VGGT-HPE: Reframing Head Pose Estimation as Relative Pose Prediction

Vasiliki Vasileiou[1,2,4] Panagiotis P. Filntisis[2,3] Petros Maragos[2,3,4] Kostas Daniilidis[1,5]

[1]Archimedes, Athena Research Center, Marousi, Greece [2]HERON – Hellenic Robotics Center of Excellence, Athens, Greece
[3]Robotics Institute, Athena Research Center, Marousi, Greece [4]School of ECE, National Technical University of Athens, Greece
[5]University of Pennsylvania

CVPR JUNE 3-7, 2026 DENVER COLORADO

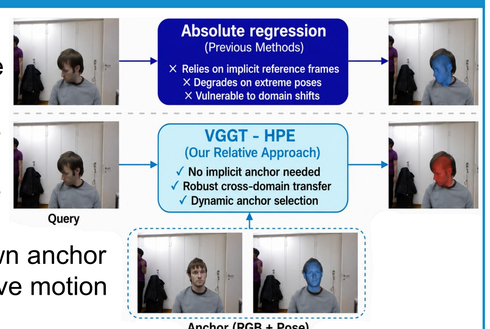
1 Motivation

🔥 **Old methods:** regress pose from one image relative to a hidden canonical frame.

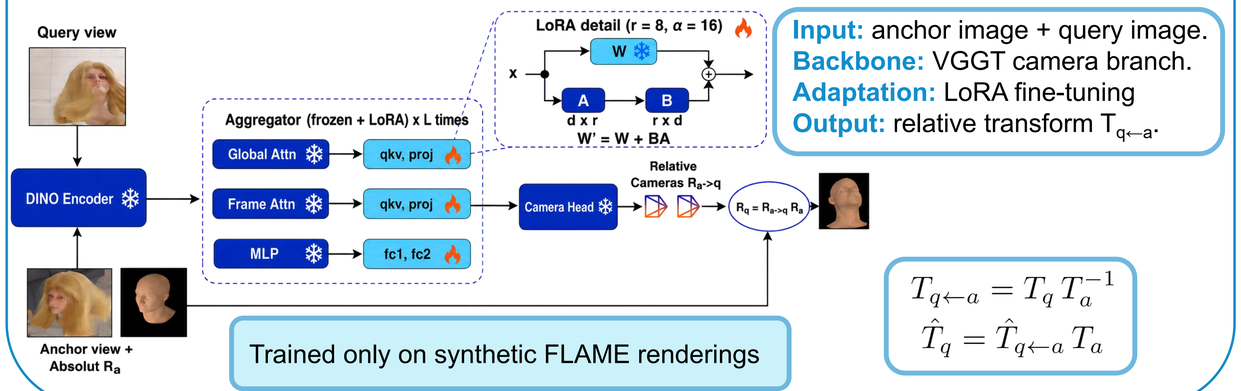
🔥 **Problem:** extreme poses and domain shifts are hard because the reference frame is only learned implicitly.

🔥 **VGGT-HPE:** uses a known anchor image and predicts the relative motion to the query.

🔥 **Benefit:** choose an easy anchor at test time to control prediction difficulty.



2 Method

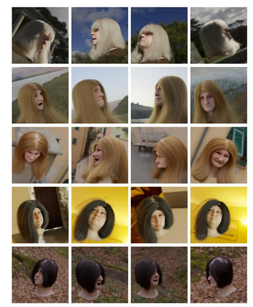


Input: anchor image + query image.
Backbone: VGGT camera branch.
Adaptation: LoRA fine-tuning
Output: relative transform $T_{q\leftarrow a}$.

Trained only on synthetic FLAME renderings

$$T_{q\leftarrow a} = T_q\, T_a^{-1}$$
$$\hat{T}_q = \hat{T}_{q\leftarrow a}\, T_a$$

3 Synthetic Training Data



- 250 identities, diverse hair-styles, expressions, viewpoints, HDRI lighting
- 25k images generated using Blender
- Supervision on rotation, translation, and field of view

4 Main Results on BIWI

Method	Yaw ↓	Pitch ↓	Roll ↓	MAE ↓	Data
Reported numbers					
6DRepNet	3.24	4.48	2.68	3.47	M
TokenHPE	3.95	4.51	2.71	3.72	M
TRG	3.04	3.44	**1.78**	**2.75**	M
VGGT-HPE (Rel., ours)	**2.24**	**3.04**	3.17	2.82	**S**
Reproduced under shared MTCNN protocol					
6DRepNet	3.74	4.95	**3.04**	3.91	M
TokenHPE-v1	5.57	6.23	3.79	5.20	M
TRG	4.58	7.18	3.68	5.15	M
VGGT-HPE-Abs (ours)	4.90	7.01	3.53	5.15	**S**
VGGT-HPE (Rel., ours)	**2.24**	**3.04**	3.17	**2.82**	**S**

5 Controlled Benchmarks

Hard pairs
8.87°
Neutral anchor / Extreme query

Easy pairs
0.96°
Near-neutral anchor / small gap

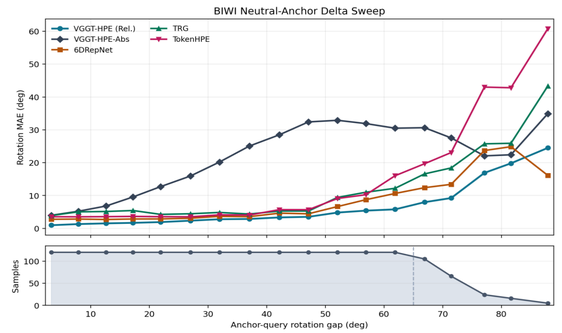
Hard benchmark

Method	MAE ↓
VGGT-HPE-Abs	31.02
TokenHPE	22.51
TRG	17.23
6DRepNet	13.33
VGGT-HPE (Rel., ours)	**8.87**

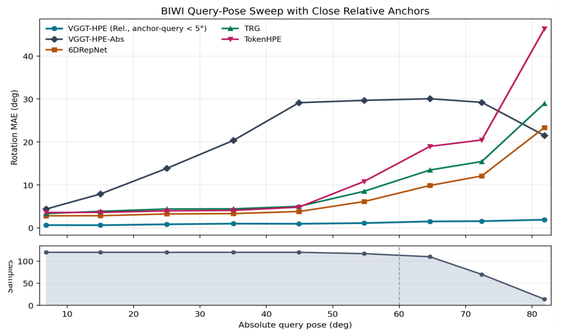
Easy benchmark

Method	MAE ↓
VGGT-HPE-Abs	3.98
TokenHPE	3.61
TRG	3.45
6DRepNet	2.80
VGGT-HPE (Rel., ours)	**0.96**

6 Error Analysis



Error vs. anchor-query rotation gap:
VGGT-HPE remains strongest across the full range.



Extreme poses remain easy with close anchors:
what matters is the relative gap, not the absolute pose.

7 Qualitative Results (BIWI)

8 Ablation

🔎 **LoRA adapts best:** full fine-tuning hurts VGGT priors; head-only overfits synthetic data.

🔎 **Predicted anchors are effective:** using a 6DRepNet anchor achieves **4.08°** on the full BIWI set and **10.26°** on the hard subset.

🔎 **Full supervision helps:** rotation-only gets **3.12°**, while rotation + translation + FoV reaches **2.82°**.

Relative formulation + LoRA + full camera supervision = best cross-domain transfer.

9 Takeaways

- ✓ Relative prediction is easier than absolute regression.
- ✓ The same backbone performs much better in relative mode.
- ✓ VGGT-HPE transfers from synthetic training to real BIWI images.
- ✓ Anchor choice at test time provides controllable difficulty.

v.vasileiou@athenarc.gr