ZInD-Tell: Towards Translating Indoor Panoramas into Descriptions

Tonmoay Deb[1]*, Lichen Wang[2], Zachary Bessinger[2], Naji Khosravan[2], Eric Penner[2], Sing Bing Kang[2]
[1]Northwestern University [2]Zillow Group
tonmoay.deb@northwestern.edu, {lichenw,zacharybe,najik,ericpe,singbingk}@zillowgroup.com

Abstract

This paper focuses on bridging the gap between natural language descriptions, 360° panoramas, room shapes, and layouts/floorplans of indoor spaces. To enable new multimodal (image, geometry, language) research directions in indoor environment understanding, we propose a novel extension to the Zillow Indoor Dataset (ZInD) which we call ZInD-Tell[1]. We first introduce an effective technique for extracting geometric information from ZInD's raw structural data, which facilitates the generation of accurate ground truth descriptions using GPT-4. A human-in-the-loop approach is then employed to ensure the quality of these descriptions. To demonstrate the vast potential of our dataset, we introduce the ZInD-Tell benchmark, focusing on two exemplary tasks: language-based home retrieval and indoor description generation. Furthermore, we propose an end-to-end, zero-shot baseline model, ZInD-Agent, designed to process an unordered set of panorama images and generate home descriptions. ZInD-Agent outperforms naïve methods in both tasks, hence, can be considered as a complement to the naïve to show potential use of the data and impact of geometry. We believe this work initiates new trajectories in leveraging Computer Vision techniques to analyze indoor panorama images descriptively by learning the latent relation between vision, geometry, and language modalities.

1. Introduction

Description generation aims to automatically generate informative and meaningful textual descriptions or narratives based on given signals such as images, videos, or other data formats. It is an emerging research topic due to its potential in various domains and tasks. Indoor description generation specifically focus on generating descriptive information about indoor spaces (e.g., homes, apartments, or offices). The goal is to generate coherent descriptions in natural language that accurately capture the layout, features, and characteristics of the indoor environment.

One of the key values of indoor home description generation lies in its applications in the real-estate industry. This technology has significant potential in enhancing the efficiency of real estate agents and improving the user experience for property seekers. Imagine a real-estate agent, focused on describing properties, captures multiple room-wise 360° panorama pictures of a multi-story building. Manually writing detailed and accurate descriptions for the property is tedious, time-consuming, and error-prone. For example, the agents want to align the panorama images together, note all specific details while writing the description. An end-to-end model that automates this process by generating coherent descriptions from the set of panorama images would significantly reduce manual effort of the agents. Also, the potential customers may search semantically relevant property by describing that in language, e.g., 'I want two bedrooms adjacent to a large dining space', customers can find properties that closely match their specifications.

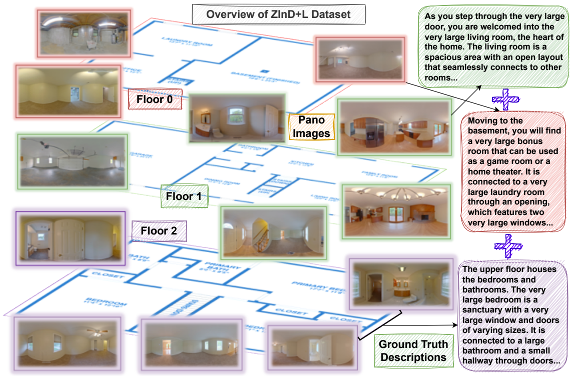
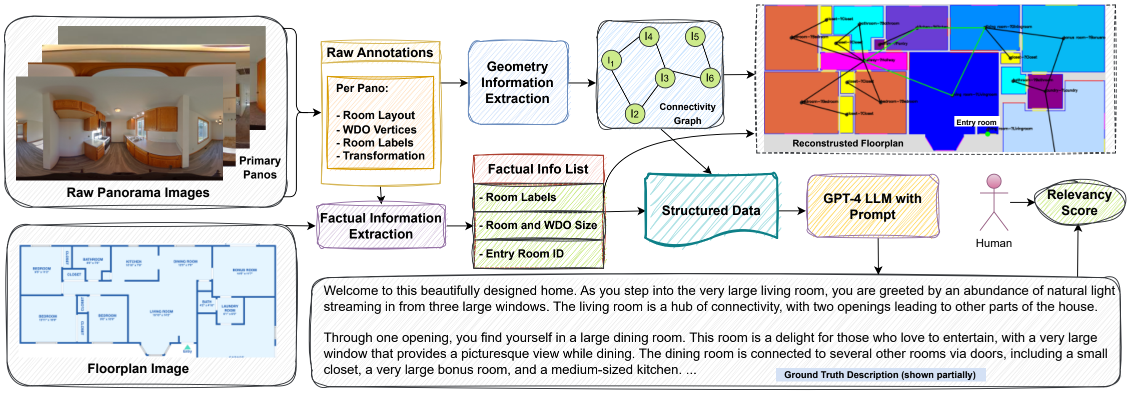

Overview of ZInD+L Dataset
Floor 0
Floor 1
Floor 2
Pano Images
Ground Truth Descriptions
As you step through the very large door, you are welcomed into the very large living room, the heart of the home. The living room is a spacious area with an open layout that seamlessly connects to other rooms...
Moving to the basement, you will find a very large bonus room that can be used as a game room or a home theater. It is connected to a very large laundry room through an opening, which features two very large windows...
The upper floor houses the bedrooms and bathrooms. The very large bedroom is a sanctuary with a very large window and doors of varying sizes. It is connected to a large bathroom and a small hallway through doors...

Figure 1. Overview of ZInD-Tell. It consists of 3150 ground-truth descriptions of 1575 homes, each having multiple unordered indoor panorama images. The descriptions contain the details about the room connectivity and coherent details of the room features across multiple floors. Each ground truth description is a combination of one or more paragraphs for each home. In this figure, we split the description floor-wise for better illustration.

There have been several prior research on generating description from indoor images [1, 5, 22, 26]. The datasets like Sentences-NYUv2 [14] facilitate image-text pairs, focused on describing an interior room contents. Some recent datasets contain indoor semantic scene graph of 3D rooms [5] to describe relations among the objects. The indoor im-

*Work done during Summer'23 internship at Zillow Group.
[1] Dataset is avaiable at https://github.com/zillow/zindtell



Welcome to this beautifully designed home. As you step into the very large living room, you are greeted by an abundance of natural light streaming in from three large windows. The living room is a hub of connectivity, with two openings leading to other parts of the house.

Through one opening, you find yourself in a large dining room. This room is a delight for those who love to entertain, with a very large window that provides a picturesque view while dining. The dining room is connected to several other rooms via doors, including a small closet, a very large bonus room, and a medium-sized kitchen. ...

Figure 2. Pipeline of generating ground truth descriptions in ZInD-Tell. The process begins with metadata extraction for each H. This stage is succeeded by factual data collection, including the entry room, room labels, and the vertices of windows, doors, and openings (W/D/O). Following this, we construct a room-room connectivity graph, then schema for each floor. These structured data serve as inputs to GPT-4, facilitating the creation of ground truth descriptions. Then, human evaluators assess these descriptions, focusing on their **relevancy score**.

age datasets mainly used for Vision-Language navigation [1, 13] contain visual landmarks instead of coherent description of a floor. To overcome this gap, we want a dataset that contains 360° panorama images of each room for every floor of a home. The description of a home should contain coherent details of each room in a floor, which highly respects both geometric and semantic constraints in that floor. Also, the description should be comprehensive to contain coherent information across all floors so that it can be considered as a holistic overview of the property. To the best of our knowledge, there is no dataset yet that meet these constraints. Hence, we propose ZInD-Tell, an extension of the Zillow Indoor Dataset (ZInD) [7] with language modality. We consider ZInD is the most appropriate for this task, given its extensive coverage of real residential homes.

We propose two strategies, schema-based and template-based, for the ground truth description generation. To this end, ZInD-Tell consists of 2 large descriptions for each residential home for 1575 homes in total. The raw datasets are initially pre-processed to meet the geometric constraints, followed by extraction of the key information based on the raw human annotation. Then, we build the structural and template-based schema to automate the description generation by leveraging GPT-4 LLM [21]. Then, we curate the descriptions by human evaluations to verify the relevancy of the ground truth descriptions to the homes. Figure 1 illustrates a brief overview of the dataset. To evaluate the effectiveness of our ZInD-Tell dataset, two tasks are proposed. The first task is language-based home retrieval. Here, given a description as a query, the trained model will retrieve the home with the closest semantic distance. Second, generation of the descriptions directly from panorama images. Here, an end-to-end trained model will take input of the unordered set of indoor panorama images per floor. The objective of the model is learn the semantic connections between the rooms in a floor and between the floors to generate a comprehensive and coherent description.

To initiate the benchmark for our ZInD-Tell dataset, we introduce a zero-shot method, ZInD-Agent. This baseline synthesizes existing large models as modules: CLIP [23] for room classification, HorizonNet [24] for room layout and window/door/opening location estimation, HoHoNet [25] for depth inference, and SaLVe [15] for predicting room-to-room connectivity graphs. These modules collectively facilitate the generation of predicted descriptions. We evaluate ZInD-Agent's performance in two tasks as mentioned earlier: 1) language-based home retrieval score at k number of homes and 2) generated description quality using standard sentence evaluation metrics. Our results indicate that ZInD-Agent outperforms naïve models in the evaluation metrics. To the best of our knowledge, this is the first dataset that includes comprehensive home descriptions in natural language. Our major technical contributions are:

- Creating ZInD-Tell, the first ever large-scale dataset that includes natural language descriptions of indoors, layouts and panoramic images, by enhancing the well-established Zillow Indoor Dataset [7].
- Offering a thorough statistical analysis and human evaluations of the dataset to assess the quality.
- Proposing ZInD-Agent, a zero-shot baseline model, for generating home descriptions from panorama images.
- Benchmarking ZInD-Tell dataset by comparing ZInD-Agent's performance with naïve baselines across two tasks: generation and retrieval.

2. Related Works

2.1. Indoor Scene Description Datasets

Several efforts have been undertaken to understand and describe indoor scenes. A pioneering dataset in this domain is the NYUv2 [19], featuring indoor RGB-D scenes with segmentation and 3D planes, including annotations of indoor furniture objects. Sentences-NYUv2 dataset [14] extended NYUv2 by creating descriptions of all 464 indoor scenes, incorporating a semantic scene graph—illustrating connections between indoor objects—and grounding these objects/relations in the descriptions. The primary aim of this dataset was to facilitate scene understanding through natural language descriptions and semantic scene graphs. Its main limitation, however, is its less image diversity and relatively small in size for training deep learning models.

Recent advancements in Multimodal AI research have increased interest in scene understanding, leading to a focus on simulated data for a richer variety of indoor scenes. For instance, the AI2THOR [13] dataset allows the generation of extensive synthetic data using a game engine. Another branch of research has extended real indoor datasets for scene understanding, such as the Spatial Commonsense Graph (SCG) [11] and ScanNet [8] datasets, featuring real-world 3D scans with scene graph annotations. Similarly, 3DSSG [28] proposed a semi-automatically generated dataset for semantic scene graph prediction. Unlike the aforementioned datasets, only the Sentences-NYUv2 dataset included sentence descriptions. Addressing this gap, the ScanRefer [5] dataset provides extensive descriptions of each RGB-D scanned indoor object within 3D bounding boxes, significantly surpassing the SentencesNYUv2 dataset in both scene graph size and number of descriptions.

Existing datasets have predominantly focused on discrete indoor rooms, not encompassing entire floors or buildings. Creating such datasets poses significant challenges, including substantial effort and legal considerations. The 3D Scene Graph [2] dataset, however, circumvents these challenges by using synthetic 3D scans of entire buildings, containing data of floor-wise and room-wise 3D indoor data, primarily 360° panorama images, along with 3D object scene graphs, but lacks explicit language descriptions. Another research direction involves creating navigational instructions in natural language to guide embodied agents. The REVERIE [22] dataset, for instance, focuses on scene-focused language navigation, such as instructing an agent to pick up a glass from a table, using First-Person View (FOV) navigation paths built on the Matterport3D simulator [4]. Similarly, datasets like R2R [1] and CVDN [26] offer indoor navigation instructions. A recent extension of the R2R dataset [29] augments descriptions in multiple languages using visual landmarks. These datasets primarily assist in object localization using language instructions.

In contrast, our proposed dataset focuses on describing entire homes coherently at both room and floor levels in natural language. The research objective is to learn the home description generation from unordered sets of panorama images, with descriptions aiding in querying and retrieving relevant homes. Since ZInD is derived from real-world unfurnished homes, its descriptions are grounded in real-world indoor contexts, adding a novel dimension to this area.

3. Problem Definition

For a given home $H \in \mathcal{H}$ from the set of all homes $\mathcal{H}$, there exists floors $f_i \in H$, where $i \geq 1$ is the floor index. Each floor f_i contains a set of indoor panorama images, denoted as $I_{f_i} = \{I_{ij}\}_{j=1}^{N}$, where i^{th} floorplan has N total images. Each image $I_{ij} \in \mathbb{R}^{3\times X\times Y}$ represents an RGB format with height X and width Y. Associated with each H are M distinct ground-truth descriptions $\{D^*_{H_j}\}_{j=1}^{M}$, satisfying $D^*_{H_j} \neq D^*_{H_k}$ for all distinct $j, k \in \{1, \ldots, M\}$. The goal of an end-to-end model is to learn an optimal set of weights θ^*, such that the generated description $D_H = f(\theta^*, H)$ approximates D^*_H. However, to the best of our knowledge, existing works do not provide D^*_H for this specific problem. Thus, subsequent sections will detail the curation of D^*_H following our proposed ZInD-Agent for predicting D_H.

4. ZInD-Tell: ZInD + Description

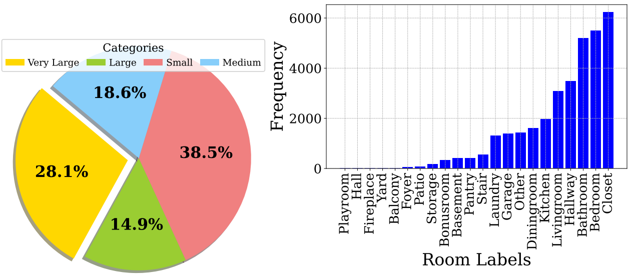
In this section, we detail the creation of ZInD-Tell, explaining the methodology employed to derive D^*_H, for $\forall H \in \mathcal{H}$ from the dataset. First, we discuss the structure and metadata of ZInD that are utilized in for deriving D^*_H. Next, we discuss the two distinct (geometry and factual) information extraction procedures from the metadata. Finally, we discuss our approach on organizing the extracted information that leads to the generation of D^*_H, followed by human evaluation. Figure 2 depicts the high-level flow of the process.

4.1. Dataset Components and Annotations

The ZInD dataset comprises 1575 real residential homes, featuring 67448 panorama images of vacant rooms. Each room includes several 360° panoramas, categorized as *primary* and *secondary*. Annotations in the *primary* panorama encompass the *layout* and *W/D/O* (window/door/opening) details of each room. Additionally, *room labels* form part of the dataset's annotations. The *primary* panoramas are arranged to ensure *co-visibility* between adjacent rooms, implying a partial overlap of images. Annotators manually assess and incorporate the *floorplans* of each home in the dataset. We denote the aggregation of all panorama images of a home H as $I_H = \bigcup_{f_i \in H} I_{f_i}$, where each home averages 1.68 floorplans. These floorplans provide precise geometric details of each floor f_i. A unique feature of the dataset is the annotation of a single entry room in each home, identified in one panorama image I_{ij}. Moreover, manual annotations include *transformations* such as translation, rotation, and scale for each room and floor [7].

4.2. Geometry Information Extraction

To analyze the internal structure of a home, understanding the *room-to-room* connectivity is crucial. While the ZInD dataset includes annotated vertices, it lacks explicit *room-to-room* connectivity data. Consequently, we utilize exist-



(a) room size (b) room label

Figure 3. **Distribution** of room size and room labels of ZInD-Tell.

ing data to construct a connectivity graph, hypothesizing that rooms in ZInD are interconnected via a *door* or an *opening* (visual edge). For each panorama image I_{ij}, we initially extract the ground truth vertices of the *room* and its corresponding *D/O* (door/opening). These vertices are then projected onto the floorplan using the transformation $\mathbf{T_{ij}} \in SE(2)$. Subsequently, door vertices are assigned to rooms based on overlapping vertex values. Connectivity between two *doors* $(d_i, d_j)_{i \neq j}$ is determined if the vertices are parallel within a specified angle threshold θ_d and distance threshold β_d, based on the projected floor-level transformations. These thresholds θ_d and β_d are empirical and consistent across all homes. A similar approach is applied for *openings* $(o_i, o_j)_{i \neq j}$, with respective thresholds θ_o and β_o. An adjacency matrix $\mathcal{M}^H_{f_i} \in \mathbb{N}_0^{R \times R \times 2}$ is then formulated for each floor f_i of a home H, where R denotes the number of rooms. Each element $m_{i,j} \in \mathcal{M}^H_{f_i}$ is a tuple of $(int, ctype)$. The first element $(m_{i,j})_1$ represents the absence (0) or presence (>0, indicating Euclidean distance) of a connection between two rooms. The second element $(m_{i,j})_2$, either D (door) or O (opening), specifies the *type* of connection. This *room-to-room* connectivity graph is bidirectional, ensuring $m_{i,j} = m_{j,i}$. The reconstructed floorplan, showcasing door (black edges) and opening-based (green edges) connectivity, is depicted in the top right image of Figure 2. The application of $\mathcal{M}^H_{f_i}$ in constructing the graph is elaborated in subsequent sections.

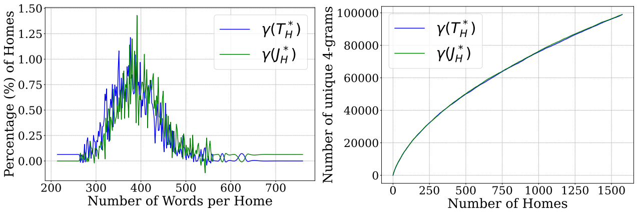
Algorithm 1 ZInD-Tell Schema Generation for Floor f_i

```
Require: Graph $\mathcal{M}^P_{f_i} \in \mathbb{N}_0^{R \times R \times 2}$, start room node $S_{f_i}$
Ensure: JSON-like schema $J_{f_i}$ for floor $f_i$
Initialize queue $Q$
Initialize visited array $V$ with False for all nodes
Initialize JSON-like schema $J_{f_i}$ as an empty structure
$Q$.enqueue($S_{f_i}$)
while there are unvisited nodes do
    $r \leftarrow Q$.dequeue()
    Initialize an empty object $O$ for node $r$
    Set $O$.id $\leftarrow r$
    Set $O$.label $\leftarrow$ label($r$)
    Set $O$.size $\leftarrow$ size($r$)
    Set $O$.wdo $\leftarrow$ WD($r$)
    Set $O$.connections $\leftarrow$ an empty list
    if not V[$r$] then
        V[$r$] $\leftarrow$ True
        for $i$ from 1 to $R$ do
            if $\mathcal{M}^P_{f_i}[r][i][1] > 0$ and not V[$i$] then
                $Q$.enqueue($i$)
                Create a connection object $C$
                Set $C$.index $\leftarrow i$
                Set $C$.type $\leftarrow \mathcal{M}^P_{f_i}[n][i][2]$
                Append $C$ to $O$.connections
    Add $O$ to $J_{f_i}$
    if $Q$ is empty and there are unvisited room then
        Find unvisited room $x$ with the highest adjacency
        $Q$.enqueue($x$)
return $J_{f_i}$
```

4.3. Factual Information Extraction

The ZInD dataset offers explicit information which we leverage for developing ZInD-Tell. Key elements utilized include room labels, layout, and *W/D/O* bounding boxes. To compute the size of *rooms* and *W/D/O*, we first calculate their areas using vertex data, expressed in the dataset's units. Then, we determine floor-level room ratios by summing the area of all **rooms** on each floor and computing the proportionate area of each. A similar approach is adopted for *W/D/O*, calculating their respective percentage ratios. Additionally, we categorize the area distribution into four size bins: small, medium, large, and very large. This binning method is also applied to the distances in the *room-to-room* connectivity analysis. Notably, our factual extraction process does not rely on explicit image features I_{ij}, as ZInD's rooms are empty and lack furnishings.

Entry Room Identification process begins with the extraction of the entry room from each home. Despite each home having exactly one entry room, denoted as I_{ij}, this information is not explicitly provided in the dataset metadata. Instead, it is indicated in the ground truth floorplan image through an upward arrow, as illustrated in Figures 2 and 5. Our analysis revealed that this arrow consistently points north and is positioned below the entry door, distinguished by a unique color. For a given home floorplan image $I^H_{f_i}$, we isolate the arrow by removing all other pixels. Subsequently, we reconstruct a 2D floorplan from the *room* and *W/D/O* vertices metadata, resulting in an image $I^H_{f'_i}$, identical to $I^H_{f_i}$ size. v_i is the set of pixel locations of door vertices in $I^H_{f'_i}$. Projecting $I^H_{f'_i}$ onto $I^H_{f_i}$, we determine the pixel location of the arrow, denoted as centroid pixel c in the projected image. We store the pixel locations of projected door ids and compute the pairwise distance (in pixels) between c and each door d_i as $D(d_i, c) = \sqrt{(x_{d_i} - x_c)^2 + (y_{d_i} - y_c)^2}$, where (x_*, y_*) represents the coordinates of d_i and c. The nearest door d_{min} is identified using $\arg\min_{d_i} D(d_i, c)$, ensuring that d_{min} is above the arrow's row pixel and almost parallel with a threshold angle θ_d. The identified door id, d_{min}, is used to locate the corresponding I_{ij}, which is then recorded as factual information $S_{f_i} = I_{ij}$. For other floors $(k \neq i)$, we set $S_{f_k} = \phi$. The top-right of Figure 2 visualizes $I^H_{f'_i}$, with a green circle indicating d_{min}, identical to $I^H_{f_i}$ shown in the bottom left. We manually verified



(a) word frequency (b) unique 4-grams

Figure 4. **Frequency distribution** of the number of unique words and total number of unique 4-grams for $\gamma(J_H^*)$ and $\gamma(T_H^*)$.

this pipeline for all homes. Out of the 16 out of 1575 ZInD homes lacking this information, we manually labeled their I_{ij}. Because, we posit that entry room information is vital for description generation, acting as a <*begin*> token. As shown in Figure 2, the generated text integrates the entry room information, illustrating its significance with phrases such as "As you step into the very large living room...".

4.4. Description Generation

In this section, we discuss the post-processing applied to the extracted data. Further, we detail the utilization of the GPT-4 LLM [21] for generating descriptions for building the final ZInD-Tell dataset. We adopted two distinct techniques.

Schema Generation: This technique entails generating a JSON-like structured data schema, denoted as J_H^*, for each home. The process involves iterating over annotation data to extract geometric and factual information (Sections 4.2 and 4.3) for each floor. Specifically, we insert the factual information of a room r, such as $label(r)$ and $size(r)$, and include a list of *W/D* with their relative sizes (in bins) for $WD(r)$. Additionally, connected rooms to r are appended based on $\mathcal{M}_{f_i}^H$. This schema generation is framed as a Breadth First Search (BFS), starting from the entry room, $S_{f_i} \in \mathcal{M}_{f_i}^H$ (assuming $S_{f_i} \neq \phi$). The process entails level-wise BFS traversal from S_{f_i}, documenting factual and connectivity information. If the BFS queue empties before visiting all rooms, indicating isolated rooms, the algorithm enqueues the room with the highest adjacency and continues the traversal until all rooms in f_i are visited. For floors where $S_{f_i} = \phi$, S_{f_i} is set to the room with the highest adjacency, following the same process. The detailed procedure is outlined in Algorithm 1. The schema for each floor J_{f_i} is then merged as $J_H^* = \bigcup_{f_i} J_{f_i}$ and used in subsequent steps.

Template-based Primitive Descriptions: While the schema J_H^* provides comprehensive information about home H, its format is not natural language, posing challenges for stochastic models such as LLMs in generating semantically consistent descriptions. For instance, interpreting room-to-room connectivity graphs from J_H^* and accurately tracking each room id, r, can be complex. To address this, we introduce a *template-based* approach to enhance the initial seed step for LLMs. This technique employs the same BFS exploration technique as Algorithm 1, exploring each floor until all rooms are visited. However, we propose semi-automatic text generation using predefined templates. For instance, a template for room factual information might read: "Room <*room_id(r)*> is labeled as <*room_label*>, with a size of <*room_size*>, constituting <*room_size_percent*>% of the total floor area." Additionally, separate templates describe the factual information of *W/D/O* for each room. We utilize three templates to delineate 1) room factual info, 2) *W/D/O* details, and 3) room-to-room connectivity. This approach results in numerous semi-automatic factual texts for each floor, providing a rich textual dataset that can be more effectively processed by stochastic LLMs for home description generation, i.e., T_H^*.

Final Descriptions Generation: To generate final descriptions, we employ GPT-4 LLM [21] with a context length of 32000 tokens. Both J_H^* and T_H^* for all homes are tokenized. The maximum token counts for J_H^* and T_H^* are 5777 and 3671, respectively. For the LLM-based generation, we use a standardized prompt[2] for each of J_H^* and T_H^*, with a maximum token length of 159. These prompts, concatenated with J_H^* and T_H^*, feed into the LLM. Given the total prompt size remains below 32000 tokens, we anticipate adequate context for the LLM to avoid hallucinations. The prompts instruct the LLM to generate descriptions with a maximum of 500 words. Hence, for each home H, two Ground Truth (GT) descriptions are produced, denoted as $D_H^* = \{\gamma(J_H^*), \gamma(T_H^*)\}$, where $\gamma(.)$ represents the output from GPT-4 LLM. The entire process is depicted in Fig. 2.

4.5. Human Evaluation of the Descriptions

As the descriptions are generated, we cross-check the quality of the generation, i.e., the correctness of the description based on the schema. Hence, we employ human evaluation to perform this task. As mentioned earlier, for each H we produce two types of home descriptions, $\{\gamma(J_H^*), \gamma(T_H^*)\}$. We designed a user interface (shown in suppl. material) that displays the floor plan panoramas on the left (with a slider to change floors) and the generated description on the right. The evaluators are given a random description from $\{\gamma(J_H^*), \gamma(T_H^*)\}$ to minimize bias. They evaluate the **relevancy** of that description of the home by inspecting the floor plan panoramas. The **relevancy score** is based on a Likert scale, ranging from 1 to 10. For each home description, we collect evaluation from at least two different subjects and then compute the average. (A third evaluator is used only if the difference between the first two scores is > 3.) After the survey, hypothetically, we remove homes with descriptions having average relevance scores < 6 from the dataset. It turned out that all the average scores we collected are > 6. Hence, we are able to use all 1575×2 descriptions in our dataset. Also, another interesting phenomenon is that the $\gamma(J_H^*)$ has on average 8.05 ± 1.41, where 7.96 ± 1.32 was

[2] To be released with the dataset

Table 1. **Unique POS tags** of the ZInD-Tell dataset

Type	#Noun	#Adj.	#Verb	#Adv.
$\gamma(J_H^*)$	793	479	380	110
$\gamma(T_H^*)$	734	435	336	120

for $\gamma(T_H^*)$, for $\forall H \in \mathcal{H}$. While the difference is small, it appears that the GPT-4 LLM learns a little better context while parsing the schema compared to parsing through a large set of templated content for ZInD-Tell.

5. Analysis of ZInD-Tell Dataset

In this section, we analyze the final generated GT dataset, ZInD-Tell, in multiple aspects. First we study the embedding space, followed by the sentence and label distributions.

Embedding Space: We encode the $\{D_H^*\}_{H\in\mathcal{H}}$, obtained using $\gamma(J_H^*)$ for each home H, with the 'all-MiniLM-L6-v2' sentence transformer[3] chosen for its efficiency and compact size. This model transforms each description into a 384-D vector, which we then project into a 3D space using t-SNE [27]. Figure 5 visualizes this 3D space, with the third axis represented by color hue. Our analysis reveals two key observations: firstly, the embeddings are closely clustered, indicating minimal spatial separation between them. The similarity metric ranges from a maximum of 0.97 to a minimum of 0.52, reflecting the limited diversity in floorplans, predominantly influenced by geometric factors. Secondly, selecting two proximal points and two distant points, we observe that visually similar floorplans correspond to nearby embeddings, while structurally distinct floorplans align with distant embeddings (Figure 5). This suggests the dataset's potential for retrieval tasks based on descriptions. Furthermore, the deployment of larger models for embedding $\gamma(J_H^*)$ might enhance retrieval performance. We also illustrate embedding space of $\gamma(T_H^*)$ in suppl. material.

Distributions: In Figure 3 and 4, key dataset distributions are depicted. The first is the room size distribution, with a notable predominance of 'small' rooms (38.5%), as highlighted in Figure 3a. The sub-figure 3b also illustrates the room label frequencies, where 'closet' emerges as the most common room type. This prevalence is attributed to properties often having multiple closets, and bedrooms typically including at least one closet. Consequently, 'closet' appears more frequently than 'bedroom', the latter being the second most common room type. This dominance of 'closet', primarily due to its smaller size, is evident in Figure 3a. In Figure 4a, the word distribution for both $\gamma(J_H^*)$ and $\gamma(T_H^*)$ across all homes is illustrated. Notably, most descriptions contain approximately 400 words, aligning with the guideline to not exceed 500 words in the prompt. Despite this, due to the stochastic nature of LLMs, some homes exhibit up to 782 words. Interestingly, the word distributions generated by both J_H^* and T_H^* are nearly identical, reflecting similar relevancy scores assigned by human evaluators. This trend is also observed in the unique 4-grams distribution in Figure 4b. The plot demonstrates the increase in unique 4-grams with a growing number of homes. Given the diversity of the descriptions, there are 98996 unique 4-grams for only 1575 homes. It is almost identical for both $\gamma(J_H^*)$ and $\gamma(T_H^*)$. The analysis of the number of unique Part-of-Speech (POS) tags is presented in Table 1. It is observed that $\gamma(J_H^*)$ encompasses a broader array of unique POS tags compared to $\gamma(T_H^*)$, with the exception of adverbs. This observation corroborates the findings from human evaluations, suggesting that the descriptions generated by $\gamma(J_H^*)$ are more diverse, evidenced by a diversity proportion of approximately 0.09 in Sec. 4.5. We discuss the room label distributions extracted directly from text in suppl. material.

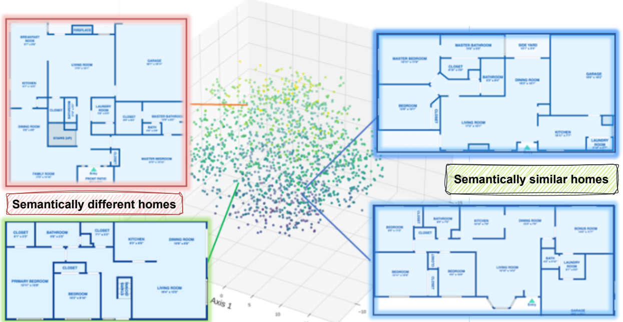


Figure 5. **Visualization of the description embedding space.** Each point in the 3D plot (in the middle) is embedding projection of a single home. The high-dimensional description embeddings are projected into 3D space using t-SNE [27]. Sampling two close embedding points result is very similar homes, i.e., floorplans (in the right), whereas, sampling two points far away corresponds to structurally dissimilar homes as shown in top-left and bottom-left.

6. ZInD-Agent: Zero-Shot Baseline Model

In this section, we discuss the proposed a zero-shot baseline model, ZInD-Agent. Essentially, this model will be based on existing pre-trained model on sevel components. The main purpose of this zero-shot model is to establish a baseline performance to carry forward further research on ZInD-Tell. ZInD-Agent generates descriptions from an unordered set of 360° panorama images, which is expected to approximate the D_H^* descriptions. It consists of multiple modules with different objectives. Hence, these modules work together on the unordered panorama images $\{I_{ij}\}_{j=1}^N$ for every $f_i \in F$ of H and generate D_H end-to-end. We discuss the module-wise performance in suppl. material.

6.1. Assumptions

We assume that the input to the model will be a unordered set of 360° panorama images. The panorama images are captured from indoor and cover entire room. Also, the im-

[3] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

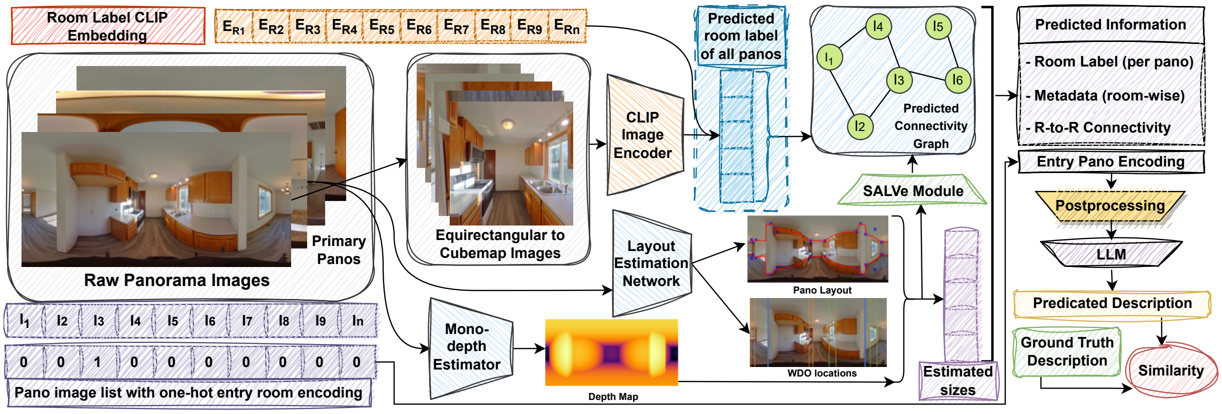


Figure 6. **ZInD-Agent.** This zero-shot model incorporates several existing modules for generating D_H. Given a set of pano images of a home as input, it initially extracts factual information from each image. Subsequently, it predicts the room connectivity graph $\tilde{\mathcal{M}}^H_{f_i}$ based on the panos in f_i. The data are then aggregated on each f_i to construct a schema, then subsequently passed into the LLM to generate D_H.

ages are in equirectangular form, i.e., captured in such a way that it has 360° horizontal and 180° vertical field-of-view of the *room* and all *W/D/O*, respectively. In addition, the image of each room partially overlaps with other adjacent room, i.e., there is a co-visibility between two adjacent rooms via D or O as all doors are opened for all rooms in ZInD dataset. To this end, we also assume that input to the model has floor-wise subset F of the panorama image set, hence $\{I_{ij}\}_{j=1}^{N} \subseteq I_H$. Also, the I_{ij}, for which $S_{f_i} = I_{ij}$, is passed to the model as one-hot encoding.

6.2. Room Classification Module

In this section, we discuss the zero-shot method for room classification. Essentially, the module will classify all I_P. However this is a challenging problem, because to classify a room, several semantic cues, e.g., furnitures play crucial roles. As all rooms in ZInD dataset are empty, it becomes a more challenging problem. We leverage CLIP [23] for this task as this model is jointly trained on a massive scale of Language and Image data and is widely used in many downstream tasks [3, 6, 18]. First, we encode all room labels using CLIP sentence encoder. We augment the label texts such as "This is a <*room_label*>". Then, for each I_{ij}, we convert the equirectangular pano image to 6 cubemap images. We then encode all images using CLIP image encoder except for ceiling and floor images, as they don't contain significant information. Hence, for a $(3 \times X \times Y)$ dimensional panorama image, we have (512×4) dimensional image embedding space. Then, we mean pool it along the 2nd dimension and then find Top-1 cosine similarity label based on the image and text embedding spaces.

6.3. Layout and Size Estimation Module

The task of this module will be to estimate the layout of each I_{ij} along with approximating *W/D/O* locations. We use modified HorizonNet [24] model; it is trained with the partial room shape geometry that can predict both floor-wall boundary and *W/D/O* approximate scores. The model is trained on Zillow's internal data. We apply the pre-trained model to all panos $\{I_H\}_{H \in \mathcal{H}}$ and store the detection results. Then, we use HoHoNet [25], a state-of-the-art mono-depth estimator model for all pano images to calculate depth. After that, we project the depth map to the boundary pixels of both *room* and *W/D/O* to approximate the actual size.

6.4. Room-to-Room Connectivity

For a floor $f_i \in H$ of a home, we want to identify the adjacency of each room, i.e., predict if I_{ij} is adjacent to I_{ik}, where $j \neq k$. As mentioned in Section 6.1, we assume that the panorama images have co-visibility as all doors are open and the pano images are captured such a way that they cover an entire room and fraction of adjacent rooms. The prediction is challenging and there has been several works recently to recover the connectivity geometry, e.g., predicting pose graphs [12, 20]. In this paper, we use state-of-the-art Semantic Alignment Verification (SALVe) [15]. The SALVe system generate multiple alignment hypotheses between (I_{ij}, I_{ik}) based on *W/D/O* by projecting panorama images to bird's eye view (BEV). Then, they verify if both of the projected images contain semantic overlap based on a learned threshold. Finally, the model predicts a pose graph and optimize using GTSAM [9]. We utilize the pretrained SALVe model to infer the floor-wise connectivity graph. In this process, we also build the connectivity-type (door or opening) based on the matched hypotheses. Hence, the predicted connectivity graph $\tilde{\mathcal{M}}^H_{f_i} \approx \mathcal{M}^H_{f_i}$.

6.5. Description Decoder

In this section, we discuss the integration of modules for generating D_H for $\forall H \in \mathcal{H}$, closely mirroring the ground truth decoding steps outlined in Section 6.5. Algorithm 1 from Section 6.5 is employed here, taking $\tilde{\mathcal{M}}^H_{f_i}$ and S_{f_i} for $\forall f_i \in F$, along with the predicted *room* label, size, and *W/D/O* sizes from Sections 6.2, and 6.3, to construct the

Table 2. **Comparative results for Description Generation and Language-Based Home Retrieval.** This table compares the zero-shot baseline model with CLIP-R for retrieval and BLIP-2 for generation tasks, using relevant metrics. Here, B@k and Emb. denote BLEU and Embedding scores, respectively. For the results columns with (↑), higher value indicates better performance, vice versa for (↓) columns.

	Language Based Home Retrieval					**Description Evaluation**					
Model	*R@1 (↑)*	*R@10 (↑)*	*R@20 (↑)*	*MdR (↓)*	*MnR (↓)*	*B@2 (↑)*	*B@4 (↑)*	*METEOR (↑)*	*CIDEr (↑)*	*ROGUE$_L$ (↑)*	*Emb. (↑)*
CLIP-R	0.95 ± 0.32	7.28 ± 0.32	15.19 ± 0.15	75.75 ± 0.25	77.27 ± 0.38	-	-	-	-	-	-
BLIP-2	-	-	-	-	-	12.92 ± 5.33	1.52 ± 1.91	15.75 ± 2.08	16.05 ± 3.65	14.32 ± 2.92	0.52 ± 0.04
ZInD-Agent	3.16 ± 0.32	16.84 ± 1.87	33.51 ± 3.65	48.32 ± 5.83	53.91 ± 3.09	27.26 ± 3.73	10.94 ± 1.89	28.16 ± 2.82	33.18 ± 3.32	32.15 ± 3.82	0.69 ± 0.08

schema J_H. Subsequently, J_H is fed into the GPT-4 LLM using the same prompt as for J_H^* to generate D_H. The subsequent sections will explore the experimental procedure employed to evaluate D_H against D_H^* on several tasks.

7. Experiments

7.1. Language Based Home Retrieval

Here, we evaluate ZInD-Agent's performance on language-based home retrieval task. Section 5 outlines the method of mapping home descriptions into an embedding space, as illustrated in Figure 5. The retrieval process involves using the model to generate home descriptions, embedding these descriptions, and calculating cosine similarity with all ground truth descriptions. We assess retrieval performance using Recall at Rank K (R@K), Median Rank (MdR), and Mean Rank (MnR) [10, 17] (see Table 2). Additionally, we compare ZInD-Agent against a naïve text-to-home retrieval method, termed CLIP-R, which involves mean-pooling embeddings of all pano images I_P extracted with the CLIP image encoder, followed by embedding the ground truth descriptions D_H^* using CLIP sentence encoders. We then compute cosine similarity between the embeddings from the image and description encoders and evaluate recall metrics. Table 2 reveals that ZInD-Agent outperforms CLIP-R by an average margin of 116.93%. This substantial improvement is attributed to effective geometry extraction, essential for accurately identifying key attributes of homes.

7.2. Home Description Generation

This task evaluates the performance of descriptions generated by ZInD-Agent. We employ standard sentence evaluation metrics such as BLEU, METEOR, CIDEr, and ROUGE$_L$, utilizing the MS-COCO toolkit[4]. Additionally, we calculate the cosine similarity between actual and predicted descriptions, with results presented in Table 2. Similar to the retrieval task, a naïve CLIP-based description generation method is implemented, utilizing BLIP-2 [16], a state-of-the-art zero-shot image captioning model. BLIP-2 employs a frozen image encoder and language decoder, linked via latent embedding. Given an image and a prompt, it generates descriptions. In this experiment, mean-pooled panoramic images were used as prompts for generating home descriptions. The performance comparison of ZInD-Agent and BLIP-2, shown in Table 2, reveals that ZInD-Agent significantly outperforms BLIP-2 across all metrics. This superior performance improvement by 180.65% on average (for all metrics) is primarily attributed to the zero-shot model's ability to accurately infer floor-level and room-level contexts from unordered image sets, complemented by its size estimation capabilities, leading to the descriptions that are syntactically and semantically more precise.
Although our proposed zero-shot baseline model surpasses naïve techniques, there is considerable room for improvement, as the current results are far from what would be deemed robust performance. Consequently, follow-up research focusing on end-to-end learning, utilizing the **ZInD-Tell** dataset, are anticipated to be pivotal in enhancing the overall performance on both tasks.

8. Limitations and Future Extensions

We acknowledge that as the ZInD dataset is limited to North American residential homes, descriptions in ZInD-Tell are also limited to that scope. Nevertheless, this research opens avenues for further exploration in a novel domain. At the dataset level, our efforts will expand into the Visual Question Answering [22] task, focusing on grounding, where the rationale for each sentence generation is linked to the corresponding image. Technically, ZInD-Tell dataset facilitates semantic matching of homes or floorplans, enabling search, retrieval, and generation tasks based on natural language queries. For instance, it allows a model to associate queries with original homes, or generating entire floorplans.

9. Conclusion

This paper addresses the novel challenge of generating natural language descriptions from unordered indoor panorama images. We present the novel ZInD-Tell dataset, created on top of ZInD, detailing information extraction, construction, and evaluation methods. To the best of our knowledge, ZInD-Tell is the first dataset for natural language descriptions of indoor homes. We then introduce ZInD-Agent, a zero-shot baseline, to initiate the benchmarking on ZInD-Tell dataset, focusing on home retrieval and description evaluation tasks. In both tasks, ZInD-Agent outperforms naïve methods, emphasizing the significance of the room-level and floor-level geometric information to semantically understand the scene for home description generation.

[4] https://github.com/tylin/coco-caption

References

[1] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 2, 3

[2] Iro Armeni, Zhi-Yang He, JunYoung Gwak, Amir R Zamir, Martin Fischer, Jitendra Malik, and Silvio Savarese. 3D scene graph: A structure for unified semantics, 3d space, and camera. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 5664–5673, 2019. 3

[3] Manuele Barraco, Marcella Cornia, Silvia Cascianelli, Lorenzo Baraldi, and Rita Cucchiara. The unreasonable effectiveness of clip features for image captioning: an experimental analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4662–4670, 2022. 7

[4] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from RGB-D data in indoor environments. *International Conference on 3D Vision (3DV)*, 2017. 3

[5] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. *Proceedings of the 16th European Conference on Computer Vision (ECCV)*, 2020. 1, 3

[6] Jaemin Cho, Seunghyun Yoon, Ajinkya Kale, Franck Dernoncourt, Trung Bui, and Mohit Bansal. Fine-grained image captioning with clip reward. *arXiv preprint arXiv:2205.13115*, 2022. 7

[7] Steve Cruz, Will Hutchcroft, Yuguang Li, Naji Khosravan, Ivaylo Boyadzhiev, and Sing Bing Kang. Zillow indoor dataset: Annotated floor plans with 360deg panoramas and 3d room layouts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2133–2143, 2021. 2, 3

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

[9] Frank Dellaert. Factor graphs and gtsam: A hands-on introduction. *Georgia Institute of Technology, Tech. Rep*, 2:4, 2012. 7

[10] Han Fang, Pengfei Xiong, Luhui Xu, and Yu Chen. Clip2video: Mastering video-text retrieval via image clip. *arXiv preprint arXiv:2106.11097*, 2021. 8

[11] Francesco Giuliari, Geri Skenderi, Marco Cristani, Yiming Wang, and Alessio Del Bue. Spatial commonsense graph for object localisation in partial scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19518–19527, 2022. 3

[12] Will Hutchcroft, Yuguang Li, Ivaylo Boyadzhiev, Zhiqiang Wan, Haiyan Wang, and Sing Bing Kang. Covispose: Co-visibility pose transformer for wide-baseline relative pose estimation in 360 indoor panoramas. In *European Conference on Computer Vision (ECCV)*, pages 615–633. Springer, 2022. 7

[13] Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Daniel Gordon, Yuke Zhu, Abhinav Gupta, and Ali Farhadi. AI2-THOR: An Interactive 3D Environment for Visual AI. *arXiv*, 2017. 2, 3

[14] Chen Kong, Dahua Lin, Mohit Bansal, Raquel Urtasun, and Sanja Fidler. What are you talking about? text-to-image coreference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 1, 2

[15] John Lambert, Yuguang Li, Ivaylo Boyadzhiev, Lambert Wixson, Manjunath Narayana, Will Hutchcroft, James Hays, Frank Dellaert, and Sing Bing Kang. Salve: Semantic alignment verification for floorplan reconstruction from sparse panoramas. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 647–664. Springer, 2022. 2, 7

[16] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 8

[17] Niluthpol Chowdhury Mithun, Juncheng Li, Florian Metze, and Amit K Roy-Chowdhury. Learning joint embedding with multimodal cues for cross-modal video-text retrieval. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, pages 19–27, 2018. 8

[18] Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021. 7

[19] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2012. 2

[20] Negar Nejatishahidin, Will Hutchcroft, Manjunath Narayana, Ivaylo Boyadzhiev, Yuguang Li, Naji Khosravan, Jana Košecká, and Sing Bing Kang. Graph-covis: Gnn-based multi-view panorama global pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6458–6467, 2023. 7

[21] OpenAI. Gpt-4 technical report, 2023. 2, 5

[22] Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. Reverie: Remote embodied visual referring expression in real indoor environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9982–9991, 2020. 1, 3, 8

[23] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763. PMLR, 2021. 2, 7

[24] Cheng Sun, Chi-Wei Hsiao, Min Sun, and Hwann-Tzong Chen. Horizonnet: Learning room layout with 1d represen-

tation and pano stretch data augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1047–1056, 2019. 2, 7

[25] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Hohonet: 360 indoor holistic understanding with latent horizontal features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2573–2582, 2021. 2, 7

[26] Jesse Thomason, Michael Murray, Maya Cakmak, and Luke Zettlemoyer. Vision-and-dialog navigation. In *Conference on Robot Learning*, pages 394–406. PMLR, 2020. 1, 3

[27] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9 (86):2579–2605, 2008. 6

[28] Johanna Wald, Helisa Dhamo, Nassir Navab, and Federico Tombari. Learning 3d semantic scene graphs from 3d indoor reconstructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3961–3970, 2020. 3

[29] Su Wang, Ceslee Montgomery, Jordi Orbay, Vighnesh Birodkar, Aleksandra Faust, Izzeddin Gur, Natasha Jaques, Austin Waters, Jason Baldridge, and Peter Anderson. Less is more: Generating grounded navigation instructions from landmarks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15428–15438, 2022. 3