ReVA-HPE: Relative 6DoF Head Pose Estimation with Synthetic Pose Supervision and Personalized Virtual Anchors
Supplementary Material

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

[1]HERON - Hellenic Robotics Center of Excellence, Athena Research Center, Greece
[2] National Technical University of Athens, School of Electrical and Computer Engineering, Greece
[3]Robotics Institute, Athena Research Center, Greece [4]University of Pennsylvania, USA
v.vasileiou@athenarc.gr

This supplementary material provides additional training and implementation details, provenance of the pretrained system components, held-out synthetic calibration of both robust-estimation stages, virtual-anchor and VA-fast details, detector controls, controlled enrollment, illumination, expression, and upstream-error analyses, runtime accounting, and additional qualitative results for ReVA-HPE.

1. Synthetic Training Data

The task-specific relative 6DoF estimator is trained exclusively on procedural FLAME renderings. The corpus contains 25,000 base images spanning identities with and without synthetic hair, varied skin textures, expressions, illumination, and viewpoints. No real face image or real head-pose annotation is used for this task-specific adaptation, nor are pseudo-labels, target-domain adaptation, or evaluation-set fine-tuning.

Figure 1 illustrates the preprocessing and augmentation pipeline. Starting from a raw FLAME render, we extract a square face crop and apply appearance-level augmentations before feeding the image pair to the relative 6DoF estimator. These augmentations modify color, contrast, blur, and noise while preserving the camera-frame pose labels and the analytically adjusted intrinsics.

Training uses a fixed one-million-pair manifest. Pair members share identity and illumination, while expression and viewpoint may differ. Figure 2 illustrates this variation across identities and views. This construction encourages the model to learn the relative geometric change between the two observations rather than memorizing an absolute canonical pose or identity-specific appearance.

2. Additional Implementation Details

Table 1 summarizes the main implementation constants used throughout the experiments. All values are fixed globally and are not tuned per benchmark.

Table 1. **Implementation details.** All settings are fixed globally for the reported experiments.

Component	Setting
Backbone	VGGT camera branch
Frozen modules	VGGT backbone, non-camera heads disabled
Trainable modules	LoRA adapters only
LoRA rank / scale	rank 8 / scale 16
Relative output	$(\Delta R, \delta_x, \delta_y, \delta_z)$
Translation target	$\delta_x = (x_q - x_a)/z_a,\ \delta_y = (y_q - y_a)/z_a,\ \delta_z = \log(z_q/z_a)$
Synthetic base images	25,000 FLAME renders
Pair sampling	fixed 1M anchor-query pair list
Crop protocol	unified pair crop, margin 0.5
Viewing directions	$D = 20$
Radius scales	$S = \{0.85, 0.95, 1.00, 1.05, 1.15\}$
Direct-1r candidates	$K = D = 20$ per query
Direct-5r candidates	$K = D \times S = 100$ per query
VA-fast candidates	20 once at enrollment one prediction per query
Bank RANSAC thresholds	$\tau_R = 10^\circ,\ \tau_t = 50$ mm
Bank RANSAC min. inliers	3
EPnP-RANSAC reproj. threshold	5 px
EPnP-RANSAC iterations	300
EPnP-RANSAC confidence	0.999
EPnP-RANSAC min. inliers	6
EPnP refinement	iterative solvePnP on RANSAC inliers

2.1. Pretrained-component provenance

Synthetic pose supervision refers specifically to the task-specific adaptation of the relative pose estimator. The complete inference pipeline also uses separately pretrained frozen modules, summarized below.

3. Robust-Estimation Thresholds

ReVA-HPE uses two RANSAC procedures with different purposes. The avatar-bank RANSAC is a pose-consensus step over complete 6DoF hypotheses predicted from different virtual anchors. The enrollment EPnP-RANSAC instead performs classical robust 2D–3D fitting over facial-landmark correspondences. All thresholds are fixed globally and kept unchanged for the BIWI and FaMoS evaluations.

3.1. Avatar-Bank Pose-Consensus RANSAC

For avatar-bank consensus, each rendered anchor produces one absolute query-pose candidate. RANSAC is there-

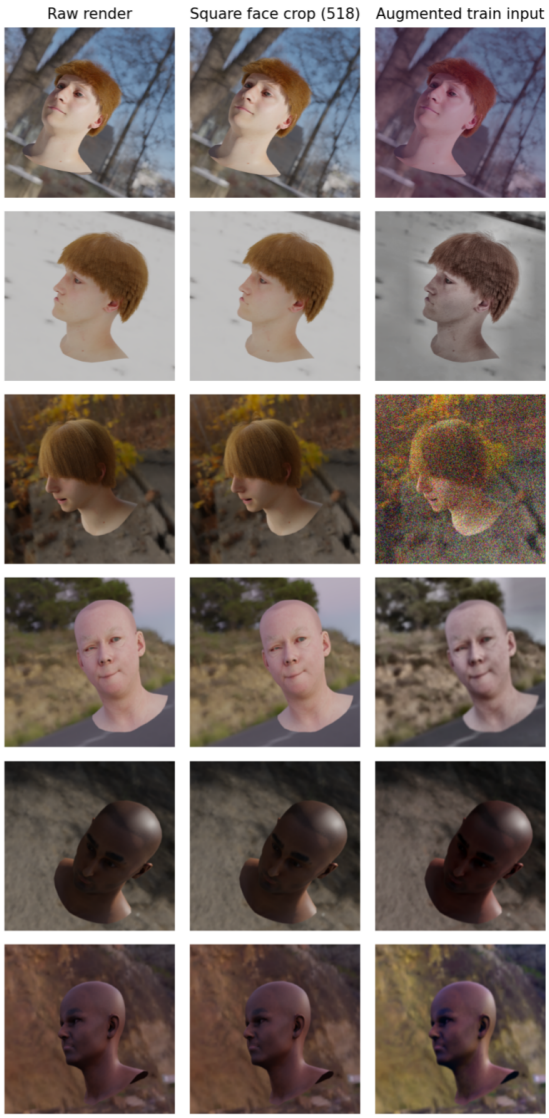


Figure 1. **Synthetic preprocessing and augmentation pipeline.** Each row shows one synthetic FLAME render before and after preprocessing: raw render, square face crop, and final augmented training input. Image-level augmentations modify appearance while preserving the camera-frame pose labels and crop-adjusted intrinsics used for supervision.

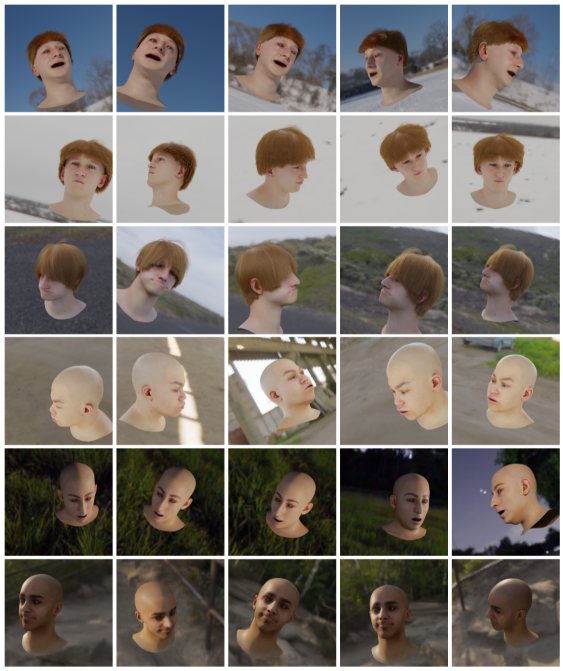


Figure 2. **Synthetic FLAME training samples.** Rows correspond to different synthetic identities and columns show different viewpoints for the same identity. The corpus includes synthetic identities with and without synthetic hair, varied textures, expressions, illumination, and pose. No real face image or real head-pose label is used for training.

Component	Provenance and role
Relative LoRA estimator	Adapted only with procedural FLAME pose supervision, no real pose labels, pseudo-labels, or target-domain fine-tuning.
VGGT-1B	Released general-purpose visual-geometry checkpoint, frozen except for the LoRA adapters.
FaceLift	Released pretrained single-image avatar model, used frozen for enrollment reconstruction.
MTCNN / RetinaFace	Released pretrained face detectors used for crop generation and detector sensitivity analysis.
FAN-style landmarks	Released pretrained 2D facial-landmark detector used for enrollment PnP and multiview alignment.
FLAME	Parametric canonical face geometry used for landmark correspondences and avatar alignment.
EPnP-RANSAC	Nonlearned geometric solver used to obtain the enrollment metric gauge.

Table 2. **Provenance of the modular inference components.** Only the task-specific relative pose adaptation uses the procedural head-pose training corpus the complete system is not an end-to-end synthetic-only pipeline.

fore not used to estimate pose from raw image correspondences, but to select a locally consistent cluster of anchor-conditioned 6DoF predictions. We use a rotation threshold $\tau_R = 10^\circ$, a translation threshold $\tau_t = 50$ mm, and require at least three inliers.

The rotation threshold is intentionally wider than the final expected pose error. Its role is to decide whether two anchor-conditioned predictions belong to the same local pose mode, rather than to define the final accuracy. A very tight gate can fragment the correct cluster under avatar-to-image mismatch, while a very loose gate can admit hypotheses from incompatible pose modes. The translation threshold is used only as an internal consistency check during bank consensus: the final translation of practical ReVA-HPE is obtained from the enrollment-PnP relative branch, not from the avatar bank.

Synthetic sensitivity analysis. We select τ_R and τ_t using a

held-out corrupted synthetic validation split. We evaluate 25 queries from five held-out synthetic identities, with 20 single-radius anchor-conditioned hypotheses per query. To stress the consensus stage, we deterministically corrupt 30% of the hypotheses using a 35° random-axis rotation and Gaussian translation noise with standard deviation 120 mm. No BIWI or FaMoS image, pose annotation, or performance result is used for this calibration.

For the criterion-count calibration in Fig. 3, we evaluate $\tau_R \in \{5°, 10°, 15°, 20°\}$ and $\tau_t \in \{25, 50, 75, 100\}$ mm, with the minimum inlier count fixed to three. Since the inlier test is joint in rotation and translation, we score each threshold pair using three predefined criteria: (i) rotation MAE within $0.12°$ of the validation minimum, (ii) internal translation L_2 within 0.05 mm of its minimum, and (iii) mean consensus support of at least 99% of its maximum. A threshold pair receives one point for each satisfied criterion.

The strict settings $\tau_R = 5°$ or $\tau_t = 25$ mm satisfy only one criterion: they either reject too many useful hypotheses or degrade translation consistency. The first threshold pair satisfying all three criteria on this coarse calibration grid is $\tau_R = 10°$, $\tau_t = 50$ mm. Larger thresholds also satisfy the criteria, but they are looser and provide no additional validation benefit. We therefore choose the most conservative valid operating point and keep it fixed for all real-data experiments. Additional neighboring settings show the same qualitative plateau, confirming that the selected point is not a narrow numerical optimum.

Minimum number of bank inliers. We require at least three inliers in the avatar-bank consensus. Three hypotheses form the smallest non-pairwise consensus cluster: this rejects isolated predictions and accidentally agreeing pairs while remaining permissive when only a local subset of virtual anchors matches a difficult query. The calibration above uses 20 single-radius hypotheses per query. The same thresholds are then reused without modification for Direct-5r, which evaluates 20 viewing directions at five radii and therefore produces 100 candidates per query.

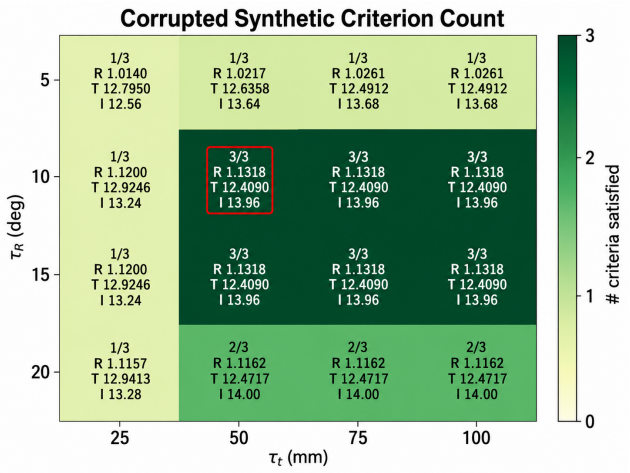


Figure 3. **Synthetic calibration of avatar-bank pose-consensus RANSAC.** We sweep the rotation threshold τ_R and translation threshold τ_t on held-out synthetic data after corrupting 30% of the anchor-conditioned hypotheses. Each cell reports how many of the three predefined validation criteria are satisfied. The outlined $10°$/50 mm setting is the first conservative operating point satisfying all three criteria. Stricter settings reject useful hypotheses or reduce translation consistency, whereas looser settings provide no additional validation benefit. No BIWI or FaMoS result is used for threshold selection.

3.2. Enrollment EPnP-RANSAC

The enrollment PnP stage is a separate robust-estimation problem. It is executed once per subject to obtain the metric translation gauge of the real enrollment image from 68 detected 2D landmarks and their corresponding canonical mean-neutral FLAME landmarks. We use EPnP inside RANSAC with 300 iterations, confidence 0.999, a 5-pixel reprojection threshold, and a minimum of six inliers, followed by iterative `solvePnP` refinement over the RANSAC inlier set.

Synthetic reprojection-threshold sensitivity. We evaluate the reprojection threshold using 2,000 synthetic views with known camera poses and intrinsics. We project the 68 canonical mean-neutral FLAME landmarks, add Gaussian localization noise with standard deviation 2 pixels, and replace 25% of the landmarks with gross perturbations having standard deviation 35 pixels. We sweep reprojection thresholds in $\{2, 3, 5, 7, 10\}$ pixels. This experiment uses only synthetic projections and does not use BIWI or FaMoS images, pose annotations, or performance results.

Very tight thresholds of 2–3 pixels retain substantially fewer valid correspondences and produce less stable pose estimates. At 5 pixels, the mean inlier support increases substantially while both reprojection and rotation errors have already entered a stable regime. Thresholds of 7–10 pixels are more permissive and may admit additional contour, expression, or landmark-detection outliers, without changing the qualitative behavior. We therefore choose 5 pixels as the most conservative value on the stable plateau.

The 300-iteration budget and confidence 0.999 provide robust sampling while adding little practical cost because enrollment PnP is executed only once per subject. Requiring six inliers places the acceptance threshold above the minimal PnP support, reducing the risk of accepting a solution from a small accidental subset while remaining permissive when some contour or self-occluded landmarks are unreliable. The final iterative refinement then optimizes the pose over the selected inlier set.

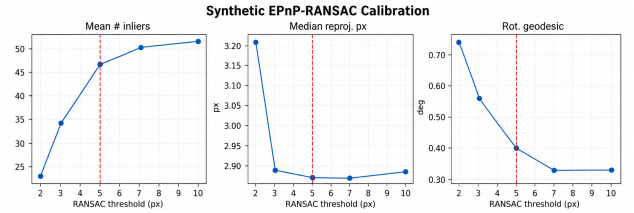


Figure 4. **Synthetic calibration of the enrollment EPnP-RANSAC reprojection threshold.** We sweep the reprojection threshold under synthetic landmark corruption. Very strict 2–3 pixel thresholds retain substantially fewer correspondences and produce less stable pose estimates. The 5-pixel setting, indicated by the dashed line, is chosen as the conservative operating point at the beginning of the stable region. The selected setting is fixed for all real-data evaluations, no BIWI or FaMoS result is used for calibration.

4. Virtual-Anchor Bank and Inference Modes

4.1. Bank construction

Here, D corresponds to the bank-size parameter M used in the main-paper ablation, while K denotes the total number of anchor-conditioned candidates after radius sampling. For each subject, we reconstruct a personalized avatar from one unlabeled enrollment image. The virtual bank is parameterized by D viewing directions and S metric-radius hypotheses, yielding $K = D \times S$ anchor-conditioned candidates.

The $D = 20$ viewing directions are farthest-point samples of a twice-subdivided icosphere restricted to the front hemisphere, seeded at the frontal view. The base radius r_0 is a per-subject estimate of the enrollment head depth, computed from the camera focal length and the enrollment FLAME fit's weak-perspective scale, and anchors are rendered with the enrollment camera intrinsics at the native frame resolution.Direct-1r uses only r_0, giving $K = 20$ candidates. Direct-5r uses the five fixed multipliers $\{0.85, 0.95, 1.00, 1.05, 1.15\}$ of r_0, giving $K = 20 \times 5 = 100$ candidates per query. The same pose-consensus thresholds are used for the single- and multi-radius configurations.

Figure 5 shows the base-radius (r_0), 20-direction slice of the bank for BIWI and FaMoS. The complete Direct-5r bank augments each of these directions with the additional radius hypotheses.

4.2. Direct inference and VA-fast

Direct inference compares every query against the complete virtual bank and performs a new query-specific consensus. Direct-5r therefore requires 100 anchor-conditioned model predictions per query. Only the consensus rotation from these bank-conditioned hypotheses is retained in the final practical pose query translation is still obtained from the real enrollment–query relative prediction decoded from the enrollment EPnP translation.

VA-fast instead uses the 20-anchor single-radius bank once during enrollment. For each virtual anchor A_m, it predicts an absolute enrollment-rotation candidate,

$$(\widehat{\Delta R}_{e\leftarrow m}, \widehat{\delta}_{e\leftarrow m}) = f_\theta(A_m, I_e), \tag{1}$$

$$\widehat{R}_e^{(m)} = \widehat{\Delta R}_{e\leftarrow m} R_m. \tag{2}$$

Pose-space consensus produces the calibrated enrollment rotation $\widehat{R}_e^{\text{bank}}$, while EPnP supplies $\widehat{t}_e^{\text{PnP}}$.

Each subsequent query requires only one real-to-real relative prediction,

$$(\widehat{\Delta R}_{q\leftarrow e}, \widehat{\delta}_{q\leftarrow e}) = f_\theta(I_e, I_q), \tag{3}$$

$$\widehat{R}_q^{\text{fast}} = \widehat{\Delta R}_{q\leftarrow e} \widehat{R}_e^{\text{bank}}, \tag{4}$$

$$\widehat{t}_q^{\text{fast}} = \text{Decode}\left(\widehat{\delta}_{q\leftarrow e}, \widehat{t}_e^{\text{PnP}}\right). \tag{5}$$

Thus, Direct-5r uses 100 bank predictions per query, whereas VA-fast uses 20 bank predictions once at enrollment and one prediction per query. Runtime and deployment-time accounting are reported in Sec. 8.

5. Dataset and Metric Details

Enrollment protocol. For each subject, the first available RGB frame is used as the unlabeled enrollment image. This frame is used to reconstruct the personalized avatar, render the virtual-anchor bank, and compute the real-enrollment PnP translation anchor. The enrollment frame is removed from the query set before evaluation.

Coordinate frame. All predicted poses are expressed in the RGB camera coordinate frame. A pose is defined as $P = (R, t)$, where $R \in SO(3)$ maps canonical head coordinates to the camera frame and $t = [x, y, z]^\top$ locates the canonical head origin.

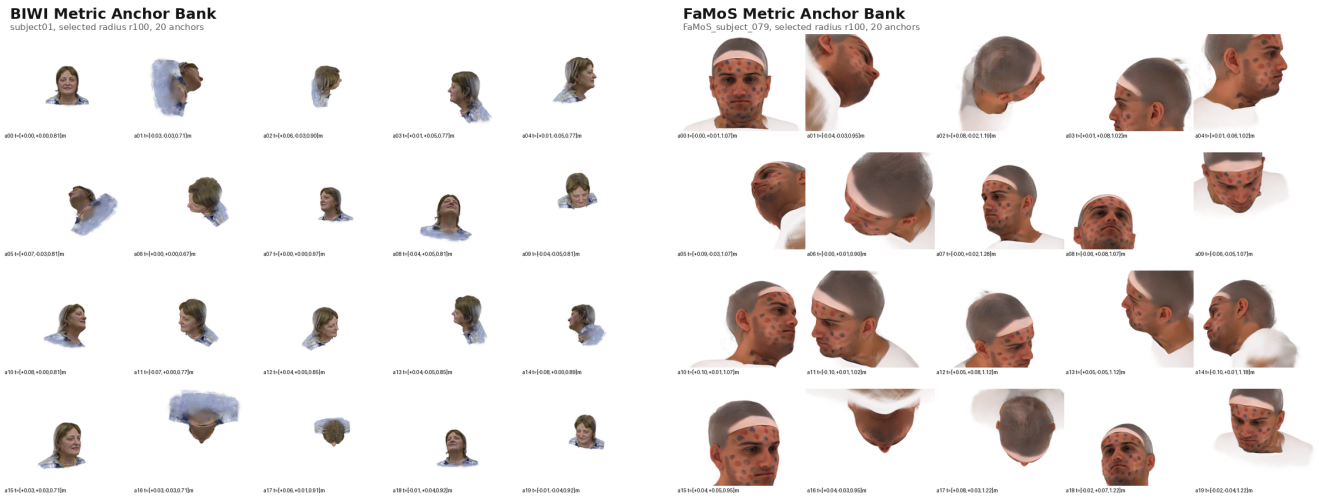
Rotation metrics. For orientation, we report yaw, pitch, and roll mean absolute error in degrees, as well as their average:

$$\text{MAE}_R = \frac{1}{3}\left(\text{MAE}_{yaw} + \text{MAE}_{pitch} + \text{MAE}_{roll}\right).$$

Translation metrics. For translation, we report per-axis mean absolute errors T_x, T_y, T_z in millimeters and their average,

$$\text{MAE}_t = \frac{1}{3}\left(T_x + T_y + T_z\right).$$

When comparing with prior work, we also report ADD by transforming the same canonical head landmarks with the



(a) BIWI example.

(b) FaMoS example.

Figure 5. **Examples of personalized metric virtual-anchor banks.** The visualization shows the base-radius r_0, 20-direction slice of the bank for one BIWI and one FaMoS subject. The selected Direct-5r protocol uses the same viewing directions at five radius hypotheses, yielding 100 candidates per query. The virtual poses are determined by the renderer and do not use dataset ground-truth head pose.

predicted and ground-truth poses and averaging the 3D point distance:

$$\mathrm{ADD} = \frac{1}{N|\mathcal{X}|} \sum_{i} \sum_{X \in \mathcal{X}} \left\| (\hat{R}_i X + \hat{t}_i) - (R_i X + t_i) \right\|_2 .$$

We additionally report the mean per-frame Euclidean translation error,

$$\text{Trans. } L_2 = \frac{1}{N} \sum_{i=1}^{N} \left\| \hat{\mathbf{t}}_i - \mathbf{t}_i \right\|_2 .$$

This is distinct from MAE_t, which remains the translation metric reported in the main-paper comparisons.

Oracle and diagnostic GT use. Rows marked with a dagger receive a ground-truth anchor pose and are non-deployable oracle references. Practical ReVA-HPE does not use dataset ground-truth pose at inference. In explicitly labelled diagnostic experiments, ground truth may be used only to construct an enrollment-pose bin or select an enrollment frame, the selected pose is never supplied to ReVA-HPE.

Ground-truth-use policy. Deployable ReVA-HPE uses one unlabeled enrollment RGB image, one query RGB image, calibrated intrinsics, detected enrollment landmarks, a FaceLift avatar reconstructed from the enrollment RGB, renderer-defined virtual-anchor poses, and frozen relative-model predictions. It does not use dataset ground-truth pose, depth maps, query ground truth, or test-error-based anchor selection during construction or inference. Dataset ground truth is used for final scoring. Experiments in which ground truth defines enrollment-pose bins, query-difficulty strata, or a camera-facing enrollment are explicitly labelled as non-deployable diagnostics.

6. Face Detector Sensitivity

The main experiments use MTCNN because it follows the established BIWI evaluation protocol of 6DRepNet and subsequent work [4–6]. To verify that this choice does not favor ReVA-HPE, we repeat the BIWI evaluation using RetinaFace [3].

From the fixed set of BIWI evaluation frames, MTCNN detects 14,980 (95.69%), while RetinaFace detects 15,514 (99.10%). Their exact intersection contains 14,980 frames. Table 3 evaluates all selected methods on this same intersection using crops produced by RetinaFace. Model checkpoints and all downstream settings remain unchanged.

ReVA-HPE remains the strongest method, achieving a rotation error of 2.54°, compared with 3.11° for Semi-UHPE and 3.31° for PnP-enrollment+Rel. Moreover, evaluating ReVA-HPE on the same frames using MTCNN crops gives 2.535°, compared with 2.538° using RetinaFace crops. Translation L_2 changes only from 37.26 to 37.70 mm. Thus, RetinaFace provides greater detection coverage, while the choice between MTCNN and RetinaFace crops has a negligible effect on ReVA-HPE accuracy.

FaMoS. Detection is considerably more difficult on FaMoS: RetinaFace detects 64.84% of the candidate anchor–query pairs, compared with 36.72% for MTCNN. We therefore repeat the detector-control comparison on the 170 pairs suc-

Method	Yaw ↓	Pitch ↓	Roll ↓	MAE$_r$ ↓	T_x ↓	T_y ↓	T_z ↓	MAE$_t$ ↓	Pose supervision
6DRepNet (AFLW2000) [4]	3.48	5.02	2.94	3.81	–	–	–	–	M
TokenHPE-v1 [6]	4.44	5.47	3.06	4.32	–	–	–	–	M
TRG [1]	4.95	7.65	4.24	5.61	14.15	14.69	143.98	57.61	M
VGGT-HPE-Abs [5]	4.73	5.88	3.59	4.73	75.13	61.64	109.11	81.96	**S**
SemiUHPE [7]	3.32	3.49	2.53	3.11	–	–	–	–	R
OPAL [2]	3.43	6.91	2.76	4.37	–	–	–	–	M
PnP-enrollment + Rel.	2.31	4.18	3.44	3.31	**9.51**	**11.51**	**30.78**	**17.27**	**S**
ReVA-HPE (ours)	**2.21**	**2.92**	**2.49**	**2.54**	**9.51**	**11.51**	**30.78**	**17.27**	**S**

Table 3. Detector-control evaluation on BIWI. All methods are evaluated on the exact-common set of $N = 14{,}980$ frames detected by both MTCNN and RetinaFace, using independently generated RetinaFace crops. Rotation errors are in degrees and translation errors are in millimeters. MAE$_t$ denotes the mean absolute translation error across the three translation axes. Pose supervision describes the data used for task-specific pose training: synthetic (S), real (R), or mixed (M).

Method	Yaw ↓	Pitch ↓	Roll ↓	MAE$_r$ ↓	T_x ↓	T_y ↓	T_z ↓	MAE$_t$ ↓	Pose supervision
6DRepNet (AFLW2000) [4]	2.18	6.09	5.01	4.42	–	–	–	–	M
TokenHPE-v1 [6]	2.83	4.70	4.24	3.92	–	–	–	–	M
TRG [1]	3.04	5.68	3.35	4.03	**5.68**	**4.38**	99.58	36.55	M
VGGT-HPE-Abs [5]	2.57	10.22	2.88	5.22	22.75	7.68	85.97	38.80	**S**
SemiUHPE [7]	1.64	**3.06**	**2.49**	**2.40**	–	–	–	–	R
OPAL [2]	2.27	4.14	3.46	3.29	–	–	–	–	M
PnP-enrollment + Rel.	1.47	6.37	3.39	3.74	10.80	5.68	**60.16**	**25.55**	**S**
ReVA-HPE (ours)	**1.25**	6.52	3.35	3.71	10.80	5.68	**60.16**	**25.55**	**S**

Table 4. Results on the exact-common FaMoS subset using RetinaFace crops. All methods are evaluated on the same 170 anchor–query pairs successfully detected by both MTCNN and RetinaFace. Rotation errors are in degrees and translation errors are in millimeters. MAE$_t$ denotes the mean absolute translation error across the three translation axes. Pose supervision describes the data used for task-specific pose training: synthetic (S), real (R), or mixed (M).

cessfully detected by both systems. On these exact-common pairs, changing only the ReVA-HPE crop from MTCNN to RetinaFace changes MAE$_r$ from $3.712°$ to $3.705°$, while translation L_2 changes from 64.88 to 63.62 mm. Thus, for ReVA-HPE, detector replacement substantially affects coverage but has little effect on accuracy for cases detected by both systems. Table 4 reports all methods using RetinaFace crops on this matched subset. This detector-control experiment uses the default enrollment protocol and is distinct from the near-frontal enrollment reported in the main paper.

Overall, RetinaFace substantially improves detector coverage, while ReVA-HPE remains highly stable to crop replacement on exact-common cases. The effect on other methods is method-dependent, therefore, the reported-versus-reproduced baseline gap cannot be attributed to MTCNN alone.

7. Enrollment and Upstream-Error Sensitivity

Protocol scope. Main-paper Tables 4–5 use benchmark-facing shared-MTCNN evaluations over their full available query sets. Sections 7.1–7.2 below are separate controlled fixed-panel diagnostics designed to isolate enrollment-pose and image-quality effects. Their values are therefore complementary to, but not numerically interchangeable with, the main-paper rows. Section 7.3 expands the landmark branch-isolation analysis reported in main-paper Table 4(b–c), while Sec. 7.4 reports a separate controlled illumination stress test.

7.1. Enrollment-pose sensitivity

We evaluate all BIWI subjects using one deterministic enrollment from each of four pose strata, where $\alpha = \max(|\text{yaw}|, |\text{pitch}|, |\text{roll}|)$. The strata are near-frontal ($\alpha \leq 15°$), moderate ($15° < \alpha \leq 30°$), off-axis ($30° < \alpha \leq 45°$), and extreme ($\alpha > 45°$). For every enrollment, we rerun FaceLift reconstruction, alignment, PnP, bank generation, and inference on the same 100 fixed query frames. Ground truth is used only to construct the diagnostic strata and score predictions.

ReVA-HPE degrades more gracefully than the single enrollment-PnP rotation: from near-frontal to extreme enrollment, ReVA-HPE changes from $2.51°$ to $4.94°$, while

Enrollment	PnP+Rel. MAE_r	ReVA MAE_r	MAE_t	Trans. L_2	Mean bank inliers
Near-frontal	3.46	**2.51**	19.32	43.35	35.28
Moderate	4.29	**3.71**	18.81	40.74	28.48
Off-axis	6.08	**4.80**	19.43	41.85	25.80
Extreme	7.22	**4.94**	21.92	48.10	21.66

Table 5. **Controlled fixed-panel BIWI enrollment-pose diagnostic.** For all subjects, one deterministic enrollment from each pose stratum is evaluated on the same 100-query panel. Translation is shared by ReVA-HPE and PnP-enrollment+Rel. by design. Ground truth is used only to define the pose strata and for scoring. This controlled protocol differs from the full-query shared-MTCNN analysis in main-paper Table 5 the two tables are not numerically interchangeable. Rotation is in degrees and translation in millimetres.

PnP-enrollment+Rel. changes from 3.46° to 7.22°. Consensus support nevertheless decreases substantially, confirming that off-axis enrollment weakens the virtual-bank agreement. The FaMoS strata in the main paper use the same definition, their varying coverage reflects whether a valid enrollment exists in each pose range.

7.2. Degraded enrollment and avatar quality

We next perturb the enrollment RGB before any processing. The study uses one near-frontal and one extreme enrollment, and five conditions: clean input, Gaussian blur with $\sigma = 2$ or $\sigma = 4$, 50% downsampling followed by restoration, and a fixed 12.5% face-box occlusion. Every condition reruns detection, FaceLift, alignment, multiview landmarks, PnP, bank construction, and ReVA-HPE inference. No clean-image or crop fallback is allowed. Of 240 expected enrollment tasks, 222 complete, all 18 failures are retained in coverage (15 enrollment-detection and three multiview-landmark failures).

Across the 174 successful paired clean/degraded tasks, avatar-landmark RMS drift correlates with rotation degradation ($\rho = 0.5656$), RANSAC-inlier loss correlates with rotation degradation ($\rho = 0.5129$), PnP-anchor translation drift correlates with final translation degradation ($\rho = 0.6650$), and enrollment-landmark NME correlates with translation degradation ($\rho = 0.6097$). This provides direct evidence of modular error propagation. The study measures sensitivity to enrollment-induced avatar degradation, it is not a direct benchmark of absolute FaceLift reconstruction quality.

7.3. Enrollment-landmark perturbation

To isolate the PnP translation branch, we perturb only the 68 enrollment landmarks while keeping the enrollment image, avatar, crops, bank rotation, and cached relative predictions fixed. Gaussian jitter uses $\sigma \in \{1, 2, 4, 8\}$ pixels, and dropout removes $\{10, 25, 40\}\%$ of the landmarks. Each non-clean condition uses five deterministic seeds over the practical BIWI evaluation set. All EPnP solves succeed.

Mild 1–2 pixel jitter and random landmark removal produce limited degradation. Larger coordinate errors primarily corrupt depth, whereas random dropout is substantially less damaging than displacement of the retained correspondences.

Enrollment	Condition	Coverage	$MAE_r \downarrow$	$MAE_t \downarrow$	Trans. $L_2 \downarrow$	ADD ↓	Mean inliers
Near	Clean	100.0%	2.64	19.38	43.52	43.52	35.29
Near	Blur, $\sigma = 2$	100.0%	2.98	24.69	57.47	57.60	33.76
Near	Blur, $\sigma = 4$	91.7%	8.03	91.83	225.66	225.85	26.36
Near	Downsample 50%	100.0%	2.74	18.64	41.21	41.15	34.97
Near	Occlusion 12.5%	100.0%	4.11	44.63	112.22	112.67	32.01
Extreme	Clean	100.0%	5.00	21.84	47.87	48.38	21.69
Extreme	Blur, $\sigma = 2$	100.0%	6.32	60.03	142.13	142.93	20.52
Extreme	Blur, $\sigma = 4$	62.5%	18.99	172.14	414.43	418.93	9.27
Extreme	Downsample 50%	100.0%	5.03	22.61	49.87	50.61	21.62
Extreme	Occlusion 12.5%	70.8%	8.59	30.60	71.58	74.66	20.04

Table 6. **Controlled fixed-panel full-pipeline degraded-enrollment stress test.** For each subject, one near-frontal and one extreme enrollment are tested under every condition using the same 20-query panel. Coverage includes genuine detection and multiview-landmark failures, failed tasks are not replaced by clean inputs or fallback crops. This controlled protocol differs from the default-first-frame, full-query analysis in main-paper Table 4(a). Rotation is in degrees and translation metrics are in millimetres.

Perturbation	Severity	MAE_r	T_x	T_y	T_z	MAE_t	Trans. L_2	ADD
None	Clean	2.53	8.43	11.14	31.74	17.10	37.82	37.72
Gaussian jitter	1 px	2.53	8.48	11.17	31.96	17.21	38.01	37.90
Gaussian jitter	2 px	2.53	8.61	11.67	31.92	17.40	38.43	38.36
Gaussian jitter	4 px	2.53	9.39	12.10	45.55	22.35	51.18	51.16
Gaussian jitter	8 px	2.53	13.24	16.97	72.16	34.12	78.78	78.76
Dropout	10%	2.53	8.44	11.26	32.52	17.41	38.44	38.35
Dropout	25%	2.53	8.53	11.20	32.59	17.44	38.55	38.43
Dropout	40%	2.53	8.58	11.49	33.83	17.97	39.75	39.64

Table 7. **Detailed enrollment-landmark error propagation on BIWI.** This table expands main-paper Table 4(b–c). Only landmarks supplied to EPnP are perturbed. Each non-clean row averages five deterministic draws. Bank rotation is fixed by construction, therefore MAE_r remains unchanged and the experiment isolates the landmark–PnP–translation pathway. Errors are in millimetres except MAE_r, which is in degrees.

7.4. Illumination sensitivity

We evaluate a controlled enrollment-only illumination stress test using one near-frontal and one extreme-pose enrollment per BIWI subject. We modify only the enrollment image using -2 EV exposure, $+2$ EV exposure, or a directional shadow covering 75% of the image, and rerun detection, avatar reconstruction, alignment, landmark detection, PnP, virtual-bank construction, and query inference. Each condition uses the same 20 clean query images, giving 144 full-pipeline evaluations, all enrollments are constructed suc-

cessfully.

As shown in Table 8, most conditions change rotation by less than 0.25° and translation L_2 by less than 1.6 mm. The largest increase is 1.01° under strong overexposure of an extreme-pose enrollment. Negative changes reflect small empirical variation and should not be interpreted as benefits of degradation.

Anchor pose	Perturbation	$\Delta\mathrm{MAE}_r$	$\Delta L_{2,t}$	Δ inliers
Near	-2 EV	$+0.11^\circ$	$+1.27$ mm	-0.78
	$+2$ EV	-0.08°	-1.52 mm	$+0.31$
	75% shadow	$+0.22^\circ$	$+1.51$ mm	-0.49
Extreme	-2 EV	-0.19°	-0.74 mm	-0.46
	$+2$ EV	$+1.01^\circ$	$+3.38$ mm	-2.38
	75% shadow	$+0.08^\circ$	-1.25 mm	-0.52

Table 8. One near-frontal and one extreme-pose enrollment are perturbed while the same clean query panel is retained. Every condition reruns the complete enrollment and inference pipeline. Values are changes relative to the corresponding clean enrollment. Ground truth is used only to select the diagnostic enrollment-pose strata and for scoring.

7.5. Expression mismatch

We isolate anchor–query expression mismatch using five held-out synthetic identities and 20 fixed pose pairs per identity. For each matched pair, identity, camera pose and translation, focal length, lighting, texture, background, and crop are held fixed only the expression of either the query or the anchor is varied. Expression amplitudes of $0.5\times$ and $1\times$ lie within the training range, while $2\times$ provides an extrapolated stress condition. The experiment contains 100 predictions per condition and achieves 800/800 successful predictions.

As shown in Table 9, expression mismatch has negligible effect on rotation within the training range. At $1\times$, MAE_r changes by only 0.0061° when varying the query expression and 0.0088° when varying the anchor expression. Even at $2\times$, MAE_r remains below 0.63°. Translation is somewhat more sensitive to severe anchor-expression extrapolation, with MAE_t increasing from 1.5016 to 1.9160 mm.

This experiment uses the synthetic ground-truth anchor pose for decoding and is therefore an oracle-anchor diagnostic of the learned relative estimator. It does not evaluate expression-induced errors in FaceLift reconstruction or enrollment PnP, and no BIWI or FaMoS data are used.

8. Runtime and Deployment Trade-offs

We measure runtime on an NVIDIA A100-SXM-64GB with CUDA synchronization and one discarded shape-matched warm-up. Enrollment-stage components are measured over 48 runs, while model-side calibration and query times are measured over eight runs. Persistent model loading is excluded because it can be amortized across subjects and queries.

Expression varied in	Level	$\mathrm{MAE}_r \downarrow$	$\mathrm{MAE}_t \downarrow$
Query	Neutral	0.5612	1.5016
	Mild, $0.5\times$	0.5659	1.4764
	Moderate, $1\times$	0.5673	1.4755
	Severe, $2\times$	0.6030	1.5458
Anchor	Neutral	0.5612	1.5016
	Mild, $0.5\times$	0.5604	1.5531
	Moderate, $1\times$	0.5700	1.6693
	Severe, $2\times$	0.6177	1.9160

Table 9. **Controlled anchor–query expression sensitivity.** Results use five held-out synthetic identities and 20 fixed pose pairs per identity, yielding 100 predictions per condition and 800/800 successful predictions overall. Only query or anchor expression is varied camera geometry and all other rendering and preprocessing factors are fixed. The $0.5\times$ and $1\times$ conditions lie within the training range, while $2\times$ is an extrapolated stress condition. Ground-truth anchor pose is used for decoding, so this experiment isolates the learned relative estimator rather than the complete avatar/PnP pipeline. No BIWI or FaMoS data are used.

The timing scopes differ from the main-paper bank-size ablation. Main-paper Table 3 reports model-only inference time for the controlled bank-size sweep, whereas Table 10 below reports broader deployment-time accounting, including the enrollment pipeline and query-side orchestration overhead. The one-time enrollment totals combine separately measured enrollment-stage and model-side components rather than one monolithic wall-clock cohort.

Protocol	Bank use	One-time enrollment (s)	Per query (s)
Direct-5r	100/query	280.44 ± 20.54	9.944 ± 0.024
VA-fast	20/enrollment + 1/query	274.28 ± 20.44	0.225 ± 0.006

Table 10. **Deployment-time accounting.** Direct-5r performs query-specific consensus over 100 candidates, whereas VA-fast calibrates with 20 anchors once at enrollment and then performs one real-to-real prediction per query. These timings include broader pipeline overhead.

Runtime trade-off. Direct-5r and VA-fast have similar one-time enrollment costs (280.44 ± 20.54 s and 274.28 ± 20.44 s, respectively), because both require personalized avatar construction. Their per-query costs differ substantially: Direct-5r requires 9.944 ± 0.024 s for query-specific 100-candidate consensus, whereas VA-fast requires only 0.225 ± 0.006 s after enrollment.

The principal one-time components reported in the rebuttal are FaceLift reconstruction (53.31 ± 14.93 s) and virtual-anchor rendering, which takes 17.92 ± 0.94 s for the 100-candidate Direct-5r bank and 9.28 ± 1.30 s for the 20-anchor VA-fast calibration bank. Thus, VA-fast is the practical option for continuous or video inference, while Direct-5r retains query-specific multi-anchor consensus at

substantially higher per-query cost.

9. Qualitative Results

Figures 6 and 7 provide additional full-6DoF qualitative results on BIWI and FaMoS. Each row shows one query example together with the enrollment image and pose overlays from the compared methods. The overlays are produced by rendering the fitted head mesh under the predicted camera-frame pose, so both rotation and metric translation errors are visible. These examples complement the quantitative results in the main paper. On BIWI, ReVA-HPE remains visually close to the ground-truth overlay across subjects and head orientations, while using no ground-truth anchor pose. On FaMoS, the examples highlight the main cross-dataset behavior: single enrollment-PnP rotation can be unreliable under shifted camera geometry, whereas avatar-bank consensus provides more stable query rotation.

10. Failure Cases and Limitations

The main remaining failure modes are caused by avatar quality, enrollment quality, and landmark reliability. First, the personalized avatar is reconstructed from a single image. Errors in identity, hair, silhouette, or unseen profile geometry can produce rendered anchors that do not match the real query well, reducing the reliability of the bank consensus. Second, the real-enrollment PnP translation anchor depends on accurate 2D landmarks and a reasonable enrollment pose. Extreme enrollment orientations or unreliable contour landmarks can bias the metric translation gauge. Third, the current system processes each query independently after enrollment. Video-based applications could therefore benefit from periodic re-enrollment. These limitations do not require ground-truth anchors and are orthogonal to the relative 6DoF formulation. Future work could improve the avatar generator for geometric matching rather than visual fidelity alone, adaptively select better enrollment frames, or combine multiple enrollment images when available.

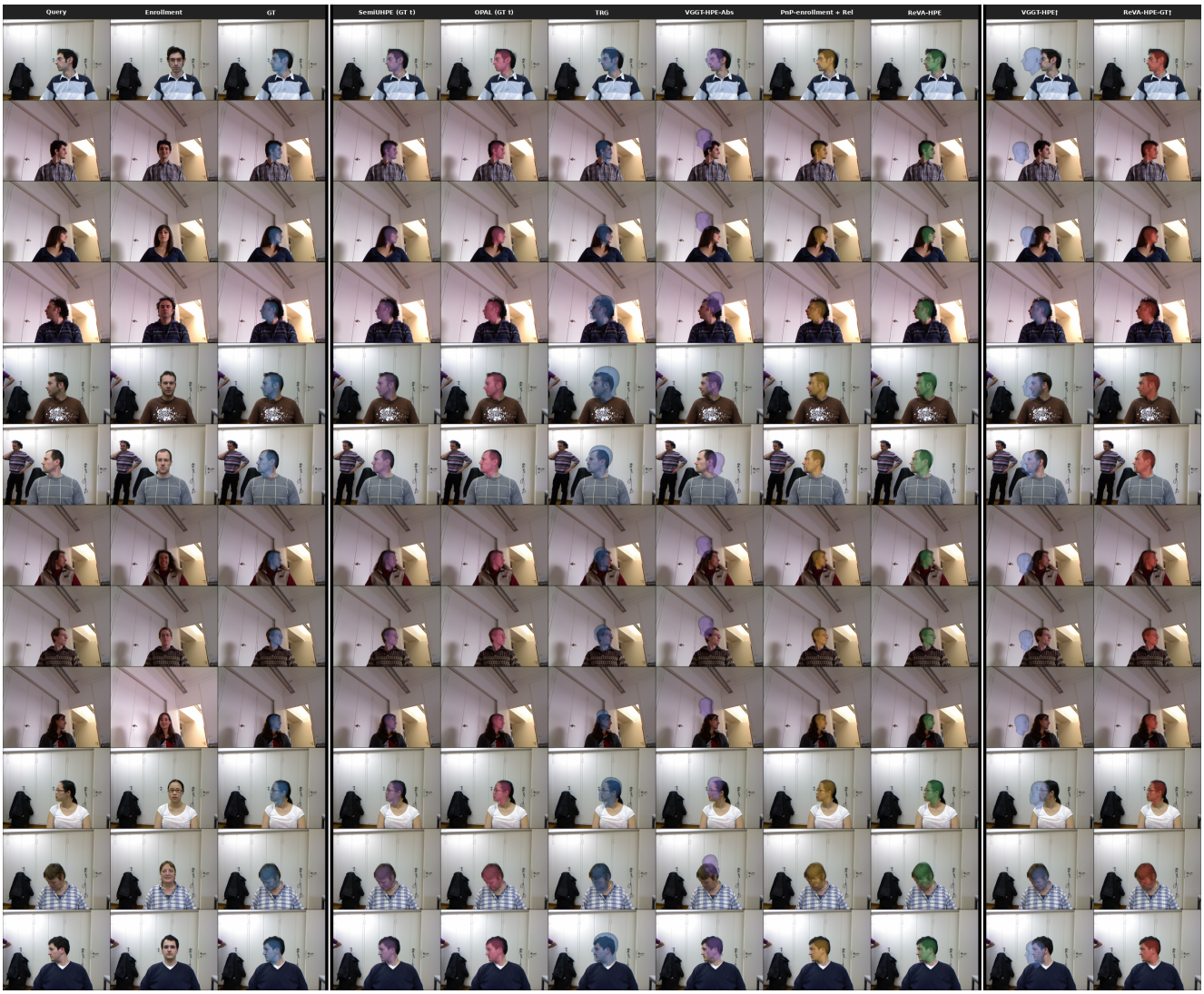


Figure 6. **Qualitative results on BIWI.** Each row shows one query example. The first columns show the query frame, the enrollment image used by ReVA-HPE, and the ground-truth pose overlay. The remaining columns compare reproduced baselines, the PnP-enrollment relative variant, ReVA-HPE, and GT-anchor oracle variants. All overlays show full 6DoF pose by rendering the fitted head mesh in the predicted camera-frame pose. ReVA-HPE produces accurate overlays across subjects and head orientations without using a ground-truth anchor pose.

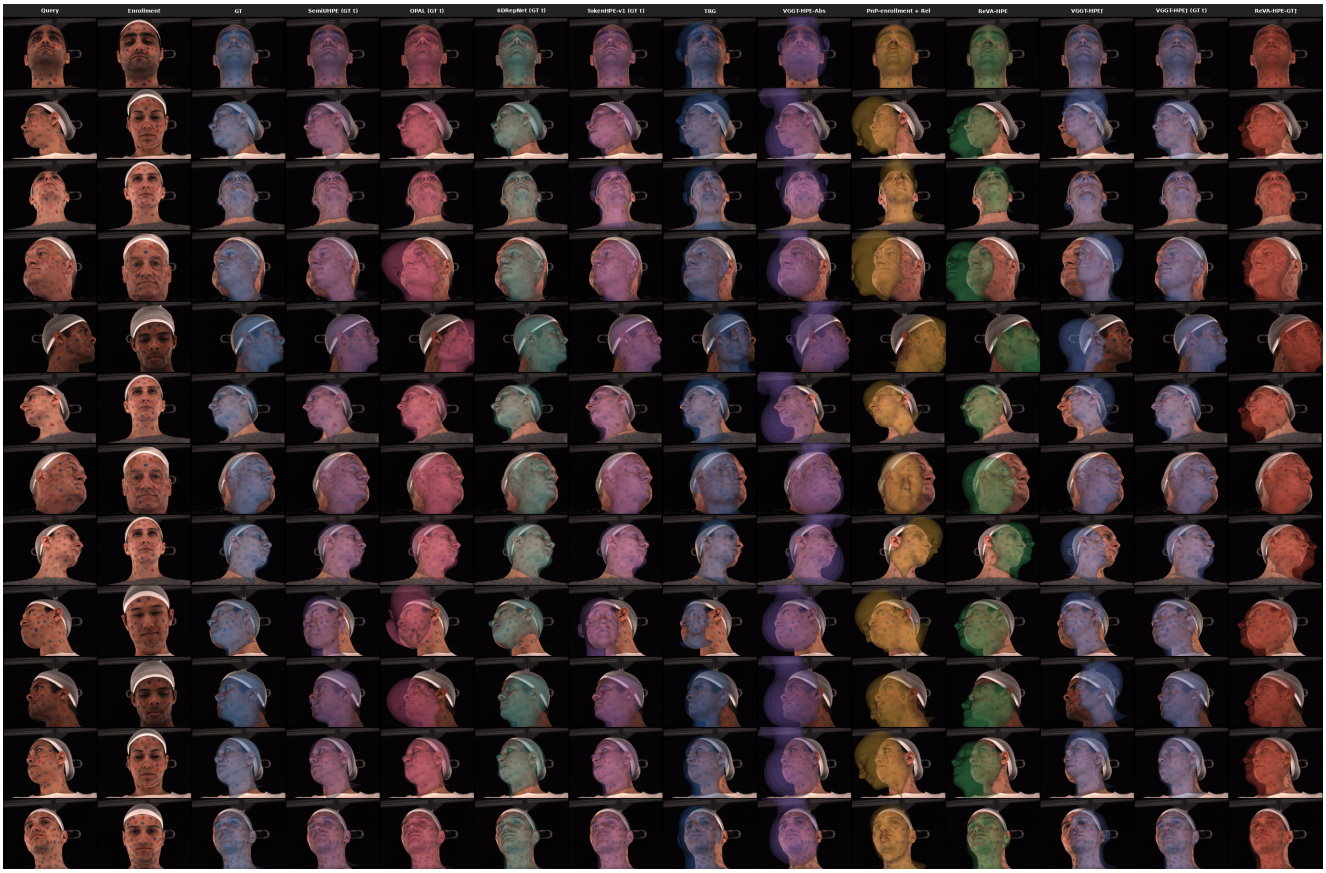


Figure 7. **Qualitative results on FaMoS.** Each row shows one query example from FaMoS together with the corresponding enrollment image and pose overlays for the compared methods. The dataset contains strong viewpoint and camera-geometry shifts, making single enrollment-PnP rotation unreliable in several cases. These examples are selected to illustrate cases in which avatar-bank rotation improves over the single enrollment-PnP rotation. They visualize the mechanism and representative failure modes but are not an unbiased estimate of average dataset performance.

References

[1] Sungho Chun and Ju Yong Chang. 6dof head pose estimation through explicit bidirectional interaction with face geometry. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XXXIII*, page 146–163, Berlin, Heidelberg, 2024. Springer-Verlag. 6

[2] 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. 6

[3] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. RetinaFace: Single-Shot Multi-Level Face Localisation in the Wild . In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5202–5211, Los Alamitos, CA, USA, 2020. IEEE Computer Society. 5

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

[5] Vasiliki Vasileiou, Panagiotis P Filntisis, Petros Maragos, and Kostas Daniilidis. Vggt-hpe: Reframing head pose estimation as relative pose prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 5464–5473, 2026. 6

[6] Cheng Zhang, Hai Liu, Yongjian Deng, Bochen Xie, and Youfu Li. TokenHPE: Learning orientation tokens for efficient head pose estimation via transformers. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8897–8906, 2023. 5, 6

[7] Huayi Zhou, Fei Jiang, Jin Yuan, Yong Rui, Hongtao Lu, and Kui Jia. Semi-supervised unconstrained head pose estimation in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(3):3082–3099, 2026. 6