Video Outpainting



Presented By
문형주

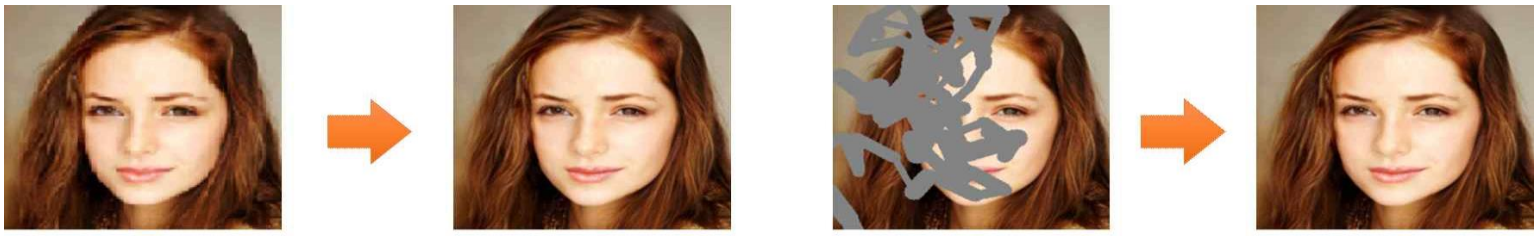
Video Outpainting

- **Task**

  - Image Outpainting

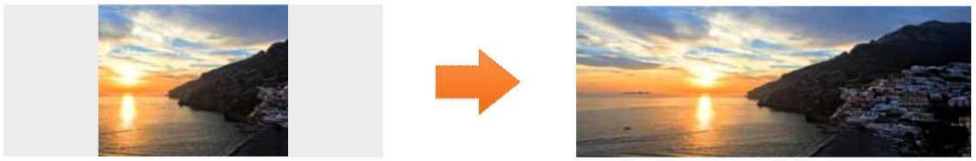
  비디오 프레임의 가장자리에 있는 빈 영역을 적절히 채우는 것을 목표로 하여 이미지의 외부 영역을 생성해내는 방법

  ⇛GANs기법과 Diffusion 모델이 적용되면서 많은 발전



(a) Super resolution

(b) Image inpainting



(c) Image outpainting

Video Outpainting

- **Task**

  - Video Outpainting

  Many frame, Long video

  ⇛채워진 영역의 temporal consistency 유지 어려움, Artifact 축적 문제

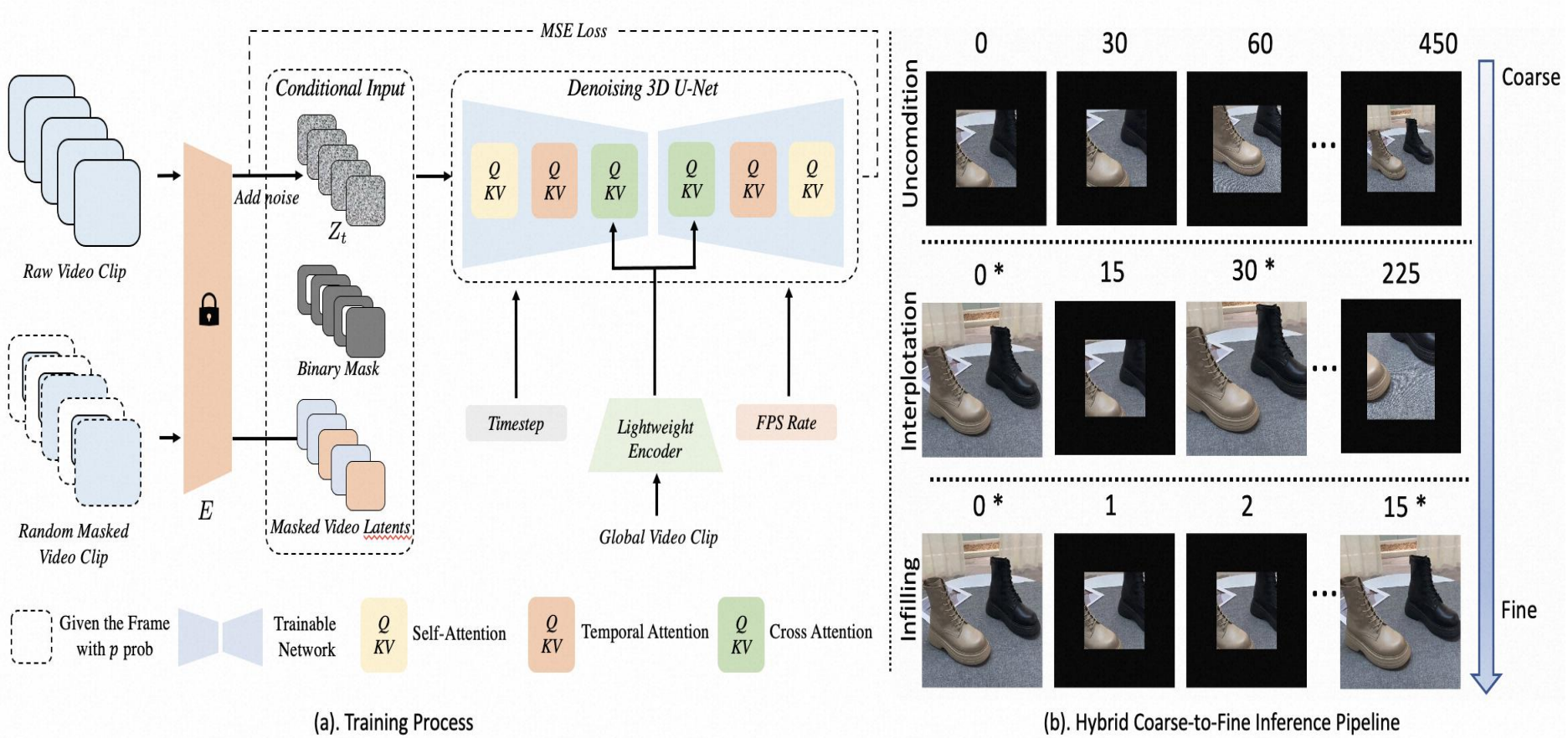


Video Outpainting

- Paper
  - Hierarchical Masked 3D Diffusion Model for Video Outpainting
    - Fanda Fan1, 2, *, Chaoxu Guo3, *, Litong Gong3, Biao Wang3, Yuning Jiang3 Chunjie Luo1, Jianfeng Zhan1, 2
  - Abstract
    - Video outpainting aims to adequately complete missing areas at the edges of video frames. Compared to image outpainting, it presents an additional challenge as the model should maintain the temporal consistency of the filled area. In this paper, we introduce a maske 3D diffusion model for video outpainting. We use the technique of mask modeling to train the 3D diffusion model. This allows us to use multiple guide frames to connect the results of multiple video clip inferences, thus ensuring temporal consistency and reducing jitter between adjacent frames. Meanwhile, we extract the global frames of the video as prompts and guide the model to obtain information other than the current video clip using cross-attention. We also introduce a hybrid coarse-to-fine inference pipeline to alleviate the artifact accumulation problem. The existing coarse-to-fine pipeline only uses the infilling strategy, which brings degradation because the time interval of the sparse frames is too large. Our pipeline benefits from bidirectional learning of the mask modeling and thus can employ a hybrid strategy of infilling and interpolation when generating sparse frames. Experiments show that our method

Video Outpainting

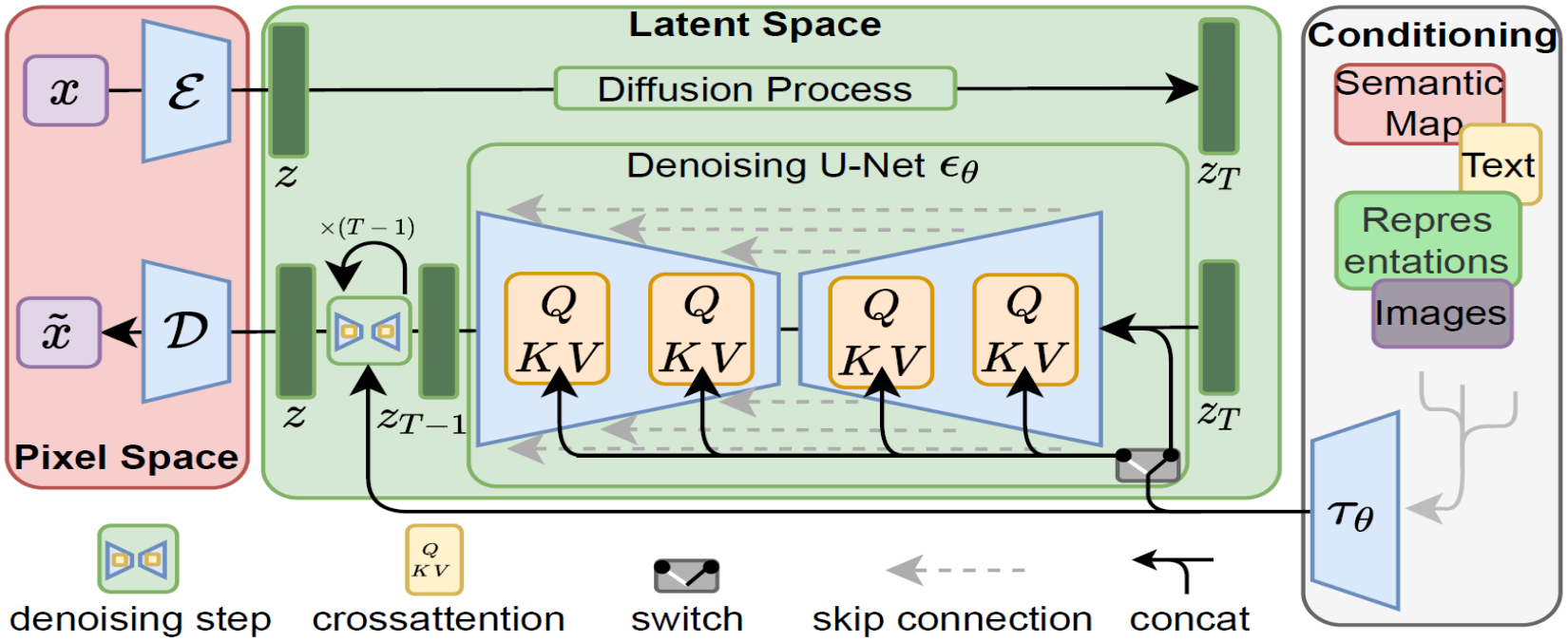
- Framework



1. masked 3D diffusion model (M3DDM)
2. hybrid coarse-to-fine inference pipeline

Video Outpainting

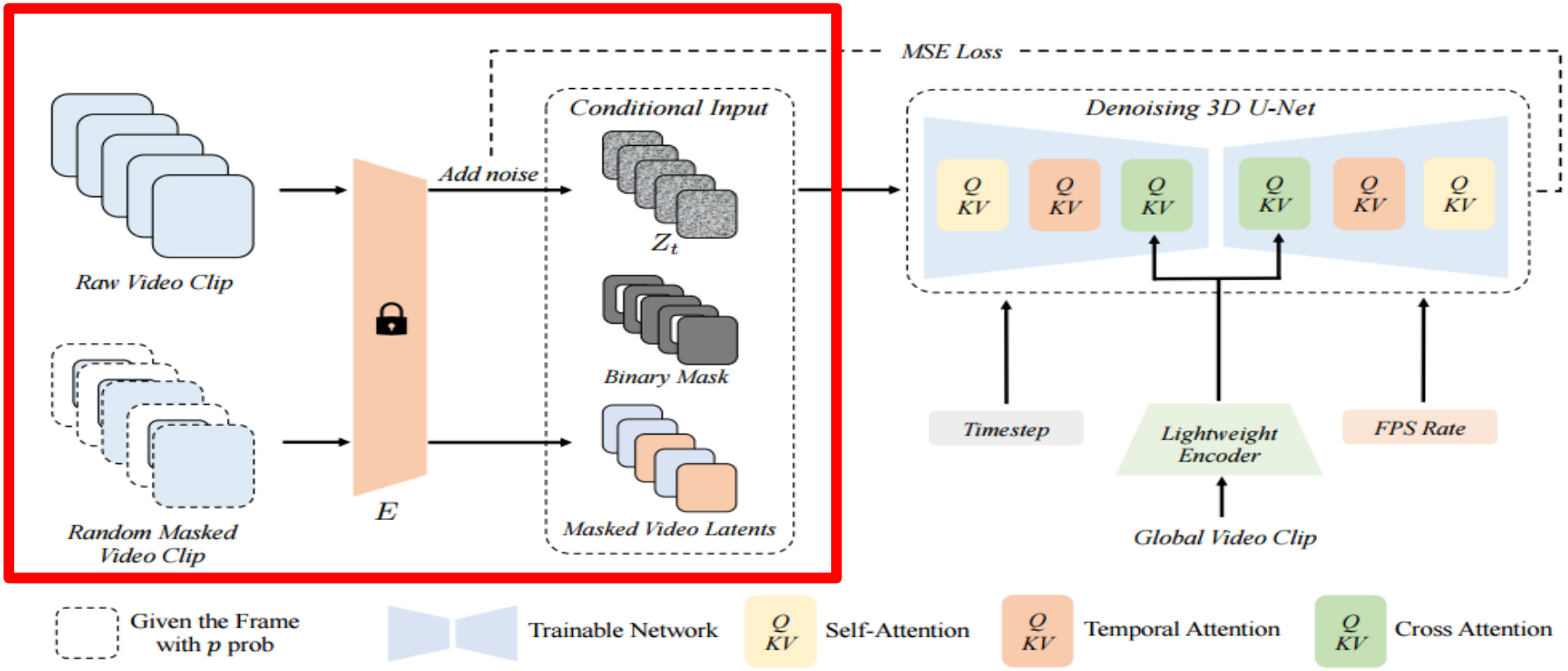
- Based Method
  - LDMs (Latent Diffusion Models)



  - LDMs의 장점
    - 비디오 프레임을 픽셀 공간이 아닌 latent space에 인코딩하여 메모리를 적게 필요로하고 더 나은 효율성을 달성
    - 사전 훈련된 LDMs는 natural image의 콘텐츠와 구조에 대한 좋은 사전 정보를 제공합니다. 모델이 비디오 아웃페인팅 작업에서 빠르게 수렴하는 데 도움이 됨.

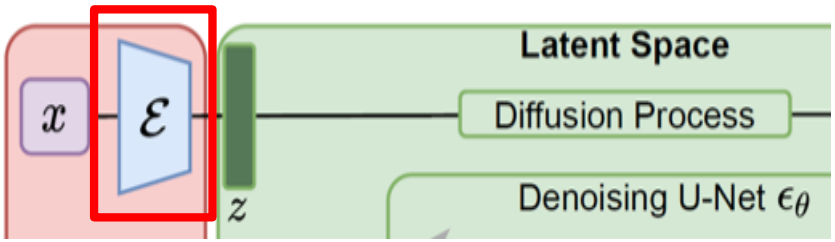
Video Outpainting

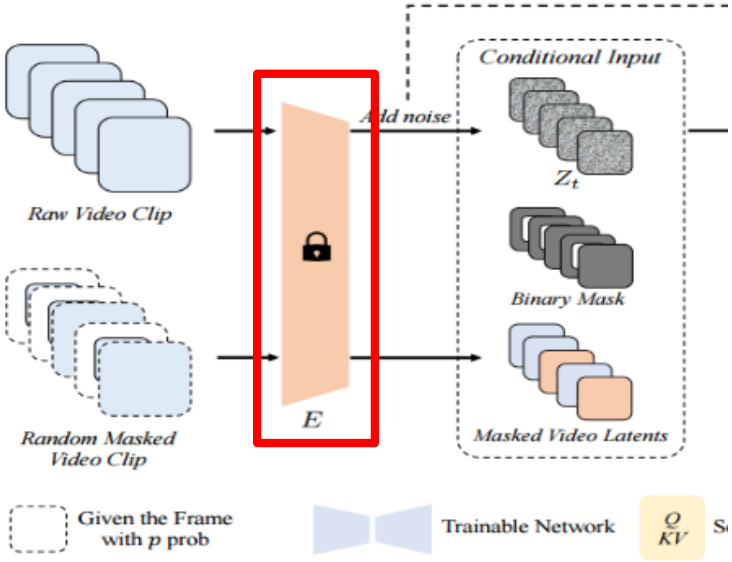
- Our Models
  - Masked 3D diffusion models



Video Outpainting

- Our Models
  - Masked 3D diffusion models





비디오 프레임을 픽셀 공간에서 잠재 공간으로 사전 훈련된 인코더 E를 통해 매핑

- 훈련 단계에서 각 컨텍스트 프레임은 인코더 E에 공급되기 전에 확률 p_frame으로 원시 비디오 프레임으로 교체

Video Outpainting

- Our Models
  - Masked 3D diffusion models
  - 변경사항 : 이전에는 모델이 훈련되는 동안 컨텍스트 프레임을 사용하여 원시 비디오 프레임을 예측하는 것이 일반적이었지만 컨텍스트 프레임 대신에 원시 비디오 프레임을 사용하여 훈련하는 것을 제안함으로써 모델은 노이즈가 있는 원시 비디오 프레임을 예측하도록 훈련됩니다.
  - 변경 사항이 가지는 두 가지 이점
    - 우리의 세부-미세 추론 파이프라인을 가능하게 하여 multiple pass에서 일관된 추론 시간을 보장합니다.
    - 첫 번째 또는 마지막 원시 프레임만 입력 조건으로 사용하는 대신, 양방향 학습은 모델이 컨텍스트 정보를 더 잘 인식하도록 돕고, 이로써 생성 품질을 향상시킨다.

Video Outpainting

- Our Models
  - Masked 3D diffusion models
    - mask strategy

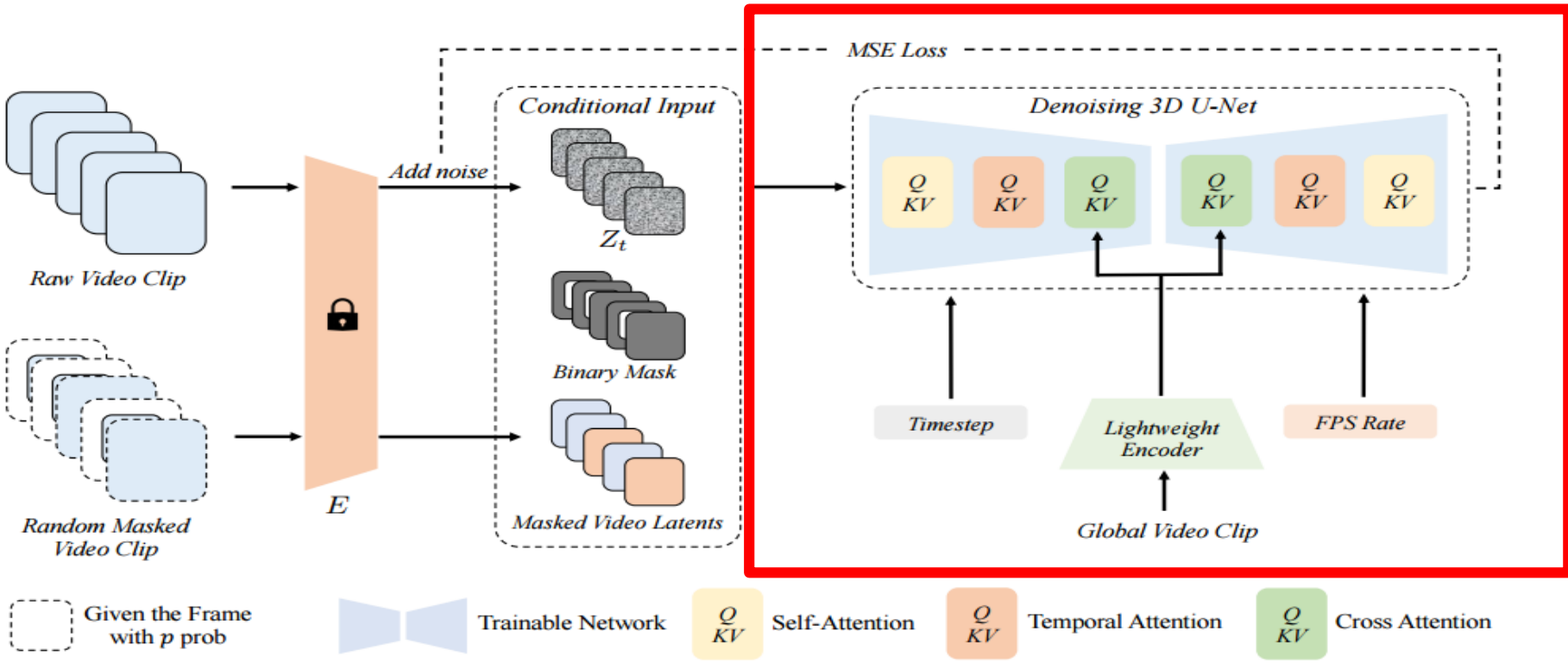


  - fourdirection, single-direction, bi-direction (left-right or top-down), random in any of four directions, and mask all
  - mask all(모두 마스킹)은 무조건적으로 생성될 수 있도록 설정

⇛모두 마스킹(모든 프레임)을 하면 어떠한 외부정보도 없이 아웃페인팅을 할 수 있어야 하기에 mask all 상황을 학습하며 여러 상황에 대응할 수 있게 되고 레이블이 주어지지 않아도 되는 Classifier-free Guidance(diffusion 모델에 좋음) 이점을 얻을 수 있음.

Video Outpainting

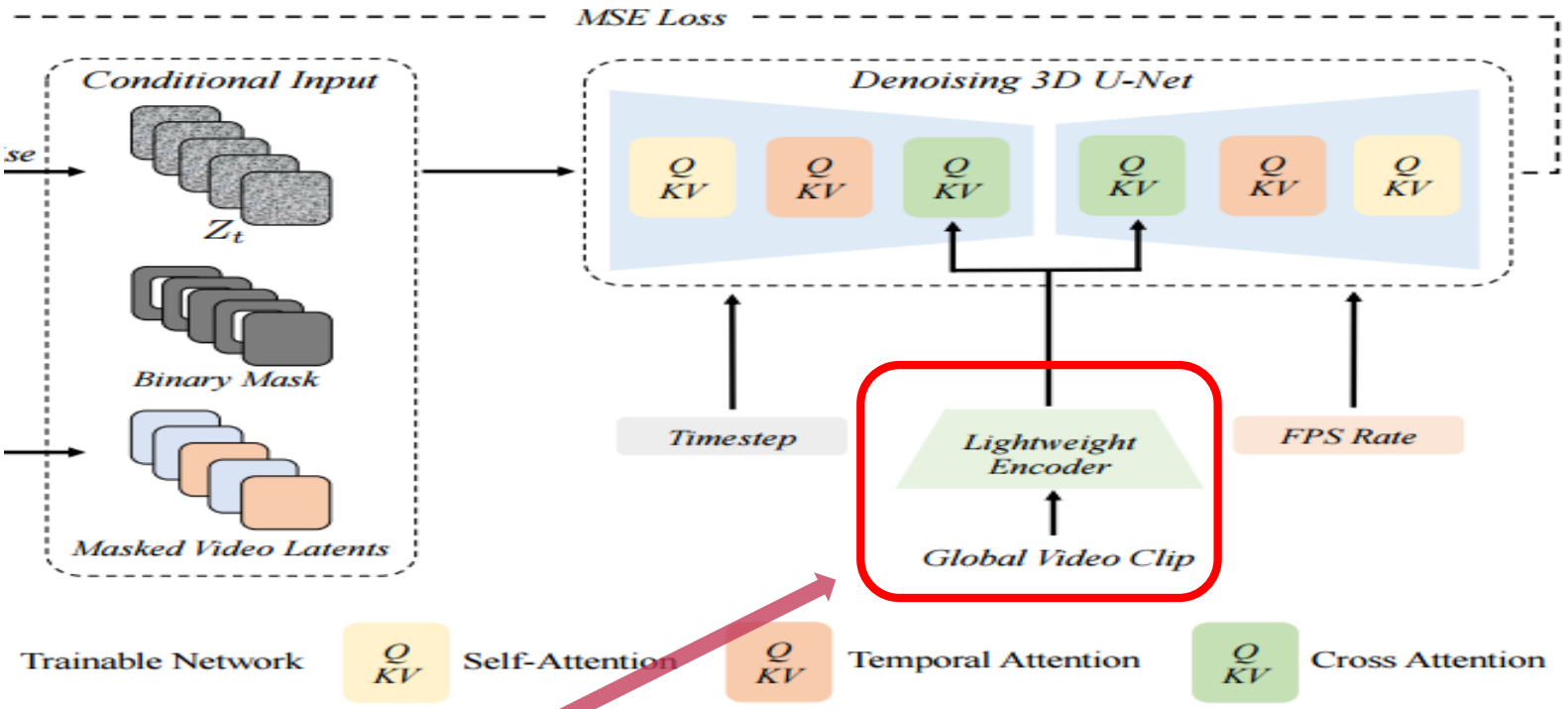
- Our Models
  - Masked 3D diffusion models



- FPS(Frames per second) : 비디오별로 FPS가 다르기 때문에 정보를 입력받아 비디오에 맞게 frame을 샘플링
- Time step : 각 프레임 간의 상대적인 시간적 위치를 나타내는 값으로 모델에게 각 프레임의 순서와 간격을 알려주어 비디오 시퀀스의 동적인 특성을 이해하는 데 도움을 줌

Video Outpainting

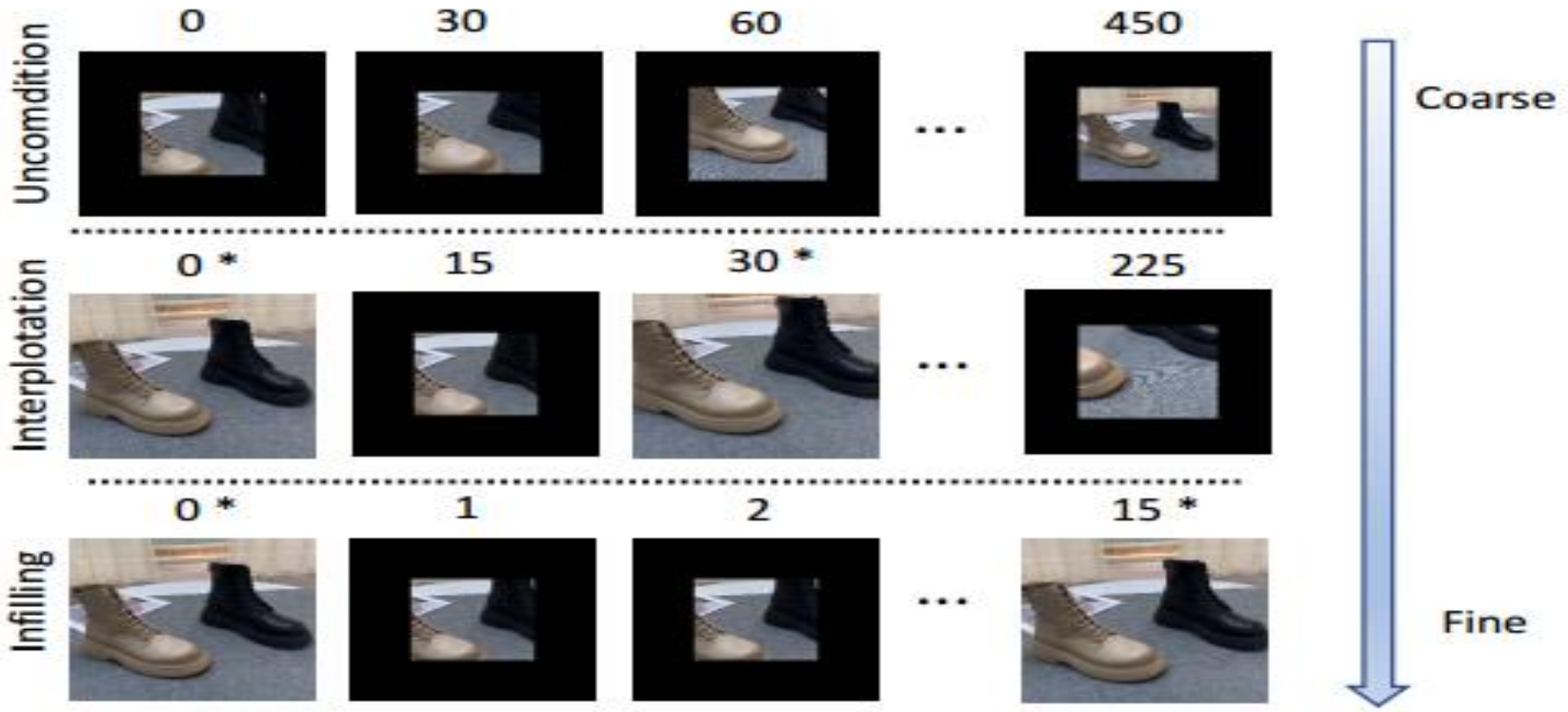
- Our Models
  - Masked 3D diffusion models



모델이 현재 클립을 넘어 Global 비디오 정보를 인식할 수 있도록하기 위해 비디오에서 g 프레임을 균일하게 샘플링

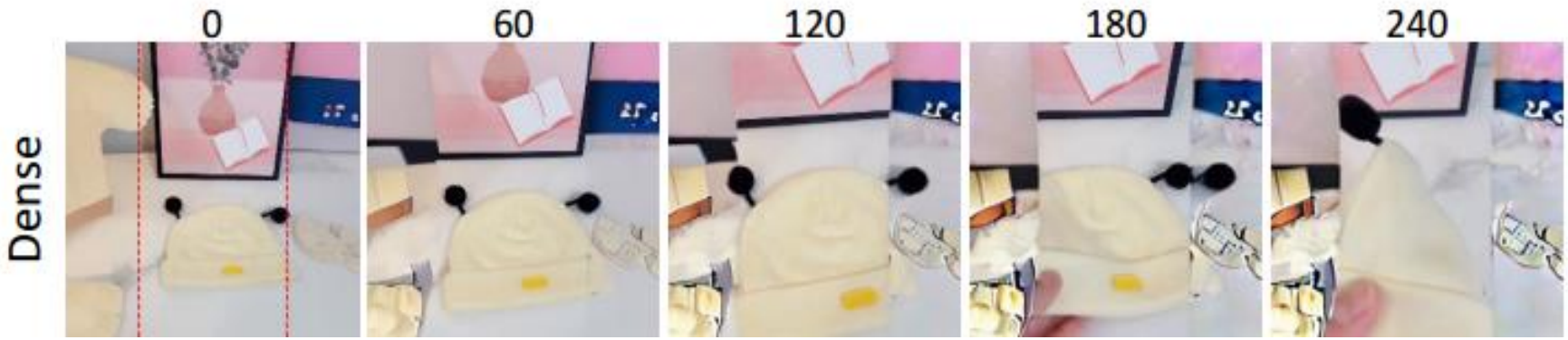
Video Outpainting

- Our Models
  - Masked 3D diffusion models
    - Hybrid Coarse-to-Fine Pipeline



Video Outpainting

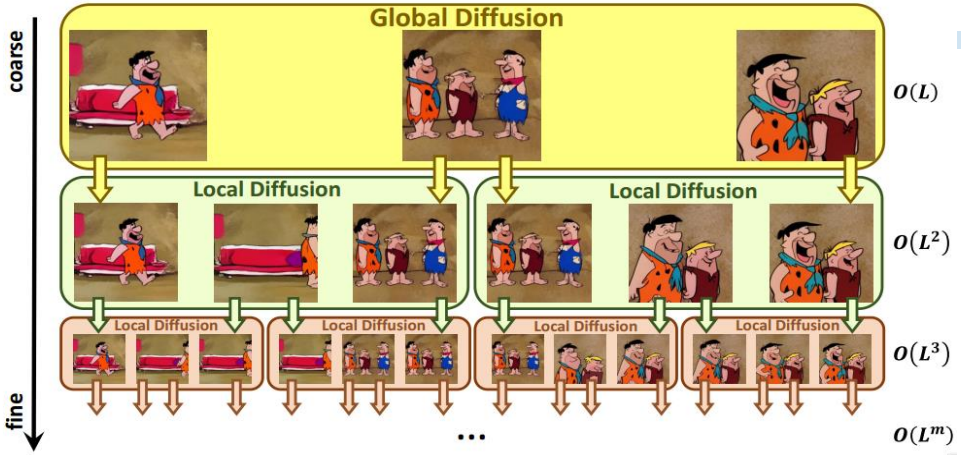
- Our Models
  - Masked 3D diffusion models
    - **Hybrid Coarse-to-Fine Pipeline**
    - 비디오 생성 작업에서 장기 비디오의 생성은 종종 Artifact가 축적되어 성능이 저하되는 결과를 낳음



  - Artifact ?
    - 본 데이터나 영상에서 예상하지 못한 형태의 노이즈, 왜곡, 또는 불규칙한 패턴으로 나타나는 defect
    - 장기간의 동영상을 생성할 때 각 프레임 간의 시간적 일관성을 유지하는 것이 어려울 수 있어 각 프레임에서 나타나는 차이가 누적되어 Artifact를 생성할 수 있다.

Video Outpainting

- Our Models
  - Masked 3D diffusion models
    - **Hybrid Coarse-to-Fine Pipeline**
    - Artifact 축적문제를 해결하기 위해 최근에 Coarse-to-fine 기법이 연구되고 있다.



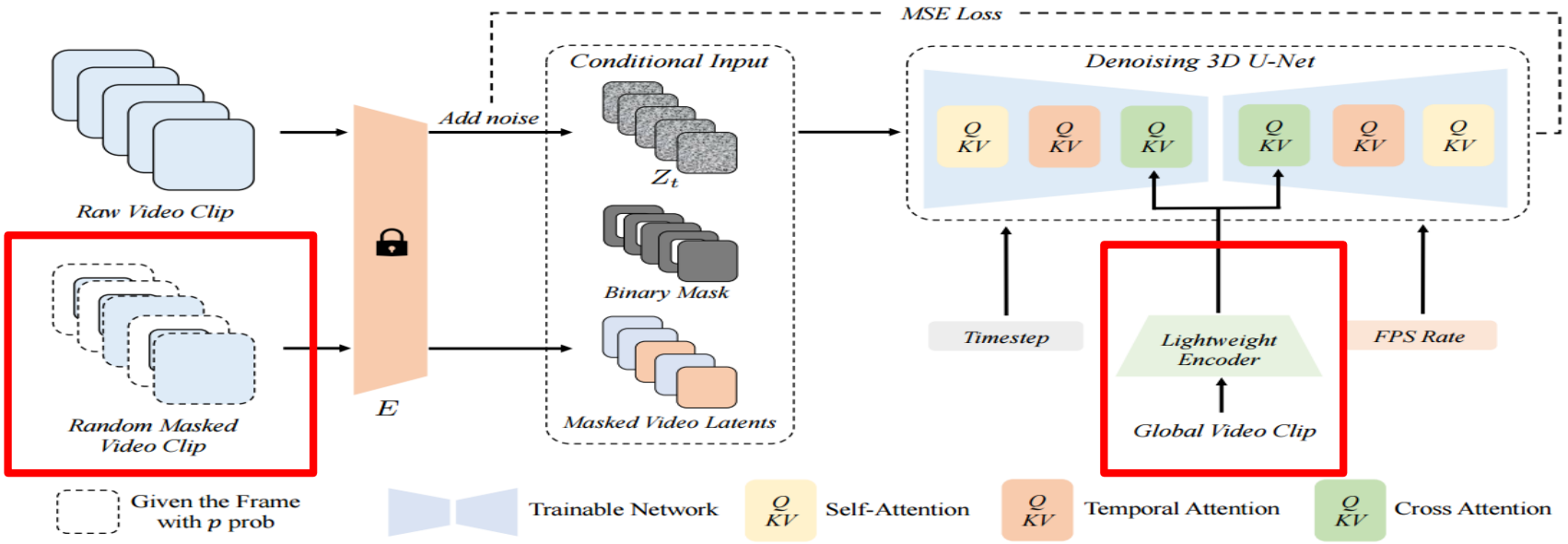
- 기존의 Coarse-to-Fine 방법은 비디오의 sparse한 키프레임(비디오 전체의 중요한 정보나 변화를 담은 프레임)을 생성한 후, 그 후에 infilling 전략을 사용하여 밀도가 높은 비디오 프레임을 채우는 계층적 구조를 사용, 첫 번째와 마지막 프레임을 가이드 프레임으로 사용하는데 infilling만 사용하면 초기 단계의 프레임 간의 간격이 크게 발생하여 데이터 부족문제나 Artifact가 발생하기 쉬움



- 이러한 Artifact 축적 문제를 해결하기 위해 Hybrid Coarse-to-Fine Pipeline방식을 사용

Video Outpainting

- Our Models
  - Masked 3D diffusion models
    - Hybrid Coarse-to-Fine Pipeline
      - infilling만 사용했을 때 문제점을 해결하기 위해 infiiling과 interpolation을 결합한 Hybrid 방법을 제시
      - ⇛두 strategy를 결합한 방법이 가능한 이유 = 양방향 학습!!

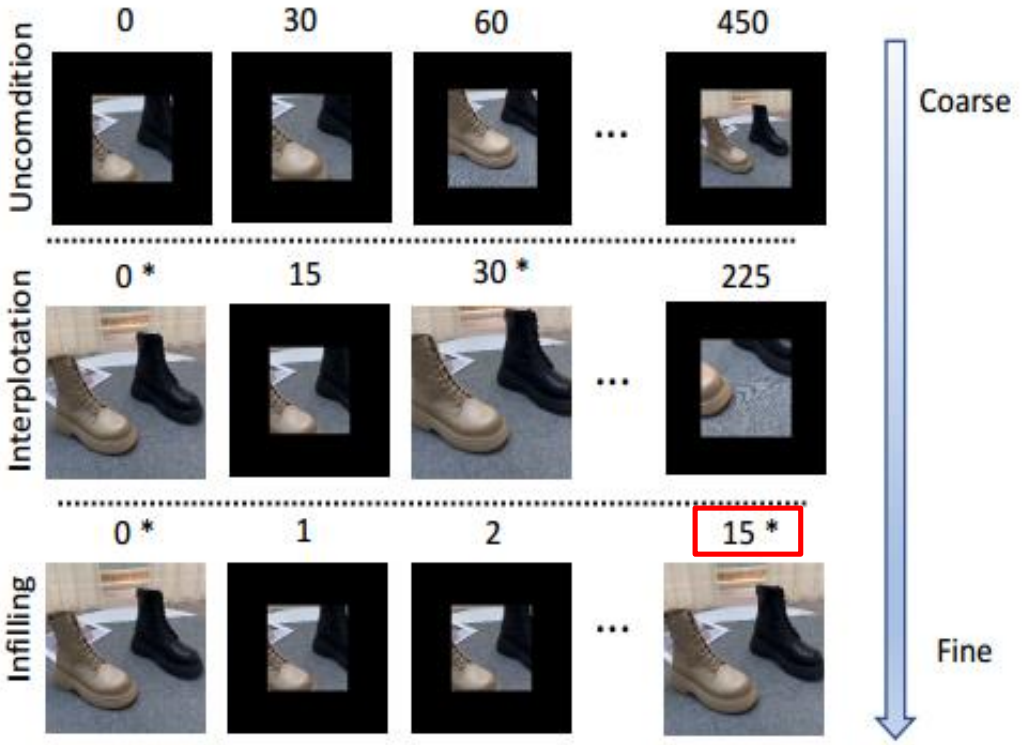


모델이 손상된 영역을 채우는 데 필요한 정보를 양방향으로 공유

모델이 전체 비디오의 Context를 고려하도록 도와주며, 앞뒤의 프레임 뿐만 아니라 비디오의 전반적인 특성을 파악하는 데 도움이 됨

Video Outpainting

- Our Models
  - Masked 3D diffusion models
    - Hybrid Coarse-to-Fine Pipeline



첫 번째 단계에서는 첫 번째 비디오 클립을 조건 없이 생성한 다음, 이전 반복의 마지막 프레임 결과를 기반으로 모든 키프레임을 반복적으로 생성합니다.

두 번째 단계에서는 첫 번째 단계에서 생성된 키프레임을 조건부 입력으로 사용하여 interpolation을 통해 더 많은 키프레임을 생성합니다.

세 번째 단계에서는 첫 번째와 마지막 프레임을 가이드 프레임으로 사용하여 프레임 간격이 1인 최종 비디오 아웃페인팅 결과를 생성

Video Outpainting

- Output
  - Masked 3D diffusion models



Video Outpainting

- Output
  - Masked 3D diffusion models
    - 다른 모델들과 비교해 봤을 때 temporal consistency가 일정하고 Artifact문제가 덜 일어남



Video Outpainting

- Contribution
  - Masked 3D diffusion models

- 1. 비디오 아웃페인팅에 masked 3D diffusion model을 처음 사용하여 최신 결과를 달성한 것으로 알고 있다.
- 2. mask modeling과 양방향 학습 방법을 제안하여 우리의 3D 확산 모델을 훈련시켰습니다. 또한, 동일한 비디오의 서로 다른 클립을 연결하기 위해 가이드 프레임을 사용하면 높은 시간적 일관성과 낮은 흔들림을 가진 video outpainting 결과를 효과적으로 생성할 수 있음을 보였습니다.
- 3. 비디오의 글로벌 프레임에서 프롬프트로서 글로벌 시간 및 공간 정보를 추출하고, 이를 교차 어텐션 형식으로 네트워크에 입력하여 모델이 더 합리적인 결과를 생성하도록 안내합니다.
- 4. sparse한 프레임을 생성할 때 infilling과 interpolation을 결합하는 hybrid coarse-to-fine generation pipeline을 제안합니다. 실험 결과, 제시한 모델의 파이프라인은 장기 비디오 아웃페인팅에서 artifact 누적을 줄이면서도 좋은 수준의 시간적 일관성을 유지할 수 있음을 보여줍니다.

Video Outpainting

The end