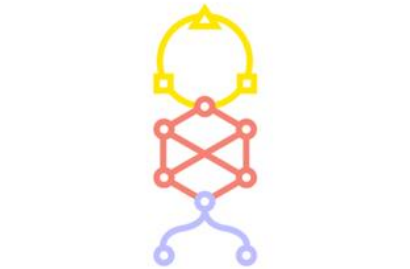
From Alpha to VGGT-Ω

Feed-Forward Geometry as a World Model for Robotics



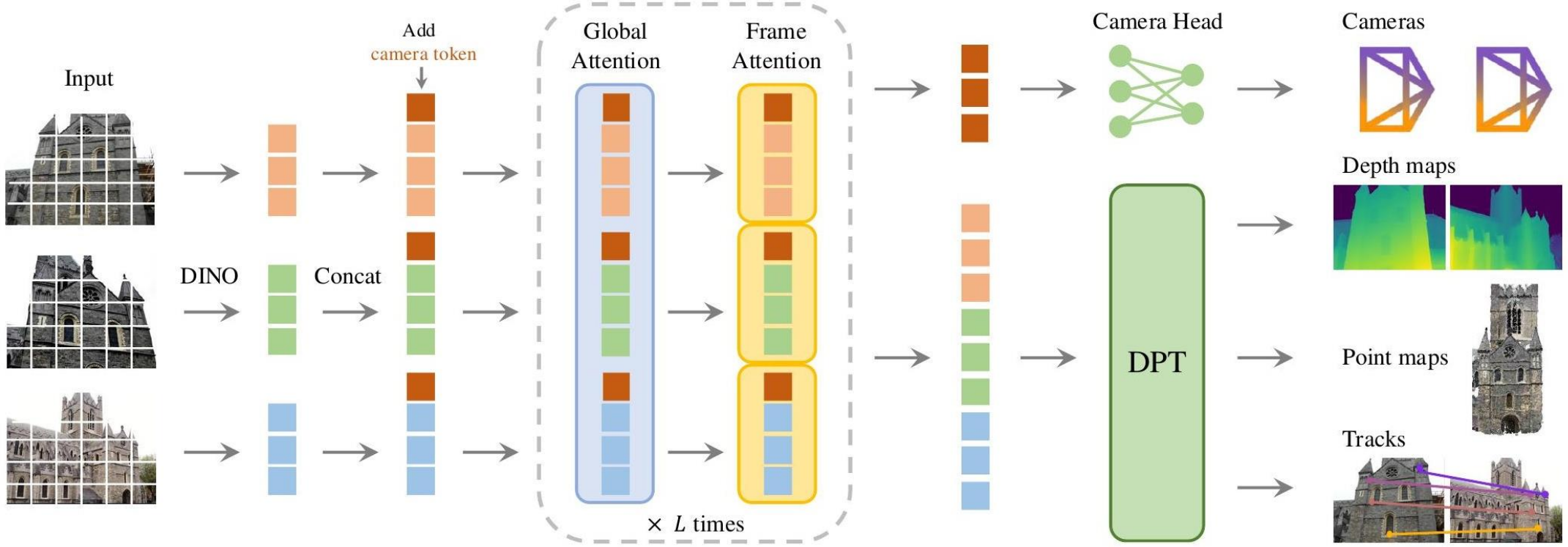
http://www.informatik.uni-hamburg.de/WTM/

Why VGGT-Ω?

https://vggt-omega.github.io/

VGGT: Visual Geometry Grounded Transformer

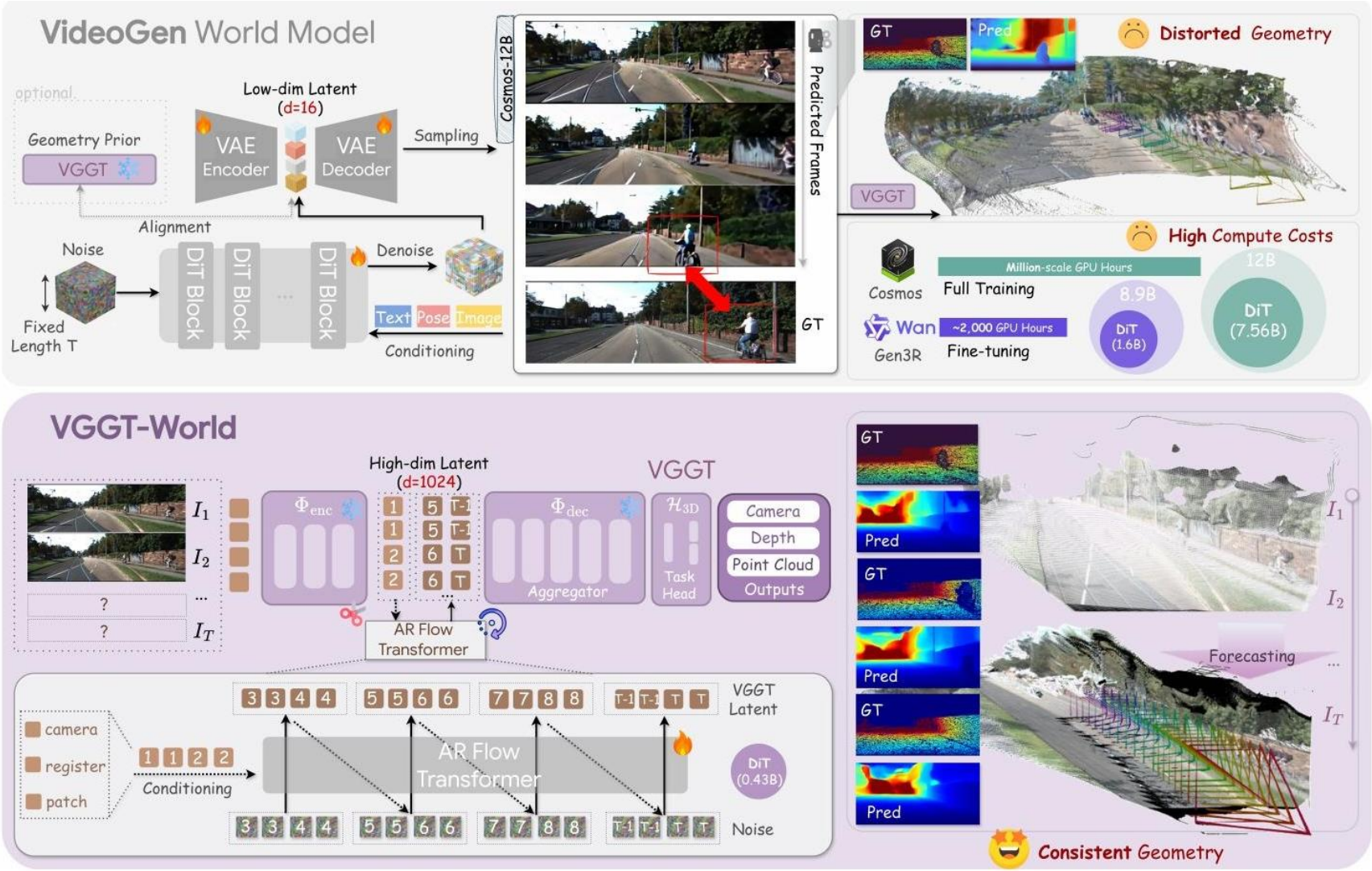
CVPR 2025



Wang et al., CVPR 2025

KNOWLEDGE TECHNOLOGY

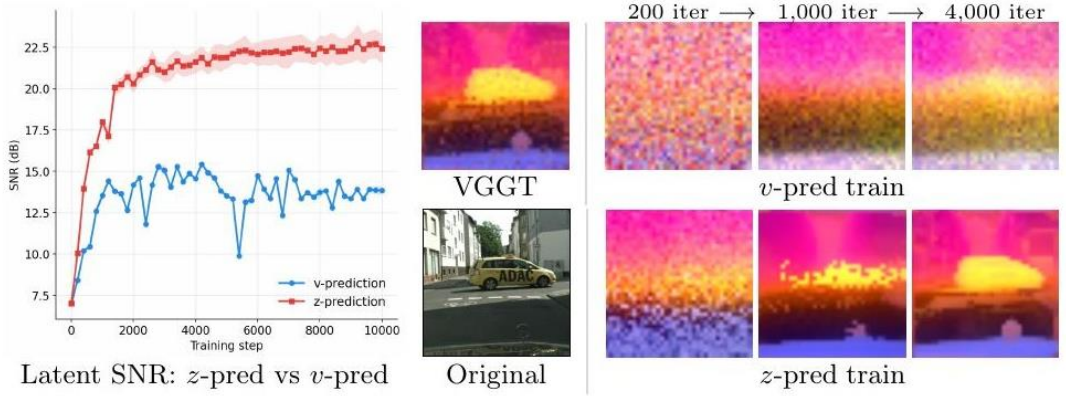
VGGT-World: Forecast Geometry, Not Pixels



Sun et al., arXiv 2026

VGGT-World: Two Challenges of a d=1024 Latent

1. Curse of Dimensionality



Sun et al., arXiv 2026

Fix: z-prediction of clean latent, instead of velocity

2. Compounding Rollout Errors

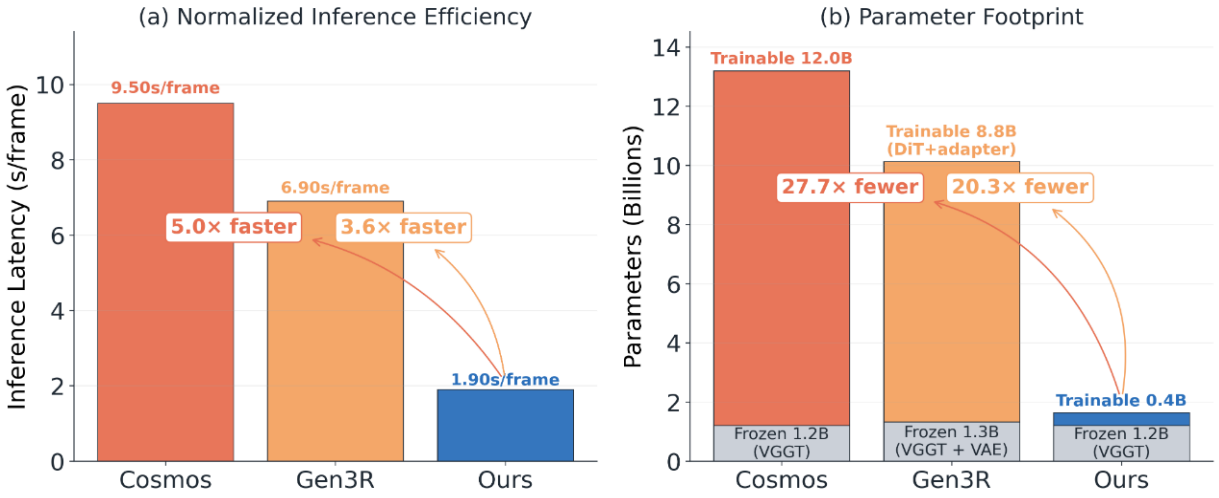
Method	AbsRel (↓)		δ_1 (↑)	
	Short	Mid	Short	Mid
Cosmos-4B [1]	0.189	0.247	76.3	70.5
Cosmos-12B [1]	0.181	0.241	78.2	72.4
COPY-LAST (VGGT [50])	0.154	0.212	84.1	77.8
VISTA_{ft} [15]	0.124	0.153	86.4	82.8
DINO-Foresight [23]	0.114	0.136	88.6	85.4
VGGT-World w/o Forcing	0.079	0.131	93.7	86.8
VGGT-World	**0.078**	**0.126**	**93.8**	**87.3**
VGGT-World ($\lambda = 0.1$)	0.084	0.142	93.1	85.0
VGGT-World ($\lambda = 0.3$)	0.089	0.148	92.6	84.9
VGGT-World ($\lambda = 0.7$)	0.109	0.165	90.1	81.4

Sun et al., arXiv 2026

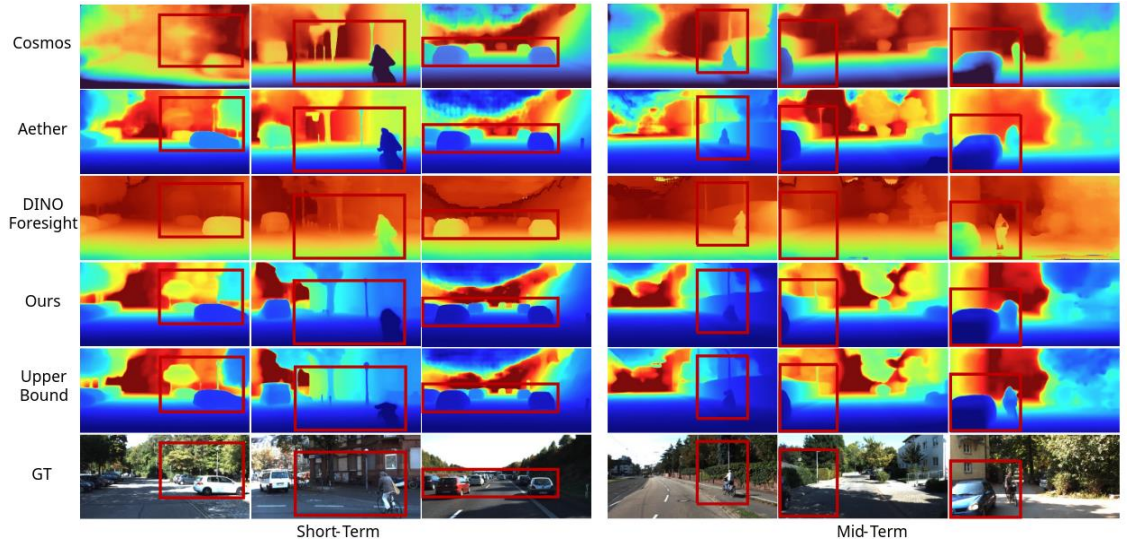
Fix: flow-forcing curriculum anneal λ: 0⟶1

KNOWLEDGE TECHNOLOGY

VGGT-World: Results



Sun et al., arXiv 2026



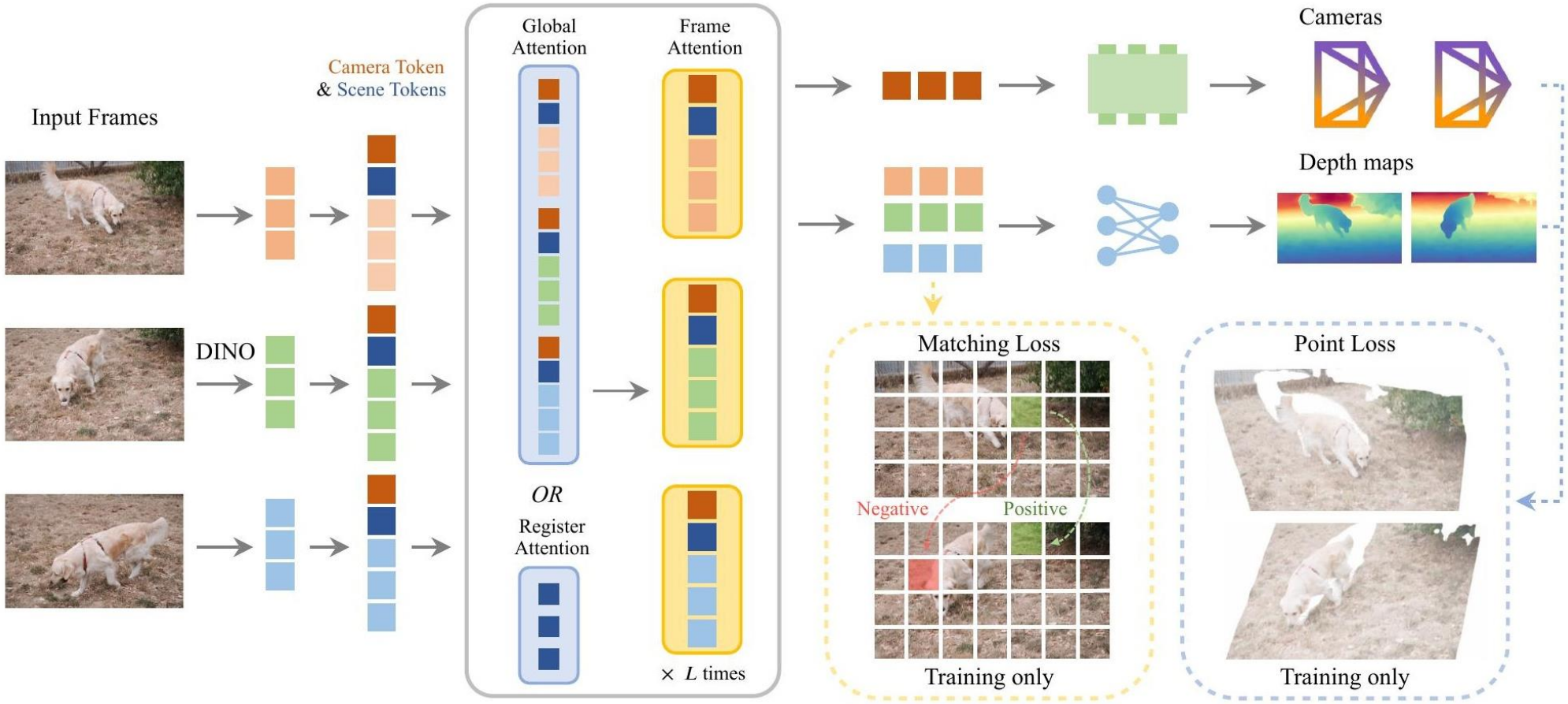
Sun et al., arXiv 2026

Method	2011_09_26_drive_0002_sync				2011_10_03_drive_0047_sync				Mean (Dataset Avg.)			
	AbsRel (↓)		δ_1 (↑)		AbsRel (↓)		δ_1 (↑)		AbsRel (↓)		δ_1 (↑)	
	Short	Mid	Short	Mid	Short	Mid	Short	Mid	Short	Mid	Short	Mid
Cosmos-4B [1]	0.138	0.148	84.0	81.4	0.155	0.156	77.5	75.2	0.155	0.165	77.1	74.1
Cosmos-12B [1]	0.131	0.139	84.5	79.7	0.149	0.176	77.6	75.9	0.154	0.185	77.5	71.5
Copy Last (VGGT [50])	0.112	0.125	89.6	82.6	0.140	0.153	84.4	79.6	0.137	0.144	83.7	78.5
Aether [44]	0.074	0.102	94.2	88.0	0.113	0.111	87.4	85.2	0.096	0.124	87.9	84.7
DINO-Foresight [23]	0.063	0.105	96.5	89.9	0.076	0.107	90.9	86.7	0.082	0.115	91.6	85.1
VGGT-World	**0.050**	**0.078**	**98.3**	**93.2**	**0.064**	**0.101**	**94.9**	**92.7**	**0.065**	**0.098**	**94.0**	**89.4**

Sun et al., arXiv 2026

KNOWLEDGE TECHNOLOGY

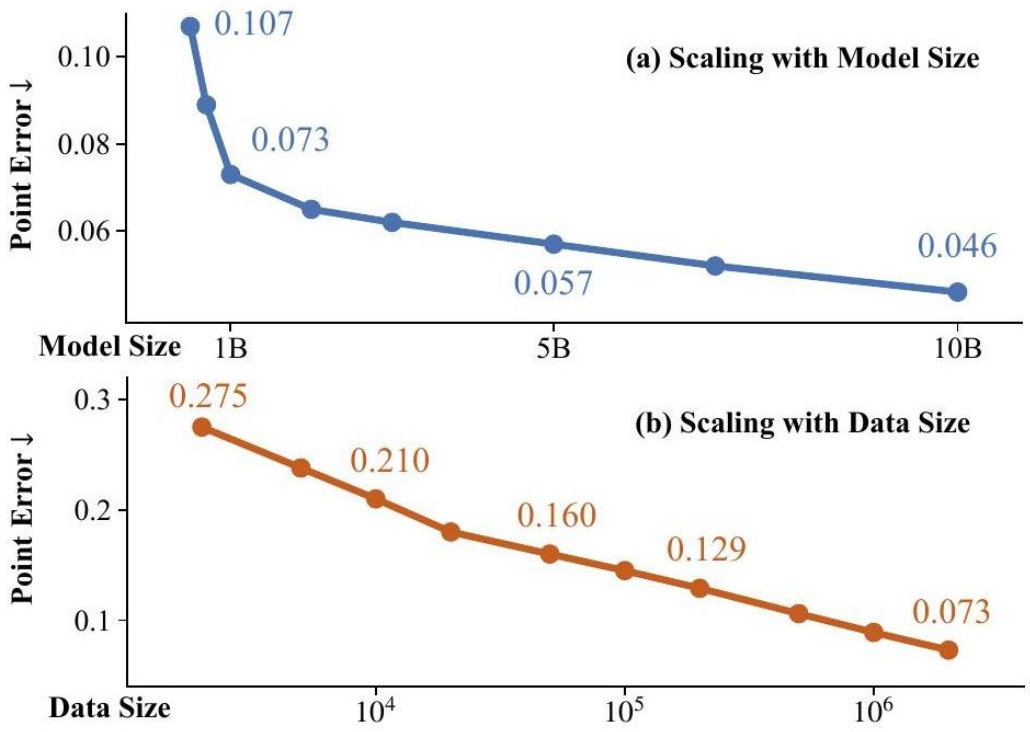
VGGT-Ω: Scene Tokens as a Bottleneck



Wang et al., CVPR 2026

KNOWLEDGE TECHNOLOGY

VGGT-Ω: Scaling Up



Wang et al., CVPR 2026

15×
more training data
(4M sequences)

70%
less GPU memory
during training

0.2B → 10B
parameters, same
architecture family

KNOWLEDGE TECHNOLOGY

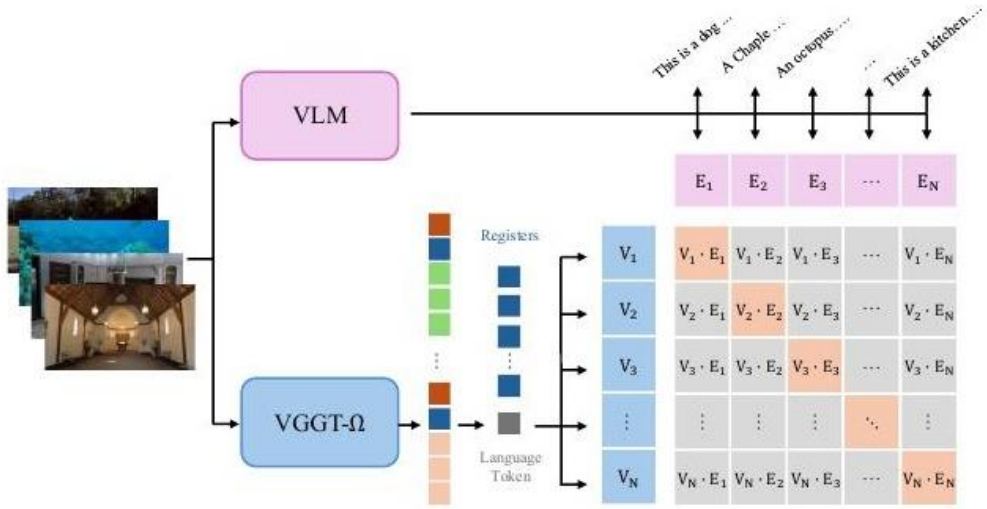
Scene Tokens Store Global Semantic Information

Robotics (VLA)

Method	Spatial SR (%)	Object SR (%)	Goal SR (%)	Long SR (%)	Average SR (%)
Diffusion Policy	78.3	92.5	68.3	50.5	72.4
TraceVLA	84.6	85.2	75.1	54.1	74.8
Octo	78.9	85.7	84.6	51.1	75.1
OpenVLA	84.7	88.4	79.2	53.7	76.5
Dita	84.2	96.3	85.4	63.8	82.4
CoT-VLA	87.5	91.6	87.6	69.0	83.9
π_0-FAST	96.4	96.8	88.6	60.2	85.5
π_0	96.8	98.8	95.8	85.2	94.2
UniVLA	96.5	96.8	95.6	92.0	95.2
OpenVLA-OFT	97.6	98.4	97.9	94.5	97.1
OpenVLA-OFT + Our Frozen Scene Tokens	**99.3**	**99.2**	**99.0**	**96.7**	**98.5**

Wang et al., CVPR 2026

Language alignment



Wang et al., CVPR 2026

76.8% top-1 video–text retrieval accuracy

Wang et al., CVPR 2026

Motion-Aware Representations as emerging property without explicit supervision

KNOWLEDGE TECHNOLOGY

Thank You

Questions?

References:

- Wang et al. VGGT: Visual Geometry Grounded Transformer. CVPR 2025.
- Sun et al. VGGT-World: Transforming VGGT into an Autoregressive Geometry World Model. arXiv 2026.
- Wang et al. VGGT-ohm. CVPR 2026.

KNOWLEDGE TECHNOLOGY