Supplementary Material for: VERA-Hand: A Verified Multidimensional Benchmark for 2D Hand Pose Estimation in Third-Person Dynamic Gesture Videos

Gibran Benitez-Garcia[1] and Hiroki Takahashi[1,2,3]

[1]Graduate School of Informatics and Engineering
[2]Artificial Intelligence eXploration Research Center (AIX)
[3]Meta-Networking Research Center (MEET)
The University of Electro-Communications, Tokyo, Japan

gibran@ieee.org, rocky@inf.uec.ac.jp

This supplementary material provides methodological details and extended results that support the main paper. Section A reports the annotation audit and characterizes the audited sample. Section B documents the estimator configurations and keypoint conversions required for reproducibility. Section C provides the complete motion-blur, hand-region, temporal-stability, annotation-bias, and benchmark size sensitivity analysis. Section D further characterizes the Sapiens family and the conditions under which cross-hand overlap occurs. Section E specifies the post-hoc cleanup procedure and presents its ablation. Section F presents qualitative comparisons between GT and the estimator predictions. Unless otherwise stated, supervised results use the bias-free subset; localization, identification, and Pose-F1 are macro-averaged over the left and right hands, whereas detection is label-agnostic. Temporal analyses are computed over the complete videos.

A. Annotation Audit

To assess the reliability of the manually verified annotations, we conducted two independent audits, each carried out by a reviewer who did not participate in the keypoint annotation. The first inspected 500 frames and the second a further 3,000, for a total of 3,500 audited frames spanning all 13 subjects and all 52 videos. Each frame was examined for four categories of error: hand presence, handedness label, gross pose configuration, and local keypoint placement. The audits were corrective rather than diagnostic: every issue identified in either pass was fixed, and the corrected annotations are the ones released and used throughout the paper.

The audited set is deliberately not a proportional sample (Table 5). Because left and bimanual hands are rare in VERA-Hand, the audit covers all 72 left-only frames and all 183 bimanual frames, together with roughly one third of the right-only frames. Left hands therefore account for 6.9% of the audited hands against 2.4% of the benchmark, so the rare and most error-prone configurations are inspected exhaustively rather than in proportion. Coverage of out-of-frame keypoints (12.5% of audited frames against 14.1%

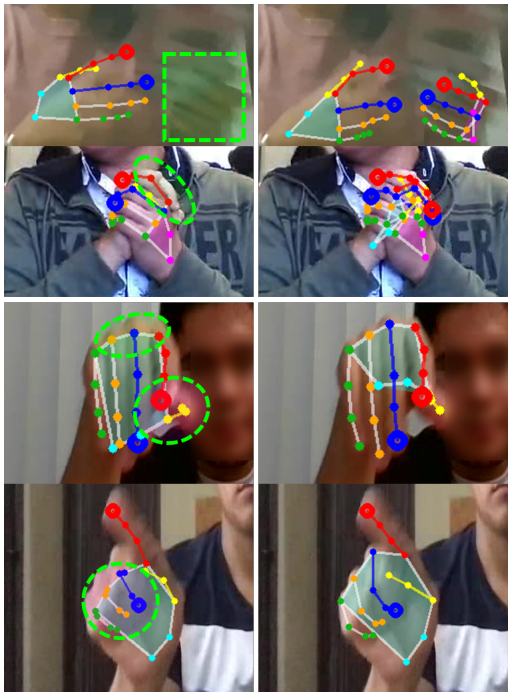


Figure 4. Representative corrections from the audit. Each row shows one frame before (left) and after (right) correction, with green markers indicating the affected region. The top two rows show previously missed hands that were added; the remaining rows show local keypoint displacements.

Property	Audited (3,500)	Full benchmark
Subjects / Videos	13 / 52	13 / 52
Frames (right / left / both)	3,245 / 72 / 183	10,252 / 72 / 183
Frames without hands	0	261
Hands (right / left)	3,428 / 255	10,435 / 255
Frames w/out-of-frame kpts	437	1,477
Blur (S / M / B)	1,447 / 1,105 / 948	4,338 / 3,379 / 2,790

Table 5. Composition of the audited set relative to the full benchmark, where S, M, and B denote the sharp, medium, and blurry classes. All left-only and bimanual frames are audited, so the rare configurations are covered exhaustively rather than proportionally. Blur is counted per frame over the frames containing hands, using the right-hand class for bimanual frames; this convention is used only for this descriptive comparison, since all results in the paper are computed per hand.

of the benchmark) and of the motion-blur classes remains close to the benchmark composition. The 261 hand-free frames fall outside this audit, since they contain no keypoint annotations for inspection; each was confirmed to contain no visible hand at selection time.

Across the 3,500 audited frames, 268 (7.7%) required a correction. Of these, 263 were small displacements of individual keypoints and five were previously missed hands that were added; no errors were found in handedness labels or gross pose configuration. Residual uncertainty is therefore concentrated in fine-grained localization rather than in hand-level validity, and the two passes agree on its scale (7.2% and 7.7% of their respective frames). Figure 4 shows representative corrections.

B. Estimator Setup and Keypoint Conversion

Each estimator is evaluated using its official off-the-shelf configuration, including the default detector, confidence thresholds, preprocessing, and input resolution, without dataset-specific tuning. The MMPose pipelines [38] use the RTMPose-m Hand5 checkpoint for RTMP-hand [40] at 256×256 and the RTMW-x Cocktail14 checkpoint for RTMW-wb [41] at 384×288. Sapiens2 [43] uses the 1B Goliath pose checkpoint at 1024×768, while MP-hands [47] and MP-holistic [39] use MediaPipe release 0.10.32. WiLoR [45] is evaluated using its publicly released checkpoint. The detectors and per-model details are summarized in Table 2 of the main paper.

Handedness is obtained differently across pipelines, which is relevant to the identification results. MP-holistic, RTMW-wb, and Sapiens2 encode the left and right hands using fixed output keypoint indices. MP-hands provides a handedness label through its native classifier, whereas WiLoR uses the side class predicted by its YOLO-based hand detector. RTMP-hand is the only evaluated pipeline whose detector returns hand bounding boxes without a side label, requiring handedness to be assigned post hoc. For this purpose, we use RTMPose-L-WholeBody from the same MMPose framework, with a 384×288 input and the same RTMDet [44] person detector used for RTMW-wb. This auxiliary pipeline is used exclusively to obtain left- and right-wrist anchors and does not alter the hand keypoints predicted by RTMP-hand. Each RTMP-hand detection is assigned the label of the nearest compatible wrist anchor. Anchors located inside the hand box are preferred, followed by anchors lying within one box diagonal of its center. When no suitable anchor is available, handedness is inferred from the predicted hand geometry using a palm cross-product. Because RTMP-hand relies on an inferred rather than native handedness label, this additional assignment step may contribute to its comparatively high label-swap rate.

All native outputs are mapped to the common 21-keypoint MediaPipe topology. Most mappings require only index selection, but Sapiens2 and WiLoR require additional transformations. In the Sapiens2 Goliath layout, each finger is ordered from tip to base, and the wrist follows the finger joints; we reverse each finger block to the MediaPipe base-to-tip order and move the wrist to the first index. WiLoR predicts 21 MANO [46] joints in 3D, which are projected onto the image plane using the model-predicted camera. For left hands, the projected coordinates are mirrored horizontally to match the annotation convention.

C. Extended Quantitative Results

This section extends the analyses summarized in the main paper by reporting complete results for all six estimators. We first examine performance across motion-blur classes and hand regions, followed by detailed temporal-stability results and an analysis of initialization bias.

C.1. Motion blur

Table 6 reports the motion-blur stratification for all six estimators, extending the four representative pipelines shown in Figure 3 of the main paper. Representative examples of the sharp, medium, and blurry classes are shown in Figure 5. Blur is assigned independently to each annotated hand, whereas the frame-level assignment in Table 5 is used only to describe the audited sample. All supervised results in this analysis are therefore computed at the hand level.

Detection is nearly invariant to blur for MP-holistic, Sapiens2-1B, and WiLoR, whereas the hand-specific pipelines degrade the most (RTMP-hand drops by 0.19 and MP-hands by 0.13) and RTMW-wb shows an intermediate decline. PCK and Pose-F1 degrade for all estimators and the rankings cross. Identification is largely unaffected by blur, and label-swap rates show no common blur-dependent trend. These results show that robustness to motion blur de-



Figure 5. Representative hand crops from the three motion-blur classes, one class per column (left: sharp; middle: medium; right: blurry). Blur is assigned per hand from optical-flow magnitude and hand-region sharpness.

pends on both the estimator and the evaluation axis, with no pipeline remaining uniformly robust across all metrics.

C.2. Hand region

Localization by hand region is evaluated on the right hand over correctly matched predictions. As shown in Table 7, the wrist provides the clearest separation among estimators, with PCK ranging from 0.16 to 0.57. The differences are smaller over the fingertips and the whole hand. The pattern is estimator-dependent: Sapiens2-1B performs substantially worse at the wrist while remaining competitive over the palm and whole hand, whereas WiLoR obtains its lowest regional score over the palm.

C.3. Temporal stability

All temporal metrics are computed on the right hand, the side that performs the gestures; the sparse left hand carries too little continuous motion to be measured meaningfully. Three patterns stand out. Motion energy E_v tracks jitter: the steadiest estimators also predict the least motion, so part of their stability reflects a low-motion prior rather than tracking accuracy (Table 8). The estimator ranking is largely consistent across keypoint subsets, with the fingertips exhibiting higher jitter than the base joints, indicating that temporal instability increases toward the distal keypoints (Table 9). Finally, the cleanup decreases the number of tracking gaps by approximately 37% to 67%, while J_{jerk} increases slightly in every case (Table 10). This effect scales with the number of gaps filled: the gap-heavy estimators gain the most continuity and pay the most in jitter, while the gap-light ones barely change. The post-hoc interpolation thus trades gap continuity for a small loss of smoothness.

C.4. Annotation bias

Table 11 quantifies the effect of evaluating MP-holistic on the bias-free subset rather than on the full annotated set. Within the bias-free subset, manual correction moved 76% of the keypoints by more than α=0.02, with a median normalized displacement of 0.06, compared with 0.008 for biased hands. Relative to the initialization, approximately 1,381 missing hands were added, and 1,387 spurious detections were removed from the corresponding frames. Evaluating MP-holistic on the full set rather than the bias-free subset primarily inflates localization performance, with substantially better PCK and MPJPE. Pose-F1 also increases, whereas Det-F1 and the label-swap rate improve only moderately. These results show that the initialization advantage is concentrated mainly in keypoint localization, which supports reporting the main supervised comparison on the bias-free subset.

C.5. Benchmark size sensitivity

To assess whether the reported comparisons depend on the size of the benchmark, we recompute the supervised metrics on nested subsets of the annotated frames. Subsampling is deterministic and applied per video: frames are grouped by video and ordered chronologically, and for a target fraction we retain $k = \text{round}(n \cdot \text{pct}/100)$ equispaced indices over that ordering, always including the first and last annotated frame of the clip. Sampling within each video preserves the proportion of frames across videos and subjects, and the equispaced selection keeps the subset distributed over the whole clip rather than concentrated at its beginning.

The relative ordering of the estimators is preserved at every fraction on all four metrics, down to 10% of the benchmark (Table 12). Detection is the most stable axis, varying by less than 0.01 across all fractions, which is expected since it is computed over every annotated hand. Cross-hand overlap is equally stable: the Sapiens2 failure remains above 82% at every fraction and stays two orders of magnitude above the other estimators, so the headline finding is not a consequence of benchmark size. Localization and Pose-F1 fluctuate more at the smallest fractions, where far fewer matched hands remain; since the subsampling is deterministic, these fluctuations reflect the particular frames retained rather than variance across repeated draws. The one effect that does depend on size is the measurement of rare failures: WiLoR's cross-hand overlap, at 0.07% on the full benchmark, reaches exactly zero at 10%, where too few frames remain for such infrequent events to appear at all.

Estimator	Det-F1 ↑			PCK ↑			Pose-F1 ↑			Swap ↓		
	S	M	B	S	M	B	S	M	B	S	M	B
MP-holistic	0.841	0.835	0.845	0.485	0.424	0.397	0.356	0.303	0.276	0.57%	0.53%	0.42%
MP-hands	0.941	0.892	0.812	0.654	0.523	0.452	0.494	0.350	0.267	3.28%	4.98%	5.49%
WiLoR	0.987	0.980	0.968	0.520	0.469	0.402	0.506	0.394	0.307	0.11%	0.09%	0.14%
RTMW-wb	0.968	0.931	0.881	0.634	0.542	0.463	0.568	0.431	0.390	0.25%	0.41%	0.28%
RTMP-hand	0.938	0.873	0.751	0.532	0.482	0.409	0.418	0.328	0.218	3.99%	3.66%	2.58%
Sapiens2-1B	0.668	0.666	0.664	0.554	0.477	0.413	0.407	0.363	0.290	0.05%	0.00%	0.03%

Table 6. Performance across motion-blur classes for all six estimators. S, M, and B denote sharp, medium, and blurry instances, respectively. PCK and Pose-F1 decrease consistently with blur, whereas detection and label-swap rates exhibit estimator-dependent patterns. Results use per-hand blur assignments and the bias-free subset; localization, Pose-F1, and identification are macro-averaged, whereas detection is label-agnostic. PCK is evaluated at α=0.10, and Pose-F1 uses α=0.10 and β=0.5.

Estimator	Wrist	Palm	Fingertips	Whole
MP-holistic	0.552	0.622	0.607	0.612
MP-hands	0.446	0.691	0.683	0.688
WiLoR	0.394	0.359	0.538	0.529
RTMW-wb	0.571	0.585	0.577	0.628
RTMP-hand	0.346	0.546	0.552	0.611
Sapiens2-1B	0.155	0.609	0.515	0.624

Table 7. PCK@0.10 by hand region for the right hand. Localization is computed over correctly matched predictions. The wrist exhibits the largest cross-estimator spread.

Estimator	Gaps ↓	J_{acc} ↓	J_{jerk} ↓	E_v
MP-holistic	3340	1.159	1.599	1.524
MP-hands	4515	1.256	1.693	1.511
WiLoR	**615**	**0.847**	**1.181**	**1.375**
RTMW-wb	2357	1.043	1.468	1.381
RTMP-hand	5221	1.312	1.764	1.427
Sapiens2-1B	831	0.937	1.302	1.399

Table 8. Temporal stability of the raw right-hand predictions over the complete videos. J_{acc}, J_{jerk}, and E_v are medians over the videos scaled by 100; Gaps is the total count of gaps. E_v (motion energy) contextualizes jitter and is not a quality metric.

Estimator	Whole ↓	Base ↓	Tips ↓
MP-holistic	1.599	1.240	1.981
MP-hands	1.693	1.292	2.068
WiLoR	**1.181**	**0.899**	**1.617**
RTMW-wb	1.468	0.990	2.049
RTMP-hand	1.764	1.386	2.296
Sapiens2-1B	1.302	0.955	1.805

Table 9. J_{jerk} by keypoint subset for the raw right-hand predictions: whole hand, base joints (wrist and MCPs), and fingertips. The fingertips are the least stable subset for every estimator.

Estimator	Gaps ↓		J_{jerk} ↓	
	raw	clean	raw	clean
MP-holistic	3340	1253	1.599	1.674
MP-hands	4515	**1481**	1.693	1.801
WiLoR	615	303	1.181	1.199
RTMW-wb	2357	966	1.468	1.517
RTMP-hand	5221	**2153**	1.764	1.880
Sapiens2-1B	831	524	1.302	1.309

Table 10. Effect of the full cleanup on right-hand temporal stability. The cleanup substantially reduces tracking gaps for every estimator but slightly increases J_{jerk}. Bold marks the two largest reductions in tracking gaps.

Set	Det-F1 ↑	PCK ↑	Pose-F1 ↑	Swap ↓	MPJPE ↓
Bias-free	0.840	0.439	0.309	0.51%	0.265
Full set	0.869	0.579	0.398	0.35%	0.202

Table 11. MP-holistic evaluated on the bias-free subset and the full annotated set. Including hands that remain close to the MP-holistic initialization substantially improves PCK, MPJPE, and Pose-F1, while also improving Det-F1 and the label-swap rate. MPJPE is normalized by hand size.

D. Sapiens Family Analysis

The cross-hand overlap reported in the main paper, computed as the percentage of annotated GT hand instances affected by an overlapping opposite-handed prediction, increases sharply from Sapiens1-1B [42] to Sapiens2-1B [43] (Table 13). Sapiens2-1B achieves substantially better localization and a lower label-swap rate, yet its raw Det-F1 falls well below that of Sapiens1-1B because the opposite-handed phantom predictions reduce precision.

Table 14 groups the annotated GT hand instances according to whether their source frame contains one or both hands, while Figure 6 shows representative predictions. Cross-hand overlap is rare when both hands are present, but

Metric	100%	75%	50%	25%	10%
Sapiens2-1B					
Det-F1 ↑	0.666	0.665	0.665	0.665	0.666
PCK ↑	0.491	0.490	0.481	0.495	0.488
Pose-F1 ↑	0.361	0.360	0.359	0.358	0.360
Cross-hand ↓	85.44%	85.74%	85.59%	85.78%	82.82%
WiLoR					
Det-F1 ↑	0.978	0.978	0.977	0.975	0.974
PCK ↑	0.470	0.475	0.463	0.454	0.431
Pose-F1 ↑	0.407	0.409	0.400	0.414	0.385
Cross-hand ↓	0.07%	0.10%	0.08%	0.11%	0%
RTMW-wb					
Det-F1 ↑	0.932	0.931	0.934	0.935	0.926
PCK ↑	0.560	0.562	0.541	0.550	0.581
Pose-F1 ↑	0.477	0.477	0.460	0.470	0.496
Cross-hand ↓	0.91%	0.88%	0.90%	0.76%	0.99%

Table 12. Sensitivity to benchmark size for three representative estimators, spanning the foundation, reconstruction-based, and whole-body paradigms. Metrics are recomputed on deterministic per-video subsets retaining the indicated fraction of annotated frames. The ordering of the estimators is preserved at every fraction on all four metrics, including the cross-hand overlap that characterizes Sapiens2.



Figure 6. Qualitative comparison of Sapiens2-1B (left column) and Sapiens1-1B (right column). In the top two rows, only Sapiens2-1B produces a cross-hand overlap over the single visible hand, yet Sapiens2 localizes the keypoints markedly better. In the bottom row, Sapiens2-1B predicts both correctly, whereas Sapiens1-1B produces the overlap on the right hand and misses the heavily blurred left one.

becomes dominant for Sapiens2-1B in single-hand frames. This pattern indicates that the failure primarily consists of an opposite-handed duplicate prediction over the visible hand rather than a failure to recover the visible hand itself. The behavior may reflect a full-body prediction prior that favors bilateral hand outputs even when only one hand is visible. Because VERA-Hand consists predominantly of single-hand frames, it provides a setting in which this failure becomes directly measurable.

The cleanup results further isolate the source of the performance loss. The spatial stage removes the cross-hand overlap entirely and raises Det-F1 from 0.666 to 0.929, while PCK and MPJPE remain effectively unchanged. The temporal stage provides a smaller additional gain in detection and removes the remaining duplicate-hand predictions. These results indicate that the dominant Sapiens2-1B failure in this benchmark is spurious hand recovery rather than inaccurate localization of the visible hand.

E. Post-hoc Cleanup: Method and Ablation

The cleanup operates exclusively on the predicted hands and uses no GT, image content, or localization metric. It comprises a spatial stage that resolves overlapping opposite-handed predictions within individual frames and a temporal stage that enforces short-term consistency across consecutive frames.

Spatial stage. For every pair of opposite-handedness predictions in a frame, we compute the bounding-box IoU (from the valid keypoints) and the center distance normalized by the larger box diagonal. The pair is treated as a cross-hand overlap when IoU ≥ 0.80 and the normalized center distance ≤ 0.15, in which case the hand with the higher mean keypoint confidence is kept and the other removed. The spatial stage primarily addresses the cross-hand overlap problem.

Temporal stage. The per-video timeline is independently processed by hand over a sliding odd window of 7 frames (a radius of 3) using a binary presence signal, with three components. Deduplication retains at most one prediction per hand label in each frame, selecting the prediction with the most valid keypoints. Phantom removal discards a prediction supported by fewer than 2 neighbors in its window, treating it as a non-persistent detection. Gap recovery fills an absent side that has at least 4 supporting neighbors, with support on both temporal directions, by linearly interpolating the keypoints between the nearest present frames; it is restricted to gaps of at most 3 frames and skipped when the motion between the endpoints is too large, to avoid in-

Variant	Cross-hand ↓	Dup. ↓	Det-F1 ↑	PCK ↑	Pose-F1 ↑	Swap ↓	MPJPE ↓
Sapiens1-1B	14.15%	2.95%	0.837	0.360	0.221	0.51%	0.230
Sapiens2-1B	85.44%	1.85%	0.666	0.491	0.361	0.03%	0.172
+ spatial	0%	0.98%	0.929	**0.492**	0.402	0.09%	**0.171**
+ spatial + temporal	0.05%	0%	**0.940**	0.490	**0.406**	0.03%	0.175

Table 13. Comparison of the Sapiens family and the effect of applying the cleanup to Sapiens2-1B. The cross-hand overlap worsens from Sapiens1 to Sapiens2, and the two-stage cleanup applied to Sapiens2 (spatial, then spatial plus temporal) removes it and recovers detection performance while leaving PCK and MPJPE nearly unchanged.

Frame context	GT hands	Sapiens1-1B	Sapiens2-1B
Both hands present	366	2.25%	2.25%
Single hand present	10,324	14.56%	**88.35%**
All annotated hands	10,690	14.15%	85.44%

Table 14. Cross-hand overlap grouped by whether the source frame contains one or both hands. Overlap is rare when both hands are present, but it explodes in single-hand frames for Sapiens2-1B.

terpolating across fast movement. The complete cleanup applies the spatial stage first, followed by all three components of the temporal stage.

Ablation. We ablate the temporal stage on MP-hands, whose gain is purely temporal since it has no cross-hand overlap, and compare against a majority-vote baseline that forces each side present whenever it appears in more than half of the window, copying the nearest hand without interpolation. Table 15 shows that deduplication removes the duplicate predictions but produces little change in the evaluation scores. Gap recovery provides the main individual gain in detection, although it also reduces PCK by introducing interpolated predictions, whereas phantom removal improves PCK and Pose-F1 at a small cost in detection by discarding non-persistent detections. Combining the three components outperforms every individual variant on both Det-F1 and Pose-F1, indicating that the components are complementary and that the detection gains outweigh the accompanying localization change. The majority-vote baseline, which forces presence without interpolation, remains below the full temporal stage on all three metrics.

Together, these stages produce the cleaned results reported throughout the paper. The spatial stage drives the Sapiens2-1B recovery shown in Table 13, while the temporal stage produces the reduction in tracking gaps and the accompanying jitter increase reported in Table 10. Applying both stages yields the raw-versus-clean comparison presented in Table 4 of the main paper.

F. Qualitative Results

This section complements the quantitative analysis with side-by-side comparisons between the ground truth and the

Variant	Det-F1 ↑	PCK ↑	Pose-F1 ↑	Dup. ↓
Raw	0.898	0.573	0.389	1.48%
+ Deduplication	0.899	0.570	0.390	0%
+ Phantom removal	0.879	**0.580**	0.430	1.47%
+ Gap recovery	0.913	0.556	0.400	1.48%
+ Full temporal stage	**0.926**	0.559	**0.442**	0%
Majority baseline	0.917	0.537	0.436	1.46%

Table 15. Temporal-stage ablation on MP-hands using the bias-free subset. Individual components are applied independently to the raw predictions, while the full temporal stage applies all three. The full temporal stage achieves the highest Det-F1 and Pose-F1, while phantom removal yields the highest PCK.

predictions of the four representative estimators shown in Figure 3 of the main paper. Each row shows the same frame under all five conditions, in the order GT, MP-hands, WiLoR, RTMW-wb, and Sapiens2-1B. Figure 7 shows frames on which the estimators largely agree with the GT, while Figure 8 shows frames on which one or more pipelines fail.

The two figures illustrate the separation between the evaluation axes. In Figure 7, every hand is recovered by every pipeline and the remaining discrepancies are confined to individual keypoints, typically at the fingertips or the wrist, which is the regime that localization metrics such as PCK are designed to capture and the only one visible to a localization-only comparison. Figure 8 shows the failure modes that such a comparison cannot express: hands that are missed entirely, spurious predictions placed on the forearm or the torso, and gross pose configurations that are wrong even though a hand was found at approximately the right location. The last row is an example of a no-hand frame, in which two pipelines still emit a hand; these frames contribute no localization measurement at all and are only scored by the detection axis.

The examples also show that failures are not evenly distributed across pipelines. A frame that is difficult for one estimator is frequently handled correctly by another, and no column is consistently the weakest across rows, which is the qualitative counterpart of the ranking reversals reported in



Figure 7. Qualitative comparison on frames where the estimators largely agree with the GT. Columns, from left to right: GT, MP-hands, WiLoR, RTMW-wb, and Sapiens2-1B. Every hand is recovered by every pipeline, and green dashed ellipses mark the residual localization deviations, which remain confined to individual keypoints.

Section 4 of the main paper. Motion blur, hands overlapping the face, and hands that partially leave the frame account for most of the cases in Figure 8, consistent with the stratified results in Section C.

References

[38] MMPose Contributors. Openmmlab pose estimation toolbox and benchmark. https://github.com/open-mmlab/mmpose, 2020. 2

[39] Ivan Grishchenko and Valentin Bazarevsky. Mediapipe holistic: Simultaneous face, hand and pose prediction, on device. *Google AI Blog*, 2020. 2

[40] Tao Jiang, Peng Lu, Li Zhang, Ningsheng Ma, Rui Han, Chengqi Lyu, Yining Li, and Kai Chen. Rtmpose: Real-time multi-person pose estimation based on mmpose. *arXiv preprint arXiv:2303.07399*, 2023. 2

[41] Tao Jiang, Xinchen Xie, and Yining Li. Rtmw: Real-time multi-person 2d and 3d whole-body pose estimation. *arXiv preprint arXiv:2407.08634*, 2024. 2

[42] Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. Sapiens: Foundation for human vision models. In *ECCV*, pages 206–228, 2024. 4

[43] Rawal Khirodkar, He Wen, Julieta Martinez, Yuan Dong, Su Zhaoen, and Shunsuke Saito. Sapiens2. In *ICLR*, 2026. 2, 4

[44] Chengqi Lyu, Wenwei Zhang, Haian Huang, Yue Zhou, Yudong Wang, Yanyi Liu, Shilong Zhang, and Kai Chen. Rtmdet: An empirical study of designing real-time object detectors. *arXiv preprint arXiv:2212.07784*, 2022. 2

[45] Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. Wilor: End-to-end 3d hand localization and reconstruction in-the-wild. In *CVPR*, pages 12242–12254, 2025. 2

[46] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM TOG*, 36(6), 2017. 2

[47] Fan Zhang, Valentin Bazarevsky, Andrey Vakunov, Andrei Tkachenka, George Sung, Chuo-Ling Chang, and Matthias



Figure 8. Qualitative comparison on frames where one or more estimators fail. Columns, from left to right: GT, MP-hands, WiLoR, RTMW-wb, and Sapiens2-1B. Green dashed rectangles mark detection or identification errors, and green dashed ellipses mark gross localization errors on a recovered hand. The last row shows a frame without hands in which two pipelines still emit a prediction. Failures are pipeline-dependent: a frame that defeats one estimator is often handled correctly by another, and the errors shown here are invisible to an evaluation restricted to matched predictions.

Grundmann. Mediapipe hands: On-device real-time hand tracking. *arXiv preprint arXiv:2006.10214*, 2020. 2