VL-N3RD-Bench: Benchmarking **V**ision-**L**anguage **N**avigation with **3**D Gaussian Splatting **R**econstruction for **D**eployment

Yu-Lun Chou*, Lu-Ching Wang*, Ying-Sheng Luo*, Yu-Tang Liu, Jeng-Lin Li

***Abstract*— Recent advances in large language models (LLMs) have expanded robotic capabilities by bridging perception and action. Vision-language navigation (VLN) enables embodied agents to follow high-level semantic instructions in complex and unseen environments. Fine-tuning VLA models requires photorealistic training data, as insufficient visual fidelity may exacerbate sim-to-real gaps and degrade action execution. However, achieving such realism in simulation is challenging, and training is often conducted using synthetic views or egocentric video datasets. Gaussian Splatting (GS) has recently emerged as a promising solution for high-fidelity 3D scene reconstruction, offering a scalable alternative for generating training environments. Nevertheless, the impact of GS reconstruction quality on downstream VLN performance has not been systematically investigated. In this work, we benchmark state-of-the-art GS pipelines using public and customized datasets. We evaluate reconstruction quality via standard image-based metrics and computational efficiency. Finally, we use *VL-N3RD-Bench* to evaluate a pretrained VLA model on a Unitree A1 robot across simulated and physical environments. Our results demonstrate that overall reconstruction quality can align with stronger indoor VLN performance, but this relationship is inconsistent, and strong image-based metrics alone do not guarantee better outdoor navigation when geometry and traversability become more important.**

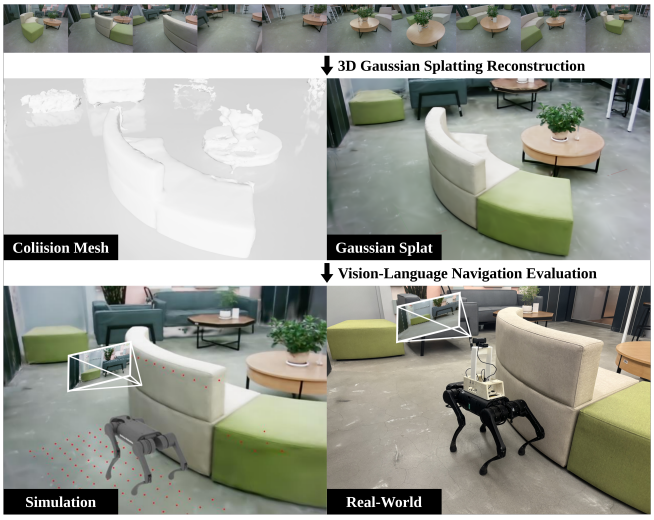

Fig. 1. Overview of the VL-N3RD-Bench pipeline. Multi-view image collections are processed by different 3DGS pipelines to reconstruct scene splats. A collision mesh is extracted from the SplaTAM [8] reconstruction and manually aligned with all splats. VL-N3RD-Bench then evaluates downstream navigation performance in simulation and the real world.

I. Introduction

Recent advances in Vision-Language Navigation (VLN) have enabled embodied agents to interpret high-level semantic instructions and navigate toward designated goals in complex, previously unseen environments. These systems can now address tasks such as following natural-language navigation commands [1] and searching for objects in indoor scenes [2]. However, training or fine-tuning VLN models requires large amounts of high-quality visual data, which are often generated in simulation. When the simulated scenes do not faithfully reflect real-world appearance and structure, a sim-to-real gap emerges and reduces the reliability of downstream action execution.

A practical way to narrow this gap is to reconstruct photorealistic scenes from captured images or videos and use them as simulation environments for embodied training and evaluation. In this setting, the key challenge is not only to render visually realistic scenes, but also to determine whether the reconstructed environments are truly useful for VLN. In particular, a reconstruction pipeline should preserve the visual cues that support language grounding, maintain sufficient geometric consistency for navigation, and remain efficient enough for practical deployment.

Creating such environments remains difficult. Synthetic-view generation [3] and scene generation methods [4] provide scalable ways to construct training data, but often sacrifice realism or scene diversity. Large-scale video datasets [5], [6] improve realism, but are less flexible when novel viewpoints or interactive environments are required. Recently, 3D Gaussian Splatting (3DGS) [7] has emerged as a promising alternative for high-fidelity 3D scene reconstruction, enabling efficient rendering of photorealistic environments from real image collections. This makes 3DGS a compelling candidate for generating realistic simulation environments and, potentially, useful training data for VLN.

Despite rapid progress in 3DGS reconstruction pipelines, their value for VLN has not been systematically studied. In particular, it remains unclear which properties of reconstruction quality matter most for downstream navigation, whether high-fidelity consistently improves sim-to-real transfer, and what practical limitations arise when reconstructed scenes are used for robotic deployment. These questions are especially important because image quality alone may not fully capture the requirements of embodied navigation.

In this work, we benchmark several state-of-the-art 3DGS pipelines on both public datasets and customized scenes. We evaluate reconstruction quality using PSNR, SSIM, and LPIPS, together with computation time and GPU mem-

*Joint first authors. All authors are with Inventec Corporation.

ory usage. We then use VL-N3RD-Bench to study how these upstream differences affect downstream navigation performance across real-world scenes and their reconstructed counterparts in simulation, as shown in Figure 1. Our results show that optimization-based pipelines consistently outperform feed-forward alternatives in reconstruction quality. More importantly, the VLN benchmark shows that navigation performance depends on scene characteristics and cannot be explained by any single reconstruction metric across all scenes. At the same time, the outdoor results show that strong visual metrics do not necessarily translate to better navigation success, indicating that appearance quality alone is insufficient to close the sim-to-real gap when geometry, traversability, and scene interaction become more critical.

Our main contributions are as follows:

- We present **VL-N3RD-Bench**, a unified benchmark for evaluating 3DGS reconstruction pipelines for VLN, jointly evaluating upstream visual fidelity and downstream navigation performance across simulation and real-world deployment.
- We compare representative GS pipelines on public benchmarks and on customized indoor and outdoor scenes, and analyze their trade-offs in visual quality, runtime, and GPU memory usage.
- We show that the relationship between reconstruction quality and VLN performance is inconsistent: stronger overall image-based reconstruction metrics can result in better indoor VLN performance, while they do not by themselves guarantee successful transfer in more complex outdoor scenes.
- We identify practical deployment challenges of current GS pipelines and VLN systems, including computational cost, sensitivity to data collection and prompt design, viewpoint mismatch between reconstruction and deployment, and the lack of a unified framework for aligning rendered scenes with collision-aware navigation geometry.

II. Related Work

Scene reconstruction methods have evolved from neural implicit representations such as NeRF [9] to Gaussian Splatting (GS) [7], driven by its explicit scene representation and capability for real-time rendering. Conventional GS pipelines [10], [8] typically rely on feature detection, extraction, and matching, followed by structure-from-motion (SfM) [11] for camera pose estimation. With the rise of transformer-based architectures, feed-forward methods [12], [13], [14] have recently gained popularity due to their near real-time inference speed for GS generation.

The availability of large-scale datasets has also significantly improved scene diversity for training and evaluation. Datasets with depth information include [15], [16], while RGB-only datasets such as [17], [18] provide both indoor and outdoor scenes. In addition, synthetic datasets [19], [3] offer scalable and controllable environments for training and benchmarking.

Vision-Language Navigation (VLN) enables embodied agents to follow natural language instructions in complex environments. Recent approaches [1], [2], [5] leverage pre-trained vision-language-action (VLA) models for improved generalization and long-horizon reasoning. However, VLN performance is highly dependent on the quality of training environments, as sim-to-real gaps can degrade navigation reliability. Prior works address this using synthetic data, video datasets, or reconstructed 3D scenes, but the effect of reconstruction fidelity—especially with Gaussian Splatting—on downstream VLN performance remains underexplored. Recent work such as SAGE-3D [20] explores training VLA models directly on 3DGS scenes for embodied navigation; however its focus is on environment construction rather than analyzing how reconstruction quality impacts sim-to-real transfer. In contrast, our work explicitly benchmarks this relationship, providing a systematic study of how visual fidelity influences real-world navigation success.

III. Experimental Setup

We present the setup of VL-N3RD-Bench for evaluating 3D Gaussian Splatting (3DGS) methods for Vision-Language Navigation (VLN). The benchmark jointly measures upstream reconstruction quality and downstream navigation performance within a unified evaluation pipeline. Specifically, we compare representative 3DGS pipelines with different pose requirements, supervision signals, and reconstruction paradigms. We then assess their scene fidelity on public datasets using standard image-based metrics. Finally, we use VL-N3RD-Bench to examine how reconstruction quality affects navigation success in both simulation and real-world environments using two custom indoor scenes and one outdoor scene. In addition to reconstruction accuracy, we also analyze computational cost and the practical challenges of deploying reconstructed scenes for collision-aware robotic navigation.

Our experiments are designed to answer the following questions:

- Which GS method performs best across indoor and outdoor scenes under the selected reconstruction metrics?
- How does reconstruction fidelity influence the sim-to-real gap in Vision-Language Navigation tasks?
- What trade-offs and limitations do current state-of-the-art GS pipelines present for robotic navigation?

A. Gaussian Splatting Pipeline Candidates

Our comparison focuses on several pipeline properties that directly affect VLN. The first is **pose dependency**, namely whether a method requires camera poses from an external structure-from-motion (SfM) system or estimates them internally. This affects data collection and deployment complexity. The second is the **reconstruction paradigm**. Optimization-based methods typically offer strong rendering quality, whereas feed-forward methods are usually faster and more scalable. The third is the **training objective**. Methods may optimize only photometric consistency or additionally incorporate depth and geometric constraints,

TABLE I
OVERVIEW OF THE EVALUATED 3DGS PIPELINES.

Method	Paradigm	Main objective	Relevant property for VLN	Metric scale
3DGUT	SfM + Optimization + RGB	Photometric reconstruction	Accessible + High rendering fidelity	✗
Vid2Sim	SfM + Optimization + RGB	Photometric + depth + geometric constraints	Improved structural accuracy + High rendering fidelity	✗
VGGT	Feed-forward + Optimization + RGB	Learned geometric prior + GS training	Fast and scalable inference	✗
SplaTAM	Online RGB-D SLAM	RGB-D tracking and mapping consistency	Geometry-aware online reconstruction	✓

which can improve structural accuracy. Finally, for embodied navigation, we consider **interaction support**, i.e., whether the reconstructed scene can be converted into a geometry suitable for collision-aware planning.

Based on these criteria, we evaluate four representative methods: 3DGUT [10], Vid2Sim [21], VGGT [22], and SplaTAM [8]. These methods span both optimization-based and feed-forward pipelines, and differ in their pose assumptions, supervision, and suitability for downstream robotic use. Table I summarizes the main characteristics of the four methods.

3DGUT is an optimization-based 3DGS pipeline that relies on camera poses estimated by an external SfM system, such as COLMAP [23] or GLOMAP [24]. It mainly minimizes photometric reconstruction error and serves as a strong baseline for high-fidelity rendering. However, it does not explicitly enforce geometric consistency beyond image reconstruction.

Vid2Sim also relies on externally estimated camera poses, but extends standard 3DGS optimization with depth supervision and geometry-aware constraints. In addition to photometric loss, it uses scale-invariant and geometry-consistent objectives to improve structural accuracy. This makes it more suitable for embodied tasks in which geometric correctness is critical.

VGGT adopts a feed-forward transformer-based framework that directly predicts camera poses and scene geometry from visual input. It does not require an external SfM pipeline. Instead, it leverages learned geometric priors to infer scene structure in a single forward pass. In our implementation, the predicted poses and point clouds are further refined through 3DGS optimization using the gsplat library [25]. To enable inference on large image sets, we use a public low-VRAM implementation of VGGT [26].

SplaTAM is an online RGB-D Gaussian SLAM method that jointly performs camera tracking and scene mapping. Unlike offline pipelines, it incrementally estimates poses and updates the scene from streaming RGB-D observations. By optimizing RGB-D consistency for both tracking and mapping, it is well suited to robotics settings with online sensor input. Its depth measurements further enable metric-scale reconstruction and provide a direct geometric basis for the recovered scene.

from RGB-D observations. Its optimization is based on RGB-D consistency for both tracking and mapping. This online design makes it especially relevant to robotics settings with streaming sensor input. In addition, its depth measurements provide a direct basis for geometric scene reconstruction

Standard 3DGS representations are designed for rendering and do not directly support agent-environment interaction. To enable collision-aware navigation, we construct Truncated Signed Distance Function (TSDF) representations [27] in SplaTAM. We then extract collision meshes and applied to reconstructed scenes across all methods for downstream evaluation. This process is particularly effective for methods such as SplaTAM that provide depth measurements directly.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

We first evaluate reconstruction quality on public benchmarks and then study how these differences affect downstream Vision-Language Navigation (VLN) on our custom scenes. Finally, we summarize the main findings with respect to the questions posed in Section III.

A. Gaussian Splatting Reconstruction Quality

Following SplaTAM [8], we evaluate reconstruction quality using PSNR, SSIM [28], and LPIPS [29] (AlexNet variant). PSNR measures pixel-wise reconstruction accuracy, SSIM measures structural consistency, and LPIPS captures perceptual similarity using learned visual features. We benchmark GS pipelines on the public ScanNet++ [15], Replica [16], DL3DV-10K [17], and Tanks&Temples (T&T) [18] datasets.

Table II summarizes reconstruction performance across indoor and outdoor scenes. Overall, Vid2Sim achieves the strongest visual fidelity. It attains the best PSNR on four of the six evaluated settings, the best SSIM on five, and the best LPIPS on all, indicating consistently strong performance in both structural and perceptual quality. 3DGUT is the closest competitor. It slightly outperforms Vid2Sim on Replica and outdoor DL3DV. In contrast, VGGT performs worse on all three visual metrics across all datasets. A likely reason is that errors in its predicted camera poses propagate to the subsequent GS optimization. In addition, although VGGT outputs COLMAP-format sparse point clouds, these are used only to initialize the Gaussians; inaccuracies in the predicted

TABLE II

COMPARISON OF RECONSTRUCTION PERFORMANCE ACROSS DIFFERENT SCENES AND PIPELINES.

Methods	Metrics	Indoor				Outdoor	
		ScanNet++ [15]	Replica [16]	DL3DV [17]	T&T [18]	DL3DV [17]	T&T [18]
3DGUT [10]	PSNR↑	33.79	**39.93**	30.60	20.07	**28.86**	20.03
	SSIM↑	0.954	**0.970**	0.927	0.697	0.857	0.685
	LPIPS↓	0.228	0.124	0.097	0.395	**0.116**	0.368
	Time (hr)	0.399	0.195	0.250	0.520	0.365	0.332
	GPU (GB)	**1.3**	**1.0**	**1.5**	**3.6**	**2.4**	**3.8**
Vid2Sim [21]	PSNR↑	**33.90**	37.90	**31.53**	**20.88**	28.56	**24.37**
	SSIM↑	**0.969**	0.966	**0.966**	**0.753**	**0.917**	**0.887**
	LPIPS↓	**0.107**	**0.112**	**0.078**	**0.373**	0.136	**0.210**
	Time (hr)	0.522	0.246	0.269	0.594	0.324	0.541
	GPU (GB)	23.2	6.2	3.8	13.1	4.4	15.7
VGGT [22]	PSNR↑	22.48	27.94	23.23	16.23	19.62	16.49
	SSIM↑	0.837	0.907	0.767	0.578	0.557	0.565
	LPIPS↓	0.369	0.203	0.246	0.645	0.344	0.582
	Time (hr)	**0.230**	**0.151**	**0.129**	**0.192**	**0.137**	**0.181**
	GPU (GB)	20.9	22.5	16.5	16.3	16.7	15.4
SplaTAM [8]	PSNR↑	24.10	20.31	–	–	–	–
	SSIM↑	0.884	0.706	–	–	–	–
	LPIPS↓	0.178	0.375	–	–	–	–
	Time (hr)	1.351	0.929	–	–	–	–
	GPU (GB)	8.3	7.3	–	–	–	–
Scene Amount		22	18	26	8	28	13

geometry can therefore degrade initialization quality and hinder convergence. SplaTAM is only evaluated on ScanNet++ and Replica because its RGB-D formulation requires depth input, which is unavailable for DL3DV and Tanks&Temples.

The methods also differ substantially in computational cost and memory usage. 3DGUT is the most memory-efficient pipeline, requiring only 1.0–3.8 GB of GPU memory across all reported scenes. VGGT is the fastest, but its feed-forward design comes with substantially higher memory demand, reaching 15.4–22.5 GB. Vid2Sim provides the best overall fidelity, but typically requires more memory and longer optimization time than 3DGUT. SplaTAM is slower than the other methods on indoor datasets and produces weaker image quality, although its RGB-D formulation provides direct metric-scale information that is useful for downstream scene alignment.

These results establish a clear upstream trend: optimization-based pipelines remain the strongest choice for high-fidelity reconstruction, with Vid2Sim offering the best overall quality and 3DGUT providing the best quality-efficiency trade-off.

B. Vision-Language Navigation Benchmark

To study how reconstruction quality affects sim-to-real transfer in VLN, we evaluate on three custom scenes: two indoor scenes, *Lobby* and *Office*, and one outdoor scene, *Park*. Compared with generic public datasets, these scenes provide a more controlled benchmark for analyzing how reconstruction fidelity influences downstream navigation. We use the pretrained VLN model from NaVILA [5] to interpret semantic instructions and generate high-level language commands. These commands are executed by a low-level locomotion policy trained for a Unitree A1 quadruped using DreamWaQ [30]. Figure 2 shows representative real-world trajectory and render result comparison across 4 pipelines in the *Office* scene.

Instruction: Walk straight and turn left near the red chair, walk to the white box and stop.

Fig. 2. Representative real-world VLN experiments in the *Office* scene, showing trajectory frames, input language instructions, and output commands, as well as comparison of render results across all pipelines.

For a fair comparison across reconstruction pipelines, we normalize the scale of all reconstructed scenes using the Kabsch–Umeyama algorithm [31]. The remaining rigid transformation is manually aligned. We use the SplaTAM reconstruction as the scale reference because its depth input preserves dimension of real-world scene more reliably. In simulation, a navigation episode is considered successful when the VLN model outputs a *finish* command and the robot is within a scene-specific distance threshold, defined as the median goal distance in that scene. While in real-world settings, we use 3 meters and 5 meters, respectively, for indoor and outdoor to consider as success. For each scene, we perform 25 tests to compute the Success Rate (SR).

In Table III, we report both success rates on reconstructed scenes in simulation and the one in real-world. The results show that the relationship between reconstruction fidelity and navigation performance is scene-dependent. In the indoor *Office* scene, the trend is particularly clear: Vid2Sim achieves the best scores in all metrics and also the highest navigation success rate in simulation (72%), which is closest to the real-world success rate of 80%. 3DGUT ranks second in both perceptual quality and simulated success rate, while SplaTAM performs substantially worse. This result suggests that, for indoor VLN, better overall reconstruction quality is generally associated with stronger downstream policy performance, though the relationship is not absolute.

The *Lobby* scene shows a similar but weaker trend. Vid2Sim again achieves the best reconstruction metrics, but 3DGUT attains a slightly higher simulated success rate (64% versus 60%). SplaTAM, despite its poor visual quality, also reaches 60% success. This indicates that in relatively

TABLE III

COMPARISON OF VISUAL LANGUAGE NAVIGATION PERFORMANCE ACROSS DIFFERENT PIPELINES WITH CUSTOMIZED DATASETS.

Methods	Metrics	Lobby 	Office 	Park 
3DGUT [10]	PSNR ↑	36.56	32.43	29.11
	SSIM ↑	0.954	0.935	0.893
	LPIPS ↓	0.191	0.180	0.095
	SR	**64%**	60%	**48%**
Vid2Sim [21]	PSNR ↑	**37.48**	**34.12**	**31.72**
	SSIM ↑	**0.982**	**0.976**	**0.980**
	LPIPS ↓	**0.138**	**0.121**	**0.073**
	SR	60%	**72%**	4%
VGGT [22]	PSNR ↑	30.91	28.02	22.08
	SSIM ↑	0.913	0.886	0.492
	LPIPS ↓	0.202	0.216	0.366
	SR	16%	44%	0%
SplaTAM [8]	PSNR ↑	15.86	14.24	12.86
	SSIM ↑	0.542	0.508	0.277
	LPIPS ↓	0.314	0.384	0.525
	SR	60%	24%	8%
Real-world SR		52%	80%	36%

simple indoor scenes, coarse geometry and a limited set of stable semantic cues may already be sufficient for successful navigation. In such cases, improvements in image fidelity do not always translate into proportional gains in task success.

The outdoor *Park* scene reveals a different trend from the indoor cases. Although Vid2Sim achieves the strongest reconstruction quality in this scene (31.72 PSNR, 0.980 SSIM, and 0.073 LPIPS), its navigation success rate in simulation is only 4%. In contrast, 3DGUT attains a substantially higher success rate of 48%, despite having weaker visual metrics overall, while the real-world success rate is 36%. This result suggests that high visual fidelity alone does not guarantee better sim-to-real transfer in more complex outdoor environments. Compared with indoor scenes, successful outdoor navigation appears to depend more strongly on factors beyond appearance quality, such as geometric consistency, traversability, and the alignment between reconstructed scenes and the navigation simulator. Overall, these results indicate that perceptual fidelity is an important factor for indoor VLN, but it is not sufficient by itself to explain downstream performance in outdoor settings.

C. Discussion

We summarize the main findings of the benchmark with respect to the research questions introduced in Section III.

a) Best GS pipeline across indoor and outdoor scenes: Across both public benchmarks and custom scenes, optimization-based pipelines consistently outperform the feed-forward VGGT baseline in reconstruction quality. Vid2Sim provides the strongest overall visual fidelity, achieving the best PSNR, SSIM, and LPIPS scores across most datasets. However, downstream VLN performance is more scene-dependent. In indoor environments, Vid2Sim generally yields the strongest navigation results, while 3DGUT remains highly competitive and achieves the best overall balance between reconstruction quality and efficiency, with by far the lowest GPU memory usage. Moreover, in the outdoor *Park* scene, 3DGUT outperforms Vid2Sim in navigation success despite weaker image-based reconstruction metrics, suggesting that stronger visual fidelity does not always translate to better downstream performance in more complex environments. VGGT offers the fastest runtime, but its lower reconstruction accuracy and high VRAM demand limit its practical value in resource-constrained settings. SplaTAM is useful when metric-scale reconstruction is required and depth input is available, but its rendering quality remains significantly below that of the optimization-based alternatives.

b) Which visual metrics matter most for VLN: Our results suggest that the relationship between image-based reconstruction quality and downstream VLN performance is inconsistent, rather than being captured by any single metric across scenes. In particular, VLN agents rely on recognizable scene layout, object appearance, and local landmarks to ground semantic instructions. The *Office* results support this view most clearly: the method with the best overall reconstruction metrics also achieves the highest navigation success rate. The *Lobby* results, however, show that this relationship is not absolute, as relatively simple scenes may remain navigable even when rendering quality is weaker. Overall, these findings indicate that the relationship between reconstruction quality and VLN performance depends on the scene; in some indoor cases, better overall image-based metrics align with stronger navigation performance, but this trend is not consistent across all scenes.

c) What limits sim-to-real transfer beyond appearance quality: The *Park* results show that appearance quality is only one part of the problem. Even when a reconstruction achieves strong PSNR, SSIM, and LPIPS, the downstream policy can still fail if the scene does not provide sufficiently accurate geometry, traversability cues, or collision-aware interaction. This finding suggests that closing the sim-to-real gap for VLN requires more than photorealistic rendering. It also requires physically meaningful scene representations that preserve scale, support navigation constraints, and remain stable under deployment conditions.

d) Practical limitations of current GS pipelines: Current GS pipelines remain computationally expensive. Both training time and VRAM usage increase rapidly with the number of input images, which limits scalability in large outdoor environments and constrains the level of detail that can be preserved in complex indoor scenes. Reconstruction quality also depends strongly on data collection. In particular, camera trajectory and viewpoint diversity have a large impact on the final result. This sensitivity becomes especially problematic in geometrically constrained settings, such as narrow spaces or long hallways, where insufficient viewpoint coverage can significantly degrade reconstruction quality. Furthermore, there is still no unified framework for aligning reconstructed splats with collision meshes, which introduces additional manual effort. Since most existing pipelines rely only on RGB input, extra rescaling is often required to

recover real-world dimensions, and alignment across reconstructions from different pipelines remains challenging.

e) VLN limitations: The downstream VLN policy also introduces limitations that are independent of scene reconstruction quality. First, navigation performance is sensitive to prompt formulation, since small changes in language instructions can alter how the model interprets goals and intermediate cues. Second, there is a mismatch between the camera viewpoint used during scene capture and reconstruction and the viewpoint observed during deployment. In particular, differences in camera pose can change the perceived layout and object appearance, which may reduce the consistency between reconstructed observations and real-world inputs. These factors introduce additional sources of sim-to-real error, even when the reconstructed scene itself is visually accurate.

V. Conclusion

In this work, we present a benchmark for studying the impact of 3D Gaussian Splatting (3DGS) reconstruction quality on Vision-Language Navigation (VLN). Through experiments on both public datasets and customized indoor and outdoor scenes, we find that optimization-based pipelines consistently outperform feed-forward alternatives in reconstruction quality. Reconstruction quality and VLN performance are not entirely related, with stronger overall image-based metrics aligning with better performance in some indoor scenes but not consistently across all settings. At the same time, our outdoor results show that strong image-based metrics do not by themselves guarantee better navigation performance, highlighting the importance of geometry, traversability, and interaction-aware scene representations for reducing the sim-to-real gap.

In future work, we will explore fine-tuning VLA models directly on high-fidelity Gaussian Splatting scenes rather than conventional triangle-mesh environments such as Matterport3D [32], to better exploit realistic visual priors. We also aim to investigate more advanced surface reconstruction methods to improve collision mesh generation and spatial reasoning beyond current TSDF-based solutions.

References

[1] Z. Zhang et al., "ActiveVLN: Towards Active Exploration via Multi-Turn RL in Vision-and-Language Navigation," Sep. 16, 2025. arXiv: 2509.12618, pre-published.

[2] L. Zhang et al., "NavA^3: Understanding Any Instruction, Navigating Anywhere, Finding Anything," Aug. 6, 2025. arXiv: 2508.04598, pre-published.

[3] H. Xia et al., "SAGE: Scalable Agentic 3D Scene Generation for Embodied AI," Feb. 20, 2026. arXiv: 2602.10116, pre-published.

[4] C. Fang et al., "SPATIALGEN: Layout-guided 3D Indoor Scene Generation," Jan. 15, 2026. arXiv: 2509.14981, pre-published.

[5] A.-C. Cheng et al., "Navila: Legged robot vision-language-action model for navigation," in *RSS*, 2025.

[6] N. Hirose, C. Glossop, D. Shah, and S. Levine, "OmniVLA: An Omni-Modal Vision-Language-Action Model for Robot Navigation," Sep. 23, 2025. arXiv: 2509.19480, pre-published.

[7] B. Kerbl, G. Kopanas, T. Leimkuehler, and G. Drettakis, "3D Gaussian Splatting for Real-Time Radiance Field Rendering," *ACM Trans. Graph.*, vol. 42, no. 4, 139:1–139:14, 26, 2023, ISSN: 0730-0301.

[8] N. Keetha et al., "SplaTAM: Splat, Track & Map 3D Gaussians for Dense RGB-D SLAM," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2024, pp. 21 357–21 366.

[9] B. Mildenhall et al., "NeRF: Representing scenes as neural radiance fields for view synthesis," *Commun. ACM*, vol. 65, no. 1, pp. 99–106, 17, 2021, ISSN: 0001-0782.

[10] Q. Wu et al., "3DGUT: Enabling Distorted Cameras and Secondary Rays in Gaussian Splatting," in *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2025, pp. 26 036–26 046.

[11] J. L. Schönberger and J.-M. Frahm, "Structure-from-Motion Revisited," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 4104–4113.

[12] H. Lin et al., "Depth Anything 3: Recovering the Visual Space from Any Views," in *The Fourteenth International Conference on Learning Representations*, Oct. 8, 2025.

[13] K. Zhang et al., "Gs-lrm: Large reconstruction model for 3d gaussian splatting," in *European Conference on Computer Vision*, Springer, 2024, pp. 1–19.

[14] C. Ziwen et al., "Long-lrm: Long-sequence large reconstruction model for wide-coverage gaussian splats," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 4349–4359.

[15] C. Yeshwanth, Y.-C. Liu, M. Nießner, and A. Dai, "Scannet++: A high-fidelity dataset of 3d indoor scenes," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 12–22.

[16] J. Straub et al., *The Replica Dataset: A Digital Replica of Indoor Spaces*, Jun. 13, 2019. arXiv: 1906.05797, pre-published.

[17] L. Ling et al., "Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 22 160–22 169.

[18] A. Knapitsch, J. Park, Q.-Y. Zhou, and V. Koltun, "Tanks and temples: Benchmarking large-scale scene reconstruction," *ACM Trans. Graph.*, vol. 36, no. 4, 78:1–78:13, 2017, ISSN: 0730-0301.

[19] Z. Lv et al., "Aria Everyday Activities Dataset," Feb. 22, 2024. arXiv: 2402.13349, pre-published.

[20] B. Miao et al., "Towards Physically Executable 3D Gaussian for Embodied Navigation," Dec. 15, 2025. arXiv: 2510.21307, pre-published.

[21] Z. Xie et al., "Vid2Sim: Realistic and Interactive Simulation from Video for Urban Navigation," in *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2025, pp. 1581–1591.

[22] J. Wang et al., "VGGT: Visual Geometry Grounded Transformer," in *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2025, pp. 5294–5306.

[23] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4104–4113.

[24] L. Pan, D. Baráth, M. Pollefeys, and J. L. Schönberger, "Global structure-from-motion revisited," in *European Conference on Computer Vision*, Springer, 2024, pp. 58–77.

[25] V. Ye et al., "Gsplat: An open-source library for gaussian splatting," *Journal of Machine Learning Research*, vol. 26, no. 34, pp. 1–17, 2025.

[26] Harry7557558. "VGGT-low-vram: Low VRAM implementation for Visual Geometry Grounded Transformer," Accessed: Apr. 13, 2026. [Online]. Available: https://github.com/harry7557558/vggt-low-vram.

[27] Q.-Y. Zhou and V. Koltun, "Dense scene reconstruction with points of interest," *ACM Transactions on Graphics (ToG)*, vol. 32, no. 4, pp. 1–8, 2013.

[28] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004, ISSN: 1941-0042.

[29] R. Zhang et al., "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 586–595.

[30] I. M. A. Nahrendra, B. Yu, and H. Myung, "Dreamwaq: Learning robust quadrupedal locomotion with implicit terrain imagination via deep reinforcement learning," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2023, pp. 5078–5084.

[31] S. Umeyama, "Least-squares estimation of transformation parameters between two point patterns," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 13, no. 4, pp. 376–380, 2002.

[32] A. Chang et al., "Matterport3d: Learning from rgb-d data in indoor environments," *International Conference on 3D Vision (3DV)*, 2017.