H-Flow: Self-supervised Human Scene Flow via Physics-inspired Joint Multi-modal Learning

Zhanbo Huang[1], **Xiaoming Liu**[1,2], **Yu Kong**[1]
[1]Michigan State University [2]University of North Carolina, Chapel Hill
[1]{huang247, yukong}@msu.edu [2]liuxm@cs.unc.edu

Abstract

Parametric human models capture global pose but cannot represent the non-rigid surface dynamics of clothing and soft tissue. Generic scene flow estimates dense motion but breaks down on articulated bodies, where pixel-level supervision is also intractable to acquire. We introduce *H-Flow*, a dense human scene flow that captures both skeletal kinematics and surface deformation. A unified multi-head transformer estimates flow from monocular video, jointly predicting pose and depth as companion outputs. The challenge lies in the lack of supervision. In place of unattainable labels, we anchor the network in the physics of human motion, encoding geometric, structural, and biomechanical priors as cross-modal training objectives. We further introduce *DynAct4D*, a high-fidelity synthetic benchmark providing dense flow annotations across diverse subjects, garments, and motions. On standard benchmarks, H-Flow outperforms scene-flow and parametric baselines, generalizes zero-shot to in-the-wild video, and transfers as an effective motion primitive to action recognition and video generation. Models, code, and resources are available at https://zbhuang.com/h-flow.

1 Introduction

Capturing how a person moves involves more than tracking joints. The body's surface deforms in ways skeletal kinematics alone cannot describe. A complete account needs both skeletal structure and dense surface detail. Parametric models [1–6] have become the dominant paradigm, with recent methods enabling robust recovery of 3D pose and shape from a single image [7, 8]. Yet their surfaces are bound to linear blend skinning, where each vertex is a weighted sum of bone transformations. This cannot represent motion that deviates from the skeleton, such as the swirl of a skirt or the deformation of soft tissue [9, 10]. These details are essential for behavior understanding and realistic modeling, yet remain out of reach for skeleton-bound methods (Fig. 1b).

Dense 3D scene flow [11–13] offers an alternative path. It estimates a motion vector for every visible point, providing the density needed to represent clothed humans. But applying it to articulated humans faces a fundamental structural barrier (Fig. 1c). Most methods rely on local rigid-body assumptions to regularize the ill-posed matching problem [14, 15]. Such rigidity priors break down on highly articulated, non-rigid human motion, smearing predictions at limb tips and garment edges. Human-region accuracy thus consistently lags behind generic-scene performance.

To close this gap, we introduce H-Flow, a dense scene flow specialized for human bodies. Every visible human point carries its own 3D motion vector, providing the per-point density needed to capture garment folds and soft-tissue deformation. Beyond density, H-Flow targets the skeletal accuracy of parametric models, expressing articulated motion without resorting to global rigid-body assumptions that contradict anatomical reality. Realizing such a representation requires acquiring human-aligned motion structure under no flow supervision. Pose, depth, and flow appear as three separate modalities, yet they are different projections of one physical system: a human body moving

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

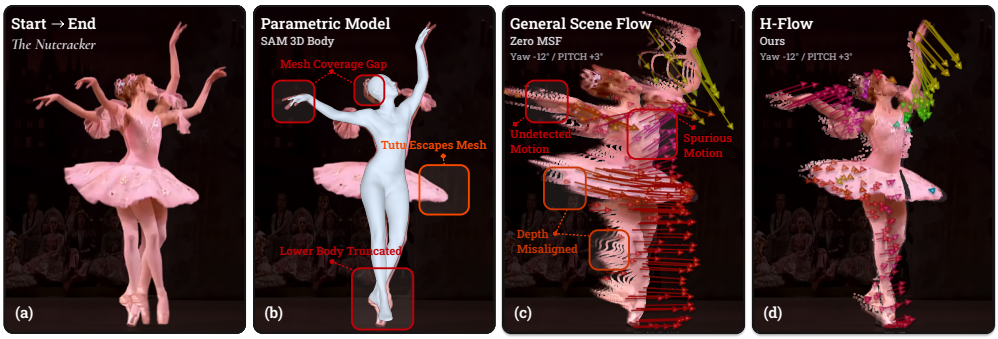

Figure 1: **Comparison of paradigms for human motion perception on an in-the-wild ballet performance.** Evaluated zero-shot on a frame from *The Nutcracker* (Mariinsky Ballet). *(a)* The input video frame. *(b)* Parametric models (SAM 3D Body [8]) recover global pose, but their mesh excludes the tutu and truncates extremities. *(c)* Generic scene flow (Zero MSF [16]) yields undetected motion on the dancer, spurious vectors in static regions, and misaligned depth. *(d)* Our H-Flow produces accurate dense 3D motion across body and garment, anchored in physical priors and learned without any flow ground truth. *The textured backgrounds in (c) and (d) are 3D point clouds reconstructed from each method's predicted depth. See Sec. F.1 for additional viewpoints.*

through space. Body geometry shapes the depth field and skeletal kinematics shape the surface flow, with both couplings arising from the body's physical structure rather than any algorithmic choice. This view turns self-supervision from constraining a modality against itself into constraining modalities against the physics that binds them. This goal lies beyond existing learning paradigms. Supervised methods require dense human flow ground truth that is not available. Self-supervised scene flow methods [14, 15] rely on the rigid-body assumption we set out to discard. We therefore design a new learning paradigm. H-Flow, pose, and depth are jointly estimated in a single architecture, with geometric, structural, and biomechanical priors encoded as cross-modal training objectives. The system trains in a fully self-supervised manner with no flow ground truth. To inject absolute metric scale, we further introduce SAM 3D Body and Depth Pro [8, 17] as anchors. During training, flow progressively aligns with human kinematics. Pose surpasses its anchor on absolute joint position, and depth significantly outperforms its anchor within the body region. The underlying principle is cross-modal mutual benefit. Pose acquires metric awareness through depth and flow, while depth concentrates capacity on the body region through pose and flow.

To quantitatively evaluate H-Flow as a dense human scene flow representation, no existing dataset suffices. Scene flow benchmarks [18–20] target rigid or generic scenes rather than human bodies, and human motion datasets [21–23] lack dense flow annotations. Acquiring human flow ground truth in 2D has been addressed by the optical flow community through synthetic benchmarks [19, 20], yet the 3D case is harder, with real-world capture intractable on non-rigid clothed bodies. Following this synthetic route, we build DynAct4D on the MetaHuman framework [24], generating photorealistic, anatomically diverse human characters with simulated cloth physics inside Unreal Engine [25]. DynAct4D features characters across genders, ethnicities, and body types, dressed in garments of varying length and fit, performing a wide range of actions across multiple scenes and viewpoints. Every frame provides per-pixel flow ground truth, offering a rigorous testbed for human scene flow.

On standard benchmarks, H-Flow outperforms scene-flow and parametric baselines, and DynAct4D provides rigorous evaluation of dense human motion. The same model generalizes zero-shot to in-the-wild ballet and everyday dance videos. Beyond direct evaluation, H-Flow transfers as a motion primitive to downstream tasks, improving both action recognition [26] and conditional video generation [27]. In summary, our contributions are fourfold.

- We introduce H-Flow, a dense scene flow specialized for human bodies that combines the skeletal accuracy of parametric models with the per-point density of generic scene flow.
- We propose a self-supervised learning paradigm that encodes geometric, structural, and biomechanical priors as cross-modal training objectives between pose, depth, and flow, training without flow labels and yielding mutual gains across all three modalities.
- We construct DynAct4D, a high-fidelity synthetic benchmark with dense flow ground truth across diverse subjects, garments, and actions.

- We demonstrate that H-Flow generalizes zero-shot to in-the-wild humans, indicating that physics-inspired cross-modal coupling captures transferable regularities of human motion rather than dataset-specific patterns.

2 Related Work

Parametric Models and Scene Flow. Parametric models like SMPL [1] and SMPL-X [2] provide strong geometric priors for 3D human shape and pose, with deep estimation methods enabling robust recovery from monocular video [7, 8, 28–30]. However, their reliance on Linear Blend Skinning (LBS) cannot model surfaces that move independently of the skeletal topology, such as loose clothing. Recent hybrid and implicit methods attempt to bypass this limitation, either by augmenting parametric meshes with learned cloth layers [31, 32], lifting single images to high-fidelity clothed humans [33, 34], or refining surface details via test-time optimization [35]. All remain anchored to parametric kinematics [36], lacking a unified treatment of non-rigid surface motion across time. Conversely, generic 3D scene flow methods such as RAFT-3D [37], FlowNet3D [12], PointPWC-Net [38], and Self-Mono-SF [39] natively capture dense temporal dynamics, with recent neural-prior [40] and ODE-based [41] formulations further improving generalization. However, these methods rely on local rigid-body assumptions to regularize the ill-posed matching problem [14, 15, 42]. Such rigidity priors break down when confronted with highly articulated, non-rigid human movements. Recent pointmap formulations [43–45] extend pairwise geometric alignment to dynamic videos, recovering 4D trajectories via test-time motion-basis optimization [46] or feed-forward tracking [47, 48] (see also concurrent work [49]). Yet without anatomical priors, generic alignment relies on low-rank bases or per-video fitting, struggling to capture rapid limb articulation and cloth deformation in one pass. Our H-Flow bridges these paradigms, providing a dense human scene flow that unifies skeletal kinematics with fine-grained surface deformation through joint multi-modal estimation.

Self-supervised Motion Learning. Estimating dense human scene flow from monocular video is inherently ill-posed due to projective and non-rigid ambiguities. The absence of real-world dense ground truth further precludes direct supervised training. To constrain the search space, self-supervised scene flow methods rely on rigidity priors, enforcing them either architecturally [15] or through pseudo-label generation [42, 50]. Such regularization causes feature tearing and unnatural stiffening on articulated, non-rigid human bodies. Recent work alternatively scales up training through mixed 3D and weak 2D supervision across diverse synthetic scenes [16, 51], achieving strong zero-shot generalization. These recipes nonetheless inherit biases toward rigid scenes and degrade on deformable surfaces such as clothing. A parallel line of work learns humans directly from monocular video, recovering animatable avatars through implicit fields [52, 53], scene decomposition [54], layered non-rigid rendering [55], or 3D Gaussian splatting [56–58]. These methods target photorealistic appearance rather than dense motion, supervising shape and texture through volume rendering without producing a dense motion field. H-Flow takes a different route to self-supervision. We integrate three classes of physics-inspired priors that govern human motion. Geometric silhouettes, structural kinematics, and biomechanical dynamics together form a unified paradigm that couples motion estimation across modalities.

Human-centric Motion Benchmarks. Existing benchmarks evaluate different facets of human motion, but none directly measure dense surface dynamics. Human3.6M [21], MPI-INF-3DHP [59], 3DPW [22], and AMASS [23] accurately capture 3D kinematics across lab, in-the-wild, and mocap settings [60–62], leaving the clothed surface unmodeled. Fit3D [63] records 611 fitness sequences ($\sim$3M frames) with four synchronized cameras and VICON mocap, yielding faithful on-body SMPL-X vertex motion in tight attire without loose garment dynamics. Dense flow benchmarks cover surface motion broadly but not clothed human dynamics specifically. KITTI [64] targets rigid driving scenes, and MPI Sintel [20] evaluates optical flow in animated scenes with generic content [19, 65]. DeformingThings4D [66] provides non-rigid ground truth, though its animations are dominated by animals and stylized CG humanoid rigs. Human-specific datasets incorporate richer body motion while remaining anchored to parametric meshes [67, 68]. SURREAL [69], AGORA [70], and M3GYM [71] render or fit SMPL bodies with varied clothing and backgrounds, yet evaluate 3D pose and shape rather than surface flow. BEDLAM [72] extends this paradigm by rendering physics-simulated garments on moving bodies, but its evaluation still targets SMPL-X body recovery rather than dense surface motion. 4D-DRESS [73] departs from synthetic pipelines [74–76], providing real-world 4D

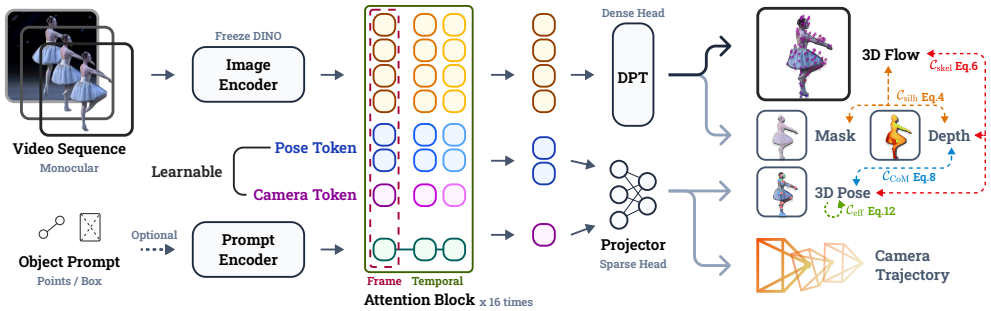

Figure 2: **H-Flow architecture.** We instantiate G_θ as a single transformer producing the outputs of $\mathcal{Y}$ from a frozen visual substrate. Patches and learnable queries (pose, camera) participate in the same attention, so cross-modal coupling exists in the representation before any constraint is computed.

scans of clothing, yet its scale is limited to a small set of subjects and outfits. DynAct4D addresses this gap. Built in a controllable simulator, it provides exact dense flow ground truth for non-rigid human motion across diverse subjects, garments, and actions, enabling rigorous evaluation of surface-level dynamics that existing benchmarks cannot supply.

3 Methodology

3.1 Problem Formulation and Architecture

H-Flow estimates depth, flow, mask, pose, and camera as joint outputs of a single network. This reflects the view that they are projections of one underlying physical system, a body in motion through space. The joint formulation also sidesteps the supervision barrier that direct flow estimation would face. Real-world per-pixel 3D motion on humans cannot be captured directly. Marker-based motion capture records only sparse anatomical landmarks, and skin sliding decouples those markers from the underlying body surface. We therefore cast the problem as joint optimization of H-Flow under a constraint set distilled from physical priors of human movement.

Let $\mathcal{X} = \{I_i\}_{i\in\mathcal{T}}$ denote a video clip indexed by $\mathcal{T}$, where I_i is the frame at index i. A network G_θ maps the clip to a sequence of per-frame estimates,

$$G_\theta : \{I_i\}_{i\in\mathcal{T}} \mapsto \{\mathcal{Y}_i\}_{i\in\mathcal{T}}. \tag{1}$$

Each per-frame tuple $\mathcal{Y}_i$ collects the five outputs jointly produced for frame i,

$$\mathcal{Y}_i = (\mathbf{D}_i,\ \mathbf{F}_i,\ \mathbf{M}_i,\ \mathbf{P}_i,\ \mathbf{C}_i), \tag{2}$$

covering dense depth $\mathbf{D}_i$ at every pixel, forward scene flow $\mathbf{F}_i$ predicting the 3D displacement of every visible point from frame i to frame $i{+}1$, soft human mask $\mathbf{M}_i$ identifying the subject, 3D joint positions $\mathbf{P}_i \in \mathbb{R}^{J\times 3}$ of J skeletal joints, and camera $\mathbf{C}_i$ packaging intrinsics and extrinsics.

We instantiate G_θ as a single transformer over a frozen DINOv3 backbone [77] (Fig. 2). Patch features and learnable queries for pose $\mathbf{P}$ and camera $\mathbf{C}$ participate in the same attention, while a DPT-style dense decoder [78] produces depth $\mathbf{D}$, flow $\mathbf{F}$, and mask $\mathbf{M}$ from the same patch features. This shared substrate routes every gradient through one backbone. Optimization signal from any constraint reaches every output head, instantiating the representation-level coupling within H-Flow. Layer counts, query parameterization, and head architecture are deferred to Sec. A.

Direct regression of flow $\mathbf{F}$ is precluded by the absence of dense supervision on humans. We instead optimize θ against differentiable constraints that encode physical priors of human movement:

$$\theta^\star = \arg\min_\theta\ \mathbb{E}_{\mathcal{X}\sim\mathcal{D}}\left[\sum_k \lambda_k\, \mathcal{C}_k\big(G_\theta(\mathcal{X})\big)\right]. \tag{3}$$

Each constraint $\mathcal{C}_k \geq 0$ measures how much the network outputs $G_\theta(\mathcal{X})$ violate physical prior k, with $\mathcal{C}_k = 0$ corresponding to full satisfaction. This recasts self-supervised motion estimation as constraint satisfaction over physical priors rather than label regression. The constraint set partitions

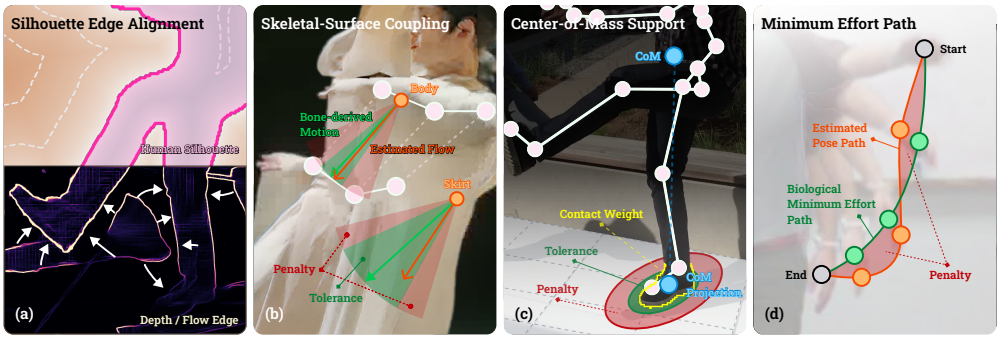

Figure 3: **Self-supervised constraints from physical priors.** Four panels illustrate the four constraints that train G_θ without scene-flow ground truth. *(a) Silhouette Edge Alignment* pulls dense edge responses in depth $\mathbf{D}$ and flow $\mathbf{F}$ toward the mask boundary, with a signed distance field penalizing both interior spurious edges and exterior leakage. *(b) Skeletal-Surface Coupling* ties surface flow to bone-induced motion via a tolerance margin, allowing loose clothing while anchoring on the skeleton. *(c) Center-of-Mass Support* requires the projected mass center to lie within the support polygon formed by ground-contact end-effectors. *(d) Minimum Effort Path* regularizes pose trajectories toward the closed-form minimum-jerk arc between endpoint poses.

into three principled scales. The geometric constraint acts on per-frame silhouette agreement. Structural constraints couple skeletal kinematics with surface flow and ground support at each instant. The biomechanical constraint regularizes multi-frame trajectories against energy-optimal motion templates. Together, the four constraints engage every output modality, shaping each head from multiple angles and instantiating the supervision-level coupling within H-Flow. Sec. 3.2 specifies the four constraints comprising $\{\mathcal{C}_k\}$, illustrated in Fig. 3.

3.2 Self-supervised Constraints from Physical Priors

Fig. 3 illustrates the four constraints described below. Throughout, the mask $\mathbf{M}$ acts as a subject selector rather than a measurement target. It provides the silhouette boundary $\partial\mathbf{M}$, the foreground indicator $\mathbb{1}_{\mathbf{M}}$, or the admissible region within which other constraints apply.

H-Flow couples modalities at two complementary levels. The architecture in Sec. 3.1 couples them at the representation level. As illustrated by the dashed lines in Fig. 2, the four constraints below couple them at the supervisory level, so every output head is shaped by gradients from at least two of them.

Geometric: Silhouette Edge Alignment. Discontinuities in dense estimates should align with the visible body boundary. A reliable depth field $\mathbf{D}$ has its sharp gradients at body edges, not in the interior. A reliable flow field $\mathbf{F}$ changes sharply across the moving subject's outline, not within a uniform shirt. We measure edge response as the gradient norm of each modality, $\|\nabla\mathbf{D}\|_2$ and $\|\nabla\mathbf{F}\|_2$. We then weight the edge response by distance to the mask boundary $\partial\mathbf{M}$. The signed distance field $\mathrm{SDF}_{\partial\mathbf{M}}$ is positive outside the mask and negative inside. Its absolute value is zero on the boundary and grows in either direction. We index pixels as $\mathbf{u} \in \Omega$ throughout, where Ω is the image grid. All dense fields (depth $\mathbf{D}$, flow $\mathbf{F}$, mask $\mathbf{M}$) take values over Ω:

$$\mathcal{C}_{\text{silh}}(\mathbf{D}, \mathbf{F}) = \frac{1}{|\Omega|} \sum_{\mathbf{u}\in\Omega} \big(\|\nabla\mathbf{D}(\mathbf{u})\|_2 + \|\nabla\mathbf{F}(\mathbf{u})\|_2\big) \cdot \big|\mathrm{SDF}_{\partial\mathbf{M}}(\mathbf{u})\big|. \tag{4}$$

Symmetric penalty on both sides of $\partial\mathbf{M}$ pulls interior spurious edges out toward the boundary and exterior leakage in (Fig. 3a). The constraint thus couples depth $\mathbf{D}$ and flow $\mathbf{F}$ through a shared geometric reference provided by mask $\mathbf{M}$.

Structural: Skeletal-Surface Coupling. Surface motion is not free. Every surface point sits within a few centimeters of an underlying bone, and its first-order motion is dominated by that bone's kinematics, even when clothing deforms non-rigidly. We encode this by lifting each foreground pixel into a 3D surface point through depth unprojection. The point is then assigned to its nearest bone, and the predicted flow $\mathbf{F}$ at that pixel is compared against the bone-induced motion derived from the joint trajectory. Let $b^* = (j_1^*, j_2^*)$ index the nearest bone, with $\alpha^* \in [0, 1]$ marking its barycentric position along that bone. The bone-induced motion is a linear interpolation of the two endpoint joint

motions. The actual flow is penalized only when its deviation exceeds an anatomical tolerance $\rho(\mathbf{u})$ that grows with distance from the skeleton:

$$\mathbf{m}_{\text{bone}}(\mathbf{u}) = (1-\alpha^*)\,\Delta\mathbf{P}_{j_1^*} + \alpha^*\,\Delta\mathbf{P}_{j_2^*}, \tag{5}$$

$$\mathcal{C}_{\text{skel}}(\mathbf{F},\mathbf{D},\mathbf{P}) = \frac{1}{N_{\mathbf{M}}}\sum_{\mathbf{u}} \mathbb{1}_{\mathbf{M}}(\mathbf{u}) \cdot \big[\, \|\mathbf{F}(\mathbf{u}) - \mathbf{m}_{\text{bone}}(\mathbf{u})\|_2 - \rho(\mathbf{u}) \,\big]_+, \tag{6}$$

where $\Delta\mathbf{P}_j$ is the per-joint motion vector and $N_{\mathbf{M}} = \sum_{\mathbf{u}} \mathbb{1}_{\mathbf{M}}(\mathbf{u})$ normalizes by foreground area (Fig. 3b). The tolerance follows a piecewise-linear form $\rho(\mathbf{u}) = \rho_{\min} + \alpha \cdot d_{\text{bone}}(\mathbf{u})$. Here $d_{\text{bone}}(\mathbf{u})$ is the 3D distance from the unprojected surface point to its nearest bone. The form admits clothing dynamics. Distant points get a wide margin, while on-skin points stay close to the skeleton. Because the unprojection step involves depth $\mathbf{D}$, the constraint backpropagates through depth, flow, and pose simultaneously, coupling all three.

Structural: Center-of-Mass Support. A standing or stepping body must keep its mass center over a base of support. The predicted center of mass $\boldsymbol{\mu}$ is a weighted average of joint positions, using anatomical mass distribution constants w_j that vary by body segment. Ground contact points are read off the pose $\mathbf{P}$ with depth-based proximity. Their convex hull on the ground plane defines the dynamic support polygon $\mathcal{S}(\mathbf{P},\mathbf{D})$. The constraint penalizes the signed distance by which the projected mass center $\boldsymbol{\mu}_{xy}$ leaves this polygon. Inside the polygon counts as zero violation by construction:

$$\boldsymbol{\mu} = \sum_{j=1}^{J} w_j\,\mathbf{P}_j, \tag{7}$$

$$\mathcal{C}_{\text{CoM}}(\mathbf{P},\mathbf{D}) = \big[\, \text{SDF}\big(\boldsymbol{\mu}_{xy},\, \partial\mathcal{S}(\mathbf{P},\mathbf{D})\big) \,\big]_+. \tag{8}$$

The polygon construction depends on depth $\mathbf{D}$ for ground geometry and on mask $\mathbf{M}$ to restrict candidate contacts to foreground regions. Full details are in Sec. B.1. Beyond stability, the constraint signals when a depth estimate places the body floating above or sinking below the ground (Fig. 3c). It thus couples pose and depth through a single physical principle.

Biomechanical: Minimum Effort Path. Voluntary human movement between two postures does not take an arbitrary detour. Decades of motor control research show that point-to-point movement follows energy-minimizing trajectories, formalized as the minimum-jerk principle [79, 80]. Closely related formulations include minimum torque-change [81] and optimal feedback control [82]. We encode this canonical model directly. For a temporal window $\mathcal{W}$ with endpoint frames i_0 and i_T, we map the endpoint poses $\mathbf{P}_{i_0}$ and $\mathbf{P}_{i_T}$ into joint-angle space through inverse kinematics, yielding $\mathbf{q}_{i_0}$ and $\mathbf{q}_{i_T}$. We then interpolate $\mathbf{q}_i^\star$ along the closed-form minimum-jerk arc between them, and forward-kinematic the result back to position space as the reference pose $\mathbf{P}_i^\star$:

$$\mathbf{q}_{i_0} = \text{IK}\big(\text{sg}(\mathbf{P}_{i_0})\big), \quad \mathbf{q}_{i_T} = \text{IK}\big(\text{sg}(\mathbf{P}_{i_T})\big), \tag{9}$$

$$\mathbf{q}_i^\star = \mathbf{q}_{i_0} + (\mathbf{q}_{i_T} - \mathbf{q}_{i_0}) \cdot \phi(\tau_i), \quad \phi(\tau) = 10\tau^3 - 15\tau^4 + 6\tau^5, \tag{10}$$

$$\mathbf{P}_i^\star = \text{FK}(\mathbf{q}_i^\star), \tag{11}$$

$$\mathcal{C}_{\text{eff}}\big(\{\mathbf{P}_i\}_{i\in\mathcal{W}}\big) = \frac{1}{|\mathcal{W}|}\sum_{i\in\mathcal{W}} \big[\, \|\mathbf{P}_i - \mathbf{P}_i^\star\|_2 - \rho_{\text{eff}} \,\big]_+, \tag{12}$$

where $\tau_i = (i - i_0)/(i_T - i_0)$ and $\text{sg}(\cdot)$ is stop-gradient. The 5th-order polynomial ϕ is the closed-form minimizer of integrated jerk between fixed endpoints. The constraint penalizes the deviation between the estimated pose $\mathbf{P}_i$ and the minimum-jerk reference $\mathbf{P}_i^\star$, beyond a tolerance margin ρ_{eff}. Stop-gradient on the endpoints prevents the network from co-adapting boundary poses to the reference, anchoring regularization on intermediate dynamics (Fig. 3d). Although the constraint nominally reads only pose $\mathbf{P}$, gradient flows back through the shared trunk and reaches every modality.

Anchoring and consistency. The four constraints above are not the complete set of optimization terms. To prevent degenerate minima admitted by self-supervision alone, we apply a margin-based external anchor on depth $\mathbf{D}$ and pose $\mathbf{P}$ against off-the-shelf predictors [8, 17], plus a single-camera consistency term on camera $\mathbf{C}$. These auxiliary terms support the optimization but do not define the physical-prior framework. Their formulation is deferred to Sec. B.2. Sec. 5 evaluates the resulting estimator across multiple datasets.

Table 1: **DynAct4D vs existing human motion and scene flow benchmarks, and representative frames.** DynAct4D pairs simulated clothed humans with vertex-level flow ground truth, while prior synthetic benchmarks either composite 2D flow with depth or omit flow entirely. "—" indicates unavailable. Right: representative frames showing diverse subjects, garments, and actions under different viewpoints.

Dataset	Domain	Source	Cloth	Hair	4D Flow GT	Frames
Human3.6M [21]	Human	Real	—	—	—	3.6M
3DPW [22]	Human	Real	SMPL fits	—	—	51K
Fit3D [63]	Human	Real	Tight SMPL-X	—	Vertex (tight)	3.0M
AMASS [23]	Human	Mocap	SMPL body	—	—	—
KITTI [64]	Scene	Real	—	—	LiDAR	400
MPI Sintel [20]	Scene	Syn.	—	—	Composite	1K
DeformingThings4D [66]	Animal	Syn.	—	—	Vertex	500K
AGORA [70]	Human	Syn.	Rigged	Static	—	17K
M3GYM [71]	Human	Syn.	SMPL fits	—	—	4M
BEDLAM [72]	Human	Syn.	Simulated	Static	Composite	380K
DynAct4D (Ours)	**Human**	**Syn.**	**Simulated**	**Groom**	**Vertex**	**720K**



4 DynAct4D Benchmark

DynAct4D bypasses the composite path used by prior synthetic benchmarks (Sec. 2) by tracking the physical surface through the renderer's skinning stage. We record the world-space position of every skinned mesh vertex at floating-point precision. A custom G-buffer pass then propagates this vertex motion to every pixel through triangle identifiers and barycentric coordinates captured at rasterization. With 24K vertices per character, the mesh resolution exceeds the pixel grid at all camera distances we render, so every pixel maps to at least three host vertices. The resulting per-pixel flow is exact rather than interpolated from sparse samples.

DynAct4D is built on 10 characters from MetaHuman [24] framework, generated within Unreal Engine [25], spanning genders, ethnicities, body types, and heights. Each character is paired with 10 garment configurations, from tight-fitted to loose-flowing. Garments run under Chaos Cloth with solver Geometry Cache vertices exported directly, and hair uses strand-based GPU Groom. Across the six loose outfits, non-rigid cloth deformation contributes 24.1% of total displacement beyond linear blend skinning, averaging 22 mm per frame and peaking at 57 mm along dress hems. Hair strands add a 10.8 mm mean offset over rigid head motion. Every subject performs 10 action categories of $\sim$15 s each, sourced from the Mixamo animation library [83] and retargeted to the MetaHuman rig. Actions cover near-static postures, locomotion, and highly articulated movements such as dancing and martial arts. We sample 100 motion sequences from the 10^3 candidate combinations. Each sampled motion is rendered at all 8 synchronized viewpoints across 3 environments spanning indoor and outdoor scenes. The result is 800 test sequences totaling $\sim$720K frames at 1920×1080 and 60 fps, alongside synchronized synthetic LiDAR point clouds. Tab. 1 shows representative frames. We position DynAct4D entirely as a test set. Controlled subject, garment, and action strata isolate individual physical factors for fine-grained diagnostic evaluation. Every method reports results under zero-shot transfer, trained on external data and evaluated on DynAct4D without exposure to its subjects, actions, scenes, or viewpoints.

5 Experiments

We pretrain H-Flow on the full Human3.6M [21] dataset to learn cross-modal human priors. Since we do not evaluate on Human3.6M, all seven subjects (S1, S5, S6, S7, S8, S9, S11) contribute to pretraining. To establish in-domain evaluation on Fit3D [63] and 3DPW [22], we then fine-tune on their respective train splits. For Fit3D we use S3, S4, S5, S7, S8, S9 for training and S10, S11 for testing; for 3DPW we use the official partition. DynAct4D serves as an out-of-domain benchmark, evaluated under zero-shot inference. Enabled by the proposed self-supervised paradigm, H-Flow trains on RGB video alone, without any modality ground truth from these datasets. We compare against representative methods spanning multiple categories of monocular motion estimation, with task-specific baselines and citations in the corresponding tables. We implement H-Flow in PyTorch with mixed-precision training over 4×10^5 iterations on 6 NVIDIA H100 GPUs.

Table 2: **Scene flow estimation on Fit3D and DynAct4D.** End-Point-Error (EPE) in millimeters; 1−Cos is one minus cosine similarity between predicted and ground-truth flow vectors; Acc.S and Acc.R are the fractions of pixels within 50 mm and 100 mm of the ground-truth flow. Best results in **bold**, second-best *underlined*.

Methods		Fit3D [63] In-Domain				DynAct4D Out-of-Domain			
Name	**Supervision**	**EPE↓** mm	**1−Cos↓**	**Acc.S↑**	**Acc.R↑**	**EPE↓** mm	**1−Cos↓**	**Acc.S↑**	**Acc.R↑**
H-Flow Ours	Self Sup.	**24.2**	**0.11**	**0.88**	**0.92**	**28.2**	**0.14**	**0.82**	**0.88**
ZeroMSF* [16]	Geom. Consist.	59.9	0.33	0.41	0.48	78.5	0.42	0.33	0.42
Self-Mono-SF* [39]	Geom. Consist.	50.2	0.31	0.38	0.40	68.0	0.39	0.28	0.32
Depth Pro + H-MoRe [17, 84]	Self Sup.	*38.3*	*0.18*	*0.81*	*0.88*	*42.2*	*0.20*	*0.72*	*0.81*
Depth Pro + RAFT [17, 85]	Geom. Consist.	85.3	0.41	0.35	0.47	122.1	0.58	0.16	0.27
SAM 3D Body [8]	N/A	49.6	0.26	0.78	0.82	56.8	0.25	0.66	0.70
WHAM [29]	Sup.	57.2	0.33	0.58	0.74	62.5	0.32	0.58	0.62

5.1 Scene Flow Estimation

We adopt Fit3D and DynAct4D as the benchmarks for this section. Fit3D combines professional motion capture with multi-view fusion to produce high-fidelity SMPL-X annotations, and all subjects wear tight-fitting attire so that the same vertex remains stable across consecutive frames; per-vertex displacement therefore serves as faithful scene flow ground truth. DynAct4D provides per-pixel scene flow ground truth and includes garment-fold dynamics, further probing how H-Flow handles fine-grained surface deformation on clothed humans. 3DPW is excluded as a scene flow source because its SMPL annotations are not directly mocap-captured and its subjects are not in tight-fitting attire; we therefore use it only for pose and depth evaluation.

We compare H-Flow against six representative baselines, two from each of three categories that produce dense scene flow. *(i) Scene flow* [16, 39] regresses 3D motion vectors from RGB pairs. *(ii) Optical flow lifting* [17, 84, 85] lifts off-the-shelf 2D optical flow into 3D using estimated depth. *(iii) Mesh residual* [8, 29] reads vertex displacement from a fitted body model and propagates it to dense pixels via nearest-neighbor.

Quantitative Results. As shown in Tab. 2, H-Flow achieves the lowest 3D end-point error on both datasets across every metric. On Fit3D, the closest competitor is the Depth Pro + H-MoRe pipeline at 38.3 mm, which H-Flow improves to 24.2 mm, a 37% reduction. The same baseline on DynAct4D reaches 42.2 mm, against H-Flow's 28.2 mm, a 33% reduction under out-of-domain transfer. The gap widens for generic scene flow methods that lack human-specific priors. ZeroMSF [16] and Self-Mono-SF [39] leave EPE above 50 mm even after Fit3D fine-tuning, as rigid regularizers do not transfer to articulated bodies (Sec. D.1 gives speed and region breakdowns).

Qualitative Results. Fig. 4 visualizes flow predictions on representative sequences from both datasets. Generic scene flow baselines produce coherent motion on torsos but smear at limb tips and clothing edges, where rigid-motion priors break. Pipelines that composite depth with optical flow accumulate edge artifacts at the silhouette boundary, exactly the regions where physical priors anchor H-Flow's predictions. Mesh residual methods preserve global pose but flatten secondary surface dynamics, missing the swirl and compression that distinguish clothed-body motion. H-Flow recovers fine-grained surface flow that aligns with both skeletal kinematics and clothing deformation.

4D Tracking, Downstream Transfer, and Efficiency. Feed-forward 3D trackers [47, 48] anchor point trajectories to a reference frame. Under a two-frame window, skeletal coupling maintains stronger directional coherence across articulated limbs (Tab. 3a). We further pass predicted flow as a motion primitive into frozen recognition [86] and video synthesis [87] backbones on Kinetics-400 and UTD-MHAD [88] (Tab. 3b). Dense 3D surface flow disentangles camera projection and retains both skeletal trajectories and garment motion, improving downstream discrimination and synthesis. Decoding all five heads in one pass also amortizes feature extraction across the shared trunk, combining real-time throughput with non-rigid surface accuracy (Tab. 3c).

5.2 Companion Modality Estimation

Pose Estimation. Pose and depth emerge as companion outputs rather than dedicated optimization targets. To isolate cross-modal coupling under a controlled budget, we retrain image-based estimators



Figure 4: **Qualitative scene flow comparison.** *(top)* Fit3D *[in-domain]*, *(bottom)* DynAct4D *[out-of-domain]*. Scene flow baselines smear on articulated limbs and clothing. Lifting pipelines accumulate edge artifacts at silhouette boundaries. Mesh residual methods recover skeletal motion but flatten surface dynamics on garments. *A higher-resolution version is provided in the supplementary material.*

Table 3: **4D tracking comparison, downstream transfer, and computational cost.** *(a)* Feed-forward 4D tracking under a two-frame window (EPE in mm, angular error in rad, Acc@25 in %). *(b)* Flow-I3D action recognition on Kinetics-400 and LGC-VD 15-frame synthesis on UTD-MHAD (†TV flow on K400, RAFT on UTD). *(c)* Parameters, GFLOPs, and FPS. Best in **bold**, second-best *underlined*.

Methods	Feed-forward 4D		
	EPE↓	Ang.↓	Acc↑
H-Flow (Ours)	**42.6**	**0.403**	**55.9**
SpatialTrackerV2 [48]	44.3	0.489	46.6
Align3R + FF [45, 89]	44.3	0.432	46.1
St4RTrack [47]	161.8	0.899	32.2

Methods	Act. Rec. (K400)		Video Gen. (UTD)	
	Top-1↑	Top-5↑	SSIM↑	FVD↓
H-Flow (Ours)	**69.4**	**88.7**	**96.2**	**105.63**
2D Flow† [85]	61.5	83.4	90.3	331.22
SAM 3D Body [8]	66.8	88.3	94.6	164.75
DP + RAFT [17, 85]	58.2	84.1	92.0	429.27

Methods	Computational Cost		
	Param M	GFLOPs	FPS
H-Flow (Ours)	878.2	3,480	5.8
Self-Mono-SF [39]	5.8	188	22.6
ZeroMSF [16]	611.6	3,064	8.2
SAM 3D Body [8]	1285.4	5,816	1.8
DP + RAFT [17, 85]	957.2	39,428	0.6

and 2D-to-3D lifters under a unified data and resolution protocol on Fit3D and active-motion 3DPW frames (Tab. 4). Under this shared protocol, H-Flow achieves the lowest MPJPE on both datasets. This gain stems from how absolute scale enters the representation. Dedicated pose estimators lift 2D observations under monocular scale ambiguity, inferring camera-to-subject distance indirectly from bone-length or camera priors. In H-Flow, shared features already encode metric 3D geometry to support scene flow, allowing pose queries to read world-space localization directly through attention. On PA-MPJPE, where global scale and rotation are removed by Procrustes alignment, methods such as HiPART [90] and MotionAGFormer [91] remain strong. These take 2D keypoints as input and benefit from the much larger pose datasets available for training relative-joint configuration. H-Flow's pose head sits competitively here as one of five joint outputs rather than a dedicated pose pipeline. Sec. D.2 further examines scaling behavior on unfiltered 3DPW.

Depth Estimation. Ground-truth depth on Fit3D and 3DPW is derived by rasterizing the official SMPL-X fits. The resulting per-pixel depth covers the body surface only, while clothing and scene depth are not annotated. We evaluate every method within this body region uniformly (Tab. 4; full metrics in Sec. D.2). This region happens to align with H-Flow's design, where the depth head is consumed only by priors that operate on the body surface. Within this region, H-Flow substantially improves over generic monocular metric depth predictors that estimate depth uniformly over the full scene. On Fit3D, MAE drops from 234.3 mm for Depth Pro [17] to 124.5 mm for H-Flow, a 47% reduction. The same pattern holds across all metrics on both Fit3D and 3DPW. Generic predictors allocate model capacity across backgrounds and clothing alike. H-Flow concentrates capacity on the human surface that matters for downstream scene flow. Joint decoding also keeps depth and pose aligned in one 3D frame, avoiding cross-modal specialist conflicts (Sec. E).

5.3 Ablations

We ablate each physical prior and the external anchor in turn (Tab. 5).

Cross-modal effects. Each prior degrades a different lead modality. $\mathcal{C}_{\text{silh}}$ mainly sharpens depth, with indirect gains on scene flow at edge regions. $\mathcal{C}_{\text{skel}}$ is the dominant scene flow regularizer, removing it doubles EPE on Fit3D and more than doubles it under out-of-domain transfer to DynAct4D. $\mathcal{C}_{\text{CoM}}$

Table 4: **Companion modality estimation on 3DPW and Fit3D.** *(a)* Pose estimation under a unified training protocol, mean per-joint position error (MPJPE) and Procrustes-aligned MPJPE (PA) in mm. Methods marked with † take 2D keypoints as input. *(b)* Depth estimation within the SMPL-X body region, mean absolute error (MAE) in mm and SiLog the scale-invariant log error. Best results in **bold**, second-best *underlined*.

Methods	3DPW		Fit3D	
	MPJPE↓ mm	PA↓ mm	MPJPE↓ mm	PA↓ mm
H-Flow (Ours)	**91.5**	48.7	**112.8**	55.4
HiPART† [90]	112.4	**44.5**	138.5	**51.2**
MotionAGFormer† [91]	108.6	45.8	135.8	*52.6*
MotionBERT† [92]	110.2	46.3	137.4	53.2
SAM 3D Body [8]	*95.0*	*45.0*	*116.0*	54.0
CameraHMR [93]	98.3	46.5	120.4	55.8
PromptHMR [94]	102.7	48.9	124.1	57.2

Methods	3DPW		Fit3D	
	MAE↓ mm	SiLog↓	MAE↓ mm	SiLog↓
H-Flow (Ours)	**158.6**	**7.2**	**124.5**	**2.9**
Depth Pro [17]	527.7	12.4	234.3	4.4
Depth Anything V2 [95]	485.3	11.8	218.7	4.1
MoGe-2 [96]	*452.1*	*11.3*	*205.4*	*3.9*
Marigold [97]	498.2	12.0	225.6	4.2
ZoeDepth [98]	512.8	12.2	228.9	4.3

Table 5: **Ablation of physical priors and the external anchor.** Rows 2–5 drop one prior each, and the last row replaces the external anchor with self-supervised cues. Metrics: EPE (mm) and 1−Cos for flow, MPJPE and PA (mm) for pose, MAE (mm) and SiLog for depth (lower is better).

Methods	Flow on Fit3D		Flow on DynAct4D		Pose on 3DPW		Depth on Fit3D	
	EPE↓ mm	1−Cos↓	EPE↓ mm	1−Cos↓	MPJPE↓ mm	PA↓ mm	MAE↓ mm	SiLog↓
H-Flow	**24.2**	**0.11**	**28.2**	**0.14**	**91.5**	**48.7**	**124.5**	**2.9**
w/o $\mathcal{C}_{\text{silh}}$	32.9	0.15	41.8	0.22	92.6	49.3	165.0	3.3
w/o $\mathcal{C}_{\text{skel}}$	48.4	0.22	63.6	0.21	98.1	52.3	164.8	3.1
w/o $\mathcal{C}_{\text{CoM}}$	26.2	0.12	30.9	0.16	104.6	54.5	133.0	3.3
w/o $\mathcal{C}_{\text{eff}}$	42.7	0.14	44.5	0.18	102.8	52.8	126.1	3.0
Anchor → self-sup.	55.1	0.13	44.8	0.17	183.3	47.2	276.8	2.8

leaves the largest mark on pose, where the support-polygon constraint contributes physical plausibility. $\mathcal{C}_{\text{eff}}$ acts on pose trajectories yet propagates through skeletal coupling. Pose error grows modestly while Fit3D flow EPE rises by 76%, as every prior shapes multiple heads via the shared trunk.

External anchor as anthropometric reference. To isolate metric scale from physical structure, we replace the external predictors with input-driven self-supervised cues (Sec. B.2). This replacement splits the metrics into two regimes. Absolute errors (EPE, MPJPE, MAE) double or triple, while scale-invariant metrics (PA-MPJPE, SiLog) remain at the full-model level. The four physical priors are scale-invariant by construction; none constrains the global metric scale of the world. The external anchor injects this missing degree of freedom during training through noisy reference predictions, and the network internalizes it as a learned anthropometric prior. At inference time the prior is self-contained, and H-Flow's out-of-domain DynAct4D results (Tab. 2) confirm that this anthropometric scale transfers across distributions without runtime calibration.

6 Conclusion

In this paper, we show that physical priors on the body's surface and skeleton supervise dense scene flow from monocular video without any 4D ground truth. A learned anthropometric prior, internalized at training, fixes the global metric scale at inference. H-Flow sets a new state of the art on absolute scene flow across three datasets while preserving competitive companion modalities at inference without runtime calibration. Two limitations remain. The four physical priors are scale-invariant by construction, and H-Flow relies on an external anchor against noisy reference predictions to fix global metric scale during training. A fully self-supervised path to anthropometric grounding stays open. Moreover, ground-support and near-bone priors assume grounded contact and bounded cloth offset, softening under aerial maneuvers or extreme hem lag. Beyond the human body, dense flow over the contact between bodies and objects is the natural next step. Extending H-Flow to human-object interaction is a direct application of its existing architecture, with new physical priors needed for contact, grasp, and release dynamics.

Acknowledgments and Disclosure of Funding

This research was partially sponsored by the Army Research Office (ARO) under Grant Number W911NF-24-1-0385 and the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via 2022-21102100004. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ARO, ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

[1] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: a skinned multi-person linear model. *ACM Transactions on Graphics (TOG)*, 34(6):1–16, 2015. 1, 3

[2] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019. 3

[3] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T. Freeman, Rahul Sukthankar, and Cristian Sminchisescu. GHUM & GHUML: Generative 3D human shape and articulated pose models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[4] Ahmed A. A. Osman, Timo Bolkart, and Michael J. Black. STAR: Sparse trained articulated human body regressor. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[5] Ahmed A. A. Osman, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. SUPR: A sparse unified part-based human representation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.

[6] Aaron Ferguson, Ahmed A. A. Osman, Berta Bescos, Carsten Stoll, Chris Twigg, Christoph Lassner, et al. MHR: Momentum human rig. *arXiv preprint arXiv:2511.15586*, 2025. 1

[7] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4D: Reconstructing and tracking humans with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14783–14794, 2023. 1, 3, 21

[8] Xitong Yang, Devansh Kukreja, Don Pinkus, Anushka Sagar, Taosha Fan, Jinhyung Park, Soyong Shin, Jinkun Cao, Jiawei Liu, Nicolas Ugrinovic, Matt Feiszli, Jitendra Malik, Piotr Dollár, and Kris Kitani. SAM 3D Body: Robust full-body human mesh recovery. *arXiv preprint arXiv:2602.15989*, 2026. 1, 2, 3, 6, 8, 9, 10, 20, 22, 24

[9] Hugo Bertiche, Meysam Madadi, and Sergio Escalera. PBNS: Physically based neural simulation for unsupervised garment pose space deformation. *ACM Transactions on Graphics (TOG)*, 40(6):1–14, 2021. 1

[10] Astitva Srivastava, Hsiao-yu Chen, Ryan Goldade, Philipp Herholz, Zhongshi Jiang, Gene Wei-Chin Lin, Lingchen Yang, Nikolaos Sarafianos, Tuur Stuyck, and Egor Larionov. PhySkin: Physics-based bone-driven neural garment simulation. *arXiv preprint arXiv:2603.27013*, 2026. 1

[11] Sundar Vedula, Simon Baker, Peter Rander, Robert Collins, and Takeo Kanade. Three-dimensional scene flow. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 1999. 1

[12] Xingyu Liu, Charles R. Qi, and Leonidas J. Guibas. FlowNet3D: Learning scene flow in 3D point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 529–537, 2019. 3

[13] Gilles Puy, Alexandre Boulch, and Renaud Marlet. FLOT: Scene flow on point clouds guided by optimal transport. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 1

[14] Yancong Lin and Holger Caesar. Icp-flow: Lidar scene flow estimation with icp. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15501–15511, 2024. 1, 2, 3

[15] Yancong Lin, Shiming Wang, Liangliang Nan, Julian Kooij, and Holger Caesar. Voteflow: Enforcing local rigidity in self-supervised scene flow. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17155–17164, 2025. 1, 2, 3

[16] Yiqing Liang, Abhishek Badki, Hang Su, James Tompkin, and Orazio Gallo. Zero-shot monocular scene flow estimation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21031–21044, 2025. 2, 3, 8, 9, 24, 25

[17] Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second. *arXiv preprint arXiv:2410.02073*, 2024. 2, 6, 8, 9, 10, 20, 22, 24

[18] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3354–3361, 2012. 2

[19] Nikolaus Mayer, Eddy Ilg, Philip Häusser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4040–4048, 2016. 2, 3

[20] Daniel J Butler, Jonas Wulff, Garrett B Stanley, and Michael J Black. A naturalistic open source movie for optical flow evaluation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 611–625. Springer, 2012. 2, 3, 7

[21] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 36(7):1325–1339, 2014. 2, 3, 7

[22] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 614–631, 2018. 3, 7, 22

[23] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5442–5451, 2019. 2, 3, 7

[24] Epic Games. Metahuman: A complete framework for photorealistic digital humans in Unreal Engine. https://dev.epicgames.com/documentation/en-us/metahuman/metahuman-documentation, 2024. Accessed: 2026-04-30. 2, 7

[25] Epic Games. Unreal Engine 5. https://www.unrealengine.com/en-US/unreal-engine-5, 2022. 2, 7

[26] Jathushan Rajasegaran, Georgios Pavlakos, Angjoo Kanazawa, Christoph Feichtenhofer, and Jitendra Malik. On the benefits of 3D pose and tracking for human action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 640–649, 2023. 2

[27] Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8153–8163, 2024. 2

[28] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7122–7131, 2018. 3

[29] Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J Black. WHAM: Reconstructing world-grounded humans with accurate 3d motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 8

[30] Zhongang Cai, Wanqi Yin, Ailing Zeng, Chen Wei, Qingping Sun, Yanjun Wang, Hui En Pang, Haiyi Mei, Mingyuan Zhang, Lei Zhang, Chen Change Loy, Lei Yang, and Ziwei Liu. SMPLer-X: Scaling up expressive human pose and shape estimation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 3

[31] Jeff Tan, Donglai Xiang, Shubham Tulsiani, Deva Ramanan, and Gengshan Yang. Dressrecon: Freeform 4d human reconstruction from monocular video. In *2025 International Conference on 3D Vision (3DV)*, pages 250–260. IEEE, 2025. 3

[32] Boyi Jiang, Yang Hong, Hujun Bao, and Juyong Zhang. SelfRecon: Self reconstruction your digital avatar from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5605–5615, 2022. 3

[33] Peng Li, Wangguandong Zheng, Yuan Liu, Tao Yu, Yangguang Li, Xingqun Qi, Xiaowei Chi, Siyu Xia, Yan-Pei Cao, Wei Xue, et al. Pshuman: Photorealistic single-image 3d human reconstruction using cross-scale multiview diffusion and explicit remeshing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16008–16018, 2025. 3

[34] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. PIFu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2304–2314, 2019. 3

[35] Navami Kairanda, Marc Habermann, Shanthika Naik, Christian Theobalt, and Vladislav Golyanik. Thin-shell-sft: Fine-grained monocular non-rigid 3d surface tracking with neural deformation fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11373–11383, 2025. 3

[36] Hang Ye, Xiaoxuan Ma, Hai Ci, Wentao Zhu, and Yizhou Wang. Freecloth: Free-form generation enhances challenging clothed human modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15987–15997, 2025. 3

[37] Zachary Teed and Jia Deng. Raft-3d: Scene flow using rigid-motion embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8375–8384, 2021. 3

[38] Wenxuan Wu, Zhi Yuan Wang, Zhuwen Li, Wei Liu, and Li Fuxin. Pointpwc-net: Cost volume on point clouds for (self-) supervised scene flow estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 88–107. Springer, 2020. 3

[39] Junhwa Hur and Stefan Roth. Self-supervised monocular scene flow estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7396–7405, 2020. 3, 8, 9

[40] Xueqian Li, Jhony Kaesemodel Pontes, and Simon Lucey. Neural scene flow prior. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 3

[41] Kyle Vedder, Neehar Peri, Ishan Khatri, Siyi Li, Eric Eaton, Mehmet Kocamaz, Yue Wang, Zhiding Yu, Deva Ramanan, and Joachim Pehserl. Neural Eulerian scene flow fields. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025. 3

[42] Ruibo Li, Chi Zhang, Guosheng Lin, Zhe Wang, and Chunhua Shen. Rigidflow: Self-supervised scene flow learning on point clouds by local rigidity prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16959–16968, 2022. 3

[43] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUSt3R: Geometric 3D vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20697–20709, 2024. 3, 22

[44] Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. MonST3R: A simple approach for estimating geometry in the presence of motion. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

[45] Jiahao Lu, Tianyu Huang, Peng Li, Zhiyang Dou, Cheng Lin, Zhiming Cui, Zhen Dong, Sai-Kit Yeung, Wenping Wang, and Yuan Liu. Align3R: Aligned monocular depth estimation for dynamic videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 3, 9

[46] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9660–9672, 2025. 3

[47] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8503–8513, 2025. 3, 8, 9

[48] Yuxi Xiao, Jianyuan Wang, Nan Xue, Nikita Karaev, Iurii Makarov, Bingyi Kang, Xin Zhu, Hujun Bao, Yujun Shen, and Xiaowei Zhou. SpatialTrackerV2: 3D point tracking made easy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. 3, 8, 9, 22

[49] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7382–7392, 2026. 3, 22

[50] Himangi Mittal, Brian Okorn, and David Held. Just go with the flow: Self-supervised scene flow estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11177–11185, 2020. 3

[51] Kyle Vedder, Neehar Peri, Nathaniel Chodosh, Ishan Khatri, Eric Eaton, Dinesh Jayaraman, Yang Liu, Deva Ramanan, and James Hays. ZeroFlow: Scalable scene flow via distillation. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024. 3

[52] Chung-Yi Weng, Brian Curless, Pratul P. Srinivasan, Jonathan T. Barron, and Ira Kemelmacher-Shlizerman. HumanNeRF: Free-viewpoint rendering of moving people from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16210–16220, 2022. 3

[53] Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. InstantAvatar: Learning avatars from monocular video in 60 seconds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16922–16932, 2023. 3

[54] Chen Guo, Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. Vid2Avatar: 3D avatar reconstruction from videos in the wild via self-supervised scene decomposition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12858–12868, 2023. 3

[55] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 21–38. Springer, 2024. 3

[56] Zhiyin Qian, Shaofei Wang, Marko Mihajlovic, Andreas Geiger, and Siyu Tang. 3DGS-Avatar: Animatable avatars via deformable 3D gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5020–5030, 2024. 3

[57] Muhammed Kocabas, Jen-Hao Rick Chang, James Gabriel, Oncel Tuzel, and Anurag Ranjan. HUGS: Human gaussian splats. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 505–515, 2024.

[58] Chen Guo, Junxuan Li, Yash Kant, Yaser Sheikh, Shunsuke Saito, and Chen Cao. Vid2Avatar-Pro: Authentic avatar from videos in the wild via universal prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 3

[59] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3D human pose estimation in the wild using improved CNN supervision. In *International Conference on 3D Vision (3DV)*, pages 506–516. IEEE, 2017. 3

[60] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. NTU RGB+D: A large scale dataset for 3D human activity analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1010–1019, 2016. 3

[61] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C. Kot. NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 42(10):2684–2701, 2020.

[62] Leonid Sigal, Alexandru O. Balan, and Michael J. Black. HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International Journal of Computer Vision (IJCV)*, 87(1–2):4–27, 2010. 3

[63] Mihai Fieraru, Mihai Zanfir, Silviu-Cristian Pirlea, Vlad Olaru, and Cristian Sminchisescu. AIFit: Automatic 3d human-interpretable feedback models for fitness training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9919–9928, 2021. 3, 7, 8, 22

[64] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3061–3070, 2015. 3, 7

[65] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, Deva Ramanan, Peter Carr, and James Hays. Argoverse 2: Next generation datasets for self-driving perception and forecasting. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 3

[66] Yang Li, Hikari Takehara, Takafumi Taketomi, Bo Zheng, and Matthias Nießner. 4DComplete: Non-rigid motion estimation beyond the observable surface. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12686–12696, 2021. 3, 7

[67] Zhongang Cai, Daxuan Ren, Ailing Zeng, Zhengyu Lin, Tao Yu, Wenjia Wang, Xiangyu Fan, Yang Gao, Yifan Yu, Liang Pan, Fangzhou Hong, Mingyuan Zhang, Chen Change Loy, Lei Yang, and Ziwei Liu. HuMMan: Multi-modal 4D human dataset for versatile sensing and modeling. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 557–577, 2022. 3

[68] Zhixuan Yu, Jae Shin Yoon, In Kyu Lee, Prashanth Venkatesh, Jaesik Park, Jihun Yu, and Hyun Soo Park. HUMBI: A large multiview dataset of human body expressions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2987–2997, 2020. 3

[69] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4627–4635, 2017. 3

[70] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13468–13478, 2021. 3, 7, 21

[71] Qingzheng Xu, Ru Cao, Xin Shen, Heming Du, Sen Wang, and Xin Yu. M3GYM: A large-scale multimodal multi-view multi-person pose dataset for fitness activity understanding in real-world settings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12289–12300, 2025. 3, 7

[72] Michael J. Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. BEDLAM: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8726–8737, 2023. 3, 7, 21, 22

[73] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 550–560, 2024. 3

[74] Qianli Ma, Jinlong Yang, Anurag Ranjan, Sergi Pujades, Gerard Pons-Moll, Siyu Tang, and Michael J. Black. Learning to dress 3D people in generative clothing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6468–6477, 2020. 3

[75] Hugo Bertiche, Meysam Madadi, and Sergio Escalera. CLOTH3D: Clothed 3D humans. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 344–359, 2020.

[76] Xingxing Zou, Xintong Han, and Waikeung Wong. CLOTH4D: A dataset for clothed human reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12847–12857, 2023. 3

[77] Oriane Siméoni et al. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025. 4, 18

[78] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12179–12188, 2021. 4, 18

[79] Tamar Flash and Neville Hogan. The coordination of arm movements: an experimentally confirmed mathematical model. *Journal of Neuroscience*, 5(7):1688–1703, 1985. 6

[80] Neville Hogan. An organizing principle for a class of voluntary movements. *Journal of Neuroscience*, 4(11):2745–2754, 1984. 6

[81] Yoji Uno, Mitsuo Kawato, and Ryoji Suzuki. Formation and control of optimal trajectory in human multijoint arm movement: minimum torque-change model. *Biological Cybernetics*, 61(2):89–101, 1989. 6

[82] Emanuel Todorov and Michael I. Jordan. Optimal feedback control as a theory of motor coordination. *Nature Neuroscience*, 5(11):1226–1235, 2002. 6

[83] Adobe Systems Incorporated. Mixamo, 2024. https://www.mixamo.com. 7

[84] Zhanbo Huang, Xiaoming Liu, and Yu Kong. H-MoRe: Learning human-centric motion representation for action analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 8

[85] Zachary Teed and Jia Deng. RAFT: Recurrent all-pairs field transforms for optical flow. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 8, 9, 22

[86] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the Kinetics dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6299–6308, 2017. 8

[87] Siyuan Yang, Lu Zhang, Yu Liu, Zhizhuo Jiang, and You He. Video diffusion models with local-global context guidance. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1640–1648, 2023. 8

[88] Chen Chen, Roozbeh Jafari, and Nasser Kehtarnavaz. UTD-MHAD: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pages 168–172, 2015. 8

[89] Zhaoyang Huang, Xiaoyu Shi, Chao Zhang, Qiang Wang, Ka Chun Cheung, Hongwei Qin, Jifeng Dai, and Hongsheng Li. FlowFormer: A transformer architecture for optical flow. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 9

[90] Hongwei Zheng, Han Li, Wenrui Dai, Ziyang Zheng, Chenglin Li, Junni Zou, and Hongkai Xiong. HiPART: Hierarchical pose autoregressive transformer for occluded 3D human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16807–16817, 2025. 9, 10

[91] Soroush Mehraban, Vida Adeli, and Babak Taati. MotionAGFormer: Enhancing 3D human pose estimation with a transformer-GCNFormer network. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 6920–6930, 2024. 9, 10

[92] Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. MotionBERT: A unified perspective on learning human motion representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 10

[93] Priyanka Patel and Michael J. Black. CameraHMR: Aligning people with perspective. In *International Conference on 3D Vision (3DV)*, 2025. 10, 22

[94] Yufu Wang, Yu Sun, Priyanka Patel, Kostas Daniilidis, Michael J. Black, and Muhammed Kocabas. PromptHMR: Promptable human mesh recovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1148–1159, 2025. 10, 22

[95] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 10, 22

[96] Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. MoGe-2: Accurate monocular geometry with metric scale and sharp details. *arXiv preprint arXiv:2507.02546*, 2025. 10, 22

[97] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9492–9502, 2024. 10, 22

[98] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. ZoeDepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023. 10, 22

[99] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5745–5753, 2019. 18

[100] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4015–4026, 2023. 19

[101] Paolo De Leva. Adjustments to Zatsiorsky-Seluyanov's segment inertia parameters. *Journal of Biomechanics*, 29(9):1223–1230, 1996. 20

[102] Reiner Birkl, Diana Wofk, and Matthias Müller. MiDaS v3.1 – a model zoo for robust monocular relative depth estimation. *arXiv preprint arXiv:2307.14460*, 2023. 22

[103] Haotong Lin, Sili Chen, Junhao Liew, Donny Y. Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2026. 22

[104] Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Permutation-equivariant visual geometry learning. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2026. 22

[105] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3D with MASt3R. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 22

[106] 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 *ACM Transactions on Graphics (TOG) – Proceedings of ACM SIGGRAPH Asia*, 2024. 22

[107] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 22

[108] Shiqi Yu, Daoliang Tan, and Tieniu Tan. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*, volume 4, pages 441–444, 2006. 22, 24

[109] Chunhui Liu, Yueyu Hu, Yanghao Li, Sijie Song, and Jiaying Liu. PKU-MMD: A large scale benchmark for continuous multi-modal human action understanding. *arXiv preprint arXiv:1703.07475*, 2017. 22, 24

[110] Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Qingkun Su, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. Champ: Controllable and consistent human image animation with 3D parametric guidance. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 23, 24

A Network Architecture

This appendix section expands on the architecture introduced in Sec. 3.1. We first detail the camera parameterization that Sec. 3.1 packages as $\mathbf{C}$. We then describe the backbone that realizes G_θ and the decoder heads that emit the outputs of $\mathcal{Y}$. Last, we cover the optional conditioning pathways for known camera parameters and user prompts.

A.1 Camera Parameterization

The per-frame camera variable $\mathbf{C}_i = (\mathbf{K}_i, \mathbf{R}_i, \mathbf{t}_i)$ packages intrinsics, rotation, and translation. The network emits $\mathbf{C}_i$ independently per frame as a 13-dimensional vector. Six dimensions parameterize rotation $\mathbf{R}_i$ via the continuous 6D representation [99], three dimensions encode translation $\mathbf{t}_i$, and four dimensions encode normalized intrinsics $(f_x/W,\ f_y/H,\ c_x/W,\ c_y/H)$ for $\mathbf{K}_i$.

The 6D representation avoids the antipodal ambiguity of quaternions and the discontinuities of Euler angles. A Gram-Schmidt step recovers a valid rotation matrix from the regressed six values. Normalizing f_x, c_x by image width W and f_y, c_y by image height H keeps the parameterization scale-invariant across resolutions, so training data of different image sizes mixes without rescaling.

We anchor the world coordinate frame to the first frame, fixing $\mathbf{R}_0 = \mathbf{I}$ and $\mathbf{t}_0 = \mathbf{0}$. The remaining $(\mathbf{R}_i, \mathbf{t}_i)$ then describe relative camera motion with respect to the first frame. This convention defines a consistent world coordinate frame even in the monocular setting where global scale and orientation are otherwise unobservable.

A single physical camera produces every frame in the clip, yet we do not hard-tie $\mathbf{K}_i$ across frames. The network outputs $\mathbf{K}_i$ independently per frame, and the camera-consistency term in Sec. B.2 softly drives them to agree. The soft formulation accommodates digital zoom or cropping, where intrinsics drift over time. When known camera parameters are available, the conditioning pathway in Sec. A.4 consumes them directly in place of these outputs.

A.2 Backbone and Trunk

We adopt DINOv3 ViT-L/16 [77] pretrained on LVD-1689M as our visual backbone and keep it frozen throughout training. The frozen backbone preserves general-purpose visual priors while concentrating trainable capacity on the trunk and downstream heads. Each input frame is resized to 512×512 and tokenized into a 32×32 patch grid, producing 1024 patch tokens per frame at hidden dimension $D = 1024$. Backbone forward passes run in bf16 precision.

A lightweight projector sits between the frozen backbone and the trunk. It applies 2 spatial-adaptation layers to patch tokens alone, performing intra-frame attention to map frozen features into the trunk input space. This module adapts the general-purpose representation toward downstream geometric tasks without unfreezing the backbone.

The trunk consists of 16 joint query-patch attention layers (Fig. 2). Pose queries and camera queries enter the attention alongside patch tokens at every layer, so queries interact with dense features throughout the stack rather than at a single final readout. Each trunk layer factorizes attention into a frame attention sub-module and a temporal attention sub-module. Frame attention performs intra-frame spatial attention to capture per-frame structure. Temporal attention aggregates across frames, applying 1D temporal RoPE on Q and K to encode frame indices and capture temporal dynamics. All trunk layers use pre-norm LayerNorm with 16 attention heads, and we run SDPA kernels with FlashAttention-2 on H100 GPUs.

For a clip of length T, the trunk processes $T \cdot (1024 + J + 1) + N_{\text{cond}}$ tokens in total. Here $J = 24$ joint queries follow the SMPL skeleton convention, and the camera query contributes one token per frame. The optional conditioning tokens N_{cond} are detailed in Sec. A.4.

A.3 Decoder Heads

The dense heads emit depth $\mathbf{D}$, flow $\mathbf{F}$, and mask $\mathbf{M}$ as pixel-level outputs. We instantiate three independent DPT-style [78] multi-scale fusion modules, one per modality. Each head taps four trunk layers at $\{4, 8, 12, 16\}$ for uniform $1/4$ spacing and fuses them at 256-d feature width. It then bilinearly upsamples from the 32×32 patch grid to the original 512×512 resolution. Output

activations differ by physical range: softplus on $\mathbf{D}$ enforces positive depth, identity on $\mathbf{F}$ outputs a signed 3D vector field, and sigmoid on $\mathbf{M}$ yields soft foreground probability.

The sparse heads emit pose $\mathbf{P}$ and camera $\mathbf{C}$ from the final-layer query states. Each pose query corresponds to one of $J = 24$ skeletal joints, and a 3-layer MLP maps its final state to $\mathbb{R}^3$ joint coordinates. The per-frame camera query passes through a 3-layer MLP that emits the 13-dimensional vector parsed under Sec. A.1's convention. Joint angles required by the minimum-effort-path constraint are derived from these joint coordinates via inverse kinematics, detailed in Sec. B.1.

The three dense heads maintain independent fusion features rather than sharing a single tower. This isolates per-modality gradients within each head, preventing high-frequency texture cues that benefit one modality from polluting the geometric structure required by another.

A.4 Optional Conditioning

When ground-truth camera parameters $(\mathbf{K}, \mathbf{R}, \mathbf{t})$ are available, the system consumes them in place of the camera head's output. Known parameters are packaged into the 13-d format of Sec. A.1, encoded by a 3-layer MLP into a single token, and added to the corresponding camera query state for that frame. The additive fusion anchors the model to accurate camera geometry when prior knowledge exists and falls back to self-prediction otherwise.

A user can optionally specify the subject of interest through 2D points and a bounding box. We follow the SAM [100] prompt encoder design, with up to 5 points and one 2-corner box per frame. Fixed random Fourier features encode their positions. Learnable label embeddings encode their types (foreground point, background point, box corner). All prompt tokens enter the trunk attention as additional tokens, interacting with patch, pose, and camera queries on equal footing.

We randomly drop conditioning inputs during training following the augmentation recipe of SAM [100]. The camera condition is replaced by a learnable null token with probability 0.5. Each prompt point slot is dropped independently with probability 0.3, and the box is dropped with probability 0.5. The model thus encounters a wide range of conditioning combinations during training, producing valid outputs at inference whether the user supplies full conditioning, partial prompts, or no conditioning at all.

B Self-supervised Learning

This appendix section expands on the four physical constraints and auxiliary anchoring terms introduced in Sec. 3.2. We first give the implementation forms of tolerances, anatomical constants, support-polygon construction, and inverse/forward kinematics (Sec. B.1). We then specify the external anchoring and camera-consistency terms (Sec. B.2), and close with a coupling footprint that summarizes how the objective set binds all five output modalities (Sec. B.3).

B.1 Loss Implementation Details

This subsection specifies the implementation forms of the four priors in Sec. 3.2. It covers tolerance values, anatomical constants, support-polygon construction, and the kinematics conventions of the minimum-effort-path constraint.

Silhouette Edge Alignment. The signed distance field $\mathrm{SDF}_{\partial\mathbf{M}}$ in Eq. (4) grows unboundedly away from the boundary and would otherwise dominate the gradient sum at distant pixels. We saturate its absolute value at $\tau = 32$ pixels via $|\widetilde{\mathrm{SDF}}_{\partial\mathbf{M}}| = \min(|\mathrm{SDF}_{\partial\mathbf{M}}|, \tau)$. This concentrates the constraint on the transition band near the boundary rather than the far field. The spatial gradients $\nabla\mathbf{D}$ and $\nabla\mathbf{F}$ are computed by a Sobel 3×3 operator over bf16 features.

Skeletal-Surface Coupling. The tolerance follows a piecewise-linear form $\rho(\mathbf{u}) = \rho_{\min} + \alpha \cdot d_{\text{bone}}(\mathbf{u})$, where $d_{\text{bone}}(\mathbf{u})$ is the 3D distance from the unprojected surface point to its nearest bone segment. We set $\rho_{\min} = 1$ cm as the on-skin floor and $\alpha = 0.5$ as the dimensionless growth rate. On-skin pixels thus stay tight to the skeleton, while loose garments such as a skirt receive a wider margin. Bone assignment $b^* = (j_1^*, j_2^*)$ and barycentric position $\alpha^* \in [0, 1]$ are obtained via $\arg\min$ over

bone segments, which is non-differentiable. We carry stop-gradient through the selection, allowing gradients only through the flow residual and the tolerance term.

Center-of-Mass Support. Anatomical mass distribution constants $\{w_j\}_{j=1}^{24}$ follow the segmental inertia table of De Leva [101], normalized so that $\sum_j w_j = 1$. The contact weight $c_j = \exp(-d_j^2/\sigma^2)$ uses end-effector height d_j above a fitted ground plane with $\sigma = 5$ cm. The ground plane is recovered from $\mathbf{D}$ by RANSAC over the lowest non-foreground point cloud, restricted to lie below the foreground bounding box. Support polygon $\mathcal{S}(\mathbf{P}, \mathbf{D})$ is the convex hull of the end-effector ground projections $\{\mathbf{P}_{j,xy}\}$ weighted by c_j, with the SDF evaluated on the weighted polygon boundary $\partial\mathcal{S}$. The mask $\mathbf{M}$ further restricts candidate contact joints to the foreground, preventing spurious contacts from background limbs.

Minimum Effort Path. Inverse kinematics traverse the kinematic chain from the pelvis root to the leaves. They encode each parent-child rotation in local axis-angle form. Bone roll is fixed to zero, which resolves the rotation ambiguity along the bone axis. Forward kinematics apply the standard parent-to-child rotation chain. The temporal window W is sampled per training iteration from the clip index set $\mathcal{T}$ with random length $|W| \in [W_{\min}, W_{\max}]$, where $W_{\min} = 8$ frames and $W_{\max} = 32$ frames. The trajectory tolerance is $\rho_{\text{eff}} = 5$ cm, beyond which deviations from the minimum-jerk reference incur penalty.

B.2 External Anchoring and Camera Consistency

This subsection specifies the external anchor and the camera consistency term mentioned at the end of Sec. 3.2.

External Anchor. The four physical priors in Sec. 3.2 are scale-invariant by construction and cannot pin down the global metric scale of the network's outputs. We anchor depth $\mathbf{D}$ and pose $\mathbf{P}$ against two off-the-shelf predictors, Depth Pro [17] for $\mathbf{D}_{\text{ref}}$ and SAM 3D Body [8] for $\mathbf{P}_{\text{ref}}$, through a margin-based hinge:

$$\mathcal{C}_{\text{anch}}(\mathbf{D}, \mathbf{P}) = \frac{1}{|\Omega|} \sum_{\mathbf{u}} \left[\, |\mathbf{D}(\mathbf{u}) - \mathbf{D}_{\text{ref}}(\mathbf{u})| - \rho_{\mathbf{D}} \,\right]_+ + \frac{1}{J} \sum_{j=1}^{J} \left[\, \|\mathbf{P}_j - \mathbf{P}_{\text{ref},j}\|_2 - \rho_{\mathbf{P}} \,\right]_+, \tag{13}$$

with $\rho_{\mathbf{D}} = 10$ cm and $\rho_{\mathbf{P}} = 3$ cm. Reference predictions are noisy, and a strict regression target would collapse H-Flow into a predictor replica. The margin instead activates the anchor only when H-Flow drifts beyond a tolerance band, leaving the physical priors free to optimize within that band. Early in training, this anchor guides coarse body geometry. Later on, it fixes metric scale and blocks trivial collapse toward a constant zero solution. In the last row of Tab. 5, we replace these external predictors with coarse masks and poses reconstructed from DINOv3 features alongside a photometric warp loss.

Camera Consistency. A single physical camera produces every frame in the clip, yet Sec. A.1 predicts $\mathbf{K}_i$ independently per frame to preserve flexibility for digital zoom or cropping. A consistency term softly drives the per-frame intrinsics to agree:

$$\mathcal{C}_{\text{cam}}(\mathbf{C}) = \frac{1}{|\mathcal{T}|^2} \sum_{i,j \in \mathcal{T}} \|\mathbf{K}_i - \mathbf{K}_j\|_2^2, \tag{14}$$

where $\mathbf{K}_i$ takes the 4-d normalized form of Sec. A.1 so that the squared norm is scale-aligned across resolutions. Per-frame rotations and translations $(\mathbf{R}_i, \mathbf{t}_i)$ are unconstrained, since they describe inter-frame camera motion by design.

B.3 Cross-Modal Coupling Footprint

Every core output head ($\mathbf{D}$, $\mathbf{F}$, $\mathbf{P}$) participates in at least two distinct physical constraints (Tab. 6). The structural terms ($\mathcal{C}_{\text{skel}}, \mathcal{C}_{\text{CoM}}$) further link pixel unprojection through camera $\mathbf{C}$ and mask $\mathbf{M}$ to 3D skeletal joints $\mathbf{P}$. $\mathcal{C}_{\text{anch}}$ leaves a tolerance band open. Within it, these overlapping constraints reconcile any geometric discrepancy between depth, pose, and flow during training. All five heads thus inhabit one mutually consistent 3D coordinate frame at inference.

Table 6: **Cross-modal coupling footprint across optimization terms.** Checkmarks (✓) denote output modalities directly entering each objective term, and dashes (—) indicate no direct dependence.

Term	Category	D	F	M	P	C	Primary Optimization Role
$\mathcal{C}_{\text{silh}}$	Geometric	✓	✓	✓	—	—	Aligns dense depth and flow discontinuities to silhouette boundary
$\mathcal{C}_{\text{skel}}$	Structural	✓	✓	✓	✓	✓	Binds unprojected 3D surface flow to nearest bone within cloth margin
$\mathcal{C}_{\text{CoM}}$	Structural	✓	—	✓	✓	✓	Constrains projected mass center inside ground-contact polygon
$\mathcal{C}_{\text{eff}}$	Biomechanical	—	—	—	✓	—	Regularizes multi-frame joint trajectories along minimum-jerk arc
$\mathcal{C}_{\text{anch}}$	Auxiliary	✓	—	—	✓	—	Fixes global metric scale via margin-based hinge against references
$\mathcal{C}_{\text{cam}}$	Auxiliary	—	—	—	—	✓	Softly aligns normalized camera intrinsics across clip frames

C Implementation Details

Training Setup. We train H-Flow on 6 NVIDIA H100 80GB GPUs with mixed precision (bf16 forward, fp32 master weights) for 4×10^5 iterations. The optimizer is AdamW with base learning rate 4×10^{-5}, weight decay 0.05, and a cosine schedule with linear warmup over the first 5K iterations. The effective batch size is 12 clips, distributed as 2 clips per GPU across 6 GPUs, with each clip containing $T = 16$ frames. Gradient clipping is applied with threshold 1.0.

Data Pipeline. Pretraining draws from all seven Human3.6M subjects. Each mini-batch randomly samples 16 consecutive frames at a stride drawn uniformly from a small range. The model thus sees a spread of inter-frame intervals rather than a single fixed frame rate. This dynamic temporal sampling lets H-Flow generalize across datasets with different native frame rates without retraining. Augmentations include foreground-centered random crops, horizontal flipping, and color jitter. For in-domain fine-tuning on Fit3D and 3DPW, we follow the splits described in Sec. 5 and reuse the pretraining hyperparameters with a shorter 1×10^5 iteration schedule.

Inference. At test time, H-Flow accepts video clips of arbitrary length. Clips at or below the training length pass through the network in a single forward call. Longer clips are processed via a sliding window of 16 frames with 4-frame overlap between adjacent windows, and predictions in the overlap region are averaged across windows. All output heads are produced in a single forward pass, with no test-time optimization or runtime calibration.

D Extended Benchmark Evaluations

This section expands the quantitative evaluations in Secs. 5.1 and 5.2, covering fine-grained scene flow breakdowns (Sec. D.1) and extended companion pose and depth results (Sec. D.2).

D.1 Fine-grained Scene Flow Analysis

Aggregate EPE averages across distinct biomechanical regimes (Tab. 7). High-velocity motion (> 100 mm/frame) and large-excursion limbs such as the forearm and upper arm combine rapid joint acceleration with secondary garment swing. Physical coupling keeps 3D flow coherent under these non-rigid dynamics, holding EPE to 62.2 mm above 100 mm/frame and 46.4 mm on the forearm. On segments governed almost purely by rigid bone rotation with minimal soft-tissue or cloth offset, such as the head and shin, rigid surface templates already capture most displacement. Residual errors concentrate at distal extremities (hands and feet), driven by self-occlusion during limb–torso contact and inertial lag along loose hems.

D.2 Extended Companion Modality Evaluation

To examine scaling with unlabeled video volume (Tab. 8a), we expand H-Flow's pretraining corpus across BEDLAM [72], AGORA [70], 4DHumans [7], and 3DPW (H-Flow*). This brings its companion pose head into the accuracy regime of dedicated mesh-recovery pipelines. It requires neither parametric body templates nor 3D pose labels. For companion depth (Tab. 8b), concentrating capacity on the SMPL-X human surface yields consistent gains over generic full-scene predictors across both absolute metric error and scale-invariant structure (δ_1 reaching 0.971 on 3DPW and 0.944 on Fit3D).

Table 7: **Fine-grained scene flow breakdown by motion speed and anatomical region.** 3D EPE (mm) over five displacement intervals (left) and eight body segments (right). Best in **bold**, second-best *underlined*.

Methods	Scene Flow EPE by Motion Speed (mm/frame)↓					Scene Flow EPE by Anatomical Region (mm)↓							
	5–10	10–25	25–50	50–100	>100	Torso	Head	U. Arm	F. Arm	Hand	Thigh	Shin	Foot
H-Flow (Ours)	**11.5**	**15.6**	**23.3**	**35.1**	**62.2**	**26.4**	43.4	**34.4**	**46.4**	**82.9**	**26.9**	31.7	**63.0**
SAM 3D Body [8]	19.9	26.4	33.9	54.0	118.2	30.7	**39.2**	43.7	66.3	88.3	29.1	**31.6**	64.8
Depth Pro + RAFT [17, 85]	62.2	67.6	77.2	97.0	152.1	75.7	82.1	74.1	83.9	111.0	74.7	80.8	113.3

Table 8: **Extended companion pose and depth evaluation.** *(a)* Pose error (mm) on unfiltered 3DPW (‡evaluated from released checkpoint; *expanded unlabeled pretraining). *(b)* Body-region depth evaluation on 3DPW and Fit3D (†affine-invariant output, including MiDaS v3.1 [102]). Best in **bold**, second-best *underlined*.

Methods	3DPW (Unfiltered)	
	MPJPE↓	PA↓
H-Flow* (Ours)	57.7	38.0
SAM 3D Body [8]	**54.8**	**33.2**
SAM 3D Body‡ [8]	60.1	41.1
CameraHMR [93]	56.0	35.1
PromptHMR [94]	58.7	36.1

Methods	3DPW [22]				Fit3D [63]			
	MAE↓	AbsRel↓	SiLog↓	δ_1↑	MAE↓	AbsRel↓	SiLog↓	δ_1↑
H-Flow (Ours)	**158.6**	**0.055**	**7.2**	**0.971**	**124.5**	**0.024**	**2.9**	**0.944**
Depth Pro [17]	527.7	0.181	12.4	0.573	234.3	0.052	4.4	0.882
Depth Anything V2 [95]	485.3	0.172	11.8	0.612	218.7	0.048	4.1	0.903
MoGe-2 [96]	452.1	0.158	11.3	0.641	205.4	0.045	3.9	0.918
Marigold [97]	498.2	0.176	12.0	0.598	225.6	0.050	4.2	0.891
ZoeDepth [98]	512.8	0.179	12.2	0.585	228.9	0.051	4.3	0.886
MiDaS v3.1† [102]	—	—	13.1	—	—	—	4.7	—

E Cross-Modal Consistency Analysis

Multi-modal perception and synthesis require geometric channels that agree within a single 3D space. H-Flow decodes depth $\mathbf{D}$, pose $\mathbf{P}$, and scene flow $\mathbf{F}$ jointly from one shared representation under physical coupling (Sec. B.3), aligning all outputs in one coordinate frame. Independent single-modality specialists lack this shared frame. Below we quantify inter-method disagreement across off-the-shelf specialists on identical clips (Sec. E.1) and measure how cross-modal consistency shapes downstream multi-modal tasks (Sec. E.2).

E.1 Disagreement Across Specialist Estimators

Fig. 5 diagnoses inter-method inconsistency across separate single-modality specialists on tight-attire studio clips [63] and simulated loose clothing [72]. H-Flow jointly predicts $\mathbf{D}$, $\mathbf{P}$, and $\mathbf{F}$ inside a single shared 3D frame without cross-model coordinate drift, leaving only physical surface-to-bone offsets ($\sim$12 mm scale). Pairing H-Flow with an external specialist would likewise reflect only that specialist's uncalibrated scale convention, so H-Flow is omitted from this cross-specialist grid. We examine four off-the-shelf families. Monocular metric depth predictors include Depth Pro [17], MoGe-2 [96], and Depth Anything 3 [103]. Multi-view pointmap models include π^3 [104], MASt3R [105], and DUSt3R [43]. Parametric mesh estimators include SAM 3D Body [8], GVHMR [106], CameraHMR [93], and PromptHMR [94]. 4D trackers include VGGT [107], SpatialTrackerV2 [48], and D4RT [49], supplemented by depth- and pointmap-lifted 2D flows in Fig. 5c,d.

Even on studio Fit3D clips, spatial depth offsets reach 1.1%–12.1% (Fig. 5a). Under clothed motion on BEDLAM2 (Fig. 5b), disagreement rises sharply across nearly all pairs: Depth Pro and SAM 3D Body differ by 5.2%, and cross-family combinations reach 10.5%–37.2%. Pairwise 3D motion magnitude similarly diverges by 21–363 mm on Fit3D and 31–1067 mm on BEDLAM2 (Fig. 5c,d). In particular, SAM 3D Body and Depth Pro-lifted flow disagree by 91 mm on Fit3D and 188 mm on BEDLAM2. Each specialist relies on an isolated scale convention: HMR uses bone-length priors, monocular depth uses scene-level depth priors, and pointmaps use pairwise rigid alignment. Combining separate specialists therefore produces contradictory 3D states with no shared physical frame to arbitrate between them.

E.2 Multi-Modal Downstream Impact

We next test how cross-modal consistency affects downstream tasks that jointly consume depth $\mathbf{D}$ and pose $\mathbf{P}$ (Fig. 6 and Tab. 9). We evaluate three benchmarks. The first is multi-modal gait recognition on CASIA-B [108] with a lightweight depth-pose fusion network (DPLite) across normal (NM), bag (BG), and coat (CL) conditions. The second is 3D action recognition on PKU-MMD [109]

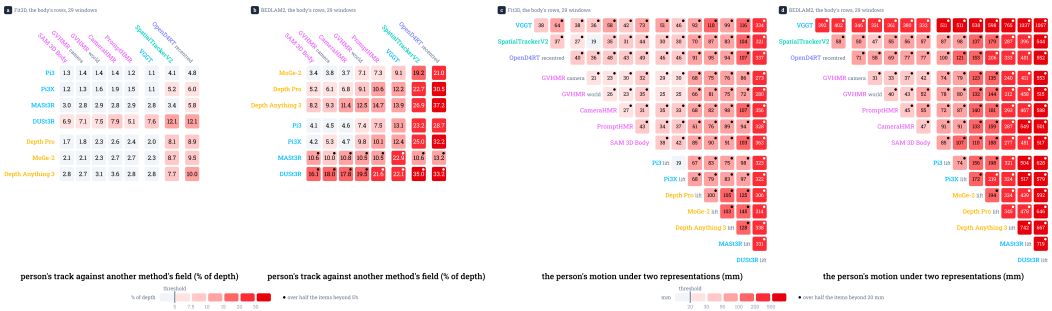

Figure 5: **Inter-method geometric and kinematic disagreement across off-the-shelf specialists.** Evaluated over 29 windows per dataset on Fit3D (*a, c*) and BEDLAM2 (*b, d*). *(a, b)* Spatial depth discrepancy (% of subject depth) projecting 3D tracks (columns) into dense depth/pointmap fields (rows); dots mark $> 5\%$ in $> 50\%$ of windows. *(c, d)* Upper-triangular pairwise 3D motion discrepancy (mm) across 15 representations, where columns follow the top-to-bottom diagonal label order (*e.g.*, top-left 38 mm in *c* compares VGGT and SpatialTrackerV2); dots mark > 20 mm in $> 50\%$ of windows.

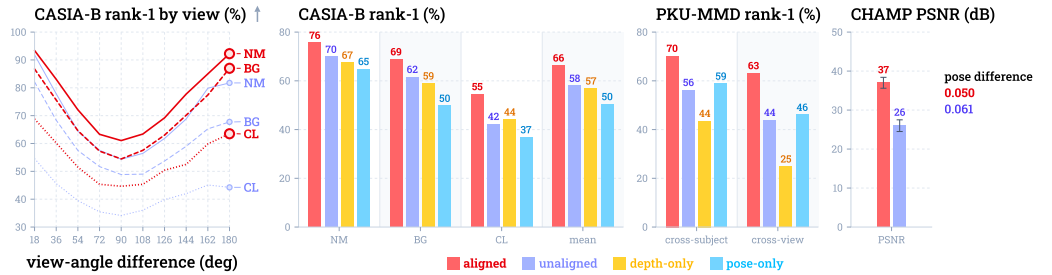

Figure 6: **Downstream multi-modal perception and synthesis under aligned vs. unaligned inputs.** *Left:* CASIA-B Rank-1 accuracy (%) across camera view-angle differences |probe − gallery| from 18° (near-same view) through 90° (orthogonal trough) to 180° (opposite front-vs.-back views, marked by circles), under NM (solid), BG (dashed), and CL (dotted). *Center-left to right:* CASIA-B condition means, PKU-MMD Rank-1 accuracy, and CHAMP PSNR (error bars show standard deviation) with pose difference.

under cross-subject and cross-view splits. The third is controllable human video animation with CHAMP [110]. *Aligned* feeds jointly consistent $(\mathbf{D}, \mathbf{P})$ from H-Flow. *Unaligned* pairs Depth Pro and SAM 3D Body directly, using the same two predictors that provide loose scale references during training, alongside single-modality *Depth-only* and *Pose-only* baselines.

Three findings stand out. First, cross-modal conflict triggers destructive fusion below single-modality performance. On PKU-MMD, adding unaligned Depth Pro maps to SAM 3D Body poses lowers accuracy below *Pose-only* on both splits (56.2% vs. 58.8% cross-subject and 43.9% vs. 46.0% cross-view). Under loose coats (CL) on CASIA-B, where garment surfaces depart furthest from the skeleton, *Unaligned* (42.3%) similarly drops below *Depth-only* (44.1%). Jointly consistent $(\mathbf{D}, \mathbf{P})$ from H-Flow turns this conflict into substantial multi-modal synergy, reaching 70.2% (+14.0) and 63.1% (+19.2) on PKU-MMD and lifting CASIA-B mean accuracy from 57.9% to 66.3% (+12.2 on CL). Second, consistency governs 3D cross-view symmetry (Fig. 6, *left*). At 18°, both cameras view the body from nearly the same side, so systematic offsets between Depth Pro and SAM 3D Body point in the same direction and largely cancel out (+1.6 points on NM and +4.7 on BG). Approaching 180°, camera-ray depth offsets invert in body-centric 3D space and compound rather than cancel: *Unaligned* recovery stalls (CL drops from 45.1% at 162° to 44.3% at 180°). H-Flow's aligned 3D geometry preserves front-back symmetry and recovers near-18° accuracy (92.3% vs. 93.4% on NM and 87.0% vs. 86.8% on BG), widening the 180° margin to +10.5 points on NM and +19.3 points on both BG and CL. Third, spatial alignment prevents conditioning conflict in synthesis. On CHAMP, co-registered depth and skeletal guidance improves PSNR by 11.0 dB (37.0 vs. 26.0 dB) and reduces pose error from 0.061 to 0.050.

Table 9: **Quantitative downstream comparison of consistent vs. conflicting multi-modal inputs.** Multi-seed mean Rank-1 accuracy (%) on CASIA-B and PKU-MMD, alongside synthesis PSNR (dB) and pose error on CHAMP. Best in **bold**, second-best *underlined*.

Input Representation	Modalities	CASIA-B Gait [108] (Rank-1 %)↑				PKU-MMD Action [109] (%)↑		CHAMP Synthesis [110]	
		NM	BG	CL	Mean	X-Sub	X-View	PSNR↑ dB	Pose Diff.↓
H-Flow (Ours) [Aligned]	**D + P**	**75.6**	**68.8**	**54.5**	**66.3**	**70.2**	**63.1**	**37.0** ± 1.4	**0.050**
Depth Pro + SAM 3D Body [8, 17] [Unaligned]	**D + P**	*69.8*	*61.6*	42.3	*57.9*	56.2	43.9	*26.0* ± 1.5	*0.061*
Depth Pro [17] [Depth-only]	**D** only	67.4	58.9	*44.1*	56.8	43.5	24.8	—	—
SAM 3D Body [8] [Pose-only]	**P** only	64.8	49.7	36.8	50.4	*58.8*	*46.0*	—	—

F Additional Qualitative Results

We close with multi-view renderings of H-Flow and Zero MSF [16] under increasing distribution shift. The sequences cover an in-the-wild ballet performance (Sec. F.1), in-domain Fit3D (Sec. F.2), and out-of-domain DynAct4D under heavy cloth deformation (Sec. F.3).

F.1 Multi-view Renderings of the Ballet Reconstruction

Fig. 1 renders only a frontal view, leaving open whether each reconstruction is genuinely volumetric or merely a frontal-facing depth sheet. Off-axis views (Fig. 7) expose this difference. Zero MSF collapses into a near-planar surface. H-Flow preserves coherent volumetric structure across viewpoints, keeping the tutu and extended limbs distinct from the background. The corresponding orbit animations are provided in the supplementary material.

F.2 Multi-view Renderings on Fit3D (In-Domain)

On in-domain Fit3D (Fig. 8), H-Flow produces flow vectors aligned with the ground truth in both direction and magnitude across torso and limbs. Zero MSF returns fragmented point clouds and sparse, misdirected vectors, leaving limb articulation largely unregistered.

F.3 Multi-view Renderings on DynAct4D (Out-of-Domain)

On out-of-domain DynAct4D (Fig. 9), a long dress introduces non-rigid cloth deformation absent from training. H-Flow recovers a coherent silhouette and consistent 3D flow across both torso and swirling skirt. Zero MSF fragments the point cloud, concentrates sparse flow on the upper body, and leaves skirt deformation largely unmodeled.

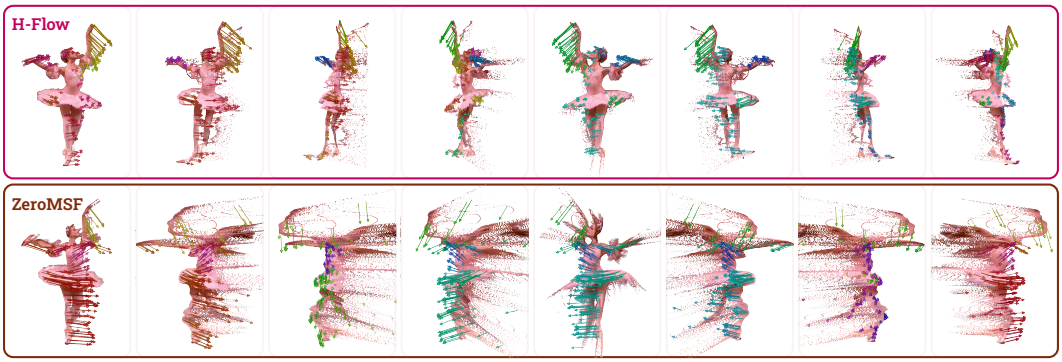

Figure 7: **Multi-view renderings of the ballet reconstruction.** 3D point clouds from Fig. 1 rotated across camera azimuths at fixed elevation and field of view (top: H-Flow; bottom: Zero MSF [16]).

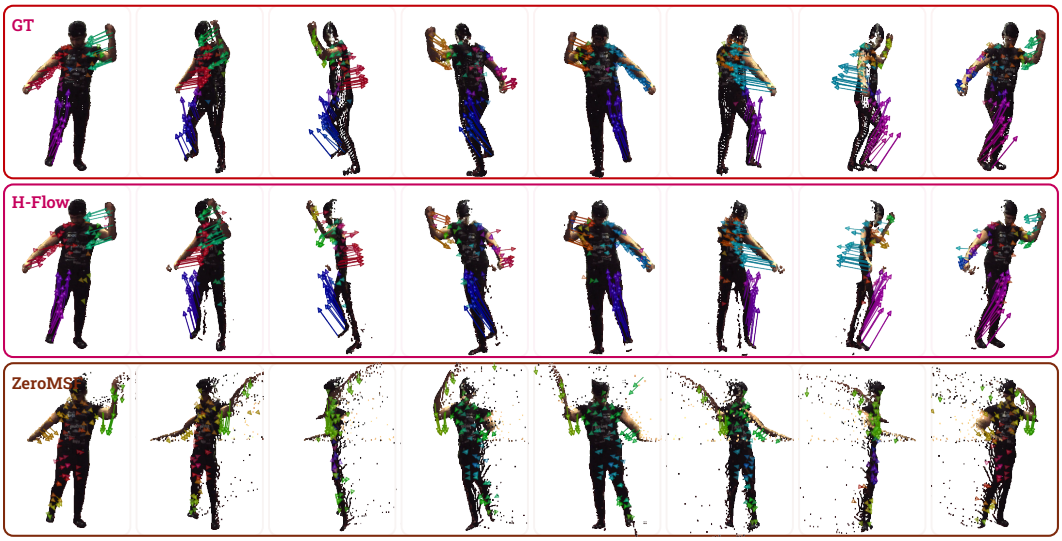

Figure 8: **Multi-view flow comparison on Fit3D.** Columns sweep camera azimuth; rows show ground truth (top), H-Flow (middle), and Zero MSF (bottom).



Figure 9: **Multi-view flow comparison on DynAct4D.** Columns sweep camera azimuth; rows show ground truth (top), H-Flow (middle), and Zero MSF (bottom).

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Sec. 1 state four contributions (the H-Flow representation, the self-supervised paradigm, the DynAct4D benchmark, and zero-shot transfer to in-the-wild video), each substantiated by experiments in Sec. 5.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Sec. 6 explicitly discusses both that the physical priors are scale-invariant (requiring an external anchor during training) and that ground-support and near-bone priors soften under aerial maneuvers or extreme loose-garment lag.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: This work is empirical in nature and does not present formal theoretical results, theorems, or proofs.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Sec. 5 specifies dataset splits, baselines, and evaluation metrics, while Secs. A to C detail the architecture, loss implementations, and training configuration sufficient to reproduce the reported results.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: As stated in the abstract, code, pretrained models, and the DynAct4D benchmark will be released upon publication, accompanied by configurations and evaluation scripts to reproduce the reported results.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Sec. 5 reports dataset splits and evaluation protocols, and Sec. C provides optimizer choice, learning rate schedule, batch size, training horizon, and data augmentation details.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Scene flow evaluation on the benchmarks we use (Fit3D, 3DPW, DynAct4D) computes deterministic per-pixel metrics over fixed test splits, and the baselines we compare against likewise report single-run results, so error bars are not part of the established evaluation protocol in this setting.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Sec. C reports the GPU type and count (6 NVIDIA H100 80GB), training horizon (4×10^5 iterations for pretraining, 1×10^5 for fine-tuning), and the mixed-precision setup used.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics, and the research, dataset construction, and experimental procedures reported in this paper conform to its requirements.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: As foundational research on dense human motion estimation, H-Flow can support beneficial applications in rehabilitation and clinical biomechanics, sports and movement analysis, animation and visual effects, and human-computer interaction.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: H-Flow is a motion estimation model and DynAct4D is fully synthetic data containing no personally identifiable information, so neither asset poses the high-risk misuse profile that this question targets.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All datasets (Human3.6M, Fit3D, 3DPW), pretrained models (DINOv3, Depth Pro, SAM 3D Body, SAM), and content libraries (MetaHuman, Unreal Engine, Mixamo) used in this work are properly cited and used in accordance with their respective licenses and terms of use.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We will release H-Flow code and pretrained checkpoints together with the DynAct4D benchmark under a research-only, non-commercial license, accompanied by dataset documentation covering subjects, garments, actions, scenes, viewpoints, file formats, and the evaluation protocol described in Sec. 4.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: This work involves no crowdsourcing and no human subject research; DynAct4D is constructed entirely from synthetic characters in a computer graphics pipeline.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This work involves no human subject research, so IRB approval is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not part of the core methodology of this paper; the H-Flow model, the self-supervised paradigm, and the DynAct4D benchmark involve no language models in their development or evaluation.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.