

ConceptFusion: Open-set Multimodal 3D Mapping

Krishna Murthy Jatavallabhula¹, Alihusein Kuwajerwala^{*2}, Qiao Gu^{*3}, Mohd Osama^{*4}, Tao Chen¹, Shuang Li¹, Ganesh Iyer^{†6}, Soroush Saryazdi^{‡7}, Nikhil Keetha⁵, Ayush Tewari¹, Joshua B. Tenenbaum¹, Celso Miguel de Melo⁸, Madhava Krishna⁴, Liam Paull², Florian Shkurti³, and Antonio Torralba¹

¹MIT, ²Université de Montréal, ³University of Toronto, ⁴IIT Hyderabad, ⁵CMU, ⁶Amazon, ⁷Matician,
⁸DEVCOM Army Research Laboratory

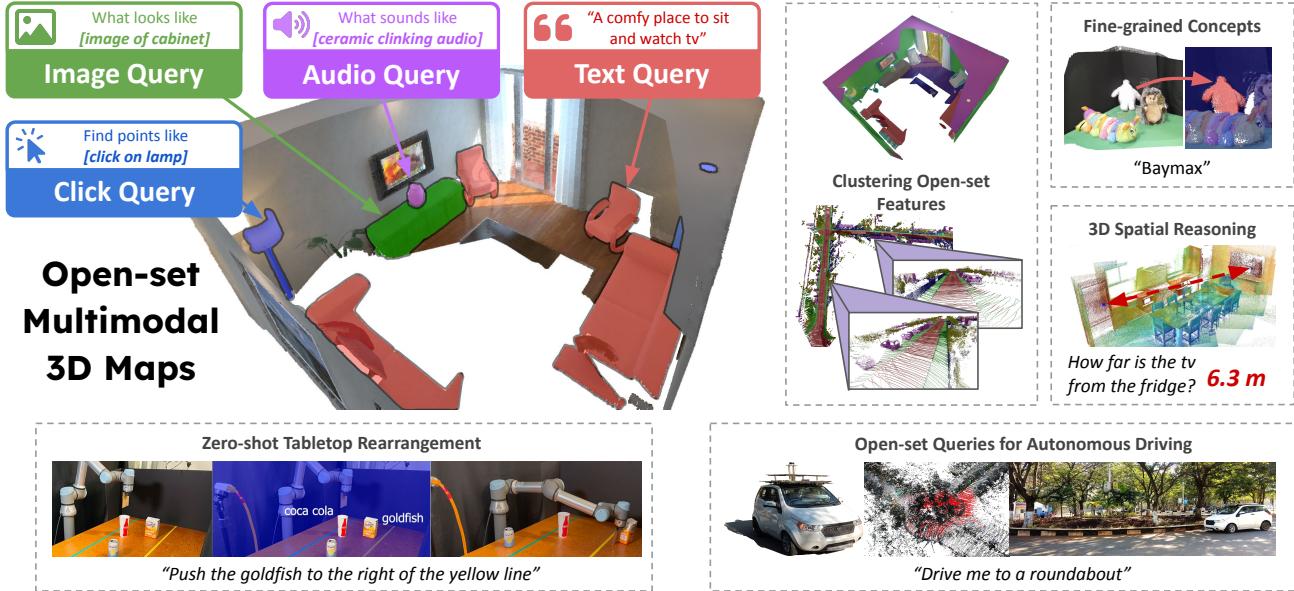


Figure 1: **ConceptFusion** builds *open-set multimodal 3D maps* from RGB images and depth, and features from foundation models like CLIP, DINO, AudioCLIP, etc. These maps are built online, and can be queried for arbitrary concepts specified as text, images, audio samples, or clicks on the 3D map. The fused features have an implicit understanding of semantic concepts, as evident by visualizing clusters obtained from a K-means algorithm. **ConceptFusion** features are significantly more adept at retaining fine-grained concepts, such as the disney character “*Baymax*”. We also build 3D spatial reasoning modules that enable reasoning about frequently observed spatial relationships. We demonstrate the applicability of **ConceptFusion** to the real-world robotic tasks of tabletop manipulation of novel objects, and an urban autonomous driving setting. ([Webpage](#))

Abstract

Building 3D maps of the environment is central to robot navigation, planning, and interaction with objects in a scene. Most existing approaches that integrate semantic concepts with 3D maps largely remain confined to the closed-set setting: they can only reason about a finite set of concepts, pre-defined at training time. Further, these maps can only be queried using class labels, or recently, using text prompts.

We address both these issues with **ConceptFusion**, a scene representation that is: (i) fundamentally open-set, enabling reasoning beyond a closed set of concepts (ii) inherently multi-modal, enabling a diverse range of possible queries to the 3D map, from language, to images, to audio, to 3D geometry, all working in concert. **ConceptFusion** leverages the

open-set capabilities of today’s foundation models that have been pre-trained on internet-scale data to reason about concepts across modalities such as natural language, images, and audio. We demonstrate that pixel-aligned open-set features can be fused into 3D maps via traditional SLAM and multi-view fusion approaches. This enables effective zero-shot spatial reasoning, not needing any additional training or finetuning, and retains long-tailed concepts better than supervised approaches, outperforming them by more than 40% margin on 3D IoU. We extensively evaluate **ConceptFusion** on a number of real-world datasets, simulated home environments, a real-world tabletop manipulation task, and an autonomous driving platform. We showcase new avenues for blending foundation models with 3D open-set multimodal mapping. We encourage the reader to view the demos on our project page: <https://concept-fusion.github.io/>

^{*}Co-second authors

[†]Work done prior to joining Amazon

[‡]Work done prior to joining Matician

1. Introduction

One of the major catalysts of continued progress in 3D spatial perception [1, 2, 3, 4] has been the advent of deep convolutional networks trained on large datasets of images. Most of these advancements have relied on a *closed-set* of concepts, a fixed set of labels available at training time. More recently, however, deep learning is witnessing yet another wave of advancements, this time with the emergence of increasingly larger and multimodal models trained on internet-scale data comprising billions of images, text, and audio [5]. Foundation models like CLIP [6], DINO [7], AudioCLIP [8], and their variants have shown impressive performance on *open-set* scenarios, where the concepts of interest are supplied only at inference time. In this work, we bridge the gap between the rich open-set capabilities enabled by large foundation models and the semantic reasoning abilities expected of futuristic 3D mapping systems.

To be as broadly applicable as possible to a diverse set of robotics tasks, map representations need to be usable zero-shot (i.e. without the need to be retrained each time reasoning capabilities for a new task are desired), and must also possess the following two capabilities: first, **3D maps should be open-set**; they should capture a large variety of concepts (orders of magnitude more than existing systems), and at varying levels of detail. For example, the concept “*can of soda*” could equivalently be “*something to drink*” or a “<particular brand of soda>” or “*a refreshment*”. Second, **3D maps should be multimodal**; they should be queryable using as many modalities as robots or end-users can leverage. For instance, the search for a particular object in a map should work equally well if the query is a single word (e.g. “*soda*”), a longer sentence with additional context (e.g. “*is there a can of soda on the kitchen table?*”), an image of a soda can, or just the ‘pop’ sound associated with opening a soda can.

Foundation models possess some of the desired traits needed to achieve open-set, multimodal representations, but are not directly applicable to 3D mapping. This major limitation exists because most foundation models consume images (e.g., CLIP [6], ALIGN [9], AudioCLIP [8]) and produce only a single vector encoding of the entire image in an embedding space. On the other hand, recent approaches trained specifically to align foundation features to 2D pixels *forget* a large number of concepts during finetuning [10] (see Fig. 4). This does not allow for the level of precise (pixel-level or object-level) reasoning robotic perception systems need across a wide range of concepts, particularly for interaction with the external 3D world (e.g., navigation, manipulation).

To this end, we propose *ConceptFusion*; an open-set and multimodal 3D mapping technique that blends advances in foundation models for images, language, and audio, with advances in dense 3D reconstruction and simultaneous localization and mapping (SLAM). We demonstrate that pixel-level foundation features may be fused into 3D maps by leveraging precisely the same surface fusion techniques as for fusing depth or color information into a 3D map. Crucially, we show that this approach is conceptually simple, principled, and effective even in the zero-shot setting (requiring no additional

training or finetuning of foundation model features). In addition, these features can be queried using computationally efficient vector similarity metrics. Our key contributions are the following:

- An approach to open-set multimodal 3D mapping that constructs map representations queryable by text, image, audio, and click queries in a zero-shot manner.
- A novel mechanism to compute pixel-aligned (local) features from foundation models that can only generate image-level (global) feature vectors. This is a key prerequisite for 3D mapping, and our approach captures long-tailed concepts significantly better than supervised or finetuned counterparts, outperforming them by a large margin ($> 40\%$ mIoU).
- A new RGB-D dataset, UnCoCo, to evaluate open-set multimodal 3D mapping. UnCoCo comprises 78 common household/office objects tagged with more than 500K queries across modalities.

We evaluate *ConceptFusion* on multiple real-world datasets and tasks, including searching for objects in the real world and simulated home environments, robot manipulation tasks, and autonomous driving.

2. Related work

Foundation models [5] are trained on vast quantities of data in a self-supervised fashion and accomplish a wide array of tasks, without the need for additional re-training or finetuning. Image-language aligned models like CLIP [6] and ALIGN [9] encode text and images into a shared concept space, and have driven progress in several open-set tasks [11, 12, 13, 14, 15, 16, 17, 18]. They have also been extended to other modalities, such as audio [8, 19] and video [20]. In a similar vein, features from image-only foundation models like DINO [7] have been employed as drop-in solutions for reasoning about visual concepts, such as classification, detection, segmentation, keypoint estimation [21, 22].

Models like CLIP only align concepts to text or images as a whole; and not to image pixels or regions [23]. This has led to a series of approaches that extract **pixel-aligned features** [24, 25, 18, 26, 23] that address poor localization quality by training or fine-tuning a pixel-aligned model over a labeled dataset. LSeg [24] leverages pixel-level labels from segmentation datasets, while OpenSeg [18], OVSeg [26], and RegionCLIP [23] use image-caption datasets and employ region-word grounding. However, as we show in Sec. 3.2, while this improves performance over concepts present in the finetuning dataset, these models do poorly at recalling concepts infrequent in the label set used for finetuning. To retain knowledge of all concepts captured by the underlying CLIP model, MaskCLIP [27] proposes a zero-shot approach that combines self-attention with engineered 1×1 convolution layers for open-set segmentation. However, it struggles with delineating object boundaries, and long-tailed concepts, as we show in Sec 3.2 and Sec 4.1. Our proposed pixel-aligned features mitigate all of the aforementioned issues by operating over region proposals and by aligning features computed across regions, enabling a variety of spatial reasoning applications.

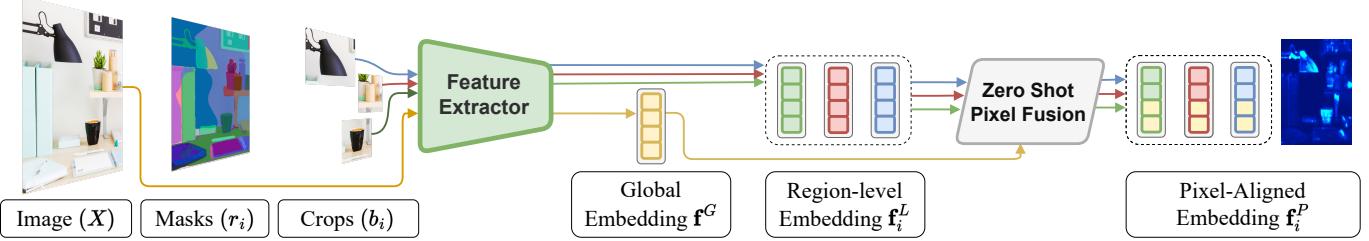


Figure 2: *ConceptFusion* constructs **pixel-aligned features** \mathbf{f}^P by: processing input images to generate generic (class-agnostic) object masks (regions) r_i , computing a bounding box for each region and extracting a local feature vector \mathbf{f}_i^L , computing a global feature \mathbf{f}^G for the input image as a whole, and fusing the region-specific features with global features as illustrated in Fig. 3 and described in Sec. 3.1.

To fuse our pixel-aligned features into a 3D map, we leverage approaches pioneered by the **dense 3D mapping** community. Approaches to dense SLAM estimate camera motion, scene geometry, and optionally, color from RGB(-D) images. At the core of nearly all of these methods is the volumetric fusion technique from Curless and Levoy [28], which has been adapted for real-time incremental capture of surface geometry and color [29, 30, 31, 32, 33]. In this work, we extend this line of work to additionally fuse pixel-aligned foundation features to 3D maps in a conceptually simple and computationally efficient manner. Noteworthy research efforts in bridging the gap between *online* 3D mapping and open-set concept acquisition include [34, 35]. In contrast, our work leverages large foundation features learned *offline*, over billions of data samples.

Recent approaches exploring **foundation models for robotics** have demonstrated impressive open-set interaction abilities [36, 37, 38, 39]. While most of these approaches focus on planning and control, we provide a complementary perspective; that of perception systems that build explicit models of scenes and are able to query for concepts. In Sec. 5, we present our outlook for how both of these classes of approaches can augment each other.

Concurrently, multiple approaches are beginning to explore the interfaces between foundation features and scene understanding. VLMaps [40], LM-Nav [41], ZSON [42], CoWs [43], and NLMap-Saycan [44] leverage the open-set features from pixel-aligned LSeg [24] models for robot navigation. CLIP-Fields encodes a 3D map and pixel (or region)-aligned foundation features (LSeg [24], Detic [45], Sentence-BERT [46]) into a compact neural network [47]. This scene-specific neural network acts as a queryable database which aligns image and language embeddings with 3D scene points, and can be applied to open-set queries specified in language. More recently, OpenScene [48] applied pixel-aligned LSeg [24] and OpenSeg [18] to open-set 3D segmentation; demonstrating that features from these 2D extractors can be distilled into neural networks operating over 3D data. Ding *et al.* [49] additionally distill features from more complex 2D open-set tasks, such as image captioning, into a model that consumes 3D data. In our work, we show that it is possible to construct 2D and 3D features zero-shot, without compromising performance on typical concepts. This zero-shot capability is also a key enabler in terms of our superior performance on long-tailed concepts and complex queries, and alignment with other modalities.

3. The *ConceptFusion* approach

The open-set multimodal 3D mapping problem: Given a sequence of image (and depth) observations of an environment $\mathcal{I} = \{I_t\}$ ($t \in \{0 \dots T\}$), we build an open-set multimodal 3D map \mathcal{M} . This map is *queryable* for concepts from multiple modalities, using query vectors $q_{\text{mode}} \in \mathbb{R}^d$. Multidimensional signals such as images, text, audio, and clicks can be encoded into such a vector space using a modality-specific encoder (a foundation model) $\mathcal{F}_{\text{mode}}$.

We first present a general feature fusion technique, extending traditional dense mapping approaches to incorporate per-pixel features in addition to color and depth information. We then present our algorithm to compute pixel-aligned features zero-shot from off-the-shelf foundation models (such as CLIP [6], AudioCLIP [8], and variants).

3.1. Fusing pixel-aligned foundation features to 3D

Map representation: We represent our open-set multimodal 3D map \mathcal{M} as an unordered set of points (indexed by k), each with the following attributes: (a) a vertex position $\bar{\mathbf{v}}_k \in \mathbb{R}^3$, (b) a normal vector $\bar{\mathbf{n}}_k \in \mathbb{R}^3$, (c) a confidence count $\bar{c}_k \in \mathbb{R}$, (d) a 3D color vector (optional), and (e) a *concept* vector \mathbf{f}_k^P enabling open-ended querying.

Frame preprocessing: Each incoming frame I_t (comprising a color image C_t and a depth image D_t) is preprocessed to compute vertex-normal maps (V_t, N_t) and camera pose estimates P_t [30, 50]. Additionally, as described in Sec. 3.2, we compute the semantic context embedding $\mathbf{f}_{u,v,t}^P \in \mathbf{f}_{X_t}^P$ for each pixel in the input image X_t .

Feature fusion: We fuse $\mathbf{f}_{u,v,t}^P$ and X_t into the global map following a 3D reconstruction pipeline [30]. First, vertex and normal maps are mapped to the global (map) coordinate frame using the camera pose P_t . We then filter out points with noisy depth values by following the depth map fusion procedure outlined in [30]. The remaining points are fused into the global map \mathcal{M} . A key departure from dense mapping approaches is the fusion of *concept vectors* $\mathbf{f}_{u,v,t}^P$ in addition to depth (and optionally, color). For each pixel $(u, v)_t$ in the image X_t that have a corresponding point p_k in \mathcal{M} , we integrate the features using the following scheme.

$$\mathbf{f}_{k,t}^P \leftarrow \frac{\bar{c}_k \mathbf{f}_{k,t-1}^P + \alpha \mathbf{f}_{u,v,t}^P}{\bar{c}_k + \alpha} \quad (1)$$

$$\bar{c}_k \leftarrow \bar{c}_k + \alpha \quad (2)$$

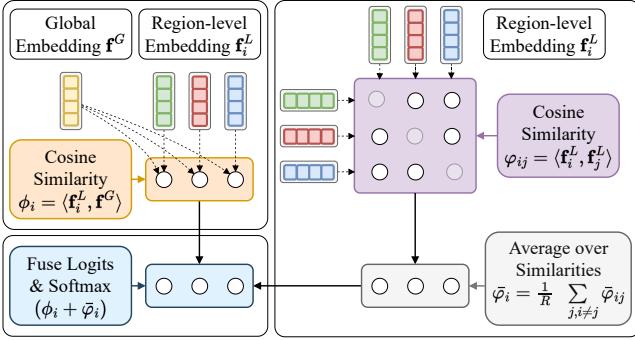


Figure 3: For each image, the global (\mathbf{f}^G) and local (\mathbf{f}^L) features are fused to obtain our **pixel-aligned** features (\mathbf{f}^P). **Top-Left:** We first compute cosine similarities between each local feature (\mathbf{f}^L) with the global feature (\mathbf{f}^G). **Right:** We compute an inter-feature similarity matrix, and compute the average similarity of each local feature to every other local feature, denoted $\bar{\varphi}_i$. **Bottom-Left:** We combine these similarities to produce weights for fusing \mathbf{f}^G and \mathbf{f}^L to obtain pixel-aligned features \mathbf{f}^P . See Sec. 3.2 for details.

where $\alpha = e^{-\gamma^2/2\sigma^2}$ is the confidence assigned to each pixel-grounded feature assigned to the vertex being aggregated, γ is the radial distance, and $\sigma = 0.6$ is a scaling term. We find empirically that having a confidence value based on the normalized radial distance to the camera center, similar to [30, 28] works well. We refer to the appendix for hyperparameter values and more details.

3.2. Computing pixel-aligned features

Although some approaches have adapted foundation models like CLIP [24, 18, 51], these require additional training on labelled image-text data. This results in pixel-alignment and stronger performance on in-dataset concepts, but we observe (see Fig. 4) that the models struggle with fine-grained concepts absent in the finetuning datasets. To mitigate this, we introduce a novel mechanism to construct pixel-aligned features that combine global (image-level) context encapsulated in models like CLIP, with local (region-level) information.

Overview: Fig. 2 gives us a broad overview of this section. Given an input image $X \in \mathbb{R}^{3 \times H \times W}$, our method uses a foundation model \mathcal{F} as a feature extractor to produce three types of embeddings, which we refer to as global (\mathbf{f}^G), pixel-aligned (\mathbf{f}^P), and local (\mathbf{f}^L). First, the global embedding $\mathbf{f}^G = \mathcal{F}(X)$ is simply the embedding of the entire image.

Local embeddings: We employ a *universal* instance segmentation model [52] to produce a set of R class-agnostic mask proposals (corresponding to image-regions that could potentially contain objects). Regions may not necessarily be contiguous, and we produce 100 mask proposals per image. Next, for each region $r_i \in \mathbf{R}$, we define a bounding box $b_i = \text{bbox}(r_i)$ around it. The bounding boxes are fed through the model \mathcal{F} to obtain local embeddings $\mathbf{f}_i^L = \mathcal{F}(b_i)$.

Fusing local and global features: The pixel-aligned embedding for each region is a weighted combination of the global embedding and the corresponding local embedding. The weight for each local embedding is based on its similarity with the global embedding \mathbf{f}^G , as well as with other embed-

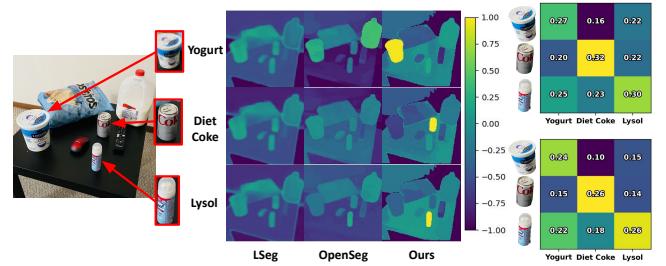


Figure 4: Our approach to computing **pixel-aligned features** is adept at capturing long-tailed and fine-grained concepts. The plots to the right show the similarity scores between the embeddings of the cropped image regions corresponding to *diet coke*, *lysol*, and *yogurt* and their text embeddings, predicted by the base CLIP model used by LSeg and OpenSeg respectively. This implies that the base CLIP models know these concepts, yet, as can be seen on the tiled plots (center), LSeg and OpenSeg are not able to retrieve these concepts; they forget the concepts when finetuned. On the other hand, our zero-shot pixel-alignment approach does not suffer this drawback, and clearly delineates the corresponding pixels.

dings \mathbf{f}_i^L . Concretely, we compute the cosine similarity

$$\phi_i = \langle \mathbf{f}_i^L, \mathbf{f}^G \rangle = \frac{(\mathbf{f}_i^L)^T \mathbf{f}^G}{\|\mathbf{f}_i^L\| \|\mathbf{f}^G\| + \epsilon} \quad (3)$$

between the local feature \mathbf{f}_i^L and the global feature \mathbf{f}^G ; and a matrix of cosine similarities between all pairs of local embeddings as $\varphi_{ij} = \langle \mathbf{f}_i^L, \mathbf{f}_j^L \rangle : \forall i, j \text{ such that } r_i, r_j \in \mathbf{R}$. Next, for each local embedding \mathbf{f}_i^L , we compute its average similarity to all other local embeddings $\bar{\varphi}_i$. This accounts for the *uniqueness* of the region for \mathbf{f}_i^L in the image.

$$\bar{\varphi}_i = \frac{1}{R} \sum_{j=1, j \neq i}^R \varphi_{ij} \quad (4)$$

We combine the two similarities above to compute the mixing weight $w_i \in [0, 1]$ (with a temperature τ , set to 1 in all reported results).

$$w_i = \frac{\exp\left(\frac{\phi_i + \bar{\varphi}_i}{\tau}\right)}{\sum_{i=1}^R \exp\left(\frac{\phi_i + \bar{\varphi}_i}{\tau}\right)} \quad (5)$$

Finally, the pixel-aligned feature for each region r_i is

$$\mathbf{f}_i^P = w_i \mathbf{f}^G + (1 - w_i) \mathbf{f}_i^L \quad (6)$$

which is normalized and mapped to the pixels (u, v) in r_i . We allow each pixel to belong to multiple regions; the corresponding pixel-aligned embedding $\mathbf{f}_{u,v}^P$ is normalized once it accumulates features \mathbf{f}_i^P from regions r_i .

Capturing long-tailed concepts: We find that our pixel-aligned embeddings capture fine-grained and long-tailed concepts significantly better than approaches like LSeg [24] and OpenSeg [18], which align CLIP features to pixels by training over smaller labelled datasets. We illustrate this in Fig. 4.



Figure 5: 3D spatial reasoning abilities: A key benefit of lifting foundation features to 3D is the ability to reason about spatial attributes. For example, the query “*how far is the refrigerator from the television*”. gets parsed by our LLM-parser into the 3DSC `howfar(refrigerator, television)`. The centroid of the point set returned by the query term `refrigerator` and `television` are shown as blue circles, and the estimated distance between them (6.303 metres) as a straight line.

We observe (right panel) that the underlying backbone CLIP models used by both LSeg and CLIP know the concepts *diet coke*, *lysol*, and *yogurt*; however the finetuned (pixel-aligned) models do not. This is due to forgetting phenomenon when finetuning CLIP-like models, as corroborated in [10]. LSeg and OpenSeg need to be finetuned on datasets with limited concepts, in order to obtain the ability of segmentation. However, this finetuning process harms their zero-shot ability to generalize to long-tailed and fine-grained concepts. In contrast, *ConceptFusion* presents a new way of mapping foundation features to pixels and 3D points, and therefore it remains zero-shot and accurately aligns long-tailed concepts to the corresponding pixels, as shown by the middle plots in Fig. 4.

3.3. Multimodal querying over 3D feature-fused maps

The 3D maps reconstructed by *ConceptFusion* can be queried using multiple modalities. Given a query q_{mode} , and a map with fused concepts \mathbf{f}_k^P , we compute a per-point score $s_k \in [-1, 1]$ as the cosine similarity defined as $s_k = \langle f_k, q_{\text{mode}} \rangle$. The computation of q_{mode} changes depending on the modality used for querying.

1. **Text query:** q_{text} is computed using the corresponding CLIP text encoder $\mathcal{F}_{\text{text}}$.
2. **Click query:** q_{click} is taken to be the fused feature vector \mathbf{f}_k^P at the clicked point.
3. **Image query:** q_{image} is computed as the image-level CLIP embedding of the query image.
4. **Audio query:** q_{audio} is the AudioCLIP [8] embedding of the query sound clip.

The similarities can then be post-processed by thresholding based on score, non-maxima suppression, and (optionally) clustering to produce 3D regions of interest.



Figure 6: Sample sequences from the **UnCoCo dataset** we captured to evaluate long-tailed reasoning over open-set multimodal 3D maps. To the right, we show sample reconstructions and semantic annotations over two sub-sequences. For each scene, UnCoCo provides 2D and 3D segmentation masks, and text, image, click, and audio queries.

3.4. Building complex 3D spatial query modules

Unique capabilities unlocked by fusing features into 3D space include the ability to reason about objects that were never co-observed in an image, and the ability to reason about spatial attributes (such as relative positions, orientations, support, containment, etc.) accessible only from 3D representations. To this end, we leverage the computed similarity scores to build a set of *3D spatial comparator* (3DSC) modules that may be further composed to recover more complex attributes.

Our set of 3DSCs all take on the relation signature $\text{RELATION}(\text{QUERY}_a, \text{QUERY}_b)$ and return a scalar or boolean value as appropriate. The complete set of 3DSCs includes (see supplementary material for more details)

1. The `HOWFAR(q_a, q_b)` 3DSC returns the distance of objects referenced by queries q_a and q_b respectively.
2. The boolean 3DSCs `ISTOTHERIGHT(q_a, q_b)`, `ISTOTHELEFT(q_a, q_b)`, `ONTOPOF(q_a, q_b)`, `UNDER(q_a, q_b)` return TRUE or FALSE depending on whether or not the object referenced by queries q_a and q_b satisfy the appropriate spatial relationship (relative to a specified viewing direction).

In Sec. 5, we optionally adopt a large language model [53] for parsing language queries to an appropriate composition of 3DSCs. For instance, the query “*what is the distance between the refrigerator and the television?*” is parsed into `HOWFAR(refrigerator, television)`. This is depicted in Fig. 5.

However, for all other results presented in this paper—unless otherwise specified—the language queries are directly fed into the CLIP text encoder without any preprocessing.

Implementation details: Our feature fusion algorithm is implemented on top of the ∇ SLAM [54] dense SLAM system, as this was one of the few implementations of the Point-Fusion algorithm [30], and for its convenience of interfacing with PyTorch for computing and accessing foundation features. For generating class-agnostic (generic) object masks, we use the Mask2Former [55] models for instance segmentation and generate 100 mask proposals per image. Our odometry and mapping approaches run at frame-rate (15 Hz). The pixel-aligned feature extraction processes run offline (10–15 seconds / image) on an NVIDIA RTX 3090 GPU.

		3D mIoU	IoU >0.15	IoU >0.25	IoU >0.5
Supervised	LSeg-3D	0.128	25%	16.66%	9.72%
	OpenSeg-3D	0.289	43.05%	36.11%	27.78%
	MaskCLIP-3D	0.091	25.97%	9.09%	1.30%
	<i>ConceptFusion</i>	0.446	77.78%	69.44%	45.83%

Table 1: Text-query based object localization performance on UnCoCo – the *structured* subset. In each column, a higher value corresponds to superior performance.

		3D mIoU	IoU >0.15	IoU >0.25	IoU >0.5
Supervised	LSeg-3D	0.134	26.88%	21.51%	9.68%
	OpenSeg-3D	0.112	23.66%	18.28%	8.60%
Zero-Shot	MaskCLIP-3D	0.094	21.51%	11.83%	4.30%
	<i>ConceptFusion</i>	0.331	54.84%	51.61%	31.18%

Table 3: Image-query based detection performance on UnCoCo – the *structured* subset. Results averaged over 3 trials.

4. Case studies

We design a systematic experimental study to investigate the following questions:

1. How do open-set multimodal 3D maps fare when queried using text, images, clicks, or audio?
2. How do we leverage the rich concept space embedded to 3D for spatial reasoning?
3. How well does *ConceptFusion* work on real-world robotics tasks?
4. What previously infeasible downstream use-cases can *ConceptFusion* enable?

Experimental setup: Our experimental benchmark comprises of sequences from multiple publicly available datasets, and sequences we collect. The benchmark comprises 20 indoor (apartment-scale) scenes from ScanNet [56, 57], Replica [58], and self-captured sequences; 5 outdoor (urban driving) scenes; 20 indoor (tabletop) scenes with common household products (UnCoCo); and 5 synthetic scenes from the ICL [59] and iTHOR benchmarks [60].

Approaches evaluated: Since there is no prior work on constructing open-set multimodal maps of the kind we build with *ConceptFusion*, we make a best-effort comparison with concurrent work in this space. Approaches such as VL-Maps [40], NLMap-SayCan [44], CoWs [43, 61], CLIP-Fields [47] leverage LSeg [24]; while OpenScene [48] experiments with both LSeg [24] and OpenSeg [18]. We therefore implement two baseline approaches that leverage LSeg and OpenSeg features respectively, and apply our feature fusion technique to obtain open-set 3D maps. *We refer to these baselines as LSeg-3D and OpenSeg-3D respectively.* Additionally, to compare with a state-of-the-art zero-shot segmentation approach, we also implement MaskCLIP-3D, which fuses per-pixel MaskCLIP [27] features into a 3D map.

4.1. Multimodal queries on the UnCoCo dataset

To the best of our knowledge, there is no existing system that supports queries as diverse and in as many modes as *ConceptFusion*; and consequently, there are no publicly available dataset to evaluate such a system on.

		3D mIoU	IoU >0.15	IoU >0.25	IoU >0.5
Supervised	LSeg-3D	0.122	31.45%	20.65%	5.65%
	OpenSeg-3D	0.153	27.26%	21.94%	11.29%
Zero-Shot	MaskCLIP-3D	0.092	20.63%	11.88%	3.06%
	<i>ConceptFusion</i>	0.378	70.16%	59.52%	34.03%

Table 2: Text-query based detection performance on UnCoCo – the *unstructured* subset. Results averaged over 20 trials. In each column, a higher value corresponds to superior performance.

		Accuracy (%)	IoU
source-ambiguous	Random	7.14%	N/A
	AudioCLIP [8]	23.81%	N/A
	<i>ConceptFusion</i>	64.29%	0.287
ecological	Random	5.56%	N/A
	AudioCLIP [8]	22.22%	N/A
	<i>ConceptFusion</i>	66.67%	0.301

Table 4: Audio-query based detection and classification performance on UnCoCo.

UnCoCo: To overcome this gap, we captured our own dataset; which we call *UnCoCo*, short for *Uncommon contexts for Common Concepts*. This real-world dataset comprises 3D scans of 78 commonly found household and office objects on a tabletop surface (see Fig. 6). There are 20 RGB-D image sequences in all, comprising a total of 12075 color and depth image pairs. For each image, we provide per-object 2D instance segmentation masks, and for each scene, a corresponding 3D segmentation mask. Importantly, UnCoCo supports a diverse set of query modalities – text, click, image, and audio. For each of these query modalities, we provide a *structured* and an *unstructured* set. The structured set comprises carefully curated and vetted queries, whereas the (much larger) unstructured set scrapes query text, and images from web-scale data. Each image in this dataset has 3-5 objects; each object has one structured text query, and 5-40 unstructured text queries (freeform queries, crowdsourced from human annotators); up to 20 structured click queries and up to 100-2000 unstructured click queries; 10 structured image queries and 10-100 unstructured image queries; and 1-5 structured audio queries. With a little over 500000 queries across modalities, UnCoCo is the only dataset to date supporting multimodal 3D query evaluations on commonly found objects.

Text query: We evaluate text-query based object localization performance on 3D maps, on the UnCoCo dataset. This task is extremely challenging due to the versatility of objects present in the dataset, ranging from extremely small objects (e.g., a 4-gram sachet of sugar, whiteboard markers), to thin objects (e.g., face masks, compact discs), to nonconvex geometries (e.g., a whisk, lego block constructions, shells). We evaluate two state-of-the-art per-pixel CLIP-aligned feature extractors in LSeg [24] and OpenSeg [18], which require additional training over a large labelled dataset; and MaskCLIP [27] – the current state-of-the-art approach for extracting zero-shot per-pixel labels based on a text prompt. Results are shown in Table 1. For each technique evaluated, we report the 3D mean intersection-over-union (IoU) metric, and also detection accuracies at IoU thresholds of 0.15, 0.25, and 0.5. We see that *ConceptFusion* outperforms all other ap-

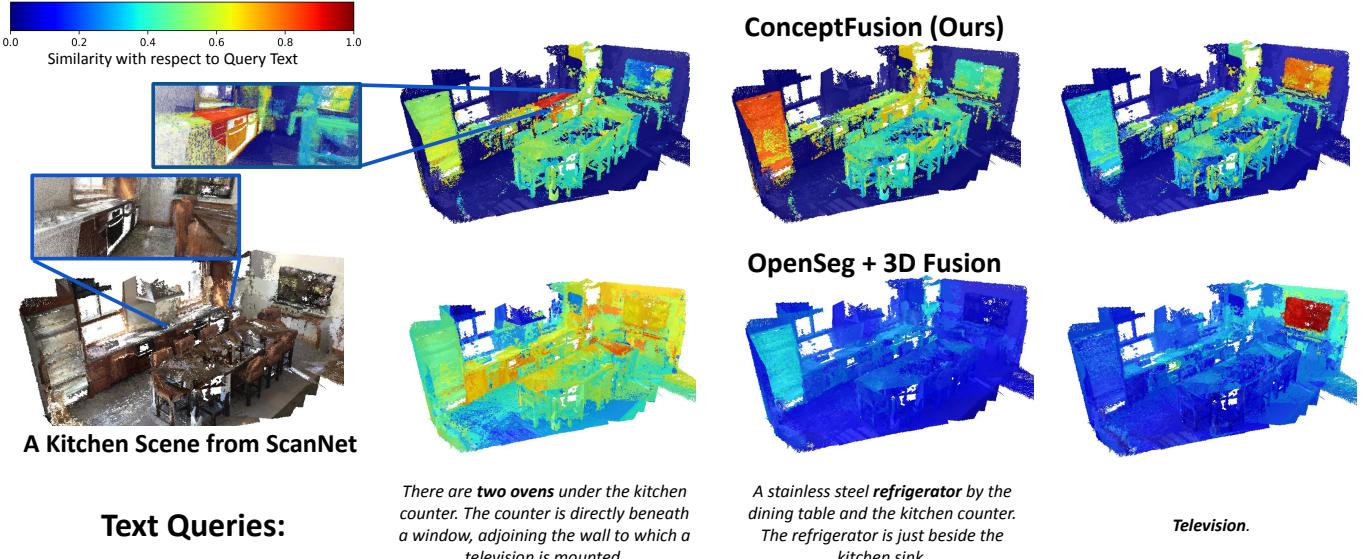


Figure 7: **Text queries over ScanNet [56]**: *ConceptFusion* is able to handle long-form text queries and accurately localize objects referenced by the query. In the first two scenarios, OpenSeg [18] is distracted by the presence of several confounding attributes (*kitchen counter*, *window*, *television*). The third scenario shows a single world query (*television*) that is part of the COCO Captions [62] dataset used to train OpenSeg, placing it at an advantage (and hence resulting in a more discriminable heatmap). *ConceptFusion*, nonetheless, accurately assigns the highest response to the map points representing the television. In each query, the referenced object is **boldfaced**.

proaches by a significant margin. We attribute this to two key characteristics of *ConceptFusion*. First, *ConceptFusion* operates on the unmodified CLIP feature space, whereas approaches like LSeg and OpenSeg specialize to the datasets they are finetuned on and end up gradually forgetting concepts that are infrequent on the finetuning set. Second, *ConceptFusion* features efficiently combine global (image-level) features with local (region-level) context; providing a rich pixel-level (and subsequently point-level) grounding. We also observe a similar trend over the unstructured text query set, as reported in Table 2. *ConceptFusion* exhibits more graceful performance degradation to unstructured queries (long sentences).

Image query: In Table 3, we report results for the scenario where the query concept is presented in the form of an image. For instance, the goal of finding a *can of soda* in a 3D scene is specified by providing a randomly picked image of a soda can from the web. Here, we again observe that *ConceptFusion* outperforms other finetuned foundation models by a significant margin in terms of both 3D mIoU and detection accuracy. Interestingly, we note that performances of most approaches across the text and image modalities remain consistent; as evident from the mIoU and detection metrics.

Audio query: A unique capability of our approach is to localize 3D objects based on audio queries. We conduct experiments on the UnCoCo dataset, which contains two classes of audio queries. *Source-ambiguous* queries comprise sounds that are caused due to material properties and geometries of objects and commonly involve object motions including scratching, scraping, rolling, crushing, tearing, etc. *Ecological* queries comprise sounds that are unique to an object (or category) – these include sounds like spraying, drilling, stirring, knocking, clicking, etc. To compensate for the lack of open-set sound source localization baselines, we implement a

privileged-information baseline by providing AudioCLIP [8] with a set of ground-truth instance boxes per image; and using the computed similarity scores to rank the relevance of each box to the query audio. Results are presented in Table 4. However, we notice that, using purely local context (as with the AudioCLIP baseline) is not performant enough; and that *ConceptFusion* features, which fuse global and local contexts, again perform better than purely-local features.

4.2. Open-Set semantic segmentation on other datasets

We also evaluate semantic segmentation performance on existing datasets with well-known concepts (classes either directly or indirectly accessible via COCO Captions [62], a dataset used for finetuning by other pixel-aligned models). In particular, we provide results on validation subsets of the ScanNet [56], Replica [58], and the SemanticKITTI [63] datasets and present results in Table 5 (refer to our appendix for further details). Of the approaches presented here, LSeg requires per-pixel CLIP features as labels, OpenSeg leverages per-image captions for labels, and CLIPSeg trains a shallow decoder atop the CLIP image encoder. MaskCLIP is the closest zero-shot baseline; we outperform it by a large margin.

4.3. Experiments on real robotic systems

Our experiments on real robotic systems may also be accessed in a video format on our [companion website](#).

Zero-shot tabletop rearrangement: To evaluate the applicability of *ConceptFusion* to real-world robotic interaction, we conduct experiments on a zero-shot tabletop rearrangement task with a UR5e manipulator and an Intel Realsense D415 RGB-D camera. The task involves a workspace (here a tabletop) with a few previously unseen objects in it. In some

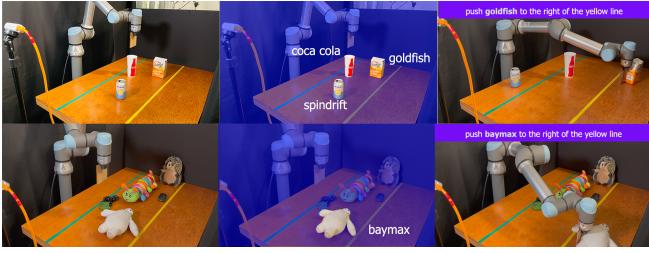


Figure 8: Real-world **tabletop rearrangement** experiments. The robot is provided with rearrangement goals involving novel objects. (*Top row*) *push goldfish to the right of the yellow line*, where *goldfish* refers to the brandname of the pack of Cheddar snack. (*Bottom row*) *push baymax to the right of the yellow line*, where *baymax* refers to the plush toy depicting the famous Disney character.

Table 5: **Open-set semantic segmentation** on ScanNet [56], Replica [58], and SemanticKITTI [63]. Queries provided in the form of text labels. **Priv.** (shaded area) indicates privileged approaches where off-the-shelf CLIP models are finetuned specifically for semantic segmentation. **ZS** (unshaded area) evaluates zero-shot approaches. We outperform the closest zero-shot approach, MaskCLIP [27] by a large margin across datasets. Furthermore, *ConceptFusion* is competitive to privileged baselines for this task.

		ScanNet	Replica	Semantic KITTI	
		mAcc	f-mIoU	mAcc	f-mIoU
Priv.	LSeg [24]	0.70	0.63	0.52	0.33
	OpenSeg [18]	0.63	0.62	0.54	0.41
	CLIPSeg (rd64-uni) [51]	0.41	0.34	0.32	0.23
	CLIPSeg (rd16-uni) [51]	0.41	0.36	0.40	0.28
	CLIPSeg (rd64-uni-refined) [51]	0.23	0.24	0.13	0.13
	MaskCLIP [27]	0.24	0.28	0.01	0.05
ZS	Mask2Former + Global CLIP feat	0.35	0.48	0.13	0.10
	<i>ConceptFusion</i>	0.63	0.58	0.31	0.24
				0.79	0.78

trials, the object set also includes distractors placed to hamper perception and/or manipulation planning. Two sides of the workspace (see Fig. 8) are tagged *left* and *right* respectively (areas on either side of the table, as indicated by the green and yellow lines). For each set of objects, a goal instruction is specified in the form of a natural language command. For instance, the two scenarios in Fig. 8 correspond to the commands *spindrift to the left; goldfish to the right; coca cola to the left* (top row) and *baymax to the right* (bottom row). This task challenges the perception capabilities of our system in multiple ways. First, *ConceptFusion* must identify the correct object corresponding to the text query. Second, the object localization must be precise, to allow the end effector to execute a successful motion plan; moving the object to its goal location, while accounting for distractors. After *ConceptFusion* identifies and segments the object of interest, we estimate the object centroid as the median of the region of interest obtained over the depth image. The robot arm then computes a motion plan (using the AIRobot library [64]) to push the object to the specified target region (i.e., to the *left* or *right*).

Text-driven autonomous navigation: We employ *ConceptFusion* in open-set autonomous navigation for a self-driving platform. Given a feature-fused map of an environment, we search the map for a best-match destination to a text query, and navigate autonomously to the location thus chosen. We used a drive-by-wire autonomous vehicle equipped

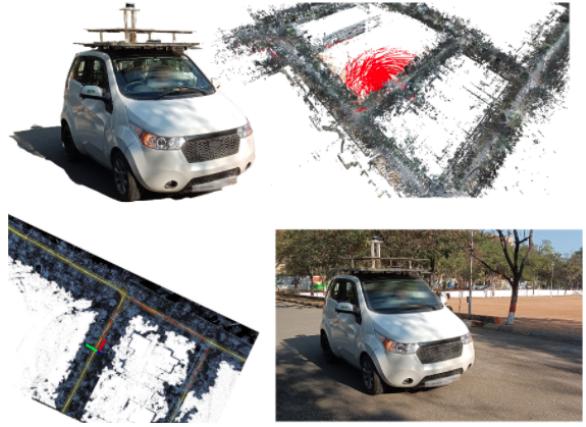


Figure 9: Real-world **autonomous navigation** experiments. (Left to right; top to bottom) Autonomous drive-by-wire platform deployed; pointcloud map of the environment with the response to the open-set text-query "football field" (shown in red); path found to the football field (shown in red); car successfully navigates to the destination autonomously. See our webpage for more results.

with a calibrated stereo camera and lidar to reconstruct a map of a 320000 square yard (4000 sq. m.) urban area. The 3D map points are used by a lidar odometry algorithm (LeGO-LOAM [65]) for global localization, and the fused features are used for open vocabulary queries to generate global waypoints. Other components of the autonomy stack include a shortest-path global planner, a Frenet [66] local planner for obstacle avoidance, and a Stanley-Controller [67] for trajectory tracking. This allows us to navigate to open-set goals specified as natural language queries. In our field trials, we evaluated multiple goals, such as *garbage bin, roundabout, entrance gate*. Fig. 9 illustrates this process over a text query *football field*. Please refer to our project page and supplementary video for viewing this demo.

5. Outlook

In this section, we provide perspectives on other emerging directions in foundation models for robotics, and how open-set multimodal 3D maps can augment and expand the capabilities of these models.

3D spatial reasoning abilities: We also evaluate the unique ability of our queryable maps to reason about object spatial relationships in 3D. We generate a set of 100 (natural language) spatial reasoning queries over 5 scenes from the ScanRefer validation set. The queries are split into 4 sets (25 queries per set), each designed to evaluate a specific type of 3D spatial relationship – distances, relative positions (e.g., *to the left of*), support (e.g., *on top of*), and containment. These queries are chosen such that (a) each query references two objects, and (b) two objects are co-observed in a single image (candidates are proposed using a distance-based threshold; and validated using image segmentation labels) for the distance and relative position query sets. The results are presented in Table 6, and compared against a baseline approach that uses only the pointcloud obtained by backpro-

ConceptFusion-Agent	GenericLLM-Agent
<p>User: I want to find some eggs. The rooms I see in the house are “bedroom”, “living room”, “kitchen”, “hallway”, and “yard”. Rank these in the decreasing order of likelihood for this task.</p> <p>LLM: 1. Kitchen 2. Hallway (if you have an egg storage container) 3. Yard (if you have chickens or hens)</p> <p>User: I am now in the kitchen. The objects I see in the room are “cabinets”, “sink”, “stove”, “trash can”, “microwave”, “fridge”, and “table”. Rank these in the decreasing order of likelihood of finding eggs.</p> <p>LLM: 1. Fridge 2. Trash can 3. Table 4. Cabinets 5. Sink 6. Stove 7. Microwave <SUCCESS></p>	<p>User: I want to find some eggs. List some places where I can find them in the house.</p> <p>LLM: 1. Grocery store 2. Local farmers market 3. Health food store 4. Restaurant 5. Deli</p> <p><FAIL></p>
 <p>(a) Agent spawns in house</p>	 <p>(b) LLM outputs “kitchen”</p>
 <p>(c) Agent goes to kitchen. LLM outputs “fridge”</p>	 <p>(d) “Eggs” found</p>
ConceptFusion-Agent	GenericLLM-Agent
<p>User: I want to find tissues. The rooms I see in the house are “bedroom”, “living room”, and “bathroom”. Rank these in the decreasing order of likelihood for this task.</p> <p>LLM: 1. Bathroom 2. Bedroom 3. Living room</p> <p>User: I am now in the bathroom. The objects I see in the room are “cabinet”, “bathtub”, “sink”, “toilet”. Rank these in the decreasing order of likelihood of finding tissue.</p> <p>LLM: 1. Cabinet 2. Sink 3. Bathtub 4. Toilet <SUCCESS></p>	<p>User: I want to find tissues List some places where I can find them in the house.</p> <p>LLM: 1. Bathroom 2. Bedroom 3. Kitchen 4. Living room 5. Garage</p> <p>User: I am now in the bathroom. Which object is most likely to contain a tissue?</p> <p>LLM: Bathroom <FAIL></p>
 <p>(a) Agent spawns in house</p>	 <p>(b) LLM outputs “bathroom”</p>
 <p>(c) Agent goes to kitchen. LLM outputs “cabinet”</p>	 <p>(d) “Tissue” found</p>

Figure 10: **Integration of a large-language model (LLM) based planner in-the-loop.** We illustrate two scenarios from the AI2-THOR [60] interactive household simulator. The GenericLLM-Agent fails to achieve the specified task since it does not have an explicit 3D map representation, deviating the LLM of the requisite context to accomplish the task. Our *ConceptFusion*-Agent modifies its response by querying the open-set map to include rooms and object locations in its immediate vicinity. The LLM receives this context to guide the agent towards success.

jecting a single RGB-D image (2.5D). This baseline performs poorly in the distance and relative position queries, where the queries require reasoning about objects that were never co-observed in image space. In the *support* and *containment* queries, the baseline shows better performance, as these are relations that can largely be gleaned from image observations alone. However, this approach still fails for room-level containment queries of type *is <OBJ> in <ROOM>*; which require additional context. Fig. 5 showcases an example of a distance query between two objects never co-observed in 2D.

Table 6: 3D lifting enables *ConceptFusion* to respond to spatial reasoning queries accurately. Each entry showcases success rates over specific query types (25 queries per type).

	Distance	Rel. position	Support	Containment
Random	24%	36%	52%	44%
2.5D (RGB-D image only)	32%	28%	76%	68%
<i>ConceptFusion</i>	84%	76%	96%	72%

Integration with large language models (LLMs): Systems such as SayCan [38] have demonstrated the viability of large-language models as queryable knowledge bases that may be used to generate task-level plans, to be executed by lower-level skills. With *ConceptFusion*, we can complement these task planning and reasoning abilities with a *perception system* that can interface via natural language (and other modalities). To demonstrate this, Fig. 10 illustrates two scenarios from the AI2-THOR [60] interactive household simulator. In each scenario, the robot is equipped with the task of finding an object of interest *that is not in its map*, because it is concealed within a receptacle. We investigate the behavior of an *GenericLLM-Agent* (a large language model [53]) and a *ConceptFusion-Agent* (*GenericLLM-Agent* equipped with a set of rooms and objects queried from our map). As seen in Fig. 10, while the *GenericLLM-Agent* is able to generate seemingly plausible subgoals to achieve the task, the lack of knowledge of the map inhibits its success. On the other hand, the *ConceptFusion*-agent successfully achieves both tasks by



Figure 11: The zero-shot nature of our approach allows integration with newer off-the-shelf foundation models without the need for finetuning. We demonstrate this ability by leveraging multilingual CLIP models, which enables us to query the same concept across multiple languages.

restricting the generative capabilities of the LLM to scene-specific contexts.

Multilingual abilities: Since *ConceptFusion* is zero-shot, it is directly applicable to any existing image-level foundation model, without requiring additional finetuning. To demonstrate this, we compute pixel-aligned features from multilingual CLIP models, and this extends the capabilities of our system to interpret and execute queries in several other languages, as illustrated in Fig. 11.

6. Conclusion

In this work, we presented *ConceptFusion* as an effective solution to the open-set multimodal 3D mapping problem. The zero-shot nature of our method enables reasoning over a significantly broad range of concepts; leveraging off-the-shelf foundation features for open-set perception. We evaluate our approach on in-house and established datasets, and on two real robotic systems (a manipulator and a self-driving vehicle). Our results indicate several promising avenues for integrating foundation models trained over web-scale data with traditional mapping systems to enable zero-shot, open-set, and multimodal perception.

Limitations: The key limitations of our method are three-fold. First, *ConceptFusion* operates over dense maps, often comprising millions of 3D points over an apartment-scale scene, and augments each point with high-dimensional concept embeddings, requiring large amounts of memory and inducing latency. Second, mask proposals used by our pixel-aligned feature computation method are currently limited to reasoning primarily about foreground objects. Third, we anticipate *ConceptFusion* to inherit the limitations and biases of LLMs [5, 68], warranting further investigations for potential harm as well as research into AI safety and alignment [69, 70].

Appendix

A1. Contribution statement

Krishna Murthy conceived the idea and led the project. Was responsible for much of the code development and wrote sections of the paper. Curated and annotated the UnCoCo dataset and helped with the tabletop robot experiments.

Ali Kuwajerwala: Organized the initial brainstorming session that kick-started this project. Collected parts of the real-world test data, curated image and text queries, implemented various features necessary for conducting experiments, created several graphics including the explainer video, and wrote sections of the paper.

Qiao Gu: Implemented key components of the system, including 3D fusion modules and 2D-3D semantic segmentation evaluation protocols. Ran important ablation experiments and contributed to the writing of the paper.

Mohd Obama: Conducted all of the autonomous driving experiments and contributed significantly to producing figures and videos for the paper.

Tao Chen: Led the tabletop rearrangement experiments and played an instrumental role in generating insights that led to the creation of the UnCoCo dataset.

Shuang Li: Led experiments integrating large language models as planners and contributed to the writing of the paper.

Ganesh Iyer: Made significant research contributions to the gradslam framework and follow-up work (prior to joining Amazon). Helped write sections of the paper.

Soroush Saryazdi: Key contributor to the gradSLAM library that *ConceptFusion* was built upon (work done prior to joining Matician).

Nikhil Keetha: Contributed to several negative results that helped shape the direction of the research and wrote sections of the paper.

Ayush Tewari and Celso de Melo: Participated in multiple brainstorming sessions that helped shape *ConceptFusion*.

Josh Tenenbaum: Provided valuable cognitive science perspectives and constructive skepticism, which informed the direction of the research and drew our attention towards potential issues (and interesting follow-up directions).

Madhava Krishna: Advised on the real-world autonomous driving experiments and suggested a crucial restructuring of the paper.

Liam Paull, Florian Shkurti, and Antonio Torralba: Involved in brainstorming and critical review throughout the project, always asking the hard questions that led to key research insights. Wrote and proofread sections of the paper.

A2. Acknowledgements

KM, JBT, and AT acknowledge funding support from ARL grant W911NF1820218. The authors thank Tamar Rott Shamam, Chuang Gan, Joanna Materzynska, Songyou Peng, and Toni Rosinol for discussions and useful feedback over the course of this project. We thank Pulkit Agrawal for providing us with a UR5e robot for real-world experiments.

A3. 3D feature fusion details

For **indoor datasets** (ScanNet [56], Replica [58], AI2-THOR [60], ICL [59], UnCoCo), we implement our 3D feature fusion algorithm on top of the ∇ SLAM dense reconstruction framework. By doing so, we leverage the point-based fusion technique proposed in [30], ensuring that points on nearby surface patches share the same *surfel*, decreasing the overall number of map elements, and also increasing the effective number of pixels that contribute to each map element. Another benefit we obtain is the ease of integration with PyTorch [71], which interfaces with a large number of foundation models. For pointfusion, we use the default hyperparameters as suggested in [30], i.e., a distance threshold of 5 cm (on positions) and an angular threshold of 20 degrees (on normals) is used to discard noisy correspondences.

On **outdoor datasets** (SemanticKITTI [63], self-captured autonomous driving sequences), we incrementally register pointclouds into a global frame using the LegoLOAM [65] technique for odometry estimation. We first compute all image points that have a valid map point by projecting the lidar depths onto the image plane. We associate the features at these pixels with the corresponding 3D locations.

A4. Pixel-aligned feature extraction

We use instance segmentation models from Mask2Former [55]; specifically the Swin-L backbone pretrained for image classification on ImageNet and subsequently finetuned for class-agnostic instance proposal generation on the COCO dataset. Note that we only use the class-agnostic instance proposal generator; and do not use any of the subsequent modules, which are explicitly trained with instance segmentation ground-truth. This results in 100 mask proposals per image. We allow each pixel to receive fused features from multiple overlapping or redundant masks. This is achieved by a running normalization whenever features from a new mask are assigned to a pixel.

A5. Foundation models used

We use two broad classes of foundation models: DINO (and associated vision transformers) [7], and CLIP (and variants) [6].

Vision transformer variants include various DINO backbones implemented in [7], as well as several vision-transformer variants explored in [21].

CLIP models used: We use open-source CLIP models from the OpenAI CLIP [72] and OpenCLIP [73] packages. We also use the publicly available AudioCLIP [8], trained on AudioSet [74].

A6. More details on our experiment setup

In all evaluations presented in the paper, we focus only on foreground objects, ignoring five background classes (*wall*, *floor*, *ceiling*, *door*, *window* for indoor scenes, and *road*, *sidewalk*, *building*) for outdoor scenes.

ScanNet: We note that most sequences from the ScanNet dataset suffer from motion blur artifacts, deviating several interesting objects of texture; or are small rooms devoid of interesting objects. We inspected every sequence (and each frame therein) over the ScanNet validation set, and identified the following sequences as being at least the scale of a one-bedroom apartment, and not suffering motion blur: scene0011, scene0050, scene0231, scene0378, scene0518. We also use scene0084 and scene0168 for debugging and tuning our reconstruction system (and consequently, these two scenes are left out of our evaluation set).

Replica: We evaluate on the following 8 replica scenes office0, office1, office2, office3, office4, room0, room1, room2.

Other datasets: We also qualitatively evaluate our mapping system over all sequences from the ICL [59] and on floorplans 9 and 402 from the AI2-THOR [60] simulator. On SemanticKITTI [63], we evaluate on all image frames containing at least one foreground object.

A7. Details of the UnCoCo dataset

The UnCoCo dataset comprises 78 commonly found objects in homes and workplaces, captured on tabletop settings over 20 RGB-D sequences. A subset of objects from UnCoCo is visualized in Fig. A.1. Of the captured 20 sequences, one was used for tuning parameters of the RGB-D reconstruction algorithm [54] and another was used for tuning hyperparameters (thresholds over cosine similarity scores); so we exclude these two sequences from evaluation.

We list the set of objects available across the 18 validation sequences in Table A.1.

References

- [1] Renato F Salas-Moreno, Richard A Newcombe, Hauke Strasdat, Paul HJ Kelly, and Andrew J Davison. Slam++: Simultaneous localisation and mapping at the level of objects. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1352–1359, 2013. [2](#)
- [2] John McCormac, Ankur Handa, Andrew Davison, and Stefan Leutenegger. Semanticfusion: Dense 3d semantic mapping with convolutional neural networks. In *2017 IEEE International Conference on Robotics and automation (ICRA)*, pages 4628–4635. IEEE, 2017. [2](#)
- [3] Nan Yang, Lukas von Stumberg, Rui Wang, and Daniel Cremers. D3vo: Deep depth, deep pose and deep uncertainty for monocular visual odometry. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1281–1292, 2020. [2](#)
- [4] Antoni Rosinol, Arjun Gupta, Marcus Abate, Jingnan Shi, and Luca Carlone. 3d dynamic scene graphs: Actionable spatial

Sequence ID	Set of objects in the sequence
Seq 03	Steel pouring mug, Ceramic coffee mug, Plastic banana, Windex spray bottle, SoftScrub
Seq 04	Baymax plush toy, Green caterpillar plush toy, Hedgehog plush toy
Seq 05	Hand-drill, Wooden spatula, Large lego block, Whiteboard marker
Seq 06	Plastic apple, Plastic grapes, Bottle of Vitamin E pills, Orange-colored bowl, Purple toy
Seq 07	Whisk, Spatula, Prongs, Silicone pastry brush
Seq 08	Paper cup, Spindrift can, Can of evaporated milk, Goldfish cheddar snack
Seq 09	Orange plastic cup, Paper cup, Steed pouring cup, Block of wood
Seq 10	Game of Bandu, Reacher grabber, Kitchen towel roll, Lysol wipes
Seq 11	Garbage bags, Cheetos, Steel measuring cup, Face mask
Seq 12	Coffee beans, Energy bar, Salted peanuts, Paper plates, Sugar sachet
Seq 13	Red hat, Magic candle, Molecule toy, Alligator toy, Blue frisbee
Seq 14	GoPro, Measuring tape, Scissors, Smartphone
Seq 15	Post-it notes, Black ceramic mug, Mustard, Tomato Ketchup
Seq 16	Bowl filled with sea shells, Ceramic vase, Large stapler
Seq 17	Stuffed mouse toy, Playing cube, Algorithms textbook, USB stick
Seq 18	USB adapter, NVIDIA Jetson board, Battery, Steel ruler
Seq 19	Compact Disk, Hard drive box, Teddy Bear, Inflatable brain toy
Seq 20	3D glasses, Spray bottle, Charger block, Purell bottle

Table A.1: List of objects from the UnCoCo sequences used for evaluation. The first two sequences (not listed here) were used for tuning hyperparameters.



Figure A.1: A subset of objects from the UnCoCo dataset. The dataset includes commonly found objects in homes and workplaces captured in a tabletop setting. Each object is annotated with 2D and 3D segmentation masks, and multimodal queries.

perception with places, objects, and humans. *Robotics Science and Systems*, 2020. 2

- [5] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. [2](#), [10](#)
 - [6] 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. *CoRR*, abs/2103.00020, 2021. [2](#), [3](#), [11](#)
 - [7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9630–9640, 2021. [2](#), [11](#)
 - [8] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Den-
...
er

gel. AudioClip: Extending clip to image, text and audio, 2021.
[2](#), [3](#), [5](#), [6](#), [7](#), [11](#)

- [9] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4904–4916. PMLR, 18–24 Jul 2021. [2](#)
 - [10] Yuxuan Ding, Lingqiao Liu, Chunna Tian, Jingyuan Yang, and Haoxuan Ding. Don’t stop learning: Towards continual learning for the clip model. 2022. [2](#), [5](#)
 - [11] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. [2](#)

- [12] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 2
- [13] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2085–2094, 2021. 2
- [14] Katherine Crowson, Stella Biderman, Daniel Kornis, Dashiell Stander, Eric Hallahan, Louis Castricato, and Edward Raff. Vqgan-clip: Open domain image generation and editing with natural language guidance. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVII*, pages 88–105. Springer, 2022. 2
- [15] Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. Point-e: A system for generating 3d point clouds from complex prompts. *arXiv preprint arXiv:2212.08751*, 2022. 2
- [16] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021. 2
- [17] Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11583–11592, 2022. 2
- [18] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVI*, pages 540–557. Springer, 2022. 2, 3, 4, 6, 7, 8
- [19] Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, and Juan Pablo Bello. Wav2clip: Learning robust audio representations from clip. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4563–4567. IEEE, 2022. 2
- [20] Hongwei Xue, Yuchong Sun, Bei Liu, Jianlong Fu, Ruihua Song, Houqiang Li, and Jiebo Luo. Clip-vip: Adapting pre-trained image-text model to video-language representation alignment. *arXiv preprint arXiv:2209.06430*, 2022. 2
- [21] Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep vit features as dense visual descriptors. *CoRR*, abs/2112.05814, 2021. 2, 11
- [22] Mark Hamilton, Zhoutong Zhang, Bharath Hariharan, Noah Snavely, and William T. Freeman. Unsupervised semantic segmentation by distilling feature correspondences. In *International Conference on Learning Representations*, 2022. 2
- [23] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16793–16803, 2022. 2
- [24] Boyi Li, Kilian Q. Weinberger, Serge J. Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. *CoRR*, abs/2201.03546, 2022. 2, 3, 4, 6, 8
- [25] Hang Zhao, Xavier Puig, Bolei Zhou, Sanja Fidler, and Antonio Torralba. Open vocabulary scene parsing. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2002–2010, 2017. 2
- [26] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. *arXiv preprint arXiv:2210.04150*, 2022. 2
- [27] Chong Zhou, Chen Change Loy, and Bo Dai. Denseclip: Extract free dense labels from CLIP. volume abs/2112.01071, 2021. 2, 6, 8
- [28] Brian Curless and Marc Levoy. A volumetric method for building complex models from range images. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 303–312, 1996. 3, 4
- [29] 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. 3
- [30] Maik Keller, Damien Lefloch, Martin Lambers, Shahram Izadi, Tim Weyrich, and Andreas Kolb. Real-time 3d reconstruction in dynamic scenes using point-based fusion. In *2013 International Conference on 3D Vision - 3DV 2013*, pages 1–8, 2013. 3, 4, 5, 11
- [31] Thomas Whelan, John McDonald, Michael Kaess, Maurice Fallon, Hordur Johannsson, and John J. Leonard. Kintinuous: Spatially extended kinectfusion. In *Proceedings of RSS '12 Workshop on RGB-D: Advanced Reasoning with Depth Cameras*, July 2012. 3
- [32] Thomas Whelan, Renato F. Salas-Moreno, Ben Glocker, Andrew J. Davison, and Stefan Leutenegger. Elasticfusion: Real-time dense SLAM and light source estimation. *Int. J. Robotics Res.*, 35(14):1697–1716, 2016. 3
- [33] Angela Dai, Matthias Nießner, Michael Zollöfer, Shahram Izadi, and Christian Theobalt. Bundlefusion: Real-time globally consistent 3d reconstruction using on-the-fly surface re-integration. *ACM Transactions on Graphics 2017 (TOG)*, 2017. 3
- [34] Ondrej Miksik, Vibhav Vineet, Morten Lidegaard, Ram Prasaath, Matthias Nießner, Stuart Golodetz, Stephen L Hicks, Patrick Pérez, Shahram Izadi, and Philip HS Torr. The semantic paintbrush: Interactive 3d mapping and recognition in large outdoor spaces. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 3317–3326, 2015. 3
- [35] Shuaifeng Zhi, Edgar Sucar, Andre Mouton, Iain Haughton, Tristan Laidlow, and Andrew J Davison. ilabel: Revealing objects in neural fields. *IEEE Robotics and Automation Letters*, 2022. 3

- [36] Stefanie Tellex, Nakul Gopalan, Hadas Kress-Gazit, and Cynthia Matuszek. Robots that use language. *Annual Review of Control, Robotics, and Autonomous Systems*, 3(1):25–55, 2020. 3
- [37] Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, et al. Experience grounds language. *arXiv preprint arXiv:2004.10151*, 2020. 3
- [38] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Daniel Ho, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang, Rosario Jauregui Ruano, Kyle Jeffrey, Sally Jesmonth, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Kuang-Huei Lee, Sergey Levine, Yao Lu, Linda Luu, Carolina Parada, Peter Pastor, Jornell Quiambao, Kanishka Rao, Jarek Rettinghouse, Diego Reyes, Pierre Sermanet, Nicolas Sievers, Clayton Tan, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Mengyuan Yan, and Andy Zeng. Do as i can and not as i say: Grounding language in robotic affordances. In *arXiv preprint arXiv:2204.01691*, 2022. 3, 9
- [39] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Tomas Jackson, Sally Jesmonth, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Kuang-Huei Lee, Sergey Levine, Yao Lu, Utsav Malla, Deeksha Manjunath, Igor Mordatch, Ofir Nachum, Carolina Parada, Jodilyn Peralta, Emily Perez, Karl Pertsch, Jornell Quiambao, Kanishka Rao, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Kevin Sayed, Jaspiar Singh, Sumedh Sonatakke, Austin Stone, Clayton Tan, Huong Tran, Vincent Vanhoucke, Steve Vega, Quan Vuong, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. Rt-1: Robotics transformer for real-world control at scale. In *arXiv preprint arXiv:2212.06817*, 2022. 3
- [40] Chenguang Huang, Oier Mees, Andy Zeng, and Wolfram Burgard. Visual language maps for robot navigation. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, London, UK, 2023. 3, 6
- [41] Dhruv Shah, Blazej Osinski, Brian Ichter, and Sergey Levine. Robotic Navigation with Large Pre-Trained Models of Language, Vision, and Action. 2022. 3
- [42] Arjun Majumdar, Gunjan Aggarwal, Bhavika Devnani, Judy Hoffman, and Dhruv Batra. Zson: Zero-shot object-goal navigation using multimodal goal embeddings. *arXiv preprint arXiv:2206.12403*, 2022. 3
- [43] Samir Yitzhak Gadre, Mitchell Wortsman, Gabriel Ilharco, Ludwig Schmidt, and Shuran Song. Clip on wheels: Zero-shot object navigation as object localization and exploration. *arXiv preprint arXiv:2203.10421*, 2022. 3, 6
- [44] Boyuan Chen, Fei Xia, Brian Ichter, Kanishka Rao, Keerthana Gopalakrishnan, Michael S. Ryoo, Austin Stone, and Daniel Kappler. Open-vocabulary queryable scene representations for real world planning. In *arXiv preprint arXiv:2209.09874*, 2022. 3, 6
- [45] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. volume abs/2201.02605, 2022. 3
- [46] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. 3
- [47] Clip-fields: Weakly supervised semantic fields for robotic memory, 2022. 3, 6
- [48] Songyou Peng, Kyle Genova, Chiyu "Max" Jiang, Andrea Tagliasacchi, Marc Pollefeys, and Thomas Funkhouser. Open-scene: 3d scene understanding with open vocabularies. In *arXiv*, 2022. 3, 6
- [49] Runyu Ding, Jihan Yang, Chuhui Xue, Wenqing Zhang, Song Bai, and Xiaojuan Qi. Language-driven open-vocabulary 3d scene understanding. *arXiv preprint arXiv:2211.16312*, 2022. 3
- [50] Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Colored point cloud registration revisited. In *Proceedings of the IEEE international conference on computer vision*, pages 143–152, 2017. 3
- [51] Timo Lüdecke and Alexander Ecker. Image segmentation using text and image prompts. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7076–7086, 2022. 4, 8
- [52] Bowen Cheng, Alexander G. Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. volume abs/2107.06278, 2021. 4
- [53] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 5, 9
- [54] 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. IEEE, 2020. 5, 11
- [55] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. 2022. 5, 11
- [56] 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. 6, 7, 8, 11
- [57] David Rozenberszki, Or Litany, and Angela Dai. Language-grounded indoor 3d semantic segmentation in the wild. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 6

- [58] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiaqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke M. Strasdat, Renzo De Nardi, Michael Goesele, Steven Lovegrove, and Richard Newcombe. The Replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 6, 7, 8, 11
- [59] A. Handa, T. Whelan, J.B. McDonald, and A.J. Davison. A benchmark for RGB-D visual odometry, 3D reconstruction and SLAM. In *IEEE Intl. Conf. on Robotics and Automation, ICRA*, Hong Kong, China, May 2014. 6, 11
- [60] Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Daniel Gordon, Yuke Zhu, Abhinav Gupta, and Ali Farhadi. AI2-THOR: An Interactive 3D Environment for Visual AI. *arXiv*, 2017. 6, 9, 11
- [61] Samir Yitzhak Gadre, Mitchell Wortsman, Gabriel Ilharco, Ludwig Schmidt, and Shuran Song. Cows on pasture: Baselines and benchmarks for language-driven zero-shot object navigation. *arXiv*, 2022. 6
- [62] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 7
- [63] J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, C. Stachniss, and J. Gall. SemanticKITTI: A Dataset for Semantic Scene Understanding of LiDAR Sequences. In *Proc. of the IEEE/CVF International Conf. on Computer Vision (ICCV)*, 2019. 7, 8, 11
- [64] Tao Chen, Anthony Simeonov, and Pulkit Agrawal. AIRobot. <https://github.com/Improbable-AI/airobot>, 2019. 8
- [65] Tixiao Shan and Brendan Englot. Lego-loam: Lightweight and ground-optimized lidar odometry and mapping on variable terrain. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4758–4765. IEEE, 2018. 8, 11
- [66] Moritz Werling, Julius Ziegler, Sören Kammel, and Sebastian Thrun. Optimal trajectory generation for dynamic street scenarios in a frenet frame. In *2010 IEEE International Conference on Robotics and Automation*, pages 987–993. IEEE, 2010. 8
- [67] Gabriel M. Hoffmann, Claire J. Tomlin, Michael Montemerlo, and Sebastian Thrun. Autonomous automobile trajectory tracking for off-road driving: Controller design, experimental validation and racing. In *2007 American Control Conference*, pages 2296–2301, 2007. 8
- [68] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623, 2021. 10
- [69] Iason Gabriel. Artificial intelligence, values, and alignment. *Minds and machines*, 30(3):411–437, 2020. 10
- [70] Eliezer Yudkowsky. The ai alignment problem: why it is hard, and where to start. *Symbolic Systems Distinguished Speaker*, 2016. 10
- [71] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. 11
- [72] OpenAI. Clip. <https://github.com/openai/CLIP>, 2021. 11
- [73] Openclip. https://github.com/mlfoundations/open_clip, 2022. 11
- [74] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017. 11