
— Supplementary Material —

Object-X: Learning to Reconstruct Multi-Modal 3D Object Representations

Anonymous Author(s)

Affiliation

Address

email

1 Overview

2 This supplementary material provides additional training details, experimental setups, visualizations,
3 and ablation studies in support of the main paper. It is organized as follows:

- 4 1. Additional details on baseline comparisons and visualizations for object-level, image-based,
5 and scene-level reconstruction (**Section A**)
- 6 2. Training procedures for pretraining, compression, and downstream adaptation (**Section B**)
- 7 3. Setup and evaluation details for visual localization using U-3DGS embeddings (**Section C**)
- 8 4. Scene alignment task setup, including sub-scene construction and evaluation metrics (**Section D**)
- 9 5. Ablation results on compression, occlusion robustness, and architectural variants (**Section E**)

11 A Baselines and Visualizations

12 This section details the implementation and evaluation protocols for all baseline methods discussed
13 in the main paper. We cover object-level, scene-level, and single-image reconstruction settings. To
14 ensure a fair and rigorous comparison across all experiments, we maintain consistent supervision
15 levels, initialization strategies, and evaluation metrics when assessing storage requirements, visual
16 quality, and geometric fidelity.

17 A.1 Object Reconstruction Baselines

18 All optimization-based methods, specifically 3D Gaussian Splatting (3DGS) (4) and 2DGS (10),
19 operate on masked RGB-D input sequences. Our comparative analysis includes three primary
20 optimization-based baselines:

- 21 1. **3DGS (Full Scene)**: Utilizes *all* available images from a given scene to serve as an upper-
22 bound reference reconstruction.
- 23 2. **3DGS (kV)**: Employs k pre-selected views that observe the target object.
- 24 3. **2DGS (kV)**: Also uses k pre-selected views observing the target object.

25 To ensure a fair comparison by providing identical starting conditions for all methods, we initialize
26 Gaussian splats directly from ground-truth object meshes. The k views for object-specific baselines
27 are selected using a k -means clustering strategy (detailed below) to promote viewpoint diversity and
28 ensure high object visibility. Crucially, evaluation is consistently performed on a disjoint test set of
29 images that were *not* utilized during the training or optimization phases of any method. For both
30 3DGS and 2DGS, we conduct optimization for 7,000 iterations using their default hyperparameter
31 settings. For experiments on the 3RScan (3) dataset, we use 12 views and, on ScanNet (1), which

typically offers fewer views per object instance, we restrict the number of selected views, k , to a maximum of 4 per object.

Regarding DepthSplat (9), we evaluate using the publicly available pre-trained model. As this model was not trained on object reconstruction tasks, we apply it to the *unmasked* image and, we post-process its output by removing any splats that fall entirely outside the masked object region.

Visual comparisons are provided in Figure 1. Alongside rendered novel views, we present mesh reconstructions derived from the 3D Gaussians using the TSDF fusion technique, as proposed in (10). These visualizations demonstrate that our proposed method, *Object-X*, achieves significantly smoother novel view syntheses and more geometrically accurate mesh reconstructions compared to the baselines.

Frame Selection Protocol. To ensure consistent and representative view selection across all relevant experiments (object-level and scene-level k -view baselines), we employ a clustering-based strategy for choosing training/optimization views. From the available set of frames for an object or scene, we first cluster their camera extrinsics (position and orientation) using k -means. Subsequently, we select one frame from each resulting cluster, prioritizing the frame that exhibits the fewest masked pixels (i.e., maximal object visibility within the frame). Any objects for which no valid test images remain after this selection process (e.g., due to insufficient visibility in all remaining frames not reserved for testing) are excluded from the evaluation set to maintain fairness.

A.2 Scene-Level Reconstruction

For full-scene evaluation, we adapt 3DGS and 2DGS to operate jointly across all objects within a scene. This is achieved by utilizing the union of the same k views per object as in the object-level experiments (specifically, 12 views for 3RScan and 4 for ScanNet), but here the RGB-D inputs are *unmasked*. Our method, *Object-X*, reconstructs the scene by independently decoding the learned U-3DGS embedding for each constituent object and then rendering their collective splats. This compositional approach requires no additional scene-level optimization. We also evaluate an augmented version, denoted as ***Object-X + Opt***. This variant leverages the compositional scene from *Object-X* as an initialization for a subsequent refinement stage. Specifically, it undergoes an additional 4,000 iterations of 3DGS optimization. To ensure stability during this fine-tuning process, all learning rates are reduced by a factor of $10\times$ compared to the standard 3DGS settings.

Visualizations of scene-level reconstructions are presented in Figure 3. While *Object-X* generally produces significantly smoother results than the baseline methods, its performance can be affected by objects missing from the input segmentations (e.g., a poster on a wall, as shown in the first row of the figure, or the objects on the desk, as shown in the second row). Additionally, fine-grained details might sometimes be diminished. However, applying 3DGS optimization as a post-processing step (***Object-X + Opt***) yields substantial improvements in accuracy, effectively recovering such lost details.

A.3 Single-Image Reconstruction

We extend our evaluation to a single-view reconstruction setting for all methods. In this scenario, 3DGS is optimized from scratch using a single masked RGB-D image (and its corresponding RGB image) for 3,000 iterations. To ensure a fair comparison, *Object-X* utilizes the same reference image. This image is selected based on criteria that maximize unmasked object coverage while minimizing cropping along the image borders. We also test our method with only RGB input with depth predicted by Metric3D (2) to generate the initial point cloud. Similarly to the scene reconstruction case, we use *Object-X* to provide an initial reconstruction which we further refine by applying an additional 1,000 iterations of 3DGS optimization, using learning rates reduced by a factor of $10\times$ (consistent with the scene-level refinement). Visual results for this setting are presented in Figure 2.

Notably, despite *Object-X* not being explicitly trained for single-image reconstruction tasks, it frequently produces visually cleaner reconstructions than 3DGS when both methods are constrained to the same single input view and 3DGS is optimized from scratch under these conditions. The reconstructed meshes are also substantially more accurate than the ones from 3DGS.

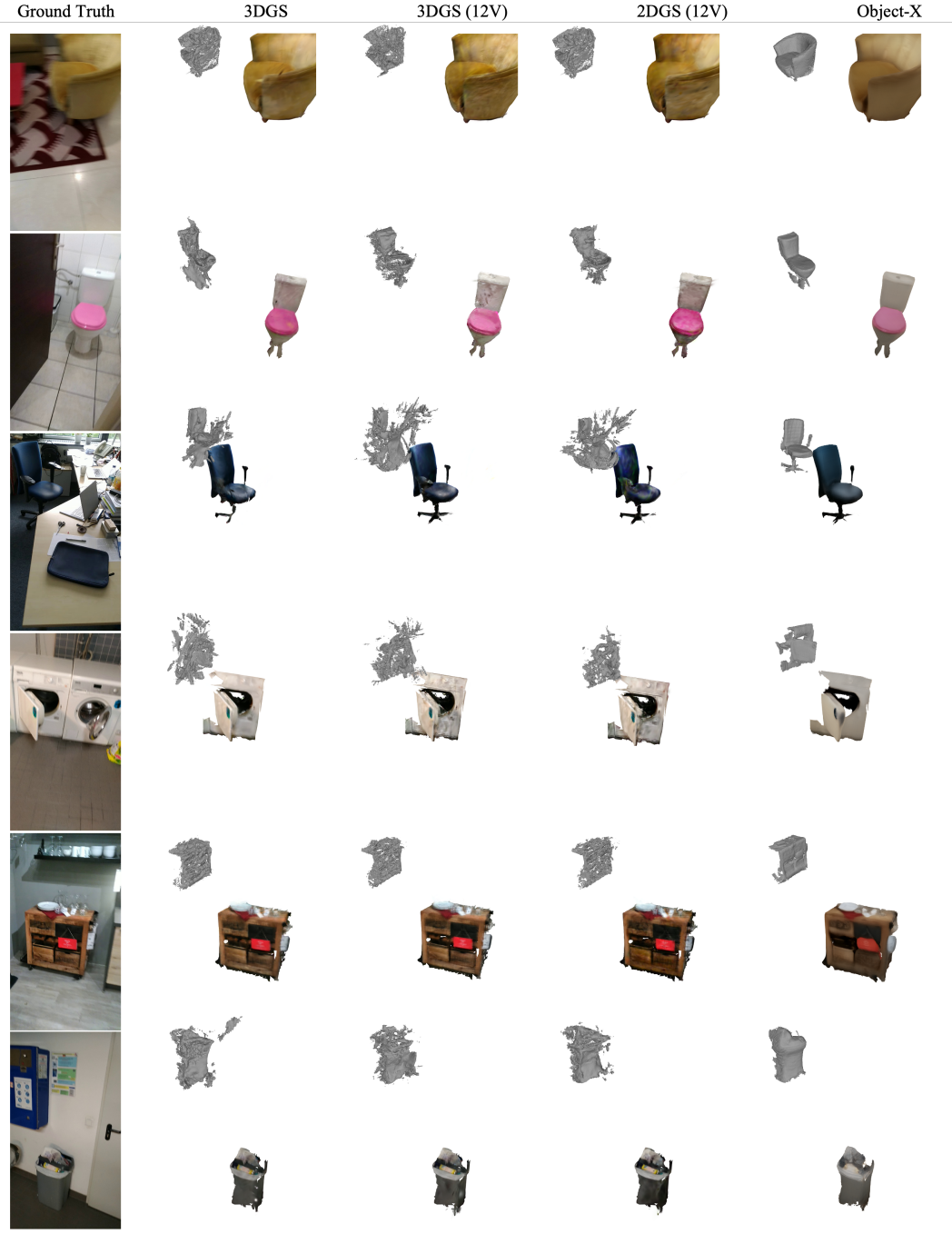


Figure 1: **Object reconstructions.** Each row shows an input object (left) and its reconstruction obtained by, from left to right: (i) 3DGS (8) optimized on all images, (ii) 3DGS or (iii) 2DGS (10) using only 12 multi-view images, and (iv) *Object-X*. For each method, we present a rendered image from the reconstructed 3D Gaussians and the corresponding mesh.

82 A.4 Evaluation Summary

83 Across all experimental settings, methods are evaluated on a fixed set of test views, distinct from
84 training/optimization views, on a per-object or per-scene basis as appropriate. We report standard
85 quantitative metrics, Peak Signal-to-Noise Ratio (PSNR), and the perceptual metric LPIPS. Qualitative

86 comparisons are provided in the relevant figures accompanying each experimental section (*e.g.*,
87 Figure 1 for object-level, Figure 3 for scene-level, and Figure 2 for single-image results). Across
88 all evaluated levels – single-view, multi-view object reconstruction, and full-scene composition –
89 *Object-X* demonstrates strong performance, simultaneously offering significant advantages in terms
90 of computational efficiency and flexibility in initialization.

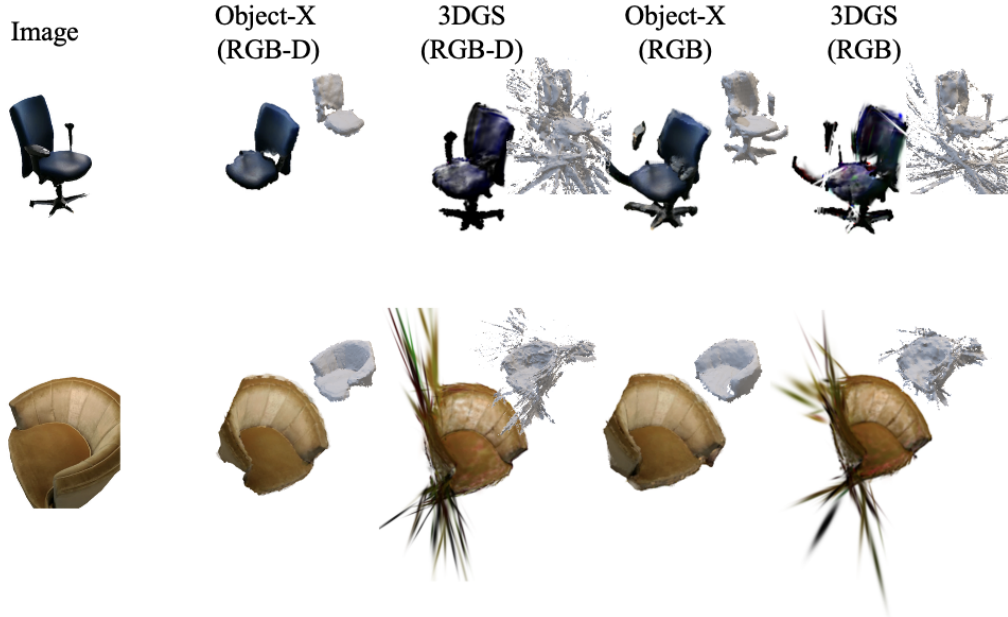


Figure 2: **Qualitative comparison for image to 3D.** We compare the proposed *Object-X* to standard 3DGS (8) on RGB and RGB-D inputs. For each method, we present the image from which the object (left column) is reconstructed and the rendered novel view together with the mesh reconstructed from the 3D Gaussians by: (2nd column) proposed *Object-X* with RGB-D input; (3rd) 3DGS with RGB-D; (4th) proposed *Object-X* with RGB input; (5th) 3DGS with RGB. The proposed method leads to significantly cleaner novel views and meshes than 3DGS applied to a single image.

91 B Training Details

92 The training procedure encompasses three primary phases: sparse representation learning, compression model training, and adaptation for downstream tasks. Each phase employs distinct optimization
93 settings to ensure both stability and efficiency. For the sparse transformer-based encoder and decoder,
94 we apply gradient clipping at a threshold of 0.01. This is crucial for stabilizing the training process
95 and preventing excessively large updates within the structured latent space. Optimization is conducted
96 using the AdamW optimizer with a learning rate of 1×10^{-4} . This learning rate is selected to strike
97 an effective balance between training stability and convergence speed. AdamW is chosen for its
98 decoupled weight decay mechanism, which aids in regularizing the model without adversely affecting
99 the gradient-based optimization updates. During the compression phase, a 3D U-Net architecture
100 is trained to map the structured latent representation to a more compact form suitable for efficient
101 storage or transmission. A higher learning rate of 1×10^{-3} is utilized in this phase. This facilitates
102 accelerated convergence while preserving reconstruction quality. Explicit gradient clipping is not
103 deemed necessary for the U-Net, as its inherent hierarchical structure and typical training dynamics
104 provide sufficient stabilization.

106 For adaptation to downstream tasks, such as object localization or instance retrieval, training is
107 performed using the AdamW optimizer with a learning rate of 1×10^{-3} when the voxel-based latent
108 representation is kept frozen. However, if the voxel representation is fine-tuned concurrently with the
109 task-specific modules, a lower learning rate of 1×10^{-4} is adopted. This approach helps to mitigate
110 the risk of catastrophic forgetting of the learned representations. Key regularization techniques

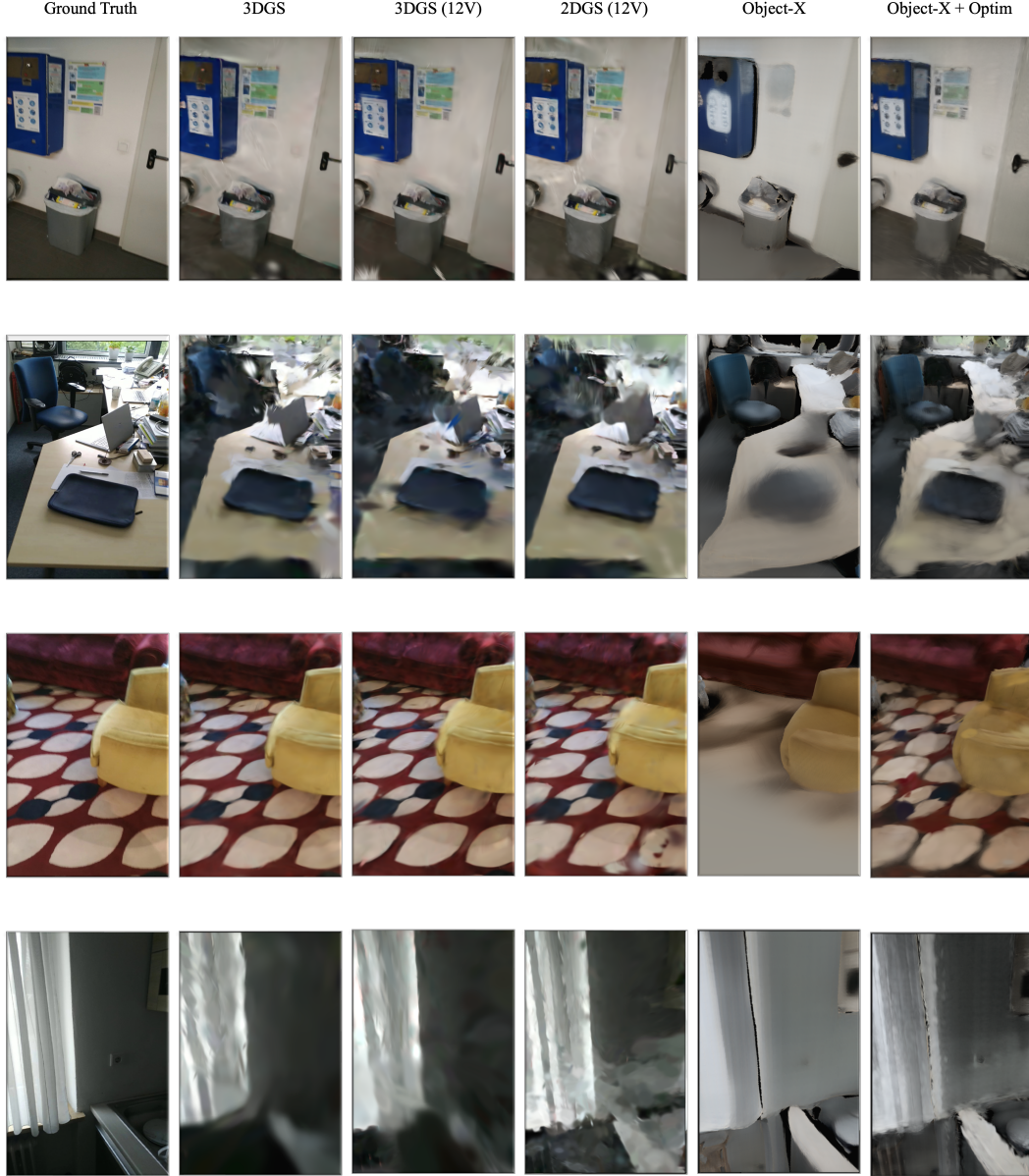


Figure 3: **Qualitative comparison for full-scene composition.** We compare the proposed *Object-X* to standard 3DGS (8) optimized on all unmasked scene images, and two 12-view baselines: 3DGS (12V) and 2DGS (12V), which optimize scenes using a subset of training images constructed by taking the union of the 12 best views selected per object.

111 employed include the aforementioned gradient clipping and structured weight decay (e.g., as provided
 112 by the AdamW optimizer).

113 C Supplementary: Coarse Visual Localization

114 We provide additional details for the visual localization experiment on the 3RScan dataset. This ex-
 115 periment is designed to evaluate the downstream utility of the U-3DGS embeddings when augmented
 116 with auxiliary modalities.

117 **Training.** The model utilized for localization is trained following the auxiliary learning setup
 118 described in the main paper. We freeze the pre-trained U-3DGS encoder and decoder. Auxiliary

encoders are then trained on object-level inputs derived from 3D scene graphs, specifically object relationships, attributes, and structural context. Each RGB query image is processed through a DINOv2 backbone followed by a patch-level encoder to generate patch-wise descriptors. A contrastive loss function aligns these image patches with the corresponding object embeddings within the scene. Concurrently, a compression loss ensures that the original U-3DGS component of the joint embedding remains accurately decodable. After this initial stage, all modules, including the U-3DGS components and auxiliary encoders, are jointly fine-tuned to enhance task-specific performance while preserving reconstruction fidelity.

Setup. Following the evaluation protocol established by SceneGraphLoc (5), we sample 123 distinct scenes from 30 rooms within the 3RScan test split. For each query image, the objective is to identify the correct scene from a candidate pool of 10 scenes, which includes the ground-truth scene. This experimental setup results in a total of 30,462 *query evaluations*.

Evaluation. At test time, each posed RGB image is encoded into a set of patch-level embeddings. For every candidate scene in the pool, these image patch embeddings are compared against all available object embeddings from that scene using cosine similarity. The final prediction for the scene is determined via a robust voting mechanism that aggregates all patch-object similarity scores. We report **Recall@K** for $K \in \{1, 3, 5\}$, which measures the frequency with which the correct scene appears among the top-K predicted scenes. This metric directly reflects the model’s capability to localize images effectively using the learned, object-centric multimodal representation.

Results. We compare our results with SceneGraphLoc (5) and the recent CrossOver method (7). Our approach demonstrates competitive localization accuracy while crucially maintaining compatibility with 3D reconstruction and other downstream applications, a benefit stemming from our jointly trained, modular representation. Detailed results are presented in Table 1. In this table, we indicate whether a given method utilizes point clouds (\mathcal{P}), images (\mathcal{I}), other modalities such as object attributes and relationships (\mathcal{O}), or the proposed U-3DGS embedding. Our method, leveraging U-3DGS embeddings in conjunction with other modalities, achieves the highest Recall@1 and Recall@3 scores. This outcome suggests that the proposed U-3DGS embeddings furnish information comparable to, or even richer than, that provided by raw point clouds or images for the task of visual localization.

Furthermore, we present an ablation study for our method (indicated with an asterisk * in Table 1) using only the U-3DGS embeddings, without any specific fine-tuning of our main encoder for this localization task. As anticipated, the auxiliary modalities (attributes, relationships, etc.) offer valuable complementary information, contributing significantly to the superior performance of the full model. Interestingly, even in this constrained setting (U-3DGS embeddings alone, without targeted training), our method performs comparably to the recent CrossOver approach (7). This highlights the inherent richness and suitability of our learned U-3DGS embeddings for visual localization tasks, even without explicit optimization for this specific application.

Method	Modalities				10 scenes		
	\mathcal{P}	\mathcal{I}	\mathcal{O}	3DSG	Recall@1	Recall@3	Recall@5
SGLoc (5)	✓	✗	✓	✗	53.6	81.9	92.8
CrossOver (7)	✗	✓	✗	✗	46.0	77.9	90.5
Object-X	✗	✗	✓	✓	56.6	82.2	91.8
Object-X*	✗	✗	✗	✓	28.7	58.5	76.5
Object-X*	✗	✗	✓	✓	44.8	72.6	85.7

Table 1: **Coarse visual localization** on the 3RScan dataset (3) using the proposed U-3DGS embedding, compared to SceneGraphLoc (5) and CrossOver (7). We report retrieval recall at 1, 3, and 5 when selecting the correct scene from 10 candidates. Evaluations are conducted using different map modalities: point cloud (\mathcal{P}), image (\mathcal{I}), other modalities (\mathcal{O}), and 3DGS. In the lower section (*), we also present results where the U-3DGS embedding is used without task-specific training.

D Supplementary: Scene Alignment

We provide additional details on the setup and evaluation procedure for the 3D Scene Alignment task on the 3RScan dataset (3), following the protocol established by SAligner (6).

Setup. To construct the evaluation data, we generate sub-scenes by selecting fixed-length sequences of consecutive RGB-D frames from the 3RScan validation set. Each such sequence is then fused into a partial 3D reconstruction using volumetric integration. This process results in a total of 848 *sub-scenes*, each representing a distinct viewpoint or region within an original, larger scene. From these sub-scenes, we create 1,906 *pairs* by selecting pairs that originate from the same ground-truth scene. These pairs are deliberately constructed to span a wide range of spatial overlap percentages (from 10% to 90%) thereby ensuring coverage of both straightforward and challenging alignment scenarios.

Evaluation. We extract object embeddings independently from each sub-scene. These embeddings are produced by the same network architecture and weights trained for the scene localization task, as detailed in Section C. During the evaluation phase, we compute the *cosine similarity* between every object embedding in one sub-scene and all object embeddings in its paired sub-scene. For each object in the first sub-scene, candidate objects from the second sub-scene are ranked based on this similarity score. We evaluate the quality of these rankings using standard retrieval metrics: **Mean Reciprocal Rank (MRR)** and **Hits@K**, where $K \in \{1, 2, \dots, 5\}$. The Hits@K metric measures the proportion of queries for which a correct match appears within the top-K ranked results, while MRR quantifies the average inverse rank of the first correct match.

E Ablation Studies

This section analyzes the impact of key components and design choices in our proposed method. Table 2 presents results as a function of the compression rate, which is defined by the resolution of the underlying voxel grid, where each voxel stores eight parameters. As a reference, we also report results for standard 3D Gaussian Splatting (3DGS). In addition to our proposed 3D U-Net architecture for compression, we evaluate a naive downsampling approach that applies max pooling followed by interpolation.

The results corresponding to a 64^3 voxel grid resolution effectively represent our Structured Latent (SLat) representation without any subsequent compression, as this directly matches the original voxel resolution described in the main paper. The ablation results demonstrate that employing naive downsampling leads to a significant degradation in accuracy as the resolution decreases. In contrast, our proposed 3D U-Net maintains high fidelity with only a marginal loss in accuracy, while substantially reducing the number of parameters required per object from $64^3 \times 8 = 2\,097\,152$ to a mere $8^3 \times 8 = 4\,096$. Based on this analysis, we adopt a resolution of 16^3 for the compressed representation in all our main experiments.

Table 3 evaluates the robustness of our method to varying degrees of occlusion by systematically removing parts of an object before it is encoded. Occlusion is simulated by selecting a random point on the object’s surface and removing all geometry within a sphere of diameter d . The diameter d is defined as a fraction of the object’s characteristic size; for example, $d = 0.4$ corresponds to approximately 40% of the object’s volume being removed. The results indicate that even under severe occlusion, our proposed method maintains high reconstruction accuracy, thereby demonstrating its resilience to incomplete or missing input data.

Resolution	Method	LPIPS (Mean \pm σ) \downarrow	Median \downarrow	PSNR (Mean \pm σ) \uparrow	Median \uparrow
3DGS	-	0.086 ± 0.082	0.060	30.15 ± 5.06	30.14
64 ³ (SLat)	-	0.094 ± 0.101	0.059	27.30 ± 6.13	27.06
32 ³	Naive 3D Unet	0.124 ± 0.126 0.099 ± 0.108	0.076 0.060	25.28 ± 5.51 27.06 ± 6.18	25.80 26.84
16 ³	Naive 3D Unet	0.189 ± 0.137 0.103 ± 0.113	0.137 0.062	21.32 ± 5.53 27.01 ± 6.29	21.41 26.71
8 ³	Naive 3D Unet	0.257 ± 0.187 0.110 ± 0.119	0.211 0.065	17.46 ± 5.26 26.74 ± 6.35	16.86 26.50

Table 2: **Ablation study on latent dimensions.** Mean and median LPIPS and PSNR on a subset of scans from the test set. We compare the standard 3DGS (as a reference), the SLat embedding without dimensionality reduction, and U-3DGS with compressed representations at 32³, 16³, and 8³. Also, we evaluate naive downscaling approaches using max pooling and interpolation alongside the proposed 3D U-Net. The 16³ resolution is selected for all other experiments as it significantly reduces storage while maintaining near-optimal reconstruction accuracy.

d	LPIPS (Mean \pm σ) \downarrow	Median \downarrow	PSNR (Mean \pm σ) \uparrow	Median \uparrow
0.0	0.104 ± 0.114	0.062	26.85 ± 6.25	26.66
0.1	0.104 ± 0.113	0.063	26.96 ± 6.32	26.69
0.2	0.106 ± 0.114	0.064	26.70 ± 6.44	26.50
0.4	0.113 ± 0.119	0.068	26.07 ± 6.70	25.93

Table 3: **Ablation study on occlusion.** Before encoding an object, we randomly select a point on its surface and remove all parts within a spherical region of diameter d . For example, $d = 0.4$ corresponds to a removal region spanning 40% of the object’s size. We report LPIPS and PSNR scores for different values of d to assess the impact of occlusion on reconstruction quality.

References

- [1] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017. Available for academic use under custom license. See <http://www.scan-net.org/> for details.
- [2] Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.
- [3] Nassir Navab Federico Tombari Matthias Niessner Johanna Wald, Armen Avetisyan. Rio: 3d object instance re-localization in changing indoor environments. 2019. Dataset licensed under CC BY-NC-SA 4.0: <https://creativecommons.org/licenses/by-nc-sa/4.0/>.
- [4] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), July 2023.
- [5] Yang Miao, Francis Engelmann, Olga Vysotska, Federico Tombari, Marc Pollefeys, and Dániel Béla Baráth. Scenegraphloc: Cross-modal coarse visual localization on 3d scene graphs, 2024.
- [6] Sayan Deb Sarkar, Ondrej Miksik, Marc Pollefeys, Daniel Barath, and Iro Armeni. Sgaligner : 3d scene alignment with scene graphs, 2023.
- [7] Sayan Deb Sarkar, Ondrej Miksik, Marc Pollefeys, Daniel Barath, and Iro Armeni. Crossover: 3d scene cross-modal alignment. *Conference on Computer Vision and Pattern Recognition*, 2025.
- [8] Johanna Wald, Helisa Dharmo, Nassir Navab, and Federico Tombari. Learning 3d semantic scene graphs from 3d indoor reconstructions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [9] Haofei Xu, Songyou Peng, Fangjinhua Wang, Hermann Blum, Daniel Barath, Andreas Geiger, and Marc Pollefeys. Depthsplat: Connecting gaussian splatting and depth. *Computer Vision and Pattern Recognition*, 2025.
- [10] Zehao Zhu, Shaohui Ding, Tianhang Wu, Yi Zhou, Ying Feng Yu, and Hao Wang. 2d gaussian splatting for geometrically accurate radiance fields. *arXiv preprint arXiv:2312.05817*, 2023.