CADENCE: Temporally Coherent Whole-Body Human Mesh Recovery that Preserves Expressive Hand Motion

Keniel Peart Jagmohan Chauhan Indu P. Bodala Shelly Vishwakarma
University of Southampton, United Kingdom

Project page: https://keniel123.github.io/cadence Code: https://github.com/keniel123/cadence

Abstract

Whole-body SMPL-X estimators recover body, hand, and face meshes from single images, but applied frame by frame to video they produce jittery motion. Smoothing the output removes the jitter at a cost that body-only benchmarks do not measure: a temporal filter strong enough to steady the torso also flattens the fast finger strokes that carry meaning in sign language and gesture. We present CADENCE, a whole-body SMPL-X system for monocular video designed around this trade-off. A per-frame estimator built on a shared C-RADIOv4 encoder, with a high-resolution hand pathway that reuses the same encoder on hand crops, produces each frame's estimate. An offline temporal refiner then filters the SMPL-X rotation sequence: attention along the kinematic tree selects, for every joint and every frame, a mixture of learned zero-phase filters, so the torso is smoothed while fast finger strokes pass through. On 3DPW, the refiner reduces acceleration error from 17.4 to 5.4 mm/s^2 while lowering PA-MPJPE from 41.9 to 40.2 mm; generic smoothers applied to the same predictions reach 6.2 to 7.7 mm/s^2. On SignAvatars/How2Sign sign-language sequences, the refiner has the lowest hand jitter among the compared smoothers while keeping a hand-speed peak amplitude of 0.98, against 0.35 for a SmoothNet retrained on the full SMPL-X vector. The per-frame estimator also obtains the lowest full-body mesh error (25.7 mm) and hand joint error (8.3 mm) among the compared methods on EHF.

1. Introduction

Sign-language understanding, avatar animation, and human motion analysis need 3D motion of the body, hands, and face together [29, 45]. SMPL-X [29] represents all three, and image-based estimators now recover detailed SMPL-X meshes from single frames [3, 5, 22, 37]. Applied independently to video, these estimators produce motion that shakes: small per-frame errors become high-frequency jitter in the limbs and fingers and flicker in global orientation. In sign language the hands carry the meaning, so unstable finger articulation is the most damaging form of this failure.

Temporal mesh recovery is well studied for body-only SMPL [7, 17, 35, 39, 41], but these models do not represent the hand, jaw, and expression parameters of SMPL-X. The usual alternative for whole-body video is to run a per-frame SMPL-X estimator and smooth its output. Generic smoothers treat the pose vector as one homogeneous signal, and this is where they fail on expressive motion. Shape is constant within a clip and the torso moves slowly, while a signer's hand holds a handshape, moves sharply, and holds again. A single temporal bandwidth that removes body jitter also attenuates these hand strokes; a bandwidth wide enough to keep the strokes leaves body jitter in place. Body-only benchmarks such as 3DPW measure only the first half of this trade-off.

Two further properties make whole-body video harder than body-only video. Fingers occupy a small part of a full-body crop, so their articulation is hard to recover from whole-body features, and strong whole-body pipelines rely on hand crops or hand-specific encoders [25, 47]. The temporal stage also has to operate on rotations in a kinematic tree, where an error at the wrist moves every finger below it.

We present CADENCE, a whole-body SMPL-X system for monocular video. A per-frame estimator built on a shared C-RADIOv4 encoder [31] and a structured SMPL-X decoder produces each frame's estimate, and a high-resolution hand pathway re-encodes hand crops with the same encoder weights to refine finger pose. An offline temporal refiner then processes the SMPL-X rotation sequence. Transformer attention over time and along the kinematic tree predicts, for every joint and every frame, mixing weights over a small bank of learned zero-phase filters, from near-identity to strongly low-pass. The selection starts low-pass for body joints and near-identity for hand, forearm, and jaw joints, so that body and hands are smoothed at different rates, and it can change within a clip as a hand moves between held handshapes and fast strokes.

On 3DPW, the refiner reduces acceleration error from 17.4 to 5.4 mm/s^2 and lowers PA-MPJPE from 41.9 to 40.2 mm relative to the per-frame input; five generic smoothers applied to the same predictions reach 6.2 to 7.7 mm/s^2. On SignAvatars/How2Sign sign-language sequences, where the trade-off between stability and hand motion is visible, the refiner has the lowest hand jitter among the compared smoothers and keeps a hand-speed peak amplitude of 0.98, while a SmoothNet retrained on the full SMPL-X vector keeps 0.35. A single-rate version of the refiner, with one bandwidth shared by all joints, reaches comparable jitter (1.50 against 1.29) but keeps a peak amplitude of only 0.50.

Our contributions are:

- A part-structured temporal refiner for SMPL-X rotation sequences that uses kinematic-tree attention to choose a learned zero-phase filter for every joint and frame (Sec. 3.4), evaluated against five generic smoothers and a single-rate variant on 3DPW and on sign-language sequences (Sec. 4.2).
- A high-resolution hand pathway that refines finger pose from hand crops with the shared encoder and no separate hand encoder (Sec. 3.2), reducing EHF hand P-MVE from 22.6 to 12.2 mm (Sec. 4.7).
- An evaluation of the full system on expressive whole-body (EHF, AGORA, UBody), body-centric video (3DPW, EMDB), and hand-object (ARCTIC) benchmarks (Secs. 4.4 to 4.6).

2. Related Work

Single-image and whole-body human mesh recovery. Human mesh recovery was first developed mainly for body-only SMPL [24], with progress driven by iterative regression, optimization-in-the-loop training, part-guided attention, and camera-aware conditioning [15, 18, 19, 21, 27]. Transformer-based models further improved monocular recovery through stronger global features and token-based reasoning [9, 12, 32]. For expressive recovery, SMPL-X [29] extends the body model with articulated hands, face, jaw, expression, and shape. Early whole-body systems such as ExPose, Hand4Whole, and PyMAF-X [8, 25, 47] used separate crops or specialised branches for body, hands, and face, improving local detail but increasing pipeline complexity. Recent unified methods, including SMPLer-X [5], OSX [22], AiOS [37], Multi-HMR [3], PromptHMR [40], PEAR [42], and SAM 3D Body [44], process the person more holistically and achieve strong single-image results.

SAM 3D Body is the closest per-frame design to CADENCE's estimator: a shared image encoder, a transformer decoder, keypoint query tokens, and a hand-specific refinement at inference. It differs in its DINOv3 backbone and in predicting a Momentum Human Rig mesh, a different body model and topology from SMPL-X, and its released pipeline wraps the network with a separate person detector and field-of-view estimator. All of these single-image methods leave temporal coherence to post-processing when applied to video, which is the problem CADENCE's temporal refiner addresses.

Hand detail and multi-scale visual features. Fine hand articulation is difficult to recover from a full-body crop because fingers occupy only a small image region. Multi-scale representations help combine semantic context with local precision [23, 36], and pose-estimation work such as ViTPose [43] shows earlier transformer features retain useful localization detail. Many whole-body methods use high-resolution hand crops and hand-specific encoders [25, 47], improving hand detail but separating the hand estimate from global body context and adding system complexity. CADENCE instead reuses its shared visual encoder on hand crops to predict finger pose, retaining the body, camera, shape, and face predictions from the whole-body estimate.

Video-based and temporal mesh recovery. Temporal mesh recovery has advanced mainly for body-only SMPL, using temporal convolutions [16], adversarial motion priors [17], past–future fusion [7], per-joint temporal modelling [41], and world- or camera-aware trajectories [35, 39]. Post-hoc methods such as SmoothNet [46] and MotionBERT [50] can also smooth existing predictions, while HTD-Refine [2] improves motion through test-time optimization. These approaches are mainly designed and evaluated for body pose, whereas expressive SMPL-X video recovery must account for hands, face, and body together.

DanceHMR [33] is the closest whole-body video method to CADENCE: it recovers temporally coherent body and hand SMPL-X motion by injecting hand evidence as a residual into a body-centric token before a single shared temporal transformer, using frozen per-frame body and hand feature extractors and a close-up-aware training strategy that synthesizes truncated crops to handle vlog- and livestream-style upper-body footage. CADENCE targets the same failure mode, a shared temporal bandwidth that over-smooths fast hand motion, but addresses it differently: it refines pose directly in SMPL-X rotation space, offline, with kinematic-tree attention that selects a learned zero-phase filter for every joint and frame, a multi-rate design in the spirit of multi-rate video models such as SlowFast [11]. We also evaluate this trade-off directly, with hand-speed peak metrics on sign-language sequences (Sec. 4.2).

3. CADENCE

Overview. Given monocular video frames $\mathbf{I}_1, \ldots, \mathbf{I}_T$ and per-frame person boxes, CADENCE estimates a sequence of SMPL-X [29] parameters. SMPL-X maps body pose,

hand pose, facial expression, shape, and camera parameters to a mesh $\mathbf{V} \in \mathbb{R}^{10475\times 3}$. At each frame, we predict a 341-dimensional parameter vector

$$\mathbf{p}_t = [\underbrace{\boldsymbol{\theta}_t^{\text{body}}}_{126}, \underbrace{\boldsymbol{\theta}_t^{\text{lh}}}_{90}, \underbrace{\boldsymbol{\theta}_t^{\text{rh}}}_{90}, \underbrace{\boldsymbol{\theta}_t^{\text{jaw}}}_{6}, \underbrace{\boldsymbol{\psi}_t}_{10}, \underbrace{\boldsymbol{\theta}_t^{\text{or}}}_{6}, \underbrace{\boldsymbol{\beta}_t}_{10}, \underbrace{\mathbf{c}_t}_{3}], \quad (1)$$

where rotations use the continuous 6D representation [49]. The body pose excludes the root orientation, which is predicted separately as $\boldsymbol{\theta}_t^{\text{or}}$.

CADENCE has three parts. First, a per-frame whole-body estimator predicts an initial SMPL-X estimate from each image. Second, a high-resolution hand pathway reuses the same visual encoder on hand crops and predicts finger pose. Third, an offline temporal refiner processes the per-frame pose sequence. The hand branch changes only finger-pose parameters; wrist, body-pose, global orientation, shape, expression, face, and camera parameters are retained from the per-frame estimate.

3.1. Per-Frame Whole-Body Estimator

CADENCE uses C-RADIOv4-SO400M [31] as a shared visual encoder. The encoder is frozen during the main training stages and is adapted with LoRA [13] in the final phase. We extract early, middle, and late visual features. Early features preserve local detail, middle features capture limb structure, and late features provide semantic context for global pose, shape, and camera estimation. Each feature map is projected to decoder width $D = 512$.

The decoder uses structured tokens. Query tokens represent the quantities to be predicted: camera, shape, global orientation, body joints, hands, jaw, and facial expression. Keypoint tokens represent where the articulated body lies in the image. The tokens are updated jointly by a 6-layer transformer decoder with self-attention, cross-attention to late visual features, and feed-forward blocks. After each layer, lightweight heads predict SMPL-X updates. Pose parameters are updated residually for coarse-to-fine refinement, while shape, expression, and camera are predicted directly.

To make refinement spatially aware, CADENCE feeds its own predictions back into the decoder. The current SMPL-X estimate is converted to 3D joints by forward kinematics, projected to the image, and used to guide keypoint-token attention and local feature sampling in later layers. The decoder therefore progressively updates where it attends as the mesh estimate changes.

3.2. High-Resolution Hand Pathway

Hands occupy only a small part of a full-body crop, so finger articulation is difficult to recover from whole-body features alone. Crop-based pipelines improve hand detail using dedicated hand encoders, but this increases complexity and can separate the hand prediction from the global body context. CADENCE instead keeps a shared-encoder design.

Given the current whole-body prediction, we crop the left and right hand regions around the predicted wrists from the original image and encode them with the same frozen C-RADIOv4 weights, adapted by a hand-specific LoRA. A whole-body 2D keypoint detector, ViTPose [43], provides the hand keypoints $\mathbf{k}_{\text{h}}$ in each crop. Crop features $\mathbf{F}_{\text{h}}^{\text{hi}}$ are bilinearly sampled at these keypoints, and a learned hand query $\mathbf{q}_{\text{h}}$ cross-attends to the sampled features to predict the finger pose of that hand:

$$\hat{\boldsymbol{\theta}}_t^{\text{h}} = \mathcal{H}\big(\mathbf{q}_{\text{h}}, \mathbf{F}_{\text{h}}^{\text{hi}}(\mathbf{k}_{\text{h}})\big), \qquad \text{h} \in \{\text{lh}, \text{rh}\}. \quad (2)$$

The prediction replaces only the 15 finger rotations of each hand in $\mathbf{p}_t$; wrist, body-pose, global-orientation, shape, expression, jaw, and camera parameters are kept from the whole-body estimate, so body metrics are unchanged whether the pathway runs or not. Both hands share encoder and branch weights, and horizontal-flip augmentation covers both orientations.

Whole-body datasets provide little reliable supervision for finger articulation, so the pathway is trained as a hand-pose estimator on hand crops. Real 3D MANO annotations come from FreiHand [51], HanCo [52], and ReInterHand [26]. On COCO and UBody hand crops, where no reliable 3D ground truth exists, we distill finger pose from HaMeR [30], add a 2D reprojection loss on COCO, and down-weight UBody pseudo ground truth. Keypoints are jittered during training so that the pathway tolerates detector noise at inference.

3.3. Kinematic-Aware Multi-Rate Temporal Refiner

3.4. Kinematic-Aware Multi-Rate Temporal Refiner

A single smoothing filter cannot serve the whole body. A bandwidth that removes body jitter also blurs fast finger motion; a bandwidth that preserves finger motion leaves body jitter untouched. Sign language makes this worse: a signer holds a handshape, moves it sharply, then holds again, so even the right bandwidth for one joint changes over time. The refiner therefore chooses a temporal bandwidth for every joint and every frame, conditioned on the motion of the whole kinematic chain.

Input. The refiner takes the SMPL-X rotation sequence for a clip of T frames and reads out $J{=}53$ per-frame rotations: global orientation, the 21 body joints, the jaw, and both hands. Each rotation $\mathbf{r}_{j,t} \in \mathbb{R}^6$ is embedded together with its frame-to-frame velocity $\mathbf{r}_{j,t} - \mathbf{r}_{j,t-1}$ and acceleration $\mathbf{r}_{j,t+1} - 2\mathbf{r}_{j,t} + \mathbf{r}_{j,t-1}$, a learned joint identity, and a temporal position embedding, giving a token $\mathbf{h}_{j,t} \in \mathbb{R}^d$. The refiner uses pose only; it receives no image features.

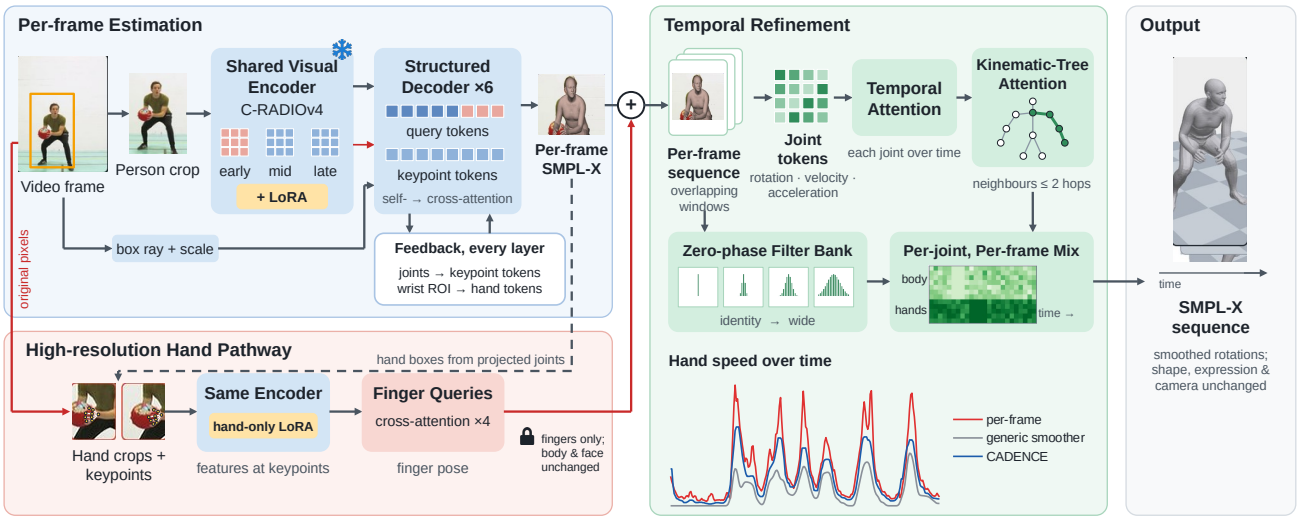


Figure 1. CADENCE overview. A shared C-RADIOv4 encoder and a structured decoder with keypoint-token feedback estimate SMPL-X per frame. The hand pathway re-encodes hand crops with the same encoder and predicts finger pose from features sampled at 2D hand keypoints, leaving the rest of the body unchanged. The temporal refiner uses attention over time and along the kinematic tree to choose, for every joint and frame, a mixture of learned zero-phase filters, smoothing the body while preserving fast hand strokes.

Kinematic-tree attention. Tokens are first updated by self-attention over time, one joint at a time, and then by self-attention over joints within a frame, masked to the SMPL-X kinematic tree:

$$\mathbf{h}_{:,t} \leftarrow \mathrm{Attn}(\mathbf{h}_{:,t}, \mathbf{h}_{:,t}, \mathbf{h}_{:,t}; M),$$
$$M_{jk} = \begin{cases} 0 & d_{\mathrm{tree}}(j,k) \le 2, \\ -\infty & \text{otherwise}, \end{cases} \tag{3}$$

where d_{tree} is the number of edges between joints j and k in the tree. A wrist token thus sees its elbow and its finger joints in one pass, so the bandwidth chosen for a finger can depend on whether the arm is moving.

Content-adaptive multi-rate filtering. A bank of $F{=}4$ learned kernels $\mathbf{w}_f \in \mathbb{R}^K$ ($K{=}9$) is applied to every joint's rotation track. Each kernel is symmetric, which makes it zero-phase, and is softmax-normalised, which makes it a non-negative low-pass filter:

$$\tilde{\mathbf{r}}^{(f)}_{j,t} = \sum_{k=-4}^{4} w_{f,k}\, \mathbf{r}_{j,t+k}. \tag{4}$$

The kernels are initialised to the identity and to Gaussians with $\sigma{=}1, 2, 4$ frames. For every joint and frame, a selector predicts mixing weights over the bank from the token and a clip-level motion context $\mathbf{c}$ (an MLP of the mean token):

$$\boldsymbol{\alpha}_{j,t} = \mathrm{softmax}\big(\mathbf{W}(\mathbf{h}_{j,t}{+}\mathbf{c}){+}\mathbf{b}_j\big), \qquad \hat{\mathbf{r}}_{j,t} = \sum_f \alpha_{j,t,f}\, \tilde{\mathbf{r}}^{(f)}_{j,t}, \tag{5}$$

followed by Gram-Schmidt re-orthonormalisation. This per-joint, per-frame choice of bandwidth is what we mean by *multi-rate*. The per-joint bias $\mathbf{b}_j$ is initialised to select the $\sigma{=}2$ kernel for body and orientation joints and the identity kernel for the jaw, forearms, and hands, and $\mathbf{W}$ is zero-initialised, so the untrained refiner is a fixed part-rate filter that keeps fast hand motion. Because every kernel is a symmetric convex filter, any mixture of them is too: the output never lags the input and never amplifies it. When a hand holds a handshape, the selector can move to a smoother kernel; during a stroke it can stay near the identity. The single-rate variant in Sec. 4.2 shares one kernel across all joints.

The refiner operates offline on 32-frame windows with a hop of 16 frames, blended with a Hann window; shape, expression, and camera pass through unchanged.

3.5. Training Objectives

The per-frame estimator is supervised at each decoder layer with a combination of rotation, 3D joint, 2D keypoint, vertex, camera, shape, and expression losses:

$$\mathcal{L}_{\mathrm{pf}} = \lambda_{\mathrm{rot}}\mathcal{L}_{\mathrm{rot}} + \lambda_{\mathrm{j3d}}\mathcal{L}_{\mathrm{j3d}} + \lambda_{\mathrm{j2d}}\mathcal{L}_{\mathrm{j2d}} + \lambda_{\mathrm{v}}\mathcal{L}_{\mathrm{v}} + \lambda_{\mathrm{cam}}\mathcal{L}_{\mathrm{cam}} + \lambda_{\beta}\mathcal{L}_{\beta} + \lambda_{\psi}\mathcal{L}_{\psi}. \tag{6}$$

The hand pathway is trained with finger rotation, hand joint, and hand vertex losses, plus the 2D reprojection term on COCO crops described above:

$$\mathcal{L}_{\mathrm{hand}} = \lambda_{\mathrm{hrot}}\mathcal{L}_{\mathrm{hrot}} + \lambda_{\mathrm{hj}}\mathcal{L}_{\mathrm{hj}} + \lambda_{\mathrm{hv}}\mathcal{L}_{\mathrm{hv}}. \tag{7}$$

The temporal refiner is trained on clips with the per-frame estimator frozen. It uses reconstruction and dynamics losses on forward-kinematic joints, together with a do-no-harm term that penalizes an increase in reconstruction error relative to the frozen per-frame estimate:

$$\mathcal{L}_{\text{temp}} = \mathcal{L}_{\text{pf}} + \lambda_{\text{dyn}}\mathcal{L}_{\text{dyn}} + \lambda_{\text{dnh}}\mathcal{L}_{\text{dnh}}. \tag{8}$$

Training proceeds in stages: first the per-frame decoder and prediction heads, then the hand pathway, then the temporal refiner on clips, and finally a short LoRA adaptation of the shared visual encoder. Section A gives the training schedule, window length, and data mixture.

4. Experiments

4.1. Experimental Setup

Datasets. The per-frame estimator is trained on AGORA [28], BEDLAM [4], UBody [22], Motion-X [48], and 3DPW [38], with 2D keypoint supervision from COCO-WholeBody [14] and MPII [1]. We add 500K frames of public 2D images from COCO-WholeBody, MPII, and InstaVariety [16] with pseudo-SMPL-X labels from the public SMPLer-X-L [5] checkpoint, rejecting frames whose SMPLer-X 2D reprojection error exceeds 20 pixels; no test images are used. The hand pathway uses the hand data in Sec. 3.2, and the temporal refiner trains on clips from the video sources. Section A.6 lists per-source counts.

Metrics. MPJPE and MPVPE (also written PVE or MVE) are mean Euclidean joint and vertex errors in millimetres; the PA- and P- prefixes denote errors after a similarity Procrustes alignment. For EHF [29], hand and face mesh errors are computed after wrist and neck alignment, following SMPLer-X [5]. 3DPW uses the standard 14-joint protocol and EMDB the 24-joint protocol; neither measures hands, absolute depth, or global placement. Acceleration error (Accel) is the mean difference between predicted and ground-truth joint accelerations, computed by central second differences over valid consecutive frames, in mm/s^2. On sign-language sequences, Jitter is the mean norm of the predicted joints' second difference, $\text{mean}_{t,j}\|\mathbf{p}_{t+2,j} - 2\mathbf{p}_{t+1,j} + \mathbf{p}_{t,j}\|_2$, in millimetres per frame step over valid three-frame runs, with body joints relative to the pelvis and hand joints relative to the wrist. It measures the smoothness of the prediction alone, without ground truth. Peak-R is the fraction of reference hand-velocity peaks that the prediction recovers, and Peak-A is the ratio of predicted to reference peak hand-velocity amplitude: 1 means matching peak size, values below 1 indicate attenuated strokes, and values above 1 overshoot. The sign-language reference motion is pseudo ground truth and contains jitter of its own, so we use these metrics to compare smoothers with each other rather than as absolute accuracy.

Table 1. Temporal smoothing on 3DPW. Every method smooths the same per-frame CADENCE predictions. Accel in mm/s^2, errors in mm.

Method	PA-MPJPE ↓	MPJPE ↓	Accel ↓
Per-frame (no smoothing)	41.9	61.7	17.4
Gaussian (σ=2)	41.5	61.1	7.0
Moving average (w=5)	41.5	61.1	6.6
One-Euro [6]	50.5	70.3	6.2
Savitzky–Golay (order 2, w=7)	41.6	61.4	7.6
SmoothNet-WB [46]	42.5	61.9	7.7
Single-rate refiner	41.7	61.2	5.7
CADENCE refiner	**40.2**	**58.1**	**5.4**

Implementation. The C-RADIOv4-SO400M encoder [31] has about 472M parameters and is frozen except for LoRA adapters added in the final training stage. The per-frame estimator has 51.8M trainable parameters, the hand pathway 21.0M, and the temporal refiner 1.88M. Person crops are 384×288. One model is used for all benchmarks, without per-benchmark fine-tuning. We chose C-RADIOv4 because it is distilled from several foundation models, including DINOv3 and SAM, which lets the whole-body decoder and the hand pathway share one backbone; we have not compared it against a human-specific backbone.

4.2. Temporal Coherence

This section asks whether the refiner stabilizes whole-body motion without erasing fast hand motion. All smoothers below take the same stored per-frame CADENCE predictions as input, so the rows differ only in temporal processing. Single-rate is the CADENCE refiner with one temporal bandwidth shared by all joints. SmoothNet-WB is SmoothNet [46] retrained on the full 341-dimensional SMPL-X vector, including the hand channels.

On 3DPW (Tab. 1), the refiner reduces acceleration error from 17.4 to 5.4 mm/s^2, the lowest in the table, and improves pose accuracy more than any other row: PA-MPJPE falls from 41.9 to 40.2 mm and MPJPE from 61.7 to 58.1 mm. The zero-phase classical filters reach 6.6 to 7.6 mm/s^2 with small accuracy changes. The causal One-Euro filter reaches 6.2 mm/s^2, but its phase lag raises PA-MPJPE to 50.5 mm. SmoothNet-WB reduces acceleration error to 7.7 mm/s^2 while slightly increasing pose error.

3DPW measures body joints only, so it cannot show what smoothing does to the hands. Table 2 does. Every smoother reduces hand jitter relative to the per-frame input (8.02), and the refiner reaches the lowest value (1.29). The smoothers

Table 2. Stability against hand motion on SignAvatars/How2Sign [45] sign-language test sequences. Errors in mm. Jitter measures residual hand acceleration; Peak-R and Peak-A measure how many hand-speed peaks survive and how much of their amplitude is kept. The reference motion is pseudo ground truth.

Method	PA-MPJPE ↓	MPJPE ↓	Jitter ↓	H-MPJPE ↓	Peak-R ↑	Peak-A ↑
Per-frame	10.72	12.64	8.021	19.04	0.93	1.18
SmoothNet-WB	10.13	11.95	1.590	17.82	0.81	0.35
Gaussian (σ=2)	10.32	12.20	1.426	18.16	0.62	0.49
One-Euro	**9.17**	**10.78**	1.894	**17.24**	0.83	0.66
Moving avg. (w=5)	10.46	12.34	2.221	18.41	0.78	0.58
Savitzky–Golay	10.66	12.56	3.250	18.82	0.83	0.76
Single-rate refiner	10.35	12.23	1.497	18.21	0.62	0.50
CADENCE refiner	10.28	12.16	**1.286**	18.68	**0.90**	**0.98**

differ in what they remove with the jitter. SmoothNet-WB and the Gaussian filter keep only 0.35 and 0.49 of the peak hand speed; the refiner keeps 0.98 and recovers 0.90 of the peaks, against 0.93 for the unsmoothed input. The single-rate refiner reaches jitter similar to the Gaussian filter (1.50) and keeps a similar peak amplitude (0.50), so the difference between the two refiner rows comes from the per-joint bandwidth. The One-Euro filter has the lowest pose errors on this benchmark but keeps less peak amplitude (0.66) at higher jitter (1.89). On a single sign-language clip where a Savitzky–Golay filter is tuned to match the refiner's residual wrist jitter, the refiner keeps 99% of the per-frame peak finger speed and the matched filter 46% (Fig. 6). Among the compared smoothers, the refiner therefore gives the lowest hand jitter while keeping nearly all hand-speed peak amplitude.

Diagnostics on matched clip subsets (Sec. D) confirm the per-joint bandwidth as the source of peak preservation: fixed filters with one shared bandwidth and with per-joint bandwidths reach the same jitter, but only the latter keeps the hand-speed peaks. The same section reports robustness to dropped frames and lower frame rates.

4.3. Body Accuracy and Smoothness on 3DPW and EMDB

Table 3 places CADENCE among published methods on the body-only protocols of 3DPW and EMDB. Accel values for other methods are taken from their papers, and dashes mark values that were not reported.

CADENCE's acceleration error (5.4 mm/s^2) is lower than that of TCMR, MPS-Net, WHAM, and TRAM (6.7 to 7.4) and slightly higher than that of GVHMR and GENMO (5.0 and 5.2), which are body-only and do not predict hands or face. CADENCE has the lowest 3DPW PVE (62.6 mm). Its joint errors are higher than those of the strongest body-focused methods: on 3DPW, SAM 3D Body and NLF-L+fit reach 33.2 and 33.6 mm PA-MPJPE against CADENCE's 40.2 mm, and DanceHMR reaches 53.6 mm MPJPE against 58.1 mm. On EMDB, CADENCE has the lowest PA-MPJPE (37.2 mm) but higher MPJPE and PVE than SAM 3D Body and DanceHMR, which points to absolute depth and placement, not articulation, as the weaker part of the estimate. These protocols score body joints only and do not measure the hand and face output.

4.4. Expressive Whole-Body Accuracy on EHF

Table 4 reports EHF, the standard expressive whole-body SMPL-X benchmark. CADENCE has the lowest full-body P-MVE (25.7 mm), body P-MPJPE (20.6 mm), and hand P-MPJPE (8.3 mm) among the listed methods. Its hand mesh error (12.2 mm) is higher than that of PyMAF-X (10.2 mm), which uses dedicated hand crops and a separate hand branch, and Multi-HMR (10.8 mm); SMPLer-X-L has the lowest face P-MVE. Lower joint error with higher mesh error suggests that the pathway recovers finger articulation better than finger surface shape.

4.5. Whole-Body Generalization on AGORA and UBody

Table 5 evaluates whole-body mesh error on AGORA validation and UBody. CADENCE has the lowest full-body PA-PVE and PVE on both datasets and the lowest UBody hand PA-PVE (5.9 mm). DanceHMR has the lowest UBody hand PVE (22.2 against 28.2 mm); since that metric is not Procrustes-aligned, it also scores wrist placement, which the CADENCE hand pathway does not change.

4.6. Hand-Object Video on ARCTIC

EHF, AGORA, and UBody use fitted or pseudo ground truth for hands. ARCTIC [10] provides accurate 3D hand ground truth on video of two-handed object manipulation, a setting absent from CADENCE's training data. We evaluate the unchanged model on 8 held-out sequences (296 frames), using crops built from ground-truth joint boxes because ARCTIC's own crops are too tight for a whole-body model. CADENCE reaches body and hand PA-MPJPE of 49.80 and 32.02 mm, well above its error on EHF, AGORA, and UBody. Close-range manipulation and object occlusion are outside the training distribution, and this result is consistent with the occlusion sensitivity measured on AGORA (Sec. 4.8).

4.7. Hand Pathway Ablation

Table 6 traces the hand pathway's development on EHF. Body P-MPJPE is identical in every row, because each variant writes only finger rotations. A first version that re-encodes hand crops at high resolution and predicts a zero-initialized residual on the finger pose reduces hand P-MVE from 22.6 to 16.6 mm; adding FreiHand and HanCo changes EHF error by only 0.1 mm, since EHF hands are already covered by the pseudo-labelled training data. The

Table 3. In-the-wild body benchmarks (3DPW 14-joint protocol, EMDB 24-joint protocol). Errors in mm, Accel in mm/s^2. CADENCE predicts whole-body SMPL-X; some reference methods are body-only. Results of other methods are as reported in their papers; DanceHMR has no public code. Best results are **bold**, second-best underlined.

		3DPW				EMDB		
Method	Repr.	PA-MPJPE ↓	MPJPE ↓	PVE ↓	Accel ↓	PA-MPJPE ↓	MPJPE ↓	PVE ↓
Whole-body SMPL-X methods								
Multi-HMR [3]	SMPL-X	41.7	61.4	–	–	–	–	–
DanceHMR [33]	SMPL-X video	35.2	**53.6**	65.1	–	40.1	64.3	75.4
CADENCE per-frame	SMPL-X	41.9	61.7	–	17.4	–	–	–
CADENCE	SMPL-X video	40.2	58.1	**62.6**	5.4	**37.2**	69.5	76.0
Body-focused per-frame or fitting methods								
PromptHMR [40]	SMPL-X	36.1	58.7	69.4	–	41.0	71.7	84.5
NLF-L+fit [32]	SMPL	33.6	54.9	63.7	–	40.9	68.4	80.6
SAM 3D Body [44]	Body	**33.2**	54.8	63.6	–	38.2	**61.7**	**72.5**
Body-only temporal methods								
TCMR [7]	SMPL video	55.8	95.0	–	6.7	–	–	–
MPS-Net [41]	SMPL video	52.1	84.3	–	7.4	–	–	–
WHAM [35]	SMPL video	35.9	57.8	68.7	6.8	–	–	–
TRAM [39]	SMPL video	35.6	59.3	69.6	7.1	–	–	–
GVHMR [34]	SMPL video	36.2	55.6	67.2	**5.0**	–	–	–
GENMO [20]	SMPL video	34.6	53.9	65.8	5.2	42.5	73.0	84.8



Figure 2. CADENCE reconstructions on samples from 3DPW [38], UBody [22], BEDLAM [4], and in-the-wild images, covering indoor and outdoor scenes, seated and standing poses, sports motion, occlusion, and expressive upper-body motion. Section F shows further examples and failure cases.

Table 4. EHF [29] whole-body evaluation (mm). Hand and face errors are wrist- and neck-aligned. CADENCE full-body P-MVE is measured without the hand pathway. Best **bold**, second underlined. Section C gives hand pathway diagnostics.

	P-MVE ↓			P-MPJPE ↓	
Method	Full	Hand	Face	Body	Hand
PyMAF-X [47]	50.2	**10.2**	5.5	52.8	10.3
SMPLer-X-L [5]	37.1	14.1	**5.0**	–	–
Multi-HMR [3]	28.2	10.8	5.3	–	–
NLF-L+fit [32]	26.0	20.7	6.3	26.1	21.8
CADENCE	**25.7**	12.2	5.1	**20.6**	**8.3**

Table 5. Whole-body mesh error (mm) on AGORA [28] validation and UBody [22]. Best per column **bold**, second underlined. DanceHMR values are as reported.

	AGORA val		UBody			
	All		All		Hand	
Method	PA-PVE	PVE	PA-PVE	PVE	PA-PVE	PVE
OSX [22]	69.4	168.6	42.4	92.4	10.8	47.7
SMPLer-X-L [5]	45.1	74.2	33.2	61.5	10.6	43.3
DanceHMR [33]	–	–	31.3	61.0	7.4	**22.2**
CADENCE	**42.4**	**70.8**	**29.3**	**58.3**	**5.9**	28.2

final pathway of Sec. 3.2 samples crop features at ViTPose hand keypoints, predicts the finger pose directly, and adds ReInterHand and HaMeR distillation; it reduces hand P-MVE to 12.2 mm and hand P-MPJPE from 14.5 to 8.3 mm. Because the last row changes several factors at once, it does

Table 6. Hand pathway development on EHF (mm). The second and third rows are cumulative versions of a residual crop pathway; the last row is the final keypoint-guided pathway, which also changes the supervision. Only finger rotations change; all other SMPL-X parameters are fixed.

	Hand	
Configuration	P-MVE ↓	P-MPJPE ↓
No hand pathway	22.6	14.5
Residual crop pathway, pseudo-GT	16.6	11.1
+ FreiHand/HanCo	16.5	11.0
Keypoint-guided pathway (final)	**12.2**	**8.3**

not isolate any single one; Sec. C compares feature source and parameterization under a matched budget, and Fig. 3 shows examples before and after the pathway.

4.8. Runtime and Occlusion

On one A100-80GB GPU in bf16, the per-frame estimator runs at 27.9 FPS at batch size 1 and 307.6 FPS at batch size 32. With the hand pathway, which encodes two extra crops per frame, the full system runs at 13.1 and 71.5 FPS. The temporal refiner processes a 32-frame window at about 6,900 FPS, so its cost is negligible, and peak memory stays below 4 GiB (Tab. 19). Applications that do not need finger detail can disable the hand pathway and keep the per-frame rate.

To measure sensitivity to occlusion, we split 500 AGORA validation samples by their annotated occlusion percentage, which on AGORA comes mostly from other people in the scene. Body PA-MPJPE rises from 32.87 mm at low occlusion ($< 20\%$, 224 samples) to 42.11 mm at medium (20–50%, 136 samples) and 57.64 mm at high occlusion ($\geq 50\%$, 140 samples); the Pearson correlation between occlusion percentage and error is 0.539.

4.9. Limitations

CADENCE processes one tracked person per box and does not detect people or resolve identity in multi-person scenes. The temporal refiner runs offline on fixed-length windows, so it is not causal and cannot be used for streaming without modification. Global orientation and limb depth remain ambiguous under the weak-perspective camera, foreshortening, and self-occlusion, which is consistent with the gap between CADENCE's MPJPE and PA-MPJPE on 3DPW and EMDB. Hands are the least stable region under motion blur, hand-object contact, and hand-hand occlusion, as the ARCTIC result shows, and the hand pathway depends on an external 2D keypoint detector. The sign-language dynamics metrics rely on pseudo ground truth. Section F.5 summarizes failure modes.

5. Conclusion

CADENCE recovers whole-body SMPL-X motion from monocular video with a per-frame estimator, a shared-encoder hand pathway, and an offline temporal refiner that smooths each joint at its own learned rate. On 3DPW, the refiner lowers acceleration error from 17.4 to 5.4 mm/s^2 while improving PA-MPJPE, and on sign-language sequences it reaches the lowest hand jitter among the compared smoothers while keeping 0.98 of the hand-speed peak amplitude, where generic and single-rate smoothing keep 0.35 to 0.50. Body-only video benchmarks do not measure this second property, and our results suggest that whole-body video methods should be evaluated on hand dynamics as well as on body acceleration. A causal version of the refiner for streaming input, and evaluation on hand dynamics with accurate rather than pseudo ground truth, are the most direct next steps.

Acknowledgements. The authors acknowledge the use of the IRIDIS X High Performance Computing Facility and the Southampton-Wolfson AI Research Machine (SWARM) GPU cluster, funded by the Wolfson Foundation, at the University of Southampton.

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 5, 14

[2] Anonymous. Natural human motion recovery by aligning high-order temporal dynamics from monocular videos. *arXiv preprint arXiv:2605.26879*, 2026. 2

[3] Fabien Baradel, Philippe Weinzaepfel, Vincent Lepetit, and Romain Brégier. Multi-HMR: Multi-person whole-body human mesh recovery in a single shot. In *ECCV*, 2024. 1, 2, 7, 15, 16, 18, 21

[4] Michael J. Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. BEDLAM: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *CVPR*, 2023. 5, 7, 14

[5] Zhongang Cai, Wanqi Ren, Ailing Zeng, Zhengyu Chen, Tian Fang, Chi Han, Haoran Pang, Haiyi Dong, Bo Jiang, Ziwei Liang, et al. SMPLer-X: Scaling up expressive human pose and shape estimation. In *NeurIPS*, 2023. 1, 2, 5, 7, 15, 16, 18, 21

[6] Géry Casiez, Nicolas Roussel, and Daniel Vogel. One euro filter: A simple speed-based low-pass filter for noisy input in interactive systems. page 2527–2530, New York, NY, USA, 2012. Association for Computing Machinery. 5

[7] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. Beyond static features for temporally consistent 3D human pose and shape from video. In *CVPR*, 2021. 1, 2, 7, 15

[8] Vasileios Choutas, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. Monocular expressive

body regression through body-driven attention. In *ECCV*, 2020. 2, 15

[9] Sai Kumar Dwivedi, Yucheng Yu, Sergey Prokudin, and Michael J. Black. TokenHMR: Advancing human mesh recovery with a tokenized pose representation. In *CVPR*, 2024. 2

[10] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J. Black, and Otmar Hilliges. ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In *CVPR*, 2023. 6

[11] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. SlowFast networks for video recognition. In *ICCV*, 2019. 2

[12] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4D: Reconstructing and tracking humans with transformers. In *ICCV*, 2023. 2

[13] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022. 3

[14] Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. Whole-body human pose estimation in the wild. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 5, 14

[15] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018. 2

[16] Angjoo Kanazawa, Jason Y. Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *CVPR*, 2019. 2, 5

[17] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. VIBE: Video inference for human body pose and shape estimation. In *CVPR*, 2020. 1, 2

[18] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. PARE: Part attention regressor for 3D human body estimation. In *ICCV*, 2021. 2

[19] Nikos Kolotouros, Georgios Pavlakos, Michael J. Black, and Kostas Daniilidis. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In *ICCV*, 2019. 2

[20] Jiaman Li, Yixuan Zhang, et al. GENMO: A generalist model for human motion. *arXiv preprint*, 2025. 7, 15

[21] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. CLIFF: Carrying location information in full frames into human pose and shape estimation. In *ECCV*, 2022. 2

[22] Jing Lin, Ailing Zeng, Haoqian Wang, Lei Zhang, and Yu Li. One-stage 3D whole-body mesh recovery with component aware transformer. In *CVPR*, 2023. 1, 2, 5, 7, 14, 21

[23] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 2

[24] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *TOG*, 34(6), 2015. 2

[25] Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Accurate 3D hand pose estimation for whole-body 3D human mesh estimation. In *CVPRW*, 2022. 1, 2

[26] Gyeongsik Moon, Shunsuke Saito, Weipeng Xu, Rohan Joshi, Julia Buffalini, et al. A dataset of relighted 3D interacting hands. In *NeurIPS Datasets and Benchmarks Track*, 2023. 3

[27] Priyanka Patel and Michael J. Black. CameraHMR: Aligning people with perspective. In *CVPR*, 2024. 2, 15

[28] Priyanka Patel, Chun-Hao P. Huang, Joachim Tesch, David T. Hoffmann, Shashank Tripathi, and Michael J. Black. AGORA: Avatars in geography optimized for regression analysis. In *CVPR*, 2021. 5, 7, 14

[29] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *CVPR*, 2019. 1, 2, 5, 7

[30] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3D with transformers. In *CVPR*, 2024. 3

[31] Mike Ranzinger, Greg Heinrich, Collin McCarthy, Jan Kautz, Andrew Tao, Bryan Catanzaro, and Pavlo Molchanov. C-radiov4 (tech report), 2026. 1, 3, 5

[32] István Sárándi and Gerard Pons-Moll. Neural localizer fields for continuous 3D human pose and shape estimation. In *NeurIPS*, 2024. 2, 7, 15

[33] Wenhao Shen, Ming Zhou, Hengyuan Zhang, Siyuan Bian, Youjiang Xu, and Yuan Zhang. Dancehmr: Hand-aware whole-body human mesh recovery from monocular videos, 2026. 2, 7, 15

[34] 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 Conference Papers*, 2024. 7, 15

[35] Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J. Black. WHAM: Reconstructing world-grounded humans with accurate 3D motion. In *CVPR*, 2024. 1, 2, 7, 15

[36] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019. 2

[37] Qingping Sun et al. AiOS: All-in-one-stage expressive human pose and shape estimation. In *CVPR*, 2024. 1, 2

[38] Timo von Marcard, Roberto Henschel, Michael J. Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3D human pose in the wild using IMUs and a moving camera. In *ECCV*, 2018. 5, 7, 14, 15, 20

[39] Yufu Wang, Soyong Shin, and Michael J. Black. TRAM: Global trajectory and motion of 3D humans from in-the-wild videos. In *ECCV*, 2024. 1, 2, 7, 15

[40] Yufu Wang, Yu Sun, Priyanka Patel, Kostas Daniilidis, Michael J. Black, and Muhammed Kocabas. PromptHMR: Promptable human mesh recovery. In *CVPR*, 2025. 2, 7, 15

[41] Lin Wei, Jian Liao, Jianbo Ren, Zhipin Piao, and Lei Zhang. Capturing humans in motion: Temporal-attentive 3D human pose and shape estimation from monocular video. In *CVPR*, 2022. 1, 2, 7, 15

[42] Jiahao Wu, Yunfei Liu, Lijian Lin, Ye Zhu, Lei Zhu, Jingyi Li, and Yu Li. Pear: Pixel-aligned expressive human mesh recovery, 2026. 2, 15

[43] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. ViTPose: Simple vision transformer baselines for human pose estimation. In *NeurIPS*, 2022. 2, 3

[44] Xitong Yang, Devansh Kukreja, Don Pinkus, et al. SAM 3D Body: Robust full-body human mesh recovery. *arXiv preprint arXiv:2602.15989*, 2026. 2, 7, 15, 18, 21

[45] Zhengdi Yu, Shaoli Huang, Yongkang Chen, and Tolga Schmid. SignAvatars: A large-scale 3D sign language holistic motion dataset and benchmark. In *ECCV*, 2024. 1, 6

[46] Ailing Zeng, Lei Yang, Xuan Ju, Jiefeng Li, Jianyi Wang, and Qiang Xu. SmoothNet: A plug-and-play network for refining human poses in videos. In *ECCV*, 2022. 2, 5

[47] 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. *PAMI*, 2023. 1, 2, 7, 15

[48] Jing Zhang, Jiahang Li, Zhongang Li, Bo Pang, Tao Jiang, Lei Yang, et al. Motion-X: A large-scale 3D expressive whole-body human motion dataset. In *NeurIPS*, 2024. 5, 14

[49] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *CVPR*, 2019. 3

[50] Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. MotionBERT: A unified perspective on learning human motion representations. In *ICCV*, 2023. 2

[51] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. FreiHAND: A dataset for markerless capture of hand pose and shape from single RGB images. In *ICCV*, 2019. 3, 14

[52] Christian Zimmermann, Max Argus, and Thomas Brox. Contrastive representation learning for hand shape estimation. In *GCPR*, 2021. 3, 14

A. Implementation and Reproducibility Details

This section lists which components are frozen and trained, the supervision used at each training stage, and the evaluation protocols.

A.1. Model Inventory and Parameter Accounting

Table 7 summarizes the main components of CADENCE. The C-RADIOv4 encoder is shared by the per-frame estimator and the high-resolution hand pathway. The backbone weights remain frozen; in the final adaptation stage, only lightweight LoRA adapter weights are trained. The temporal refiner is trained after the per-frame estimator and hand pathway are fixed.

Table 7. CADENCE implementation inventory. The frozen visual backbone is reported separately from the trainable task-specific modules.

Component	Role	Implementation	Trainable
C-RADIOv4-SO400M encoder	Shared image encoder	Processes full-person crops and hand crops; provides early, middle, and late visual features.	No
Structured SMPL-X decoder	Per-frame prediction	Uses structured query tokens for camera, shape, global orientation, body, hands, jaw, and expression, with keypoint-token spatial feedback.	Yes
Prediction heads	Parameter regression	Predict residual pose updates and direct shape, expression, and camera parameters at each decoder layer.	Yes
High-resolution hand pathway	Local hand refinement	Crops hands from the original image, reuses the same encoder weights, and predicts finger rotations only.	Yes
LoRA adapters	Lightweight backbone adaptation	Inserted into the C-RADIOv4 encoder's attention and feed-forward layers in the final training stage; the encoder's original pretrained weights stay frozen and only the added low-rank adapter weights are trained.	Yes
Kinematic temporal refiner	Temporal coherence	Refines stored per-frame SMPL-X sequences: temporal and kinematic-tree attention select a learned zero-phase filter for every joint and frame.	Yes

Parameter accounting. Table 8 reports the parameter count for each component. The frozen C-RADIOv4-SO400M encoder contains 472.35M parameters and is not updated during the main training stages. The per-frame SMPL-X estimator contains 51.8M trainable parameters, including the structured decoder, prediction heads, spatial feedback modules, and lightweight adaptation parameters where enabled. The high-resolution hand pathway adds 21.0M trainable parameters. The temporal refiner is lightweight, adding only 1.88M parameters. Thus, the per-frame and hand modules contain 72.8M trainable parameters, and full CADENCE including the temporal refiner contains 74.7M trainable parameters.

Table 8. CADENCE parameter accounting. The frozen visual backbone is reported separately from the trainable task-specific modules.

Component	Parameters	Trainable
C-RADIOv4-SO400M visual encoder	472.35M	No
Per-frame SMPL-X estimator	51.8M	Yes
High-resolution hand pathway	21.0M	Yes
Kinematic temporal refiner	1.88M	Yes
Per-frame + hand modules	72.8M	Yes
Full CADENCE trainable total	74.7M	Yes

Per-frame estimator. The per-frame estimator takes a person crop, resized to 384×288, and predicts an initial SMPL-X parameter vector. We use C-RADIOv4-SO400M as the shared visual encoder and extract multi-scale visual features. Early features preserve local spatial detail, middle features capture limb structure, and late features provide semantic context. Each feature map is projected to the decoder width before cross-attention.

The decoder uses structured query tokens rather than a single global regression token. Separate tokens represent camera, shape, global orientation, body joints, left hand, right hand, jaw, and expression. Keypoint tokens represent the projected articulated body in image space. After each decoder layer, lightweight heads predict intermediate SMPL-X updates. The current SMPL-X estimate is converted to joints by forward kinematics, projected to the image, and used to update keypoint-token attention and local feature sampling in later decoder layers. This creates an iterative spatial-feedback loop in which the model refines both the parameter estimate and the image locations it uses for evidence.

A.2. SMPL-X Parameterization

CADENCE predicts a 341-dimensional SMPL-X parameter vector for each frame:

$$\mathbf{p}_t = [\boldsymbol{\theta}_t^{body}, \boldsymbol{\theta}_t^{lh}, \boldsymbol{\theta}_t^{rh}, \boldsymbol{\theta}_t^{jaw}, \boldsymbol{\psi}_t, \boldsymbol{\theta}_t^{or}, \boldsymbol{\beta}_t, \mathbf{c}_t],$$

where body pose contains 21 body joints in 6D rotation form ($21{\times}6 = 126$), the left and right hands contain 15 joints each ($15{\times}6 = 90$ per hand), the jaw is represented by one 6D rotation, expression has 10 dimensions, global orientation is one 6D rotation, body shape has 10 dimensions, and the weak-perspective camera has 3 dimensions. Rotations use the continuous 6D representation during learning and are projected to valid rotation matrices before SMPL-X forward kinematics.

Table 9. CADENCE per-frame SMPL-X parameterization.

Parameter group	Dim.	Description
Body pose $\boldsymbol{\theta}^{body}$	126	21 body joints, 6D rotations
Left hand $\boldsymbol{\theta}^{lh}$	90	15 hand joints, 6D rotations
Right hand $\boldsymbol{\theta}^{rh}$	90	15 hand joints, 6D rotations
Jaw $\boldsymbol{\theta}^{jaw}$	6	1 jaw joint, 6D rotation
Expression $\boldsymbol{\psi}$	10	SMPL-X expression coefficients
Global orient. $\boldsymbol{\theta}^{or}$	6	Root orientation, 6D rotation
Shape $\boldsymbol{\beta}$	10	SMPL-X shape coefficients
Camera $\mathbf{c}$	3	Weak-perspective camera
Total	341	Full CADENCE vector

A.3. High-Resolution Hand Pathway

Section 3.2 describes the pathway. FreiHand and HanCo contain right hands only; horizontal-flip augmentation exposes the shared branch to both orientations, so no separate left-hand branch or explicit mirroring step is needed. Figure 3 shows examples before and after the pathway.

A.4. Temporal Refiner Details

The temporal refiner is trained after the per-frame and hand components are fixed. It operates on stored per-frame CADENCE predictions rather than raw images. Each SMPL-X joint rotation is treated as a token over time. Tokens are first updated with temporal self-attention and then with attention restricted to the SMPL-X kinematic tree, allowing information to pass between each joint, its parent, and its children.

Each joint chooses its bandwidth per frame from a bank of four learned zero-phase kernels, initialised as the identity and Gaussians with $\sigma{=}1, 2, 4$ frames. The selector's per-joint prior starts body and global-orientation joints on the $\sigma{=}2$ kernel and the jaw, forearms, and hands on the identity kernel, and its learned weights are zero-initialised, so training starts from a fixed part-rate filter. Every kernel is symmetric and non-negative with unit sum, so the output cannot lag or amplify the input. The filtered rotations are re-orthonormalised before SMPL-X forward kinematics. The refiner has 1.88M parameters, uses pose only, and runs on 32-frame windows with a hop of 16 frames and Hann-weighted overlap.

A.5. Losses and Training Schedule

The per-frame estimator is supervised at each decoder layer using rotation, joint, vertex, 2D keypoint, camera, shape, and expression losses:

$$\mathcal{L}_{pf} = \lambda_{rot}\mathcal{L}_{rot} + \lambda_{j3d}\mathcal{L}_{j3d} + \lambda_{j2d}\mathcal{L}_{j2d} + \lambda_v\mathcal{L}_v + \lambda_{cam}\mathcal{L}_{cam} + \lambda_\beta\mathcal{L}_\beta + \lambda_\psi\mathcal{L}_\psi.$$

The hand pathway adds losses on finger rotations, hand joints, and hand vertices:

$$\mathcal{L}_{hand} = \lambda_{hrot}\mathcal{L}_{hrot} + \lambda_{hj}\mathcal{L}_{hj} + \lambda_{hv}\mathcal{L}_{hv}.$$

The temporal refiner is trained with reconstruction, velocity, acceleration, and do-no-harm objectives:

$$\mathcal{L}_{temp} = \mathcal{L}_{pf} + \lambda_{dyn}\mathcal{L}_{dyn} + \lambda_{dnh}\mathcal{L}_{dnh}.$$

The do-no-harm term is a hinge loss that penalizes temporal predictions when they increase reconstruction error relative to the frozen per-frame estimate.

Table 10. Loss terms used to train CADENCE. Losses are enabled according to annotation availability for each dataset and training stage.

Loss	Applied to	Purpose
$\mathcal{L}_{rot}$	SMPL-X rotations	Pose supervision
$\mathcal{L}_{j3d}$	3D joints	Body/hand joint alignment
$\mathcal{L}_{j2d}$	Projected 2D joints	Image-space alignment
$\mathcal{L}_{v}$	SMPL-X vertices	Mesh alignment
$\mathcal{L}_{cam}$	Camera parameters	Weak-perspective camera supervision
$\mathcal{L}_{\beta}$	Shape	Shape regularization/supervision
$\mathcal{L}_{\psi}$	Expression	Face expression supervision
$\mathcal{L}_{hrot}$	Hand rotations	Finger articulation
$\mathcal{L}_{hj}$	Hand joints	Hand joint alignment
$\mathcal{L}_{hv}$	Hand vertices	Hand mesh alignment
$\mathcal{L}_{dyn}$	Joint velocity/acceleration	Temporal smoothness
$\mathcal{L}_{dnh}$	Reconstruction error margin	Guarded temporal residual

Training stages. CADENCE is trained in stages so that each module has a clear role. Table 11 summarizes the training sequence.

Table 11. CADENCE training stages. Later stages reuse the same shared visual encoder and preserve the previously learned whole-body prediction unless explicitly stated.

Stage	Trained modules	Supervision	Purpose
Stage 1	SMPL-X decoder and prediction heads	Whole-body SMPL-X, 3D joints, 2D keypoints, vertices, shape, expression, camera	Learn the base per-frame whole-body estimator
Stage 2	High-resolution hand pathway	Whole-body hand supervision and real hand-3D supervision from FreiHand/HanCo	Improve finger articulation from hand crops
Stage 3	Temporal refiner	Stored per-frame predictions and clip-level motion supervision	Offline temporal processing of the frozen per-frame estimate
Stage 4	LoRA adapters and task heads	Mixed whole-body and hand supervision	Lightweight final adaptation of the shared encoder and heads

Optimization and augmentation. Training uses AdamW with mixed-precision acceleration. The decoder and prediction heads use a higher learning rate than the LoRA parameters during the final adaptation stage. We apply standard image augmentations including random scale, translation, horizontal flip, color jitter, and synthetic occlusion where appropriate. For video clips, we sample fixed-length windows and apply temporal subsampling during training to improve robustness to different motion speeds.

Table 12. Optimization summary for the training configuration.

Setting	Value
Optimizer	AdamW
Precision	Mixed precision; bf16 where available
Decoder/head LR	2×10^{-4}
LoRA LR	1×10^{-4}
Training length	300k iterations
Temporal window	32 frames
Image augmentations	Scale, translation, flip, color jitter, occlusion
Temporal augmentations	Window sampling, temporal subsampling

A.6. Training Datasets

CADENCE is trained from a mixture of whole-body SMPL-X data, in-the-wild video data, 2D keypoint supervision, and real hand-3D data. Whole-body datasets provide the main SMPL-X supervision. COCO-WholeBody and MPII provide additional 2D keypoint supervision. FreiHand and HanCo provide real hand-3D supervision for the canonical hand training mode. The temporal refiner is trained on clips from available video sources using stored per-frame CADENCE predictions.

Table 13. Training data used by CADENCE. The table reports per-dataset image/frame counts, modality, and the role each source plays in supervision.

Dataset	Images / frames	Video	SMPL-X / 3D	2D keypoints	Usage
AGORA [28]	114,486	No	Yes	Optional	Whole-body SMPL-X supervision
BEDLAM [4]	3,500,000	Yes	Yes	Optional	Whole-body and video supervision
UBody [22]	542,000	Yes	Yes	Yes	Whole-body and expressive motion supervision
Motion-X [48]	2,931,318	Yes	Yes	Optional	Motion and temporal supervision
3DPW [38]	22,735	Yes	Body 3D	Optional	In-the-wild body/video supervision
COCO-WholeBody [14]	118,287	No	No	Yes	2D whole-body keypoint supervision
MPII [1]	28,880	No	No	Yes	Additional 2D keypoint supervision
FreiHand [51]	130,240	No	Hand 3D	Optional	Canonical hand training
HanCo [52]	860,304	No	Hand 3D	Optional	Canonical hand training

Sampling. During per-frame training, batches are sampled from a mixture of whole-body 3D datasets and 2D keypoint datasets. Loss masks ensure that each sample only contributes to losses for which annotations are available. During hand-pathway training, whole-body samples and canonical hand-only samples are mixed so that the hand branch learns both body-context hand refinement and detailed real hand articulation. During temporal training, the image encoder and per-frame decoder are frozen and the temporal refiner is trained from stored predictions, ensuring that controlled studies change only the temporal processing stage.

A.7. Evaluation Protocols

EHF is evaluated with Procrustes-aligned full-body mesh error and regional hand/face metrics. AGORA and UBody use whole-body mesh metrics with full-body and regional breakdowns. 3DPW uses the standard 14-joint body protocol, and EMDB uses the 24-joint body protocol. Acceleration error is computed from the difference of predicted and ground-truth central second differences over valid consecutive frames at the declared FPS.

CADENCE predicts whole-body SMPL-X, but 3DPW and EMDB evaluate only the body quantities specified by their protocols. These benchmarks therefore do not measure hand accuracy, absolute depth, or global placement.

B. Extended Benchmark Results

This section extends Tab. 3 with additional 3DPW methods and compares SAM 3D Body on the expressive benchmarks.

B.1. 3DPW Extended Results

3DPW is a body-centric benchmark. Although CADENCE predicts full SMPL-X parameters, evaluation follows the body protocol used by prior work. Table 14 groups methods by representation to make the comparison clearer: whole-body SMPL-X methods, body-focused per-frame or fitting methods, and body-only temporal methods. Acceleration errors are listed in Tab. 3.

Table 14. Extended 3DPW [38] body evaluation in millimetres. CADENCE predicts whole-body SMPL-X but is evaluated with the standard 3DPW body protocol. Dashes indicate values not reported by the corresponding method.

Method	Repr.	PA-MPJPE ↓	MPJPE ↓	PVE ↓
Whole-body SMPL-X methods				
ExPose [8]	SMPL-X	60.7	93.4	–
PyMAF-X [47]	SMPL-X	45.3	74.2	–
SMPLer-X-H [5]	SMPL-X	46.6	76.7	91.8
PEAR [42]	SMPL-X	45.3	71.7	–
Multi-HMR [3]	SMPL-X	41.7	61.4	–
DanceHMR [33]	SMPL-X video	35.2	**53.6**	65.1
CADENCE	SMPL-X video	40.2	58.1	**62.6**
Body-focused per-frame or fitting methods				
CameraHMR [27]	SMPL	35.1	56.0	65.9
PromptHMR [40]	SMPL-X	36.1	58.7	69.4
NLF-L+fit [32]	SMPL	<u>33.6</u>	54.9	63.7
SAM 3D Body [44]	Body	**33.2**	54.8	<u>63.6</u>
Body-only temporal methods				
TCMR [7]	SMPL video	55.8	95.0	–
MPS-Net [41]	SMPL video	52.1	84.3	–
WHAM [35]	SMPL video	35.9	57.8	68.7
TRAM [39]	SMPL video	35.6	59.3	69.6
GVHMR [34]	SMPL video	36.2	55.6	67.2
GENMO [20]	SMPL video	34.6	<u>53.9</u>	65.8

SAM 3D Body on EHF, AGORA, and UBody. Table 3 compares SAM 3D Body [44] with CADENCE on 3DPW and EMDB; here we extend the comparison to EHF, AGORA, and UBody. SAM 3D Body predicts a Momentum Human Rig mesh, a different topology from SMPL-X with no public vertex regressor to SMPL-X, so we compare body joints instead: SAM 3D Body's own shipped correspondence from its 70-keypoint output to the 13-joint LSP layout, paired with the same 13 SMPL-X joints via the standard SMPL-X joint regressor (7 upper-body joints only on UBody, whose close-up framing leaves hips, knees, and ankles out of frame in many clips). On this joint protocol, SAM 3D Body reaches 38.06/59.94/53.77 mm body PA-MPJPE on EHF/AGORA/UBody, next to CADENCE's own published number on each dataset's own protocol of 20.58/42.4/29.25 mm (EHF: 14-joint PA-MPJPE; AGORA/UBody: vertex PA-PVE). These are different metrics on different joint sets, shown for context, not as a matched comparison.

C. Hand Pathway Diagnostics

All variants in this section use the same per-frame whole-body estimator, shared encoder, and EHF protocol, and change only the hand pathway. Because the pathway writes only finger rotations, body metrics are identical across variants.

Table 15 compares design choices under an identical single-stage training budget, with the final pathway from Tab. 6 as reference. The feature source has the largest effect: sampling late-layer crop features gives 12.89 mm hand P-MVE and 8.72 mm hand P-MPJPE, against 14.93 and 10.07 mm with early-layer features only. The parameterization matters less: removing the zero-initialized residual or predicting the full hand pose directly changes either error by less than 0.5 mm.

C.1. Per-Finger Diagnostics on Sign Language

Table 16 reports per-finger wrist-aligned PA-MPJPE on an internal UBody sign-language test split with 18,130 samples, chosen because signing contains frequent and varied hand articulation. Errors are largest at the fingertips, especially of the

Table 15. Hand pathway design diagnostics on EHF (mm). Rows below the rule are trained under an identical single-stage budget and each changes one factor.

Variant	Hand P-MVE ↓	Hand P-MPJPE ↓
Final pathway (Tab. 6)	**12.2**	**8.3**
Late-layer crop features	12.89	8.72
Early-layer crop features only	14.93	10.07
No zero-initialized residual	12.45	8.47
Predict full hand pose directly	12.85	8.69

thumb and index finger, where opposition, self-occlusion, and fast motion make articulation hardest. CADENCE has lower error than SMPLer-X-H [5] and Multi-HMR-L [3] on every finger, with hand PA-MPVPE of 6.58 mm against 7.63 and 7.93 mm, and the fraction of joints within 5 mm (F@5 mm) is 7 to 9 points higher.

Table 16. Per-finger wrist-aligned PA-MPJPE (mm) on the UBody sign-language internal test split (18,130 samples). Values are averaged across left and right hands. Best results are **bold**.

	CADENCE				SMPLer-X-H				Multi-HMR-L			
Finger	MCP	PIP	Tip	Avg	MCP	PIP	Tip	Avg	MCP	PIP	Tip	Avg
Thumb	**4.45**	**3.76**	**8.21**	**5.47**	5.23	4.31	9.48	6.34	5.13	4.46	9.86	6.48
Index	**3.06**	**4.35**	**8.02**	**5.15**	3.77	4.88	9.40	6.02	3.52	5.38	10.97	6.62
Middle	**2.79**	**3.51**	**6.65**	**4.32**	3.37	3.94	7.54	4.95	3.34	4.23	7.58	5.05
Ring	**2.49**	**2.34**	**5.30**	**3.38**	2.96	2.68	6.26	3.96	2.95	2.94	6.30	4.06
Pinky	**2.79**	**3.27**	**6.13**	**4.06**	3.35	3.50	6.36	4.40	3.29	3.78	6.94	4.67
Avg	**3.12**	**3.45**	**6.86**	**4.48**	3.74	3.86	7.81	5.13	3.65	4.16	8.33	5.38
Hand PA-MPVPE	**6.58**				7.63				7.93			
F@5 mm (%↑)	**71.11**				64.16				61.88			
F@15 mm (%↑)	**96.85**				95.91				95.54			

D. Temporal Refiner Diagnostics

This section complements Sec. 4.2. All variants start from the same stored per-frame CADENCE predictions, evaluation clips, and metric implementation, and change only the temporal stage.

D.1. Single-Rate, Multi-Rate, and Component Ablations

Table 17 uses matched 32-frame clip subsets of 3DPW and SignAvatars/How2Sign, with identical stored inputs within each subset. The first block compares fixed filters with one shared bandwidth (single-rate) and with per-joint bandwidths (multi-rate). The second block removes one component at a time from an earlier version of the refiner, which added a residual rotation correction gated by a learned HOLD/STROKE phase on hand and jaw joints; the released refiner (Sec. 3.4) does not use these components. Clip pools and conditioning signals differ from Sec. 4.2, so absolute values are not comparable with Tabs. 1 and 2.

The single-rate and multi-rate filters have identical body accuracy, acceleration error, and jitter, but the multi-rate filter keeps more of the hand motion: Peak-R rises from 0.63 to 0.76 and Peak-A from 0.55 to 0.63. At equal smoothness, the per-joint bandwidth is therefore what preserves hand-speed peaks.

In the earlier refiner, the component rows change body accuracy and acceleration error by at most 0.06 mm and 0.03 mm/s^2. Removing the phase gate, the do-no-harm term, or the residual's conditioning leaves all metrics within 0.01, which is why the released refiner drops the residual and the gate and keeps the filter selection. Removing kinematic-tree attention raises Peak-A from 0.61 to 0.74 at higher jitter (0.69 against 0.64).

D.2. Low-Frame-Rate and Missing-Frame Robustness

Temporal filters can be sensitive to frame rate and missing observations. We therefore evaluate CADENCE temporal refinement under simple test-time degradations: temporal downsampling and random frame dropping. In each case, the same

Table 17. Diagnostic component ablations of the temporal refiner on matched 32-frame clip pools. Within each dataset, all rows use identical stored per-frame inputs and the same metric implementation. The 3DPW columns are evaluated on a matched 3DPW diagnostic subset, and the SignAvatars/How2Sign columns are evaluated on a matched sign-language diagnostic subset. These rows are not intended to replace the headline benchmark results in the main paper; absolute values differ slightly from the main tables because the clip pools and conditioning signals are restricted for controlled ablation.

Variant	3DPW diagnostic subset			SignAvatars/How2Sign diagnostic subset		
	PA-MPJPE ↓	MPJPE ↓	Accel ↓	Jitter ↓	Peak-R ↑	Peak-A ↑
CADENCE per-frame	42.07	61.74	16.96	3.47	0.93	1.15
Single-rate temporal filter	41.67	61.20	5.72	0.63	0.63	0.55
Multi-rate temporal filter	41.67	61.20	5.72	0.63	0.76	0.63
Earlier refiner, matched subset	41.71	61.20	5.72	0.64	0.75	0.61
w/o kinematic-tree attention	41.73	61.26	5.75	0.69	0.74	0.74
w/o HOLD/STROKE phase gate	41.68	61.19	5.72	0.64	0.76	0.61
w/o do-no-harm objective	41.70	61.19	5.72	0.64	0.76	0.61
unconditioned residual	41.66	61.18	5.72	0.64	0.75	0.58

per-frame model is used and only the temporal input sequence is degraded. For downsampling, we evaluate on the retained frames. For missing frames, dropped frames are masked during temporal refinement and metrics are computed on frames with available ground truth.

Table 18. Robustness of CADENCE temporal refinement to low frame rate and missing frames. Test-time degradations are applied to the temporal input sequence only. Body accuracy is reported on 3DPW, while expressive hand-motion preservation is reported on SignAvatars/How2Sign using Peak-A. Acceleration error increases under downsampling because retained frames are farther apart.

	3DPW body			SignAvatars/How2Sign
Setting	PA-MPJPE ↓	MPJPE ↓	Accel ↓	Peak-A ↑
Original frame rate	40.2	58.1	5.4	1.03
Half frame rate	42.1	59.5	12.3	1.10
One-third frame rate	45.0	61.9	21.6	1.06
Random drop 10%	40.4	58.2	7.1	1.06
Random drop 25%	40.8	58.6	10.0	1.17

The results show graceful degradation rather than collapse. Random frame dropping is handled robustly: PA-MPJPE rises from 40.2 to 40.8 mm when a quarter of frames are removed. Temporal downsampling is more demanding because the inter-frame interval grows, which both widens the motion gap the refiner must bridge and increases the measured finite-difference acceleration. Even so, accuracy degrades smoothly, with PA-MPJPE changing from 40.2 to 45.0 mm from full to one-third frame rate. Peak-A remains near or above 1.0, indicating that expressive hand-motion peaks are not collapsed under these degradations. Peak-A values above 1 indicate peak hand speeds above the reference.

D.3. Rotation Interpolation and Temporal Filtering Details

CADENCE applies temporal processing to articulated SMPL-X rotations rather than to unconstrained 3D joint coordinates. Rotations are represented using the continuous 6D representation during learning, converted to valid rotation matrices before forward kinematics, and re-orthonormalized after temporal updates. The deployed temporal model does not require linear interpolation of reduced-rate 6D rotations back to the full frame rate; the per-joint temporal filters operate over the full input window with different effective bandwidths for different joint groups.

For diagnostic variants that explicitly resample time, interpolation is applied on valid rotations rather than on arbitrary unnormalized 6D vectors. In practice, we convert 6D rotations to rotation matrices, interpolate rotations using a rotation-aware procedure, and then convert back to the representation used by the model. This avoids treating the 6D representation as an unconstrained Euclidean signal and prevents invalid rotations from entering the SMPL-X forward kinematics.

E. Runtime Accounting

Table 19 reports throughput for the main inference configurations and the dominant runtime components. All measurements are obtained on one A100-80GB GPU using bf16 inference. Batch-1 throughput measures low-latency inference, while batch-32 throughput measures batched throughput.

The base per-frame estimator processes the full-person crop and predicts whole-body SMPL-X parameters. Enabling the high-resolution hand pathway adds two hand-crop encoder passes, improving finger articulation but reducing throughput. The temporal refiner is inexpensive relative to image encoding, running at approximately 6,900 FPS for a 32-frame window.

Table 19. A100-80GB bf16 runtime and memory accounting. The visual encoder dominates runtime, the high-resolution hand pathway trades throughput for improved finger articulation, and the temporal refiner adds negligible overhead.

Configuration	Batch 1 FPS	Batch 32 FPS	Notes
Visual encoder only	62.8	367.0	Shared C-RADIOv4 encoder
Base per-frame estimator	27.9	307.6	Full-person crop; encoder + decoder + heads
Full CADENCE	13.1	71.5	Per-frame estimator + hand pathway + temporal refiner
Temporal refiner only	∼6,900	∼6,900	32-frame window; negligible overhead
Peak GPU memory	3.73 GiB		Batch 32, bf16

Interpretation. CADENCE is not bottlenecked by the temporal refiner. Runtime is dominated by visual encoding and, when enabled, the high-resolution hand pathway. This is expected: the hand pathway deliberately trades throughput for improved finger articulation by reusing the same high-capacity encoder on hand crops. Applications that require maximum throughput can run the per-frame estimator with the temporal refiner and skip the hand pathway.

F. Additional Qualitative Results

This section describes layer-by-layer decoder refinement, a comparison with whole-body baselines, temporal examples, and failure cases. The corresponding figures are collected in Sec. G.

F.1. Whole-Body Qualitative Results

Figure 4 shows additional CADENCE reconstructions on diverse 3DPW sequences, visualized as a layer-by-layer refinement progression. The structured decoder starts from a coarse initial estimate and progressively updates the mesh and joint locations across its six layers using iterative spatial feedback. The examples include walking, running, stair climbing, and seated poses with partial occlusion. They illustrate how the decoder refines global body structure and projected joint alignment over successive layers.

F.2. Multi-Baseline Qualitative Comparison on Expressive Scenes

Figure 5 compares CADENCE qualitatively with recent whole-body or whole-human recovery methods on expressive UBody scenes. We compare against SMPLer-X [5], Multi-HMR [3], and SAM 3D Body [44]. To make the visual comparison controlled, all methods are run in single-person mode using the same detected person box where possible. Each panel shows the full-person overlay and a zoom strip containing the face and both hands cropped from the same image regions.

The methods use different representations and processing assumptions in expressive regions such as hands and face. SAM 3D Body uses a different mesh representation from SMPL-X. We therefore show the panels only as qualitative context and avoid topology-specific or hand-accuracy claims. CADENCE's high-resolution branch is finger-only and retains a whole-body SMPL-X output.

F.3. Temporal Consistency Examples

Figure 6 compares per-frame predictions, a generic temporal smoother, and the deployed CADENCE temporal refiner on a sign-language clip with rapid finger articulation. Per-frame predictions exhibit large frame-to-frame acceleration. To isolate *what* is being smoothed rather than only *how much*, we tune the generic Savitzky–Golay smoother so that its residual wrist jitter matches the CADENCE refiner's (2.5 vs. 2.6 mm/frame^2). At this matched stability level, the generic smoother attenuates the hand-motion peaks, whereas CADENCE preserves them more closely: on this clip, CADENCE retains 99% of the per-frame peak finger speed compared with 46% for the matched generic smoother.

F.4. Consecutive-Frame Visualizations Across Diverse Activities

Figure 7 shows CADENCE reconstructions on clips spanning different activity types, including dynamic athletic motion, classroom interaction, and expressive sign-language sequences. For each clip, consecutive frames are shown from left to right with the recovered SMPL-X overlay. These examples show how the recovered body and hand configurations evolve across neighbouring frames in diverse scenes.

Videos. Video results, including continuous two-handed sign-language sequences, are on the project page, https://keniel123.github.io/cadence.

F.5. Failure Cases

Table 20 summarizes the main failure modes, their likely causes, and possible future directions. These examples are consistent with the limitations in Sec. 4.9. The most common failure modes arise from monocular ambiguity, severe occlusion, unusual camera viewpoints, motion blur, inaccurate person boxes, and difficult hand-object or hand-hand interactions.

Table 20. Common CADENCE failure modes, likely causes, and possible future directions.

Failure mode	Typical cause	Possible direction
Incorrect global orientation	Monocular camera ambiguity and weak-perspective assumptions	Full-perspective camera modelling and stronger visual-temporal cues
Wrong limb depth	Foreshortening, self-occlusion, side views, or ambiguous self-contact	Depth-aware conditioning and multi-frame geometric constraints
Unstable hands	Small hand image regions, motion blur, hand-object interaction, or hand-hand occlusion	Higher-resolution evidence and hand-object-aware supervision
Hand or face temporal error	Rapid motion, small image regions, or occlusion	Motion-aware hand and face modelling
Incorrect person scale or placement	Person-box errors, crop/camera mismatch, or weak-perspective scale ambiguity	Better box tracking and camera-aware crop normalization
Multi-person confusion	The current system assumes a single tracked person crop	Multi-person tracking and identity-aware temporal modelling

G. Qualitative Figures

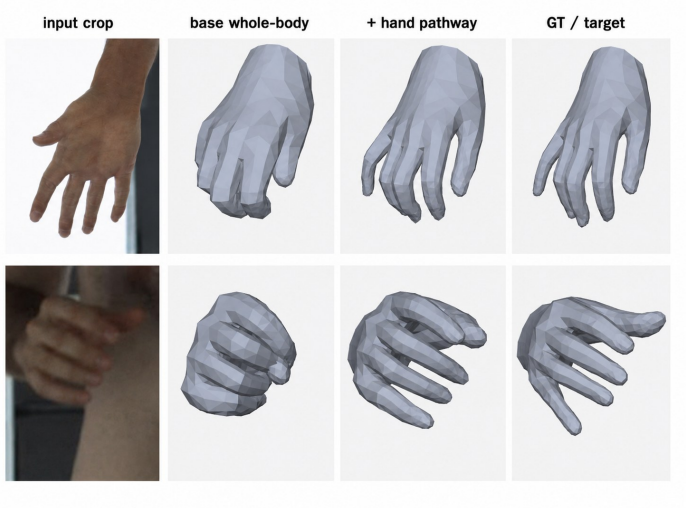


Figure 3. Effect of the high-resolution hand pathway. Each row shows the input hand crop, the base whole-body hand prediction, the prediction after the hand pathway, and the target. The base estimate captures coarse hand placement but can miss finger articulation. The hand pathway reuses the shared visual encoder on hand crops and predicts finger pose while retaining the remaining whole-body parameters.

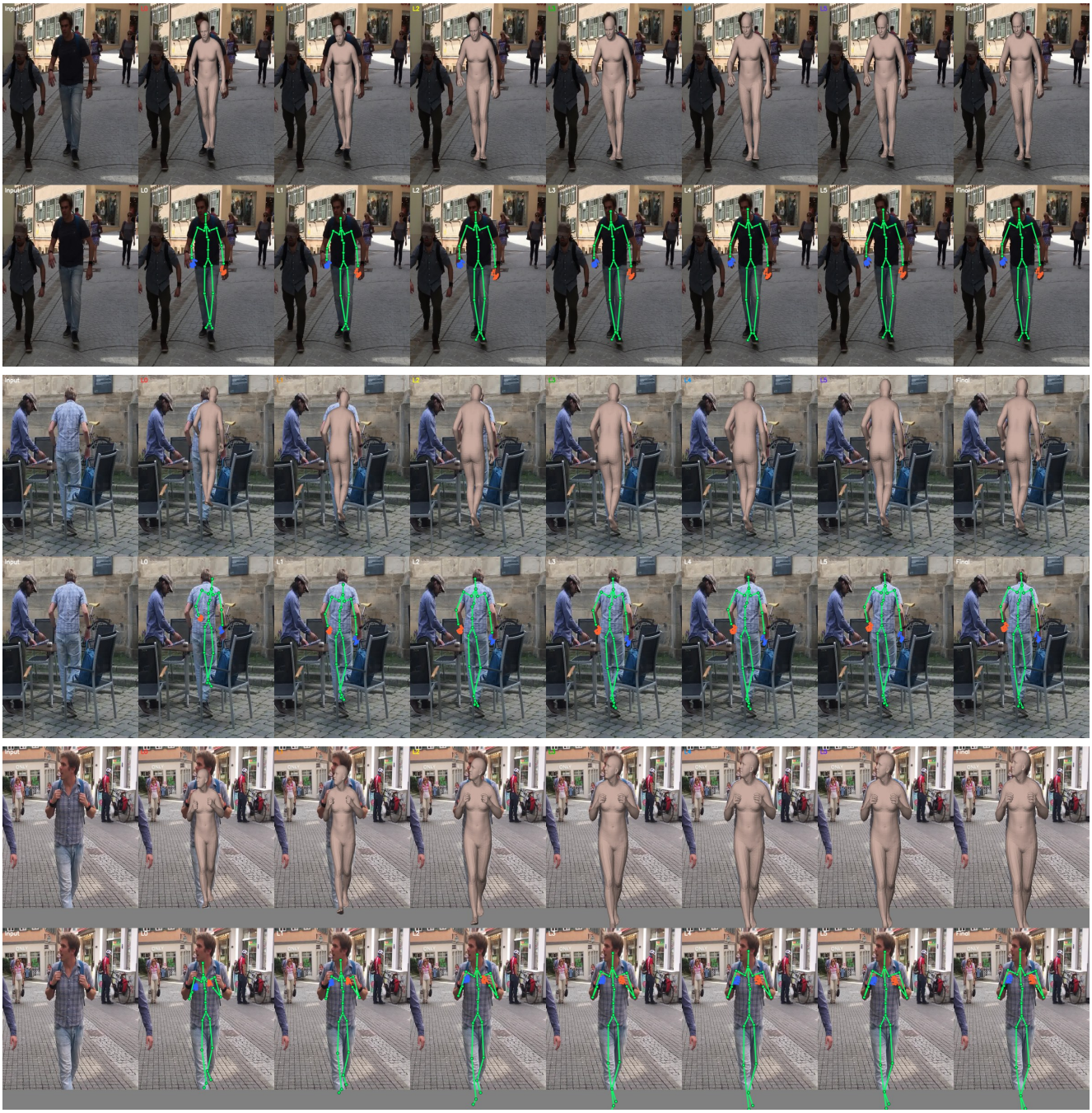


Figure 4. Whole-body qualitative results with layer-by-layer refinement on 3DPW [38] test sequences. For each example, the top row shows the SMPL-X mesh overlay across decoder layers (Input→L0→L1→L2→L3→L4→L5→Final), and the bottom row shows the corresponding skeleton keypoints. The iterative decoder progressively improves mesh alignment and joint localization across layers. Please zoom in for details.

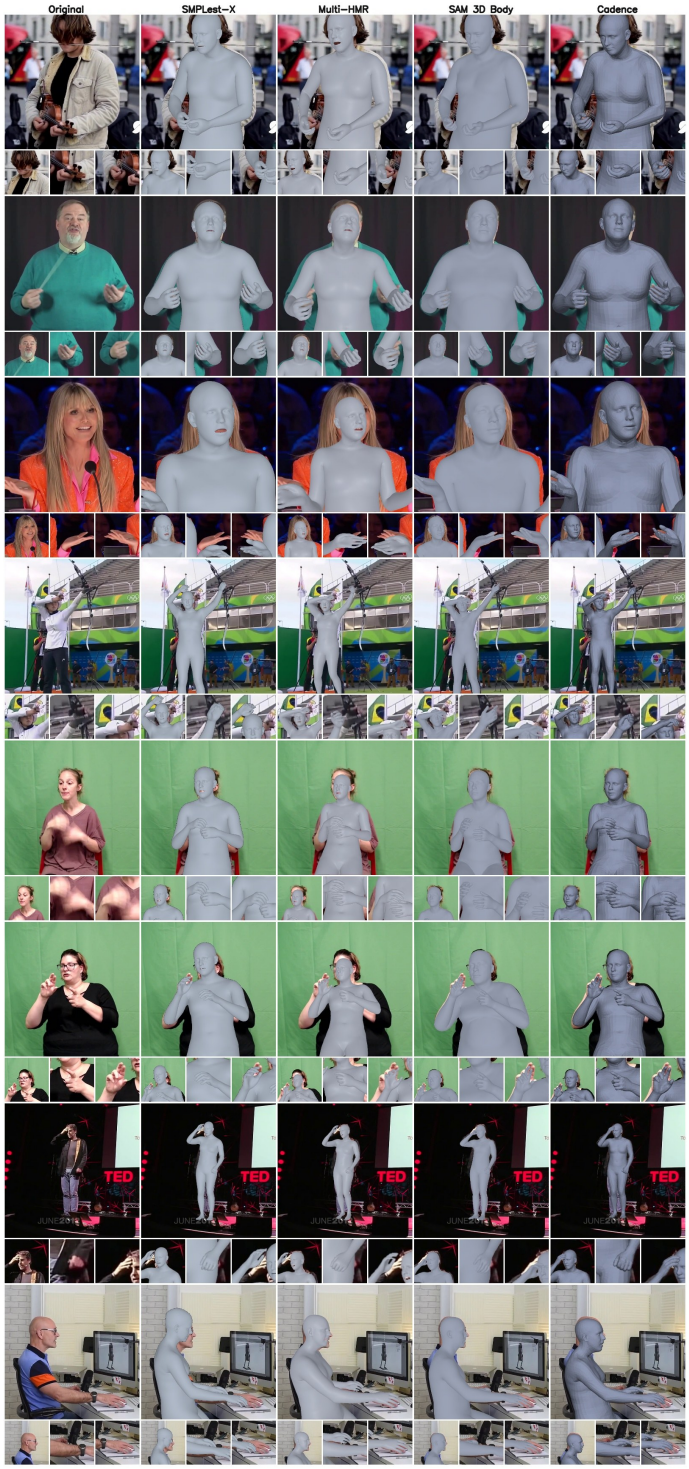


Figure 5. Qualitative comparison on expressive UBody [22] scenes. Columns show the input frame, SMPLer-X [5], Multi-HMR [3], SAM 3D Body [44], and CADENCE with the hand pathway. Each method is shown with a full-person overlay and a zoom strip for the face and both hands using shared crop regions. This figure is qualitative context rather than a hand-accuracy comparison. SAM 3D Body is shown with a non-SMPL-X topology. Please zoom in for details.

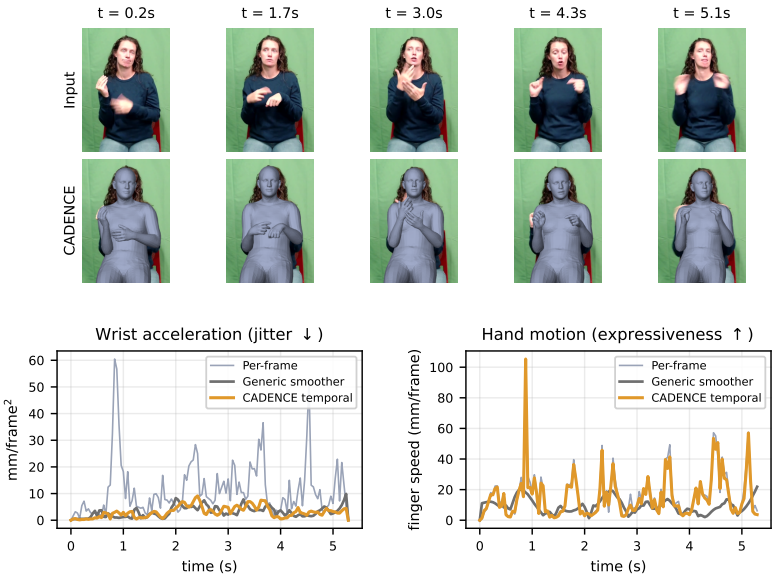


Figure 6. Temporal qualitative comparison on a sign-language clip with rapid finger articulation. **Top:** input frames and the corresponding CADENCE whole-body reconstruction over time; the per-frame CADENCE estimate tracks the signer's pose and hands closely. **Bottom left:** wrist acceleration, used as a jitter proxy; per-frame predictions spike sharply, while the generic smoother (Savitzky–Golay) and the CADENCE temporal refiner reduce jitter to a matched level (2.5 vs. 2.6 mm/frame^2). **Bottom right:** finger speed over time; at equal jitter the generic smoother attenuates hand-motion peaks, whereas CADENCE preserves them more closely, retaining 99% of the per-frame peak finger speed versus 46% for the matched generic smoother. CADENCE therefore improves temporal stability while better preserving expressive hand dynamics.

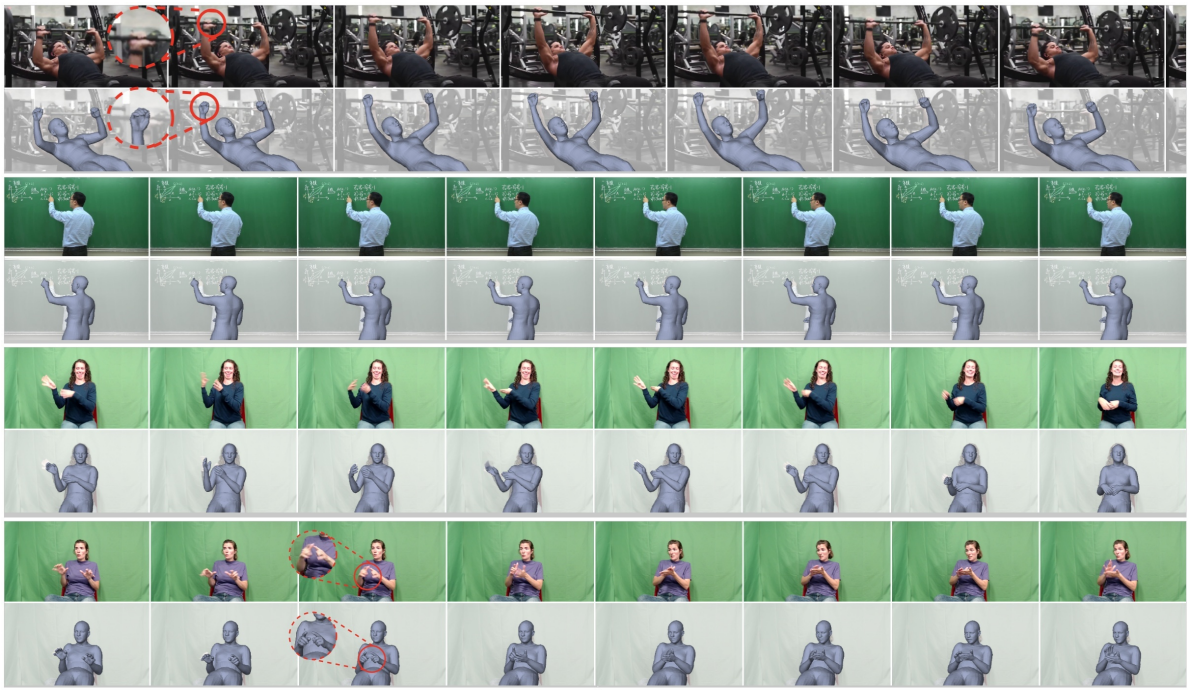


Figure 7. Consecutive-frame CADENCE visualizations across diverse activities. Each clip is shown as a pair of rows: input video frames and CADENCE SMPL-X overlays, sampled at consecutive timesteps from left to right. The examples include dynamic body motion, classroom interaction, and expressive sign-language sequences. Highlighted insets mark hand regions across neighbouring frames. Please zoom in for details.