Semantic Abstraction: Open-World 3D Scene Understanding from 2D Vision-Language Models

Anonymous Author(s)
Affiliation
Address
email

Abstract: We study open-world 3D scene understanding, a family of tasks that require agents to reason about their 3D environment with an open-set vocabulary and out-of-domain visual inputs – a critical skill for robots to operate in the unstructured 3D world. Towards this end, we propose Semantic Abstraction (SemAbs), a framework that equips 2D Vision-Language Models (VLMs) with new 3D spatial capabilities, while maintaining their zero-shot robustness. We achieve this abstraction using relevancy maps extracted from CLIP, and learn 3D spatial and geometric reasoning skills on top of those abstractions in a semantic-agnostic manner. We demonstrate the usefulness of SemAbs on two open-world 3D scene understanding tasks: 1) completing partially observed objects and 2) localizing hidden objects from language descriptions. Experiments show that SemAbs can generalize to novel vocabulary, materials/lighting, classes, and domains (i.e., real-world scans) from training on limited 3D synthetic data.

Keywords: 3D scene understanding, out-of-domain generalization, language

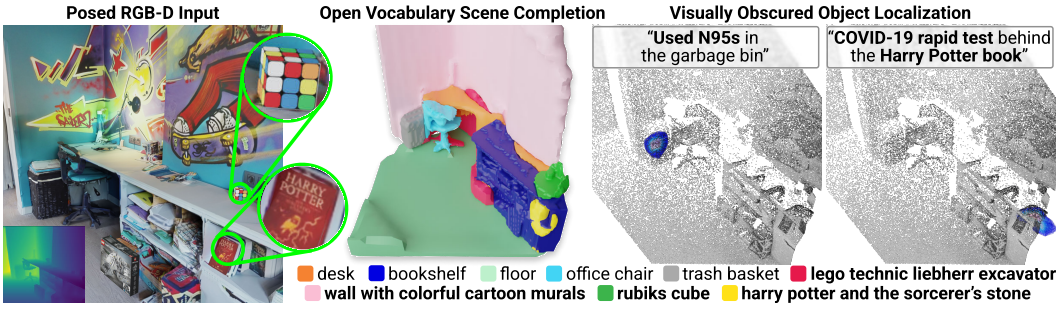

Figure 1: **Open-World 3D Scene Understanding**. Our approach, Semantic Abstraction, unlocks 2D VLM's capabilities to 3D scene understanding. Trained with a small simulated dataset, our model generalizes to unseen classes in a novel domain (*i.e.*, real world scans), even for small objects like "rubiks cube", long-tail concepts like "harry potter", and hidden objects like the "used N95s in the garbage bin".

1 Introduction

To assist people in their daily life, robots must recognize a large, context-dependent, and dynamic set of semantic categories from their visual sensors. However, if their 3D training environment contains only *common* categories like desks, office chairs, and bookshelves, how could they represent and localize *long-tail* objects, such as the "used N95s in the garbage bin" (Fig 1)?

This is an example of an open-world 3D scene understanding task (Fig. 2), a family of 3D vision-language tasks which encompasses open-set classification (evaluation on novel classes) with extra generalization requirements to novel vocabulary (*e.g.* synonyms of seen vocabulary), visual properties (*e.g.* lighting, textures), and domains (*e.g.* sim v.s. real). The core challenge in such tasks is the limited data: all existing 3D datasets [1, 2, 3, 4, 5, 6] are limited in diversity and scale compared to their internet-scale 2D counterparts [7, 8, 9], so training on them does not prepare robots for the open 3D world.

On the other hand, large-scale 2D vision language models (VLM) [7, 10, 9, 11] have demonstrated impressive performance on open-set image classification. Through exposure to many image-caption pairs

Submitted to the 6th Conference on Robot Learning (CoRL 2022). Do not distribute.

from the internet, they have acquired an unprecedented level of robustness to visual distributional shifts and a large repertoire of visual-semantic concepts and vocabulary. However, fine-tuning these pretrained models reduces their open-set classification robustness [7, 10, 12, 13]. Compared to internet-scale image-caption data, the finetuning datasets, including the existing 3D datasets [1, 2, 3, 4, 5, 6], show a significant reduction in scale and diversity, which causes finetuned models to become task-specialized and lose their generality. Hence, we investigate the following question:

How can we equip 2D VLMs with new 3D capabilities,
while maintaining their zero-shot robustness?

We propose **Semantic Abstraction (SemAbs)**, a framework for tackling visual-semantic reasoning in open-world 3D scene understanding tasks using 2D VLMs. We hypothesize that, while open-world visual-semantic reasoning requires exposure to internet-scale datasets, 3D spatial and geometric reasoning is tractable even with a limited synthetic dataset, and could generalize better if learned in a semantic-agnostic manner. For instance, instead of learning the concept of "behind the Harry Potter book", the 3D localization model only needs to learn the concept of "behind *that object*".

To achieve this abstraction, we leverage the relevancy maps extracted from 2D VLMs and used them as the "abstracted object" representation that is agnostic to their semantic labels. The SemAbs module factorizes into two submodules: 1) A semantic-aware wrapper that takes an input RGB-D image and object category label and outputs the relevancy map of a pre-trained 2D VLM model (i.e., CLIP [7]) for that label, and 2) a semantic-abstracted 3D module that uses the relevancy map to predict 3D occupancy. This 3D occupancy can either represent the 3D geometry of a partially observed object or the possible 3D locations of a hidden object (*e.g.* mask in the trash) conditioned on the language input.

While we only train the 3D network on a limited synthetic 3D dataset, it generalizes to any novel semantic labels, vocabulary, visual properties, and domains that the 2D VLM can generalize to. As a result, the SemAbs module inherits the VLM's visual robustness and open-set classification abilities while adding the spatial reasoning abilities it lacks [14, 15, 16]. In summary, our contributions are three-fold[1]:

- **Semantic Abstraction**, a framework for augmenting 2D VLMs with 3D reasoning capabilities for open-world 3D scene understanding tasks. By abstracting semantics away using relevancy maps, our SemAbs module generalizes to novel semantic labels, vocabulary, visual properties, and domains (*i.e.*, sim2real) despite being trained on only a limited synthetic 3D dataset.
- **Efficient Multiscale Relevancy Extraction.** To support Semantic Abstraction, we propose a multi-scale relevancy extractor for vision transformers [17] (ViTs), which robustly detects small, long-tail objects and achieves over $\times$**60** speed up from prior work [18].
- Two novel **Open-world 3D Scene Understanding** tasks (open-vocabulary semantic scene completion and visually obscured object localization), a data generation pipeline with AI2-THOR [1] simulator for the tasks, and a systematic open-world evaluation procedure. We believe these two tasks are important primitives for bridging existing robotics tasks to the open-world.

2 Related Works

2D Visual Language Models. The recent advancements in scaling up contrasting learning have enabled the training of large 2D vision language models (VLM). Through their exposure to millions of internet image-caption pairs, these VLMs [7, 10, 9, 11] acquire a remarkable level of zero-shot robustness – they can recognize a diverse set of long-tail semantic concepts robustly under distributional shifts (*e.g.* , video game renders, real-world photographs). However, the learned representations are often at the image level and, hence, not directly applicable to spatial-reasoning tasks which require pixel-level information. To address this issue, approaches were proposed to extract dense features [19, 20, 21] or relevancy

[1] All code, data, and models will be publicly available.

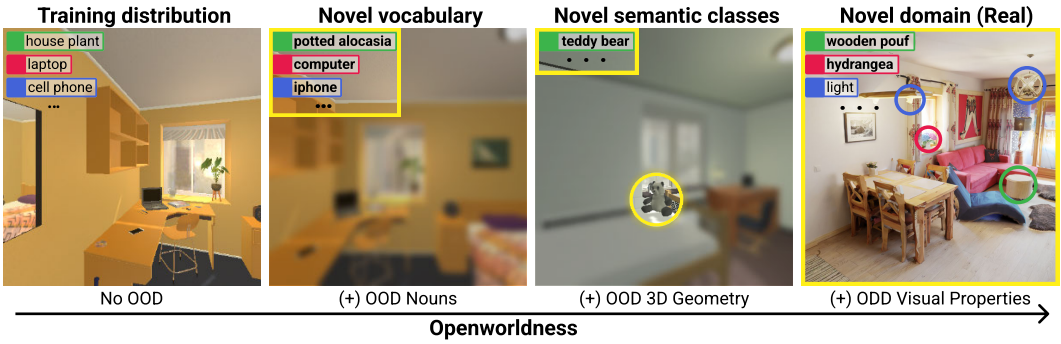

Figure 2: **Open-world generalization requirements** can build on top of each other, forming tiers of open-worldness (distinct property of each tier outlined yellow).

maps [18, 22]. These advancements complement our contribution, which leverages such techniques for interfacing 2D VLMs to open-world 3D scene understanding tasks.

Transferring 2D VLMs to 2D Applications. Encouraged by their impressive cabilities, many downstream tasks build around these pretrained VLMs [23, 24, 25, 14, 26, 27, 28], often finetuning them (either end-to-end or train learnable weights on top of the pretrained encoder) on a small task-specific dataset. However, many recent studies show that finetuning these VLMs significantly weakens their zero-shot robustness [7, 12, 10, 13, 19], which motivates a new paradigm for using pretrained models. By combining a *fixed* pretrained VLM with prompting [29] or attention extraction [19, 20, 21, 15], the new generation of zero-shot transfer techniques can inherit the VLM's robustness in their downstream tasks without overfitting to any visual domain. However, all existing zero-shot transfer techniques have been confined to the realms of 2D vision.

Closed-World 3D Scene Understanding Understanding the semantics and 3D structure of an unstructured environment is a fundamental capability for robots. Prior work has studied 3D object detection [30, 31, 32], semantic scene completion [33, 34, 35], language-informed object localization [6, 36]. However, such works have been limited to a small range of visual diversity and a set of common object categories, a limitation imposed by their 3D training data. Despite great efforts from the community [1, 2, 3, 4, 5, 6], existing 3D datasets' scale, diversity, and coverage pale in comparison to internet-scale image-text pairs [7, 8, 9]. However, our key insight is that visual-semantic reasoning can be learned separately from 3D reasoning. That is, with the right abstraction, the complex visual-semantic reasoning can be offloaded to 2D VLMs, while the 3D model specializes in semantic-agnostic spatial and geometry reasoning.

3 Method: Semantic Abstraction

Evidence from linguistics and cognitive science suggests that semantic concepts are much more diverse than spatial and geometric concepts. For example, while the average adult knows roughly 10,000 nouns for object categories, there are only 90 words in English for describing spatial relations [37]. In fact, there are so few spatial prepositions that they are typically considered a closed-class vocabulary [38]. Similarly for 3D geometry, the idea that our 3D world (especially of man-made objects) can be decomposed into sets of primitives [39] has been applied in many works, such as generalized cylinders [40] and block worlds [41]. In contrast, the space of nouns for semantic concepts is large and often cannot be further decomposed into common primitives. This suggests that while we may need a large and diverse dataset for semantic and visual concept learning, a relatively small 3D dataset could cover spatial and geometric reasoning. This observation motivates our approach, Semantic Abstraction.

3.1 Abstraction via Relevancy

The Semantic Abstraction (SemAbs) module (Fig 3c) takes as input an RGB-D image $\mathcal{I} \in \mathbb{R}^{H \times W}$ and an object class text label $\mathcal{T}$ (*e.g.* "biege armchair") and outputs the 3D occupancy $\mathcal{O}$ for objects of class $\mathcal{T}$. It factorizes into two submodules:

The semantic-aware wrapper (Fig 3c, green background) abstracts $\mathcal{I}$ and $\mathcal{T}$ into a relevancy map $\mathcal{R} \in \mathbb{R}^{H \times W}$, where each pixel's value denotes that pixel's contribution to CLIP [7]'s classification score for $\mathcal{T}$. Introduced for model explainability [22, 18], relevancy maps can be treated as a coarse localization

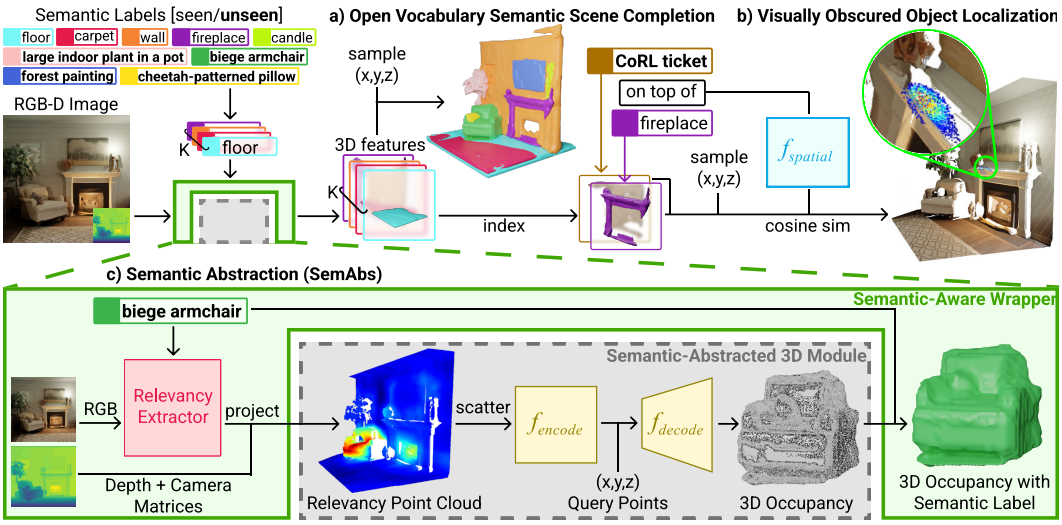

Figure 3: **Semantic Abstraction Overview.** Our framework can be applied to open-world 3D scene understanding tasks (a-b) using the SemAbs module (c). It consists of a **semantic-aware wrapper** (green background) that abstracts the input image and semantic label into a relevancy map, and a **semantic-abstracted 3D module** (grey background) that completes the projected relevancy map into a 3D occupancy. This abstraction allows our approach to generalize to long-tail semantic labels unseen (bolded) during 3D training, such as the "CoRL ticket on top of the fireplace".

of the text label. Using the depth image and camera matrices, $\mathcal{R}$ is projected into a 3D point cloud, $\mathcal{R}^{\text{proj}} = \{r_i\}_{i=1}^{H \times W}$ where $r_i \in \mathbb{R}^4$ (a 3D location with a scalar relevancy value). Only $\mathcal{R}^{\text{proj}}$, but neither the text $\mathcal{T}$ nor the image $\mathcal{I}$, is passed to the second submodule.

The semantic-abstracted 3D module (Fig 3c, grey background) treats the relevancy point cloud as the localization of a partially observed object and completes it into that object's 3D occupancy. To do this, we first scatter $\mathcal{R}^{\text{proj}}$ into a 3D voxel grid $\mathcal{R}^{\text{vox}} \in \mathbb{R}^{32 \times 32 \times 32}$. Then, we encode $\mathcal{R}^{\text{vox}}$ as a 3D feature volume: $f_{\text{encode}}(\mathcal{R}^{\text{vox}}) \mapsto Z \in \mathbb{R}^{D \times 32 \times 32 \times 32}$, where f_{encode} is a 3D CNN. Z can be sampled using trilinear interpolation to produce local features $\phi_q^Z \in \mathbb{R}^D$ for any 3D query point $q \in \mathbb{R}^3$. Finally, decoding ϕ_q^Z with a learned MLP f_{decode} gives us an occupancy probability for each point q, $f_{\text{decode}}\left(\phi_q^Z\right) \mapsto o(q) \in [0,1]$. In this submodule, both f_{encode} and f_{decode} (Fig. 3c, yellow boxes) are trained with the 3D dataset.

With $\mathcal{T}$ from the semantic-aware wrapper, we can interpret the semantic-abstracted 3D module's output as the 3D occupancy for the semantic label $\mathcal{T}$. Since the semantic-abstracted 3D module is agnostic to semantic labels – it only observes relevancy abstractions of the labels, but not semantic labels themselves – it generalizes to any semantic label that can be recognized by the 2D VLMs' relevancy maps, even though it was trained on a limited 3D dataset. In the next sections, we explain how to ensure the visual-semantic concepts are reliably recognized by the relevancy maps (Sec. 3.2), and how to apply SemAbs to 3D scene understanding tasks (Sec. 3.3).

3.2 An Efficient Multi-Scale Relevancy Extractor

We choose to use ViT-based CLIP [7] due to their superior performance over the ResNet variants. However, existing ViT relevancy techniques [18] often produce noisy, low-resolution maps that highlight irrelevant regions or miss small objects [42, 27].

To reduce noise from irrelevant regions, we use horizontal flipping and RGB augmentations [27] of $\mathcal{I}$ and obtain a relevancy map for each augmentation. To reliably detect small objects, we propose a multi-scale relevancy extractor that *densely* computes relevancy maps at different scales and locations in a sliding window fashion (an analogous relevancy-

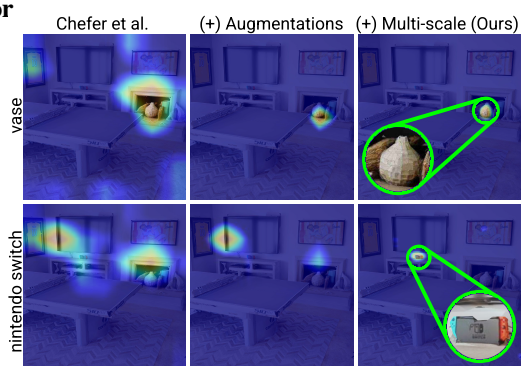

Figure 4: Our relevancy extractor robustly detects even small, long-tail objects, like the "nintendo switch".

based approach to Li et al.'s local attention-pooling). The final relevancy map is averaged across all augmentations and scales.

Qualitatively, while augmentations help in reducing irrelevant regions (*e.g.* Fig. 4, from localizing the entire fireplace in the 1st column to only the vase in the 2nd column), they still often miss small objects. However, with our multi-scale relevancy extractor, our relevancy maps can detect small objects such as the tiny "nintendo switch" (Fig. 4, bottom right).

To support efficient multi-scale relevancy extraction, we implemented a batch-parallelized ViT relevancy extractor. At 0.4 seconds per text label on average[2], we achieve over $\times$ **60** speed up compared to the non-parallelized implementation. With the rise in popularity of ViTs, we believe this efficient ViT relevancy extractor implementation will be a useful primitive for the research community and will be releasing it along with our dataset and codebase.

3.3 Applications to Open-world 3D Completion and Localization

In this section, we present how SemAbs can be applied to two open-world 3D scene understanding tasks that completes partially observed objects (OVSSC §3.3.1) and localize hidden objects (VOOL §3.3.2), regardless of whether it has seen those object classes during the 3D training process. When OVSSC's and VOOL's outputs are taken together, we get a simple but powerful open-world 3D scene representation. Through this scene representation, SemAbs unlocks CLIP's large repertoire of visual-semantic concepts and visual robustness to the realm of open-world 3D scene understanding.

3.3.1 Open Vocabulary Semantic Scene Completion (OVSSC).

Task: Given as input a single RGB-D image $\mathcal{I}$ and a set of K object class labels, represented as open-vocabulary text $\{\mathcal{T}_k\}_{k=1}^{K}$, the task is to complete the 3D geometry of all the partially observed objects referred by $\{\mathcal{T}_k\}_{k=1}^{K}$. In contrast to the closed-world semantic scene completion task [33], where the object list is fixed for all scenes, OVSSC formulation accounts for open-world context-dependency, which allows the class labels to change from scene to scene and from training to testing.

Applying SemAbs to OVSSC: We can perform this task by simply applying the SemAbs module K times, once per object label. First, the semantic-aware wrapper abstracts $\mathcal{I}$ and $\{\mathcal{T}_k\}_{k=1}^{K}$ into K relevancy point clouds $\{\mathcal{R}_k^{\text{proj}}\}_{k=1}^{K}$. Then, the semantic-abstracted 3D module extracts K feature volumes $\{Z_k\}_{k=1}^{K}$, used to decode K 3D occupancies $\{\mathcal{O}_k\}_{k=1}^{K}$. The final OVSSC output is a point-wise argmax of occupancy probability over the K classes. Query points with the maximum occupancy probability less than a threshold c are labeled as empty.

Data generation and training: We generate training data for f_{encode} and f_{decode} using our custom version of the AI2-THOR [1] simulator, which adds the extra functionality for outputting occupancy and semantic category for query 3D points. To generate views in the rooms, we spawn the robot at random locations and Z-rotations and render RGB-D images. We filter out views with too few objects. We sample 15000 query points in the view bound to supervise the occupancy prediction with Binary Cross Entropy loss.

3.3.2 Visually Obscured Object Localization (VOOL)

Motivation: Even after completing partially visible objects in the scene, OVSSC's output may not be complete – the scene may contain visually obscured or hidden objects (e.g., the ticket on the fireplace in Fig. 3b). However those objects might be important for the human user in downstream tasks (e.g., fetch the ticket). In such cases, the human user can provide additional hints to the robot using natural language to help localize the object, which motivates our VOOL task.

Task: Given as input a single RGB-D image $\mathcal{I}$ of a 3D scene and a description $\mathcal{D}$ of an object's location in the scene, the task is to predict the possible occupancies of the referred object. Similar to prior works [43], we assume that the description is in the standard linguistic representation of an object's place [37]. Specifically, $\mathcal{D} = \langle \mathcal{T}_{\text{target}}, \mathcal{S}, \mathcal{T}_{\text{ref}} \rangle$, where the open-vocabulary target object label $\mathcal{T}_{\text{target}}$ (*e.g.*

[2] measured over 10 repetitions with 100 text labels on a GTX 1080 Ti machine with a light load. We provide extra details in the appendix.

"rapid test") is the visually obscured object to be located, the open-vocabulary reference object label $\mathcal{T}_{\text{ref}}$ (*e.g.* "harry potter book") is presumably a visible object, and the closed-vocabulary spatial preposition $\mathcal{S}$ describes $\mathcal{T}_{\text{target}}$'s location with respect to $\mathcal{T}_{\text{ref}}$ (*e.g.* "behind").

Applying SemAbs to VOOL: We propose to learn an embedding $f_{\text{spatial}}(\mathcal{S}) \in \mathbb{R}^{2D}$ for each spatial relation (Fig. 3b, blue box), since the set of spatial prepositions is small and finite [37], and use them as follows. As in OVSSC (§3.3.1), we can get occupancy feature volumes Z_{target} and Z_{ref} for the target and reference objects respectively. Given a set of fixed query points $Q = \{q_i\}_{i=1}^{N}$, we extract their local feature point clouds $\phi_Q^{Z_{\text{target}}}, \phi_Q^{Z_{\text{ref}}} \in R^{N \times D}$, then concatenate them to get a new feature pointcloud $\phi_Q^{Z_{\text{target}} \| Z_{\text{ref}}} \in R^{N \times 2D}$. Finally, to get the 3D occupancy $\mathcal{O}$ (*i.e.*, the localization output for this task), we perform a point-wise cosine-similarity between $\phi_Q^{Z_{\text{target}} \| Z_{\text{ref}}}$ and $f_{\text{spatial}}(\mathcal{S})$. For this task, we only need to learn the spatial embeddings f_{spatial} conditioned on the occupancy information encoded in Z_{target} and Z_{ref}, since semantic reasoning of $\mathcal{T}_{\text{target}}$ and $\mathcal{T}_{\text{ref}}$ has been offloaded to the SemAbs (and, therefore, to CLIP).

Data generation and training: We focus on six common spatial prepositions: behind, left of, right of, in front, on top of, and inside. As in OVSSC (§3.3.1), we use our custom AI2-THOR [1] simulator to generate training data. Specifically, we define "on top of" and "inside" using AI2-THOR's receptacle information, while "behind", "left of", "right of", and "in front" are defined in a viewer-centric fashion (more details in the supplementary material). To handle ambiguous descriptions (*e.g.* "hand sanitizer in the cabinet" when there are multiple cabinets), our ground truth positives contain all points consistent with the description (*e.g.* points in the "inside" receptacle for all cabinets in view are labeled positive).

4 Experiments

We design experiments to systemically investigate Semantic Abstraction's open-world generalization abilities in 3D scene understanding. We categorize generalization scenarios along two axes: 1) **Visual generalization**: This axis encompasses generalization to novel 2D visual properties (*i.e.*, novel lighting + materials) as well as 3D geometry (*i.e.*, unseen meshes of the same class). 2) **Vocabulary generalization**: This axis evaluates the learner on novel words, which could be synonyms of seen words, parent categories (*e.g.* "tomato" to "fruit"), child categories (*e.g.* "house plant" to "hydrangea"), novel referred attributes (*e.g.* "sofa" to "leather sofa"). Some scenarios demand generalization along both of these axes, such as novel semantic categories (both unseen 3D geometry and unseen object text label).

Benchmark design for OOD evaluation. Here, we define "novel" as to whether a concept was observed during the *3D* training process. We do not control or change the 2D VLM training process.

- **Novel Rooms**: we follow AI2-THOR [1]'s split and hold out the last 5 rooms in each room type for testing, while using the rest as training rooms.
- **Novel Visual**: we use AI2-THOR's randomize material and lighting functionality to generate more varied 2D visual conditions of the novel rooms.
- **Novel Vocab**: we manually picked out 25 classes in the simulator which had natural synonyms (*e.g.* "cell phone" to "mobile phone", complete list included in supplementary materials) and replace their occurrences in text inputs (*i.e.*, object label in OVSSC, description in VOOL) to these synonyms.
- **Novel Class**: we hold out 6 classes out of the 202 classes in AI2-THOR. For OVSSC, any view which contains one of these novel classes is held out for testing. FOr VOOL, any scene with a description that contains a novel class (as either the target or reference object) is held out.
- **Novel Domain**: We also provide qualitative evaluations (Fig. 5) using real world scans from HM3d [4].

Metrics & Baselines. For both tasks, we measure the learner's performance with voxel IoU of dimension $32 \times 32 \times 32$, where the ground truth of each task is generated as described in §3.3.1 for OVSSC and §3.3.2 for VOOL. We compare with following categories of baselines:

- **Semantic-Aware (SemAware).** To evaluate the effectiveness of semantic abstraction, we design baselines that perform visual-semantic reasoning themselves by taking RGB point clouds as input (instead of relevancy point clouds). For OVSSC, the baseline's feature point cloud is supervised with BCE to be aligned (*i.e.*, using cosine similarity) with the correct object class text embedding. Similarly,

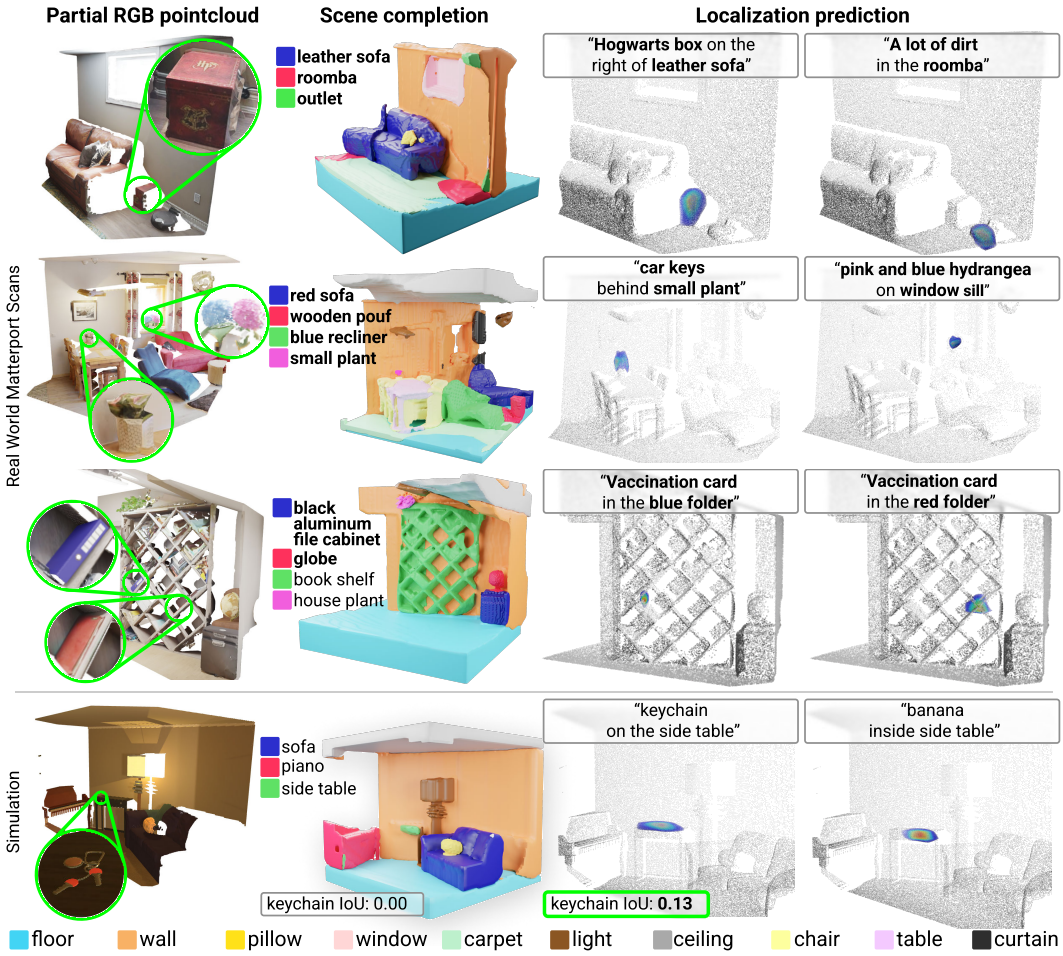

Figure 5: **Semantic Abstraction inherits CLIP's visual-semantic reasoning skils** From distinguishing colors (*e.g.* "red folder" v.s. "blue folder") to recognizing cultural (*e.g.* "hogwarts box") and long-tail semantic concepts (*e.g.* "roomba", "hydrangea"), our approach offloads such visual-semantic reasoning challenges to CLIP. Indoing so, its learned 3D spatial and geometric reasoning skills transfers sim2real in a zero-shot manner.

for VOOL, the feature point cloud's cosine similarity with an embedding of the entire localization description (*e.g.* "N95s inside the trash") is supervised using BCE to the target object occupancy.

- **Semantic & Spatial Abstraction (ClipSpatial).** What if we off-load both 3D and visual-semantic reasoning to 2D VLMs (instead of just the latter, as in our approach)? To answer this, we designed CLIPSpatial, a VOOL baseline that uses relevancy maps for the entire description (*e.g.* "N95s inside the trash") as input. We expect this approach to perform poorly since it has been demonstrated that current 2D VLMs [7, 8] struggles with spatial relations [14, 15, 16].
- **Naive Relevancy Extraction. (SemAbs + [18])** To investigate the effects of relevancy quality, we replace our multi-scale relevancy extractor with Chefer et al. [18]'s approach.

The first two baselines can be seen as two extremes on a spectrum of how much we rely on pretrained 2D VLMs. While SemAware tries to learn both semantics and 3D spatial reasoning itself, CLIPSpatial delegates both to pretrained 2D VLMs. Our approach is in between these two extremes, designed such that it builds on 2D VLMs' visual robustness and open-set classification strengths while addressing its spatial reasoning weaknesses.

4.1 Results

Tab. 1 and 2 summarize the quantitative results, and Fig. 5 shows real world qualitative results tested on Matterport scenes [4]. More results and comparisons can be found in supp. material.

Semantic Abstraction simplifies open-world generalization. As shown in Tab. 1 and Tab. 2, SemAware baselines especially struggle to generalize to novel semantic classes. We hypothesize that observing text embeddings during training causes these baselines to specialize to this distribution of embeddings, such that some synonyms (*i.e.*, novel vocab) and most novel class labels are out-of-distribution inputs. Our approach significantly outperforms the SemAware baseline across the board, suggesting that Semantic Abstraction not only simplifies open-world generalization but is also a strong inductive bias for learning the task.

Approach	Novel			
	Room	**Visual**	**Vocab**	**Class**
SemAware	37.6	37.4	13.0	2.0
SemAbs+[18]	40.2	39.8	6.6	4.5
Ours	**61.4**	**58.0**	**24.3**	**32.3**

Table 1: **Open Vocabulary Semantic Scene Completion**

Semantic & Spatial Abstraction generalizes poorly. The CLIPSpatial baseline consistently achieves over 10.0 IoU worse than our approach in all generalization scenarios (Tab. 2). This indicates that our design of learning spatial reasoning instead of relying on CLIP was crucial to semantic abstraction's performance.

Semantic Abstraction requires high quality relevancy maps. Intuitively, our model can only do well if its only input – the relevancy maps – robustly highlights the target object, especially when the object is small. The average IoU doubles (novel room and visual, Table 1), triples (across the board for Table 2), quadruples (novel vocab, Table 1) or more (novel class, Table 1) from [SemAbs + [18]] to our approach, which demonstrates the importance of our multi-scale saliency extractor in a successful application of Semantic Abstraction.

Approach	Novel			
	Room	**Visual**	**Vocab**	**Class**
SemAware	19.0	18.6	18.3	12.2
SemAbs+[18]	8.5	8.5	10.8	11.7
ClipSpatial	18.1	15.5	20.2	27.1
Ours	**28.7**	**25.9**	**31.1**	**34.0**

Table 2: **Visually Obscured Object Localization**

Inheriting VLM zero-shot robustness enables Sim2Real transfer. Despite training in simulation, our approach can perform completion and localization from real-world matterport scans [4]. From novel vocabulary for describing materials (*e.g.* "leather sofa", "wooden pouf"), colors (*e.g.* "red folder", "blue folder"), conjunctions of them (*e.g.* "black aluminum file cabinet") to novel semantic classes (*e.g.* "roomba", "Hogwarts box"), our approach unlocks CLIP's visual robustness and large repertoire of visual-semantic concepts to 3D scene understanding. When OVSSC's completion misses objects (*e.g.* "hydrangea" in the second row, "keychain" in the bottom row), VOOL can still localize them (*e.g.* +0.13 IoU for keychain).

4.2 Limitations and assumptions

Our key assumption is that geometry and spatial reasoning can be done in a semantic agnostic manner. This assumption does not hold when we care about completing detailed object geometry, which is often semantic-dependent. However, we believe that coarse 3D reasoning with paired semantic labels may already useful for open-world robotic applications such as object retrieval. Second, while we have assumed a simple description syntax for object localization, incorporating natural language understanding into open-world 3D scene understanding is an important extension of our work. Lastly, we assumed that relevancy maps highlight only relevant regions. Limited by current VLMs, our relevancy extractor can sometimes fail to distinguish highly related concepts (*e.g.* book v.s. bookshelf). However, since Semantic Abstraction is VLM-agnostic, this limitation could be addressed by using other VLMs.

5 Conclusion

We proposed Semantic Abstraction, a framework that equips 2D VLMs with 3D spatial capabilities for open-world 3D scene understanding tasks. By abstracting RGB and semantic label inputs as relevancy maps, our learned 3D spatial and geometric reasoning skills generalizes to novel vocabulary, visual properties, semantic classes, and domains. Since scene completion and 3D object localization are fundamental primitives for many robotic tasks, we believe our approach can assist in bridging currently close-world robotics tasks, such as object retrieval, into the open-world formulation. More importantly, we hope our framework encourages roboticists to embrace building systems around large pretrained models, as a path towards open-world robotics.

References

[1] E. Kolve, R. Mottaghi, W. Han, E. VanderBilt, L. Weihs, A. Herrasti, D. Gordon, Y. Zhu, A. Gupta, and A. Farhadi. AI2-THOR: An Interactive 3D Environment for Visual AI. *arXiv*, 2017.

[2] S. Song, S. P. Lichtenberg, and J. Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 567–576, 2015.

[3] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015.

[4] S. K. Ramakrishnan, A. Gokaslan, E. Wijmans, O. Maksymets, A. Clegg, J. M. Turner, E. Undersander, W. Galuba, A. Westbury, A. X. Chang, M. Savva, Y. Zhao, and D. Batra. Habitat-matterport 3d dataset (HM3d): 1000 large-scale 3d environments for embodied AI. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. URL https://openreview.net/forum?id=-v4OuqNs5P.

[5] A. Szot, A. Clegg, E. Undersander, E. Wijmans, Y. Zhao, J. Turner, N. Maestre, M. Mukadam, D. Chaplot, O. Maksymets, A. Gokaslan, V. Vondrus, S. Dharur, F. Meier, W. Galuba, A. Chang, Z. Kira, V. Koltun, J. Malik, M. Savva, and D. Batra. Habitat 2.0: Training home assistants to rearrange their habitat. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[6] D. Z. Chen, A. X. Chang, and M. Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *European Conference on Computer Vision*, pages 202–221. Springer, 2020.

[7] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[8] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in Neural Information Processing Systems*, 34, 2021.

[9] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022.

[10] H. Pham, Z. Dai, G. Ghiasi, H. Liu, A. W. Yu, M.-T. Luong, M. Tan, and Q. V. Le. Combined scaling for zero-shot transfer learning. *arXiv preprint arXiv:2111.10050*, 2021.

[11] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021.

[12] M. Wortsman, G. Ilharco, M. Li, J. W. Kim, H. Hajishirzi, A. Farhadi, H. Namkoong, and L. Schmidt. Robust fine-tuning of zero-shot models. *arXiv preprint arXiv:2109.01903*, 2021.

[13] A. Andreassen, Y. Bahri, B. Neyshabur, and R. Roelofs. The evolution of out-of-distribution robustness throughout fine-tuning. *arXiv preprint arXiv:2106.15831*, 2021.

[14] S. Subramanian, W. Merrill, T. Darrell, M. Gardner, S. Singh, and A. Rohrbach. Reclip: A strong zero-shot baseline for referring expression comprehension. *arXiv preprint arXiv:2204.05991*, 2022.

[15] S. Y. Gadre, M. Wortsman, G. Ilharco, L. Schmidt, and S. Song. Clip on wheels: Zero-shot object navigation as object localization and exploration. *arXiv preprint arXiv:2203.10421*, 2022.

[16] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

[17] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[18] H. Chefer, S. Gur, and L. Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 397–406, 2021.

[19] C. Zhou, C. C. Loy, and B. Dai. Denseclip: Extract free dense labels from clip. *arXiv preprint arXiv:2112.01071*, 2021.

[20] Y. Rao, W. Zhao, G. Chen, Y. Tang, Z. Zhu, G. Huang, J. Zhou, and J. Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18082–18091, 2022.

[21] J. Li, G. Shakhnarovich, and R. A. Yeh. Adapting clip for phrase localization without further training. *arXiv preprint arXiv:2204.03647*, 2022.

[22] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[23] R. Mokady, A. Hertz, and A. H. Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021.

[24] H. Song, L. Dong, W.-N. Zhang, T. Liu, and F. Wei. Clip models are few-shot learners: Empirical studies on vqa and visual entailment. *arXiv preprint arXiv:2203.07190*, 2022.

[25] Z. Gao, J. Liu, S. Chen, D. Chang, H. Zhang, and J. Yuan. Clip2tv: An empirical study on transformer-based methods for video-text retrieval. *arXiv preprint arXiv:2111.05610*, 2021.

[26] M. Wang, J. Xing, and Y. Liu. Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*, 2021.

[27] N. Zabari and Y. Hoshen. Semantic segmentation in-the-wild without seeing any segmentation examples. *CoRR*, abs/2112.03185, 2021. URL https://arxiv.org/abs/2112.03185.

[28] 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.

[29] A. Zeng, A. Wong, S. Welker, K. Choromanski, F. Tombari, A. Purohit, M. Ryoo, V. Sindhwani, J. Lee, V. Vanhoucke, et al. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv preprint arXiv:2204.00598*, 2022.

[30] Shuran Song and J. Xiao. Sliding Shapes for 3D object detection in depth images. 2014.

[31] Shuran Song and J. Xiao. Deep sliding shapes for amodal 3D object detection in rgb-d images. 2016.

[32] J. Hou, A. Dai, and M. Nießner. 3d-sis: 3d semantic instance segmentation of rgb-d scans. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2019.

[33] S. Song, F. Yu, A. Zeng, A. X. Chang, M. Savva, and T. Funkhouser. Semantic scene completion from a single depth image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1746–1754, 2017.

[34] A. Dai, D. Ritchie, M. Bokeloh, S. Reed, J. Sturm, and M. Nießner. Scancomplete: Large-scale scene completion and semantic segmentation for 3d scans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4578–4587, 2018.

[35] A. Avetisyan, M. Dahnert, A. Dai, M. Savva, A. X. Chang, and M. Nießner. Scan2cad: Learning cad model alignment in rgb-d scans. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 2614–2623, 2019.

[36] J. Roh, K. Desingh, A. Farhadi, and D. Fox. Languagerefer: Spatial-language model for 3d visual grounding. In *Conference on Robot Learning*, pages 1046–1056. PMLR, 2022.

[37] L. Barbara and R. Jackendoff. "what" and "where" in spatial language and spatial cognition. *Behavioral and brain sciences*, 16:217–238, 1993.

[38] W. G. Hayward and M. J. Tarr. Spatial language and spatial representation. *Cognition*, 55(1):39–84, 1995.

[39] S. Tulsiani, H. Su, L. J. Guibas, A. A. Efros, and J. Malik. Learning shape abstractions by assembling volumetric primitives. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2635–2643, 2017.

[40] I. Binford. Visual perception by computer. In *IEEE Conference of Systems and Control*, 1971.

[41] A. Gupta, A. A. Efros, and M. Hebert. Blocks world revisited: Image understanding using qualitative geometry and mechanics. In *European Conference on Computer Vision*, pages 482–496. Springer, 2010.

[42] S. Shen, L. H. Li, H. Tan, M. Bansal, A. Rohrbach, K.-W. Chang, Z. Yao, and K. Keutzer. How much can clip benefit vision-and-language tasks? *arXiv preprint arXiv:2107.06383*, 2021.

[43] N. Liu, S. Li, Y. Du, J. Tenenbaum, and A. Torralba. Learning to compose visual relations. *Advances in Neural Information Processing Systems*, 34, 2021.