

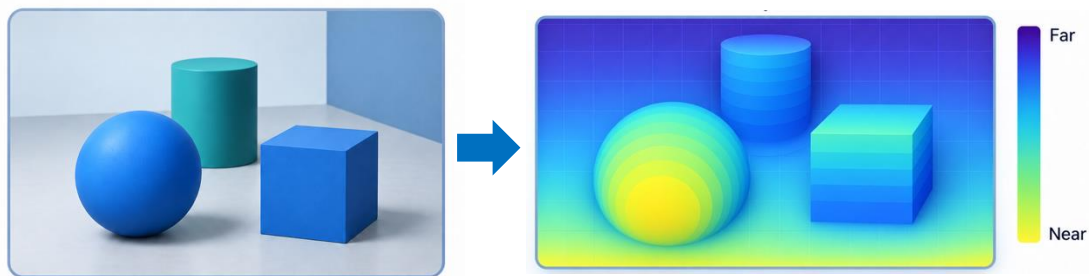
Building and Exploiting Vision Models for Spatial Intelligence

Xingang Pan

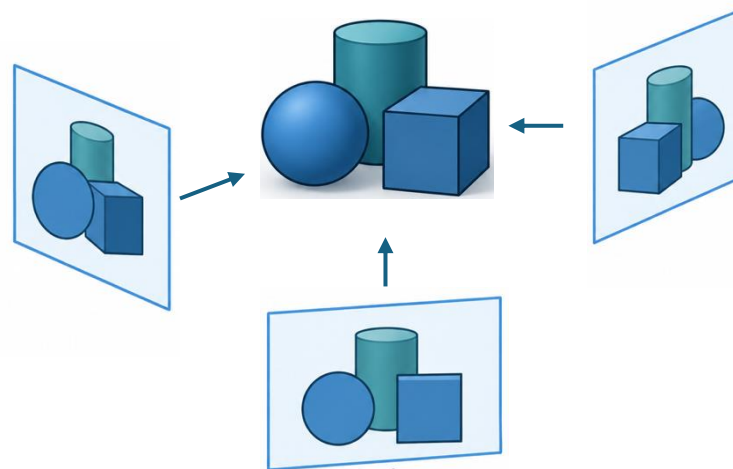
Assistant Professor

Nanyang Technological University

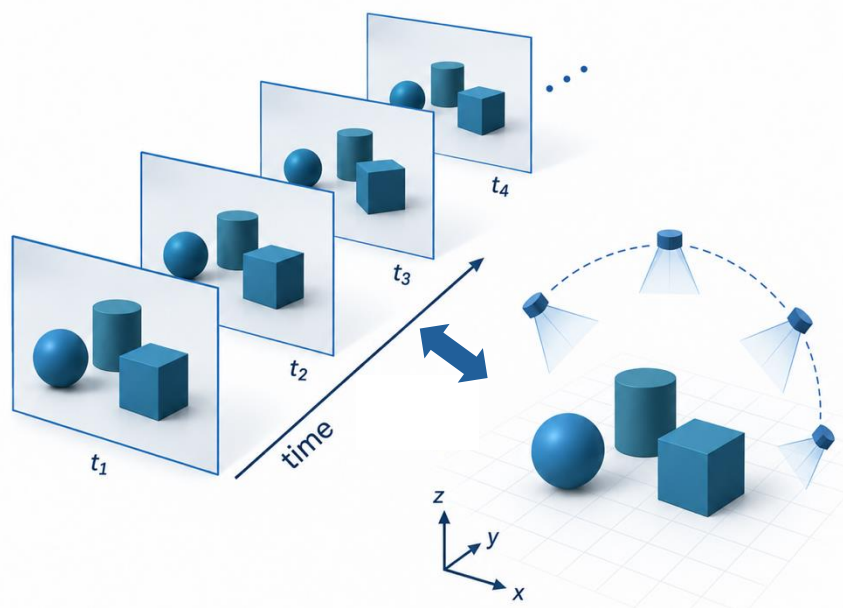
<https://xingangpan.github.io/>



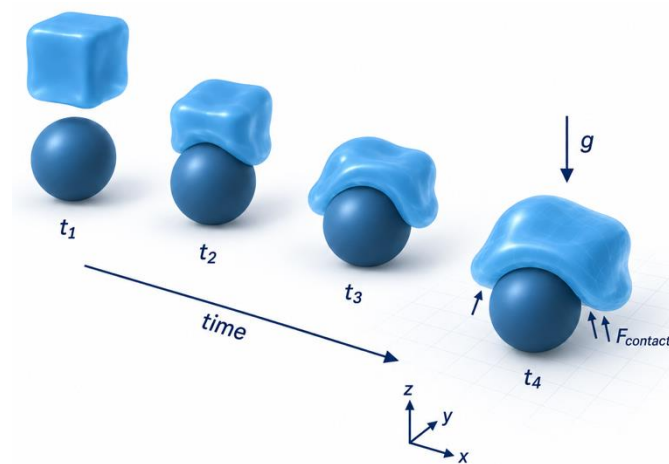
Depth Estimation



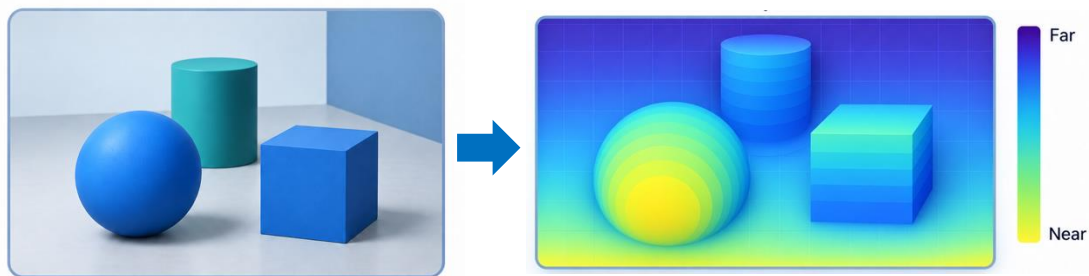
3D Reconstruction



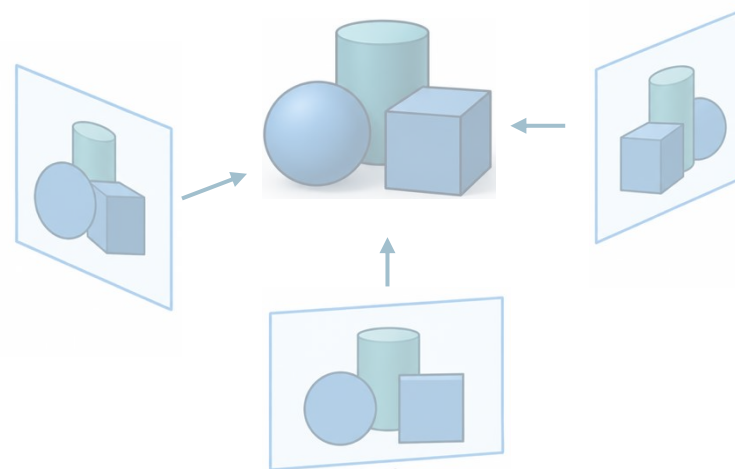
3D in Video Generative Models



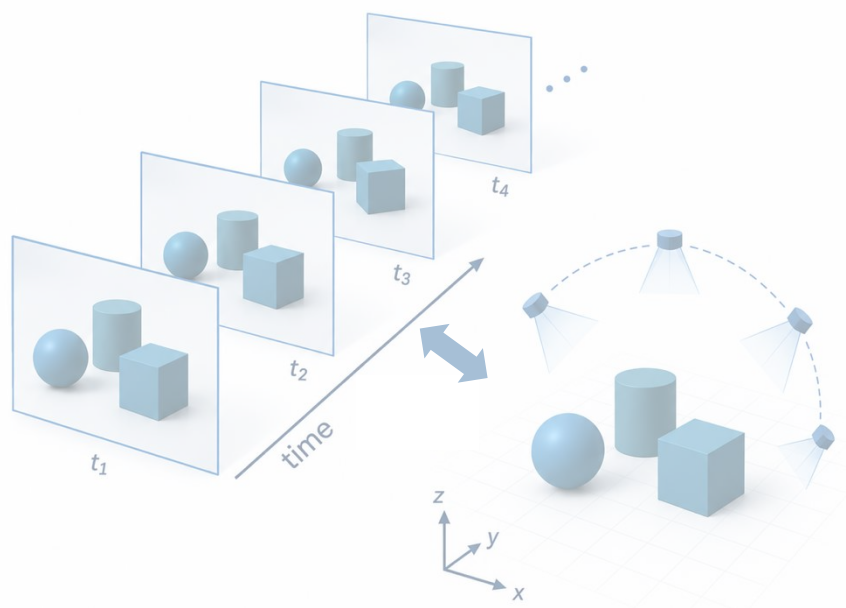
3D Simulation



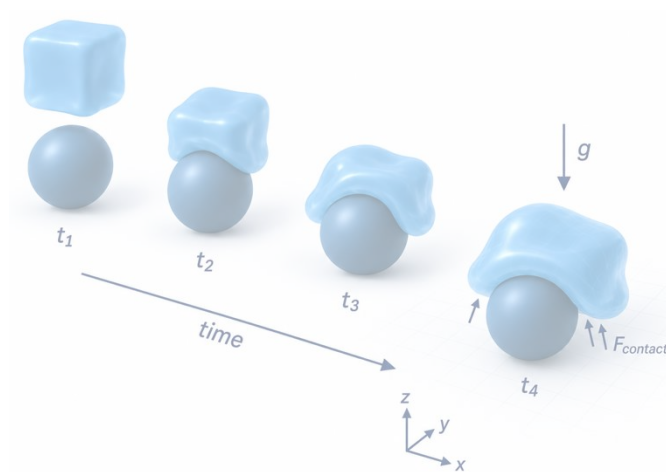
Depth Estimation



3D Reconstruction

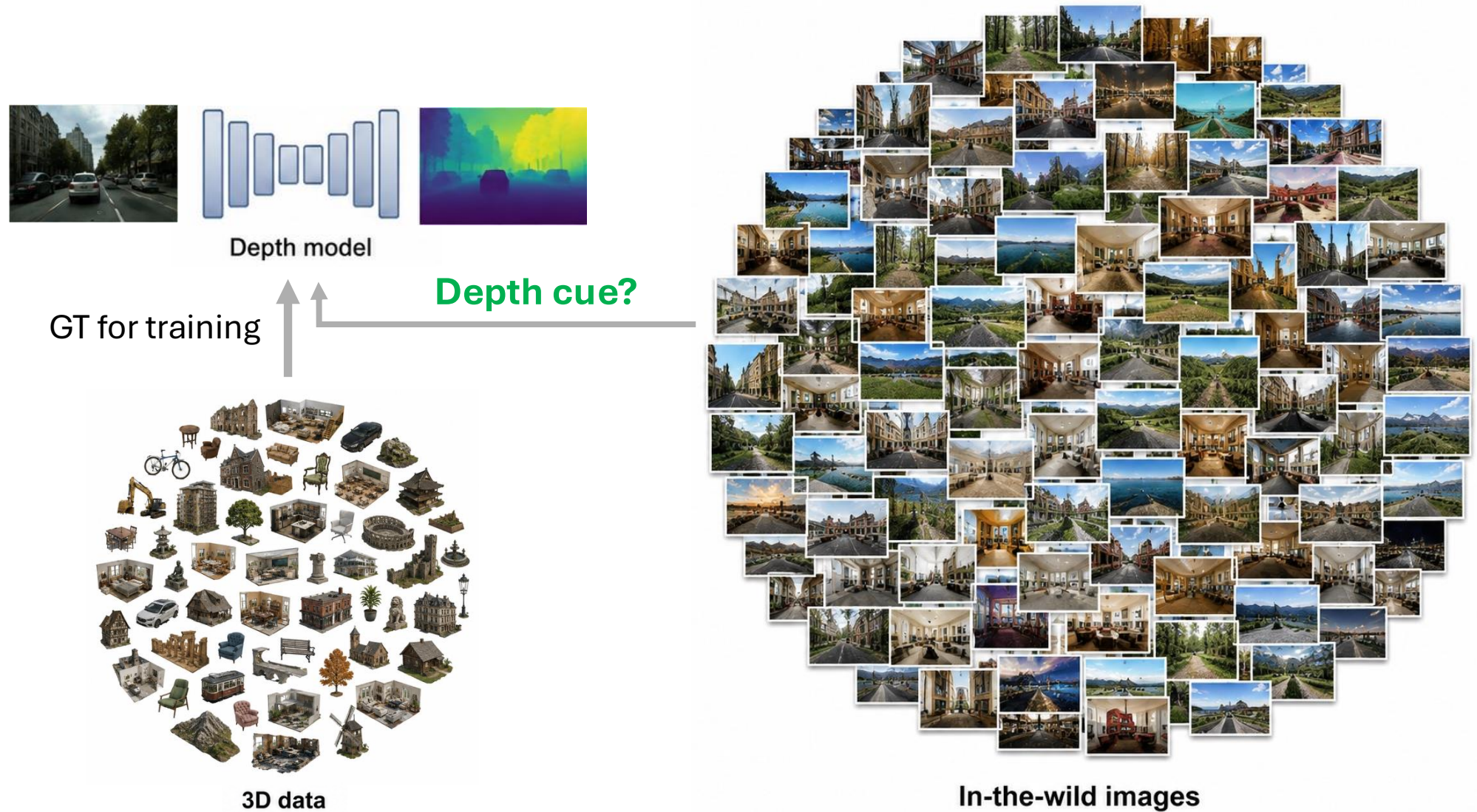


3D in Video Generative Models



3D Simulation

Depth Estimation Data Source



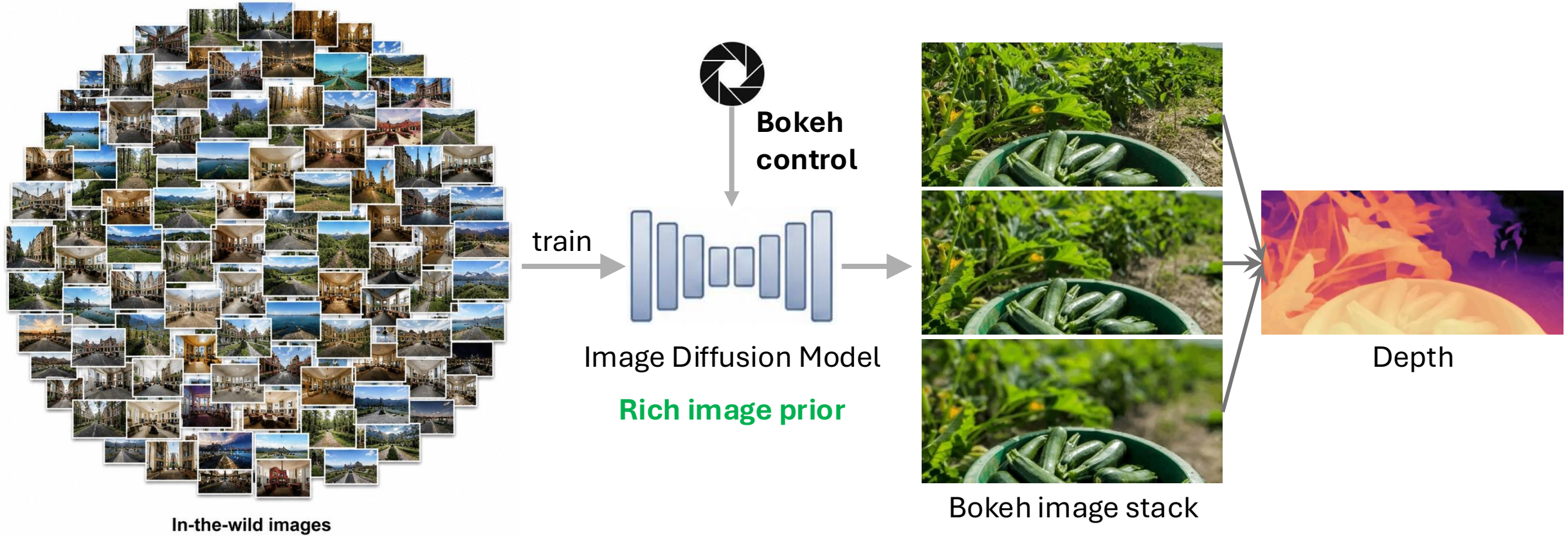
Depth from Defocus



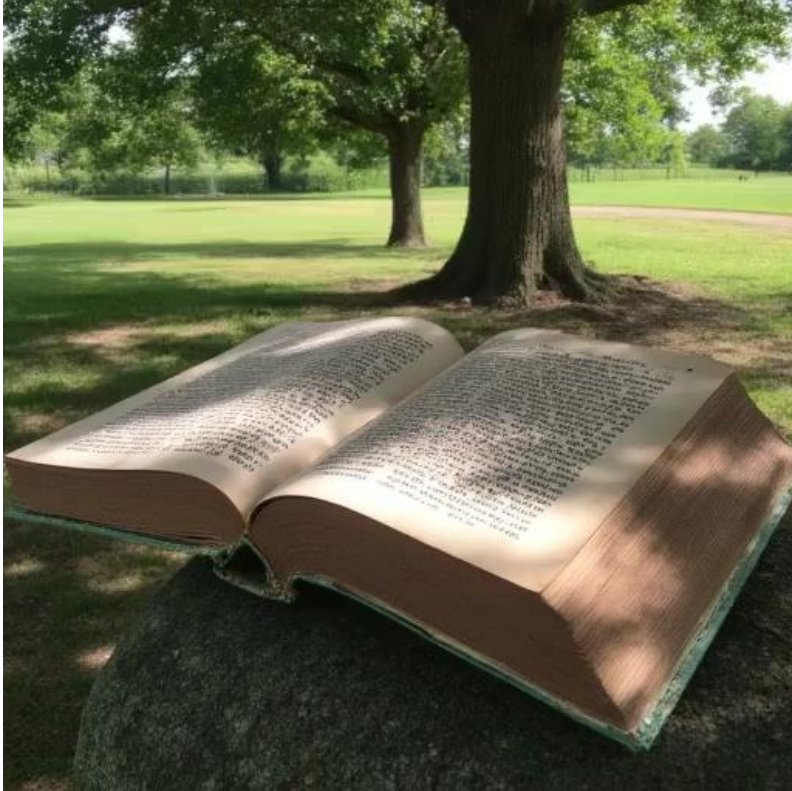
Bokeh Effects Widely Exist ...



Motivation



Bokeh Diffusion



Well-loved book lies forgotten
beneath a towering tree



Train track cutting through a deep
valley, dynamic perspective



Skateboards leaning on a wall outside
a skatepark

Data source 1: In-the-wild images with EXIF data



```
{  
  FNumber: 11,  
  FocalLength: 18mm,  
  FocalLength35mm: 36mm,  
  ...  
}
```



```
{  
  FNumber: 3.2,  
  FocalLength: 9mm,  
  FocalLength35mm: 53mm,  
  ...  
}
```



```
{  
  FNumber: 4,  
  FocalLength: 80 mm,  
  FocalLength35mm: 120mm,  
  ...  
}
```



```
{  
  FNumber: 7.1,  
  FocalLength: 55mm,  
  FocalLength35mm: 82mm,  
  ...  
}
```



```
{  
  FNumber: 4.5,  
  FocalLength: 45mm,  
  FocalLength35mm: 45mm,  
  ...  
}
```



```
{  
  FNumber: 5,  
  FocalLength: 116mm,  
  FocalLength35mm: 116mm,  
  ...  
}
```

Synthetic Defocus Blur Augmentations

(1) Select AiF Images



Image data stack

$$\sim S_1$$

Empirical focus
distance

(2) Calculate defocus map
& sample blur radius K



Defocus map

$$K \sim \mathcal{U}$$

Sampled blur radius

(3) Render synthetic
augmentations



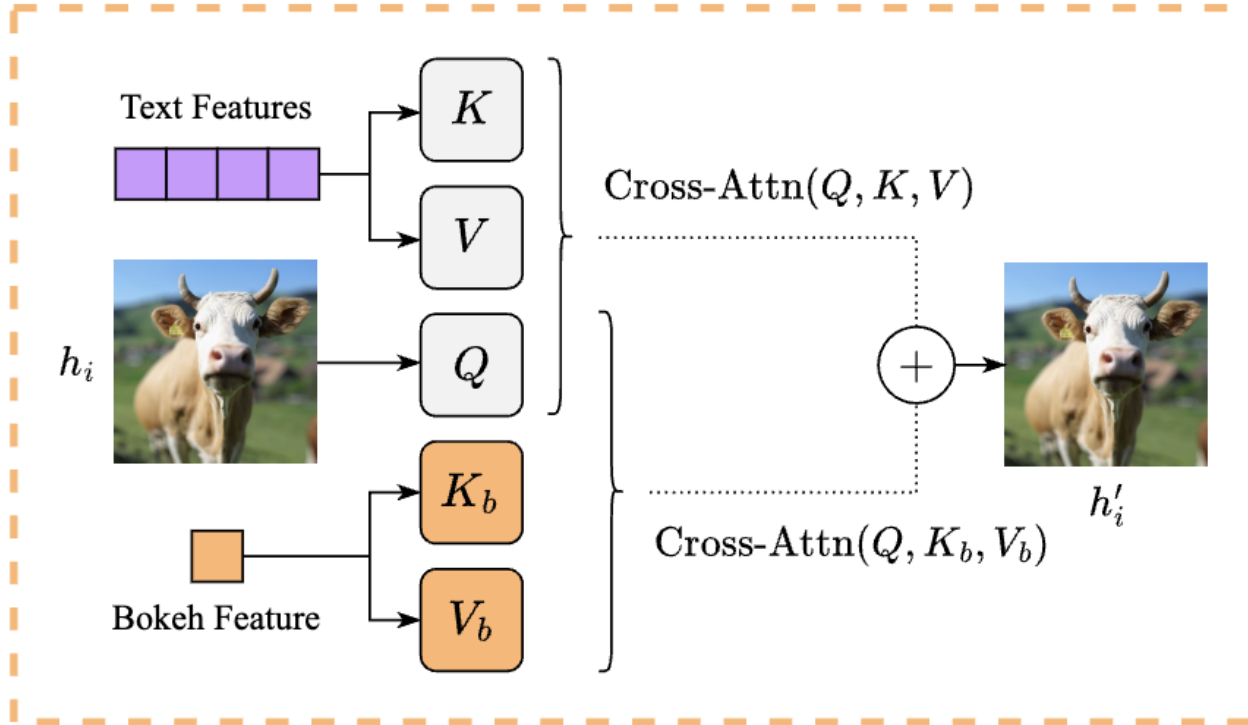
...



Contrastive Batch

Defocus Blur Conditioning

Text & Bokeh Cross-Attention (Sec. 3.3)



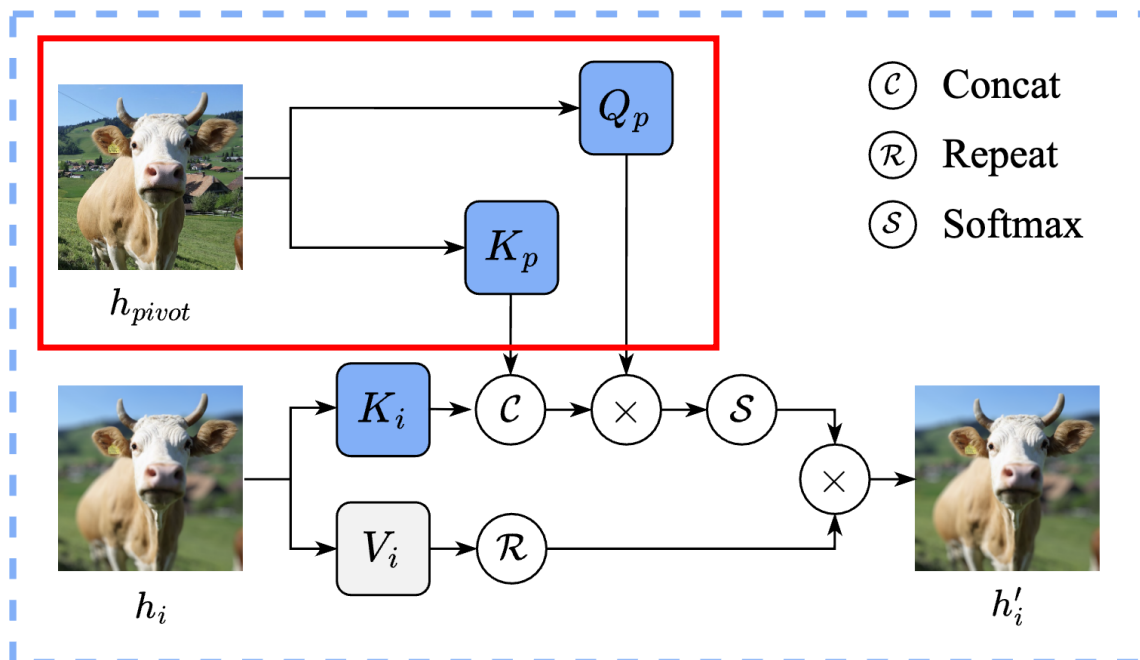
$$\text{Attention}(Q, K_{txt}, V_{txt}) + \lambda \text{Attention}(Q, K_{bkh}, V_{bkh})$$



Defocus Blur Conditioning

Imposing Scene Consistency

Grounded Self-Attention (Sec. 3.4)



$$\text{Attention}(Q_{tgt}, K_{tgt}, V_{tgt})$$

$$\text{Attention}(\boxed{Q_{piv}}, [K_{tgt}, \boxed{K_{piv}}], [V_{tgt}, V_{tgt}])$$



Defocus Blur Conditioning



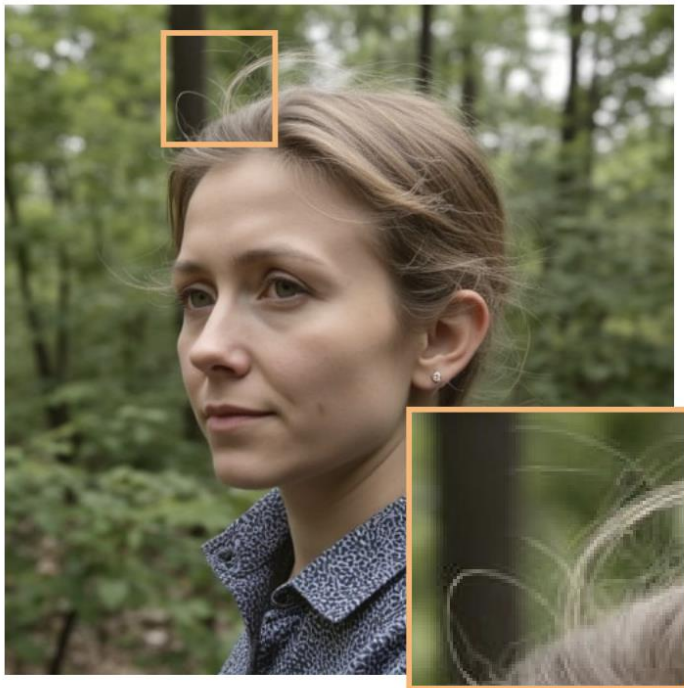
Imposing Scene Consistency

Bokeh Diffusion – Bokeh strength control



Comparison with Conventional Pipeline (BokehMe)

Base Image
(Ours)



Depth Estimation
+ Bokeh Renderer



Bokeh Generation
(Ours)

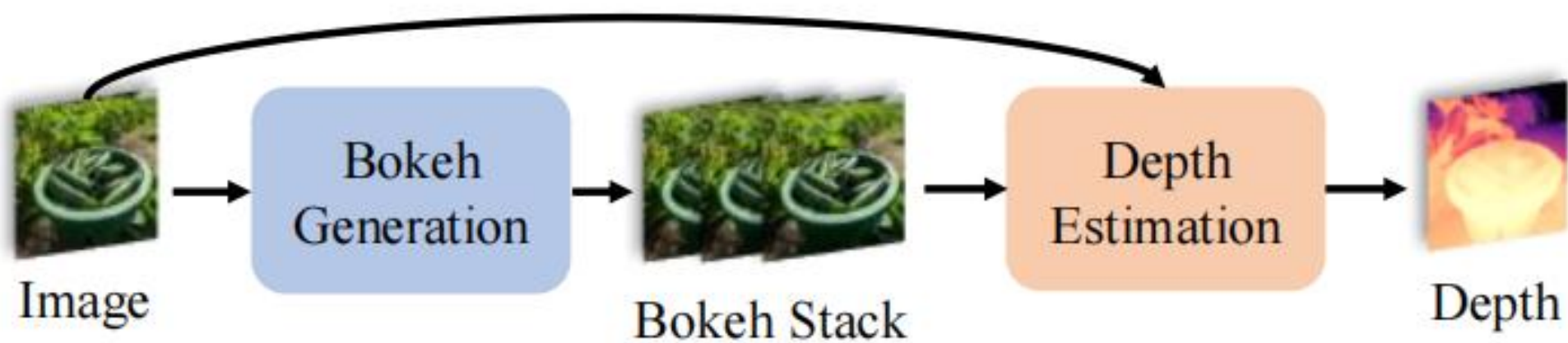


Fig. 3. Comparison between our generative method and BokehMe [\[Peng et al. 2022\]](#), which depends on external depth estimation.



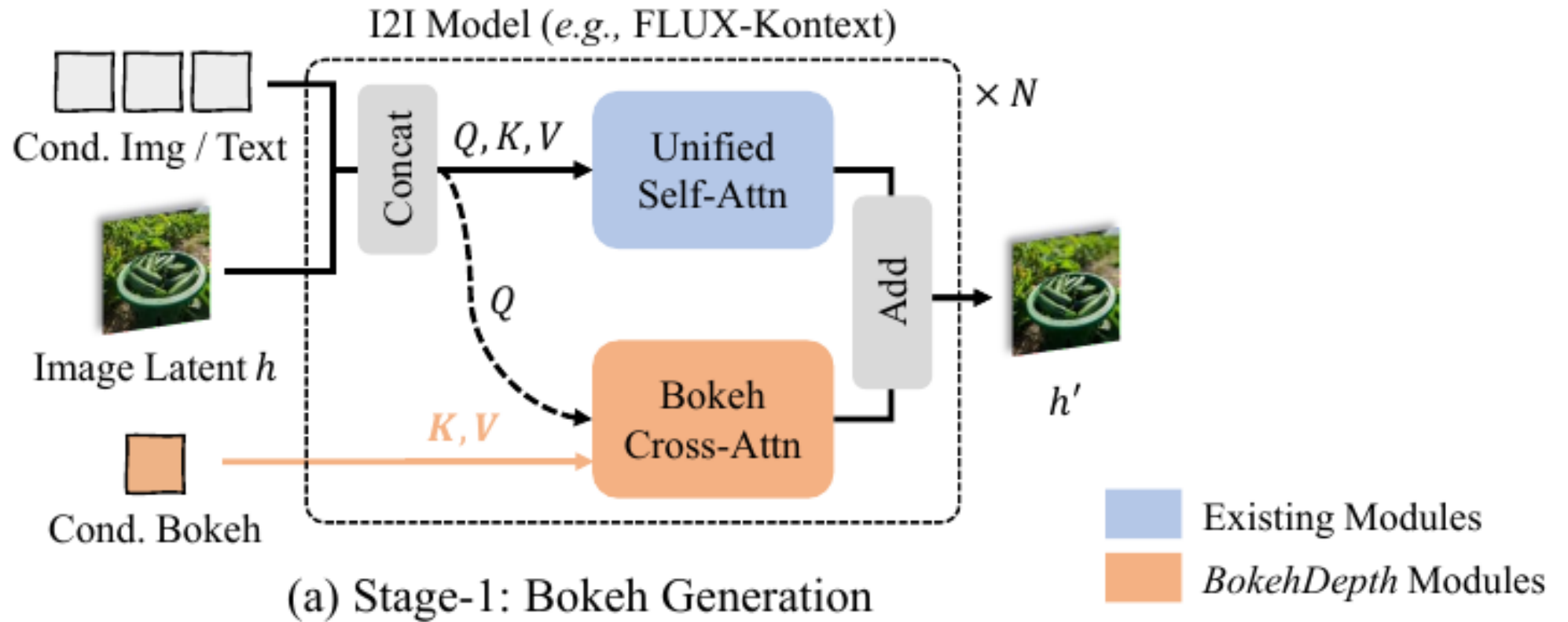
Fig. 9. Additional comparison between our generative method and BokehMe [Peng et al. 2022], which depends on external depth estimation models.

BokehDepth



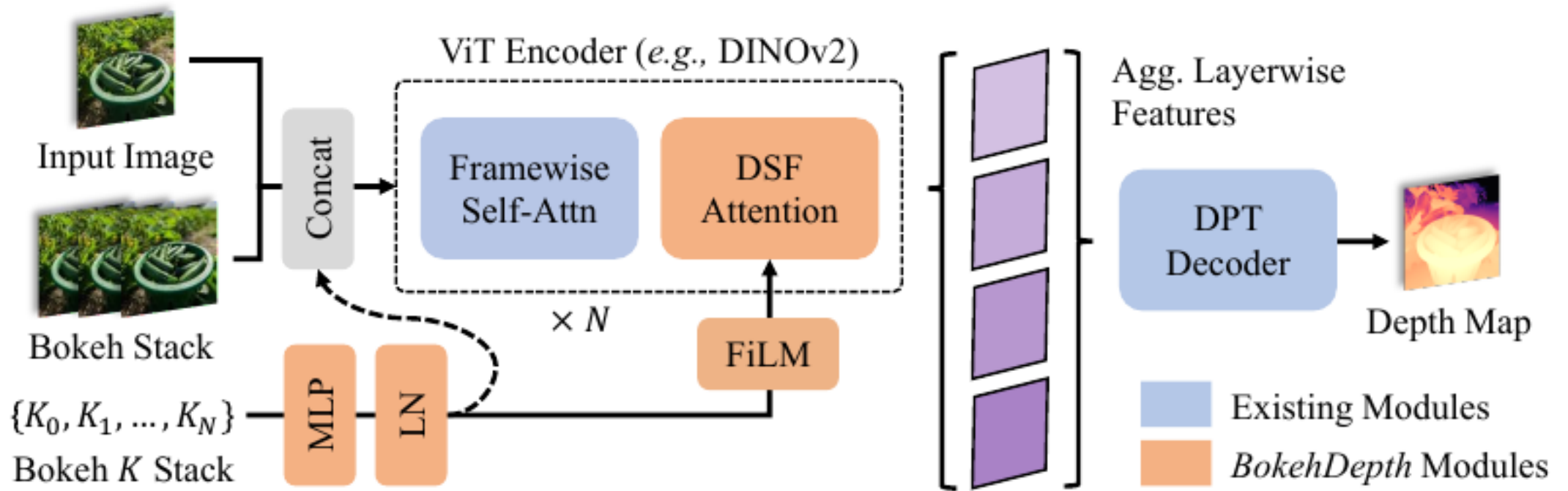
Overview of *BokehDepth*

BokehDepth Overview



Stage 1 adds K-conditioned bokeh cross-attention to a pretrained image editor.

BokehDepth Overview



(b) Stage-2: Bokeh Stack Fusion for Depth

Stage 2 inserts Divided Space Focus Attention (DSFA) into the encoder while keeping the DPT decoder unchanged.

Qualitative Results – Stage 1 + Stage 2



Quantitative Results – Stage 2

Method	HAMMER [42]		IBims-1 [44]		Middlebury [85]		Make3D [84]		Sintel [10]		ETH3D [87]	
	$\delta_1 \uparrow$	Abs_rel $\downarrow$	$\delta_1 \uparrow$	Abs_rel $\downarrow$	$\delta_1 \uparrow$	Abs_rel $\downarrow$	$\delta_1 \uparrow$	Abs_rel $\downarrow$	$\delta_1 \uparrow$	Abs_rel $\downarrow$	$\delta_1 \uparrow$	Abs_rel $\downarrow$
ZoeDepth [6]	0.009	0.943	0.672	0.174	0.538	<u>0.214</u>	0.649	0.221	0.078	0.946	0.338	0.500
Metric3Dv2 [34]	0.653	0.357	0.941	0.100	0.299	0.450	0.512	0.394	0.383	0.370	0.987	0.042
DepthPro [8]	0.630	0.391	0.823	0.159	0.605	0.251	0.728	0.254	0.400	0.508	0.415	0.327
DAv2 [113]	0.828	0.134	0.938	0.080	0.618	0.240	0.726	<u>0.217</u>	0.538	0.569	0.910	0.091
+ <i>BokehDepth</i>	<u>0.894</u>	<u>0.105</u>	<u>0.960</u>	<u>0.064</u>	<u>0.675</u>	0.215	<u>0.740</u>	0.206	<u>0.542</u>	0.413	0.954	0.077
UDv2 [74]	0.645	0.293	0.945	0.082	0.535	0.288	0.739	0.263	0.344	0.496	0.852	0.160
+ <i>BokehDepth</i>	0.895	0.094	0.978	0.039	0.716	0.205	0.786	0.228	0.671	<u>0.391</u>	<u>0.963</u>	<u>0.074</u>

Table 2. Zero-shot metric-depth comparison across diverse indoor and outdoor datasets. **Boldface** denotes the best result and underline denotes the second best. Our method consistently improves accuracy over each corresponding base model.

Qualitative Results

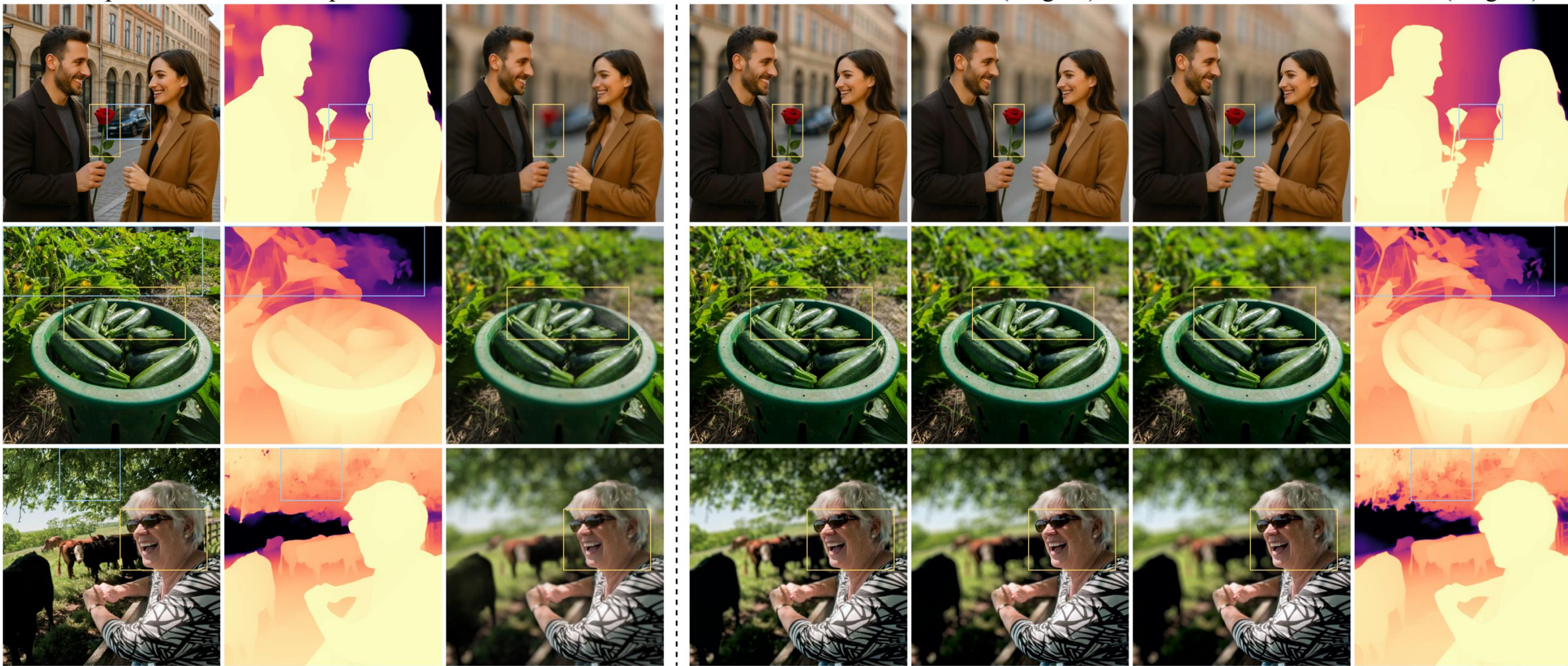
Input

Unidepth V2

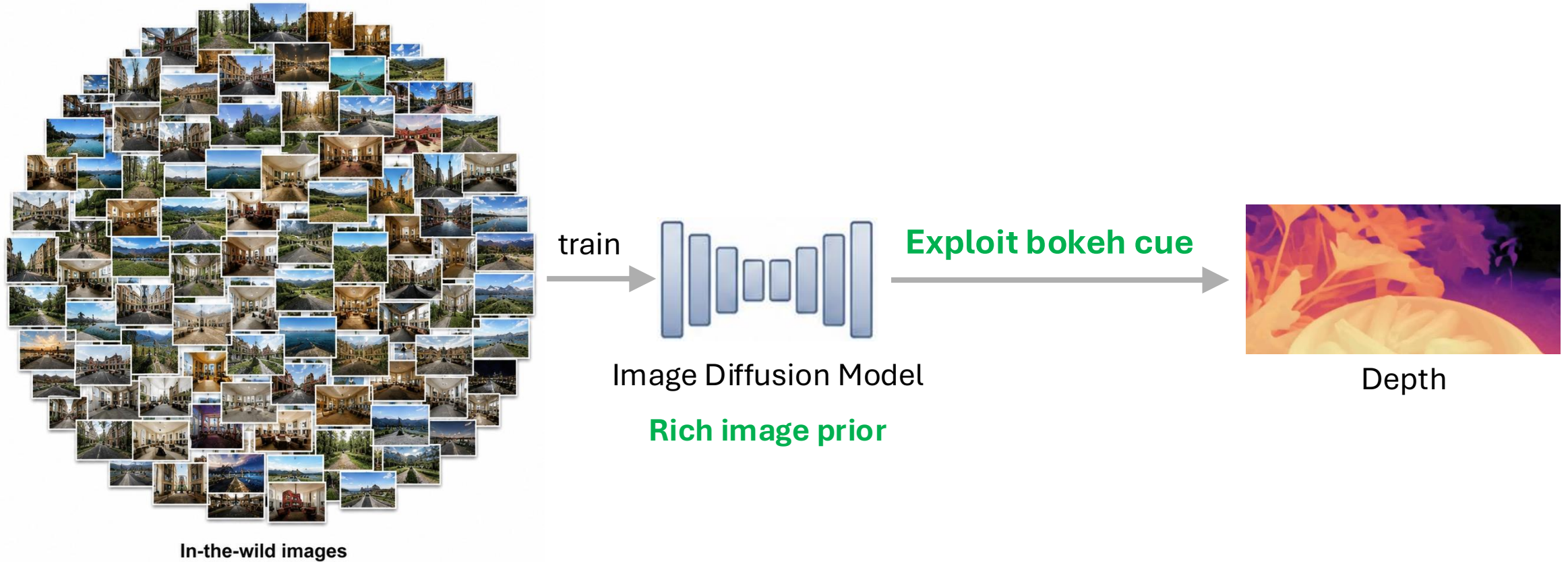
Dr. Bokeh

Ours (Stage-1)

Ours (Stage-2)

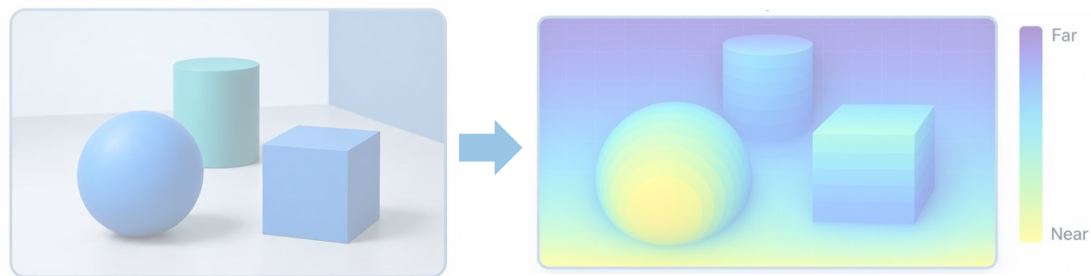


Takeaways and Limitations

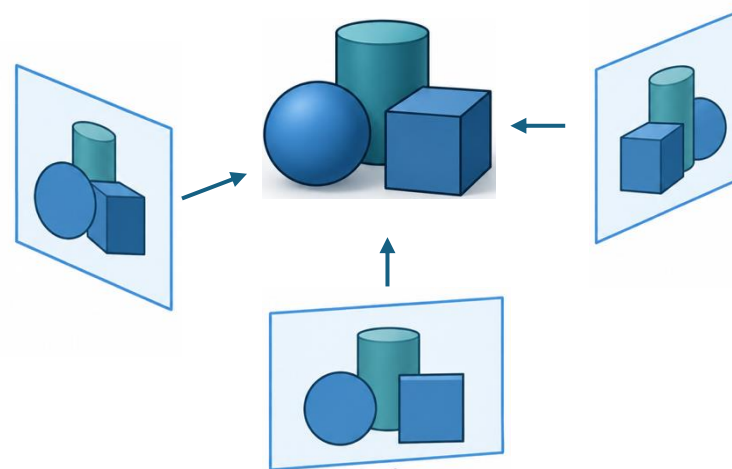


BokehDepth provides a way to **exploit rich depth cues** existing in the large-scale **in-the-wild images**

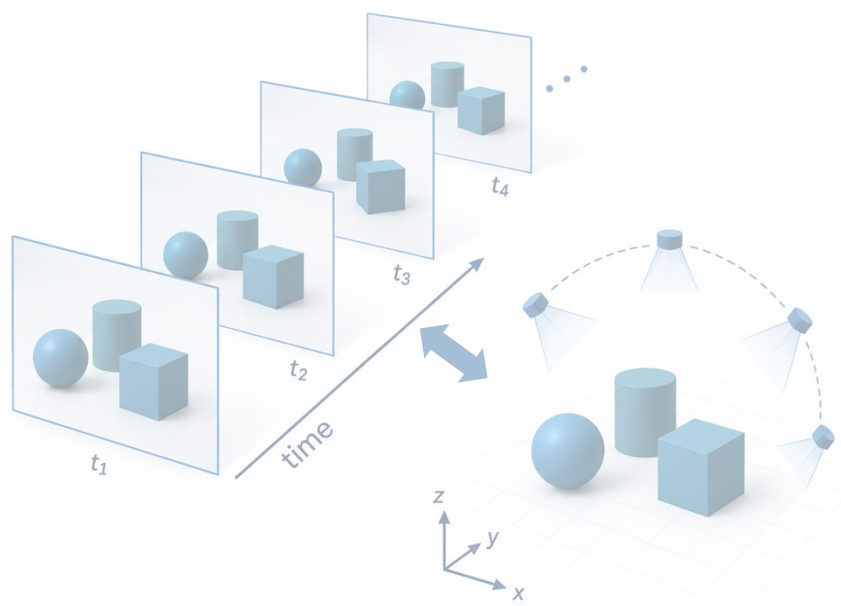
Limitation: Efficiency, 3 x 30 steps diffusion



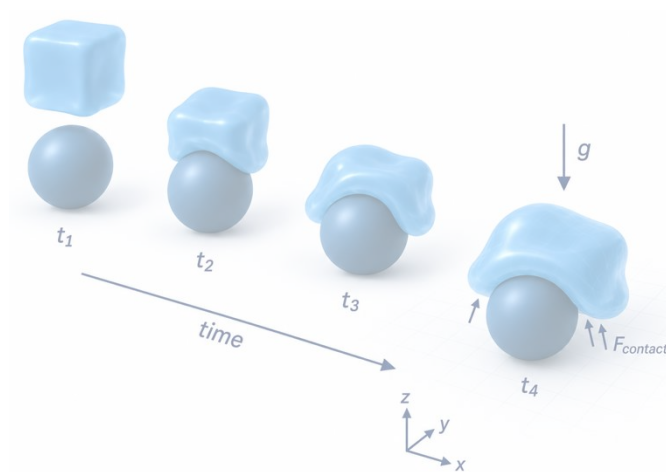
Depth Estimation



3D Reconstruction

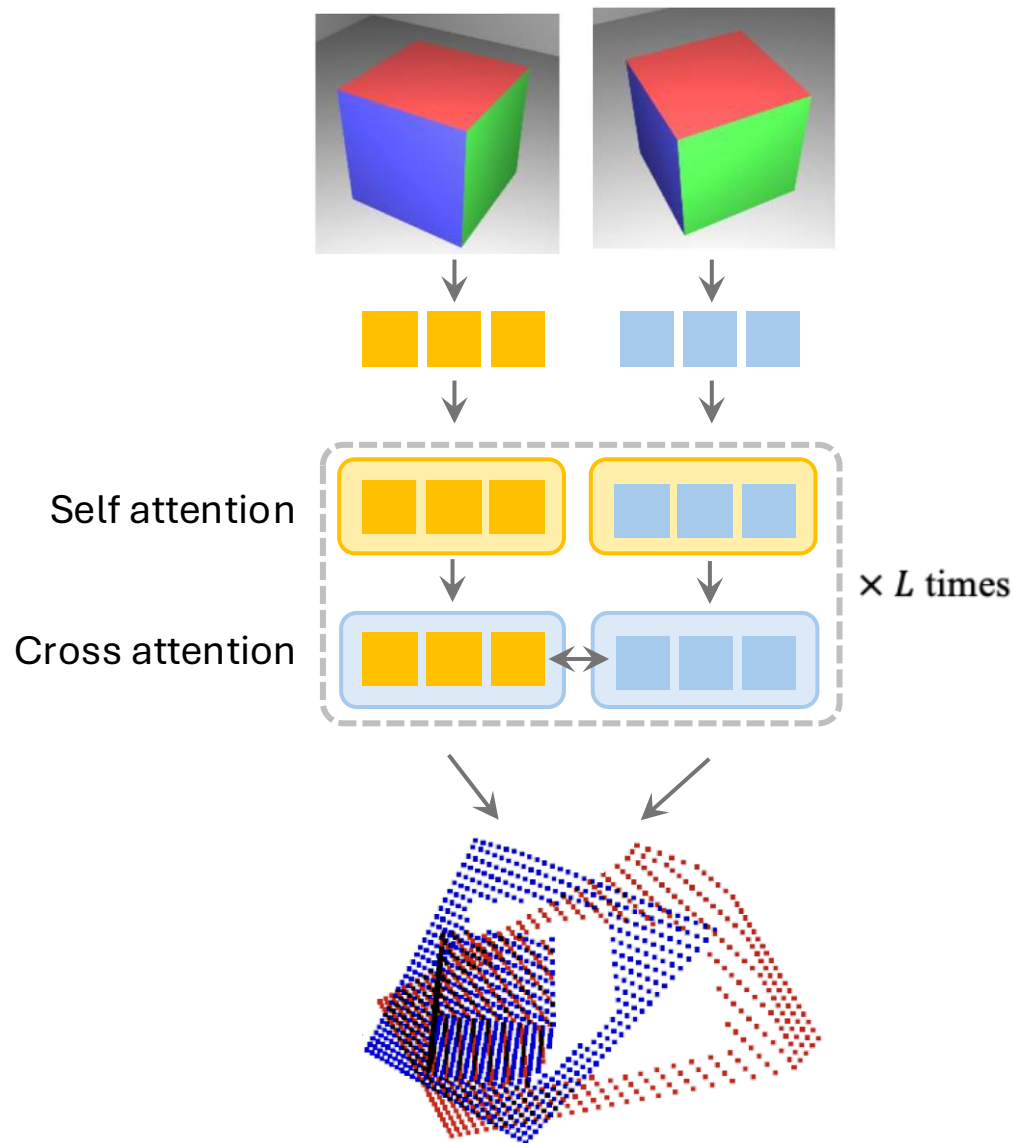


3D in Video Generative Models

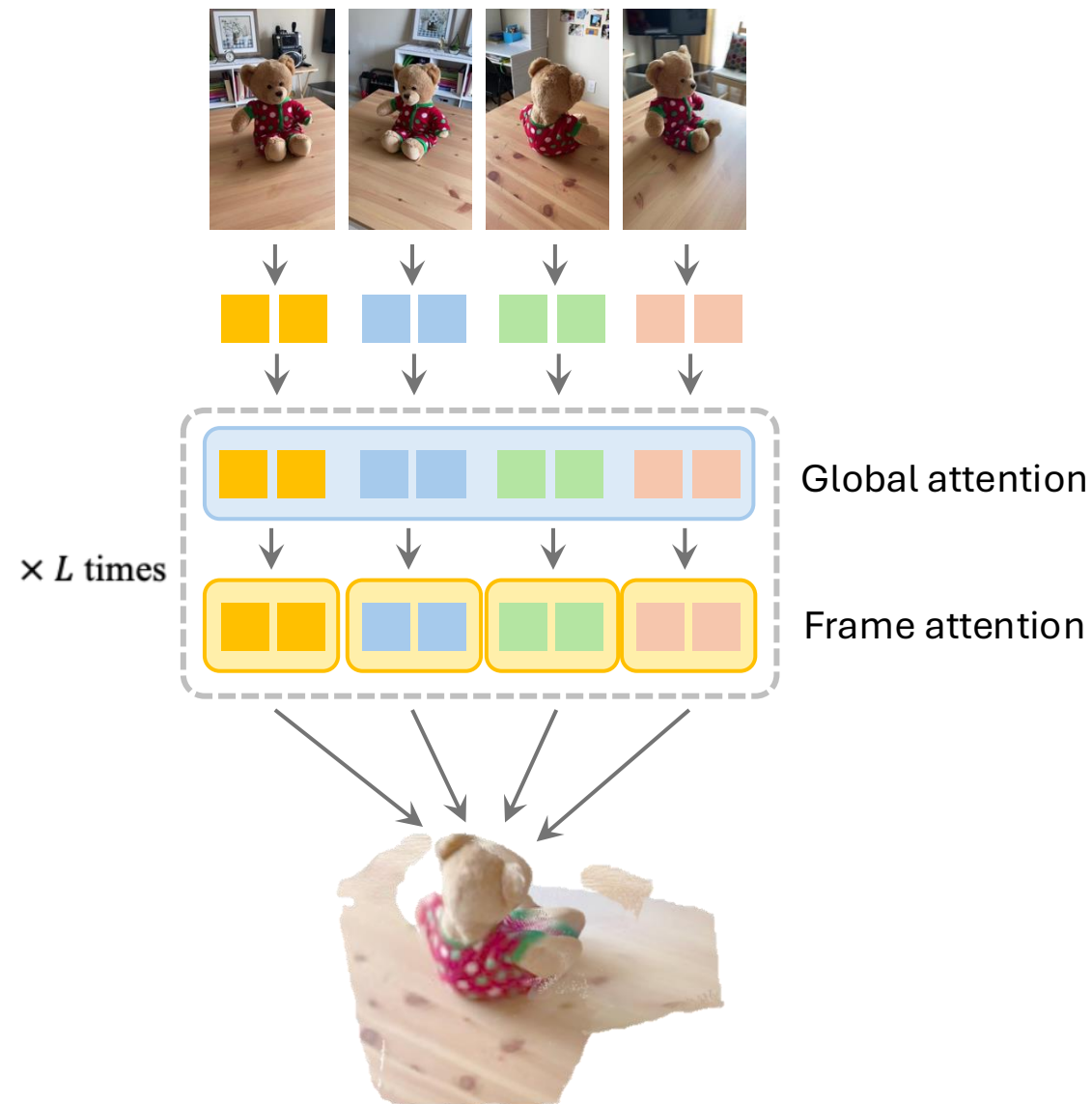


3D Simulation

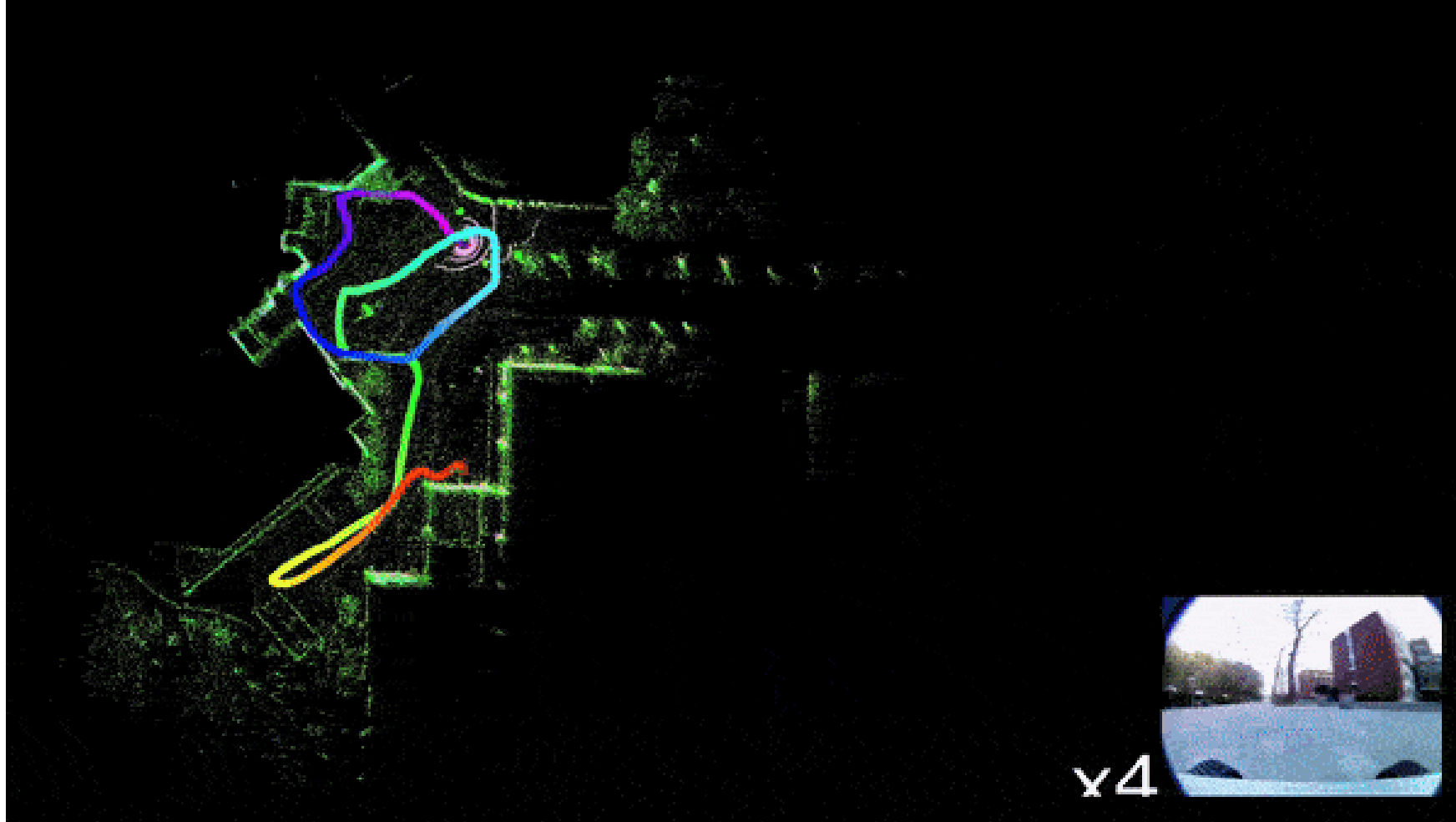
Dust3R



VGGT

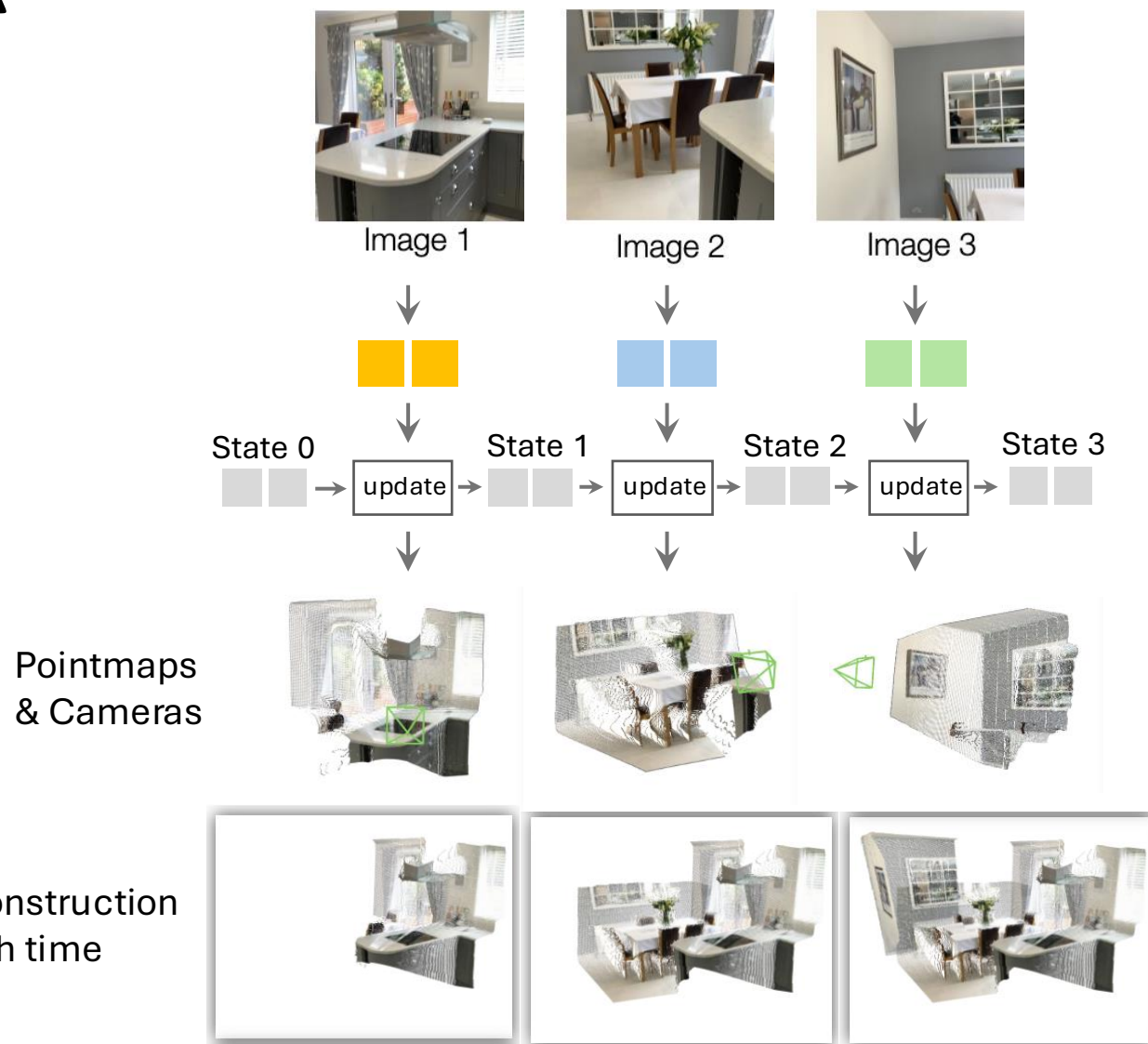


3D Reconstruction from Streaming Input



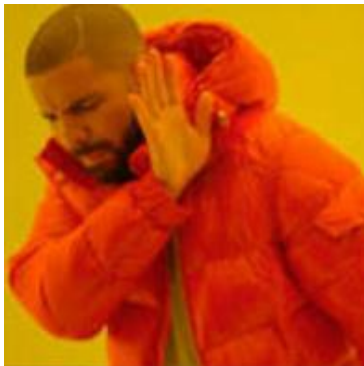
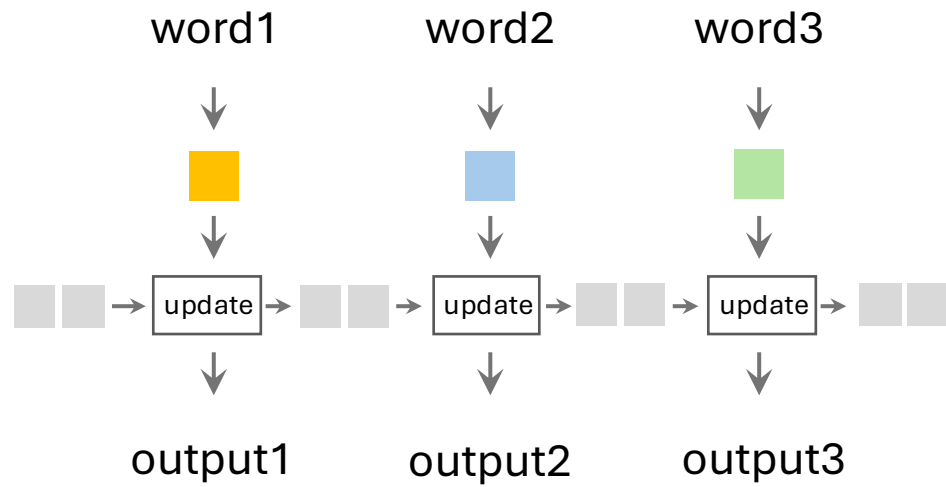
<https://learnopencv.com/monocular-slam-in-python/>

Cut3R

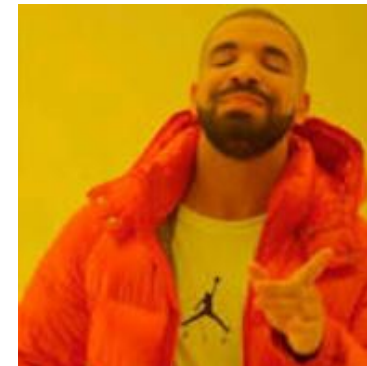
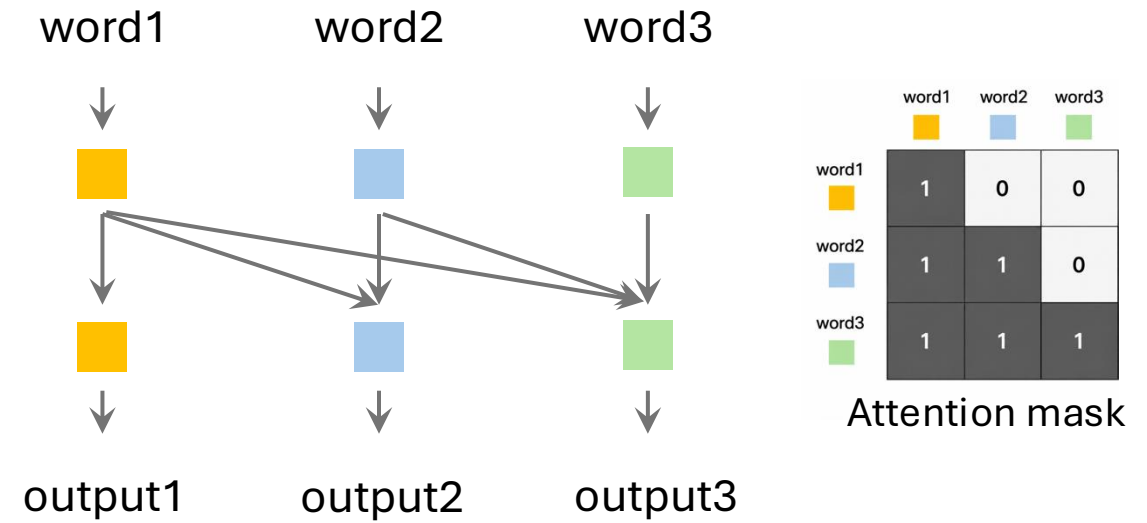


Recurrent Neural Network

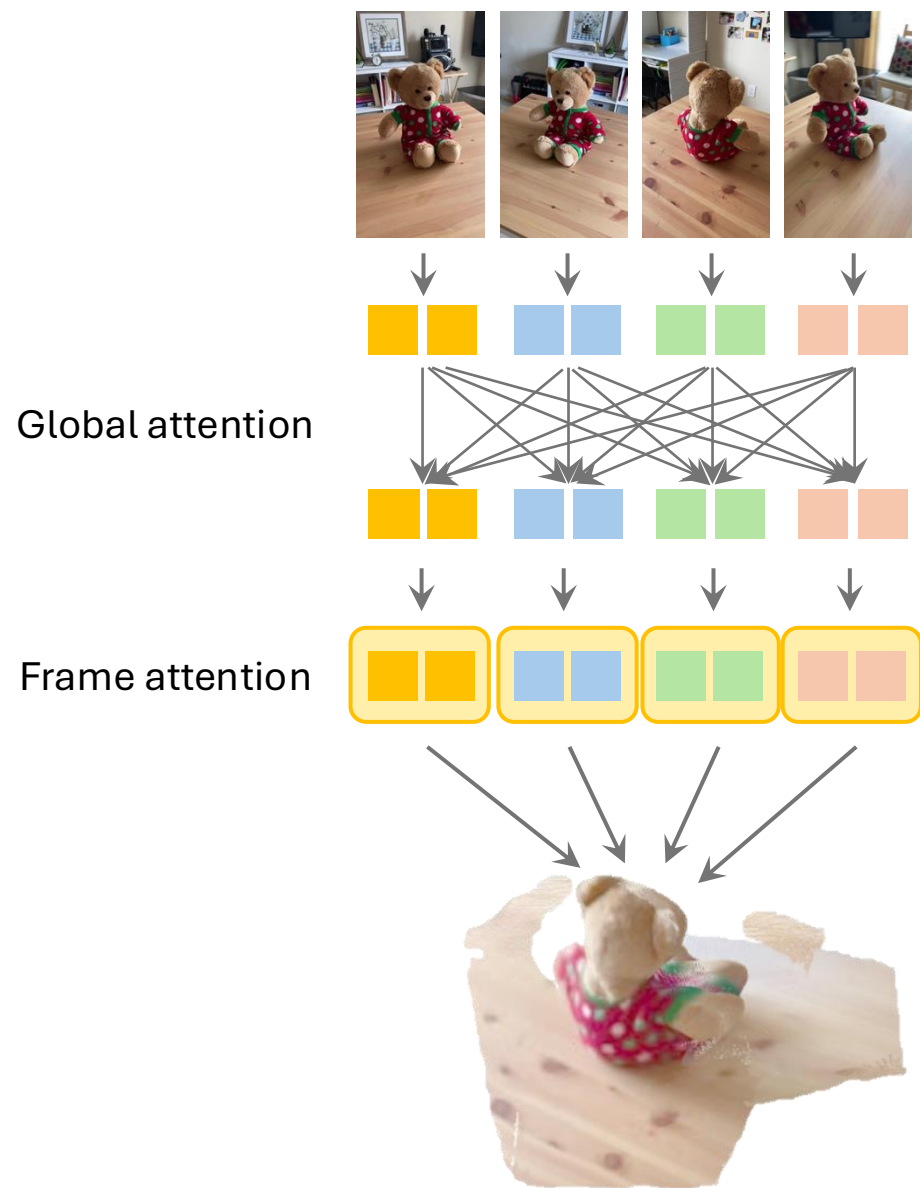
Recurrent Neural Network



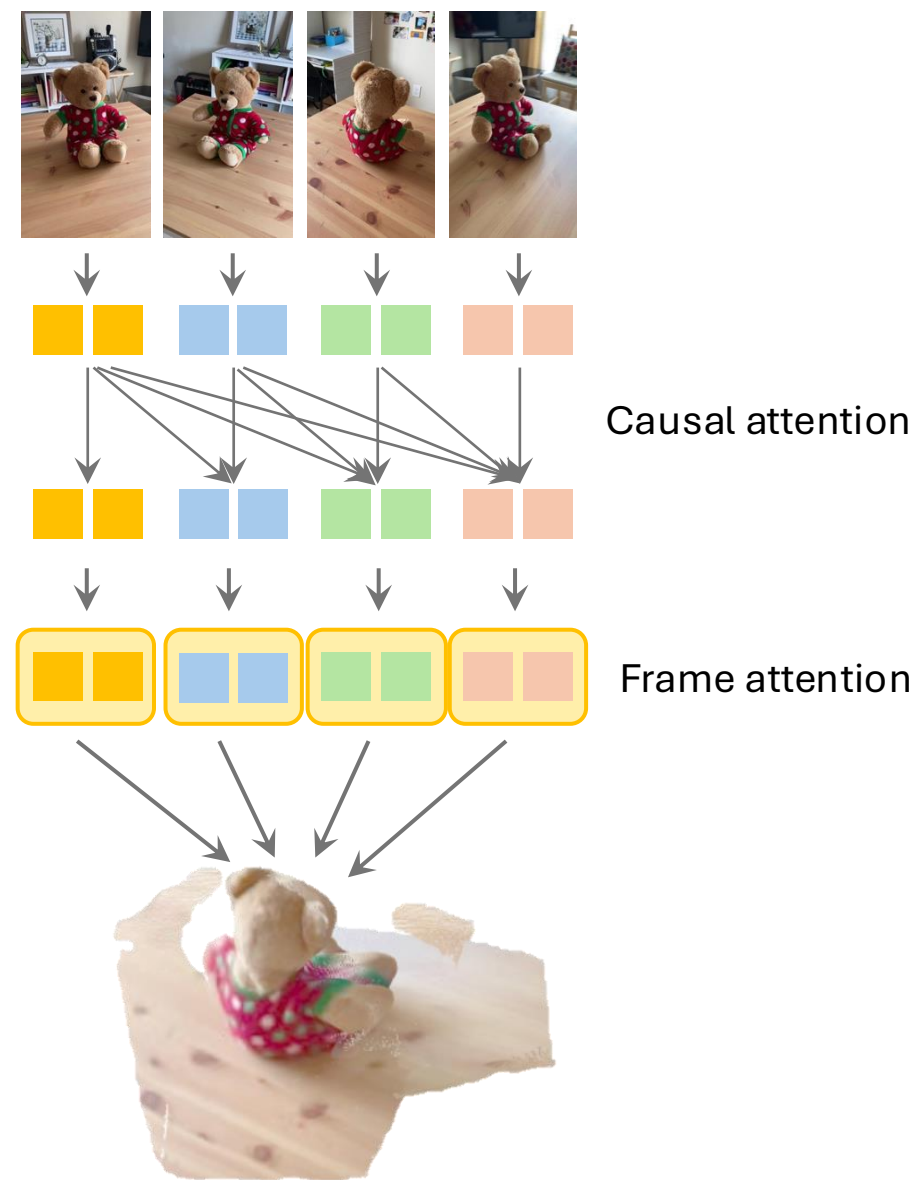
Causal Attention + KV Cache



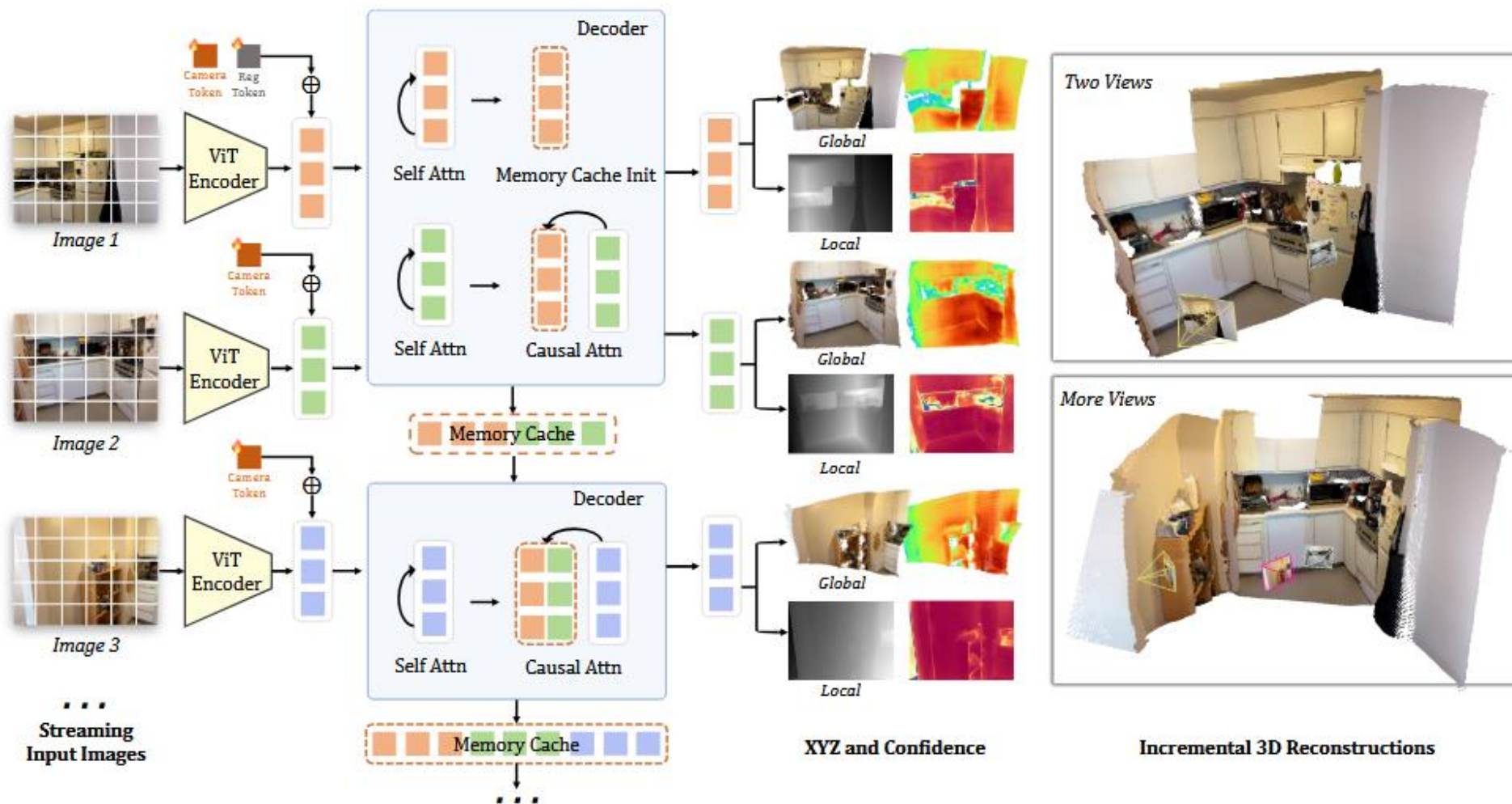
VGGT



STream3R



SStream3R



Memory Cache + Causal Transformer



Input Images



Global 3D Points

Video Input



Streaming reconstruction

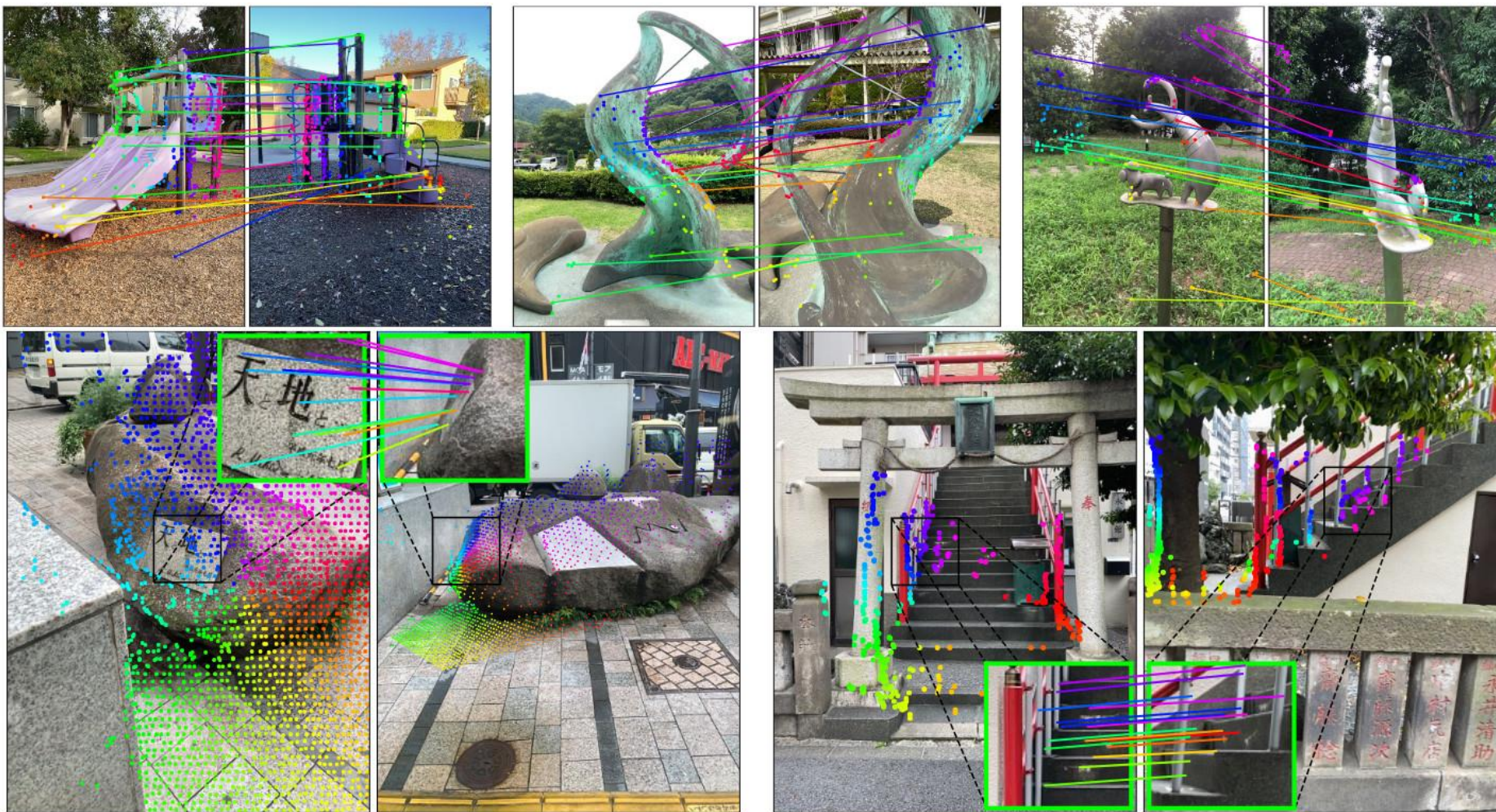


STream3R: Evaluation on 3D Reconstruction

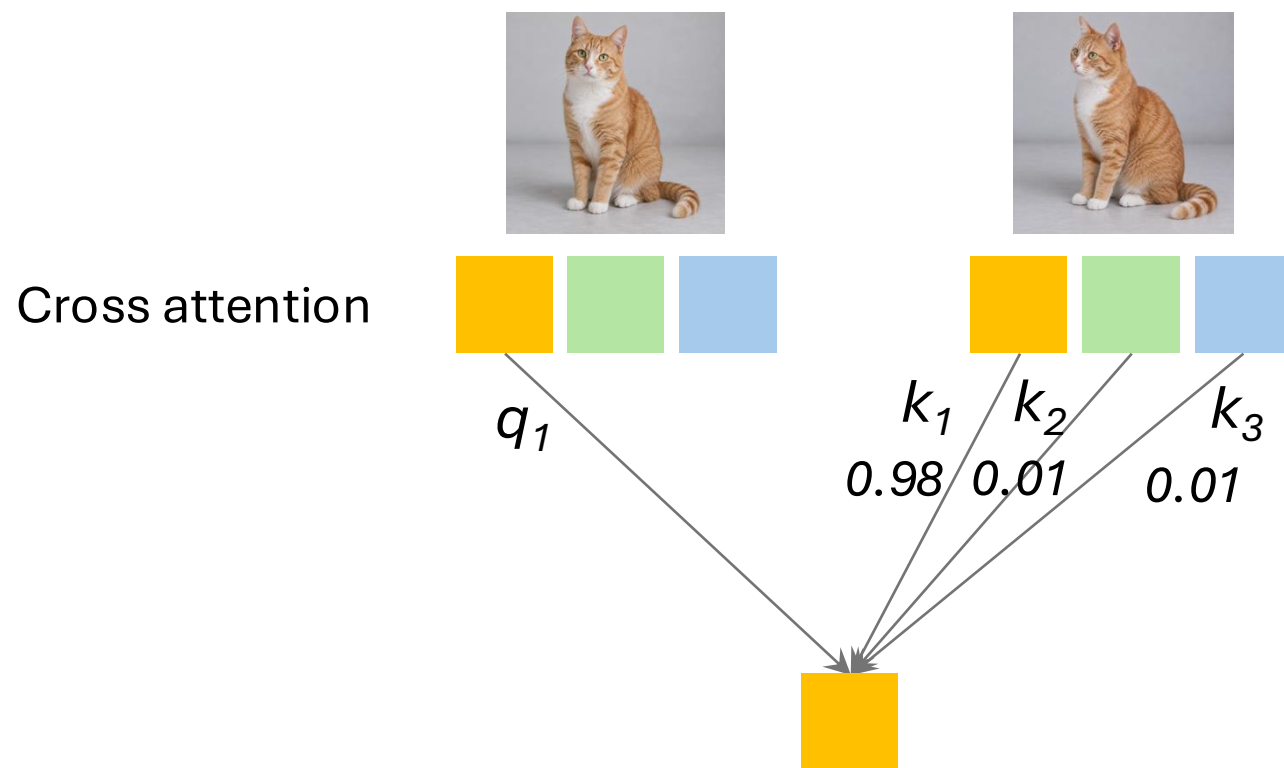
Method	Type	Acc↓		Comp↓		NC↑	
		Mean	Med.	Mean	Med.	Mean	Med.
VGG-T [13]	FA	0.073	0.018	0.077	0.021	0.910	0.990
DUS _t 3R-GA [10]	Optim	0.144	<u>0.019</u>	0.154	<u>0.018</u>	<u>0.870</u>	<u>0.982</u>
MASt3R-GA [11]	Optim	<u>0.085</u>	0.033	<u>0.063</u>	0.028	0.794	0.928
MonST3R-GA [49]	Optim	0.272	0.114	0.287	0.110	0.758	0.843
Spann3R [15]	Stream	0.416	0.323	0.417	0.285	0.684	0.789
CUT3R [17]	Stream	0.099	<u>0.031</u>	0.076	<u>0.026</u>	0.837	0.971
VGG-T [Streaming] [13]	Stream	0.219	0.102	0.212	0.105	0.797	0.936
STREAM3R ^{β}	Stream	0.057	0.014	0.028	0.013	0.910	0.993

3D Reconstruction Comparison on NRGBD

Mast3R



Attention aligns with Matching





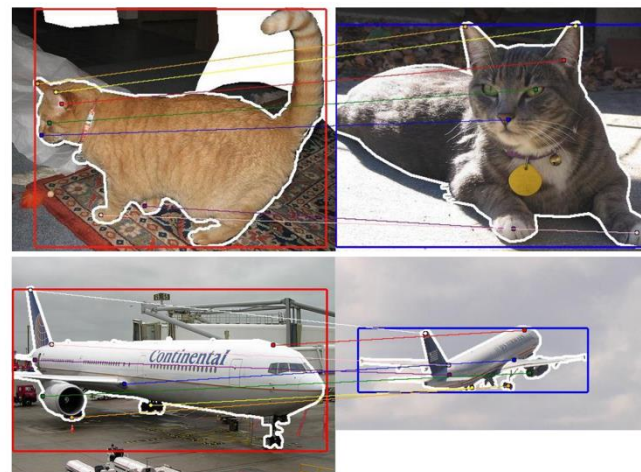
What if ...?



Dataset

SPair-71k

- Human annotated sparse point correspondence
- 71k image pairs

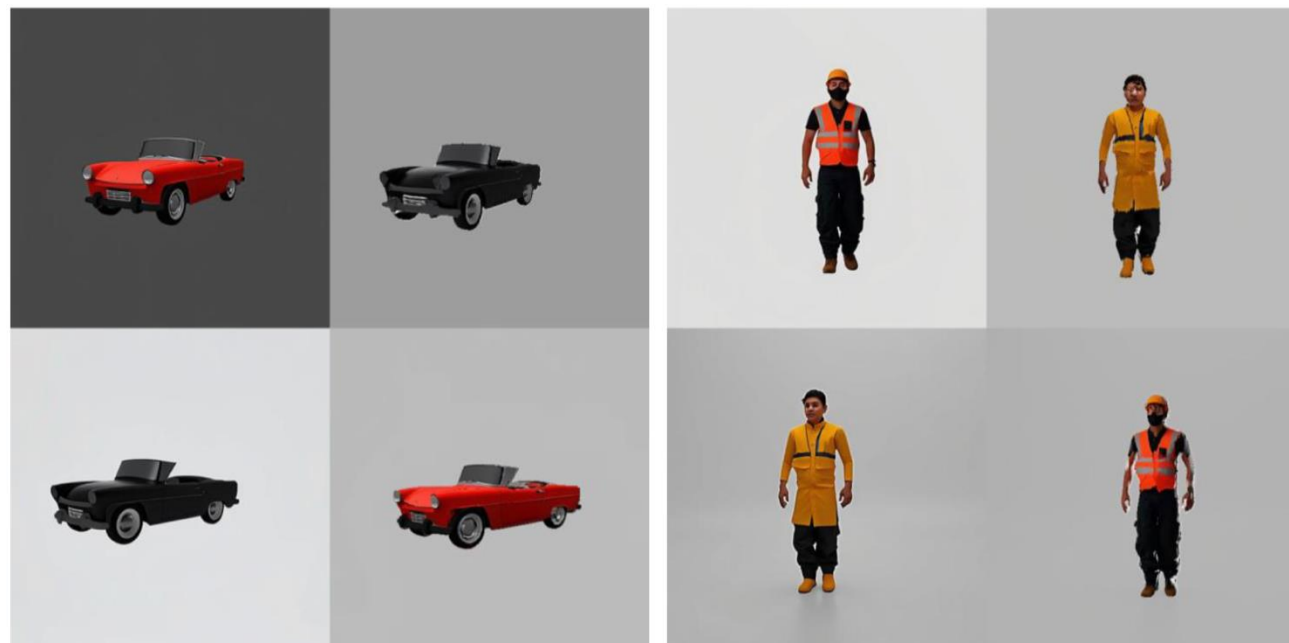


View-point diff.: Medium
Scale diff.: Easy
Truncation: Target truncated
Occlusion: None
Keypoints: 8

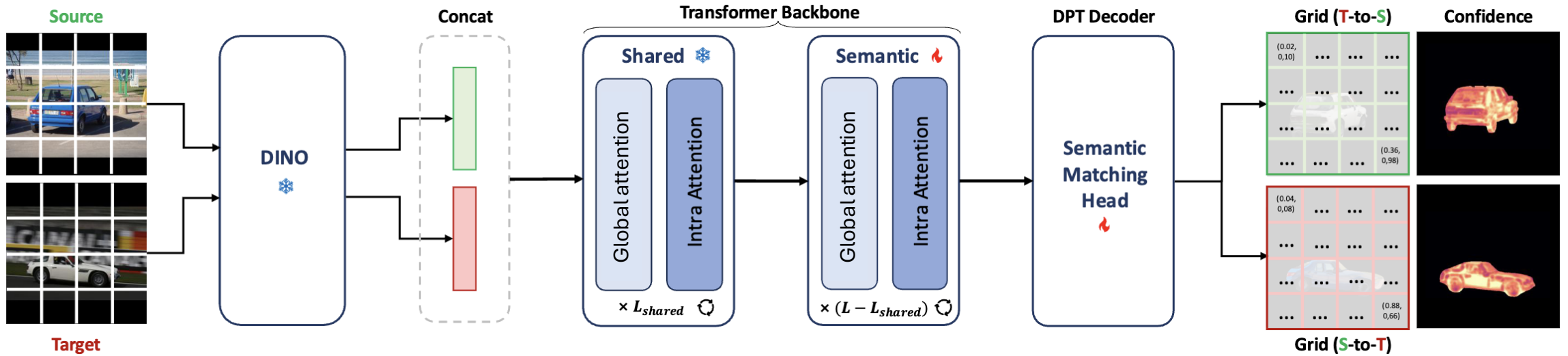
View-point diff.: Medium
Scale diff.: Medium
Truncation: Source truncated
Occlusion: None
Keypoints: 8

Synthetic paired data

- Created via 3D object rendering + Flux appearance diversification
- 500 image pairs
- Dense matching

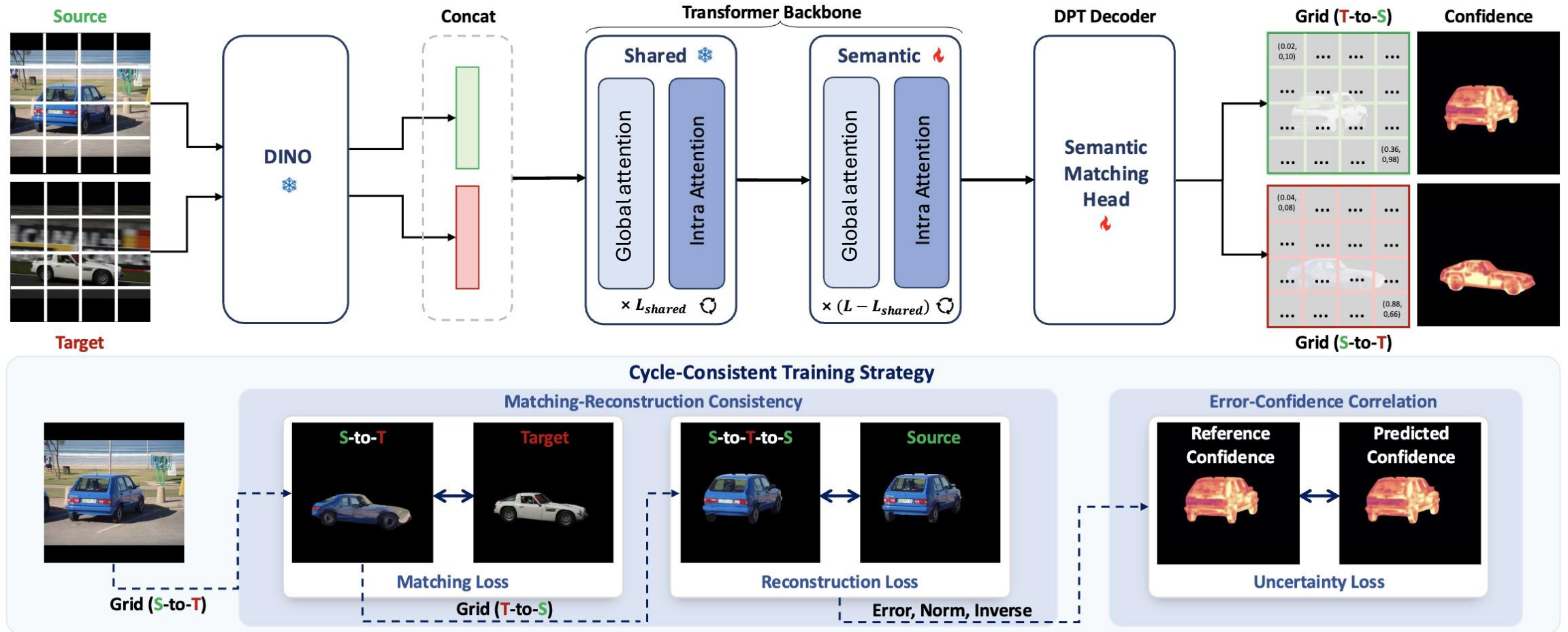


Adapting VGGT for Semantic Matching



- **Feature Refinement:** VGGT-adapted backbone
- **Correspondence Prediction:** Semantic matching head outputs pixel-wise grid correspondences

Adapting VGGT for Semantic Matching



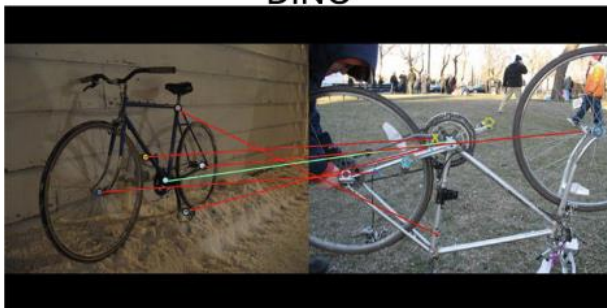
- **Cycle-Consistent Training:** Round-trip mapping reconstructs the original image
- **Error-Confidence Correlation:** Reconstruction error aligns with predicted confidence for reliable predictions

Qualitative Results: Sparse point matching

SD



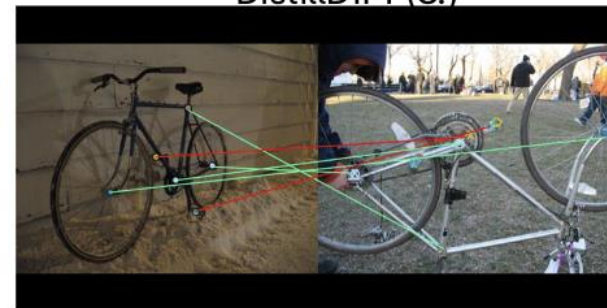
DINO



DINO+SD



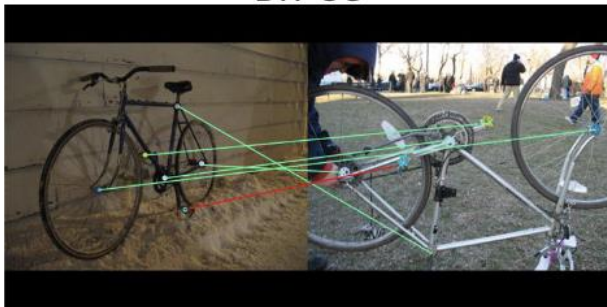
DistillDIFT (S.)



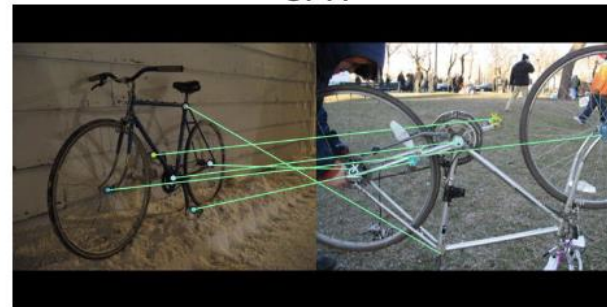
Geo-SC (S.)



DIY-SC



SPH



Ours



Qualitative Results: Dense semantic matching

Texture swap

Image 1



Image 2



Qualitative Results: Dense semantic matching

Texture swap

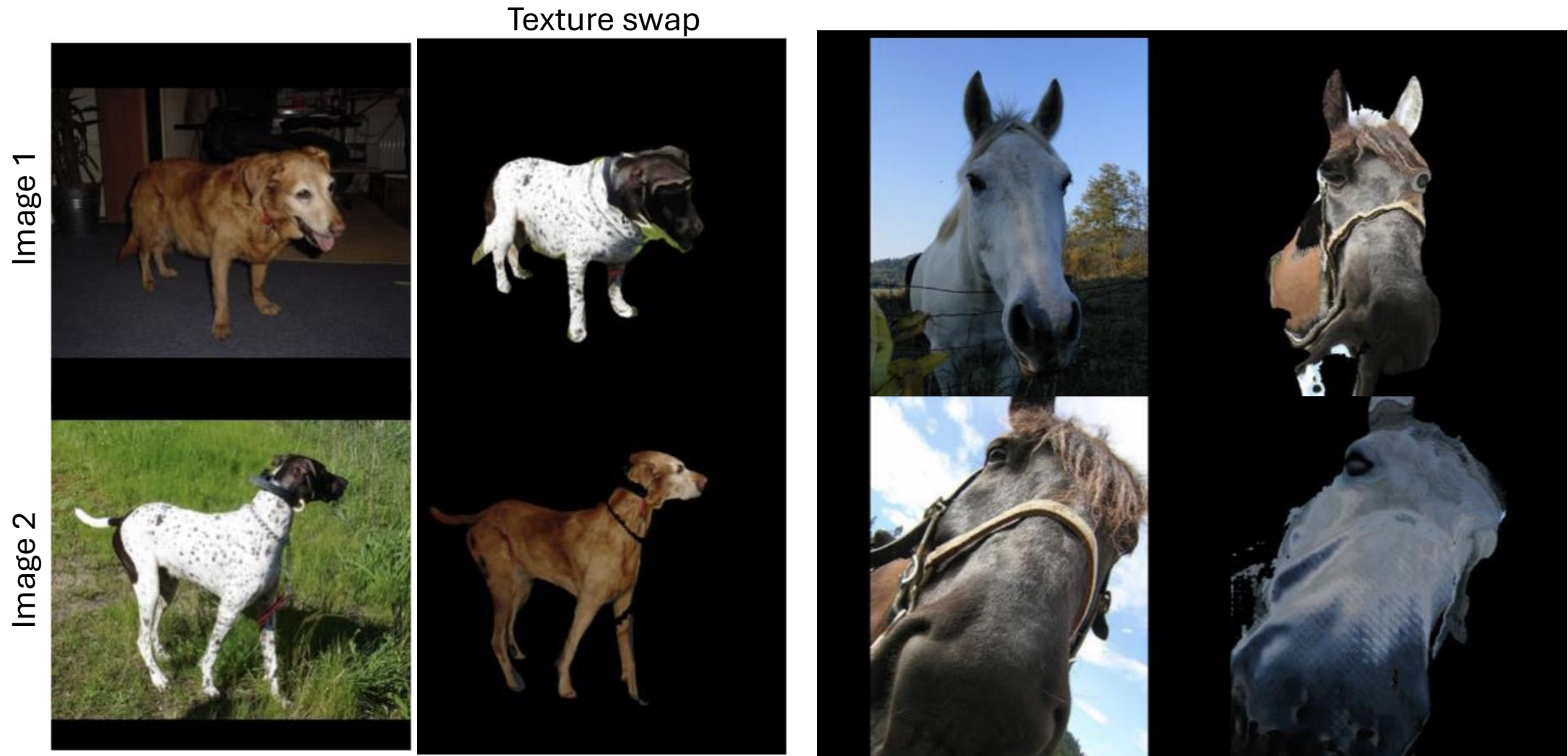
Image 1



Image 2



Qualitative Results: Dense semantic matching



Qualitative Results: Dense semantic matching

SD

DINO

DINO+SD

DistillDIFT (S.)

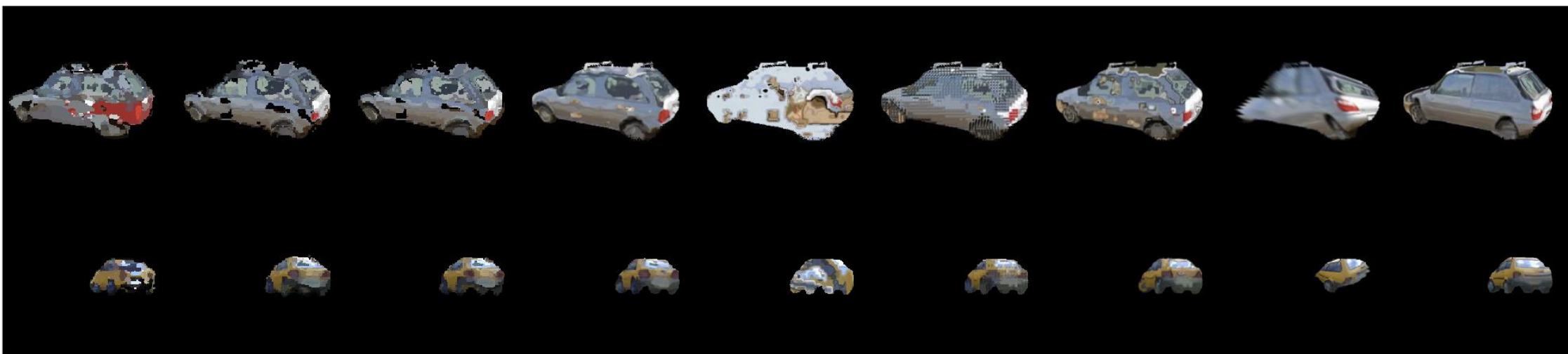
Geo-SC (S.)

DIY-SC

SPH

SpaceJAM

Ours



Quantitative Results: Point Matching

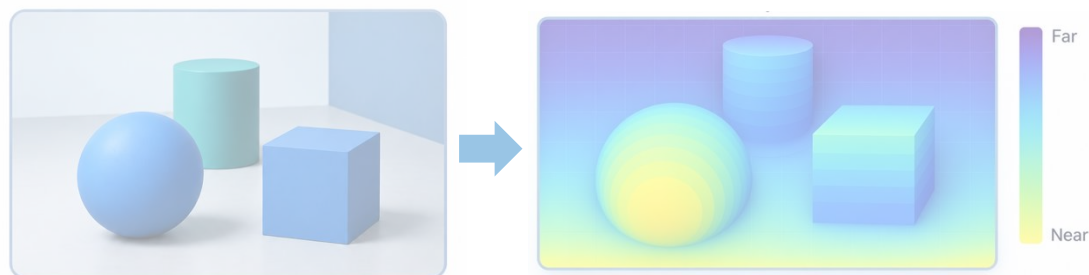
Table 1: Quantitative results on SPair-71k [30] and AP-10k [57]. 0.05* indicates additional aggregation on symmetric keypoints (*e.g.*, front/rear wheels).

Models		SPair-71k (PCK $\uparrow$)				AP-10k (PCK@0.1 $\uparrow$)			Synthetic $\downarrow$	DINO $\downarrow$	SD $\downarrow$	FID $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
		0.1	0.05	0.01	0.05*	I.S	C.S.	C.F.							
SD + DINO	[59]	59.9	44.7	7.9	40.3	62.9	59.3	48.3	0.20	0.35	0.28	28.73	17.88	0.59	0.249
DistillDIFT (U.S.)	[8]	60.8	45.4	8.0	44.7	65.8	64.2	56.1	0.16	0.21	0.18	32.04	18.08	0.62	0.242
Geo-SC (Flip)	[58]	65.4	49.1	9.9	51.0	68.7	64.6	52.7	0.14	0.23	0.16	29.45	18.02	0.63	0.245
DistillDIFT (S.)	[8]	65.3	49.9	8.9	49.8	66.9	64.7	58.0	0.15	0.17	0.16	<u>18.98</u>	18.95	0.63	0.221
Geo-SC (S.)	[58]	<u>82.9</u>	<u>72.6</u>	<u>21.6</u>	<u>73.5</u>	70.1	68.3	<u>58.4</u>	0.26	0.51	0.28	128.70	15.72	0.49	0.318
DIY-SC	[7]	71.6	54.2	10.1	53.8	<u>70.6</u>	<u>69.1</u>	57.8	0.11	0.15	0.14	25.46	17.22	0.56	0.268
SPH	[27]	64.4	48.2	8.4	49.3	65.4	63.1	51.0	0.10	0.17	0.15	23.89	16.51	0.53	0.289
ASIC	[9]	36.9	34.2	5.8	32.3	32.5	29.1	23.0	0.09	0.11	0.10	29.21	14.99	0.50	0.287
SpaceJAM	[2]	44.5	34.6	6.7	33.8	42.7	39.9	35.2	<u>0.08</u>	<u>0.10</u>	<u>0.07</u>	23.06	<u>21.13</u>	<u>0.71</u>	<u>0.178</u>
Ours		84.2	76.3	25.9	77.8	73.2	71.4	60.9	0.07	0.06	0.06	3.58	23.84	0.78	0.142

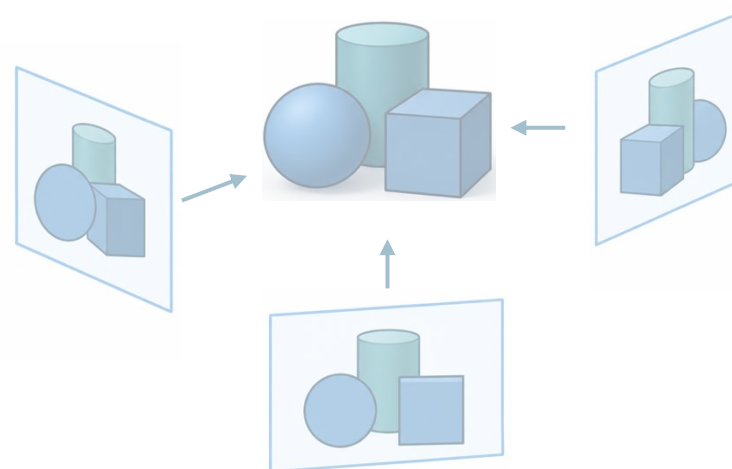
Ablation Study

Table 3: Ablation of the proposed training recipe. We evaluate the contribution of each component step-by-step.

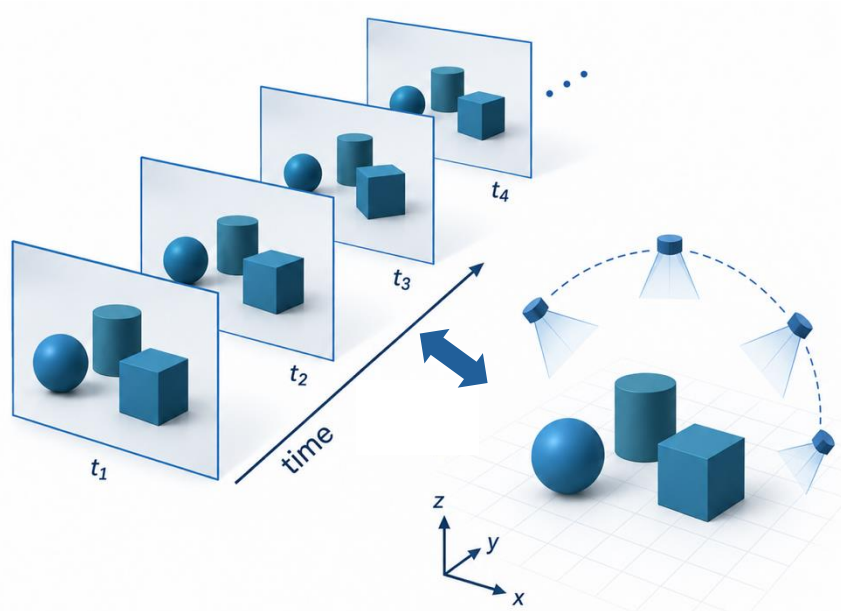
Settings	SPair-71k (PCK@0.1) $\uparrow$	Synthetic Dense $\downarrow$
DUSt3R	58.7	0.15
VGGT (Random Init.)	69.2	0.13
No Tuning	9.0	0.49
+ Sparse Supervision	79.4	0.18
+ Synthetic Data	79.3	0.11
+ Smoothness Loss	79.1	0.10
+ Reconstruction Loss	80.5	0.10
+ Matching Loss (VGG)	81.8	0.09
+ Matching Loss (DINO)	84.2	0.07



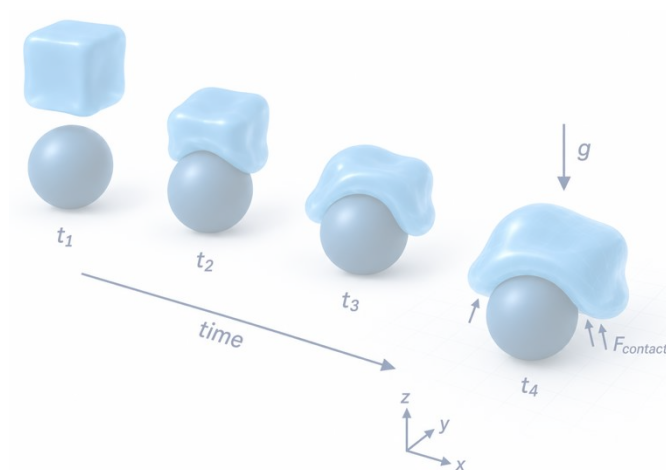
Depth Estimation



3D Reconstruction



3D in Video Generative Models



3D Simulation

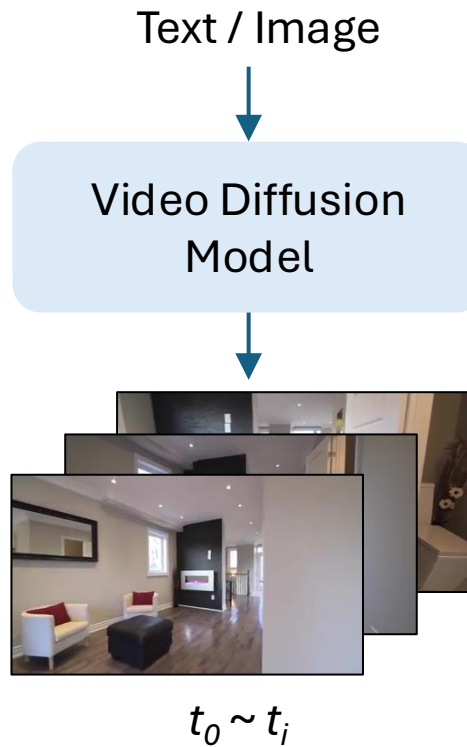
Video World Models



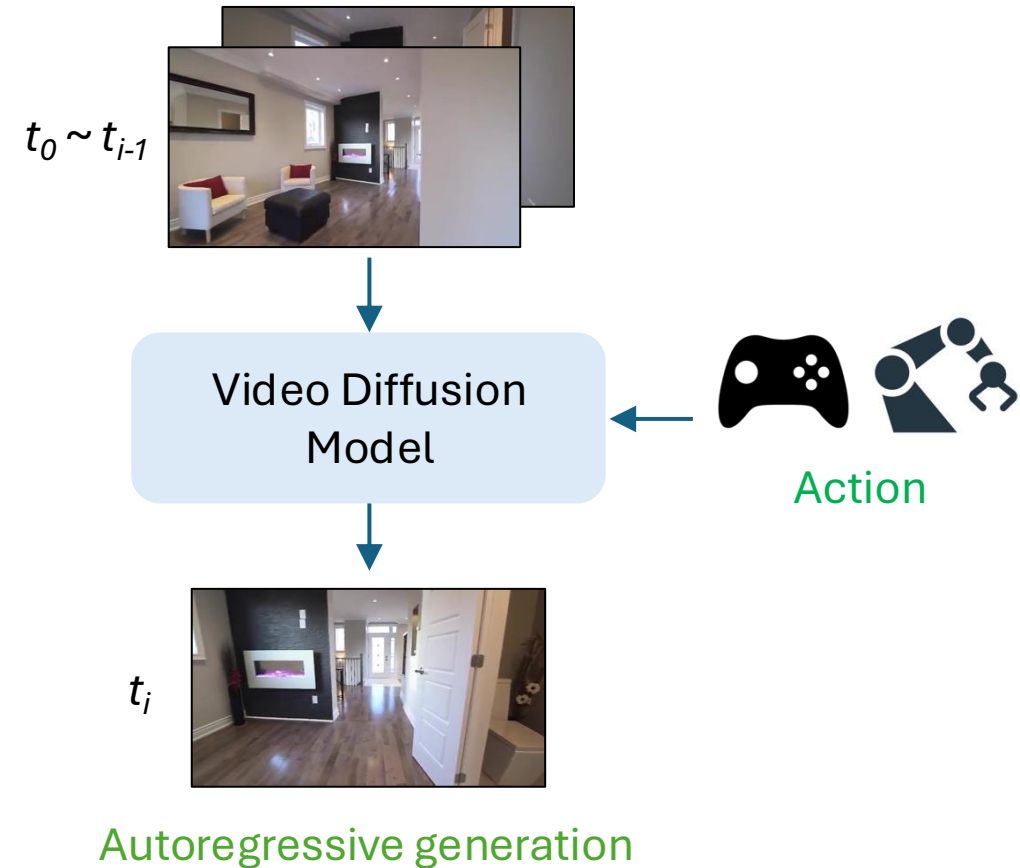
Oasis, <https://oasis-model.github.io/>

Video World Models

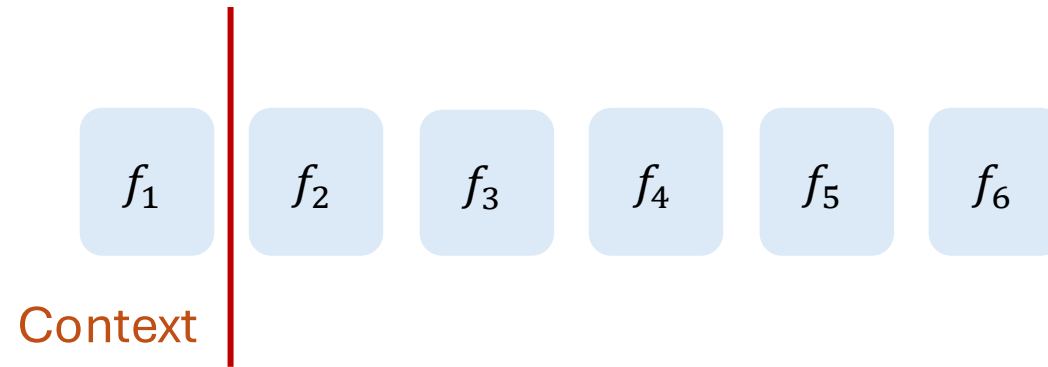
Vanilla Video Diffusion Models



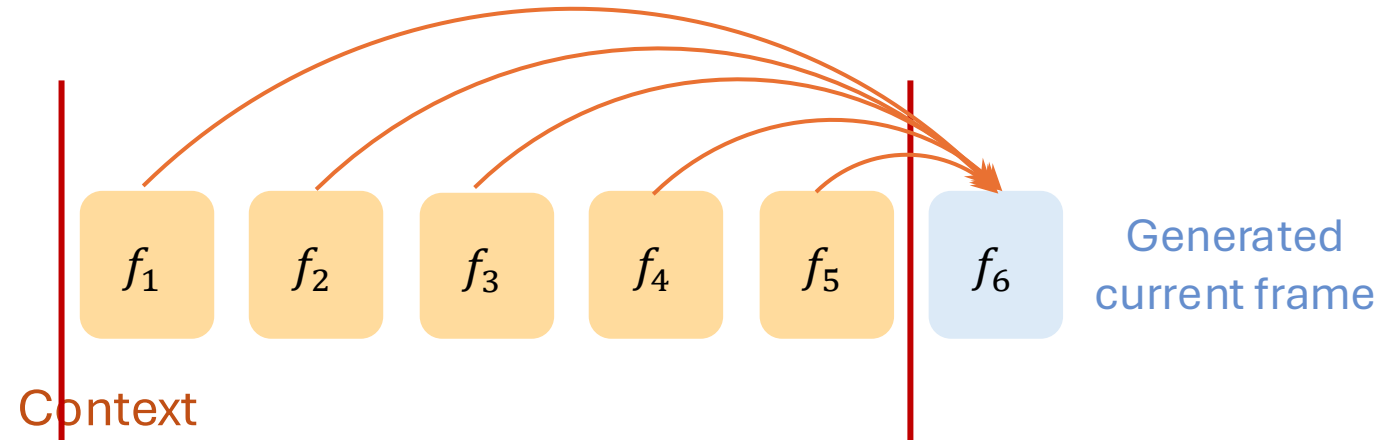
Video World Models



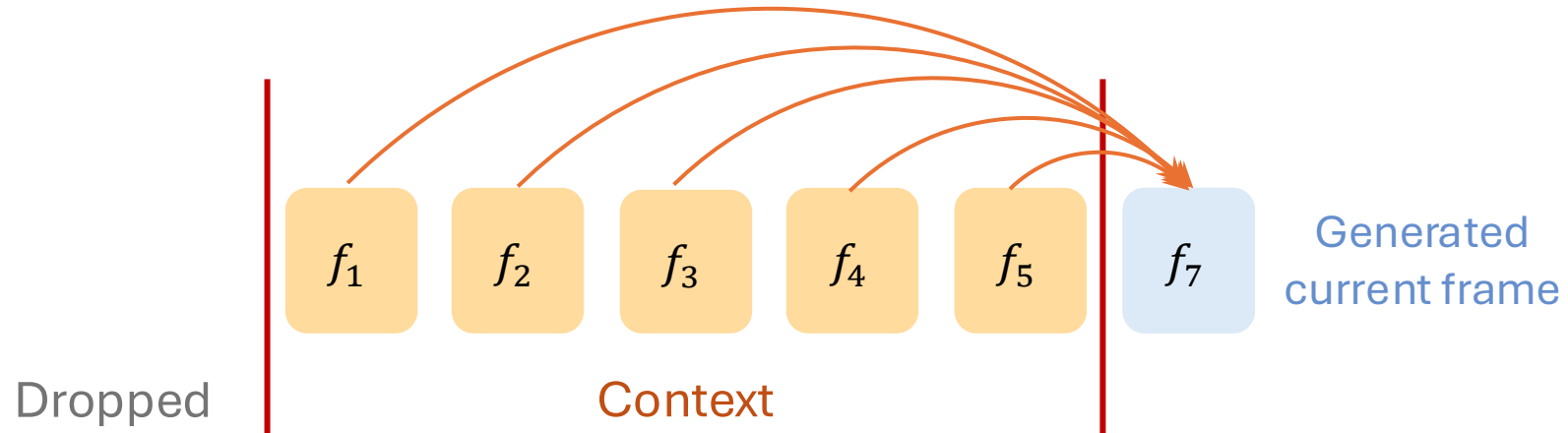
Autoregressive Generation



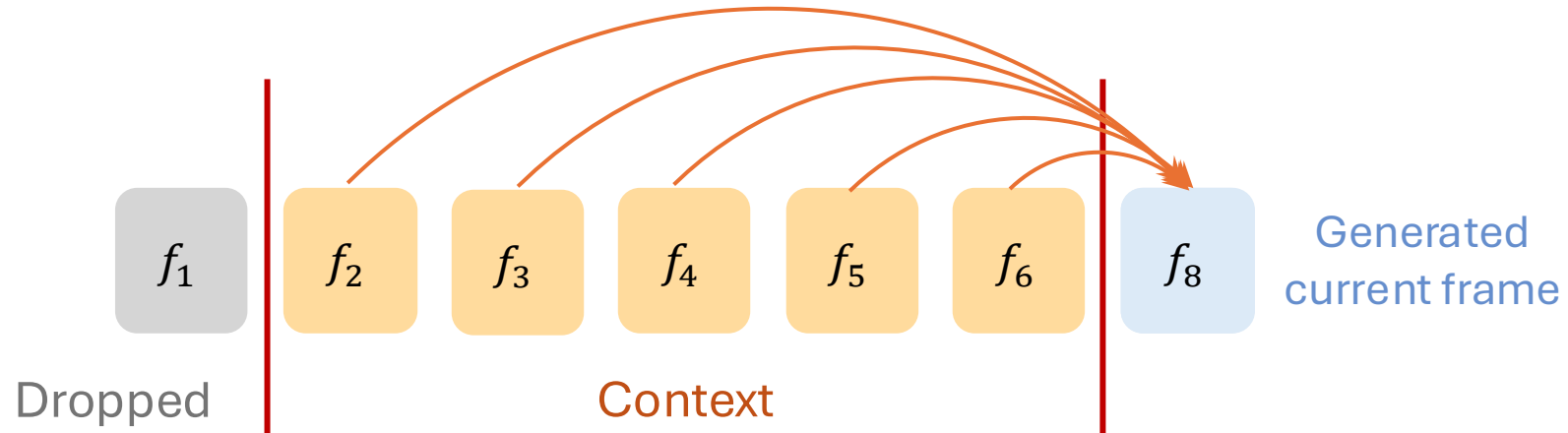
Autoregressive Generation



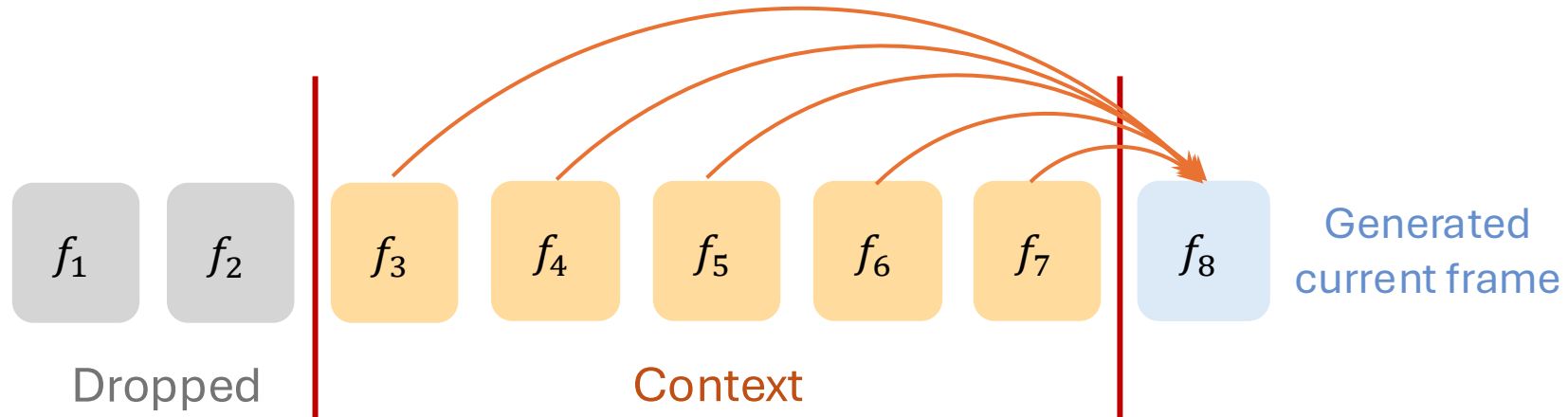
Autoregressive Generation



Autoregressive Generation



Autoregressive Generation

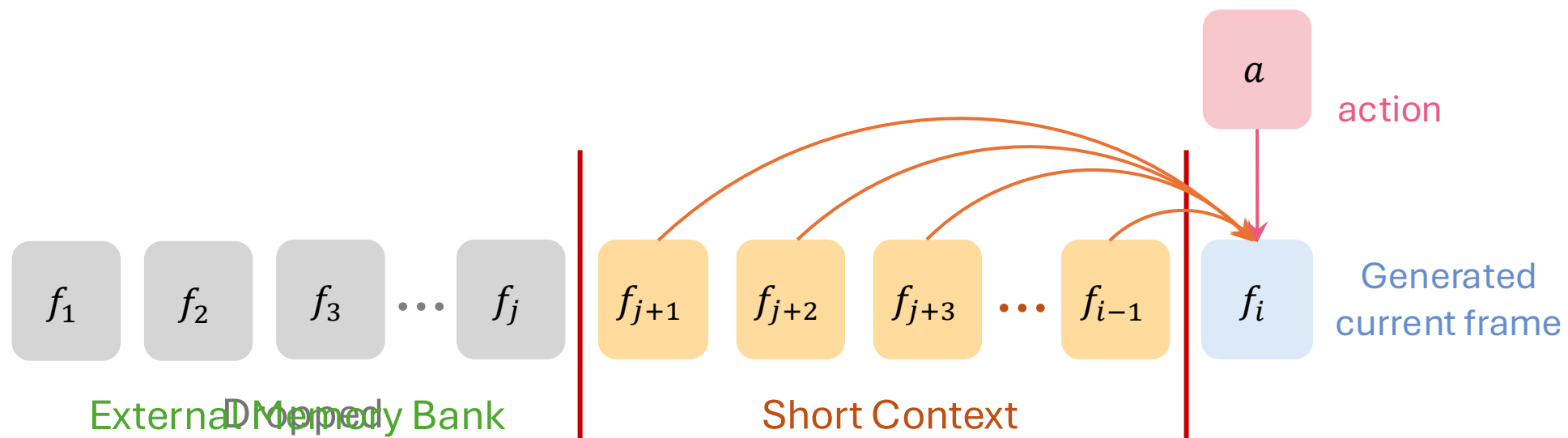


Inconsistency from Limited Context Window



Oasis, <https://oasis-model.github.io/>

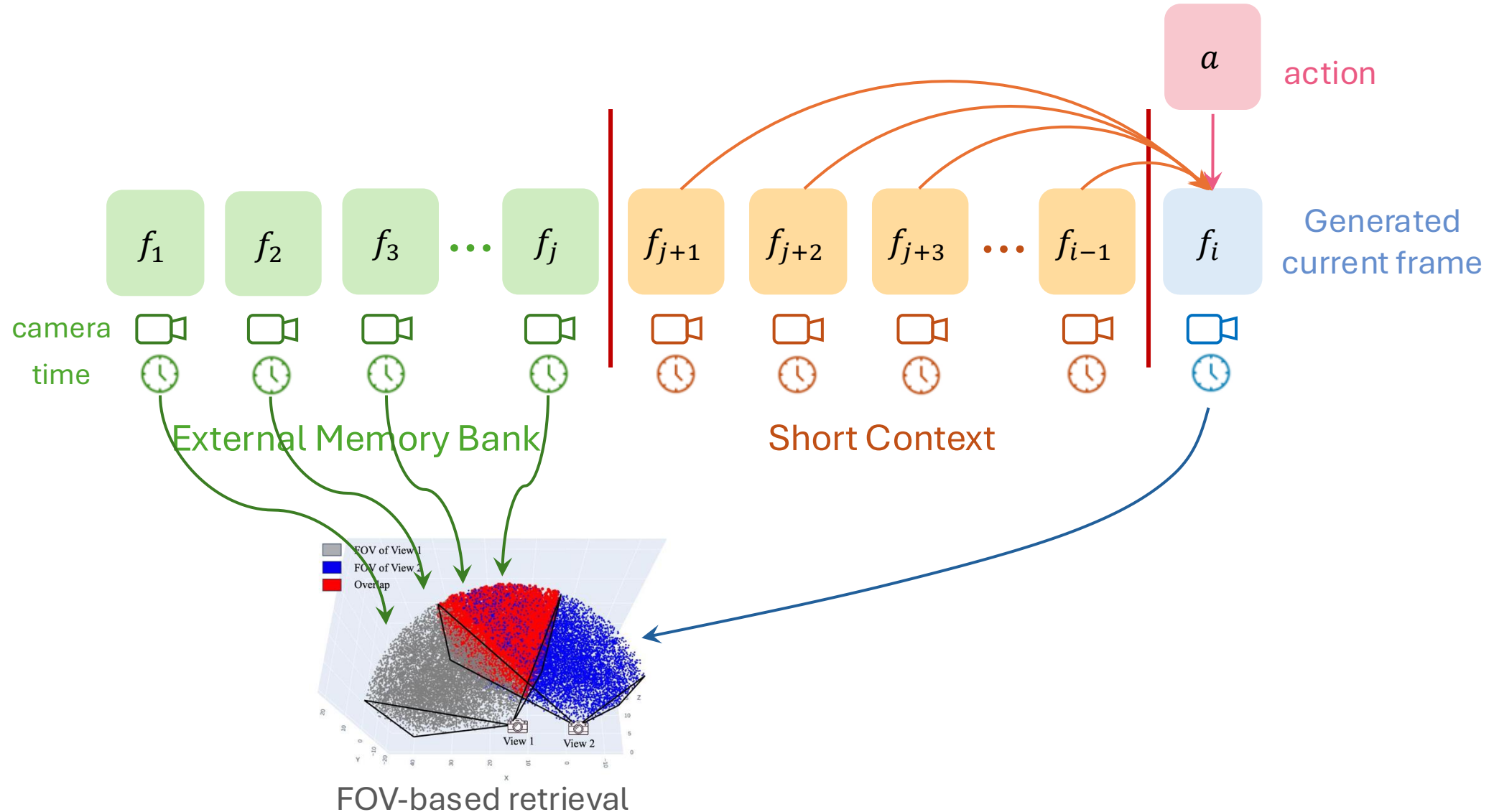
Motivation: Design External Memory



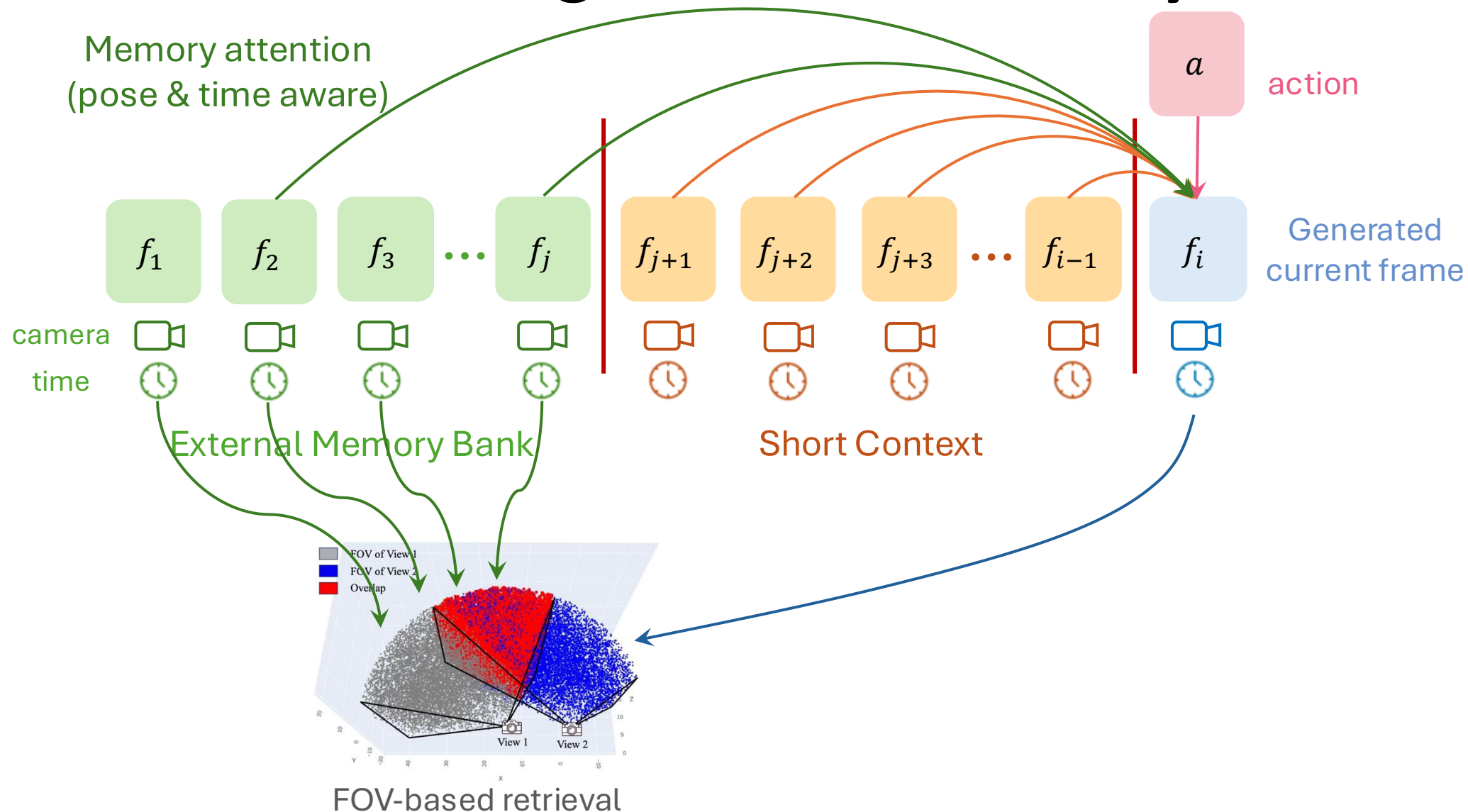
Which Memory Frame to Retrieve?

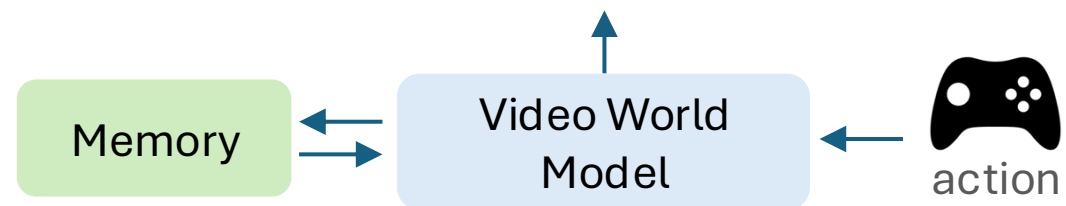


WorldMem: Design External Memory



WorldMem: Design External Memory

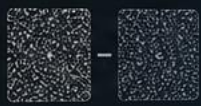




Hand2World

Text Prompt
"a person opens
a box on a desk"

Diffusion
Process



Generated
Video



VIDEO DIFFUSION MODEL

FRAME t

... FRAME $t+k$

High-Dimensional
Latent Space

Denoising
Step by Step

Temporal
Consistency

Realistic
Video Output

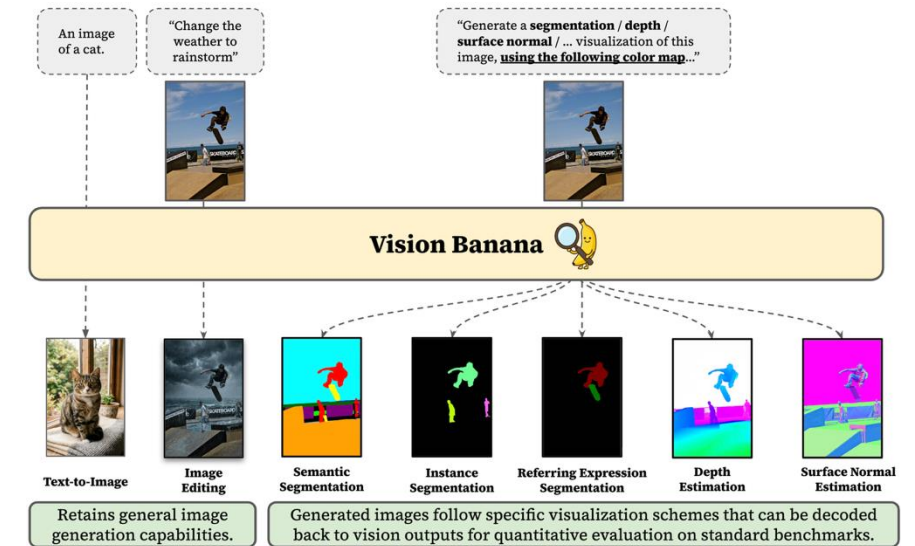
Rethinking Video Diffusion Models ...



WAN^[1] finetuned on Arctic dataset^[2]

The fact that they can generate implies that they *understand* the

- **structure**
- **motion**
- **correspondence**



Vision Banana^[3]

[1] Team Wan, Wan: Open and Advanced Large-Scale Video Generative Models, 2025

[2] Zicong Fan et al, ARCTIC: A Dataset for Dexterous Bimanual Hand-Object Manipulation, CVPR 2023

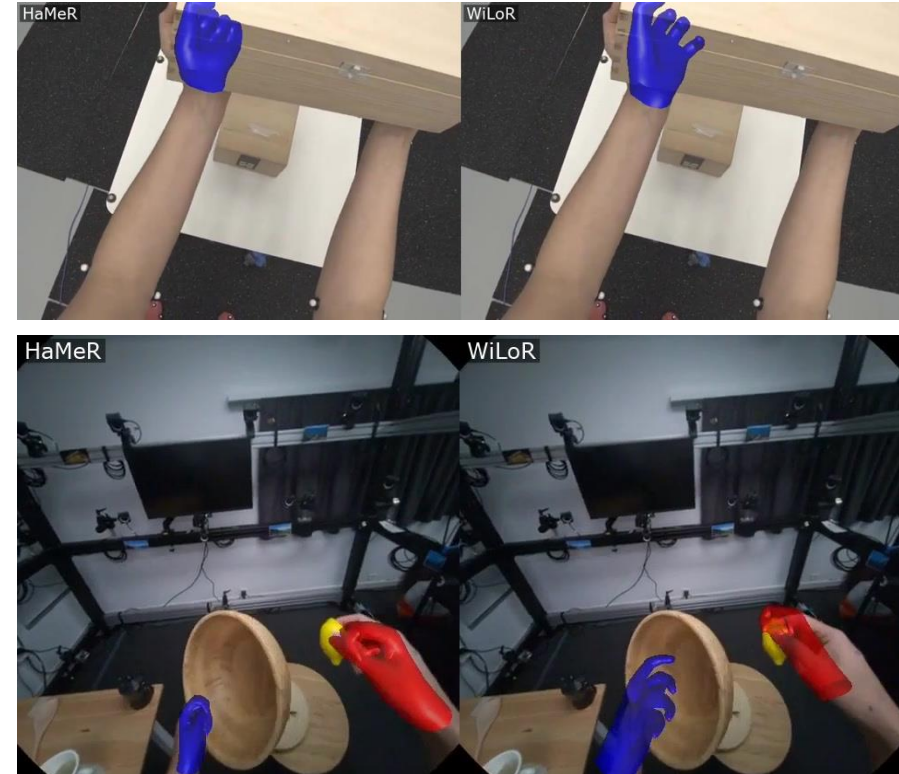
[3] Valentin Gabeur, Shangbang Long, Songyou Peng et al, Image Generators are Generalist Vision Learners, arXiv preprint, 2026

On the other hand ...



Ego4D^[1]

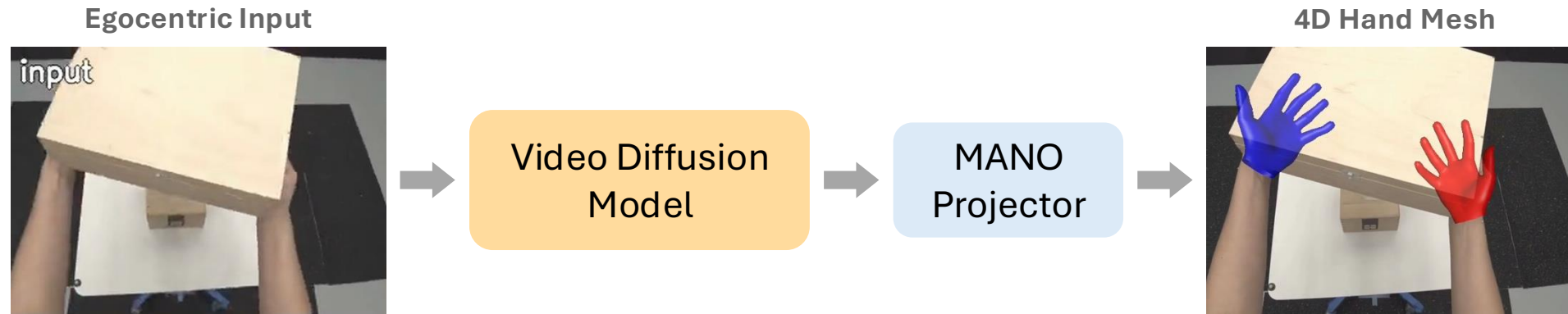
Egocentric vision is increasingly important



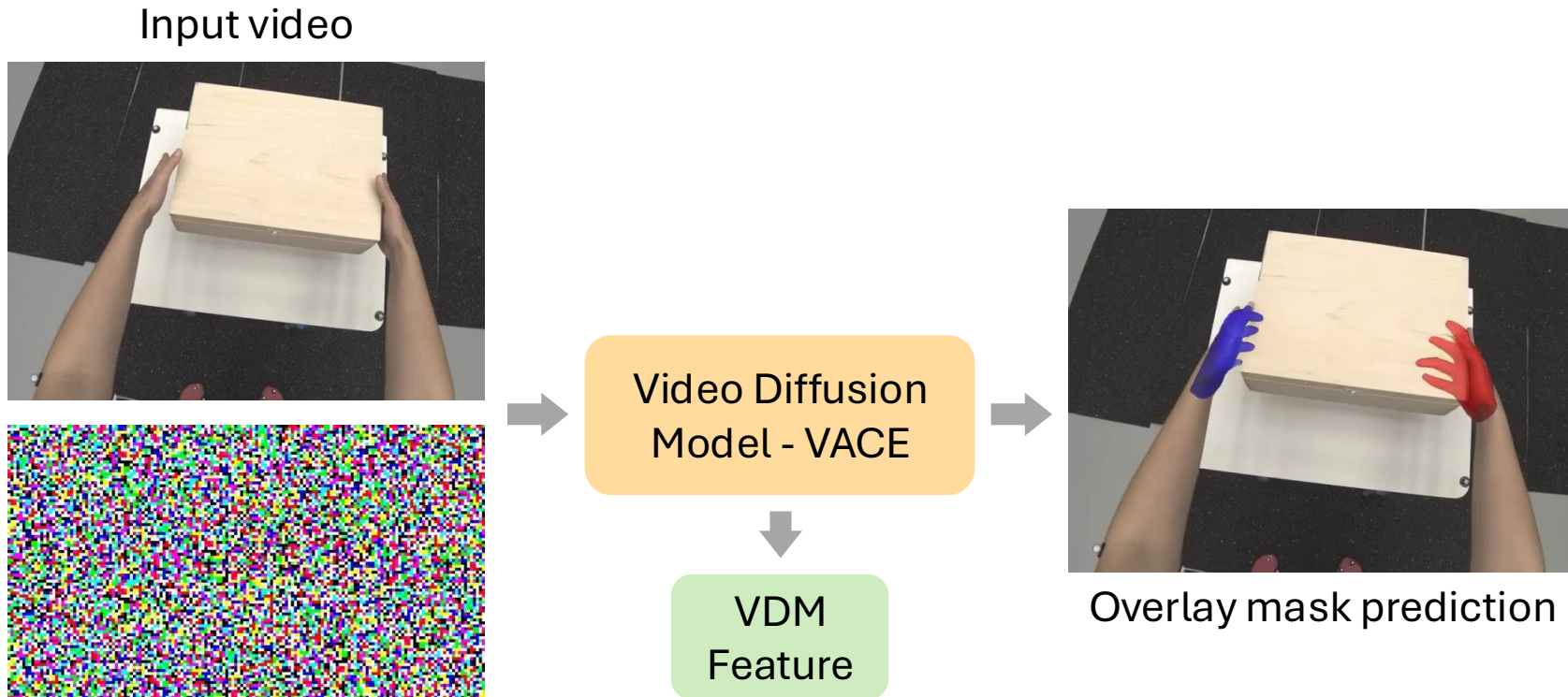
Yet understanding hand motion is still challenging ...

[1] Kristen Grauman et al, Ego4D: Around the World in 3,000 Hours of Egocentric Video, CVPR 2022

Leveraging VDMs for 3D Hand Motion Capture?

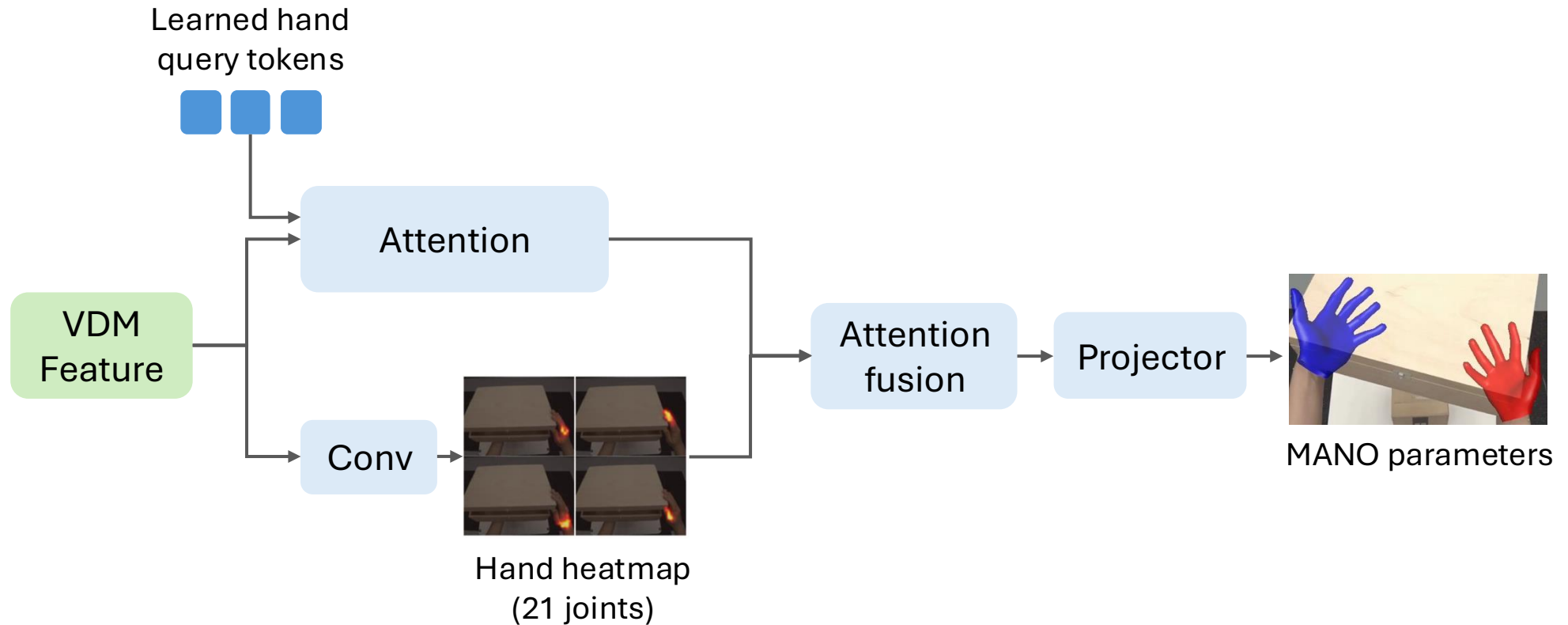


ViDiHand Stage 1: V2V overlay finetuning



Steer the video diffusion representation to focus more on hands

ViDiHand Stage 2: MANO Decoder Training



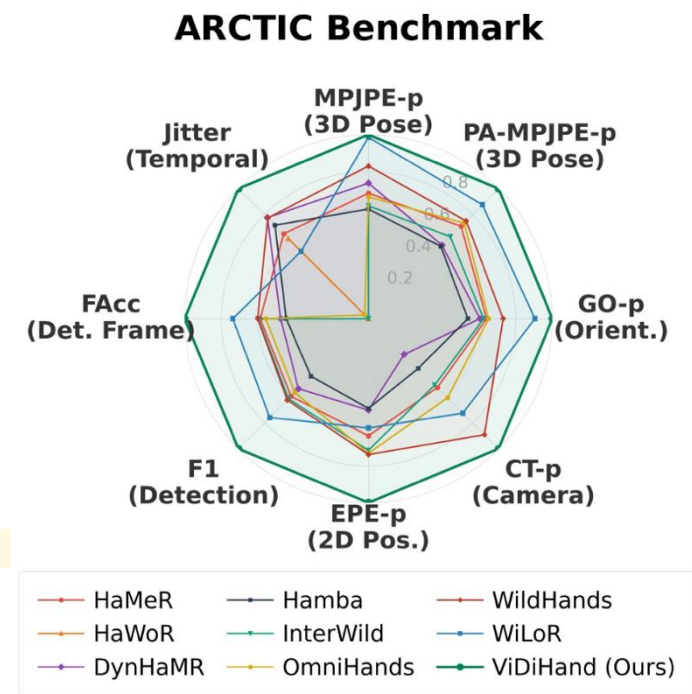
Quantitative Results: ARCTIC Benchmark

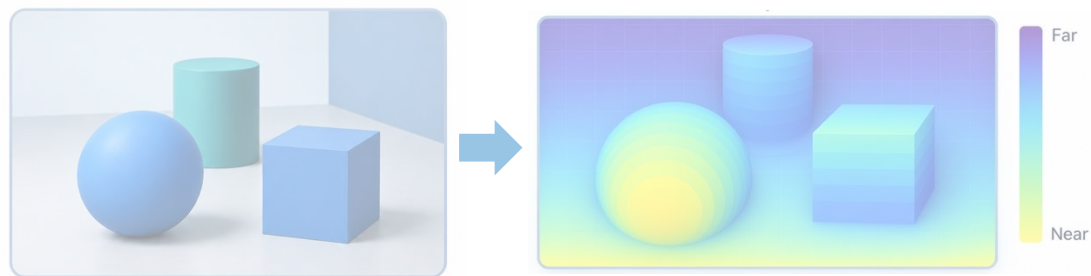
Training: ARCTIC, HOT3D

Testing:

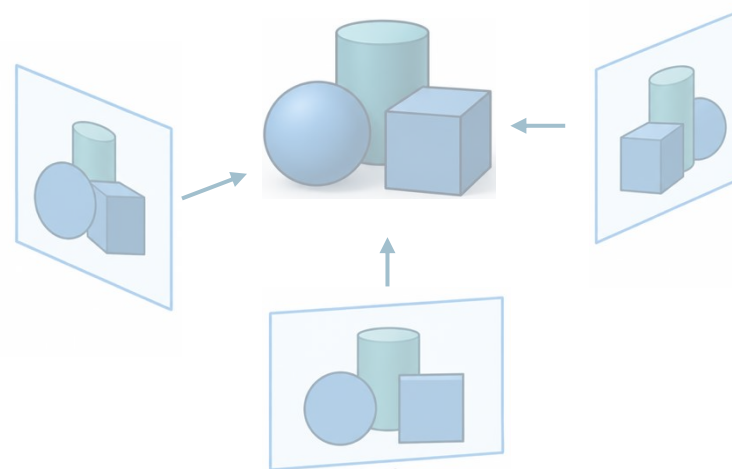
	Method	Detection			3D Pose		Orient. & Position			Temporal
		FAcc	Recall	F1	MPJPE-p	PA-MPJPE-p	EPE-p	GO-p	CT-p	Jitter
ARCTIC	InterWild [16]	0.878	0.943	0.959	30.817	15.952	53.888	25.386	0.097	46.577
	HaMeR [18]	0.875	0.943	0.957	29.197	14.596	65.289	24.907	0.095	18.279
	Hamba [3]	0.833	0.912	0.941	31.233	17.168	87.047	27.822	0.110	15.357
	WildHands [20]	0.879	0.946	0.960	25.704	13.941	50.517	22.320	0.058	12.972
	OmniHands [14]	0.866	0.949	0.954	29.674	14.203	51.505	24.580	0.087	45.312
	WiLoR [19]	0.919	0.951	0.974	22.012	11.873	71.527	17.358	0.075	24.091
	Dyn-HaMR [32]	0.842	0.918	0.951	27.904	17.017	85.723	25.951	0.121	12.840
	HaWoR [34]	0.700	0.817	0.895	45.357	26.375	158.062	43.325	0.149	19.789
	ViDiHand (Ours)	0.997	0.999	0.999	21.668	9.821	12.407	14.642	0.047	3.183

- Frame Accuracy 0.997 - **27x** lower error rate than SOTA (WiLoR)
- EPE-p 12.4 px - **4.07x** lower than SOTA (WildHands)
- Jitter 3.18 mm - **3.93x** lower than SOTA (Dyn-HaMR)

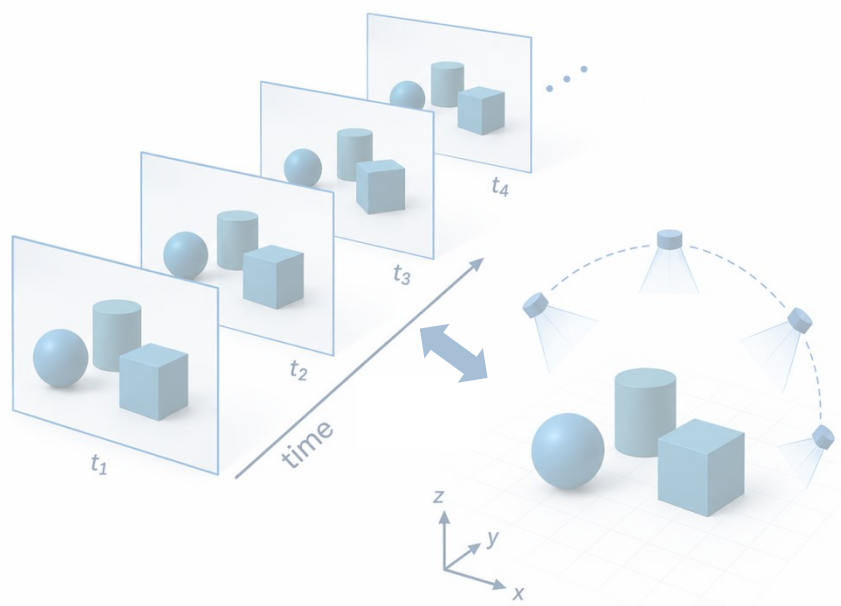




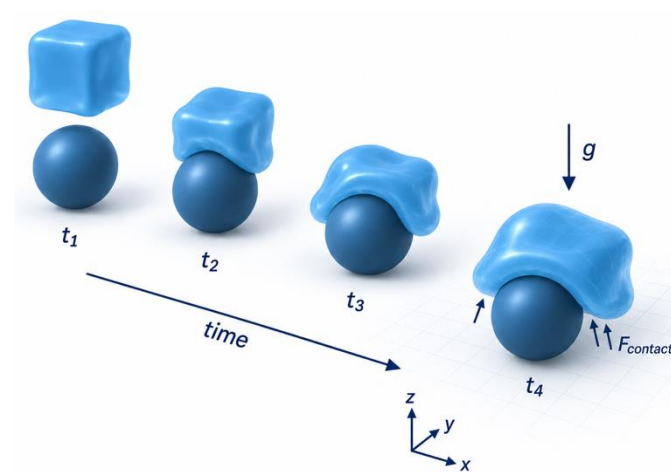
Depth Estimation



3D Reconstruction



3D in Video Generative Models



3D Simulation

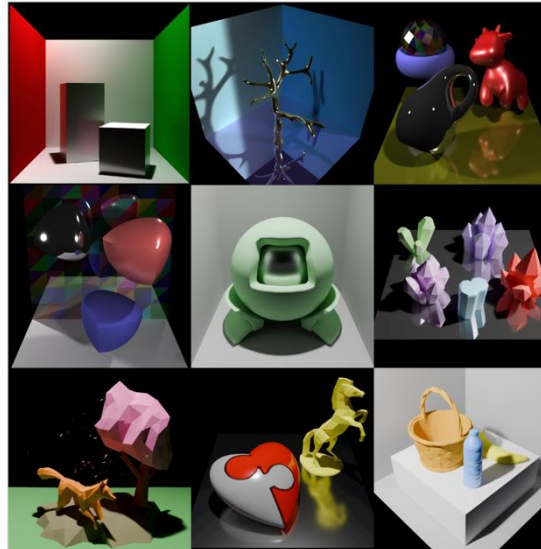
Background

- **Trend:** Transformers are taking over visual effects that traditionally rely on hand-crafted graphics pipelines.



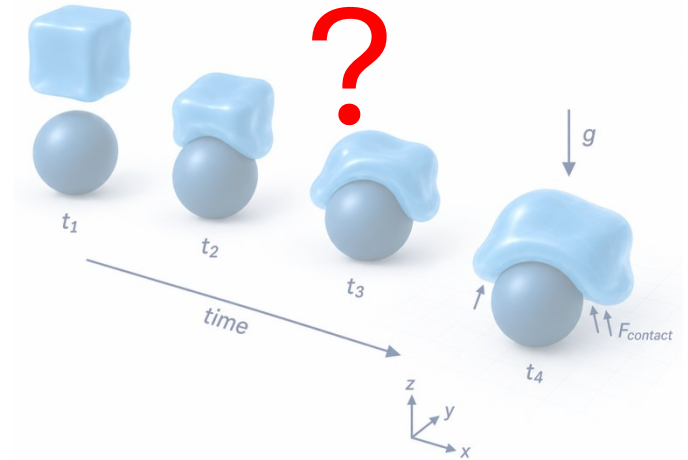
3D Consistency

Seedance 2.5, Bytedance



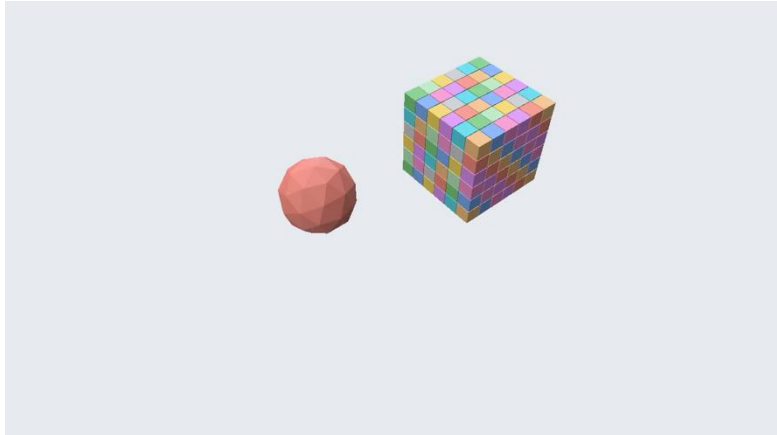
Rendering

Chong et al, RenderFormer,
SIGGRAPH2025



Simulation?

Related Works



RigidFormer



PhysiFormer



ClothTransformer

Zhiyang Dou et al, RigidFormer: Learning Rigid Dynamics using Transformers, arXiv 2026

Yiming Chen et al, PhysiFormer: Learning to Simulate Mechanics in World Space, arXiv 2026

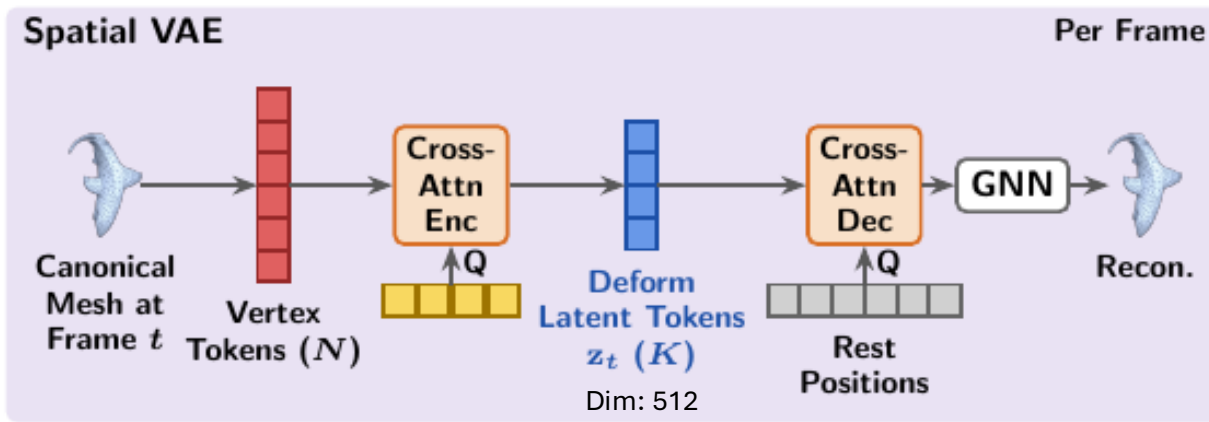
Yu Zhang et al, ClothTransformer: Unified Latent-Space Transformers for Scalable Cloth Simulation, arXiv 2026

Need of a More Compact Latent Space

N	tokens ($T \times N$)	PhysiFormer (DiT-L) GiB / ms	PhysiFormer-B (DiT-B) GiB / ms
500	32k	64.4 / 669	24.0 / 212
1,000	64k	122.1 / 1317	46.0 / 412
1,200	77k	OOM	—
2,000	128k	OOM	89.9 / 879
2,500	160k	OOM	111.8 / 1147
2,750	176k	OOM	123 / 1289
3,000	192k	OOM	OOM
3,500	224k	OOM	OOM
5,000	320k	OOM	OOM

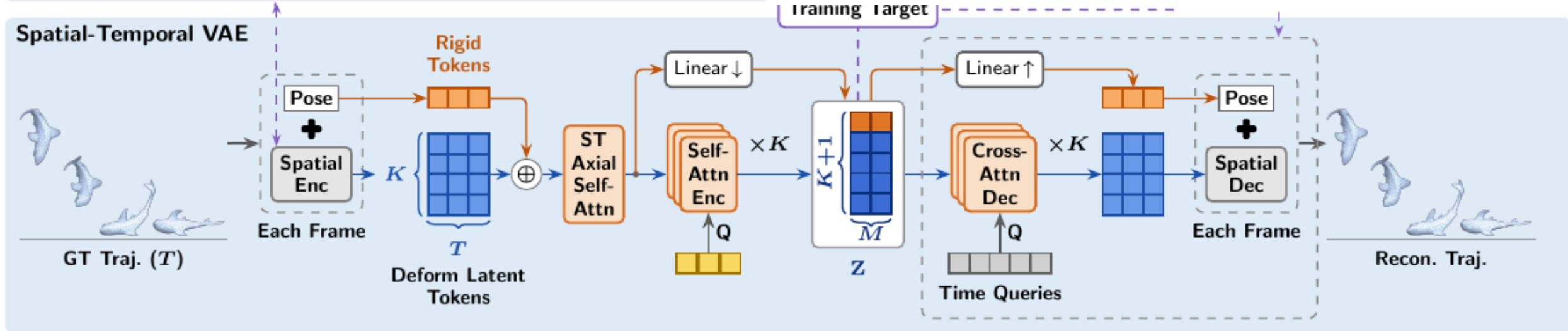
Table 10: PhysiFormer’s memory wall on a 140 GiB H200: peak training memory (GiB) and GPU time per training sample (ms) at batch 1, $T=64$, $b_f=16$, flash attention, as a function of vertex count N . OOM: out of memory at batch 1. For reference, our generator trains at ~ 4.4 GiB and ~ 27 ms per sample at any N (Table 3).

PhysLDM



- Transformer-based S-T VAE
- Factor coordinate into a global rigid transform acting on a canonical deformation

$$\mathbf{x}_t = \mathbf{R}_t (\mathbf{r} + \mathbf{d}_t) + \mathbf{c}_t$$

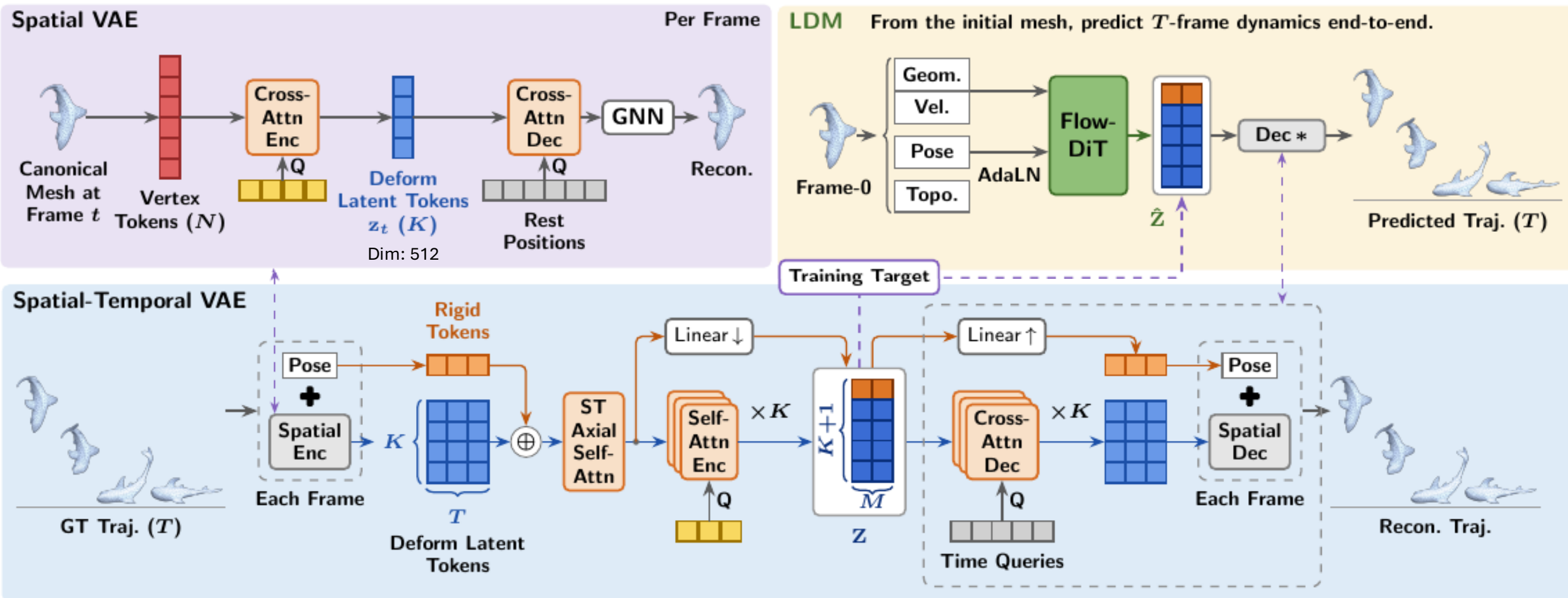


Spatial compression ratio: x 9.8 for 5000 vertices

Temporal compression ratio: x 8

Total compression ratio: x 78

PhysLDM



Spatial compression ratio: x 9.8 for 5000 vertices

Temporal compression ratio: x 8

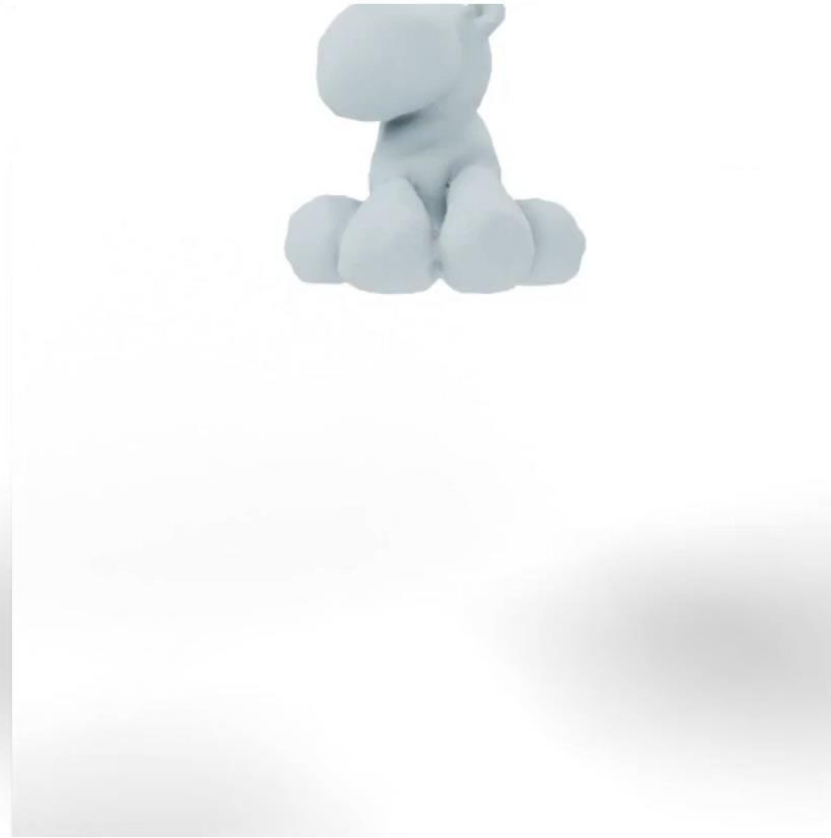
Total compression ratio: x 78

ST-VAE Reconstruction Results

unseen GSO scan · 6,067 tet nodes · 95× token compression · 2× slow



Ground truth



Autoencoder reconstruction

ST-VAE Reconstruction Results



Ground truth

Autoencoder reconstruction

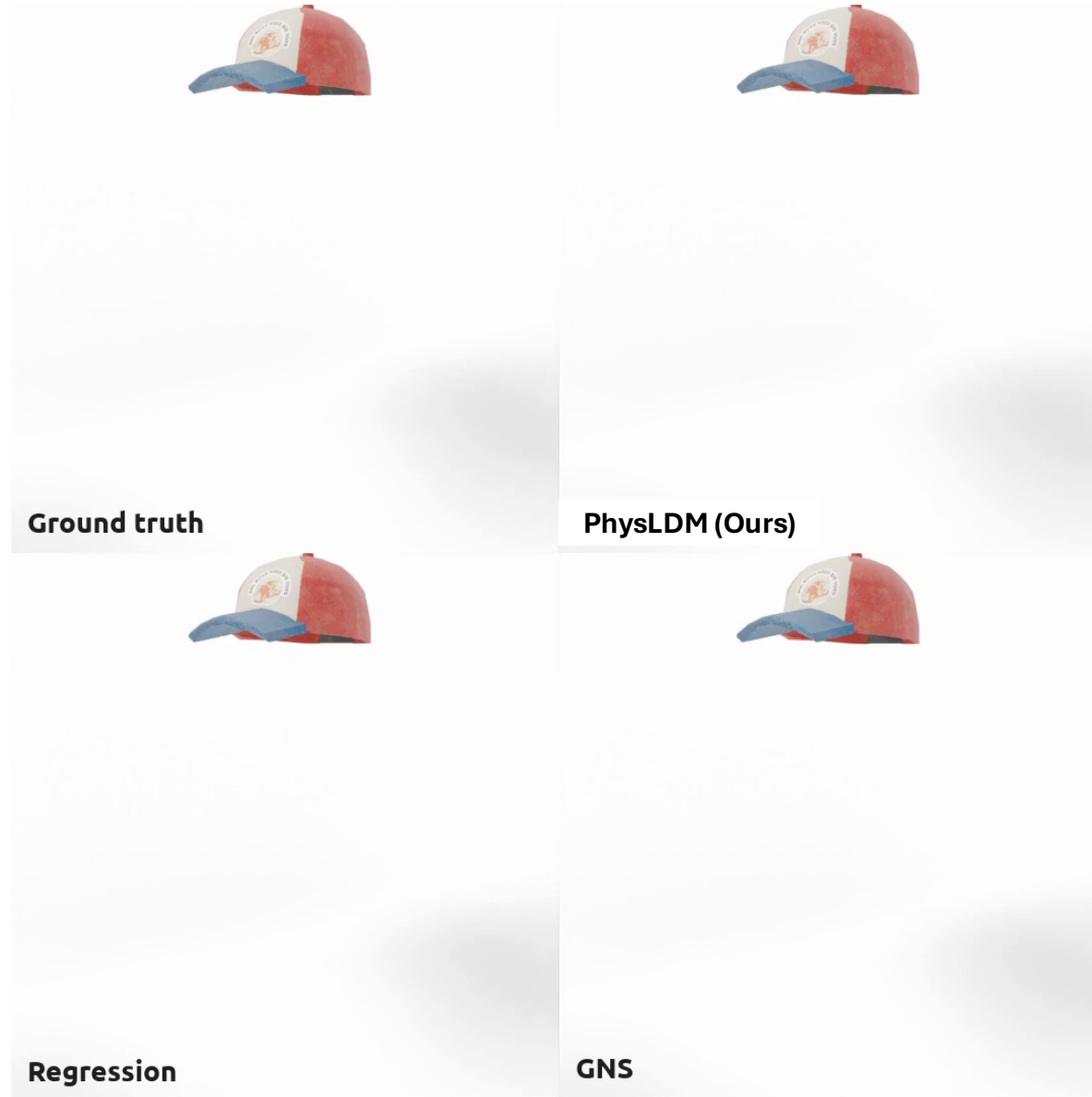
ST-VAE Reconstruction Results



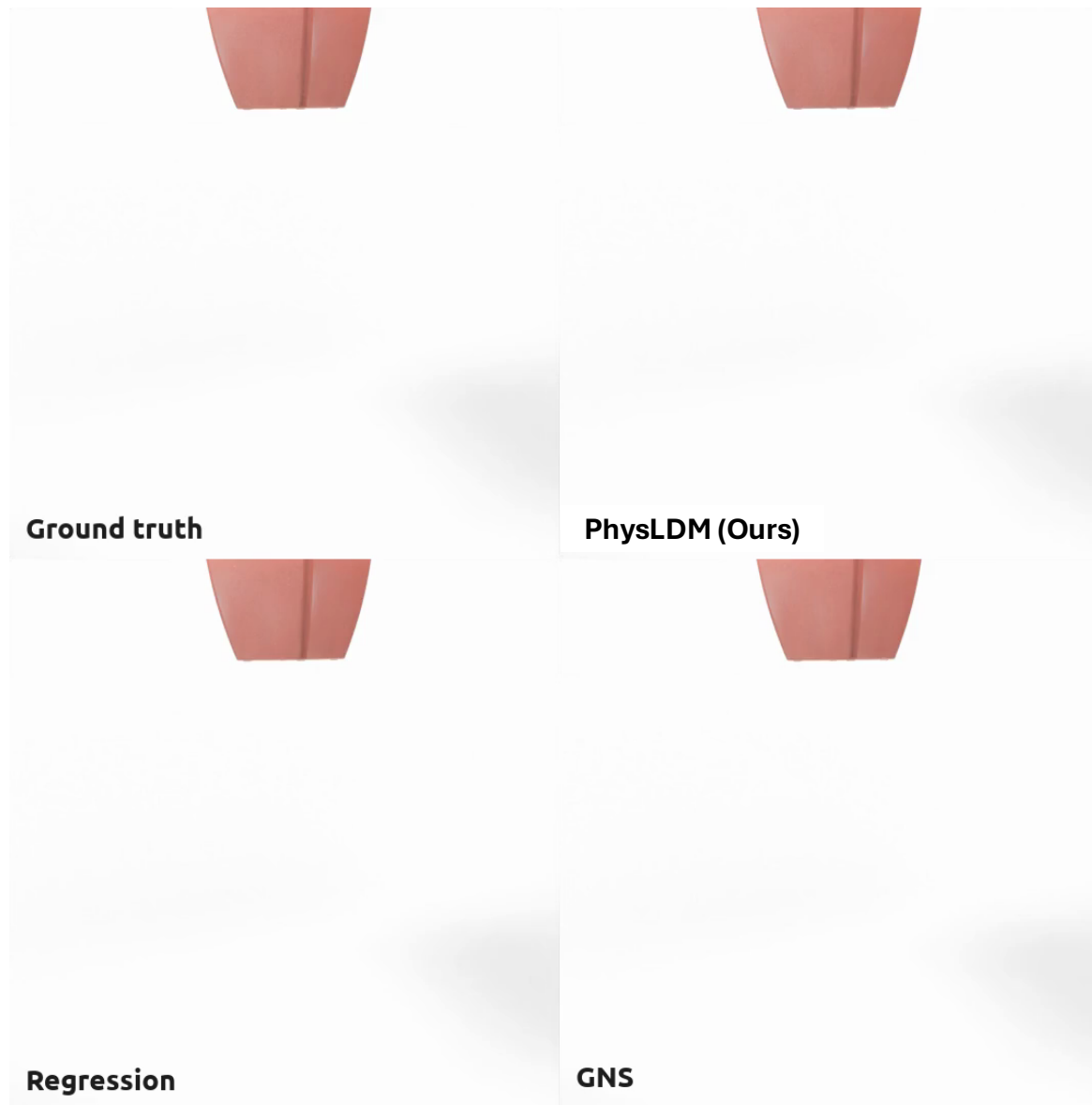
Ground truth

Autoencoder reconstruction

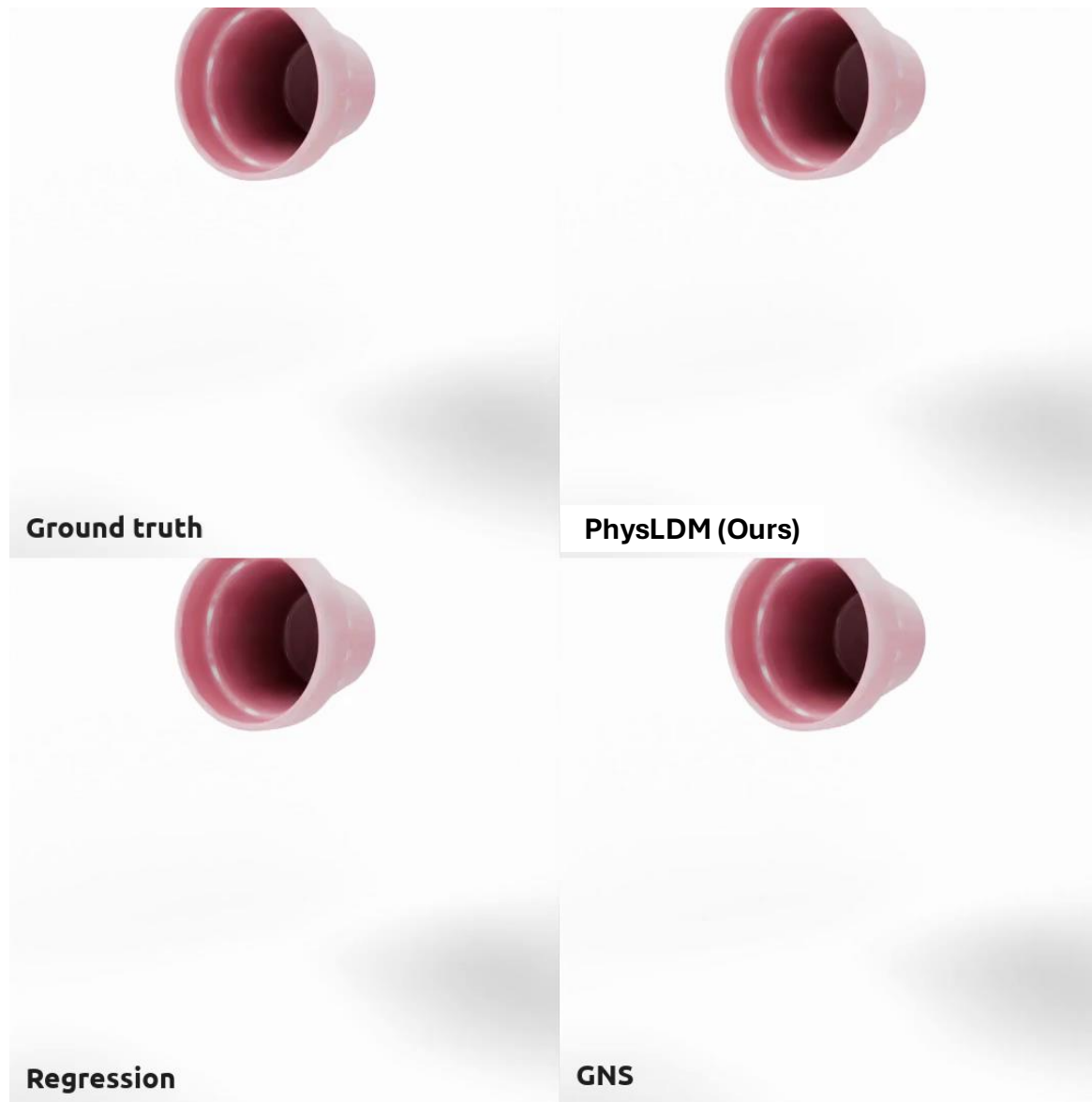
LDM Simulation - Objaverse held-out



LDM Simulation – OOD: GSO and toys4d

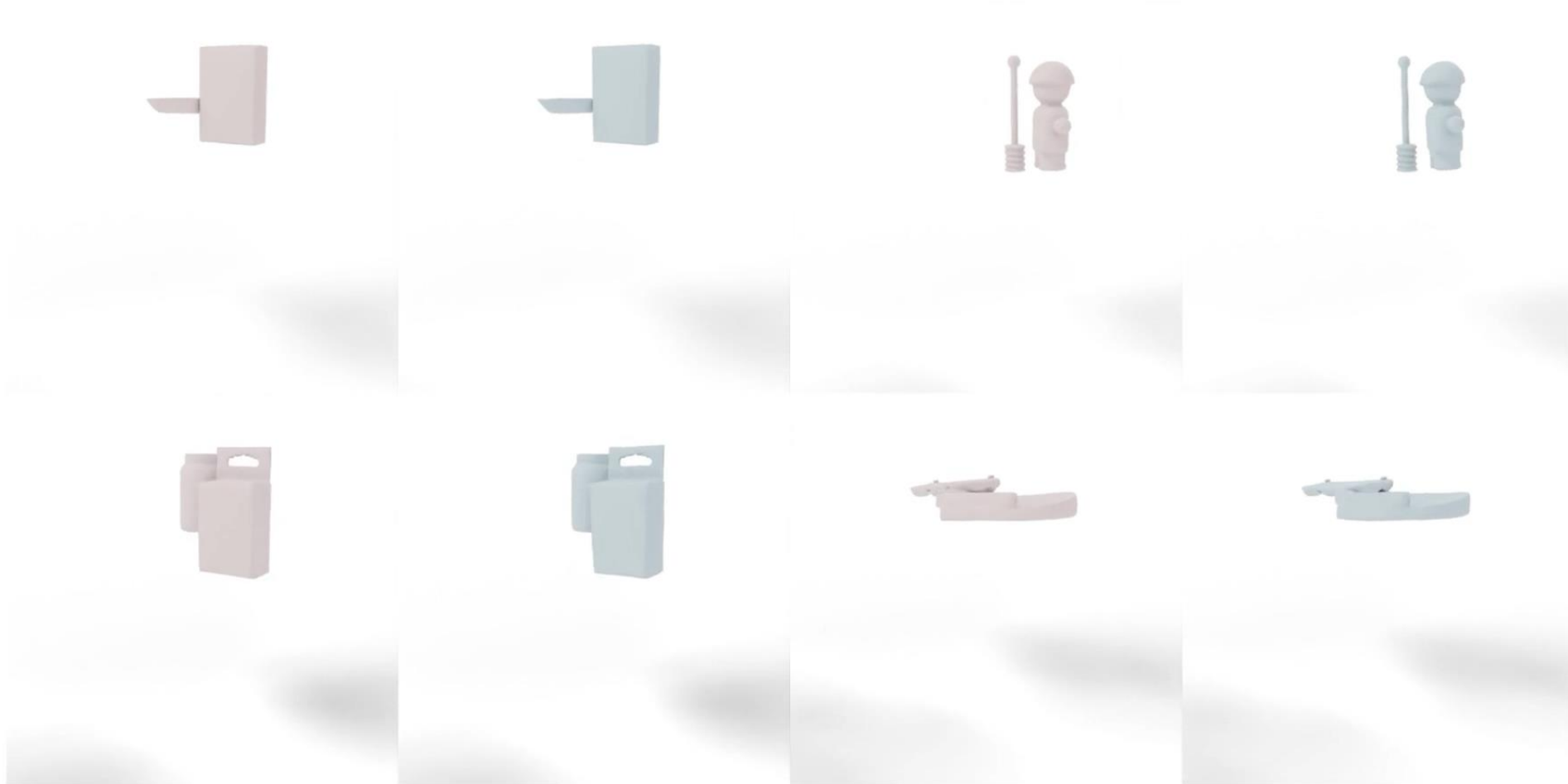


LDM Simulation – Challenging Cases



Extension to Multi-body

ST-VAE Reconstruction



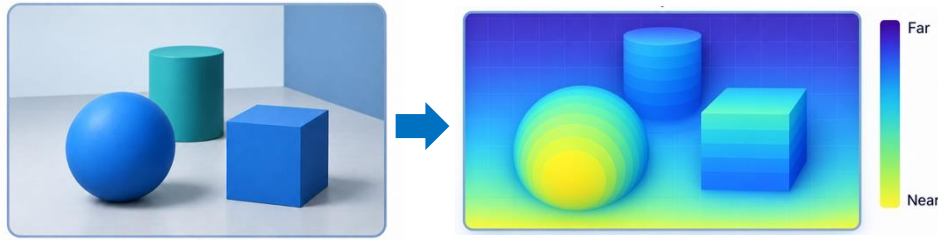
Rose = Ground truth
Blue = ST-VAE reconstruction

Solving Inverse Problem: Material Estimation

Material Estimation

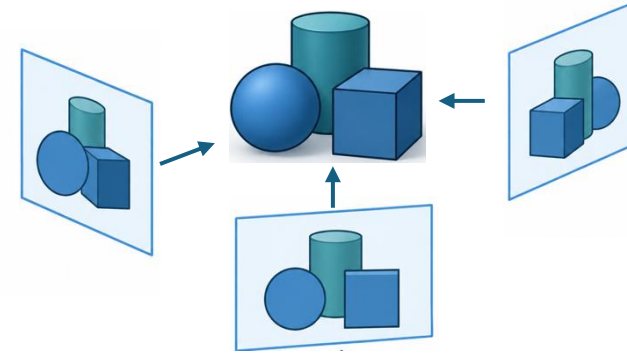
Recover an unknown stiffness from one observed trajectory.

Depth Estimation



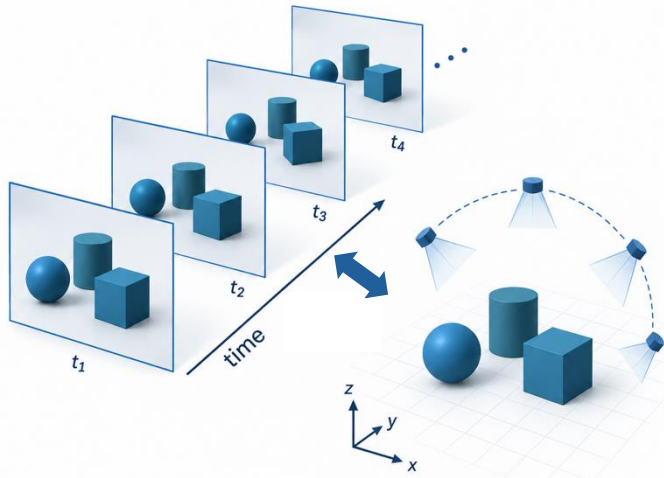
Exploit depth cue from **bokeh** effect in *image diffusion models*

3D Reconstruction



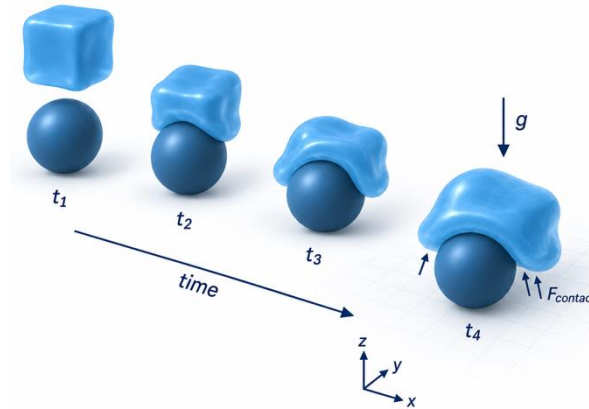
Build streaming reconstruction model via **causal** attention
Exploit VGGT for dense semantic matching

3D in Video Generative Model



Build 3D consistent video world model via **memory**
Exploit video diffusion models for 3D hand motion capture

3D Simulation



Build latent diffusion model for 3D simulation

Acknowledgement



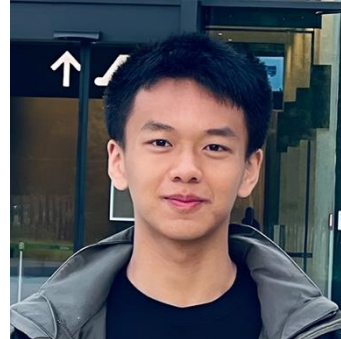
Armando Teles Fortes



Hangwei Zhang



Yushi Lan



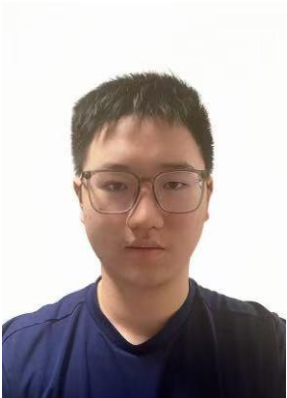
Yihang Luo



Songlin Yang



Zeqi Xiao



Yuxi Wang



Chengkai Jin



Yu Zhang



Tianyi Wei



Shangchen Zhou



Xudong Xu

Q & A