

Technical Paper

Seeing Machines HMR: Real-time 3D Human Mesh Recovery at up to 180 FPS on NVIDIA Jetson Thor

Bo Chen • Yuetian Weng • Edward Guilfoyle • Pratyush Sahay • Mathieu Ghatti • Akshay Asthana



Executive summary

Delivering state-of-the-art, embeddable 3D human geometry to Physical AI for safer operation, more capable interaction and productive deployment.

Robots working alongside people need timely information about human posture, movement and occupied space to make appropriate decisions as interactions unfold. Bringing human mesh recovery onto the embedded device in real-time makes this information available locally during operation.

Human Mesh Recovery (HMR) is one of the leading features of Seeing Machines' [Human-Centred Physical AI Platform](#), launched in August 2026¹. Available through the platform, it reconstructs 3D persons from a single-view RGB or RGB-D input (comprising an RGB image and the corresponding depth map) in real-time on embedded systems (**Figure 1**).

Given a bounding box around a person, the system predicts articulated dense 3D pose, 3D body shape, and metric camera-frame placement achieving state-of-the-art 3D pose estimation accuracy with significantly lower latency (**Figure 2**) than frontier models such as Meta's SAM-3D-Body² and Fast-SAM-3D-Body³. Both HMR and HMR-Lite support RGB and RGB-D operation without increasing their respective parameter counts.

On NVIDIA Jetson Thor,

- **HMR:**
 - Achieves **15.8ms** median latency for RGB input and **16.5ms** median latency for RGB-D input with only **386M parameters**.
 - Is similar in accuracy to **SAM-3D-Body** with **~33× speed-up** and **~70% fewer parameters**.
 - Is more accurate than **Fast-SAM-3D-Body** with **~7× speed-up** and **~70% fewer parameters**.
- **Similarly, HMR-Lite:**
 - Achieves **5.5ms** median latency for RGB input and **6.2ms** median latency for RGB-D input with only **146M parameters**.
 - Is similar in accuracy to **Fast-SAM-3D-Body** with **~20× speed-up** and **~88% fewer parameters**.

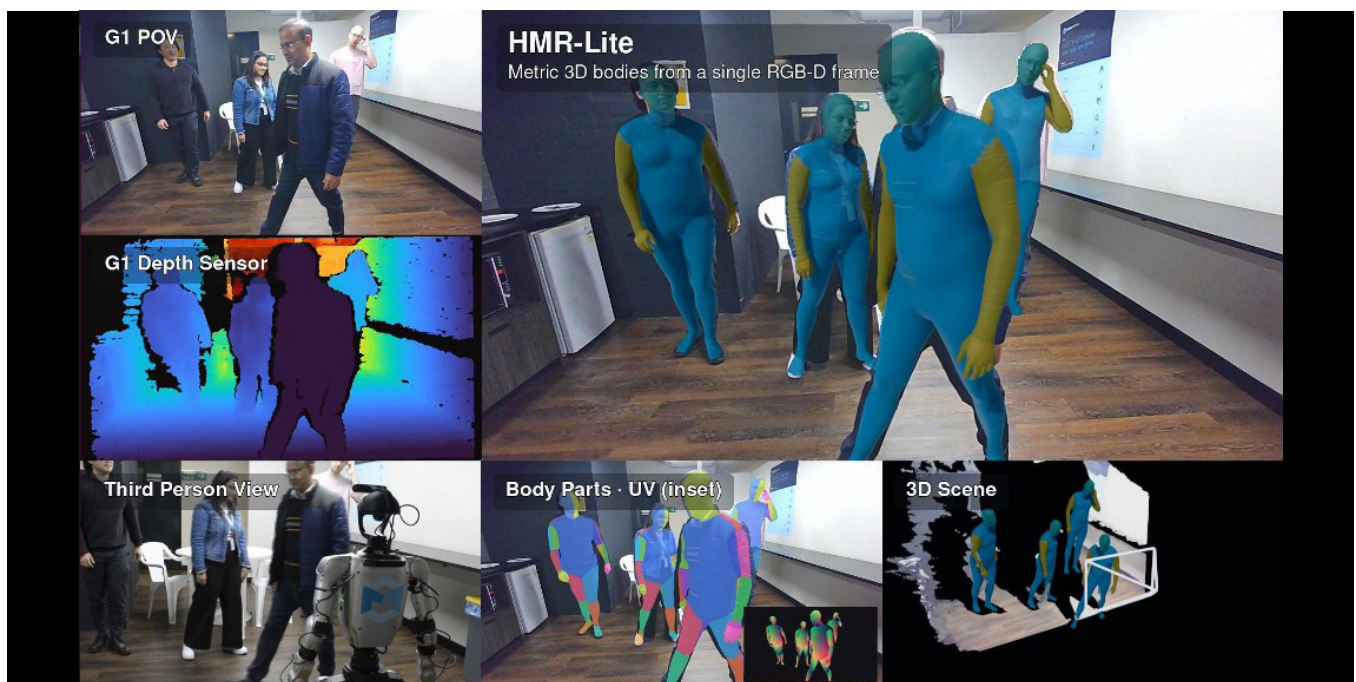


Figure 1: Unitree G1 in a busy kitchen setting.

Accuracy, Latency and Size Comparison

RGB inputs | Harmony4D MPJPE | NVIDIA Jetson Thor

Lower MPJPE and latency are better. Bubble area is proportional to parameter count.

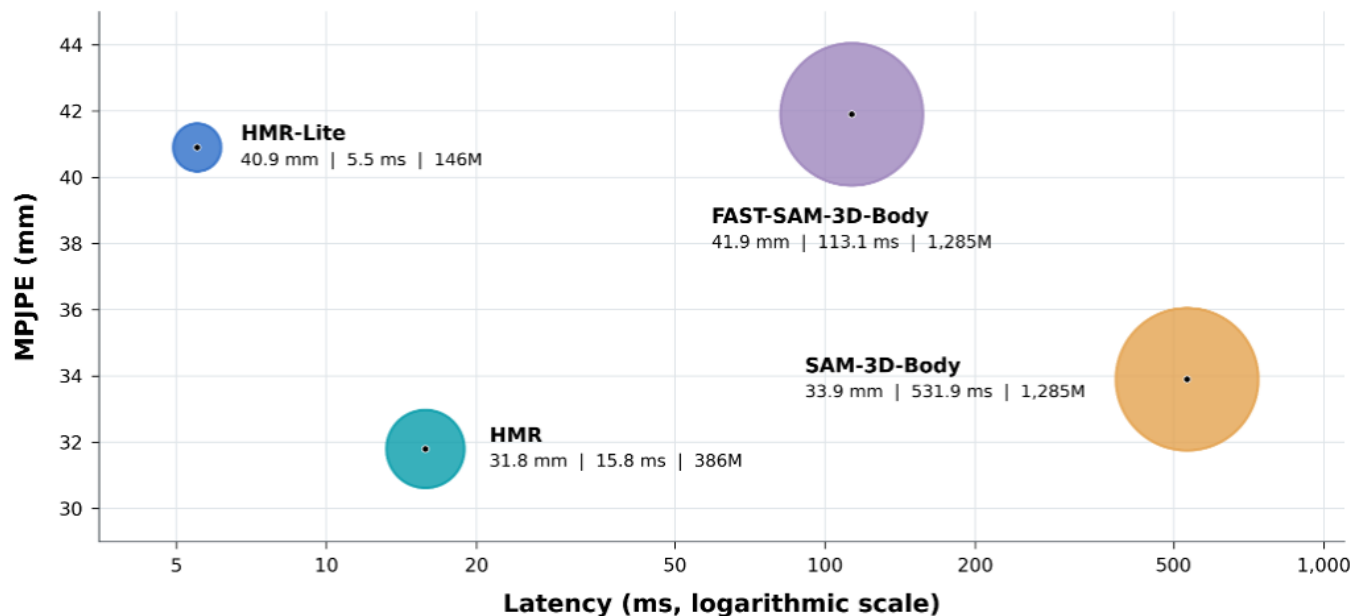


Figure 2: Accuracy, Latency and Size comparison. Accuracy is MPJPE on Harmony4D leave-one-out split^{2,4}. Latency is median full-body pipeline latency, excluding the person detection bounding box latency, on NVIDIA Jetson Thor. Bubble area is proportional to model parameter count.

The significance extends beyond 3D mesh recovery speed. More broadly, HMR demonstrates how rich human understanding can be brought to embedded Physical AI systems in real time, enabling robots to reason about people rather than simply detect them.

A locally available, frequently updated body representation can supply the human geometry needed for proximity assessment, motion forecasting, coordinated handover, human-aware navigation, teleoperation and robot learning.

Combined with scene perception, temporal tracking, robot state and appropriate supervision, it provides a foundation for three operational goals:

- **Safer shared-space operation.**
- **More capable human-robot interaction.**
- **Productive deployment at scale.**

For Seeing Machines, HMR creates an opportunity to extend its human-perception and embedded-software capabilities across Physical AI including robots (humanoids, collaborative robots, industrial automation and autonomous mobile robots) and other adjacent markets. Broadly speaking, it represents capability to understand human behaviour and movement in real time on embedded platforms. As Physical AI systems move from controlled environments into human-centric spaces, these capabilities will become increasingly important for safe, effective and scalable deployment.

HMR is also an example of a broader transition within Seeing Machines: applying decades of expertise in understanding human behaviour inside safety-critical systems to the emerging Physical AI market. While this paper focuses on 3D human mesh recovery, the same underlying approach can support a wider range of human-centric perception capabilities required for next-generation robots and autonomous systems.

HMR demonstrates how rich human understanding can be brought to embedded Physical AI systems in real time, enabling robots to reason about people rather than simply detect them.

What does a 3D human mesh deliver?

Human mesh recovery provides a unified 3D representation of a person's body surface, shape and articulated configuration. From the mesh and its associated body-model parameters, applications can obtain multiple complementary outputs (**Figure 3**):

- **Anatomical landmarks and joints:** Estimated 3D locations of shoulders, elbows, wrists, hips, knees and other reference points.
- **Body pose and articulation:** Body orientation, limb configuration and joint rotations for understanding posture and mapping human movement to robot motion.
- **Body shape and proportions:** Estimated body dimensions and surface geometry, providing spatial detail beyond a skeleton or bounding box.
- **Body-part geometry:** The position and extent of individual body regions, supporting assessments of reaching, leaning and proximity.
- **Spatial placement and occupancy:** Estimated camera-frame position and occupied volume, which can be related to robots, objects and workspaces through calibration.
- **Movement over time:** With temporal tracking, successive reconstructions support trajectories, joint velocities and posture changes that can feed motion forecasting and activity models.

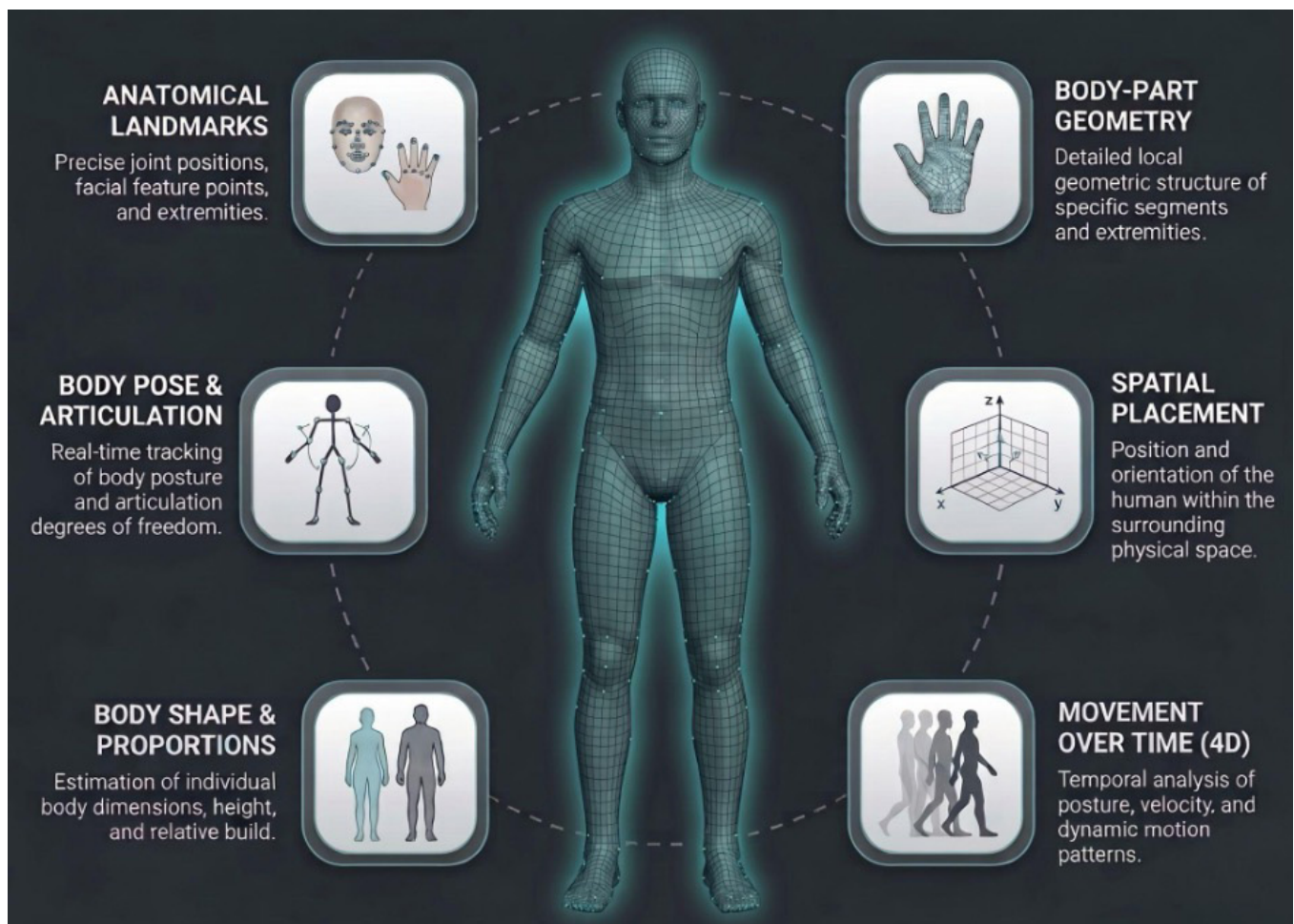


Figure 3: 3D Human Mesh outputs rich geometric and semantic information.

How does HMR work?

HMR combines the **Seeing Machines Geometry-Aware Foundation Model** with a task-specific Human Mesh Decoder module. The pipeline (**Figure 4**) converts RGB or RGB-D input, optionally supported by multi-view imagery, into parameters describing a person's 3D body pose, shape and placement. HMR and HMR-Lite each support both input modes. Throughout this paper, **RGB** denotes operation using colour imagery alone, while **RGB-D** denotes operation using colour imagery and a corresponding depth map. Both modes recover a 3D human mesh.

Geometry-Aware feature extraction

Seeing Machines has developed a **family of embedding friendly and geometry-aware foundational models** that provide a shared geometric foundation for its Physical AI platform, with development spanning multi-view understanding and multimodal fusion. For HMR, the input image is processed by this geometry-aware foundational model to produce a representation that captures visual appearance and spatial structure.

In both HMR and HMR-Lite, an optional depth map enters the foundational backbone alongside the monocular or multi-view images, providing additional spatial information to support mesh recovery.

Multi-view and depth inputs are optional, allowing HMR to operate from monocular RGB imagery alone. HMR represents one application of this foundation, demonstrating how a common perception stack can support multiple Physical AI capabilities.

Human Mesh Decoder

The Human Mesh Decoder module uses a Transformer-based prediction head to map foundational backbone features to body mesh parameters. This task-specific stage translates the foundation model's representation into estimates of body shape, articulated pose and camera-frame placement. These parameters define a coherent 3D body representation from which the human mesh and associated anatomical joints can be recovered.

Reconstruction and downstream outputs

The resulting representation supports multiple outputs, including body-surface geometry, anatomical landmarks, joint configurations, body proportions and spatial extent. Downstream applications can use the complete mesh or derive simpler representations, such as skeletal pose and body-part volumes, according to their requirements. This shared representation supports applications ranging from motion retargeting and demonstration processing to human-aware interaction and proximity assessment.

Extension to temporal understanding

Our geometry-aware foundational model development is currently approaching the end of its multimodal-fusion stage. A planned extension will introduce temporal modelling to connect observations across time, targeting more consistent representations of changing scenes and human movement.

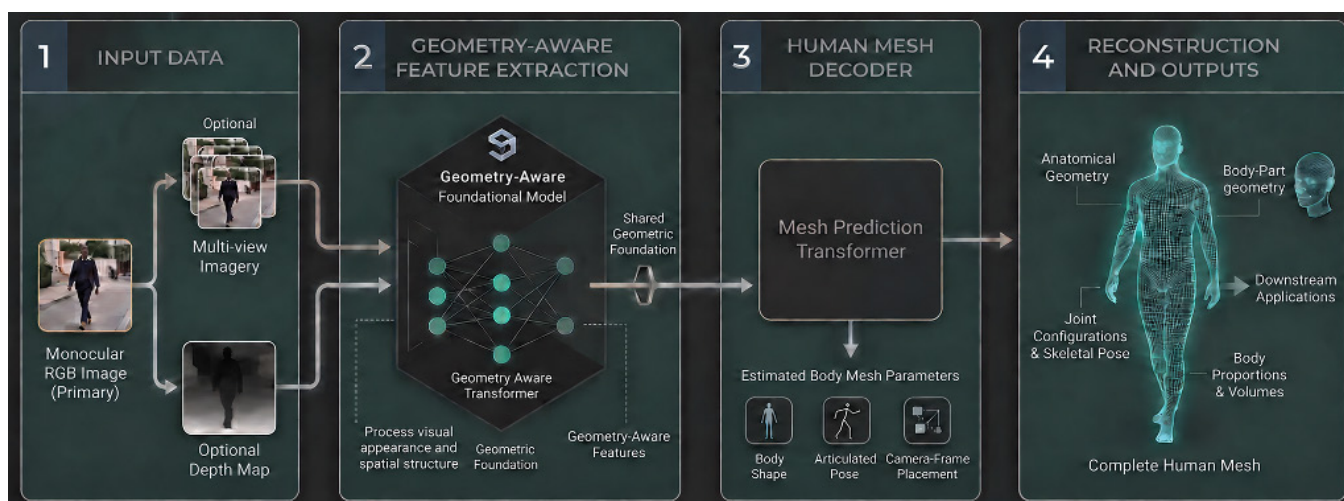


Figure 4. HMR design overview. Both HMR and HMR-Lite support RGB and RGB-D inputs. Incorporating depth adds minimal processing overheads on NVIDIA Jetson Thor without increasing model parameter count.

Compelling combination of accuracy, latency and size

Embedded human perception must balance reconstruction accuracy, processing latency and model size. HMR and HMR-Lite combine state-of-the-art joint-position accuracy with extremely low embedded latency and compact model size relative to other publicly available frontier models^{2,3}.

In this section, the treatment of the person detection bounding box is excluded from the latency measurements. Measurements used a NVIDIA Jetson AGX Thor Developer Kit in the MAXN power mode with

jetson_clocks applied, which pins all 14 CPU cores at 2.6 GHz and the GPU at 1575 MHz. The software stack was JetPack 7.2, CUDA 13.0, and TensorRT 11.1.0. Each latency value is the median of 100 CUDA-event-timed executions at batch size 1 after 20 warm-up iterations; the latency spread was within 2% of the median for all models. All experiments are conducted using TensorRT FP16 mixed precision denoting mixed-precision execution using FP16 with minimal FP32 pinning: selected operations were explicitly retained in FP32.

Model	Input	#Parameters	Latency
SAM-3D-Body	RGB	1285M	531.9ms
FAST-SAM-3D-Body	RGB	1285M	113.1ms
HMR	RGB	386M	15.8ms
	RGB-D		16.5ms
HMR-Lite	RGB	146M	5.5ms
	RGB-D		6.2ms

Table 1. Measured latency of SAM-3D-Body, FAST-SAM-3D-Body, HMR and HMR-Lite on NVIDIA Jetson Thor.

We report Mean Per-Joint Position Error (MPJPE) and Per Vertex Error (PVE) in millimeters on Harmony4D leave-one-out split^{2,4}. The accuracy results below apply to RGB input.

HMR prioritises accuracy while maintaining low embedded latency. It achieves MPJPE of 31.8mm and PVE of 38.8mm with a **15.8ms** latency on **NVIDIA Jetson Thor** with **386 million parameters**. In comparison, SAM-3D-Body achieves MPJPE of 33.9mm and PVE of 41.0mm with 531.9ms latency on the same device with 1285M parameters.

On the other hand, HMR-Lite offers a smaller model with further latency savings. It achieves MPJPE of 40.9mm and PVE of 49.9mm on the same dataset with a **reduced latency of 5.5ms** with **146 million parameters count**. In comparison, FAST-SAM-3D-Body achieves MPJPE of 41.9mm and PVE of 46.9mm with 113.1ms latency on the same device with 1285M parameters.

Our milestone is the combination of high accuracy in 3D mesh recovery, low embedded latency and significantly fewer parameters. Lower latency leaves more of the application's response budget available for tracking, scene understanding and downstream decisions. Fewer parameters further improve deployability by reducing memory, bandwidth and power demands, allowing the model to operate alongside other physical AI workloads on resource-constrained embedded platforms.

Support for RGB and RGB-D inputs adds deployment flexibility: applications can use RGB imagery alone or incorporate available depth information without increasing the model parameter count. Together, these characteristics make HMR and HMR-Lite compelling options for integrating detailed human geometry into embedded Physical AI systems. See Appendix for detailed qualitative results (**Figures 6-11**).

Why does real-time embedded HMR matter?

Detailed geometry supports more specific decisions

A robot approaching a person, exchanging an object or moving through a shared workspace needs information about the space the person occupies and how that space is changing. An articulated body model provides estimated body shape, posture and limb configuration.

Combined with robot and scene geometry, it can support reasoning about an extended arm, a leaning torso, a reaching movement or a person crossing the robot's intended path. While key-points and occupancy estimates are sufficient for some applications, rich body geometry captures body extent, articulation and posture simultaneously, enabling finer-grained reasoning about human movement and interaction.

Consider a worker reaching into a shared workcell while their feet and torso remain largely stationary. A body representation can supply the estimated arm configuration needed to assess that movement relative to the robot and surrounding objects. A downstream system can then evaluate whether the situation calls for continued operation, reduced speed, yielding or stopping.

The value of a mesh depends on the application. Some tasks can be served effectively by keypoints, body capsules or occupancy estimates. HMR's opportunity is to demonstrate where its combination of articulated geometry, body extent and responsiveness improves the decisions that matter.

Low latency keeps the representation relevant

People continue moving while perception runs. As a result, 3D mesh quality and information age both affect the usefulness of an estimate. For illustration, a body part moving at 1 metre per second travels 10 centimetres in 100 milliseconds.

Lower perception latency can make more recent geometry available to motion prediction, interaction planning and supervision. HMR's reported latency is therefore significant as an enabling contribution to the robot's response budget.

Embedded execution makes human geometry available locally

Running 3D mesh recovery on the embedded device removes the need for a network round trip for that stage of perception. It also allows integration with locally available camera streams, robot telemetry and other perception models.

The central proposition is: Real-time embedded HMR makes detailed human geometry available while interactions are unfolding, creating opportunities for more responsive and context-aware robot behaviour.



Applications: from human geometry to operational value

Real-time embedded HMR supplies detailed human geometry for responsive robot behaviour. Combined with temporal tracking, scene understanding and robot state, it can support safer operation, more capable interaction and productive deployment. Building on this foundation, proposed applications include:

- **Safer shared-space operation:** Body and limb proximity, dynamic exclusion zones, human-object-robot risk assessment, motion forecasting and fall-event monitoring can inform earlier decisions to slow, stop, yield or reroute. The objective is improved protection with fewer avoidable interruptions.
- **More capable interaction:** Body configuration and movement can support human-aware navigation, gesture interpretation, coordinated handover and assistance. Separate gaze, attention and object-state inputs can add interaction context.
- **Learning and simulation:** Reconstructed motion can support teleoperation, teach by demonstration, candidate 3D annotations, policy training and scenario generation. Live human-state estimates can also provide inputs to robot policies and VLA systems.
- **Fleet intelligence:** Combining body-state sequences with robot actions and scene context can support incident replay, intervention analysis and deployment improvement. Selected, permissioned data can inform model updates through controlled review and validation.



Figure 5. Real-time Teleoperation of Unitree G1 using Seeing Machines' HMR on NVIDIA Jetson Thor.

Future work

From real-time human geometry to human-centred physical AI

HMR establishes a foundation for bringing detailed human geometry into responsive robotics systems. Future papers will extend HMR towards broader embedded deployment, exploring low-precision models and production-grade NPUs to balance accuracy, responsiveness and power efficiency.

Another promising direction is contact-point estimation: combining body-surface geometry with scene context to support large-scale annotation of egocentric video and retargeting that better preserves human interactions with objects and supporting surfaces.

Beyond human mesh recovery, we will showcase other applications of our geometry-aware foundational model including real-time and embedding friendly 3D object detection and human motion forecasting, opening opportunities to connect human and scene understanding with anticipation of how people move and interact.

Partner with Seeing Machines

Seeing Machines welcomes discussions with prospective customers and partners interested in integrating HMR into robotics and Physical AI systems.

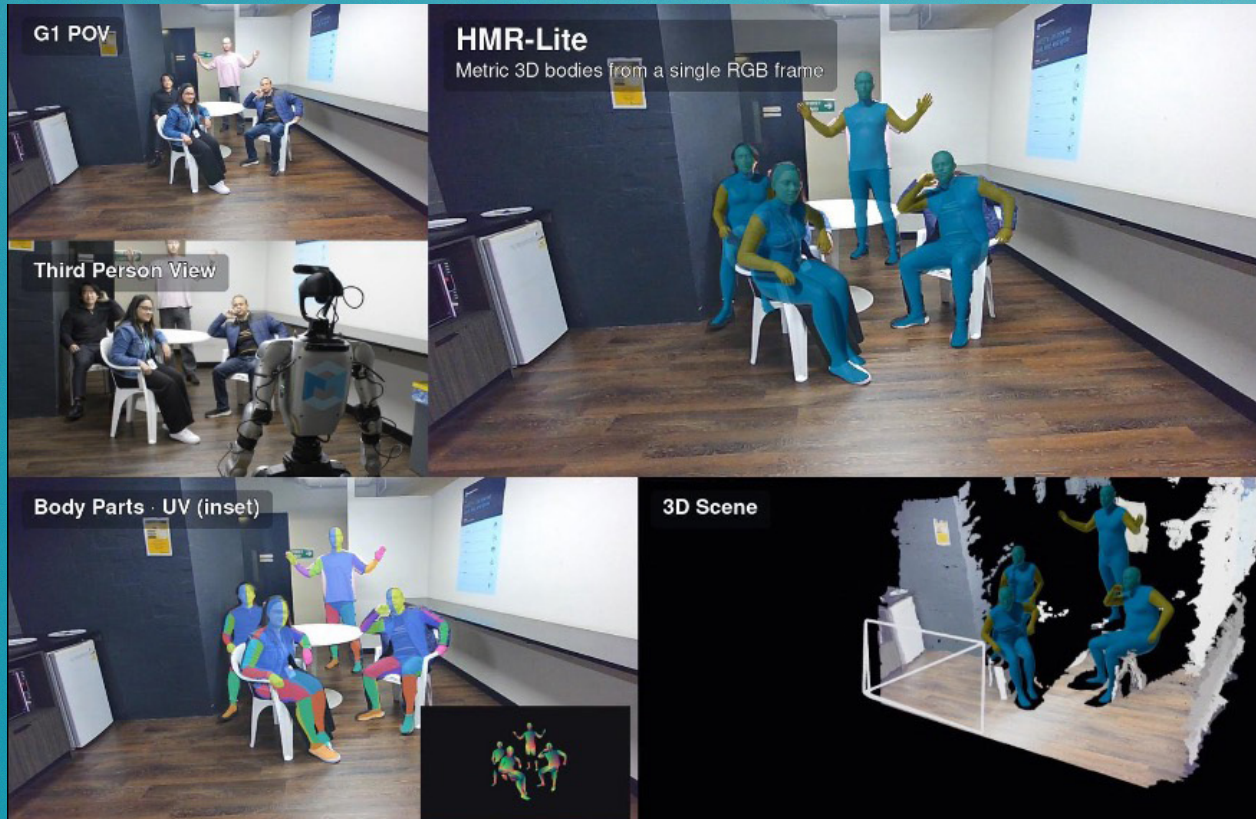
To learn more about our human-centred perception capabilities and explore collaboration opportunities, visit **seeingmachines.com**.

Acknowledgement: We thank Raul Latorre Fortes for his comments on this whitepaper.

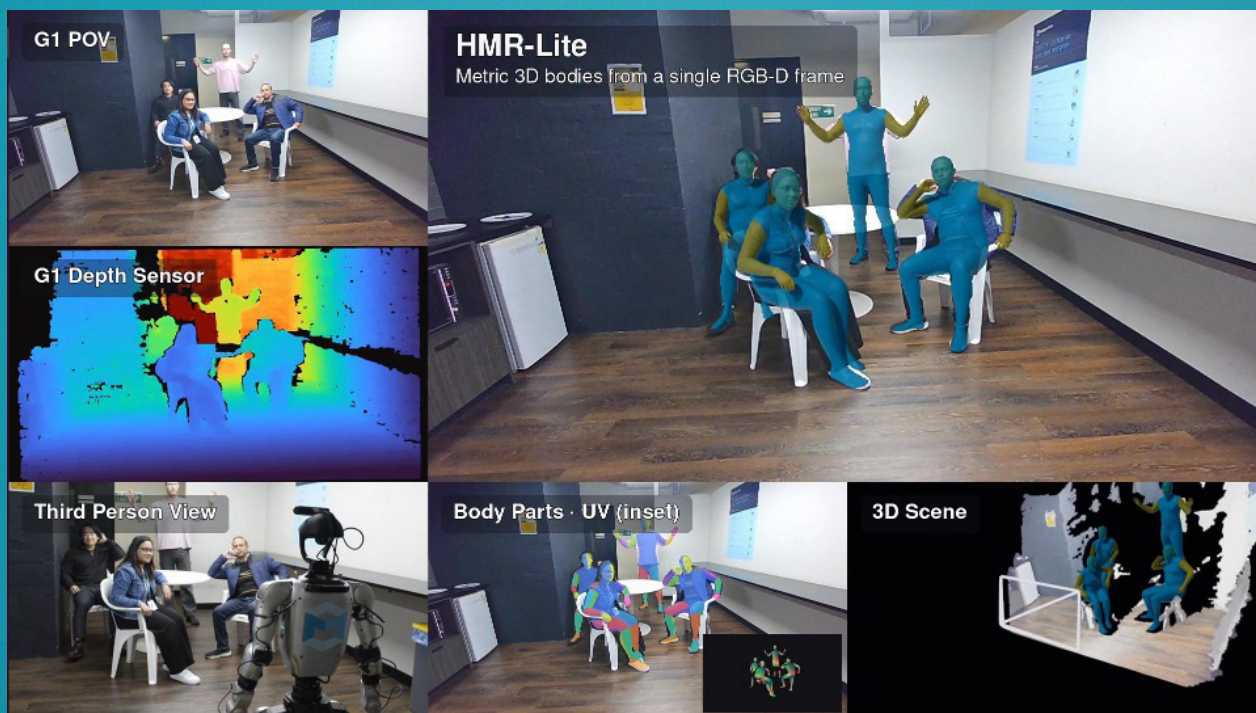
References

1. Seeing Machines. (2026, August 17). *Seeing Machines launches Physical AI Platform to power the next generation of humanoid robots*. [Seeing Machines](#).
2. Yang, X., Kukreja, D., Pinkus, D., Fan, T., Park, J., Shin, S., Cao, J., Liu, J.-W., Ugrinovic, N., Sagar, A., Malik, J., Feiszli, M., Dollár, P., & Kitani, K. (2026). SAM 3D Body: Robust full-body human mesh recovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 7209–7219).
3. Yang, T., He, S., Jing, H., Yang, J., Liu, Z., Zou, C., & Wang, Y. (2026). Fast SAM 3D Body: Accelerating SAM 3D Body for real-time full-body human mesh recovery. In P. Favaro, Z. Kukelova, A. Maki, A. Rohrbach, K. Schindler, & F. Tombari (Eds.), *Computer vision – ECCV 2026* (Lecture Notes in Computer Science, Vol. 17012). Springer. https://doi.org/10.1007/978-3-032-37044-0_15
4. Khirodkar, R., Song, J.-T., Cao, J., Luo, Z., & Kitani, K. (2024). Harmony4D: A video dataset for in-the-wild close human interactions. *Advances in Neural Information Processing Systems*, 37, 107270–107285. <https://doi.org/10.52202/079017-3407>

Appendix: Qualitative results



(a) HMR-Lite with RGB input



(b) HMR-Lite with RGB-D input

Figure 6. Unitree G1 in the Kitchen.

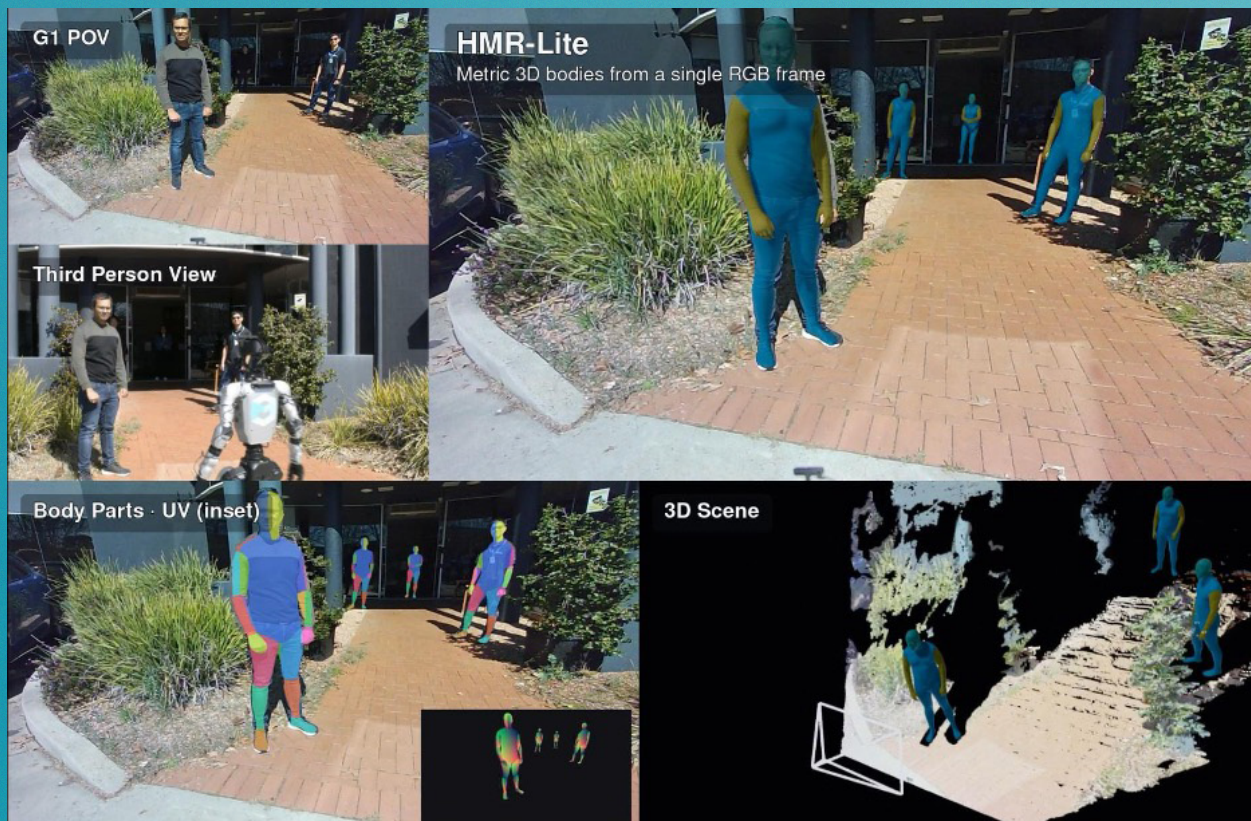


(a) HMR-Lite with RGB input

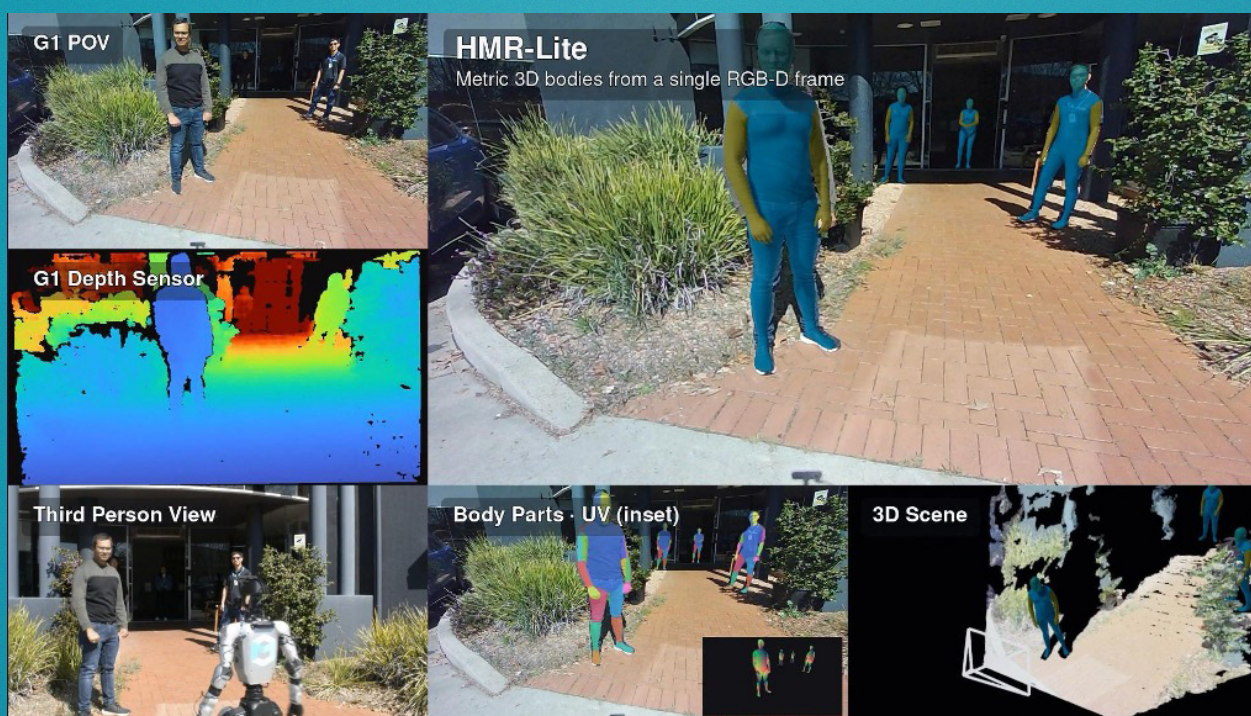


(b) HMR-Lite with RGB-D input

Figure 7. Unitree G1 in the Kitchen.

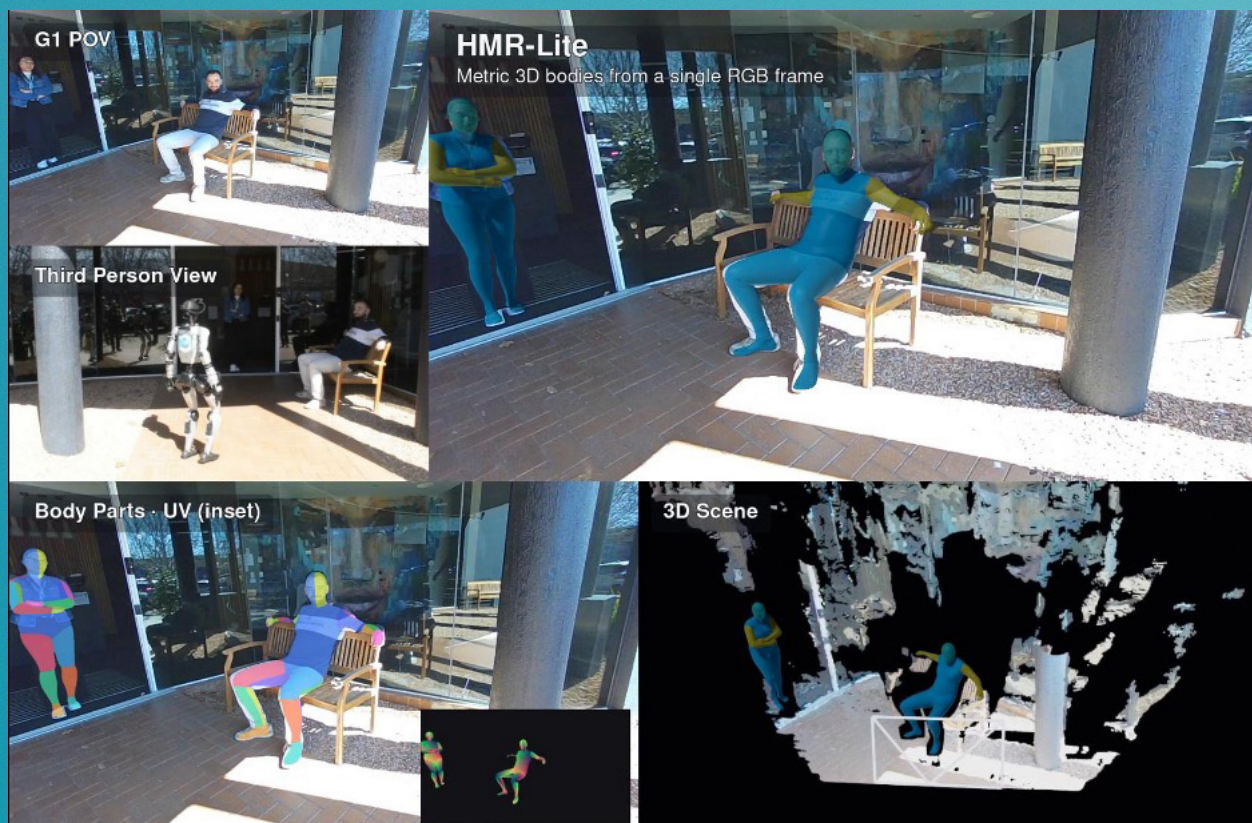


(a) HMR-Lite with RGB input

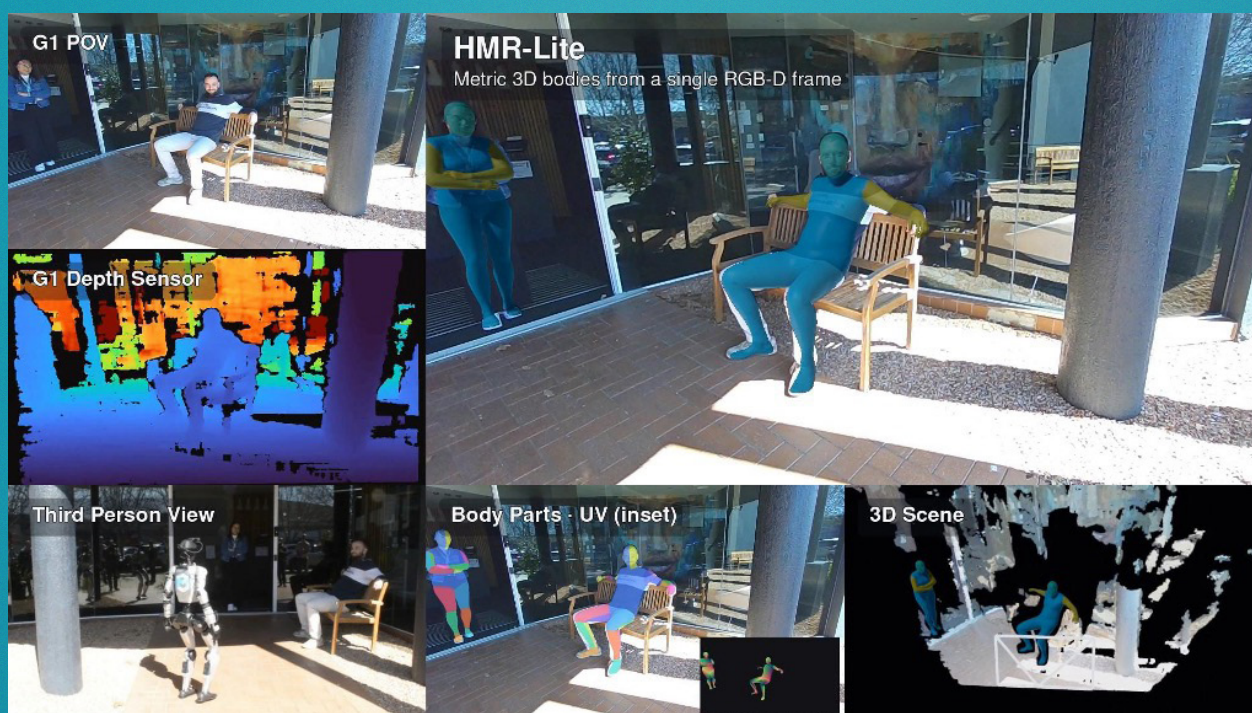


(b) HMR-Lite with RGB-D input

Figure 8. Unitree G1 in the outdoor settings under tough illumination conditions.

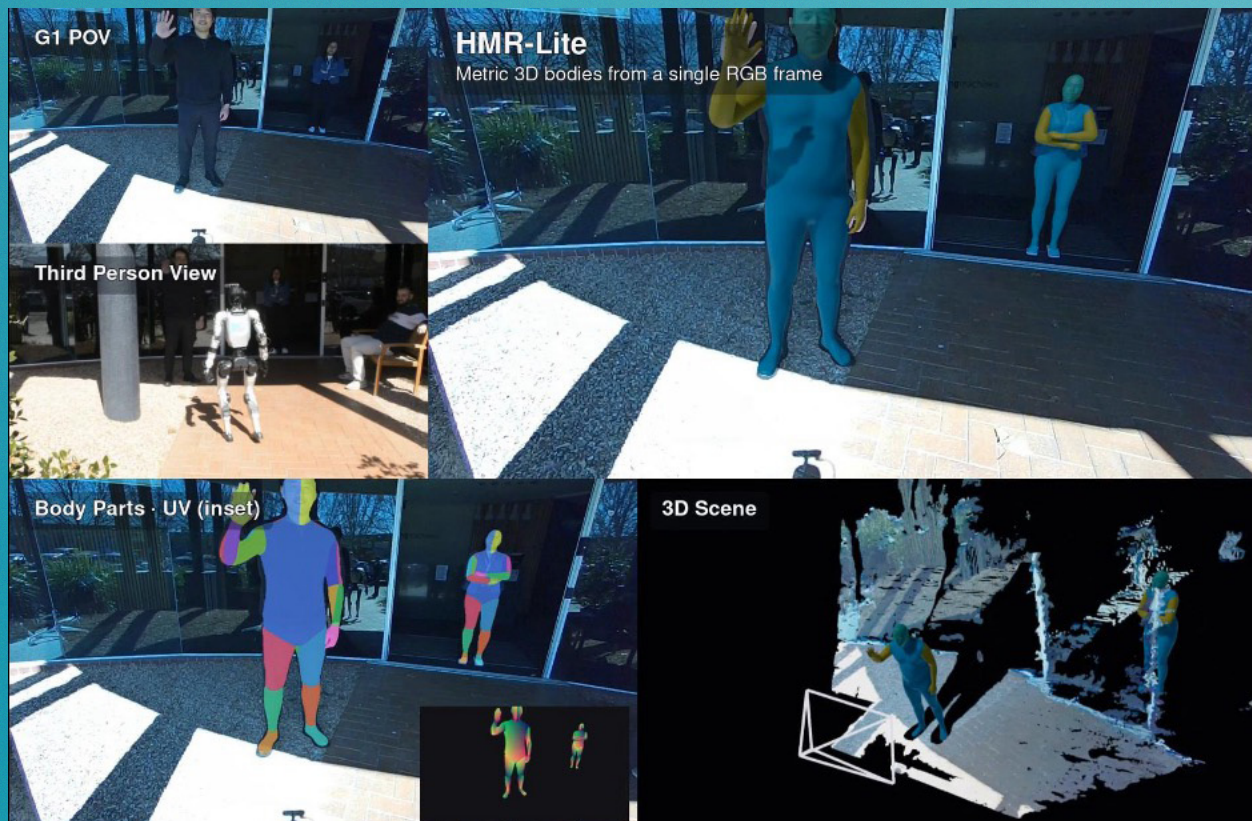


(a) HMR-Lite with RGB input

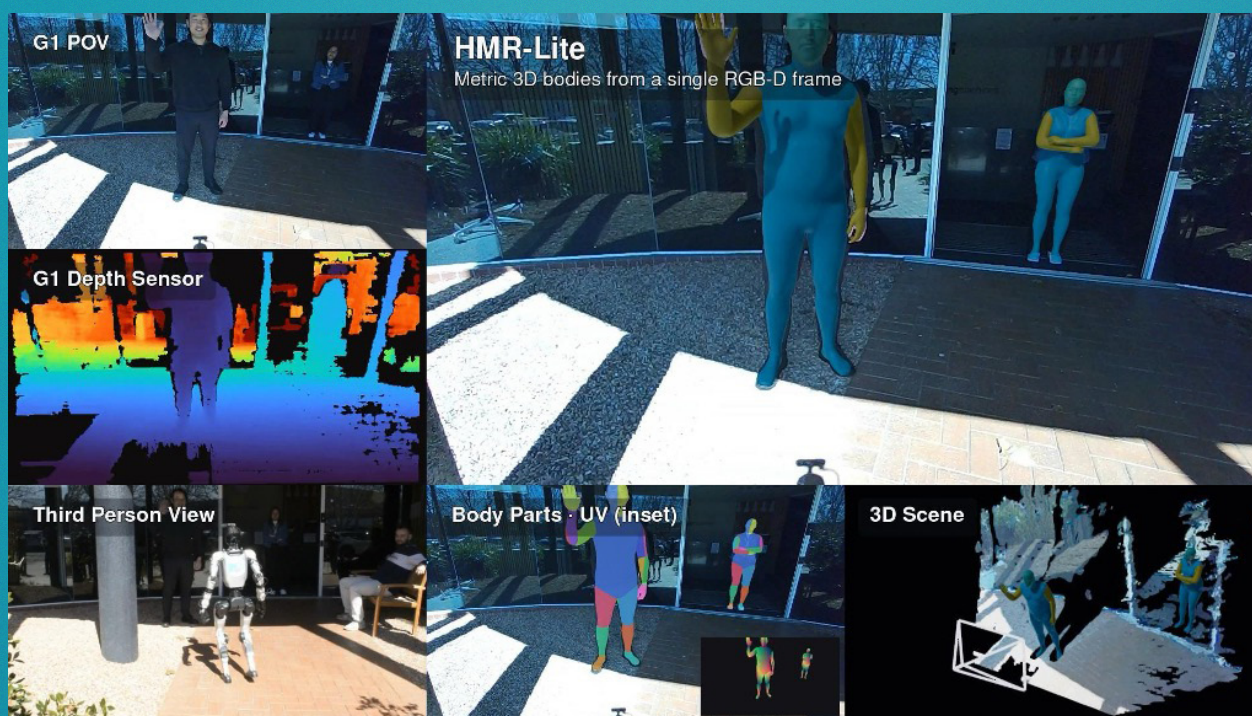


(b) HMR-Lite with RGB-D input

Figure 9. Unitree G1 in the outdoor settings under tough illumination conditions.

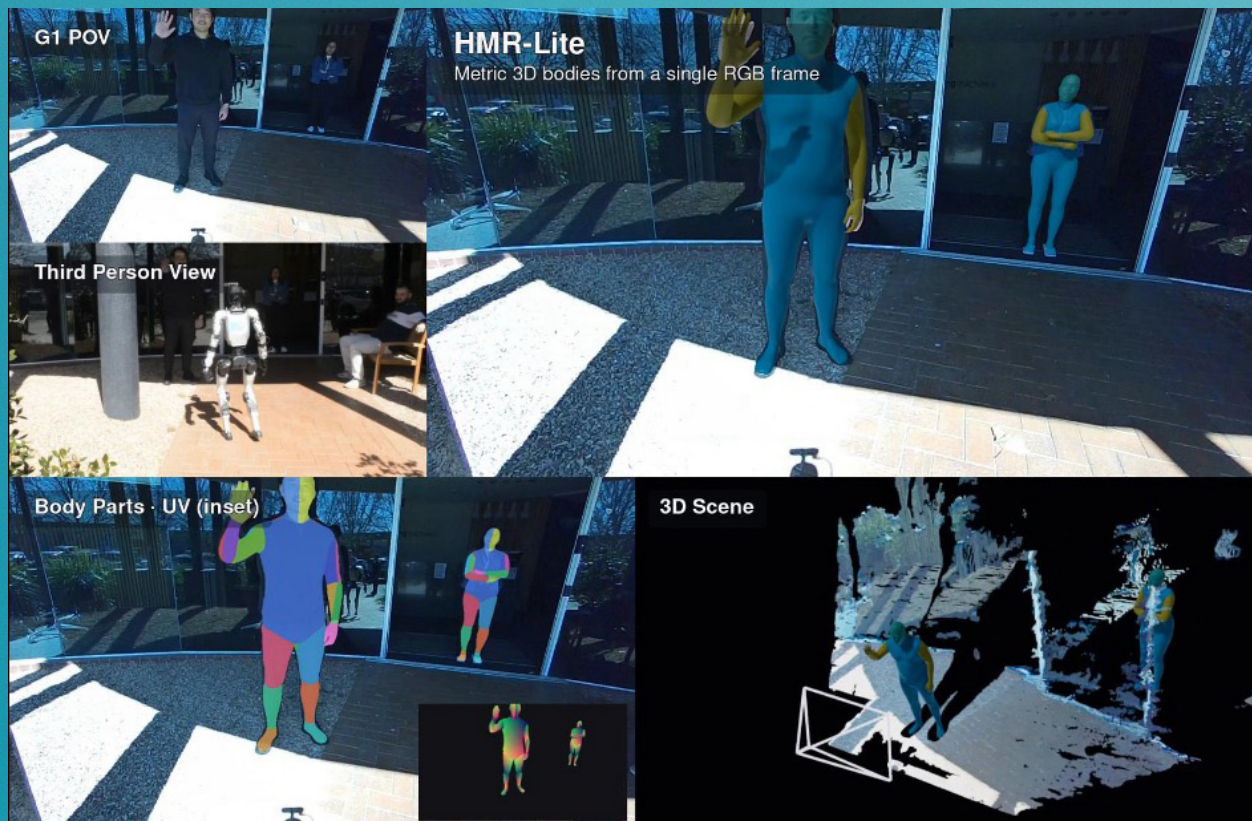


(a) HMR-Lite with RGB input

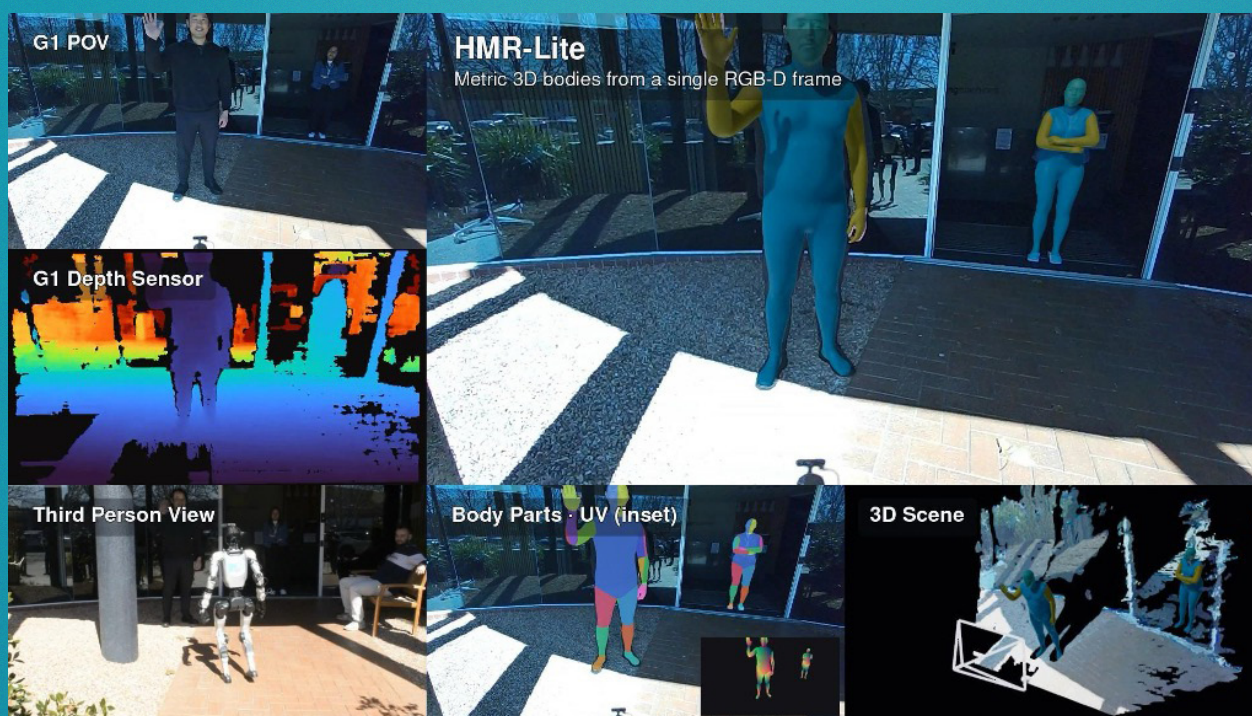


(b) HMR-Lite with RGB-D input

Figure 10. Unitree G1 in the outdoor settings under tough illumination conditions.



(a) HMR-Lite with RGB input



(b) HMR-Lite with RGB-D input

Figure 11. Unitree G1 in the indoor settings under tough illumination conditions.

