3D-SCENE-ENTITIES: USING PHRASE-TO-3D-OBJECT CORRESPONDENCES FOR RICHER VISIO-LINGUISTIC MODELS IN 3D SCENES

Anonymous authors
Paper under double-blind review

ABSTRACT

Recently, there has been significant progress in connecting natural language to real-world 3D scenes. Namely, for the problems of reference disambiguation and discriminative reference production for objects in 3D scenes, various deep-learning-based approaches have been explored by tapping into novel datasets such as ScanRefer (Chen et al., 2019) and ReferIt3D (Achlioptas et al., 2020). In this paper, we curate a large-scale and complementary dataset extending both the aforementioned ones by associating *all* objects mentioned in a referential sentence to their underlying instances in a 3D scene. Specifically, our **3D Sc**ene **Ent**ities (***3D-Scent***) dataset provides an explicit correspondence between 369,039 objects, spanning 705 scenes, over 84,015 natural referential sentences. Crucially, we show that by incorporating simple and intuitive losses that enable learning from this new dataset, we can significantly improve the performance of several recently introduced *neural listening* architectures, including *improving the SoTA* by **5.0%** in *both* the ScanRefer and Nr3D benchmarks. Moreover, we experiment with competitive baseline methods for the task of *language generation* and show that, as with neural-listeners, 3D neural-speakers can also noticeably benefit by training with 3D-Scent. Last but not least, our carefully conducted experimental studies strongly support the conclusion that, by learning on 3D-Scent, commonly used visio-linguistic 3D architectures can become more semantically *robust* in their generalization without needing to provide these newly collected annotations at test time.

1 INTRODUCTION

"The limits of my language mean the limits of my world."

— Ludwig Wittgenstein.

As the amount of available data from both the linguistic and 3D domains has increased drastically in recent years, so too has an interest in sophisticated techniques to combine, understand, and exploit this data to solve outstanding problems involving both domains. In particular, there has been flourishing interest in and work towards connecting natural *language* to *object-centric 3D scene understanding*, a task crucial to solving fundamental problems concerning objects in real-world 3D scenes. Specifically, the advent of large-scale multi-modal datasets (ScanRefer and Nr3D) catalyzed a series of learning-based solutions to problems, which range from language-assisted object localization and fine-grained object identification ((Chen et al., 2019; Achlioptas et al., 2020)), to object captioning (Chen et al., 2021), scene-based Q/A (Azuma et al., 2021), and language-based semantic segmentation (Rozenberszki et al., 2022).

At the heart of all current methods addressing these problems lies the exploitation of referential language that distinguishes one ("target") object from the remaining objects/entities co-existing in a 3D scene. This is understandable, as when humans naturally produce such discriminative referential language, they do *not* merely enumerate properties of the target project in isolation, such as its ego-centric properties of its color and geometry. Instead, they typically contextualize their description and present explicit associations between the target and other objects in the scene (e.g., *the tall chair*

$\rightarrow$ *the tall chair between the table and the fireplace*). Despite this crucial fact, current methods are constrained due to lack of available grounding data, to largely ignore (or at best, indirectly and/or weakly induce) most of these other mentioned ("anchor") entities that together with the target object contribute to the discriminative nature of each reference.

This work addresses this oversight concerning the utilization of anchor entities in two ways. First, by curating and sharing with the research community grounding annotations that go *beyond* each target object, and explicitly detail the correspondences between *all* 3D objects and any of their mentions, for both Nr3D and ScanRefer. Second, by posing and verifying the hypothesis that 3D visio-linguistic architectures *can and should* model such pairwise or higher-order object-to-object relations in order to become more robust learners. We demonstrate this intuitive hypothesis by a variety of experimental results concerning *two* cornerstone tasks for language-based 3D scene understanding: experiments addressing object-centric language comprehension (a.k.a. 'neural listening'), and experiments concerning object-centric language production (a.k.a. 'neural speaking'). Specifically, we demonstrate that the incorporation of our loss functions, which disentangle and localize the various objects mentioned in a given referential utterance during training, is: i) *effective*, as it results in significantly improved accuracy for both tasks in well-established benchmarks; ii) *generic*, as it has a positive performance effect across all the (many) evaluated architectures, and iii) its learning effect is intuitive and *natural* – we show that the primary cause of the quantitative gains we attain is learning more and/or better object-to-object relations expressed in the referential language.

To summarize, our main contributions are the following:

- We introduce a large-scale dataset extending both Nr3D and ScanRefer by grounding all objects mentioned in their referential utterances to their underlying 3D scenes. Our ***3D-Scent*** dataset (3D Scenes Entities) includes an additional 369,039 language-to-object correspondences, more than **four times** the number from the original works.
- We demonstrate that, by incorporating appropriate training losses exploiting the new annotations, we can significantly improve the performance of several 3D neural listening architectures, including **improving** the SoTA in Nr3D and ScanRefer by $\sim$ **5.0**% (55.1% to 60.1% and 54.8% to 60.8%, respectively). Crucially, we note that to keep things fair and comparable with existing works, we do *not* train our networks with more referential utterances (or scenes), or use 3D-Scent's annotations during inference. Instead, we rely solely on leveraging all objects mentioned in referential utterances during training.
- We also demonstrate the utility of 3D-Scent by using it to train and improve the generalization error of several neural speaking architectures in challenging benchmarks, as measured with standard captioning metrics (e.g., BLEU, METEOR, ROUGE and CIDEr). For instance, we **improve** the SoTA for neural speaking with Nr3D, per CIDEr, by **+13.2**.
- Last but not least, we present quantitative and qualitative results indicating that by training with 3D-Scent, different neural speaking (or listening) architectures not only attain improved performance, but do so in a *natural and interpretable* manner, i.e., by better learning high-order (primarily, spatial) object-to-object relations.

2 Related Work and Background

3D Datasets. Large and labeled datasets are crucial for deep learning tasks. Compared to 2D images, 3D data is inherently more difficult to collect and annotate, resulting in relatively immature progress in 3D tasks while 2D counterparts have already been fully explored. To mitigate the difficulty, synthetic datasets are first proposed in object-level (Chang et al., 2015) and scene-level (Handa et al., 2016; Fu et al., 2021) . However, the domain gap between real and synthetic data is usually ineligible. To further facilitate the applicability, datasets collected and annotated in real worlds are proposed for objects (Reizenstein et al., 2021) and scenes (Silberman et al., 2012; Xiao et al., 2013; Hua et al., 2016; Dai et al., 2017). Based on these efforts, supplemental annotations are collected to further enrich the semantic exploitation of the dataset. Caption and referring expressions (Thomason et al., 2021; Chen et al., 2018; 2019; Achlioptas et al., 2020) are annotated based on ShapeNet and ScanNet. In these descriptions of objects in scenes, we observe that "anchor" objects are commonly used serving as referential entities. These anchor objects, however, are in large being ignored by existing methods as well as datasets.

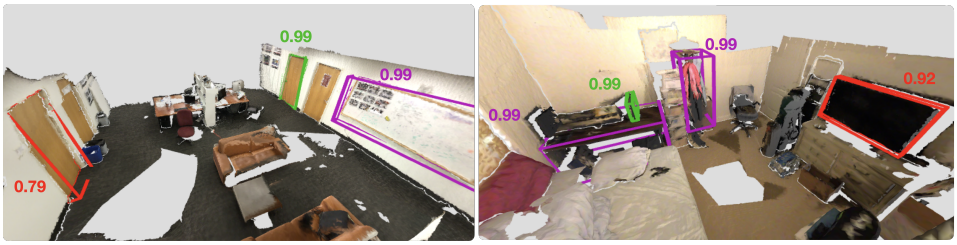


Figure 1: **Examples demonstrating the effect of training modern neural listeners with 3D-Scent.** For the two shown utterances (taken from the test split of Nr3D), the current SoTA listener model (MVT (Huang et al., 2022), wrongly (and confidently) predicts as targets the objects shown inside in each red box (respectively, for each utterance and 3D scene). When the same model is trained with 3D-Scent, it identifies the actual targets very confidently (green). In addition, the fact that it uses our dataset means that can employ a new loss that it enables it to **separately** learn to predict anchor objects (purple) **independently** of the target. Note that the confidences of each prediction have the same color and are placed nearby the corresponding boxes.

Modern visio-linguistic tasks for 3D scenes. Vision and language are two common and collaborative modalities for human beings. With the rapid development in both computer vision and natural language processing, researchers have extensively studied tasks that require joint understanding of both modalities. Thanks to the recent 3D datasets along with linguistic annotations, various tasks are unleashed, including captioning in object-level (Han et al., 2020) and scene level (Chen et al., 2021; Yuan et al., 2022), object identification in 3D scene(Chen et al., 2019; Achlioptas et al., 2020; Yuan et al., 2022), language-based semantic segmentation (Rozenberszki et al., 2022; Hou et al., 2021), and 3D question answering(Gordon et al., 2018; Kolve et al., 2017; Wijmans et al., 2019; Yu et al., 2019; Azuma et al., 2021). In this work, we target at 3D visual grounding (Huang et al., 2022; Yang et al., 2021; Roh et al., 2021; Yuan et al., 2021; He et al., 2021; Zhao et al., 2021) that associate objects in a scene given referential sentences.

3D Visual Grounding. Visual grounding aims to identify the target object given a natural language query. Visual grounding in 2D image has long been studied (Kazemzadeh et al., 2014; Mao et al., 2016; Plummer et al., 2015; Yu et al., 2016; Yang et al., 2020b; 2019; Yu et al., 2018). On the contrary, 3D visual grounding is still in its infancy due to the lack of data. Recently, Referit3D (Achlioptas et al., 2020) and ScanRef propose datasets for 3D visual grounding based on the ScanNet (Dai et al., 2017) dataset. With the 3D visual grounding datasets, explorations have been made to explore different designs and formulation (Cai et al., 2022; Wang et al., 2022; Yang et al., 2021; Abdelreheem et al., 2022; Feng et al., 2021; Roh et al., 2021). In this work, we adopt and improve two state-of-the art methods: MVT (Huang et al., 2022) that leverages additional information from different views, and SAT (Yang et al., 2021) that uses 2D images as auxiliary semantic input.

3 3D-Scene-Entities Dataset

In this section we first briefly describe our pipeline for curating the annotations for 3D-Scent. We then present its key characteristics in more detail. In the following sections, we describe the methods we developed to employ it (Sec. 4), then present and discuss our experimental results (Sec. 5).

3.1 Curating human annotations

Curating all correspondences between each noun phrase in a referential sentence and their underlying objects within a 3D scene is generally an error-prone task. First, it requires the annotators to be familiar with (albeit simple) linguistic and syntactic rules in the target language to parse the sentence. Second, they must be able to carefully navigate inside a complicated (and, possibly, poorly reconstructed) scene, which typically contains multiple objects of the same fine-grained object class (e.g., multiple kitchen cabinets, as in the lower-right example in Figure 2), so as to select *all and*

Figure 2: **Typical annotation examples from 3D-Scent**. We color-code the noun-phrases corresponding to the mentioned anchor entities for each of the four referential utterances portrayed. We use the same color to visualize the corresponding 3D objects inside bounding boxes surrounding them, and provide an exemplar view of each of the underlying 3D scenes. Our annotations are diverse with regards to the categories of anchor objects indicated, providing a rich context for each utterance/scene they outline.

only the correct referenced objects. In order to ensure the curation of high-quality correspondences with a low error rate and high coverage, we took several critical steps. First, we developed a custom web-based UI for 3D scene navigation, which was interactive, light-weight (i.e., fast), user-friendly, and which allowed for maintaining an active dialogue with the annotators. Second, we coordinated with a team of *professional* data labelers to ensure the collection of sufficiently accurate labels for 3D-Scent to validate our approach.

While a common approach to large-scale data collection today is to use crowd-sourcing techniques with platforms such as Amazon Mechanical Turk (AMT) (Crowston, 2012), we note that we conducted an AMT-based *pilot* study to determine whether such an approach is sufficient, given the aforementioned complexity and specificity of this task. We found that the error rate within the collected annotations was significantly higher than that in the annotations provided by the professional labelers (error rates of 16% vs. $< 5\%$, respectively). Rather than attempt to evaluate our approach using data with such a high level of erroneous labels, we ultimately decided to employ the professional annotators, which significantly improved the attained quality of 3D-Scent.

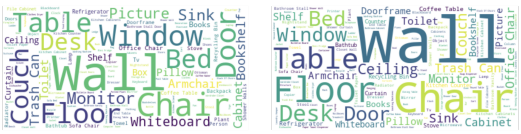
Finally, we split the curation process into two phases; the annotation phase and the verification phase. The verification phase also involved correcting the mistakes found so as to provide high-quality annotations. In Figure 2, we show examples from 3D-Scent dataset for Nr3D and ScanRefer, which demonstrate that our annotations cover different classes of anchor objects, and that our annotations provide rich contexts for these utterances.

3.2 Key Characteristics of 3D-Scent

In this section, we present some key characteristics of our 3D-Scent dataset (3D **Sc**enes **Ent**ities). A scene entity for a given utterance is a pair of words or short phrases (e.g., *tables*, *trash can*) and the 3D objects that correspond to these words, as shown in Figure 2. In Table 1, we collect annotations for 37,842 examples from the Nr3D dataset and 46,173 examples from the ScanRefer dataset. We observe that in general, ScanRefer examples provide more entities per single example compared to Nr3D (182,300 vs. 96,032, respectively) as ScanRefer utterances are usually longer than Nr3D ones (on average, there are 20.3 words per utterance in ScanRefer, vs. 11.4 in Nr3D). In Figure 3, we portray the classes most used as anchor objects for both Nr3D and ScanRefer datasets.

We also calculate how frequently an object is used as an anchor object when it is the only example from its class in the scene (e.g., single TV in a 3D scene). We find that 24.3% of all the anchor objects are unique in Nr3D and 39.1% of all anchor objects are unique in ScanRefer. As these unique anchors represent salient objects in their scenes, they are quite useful in locating the target object.

For instance, in Figure 1, the presence of the unique anchor objects (e.g., *whiteboard* and *desk*) make it easier to predict the target object.



(*a*) (*b*)

Figure 3: **Wordclouds depicting the most common anchor object classes in (a) Nr3D and (b) ScanRefer datasets.** The font size of each printed class name is proportional to its underlying frequency (better seen by zooming in).

	Nr3D	ScanRefer
Annotated Utterances	37,842	46,173
No. of Entities	96,032	182,300
Avg. No. of Entities per Utterance	2.55	3.95
Avg. No. of Objects per Entity	1.32	1.33
Avg. No. of Words per Entity	1.11	1.13

Table 1: **Basic statistics from 3D-Scent for the Nr3D and ScanRefer datasets.** ScanRefer has, by construction, more verbose utterances compared to the more parsimonious Nr3D. This distinction is clearly reflected in these statistics from 3D-Scent.

4 METHOD

In this section, we describe how the proposed 3D-Scent dataset can be used to improve the performance of current neural listeners and neural speakers. For neural listeners, we propose three simple, intuitive, yet surprisingly effective loss functions to leverage the additional data. These loss functions can serve as auxiliary add-ons to existing neural listeners, e.g., SAT (Yang et al., 2021), and MVT (Huang et al., 2022). For neural speakers, we apply a similar approach to the Show, Attend, and Tell model (Xu et al., 2015) and the $\mathcal{X}-$Trans2Cap model (Yuan et al., 2022).

4.1 NEURAL LISTENING

The goal of the neural listener is to identify the target object correctly given an input description. Following (Achlioptas et al., 2020), the input to a neural listener is a set of M 3D object proposals present in a particular scene, in which each proposal is represented as a 3D point cloud, and an input utterance that describes a target object and is represented as a set of N words. The recent neural listeners are transformer-based models (Huang et al., 2022; Yang et al., 2021; Zhao et al., 2021), each of which applies multi-modal attention layers between the features of the 3D objects and the features of the words of the input utterance. We propose three auxiliary losses to enhance existing neural listeners.

4.1.1 ANCHOR PREDICTION LOSS

We introduce an anchor prediction loss $\mathcal{L}_{\text{anc}}$ to guide the neural listener to predict the anchor (non-target) objects that are mentioned in an input utterance. First, we obtain the object feature vectors $F_L = \{f_0, f_i, \ldots, f_M\}$. For the MVT model (Huang et al., 2022), F_L is obtained from a sequence of transformer decoder layers followed by aggregation over multiple views. For the SAT model (Yang et al., 2021), F_L is obtained from a sequence of multi-modal attention layers. Then we obtain $X_{\text{anc}} = \phi(F_L)$ with an MLP consisting of two fully connected layers $\phi(.)$, where X_{anc} is a logits vector of shape $M \times 1$. We apply a binary cross entropy loss as in Equation 1, where Y_{anc} is a ground truth vector of shape $M \times 1$.

$$\mathcal{L}_{\text{anc}} = BCE(X_{\text{anc}}, Y_{\text{anc}}) \tag{1}$$

4.1.2 SAME-CLASS DISTRACTOR PREDICTION LOSS

We apply an intuitive loss for guiding the neural listener to predict the same-class distractor objects ($\mathcal{L}_{\text{dis}}$). A same-class distractor object is one from the same class as the target object co-existing in the scene but is not the one referred to in the input utterance. As with the anchor objects, we treat the same-class distractor prediction problem as a multi-label classification problem. Thus, we use an approach similar to the anchor prediction loss discussed in Section 4.1.1. We obtain the logits for predicting the same-class distractor $X_{\text{dis}} = \psi(F_L)$ of shape $M \times 1$ with an MLP $\psi(.)$ The loss is binary cross entropy as in Equation 2, where Y_{dis} is a multi-hot target vector of shape $M \times 1$. Note that a same-class distractor object may not be mentioned in the given input utterance.

$$\mathcal{L}_{\text{dis}} = BCE(X_{\text{dis}}, Y_{\text{dis}}) \tag{2}$$

4.1.3 CROSS-ATTENTION MAP LOSS

We introduce a grounding loss that encourages the network to attain high relevance values between the objects and words belonging to the same scene entity. It operates on the cross-attention maps A (before applying the softmax operation) between the features of the input scene 3D objects and the word tokens of the input utterance, where A is of shape $M \times N$. The target matrix Y_{attn} is a binary matrix of shape $M \times N$, where a cell $(y_{i,j})$ has a value of 1 if the ith object and the jth word correspond to one another. For each row R_i of shape $1 \times N$ and the corresponding row Y^i_{attn} in the target matrix, the cross-attention map loss ($\mathcal{L}_{\text{attn}}$) is measured as:

$$\mathcal{L}_{\text{attn}} = \frac{1}{M} \sum_{i=1}^{M} BCE(R_i, Y^i_{\text{attn}}) \tag{3}$$

4.1.4 TRAINING OBJECTIVE FUNCTION

The proposed losses can serve as auxiliary add-ons to the original loss term ($\mathcal{L}_{\text{org}}$) of existing neural listeners, such as the MVT and SAT models. We train these models in an end-to-end fashion, as:

$$\mathcal{L} = \mathcal{L}_{\text{org}} + \mathcal{L}_{\text{aux}}, \quad \text{where} \quad \mathcal{L}_{\text{aux}} = \alpha\mathcal{L}_{\text{anc}} + \beta\mathcal{L}_{\text{attn}} + \gamma\mathcal{L}_{\text{dis}} \tag{4}$$

4.2 NEURAL SPEAKING

In addition to the neural listening task, we demonstrate that the extra annotations in 3D-Scent can benefit neural speakers. First, We adopt a simple baseline based on the well-known "Show, Attend, and Tell" model (Xu et al., 2015), referred to here as SATCap-Scent. SATCap-Scent is an encoder-decoder network with pre-trained PointNet++ (Qi et al., 2017) layers and the 3D object self-attention layers in the MVT (Huang et al., 2022) network as the encoder, and an LSTM (Greff et al., 2017) as the decoder. The encoder part is given as inputs the ground-truth objects in a similar manner to the neural listener. The speaker model is trained with a teacher-forcing approach. We apply our proposed entity prediction loss during the decoding steps. At each decoding step, if the current word to be predicted corresponds to a scene entity, then the attention to the objects corresponding to that scene entity should be the highest among other objects that are present in the input scene.

Second, we try a similar approach on the $\mathcal{X}-$Trans2Cap model (Yuan et al., 2022), referred to as M2Cap-Scent. We introduce the following two changes to $\mathcal{X}-$Trans2Cap architecture. First, we use a pre-trained PointNet++ encoder followed by the pre-trained 3D object self-attention layers in the MVT (Huang et al., 2022) network. Second, we add a new cross-attention layer after the captioning layer. The layer applies a cross-attention operation between the features of the 3D objects of shape $M \times d$ and the features of the predicted tokens $N \times d$, where d is the latent feature dimension, to obtain new enhanced features F_c of shape $M \times d$. Finally, the logit vector is obtained with an MLP $\theta(.)$, representing a confidence value for each object as to whether it is mentioned in the target caption or not. A binary cross-entropy loss $\mathcal{L}_{\text{men}} = BCE(\theta(F_c), Y_{\text{men}})$ is employed, in which the target vector Y_{men} is a multi-hot vector (y^i_{men} is 1 if the ith object is mentioned in the target caption).

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Datasets. We use the Nr3D (Achlioptas et al., 2020) and ScanRefer (Chen et al., 2019) datasets with their original annotations as well as our additional annotations provided with the proposed 3D-Scent dataset. We use the official ScanNet (Dai et al., 2017) training and validation splits.

Metrics. In the neural listening experiments, we report the target referential accuracy. In the neural speaking experiment, we report CIDEr (Vedantam et al., 2015), BLEU-4 (Papineni et al., 2002), METEOR (Banerjee & Lavie, 2005), and ROUGE (Lin, 2004).

5.2 NEURAL LISTENING

We demonstrate the effectiveness of the proposed 3D-Scent by comparing state-of-the-art models trained with and without the additional annotations. For all experiments, we note that our dataset only

Arch.	Overall	Easy	Hard	View-dep.	View-indep.
ReferIt3DNet (Achlioptas et al., 2020)	35.6%±0.7%	43.6%±0.8%	27.9%±0.7%	32.5%±0.7%	37.1%±0.8%
InstanceRefer (Yuan et al., 2021)	38.8%±0.4%	46.0%±0.5%	31.8%±0.4%	34.5%±0.6%	41.9%±0.4%
3DRefTransformer (Abdelreheem et al., 2022)	39.0%±0.2%	46.4%±0.4%	32.0%±0.3%	34.7%±0.3%	41.2%±0.4%
3DVG-Transformer (Zhao et al., 2021)	40.8%±0.2%	48.5%±0.2%	34.8%±0.4%	34.8%±0.7%	43.7%±0.5%
FFL-3DOG (Feng et al., 2021)	41.7%	48.2%	35.0%	37.1%	44.7%
TransRefer3D (He et al., 2021)	42.1%±0.2%	48.5%±0.2%	36.0%±0.4%	36.5%±0.6%	44.9%±0.3%
LanguageRefer (Roh et al., 2021)	43.9%	51.0%	36.6%	41.7%	45.0%
SAT (Yang et al., 2021)	49.2%±0.3%	56.3%±0.5%	42.4%±0.4%	46.9%±0.3%	50.4%±0.3%
3D-SPS (Luo et al., 2022)	51.5%±0.2%	58.1%±0.3%	45.1%±0.4%	48.0%±0.2%	53.2%±0.3%
MVT (Huang et al., 2022)	55.1%±0.3%	61.3%±0.4%	49.1%±0.4%	54.3%±0.5%	55.4%±0.3%
SAT-Scent (ours)	52.5%±0.2% (+3.3%)	59.8%±0.2% (+3.6%)	45.6%±0.3% (+3.2%)	51.3%±0.5% (+4.4%)	53.2%±0.1% (+2.8%)
MVT-Scent (ours)	**60.1%±0.4%** (+5.0%)	**66.1%±0.5%** (+4.8%)	**54.1%±0.3%** (+5.0%)	**60.3%±0.7%** (+6.0%)	**60.1%±0.3%** (+4.7%)

Table 2: **Listening performance on Nr3D dataset.** The neural listeners are trained with or without our 3D-Scent dataset and our proposed losses. The numbers in green are the relative improvements over their original counterparts.

leads to modifications at training time. At inference time, our trained models and their respective baseline models use the same input data.

First, we demonstrate that neural listeners trained with 3D-Scent achieve state-of-the-art performance. As shown in Table 2 and Table 3, our MVT-Scent model, which trained with our proposed dataset and our auxiliary losses, achieves the state-of-the-art, outperforming the current SoTA models. MVT-Scent outperforms the original MVT (Huang et al., 2022) on both the Nr3D (+5.0%) and ScanRefer (+5.0%) datasets, while SAT-Scent model similarly outperforms the original SAT (Yang et al., 2021) model on both the Nr3D (+3.3%) and ScanRefer (+2.4%) datasets. The results illustrate that the proposed loss terms are generic enough to be applied as plug-ins to other models.

Next, we observe that neural listeners trained with 3D-Scent are semantically more robust. We analyze the predictions generated by our best-performing MVT-Scent model. As shown in Figure 4, first, we find that the MVT-Scent model is less likely to mistakenly predict one of the same-class distractor objects as the target object for a given input example than the original MVT (Huang et al., 2022) model by 3.5% (20.3% vs. 23.8%). Second, we also find that the MVT-Scent model is less likely to mistakenly predict an object present in the input scene but not mentioned in the input utterance as the target object than the MVT model by 1.5% (14.7% vs. 16.2%).

Furthermore, we observe considerable improvements in each context for Nr3D, particularly in the view-dependent and hard contexts (6.0% and 5.0% as in Table 2, respectively). In addition, we report the F_1 score (Powers, 2020), which measures the overall accuracy of a test taking into account its precision and recall, of the anchor object classification in the MVT-Scent model. The F_1 score of 0.64 (out of a possible maximum of 1) suggests that the full potential value of our proposed dataset 3D-Scent may still be attained with the development of more sophisticated losses, a promising area for future work.

Arch.	Acc.
ScanRefer (Chen et al., 2019)	44.5%
ReferIt3DNet (Achlioptas et al., 2020)	46.9%±0.2%
SAT (Yang et al., 2021)	53.8%±0.1%
MVT (Huang et al., 2022)	54.8%±0.1%
SAT-Scent (ours)	56.2%±0.2%
MVT-Scent (ours)	**60.8%±0.2%**

Table 3: **Listening performance on the ScanRefer dataset.** The neural listeners are trained using the ground truth boxes as input with or without using the additional annotations from the 3D-Scent dataset and our proposed losses.

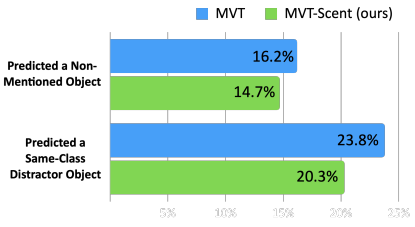


Figure 4: **Comparison of incorrect predictions on the Nr3D dataset.** The neural listeners trained with 3D-Scent are semantically more robust.

Finally, In Figure 5, we present examples of how knowing the anchor objects allows the model to identify the target object correctly. Comparing the proposed model MVT-Scent and the current state-of-the-art method MVT, we demonstrate that guiding our network to understand the anchor entities mentioned in the input utterances allows the listener to better identify the target object. In the third column, we illustrate the predicted target object and the predicted anchor objects by MVT-Scent in green and purple bounding boxes, respectively.

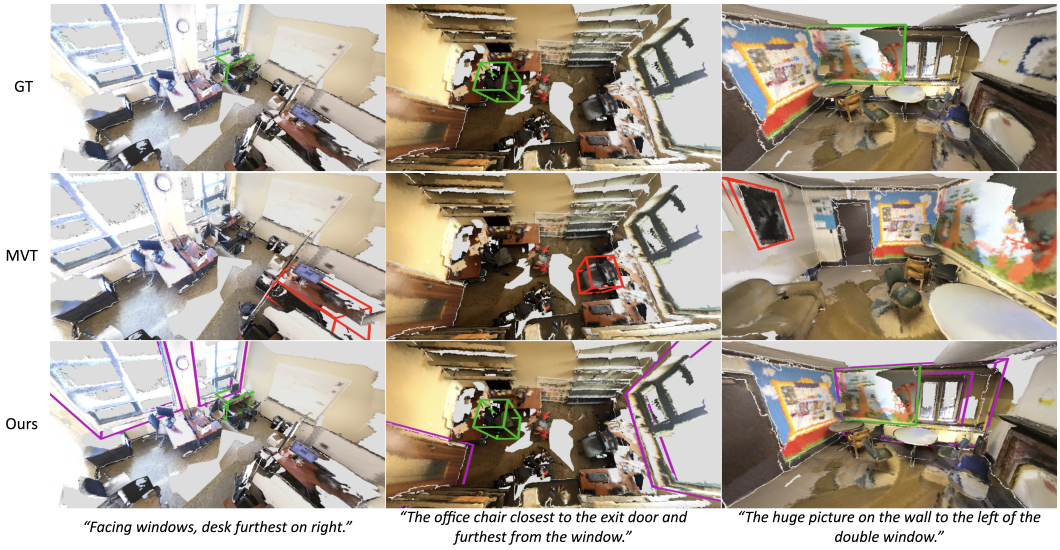


Figure 5: **Qualitative results for our proposed model (MVT-Scent) compared to the MVT model.** The rows from top to bottom show the ground-truth (green box), the target object predicted by MVT (red box), the predicted target object predicted by MVT-Scent (green box) along with the predicted anchor objects (purple boxes), and the input utterance. The above examples show that the model can accurately predict the target object by correctly understanding the underlying anchor objects mentioned in the input utterance.

5.3 Neural Speaking

With the proposed 3D-Scent dataset, the modified speaker models, SATCap-Scent and M2Cap-Scent, improve significantly against their corresponding baseline, as shown in Table 4. The encoder networks in SATCap and M2Cap models are the pretrained encoder weights of an original MVT neural listener trained without 3D-Scent, while the encoder networks in SATCap-Scent and M2Cap-Scent are the pretrained weights of an MVT-Scent listener. We observe that 3D-Scent helps our speaker models to provide better captions for Nr3D and ScanRefer across all metrics (BLEU, CIDEr, METEOR, and ROUGE). The M2Cap-Scent model improves the SoTA for neural speaking with Nr3D, per CIDEr, by **+13.2**. In all experiments, we use the ground truth instances as input. Also, we do not provide the extra 2D modality during the inference phase, and do not use the extra CIDEr loss in the final objective function found in (Yang et al., 2020a).

5.4 Ablation Study

We conduct an ablation study by applying different combinations of our proposed auxiliary losses to see their effect on the neural listener performance. We try each possible combination of our proposed losses ($\mathcal{L}_{\text{anc}}$, $\mathcal{L}_{\text{attn}}$, and $\mathcal{L}_{\text{dis}}$) with the MVT (Huang et al., 2022) architecture and report their performance on the Nr3D dataset, as shown in table 5. When applying the cross-attention map loss alone, we obtain an overall boost of 1.5% over the baseline MVT model (using none of our proposed losses). As mentioned earlier, the boost is considered small for the following reason: forcing the cross-attention maps to be closer to the sparse ground truth matrix prevents the attention layer from attending to other important words (like prepositions, spatial relation words, and object attributes). We also observe that, in general, using the cross-attention map loss in any experiment hurts the performance more than when not using it. We observe that incorporating the same-class distractor

Arch.	Nr3D				ScanRefer			
	C	B-4	M	R	C	B-4	M	R
Scan2Cap Chen et al. (2021)	61.89	32.02	28.88	64.17	64.44	36.89	28.42	60.42
$\mathcal{X}-$Trans2Cap Yuan et al. (2022)	80.02	37.90	30.48	67.64	87.09	44.12	30.67	64.37
SATCap (ours)	76.57	29.12	24.97	55.62	80.98	37.47	26.91	56.98
SATCap-Scent (ours)	84.37	30.73	25.90	56.57	84.81	38.85	27.18	57.62
M2Cap (ours)	86.15	37.03	30.63	67.00	85.75	44.02	30.74	64.80
M2Cap-Scent (ours)	**93.25**	**39.33**	**31.55**	**68.33**	**87.20**	**44.81**	**30.93**	**65.24**

Table 4: **Speaking performance on Nr3D and ScanRefer datasets.** The results of incorporating 3D-Scent dataset in our proposed approaches for the speaking (captioning) task. A speaking model trained with our rich annotations performs better than one trained without in both Nr3D and ScanRefer datasets.

prediction loss helps in improving the referential performance. We obtain an improvement in the performance of 1.8% upon applying the same-class distractor loss alone. This result is unsurprising, as we find that the same-class distractors are mentioned in 17.2% of the utterances in the Nr3D and 12.4% in the ScanRefer datasets. Applying the anchor prediction loss gives the best boost in every experiment where it is applied compared to the other losses. We observe that incorporating the anchor prediction loss is useful for all the Nr3D contexts, especially the hard contexts. The aforementioned result demonstrates how useful the knowledge of the anchor objects mentioned in the input utterance is. The best performing model applies both the anchor and same-class distractor prediction losses together, and the performance is better than using all three losses combined by 0.5%.

$\mathcal{L}_{attn}$	$\mathcal{L}_{anc}$	$\mathcal{L}_{dis}$	Overall	Easy	Hard	View-dep.	View-indep.
			55.1%±0.3%	61.3%±0.4%	49.1%±0.4%	54.3%±0.5%	55.4%±0.3%
✓			56.6%±0.2%	63.0%±0.3%	50.5%±0.3%	55.4%±0.4%	57.2%±0.2%
		✓	56.9%±0.3%	63.5%±0.3%	50.6%±0.3%	55.3%±0.4%	57.8%±0.4%
✓		✓	56.9%±0.3%	63.6%±0.4%	50.6%±0.4%	56.0%±0.1%	57.4%±0.4%
✓	✓		58.0%±0.2%	64.2%±0.3%	52.1%±0.2%	57.1%±0.6%	58.5%±0.1%
	✓		58.9%±0.3%	65.1%±0.4%	52.9%±0.2%	57.4%±0.2%	59.6%±0.4%
✓	✓	✓	59.6%±0.3%	65.5%±0.5%	53.9%±0.3%	57.8%±0.5%	**60.4%±0.2%**
	✓	✓	**60.1%±0.4%**	**66.1%±0.5%**	**54.1%±0.3%**	**60.3%±0.7%**	60.1%±0.3%

Table 5: **Ablation study on neural listeners.** We ablate different combinations of our proposed auxiliary losses on the MVT neural listener, trained on Nr3D using our proposed 3D-Scent dataset.

6 Conclusion

Humans describe or locate objects in 3-dimensional scenes by understanding and utilizing their relationships to other, co-existing objects. This work takes substantial initial steps to bring such object-to-object interactions, *grounded in language*, to the frontline of relevant learning-based methods. First, we curate and share a set of rich correspondences covering all referential entities mentioned in Nr3D and ScanRefer to their underlying environments. Second, we employ these (meta) annotations to effectively train neural networks that better understand 3D objects and their language-based grounding. Interestingly, we find that using this new grounding data and integrating our proposed *simple* and *intuitive* losses enables *state-of-the-art* results in neural listening *and* speaking tasks.

By better incorporating human-like comprehension of contextual information to describe and interact with scenes, these insights will open new opportunities to advance related multimodal tasks. These could include improving the ability of robots to understand and interact with objects in real environments and perform complex language-driven tasks; describing and manipulating large-scale outdoor environments (e.g., to caption real environments or design virtual worlds); and related tasks for which the relationship between scene entities is crucial to attain human-like performance.

REFERENCES

Ahmed Abdelreheem, Ujjwal Upadhyay, Ivan Skorokhodov, Rawan Al Yahya, Jun Chen, and Mohamed Elhoseiny. 3DRefTransformer: Fine-grained object identification in real-world scenes using natural language. *WACV*, 2022.

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 *European Conference on Computer Vision (ECCV)*, 2020.

Daich Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoki Kawanabe. ScanQA: 3d question answering for spatial scene understanding. *ArXiv*, abs/2112.10482, 2021.

Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *IEEvaluation@ACL*, 2005.

Daigang Cai, Lichen Zhao, Jing Zhang, Lu Sheng, and Dong Xu. 3djcg: A unified framework for joint dense captioning and visual grounding on 3d point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16464–16473, 2022.

Angel X. Chang, Thomas A. Funkhouser, Leonidas J. Guibas, Pat Hanrahan, Qi-Xing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. ShapeNet: An information-rich 3D model repository. *Computing Research Repository (CoRR)*, abs/1512.03012, 2015.

Kevin Chen, Christopher B Choy, Manolis Savva, Angel X Chang, Thomas Funkhouser, and Silvio Savarese. Text2shape: Generating shapes from natural language by learning joint embeddings. *Computing Research Repository (CoRR)*, abs/1803.08495, 2018.

Z. Dave Chen, Angel X. Chang, and Matthias Nießner. ScanRefer: 3D object localization in RGB-D scans using natural language. *Computing Research Repository (CoRR)*, abs/1912.08830, 2019.

Zhenyu Chen, Ali Gholami, Matthias Nießner, and Angel X Chang. Scan2Cap: Context-aware dense captioning in rgb-d scans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3193–3203, 2021.

Kevin Crowston. Amazon mechanical turk: A research tool for organizations and information systems scholars. In Anol Bhattacherjee and Brian Fitzgerald (eds.), *Shaping the Future of ICT Research. Methods and Approaches*, pp. 210–221, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. ISBN 978-3-642-35142-6.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Nießner. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

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)*, pp. 3702–3711, 2021.

Huan Fu, Bowen Cai, Lin Gao, Ling-Xiao Zhang, Jiaming Wang, Cao Li, Qixun Zeng, Chengyue Sun, Rongfei Jia, Binqiang Zhao, et al. 3d-front: 3d furnished rooms with layouts and semantics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10933–10942, 2021.

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*, pp. 4089–4098, 2018.

Klaus Greff, Rupesh Kumar Srivastava, Jan Koutník, Bas R. Steunebrink, and Jürgen Schmidhuber. Lstm: A search space odyssey. *IEEE Transactions on Neural Networks and Learning Systems*, 28: 2222–2232, 2017.

Zhizhong Han, Chao Chen, Yu-Shen Liu, and Matthias Zwicker. Shapecaptioner: Generative caption network for 3d shapes by learning a mapping from parts detected in multiple views to sentences. In *ACM International Conference on Multimedia (MM)*, 2020.

Ankur Handa, Viorica Pătrăucean, Simon Stent, and Roberto Cipolla. Scenenet: An annotated model generator for indoor scene understanding. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5737–5743. IEEE, 2016.

Dailan He, Yusheng Zhao, Junyu Luo, Tianrui Hui, Shaofei Huang, Aixi Zhang, and Si Liu. TransRefer3D: Entity-and-relation aware transformer for fine-grained 3D visual grounding. *Computing Research Repository (CoRR)*, abs/2108.02388, 2021.

Ji Hou, Benjamin Graham, Matthias Nießner, and Saining Xie. Exploring data-efficient 3d scene understanding with contrastive scene contexts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15587–15597, 2021.

Binh-Son Hua, Quang-Hieu Pham, Duc Thanh Nguyen, Minh-Khoi Tran, Lap-Fai Yu, and Sai-Kit Yeung. Scenenn: A scene meshes dataset with annotations. In *2016 fourth international conference on 3D vision (3DV)*, pp. 92–101. Ieee, 2016.

Shijia Huang, Yilun Chen, Jiaya Jia, and Liwei Wang. Multi-view transformer for 3d visual grounding. In *CVPR*, 2022.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 787–798, 2014.

Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Daniel Gordon, Yuke Zhu, Abhinav Gupta, and Ali Farhadi. Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474*, 2017.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *ACL 2004*, 2004.

Junyu Luo, Jiahui Fu, Xianghao Kong, Chen Gao, Haibing Ren, Hao Shen, Huaxia Xia, and Si Liu. 3d-sps: Single-stage 3d visual grounding via referred point progressive selection. *arXiv preprint arXiv:2204.06272*, 2022.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 11–20, 2016.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

David M. W. Powers. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *CoRR*, abs/2010.16061, 2020. URL `https://arxiv.org/abs/2010.16061`.

Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. PointNet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10901–10911, 2021.

Junha Roh, Karthik Desingh, Ali Farhadi, and Dieter Fox. LanguageRefer: Spatial-language model for 3D visual grounding. *Computing Research Repository (CoRR)*, abs/2107.03438, 2021.

Dávid Rozenberszki, Or Litany, and Angela Dai. Language-grounded indoor 3d semantic segmentation in the wild. *ArXiv*, abs/2204.07761, 2022.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European conference on computer vision*, pp. 746–760. Springer, 2012.

Jesse Thomason, Mohit Shridhar, Yonatan Bisk, Chris Paxton, and Luke Zettlemoyer. Language grounding with 3d objects. *Computing Research Repository (CoRR)*, abs/2107.12514, 2021.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4566–4575, 2015.

Heng Wang, Chaoyi Zhang, Jianhui Yu, and Weidong (Tom) Cai. Spatiality-guided transformer for 3d dense captioning on point clouds. In *IJCAI*, 2022.

Erik Wijmans, Samyak Datta, Oleksandr Maksymets, Abhishek Das, Georgia Gkioxari, Stefan Lee, Irfan Essa, Devi Parikh, and Dhruv Batra. Embodied question answering in photorealistic environments with point cloud perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6659–6668, 2019.

Jianxiong Xiao, Andrew Owens, and Antonio Torralba. Sun3d: A database of big spaces reconstructed using sfm and object labels. In *Proceedings of the IEEE international conference on computer vision*, pp. 1625–1632, 2013.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning (ICML)*, 2015.

Muli Yang, Cheng Deng, Junchi Yan, Xianglong Liu, and Dacheng Tao. Learning unseen concepts via hierarchical decomposition and composition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020a.

Zhengyuan Yang, Boqing Gong, Liwei Wang, Wenbing Huang, Dong Yu, and Jiebo Luo. A fast and accurate one-stage approach to visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4683–4693, 2019.

Zhengyuan Yang, Tianlang Chen, Liwei Wang, and Jiebo Luo. Improving one-stage visual grounding by recursive sub-query construction. In *European Conference on Computer Vision*, pp. 387–404. Springer, 2020b.

Zhengyuan Yang, Songyang Zhang, Liwei Wang, and Jiebo Luo. SAT: 2d semantics assisted training for 3D visual grounding. *Computing Research Repository (CoRR)*, abs/2105.11450, 2021.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *European Conference on Computer Vision*, pp. 69–85. Springer, 2016.

Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1307–1315, 2018.

Licheng Yu, Xinlei Chen, Georgia Gkioxari, Mohit Bansal, Tamara L Berg, and Dhruv Batra. Multi-target embodied question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6309–6318, 2019.

Zhihao Yuan, Xu Yan, Yinghong Liao, Ruimao Zhang, Zhen Li, and Shuguang Cui. InstanceRefer: Cooperative holistic understanding for visual grounding on point clouds through instance multi-level contextual referring. *International Conference on Computer Vision (ICCV)*, 2021.

Zhihao Yuan, Xu Yan, Yinghong Liao, Yao Guo, Guanbin Li, Shuguang Cui, and Zhen Li. X-trans2cap: Cross-modal knowledge transfer using transformer for 3d dense captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8563–8573, June 2022.

Lichen Zhao, Daigang Cai, Lu Sheng, and Dong Xu. 3DVG-Transformer: Relation modeling for visual grounding on point clouds. In *ICCV*, 2021.

A Appendix

A.1 Dataset Collection

This section discusses in detail the two phases of our 3D-Scent curation. Figure 6 shows the user interface we implemented.

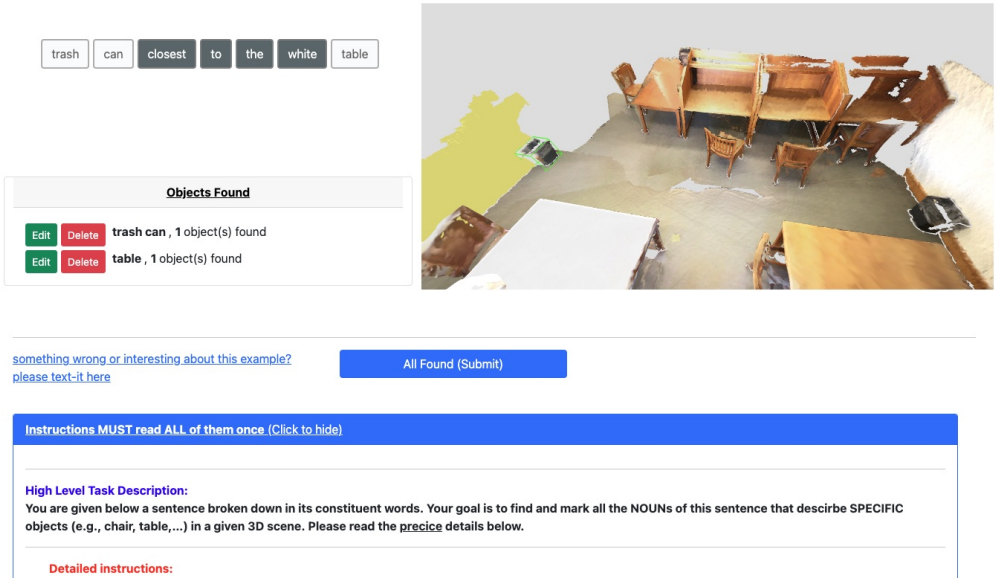


Figure 6: The user interface for 3D-Scent dataset collection.

Annotation Phase. A human labeler is given an utterance and a 3D scene. While the utterance generally describes one specific object in the 3D scene, the labeler is first asked to mark all the nouns (entities) that describe specific objects in the given 3D scene (e.g., chair, table, etc.) in the utterance. Then, for each selected entity in the given utterance, the labeler should highlight the corresponding 3D objects in the given 3D scene. The labeler can zoom, pan or rotate the 3D scene to find the corresponding 3D objects. Each labeler is provided with one random utterance at a time. We assign one labeler for each example.

Review Phase. A reviewer is given one annotated example randomly and is asked to determine whether the example was correctly annotated. If the example was annotated incorrectly, the reviewer is then requested to correct and fix the annotation. The reviewer is shown a similar user interface to the labeler. Each annotation is reviewed by one reviewer.

A.2 Implementation Details

For the listening experiments, We used the same hyper-parameters specified in MVT (Huang et al., 2022) and SAT (Yang et al., 2021). We use one NVidia V100 GPU in each of our experiments. For the listening experiments, we use $\alpha = 3.0$, $\gamma = 2.0$, and $\beta = 0$. We use the same hyper-parameters found in Yang et al. (2020a) for the neural speakers.