A Appendix / supplemental material

A.1 Implementation Details

Data Preprocessing All sequences are first normalised to a common training resolution of $512{\times}512$ pixels. Following the protocol of BANMo, each 960×720 RGB frame is centre-cropped and downsampled, while its paired 256×192 depth image is bilinearly up-scaled. To stabilise early optimisation, we apply a global scale of 0.2 to both (i) the raw depth values and (ii) the translation component of the ARKit camera extrinsics that initialise the background root pose G_o^t. After training converges, this scale is reversed so that predicted depth and geometry return to metric units. All quantitative evaluations are finally performed on renderings resampled to 480×360 resolution.

Dataset Details Our experiments are conducted on a newly captured dataset comprising 11 sequences recorded with a stereo camera setup at 30fps, featuring diverse scenes with complex interactions between humans and animals. Each sequence is approximately 0.5-1 minutes long, containing between 400 and 900 frames. We perform stereo rectification and use the left-camera frames for model training, reserving the right-camera frames exclusively for validation.

Evaluation Metrics We adopt standard visual quality metrics (LPIPS, PSNR, SSIM) and depth accuracy metrics (Acc@0.1m and RMS depth error). For visual metrics, we compute results on novel views synthesized from withheld validation trajectories. Depth accuracy metrics utilize stereo-derived depth maps as ground truth.

Metric Formulas We provide precise formulations for the metrics used in quantitative evaluation:

- **PSNR**: $\text{PSNR} = 10 \cdot \log_{10}\left(\frac{\text{MAX}_I^2}{\text{MSE}}\right)$, where $\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(I_i - \hat{I}_i)^2$.
- **SSIM**: $\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}$, following standard definitions.
- **LPIPS**: Utilizes a pre-trained neural network to measure perceptual similarity.
- **Acc@0.1m**: Defined as the proportion of predicted depth values within 0.1 meters of the ground truth.
- **RMS depth error**: $\sqrt{\frac{1}{N}\sum_{i=1}^{N}(D_i - \hat{D}_i)^2}$, measuring mean depth deviation.

Deformation Network Initialization Dynamic Gaussian Splatting is notoriously sensitive to its starting configuration: poorly placed Gaussians or mis-estimated skeletal poses readily trap optimisation in severe local minima, producing results that are hardly better than a naïve Deformable-GS baseline. To avoid this collapse we adopt the two–stage scheme described in the main paper: (i) a *neural-SDF pre-fit* jointly refines camera intrinsics, skeletal articulation, and soft deformation; (ii) Gaussians are then sampled on the resulting neural SDF canonical surface and the warping network is continued to be optimized while we switch the objective to dynamic Gaussian splatting. This warm-start supplies accurate joint positions, correct scale, and well-distributed primitives, allowing subsequent learning to focus on fine non-rigid motion rather than coarse alignment. Ablations in Table 5 confirm that removing this initialisation causes up to a 35% drop in PSNR and depth accuracy on articulated human/animal sequences.

Network Architecture For the deformation networks, we adopt multi-layer perceptrons (MLPs) with sinusoidal Fourier features for positional encoding. Specifically, our global and object-root transformations use MLPs with 5 hidden layers, each containing 256 neurons, activated with ReLU functions. The neural soft deformation network, modeled with a flow-based architecture inspired by RealNVP, comprises 4 coupling layers to ensure invertibility.

Training and Optimization We implemented our model using PyTorch and optimized all networks using Adam with an initial learning rate of 10^{-4}, exponentially decayed by a factor of 0.5 every 2,000 iterations. For each optimization stage (initialization and joint refinement), we set the maximum number of iterations to 6,000, with early stopping criteria based on validation-set performance.

Method	PSNR↑	SSIM↑	LPIPS↓	Depth Acc↑	Depth Err↓
Ours (full)	21.31	0.747	0.263	0.901	0.127
w/o initialization	17.30	0.552	0.425	0.742	0.251

Table 5: **Effect of initialization.** Higher is better for PSNR / SSIM / Depth Acc; lower is better for Depth Err.

Computational Cost Our proposed method significantly reduces computational requirements compared to NeRF-based methods. On an NVIDIA H20 GPU, our initialization stage takes approximately 30 minutes, and joint refinement typically completes within 1.5 hours for sequences with around 800 frames. Inference for novel view synthesis operates at interactive frame rates (20fps on average).

Because TOTAL-RECON reports training times on an RTX A6000, we re-ran our training on the same A6000. Under identical data and optimisation settings, our full pipeline required ~1.2 hours, whereas TOTAL-RECON took ~12 hours to reach comparable visual quality, confirming a $\approx 10\times$ speed-up while maintaining (and improving) reconstruction fidelity.

A.2 Additional Visual Qualitative Comparison

Previous work on Dynamic Gaussian Splatting encompasses a variety of architectures and settings. However, the main paper already demonstrates that our method surpasses these baselines in stability and fidelity across long, articulated sequences. Here, we therefore focus on the most competitive prior art, TOTAL-RECON, which similarly targets long-range, high-quality reconstructions. Comprehensive side-by-side renderings and depth maps (7, 8, 9, 10, 11) show that our approach produces sharper silhouettes, fewer temporal artifacts, and consistently lower photometric and geometric error. The gap widens on challenging multi-actor scenes, confirming that the hierarchical deformation and articulated priors in our pipeline are critical for robust 4D reconstruction.

B Limitations and Future Work

Handling Discontinuous Motions Although our model effectively captures continuous articulated motions, handling abrupt discontinuities remains challenging due to our smooth deformation field assumption. Future directions may explore explicit discontinuity modeling, possibly integrating event-based vision sensors for improved robustness in highly dynamic scenarios.

Improved Initialization Exploring advanced initialization methods, potentially leveraging parametric body models (such as SMPL for humans or animal-specific skeletal models), could further enhance reconstruction quality and reduce sensitivity to initialization.

C Broader Impacts

Our method has potential positive impacts in AR/VR applications, enhancing realism in interactive systems. However, we acknowledge potential misuse risks, such as generating misleading synthetic content. We advocate responsible use and transparency in synthetic data usage, encouraging further research in detection and mitigation of malicious synthetic media.

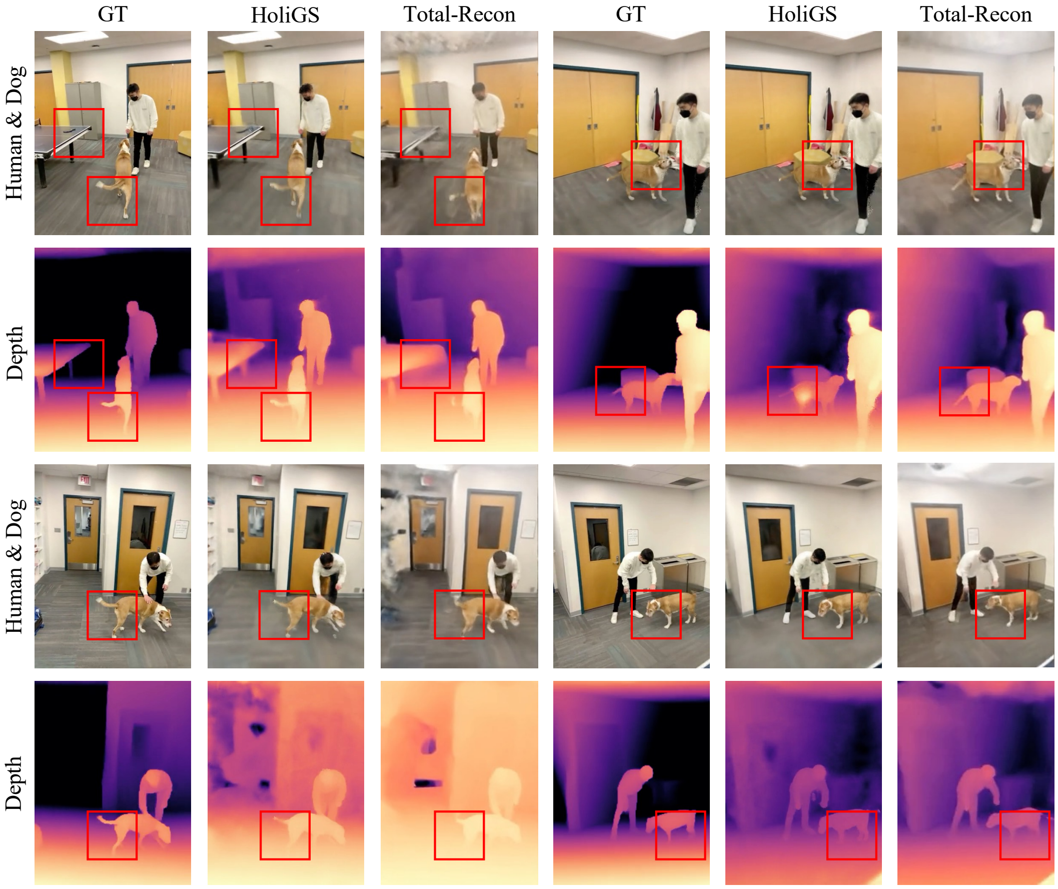

Figure 7: **NVS comparisons with Total-Recon.**

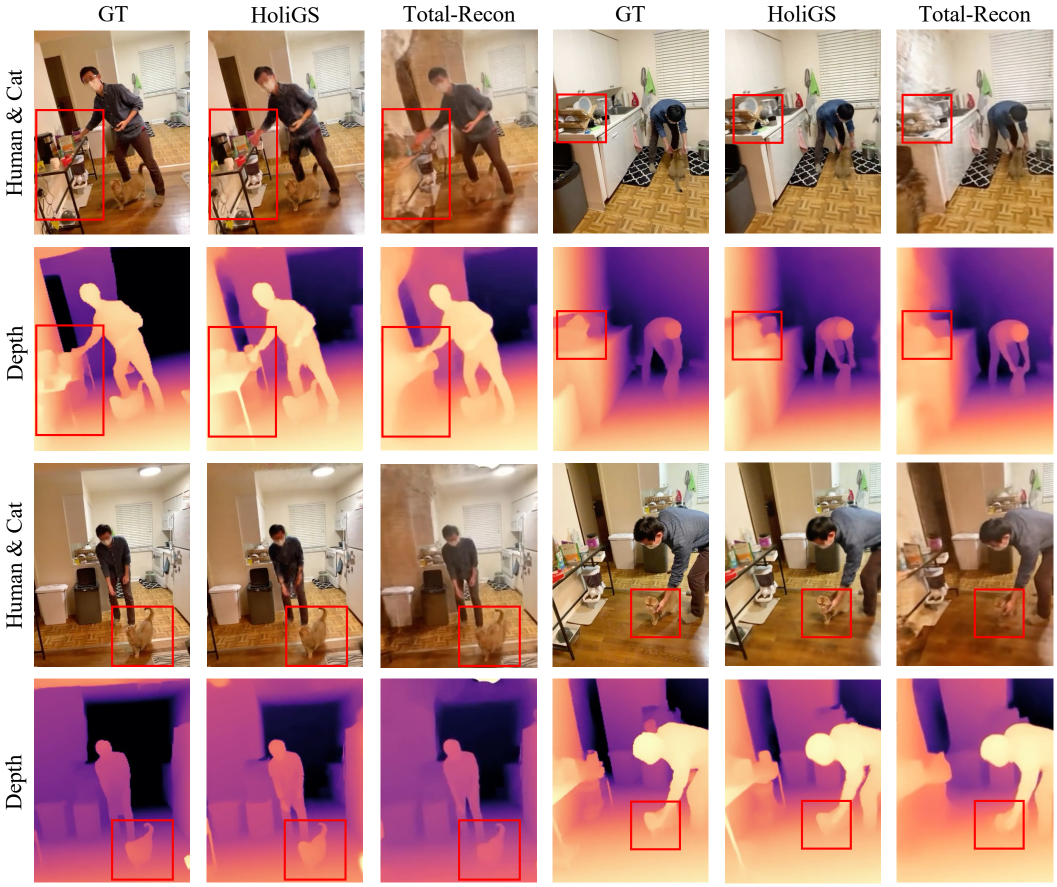

Figure 8: **NVS comparisons with Total-Recon.**

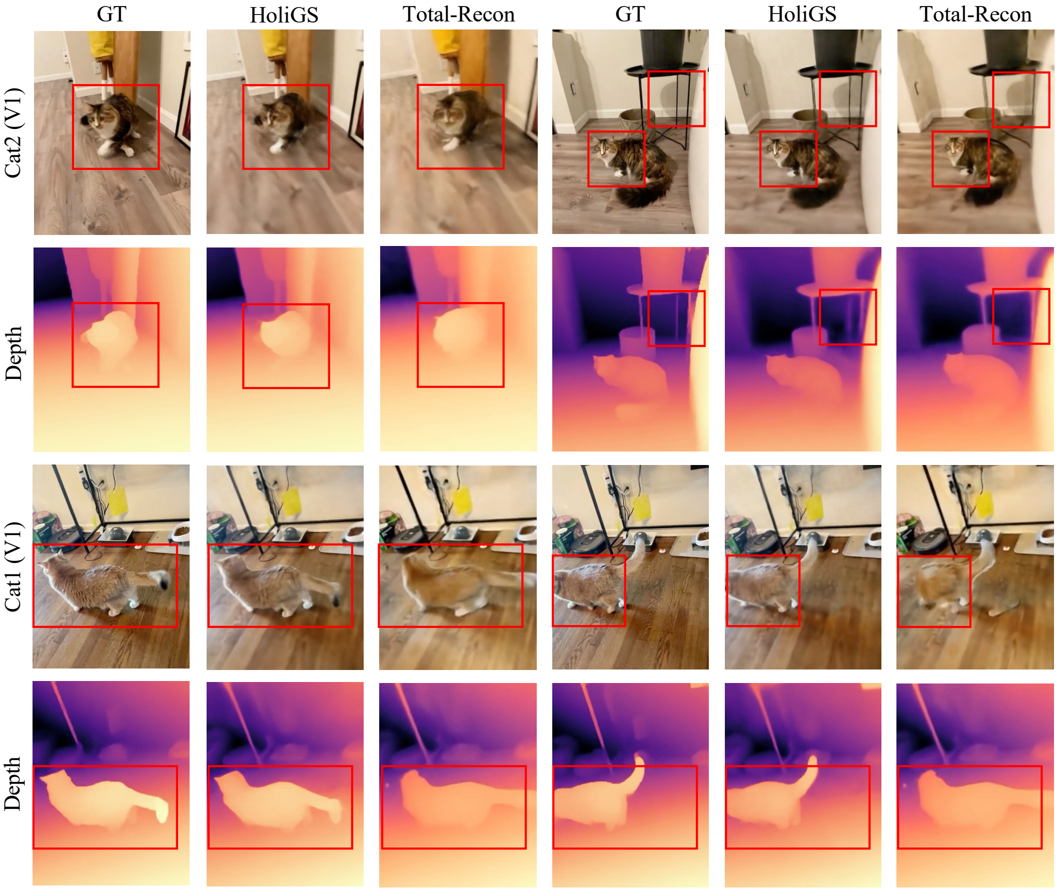

Figure 9: **NVS comparisons with Total-Recon.**

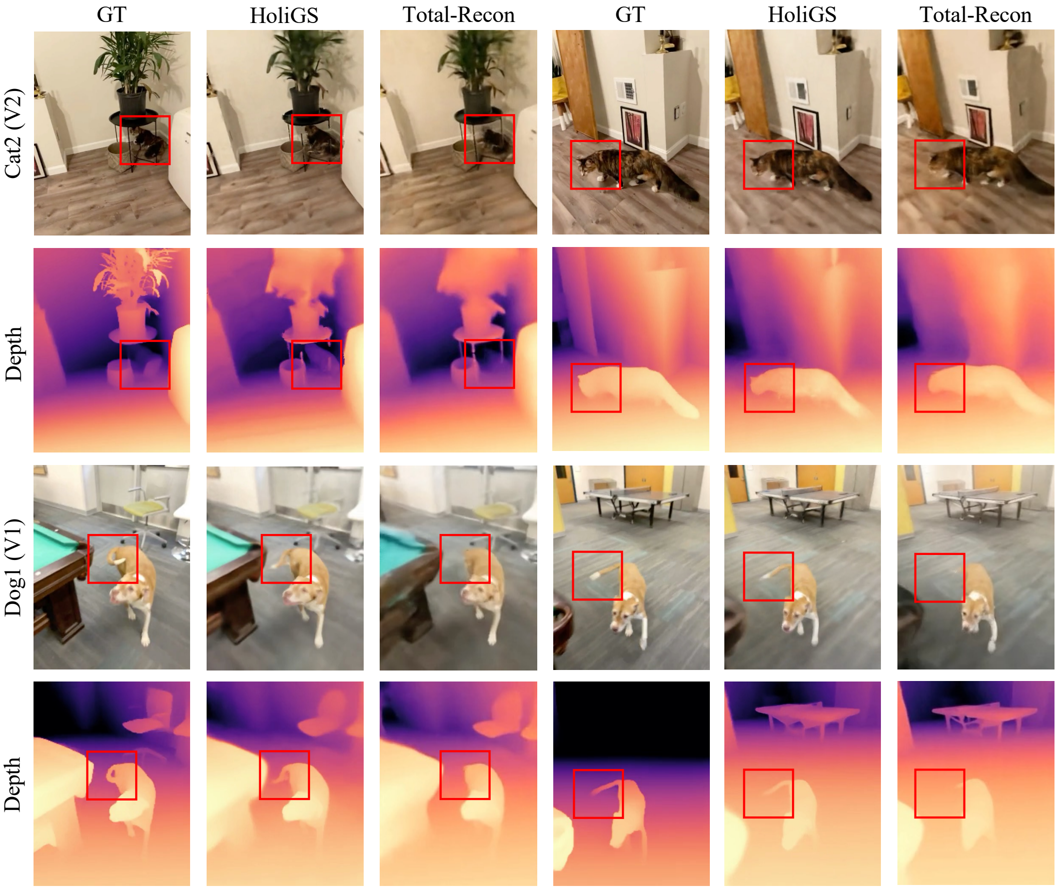

Figure 10: **NVS comparisons with Total-Recon.**

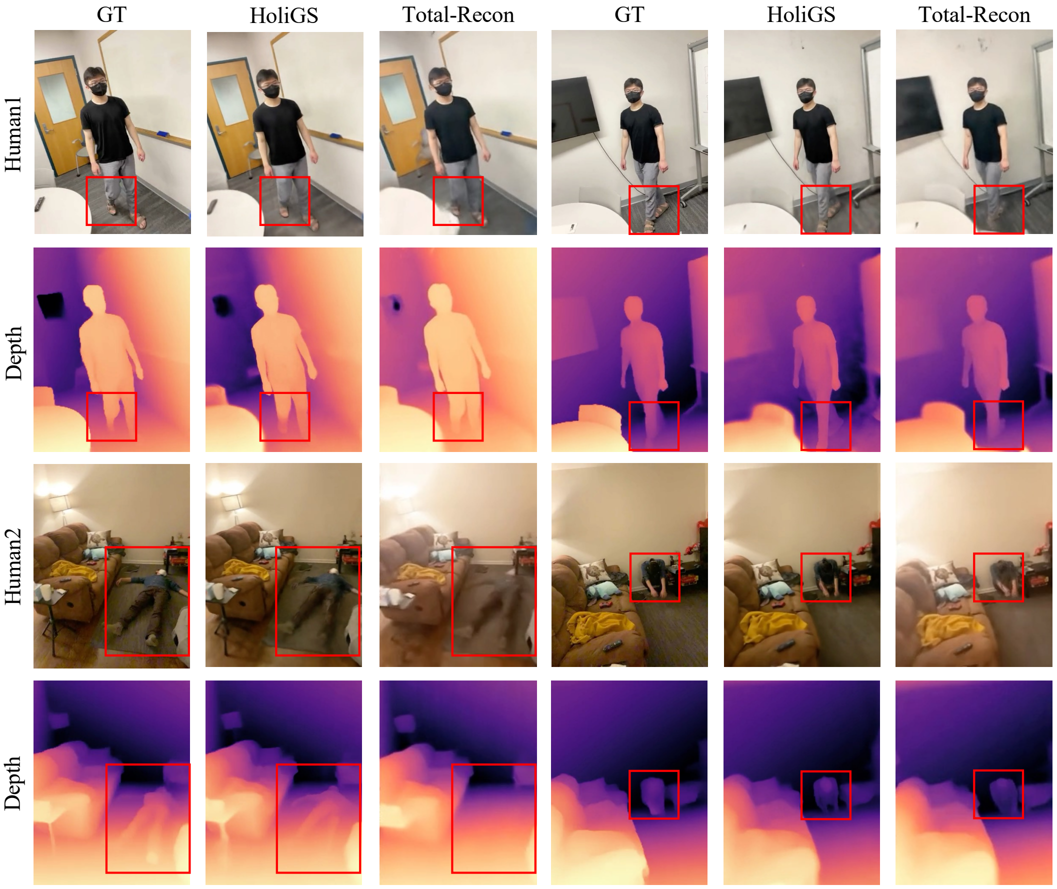

Figure 11: **NVS comparisons with Total-Recon.**