

무은재학부 이수겸
IDS 장승현

Intro



1.물리 법칙과 현실성의 발전
->현실세계 물리 법칙에 대한 이해

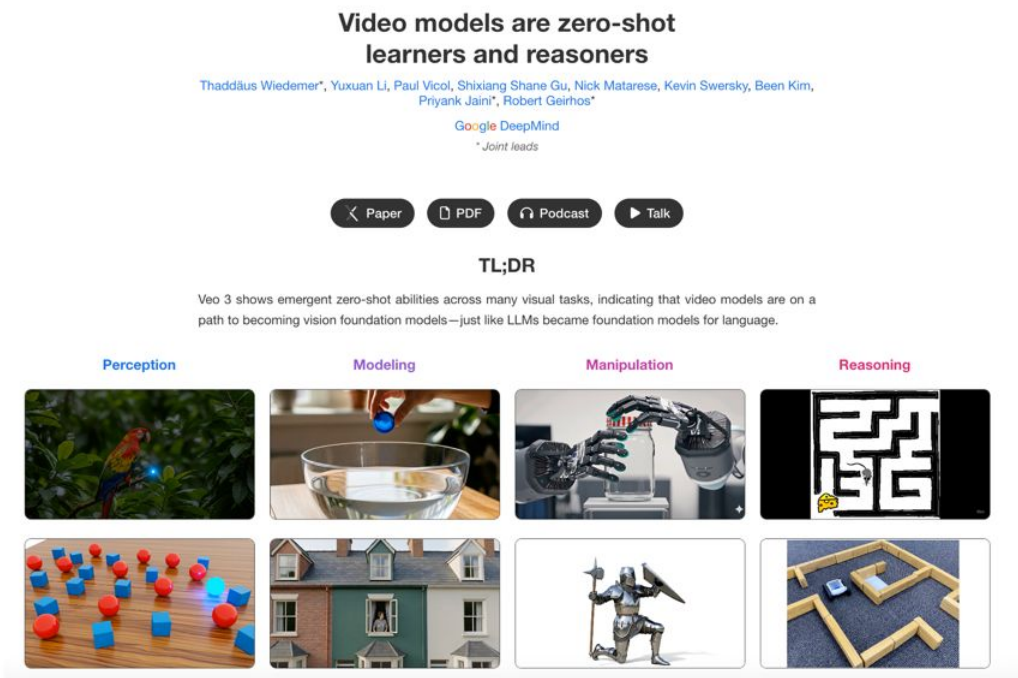
2.통합 오디오 생성
->영상과 소리 동시 생성

3.실세계 객체 주입 및 개인화
->텍스트나 이미지를 기반 영상생성

4.향상된 조작성
->복잡한 프롬프트 이해 향상

https://openai.com/ko-KR/index/sora-2/

Intro



https://video-zero-shot.github.io/

범용적인 제로샷(Zero-Shot) 능력의 발현

1.지각(Perceiving)
->객체 분할, 엣지 감지

2.모델링 (Modeling)
->Understanding physical properties, Simulating tool use

3.조작 (Manipulation)
->이미지/장면 편집, 3D 및 포즈 조작

4.추론 (Reasoning)
->미로 찾기, 논리 및 문제 해결

=Emergent Capabilities의 등장으로 범용 비전 모델의 foundation model의 등장

Intro

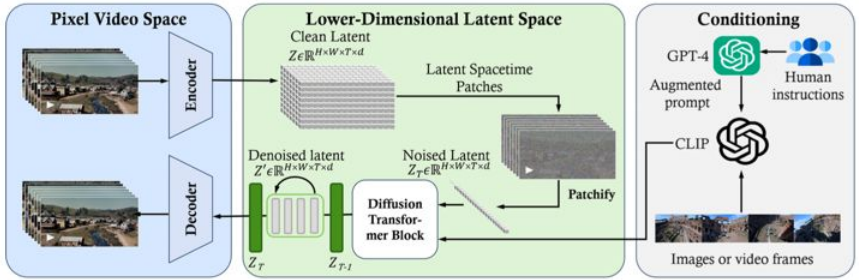
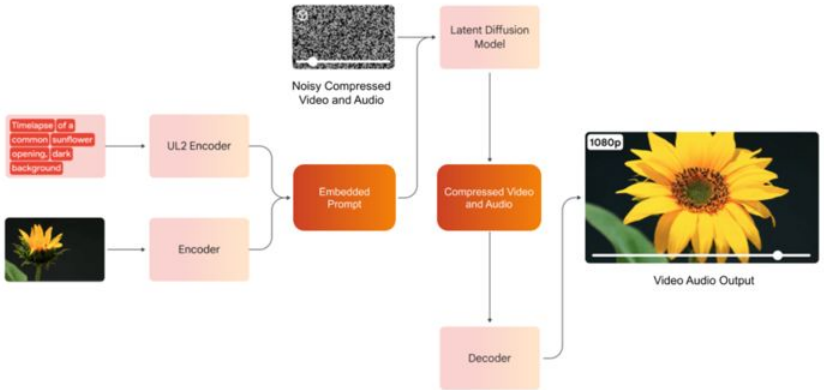


Figure 4: **Reverse Engineering**: Overview of `Sora` framework



Intro

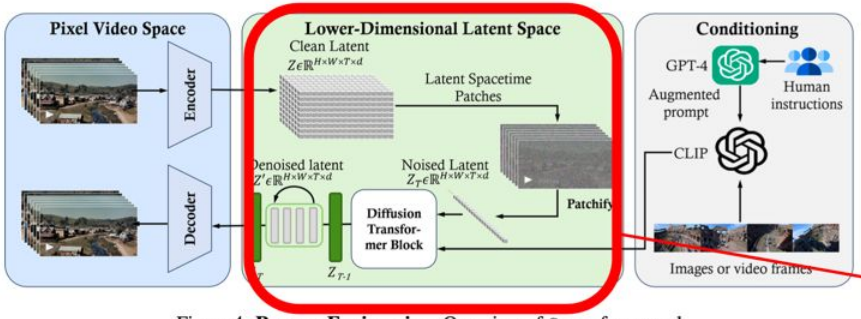
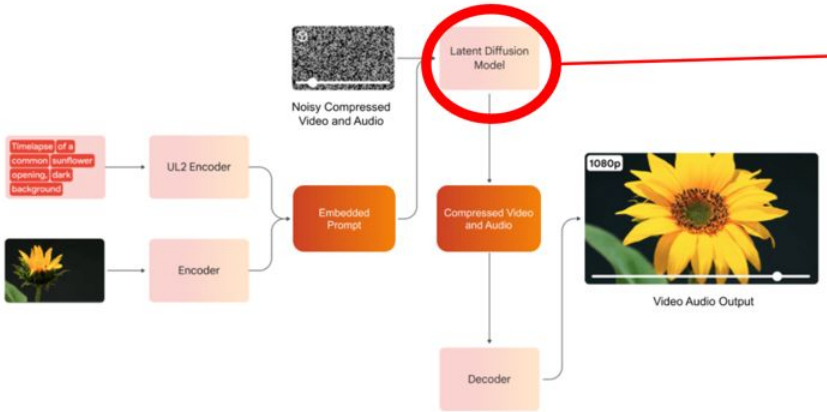


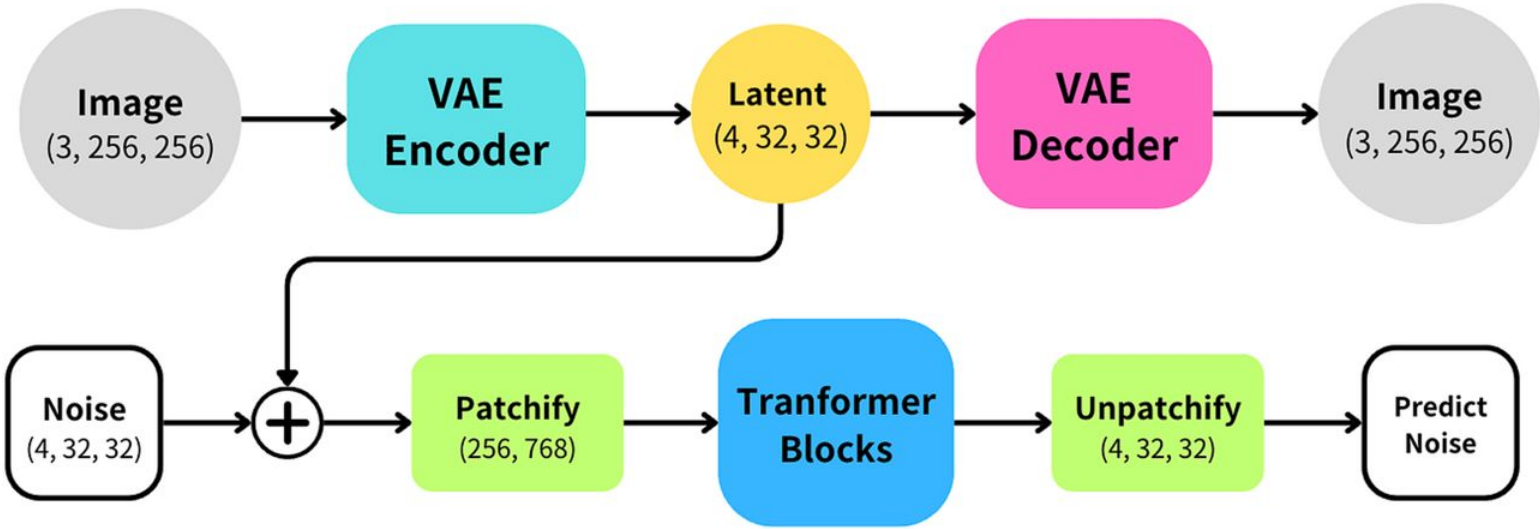
Figure 4: **Reverse Engineering**: Overview of `Sora` framework



Latent Diffusion Transformer

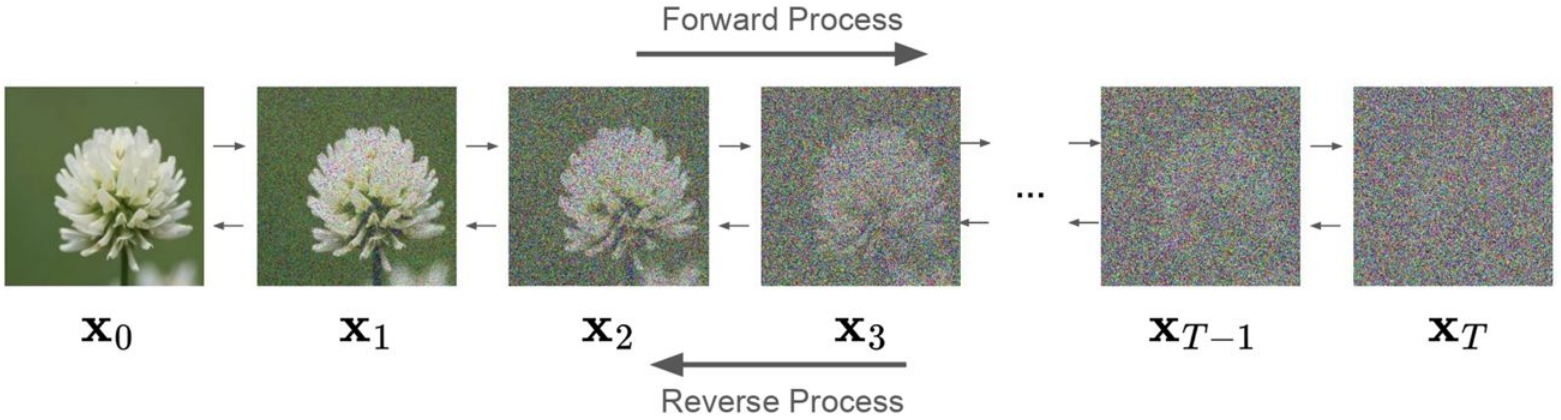
Background

Latent diffusion transformer



Latent diffusion transformer

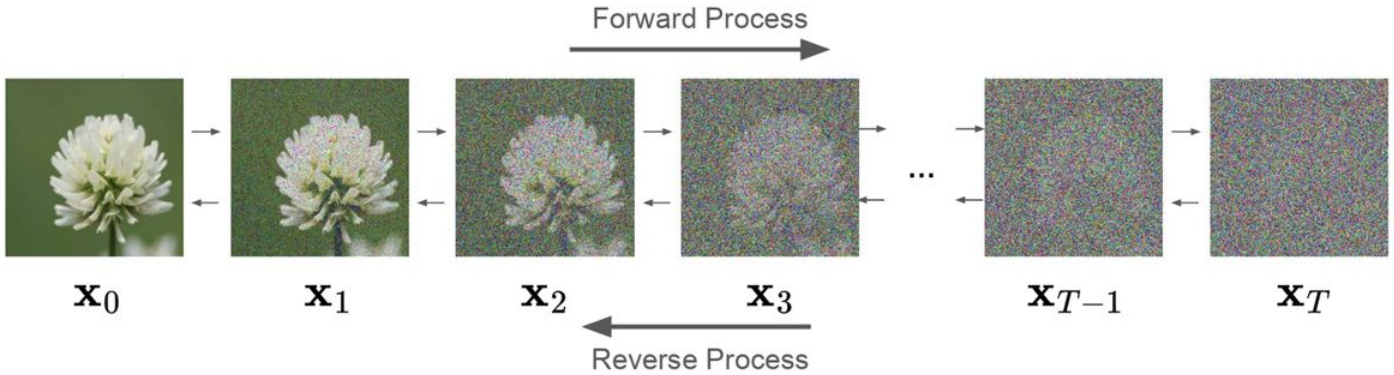
-Diffusion model



Forward Process: adding gausian noise
Reverse Process: denoising process

Latent diffusion transformer

-Diffusion model



Distribution of the noised images — Output — Mean μ_t — Variance Σ_t

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t;\ \sqrt{1-\beta_t}x_{t-1},\ \beta_t I)$$

Notations:

- t : time step (from 0 to T)
- x_0 : a data sampled from the real data distrition $q(x)$ (i.e. $x_0 \sim q(x)$)
- β_t : variance schedule ($0 \le \beta_t \le 1$, and β_0 = small number, β_T = large number)
- I : identity matrix

Since $q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = N\left(\boldsymbol{x}_t|\sqrt{1-\beta_t}\boldsymbol{x}_{t-1}, \beta_t \boldsymbol{I}\right)$

Then $\boldsymbol{x}_t = \sqrt{1-\beta_t}\boldsymbol{x}_{t-1} + \sqrt{\beta_t}\boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon} \sim N(\boldsymbol{\epsilon}|\boldsymbol{0}, \boldsymbol{I})$

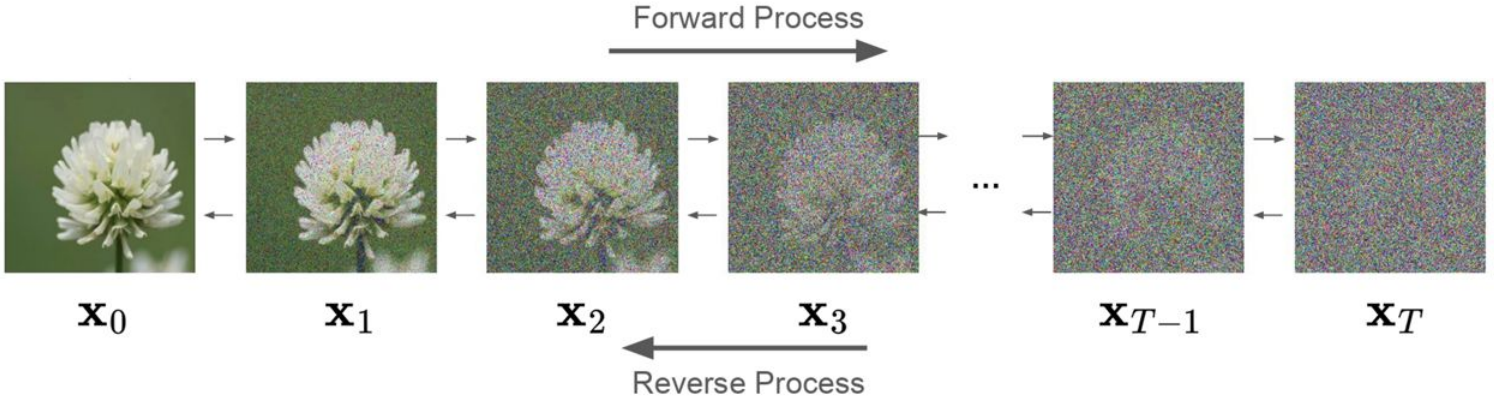
$$\boldsymbol{x}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}$$

where $\boldsymbol{\epsilon} \sim N(\boldsymbol{\epsilon}|\boldsymbol{0}, \boldsymbol{1})$

$\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$ and $\alpha_i = 1 - \beta_i$

Latent diffusion transformer

-Diffusion model



$$\log p(\boldsymbol{x}) \geq \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\right]$$

$$= \underbrace{\mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)\right]}_{\text{reconstruction term}} - \underbrace{D_{\mathrm{KL}}(q(\boldsymbol{x}_T|\boldsymbol{x}_0) \parallel p(\boldsymbol{x}_T))}_{\text{prior matching term}} - \sum_{t=2}^{T} \underbrace{\mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}\left[D_{\mathrm{KL}}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) \parallel p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t))\right]}_{\text{denoising matching term}}$$

Latent diffusion transformer

-Diffusion model

$$\log p(\boldsymbol{x}) \geq \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\right]$$

$$= \underbrace{\mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)\right]}_{\text{reconstruction term}} - \underbrace{D_{\mathrm{KL}}(q(\boldsymbol{x}_T|\boldsymbol{x}_0) \parallel p(\boldsymbol{x}_T))}_{\text{prior matching term}} - \sum_{t=2}^{T} \underbrace{\mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}\left[D_{\mathrm{KL}}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) \parallel p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t))\right]}_{\text{denoising matching term}}$$

Reconstruction term : step 1과 0의 차이

Prior matching term: step T와 가우시안 부포의 차이

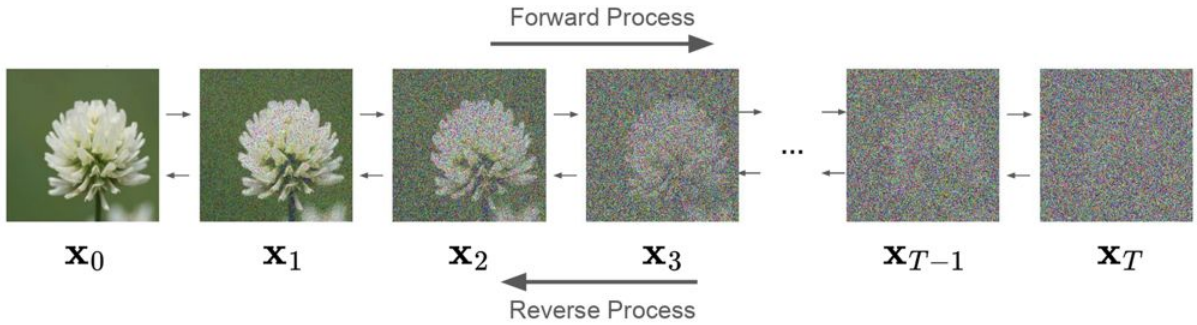
->둘다 상수취급 가능

몇가지 조건 하에 time step t에서 노이즈와 전체 노이즈차의 MSE term과 같은 최적해로 수렴함

$$L_{\text{simple}} = \mathbb{E}\left[\|\epsilon_\theta(x_t, t) - \epsilon\|^2\right]$$

Latent diffusion transformer

-Diffusion model



1. 훈련 과정 (Training Phase): 노이즈 예측 학습

학습 단계의 목표는 모델이 **임의의 오염된 상태(t)에서 노이즈의 정체(ϵ)를 정확히 예측하는 것**입니다. 이 과정은 효율성을 위해 **단 한 번의 예측**으로 끝납니다.

1. **시점 및 데이터 선택:** $t = 1$부터 $t = 1000$ 사이의 임의의 시간(t)을 무작위로 선택합니다.
2. **문제 출제:** 원본 비디오(x_0)를 **VAE 인코더**로 잠재 공간(z_0)에 압축한 뒤, 선택된 시간(t)에 해당하는 **정답 노이즈(ϵ)**를 섞어 **오염된 잠재 데이터(z_t)**를 만듭니다.
3. **모델 예측:** `Diffusion Transformer (DiT)` 는 오염된 데이터(z_t)와 조건 정보(클래스 토큰 등)를 **AdaLN 블록**으로 주입받아 **예측 노이즈(ϵ_θ)**를 출력합니다.
4. **피드백 및 훈련:** **[정답 노이즈(ϵ)]**와 **[예측 노이즈(ϵ_θ)]**의 차이(Loss)를 계산하여 모델의 파라미터를 수정합니다. 학습할 때는 $t - 1$로 가는 복잡한 과정을 반복하지 않고, **랜덤한 시점(t)에서 노이즈만 잘 맞추도록** 훈련에 집중합니다.

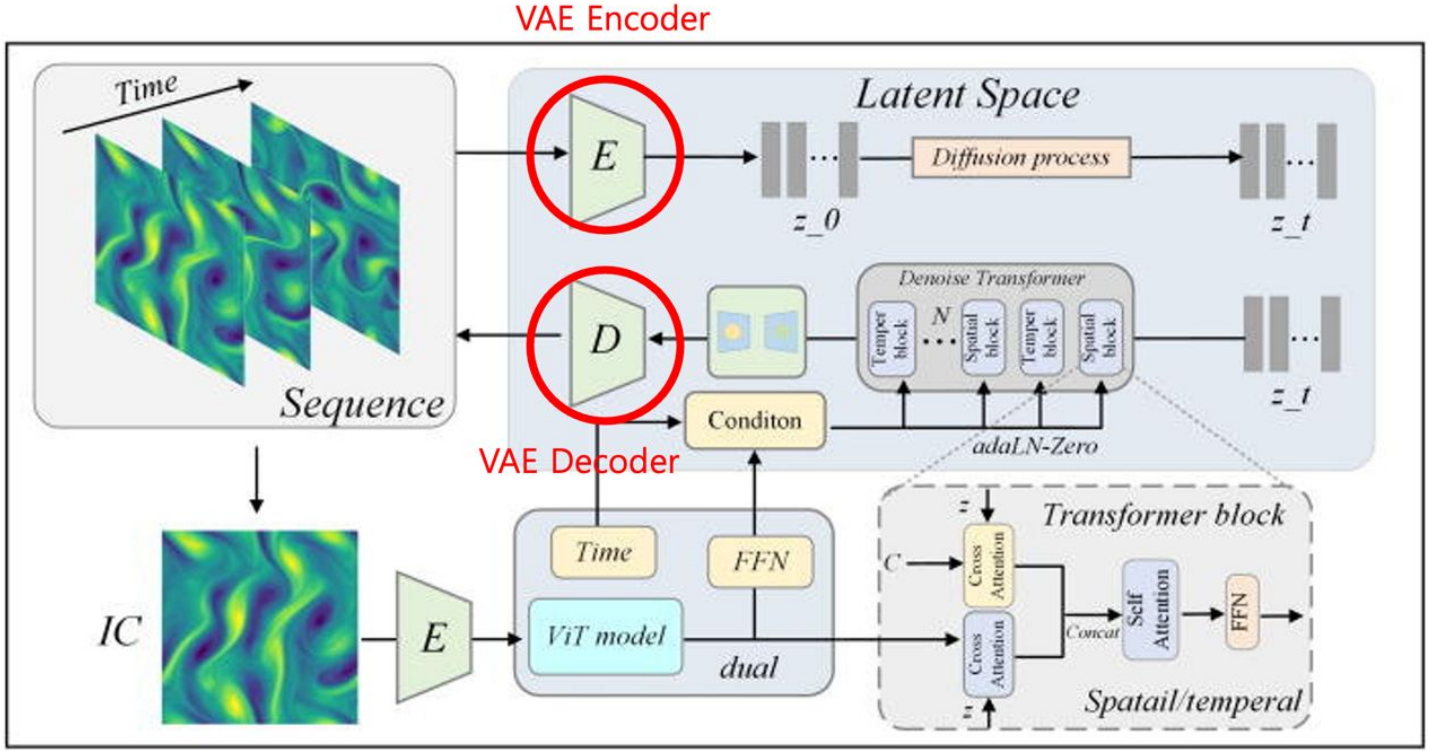
2. 생성 과정 (Inference Phase): 프레임의 사슬로 완성

생성 단계의 목표는 **학습된 모델의 예측을 활용하여 랜덤 노이즈에서 점진적으로 깨끗한 비디오를 복원**하는 것입니다.

1. **시작:** 가장 노이즈가 많은 상태(T)의 **랜덤 노이즈(z_T)**에서 출발합니다.
2. **방향성 설정:** **클래스 토큰이나 텍스트 프롬프트**를 AdaLN 블록에 주입하여, 디노이징이 원하는 주제나 스타일 방향으로 진행되도록 유도합니다.
3. **반복적 노이즈 제거:**
   - **예측:** 현재 시점(t)에서 DiT는 **전체 노이즈(ϵ_θ)**를 예측합니다.
   - **일부 제거:** 예측된 전체 노이즈를 한 번에 다 빼지 않고, **스케줄러가 정한 안전한 비율만큼만** 걷어내어 다음 단계인 **z_{t-1}**로 이동합니다.
   - **반복:** $t - 1$ 상태에서 다시 노이즈를 예측하고 제거하는 과정을 $T \to 0$까지 반복합니다. 이 **연속적인 과정**을 **프레임의 사슬(Chain-of-Frames)**이라 볼 수 있습니다.
4. **최종 복원:** $t = 0$에 도달하여 깨끗해진 잠재 데이터(z_0)를 **VAE 디코더**에 통과시켜 최종 비디오(픽셀)를 완성합니다.

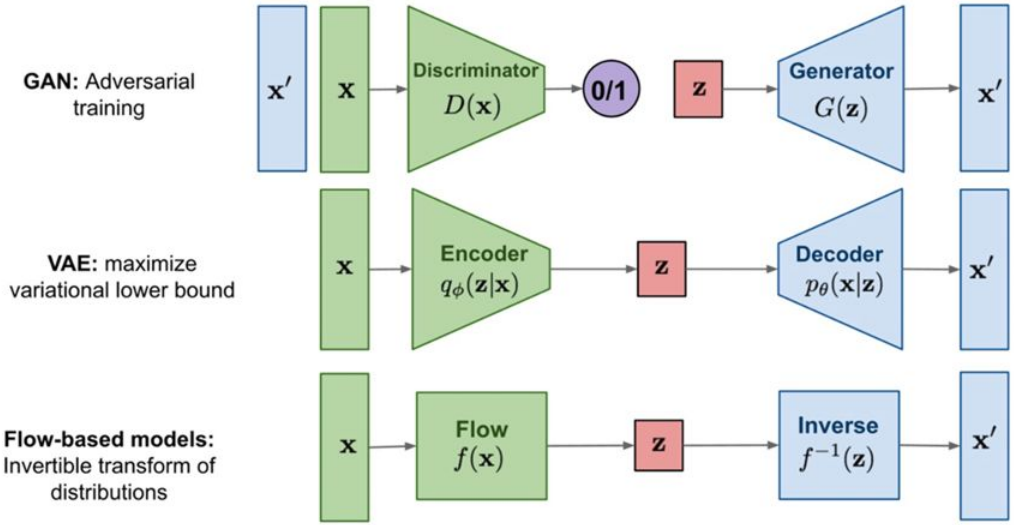
Latent diffusion transformer

-Latent Diffusion model



Latent diffusion transformer

-Latent Diffusion model



From https://lilianweng.github.io/

-LDM에 사용되는 LDM Encoder는 영상 생성 모델인 VAE 모델의 Encoder와 유사

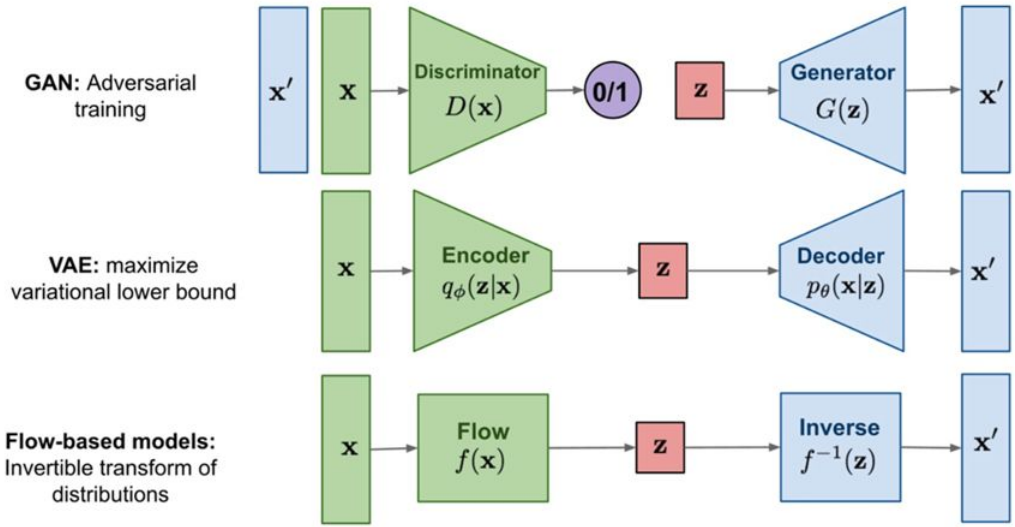
VAE Encoder 모델은latent space로 이미지를 압축 -> 디코더로 이미지 재구성 방식

CNN 기반 이미지 feature을 잘반영해서 압축하고 연속적이고 매끄럽게 압축가능

고품질 이미지 생성 난항

Latent diffusion transformer

-Latent Diffusion model



From https://lilianweng.github.io/

-VAE 모델의 인코더와 유사

Transformer 기반 Diffusion에서는 token 수에 대해 연산량이 $O(N^2)$이므로 VAE 인코더를 통해 연산량 감소

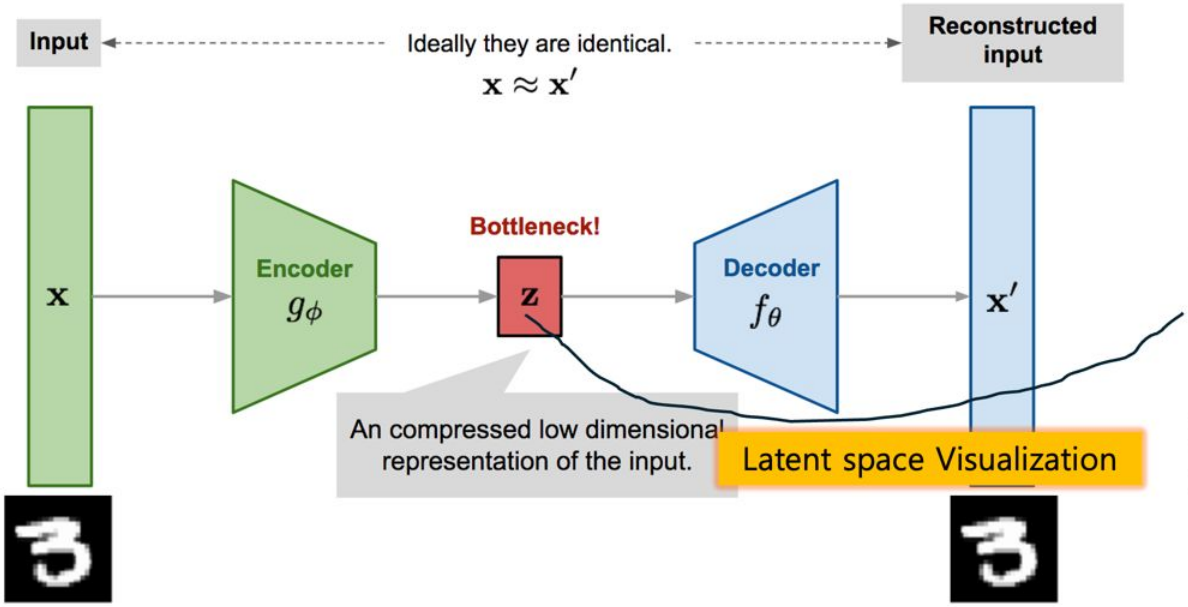
-semantic information만 남기고 불필요한 high-frequency detail을 제거

Example)
사진:
256×256×3 -> 32×32×4

동영상:
[16, 256, 256, 3] -> [16, 32, 32, 4]
[프레임 수 (T), 높이 (H), 너비 (W), 채널 (C)]

Latent diffusion transformer

-Latent Diffusion model



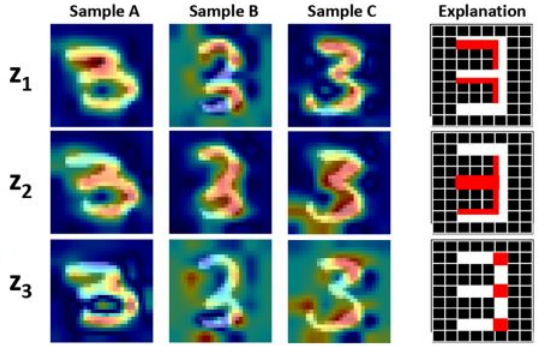


Figure 1. We propose to visually explain variational autoencoders. Each element in the latent vector (here $z_1 - z_3$) can be explained separately with our attention maps, visualizing consistent explanations across different samples.

Attention map을 통해 재구성한 사진

VAE Encoder는 Latent space에서는 특징과 feature을 잘 살려 압축함

Latent diffusion transformer -Latent Diffusion transformer

Published in Transactions on Machine Learning Research (03/2025)

Latte: Latent Diffusion Transformer for Video Generation

Xin Ma[1†], **Yaohui Wang**[2‡], **Xinyuan Chen**[2], **Gengyu Jia**[3],
Ziwei Liu[4], **Yuan-Fang Li**[1], **Cunjian Chen**[1] **Yu Qiao**[2]
[1]Department of Data Science & AI, Faculty of Information Technology, Monash University [2]Shanghai AI Laboratory
[3]Nanjing University of Posts and Telecommunications [4]S-Lab, Nanyang Technological University

Reviewed on OpenReview: https://openreview.net/forum?id=ntGPYNUF3t

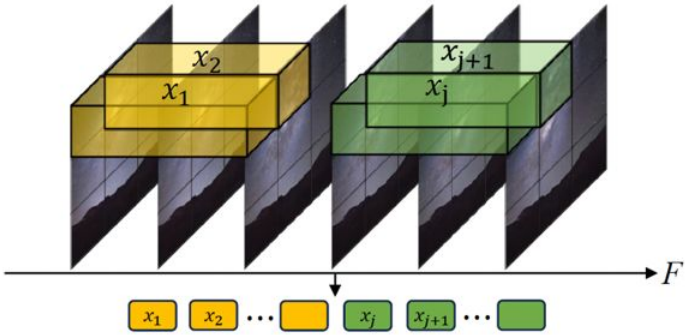
-Latent Diffusion transformer의 아키텍쳐로는 위 논문을 참고함

-허나 위 논문은 Sora나 veo직접적 연곤성은 없으므로 위논문에서 제시하는 핵심 아키텍쳐 위주로 참고함

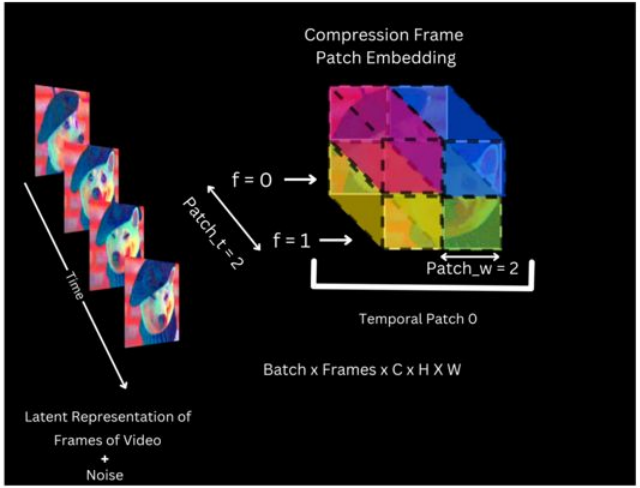
Latent diffusion transformer

-Latent Diffusion transformer (video)

-compression frame patch embedding



(b) compression frame patch embedding

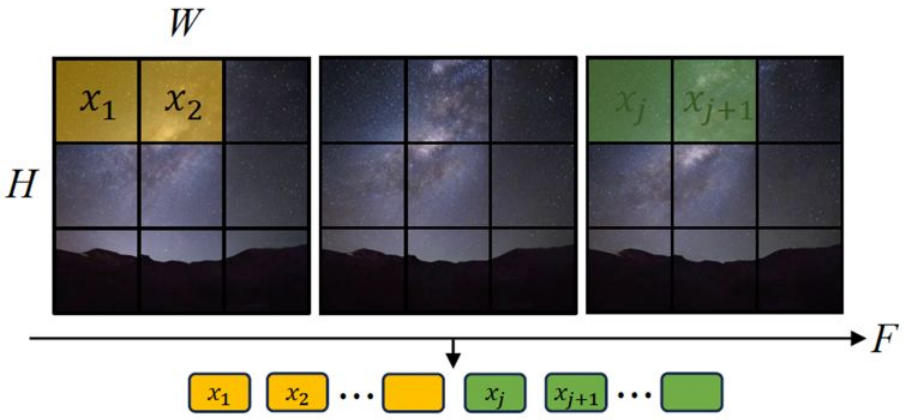


Transformer는 시퀀스 기반 모델이므로 patch embedding이 요구됨

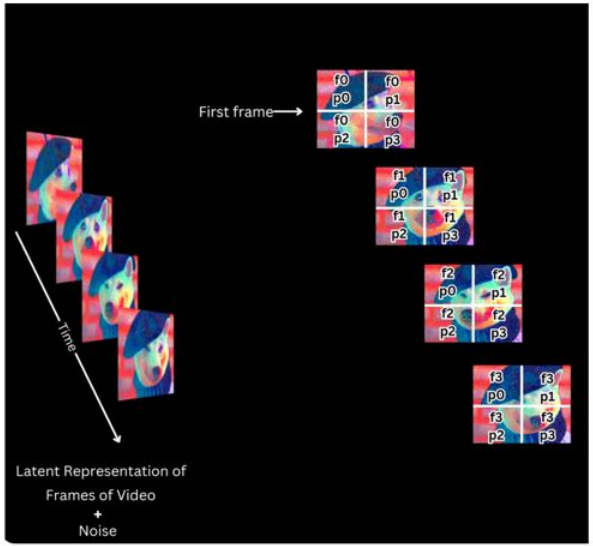
Latent diffusion transformer

-Latent Diffusion transformer (video)

-Uniform patch embedding



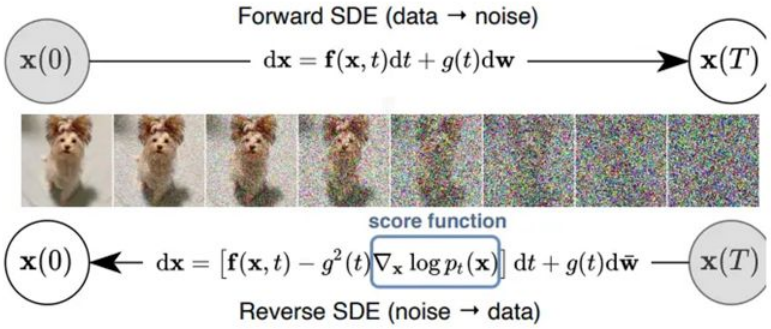
(a) uniform frame patch embedding



다양한 패치 embedding 기법이 존재함

Latent diffusion transformer -Latent Diffusion transformer

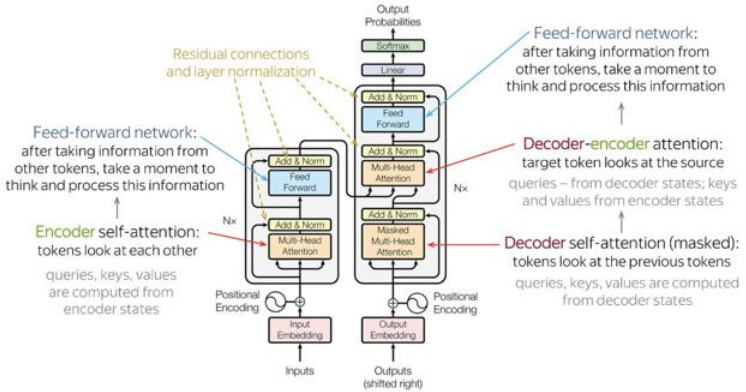
Recap of diffusion model

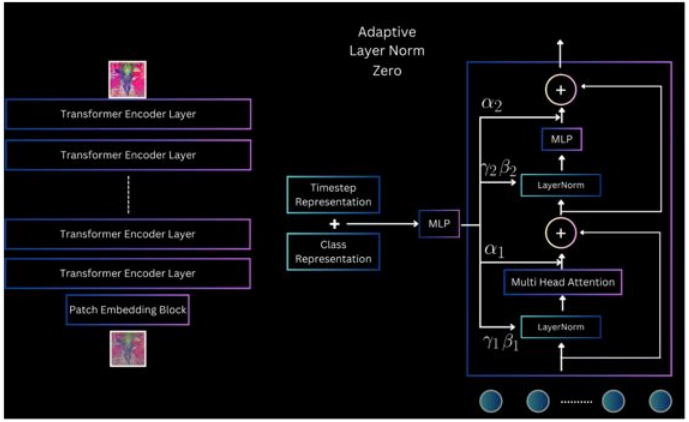


1. 훈련 과정 (Training Phase): 노이즈 예측 학습

학습 단계의 목표는 모델이 **임의의 오염된 상태(t)에서 노이즈의 정체(ϵ)를 정확히 예측하는 것**입니다. 이 과정은 효율성을 위해 **단 한 번의 예측**으로 끝납니다.

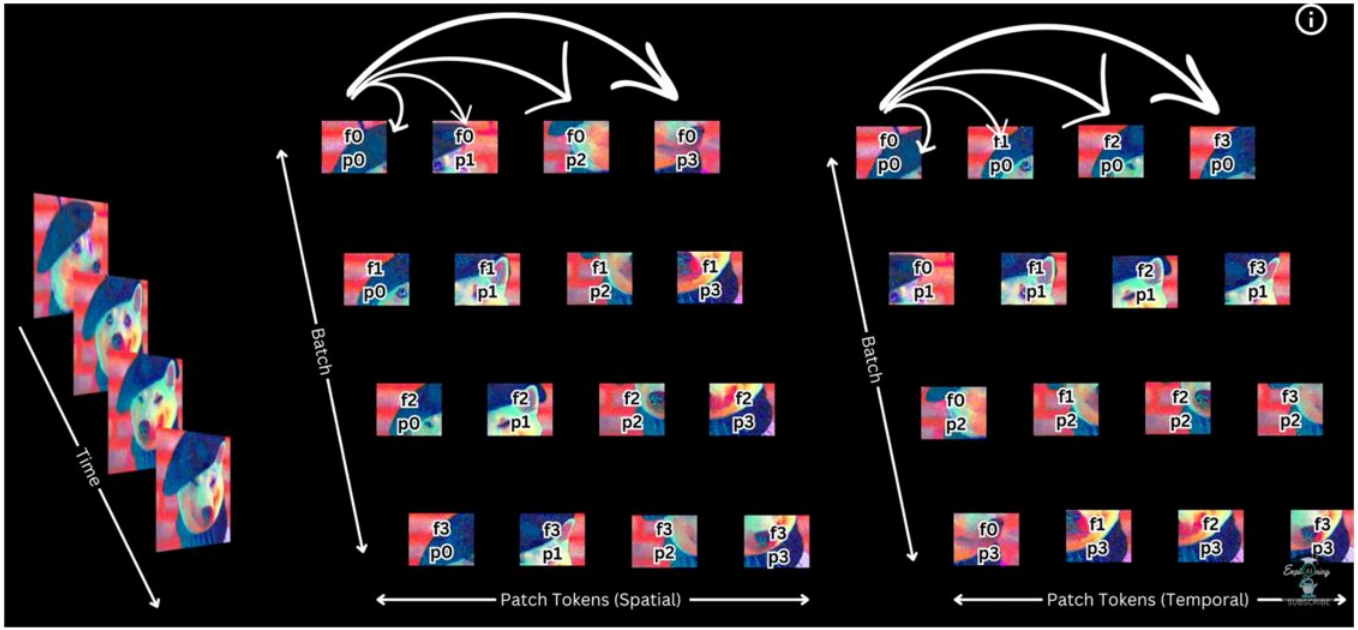
1. **시점 및 데이터 선택:** $t = 1$부터 $t = 1000$ 사이의 임의의 시간(t)을 무작위로 선택합니다.
2. **문제 출제:** 원본 비디오(x_0)를 **VAE 인코더**로 잠재 공간(z_0)에 압축한 뒤, 선택된 시간(t)에 해당하는 **정답 노이즈(ϵ)**를 섞어 **오염된 잠재 데이터(z_t)**를 만듭니다.
3. **모델 예측:** `Diffusion Transformer (DiT)` 는 오염된 데이터(z_t)와 조건 정보(클래스 토큰 등)를 **AdaLN 블록**으로 주입받아 **예측 노이즈(ϵ_θ)**를 출력합니다.
4. **피드백 및 훈련:** **[정답 노이즈(ϵ)]**와 **[예측 노이즈(ϵ_θ)]**의 차이(Loss)를 계산하여 모델의 파라미터를 수정합니다. 학습할 때는 $t - 1$로 가는 복잡한 과정을 반복하지 않고, **랜덤한 시점(t)에서 노이즈만 잘 맞추도록** 훈련에 집중합니다.





Latent diffusion transformer

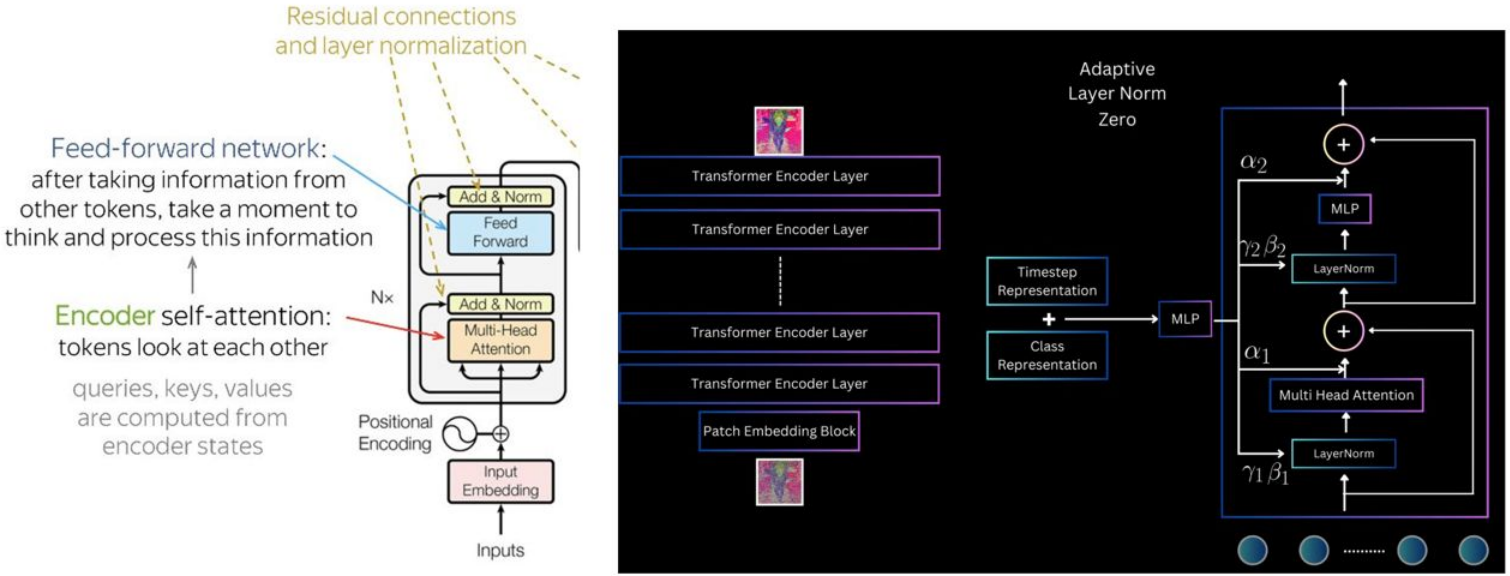
-Latent Diffusion transformer



비디오이므로 시간축과 공간축에 대해서 spatial & temporal attention 진행해야함

Latent diffusion transformer

-Latent Diffusion transformer



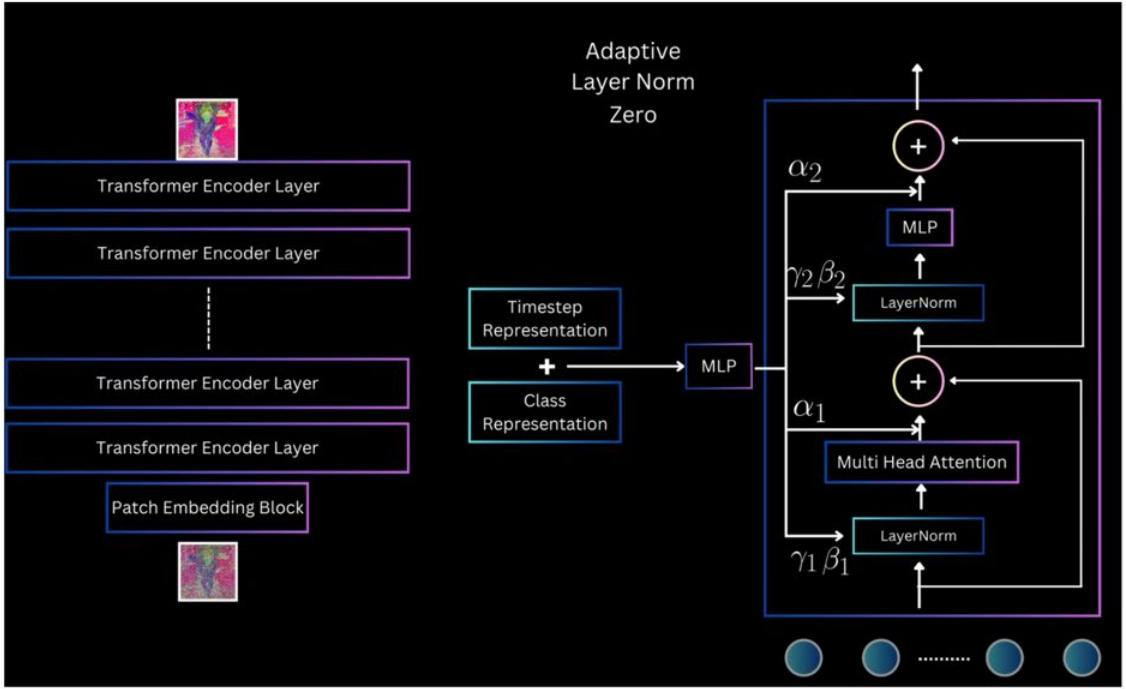
기존 이미지 Dit와 다르게 video model에서는 시간(frame)도 같이 계산해줘야 함

class representation 에 나중에 프롬프트 임베딩을 통한 text representation으로 바꾸면 text to video generation 구현가능

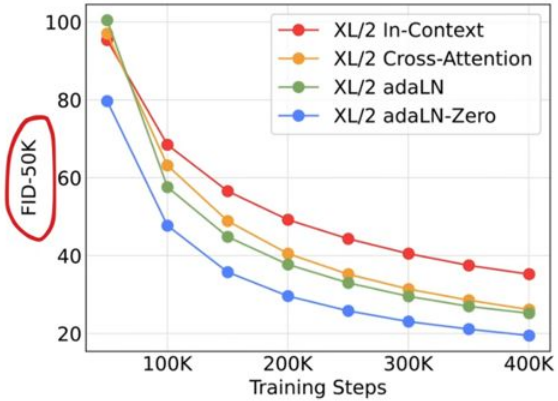
Latent diffusion transformer

-Latent Diffusion transformer (video)

-AdaLN(Adaptive Layer Norm zero in image transformer)



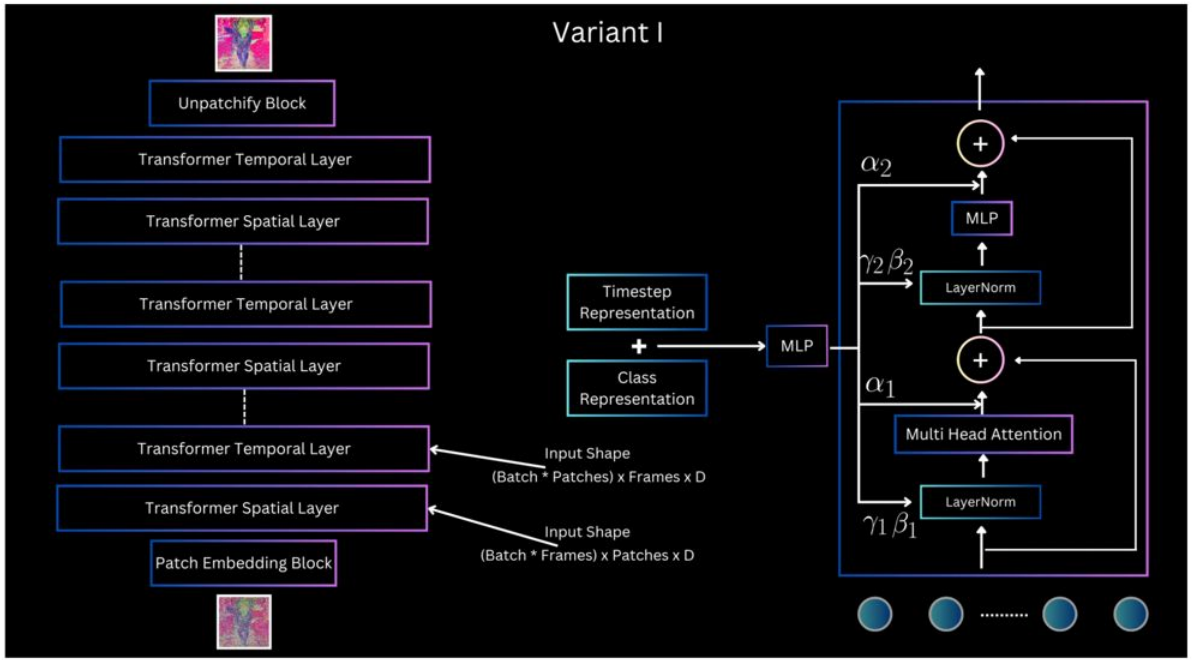
Class Representation과 Time step representation의 학습을 통해 성능이 많이 올라감을 알 수 있다.



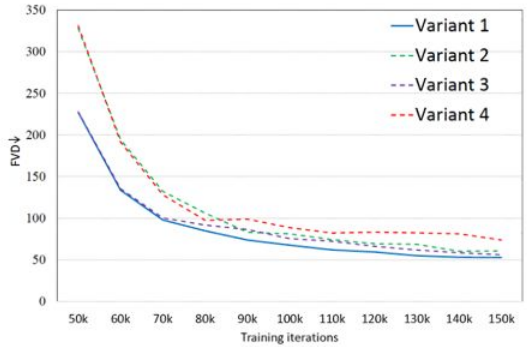
Latent diffusion transformer

-Latent Diffusion transformer (video)

-AdaLN(Adaptive Layer Norm zero in video transformer)



디퓨전 트랜스포머 모델에서도 Class Representation을 잘 학습시키면 성능이 많이 올라감을 알 수 있다.



(f) Model variants

Veo3 / Sora Report

Video Generation Architecture - Sora

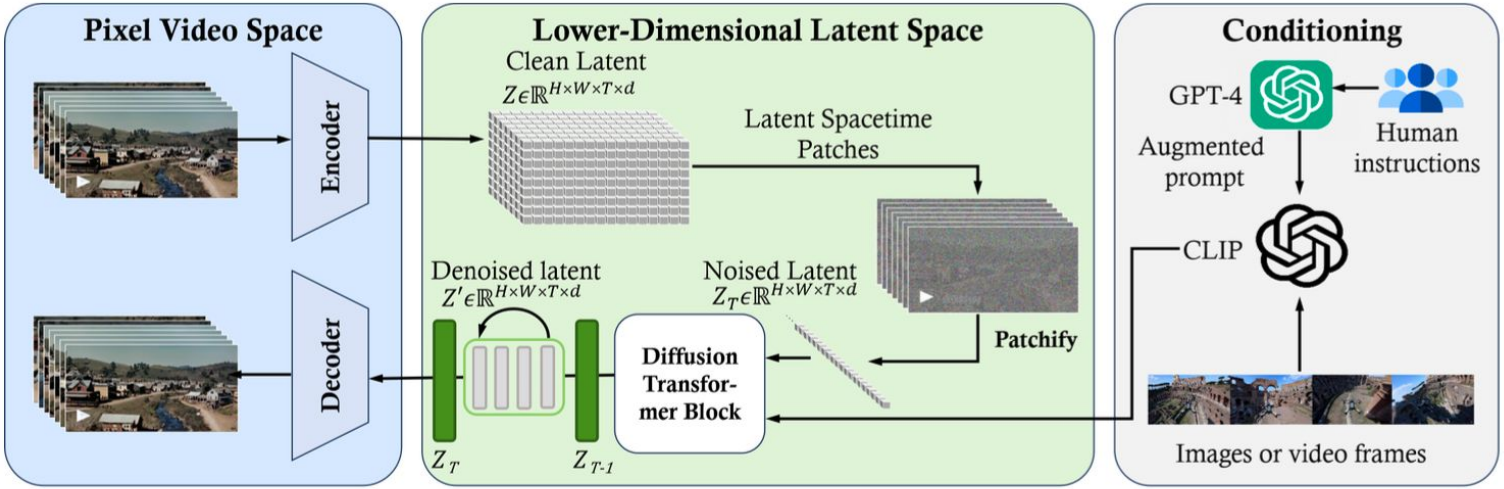
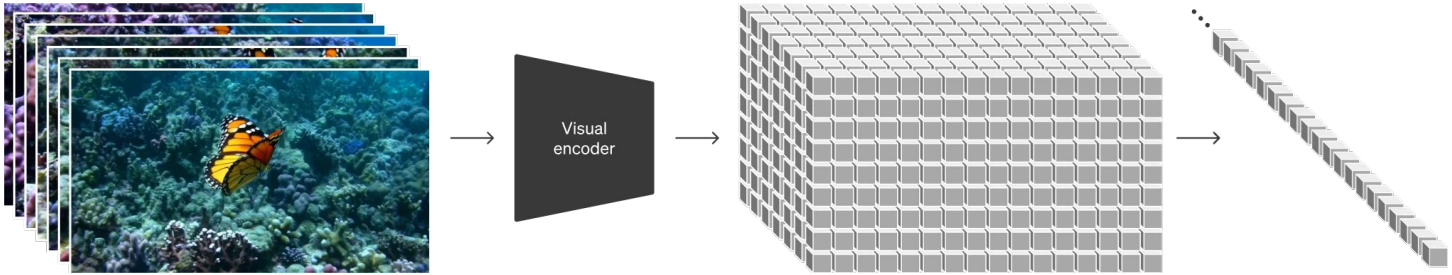


Figure 4: **Reverse Engineering**: Overview of `Sora` framework

https://openai.com/index/video-generation-models-as-world-simulators
https://j-qi.medium.com/openai-soras-technical-review-a8f85b44cb7f - **OpenAI Sora's Technical Review**

Characteristics of Sora

Turning visual data into patches - Video Compression Network



Sora는 raw pixel 영상을 직접 모델링하지 않고, Visual Encoder를 통해 시공간(latent) 표현으로 압축

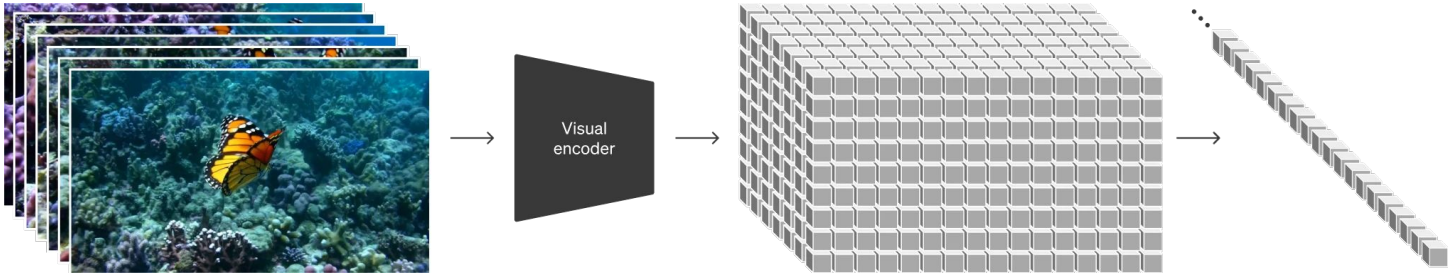
→ 효율적인 대규모 학습과 변형 가능한 영상 생성을 가능하게 함

video compression network(vision encoder + decoder) 를 훈련

→ latent representation은 이후에 **spacetime patches로 분할되어 Transformer token으로 사용**

Characteristics of Sora

Turning visual data into patches - Spacetime Latent patches



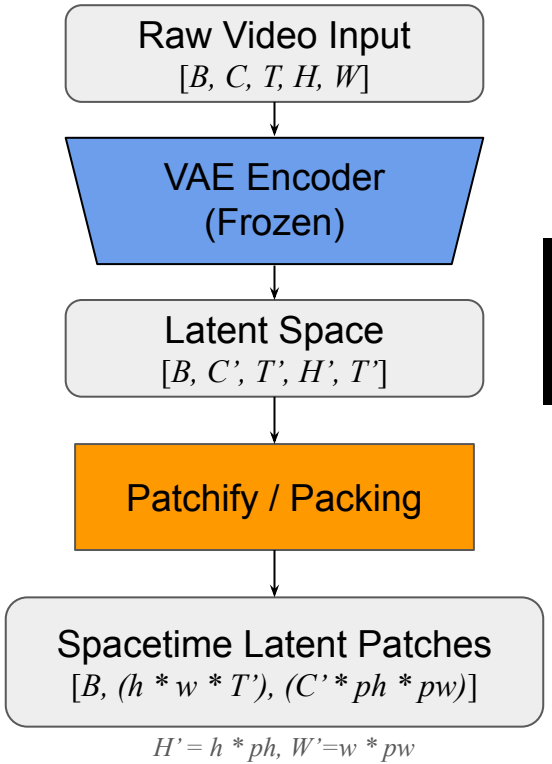
Spacetime Latent Patch: **Sora가 비디오 데이터를 입력으로 받아들이는 단위 표현 방식** → Image와 달리 Video에는 time-axis 존재

→ 비디오를 **공간(spatial)** 과 **시간(temporal)** 을 동시에 고려하는 **3D 토큰화(patchification)** 진행

→ **다양한 해상도/비율의 영상 처리 가능** → 사전 Resizing/Padding 필요 없음

Characteristics of Sora

Compression Structure based Latent Space (based on open-sora)




```python
def pack(x: Tensor, patch_size: int = 2) -> Tensor:
    return rearrange(
        x, "b c t (h ph) (w pw) -> b (t h w) (c ph pw)", ph=patch_size, pw=patch_size
    )
```

einops의 rearrange 문법에 따라 (t, h, w) 중, t가 가장 앞에 위치
→ 시간 순서, 즉 frame의 순서를 유지한 상태로 순서 재배열

https://openai.com/index/video-generation-models-as-world-simulators
https://j-qi.medium.com/openai-soras-technical-review-a8f85b44cb7f - **OpenAI Sora's Technical Review**

Characteristics of Sora

Turning visual data into patches - Spacetime Latent patches

Image (ViT 기준)	Attributes	Video
2D	데이터 차원	3D
공간	패치 구성	공간 + 시간
(H x W)	dim 구성	(H x W x T)
2D VAE	Encoding Structure	3D VAE
2D grid attention	Attention 범위	3D (spatial + temporal)

Latent Patch 비교 (Image vs Video)

Characteristics of Sora

Text → Rich Captioning based Training

기존 Video Dataset - Caption이 너무 짧았음

→ DALL-E 3의 re-captioning 방식을 video에 적용

→ GPT-4를 highly descriptive re-captioner로 train 시켜서 “Descriptive Caption”을 새로 생성

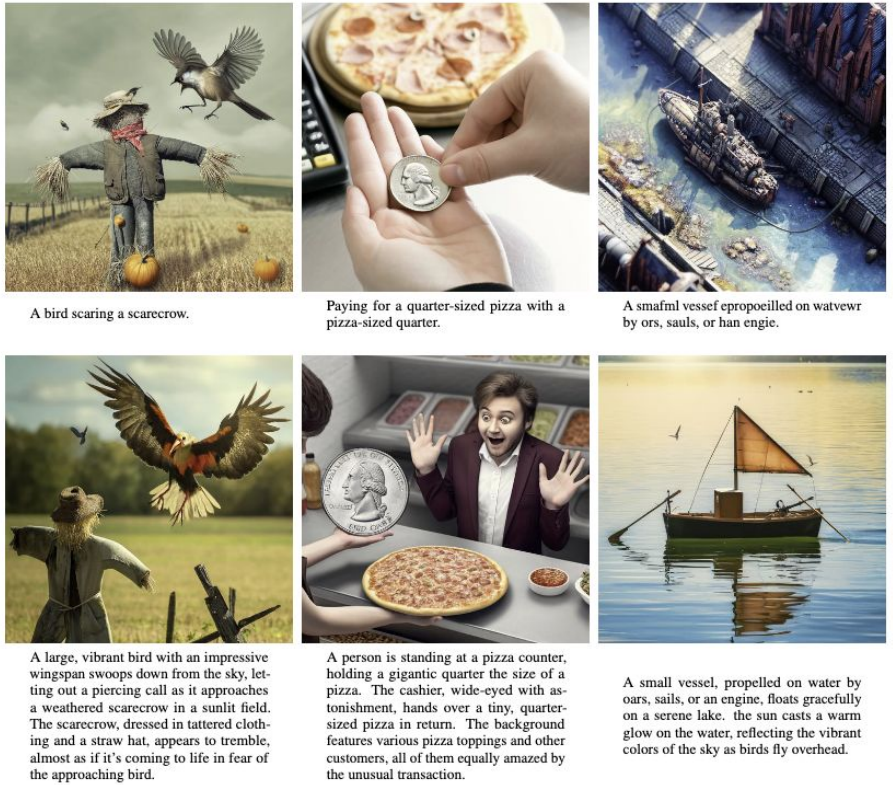


Figure 6 – Effect of using "upsampled" drawbench captions to create samples with DALL-E 3. Original drawbench captions on top, upsampled captions on bottom. Images are best of 4 for each caption.

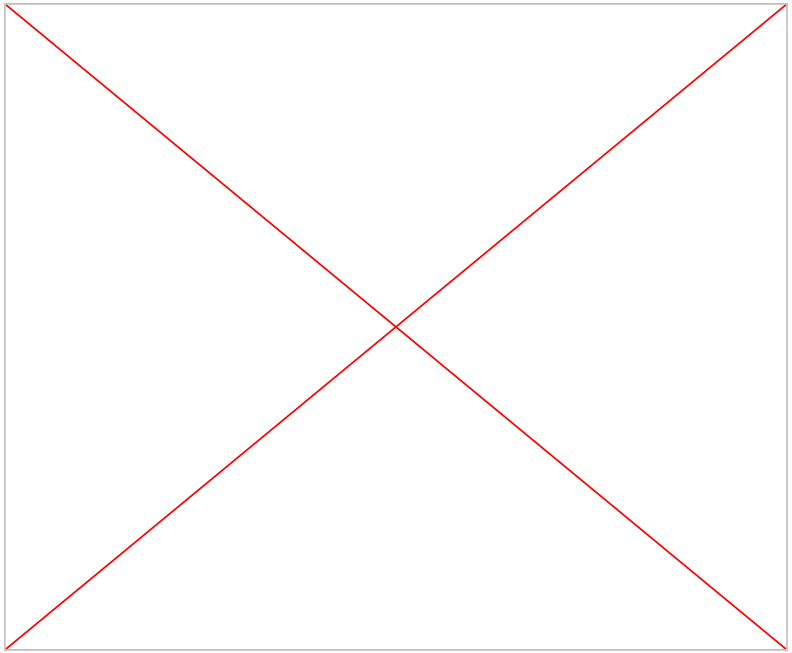
Betker, J. et al. (no date) “Improving Image Generation with Better Captions.”

Capabilities of sora

Animating DALL·E images

- 이미지를 영상화(Image-to-video) 하는 능력

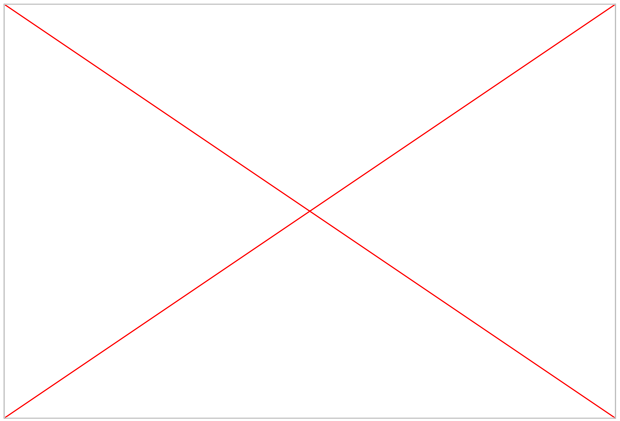


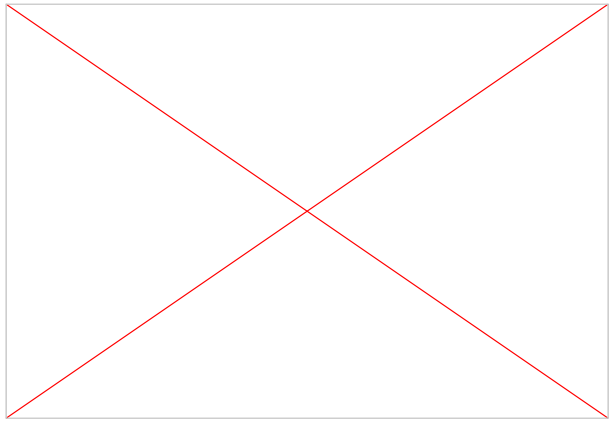


Capabilities of sora

Extending generated videos

하나의 짧은 비디오를 입력하면, 앞으로 또는 뒤로 시간축을 확장하여 긴 영상으로 만들 수 있는 기능







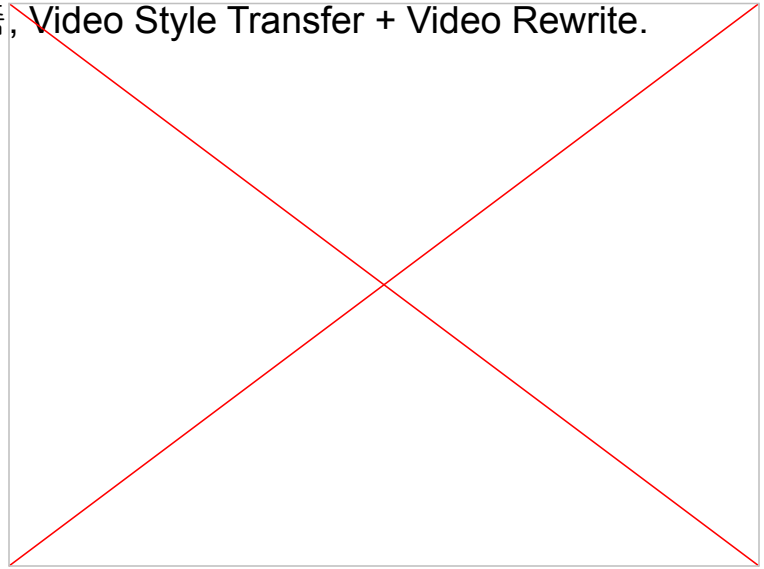
Sora 내부 Transformer 구조에서 이전/이후 프레임 예측하여 이어붙이는 형태로 연장했을 것으로 추정됨

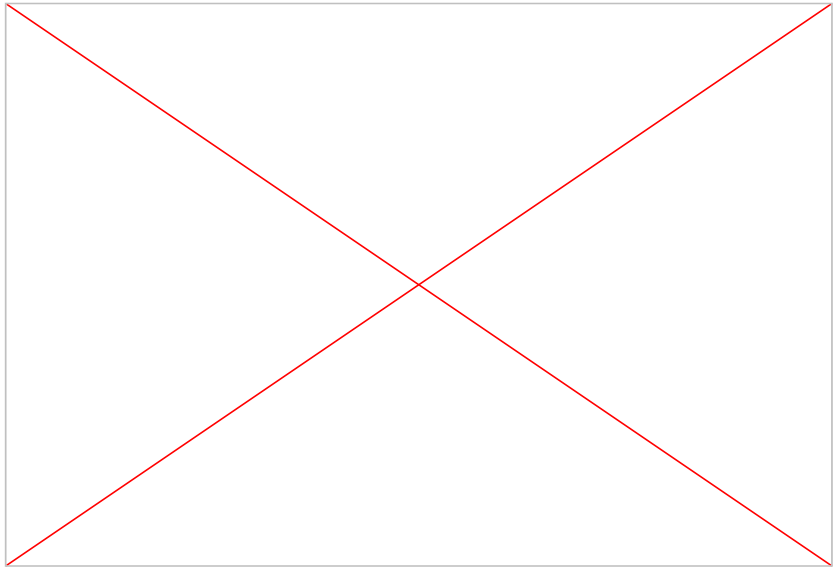
Capabilities of sora

Video-to-Video editing

기존 비디오를 입력하면, 장면 스타일·환경·조명·배경·예술 스타일을 바꾼 새로운 비디오로 변환하는 기능.

→ 즉, Video Style Transfer + Video Rewrite.





Capabilities of sora

Connecting videos

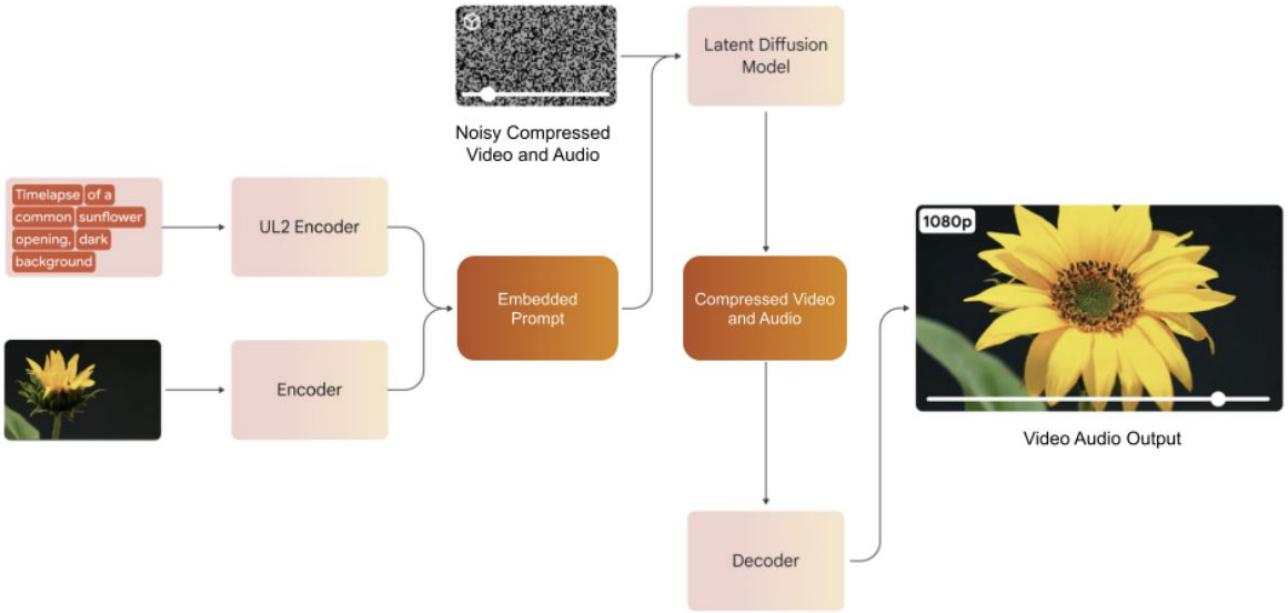




입력된 latent 시퀀스의 중간 구간(intermediate latent)을 생성하는 형태로 **latent interpolation**이 이뤄질 가능성

모델이 “A → B” 변화를 자연스럽게 연결하도록 학습되었거나, inference 시 “A로 시작해서 B로 끝나는 latent 토큰 시퀀스”를 생성하게 설계됨

Characteristics of Veo3



A high-level diagram of Veo

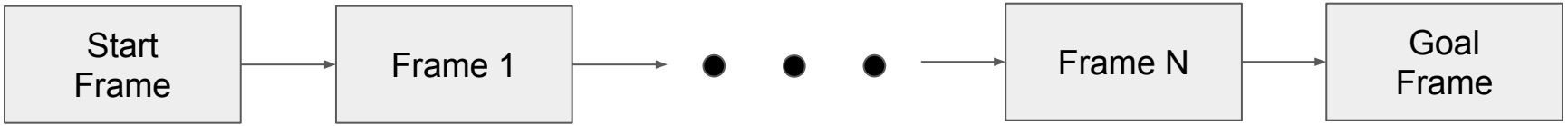
Characteristics of Veo3

Chain of Frames이란

- Frame을 순서대로 생성과정을 문제를 논리적으로 풀어나가는 추론 과정으로 간주
  - 3가지 Reasoning Task로 평가 시도
- LLM의 CoT(Chain of Thought)의 아이디어를 Vision Domain으로 가져오려는 시도
  - LLM에게 문제를 해결하는 과정의 중간 단계를 명시적으로 보여주거나 생각하도록 유도하는 방법 (문제 해결 단계 제시, 예시 제공 etc.)



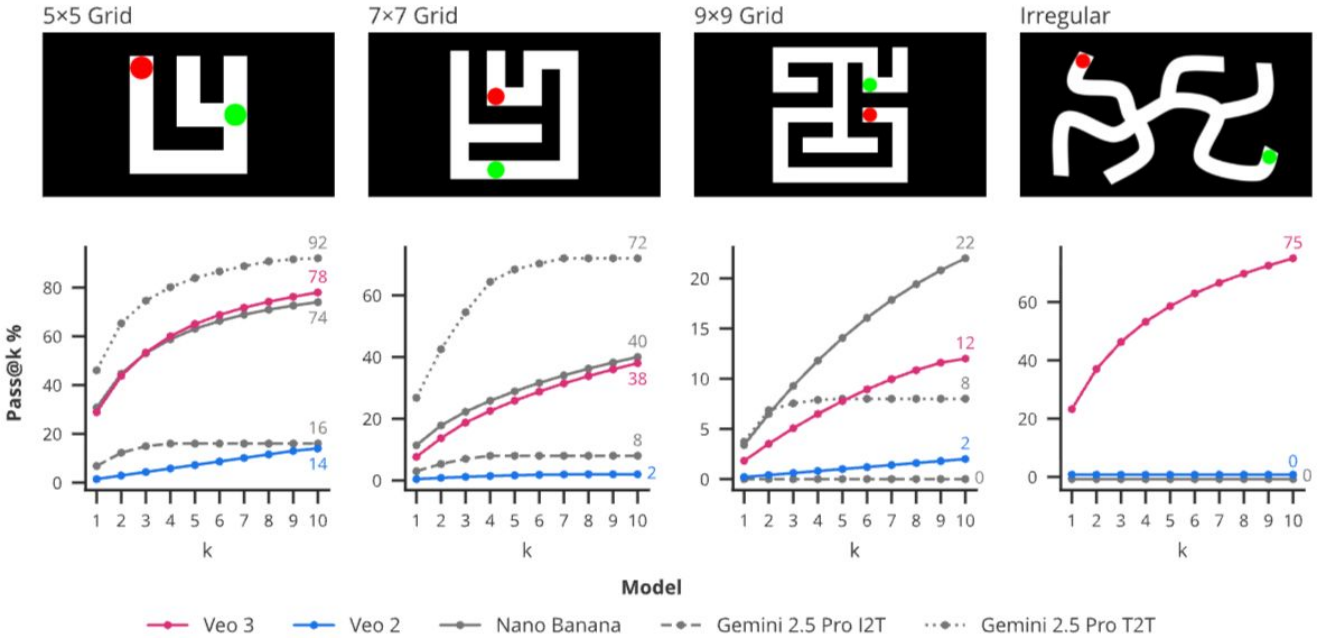
Chain of Thought



Chain of Frames

Characteristics of Veo3

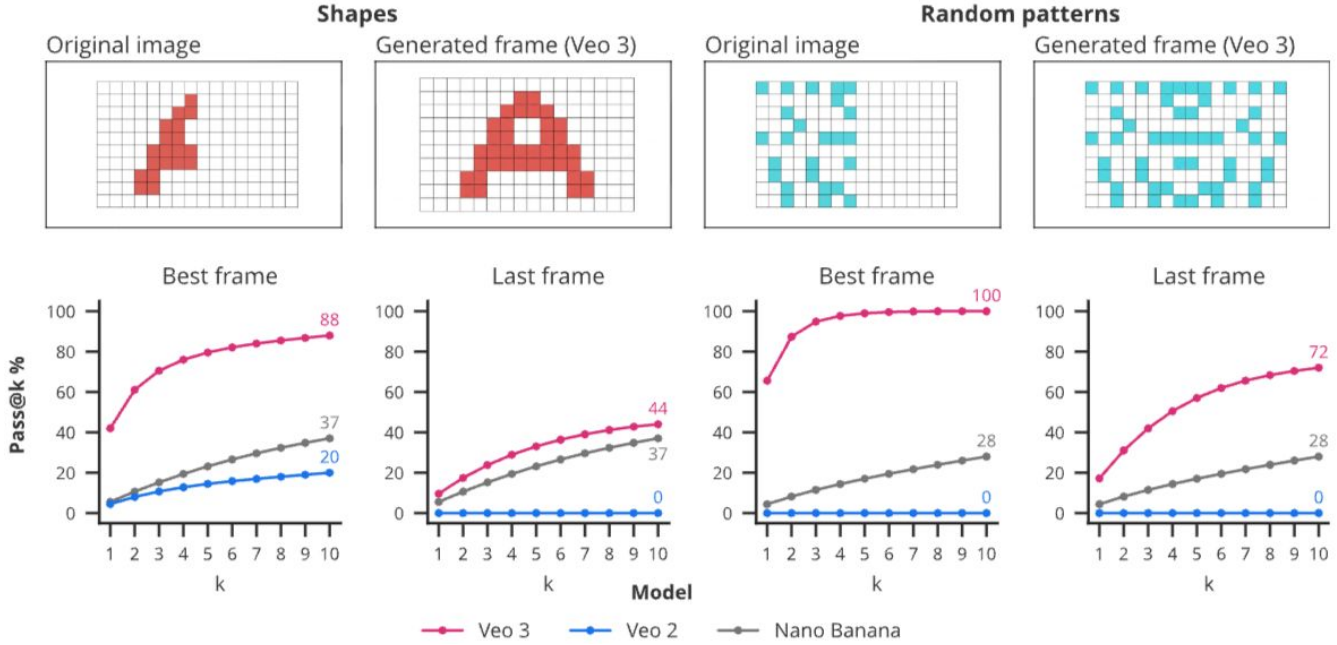
Chain of Frames



Maze Solving: 빨간색 원이 벽을 가로지르지 않고 목표 위치(초록색 원)로 이동해야 하는 Task
Veo 2가 illegal moves으로 작은 미로조차 해결 X, Veo 3는 5x5 그리드에서 pass@10 78%를 달성
특히 irregular maze에서 veo3만이 유의미한 성과 → CoF 시각적 추론의 초기 형태 중 하나

Characteristics of Veo3

Chain of Frames

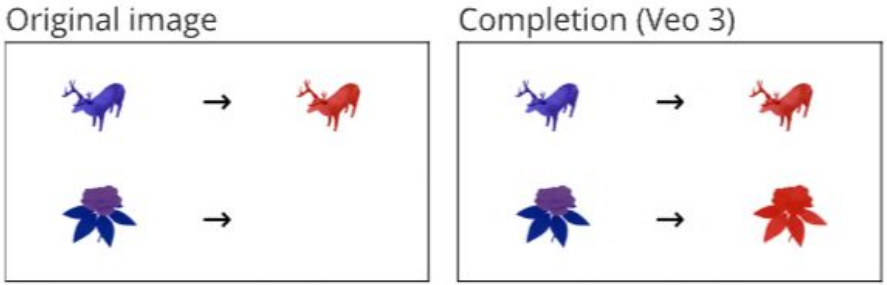


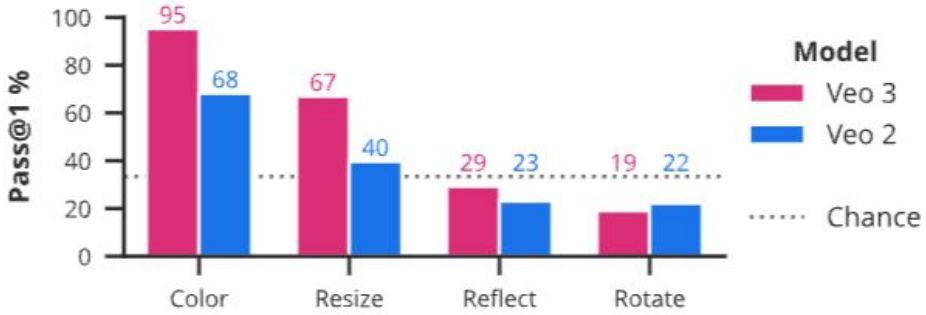
Visual Symmetry Solving: 주어진 패턴을 대칭적으로 완성하는 태스크, 모델의 공간추론 능력 평가

Veo 3는 Veo 2, Nano Banana 큰 차이로 앞섬, **공간 추론 능력**에서 탁월한 개선
→ “대칭”이라는 시각적 상태를 이해, 최종적인 목표 상태(완성된 대칭 패턴)를 달성하는 단계별 시각적 추론 및 생성 능력

Characteristics of Veo3

Chain of Frames:



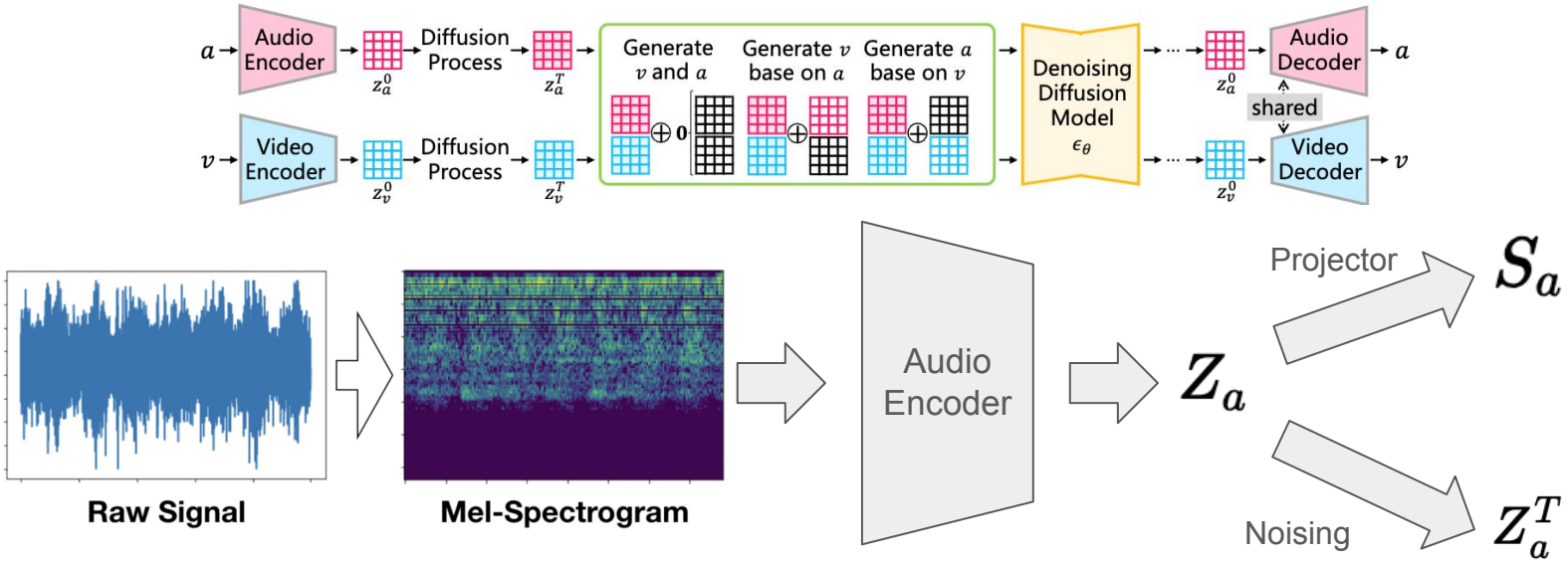


Visual Analogy Completion - 주어진 시각적 관계나 변환 규칙을 이해, 이를 새로운 상황에 적용, 빠진 부분을 완성 작업

Veo 2: 예시 문제로부터 관계 추론하는 과정에서 어려움 보임
Veo 3: 색상 및 크기 조정 유추를 정확하게 완성하는 초기 수준의 추상적 추론 달성

Characteristics of Veo3

Inherent Multimodal Integration of Audio and Video - 추정되는 통합 방식

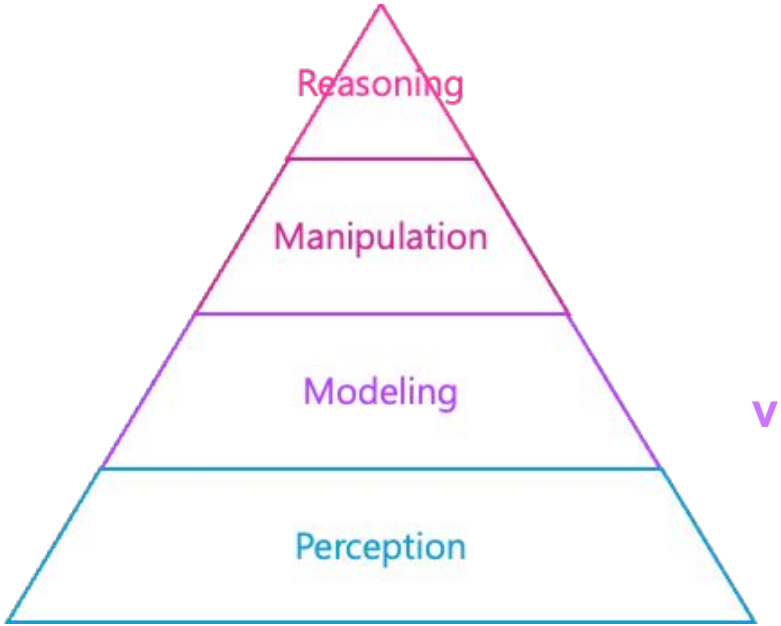


Process of Audio Encoding in MM-LDM

Audio와 Video 입력을 Unified type으로 처리
→ Audio를 2D image(Mel-Spectrogram, 시간적 정보 포함)로 변형해서 통합 생성

Sun, M. et al. (2024) “MM-LDM: Multi-Modal Latent Diffusion Model for Sounding Video Generation,” ACM Multimedia [Preprint].

LVM(Veo 3)’s four hierarchical capabilities



시공간에 따른
Manipulation의 순차적 수행
인식되고 모델링된 visual world를 변형하는 능력

Perception 기반
visual world의 모델 구축 능력

시각 정보의 '입력' 및 '해석'
(가장 기본적인 감각)

Task 복잡도

Veo3 저자 - 위의 4단계의 Vision Task의 층위 제안
이를 기반으로 범용 LVM의 잠재력을 가졌음을 입증하고자 함 → **LLM의 역량과 비교**

LVM(Veo 3)’s four hierarchical capabilities

Perception

- 다양한 기초 CV 태스크 제로샷 수행: Edge Detection, Segmentation
- LLM 비유: GPT가 문법, 단어 의미, 문장 구조를 이해하는 것과 같음.
- LVM의 의미: 범용적인 시각 '입력 모듈'로서의 강력한 기반 능력 확보.



Edge detection



Segmentation

https://video-zero-shot.github.io/

LVM(Veo 3)’s four hierarchical capabilities

Modeling

- 직관적 물리 법칙 이해: 부력, 중력, 강체/연체 역학, 공기 저항, 재료 속성(가연성, 광학 현상) 시뮬레이션
- LLM과의 비교: GPT가 세상의 사실, 상식, 개념, 논리법칙 등을 학습하는 것과 같음.
- LVM의 의미: Vision World에 대한 '이해의 틀' 구축, 복잡한 시나리오 예측 가능성.



물체 연소



Object Packing



부력

https://video-zero-shot.github.io/

LVM(Veo 3)’s four hierarchical capabilities

Manipulation

- **다양한 이미지/비디오 편집**: 배경 제거, 스타일 변환, 인페인팅/아웃페인팅, 3D-aware reposing
- **객체 추출 및 재배치:** 특정 객체만 추출하여 재구성
- LLM과 비교: GPT가 글쓰기, 요약, 번역 등 **언어를 조작하는 '행동'하는 것과 같음**
- LVM의 의미: **Vision World를 '통제'하고 '변형'**할 수 있는 능동적인 개입자로서의 잠재력.



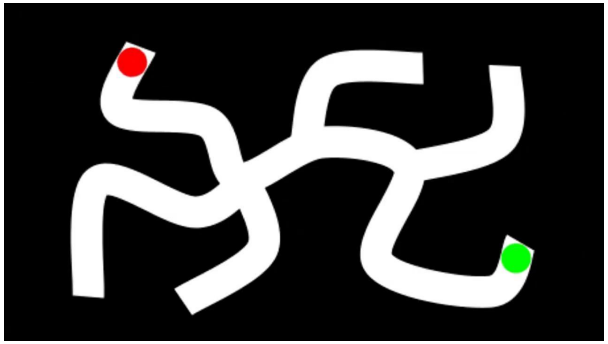


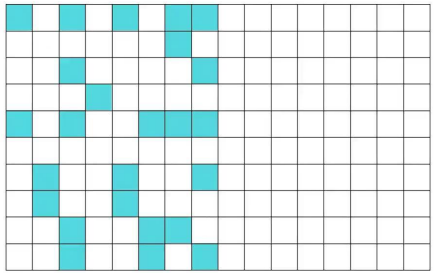
https://video-zero-shot.github.io/

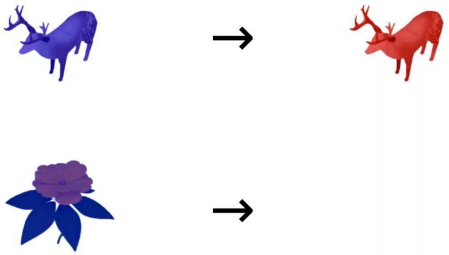
LVM(Veo 3)’s four hierarchical capabilities

Reasoning

- 복잡한 시각적 문제 해결: 미로 풀기, 시각적 대칭 완성, 시각적 유추 완성, 도구 사용, 스도쿠 풀이







- LLM과 비교: GPT가 코딩, 수학, 복잡한 질문 답변을 통해 **'추론'**하는 것과 유사한 task라고 주장

https://video-zero-shot.github.io/

- LVM의 의미: Visual World 내에서 **Frame의 순서를 통해 '생각하고 문제 해결'**하는 진정한

Applications

Applications of Video Generation Model

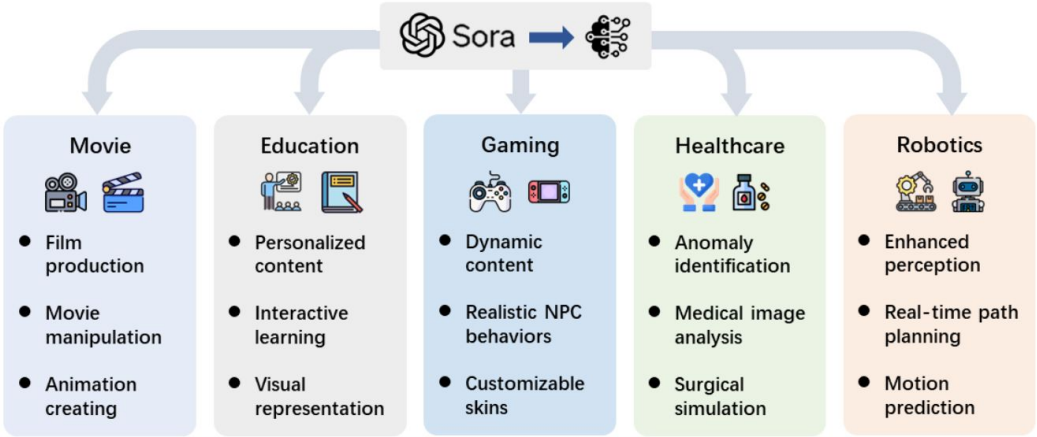


Figure 18: Applications of Sora.

(3) *The Need for Interdisciplinary Collaboration*: Ensuring the safety of models is not just a technical issue but also requires cross-disciplinary cooperation. To address these challenges, experts from various fields such as law [131] and psychology [132] need to work together to develop appropriate norms (e.g., what's the safety and what's unsafe?), policies, and technological solutions. The need for interdisciplinary collaboration significantly increases the complexity of solving these issues.

https://www.oxen.ai/blog/road-to-sora-reading-list?utm_source=pytorchkr

Applications of Video Generation - Movie





https://youtube.com/shorts/4VALgDNeWMQ?si=Ic_Pfiu-crzLo_W9

Applications of Video Generation - Movie



https://www.oxen.ai/blog/road-to-sora-reading-list?utm_source=pytorchkr

Applications of Video Generation - Movie

AI Platforms & Assistants

This movie is fully AI-generated and has a fully SAG-AFTRA cast – here's 3 things you need to know about Echo Hunter

News By Eric Hal Schwartz published May 31, 2025

Virtual films made real

Comments (1)

When you purchase through links on our site, we may earn an affiliate commission. Here's how it works.



(Image credit: Echo Hunter)

조회수 338,914회 2025. 5. 29.
Arcana Labs Presents - A Phantom X Production - Echo Hunter
A Kavan the Kid Film

Phantom X
https://www.phantomx.ai/contact

Arcana Labs
Made with the tools inside Arcana Labs: https://www.arcanalabs.ai/echo-hunter...

In a world where the elite harvest clones for spare parts, a relentless hunter begins to question his mission while confronting his forgotten past. He'll have to choose between duty and a devastating truth.

Starring -
Breckin Meyer, Taylor John Smith, Danielle Bisutti, Hanna Balicki, Gedeon Burkhard and Xander Bailey.

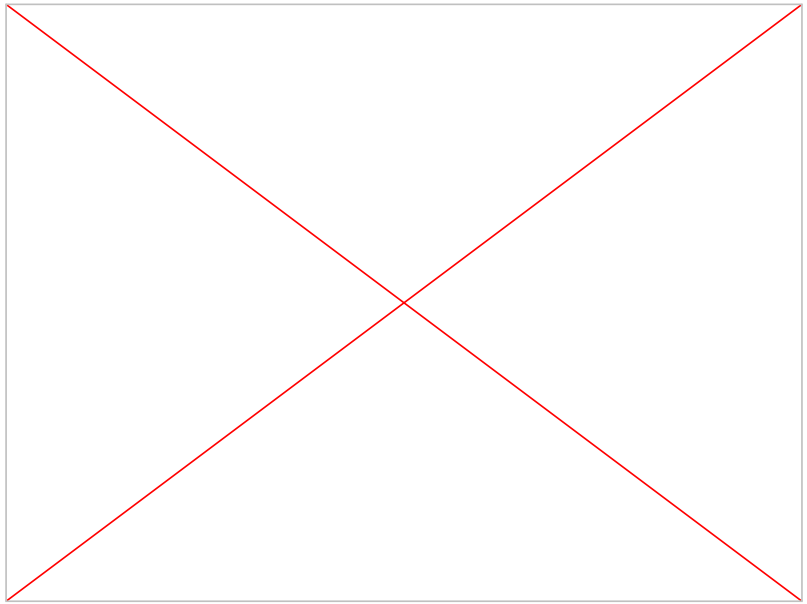
Executive Producer - Jonathan Yunger
Producer - Mike J Mitch

Music Composed By - Miguel Johnson
VFX By - Sav

시나리오, 목소리, - SAG-AFTRA actors
영상 내 화면 구성 - ARCANA AI
→ 영화 제작에 있어 사람의 참여 최소화하는 도전

https://www.techradar.com/computing/artificial-intelligence/this-movie-is-fully-ai-generated-and-has-a-fully-sag-aftra-cast-heres-3-things-you-need-to-know-about-echo-hunter

Applications of Video Generation - Gaming



Prompt: The video shows a first person perspective of someone navigating difficult terrain in the middle of a volcanic area.
This is a real world video shot from the perspective of a wheeled robot that needs to traverse across a terrain.
The vehicle has chunky offroad tires that crunch under the blackened rock.
The camera is an egocentric camera mounted to the vehicle, and you can see the front tires just on the bottom of the camera along with the body of the robot.
In the distance you can see smoke and lava flowing from the volcano.
There are no other visible signs of life.
There are lava pools that the agent is trying to avoid and random rock formations.
The sky is a vivid blue.

Deepmind Genie 3 - Image + Prompt → Interactive world 생성
Real-Time Interactivity + 높은 일관성 → 기존 게임 엔진을 대체 / 확장하는 생성형 게임 엔진의 초기 형태

https://youtu.be/PDKhUknuQDg?si=OJbQhv3FWUei2MHC

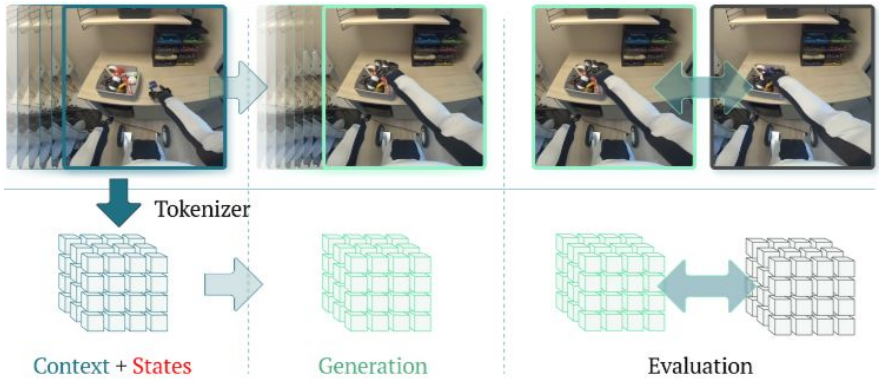
Applications of Video Generation - Gaming



Prompt: A vibrant 3D style, an adorable, fluffy creature bounding across a vibrant rainbow bridge in a fantastical landscape. The creature is small and compact, with fur that mimics the warm hues of a sunrise – oranges, yellows, and pinks blending seamlessly together. Its most striking feature is a pair of large, perked ears, shaped like those of a German Shepherd, adding a touch of playful contrast to its otherwise rounded form. As it runs on four short legs across the …

https://youtu.be/PDKhUknuQDg?si=OJbQhv3FWUei2MHC

Applications of Video Generation - Robotics





1x World Model

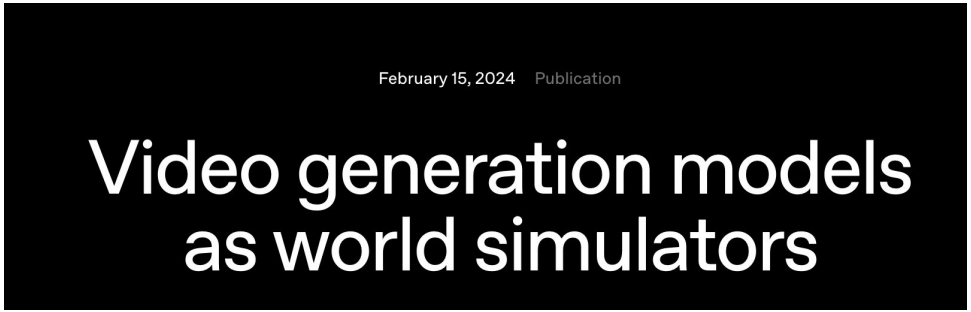
- **로봇 학습의 핵심 기반** - 물리 법칙, 인과관계, 동역학을 이해하는 “World model” 필요
- **현실 데이터 수집의 한계**: 고비용·고위험·저효율로 대규모 학습 데이터 확보가 어려움
- **기존 영상 생성 모델의 문제**: 시각적 사실성 위주, 물리적 정확성·상호작용 모델링 부족
- 물리법칙 내장한 Video Generation Model

→ **실제와 같은 물리적 반응 시뮬레이션 가능**

https://youtu.be/9Uch931cDx8

Road to World Model





https://deepmind.google/blog/genie-3-a-new-frontier-for-world-models/
https://openai.com/index/video-generation-models-as-world-simulators/

World Model: 텍스트나 이미지를 넘어 현실 세계의 물리적 법칙, 공간, 시간, 상호작용 등을 이해하고 시뮬레이션할 수 있도록 구축된 파운데이션 모델

A key stepping stone on the path to AGI

- Make it possible to train AI agents in **an unlimited curriculum** of rich simulation environments.
- Generate new environments for agents

Common Direction

- Google Deepmind, Openai sora 모두 World model을 지향점으로 하고 있음.