Language Embedded Radiance Fields for Zero-Shot Task-Oriented Grasping

Anonymous Author(s)
Affiliation
Address
email

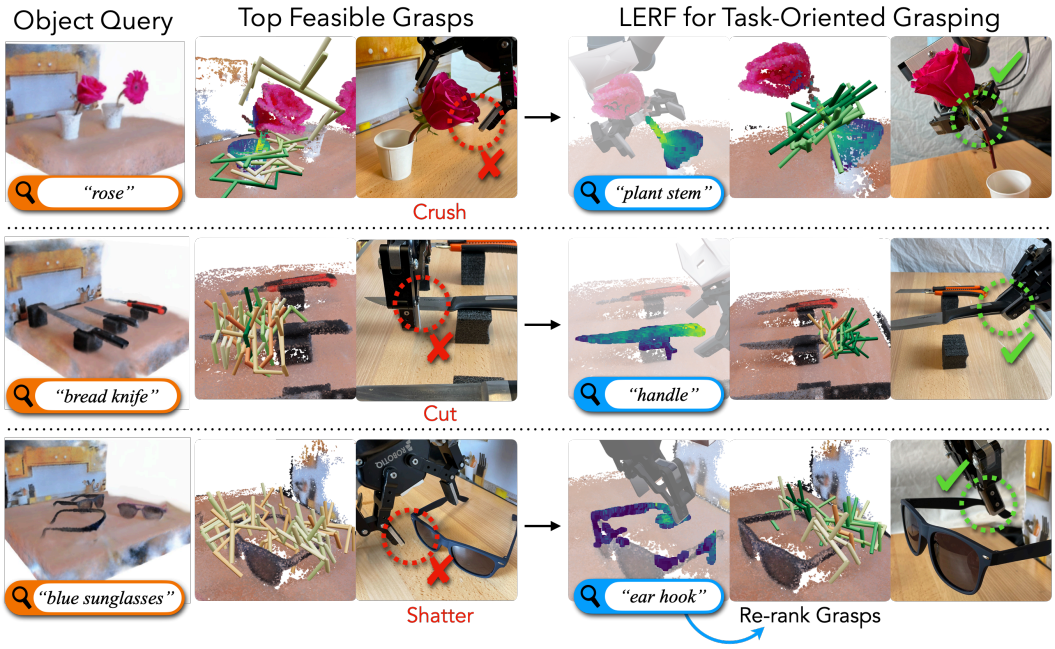

Figure 1: Learning-based grasp planners primarily consider object geometry, potentially leading to dangerous grasps for robots and their environment. LERF-TOGO uses natural language to select the target object with LERF [1] (in **orange**), and resamples grasps towards specific object parts using conditional LERF queries (in **blue**) for safe, task-oriented grasps.

Abstract: Grasping objects by a specific part is often crucial for safety and for executing downstream tasks. Yet, learning-based grasp planners lack this behavior unless they are trained on specific object part data, making it a significant challenge to scale object diversity. Instead, we propose LERF-TOGO, which uses vision-language models zero-shot to output a grasp distribution over an object given a natural language query. To accomplish this we first reconstruct a Language Embedded Radiance Field (LERF) of the scene, which distills CLIP embeddings into a multi-scale 3D language field queryable with text. However, LERF has no sense of objectness, meaning its relevancy outputs often return incomplete activations over an object which are insufficient for subsequent part queries. LERF-TOGO mitigates this lack of spatial grouping by extracting a 3D object mask via DINO features, then conditionally querying LERF on this mask to obtain a grasp distribution over the object. We evaluate LERF-TOGO's ability to grasp task-oriented parts on 31 different physical objects, and find it selects grasps on the correct part in 81% of all trials and grasps successfully in 69%.

Keywords: NeRF, Grasping, Semantics, Natural Language

Submitted to the 7th Conference on Robot Learning (CoRL 2023). Do not distribute.

1 Introduction

Many common objects must be grasped appropriately to avoid damage or facilitate performing a task: a knife by its handle, a flower by its stem, or sunglasses by their frame. Learning-based grasping systems exhibit impressive robustness on grasping arbitrary objects [2, 3, 4, 5, 6, 7, 8, 9, 10], but these systems typically measure grasp success based on whether the object was lifted [11, 12, 13, 14, 15]. Critically, these methods ignore an object's semantic properties: even if a robot could locate your favorite sunglasses, rather than safely grasp at the frame it may shatter the lenses. This ability to grasp an object part based on a desired task and constraints is called task-oriented grasping, and while well studied [16, 17, 18, 19, 20, 21], previous methods collect specific object affordance datasets and struggle to scale to a diverse set of objects. Instead, the flexibility of natural language makes it a fitting interface for specifying what and where to grasp. In this work, we propose *LERF* for *Task-Oriented Grasping on Objects* (LERF-TOGO), a method which enables task-oriented grasping through natural language by using large vision-language models in a zero-shot manner.

LERF-TOGO takes as input an object and object part natural language description (i.e. *"flower; stem"*), and outputs a ranking over viable grasps on this object from which the robot should grasp. We build on recent work Language Embedded Radiance Fields (LERF) [1], which takes in calibrated RGB images and trains a standard NeRF in tandem with a scale-conditioned CLIP [22] feature field. Given a sentence prompt query, it outputs a 3D *relevancy* heatmap representing similarity to the query. However, these heatmap may fail to highlight the full object (e.g., highlight only the bristles of a brush), which may cause issues when directly deployed to a task-oriented grasping task (grasp the "handle" of a brush). LERF-TOGO improves upon LERF's capabilities by predicting a 3D object mask using 3D DINO [23] features explicitly during inference. We propose a method of conditional LERF querying which restricts an object sub-part query to the object mask, leveraging the multi-scale nature of LERF to isolate specific regions within an object.

We build a system with appropriate regularizations which allows LERF-TOGO to operate on a physical robot and evaluate its semantic grasping capabilities on a wide range of objects. In our experiments, 96% of generated grasps are on the correct object, 81% on the correct object part, and 69% result in a successful grasp.

This work contributes LERF-TOGO, a method for producing task-based semantic grasp distributions over an object by first extracting a coarse 3D object mask with DINO [23] features to locally grow a relevant region, then conditioning a LERF query on this mask to isolate sub-parts of an object. We design a robotic system which integrates LERF-TOGO on a physical robot to reconstruct a LERF of a scene, then execute task-oriented grasps through natural language to grasp semantically meaningful object parts. Code and data will be released upon publication.

2 Related Work

2.1 Task-Oriented Grasping

Task oriented grasping studies how to grasp objects by specific parts based on a use case. It has been studied by probabilistically modeling human grasps [19], extracting geometric features from labeled object parts [21], training on part-affordance datasets in simulation [24], or transferring category-specific part grasps to new instances [25]. Recent works [26, 17] train object-part grasp networks by leveraging object part and manipulation affordance datasets for a range of household objects. In contrast, LERF-TOGO's usage of off-the-shelf vision-language models trained at scale does not require training on affordance datasets, capturing long-tail objects more easily. Given 3D grasp regions on an object, Song et al. [27] demonstrates that biasing an off-the-shelf grasp planner towards these regions is a viable approach to sampling task-oriented semantic grasps. Kokic et al. [18] uses videos of humans interacting with objects to guide grasps towards the same part, Decomposing objects into parts has also long been studied as a co-segmentation task in vision [28, 29]. Recent approaches use pretrained vision features to discover common parts within sets of objects [30]. This technique has been applied at scale to segment parts of objects based on a canonical object [31] or

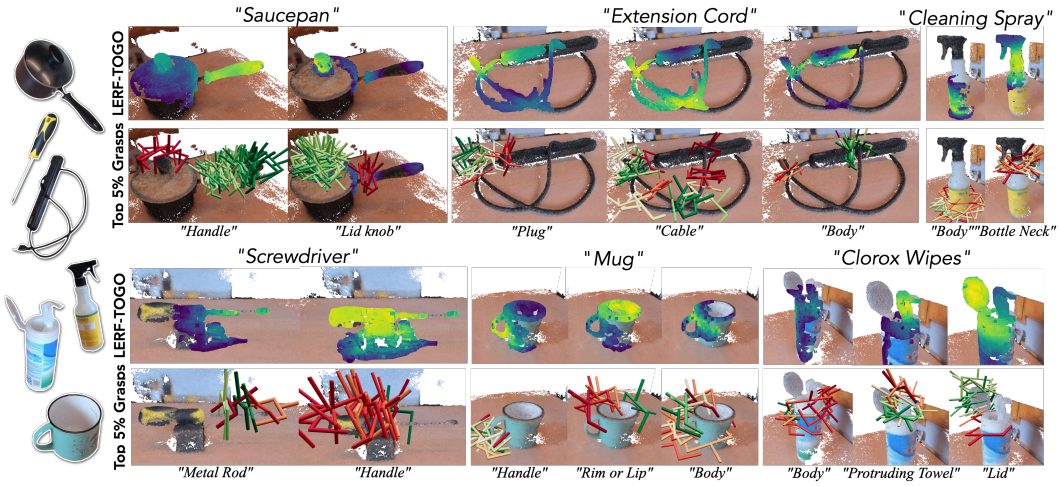

Figure 2: **Task-Oriented Grasps**: LERF-TOGO grasps target objects by different parts with different natural language queries (in quotes). Left: Crops of target objects. Right: The top row visualizes the relevancy distribution across the object mask, and bottom the top 5% of resampled grasps.

detect object affordances from example images of human usage [16]. Though effective, it assumes access to a canonical image of each object and pre-existing part labels or demonstrations, which are restrictive in real-world application. LERF-TOGO instead uses free-form language queries to isolate parts of objects in a 3D reconstruction.

2.2 Neural Radiance Fields (NeRF)

Neural Radiance Fields (NeRF) [32] are an attractive representation for high quality scene reconstruction from pose RGB images, with an explosion of recent work on visual quality [33, 34, 35, 36, 37, 38, 39], large-scale scenes [40, 41, 42], optimization speed [43, 44, 45, 46], dynamic scenes [47, 48, 49], and more. Because of its high-quality reconstruction and differentiable properties, NeRF has been widely explored in robotics for navigation and mapping [33, 50, 51, 52], manipulation [53, 54, 55, 56], and for synthetic data generation [57]. This work is most similar to works such as Evo-NeRF [55] which use NeRF as a real-time scene reconstruction to grasp objects. However, in contrast to previous works which only use RGB information, in this work we must include additional semantic information in 3D to select grasps falling on relevant target objects.

Several prior works explore using semantic outputs inside NeRF. Semantic-NeRF [58], Panoptic Lifting [59], and Panoptic Neural Fields [60] distill semantic categories from semantic segmentation networks into 3D to improve the 3D consistency of labels, particularly noting the denoising effect of averaging multiple views. Other works such as Distilled Feature Fields [61] or Neural Feature Fusion Fields [62] distill feature vectors from DINO and LSeg [63], and show they can be used for editing and scene segmentation. We build off of LERF [1], a recent work which proposes a method of distilling CLIP features into 3D without fine-tuning or mask proposals, while capturing semantics at a multitude of different scales.

2.3 Open-Vocabulary Detectors

Open-vocabulary object detection attempts to output mask or bounding box detections given text prompts as input. OVD [64] proposes a two-stage training pipeline, where they first learn a visual-semantic stage using image-text pairs and then learn object detection using object annotations from a set of base classes. OpenSeg [65] employs a similar joint learning framework for simultaneously training a mask prediction model and learning visual-semantic alignments for each mask by aligning captions and masks. Some methods [66, 67, 68] distill CLIP embeddings into a visual encoder by

using labels from detection or classification datasets, or fine-tune a lightweight detection head on top of a pretrained backbone [69]. However, fine-tuning on detection datasets severely bottlenecks the language understanding capability of detection models [1, 61, 70], and as noted by the authors of OV-Seg [67], these networks face challenges in decomposing hierarchical components of the original masks, including object parts. LERF circumvents the usage of region proposals or fine-tuning by incorporating language embeddings densely in 3D, making it an attractive approach for effective part decomposition.

2.4 Natural Language in Robotics

With the advent of large pretrained language and vision-language models, several works have explored building 3D map representations to guide robot navigation. VL-Maps [71] and OpenScene [72] build a 3D language embedding from pretrained open-vocab detectors [73, 63], which can be used to navigate to target queries. CLIP-Fields [74], ConceptFusion [75], and NLMaps-SayCan [76] take more of a region-proposal based approach, querying CLIP on the outputs of some region proposal methods and fusing them into 3D point clouds for downstream navigation tasks. Region-based zero-shot methods retain more language understanding than fine-tuned features but run the risk of missing objects by insufficiently masking input images. LERF-TOGO avoids this issue by densely building a multi-scale field which can be queried for for sub-object parts. Language has also been studied in the context of robot manipulation. Mees and Burgard [77] uses the MAttNet [78] vision-language model for object rearrangement, and CLIPort [79] uses language understanding from CLIP to train a language-conditioned pick and place module from demonstrations. PerAct [80] uses language-conditioned demonstrations with a 3D scene transformer to learn diverse tasks, MOO [81] uses the outputs from Owl-ViT to condition a manipulation policy for grasping objects, and large-scale demonstration datasets like RT-1 [82] train on massive language-conditioned demonstration trajectories. In contrast to many other language-conditioned approaches, LERF-TOGO uses internet-scale vision models purely zero-shot, not requiring fine-tuning on demonstrations or robot exploration.

3 Problem and Assumptions

Given a planar surface (table or workbench) containing a set of objects, the objective is for a robot to grasp and lift a target object specified using natural language. The query includes both the object and the part to grasp the object by, such as *"sunglasses; ear hooks."* We assume access to a robot manipulator with a parallel jaw gripper and calibrated wrist-mounted RGB camera, and that the objects in the scene are graspable by the robot. We also assume the target query specifies an object that actually exists in the scene.

4 Method

Given an object and object part query, LERF-TOGO outputs a ranking of viable grasps on the object part. To accomplish this, it first captures the scene and reconstructs a LERF, then extracts a 3D object mask and conditions the object part query over this mask to produce a semantic grasp distribution.

4.1 3D Object Extraction

An important limitation of LERF is its lack of spatial grouping within objects: for example given *"can opener"*, LERF tends to highlight regions of the scene that most obviously represent that object, (e.g. the metal cogs), as in Fig. 3. This is problematic when the natural graspable part of the object (e.g. the handle), is far away. LERF inherently exhibits such local behavior because it trains on local crops of input images, causing CLIP embeddings surrounding the handle to be unaware if it belongs to a can opener. LERF-TOGO overcomes this by finding a 3D object mask given a language query, which groups the object part together with the LERF activation. To do this, we use 3D distilled DINO embeddings [23] (self-**DI**stillation with **NO** labels) during inference, which

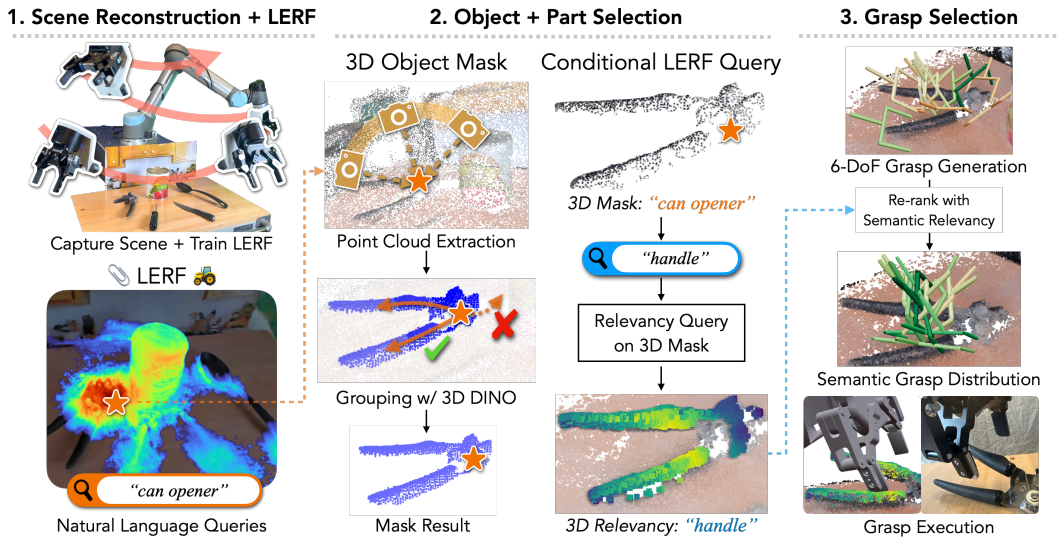

Figure 3: **LERF-TOGO Pipeline**: After reconstructing the scene with a wrist-mounted camera, we render an object-centric point cloud around the highest LERF activation. We next extract a 3D object mask by flood-filling the DINO features in this point cloud, condition an object part query on this object mask. Finally, we sample grasps and re-rank them according to 3D object part relevancy.

have been shown to exhibit strong object awareness and foreground-background distinction within their first principle components [30, 23, 83]. To extract the first DINO principle component for a scene, we render a top-down view observing the whole workspace and calculate the first principal component, f. Next, we query LERF for the target object and find the highest activation in 3D (Fig 3, orange star). We render an object-centric point cloud around this 3D point and use it to initiate a flood-fill on the point cloud. This flood-fill propagates to neighboring points within a threshold DINO similarity along the f, which produces a 3D object mask.

4.2 Conditional LERF Queries

Another important challenge of using CLIP is its tendency to behave as a bag-of-words [22, 1]: the activation for *"mug"* behaves very similarly to *"mug handle"* because CLIP latches onto individual words, not the grammatical structure of sentences. To mitigate this phenomenon, LERF-TOGO introduces a conditional method of querying LERF relevancy by composing two related queries. Because LERF is scale-conditioned, it conducts a grid-search over physical scales for a given query during relevancy computation, returning the relevancy at the scale which produced the highest activation. To condition a LERF query, we conduct this search only over the points from the 3D object mask. Intuitively, this results in a distribution over the object's 3D geometry representing the likelihood that a given point is the desired object part, which can be used for biasing grasps towards this region.

4.3 Grasping

Grasp Sampling Ensuring complete coverage of grasps on objects is critical to avoid missing specific object parts. We use GraspNet [15], which can generate 6-DOF parallel jaw grasps from a monocular RGBD point cloud, but from a single view it often misses key grasps on target object parts. To mitigate this, and to leverage the full 3D geometry available within NeRF, we create a hemisphere of virtual cameras oriented towards the scene's center. For every virtual camera, we convert the scene's point cloud to the camera coordinate frames before providing it as input to the pretrained GraspNet model. To obtain the final set of grasps for the scene, we combine the generated grasps from the virtual cameras using non-maximum suppression to remove duplicates.

Figure 4: **Left**: Select scenes used during experiments, **Right**: example grasps using LERF-TOGO.

Grasp Ranking Given the grasps sampled in the previous step (the *geometric* distribution), we now combine it with the semantic distribution across an object obtained from LERF-TOGO. The semantic score s_{sem} for a given grasp is computed as the median LERF relevancy of points within the grasp volume. The geometric score s_{geom} is the confidence output from GraspNet, indicating grasp quality based on geometric cues. To balance relevance and success likelihood, we combine the grasp score $s = 0.95s_{sem} + 0.05s_{geom}$ to ensure that we consider the most relevant grasps while slightly biasing towards confident grasps.

4.4 Scene Reconstruction

Robot Capture The robot uses a wrist-mounted camera to capture the scene with a hemispherical trajectory centered at the workspace, similar to Evo-NeRF [55]. The capture has a radius of 45 cm and arcs from $\pm 100^\circ$ around the workspace horizontally and an inclination range of 30° to 75°. We capture images while the arm moves at 15 cm/sec at a rate of 3 hz, resulting in around 60 images per capture. We discard blurry images by analyzing the variance of the image Laplacian, ensuring the images are high quality. While the robot moves, we pre-process each image to extract DINO features, multi-scale CLIP, and ZoeDepth [84] (see below) to be used during LERF training.

NeRF Regularization NeRF encounters difficulties in reconstructing texture-less planar surfaces, especially in the presence of specularities. This limitation is prominent in our table-top scenes, where the glossy surface and metallic objects can result in depth renderings with jagged missing regions. These missing regions can cause LERF renderings to spuriously activate and degrade the performance of grasp networks, so we apply depth regularization to mitigate this issue. We adopt the local depth ranking loss proposed in SparseNeRF [85] and use ZoeDepth [84] as the underlying depth model. We found this performs better than smoothness priors [86, 46] because it retains more fine-grained geometry. Additionally, we use the gradient scaling approach from Philip and Deschaintre [39], which significantly reduces the number of near-camera floaters and enables more robust grasping directly from point clouds rendered from the NeRF.

Poses obtained from cameras in motion are slightly inaccurate, which we found could result in oversmoothed geometry with depth regularization. To overcome this, we optimize the NeRF for the first 500 steps without any regularization to allow the camera poses to settle, then anneal the depth regularization loss term from 0 to 100% over the next 1500 steps. Interestingly, we find staged training not only preserves thin features better but also speeds LERF optimization. We hypothesize this is because supervising the language field on un-converged density in free space results in a poor network initialization, while beginning LERF optimization after geometry has been largely removed from free space allows a smoother learning signal.

5 Experiments

Evaluation We evaluate LERF-TOGO on a wide variety of 31 different objects and 49 total part queries (Fig. 4). For each object, we select an object query by describing it sufficiently to un-

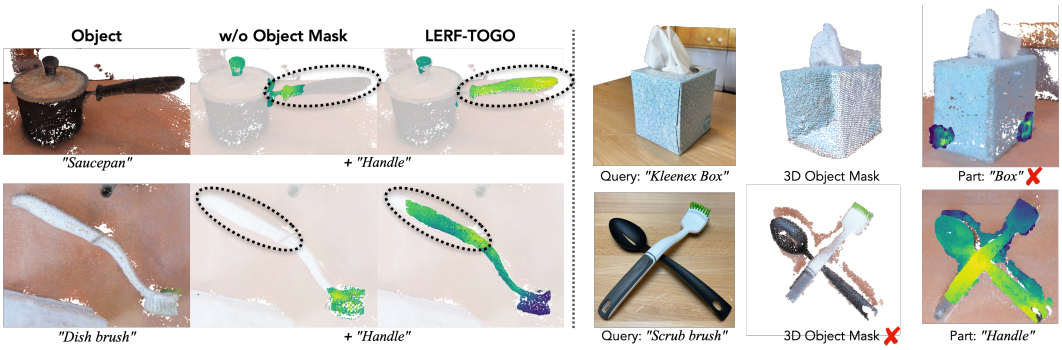

Figure 5: **Ablation and Limitations**: *Left*: Without 3D object masking and conditional querying, LERF cannot capture oblong object shapes. *Right*: CLIP can sometimes fail on generic prompt queries, like the poor activation on the box. Additionally, LERF-TOGO struggles with groups of connected objects as the 3D object mask groups them all together.

	ConceptFusion [70]	LERF-TOGO
Grasp on Correct Part	39%	82%
Grasp on Correct Object	77%	96%
Physically Successful	–	69%

Table 1: **Aggregated Grasping Results:** Results are reported across 49 different prompts and 12 scenes. See the Appendix for a complete list of scenes and queries.

ambiguously differentiate between other objects in the scene. We use semantic descriptions when possible, and add visual descriptions *only* when such descriptions are ambiguous (i.e using color to differentiate multiple mugs in a scene). We provide a part query for each object by describing a natural place for a robot to grasp and lift (i.e. "handle", "plant stem", "ear hook", "frame"). In addition, several objects include multiple different grasp locations. A grasp is successful if it lifts the correct object using the appropriate subpart at least 10cm vertically, and the object remains securely within the gripper jaws throughout. For each query, we measure 1) whether the selected grasp was on the correct *object*, 2) whether the selected grasp was on the correct object *part*, and 3) whether the grasp successfuly *lifted* the object from the table. For a full list of object queries see the Appendix.

ConceptFusion Comparison ConceptFusion [70] generates a multimodal point cloud of a scene by fusing RGBD images and their extracted features together. We provide ConceptFusion with depth generated from the NeRF, which results in the same high-quality point cloud we use for grasping. To fairly represent the paper, we use the OpenCLIP ViT-H/14 model, which is several times larger than the ViT-B/16 model for LERF-TOGO. To query ConceptFusion, we provide it the concatenated object and part prompt (i.e. *"mug handle"*) and rank grasps via the highest similarity. We report the object and part success without physical evaluation.

6 Grasping Results

LERF-TOGO overall achieves a 69% success rate at physically grasping and lifting objects by the correct part. The selected grasp was located around the correct object part 81% of the time, with the remaining failures being grasp execution failures. For context, the highest confidence geometric grasp on an object mask only lies on the correct part 18% of the time, suggesting LERF-TOGO meaningfully biases the grasp distribution to the object part. Selected task-oriented queries are visualized in Fig. 2: the distribution of grasps drastically shifts based on the given part query, and can focus task-oriented grasps on multiple different regions per object based on the language prompt. LERF-TOGO shows strong language understanding performance for object selection (96%), able to differentiate between very fine-grained language queries like color, appearance ("matte" vs "shiny"), or semantically similar categories (*"popsicle"* vs *"lolipop"*). It also can recognize long-tail object queries like *"ethernet dongle"*, *"cork"*, or *"martini glass"*, owing to its usage of CLIP zero-shot.

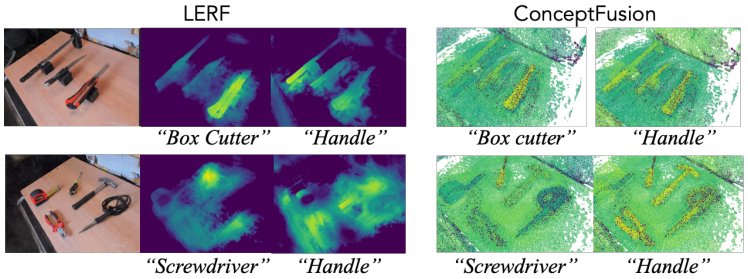

Figure 6: **ConceptFusion Comparison**: ConceptFusion performs well on object-level queries, however struggles with sub-object part queries because of its lack of multi-scale semantics.

LERF-TOGO out-performs ConceptFusion by 43% at task-oriented grasping because it can capture multi-scale semantics, while ConceptFusion is limited to one CLIP embedding per point. This makes hierarchical querying difficult, and is reflected by the fact that ConceptFusion performs similarly to LERF-TOGO at selecting the correct object, but suffers at selecting the right object part. Due to its lack of scale-conditioning, ConceptFusion frequently emphasizes large sections of the table due to the inclusion of both the objects and the table itself in the mask proposals (Fig. 6).

Failures The primary failure modes of LERF-TOGO are mistaking visually similar object parts for one another (eg teapot spout for a handle), missing subtle geometries like the small teacup handle or spray trigger, or confusing very close categories like steak and bread knives. We also observe prompt sensitivity with part queries: for example *"plant stem"* more strongly localizes grasps than *"stem"*, and without more prompt tuning *"body"* sometimes fails to highlight the bases of bottles.

Object Extraction Ablation: Without 3D object masking and conditional querying, LERF-TOGO suffers with oblong objects, as shown in Fig. 5. We compare against querying LERF individually for the object and part, and multiplying their results together. This produces fragmented results which can ignore relevant parts of the object for part queries.

7 Limitations and Future Work

One limitation of LERF-TOGO is speed: the entire end-to-end the process takes a few minutes which can be impractical for time-sensitive applications. Future work on additional regularizations and optimizations to LERF training may reduce computation time. Another key limitation of LERF-TOGO is with groups of connected foreground objects, for example a bouquet of multiple flowers. The DINO flood-fill will create a foreground group containing all flowers, after which isolating the stem of a specific individual type of flower (i.e *"daisy"* vs *"rose"*) is challenging. Supporting hierarchy within foreground groups is critical to enable such cases. In addition, we assume access to the object part query, but automatically obtaining good object part prompts would be critical for real-world usage. Future work might study integration with text-based LLMs to bridge this gap.

8 Conclusion

This paper presents LERF-TOGO, a method for using vision-language models zero-shot with Language Embedded Radiance Fields to grasp objects and their parts via language. By improving the spatial grouping of LERF relevancy outputs, LERF-TOGO can support hierarchical part queries conditioned on the full object. Results indicate it performs strongly at language-guided grasping, with grasps landing on the correct object 96% of the time, and furthermore can direct grasps to the correct object parts 81% of the time. Code and datasets will be released after the submission process.

References

[1] J. Kerr, C. M. Kim, K. Goldberg, A. Kanazawa, and M. Tancik. Lerf: Language embedded radiance fields. *arXiv preprint arXiv:2303.09553*, 2023.

[2] K. Kleeberger and M. F. Huber. Single shot 6d object pose estimation, 2020.

[3] K. Bousmalis, A. Irpan, P. Wohlhart, Y. Bai, M. Kelcey, M. Kalakrishnan, L. Downs, J. Ibarz, P. Pastor, K. Konolige, S. Levine, and V. Vanhoucke. Using simulation and domain adaptation to improve efficiency of deep robotic grasping, 2017.

[4] S. Song, A. Zeng, J. Lee, and T. Funkhouser. Grasping in the wild: Learning 6dof closed-loop grasping from low-cost demonstrations. *Robotics and Automation Letters*, 2020.

[5] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world, 2017.

[6] D. Kalashnikov, A. Irpan, P. Pastor, J. Ibarz, A. Herzog, E. Jang, D. Quillen, E. Holly, M. Kalakrishnan, V. Vanhoucke, and S. Levine. Qt-opt: Scalable deep reinforcement learning for vision-based robotic manipulation, 2018.

[7] S. Levine, P. Pastor, A. Krizhevsky, J. Ibarz, and D. Quillen. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. *The International Journal of Robotics Research*, 37(4-5):421–436, 2018. doi:10.1177/0278364917710318.

[8] A. Mousavian, C. Eppner, and D. Fox. 6-dof graspnet: Variational grasp generation for object manipulation, 2019.

[9] J. Mahler, M. Matl, V. Satish, M. Danielczuk, B. DeRose, S. McKinley, and K. Goldberg. Learning ambidextrous robot grasping policies. *Science Robotics*, 4(26):eaau4984, 2019.

[10] L. Pinto and A. Gupta. Supersizing self-supervision: Learning to grasp from 50k tries and 700 robot hours. In *2016 IEEE international conference on robotics and automation (ICRA)*, pages 3406–3413. IEEE, 2016.

[11] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár. Microsoft coco: Common objects in context, 2015.

[12] A. Depierre, E. Dellandréa, and L. Chen. Jacquard: A large scale dataset for robotic grasp detection. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3511–3516. IEEE, 2018.

[13] I. Lenz, H. Lee, and A. Saxena. Deep learning for detecting robotic grasps. *The International Journal of Robotics Research*, 34(4-5):705–724, 2015.

[14] B. Calli, A. Singh, A. Walsman, S. Srinivasa, P. Abbeel, and A. M. Dollar. The ycb object and model set: Towards common benchmarks for manipulation research. In *2015 international conference on advanced robotics (ICAR)*, pages 510–517. IEEE, 2015.

[15] H.-S. Fang, C. Wang, M. Gou, and C. Lu. Graspnet-1billion: A large-scale benchmark for general object grasping. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11444–11453, 2020.

[16] G. Li, V. Jampani, D. Sun, and L. Sevilla-Lara. Locate: Localize and transfer object parts for weakly supervised affordance grounding. In *Proceedings of the IEEE International Conference on Computer Vision*, 2023.

[17] A. Murali, W. Liu, K. Marino, S. Chernova, and A. Gupta. Same object, different grasps: Data and semantic knowledge for task-oriented grasping. In *Conference on Robot Learning*, 2020.

[18] M. Kokic, D. Kragic, and J. Bohg. Learning task-oriented grasping from human activity datasets. *IEEE Robotics and Automation Letters*, 5(2):3352–3359, 2020.

[19] D. Song, C. H. Ek, K. Huebner, and D. Kragic. Task-based robot grasp planning using probabilistic inference. *IEEE transactions on robotics*, 31(3):546–561, 2015.

[20] R. Detry, J. Papon, and L. Matthies. Task-oriented grasping with semantic and geometric scene understanding. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3266–3273. IEEE, 2017.

[21] A. Myers, C. L. Teo, C. Fermüller, and Y. Aloimonos. Affordance detection of tool parts from geometric features. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1374–1381. IEEE, 2015.

[22] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision, 2021.

[23] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[24] F.-J. Chu, R. Xu, and P. A. Vela. Learning affordance segmentation for real-world robotic manipulation via synthetic images. *IEEE Robotics and Automation Letters*, 4(2):1140–1147, 2019. doi:10.1109/LRA.2019.2894439.

[25] N. Vahrenkamp, L. Westkamp, N. Yamanobe, E. E. Aksoy, and T. Asfour. Part-based grasp planning for familiar objects. In *2016 IEEE-RAS 16th International Conference on Humanoid Robots (Humanoids)*, pages 919–925, 2016. doi:10.1109/HUMANOIDS.2016.7803382.

[26] Y. Song, P. Sun, Y. Ren, Y. Zheng, and Y. Zhang. Learning 6-dof fine-grained grasp detection based on part affordance grounding. *arXiv preprint arXiv:2301.11564*, 2023.

[27] J. Song, A. Tanwani, J. Ichnowski, M. Danielczuk, K. Sanders, J. Chui, J. A. Ojea, and K. Goldberg. Robust task-based grasping as a service. In *2020 IEEE 16th International Conference on Automation Science and Engineering (CASE)*, pages 22–28. IEEE, 2020.

[28] D. S. Hochbaum and V. Singh. An efficient algorithm for co-segmentation. In *2009 IEEE 12th International Conference on Computer Vision*, pages 269–276. IEEE, 2009.

[29] Y. Chai, V. Lempitsky, and A. Zisserman. Bicos: A bi-level co-segmentation method for image classification. In *2011 International conference on computer vision*, pages 2579–2586. IEEE, 2011.

[30] S. Amir, Y. Gandelsman, S. Bagon, and T. Dekel. Deep vit features as dense visual descriptors. *ECCVW What is Motion For?*, 2022.

[31] P. Sun, S. Chen, C. Zhu, F. Xiao, P. Luo, S. Xie, and Z. Yan. Going denser with open-vocabulary part segmentation. *arXiv preprint arXiv:2305.11173*, 2023.

[32] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision*, pages 405–421. Springer, 2020.

[33] M. Adamkiewicz, T. Chen, A. Caccavale, R. Gardner, P. Culbertson, J. Bohg, and M. Schwager. Vision-only robot navigation in a neural radiance world. *IEEE Robotics and Automation Letters*, 7(2):4606–4613, 2022.

[34] J. T. Barron, B. Mildenhall, M. Tancik, P. Hedman, R. Martin-Brualla, and P. P. Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5855–5864, 2021.

[35] J. T. Barron, B. Mildenhall, D. Verbin, P. P. Srinivasan, and P. Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5470–5479, 2022.

[36] L. Ma, X. Li, J. Liao, Q. Zhang, X. Wang, J. Wang, and P. V. Sander. Deblur-nerf: Neural radiance fields from blurry images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12861–12870, 2022.

[37] X. Huang, Q. Zhang, Y. Feng, H. Li, X. Wang, and Q. Wang. Hdr-nerf: High dynamic range neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18398–18408, 2022.

[38] S. Sabour, S. Vora, D. Duckworth, I. Krasin, D. J. Fleet, and A. Tagliasacchi. Robustnerf: Ignoring distractors with robust losses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20626–20636, 2023.

[39] J. Philip and V. Deschaintre. Radiance field gradient scaling for unbiased near-camera training. *arXiv preprint arXiv:2305.02756*, 2023.

[40] M. Tancik, E. Weber, E. Ng, R. Li, B. Yi, J. Kerr, T. Wang, A. Kristoffersen, J. Austin, K. Salahi, et al. Nerfstudio: A modular framework for neural radiance field development. *arXiv preprint arXiv:2302.04264*, 2023.

[41] P. Wang, Y. Liu, Z. Chen, L. Liu, Z. Liu, T. Komura, C. Theobalt, and W. Wang. F2-nerf: Fast neural radiance field training with free camera trajectories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4150–4159, 2023.

[42] J. T. Barron, B. Mildenhall, D. Verbin, P. P. Srinivasan, and P. Hedman. Zip-nerf: Anti-aliased grid-based neural radiance fields. *arXiv preprint arXiv:2304.06706*, 2023.

[43] T. Müller, A. Evans, C. Schied, and A. Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)*, 41(4):1–15, 2022.

[44] A. Chen, Z. Xu, A. Geiger, J. Yu, and H. Su. Tensorf: Tensorial radiance fields. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXII*, pages 333–350. Springer, 2022.

[45] S. Fridovich-Keil, G. Meanti, F. R. Warburg, B. Recht, and A. Kanazawa. K-planes: Explicit radiance fields in space, time, and appearance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12479–12488, 2023.

[46] A. Yu, S. Fridovich-Keil, M. Tancik, Q. Chen, B. Recht, and A. Kanazawa. Plenoxels: Radiance fields without neural networks. *arXiv preprint arXiv:2112.05131*, 2021.

[47] K. Park, U. Sinha, P. Hedman, J. T. Barron, S. Bouaziz, D. B. Goldman, R. Martin-Brualla, and S. M. Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *arXiv preprint arXiv:2106.13228*, 2021.

[48] Z. Li, Q. Wang, F. Cole, R. Tucker, and N. Snavely. Dynibar: Neural dynamic image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4273–4284, 2023.

[49] A. Pumarola, E. Corona, G. Pons-Moll, and F. Moreno-Noguer. D-NeRF: Neural Radiance Fields for Dynamic Scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[50] Z. Zhu, S. Peng, V. Larsson, W. Xu, H. Bao, Z. Cui, M. R. Oswald, and M. Pollefeys. Nice-slam: Neural implicit scalable encoding for slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12786–12796, June 2022.

[51] E. Sucar, S. Liu, J. Ortiz, and A. J. Davison. imap: Implicit mapping and positioning in real-time. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6229–6238, October 2021.

[52] A. Rosinol, J. J. Leonard, and L. Carlone. Nerf-slam: Real-time dense monocular slam with neural radiance fields, 2022.

[53] Y. Li, S. Li, V. Sitzmann, P. Agrawal, and A. Torralba. 3d neural scene representations for visuomotor control. In *Conference on Robot Learning*, pages 112–123. PMLR, 2022.

[54] D. Driess, I. Schubert, P. Florence, Y. Li, and M. Toussaint. Reinforcement learning with neural radiance fields. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[55] J. Kerr, L. Fu, H. Huang, Y. Avigal, M. Tancik, J. Ichnowski, A. Kanazawa, and K. Goldberg. Evo-neRF: Evolving neRF for sequential robot grasping of transparent objects. In *6th Annual Conference on Robot Learning*, 2022. URL https://openreview.net/forum?id=Bxr45keYrf.

[56] J. Ichnowski*, Y. Avigal*, J. Kerr, and K. Goldberg. Dex-NeRF: Using a neural radiance field to grasp transparent objects. In *Conference on Robot Learning (CoRL)*, 2020.

[57] A. Byravan, J. Humplik, L. Hasenclever, A. Brussee, F. Nori, T. Haarnoja, B. Moran, S. Bohez, F. Sadeghi, B. Vujatovic, and N. Heess. Nerf2real: Sim2real transfer of vision-guided bipedal motion skills using neural radiance fields, 2022.

[58] S. Zhi, T. Laidlow, S. Leutenegger, and A. J. Davison. In-place scene labelling and understanding with implicit scene representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15838–15847, 2021.

[59] Y. Siddiqui, L. Porzi, S. R. Buló, N. Müller, M. Nießner, A. Dai, and P. Kontschieder. Panoptic lifting for 3d scene understanding with neural fields, 2022.

[60] A. Kundu, K. Genova, X. Yin, A. Fathi, C. Pantofaru, L. Guibas, A. Tagliasacchi, F. Dellaert, and T. Funkhouser. Panoptic Neural Fields: A Semantic Object-Aware Neural Scene Representation. In *CVPR*, 2022.

[61] S. Kobayashi, E. Matsumoto, and V. Sitzmann. Decomposing nerf for editing via feature field distillation. In *NeurIPS*, volume 35, 2022. URL https://arxiv.org/pdf/2205.15585.pdf.

[62] V. Tschernezki, I. Laina, D. Larlus, and A. Vedaldi. Neural feature fusion fields: 3d distillation of self-supervised 2d image representations. In *Proceedings of the International Conference on 3D Vision (3DV)*, 2022.

[63] B. Li, K. Q. Weinberger, S. Belongie, V. Koltun, and R. Ranftl. Language-driven semantic segmentation. *arXiv preprint arXiv:2201.03546*, 2022.

[64] A. Zareian, K. D. Rosa, D. H. Hu, and S.-F. Chang. Open-vocabulary object detection using captions, 2021.

[65] G. Ghiasi, X. Gu, Y. Cui, and T.-Y. Lin. Scaling open-vocabulary image segmentation with image-level labels, 2022.

[66] X. Gu, T.-Y. Lin, W. Kuo, and Y. Cui. Open-vocabulary object detection via vision and language knowledge distillation, 2022.

[67] F. Liang, B. Wu, X. Dai, K. Li, Y. Zhao, H. Zhang, P. Zhang, P. Vajda, and D. Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip, 2023.

[68] X. Zhou, R. Girdhar, A. Joulin, P. Krähenbühl, and I. Misra. Detecting twenty-thousand classes using image-level supervision, 2022.

[69] M. Minderer, A. Gritsenko, A. Stone, M. Neumann, D. Weissenborn, A. Dosovitskiy, A. Mahendran, A. Arnab, M. Dehghani, Z. Shen, X. Wang, X. Zhai, T. Kipf, and N. Houlsby. Simple open-vocabulary object detection with vision transformers, 2022.

[70] K. Jatavallabhula, A. Kuwajerwala, Q. Gu, M. Omama, T. Chen, S. Li, G. Iyer, S. Saryazdi, N. Keetha, A. Tewari, J. Tenenbaum, C. de Melo, M. Krishna, L. Paull, F. Shkurti, and A. Torralba. Conceptfusion: Open-set multimodal 3d mapping. *arXiv*, 2023.

[71] C. Huang, O. Mees, A. Zeng, and W. Burgard. Visual language maps for robot navigation. *arXiv preprint arXiv:2210.05714*, 2022.

[72] S. Peng, K. Genova, C. Jiang, A. Tagliasacchi, M. Pollefeys, T. Funkhouser, et al. Openscene: 3d scene understanding with open vocabularies. *arXiv preprint arXiv:2211.15654*, 2022.

[73] G. Ghiasi, X. Gu, Y. Cui, and T.-Y. Lin. Open-vocabulary image segmentation. *arXiv preprint arXiv:2112.12143*, 2021.

[74] N. M. M. Shafiullah, C. Paxton, L. Pinto, S. Chintala, and A. Szlam. Clip-fields: Weakly supervised semantic fields for robotic memory. *arXiv preprint arXiv:2210.05663*, 2022.

[75] K. M. Jatavallabhula, A. Kuwajerwala, Q. Gu, M. Omama, T. Chen, S. Li, G. Iyer, S. Saryazdi, N. Keetha, A. Tewari, et al. Conceptfusion: Open-set multimodal 3d mapping. *arXiv preprint arXiv:2302.07241*, 2023.

[76] B. Chen, F. Xia, B. Ichter, K. Rao, K. Gopalakrishnan, M. S. Ryoo, A. Stone, and D. Kappler. Open-vocabulary queryable scene representations for real world planning. *arXiv preprint arXiv:2209.09874*, 2022.

[77] O. Mees and W. Burgard. Composing pick-and-place tasks by grounding language, 2021.

[78] L. Yu, Z. Lin, X. Shen, J. Yang, X. Lu, M. Bansal, and T. L. Berg. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1307–1315, 2018.

[79] M. Shridhar, L. Manuelli, and D. Fox. Cliport: What and where pathways for robotic manipulation. In *Proceedings of the 5th Conference on Robot Learning (CoRL)*, 2021.

[80] M. Shridhar, L. Manuelli, and D. Fox. Perceiver-actor: A multi-task transformer for robotic manipulation. In *Proceedings of the 6th Conference on Robot Learning (CoRL)*, 2022.

[81] A. Stone, T. Xiao, Y. Lu, K. Gopalakrishnan, K.-H. Lee, Q. Vuong, P. Wohlhart, B. Zitkovich, F. Xia, C. Finn, et al. Open-world object manipulation using pre-trained vision-language models. *arXiv preprint arXiv:2303.00905*, 2023.

[82] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022.

[83] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[84] S. F. Bhat, R. Birkl, D. Wofk, P. Wonka, and M. Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023.

[85] Guangcong, Z. Chen, C. C. Loy, and Z. Liu. Sparsenerf: Distilling depth ranking for few-shot novel view synthesis. *Technical Report*, 2023.

[86] M. Niemeyer, J. T. Barron, B. Mildenhall, M. S. M. Sajjadi, A. Geiger, and N. Radwan. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs, 2021.