SVG: 3D Stereoscopic Video Generation via Denoising Frame Matrix

Peng Dai, Feitong Tan[*], Qiangeng Xu[*], David Futschik, Ruofei Du, Sean Fanello, Xiaojuan Qi, Yinda Zhang

Google, The University of Hong Kong

ICLR 2025

香 港 大 學

THE UNIVERSITY OF HONG KONG

SVG: 3D Stereoscopic Video Generation via Denoising Frame Matrix

Left View

Right View

Given left-view video → Generate right-view video

Motivation

VR headset

Stereoscopic video is highly desirable

Monocular video generation

Stereoscopic/multi-view video generation is under-explored

1. Lack data. Compared to monocular videos, stereoscopic videos are scarce

2. Semantic consistency between left view and right view

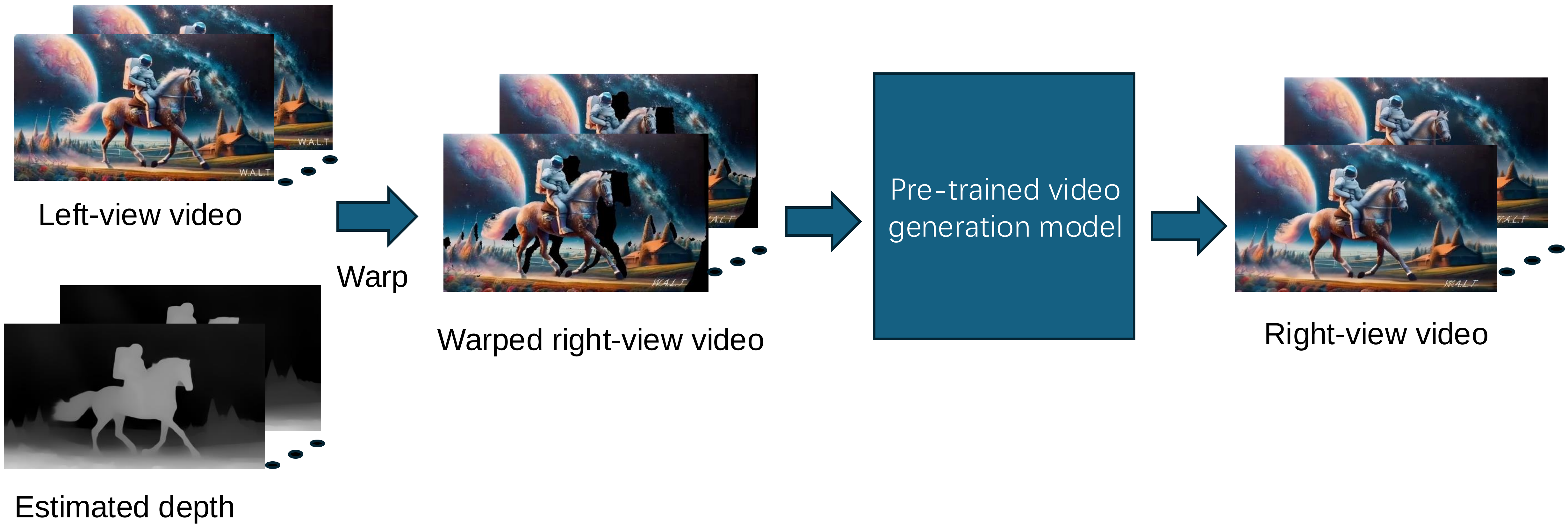
Left view

Warped right view

Inpainted right view

Method

Lack training data => Zero-shot stereoscopic video generation

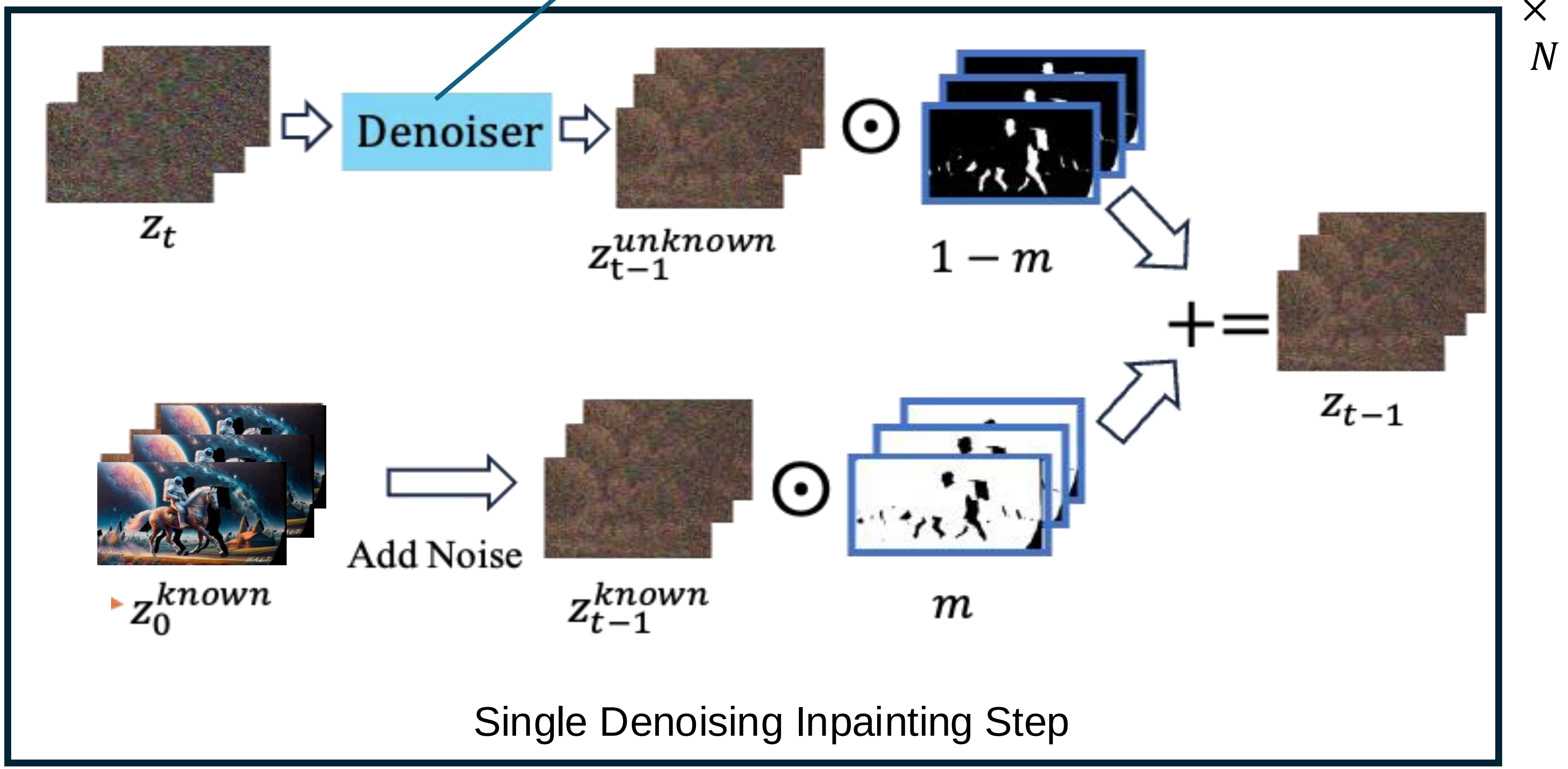


Convert monocular video to stereoscopic video, leveraging pre-trained large video generation model

Method

Details of zero-shot stereoscopic video generation

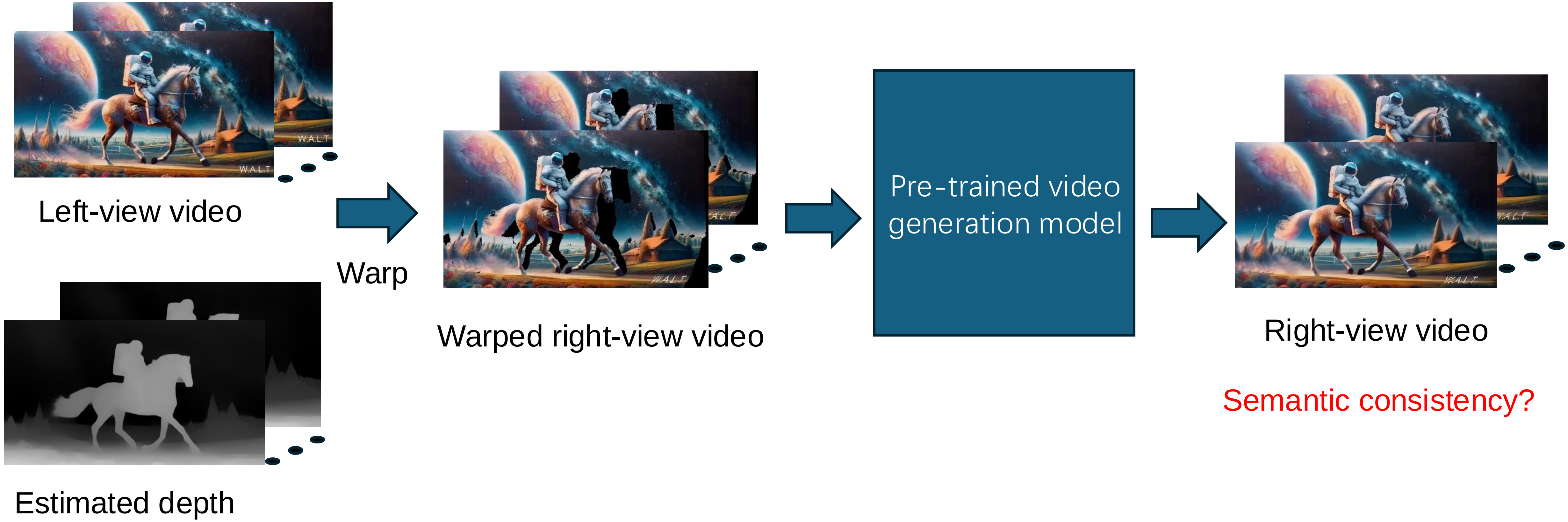
Pre-trained video generation/diffusion model



Fill/generate unknown regions while retaining known regions

Method

Lack training data => Zero-shot stereoscopic video generation



Convert monocular video to stereoscopic video, leveraging pre-trained large video generation model

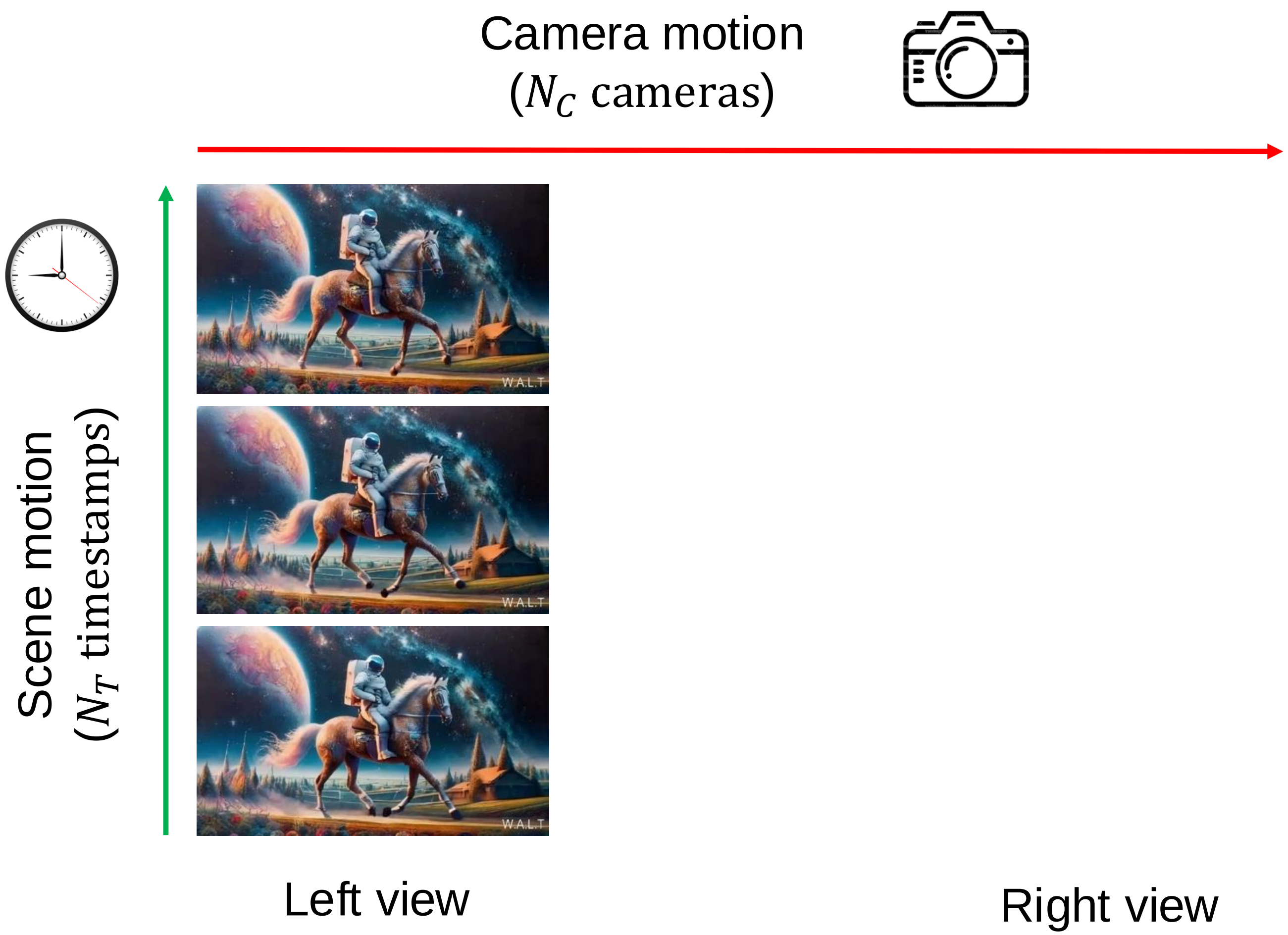
Method

Enhance semantic consistency => Denoise frame matrix

Frame matrix representation

Method

Enhance semantic consistency => Denoise frame matrix



Frame matrix representation

Method

Enhance semantic consistency => Denoise frame matrix

Camera motion
(N_C cameras)

Scene motion
(N_T timestamps)

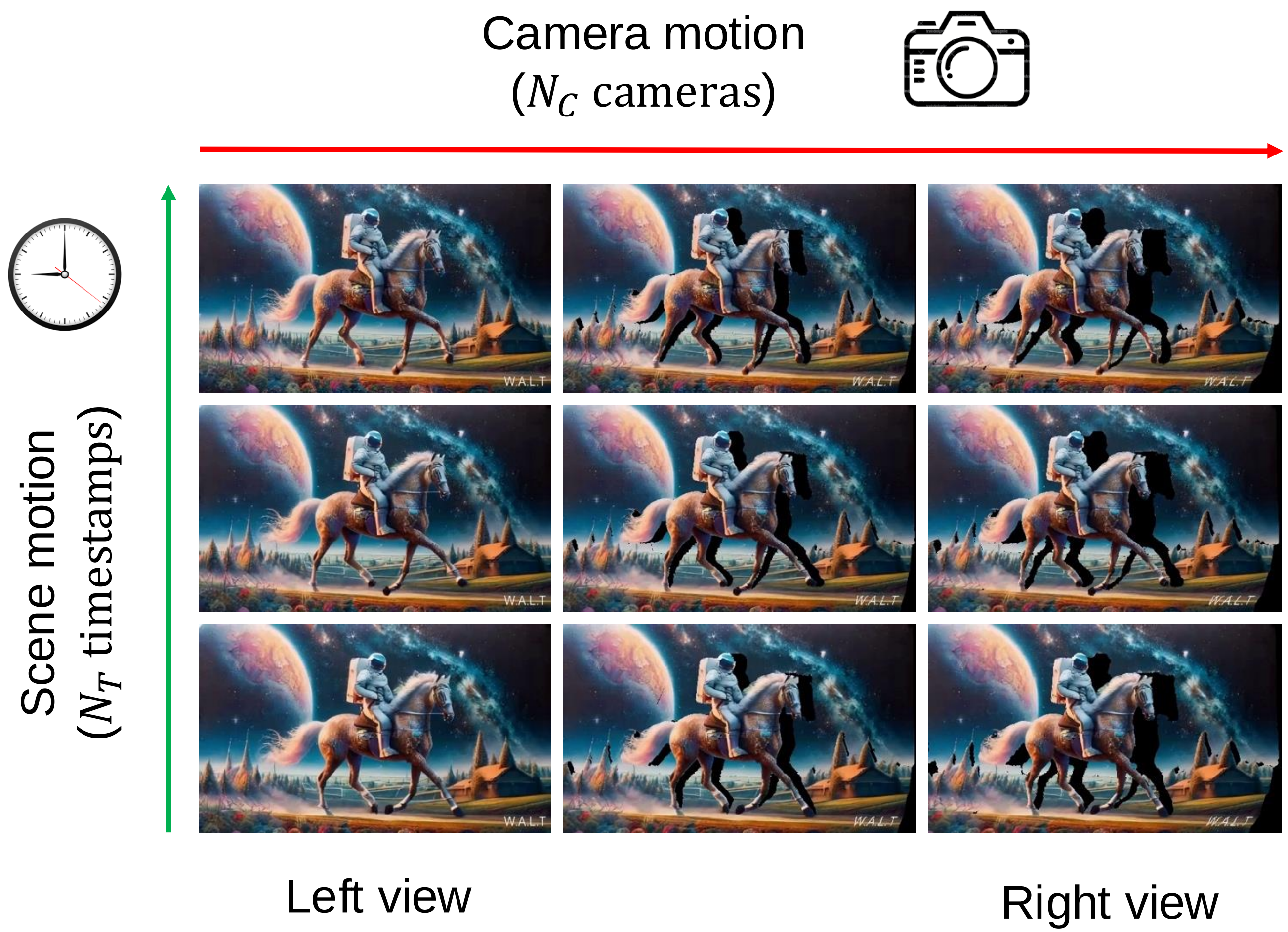
Left view

Right view

Frame matrix representation

Method

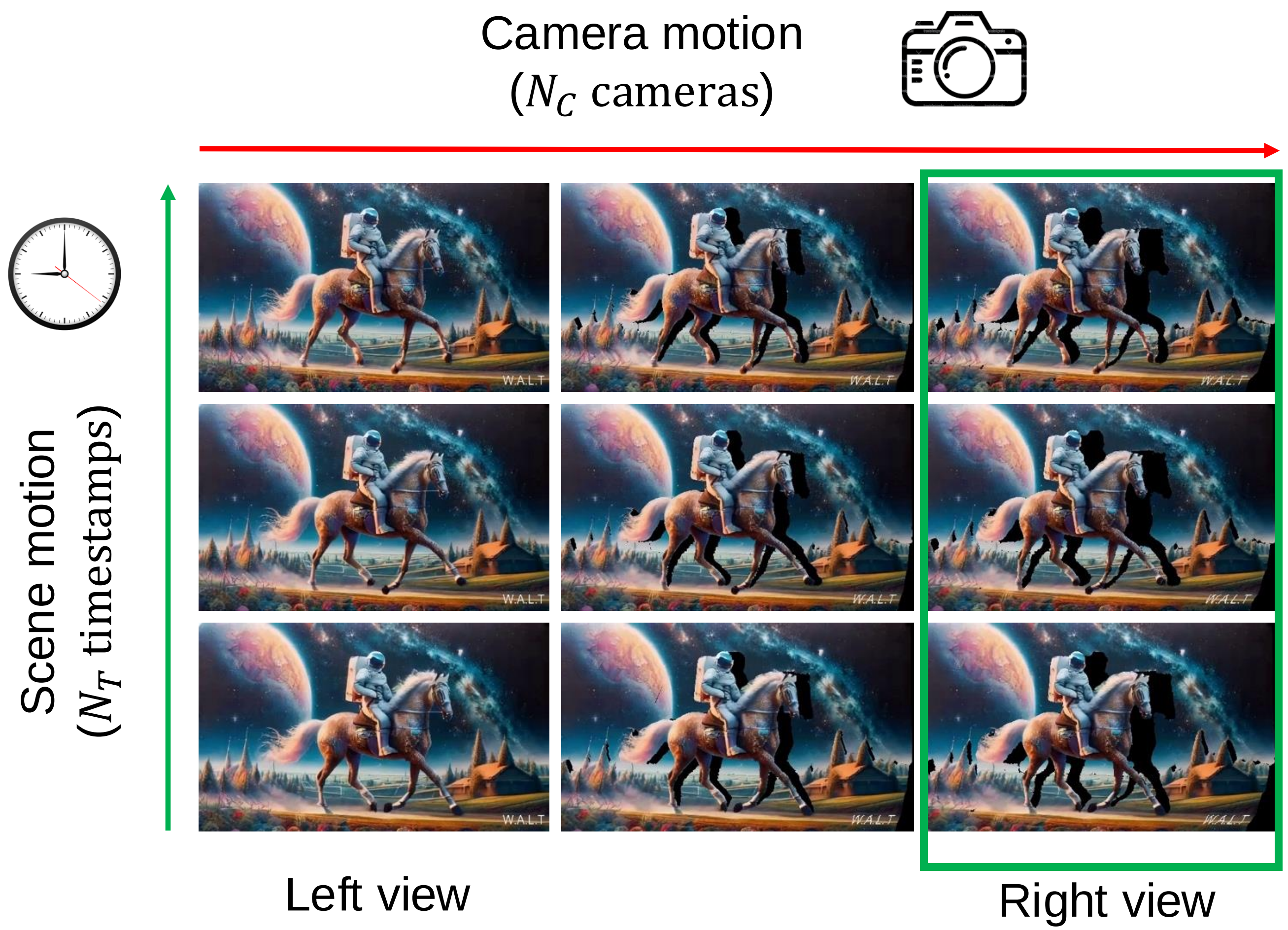
Enhance semantic consistency => Denoise frame matrix



Frame matrix representation

Method

Enhance semantic consistency => Denoise frame matrix

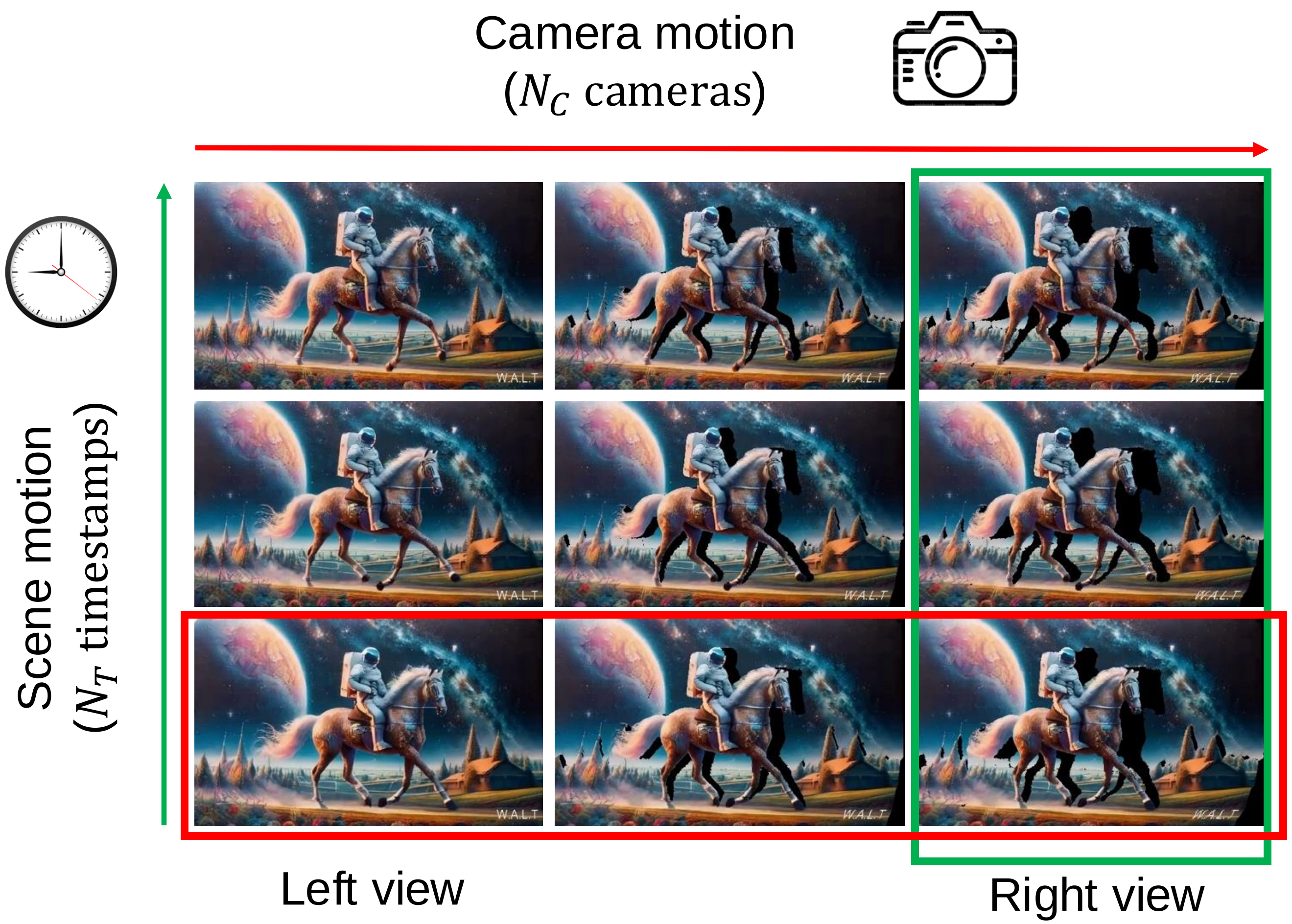


Frame matrix representation

Time direction

Method

Enhance semantic consistency => Denoise frame matrix



Frame matrix representation

Time direction

Spatial direction
(connect left and right views)

Method

One example of spatial direction denoising inpainting

Left view

Right view

Semantically consistent across different views/frames

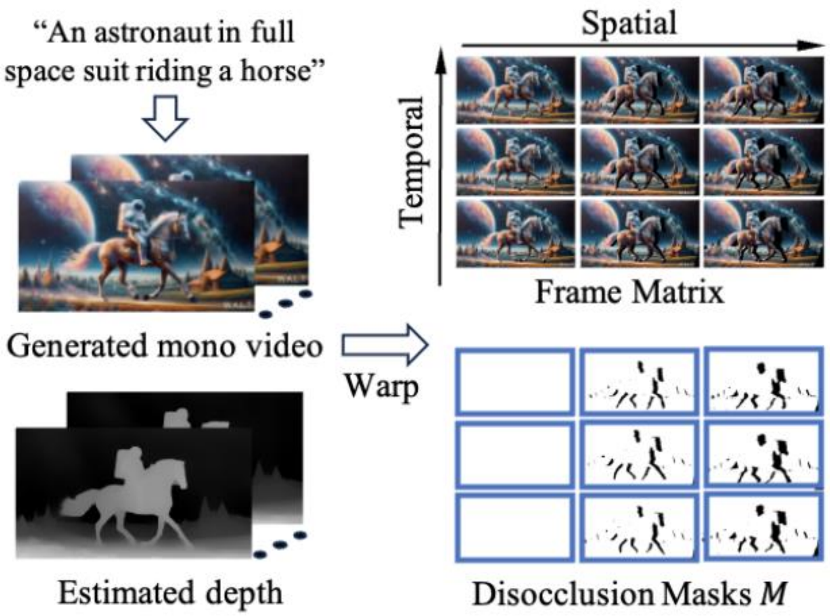
Method

Overview

Begin with monocular video and estimated video depth

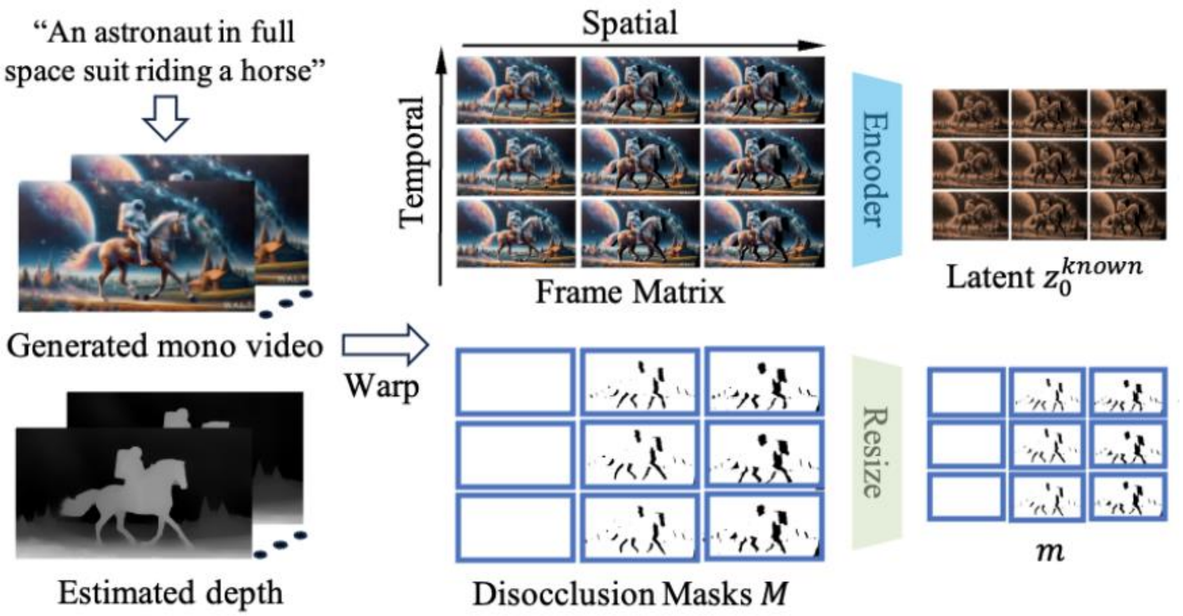
Method

Overview



Construct frame matrix

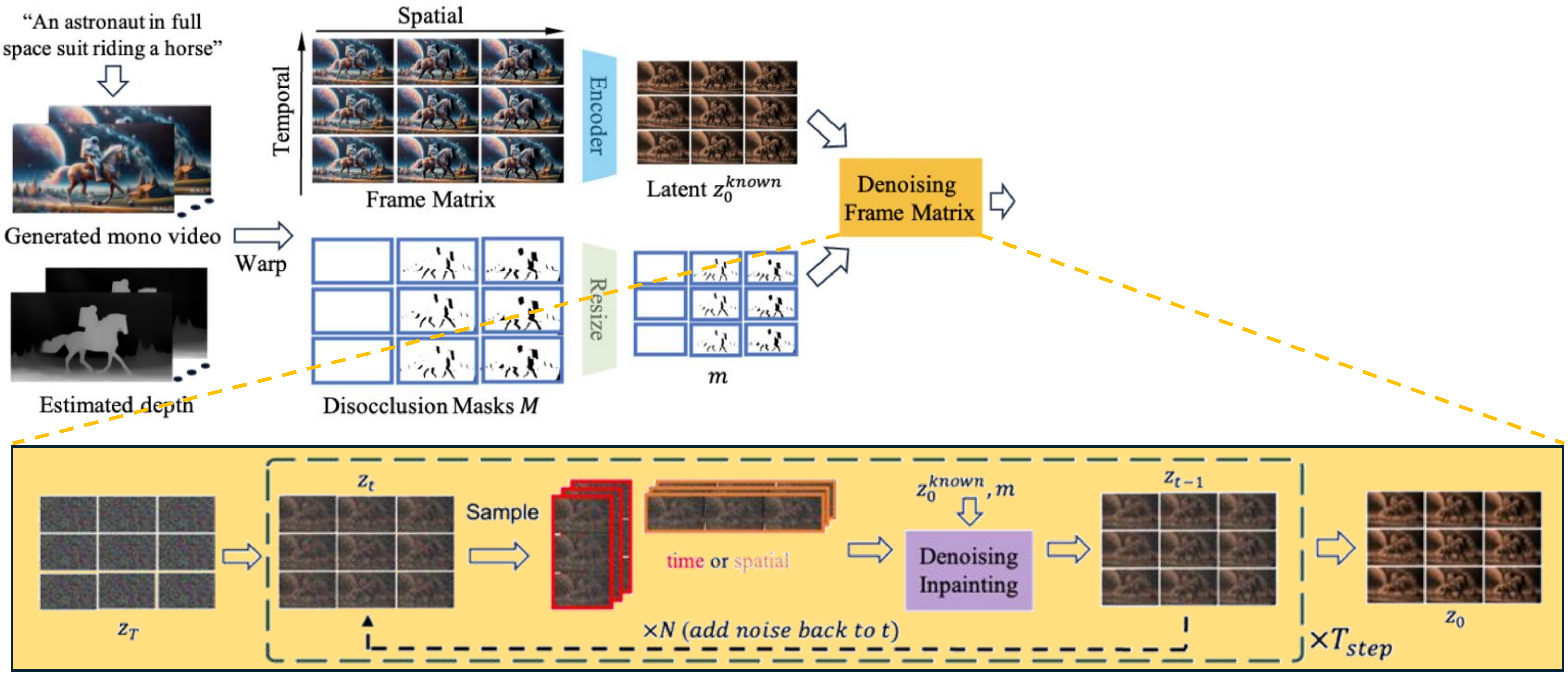
Overview



Encode frame matrix into latent space

Method

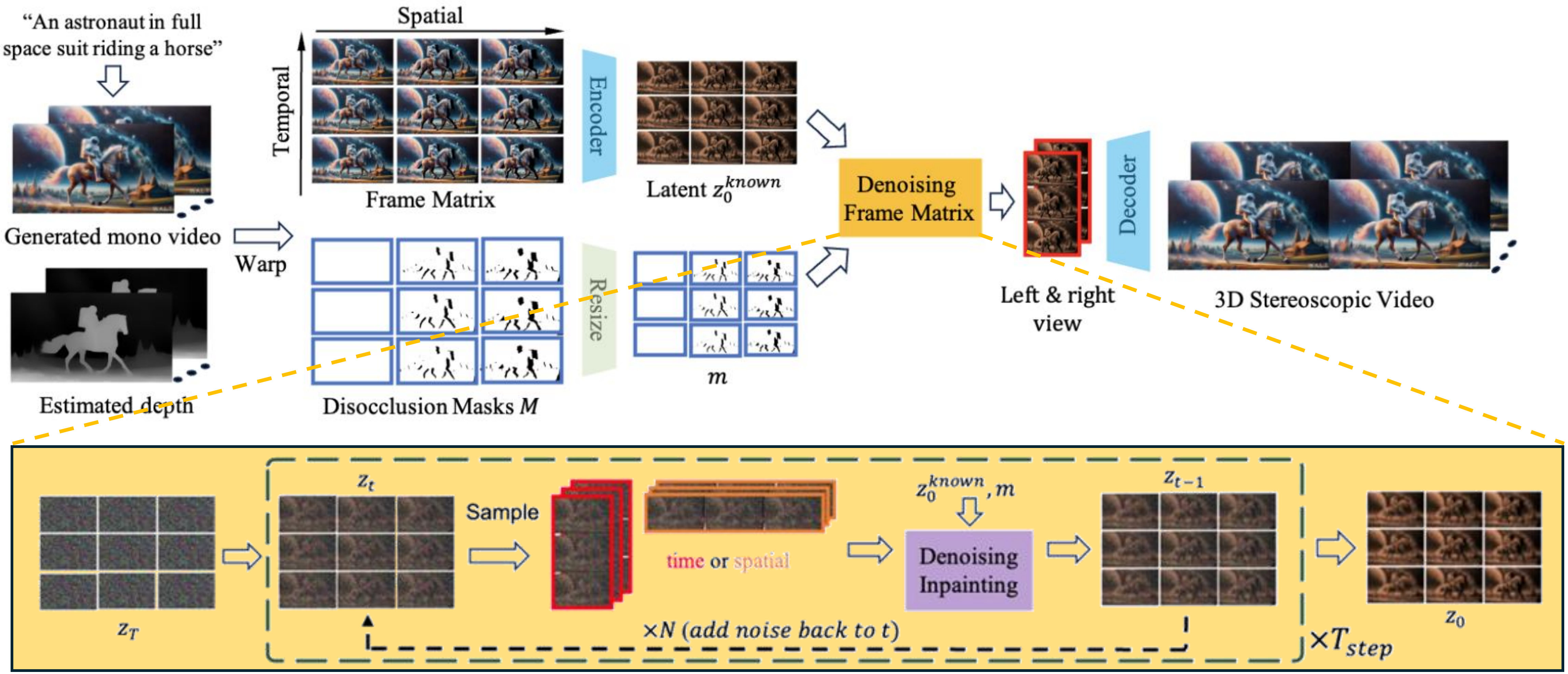
Overview



Iteratively denoising in spatial and temporal directions to fill unknown regions within frame matrix

Method

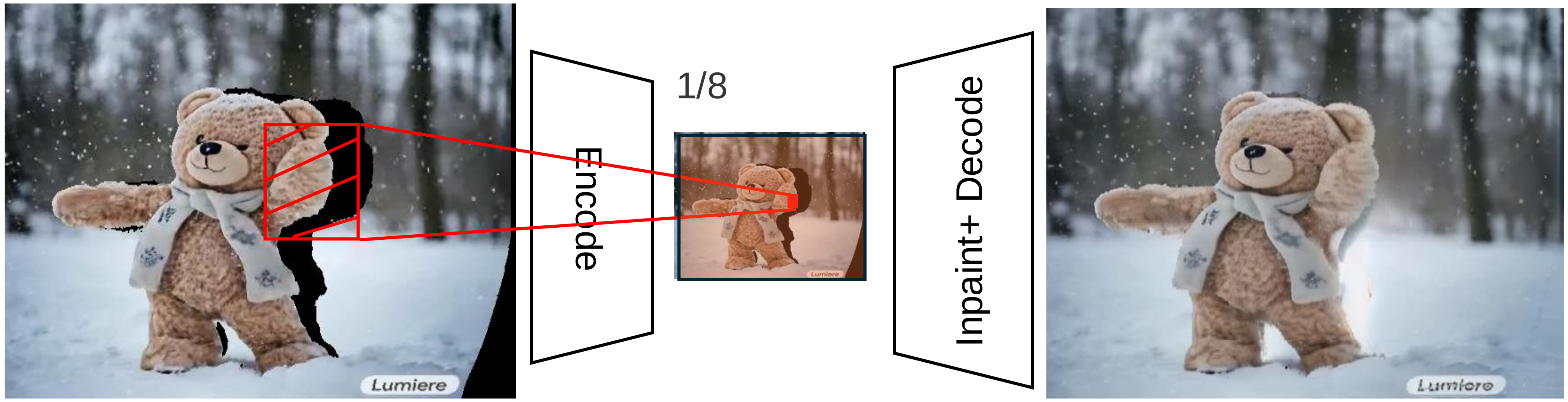
Overview



Choose the left-most and right-most views for stereoscopic video generation
Choose the entire frame matrix for multi-view video generation

Method

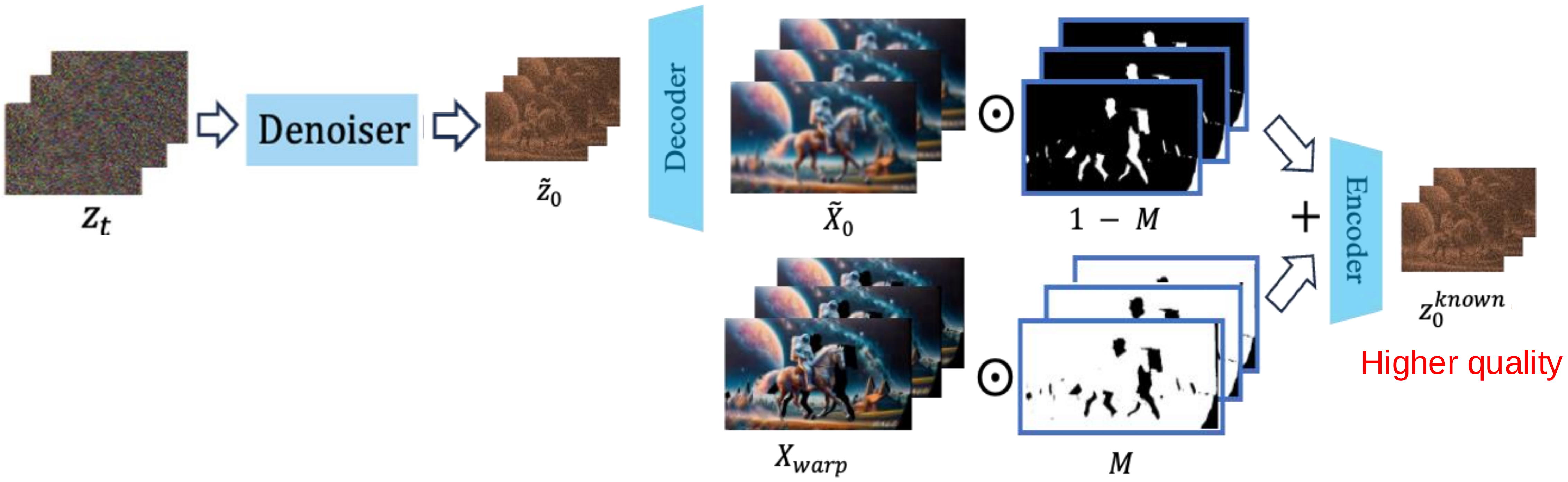
Another challenge when adopting latent diffusion



Disoccluded regions leak into known regions → Artifacts around disocclusion boundary

Method

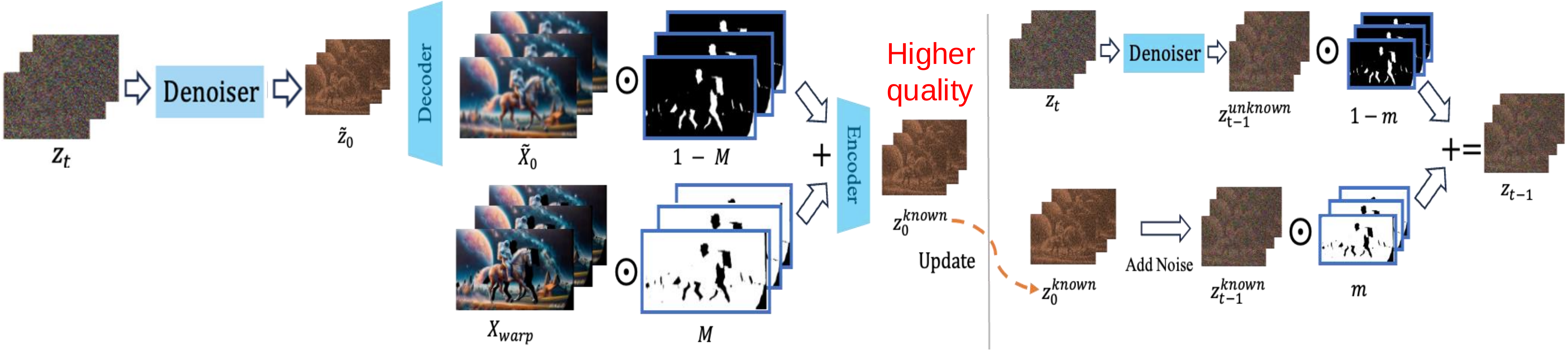
Update features around disocclusion boundary



Replace disoccluded regions with predicted content in image space, then encode again to obtain better latent features.

Method

Disocclusion boundary re-injection



Inject the updated latent features into denoising process

Results

Different scenarios

Generated video

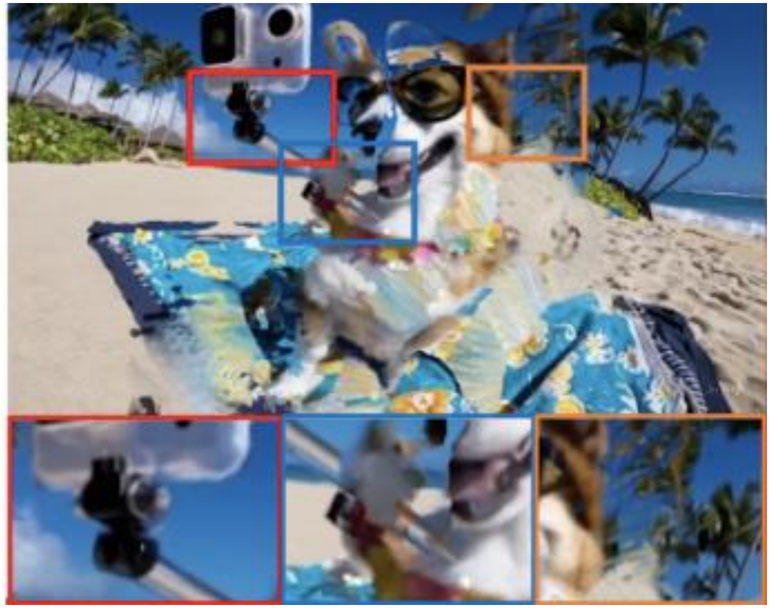
Fast moving objects

Real person

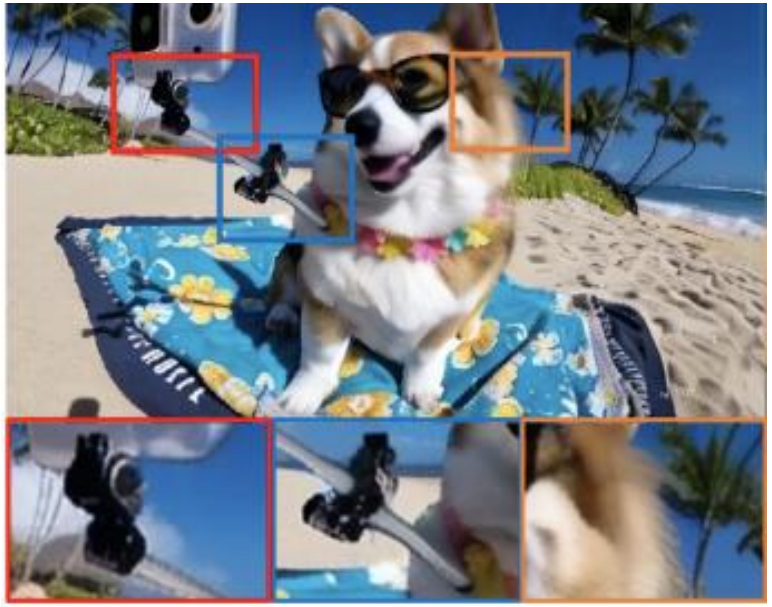
Results

Qualitative comparisons (Dynamic novel view synthesis)

RoDynRF (wide stereo baseline)



DynIBaR

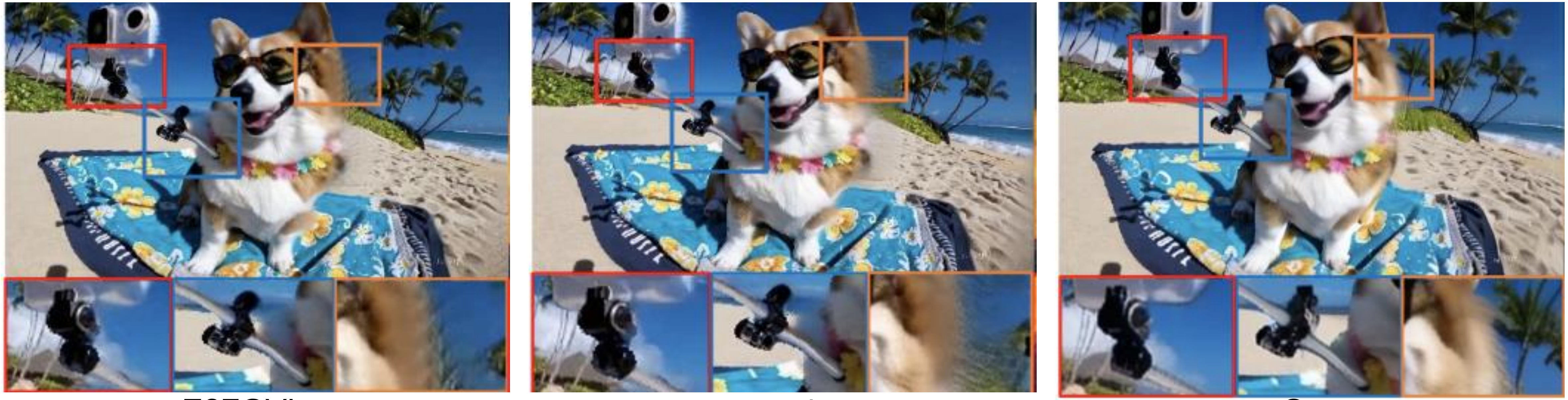


Ours

Dynamic NVS cannot hallucinate occluded regions, and requires accurate camera pose estimation

Results

Qualitative comparisons (Video inpainting)



E2FGVI ProPainter Ours

Existing video inpainting methods produce blurry results

Results

Qualitative video comparisons


Left view


E2FGVI


ProPainter


RoDynRF


DynIBaR


Ours

Baseline methods: blurry, contain unknown regions, require accurate camera poses

Results

Quantitative comparisons

	E2FGVI	ProPainter	RoDynRF	DynIBaR	Ours
Stereo Effect ↑	4.79 (1.08)	4.81 (1.13)	2.97 (1.34)	1.86 (1.25)	**5.24** (0.94)
Temporal Consistency ↑	4.74 (1.33)	4.74 (1.22)	3.35 (1.66)	1.89 (1.33)	**5.15** (1.22)
Image Quality ↑	4.42 (1.27)	4.38 (1.28)	2.84 (1.60)	1.67 (1.07)	**5.12** (1.33)
Overall Experience ↑	4.67 (1.04)	4.66 (1.09)	2.92 (1.43)	1.72 (1.06)	**5.35** (0.99)

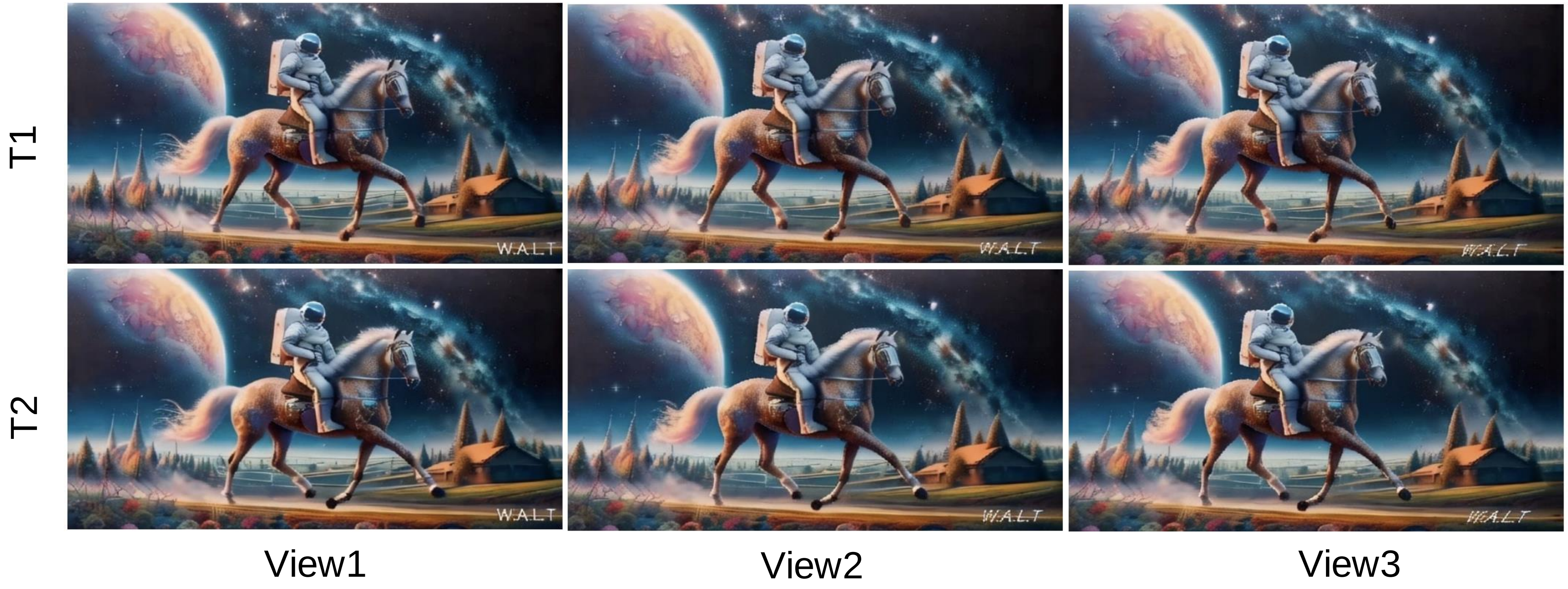
Human perception studies

Method	E2FGVI	ProPainter	RoDynRF	DynIBaR	Ours - FM	Ours - DBR	Ours
CLIP ↑	94.34	95.29	96.03	93.24	95.81	95.60	**96.44**
Aesthetic ↑	5.06	5.07	4.97	4.66	5.25	5.18	**5.27**
DOVER ↑	0.547	0.535	0.352	0.365	0.565	0.560	**0.584**
FVD ↓	638	606	727	1208	614	699	**599**

Video quality measurement

Results

Multi-view video generation



Choose >2 views from frame matrix to obtain multi-view videos

Results

Fix time, change view
(row of frame matrix)

Time 1

Time 2

Fix view, change time
(column of frame matrix)

View 1

View 2

Results

Fix time, change view
(row of frame matrix)

Time 1

Time 2

Fix view, change time
(column of frame matrix)

View 1

View 2

Results

Efficacy of frame matrix

Left view

Without frame matrix

Ours

Improved semantic consistency

Results

Efficacy of disocclusion boundary re-injection

Without disocclusion boundary re-injection

Ours

Improved video quality

Results

Utilize unobserved content

Character “R” is correctly inpainted

Thanks for your attention !