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

Abstract

Monocular head pose estimation is traditionally formulated as direct regression from a single image to an absolute pose. This paradigm forces the network to implicitly internalize a dataset-specific canonical reference frame. In this work, we argue that predicting the relative rigid transformation between two observed head configurations is a fundamentally easier and more robust formulation. We introduce VGGT-HPE, a relative head pose estimator built upon a general-purpose geometry foundation model. Fine-tuned exclusively on synthetic facial renderings, our method sidesteps the need for an implicit anchor by reducing the problem to estimating a geometric displacement from an explicitly provided anchor with a known pose. As a practical benefit, the relative formulation also allows the anchor to be chosen at test time — for instance, a near-neutral frame or a temporally adjacent one — so that the prediction difficulty can be controlled by the application. Despite zero real-world training data, VGGT-HPE achieves state-of-the-art results on the BIWI benchmark, outperforming established absolute regression methods trained on mixed and real datasets. Through controlled easy- and hard-pair benchmarks, we also systematically validate our core hypothesis: relative prediction is intrinsically more accurate than absolute regression, with the advantage scaling alongside the difficulty of the target pose. Project page and code: https://vasilikivas.github.io/VGGT-HPE

1. Introduction

Head pose estimation - recovering the orientation of a human head from a single image - is a fundamental building block in computer vision [1, 28]. Accurate head pose drives a wide range of downstream applications, from gaze estimation [43] and driver attention monitoring [17] to sign language recognition [20], facial behavior analysis in the wild [21], human-robot interaction [27], and augmented reality [26]. Despite decades of progress, robust head pose estimation in unconstrained settings remains challenging due to appearance variability, partial occlusions, and the inherent ambiguity of monocular 3D inference.

The dominant paradigm in learned head pose estimation treats the problem as direct regression from a single image to an absolute pose vector — typically three Euler angles (yaw, pitch, roll) and, less commonly, a 3D translation. Methods such as HopeNet [29], 6DRepNet [13], and WHENet [44] train deep networks to map face crops directly to global orientation. This requires the model to internalize a canonical reference frame from the training data - one that is never explicitly provided at test time. More recent works improve robustness through better rotation representations [13], transformer-based modeling [40], or ex-

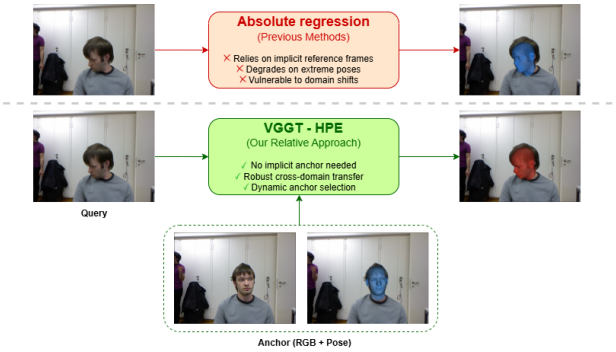

Figure 1. Absolute methods (top) regress pose from a single image relative to an implicit canonical frame learned during training, making them brittle under domain shift and extreme poses. VGGT-HPE (bottom) predicts the relative transformation between an anchor with known pose and a query, recovering the target pose by composition. This sidesteps the need for an implicit reference frame, transfers better from synthetic to real data, and allows the anchor to be chosen at test time to control prediction difficulty.

*The research work of P. P. Filntisis and P. Maragos was supported by the project "Applied Research for Autonomous Robotic Systems" (MIS 5200632), implemented within the framework of the National Recovery and Resilience Plan "Greece 2.0" (Measure: 16618 – Basic and Applied Research), and funded by the European Union – NextGenerationEU.

plicit facial geometry [6], but the core formulation remains unchanged. As we show in our experiments, these models tend to degrade for extreme poses, where the implicit anchor - a neutral, frontal configuration baked into the learned weights - lies far from the target.

We argue that relative head pose estimation is a fundamentally easier problem. Predicting the rigid transformation between two observed head configurations sidesteps the need for an implicit canonical frame and reduces the task to estimating a displacement. Because it observes both states, the network can solve this via implicit feature matching—shifting the paradigm from absolute classification to direct image matching. This is simpler and transfers better across domains, since relative geometry does not depend on absolute coordinate conventions or appearance distributions specific to the training set. The relative formulation also introduces a practical degree of freedom that absolute methods lack: the anchor can be chosen at test time. In video, for instance, a temporally close frame with known pose keeps the relative rotation gap small, placing the prediction well within the model's reliable operating range.

In this work we introduce VGGT-HPE, a head pose estimator built on the Visual Geometry Grounded Transformer (VGGT) [33], a large-scale geometry foundation model pretrained for general-purpose camera pose and 3D reconstruction. VGGT already learns to estimate relative camera poses between arbitrary view pairs, making it a natural backbone for our relative formulation. We fine-tune VGGT with LoRA [15] on synthetic two-view head pairs rendered from FLAME [24]. At test time, we compose the predicted relative pose with a known anchor pose to recover the absolute target pose in any desired coordinate frame. Despite training exclusively on synthetic data, VGGT-HPE achieves state-of-the-art results on the BIWI benchmark [10] outperforming methods trained on mixed and real datasets.

Our contributions are as follows:

- A relative formulation for head pose estimation that reframes the problem as rigid displacement prediction between two views, eliminating the dependence on an implicit canonical frame and enabling strategic anchor selection at test time.
- A lightweight architecture that adapts VGGT [33], a general-purpose geometry foundation model, to head-specific relative pose estimation via LoRA fine-tuning [15] on synthetic FLAME [24] renderings, requiring no real-world training data.
- State-of-the-art results on BIWI [10] using only synthetic training data, and a detailed analysis through controlled benchmarks - easy-pair, hard-pair, and binned error-vs-rotation-gap sweeps - that directly validates the hypothesis that relative prediction is fundamentally easier than absolute regression, with the advantage growing as the anchor-target displacement increases.

2. Related Work

Monocular Head Pose Estimation. Early head pose estimation methods relied on facial landmarks and 3D model fitting, typically recovering pose through geometric alignment or PnP-style optimization. More recent learning-based approaches instead regress pose directly from a face crop. HopeNet [29] introduced a combined classification-and-regression strategy for Euler-angle prediction, showing that direct image-to-pose learning can be competitive without explicit landmark fitting. Subsequent works improved robustness to wider pose ranges and better rotation representations. WHENet [44] targeted wide-range yaw estimation with a lightweight design suitable for real-time use, while 6DRepNet [13] replaced Euler-angle regression with a continuous 6D rotation representation and a geodesic loss, improving stability under large rotations. Transformer-based methods such as TokenHPE [40] further demonstrated that richer relational modeling over facial appearance can improve monocular head pose estimation. More recent 6DoF methods, including TRG [6], also incorporate facial geometry explicitly in order to improve translation and full pose prediction. Despite these differences, the dominant paradigm remains the same: a single image is mapped directly to an absolute pose defined in a dataset-specific canonical frame. As a result, the network must implicitly learn both the rigid orientation and the hidden reference frame relative to which that orientation is expressed. Furthermore, training these absolute regressors typically requires massive, real-world datasets often expanded via synthetic face profiling (e.g., 300W-LP [45]), which leaves them vulnerable to domain shifts. Our relative formulation, by contrast, treats pose as an image matching task, allowing it to generalize across domains even when trained exclusively on synthetic data.

Relative Pose Formulations and Geometric Composition. Estimating the relative pose between two views is a standard formulation in visual odometry, SLAM, and rigid object tracking [9, 31, 35]. In these settings, predicting the relative transformation between a query and a reference view tends to be more robust and to generalize better than regressing an absolute pose in a fixed coordinate system. In head pose estimation, however, most learning-based methods still operate in the absolute, single-image regime. Temporal or multi-frame information is sometimes used in video-based head pose estimation, but typically for recurrent feature aggregation, temporal smoothing, or heuristic tracking [12, 36]—not for explicit geometric composition. A few works explore delta-pose or relative rotation for specific sub-tasks, such as enforcing temporal consistency or regularizing semi-supervised training [22], but they do not output explicit relative pose transformations that can be

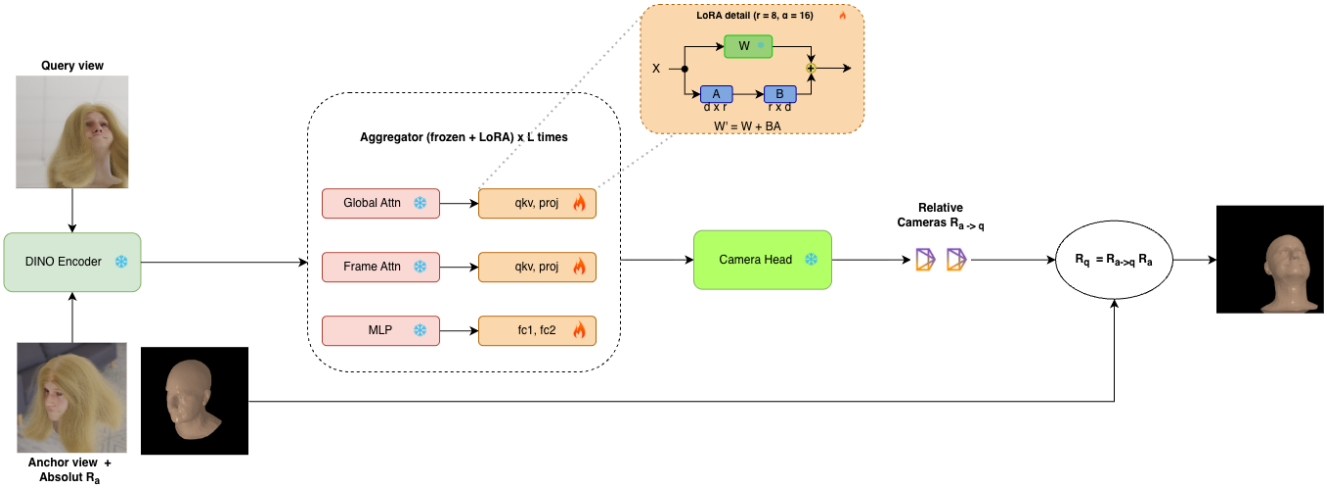

Figure 2. **Overview of the VGGT-HPE architecture.** The model takes an anchor image (I_a) and a query image (I_q) to predict the relative rigid transformation ($T_{q \leftarrow a}$) between them. We use a frozen, pre-trained VGGT backbone efficiently fine-tuned with LoRA for the facial domain. The final absolute query pose ($\hat{T}_q$) is recovered by composing the predicted relative displacement with the known anchor pose.

composed at test time. Our work brings the relative formulation into head pose estimation. By predicting the relative pose between an anchor with known pose and a target frame, we remove the dependence on an implicit, dataset-specific canonical frame, connecting head pose estimation to the reference-based matching formulations used in general 3D vision.

Geometry Foundation Models. Recent progress in 3D computer vision has moved from task-specific architectures toward large-scale geometry foundation models. CroCo [37], DUSt3R [34], and MASt3R [23] showed that general-purpose 3D priors—dense point maps, depth, and relative camera poses—can be learned through pre-training on large, diverse multi-view datasets. VGGT [33] extends this line of work with a feed-forward transformer that jointly infers camera parameters and 3D scene structure from uncalibrated views. Traditional head pose estimators must learn perspective projection and 3D rotation from scratch on limited facial datasets. Models like VGGT, by contrast, already encode an understanding of epipolar geometry and relative spatial transformations from their pre-training data. Our approach builds on this: rather than designing a new architecture for faces, we adapt the pre-trained geometric reasoning of VGGT, showing that a general 3D foundation model can be efficiently specialized to head pose estimation.

3. Methodology

Overview. Our goal is to estimate the pose of a query head image by predicting its rigid transformation relative to an anchor image with known pose. The key intuition is that absolute methods must map a single crop to a pose defined in an implicit canonical frame that is never provided at test time, while the relative formulation turns the problem into image matching between two visible inputs — a task that is geometrically simpler and naturally sidesteps the domain gap. We build on VGGT [33], a geometry foundation model whose camera pose branch already estimates relative rigid transformations between two images, and specialize it to the facial domain via LoRA fine-tuning on synthetic FLAME renderings. An overview of our architecture can be seen in Fig. 2.

Backbone and Adaptation. VGGT is a feed-forward transformer pretrained on large-scale multi-view data to jointly predict camera parameters, depth, point maps, and correspondences. We retain only the camera branch and disable all other heads. Each camera pose is encoded as

$$p = [t, q, \phi_h, \phi_w], \tag{1}$$

where $t \in \mathbb{R}^3$ is translation, $q \in \mathbb{R}^4$ is a quaternion rotation, and ϕ_h, ϕ_w are the horizontal and vertical fields of view.

To adapt VGGT without destroying its pretrained priors, we use LoRA [15]: for each pretrained linear layer W, a low-rank update $\Delta W = BA$ is learned while W remains frozen, yielding $W' = W + BA$. LoRA modules are inserted into the attention and MLP projections of both the frame-level and global transformer blocks, with rank 8 and scaling factor 16.

Relative Formulation. Given an anchor image I_a with known pose $T_a \in SE(3)$ and a query image I_q with unknown pose T_q, the model predicts the relative transforma-

tion

$$T_{q\leftarrow a} = T_q\, T_a^{-1}. \tag{2}$$

The absolute query pose is then recovered by composition:

$$\hat{T}_q = \hat{T}_{q\leftarrow a}\, T_a. \tag{3}$$

The advantage over absolute regression is twofold. First, absolute methods must map a single crop to a pose defined in a canonical frame that is never provided — it exists only implicitly in the training data. The relative formulation removes this requirement: the model just estimates how much the head moved between two images it can see. Second, by conditioning on a pair of images, the model can perform implicit feature matching between anchor and query, grounding its prediction in visual correspondence rather than memorized pose distributions. This makes the task closer to image matching than to classification, which is both geometrically simpler and naturally robust to domain shift — the relative displacement between two views of a head looks the same whether the images are synthetic or real.

During training, we normalize each pair to the anchor frame: given extrinsics $\{T_1, T_2\}$, we set $\tilde{T}_i = T_1^{-1} T_i$ so the anchor becomes the identity. In contrast with the original VGGT [33], which supervises the poses of all input frames, we supervise only the second frame's pose. We also study an absolute baseline (VGGT-HPE-Abs) using the same backbone and data but without relative normalization, to isolate the effect of the formulation.

Synthetic Training Data. We generate training pairs using FLAME [24] head meshes rendered in Blender [3], following a pipeline similar to [11]. The dataset comprises 250 identities — 200 with hairstyles from HAAR [30] and 50 bald — each rendered under two HDRI environment maps with ten FLAME expressions and five viewpoints per expression, yielding 25k images. Albedo is sampled from 54 textures covering diverse skin tones. Each pair shares identity and lighting while expression and viewpoint vary freely. At training time, we crop a square face region with random margin and shift, updating intrinsics accordingly, and apply pairwise appearance augmentation (color jitter, gamma, CLAHE, RGB shift, blur, noise) coherently across both views. We show representative samples of our rendered scenes in Fig. 3.

Training Objective. We supervise not only rotation but also translation and field of view, as we found that jointly predicting all camera parameters improves rotation accuracy even when only orientation is needed at test time. VGGT produces pose predictions at K stages. We supervise all stages with exponentially decayed weights:

$$\mathcal{L}_{\text{cam}} = \frac{1}{K} \sum_{k=1}^{K} \gamma^{K-k} \Big(\lambda_T \mathcal{L}_T^{(k)} + \lambda_R \mathcal{L}_R^{(k)} + \lambda_F \mathcal{L}_F^{(k)} \Big), \tag{4}$$

Figure 3. Samples from our synthetic training set. Each row shows a single identity rendered under varying viewpoints, expressions, and HDRI environment maps.

where γ is a decay factor. The translation loss is $\mathcal{L}_T = \|\hat{t} - t\|_1$ and the rotation loss operates in quaternion space: $\mathcal{L}_R = \|\hat{q} - q\|_1$.

For field-of-view supervision, we supervise the inter-frame focal ratio rather than absolute focal lengths per frame. Defining $r(\phi) = \log \tan(\phi/2)$, the focal loss is

$$\mathcal{L}_F = \left\| \big(\hat{r}(\phi_2) - \hat{r}(\phi_1)\big) - \big(r(\phi_2) - r(\phi_1)\big) \right\|_1. \tag{5}$$

We set $\lambda_T = 1.0$, $\lambda_R = 1.0$, and $\lambda_F = 0.5$.

Optimization. We fine-tune only the LoRA parameters with AdamW and mixed-precision training, keeping the full VGGT backbone frozen. Training runs for 200 epochs on a single A100 64 GB GPU and completes in approximately one day.

4. Experiments

4.1. Experimental Setup

Evaluation Benchmark. We evaluate on the BIWI Kinect Head Pose Database [10], which provides ground-truth 6DoF head poses for 24 subjects recorded with an RGB-D sensor. Since BIWI annotations are defined in the depth-camera coordinate frame, we transform all poses to the RGB-camera frame using the per-subject calibration parameters before evaluation. Face detection is performed with a shared MTCNN [42] detector for all methods, following the evaluation protocol of [13].

Baselines. We compare against multiple head pose estimation methods evaluated on BIWI, spanning both rotation-only and 6DoF approaches: HopeNet [29], FSA-Net [39], WHENet-V [44], 6DRepNet [13], TriNet [5], To-

Table 1. Cross-domain evaluation on BIWI. Rotation errors in degrees (lower is better). Best in **bold**, second-best underlined. Top: numbers reported in original papers (from [6]). Bottom: reproduced under our shared MTCNN detection protocol. "Data" indicates the training-data regime: synthetic-only (S), real-only (R), or mixed (M; real images combined with synthetic augmentation / synthetic pose expansion, e.g., 300W-LP).

Method	Yaw ↓	Pitch ↓	Roll ↓	MAE ↓	Data
Reported numbers					
Dlib [19]	11.86	13.00	19.56	14.81	R
3DDFA [45]	5.50	41.90	13.22	19.07	M
EVA-GCN [38]	4.01	4.78	2.98	3.92	M
HopeNet [29]	4.81	6.61	3.27	4.89	M
QuatNet [14]	4.01	5.49	2.94	4.15	M
Liu *et al*. [25]	4.12	5.61	3.15	4.29	M
FSA-Net [39]	4.27	4.96	2.76	4.00	M
HPE [16]	4.57	5.18	3.12	4.29	M
WHENet-V [44]	3.60	4.10	2.73	3.48	M
RetinaFace [8]	4.07	6.42	2.97	4.49	R
FDN [41]	4.52	4.70	2.56	3.93	M
MNN [32]	3.98	4.61	2.39	3.66	M
TriNet [5]	3.05	4.76	4.11	3.97	M
6DRepNet [13]	3.24	4.48	2.68	3.47	M
Cao *et al*. [4]	4.21	3.52	3.10	3.61	M
TokenHPE [40]	3.95	4.51	2.71	3.72	M
Cobo *et al*. [7]	4.58	4.65	2.71	3.98	M
img2pose [2]	4.57	3.55	3.24	3.79	M
PerspNet [18]	3.10	3.37	2.38	2.95	R
TRG [6]	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 [13]	3.74	4.95	**3.04**	3.91	M
TokenHPE-v1 [40]	5.57	6.23	3.79	5.20	M
TRG [6]	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**

kenHPE [40], img2pose [2], PerspNet [18], and TRG [6]. All baselines are trained on real-world data or mixed regimes (e.g., real images expanded via synthetic pose profiling like 300W-LP [45], ARKitFace [18], or combinations thereof).

Evaluation Protocol. We report mean absolute error (MAE) in degrees for yaw, pitch, and roll, along with the overall MAE averaged across the three axes. For the relative model (VGGT-HPE), inference proceeds in two-view mode: an anchor frame with known ground-truth pose is paired with each target frame, the model predicts the relative rigid transformation, and the final target pose is recovered by composing the prediction with the anchor pose.

4.2. Cross-Domain Evaluation on BIWI

Table 1 presents the main cross-domain evaluation, split into two sections. The top section collects numbers reported in the original publications (following the compilation of [6]); however, these results are not directly comparable, as each method uses its own face detector, crop strategy, and evaluation subset. To enable a fair comparison, the bottom section reproduces all methods under a shared evaluation protocol using MTCNN [42] detection, following the widely adopted setup of 6DRepNet [13]. For TokenHPE, only the v1 checkpoint is publicly available, so we report results with that model. For TRG, we observed that the authors evaluate only on the subset of BIWI frames for which their preprocessing pipeline (based on FAN and MTCNN) successfully detects a face; since this filtering is not documented, we re-ran TRG under our shared protocol for consistency. For VGGT-HPE (Rel.), we use as a fixed anchor throughout each BIWI subject sequence the first frame of each recording, requiring essentially *only a single ground-truth pose per subject* rather than per-frame annotations.

In the reported-numbers section, VGGT-HPE (Rel.) achieves the lowest yaw ($2.24°$) and pitch ($3.04°$) errors among all methods, and the second-lowest MAE ($2.82°$), behind only TRG ($2.75°$), while being the only method trained exclusively on synthetic data. Under the controlled shared protocol, VGGT-HPE (Rel.) achieves the lowest MAE ($2.82°$) among the baselines.

The key takeaway emerges from comparing our two variants: VGGT-HPE-Abs, which shares the same backbone and synthetic training data but operates in absolute single-image mode, reaches only $5.15°$, while switching to the relative formulation yields the best result in the table. This confirms that the advantage is structural rather than architectural. Absolute methods must internalize a canonical reference frame from the training data, making them sensitive to domain shift. The relative formulation sidesteps this by predicting a displacement between two explicitly provided views, removing the dependence on an implicit anchor.

In Fig. 4 we also show qualitative results among various methods on the BIWI benchmark.

Table 2. BIWI hard benchmark (360 neutral-anchor / extreme-query pairs). Rotation errors in degrees (lower is better). Best in **bold**, second-best underlined. "Data" indicates the training regime (S: synthetic-only, M: mixed) as defined in Table 1.

Method	Yaw ↓	Pitch ↓	Roll ↓	MAE ↓	Data
VGGT-HPE-Abs	40.74	18.65	33.66	31.02	**S**
TokenHPE [40]	21.85	26.35	19.34	22.51	M
TRG [6]	8.95	33.88	8.87	17.23	M
6DRepNet [13]	14.27	18.91	**6.81**	13.33	M
VGGT-HPE (Rel., ours)	**3.81**	**15.87**	6.93	**8.87**	**S**

4.3. Controlled Anchor Benchmarks: Easy and Hard Pairs

To study how prediction difficulty scales with the anchor–target rotation gap, we construct two complementary bench-

Figure 4. Qualitative results on BIWI. Each row shows a different subject. From left to right: the query frame, the anchor frame with its known pose overlay, the ground-truth pose, our prediction (VGGT-HPE), and three baselines (6DRepNet, TokenHPE, TRG). Our method produces pose estimates that are visually closer to the ground truth across a range of head orientations, including challenging near-profile and upward-looking poses where absolute methods exhibit larger deviations.

marks from BIWI, each with 360 pairs per subject. In both cases the anchor is near-neutral, so the absolute baselines-which ignore the anchor entirely-are evaluated directly on the target frames. In the hard benchmark these targets are extreme poses; in the easy one they are near-frontal. This lets us study how the different absolute and relative models behave in both easy and hard (extreme) poses.

Hard-pair benchmark. Each pair is formed by selecting a near-neutral anchor frame and an extreme-pose query frame, with a mean anchor–target rotation gap of approximately $70°$. Results are shown in Table 2. The performance gap between VGGT-HPE (Rel.) and the absolute baselines widens substantially under these conditions: while 6DRepNet, the strongest absolute baseline, achieves $13.33°$ MAE, VGGT-HPE (Rel.) reaches $8.87°$, a 33% relative improvement. The absolute variant of our model collapses entirely to $31.02°$, confirming once more that the benefits stem from the relative formulation rather than the backbone.

Easy-pair benchmark. Here the situation is reversed: anchor and target are both near-neutral, with a mean rotation gap of only $3.82°$. In this regime the absolute baselines are evaluated on their easiest samples — target poses lie close to the implicit canonical frame that these methods internalize during training. Results are shown in Table 3. Even in this favorable setting for absolute methods, VGGT-HPE (Rel.) achieves $0.96°$ MAE, nearly three times better than 6DRepNet ($2.80°$) and the absolute VGGT variant ($3.98°$). This result highlights a key practical advantage of the relative formulation: when the anchor–target gap is small, the prediction problem becomes almost trivially easy.

4.4. Error Analysis: Relative Gap vs. Absolute Pose

The easy- and hard-pair benchmarks test two ends of the difficulty range. Here we look at the full picture by binning results in $5°$ increments along two axes: the anchor–query rotation gap and the absolute query orientation.

Table 3. BIWI easy neutral-anchor benchmark (360 pairs; pair delta mean: 3.82°). Rotation errors in degrees (lower is better). Best in **bold**, second-best underlined. "Data" indicates the training regime (S: synthetic-only, M: mixed).

Method	Yaw ↓	Pitch ↓	Roll ↓	MAE ↓	Data
VGGT-HPE-Abs	5.62	2.26	4.06	3.98	**S**
TokenHPE [40]	3.41	5.99	1.43	3.61	M
TRG [6]	3.59	4.24	2.52	3.45	M
6DRepNet [13]	2.31	4.93	1.14	2.80	M
VGGT-HPE (Rel., ours)	**1.17**	**0.74**	**0.97**	**0.96**	**S**

Table 4. Ablation study on synthetic validation data and BIWI (cross-domain). Rotation errors in degrees, sorted by decreasing MAE on BIWI. Best in **bold**, second-best underlined. The absolute single-image variant (Abs. Single) fits the synthetic distribution best but transfers worst to BIWI, while the full relative model (VGGT-HPE) reverses this ranking — evidence that the relative formulation transfers better across domains.

	Synthetic				BIWI			
Variant	Yaw ↓	Pitch ↓	Roll ↓	MAE ↓	Yaw ↓	Pitch ↓	Roll ↓	MAE ↓
Adaptation strategy								
Full finetune	38.00	33.76	31.12	34.29	23.07	17.90	10.05	17.00
From scratch	9.21	15.07	14.05	12.78	7.71	8.12	6.94	7.59
Head-only	3.82	6.50	5.89	5.40	18.08	15.17	8.25	13.83
LoRA (ours)	**2.46**	**4.65**	**4.51**	**3.87**	**2.24**	**3.04**	**3.17**	**2.82**
Loss and formulation variants (all LoRA)								
Small-Gap	17.03	23.20	21.77	20.66	7.38	7.07	3.88	6.11
Abs. Pair	3.12	4.59	7.01	4.91	3.61	6.99	6.49	5.69
Abs. Single	2.44	4.03	**3.24**	**3.24**	4.90	7.01	3.53	5.15
T-Aux, No FoV	2.94	5.45	6.01	4.80	2.94	3.91	3.90	3.59
No FoV	2.39	4.13	4.19	3.57	2.64	4.05	3.61	3.43
T-Aux	2.43	4.37	4.48	3.76	2.76	3.10	3.90	3.25
Geo Loss	2.28	**3.99**	3.76	3.34	2.95	3.24	3.32	3.17
Rot.-Only	2.40	4.31	4.06	3.59	2.58	3.48	3.32	3.12
Baseline	**2.19**	4.15	3.96	3.43	2.61	3.17	3.31	3.03
VGGT-HPE	2.46	4.65	4.51	3.87	**2.24**	**3.04**	**3.17**	**2.82**

Scaling with the Anchor–Query Gap. In Figure 5 we bin anchor–target pairs by their ground-truth rotation gap in 5° increments and plot the per-bin MAE for each method. Since for this test the anchor is always near-neutral, the gap directly reflects how far the query deviates from frontal. All methods exhibit increasing error as the gap grows, but VGGT-HPE (Rel.) starts from a lower baseline and rises more gradually than every absolute baseline, including our own absolute variant. This directly validates the hypothesis that relative head pose prediction is fundamentally easier, and that the advantage of the relative formulation grows with the magnitude of the displacement being estimated.

Robustness to Extreme Poses via Close Anchors. Absolute methods have no way to adapt at test time — they always regress from a single crop relative to a fixed canonical frame learned during training, and performance drops as the target moves further from that reference. The relative formulation does not have this limitation: the anchor can be chosen to keep the prediction gap small, regardless of how

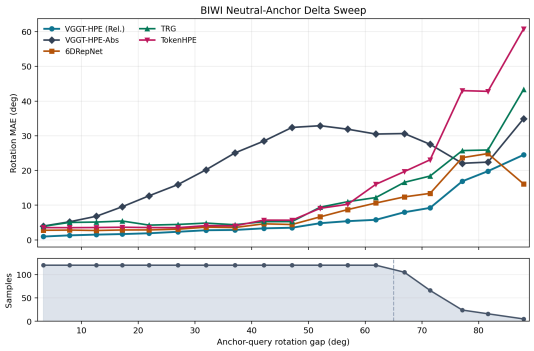

Figure 5. BIWI neutral-anchor evaluation as a function of anchor-query rotation gap. The upper plot reports rotation MAE, while the lower band shows the number of sampled pairs per bin. VGGT-HPE remains the strongest method across the full range.

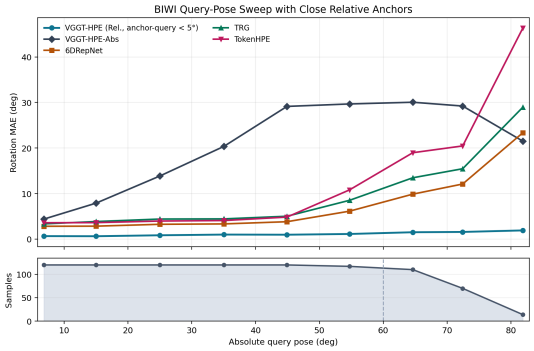

Figure 6. BIWI query-pose evaluation as a function of absolute query pose. For VGGT-HPE, each query is paired with a same-subject anchor whose anchor-query geodesic gap is below 5°. The upper plot reports rotation MAE, while the lower band shows the number of sampled pairs per bin.

extreme the target pose is. Figure 6 tests this directly. We sweep over absolute query pose but pair each query with a same-subject anchor within 5° geodesic distance. The absolute baselines degrade beyond 60°, while VGGT-HPE (Rel.) stays below 2° MAE across the full range — the curve is nearly flat. This shows that what matters for relative prediction is the anchor–query gap, not the absolute pose of either view. With a close enough anchor, extreme poses are no harder than frontal ones.

4.5. Ablation Studies

Table 4 presents ablation results on both the synthetic validation set and BIWI. The top section compares adaptation strategies for the VGGT backbone. Full fine-tuning destroys the pretrained representations and performs worst on both datasets. Training from scratch (i.e., without pretrained weights) does better but still lags far behind LoRA on BIWI (7.59° vs. 2.82°). Head-only tuning is an interesting case: it achieves reasonable synthetic error (5.40°)

Table 5. Relaxing the ground-truth anchor assumption on the full BIWI set. The anchor pose is provided by an external absolute model instead of ground truth. Rotation errors in degrees (lower is better).

Method	Yaw ↓	Pitch ↓	Roll ↓	MAE ↓
VGGT-HPE (Rel., GT anchor)	**2.24**	**3.04**	**3.17**	**2.82**
VGGT-HPE (Rel., VGGT-HPE-Abs anchor)	4.56	5.28	3.42	4.42
VGGT-HPE (Rel., 6DRepNet anchor)	2.89	6.05	3.29	4.08
VGGT-HPE (Rel., TokenHPE anchor)	3.35	7.30	3.91	4.85
VGGT-HPE (Rel., TRG anchor)	4.02	12.03	7.13	7.73

Table 6. Relaxing the ground-truth anchor assumption on the hard BIWI subset (360 neutral-anchor / extreme-query pairs). The anchor pose is provided by an external absolute model instead of ground truth. Rotation errors in degrees (lower is better).

Method	Yaw ↓	Pitch ↓	Roll ↓	MAE ↓
VGGT-HPE (Rel., GT anchor)	**3.81**	**15.87**	6.93	**8.87**
VGGT-HPE (Rel., VGGT-HPE-Abs anchor)	6.93	17.36	8.68	10.99
VGGT-HPE (Rel., 6DRepNet anchor)	4.81	19.36	**6.62**	10.26
VGGT-HPE (Rel., TokenHPE anchor)	5.92	20.31	7.44	11.22
VGGT-HPE (Rel., TRG anchor)	7.07	21.82	17.00	15.30

but collapses on BIWI (13.83°), suggesting it overfits to the synthetic pose distribution without learning transferable features. LoRA strikes the best balance, preserving the pretrained geometric priors while adapting to the facial domain.

The bottom section ablates loss and formulation variants. The absolute single-image variant achieves the best synthetic MAE (3.24°) but degrades to 5.15° on BIWI, while VGGT-HPE reaches 2.82° — a clear sign that the relative formulation transfers better across domains. The Small-Gap variant performs worst on both datasets, confirming that sufficient anchor–query displacement is needed during training. Among the loss components, removing field-of-view prediction (No FoV; 3.43°) and translation supervision (T-Aux, No FoV; 3.59°) each hurt cross-domain performance. The geodesic loss and rotation-only variants are competitive, but the full combination in VGGT-HPE yields the best cross-domain result.

4.6. Relaxing the Ground-Truth Anchor Assumption

A practical limitation of the relative formulation is that it requires a known anchor pose at test time. To understand how sensitive our method is to the anchor pose quality, we replace the ground-truth anchor pose with the prediction of an external absolute model. For each BIWI anchor–query pair, we estimate the anchor pose using one of VGGT-HPE-Abs, 6DRepNet, TokenHPE, or TRG, and use that prediction as the reference for VGGT-HPE (Rel.). The model then predicts the relative transformation with respect to the predicted anchor and composes it to recover an absolute query pose. Since any error in the anchor estimate propagates directly into the composed query pose as a constant offset, this setting is expected to increase the overall error proportionally to the accuracy of the anchor model. Results on the full BIWI set and the hard subset are shown in Tables 5 and 6, respectively.

Performance naturally degrades compared to using the ground-truth anchor, but the drop is moderate when the anchor estimator is reasonably accurate. With a 6DRepNet anchor, for instance, VGGT-HPE (Rel.) reaches 4.08° on the full set — worse than with ground truth (2.82°), but still competitive with the absolute baselines in Table 1. On the hard subset the same trend holds: the relative pipeline with a 6DRepNet anchor (10.26°) still outperforms every standalone absolute method except 6DRepNet itself. These results suggest that the ground-truth anchor requirement is not a hard constraint — a reasonable absolute estimator can provide a sufficient reference frame, and the relative formulation remains beneficial even when the anchor is imperfect.

5. Limitations

The relative formulation requires a known anchor pose at test time. As shown in Section 4.6, this requirement can be partially relaxed by using the prediction of an absolute model as the anchor, but performance degrades with anchor quality. In sequential settings—such as video conferencing, sign language recognition, and driver monitoring—prior frames naturally serve as anchors, though using the model's own predictions auto-regressively can lead to error accumulation and drift over time. The flat error curve in Figure 6 relies on access to geometrically close ground-truth anchors; deploying the method in continuous temporal settings would require periodic recalibration, multi-anchor consensus, or external stabilization to prevent drift.

6. Conclusions

We presented VGGT-HPE, a relative head pose estimator that reframes the problem as relative pose prediction between two views. Built on the VGGT geometry foundation model, our method achieves the best results on BIWI among all compared methods, despite training only on synthetic data, while the same backbone in absolute mode ranks among the weakest. This contrast alone highlights the power of the relative formulation. Our controlled easy-pair and hard-pair benchmarks show that the advantage holds across the full difficulty spectrum and grows as the anchor–target gap increases. The error-vs-gap analysis directly validates the core thesis of this work: relative prediction is fundamentally easier than absolute regression. Moreover, unlike absolute methods whose implicit reference frame is fixed at training time, the relative formulation offers the flexibility to choose the anchor at test time, adapting the difficulty of the prediction to the application at hand.

Acknowledgements

The work of V. Vasileiou and K. Daniilidis has been partially supported by project MIS 5154714 of the National Recovery and Resilience Plan Greece 2.0, funded by the European Union under the NextGenerationEU Program. The work of V. Vasileiou, P. P. Filntisis, and P. Maragos has also been partially funded by the European Union under Horizon Europe (grant No. 101136568 – HERON). We acknowledge EuroHPC Joint Undertaking for awarding us access to Leonardo at CINECA, Italy, under project EHPC-DEV-2026D01-089.

References

[1] Andrea F. Abate, Carmen Bisogni, Arcangelo Castiglione, and Michele Nappi. Head pose estimation: An extensive survey on recent techniques and applications. *Pattern Recognition*, 127:108591, 2022. 1

[2] Vítor Albiero, Xingyu Chen, Xi Yin, Guan Pang, and Tal Hassner. img2pose: Face alignment and detection via 6DoF, face pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 5

[3] Blender Online Community. Blender – a 3D modelling and rendering package. Blender Foundation, 2018. 4

[4] Zhiwen Cao, Dongfang Liu, Qijun Wang, and Yingjie Chen. Towards unbiased label distribution learning for facial pose estimation using anisotropic spherical Gaussian. In *ECCV*. 5

[5] Zhiwen Cao, Zongcheng Chu, Dongfang Liu, and Yingjie Chen. A vector-based representation to enhance head pose estimation. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2021. 4, 5

[6] Sungho Chun and Ju Yong Chang. 6DoF head pose estimation through explicit bidirectional interaction with face geometry. In *European Conference on Computer Vision (ECCV)*, 2024. 2, 5, 7

[7] Alejandro Cobo, Roberto Valle, José M Buenaposada, and Luis Baumela. On the representation and methodology for wide and short range head pose estimation. *Pattern Recognition*, 149:110263, 2024. 5

[8] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. RetinaFace: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 5

[9] Jakob Engel, Thomas Schöps, and Daniel Cremers. Lsd-slam: Large-scale direct monocular slam. In *European conference on computer vision*, pages 834–849. Springer, 2014. 2

[10] Gabriele Fanelli, Matthias Dantone, Juergen Gall, Andrea Fossati, and Luc Van Gool. Random forests for real time 3D face analysis. *International Journal of Computer Vision*, 101:437–458, 2013. 2, 4

[11] Panagiotis P. Filntisis, George Retsinas, Radek Daneček, Vanessa Sklyarova, Petros Maragos, and Timo Bolkart. MOCHI: Registration-free learnable multi-view capture of faces in dense semantic correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. 4

[12] Jinwei Gu, Xiaodong Yang, Shalini De Mello, and Jan Kautz. Dynamic facial analysis: From bayesian filtering to recurrent neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1548–1557, 2017. 2

[13] Thorsten Hempel, Ahmed A. Abdelrahman, and Ayoub Al-Hamadi. 6D rotation representation for unconstrained head pose estimation. In *IEEE International Conference on Image Processing (ICIP)*, pages 2496–2500, 2022. 1, 2, 4, 5, 7

[14] Hao-Wei Hsu, Ting-Yang Wu, Shen Wan, Wing Hung Wong, and Chen-Yi Lee. QuatNet: Quaternion-based head pose estimation with multiregression loss. *IEEE Transactions on Multimedia*, 21(4):1035–1046, 2018. 5

[15] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shanen Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022. 2, 3

[16] Bin Huang, Renwen Chen, Wang Xu, and Qinbang Zhou. Improving head pose estimation using two-stage ensembles with top-k regression. page 103827, 2020. 5

[17] Sumit Jha and Carlos Busso. Estimation of driver's gaze region from head pose using a single RGB camera. *IEEE Transactions on Intelligent Transportation Systems*, 23(10): 17907–17918, 2022. 1

[18] Yueying Kao, Bowen Pan, Miao Xu, Jiangjing Lyu, Xiangyu Zhu, Yanbo Chang, Xiaobo Li, and Zhen Lei. Toward 3D face reconstruction in perspective projection: Estimating 6DoF face pose from monocular image. *IEEE Transactions on Image Processing*, 32:3080–3091, 2023. 5

[19] Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 5

[20] Oscar Koller. Quantitative survey of the state of the art in sign language recognition. *arXiv preprint arXiv:2008.09918*, 2020. 1

[21] Dimitrios Kollias and Stefanos Zafeiriou. Affect analysis in-the-wild: Valence-arousal, expressions, action units and a unified framework. *arXiv preprint arXiv:2103.15792*, 2021. 1

[22] Lukas Kuhn, Markus Müller, et al. Domain adaptation for head pose estimation using relative pose consistency. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2023. 2

[23] Vincent Leroy, Yohann Cabon, and Jérome Revaud. MASt3R: Matching and stereo 3d reconstruction. *arXiv preprint arXiv:2406.09756*, 2024. 3

[24] Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. In *ACM SIGGRAPH Asia*, 2017. 2, 4

[25] Zhiwen Liu, Zhihang Chen, Jing Bai, Shanshan Li, and Shiguo Lian. Facial pose estimation by deep learning from label distributions. In *IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2019. 5

[26] Eric Marchand, Hideaki Uchiyama, and Fabien Spindler. Pose estimation for augmented reality: A hands-on survey. *IEEE Transactions on Visualization and Computer Graphics*, 22(12):2633–2651, 2016. 1

[27] Nikolaos Mavridis. A review of verbal and non-verbal human–robot interactive communication. *Robotics and Autonomous Systems*, 63:22–35, 2015. 1

[28] Erik Murphy-Chutorian and Mohan Manubhai Trivedi. Head pose estimation in computer vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(4): 607–626, 2009. 1

[29] Nataniel Ruiz, Eunji Chong, and James M. Rehg. Fine-grained head pose estimation without keypoints. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018. 1, 2, 4, 5

[30] Vanessa Sklyarova, Egor Zakharov, Otmar Hilliges, Michael J. Black, and Justus Thies. Text-conditioned generative model of 3D strand-based human hairstyles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4703–4712, 2024. 4

[31] Zachary Teed and Jia Deng. DROID-SLAM: Deep visual SLAM for monocular, stereo, and RGB-D cameras. In *Advances in Neural Information Processing Systems*, pages 7166–7177, 2021. 2

[32] Roberto Valle, José M. Buenaposada, and Luis Baumela. Multi-task head pose estimation in-the-wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(8): 2874–2881, 2020. 5

[33] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5294–5306, 2025. 2, 3, 4

[34] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jérome Revaud. DUSt3R: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20709, 2024. 3

[35] Wenshan Wang, Yaoyu Zhu, Xin Wang, Yuwei Zeng, and Mingyu Ding. Tartanvo: A generalizable learning-based vo. In *Conference on Robot Learning*, pages 1761–1772. PMLR, 2021. 2

[36] Xinyu Wang et al. Temporal modeling and structure aggregation for video head pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 2

[37] Philippe Weinzaepfel, Vincent Leroy, Thomas Lucas, Romain Brégier, Yohann Cabon, Vaibhav Arora, Leonid Antsfeld, Boris Chidlovskii, Grégory Rogez, and Jérome Revaud. CroCo: Cross-view completion for 3d vision. In *Advances in Neural Information Processing Systems*, pages 17424–17438, 2022. 3

[38] Miao Xin, Shuangtao Mo, and Yaoyang Lin. EVA-GCN: Head pose estimation based on graph convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 5

[39] Tsun-Yi Yang, Yi-Ting Chen, Yin-Yu Lin, and Yung-Yu Chuang. FSA-Net: Learning fine-grained structure aggregation for head pose estimation from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1087–1096, 2019. 4, 5

[40] Cheng Zhang, Hai Liu, Yongjian Deng, Bochen Xie, and Youfu Li. TokenHPE: Learning orientation tokens for efficient head pose estimation via transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8897–8906, 2023. 1, 2, 5, 7

[41] Hao Zhang, Mingyi Wang, Yonggenui Liu, and Yi Yuan. FDN: Feature decoupling network for head pose estimation. In *AAAI Conference on Artificial Intelligence*, 2020. 5

[42] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23 (10):1499–1503, 2016. 4, 5

[43] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. MPIIGaze: Real-world dataset and deep appearance-based gaze estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(1):162–175, 2019. 1

[44] Yijun Zhou and James Gregson. WHENet: Real-time fine-grained estimation for wide range head pose. *arXiv preprint arXiv:2005.10353*, 2020. 1, 2, 4, 5

[45] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z. Li. Face alignment across large poses: A 3D solution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 146–155, 2016. 2, 5