LIRA: Reasoning Reconstruction via Multimodal Large Language Models

Zhen Zhou[1], Tong Wang[2], Yunkai Ma[1,*], Xiao Tan[2], and Fengshui Jing[1,*]
[1]Institute of Automation, Chinese Academy of Sciences, [2]Baidu Inc.
E-mail:zhouzhen2021@ia.ac.cn, *Corresponding author.

Abstract

*Existing language instruction-guided online 3D reconstruction systems mainly rely on explicit instructions or queryable maps, showing inadequate capability to handle implicit and complex instructions. In this paper, we first introduce a reasoning reconstruction task. This task inputs an implicit instruction involving complex reasoning and an RGB-D sequence, and outputs incremental 3D reconstruction of instances that conform to the instruction. To handle this task, we propose LIRA: **L**anguage **I**nstructed **R**econstruction **A**ssistant. It leverages a multimodal large language model to actively reason about the implicit instruction and obtain instruction-relevant 2D candidate instances and their attributes. Then, candidate instances are back-projected into the incrementally reconstructed 3D geometric map, followed by instance fusion and target instance inference. In LIRA, to achieve higher instance fusion quality, we propose TIFF, a **T**ext-enhanced **I**nstance **F**usion module operating within **F**ragment bounding volume, which is learning-based and fuses multiple keyframes simultaneously. Since the evaluation system for this task is not well established, we propose a benchmark ReasonRecon comprising the largest collection of scene-instruction data samples involving implicit reasoning. Experiments demonstrate that LIRA outperforms existing methods in the reasoning reconstruction task and is capable of running in real time. Code and benchmark are available at https://github.com/zhen6618/LIRA.*

1. Introduction

Online 3D reconstruction guided by language instructions serves as a key task for embodied agents to understand environment and human intentions, enabling many applications such as navigation, manipulation and human–robot interaction. However, existing systems [15, 27, 46, 47] mainly rely on explicit instructions, such as explicitly indicating target objects or categories, to reconstruct instruction-relevant regions, while implicit instruction reasoning is more important for human intention understanding. Humans tend to directly express their demands involving implicit and complex descriptions, e.g., "*I am hungry. Find something to eat*" or "*Fixtures used for hygiene. Find a smaller one suitable for a child*", rather than providing explicit step-by-step instructions. Furthermore, humans in unfamiliar environments may not know what objects exist and thus can only give implicit instructions. Given the crucial importance of implicit instruction reasoning, this work primarily focuses on online 3D reconstruction guided by implicit instructions.

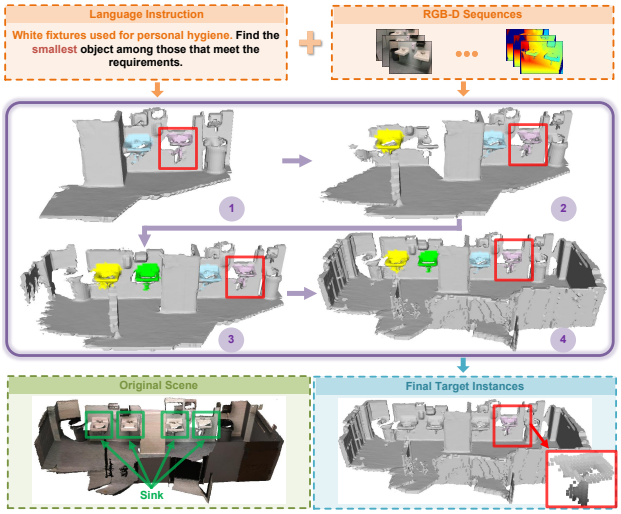

Figure 1. Online reasoning reconstruction results of LIRA. It inputs RGB-D sequences and reconstructs instruction-relevant instances and background environment. As the process is online, LIRA can halt at any time step between 1 and 4, and identify the target "*the smallest object*" within the current map. Different instruction-relevant candidate instances in the purple box are distinguished by colors. The target instance based on the current map is within the red box. We visualize double-layered mesh.

In this work, we first refer to the online 3D reconstruction task guided by implicit instructions as *reasoning reconstruction*, which requires reconstruction of instances that conform to implicit instructions involving complex reasoning, and geometric reconstruction of background environment (see Fig. 1). Although reconstruction of background environment is not the core of the task, it serves as crucial reference information for path planning by embodied

agents. Since geometric reconstruction can be accurately obtained by systems such as Simultaneous Localization and Mapping (SLAM) [12, 16, 17, 55], the main challenge in reasoning reconstruction from RGB-D sequences is how to accurately segment instruction-relevant instances from the incrementally reconstructed 3D geometry map.

Existing methods [13, 15, 27, 46] mainly rely on explicit instructions or queryable maps to achieve instruction-guided online 3D reconstruction. They first apply SLAM systems for geometric reconstruction and extract all instance features and masks. Then, instructions are matched with the pre-defined maps. To enhance comprehension, some approaches [8, 13, 22] employ large language models (LLMs) for pre-processing or post-processing. However, these methods contain much instruction-irrelevant information, and exhibit limited interaction and reasoning between target instance features and instruction features. Particularly for implicit instructions involving complex reasoning, they are more difficult to handle.

To handle the reasoning reconstruction task, we propose a **L**anguage **I**nstructed **R**econstruction **A**ssistant (LIRA), which applies a multimodal LLM (MLLM) to actively reason about implicit instructions and obtain instruction-relevant 2D candidate instances and their attributes. These candidate instances are then back-projected into the incrementally reconstructed 3D geometric map, followed by instance fusion. Existing instance fusion strategies mainly rely on image features and geometric features, and perform frame-by-frame fusion. To achieve higher-quality instance fusion, we propose TIFF, a **T**ext-enhanced **I**nstance **F**usion module operating within a **F**ragment bounding volume (FBV), which is learning-based and fuses multiple keyframes simultaneously. Finally, another LLM is used to infer target instances based on fused candidate instances. Compared with methods using queryable maps, LIRA achieves better reasoning reconstruction performance by leveraging a MLLM to perform comprehensive vision-instruction fusion and reasoning, and reason about instances that are only relevant to implicit instructions (e.g., LIRA only detects "*sink*" in Fig. 1). Supplementary material gives a visual comparison. Moreover, by applying model acceleration strategies, LIRA is capable of running in real time.

Since the benchmark for reasoning reconstruction evaluation is not well established, we further propose a benchmark *ReasonRecon*, which comprises the largest collection of scene-instruction data samples ($> 5k$) involving implicit and complex reasoning. It supports diverse instance output formats including multi-class, multi-target, single-target, and zero-target configurations. Furthermore, ReasonRecon incorporates rich and high-quality instance-level annotations encompassing both 2D and 3D modalities. In summary, our major contributions are as follows:

- We introduce the *reasoning reconstruction* task, which requires online 3D reconstruction guided by implicit and complex instructions. Also, we propose a reasoning reconstruction method LIRA, which outperforms existing methods and is capable of running in real time.
- To achieve higher instance fusion quality, we propose a **T**ext-enhanced **I**nstance **F**usion module operating within **F**BV (TIFF), which is learning-based and fuses multiple keyframes simultaneously.
- We establish a reasoning reconstruction benchmark, *ReasonRecon*, which contains the largest collection of scene-instruction data samples ($> 5k$) involving implicit and complex reasoning, and diverse instance output formats.

2. Related Work

2.1. Language Instruction-Guided Reconstruction

Existing language instruction-guided reconstruction methods [8, 9, 15, 18, 22, 27–29, 46, 47, 51] mainly rely on explicit instructions or queryable maps, which first apply SLAM systems for geometric reconstruction and extract all object features or structured attributes. Then, they match language instructions with the pre-defined queryable maps. Based on open-vocabulary instructions, OVIR-3D [27] extracted text-aligned features from the open-vocabulary instance segmenter Detic [57] to construct a text-queryable feature map. Open-Fusion [46] employed SEEM [59] to extract region-based features and mapped them to a semantic truncated signed distance function (TSDF) volume, thereby constructing a queryable scene representation.

To enhance comprehension, some approaches [8, 13, 22] employ LLMs for pre-processing or post-processing. VLMaps [13] first employed a LLM to decompose language instructions into multiple landmarks, then extracted visual features from LSeg [21] for queryable map construction. ConceptGraphs [8] and BBQ [22] leveraged segmentation models, such as Segment Anything Model (SAM) [19], to extract all object features, followed by the utilization of LLaVA [23] to capture the visual and spatial attributes, thereby generating queryable maps. Then, they used Llama3 [7] or GPT-4 [33] to infer targets based on the queryable maps and input instructions. However, these methods contain much instruction-irrelevant information, and exhibit limited interaction and reasoning between target instance features and instruction features. Particularly for implicit instructions involving complex reasoning, they are more difficult to handle. LIRA leverages a MLLM to perform comprehensive vision-instruction fusion and reasoning and only reason about instruction-relevant instances, demonstrating better reasoning reconstruction performance.

2.2. 3D Instance Fusion

Existing 3D instance fusion strategies [24, 27, 30, 44, 45, 50, 52] based on RGB-D inputs primarily rely on image

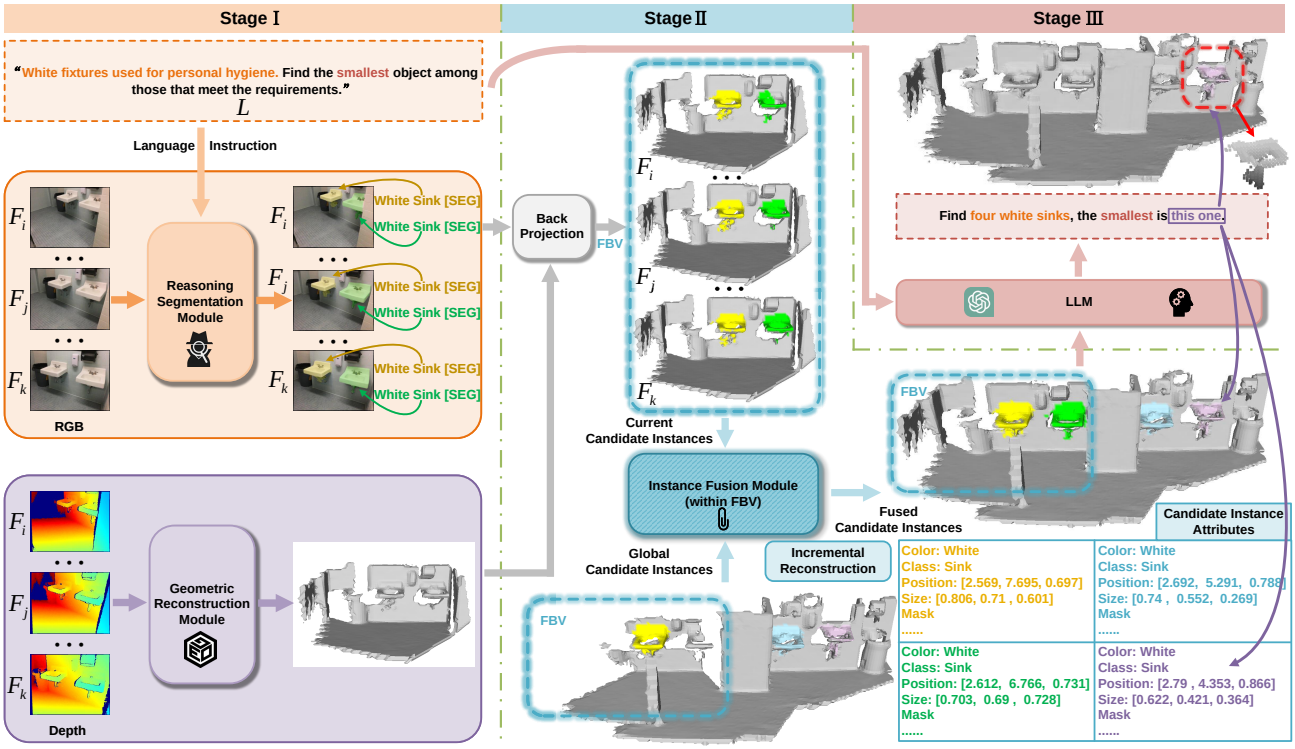

Figure 2. Overview of LIRA. Stage I: LIRA uses depth maps for geometric reconstruction and applies a MLLM to infer instruction-relevant 2D candidate instance masks and their attributes based on the implicit instruction and RGB images. Stage II: It back-projects candidate instances into the reconstructed 3D geometric map and performs instance fusion. Stage III: LLM infers target instances according to fused candidate instance attributes from a global perspective. Different instruction-relevant instances are distinguished by colors.

features and geometric features. For example, OVIR-3D [27] used image features and multiple types of geometric features such as detection rate for instance fusion. To enhance the efficiency of instance fusion, EmbodiedSAM [45] introduces a learning-based fusion strategy to compute geometric, contrastive, and semantic similarities. To further enhance the quality of instance fusion, we innovatively introduce text features into the learning-based fusion process. In addition, since current methods [27, 30, 45, 52] mainly perform instance fusion frame by frame, LIRA simultaneously fuses multiple keyframes within a FBV to learn richer features, achieving better instance fusion performance.

3. Method

The overview of LIRA is depicted in Fig. 2. Given an implicit and complex instruction $\mathcal{L}$ and posed RGB-D sequences as input, LIRA first incrementally performs geometric reconstruction, and leverages a MLLM to actively reason about $\mathcal{L}$ and obtain instruction-relevant 2D candidate instances and their attributes (Stage I). Then, candidate instances are back-projected into the reconstructed 3D geometric map, followed by instance fusion (Stage II). Finally, a LLM infers target instances according to fused candidate instance attributes on the global map (Stage III).

This processing approach is inspired by the human thinking pattern. For example, when given the instruction $\mathcal{L}$ in Fig. 2, humans will scan the environment using their limited binocular vision to gather information. In their brains, they record objects related to the instruction in their limited field of view at each moment. They will look for "*white sink*" (i.e., candidate instances) in each view. The perceptual information is progressively constructed into a global map containing multiple candidate instances in the brain. Ultimately, humans deduce the target instance ("*the smallest object*") through comprehensive reasoning based on these candidate instances from a global perspective.

3.1. Stage I: Geometric Reconstruction and 2D Reasoning Segmentation

3.1.1. Incremental Geometric Reconstruction

To provide enough motion parallax and keep multi-view covisibility for language-instructed reconstruction, following the strategy of [3, 58], a frame is selected as a key frame if its relative translation is greater than t_{max} and the relative rotation angle is greater than R_{max}. A window with N keyframes is defined as a local fragment, and the global reconstruction result is obtained by fusing all local fragments. The maximum visible depth of each view is set to d_{max}, and

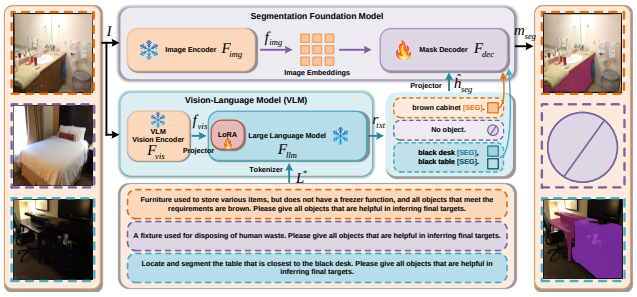

Figure 3. Architecture of the proposed 2D reasoning segmentation module. Three examples are given (corresponding to the dashed boxes of different colors). Please zoom in to see.

all view frustums in a fragment are limited to a cubic-shaped and voxelized FBV $\mathcal{B}_t$. The geometric reconstruction process of LIRA is mainly conducted within each FBV.

For an incoming FBV $\mathcal{B}_t$ at time t and depth maps from N views, LIRA performs depth map fusion using standard TSDF integration proposed in [4, 31] with a truncation distance λ. To perform global TSDF fusion, only the global TSDF values within the current FBV $\mathcal{B}_t$ are updated. For the previous global TSDF fusion result, its part within $\mathcal{B}_t$ is involved in the fusion process. We incrementally recover scene geometry and only retain TSDF values smaller than λ. Marching Cubes algorithm [26] is performed to reconstruct the scene mesh. Based on geometric reconstruction results, LIRA further segments target instances that conform to $\mathcal{L}$. The supplementary material provides visualizations.

3.1.2. 2D Reasoning Segmentation within the FBV

To reason about implicit and complex language instructions, inspired by [20, 42, 48], we leverage a vision-language model (VLM) to perform instruction reasoning and a segmentation foundation model to perform segmentation. The proposed 2D reasoning segmentation module is in Fig. 3.

The key frame images $\mathcal{I}$ within $\mathcal{B}_t$ are respectively transformed into f_{vis} and f_{img} by the vision encoder $\mathcal{F}_{\text{vis}}$ of the VLM and the image encoder $\mathcal{F}_{\text{img}}$ of the segmentation foundation model. The process can be formulated as

$$f_{vis} = \mathcal{F}_{\text{vis}}(\mathcal{I}), \quad f_{img} = \mathcal{F}_{\text{img}}(\mathcal{I}). \tag{1}$$

We first add a prompt "*Please give all objects that are helpful in inferring final targets*" to $\mathcal{L}$ to obtain $\mathcal{L}^*$. The instruction $\mathcal{L}^*$ is tokenized and sent to the LLM $\mathcal{F}_{\text{llm}}$ of the VLM along with f_{vis} for reasoning and understanding. Then, VLM identifies instances in images that conform to $\mathcal{L}^*$. LoRA [11] is used for efficient fine-tuning.

An image can only provide instance information within a local field of view, and the complex language instruction requires reasoning based on the global map. In the 2D reasoning segmentation part, we need to infer instances (i.e., candidate instances) from the current field of view that are helpful for reasoning final target instances. For each candidate instance, VLM infers its structured attributes in text form r_{txt}, such as color and class.

$$r_{txt} = \mathcal{F}_{\text{llm}}(\mathcal{L}^*, f_{vis}). \tag{2}$$

Multiple candidate instances are output sequentially. Inspired by [20], we expand the VLM vocabulary with a new token, i.e., [SEG]. To output the segmentation mask of the candidate instance, the [SEG] token is predicted after the attribute information of each candidate instance. The VLM last-layer embedding h_{seg} corresponding to the [SEG] token is extracted and transformed into $\hat{h}_{seg}$ by a simple Multi-Layer Perceptron (MLP) projection layer $\mathcal{F}_{\text{proj}}$.

$$\hat{h}_{seg} = \mathcal{F}_{\text{proj}}(h_{seg}). \tag{3}$$

Then, f_{img} and $\hat{h}_{seg}$ (text feature prompt) are input into the mask decoder $\mathcal{F}_{\text{dec}}$ of the segmentation foundation model to output the binary mask m_{seg} of the candidate instance. Each [SEG] token outputs a mask.

$$m_{seg} = \mathcal{F}_{\text{dec}}(f_{img}, \hat{h}_{seg}). \tag{4}$$

Compared with previous works [20, 42, 48], our 2D reasoning segmentation module predicts attributes for each instance and supports the output format of $0/1/n$ $(n > 1)$ instances. If there are no candidate instances in the image, this module returns "*No object*". Furthermore, it supports outputting different instances of candidate objects with the same attributes. For example, our module returns "*black chair [SEG] and black chair [SEG]*", where masks are predicted in a fixed regular order according to their pixel positions. The experiments in Section 4.6 and supplementary material demonstrate the effectiveness of the above designs.

3.2. Stage II: Text-Enhanced Instance Fusion within the FBV

Based on depth maps, candidate instance masks and their attributes are back-projected into the TSDF volume within the current FBV $\mathcal{B}_t$. The flowchart of candidate instance fusion within $\mathcal{B}_t$ is shown in Fig. 4. We first extract voxel features $\mathcal{F}_{voxel}$, text features $\mathcal{F}_{text}$, and axis-aligned 3D bounding boxes $\{x, y, z, w, h, l\}$ for each candidate instance, where $\{x, y, z\}$ and $\{w, h, l\}$ are the center point position and side lengths in the global coordinate system, respectively.

For voxel feature extraction, we back-project image features from N views into voxels of $\mathcal{B}_t$ and take the average of these features. Compared with frame-by-frame fusion, richer features are obtained from multiple views. The image features directly use the image embeddings f_{img} of the segmentation foundation model in the 2D reasoning segmentation module. For text feature extraction, the VLM last-layer embedding h_{seg} corresponding to the [SEG] token is used.

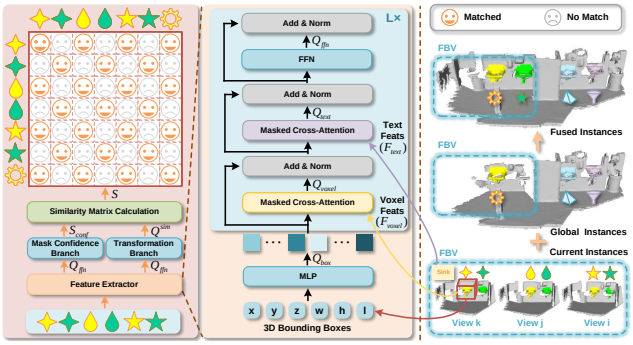

Figure 4. Illustration of the proposed instance fusion module. Different colors represent different instances, and different icons represent instances at different frames. Please zoom in to see.

In addition, for 3D bounding box extraction, since the 3D mask of each candidate instance is obtained, we directly extract the axis-aligned 3D bounding box of each mask.

We design a network with L repeated transformer blocks to aggregate these features. The M 3D bounding boxes are first passed through a MLP to obtain $\mathcal{Q}_{box} \in \mathbb{R}^{M \times C}$ (M is the number of instances and C is the feature dimension).

$$\mathcal{Q}_{box} = \text{MLP}(\{x, y, z, w, h, l\}^{\times M}). \tag{5}$$

Then, a masked cross-attention is applied to integrate voxel features $\mathcal{F}_{voxel}$. The masked attention mechanism limits the attention range to the foreground region of the 3D mask of each candidate instance and ensures that the attention process only interacts with the current instance region. For convenience, *Add* and *LayerNorm* layers are ignored.

$$\mathcal{Q}_{voxel} = \text{MaskedCrossAttn}(\mathcal{Q}^Q_{box}, \mathcal{F}^K_{voxel}, \mathcal{F}^V_{voxel}). \tag{6}$$

Fourier positional encodings [37] based on voxel positions are used. Subsequently, masked cross-attention is also utilized to integrate text features $\mathcal{F}_{text}$.

$$\mathcal{Q}_{text} = \text{MaskedCrossAttn}(\mathcal{Q}^Q_{voxel}, \mathcal{F}^K_{text}, \mathcal{F}^V_{text}). \tag{7}$$

The masked attention mechanism makes $\mathcal{Q}_{voxel}$ interact only with the text features of the current candidate instance. A feed-forward network is used to aggregate features.

$$\mathcal{Q}_{ffn} = \text{FFN}(\mathcal{Q}_{text}). \tag{8}$$

The aggregated features $\mathcal{Q}_{ffn}$ are then passed through two network branches, both composed of MLPs. The mask confidence branch predicts a confidence score $\mathcal{S}_{conf}$ for each instance. If the confidence score is lower than a threshold θ_{conf}, the corresponding candidate instance is discarded. The transformation branch further transforms $\mathcal{Q}_{ffn}$ into $\mathcal{Q}^{sim}$ for similarity calculation between instances.

Similar to the geometric reconstruction process, when fusing with global candidate instances, only global instances within the current FBV $\mathcal{B}_t$ are involved in the fusion process (i.e., within the blue dashed box in Fig. 2 and Fig. 4). A similarity matrix $\mathcal{S}$ is constructed and each element s_{ij} is the cosine similarity between candidate instance i and candidate instance j, where candidate instances come from either the global candidate instances or candidate instances obtained from N current keyframe images within $\mathcal{B}_t$.

$$s_{ij} = \cos < \mathcal{Q}^{sim}_i, \mathcal{Q}^{sim}_j > . \tag{9}$$

A s_{ij} greater than a predefined threshold θ_{sim} is judged to be the same instance. The DBSCAN [6] algorithm is applied to the similarity matrix $\mathcal{S}$ to fuse multiple instances simultaneously. Similar to [45], our instance fusion process is also learning-based and thus computationally efficient.

For instance attribute updates on the global map, weighted averages are used to update the continuous-valued attributes (e.g., $\mathcal{Q}^{sim}$). The attributes with the highest occurrence frequency are used to update the discrete-valued attributes (e.g., class). Since we do not need to store high-dimensional feature vectors for each voxel or maintain lengthy redundant attribute documents for each instance, as are required in queryable maps [8, 22], our global map has relatively low storage requirements and stronger scalability.

3.3. Stage III: LLM Inference

After instance fusion, we obtain all candidate instances and their attributes on the global map, which are used to infer target instances that conform to the input language instruction $\mathcal{L}$. Another LLM, such as ChatGPT-4o or Qwen2.5 [35], is applied to perform global reasoning. For example, given the instruction $\mathcal{L}$ and four candidate instances with their attributes in Fig. 2, the LLM infers "*white fixtures used for personal hygiene*" (*white sink*) and identifies "*the smallest one*" (painted purple in the figure). The detailed form of the prompt input to the LLM is as follows:

The followings are some objects and their attribute information: {Object ID, Color, Class, Position, Size, ...}, {...}. Here is now a language instruction. Find objects and their IDs that match the instruction based on the above object information. Let's think step by step.

In addition, when an additional prompt "*Finally, write a brief sentence to summarize*" is provided to the LLM, it can also generate an additional explanatory text description.

3.4. Benchmark

To establish a comprehensive evaluation system suitable for the reasoning reconstruction task, a benchmark *ReasonRecon* is constructed and the data collection pipeline is shown in Fig. 5. It uses RGB-D sequences and 3D instance segmentation annotations in the ScanNetV2 dataset [5].

First, we extend the attributes of instances in ScanNetV2. The newly added attributes are frequently referenced in human language expressions [2]. Specifically, for an instance in a given scene, its point cloud is projected into an image

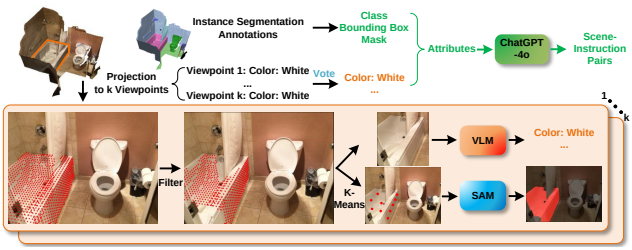

Figure 5. Illustration of the data collection process. Red dots are projected points. Extended attributes are indicated in orange font.

Dataset	Scene-Instruction Pairs	Implicit Instruction Scale $> 5k$	High-Quality 2D Annotations	M., M., Z. Outputs
ScanNetV2 [5]	-	-	✗	-
ScanNet++ [49]	-	-	✗	-
Sr3D/Nr3D [1]	83,572/41,503	✗	✗	✗
ScanRefer [2]	51,583	✗	✗	✗
Multi3DRefer [56]	61,926	✗	✗	✗
Instruct3D [10]	2,565	✗	✗	✗
ReasonRecon	12,500	✓	✓	✓

Table 1. Comparison of ReasonRecon with related datasets. "M., M., Z." indicates multi-class, multi-target, and zero-target.

from a certain viewpoint. Erroneous projected pixels caused by occlusion are filtered out. Then, we crop out the instance region and utilize a VLM (i.e., Qwen2-VL [39]) to infer attributes, e.g., color. Inference results from all viewpoints are voted to determine the attributes of the instance.

Second, we construct scene-instruction pairs. ChatGPT-4o is utilized to design a series of instruction templates and generate scene-instruction pairs (see supplementary material) based on instance attributes in a scene, which include instance segmentation annotations (e.g., class, bounding box, and mask) and attributes inferred by Qwen2-VL (e.g., color). The generated instructions are further polished by ChatGPT-4o to make them more consistent with human language expression. Human corrections are also incorporated to enhance the quality of these instructions. The training set and test set are divided into $8:2$.

Third, we generate 2D segmentation masks for instances in each scene-instruction pair. After obtaining the pixels of an instance projected into an image from a certain viewpoint in the first step, these pixels are subjected to K-Means clustering. Then, we use the cluster centers as prompts and apply SAM to generate the corresponding segmentation mask. To evaluate 2D mask quality, we randomly sample 2000 images (about 1%) for manual annotation. The mean Intersection over Union (mIoU) between manually annotated masks and automatically generated masks is 95.6%. Based on a normal distribution approximation, at a 95% confidence level, the mIoU for the entire ReasonRecon is estimated to be between 94.7% and 96.5%. This demonstrates the high quality of the automatically generated 2D mask annotations.

The comparison of ReasonRecon with related datasets is shown in Tab. 1. ReasonRecon has the largest-scale implicit scene-instruction pairs ($> 5k$) and supports a wide range of output types, including multi-class, multi-target, zero-target and single-target outputs. Also, we obtain high-quality 2D segmentation annotations. In addition, ReasonRecon contains many instructions that require spatial reasoning, involving reasoning based on the size and relative positions of objects. More details are in the supplementary material.

Method	Online	AP	AP_{50}	AP_{25}
OpenScene [34] + DBSCAN [6]	✗	3.04	6.17	10.60
OpenScene [34] + Mask3D [37]	✗	8.88	10.59	12.23
OpenMask3D [38]	✗	10.37	15.34	16.87
OpenIns3D [14] (ODISE [43])	✗	8.08	10.74	11.80
Open3DIS [32] (2D and 3D)	✗	10.60	16.08	19.12
3D-STMN [40]	✗	10.93	17.80	25.67
VLMaps-3D [13]	✓	8.49	11.06	15.04
OVIR-3D [27]	✓	9.25	13.86	18.99
MaskClustering [47]	✓	10.26	15.24	21.61
Open-Fusion [46]	✓	9.51	14.19	19.50
ConceptFusion [15]	✓	10.33	15.38	21.88
ConceptGraphs [8]	✓	11.13	19.84	30.55
BBQ [22]	✓	11.52	22.17	35.86
LIRA	✓	11.57	34.39	66.24
LIRA*	✓	**12.49**	**35.32**	**67.34**

Table 2. Quantitative results of reasoning reconstruction.

4. Experiments

4.1. Implementation Details

The voxel size is set to 4cm, and 9 keyframe sequences make up a FBV. We use 3 transformer blocks with a feature dimension of 128 in the instance fusion module. Unless otherwise specified, in the 2D reasoning segmentation module, we use LLaVA-7B [23] as the base VLM, and adopt SAM with ViT-H backbone [19] as the segmentation foundation model. The minimum number of neighbors of a core point in the DBSCAN algorithm is set to 1. The LLM in stage III is ChatGPT-4o-mini. Loss functions and more implementation details are described in the supplementary material.

4.2. Evaluation Metrics

Since reconstruction of environment is not the core of reasoning reconstruction, the reconstruction performance of instruction-relevant instances is primarily evaluated. We evaluate using standard Average Precision (AP) metrics at IoU thresholds of 50% and 25%, and also calculate mean score across IoU thresholds from 50% to 95% in 5% increments. These metrics evaluate the performance of both geometric reconstruction and instance matching.

4.3. Reasoning Reconstruction Results

As shown in Tab. 2, we begin by comparing with online instruction-guided reconstruction methods. Some of them are improved to support multi-instance outputs for a fair

Method	Online	AP	AP_{50}	AP_{25}
OpenScene [34] + DBSCAN [6]	✗	6.94	15.84	23.66
OpenScene [34] + Mask3D [37]	✗	14.01	20.96	28.57
OpenMask3D [38]	✗	19.62	30.17	42.62
OpenIns3D [14] (ODISE [43])	✗	12.90	18.32	25.41
Open3DIS [32] (2D and 3D)	✗	**22.17**	35.12	46.98
3D-STMN [40]	✗	19.73	31.45	42.17
VLMaps-3D [13]	✓	13.58	21.03	35.34
OVIR-3D [27]	✓	18.11	33.65	40.44
MaskClustering [47]	✓	18.96	35.40	42.29
Open-Fusion [46]	✓	19.00	35.11	42.67
ConceptFusion [15]	✓	18.30	34.62	41.95
LIRA	✓	20.38	41.44	71.02
LIRA*	✓	21.19	**42.80**	**72.39**

Table 3. Results of explicit instruction-guided reconstruction.

Method	KFPS	AP	AP_{50}	AP_{25}
OVIR-3D [27]	2.56	9.25	13.86	18.99
Open-Fusion [46]	4.06	9.51	14.19	19.50
LIRA-Fast	**5.63**	**10.47**	**31.67**	**61.75**

Table 4. Runtime analysis of reasoning reconstruction.

comparison. VLMaps is extended to a 3D map by canceling top-down projection. LIRA* represents that LIRA uses LLaVA-13B and applies ChatGPT-4o for reasoning in stage III. Experiments demonstrate that LIRA outperforms existing methods across all accuracy metrics. Compared to methods based mainly on explicit instructions or queryable maps, LIRA exhibits superior reasoning and comprehension capabilities for implicit and complex instructions.

Subsequently, we compare LIRA with recent point cloud segmentation methods. These methods perform segmentation on the complete scene point clouds reconstructed from RGB-D sequences, and thus can be regarded as offline instruction-guided reconstruction. 3D-STMN [40] is retrained on ReasonRecon. Although these methods usually utilize visual-language feature extractors such as CLIP [36] to comprehend instructions from a global or local view, their performance for implicit instructions is inferior to LIRA.

Another finding is that LIRA significantly outperforms other methods in terms of low IoU threshold metrics, indicating that LIRA has a strong ability to find target instances, though its high-precision reconstruction ability needs further improvement. The performance improvement of LIRA* indicates that using more powerful MLLMs further enhances reasoning reconstruction performance.

4.4. Explicit Instruction-Guided Reconstruction

For a fair comparison with existing methods, we also evaluate reconstruction results guided by explicit and relatively simple instructions. Specifically, we modify ReasonRecon by replacing all implicit and complex instructions with corresponding explicit vocabulary words. For example, an instruction "*Appliances or furniture used to store food*" is replaced with "*Cabinet, Refrigerator*". The generated explicit instruction format is consistent with the instruction format used in existing online methods for ScanNet scenes. In Tab. 3, experimental results demonstrate that LIRA still achieves superior performance, particularly showing a significant lead on AP_{50} and AP_{25} metrics.

Stage	Method	AP	AP_{50}	AP_{25}
I	Replace with SEEM [59]	3.68	11.00	19.57
	Replace with Grounded-SAM [25]	3.06	10.12	18.26
	Replace with LISA [20] (ft)	6.01	22.66	42.40
	Replace with LISA++ [48] (ft)	11.17	33.34	64.61
	Replace with GSVA [42] (ft)	10.68	31.07	60.53
	Replace with LLaVA-Grounding [54] (ft)	**12.12**	34.26	64.88
	The final model	11.57	**34.39**	**66.24**
II	Remove bbox features	10.33	33.21	64.07
	Remove voxel features	10.29	33.68	63.29
	Remove text features	10.49	33.13	64.79
	Remove confidence	8.20	29.30	61.05
	Remove FBV fusion	11.44	34.15	65.87
	The final model	**11.57**	**34.39**	**66.24**
III	Replace with Llama3 [7]	11.06	33.27	65.10
	Replace with Qwen2.5 [35]	11.25	33.58	65.45
	Replace with ChatGPT-4o [33]	**11.81**	**34.74**	**66.45**
	The final model	11.57	34.39	66.24

Table 5. Ablation studies of the three stages of LIRA.

4.5. Runtime Analysis

For a fair comparison, runtime evaluation is performed on a single NVIDIA Tesla A800 GPU. We measure time consumption for reasoning reconstruction in KFPS (key frames per second). The average inference time for each RGB-D keyframe in the FBV is provided. To achieve real-time inference, we propose LIRA-Fast. LIRA-Fast utilizes parallel TSDF fusion, 8-bit quantized 2D reasoning segmentation, 8-bit quantized instance fusion, MobileSAM [53], and *qwen2.5-7b-instruct* API. In stage III, the prompt for the LLM is changed from "*Let's think step by step*" to "*Let's respond briefly and quickly*" to accelerate inference.

As shown in Tab. 4, consistent with the evaluation criteria in [3, 41, 58], since keyframes are created at a far lower frequency than the framerate (approximately 8.7 times lower based on our keyframe selection method), LIRA-Fast still achieves a real-time reasoning reconstruction. Compared with other methods, our LIRA-Fast has advantages in both reasoning reconstruction speed and accuracy.

4.6. Ablation Studies

To verify the effectiveness of the components in LIRA, we observe the performance changes after replacing or removing certain modules in all three stages, as shown in Tab. 5.

Stage I. We first replace the 2D reasoning segmentation module of LIRA with some other segmentation methods. "ft" denotes the model is finetuned on the training set of ReasonRecon. Experiments demonstrate the effectiveness

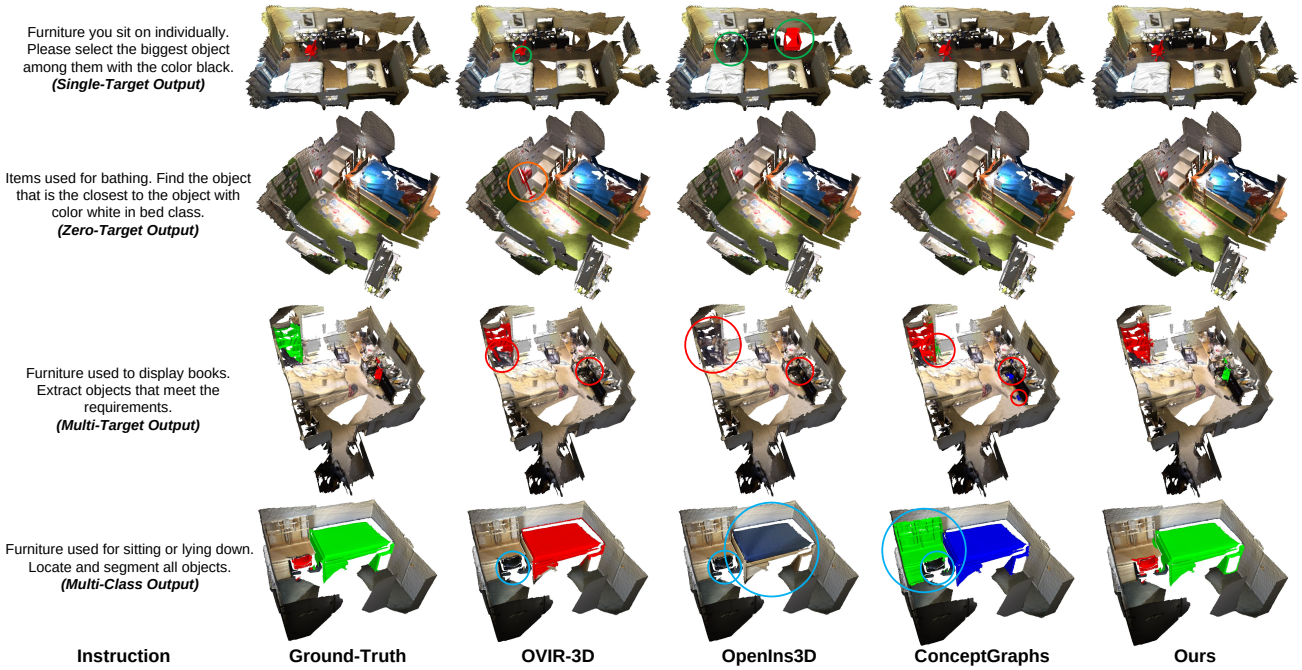

Figure 6. Visualization results of different reasoning reconstruction methods on the ReasonRecon test set. The reconstructed geometric results are augmented with image textures. Single-layered mesh is visualized.

of our design of the 2D reasoning segmentation module. Compared with LISA [20], LISA++ [48] and GSVA [42], which have a similar architecture to our method, our method achieves better reasoning reconstruction accuracy. This improvement is primarily attributed to our more effective handling of multi-target and zero-target outputs.

Stage II. Experiments show that each design is important for the quality of instance fusion. Since text features are responsible for distinguishing different instances within a single view and guiding the generation of masks for candidate instances, they store rich semantic information about different instances. Text features further enhance the instance fusion quality. We also attempt not to perform the fusion of multiple keyframe instances simultaneously within a FBV, instead adopting the fusion strategy commonly used in existing methods [27, 45], where one keyframe is fused at a time. However, this strategy results in reduced fusion performance, as the model learns richer and more effective features when multiple keyframes are fused within the FBV. The mask confidence branch is necessary since it eliminates many low-quality candidate instances. The multidimensional scaling (MDS) visualization in the supplementary material validates that the proposed TIFF learns discriminative feature representation for object matching.

Stage III. ChatGPT-4o-mini is replaced with other mainstream LLMs. Experimental results show that ChatGPT series outperform other LLMs. ChatGPT-4o-mini is ultimately selected due to its significantly faster inference speed compared to ChatGPT-4o, while maintaining comparable reasoning reconstruction accuracy. In addition, more ablation studies are provided in the supplementary material.

4.7. Qualitative Results

As depicted in Fig. 6, a visual comparison with existing related works is provided. We give four different types of outputs, including single-target, zero-target, multi-target, and multi-class cases. LIRA exhibits superior recognition and reasoning reconstruction capabilities for implicit instruction-relevant target instances. More visualization results are given in the supplementary material.

5. Conclusion

This paper introduces the *reasoning reconstruction* task, which focuses on online 3D reconstruction guided by implicit and complex language instructions. Also, we propose LIRA, an effective framework designed for addressing the reasoning reconstruction task. In LIRA, a learning-based TIFF is proposed to improve instance fusion quality. In addition, we propose a benchmark *ReasonRecon* comprising the largest collection of scene-instruction data samples involving implicit and complex reasoning. Experiments demonstrate that LIRA achieves superior reasoning reconstruction performance and is capable of running in real time. One limitation is that LIRA exhibits relatively low performance in high-precision reconstruction. Future work will consider further optimization in 3D space. We hope that our work can provide insights for embodied agents to better understand complex physical environments.

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 *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 422–440, 2020. 6

[2] Dave Zhenyu Chen, Angel X. Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 202–221, 2020. 5, 6

[3] Xi Chen, Jiaming Sun, Yiming Xie, Hujun Bao, and Xiaowei Zhou. Neuralrecon: Real-time coherent 3d scene reconstruction from monocular video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):7542–7555, 2024. 3, 7

[4] Brian Curless and Marc Levoy. A volumetric method for building complex models from range images. In *SIGGRAPH*, page 303–312, 1996. 4

[5] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Niessner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 5, 6

[6] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Knowledge Discovery and Data Mining*, pages 226–231, 1996. 5, 6, 7

[7] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, et al. The llama 3 herd of models. In *arXiv preprint arXiv:2407.21783*, 2024. 2, 7

[8] Qiao Gu, Ali Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, Chuang Gan, Celso Miguel de Melo, Joshua B. Tenenbaum, Antonio Torralba, Florian Shkurti, and Liam Paull. Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5021–5028, 2024. 2, 5, 6

[9] Huy Ha and Shuran Song. Semantic abstraction: Open-world 3D scene understanding from 2D vision-language models. In *Proceedings of the 2022 Conference on Robot Learning*, 2022. 2

[10] Shuting He, Henghui Ding, Xudong Jiang, and Bihan Wen. Segpoint: Segment any point cloud via large language model. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 349–367, 2025. 6

[11] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *arXiv preprint arXiv:2106.09685*, 2021. 4

[12] Jiarui Hu, Xianhao Chen, Boyin Feng, Guanglin Li, Liangjing Yang, Hujun Bao, Guofeng Zhang, and Zhaopeng Cui. Cg-slam: Efficient dense rgb-d slam in a consistent uncertainty-aware 3d gaussian field. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 93–112, 2025. 2

[13] Chenguang Huang, Oier Mees, Andy Zeng, and Wolfram Burgard. Visual language maps for robot navigation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 10608–10615, 2023. 2, 6, 7

[14] Zhening Huang, Xiaoyang Wu, Xi Chen, Hengshuang Zhao, Lei Zhu, and Joan Lasenby. Openins3d: Snap and lookup for 3d open-vocabulary instance segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 169–185, 2025. 6, 7

[15] Krishna Murthy Jatavallabhula, Alihusein Kuwajerwala, Qiao Gu, Mohd Omama, Tao Chen, Shuang Li, Ganesh Iyer, Soroush Saryazdi, Nikhil Keetha, Ayush Tewari, Joshua B. Tenenbaum, Celso Miguel de Melo, Madhava Krishna, Liam Paull, Florian Shkurti, and Antonio Torralba. Conceptfusion: Open-set multimodal 3d mapping. *Robotics: Science and Systems (RSS)*, 2023. 1, 2, 6, 7

[16] Krishna Murthy Jatavallabhula, Ganesh Iyer, and Liam Paull. slam: Dense slam meets automatic differentiation. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2130–2137, 2020. 2

[17] Nikhil Keetha, Jay Karhade, Krishna Murthy Jatavallabhula, Gengshan Yang, Sebastian Scherer, Deva Ramanan, and Jonathon Luiten. Splatam: Splat track & map 3d gaussians for dense rgb-d slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21357–21366, 2024. 2

[18] Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lerf: Language embedded radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19729–19739, 2023. 2

[19] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4015–4026, 2023. 2, 6

[20] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9579–9589, 2024. 4, 7, 8

[21] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and Rene Ranftl. Language-driven semantic segmentation. In *International Conference on Learning Representations*, 2022. 2

[22] Sergey Linok, Tatiana Zemskova, Svetlana Ladanova, Roman Titkov, Dmitry Yudin, Maxim Monastyrny, and Aleksei Valenkov. Beyond bare queries: Open-vocabulary object grounding with 3d scene graph. In *arXiv preprint arXiv:2406.07113*, 2024. 2, 5, 6

[23] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, pages 34892–34916, 2023. 2, 6

[24] Leyao Liu, Tian Zheng, Yun-Jou Lin, Kai Ni, and Lu Fang. Ins-conv: Incremental sparse convolution for online 3d segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18975–18984, 2022. 2

[25] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 38–55, 2025. 7

[26] William E. Lorensen and Harvey E. Cline. Marching cubes: A high resolution 3d surface construction algorithm. In *SIGGRAPH*, pages 163–169, 1987. 4

[27] Shiyang Lu, Haonan Chang, Eric Pu Jing, Abdeslam Boularias, and Kostas Bekris. Ovir-3d: Open-vocabulary 3d instance retrieval without training on 3d data. In *Proceedings of The 7th Conference on Robot Learning*, pages 1610–1620, 2023. 1, 2, 3, 6, 7, 8

[28] Dominic Maggio, Yun Chang, Nathan Hughes, Matthew Trang, Dan Griffith, Carlyn Dougherty, Eric Cristofalo, Lukas Schmid, and Luca Carlone. Clio: Real-time task-driven open-set 3d scene graphs. *IEEE Robotics and Automation Letters*, 9(10):8921–8928, 2024.

[29] Nur (Mahi)Shafiullah, Chris Paxton, Lerrel Pinto, Soumith Chintala, and Arthur Szlam. Clip-fields: Weakly supervised semantic fields for robotic memory. In *Robotics: Science and Systems XIX*, 2023. 2

[30] Gaku Narita, Takashi Seno, Tomoya Ishikawa, and Yohsuke Kaji. Panopticfusion: Online volumetric semantic mapping at the level of stuff and things. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4205–4212, 2019. 2, 3

[31] Richard A. Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J. Davison, Pushmeet Kohi, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *2011 10th IEEE International Symposium on Mixed and Augmented Reality*, pages 127–136, 2011. 4

[32] Phuc Nguyen, Tuan Duc Ngo, Evangelos Kalogerakis, Chuang Gan, Anh Tran, Cuong Pham, and Khoi Nguyen. Open3dis: Open-vocabulary 3d instance segmentation with 2d mask guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4018–4028, 2024. 6, 7

[33] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, et al. Gpt-4 technical report. In *arXiv preprint arXiv:2303.08774*, 2024. 2, 7

[34] 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)*, pages 815–824, 2023. 6, 7

[35] Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, et al. Qwen2.5 technical report. In *arXiv preprint arXiv:2412.15115*, 2025. 5, 7

[36] 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. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763, 2021. 7

[37] Jonas Schult, Francis Engelmann, Alexander Hermans, Or Litany, Siyu Tang, and Bastian Leibe. Mask3d: Mask transformer for 3d semantic instance segmentation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8216–8223, 2023. 5, 6, 7

[38] Ayça Takmaz, Elisabetta Fedele, Robert W. Sumner, Marc Pollefeys, Federico Tombari, and Francis Engelmann. OpenMask3D: Open-Vocabulary 3D Instance Segmentation. In *Advances in Neural Information Processing Systems*, 2023. 6, 7

[39] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. In *arXiv preprint arXiv:2409.12191*, 2024. 6

[40] Changli Wu, Yiwei Ma, Qi Chen, Haowei Wang, Gen Luo, Jiayi Ji, and Xiaoshuai Sun. 3d-stmn: Dependency-driven superpoint-text matching network for end-to-end 3d referring expression segmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(6):5940–5948, 2024. 6, 7

[41] Dong Wu, Zike Yan, and Hongbin Zha. Panorecon: Real-time panoptic 3d reconstruction from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21507–21518, 2024. 7

[42] Zhuofan Xia, Dongchen Han, Yizeng Han, Xuran Pan, Shiji Song, and Gao Huang. Gsva: Generalized segmentation via multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3858–3869, 2024. 4, 7, 8

[43] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2955–2966, 2023. 6, 7

[44] Xiuwei Xu, Chong Xia, Ziwei Wang, Linqing Zhao, Yueqi Duan, Jie Zhou, and Jiwen Lu. Memory-based adapters for online 3d scene perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21604–21613, 2024. 2

[45] Xiuwei Xu, Huangxing Chen, Linqing Zhao, Ziwei Wang, Jie Zhou, and Jiwen Lu. EmbodiedSAM: Online segment any 3d thing in real time. In *The Thirteenth International Conference on Learning Representations*, 2025. 2, 3, 5, 8

[46] Kashu Yamazaki, Taisei Hanyu, Khoa Vo, Thang Pham, Minh Tran, Gianfranco Doretto, Anh Nguyen, and Ngan Le. Open-fusion: Real-time open-vocabulary 3d mapping and queryable scene representation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9411–9417, 2024. 1, 2, 6, 7

[47] Mi Yan, Jiazhao Zhang, Yan Zhu, and He Wang. Maskclustering: View consensus based mask graph clustering for open-vocabulary 3d instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 28274–28284, 2024. 1, 2, 6, 7

[48] Senqiao Yang, Tianyuan Qu, Xin Lai, Zhuotao Tian, Bohao Peng, Shu Liu, and Jiaya Jia. Lisa++: An improved baseline for reasoning segmentation with large language model. In *arXiv preprint arXiv:2312.17240*, 2024. 4, 7, 8

[49] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12–22, 2023. 6

[50] Yingda Yin, Yuzheng Liu, Yang Xiao, Daniel Cohen-Or, Jingwei Huang, and Baoquan Chen. Sai3d: Segment any instance in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3292–3302, 2024. 2

[51] Justin Yu, Kush Hari, Kishore Srinivas, Karim El-Refai, Adam Rashid, Chung Min Kim, Justin Kerr, Richard Cheng, Muhammad Zubair Irshad, Ashwin Balakrishna, Thomas Kollar, and Ken Goldberg. Language-embedded gaussian splats (legs): Incrementally building room-scale representations with a mobile robot. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 13326–13332, 2024. 2

[52] Tong He Hengshuang Zhao Yunhan Yang, Xiaoyang Wu and Xihui Liu. Sam3d: Segment anything in 3d scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2023. 2, 3

[53] Chaoning Zhang, Dongshen Han, Yu Qiao, Jung Uk Kim, Sung-Ho Bae, Seungkyu Lee, and Choong Seon Hong. Faster segment anything: Towards lightweight sam for mobile applications. In *arXiv preprint arXiv:2306.14289*, 2023. 7

[54] Hao Zhang, Hongyang Li, Feng Li, Tianhe Ren, Xueyan Zou, Shilong Liu, Shijia Huang, Jianfeng Gao, Leizhang, Chunyuan Li, and Jainwei Yang. Llava-grounding: Grounded visual chat with large multimodal models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 19–35, 2025. 7

[55] Shishun Zhang, Longyu Zheng, and Wenbing Tao. Survey and evaluation of rgb-d slam. *IEEE Access*, 9:21367–21387, 2021. 2

[56] Yiming Zhang, ZeMing Gong, and Angel X. Chang. Multi3drefer: Grounding text description to multiple 3d objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15225–15236, 2023. 6

[57] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 350–368, 2022. 2

[58] Zhen Zhou, Yunkai Ma, Junfeng Fan, Shaolin Zhang, Fengshui Jing, and Min Tan. Eprecon: An efficient framework for real-time panoptic 3d reconstruction from monocular video. 2024. 3, 7

[59] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. In *Advances in Neural Information Processing Systems*, pages 19769–19782, 2023. 2, 7