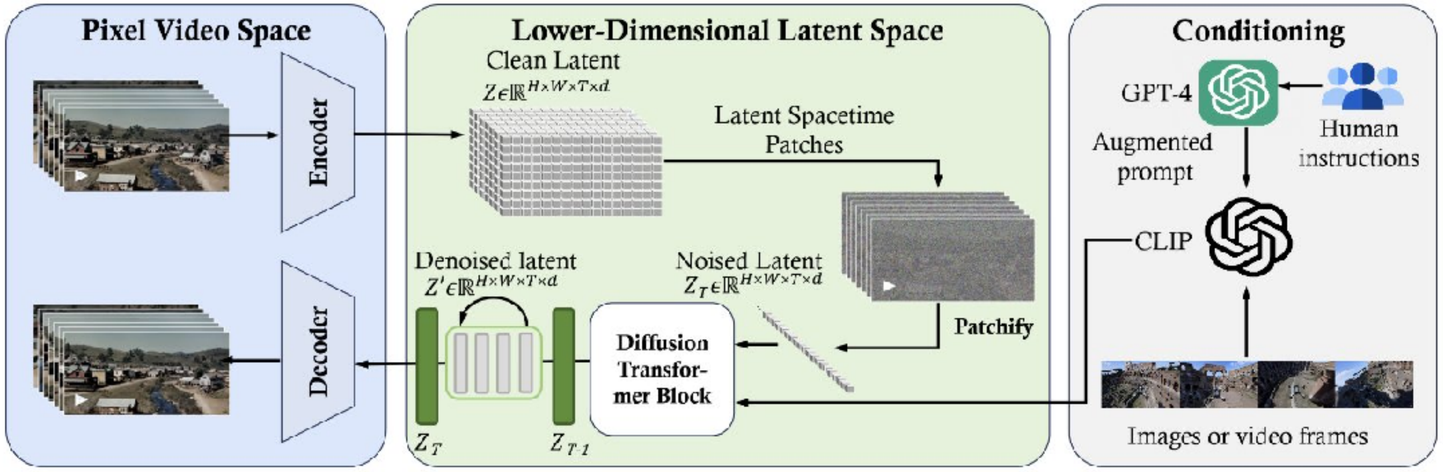


Videó Generálás

(Open)Sora

Sora

https://github.com/hpcaitech/Open-sora



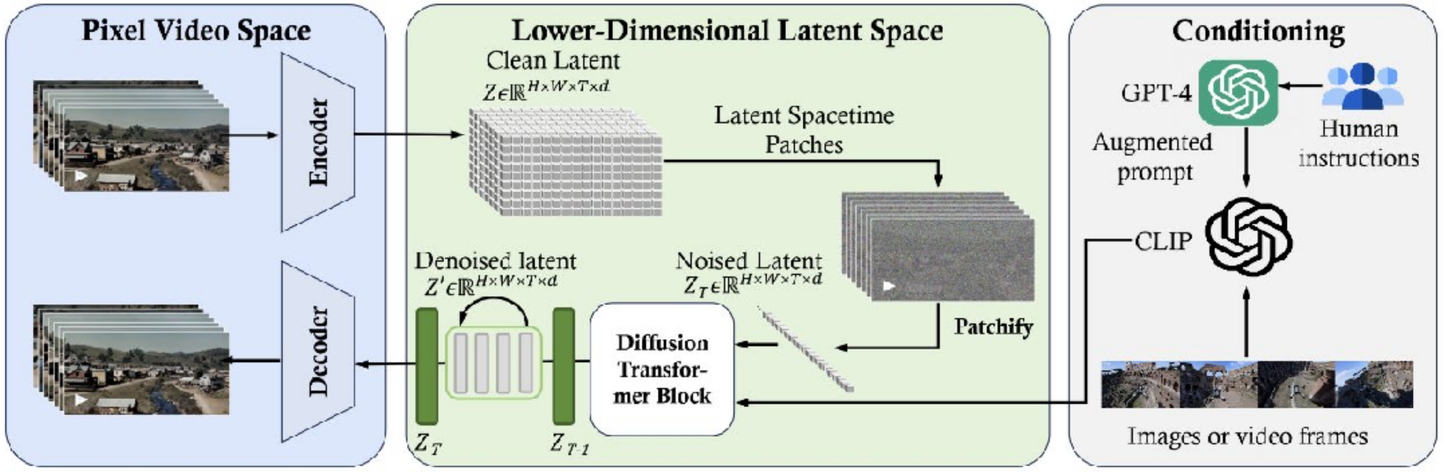
Látens “tér-idő” patch-ek

Videó Generálás

(Open)Sora

Sora

https://github.com/hpcaitech/Open-sora



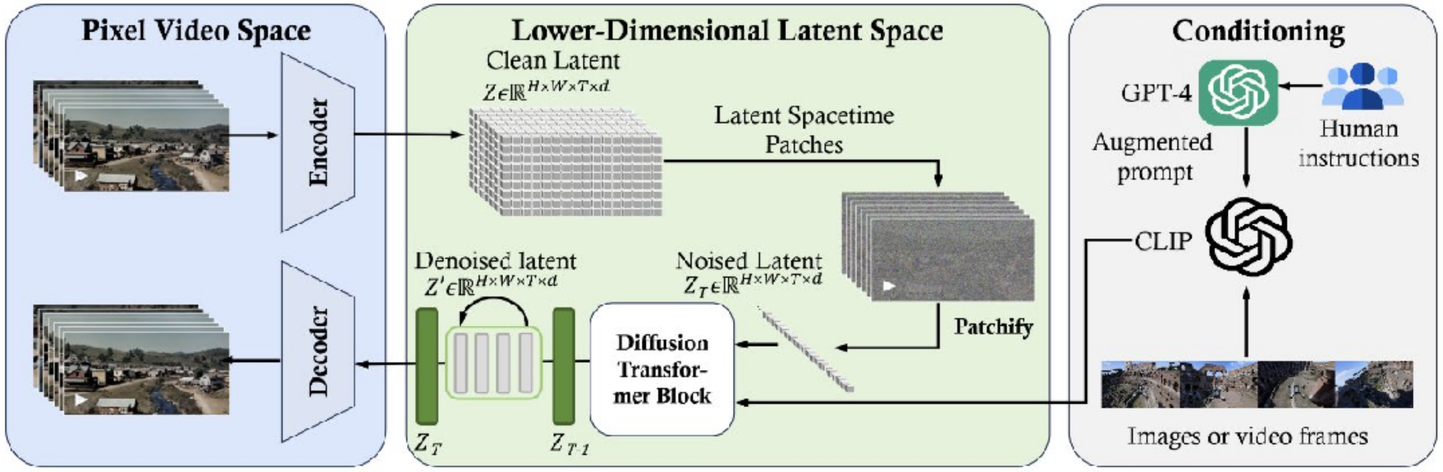
Látens “tér-idő” patch-ek

Videó Generálás

(Open)Sora

Sora

https://github.com/hpcaitech/Open-sora



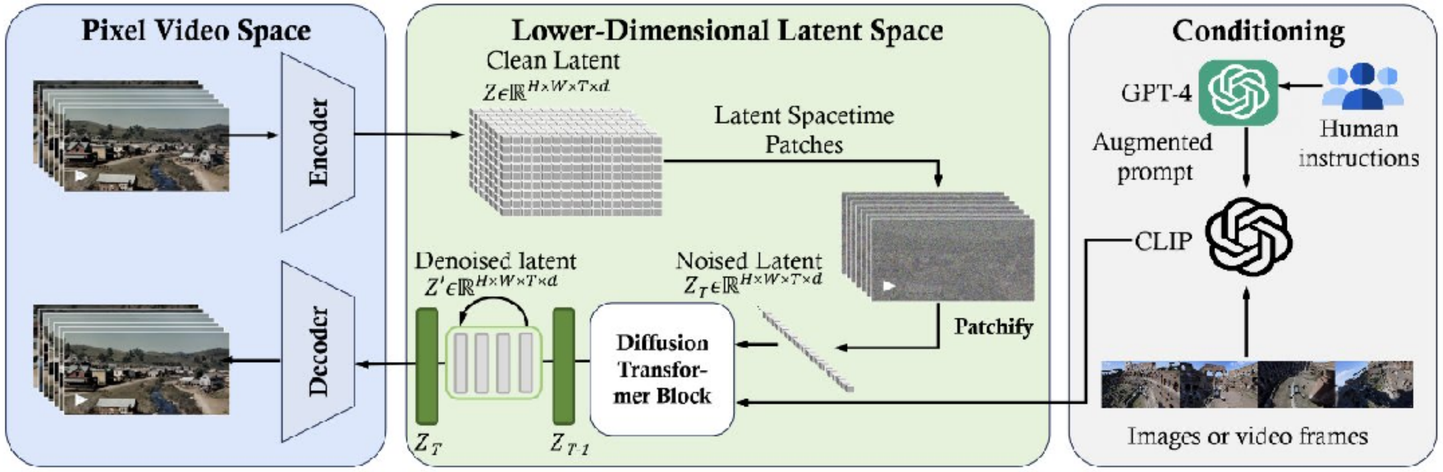
Látens “tér-idő” patch-ek

Videó Generálás

(Open)Sora

Sora

https://github.com/hpcaitech/Open-sora



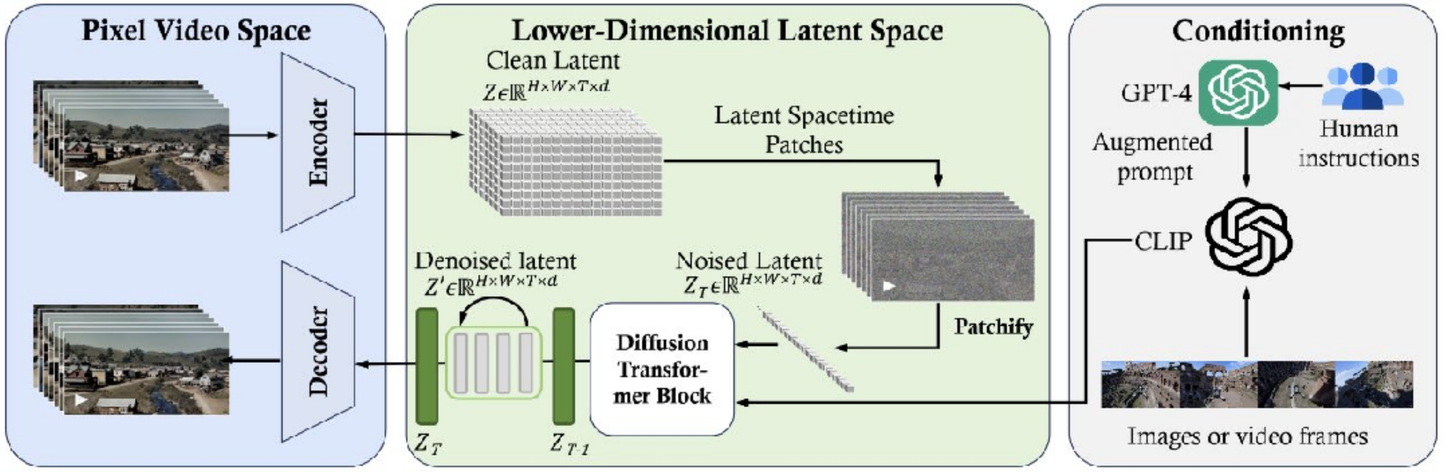
Látens “tér-idő” patch-ek

Videó Generálás

(Open)Sora

Sora

https://github.com/hpcaitech/Open-sora

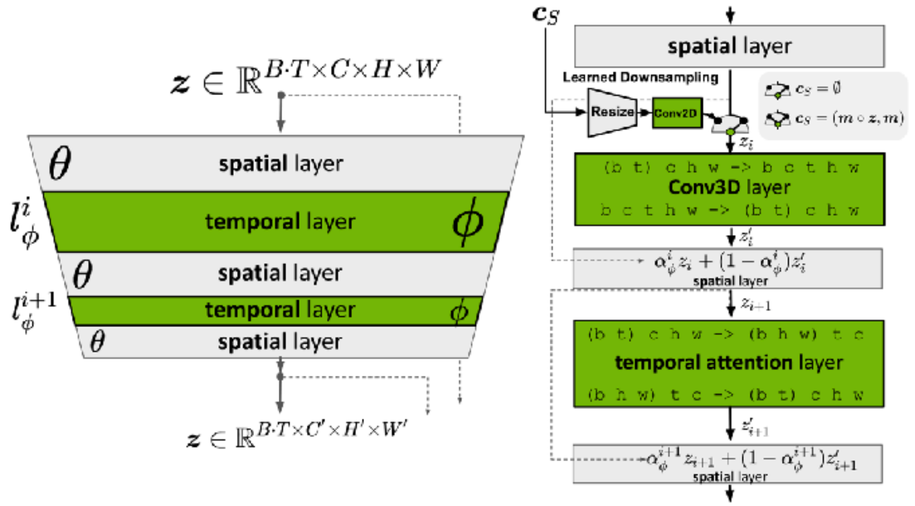


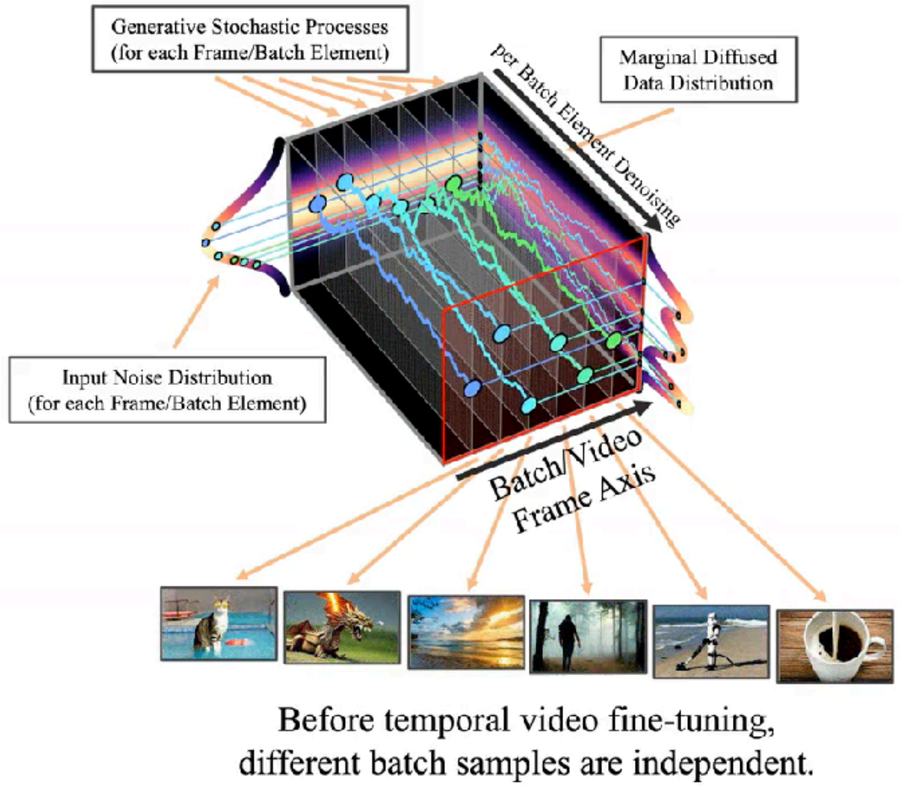
Látens “tér-idő” patch-ek

Videó Generálás

Align Your Latents

- Probléma: az egymástól függetlenül generált látens frame-ek nem konzisztensek
- Előtanított képgenerátor finomhangolása temporális rétegekkel



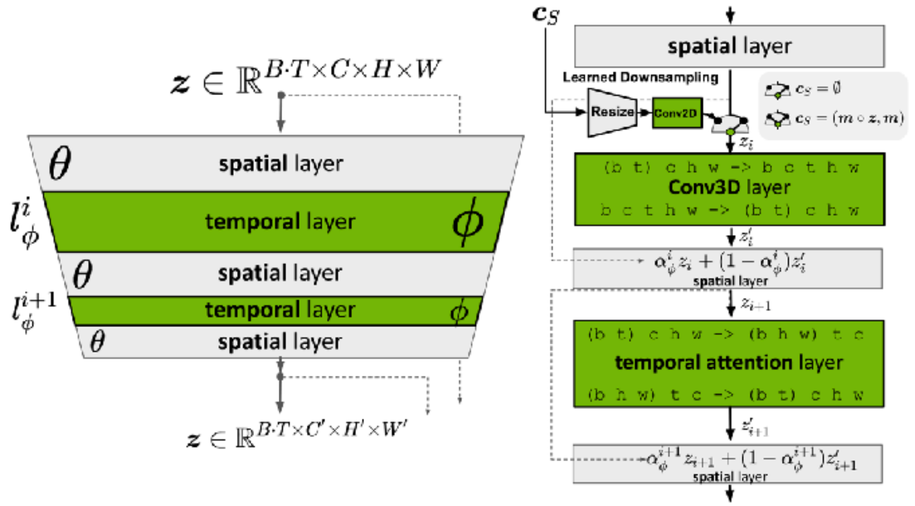


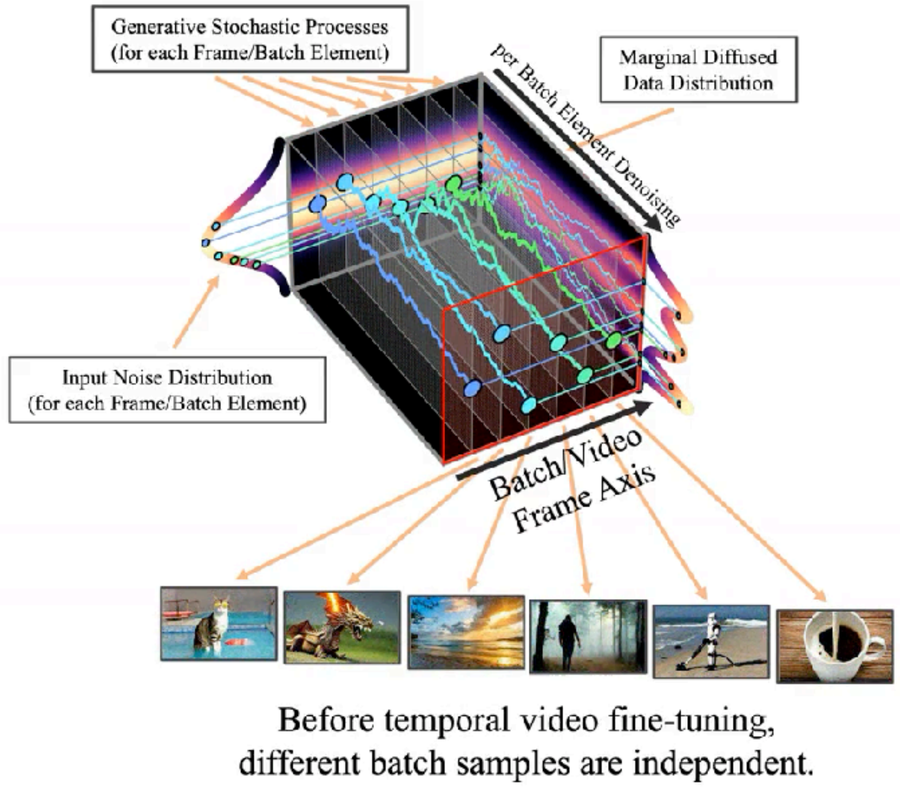
Before temporal video fine-tuning,
different batch samples are independent.

Videó Generálás

Align Your Latents

- Probléma: az egymástól függetlenül generált látens frame-ek nem konzisztensek
- Előtanított képgenerátor finomhangolása temporális rétegekkel





Before temporal video fine-tuning,
different batch samples are independent.

Videó Generálás

Meta Movie Gen

Meta

Movie Gen: A Cast of Media Foundation Models

The Movie Gen team @ Meta[1]

https://ai.meta.com/research/movie-gen/

Videó Generálás

Meta Movie Gen

Meta

Movie Gen: A Cast of Media Foundation Models

The Movie Gen team @ Meta[1]

https://ai.meta.com/research/movie-gen/

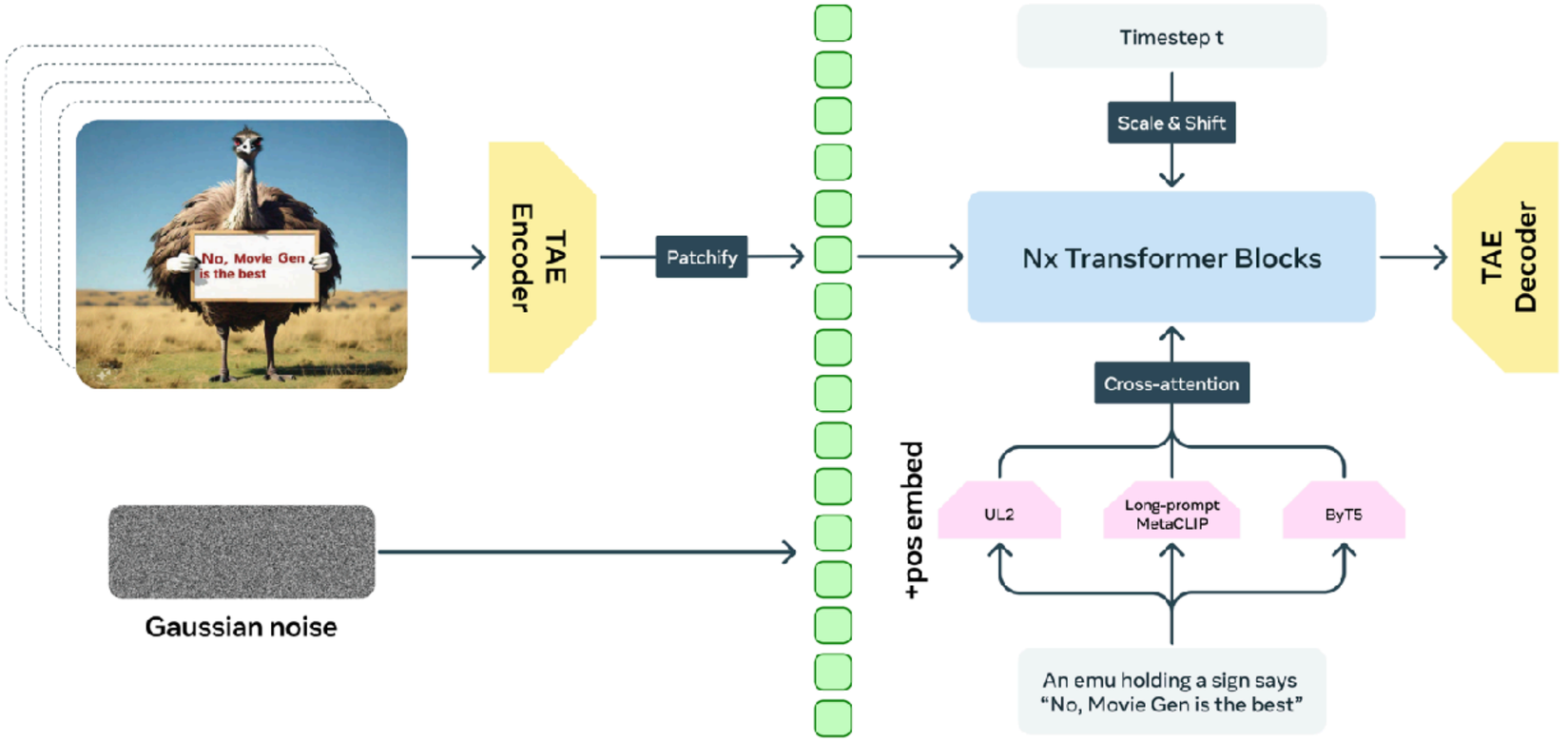
Videó Generálás

Meta Movie Gen

Meta

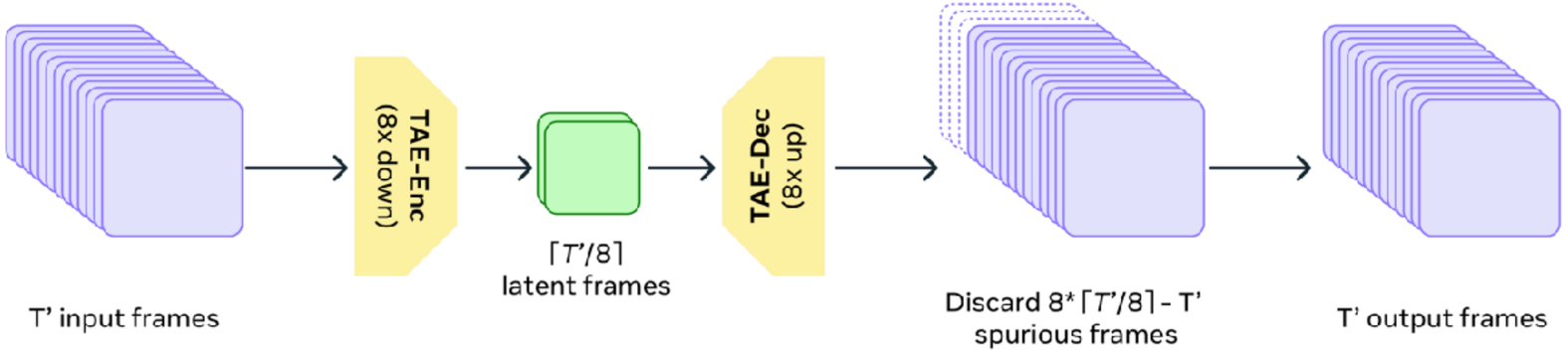
Movie Gen: A Cast of Media Foundation Models

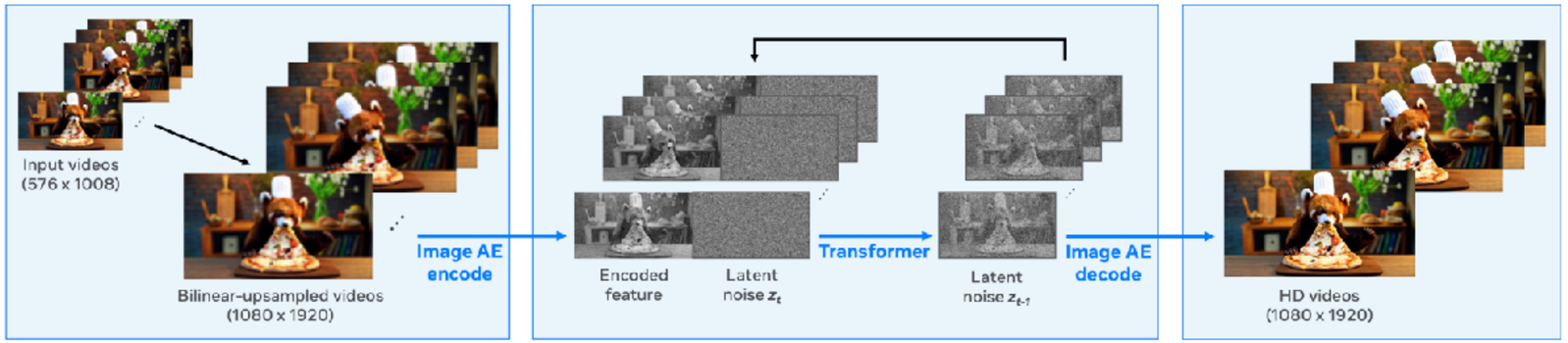
The Movie Gen team @ Meta[1]

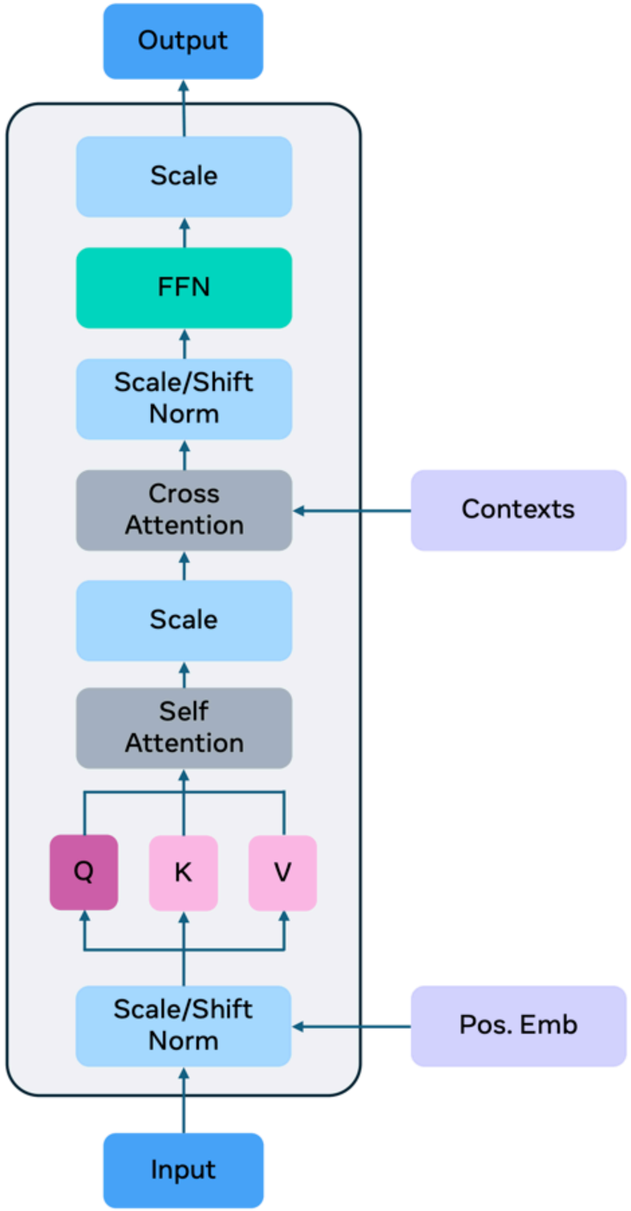


Videó Generálás

Meta Movie Gen

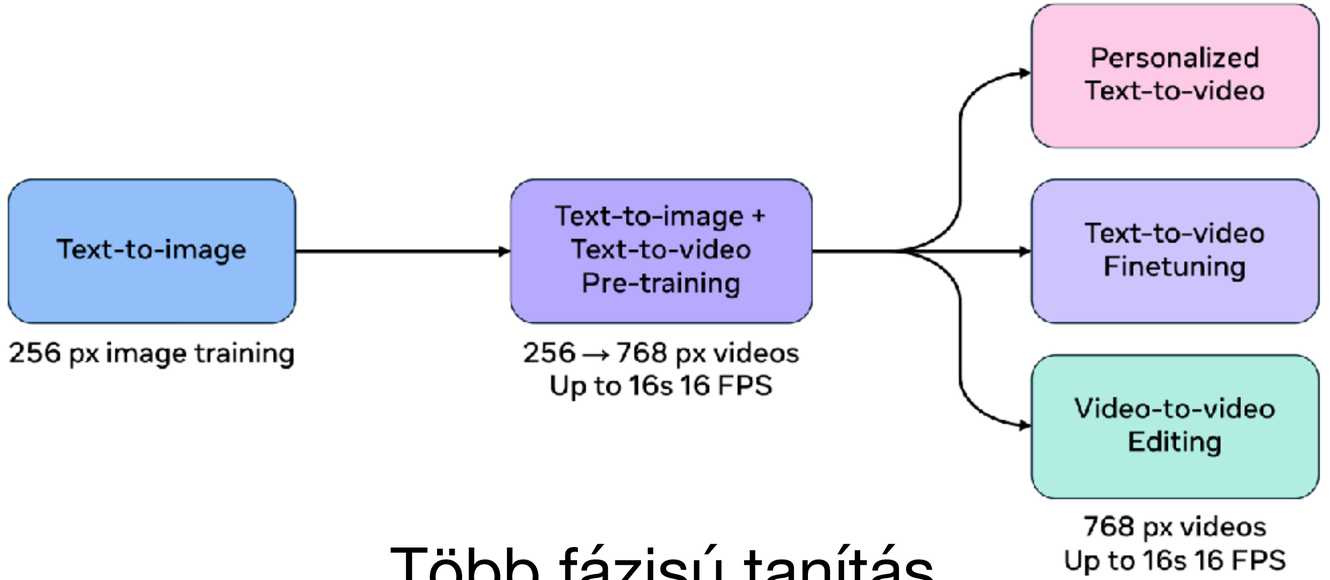




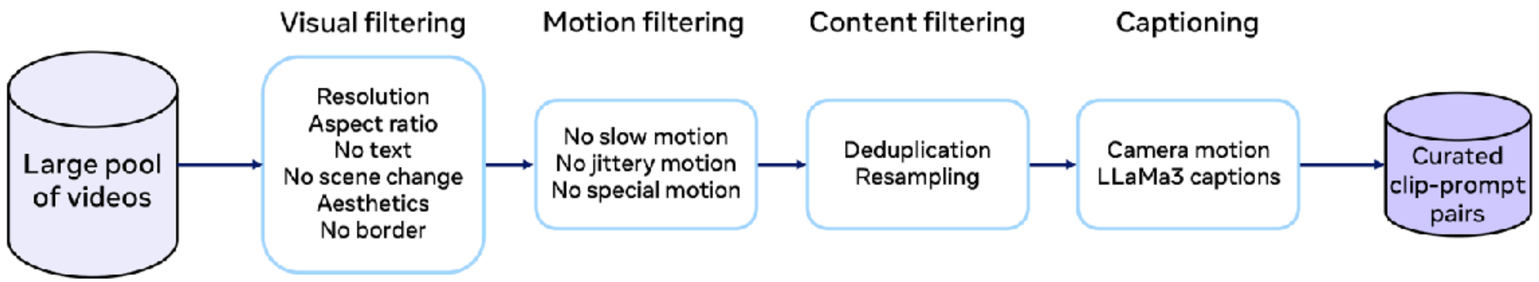


Videó Generálás

Meta Movie Gen — Tanítás



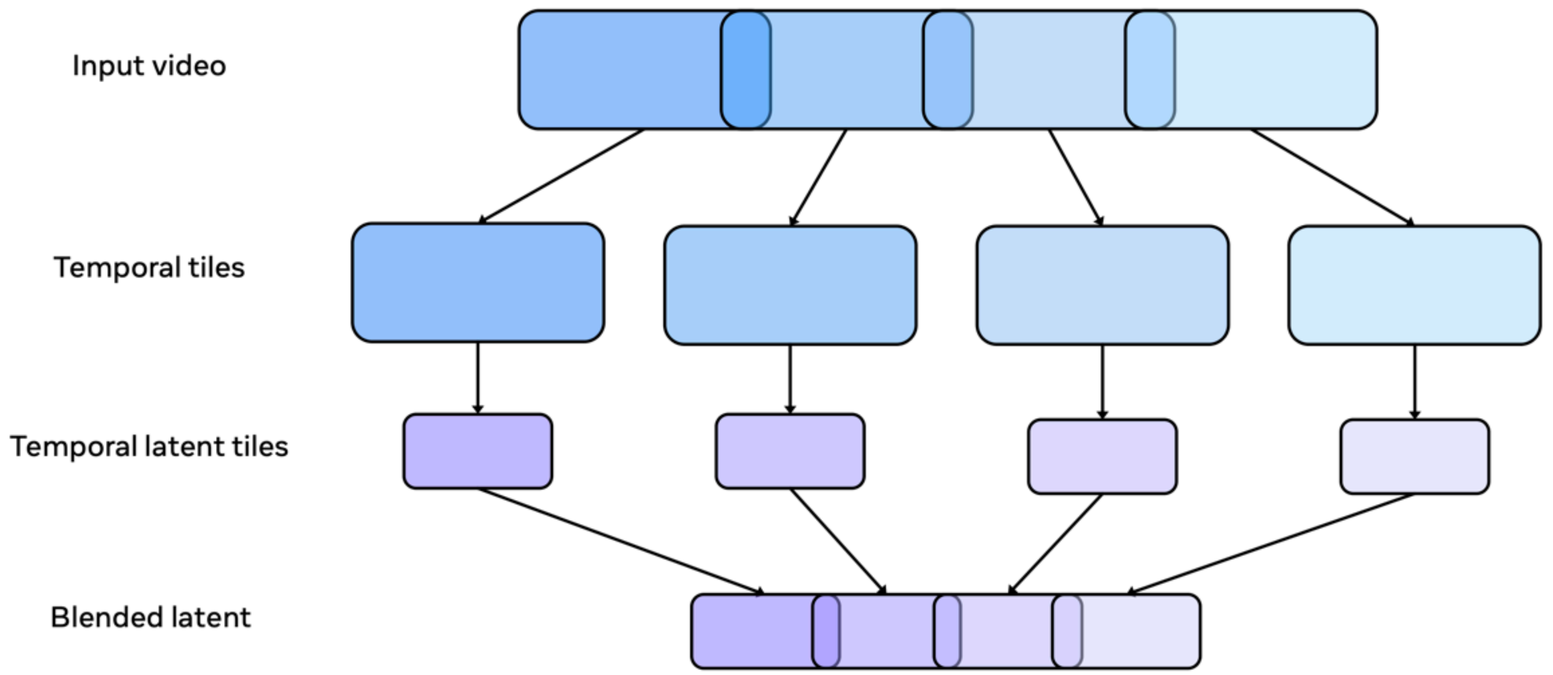
Több fázisú tanítás



Adathalmaz kurálása

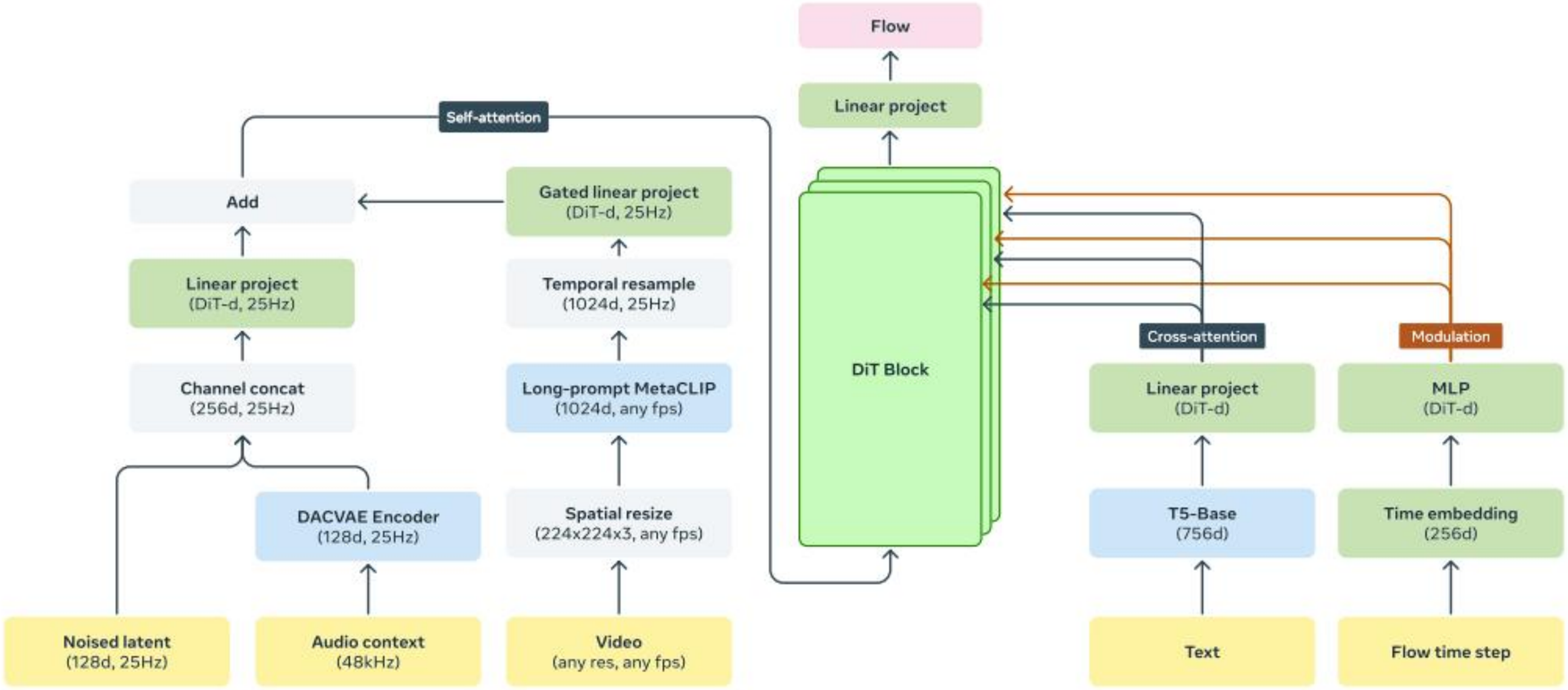
Videó Generálás

Meta Movie Gen — Tiling



Videó Generálás

Meta Movie Gen — Hang generálás

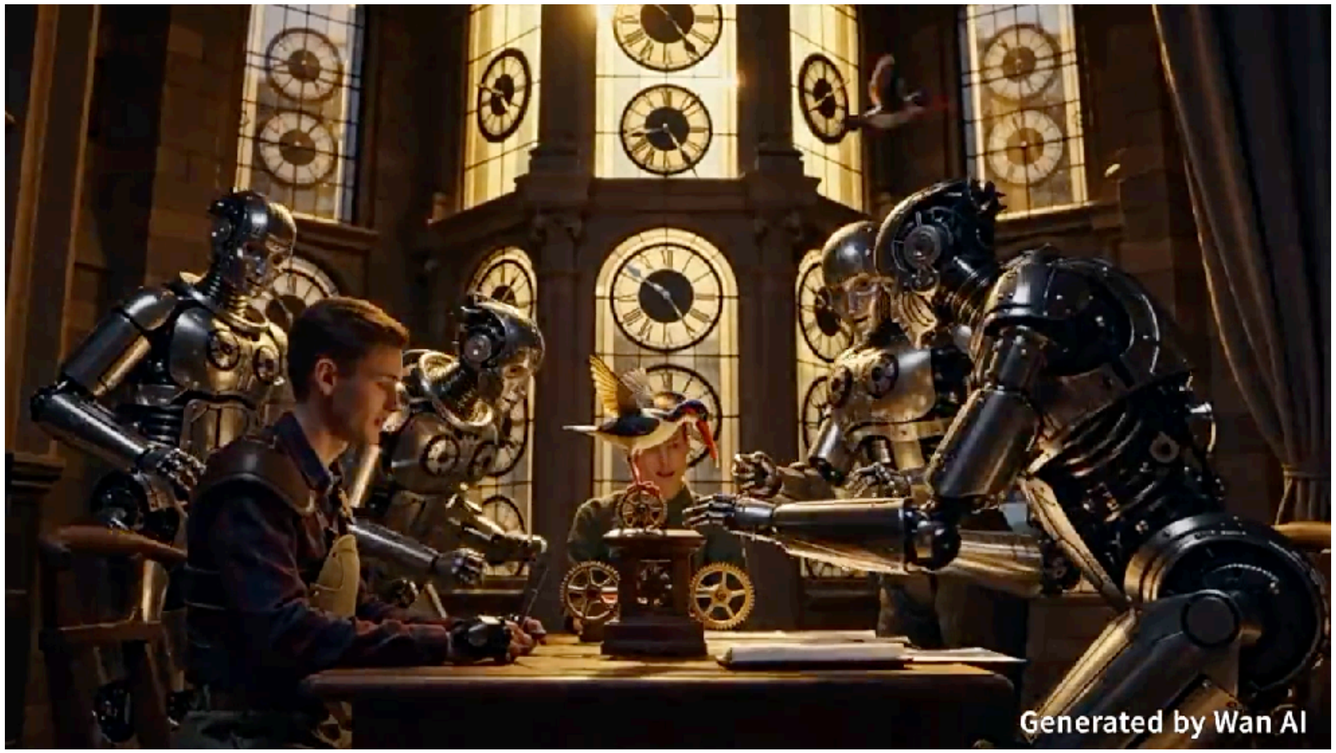


Videó Generálás

Alibaba Wan

WAN: OPEN AND ADVANCED LARGE-SCALE VIDEO GENERATIVE MODELS

Wan Team, Alibaba Group



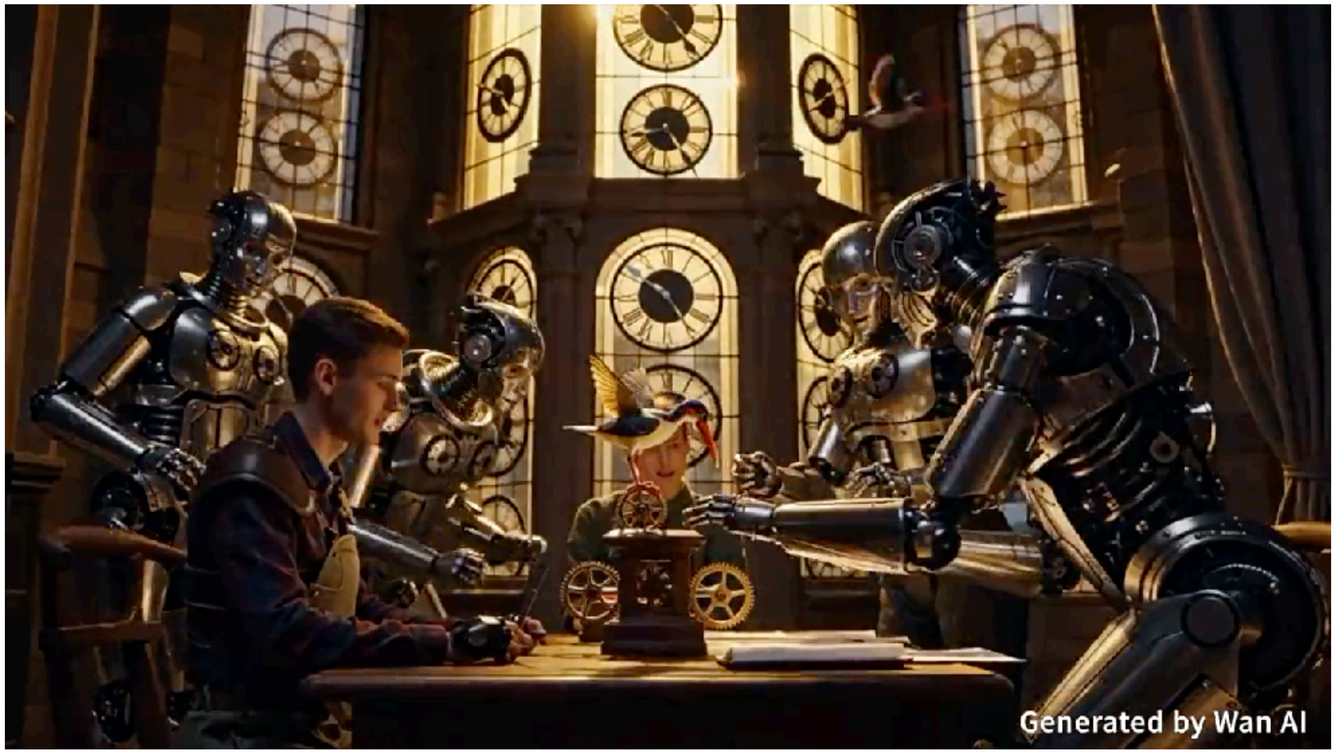
https://wan.video/

Videó Generálás

Alibaba Wan

WAN: OPEN AND ADVANCED LARGE-SCALE VIDEO GENERATIVE MODELS

Wan Team, Alibaba Group



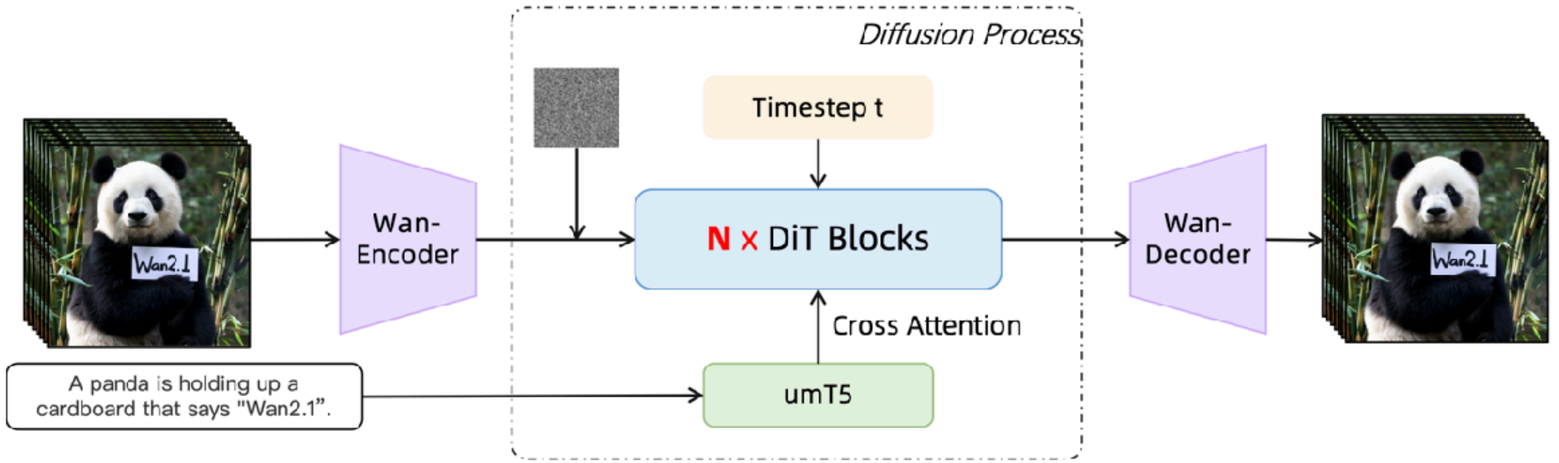
https://wan.video/

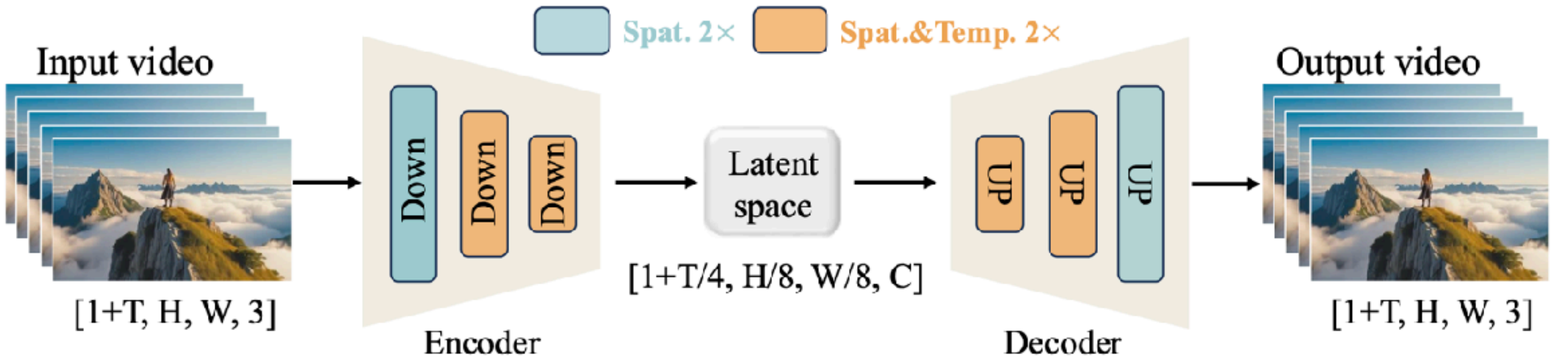
Videó Generálás

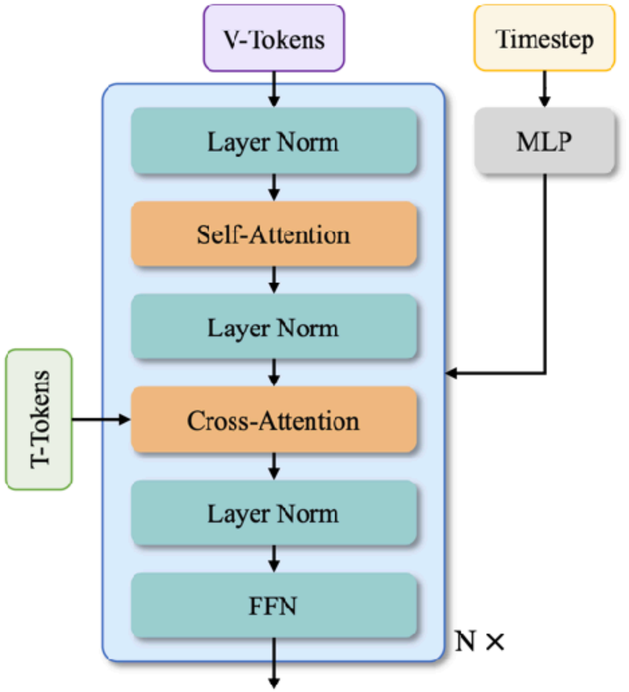
Alibaba Wan

WAN: OPEN AND ADVANCED LARGE-SCALE VIDEO GENERATIVE MODELS

Wan Team, Alibaba Group

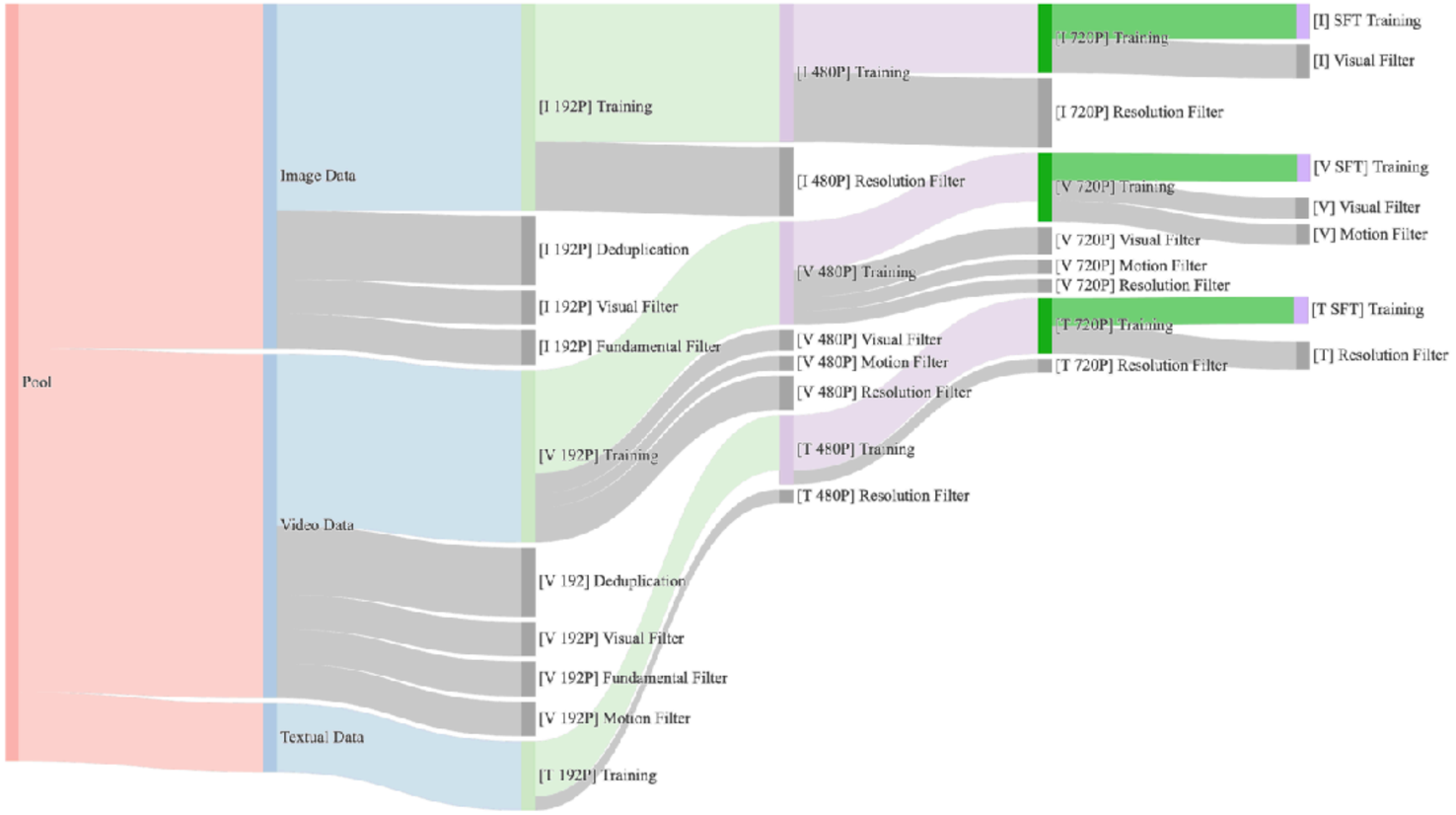




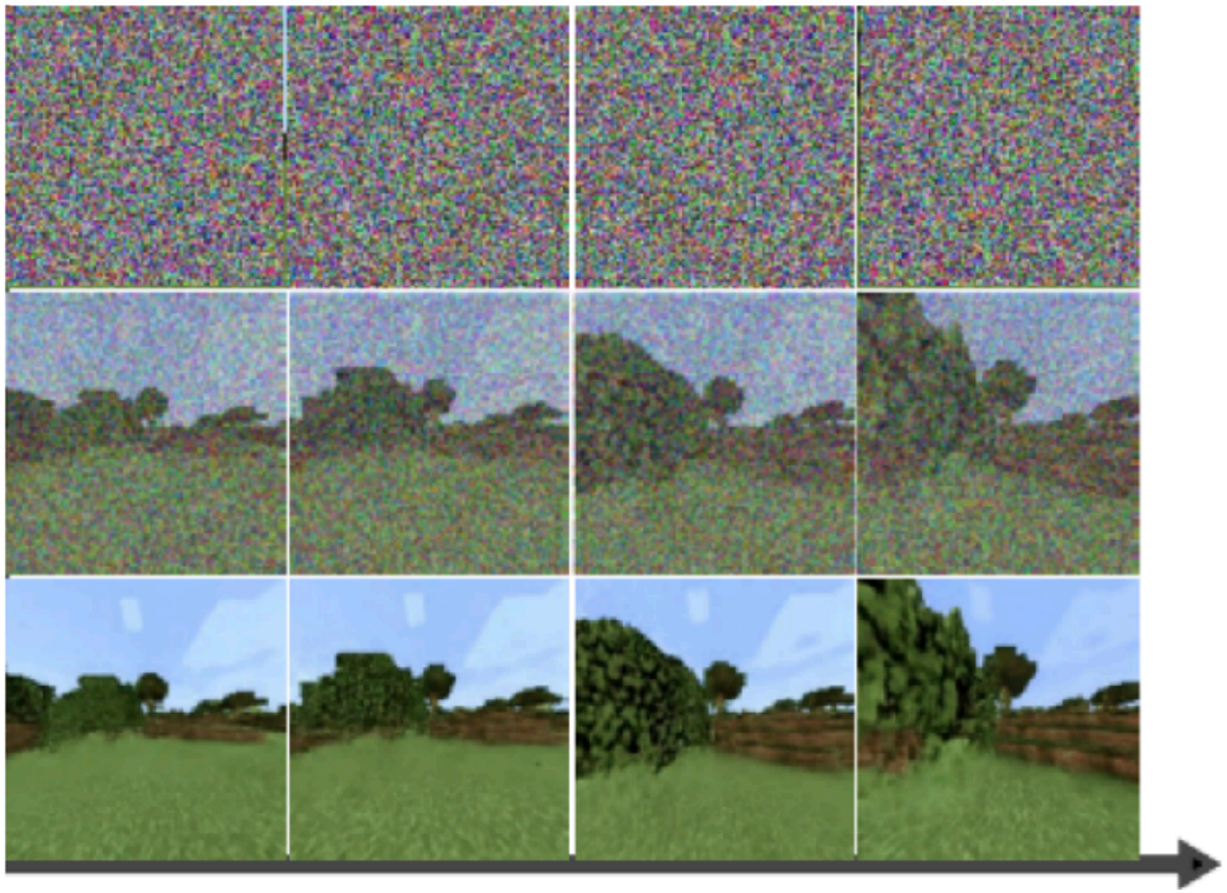


Videó Generálás

Alibaba Wan — Tanító adathalmaz



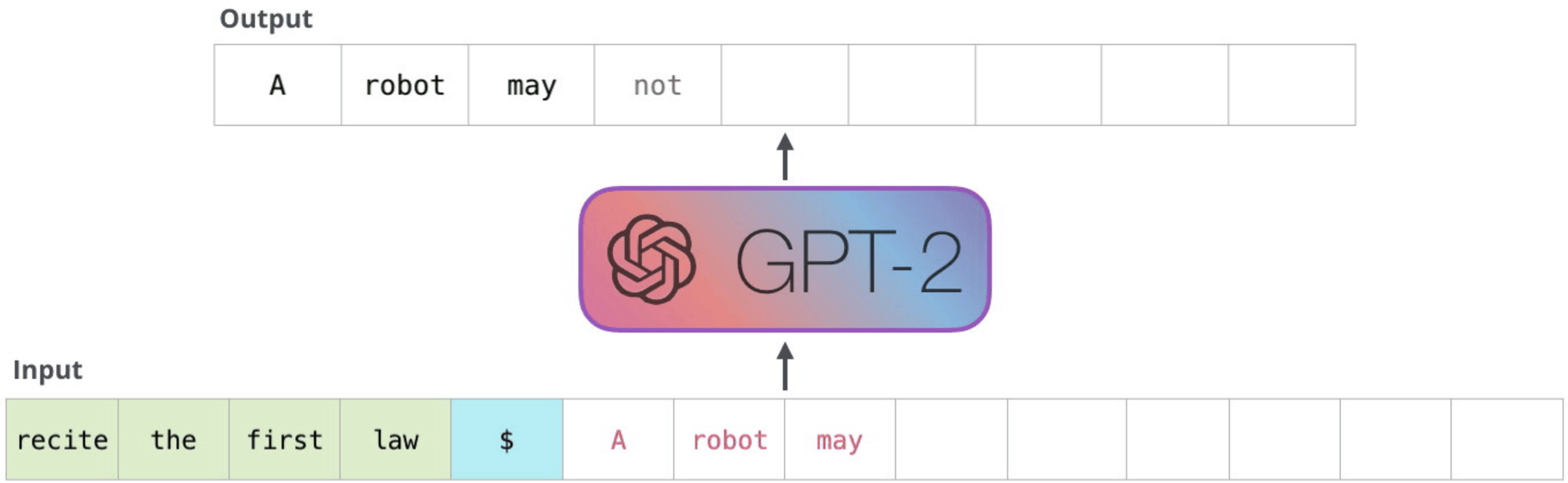
Videó Generálás



Eddig: diffúzió a teljes képszekvenciára — lehet máshogy is?

Videó Generálás

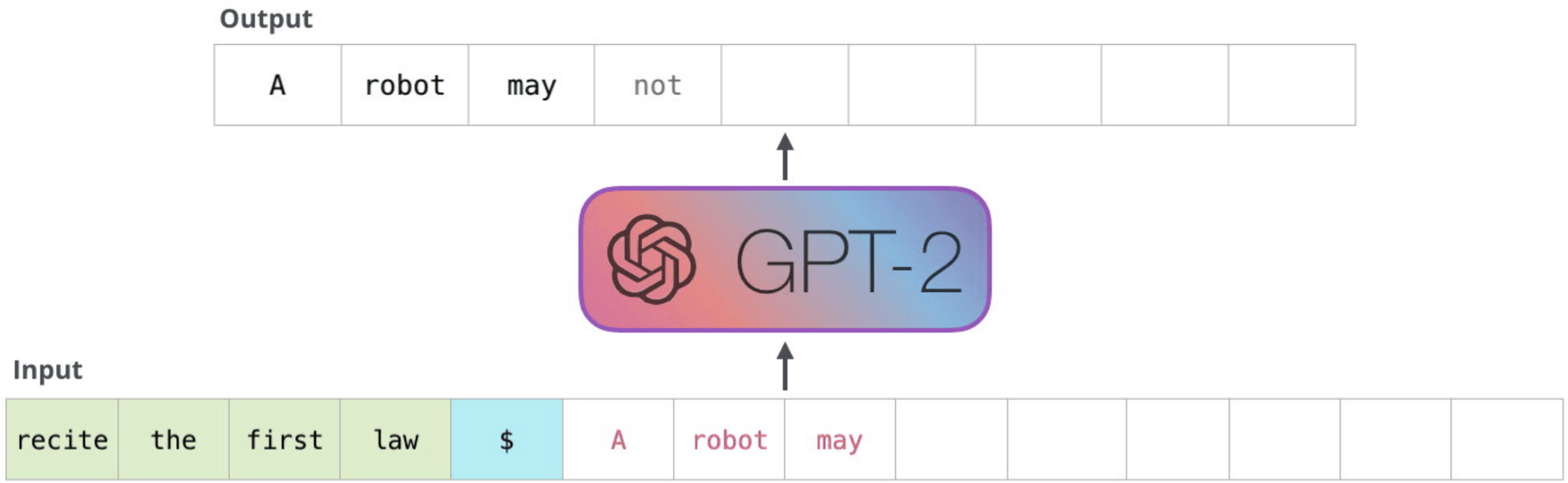
Autoregresszió?



Egy videó képek időben rendezett szekvenciája…
Lehetne autoregressziót használni a szövegek mintájára?

Videó Generálás

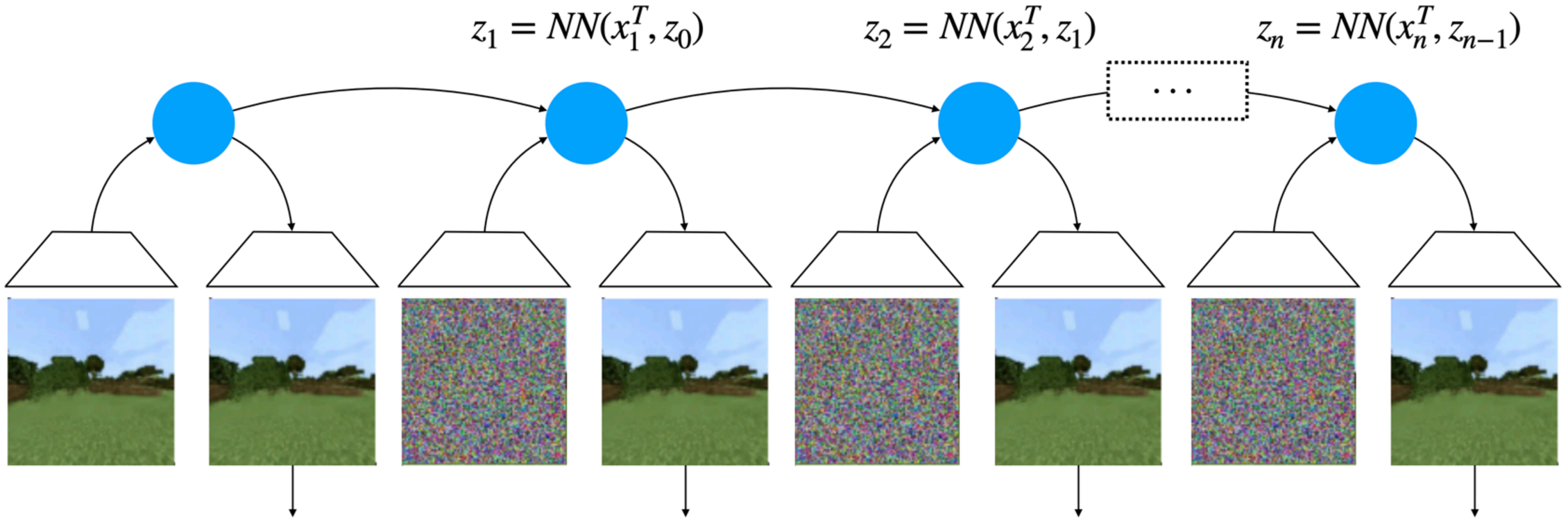
Autoregresszió?



Egy videó képek időben rendezett szekvenciája…
Lehetne autoregressziót használni a szövegek mintájára?

Videó Generálás

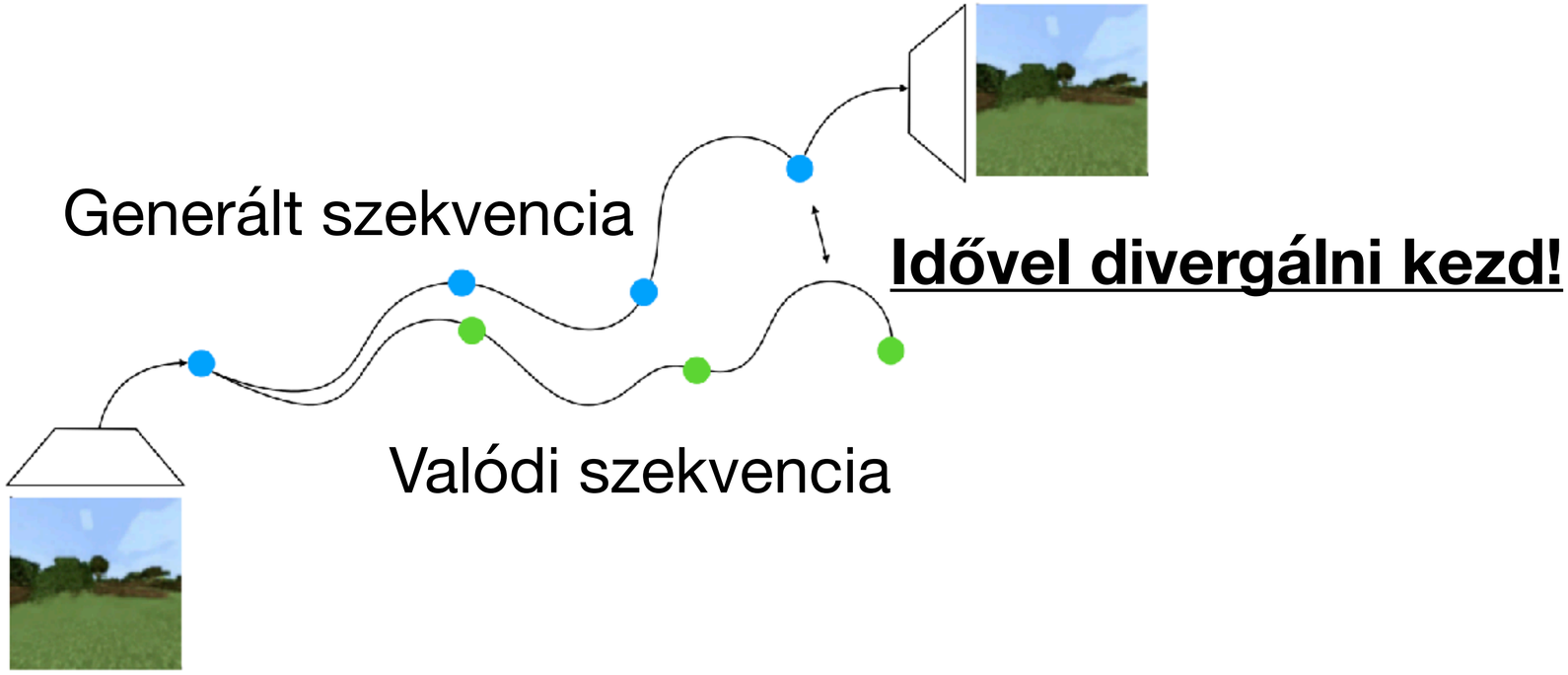
Autoregresszió — Szabad futás



Tanítás szabad futással (free running): futtassuk az aktuális modellt a felügyelt pillanatig

Videó Generálás

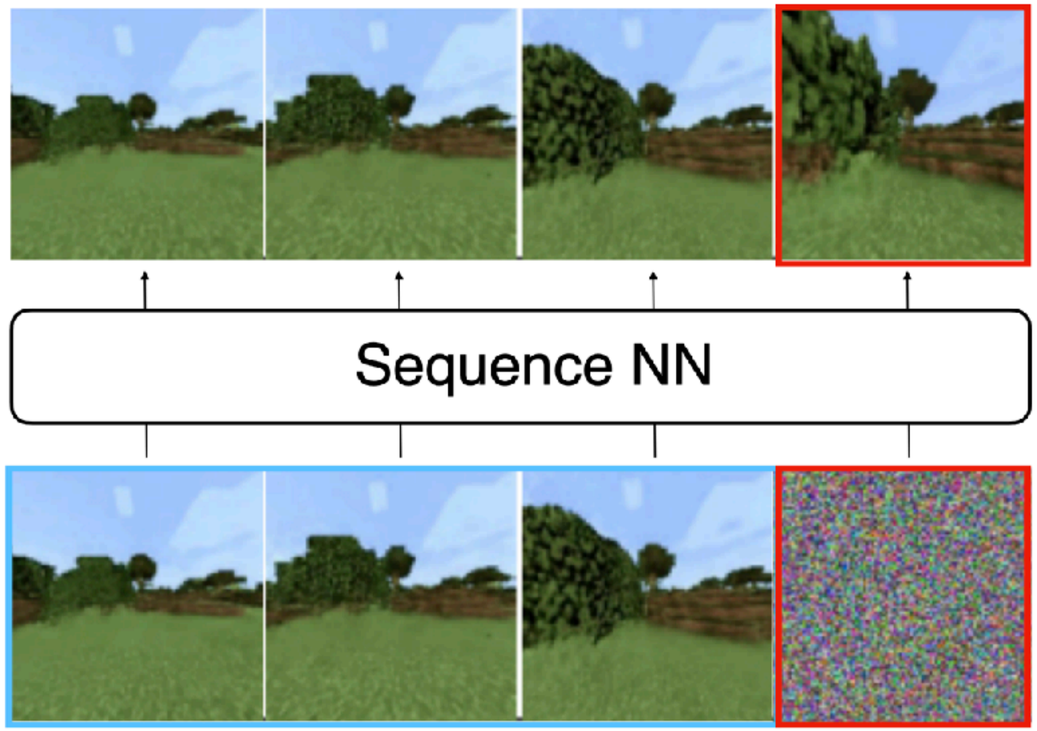
Autoregresszió — Szabad futás



Szabad futással a hibák könnyen halmozódnak, nehéz stabilan tanítani…

Videó Generálás

Autoregresszió — Teacher forcing

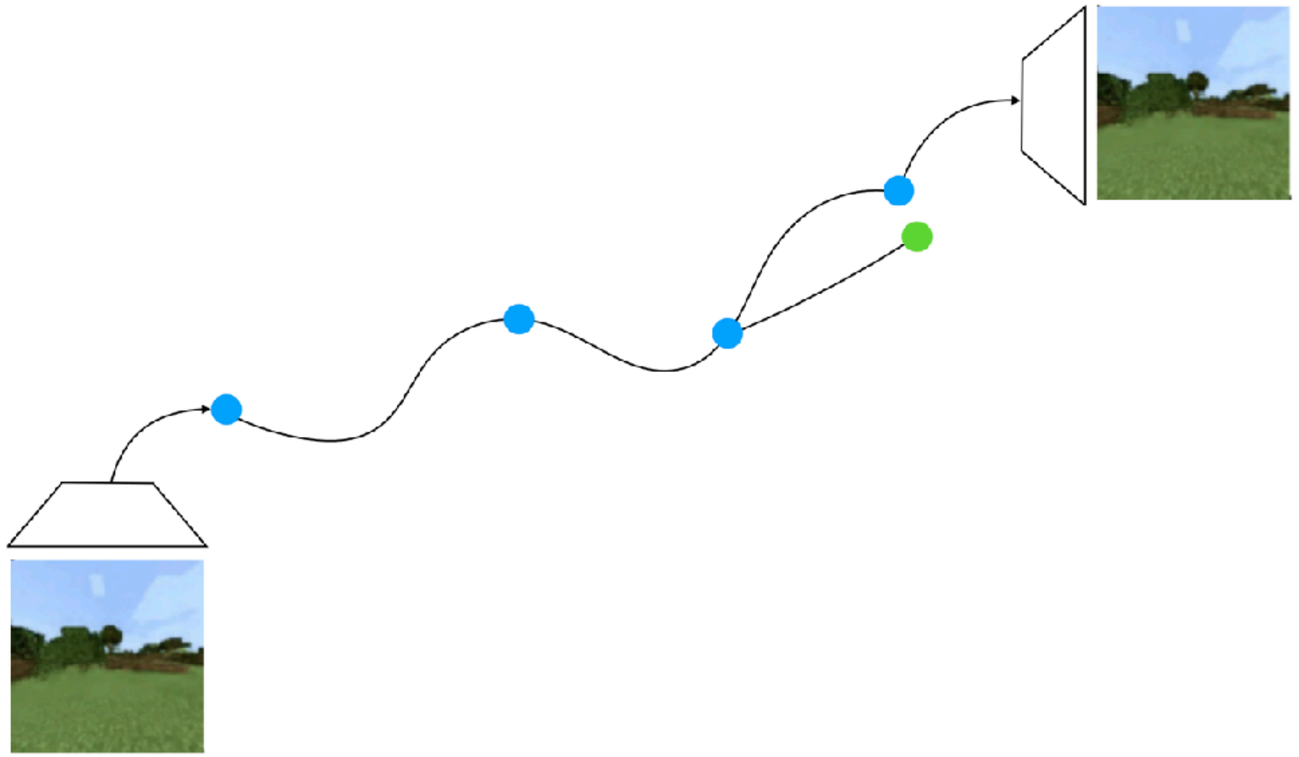


“Teacher forcing”: tanítás során mindig csak egy tokent (frame-et) prediktálunk, a korábbiakat a “ground truth” adatokból vesszük!

Videó Generálás

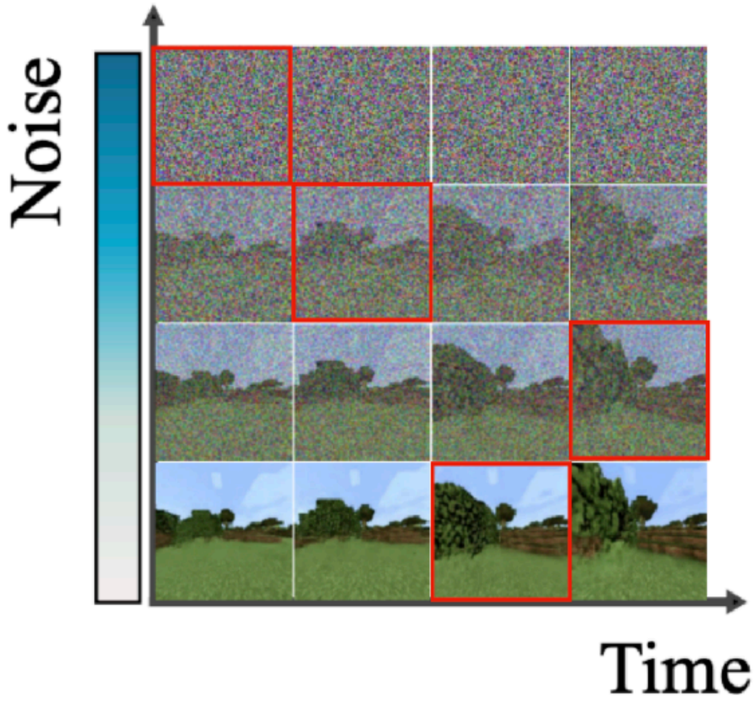
Autoregresszió — Teacher forcing

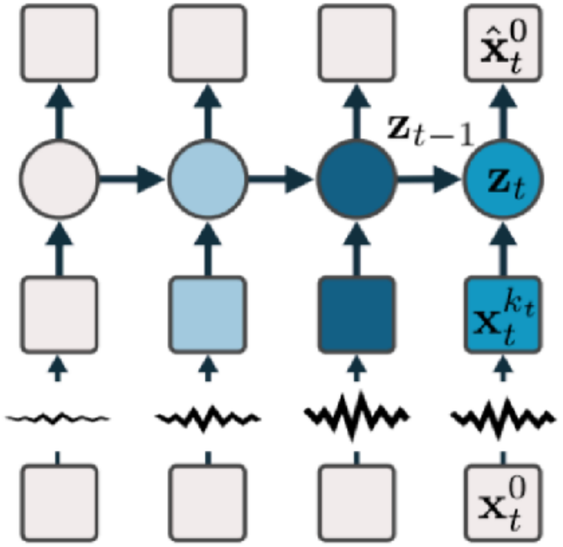
- Teacher forcing csökkenti a divergenciát…
- De “élesben” már generált frame-ek alapján kell újabbakat generálni — “out of distribution” a tanított modell szempontjából!
- (Szöveg generálásnál ez kisebb probléma a diszkrét tokenek miatt…)



Videó Generálás

Diffusion forcing

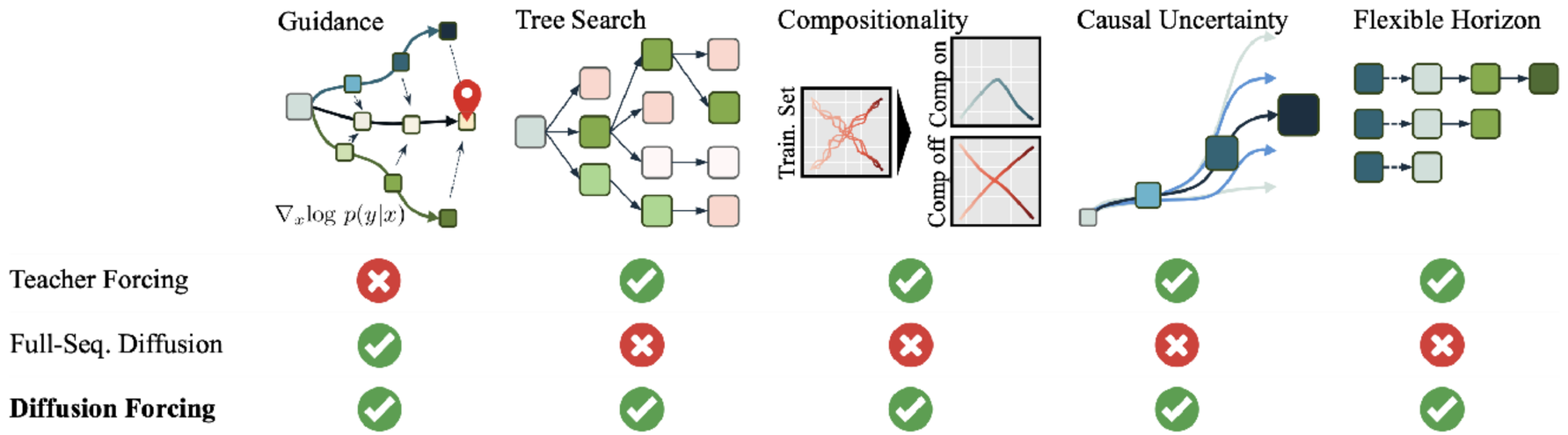




Ötlet: válasszunk különböző (random) mennyiségű zajt az egyes frame-ekre — **diffusion forcing**

Videó Generálás

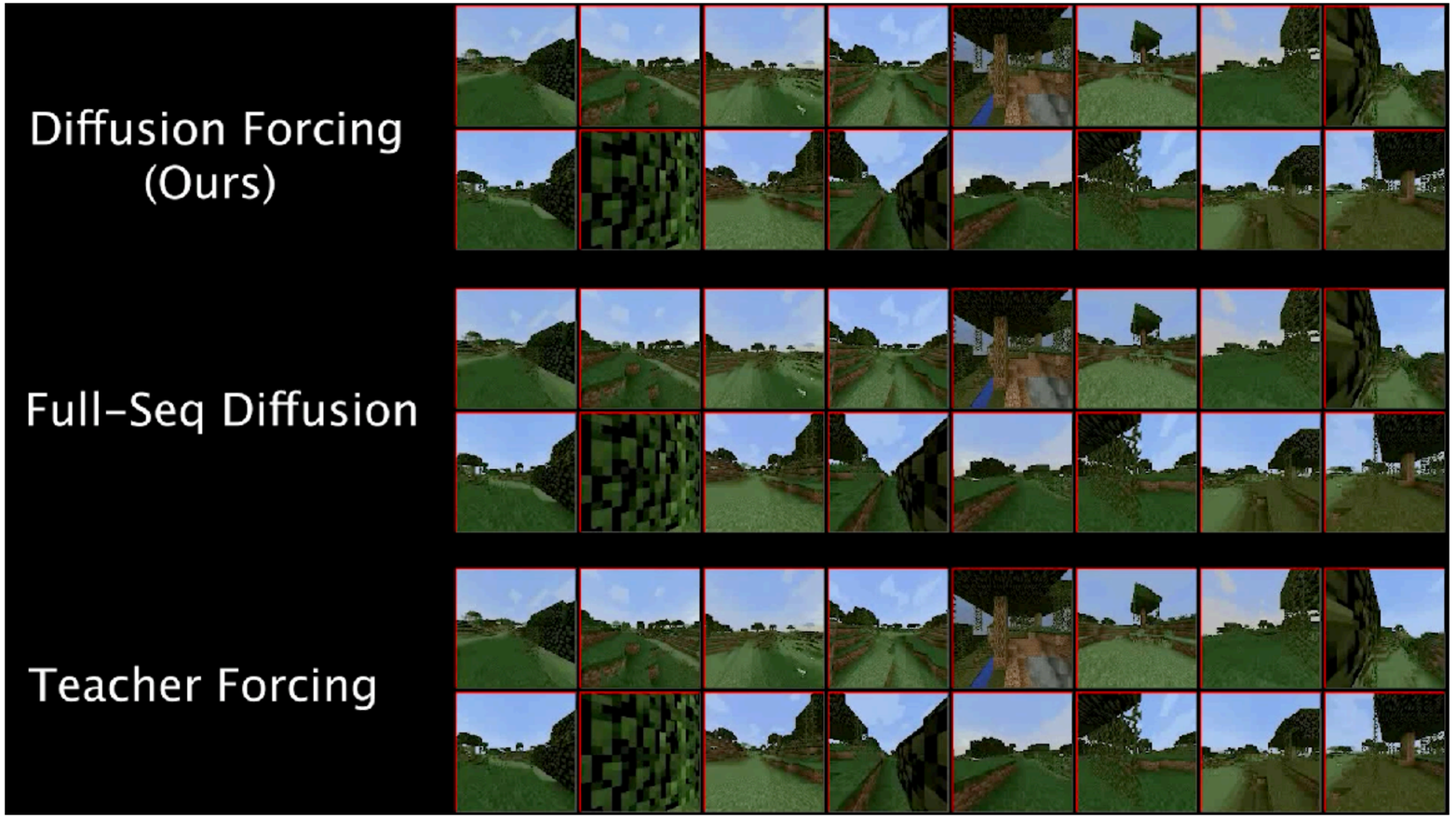
Diffusion forcing



Diffusion forcing: kombinálja a diffúzió és az autoregresszió előnyeit!

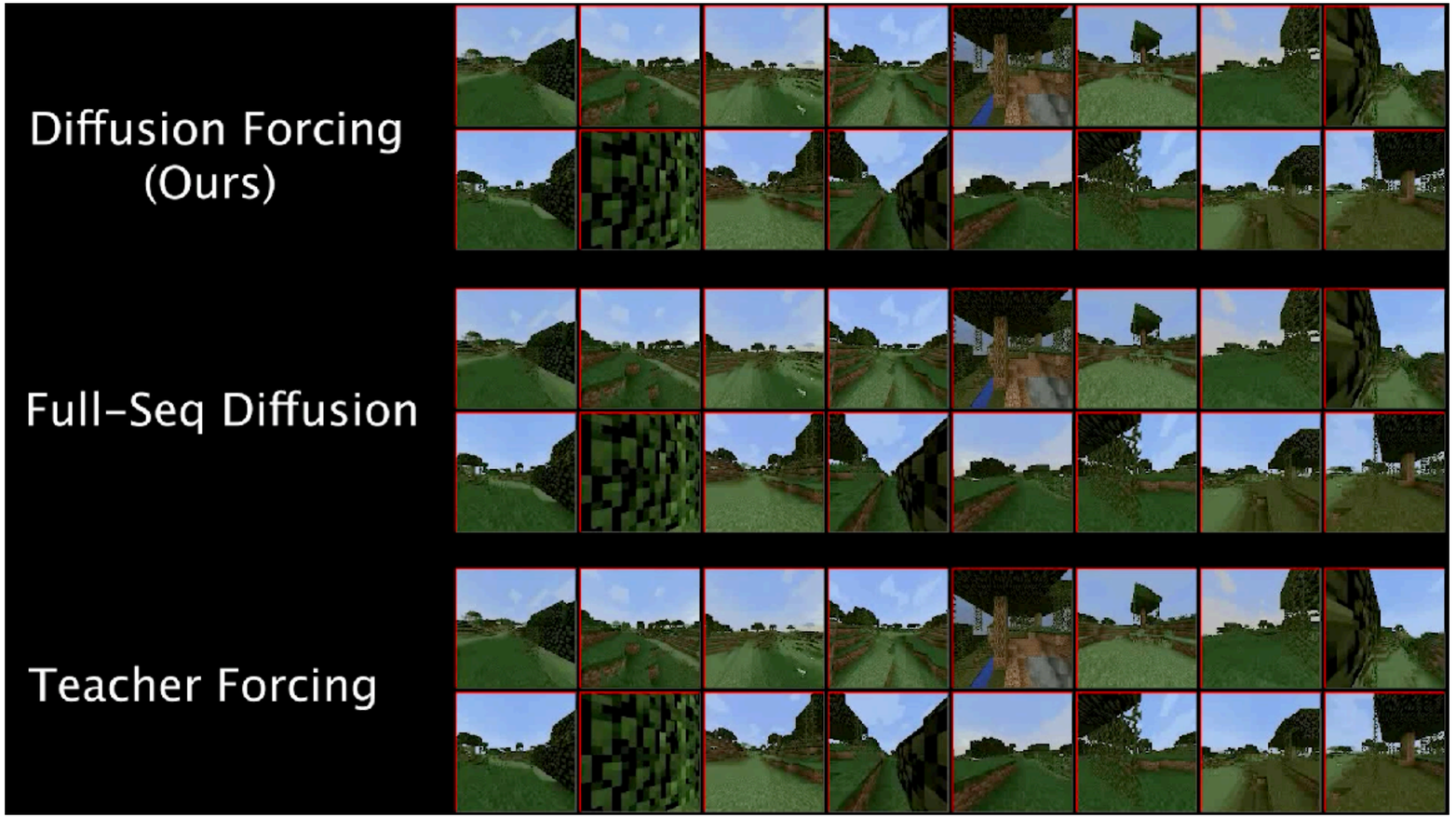
Videó Generálás

Diffusion forcing



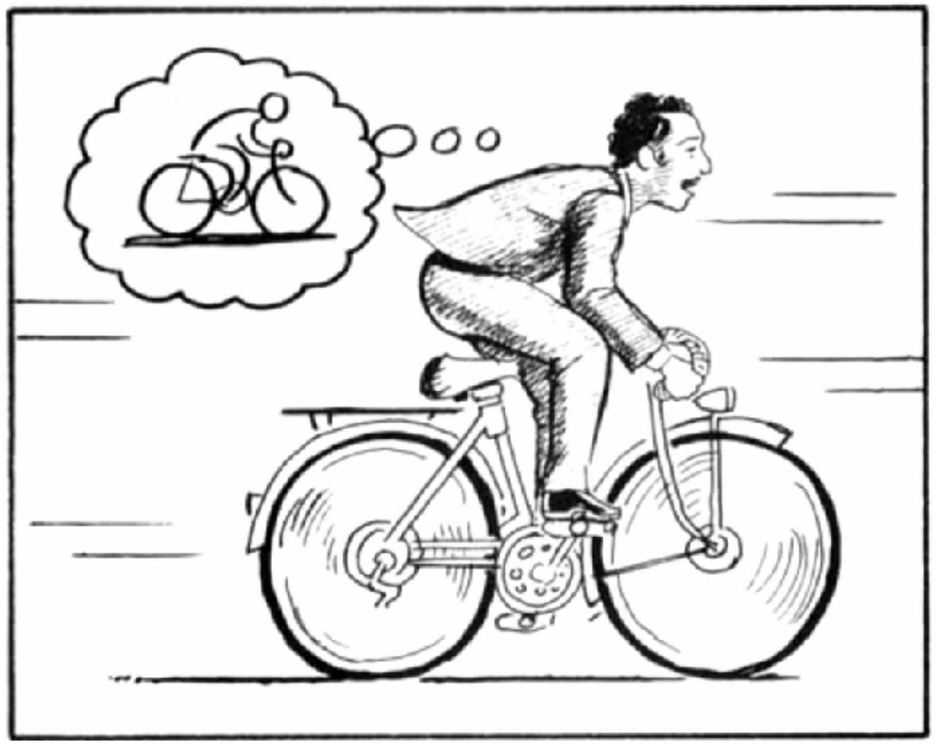
Videó Generálás

Diffusion forcing

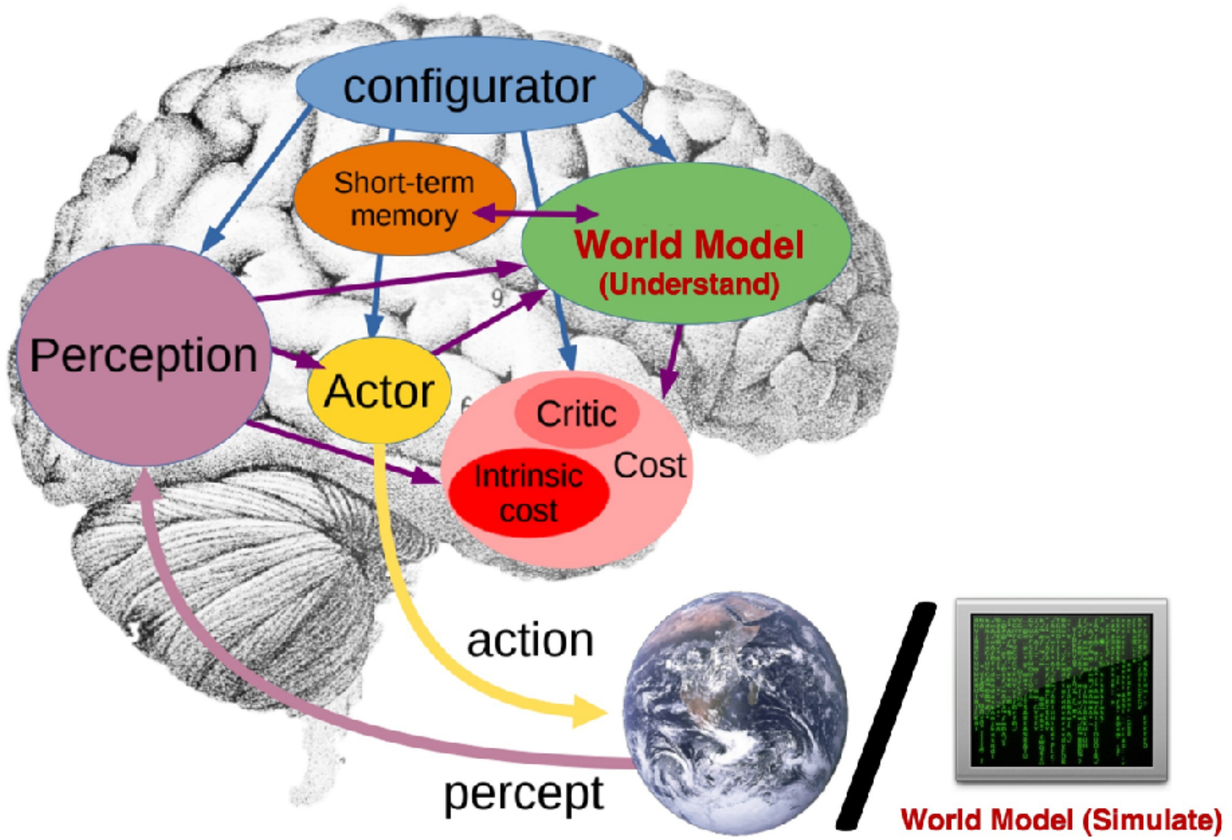


Videó Generálás

Világ modellek



https://worldmodels.github.io/



Videó Generálás

Világ modellek — Google Genie

https://deepmind.google/models/genie/

Videó Generálás

Világ modellek — Google Genie

https://deepmind.google/models/genie/

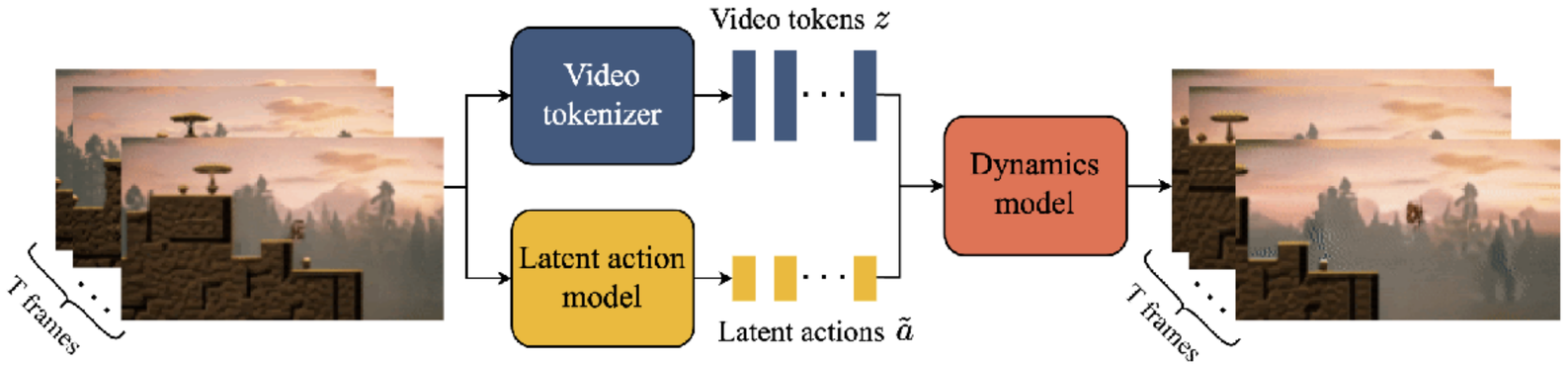
Videó Generálás

Világ modellek — Google Genie

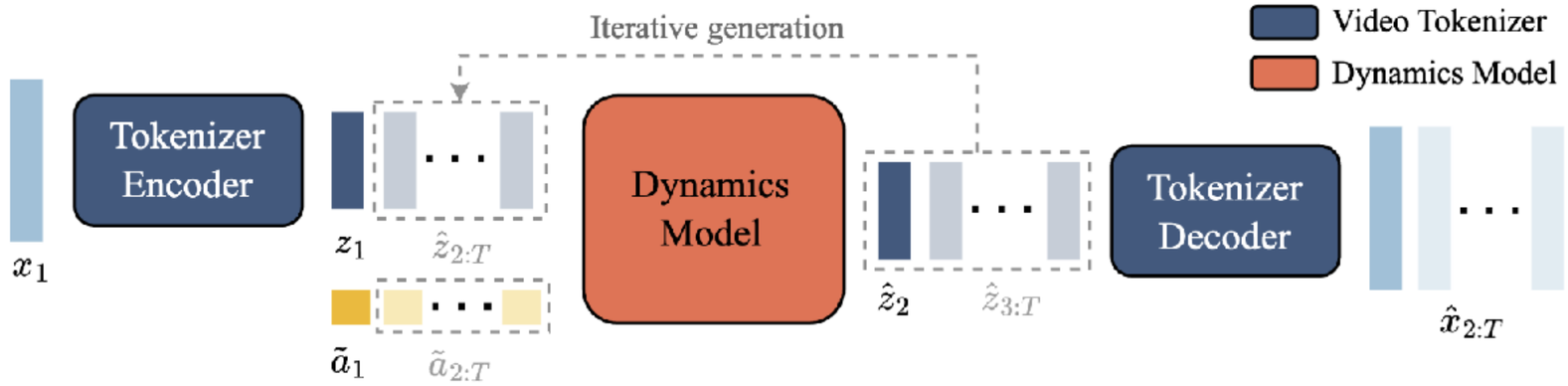
https://deepmind.google/models/genie/

Videó Generálás

Világ modellek — Google Genie



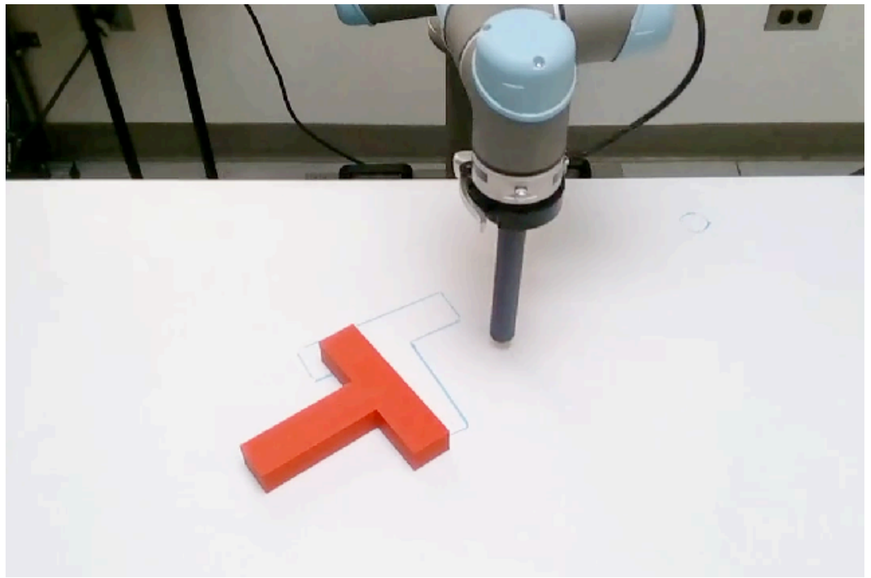
Tanítás



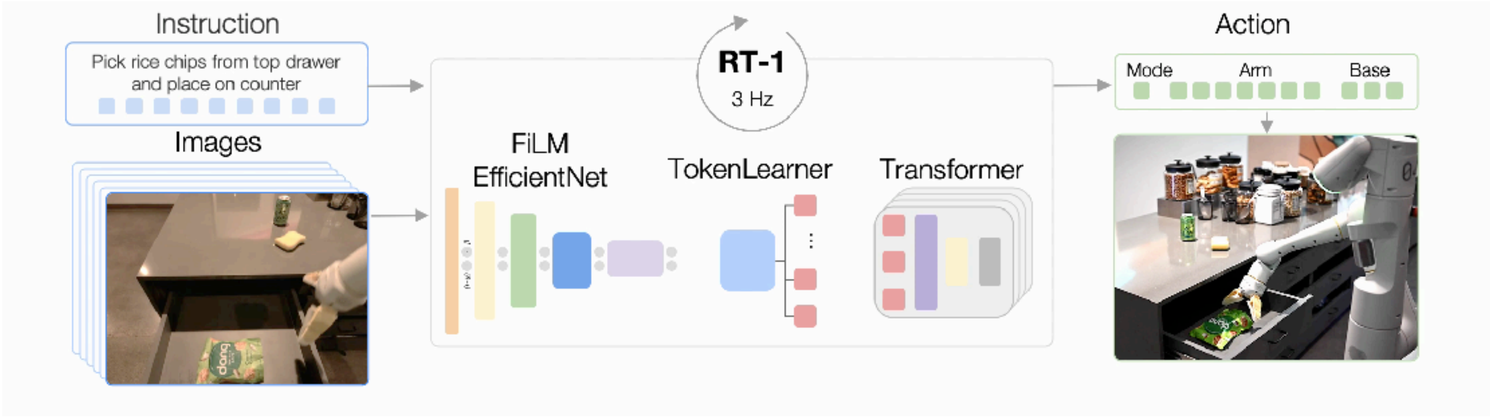
Generálás

Videó Generálás

Kitérő — Vision-Language-Action (VLA) Modellek*



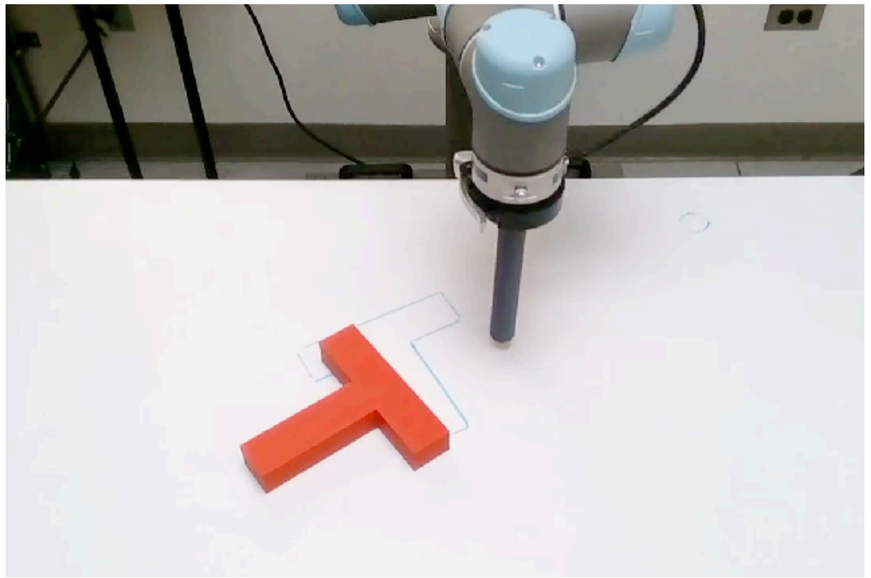
https://diffusion-policy.cs.columbia.edu/



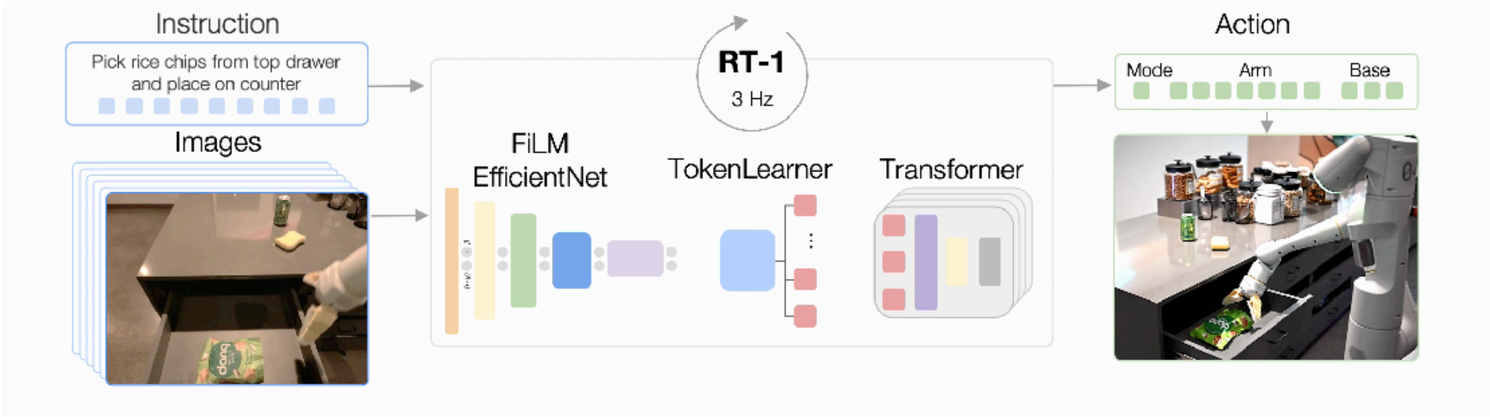
https://robotics-transformer1.github.io/

Videó Generálás

Kitérő — Vision-Language-Action (VLA) Modellek*



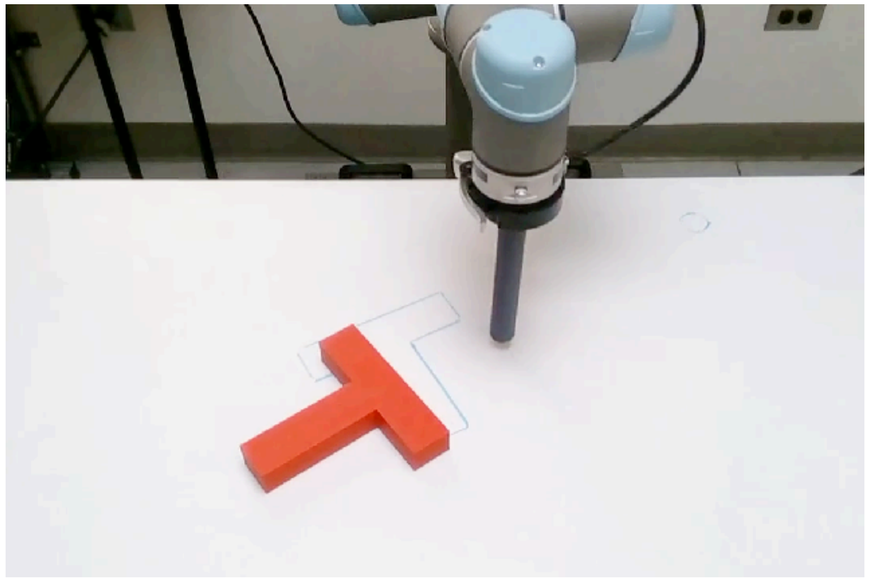
https://diffusion-policy.cs.columbia.edu/



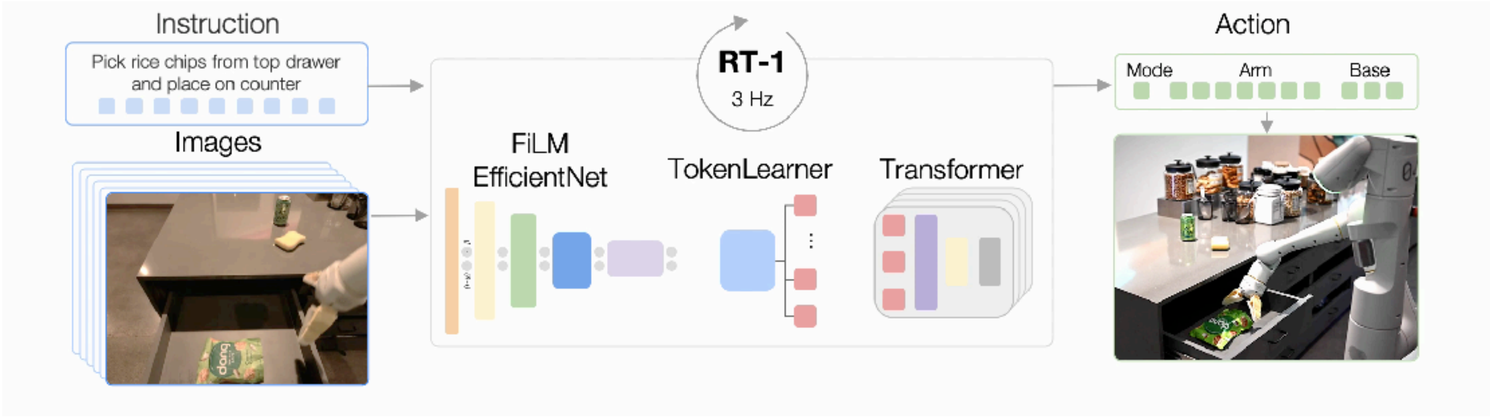
https://robotics-transformer1.github.io/

Videó Generálás

Kitérő — Vision-Language-Action (VLA) Modellek*



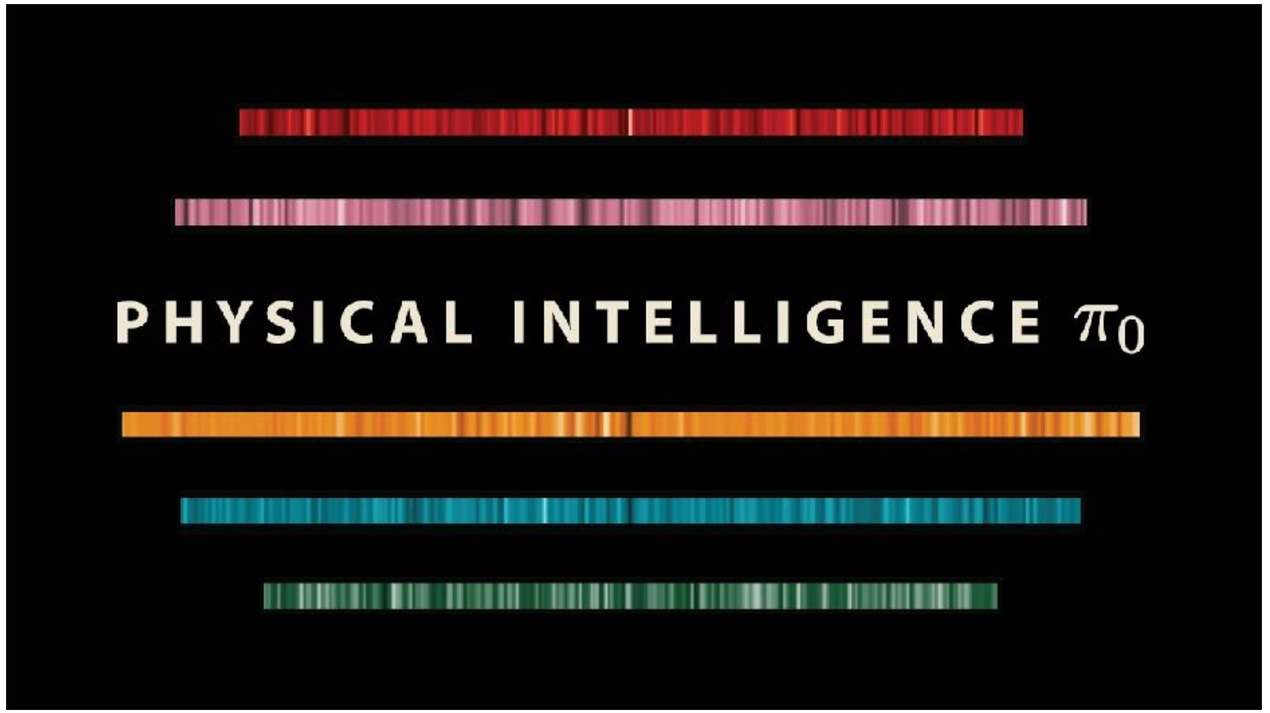
https://diffusion-policy.cs.columbia.edu/



https://robotics-transformer1.github.io/

Videó Generálás

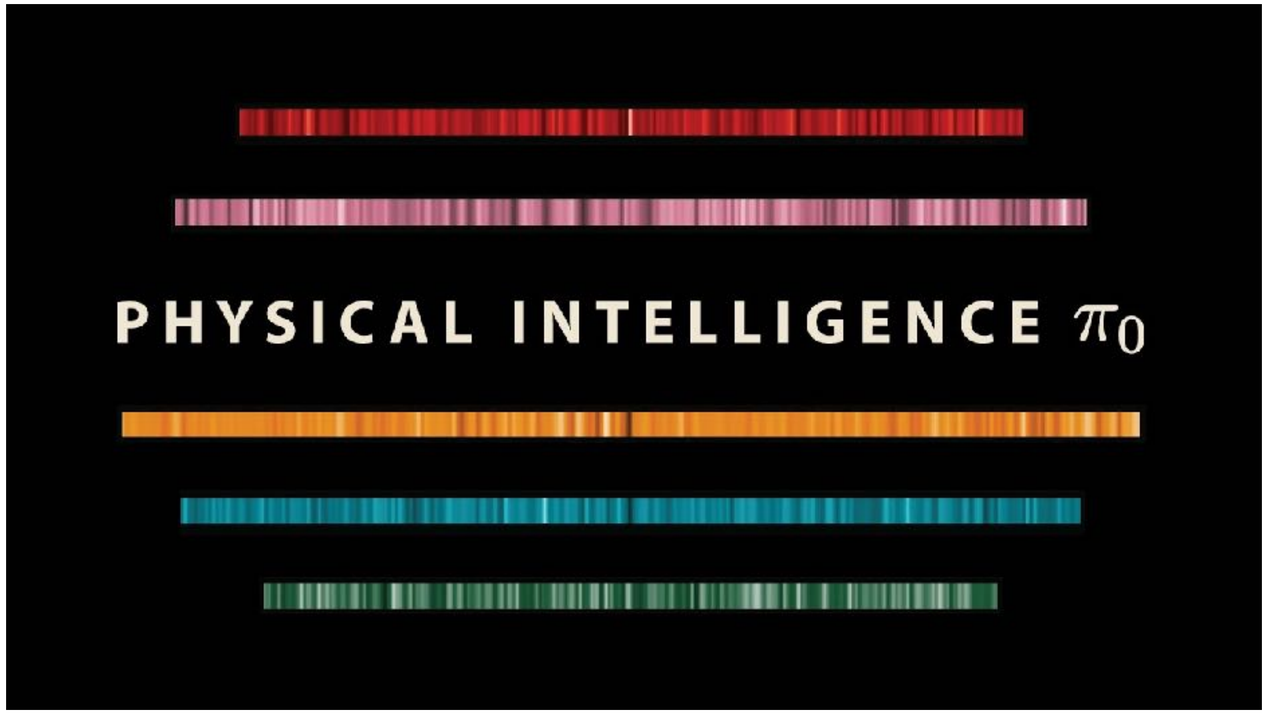
Kitérő — Vision-Language-Action (VLA) Modellek*



https://www.pi.website/blog/pi0

Videó Generálás

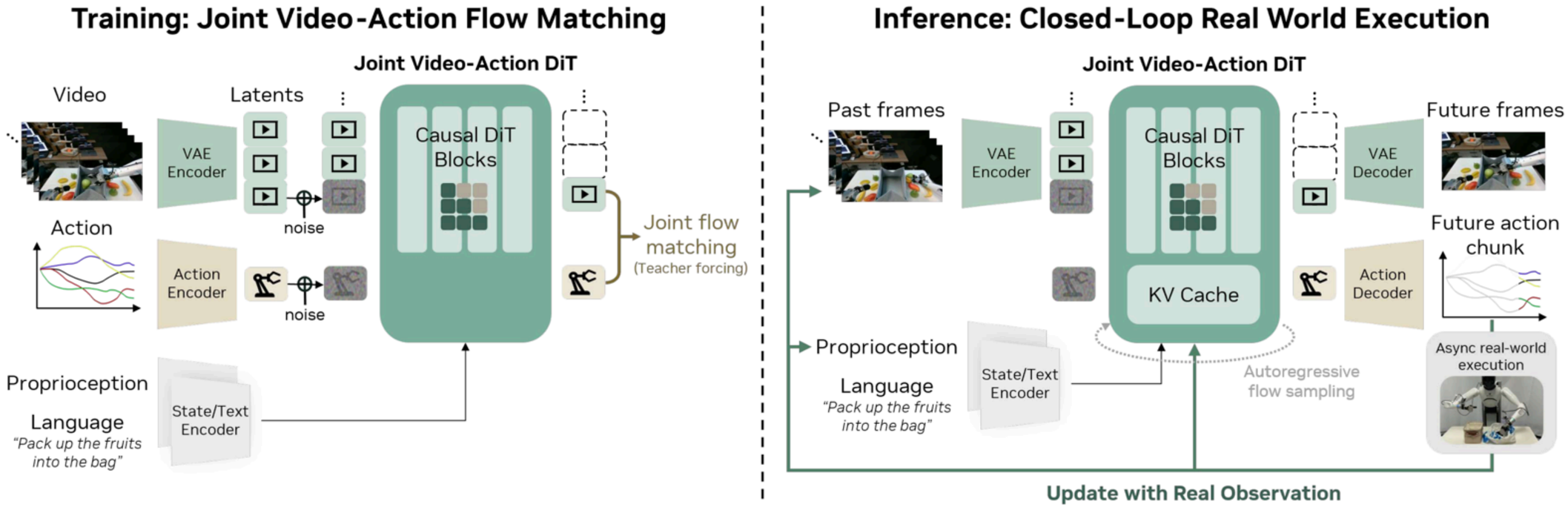
Kitérő — Vision-Language-Action (VLA) Modellek*



https://www.pi.website/blog/pi0

Videó Generálás

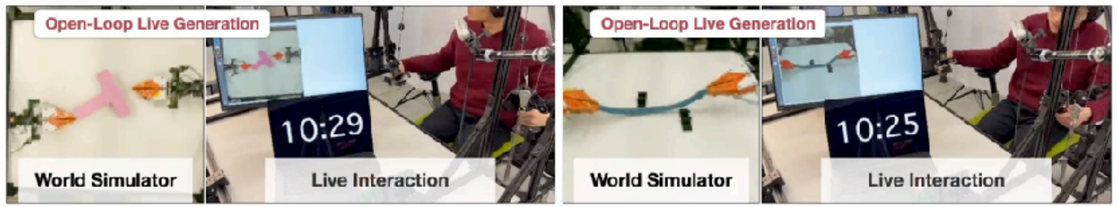
Világ modellek — NVIDIA DreamZero



https://dreamzero0.github.io/

Videó Generálás

Világ modellek — Interaktív szimuláció



Interactive World Simulator

for Robot Policy Training and Evaluation

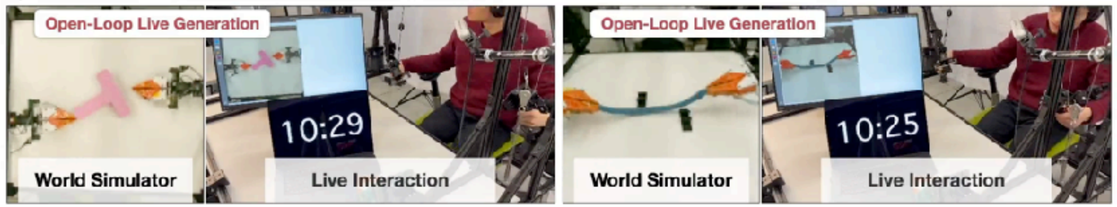
Yixuan Wang Rhythm Syed Fangyu Wu Mengchao Zhang Aykut Onol
Jose Barreiros Hooshang Nayyeri Tony Dear Huan Zhang Yunzhu Li

COLUMBIA ENGINEERING TOYOTA RESEARCH INSTITUTE

https://www.yixuanwang.me/interactive_world_sim/

Videó Generálás

Világ modellek — Interaktív szimuláció



Interactive World Simulator

for Robot Policy Training and Evaluation

Yixuan Wang Rhythm Syed Fangyu Wu Mengchao Zhang Aykut Onol
Jose Barreiros Hooshang Nayyeri Tony Dear Huan Zhang Yunzhu Li

COLUMBIA ENGINEERING TOYOTA RESEARCH INSTITUTE

https://www.yixuanwang.me/interactive_world_sim/

Videó Generálás

Világ modellek — 3D modell + Videó?

https://marble.worldlabs.ai/

World Labs

https://research.nvidia.com/labs/sil/projects/lyra2/

NVIDIA

Videó Generálás

Világ modellek — 3D modell + Videó?

https://marble.worldlabs.ai/

World Labs

https://research.nvidia.com/labs/sil/projects/lyra2/

NVIDIA

Videó Generálás

Világ modellek — 3D modell + Videó?

https://marble.worldlabs.ai/

World Labs

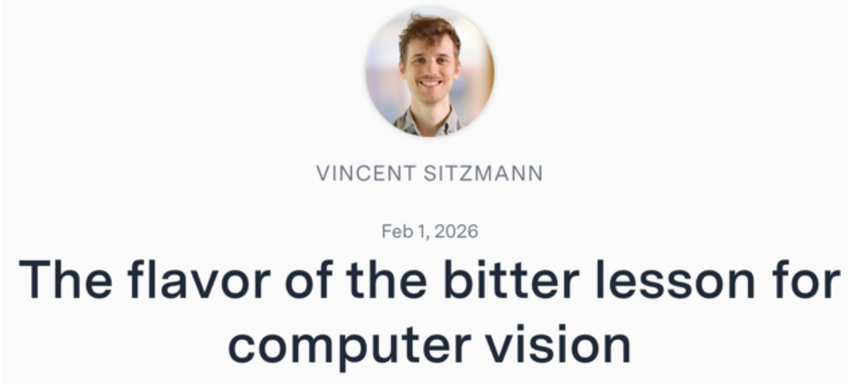
https://research.nvidia.com/labs/sil/projects/lyra2/

NVIDIA

Videó Generálás

Világ modellek — Quo vadis?

- Továbbra sem könnyű az idő- és térbeli konzisztenciát biztosítani videógenerálás során…
- Mi lenne, ha direkt frame-generálás helyett 3D geometriára alapoznánk (“grounding”)?
- Jó minőségű 3D (pláne összetartozó 2D-3D adatok) adatok sajnos nem nagyon állnak rendelkezésre…
  - Szintetikus adatokat persze könnyű generálni (pl. játékokból, szimulátorokból)…
- Nyers képek/videók ellenben szinte végtelen mennyiségben rendelkezésre állnak (“bitter lesson”?)



Érdekes vitaindító írás:
https://www.vincentsitzmann.com/blog/bitter_lesson_of_cv/

Következő előadás:

3D Geometria

- 3D geometria reprezentációi
- 3D adatok gyűjtése, mérése
- Konverzió reprezentációk között