SAB3R: Semantic-Augmented Backbone in 3D Reconstruction

Anonymous CVPR submission

Paper ID 0000

Abstract

*We introduce a new task, **Map and Locate**, which unifies the traditionally distinct objectives of open-vocabulary segmentation—detecting and segmenting object instances based on natural language queries—and 3D reconstruction, the process of estimating a scene's 3D structure from visual inputs. Specifically, Map and Locate involves generating a point cloud from an unposed video and segmenting object instances based on open-vocabulary queries. This task serves as a critical step toward real-world embodied AI applications and introduces a practical task that bridges reconstruction, recognition and reorganization.*

*To tackle this task, we introduce a simple yet effective baseline, which we denote as **SAB3R**. Our approach builds upon MASt3R, a recent breakthrough in 3D computer vision, and incorporates a lightweight distillation strategy. This method transfers dense, per-pixel semantic features from 2D vision backbones (*e.g.*, CLIP and DINOv2) to enhance MASt3R's capabilities. Without introducing any auxiliary frozen networks, our model generates per-pixel semantic features and constructs cohesive point maps in a single forward pass.*

Compared to separately deploying MASt3R and CLIP, our unified model, SAB3R, achieves superior performance on the Map and Locate benchmark. Furthermore, we evaluate SAB3R on both 2D semantic segmentation and 3D tasks to comprehensively validate its effectiveness.

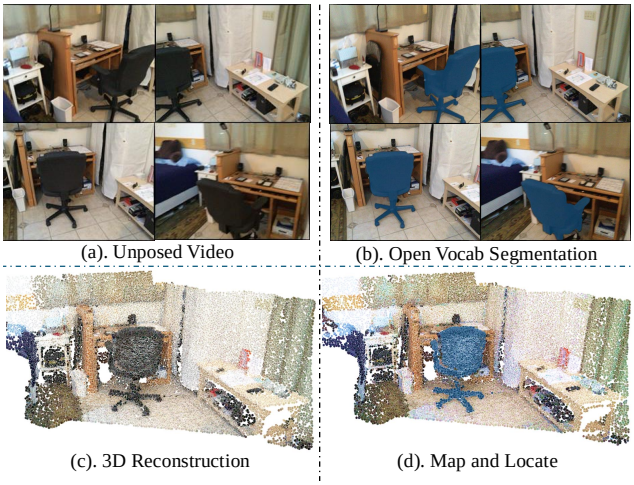

Figure 1. Given an unposed input video (a), we show ground truth for: (b) open-vocab semantic segmentation (per-pixel labels for the prompt "a black office chair"), (c) 3D reconstruction (ground-truth point cloud), and (d) the proposed *Map and Locate* task (open-vocab segmentation for the prompt "a black office chair" and point cloud). The Map and Locate task: (1) encompasses both 2D and 3D tasks, (2) bridges reconstruction and recognition, and (3) introduces practical questions in robotics and embodied AI. The Map and Locate generalizes both 2D and 3D tasks, and we expect this unified approach to present novel challenges and enable innovative new methods.

1. Introduction

Current 3D open-vocabulary segmentation methods [42, 51, 68] typically assume access to complete, high-quality point clouds—an assumption that rarely holds in real-world embodied AI scenarios. One major challenge lies in the high cost and complexity of curating large-scale 3D open-vocabulary datasets, even with prior efforts such as ScanRefer [9] and ReferIt3D [1], which remain limited in both scale and diversity. Additionally, existing methods either depend on precise camera poses and sensor calibration for accurate point cloud reconstruction, an impractical requirement in continuously changing environments, or rely on test-time optimization techniques [24, 37], which are computationally expensive and unsuitable for real-time applications. Despite these challenges, human perception effortlessly integrates 2D visual semantics with 3D structural understanding, leveraging depth cues and object motion over a lifetime of interaction [26]. Thus, we aim to explore how a model can simultaneously perform 2D semantic understanding and 3D reconstruction, bridging the gap between segmentation and spatial reasoning in open-vocab settings.

Malik et al. [36] categorize vision tasks into recognition, reconstruction, and reorganization. Recognition involves assigning semantic categories to images, reconstruction focuses on estimating 3D structures, and reorganization deals

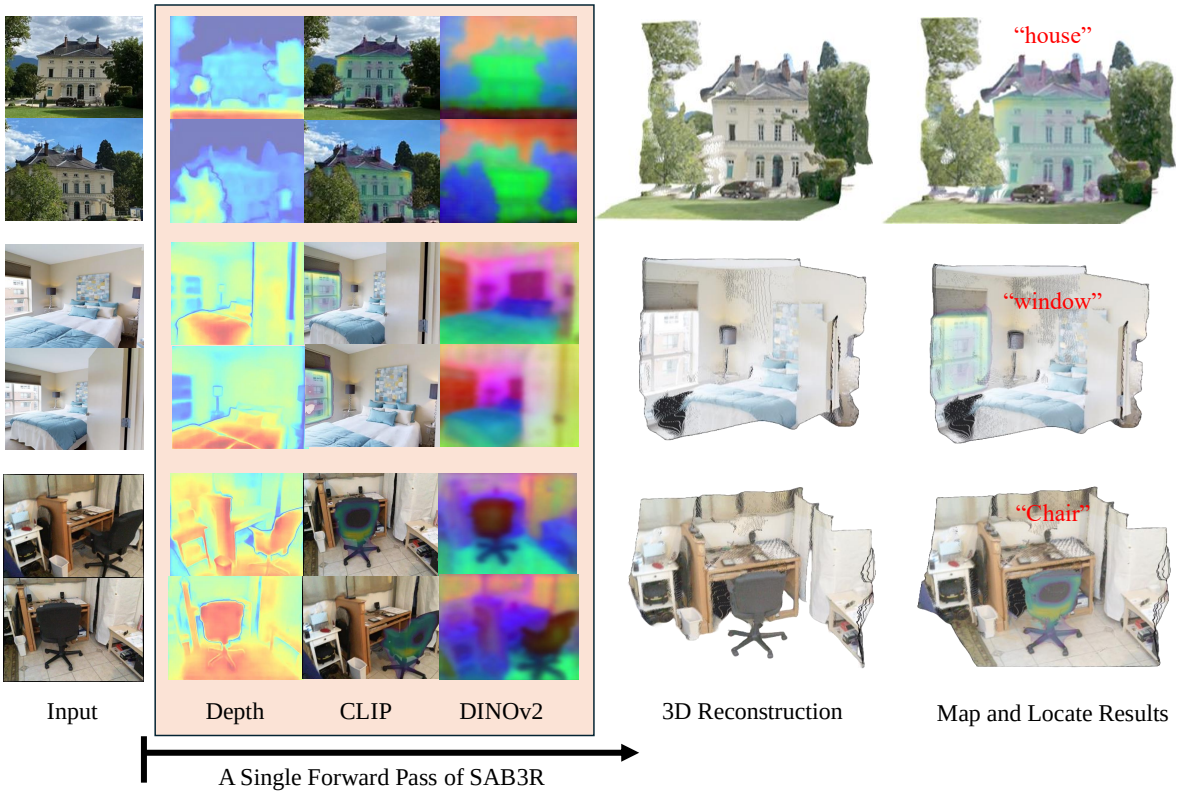

Figure 2. Our method, SAB3R , a semantic-augmented backbone for 3D reconstruction, enables zero-shot open-vocabulary segmentation and 3D reconstruction from unposed images in a single forward pass. By jointly performing reconstruction and open-vocabulary semantic segmentation, SAB3R introduces a novel capability that unifies these tasks within a single framework.

with grouping and segmenting images based on spatial or perceptual similarity. Ideally, these tasks should mutually benefit one another. Moreover, maintaining separate models for different vision tasks is inefficient, incurring high memory and runtime costs [48]. This raises a critical question: *Can 3D open-vocabulary segmentation and 3D reconstruction be effectively reconciled?*

Therefore, our work addresses this challenge by drawing inspiration from human visual perception. As human can seamlessly interpret images by combining 2D visual information with an intuitive understanding of 3D structure. While existing methods take in posed RGB-D sequences or pre-scanned environments, we propose using unposed video as input—a natural and accessible modality for embodied agents operating in real-world settings. As illustrated in Figure 1, our Map and Locate task jointly constructs a 3D geometric map and segments objects specified through open-vocabulary queries. This approach enables simultaneous spatial mapping, semantic understanding, and segmentation without requiring pre-processed point clouds. To this end, we introduce a simple yet effective baseline SAB3R , as shown in Figure 2, which takes unposed images as input and predicts a point map, dense CLIP features, and dense DINOv2 features in a single forward pass.

This integration offers three key advantages. First, it eliminates the reliance on high-quality, pre-scanned point clouds by taking in unposed video as input. Second, it removes the dependence on precise camera poses and sensor calibrations, making 3D segmentation and reconstruction feasible in real-world environments without test-time optimization, which is often computationally prohibitive. Third, it unifies recognition, reorganization and reconstruction into a single model, reducing memory and runtime overhead. By bridging the gap between open-vocab segmentation and reconstruction, our approach offers a more practical and scalable solution for embodied perception.

In summary, our contributions are:

- *Map and Locate* Benchmark: We introduce a novel benchmark for multi-view 3D semantic segmentation that jointly addresses the tasks of reconstruction, reorganization, and recognition. The benchmark is accompanied by a large-scale dataset, clearly defined evaluation protocols, and standardized metrics.
- SAB3R : We propose a unified framework that concurrently performs open-vocabulary segmentation and 3D reconstruction from unposed images via an efficient distillation strategy. We present it as a baseline due to its performance and computational efficiency.

2. Related Work

2.1. 3D Reconstruction

The landscape of 3D reconstruction has evolved from traditional geometric methods like SfM [2, 52] and SLAM [6, 38] to learning-based approaches that leverage data-driven priors [59, 63]. DUSt3R [64] pioneered a paradigm shift by predicting dense point maps from image pairs in a shared coordinate frame, removing the need for explicit pose supervision. However, its reliance on stereo inputs limits its applicability to multi-view settings. More recently, MASt3R [29] extended this idea by learning viewpoint-invariant representations for dense point prediction across multiple images, significantly improving robustness in unposed scenarios. While these advances enable reconstructing 3D geometry from unconstrained image sequences, they primarily focus on geometric consistency and do not incorporate high-level semantics.

Our work builds upon MASt3R and extends it to the novel Map and Locate task, which bridges 3D reconstruction with open-vocabulary segmentation. Unlike prior methods that treat reconstruction and recognition as separate problems, we introduce a unified approach that simultaneously maps the environment and segments objects based on free-form queries. This perspective transforms 3D perception into a richer and more interactive task, opening new avenues for embodied AI and scene understanding beyond purely geometric reconstruction.

2.2. Leveraging 2D for 3D Vision

Most 3D visual-language models operate directly on 3D point clouds without leveraging 2D pre-trained features. SAT-2D [69] was one of the first 3D visual grounding models to incorporate 2D visual features, aligning 2D and 3D representations during training and achieving significant improvements over versions without 2D features. More recent approaches, such as 3DLLM [21] in 3D Question Answering, use multi-view 2D features with LLMs to decode answers, but have yet to fully address 3D visual grounding tasks. Similarly, PQ3D [81] integrates various visual backbones, including a 2D feature backbone from OpenScene [42].

EFM3D [56] lifts 2D image features into 3D feature volumes, but focuses on 3D object detection and surface reconstruction. ODIN [23] proposes an interleaved 2D-3D backbone with pre-trained 2D weights, but is limited to object detection. Fit3D [73], which lifts 2D semantic features into 3D Gaussian representations, injects 3D awareness when training 2D foundation models—a complementary approach to ours.

2.3. 3D Open-Vocabulary Segmentation

Our work is closely related to recent efforts in distilling 2D semantic features into 3D representations for open-vocabulary segmentation. These approaches often utilize neural rendering techniques, such as NeRF [37] and Gaussian Splatting [24], to aggregate multi-view information. For instance, Semantic NeRF [78] and Panoptic Lifting [54] embed 2D semantics into 3D volumes, enabling dense scene understanding.

More recent works, such as LeRF [25], Distilled Feature Fields [53], NeRF-SOS [15], and Neural Feature Fusion Fields [60], further distill features from strong 2D models like LSeg [30] and DINO [7] into view-consistent 3D representations. Featured 3DGS [80] extends this paradigm to the Gaussian Splatting framework, enabling efficient distillation of 2D pre-trained models into 3D point-based representations.

While prior methods have demonstrated strong performance in 3D open-vocabulary segmentation, they typically depend on posed multi-view images and scene-specific optimization, which constrains their applicability in real-world settings. In contrast, our approach eliminates the need for pose supervision by directly distilling 2D features into point maps, enabling broader generalization across diverse and unstructured environments.

Similarly, LSM [16] jointly estimates geometry, appearance, and semantics in a single feed-forward pass and is capable of synthesizing diverse label maps. However, it employs a frozen language segmentation backbone and restricts input to only two images due to its reliance on point transformer [65].

3. A Novel Task: *Map and Locate*

Task Setting In this novel task, termed *Map and Locate*, the model receives multiview inputs and a set of semantic labels to reconstruct a 3D scene and localize target objects based on text prompts. This task extends beyond independent depth estimation for each image, requiring the model to infer relative camera poses across views and classify the semantic category of each predicted 3D point.

The task is defined as follows: given n input images ($n \geq 2$) and a set of grounding queries $\mathcal{L} = \{0, \ldots, L-1\}$, the goal is to map each pixel i to a pair $(X_i, l_i) \in \mathbb{R}^3 \times \mathcal{L}$, where $X_i = (x_i, y_i, z_i)$ represents the 3D coordinates of the point corresponding to pixel i, and l_i denotes its semantic class. For an image I of resolution $W \times H$, this establishes a one-to-one mapping between pixels and 3D scene points with semantic labels, i.e., $I_{i,j} \leftrightarrow (X_{i,j}, l_{i,j})$, for all $(i, j) \in \{1, \ldots, W\} \times \{1, \ldots, H\}$. We assume each camera ray intersects only a single 3D point, excluding cases like translucent surfaces. Ambiguous or out-of-class pixels are assigned a void label in the annotations.

For implementation, we adopt MaskCLIP [12] enhanced with FeatUp [18], combined with the MASt3R [29] pipeline as our baseline method. MaskCLIP and MASt3R act as teacher models for SAB3R , guiding the distillation process to achieve both 3D reconstruction and open-vocabulary semantic segmentation.

Data Curation Our data is sourced from ScanNet [10], a large-scale indoor scene dataset that provides RGB-D sequences, camera poses, and semantic and instance-level annotations. From the validation split, we curate a subset of 24 diverse scenes, selected based on their unique object layouts and camera trajectories. For each scene, we create image groups containing 2, 3, or 4 views, ensuring that each image overlaps with at least one other in the group. This overlap guarantees shared visual context, enabling robust evaluation of 3D reconstruction and localization tasks. To balance evaluation time and dataset diversity, we limit our selection to 24 scenes, which already requires approximately 10 hours for the evaluation to complete.

For semantic classification, we map ground-truth annotations to the widely used NYU40 class taxonomy [39]. The curated dataset includes a wide range of objects with both semantic and instance-level annotations. Each image group is paired with its corresponding RGB images, depth maps, camera poses (intrinsics and extrinsics), and semantic and instance labels. Detailed data statistics, example image groups, and the full data curation process, including selection criteria and preprocessing steps, are provided in the supplementary materials.

Evaluation Metrics For the *Map and Locate* task, we evaluate model performance using several key metrics, and in all metrics, higher values consistently indicate better performance. Additionally, before evaluating these metrics, models are required to compute pair (X, l) for every pixel in each image, using only the image inputs without any ground truth data, such as intrinsic or extrinsic matrices, then use one ground truth image's depth and pose for scaling and alignment to the ground truth coordinates.

mIoU (mean intersection over union) quantifies the overlap between predicted and ground truth points, calculated as the ratio of correctly predicted points to the union of predicted and ground truth points. This metric provides an overall measure of segmentation accuracy. In our task, we compute the mIoU by finding the nearest predicted point for each ground truth point and using its label to evaluate against the ground truth labels.

Acc (accuracy) is defined as the proportion of correctly predicted points relative to the total ground truth points, indicating the model's effectiveness in assigning correct semantic classes to 3D points. In our setting, similar to mIoU, we calculate Acc using the same approach.

mComp (Mean Completeness) measures how comprehensively the predicted points cover the ground truth point cloud. After aligning the predicted points with the ground truth pose, we compute the average distance from each predicted point to its nearest neighbor in the ground truth, offering a general sense of the reconstruction's completeness. For our task, we filter points based on each test label in both the ground truth and the predictions, then calculate the mComp metric accordingly.

mdComp (Median Completeness) is similar to mean completeness but calculates the median of nearest-neighbor distances instead. This approach reduces the impact of outliers, providing a more stable indication of coverage consistency across samples.

4. Method

In this section, we present SAB3R , a simple baseline method that distills dense 2D semantic features from foundation models into a 3D reconstruction framework. Building on a base 3D reconstruction model, we transfer knowledge from 2D foundation features—enhanced via FeatUp [18]—to integrate semantic understanding into the 3D domain. Our objective is to unify 2D and 3D representations within a shared backbone, enabling joint 3D reconstruction and open-vocabulary semantic segmentation.

To facilitate understanding, this section is organized as follows: Sec. 4.1 reviews the core 3D reconstruction backbone, Sec. 4.2 details the distillation process of 2D semantic features, and Sec. 4.3 outlines how additional features can be incorporated to further enrich the model's capabilities.

4.1. Foundational Components

DUSt3R [64] is a recent method that addresses a range of 3D tasks using unposed images as input, including camera calibration, depth estimation, pixel correspondence, camera pose estimation, and dense 3D reconstruction. It uses a transformer-based network to generate *local* 3D reconstructions from two input images, producing dense 3D point clouds $X^{1,1}$ and $X^{2,1}$, referred to as *pointmaps*.

A pointmap $X^{a,b} \in \mathbb{R}^{H\times W\times 3}$ represents a 2D-to-3D mapping from each pixel $i = (u, v)$ in image I^a to its corresponding 3D point $X^{a,b}_{u,v} \in \mathbb{R}^3$ in the coordinate system of camera C^b. By jointly regressing two pointmaps, $X^{1,1}$ and $X^{2,1}$, expressed in the coordinate system of camera C^1, DUSt3R simultaneously performs calibration and 3D reconstruction. For multiple images, a global alignment step merges all pointmaps into a unified coordinate system.

Images are encoded in a Siamese manner using a ViT [13], producing representations H^1 and H^2:

$$H^1 = \text{Encoder}(I^1), \quad H^2 = \text{Encoder}(I^2).$$

Two intertwined decoders process these representations, exchanging information via cross-attention to capture spatial

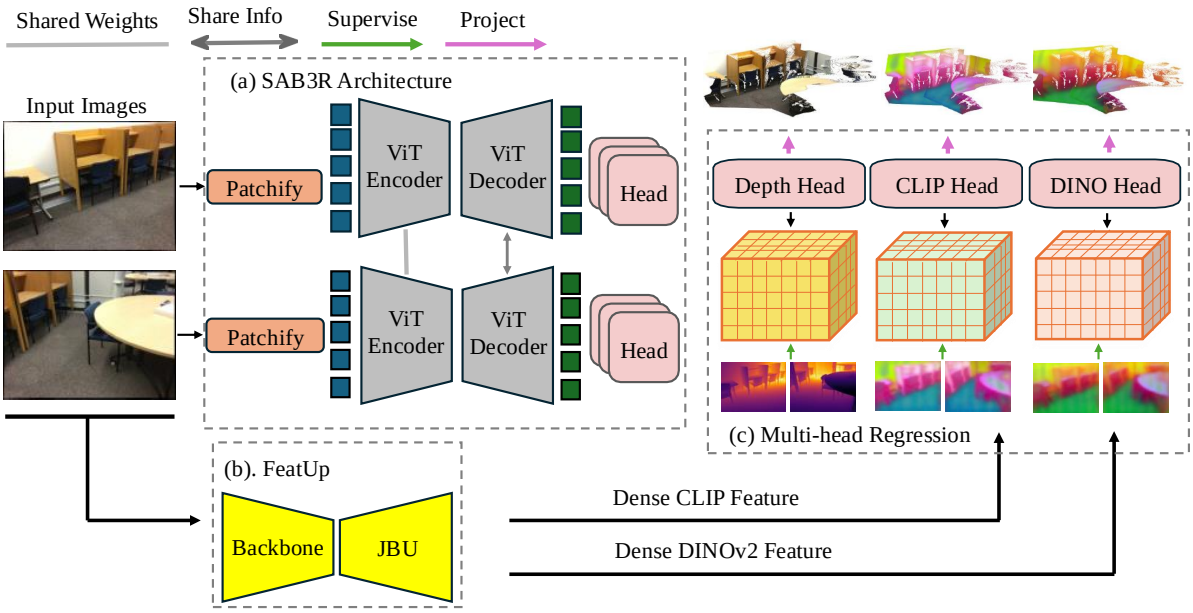

Figure 3. **Methods Architecture.** We distill dense features from CLIP and DINO into the MASt3R framework, enriching it with 2D semantic understanding. Each encoder-decoder pair operates on multi-view images, sharing weights and exchanging information to ensure consistent feature extraction across views. The model simultaneously generates depth, dense DINOv2, and dense CLIP features, which are then used for multi-view 3D reconstruction and semantic segmentation. This architecture enables SAB3R to seamlessly integrate 2D and 3D representations, achieving both geometric and semantic comprehension in a unified model.

relationships and global 3D geometry. The enhanced representations are denoted H'^1 and H'^2:

$$H'^1, H'^2 = \text{Decoder}(H^1, H^2).$$

Finally, prediction heads regress the pointmaps and confidence maps:

$$X^{1,1}, C^1 = \text{Head}^1_{3D}([H^1, H'^1]), \quad (1)$$
$$X^{2,1}, C^2 = \text{Head}^2_{3D}([H^2, H'^2]). \quad (2)$$

4.2. Distilling 2D Semantic Features

To integrate 2D semantic information into the model while retaining its 3D capabilities, we design a multitask framework that prevents catastrophic forgetting. This framework enables the model to simultaneously learn both 2D and 3D features. We adopt the MASt3R [29] architecture, which consists of a ViT-Large encoder, a ViT-Base decoder, and DPT heads. To distill dense 2D features, we introduce new heads to regress features from DINO [41] and CLIP [43].

Following DUSt3R [64] and MASt3R [28], the new heads leverage either a DPT architecture or a simpler MLP structure. The DPT design is particularly effective for dense prediction tasks like depth estimation and semantic feature extraction. In addition to the depth and descriptor heads ($\text{Head}^{1,2}_{3D}$ and $\text{Head}^{1,2}_{desc}$), we introduce two new heads, $\text{Head}^{1,2}_{\text{2D feature}}$, for distilling 2D features:

$$S^1 = \text{Head}^1_{\text{2D feature}}([H^1, H'^1]), \quad (3)$$
$$S^2 = \text{Head}^2_{\text{2D feature}}([H^2, H'^2]). \quad (4)$$

Here, H^1 and H^2 are embeddings from the encoder, and H'^1, H'^2 are enhanced representations from the decoder. The concatenation $[H, H']$ combines multi-scale features from each view.

To preserve depth estimation capabilities, we retain the regression loss $\mathcal{L}_{conf}$ from DUSt3R and the matching loss $\mathcal{L}_{match}$ from MASt3R. Additionally, we introduce a regression loss for the 2D features, guiding the model to learn semantic information:

$$\mathcal{L}_{\text{2D}} = \left\| S^v - \hat{S}^v \right\|, \quad v \in \{1, 2\}, \quad (5)$$

where $\hat{S}^v$ is the target 2D feature extracted from foundation models for the corresponding view v. Dense pixel features from FeatUp [18] are used as supervision.

The total loss combines all components, weighted by hyper-parameters β and γ:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{conf} + \beta \mathcal{L}_{match} + \gamma \mathcal{L}_{\text{2D}}. \quad (6)$$

4.3. Incorporating Additional Features

Our distillation pipeline is designed to flexibly incorporate multiple 2D features into the 3D foundation model, enhancing its capabilities. For each additional feature, we add a

dedicated head and regression loss, resulting in an updated training objective:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{conf} + \beta\mathcal{L}_{match} + \gamma_1\mathcal{L}_{2D_1} + \gamma_2\mathcal{L}_{2D_2}. \quad (7)$$

Here, $\mathcal{L}_{2D_1}$ and $\mathcal{L}_{2D_2}$ are regression losses for individual 2D features, with γ_1 and γ_2 controlling their contributions. MaskCLIP and DINOv2 features are integrated into the 3D backbone through this framework, with dedicated heads for each feature.

5. Experiments

In this section, we showcase the effectiveness of our simple baseline SAB3R for distilling 2D foundation models into a 3D reconstruction model. The section is organized into five parts. In Sec.5.1, we provide details of our implementation for SAB3R . Sec.5.2 analyzes how SAB3R retains 3D performance compared to the teacher models. In Sec.5.3, we demonstrate our method's zero-shot semantic segmentation performance, achieving results comparable to the teacher models. Finally, in Sec.5.4, we presents results and analysis for the novel task, *Map and Locate*.

5.1. Implementation Details

We fine-tune our model based on pre-trained MASt3R [29] with datasets from DUSt3R [64] and MASt3R [29], including Habitat [58], ScanNet++ [71], ARKitScenes [3], Co3Dv2 [46], and BlenderMVS [70]. Data preprocessing adheres to the guidelines of each dataset. To avoid the impracticality of storing dense 2D VFM features locally, which would require over 60 TB of storage, we leverage FeatUp to dynamically generate these features during training. Additional details on the datasets and preprocessing steps are provided in the supplementary materials.

Training We adopt MASt3R [29] as the base 3D foundation model. During training, we unfreeze the encoder to improve its ability to extract semantically meaningful 2D features while preserving depth estimation accuracy. For distillation using only MaskCLIP features, we set the loss weights to $\beta = 0.75$ and $\gamma = 20$. When distilling both MaskCLIP and DINOv2 features, we modify the weights to $\beta = 0.75$, $\gamma_1 = 20$, and $\gamma_2 = 4$. Based on our empirical observations, these hyperparameters are highly sensitive—small deviations can result in modality collapse.

5.2. Zero-Shot 3D Tasks

Monocular Depth Estimation We benchmark SAB3R on both an indoor dataset, NYUv2 [39], and an outdoor dataset, KITTI [19], comparing its performance to state-of-the-art methods in Tab. 1. For monocular depth evaluation, we use two commonly applied metrics following DUSt3R [64] and recent studies [4, 55].

Methods	Train	NYUD-v2 (Indoor)		KITTI (Outdoor)	
		Rel↓	$\delta_{1.25}$ ↑	Rel↓	$\delta_{1.25}$ ↑
DPT-BEiT[45]	D	**5.40**	**96.54**	9.45	89.27
NeWCRFs[72]	D	6.22	95.58	**5.43**	**91.54**
Monodepth2 [20]	SS	16.19	74.50	11.42	86.90
SC-SfM-Learners [5]	SS	13.79	79.57	11.83	86.61
SC-DepthV3 [57]	SS	**12.34**	**84.80**	11.79	86.39
MonoViT [77]	SS	-	-	**9.92**	**90.01**
RobustMIX [40]	T	11.77	90.45	18.25	76.95
SlowTv [55]	T	11.59	87.23	(6.84)	(56.17)
DUSt3R 224-NoCroCo	T	14.51	81.06	20.10	71.21
DUSt3R 224	T	10.28	88.92	16.97	77.89
DUSt3R 512	T	**6.51**	**94.09**	**12.02**	**83.43**
MASt3R	T	8.17	92.59	**8.28**	**93.27**
SAB3R (C)	T	7.80	92.67	11.63	86.74
SAB3R (CD)	T	**7.67**	**92.82**	12.53	83.51

Table 1. **Monocular depth estimation on NYU-v2 and KITTI datasets.** D = Supervised, SS = Self-supervised, T = Transfer (zero-shot). (Parentheses) refers to training on the same set. SAB3R (C) represents our model distilled with CLIP features, while SAB3R (CD) builds upon this by integrating both CLIP and DINO features during distillation. This notation is used consistently throughout the paper.

Methods	RRA@15↑	RTA@15↑	mAA(30)↑
Colmap+SG [11, 49]	36.1	27.3	25.3
PixSfM [33]	33.7	32.9	30.1
RelPose [75]	57.1	-	-
PosReg [62]	53.2	49.1	45.0
PoseDiff [62]	80.5	79.8	66.5
RelPose++ [32]	(85.5)	-	-
RayDiff [76]	(93.3)	-	-
DUSt3R-GA [64]	**96.2**	86.8	76.7
DUSt3R [64]	94.3	88.4	77.2
MASt3R	94.2	**88.6**	**81.1**
SAB3R (C)	92.6	87.3	79.7
SAB3R (CD)	92.9	87.8	80.3

Table 2. **Multi-view pose regression on the CO3Dv2 [46] dataset using 10 random frames.** Results in parentheses denote methods evaluated on 8 views, as they do not report results for the 10-view setup. We distinguish multi-view and pairwise methods for clarity.

As shown in Tab. 1, SAB3R demonstrates strong adaptability to both indoor and outdoor environments. Distilling dense features from MaskCLIP or DINOv2 into the MASt3R backbone does not degrade the model's performance or induce catastrophic forgetting for indoor setting. Therefore, SAB3R is still capable of making accurate depth prediction. Interestingly, SAB3R trained with MaskCLIP, or with both MaskCLIP and DINOv2, outperforms the base model MASt3R on the NYUv2 indoor dataset [39]. However, our approach performs less effectively in outdoor scenarios, likely due to the indoor-focused nature of our training data.

Model	Params	FLOPs	Sparse View = 2				Sparse View = 3				Sparse View = 4			
			mIoU	**Acc.**	**mComp.**	**mdComp.**	**mIoU**	**Acc.**	**mComp.**	**mdComp.**	**mIoU**	**Acc.**	**mComp.**	**mdComp.**
Baseline	838M	248G	4.57	18.10	0.64	0.67	6.03	21.26	0.68	0.71	5.12	19.31	0.68	0.70
LSM [16]	1B	> 592G	**21.40**	42.34	0.72	**0.80**	-	-	-	-	-	-	-	-
SAB3R (C)	729M	218G	17.26	41.11	**0.73**	0.75	22.83	**53.19**	**0.78**	**0.81**	19.92	**48.07**	**0.77**	**0.80**
SAB3R (CD)	729M	218G	17.50	**42.72**	**0.73**	0.76	**22.94**	52.86	0.77	0.80	**20.31**	46.26	0.75	0.78

Table 3. Performance comparison across different sparse view configurations (2, 3, and 4 views) using mIoU, Accuracy, Mean Completeness, and Median Completeness. Params and FLOPs refer to the number of parameters and computational cost per frame.

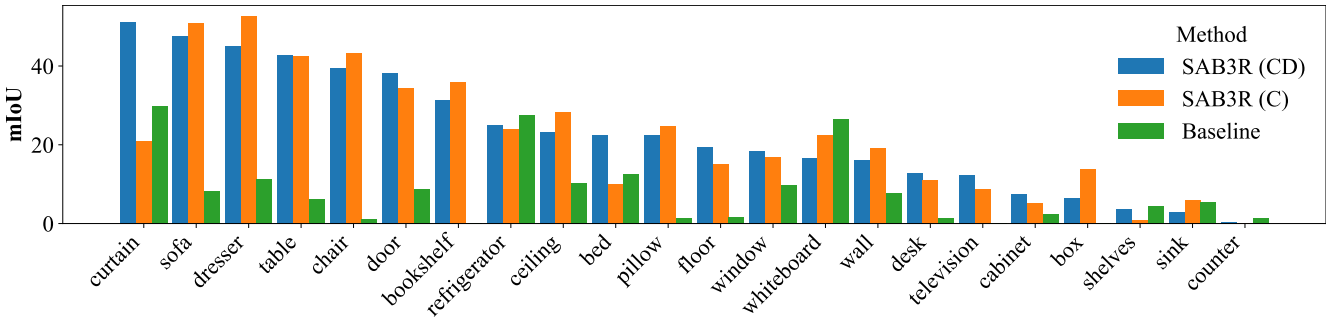

Figure 4. **mIoU Analysis on Frequently Occurring Objects Across Three Methods (Sparse View = 3).** This plot compares mIoU values for frequently appearing objects, illustrating performance differences between our methods and the pipeline approaches and providing insights into the superior results achieved by our methods.

Relative Camera Pose Next, we evaluate for the task of relative pose estimation on the CO3Dv2 [46] dataset. CO3Dv2 contains 6 million frames extracted from approximately 37k videos, covering 51 MS-COCO categories.

We compare our method's Relative Camera Pose results with popular approaches like RelPose [75], RelPose++ [32], PoseReg and PoseDiff [62], RayDiff [76], DUSt3R [64] and MASt3R [29] in Tab. 2. Our experiments show that our method performs comparably to the original MASt3R [29], indicating that catastrophic forgetting is not an issue. These results reinforce that SAB3R retains strong relative camera pose capabilities and can reliably estimate camera poses from unposed images. However, in both 3D tasks, incorporating DINO features does not improve the model's 3D reasoning capabilities.

5.3. Zero-Shot Open Vocabulary Tasks

Zero-Shot Transfer to Semantic Segmentation We evaluate the semantic features learned by SAB3R through zero-shot semantic segmentation on two standard benchmarks: Pascal VOC [14] and ADE20K [79]. As shown in Table 4, we follow the evaluation protocol of SAM-CLIP [61], with the key distinction that SAB3R produces dense, pixel-level predictions. Notably, SAB3R outperforms SAM-CLIP on the more challenging ADE20K dataset, which includes 150 semantic categories. While it does not surpass SAM-CLIP on Pascal VOC, it achieves competitive results and exceeds the performance of the teacher model, FeatUp-upsampled MaskCLIP [12]. We attribute these gains primarily to improved segmentation of large, structurally coherent objects (e.g., curtain, floor, desk). This observation aligns with findings from LeRF [25], which suggest that models with 3D reasoning capabilities tend to yield stronger semantic segmentation performance. Additional qualitative results, including PCA visualizations of the learned 2D feature space, are included in the supplementary material.

Model	Arch	VOC↑	ADE20k↑
GroupViT [66]	ViT-S	52.3	-
ViewCo [47]	ViT-S	52.4	-
ViL-Seg [34]	ViT-B	37.3	-
OVS [67]	ViT-B	53.8	-
CLIPpy [44]	ViT-B	52.2	13.5
TCL [8]	ViT-B	51.2	14.9
SegCLIP [35]	ViT-B	52.6	8.7
SAM-CLIP [61]	ViT-B	**60.6**	17.1
FeatUp (MaskCLIP)	-	51.2	14.3
SAB3R (C)	ViT-B	55.4	18.3
SAB3R (CD)	ViT-B	56.4	**19.0**

Table 4. **Zero-shot Semantic Segmentation Comparison.** Performance comparison of zero-shot semantic segmentation with recent state-of-the-art methods. **Note:** Results for SAB3R are based solely on the CLIP-head output.

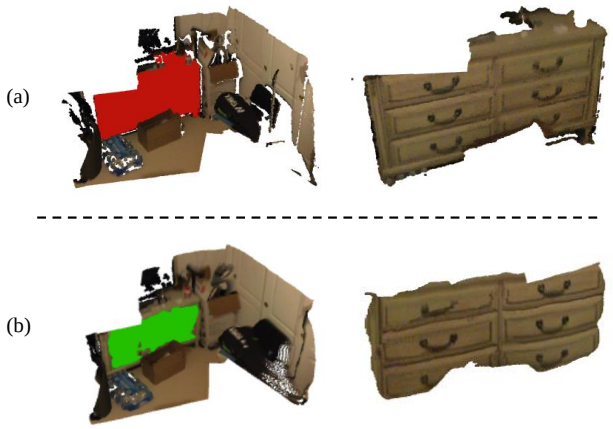

Figure 5. **Qualitative Example of *Map and Locate*.** This figure illustrates an example from our benchmark. In (a), the ground truth annotation for the scene is highlighted in red, with the dresser segmented from the rest of the scene on the left. In (b), the predictions from SAB3R are highlighted in green, and the predicted dresser is similarly segmented on the right. These segmented results are subsequently used to compute evaluation metrics.

5.4. A Novel Task - *Map and Locate*

We use MASt3R [29] and FeatUp [18] as teacher models and adopt them as our primary baselines. Additionally, we report the performance of LSM [16] on this new task for comparison. We present the results in Table 3. Our method, SAB3R , consistently outperforms the baseline across all sparse view settings (views = 2, 3, 4) and evaluation metrics, demonstrating strong performance on the *Map and Locate* task. Notably, SAB3R achieves a 3× speedup in inference compared to the baseline, as it operates in an end-to-end manner, whereas the baseline relies on a two-stage pipeline involving separate models for reconstruction and segmentation. In terms of semantic quality, measured by mIoU and accuracy, SAB3R surpasses the baseline by a substantial margin, highlighting its effectiveness in jointly performing 3D reconstruction and open-vocabulary segmentation without pose supervision. For completion metrics, which assess the geometric fidelity of reconstructed semantic objects, SAB3R also consistently outperforms the baseline under all sparse view configurations. Interestingly, we observe no clear correlation between the number of input views and overall performance. We hypothesize that additional views improve results when they focus on overlapping regions or specific objects, enabling the model to better infer structure and semantics. However, performance may degrade when added views are sparsely distributed across unrelated parts of the scene, leading to reduced overlap and more fragmented supervision during reconstruction.

In Fig. 4, our model demonstrates significant improvements over the baseline in large furniture categories such as sofas, dressers, tables, and chairs. It also successfully recognizes items like bookshelves and televisions, which the baseline fails to detect. Across most categories, our model achieves substantially higher scores, showcasing its strong semantic understanding and superior 3D reconstruction capabilities. Furthermore, it exhibits the ability to identify smaller objects and less common items, underscoring its versatility and robustness. In Fig. 5, we showcase an example of mapping and locating a *dresser* across two images. In part (b) of the qualitative example, the predicted segmentation demonstrates remarkable accuracy compared to the ground truth shown in part (a), highlighting the effectiveness of our model SAB3R .

LSM [16] demonstrates strong performance when operating on two input views, benefiting from its ability to jointly estimate geometry, semantics, and appearance in a single feed-forward pass. However, extending LSM to more than two views is non-trivial, as its point transformer architecture and Gaussian fusion strategy are designed specifically for dual-view inputs. Moreover, while LSM employs a powerful frozen segmentation backbone that contributes to its accuracy, this comes at the cost of significantly higher computational complexity—both in terms of FLOPs and parameter count—compared to our more lightweight and efficient baseline model SAB3R .

6. Conclusion

Our experiments validate the central insight of this work: 3D open-vocabulary segmentation and 3D reconstruction can be effectively unified through the proposed *Map and Locate* task. Unlike existing approaches that rely on pre-scanned point clouds or posed RGB-D sequences, our formulation accepts unposed video as input—offering a more realistic and scalable setting for embodied agents.

The *Map and Locate* benchmark demonstrates how spatial mapping and semantic understanding can be performed simultaneously, requiring models to reason over both 3D structure and 2D semantics. We present SAB3R , a simple yet effective baseline that distills 2D foundation models into a unified model capable of predicting 3D point maps along with dense CLIP and DINOv2 features in a single forward pass. Despite its simplicity, SAB3R performs competitively across both reconstruction and segmentation metrics, while remaining significantly more efficient than multi-stage baselines.

Overall, our findings demonstrate the feasibility of unifying recognition, reconstruction, and reorganization within a single model, offering a more efficient and scalable approach to 3D scene understanding. We hope the *Map and Locate* task serves as a testbed for advancing real-world embodied perception research.

References

[1] Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas Guibas. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 422–440. Springer, 2020. 1

[2] Sameer Agarwal, Yasutaka Furukawa, Noah Snavely, Ian Simon, Brian Curless, Steven M Seitz, and Richard Szeliski. Building rome in a day. *Communications of the ACM*, 54 (10):105–112, 2011. 3

[3] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, and Elad Shulman. ARKitscenes - a diverse real-world dataset for 3d indoor scene understanding using mobile RGB-d data. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021. 6, 1

[4] Jia-Wang Bian, Huangying Zhan, Naiyan Wang, Tat-Jun Chin, Chunhua Shen, and Ian Reid. Auto-rectify network for unsupervised indoor depth estimation, 2021. 6

[5] Jia-Wang Bian, Huangying Zhan, Naiyan Wang, Zhichao Li, Le Zhang, Chunhua Shen, Ming-Ming Cheng, and Ian Reid. Unsupervised scale-consistent depth learning from video. *International Journal of Computer Vision (IJCV)*, 2021. 6

[6] Cesar Cadena, Luca Carlone, Henry Carrillo, Yasir Latif, Davide Scaramuzza, Jose Neira, Ian Reid, and John J. Leonard. Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. *IEEE Transactions on Robotics*, 32(6):1309–1332, 2016. 3

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers, 2021. 3

[8] Junbum Cha, Jonghwan Mun, and Byungseok Roh. Learning to generate text-grounded mask for open-world semantic segmentation from only image-text pairs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11165–11174, 2023. 7

[9] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *European conference on computer vision*, pages 202–221. Springer, 2020. 1

[10] 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. 4, 2

[11] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised Interest Point Detection and Description. In *CVPR*, 2018. 6

[12] Xiaoyi Dong, Jianmin Bao, Yinglin Zheng, Ting Zhang, Dongdong Chen, Hao Yang, Ming Zeng, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, and Nenghai Yu. Maskclip: Masked self-distillation advances contrastive language-image pretraining, 2023. 4, 7, 1

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 4

[14] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, 2015. 7

[15] Zhiwen Fan, Peihao Wang, Yifan Jiang, Xinyu Gong, Dejia Xu, and Zhangyang Wang. Nerf-sos: Any-view self-supervised object segmentation on complex scenes, 2022. 3

[16] Zhiwen Fan, Jian Zhang, Wenyan Cong, Peihao Wang, Renjie Li, Kairun Wen, Shijie Zhou, Achuta Kadambi, Zhangyang Wang, Danfei Xu, Boris Ivanovic, Marco Pavone, and Yue Wang. Large spatial model: End-to-end unposed images to semantic 3d, 2024. 3, 7, 8

[17] Stephanie Fu, Mark Hamilton, Laura Brandt, Axel Feldman, Zhoutong Zhang, and William T. Freeman. Featup: A model-agnostic framework for features at any resolution, 2024. 1

[18] Stephanie Fu, Mark Hamilton, Laura E. Brandt, Axel Feldmann, Zhoutong Zhang, and William T. Freeman. Featup: A model-agnostic framework for features at any resolution. In *The Twelfth International Conference on Learning Representations*, 2024. 4, 5, 8, 1

[19] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, 2013. 6, 1

[20] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J. Brostow. Digging into self-supervised monocular depth prediction. 2019. 6

[21] Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models, 2023. 3

[22] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. OpenCLIP, 2021. 1

[23] Ayush Jain, Pushkal Katara, Nikolaos Gkanatsios, Adam W. Harley, Gabriel Sarch, Kriti Aggarwal, Vishrav Chaudhary, and Katerina Fragkiadaki. Odin: A single model for 2d and 3d segmentation, 2024. 3

[24] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42 (4), 2023. 1, 3

[25] Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lerf: Language embedded radiance fields, 2023. 3, 7

[26] Michael Landy and J. Anthony Movshon. *The Plenoptic Function and the Elements of Early Vision*, pages 3–20. 1991. 1

[27] Donghwan Lee, Soohyun Ryu, Suyong Yeon, Yonghan Lee, Deokhwa Kim, Cheolho Han, Yohann Cabon, Philippe Weinzaepfel, Nicolas Guérin, Gabriela Csurka, and Martin Humenberger. Large-scale localization datasets in crowded indoor spaces, 2021. 1

[28] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r, 2024. 5

[29] Vincent Leroy, Yohann Cabon, and Jerome Revaud. Grounding image matching in 3d with mast3r, 2024. 3, 4, 5, 6, 7, 8, 1

[30] Boyi Li, Kilian Q. Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation, 2022. 3

[31] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

[32] Amy Lin, Jason Y. Zhang, and Shubham Tulsiani. Relpose++: Recovering 6d poses from sparse-view observations. *CoRR*, abs/2305.04926, 2023. 6, 7

[33] Philipp Lindenberger, Paul-Edouard Sarlin, Viktor Larsson, and Marc Pollefeys. Pixel-perfect structure-from-motion with featuremetric refinement. In *ICCV*, 2021. 6

[34] Quande Liu, Youpeng Wen, Jianhua Han, Chunjing Xu, Hang Xu, and Xiaodan Liang. Open-world semantic segmentation via contrasting and clustering vision-language embedding. In *European Conference on Computer Vision*, pages 275–292. Springer, 2022. 7

[35] Huaishao Luo, Junwei Bao, Youzheng Wu, Xiaodong He, and Tianrui Li. Segclip: Patch aggregation with learnable centers for open-vocabulary semantic segmentation. In *International Conference on Machine Learning*, pages 23033–23044. PMLR, 2023. 7

[36] Jitendra Malik, Pablo Arbeláez, João Carreira, Katerina Fragkiadaki, Ross Girshick, Georgia Gkioxari, Saurabh Gupta, Bharath Hariharan, Abhishek Kar, and Shubham Tulsiani. The three r's of computer vision: Recognition, reconstruction and reorganization. *Pattern Recognition Letters*, 72:4–14, 2016. Special Issue on ICPR 2014 Awarded Papers. 1

[37] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 1, 3

[38] Raul Mur-Artal, J. M. M. Montiel, and Juan D. Tardos. Orb-slam: A versatile and accurate monocular slam system. *IEEE Transactions on Robotics*, 31(5):1147–1163, 2015. 3

[39] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 4, 6, 1

[40] Jonas Ngnawe, Marianne Abemgnigni Njifon, Jonathan Heek, and Yann Dauphin. Robustmix: Improving robustness by regularizing the frequency bias of deep nets, 2024. 6

[41] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. 5, 1, 3

[42] Songyou Peng, Kyle Genova, Chiyu "Max" Jiang, Andrea Tagliasacchi, Marc Pollefeys, and Thomas Funkhouser. Openscene: 3d scene understanding with open vocabularies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 3

[43] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 5, 1, 2, 3

[44] Kanchana Ranasinghe, Brandon McKinzie, Sachin Ravi, Yinfei Yang, Alexander Toshev, and Jonathon Shlens. Perceptual grouping in contrastive vision-language models. ICCV, 2023. 7, 2

[45] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *ICCV*, 2021. 6

[46] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *International Conference on Computer Vision*, 2021. 6, 7, 1, 2

[47] Pengzhen Ren, Changlin Li, Hang Xu, Yi Zhu, Guangrun Wang, Jianzhuang Liu, Xiaojun Chang, and Xiaodan Liang. Viewco: Discovering text-supervised segmentation masks via multi-view semantic consistency. *arXiv preprint arXiv:2302.10307*, 2023. 7

[48] Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Tali Bers, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. Multitask prompted training enables zero-shot task generalization, 2022. 2

[49] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning feature matching with graph neural networks. In *CVPR*, 2020. 6

[50] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A platform for embodied ai research, 2019. 1

[51] Jonas Schult, Francis Engelmann, Alexander Hermans, Or Litany, Siyu Tang, and Bastian Leibe. Mask3D: Mask Transformer for 3D Semantic Instance Segmentation. 2023. 1

[52] Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4104–4113, 2016. 3

[53] William Shen, Ge Yang, Alan Yu, Jansen Wong, Leslie Pack Kaelbling, and Phillip Isola. Distilled feature fields enable few-shot language-guided manipulation, 2023. 3

[54] Yawar Siddiqui, Lorenzo Porzi, Samuel Rota Bulò, Norman Müller, Matthias Nießner, Angela Dai, and Peter Kontschieder. Panoptic lifting for 3d scene understanding with neural fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9043–9052, 2023. 3

[55] Jaime Spencer, Chris Russell, Simon Hadfield, and Richard Bowden. Kick back relax: Learning to reconstruct the world by watching slowtv, 2023. 6

[56] Julian Straub, Daniel DeTone, Tianwei Shen, Nan Yang, Chris Sweeney, and Richard Newcombe. Efm3d: A benchmark for measuring progress towards 3d egocentric foundation models, 2024. 3

[57] Libo Sun, Jia-Wang Bian, Huangying Zhan, Wei Yin, Ian Reid, and Chunhua Shen. Sc-depthv3: Robust self-supervised monocular depth estimation for dynamic scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2023. 6

[58] Andrew Szot, Alex Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Chaplot, Oleksandr Maksymets, Aaron Gokaslan, Vladimir Vondrus, Sameer Dharur, Franziska Meier, Wojciech Galuba, Angel Chang, Zsolt Kira, Vladlen Koltun, Jitendra Malik, Manolis Savva, and Dhruv Batra. Habitat 2.0: Training home assistants to rearrange their habitat. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 6

[59] Zachary Teed and Jia Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras, 2021. 3

[60] Vadim Tschernezki, Iro Laina, Diane Larlus, and Andrea Vedaldi. Neural feature fusion fields: 3d distillation of self-supervised 2d image representations, 2022. 3

[61] Haoxiang Wang, Pavan Kumar Anasosalu Vasu, Fartash Faghri, Raviteja Vemulapalli, Mehrdad Farajtabar, Sachin Mehta, Mohammad Rastegari, Oncel Tuzel, and Hadi Pouransari. Sam-clip: Merging vision foundation models towards semantic and spatial understanding, 2024. 7, 2

[62] Jianyuan Wang, Christian Rupprecht, and David Novotny. PoseDiffusion: Solving pose estimation via diffusion-aided bundle adjustment. 2023. 6, 7

[63] Jianyuan Wang, Nikita Karaev, Christian Rupprecht, and David Novotny. Vggsfm: Visual geometry grounded deep structure from motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21686–21697, 2024. 3

[64] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy, 2023. 3, 4, 5, 6, 7

[65] Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer v3: Simpler, faster, stronger. In *CVPR*, 2024. 3

[66] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18134–18144, 2022. 7

[67] Jilan Xu, Junlin Hou, Yuejie Zhang, Rui Feng, Yi Wang, Yu Qiao, and Weidi Xie. Learning open-vocabulary semantic segmentation models from natural language supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2935–2944, 2023. 7

[68] Jianing Yang, Xuweiyi Chen, Shengyi Qian, Nikhil Madaan, Madhavan Iyengar, David F Fouhey, and Joyce Chai. Llm-grounder: Open-vocabulary 3d visual grounding with large language model as an agent. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7694–7701. IEEE, 2024. 1

[69] Zhengyuan Yang, Songyang Zhang, Liwei Wang, and Jiebo Luo. Sat: 2d semantics assisted training for 3d visual grounding, 2021. 3

[70] Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. *Computer Vision and Pattern Recognition (CVPR)*, 2020. 6, 1

[71] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes, 2023. 6, 1

[72] Weihao Yuan, Xiaodong Gu, Zuozhuo Dai, Siyu Zhu, and Ping Tan. New crfs: Neural window fully-connected crfs for monocular depth estimation, 2022. 6

[73] Yuanwen Yue, Anurag Das, Francis Engelmann, Siyu Tang, and Jan Eric Lenssen. Improving 2d feature representations by 3d-aware fine-tuning, 2024. 3

[74] Amir R. Zamir, Tilman Wekel, Pulkit Agrawal, Colin Wei, Jitendra Malik, and Silvio Savarese. *Generic 3D Representation via Pose Estimation and Matching*, page 535–553. Springer International Publishing, 2016. 1

[75] Jason Y. Zhang, Deva Ramanan, and Shubham Tulsiani. Relpose: Predicting probabilistic relative rotation for single objects in the wild. In *ECCV*, 2022. 6, 7

[76] Jason Y Zhang, Amy Lin, Moneish Kumar, Tzu-Hsuan Yang, Deva Ramanan, and Shubham Tulsiani. Cameras as rays: Pose estimation via ray diffusion. In *International Conference on Learning Representations (ICLR)*, 2024. 6, 7

[77] Chaoqiang Zhao, Youmin Zhang, Matteo Poggi, Fabio Tosi, Xianda Guo, Zheng Zhu, Guan Huang, Yang Tang, and Stefano Mattoccia. Monovit: Self-supervised monocular depth estimation with a vision transformer. In *International Conference on 3D Vision*, 2022. 6

[78] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew J. Davison. In-place scene labelling and understanding with implicit scene representation. In *ICCV*, 2021. 3

[79] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127(3):302–321, 2019. 7

[80] Shijie Zhou, Haoran Chang, Sicheng Jiang, Zhiwen Fan, Zehao Zhu, Dejia Xu, Pradyumna Chari, Suya You, Zhangyang Wang, and Achuta Kadambi. Feature 3dgs: Supercharging 3d gaussian splatting to enable distilled feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21676–21685, 2024. 3

[81] Ziyu Zhu, Zhuofan Zhang, Xiaojian Ma, Xuesong Niu, Yixin Chen, Baoxiong Jia, Zhidong Deng, Siyuan Huang, and Qing Li. Unifying 3d vision-language understanding via promptable queries. *ECCV*, 2024. 3

SAB3R: Semantic-Augmented Backbone in 3D Reconstruction

Supplementary Material

In Sec. A, we provide additional details about the experiments conducted in this work, including a discussion of the software used in SAB3R and a detailed breakdown of each experiment. Comprehensive analysis and visualizations of our novel task, *Map and Locate*, are provided in Sec. B, including both successful and failure cases from our experiments. Sec. C presents supplementary visualizations of the features generated by CLIP [43] and DINOv2 [41]. Finally, we discuss the limitations of our approach in Sec. D.

A. More Experiment Details

A.1. Teacher Models and Frameworks

CLIP & MaskCLIP Vision and language models are trained to generate aligned feature embeddings using a contrastive objective. The original CLIP family of models was proposed by Radford et al. [43] and included a wide variety of architectures in a private dataset of 400M image-text pairs called WIT. More recently, Ilharco et al. [22] trained several CLIP models using several architectures trained on publicly available datasets. In SAB3R , we used MaskCLIP [12], which enhances CLIP pretraining by introducing masked self-distillation. This transfers knowledge from full-image representations to masked-image predictions. This approach complements the vision-language contrastive objective by focusing on local patch representations while aligning features with indirect supervision from language. Additionally, MaskCLIP incorporates local semantic supervision into the text branch, further improving pretraining performance. We follow suggestions from FeatUp [18] that MaskCLIP [12] has better local semantic feature compare with CLIP [43].

MASt3R MASt3R [29] was trained on an extensive multi-view dataset comprising 5.3 million real-world image pairs and 1.8 million synthetic pairs. The real-world data includes diverse scenarios from ARKitScenes [3], MegaDepth [31], 3DStreetView [74], and IndoorVL [27]. The synthetic data was generated using the Habitat simulator [50], covering indoor, outdoor, and landmark environments.

Our model is finetuned on top of MASt3R, leveraging Habitat-Sim [50], ScanNet++[71], and Co3Dv2[46], ARKitScenes [3] and BlenderMVS [70].

FeatUp FeatUp [17] is a framework designed to enhance spatial resolution in deep features for tasks like segmentation and depth prediction. It addresses the loss of spatial detail caused by pooling in traditional networks using two approaches: guided upsampling with high-resolution signals in a single pass and reconstructing features at arbitrary resolutions with an implicit model. Both methods use a multi-view consistency loss inspired by NeRFs to maintain feature semantics.

FeatUp integrates seamlessly into existing pipelines, boosting resolution and performance without re-training. Experiments demonstrate its superiority over other methods in tasks such as segmentation, depth prediction, and class activation map generation. In SAB3R , we find the MaskCLIP variant of FeatUp model can also perform zero-shot semantic segmentation and we use it as our teacher model for distillation.

Table 5. **Checkpoint Details.** Information about the pre-trained checkpoints used in this work, including source and license.

Checkpoint	Source Link	License
FeatUp MaskCLIP	MaskCLIP	MIT
MASt3R	MASt3R	CC BY-NC-SA 4.0

We list the checkpoints used in SAB3R in Tab. 5, detailing the FeatUp MaskCLIP variant and MASt3R, along with their source links and license information.

A.2. Experiments Details

Monocular Depth In the main text, we benchmark SAB3R on the outdoor dataset KITTI [19] and the indoor dataset NYUv2 [39]. Here, we provide a detailed discussion of the evaluation metrics. Following DUSt3R, we use two commonly adopted metrics in monocular depth estimation:

- Absolute Relative Error (AbsRel): This measures the relative error between the ground truth depth y and the predicted depth $\hat{y}$, defined as:

$$\text{AbsRel} = \frac{|y - \hat{y}|}{y}.$$

- Prediction Threshold ($\delta_{1.25}$): This evaluates the fraction of predictions within a given threshold and is defined as:

$$\delta_{1.25} = \frac{\max\left(\frac{\hat{y}}{y}, \frac{y}{\hat{y}}\right) < 1.25}{\text{Total Predictions}}.$$

These metrics allow for comprehensive evaluation of depth prediction accuracy and robustness across different datasets.

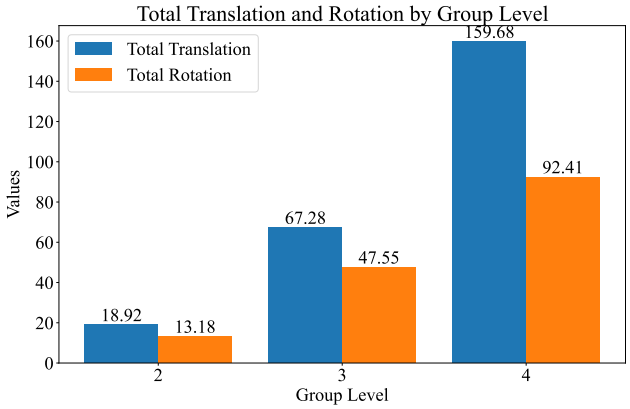

Figure 6. **Camera Distributions.** Camera translation differences and rotation differences at different group levels.

Relative Camera Pose We evaluate SAB3R on the task of relative pose estimation using the CO3Dv2 dataset [46]. To assess the relative pose error for each image pair, we report the Relative Rotation Accuracy (RRA) and Relative Translation Accuracy (RTA). For evaluation, we select a threshold $\tau = 15°$ and report RRA@15 and RTA@15, representing the percentage of image pairs where the errors in rotation and translation are below the threshold τ.

The rotation error e_{rot} and translation error e_{trans} for each image pair are computed as:

$$e_{rot} = \arccos\left(\frac{\text{trace}(\mathbf{R}^\top \hat{\mathbf{R}}) - 1}{2}\right),$$

$$e_{trans} = \arccos\left(\frac{\mathbf{t}^\top \hat{\mathbf{t}}}{\|\mathbf{t}\|\|\hat{\mathbf{t}}\|}\right),$$

where $\mathbf{R}$ and $\hat{\mathbf{R}}$ are the ground truth and predicted rotation matrices, and $\mathbf{t}$ and $\hat{\mathbf{t}}$ are the ground truth and predicted translation vectors.

We also report the mean Average Accuracy (mAA@30), defined as the area under the accuracy curve of the angular differences for $\min(\text{RRA@30}, \text{RTA@30})$. The mAA@30 is calculated as:

$$\text{mAA@30} = \frac{1}{30}\int_0^{30} \min(\text{RRA@}\theta, \text{RTA@}\theta)\, d\theta,$$

where θ represents the threshold angle in degrees.

Zero-Shot Semantic Segmentation For zero-shot semantic segmentation, we largely follow the approach outlined by Ranasinghe et al.[44], utilizing 80 prompt templates introduced by Radford et al .[43, 61]. Class names are embedded into these prompts, and text embeddings are generated using the text encoder. We then compute the cosine similarity between each text embedding and the corresponding pixel feature—extracted directly from the CLIP head. The class with the highest cosine similarity is assigned as the predicted class for each pixel.

The class predictions are subsequently resized to match the original image dimensions, and the mean Intersection over Union (mIoU) is computed for evaluation. Unlike prior methods, our approach eliminates the concept of patches. Instead, because the CLIP head directly generates per-pixel features, we can seamlessly perform top-1 matching between semantic classes and pixel features, bypassing the need for patch-based processing.

B. Additional *Map and Locate* Details

B.1. Dataset Summary

We evaluate our *Map and Locate* framework using the ScanNet dataset [10], a large-scale indoor scene dataset that provides RGB-D sequences, camera poses, semantic and instance annotations. Specifically, we select 24 scenes from the validation split, each containing diverse object layouts and camera trajectories. Across these 10 scenes, there are a total of 942 objects with semantic and instance-level ground truth annotations.

For evaluation, we construct 2 sets of image groups for each scene, where each group comprises 2, 3, or 4 images. The image selection ensures:

- Object visibility: Objects in each group are visible across multiple images to ensure reliable localization and mapping.
- Viewpoint diversity: Selected images capture varying camera viewpoints to test robustness to occlusion and perspective changes.

In total, this results in 144 image groups (2 sets per scene $\times$ 24 scenes $\times$ 3 group sizes). Each group is paired with its corresponding rgb images, depth maps, camera poses (intrinsics and extrinsic) , and semantic and instance labels, providing a comprehensive benchmark for evaluating both mapping accuracy and object localization performance.

B.2. Dataset Visualizations

We present a dataset statistics visualization in Fig. 6, showing camera translation differences and rotation differences. Translation differences are computed as the Euclidean distance between translation vectors, $d_{\text{translation}} = \|\mathbf{t}_1 - \mathbf{t}_2\|_2$, and rotation differences are calculated as the geodesic distance on $SO(3)$, $d_{\text{rotation}} = \|\mathbf{r}_\Delta\|_2$, where $\mathbf{r}_\Delta$ is the axis-angle representation of the relative rotation $\mathbf{R}_\Delta = \mathbf{R}_1^{-1}\mathbf{R}_2$. These metrics highlight the variability in camera poses across the dataset. We observe that as the number of views increases, both camera translation differences and rotation differences grow. Despite this, our results demonstrate consistent performance across all group levels, highlighting the robustness of our algorithm.

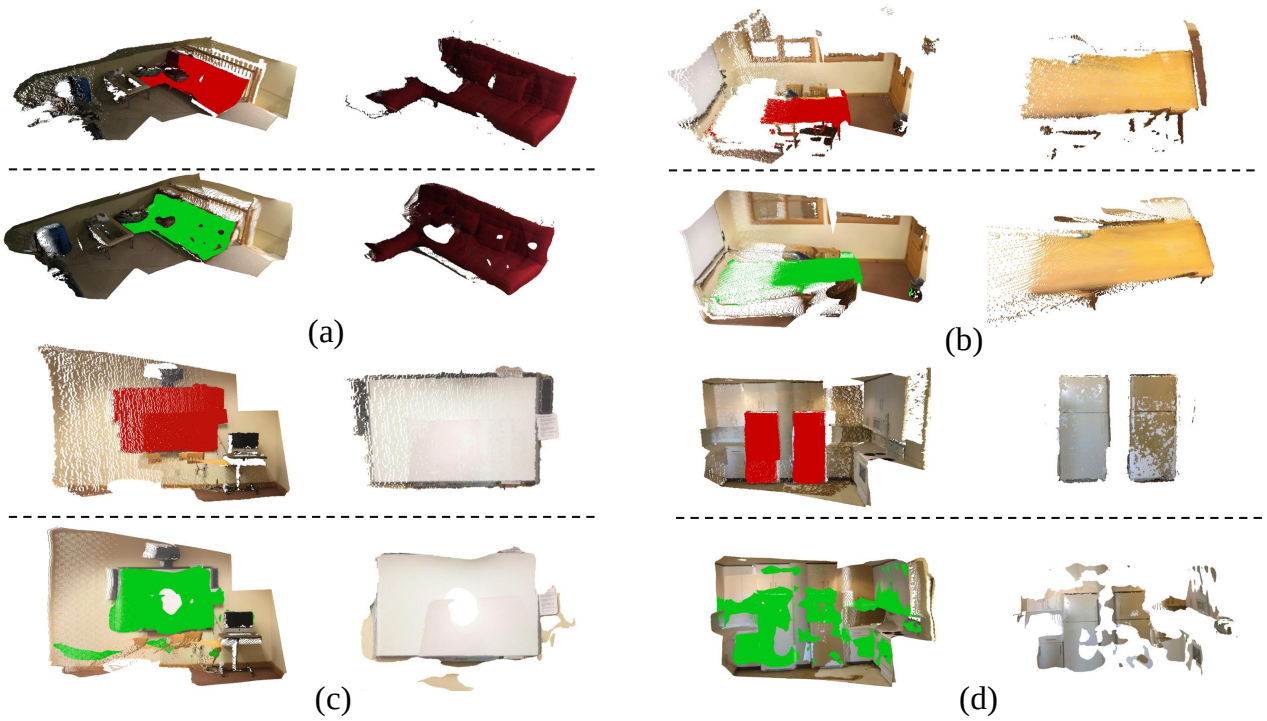

Figure 7. **Qualitative Examples of *Map and Locate* with SAB3R .** Panels (a), (b), and (c) illustrate successful examples of 3D scene reconstruction and accurate object segmentation. In each sub-group, the top row shows the ground truth, with the target objects highlighted in red, accompanied by visualizations of segmented objects for each ground truth target. The bottom row presents the predicted results, where the segmented objects are shown in green, with the extracted objects displayed on the right for clarity. Panel (d) provides an example of a failure case.

B.3. More Qualitative Examples

Fig. 7 presents additional qualitative examples demonstrating the performance of *Map and Locate* with SAB3R .

C. Additional visualization

Fig. 8 presents additional visualizations of 3D features from DINO [41] and CLIP [43]. The visualizations highlight distinct features for different objects. Predicted RGB is provided as a reference.

D. Limitations

Our study is constrained by limited computational resources, which restricted us from training the model for more epochs, potentially resulting in under-trained checkpoints. Additionally, predicting dense features significantly increases vRAM requirements, further limiting our ability to optimize the model fully. Due to these resource constraints, we were unable to use the entire pre-training dataset for fine-tuning, which may have prevented the model from achieving its best possible performance. Our novel task, *Map and Locate*, relies on the ScanNet dataset, which, despite its comprehensiveness, is primarily biased toward indoor environments. Extending this work to more diverse datasets, including outdoor or dynamic scenes, represents an interesting direction for future works.

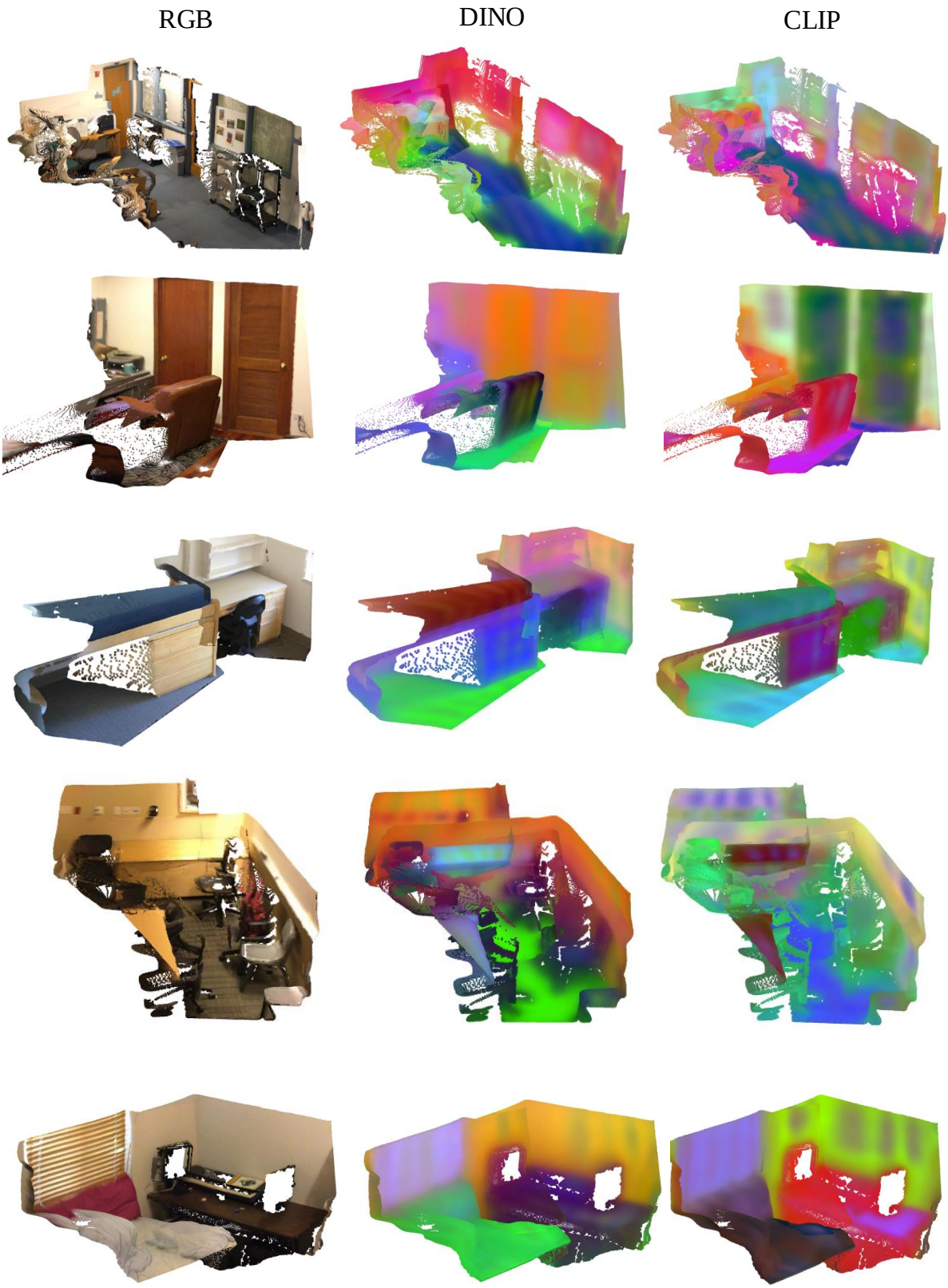

Figure 8. **3D Feature Visualizations.** Additional visualizations of 3D features are presented for DINO and CLIP, alongside the original RGB 3D point map for reference.