Supplementary Material of GeoMan: Temporally Consistent Human Geometry Estimation using Image-to-Video Diffusion

Gwanghyun Kim[1,2*], Xueting Li[1], Ye Yuan[1], Koki Nagano[1], Tianye Li[1],
Jan Kautz[1], Se Young Chun[2], Umar Iqbal[1]
[1]NVIDIA [2]Seoul National University
https://research.nvidia.com/labs/dair/geoman

In this documents, we provide additional results and details. We highly encourage readers to view our supplementary video introduced in our project page. We include more experimental and evaluation details in Sec. A. Additional results, including geometry estimation on long videos and multi-person videos, quantitative evaluation on the Goliath dataset [S8] and more ablation studies are present in Sec. B. Finally, we discuss the limitations of GeoMan and future works in Sec. C.

A. Experimental Details

A.1. Implementation Details

We implemented the Image Geometry Diffusion (I2G) model based on the pretrained weights of Stable Diffusion 2 and the Video Geometry Diffusion (V2G) model using the pretrained weights of I2VGen-XL [S12], initializing Video ControlNet with CtrlAdapter [S7], and leveraging the `diffusers` library [S9].

As shown in Fig. 2 of the main paper, the VAE encoder-decoder and Video ControlNet remain frozen, while only the diffusion denoiser, V2G, is fine-tuned.

Note that we learn separate diffusion models for image-based normal and depth map estimation. This is because predicting depth and normal from RGB images requires a diffusion model to map an RGB image to two significantly different outputs, thus separate models outperform a single model.

Both models are trained at 512×512 resolution using the Adam optimizer [S6] with a learning rate of $1e{-}5$. I2G is trained for 20K iterations with a batch size of 144, taking approximately two days on 4 NVIDIA A100 GPUs. V2G is trained on 12-frame sequences for 30K iterations with a batch size of 8, requiring approximately three days on 8 A100 GPUs. Evaluation is performed on 32-frame inputs for moving subject videos and 16-frame inputs for moving camera videos. For all experiments, we set the number of denoising steps to 100 and use an ensemble size of 8.

We preprocess input images by removing backgrounds, cropping the human region, and resizing images to a resolution of 512×512. For inference on in-the-wild videos, we employ video matting method, BiRefNet [S13], to remove backgrounds. This preprocessing ensures that our model focuses solely on human-centric geometry while maintaining robustness across diverse scenarios. To leverage the static 3D scans in THuman-2.0 for training the V2G model, we transform multiview images into simulated videos that mimic camera rotation.

A.2. Evaluation Details

Evaluation Dataset The original ActorsHQ dataset includes eight actors performing 15 sequences, captured with 160 cameras at 25 fps. For the evaluation, we generate 192 videos of moving subjects with static cameras (eight actors, 24 cameras, 32 frames per video) at various body scales. we also construct 256 videos with moving cameras and static subjects (32 cameras, 16 frames per video).

Depth Alignment For 'Optimizing shift' in the comparison of depth estimation methods, we optimized the shift per frame with the ground truth depth maps . For 'Optimizing scale + shift,' we optimized the scale per sequence following [S2] and optimized the shift per frame.

Point Cloud Reconstruction Root depth is estimated using a 3D pose estimation method [S11] to convert predictions into metric depth. Point cloud reconstruction is performed using the dataset's camera intrinsic parameters. While we present un-

*Work done during an internship at NVIDIA.

aligned reconstructions, we also align depth predictions to ground truth using an optimized shift to highlight scale differences, as shown in Figs. S5–S8.

Fine-tuning Sapiens on Our Training Dataset For the fine-tuning of Sapiens*, we use the pretrained Sapiens-2B model, which serves as the foundation model trained on a large-scale human dataset. We fine-tune this model specifically for the normal and root-relative depth estimation using our training dataset, following their fine-tuning pipeline [1].

A.3. Metrics

For evaluating the predicted depth, we follow [S1] and use the following metrics:

$$
\begin{aligned}
&\text{Abs Relative: } \frac{1}{\sum(K_t==1)} \sum_{k_t \in K, d_t \in D_t} k_t \left\| \frac{d_t - d_t^{gt}}{d_t^{gt}} \right\| \\
&\text{Squared Relative: } \frac{1}{\sum(K_t==1)} \sum_{k_t \in K, d_t \in D_t} \frac{\left\| d_t - d_t^{gt} \right\|^2}{d_t^{gt}} \\
&\text{RMSE (linear): } \sqrt{\frac{1}{\sum(K_t==1)} \sum_{k_t \in K, d_t \in D_t} \left\| d_t - d_t^{gt} \right\|^2} \\
&\text{RMSE (log): } \sqrt{\frac{1}{\sum(K_t==1)} \sum_{k_t \in K, d_t \in D_t} \left\| \log d_t - \log d_t^{gt} \right\|^2} \\
&\delta < \text{thr: } \frac{1}{\sum(K_t==1)} \sum_{k_t \in K, d_t \in D_t} K_t \left[\text{Max}\left(\frac{D_t}{D_t^{gt}}, \frac{D_t^{gt}}{D_t} \right) < \text{thr} \right] \\
&\delta 1\text{: } \frac{1}{\sum(K_t==1)} \sum_{k_t \in K, d_t \in D_t} K_t \left[\text{Max}\left(\frac{D_t}{D_t^{gt}}, \frac{D_t^{gt}}{D_t} \right) < 1.25 \right] \\
&\text{SI-log10: } \sqrt{\frac{1}{\sum(K_t==1)} \sum_{k_t \in K, d_t \in D_t} \left\| \log_{10}(d_t) - \log_{10}(d_t^{gt}) \right\|^2}
\end{aligned}
\tag{S1}
$$

where K_t is a depth validity mask, D_t is the predicted depth for image I_t, and D_t^{gt} is the ground-truth depth.

For evaluating the predicted normal maps, we use the metrics from Sapiens. The angular error is computed as the mean and median angular errors between predicted normal vectors and ground-truth normal vectors:

$$
\text{Angular Error} = \frac{1}{n} \sum_{i=1}^{n} \cos^{-1}\left(\frac{\mathbf{n}_i \cdot \mathbf{n}_i^{gt}}{|\mathbf{n}_i||\mathbf{n}_i^{gt}|} \right) \tag{S2}
$$

Additionally, we compute the percentage of pixels where the angular error is below specific thresholds:

$$
\text{Percentage of Pixels with Angular Error} < t^\circ = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\left(\cos^{-1}\left(\frac{\mathbf{n}_i \cdot \mathbf{n}_i^{gt}}{|\mathbf{n}_i||\mathbf{n}_i^{gt}|} \right) < t \right) \tag{S3}
$$

where $t \in \{11.5^\circ, 20^\circ\}$.

For evaluating temporal consistency, we introduce optical flow-based metrics. The optical flow-based warping metric (OPW) is defined by [S10] as:

$$
OPW = \frac{1}{n} \sum_{i=1}^{n} \sum_{t=0}^{T-1} W_{t+1 \Rightarrow t}^{(i)} \left\| D_{t+1}^{(i)} - \hat{D}_t^{(i)} \right\|_1, \tag{S4}
$$

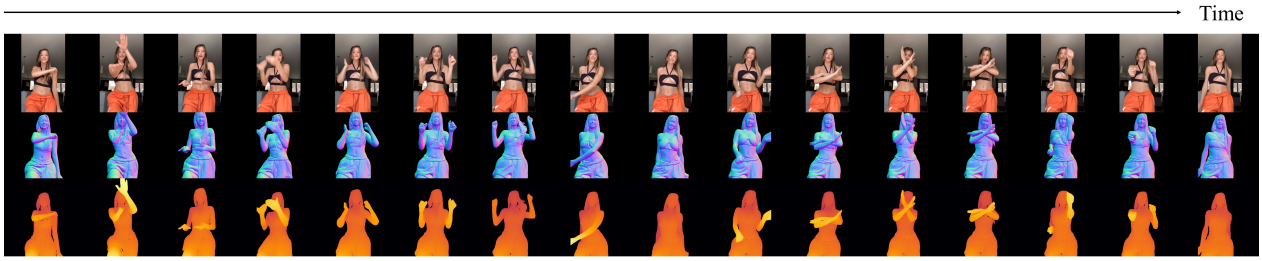
where $W_{t+1 \Rightarrow t}^{(i)}$ is the optical flow-based visibility mask calculated from the warping discrepancy between subsequent frames, as explained in [S10]. We use the optical flow generated by the latest SOTA FlowFormer [S3]. This metric is applied to both depth and normal frames, where the depth and normal values are warped between consecutive frames.

To further encourage comprehensive evaluation of temporal consistency, we introduce additional metrics. For normal frames, we use the following:

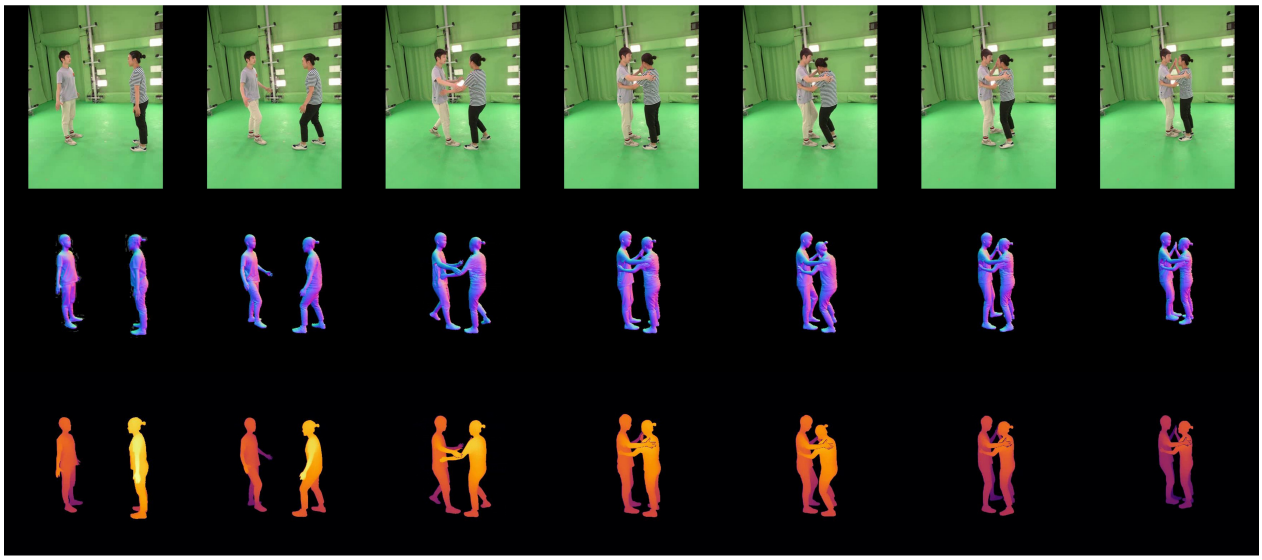
$$
\text{TC-Mean (Optical Flow-based Angular Error)} = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{T-1} \sum_{t=0}^{T-1} W_{t+1 \Rightarrow t}^{(i)} \cos^{-1}\left(\frac{\hat{\mathbf{n}}_t^{(i)} \cdot \mathbf{n}_{t+1}^{(i)}}{|\hat{\mathbf{n}}_t^{(i)}||\mathbf{n}_{t+1}^{(i)}|} \right) \tag{S5}
$$

$$
\text{TC-11.25}^\circ \text{ (Optical Flow-based Angular Error)} = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{T-1} \sum_{t=0}^{T-1} W_{t+1 \Rightarrow t}^{(i)} \mathbb{I}\left(\cos^{-1}\left(\frac{\hat{\mathbf{n}}_t^{(i)} \cdot \mathbf{n}_{t+1}^{(i)}}{|\hat{\mathbf{n}}_t^{(i)}||\mathbf{n}_{t+1}^{(i)}|} \right) < 11.25^\circ \right) \tag{S6}
$$

[1] https://github.com/facebookresearch/sapiens



(a) Geometry estimation on a long video



(b) Geometry estimation on a multi-human video

Figure S1. Results of geometry estimation on long and multi-person videos.

where $\hat{\mathbf{n}}_t^{(i)}$ is the warped normal for the i-th sample at frame t, and $\mathbf{n}_{t+1}^{(i)}$ is the depth from the next frame.

For depth frames, we define the following temporal consistency metric:

$$\text{TC-RMSE (Optical Flow-based Temporal Consistency)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{T-1}\sum_{t=0}^{T-1} W_{t+1\Rightarrow t}^{(i)}\left\|\hat{D}_t^{(i)} - D_{t+1}^{(i)}\right\|^2} \tag{S7}$$

$$\text{TC-}\delta_1\text{ (Optical Flow-based Depth Consistency)} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{T-1}\sum_{t=0}^{T-1} W_{t+1\Rightarrow t}^{(i)}\mathbb{I}\left(\max\left(\frac{\hat{D}_t^{(i)}}{D_{t+1}^{(i)}}, \frac{D_{t+1}^{(i)}}{\hat{D}_t^{(i)}}\right) < 1.25\right) \tag{S8}$$

where $\hat{D}_t^{(i)}$ is the warped depth for the i-th sample at frame t, and $D_{t+1}^{(i)}$ is the depth from the next frame.

B. Additional Qualitative and Quantitative Results

B.1. Geometry Estimation on Long Videos and Multi-Person Videos

Our model effectively generalizes to long video sequences and multi-person scenarios (Fig. S1). Despite being trained on fixed-length sequences, it supports inference on 64-frame sequences using an NVIDIA A100, leveraging robust temporal modeling and extending further via autoregressive inference.

To ensure smooth transitions across segments, we adopt the mortise-and-tenon-style latent interpolation from [S2]. Overlapping frames between segments are interpolated with linearly decreasing weights, and the final depth and normal sequences are reconstructed by decoding the stitched latent representations via the VAE decoder.

For multi-person scenarios, our approach remains effective despite training on single-person data. Normal estimation requires no modification, as pretrained priors enable natural adaptation. For depth estimation, we address root-relative ambiguity via per-subject masking and result aggregation. Human masks segment individuals, root-relative depths are estimated, and metric depths are reconstructed via 3D pose estimation, ensuring globally consistent depth while preserving per-subject geometric integrity.

B.2. Qualitative Results on In-the-Wild Videos

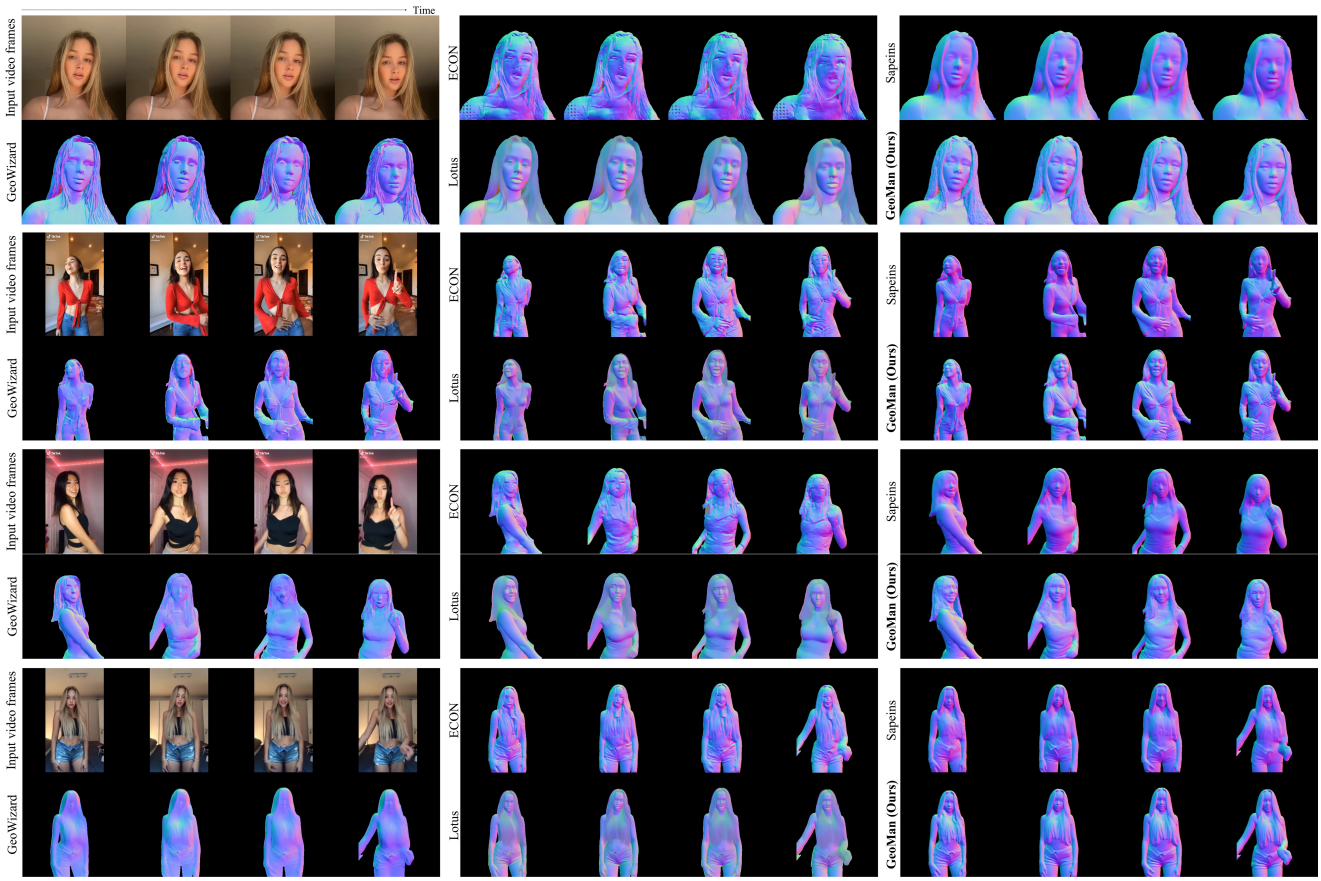


Figure S2. Comparison on zero-shot normal estimation results on in-the-wild videos.

To evaluate generalization to in-the-wild videos, we perform a qualitative comparison on short-form videos collected in Champ [S14], which compiles content from platforms such as Bilibili, Kuaishou, TikTok, YouTube, and Xiaohongshu. This dataset includes a diverse range of individuals in full-body, half-body, and close-up shots, with varying ages, ethnicities, genders, and backdrops.

As shown in Fig. S2 and S3, our results demonstrate robust performance and high-quality geometry in these in-the-wild videos.

B.3. Quantitative Results on Goliath Dataset

We provide additional quantitative results on the Goliath [S8] dataset, a multiview video dataset designed for studying complete avatars, including full-body geometry and underlying body shape. Each instance consists of eight captures of the same person, with four sequences featuring four subjects in full-body capture. For our evaluation, we used 10 views with 32 frames as the test set.

As shown in Table S2 and S1, while GeoMan significantly outperforms other methods, the pre-trained version of Sapiens [S5] remains competitive. However, it is important to note that Sapiens was trained on a proprietary dataset. To ensure a

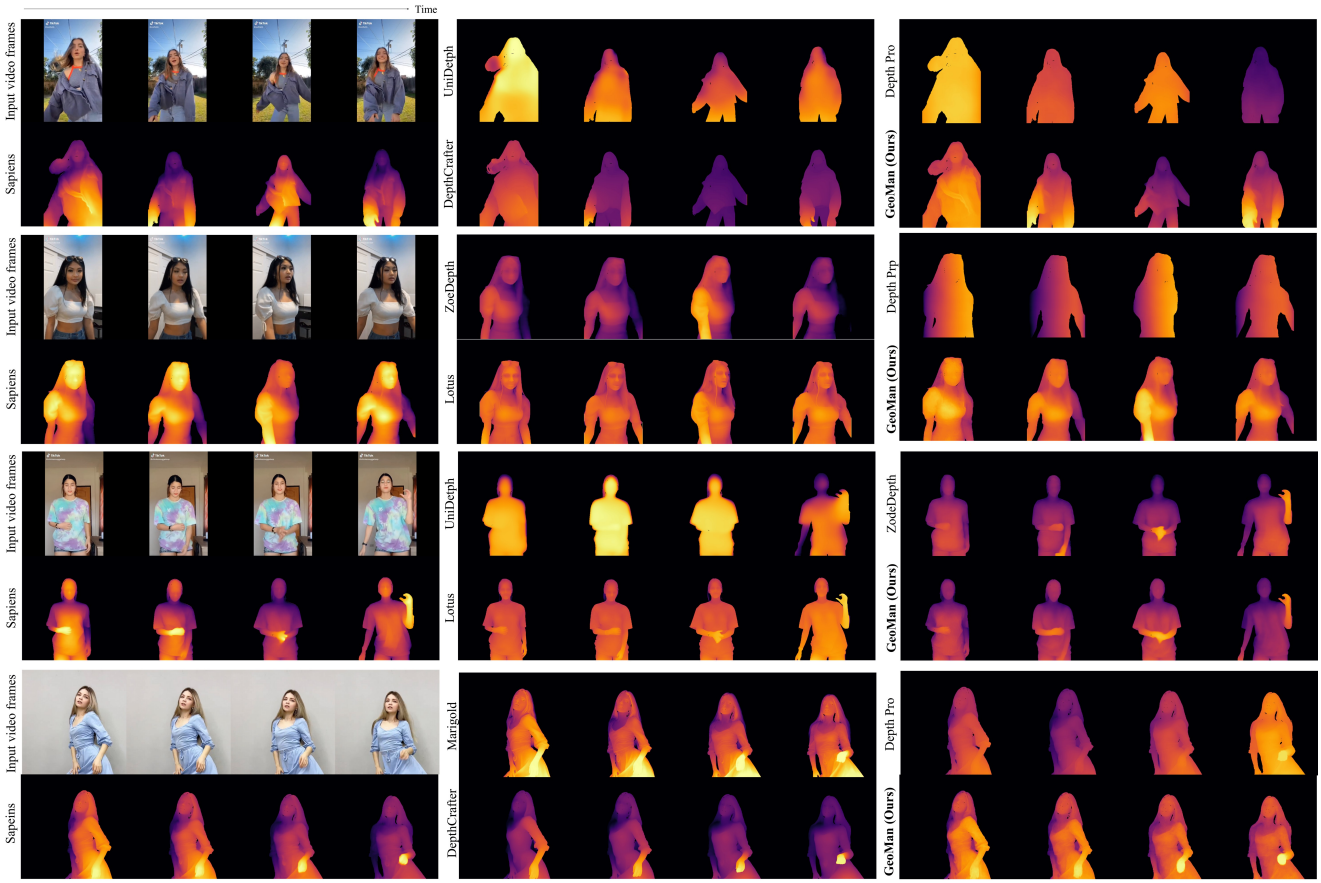


Figure S3. Comparison on zero-shot depth estimation results on in-the-wild videos.

	Performance on moving subject videos						Temporal consistency		
	Optimizing shift			Optimizing scale + shift					
	AbsRel↓	δ <1.05↑	δ1↑	AbsRel↓	δ <1.05↑	δ1↑	OPW↓	TC-RMSE↓	TC-δ1↑
Sapiens	0.030	0.807	**1.000**	0.007	0.994	**1.000**	0.077	0.024	0.754
Sapiens*	0.017	0.954	**1.000**	0.011	0.986	**1.000**	0.064	0.022	0.855
Marigold	0.022	0.900	**1.000**	0.020	0.904	**1.000**	0.098	0.031	0.605
GeoWizard	0.018	0.933	**1.000**	0.015	0.955	0.999	0.061	0.024	0.709
Lotus	0.023	0.888	**1.000**	0.022	0.898	**1.000**	0.097	0.031	0.680
Depthcrafter	0.015	0.957	**1.000**	0.011	0.985	**1.000**	0.039	0.019	**0.919**
Depth any video	0.015	0.960	**1.000**	0.013	0.979	0.999	0.044	0.020	0.866
GeoMan (Ours)	**0.011**	**0.987**	**1.000**	0.010	0.989	**1.000**	**0.038**	**0.018**	0.909

Table S1. Comparison with affine-invariant depth estimation methods on zero-shot settings on Goliath. Sapiens*: Finetuned on our dataset.

fair comparison, we fine-tuned Sapiens's pre-trained MAE model on our dataset. Under this fair setting, GeoMan outperforms Sapiens by a notable margin.

B.4. Additional Results on ActorsHQ

Fig. S4 presents additional qualitative comparisons in surface normal estimation, demonstrating that our method not only achieves superior accuracy over baselines but also ensures greater temporal consistency and higher fidelity.

Figs. S5–S10 provide further qualitative comparisons with baseline methods forboth moving camera and moving subject videos. To ensure consistency, all predicted depth maps are renormalized using sequence-wise min-max scaling within the human mask. Compared to per-frame estimation models, our approach produces more temporally stable results.

For point cloud reconstruction, we present both unaligned reconstructions and those with depth aligned to ground truth

	Performance on moving subject videos				Temporal consistency		
	Mean↓	Median↓	11.25°↑	30°↑	OPW↓	TC-Mean↓	TC-11.25°↑
ECON	25.498	20.129	21.800	72.345	0.196	16.280	53.385
Sapiens	**11.205**	**8.856**	**64.284**	**96.091**	0.104	9.653	77.615
Sapiens*	13.531	10.605	53.533	93.712	0.113	10.812	74.594
GeoWizard	19.622	16.167	30.789	82.829	0.190	18.037	41.474
Lotus	17.999	14.766	35.198	86.666	0.104	10.155	75.924
ICLight	27.664	24.571	14.084	63.199	0.169	14.513	56.882
GeoMan (Ours)	12.831	10.034	56.665	94.308	**0.084**	**8.128**	**82.811**

Table S2. Comparison on zero-shot normal estimation on Goliath. Sapiens*: Fine-tuned on our dataset.

	Normal					Depth					
	Angular difference		%within t°	Temporal consistency		Optimizing shift		Optimizing scale + shift		Temporal consistency	
	Mean↓	Median↓	11.25°↑	OPW↓	TC-Mean↓	AbsRel↓	δ <1.05↑	AbsRel↓	δ <1.05↑	OPW↓	TC-RMSE↓
(a) Naïve extension vs GeoMan fine-tuning											
Naïve-5K-steps	25.305	22.732	15.602	0.076	7.458	0.038	0.720	0.028	0.856	0.049	0.019
Naïve-10K steps	23.550	20.839	18.601	0.071	6.938	0.026	0.866	0.022	0.907	0.046	0.017
Naïve-20K steps	21.137	18.435	22.488	0.070	6.855	0.021	0.915	0.018	0.939	0.040	0.016
Naïve-30K steps	19.669	16.628	28.405	0.072	7.169	0.020	0.926	0.017	0.943	0.040	0.016
Naïve-50K steps	20.307	17.131	27.754	0.075	7.274	0.017	0.945	0.016	0.958	0.041	0.535
GeoMan-5K steps	17.499	13.541	40.340	0.069	6.781	0.014	0.964	0.013	0.970	0.036	**0.014**
GeoMan-10K steps	17.255	13.500	40.094	0.071	6.881	0.012	0.975	0.012	0.981	0.033	**0.014**
GeoMan-20K steps	16.548	12.719	43.483	**0.066**	**6.502**	**0.012**	0.977	**0.011**	**0.984**	**0.032**	**0.014**
GeoMan-30K steps	16.185	12.334	45.217	0.070	6.876	**0.012**	**0.978**	**0.011**	**0.984**	**0.032**	**0.014**
GeoMan-50K steps	**16.105**	**12.144**	**45.995**	0.067	6.599	**0.012**	0.975	**0.011**	0.983	0.033	0.015
(b) I2G vs V2G											
I2G	**16.033**	**12.068**	**46.222**	0.118	11.261	0.013	0.970	0.013	0.974	0.040	0.016
I2G+V2G	16.185	12.334	45.217	**0.070**	**6.876**	**0.012**	**0.978**	**0.011**	**0.984**	**0.032**	**0.014**
(c) Multimodality of I2G											
Unimodal	**16.033**	**12.068**	**46.222**	0.118	11.261	0.013	0.970	0.013	0.974	0.040	0.016
Multimodal	18.442	14.452	36.985	0.086	9.930	0.017	0.946	0.016	0.956	0.040	0.017
(d) Training data source of V2G											
Only 3D data	24.068	20.184	21.938	0.107	[9.966	0.023	0.883	0.024	0.883	0.048	0.017
Only 4D data	16.387	12.441	44.787	**0.062**	**6.198**	0.013	0.967	0.012	0.980	0.035	0.015
Combined	**16.185**	**12.334**	**45.217**	0.070	6.876	**0.012**	**0.978**	**0.011**	**0.984**	**0.032**	**0.014**
(e) Effectiveness of human area crop											
w/o Human area crop	16.821	12.891	42.801	**0.070**	6.974	**0.012**	0.977	0.012	0.982	0.037	0.015
w Human area crop	**16.185**	**12.334**	**45.217**	**0.070**	**6.876**	**0.012**	**0.978**	**0.011**	**0.984**	**0.032**	**0.014**
(f) Source of the First Frame for V2G											
1/4 Downscaled I2G	17.843	13.510	40.593	0.072	8.118	0.016	0.964	0.015	0.968	0.040	0.016
I2G	16.185	12.334	45.217	**0.070**	**6.876**	0.012	0.978	0.011	0.984	**0.032**	**0.014**

Table S3. Additional ablation studies. We validate the impact of various design choices.

using an optimized shift to emphasize scale differences. Our method effectively preserves human scale, demonstrating strong geometric consistency.

B.5. Additional Ablation Studies and Analysis

Naïve extension vs. GeoMan Fine-Tuning. In our experiments, we rigorously compare the performance of our proposed image-to-video formulation of GeoMan against a naïve extension of [S4, S2]. Despite training for further 50K steps, the naïve extension strategy consistently yields suboptimal results across multiple metrics. As highlighted in Tab. S3(a), the performance gap remains significant, reinforcing the superior efficiency and effectiveness of our GeoMan approach in handling complex video generation tasks. This stark contrast in performance emphasizes the importance of a more structured, model-specific fine-tuning strategy, which our GeoMan method embodies, leading to more robust and reliable results.

Fine-Tuning Video ControlNet vs. Fine-Tuning UNet. Further investigation into the finer details of our approach reveals a compelling insight into the critical components involved in fine-tuning. As illustrated in Fig. S11, we demonstrate that fine-tuning solely the Video ControlNet component is not sufficient to effectively adapt image-to-video diffusion models to the geometry estimation method. In contrast, our approach of fine-tuning the UNet component proves far more effective, facilitating a seamless transition from image-based generation to accurate video geometry estimation.

I2G vs. V2G. We present a detailed comparison between the per-frame estimations of the image-to-geometry (I2G) and the

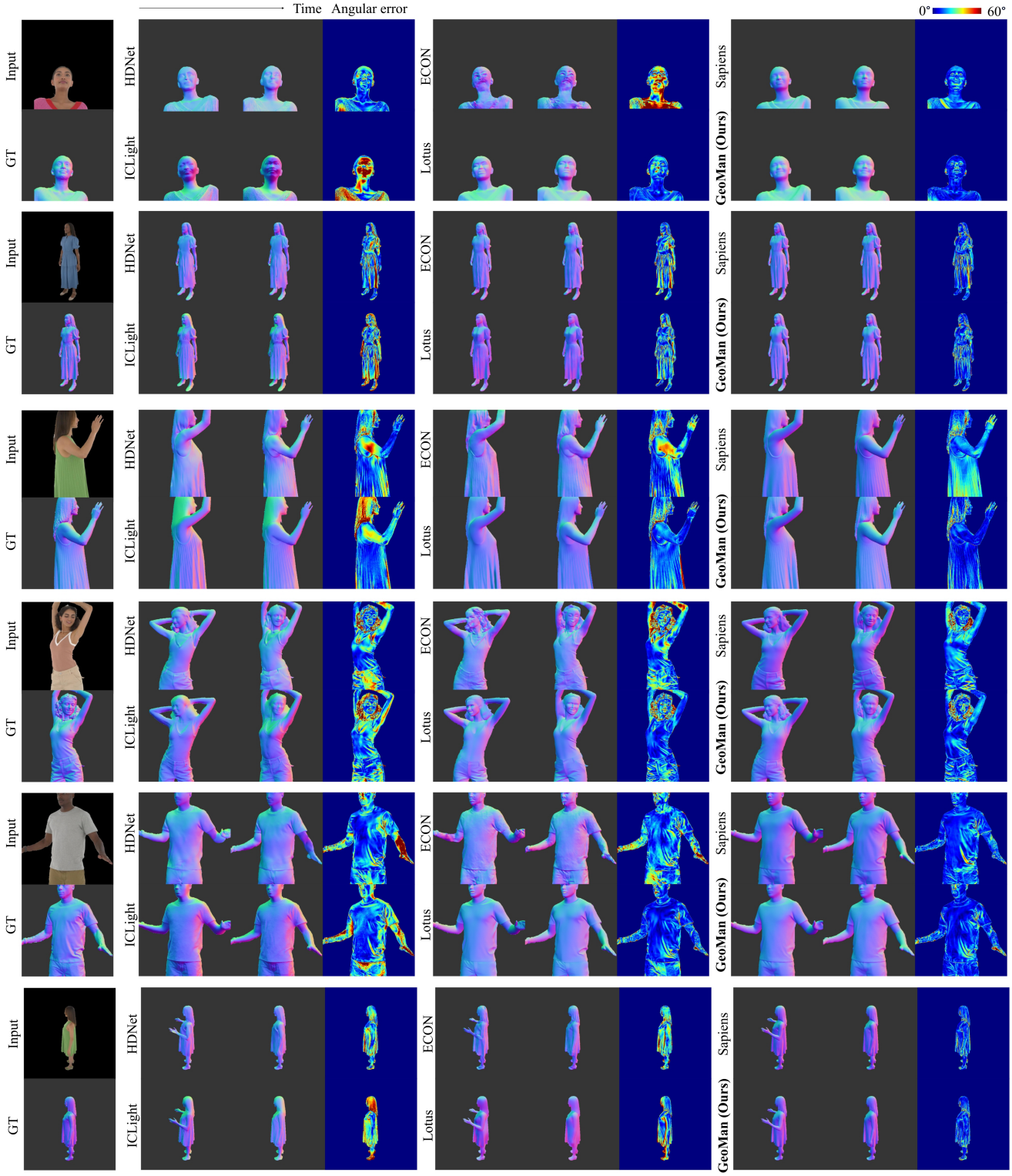


Figure S4. Comparison on zero-shot normal estimation on ActorsHQ. **Left**: Predicted normal. **Right**: Visualization of angular error.

full GeoMan pipeline in Tab. S3(b). While I2G and GeoMan (I2G+V2G) exhibits similar accuracy, GeoMan outperforms

Figure S5. Comparison with the existing depth estimation models. GeoMan produces the state-of-the-art results in both temporal consistency and fidelity.

I2G in temporal consistency. GeoMan pipeline excels in maintaining stability and coherence across frames, thereby ensuring smoother, more realistic video sequences. This superior temporal consistency is essential for high-quality video generation and demonstrates the power of GeoMan in addressing the challenges posed by dynamic visual content.

Multimodality of I2G. I2G faces challenges in multimodal modeling, as shown in Tab.S3(c), requiring modifications to input and output layers, which limits performance. While the V2G framework, involving similar tasks in image-to-video

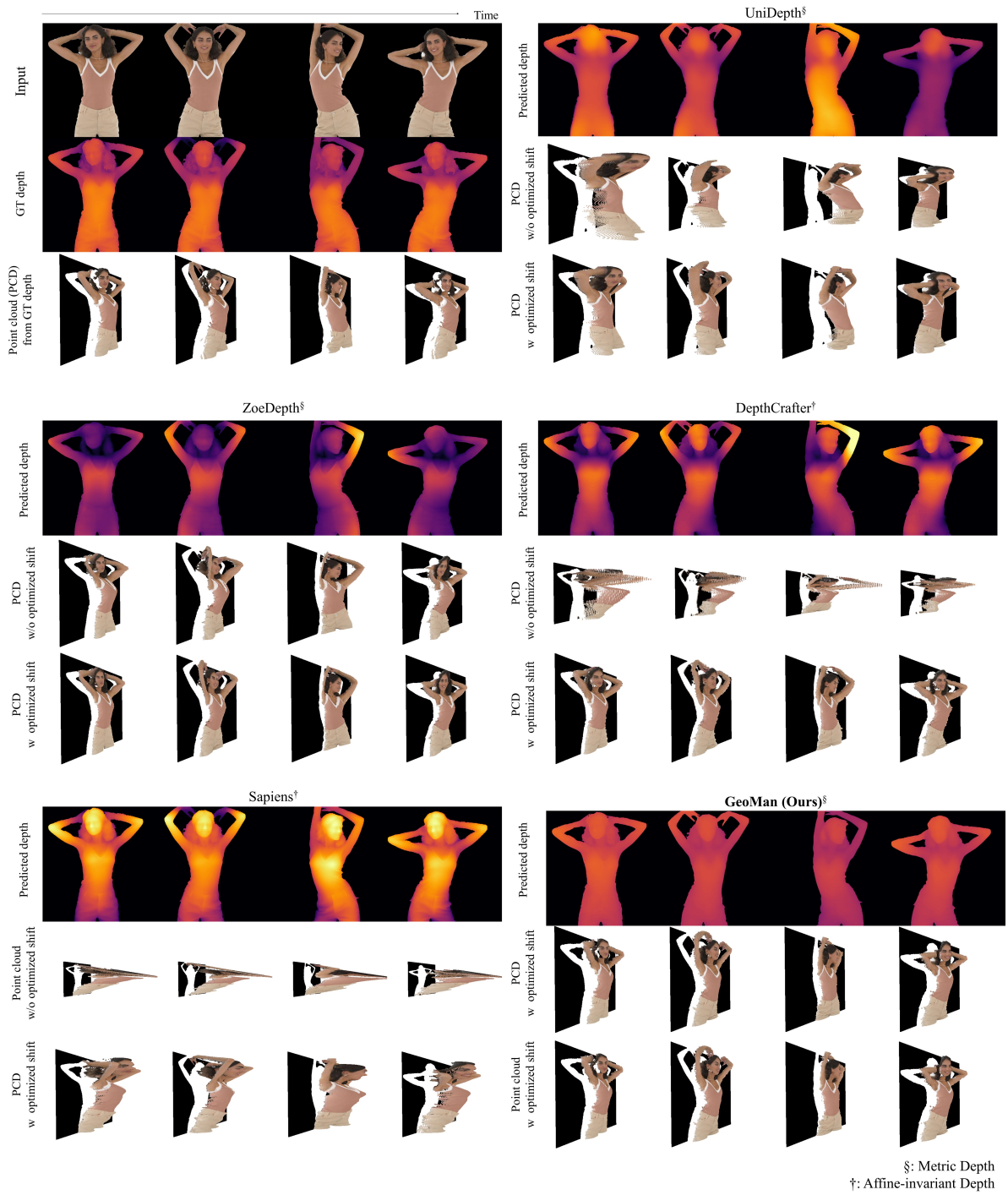


Figure S6. Comparison with the existing depth estimation models. GeoMan produces the state-of-the-art results in both temporal consistency and fidelity.

translation, benefits from fine-tuning a single model, enhancing efficiency and performance, I2G's task divergence demands separate training for each modality.

Training Data Sources for V2G. In Tab. S3(d), we evaluate the impact of different training datasets on V2G. Training only on rotating videos generated from static 3D scans introduces artifacts caused by motion bias. On the other hand, using only

Figure S7. Comparison with the existing depth estimation models. GeoMan produces the state-of-the-art results in both temporal consistency and fidelity.

low-resolution 4D data results in a lack of fine details, limiting the quality of predictions. By combining both datasets, we achieve the best results, with improved visual quality and enhanced temporal consistency.

Effectiveness of Human Area Crop. As demonstrated in Tab. S3(e), the use of human area cropping ensures that the model effectively targets relevant regions, thereby optimizing computational efficiency and accuracy. The positive impact of human

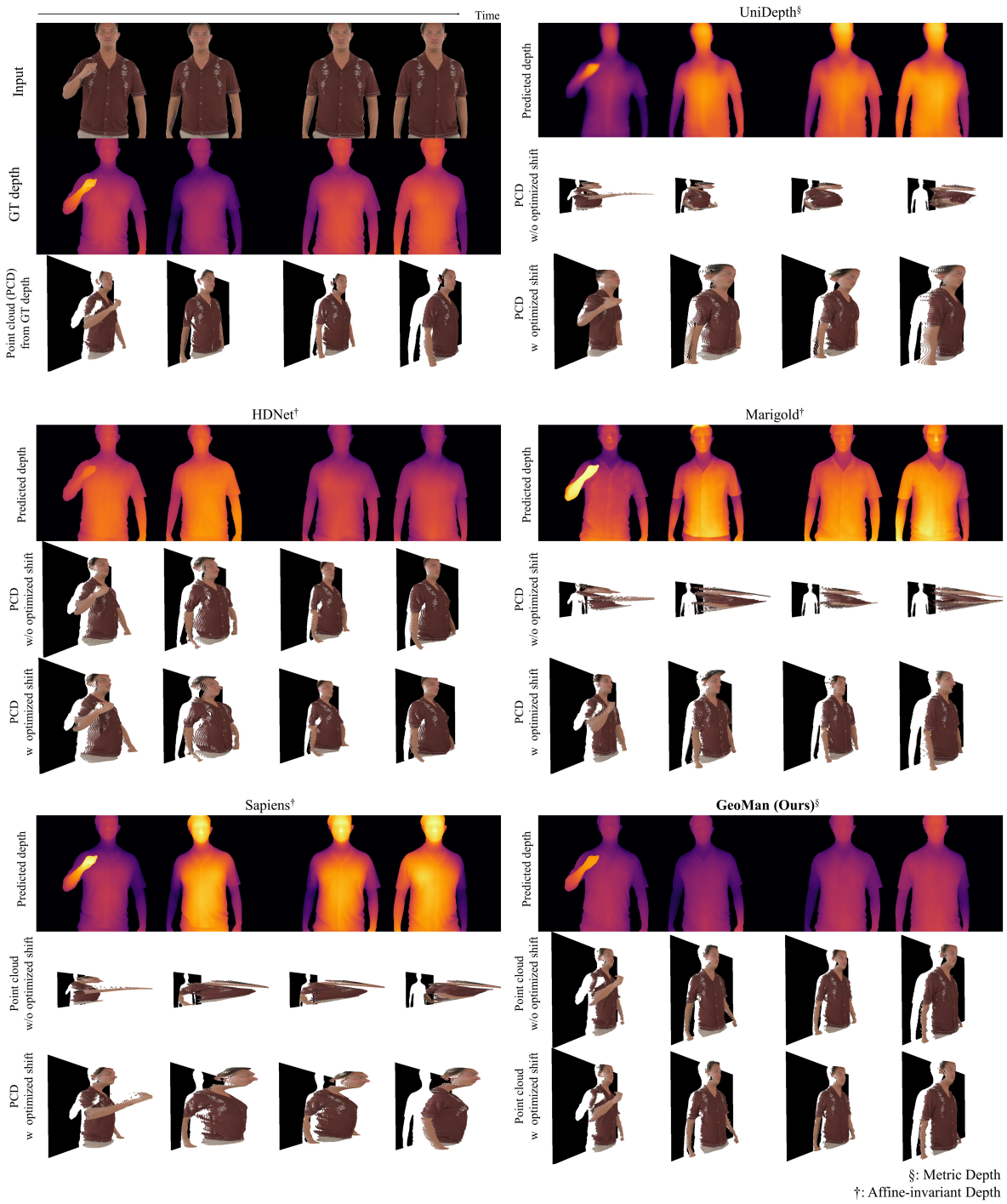


Figure S8. Comparison with the existing depth estimation models. GeoMan produces the state-of-the-art results in both temporal consistency and fidelity.

area cropping confirms that focusing on key areas within the video frames is important, ensuring that unnecessary background noise does not interfere with the core task of human geometry estimation.

First-Frame Dependency. Replacing I2G predictions with ground truth improves performance as shown in Tab. 4 in the paper. Degrading the first frame by reducing its resolution by 4 times resulted in only a slight increase in error (Tab. S3(f),

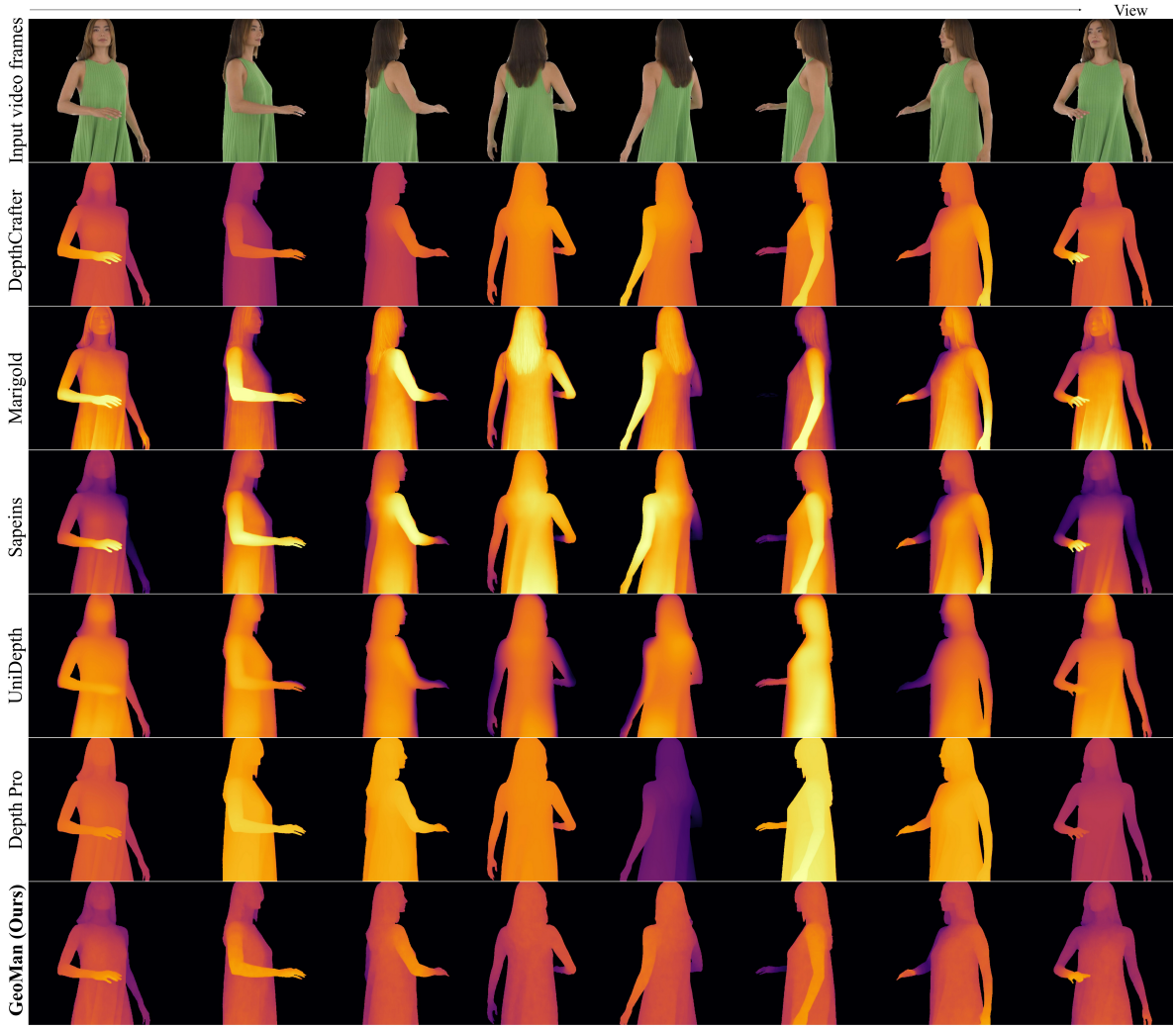


Figure S9. Comparison with the existing depth estimation models. GeoMan produces the state-of-the-art results in both temporal consistency and fidelity.

indicating the robustness of our model. The degraded version still outperforms other baselines, including Sapiens.

Robustness & Generalizability. We've shown generalization to long videos, multiple persons, and drastic poses in the supp. video. Fig. S12 additionally highlights GeoMan's robustness to diverse conditions, including occlusions, poor lighting, and atypical human types.

Method	GeoMan	Marigold	GeoWizard	IC-Light	ECON
Time (s)	10.57	8.44	23.34	7.854	134
Method	Depthcrafter	Sapiens	Metric3Dv2	UniDepth	DepthPro
Time (s)	3.59	1.92	0.31	0.56	0.90

Table S4. Methods colored in red are diffusion-based, in violet is optimization-based, in blue are feed-forward.

Efficiency Analysis. As shown in Tab. S4, GeoMan achieves competitive inference times compared to existing *diffusion-based* methods, though it is slower than some other baselines. However, Geoman has clear benefits: significantly improved temporal consistency, richer 3D detail, and enhanced human-specific understanding. While inference time increases with ensemble size or the number of diffusion steps, this trade-off is adjustable and can be tuned per application. We envision GeoMan as a practical tool for acquiring large-scale, real-world supervision, enabling its distillation into the next generation of faster, lightweight human geometric estimation models.

Figure S10. Comparison with the existing depth estimation models. GeoMan produces the state-of-the-art results in both temporal consistency and fidelity.

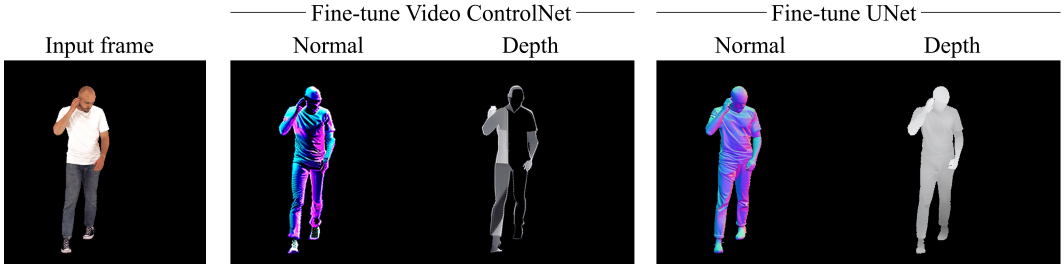


Figure S11. Fine-tuning the only Video ControlNet part is insufficient for reframing image-to-video diffusion models, whereas fine-tuning the UNet successfully transforms image-to-video generation into video geometry estimation.

C. Limitations and Future Works

For in-the-wild applications, our method relies on matting techniques to separate the subject from the background, making its performance inherently dependent on matting accuracy. Additionally, metric depth reconstruction is constrained by the precision of 3D human pose estimation, particularly the accuracy of pelvis root depth estimation. Due to VRAM limitations,

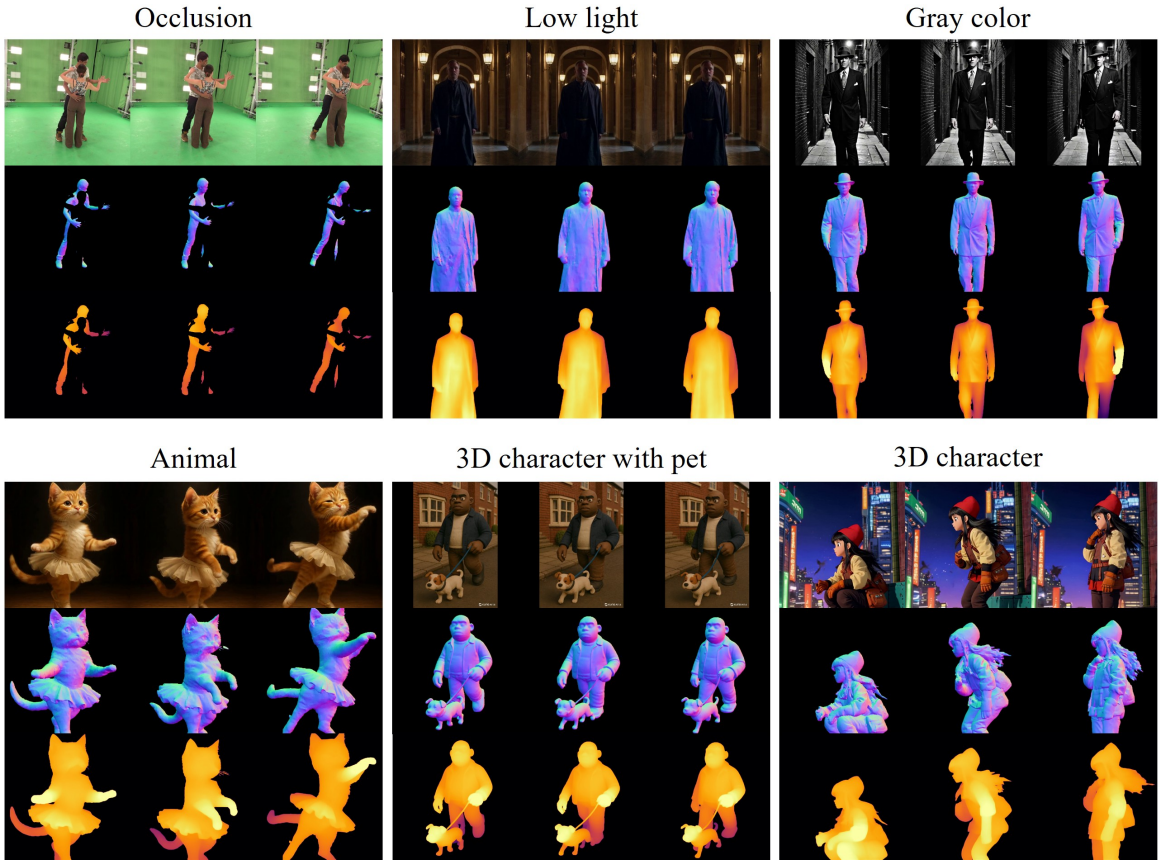


Figure S12. GeoMan demonstrates generalizablity across diverse scenarios, including occlusions, low lighting, and atypical human types.

we trained our model at a maximum resolution of 512, fitting within 80GB of NVIDIA A100 GPUs with a batch size of 1. While our approach produces detailed and high-quality geometry, we plan to explore lightweight models or alternative training strategies to enable higher-resolution generation.

References

[S1] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *NeurIPS*, 2014. 2

[S2] Wenbo Hu, Xiangjun Gao, Xiaoyu Li, Sijie Zhao, Xiaodong Cun, Yong Zhang, Long Quan, and Ying Shan. DepthCrafter: Generating consistent long depth sequences for open-world videos. *arXiv preprint arXiv:2409.02095*, 2024. 1, 3, 6

[S3] 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 *ECCV*, pages 668–685. Springer, 2022. 2

[S4] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. *CVPR*, 2024. 6

[S5] Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. Sapiens: Foundation for human vision models. *ECCV*, 2024. 4

[S6] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[S7] Han Lin, Jaemin Cho, Abhay Zala, and Mohit Bansal. Ctrl-Adapter: An efficient and versatile framework for adapting diverse controls to any diffusion model. *arXiv preprint arXiv:2404.09967*, 2024. 1

[S8] Julieta Martinez, Emily Kim, Javier Romero, Timur Bagautdinov, Shunsuke Saito, Shoou-I Yu, Stuart Anderson, Michael Zollhöfer, Te-Li Wang, Shaojie Bai, Chenghui Li, Shih-En Wei, Rohan Joshi, Wyatt Borsos, Tomas Simon, Jason Saragih, Paul Theodosis, Alexander Greene, Anjani Josyula, Silvio Mano Maeta, Andrew I. Jewett, Simon Venshtain, Christopher Heilman, Yueh-Tung Chen, Sidi Fu, Mohamed Ezzeldin A. Elshaer, Tingfang Du, Longhua Wu, Shen-Chi Chen, Kai Kang, Michael Wu, Youssef Emad, Steven Longay, Ashley Brewer, Hitesh Shah, James Booth, Taylor Koska, Kayla Haidle, Matt Andromalos, Joanna Hsu, Thomas Dauer, Peter Selednik, Tim Godisart, Scott Ardisson, Matthew Cipperly, Ben Humberston, Lon Farr, Bob Hansen, Peihong Guo, Dave Braun, Steven Krenn, He Wen, Lucas Evans, Natalia Fadeeva, Matthew Stewart, Gabriel Schwartz, Divam Gupta, Gyeongsik Moon, Kaiwen Guo, Yuan Dong, Yichen Xu, Takaaki Shiratori, Fabian Prada, Bernardo R. Pires, Bo Peng, Julia Buffalini, Autumn Trimble, Kevyn McPhail, Melissa Schoeller, and Yaser Sheikh. Codec Avatar Studio: Paired Human Captures for Complete, Driveable, and Generalizable Avatars. *NeurIPS*, 2024. 1, 4

[S9] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, William Berman, Yiyi Xu, Steven Liu, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022. 1

[S10] Yiran Wang, Zhiyu Pan, Xingyi Li, Zhiguo Cao, Ke Xian, and Jianming Zhang. Less is more: Consistent video depth estimation with masked frames modeling. In *ACM MM*, 2022. 2

[S11] Yufu Wang, Ziyun Wang, Lingjie Liu, and Kostas Daniilidis. TRAM: Global trajectory and motion of 3d humans from in-the-wild videos. In *Proc. European Conf. on Computer Vision (ECCV)*, pages 467–487, 2024. 1

[S12] Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv preprint arXiv:2311.04145*, 2023. 1

[S13] Peng Zheng, Dehong Gao, Deng-Ping Fan, Li Liu, Jorma Laaksonen, Wanli Ouyang, and Nicu Sebe. Bilateral reference for high-resolution dichotomous image segmentation. *CAAI Artificial Intelligence Research*, 3:9150038, 2024. 1

[S14] Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. Champ: Controllable and consistent human image animation with 3d parametric guidance. *ECCV*, 2024. 4