DENSE4D: PIXEL-LEVEL 4D HUMAN CORRESPONDENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Reconstructing human motion in 4D requires identifying the same surface points through articulation, deformation and changes in visibility. We introduce Dense4D, to our knowledge the first feed-forward model dedicated to dense 4D correspondence across the body, clothing and hair from RGB images, without a body template. Trained specifically for human geometry, Dense4D jointly processes person-centred crops with known image-to-crop transforms to predict Human Point Maps (HPMs) and depth for each image. An HPM assigns each person pixel the 3D position of its observed surface point at a common reference time. These maps support amodal correspondence: points hidden in the reference image can be localised from observations in other frames. Matching HPM coordinates across crops establishes correspondence without predicting camera poses or large scene displacements; per-image depth then lifts the matches into 3D tracks. Experiments on real people demonstrate substantially higher tracking accuracy and closer agreement with scanned surfaces than the evaluated baselines. Points from different frames assemble into coherent human geometry in the reference pose, providing visible evidence of consistent dense correspondence through articulation and clothing deformation. Videos of dense 4D correspondence and tracking: https://anonymous.4open.science/w/research-materials-F6D0/.

1 INTRODUCTION

A video reveals different parts of a person as they move and turn. Reconstructing geometry frame by frame does not establish which observations belong to the same surface point. A patch of clothing may deform, disappear and reappear elsewhere. We study *dense 4D human correspondence*: matching points across the body, clothing and hair in different images and recovering their three-dimensional trajectories.

Human correspondence methods commonly map image pixels to a shared body template (Güler et al., 2018). Such coordinates describe the underlying body, but do not directly represent the geometry and independent motion of loose garments and hair. In parallel, many 4D dynamic scene reconstruction systems jointly estimate geometry, motion and cameras across diverse scenes (Sucar et al., 2026; Zhang et al., 2026). We introduce Dense4D for dense 4D human correspondence without a body template.

Dense4D follows a top-down approach: we isolate each person in the scene and jointly process their crops across time (Fig. 2). Using the focal length and person mask, a known virtual-camera warp reorients each crop towards the person, and the crop zoom is supplied to the network. A Human Point Map (HPM) assigns each person pixel the 3D position of its observed surface point at a selected reference time, expressed in the reference crop's camera coordinates. These maps support *amodal correspondence* to the reference surface, including regions hidden in the reference image but observed in other frames. Their union samples this reference-time geometry (Fig. 1, right).

By expressing correspondence in person-centred crop coordinates, Dense4D separates surface matching from the recovery of camera motion and the person's displacement through the scene. This simplifies the prediction task: each observed point is located on the person's reference geometry, with crop transforms and zoom already known. Where recent dynamic scene reconstruction methods such as V-DPM predict point motion in a shared camera frame (Sucar et al., 2026),

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: **Dense 4D correspondence of the body, clothing and hair.** Top left: input frames show the person at different times, with each coloured box identifying a source frame. Below each frame, its HPM maps the visible person pixels to 3D points in the reference pose (0.0 s), using the same source-frame colour. Right: combining these maps produces the HPM union, shown from three viewpoints with source-frame colours (top), a 3D chequerboard (middle) and input RGB colours (bottom). Although the person moves, the mapped observations align into a coherent human in the reference pose. The chequerboard is transferred from reference space to the input images (third row, left), showing dense correspondences. Matching HPM coordinates and lifting the matched pixels with each frame's predicted depth gives the 3D trajectories shown at bottom left.

Dense4D obtains each track position by matching HPM coordinates and lifting the matched pixel with its image's predicted depth. Scene camera poses enter only for subsequent world placement.

We use the image encoder and aggregator architecture of Wang et al. (2026). An additional crop token conditions the aggregator, which alternates attention within and across frames. Two dense heads predict HPMs and per-image depth in a single forward pass. Training uses human geometry and correspondence supervision from synthetic motion and multi-view renderings of real scans.

Across articulated motion and clothing deformation, Dense4D brings observations from different frames into coherent reference-time human geometry, with visibly aligned body parts and fine surface details (Sec. 4.4). On real footage, Dense4D recovers more accurate human trajectories and better matches the scanned reference geometry than the methods we compare against (Secs. 4.2 and 4.3). The shared reference representation also supports tracking across long intervals, as demonstrated on an 86 s video (Fig. 12).

Our contributions are:

- We introduce Dense4D, to our knowledge the first feed-forward model dedicated to template-free dense 4D human correspondence. Its predictions assemble observations of the body, clothing and hair into a coherent reference-time geometry, including regions hidden in the reference image.
- We decouple correspondence estimation from camera-pose and scene-motion recovery. Dense4D supports direct matching across long temporal intervals and integrates with scene reconstruction through shared image pixels.
- We introduce an amodal 3D localisation protocol on 4D-DRESS, measuring how closely observations across time align with the full reference-time scan. We also evaluate 4D point tracking on the human queries of the existing TAPVid-3D Panoptic Studio benchmark.

2 RELATED WORK

Human correspondence and parametric reconstruction. A shared body template allows pixels on different people to be related through common anatomical coordinates (Güler et al., 2018). BodyMap and TemporalUV extend this association to clothing and hair (Ianina et al., 2022; Xie et al., 2022), while HumanGPS learns scan-supervised embeddings shared across subjects (Tan et al., 2021). These coordinates and embeddings establish correspondence but do not themselves reconstruct the individual's body, garment or hair geometry. Mesh reconstruction methods such as 4DHumans (Goel et al., 2023) achieve body correspondence thanks to the fixed topology of the underlying parametric model, but do not represent the geometry and independent motion of garments or hair. Our goal is *personalised correspondence*: HPMs reconstruct points on the person's reference-time surface and locate observations of that same person on it. The correspondence space follows the individual's geometry without imposing a shared body template.

Deformable human reconstruction. ReLoo and DressRecon fit personalised surfaces, appearance and body/clothing deformation models to video through differentiable rendering (Guo et al., 2024; Tan et al., 2025). Their temporal correspondences follow the fitted 3D deformation fields used to render the reconstructed motion. We learn pixel-to-reference-surface correspondence directly from dense 3D supervision, together with per-image depth. This formulation handles the body, clothing and hair together, without decomposing their motion into body and garment deformation models.

Point tracking and amodal correspondence. Point trackers estimate trajectories for image queries, from the long-range formulation of Particle Video (Sand & Teller, 2006) to learned tracking through occlusion (Harley et al., 2022; Doersch et al., 2023; Karaev et al., 2025). Three-dimensional trackers combine point correspondence with depth or geometric estimation (Xiao et al., 2024; 2025; Ngo et al., 2025; Zhang et al., 2025). HPMs also reconstruct points on the reference surface: a point observed in another frame can be located in 3D at the reference time even when hidden in the reference image.

Scene reconstruction. Point-map models represent pixels from different images as 3D points in a shared reference camera frame (Wang et al., 2024a; Leroy et al., 2024; Wang et al., 2025a). DualPM pairs posed and canonical rest-pose maps for quadrupeds (Kaye et al., 2025). Face Anything predicts depth and canonical facial coordinates, using nearest-neighbour matching in a template-aligned canonical space for correspondence (Kocasari et al., 2026). HPMs instead use the individual's full human surface at a selected input time as the reference.

Feed-forward models extend point-based reconstruction to dynamic scenes (Feng et al., 2025; Zhang et al., 2026), often jointly training geometry, motion and camera predictors (Karhade et al., 2025; Luo et al., 2026; Sucar et al., 2026). Dynamic Point Maps and V-DPM are closest to our representation: they express observations at a shared reference time and viewpoint (Sucar et al., 2025; 2026). Dense4D adopts this principle within person-centred crops with known transforms. The model learns to locate points on the person's reference geometry, separating the correspondence task from camera motion and potentially large displacements through the scene.

3 METHOD

3.1 DENSE 4D CORRESPONDENCE WITH HUMAN POINT MAPS

Given images $\{\mathbf{I}_t\}_{t=1}^{S}$ of one person, person masks and focal lengths f_t, we predict the geometry and correspondence of the person's body, clothing and hair across time. Using each mask and focal length, we resample the person into a virtual camera aimed at the centre of the person's bounding box, keeping the camera centre fixed. At inference, focal length is Dense4D's only supplied camera parameter; the model neither receives nor predicts world-camera poses. Below, pixel coordinates and foreground sets $\mathcal{M}_t$ refer to these crops. Ground-truth masks define the training targets; at inference, the masks specify the person to be processed.

The first input is the reference, $r = 1$; input order need not be chronological. For $p \in \mathcal{M}_t$, the HPM target $\mathbf{X}_t(p) \in \mathbb{R}^3$ is the position at the reference time of the surface point observed at p in frame

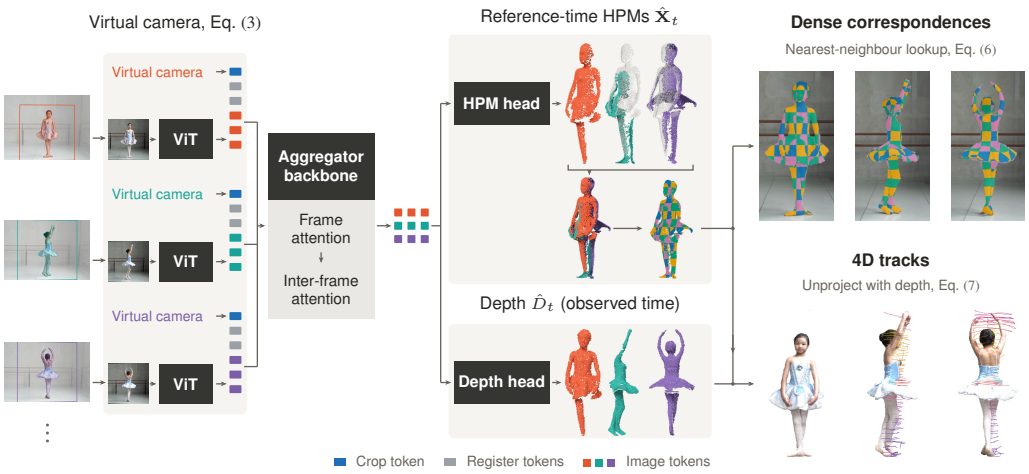

Figure 2: **Dense 4D human correspondence.** An extra crop token conditions the aggregator on cropping geometry (Eq. (3)). The heads predict reference-time HPMs and per-image depth, visualised as HPM point clouds and unprojected depth maps. Nearest-neighbour lookup matches HPM coordinates within the target frame (Eq. (6)). A 3D chequerboard defined in reference space colours the union and is transferred to the images through HPM coordinates, visualising dense correspondence. Unprojecting the matched pixels with predicted depth gives 4D tracks (Eq. (7) and Sec. 3.4), shown as coloured trajectories.

t. This position is expressed in the reference image's virtual camera coordinates. The scalar depth target $D_t(p)$ is the same point's optical-axis depth in frame t's virtual camera at time t. Both targets use a shared normalising scale s.

To define these targets formally, let $\mathbf{x}_{t\to u}(p)$ denote the world-space position at time u of the surface point observed at pixel p in image t. For supervision, the known virtual-camera extrinsics $\mathbf{E}_{v,t} \in \mathbb{R}^{3\times 4}$ map homogeneous world points into the crop's camera coordinates. The HPM and depth targets are therefore

$$\mathbf{X}_t(p) = \frac{1}{s}\mathbf{E}_{v,r}\begin{bmatrix}\mathbf{x}_{t\to r}(p)\\ 1\end{bmatrix}, \qquad D_t(p) = \frac{1}{s}\begin{bmatrix}0 & 0 & 1\end{bmatrix}\mathbf{E}_{v,t}\begin{bmatrix}\mathbf{x}_{t\to t}(p)\\ 1\end{bmatrix}. \tag{1}$$

The row vector $[0\ 0\ 1]$ selects the optical-axis coordinate of the point in the virtual camera. For the training targets, s is the mean distance of the reference-time surface samples from the reference virtual camera centre:

$$s = \frac{1}{\sum_t |\mathcal{M}_t|}\sum_t \sum_{p\in\mathcal{M}_t}\left\|\mathbf{E}_{v,r}\begin{bmatrix}\mathbf{x}_{t\to r}(p)\\ 1\end{bmatrix}\right\|_2. \tag{2}$$

All HPMs share the reference virtual camera's origin and axes, and both outputs use the same normalised scale. The network predicts in this normalised space; absolute body size and camera distance are not recovered.

Observations of the same surface point have equal HPM targets, providing a shared coordinate for correspondence across time. The union of these targets samples the person's surface at the reference time, bringing together the body, clothing and hair seen across the input clip. It includes regions hidden in the reference image but observed elsewhere, without fitting a body or surface template. The network learns to reproduce this common reference surface.

3.2 Crop-aware prediction

We construct the image intrinsics $\mathbf{K}_t$ from f_t, assuming square pixels, zero skew and an image-centred principal point. The virtual crop is square with side $\ell = 512$, intrinsics $\mathbf{K}_{v,t}$ and a crop-centred principal point. The rotation $\mathbf{R}_{v,t}$ maps the ray through the person's box centre from the

original camera coordinates to the virtual camera's optical axis. With a fixed camera centre, the original and virtual image planes are related by a rotation-induced planar homography independent of scene depth (Yu et al., 2021a):

$$\mathbf{H}_t = \mathbf{K}_{v,t}\mathbf{R}_{v,t}\mathbf{K}_t^{-1}, \qquad \mathbf{c}_t = \left[(\mathbf{R}_{v,t})_{3,:}, \log(f_{v,t}/\ell)\right] \in \mathbb{R}^4. \tag{3}$$

Here $f_{v,t} = f_t\ell/\max(c_w, c_h)$, where c_w and c_h are the source-image box width and height. The first three entries encode the unit box-centre ray in the original camera coordinates; the fourth is the logarithm of the normalised virtual focal length. Masks define the crops and restrict matching to person pixels. The network receives the RGB crops and an embedding of these four parameters. CLIFF and CameraHMR (Li et al., 2022; Patel & Black, 2025) also supply focal-normalised crop position and size for human mesh recovery. Appendix D.1 analyses the respective roles of crop zoom and the centre ray.

The *crop embedding* supplies this geometry to the network: an MLP maps $\mathbf{c}_t$ to one additional token per image (Fig. 2). This token joins the patch tokens from a ViT image encoder before the first attention block. The aggregator alternates attention within and across frames to process all observations jointly. Two dense heads read the aggregator features: the HPM head predicts reference-time surface coordinates $\hat{\mathbf{X}}_t$, and the depth head predicts depth $\hat{D}_t$ for each input image. Each head also predicts a confidence map.

We initialise the ViT image encoder and aggregator with pretrained VGGT-Ω weights (Wang et al., 2026). The HPM decoder uses weights from VGGT-Ω's dense prediction head. Its final layer is initialised from scratch to predict 3D coordinates. The depth head is trained from scratch, and the crop embedding MLP's final layer is initialised to zero. The model has approximately 974 M parameters; further architecture and crop-implementation details are given in Appendix B.

3.3 Surface supervision and training

We train with animated surfaces from BEDLAM2.0 (BEDLAM2; Tesch et al., 2025), multi-view renderings of static THuman2.1 scans (Yu et al., 2021b), and MammaSyn body-surface tracks (Cuevas-Velasquez et al., 2026) (Appendix B.3). Figure 8 in the supplementary material illustrates the supervision surfaces and input crops. All three sources provide meshes with consistent vertex and face indices across the input images. For BEDLAM2, these include body and garment meshes, together with envelope meshes derived from hair grooms and transported with the posed head. We store each supervised pixel's rasterised triangle index and barycentric coordinates, as also done in Syn4D (Jiang et al., 2026). Applying the corresponding perspective-correct weights $\omega_{t,p,m}$ to the reference vertices gives the reference-time position:

$$\mathbf{x}_{t\to r}(p) = \sum_{m\in F[\varphi_t(p)]} \omega_{t,p,m}\mathbf{V}_r[m]. \tag{4}$$

Here $\varphi_t(p)$ is the triangle index, $F[\varphi_t(p)]$ lists its three vertex indices, and $\mathbf{V}_r[m]$ is vertex m in world coordinates at the reference time. The same barycentric interpolation (Laine et al., 2020; Cole et al., 2021), evaluated on frame t's vertices, gives $\mathbf{x}_{t\to t}(p)$ for the depth target in Eq. (1). Dataset preparation is detailed in Appendix A.

For supervised pixels $i = (t, p)$ with $p \in \mathcal{M}_t$, we minimise the point loss $\mathcal{L}_{\mathrm{pt}}$ and depth loss $\mathcal{L}_{\mathrm{dp}}$ with equal weight. For each head $k \in \{\mathrm{pt}, \mathrm{dp}\}$, let σ_i^k be its predicted confidence:

$$\begin{aligned} \mathcal{L}_k &= \mathcal{Q}_{.98}\big(\{r_i^k\sigma_i^k - 0.2\log\sigma_i^k\}_i\big) + \mathcal{Q}_{.98}\big(\{r_i^k\}_i\big)\,, \\ r_i^{\mathrm{pt}} &= \|\hat{\mathbf{X}}_i - \mathbf{X}_i\|_2, \qquad r_i^{\mathrm{dp}} = |\hat{D}_i - D_i|, \qquad \sigma_i^k \geq 1. \end{aligned} \tag{5}$$

This confidence-weighted objective follows DUSt3R and VGGT (Wang et al., 2024a; 2025a). The operator $\mathcal{Q}_{.98}$ trims the upper tail by averaging only values below the 98th percentile, independently for the confidence-weighted and unweighted terms. During training, frames are sampled in random order so that any labelled observation can serve as the reference. Implementation and training details are given in Appendix B.3.

3.4 Dense 4D tracking and reconstruction

A single network forward pass predicts HPMs and depth jointly for all input frames. Each HPM sample retains its source frame and pixel index when the maps are combined into a union (Fig. 2).

For a query pixel p in frame t, we search only the HPM samples belonging to the person mask in target frame u:

$$\hat{q}(p; t \to u) = \arg\min_{q \in \mathcal{M}_u} \|\hat{\mathbf{X}}_u(q) - \hat{\mathbf{X}}_t(p)\|_2. \tag{6}$$

We estimate a match as visible when the residual at $\hat{q}$ is below $\tau = 0.005$. This allows a small tolerance between predictions of the same surface point in normalised reference coordinates. An estimate of the metric scale could instead support a tolerance specified in centimetres; we use a fixed normalised threshold. The predicted depth lifts the matched pixel into the target frame's original camera coordinates, using the crop intrinsics and rotation:

$$\mathbf{P}_u(q) = \mathbf{R}_{v,u}^{\top} \hat{D}_u(q) \mathbf{K}_{v,u}^{-1} \tilde{q}, \qquad \hat{\mathbf{T}}_{t,p}(u) = \mathbf{P}_u(\hat{q}(p; t \to u)) \,. \tag{7}$$

Here $\tilde{q} = [q_x, q_y, 1]^{\top}$, with q_x, q_y denoting pixel-centre coordinates. Repeating the lookup gives the query point's trajectory and visibility estimates in per-frame camera coordinates, with a shared normalised scale. Each 3D position uses the depth predicted at the matched target pixel, so tracks and per-frame depth are consistent by construction.

The known crop warps and person masks align our predictions with the image pixels used by a scene reconstruction model. Scene depth within the person masks sets the person-to-scene scale; scene camera poses then place our human geometry and tracks in the reconstructed world (Appendix B.2). Our predicted depth and HPMs supply the human geometry and correspondences. For internet video, we use VGGT-Ω (Wang et al., 2026) for scene reconstruction and SAM 3 (Carion et al., 2025) for person masks. Benchmark-specific inter-person scaling is described in Appendix C.1.

4 Experiments

We evaluate the 4D representation through point tracking on real moving people and an amodal 3D localisation protocol on 4D-DRESS. Crop-conditioning ablations and temporal analysis are reported in Appendix D, with additional qualitative results in Appendix E.

4.1 Evaluation setup

We train Dense4D cumulatively using all 27,727 BEDLAM2 (Tesch et al., 2025) training tracks, 2,200 of the 2,445 THuman2.1 (Yu et al., 2021b) scans and 15,710 MammaSyn (Cuevas-Velasquez et al., 2026) person tracks. Every tenth THuman2.1 scan is held out; the training budget and sampling schedule are given in Appendix B.3. We use BEDLAM2's official test split for in-domain ablations, while our main evaluations assess transfer to real footage.

4.2 4D point tracking on Panoptic Studio Humans

Protocol. We evaluate the human-query subset of TAPVid-3D Panoptic Studio (Koppula et al., 2024; Joo et al., 2019), which we refer to as *Panoptic Studio Humans*. We retain all 156 clips of 150 frames, evaluating the same set of queries on people across methods. We report average Jaccard (AJ), average percentage of points within the depth-adaptive thresholds (APD), and occlusion accuracy (OA), using the official evaluator. VGGT-Ω depth sets the relative scale of people; our model supplies their geometry and correspondences. Further protocol details and baseline configurations are given in Appendices C.1 and C.2, respectively.

Results. Our model obtains 38.4 AJ, 50.3 APD and 89.6 OA (Tab. 1). The strongest evaluated baselines are TAPIP3D (17.0 AJ, 27.1 APD) and V-DPM (16.9 AJ, 27.6 APD). Our model improves on the best evaluated baseline for AJ by 21.4 points and for APD by 22.7 points. Dense4D's advantage over the compared systems spans both point localisation and occlusion prediction. Figure 3 shows our tracks and reference-time HPM unions, with further examples in Fig. 15.

4.3 Amodal 3D localisation on 4D-DRESS

Localising observations at a common reference time requires predictions for regions that may be hidden in the reference image. We introduce an *amodal 3D localisation* evaluation on 4D-DRESS (Wang et al., 2024b), whose synchronised images and scans let us test whether points observed across time agree with the person's full reference-time geometry. Because the scan covers

Method	AJ↑	APD↑	OA↑
Ours	**38.4**	**50.3**	**89.6**
TAPIP3D	17.0	27.1	85.2
V-DPM	16.9	27.6	76.7
SpatialTrackerV2	15.5	24.4	78.7
4RC	14.4	24.1	76.6
Shape of Motion	12.5	21.3	80.7
Any4D	7.3	13.6	76.6

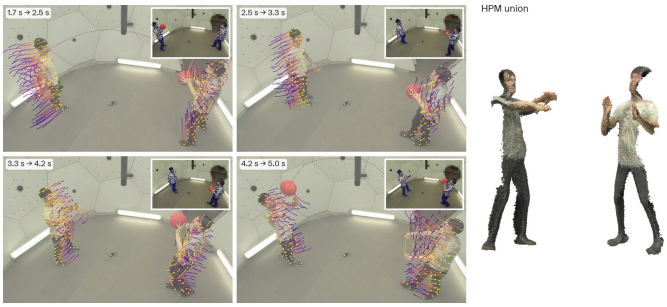

Table 1 and Figure 3: **Panoptic Studio Humans.** Left: tracking accuracy on the subset of benchmark points on human bodies. Centre: our tracks from 1.7 to 5.0 s. Far right: HPM unions combining person pixels from across the input sequence in each person's reference pose. Despite the motion across frames, these points assemble into coherent bodies with aligned limbs and surface detail, making the quality of temporal alignment directly visible.

regions hidden in the reference image, observations revealed later in the video can also be evaluated in the reference pose. Small distances indicate close agreement with measured geometry, providing a stricter test than coarse localisation near the body.

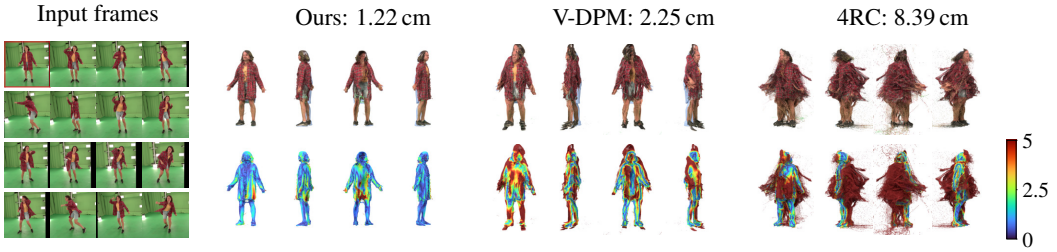

Figure 4: **Reference-time reconstruction from sixteen observations on 4D-DRESS.** Left: input frames, with the reference boxed. For each method, four viewpoints show the union coloured by source appearance (top) and distance to the scan from 0 to 5 cm (bottom). The measured scan is overlaid in blue-grey in the top row.

We evaluate 128 distinct dynamic takes, with one sixteen-frame window per take beginning at its first frame. All methods receive the same person crops from one camera, with the first frame as the reference. HPMs directly predict the reference-time positions of all foreground pixels; V-DPM (Sucar et al., 2026) and 4RC (Luo et al., 2026) decode each frame at that same reference time. We combine the sixteen point maps into one cloud and report the mean one-sided point-to-scan distance, using the full reference scan. The distance is one-sided because the input views need not cover the full reference scan. Each method discards its own least confident 5 % of points across the whole union; details are in Appendix C.3. This measures geometric agreement across time and provides indirect evidence of correspondence quality; it does not test whether each prediction reaches the correct surface point (Sec. 5).

Results. Our union lies less than 1 cm from the measured surface on average: 0.86 cm, compared with 2.31 cm for V-DPM and 4.84 cm for 4RC (Tab. 2). These sequences involve little translation of the person in world coordinates, yet Dense4D is substantially more accurate than both dynamic scene models. This gap highlights the challenge of fine surface localisation for models that predict human motion in scene coordinates. Points from non-reference frames are nearly as accurate as those from the reference frame (0.87 versus 0.80 cm), and only 6.8 % lie more than 2 cm from the surface, compared with 41.5 % and 53.4 % for the baselines. For both baselines, reference-frame predictions have lower error than those from other frames: 1.64 versus 2.36 cm for V-DPM and 4.07 versus 4.89 cm for 4RC.

Figure 5 shows how error varies with time and confidence. Our error stays within approximately 0.1 cm of the reference-frame value across the plotted temporal range. V-DPM's error rises by about

0.8 cm within the first second and then levels off, while 4RC's grows gradually, by about 1.3 cm over the first two seconds. Increasing the confidence cut from 0 to 50 % reduces 4RC's error from 5.86 to 2.78 cm and V-DPM's from 2.70 to 1.85 cm, while ours changes from 0.93 to 0.77 cm. Confidence filtering therefore reduces the baselines' errors more strongly, while our unfiltered predictions already lie within 1 cm of the surface on average. Figure 4 shows one test case for the three methods, with further examples in Fig. 16.

Method	Mean (cm)↓	> 2 cm (%)↓	> 5 cm (%)↓
V-DPM	2.31	41.5	9.1
4RC	4.84	53.4	20.7
Ours	**0.86**	**6.8**	**0.1**

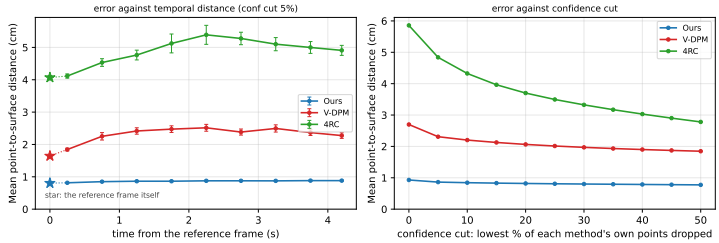

Table 2 and Figure 5: **Amodal 3D localisation on 4D-DRESS.** Left: mean distance to scan and percentages of points more than 2 and 5 cm from the surface. The plots show error versus time (centre) and confidence cut (right), with standard-error bars. The table and temporal plot use a union-wide 5 % confidence cut.

4.4 Human surface consistency in the wild



Figure 6: **Dense correspondences and 4D tracks.** Each row follows a person through three instants, with matching times shown in A and B. (A) The same reference-space 3D chequerboard colours every image, making it possible to follow surface regions through body rotation and garment deformation. (B) The reconstructed person and coloured 3D trajectories, integrated with VGGT-Ω's scene points. Together, the panels connect correspondence on the image with the motion of human surface points in 3D.

On DAVIS (Pont-Tuset et al., 2017), Dense4D's transferred points follow the moving person more faithfully than those of V-DPM and 4RC, whose transfers exhibit displaced or repeated body parts. Figure 7 summarises this visual advantage with matched forward-transfer overlays and reference-time unions, where our predictions form a more coherent human surface. Additional sequences illustrate body rotation and clothing deformation (Figs. 1, 6 and 13). The same reference-space matching supports tracking over long intervals, as demonstrated on an 86 s video with 1,287 frames processed in a single forward pass (Fig. 12).

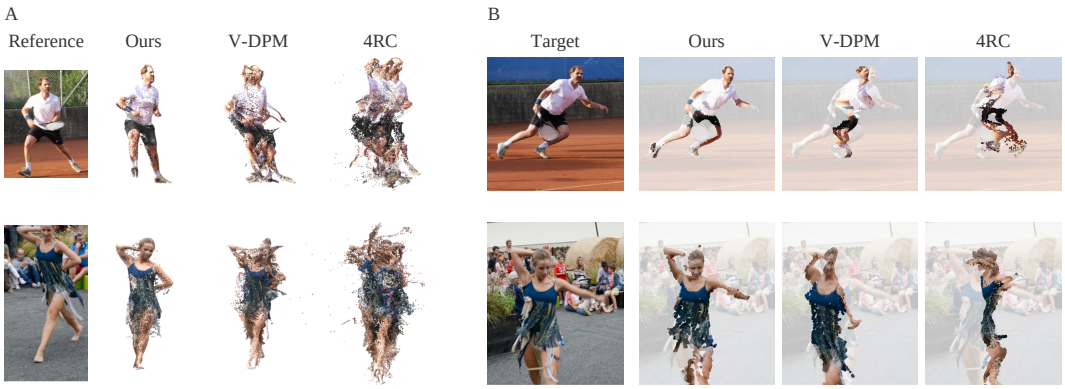

Figure 7: **Human surface consistency on DAVIS.** (A) Reference image and unions of person points from multiple frames, all expressed at frame zero. Notice the distinct body outline preserved by Dense4D, compared with the overlapping limbs and dispersed points in the baseline unions. (B) Frame-zero points transferred to the target image, with their original colours retained against a dimmed background. These colours allow direct inspection of alignment between corresponding body regions. Dense4D anchors transfers to pixels within the target person mask and uses their predicted depth, ensuring consistency with its reconstructed geometry by construction.

The comparisons distinguish plausible reconstruction in each frame from consistent transfer of the same surface points across time. The additional views in Fig. 14 show this directly: the evaluated dynamic scene systems can reconstruct recognisable bodies yet place transferred points away from their own surfaces.

5 Discussion and Conclusion

Dense4D recovers dense 4D correspondence across the body, clothing and hair without a body template. Its Human Point Maps assemble observations from different frames into coherent human geometry in the reference pose, with aligned body parts and fine surface detail. Dense4D separates correspondence from camera-pose and scene-motion estimation while keeping tracks consistent with its predicted human geometry by construction. Compared with the baselines, Dense4D produces more precise trajectories and a better fit to measured human geometry, with videos showing robust tracking through articulation and clothing deformation. Dense4D can also be integrated with scene point clouds simply through shared image pixels.

Limitations. Self-contact can make reference-space correspondence ambiguous when distinct surface points coincide. Selecting a reference frame without self-contact, when available, is a relatively straightforward mitigation that we do not pursue here. A broader challenge is evaluating correspondence on real people: 4D-DRESS measures proximity to scans rather than point identity, while Panoptic Studio Humans relies on reconstruction-derived trajectory labels that may inherit reconstruction errors. Synthetic data provides exact correspondence, but domain differences limit its relevance to real images. Evaluation of scene integration remains open as well: our visualisations demonstrate effective integration with scene point clouds, but its accuracy has not yet been measured quantitatively.

Future work. Temporal and multi-view consistency losses on large corpora of real video could improve correspondence and reduce reliance on synthetic supervision. Developing real-data benchmarks with independently measured surface-point identities would strengthen evaluation beyond surface proximity and reconstruction-derived tracks. The alignment provided by Dense4D could support learning appearance models based on Gaussian splatting (Kerbl et al., 2023) from humans in motion, while its accurate tracks could control human motion in generated video (Geng et al., 2025). Dense4D provides a foundation for augmenting scene understanding systems with more faithful human geometry, dense correspondence and robust tracking. Building on this foundation, future work could develop richer models of people and their surroundings that remain coherent through prolonged motion and interaction.

USE OF LARGE LANGUAGE MODELS

LLM-based tools assisted with implementing components of the system, running and organising experiments, and writing scripts for visualisations using 3D graphics tools. They also assisted with editing and polishing the manuscript and occasionally supported literature and dataset searches. The authors conceived the research, developed the representation and methodology, designed the experiments and interpreted the results. They directed the computational work, reviewed the resulting code and manuscript, and take full responsibility for the paper's claims and conclusions.

ETHICS STATEMENT

This work uses existing research datasets and publicly available video footage, including clips from Pexels (n.d.), in accordance with their respective terms of use. No new recordings of people were collected. Training uses BEDLAM2, THuman2.1 and MammaSyn; evaluation includes existing benchmarks of real videos and 3D scans. Dense4D estimates human geometry and correspondence without predicting biometric identity. Applications should respect the consent and privacy of the people depicted.

Our experiments do not assess demographic differences in performance or suitability for consequential decisions about individuals. The underlying datasets and pretrained models remain subject to their respective terms of use.

REPRODUCIBILITY STATEMENT

The representation, training targets and inference procedure are defined in Sec. 3. The supplementary material details data preparation (Appendix A), implementation and training (Appendix B), and evaluation protocols and baseline settings (Appendix C). We will publicly release our inference code and evaluation pipelines, together with code to reproduce the image and video visualisations.

REFERENCES

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment Anything with Concepts. *arXiv preprint arXiv:2511.16719*, 2025.

Forrester Cole, Kyle Genova, Avneesh Sud, Daniel Vlasic, and Zhoutong Zhang. Differentiable Surface Rendering via Non-Differentiable Sampling. In *ICCV*, 2021.

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 *CVPR*, 2026.

Carl Doersch, Yi Yang, Mel Vecerik, Dilara Gokay, Ankush Gupta, Yusuf Aytar, Joao Carreira, and Andrew Zisserman. TAPIR: Tracking Any Point with per-frame Initialization and temporal Refinement. In *ICCV*, 2023.

Haiwen Feng, Junyi Zhang, Qianqian Wang, Yufei Ye, Pengcheng Yu, Michael J. Black, Trevor Darrell, and Angjoo Kanazawa. St4RTrack: Simultaneous 4D Reconstruction and Tracking in the World. In *ICCV*, 2025.

Michael Garland and Paul S. Heckbert. Surface Simplification Using Quadric Error Metrics. In *SIGGRAPH*, 1997.

Daniel Geng, Charles Herrmann, Junhwa Hur, Forrester Cole, Serena Zhang, Tobias Pfaff, Tatiana Lopez-Guevara, Carl Doersch, Yusuf Aytar, Michael Rubinstein, Chen Sun, Oliver Wang, Andrew Owens, and Deqing Sun. Motion Prompting: Controlling Video Generation with Motion Trajectories. In *CVPR*, 2025.

Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4D: Reconstructing and Tracking Humans with Transformers. In *ICCV*, 2023.

Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. DensePose: Dense Human Pose Estimation In The Wild. In *CVPR*, 2018.

Chen Guo, Tianjian Jiang, Manuel Kaufmann, Chengwei Zheng, Julien Valentin, Jie Song, and Otmar Hilliges. ReLoo: Reconstructing Humans Dressed in Loose Garments from Monocular Video in the Wild. In *ECCV*, 2024.

Adam W. Harley, Zhaoyuan Fang, and Katerina Fragkiadaki. Particle Video Revisited: Tracking Through Occlusions Using Point Trajectories. In *ECCV*, 2022.

Anastasia Ianina, Nikolaos Sarafianos, Yuanlu Xu, Ignacio Rocco, and Tony Tung. BodyMap: Learning Full-Body Dense Correspondence Map. In *CVPR*, 2022.

Zeren Jiang, Yushi Lan, Yihang Luo, Yufan Deng, Zihang Lai, Edgar Sucar, Christian Rupprecht, Iro Laina, Diane Larlus, Chuanxia Zheng, and Andrea Vedaldi. Syn4D: A Multiview Synthetic 4D Dataset. In *ECCV*, 2026.

Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Godisart, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic Studio: A Massively Multiview System for Social Interaction Capture. *IEEE TPAMI*, 2019.

Nikita Karaev, Iurii Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. CoTracker3: Simpler and Better Point Tracking by Pseudo-Labelling Real Videos. In *ICCV*, 2025.

Jay Karhade, Nikhil Keetha, Yuchen Zhang, Tanisha Gupta, Akash Sharma, Sebastian Scherer, and Deva Ramanan. Any4D: Unified Feed-Forward Metric 4D Reconstruction. *arXiv preprint arXiv:2512.10935*, 2025.

Ben Kaye, Tomas Jakab, Shangzhe Wu, Christian Rupprecht, and Andrea Vedaldi. DualPM: Dual Posed-Canonical Point Maps for 3D Shape and Pose Reconstruction. In *CVPR*, 2025.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics*, 42(4), 2023. URL `https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/`.

Umut Kocasari, Simon Giebenhain, Richard Shaw, and Matthias Nießner. Face Anything: 4D Face Reconstruction from Any Image Sequence. In *ECCV*, 2026.

Skanda Koppula, Ignacio Rocco, Yi Yang, Joe Heyward, João Carreira, Andrew Zisserman, Gabriel Brostow, and Carl Doersch. TAPVid-3D: A Benchmark for Tracking Any Point in 3D. In *NeurIPS Datasets and Benchmarks*, 2024.

Samuli Laine, Janne Hellsten, Tero Karras, Yeongho Seol, Jaakko Lehtinen, and Timo Aila. Modular Primitives for High-Performance Differentiable Rendering. *ACM Transactions on Graphics*, 39 (6), 2020.

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding Image Matching in 3D with MASt3R. In *ECCV*, 2024.

Zhengqi Li, Richard Tucker, Forrester Cole, Qianqian Wang, Linyi Jin, Vickie Ye, Angjoo Kanazawa, Aleksander Holynski, and Noah Snavely. MegaSaM: Accurate, Fast, and Robust Structure and Motion from Casual Dynamic Videos. In *CVPR*, 2025.

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.

Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. RefineNet: Multi-Path Refinement Networks for High-Resolution Semantic Segmentation. In *CVPR*, 2017.

William E. Lorensen and Harvey E. Cline. Marching Cubes: A High Resolution 3D Surface Construction Algorithm. In *SIGGRAPH*, 1987.

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *ICLR*, 2019.

Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. Dynamic 3D Gaussians: Tracking by Persistent Dynamic View Synthesis. In *3DV*, 2024.

Yihang Luo, Shangchen Zhou, Yushi Lan, Xingang Pan, and Chen Change Loy. 4RC: 4D Reconstruction via Conditional Querying Anytime and Anywhere. In *ICML*, 2026.

Tuan Duc Ngo, Peiye Zhuang, Chuang Gan, Evangelos Kalogerakis, Sergey Tulyakov, Hsin-Ying Lee, and Chaoyang Wang. DELTA: Dense Efficient Long-range 3D Tracking for any video. In *ICLR*, 2025.

Priyanka Patel and Michael J. Black. CameraHMR: Aligning People with Perspective. In *3DV*, 2025.

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.

Pexels. Pexels Videos, n.d. URL `https://www.pexels.com/videos/`. Accessed 26 September 2026.

Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 DAVIS Challenge on Video Object Segmentation. *arXiv preprint arXiv:1704.00675*, 2017.

Peter Sand and Seth Teller. Particle Video: Long-Range Motion Estimation Using Point Trajectories. In *CVPR*, pp. 2195–2202, 2006.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

Edgar Sucar, Zihang Lai, Eldar Insafutdinov, and Andrea Vedaldi. Dynamic Point Maps: A Versatile Representation for Dynamic 3D Reconstruction. In *ICCV*, 2025.

Edgar Sucar, Eldar Insafutdinov, Zihang Lai, and Andrea Vedaldi. V-DPM: 4D Video Reconstruction with Dynamic Point Maps. In *CVPR*, 2026.

Feitong Tan, Danhang Tang, Mingsong Dou, Kaiwen Guo, Rohit Pandey, Cem Keskin, Ruofei Du, Deqing Sun, Sofien Bouaziz, Sean Fanello, Ping Tan, and Yinda Zhang. HumanGPS: Geodesic PreServing Feature for Dense Human Correspondences. In *CVPR*, 2021.

Jeff Tan, Donglai Xiang, Shubham Tulsiani, Deva Ramanan, and Gengshan Yang. DressRecon: Freeform 4D Human Reconstruction from Monocular Video. In *3DV*, 2025.

Joachim Tesch, Giorgio Becherini, Prerana Achar, Anastasios Yiannakidis, Muhammed Kocabas, Priyanka Patel, and Michael J. Black. BEDLAM2.0: Synthetic Humans and Cameras in Motion. In *NeurIPS Datasets and Benchmarks*, 2025.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual Geometry Grounded Transformer. In *CVPR*, 2025a.

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. VGGT-Ω. In *CVPR*, 2026.

Qianqian Wang, Vickie Ye, Hang Gao, Weijia Zeng, Jake Austin, Zhengqi Li, and Angjoo Kanazawa. Shape of Motion: 4D Reconstruction from a Single Video. In *ICCV*, 2025b.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUSt3R: Geometric 3D Vision Made Easy. In *CVPR*, 2024a.

Wenbo Wang, Hsuan-I Ho, Chen Guo, Boxiang Rong, Artur Grigorev, Jie Song, Juan Jose Zarate, and Otmar Hilliges. 4D-DRESS: A 4D Dataset of Real-World Human Clothing with Semantic Annotations. In *CVPR*, 2024b.

Yuxi Xiao, Qianqian Wang, Shangzhan Zhang, Nan Xue, Sida Peng, Yujun Shen, and Xiaowei Zhou. SpatialTracker: Tracking Any 2D Pixels in 3D Space. In *CVPR*, 2024.

Yuxi Xiao, Jianyuan Wang, Nan Xue, Nikita Karaev, Yuri Makarov, Bingyi Kang, Xing Zhu, Hujun Bao, Yujun Shen, and Xiaowei Zhou. SpatialTrackerV2: 3D Point Tracking Made Easy. In *ICCV*, 2025.

You Xie, Huiqi Mao, Angela Yao, and Nils Thuerey. TemporalUV: Capturing Loose Clothing with Temporally Coherent UV Coordinates. In *CVPR*, 2022.

Frank Yu, Mathieu Salzmann, Pascal Fua, and Helge Rhodin. PCLs: Geometry-Aware Neural Reconstruction of 3D Pose With Perspective Crop Layers. In *CVPR*, pp. 9064–9073, 2021a.

Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. Function4D: Real-time Human Volumetric Capture from Very Sparse Consumer RGBD Sensors. In *CVPR*, 2021b.

Bowei Zhang, Lei Ke, Adam W. Harley, and Katerina Fragkiadaki. TAPIP3D: Tracking Any Point in Persistent 3D Geometry. In *NeurIPS*, 2025.

Chuhan Zhang, Guillaume Le Moing, Skanda Koppula, Ignacio Rocco, Liliane Momeni, Junyu Xie, Shuyang Sun, Rahul Sukthankar, Joëlle K. Barral, Raia Hadsell, Zoubin Ghahramani, Andrew Zisserman, Junlin Zhang, and Mehdi S. M. Sajjadi. Efficiently Reconstructing Dynamic Scenes One D4RT at a Time. In *CVPR*, 2026.

A DATA AND SURFACE SUPERVISION

All three training sources use the surface-label construction in Sec. 3.3: triangle indices and barycentric coordinates provide both HPM and depth targets. We rasterise all people together with a shared depth buffer to account for inter-person occlusion. Figure 8 illustrates the supervision surfaces and input crops.

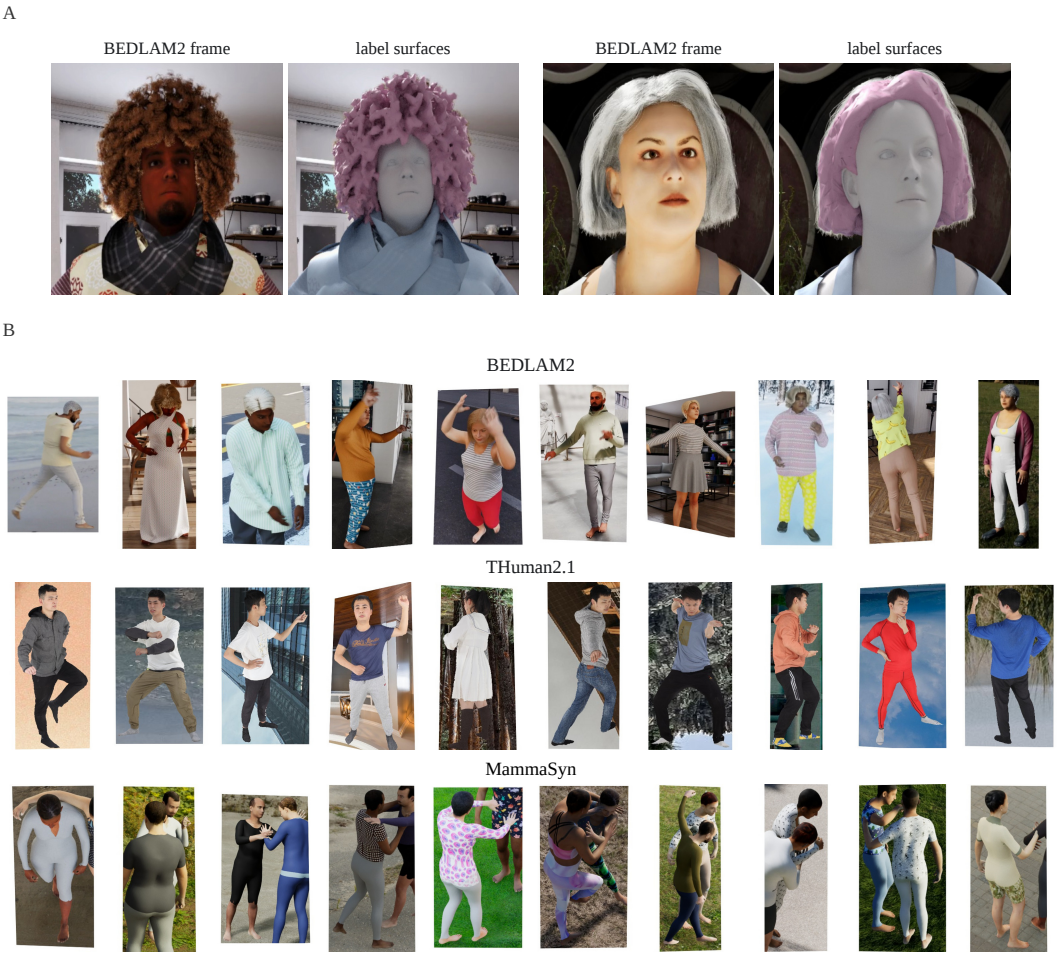

Figure 8: **Training surfaces and samples.** (A) BEDLAM2 images and the body, garment and hair surfaces used for supervision. (B) Training crops from BEDLAM2, THuman2.1 and MammaSyn, each from a different clip or scan.

A.1 BEDLAM2

Holding out every tenth BEDLAM2 (Tesch et al., 2025) track leaves 27,727 training tracks (1,250,307 labelled frames) and 3,081 development tracks, with 1,454 subjects across the two subsets. The official test split contains 7,157 tracks and 181,263 labelled frames from 1,223 subjects.

Body and camera labels are available every fifth frame of the 30 fps video, giving supervision at 6 Hz. The surfaces comprise neutral SMPL-X bodies (Pavlakos et al., 2019), garments with fixed topology across time, and hair envelopes derived from strand grooms.

For each groom, we extract the 0.3 level set of a strand-density field, offset it 3 mm inward and decimate it to 25 % of its faces (Lorensen & Cline, 1987; Garland & Heckbert, 1997). We observe no independently simulated hair motion in the BEDLAM2 renders. The envelope vertices are bound to head triangles and transported with the posed head. Figure 9 shows the envelopes used in Fig. 8(A).

Compared with the renderer's hair mattes, the envelopes achieve approximately 92 % precision and 84 % recall.

Body, garment and hair pixels account for 21, 71 and 8 % of training labels and 18, 72 and 10 % of test labels, estimated from 1,541 training and 1,431 test tracks (Fig. 10).

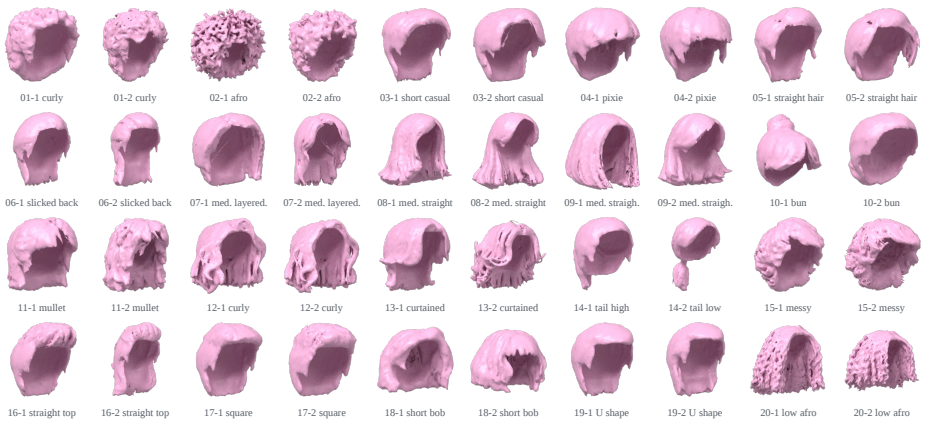

Figure 9: **Hair envelope meshes.** Envelopes derived from BEDLAM2's strand grooms, shown from a common viewpoint with consistent lighting. Labels identify the source grooms.

A.2 THUMAN2.1

We train on 2,200 of the 2,445 textured THuman2.1 (Yu et al., 2021b) scans, holding out every tenth scan. Each scan is rendered from 50 cameras on a jittered orbit: elevation $[-10°, 25°]$, log-uniform vertical field of view $[25°, 90°]$, off-centre aim and projected extent of 55 to 85 % of image height. Images are 1280×720 with panorama backgrounds sampled by camera ray. RGB and labels share the same rasterisation.

A.3 MAMMASYN

We train on 15,710 person tracks in 7,856 views from MammaSyn, MAMMA's synthetic data family (Cuevas-Velasquez et al., 2026). We retain cameras pitched downwards by at least $15°$ and every fourth frame. The photorealistic renderings depict SMPL-X bodies with clothing represented only in the texture, without separate garment or hair meshes.

B IMPLEMENTATION AND TRAINING

B.1 ARCHITECTURE AND VIRTUAL CAMERAS

A $4 \to 1024 \to 1024$ MLP with GELU activation embeds the crop geometry. The ViT image encoder uses DINOv3-L (Siméoni et al., 2025) with patch size 16. Each 512×512 crop supplies 1,024 patch tokens of width 1,024, which are joined by one camera token, 16 register tokens and the crop token. Separate learned camera and register tokens identify the first frame as the reference; the remaining frames share a second set. Camera tokens are internal features, with no camera-prediction head or camera-regression loss. There are 24 pairs of frame and inter-frame attention blocks. Five inter-frame blocks (indices 2, 6, 9, 14 and 20, zero-indexed) update only the camera, register and crop tokens; patch tokens pass through these steps unchanged. The remaining 19 inter-frame blocks attend across all tokens.

Both dense heads use features from blocks 4, 11, 17 and 23 (zero-indexed), projected to resolutions 128^2, 64^2, 32^2 and 16^2, then fused by a 256-channel RefineNet decoder (Lin et al., 2017) and pixel-shuffled to 512^2. HPM coordinates have no output activation. The heads run in fp32 and the backbone in bf16; each head has 32.3 M parameters.

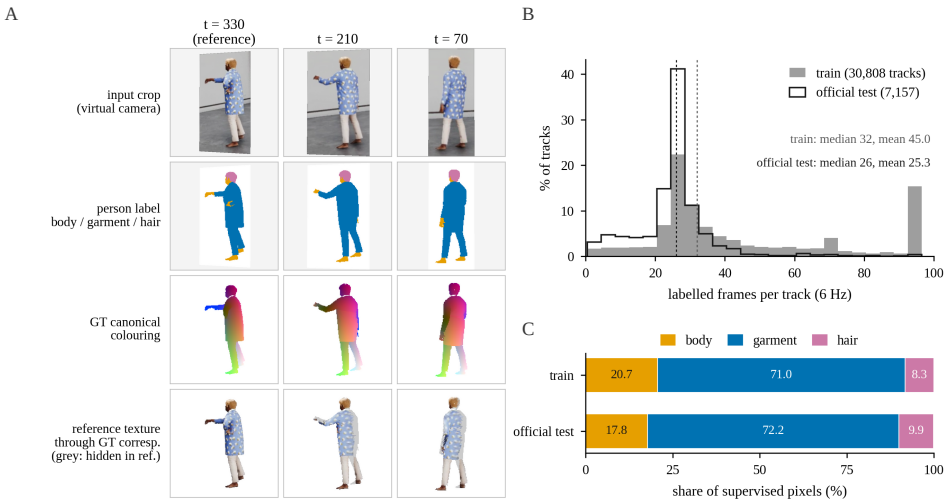

Figure 10: **Surface supervision for 4D correspondence.** A human sequence with body, garment and hair labels, reference-coordinate colours and texture transfer. The plots summarise track length and labelled-pixel composition; the histogram's training pool includes the development tracks.

Table 3: Inference time and memory of the full pipeline on one GH200, for a 1280×720 internet video, with each stage at its own input size. SAM 3 tracks the person on the full frames; VGGT-Ω gives the cameras and scene depth from the full frames at 512×288; Dense4D runs on 512×512 person crops, which carry $1.8\times$ the tokens of VGGT-Ω's frames, so its forward pass is the longer one although the two networks use the same backbone architecture. "Crops, read-out" is CPU time for building the crops and writing the outputs. All models stay loaded in one process (loading, 27 s, is not included); peak memory is the GPU allocator maximum, including 11.4 GB of model weights.

	SAM 3	VGGT-Ω	Dense4D, 512×512 crops			
Frames	1280×720	512×288	crops, read-out	forward	Total	Peak memory
100	10.6 s	1.8 s	0.8 s	4.7 s	17.9 s	19.5 GB
150	18.2 s	3.3 s	1.4 s	10.0 s	32.9 s	21.9 GB
300	33.0 s	11.0 s	2.6 s	36.6 s	83.2 s	29.1 GB

The virtual-camera warp follows Eq. (3), with supervision extrinsics $\mathbf{E}_{v,t} = \mathbf{R}_{v,t}\mathbf{E}_t$, where $\mathbf{E}_t$ is the original world-to-camera transform. Training boxes have 12 px padding; test boxes are smoothed over seven frames and expanded by 12 % on each side. Images are warped bilinearly and labels/masks by nearest neighbour.

B.2 Integration with dense scene reconstruction

To place our human geometry in the scene, we use VGGT-Ω's full-frame depth D_t^Ω and world-to-camera pose $(\mathbf{R}_t^\Omega, \mathbf{t}_t^\Omega)$. The crop warp and person mask pair scene-depth samples with the real-camera depths of our predicted points $\mathbf{P}_t$ from Eq. (7). These depth pairs determine a person-to-scene scale $s_{w,t}$, giving

$$\mathbf{X}_t^{\text{world}} = \mathbf{R}_t^{\Omega\top}(s_{w,t}\mathbf{P}_t - \mathbf{t}_t^\Omega). \tag{8}$$

Scene depth sets the scale only; our predicted depth supplies the human geometry. Background points come from the scene reconstruction outside the person masks. For Panoptic Studio Humans we fit one scale per person over the clip and score tracks in the benchmark cameras (Appendix C.1). For the moving-camera visualisations, $s_{w,t}$ is the per-frame median ratio of VGGT-Ω depth to our depth over the eroded person mask. Table 3 reports inference time and peak GPU memory for the full pipeline.

B.3 Training schedule

Dense4D is trained for a cumulative 686,000 steps, taking approximately 490 hours on four GH200 GPUs. We report the final checkpoint, without weight averaging. AdamW (Loshchilov & Hutter, 2019) uses weight decay 0.05 and peak learning rate 2×10^{-5}, with linear warm-up and cosine decay. Gradient norms are clipped to 1.0 separately for the backbone and each head; activation checkpointing covers attention-block pairs.

For the datasets in Appendix A, the number of input frames per person is sampled from $S \in \{2, 3, 4, 6, 8, 12, 24\}$ with weights $1, 1, 1, 1, 2, 3, 3$. The batch contains 96 images per step, consistently packed as 24 images per GPU; the number of person sequences varies with S. Frames are sampled without replacement where the track length permits.

Photometric augmentation uses colour jitter with probability 0.9 and greyscale with probability 0.05; jitter is shared across frames with probability 0.7. Brightness, contrast and saturation factors lie in $[0.5, 1.5]$, with hue in $[-0.1, 0.1]$.

C Evaluation protocols

C.1 Panoptic Studio Humans

We evaluate 40,779 human queries across all 156 clips of 150 frames, retaining the Dynamic-3DGS pseudo-labels (Luiten et al., 2024). Queries on held objects or with invalid annotations are excluded consistently across methods. The official TAPVid-3D evaluator uses benchmark intrinsics, depth-adaptive thresholds and one median rescaling per clip; scores are averaged over clips. The dataset contains 49 two-person and 107 single-person clips.

SAM 3 (Carion et al., 2025) supplies person masks, and the frame with the largest mask is selected as each person's reference. We fit one scale per person across the clip using pixel-aligned VGGT-Ω and predicted human depths on confident mask interiors. The fit selects the scale supported by the most depth ratios within 5 %, then refines it using the median inlier ratio. VGGT-Ω sets the scale only; our predicted depth supplies the track geometry, without temporal smoothing.

A match is visible when its reference-space residual is below $\tau = 0.005$. Undoing the normalisation in Eq. (2) gives a tolerance of $s\tau$: for example, $s = 2$ m gives 1 cm.

C.2 Baseline implementations

All baselines use the same human queries and evaluator.

SpatialTrackerV2. We use the released front-end and offline tracker (Xiao et al., 2025) in static-camera mode with bidirectional tracking. Its predicted depth and 2D tracks are unprojected using benchmark intrinsics.

TAPIP3D. We use the authors' evaluation code (Zhang et al., 2025) with MegaSaM depth and camera poses (Li et al., 2025), six refinement iterations, no support grid (the default of the authors' evaluation code) and a fixed visibility threshold of 0.5. Published MegaSaM annotations cover the 106 full-split clips; the released script supplies annotations for the 50 minival clips. The published annotations match our TAPVid-3D copy in frame count, resolution and intrinsics. Predicted world-space tracks are transformed into each frame's estimated camera for scoring.

4RC. 4RC (Luo et al., 2026) encodes all 150 frames jointly and decodes positions for each query and target time. Predicted points are projected using its own intrinsics and unprojected using benchmark intrinsics at unchanged depth.

V-DPM. For V-DPM (Sucar et al., 2026), we process pairs of temporal chunks with reference frame zero and align their scales using shared reference pixels. Focal and principal-point corrections use query-frame reprojections and query pixels, without ground-truth 3D.

Other baselines. Any4D (Karhade et al., 2025) groups queries by reference frame and uses all target frames, with implicit intrinsics estimated from predicted rays and corrected to the benchmark convention. Shape of Motion (Wang et al., 2025b) uses its default optimisation.

C.3 4D-DRESS: AMODAL 3D LOCALISATION

Each of the 128 dynamic takes supplies sixteen frames from one camera, beginning at its first frame. All methods receive identical crops and predict points at the first frame's reference time.

For each method, one similarity transform is robustly fitted using only reference-frame pixel correspondences to the scan, then applied unchanged to every frame's predictions. Each method discards its least confident 5 % of points across the sixteen-frame union; reference-frame and temporal statistics use this same retained point set. We report mean one-sided point-to-scan distance in centimetres and the fractions exceeding 2 and 5 cm, averaging these statistics over takes. The distance is one-sided because the observations do not cover the full scan; regions hidden in the reference image are included when observed in other frames. Error bars are standard errors over takes.

C.4 EVALUATION USING SYNTHETIC DATA

The conditioning ablation uses 128 tracks selected at a fixed stride through BEDLAM2's sorted official test split, with four evenly spaced frames per track. We measure Euclidean distance from each predicted point to its ground-truth reference-time position. Known reference-camera geometry and each clip's ground-truth scale (Eq. (2)) convert errors to centimetres before averaging, without fitted alignment or confidence filtering. The temporal analysis uses the same error and scale convention (Appendix D.2).

D ABLATIONS AND ANALYSIS

D.1 CROP ZOOM AND VIEWING DIRECTION

The crop embedding encodes both the centre ray and crop zoom (Eq. (3)); we examine their roles separately. Removing the entire embedding increases BEDLAM2 pointwise reconstruction error from 2.99 to 4.06 cm and reduces Panoptic Studio Humans tracking from 38.4 to 30.4 AJ and from 50.3 to 41.4 APD (Tab. 4). The unconditioned variant has higher reconstruction error on 111 of 128 test tracks and lower tracking AJ on 150 of 156 clips.

Retaining crop zoom while fixing the three centre-ray channels to $(0, 0, 1)$ at inference gives equivalent reconstruction and tracking performance at the reported precision. This is expected because the rotation-induced homography already reorients each crop towards the person. The result suggests that crop zoom carries the useful conditioning information after this rectification. The reported checkpoint retains the centre-ray inputs used during training, but their variation provides no measurable benefit in this experiment.

Variant	Panoptic Studio Humans			BEDLAM2
	AJ↑	APD↑	OA↑	(cm)↓
Without crop conditioning	30.4	41.4	89.3	4.06
Centre ray fixed to $(0, 0, 1)$	**38.4**	**50.3**	**89.6**	**2.99**
Final model	**38.4**	**50.3**	**89.6**	**2.99**

Table 4: **Crop-geometry conditioning.** The fixed-ray row modifies only the final model's inference input. Evaluation follows Tab. 1 for Panoptic Studio Humans and Appendix C.4 for BEDLAM2.

D.2 TEMPORAL EXTENT AND INPUT DENSITY

We use all official-test BEDLAM2 tracks lasting at least ten seconds (155 tracks, 139 subjects), including 25 that last sixteen seconds. Windows start at the reference frame and span 1 to 16 s, sampled at 6 fps for up to 96 frames per forward pass. Evaluation follows Appendix C.4; Fig. 11

reports means and standard errors over tracks. Over a fixed sixteen-second interval, varying the input from 4 to 96 frames gives reconstruction errors of 3.5 to 3.7 cm, including frame counts above the usual training maximum of 24. Extending the interval from one to sixteen seconds increases error from 2.5 to 3.6 cm, while reference-frame error remains approximately 2.0 cm. Point-to-surface distance remains between 0.87 and 0.92 cm, suggesting that the additional pointwise error primarily concerns localisation along the surface.

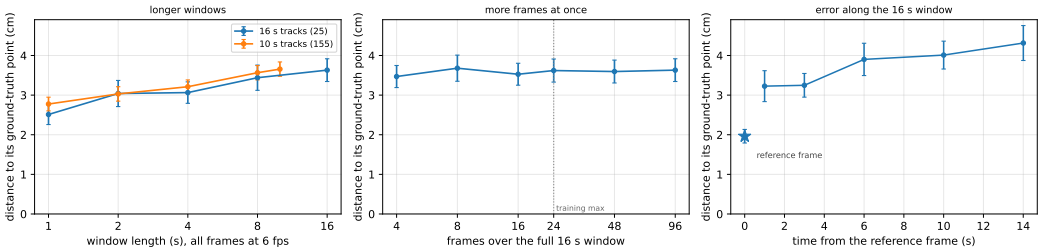

Figure 11: **Temporal extent and input density on BEDLAM2.** Pointwise reconstruction error in centimetres, with means and standard errors over tracks. Left: windows sampled at 6 fps on tracks lasting at least ten seconds (155 tracks) or sixteen seconds (25 tracks). Middle: 4 to 96 frames over the same sixteen-second window. Right: error by temporal offset within that window; the star marks the reference.

E Additional qualitative results

E.1 Tracking over a long video

The video is sampled at 15 fps, using 1280×720 images and 512×512 person crops. VGGT-Ω supplies the scene reconstruction (Appendix B.2).

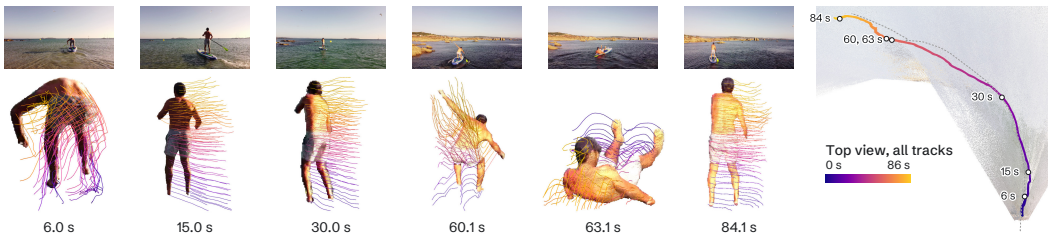

Figure 12: **Tracking over 86 s.** Left: input frames and one-second track segments coloured by body height. Right: full-video tracks coloured by time in VGGT-Ω's scene. The dashed line shows the camera path; circles mark the selected frames.

E.2 ADDITIONAL DAVIS VISUALISATIONS

For Fig. 7, we discard each method's least confident 5 % of points and use identical crops and display scales within each forward-transfer comparison. Overlays are clipped at the target-image boundaries.

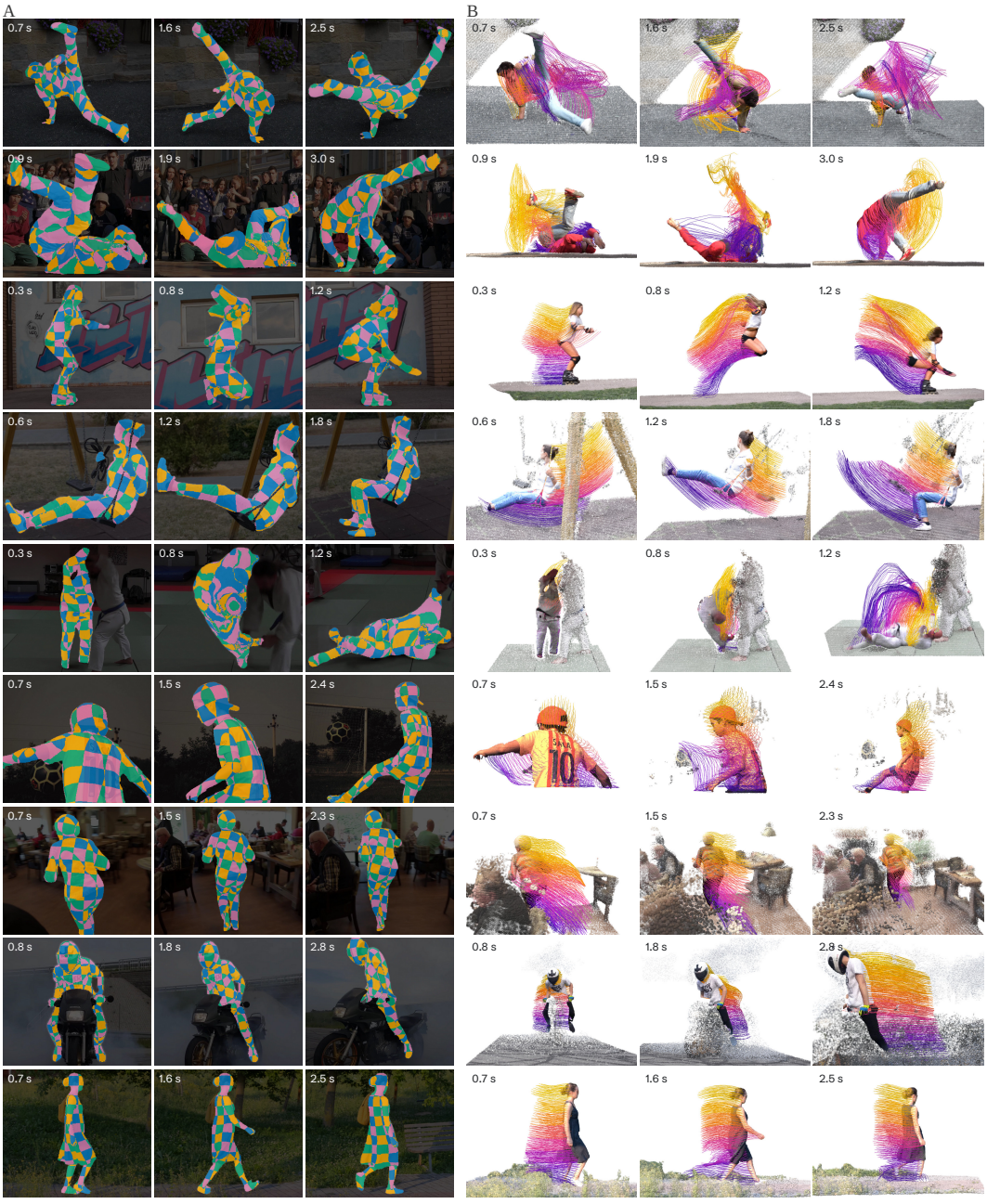

Figure 13: **Dense4D correspondences and 4D tracks on DAVIS.** Each row shows three instants from one clip. (A) A reference-space 3D chequerboard is transferred to the images through HPMs. (B) Our 3D geometry and tracks in the reconstructed scene.

E.3 FORWARD-TRANSFER CONSISTENCY ON DAVIS

Figure 14 compares frame-0 points transferred by Dense4D, V-DPM, 4RC and Any4D with each method's own target-time surface, extending the overlays in Fig. 7. Several methods reconstruct recognisable human geometry at the target time, yet place transferred points away from that geometry. This reveals inconsistency between their point-transfer and depth predictions, even when the target surface itself looks plausible. Dense4D instead matches target-image pixels and lifts them with their predicted depth, so transferred points lie on its target surface by construction. The height colours help assess alignment with the corresponding body regions; proximity to a method's own surface alone does not establish correspondence accuracy.



Figure 14: **Frame 0 carried forward on DAVIS.** Left: frame 0 coloured by body height (head red, feet blue), and the target frame t. Right: each method's surface at t (grey), overlaid with points from frame 0 transferred to t and retaining their height colours. Dense4D preserves the head-to-foot colour ordering through the displayed motions. The baselines show scattered transferred points or displacement from their own target-time geometry.

E.4 Additional Panoptic Studio Humans visualisations

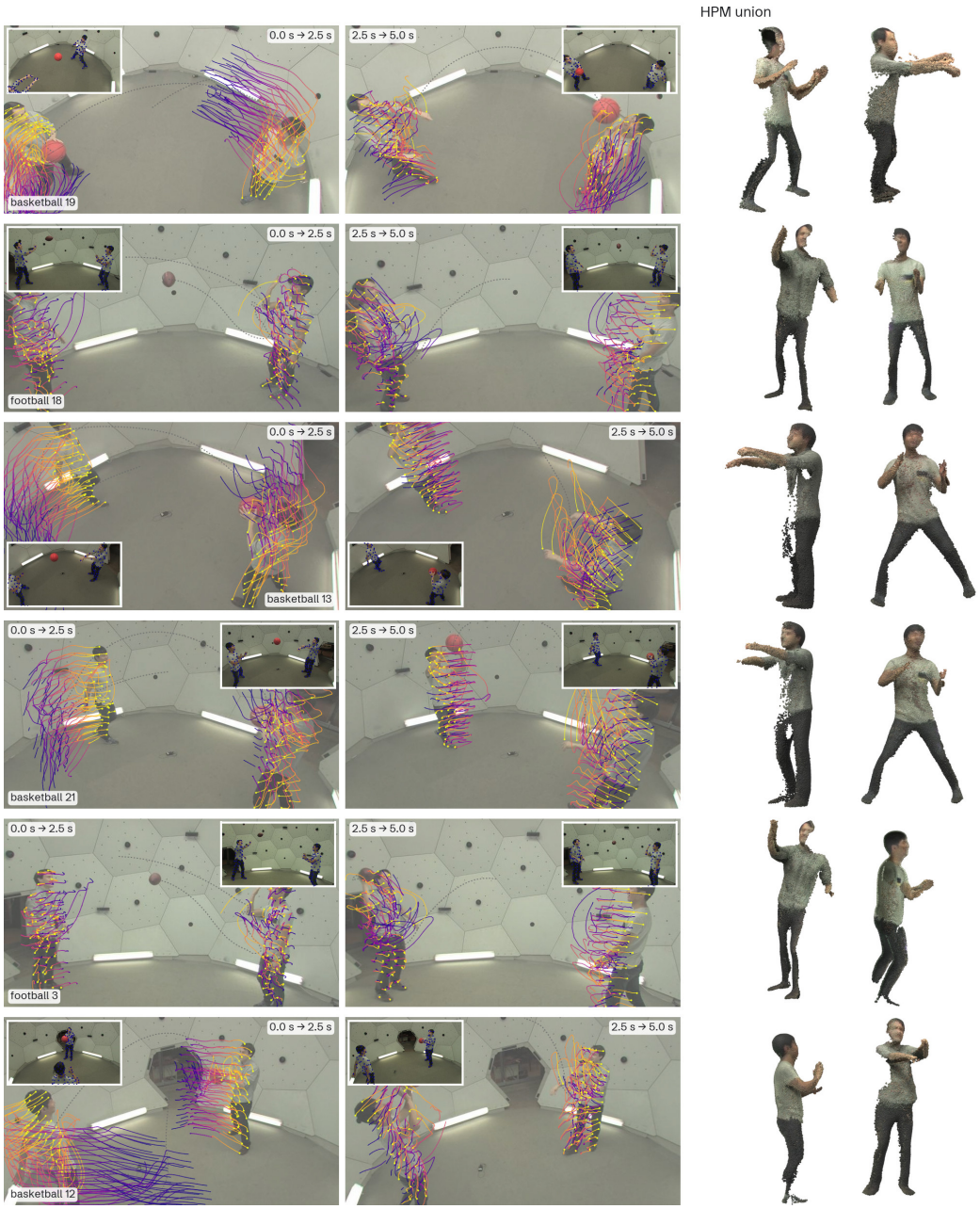

Figure 15: **Additional 4D tracking visualisations on Panoptic Studio Humans.** Each row shows one clip, with tracks over 0.0–2.5 s and 2.5–5.0 s (left) and the two people's reference-time HPM unions (right). Insets show image observations.

E.5 ADDITIONAL 4D-DRESS VISUALISATIONS



Figure 16: **Additional reference-time reconstructions on 4D-DRESS.** Each row shows input images (reference boxed in red) and the unions predicted by Dense4D (HPM), V-DPM and 4RC. For each method, two appearance views are followed by distance to the scan (0–5 cm). Each method discards its least confident 5 % of points. The scan is overlaid in pale blue in the side view.