



Sora: A Review on Background, Technology, Limitations, and Opportunities of Large Vision Models

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,

 <https://openai.com/research/video-generation-models-a-s-world-simulators>

OpenAI 공식 technical report



! Official Paper는 아님 !



Sora

- 2024년 2월에 OpenAI에 의해 발표된 텍스트-비디오 생성 인공지능 모델
 - ↳ 텍스트 지시에 기반하여 현실적이거나 상상의 장면을 생성
- 해당 논문은 공개 기술 보고서와 역공학을 기반으로 이 모델의 배경/ 관련 기술/ 응용 분야/ 남아 있는 과제 및 텍스트-비디오 AI 모델의 미래 방향에 대한 포괄적인 검토를 제시

1. Introduction

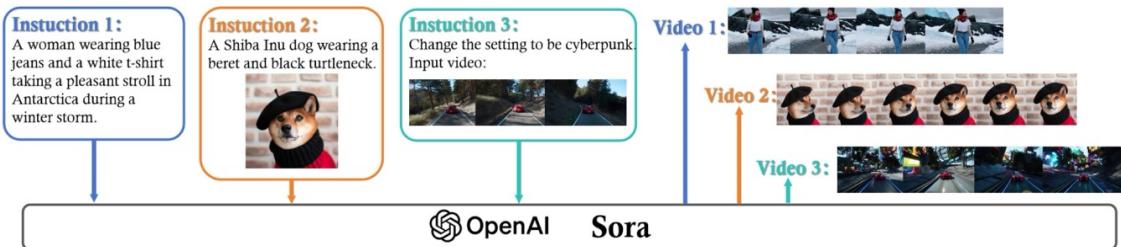


Figure 2: Examples of Sora in text-to-video generation. Text instructions are given to the OpenAI Sora model, and it generates three videos according to the instructions.

• Sora

- 텍스트 프롬프트에서 현실적이거나 상상적인 장면의 비디오를 생성할 수 있는 텍스트-비디오 생성 AI 모델
- 복잡한 사용자 지침을 해석하고, 이러한 이해를 적용하여 역동적이고 상황에 맞는 풍부한 시뮬레이션을 생성
- 높은 시각적 품질과 강력한 시각적 일관성을 유지하면서 최대 1분 길이의 비디오를 저장할 수 있음

⇒ 사용자가 텍스트 내러티브를 풍부한 시각적 스토리로 변환할 수 있음

Technology

- 사전 훈련된 확산 트랜스포머(diffusion transformer)
 - Sora는 텍스트를 구문 분석하고 복잡한 사용자 지시를 이해할 수 있음
 - 비디오 생성의 계산 효율성을 높이기 위해 공간-시간 잠재 패치를 사용
 - 원시 입력 비디오를 잠재 공간-시간 표현으로 압축
 - 축된 비디오에서 잠재 공간-시간 패치의 시퀀스가 추출되어 시각적 외형과 생성된 비디오가 원하는 내용 및 품질과 더 일치하도록 다단계 정제 프로세스를 통해 나타냄

Highlights

시뮬레이션 능력 향상

- 명시적인 3D 모델링이 없더라도 Sora는 동적 카메라 이동과 먼 거리 일관성을 포함한 3D 일관성을 나타내며, 물체 지속성과 세계와의 간단한 상호 작용을 모방 가능
- ⇒ 물리적 및 디지털 세계의 복잡성을 시뮬레이션

창의성 촉진

텍스트를 통해 개념을 개요로 그린 다음 몇 초 내에 현실적이거나 고풍스러운 비디오를 생성해 냄

교육 혁신 촉진

Sora를 사용하면 교육자는 쉽게 텍스트에서 비디오로 여러 자료들을 시뮬레이션 할 수 있음

접근성 향상

텍스트 설명을 시각적 콘텐츠로 변환

⇒ 시각 장애가 있는 사람을 포함한 모든 개인이 콘텐츠 생성에 참여하고 더 효과적으로 다른 사람과 상호 작용할 수 있도록 능력을 부여

신흥 응용 프로그램 육성

- 마케터: 특정 대상 설명에 맞춘 동적 광고를 만들기 위해 이를 사용할 수 있음
- 게임 개발자: 플레이어 이야기에서 맞춤형 시각적/캐릭터 작업 생성

Limitations/Opportunities

- 복잡한 동작 묘사, 미묘한 얼굴 표정을 잡아내는 것
- 윤리 문제
 - 편향을 완화하고 유해한 시각적 결과물을 방지

2. Background

2-1. History

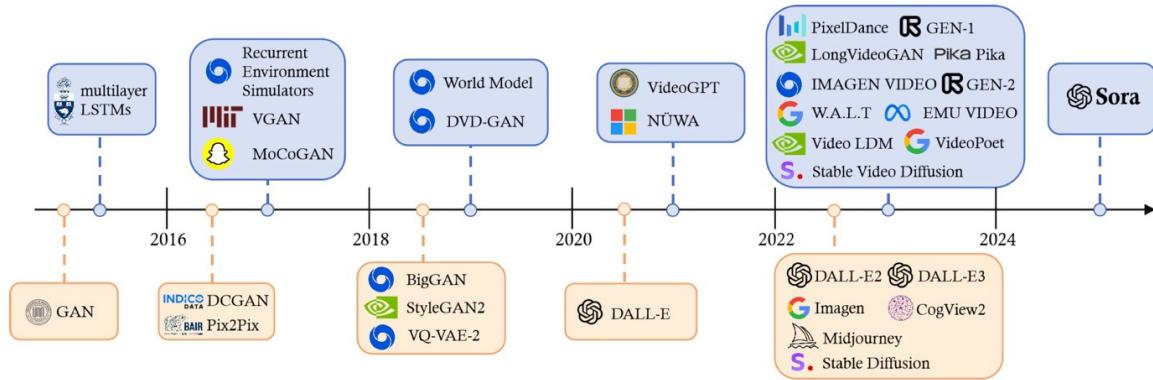


Figure 3: History of Generative AI in Vision Domain.

- 전통적인 이미지 생성 기술은 텍스처 합성 등에 의존하였음
 - texture synthesis, texture mapping 등
 - 복잡, 이미지 생성에 한계
- 생성적 적대 신경망(GAN)의 도입, VAE(변형 자동 인코더)
- flow model, diffusion model
 - 향상된 디테일과 품질로 이미지 생성이 더욱 향상되었음
- 최근) AIGC(인공지능 생성 콘텐츠)

⇒ 사용자가 간단한 텍스트 지침을 통해 원하는 콘텐츠를 생성할 수 있게 되었음
- Transformer 아키텍처의 성공적인 적용 이후 급격하게 발전
 - NLP: BERT, GPT
 - CV: ViT(Vision Transformer), Swin Transformer, Diffusion Model, U-Net
- 2020년) 다중 모드 모델로 알려진 인간의 지시를 해석할 수 있는 생성 언어 및 비전 모델의 개발/발전
 - CLIP
 - 변환기 아키텍처와 시각적 요소를 결합하여 방대한 텍스트 및 이미지 데이터 세트에 대한 교육을 용이하게 하는 선구적인 비전 언어 모델
 - 처음부터 시각적 및 언어적 지식을 통합함으로써 다중 모드 생성 프레임워크 내에서 이미지 인코더로 기능
 - Stable Diffusion
 - 적응성과 사용 용이성으로 유명한 다목적 텍스트-이미지 AI 모델

- 변환기 아키텍처와 잠재 확산 기술을 사용
→ 텍스트 입력을 디코딩하고 다양한 스타일의 이미지를 생성하며 multi-modal AI의 발전을 더욱 잘 보여줌
- 최근) 사용자가 간단한 텍스트 프롬프트를 통해 고해상도 및 품질의 새로운 이미지를 생성할 수 있게 되었음
 - Stable Diffusion, Midjourney, DALL-E 3

Sora는 인간의 지시에 따라 최대 1분 길이의 비디오를 생성할 수 있는 최초의 모델이다!

2-2. Advanced Concepts

비전 모델의 확장 법칙

- LLM처럼 비전 모델도 유사한 확장 법칙을 따를까?
→ ViT의 확장 속도를 보니까 그런 것 같다!
- 고정 모델을 사용하여 임베딩을 생성한 다음 그 위에 얇은 레이어를 훈련하면 뛰어난 성능을 얻을 수 있음

긴급 능력

- 개발자가 명시적으로 프로그래밍하거나 예상하지 않은 특정 규모(종 모델 매개변수의 크기와 연결됨)에서 나타나는 정교한 동작 또는 기능에 대한 대처 능력
⇒ 좋은 것 같다~

3. Technology

3-1. Overview of Sora

In the core essence, Sora is a diffusion transformer [4] with flexible sampling dimensions as shown in Figure 4. It has three parts: (1) A time-space compressor first maps the original video into latent space. (2) A ViT then processes the tokenized latent representation and outputs the denoised latent representation. (3) A CLIP-like [26] conditioning mechanism receives LLM-augmented user instructions and potentially visual prompts to guide the diffusion model to generate styled or themed videos. After many denoising

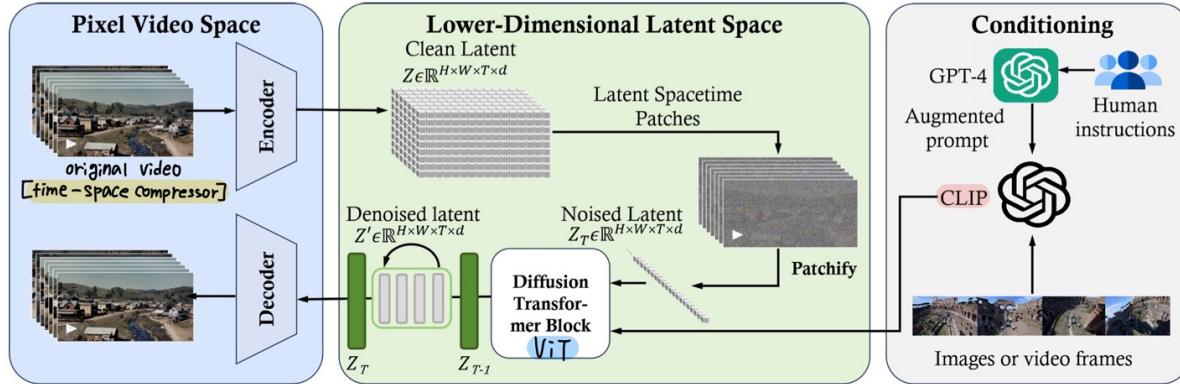


Figure 4: **Reverse Engineering:** Overview of Sora framework

1. 시간-공간 압축기: 원본 비디오를 잠재 공간으로 맵핑
2. ViT: 토큰화된 잠재 표현을 처리하고 노이즈 제거된 잠재 표현을 출력
3. 조건화 메커니즘: CLIP과 유사, LLM으로 증강된 사용자 지시와 잠재적으로 시각적 프롬프트를 수신하여 확산 모델이 스타일이나 테마가 있는 비디오를 생성하도록 안내
 - 다수의 노이즈 제거 단계를 거친 후, 생성된 비디오의 잠재 표현이 얻어지고 해당 디코더와 함께 픽셀 공간으로 다시 맵핑

3-2. Data Pre-processing

3-2-1. Variable Durations, Resolutions, Aspect Ratios



(a) Training on videos that are cropped to squares leads to unnatural compositions and framing.
(b) Training in native sizes improves framing.

Figure 6: A comparison between Sora (right) and a modified version of the model (left), which crops videos to square shapes—a common practice in model training—highlights the advantages.

- 비디오와 이미지를 원래 크기에서 학습하고 이해하며 생성할 수 있음
 - 가변적인 크기의 데이터를 처리할 수 있음
 - 이로 인한 정보 손실 등의 문제를 개선
- ⇒ 전통적인 인공지능의 인간 기반 추상화에 의존하지 않음

3-2-2. Unified Visual Representation

- 다양한 지속 시간, 해상도, 종횡비를 갖는 이미지와 비디오를 효과적으로 처리하기 위해서는 모든 형태의 시각적 데이터를 통일된 표현으로 변환하여 대규모로 생성 모델을 훈련시켜야 함
- Sora 는 비디오를 먼저 낮은 차원의 잠재 공간으로 압축한 다음 해당 표현을 시공간 패치로 분해하여 비디오를 패치화

3-2-3. Video Compression Network

(도대체 어디까지가 official이고 어디가 애내들의 뇌피셜인걸까)

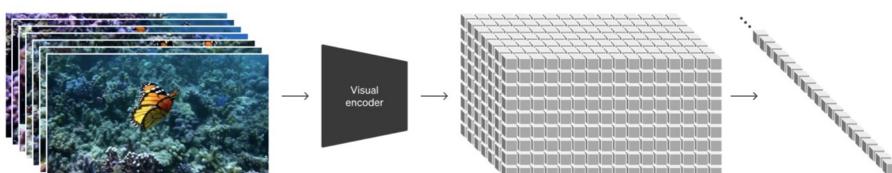


Figure 7: At a high level, Sora turns videos into patches by first compressing videos into a lower-dimensional latent space, and subsequently decomposing the representation into spacetime patches. Source: Sora's technical report [3].

- Sora의 비디오 압축 네트워크는 입력 데이터의 차원을 줄이고 압축된 시간적 및 공간적 잠재 표현을 출력하고자 함

- VAE 또는 VQ-VAE 기반의 압축 네트워크는 크기 조정 및 자르기를 사용하지 않으면 모든 크기의 시각 데이터를 통일된 크기의 잠재 공간으로 매핑하는 것이 어려움

⇒ 문제 해결을 위해 논문에서 연구자들은 두 가지 구현 방법을 제시
(official 아님..?)

1. 공간-패치 압축

- 비디오 프레임을 고정 크기 패치로 변환한 다음 잠재 공간으로 인코딩하기 전에 처리
- 다양한 해상도와 종횡비의 비디오를 수용하기에 효과적이며, 시간 차원 가변성, 사전 훈련된 시각 인코더 활용, 시간 정보 집계 등을 강조

2. 공간-시간-패치 압축

- 비디오 데이터의 공간 및 시간 차원을 모두 고려하여 표현하는 기술로, 3D 컨볼루션을 활용
- 시간 정보 집계에 더 중점을 두는 공간-패치 압축과 비교하여 동적인 측면을 포착하는 것을 강조
- 이러한 접근 방식은 VAE나 VQ-VQE와 같은 변형을 기반으로 하며, 큰 패치 크기를 사용하여 고품질 비디오를 생성하려는 목표에 따라 고정 크기의 패치를 사용할 것으로 기대됨
 - 다양한 크기의 패치를 사용할 경우 무효화된 위치 인코딩 및 디코더에서 크기가 다른 잠재 패치로 비디오를 생성하는 어려움이 발생할 수 있음

3-2-4. Spacetime Latent Patches

- 압축 네트워크 부분에서 남은 중요한 고민은 다양한 비디오 유형에서 나오는 잠재 피처 청크 또는 패치의 수에 대한 잠재 공간 차원의 변이를 어떻게 처리할 것인가임(⇒ 가변적인 길이)
- ⇒ **PNP(patch n' pack)**로 해결할 수 있지 않을까?
- PNP는 여러 이미지에서 추출한 패치를 단일 시퀀스에 패킹하는 방법
 - 가변 길이 입력에 효과적인 교육을 제공하는 자연어 처리의 예제 패킹에서 영감
- 패치화 및 토큰 임베딩 단계는 압축 네트워크에서 완료되어야 하며, 시퀀스 내의 토큰을 효율적으로 패킹하고 어떤 토큰을 삭제할지에 대한 고민이 필요

- 이를 위해 Sora는 PNP와 유사한 방법을 사용할 것으로 예상되며, 토큰 패킹에 대한 간단한 그리디 접근 방식을 사용하면서 패딩을 제한하기 위해 샘플링된 해상도와 프레임을 조절할 것으로 예상됨
- 또한, Sora는 세부 정보 손실을 최소화하기 위해 모든 토큰을 패킹하는 슈퍼 긴 컨텍스트 창을 사용할 것으로 예상됨

3-2-5. Discussion

- Sora의 데이터 전처리에는 두 가지 기술적 해결책이 존재
 - 이전 방법과 달리 비디오를 표준 크기로 조정, 자르기 또는 다듬지 않고 Sora는 데이터를 원본 크기에서 훈련
 - 이를 위해 Sora는 시각 패치를 낮은 차원의 잠재 표현으로 압축하고 이를 시퀀스에 배열한 후 확산 트랜스포머의 입력 레이어에 주입하기 전에 이러한 잠재 패치에 노이즈를 추가한다고 예상됨
 - Sora가 채택한 공간-시간 패치화는 구현이 간단하며 고밀도 정보를 가진 토큰을 사용하여 컨텍스트 길이를 줄이고 시간 정보를 모델링하는 복잡성을 감소시킬 수 있음
- ⇒ 비디오 모델링의 효과성과 효율성 간의 균형 유지에 대해 고민해야 함

3-3. Modeling

3-3-1. Diffusion Transformer

이미지 디퓨전 트랜스포머

- 전통적인 디퓨전 모델은 대부분 다운샘플링과 업샘플링 블록을 포함하는 컨볼루션 U-Net을 활용
 - 그러나 최근 연구에 따르면 U-Net 아키텍처는 디퓨전 모델의 성능에 중요하지 않는다는 것이 확인됨
 - 더 유연한 트랜스포머 아키텍처를 통합 ⇒ DiT와 U-ViT

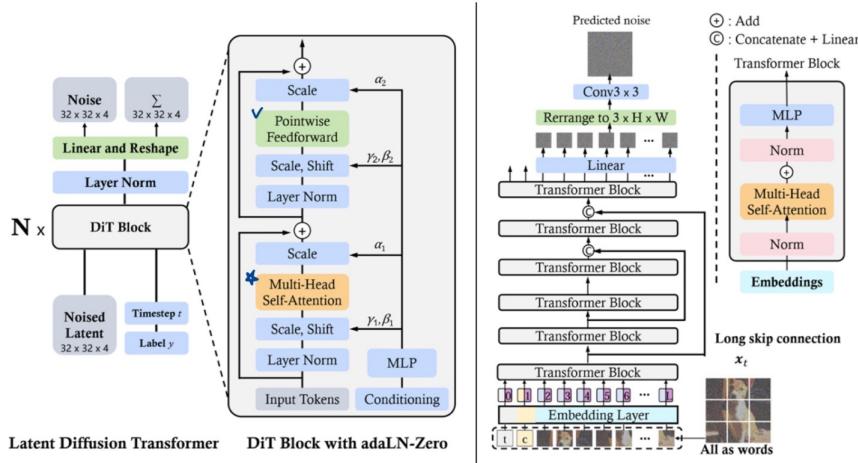


Figure 11: The overall framework of DiT (left) and U-ViT (right)

- **DiT**는 ViT와 유사하게 멀티헤드 셀프어텐션 레이어와 pointwise feed-forward 네트워크를 사용하며 일부 레이어 정규화 및 스케일링 레이어와 함께 엮여 있음
 - 추가로 MLP 레이어를 사용하여 조건부로 adaptive layer norm (AdaLN)을 통합
 - 이는 각 잔여 블록을 항등 함수로 초기화하여 훈련 과정을 안정화
- **U-ViT**는 모든 입력(시간, 조건 및 노이지 이미지 패치 포함)을 토큰으로 취급하고 얕은/깊은 트랜스포머 레이어 간에 긴 스kip 연결을 제안
 - CNN 기반 U-Net의 다운샘플링 및 업샘플링 연산자가 항상 필요하지 않음을 시사
- **Masked Diffusion Transformer(MDT)**는 이미지 합성에서 물체 의미 부분 간의 맥락적 관계 학습을 명시적으로 향상시키기 위해 마스크 잠재 모델링을 통합
⇒ DiT보다 뛰어난 성능과 더 빠른 학습 속도

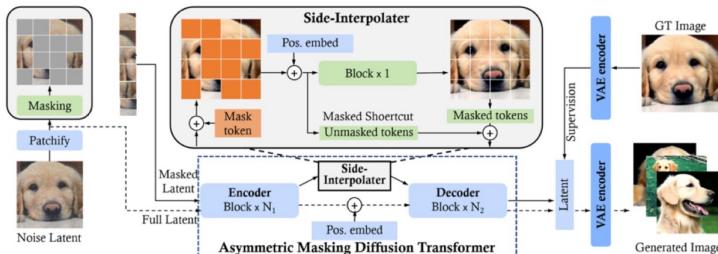


Figure 12: The overall framework of Masked Diffusion Transformer (MDT). A solid/dotted line indicates the training/inference process for each time step. Masking and side-interpolator are only used during training and are removed during inference.

- **Diffusion Vision Transformers(DiffiT)**는 AdaLN을 사용하는 대신 시간 조건 모델링을 위해 시간 종속 셀프 어텐션(TMSA) 모듈을 도입
 - 공간 및 잠재 공간에서 효율적인 노이징을 위해 두 가지 하이브리드 계층 구조를 사용

비디오 디퓨전 트랜스포머

- 최근 연구는 텍스트에서 비디오(**T2V**) 생성 작업에 대한 diffusion transformer의 잠재력을 실현하는 데 중점을 두고 있음
- 비디오의 **시간적(temporal)** 특성으로 인해 DiT를 비디오 도메인에 적용하는 데는 몇 가지 도전 과제가 있음
 - 비디오에서 효율적으로 노이즈를 제거하기 위한 잠재 공간으로의 시간 및 공간 압축
 - 압축된 잠재(latent)를 패치로 변환하고 트랜스포머에 공급하는 방법
 - 장거리 시간 및 공간 종속성을 다루고 콘텐츠 일관성을 보장하는 것

⇒ 해당 논문에서는 DiT를 사용하여 공간/시간적으로 압축된 잠재 공간에서 작동하는 노이즈 제거 네트워크 아키텍처에 중점을 두고, 이와 관련된 두 가지 주요 작업인 **Imagen Video**와 **Video LDM**에 대한 상세 검토를 제시

▼ Imagen Video

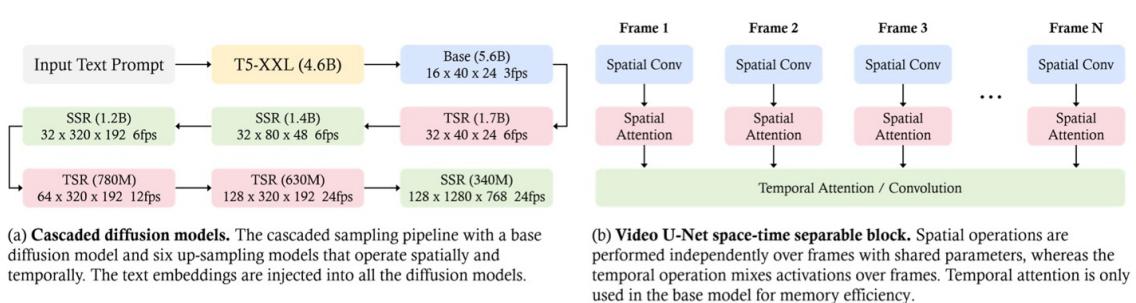
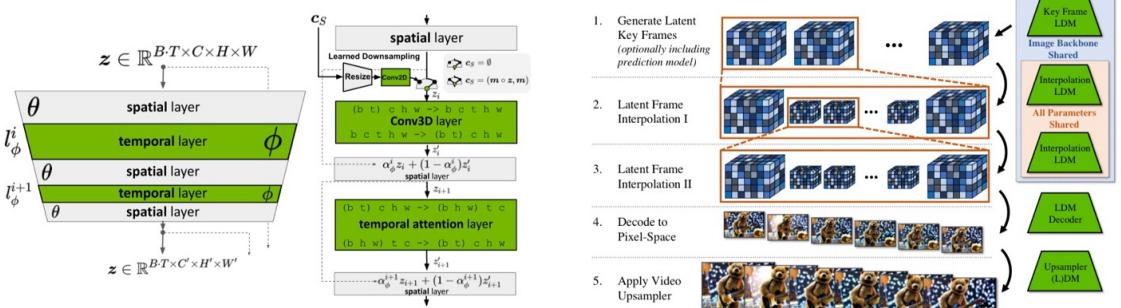


Figure 13: The overall framework of Imagen Video. Source: Imagen Video [29].

- Google Research에서 개발한 텍스트에서 비디오 생성 시스템
- 7개 하위 모델의 디퓨전 모델 연속을 사용하여 텍스트 프롬프트를 고해상도 비디오로 변환
 - frozen T5 텍스트 인코더에 의해 생성된 문맥 임베딩 기반
 - 저해상도 비디오 생성을 위해 기본 모델에 주입된 후 디퓨전 모델을 통해 해상도를 향상
 - 공간 및 시간 분리형 3D U-Net 아키텍처를 사용하여 프레임 간 종속성을 효과적으로 캡처하며, 이미지 및 비디오에 대한 공동 훈련을 통해 높은 충실도와 풍부한 제어성을 가짐

▼ Video LDM



(a) **Additional temporal layer.** A pre-trained LDM is turned into a video generator by inserting temporal layers that learn to align frames into temporally consistent sequences. During optimization, the image backbone θ remains fixed and only the parameters ϕ of the temporal layers l_ϕ^i are trained.

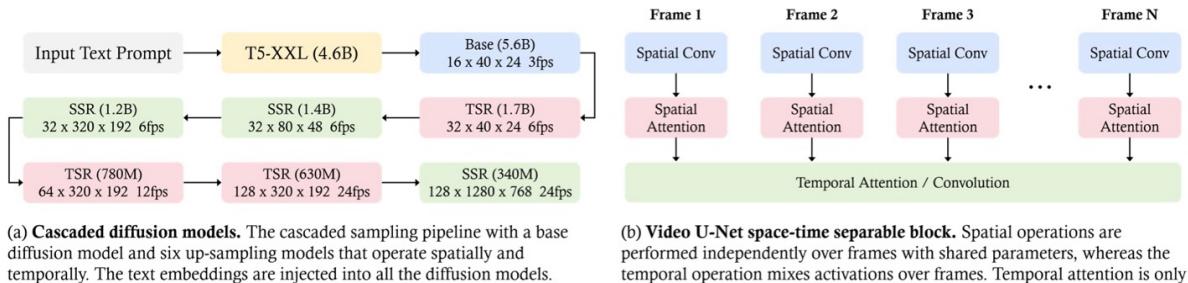
(b) **Video LDM stack.** Video LDM first generates sparse key frames and then temporally interpolates twice with the same latent diffusion models to achieve a high frame rate. Finally, the latent video is decoded to pixel space, and optionally, a video upsampler diffusion model is applied.

Figure 14: The overall framework of Video LDM. Source: Video LDM [36].

- 2D Latent Diffusion Model → Video Latent Diffusion Model로의 변환을 제안 (공간감 추가?)
- 공간 레이어에 몇 개의 시간 레이어를 추가하여 각 프레임을 정렬하며, 디코더는 시간적 일관성 및 공간 해상도 향상을 위해 미세 조정됨
- 훈련 데이터에서 인코딩된 비디오를 기반으로 시간 레이어를 훈련하면서 이미지 데이터셋을 사전 훈련에 활용
 - Video LDM 출력을 4배로 확장하여 고 공간 해상도를 유지하면서 시간적 일관성을 유지

⇒ 계산 효율적으로 긴 비디오를 생성하며, 이미지 LDM을 텍스트에서 비디오로 변환할 수 있는 능력도 보여줌

3-3-2. Discussion



(a) **Cascaded diffusion models.** The cascaded sampling pipeline with a base diffusion model and six up-sampling models that operate spatially and temporally. The text embeddings are injected into all the diffusion models.

(b) **Video U-Net space-time separable block.** Spatial operations are performed independently over frames with shared parameters, whereas the temporal operation mixes activations over frames. Temporal attention is only used in the base model for memory efficiency.

Figure 13: The overall framework of Imagen Video. Source: Imagen Video [29].

공간 및 시간 업샘플링을 위한 캐스케이드 확산 모델

- Sora는 기본 모델과 여러 공간-시간 refiner 모델로 구성된 케스케이드 확산 모델 아키텍처를 활용하는 것으로 판단됨
- 또한 기존의 latent x 또는 노이즈 ϵ 을 예측하는 다른 변형체와 비교 시 뛰어난 성능을 보이는 v-파라미터화 확산 모델을 사용할 가능성이 높음

잠재 인코더 측면

- 훈련 효율성을 위해 대부분의 기존 작업은 Stable Diffusions의 사전 훈련된 2D 확산 모델의 VAE 인코더를 초기화된 모델 체크포인트로 활용
 - 그러나 해당 인코더는 시간적 압축 능력이 부족
- 기존의 사전 훈련된 VAE 인코더 대신 Sora는 비디오 데이터에 대해 처음부터 훈련된 space-time VAE 인코더를 사용할 가능성이 높다고 판단됨

3-4. Language Instruction Following

- 사용자는 생성형 AI 모델과 주로 텍스트 프롬프트로 알려진 자연어 지시를 통해 상호 작용함
 - 모델 지시 튜닝은 AI 모델이 정확하게 프롬프트를 따르도록 성능을 향상시키는 것이 목표
 - 모델이 자연어 질문에 대한 인간의 응답과 더 가깝게 출력을 생성할 수 있도록 하고자 함

→ 대형 언어 모델(LLMs) 및 DALL·E 3와 같은 텍스트-이미지 모델에 대한 명령(지시) 따르기 기술을 검토

- 텍스트-비디오 모델의 텍스트 지시 따르기 능력을 향상시키기 위해 Sora는 DALL·E 3 와 유사한 접근 방식을 활용
 - 묘사적 캡션 생성자를 훈련하고 캡션 생성자가 생성한 데이터를 fine-tuning에 활용

⇒ 다양한 사용자 요청 수용 + 지시 세부 사항 준수 + 사용자의 요구에 정확히 부합하는 비디오를 생성해 냄

3-4-1. Large Language Model

- 지시 튜닝을 통해 LLM을 지시로 포맷된 작업의 혼합물로 fine tuning을 수행
- 보이지 않는(reference가 없는?) 작업에 대해서도 지시만으로 일정 수준 이상의 결과물을 생성해 냄
(무에서 유를 창조했다는 건가..)

3-4-2. Text-to-Image

- 노이즈가 많거나 대부분의 시각 정보를 생략하는 짧은 캡션 등의 저품질 데이터의 경우에는 키워드와 단어 순서를 무시하고 사용자 의도를 오해하는 등 여러 문제를 야기함
- 이미지의 자세한 캡션을 생성하기 위해 이미지 캡션 개선 방법을 사용
 - 이미지 캡션 생성자를 훈련시키고 그 결과를 텍스트-이미지 모델을 미세 조정하는데 활용
 - 이를 통해 모델은 사용자의 입력을 적절하게 캡처
- 이미지 캡션 개선 방법은 데이터의 불일치 문제를 야기
 - 업샘플링을 통해 해결
 - 이 과정에서 LLMs를 사용하여 짧은 사용자 프롬프트를 상세하고 긴 지시로 다시 쓰는 데 활용하여 모델의 텍스트 입력 일관성을 유지

3-4-3. Text-to-Video

- Sora는 지시 따르기 능력을 향상시키기 위해 유사한 캡션 개선 접근 방식을 채택
 - 먼저 비디오에 대한 자세한 설명을 생성할 수 있는 비디오 캡션 생성기를 훈련시킴
 - 이후 이 비디오 캡션 생성기를 훈련 데이터의 모든 비디오에 적용하여 고품질 (비디오, 설명형 캡션) 쌍을 생성
- **Sora의 기술 보고서에서는 비디오 캡션 생성기의 교육에 대한 구체적인 내용이 나오지 않음**
(해당 리뷰 논문의 뇌피셜..?)
 - VideoCoCa: CoCa 아키텍처를 비디오 캡션에 활용
 - mPLUG-2, GIT, FrozenBiLM 등도 고려될 수 있음
- 사용자 지시가 훈련 데이터의 설명형 캡션 형식과 일치하도록 하기 위해 Sora는 GPT-4V를 사용하여 사용자 입력을 자세한 설명형 프롬프트로 확장하는 추가적인 프롬프트 확장 단계를 수행

3-5. Prompt Engineering

- 생성 모델의 맥락을 특정하거나 최적화된 출력을 얻기 위해 AI 시스템에 제공되는 입력을 설계/개선하는 과정
⇒ 모델이 가장 정확하고 관련성 있으며, 일관된 응답을 생성하도록 입력을 조작하는 작업

- 텍스트, 이미지, 비디오 프롬프트의 복합적 활용
→ 시각적으로 잘 정렬된 결과물 도출 가능
 - 기존에는 텍스트 or 이미지까지만 활용했는데, Sora에서는 비디오 프롬프팅까지 활용하지 않았을까?

3-5-1. Text Prompt



Figure 15: A case study on ^{text}**prompt** engineering for text-to-video generation, employing color coding to delineate the creative process. The text highlighted in blue describes the elements generated by Sora, such as the depiction of a stylish woman. In contrast, the text in yellow accentuates the model's interpretation of actions, settings, and character appearances, demonstrating how a meticulously crafted prompt is transformed into a vivid and dynamic video narrative.

장면의 행동, 설정, 캐릭터 외형, 심지어 원하는 분위기와 대기까지 자세히 명시됨

- 텍스트에서 비디오로의 모델을 지도하여 시각적으로 강렬하면서도 사용자의 명세를 정확하게 충족시키도록 하는 데 중요
 - 인간의 창의력과 AI의 실행 능력 사이의 간극을 최소화
- 프롬프트 엔지니어링의 품질은 단어의 신중한 선택, 제공된 세부 사항의 구체성 및 이러한 세부 사항이 모델 출력에 미치는 영향을 이해하는 데에 달려 있음

3-5-2. Image Prompt

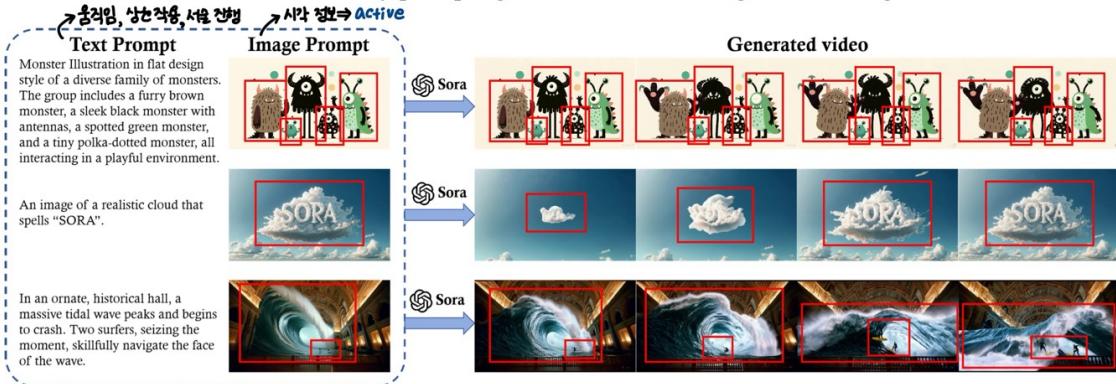


Figure 16: This example illustrates the image prompts to guide Sora’s text-to-video model to generation. The red boxes visually anchor the key elements of each scene—monsters of varied designs, a cloud formation spelling “SORA”, and surfers in an ornate hall facing a massive tidal wave.

- 이미지 프롬프트는 생성될 비디오의 내용 및 캐릭터, 설정, 분위기 등과 같은 다른 요소에 대한 시각적 기준을 제시하는 역할
- 이미지 프롬프트의 사용은 Sora에게 시각적 및 텍스트 정보를 활용하여 정직 이미지를 동작이고 서술 주도적인 비디오로 변환할 수 있도록 함
- DALL-E에서 생성된 이미지를 Sora의 이미지 프롬프팅으로 활용하여 비디오 생성 능력을 향상시킴

3-5-3. Video Prompt

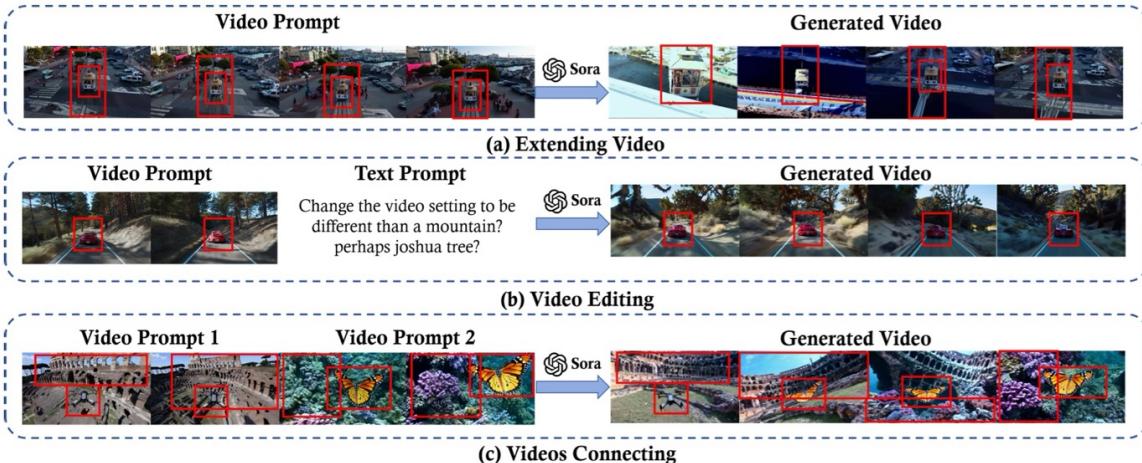


Figure 17: These examples illustrate the video prompt techniques for Sora models: (a) Video Extension, where the model extrapolates the sequence backward the original footage, (b) Video Editing, where specific elements like the setting are transformed as per the text prompt, and (c) Video Connection, where two distinct video prompts are seamlessly blended to create a coherent narrative. Each process is guided by a visual anchor, marked by a red box, ensuring continuity and precision in the generated video content.

- 최근 연구들(Moonshot, Fast-Vid2Vid)은 좋은 비디오 프롬프트가 구체적이고 유연해야 한다는 점을 시사

- 모델이 특정한 객체와 시각적 테마의 묘사와 같은 명확한 목표에 대한 명확한 지침을 받을 수 있을 뿐만 아니라 최종 출력에서 창의적인 변화도 허용될 수 있도록 하기 위함

4. Applications

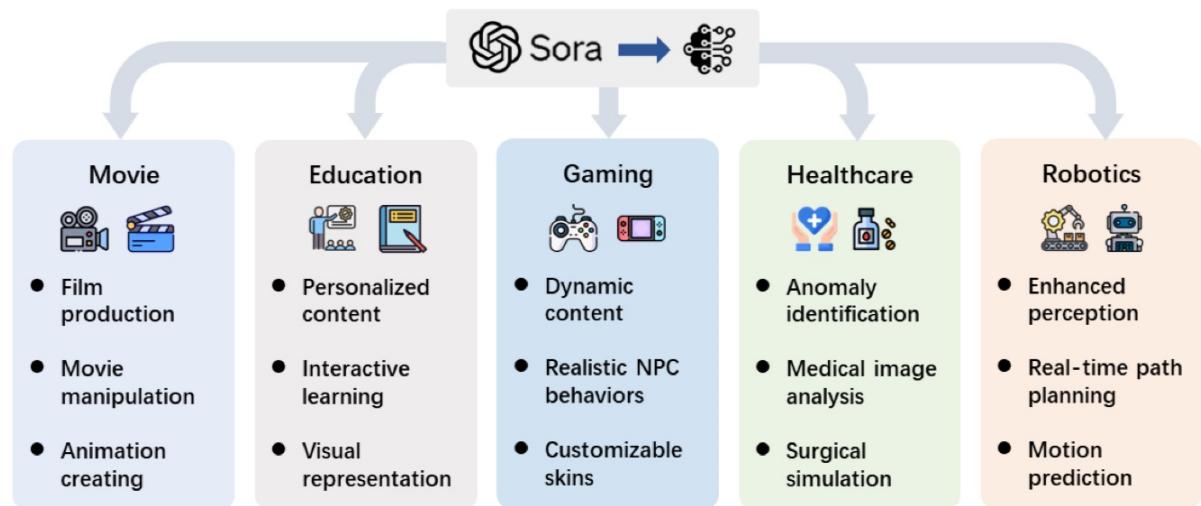


Figure 18: Applications of Sora.

5. Discussion

- Sora는 정교하게 구성된 환경 속에서 다양한 캐릭터를 활용하여 세밀한 비디오를 생성
 - 일관된 스토리텔링 능력을 보장하면서 최대 1분까지의 비디오를 생성 가능

5-1. Limitations

물리적 현실정에 대한 도전

- 복잡한 장면 내에서 물리적 원리를 일관되게 처리하지 못하며, 특정한 예의 인과 관계를 정확하게 복사하지 못하는 한계가 존재
 - ex) 쿠키의 일부를 섭취하면 해당 부분에 대응하는 물방울이 생기지 않음
- 움직임 시뮬레이션
 - 물체의 비정상적인 변형이나 의자와 같은 강체 구조물의 잘못된 시뮬레이션 등

공간/시간 복잡성

- 주어진 프롬프트 내에서 객체 및 캐릭터의 배치 또는 정렬과 관련된 지시사항을 오해하며 방향에 대한 혼란을 야기
 - 왼쪽을 오른쪽으로 혼동
- 이벤트의 시간적 정확성을 유지하는 데 어려움을 겪기도 함
 - 의도된 장면의 시간적 흐름에서 벗어나게 될 수 있음
- 캐릭터나 요소가 다수 포함된 복잡한 시나리오에서는 Sora가 관련 없는 동물이나 인물을 삽입하는 경향이 있음

인간-컴퓨터 상호 작용(HCI)

- 비디오 생성 영역에서 잠재력을 보이지만, HCI에서는 상당한 제약 사항이 존재
- 복잡한 언어 지시사항을 이해하거나 미묘한 의미 차이를 포착하는 능력의 제약
→ 사용자 기대/요구를 완전히 충족하지 못하는 비디오 콘텐츠의 생성

5-2. Opportunities

- 학계, 산업, 사회 여러 분야에 적용되어 발전 가능성이 높음