CoHSPAD: Continuous Human Gaussian Reconstruction from Raw Monocular 1-bit SPAD Streams

Anonymous authors
Paper under double-blind review

Abstract

Monocular human reconstruction from conventional RGB video is limited by finite frame rates and exposure time. Rapid motion causes severe blur, degrading the image quality needed for reliable pose estimation and appearance recovery. SPAD cameras offer an alternative: binary photon-arrival observations at very high temporal resolution, which remain informative under low light and rapid motion. However, a single 1-bit frame is extremely sparse, and independent per-frame optimization of SMPL parameters and 3D Gaussians is easily destabilized by noisy photon gradients. We present CoHSPAD, a monocular framework for 3D Gaussian reconstruction of dynamic humans and static scenes with continuous human motion recovered directly from raw SPAD streams. Built on a differentiable SPAD imaging model, we supervise rendering directly in the 1-bit raw sensor domain. Instead of independent per-frame poses, we model motion as a continuous SMPL trajectory with sparse temporal knots that share supervision across the stream and stabilize pose gradients. We introduce a Multi-Scale Temporal Fusion Network (MTF-Net) to aggregate long and short streams per-timestamp, providing a more stable signal than a single binary frame while avoiding blur from naive temporal averaging. Experiments on our newly collected CoH-Synthesis and CoH-Real show that CoHSPAD remains robust and temporally consistent under inaccurate initial SMPL poses and rapid motion.

1 Introduction

Monocular dynamic human reconstruction aims to recover a dynamic 3D human representation from continuous observations captured by a fixed camera, enabling novel-view synthesis and human animation. Recent advances based on Neural Radiance Fields (NeRFs) (Mildenhall et al., 2020) and 3D Gaussian Splatting (3DGS) (Kerbl et al., 2023) have achieved promising results in recovering photorealistic human appearance and geometry from monocular videos. However, these methods typically rely on clean and stable intensity observations from conventional RGB cameras. Under fast human motion, finite frame rates and exposure times inevitably cause motion blur, degrading silhouettes and textures needed for reliable pose estimation and appearance recovery. These degradations propagate to both geometry and motion estimation, making dynamic human reconstruction particularly challenging under rapid motion.

SPAD (Single-Photon Avalanche Diode) cameras provide a different paradigm. Unlike conventional RGB cameras that integrate incoming light over an exposure interval, SPAD cameras record binary photon detections at high temporal resolution, preserving temporal information under low light and rapid motion. However, this high temporal resolution comes at the cost of extremely limited photon information at each instant. Each SPAD pixel produces a binary observation that can be modeled as a Bernoulli trial, taking a value of 1 when one or more photons are detected within a short interval and 0 otherwise. Consequently, individual 1-bit SPAD frames are extremely sparse and dominated by photon shot noise. As illustrated in Fig. 1, conventional exposure loses temporal detail through motion blur, whereas SPAD preserves fast motion but yields extremely sparse observations. This sparsity makes existing RGB-based reconstruction pipelines difficult to apply directly. In particular, per-frame optimization of SMPL pose and 3D Gaussians becomes highly sensitive to noisy photon gradients, while accumulating neighboring SPAD frames improves photon statistics at the

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

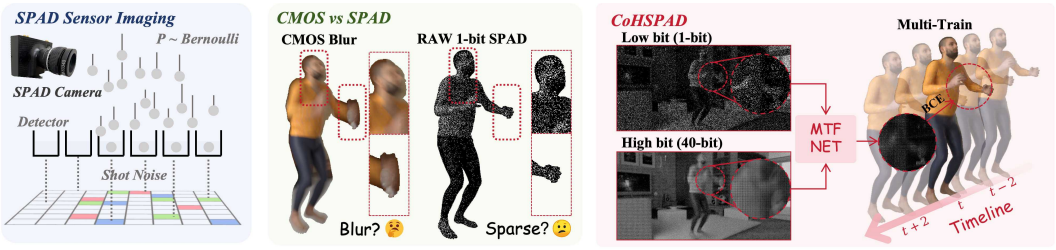

Figure 1: **SPAD imaging and CoHSPAD framework.** Left: Raw SPAD imaging follows Bernoulli photon detection with shot noise. Middle: CMOS exposure causes motion blur, while 1-bit SPAD streams are temporally precise but extremely sparse. Right: CoHSPAD combines continuous SMPL poses and MTF-Net to aggregate sparse SPAD observations for human and scene reconstruction.

cost of reintroducing motion blur. The key challenge is therefore to obtain reliable reconstruction supervision from sparse 1-bit observations without sacrificing SPAD's high temporal resolution.

To address this challenge, we present **CoHSPAD**, a monocular framework that reconstructs a dynamic human Gaussian representation and a static scene from a continuous SPAD stream. A differentiable SPAD imaging model connects 3D Gaussians with binary photon detection, enabling end-to-end optimization in the raw (Mildenhall et al., 2022) 1-bit sensor domain. Our key observation is that human motion is inherently continuous. Unlike exposure-based motion models that parameterize motion separately within individual input frames, we model a single continuous SMPL trajectory over the entire SPAD stream with sparse temporal control points. This allows photon observations across time to jointly constrain shared motion parameters and stabilize pose and Gaussian optimization. To further strengthen the limited supervision at each timestamp without direct frame accumulation, we propose Multi-Scale Temporal Fusion Network (MTF-Net), which fuses complementary temporal contexts toward the target timestamp to provide stable photon supervision with reduced temporal blur. Our main contributions are summarized as follows:

- We propose **CoHSPAD**, a monocular framework for dynamic human and static scene reconstruction from SPAD streams, with a differentiable SPAD imaging model and **SPAD-aware densification** enabling robust optimization in the raw 1-bit sensor domain.
- We introduce a **continuous SMPL pose trajectory** parameterized by sparse temporal control points, enabling cross-time photon supervision to stabilize pose and Gaussian optimization under sparse observations.
- We propose **MTF-Net** to fuse complementary temporal contexts for stronger photon supervision with reduced temporal blur, and introduce the new **CoH-Synthesis** and **CoH-Real** datasets for synthetic and real-world evaluation.

2 Related Work

2.1 Monocular Human and Scene Reconstruction

NeRF and 3DGS have significantly advanced monocular human reconstruction. Gaussian-based methods, including GauHuman (Hu et al., 2024b), GaussianAvatar (Hu et al., 2024a), 3DGS-Avatar (Qian et al., 2024), and Animatable Gaussians (Li et al., 2024), typically reconstruct humans in canonical space and deform them to target poses using SMPL priors and Linear Blend Skinning (LBS). Most methods (Wen et al., 2024; Shao et al., 2024; Pang et al., 2024; Moon et al., 2024; Moreau et al., 2024; Tang et al., 2025; Saunders et al., 2025) focus primarily on foreground humans. HUGS (Kocabas et al., 2024) extends this paradigm by jointly reconstructing human and scene Gaussians with per-frame SMPL pose refinement. ODHSR (Zhang et al., 2025) and SkinningGS (Li et al., 2025) further explore joint human-scene reconstruction, and E-Sem3DGS (Yin et al., 2025) incorporates high-temporal-resolution event observations to handle fast motion. Departing from these frame-based formulations, CoHSPAD models human motion continuously over the photon stream.

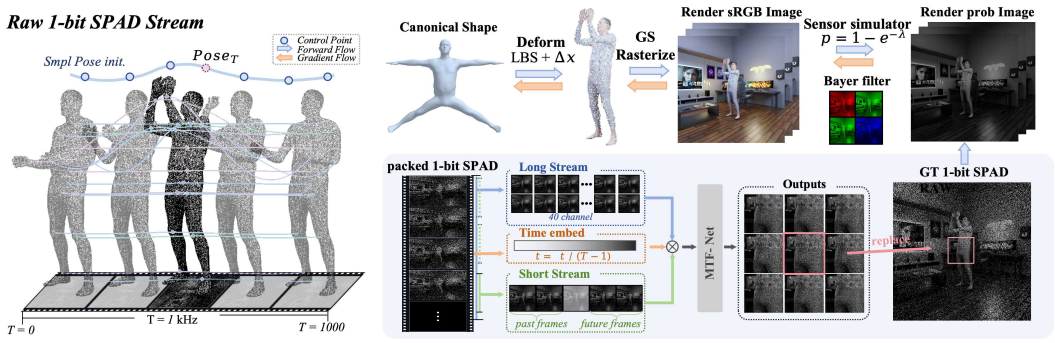

Figure 2: **Overview of CoHSPAD.** CoHSPAD reconstructs dynamic humans and static scenes from monocular 1-bit SPAD streams via continuous SMPL trajectories, differentiable Gaussian rendering and SPAD simulation, and an MTF-Net supervised by long and short streams.

2.2 Photon-based Reconstruction

High-temporal-resolution, low-bit sensing (Seets et al., 2021; Ma et al., 2020; Sundar et al., 2025; Jungerman et al., 2025a; Dutson et al., 2022) has recently been explored for NeRF and 3DGS reconstruction. SpikeGS (Zhang et al., 2024; Yu et al., 2024) combines instantaneous and accumulated spike signals for static scene reconstruction. USP-Gaussian (Chen et al., 2025) uses a spike reconstruction network with short and long spike inputs, jointly optimized with Gaussian Splatting. Nope-SGS (Guo et al., 2026) extends spike-based reconstruction to unposed streams. Event-based methods (Muglikar et al., 2025) similarly exploit asynchronous brightness changes for fast motion reconstruction. However, SPAD cameras follow a distinct physical image formation model. PhotonSplat (Teja et al., 2025) directly optimizes 3D Gaussian from binary SPAD observations with sensor-domain supervision, focusing on static scenes. In parallel, bit2bit (Liu et al., 2024) and QUIVER (Chennuri et al., 2024) recover 2D intensity videos from sparse photon measurements.

2.3 Deblur Reconstruction

Recent works (Srinivasan et al., 2017; Ma et al., 2022; Lee et al., 2023) have incorporated continuous motion modeling into reconstruction from motion-blurred RGB images. BAD-Gaussians (Zhao et al., 2024) models continuous camera trajectories during exposure, and CoMoGaussian (Lee et al., 2025) further represents camera motion with a Neural ODE. For dynamic humans, Deblur-Avatar (Luo et al., 2025) interpolates SMPL poses over each exposure interval. MAD-Avatar (Niu et al., 2026) instead adopts joint-level splines to model intra-exposure motion. In these methods, continuous trajectories primarily describe latent motion within individual RGB exposures.

3 Method

Given a monocular 1-bit SPAD stream $\{\mathbf{B}_t\}_{t=1}^{T}$ and initial SMPL estimates, our goal is to jointly reconstruct an animatable human representation, a static scene, and continuous human motion. We represent both the human and scene with 3D Gaussians and directly supervise their rendering in the raw SPAD sensor domain through a differentiable photon-detection model.

3.1 Preliminary

3D Gaussian Human Representation. We represent the dynamic human and static scene with 3D Gaussian sets $\mathcal{G}^h$ and $\mathcal{G}^s$. Each Gaussian $\mathcal{G}_i$ is parameterized by its center $\boldsymbol{\mu}_i \in \mathbb{R}^3$, covariance $\boldsymbol{\Sigma}_i \in \mathbb{R}^{3\times3}$, opacity α_i, and appearance coefficients $\mathbf{c}_i$, where $\boldsymbol{\Sigma}_i = \mathbf{R}_i\mathbf{S}_i\mathbf{S}_i^\top\mathbf{R}_i^\top$ with $\mathbf{R}_i$ and $\mathbf{S}_i$ specifying rotation and scale. Human Gaussians are defined in canonical space and posed at time t via SMPL linear blend skinning (LBS):

$$\boldsymbol{\mu}_{i,t}^{\mathrm{LBS}} = \sum_{j=1}^{J} w_{i,j}\, \mathbf{T}_j(\boldsymbol{\theta}(t), \boldsymbol{\beta})\, \tilde{\boldsymbol{\mu}}_i^c, \qquad \sum_{j=1}^{J} w_{i,j} = 1, \tag{1}$$

where $\tilde{\boldsymbol{\mu}}_i^c$ is the homogeneous canonical center, $w_{i,j}$ are the skinning weights, $\boldsymbol{\beta}$ denotes the SMPL shape parameters, and $\mathbf{T}_j$ is the corresponding SMPL joint transformation. The posed human and static Gaussians are combined as $\mathcal{G}(t) = \mathcal{G}^h(t) \cup \mathcal{G}^s$ and rendered via standard Gaussian splatting:

$$\mathbf{I}_t(\mathbf{p}) = \sum_{i \in \mathcal{N}(\mathbf{p})} \mathbf{c}_i \, \tilde{\alpha}_i(\mathbf{p}) \prod_{j<i} (1 - \tilde{\alpha}_j(\mathbf{p})) \,, \tag{2}$$

where $\mathcal{N}(\mathbf{p})$ contains depth-ordered Gaussians at pixel $\mathbf{p}$, and $\tilde{\alpha}_i(\mathbf{p})$ denotes projected opacity.

SPAD Image Formation. A 1-bit SPAD records whether a photon is detected per pixel during a short exposure. Let $\mathbf{B}_t(\mathbf{p}) \in \{0, 1\}$ denote the binary observation at pixel $\mathbf{p}$ and time t. Assuming Poisson photon arrivals, we map rendered linear intensity to the expected photon count as

$$\lambda_t(\mathbf{p}) = e^{\eta} \, \mathcal{M}_{\mathrm{CFA}}(\mathbf{I}_t)(\mathbf{p}), \qquad P_t(\mathbf{p}) = 1 - e^{-\lambda_t(\mathbf{p})}, \tag{3}$$

where $\mathcal{M}_{\mathrm{CFA}}$ denotes sampling under the raw Bayer layout and η is a learnable global exposure parameter. The resulting $P_t(\mathbf{p})$ defines the Bernoulli photon-detection probability, allowing the rendered Gaussians to be directly supervised by raw 1-bit observations.

3.2 Continuous Articulated Representation

Shared Pose Trajectory. Instead of assigning an independent body pose to each SPAD timestamp, we represent human articulation by a continuous trajectory that can be queried at any time. The 23 SMPL body joints are represented by axis-angle rotations, forming a pose vector $\boldsymbol{\theta}(t) \in \mathbb{R}^{69}$. We parameterize the trajectory using M learnable pose controls $\{(\tau_k, \boldsymbol{\theta}_k)\}_{k=1}^{M}$ distributed over the sequence. These controls are initialized from estimated SMPL poses and refined during reconstruction. For any timestamp $t \in [\tau_k, \tau_{k+1}]$, the pose backbone is obtained through B-spline interpolation:

$$\bar{\boldsymbol{\theta}}(t) = \mathcal{S}(t; \, \boldsymbol{\theta}_{k-1}, \boldsymbol{\theta}_k, \boldsymbol{\theta}_{k+1}, \boldsymbol{\theta}_{k+2}) \,, \tag{4}$$

where $\mathcal{S}$ denotes B-spline interpolation. This representation shares motion parameters across neighboring observations while allowing the SMPL pose to be evaluated at arbitrary timestamps. Sparse photon observations across time therefore jointly constrain a common motion trajectory.

Shared Pose Residual. The B-spline trajectory provides temporal sharing, while the sparse controls may limit local motion flexibility. We therefore introduce a residual function f_{pose} shared across the sequence. Given the interpolated pose and its temporal velocity, we obtain

$$\Delta\boldsymbol{\theta}(t) = f_{\mathrm{pose}}\Big(\bar{\boldsymbol{\theta}}(t), \dot{\bar{\boldsymbol{\theta}}}(t)\Big) \,, \qquad \boldsymbol{\theta}(t) = \bar{\boldsymbol{\theta}}(t) + \Delta\boldsymbol{\theta}(t), \tag{5}$$

where $\dot{\bar{\boldsymbol{\theta}}}(t) = \partial\bar{\boldsymbol{\theta}}(t)/\partial t$. The residual provides a sequence-shared correction for local motion variations without introducing independent poses at individual timestamps.

Post-LBS Gaussian Deformation. The SMPL trajectory models skeletal motion, while local non-rigid deformation is captured by a lightweight residual field after LBS. We encode the motion state as $\mathbf{e}_t = E_{\mathrm{pose}}(\boldsymbol{\theta}(t), \dot{\bar{\boldsymbol{\theta}}}(t))$ and deform each human Gaussian as

$$\boldsymbol{\mu}_{i,t} = \boldsymbol{\mu}_{i,t}^{\mathrm{LBS}} + f_{\mathrm{def}}(\gamma(\boldsymbol{\mu}_i^c), \mathbf{e}_t) \,, \tag{6}$$

where E_{pose} and f_{def} denote the pose encoder and deformation field, respectively, and $\gamma(\boldsymbol{\mu}_i^c)$ encodes the canonical Gaussian position. The shared deformation field allows the human representation to be evaluated at arbitrary timestamps without independent Gaussian offsets.

3.3 Multi-Scale Temporal Fusion Network

While the continuous trajectory provides pose supervision across time, photon information at each timestamp remains extremely sparse. Temporal accumulation improves photon statistics, but a fixed window mixes body states and introduces motion blur. This trade-off is spatially varying: slowly changing regions benefit from broader temporal evidence, whereas fast-moving regions require observations closer to the target time to preserve motion boundaries. We therefore introduce the **Multi-Scale Temporal Fusion Network (MTF-Net)** $\mathcal{A}_\psi$ to spatially adapt photon aggregation. By combining long and short temporal contexts, MTF-Net produces a denser target-aligned photon probability with reduced motion blur.

Long and Short Temporal Context Encoding. Inspired by the multi-scale temporal design of USP-Gaussian, we construct two complementary contexts from raw SPAD observations within a human-centered crop. For target time t, the long temporal context $\mathcal{B}_t^l = \{\mathbf{B}_\tau\}_{\tau\in\mathcal{W}_l(t)}$ spans a broad temporal window, while the short temporal context $\mathcal{B}_t^s = \{\mathbf{B}_\tau\}_{\tau\in[t-r,t+r]\setminus\{t\}}$ contains observations near t. Here, $\mathcal{W}_l(t)$ denotes the long temporal window around t, r the short temporal radius, and $\mathbf{B}_\tau \in \{0,1\}^{h\times w}$ a cropped raw SPAD observation of resolution $h \times w$. Rather than averaging the long context, we preserve its temporal ordering through ordered sub-bin summaries before encoding it as $\mathbf{F}^l = E_l(\mathcal{B}_t^l)$, enabling content-dependent temporal aggregation instead of a fixed accumulation window. E_l denotes the long temporal encoder. For each anchor a_m, the short temporal encoder E_s extracts $\mathbf{F}^s_{a_m} = E_s(\mathcal{B}^s_{a_m})$ from its temporal neighborhood. Target observations are withheld from corresponding inputs to prevent shortcut prediction.

Multi-Anchor Prediction. We select A anchors $\{a_m\}_{m=1}^A$ within each long temporal window. For each anchor, its short temporal feature is fused with the shared long temporal feature and a normalized temporal encoding: $\widehat{\mathbf{Q}}^h_{a_m} = H\big(\mathrm{Cat}[\mathbf{F}^l, \mathbf{F}^s_{a_m}, \mathbf{T}_{a_m}]\big)$, where H denotes the prediction head and $\mathbf{T}_{a_m}$ is the normalized temporal encoding of anchor a_m. Here, $\widehat{\mathbf{Q}}^h_{a_m}$ represents the target-aligned photon detection probability at a_m. For the center anchor $a_c = t$, we use $\mathbf{Q}^h_t = \widehat{\mathbf{Q}}^h_{a_c}$ as the target-time prediction, while the remaining anchors enforce temporal consistency during self-supervised training. We combine this denser human-region target with raw observation as

$$\widetilde{\mathbf{B}}_t(\mathbf{p}) = \begin{cases} \mathbf{Q}^h_t(\mathbf{p}), & \mathbf{p} \in \Omega^h_t, \\ \mathbf{B}_t(\mathbf{p}), & \text{otherwise}, \end{cases} \tag{7}$$

where Ω^h_t denotes the human region. Thus, MTF-Net adaptively fuses temporal photon evidence for the dynamic human, while the static scene retains the original target-time 1-bit observation. The self-supervised MTF-Net objective is described in Sec. 3.4.

3.4 Optimization and Densification

Self-Supervised MTF-Net Optimization. MTF-Net is trained solely from raw 1-bit observations. All A blind-anchor predictions are used for temporal supervision, while the center prediction additionally guides human reconstruction. For R nested anchor intervals $[a_{i_r}, a_{j_r}]$, we match the mean predicted probability to the observed photon frequency:

$$\mathcal{L}_{\text{long}} = \frac{1}{R}\sum_{r=1}^{R}\left\| \frac{\sum_{m=i_r}^{j_r}\widehat{\mathbf{Q}}^h_{a_m}}{j_r - i_r + 1} - \frac{\sum_{\tau=a_{i_r}}^{a_{j_r}}\mathbf{B}^h_\tau}{a_{j_r} - a_{i_r} + 1}\right\|_1. \tag{8}$$

Each anchor is further matched to its local five-frame photon frequency:

$$\mathcal{L}_{\text{short}} = \frac{1}{A}\sum_{m=1}^{A}\left\|\widehat{\mathbf{Q}}^h_{a_m} - \frac{1}{5}\sum_{\tau=a_m-2}^{a_m+2}\mathbf{B}^h_\tau\right\|_1. \tag{9}$$

The MTF-Net loss is $\mathcal{L}_{\text{anchor}} = \mathcal{L}_{\text{long}} + \lambda_{\text{s}}\mathcal{L}_{\text{short}} + \lambda_{\text{tv}}\mathcal{L}_{\text{tv}}$, where $\mathcal{L}_{\text{tv}}$ provides weak smoothness.

SMPL and Gaussian Optimization. At each iteration, we sample K consecutive timestamps and render each timestamp independently using its continuous pose. This exposes the shared trajectory to neighboring photon observations without temporally averaging the rendered motion. Using the mixed target from equation 7, the Bernoulli reconstruction loss is

$$\mathcal{L}_{\text{bce}} = \frac{1}{K}\sum_{t\in\mathcal{T}}\mathrm{BCE}\Big(\mathbf{P}_t, \mathrm{sg}[\widetilde{\mathbf{B}}_t]\Big), \qquad |\mathcal{T}| = K, \tag{10}$$

where $\mathrm{sg}[\cdot]$ stops gradients toward MTF-Net. The reconstruction objective is

$$\mathcal{L}_{\text{photo}} = \lambda_{\text{bce}}\mathcal{L}_{\text{bce}} + \lambda_{\text{mask}}\mathcal{L}_{\text{mask}} + \lambda_{\text{limit}}\mathcal{L}_{\text{limit}} + \lambda_{\text{smooth}}\mathcal{L}_{\text{smooth}}, \tag{11}$$

where $\mathcal{L}_{\text{mask}}$ supervises human opacity, while $\mathcal{L}_{\text{smooth}}$ and $\mathcal{L}_{\text{limit}}$ regularize the pose controls. Their definitions are provided in the supplementary material. MTF-Net is optimized independently with $\mathcal{L}_{\text{anchor}}$, and its prediction serves as a detached reconstruction target without receiving gradients from Gaussian optimization.

SPAD-Aware Densification. Standard 3DGS densification accumulates image-space position gradients as $s_i = \frac{1}{N_i}\sum_{n=1}^{N_i}\left\|\frac{\partial \mathcal{L}}{\partial \boldsymbol{\mu}_{i,n}^{2D}}\right\|_2$, where N_i is the number of valid observations. Sparse photon supervision can produce unstable gradients from isolated detections. We therefore use separate SPAD-aware densification for human and scene Gaussians.

For **human Gaussians**, we accumulate standard 3DGS densification gradients only for visible Gaussians within a dilated human silhouette. To reduce unreliable growth under sparse SPAD observations, we further assess gradient consistency across neighboring timestamps. Gaussians with persistent gradients retain higher densification scores, while isolated spikes are suppressed before standard cloning and splitting. For **scene Gaussians**, neighboring SPAD observations can be safely aggregated because the background is static. We form the soft photon frequency $\mathbf{B}_t^{\text{soft}} = \frac{1}{K}\sum_{\tau\in\mathcal{T}}\mathbf{B}_\tau$ and compare it with the rendered photon probability in each Bayer sub-plane c:

$$\mathbf{R}_t^c = \mathrm{DoG}(\mathcal{V}(\mathbf{P}_t^c)) - \mathrm{DoG}\Big(\mathcal{V}(\mathbf{B}_t^{\text{soft},c})\Big), \qquad \mathcal{V}(x) = 2\arcsin\sqrt{x}, \tag{12}$$

where DoG emphasizes structural differences and $\mathcal{V}$ stabilizes photon variance. The resulting loss $\mathcal{L}_{\text{str}} = \sum_c \|\mathbf{R}_t^c\|_1$ is used only for scene densification and is excluded from the reconstruction objective. We retain the highest-scoring scene Gaussians under a fixed budget, followed by standard 3DGS cloning, splitting, and pruning.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate CoHSPAD on synthetic (**CoH-Synthesis**) and real (**CoH-Real**) SPAD data, each comprising five scenes. CoH-Synthesis contains 2-s sequences rendered at 1kHz and 1024^2 resolution in linear sensor space. We synthesize 1-bit SPAD measurements using the observation model in Sec. 3 and retain clean renderings as ground truth. CoH-Real contains 1-s sequences captured at 8kHz, yielding 8,000 binary 1024^2 Bayer frames each. Initial SMPL poses for both datasets are estimated by GVHMR (Shen et al., 2024) without ground-truth poses. Human masks are available for CoH-Synthesis and estimated by Cutie (Cheng et al., 2024) for CoH-Real. During training, the CFA bad-pixel map excludes defective and persistently active pixels from sensor supervision without replacement or inpainting. See the supplementary material for acquisition details.

Baselines and Evaluation. We compare with MAD-Avatar and HUGS under two input settings. For accumulated SPAD, every 40 synthetic or 320 real frames are combined into one observation, corresponding to 40 ms temporal support and approximately 25 observations per second while preserving the Bayer layout. We replace the original RGB L1 supervision with Bernoulli BCE in the SPAD domain. SSIM and LPIPS are computed on demosaiced sRGB images converted from predicted linear intensity. For denoised RGB, the accumulated measurements are converted to intensity, demosaiced, and denoised with Restormer (Zamir et al., 2022) before being passed to the standard RGB pipelines of MAD-Avatar and HUGS. The main comparison evaluates reconstruction at common exposure-aligned timestamps under each method's input formulation. Ablations and temporal analyses instead use held-out timestamps whose raw SPAD observations are excluded from CoHSPAD optimization, thereby evaluating temporal interpolation rather than fitting quality.

Implementation Details. We optimize each sequence with $M = 200$ cubic pose controls and supervise $K = 5$ consecutive timestamps. MTF-Net uses 128×128 human-centered crops with $A = 9$ blind anchors and is warmed up for 2,000 updates before Gaussian optimization. For both datasets, its 320-frame long context uses 10 temporal bins with 4 ordered sub-bins each, while each anchor uses neighboring observations as short context. Sensor-domain optimization preserves the Bayer layout and learns the global exposure parameter in equation 3.

4.2 Results on CoH-Synthesis

Quantitative Comparison. Table 1 compares reconstruction quality on CoH-Synthesis. For accumulated SPAD, HUGS and MAD-Avatar use Bernoulli BCE instead of their original RGB photometric loss, while their pseudo-RGB variants operate on accumulated and demosaiced Bayer observations. All methods are evaluated at the same exposure-aligned timestamps using identical clean

Table 1: Reconstruction comparison on CoH-Synthesis for **(a)** human-region and **(b)** full-image reconstruction. Best and second-best results are highlighted.

		Catch			Jog			Jump			Sit			Walk			Average		
Method	**Input**	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
(a) Human-region reconstruction																			
HUGS	Accum. SPAD	26.661	0.890	0.145	26.112	0.858	0.190	26.217	0.876	0.193	21.885	0.800	0.205	24.256	0.862	0.149	25.026	0.857	0.176
MAD-Avatar	Accum. SPAD	30.253	0.914	0.140	26.091	0.798	0.238	28.479	0.838	0.199	26.042	0.846	0.173	29.101	0.893	0.156	27.993	0.858	0.181
HUGS	Restormer RGB	26.385	0.886	0.138	25.851	0.853	0.200	22.380	0.837	0.252	21.269	0.788	0.203	24.205	0.859	0.154	24.018	0.845	0.189
MAD-Avatar	Restormer RGB	31.338	0.946	0.117	28.416	0.886	0.172	29.610	0.907	0.155	25.937	0.874	0.160	27.512	0.919	0.143	28.563	0.906	0.149
Ours	**Raw SPAD**	**32.353**	**0.950**	**0.082**	**30.435**	**0.921**	**0.131**	**32.019**	**0.930**	**0.120**	**27.533**	**0.880**	**0.148**	**30.890**	**0.943**	**0.089**	**30.646**	**0.925**	**0.114**
(b) Full-image reconstruction																			
HUGS	Accum. SPAD	29.638	0.900	0.310	31.097	0.874	0.419	29.838	0.838	0.370	27.911	0.837	0.334	30.615	0.924	0.303	29.820	0.875	0.347
HUGS	Restormer RGB	29.797	0.905	0.314	30.848	0.877	0.423	28.155	0.837	0.374	26.618	0.823	0.368	30.540	0.921	0.306	29.191	0.872	0.357
Ours	**Raw SPAD**	**35.648**	**0.954**	**0.208**	**34.550**	**0.912**	**0.355**	**33.250**	**0.897**	**0.259**	**32.852**	**0.904**	**0.236**	**36.193**	**0.974**	**0.172**	**34.499**	**0.928**	**0.246**

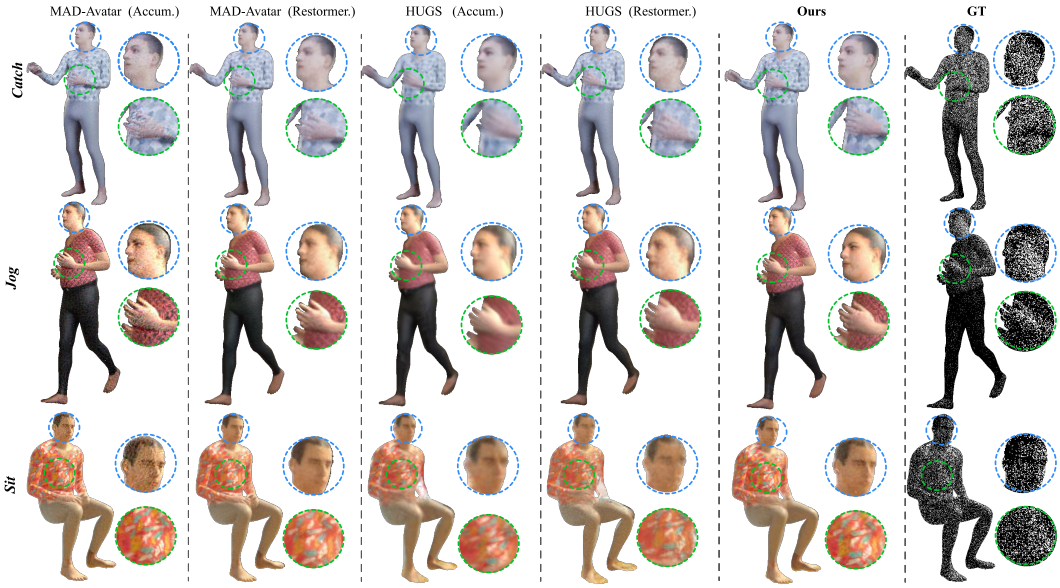

Figure 3: **Human reconstruction on CoH-Synthesis.** Comparison with baselines across human motions. CoHSPAD recovers clearer appearance and motion boundaries from raw SPAD data.

references and metrics. CoHSPAD consistently achieves better reconstruction, with larger gains in the dynamic human region. HUGS remains sensitive to photon noise on accumulated SPAD, while pseudo-RGB preprocessing suppresses noise at the cost of temporal information. MAD-Avatar handles motion blur more effectively, but its frame-based motion modeling remains challenging under sparse SPAD observations. Fig. 3 shows clearer human boundaries and more consistent appearance under rapid motion, supporting the benefit of exploiting temporal information directly from raw SPAD rather than collapsing the stream before reconstruction. Full-image results in Fig. 4 likewise show better preserved scene structures and human details, consistent with the quantitative results. MAD-Avatar is omitted from full-image evaluation as a human-only method.

4.3 Results on CoH-Real

Fig. 5 compares human reconstruction on five real SPAD streams. For each scene, we show the direct linear rendering and its display-corrected result. Since the real sensor follows a BGGR Bayer layout, the direct rendering exhibits a noticeable green cast. For visualization, we apply pre-calibrated white-balance gains to the reconstructed linear RGB image, followed by tone mapping and conversion to sRGB. These gains are used only for visualization and are not involved in training. The real results reveal a clear trade-off between photon support and temporal resolution. HUGS generally produces the most blurred reconstruction because it does not explicitly account for motion blur. MAD-Avatar benefits from the higher photon density provided by temporal accumulation and can preserve fine appearance details, such as clothing texture, in regions with limited motion. However,

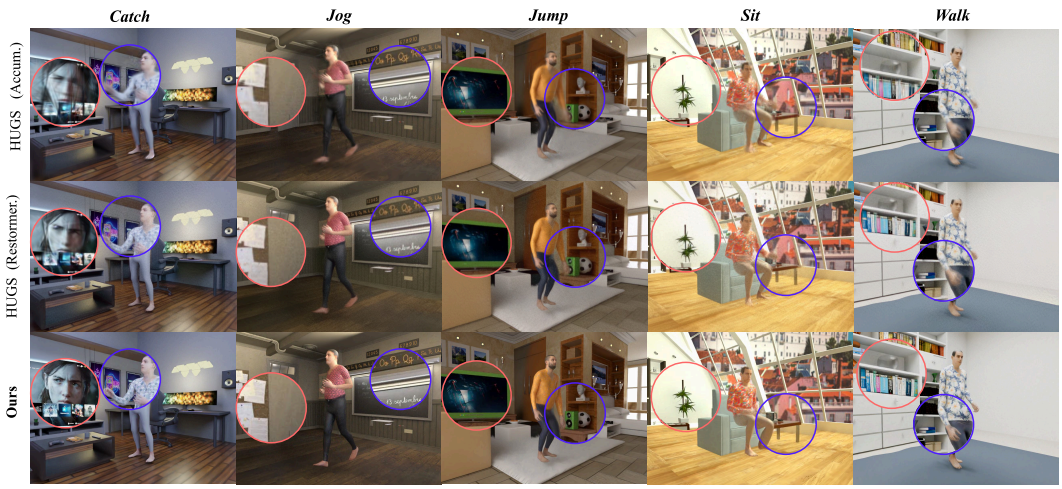

Figure 4: **Full-image reconstruction on CoH-Synthesis.** Comparison of joint human-scene reconstruction. CoHSPAD favorably reconstructs the dynamic human and static background.

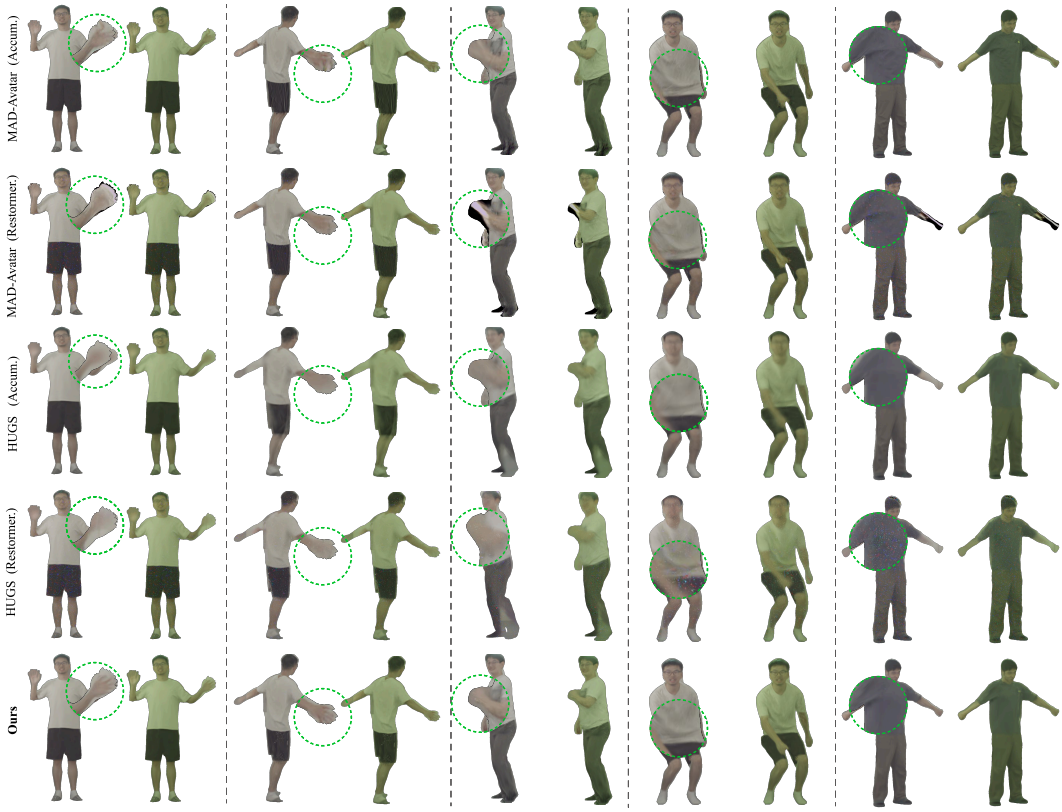

Figure 5: **Human reconstruction on CoH-Real.** Comparison with accumulated SPAD and pseudo RGB baselines. CoHSPAD produces more consistent appearance and cleaner body structures.

temporal accumulation mixes observations during rapid motion, and fine structures such as moving hands are often less completely recovered. Its pseudo-RGB variant further introduces information loss through Bayer demosaicing and preprocessing. CoHSPAD avoids collapsing the raw temporal observations and provides a more balanced reconstruction, preserving coherent appearance while recovering cleaner structures in rapidly moving regions. These results are consistent across the five real captures and complement the observations on CoH-Synthesis.

Table 2: Core ablation at held-out times.

Cont. SMPL	MTF-Net	PSNR↑	SSIM↑	LPIPS↓
×	×	26.542	0.873	0.146
×	✓	26.710	0.877	0.144
✓	×	28.672	0.901	0.123
✓	✓	**28.795**	**0.904**	**0.122**

Table 3: Additional ablation at held-out times.

Variant	PSNR↑	SSIM↑	LPIPS↓
w/o Post-LBS Res.	27.959	0.893	0.130
w/o Pose Res.	28.360	0.897	0.131
w/o Mask Loss	28.604	0.899	0.129
Ours	**28.795**	**0.904**	**0.122**



Figure 6: **Qualitative ablation on CoH-Synthesis.** From left to right: scene densification on *Catch*, mask supervision on *Jog*, and multi-frame training on *Jump*.

4.4 Ablation Study

Continuous Trajectory and Temporal Fusion. We evaluate all ablations on the same held-out timestamps, whose raw SPAD observations are excluded from optimization. Table 2 evaluates the two core components of CoHSPAD. Continuous SMPL provides the largest individual improvement, showing that independent poses are poorly constrained by sparse 1-bit observations. The shared cubic controls instead allow observations across time to jointly constrain a compact motion trajectory. MTF-Net provides a complementary improvement by strengthening target-time photon supervision with temporal context without direct frame accumulation. Their combination achieves the best performance, showing complementary benefits of continuous motion modeling and temporal photon fusion. Table 3 evaluates the motion refinement components. The pose residual MLP is shared across the trajectory, allowing gradients from each 1-bit observation to propagate beyond the four cubic controls involved in interpolation and improve trajectory optimization under sparse supervision. The post-LBS residual further captures local non-rigid deformation beyond skeletal motion. Removing either component degrades reconstruction, with the larger drop without post-LBS refinement showing the benefit of this additional deformation capacity.

Reconstruction Refinements. Figure 6 illustrates the effects of scene densification, mask supervision, and multi-frame training. In *Catch*, standard gradient-based densification provides weak, unstable cues under sparse photon observations, making fine structures difficult to recover. This is particularly visible in the TV face, whose mid-to-high frequency details are missing with standard densification. The proposed scene densification aggregates structural evidence across neighboring observations, providing more reliable cues for Gaussian growth and recovering facial details more faithfully. Since CoHSPAD represents the human and scene with separate Gaussians but optimizes them through joint rendering, full-image supervision cannot ensure their separation. This ambiguity is pronounced in *Jog*, where dark clothing resembles the background and the moving human is harder to optimize than the static scene. Without mask supervision, scene Gaussians can explain human regions, resulting in transparent legs. The mask loss maintains cleaner human–scene separation. In *Jump*, single-timestamp optimization produces highly stochastic motion gradients. Multi-frame training uses five neighboring timestamps sharing largely the same cubic controls, allowing their gradients to jointly constrain the trajectory within each optimization step. This reduces reliance on any single 1-bit observation and produces more stable human structure during rapid motion.

5 Conclusion

We present CoHSPAD for continuous human and scene reconstruction from raw monocular 1-bit SPAD streams. By jointly modeling continuous human motion and temporal photon information, CoHSPAD preserves high temporal resolution while enabling coherent reconstruction. Our results demonstrate their complementary benefits under sparse photon observations and the potential of photon sensing for dynamic human reconstruction under challenging motion.

AI USE STATEMENT

We used generative AI tools to assist with code editing and debugging, and to improve the clarity and readability of the manuscript. All AI-assisted content was carefully reviewed and verified by the authors. The authors take full responsibility for the final paper, including all reported results and claims.

ETHICS STATEMENT

CoH-Real contains human participants captured in a controlled indoor setting for research purposes. All participants consented to the data collection and research use of the captured data. The dataset is used in this work to evaluate dynamic human reconstruction and does not contain annotations of identity or other sensitive personal attributes.

REPRODUCIBILITY STATEMENT

The main paper presents the formulation of CoHSPAD and the experimental protocol, while the supplementary material provides the details needed to reproduce the data preparation, implementation, and evaluation. Building on the dataset description in Sec. 4.1, Sec. A details the construction of CoH-Synthesis and the acquisition of CoH-Real, together with synthetic SPAD generation, baseline preprocessing, initialization, and sensor calibration. For the method described in Sec. 3, Sec. B provides the corresponding implementation details, including the continuous trajectory parameterization, MTF-Net configuration, optimization schedule, and loss weights. To further characterize the temporal setting used in our experiments, Sec. C analyzes human motion over the accumulation window. Complementing the ablations in Sec. 4.4, Sec. D further examines photon-driven pose refinement, robustness to pose initialization, continuous trajectory parameterization, and held-out reconstruction. Finally, Sec. E extends the CoH-Real results in Sec. 4.3 with evaluations of silhouette consistency and human–scene separation.

REFERENCES

Michael J. Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. BEDLAM: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 8726–8737, June 2023.

Kang Chen, Jiyuan Zhang, Zecheng Hao, Yajing Zheng, Tiejun Huang, and Zhaofei Yu. Usp-gaussian: unifying spike-based image reconstruction, pose correction and gaussian splatting. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16609–16618. IEEE, 2025.

Ho Kei Cheng, Seoung Wug Oh, Brian Price, Joon-Young Lee, and Alexander Schwing. Putting the object back into video object segmentation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3151–3161. IEEE, 2024.

Prateek Chennuri, Yiheng Chi, Enze Jiang, GM Dilshan Godaliyadda, Abhiram Gnanasambandam, Hamid R Sheikh, Istvan Gyongy, and Stanley H Chan. Quanta video restoration. In *European Conference on Computer Vision*, pp. 152–171. Springer, 2024.

Matthew Dutson, Yin Li, and Mohit Gupta. Event neural networks. In *European Conference on Computer Vision*, pp. 276–293. Springer, 2022.

Yijia Guo, Tong Hu, Liwen Hu, Lei Ma, and Tiejun Huang. 3d gaussian splatting from unposed spike stream. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 41002–41011, 2026.

Liangxiao Hu, Hongwen Zhang, Yuxiang Zhang, Boyao Zhou, Boning Liu, Shengping Zhang, and Liqiang Nie. Gaussianavatar: Towards realistic human avatar modeling from a single video via animatable 3d gaussians. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 634–644. IEEE, 2024a.

Shoukang Hu, Tao Hu, and Ziwei Liu. Gauhuman: Articulated gaussian splatting from monocular human videos. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20418–20431. IEEE, 2024b.

Sacha Jungerman, Aryan Garg, and Mohit Gupta. Radiance fields from photons. *ACM Transactions on Graphics*, 45(1):1–16, 2025a.

Sacha Jungerman, Shantanu Gupta, Kaustubh Sadekar, Max Leblang, and Mohit Gupta. Visionsim, 2025b. URL `https://visionsim.readthedocs.io/en/latest/tutorials/large-dataset.html`.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, George Drettakis, et al. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.

Muhammed Kocabas, Jen-Hao Rick Chang, James Gabriel, Oncel Tuzel, and Anurag Ranjan. Hugs: Human gaussian splats. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 505–515. IEEE, 2024.

Dogyoon Lee, Minhyeok Lee, Chajin Shin, and Sangyoun Lee. Dp-nerf: Deblurred neural radiance field with physical scene priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12386–12396, 2023.

Jungho Lee, Donghyeong Kim, Dogyoon Lee, Suhwan Cho, Minhyeok Lee, Wonjoon Lee, Taeoh Kim, Dongyoon Wee, and Sangyoun Lee. Comogaussian: Continuous motion-aware gaussian splatting from motion-blurred images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 26415–26424, 2025.

Da Li, Donggang Jia, Markus Hadwiger, and Ivan Viola. Skinninggs: Editable dynamic human scene reconstruction using gaussian splatting based on a skinning model. *arXiv preprint arXiv:2506.21632*, 2025.

Zhe Li, Zerong Zheng, Lizhen Wang, and Yebin Liu. Animatable gaussians: Learning pose-dependent gaussian maps for high-fidelity human avatar modeling. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19711–19722. IEEE, 2024.

Yehe Liu, Alexander Krull, Hector Basevi, Aleš Leonardis, and Michael Jenkins. bit2bit: 1-bit quanta video reconstruction via self-supervised photon prediction. *Advances in Neural Information Processing Systems*, 37:88443–88485, 2024.

Xianrui Luo, Juewen Peng, Zhongang Cai, Lei Yang, Fan Yang, Zhiguo Cao, and Guosheng Lin. Deblur-avatar: Animatable avatars from motion-blurred monocular videos. *arXiv preprint arXiv:2501.13335*, 2025.

Li Ma, Xiaoyu Li, Jing Liao, Qi Zhang, Xuan Wang, Jue Wang, and Pedro V Sander. Deblur-nerf: Neural radiance fields from blurry images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12861–12870, 2022.

Sizhuo Ma, Shantanu Gupta, Arin C Ulku, Claudio Bruschini, Edoardo Charbon, and Mohit Gupta. Quanta burst photography. *Acm Transactions on Graphics (Tog)*, 39(4):79–1, 2020.

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*, pp. 5442–5451, 2019.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision*, pp. 405–421. Springer, 2020.

Ben Mildenhall, Peter Hedman, Ricardo Martin-Brualla, Pratul P Srinivasan, and Jonathan T Barron. Nerf in the dark: High dynamic range view synthesis from noisy raw images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16190–16199, 2022.

Gyeongsik Moon, Takaaki Shiratori, and Shunsuke Saito. Expressive whole-body 3d gaussian avatar. In *European Conference on Computer Vision*, pp. 19–35. Springer, 2024.

Arthur Moreau, Jifei Song, Helisa Dhamo, Richard Shaw, Yiren Zhou, and Eduardo Pérez-Pellitero. Human gaussian splatting: Real-time rendering of animatable avatars. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 788–798. IEEE, 2024.

Manasi Muglikar, Siddharth Somasundaram, Akshat Dave, Edoardo Charbon, Ramesh Raskar, and Davide Scaramuzza. Event cameras meet spads for high-speed, low-bandwidth imaging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(9):7886–7897, 2025.

Muyao Niu, Yifan Zhan, Qingtian Zhu, Zhuoxiao Li, Wei Wang, Zhihang Zhong, Xiao Sun, and Yinqiang Zheng. Motion-aware animatable gaussian avatars deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 40140–40151, 2026.

Haokai Pang, Heming Zhu, Adam Kortylewski, Christian Theobalt, and Marc Habermann. Ash: Animatable gaussian splats for efficient and photoreal human rendering. In *2024 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 1165–1175. IEEE, 2024.

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 *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13463–13473. IEEE, 2021.

Zhiyin Qian, Shaofei Wang, Marko Mihajlovic, Andreas Geiger, and Siyu Tang. 3dgs-avatar: Animatable avatars via deformable 3d gaussian splatting. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5020–5030. IEEE, 2024.

Jack Saunders, Charlie Hewitt, Yanan Jian, Marek Kowalski, Tadas Baltrusaitis, Yiye Chen, Darren Cosker, Virginia Estellers, Nicholas Gydé, Vinay P Namboodiri, et al. Gasp: Gaussian avatars with synthetic priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 271–280, 2025.

Trevor Seets, Atul Ingle, Martin Laurenzis, and Andreas Velten. Motion adaptive deblurring with single-photon cameras. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1944–1953. IEEE, 2021.

Zhijing Shao, Zhaolong Wang, Zhuang Li, Duotun Wang, Xiangru Lin, Yu Zhang, Mingming Fan, and Zeyu Wang. Splattingavatar: Realistic real-time human avatars with mesh-embedded gaussian splatting. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1606–1616. IEEE, 2024.

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

Pratul P. Srinivasan, Ren Ng, and Ravi Ramamoorthi. Light field blind motion deblurring. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2354–2362, 2017.

Varun Sundar, Tianyi Zhang, Sacha Jungerman, and Mohit Gupta. Quanta neural networks: from photons to perception. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5091–5101. IEEE, 2025.

Jiapeng Tang, Davide Davoli, Tobias Kirschstein, Liam Schoneveld, and Matthias Niessner. Gaf: Gaussian avatar reconstruction from monocular videos via multi-view diffusion. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5546–5558. IEEE, 2025.

Sai Sri Teja, Sreevidya Chintalapati, Vinayak Gupta, T Mukund Varma, Haejoon Lee, Aswin Sankaranarayanan, and Kaushik Mitra. Photonsplat: 3d scene reconstruction and colorization from spad sensors. In *2025 IEEE International Conference on Computational Photography (ICCP)*, pp. 1–17. IEEE, 2025.

Jing Wen, Xiaoming Zhao, Zhongzheng Ren, Alexander G Schwing, and Shenlong Wang. Gomavatar: Efficient animatable human modeling from monocular video using gaussians-on-mesh. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2059–2069. IEEE, 2024.

Xiaoting Yin, Hao Shi, Kailun Yang, Jiajun Zhai, Shangwei Guo, and Kaiwei Wang. E-sem3dgs: Monocular human and scene reconstruction via event-aided semantic 3dgs. *Sensors*, 26(1):188, 2025.

Jinze Yu, Xin Peng, Zhengda Lu, Laurent Kneip, and Yiqun Wang. Spikegs: Learning 3d gaussian fields from continuous spike stream. In *Asian Conference on Computer Vision*, pp. 159–177. Springer, 2024.

Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *CVPR*, 2022.

Jiyuan Zhang, Kang Chen, Shiyan Chen, Yajing Zheng, Tiejun Huang, and Zhaofei Yu. Spikegs: 3d gaussian splatting from spike streams with high-speed camera motion. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 9194–9203, 2024.

Zetong Zhang, Manuel Kaufmann, Lixin Xue, Jie Song, and Martin R Oswald. Odhsr: Online dense 3d reconstruction of humans and scenes from monocular videos. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21824–21835. IEEE, 2025.

Lingzhe Zhao, Peng Wang, and Peidong Liu. Bad-gaussians: Bundle adjusted deblur gaussian splatting. In *European Conference on Computer Vision*, pp. 233–250. Springer, 2024.

A Dataset and Acquisition Details

This section provides additional details on the datasets and implementation of CoHSPAD. We evaluate CoHSPAD on **CoH-Synthesis** and **CoH-Real**, each containing five monocular human SPAD streams captured from a fixed viewpoint. CoH-Synthesis provides temporally aligned clean references for quantitative evaluation, while CoH-Real contains raw observations acquired with a physical SPAD camera. The dataset configurations are summarized in Table A.1.

Table A.1: **Dataset configuration.**

Dataset	Resolution	CFA	Rate	Interval	Duration	Frames	Clean Ref.
CoH-Synthesis	1024×1024	RGGB	1 kHz	1 ms	2 s	2,000	Yes
CoH-Real	1024×1024	BGGR	8 kHz	125 μs	1 s	8,000	No

A.1 CoH-Synthesis

Data Construction. CoH-Synthesis consists of five sequences: Catch, Jog, Jump, Sit, and Walk. We construct the scenes in Blender with VisionSim (Jungerman et al., 2025b), using clothed human assets and textures from BEDLAM (Black et al., 2023) and AGORA (Patel et al., 2021), and motions sampled from AMASS (Mahmood et al., 2019). The renderer provides temporally aligned clean images, human masks, camera parameters, and ground-truth SMPL motion. Ground-truth SMPL is used only for data generation, reconstruction is initialized with GVHMR (Shen et al., 2024), following the same procedure as for CoH-Real. Figure A.1 visualizes all five CoH-Synthesis sequences, including the raw 1-bit SPAD observations, temporally accumulated observations, CoHSPAD reconstructions, and aligned ground-truth sRGB renderings.

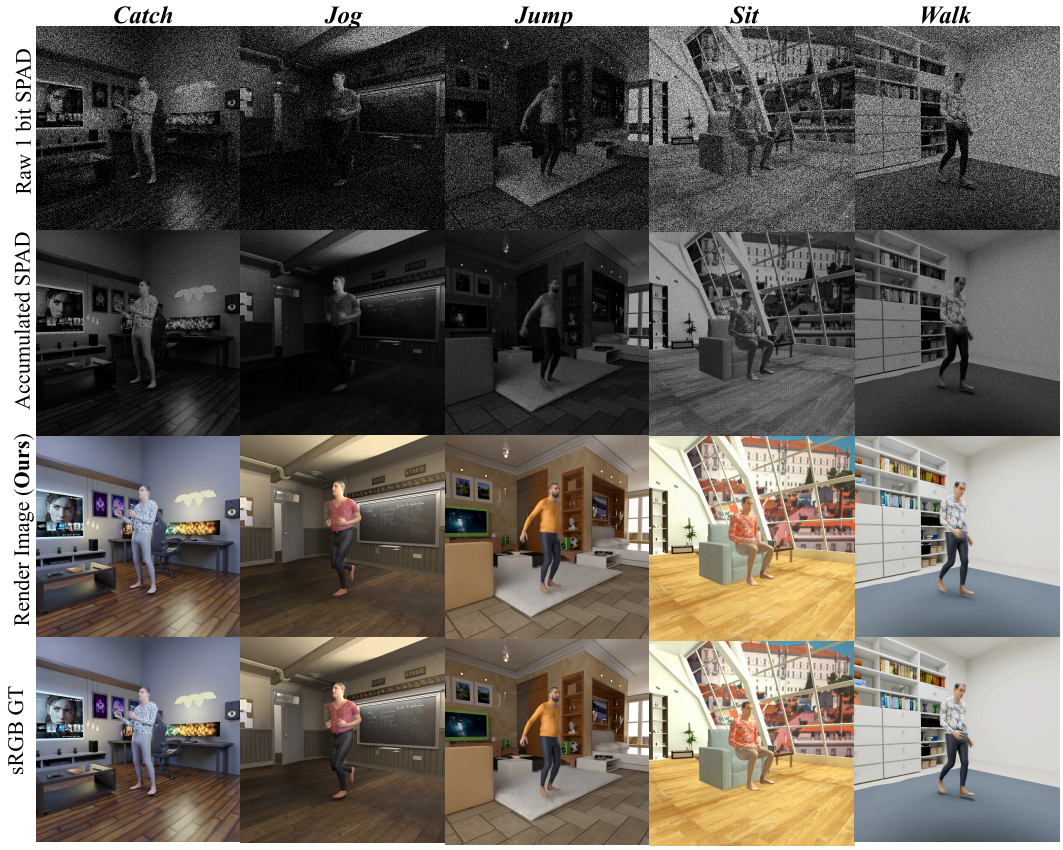

Figure A.1: **CoH-Synthesis visualization.** From top to bottom: raw 1-bit SPAD observations, 40-observation accumulations, CoHSPAD renderings, and ground-truth sRGB renderings.

SPAD Observation Generation. We generate synthetic SPAD observations from the clean Blender renderings using the image formation model described in the main paper with fixed unit exposure. The expected photon count at timestamp t is

$$\lambda_t(\mathbf{p}) = \mathcal{M}_{\mathrm{RGGB}}(\mathbf{I}_t)\,(\mathbf{p}), \tag{A.1}$$

where $\mathcal{M}_{\mathrm{RGGB}}$ applies the RGGB Bayer sampling pattern. We sample $N_t(\mathbf{p}) \sim \mathrm{Poisson}(\lambda_t(\mathbf{p}))$ and form the binary observation as $B_t(\mathbf{p}) = \mathbb{I}[N_t(\mathbf{p}) > 0]$.

A.2 CoH-Real

SPAD Acquisition. CoH-Real consists of five sequences captured with SPADAlpha from a fixed camera. Raw 1-bit BGGR observations are stored in packed binary form and decoded without temporal accumulation for CoHSPAD optimization. Camera intrinsics are estimated by GVHMR, while a fixed camera pose is used throughout each sequence, the resulting camera parameters are stored in COLMAP format. Figure A.2 visualizes all five CoH-Real sequences, showing the raw 1-bit SPAD observations, temporally accumulated observations, and CoHSPAD reconstructions before and after calibrated white balance.

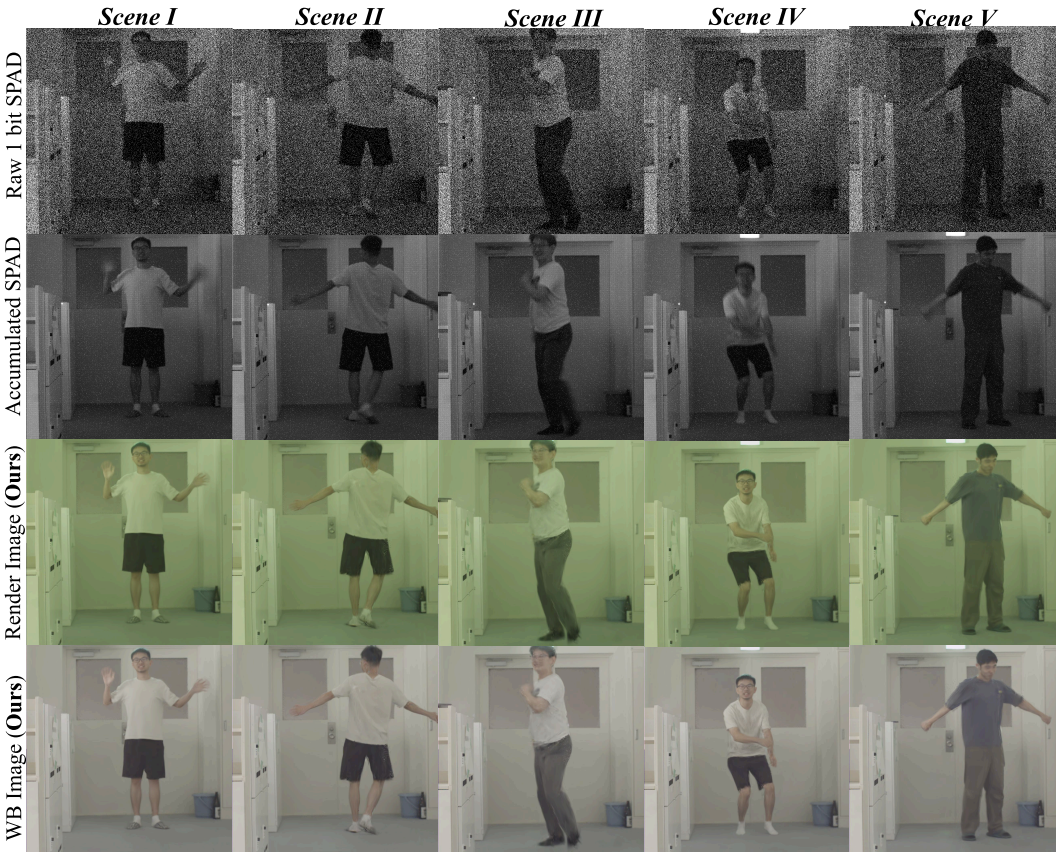

Figure A.2: **CoH-Real visualization.** From top to bottom: raw 1-bit SPAD observations, 320-observation accumulations, CoHSPAD renderings before and after calibrated white balance.

Baseline Preprocessing. For baseline construction, we average K consecutive raw SPAD observations to obtain the empirical detection frequency $\overline{P}_t^{(K)}$. Accumulated-SPAD baselines directly use this representation while preserving the Bayer layout. For pseudo-RGB baselines, we convert the detection frequency to photon intensity by inverting the Bernoulli observation model:

$$\widetilde{\lambda}_t^{(K)}(\mathbf{p}) = -\log\left[1 - \mathrm{clip}\left(\overline{P}_t^{(K)}(\mathbf{p}), 0, 1-\epsilon\right)\right], \tag{A.2}$$

where ϵ prevents numerical instability near unit detection probability. We use $K = 40$ for CoH-Synthesis and $K = 320$ for CoH-Real, corresponding to the same $40\,\mathrm{ms}$ temporal support and

approximately 25 accumulated observations per second. For pseudo-RGB baselines, the resulting Bayer measurements are further demosaiced to obtain RGB inputs. **This temporal accumulation is used only to construct the corresponding baseline inputs, CoHSPAD does not use these accumulated or pseudo-RGB observations as reconstruction targets.**

CoHSPAD Initialization. For CoH-Real, every 20 consecutive raw observations are averaged for initialization. GVHMR (Shen et al., 2024) provides initial SMPL estimates, with joint rotations interpolated in $SO(3)$ via SLERP to the raw SPAD timestamps. Cutie (Cheng et al., 2024) provides human masks for supervision and MTF-Net cropping. Reconstruction remains directly supervised by the raw 1-bit Bayer observations at their original temporal resolution.

White Balance and Visualization. For CoH-Real, we apply the same precalibrated white balance gains to all methods to account for the sensor channel response. The gains remain fixed throughout reconstruction and evaluation and are shared across CoHSPAD and all baselines. For visualization, the reconstructed linear RGB values are white balanced, and tone mapped for display. Direct visualization of the raw linear output can therefore appear overexposed in high-intensity regions, as illustrated in Fig. A.3.

Dark Noise Calibration and Invalid Pixels. We capture fully blocked observations at multiple exposure settings and temporally average them to estimate dark event rates. The resulting bad pixel map contains 32,623 invalid pixels out of 1,048,576, corresponding to 3.111% of the sensor area. The four $H/2 \times W/2$ maps associated with the 2×2 CFA locations are interleaved into a full-resolution validity mask V. Invalid pixels are treated as missing observations and excluded from sensor supervision. For a sensor loss $\ell(\mathbf{p})$ over region S, we compute

$$\mathcal{L}_{\text{valid}} = \frac{\sum_{\mathbf{p}} V(\mathbf{p}) S(\mathbf{p}) \ell(\mathbf{p})}{\sum_{\mathbf{p}} V(\mathbf{p}) S(\mathbf{p}) + \epsilon}, \tag{A.3}$$

where $V(\mathbf{p}) \in \{0, 1\}$ indicates pixel validity and $S(\mathbf{p}) \in \{0, 1\}$ the supervision region. This masking is applied to SPAD reconstruction, MTF-Net supervision, and SPAD-aware densification. Figure A.3 further illustrates the effects of real-data calibration and visualization. Invalid pixels can be absorbed into the Gaussian appearance when included in sensor supervision, while directly displaying the linear reconstruction without tone mapping leads to overexposed highlights. Our final visualization applies the calibrated processing described above.



(a) Bad pixel (b) Raw Linear (c) Ours

Figure A.3: **Real data calibration and visualization.** (a) Bad pixel mask and reconstructions without masking before and after white balance, showing persistent sensor artifacts. (b) Raw linear reconstruction without tone mapping, with overexposed highlights. (c) Final reconstruction with invalid pixel masking and calibrated visualization.

B Implementation Details

CoHSPAD is implemented in PyTorch and optimized separately for each sequence on a single NVIDIA RTX 4090 GPU. We train MTF-Net from scratch with a 2,000 update warm-up, followed by 14,000 reconstruction iterations. At each reconstruction iteration, we supervise $K = 5$ consecutive physical timestamps. We use Adam with learning rates of 2×10^{-5} for the pose trajectory, 10^{-4} for post-LBS deformation, 10^{-3} for MTF-Net, and 10^{-2} for the global exposure parameter.

For the continuous human trajectory, we use $M = 200$ uniformly spaced SMPL pose controls with cubic B-spline interpolation in axis-angle space. The pose residual MLP has two hidden layers with 256 channels and is initialized near zero. We regularize the pose controls using a Huber penalty on their second-order differences and penalize axis-angle magnitudes above $150°$. The post-LBS

deformation field uses a 64-D motion feature and bounds the positional residual to $5\,\mathrm{mm}$. We follow standard 3DGS optimization for the remaining Gaussian parameters. Human densification uses temporal support and gradient variation to suppress isolated gradients, while scene densification uses the structural residual in equation 12.

MTF-Net operates on 128×128 human-centered crops. Its 320-observation long context is represented by 10×4 ordered temporal sub-bins, yielding 40 long-context input channels. The short context uses 20 observations on each side of the target, excluding the target itself. The long and short encoders each produce 32-channel features using a 3×3 convolution, which are fused with normalized temporal encoding by a lightweight convolutional prediction head. We use $A = 9$ blind anchors for self-supervision.

The reconstruction loss weights are $\lambda_{\mathrm{bce}} = 0.8$, $\lambda_{\mathrm{mask}} = 0.05$, $\lambda_{\mathrm{limit}} = 0.02$, and $\lambda_{\mathrm{smooth}} = 0.005$. For MTF-Net, we use $\lambda_{\mathrm{s}} = 0.25$ and $\lambda_{\mathrm{tv}} = 0.01$, where $\mathcal{L}_{\mathrm{tv}}$ is spatial total variation on the predicted probability maps. MTF-Net is optimized independently, and its predictions are detached when used as reconstruction targets. All reported sequences use the same configuration without scene-specific hyperparameter tuning.

C Motion Characteristics of CoHSPAD

We characterize the motion in CoH-Real and CoH-Synthesis from both temporal and spatial perspectives. Since the accumulated SPAD baselines integrate observations over a temporal window of $\Delta t = 40\,\mathrm{ms}$, we measure the projected human motion over the same interval. Let $\mathbf{J}_{t,j}$ denote the 3D position of the j-th SMPL joint at time t, and let $\pi(\cdot)$ denote the camera projection. We define the mean projected joint displacement over Δt as

$$d_t^{\Delta t} = \frac{1}{J} \sum_{j=1}^{J} \left\| \pi(\mathbf{J}_{t+\Delta t,j}) - \pi(\mathbf{J}_{t,j}) \right\|_2, \qquad \Delta t = 40\,\mathrm{ms}, \tag{C.1}$$

where $J = 24$ includes the root joint, and the displacement is measured in image-space pixels. We use the optimized SMPL trajectories for CoH-Real and the ground-truth SMPL motions for CoH-Synthesis. Figure C.1 reports $d_t^{\Delta t}$ for all sequences. Each accumulation window contains 320 raw observations for CoH-Real and 40 for CoH-Synthesis, matching the temporal support of the accumulated SPAD baselines. The curves therefore quantify the amount of human motion occurring within each accumulated observation. The dashed lines indicate the 95th-percentile displacement, while the shaded yellow regions mark the intervals with the largest motion in each sequence. Substantial displacement can occur within a single accumulation window, particularly during fast motion. Figure C.2 further illustrates the spatial variation of human motion. We compute the displacement of body surface points between consecutive timestamps and normalize the displacement magnitudes to $[0, 1]$ for visualization. Motion varies substantially across the body, with rapidly moving regions often coexisting with relatively static ones. Together, Figs. C.1 and C.2 illustrate both the temporal extent and the spatial variation of human motion in our sequences.

D Continuous Motion Analysis

D.1 Pose Refinement and Initialization Robustness

We conduct two analyses to examine the role of SMPL pose optimization in CoHSPAD. We first verify whether the SMPL trajectory is effectively refined by raw SPAD supervision rather than simply inheriting its initialization. We then evaluate the sensitivity of the reconstruction to errors in the initial SMPL pose. Both analyses are performed on CoH-Synthesis, where clean SMPL ground truth is available for evaluation but is not used during reconstruction.

Photon-driven pose refinement. To verify that the SMPL pose is effectively optimized, we compare the initial trajectory with the final trajectory on held-out times. All pose metrics are root aligned to evaluate body-pose accuracy independently of global translation. As shown in Table D.1, the optimized trajectory achieves lower average errors across all pose metrics. The results indicate that raw SPAD supervision provides effective pose refinement rather than simply preserving the initial

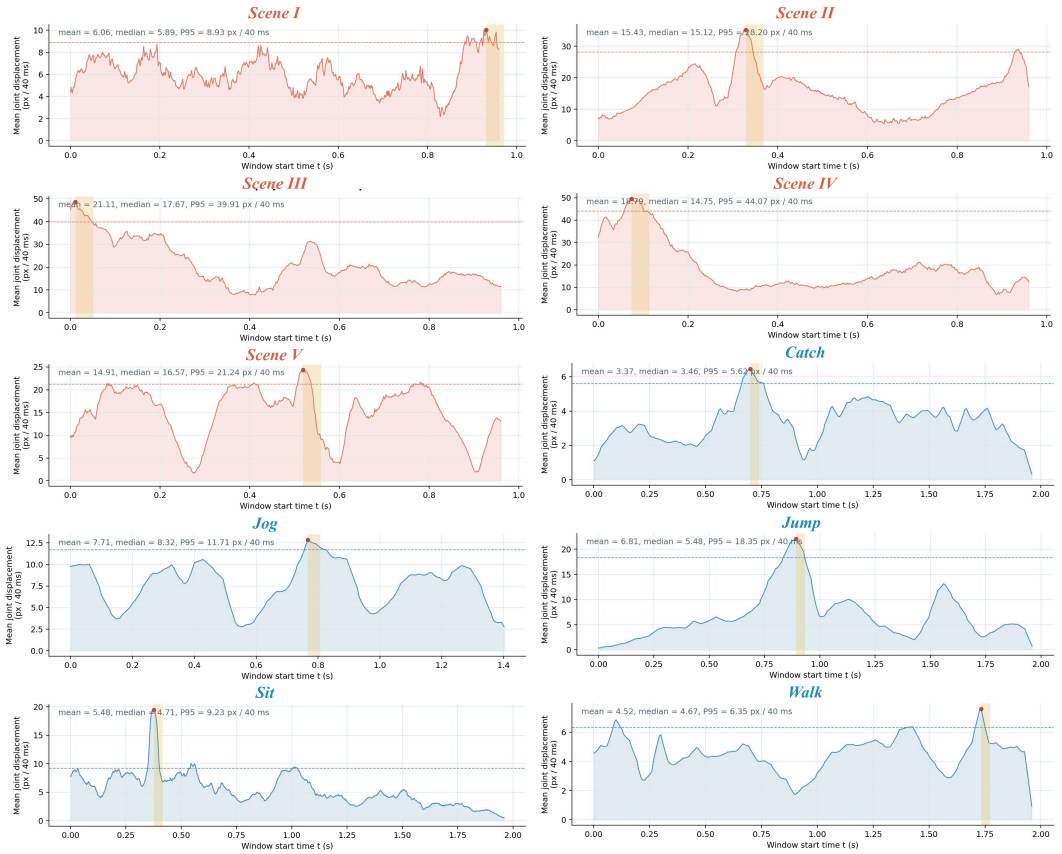

Figure C.1: **Human motion over the accumulation window.** Mean projected SMPL joint displacement over 40 ms windows for five CoH-Real (red) and CoH-Synthesis (blue) sequences. Dashed lines show the 95th percentile, and yellow regions mark intervals with the largest motion.



Figure C.2: **Spatial distribution of human motion.** Normalized body surface displacement for CoH-Real (top) and CoH-Synthesis (bottom), with warmer colors indicating larger motion.

SMPL estimates. Figure D.1 provides a representative visualization of this refinement. The optimized pose corrects discrepancies in the initial alignment and better follows the observed human configuration. The joint-wise gradient response further shows that the raw SPAD reconstruction loss provides direct optimization signals to the SMPL pose parameters.

Table D.1: **SMPL pose refinement on CoH-Synthesis.** Root-aligned pose errors are evaluated against ground truth at held-out timestamps, following the protocol in Table D.4. Lower is better.

SMPL State	Metric	Catch	Jog	Jump	Sit	Walk	Avg.
Init.	MPJPE ↓	27.24	22.44	**25.32**	52.97	25.78	30.75
	PA-MPJPE ↓	19.90	17.31	22.26	37.06	**18.41**	22.99
	PVE ↓	35.46	25.76	**36.59**	66.27	39.69	40.76
Optimized	MPJPE ↓	**21.16**	**18.42**	25.47	**46.24**	**25.62**	**27.38**
	PA-MPJPE ↓	**14.70**	**15.30**	**22.12**	**31.79**	19.80	**20.74**
	PVE ↓	**25.72**	**25.10**	38.02	**61.11**	**38.04**	**37.60**

Table D.2: **Robustness to SMPL pose initialization on CoH-Synthesis.** Body-joint rotations are perturbed before optimization, with results evaluated at held-out timestamps following Table D.4.

	Human Region			Full Image		
Pose Noise	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
0°	**28.795**	**0.904**	**0.122**	**33.744**	**0.925**	**0.247**
$\pm 1^\circ$	28.699	0.903	0.122	33.708	**0.925**	**0.247**
$\pm 3^\circ$	28.139	0.896	0.132	33.446	0.924	0.248
$\pm 5^\circ$	27.269	0.885	0.145	32.937	0.922	0.251



(a) Init. SMPL Pose (b) Optimized SMPL Pose (c) Pose Refinement (d) Joint-Wise Gradient

Figure D.1: **Pose refinement from raw SPAD supervision.** From left to right we show the initial SMPL pose in green and the optimized SMPL pose in blue on the same reconstruction. Their overlay highlights the pose update during optimization. The last panel visualizes the joint-wise gradient response for the 24 SMPL joints under the raw SPAD objective.

Robustness to pose initialization. We perturb each body-joint rotation of the initial SMPL trajectory with independent uniform axis-angle noise in $[-\theta, \theta]$, where $\theta \in 0^\circ, 1^\circ, 3^\circ, 5^\circ$, and evaluate all noise levels under the same reconstruction settings. As shown in Table D.2, reconstruction quality remains stable under small pose perturbations and degrades gradually as the initialization noise increases. Human-region metrics are more affected than full-image metrics, since pose errors primarily influence the dynamic human.

Figure D.2 illustrates the corresponding optimization behavior. Increasing noise progressively displaces the initial body pose, while the optimized pose consistently moves back toward the original human configuration. The pose overlays make this correction directly visible, and the joint-wise gradients show that raw SPAD supervision continues to provide pose optimization signals under perturbed initialization. The final renderings remain visually stable for moderate perturbations, which is consistent with the quantitative trend.

D.2 Continuous Trajectory Parameterization

We further study the interpolation used to recover continuous SMPL poses from sparse temporal pose controls. We compare Catmull–Rom, linear interpolation, SLERP, and B-spline while keeping all other reconstruction settings unchanged. Catmull–Rom interpolates the pose controls with

Table D.3: **Ablation of continuous pose interpolation on CoH-Synthesis.** Human-region reconstruction under different interpolation schemes. Best results are in **bold**.

Interp.	Catch			Jog			Jump			Sit			Walk			Average		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Catmull-Rom	30.149	0.930	0.093	28.669	0.894	0.144	30.139	0.913	0.130	26.033	0.857	0.156	28.424	0.920	0.096	28.683	0.903	0.124
Linear	30.326	0.931	0.091	28.691	0.894	0.142	30.225	0.914	**0.127**	**26.080**	0.857	0.156	28.307	0.918	0.101	28.726	0.903	0.123
SLERP	30.301	0.931	0.092	**28.780**	**0.895**	0.141	30.248	0.914	0.128	26.060	**0.858**	**0.155**	28.433	0.920	0.096	28.764	0.904	0.123
B-spline (Ours)	**30.407**	**0.932**	**0.090**	28.705	0.895	**0.139**	**30.259**	**0.914**	0.128	26.061	0.857	0.157	**28.545**	**0.921**	**0.095**	**28.795**	**0.904**	**0.122**



(a) Init. Pose (b) Noisy Pose (c) Optimized Pose (d) Pose Refinement (e) Joint-Wise Gradient (f) Full Image

Figure D.2: **Pose refinement under perturbed initialization.** Representative examples compare the noisy initial SMPL pose with the pose obtained after raw SPAD optimization. Identical views and timestamps are used for direct comparison.

piecewise cubic curves and passes through the control poses, providing smooth temporal transitions. Linear interpolation directly interpolates between adjacent pose controls, resulting in a piecewise linear trajectory. SLERP performs joint-wise interpolation in quaternion space and follows the shortest rotational path between adjacent controls. B-spline combines multiple neighboring pose controls through smooth basis functions, producing a locally controlled and temporally smooth trajectory.

As shown in Table D.3, all variants achieve comparable reconstruction quality, indicating that our framework is not sensitive to the specific interpolation scheme. B-spline achieves the best overall performance and is therefore adopted in our final model.

D.3 Held-Out Reconstruction

We further evaluate reconstruction at the held-out timestamps used in the main-paper ablations. For CoHSPAD, observations at these timestamps are excluded from optimization, and the corresponding poses are obtained from the continuous trajectory. For comparison at the same timestamps, MAD-Avatar interpolates its intra-exposure motion, while HUGS uses the initial SMPL poses to drive the optimized Gaussians. As shown in Table D.4, CoHSPAD achieves the best average performance on both human-region and full-image reconstruction. The gain is more pronounced in the human region, where the evaluation directly depends on recovering the human state at the held-out times-

Table D.4: Held-out reconstruction on CoH-Synthesis for **(a)** human-region and **(b)** full-image reconstruction. Best and second-best results are highlighted.

		Catch			Jog			Jump			Sit			Walk			Average		
Method	**Input**	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
(a) Human-region reconstruction																			
HUGS	Accum. SPAD	26.740	0.892	0.144	26.088	0.857	0.192	26.178	0.876	0.193	21.852	0.799	0.206	24.299	0.863	0.147	25.031	0.857	0.176
MAD-Avatar	Accum. SPAD	30.078	0.913	0.162	25.188	0.776	0.271	27.731	0.827	0.224	25.501	0.835	0.178	**29.086**	0.894	0.172	27.517	0.849	0.202
HUGS	Restormer RGB	26.116	0.882	0.166	25.896	0.849	0.204	26.040	0.867	0.201	21.269	0.787	0.204	24.179	0.855	0.181	24.700	0.848	0.191
MAD-Avatar	Restormer RGB	28.166	**0.935**	0.145	25.712	0.841	0.212	26.152	0.871	0.196	22.807	0.832	0.176	24.738	0.891	0.162	25.515	0.874	0.178
Ours	**Raw SPAD**	**30.407**	0.932	**0.090**	**28.705**	**0.895**	**0.139**	**30.259**	**0.914**	**0.128**	**26.061**	**0.857**	**0.157**	28.545	**0.921**	**0.095**	**28.795**	**0.904**	**0.122**
(b) Full-image reconstruction																			
HUGS	Accum. SPAD	29.655	0.900	0.310	31.068	0.874	0.419	29.832	0.838	0.370	27.971	0.838	0.333	30.618	0.924	0.303	29.829	0.875	0.347
HUGS	Restormer RGB	29.630	0.904	0.354	30.179	0.867	0.445	29.122	0.829	0.408	26.691	0.824	0.368	30.154	0.930	0.368	29.155	0.871	0.388
Ours	**Raw SPAD**	**34.865**	**0.951**	**0.209**	**33.849**	**0.909**	**0.356**	**32.854**	**0.894**	**0.259**	**32.190**	**0.901**	**0.236**	**34.959**	**0.970**	**0.173**	**33.744**	**0.925**	**0.247**

tamp. Together with the exposure-aligned results in the main paper, this comparison provides a complementary evaluation at held-out temporal states.

E Real-Data Evaluation

CoH-Real has no clean RGB ground truth, which prevents direct image-space evaluation of the real sequences. We therefore analyze whether each method reconstructs the human at the correct image location and whether the learned human Gaussians remain separated from the static background.

The following analyses use two different masks for different purposes. For silhouette evaluation, the reference mask is the Cutie mask used for human supervision in the main experiments, while the predicted mask is obtained by thresholding each method's rendered human alpha. This evaluation therefore measures consistency with the human supervision rather than independent ground truth. For human and scene separation, no Cutie mask is used for visualization. Instead, each method's own rendered human alpha is applied to its full-image rendering to expose the appearance stored in its human representation.

E.1 Silhouette Consistency

We evaluate whether the reconstructed human support agrees with the Cutie masks used for human supervision. At each timestamp t, the predicted mask P_t is obtained by thresholding the rendered human alpha at 0.5, while R_t denotes the corresponding Cutie mask. All metrics are computed at the original 1024×1024 resolution. We evaluate 25 matched timestamps per sequence, corresponding to the centers of the 320-observation windows.

Single-frame consistency. We evaluate foreground overlap, boundary alignment, and foreground area. Mask IoU measures the overlap between the predicted mask P_t and the reference mask R_t:

$$\mathrm{IoU}_t = \frac{|P_t \cap R_t|}{|P_t \cup R_t|}. \tag{E.1}$$

Boundary F1 (BF1) measures contour agreement with a 2-pixel tolerance. We extract one-pixel boundaries by 3×3 erosion and match them after dilation with a 5×5 elliptical kernel:

$$\mathrm{BF1}_t = \frac{2\,\mathrm{Prec}_t\,\mathrm{Rec}_t}{\mathrm{Prec}_t + \mathrm{Rec}_t}. \tag{E.2}$$

Relative Area Error (RAE) measures the relative difference in foreground area:

$$\mathrm{RAE}_t = \frac{\big||P_t| - |R_t|\big|}{|R_t|}. \tag{E.3}$$

Temporal consistency. We further evaluate temporal silhouette changes over the 24 adjacent timestamp pairs in each sequence. Let c_t and A_t denote the centroid and foreground area of a binary mask. Centroid Velocity Error measures the difference in image-space silhouette displacement:

$$E_{\mathrm{centroid},t} = \left\|(c^P_{t+1} - c^P_t) - (c^R_{t+1} - c^R_t)\right\|_2. \tag{E.4}$$

Table E.1: **Silhouette consistency on CoH-Real.** Predicted masks are rendered from each method's human alpha and compared with the Cutie masks used for human supervision. The first three metrics measure single-frame agreement and the last three measure temporal silhouette changes.

		Silhouette			Silhouette Dynamics		
Method	**Input**	**Mask IoU ↑**	**BF1$_{2px}$ ↑**	**Area Err. ↓**	**Centroid Vel. ↓**	**Area Change ↓**	**Motion IoU ↑**
MAD-Avatar	Accum. SPAD	0.5520	0.0148	0.8011	7.9378	0.0223	0.0221
MAD-Avatar	Restormer RGB	0.9132	0.5982	0.0736	5.9052	0.0192	0.4830
HUGS	Accum. SPAD	0.3379	0.0013	2.5031	8.0568	0.0262	0.0109
HUGS	Restormer RGB	0.8856	0.3649	0.0312	5.0152	0.0174	0.3945
Ours	**Raw SPAD**	**0.9469**	**0.6202**	**0.0111**	**2.6027**	**0.0091**	**0.6066**

Log Area Change Error measures the difference in relative foreground-area change:

$$E_{\mathrm{area},t} = \left| \log \frac{A^P_{t+1} + 1}{A^P_t + 1} - \log \frac{A^R_{t+1} + 1}{A^R_t + 1} \right|. \tag{E.5}$$

Motion Region IoU measures whether the predicted and reference silhouettes change in the same image regions. For a binary mask sequence M, we define

$$D_t = \mathrm{Dilate}(M_t \oplus M_{t+1}), \tag{E.6}$$

using a 5×5 elliptical kernel, and compute the IoU between D^P_t and D^R_t. Table E.1 shows that CoHSPAD achieves the best agreement with the Cutie masks in both single-frame support and temporal silhouette changes. Since the Cutie masks are used for human supervision, these metrics measure consistency with the training masks rather than independent reconstruction accuracy.

E.2 Human and Scene Separation

We further analyze the foreground-background decomposition learned by each method. Unlike the silhouette evaluation above, this analysis does not use the Cutie mask for human extraction. For each method, we render the full image and extract the human component using its predicted human alpha. This visualization therefore reflects the appearance attributed to the learned human representation itself. As shown in Fig. E.1, the accumulated-SPAD baselines exhibit less accurate foreground-background separation. MAD-Avatar produces elongated Gaussian artifacts around the human body, indicating that part of the background appearance is represented by the human Gaussians. A similar background leakage is observed for HUGS with accumulated SPAD. In contrast, the Restormer RGB variants substantially reduce these artifacts and yield cleaner human representations. CoHSPAD also maintains a compact human representation with limited background contamination under direct raw SPAD supervision.

This analysis reveals differences in the learned decomposition that are not necessarily apparent from the composite full-image rendering. In particular, CoHSPAD better preserves the separation between dynamic human content and the static scene under raw SPAD supervision, consistent with its explicit human-mask constraint during optimization.

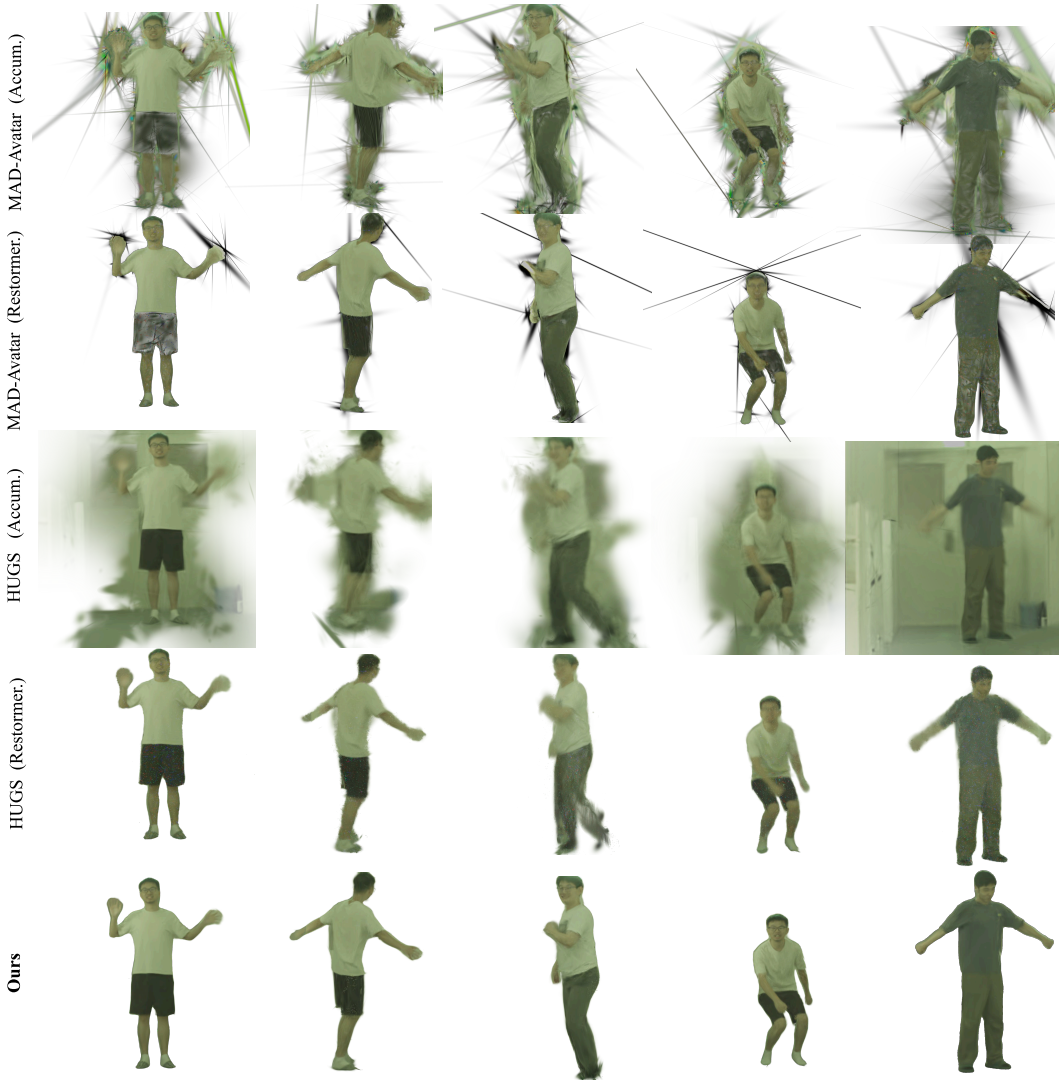

Figure E.1: **Human and scene separation on CoH-Real.** Human appearance is extracted from the full rendering using each method's own rendered human alpha without applying the Cutie mask. Accumulated-SPAD baselines retain more background appearance in the human representation, while Restormer RGB baselines and CoHSPAD produce cleaner separation.