

DeWorldSG: Depth-Aware 3D Semantic Scene Graph Generation via World-Model Priors

Seok-Young Kim¹, Abdelrahman Elskhawy^{2,3}, Taewook Ha¹, Dooyoung Kim⁴, Eunjae Shin¹, Benjamin Busam^{2,3†}, and Woontack Woo^{1†}

¹ Korea Advanced Institute of Science and Technology

² Technical University of Munich

³ Munich Center for Machine Learning (MCML)

⁴ La Trobe University

seokyoung@kaist.ac.kr

<https://deworldsg2026.github.io>

Abstract. We present *DeWorldSG*, a novel framework that generates spatio-temporally robust 3D Semantic Scene Graphs from RGB-D sequences. Existing methods often struggle to construct reliable 3D scene graphs due to unstable 3D object representations and missing relations caused by frame-wise inference. DeWorldSG addresses these issues by estimating instance-level geometric 3D Gaussian distributions through depth-guided filtering and representing each object as a probabilistic 3D node rather than a single projected point. To mitigate relational sparsity from frame-wise inference, our framework further aggregates spatiotemporal evidence across object pairs and refines relations using contextual priors derived from a world model (V-JEPA 2). Experiments on the 3DSSG and ReplicaSSG datasets demonstrate state-of-the-art (SoTA) performance in both object and predicate prediction, while producing temporally consistent scene structures. In particular, our method improves triplet recall by 77.4% and predicate recall by 23.2% over prior SoTA approaches, making it suitable for robotic manipulation and AR applications. Our code and models are open-sourced.

Keywords: 3D Semantic Scene Graph · World Models · 3D Scene Understanding

1 Introduction

For robotic intelligence and Augmented Reality (AR) systems to operate reliably in real-world environments, it is necessary to go beyond recognizing individual objects and instead reason about the complex spatial relationships among them. To fundamentally perceive 3D space and perform high-level tasks, a compact structural representation that captures scene entities and their mutual relations is essential [16, 34]. 3D Semantic Scene Graph (SSG) generation [42] addresses

[†] Co-corresponding authors.

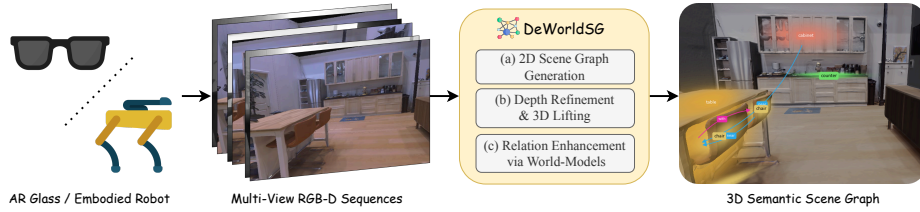


Fig. 1: DeWorldSG generates a spatio-temporally robust 3D semantic scene graph from RGB-D observations. It combines 2D scene graph generation, depth-aware 3D lifting, and world-model-based relation enhancement to produce stable object nodes and context-aware relational edges for AR and embodied robotic systems.

this need by transforming complex 3D environments into a compact graph structure consisting of object-level nodes and relational edges. Such a structural representation preserves both geometric information and semantic context of the scene, thereby providing a fundamental basis for higher-level cognitive tasks such as agent planning [4, 9, 39], 3D content synthesis [15, 19, 49–51], and object augmentation [20, 25, 26] in AR. However, current 3D SSG generation approaches still suffer from unstable 3D geometry and inconsistent relational inference across video frames, often resulting in missing or unreliable relations in the final graph.

Recently, with the goal of enabling these interactive applications, increasing attention has been given to approaches [10, 13, 21] that first extract 2D scene graphs from RGB(-D) sequences and then progressively lift and merge them into a global 3D scene graph. This strategy reduces dependency on full 3D reconstruction, maintains robustness to geometric noise, and lessens sensitivity to camera pose errors, making it suitable for stable operation in real-world settings. Despite these advantages, unstable depth estimates near object boundaries can distort lifted object geometry and destabilize object-level 3D representations, leading to erroneous merging of distinct objects or fragmentation of a single one [13]. Moreover, because existing approaches [10, 13, 21, 44–46] infer relationships independently for each frame, limited observations caused by occlusion or low-resolution input often result in missing essential edges. As a result, the generated scene graphs tend to reflect local geometric proximity rather than coherent scene-level context.

The root cause of these issues is treating scene graph generation as a sequence of independent frame-level predictions. In real-world environments, object geometry must be modeled with uncertainty during 2D-to-3D lifting, and semantic relationships should be inferred by accumulating evidence over time rather than from single observations. This calls for an incremental 3D SSG framework that maintains stable object representations while refining inter-object relations using spatiotemporal context.

In this work, we propose *DeWorldSG*, a novel framework for context-grounded 3D SSG generation that combines depth-aware probabilistic object modeling with spatiotemporal relation reasoning (Fig. 1). Our approach constructs per-

frame 2D scene graphs from RGB-D sequences and incrementally builds a global 3D graph while explicitly accounting for uncertainty in lifted object geometry. To enhance relational completeness and consistency, it aggregates relation evidence across temporally dispersed observations and incorporates video-based priors from a world model [1] when local visual evidence is insufficient. Finally, confidence-based probabilistic fusion balances the influence of these priors against direct visual and geometric cues, suppressing hallucinated relations and yielding a more globally consistent 3D scene graph.

We evaluate DeWorldSG on the 3DSSG [42] and ReplicaSSG [13] datasets, which provide object class and relation annotations for indoor environments. Our method consistently outperforms existing baselines in both object classification and relationship prediction, achieving state-of-the-art performance across all metrics. Ablation studies further validate each component’s contribution and demonstrate robust graph generation even without ground-truth camera poses, while sustaining incremental graph updates at interactive rates. These results indicate that DeWorldSG provides a reliable structural representation with high-fidelity relational reasoning, which is essential for downstream spatial reasoning applications such as robotic manipulation [28,47] and context-aware AR anchoring [19].

The main contributions of this work are summarized as follows:

1. We present *DeWorldSG*, a novel framework for depth-aware 3D semantic scene graph generation from RGB-D sequences that represents each object as an instance-level probabilistic 3D Gaussian distribution derived from mask-guided depth observations.
2. We propose Dual-Domain Depth Refinement (DR), a sequential filtering mechanism that effectively removes spatial and distributional depth noise, demonstrating robust instance-level 3D Gaussian estimation and stabilizing global 3D scene graph construction.
3. We introduce a spatiotemporal relation reasoning scheme that aggregates relational evidence across frames and leverages world model priors to contextualize local observations, enhancing relational completeness and consistency.

2 Related Work

2.1 Geometry-Explicit 3D SSG Generation

3D Semantic Scene Graphs [31,42,43] represent a 3D environment as a structured graph, where nodes correspond to object instances and edges encode semantic and spatial relationships, providing a compact yet expressive abstraction for embodied AI [36,37] and robotics reasoning [3,12,35]. Early 3D SSG generation methods [32,42,44,45] follow a geometry-first paradigm, explicitly leveraging point-clouds or SLAM-estimated 3D structure. In [42], a learned pipeline regresses a semantic scene graph directly from reconstructed indoor point-clouds, extracting features for object instances and pairwise union regions and refining

them with a GCN [22] to jointly predict objects and relations. Similarly, [44, 45] leverage Visual-SLAM [5, 40] to estimate point-clouds and 3D bounding boxes from RGB(-D) sequences, and then construct scene graphs from the reconstructed 3D structure using GNN-based reasoning [11]. The main advantage of this line of work is that relations are grounded in explicit 3D geometry from the outset, which historically made it effective for spatial reasoning. However, graph quality is tightly coupled to reconstruction quality, making these methods sensitive to pose drift, incomplete geometry, and computational overhead. DeWorldSG mitigates this bottleneck by estimating instance-level probabilistic geometry directly from RGB-D observations, enabling stable object nodes without requiring dense reconstruction.

2.2 2D-to-3D Lifted 3D SSG Generation

Lifting-based 3D SSG methods first predict per-frame 2D entities and relations from RGB(-D) observations and then align them in a shared 3D coordinate frame. [10] generates 3D scene graphs from multi-view image sequences by first estimating object 3D locations and occupancies from matched 2D detections via multi-view geometry, followed by relation prediction. Also, [21] enables incremental construction and updating of a global 3D scene graph from sequential observations for intelligent-agent reasoning. More recently, FROSS [13] directly lifts 2D scene graphs to 3D and merges objects as 3D Gaussian distributions [7], facilitating online 3D SSG generation. Unlike earlier lifting pipelines that rely primarily on geometric alignment, FROSS introduces probabilistic Gaussian representations for object merging, improving robustness during incremental graph construction. However, because the 2D-to-3D lifting pipeline of FROSS does not explicitly model object geometry or depth distributions, the generated scene graphs often contain spatially inconsistent or missing relations, which limits fine-grained spatial reasoning. In contrast, DeWorldSG explicitly estimates mask-guided instance-level depth distributions from filtered RGB-D evidence, yielding geometrically grounded 3D Gaussians that support more reliable graph reasoning than weakly constrained lifting.

2.3 World Models for 3D Scene Understanding

A world model is a predictive model that (i) infers and completes a latent state of the environment from partial observations [1, 2, 6] and (ii) forecasts the evolution of this state to support internal rollouts for planning [24]. In this work, we focus on latent-state inference and completion, exploiting it as a contextual prior for 3D scene understanding. Recent works exploit this ability for 3D scene understanding by recovering occluded structure with explicit 3D Gaussian world models [14], imagining plausible unseen views with controllable video world models [18, 48], or learning geometry-aware cross-modal latent representations [52].

While these approaches primarily target geometric reconstruction, view synthesis, or representation learning, we instead incorporate world-model priors into a scene graph structure for structured 3D spatial reasoning. In scene graph

generation, relations between instances are often under-constrained by instantaneous visual evidence due to occlusion, noisy geometry, and viewpoint-dependent ambiguities. A world model learns predictive representations that complete unobserved aspects of the current state and encode regularities and commonsense constraints over object interactions beyond per-frame appearance [24]. Building on this insight, DeWorldSG leverages the contextual understanding of a pre-trained video world model to infer relational priors across clips and fuses these priors with geometric evidence to construct a spatio-temporally consistent 3D semantic scene graph.

3 Preliminary

3D Semantic Scene Graph. A 3D Semantic Scene Graph (SSG) abstracts a scene into a directed graph $G = (V, E)$ composed of object nodes and inter-object relation edges. Each node $v_i \in V$ represents the i -th object instance and is defined by its 3D geometric parameters (position, scale, and spatial uncertainty) and a semantic category label c_i . Each directed edge $e_{i \rightarrow j} \in E$ encodes a spatial relation between a subject v_i and an object v_j , where the predicate type $r_{i \rightarrow j}$ is selected from a predefined predicate set R (e.g., *attached to*, *under*). Depending on downstream scene-graph applications, nodes can be extended with additional attributes such as visual descriptors, text embeddings, or temporal information. In this paper, we focus on the core components required for semantic graph reasoning: $(c_i, \mu_i^{3D}, \Sigma_i^{3D})$ and $r_{i \rightarrow j}$. Accordingly, our goal is to construct a 3D SSG G from an RGB-D sequence $\{(I_t, D_t)\}_{t=1}^T$, given camera intrinsics K , and per-frame camera poses $T_t = [R_t \mid t_t]$.

Object Representation with 3D Gaussians. In 3D SSG, each object node represents an instance in the scene, which can be defined using various geometric representations such as a 3D bounding box, a point-cloud or mesh, or a probability distribution. Inspired by [13], we model the observation of object instance i at frame t as a 3D Gaussian $\mathcal{N}(\mu_{i,t}^{3D}, \Sigma_{i,t}^{3D})$. Compared with point-cloud or mesh-based representations, this probabilistic form summarizes an object’s position and spatial extent using low-dimensional statistics, resulting in improved computational efficiency. In contrast to rigid 3D bounding-box representations, it provides a more faithful approximation of the underlying object’s geometric structure. Furthermore, this representation enables concise integration of multi-view observations through iterative updates of the mean and covariance, while facilitating stable object association during global graph merging via distribution-based distance metrics. Unlike prior 2D-to-3D lifting approaches that approximate object geometry from projected 2D bounding boxes, we directly estimate instance-level 3D Gaussian distributions from depth observations. This probabilistic formulation provides a robust spatial foundation for downstream relation reasoning and global graph construction.

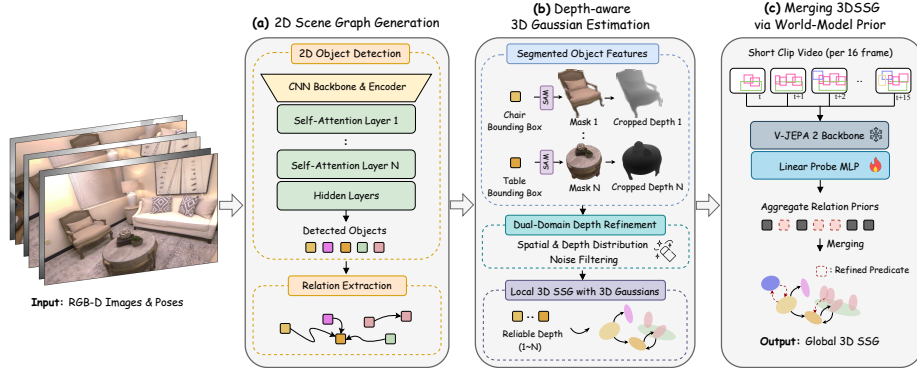


Fig. 2: Overview of DeWorldSG. Given RGB-D image sequences, DeWorldSG first constructs frame-wise 2D scene graphs by detecting objects and extracting their relationships (a). The detected instances are lifted to depth-aware probabilistic 3D Gaussians through mask-guided depth filtering and dual-domain depth refinement (b). Finally, spatiotemporal predicate priors from a video world model are aggregated and fused with visual predictions to incrementally merge local graphs into a globally consistent 3D semantic scene graph (c).

4 Method

Overview. We propose *DeWorldSG*, a framework for generating spatially coherent 3D semantic scene graphs from RGB-D sequences. As illustrated in Fig. 2, the core of DeWorldSG is to represent each object as a probabilistic 3D Gaussian estimated from filtered instance-level depth observations. For each frame, we first infer 2D object instances and pairwise relations from the input RGB-D sequence (Sec. 4.1). We then refine the depth of each instance using SAM-guided masks and dual-domain filtering, and estimate a depth-aware 3D Gaussian that serves as the local object node. These local graphs are incrementally merged into a global 3D scene graph based on semantic consistency and Gaussian similarity (Sec. 4.2). Finally, ambiguous relational edges are refined by fusing temporally aggregated visual evidence with priors from a video world model through confidence-aware probabilistic fusion (Sec. 4.3). This process yields stable, globally consistent 3D Semantic Scene Graphs that capture both geometric accuracy and context-aware relational structure.

4.1 Depth-aware Local 3D Gaussian Representation

Relying on a single representative depth value for 2D-to-3D lifting inherently limits the stability of the approximated 3D Gaussian distribution. In scenes with significant object overlap, boundary depth distortions and flying pixels introduce high variance into μ_{3D} and Σ_{3D} , often leading to fragmented or incorrectly merged instances in the global graph. To overcome this limitation, *DeWorldSG* reformulates object lifting as direct estimation of a 3D Gaussian distribution

from instance-level depth observations, instead of inferring geometry from lifting 2D bounding boxes.

2D Scene Graph Generation. For each frame t of the input RGB-D sequence, we extract a 2D scene graph G_t^{2D} from the RGB image I_t . Following the strategy of [13], we use an object detector [27, 53] to obtain a per-frame instance set $\{o_{i,t}\}$, and predict a bounding box $b_{i,t} = (c_x, c_y, w, h)$ and class probability $p(c_{i,t})$ for each object. A 2D relation inference model [17] is then applied to each ordered object pair (i, j) within the frame, producing relation logits $p(r_{i \rightarrow j, t})$. As a result, the 2D scene graph at frame t consists of nodes $\{(b_{i,t}, p(c_{i,t}))\}$ and edges $\{p(r_{i \rightarrow j, t})\}$, which are subsequently reprojected into 3D and serve as input for global graph construction.

Box-Conditioned Instance Extraction. Given the estimated bounding box $b_{i,t}$, we generate an object-wise binary mask $M_{i,t}$ via semantic segmentation [33]. If masks from different classes assign the same pixel within a frame, we identify it as an ambiguous pixel and discard it from all masks to prevent depth mixing across objects. Moreover, if the number of valid pixels in a mask is too small (< 10 pixels), covariance estimation becomes numerically unstable. In this case, we deem the instance as an invalid observation and exclude it from subsequent steps. This mask-based conditioning exploits object shape to produce cleaner depth samples for subsequent Gaussian estimation.

Dual-Domain Depth Refinement. Despite the use of instance masks $M_{i,t}$, the corresponding depth map D_t often contains outliers from multi-surface observations or overlapping objects, which bias the estimated 3D Gaussians. We address this with Dual-Domain Refinement (DR), a sequential filtering algorithm operating in both *spatial* and *depth-distribution* domains. In the *spatial domain*, we enforce local depth consistency to preserve continuous geometric structures. For each pixel $p \in M_{i,t}$, we compare its depth $D_t(p)$ to the median depth of its 3×3 neighborhood $\Omega(p)$. Pixels that deviate significantly from the local median are filtered out as follows:

$$M_{\text{spatial}} = \left\{ p \in M_{i,t} \mid \left| D_t(p) - \tilde{D}_t(p) \right| < \tau \tilde{D}_t(p) + \epsilon \right\}, \quad (1)$$

where $\tilde{D}_t(p) = \text{median}_{q \in \Omega(p)} D_t(q)$, and τ, ϵ are tolerance hyperparameters. This step suppresses discontinuous boundary noise (e.g., flying pixels) and yields a more coherent depth distribution around the object.

In the *depth-distribution domain*, we further refine the spatial inliers via 1D clustering along the depth axis. Let $Z_{\text{spatial}} = \{D_t(p) \mid p \in M_{\text{spatial}}\}$ be the set of refined depth values, and define the reference depth as $d_{\text{ref}} = \text{median}(Z_{\text{spatial}})$. To adapt to varying depth scales, the clustering threshold is dynamically determined as:

$$\Delta_{\text{thr}} = \max(\gamma_{\text{base}}, \alpha \cdot \text{median}(\Delta z), \beta \cdot d_{\text{ref}}), \quad (2)$$

where Δz denotes adjacent depth differences in the sorted depth set, γ_{base} enforces a minimum separation, α scales the typical local depth variation, and β adjusts the threshold according to the global depth scale. We then partition the sorted depth sequence using Δ_{thr} , and retain the cluster whose mean depth is closest to d_{ref} as the final inlier set P_{final} . DR therefore isolates a dominant, self-consistent depth layer per instance, which is crucial for stable Gaussian parameter estimation and erroneous global merging prevention.

Depth-guided 3D Gaussian Estimation. Given the refined inlier pixel set $P_{\text{final}} \subseteq M_{i,t}$, each pixel $(u, v) \in P_{\text{final}}$ is back-projected into 3D camera coordinates using depth $z = D_t(u, v)$, and then transformed into the world frame to form an instance-specific 3D point set $P_{i,t}$. With camera intrinsics K , each pixel is mapped to a camera-coordinate point $\tilde{\mathbf{x}}$, and using the frame pose $T_t = [R_t \mid t_t]$ yields the world-coordinate point $\mathbf{x}_n$:

$$\tilde{\mathbf{x}} = z \mathbf{K}^{-1} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix}, \quad \mathbf{x}_n = \mathbf{R}_t \tilde{\mathbf{x}} + \mathbf{t}_t, \quad \mathbf{x}_n \in \mathbb{R}^3. \quad (3)$$

$P_{i,t}$ thus aggregates multiple refined depth observations, through mask-based instance extraction and DR, for instance i in frame t . In contrast to representative-depth lifting, which relies on a single depth value per object, we explicitly model geometric uncertainty and surface spread by estimating a 3D Gaussian from $P_{i,t}$ via the sample mean and covariance:

$$\mu_{i,t}^{3D} = \frac{1}{|P_{i,t}|} \sum_{\mathbf{x}_n \in P_{i,t}} \mathbf{x}_n, \quad \Sigma_{i,t}^{3D} = \frac{1}{|P_{i,t}| - 1} \sum_{\mathbf{x}_n \in P_{i,t}} (\mathbf{x}_n - \mu_{i,t}^{3D})(\mathbf{x}_n - \mu_{i,t}^{3D})^\top + \epsilon \mathbf{I}. \quad (4)$$

where $\epsilon \mathbf{I}$ is a regularization term for numerical stability. Each object node is therefore represented as a probabilistic 3D Gaussian derived from instance-level depth-refined observations. Together with frame-wise relation predictions, these nodes form the local 3D scene graph $G_t^{3D} = (V_t, E_t)$ at frame t .

4.2 Incremental Merging into the Global 3D SSG

DeWorldSG incrementally integrates each local 3D graph $G_t^{3D} = (V_t, E_t)$ into the global graph G^{3D} . A key challenge is establishing cross-frame object identity between 3D Gaussian nodes and merging those that correspond to the same instance. For each local node $v_i^t \in V_t$, we therefore search for a correspondence in the global node set V^{3D} . To compare geometric similarity, we compute the Hellinger distance [29] between two Gaussians $\mathcal{N}(\mu_i, \Sigma_i)$ and $\mathcal{N}(\mu_j, \Sigma_j)$. If this distance is below a threshold δ_g , the nodes are considered to have compatible spatial location and dispersion structure. In parallel, for class probability distributions $p_i, p_j \in \mathbb{R}^{|C|}$, we define an inner-product-based class distance:

$$D_c(i, j) = 1 - \sum_{k=1}^{|C|} p_i^{(k)} p_j^{(k)}. \quad (5)$$

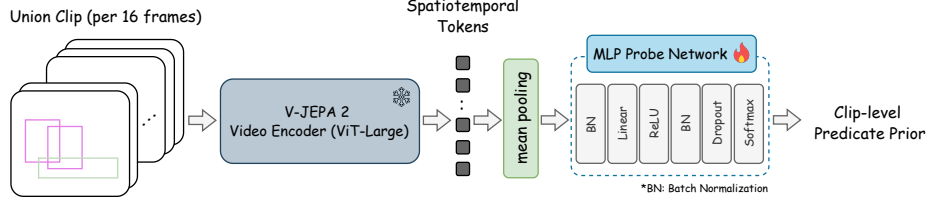


Fig. 3: For each object pair, we extract a 16-frame union-crop clip and encode it with a frozen V-JEPA 2 backbone. The resulting spatiotemporal tokens are mean-pooled and passed to an MLP probe to produce a clip-level predicate distribution.

This distance is small when two nodes assign high probability mass to the same class. We merge nodes only when both geometric and semantic criteria are satisfied:

$$HD(i, j) < \delta_g \quad \text{and} \quad D_c(i, j) < \delta_c. \quad (6)$$

When these conditions hold, we fuse the two Gaussians using weighted averaging. Let w_i and w_j denote their respective observation frequencies. The merged node v_k is updated as:

$$\mu_k = \frac{w_i \mu_i + w_j \mu_j}{w_i + w_j}, \quad \Sigma_k = \frac{w_i \Sigma_i + w_j \Sigma_j}{w_i + w_j} + \frac{w_i w_j}{(w_i + w_j)^2} (\mu_i - \mu_j)(\mu_i - \mu_j)^\top. \quad (7)$$

where μ_i and μ_j are the mean vectors, Σ_i and Σ_j are the associated covariance matrices of the Gaussians being merged.

For relation edges, we reassign all edges connected to the merged nodes and aggregate the relation prediction distributions for each object pair. The final relation is obtained by accumulating prediction probabilities across multiple frames and selecting the most probable relation. This temporal aggregation mitigates noisy single frame predictions and yields a more stable and temporally consistent 3D scene graph.

4.3 3D Semantic Scene Graph Meets World Models

DeWorldSG is motivated by the premise that scene-graph relations should be refined through aggregated visual evidence, rather than inferred from a single-frame prediction. To this end, we integrate geometric constraints from the global 3D scene graph with temporal visual context encoded by a world model. In particular, V-JEPA 2 [1] captures structural regularities and physical consistency beyond frame-level feature extraction. We leverage it as a *visual evidence collector* to refine relational edges in the global graph.

Geometric Neighbor Pairing with Temporal Context. To reduce hallucinated relations, we do not treat all mask pairs within a frame as candidate edges. Instead, we restrict inference to geometrically proximal object pairs. Following [8], we adopt a geometric proximity cost that combines the normalized

2D center distance $d_{2D}(i, j)$ and the depth disparity:

$$\text{GeoCost}(i, j) = 0.5 \cdot \frac{d_{2D}(i, j)}{d_{\text{diag}}} + |z_i - z_j|, \quad \text{GeoCost}(i, j) < 0.45. \quad (8)$$

In Eq. (8), d_{diag} denotes the image diagonal length and $|z_i - z_j|$ is the absolute depth difference between the two objects. Only pairs satisfying this constraint are retained.

For each selected pair (i, j) , we compute their union bounding box and extract the corresponding region from a sliding 16-frame buffer to construct a short video clip. If one object becomes occluded or undetected, the clip is truncated at the last reliable observation. When the valid interval is shorter than required, the most recent valid frame is repeated to maintain a fixed input length. Using short temporal sequences instead of a single frame encourages predicates to stabilize based on recurring structural patterns rather than instantaneous geometric noise.

Spatiotemporal Predicate Refinement with World-Model. From each 16-frame union-crop clip, we extract temporal representations that encode structural regularities and spatiotemporal context, which serve as evidence for predicate refinement. We freeze the pretrained V-JEPA 2 backbone and attach a lightweight probe network consisting of two fully connected layers. The probe is trained on the training split of each dataset to predict relation distributions. Given the temporal representation, the MLP produces predicate probabilities $p_{\text{WM}}(r_{i \rightarrow j})$ over predicate classes defined in the target dataset [13, 42].

Rather than directly applying frame-wise predictions, we accumulate them at the class-pair level. For each (c_s, c_o) pair, relation distributions are aggregated over time to obtain a stabilized relational prior $P_{\text{WM}}(r \mid c_s, c_o)$. During inference, this prior is estimated without accessing ground-truth relations and is updated solely from model predictions. This temporal accumulation suppresses transient mispredictions and reinforces consistently supported predicates. In parallel, we compute the visually aggregated relation distribution by summing per-frame logits:

$$P_{\text{vis}}(r \mid i, j) = \text{Softmax} \left(\sum_{t \in \mathcal{T}_{ij}} \ell_t(r \mid i, j) \right). \quad (9)$$

We estimate edge uncertainty via the entropy of P_{vis} . The world-model prior is injected only when this entropy exceeds a predefined threshold, indicating relational ambiguity. The final distribution is obtained via uncertainty-aware additive fusion:

$$P_{\text{new}}(r \mid i, j) \propto P_{\text{vis}}(r \mid i, j) + \alpha_{ij} P_{\text{WM}}(r \mid c_s, c_o), \quad (10)$$

where α_{ij} is activated only under high uncertainty. The result is normalized to produce a valid probability distribution. This design complements missing relation probabilities while avoiding overrides of strong visual and geometric evidence.

5 Experiments

Datasets. We evaluate DeWorldSG on the 3DSSG [42] and ReplicaSSG [13] benchmarks. 3DSSG extends 3RScan [41] with 3D semantic scene graph annotations, comprising 1,482 real-world RGB-D scans of 478 indoor environments with instance-level annotations and directional relationships. ReplicaSSG [13] augments the Replica dataset [38] with relational annotations and adopts the Visual Genome [23] object category system to enable zero-shot transfer learning. It uses 7 scenes for validation and 11 scenes for testing.

Evaluation Metrics. We follow the evaluation protocol used in prior 3D Semantic Scene Graph work [10, 13, 21, 42, 44–46]. We report Object Recall, Predicate Recall, Relationship Recall, and mRecall to account for class imbalance. Object Recall measures the fraction of ground-truth object instances that are correctly localized and classified. Predicate Recall measures the fraction of correctly predicted predicates among detected object pairs. Relationship Recall measures the fraction of correctly predicted *subject-predicate-object* triplets for connected object pairs.

Baselines. We compare DeWorldSG against representative existing 3D SSG generation methods. These include methods that explicitly utilize 3D geometry information [42, 44, 45] and lifting-based methods that first estimate a 2D scene graph and then extend it into 3D space [10, 13, 21, 46]. All methods are evaluated under the same label mapping and evaluation criteria to ensure fair comparison.

Implementation Details. All experiments including training, evaluation, and visualization are conducted on a single NVIDIA RTX A6000 GPU with 48 GB of memory. Following [13], we apply a confidence threshold of 0.7 for object filtering and retain only the top 10 relationships per frame. In the DR stage (Eq. (1), Eq. (2)), we set $\tau = 0.05$, $\epsilon = 1e-3$, $\gamma_{\text{base}} = 0.03$, $\alpha = 10$, and $\beta = 0.02$. During global graph merging, we use a Hellinger distance threshold of $\delta_g = 0.7$, and a class distance threshold of $\delta_c = 0.8$. The uncertainty-based relation fusion coefficient is set to $\alpha_{ij} = 0.5$, and the entropy threshold is set to 1.0. For more details, please refer to the Supplementary Material.

5.1 Quantitative Results

We evaluate DeWorldSG on the 3DSSG benchmark, and report the results in Tab. 1. DeWorldSG outperforms prior state-of-the-art across all metrics, achieving +77.4% improvement in Relation Recall, +20.2% improvement in Object Recall, and +23.2% improvement in Predicate Recall. Notably, pointcloud-based methods [42, 44, 45] typically excel at relational reasoning due to explicit 3D geometry. We surpass these methods leveraging only depth-guided lifting and temporal relation accumulation, without scene-level point-cloud inference. Moreover, under class-balanced mRecall, DeWorldSG demonstrates robust 3D scene graph generation despite severe class imbalance. On ReplicaSSG Tab. 2, DeWorldSG substantially outperforms the baseline model, gaining +71.3% in relationship, +35.6% in Object Recall, and +62.9% in Predicate Recall. These results indicate that DeWorldSG’s incremental merging and spatiotemporal relation aggregation

Table 1: Performance comparison of 3D SSG generation methods on the 3DSSG dataset. Note that MonoSSG [44] supports RGB inputs, the reported results are based on predictions using dense reconstruction from RGB-D sequences.

Method	Input Modality	<i>Recall (%)</i>			<i>mRecall (%)</i>	
		Rel.	Obj.	Pred.	Obj.	Pred.
IMP [46]	RGB-D	19.7	49.5	20.9	34.7	13.8
VGFM [10]	RGB	19.6	50.0	20.4	34.8	11.0
3DSSG [42]	Point Cloud	12.9	37.4	22.0	26.2	14.4
SGFN [45]	RGB-D	22.0	51.6	27.5	37.7	24.0
Kim [21]	RGB-D	9.1	59.0	7.1	51.0	8.0
MonoSSG [44]	RGB-D	23.3	53.8	28.4	43.8	26.6
FROSS [13]	RGB-D	27.9	62.4	33.0	63.8	18.0
DeWorldSG (Ours)	RGB-D	50.2	75.0	57.3	70.6	36.0

Table 2: Performance comparison using predicted camera trajectories from the ReplicaSSG dataset and ground-truth camera trajectories as inputs. ‘w/o GT Pose’ denotes the DeWorldSG version utilizing trajectories estimated by ORB-SLAM3 [5].

Method	<i>Recall (%)</i>			<i>mRecall (%)</i>	
	Rel.	Obj.	Pred.	Obj.	Pred.
FROSS [13]	22.3	26.1	27.8	28.8	20.4
Ours	38.2	35.4	45.3	38.1	27.3
FROSS [13] w/o GT Pose	22.7	25.8	27.2	27.7	20.1
Ours w/o GT Pose	37.4	33.8	42.8	37.6	26.9

strategy across consecutive frames mitigate the isolation of objects and relations that commonly arises in single-frame inference, thereby improving scene-level relational consistency.

5.2 Qualitative Results

Fig. 4 qualitatively compares DeWorldSG’s predictions against the previous SoTA method FROSS [13] on ReplicaSSG. DeWorldSG produces more consistent object nodes and spatial extents via depth-refined 3D Gaussians (see Fig. 5), and forms semantically coherent relation graphs through temporal evidence accumulation from world models. In contrast, FROSS frequently exhibits failure cases in which the same object is redundantly instanti-

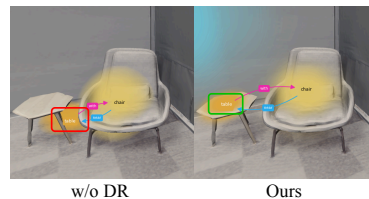


Fig. 5: 3D SSG results under the same camera viewpoint without (left) and with (right) Dual-Domain Refinement (DR). The red rectangle highlights the unstable object position.

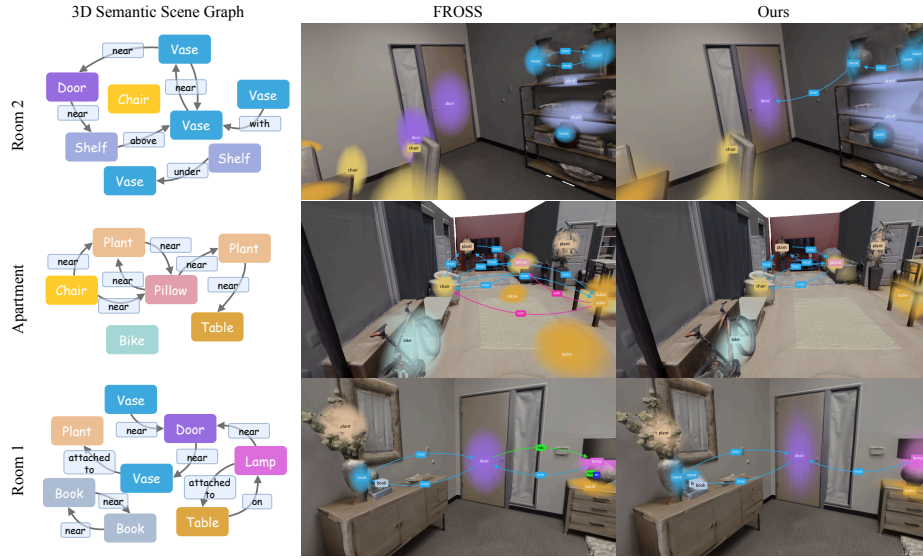


Fig. 4: Qualitative comparison of 3D SSG generation results on ReplicaSSG. (Zoom for details)

ated or object locations are inaccurately projected, leading to overly dense or incorrectly connected relation edges. Additional qualitative comparisons are provided in the Supplementary Material.

5.3 Ablation Study

We analyze DeWorldSG’s key components through four configurations as presented in Tab. 3. The baseline model (a), which uses only 2D scene graph lifting without additional refinement, yields the lowest performance across all metrics. Adding box-conditioned instance masking module (b) showed overall performance improvements across all metrics, confirming that mask-guided depth sampling stabilizes 3D Gaussian estimation for both localization and relation reasoning. Subsequently, the dual-domain depth refinement module (c) further boosts performance by removing depth noise and flying pixels. This validates the reliability of using spatial and depth filtering for global merging. Configuration (d) replaces the V-JEPA 2 video world-model prior with DINOv2 [30] while keeping the same probe-training and relation-injection pipeline. This setting isolates the effect of the temporal world-model prior from generic pretrained visual representations. Although the DINOv2 baseline yields modest predicate gains, DeWorldSG’s full temporal accumulation achieves stronger relation performance, demonstrating that spatiotemporal evidence collection provides more informative cues for relation inference than static image-level representations.

To assess camera pose robustness, we replace ground-truth poses with ORB-SLAM3 [5] estimated poses on ReplicaSSG and report the results in Tab. 2.

Table 3: Ablation results on the 3DSSG benchmark. We evaluate four configurations by progressively enabling instance mask-based depth distribution extraction, Dual-Domain Depth Refinement (DR), predicate refinement with world-model, and DINOv2 [30]-based refinement.

Model	Instance Mask	DR	World Model	DINOv2	Recall (%)			mRecall (%)	
					Rel.	Obj.	Pred.	Obj.	Pred.
(a)	✗	✗	✗	–	28.5	63.4	33.4	64.3	20.5
(b)	✓	✗	✗	–	40.3	71.7	46.7	68.1	26.3
(c)	✓	✓	✗	–	47.8	74.6	52.9	70.8	30.1
(d)	✓	✓	–	✓	49.5	74.6	56.6	69.5	33.7
(e)	✓	✓	✓	–	50.2	75.0	57.3	70.6	36.0

Our model maintains a strong performance (89.2% of GT-pose Object Recall and 87.6% Relationship Recall), confirming stable graph generation under pose noise typical of online AR/robotics systems.

5.4 Runtime

We analyze DeWorldSG’s runtime in the same environment as Sec. 5. The results are averaged over 14,400 frames from four test scenes in the ReplicaSSG dataset to assess practical online feasibility. Module-specific runtime results show that the object detection and instance processing stage takes an average of 4.42 ms per frame, relation inference 7.28 ms, and global 3D SSG merging 2.23 ms per frame. The main computational overhead comes from SAM-based mask segmentation (25.26 ms) and relation refinement process (69.39 ms). Overall, the full pipeline has an average latency of 108.53 ms per frame, demonstrating practical feasibility for incremental 3D scene graph updates in interactive settings.

5.5 Limitations

DeWorldSG relies on a 2D object detector for object class prediction, which can propagate frame-level recognition errors into the constructed 3D scene graph. As shown in Fig. 6, when the same physical object (e.g., a table) receives inconsistent labels across frames (e.g., bench), the system may treat them as distinct instances during graph construction, creating duplicated nodes and fragmented relational structures. Such inconsistencies are more likely under partial occlusion or large viewpoint changes. Future work could address

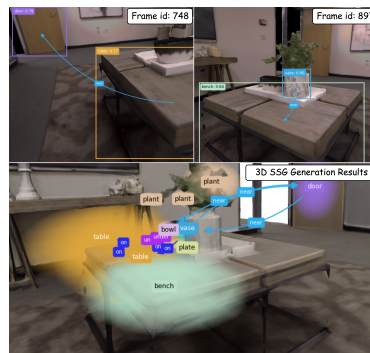


Fig. 6: Failure case caused by inconsistent 2D object classification across frames.

this through cross-frame semantic consistency mechanisms or more robust 2D object detectors.

6 Conclusion

We presented *DeWorldSG*, a framework for generating spatio-temporally coherent 3D semantic scene graphs from RGB-D sequences. By explicitly modeling instance-level depth distributions and leveraging world-model priors for relational reasoning, DeWorldSG enables stable object representation and consistent relation inference under noisy observations. Extensive experiments demonstrate state-of-the-art performance on 3D SSG benchmarks and show that our approach maintains robust graph construction even without ground-truth camera poses.

Beyond benchmark improvements, DeWorldSG provides a practical foundation for 3D scene understanding in embodied systems. The ability to incrementally construct structured scene representations opens new opportunities for applications such as robotic perception, embodied AI agents, and spatially-aware AR systems that require reliable reasoning over objects and their relationships. We believe that integrating probabilistic geometric modeling with predictive world-model priors represents a promising direction toward scalable and interactive 3D scene understanding in real-world environments.

7 Acknowledgements

I would like to extend my sincere gratitude to all of my co-authors for their invaluable contributions to this manuscript. The figures in this paper were designed using resources from Flaticon.com. This paper was supported by Korea Institute for Advancement of Technology(KIAT) grant funded by the Korea Government(MOTIE) (RS-2025-02304167, HRD Program for Industrial Innovation). This work was supported by the IITP (Institute of Information & Communications Technology Planning & Evaluation)-ITRC (Information Technology Research Center) grant funded by the Korea government(Ministry of Science and ICT)(IITP-2026-RS-2024-00436398). This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (RS-2026-25523396, Core Technology Development for Immersive Content).

References

1. Assran, M., Bardes, A., Fan, D., Garrido, Q., Howes, R., Muckley, M., Rizvi, A., Roberts, C., Sinha, K., Zholus, A., et al.: V-jepa 2: Self-supervised video models enable understanding, prediction and planning. arXiv preprint arXiv:2506.09985 (2025)
2. Bardes, A., Garrido, Q., Ponce, J., Chen, X., Rabbat, M., LeCun, Y., Assran, M., Ballas, N.: V-jepa: Latent video prediction for visual representation learning (2023)

3. Booker, M., Byrd, G., Kemp, B., Schmidt, A., Rivera, C.: Embodiedrag: Dynamic 3d scene graph retrieval for efficient and scalable robot task planning. arXiv preprint arXiv:2410.23968 (2024)
4. Büchner, M., Röfer, A., Engelbracht, T., Welschehold, T., Bauer, Z., Blum, H., Pollefeys, M., Valada, A.: Articulated 3d scene graphs for open-world mobile manipulation. arXiv preprint arXiv:2602.16356 (2026)
5. Campos, C., Elvira, R., Rodríguez, J.J.G., Montiel, J.M., Tardós, J.D.: Orb-slam3: An accurate open-source library for visual, visual-inertial, and multimap slam. *IEEE transactions on robotics* **37**(6), 1874–1890 (2021)
6. Chen, D., Shukor, M., Moutakanni, T., Chung, W., Yu, J., Kasarla, T., Bolourchi, A., LeCun, Y., Fung, P.: Vl-jepa: Joint embedding predictive architecture for vision-language. arXiv preprint arXiv:2512.10942 (2025)
7. Crouse, D.F., Willett, P., Pattipati, K., Svensson, L.: A look at gaussian mixture reduction algorithms. In: 14th International Conference on Information Fusion. pp. 1–8. IEEE (2011)
8. Elskhawy, A., Li, M., Navab, N., Busam, B.: Prism-0: A predicate-rich scene graph generation framework for zero-shot open-vocabulary tasks. arXiv preprint arXiv:2504.00844 (2025)
9. Engelbracht, T., Zurbrügg, R., Pollefeys, M., Blum, H., Bauer, Z.: Spotlight: Robotic scene understanding through interaction and affordance detection. In: 2025 IEEE-RAS 24th International Conference on Humanoid Robots (Humanoids). pp. 1–8. IEEE (2025)
10. Gay, P., Stuart, J., Del Bue, A.: Visual graphs from motion (vgfm): Scene understanding with object geometry reasoning. In: Asian Conference on Computer Vision. pp. 330–346. Springer (2018)
11. Gilmer, J., Schoenholz, S.S., Riley, P.F., Vinyals, O., Dahl, G.E.: Neural message passing for quantum chemistry. In: International conference on machine learning. pp. 1263–1272. Pmlr (2017)
12. Gu, Q., Kuwajerwala, A., Morin, S., Jatavallabhula, K.M., Sen, B., Agarwal, A., Rivera, C., Paul, W., Ellis, K., Chellappa, R., et al.: Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 5021–5028. IEEE (2024)
13. Hou, H.Y., Lee, C.Y., Sonogashira, M., Kawanishi, Y.: Fross: Faster-than-real-time online 3d semantic scene graph generation from rgb-d images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 28818–28827 (2025)
14. Hu, W., Wen, X., Li, X., Wang, G.: Dsg-world: Learning a 3d gaussian world model from dual state videos. arXiv preprint arXiv:2506.05217 (2025)
15. Huang, Z., Wu, X., Zhong, F., Zhao, H., Nießner, M., Lasenby, J.: Litereality: graphics-ready 3d scene reconstruction from rgb-d scans. *Advances in Neural Information Processing Systems* (2025)
16. Hughes, N., Chang, Y., Carlone, L.: Hydra: A real-time spatial perception system for 3d scene graph construction and optimization. *Proceedings of Robotics: Science and Systems* (2022)
17. Im, J., Nam, J., Park, N., Lee, H., Park, S.: Egtr: Extracting graph from transformer for scene graph generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 24229–24238 (2024)
18. Kim, B., Kim, T., Lee, J., Joo, H.: Dexterous world models. arXiv preprint arXiv:2512.17907 (2025)

19. Kim, S.Y., Kim, D., Cho, W., Song, H., Kang, S., Woo, W.: Scenelinker: Compositional 3d scene generation via semantic scene graph from rgb sequences. arXiv preprint arXiv:2602.02974 (2026)
20. Kim, S., Kim, D., Son, T., Woo, W.: Realitycrafter: User-guided editable 3d scene generation from a single image in mixed reality. In: Adjunct Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology. pp. 1–3 (2025)
21. Kim, U.H., Park, J.M., Song, T.J., Kim, J.H.: 3-d scene graph: A sparse and semantic representation of physical environments for intelligent agents. *IEEE transactions on cybernetics* **50**(12), 4921–4933 (2019)
22. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016)
23. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision* **123**(1), 32–73 (2017)
24. LeCun, Y., et al.: A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review* **62**(1), 1–62 (2022)
25. Lee, J., Aleotti, F., Mazala, D., Garcia-Hernando, G., Vicente, S., Johnston, O.J., Kraus-Liang, I., Powierza, J., Shin, D., Froehlich, J.E., et al.: Imaginatear: Ai-assisted in-situ authoring in augmented reality. In: Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology. pp. 1–21 (2025)
26. Liu, J., Zhang, R., Butaslac, I., Sawabe, T., Fujimoto, Y., Kanbara, M., Kato, H.: Everywherear: A visual authoring system for creating adaptive ar game scenes. *IEEE Transactions on Visualization and Computer Graphics* (2025)
27. Lv, W., Zhao, Y., Chang, Q., Huang, K., Wang, G., Liu, Y.: Rt-detr2: Improved baseline with bag-of-freebies for real-time detection transformer. arXiv preprint arXiv:2407.17140 (2024)
28. Miao, R., Jia, Q., Sun, F.: Long-term robot manipulation task planning with scene graph and semantic knowledge. *Robotic Intelligence and Automation* **43**(1), 12–22 (2023)
29. Murrugarra-Llerena, J., Kirsten, L.N., Zeni, L.F., Jung, C.R.: Probabilistic intersection-over-union for training and evaluation of oriented object detectors. *IEEE Transactions on Image Processing* **33**, 671–681 (2024)
30. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: DINOv2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
31. Özsoy, E., Czempiel, T., Örneke, E.P., Eck, U., Tombari, F., Navab, N.: Holistic or domain modeling: a semantic scene graph approach. *International Journal of Computer Assisted Radiology and Surgery* **19**(5), 791–799 (2024)
32. Pham, Q.P., Nguyen, K.T., Ngo, L.C., Do, T., Song, D., Hy, T.S.: Tesgnn: Temporal equivariant scene graph neural networks for efficient and robust multi-view 3d scene understanding. arXiv preprint arXiv:2411.10509 (2024)
33. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
34. Rosinol, A., Violette, A., Abate, M., Hughes, N., Chang, Y., Shi, J., Gupta, A., Carloni, L.: Kimera: From slam to spatial perception with 3d dynamic scene graphs. *The International Journal of Robotics Research* **40**(12-14), 1510–1546 (2021)

35. Saxena, S., Buchanan, B., Paxton, C., Liu, P., Chen, B., Vaskevicius, N., Palmieri, L., Francis, J., Kroemer, O.: Grapheqa: Using 3d semantic scene graphs for real-time embodied question answering. arXiv preprint arXiv:2412.14480 (2024)
36. Singh, K.P., Salvador, J., Weihs, L., Kembhavi, A.: Scene graph contrastive learning for embodied navigation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10884–10894 (2023)
37. Song, Y., Sun, P., Liu, H., Li, Z., Song, W., Xiao, Y., Zhou, X.: Scene-driven multimodal knowledge graph construction for embodied ai. IEEE Transactions on Knowledge and Data Engineering **36**(11), 6962–6976 (2024)
38. Straub, J., Whelan, T., Ma, L., Chen, Y., Wijmans, E., Green, S., Engel, J.J., Mur-Artal, R., Ren, C., Verma, S., et al.: The replica dataset: A digital replica of indoor spaces. arXiv preprint arXiv:1906.05797 (2019)
39. Tahara, T., Seno, T., Narita, G., Ishikawa, T.: Retargetable ar: Context-aware augmented reality in indoor scenes based on 3d scene graph. In: 2020 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct). pp. 249–255. IEEE (2020)
40. Tateno, K., Tombari, F., Navab, N.: Real-time and scalable incremental segmentation on dense slam. In: 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 4465–4472. IEEE (2015)
41. Wald, J., Avetisyan, A., Navab, N., Tombari, F., Nießner, M.: Rio: 3d object instance re-localization in changing indoor environments. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7658–7667 (2019)
42. Wald, J., Dharm, H., Navab, N., Tombari, F.: Learning 3d semantic scene graphs from 3d indoor reconstructions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3961–3970 (2020)
43. Wang, Z., Cheng, B., Zhao, L., Xu, D., Tang, Y., Sheng, L.: Vl-sat: Visual-linguistic semantics assisted training for 3d semantic scene graph prediction in point cloud. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 21560–21569 (2023)
44. Wu, S.C., Tateno, K., Navab, N., Tombari, F.: Incremental 3d semantic scene graph prediction from rgb sequences. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5064–5074 (2023)
45. Wu, S.C., Wald, J., Tateno, K., Navab, N., Tombari, F.: Scenegrphfusion: Incremental 3d scene graph prediction from rgb-d sequences. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7515–7525 (2021)
46. Xu, D., Zhu, Y., Choy, C.B., Fei-Fei, L.: Scene graph generation by iterative message passing. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5410–5419 (2017)
47. Yan, Z., Li, S., Wang, Z., Wu, L., Wang, H., Zhu, J., Chen, L., Liu, J.: Dynamic open-vocabulary 3d scene graphs for long-term language-guided mobile manipulation. IEEE Robotics and Automation Letters (2025)
48. Yang, Y., Liu, J., Zhang, Z., Zhou, S., Tan, R., Yang, J., Du, Y., Gan, C.: Mind-journey: Test-time scaling with world models for spatial reasoning. arXiv preprint arXiv:2507.12508 (2025)
49. Yang, Z., Lu, K., Zhang, C., Qi, J., Jiang, H., Ma, R., Yin, S., Xu, Y., Xing, M., Xiao, Z., et al.: Mmgdreamer: Mixed-modality graph for geometry-controllable 3d indoor scene generation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 9391–9399 (2025)

50. Zhai, G., Örnek, E.P., Chen, D.Z., Liao, R., Di, Y., Navab, N., Tombari, F., Busam, B.: Echoscene: Indoor scene generation via information echo over scene graph diffusion. In: European Conference on Computer Vision. pp. 167–184. Springer (2024)
51. Zhai, G., Örnek, E.P., Wu, S.C., Di, Y., Tombari, F., Navab, N., Busam, B.: Commonsences: Generating commonsense 3d indoor scenes with scene graph diffusion. *Advances in Neural Information Processing Systems* **36**, 30026–30038 (2023)
52. Zhang, Y., Wu, X., Lao, Y., Wang, C., Tian, Z., Wang, N., Zhao, H.: Concerto: Joint 2d-3d self-supervised learning emerges spatial representations. *arXiv preprint arXiv:2510.23607* (2025)
53. Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., Liu, Y., Chen, J.: Detrs beat yolos on real-time object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16965–16974 (2024)