

Article

Reliability-Conditioned Virtual-View Pose Fusion for Monocular 3D Human Pose Estimation with Real 2D Detector Inputs

Phalakron Nilkhet¹  and Thanaruk Theeramunkong^{1,2,*} 
¹ Sirindhorn International Institute of Technology, Thammasat University, Pathum Thani 12120, Thailand; d5522300010@siit.tu.ac.th

² The Royal Society of Thailand, 2–4, Building C, Government Complex, 120 Chaeng Watthana Road, Thung Song Hong, Lak Si, Bangkok 10210, Thailand

* Correspondence: thanaruk@siit.tu.ac.th; Tel.: +66-2-986-9010

Abstract

Monocular 3D human pose estimation is sensitive to depth ambiguity and upstream 2D-detector error. We investigate a detector-conditioned virtual 2D pose as an auxiliary hypothesis rather than an independent physical measurement. Joint reliability combines training-subject-only monotonic confidence calibration with detector-, joint-, and confidence-specific residual risk before a compact pose lifter. A four-arm factorial design separates observed-only (O), observed plus virtual (OV), observed plus reliability (OR), and observed plus virtual plus reliability (OVR). Across three subject-disjoint splits, three seeds, two RGB detectors, and 48 subject–sequence–camera clusters, OVR improved mean per-joint position error (MPJPE) by 3.049 mm (95% cluster-aware confidence interval (CI): 0.583–4.976); reliability-only improvement was supported, whereas the conditional virtual increment and interaction were not. Validation-only sensitivity covered eight yaw /pitch hypotheses and five multi-view policies; no multi-view policy improved primary MPJPE over the locked -15° view. A component-level audit of two synchronized directed camera pairs favored the generator over the observed-pose baseline in all 24 stratified rows, with only two sequence clusters per row. The locked MPI-INF-3DHP Official TS1–TS6 result improved MPJPE by approximately 0.97%, below its prespecified 2% practical threshold, with a CI crossing zero. On an additional external 3DPW cohort (13,579 eligible frames, 24 sequences, YOLO11L-Pose only), OVR improved MPJPE by 1.004 mm (95% CI: -2.965 – 5.016), while its Procrustes-aligned MPJPE point estimate worsened by 0.640 mm; neither contrast was conclusive. Thus, calibrated reliability is the most consistently supported mechanism, whereas virtual-view dominance, official superiority, and cross-dataset superiority are not established.

Keywords: monocular 3D human pose estimation; virtual view; reliability calibration; pose lifting; detector uncertainty; subject-disjoint evaluation; factorial ablation; MPI-INF-3DHP; 3DPW



Academic Editor: Danilo Avola

Received: 4 August 2026

Revised: 28 August 2026

Accepted: 3 September 2026

Published: 5 September 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Recovering a three-dimensional human skeleton from one RGB frame is an inverse problem with limited information. Perspective projection removes depth, self-occlusion hides joints, and several three-dimensional configurations can explain nearly identical two-dimensional keypoints. A common practical pipeline, therefore, separates the task into an external 2D detector and a learned 2D-to-3D pose lifter [1,2]. This decomposition is

modular, but it also transfers detector errors directly to the 3D prediction; only the pose-lifting module, not the complete detector–generator–lifter pipeline, is shown to be compact.

Multi-view supervision and multi-hypothesis models address the missing information in different ways. Real synchronized views provide geometric consistency during training [3,4], whereas diffusion and flow-based methods represent several plausible 3D explanations [5–8]. A third possibility is to remain entirely in pose space: predict how the observed 2D skeleton would appear from nearby virtual cameras and use those learned projections for 3D lifting, as demonstrated by the Multi-View Pose Generator of Sun et al. [9]. This approach avoids image-domain novel-view synthesis, but a virtual pose is not a new physical measurement. If the observed keypoints are unreliable, the generated view may simply transform and amplify their error.

This paper investigates that failure mode explicitly. The method uses two pretrained RGB pose detectors, creates one detector-conditioned virtual pose at a locked yaw of -15° , calibrates joint confidence without using the held-out subject, and estimates a bounded reliability value from calibrated confidence and a nonparametric residual-risk bank. The observed pose, virtual pose, their reliability-modulated versions, and the reliability vector are then passed to a compact lifter.

The main contributions are as follows:

1. A detector-conditioned virtual 2D pose treated as a reliability-modulated auxiliary hypothesis, not an additional physical camera or independent measurement.
2. Training-subject-only monotonic confidence calibration and residual-risk estimation, audited with held-out ECE, Brier score, and calibration slope for two real RGB detector families.
3. A 72-model four-arm factorial evaluation across subject-disjoint splits, seeds, and detectors, with whole subject–sequence–camera cluster resampling and a separately quarantined official partition.
4. Frozen-checkpoint validation sensitivity for eight viewpoints and five aggregation policies, together with a component-level synchronized-camera audit of the virtual-pose generator.
5. A frozen external 3DPW evaluation with explicit joint mapping, sample attrition, 24 sequence clusters, paired uncertainty, joint-level contrasts, and transparent reporting of mixed or inconclusive outcomes.

Figure 1 summarizes the implemented pipeline. The central distinction is between an externally pretrained 2D detector, the locked virtual-view generator, a non-learned reliability construction fitted on development data, and lifter weights trained for each factorial arm.

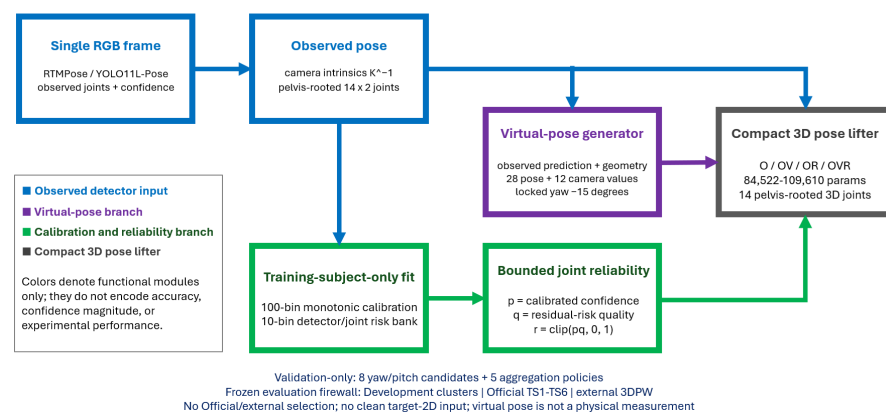


Figure 1. Implemented reliability-conditioned pose-space pipeline. RTMPose or YOLO11L-Pose Supplies observed keypoints and confidence from one RGB frame. One detector-conditioned virtual

pose is generated at the pre-existing locked yaw -15° ; a separate eight-candidate frozen validation screen is reported as sensitivity evidence, not test-set angle selection. Joint reliability is computed from leakage-safe calibrated confidence and a residual-risk quality factor. The full OVR representation concatenates observed pose, virtual pose, their reliability-modulated forms, and reliability itself before 3D lifting.

2. Related Work

2.1. Monocular 2D-to-3D Lifting

The residual multilayer perceptron of Martinez et al. [1] demonstrated that a carefully normalized 2D skeleton can support a strong 3D baseline. Later work introduced skeletal graph convolution [10,11], temporal convolution [2], spatial-temporal attention [12,13], and large-scale motion pretraining [14]. These approaches increase model capacity or temporal context, whereas our study intentionally retains a small framewise lifter so that changes can be attributed to virtual-view and reliability information rather than to a large backbone.

Probabilistic models treat monocular ambiguity as a distribution rather than a single regression target. Normalizing-flow methods incorporate 2D uncertainty [6]; D3DP and FinePOSE use diffusion-based multi-hypothesis generation [5,7]; and FMPose3D uses flow matching with reprojection-based aggregation [8]. Our virtual 2D pose is also a learned hypothesis, but it is deterministic, single-view, and explicitly tied to a target camera encoding. We therefore do not claim to model the full posterior over 3D poses.

2.2. Multi-View Supervision and Virtual Views

Multi-view supervision can constrain monocular inference without requiring multiple cameras at deployment. Rhodin et al. [3] learn from multi-view images, and Epipolar-Pose [4] exploits epipolar geometry for self-supervision. Geometry-aware representation learning synthesizes skeletons between viewpoints [15], while CanonPose factorizes pose and viewpoint in a canonical space [16]. Ingwersen et al. [17] use additional views and multi-view consistency for domain-specific fine-tuning while keeping inference monocular.

The closest pose-space precedent is the Multi-View Pose Generator of Sun et al. [9], which predicts multiple 2D views from a monocular 2D pose and passes the synthesized poses to a 3D lifting network. We therefore do not claim that generating virtual 2D poses for monocular lifting is itself new. Our study instead asks whether such a learned pose should be trusted when its source is a real RGB detector. Unlike direct use of all synthesized poses, the proposed representation combines a fixed geometry-conditioned hypothesis with leakage-safe joint reliability and separately measures the virtual-view effect, the reliability effect, and their interaction in a four-arm factorial protocol.

Recent work further narrows the distinction between monocular and multi-view lifting. Mocap-2-to-3 uses multi-view lifting and 2D pretraining to recover monocular motion and absolute position [18]. Our scope is different: a framewise root-relative skeleton, one nearby virtual yaw, and a factorial study of whether a synthesized pose is still useful after realistic detector error and confidence calibration. Accordingly, no state-of-the-art ranking is claimed because protocols, temporal context, joint layouts, and 2D inputs are not matched.

2.3. Detector Reliability and Robustness

The RGB-to-3D pipeline depends on the upstream keypoints. RTMPose targets efficient high-performance pose estimation [19], and YOLO11 supports real-time pose inference in the Ultralytics framework [20]. However, a detector's raw confidence need not equal the probability that a joint is geometrically accurate. Modern neural networks can be miscalibrated even when their ranking accuracy is strong [21]. Hoang et al. [22] previously incorporated detector confidence through a confidence-aware convolution and evaluated

robustness to corrupted inputs. Accordingly, our contribution is not the first use of detector confidence in a 3D lifter. It is the leakage-safe construction of calibrated, residual-risk-adjusted joint reliability and its factorial separation from a virtual-view branch. Pose augmentation and cross-dataset adaptation broaden the training distribution [23,24], but they do not by themselves identify whether a particular joint or synthesized view should be trusted.

3. Materials and Methods

3.1. Problem Formulation, Joint Mapping, and Coordinate Conventions

For an RGB frame I , detector D_d of family d returns pixel coordinates and raw confidence,

$$D_d(I) = \{(\mathbf{u}_j, c_j)\}_{j=0}^{14}, \quad \mathbf{u}_j = (u_j^x, u_j^y)^T, \quad c_j \in [0, 1]. \quad (1)$$

Table 1 specifies the canonical 15-joint correspondence, the COCO-style detector source, and the external SMPL-24 source. The 3D pelvis is the root and the remaining 14 joints are evaluated. Detector pelvis and neck are derived from observable hips and shoulders:

$$\mathbf{u}_{\text{pelvis}} = \frac{1}{2}(\mathbf{u}_{\text{left hip}} + \mathbf{u}_{\text{right hip}}), \quad c_{\text{pelvis}} = \min(c_{\text{left hip}}, c_{\text{right hip}}), \quad (2)$$

$$\mathbf{u}_{\text{neck}} = \frac{1}{2}(\mathbf{u}_{\text{left shoulder}} + \mathbf{u}_{\text{right shoulder}}), \quad c_{\text{neck}} = \min(c_{\text{left shoulder}}, c_{\text{right shoulder}}). \quad (3)$$

The detector head uses the nose proxy, whereas external 3D targets use the SMPL head; this semantic mismatch is an explicit limitation. With camera focal lengths f_x, f_y and principal point (c_x, c_y) ,

$$\mathbf{K} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}, \quad \tilde{\mathbf{u}}_j = \mathbf{K}^{-1} \begin{bmatrix} u_j^x \\ u_j^y \\ 1 \end{bmatrix}. \quad (4)$$

Normalized camera coordinates are pelvis-rooted to form the non-root observed pose $\mathbf{o} \in \mathbb{R}^{14 \times 2}$; the target $\mathbf{Y} \in \mathbb{R}^{14 \times 3}$ is pelvis-rooted in millimetres. Raw physical-camera annotations and RGB-detector pixel coordinates are distinct coordinate systems and are never treated as interchangeable.

Table 1. Canonical 15-joint mapping for the COCO-style detector and the frozen external SMPL-24 target. Detector indices are zero-based; pelvis and neck are derived.

Canonical Joint	Detector Source (COCO Index)	3DPW SMPL-24 Index
pelvis	midpoint (left hip, right hip)	0 (pelvis)
r_hip	right hip (12)	2 (right_hip)
r_knee	right knee (14)	5 (right_knee)
r_ankle	right ankle (16)	8 (right_ankle)
l_hip	left hip (11)	1 (left_hip)
l_knee	left knee (13)	4 (left_knee)
l_ankle	left ankle (15)	7 (left_ankle)
neck	midpoint (left shoulder, right shoulder)	12 (neck)
head	nose proxy (0)	15 (head)
l_shoulder	left shoulder (5)	16 (left_shoulder)
l_elbow	left elbow (7)	18 (left_elbow)
l_wrist	left wrist (9)	20 (left_wrist)
r_shoulder	right shoulder (6)	17 (right_shoulder)
r_elbow	right elbow (8)	19 (right_elbow)
r_wrist	right wrist (10)	21 (right_wrist)

3.2. Detector-Conditioned Virtual View

The virtual-view generator receives a root-relative observed pose and a relative camera code. The 12-dimensional code contains the nine entries of a 3×3 relative rotation and the three entries of a relative translation; the pose contributes $14 \times 2 = 28$ coordinates, so $28 + 9 + 3 = 40$ generator inputs. The model predicts a residual 2D pose for the target view. The locked generator uses a 40-dimensional input (28 pose coordinates and 12 geometry values), width 1024, three residual blocks, GELU activations, layer normalization, and dropout 0.10. It was trained on calibrated multi-view motion-capture pairs with an out-of-fold subject policy and was not fine-tuned during the reliability study.

The selected inference policy uses exactly one virtual view:

$$\mathbf{v} = G_{\theta}(\mathbf{o}, \mathbf{g}_{-15^{\circ}}), \quad (5)$$

where $\mathbf{g}_{-15^{\circ}}$ represents a yaw rotation of -15° . The generator input is the detector prediction, not ground truth 2D. Separate cached virtual poses are produced from the matched RTMPose or YOLO11L-Pose prediction; thus, detector-specific input errors condition the generator, although the generator can transform or partially correct those errors. No clean target-conditioned shortcut or second physical measurement enters inference. The -15° configuration was already fixed in the locked pipeline; seven additional yaw/pitch candidates (eight viewpoints in total) and five view-aggregation policies were subsequently screened on frozen validation checkpoints only, without using Development Test, Official TS1–TS6, or 3DPW for selection.

3.3. Leakage-Safe Joint Reliability

Raw detector confidence is calibrated separately for each cross-subject split and detector. Calibration uses training subjects only. A joint is defined as correct when its normalized keypoint error is below 0.10 torso lengths. One hundred equal-frequency bins are followed by weighted pair-adjacent-violators monotonic regression, yielding

$$p_j = \mathcal{C}_{s,d}(c_j), \quad (6)$$

where s indexes the subject-disjoint split, d indexes RTMPose or YOLO11L-Pose, and p_j is the calibrated probability that joint j satisfies the declared normalized-error threshold.

A nonparametric residual bank stores normalized 2D error vectors by split, detector, joint, and confidence bin. Let ρ_j be the mean residual magnitude for the relevant bank cell, and let m_j be the median positive training-bank risk for the same detector and joint. A bounded quality factor and final reliability are

$$q_j = \frac{1}{1 + \rho_j / \max(m_j, 10^{-6})}, \quad r_j = \text{clip}(p_j q_j, 0, 1). \quad (7)$$

Here, q_j is residual-risk quality, r_j is the bounded reliability applied to both observed and virtual joint features, and $\text{clip}(z, 0, 1) = \min(\max(z, 0), 1)$. The constant 10^{-6} is a fixed numerical floor that prevents division by zero; it is not a selected performance threshold. The risk bank uses 10 confidence bins, calibration fitting uses 100 equal-frequency bins, and held-out ECE evaluation uses 15 bins. Thus, high raw confidence is insufficient when calibration indicates overconfidence or the residual bank predicts large joint error.

Figure 2 shows the validation audit. Calibration reduced expected calibration error (ECE) and Brier score in all six split-detector pairs; no validation or test subject was used to fit its own calibrator. The corresponding cohort-level calibration values are reported in Table 2.

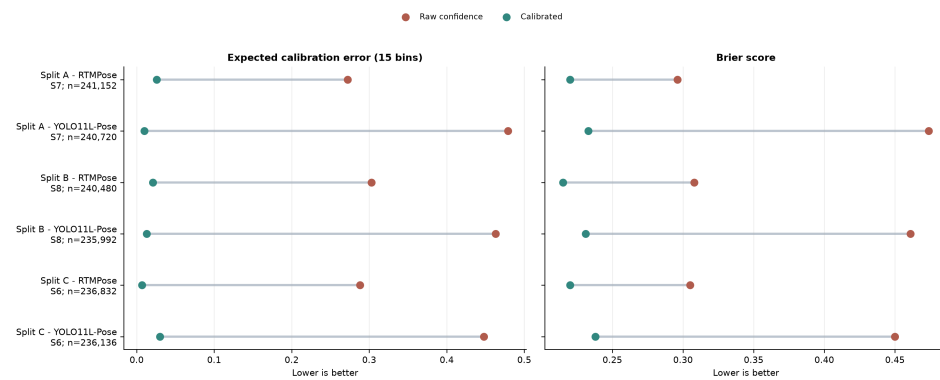


Figure 2. Held-out validation calibration for six explicitly labeled split-detector cohorts. Panels show ECE using 15 evaluation bins and Brier score before versus after monotonic calibration. Labels include the held-out subject and exact joint-observation count; calibration fitting used 100 bins and residual-risk lookup used 10 bins.

Table 2. Held-out calibration by split, detector, validation subject, and joint-observation count. Brier score and ECE decrease in all six cohorts; slopes near one indicate improved calibration.

Split/Detector	Subject	Observations	Raw ECE	Cal. ECE	Raw Brier	Cal. Brier	Cal. Slope
A/RTMPose	S7	241,152	0.272	0.026	0.296	0.220	1.020
A/YOLO11L-Pose	S7	240,720	0.479	0.010	0.474	0.233	1.035
B/RTMPose	S8	240,480	0.303	0.021	0.308	0.215	1.028
B/YOLO11L-Pose	S8	235,992	0.463	0.013	0.461	0.231	0.877
C/RTMPose	S6	236,832	0.288	0.007	0.305	0.220	0.999
C/YOLO11L-Pose	S6	236,136	0.448	0.030	0.450	0.238	0.844

The mean Brier score decreases from 0.3823 to 0.2261 across the six held-out cohorts, and calibrated slopes range from 0.844 to 1.035.

3.4. Factorial Feature Arms and Pose Lifter

The four prespecified arms form a 2×2 design: O is observed-only, OV is observed plus an ungated virtual pose, OR is observed plus reliability, and OVR is observed plus the virtual pose and reliability:

$$\mathbf{f}_O = [\text{vec}(\mathbf{o})], \quad \dim(\mathbf{f}_O) = 28, \quad (8)$$

$$\mathbf{f}_{OV} = [\text{vec}(\mathbf{o}), \text{vec}(\mathbf{v})], \quad \dim(\mathbf{f}_{OV}) = 56, \quad (9)$$

$$\mathbf{f}_{OR} = [\text{vec}(\mathbf{o}), \text{vec}(\mathbf{r} \odot \mathbf{o}), \mathbf{r}], \quad \dim(\mathbf{f}_{OR}) = 70, \quad (10)$$

$$\mathbf{f}_{OVR} = [\text{vec}(\mathbf{o}), \text{vec}(\mathbf{v}), \text{vec}(\mathbf{r} \odot \mathbf{o}), \text{vec}(\mathbf{r} \odot \mathbf{v}), \mathbf{r}], \quad \dim(\mathbf{f}_{OVR}) = 126. \quad (11)$$

Here, $\mathbf{r} \odot \mathbf{o}$ and $\mathbf{r} \odot \mathbf{v}$ apply the joint reliability to both coordinates of each joint. This is feature modulation, not a convex average of two predicted 3D poses.

Each arm uses the same compact lifter family: a fully connected layer of width 256, ReLU, layer normalization, a second width-256 ReLU layer, and a final linear layer reshaped to 14×3 . Models are trained from scratch for each split, seed, detector, and arm with mean squared error in metres, AdamW, learning rate 10^{-3} , weight decay 10^{-4} , batch size 2048, gradient clipping at 5, a maximum of 50 epochs, and early-stopping patience 8. Checkpoints minimize the geometric mean of validation MPJPE and PA-MPJPE for the matched detector. Calibrators, residual-risk banks, and lifters are detector-specific; a lifter trained with one detector is not evaluated with the other detector's outputs. Verified trainable pose-lifter parameter counts are 84,522 (O), 91,690 (OV), 95,274 (OR), and 109,610 (OVR). These counts

exclude the RGB detector and virtual generator, and no end-to-end latency or FLOP claim is made; accordingly, only the pose lifter is described as compact.

3.5. Dataset, RGB Detector Audit, and Splits

MPI-INF-3DHP supplies synchronized RGB and motion-capture data for eight training subjects and a six-subject Official Test partition [25]. The development RGB audit sampled 160,728 frames at 5 Hz from 128 videos: two sequences for each of subjects S1–S8 and eight available camera indices (0, 1, 2, 4, 5, 6, 7, and 8). Person selection used detector-native confidence with no ground truth assistance.

Two independently implemented detectors were used: RTMPose in the `rtmlib` performance configuration (version 0.0.15) and `yolo11l-pose.pt` through Ultralytics (version 8.4.112). Table 3 reports the locked detector audit. Median normalized error and PCK2D favoured RTMPose, while the mean was affected by a long error tail for both detectors. The purpose of this audit was to characterize detector error and fit the reliability mechanism, not to claim improved 3D performance.

Table 3. Development RGB detector audit on 160,728 frames. NME is normalized by ground-truth pelvis-to-neck torso length; common-12 excludes derived pelvis/neck and the head proxy.

Detector	Success (%)	Mean NME	Median NME	PCK2D@0.10 (%)	corr($c, -e$)	Raw ECE
RTMPose performance	100.000	0.311	0.102	49.103	0.435	0.285
YOLO11L-Pose	99.157	0.268	0.116	42.607	0.297	0.470

Three subject-disjoint protocols were fixed before the final confirmation (Table 4). Seeds were 42, 123, and 2026. The matrix contains 3 splits $\times$ 3 seeds $\times$ 2 detectors $\times$ 4 arms = 72 trained lifters. Development Test was post-selection evidence and was not used for check-point or arm selection.

Table 4. Locked cross-subject development splits. Calibration and residual-risk fitting use only the training subjects in each row.

Split	Training Subjects	Validation Subject	Development Test Subject
A	S1–S6	S7	S8
B	S1–S5, S7	S8	S6
C	S1–S5, S8	S6	S7

3.6. Outcome Definitions, Sampling Units, and Leakage Boundaries

MPJPE is the primary error endpoint; PA-MPJPE, 3DPCK, AUC, and calibration statistics are complementary outcomes. For N frames and $J = 14$ non-root canonical joints,

$$\text{MPJPE} = \frac{1}{NJ} \sum_{n=1}^N \sum_{j=1}^J \|\hat{\mathbf{Y}}_{n,j} - \mathbf{Y}_{n,j}\|_2, \quad (12)$$

$$\text{PA-MPJPE} = \frac{1}{NJ} \sum_{n=1}^N \sum_{j=1}^J \|T_n^*(\hat{\mathbf{Y}}_{n,j}) - \mathbf{Y}_{n,j}\|_2, \quad (13)$$

$$\text{3DPCK}_{150} = \frac{1}{NJ} \sum_{n=1}^N \sum_{j=1}^J \mathbf{1}[\|\hat{\mathbf{Y}}_{n,j} - \mathbf{Y}_{n,j}\|_2 < 150 \text{ mm}], \quad (14)$$

$$\text{Gain}_{A \rightarrow B} = \text{MPJPE}(A) - \text{MPJPE}(B). \quad (15)$$

Here, T_n^* is the best-fitting per-pose Procrustes similarity transform; a positive gain favors B . AUC integrates 3DPCK over 0–150 mm. For correctness indicators z_i and calibrated

probabilities p_i , $\text{Brier} = N^{-1} \sum_i (p_i - z_i)^2$; ECE aggregates absolute binwise confidence–accuracy differences across 15 bins. The reported calibration slope is estimated on held-out validation observations.

The development analysis contains 48 subject–sequence–camera clusters and resamples split/test subject, seed within split, matched detector, and complete sequence–camera clusters (20,000 paired replicates). Official inference has six subjects and uses subject, split, seed within split, and detector as resampling levels. The external analysis resamples 24 complete 3DPW sequences and nine matched split–seed checkpoint units (10,000 paired replicates). Neither Official labels nor external outcomes selected a yaw, checkpoint, arm, detector, calibrator, threshold, aggregation policy, or qualitative example; FreeMan and Q1 fitted state were excluded.

3.7. Frozen External 3DPW Evaluation Protocol

3DPW provides an independent in-the-wild benchmark with 3D pose supervision derived from a moving camera and inertial measurements [26]. The frozen test-only protocol first enumerated 26,240 annotated frames, of which 26,214 had a YOLO11L-Pose detection. Requiring exactly one camera-pose-valid annotated person and a detector-selected person retained 13,579 eligible frames (51.75% of the annotated cohort) across 24 sequences. This conditioning is reported as sample attrition rather than full-benchmark coverage.

The original training splits, three seeds, fixed -15° policy, existing calibrators, and all 27 frozen checkpoints were reused without tuning on 3DPW. The external scope includes only one detector and three available arms (O, OV, and OVR); the OR arm and a second detector were not evaluated externally. The canonical SMPL-24 correspondence is fixed in Table 1, pelvis-rooted 3D comparisons use 14 non-root joints, and matched checkpoint units remain paired across all contrasts.

3.8. Locked Official TS1–TS6 Protocol

Only after the development and factorial decisions were frozen was the new RGB-detector-input method evaluated on TS1–TS6. This was the first locked Official evaluation of the reliability-conditioned RGB-detector method, although TS1–TS6 had previously been inspected using a legacy ground-truth-2D system. The present evaluation is therefore not described as a pristine project-level first look.

All 2875 frames marked valid by the dataset were retained. Both detectors produced a person for every Official frame. Detector-native person selection was used; no ground-truth 2D input or ground-truth-assisted person selection was allowed. The 3D annotation was unsealed only after the detector cache was locked. Predictions and targets were pelvis-rooted and evaluated on 14 non-root canonical joints. MPJPE is the primary manuscript endpoint; PA-MPJPE, 3DPCK at 150 mm, and AUC over thresholds 0–150 mm in 5 mm steps are secondary. The previously locked Official decision rule additionally required positive uncertainty support for both MPJPE and PA-MPJPE before a confirmatory Official claim. The primary aggregate gives equal weight to subject $\times$ split $\times$ seed $\times$ detector units.

Paired uncertainty was estimated with 20,000 hierarchical bootstrap replicates. Development resampling followed split, seed, and detector; Official resampling followed subject, split, seed-within-split, and detector. A positive reported improvement means lower error for the named challenger. The strongest predeclared Official claim additionally required both MPJPE and PA-MPJPE paired 95% CIs to exclude zero, both detector-specific point means to improve, at least 90% detector coverage, and at least a 2% relative MPJPE improvement. The 2% practical threshold belongs to the locked Official protocol only and was not retrospectively declared for Development or 3DPW. The result was not rerun for model selection after unsealing.

3.9. Use of Generative Artificial Intelligence

During preparation of this manuscript, P.N. used OpenAI ChatGPT Work with Codex (GPT-5.6) (accessed on 4 August 2026) to assist with English-language drafting, LaTeX migration, literature-query organization, and plotting-code preparation. No experimental result, detector output, confidence interval, or scientific gate was generated by these tools. Numerical statements were read from locked result artifacts. The authors reviewed and edited the outputs and take full responsibility for the content of the manuscript.

4. Results

4.1. Development Confirmation

Across the 18 matched split–seed–detector units, OVR reduced MPJPE from 128.943 to 125.894 mm and PA-MPJPE from 85.916 to 84.320 mm. The mean improvements were 3.049 mm (2.269%) for MPJPE and 1.596 mm (1.831%) for PA-MPJPE. Both finalized whole-cluster hierarchical 95% CIs were positive: MPJPE 0.583–4.976 mm and PA-MPJPE 0.736–2.375 mm (Table 5). Mean improvements were positive for both detectors and for all three subject splits, although split B was substantially weaker than splits A and C.

Table 5. Development Test factorial results from the finalized paired, whole-cluster reanalysis. Absolute values aggregate 18 matched split–seed–detector units; confidence intervals resample all 48 subject–sequence–camera clusters.

Arm or Effect	MPJPE (mm)	PA-MPJPE (mm)	MPJPE Gain [95% CI]	PA Gain [95% CI]
O: observed-only	128.943	85.916	—	—
OV: observed + virtual	128.790	85.799	0.153 [−1.359, 1.548]	0.117 [−0.661, 0.997]
OR: observed + reliability	126.292	84.793	2.650 [0.572, 4.506]	1.123 [0.300, 2.035]
OVR: observed + virtual + reliability	125.894	84.320	3.049 [0.583, 4.976]	1.596 [0.736, 2.375]
Conditional virtual: OVR − OR	—	—	0.399 [−0.651, 1.498]	0.473 [−0.116, 1.074]
Interaction: (OVR − OR) − (OV − O)	—	—	0.246 [−1.364, 1.728]	0.355 [−0.814, 1.308]

The factorial result localizes the stable development evidence. Reliability-only OR has positive intervals for both primary metrics. In contrast, ungated OV is near zero, the incremental effect of adding the virtual view to OR is positive in point estimate but statistically unresolved, and the difference-in-differences interaction crosses zero. Figure 3 visualizes these effects. The correct development claim is therefore that the selected full representation improves the baseline and that reliability is the more consistently supported component; the study does not establish a strong virtual–reliability synergy.

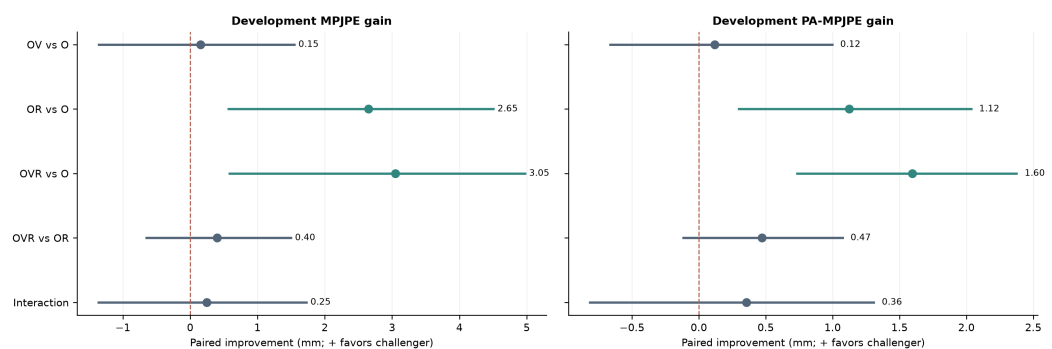


Figure 3. Development factorial attribution. The full OVR and reliability-only OR effects have positive 95% intervals for both primary metrics. The conditional virtual-view increment and the virtual–reliability interaction remain unresolved.

4.2. Frozen Validation Viewpoint and Multi-View Sensitivity

The locked yaw -15° was assessed against seven additional frozen-checkpoint validation candidates without retraining or examining Official/external outcomes. All eight values are reported in Table 6; yaw -15° had the lowest validation MPJPE (125.339 mm). This is a checkpoint-frozen validation sensitivity finding, not a trained-angle comparison or a globally optimal viewpoint.

Table 6. All eight frozen-checkpoint, validation-only virtual-view candidates. No Development Test, Official, or external labels selected the policy.

Candidate	Validation MPJPE (mm)	Validation PA-MPJPE (mm)
Yaw -15° (locked policy)	125.339	83.981
Pitch -10°	127.359	85.547
Yaw -30°	128.762	85.291
Pitch $+10^\circ$	131.527	87.276
Yaw -45°	136.167	88.873
Yaw $+15^\circ$	136.382	90.752
Yaw $+30^\circ$	149.757	97.955
Yaw $+45^\circ$	164.815	105.148

No multi-view policy improved primary MPJPE over the single locked yaw. Table 7 reports all five policies, and Figure 4 reports the complete positive and negative sensitivity evidence without selecting a new checkpoint or view.

Table 7. All five prespecified frozen validation aggregation policies. Lower MPJPE is the primary comparison; the triple policy's slightly lower PA-MPJPE does not overturn its higher primary MPJPE.

Policy	Validation MPJPE (mm)	Validation PA-MPJPE (mm)
Single: yaw -15°	125.339	83.981
Triple: yaw $-15^\circ/-30^\circ$ + pitch -10°	125.360	83.956
Diverse: yaw -15° + pitch -10°	125.635	84.367
Nearby: yaw $-15^\circ/-30^\circ$	126.291	84.229
Symmetric: yaw $-15^\circ/+15^\circ$	128.267	85.821

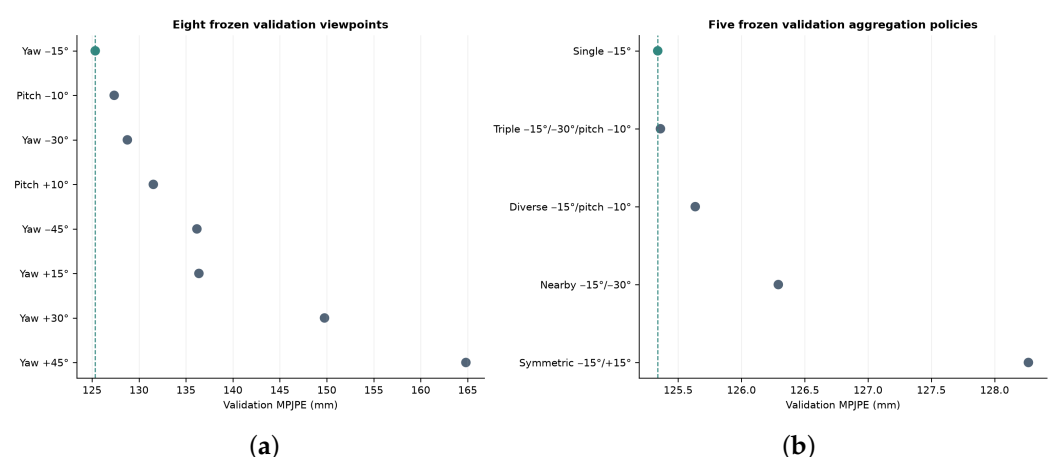


Figure 4. Frozen validation-only sensitivity: (a) all eight yaw/pitch candidates; (b) all five prespecified aggregation policies. The dashed line marks the locked single-yaw reference. This descriptive screen does not retrain candidates or authorize Official/external model selection.

4.3. Independent Synchronized-Camera Generator Audit

The frozen generator was additionally compared with physical paired-camera 2D reference annotations using two available directed camera pairs, $c05 \rightarrow c12$ and $c07 \rightarrow c09$.

The comparator is the audit-defined observed-pose baseline in the declared paired-reference coordinate convention; the generator produces the geometry-conditioned target-view hypothesis. Across three subject splits, validation/test partitions, two detectors, and two directed pairs, all 24 stratified rows had smaller generator NME than their matched observed-pose baseline (Table 8; Figure 5).

Table 8. Physical paired-camera generator audit summarized across split and partition. Each detector–pair group contains six stratified rows, but each row has only two sequence clusters; rows are not independent physical replications.

Directed Pair	Detector	Rows	Generator NME Range	Observed NME Range	Clusters/Row
c05 → c12	RTMPose	6	0.302–0.354	1.985–2.183	2
c05 → c12	YOLO11L-Pose	6	0.316–0.373	1.980–2.167	2
c07 → c09	RTMPose	6	0.291–0.340	0.696–0.885	2
c07 → c09	YOLO11L-Pose	6	0.301–0.351	0.664–0.836	2

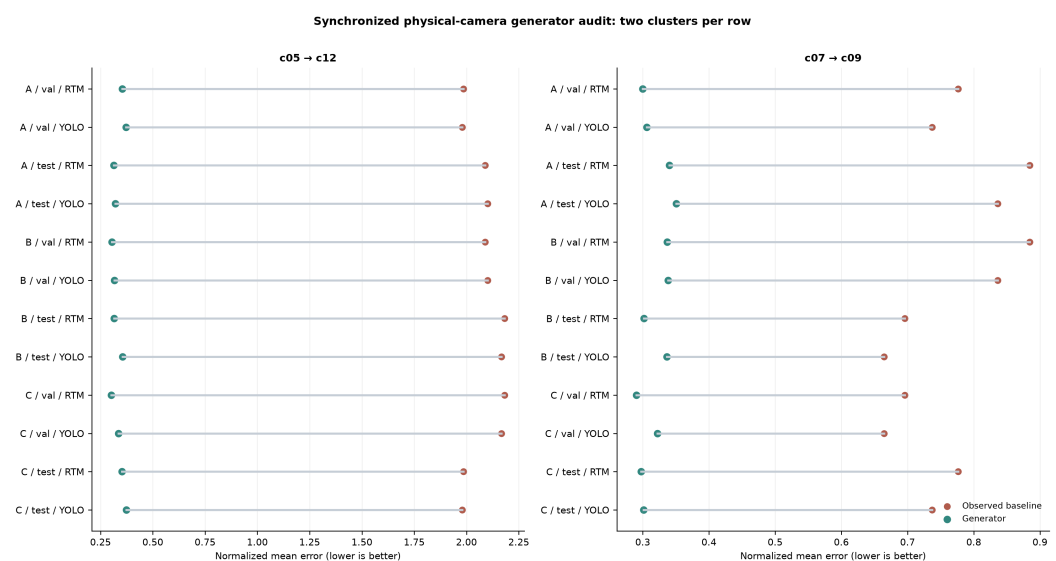


Figure 5. Component-level synchronized-camera generator validation for all 24 split/partition/detector/pair rows. Generator and observed-baseline NME are paired within each row. Only two directed camera pairs and two sequence clusters per row are available.

Only these two directed pairs are available; reverse or other camera directions must not be inferred. The comparison validates the virtual-pose generator component against physical paired-camera annotations. It does not validate end-to-end RGB-detector accuracy on the target camera, does not establish equivalence of raw physical and detector pixel coordinates, and cannot support robust population-level uncertainty from only two sequence clusters per row.

4.4. Locked Official Factorial Result with Real Detector Inputs

Table 9 and Figure 6 report the primary Official aggregate. OVR achieved the lowest MPJPE and PA-MPJPE among the four arms. Relative to O, the point improvements were 1.231 mm (0.967%) MPJPE, 2.327 mm PA-MPJPE, and 0.564 percentage points 3DPCK. AUC decreased by 0.185 percentage points. Thus, not every metric favoured the full representation.

The hierarchical bootstrap did not confirm superiority. OVR-versus-O MPJPE improvement had mean 1.237 mm and 95% CI −1.234 to 3.644 mm; PA-MPJPE improvement had mean 2.325 mm and 95% CI −0.229 to 4.954 mm. Both paired 95% CIs included zero, and the predeclared interval-exclusion criterion was not met.

Table 9. Locked MPI-INF-3DHP TS1–TS6 aggregate using real RGB detector predictions and no ground-truth 2D model input. Lower is better for MPJPE and PA-MPJPE; higher is better for 3DPCK and AUC. Bold values indicate the best point estimate in each metric column.

Arm	MPJPE (mm)	PA-MPJPE (mm)	3DPCK (%)	AUC (%)
O	127.295	84.176	71.028	34.560
OV	127.468	83.856	71.096	34.228
OR	126.481	82.188	71.379	34.494
OVR	126.065	81.849	71.592	34.375

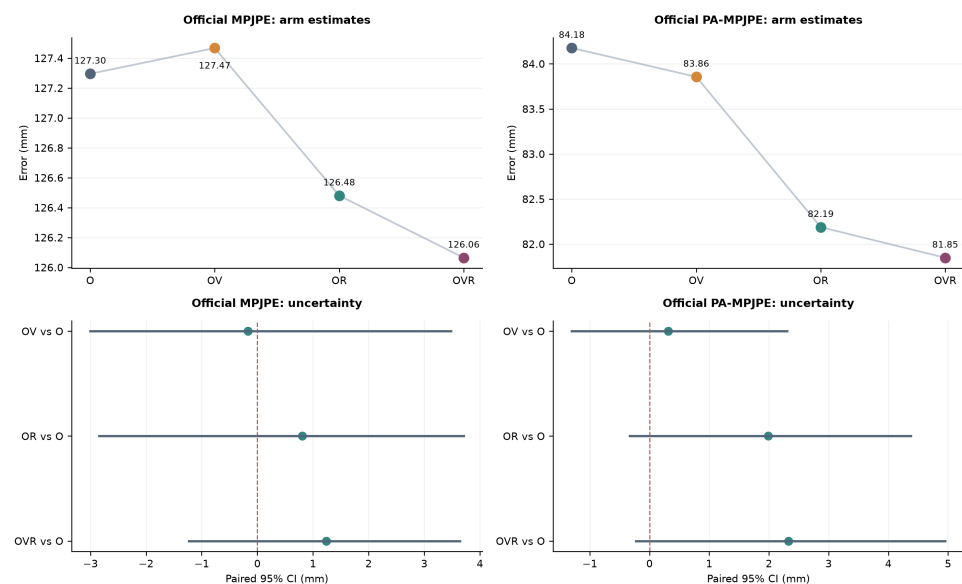


Figure 6. Locked Official TS1–TS6 arm point estimates and paired 95% hierarchical confidence intervals relative to O. The dashed vertical line in each effect panel marks zero difference relative to O. The right-hand effect panels include zero for every available Official comparison; descriptive arm differences are not proof of superiority.

Official factorial attribution was also unresolved (Figure 7). Reliability-only OR versus O had point gains of 0.810 mm MPJPE and 1.986 mm PA-MPJPE, but both intervals included zero. Ungated OV versus O was -0.166 mm MPJPE and 0.314 mm PA-MPJPE. The conditional virtual increment OVR versus OR was 0.427 mm MPJPE and 0.339 mm PA-MPJPE. The interaction estimate was 0.593 mm MPJPE and 0.025 mm PA-MPJPE. Every official factorial 95% interval crossed zero.

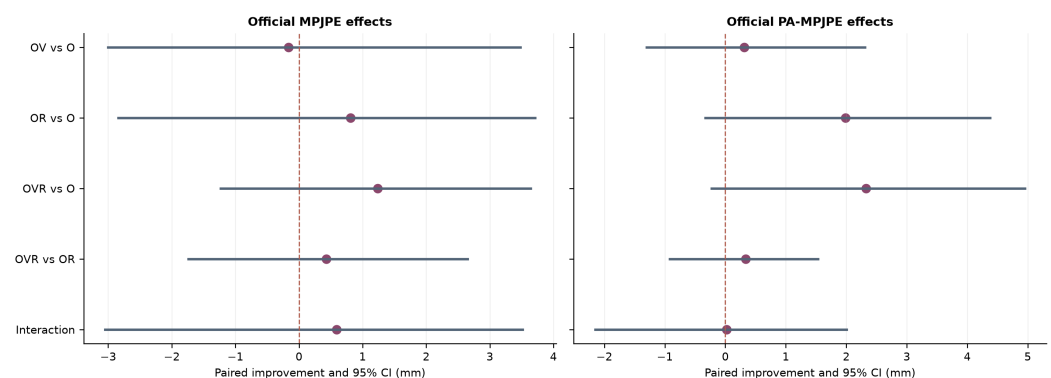


Figure 7. Paired hierarchical Official effect estimates. The dashed vertical line denotes zero effect. All intervals cross the zero-effect line, so Official superiority, a standalone reliability effect, a standalone virtual-view effect, and synergy are not confirmed.

4.5. Detector-Specific Behavior

Both detector-specific OVR point estimates improved both primary errors (Figure 8). RTMPose improved MPJPE by 1.426 mm and PA-MPJPE by 2.258 mm; YOLO11L-Pose improved MPJPE by 1.036 mm and PA-MPJPE by 2.396 mm. Detection coverage was 100% for both detectors on all 2875 official frames. The consistent direction across detector families is encouraging, but detector-specific point estimates do not replace the hierarchical uncertainty analysis.

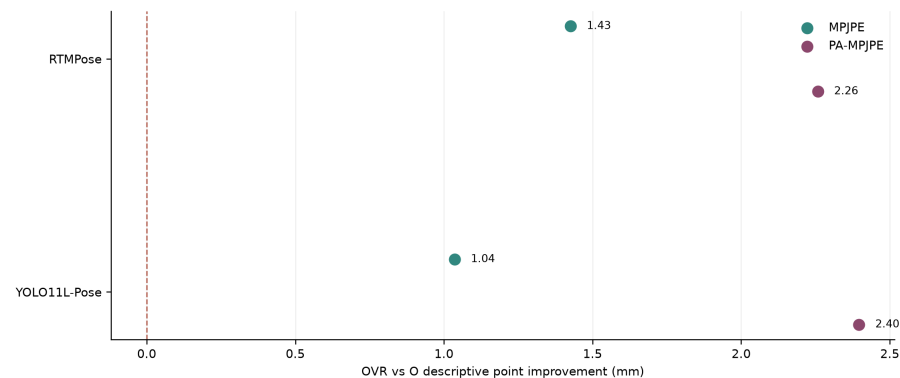


Figure 8. Detector-specific Official point improvements for OVR versus O. The dashed vertical line denotes zero point improvement. Both primary errors improve for RTMPose and YOLO11L-Pose, while the overall hierarchical intervals in Figure 7 remain non-confirmatory.

4.6. Evidence-Grounded Qualitative Example

Figure 9 shows a deterministic median-case example from the locked Official archive. The sample is the RTMPose TS5 frame whose O-minus-OVR MPJPE change is nearest the TS5 median; it was selected only after the result lock for visualization and was not used for model selection. The corresponding original RGB frame was retrieved from the MPI-INF-3DHP dataset using the image path recorded in the locked detector cache. Panel (a) overlays the actual cached RTMPose keypoints used as model input on that frame. Panels (b)–(d) use the saved MoCap target and ensemble predictions.

Locked median-case Official example: TS5, frame 49



Figure 9. Locked median-case Official example (TS5, frame 49). (a) Original MPI-INF-3DHP RGB frame overlaid with the actual RTMPose keypoints used as model input. (b) MoCap target. (c) Observed-only ensemble prediction. (d) Full OVR ensemble prediction. The full method reduces MPJPE from 86.9 to 82.9 mm and PA-MPJPE from 46.7 to 45.0 mm. The sample was selected after the result lock for visualization only and was not used for model selection.

4.7. Frozen External 3DPW Results and Joint-Level Effects

The independent 3DPW cohort comprises 13,579 eligible frames from 26,240 annotated frames across 24 sequences. Table 10 reports all 27 frozen checkpoints (nine per arm) using YOLO11L-Pose only, and Table 11 reports the paired external contrasts. OVR has a lower

MPJPE point estimate than O, but its PA-MPJPE point estimate is higher; therefore, the external directions are mixed.

Table 10. Frozen external 3DPW arm means over nine locked split-seed checkpoints per arm. Only YOLO11L-Pose and arms O, OV, and OVR were available; OR and a second detector were not evaluated.

Arm	Checkpoints	MPJPE (mm)	PA-MPJPE (mm)	3DPCK@150 (%)	AUC (%)
O	9	140.182	89.334	69.747	27.983
OV	9	139.467	88.797	70.343	27.977
OVR	9	139.178	89.975	70.417	28.118

Table 11. Paired external improvements and 95% two-level cluster bootstrap intervals. Positive values favor the first named challenger; 3DPCK and AUC are percentage-point differences. Every interval includes zero. PA-MPJPE comes from the matched frozen 3X analysis; MPJPE/3DPCK/AUC come from the corrected 3Z analysis on the same cohort.

Contrast	MPJPE Gain (mm)	PA-MPJPE Gain (mm)	3DPCK Gain (pp)	AUC Gain (pp)
OV versus O	0.715 [−1.156, 2.873]	0.537 [−1.251, 2.115]	0.596 [−0.497, 1.696]	−0.007 [−0.592, 0.607]
OVR versus OV	0.289 [−3.617, 4.147]	−1.178 [−2.938, 0.688]	0.075 [−1.818, 2.014]	0.141 [−1.113, 1.439]
OVR versus O	1.004 [−2.965, 5.016]	−0.640 [−2.040, 0.766]	0.671 [−1.206, 2.708]	0.135 [−1.008, 1.330]

The OVR-versus-O MPJPE gain is 1.004 mm (95% CI −2.965 to 5.016), whereas the corresponding PA-MPJPE gain is −0.640 mm (95% CI −2.040 to 0.766). Thus, neither overall external superiority nor reliable alignment improvement is established. Figure 10 reports all paired contrasts, and Figure 11 reports the exploratory 14 non-root external joint-level net effects. Joint labels were corrected to the canonical 15-joint order after correcting a one-position offset; these external joint findings must not be represented as Development all-arm joint evidence.

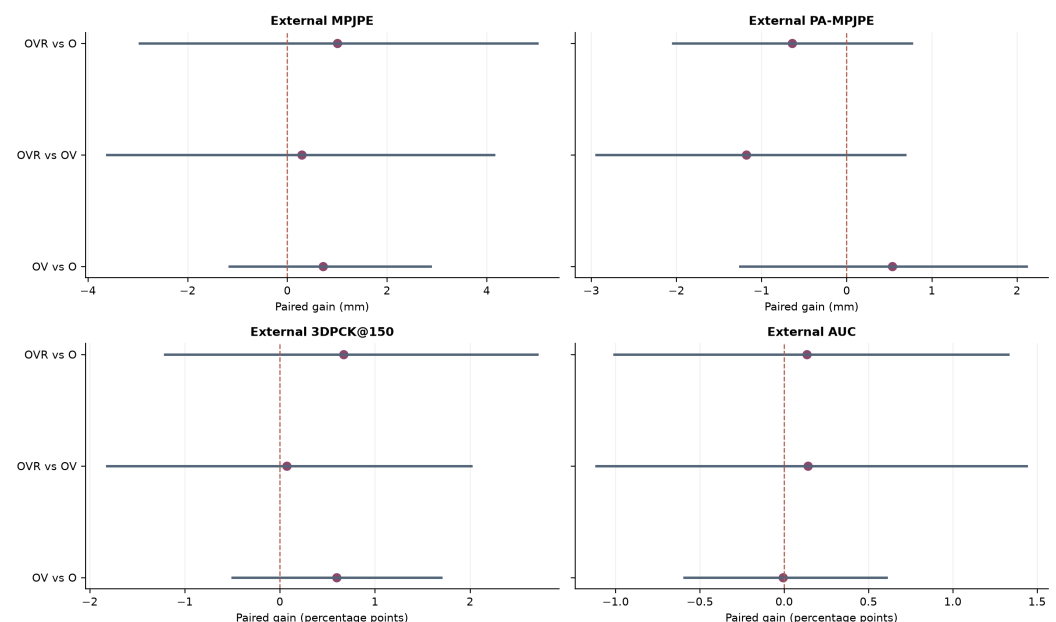


Figure 10. Frozen 3DPW paired point-and-interval contrasts for MPJPE, PA-MPJPE, 3DPCK@150, and AUC. The dashed vertical line denotes zero gain. Complete sequence clusters and matched split-seed units are resampled jointly; all intervals cross zero.

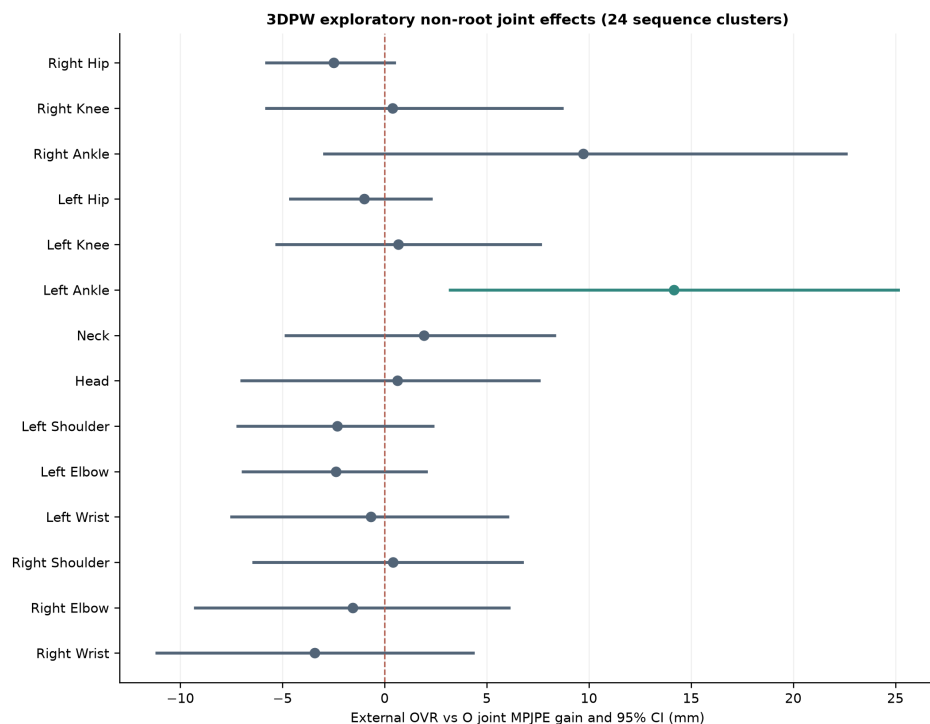


Figure 11. Exploratory 3DPW OVR-versus-O joint-level MPJPE improvements with two-level cluster-aware 95% intervals for all 14 non-root canonical joints. The dashed vertical line denotes zero joint-level improvement. The pelvis is excluded and corrected joint labels are used; this is not a Development joint-level factorial analysis.

5. Discussion

5.1. What Is Supported

The development evidence supports the full reliability-conditioned representation under matched detector and subject-disjoint evaluation. The result is not a single-seed observation: it aggregates three splits, three seeds, and two detectors, with checkpoint selection restricted to matched validation data. It also supports a more specific mechanistic statement: reliability-only OR produces the most stable independent contribution to development data.

The held-out MPI-INF-3DHP official result supports a narrower within-benchmark claim. OVR has better MPJPE and PA-MPJPE point estimates than O in both detector-specific pipelines, and the PA-MPJPE trend is larger than the MPJPE trend. Nevertheless, both primary 95% CIs include zero, and the relative MPJPE gain is below the predeclared 2% practical threshold. The method should therefore be described as showing a modest two-detector replication of the point-estimate direction, not as outperforming the baseline on the official partition or transferring to an unseen detector.

5.2. Why the Virtual View Is Not Yet Independently Established

An ungated virtual pose adds a second learned representation, but it does not add independent physical evidence. OV is close to O on development data and slightly worsens official MPJPE. Once reliability features are present, OVR improves on OR in point estimate, but the conditional virtual intervals still cross zero. The interaction is also unresolved. These results are compatible with at least three explanations: the -15° view contains a small useful geometric cue, the additional feature capacity acts as regularization, or the virtual branch helps only for a subset of poses and joints. The current factorial design rules out a strong general synergy claim but does not distinguish these weaker mechanisms. The frozen eight-angle screen supports the retained -15° validation choice, while all

five multi-view policies show that additional hypotheses did not improve primary MPJPE. The synchronized-camera experiment supports geometry-conditioned generator behavior at the component level, but its two-pair/two-cluster scope cannot establish full-pipeline physical-view generalization.

5.3. Why Calibration Matters

Raw confidence was substantially miscalibrated, particularly for YOLO11L-Pose. Monotonic calibration reduced validation ECE in every split-detector pair, and the residual-risk quality factor further penalized confidence bins associated with large geometric errors. Reliability, therefore, encodes both probability of correctness and expected error severity. The positive development OR effect indicates that this information is useful even without a virtual view. However, the official OR intervals cross zero, and external 3DPW OVR-versus-O PA-MPJPE moves in the unfavorable direction. Development calibration quality, therefore, does not guarantee an official or cross-dataset 3D improvement.

5.4. Position Relative to Recent Work

Recent diffusion and flow-matching methods model multiple 3D hypotheses directly [5,8], while multi-view consistency and Mocap-2-to-3 exploit richer spatial or temporal supervision [17,18]. Sun et al. [9] are more directly related because they generate virtual 2D poses for monocular lifting, while Hoang et al. [22] show that detector confidence can improve robustness. The present method is smaller and framewise, and its contribution is methodological rather than leaderboard-oriented: it treats a geometry-conditioned virtual pose as a reliability-conditioned auxiliary feature and uses a leakage-safe factorial protocol with real detector inputs to separate virtual-view information, reliability, and their interaction. A fair accuracy ranking would require common 2D detections, joint mapping, temporal context, training data, and aggregation; those conditions are not available here.

6. Limitations

First, external validation uses a single additional benchmark, one YOLO11L-Pose detector, three available arms, 24 sequences, and 13,579 eligible frames from 26,240 annotated frames; it is neither a complete four-arm/two-detector external factorial nor demonstrated transfer to an unseen detector. Its MPJPE and PA-MPJPE point directions are mixed, and all paired intervals include zero. Second, the eight-angle and five-policy experiments reuse frozen checkpoints and are descriptive validation-only screens, not independently trained angle policies or evidence of a globally optimal viewpoint. Third, physical generator validation contains only two directed camera pairs and two sequence clusters per row; its 24 stratified rows are not independent experiments and do not establish end-to-end RGB-detector validity. Fourth, the detector head is a nose proxy while the external target uses an anatomical SMPL head, and raw physical-camera 2D coordinates cannot be equated with detector pixels. Fifth, the pose lifter is framewise and compact, but the RGB detector and generator are excluded from its parameter counts, and no controlled end-to-end latency or FLOP benchmark was available. Sixth, the official partition has six subjects, both confirmatory intervals cross zero, and its approximately 0.97% relative MPJPE improvement is below the protocol-specific 2% practical-effect criterion. Development of all-arm joint-level tensors was not available, so joint-level findings are explicitly restricted to 3DPW.

7. Conclusions

Reliability-conditioned monocular pose lifting can use a detector-conditioned virtual 2D pose as an auxiliary hypothesis without treating it as an independent camera measurement. Across subject-disjoint development evaluation with two real detectors, the

complete OVR representation and the reliability-only OR arm improved the observed-only baseline with finalized whole-cluster confidence intervals, making calibrated residual-risk-aware reliability the most consistently supported factor. The conditional virtual increment and virtual–reliability interaction remain unresolved. Frozen validation favored the already locked -15° hypothesis among eight candidates, while none of five aggregation policies improved primary MPJPE; synchronized-camera annotations provided limited component-level support for two generator-view pairs. The six-subject Official partition and the independent 24-sequence 3DPW benchmark yielded non-confirmatory intervals, and the external PA-MPJPE direction was unfavorable. The supported contribution is therefore a compact, detector-aware pose-lifting methodology with transparent mechanistic attribution, validation sensitivity, bounded physical-generator evidence, and candid cross-dataset reporting—not established virtual-view dominance, practical official superiority, real-time whole-pipeline efficiency, or cross-dataset superiority.

Author Contributions: Conceptualization, P.N. and T.T.; methodology, P.N.; software, P.N.; validation, P.N.; formal analysis, P.N.; investigation, P.N.; data curation, P.N.; writing—original draft preparation, P.N.; writing—review and editing, P.N. and T.T.; visualization, P.N.; supervision, T.T.; project administration, P.N. and T.T.; funding acquisition, T.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research was financially supported by the TAIST–Tokyo Tech Program, a collaboration among the Thailand Advanced Institute of Science and Technology, the National Science and Technology Development Agency, and the Tokyo Institute of Technology; the Thammasat University Research Fund; the Center of Excellence in Intelligent Informatics, Speech and Language Technology and Service Innovation (CILS); the Intelligent Informatics and Service Innovation (IISI) Research Center; the Thailand Research Fund under grant numbers RTA6080013 and RTA6280015; the Thammasat University Fund for Research on Intelligent Informatics for Political Data Analysis; and the STEM Workforce Fund of the National Science and Technology Development Agency (NSTDA); and a research project supported by the Career for the Future Academy (CFA), National Science and Technology Development Agency (NSTDA), on behalf of Sirindhorn International Institute of Technology, Thammasat University, Thailand.

Institutional Review Board Statement: Not applicable. This study reanalyzed an existing benchmark dataset and involved no new collection of data from human participants.

Informed Consent Statement: Not applicable.

Data Availability Statement: The MPI-INF-3DHP dataset analyzed in this study is available from its original provider at <https://vcai.mpi-inf.mpg.de/3dhp-dataset/> (accessed on 2 September 2026). The additional 3DPW benchmark is described in Reference [26]. Both datasets remain subject to their providers' applicable license and access conditions. The source code, locked experimental configurations, supporting evidence, run provenance, and reproducibility materials associated with this study are publicly available at <https://github.com/EDENASINC/RCVV-Pose-Reproducibility> (accessed on 2 September 2026). The versioned release corresponding to this manuscript is available at <https://github.com/EDENASINC/RCVV-Pose-Reproducibility/releases/tag/v1.0.0-paper-information-4512788> (accessed on 2 September 2026). The original MPI-INF-3DHP and 3DPW images, videos, and annotations are not redistributed by the authors because they remain subject to the dataset provider's terms and conditions.

Acknowledgments: The authors thank the providers of the MPI-INF-3DHP dataset for making the benchmark data available to the research community.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

2D	Two-dimensional
3D	Three-dimensional
AUC	Area under the curve
3DPW	Three-Dimensional Poses in the Wild
CI	Confidence interval
ECE	Expected calibration error
MPJPE	Mean per-joint position error
NME	Normalized mean error
O	Observed-only arm
OR	Observed plus reliability arm
OV	Observed plus virtual-view arm
OVR	Observed plus virtual-view plus reliability arm
PA-MPJPE	Procrustes-aligned mean per-joint position error
PCK	Percentage of correct keypoints

References

1. Martinez, J.; Hossain, R.; Romero, J.; Little, J.J. A Simple Yet Effective Baseline for 3D Human Pose Estimation. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.
2. Pavllo, D.; Feichtenhofer, C.; Grangier, D.; Auli, M. 3D Human Pose Estimation in Video with Temporal Convolutions and Semi-Supervised Training. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
3. Rhodin, H.; Spörri, J.; Katircioglu, I.; Constantin, V.; Meyer, F.; Müller, E.; Salzmann, M.; Fua, P. Learning Monocular 3D Human Pose Estimation from Multi-View Images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018.
4. Kocabas, M.; Karagoz, S.; Akbas, E. Self-Supervised Learning of 3D Human Pose Using Multi-View Geometry. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
5. Shan, W.; Liu, Z.; Zhang, X.; Wang, Z.; Han, K.; Wang, S.; Ma, S.; Gao, W. Diffusion-Based 3D Human Pose Estimation with Multi-Hypothesis Aggregation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023.
6. Wehrbein, T.; Rudolph, M.; Rosenhahn, B.; Wandt, B. Probabilistic Monocular 3D Human Pose Estimation with Normalizing Flows. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021.
7. Xu, J.; Guo, Y.; Peng, Y. FinePOSE: Fine-Grained Prompt-Driven 3D Human Pose Estimation via Diffusion Models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024; pp. 561–570.
8. Wang, T.; Yu, X.; Mathis, M.W. FMPose3D: Monocular 3D Pose Estimation via Flow Matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Denver, CO, USA, 2–6 June 2026.
9. Sun, J.; Wang, M.; Zhao, X.; Zhang, D. Multi-View Pose Generator Based on Deep Learning for Monocular 3D Human Pose Estimation. *Symmetry* **2020**, *12*, 1116. [[CrossRef](#)]
10. Zhao, L.; Peng, X.; Tian, Y.; Kapadia, M.; Metaxas, D.N. Semantic Graph Convolutional Networks for 3D Human Pose Regression. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
11. Zhao, W.; Wang, W.; Tian, Y. GraFormer: Graph-Oriented Transformer for 3D Pose Estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022.
12. Zheng, C.; Zhu, S.; Mendieta, M.; Yang, T.; Chen, C.; Ding, Z. 3D Human Pose Estimation with Spatial and Temporal Transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021.
13. Zhang, J.; Tu, Z.; Yang, J.; Chen, Y.; Yuan, J. MixSTE: Seq2seq Mixed Spatio-Temporal Encoder for 3D Human Pose Estimation in Video. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022.
14. Zhu, W.; Ma, X.; Liu, Z.; Liu, L.; Wu, W.; Wang, Y. MotionBERT: A Unified Perspective on Learning Human Motion Representations. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023.

15. Chen, X.; Lin, K.-Y.; Liu, W.; Qian, C.; Lin, L. Weakly-Supervised Discovery of Geometry-Aware Representation for 3D Human Pose Estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
16. Wandt, B.; Rudolph, M.; Zell, P.; Rhodin, H.; Rosenhahn, B. CanonPose: Self-Supervised Monocular 3D Human Pose Estimation in the Wild. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021.
17. Ingwersen, C.K.; Tirsgaard, R.; Nylander, R.; Jensen, J.N.; Dahl, A.B.; Hannemose, M.R. Two Views Are Better than One: Monocular 3D Pose Estimation with Multiview Consistency. *arXiv* **2023**, arXiv:2311.12421.
18. Wang, Z.; Hu, Z.; Guo, R.; Pi, H.; Feng, Z.; Zhang, L.; Pei, M.; Huang, S. Mocap-2-to-3: Multi-View Lifting for Monocular Motion Recovery with 2D Pretraining. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Denver, CO, USA, 2–6 June 2026.
19. Jiang, T.; Lu, P.; Zhang, L.; Ma, N.; Han, R.; Lyu, C.; Li, Y.; Chen, K. RTMPose: Real-Time Multi-Person Pose Estimation Based on MMPose. *arXiv* **2023**, arXiv:2303.07399.
20. Ultralytics. YOLO11 Model Documentation. Available online: <https://docs.ultralytics.com/models/yolo11> (accessed on 4 August 2026).
21. Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K.Q. On Calibration of Modern Neural Networks. In Proceedings of the 34th International Conference on Machine Learning (ICML), Sydney, Australia, 6–11 August 2017; pp. 1321–1330.
22. Hoang, T.-H.; Zehni, M.; Phan, H.; Vo, D.M.; Do, M.N. Improving the Robustness of 3D Human Pose Estimation: A Benchmark Dataset and Learning from Noisy Input. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 17–18 June 2024; pp. 113–123. [[CrossRef](#)]
23. Gong, K.; Zhang, J.; Feng, J. PoseAug: A Differentiable Pose Augmentation Framework for 3D Human Pose Estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021.
24. Gholami, M.; Wandt, B.; Rhodin, H.; Ward, R.; Wang, Z.J. AdaptPose: Cross-Dataset Adaptation for 3D Human Pose Estimation by Learnable Motion Generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022.
25. Mehta, D.; Rhodin, H.; Casas, D.; Fua, P.; Sotnychenko, O.; Xu, W.; Theobalt, C. Monocular 3D Human Pose Estimation in the Wild Using Improved CNN Supervision. In Proceedings of the International Conference on 3D Vision (3DV), Qingdao, China, 10–12 October 2017.
26. von Marcard, T.; Henschel, R.; Black, M.J.; Rosenhahn, B.; Pons-Moll, G. Recovering Accurate 3D Human Pose in The Wild Using IMUs and a Moving Camera. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 601–617.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.