[arXiv 2025]

STABLE VIRTUAL CAMERA: Generative View Synthesis with Diffusion Models

Jensen (Jinghao) Zhou[1,2,*,†] Hang Gao[1,3,*,†]
Vikram Voleti[1] Aaryaman Vasishta[1] Chun-Han Yao[1] Mark Boss[1]
Philip Torr[2] Christian Rupprecht[2] Varun Jampani[1]

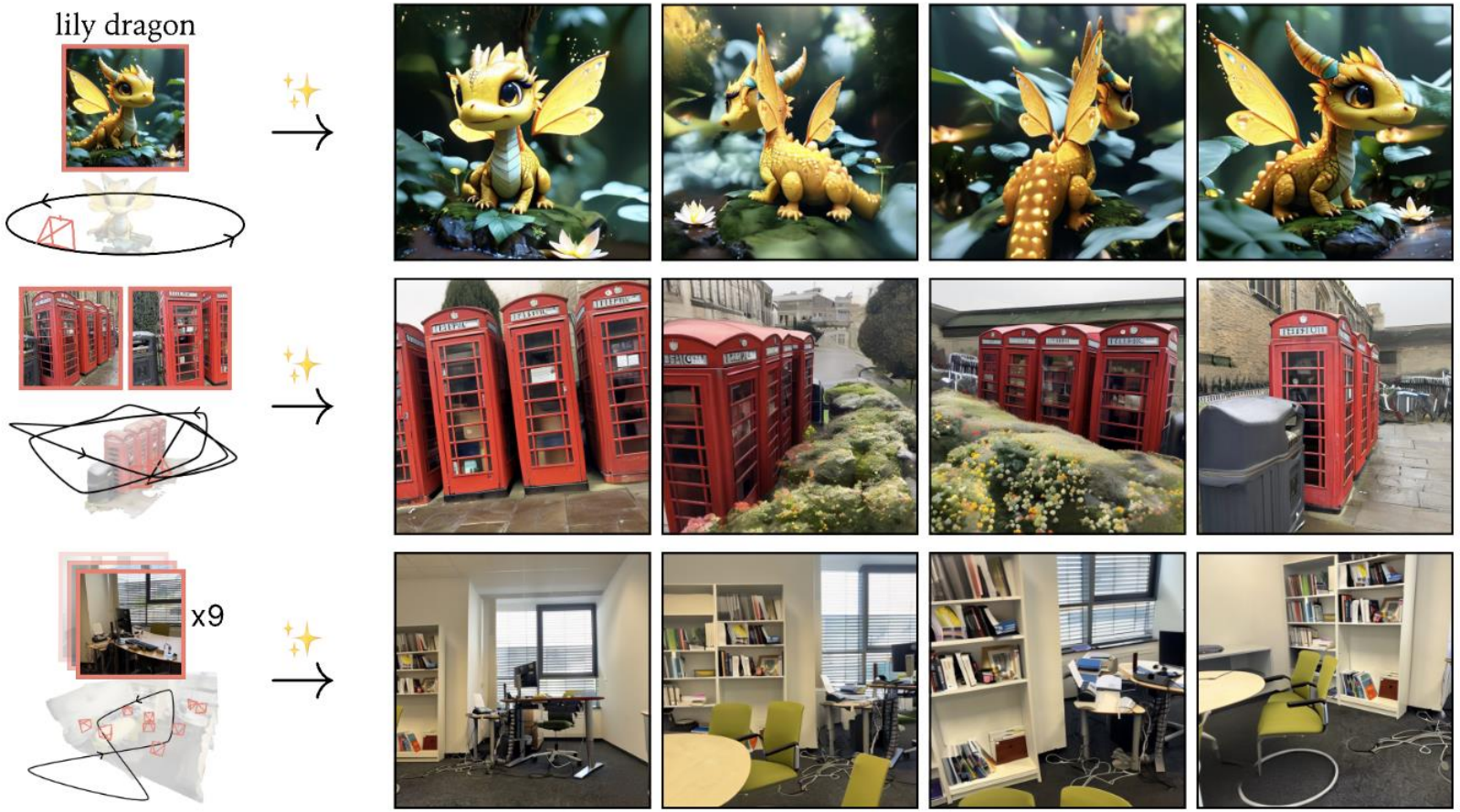
[1]Stability AI [2]University of Oxford [3]University of California, Berkeley

Presenter: Gyeongsu Cho
Fri Feb 21, 2025

Contents

Introduction

SEVA



SEVA generates novel views from any number of input views and target cameras, which the user can specify anywhere

Motivation

Virtual camera system

Article Talk

From Wikipedia, the free encyclopedia

In 3D video games, a **virtual camera system** aims at controlling a camera or a set of cameras to display a view of a 3D virtual world. Camera systems are used in video games where their purpose is to show the action at the best possible angle; more generally, they are used in 3D virtual worlds when a third-person view is required.

As opposed to filmmakers, virtual camera system creators have to deal with a world that is interactive and unpredictable. It is not possible to know where the player character is going to be in the next few seconds; therefore, it is not possible to plan the shots as a filmmaker would do. To solve this issue, the system relies on certain rules or artificial intelligence to select the most appropriate shots.

SEVA aims to generate realistic images from new camera viewpoints, given a few inputs.

Challenges in NVS

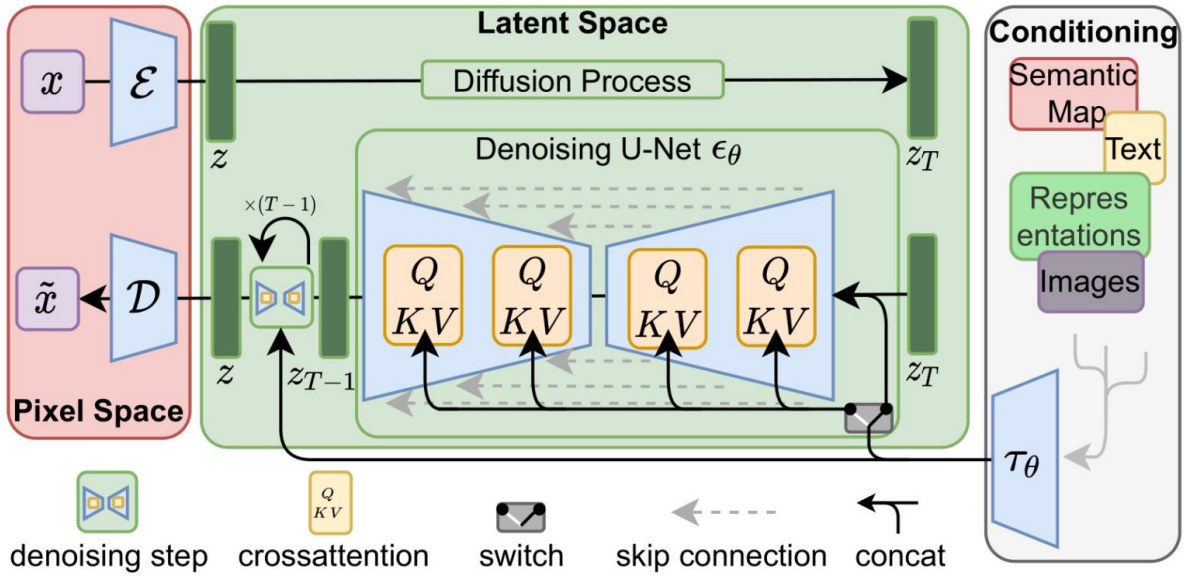
- Sparse input views
- Large viewpoint gaps
- Temporal inconsistency

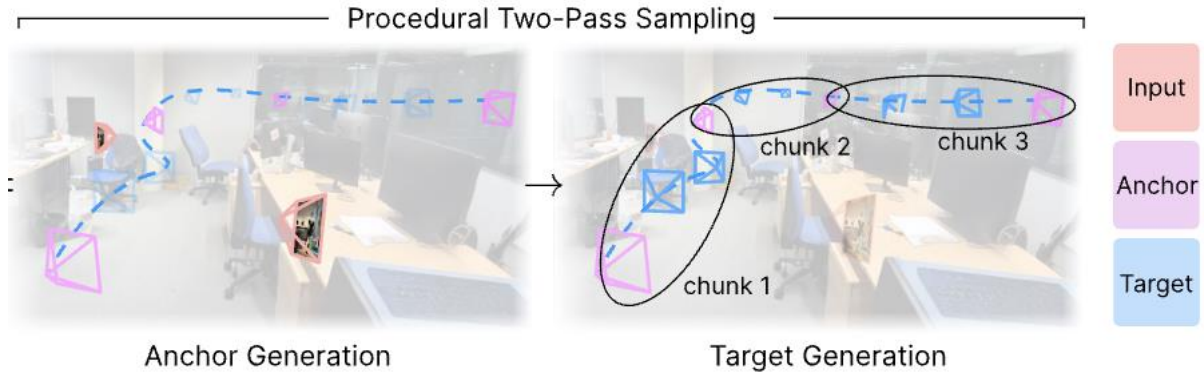
- Limitations of NeRFs with diffusion models

[NeurIPS 2024] CAT3D

Existing methods have limited inputs and outputs or need 3D representation (NeRF, 3D pcd, 3DGS)

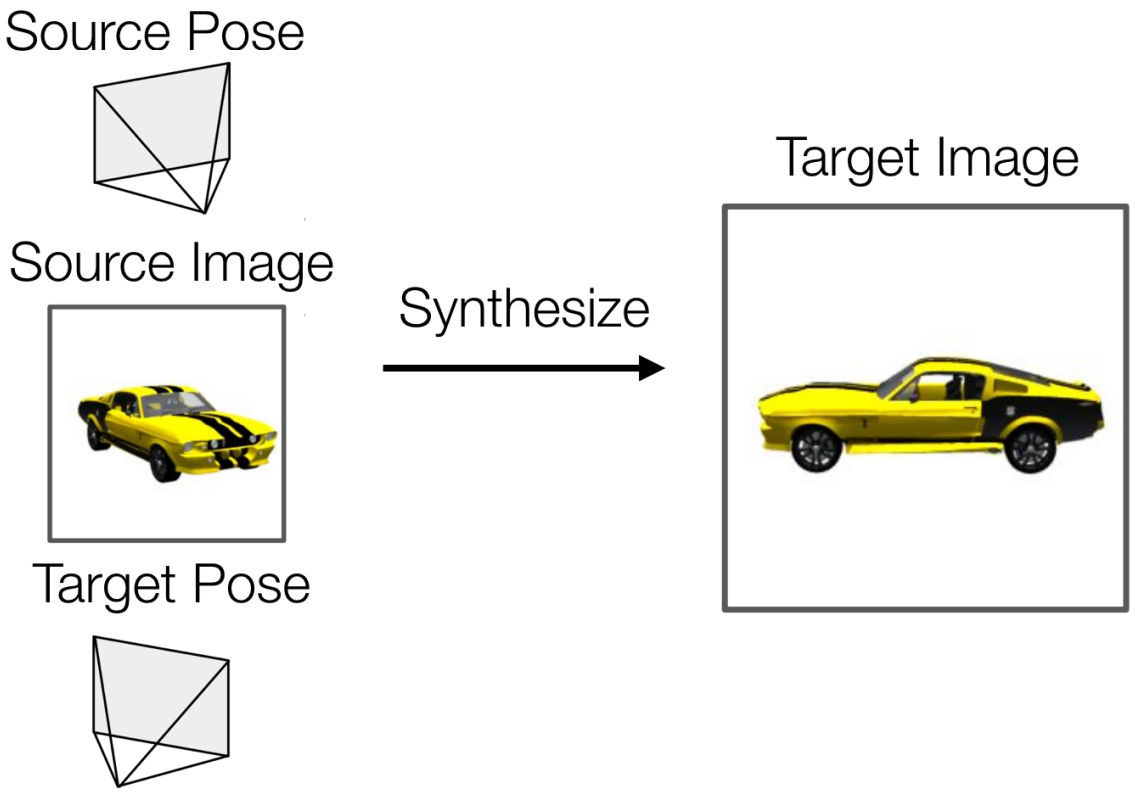
Key Idea: Diffusion-Based Virtual Camera Without 3D Supervision





- Treat novel view synthesis (NVS) as a conditional image generation problem.
- Enable *flexible camera trajectories* and input-output configurations via **procedural sampling**.

Preliminaries: Novel View Synthesis

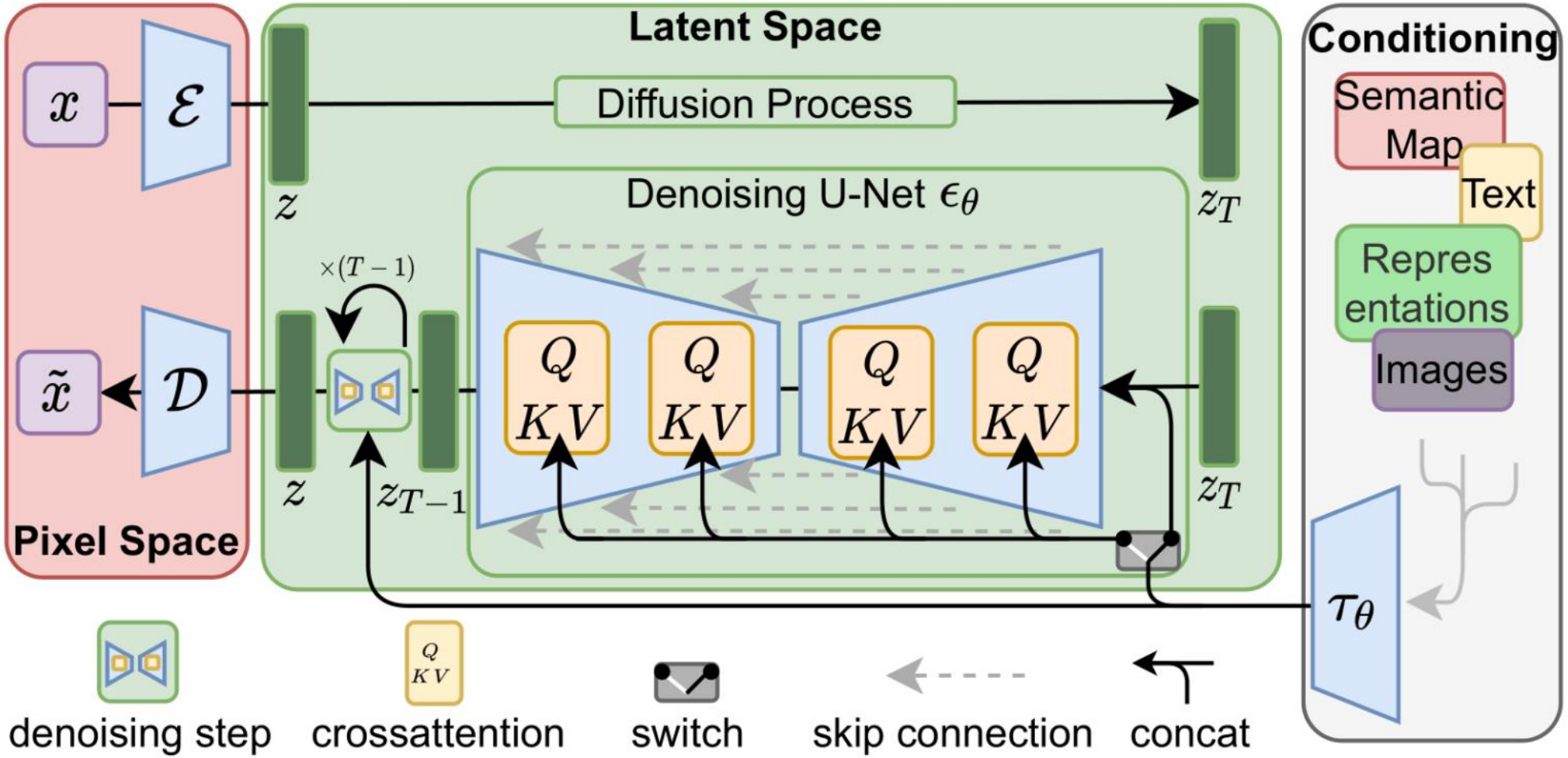


Set NVS

Trajectory NVS

The goal in NVS is to synthesize images from new, unseen viewpoints given a set of input images and their corresponding camera parameters.

Preliminaries: Latent diffusion model

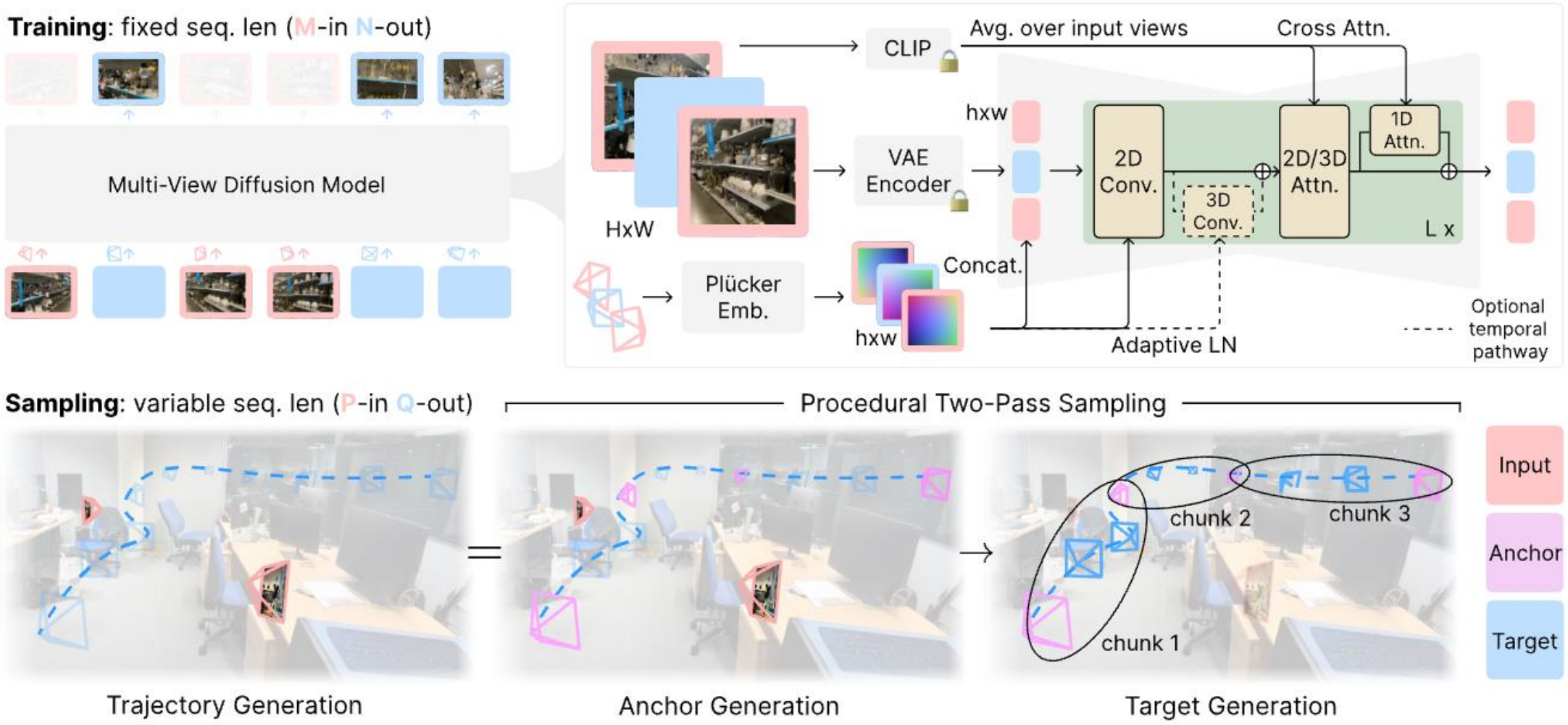


Latent Diffusion Models (LDMs):

- Operate in compressed *latent space* instead of pixel space.
- Use a pre-trained **VAE** to encode/decode images.

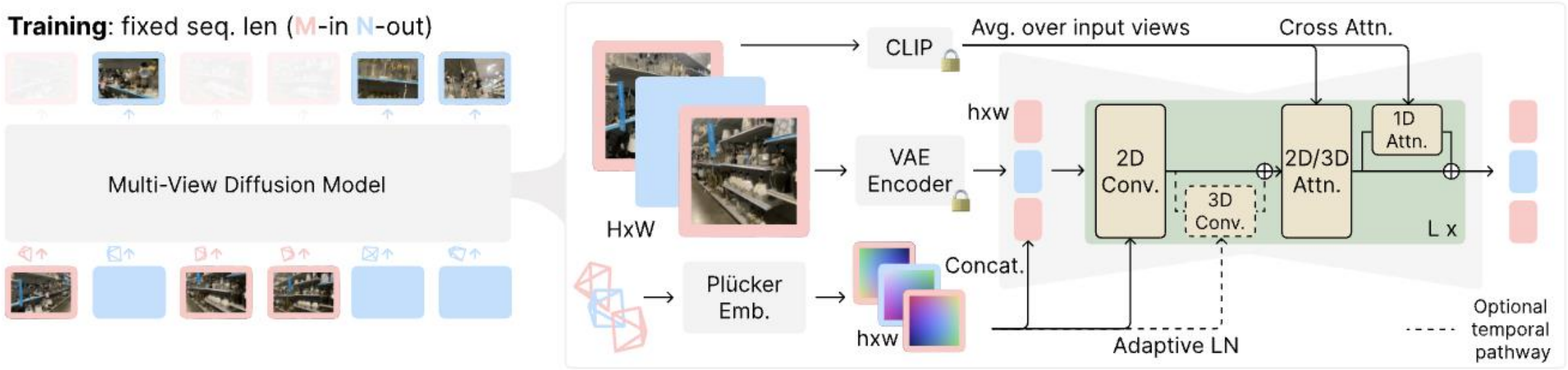
Method

Overview



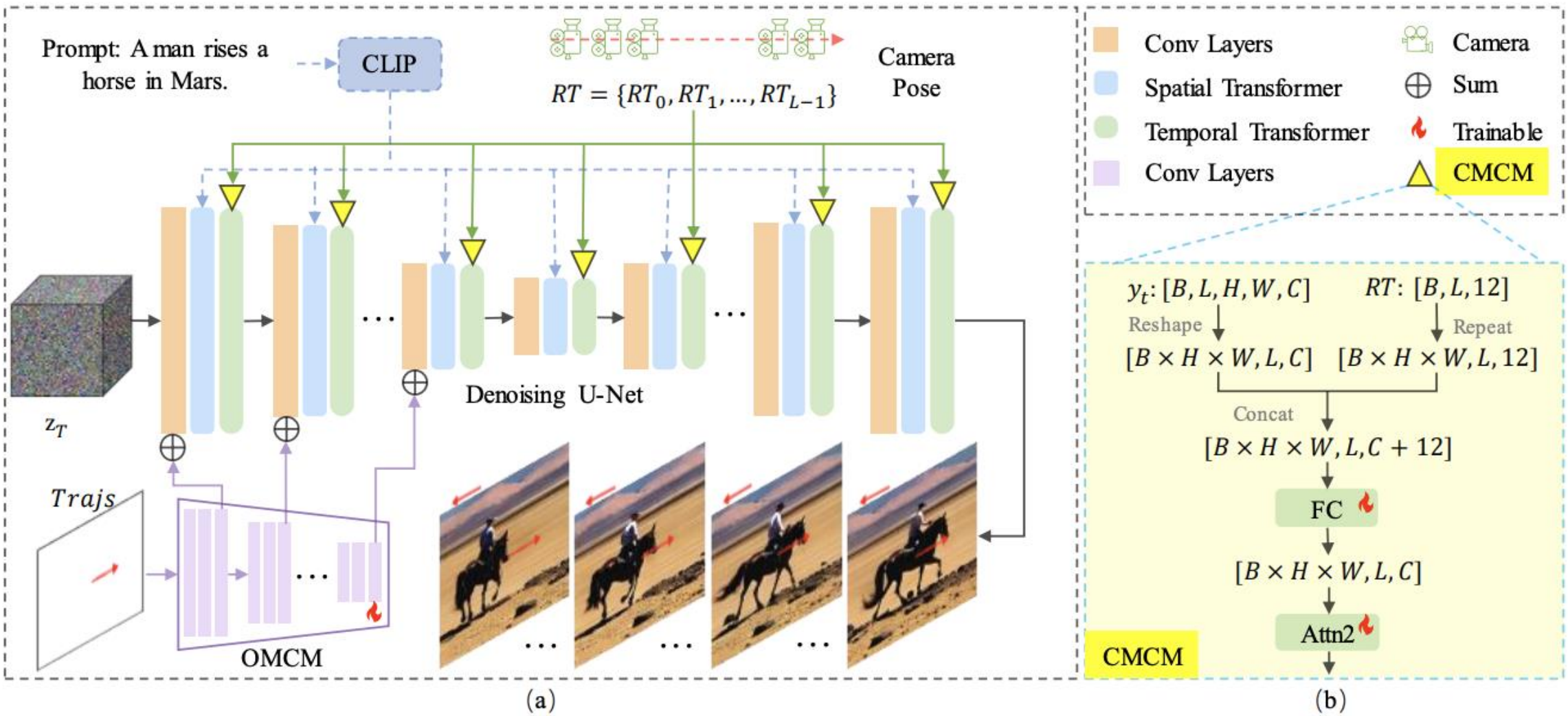
SEVA builds on a latent diffusion model (SD 2.1) with modifications

Overview



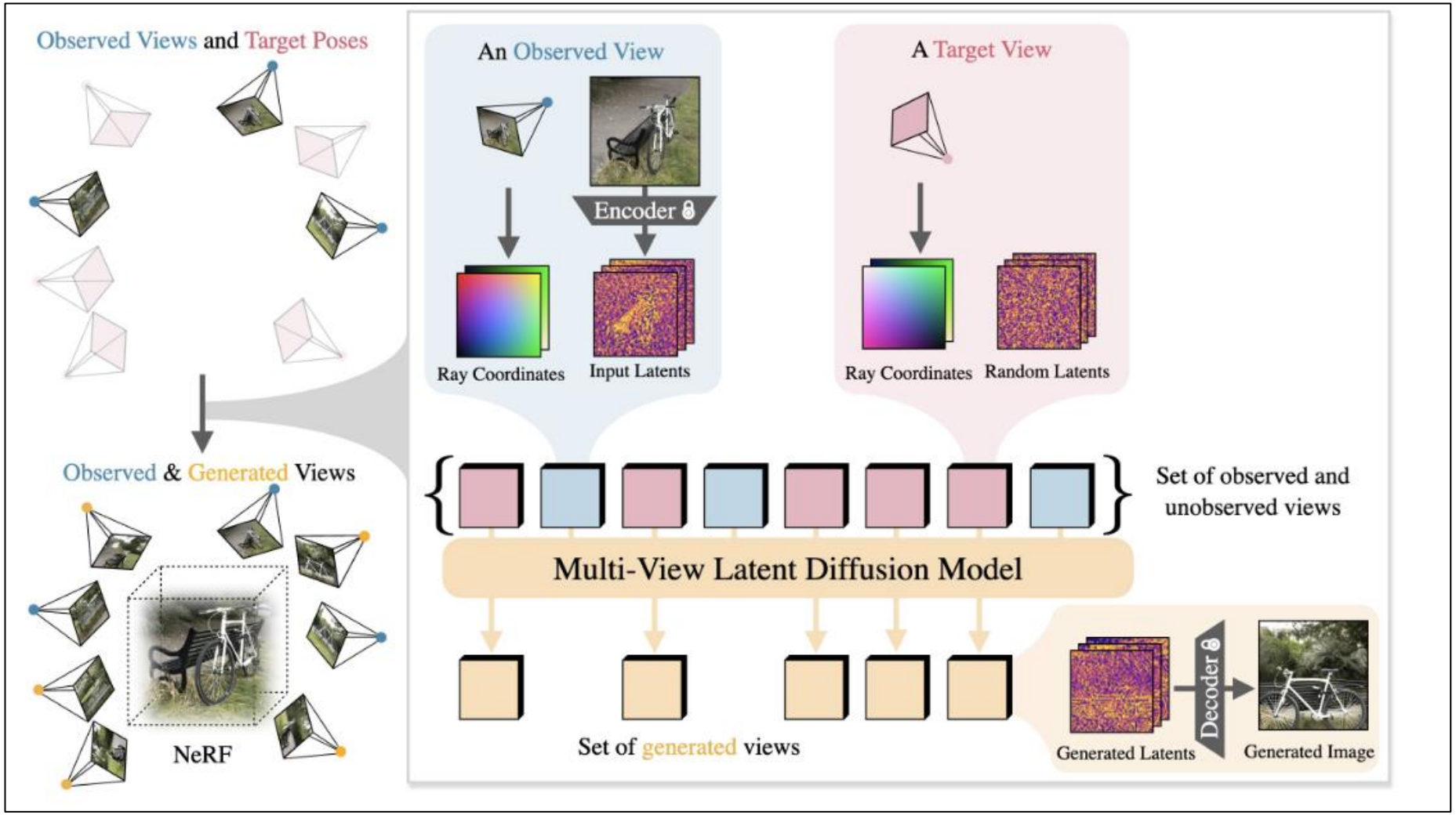
SEVA builds on a latent diffusion model (SD 2.1) with modifications

Overview



MotionCtrl Framework

Overview

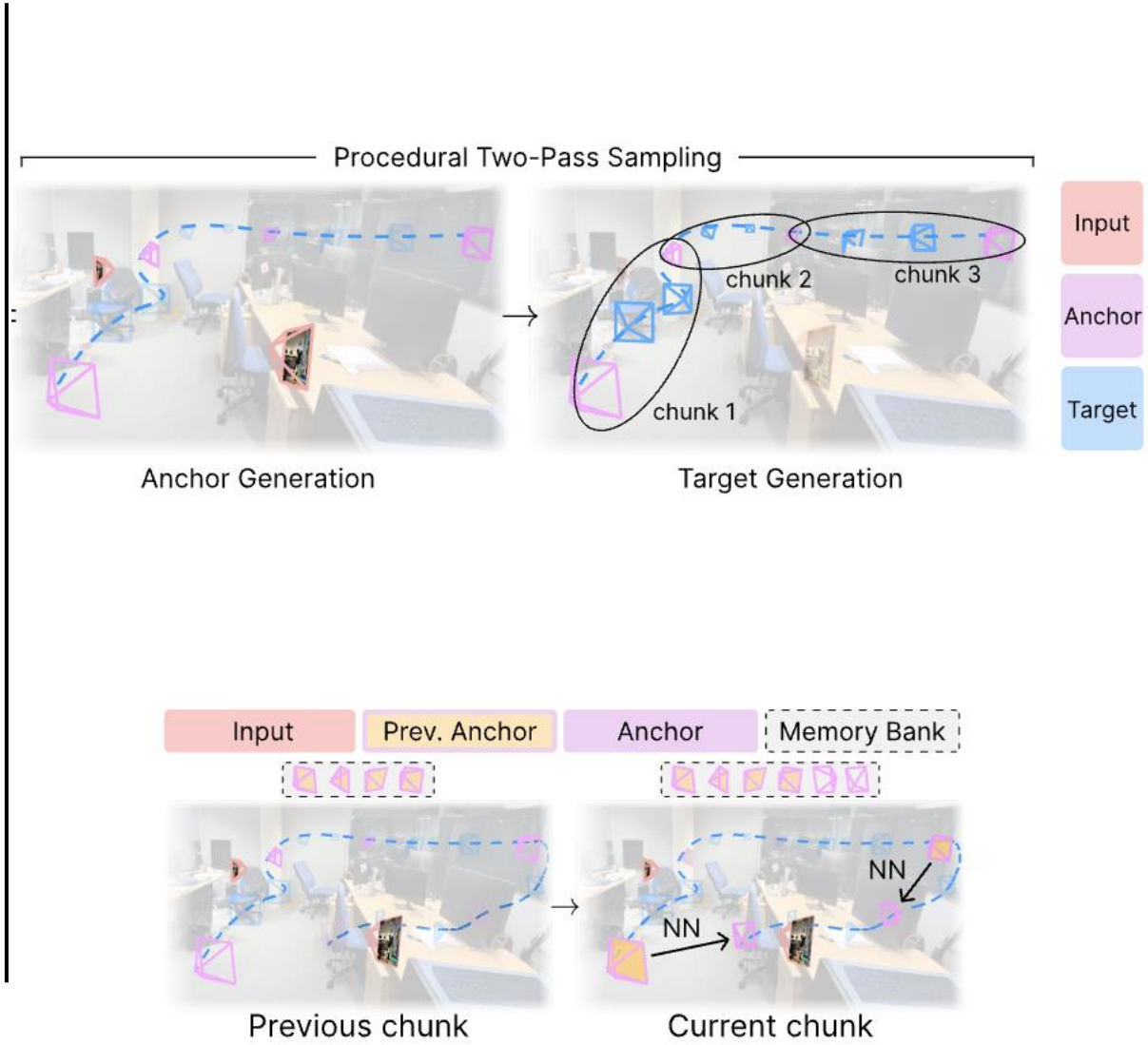


CAT3D uses a **multi-view latent diffusion model** to generate novel views of the scene.

Overview

SEVA builds on a latent diffusion model (SD 2.1) with modifications

Training Strategies



- Two-stage training (T=8 and T=21)
- Input-output sampling strategy

Experiments

Experiments

- Dataset
  - 10 datasets
- Baseline
  - Set NVS
  - Trajectory NVS
- Evaluation metrics
  - PSNR, LPIPS, TSED

Experiments

Method	dataset	OO3D	GSO	RE10K					LLFF		DTU		CO3D		WRGBD		Mip360	DL3DV		T&T	
	split	O	O	D [12]	V [9]	P [16]	R [7]		R [7]		R [7]		V [9]	R [7]	O_e	O_h	R [7]	O	L [18]	V [9]	L [18]
	P	3	3	1	1	2	1	3	1	3	1	3	1	3	3	6	6	6	32	1	32
Regression-based models																					
Long-LRM [18]		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	23.86	-	18.20
MVSplat [17]		14.78	15.21	20.42	20.32	26.39	21.56	25.64	11.23	12.50	13.87	15.52	12.52	13.52	14.56	12.54	13.56	14.34	16.24	13.22	12.63
DepthSplat [45]		15.67	16.52	20.90	19.24	27.44	21.87	22.54	12.07	12.62	14.15	16.24	13.23	13.77	15.93	14.23	14.01	15.72	16.78	14.35	13.12
LVSM [11]		-	-	-	-	29.67	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Diffusion-based models																					
MotionCtrl [46]		-	-	12.74	16.29	-	-	-	-	-	-	-	15.46	-	-	-	-	-	-	13.29	-
4DiM [12]		-	-	17.08	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
ViewCrafter [9]		14.64	15.93	20.43	22.04	21.42	20.88	22.81	10.53	13.52	12.66	16.40	18.96	14.72	16.42	12.66	14.59	13.78	-	18.07	-
SEVA		30.30	31.53	17.99	18.56	25.66	18.11	27.57	14.03	19.48	14.47	20.82	18.40	19.25	19.75	18.91	16.70	17.80	20.96	15.16	16.50

Qualitative comparisons of SEVA

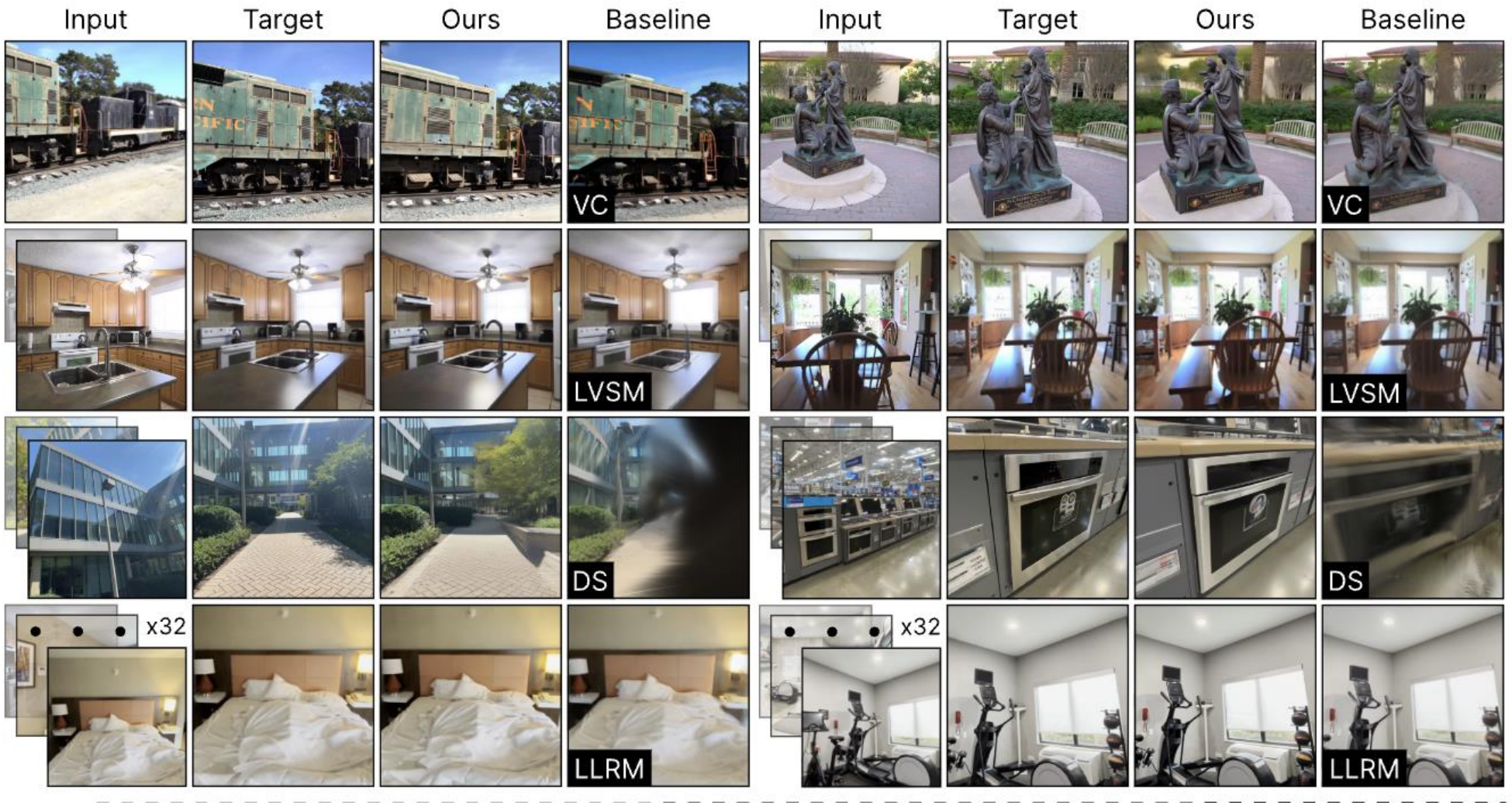
Experiments

Method	dataset	OO3D	GSO	CO3D	WRGBD		Mip360		DL3DV		T&T			
	split	S [19]	S [19]	R [7]	O_h		R [7]		O		O			
	P	1	1	1	1	3	1	3	1	3	1	3	6	9
SV3D [19]		19.28	20.38	-	-	-	-	-	-	-	-	-	-	-
DepthSplat [45]		11.56	12.32	10.42	9.35	13.53	10.49	12.54	9.63	12.52	8.63	9.78	10.12	11.20
CAT3D [8]		-	-	-	-	-	-	15.15	-	-	-	-	-	-
ViewCrafter [9]		10.56	11.42	10.11	9.12	13.45	9.79	10.34	8.97	11.50	9.23	9.88	10.32	11.08
SEVA		19.25	20.65	15.30	14.37	17.28	12.93	15.78	13.01	15.95	11.28	12.65	13.80	14.72

Method	small-viewpoint				large-viewpoint
	RE10K	LLFF	DTU	CO3D	Mip360
ZipNeRF [47]	20.77	17.23	9.18	14.34	12.77
ZeroNVS [13]	19.11	15.91	16.71	17.13	14.44
ReconFusion [7]	25.84	21.34	20.74	19.59	15.50
CAT3D [8]	26.78	21.58	22.02	20.57	16.62
SEVA	27.95	21.88	22.68	21.88	17.82

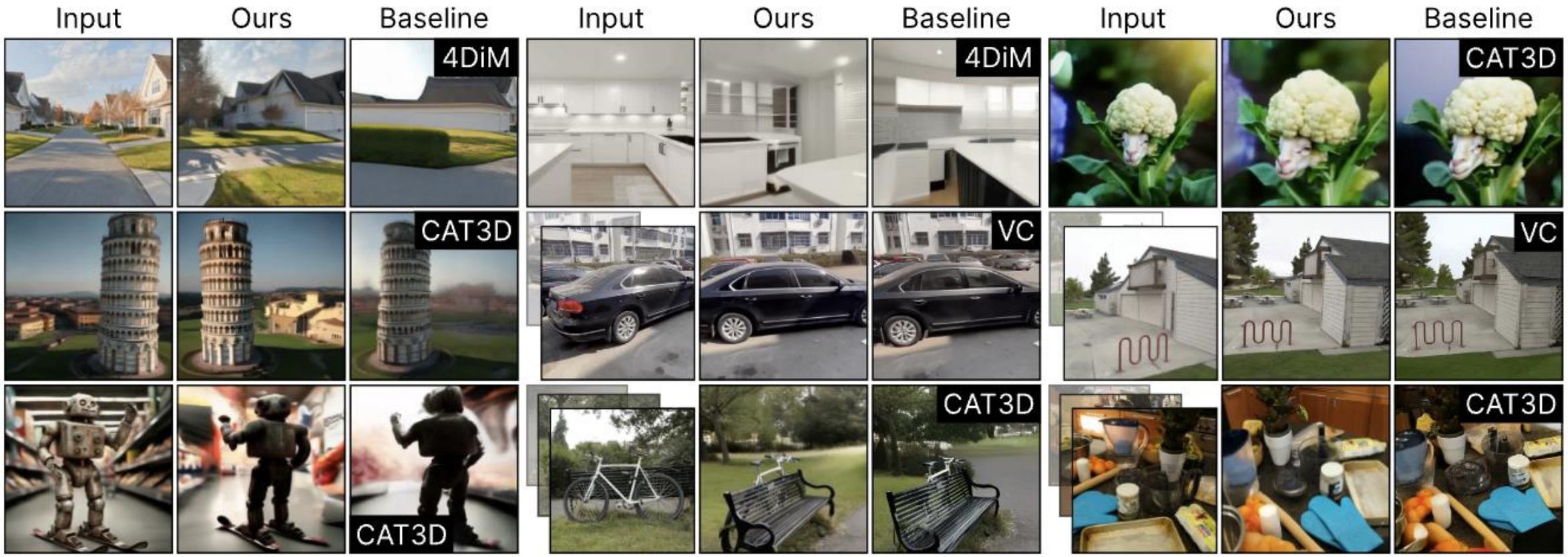
Qualitative comparisons of SEVA

Experiments



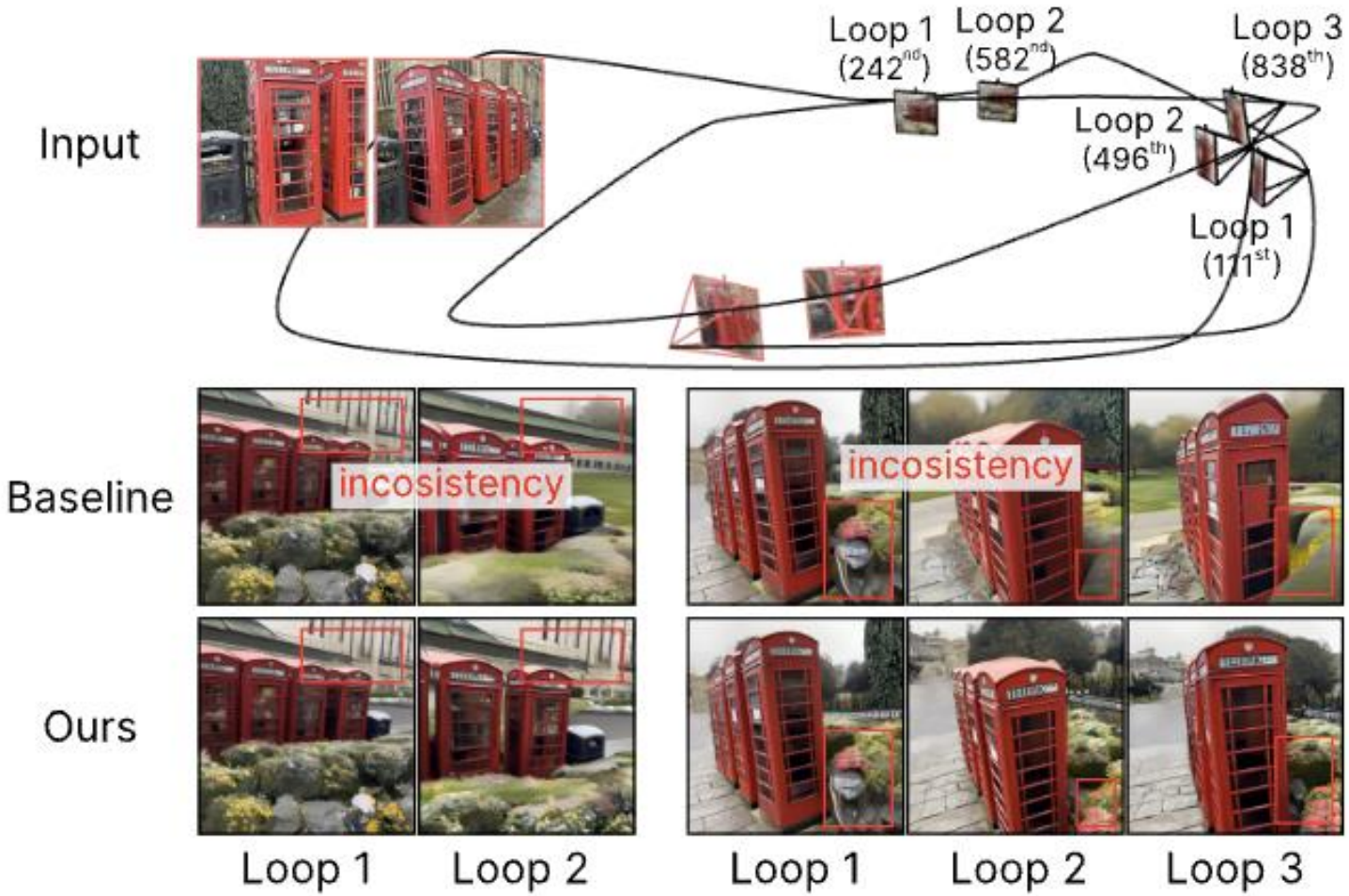
SOTA comparison

Experiments



SOTA comparison

Experiments



Long-range 3D consistency

Conclusion

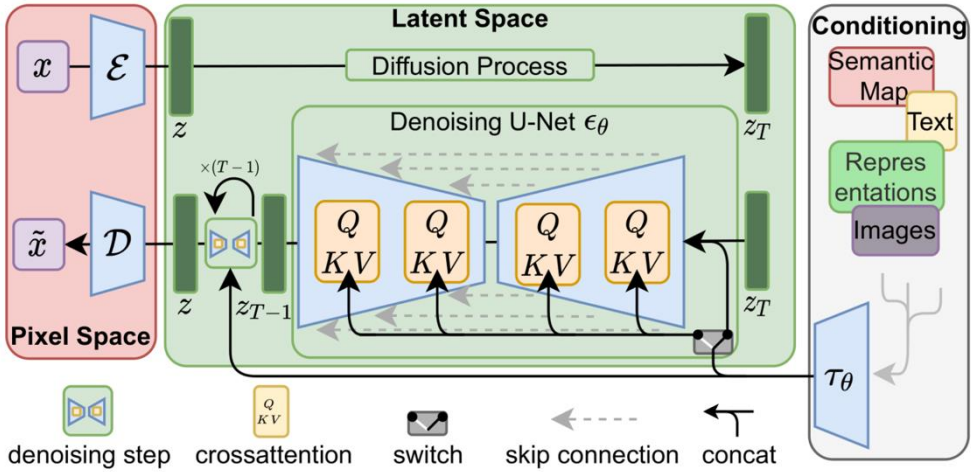
Conclusion

Input: Observed View

Output: Novel Views

Given any number of input views and their cameras, SEVA generates novel views of a scene at any target camera of interest

Conclusion



[CVPR 2022] Latent Diffusion

[ICCV 2023] ControlNet

[ICCV 2023] Zero123

[CVPR 2023] Align your latents

SORA ..

I believe SEVA has the potential to be impactful in a way that resembles how Stable Diffusion revolutionized image generation.

Appendix

Appendix

model	training data	generation capacity	interpolation smoothness	input flexibility
Regression-based				
pixelNeRF [4]	📦	✗	✓	sparse (1)
pixelSplat [16]	▲	✗	✓	sparse (2)
MVSplat [17]	▲	✗	✓	sparse (2)
Long-LRM [18]	▲	✗	✓	semi-dense ({16, 32})
LVSM [11]	📦 / ▲	✗	✓	sparse ({2, 4})
Diffusion-based: image models				
Zero123 [6]	📦	✓	✗	sparse (1)
ZeroNVS [13]	📦 ▲	✓	✗	sparse (1)
ReconFusion [7]	📦 ▲	✓	✗	sparse (3)
CAT3D [8]	📦 ▲	✓	✗	sparse ([1, 9])
Diffusion-based: video models				
SV3D [19]	📦	✗	✓	sparse (1)
MotionCtrl [10]	▲	✗	✓	sparse (1)
ViewCrafter [9]	▲	✗	✓	sparse (2)
4DiM [12]	▲	✓	✓	sparse ({1, 2, 8})
SEVA	📦 ▲	✓	✓	sparse ([1, 8]), semi-dense ([9, 32*])

	type	split	#scene	$(\mathbf{I}^{inp}, \mathbf{I}^{tgt}) \sim \mathcal{V}$	P	$\mathcal{D}_{\text{CLIP}}(\mathbf{I})$
Small-viewpoint NVS						
OmniObject3D [35]	📦	O (dynamic orbit)	308	✓	3	0.11
GSO [36]	📦	O (dynamic orbit)	300	✓	3	0.11
RealEstate10K [41]	▲	D [12]	128	✓	1	0.09
		R [7]	10	✓	1 3	0.08 0.03
		P [16]	6474	✓	2	0.04
		V [9]	10	✓	2	0.11
LLFF [37]	▲	R [7]	8	✓	1 3	0.04 0.03
DTU [38]	▲	R [7]	15	✓	1 3	0.07 0.06
CO3D [39]	▲	R [7]	20	✓	3	0.09
		V [9]	10	✓	2	0.09
WildRGB-D [40]	▲	O_e (1/3 orbit) O_h (full orbit)	20	✓	3 6	0.07 0.11
Mip-NeRF360 [42]	▲	R [7]	9	✗	6	0.11
DL3DV-140 [43]	▲	O	10	✓	6	0.10
		L [18]	140	✓	32	0.05
Tanks and Temples [44]	▲	V [9]	22	✓	2	0.10
		L [18]	2	✓	32	0.10
Large-viewpoint NVS						
OmniObject3D [35]	📦	S [19] (dynamic orbit)	308	✓	1	0.16
GSO [36]	📦	S [19] (dynamic orbit)	300	✓	1	0.18
CO3D [39]	▲	R [7]	20	✓	1	0.15
WildRGB-D [40]	▲	O_h (full orbit)	20	✓	1 3	0.19 0.14
Mip-NeRF360 [42]	▲	R [7]	9	✗	1 3	0.19 0.13
DL3DV-140 [43]	▲	O	10	✓	1 3	0.21 0.12
Tanks and Temples [44]	▲	O	2	✓	1 3 6 9	0.21 0.18 0.16 0.14