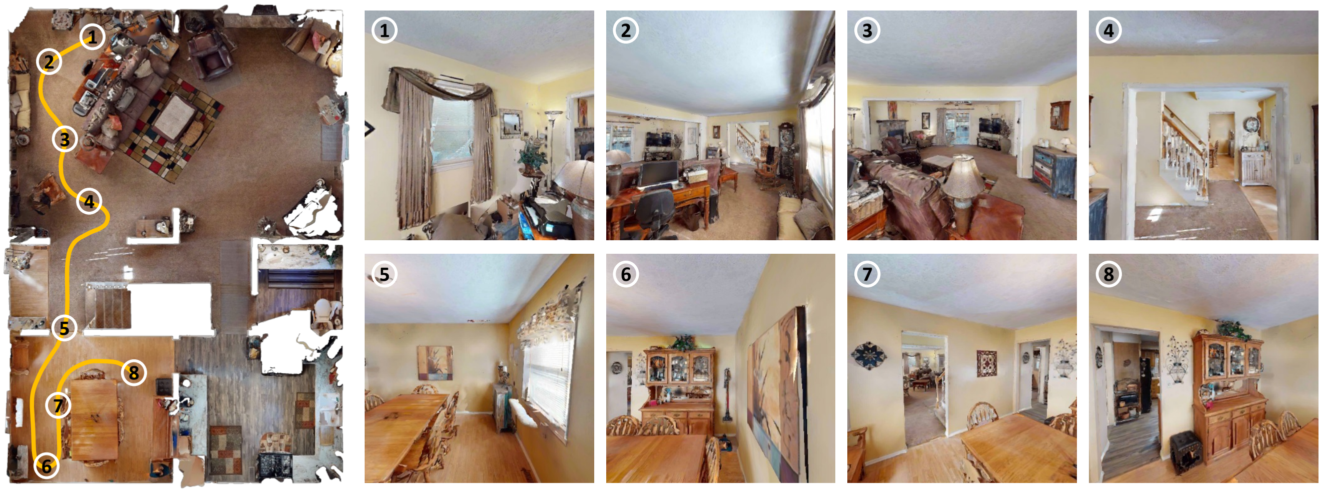
3D Concept Learning and Reasoning from Multi-View Images

Yining Hong[1], Chunru Lin[2], Yilun Du[3],
Zhenfang Chen[5], Joshua B. Tenenbaum[3], Chuang Gan[4, 5]
[1]UCLA, [2]Shanghai Jiaotong University, [3]MIT CSAIL,
[4]UMass Amherst, [5]MIT-IBM Watson AI Lab
https://vis-www.cs.umass.edu/3d-clr/



Concept:
Q: Are there any televisions? **A:** Yes

Q: Is there a sofa in the room with a printer? **A**: Yes

Counting:
Q: How many chairs are close to the table in the room with plant on the cabinet? **A**: 6

Q: How many rooms have sofas? **A**: 1

Relation:
Q: Facing the computer from the curtain, is there a lamp on the right? **A**: Yes

Q: What's on the cabinet in the smaller room? **A:** Plant

Comparison:
Q: Are there fewer pictures in the larger room than the other room? **A**: No
Q: Is the computer closer to a printer or a lamp? **A:** Printer

Figure 1. An exemplar scene with multi-view images and question-answer pairs of our 3DMV-VQA dataset. 3DMV-VQA contains four question types: concept, counting, relation, comparison. Orange words denote semantic concepts; blue words denote the relations.

Abstract

Humans are able to accurately reason in 3D by gathering multi-view observations of the surrounding world. Inspired by this insight, we introduce a new large-scale benchmark for 3D multi-view visual question answering (3DMV-VQA). This dataset is collected by an embodied agent actively moving and capturing RGB images in an environment using the Habitat simulator. In total, it consists of approximately 5k scenes, 600k images, paired with 50k questions. We evaluate various state-of-the-art models for visual reasoning on our benchmark and find that they all perform poorly. We suggest that a principled approach for 3D reasoning from multi-view images should be to infer a compact 3D representation of the world from the multi-view images, which is further grounded on open-vocabulary semantic concepts, and then to execute reasoning on these 3D representations. As the first step towards this approach, we propose a novel 3D concept learning and reasoning (3D-CLR) framework that seamlessly combines these components via neural fields, 2D pre-trained vision-language models, and neural reasoning operators. Experimental results suggest that our framework outperforms baseline models by a large margin, but the challenge remains largely unsolved. We further perform an in-depth analysis of the challenges and highlight potential future directions. .

1. Introduction

Visual reasoning, the ability to composite rules on internal representations to reason and answer questions about visual scenes, has been a long-standing challenge in the field of artificial intelligence and computer vision. Several datasets [23, 33, 69] have been proposed to tackle this challenge. However, they mainly focus on visual reasoning on 2D single-view images. Since 2D single-view images only cover a limited region of the whole space, such reasoning inevitably has several weaknesses, including occlusion, and failing to answer 3D-related questions about the entire scene that we are interested in. As shown in Fig. 1, it's difficult, even for humans, to count the number of chairs in a scene due to the object occlusion, and it's even harder to infer 3D relations like "closer" from a single-view 2D image.

On the other hand, there's strong psychological evidence that human beings conduct visual reasoning in the underlying 3D representations [55]. Recently, there have been several works focusing on 3D visual question answering [2,16,62,64]. They mainly use traditional 3D representations (*e.g.,* point clouds) for visual reasoning. This is inconsistent with the way human beings perform 3D reasoning in real life. Instead of being given an entire 3D representation of the scene at once, humans will actively walk around and explore the whole environment, ingesting image observations from different views and converting them into a holistic 3D representation that assists them in understanding and reasoning about the environment. Such abilities are crucial for many embodied AI applications, such as building assistive robots.

To this end, we propose the novel task of 3D visual reasoning from multi-view images taken by active exploration of an embodied agent. Specifically, we generate a large-scale benchmark, 3DMV-VQA (3D multi-view visual question answering), that contains approximately 5k scenes and 50k question-answering pairs about these scenes. For each scene, we provide a collection of multi-view image observations. We generate this dataset by placing an embodied agent in the Habitat-Matterport environment [47], which actively explores the environment and takes pictures from different views. We also obtain scene graph annotations from the Habitat-Matterport 3D semantics dataset (HM3DSem) [61], including ground-truth locations, segmentations, semantic information of the objects, as well as relationships among the objects in the environments, for model diagnosis. To evaluate the models' 3D reasoning abilities on the entire environment, we design several 3D-related question types, including concept, counting, relation and comparison.

Given this new task, the key challenges we would like to investigate include: 1) how to efficiently obtain the compact visual representation to encode crucial properties (*e.g.,* semantics and relations) by integrating all incomplete observations of the environment in the process of active exploration for 3D visual reasoning? 2) How to ground the semantic concepts on these 3D representations that could be leveraged for downstream tasks, such as visual reasoning? 3) How to infer the relations among the objects, and perform step-by-step reasoning?

As the first step to tackling these challenges, we propose a novel model, 3D-CLR (3D Concept Learning and Reasoning). First, to efficiently obtain a compact 3D representation from multi-view images, we use a neural-field model based on compact voxel grids [57] which is both fast to train and effective at storing scene properties in its voxel grids. As for concept learning, we observe that previous works on 3D scene understanding [1,3] lack the diversity and scale with regard to semantic concepts due to the limited amount of paired 3D-and-language data. Although large-scale vision-language models (VLMs) have achieved impressive performances for zero-shot semantic grounding on 2D images, leveraging these pretrained models for effective open-vocabulary 3D grounding of semantic concepts remains a challenge. To address these challenges, we propose to encode the features of a pre-trained 2D vision-language model (VLM) into the compact 3D representation defined across voxel locations. Specifically, we use the CLIP-LSeg [37] model to obtain features on multi-view images, and propose an alignment loss to map the features in our 3D voxel grid to 2D pixels. By calculating the dot-product attention between the 3D per-point features and CLIP language embeddings, we can ground the semantic concepts in the 3D compact representation. Finally, to answer the questions, we introduce a set of neural reasoning operators, including FILTER, COUNT, RELATION operators and so on, which take the 3D representations of different objects as input and output the predictions.

We conduct experiments on our proposed 3DMV-VQA benchmark. Experimental results show that our proposed 3D-CLR outperforms all baseline models a lot. However, failure cases and model diagnosis show that challenges still exist concerning the grounding of small objects and the separation of close object instances. We provide an in-depth analysis of the challenges and discuss potential future directions.
To sum up, we have the following contributions in this paper.

- We propose the novel task of 3D concept learning and reasoning from multi-view images.
- By having robots actively explore the embodied environments, we collect a large-scale benchmark on 3D multi-view visual question answering (3DMV-VQA).
- We devise a model that incorporates a neural radiance field, 2D pretrained vision and language model, and neural reasoning operators to ground the concepts and perform 3D reasoning on the multi-view images. We illustrate that our model outperforms all baseline models.
- We perform an in-depth analysis of the challenges of this new task and highlight potential future directions.

2. Related Work

Visual Reasoning There have been numerous tasks focusing on learning visual concepts from natural language, including visually-grounded question answering [18, 19], text-image retrieval [59] and so on. Visual reasoning has drawn much attention recently as it requires human-like understanding of the visual scene. A wide variety of benchmarks have been created over the recent years [7, 8, 23, 27, 33, 69]. However, they mainly focus on visual reasoning from 2D single-view images, while there's strong psychological evidence that human beings perform visual reasoning on the underlying 3D representations. In this paper, we propose the novel task of visual reasoning from multi-view images, and collect a large-scale benchmark for this task. In recent years, numerous visual reasoning models have also been proposed, ranging from attention-based methods [5, 30], graph-based methods [28], to models based on large pretrained vision-language model [9, 38]. These methods model the reasoning process implicitly with neural networks. Neural-symbolic methods [6, 40, 65] explicitly perform symbolic reasoning on the objects representations and language representations. They use perception models to extract 2D masks as a first step, and then execute operators and ground concepts on these pre-segmented masks, but are limited to a set of pre-defined concepts on simple scenes. [26] proposes to use the feature vectors from occupancy networks [42] to do visual reasoning in the 3D space. However, they also use a synthetic dataset, and learn a limited set of semantic concepts from scratch. We propose to learn 3D neural field features from 2D multi-view real-world images, and incorporate a 2D VLM for open-vocabulary reasoning.

3D Reasoning Understanding and reasoning about 3D scenes has been a long-standing challenge. Recent works focus on leveraging language to explore 3D scenes, such as object captioning [3, 4] and object localization from language [1, 17, 29]. Our work is mostly related to 3D Visual Question Answering [2, 16, 62, 64] as we both focus on answering questions and reasoning about 3D scenes. However, these works use point clouds as 3D representations, which diverts from the way human beings perform 3D reasoning. Instead of being given an entire 3D representation all at once, human beings would actively move and explore the environment, integrating multi-view information to get a compact 3D representation. Therefore, we propose 3D reasoning from multi-view images. In addition, since 3D assets paired with natural language descriptions are hard to get in real-life scenarios, previous works struggle to ground open-vocabulary concepts. In our work, we leverage 2D VLMs for zero-shot open-vocabulary concept grounding in the 3D space.

Embodied Reasoning Our work is also closely related to Embodied Question Answering (EQA) [11, 67] and Interactive Question Answering (IQA) [22, 35], which also involve an embodied agent exploring the environment and answering the question. However, the reasoning mainly focuses on the outcome or the history of the navigation on 2D images and does not require a holistic 3D understanding of the environment. There are also works [12, 20, 51, 54, 56, 68] targeting instruction following in embodied environments, in which an agent is asked to perform a series of tasks based on language instructions. Different from their settings, for our benchmark an embodied agent actively explores the environment and takes multi-view images for 3D-related reasoning.

Neural Fields Our approach utilizes neural fields to parameterize an underlying 3D compact representations of scenes for reasoning. Neural field models (*e.g.,* [43]) have gained much popularity since they can reconstruct a volumetric 3D scene representation from a set of images. Recent works [21, 24, 57, 66] have pushed it further by using classic voxel-grids to explicitly store the scene properties (*e.g.*, density, color and feature) for rendering, which allows for real-time rendering and is utilized by this paper. Neural fields have also been used to represent dynamic scenes [14, 44], appearance [43, 45, 49, 53, 63], physics [34], robotics [32, 52], acoustics [39] and more general multi-modal signals [13]. There are also some works that integrate semantics or language in neural fields [31, 60]. However, they mainly focus on using language for manipulation, editing or generation. [26] leverages neural descriptor field [52] for 3D concept grounding. However, they require ground-truth occupancy values to train the neural field, which can not be applied to real-world scenes. In this paper, we propose to leverage voxel-based neural radiance field [57] to get the compact representations for 3D visual reasoning.

3. Dataset Generation

3.1. Multi-View Images

Our dataset includes 5k 3D scenes from the Habitat-Matterport 3D Dataset (HM3D) dataset [47], and approximately 600k images rendered from the 3D scenes. The images are rendered via Habitat [50, 58].

Scene Generation We build our benchmark on top of the HM3DSem dataset [61], which is a large-scale dataset of 3D real-world indoor scenes with densely annotated semantics. It consists of 142,646 object instance annotations across 216 3D spaces and 3,100 rooms within those spaces. HM3D dataset uses texture information to annotate pixel-accurate object boundaries, which provides large-scale object annotations and ensures the scale, quality, and diversity of 3D visual reasoning questions of our benchmark.

To construct a benchmark that covers questions of different difficulty levels, it's crucial that we include 3D scenes of different scales in our benchmark. We start with single rooms in HM3D scenes, which has an appropriate amount of semantic concepts and relationships to base some simple questions on. To get the scale of single rooms, we calculate bounding boxes of rooms according to floor instance segmentations. We then proceed to generate bounding boxes for scenes with multiple adjacent rooms. For more complex holistic scene understanding, we also include whole-house scenes, which may contain tens of rooms. Overall, the 3DMV-VQA benchmark contains three levels of scenes (2000 single-room scenes, 2000 multi-room scenes and 100 whole-house scenes).

Image Rendering After we get the bounding box of each scene, we load the scene into the Habitat simulator. We also put a robot agent with an RGB sensor at a random initial point in the bounding box. The data is collected via exploration of the robot agent. Specifically, at each step of the data collection process, we sample a navigable point and make the agent move to the point along the shortest path. When the agent has arrived at a point, we rotate the agent 30° along z-axis for 12 times so that the agent can observe the 360° view of the scene at the position. It can also look up and down, with a random mild angle from $[-10°, 10°]$ along the x-axis. A picture is taken each time the agent rotates to a new orientation. In total 12 pictures are taken from each point. While traveling between points, the robot agent further takes pictures. We also exploit a policy such that when the camera is too far from or too close to an object and thus the agent cannot see anything, we discard the bad-view images.

3.2. Questions and Answers

We pair each scene with machine-generated questions from pre-defined templates. All questions are open-ended and can be answered with a single word (samples in Fig. 1).

Concepts and Relationships To generate questions and answers, we utilize the semantic annotations of HM3DSem [61] to get the semantic concepts and their bounding boxes, as well as the bounding boxes of the rooms. We merge semantic concepts with similar meanings (*e.g.,*, L-shaped sofa to sofa, desk chair / computer chair e.g. to chair). We also define 11 relationships: inside, above, below, on the top of, close, far, large, small, between, on the left, and on the right. Before generating questions, we first generate a scene graph for each scene containing all concepts and relationships.

Question Types We define four types of questions: concept, counting, relation and comparison.

- **Concept.** Conceptual questions query if there's an object of a certain semantic concept in the scene, or whether there's a room containing the objects of the semantic concept.
- **Counting.** Counting-related questions ask about how many instances of a semantic concept are in the scene, or how many rooms contain objects of the semantic concept.
- **Relation.** Relational questions ask about the 11 relationships and their compositions. Based on the number of relations in a question, we have one-hop to three-hop questions for the relation type.
- **Comparison.** The comparison question type focuses on the comparison of two objects, two semantic concepts or two rooms. It can be combined with the relational concepts to compare two objects (*e.g.,* larger, closer to, more left *etc*). It also compares the number of instances of two semantic concepts, or the number of objects of certain concepts in different rooms.

Bias Control. Similar to previous visual reasoning benchmarks [26, 33], we use machine-generated questions since the generation process is fully controllable so that we can avoid dataset bias. Questions are generated from pre-defined templates, and transformed into natural language questions with associated semantic concepts and relationships from the scene. We manually define 41 templates for question generation. We use depth-first search to generate questions. We perform bias control based on three perspectives: template counts, answer counts, and concept counts. For selecting templates, we sort the templates each time we generate a question to ensure a balanced question distribution. We force a flat answer distribution for each template by rejection sampling. Specifically, once we generate a question and an answer, if the number of the questions having the same answer and template is significantly larger than other answers, we discard it and continue searching. Once we find an answer that fits in the ideal answer distribution, we stop the depth-first searching for this question. We also force a flat concept distribution for each template using the same method. In addition to controlling the number of concepts mentioned in the templates, we also control the number of relation tuples consisting of the same concept sets.

4. Method

Fig. 2 illustrates an overview of our framework. Specifically, our framework consists of three steps. First, we learn a 3D compact representation from multi-view images using neural field. And then we propose to leverage pre-trained 2D vision-and-language model to ground concepts on 3D space. This is achieved by 1) generating 2D pixel features using CLIP-LSeg; 2) aligning the features of 3D voxel grid and 2D pixel features from CLIP- LSeg [37]; 3) dot-product attention between the 3D features and CLIP language features [37]. Finally, to perform visual reasoning, we propose neural reasoning operators, which execute the question step by step on the 3D compact representation and outputs a final answer. For example, we use FILTER operators to ground semantic concepts on the 3D representation, GET_INSTANCE to get all instances of a semantic class, and COUNT_RELATION to count how many pairs of the two semantic classes have the queried relation.

4.1. Learning 3D Compact Scene Representations

Neural radiance fields [43] are capable of learning a 3D representation that can reconstruct a volumetric 3D scene representation from a set of images. Voxel-based meth-

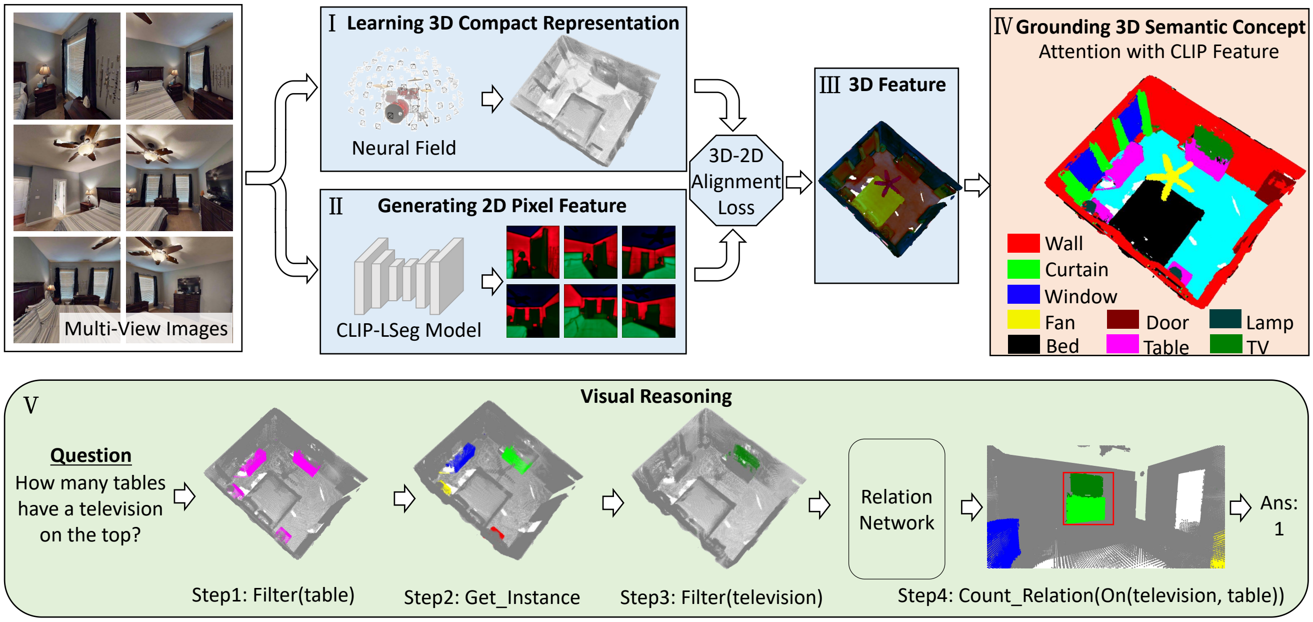

Multi-View Images
I Learning 3D Compact Representation
Neural Field
II Generating 2D Pixel Feature
CLIP-LSeg Model
3D-2D Alignment Loss
III 3D Feature
IV Grounding 3D Semantic Concept
Attention with CLIP Feature
Wall
Curtain
Window
Fan
Door
Lamp
Bed
Table
TV
V
Visual Reasoning
Question
How many tables have a television on the top?
Relation Network
Ans: 1
Step1: Filter(table)
Step2: Get_Instance
Step3: Filter(television)
Step4: Count_Relation(On(television, table))

Figure 2. An overview of our 3D-CLR framework. First, we learn a 3D compact scene representation from multi-view images using neural fields (I). Second, we use CLIP-LSeg model to get per-pixel 2D features (II). We utilize a 3D-2D alignment loss to assign features to the 3D compact representation (III). By calculating the dot-product attention between the 3D per-point features and CLIP language embeddings, we could get the concept grounding in 3D (IV). Finally, the reasoning process is performed via a set of neural reasoning operators, such as FILTER, GET_INSTANCE and COUNT_RELATION (V). Relation operators are learned via relation networks.

ods [21, 24, 57, 66] speed up the learning process by explicitly storing the scene properties (*e.g.*, density, color and feature) in its voxel grids. We leverage Direct Voxel Grid Optimization (DVGO) [57] as our backbone for 3D compact representation for its fast speed. DVGO stores the learned density and color properties in its grid cells. The rendering of multi-view images is by interpolating through the voxel grids to get the density and color for each sampled point along each sampled ray, and integrating the colors based on the rendering alpha weights calculated from densities according to quadrature rule [41]. The model is trained by minimizing the L2 loss between the rendered multi-view images and the ground-truth multi-view images. By extracting the density voxel grid, we can get the 3D compact representation (*e.g.,* By visualizing points with density greater than 0.5, we can get the 3D representation as shown in Fig. 2 I.)

4.2. 3D Semantic Concept Grounding

Once we extract the 3D compact representation of the scene, we need to ground the semantic concepts for reasoning from language. Recent work from [26] has proposed to ground concepts from paired 3D assets and question-answers. Though promising results have been achieved on synthetic data, it is not feasible for open-vocabulary 3D reasoning in real-world data, since it is hard to collect large-scale 3D vision-and-language paired data. To address this challenge, our idea is to leverage pre-trained 2D vision and language model [46, 48] for 3D concept grounding in real-world scenes. But how can we map 2D concepts into 3D neural field representations? Note that 3D compact representations can be learned from 2D multi-view images and that each 2D pixel actually corresponds to several 3D points along the ray. Therefore, it's possible to get 3D features from 2D per-pixel features. Inspired by this, we first add a feature voxel grid representation to DVGO, in addition to density and color, to represent 3D features. We then apply CLIP-LSeg [37] to learn per-pixel 2D features, which can be attended to by CLIP concept embeddings. We use an alignment loss to align 3D features with 2D features so that we can perform concept grounding on the 3D representations.

2D Feature Extraction. To get per-pixel features that can be attended by concept embeddings, we use the features from language-driven semantic segmentation (CLIP-LSeg) [37], which learns 2D per-pixel features from a pre-trained vision-language model (*i.e.,* [46]). Specifically, it uses the text encoder from CLIP, trains an image encoder to produce an embedding vector for each pixel, and calculates the scores of word-pixel correlation by dot-product. By outputting the semantic class with the maximum score of each pixel, CLIP-LSeg is able to perform zero-shot 2D semantic segmentation.

3D-2D Alignment. In addition to density and color, we also store a 512-dim feature in each grid cell in the compact representation. To align the 3D per-point features with 2D per-pixel features, we calculate an L1 loss between each pixel and each 3D point sampled on the ray of the pixel. The overall L1 loss along a ray is the weighted sum of all the pixel-point alignment losses, with weights same as the rendering weights: $\mathcal{L}_{\text{feature}} = \sum_{i=1}^{K} w_i(\|\boldsymbol{f_i} - F(\boldsymbol{r})\|)$, where $\boldsymbol{r}$ is a ray corresponding to a 2D pixel, $F(\boldsymbol{r})$ is the 2D feature from CLIP-LSeg, K is the total number of sampled points along the ray and $\boldsymbol{f_i}$ is the feature of point i by interpolating through the feature voxel grid, w_i is the rendering weight.

Concept Grounding through Attention. Since our feature voxel grid representation is learnt from CLIP-LSeg, by calculating the dot-product attention $< \boldsymbol{f}, \boldsymbol{v} >$ between per-point 3D feature $\boldsymbol{f}$ and the CLIP concept embeddings $\boldsymbol{v}$, we can get zero-shot view-independent concept grounding and semantic segmentations in the 3D representation, as is presented in Fig. 2 IV.

4.3. Neural Reasoning Operators

Finally, we use the grounded semantic concepts for 3D reasoning from language. We first transform questions into a sequence of operators that can be executed on the 3D representation for reasoning. We adopt a LSTM-based semantic parser [65] for that. As [26, 40], we further devise a set of operators which can be executed on the 3D representation. Please refer to **Appendix** for a full list of operators.

Filter Operators. We filter all the grid cells with a certain semantic concept.

Get_Instance Operators. We implement this by utilizing DBSCAN [15], an unsupervised algorithm which assigns clusters to a set of points. Specifically, given a set of points in the 3D space, it can group together the points that are closely packed together for instance segmentation.

Relation Operators. We cannot directly execute the relation on the 3D representation as we have not grounded relations. Thus, we represent each relation using a distinct neural module (which is practical as the vocabulary of relations is limited [36]). We first concatenate the voxel grid representations of all the referred objects and feed them into the relation network. The relation network consists of three 3D convolutional layers and then three 3D deconvolutional layers. A score is output by the relation network indicating whether the objects have the relationship or not. Since vanilla 3D CNNs are very slow, we use Sparse Convolution [10] instead. Based on the relations asked in the questions, different relation modules are chosen.

5. Experiments

5.1. Experimental Setup

Evaluation Metric. We report the visual question answering accuracy on the proposed 3DMV-VQA dataset w.r.t the four types of questions. The train/val/test split is 7:1:2.

Implementation Details For 3D compact representations, we adopt the same architectures as DVGO, except skipping the coarse reconstruction phase and directly training the fine reconstruction phase. After that, we freeze the density voxel grid and color voxel grid, for the optimization of the feature voxel grid only. The feature grid has a world size of 100 and feature dim of 512. We train the compact representations for 100,000 iterations and the 3D features for another 20,000 iterations. For LSeg, we use the official demo model, which has the ViT-L/16 image encoder and CLIP's ViT-B/32 text encoder. We follow the official script for inference and use multi-scale inference. For DBSCAN, we use an epsilon value of 1.5, minimum samples of 2, and we use L1 as the clustering method. For the relation networks, each relation is encoded into a three-layer sparse 3D convolution network with hidden size 64. The output is then fed into a one-layer linear network to produce a score, which is normalized by sigmoid function. We use cross-entropy loss to train the relation networks, and we use the one-hop relational questions with "yes/no" answers to train the relation networks.

5.2. Baselines

Our baselines range from vanilla neural networks, attention-based methods, fine-tuned from large-scale VLM, and graph-based methods, to neural-symbolic methods.

- **LSTM**. The question is transferred to word embeddings which are input into a word-level LSTM [25]. The last LSTM hidden state is fed into a multi-layer perceptron (MLP) that outputs a distribution over answers. This method is able to model question-conditional bias since it uses no image information.
- **CNN+LSTM**. The question is encoded by the final hidden states from LSTM. We use a resnet-50 to extract frame-level features of images and average them over the time dimension. The features are fed to an MLP to predict the final answer. This is a simple baseline that examines how vanilla neural networks perform on 3DMV-VQA.
- **3D-Feature+LSTM**. We use the 3D features we get from 3D-2D alignment and downsample the voxel grids using 3D-CNN as input, concatenated with language features from LSTM and fed to an MLP.
- **MAC** [30]. MAC utilizes a Memory, Attention and Composition cell to perform iterative reasoning process. Like CNN+LSTM, we use the average pooling over multi-view images as the feature map.
- **MAC(V)**. We treat the multi-view images along a trajectory as a video. We modify the MAC model by applying a temporal attention unit across the video frames to generate a latent encoding for the video.
- **NS-VQA** [65]. This is a 2D version of our 3D-CLR model. We use CLIP-LSeg to ground 2D semantic concepts from multi-view images, and the relation network also takes the 2D features as input. We execute the operators on each image and max pool from the answers to get our final predictions.

Methods	Concept	Counting	Relation	Comparison	Overall
Q-type (rand.)	49.4	10.7	21.6	49.2	26.4
LSTM	53.4	15.3	24.0	55.2	29.8
CNN+LSTM	57.8	22.1	35.2	59.7	37.8
MAC	62.4	19.7	47.8	62.3	46.7
MAC(V)	60.0	24.6	51.6	65.9	50.0
NS-VQA	59.8	21.5	33.4	61.6	38.0
ALPRO	65.8	12.7	42.2	68.2	43.3
LGCN	56.2	19.5	35.5	66.7	39.1
3D-Feature+LSTM	61.2	22.4	49.9	61.3	48.2
3D-CLR (Ours)	**66.1**	**41.3**	**57.6**	**72.3**	**57.7**

Table 1. Question-answering accuracy of 3D visual reasoning baselines on different question types.

- **ALPRO** [38]. ALPRO is a video-and-language pre-training framework. A transformer model is pretrained on large webly-source video-text pairs and can be used for downstream tasks like Video Question answering.
- **LGCN** [28]. LGCN represents the contents in the video as a location-aware graph by incorporating the location information of an object into the graph construction.

5.3. Experimental Results

Result Analysis. We summarize the performances for each question type of baseline models in Table 1. All models are trained on the training set until convergence, tuned on the validation set, and evaluated on the test set. We provide detailed analysis below.

First, for the examination of language-bias of the dataset, we find that the performance of LSTM is only slightly higher than random and frequency, and all other baselines outperform LSTM a lot. This suggests that there's little language bias in our dataset. Second, we observe that encoding temporal information in MAC (*i.e.,* MAC(V)) is better than average-pooling of the features, especially in counting and relation. This suggests that average-pooling of the features may cause the model to lose information from multi-view images, while attention on multi-view images helps boost the 3D reasoning performances. Third, we also find that fine-tuning on large-scale pretrained model (*i.e.,* ALPRO) has relatively high accuracies in concept-related questions, but for counting it's only slightly higher than the random baseline, suggesting that pretraining on large-scale video-language dataset may improve the model's perception ability, but does not provide the model with the ability to tackle with more difficult reasoning types such as counting. Next, we find that LGCN has poor performances on the relational questions, indicating that building a location-aware graph over 2D objects still doesn't equip the model with 3D location reasoning abilities. Last but not least, we find that 3D-based baselines are better than their 2D counterparts. 3D-Feature+LSTM performs well on the 3D-related questions, such as counting and relation, than most of the image-based baselines. Compared with 3D-CLR, NS-VQA can perform well in the conceptual questions. However, it underperforms 3D-CLR a lot in counting and relation, suggesting that these two types of questions require the holistic 3D understanding of the entire 3D scenes. Our 3D-CLR outperforms other baselines by a large margin, but is still far from satisfying. From the accuracy of the conceptual question, we can see that it can only ground approximately 66% of the semantic concepts. This indicates that our 3DMV-VQA dataset is indeed very challenging.

Qualitative Examples. In Fig. 3, we show four qualitative examples. From the examples, we show that our 3D-CLR can infer an accurate 3D representation from multi-view images, as well as ground semantic concepts on the 3D representations to get the semantic segmentations of the entire scene. Our 3D-CLR can also learn 3D relationships such as "close", "largest", "on top of" and so on. However, 3D-CLR also fails on some questions. For the third scene in the qualitative examples, it fails to ground the concepts "mouse" and "printer". Also, it cannot accurately count the instances sometimes. We give detailed discussions below.

5.4. Discussions

We perform an in-depth analysis to understand the challenge of this dataset. We leverage the modular design of our 3D-CLR, replacing individual components of the framework with ground-truth annotations for model diagnosis. The result is shown in Fig 4. 3D-CLR w/ Semantic denotes our model with ground-truth semantic concepts from HM3DSem annotations. 3D-CLR w/ Instance denotes that we have ground-truth instance segmentations of semantic concepts. From Fig. 3 and Fig. 4, we summarize several key challenges of our benchmark:

Very close object instances From Fig. 4, we can see that even with ground-truth semantic labeling of the 3D points, 3D-CLR still has unsatisfying results on counting questions. This suggests that the instance segmentations provided by DBSCAN are not accurate enough. From the top two qualitative examples in Fig. 3, we can also see that if two chairs

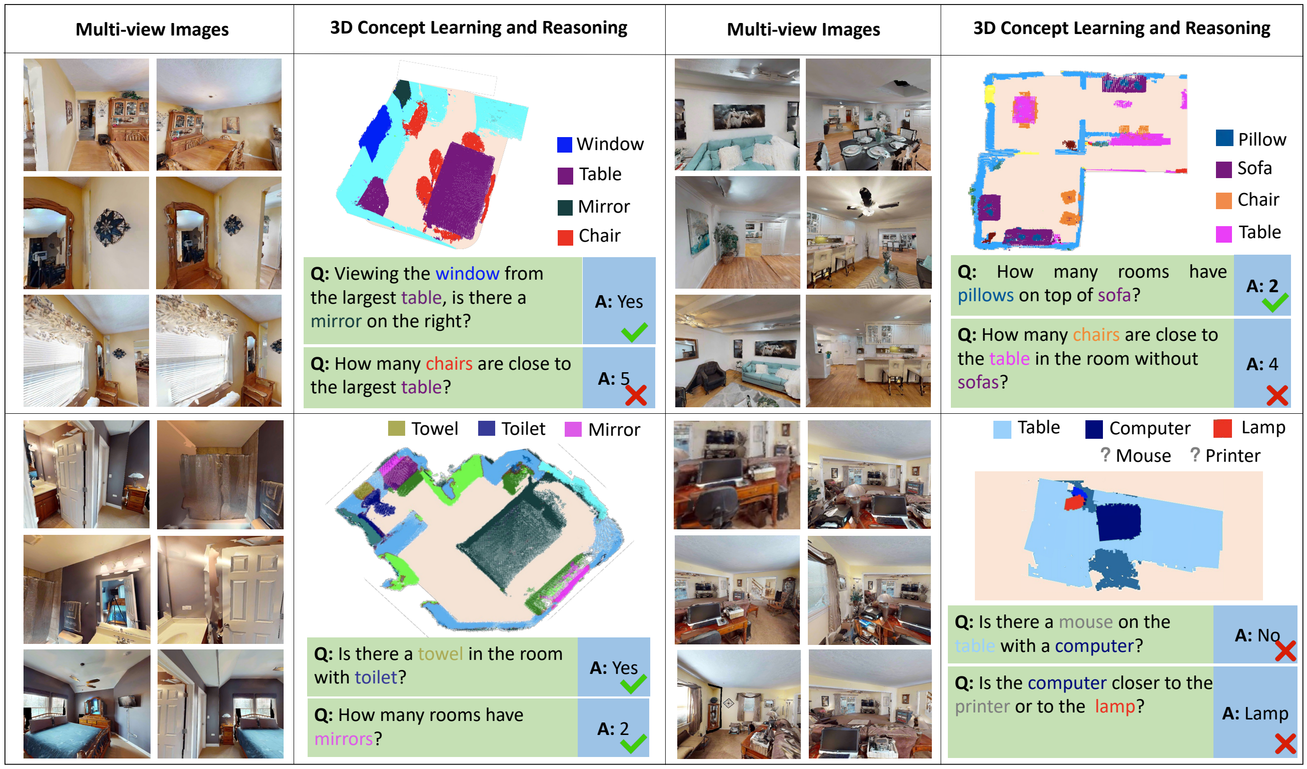

Multi-view Images
3D Concept Learning and Reasoning
Multi-view Images
3D Concept Learning and Reasoning
🟦 Window
🟪 Table
■ Mirror
🟥 Chair
Q: Viewing the window from the largest table, is there a mirror on the right?
A: Yes ✓
Q: How many chairs are close to the largest table?
A: 5 ✗
🟦 Pillow
🟪 Sofa
🟧 Chair
🟪 Table
Q: How many rooms have pillows on top of sofa?
A: 2 ✓
Q: How many chairs are close to the table in the room without sofas?
A: 4 ✗
■ Towel
🟦 Toilet
🟪 Mirror
Q: Is there a towel in the room with toilet?
A: Yes ✓
Q: How many rooms have mirrors?
A: 2 ✓
🟦 Table
🟦 Computer
🟥 Lamp
? Mouse
? Printer
Q: Is there a mouse on the table with a computer?
A: No ✗
Q: Is the computer closer to the printer or to the lamp?
A: Lamp ✗

Figure 3. Qualitative examples of our 3D-CLR. We can see that 3D-CLR can ground most of the concepts and answer most questions correctly. However, it still fails sometimes, mainly because it cannot separate close object instances and ground small objects.

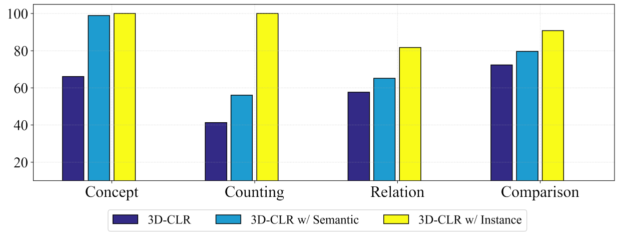


Figure 4. Model diagnosis of our 3D-CLR.

contact each other, DBSCAN will not tell them apart and thus have poor performance on counting. One crucial future direction is to improve unsupervised instance segmentations on very close object instances.

Grounding small objects Fig. 4 suggests that 3D-CLR fails to ground a large portion of the semantic concepts, which hinders the performance. From the last example in Fig. 3, we can see that 3D-CLR fails to ground small objects like "computer mouse". Further examination indicates there are two possible reasons: 1) CLIP-LSeg fails to assign the right features to objects with limited pixels; 2) The resolution of feature voxel grid is not high enough and therefore small objects cannot be represented in the compact representation. An interesting future direction would be learning exploration policies that enable the agents to get closer to uncertain objects that cannot be grounded.

Ambiguity on 3D relations Even with ground-truth semantic and instance segmentations, the performance of the relation network still needs to be improved. We find that most of the failure cases are correlated to the "inside" relation. From the segmentations in Fig. 3, we can see that 3D-CLR is unable to ground the objects in the cabinets. A potential solution can be joint depth and segmentation predictions.

6. Conclusion

In this paper, we introduce the novel task of 3D reasoning from multi-view images. By placing embodied robot that actively explores indoor environments, we collect a large-scale benchmark named 3DMV-VQA. We also propose a new 3D-CLR model that incorporates neural field, 2D VLM, as well as reasoning operators for this task and illustrate its effectiveness. Finally, we perform an in-depth analysis to understand the challenges of this dataset and also point out potential future directions. We hope that 3DMV-VQA can be used to push the frontiers of 3D reasoning.

Acknowledgements. This work was supported by the MIT-IBM Watson AI Lab, DARPA MCS, DSO grant DSOCO21072, and gift funding from MERL, Cisco, Sony, and Amazon. We would also like to thank the computation support from AiMOS, a server cluster for the IBM Research AI Hardware Center.

References

[1] Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas J. Guibas. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In *ECCV*, 2020. 2, 3

[2] Daich Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19107–19117, 2022. 2, 3

[3] Dave Zhenyu Chen, Angel X. Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *ECCV*, 2020. 2, 3

[4] Dave Zhenyu Chen, Ali Gholami, Matthias Nießner, and Angel X. Chang. Scan2cap: Context-aware dense captioning in rgb-d scans. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3192–3202, 2021. 3

[5] Z Chen, L Ma, W Luo, and KKY Wong. Weakly-supervised spatio-temporally grounding natural sentence in video. In *ACL*, 2019. 3

[6] Zhenfang Chen, Jiayuan Mao, Jiajun Wu, Kwan-Yee Kenneth Wong, Joshua B Tenenbaum, and Chuang Gan. Grounding physical concepts of objects and events through dynamic visual reasoning. *ICLR*, 2021. 3

[7] Zhenfang Chen, Peng Wang, Lin Ma, Kwan-Yee K Wong, and Qi Wu. Cops-ref: A new dataset and task on compositional referring expression comprehension. In *CVPR*, 2020. 3

[8] Zhenfang Chen, Kexin Yi, Yunzhu Li, Mingyu Ding, Antonio Torralba, Joshua B Tenenbaum, and Chuang Gan. Comphy: Compositional physical reasoning of objects and events from videos. In *ICLR*, 2022. 3

[9] Zhenfang Chen, Qinhong Zhou, Yikang Shen, Yining Hong, Hao Zhang, and Chuang Gan. See, think, confirm: Interactive prompting between vision and language models for knowledge-based visual reasoning. *arXiv preprint arXiv:2301.05226*, 2023. 3

[10] Spconv Contributors. Spconv: Spatially sparse convolution library. https://github.com/traveller59/spconv, 2022. 6

[11] Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied question answering. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2135–213509, 2018. 3

[12] Mingyu Ding, Yan Xu, Zhenfang Chen, David Daniel Cox, Ping Luo, Joshua B Tenenbaum, and Chuang Gan. Embodied concept learner: Self-supervised learning of concepts and mapping through instruction following. In *CoRL*. 3

[13] Yilun Du, M. Katherine Collins, B. Joshua Tenenbaum, and Vincent Sitzmann. Learning signal-agnostic manifolds of neural fields. In *Advances in Neural Information Processing Systems*, 2021. 3

[14] Yilun Du, Yinan Zhang, Hong-Xing Yu, Joshua B. Tenenbaum, and Jiajun Wu. Neural radiance flow for 4d view synthesis and video processing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 3

[15] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, 1996. 6

[16] Yasaman Etesam, Leon Kochiev, and Angel X Chang. 3dvqa: Visual question answering for 3d environments. In *2022 19th Conference on Robots and Vision (CRV)*, pages 233–240. IEEE, 2022. 2, 3

[17] Mingtao Feng, Zhen Li, Qi Li, Liang Zhang, Xiangdong Zhang, Guangming Zhu, Hui Zhang, Yaonan Wang, and Ajmal S. Mian. Free-form description guided 3d visual graph network for object grounding in point cloud. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3702–3711, 2021. 3

[18] Chuang Gan, Yandong Li, Haoxiang Li, Chen Sun, and Boqing Gong. Vqs: Linking segmentations to questions and answers for supervised attention in vqa and question-focused semantic segmentation. In *ICCV*, pages 1811–1820, 2017. 3

[19] Siddha Ganju, Olga Russakovsky, and Abhinav Kumar Gupta. What's in a question: Using visual questions as a form of supervision. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6422–6431, 2017. 3

[20] Xiaofeng Gao, Qiaozi Gao, Ran Gong, Kaixiang Lin, Govind Thattai, and Gaurav S Sukhatme. Dialfred: Dialogue-enabled agents for embodied instruction following. *arXiv*, 2022. 3

[21] Stephan J. Garbin, Marek Kowalski, Matthew Johnson, Jamie Shotton, and Julien P. C. Valentin. Fastnerf: High-fidelity neural rendering at 200fps. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14326–14335, 2021. 3, 5

[22] Daniel Gordon, Aniruddha Kembhavi, Mohammad Rastegari, Joseph Redmon, Dieter Fox, and Ali Farhadi. Iqa: Visual question answering in interactive environments. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4089–4098, 2018. 3

[23] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6325–6334, 2017. 2, 3

[24] Peter Hedman, Pratul P. Srinivasan, Ben Mildenhall, Jonathan T. Barron, and Paul E. Debevec. Baking neural radiance fields for real-time view synthesis. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5855–5864, 2021. 3, 5

[25] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9:1735–1780, 1997. 6

[26] Yining Hong, Yilun Du, Chunru Lin, Joshua B Tenenbaum, and Chuang Gan. 3d concept grounding on neural fields. *arXiv preprint arXiv:2207.06403*, 2022. 3, 4, 5, 6

[27] Yining Hong, Li Yi, Joshua B. Tenenbaum, Antonio Torralba, and Chuang Gan. Ptr: A benchmark for part-based conceptual, relational, and physical reasoning. In *NeurIPS*, 2021. 3

[28] Deng Huang, Peihao Chen, Runhao Zeng, Qing Du, Mingkui Tan, and Chuang Gan. Location-aware graph convolutional networks for video question answering. In *AAAI*, 2020. 3, 7

[29] Pin-Hao Huang, Han-Hung Lee, Hwann-Tzong Chen, and Tyng-Luh Liu. Text-guided graph neural networks for referring 3d instance segmentation. In *AAAI*, 2021. 3

[30] D. A. Hudson and Christopher D. Manning. Compositional attention networks for machine reasoning. *ICLR*, 2018. 3, 6

[31] Ajay Jain, Ben Mildenhall, Jonathan T. Barron, P. Abbeel, and Ben Poole. Zero-shot text-guided object generation with dream fields. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 857–866, 2022. 3

[32] Zhenyu Jiang, Yifeng Zhu, Maxwell Svetlik, Kuan Fang, and Yuke Zhu. Synergies between affordance and geometry: 6-dof grasp detection via implicit representations. *ArXiv*, abs/2104.01542, 2021. 3

[33] J. Johnson, Bharath Hariharan, L. V. D. Maaten, Li Fei-Fei, C. L. Zitnick, and Ross B. Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1988–1997, 2017. 2, 3, 4

[34] Stefan Kollmannsberger, Davide D'Angella, Moritz Jokeit, and Leon Alexander Herrmann. Physics-informed neural networks. *Deep Learning in Computational Mechanics*, 2021. 3

[35] Natalia Konstantinova and Constantin Orasan. Interactive question answering. In *EMNLP*. IGI Global, 2013. 3

[36] Barbara Landau and Ray Jackendoff. "what" and "where" in spatial language and spatial cognition. *Behavioral and Brain Sciences*, 16:217–238, 1993. 6

[37] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. *ICLR*, 2022. 2, 4, 5

[38] Dongxu Li, Junnan Li, Hongdong Li, Juan Carlos Niebles, and Steven C. H. Hoi. Align and prompt: Video-and-language pre-training with entity prompts. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4943–4953, 2022. 3, 7

[39] Andrew Luo, Yilun Du, Michael J Tarr, Joshua B Tenenbaum, Antonio Torralba, and Chuang Gan. Learning neural acoustic fields. *arXiv preprint arXiv:2204.00628*, 2022. 3

[40] Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B. Tenenbaum, and Jiajun Wu. The neuro-symbolic concept learner: Interpreting scenes words and sentences from natural supervision. *ArXiv*, abs/1904.12584, 2019. 3, 6

[41] Nelson L. Max. Optical models for direct volume rendering. *IEEE Trans. Vis. Comput. Graph.*, 1:99–108, 1995. 5

[42] Lars M. Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4455–4465, 2019. 3

[43] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *Proc. ECCV*, 2020. 3, 4

[44] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Occupancy flow: 4d reconstruction by learning particle dynamics. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5379–5389, 2019. 3

[45] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *Proc. CVPR*, 2020. 3

[46] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 5

[47] Santhosh K. Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alexander Clegg, John Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X. Chang, Manolis Savva, Yili Zhao, and Dhruv Batra. Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai. *ArXiv*, abs/2109.08238, 2021. 2, 3

[48] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. *ArXiv*, abs/2102.12092, 2021. 5

[49] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proc. ICCV*, pages 2304–2314, 2019. 3

[50] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A Platform for Embodied AI Research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 3

[51] Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In *CVPR*, 2020. 3

[52] Anthony Simeonov, Yilun Du, Andrea Tagliasacchi, Joshua B Tenenbaum, Alberto Rodriguez, Pulkit Agrawal, and Vincent Sitzmann. Neural descriptor fields: Se (3)-equivariant object representations for manipulation. *arXiv preprint arXiv:2112.05124*, 2021. 3

[53] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. In *Proc. NeurIPS 2019*, 2019. 3

[54] Chan Hee Song, Jihyung Kil, Tai-Yu Pan, Brian M Sadler, Wei-Lun Chao, and Yu Su. One step at a time: Long-horizon vision-and-language navigation with milestones. *arXiv preprint arXiv:2202.07028*, 2022. 3

[55] Elizabeth S Spelke, Karen Breinlinger, Kristen Jacobson, and Ann Phillips. Gestalt Relations and Object Perception: A Developmental Study. *Perception*, 22(12):1483–1501, 1993. 2

[56] Alessandro Suglia, Qiaozi Gao, Jesse Thomason, Govind Thattai, and Gaurav Sukhatme. Embodied bert: A transformer model for embodied, language-guided visual task completion. *arXiv preprint arXiv:2108.04927*, 2021. 3

[57] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Direct voxel grid optimization: Super-fast convergence for radiance fields

reconstruction. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5449–5459, 2022. 2, 3, 5

[58] Andrew Szot, Alex Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Chaplot, Oleksandr Maksymets, Aaron Gokaslan, Vladimir Vondrus, Sameer Dharur, Franziska Meier, Wojciech Galuba, Angel Chang, Zsolt Kira, Vladlen Koltun, Jitendra Malik, Manolis Savva, and Dhruv Batra. Habitat 2.0: Training home assistants to rearrange their habitat. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 3

[59] Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. Order-embeddings of images and language. *CoRR*, abs/1511.06361, 2016. 3

[60] Can Wang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. Clip-nerf: Text-and-image driven manipulation of neural radiance fields. *ArXiv*, abs/2112.05139, 2021. 3

[61] Karmesh Yadav, Ram Ramrakhya, Santhosh Kumar Ramakrishnan, Theo Gervet, John Turner, Aaron Gokaslan, Noah Maestre, Angel Xuan Chang, Dhruv Batra, Manolis Savva, et al. Habitat-matterport 3d semantics dataset. *arXiv preprint arXiv:2210.05633*, 2022. 2, 3, 4

[62] Xu Yan, Zhihao Yuan, Yuhao Du, Yinghong Liao, Yao Guo, Zhen Li, and Shuguang Cui. Clevr3d: Compositional language and elementary visual reasoning for question answering in 3d real-world scenes. *arXiv preprint arXiv:2112.11691*, 2021. 2, 3

[63] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. *Proc. NeurIPS*, 2020. 3

[64] Shuquan Ye, Dongdong Chen, Songfang Han, and Jing Liao. 3d question answering. *ArXiv*, abs/2112.08359, 2021. 2, 3

[65] Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Joshua B. Tenenbaum. Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. In *NeurIPS*, 2018. 3, 6

[66] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. Plenoctrees for real-time rendering of neural radiance fields. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5732–5741, 2021. 3, 5

[67] Licheng Yu, Xinlei Chen, Georgia Gkioxari, Mohit Bansal, Tamara L Berg, and Dhruv Batra. Multi-target embodied question answering. In *CVPR*, pages 6309–6318, 2019. 3

[68] Kaizhi Zheng, Xiaotong Chen, Odest Chadwicke Jenkins, and Xin Eric Wang. Vlmbench: A compositional benchmark for vision-and-language manipulation. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2022. 3

[69] Yuke Zhu, O. Groth, Michael S. Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4995–5004, 2016. 2, 3