Paper seminar

2025-08-22

Department of Metaverse Convergence at Chung-Ang University
Myungjun Yun

Background

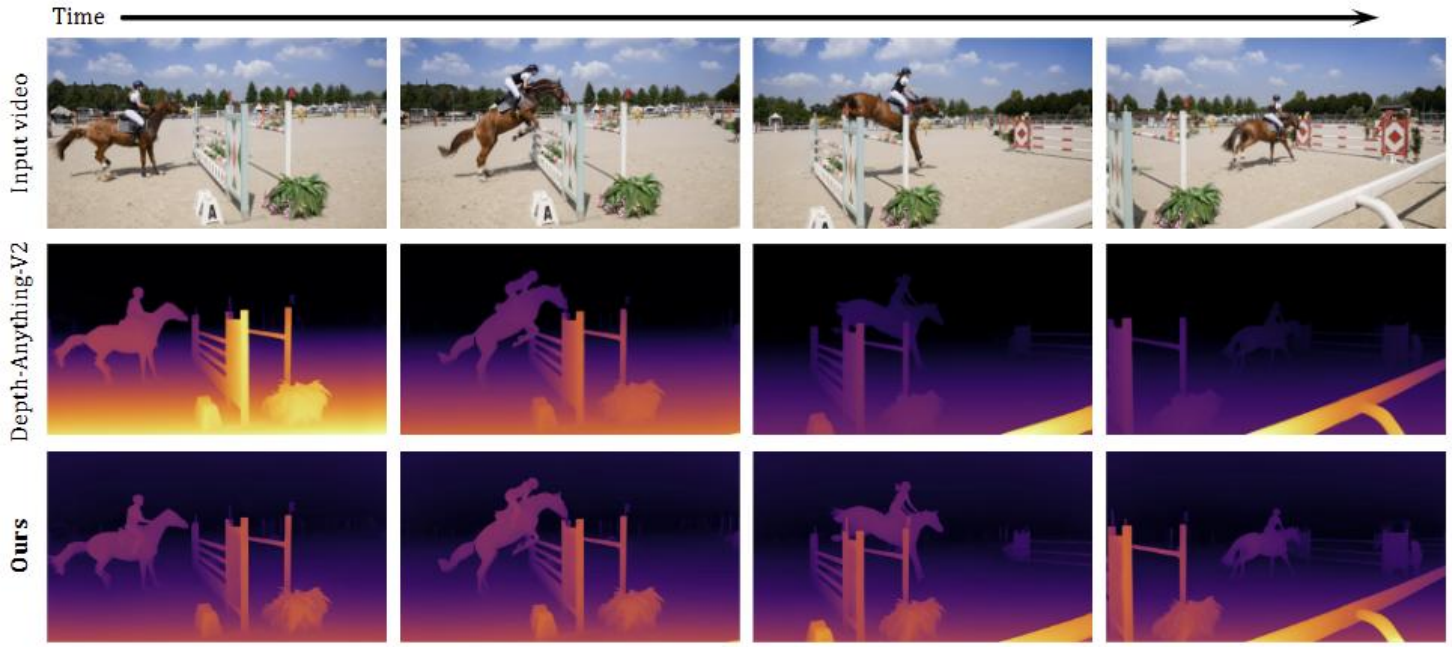
Monocular depth estimation

- **Single view**를 입력으로 받아 2D image에서 **3D depth value를 알아내는 task**
- Metaverse, Autonomous driving, Robotics등 다양한 분야에서 응용
- 최근에는 Foundation model의 발전으로 인해 높은 성능을 달성

DepthCrafter

Background

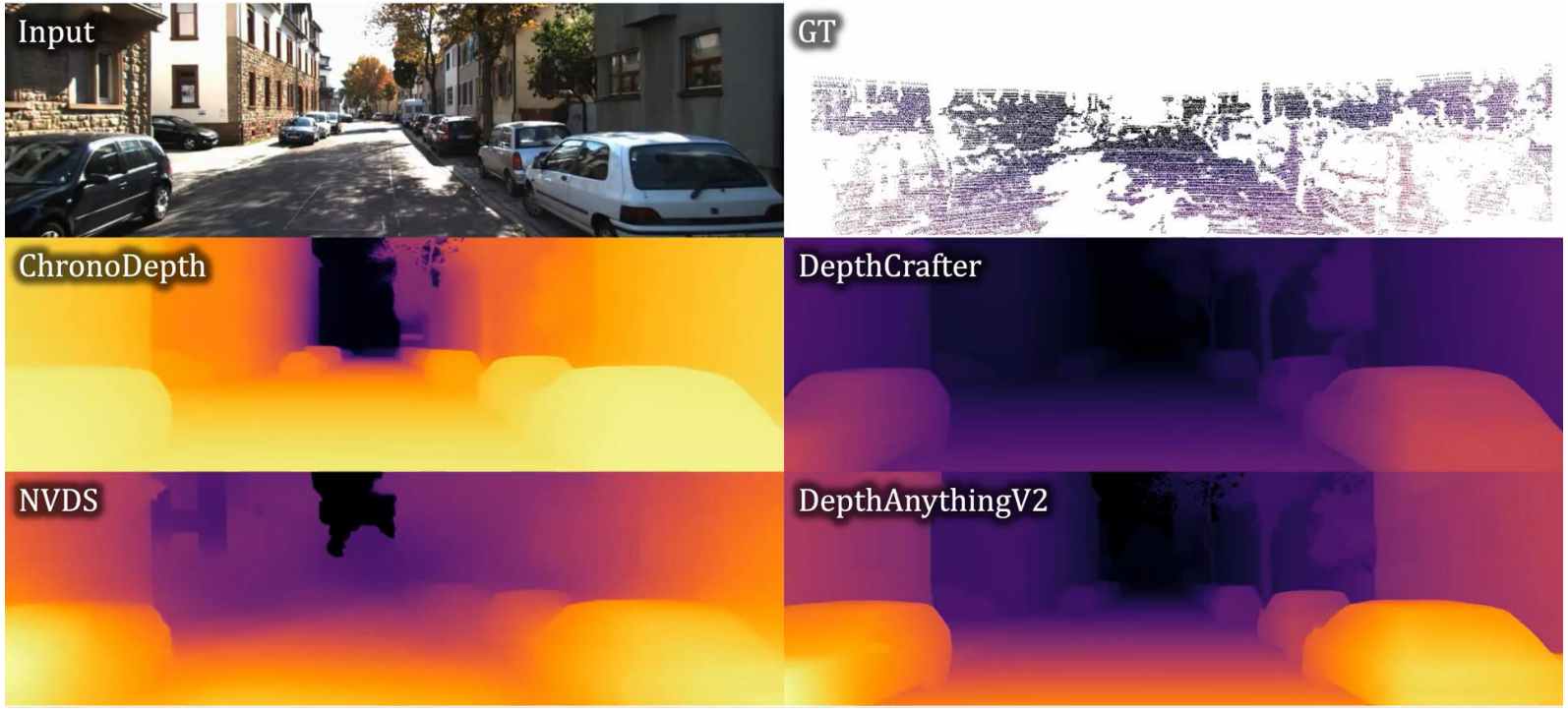
- Open-world scenarios에서 Video depth estimation은 어려움(객체의 움직임, 길이)
- 기존의 방법은 Video**의 시간적 정보 고려를 하지 않아 Flickering** 발생
- Camera movement, long sequences, motion 같은 추가 정보 없이 depth sequence를 생성하는 방법 필요



DepthCrafter

Purpose

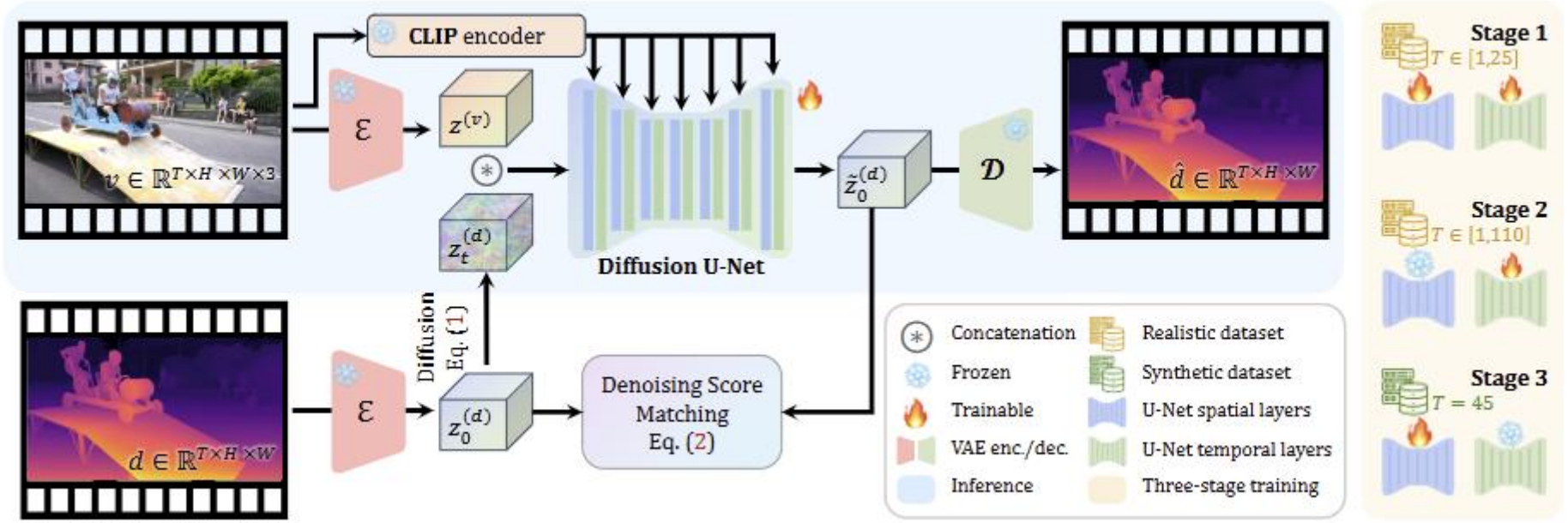
- **Detail & 시간적으로 일관된** long depth sequences를 생성
- 길고 가변적인 시간적 맥락을 가진 long depth sequences 생성을 가능하게 하는 **3단계 훈련 전략**
- 110 frames 초과 비디오는 **Segment 단위로 처리 후 매끄럽게 연결하는** 추론 전략



DepthCrafter

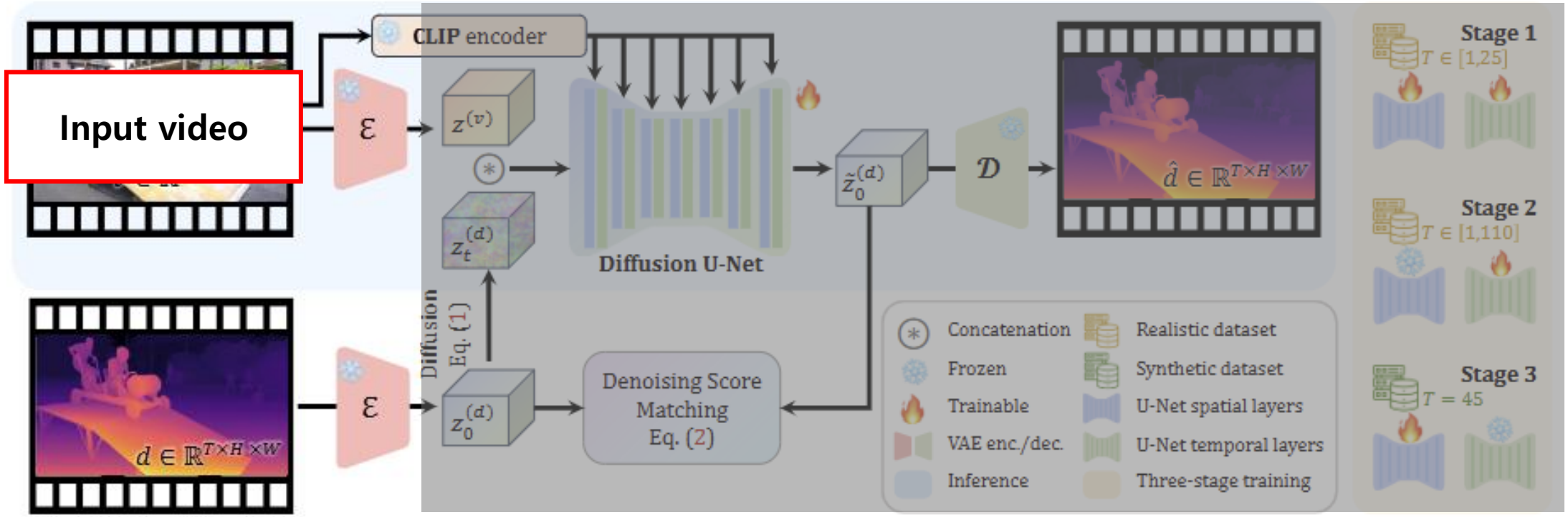
Method

- **Detail & 시간적으로 일관된** long depth sequences를 생성
- 길고 가변적인 시간적 맥락을 가진 long depth sequences 생성을 가능하게 하는 **3단계 훈련 전략**
- 110 frames 초과 비디오는 **Segment 단위로 처리 후 매끄럽게 연결하는** 추론 전략



DepthCrafter

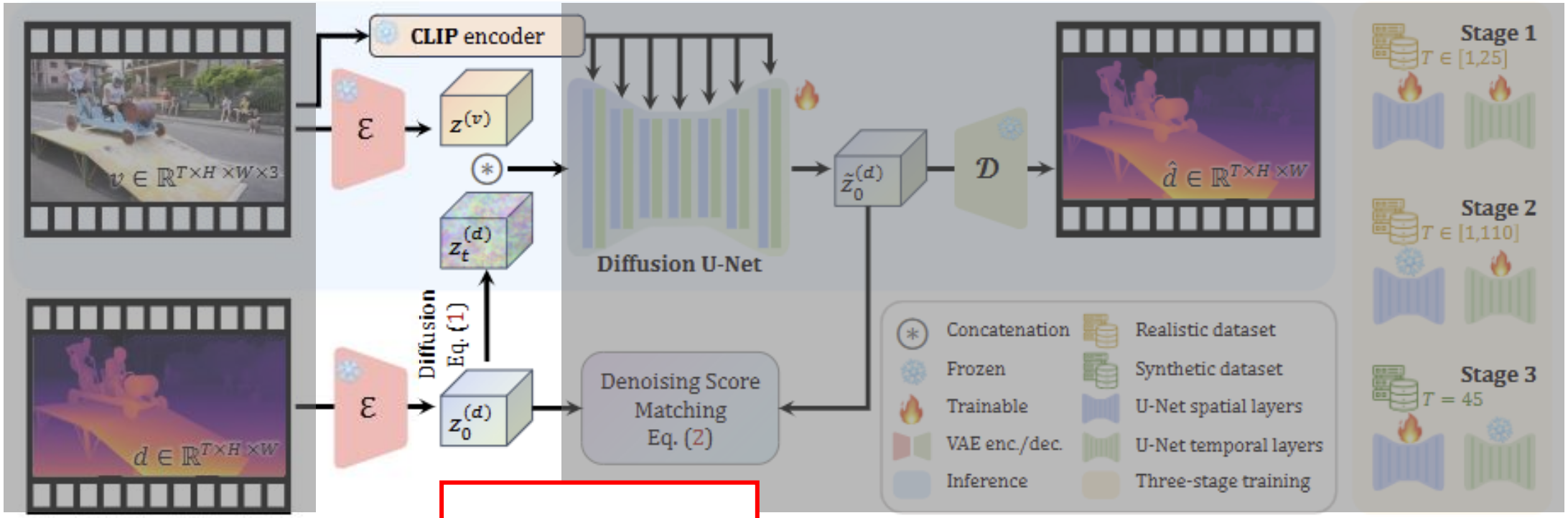
Method



DepthCrafter

Method

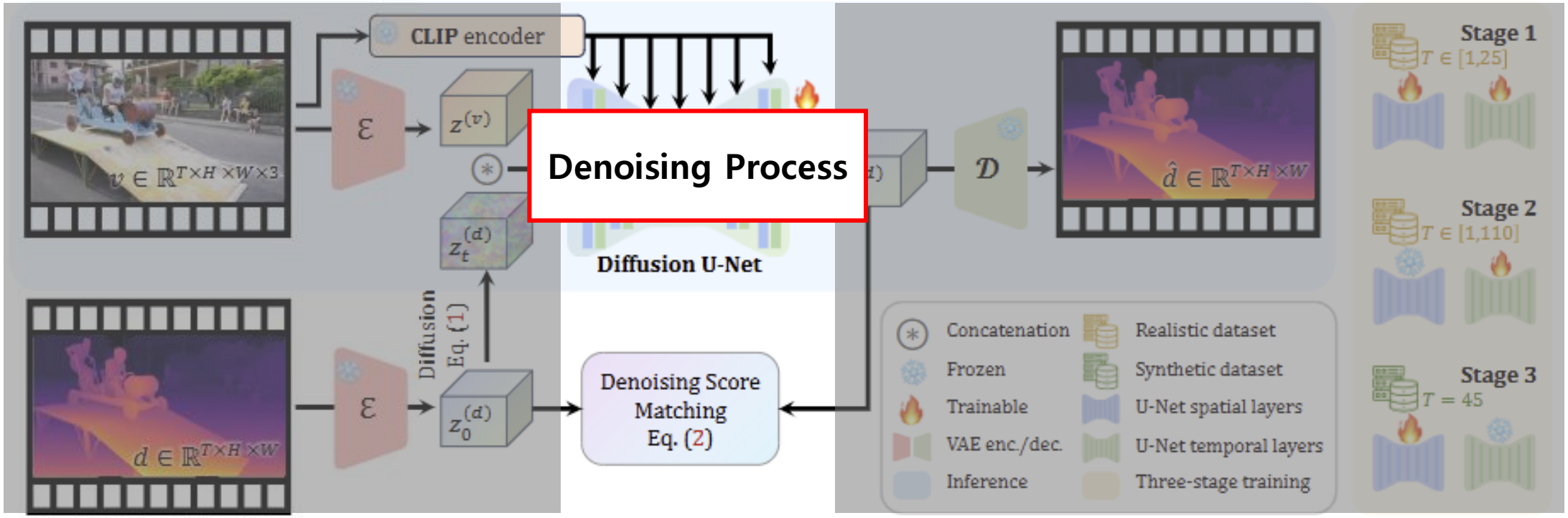
Encoding & Condition Information Extraction



Diffusion process

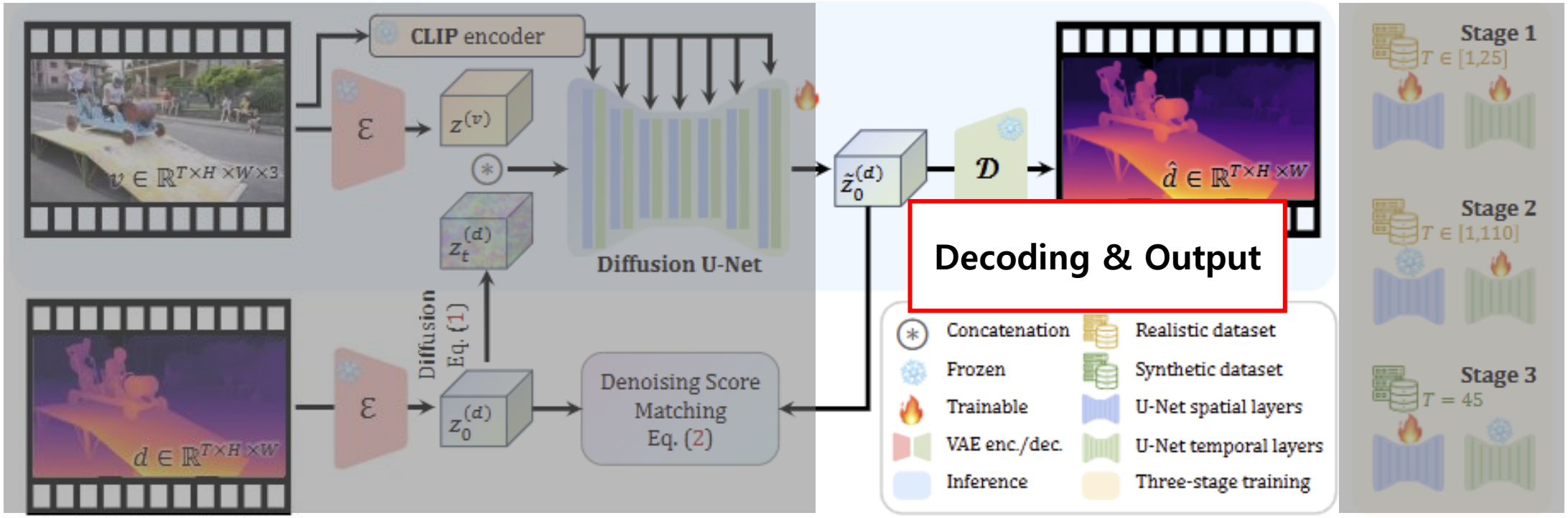
DepthCrafter

Method



DepthCrafter

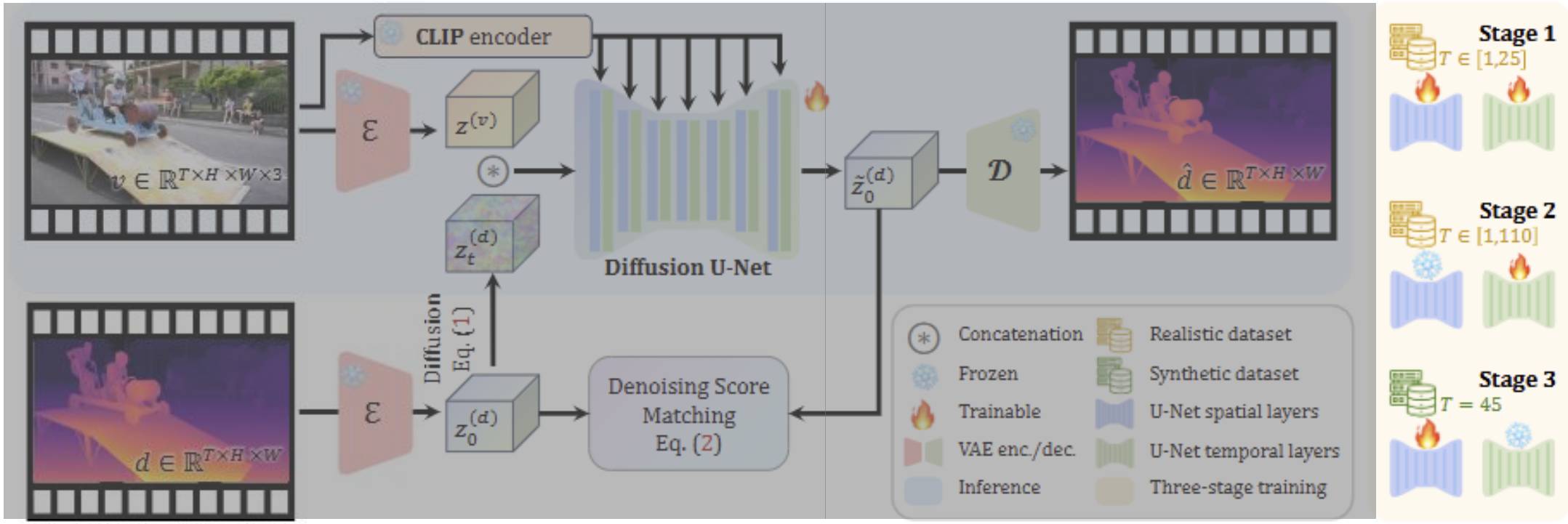
Method



DepthCrafter

Method

Three-stage training



DepthCrafter

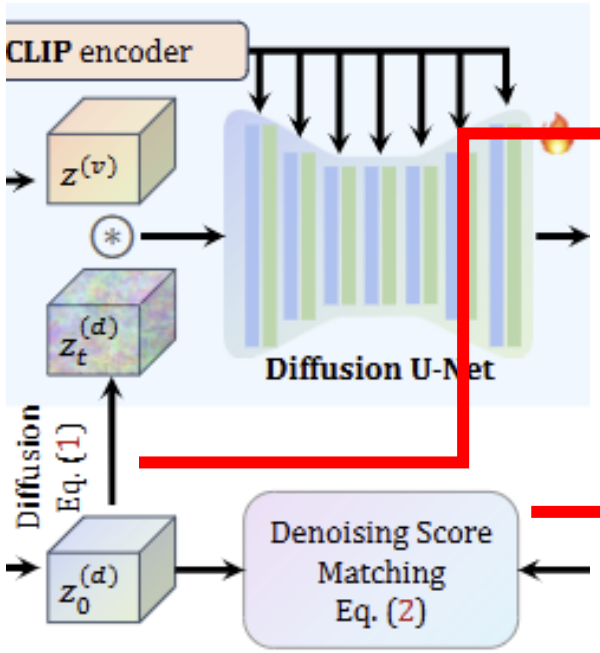
Method

Diffusion model

- SVD(stable video diffusion)
- Video Conditioning

$\boldsymbol{x_t}$: Noised data
$\boldsymbol{x_0}$: raw data
$\boldsymbol{\sigma_t}$:Noise level
$\boldsymbol{\epsilon}$: Normal distribution.
$\boldsymbol{E_{x_t}}$: Expected value.
$\boldsymbol{p(X;\sigma_t)}$: The probability distribution of the noised data x_t given the noise coefficient.
$\boldsymbol{p(\sigma)}$:Distribution sampling noise coefficient during training

$\boldsymbol{\lambda_{\sigma_t}}$: Weighting based on noise coefficient
$\boldsymbol{D_\theta}$: Denoiser function
$\boldsymbol{c}$: Condition
$\boldsymbol{c_{skip}}$: Adjust the strength of skip connections based on the noise level.
$\boldsymbol{c_{out}}$: Scale the output of F_θ
$\boldsymbol{c_{in}}$: Adjust data x_t with added noise in the input to U-net.
$\boldsymbol{c_{noise}}$: Mapping noise levels to U_net conditioning inputs



$$\mathbf{x}_t = \mathbf{x}_0 + \sigma_t^2 \epsilon, \quad \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

$$\mathbb{E}_{\mathbf{x}_t \sim p(\mathbf{x};\sigma_t), \sigma_t \sim p(\sigma)} \left[\lambda_{\sigma_t} \left\| D_\theta(\mathbf{x}_t; \sigma_t; \mathbf{c}) - \mathbf{x}_0 \right\|_2^2 \right]$$

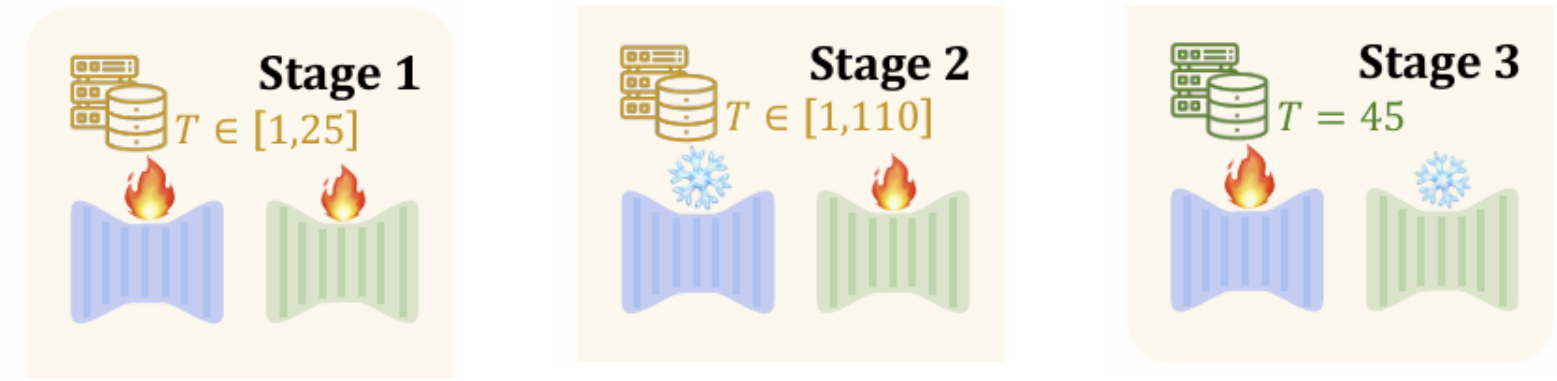
$$D_\theta(\mathbf{x}_t; \sigma_t; \mathbf{c}) = c_{\text{skip}}(\sigma_t)\mathbf{x}_t + c_{\text{out}}(\sigma_t) F_\theta\left(c_{\text{in}}\mathbf{x}_t; c_{\text{noise}}(\sigma_t); \mathbf{c}\right)$$

DepthCrafter

Method

Three-stage training

- 다양한 길이의 Open-world video에 대한 고품질 깊이 시퀀스를 생성
- 가변적인 긴 시간적 문맥의 어려움

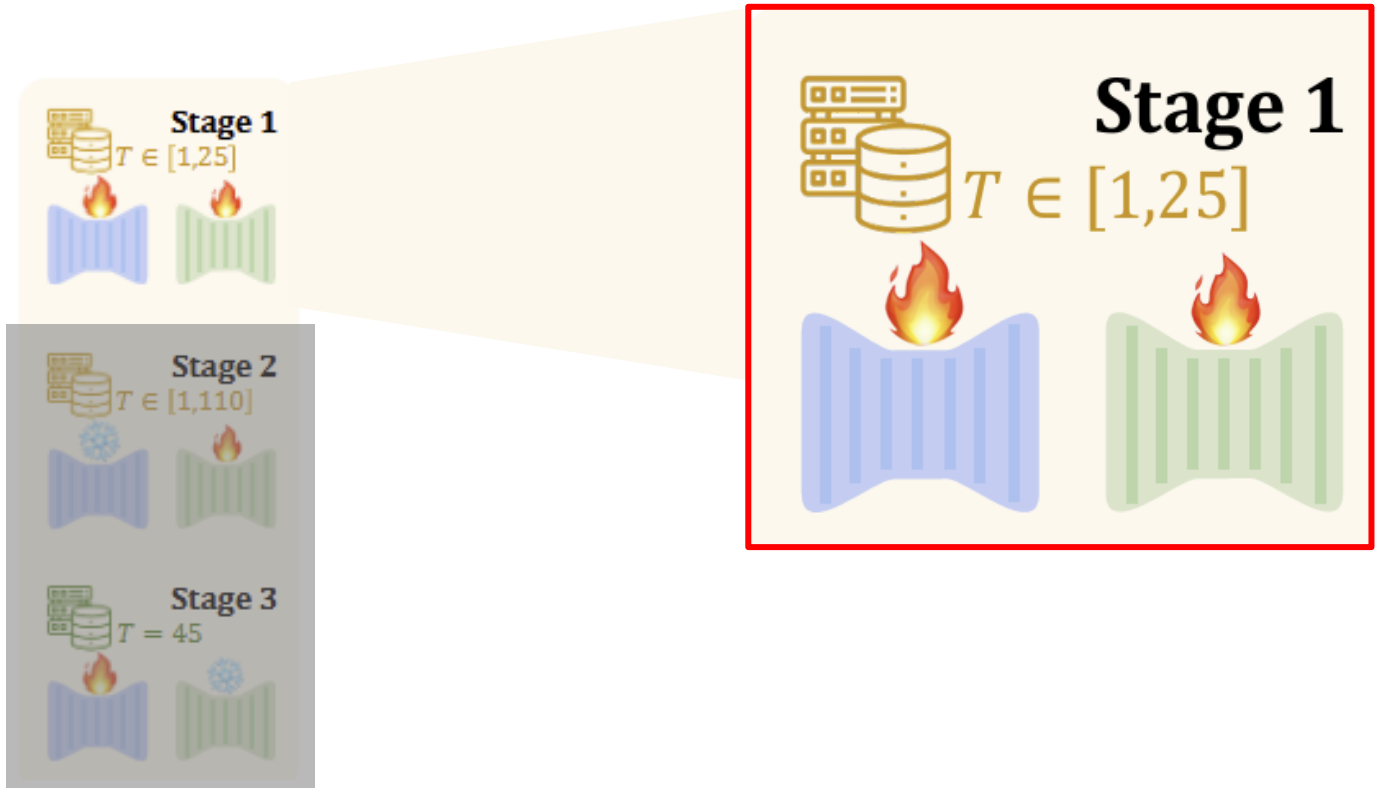


DepthCrafter

Method

Stage 1

- Pre-trained SVD 모델을 비디오-깊이 생성 작업에 적응
- Large realistic dataset

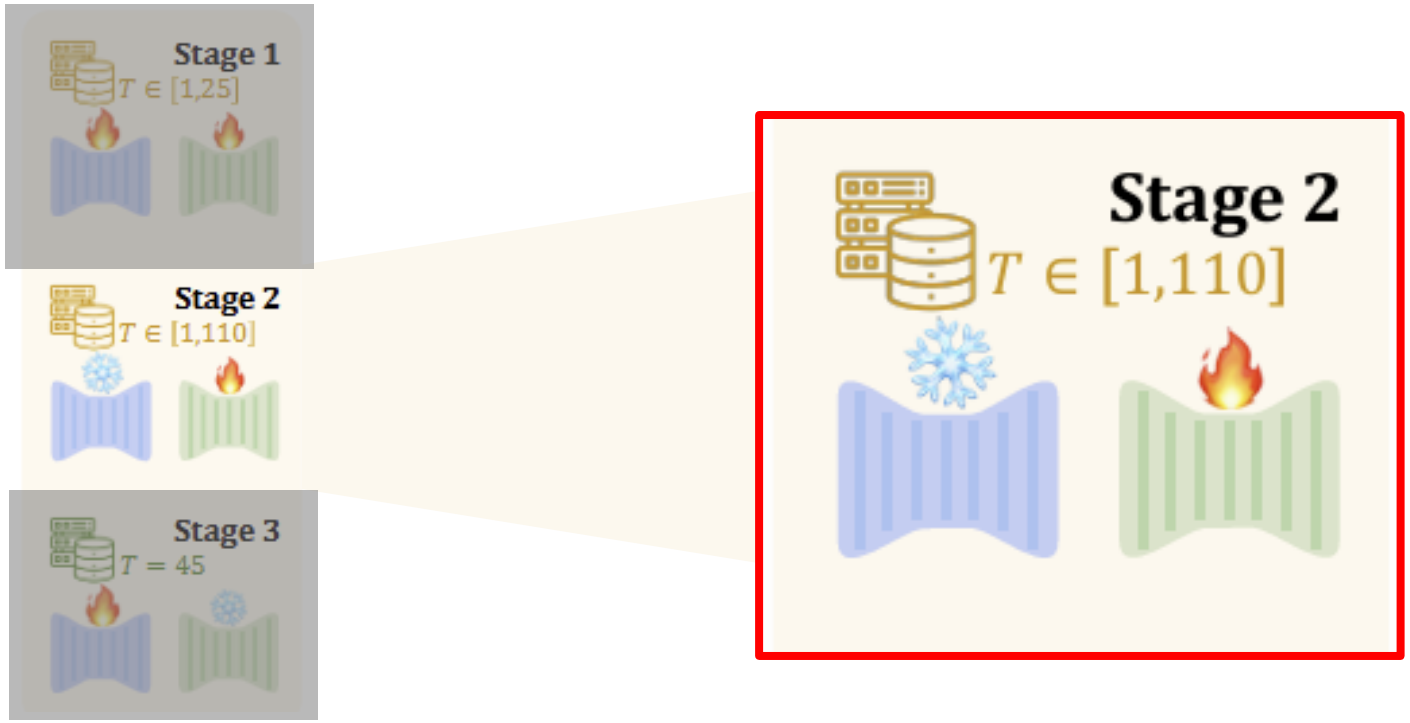


DepthCrafter

Method

Stage 2

- 긴 시간적 문맥을 처리하고, 길고 가변적인 시퀀스에 대한 전체 깊이 분포를 정확하게 배열
- Large realistic dataset

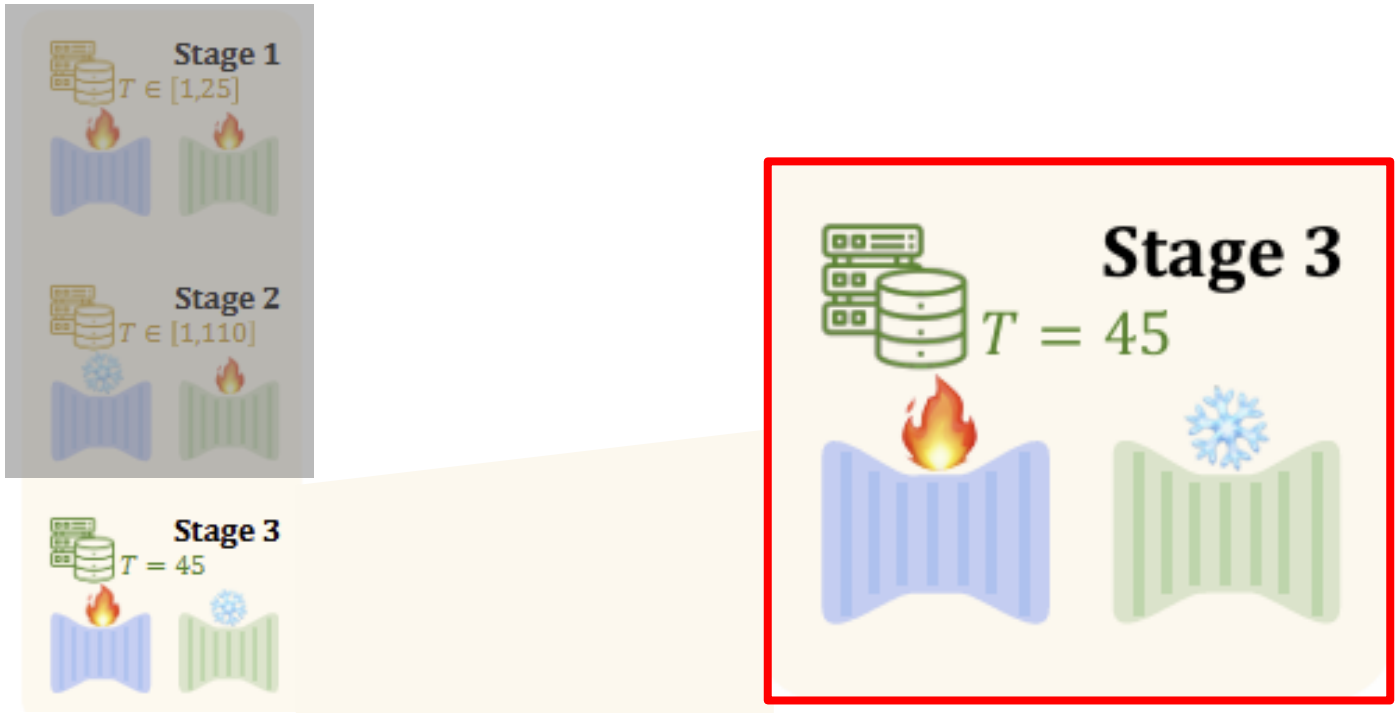


DepthCrafter

Method

Stage 3

- 합성 데이터셋의 장점을 활용하여 더 정밀한 깊이 디테일을 학습
- Small synthetic dataset

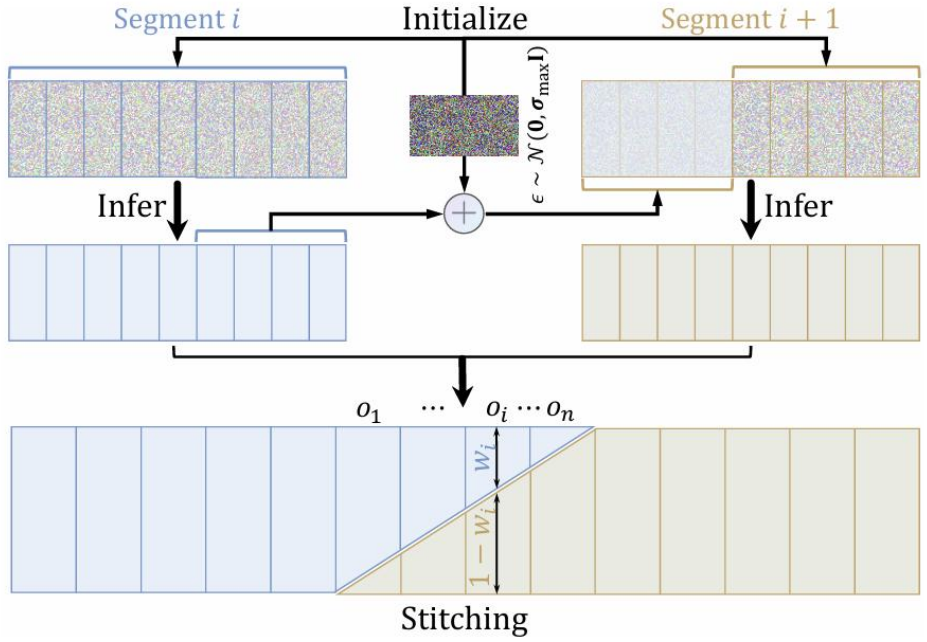


DepthCrafter

Method

Inference for Extremely Long Videos

- 모델이 학습 후 110 Frames 까지 추정할 수 있지만 Open-World Video에서는 여전히 부족
- Long depth sequence를 Segment 방식으로 추론, 연결

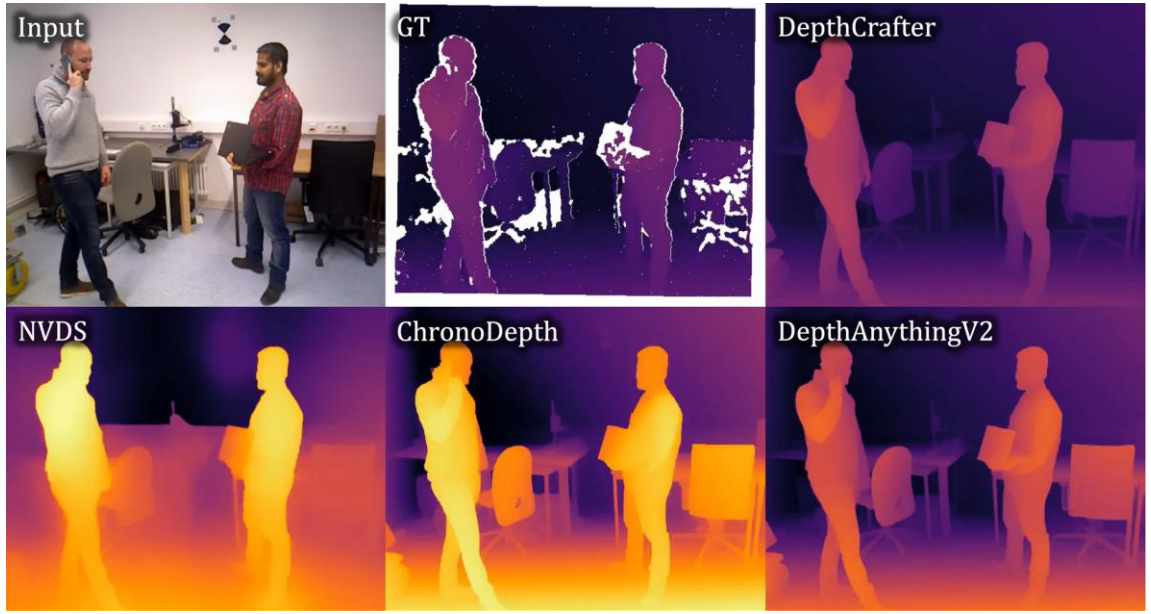


DepthCrafter

Result

Qualitative results

- 다양한 Open-world videos에서 시간적으로 일관된 깊이 시퀀스를 생성



DepthCrafter

Result

Qualitative results

- 여러 데이터셋에서 SOTA 달성
- KITTI, Sintel: Camera motion, Fast-moving object
- Scannet, Bonn: Indoor dataset
- Three-stage training이 진행될수록 높은 성능 달성

	stage 1	stage 2	stage 3
AbsRel ↓	0.322	0.316	**0.270**
δ_1 ↑	0.626	0.675	**0.697**

Method	Sintel (~50 frames)		Scannet (90 frames)		KITTI (110 frames)		Bonn (110 frames)		NYU-v2 (1 frame)	
	AbsRel ↓	δ_1 ↑	AbsRel ↓	δ_1 ↑	AbsRel ↓	δ_1 ↑	AbsRel ↓	δ_1 ↑	AbsRel ↓	δ_1 ↑
NVDS [63]	0.408	0.483	0.187	0.677	0.253	0.588	0.167	0.766	0.151	0.780
ChronoDepth [54]	0.587	0.486	0.159	0.783	0.167	0.759	0.100	0.911	0.073	0.941
Marigold [32]	0.532	0.515	0.166	0.769	0.149	0.796	0.091	0.931	0.070	0.946
Depth-Anything [67]	0.325	0.564	0.130	0.838	0.142	0.803	0.078	0.939	**0.042**	**0.981**
Depth-Anything-V2 [68]	0.367	0.554	0.135	0.822	0.140	0.804	0.106	0.921	0.043	0.978
DepthCrafter (Ours)	**0.270**	**0.697**	**0.123**	**0.856**	**0.104**	**0.896**	**0.071**	**0.972**	0.072	0.948

DepthCrafter

Conclusion

- Video diffusion model을 활용하는 새로운 Open-world video 깊이 추정 방법인 DepthCrafter 제안
- **추가 정보 없이** 다양한 모션 및 카메라 움직임을 가진 비디오에 대해 **디테일과 시간적으로 일관된 깊이 시퀀스**를 생성
- 1 Frame에서 긴 비디오에 이르기까지 **다양한 길이의 비디오 지원**

DepthCrafter

Discussion

- 높은 계산량 및 메모리 소비량
- CLIP 모델이 사전에 학습하지 못한 특정 도메인에서는 낮은 정확도 우려
- dToF 같은 센서 데이터를 사용하지 않아서 거울이나 눈밭 같이 텍스처가 없는 환경에서 작동 우려

Table 3. Inference time per frame (ms) of our model, Depth-Anything (V2), and Marigold, with the resolution of 1024×576.

Method	Encoding	Denoising	Decoding	All
Depth-Anything (V2)	N/A	N/A	N/A	180.46
Marigold	256.40	114.53	699.36	1070.29
DepthCrafter (Ours)	51.85	160.93	253.06	465.84

Inference

DepthCrafter Inference

Input video

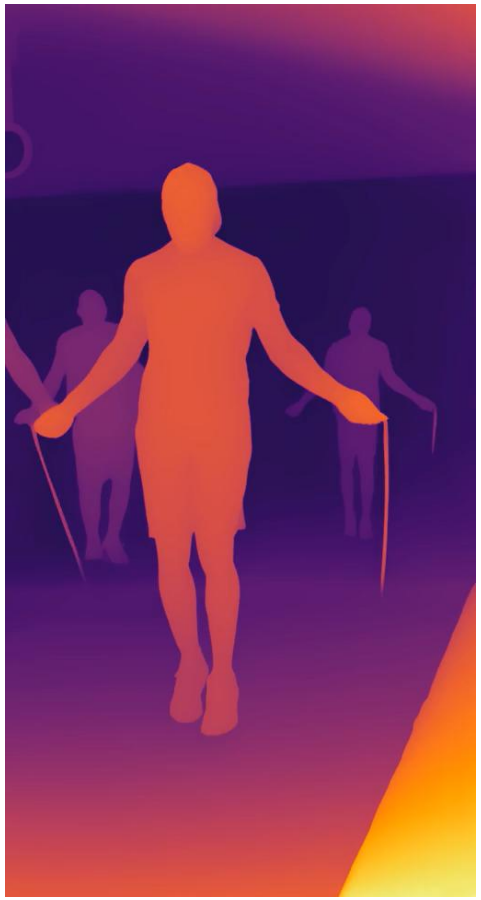
DepthCrafter Output

DepthCrafter Output

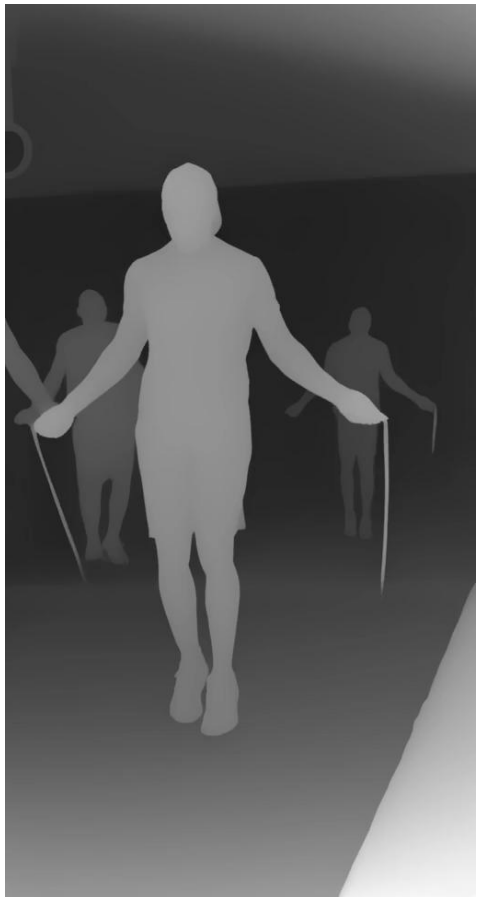
Inference

DepthCrafter Inference

Input video



DepthCrafter Output



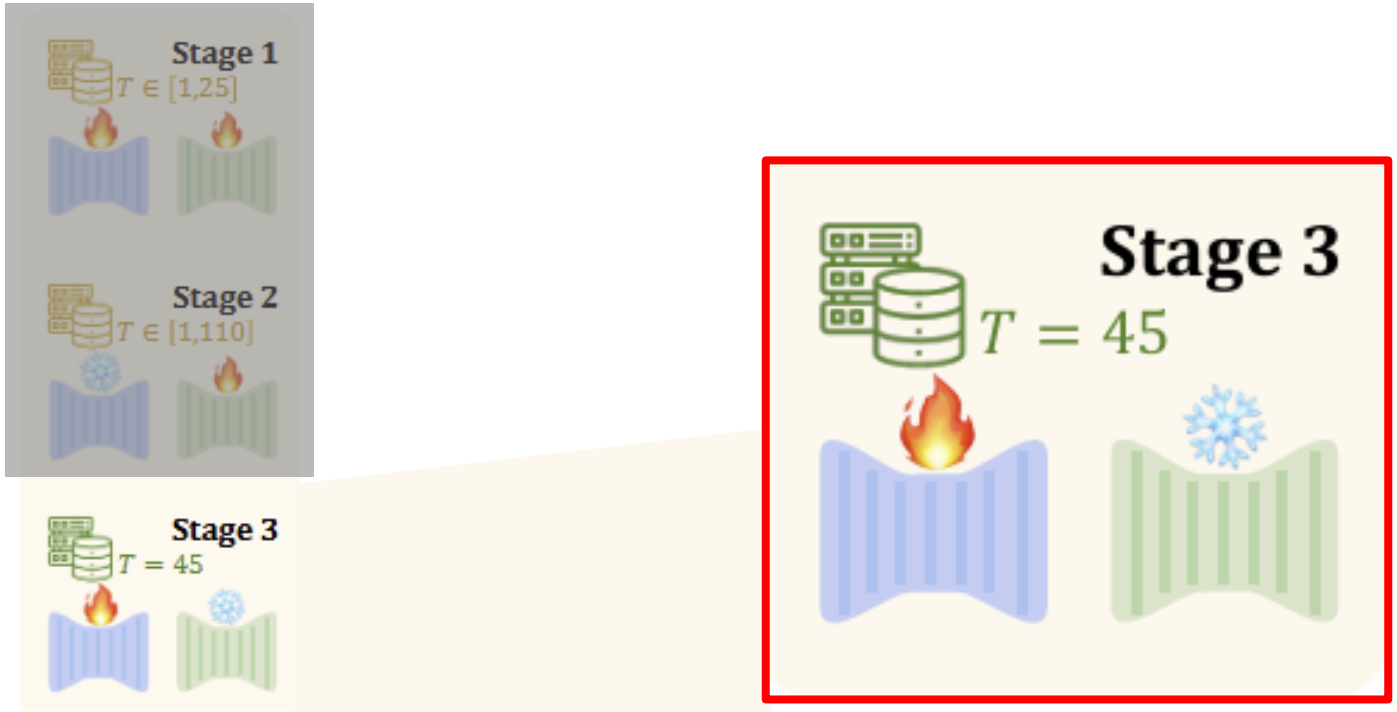
DepthCrafter Output

DepthCrafter

Method

Stage 3

- 합성 데이터셋의 장점을 활용하여 더 정밀한 깊이 디테일을 학습
- Small synthetic dataset → Depth를 추정하기 어려운 dataset 포함?



감사합니다.