EvUBody: Resolving Kinematic Ambiguity in Event-Assisted Joint Upper-Body and Hand Reconstruction

Anonymous authors
Paper under double-blind review

Abstract

Extracting precise human manipulation trajectories from video is critical for scaling embodied robot learning. While sparse RGB frames frequently miss high-frequency kinematics and suffer from motion blur, event streams have emerged as a powerful modality to capture asynchronous dynamic changes. However, existing event-based paradigms are overwhelmingly restricted to isolated hand motion, failing to account for the macro upper-body movements that induce complex, ambiguous event patterns across the hand during real-world manipulation. Transitioning to a joint upper-body and hand framework is strictly necessary, yet it has been severely bottlenecked by a fundamental data collection conflict: obtaining precise, continuous 3D joint annotations typically requires invasive wearable sensors or dense optical markers, which destroy the natural visual appearance of the human and suffer from severe occlusion during hand-object interaction. Furthermore, merging separated body and hand tracking systems introduces persistent spatial discontinuities at the wrist. To solve this, we introduce EvUBody, the first comprehensive dataset and benchmark that successfully balances natural RGB-Event visual observation with high-fidelity, spatially aligned 3D kinematic references for the coupled arm-and-hand system. Comprising both simulated and real-world subsets, EvUBody enables our proposed causal reconstruction framework, which fuses sparse RGB spatial anchors with continuous event streams via joint spatio-temporal modeling. By evaluating across varying frame rates and lighting conditions, we demonstrate that jointly modeling the upper body resolves event ambiguities, allowing our framework to recover true manipulation trajectories that RGB-only priors and hand-isolated approaches fail to capture.

1 Introduction

Human demonstrations offer a rich source of manipulation experience for embodied learning, motivating approaches that learn robot policies and vision–language–action models from human videos (Lepert et al., 2025; Yang et al., 2025). For tabletop manipulation, finger articulation alone is insufficient to characterize how an action is executed: arm movements transport the hands, wrist poses determine their placement and orientation, and finger joints enable precise grasping and manipulation. Coarse body skeletons omit these fine-grained details, while hand-local representations do not fully describe how the hands are positioned and coordinated with the upper limbs throughout an operation (Huang et al., 2024). Faithful reconstruction therefore calls for a joint representation of upper-body and articulated two-hand motion, preserving the coordinated trajectory from reaching and grasping to sustained manipulation.

Acquiring reference data for such coordinated motion is itself challenging. The difficulty lies in obtaining detailed, temporally continuous, and spatially consistent annotations while preserving the appearance of uninstrumented manipulation. Capture suits and data gloves alter visible body and hand appearance, while combining tracking devices requires calibration and registration of their coordinate systems, as illustrated by DexCap (Wang et al., 2024a). Marker-based optical capture faces a different difficulty: occlusion during close interactions can cause missing tracks or ambiguous marker identities, increasing manual cleanup. These limitations and the associated post-processing burden are documented in MAMMA (Cuevas Velasquez et al., 2026). Together, they make it costly

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

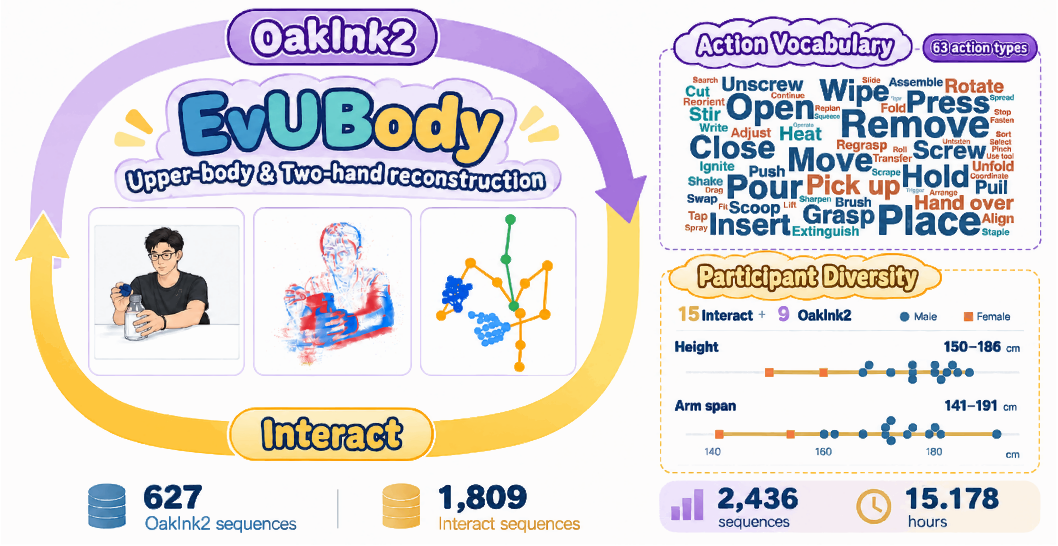

Figure 1: Overview of EvUBody, a dataset and benchmark for joint upper-body and two-hand reconstruction from sparse RGB and event streams. EvUBody combines 627 EvUBody-OakInk2 sequences augmented with simulated events and 1,809 EvUBody-Interact sequences captured with real RGB–event sensors and multiview markerless reference annotations, totaling 2,436 sequences and 15.178 hours. The two subsets, OakInk2 and Interact, include 9 and 15 participants, respectively. The central examples illustrate RGB appearance, event observations, and the target body–hand representation. The right panels summarize a vocabulary of 63 manipulation actions and participant diversity, with reported heights of 150–186 cm and arm spans of 141–191 cm.

to acquire jointly annotated body–hand trajectories throughout complete operations while retaining natural body and hand appearance.

Obtaining reference trajectories, however, does not eliminate the difficulty of recovering them from limited visual observations. Sparse RGB frames capture only selected moments, so similar observed poses can be consistent with different intermediate trajectories. Brief finger adjustments or reversals of wrist motion may consequently remain unobserved. Low-light imaging introduces a further trade-off: shorter exposures reduce the available signal, whereas longer exposures can increase motion blur (Kim et al., 2024). Temporal models can use motion priors to infer plausible transitions, but a plausible trajectory need not match the executed action. The central challenge is therefore missing or degraded motion evidence, rather than merely an insufficient output frame rate.

Event cameras offer complementary observations within these intervals. They asynchronously record timestamped changes in log intensity with fine temporal resolution and high dynamic range. RGB provides appearance and spatial context, whereas events constrain when and where visible changes occur between frames. Events are not direct measurements of 3D motion, however: regions with little intensity change provide few events, while illumination changes can generate responses unrelated to subject motion (Gallego et al., 2020). Existing resources address complementary aspects of this problem. OakInk2 and EgoDex provide annotated manipulation recordings (Zhan et al., 2024; Hoque et al., 2026); event-based work explores RGB–event hand reconstruction, motion capture, and robotic action prediction (Jiang et al., 2024b; Wu et al., 2026b; Zhai et al., 2026). Studying event assistance at the level of complete manipulation demonstrations requires RGB–event observations paired with joint upper-body and detailed two-hand reference trajectories, together with an evaluation protocol designed for this setting.

We introduce **EvUBody**, a dataset and benchmark for RGB–event reconstruction of coordinated upper-body and articulated two-hand motion. **EvUBody-OakInk2** augments existing manipulation recordings with simulated events, while **EvUBody-Interact** provides newly captured tabletop manipulation from 15 participants under normal and low-light conditions. The real capture inventory contains 1,809 video sequences totaling approximately 5.88 hours and covers 120 tasks in 12 cat-

egories. A MAMMA-based multi-view pipeline (Cuevas Velasquez et al., 2026) generates joint 3D reference annotations independently of the DVSync reconstruction inputs. Task metadata and sequence-level language descriptions retain semantic context alongside motion; they are not model inputs. Capture statistics and annotation validation are distinguished in Section 3.

Using EvUBody, we study offline reconstruction under causal input constraints: each target pose uses only RGB observations at or before that timestamp and events preceding it. Our model combines *RGB-anchored hand regions*, quality-conditioned cross-modal fusion, and joint-level spatiotemporal modeling. Predictions follow the native 30 Hz RGB-aligned annotation timeline. This inference constraint is distinct from the offline procedures used to generate simulated events and multi-view reference annotations.

We evaluate reconstruction on EvUBody-OakInk2 and EvUBody-Interact under simulated-event and real-sensor conditions, assessing body and hand poses according to each method's supported outputs. Further experiments on EvUBody-OakInk2 use an independently trained RGB-only baseline, architectural variants, and RGB exposure perturbations to examine the contributions of event information and model design. Separate evaluation at timestamps without current RGB observations, together with interval displacement error, characterizes inter-frame motion recovery. The results support the complementary value of events for reconstructing coordinated body–hand motion from sparse RGB observations.

Our contributions are threefold:

- **EvUBody: Joint Motion Data with Natural Appearance.** We introduce EvUBody, a dataset and benchmark for event-assisted upper-body and hand motion reconstruction. Its simulated and real-world subsets pair RGB–event observations with joint 3D reference trajectories, using markerless multiview capture for real recordings to preserve natural appearance and avoid registering separate body and hand tracking systems.
- **Causal Body–Hand Reconstruction.** We propose a causal framework that jointly models upper-body motion and hand articulation. RGB-anchored hand regions, quality-conditioned cross-modal fusion, and joint-level spatiotemporal modeling integrate sparse appearance observations with intervening events to reconstruct coordinated motion.
- **Targeted Evaluation of Inter-Frame Reconstruction.** We establish a temporally constrained protocol that separately evaluates timestamps without current RGB observations. Comparisons with an independently trained RGB-only baseline and evaluations under RGB exposure perturbations quantify the contribution of events to pose accuracy and inter-frame motion recovery.

2 Related Work

Manipulation data and joint motion capture. Human videos support imitation learning and VLA pretraining through shared representations, action supervision, and human–robot transfer (Kareer et al., 2025a; Yang et al., 2025; Li et al., 2025b; Kim et al., 2026; Kareer et al., 2025b). PhysBrain and Human-as-Humanoid further connect human observations to embodied supervision and humanoid actions (Lin et al., 2025; 2026). For motion data, OakInk2 and EgoDex associate manipulation recordings with articulated trajectories and language (Zhan et al., 2024; Hoque et al., 2026). InterCap addresses joint human–object capture, DexCap uses wearable tracking for dexterous demonstrations, and MAMMA provides markerless multi-view reconstruction (Huang et al., 2024; Wang et al., 2024a; Cuevas Velasquez et al., 2026). EvUBody adds simulated and recorded events to the study of coordinated upper-body and two-hand motion; MAMMA supplies real-data reference annotations rather than our reconstruction algorithm.

RGB body–hand reconstruction. Image-based methods recover hand detail and expressive body pose through local representations and cross-part fusion (Pavlakos et al., 2024; Choutas et al., 2020; Zhang et al., 2023; Moon, 2026). Video methods incorporate temporal context, motion priors, and geometric constraints to improve consistency or recover global trajectories (Kocabas et al., 2020; Shin et al., 2024; Shen et al., 2024). These techniques provide spatial and temporal foundations for reconstruction, but motion between retained images remains constrained by the available RGB

observations. We study whether interval events improve agreement with reference motion under explicitly restricted image access.

Event and RGB–event reconstruction. Event-based methods address body pose, single hands, and interacting hands (Scarpellini et al., 2021; Zou et al., 2021; Rudnev et al., 2021; Jiang et al., 2024a; Millerdurai et al., 2024). EvRGBHand fuses real RGB–event recordings for hand reconstruction, while EVT studies body reconstruction with simulated events (Jiang et al., 2024b; Lang & Chuah, 2025). RELI11D includes multiple sensing modalities, and its LEIR baseline additionally uses LiDAR (Yan et al., 2024). These settings differ in output anatomy, sensor inputs, and temporal access. EvUBody brings sparse RGB observations, event streams, joint reconstruction of the upper body and fine-grained articulation of both hands, and inter-frame evaluation into a unified benchmark.

3 EvUBody: RGB–Event Upper-Body and Hand Motion

Complementary data sources. EvUBody organizes two sources around joint upper-body and articulated two-hand reconstruction (Figure 1). EvUBody-OakInk2 reuses existing manipulation videos and motion annotations to construct simulated events. EvUBody-Interact records real sensor observations alongside a separate multi-view reference-annotation system. The former supports controlled reconstruction experiments; the latter provides captured data for studying sensor and annotation effects beyond video simulation.

Video-derived events. We use the allocentric top view of OakInk2 (Zhan et al., 2024), deriving the body–hand representation from its released SMPL-X/MANO annotations. Events are generated with v2e (Hu et al., 2021) and SuperSloMo from the complete source RGB sequence before restricting the images visible to the reconstruction model. Thus, simulation uses hidden source frames and may use subsequent frames for interpolation: its events are not independent physical observations of inter-frame motion. Supervision remains at the original RGB-aligned 30 Hz timestamps. Appendix B specifies simulation parameters and temporal mapping.

Real tabletop capture. EvUBody-Interact records 15 participants performing 120 tasks across 12 categories under normal and low-light conditions. Tasks span object transport, precision grasping, tool use, bimanual assembly, handover, constrained manipulation, and failure recovery. Specifications define objects, action order, variants, and completion criteria. Participant metadata include age, height, arm span, hand length, and handedness. All sequences include task metadata and language descriptions that provide semantic context. The capture inventory totals 21,151.93 seconds. Sequence durations and the participant-disjoint split are detailed in Appendix A.

Multi-view reference annotation. Eight RealSense D435i RGB views drive a MAMMA-based pipeline (Cuevas Velasquez et al., 2026): human masks and dense point predictions support cross-view association, triangulation, and sequence-level SMPL-X fitting. DVSync APS RGB and DVS events are excluded from pose fitting; APS projection is a separate visualization step. Multi-view RGB and exported poses use a software-aligned nominal 30 Hz timeline, without hardware exposure triggering. Reference fitting uses whole-sequence information, whereas the reconstruction model obeys causal input constraints. Validation against a full-body optical motion capture system yields a mean per-joint position error (MPJPE) of 0.8 mm. Calibration and temporal alignment are detailed in Appendix A.

4 Method

4.1 Problem Formulation

We reconstruct upper-body and articulated two-hand motion offline under causal observation access. Let $\{\tau_t\}_{t=0}^{T-1}$ be target timestamps, $\mathcal{R}$ the retained RGB indices, and $\mathcal{E} = \{(\mathbf{x}_i, u_i, p_i)\}_i$ the events, with location $\mathbf{x}_i$, time u_i, and polarity p_i. Given camera intrinsics K, the model predicts

$$\widehat{Y}_t = f_\theta\big(\{I_r : r \in \mathcal{R},\ r \le t\}, \mathcal{E}_{<\tau_t}, K\big) \in \mathbb{R}^{57\times 3}. \tag{1}$$

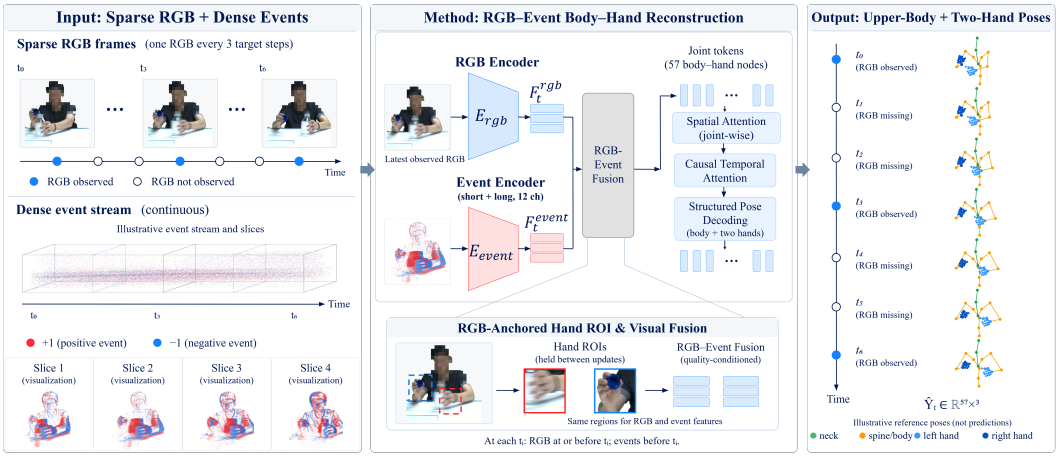

Figure 2: Overview of causal RGB–event reconstruction. The latest available RGB observation anchors hand regions, which are held until the next RGB update and shared by RGB and event feature sampling. Short- and long-window event features undergo quality-conditioned fusion with RGB features. Joint-wise spatial attention, causal temporal attention, and structured decoding recover upper-body and two-hand poses at every target timestamp. The one-in-three RGB pattern illustrates the experimental protocol; event slices and reference skeletons illustrate the inputs and target representation, not model predictions.

The representation contains 15 body nodes and 21 nodes per hand. In the EvUBody benchmark, coordinates are metric 3D positions in the input RGB camera frame, and targets follow the native 30 Hz RGB-aligned timeline. Language is annotation metadata, not an input. Figure 2 summarizes three stages: encoding the latest available RGB observation and preceding events, fusing global and RGB-anchored hand features, and decoding 57 joint tokens through spatial and causal temporal attention. The structured output combines upper-body position with wrist-relative hand articulation.

4.2 RGB Anchors and Event Features

At each retained image, a coarse RGB path followed by spatial refinement produces a predicted skeleton P_r^R and appearance features. A ViT-H/16 encoder supplies final patch features, which are adapted to multiple spatial scales for our 2.5D refinement head. The head refines the coarse pose using camera intrinsics. The most recent prediction, indexed by $a(t) = \max\{r \in \mathcal{R} : r \leq t\}$, anchors subsequent reconstruction. Historical RGB pose differences remain motion priors for temporal modeling; they do not move the sampling regions. Feature adaptation is detailed in Appendix C.

Short and long event windows represent recent and accumulated brightness changes. Each window contains four count channels (two time bins and two polarities) and two last-event time surfaces, giving 12 channels in total. A global encoder captures scene-level changes; local encoders preserve hand detail. All windows exclude the target timestamp, and empty inputs are explicitly masked. Exact boundaries, normalization, and the event-generation model are given in Appendix B.

4.3 RGB-Anchored Hand Regions and Fusion

For each hand, projected joints from the latest retained RGB prediction determine a square primary region and a larger concentric context region. Both remain fixed until the next RGB update. RGB features and primary event features are sampled from the same primary region, while the expanded region provides additional event context. Sampling uses these predicted regions, with no ground-truth boxes or motion-based relocation at inference. An RGB-driven auxiliary localization branch remains supervised, and its confidence enters the fusion conditions; its predicted center and scale do not determine the actual crop. Appendix C specifies the region geometry and auxiliary supervision.

Global and hand features undergo separate quality-conditioned fusion. Local correlations predict bounded offsets for resampling RGB features R into $\widetilde{R}$. Channel and spatial gates combine these

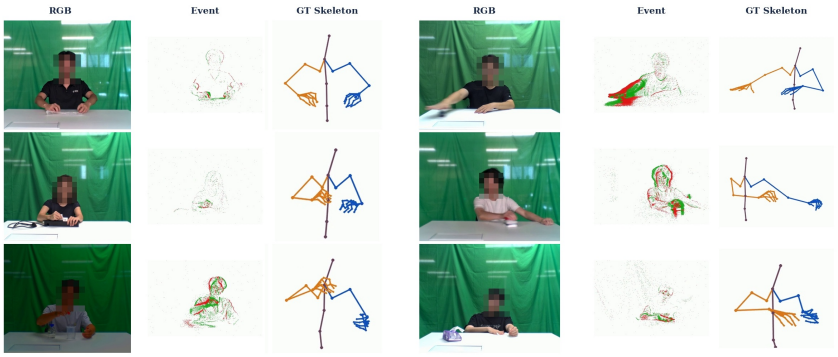

Figure 3: Examples from EvUBody-Interact. RGB images, event visualizations, and corresponding reference skeletons illustrate tabletop manipulation by different participants.

features with events E:

$$F^{\text{fuse}} = R + m_E \, \psi_f\big([G_R \odot \widetilde{R},\ G_E \odot E]\big). \tag{2}$$

The gates depend on visual features and brightness statistics, observation age, event presence, and region confidence. The mask m_E suppresses the event residual for empty inputs. The RGB residual preserves appearance information, while bounded resampling accommodates local feature misalignment.

4.4 Joint-Level Reconstruction and Learning

RGB pose projections guide body and hand token aggregation. Tokens encode joint identity, reference position, normalized historical pose differences, and observation timing. Spatial attention first models relationships across the 57 nodes; causal temporal attention then aggregates each node's current and past states. A body decoder predicts a residual from the RGB anchor, while a shared hand Transformer refines wrist-relative articulation:

$$\widehat{B}_t = P^R_{a(t),B} + \Delta B_t, \qquad \widehat{Y}_{t,h,j} = \widehat{B}_{t,w(h)} + \widehat{H}_{t,h,j}, \quad \widehat{H}_{t,h,0} = \mathbf{0}. \tag{3}$$

Here $w(h)$ denotes the corresponding body wrist. This representation gives body and hand wrist nodes identical coordinates by construction, while allowing local articulation to be learned separately from global placement.

All parameters are jointly optimized with absolute pose, wrist-relative hand, interval displacement, and adjacent displacement supervision:

$$\mathcal{L} = \mathcal{L}_{\text{abs}} + \lambda_h \mathcal{L}_{\text{hand}} + \lambda_i \mathcal{L}_{\text{interval}} + \lambda_\Delta \mathcal{L}_{\text{increment}} + \mathcal{L}_{\text{aux}}. \tag{4}$$

The motion terms match reference displacement rather than penalizing motion magnitude. Auxiliary terms supervise coarse hand predictions, the RGB branch, hand bone lengths, and localization. Validity and clip-ownership masks exclude invalid nodes and duplicated context supervision. Appendix C gives the weights, normalization, and training settings.

5 Experiments

5.1 Datasets and Compared Methods

We evaluate on EvUBody-OakInk2 and EvUBody-Interact, the simulated-event and captured tabletop manipulation subsets of EvUBody. The benchmark targets upper-body joints and articulated hand joints. Its two sources distinguish reconstruction with video-derived events from reconstruction with recorded sensor events. EvUBody-Interact covers 120 tasks in 12 categories, including object transport, precision grasping, tool use, bimanual assembly, handover, constrained manipulation, and failure recovery. Figure 3 presents examples from EvUBody-Interact.

Table 1: Input modalities, output coverage, and reconstruction errors on EvUBody. The evaluated region is specified for each row. Body and hand results are compared separately.

	Input		Output			EvUBody-OakInk2		EvUBody-Interact	
Method	RGB	Ev.	B	H	Eval.	M↓	PA↓	M↓	PA↓
EvRGBHand (Jiang et al., 2024b)	✓	✓	✗	✓	Hands	46.92	14.42	67.02	10.73
Ev2Hands (Millerdurai et al., 2024)	✗	✓	✗	✓	Hands	37.87	9.78	130.83	12.73
EventHands (Rudnev et al., 2021)	✗	✓	✗	✓	Hands	28.28	**7.88**	105.39	10.91
LME (Scarpellini et al., 2021)	✗	✓	✓	✗	Body	45.97	27.38	99.01	51.03
AiOS (Sun et al., 2024)	✓	✗	✓	✓	Body	38.21	15.73	39.18	20.53
SMPLest-X (Yin et al., 2025)	✓	✗	✓	✓	Body	21.22	**12.12**	42.79	20.70
SAM-Body4D^{v} (Gao et al., 2025)	✓	✗	✓	✓	Body	103.79	47.39	164.05	48.25
Hand4Whole++ (Moon, 2026)	✓	✗	✓	✓	Body	100.41	46.18	53.16	27.79
PEAR (Wu et al., 2026a)	✓	✗	✓	✓	Body	72.55	31.88	81.51	30.30
GEM-X^{v} (Li et al., 2025a)	✓	✗	✓	✓	Body	46.18	31.04	53.61	32.37
Ours	✓	✓	✗	✓	Hands	**20.21**	12.84	**36.36**	**7.89**
Ours	✓	✓	✓	✓	Body	**20.97**	13.57	**37.80**	**19.88**

Ev.: events; B/H: body/hand output; Eval.: evaluated region; v: RGB video. M/PA: MPJPE/PA-MPJPE (mm); lower is better. Bold indicates the lowest error for each metric within each evaluated region and subset. It should be noted that SAM-Body4D is originally training-free and we did not finetune it during evaluation.

We group baselines by their actual visual inputs and anatomical outputs. EvRGBHand (Jiang et al., 2024b) uses RGB and events for hand reconstruction. EventHands (Rudnev et al., 2021) and Ev2Hands (Millerdurai et al., 2024) use events to reconstruct hands, whereas LME (Scarpellini et al., 2021) estimates body joints without articulated fingers. AiOS (Sun et al., 2024), SMPLest-X (Yin et al., 2025), SAM-Body4D (Gao et al., 2025), Hand4Whole++ (Moon, 2026), PEAR (Wu et al., 2026a), and GEM-X (Li et al., 2025a) use RGB images or videos and provide both body and hand outputs. The evaluated region for each method is specified in Table 1. Among these baselines, none jointly reconstructs both regions using RGB and events. EvUBody and our model address this combination through a paired benchmark and reconstruction framework.

5.2 Baseline Adaptation and Evaluation

We adapt trainable baseline components on both subsets using official implementations and training recipes, retaining each method's input modality and configuration. Table 1 reports MPJPE and PA-MPJPE separately for the indicated hand and body regions; computational budgets are method-specific. Both are measured in millimeters.

For hand reconstruction, our method achieves the lowest MPJPE on both subsets, with errors of 20.21 mm on EvUBody-OakInk2 and 36.36 mm on EvUBody-Interact. These results correspond to reductions of 28.5% and 45.7% relative to the strongest baseline on each subset, respectively. On EvUBody-Interact, our method also achieves the lowest PA-MPJPE of 7.89 mm, compared with 10.73 mm for the best baseline.

For body reconstruction, our method obtains the lowest MPJPE on both subsets, reaching 20.97 mm on EvUBody-OakInk2 and 37.80 mm on EvUBody-Interact, compared with the best baseline results of 21.22 mm and 39.18 mm, respectively. Our method also achieves the lowest PA-MPJPE on EvUBody-Interact at 19.88 mm, while SMPLest-X performs best on EvUBody-OakInk2 with 12.12 mm, compared with our 13.57 mm. Overall, our method achieves the lowest MPJPE across both evaluated regions and subsets, and leads in both metrics for hand and body reconstruction on EvUBody-Interact.

Together, these results demonstrate consistently lower MPJPE across both hand and body reconstruction on the two EvUBody subsets, simulated and recorded event settings. Figure 4 complements these quantitative results with qualitative comparisons of our predictions and selected baselines against reference skeletons. Additional visualizations are provided in Appendix F.

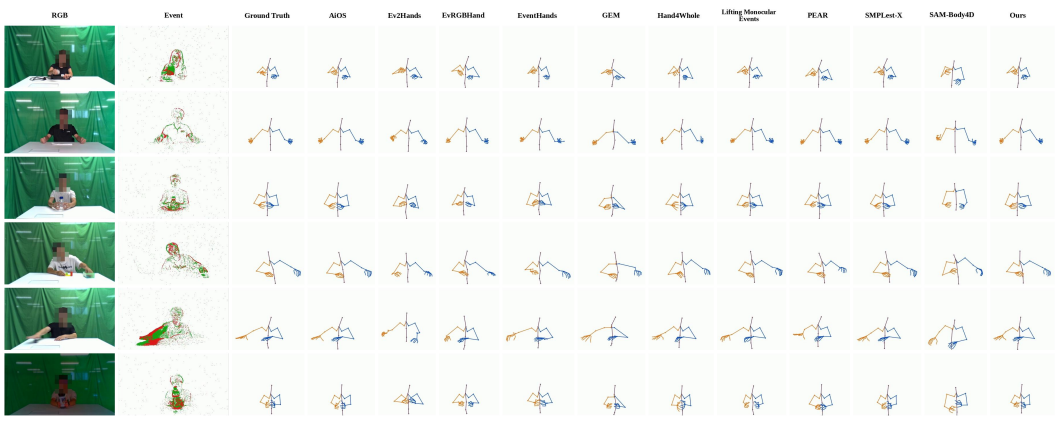

Figure 4: Qualitative comparison of skeleton reconstruction. RGB images and event visualizations are shown alongside reference skeletons and predictions from selected baseline methods. Additional comparisons are provided in the appendix.

5.3 Implementation Details

We jointly train the RGB–event model using a ViT-H/16 spatial encoder, dual-window event features, and causal temporal attention. The EvUBody-OakInk2 configuration uses AdamW with learning rates of 10^{-5} for the encoder and 10^{-4} for the remaining parameters. Input resolutions, architecture settings, optimization schedules, and hardware details are provided in Appendix C.

Sparse observations and exposure. The sparse-observation protocol retains approximately one in three RGB frames, rotates training phases across epochs, and fixes evaluation to phase zero. Supervision covers all original RGB-aligned timestamps. Each update averages normal, underexposed, and overexposed RGB conditions; events and annotations remain fixed. This augmentation is distinct from the recorded normal/low-light conditions in EvUBody-Interact.

5.4 Ablation Study

We compare the full model with an RGB-only baseline and four architectural variants on the EvUBody-OakInk2 validation split. All configurations are trained for 40 epochs with seed 42 under the $k = 3$ sparse-RGB protocol. Table 2 evaluates the 57-node body–hand representation at withheld-RGB timestamps under normal exposure (EV0). Appendix E.1 specifies the evaluated configurations, including their region-selection policies, and the alignment protocol.

Table 2: Reconstruction at withheld-RGB timestamps on EvUBody-OakInk2 validation ($k = 3$, EV0). Errors cover all valid nodes of the 57-node representation and are in mm. Bold indicates the lowest error in each column.

Configuration	MPJPE ↓	PA-MPJPE ↓
RGB only	23.40	14.53
Without global events	22.16	13.59
Without hand events	22.24	13.73
Concatenation fusion	23.09	13.94
Without hand refinement	22.46	13.76
Full model	**20.97**	**13.57**

Event representations and fusion. The full model attains 20.97 mm MPJPE and 13.57 mm PA-MPJPE, respectively 10.38% and 6.61% lower than the RGB-only baseline (23.40 mm and 14.53 mm). The variants without global or hand event features have higher errors than the full model on both metrics. Concatenation fusion gives 23.09 mm and 13.94 mm. Body, left-hand, and

right-hand MPJPE decrease by 9.61%, 10.00%, and 11.42%, respectively, relative to RGB only (Table 3).

Hand refinement. The full model has independently aligned left- and right-hand PA-MPJPE of 5.91 mm and 6.11 mm, respectively 0.28 mm and 0.22 mm lower than the variant without hand refinement (6.19 mm and 6.33 mm; Table 3). These measurements describe local hand structure after independent similarity alignment.

Table 3: Regional errors at withheld-RGB timestamps under EV0. Body, Left, and Right denote MPJPE for the upper body and the two hands; Left PA and Right PA use independent hand alignment.

Configuration	Body	Left	Right	Left PA	Right PA
RGB only	25.07	22.59	23.02	6.32	6.48
Without global events	24.05	22.80	21.17	6.11	6.16
Without hand events	23.73	23.13	21.28	6.19	6.19
Concatenation fusion	25.17	23.25	22.45	6.12	6.23
Without hand refinement	25.52	22.62	21.12	6.19	6.33
Full model	**22.66**	**20.33**	**20.39**	**5.91**	**6.11**

Exposure and motion recovery. The full model has the lowest observed MPJPE and PA-MPJPE at all seven RGB exposure settings from EV-3 to EV$+3$. Reductions relative to RGB only range from 8.41–10.71% and 5.31–7.24%, respectively (Appendix Table A3). Both metrics improve over RGB only at retained- and withheld-RGB timestamps (Appendix Table A4). At withheld-RGB timestamps, interval displacement and bone-length errors are 24.91% and 9.38% lower than RGB only; the full model has the lowest observed error on both metrics (Appendix Table A5).

6 Conclusion

We introduce EvUBody, a dataset and benchmark for joint reconstruction of upper-body motion and fine-grained two-hand articulation from sparse RGB observations and event streams. Its simulated and real subsets pair RGB–event observations with joint 3D motion annotations and sequence-level language descriptions, while markerless multiview capture preserves natural manipulation appearance. Our causal framework combines RGB-anchored hand regions, quality-conditioned cross-modal fusion, and joint-level spatiotemporal modeling to recover coordinated body–hand motion. Comparisons on both subsets show improved body and hand joint-position accuracy, while inter-frame evaluation and RGB exposure perturbation experiments further support the complementary value of events for reconstruction from sparse RGB observations.

We will publicly release the EvUBody datasets and our method implementation to provide reusable resources for event-assisted human manipulation data construction and coordinated motion reconstruction.

Ethics and data availability. All participants provided informed consent for data collection and public release in accordance with applicable ethical requirements. Identifying visual information is masked to protect privacy. Upon acceptance, we will release the full EvUBody dataset and benchmark implementations. Data samples are available in the supplementary material.

AI Use Statement

We used generative AI tools to assist with general coding tasks, translation, manuscript editing, and scientific figure preparation. The research questions, methodological choices, experimental design, model training, and interpretation of results were developed and carried out by the authors without the involvement of generative AI. We subjected all AI-assisted outputs to two rounds of rigorous cross-checking to ensure the accuracy and integrity of the final manuscript, code, and figures.

REFERENCES

Vasileios Choutas, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J Black. Monocular expressive body regression through body-driven attention. In *European Conference on Computer Vision*, pp. 20–40. Springer, 2020.

Hanz Cuevas Velasquez, Anastasios Yiannakidis, Soyong Shin, Giorgio Becherini, Markus Höschle, Joachim Tesch, Taylor Obersat, Tsvetelina Alexiadis, Eni Halilaj, and Michael J. Black. MAMMA: Markerless Accurate Multi-person Motion Acquisition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J Davison, Jörg Conradt, Kostas Daniilidis, et al. Event-based vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(1):154–180, 2020.

Mingqi Gao, Yunqi Miao, and Jungong Han. Sam-body4d: Training-free 4d human body mesh recovery from videos. *arXiv preprint arXiv:2512.08406*, 2025. URL `https://arxiv.org/abs/2512.08406`.

Ryan Hoque, Peide Huang, David Yoon, Mouli sivapurapu, and Jian Zhang. Egodex: Learning dexterous manipulation from large-scale egocentric video. In *International Conference on Learning Representations*, volume 2026, pp. 4218–4237, 2026.

Yuhuang Hu, Shih-Chii Liu, and Tobi Delbruck. v2e: From video frames to realistic dvs events. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1312–1321. IEEE, 2021.

Yinghao Huang, Omid Taheri, Michael J Black, and Dimitrios Tzionas. Intercap: Joint markerless 3d tracking of humans and objects in interaction from multi-view rgb-d images. *International Journal of Computer Vision*, 132(7):2551–2566, 2024.

Zhewei Huang, Tianyuan Zhang, Wen Heng, Boxin Shi, and Shuchang Zhou. Real-time intermediate flow estimation for video frame interpolation. In *European conference on computer vision*, pp. 624–642. Springer, 2022.

Jianping Jiang, Jiahe Li, Baowen Zhang, Xiaoming Deng, and Boxin Shi. Evhandpose: Event-based 3d hand pose estimation with sparse supervision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(9):6416–6430, 2024a.

Jianping Jiang, Xinyu Zhou, Bingxuan Wang, Xiaoming Deng, Chao Xu, and Boxin Shi. Complementing event streams and rgb frames for hand mesh reconstruction. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24944–24954. IEEE, 2024b.

Simar Kareer, Dhruv Patel, Ryan Punamiya, Pranay Mathur, Shuo Cheng, Chen Wang, Judy Hoffman, and Danfei Xu. Egomimic: Scaling imitation learning via egocentric video. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 13226–13233. IEEE, 2025a.

Simar Kareer, Karl Pertsch, James Darpinian, Judy Hoffman, Danfei Xu, Sergey Levine, Chelsea Finn, and Suraj Nair. Emergence of human to robot transfer in vision-language-action models. *arXiv preprint arXiv:2512.22414*, 2025b.

Ji Woong Kim, Ke Wang, Zipeng Fu, Sirui Chen, Jeff Lai, Chelsea Finn, et al. Ego-pi: Vla finetuning for ego-centric human and robot data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1515–1524, 2026.

Taewoo Kim, Jaeseok Jeong, Hoonhee Cho, Yuhwan Jeong, and Kuk-Jin Yoon. Towards real-world event-guided low-light video enhancement and deblurring. In *European Conference on Computer Vision*, pp. 433–451. Springer, 2024.

Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5252–5262. IEEE, 2020.

Bo Lang and Mooi Choo Chuah. Event-guided video transformer for end-to-end 3d human pose estimation. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5114–5124. IEEE, 2025.

Marion Lepert, Jiaying Fang, and Jeannette Bohg. Phantom: Training robots without robots using only human videos. In *Conference on Robot Learning (CoRL)*, Seoul, Korea, 2025.

Jiefeng Li, Jinkun Cao, Haotian Zhang, Davis Rempe, Jan Kautz, Umar Iqbal, and Ye Yuan. Genmo: A generalist model for human motion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025a.

Qixiu Li, Yu Deng, Yaobo Liang, Lin Luo, Lei Zhou, Chengtang Yao, Lingqi Zeng, Zhiyuan Feng, Huizhi Liang, Sicheng Xu, et al. Scalable vision-language-action model pretraining for robotic manipulation with real-life human activity videos. *arXiv preprint arXiv:2510.21571*, 2025b.

Xiaopeng Lin, Shijie Lian, Bin Yu, Ruoqi Yang, Zhaolong Shen, Changti Wu, Yuzhuo Miao, Yurun Jin, Yukun Shi, Jiyan He, et al. Physbrain: Human egocentric data as a bridge from vision language models to physical intelligence. *arXiv preprint arXiv:2512.16793*, 2025.

Xiaopeng Lin, Ruoqi Yang, Shijie Lian, Zhaolong Shen, Bin Yu, Changti Wu, Haibao Liu, Yuxiang Zhang, Hong Li, Qiyuan Su, et al. Human-as-humanoid: Enabling zero-shot humanoid learning from ego-exo human videos with human-aligned embodiments. *arXiv preprint arXiv:2606.32009*, 2026.

Christen Millerdurai, Diogo Luvizon, Viktor Rudnev, André Jonas, Jiayi Wang, Christian Theobalt, and Vladislav Golyanik. 3d pose estimation of two interacting hands from a monocular event camera. In *2024 International Conference on 3D Vision (3DV)*, pp. 291–301. IEEE, 2024.

Gyeongsik Moon. Enhancing hands in 3d whole-body pose estimation with conditional hands modulator. *arXiv preprint arXiv:2603.14726*, 2026.

Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3d with transformers. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9826–9836. IEEE, 2024.

Viktor Rudnev, Vladislav Golyanik, Jiayi Wang, Hans-Peter Seidel, Franziska Mueller, Mohamed Elgharib, and Christian Theobalt. Eventhands: Real-time neural 3d hand pose estimation from an event stream. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12365–12375. IEEE, 2021.

Gianluca Scarpellini, Pietro Morerio, and Alessio Del Bue. Lifting monocular events to 3d human poses. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1358–1368. IEEE, 2021.

Zehong Shen, Huaijin Pi, Yan Xia, Zhi Cen, Sida Peng, Zechen Hu, Hujun Bao, Ruizhen Hu, and Xiaowei Zhou. World-grounded human motion recovery via gravity-view coordinates. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–11, 2024.

Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J Black. Wham: Reconstructing world-grounded humans with accurate 3d motion. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2070–2080. IEEE, 2024.

Qingping Sun, Yanjun Wang, Ailing Zeng, Wanqi Yin, Chen Wei, Wenjia Wang, Haiyi Mei, Chi-Sing Leung, Ziwei Liu, Lei Yang, et al. Aios: All-in-one-stage expressive human pose and shape estimation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1834–1843. IEEE, 2024.

Chen Wang, Haochen Shi, Weizhuo Wang, Ruohan Zhang, Li Fei-Fei, and C Karen Liu. Dexcap: Scalable and portable mocap data collection system for dexterous manipulation. *arXiv preprint arXiv:2403.07788*, 2024a.

Yihan Wang, Lahav Lipson, and Jia Deng. Sea-raft: Simple, efficient, accurate raft for optical flow. In *European Conference on Computer Vision*, pp. 36–54. Springer, 2024b.

Jiahao Wu, Yunfei Liu, Lijian Lin, Ye Zhu, Lei Zhu, Jingyi Li, and Yu Li. Pear: Pixel-aligned expressive human mesh recovery. *arXiv preprint arXiv:2601.22693*, 2026a.

Zekai Wu, Shuqi Fan, Mengyin Liu, Yuhua Luo, Xincheng Lin, Ming Yan, Junhao Wu, Xiuhong Lin, Yuexin Ma, Chenglu Wen, et al. Flashcap: Millisecond-accurate human motion capture via flashing leds and event-based vision. *arXiv preprint arXiv:2603.19770*, 2026b.

Ming Yan, Yan Zhang, Shuqiang Cai, Shuqi Fan, Xincheng Lin, Yudi Dai, Siqi Shen, Chenglu Wen, Lan Xu, Yuexin Ma, et al. Reli11d: A comprehensive multimodal human motion dataset and method. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2250–2262. IEEE, 2024.

Ruihan Yang, Qinxi Yu, Yecheng Wu, Rui Yan, Borui Li, An-Chieh Cheng, Xueyan Zou, Yunhao Fang, Xuxin Cheng, Ri-Zhao Qiu, et al. Egovla: Learning vision-language-action models from egocentric human videos. *arXiv preprint arXiv:2507.12440*, 2025.

Wanqi Yin, Zhongang Cai, Ruisi Wang, Ailing Zeng, Chen Wei, Qingping Sun, Haiyi Mei, Yanjun Wang, Hui En Pang, Mingyuan Zhang, et al. Smplest-x: Ultimate scaling for expressive human pose and shape estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Jiajun Zhai, Hao Shi, Shangwei Guo, Kailun Yang, and Kaiwei Wang. E-vla: Event-augmented vision-language-action model for dark and blurred scenes. *arXiv preprint arXiv:2604.04834*, 2026.

Xinyu Zhan, Lixin Yang, Yifei Zhao, Kangrui Mao, Hanlin Xu, Zenan Lin, Kailin Li, and Cewu Lu. Oakink2: A dataset of bimanual hands-object manipulation in complex task completion. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 445–456. IEEE, 2024.

Hongwen Zhang, Yating Tian, Yuxiang Zhang, Mengcheng Li, Liang An, Zhenan Sun, and Yebin Liu. Pymaf-x: Towards well-aligned full-body model regression from monocular images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12287–12303, 2023.

Shihao Zou, Chuan Guo, Xinxin Zuo, Sen Wang, Pengyu Wang, Xiaoqin Hu, Shoushun Chen, Minglun Gong, and Li Cheng. Eventhpe: Event-based 3d human pose and shape estimation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10976–10985. IEEE, 2021.



(a) Camera arrangement (b) Calibration scene

Figure 5: Capture and annotation systems. (a) Eight RealSense D435i cameras (red) provide multi-view RGB for offline MAMMA reference annotation; DVSync (blue) records APS RGB and events for reconstruction. The streams use software time alignment, not shared hardware exposure triggering. (b) Calibration scene.

A Capture, Reference Annotation, and Data Splits

Inventory and task scope. The real capture inventory contains 1,809 video groups totaling 21,151.93 seconds, with mean/median duration 11.69/10.00 seconds and range 4.20–51.73 seconds. The 12 task categories are pick-and-place, precision grasping, opening/disassembly, rotation/tool use, pouring/transfer, continuous control, surface cleaning, deformable-object manipulation, bimanual assembly, handover/regrasping, constrained-space manipulation, and failure recovery. Operating hand, speed, target position, and occlusion vary within task specifications.

Table A1: Real-data video inventory by participant-disjoint split. Task categories are shared across splits. Counts precede independent reference-pose quality validation.

Split	Participants	Video groups	Duration
Train	10	1,199	3:54:20
Validation	2	238	0:49:21
Test	3	372	1:08:51

Sensors, timing, and coordinates. As shown in Figure 5, inspected capture metadata record eight 1920×1080 RGB streams at 30 fps, DVSync APS at 3840×2160, and DVS at 1280×720, with on/off bias values of 40/40. APS acquisition rate is not inferred from the multi-view RGB rate. No depth input is used in the documented annotation pipeline. RGB groups use software timestamps and a common nominal 30 Hz presentation timeline; hardware triggering is disabled. A sampled sequence has a maximum grouping skew of 20.456 ms and a 95th percentile of 20.448 ms. These are software grouping statistics, not exposure-time errors or RGB–event synchronization accuracy. Per-camera intrinsics and multi-camera extrinsics exist.

Reference fitting. The deployed MAMMA pipeline uses SAM2 masks and MammaNet dense predictions (512 points, visibility, and contact probabilities), followed by cross-view association, triangulation, and SMPL-X optimization. Cameras 01–08 provide fitting observations; APS is output-only and DVS is excluded. The executed optimizer is LBFGS in four stages, despite an Adam field in the saved configuration. Table A2 records the main settings. Body shape is shared over a sequence; temporal regularization and possible backward mask propagation use future information. The references are consequently offline, noncausal annotations.

The angular-acceleration implementation retains its default $dt = 1/20$; the caller does not override it for 30 Hz input. The translation temporal term compares a square-root displacement quantity with 0.02^2, and is not interpreted as a 2 cm hard filter. These settings describe the annotations used, not

Table A2: Recorded MAMMA fitting stages. Iterations are outer-loop limits; weights describe the deployed configuration.

Stage	Limit	Parameters and main constraints
1	50	Global rotation/translation; reprojection 20, triangulated 3D 100.
2	500	Pose including hands (excluding jaw), shape, translation; reprojection 100, shape prior 1.
3	100	Pose, shape, translation; reprojection 120, angular acceleration 0.003, 3D temporal term 0.03.
4	25	Body-parameter gradient mask; reprojection 300, angular acceleration 0.005, 3D temporal term 0.05, configured intersection weight 1.

corrected labels. APS overlays separately map frame times, interpolate adjacent reference joints, and project them.

B Event Simulation and Input Representation

Simulation. OakInk2 source RGB is nominally 30 fps at 848×480; events use a 640×480 grid. Recorded v2e settings are seed 1, SuperSloMo enabled with `SuperSloMo39.ckpt`, batch size 16, automatic timestamp resolution disabled, and requested resolution 0.01 seconds. The noisy preset specifies positive/negative thresholds 0.2/0.2, threshold standard deviation 0.05, cutoff 30 Hz, leak 0.1 Hz, shot noise 5 Hz, and zero refractory period. Saved examples use fourfold interpolation (8.33 ms intervals). Integer microsecond storage does not imply microsecond physical observations or supervision. The exact v2e revision remains to be recorded.

Under the ideal event model (Gallego et al., 2020),

$$L(\mathbf{x}_i, u_i) - L(\mathbf{x}_i, u_i^-) = p_i C, \qquad L = \log I, \tag{5}$$

where u_i^- is the preceding event time at that pixel and C is a contrast threshold. The simulator approximates intensity changes from interpolated source video. Events are generated before RGB restriction, and interpolation may depend on subsequent source images.

Clock mapping and windows. For the recorded generation recipe, raw event seconds are mapped to source-frame coordinates by $u = 30t_{\mathrm{raw}}(M-1)/M + 0.125$, with $M = 4(N-1)$ and N source images. This calibration is recipe-specific. The short window is $[\max(t-1, 0), t)$; the long window starts at the latest retained RGB index when RGB is withheld, and at the preceding retained index when RGB is present. Both are empty at $t = 0$ and exclude events exactly at t.

Channels. For a window beginning at ℓ, normalized event time is $a = \mathrm{clip}((u-\ell)/\max(t-\ell, 1), 0, 1-10^{-7})$. The time bin is $\lfloor 2a \rfloor$. Channels are ordered ON-early, ON-late, OFF-early, OFF-late, ON-last-time, OFF-last-time; short-window channels precede long-window channels. Counts use $\log(1+n)/\log 256$ without count clipping. Last-time surfaces store the maximum a per pixel and polarity, with zero for empty locations. They are not exponentially decaying surfaces. Maps are constructed on the native event grid before resizing or hand sampling.

C Network and Training Details

RGB and event encoders. RGB inputs are cached at 384×384. The coarse path average-pools to 128×128 and uses three convolutional blocks with 32/64/128 channels, GroupNorm, and SiLU. A pose head predicts body nodes and hand regions, followed by shared coarse hand decoding.

Spatial refinement uses a ViT-H/16 encoder with 32 Transformer blocks, 16 attention heads, a feature dimension of 1280, and 80 task tokens. Cached RGB images are resized to 256×192 using nearest-neighbor interpolation, with values in $[0, 1]$ and no ImageNet mean/std normalization. The encoder produces a final $1280 \times 16 \times 12$ patch feature map. Task tokens participate in encoder attention, but their output embeddings are not passed to the pose decoder; the 57 joint tokens are constructed separately downstream.

Three trainable 1×1 projections map this final patch feature map to 128, 256, and 512 channels. The resulting maps are bilinearly resampled to 48×48, 24×24, and 12×12, respectively, with `align_corners=False`. Lateral projections map each scale to 64 channels before resampling to 48×48, summation, and refinement. Heatmaps, projection priors, and joint embeddings support 2.5D corrections and metric recovery with camera intrinsics. A final 1×1 projection produces 128-channel appearance features for RGB–event fusion. All parameters are jointly optimized.

Global event maps are resized to 128×128. Primary and context hand maps use bilinear sampling at 192×192 and 96×96, with zero padding and `align_corners=False`. Fusion has 128 channels. Spatial and temporal Transformers each use one layer, four heads, and 256-dimensional feed-forward blocks. Temporal attention sees the current and preceding 32 original timestamps. The hand-refinement Transformer has two layers shared by the hands.

Region geometry and auxiliary branch. Let $\mathbf{l}, \mathbf{h}$ be coordinatewise extrema of the 21 projected hand nodes from the latest retained RGB prediction. In normalized image coordinates,

$$\mathbf{c} = (\mathbf{l} + \mathbf{h})/2, \qquad s = \mathrm{clip}(1.5 \max(\mathbf{h} - \mathbf{l}), 0.10, 0.90). \tag{6}$$

The primary square is $[\mathbf{c} - s/2, \mathbf{c} + s/2]$; the context square is concentric with side $1.75s$. Crops stay fixed between RGB updates. Auxiliary logits are predicted from 32×32 RGB features and combined with a spatial prior of $\sigma = 0.25$. A separate RGB-conditioned predictor uses observation age and period to estimate region parameters and confidence. Auxiliary centers and scales do not control actual sampling. Confidence is supervised against the primary-box IoU, and its mean over hands enters fusion. Historical pose differences are divided by $\max(\mathrm{period}, 1)$ and clipped coordinatewise to ± 0.1 before temporal-token construction.

Initialization and context. The first RGB frame is always retained. Previous and current anchors initially coincide, with zero motion difference and empty event windows. Each clip contains 448 owned supervision timestamps, 32 history positions, and three anchor-context positions. Validity masks suppress padded context in attention and scoring; context is not supervised twice. An empty event window does not invalidate an otherwise valid target. The protocol does not test RGB-free cold starts.

Losses. The implemented weights are

$$\begin{aligned} \mathcal{L} = \mathcal{L}_{\mathrm{abs}} + \mathcal{L}_{\mathrm{hand}} + 0.2\mathcal{L}_{\mathrm{coarse}} + 0.5\mathcal{L}_{\mathrm{rgb}} \\ + 0.5\mathcal{L}_{\mathrm{interval}} + 0.2\mathcal{L}_{\mathrm{increment}} + 0.05\mathcal{L}_{\mathrm{bone}} + 0.1\mathcal{L}_{\mathrm{roi}}. \end{aligned} \tag{7}$$

Coordinates are in meters. Absolute pose distances are normalized by 0.05; wrist-relative fine/coarse hand distances by 0.02; interval, adjacent displacement, and hand bone-length errors by 0.01. The RGB term sums its normalized absolute and wrist-relative errors. Region supervision combines center L1, box L1, 0.1 times confidence BCE, and 0.1 times target-to-predicted heatmap KL. The IoU confidence target is detached. Ground-truth heatmap standard deviations are clipped to $[0.04, 0.4]$.

Each term first averages over its valid nodes, edges, or hands at a timestamp, then normalizes by owned target count. Invalid terms contribute zero. Hand terms require a valid wrist. Interval terms require a withheld target and a valid in-clip anchor; adjacent terms require consecutive original indices and valid endpoints. These are displacement losses, without division by seconds. RGB supervision is normalized by owned targets rather than retained-image count. This training reduction differs from the valid-node-weighted evaluation.

Optimization and exposure. The EvUBody-OakInk2 configuration uses single-stage joint optimization on eight RTX PRO 6000 Blackwell GPUs with BF16 mixed precision. Each global batch contains eight clips, each with 448 supervised timestamps and 35 context positions (32 history and three anchor-context positions). Ablation runs use 40 training epochs and the final checkpoint.

AdamW starts at 10^{-5} for the ViT-H/16 encoder and 10^{-4} for all remaining parameters, including the feature adapters, with weight decay 10^{-4} and gradient norm clipping at 1.0. There are 280 updates per epoch and a global batch of eight clips. Warm-up lasts 560 updates; cosine decay reaches 0.1 of the base rates at epoch 40, which are held through epoch 60. The final checkpoint is used. Our model has 635,345,709 trainable parameters (approximately 635.35M), including 631,014,400 in the ViT-H/16 encoder.

For nonzero exposure e, with $X = T_{\mathrm{sRGB}\to\mathrm{linear}}(I)$,

$$I' = T_{\mathrm{linear}\to\mathrm{sRGB}}\Big(\mathrm{clip}\left[2^e X + \sqrt{10^{-4}2^e X}\,\epsilon_s + 0.003\epsilon_r, 0, 1\right]\Big), \tag{8}$$

where ϵ_s, ϵ_r are independent standard Gaussians; $e = 0$ returns the original image. Training equally weights original, underexposed ($e \sim U[-3,-1]$), and overexposed ($e \sim U[1,3]$) losses. Final evaluation uses $e \in \{-3,-2,-1,0,1,2,3\}$ with deterministic sequence/frame/condition noise based on seed 314159. Events and references remain fixed. This is a signal-dependent Gaussian approximation, not Poisson sampling or a joint physical sensor simulation.

D Skeleton and Metric Definitions

Node mapping. The 57-node order is 15 body nodes, 21 left-hand nodes, and 21 right-hand nodes. Body SMPL-X source indices are

```
0, 3, 9, 12, 15, 16, 18, 20, 17, 19, 21, 20, 21, 13, 14.
```

Body positions 11/12 are replaced by the mean of the corresponding hand's non-thumb MCP nodes (local indices 5,9,13,17). The left-hand source indices are

```
20, 37, 38, 39, 66, 25, 26, 27, 67, 28, 29, 30, 68, 34, 35,
36, 69, 31, 32, 33, 70,
```

and the right-hand indices are

```
21, 52, 53, 54, 71, 40, 41, 42, 72, 43, 44, 45, 73, 49, 50,
51, 74, 46, 47, 48, 75.
```

Each hand contains the wrist followed by thumb, index, middle, ring, and little finger chains. Tips come from extended model joints. Body wrists at indices 7/10 duplicate hand wrists at global indices 15/36. Duplicates and derived centers retain their weights in overall scoring. The hand-bone parent list for children 1–20 is

```
0,1,2,3, 0,5,6,7, 0,9,10,11, 0,13,14,15, 0,17,18,19.
```

Coordinate and alignment rules. OakInk2 uses the allocentric top RGB camera (serial 104422071041), metric coordinates, and conversion to millimeters for reporting. Valid reference nodes are finite and in front of the camera; this is not an occlusion-visibility label. MPJPE is $1000\sum_{t,j} m_{t,j}\|\widehat{Y}_{t,j} - Y_{t,j}\|_2 / \sum_{t,j} m_{t,j}$. Overall PA applies one per-frame similarity transform with determinant-positive rotation to valid 57-node sets; independent hand PA fits each 21-node hand separately. Wrist-relative error subtracts each prediction's and reference's own wrist without rotation or scale fitting. Frames with fewer than three usable points or insufficient rank are excluded from PA. Region results use body indices 0–14, left 15–35, right 36–56, and body wrists 7/10.

Motion and structure. For a withheld timestamp t and latest retained anchor $a(t)$, interval error averages

$$1000\left\|(\widehat{Y}_{t,j} - \widehat{Y}_{a(t),j}) - (Y_{t,j} - Y_{a(t),j})\right\|_2 \tag{9}$$

over valid endpoint pairs with an in-clip anchor. Bone-length error averages the absolute difference between predicted and reference edge lengths over the 40 hand edges. Both are in mm, not mm/s. All metrics sum errors and valid counts across sequences before division. Padding and non-owned targets are excluded; no unweighted mean of sequence means is used.

E Supplementary Ablation Results

E.1 Evaluation Protocol and Configuration Scope

Data and checkpoints. We use the EvUBody-OakInk2 validation split (71 sequences; 126,571 timestamps). With $k = 3$ and phase zero, 42,216 timestamps retain RGB and 84,355 withhold it.

All runs use seed 42 and the final epoch-40 checkpoint; the original learning-rate schedule spans 40 epochs. Unless stated otherwise, evaluation uses withheld-RGB timestamps at EV0. These single-seed validation results do not include multi-seed uncertainty estimates, an independent test split, or EvUBody-Interact measurements.

Configurations. The full model derives hand regions from the latest RGB-predicted joints and holds them until the next RGB update, retaining auxiliary localization, confidence supervision, and associated fusion inputs. The four component variants retain the adaptive policy in Section 4; comparisons therefore also involve a region-policy change. Matched RGB-anchored variants have not been evaluated. Concatenation fusion retains both modalities, convolution, and an RGB residual. Removing hand refinement preserves the prediction head and wrist consistency. The historical full model with adaptive regions is excluded.

Alignment and aggregation. Overall errors cover 57 nodes (15 body and 21 per hand), including duplicated wrists and derived hand centers. Overall PA-MPJPE uses one reflection-free similarity alignment per frame; hand PA-MPJPE aligns each hand independently. Valid-joint weighting means all-timestamp errors need not equal the unweighted mean of retained and withheld errors. This overall scope differs from the hand-only baseline metrics in Table 1. Errors are in mm; lower is better. Bold marks column minima.

Table A3: Exposure perturbations at withheld-RGB timestamps. Panels report (a) MPJPE and (b) overall PA-MPJPE over the 57-node representation. Events and annotations remain fixed.

(a) MPJPE

Configuration	EV−3	EV−2	EV−1	EV0	EV+1	EV+2	EV+3
RGB only	24.75	24.36	23.95	23.40	23.44	24.28	27.70
Without global events	23.84	22.90	22.59	22.16	22.89	22.92	26.67
Without hand events	23.99	23.03	22.87	22.24	22.58	23.09	26.72
Concatenation fusion	24.88	23.98	23.58	23.09	23.45	23.67	26.89
Without hand refinement	24.35	23.38	23.03	22.46	22.73	23.13	27.12
Full model	**22.43**	**21.75**	**21.43**	**20.97**	**21.19**	**21.74**	**25.37**

(b) PA-MPJPE

Configuration	EV−3	EV−2	EV−1	EV0	EV+1	EV+2	EV+3
RGB only	15.32	15.05	14.78	14.53	14.55	14.82	16.77
Without global events	14.73	14.06	13.83	13.59	13.65	14.04	15.91
Without hand events	14.67	14.10	14.00	13.73	13.85	14.15	15.99
Concatenation fusion	14.99	14.28	14.18	13.94	14.00	14.29	16.31
Without hand refinement	14.83	14.18	14.03	13.76	13.85	14.04	16.12
Full model	**14.44**	**13.96**	**13.74**	**13.57**	**13.62**	**13.89**	**15.88**

E.2 Robustness to RGB Exposure Perturbations

Table A3 evaluates seven exposure settings using the same checkpoint for each configuration. Perturbations modify only RGB inputs; events and annotations remain unchanged. They measure sensitivity to RGB exposure degradation, rather than jointly simulating both sensors' responses to physical lighting changes. The full model has the lowest observed MPJPE and PA-MPJPE at every setting, with reductions of 8.41–10.71% and 5.31–7.24%, respectively, relative to RGB only. At EV−3 and EV+3, its MPJPE is 22.43 mm and 25.37 mm, compared with 24.75 mm and 27.70 mm for RGB only.

E.3 Retained- and Withheld-RGB Timestamps

Table A4 reports all timestamps and the two observation subsets. The full model achieves MPJPE of 20.81 mm overall, 20.50 mm at retained-RGB timestamps, and 20.97 mm at withheld-RGB timestamps, with corresponding PA-MPJPE of 13.50 mm, 13.36 mm, and 13.57 mm. It has the lowest MPJPE in all three subsets and the lowest PA-MPJPE over all and withheld-RGB timestamps. At

retained-RGB timestamps, its PA-MPJPE ties with the variant without global events at the reported precision (13.36 mm). Both metrics remain lower than RGB only in all three subsets, showing that the observed gains extend to timestamps with and without current RGB input.

Table A4: Errors under EV0 over all, retained-RGB (Ret.), and withheld-RGB (With.) timestamps. M denotes MPJPE and PA denotes overall PA-MPJPE. Each subset is aggregated by valid-joint counts.

Configuration	All M	All PA	Ret. M	Ret. PA	With. M	With. PA
RGB only	23.09	14.41	22.48	14.16	23.40	14.53
Without global events	22.00	13.51	21.67	13.36	22.16	13.59
Without hand events	22.09	13.66	21.79	13.51	22.24	13.73
Concatenation fusion	22.94	13.86	22.66	13.72	23.09	13.94
Without hand refinement	22.30	13.69	21.99	13.55	22.46	13.76
Full model	**20.81**	**13.50**	**20.50**	**13.36**	**20.97**	**13.57**

E.4 Motion Recovery and Skeletal Structure

Table A5 complements position errors with interval displacement and bone-length errors at withheld-RGB timestamps under EV0. Interval displacement measures motion relative to the latest retained RGB observation; it is a displacement error in mm, not a velocity or jitter metric. Relative to RGB only, the full model reduces displacement error from 2.93 mm to 2.20 mm (24.91%) and bone-length error from 1.92 mm to 1.74 mm (9.38%). It has the lowest observed error on both metrics among these checkpoints.

Table A5: Interval displacement and bone-length errors at withheld-RGB timestamps under EV0. Displacement is referenced to the latest retained RGB observation; bone-length error uses 20 edges per hand. Both metrics are in mm.

Configuration	Interval displacement	Bone-length error
RGB only	2.93	1.92
Without global events	2.22	1.79
Without hand events	2.22	1.81
Concatenation fusion	2.21	1.78
Without hand refinement	2.35	1.76
Full model	**2.20**	**1.74**

F SUPPLEMENTARY QUALITATIVE EXPORTS

F.1 ILLUSTRATIVE SKELETON OUTPUTS

We show the available qualitative exports here, with RGB images, event renderings, reference skeletons, and method outputs. These exports retain their original renderer labels.

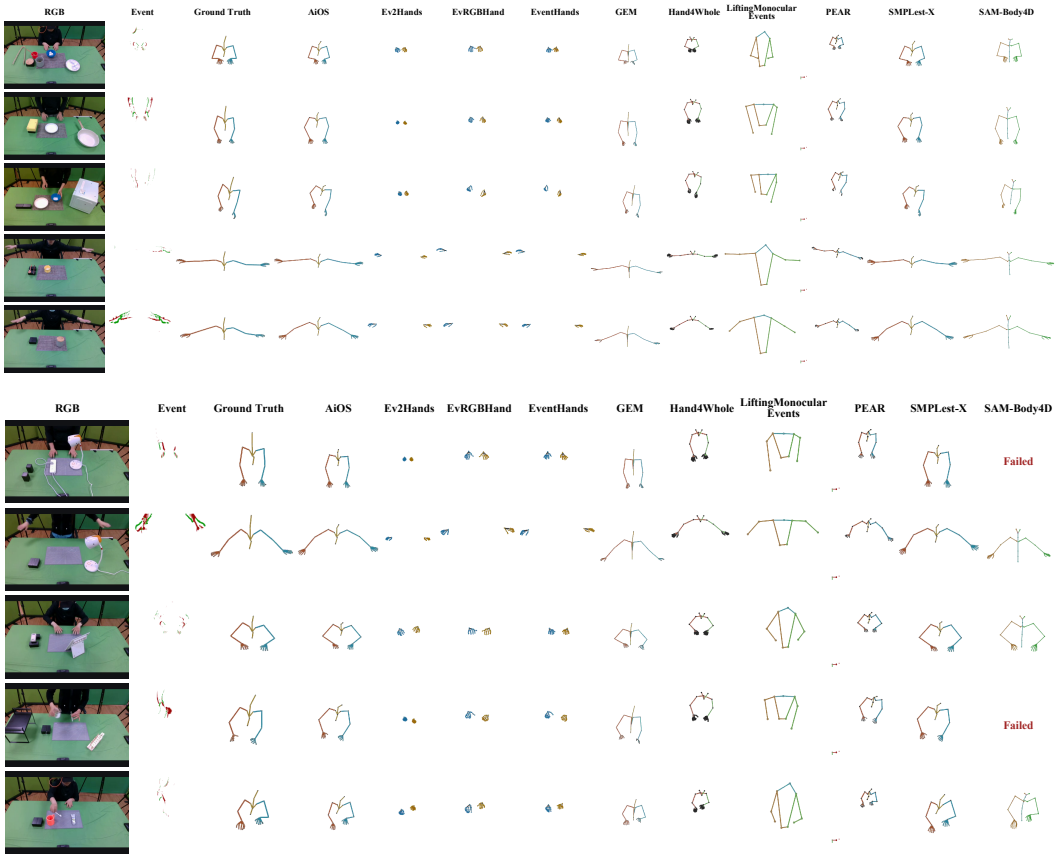

Figure 6: Additional qualitative skeleton exports with RGB and event context. Columns use the export's method names and anatomical outputs; they should not be interpreted as a common 57-node quantitative comparison. Further examples from the archived method-output visualization are shown below.

F.2 VISUALIZATION OF EVUBODY-INTERACT

The Interact panels place APS imagery, DVS renderings, and skeletal annotation visualizations side by side. They demonstrate the available data modalities,

G SUPPLEMENTARY QUANTITATIVE EXPERIMENTS

G.1 INTERPOLATION AND HISTORICAL EXTRAPOLATION

RIFE (Huang et al., 2022) creates internal withheld images from adjacent retained endpoints. SEA-RAFT (Wang et al., 2024b) uses the current and previous retained RGB frames for historical flow extrapolation. The former is an offline input condition; the latter does not access future RGB. Both conditions freeze the sparse-trained reconstruction weights and keep the original retained/withheld masks.

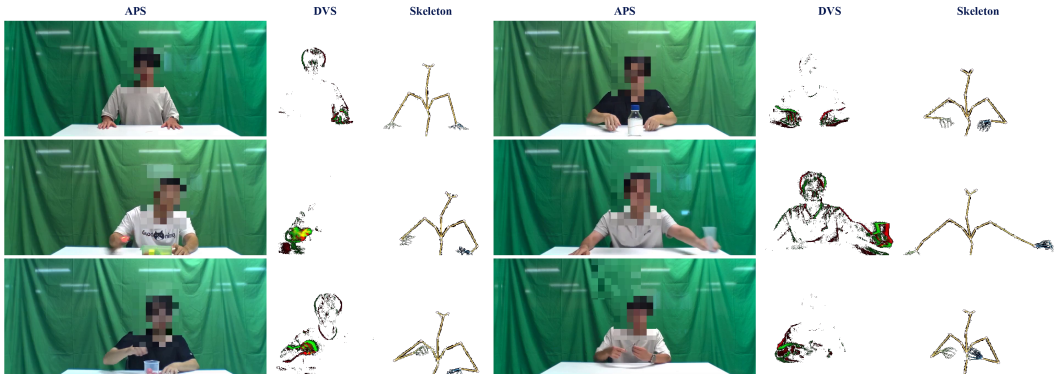

Figure 7: Additional Interact capture examples, with APS imagery, DVS renderings, and skeletal annotation visualizations in each panel. These examples illustrate appearance and pose diversity.

Table A6: Frozen-weight input diagnostics on OakInk2 validation. Withheld MPJPE is averaged equally over seven RGB EV settings (mm). Δ is the change from Sparse (negative is better). RIFE and SEA-RAFT have different temporal access.

Method	Sparse	RIFE	SEA-RAFT	Δ RIFE	Δ SEA
GEM-X	46.0384	45.7480	46.1798	-0.2904	$+0.1413$
AiOS	38.0843	36.9966	38.2149	-1.0878	$+0.1306$
Hand4Whole++	99.3106	99.1252	100.4067	-0.1854	$+1.0961$
PEAR	72.0086	71.3781	72.5544	-0.6306	$+0.5457$
SMPLest-X	21.5967	20.5500	21.2245	-1.0466	-0.3721
EvRGBHand	46.9812	46.6635	46.9244	-0.3177	-0.0568

Boundary and execution effects. RIFE holds the final anchor where no future endpoint exists. SEA-RAFT holds the anchor until two real retained observations are available: OakInk2 has 142 bootstrap-held and 84,213 historically extrapolated withheld targets per EV.

G.2 Detailed Benchmark Results by Exposure and Temporal Subset

G.2.1 Interact: SMPLest-X

Table A7: Interact test set: SMPLest-X. Absolute and aligned errors by exposure and temporal subset (mm).

EV	Subset	MPJPE	PA	Body	Left	Right	Left PA	Right PA
-3	All	44.813	21.835	48.303	41.301	45.833	7.215	7.642
-3	Retained	43.636	20.781	47.984	40.171	43.996	7.190	7.607
-3	Hidden	45.399	22.359	48.462	41.864	46.746	7.227	7.659
-2	All	43.407	21.200	47.362	39.959	44.029	7.136	7.526
-2	Retained	42.181	20.125	47.022	38.771	42.134	7.109	7.490
-2	Hidden	44.016	21.734	47.531	40.550	44.972	7.150	7.544
-1	All	43.045	20.834	46.994	39.650	43.620	7.103	7.471
-1	Retained	41.718	19.735	46.594	38.349	41.605	7.076	7.434
-1	Hidden	43.705	21.381	47.192	40.297	44.622	7.117	7.489
0	All	42.791	20.697	46.733	39.377	43.388	7.073	7.432
0	Retained	41.492	19.593	46.351	38.102	41.412	7.044	7.393
0	Hidden	43.436	21.246	46.924	40.010	44.371	7.087	7.451
+1	All	43.098	20.785	46.622	39.871	43.807	7.077	7.420
+1	Retained	41.778	19.686	46.221	38.574	41.809	7.048	7.381
+1	Hidden	43.754	21.331	46.821	40.516	44.801	7.091	7.439
+2	All	44.020	21.710	47.380	40.695	44.946	7.247	7.657
+2	Retained	42.705	20.643	46.974	39.398	42.962	7.223	7.620
+2	Hidden	44.675	22.240	47.582	41.340	45.933	7.259	7.675
+3	All	48.230	24.596	48.707	45.158	50.961	7.765	8.277
+3	Retained	47.046	23.630	48.346	44.007	49.157	7.746	8.252
+3	Hidden	48.819	25.076	48.887	45.731	51.858	7.775	8.289

G.2.2 INTERACT: AIOS

Table A8: Interact test set: AiOS. Absolute and aligned errors by exposure and temporal subset (mm).

EV	Subset	MPJPE	PA	Body	Left	Right	Left PA	Right PA
-3	All	41.778	22.605	45.783	41.006	39.689	6.666	6.954
-3	Retained	39.898	21.503	45.071	39.165	36.937	6.639	6.911
-3	Hidden	42.712	23.153	46.137	41.922	41.057	6.680	6.976
-2	All	40.455	21.525	44.919	39.320	38.401	6.504	6.790
-2	Retained	38.514	20.377	44.185	37.418	35.560	6.476	6.742
-2	Hidden	41.420	22.097	45.284	40.265	39.813	6.518	6.814
-1	All	39.812	21.000	44.323	38.510	37.892	6.426	6.718
-1	Retained	37.839	19.828	43.574	36.579	35.002	6.397	6.670
-1	Hidden	40.794	21.583	44.695	39.471	39.330	6.441	6.742
0	All	39.180	20.527	43.778	37.427	37.649	6.361	6.676
0	Retained	37.169	19.328	43.019	35.449	34.711	6.330	6.627
0	Hidden	40.180	21.123	44.155	38.410	39.109	6.376	6.700
+1	All	39.483	20.842	43.811	38.010	37.864	6.405	6.677
+1	Retained	37.486	19.659	43.054	36.064	34.932	6.374	6.628
+1	Hidden	40.476	21.431	44.188	38.978	39.322	6.420	6.702
+2	All	40.037	21.588	44.283	38.521	38.522	6.645	6.829
+2	Retained	38.104	20.441	43.549	36.622	35.696	6.621	6.784
+2	Hidden	40.999	22.159	44.648	39.465	39.927	6.657	6.852
+3	All	42.429	24.243	45.541	40.857	41.779	7.321	7.699
+3	Retained	40.651	23.201	44.856	39.126	39.172	7.307	7.670
+3	Hidden	43.314	24.762	45.882	41.718	43.075	7.328	7.714

G.2.3 INTERACT: GEM-X

Table A9: Interact test set: GEM-X. Absolute and aligned errors by exposure and temporal subset (mm).

EV	Subset	MPJPE	PA	Body	Left	Right	Left PA	Right PA
-3	All	56.443	33.153	70.268	49.082	53.928	9.891	10.426
-3	Retained	55.305	32.511	69.881	47.863	52.334	9.867	10.403
-3	Hidden	57.008	33.472	70.461	49.687	54.720	9.903	10.437
-2	All	54.780	32.664	69.401	47.082	52.034	9.850	10.397
-2	Retained	53.671	32.049	69.027	45.901	50.473	9.829	10.376
-2	Hidden	55.331	32.971	69.586	47.669	52.811	9.861	10.408
-1	All	54.073	32.454	68.922	46.427	51.113	9.822	10.377
-1	Retained	52.975	31.850	68.553	45.260	49.563	9.801	10.357
-1	Hidden	54.619	32.754	69.105	47.008	51.883	9.832	10.387
0	All	53.607	32.375	68.361	45.976	50.701	9.805	10.369
0	Retained	52.534	31.790	68.000	44.834	49.186	9.786	10.350
0	Hidden	54.141	32.666	68.540	46.543	51.454	9.814	10.379
+1	All	54.035	32.432	68.511	47.007	50.724	9.844	10.429
+1	Retained	52.950	31.840	68.149	45.862	49.183	9.826	10.410
+1	Hidden	54.575	32.727	68.692	47.576	51.490	9.853	10.438
+2	All	56.095	33.341	68.803	50.044	53.070	9.933	10.518
+2	Retained	55.031	32.771	68.437	48.948	51.539	9.916	10.500
+2	Hidden	56.625	33.624	68.985	50.589	53.832	9.941	10.527
+3	All	60.204	35.238	70.651	54.483	58.462	10.181	10.834
+3	Retained	59.135	34.661	70.276	53.430	56.883	10.165	10.818
+3	Hidden	60.735	35.524	70.838	55.006	59.247	10.188	10.842

G.2.4 INTERACT: PEAR

Table A10: Interact test set: PEAR. Absolute and aligned errors by exposure and temporal subset (mm).

EV	Subset	MPJPE	PA	Body	Left	Right	Left PA	Right PA
-3	All	77.349	35.105	69.515	78.557	81.737	9.381	9.660
-3	Retained	76.341	34.473	69.156	77.544	80.270	9.378	9.642
-3	Hidden	77.850	35.420	69.694	79.061	82.466	9.383	9.668
-2	All	77.415	32.833	72.585	77.439	80.840	9.012	9.210
-2	Retained	76.388	32.153	72.234	76.405	79.337	9.009	9.191
-2	Hidden	77.925	33.171	72.759	77.954	81.587	9.013	9.220
-1	All	76.975	31.462	73.582	76.585	79.789	8.812	8.995
-1	Retained	75.892	30.732	73.211	75.516	78.183	8.808	8.976
-1	Hidden	77.514	31.826	73.767	77.117	80.587	8.814	9.005
0	All	81.510	30.296	79.132	80.733	83.986	8.455	8.799
0	Retained	80.455	29.569	78.790	79.691	82.409	8.448	8.783
0	Hidden	82.035	30.658	79.303	81.251	84.770	8.458	8.807
+1	All	77.932	29.947	76.393	76.981	79.981	8.537	8.830
+1	Retained	76.810	29.184	76.032	75.857	78.320	8.532	8.811
+1	Hidden	78.489	30.326	76.573	77.541	80.807	8.539	8.839
+2	All	80.916	31.536	78.502	80.521	83.037	8.671	9.100
+2	Retained	79.889	30.811	78.171	79.474	81.530	8.667	9.082
+2	Hidden	81.427	31.897	78.666	81.041	83.786	8.673	9.110
+3	All	85.954	37.017	76.598	86.788	91.804	9.122	9.897
+3	Retained	85.155	36.436	76.330	85.949	90.664	9.119	9.886
+3	Hidden	86.352	37.305	76.731	87.205	92.371	9.124	9.903

G.2.5 INTERACT: HAND4WHOLE++

Table A11: Interact test set: Hand4Whole++. Absolute and aligned errors by exposure and temporal subset (mm).

EV	Subset	MPJPE	PA	Body	Left	Right	Left PA	Right PA
-3	All	94.756	32.844	87.583	96.352	98.284	12.086	13.017
-3	Retained	93.701	32.089	87.299	95.322	96.652	12.078	13.012
-3	Hidden	95.281	33.220	87.724	96.864	99.095	12.091	13.019
-2	All	65.166	29.342	61.226	65.753	67.393	12.251	13.370
-2	Retained	63.912	28.505	60.815	64.539	65.497	12.245	13.364
-2	Hidden	65.790	29.758	61.431	66.357	68.336	12.254	13.373
-1	All	55.252	28.140	52.631	55.148	57.230	12.430	13.561
-1	Retained	53.913	27.272	52.171	53.858	55.212	12.423	13.556
-1	Hidden	55.919	28.571	52.859	55.790	58.233	12.433	13.563
0	All	53.161	27.789	50.867	52.669	55.291	12.610	13.529
0	Retained	51.773	26.909	50.374	51.336	53.209	12.604	13.526
0	Hidden	53.851	28.227	51.113	53.332	56.327	12.613	13.531
+1	All	53.409	27.900	50.887	53.091	55.529	12.588	13.577
+1	Retained	52.030	27.020	50.403	51.762	53.461	12.583	13.572
+1	Hidden	54.095	28.338	51.128	53.752	56.556	12.591	13.579
+2	All	53.710	28.630	52.121	52.678	55.877	12.736	13.747
+2	Retained	52.365	27.793	51.648	51.384	53.858	12.732	13.742
+2	Hidden	54.379	29.046	52.356	53.322	56.880	12.738	13.749
+3	All	83.774	34.814	79.080	84.487	86.415	12.882	14.021
+3	Retained	82.709	34.128	78.719	83.464	84.804	12.882	14.013
+3	Hidden	84.304	35.155	79.260	84.996	87.216	12.883	14.025

G.2.6 INTERACT: EVRGBHAND

Table A12: Interact test set: EvRGBHand. Absolute and aligned errors by exposure and temporal subset (mm).

EV	Subset	MPJPE	PA	Body	Left	Right	Left PA	Right PA
-3	All	68.391	35.997	68.783	67.163	69.338	10.602	11.041
-3	Retained	67.634	35.691	68.534	66.311	68.314	10.583	11.048
-3	Hidden	68.767	36.149	68.907	67.587	69.847	10.611	11.038
-2	All	67.461	35.671	69.490	65.521	67.951	10.520	10.953
-2	Retained	66.704	35.363	69.245	64.662	66.932	10.500	10.960
-2	Hidden	67.837	35.824	69.612	65.948	68.458	10.529	10.950
-1	All	67.461	35.636	70.556	65.282	67.429	10.506	10.938
-1	Retained	66.704	35.326	70.309	64.418	66.413	10.486	10.946
-1	Hidden	67.838	35.790	70.679	65.712	67.934	10.516	10.935
0	All	68.456	35.623	72.488	66.274	67.757	10.499	10.951
0	Retained	67.696	35.314	72.241	65.402	66.744	10.480	10.960
0	Hidden	68.833	35.776	72.611	66.707	68.261	10.508	10.947
+1	All	69.675	35.690	73.706	67.943	68.528	10.515	10.994
+1	Retained	68.920	35.381	73.467	67.074	67.519	10.496	11.004
+1	Hidden	70.050	35.843	73.824	68.375	69.030	10.525	10.989
+2	All	70.528	36.169	73.950	69.257	69.356	10.591	11.059
+2	Retained	69.794	35.879	73.711	68.423	68.367	10.571	11.070
+2	Hidden	70.893	36.313	74.068	69.671	69.848	10.601	11.053
+3	All	77.084	37.170	76.410	77.632	77.016	11.009	11.373
+3	Retained	76.330	36.882	76.156	76.795	75.988	10.991	11.382
+3	Hidden	77.458	37.312	76.535	78.049	77.527	11.018	11.369

G.2.7 INTERACT: EV2HANDS

Table A13: Interact test set: Ev2Hands. Absolute and aligned errors by exposure and temporal subset (mm). Because its input only includes event streams and not RGB images, there is no experimental data available for different EV values.

EV	Subset	MPJPE	PA	Body	Left	Right	Left PA	Right PA
Event	All	119.027	63.069	85.973	133.723	127.943	11.313	14.148
Event	Retained	118.957	63.028	86.001	133.758	127.695	11.305	14.141
Event	Hidden	119.063	63.090	85.958	133.705	128.066	11.317	14.151

G.2.8 INTERACT: EVENTHANDS

Table A14: Interact test set: EventHands. Absolute and aligned errors by exposure and temporal subset (mm). Because its input only includes event streams and not RGB images, there is no experimental data available for different EV values.

EV	Subset	MPJPE	PA	Body	Left	Right	Left PA	Right PA
Event	All	97.662	51.895	76.036	109.034	101.736	10.615	11.200
Event	Retained	97.622	51.706	76.117	109.010	101.594	10.613	11.205
Event	Hidden	97.682	51.989	75.996	109.046	101.807	10.616	11.198

G.2.9 INTERACT: LME

Table A15: Interact test set: LME. Absolute and aligned errors by exposure and temporal subset (mm). Because its input only includes event streams and not RGB images, there is no experimental data available for different EV values.

EV	Subset	MPJPE	PA	Body	Left	Right	Left PA	Right PA
Event	All	99.009	51.033	81.695	110.840	99.545	11.831	12.546
Event	Retained	98.793	50.826	81.642	110.771	99.067	11.831	12.536
Event	Hidden	99.116	51.136	81.721	110.874	99.782	11.831	12.550

G.2.10 INTERACT: SAM-BODY4D

Table A16: Interact test set: SAM-Body4D. Absolute and aligned errors by exposure and temporal subset (mm).

EV	Subset	MPJPE	PA	Body	Left	Right	Left PA	Right PA
-3	All	158.890	51.382	181.549	151.528	150.067	12.759	13.484
-3	Retained	158.234	50.885	181.398	150.807	149.114	12.757	13.482
-3	Hidden	159.216	51.630	181.624	151.886	150.540	12.759	13.485
-2	All	163.196	49.811	186.980	153.919	155.485	12.747	13.477
-2	Retained	162.557	49.288	186.842	153.204	154.562	12.745	13.475
-2	Hidden	163.514	50.072	187.049	154.274	155.944	12.748	13.478
-1	All	165.379	49.097	189.545	155.126	158.370	12.759	13.485
-1	Retained	164.757	48.556	189.416	154.425	157.475	12.757	13.483
-1	Hidden	165.688	49.366	189.609	155.475	158.814	12.759	13.486
0	All	163.750	48.248	188.394	151.452	158.444	12.913	13.572
0	Retained	163.137	47.691	188.274	150.744	157.576	12.911	13.570
0	Hidden	164.054	48.525	188.454	151.804	158.876	12.913	13.573
+1	All	165.708	49.088	191.275	153.883	159.270	12.817	13.546
+1	Retained	165.079	48.547	191.139	153.158	158.387	12.816	13.545
+1	Hidden	166.020	49.358	191.342	154.244	159.709	12.818	13.547
+2	All	159.654	54.070	189.656	145.309	152.570	12.833	13.606
+2	Retained	158.985	53.611	189.493	144.536	151.642	12.832	13.604
+2	Hidden	159.988	54.298	189.736	145.694	153.032	12.834	13.606
+3	All	161.839	59.877	197.024	150.590	147.955	12.894	13.679
+3	Retained	161.231	59.464	196.894	149.918	147.071	12.894	13.678
+3	Hidden	162.141	60.082	197.088	150.925	148.395	12.895	13.680

G.2.11 INTERACT: OURS

Table A17: Interact test set: Ours. Absolute and aligned errors by exposure and temporal subset (mm).

EV	Subset	MPJPE	PA	Body	Left	Right	Left PA	Right PA
-3	All	40.095	23.056	44.075	36.368	40.977	7.795	8.184
-3	Retained	38.425	22.093	43.448	34.372	38.891	7.750	8.123
-3	Hidden	40.925	23.535	44.387	37.361	42.015	7.817	8.214
-2	All	38.410	22.389	42.358	35.003	38.998	7.745	8.142
-2	Retained	36.690	21.391	41.718	32.938	36.851	7.700	8.082
-2	Hidden	39.265	22.885	42.676	36.030	40.065	7.768	8.172
-1	All	37.756	22.112	41.707	34.624	38.066	7.709	8.129
-1	Retained	36.008	21.099	41.059	32.533	35.875	7.664	8.068
-1	Hidden	38.625	22.616	42.030	35.664	39.155	7.731	8.160
0	All	37.809	22.218	41.870	34.748	37.969	7.657	8.123
0	Retained	36.057	21.207	41.221	32.652	35.774	7.612	8.059
0	Hidden	38.680	22.721	42.193	35.790	39.061	7.679	8.154
+1	All	37.665	22.072	41.345	34.611	38.090	7.693	8.121
+1	Retained	35.924	21.057	40.694	32.530	35.912	7.649	8.057
+1	Hidden	38.531	22.577	41.668	35.647	39.174	7.715	8.152
+2	All	40.764	23.396	43.889	37.543	41.753	7.885	8.273
+2	Retained	39.150	22.455	43.283	35.601	39.746	7.845	8.213
+2	Hidden	41.567	23.864	44.191	38.508	42.752	7.906	8.303
+3	All	48.098	26.897	45.457	46.733	51.349	8.454	9.016
+3	Retained	46.704	26.077	44.909	45.063	49.627	8.413	8.960
+3	Hidden	48.791	27.305	45.730	47.563	52.206	8.475	9.044