

Ego2Sim: Photorealistic Contact-Enabled Simulation from Egocentric Data

Anonymous Author(s)

Affiliation

Address

email



Fig. 1: Ego2Sim converts a single walk captured by head-worn glasses into a *contact-enabled hybrid simulation scene*: a Gaussian splat for what a robot sees and a collision mesh for contact. The Spot robots in the image (from Isaac Sim [1]) walk in a scene built from a 2.6-minute egocentric recording and make contact with the environment, as shown above.

Abstract: Vision-based navigation and locomotion policies are increasingly trained in digital twins of real environments. However, a training environment with both the visual appearance and the contact geometry of a scene often requires a dedicated scanner or a manual registration step. Egocentric glasses offer a lightweight alternative: they capture a scene hands-free as the wearer walks through it, and recordings from such glasses already serve as human demonstrations for robot learning. In this paper, we present a real-to-sim pipeline that builds a realistic and contact-enabled digital twin in Isaac Sim from a single walk recorded with Meta Aria Gen2 glasses. The resulting hybrid scene pairs a Gaussian splat for appearance with a collision mesh derived from the glasses’ own SLAM points. Compared against same-day iPad LiDAR scans, the collision mesh achieves an F-score of 0.84–0.94 at 10 cm within 1 m of the ground, taking 51–160 $\times$ less time to build than a multi-view stereo mesh (MVS). Website: ego2sim.xyz

Keywords: Real-to-sim, Egocentric data, Collision geometry, Digital twin, Gaussian Splatting

1 Introduction

Navigation and locomotion policies are often trained in simulation as it is cheaper and safer than on hardware, and a growing line of work builds these simulators from real places [2, 3, 4, 5]. Following Vid2Sim [2], we call such an environment a *hybrid scene*: a 3D Gaussian splat [6] for appearance plus a collision mesh (the *collider*) for contact (Fig. 1). Image-only pipelines build such a scene

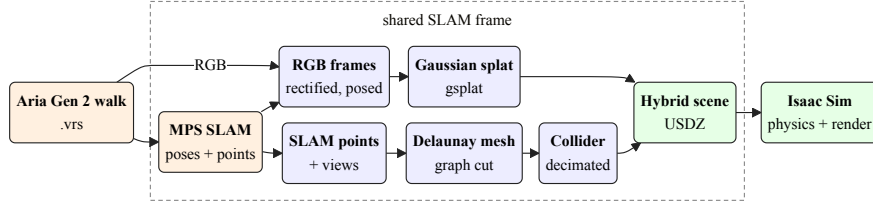


Fig. 2: Overview of the Ego2Sim pipeline. The top row builds the Gaussian splat and the bottom row builds the collision mesh.

from just a phone or a video, but structure-from-motion recovers neither true scale nor the direction of gravity, and both must be set some other way, such as aligning each scene to the simulator by hand [3]. Depth-equipped phones [5] and LiDAR scanners [7] measure true scale but are handheld, and pairing a scan with an egocentric recording takes a second device and a later alignment [8].

Egocentric glasses offer an alternative: they record the world hands-free from the wearer’s view-point [9], and their recordings already supply human demonstrations for robot learning [10]. We use Aria Gen 2 glasses [11], Meta’s egocentric research platform, whose Machine Perception Services (MPS) [12] return a metric trajectory and 3D points in a gravity-aligned *SLAM frame*. Building on egocentric splatting [13], we reconstruct a contact-enabled scene from egocentric capture alone, and we measure how closely its appearance and its contact geometry match the real place – which is the part of the sim-to-real gap that the scene itself introduces. Our contributions are as follows:

- **An egocentric real-to-sim pipeline.** To our knowledge, Ego2Sim is the first real-to-sim pipeline that builds a photoreal, collidable simulation scene from egocentric glasses alone, with no depth sensor or separate scanner. Its collider is meshed directly from the glasses’ own SLAM points, so it shares the splat’s frame and needs no manual registration.
- **An evaluation of both halves of the scene.** We compare our collider against same-day iPad LiDAR scans of five walks: our collider achieves an F-score of 0.84–0.94 at 10 cm on geometry up to 1 m above the ground, within 0.031 of a COLMAP multi-view-stereo (MVS) collider that takes 51–160× longer to build; on held-out frames, the splats reach a mean PSNR of 27.0–30.7 dB per walk.

2 Related Work

Real-to-sim from video. Vid2Sim [2] turns web video into hybrid scenes for urban navigation, but its metric accuracy is not evaluated. VR-Robo [3] builds a splat and collider in Isaac Sim from phone photos and aligns them to the simulator by hand. EmbodiedSplat [5] captures scenes with an iPhone and checks its meshes against laser scans, and GaussGym [4] renders splats inside a vectorized physics simulator. **Scene reconstruction from egocentric capture.** Lv et al. [13] reconstruct photoreal Gaussian splats from Aria Gen 1 recordings, and EgoLifter [14] decomposes Aria scenes into per-object Gaussians, but neither builds collision geometry or a simulator scene. FunREC [15] builds articulated mesh twins that export to simulators, but from a head-mounted RGB-D camera and without a splat, and EgoHTR [8] records terrain traversal with Aria glasses but scans the terrain with a separate laser scanner. To our knowledge, none builds a photoreal, collidable scene from the glasses alone.

3 Method

As shown in Fig. 2, an Aria Gen 2 capture yields MPS SLAM points and trajectory, from which we train a Gaussian splat and build a collider mesh. Both are exported into one USDZ scene for Isaac Sim and are aligned by construction because they share the SLAM frame.

58 **Capture:** The wearer walks through the scene for a few minutes. MPS [12] returns the trajectory at
59 1 kHz and the semi-dense SLAM points, each with a depth uncertainty.

60 **Collider from the SLAM points:** We keep low-uncertainty points from Aria’s MPS SLAM, treat
61 each SLAM camera at each observation time as a view, and pass both to COLMAP’s Delaunay
62 mesher [16], which labels tetrahedra free or occupied with a visibility graph cut [17]: each camera-
63 to-point ray votes for the space it crosses to be free. We drop triangles with edges over 2 m to
64 prevent the mesh spanning unobserved space, and decimate with quadric error metrics [18, 19]
65 (median deviation below 1 cm) to 17k–90k triangles.

66 **Gaussian splat:** We adapt Meta’s Aria splat trainer [13], built on gsplat [20], to Gen 2 and train 30k
67 steps at full resolution. MCMC densification [21] recycles faded Gaussians and adds new ones up
68 to 1.5 M; larger SLAM-point initializations are kept (final: 1.5–2.9 M). From step 10k, we model
69 the rolling shutter of the RGB camera (image rows exposed at slightly different times) and keep the
70 factory RGB calibration, which MPS does not refine.

71 **Hybrid scene export:** Adapting the 3dgrut exporter [22], we write the splat in NVIDIA’s NuRec
72 format, which Isaac Sim [1] renders with 3DGUT [23], and package it with the decimated mesh as
73 an invisible static collider, both with an identity transform, in one USDZ file.

74 4 Experimental Setup

75 **Data collection.** We recorded five walks of 4–6 min with Aria Gen 2: two indoor walks of one
76 bedroom (Room-1, Room-2) and three outdoor walks of a front yard (Yard) and two parking areas
77 (Hall, Parking Lot). On the same day, we scanned each place with iPad LiDAR using the RTAB-Map
78 app (v0.22.1) [24] as a reference.

79 **Reference geometry.** Consumer LiDAR scans are not survey-grade [25], so we quantify their error.
80 We align each scan to the SLAM frame by leveling its floor, searching heading and translation on
81 top-down occupancy maps, and refining with point-to-plane ICP [26] of the confident MPS points
82 onto the scan, instead of the collider being evaluated. The reference error $\sigma = (\sigma_{\text{scan}}^2 + \sigma_{\text{align}}^2)^{1/2}$,
83 combining the scan’s scatter around fitted planes with the ICP residual, is 2.3–3.0 cm, so we trust
84 only $\tau \geq 2\sigma$ and report $\tau = 10$ cm. On Yard, Hall and Parking Lot, an independent feature-matching
85 alignment agrees within 0.3°.

86 **Metrics.** Following Tanks and Temples [27], we report precision P (the fraction of the collider
87 within τ of the scan), recall R (the fraction of the scan within τ of the collider) and their harmonic
88 mean, the F-score $F = 2PR/(P + R)$, within each scan’s footprint and a *contact band* up to 1 m
89 above the local ground, which holds the geometry a wheeled or legged robot touches. *Ground recall*
90 is the fraction of the scanned ground within τ of the collider. As a baseline, we replace the SLAM
91 points with dense points from COLMAP MVS [28], computed from 972–1362 RGB keyframes per
92 walk with the MPS camera poses, and mesh and decimate them as in Sec. 3, so the two colliders
93 differ only in their input points. For the splat, we follow Lv et al. [13]: we hold out every 8th RGB
94 frame and compare each with its render using PSNR, SSIM [29] and LPIPS (AlexNet) [30].

95 5 Results

96 **Geometric accuracy.** Against the scans, an ICP fit with free scale puts the MPS scale within 0.1–
97 0.3% indoors and within about 1.5% outdoors, where the estimate shifts with fit direction, and MPS
98 gravity agrees with the scanned floor to within 1°, versus up to 2.6° for the scanning app’s own
99 “up”, so the SLAM frame is metric and gravity-aligned without any registration. On this frame, the
100 contact-band F at 10 cm is 0.84–0.94 and ground recall is 90.6–99.9% (Table 1). Indoors, precision
101 and recall are 0.90–0.93, and the two Room walks differ by 0.003 in F . Outdoors, recall is 0.95–
102 0.98, but precision falls to 0.75 on Yard and Hall, mostly from unconfirmed surfaces near the scan’s
103 edge and, on Yard, around the hedge and car. In both settings, the collider captures nearly all of the
104 scanned geometry that a robot would touch. As a qualitative check, Isaac Sim’s Spot example [1]

Table 1: Comparison of each walk’s contact-band accuracy at $\tau = 10$ cm and collider build time. Ours: collider from the SLAM points. ICP inliers and σ (reference error) describe each walk’s scan alignment. [†]Dense stereo ran on RTX 4090, all others on RTX 4080 SUPER.

Walk	Collider	ICP inliers	σ	Precision $\uparrow$	Recall $\uparrow$	F $\uparrow$	Ground recall $\uparrow$	Build time $\downarrow$
Room-1	Ours	89%	2.3 cm	0.899	0.925	0.912	95.1%	2.9 min
	MVS			0.945	0.913	0.929	87.9%	4.7 h [†]
Room-2	Ours	88%	2.3 cm	0.909	0.910	0.909	90.6%	2.4 min
	MVS			0.954	0.914	0.934	88.7%	5.1 h [†]
Yard	Ours	71%	3.0 cm	0.748	0.970	0.845	97.9%	2.5 min
	MVS			0.755	0.982	0.854	99.3%	6.1 h
Hall	Ours	77%	2.6 cm	0.754	0.948	0.840	98.5%	1.8 min
	MVS			0.788	0.972	0.871	99.6%	4.8 h [†]
Parking Lot	Ours	86%	2.5 cm	0.893	0.982	0.935	99.9%	7.0 min
	MVS			0.898	0.977	0.936	99.9%	5.9 h

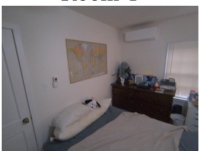
		Room-1	Yard	Hall				
original					Walk	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
					Room-1	30.70	0.913	0.463
					Room-2	30.24	0.913	0.466
					Yard	26.97	0.886	0.337
					Hall	29.27	0.937	0.382
render					Parking Lot	29.72	0.947	0.197
					Mean	29.38	0.919	0.369

Fig. 3: Splat quality on held-out frames (every 8th RGB frame). Left: example frames (top) and renders of the same views (bottom), with each frame’s PSNR in dB. Right: per-walk means.

105 with its unchanged flat-terrain policy walks on the collider of an earlier 2.6-minute front-yard scene
 106 and touches the car and the hedge (Fig. 1).

107 **Baseline comparison.** Across the five walks, MVS raises F by 0.001–0.031, mostly through preci-
 108 sion indoors (0.945–0.954 vs. 0.899–0.909) and on Hall (0.788 vs. 0.754), but lowers indoor ground
 109 recall (87.9–88.7% vs. 90.6–95.1%). It took 4.7–6.1 h per walk versus 2–7 min for ours. Our collider
 110 is thus far faster to build, with nearly the same accuracy and better indoor ground recall.

111 **Splat quality.** On held-out frames, the splats reach a mean PSNR of 27.0–30.7 dB (mean 29.4;
 112 Fig. 3), comparable to the 24.0–30.0 dB (mean 26.6) that Lv et al. [13] report on 12 Aria Gen 1
 113 scenes with the same protocol and code, despite our factory calibration; the scenes differ, so this is
 114 a reference point, not a head-to-head. Held-out frames also lie a median of 33 ms from a training
 115 frame, so our PSNRs thus measure fidelity along the walk, not from novel viewpoints such as a
 116 robot’s lower camera.

117 6 Conclusion

118 We presented a pipeline that builds contact-enabled Isaac Sim scenes from a single egocentric walk
 119 with Aria Gen 2 glasses. Across five indoor and outdoor walks, the collider reaches a contact-band
 120 F -score of 0.84–0.94 at 10 cm, with nearly the same accuracy as an MVS collider that takes hours
 121 instead of minutes to build. However, our study has two limitations. First, we neither train nor
 122 evaluate a policy in these scenes, and we test no real robot. Second, five walks is a small sample.
 123 In future work, we will train policies on teleoperation demonstrations in these scenes and on human
 124 demonstrations from the same recordings, and deploy them on a real robot. We will also capture
 125 more diverse indoor and outdoor places and release the reconstructed scenes as a benchmark.

References

- [1] NVIDIA. Isaac Sim. <https://github.com/isaac-sim/IsaacSim>, 2025. Version 5.1.0, software. Accessed 2026-10-07.
- [2] Z. Xie, Z. Liu, Z. Peng, W. Wu, and B. Zhou. Vid2Sim: Realistic and interactive simulation from video for urban navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1581–1591, 2025. doi:10.1109/CVPR52734.2025.00155.
- [3] S. Zhu, L. Mou, D. Li, B. Ye, R. Huang, and H. Zhao. VR-Robo: A real-to-sim-to-real framework for visual robot navigation and locomotion. *IEEE Robotics and Automation Letters*, 10(8):7875–7882, 2025. doi:10.1109/LRA.2025.3575648.
- [4] A. Escontrela, J. Kerr, A. Allshire, J. Frey, R. Duan, C. Sferrazza, and P. Abbeel. GaussGym: An open-source real-to-sim framework for learning locomotion from pixels. *arXiv preprint arXiv:2510.15352*, 2025.
- [5] G. Chhablani, X. Ye, M. Z. Irshad, and Z. Kira. EmbodiedSplat: Personalized real-to-sim-to-real navigation with Gaussian splats from a mobile device. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 25431–25441, 2025. doi:10.1109/ICCV51701.2025.02359.
- [6] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis. 3D Gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4):1–14, July 2023. doi:10.1145/3592433. URL <https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/>.
- [7] X. Liu, J. Li, Y. Deng, R. Chen, Y. Zhang, Y. Ma, L. Guo, Y. Li, J. Zhang, and C. Feng. Wanderland: Geometrically grounded simulation for open-world embodied AI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1041–1052, June 2026.
- [8] A. Brandes, H. C. G. Sajelian, M. Patel, D. Hollidt, C. Li, M. Heyrman, O. Hausdoerfer, M. Kaufmann, X. Wang, J. Frey, A. P. Schoellig, C. Holz, M. Pollefeys, and M. Hutter. EgoHTR: Egocentric 4D demonstrations of human terrain traversal. *arXiv preprint arXiv:2607.13472*, 2026.
- [9] J. Engel, K. Somasundaram, M. Goesele, A. Sun, A. Gamino, A. Turner, A. Talattof, A. Yuan, B. Souti, B. Meredith, C. Peng, C. Sweeney, C. Wilson, D. Barnes, D. DeTone, D. Caruso, et al. Project Aria: A new tool for egocentric multi-modal AI research, 2023. URL <https://arxiv.org/abs/2308.13561>.
- [10] S. Kareer, D. Patel, R. Punamiya, P. Mathur, S. Cheng, C. Wang, J. Hoffman, and D. Xu. EgoMimic: Scaling imitation learning via egocentric video. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 13226–13233, 2025.
- [11] Project Aria Team at Meta. Aria Gen 2: An advanced research device for egocentric AI research. Whitepaper, <https://www.projectaria.com/ariagen2devicepaper>, 2025. Accessed: 2026-10-15.
- [12] Meta Platforms, Inc. Project Aria machine perception services (MPS). Project Aria Docs (Aria Gen 2), https://facebookresearch.github.io/projectaria_tools/gen2/ark/mps, 2026. Accessed: 2026-10-15.
- [13] Z. Lv, M. Monge, K. Chen, Y. Zhu, M. Goesele, J. Engel, Z. Dong, and R. Newcombe. Photoreal scene reconstruction from an egocentric device. In *SIGGRAPH Conference Papers '25*, pages 1–11, New York, NY, USA, 2025. ACM. doi:10.1145/3721238.3730753.

- [14] Q. Gu, Z. Lv, D. Frost, S. Green, J. Straub, and C. Sweeney. EgoLifter: Open-world 3D segmentation for egocentric perception. In *Computer Vision – ECCV 2024, Part XLIII*, Lecture Notes in Computer Science, pages 382–400. Springer, 2024. doi:10.1007/978-3-031-72775-7_22.
- [15] A. Delitzas, C. Zhang, A. Gavryushin, T. Di Mario, B. Sun, R. Dabral, L. Guibas, C. Theobalt, M. Pollefeys, F. Engelmann, and D. Barath. FunREC: Reconstructing functional 3D scenes from egocentric interaction videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 28446–28457, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Delitzas_FUN_REC_Reconstructing_Functional_3D_Scenes_from_Egocentric_Interaction_CVPR_2026_paper.html.
- [16] J. L. Schönberger and J.-M. Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4104–4113, 2016. doi:10.1109/CVPR.2016.445.
- [17] P. Labatut, J.-P. Pons, and R. Keriven. Robust and efficient surface reconstruction from range data. *Computer Graphics Forum*, 28(8):2275–2290, 2009. doi:10.1111/j.1467-8659.2009.01530.x.
- [18] M. Garland and P. S. Heckbert. Surface simplification using quadric error metrics. In *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '97)*, pages 209–216. ACM Press, 1997. doi:10.1145/258734.258849.
- [19] Q.-Y. Zhou, J. Park, and V. Koltun. Open3D: A modern library for 3D data processing. *arXiv preprint arXiv:1801.09847*, 2018.
- [20] V. Ye, R. Li, J. Kerr, M. Turkulainen, B. Yi, Z. Pan, O. Seiskari, J. Ye, J. Hu, M. Tancik, and A. Kanazawa. gsplat: An open-source library for Gaussian splatting. *Journal of Machine Learning Research*, 26(34):1–17, 2025. URL <http://jmlr.org/papers/v26/24-1476.html>.
- [21] S. Kheradmand, D. Rebain, G. Sharma, W. Sun, Y.-C. Tseng, H. Isack, A. Kar, A. Tagliasacchi, and K. M. Yi. 3D Gaussian splatting as Markov chain Monte Carlo. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 80965–80986. Curran Associates, Inc., 2024. doi:10.52202/079017-2573.
- [22] NVIDIA Spatial Intelligence Lab. 3DGRUT: Official implementations of 3D Gaussian ray tracing (3DGRT) and 3D Gaussian unscented transform (3DGUT). <https://github.com/nv-tlabs/3dgrut>, 2026. GitHub repository nv-tlabs/3dgrut, Apache-2.0 license. Accessed: 2026-10-07.
- [23] Q. Wu, J. Martinez Esturo, A. Mirzaei, N. Moenne-Loccoz, and Z. Gojcic. 3DGUT: Enabling distorted cameras and secondary rays in Gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26036–26046, 2025. doi:10.1109/CVPR52734.2025.02425. URL <https://arxiv.org/abs/2412.12507>.
- [24] M. Labbé and F. Michaud. RTAB-Map as an open-source lidar and visual simultaneous localization and mapping library for large-scale and long-term online operation. *Journal of Field Robotics*, 36(2):416–446, 2019. doi:10.1002/rob.21831.
- [25] D. Treccani, A. Adami, and L. Fregonese. Assessing the effectiveness of LiDAR-based apps on Apple devices to survey indoor and outdoor medium sized areas. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLVIII-2/W8-2024: 431–438, 2024. doi:10.5194/isprs-archives-XLVIII-2-W8-2024-431-2024. URL <https://isprs-archives.copernicus.org/articles/XLVIII-2-W8-2024/431/2024/>.

- 217 [26] Y. Chen and G. Medioni. Object modelling by registration of multiple range images. *Image*
218 *and Vision Computing*, 10(3):145–155, Apr. 1992. doi:10.1016/0262-8856(92)90066-C.
- 219 [27] A. Knapitsch, J. Park, Q.-Y. Zhou, and V. Koltun. Tanks and temples: Benchmarking large-
220 scale scene reconstruction. *ACM Transactions on Graphics*, 36(4):1–13, July 2017. doi:10.
221 1145/3072959.3073599. URL <https://www.tanksandtemples.org/>.
- 222 [28] J. L. Schönberger, E. Zheng, J.-M. Frahm, and M. Pollefeys. Pixelwise view selection for
223 unstructured multi-view stereo. In *Computer Vision – ECCV 2016*, volume 9907 of *Lecture*
224 *Notes in Computer Science*, pages 501–518. Springer, 2016. doi:10.1007/978-3-319-46487-9_
225 31.
- 226 [29] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: From
227 error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–
228 612, Apr. 2004. doi:10.1109/TIP.2003.819861.
- 229 [30] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effec-
230 tiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Con-*
231 *ference on Computer Vision and Pattern Recognition (CVPR)*, pages 586–595, 2018. doi:
232 10.1109/CVPR.2018.00068.