Sora & Genie

2024.03.06
Lab Seminar
Park Kieun

HUMAN
CENTERED
COMPUTING
LABORATORY

OpenAI

[Sora]

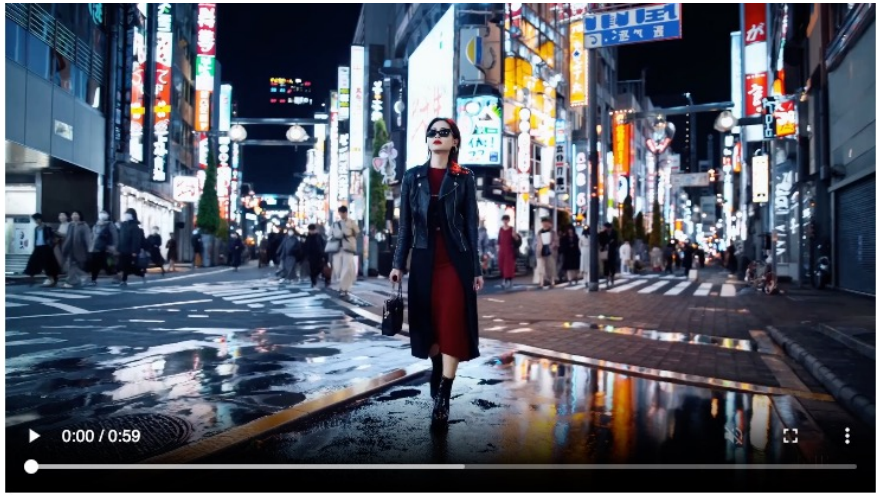
Research

Video generation models as world simulators

We explore large-scale training of generative models on video data. Specifically, we train text-conditional diffusion models jointly on videos and images of variable durations, resolutions and aspect ratios. We leverage a transformer architecture that operates on spacetime patches of video and image latent codes. Our largest model, Sora, is capable of generating a minute of high fidelity video. Our results suggest that scaling video generation models is a promising path towards building general purpose simulators of the physical world.

Sora Key-points

- text-to-video model
  + take an existing still image and generate a video from it
  + take an existing video and extend it or fill in missing frames
- Model
  - diffusion model
  - transformer architecture
- Model & implementation details 비공개
- Demo



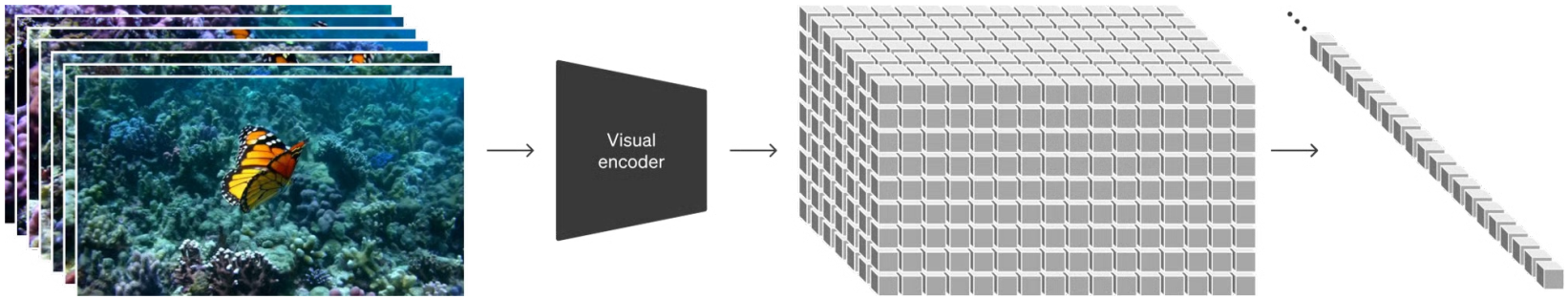
Prompt: A stylish woman walks down a Tokyo street filled with warm glowing neon and animated city sign age. She wears a black leather jacket, a long red dre ss, and black boots, and carries a black purse. She wears sunglasses and red lipstick. She walks confid ently and casually. The street is damp and reflective, creating a mirror effect of the colorful lights. Many pe destrians walk about.

https://openai.com/sora

Turning visual data into spacetime patches

- LLM의 token 처럼 spacetime patch 를 만들자

1. Video →(visual encoder)→ lower-dimensional latent space
2. Representation →(decomposing)→ spacetime patches



https://openai.com/research/video-generation-models-as-world-simulators

Diffusion transformer

- Given input noisy patches → trained to predict the original 'clean' pathches
- Diffusion transformer는 video model에도 아주 잘 적용됨
- 학습이 진행됨에 따라 결과가 좋아짐 (demo)

https://openai.com/research/video-generation-models-as-world-simulators

Language understanding

- Dall-E 3 의 recaption technique 사용
  1. train a highly descriptive captioner model
  2. 이를 이용하여 training set의 모든 비디오에 대한 캡션 생성
  3. 생성한 training set을 이용해 Training text-to-video generation systems
- GPT를 활용하여 사용자 프롬프트를 더 길고 자세한 캡션으로 전환
- Demo

an old man wearing blue jeans and a white t-shirt taking a pleasant stroll in Johannesburg, South Africa during a beautiful sunset

Additional capabilities

- Variable durations, resolutions, aspect ratios
  - 과거 : 256x256p 4초 비디오
  - Sora : 1920x1080p (wide screen) ~ 1080x1920p (vertical) 사이의 모든 것 가능
- Image to video
- Extend video
- Video-to-video editing
- Demo

https://openai.com/research/video-generation-models-as-world-simulators

Additional capabilities

- Connecting videos
- Image generation
- Simple interaction
- Dynamic camera motion
- Simulating digital worlds
- Demo

https://openai.com/research/video-generation-models-as-world-simulators

Weakness

- 복잡한 장면의 물리적 상호능력 표현 부족
- 원인과 결과의 구체적 사례 이해 부족
- 공간적 방향성과 시간에 따른 정확한 추적능력 부족
- 예시 (Demo)
  - 쿠키를 먹었지만 먹고 난 후에 쿠키가 그대로 (영상 없음)
  - 물리적으로 비현실한 움직임 생성 (트레드밀을 거꾸로 타기)
  - 동물, 사람, 사물 등이 갑자기 나타남
  - 물리 모델링 부정확 및 부자연스러운 객체 변형이 보임
  - 이상한 물리적 상호작용 초래
  - 여러 객체 및 캐릭터간 복잡한 상호작용 시뮬레이션 부족

https://openai.com/sora

반응

- 한 회사가 너무 많은 권력을 가지는 것에 대한 두려움
- 이번에는 매우 현실적으로 느껴진다고 함
- 이미지/비디오 측면을 넘어서 물리학과 객체 간의 관계에 대한 이해를 보여주는 모델의 중요성을 강조
- 업계 많은 사람들이 AI 도구의 발전으로 인해 직업에 대한 두려움을 느낌
- 현재 공개된 모델들을 훨씬 뛰어넘는 성과
- Gemini 1.5 모델이 구식이 되었다
- 비디오의 품질에 대해서는 믿을 수 없을 정도로 인상적
- 비디오 생성 방식에 대한 기술적인 질문을 던지며, 모델이 장면의 기하학적 구조와 카메라를 분리하는 방식에 대해 궁금해함
- 문화적 변화를 가져올 것을 기대

https://discuss.pytorch.kr/t/gn-openai-sora-ai/3519

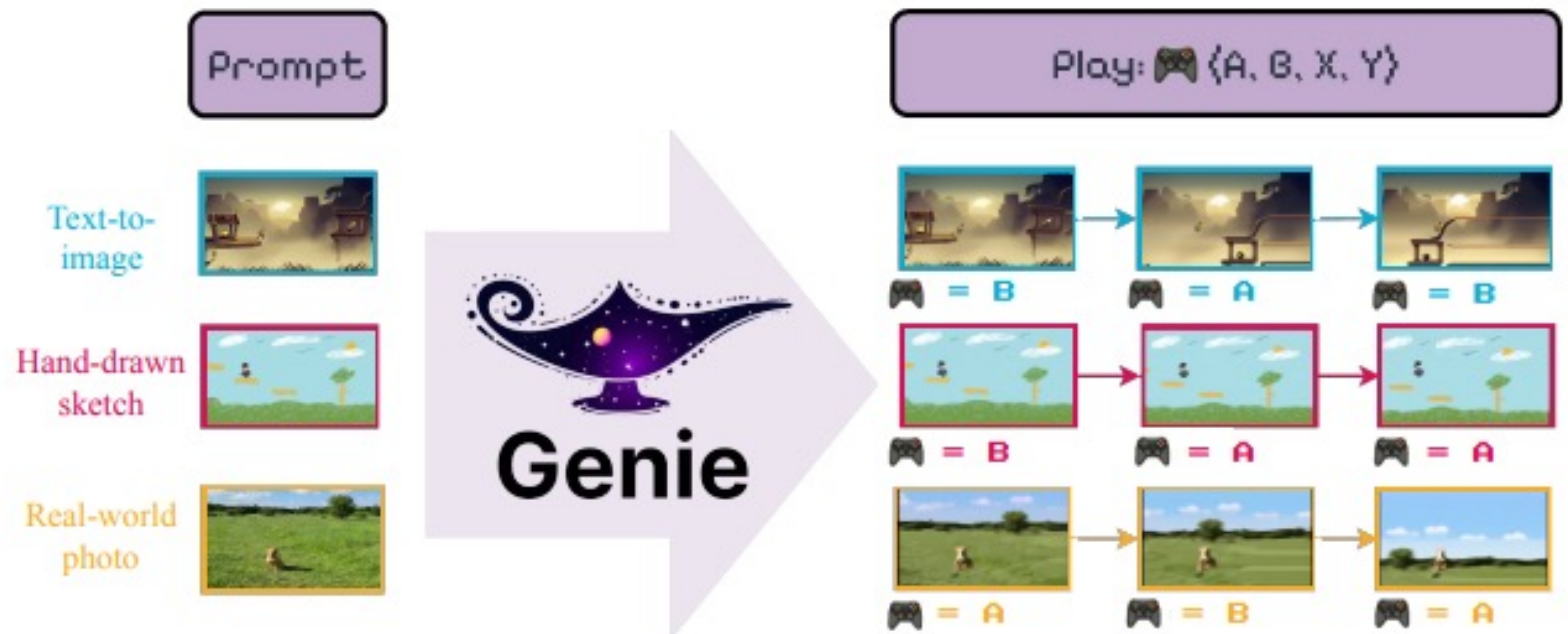
arXiv:2402.15391v1 [cs.LG] 23 Feb 2024

Google DeepMind 2024-2-26

Genie: Generative Interactive Environments

Jake Bruce[*,1], Michael Dennis[*,1], Ashley Edwards[*,1], Jack Parker-Holder[*,1], Yuge (Jimmy) Shi[*,1], Edward Hughes[1], Matthew Lai[1], Aditi Mavalankar[1], Richie Steigerwald[1], Chris Apps[1], Yusuf Aytar[1], Sarah Bechtle[1], Feryal Behbahani[1], Stephanie Chan[1], Nicolas Heess[1], Lucy Gonzalez[1], Simon Osindero[1], Sherjil Ozair[1], Scott Reed[1], Jingwei Zhang[1], Konrad Zolna[1], Jeff Clune[1,2], Nando de Freitas[1], Satinder Singh[1] and Tim Rocktäschel[*,1]

[*]Equal contributions, [1]Google DeepMind, [2]University of British Columbia



https://arxiv.org/pdf/2402.15391.pdf

간단 소개

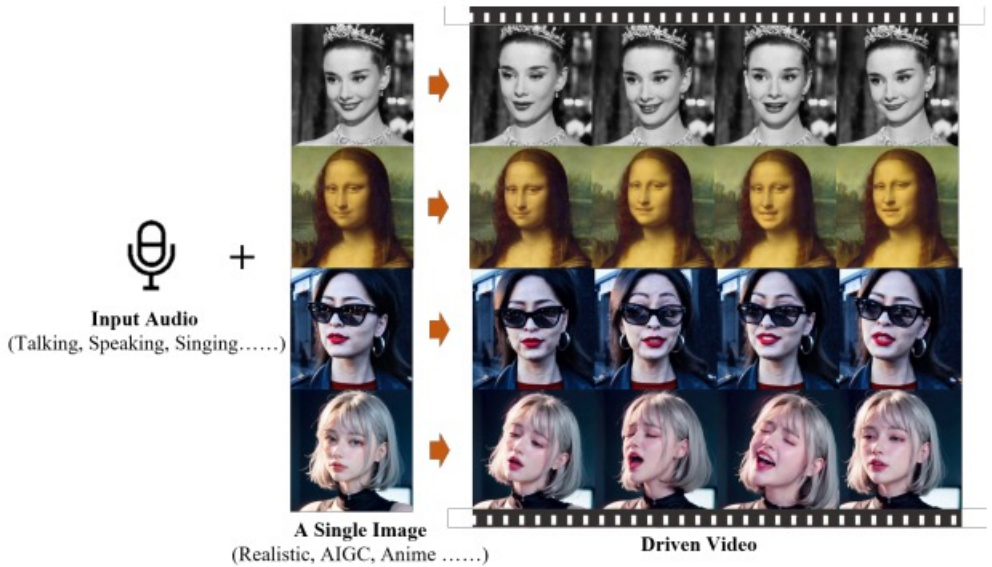
- **Genie: Generative Interactive Environments** (Jake Bruce, et al., 24.02.23)
- Google DeepMind https://sites.google.com/view/genie-2024/home
- Unsupervised training, unlabelled internet video
- Text, synthetic image, photo, sketch 로부터 생성 가능
- 구성
  - spatiotemporal video tokenizer
  - autoregressive dynamics model
  - simple and scalable latent action model
- Demo

arXiv:2402.17485v1 [cs.CV] 27 Feb 2024

EMO: Emote Portrait Alive - Generating Expressive Portrait Videos with Audio2Video Diffusion Model under Weak Conditions

Linrui Tian, Qi Wang, Bang Zhang, and Liefeng Bo

Institute for Intelligent Computing, Alibaba Group
{tianlinrui.tlr, wilson.wq, zhangbang.zb, liefeng.bo}@alibaba-inc.com
https://humanaigc.github.io/emote-portrait-alive/



https://arxiv.org/pdf/2402.17485.pdf

간단 소개

- **EMO: Emote Portrait Alive - Generating Expressive Portrait Videos with Audio2Video Diffusion Model under Weak Conditions** (Linrui Tian, et al., 24.02.27)
- Alibaba https://humanaigc.github.io/emote-portrait-alive/?
- Facial movement 와 audio cue 간의 nuanced relationship 에 초점
- Direct audio-to-video synthesis approach (3D model 이나 facial landmark 불필요)
- Speaking + Singing 모두 가능
- Demo

QA

HUMAN
CENTERED
COMPUTING
LABORATORY