ConceptFusion: Open-set Multimodal 3D Mapping

Krishna Murthy Jatavallabhula[1], Alihusein Kuwajerwala[*2], Qiao Gu[*3], Mohd Omama[*4], Tao Chen[1], Alaa Maalouf[1], Shuang Li[1], Ganesh Iyer[††6], Soroush Saryazdi[†7], Nikhil Keetha[5], Ayush Tewari[1], Joshua B. Tenenbaum[1], Celso Miguel de Melo[8], Madhava Krishna[4], Liam Paull[2], Florian Shkurti[3], and Antonio Torralba[1]

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

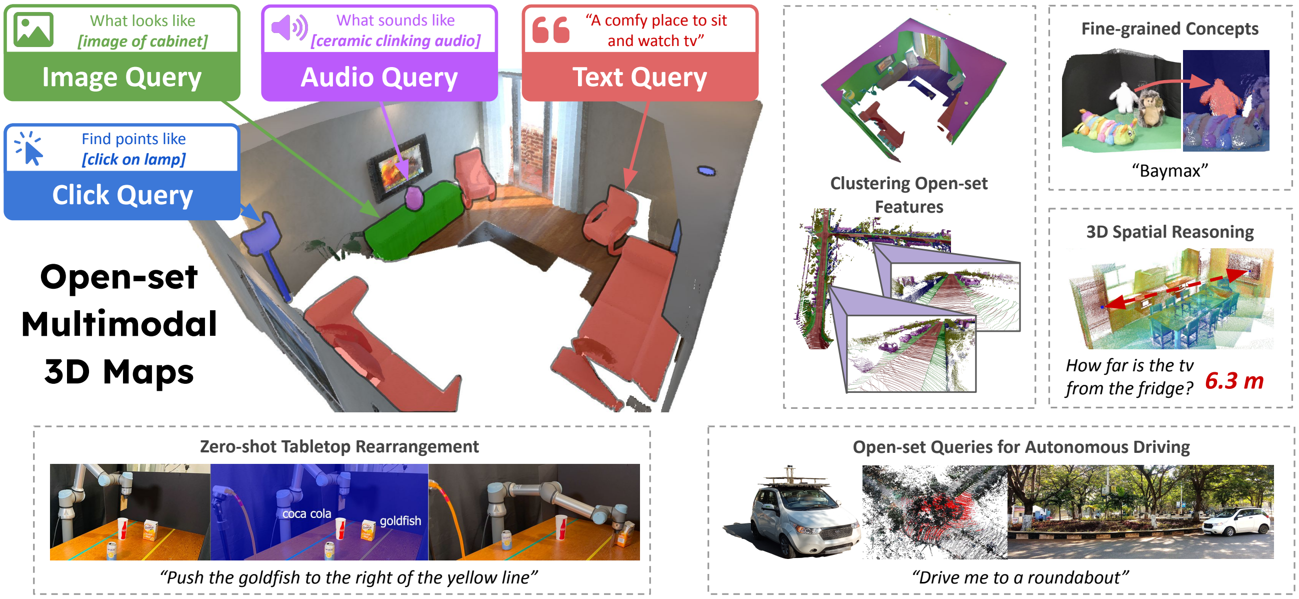
What looks like
[image of cabinet]
Image Query
What sounds like
[ceramic clinking audio]
Audio Query
"A comfy place to sit and watch tv"
Text Query
Find points like
[click on lamp]
Click Query
Open-set Multimodal 3D Maps
Clustering Open-set Features
Fine-grained Concepts
"Baymax"
3D Spatial Reasoning
How far is the tv from the fridge?
6.3 m
Zero-shot Tabletop Rearrangement
coca cola
goldfish
"Push the goldfish to the right of the yellow line"
Open-set Queries for Autonomous Driving
"Drive me to a roundabout"

Figure 1: ***ConceptFusion*** is a real-time approach to building *open-set multimodal 3D maps* from RGB-D images, and features from foundation models like CLIP and DINO. These maps, built on-the-fly, can be queried for arbitrary concepts specified as text, images, audio samples, or clicks on the 3D map. The fused features implicitly capture 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 current affiliation

1. Introduction

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 posess 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.

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 [1], ALIGN [2], AudioCLIP [3]) 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 [4] (see Fig. **??**). 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 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).

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. The *ConceptFusion* approach

The open-set multimodal 3D mapping problem: Given a sequence of image (and depth) observations of an environment $\mathcal{I} = \{I_t\}$ $(t \in \{0 \cdots T\})$, 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) $\mathcal{F}_{\text{mode}}$.

Components of *ConceptFusion*: The three primary components of the *ConceptFusion* include (a) a universal, instance segmentation module (such as Mask2Former [5] or SAM [6]), (b) a pixel-aligned feature extraction mechanism (which takes an image-level foundation model like CLIP and computes pixel-level features), and (c) a 3D feature fusion module (which fuses the pixel-aligned features to a 3D map).

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 $\overline{\boldsymbol{v}}_k \in \mathbb{R}^3$, (b) a normal vector $\overline{\boldsymbol{n}}_k \in \mathbb{R}^3$, (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.

Pixel-aligned feature extraction: For each image, we first extract class-agnostic instance masks by passing it through a universal segmentation model [5, 6]. For each mask, we compute local CLIP [1] features $\mathbf{f}^L$ of a bounding box sampled around the masked out region (blanking out all pixels that are not a part of the mask). We also compute the image-level CLIP feature vector $\mathbf{f}^G$. We fuse $\mathbf{f}^L$ and $\mathbf{f}^G$ to compute pixel aligned features $\mathbf{f}^P$ by computing the cosine similarity $\phi_i = \left\langle \mathbf{f}_i^L, \mathbf{f}^G \right\rangle = \frac{(\mathbf{f}_i^L)^T \mathbf{f}^G}{\|\mathbf{f}_i^L\| \|\mathbf{f}^G\| + \epsilon}$. Additionally, we compute a uniqueness score for each mask, μ_i, which is the average cosine similarity between mask i and every other mask. We combine the two similarities above to compute a mixing weight $w_i \in [0, 1]$ (with a temperature τ, set to 1 in all reported results). $w_i = \dfrac{\exp\left(\dfrac{\phi_i + \bar{\mu}_i}{\tau}\right)}{\sum_{i=1}^{R} \exp\left(\dfrac{\phi_i + \bar{\mu}_i}{\tau}\right)}$ Finally, the pixel-aligned feature for each region r_i is $\mathbf{f}_i^P = w_i \mathbf{f}^G + (1 - w_i) \mathbf{f}^L$ which is normalized and mapped to the pixels (u, v) in r_i.

3D feature fusion: We fuse $\mathbf{f}_{u,v,t}^P$ into the global map following a 3D reconstruction pipeline [7]. First, vertex and normal maps for each RGB-D image 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 [7]. The remaining points are fused into the global map. 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 a weighted averaging scheme $\mathbf{f}_{k,t}^P \leftarrow \frac{\bar{c}_k \mathbf{f}_{k,t-1}^P + \alpha \mathbf{f}_{u,v,t}^P}{\bar{c}_k + \alpha}$ and $\bar{c}_k \leftarrow \bar{c}_k + \alpha$, 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

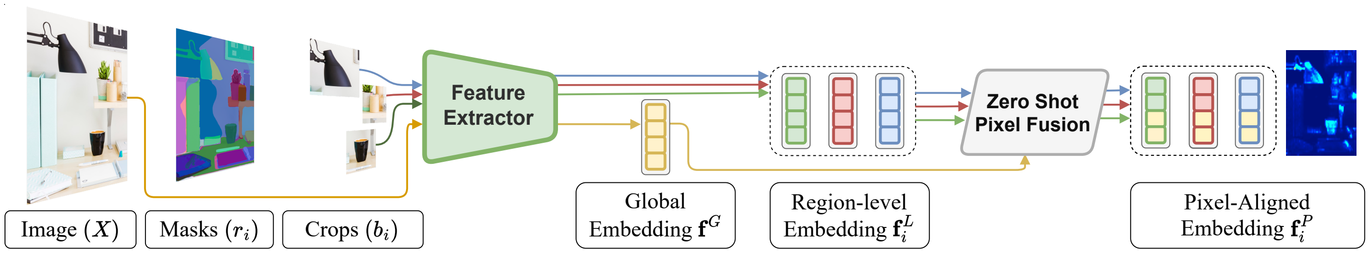
Feature Extractor
Zero Shot Pixel Fusion
Image (X)
Masks (r_i)
Crops (b_i)
Global Embedding $\mathbf{f}^G$
Region-level Embedding $\mathbf{f}_i^L$
Pixel-Aligned Embedding $\mathbf{f}_i^P$

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. **??** and described in Sec. **??**.

empirically that having a confidence value based on the normalized radial distance to the camera center, similar to [7, 8] works well. We refer to the appendix for hyperparameter values and more details.

Implementation details: Our feature fusion algorithm is implemented on top of the $\nabla SLAM$ [9] dense SLAM system, as this was one of the few implementations of the PointFusion algorithm [7], 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 [5] models for instance segmentation and generate 100 mask proposals per image. Our odometry and mapping approaches run at frame-rate (15 Hz). For frame-rate feature extraction, we quantize all foundation models used, and compile them to a static computation graph.

3. 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 well does *ConceptFusion* work on real-world robotics tasks?

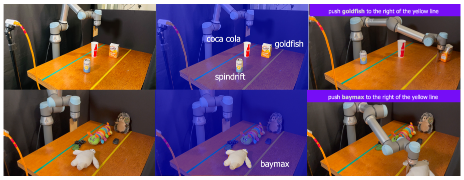
coca cola
goldfish
spindrift
push goldfish to the right of the yellow line
baymax
push baymax to the right of the yellow line

Figure 3: Real-world **tabletop rearrangement** experiments. The robot is provided with rearrangment 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.

Experimental setup: Our experimental benchmark comprises comprises 20 RGB-D scenes, spanning 78 commonly found household and office objects on a tabletop surface (see appendix). We crowdsource text, image, audio, and click queries for each object, resulting in a little over 500000 queries.

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 [10], NLMap-SayCan [11], CoWs [12, 13], CLIP-Fields [14] leverage LSeg [15]; while OpenScene [16] experiments with both LSeg [15] and OpenSeg [17]. 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 [18] features into a 3D map.

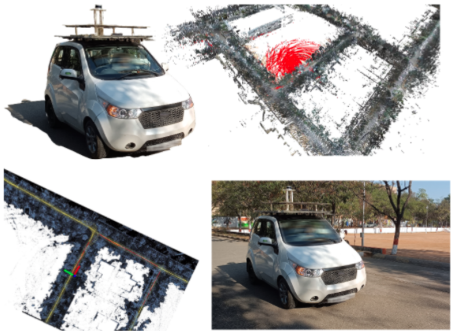

Figure 4: Real-world **autonomous navigation** experiments. (Left to right; top to bottom) Autonomous drive-by-wire platform deployed; pointcloud map of the environment with the response to the openset text-query ”football field” (shown in red); path found to the football field (shown in red); car successfully navigates to the destination autonomously. See our webpage for more results.

Discussion: We evaluate text query (structured and unstructured), image query, and audio query based object localization performance 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). Results are shown in Tables 1 through 4. 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 approaches 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

		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%
	ConceptFusion	**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%
	ConceptFusion	**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.

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.

Additional datasets: We also evaluate our approach on existing datasets like ScanNet [19], Replica [20], and Semantic KITTI [21]. Fig. 5 illustrates the performance of *ConceptFusion* on a sequence from the ScanNet dataset. *ConceptFusion* works out-of-the-box, even on free form text queries, while other approaches like OpenSeg [17] fail on all but the simplest of text queries.

Real-world experiments: Videos of our real robot experiments are accessed on our companion website. We apply *ConceptFusion* to the tasks of zero-shot tabletop rearrangement (3) and text-goal based autonomous navigation (4). 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 trials, the object set also includes distractors placed to hamper perception and/or manipulation planning. Two sides of the workspace (see Fig. 3) 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. 3 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). The autonomous navigation experiments are conducted on a self-driving vehicle. 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 with a calibrated stereo camera and lidar to reconstruct a map of a 320000 square yard (4000 sq. m.) urban area.

		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%
	ConceptFusion	**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 [3]	23.81%	N/A
	ConceptFusion	**64.29%**	0.287
ecological	Random	5.56%	N/A
	AudioCLIP [3]	22.22%	N/A
	ConceptFusion	**66.67%**	0.301

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

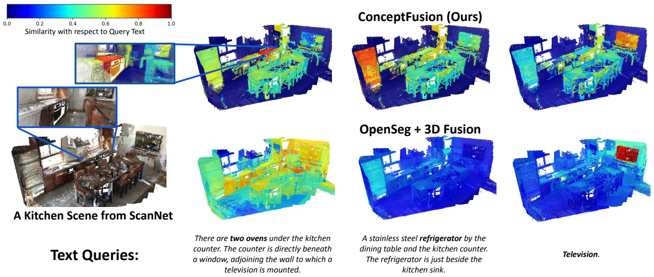
0.0
0.2
0.4
0.6
0.8
1.0
Similarity with respect to Query Text
ConceptFusion (Ours)
OpenSeg + 3D Fusion
A Kitchen Scene from ScanNet
Text Queries:
There are two ovens under the kitchen counter. The counter is directly beneath a window, adjoining the wall to which a television is mounted.
A stainless steel refrigerator by the dining table and the kitchen counter. The refrigerator is just beside the kitchen sink.
Television.

Figure 5: **Text queries over ScanNet [19]**: *ConceptFusion* handles long-form text queries and accurately localizes concepts. In the first two scenarios, OpenSeg [17] 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 [22] 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**.

4. 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 of our method are threefold. First, *ConceptFusion* operates over dense maps, comprising millions of 3D points over an apartment-scale scene, and augments each point with high-dimensional concept embeddings, requiring large amounts of memory. Third, we anticipate *ConceptFusion* to inherit the limitations and biases of foundation models [23, 24], warranting further investigations for potential harm as well as research into AI safety and alignment [25, 26].

References

[1] 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, 7

[2] 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

[3] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Audioclip: Extending clip to image, text and audio, 2021. 2, 4, 7

[4] 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] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 2022. 2, 3, 7

[6] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 2

[7] 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. 2, 3, 7

[8] 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

[9] 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. 3, 7

[10] 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

[11] 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

[12] 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

[13] 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. 3

[14] Clip-fields: Weakly supervised semantic fields for robotic memory, 2022. 3

[15] Boyi Li, Kilian Q. Weinberger, Serge J. Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. *CoRR*, abs/2201.03546, 2022. 3

[16] Songyou Peng, Kyle Genova, Chiyu "Max" Jiang, Andrea Tagliasacchi, Marc Pollefeys, and Thomas Funkhouser. Openscene: 3d scene understanding with open vocabularies. In *arXiv*, 2022. 3

[17] 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. 3, 4

[18] Chong Zhou, Chen Change Loy, and Bo Dai. Denseclip: Extract free dense labels from CLIP. In *ECCV*, 2021. 3

[19] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017. 4, 7

[20] 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. 4, 7

[21] 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. 4, 7

[22] 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. 4

[23] 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. 4

[24] 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. 4

[25] Iason Gabriel. Artificial intelligence, values, and alignment. *Minds and machines*, 30(3):411–437, 2020. 4

[26] Eliezer Yudkowsky. The ai alignment problem: why it is hard, and where to start. *Symbolic Systems Distinguished Speaker*, 2016. 4

[27] 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. 7

[28] 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. 7

[29] 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. 7

[30] 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. 7

[31] 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. 7

[32] Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep vit features as dense visual descriptors. *CoRR*, abs/2112.05814, 2021. 7

[33] OpenAI. Clip. `https://github.com/openai/CLIP`, 2021. 7

[34] Openclip. `https://github.com/mlfoundations/open_clip`, 2022. 7

[35] 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. 7

[36] Sanjay Subramanian, Will Merrill, Trevor Darrell, Matt Gardner, Sameer Singh, and Anna Rohrbach. Reclip: A strong zero-shot baseline for referring expression comprehension. *arXiv preprint arXiv:2204.05991*, 2022. 7

[37] Nan Liu, Shuang Li, Yilun Du, Josh Tenenbaum, and Antonio Torralba. Learning to compose visual relations. *Advances in Neural Information Processing Systems*, 34:23166–23178, 2021. 7

[38] Shuang Li, Yilun Du, Joshua B Tenenbaum, Antonio Torralba, and Igor Mordatch. Composing ensembles of pre-trained models via iterative consensus. *arXiv preprint arXiv:2210.11522*, 2022. 7

[39] 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. 7

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 Omama: 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 (prior to joining Amazon) and follow-up work. 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 Shaham, 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 [19], Replica [20], AI2-THOR [27], ICL [28], 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 [7], 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 [29], which interfaces with a large number of foundation models. For pointfusion, we use the default hyperparameters as suggested in [7], 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 [21], self-captured autonomous driving sequences), we incrementally register pointclouds into a global frame using the LegoLOAM [30] 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 [5]; 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 each pixel to recieve 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) [31], and CLIP (and variants) [1].

Vision transformer variants include various DINO backbones implemented in [31], as well as several vision-transformer variants explored in [32].

CLIP models used: We use open-source CLIP models from the OpenAI CLIP [33] and OpenCLIP [34] packages. We also use the publicly available AudioCLIP [3], trained on AudioSet [35].

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, devoiding 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 [28] and on floorplans 9 and 402 from the AI2-THOR [27] simulator. On SemanticKITTI [21], 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 [9] 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.

A8. Details on 3D spatial query modules

In Sec. **??**, we described our approach to handling 3D spatial reasoning queries. As investigated in [36, 37, 38], CLIP does not inherently capture spatial relationships or compositions. We therefore implement a set of primitive 3D spatial comparator (3DSC) modules, and rely on a large language model [39] (LLM) for parsing a natural language query into the function signatures (function name and input arguments) of the corresponding 3DSC. In particular, we use the `text-davinci-003` model from OpenAI, with the default settings (temperature of 0.7, maximum length of 256, and TopP equalling 1). We first condition the LLM by presenting a list of available 3DSCs and a brief natural language description of their behavior, followed by one example query and response. Here is the exact text prompt used.

```
Here is a set of available functions:
1. howFar(object1, object2): returns the distance
    between object1 and object2
2. isToTheRight(object1, object2): returns true if
    object1 is to the right of object2
```

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.

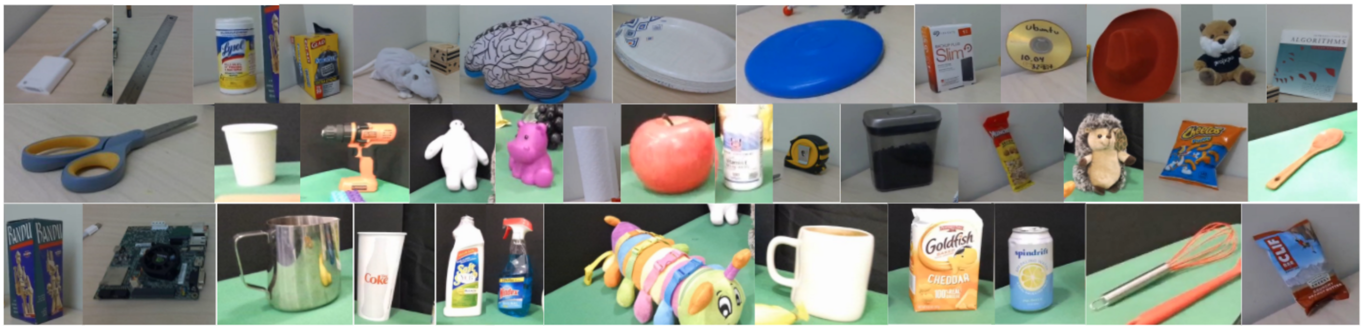

Figure A.1: A subset of objects from the UnCoCo dataset. The dataset includes commonly found objects in homes and workplaces captured in a tabletop setting. Each object is annotated with 2D and 3D segmentation masks, and multimodal queries.

```
3. isToTheLeft(object1, object2): returns true if
    object1 is to the left of object2
4. isContained(object1, object2): returns true if
    object1 is contained in object2
5. onTopOf(object1, object2): returns true if
    object1 is on top of object 2
6. under(object1, object2): returns true if object1
     is underneath object 2
7. isBigger(object1, object2): returns true if
    object1 is bigger than object2
8. canFitInside(object1, object2): returns true if
    object1 can fit inside object2
Parse the provided queries into one of the above
    function formats.

Query: How close is the chair from the sofa?
Response: howFar(chair, sofa)
```

Listing 1: Base prompt used to condition for spatial queries

Once conditioned with this prompt, the model is able to parse language queries into function signatures. We directly execute these function signatures as is. We find LLMs to be very effective at parsing: of the 100 queries we used, each one was parsed correctly. Shown below are a few outputs.

```
Query: Is the bread inside the bowl?
Response: isContained(bread, bowl)

Query: Is the apple on the table?
Response: onTopOf(apple, table)

Query: How far is the sanitizer from the door?
Response: howFar(sanitizer, door)

Query: I want to know the distance between the door
     and the window.
Response: howFar(door, window)

Query: Where is the closest restroom from my
    location?
Response: howFar(restroom, my location)

Query: I want to grab a can of soda and put this
    into a bag.
Response: canFitInside(soda, bag)

Query: Is the soda inside the bag?
Response: isContained(soda, bag)
```

Listing 2: Sample outputs from the LLM after conditioning