Novelty Assessment Report

Paper: Towards Physically Executable 3D Gaussian for Embodied Navigation

PDF URL: https://openreview.net/pdf?id=HB6KvsqcAn

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2025-12-27

Abstract

3D Gaussian Splatting (3DGS), a 3D representation method with photorealistic real-time rendering capabilities, is regarded as an effective tool for narrowing the sim-to-real gap. However, it lacks fine-grained semantics and physical executability for Visual-Language Navigation (VLN). To address this, we propose **SAGE-3D** (**S**emantically and Physically **A**ligned **G**aussian **E**nvironments for **3D** Navigation), a new paradigm that upgrades 3DGS into an executable, semantically and physically aligned environment. It comprises two components: **(1) Object-Centric Semantic Grounding**, which adds object-level fine-grained annotations to 3DGS; and **(2) Physics-Aware Execution Jointing**, which embeds collision objects into 3DGS and constructs rich physical interfaces. We release **InteriorGS**, containing 1K object-annotated 3DGS indoor scene data, and introduce **SAGE-Bench**, the first 3DGS-based VLN benchmark with 2M VLN data. Experiments show that 3DGS scene data is more difficult to converge, while exhibiting strong generalizability, improving baseline performance by 31% on the VLN-CE Unseen task.

Core Task Landscape

This paper addresses: **Visual-Language Navigation in 3D Gaussian Splatting Environments**

A total of **32 papers** were analyzed and organized into a taxonomy with **12 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Semantic 3D Gaussian Splatting Representations**
- **Navigation Frameworks and Policies**
- **Sim-to-Real Transfer and Embodied AI Platforms**
- **Related Embodied AI Applications**

Complete Taxonomy Tree

- Visual-Language Navigation in 3D Gaussian Splatting Environments Survey Taxonomy
- Semantic 3D Gaussian Splatting Representations
  - Open-Vocabulary Semantic Grounding (5 papers)
  - [2] 3d gaussian map with open-set semantic grouping for vision-language navigation (J Gao, 2025) View paper
  - [4] Fmgs: Foundation model embedded 3d gaussian splatting for holistic 3d scene understanding (Xingxing Zuo, 2025) View paper
  - [21] 3D Vision-Language Gaussian Splatting (Peng, 2024) View paper
  - [26] OVSG-SLAM: Open-Vocabulary Semantic Gaussian Splatting SLAM (Z Liu, 2025) View paper
  - [31] CoLaSplat: Compact Language 3D Gaussian Splatting (J Xu, n.d.) View paper
  - Multi-Granularity Semantic Representations (3 papers)
  - [14] A Neural Representation Framework with LLM-Driven Spatial Reasoning for Open-Vocabulary 3D Visual Grounding (Liu ZhenYang, 2025) View paper
  - [22] FMLGS: Fast Multilevel Language Embedded Gaussians for Part-level Interactive Agents (Tan Xin, 2025) View paper
  - [23] ReasonGrounder: LVLM-Guided Hierarchical Feature Splatting for Open-Vocabulary 3D Visual Grounding and Reasoning (Liu ZhenYang, 2025) View paper
  - Temporal and Dynamic Scene Representations (2 papers)
  - [9] 4d langsplat: 4d language gaussian splatting via multimodal large language models (Li Wanhua, 2025) View paper
  - [28] Towards Integrated Multimodal Interaction: Merging Immersive 3D Worlds with Language Based Retrieval for 3D Scene Understanding (Matuszek, 2025) View paper
- Navigation Frameworks and Policies
  - Vision-Language Navigation in Continuous Environments
  - Language-Guided Task Execution ★ (4 papers)
    - [0] Towards Physically Executable 3D Gaussian for Embodied Navigation (Anon et al., 2026) View paper
    - [5] ATLAS Navigator: Active Task-driven LAnguage-embedded Gaussian Splatting (Tao, 2025) View paper
    - [18] LagMemo: Language 3D Gaussian Splatting Memory for Multi-modal Open-vocabulary Multi-goal Visual Navigation (Zhou Hao-tian, 2025) View paper
    - [20] RoboTidy: A 3D Gaussian Splatting Household Tidying Benchmark for Embodied Navigation and Action (Xiaoquan Sun, 2025) View paper
  - Trajectory Planning and Viewpoint Synthesis (2 papers)
    - [7] Unitedvln: Generalizable gaussian splatting for continuous vision-language navigation (Dai, 2024) View paper
    - [8] YOPO-Nav: Visual Navigation using 3DGS Graphs from One-Pass Videos (Ryan Meegan, 2025) View paper
  - Image-Goal and Instance-Level Navigation (5 papers)
  - [6] Gaussnav: Gaussian splatting for visual navigation (Xiaohan Lei, 2025) View paper

- [11] IGL-Nav: Incremental 3D Gaussian Localization for Image-goal Navigation (Guo Wenxuan, 2025) View paper
- [12] GSplatVNM: Point-of-View Synthesis for Visual Navigation Models Using Gaussian Splatting (Honda Kohei, 2025) View paper
- [24] SplatSearch: Instance Image Goal Navigation for Mobile Robots using 3D Gaussian Splatting and Diffusion Models (Siddarth Narasimhan, 2025) View paper
- [25] BEINGS: Bayesian Embodied Image-goal Navigation with Gaussian Splatting (Wugang Meng, 2024) View paper
- Safe and Real-Time Robot Navigation (2 papers)
- [1] Splat-nav: Safe real-time robot navigation in gaussian splatting maps (Timothy Chen, 2025) View paper
- [27] Go-SLAM: Grounded Object Segmentation and Localization with Gaussian Splatting SLAM (Pham Phu, 2024) View paper
- Active Exploration and Task-Relevant Mapping (2 papers)
- [10] VISTA: Open-Vocabulary, Task-Relevant Robot Exploration with Online Semantic Gaussian Splatting (Nagami, 2025) View paper
- [30] LiLMaps: Learnable Implicit Language Maps (Evgenii Kruzhkov, 2025) View paper
- Aerial and Drone Navigation (2 papers)
- [17] GRaD-Nav++: Vision-Language Model Enabled Visual Drone Navigation with Gaussian Radiance Fields and Differentiable Dynamics (Chen, 2025) View paper
- [19] OpenFly: A Comprehensive Platform for Aerial Vision-Language Navigation (Gao Yunpeng, 2025) View paper

- Sim-to-Real Transfer and Embodied AI Platforms
  - Real-to-Sim-to-Real Pipelines (2 papers)
  - [15] EmbodiedSplat: Personalized Real-to-Sim-to-Real Navigation with Gaussian Splats from a Mobile Device (Chhablani, 2025) View paper
  - [16] VR-Robo: A Real-to-Sim-to-Real Framework for Visual Robot Navigation and Locomotion (Shaoting Zhu, 2025) View paper
  - Benchmarks and Evaluation Platforms (2 papers)
  - [13] RealMirror: A Comprehensive, Open-Source Vision-Language-Action Platform for Embodied AI (Zheng Zhao-yu, 2025) View paper
  - [32] Visually-grounded Humanoid Agents (H Ye, n.d.) View paper
- Related Embodied AI Applications (2 papers)
  - [3] End-to-end autonomous driving: Challenges and frontiers (Chitta Kashyap, 2024) View paper
  - [29] Embodied Navigation in Unknown Environments with Implicit Scene Memory and Target-aware Memory Retrieval (Qiming Liu, 2025) View paper

Narrative

Core task: Visual-Language Navigation in 3D Gaussian Splatting environments. This emerging field sits at the intersection of neural scene representations and embodied AI, where agents must interpret natural language instructions to navigate photorealistic 3D spaces reconstructed via Gaussian splatting. The taxonomy reveals four main branches: semantic 3D Gaussian splatting representations that enrich scene geometry with language-grounded features (e.g., Vision-Language Gaussian Splatting[21], FMGS[4]); navigation frameworks and policies that design control strategies for continuous or discrete movement in these environments (e.g., Splat-nav[1], GaussNav[6]); sim-to-real transfer and embodied AI platforms addressing the gap between synthetic training and physical deployment (e.g., BEINGS[25], RealMirror[13]); and related embodied AI applications extending beyond pure navigation to manipulation and multi-task scenarios (e.g., RoboTidy[20], VR-Robo[16]). These branches collectively capture how Gaussian splatting's rendering efficiency and geometric fidelity enable richer vision-language grounding compared to traditional mesh or voxel representations.

Recent work has concentrated on two contrasting themes: memory-augmented architectures that maintain spatial-semantic histories for long-horizon tasks (e.g., LagMemo[18], ATLAS Navigator[5]) versus end-to-end policies that directly map observations to actions without explicit memory modules. Physically Executable Gaussian[0] falls within the language-guided task execution cluster, emphasizing the generation of physically plausible action sequences grounded in Gaussian-based scene understanding. Compared to LagMemo[18], which prioritizes episodic memory for multi-step reasoning, and ATLAS Navigator[5], which focuses on hierarchical planning with topological maps, Physically Executable Gaussian[0] appears to stress the executability constraint—ensuring that predicted trajectories respect physical dynamics and scene affordances. This positions it as a bridge between high-level language grounding and low-level motion feasibility, a trade-off that remains an open question as the field scales to more complex real-world environments.

Related Works in Same Category

The following **3 sibling papers** share the same taxonomy leaf node with the original paper:

1. ATLAS Navigator: Active Task-driven LAnguage-embedded Gaussian Splatting

Authors: Tao, Yuezhan, Murali Varun, Spasojevic, Igor, et al. (7 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

We address the challenge of task-oriented navigation in unstructured and unknown environments, where robots must incrementally build and reason on rich, metric-semantic maps in real time. Since tasks may require clarification or re-specification, it is necessary for the information in the map to be rich enough to enable generalization across a wide range of tasks. To effectively execute tasks specified in natural language, we propose a hierarchical representation built on language-embedded Gauss...

Relationship Analysis

Both papers belong to the Language-Guided Task Execution category, focusing on frameworks that execute language-specified tasks through hierarchical planning and semantic reasoning in 3DGS environments. The original paper (SAGE-3D) overlaps with ATLAS Navigator in using 3D Gaussian Splatting for vision-language navigation with semantic grounding and hierarchical planning, but differs significantly in its approach: SAGE-3D emphasizes object-level manual annotation, physics-aware execution with collision bodies, and a benchmark for VLN evaluation, while ATLAS Navigator focuses on active task-driven navigation with compressed language features via PCA, submapping for large-scale operation, and real-time task re-specification without requiring manual object annotations.

---

2. LagMemo: Language 3D Gaussian Splatting Memory for Multi-modal Open-vocabulary Multi-goal Visual Navigation

Authors: Zhou Hao-tian, Wang Xiao-le, Li He, Sun Fusheng, Guo Shengyu, et al. (7 authors total) | **Year/Venue**: 2025 • arXiv (Cornell University) | **URL**: View paper

Abstract

Navigating to a designated goal using visual information is a fundamental capability for intelligent robots. Most classical visual navigation methods are restricted to single-goal, single-modality, and closed set goal settings. To address the practical demands of multi-modal, open-vocabulary goal queries and multi-goal visual navigation, we propose LagMemo, a navigation system that leverages a language 3D Gaussian Splatting memory. During exploration, LagMemo constructs a unified 3D language mem...

Relationship Analysis

Both papers belong to the Language-Guided Task Execution category, focusing on frameworks that execute language-specified tasks through hierarchical planning and semantic reasoning in 3DGS-based environments. They overlap in using 3D Gaussian Splatting for visual-language navigation with semantic grounding and multi-modal goal specification. However, the original paper (SAGE-3D) emphasizes physically executable 3DGS environments with collision bodies and physics simulation for embodied navigation training, while the candidate paper (LagMemo) focuses on language-embedded 3DGS memory for multi-goal localization and navigation through codebook-based feature retrieval without explicit physics simulation.

---

3. RoboTidy: A 3D Gaussian Splatting Household Tidying Benchmark for Embodied Navigation and Action

Authors: Xiaoquan Sun, Ruijian Zhang, Kang Pang, Bingchen Miao, Yuxiang Tan, et al. (8 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Household tidying is an important application area, yet current benchmarks neither model user preferences nor support mobility, and they generalize poorly, making it hard to comprehensively assess integrated language-to-action capabilities. To address this, we propose RoboTidy, a unified benchmark for language-guided household tidying that supports Vision-Language-Action (VLA) and Vision-Language-Navigation (VLN) training and evaluation. RoboTidy provides 500 photorealistic 3D Gaussian Splatting...

Relationship Analysis

Both papers belong to the Language-Guided Task Execution category, focusing on executing language-specified tasks through hierarchical planning and semantic reasoning in 3DGS environments. They overlap in using 3DGS representations for embodied navigation, semantic grounding of objects, and physics-aware execution with collision handling. The key difference is that the original paper (SAGE-3D) focuses on general VLN tasks with 2M trajectory-instruction pairs across diverse indoor scenes, while the candidate paper (RoboTidy) specializes in household tidying tasks with object-container sorting, providing 6.4k manipulation demonstrations and real-world sim2real transfer experiments for bimanual manipulation.

Contributions Analysis

Overall novelty summary. The paper introduces SAGE-3D, a paradigm that enhances 3D Gaussian Splatting with object-level semantic annotations and physics-aware collision interfaces for Visual-Language Navigation. It resides in the 'Language-Guided Task Execution' leaf alongside three sibling papers (LagMemo, ATLAS Navigator, and one other), forming a small cluster within the broader 'Vision-Language Navigation in Continuous Environments' branch. This leaf represents a focused research direction within a taxonomy of 32 papers across 12 leaf nodes, suggesting moderate but not overwhelming prior work in this specific intersection of semantic grounding and task execution.

The taxonomy tree reveals that SAGE-3D sits adjacent to 'Trajectory Planning and Viewpoint Synthesis' (2 papers) and 'Image-Goal and Instance-Level Navigation' (5 papers), both under the same parent branch. Neighboring branches include 'Semantic 3D Gaussian Splatting Representations' (10 papers across three leaves) and 'Sim-to-Real Transfer and Embodied AI Platforms' (4 papers). The scope notes clarify that SAGE-3D's emphasis on physical executability and object-centric grounding distinguishes it from purely semantic representation methods (excluded from this leaf) and from trajectory synthesis approaches that lack explicit task-level reasoning.

Among 24 candidates examined across three contributions, no clearly refutable prior work was identified. The SAGE-3D paradigm examined 4 candidates with 0 refutations; InteriorGS dataset examined 10 candidates with 0 refutations; SAGE-Bench benchmark examined 10 candidates with 0 refutations. This limited search scope—top-K semantic matches plus citation expansion—suggests that within the examined literature, the combination of object-level semantic grounding, physics-aware execution interfaces, and a dedicated VLN benchmark appears relatively unexplored. However, the analysis does not claim exhaustive coverage of all possible prior work.

Given the constrained search scope (24 candidates, not hundreds), the contributions appear to occupy a niche where semantic 3DGS, physical executability, and VLN benchmarking converge. The taxonomy structure indicates this is a moderately populated research area with clear boundaries separating representation methods from navigation policies. The absence of refutable candidates among examined papers suggests potential novelty, though a broader literature review would be needed to confirm whether similar integrations exist outside the top-K semantic neighborhood.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: SAGE-3D paradigm for semantically and physically aligned 3D Gaussian environments

Description: The authors introduce SAGE-3D, a paradigm that upgrades 3D Gaussian Splatting from a rendering-only representation into an executable environment foundation by adding object-level semantics and physics-aware execution capabilities for embodied navigation tasks.

This contribution was assessed against **4 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Feature splatting: Language-driven physics-based scene synthesis and editing

URL: View paper

Brief Assessment

Feature Splatting[44] focuses on language-driven physics simulation for scene editing using 3D Gaussians with semantic features, but does not address embodied navigation tasks or the specific paradigm of creating executable environments for VLN benchmarking that SAGE-3D proposes.

---

2. Three Dimensional Gaussian Splatting as a Foundation for Multitask Scene Modeling Spanning Segmentation Editing and Generation

URL: View paper

Brief Assessment

Multitask Scene Modeling[45] focuses on segmentation, editing, and generation tasks using 3DGS, not on creating executable environments for embodied navigation with physics-aware execution capabilities.

---

3. Enhancing 3D Gaussian splatting for low-quality images: semantically guided training and unsupervised quality assessment

URL: View paper

Brief Assessment

Semantically Guided Splatting[43] focuses on enhancing 3D Gaussian splatting for low-quality images through semantic guidance and unsupervised quality assessment, not on creating executable environments with physics simulation for embodied navigation tasks.

---

4. Scan, Materialize, Simulate: A Generalizable Framework for Physically Grounded Robot Planning

URL: View paper

Brief Assessment

Scan Materialize Simulate[46] focuses on physics-informed robot planning using 3DGS for scene reconstruction and physics simulation, not on creating executable environments for embodied navigation with semantic grounding and VLN benchmarks.

Contribution 2: InteriorGS dataset with object-level annotated 3DGS scenes

Description: The authors release InteriorGS, a dataset containing 1,000 manually annotated 3D Gaussian Splatting indoor scenes with over 554,000 object instances across 755 categories, providing fine-grained object-level semantics including instance IDs, categories, and bounding boxes.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Open-vocabulary functional 3d scene graphs for real-world indoor spaces

URL: View paper

Brief Assessment

Functional Scene Graphs[34] focuses on functional 3D scene graphs with interactive elements and their relationships, not on large-scale 3DGS datasets with dense object-level annotations. Their dataset (fungraph3d) contains only 14 scenes with 146 objects, compared to InteriorGS's 1,000 scenes with 554,000 object instances.

---

2. Language-grounded indoor 3d semantic segmentation in the wild

URL: View paper

Brief Assessment

Language-grounded Indoor Segmentation[38] focuses on 3D semantic segmentation with 200 class categories on ScanNet data, not on creating object-level annotated 3DGS scenes. The candidate addresses semantic segmentation benchmarks rather than 3DGS scene reconstruction with instance-level annotations.

---

3. ToF-360 - A Panoramic Time-of-Flight RGB-D Dataset for Single Capture Indoor Semantic 3D Reconstruction

URL: View paper

Brief Assessment

ToF-360[36] focuses on panoramic Time-of-Flight RGB-D scanning with 2D/3D semantic segmentation tasks, not 3D Gaussian Splatting scenes with object-level annotations for embodied navigation.

---

4. Mobile Robot Oriented Large-Scale Indoor Dataset for Dynamic Scene Understanding

URL: View paper

Brief Assessment

Mobile Robot Dataset[40] focuses on dynamic scene understanding with 2D/3D bounding boxes for mobile robots, not 3D Gaussian Splatting reconstructions with object-level semantics for embodied navigation.

---

5. Learning 3d semantic scene graphs from 3d indoor reconstructions

URL: View paper

Brief Assessment

Semantic Scene Graphs[39] focuses on semantic scene graphs from 3D reconstructions with relationship annotations, not on 3D Gaussian Splatting scenes. The candidate uses traditional mesh-based reconstructions from 3RScan, while the original paper specifically addresses 3DGS-based representations with different annotation goals (object instances for VLN tasks vs. semantic relationships for scene understanding).

---

6. IRef-VLA: A Benchmark for Interactive Referential Grounding with Imperfect Language in 3D Scenes

URL: View paper

Brief Assessment

IRef-VLA[37] focuses on scanned 3D rooms with semantic relations and referential statements for language-guided navigation, not on 3D Gaussian Splatting scenes with object-level annotations.

---

7. HOI4D: A 4D Egocentric Dataset for Category-Level Human-Object Interaction

URL: View paper

Brief Assessment

HOI4D[41] focuses on egocentric human-object interaction with 4D temporal sequences and hand pose tracking, not on static 3D Gaussian Splatting scene reconstructions with object-level semantics for navigation tasks.

---

8. CACE: Sim-to-Real Indoor 3D Semantic Segmentation via Context-Aware Augmentation and Consistency Enforcement

URL: View paper

Brief Assessment

CACE[35] focuses on sim-to-real semantic segmentation using ScanNet, which contains 1,613 3D scans with 18 categories. This is fundamentally different from InteriorGS's 1,000 3DGS scenes with 554k object instances across 755 categories, and CACE[35] does not provide object-level 3DGS annotations or instance-level semantics.

---

9. Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation

URL: View paper

Brief Assessment

Omniobject3d[33] focuses on scanned 3D objects (6,000 individual objects across 190 categories) rather than complete indoor scenes with spatial relationships. The candidate does not provide scene-level annotations or indoor environment reconstructions comparable to InteriorGS's 1,000 annotated 3DGS indoor scenes.

---

10. 3D-MoRe: Unified Modal-Contextual Reasoning for Embodied Question Answering

URL: View paper

Brief Assessment

3D-MoRe[42] focuses on generating QA pairs and object descriptions from existing ScanNet scenes, not on creating object-level annotated 3DGS datasets. The candidate uses ScanNet data with text annotations, while the original contribution is about manually annotating 3DGS indoor scenes with instance IDs, categories, and bounding boxes.

Contribution 3: SAGE-Bench VLN benchmark with hierarchical instructions and continuity metrics

Description: The authors introduce SAGE-Bench, the first 3DGS-based Vision-Language Navigation benchmark featuring 2 million trajectory-instruction pairs, hierarchical instruction generation combining high-level semantic goals with low-level actions, and three novel navigation natural continuity metrics (Continuous Success Ratio, Integrated Collision Penalty, and Path Smoothness).

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Structured Preference Optimization for Vision-Language Long-Horizon Task Planning

URL: View paper

Brief Assessment

Structured Preference Optimization[49] focuses on long-horizon task planning with preference-based optimization in VirtualHome and Habitat 2.0, not on 3DGS-based VLN benchmarks with hierarchical instructions and continuity metrics.

---

2. Instruction-aligned hierarchical waypoint planner for vision-and-language navigation in continuous environments

URL: View paper

Brief Assessment

Hierarchical Waypoint Planner[51] focuses on waypoint planning methods for VLN in continuous environments, not on benchmark construction with hierarchical instructions and continuity metrics. The candidate paper's full text is not available for comparison.

---

3. SENTINEL: A Multi-Level Formal Framework for Safety Evaluation of LLM-based Embodied Agents

URL: View paper

Brief Assessment

SENTINEL[54] focuses on formal safety evaluation of LLM-based embodied agents using temporal logic across semantic, plan, and trajectory levels. It does not address vision-language navigation benchmarks, hierarchical instruction generation, or continuity metrics for navigation tasks.

---

4. Integrating Chain-of-Thought for Multimodal Alignment: A Study on 3D Vision-Language Learning

URL: View paper

Brief Assessment

Chain-of-Thought 3D Learning[50] focuses on integrating chain-of-thought reasoning into 3D vision-language alignment with a benchmark for shape recognition and functional inference, not vision-language navigation with hierarchical instructions and continuity metrics for embodied agents.

---

5. MobileVLA-R1: Reinforcing Vision-Language-Action for Mobile Robots

URL: View paper

Brief Assessment

MobileVLA-R1[53] focuses on quadruped robot control with chain-of-thought reasoning for VLA tasks, not on creating VLN benchmarks with hierarchical instructions or continuity metrics for 3DGS-based navigation environments.

---

6. Think Hierarchically, Act Dynamically: Hierarchical Multi-modal Fusion and Reasoning for Vision-and-Language Navigation

URL: View paper

Brief Assessment

Hierarchical Multi-modal Fusion[48] focuses on multi-level feature fusion and reasoning architectures for VLN, not on benchmark construction with hierarchical instructions or continuity metrics. The candidate does not present a benchmark dataset or propose navigation continuity metrics.

---

7. SEER-VAR: Semantic Egocentric Environment Reasoner for Vehicle Augmented Reality

URL: View paper

Brief Assessment

SEER-VAR[55] focuses on vehicle-based augmented reality with egocentric SLAM and AR overlay generation for driving scenarios, not vision-language navigation benchmarks with hierarchical instructions or continuity metrics for embodied agents.

---

8. Hierarchical semantic-augmented navigation: Optimal transport and graph-driven reasoning for vision-language navigation

URL: View paper

Brief Assessment

Semantic-augmented Navigation[47] focuses on hierarchical semantic scene graphs and optimal transport-based planning for VLN-CE, not on creating a 3DGS-based benchmark with continuity metrics. The candidate's hierarchical instructions serve a different purpose (semantic reasoning) than SAGE-Bench's hierarchical scheme (combining high-level semantic goals with low-level actions and introducing continuity metrics).

---

9. MLANet: Multi-Level Attention Network with Sub-instruction for Continuous Vision-and-Language Navigation

URL: View paper

Brief Assessment

MLANet[56] focuses on sub-instruction generation and multi-level attention mechanisms for continuous VLN, but does not present a benchmark with hierarchical instructions or continuity metrics like SAGE-Bench's Continuous Success Ratio, Integrated Collision Penalty, and Path Smoothness.

---

10. VisuCraft: Enhancing Large Vision-Language Models for Complex Visual-Guided Creative Content Generation via Structured Information Extraction

URL: View paper

Brief Assessment

VisuCraft[52] focuses on enhancing vision-language models for creative content generation (story generation, poetry composition) with structured information extraction, not on vision-language navigation benchmarks or embodied agent navigation tasks.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] Towards Physically Executable 3D Gaussian for Embodied Navigation View paper
- [1] Splat-nav: Safe real-time robot navigation in gaussian splatting maps View paper
- [2] 3d gaussian map with open-set semantic grouping for vision-language navigation View paper
- [3] End-to-end autonomous driving: Challenges and frontiers View paper
- [4] Fmgs: Foundation model embedded 3d gaussian splatting for holistic 3d scene understanding View paper
- [5] ATLAS Navigator: Active Task-driven LAnguage-embedded Gaussian Splatting View paper
- [6] Gaussnav: Gaussian splatting for visual navigation View paper
- [7] Unitedvln: Generalizable gaussian splatting for continuous vision-language navigation View paper
- [8] YOPO-Nav: Visual Navigation using 3DGS Graphs from One-Pass Videos View paper
- [9] 4d langsplat: 4d language gaussian splatting via multimodal large language models View paper
- [10] VISTA: Open-Vocabulary, Task-Relevant Robot Exploration with Online Semantic Gaussian Splatting View paper
- [11] IGL-Nav: Incremental 3D Gaussian Localization for Image-goal Navigation View paper
- [12] GSplatVNM: Point-of-View Synthesis for Visual Navigation Models Using Gaussian Splatting View paper
- [13] RealMirror: A Comprehensive, Open-Source Vision-Language-Action Platform for Embodied AI View paper
- [14] A Neural Representation Framework with LLM-Driven Spatial Reasoning for Open-Vocabulary 3D Visual Grounding View paper
- [15] EmbodiedSplat: Personalized Real-to-Sim-to-Real Navigation with Gaussian Splats from a Mobile Device View paper
- [16] VR-Robo: A Real-to-Sim-to-Real Framework for Visual Robot Navigation and Locomotion View paper
- [17] GRaD-Nav++: Vision-Language Model Enabled Visual Drone Navigation with Gaussian Radiance Fields and Differentiable Dynamics View paper
- [18] LagMemo: Language 3D Gaussian Splatting Memory for Multi-modal Open-vocabulary Multi-goal Visual Navigation View paper
- [19] OpenFly: A Comprehensive Platform for Aerial Vision-Language Navigation View paper
- [20] RoboTidy: A 3D Gaussian Splatting Household Tidying Benchmark for Embodied Navigation and Action View paper
- [21] 3D Vision-Language Gaussian Splatting View paper
- [22] FMLGS: Fast Multilevel Language Embedded Gaussians for Part-level Interactive Agents View paper
- [23] ReasonGrounder: LVLM-Guided Hierarchical Feature Splatting for Open-Vocabulary 3D Visual Grounding and Reasoning View paper
- [24] SplatSearch: Instance Image Goal Navigation for Mobile Robots using 3D Gaussian Splatting and Diffusion Models View paper
- [25] BEINGS: Bayesian Embodied Image-goal Navigation with Gaussian Splatting View paper
- [26] OVSG-SLAM: Open-Vocabulary Semantic Gaussian Splatting SLAM View paper
- [27] Go-SLAM: Grounded Object Segmentation and Localization with Gaussian Splatting SLAM View paper
- [28] Towards Integrated Multimodal Interaction: Merging Immersive 3D Worlds with Language Based Retrieval for 3D Scene Understanding View paper
- [29] Embodied Navigation in Unknown Environments with Implicit Scene Memory and Target-aware Memory Retrieval View paper
- [30] LiLMaps: Learnable Implicit Language Maps View paper
- [31] CoLaSplat: Compact Language 3D Gaussian Splatting View paper
- [32] Visually-grounded Humanoid Agents View paper
- [33] Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation View paper
- [34] Open-vocabulary functional 3d scene graphs for real-world indoor spaces View paper
- [35] CACE: Sim-to-Real Indoor 3D Semantic Segmentation via Context-Aware Augmentation and Consistency Enforcement View paper
- [36] ToF-360 - A Panoramic Time-of-Flight RGB-D Dataset for Single Capture Indoor Semantic 3D Reconstruction View paper
- [37] IRef-VLA: A Benchmark for Interactive Referential Grounding with Imperfect Language in 3D Scenes View paper
- [38] Language-grounded indoor 3d semantic segmentation in the wild View paper
- [39] Learning 3d semantic scene graphs from 3d indoor reconstructions View paper
- [40] Mobile Robot Oriented Large-Scale Indoor Dataset for Dynamic Scene Understanding View paper
- [41] HOI4D: A 4D Egocentric Dataset for Category-Level Human-Object Interaction View paper
- [42] 3D-MoRe: Unified Modal-Contextual Reasoning for Embodied Question Answering View paper
- [43] Enhancing 3D Gaussian splatting for low-quality images: semantically guided training and unsupervised quality assessment View paper
- [44] Feature splatting: Language-driven physics-based scene synthesis and editing View paper
- [45] Three Dimensional Gaussian Splatting as a Foundation for Multitask Scene Modeling Spanning Segmentation Editing and Generation View paper
- [46] Scan, Materialize, Simulate: A Generalizable Framework for Physically Grounded Robot Planning View paper
- [47] Hierarchical semantic-augmented navigation: Optimal transport and graph-driven reasoning for vision-language navigation View paper
- [48] Think Hierarchically, Act Dynamically: Hierarchical Multi-modal Fusion and Reasoning for Vision-and-Language Navigation View paper
- [49] Structured Preference Optimization for Vision-Language Long-Horizon Task Planning View paper
- [50] Integrating Chain-of-Thought for Multimodal Alignment: A Study on 3D Vision-Language Learning View paper
- [51] Instruction-aligned hierarchical waypoint planner for vision-and-language navigation in continuous environments View paper
- [52] VisuCraft: Enhancing Large Vision-Language Models for Complex Visual-Guided Creative Content Generation via Structured Information Extraction View paper
- [53] MobileVLA-R1: Reinforcing Vision-Language-Action for Mobile Robots View paper
- [54] SENTINEL: A Multi-Level Formal Framework for Safety Evaluation of LLM-based Embodied Agents View paper
- [55] SEER-VAR: Semantic Egocentric Environment Reasoner for Vehicle Augmented Reality View paper
- [56] MLANet: Multi-Level Attention Network with Sub-instruction for Continuous Vision-and-Language Navigation View paper