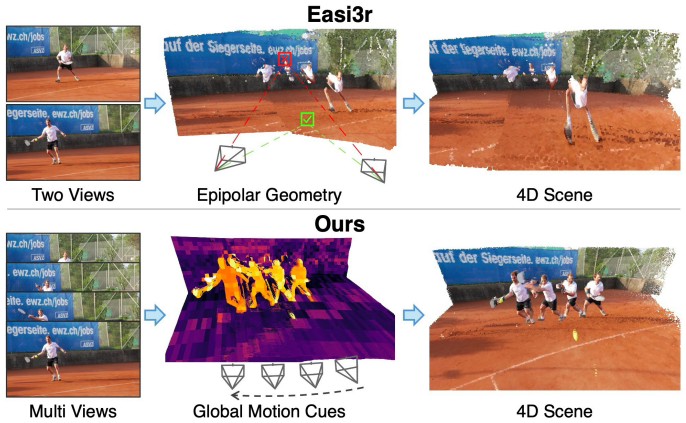


VGGT4D: Mining Motion Cues in Visual Geometry Transformers for 4D Scene Reconstruction

Yu Hu[1*], Chong Cheng[1,2*], Sicheng Yu[1*], Xiaoyang Guo[2], Hao Wang[1†]

The Hong Kong University of Science and Technology (Guangzhou)[1], Horizon Robotics[2]

CVPR 2026

Presenter: 송우석

- **site:** https://3dagentworld.github.io/vggt4d/
- **github:** https://github.com/3DAgentWorld/VGGT4D
- **paper:** https://arxiv.org/abs/2511.19971

RAISE LAB

Index

- Problem & Motivation
- Method
- Experiments
- Limitation

Why should we handle dynamics?

- 4D scene reconstruction with dynamic object

  → has been a challenging task

- moving object
  - degrade pose estimation & interfere with the background geometry modeling
  - their motion is often entangled with camera motion

    → leading to severe artifacts in 3D scene representations

⇒ **Therefore, how to model dynamics is crucial for robust 4D reconstruction**

What matters if we didn't handle dynamics?

- Traditional SfM, MVS methods
  - rely on multi-view rigidity & photometric constancy

  → Dynamic regions violate these assumptions, degrading correspondences and bundle adjustment, and often causing failure

- 3D foundation models (e.g. VGGT)
  - deliver fast, accurate 3D geometry and camera pose estimation

  → they are largely trained & inferred under static-scene assumptions & lack an explicit mechanism to disentangle moving objects

  ⇒ When dynamics dominate, this coupling of dynamics and statics leads to brittle reconstruction & pose drift

What's the problem of 4D Reconstruction?

- two limitations (existing methods)
  1. heavy iterative refinement

     → that incurs substantial runtime and memory overhead
  2. reliance on external modules (e.g. optical flow, depth, semantic segmentation)

     → which complicates integration and makes performance sensitive to module quality and domain shift
- recent methods
  - efficient feed-forward architectures

    → but, still require large-scale training and fine-tuning on high-quality dynamic datasets

  ⇒ costly to curate and scarce at scale

- Easi3R (preliminary method)
  - training-free extension of DUSt3R
  - segments dynamic masks by analyzing the spatial and temporal statistics of decoder attention
  - pairwise cross-attention architecture captsures only local feature interactions

    → has short temporal horizon, yielding masks that are inconsistent across frames with boundary errors at dynamic–static interfaces

    → cause depth drift and floating artifacts in the reconstructed point clouds

  ⇒ its core assumption that tokens violating epipolar geometry receive low attention **does not generalize to VGGT**

Related Works

- 3D Foundation Models

  - **DUST3R:** largescale pretraining for feed-forward 3D foundation models by predicting dense pointmaps from image pairs
  - **MASt3R:** strengthened correspondence quality of DUSt3R
  - **Reloc3R:** directly regressed 6-DoF poses to sharpen camera estimation

  → pairwise inputs make inference cost grow quadratically with sequence length

  multi-image variants (memory encoders, subgraph fusion) aggregate context without exhaustive pairing

- further works ?

  - **VGGT, Fast3R:** further employ global attention for cross-view reasoning
  - **MV-DUSt3R+, FLARE:** couple such priors with 3D Gaussian Splatting for end-to-end reconstruction from sparse views

- recent follow-ups

  - **Dens3r:** unified dense-geometry prediction
  - **Stream3r, StreamVGGT:** causal streaming for long sequences
  - **π^3:** permutation equivariant, reference-free design
  - **Fastvggt:** training-free token-merging acceleration for VGGT

Related Works

- 4D Scene Reconstruction
  - early methods

    - self-supervised variants
    - **CasualSAM, Robust-CVD:** strengthened correspondence quality of DUSt3R
    - **MegaSaM:** achieve robust poses and reconstructions on dynamic videos by leveraging monocular depth priors
    - **Uni4D:** integrates multiple visual foundation models with multi-stage bundle adjustment for high-quality 4D results

    → effective but rely on strong priors and heavy test-time optimization which limits scalability

  - recent works

    - **MonST3R:** fine-tunes on dynamic data and exploits optical flow
    - **DAS3R:** augments a DPT head for feed-forward motion masks
    - **CUT3R:** ~~fine-tunes MASt3R on mixed static/dynamic data for fast reconstruction~~
    - **Easi3R:** adapts attention during inference to perform training-free 4D reconstruction on DUSt3R

    → most pipelines require fine-tuning, second-stage post-processing

    training-free variants are often tightly coupled to specific backbones

    → become unstable on long sequences

VGGT4D ✓

- pretrained VGGT model → 4D scene reconstruction (without further retraining)
  - multi-frame, layer-aware attention mining

    → yields globally consistent, robust dynamic masks

  - temporal window, forming a dynamic saliency signal

    → refined by a projection gradient-aware strategy, yielding sharp and robust masks

    → explicitly decouple dynamic and static regions

  - apply these masks by suppressing dynamic image tokens only in the shallow & mid layers

    → which mitigates motion interference and yields undisturbed pose estimation and 4D reconstruction

Can VGGT perceive dynamic pixels?

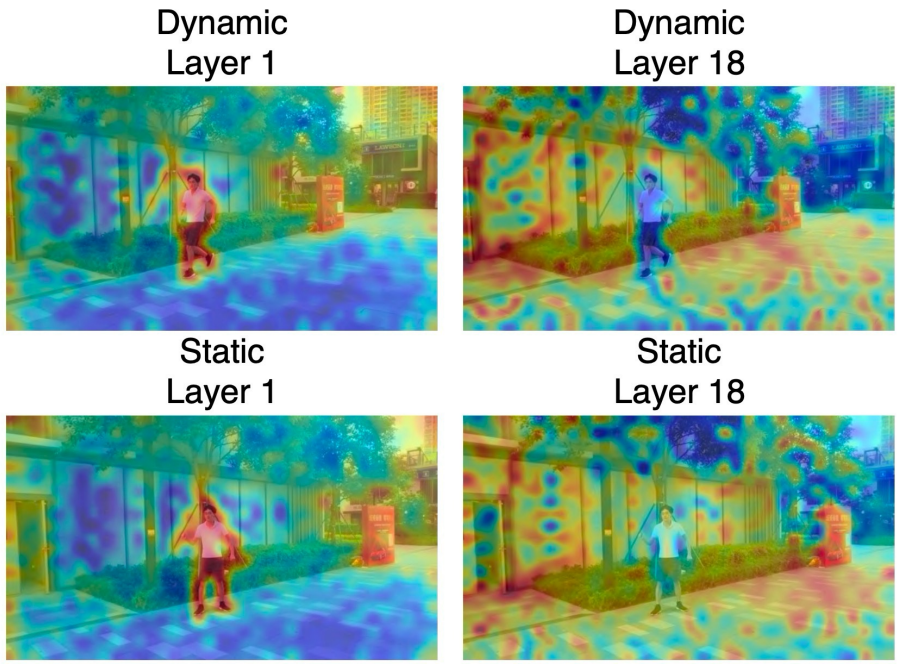


Figure 3. Visualization of VGGT's standard camera-image attention A^{QK}

however, this phenomenon is,

- highly scene-dependent
- occurs only at certain tokens and layers

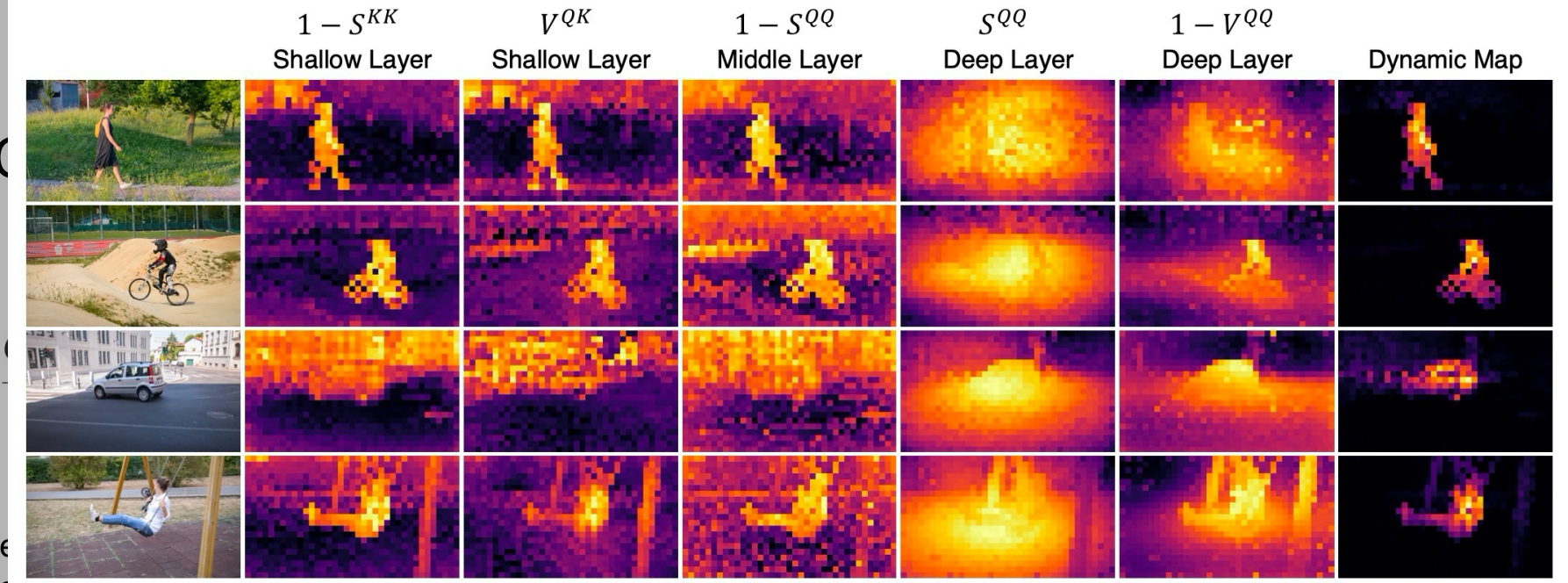
+ attention maps among image tokens contain substantial high-level semantic noise

⇒ making direct extraction unreliable

- Attention map
  - **shallow layers:** respond strongly to semantically dynamic objects (e.g. people)
  - **deeper layers:** gradually suppress pixels with inconsistent multi-view geometry

  ⇒ VGGT is sensitive to physical motion and implicitly encodes dynamic cues

Dynamic C



- Easi3R

$$A^{QK}_{l,t,s} =$$

- $A^{Q,K}$ inhe
  semantic responses

  → reducing the clarity of motion cues

- Gram Similarity

$$A^{QQ}_{l,t,s} = \frac{Q_{l,t}Q^{\top}_{l,s}}{\sqrt{c}}, \quad A^{KK}_{l,t,s} = \frac{K_{l,t}K^{\top}_{l,s}}{\sqrt{c}}. \tag{2}$$

- self-similarity matrices($QQ^{\top}, KK^{\top}$) effectively enhance dynamic cues otherwise encoded implicitly within Q and K
- inter frame sliding-window strategy to aggregate temporal information

$$\mathcal{W}(t) = \{t-n, \ldots, t-1, t+1, \ldots, t+n\}$$

- $\mathrm{X} \in \{QQ, QK, KK\}$
- i: start layer indicies
- j: end layer indicies

ilarities

, (3)

(4)

- construct dynamic saliency map

$$\mathrm{Dyn} = w_{\text{shallow}} \odot w_{\text{middle}} \odot w_{\text{deep}} \tag{5}$$

- factors

$$w_{\text{shallow}} = (1 - S^{KK}_{\text{shallow}}) \odot V^{QK}_{\text{shallow}}, \tag{6}$$
$$w_{\text{middle}} = 1 - S^{QQ}_{\text{middle}}, \tag{7}$$
$$w_{\text{deep}} = (1 - V^{QQ}_{\text{deep}}) \odot S^{QQ}_{\text{deep}}. \tag{8}$$

- w_{shallow}: captures semantic saliency
- w_{middle}: identifies motion instability
- w_{deep}: acts as a spatial prior to suppress outliers

- final mask

$$M_t = [\mathrm{Dyn} > \alpha], \text{ (+ feature clustering)}$$

RAISE LAB

Supplementary Material

Standard Attention vs Gram Similarity

- Standard Attention
  - $QK^\top$

  ⇒ strong semantic bias washes out dynamic motion signal
- Gram Similarity
  - $QQ^\top, KK^\top$

  ⇒ make physically dynamic regions salient preserving motion cue

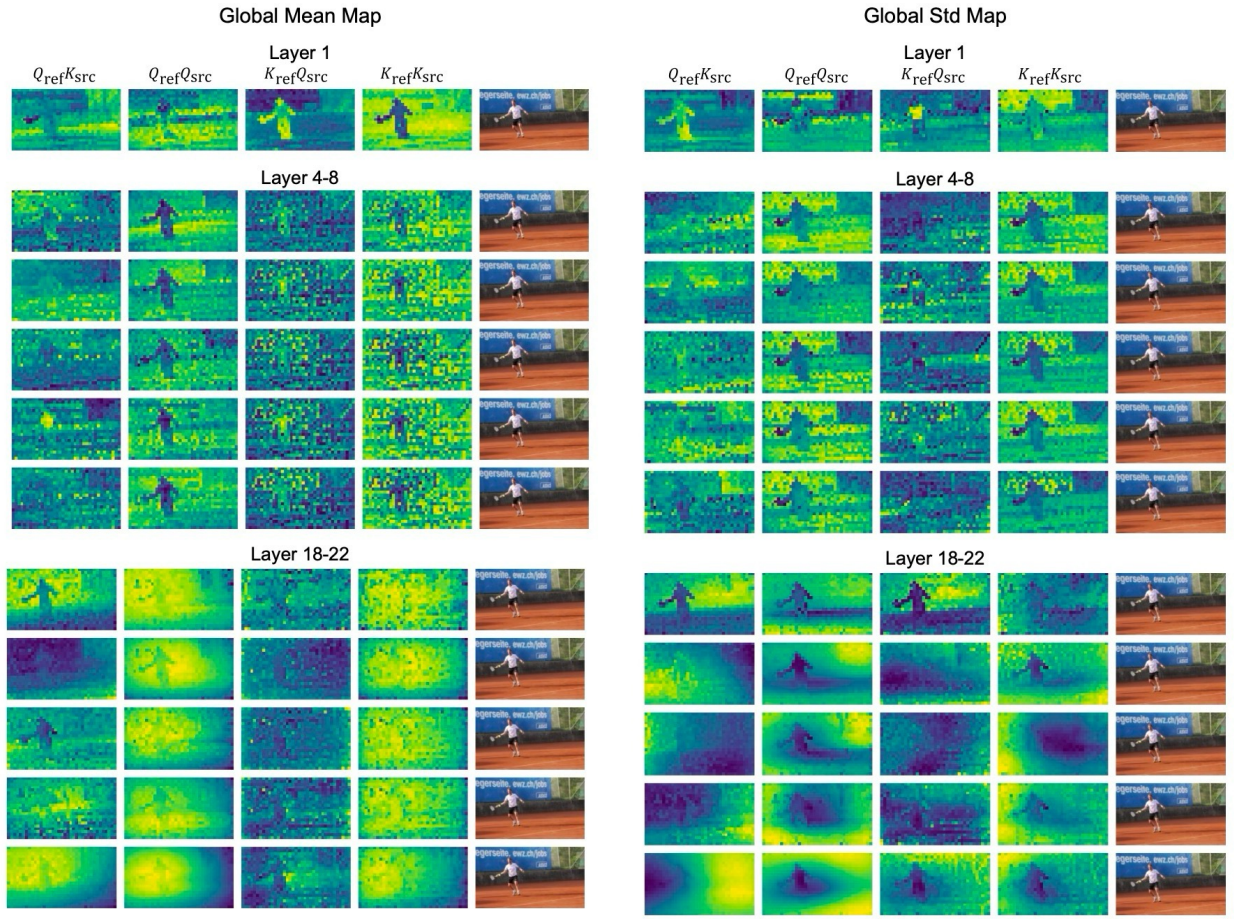
Layer-wise Dynamic Cues

- Shallow Layers (e.g. Layer 1 - $KK^\top$)

  → strong semantic bias (don't handle motion state)
- Middle Layers (e.g. Layer 4-8 - $QQ^\top$)

  → encode motion variability
- Deep Layers (e.g. Layer 18-22 - $QQ^\top$)

  → spatial priors dominate (suppress noisy responses from earlier layers)



Mask Refinement

- Why?
  - directly extracted mask are coarse
  - led to “floaters” in 4D reconstruction

  → need for refinement method to improve boundary accuracy

- How?
  - **core idea:** 3D points from dynamic objects will have large geometric and photometric errors (when projected onto the static regions of other views)

$$\mathcal{L}_{proj} = \frac{1}{2}\mathbb{I}_i(1 - M_i)\|r_{d,i}\|_2^2, \tag{9}$$

depth residual

- $r_{d,i} = d_i - D_i(u_i, v_i)$
  - d_i: projected depth
  - D_i: depth map
- M_i: initial dynamic mask
- $\mathbb{I}_i$: visibility mask

- point gradients aggregation

depth gradient

$$\mathbf{agg}^{\text{proj}} = \frac{1}{N}\sum_i^N \|w_i r_{d,i} \nabla r_{d,i}\|, \tag{10}$$
$$\text{where} \quad w_i = \mathbb{I}_i(1 - M_i).$$

- c: point's color
- C_i: sampled color in view i

for textureless regions (e.g. flat walls or floors)

$$\mathbf{agg}^{\text{photo}} = \frac{1}{N}\sum_i^N \|w_i(c - C_i(u_i, v_i))\|, \tag{11}$$

- combine aggregation score

$$\mathbf{agg}^{\text{total}} = \mathbf{agg}^{\text{proj}} + \lambda \mathbf{agg}^{\text{photo}}. \tag{12}$$

- dynamic point

$\mathbf{agg}^{\text{total}} > \tau$, (+ point cloud filtering, spatial clustering)

⇒ sharpen mask boundaries

Early-Stage Masking

- naively masking (all dynamic tokens in all layers)

  → detrimental

  - VGGT already learns to partially attenuate dynamic signals

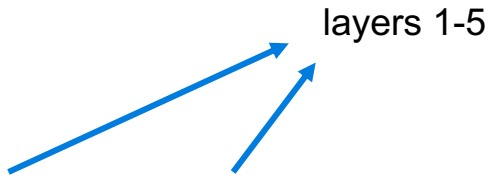
    → full-scale masking pushes the model into an out-of-distribution state

    ⇒ amplifying errors and removing valid static regions

- early-stage masking ✓

  - masking dynamic image tokens only in the shallow semantic and mid-level layers
  - suppressing the Key (K) vectors of dynamic tokens in these layers

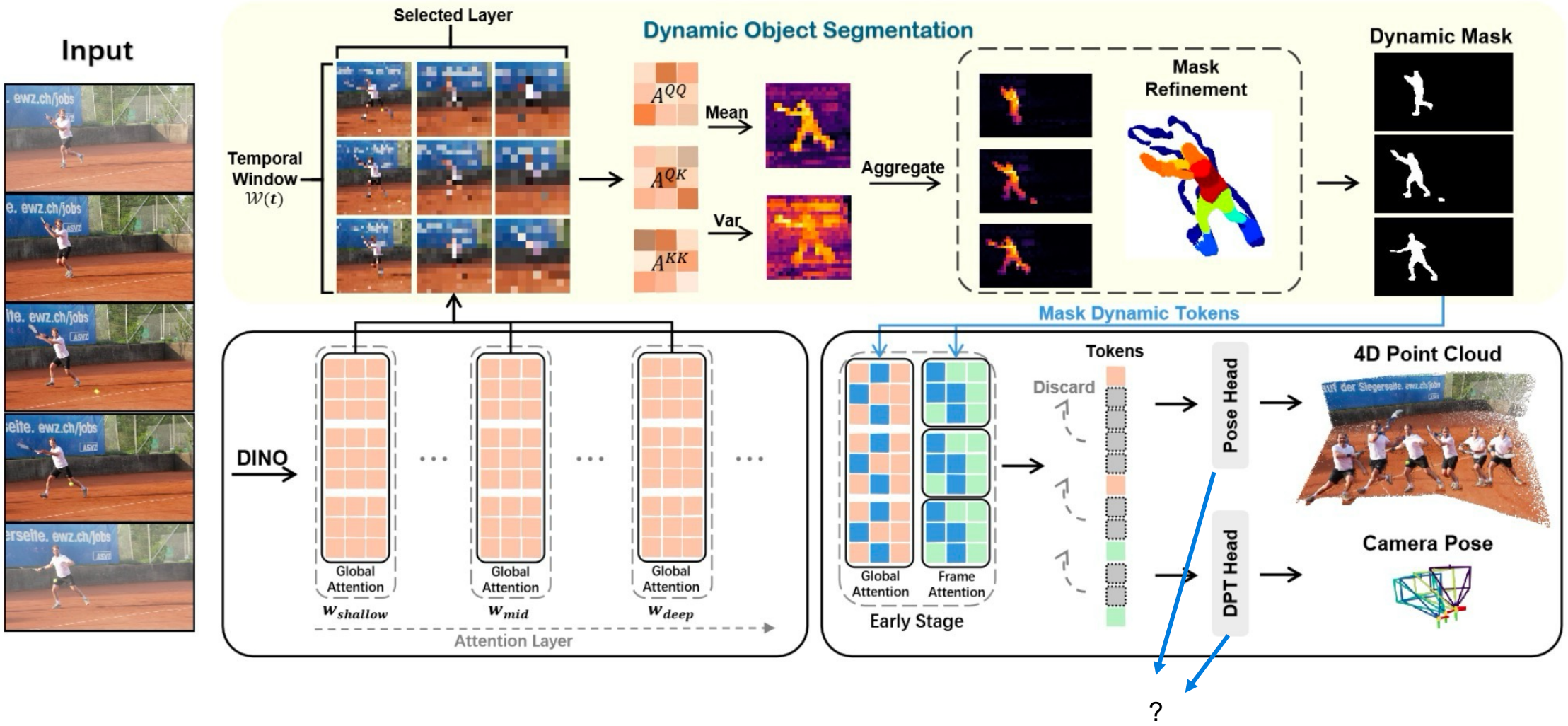
    ⇒ stable pose & clean, disentagled dynamic and static point clouds



Method	ATE↓	RTE↓	RRE↓
Full Mask	0.0302	0.0213	1.0660
VGGT	0.0131	0.0082	0.4185
Ours	**0.0106**	**0.0072**	**0.3746**

Ablation Study
(Camera Pose Estimation)

Pipeline


Input
Selected Layer
Temporal Window $\mathcal{W}(t)$
Dynamic Object Segmentation
A^{QQ}
A^{QK}
A^{KK}
Mean
Var
Aggregate
Mask Refinement
Dynamic Mask
Mask Dynamic Tokens
DINO
Global Attention
$w_{shallow}$
w_{mid}
w_{deep}
Attention Layer
Global Attention
Frame Attention
Early Stage
Tokens
Discard
Pose Head
DPT Head
4D Point Cloud
Camera Pose
?

Experiments

- Dynamic Object Segmentation
- Camera Pose Estimation
- 4D Reconstruction
- Ablation Study

Implementation Details

- single NVIDIA A6000 GPU
- modified VGGT's attention operator

Parameter	Value
Temporal Window	6 source frames (stride 2)
Layers for w_{shallow}	Layer 1
Layers for w_{middle}	Layers 4–8
Layers for w_{deep}	Layers 19–20 (Var), 18–22 (Mean)
Early-stage Masking Layers	Layers 1–5
SOR Neighbors (k)	20
SOR Std. Dev. Mult. (σ)	2.5

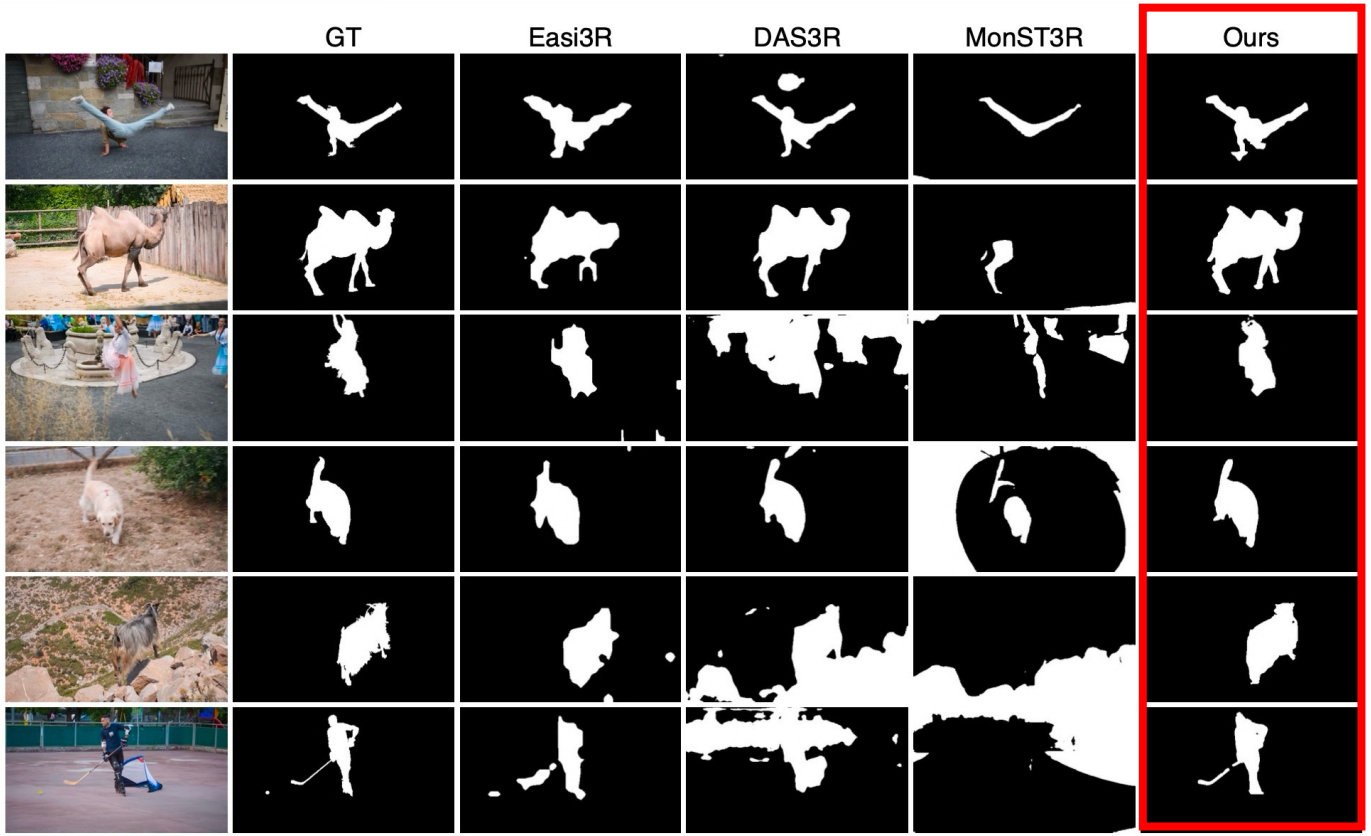
hyperparameter details

Dynamic Mask Estimation

Dynamic Object Segmentation

- Dataset
  - DAVIS-2016, DAVIS-2017, DAVIS-all
- Metric
  - JM (IoU Mean)
  - JR (IoU > 0.5)
  - FM (Boundary F-measure)
  - FR (Boundary F-measure > 0.5)

Method	DAVIS-2016				DAVIS-2017				DAVIS-all			
	JM↑	JR↑	FM↑	FR↑	JM↑	JR↑	FM↑	FR↑	JM↑	JR↑	FM↑	FR↑
$Easi3R_{dust3r}$	50.10	55.77	43.40	37.25	46.86	50.54	39.06	30.05	44.10	50.85	35.16	27.24
$Easi3R_{monst3r}$	54.93	68.00	45.29	47.30	54.75	**66.16**	44.09	42.36	**51.64**	**63.06**	40.98	38.49
MonST3R	40.42	40.39	49.54	52.12	38.07	36.05	48.24	49.01	36.98	34.52	47.03	**46.72**
DAS3R	41.13	38.67	44.50	36.94	44.51	43.95	46.71	44.96	43.33	38.93	45.24	38.78
Ours	**62.12**	**76.80**	**56.04**	**67.49**	**56.45**	65.62	**51.09**	**56.85**	50.75	55.59	**47.04**	46.43



RAISE LAB

Camera Pose Estimation

- Dataset
  - Sinetel, TUM-Dynamics, VKITTI
  - Point Odyssey (long-seq)
- Metric
  - ATE, RTE, RRE

Method	Sintel			TUM-Dynamics			VKITTI		
	ATE↓	RTE↓	RRE↓	ATE↓	RTE↓	RRE↓	ATE↓	RTE↓	RRE↓
Easi3R$_{dust3r}$	0.372	0.227	10.356	0.063	0.046	2.523	2.789	0.506	0.108
Easi3R$_{monst3r}$	0.109	0.051	0.277	0.133	0.120	4.366	2.036	0.173	0.124
MonST3R	0.151	0.034	0.258	0.156	0.103	12.041	2.272	0.180	0.091
POMATO	0.557	0.158	0.878	0.153	0.097	11.102	1.377	0.232	0.119
SpatialTrackerV2	**0.073**	0.035	0.340	0.054	0.041	2.530	0.720	0.160	0.127
DAS3R	0.125	**0.030**	**0.185**	0.155	0.102	12.038	2.043	0.169	0.114
CUT3R	0.152	0.077	0.454	0.054	0.041	5.346	5.583	0.381	0.174
VGGT	0.081	0.045	0.287	0.017	**0.020**	0.617	0.170	0.065	**0.062**
Ours	0.076	0.043	0.273	**0.016**	**0.020**	**0.612**	**0.164**	**0.064**	**0.062**

still great

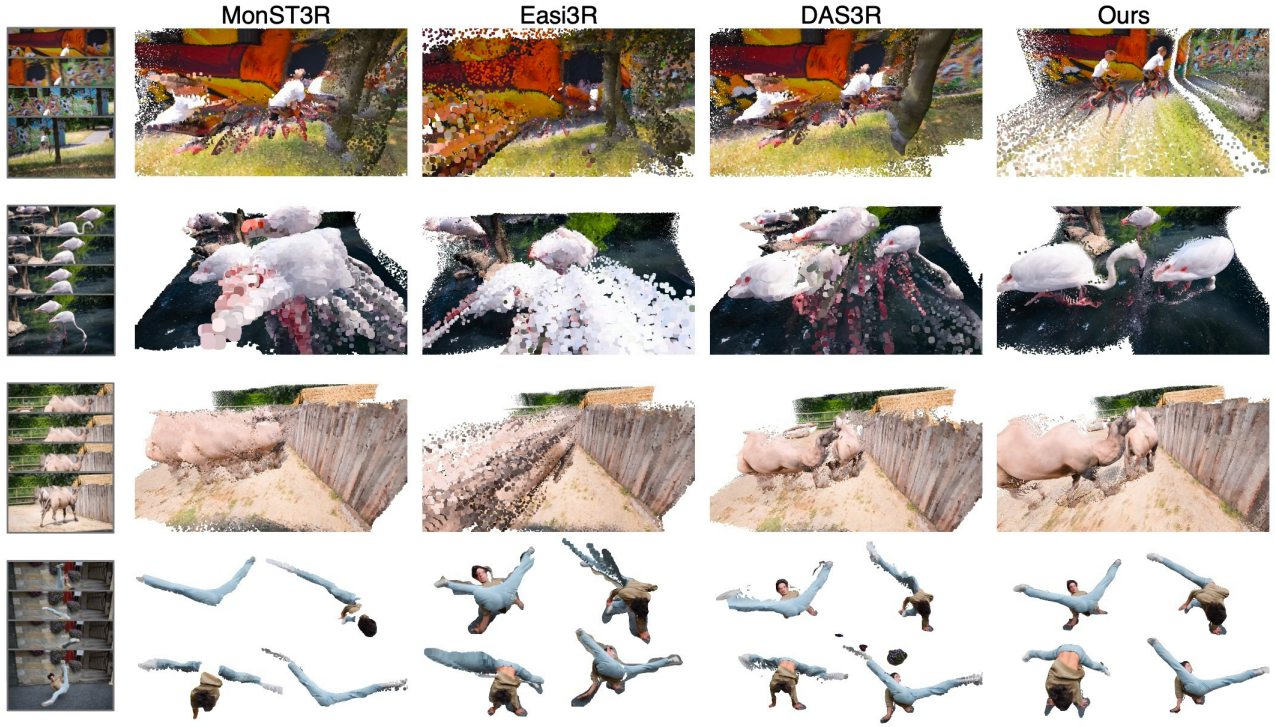
Method	ATE↓	RTE↓	RRE↓
Easi3R$_{dust3r}$	-	-	-
Easi3R$_{monst3r}$	-	-	-
MonST3R	-	-	-
POMATO	-	-	-
DAS3R	-	-	-
SpatialTrackerV2	-	-	-
CUT3R	0.417	0.028	0.605
FastVGGT	0.026	0.017	0.380
VGGT	0.022	0.015	0.344
Ours	**0.019**	**0.009**	**0.290**

4D Reconstruction

- Dataset
  - DyCheck
- Metric
  - ATE, RTE, RRE
  - Accuracy, Completeness, Distance

still great

Method	Pose Estimation			Accuracy		Completeness		Distance	
	ATE↓	RTE↓	RRE↓	Mean↓	Median↓	Mean↓	Median↓	Mean↓	Median↓
$\text{Easi3R}_{\text{dust3r}}$	0.022	0.009	0.806	0.070	0.044	0.060	0.033	0.194	0.132
$\text{Easi3R}_{\text{monst3r}}$	0.032	0.008	1.075	0.100	0.050	0.121	0.082	0.289	0.270
MonST3R	0.038	0.010	1.172	0.090	0.033	0.113	0.064	0.279	0.234
POMATO	0.128	0.027	3.648	0.960	0.950	0.814	0.776	1.484	1.434
SpatialTrackerV2	0.011	**0.006**	**0.347**	0.115	0.064	0.052	0.026	0.421	0.304
DAS3R	0.052	0.012	1.560	0.192	0.142	0.250	0.108	0.428	0.336
CUT3R	0.036	0.013	0.860	0.073	0.054	0.133	0.049	0.328	0.224
VGGT	0.013	0.008	0.418	0.028	0.009	0.063	0.019	0.150	0.055
Ours	**0.010**	0.007	0.374	**0.022**	**0.004**	**0.051**	**0.012**	**0.123**	**0.050**



Ablation

- Dataset, Metric
  - Dynamic Object Segmentation
  - Camera Pose Estimation

poor performance

already strong

best performance

Method	DAVIS-2016				DAVIS-2017				DAVIS-all			
	JM↑	JR↑	FM↑	FR↑	JM↑	JR↑	FM↑	FR↑	JM↑	JR↑	FM↑	FR↑
Easi3R$_{vggt}$	7.51	0.12	12.78	0.00	10.61	0.08	16.73	0.08	10.58	0.12	17.36	0.24
w/o refine	59.74	73.10	50.64	58.30	54.26	62.72	46.37	47.32	47.71	50.39	40.00	32.40
Ours	**62.12**	**76.80**	**56.04**	**67.49**	**56.45**	**65.62**	**51.09**	**56.85**	**50.75**	**55.59**	**47.04**	**46.43**

Dynamic Object Segmentation

harmful

already strong

optimal

Method	ATE↓	RTE↓	RRE↓
Full Mask	0.0302	0.0213	1.0660
VGGT	0.0131	0.0082	0.4185
Ours	**0.0106**	**0.0072**	**0.3746**

Camera Pose Estimation

Method	DAVIS-2016			
	JM↑	JR↑	FM↑	FR↑
w/o $w_{shallow}$	54.15	62.44	46.43	44.27
w/o w_{middle}	56.13	57.12	44.07	41.90
w/o w_{deep}	46.85	48.89	41.52	45.30
w/o refinement	59.74	73.10	50.64	58.30
Ours	**62.12**	**76.80**	**56.04**	**67.49**

layer-wise comparison

Conclusion

- VGGT → 4D Reconstruction
  - VGGT has inherent ability to perceive dynamic objects
  - leverage Gram Similarity signals to mine and amplify dynamic cues

    → enable dynamic disentanglement without relying on any external segmentation modules
  - projection gradient based refinement strategy to sharpen dynamic mask
  - integrate the refined dynamic masks into VGGT's early inference

    → effectively suppressing dynamic interference and improving both pose estimation and geometric reconstruction

Limitation

- computational overhead from computing Gram Similarities
- mask refinement depends on initial depth estimation and assumes rigid motion for projection checks