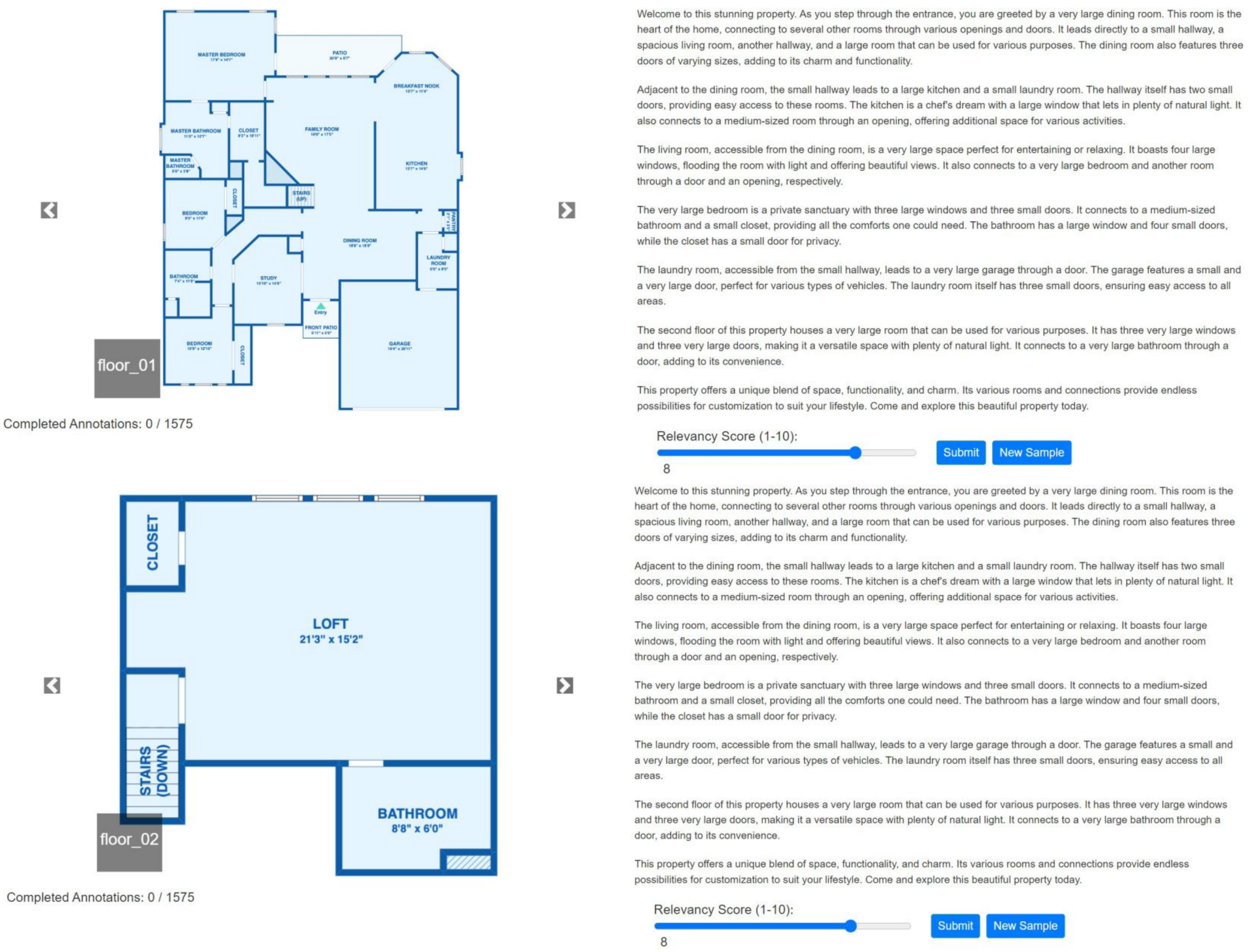
ZInD-Tell: Towards Translating Indoor Panoramas into Descriptions

Supplementary Material



Completed Annotations: 0 / 1575

Welcome to this stunning property. As you step through the entrance, you are greeted by a very large dining room. This room is the heart of the home, connecting to several other rooms through various openings and doors. It leads directly to a small hallway, a spacious living room, another hallway, and a large room that can be used for various purposes. The dining room also features three doors of varying sizes, adding to its charm and functionality.

Adjacent to the dining room, the small hallway leads to a large kitchen and a small laundry room. The hallway itself has two small doors, providing easy access to these rooms. The kitchen is a chef's dream with a large window that lets in plenty of natural light. It also connects to a medium-sized room through an opening, offering additional space for various activities.

The living room, accessible from the dining room, is a very large space perfect for entertaining or relaxing. It boasts four large windows, flooding the room with light and offering beautiful views. It also connects to a very large bedroom and another room through a door and an opening, respectively.

The very large bedroom is a private sanctuary with three large windows and three small doors. It connects to a medium-sized bathroom and a small closet, providing all the comforts one could need. The bathroom has a large window and four small doors, while the closet has a small door for privacy.

The laundry room, accessible from the small hallway, leads to a very large garage through a door. The garage features a small and a very large door, perfect for various types of vehicles. The laundry room itself has three small doors, ensuring easy access to all areas.

The second floor of this property houses a very large room that can be used for various purposes. It has three very large windows and three very large doors, making it a versatile space with plenty of natural light. It connects to a very large bathroom through a door, adding to its convenience.

This property offers a unique blend of space, functionality, and charm. Its various rooms and connections provide endless possibilities for customization to suit your lifestyle. Come and explore this beautiful property today.

Relevancy Score (1-10):

8

Completed Annotations: 0 / 1575

Welcome to this stunning property. As you step through the entrance, you are greeted by a very large dining room. This room is the heart of the home, connecting to several other rooms through various openings and doors. It leads directly to a small hallway, a spacious living room, another hallway, and a large room that can be used for various purposes. The dining room also features three doors of varying sizes, adding to its charm and functionality.

Adjacent to the dining room, the small hallway leads to a large kitchen and a small laundry room. The hallway itself has two small doors, providing easy access to these rooms. The kitchen is a chef's dream with a large window that lets in plenty of natural light. It also connects to a medium-sized room through an opening, offering additional space for various activities.

The living room, accessible from the dining room, is a very large space perfect for entertaining or relaxing. It boasts four large windows, flooding the room with light and offering beautiful views. It also connects to a very large bedroom and another room through a door and an opening, respectively.

The very large bedroom is a private sanctuary with three large windows and three small doors. It connects to a medium-sized bathroom and a small closet, providing all the comforts one could need. The bathroom has a large window and four small doors, while the closet has a small door for privacy.

The laundry room, accessible from the small hallway, leads to a very large garage through a door. The garage features a small and a very large door, perfect for various types of vehicles. The laundry room itself has three small doors, ensuring easy access to all areas.

The second floor of this property houses a very large room that can be used for various purposes. It has three very large windows and three very large doors, making it a versatile space with plenty of natural light. It connects to a very large bathroom through a door, adding to its convenience.

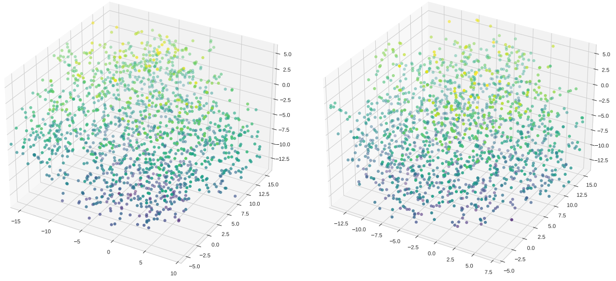
This property offers a unique blend of space, functionality, and charm. Its various rooms and connections provide endless possibilities for customization to suit your lifestyle. Come and explore this beautiful property today.

Relevancy Score (1-10):

8

Figure 7. The User Interface (UI) is designed to collect **relevancy scores** from human evaluators. The left pane displays the floorplan images in a carousel, accompanied by their respective `floor_num`. The bottom section indicates the number of homes being annotated. On the right, evaluators are presented with a random description from $\gamma(J_H^*)$ or $\gamma(T_H^*)$. Evaluators review the description and examine the corresponding floorplan(s) in the carousel (shown in two images: `floor_01`, `floor_02`). They then select a 'relevancy score' and click 'submit'. After that, the UI refreshes, presenting a new home and description. If an evaluator wishes to annotate at a later time, they can select "New Sample" to view a different home.

In the supplementary material, we investigate additional aspects of the dataset, followed by ablation study experiments. Initially, the construction of the User Interface (UI) and evaluation collection procedures are discussed. Subsequently, a comprehensive experiment on the embedding space of the ZInD-Tell dataset is conducted. This is followed by an analysis of the room label distribution in the dataset and its correlation with the J_H^* and $\gamma(J_H^*)$. Furthermore, the module-wise performance of ZInD-Agent is examined, along with a detailed exposition of the implementation steps for the naïve methods. Finally, we present a home (panorama and floorplan images) with the corresponding descriptions predictions from both ZInD-Agent and the naïve methods, followed by syntactic and semantic evaluation scores.



(a) $\gamma(T_H^*)$ (b) $\gamma(J_H^*)$

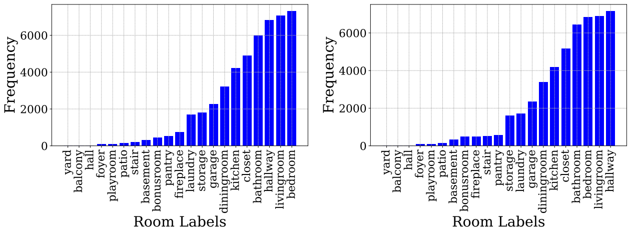
Figure 8. **Embedding Space Visualization** using ‘all-MiniLM-L6-370v2’ in 3-D space using for the descriptions $\gamma(J_H^*)$, $\gamma(T_H^*)$.

10. Manual Evaluation of Home Descriptions

The user interface (UI) used for manual evaluation of home descriptions is shown in Figure 7. The left pane displays floorplan images in a carousel format. Given the possibility of multiple floorplans, evaluators can navigate through these images. The floorplan id is displayed in the bottom left corner of each image. Subsequently, a random description from the set $\{\gamma(J_H^*), \gamma(T_H^*)\}$ is chosen and displayed in the right pane, which continues for 1575 homes. Hence, each evaluator evaluates 3150 descriptions. They are tasked with assessing the consistency of the description with the corresponding floorplans and assigning a **relevancy score** on a Likert scale ranging from 1 to 10. Upon submission, the score is recorded, and the floorplan’s description is not presented to the evaluator again. This process is repeated for each home. The bottom left of Figure 7 shows the annotation progress for homes, with a home marked as complete once both descriptions in $\{\gamma(J_H^*), \gamma(T_H^*)\}$ are evaluated. Notably, evaluators are unaware of the specific source ($\{\gamma(J_H^*)$ or $\gamma(T_H^*)\}$) of each description, ensuring an unbiased assessment. Data records are managed in JSON format.

11. More Analysis on the Embedding Space

In this section, we revisit the embedding visualization conducted in Figure 5 of the main paper. We initially mapped $\gamma(J_H^*)$ into a 384-dimensional embedding space using the ‘all-MiniLM-L6-370-v2’ sentence encoder[5], followed by a reduction to three dimensions using t-SNE. As a follow-up experiment, we undertake two tasks: first, visualizing the distribution of embeddings using $\gamma(T_H^*)$, and second, computing the average cosine similarity between pairs of embedding points within the same home. As depicted in Figure 8, the embedding spaces of both $\gamma(J_H^*)$ and $\gamma(T_H^*)$ appear nearly identical visually. The pairwise cosine similarity is 0.89 ± 0.04, suggesting that the descriptions are semantically very similar. We further extend the embedding space analysis to $\{\gamma(J_H^*), \gamma(T_H^*)\}$ using three larger models: T5 [22], DeBERTa [12], and MPNet [26]. The cosine similarities obtained are 0.97 ± 0.01 for T5, 0.97 ± 0.02 for DeBERTa, and 0.92 ± 0.03 for MPNet-based embeddings. These results corroborate our assertion in Section 5 regarding the superior semantic discernment of larger models.



(a) $\gamma(T_H^*)$ (b) $\gamma(J_H^*)$

Figure 9. **Room label distribution** in corpora containing descriptions of all homes $\gamma(J_H^*)$, $\gamma(T_H^*)$.

12. Syntactic Similarity Analysis

In this section, we conduct further analyses on $\gamma(J_H^*)$ and $\gamma(T_H^*)$. Unlike the computation of pairwise cosine distances, i.e., semantic similarities in the embedding space, we undertake a comparative assessment using conventional sentence evaluation metrics. This approach aims to examine the syntactic (structural) similarity of descriptions for each home. Such evaluation is crucial for ensuring syntactic diversity between $\gamma(J_H^*)$ and $\gamma(T_H^*)$. We utilize the same metrics previously applied in Section 7.2. The performance outcomes for the entire dataset (1575 homes) and the test set (158 homes) are presented in Table 3. These results indicate that while the descriptions exhibit high semantic similarity (up to 0.97), they are not syntactically identical. This observation suggests potential implications for future experiments with ZInD-Tell. Despite both $\gamma(J_H^*)$ and $\gamma(T_H^*)$ conforming to similar word and n-gram distributions, as observed in Figures 4b, 4a, they differ significantly in a one-to-one syntactic comparison, thus offering sufficient diversity for an end-to-end model to learn from the descriptions.

13. Room Label Distribution Across ZInD-Tell

In this section, we present an alternative technique for computing the distribution of room labels. As depicted in Figure 3b, the initial study computed these distributions directly

Table 3. **Pair-wise Syntactic Similarity Analysis** of ZInD-Tell

Set	BLEU-2	BLEU-4	ROUGE$_\text{L}$	METEOR	CIDEr
Full	39.10 ± 5.59	18.18 ± 4.25	33.17 ± 4.72	37.19 ± 4.34	41.38 ± 3.67
Test	46.67 ± 6.68	27.29 ± 6.58	44.62 ± 6.47	42.78 ± 5.59	46.91 ± 5.47

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

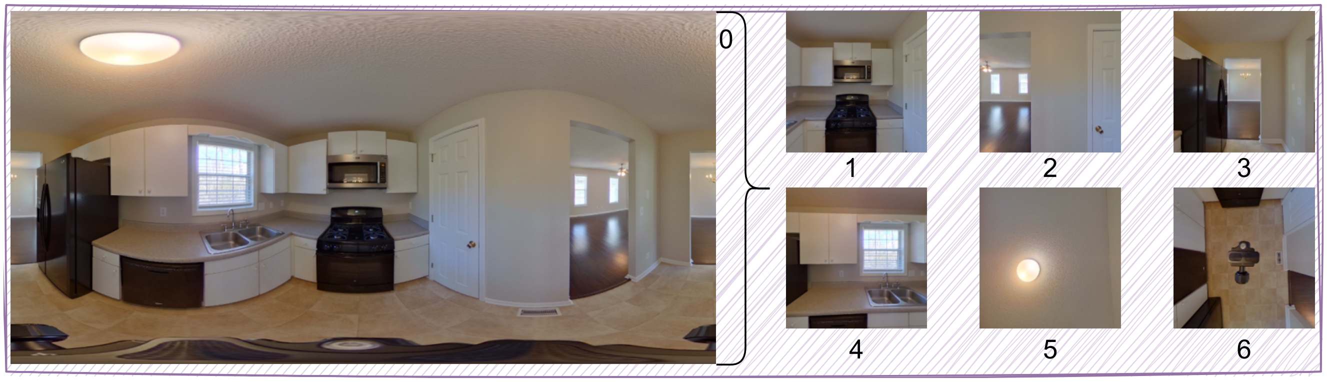


Figure 10. **Conversion of Equirectangular Panorama Image into Cubefaces.** On the left, the equirectangular panorama image, denoted as I_{ij} and indexed as 0, is decomposed into six (256×256) cubeface images, as depicted on the right. These cubefaces are assigned indices, which are referenced in Tables 5 and 6 for discussing classification accuracy.

from J_H^*. Our analysis investigates whether $\gamma(J_H^*)$ and $\gamma(T_H^*)$ yield similar distributions. This examination is pivotal in ascertaining that the LLM does not favor frequently occurring labels in generating descriptions. We conduct a quantitative analysis to confirm the impartiality of the descriptions, despite the high **relevance scores** from evaluators, as outlined in Table 1. Our method employs a Part-of-Speech (POS) tagging approach for extracting nouns as potential room indicators and quantifying their occurrence relative to the label dictionary, as shown in Figure 3b. This approach is uniformly applied to $\gamma(J_H^*)$ and $\gamma(T_H^*)$ corpora. In Figure 9, we present the distributions. Notably, the original distribution in J_H^* (Figure 3b) predominantly features 'closet', while in $\gamma(J_H^*)$ and $\gamma(T_H^*)$, it ranks fifth. This shift suggests that the generated descriptions diversify the focus on various rooms. For instance, the top three room labels in both $\gamma(J_H^*)$ and $\gamma(T_H^*)$ are 'bedroom', 'living room', and 'hallway', illustrating a comprehensive interior description. Additionally, $\gamma(J_H^*)$ and $\gamma(T_H^*)$ exhibit similar label frequencies, with notable deviations. Specifically, $\gamma(J_H^*)$ describes 'hallway' 6813 times, while $\gamma(T_H^*)$ emphasizes it 7157 times. We attribute these differences to the syntactic and semantic variances discussed in Sections 12 and 11.

14. Module-wise Performance of ZInD-Agent

In this section, we evaluate the performance of each component within the zero-shot ZInD-Agent baseline.

Room Classification: This zero-shot module, implemented using CLIP [25] with 'ViT-B/32' image and text encoder, processes mean-pooled cubeface images extracted from indoor panorama images, excluding floor and ceiling images due to their limited informational content (discussed in Section 6.2). We experimented with all combinations of 7 images (one panoramic and six (256×256) cubefaces, detailed in Figure 10), employing mean pooling to ascertain the optimal arrangement for classification accuracy. Results are presented in Table 5. Notably, omitting ceiling images enhances accuracy, reaching a peak of 30.38%, on predicting 22 room labels. Directly utilizing equirectangular images yields a comparable accuracy of 30.08%. A subsequent experiment focusing on top-5 classification accuracy (Table 6) revealed the highest accuracy of 69.68% when combining the equirectangular image with front and back images. This leads to two key insights: first, the modest top-1 room classification accuracy underscores the need for further research in this area; second, while cubemap images are crucial for top-1 classification, equirectangular images significantly influence top-5 prediction accuracy. The sentences per room label used for producing CLIP text embeddings are shown in Table 4.

Layout Estimation and W/D/O Approximation: As delineated in Section 6.3, we use a modified HorizonNet [27] trained for simultaneous room and W/D/O layout prediction. Utilizing the pre-trained module, we adhere to its reported performance in [16]. To assess layout estimation accuracy, the research compares the 2D Intersection over Union (IoU) between predicted and actual layouts. The W/D/O evaluation, following 1D projection, is based on 1D IoU accuracy. On the ZInD test set, the module achieved an IoU accuracy of 85% for room layout estimation. For W/D/O detection, it attained a 70% accuracy threshold in 1D IoU, with F1 scores of 0.91, 0.89, and 0.67 for windows, doors, and openings, respectively. This indicates a relatively lower score in openings detection, attributed by the authors to annotation errors in the ZInD dataset.

Room-to-Room Connectivity: We adopt SALVe [16] as our core implementation, as delineated in Section 6.4. SALVe, rigorously trained on the ZInD training set and evaluated against its test set, is tasked with predicting global

Table 4. **CLIP Descriptions for Rooms**. Descriptions assigned to each room facilitate sentence embeddings, aiding in room classification.

Room	Description
Bedroom	a space with potential for a cozy bedroom setup, typically including areas for a bed and wardrobe
Bathroom	an empty bathroom space, commonly featuring areas for a toilet, shower, and sink
Basement	a spacious and empty basement area, often used for storage or recreational purposes
Playroom	an open space suitable for a playroom setup, ideal for children's activities and toys
Storage	a room designated for storage, potentially with shelves or cabinets for organization
Laundry	an area for laundry, typically with connections for a washer and dryer
Livingroom	a large empty space ideal for a living room, often featuring areas for sofas and a TV
Kitchen	an empty space with potential for a kitchen, usually including fittings for a sink, cabinets, and appliances
Hallway	a connecting hallway, spacious and empty, typically leading to other rooms
Pool	an area for an indoor or outdoor pool, often accompanied by poolside fittings
Balcony	an open balcony space, empty but with potential for outdoor seating and a view
Closet	a smaller room for storage, possibly a walk-in closet with shelving and hanging space
Stair	a space with staircases, connecting different floors or levels
Diningroom	an open area suitable for a dining setup, typically including space for a dining table and chairs
Garage	a spacious garage area, empty, often used for vehicle parking and storage
Yard	an outdoor yard space, open and versatile for landscaping or outdoor activities
Other	a room with unspecified or variable characteristics, adaptable to different uses
Fireplace	a space centered around a fireplace, potentially a focal point in a living or sitting room
Pantry	a smaller area designated for pantry or food storage, often with shelving
Hall	an empty hall, versatile for various uses, possibly connecting different areas
Foyer	an entrance foyer, spacious and welcoming, often leading to main living areas
Patio	an outdoor patio space, open and adaptable for outdoor furniture and activities

poses from an unordered collection of floor-level panorama images. This inference facilitates the construction of a pose graph, subsequently refined through GTSAM optimization [9]. The technique method achieved a localization accuracy of 60.70%, a pivotal metric influencing the accuracy of the ensuing room-to-room connectivity graph. Although 60.70% does not represent an exemplary level of accuracy, it is currently the state-of-the-art approach that exploits co-visibility among panoramas to estimate the pose graph, and by extension, the connectivity graph. Importantly, the accuracy of this module significantly affects the description generation phase of ZInD-Agent, as the predicted descriptions should align closely with the spatial floor-level geometry.

15. More Details on the Naïve Methods

Here, we provide additional details about the implementation of the naive methods discussed in Section 7.

CLIP-R: As previously stated, we aggregate (mean-pool) the embeddings of all panorama images associated with each home. For text encoding, the limitation of CLIP models to 77 tokens necessitates splitting each home's description into segments not exceeding this token count. Subsequently, we employ the CLIP sentence encoder on these segments to obtain their embedding vectors, which are then aggregated through mean-pooling to facilitate the final vector for performing home retrieval evaluation.

BLIP-2: This model generally produces a single caption per image. Therefore, we input each panorama image of a home into the model, yielding a series of captions that individually describe each image. These captions are subsequently consolidated into a single paragraph. The resulting combined description is then used for benchmarking on the ZInD-Tell dataset, with results presented in Table 2.

16. Empirical Parameters Setup

In Section 4.2, we discussed various parameters for extracting geometry information. The door-based connectivity angle threshold is set to $\theta_d = 10°$, with a corresponding distance threshold of $\beta_d = 0.1$ units. For opening-based connectivity, these thresholds are $\theta_o = 5°$ and $\beta_o = 0.1$ units,

Table 5. **Top-1 CLIP Classification Accuracy** on the ZInD-Tell Test Set. The **Comb.** column represents the set of images used for mean-pooling, while the **Acc.** column indicates the corresponding accuracy values. Image 0 is the equirectangular image, and the remaining six are the extracted cubeface images from image 0 (an example shown in Figure 10).

Comb.	**Acc.**	**Comb.**	**Acc.**	**Comb.**	**Acc.**	**Comb.**	**Acc.**	**Comb.**	**Acc.**	**Comb.**	**Acc.**	**Comb.**	**Acc.**
(0,)	24.62	(2, 4)	25.22	(0, 3, 5)	22.20	(2, 4, 6)	27.30	(0, 2, 4, 5)	26.97	(2, 3, 5, 6)	23.56	(1, 2, 3, 4, 6)	**30.38**
(1,)	25.16	(2, 5)	20.27	(0, 3, 6)	23.68	(2, 5, 6)	23.38	(0, 2, 4, 6)	27.81	(2, 4, 5, 6)	26.79	(1, 2, 3, 5, 6)	27.63
(2,)	20.54	(2, 6)	24.40	(0, 4, 5)	24.68	(3, 4, 5)	23.08	(0, 2, 5, 6)	25.73	(3, 4, 5, 6)	24.40	(1, 2, 4, 5, 6)	28.93
(3,)	15.54	(3, 4)	23.47	(0, 4, 6)	26.06	(3, 4, 6)	24.89	(0, 3, 4, 5)	25.61	(0, 1, 2, 3, 4)	29.38	(1, 3, 4, 5, 6)	28.02
(4,)	21.57	(3, 5)	16.83	(0, 5, 6)	22.59	(3, 5, 6)	19.70	(0, 3, 4, 6)	26.94	(0, 1, 2, 3, 5)	27.21	(2, 3, 4, 5, 6)	26.97
(5,)	11.34	(3, 6)	19.22	(1, 2, 3)	26.91	(4, 5, 6)	23.59	(0, 3, 5, 6)	23.38	(0, 1, 2, 3, 6)	28.69	(0, 1, 2, 3, 4, 5)	29.20
(6,)	16.11	(4, 5)	20.78	(1, 2, 4)	28.93	(0, 1, 2, 3)	27.96	(0, 4, 5, 6)	25.85	(0, 1, 2, 4, 5)	28.51	(0, 1, 2, 3, 4, 6)	30.08
(0, 1)	26.33	(4, 6)	23.95	(1, 2, 5)	25.22	(0, 1, 2, 4)	28.81	(1, 2, 3, 4)	29.32	(0, 1, 2, 4, 6)	28.81	(0, 1, 2, 3, 5, 6)	28.02
(0, 2)	25.49	(5, 6)	16.95	(1, 2, 6)	28.08	(0, 1, 2, 5)	26.61	(1, 2, 3, 5)	26.61	(0, 1, 2, 5, 6)	27.45	(0, 1, 2, 4, 5, 6)	28.99
(0, 3)	21.99	(0, 1, 2)	27.12	(1, 3, 4)	27.81	(0, 1, 2, 6)	27.96	(1, 2, 3, 6)	28.33	(0, 1, 3, 4, 5)	28.11	(0, 1, 3, 4, 5, 6)	28.42
(0, 4)	25.76	(0, 1, 3)	26.55	(1, 3, 5)	23.86	(0, 1, 3, 4)	28.14	(1, 2, 4, 5)	28.14	(0, 1, 3, 4, 6)	28.48	(0, 2, 3, 4, 5, 6)	28.11
(0, 5)	22.23	(0, 1, 4)	27.24	(1, 3, 6)	25.94	(0, 1, 3, 5)	25.82	(1, 2, 4, 6)	29.44	(0, 1, 3, 5, 6)	26.64	(1, 2, 3, 4, 5, 6)	29.44
(0, 6)	23.50	(0, 1, 5)	24.31	(1, 4, 5)	25.97	(0, 1, 3, 6)	27.06	(1, 2, 5, 6)	27.39	(0, 1, 4, 5, 6)	27.39	(0, 1, 2, 3, 4, 5, 6)	29.62
(1, 2)	26.24	(0, 1, 6)	25.88	(1, 4, 6)	27.75	(0, 1, 4, 5)	26.33	(1, 3, 4, 5)	27.03	(0, 2, 3, 4, 5)	27.75		
(1, 3)	24.22	(0, 2, 3)	25.40	(1, 5, 6)	24.59	(0, 1, 4, 6)	28.11	(1, 3, 4, 6)	28.75	(0, 2, 3, 4, 6)	28.08		
(1, 4)	26.76	(0, 2, 4)	27.51	(2, 3, 4)	25.76	(0, 1, 5, 6)	25.88	(1, 3, 5, 6)	24.74	(0, 2, 3, 5, 6)	25.67		
(1, 5)	22.08	(0, 2, 5)	24.68	(2, 3, 5)	22.05	(0, 2, 3, 4)	28.14	(1, 4, 5, 6)	27.45	(0, 2, 4, 5, 6)	27.63		
(1, 6)	25.43	(0, 2, 6)	26.15	(2, 3, 6)	24.04	(0, 2, 3, 5)	24.98	(2, 3, 4, 5)	25.82	(0, 3, 4, 5, 6)	26.09		
(2, 3)	22.32	(0, 3, 4)	26.21	(2, 4, 5)	25.43	(0, 2, 3, 6)	26.15	(2, 3, 4, 6)	27.27	(1, 2, 3, 4, 5)	28.96		

Table 6. **Top-5 CLIP Classification Accuracy** on the ZInD-Tell Test Set. The **Comb.** column represents the set of images used for mean-pooling, while the **Acc.** column indicates the corresponding accuracy values. Image 0 is the equirectangular image, and the remaining six are the extracted cubeface images from image 0 (an example shown in Figure 10).

Comb.	**Acc.**	**Comb.**	**Acc.**	**Comb.**	**Acc.**	**Comb.**	**Acc.**	**Comb.**	**Acc.**	**Comb.**	**Acc.**	**Comb.**	**Acc.**
(0,)	64.10	(2, 4)	64.04	(0, 3, 5)	64.01	(2, 4, 6)	60.87	(0, 2, 4, 5)	64.83	(2, 3, 5, 6)	57.25	(1, 2, 3, 4, 6)	65.25
(1,)	66.79	(2, 5)	55.75	(0, 3, 6)	64.62	(2, 5, 6)	53.00	(0, 2, 4, 6)	65.49	(2, 4, 5, 6)	58.16	(1, 2, 3, 5, 6)	62.65
(2,)	61.57	(2, 6)	56.83	(0, 4, 5)	63.53	(3, 4, 5)	59.25	(0, 2, 5, 6)	61.03	(3, 4, 5, 6)	56.50	(1, 2, 4, 5, 6)	62.23
(3,)	56.08	(3, 4)	62.78	(0, 4, 6)	63.53	(3, 4, 6)	60.78	(0, 3, 4, 5)	65.07	(0, 1, 2, 3, 4)	68.72	(1, 3, 4, 5, 6)	62.20
(4,)	60.33	(3, 5)	52.10	(0, 5, 6)	54.30	(3, 5, 6)	51.55	(0, 3, 4, 6)	65.70	(0, 1, 2, 3, 5)	67.06	(2, 3, 4, 5, 6)	60.06
(5,)	38.82	(3, 6)	54.81	(1, 2, 3)	66.76	(4, 5, 6)	52.91	(0, 3, 5, 6)	60.81	(0, 1, 2, 3, 6)	68.39	(0, 1, 2, 3, 4, 5)	66.76
(6,)	42.75	(4, 5)	55.60	(1, 2, 4)	67.57	(0, 1, 2, 3)	69.59	(0, 4, 5, 6)	60.36	(0, 1, 2, 4, 5)	67.27	(0, 1, 2, 3, 4, 6)	67.81
(0, 1)	68.66	(4, 6)	56.26	(1, 2, 5)	63.50	(0, 1, 2, 4)	69.17	(1, 2, 3, 4)	66.79	(0, 1, 2, 4, 6)	67.93	(0, 1, 2, 3, 5, 6)	65.88
(0, 2)	67.90	(5, 6)	44.10	(1, 2, 6)	63.74	(0, 1, 2, 5)	67.48	(1, 2, 3, 5)	64.13	(0, 1, 2, 5, 6)	65.55	(0, 1, 2, 4, 5, 6)	65.97
(0, 3)	67.81	(0, 1, 2)	69.65	(1, 3, 4)	67.30	(0, 1, 2, 6)	67.75	(1, 2, 3, 6)	64.92	(0, 1, 3, 4, 5)	67.18	(0, 1, 3, 4, 5, 6)	65.91
(0, 4)	66.52	(0, 1, 3)	**69.68**	(1, 3, 5)	63.50	(0, 1, 3, 4)	68.72	(1, 2, 4, 5)	64.40	(0, 1, 3, 4, 6)	67.66	(0, 2, 3, 4, 5, 6)	64.28
(0, 5)	58.94	(0, 1, 4)	68.81	(1, 3, 6)	64.49	(0, 1, 3, 5)	67.36	(1, 2, 4, 6)	64.46	(0, 1, 3, 5, 6)	65.58	(1, 2, 3, 4, 5, 6)	62.99
(0, 6)	58.73	(0, 1, 5)	66.18	(1, 4, 5)	62.62	(0, 1, 3, 6)	68.39	(1, 2, 5, 6)	59.67	(0, 1, 4, 5, 6)	65.01	(0, 1, 2, 3, 4, 5, 6)	65.82
(1, 2)	67.33	(0, 1, 6)	66.06	(1, 4, 6)	63.47	(0, 1, 4, 5)	66.94	(1, 3, 4, 5)	64.28	(0, 2, 3, 4, 5)	65.13		
(1, 3)	67.00	(0, 2, 3)	68.45	(1, 5, 6)	55.48	(0, 1, 4, 6)	66.67	(1, 3, 4, 6)	65.52	(0, 2, 3, 4, 6)	66.33		
(1, 4)	67.03	(0, 2, 4)	67.66	(2, 3, 4)	64.13	(0, 1, 5, 6)	62.93	(1, 3, 5, 6)	60.12	(0, 2, 3, 5, 6)	63.41		
(1, 5)	59.94	(0, 2, 5)	64.34	(2, 3, 5)	60.39	(0, 2, 3, 4)	67.24	(1, 4, 5, 6)	60.00	(0, 2, 4, 5, 6)	63.29		
(1, 6)	59.97	(0, 2, 6)	64.65	(2, 3, 6)	60.72	(0, 2, 3, 5)	65.31	(2, 3, 4, 5)	62.08	(0, 3, 4, 5, 6)	62.99		
(2, 3)	62.81	(0, 3, 4)	67.24	(2, 4, 5)	60.97	(0, 2, 3, 6)	66.64	(2, 3, 4, 6)	62.50	(1, 2, 3, 4, 5)	64.56		

respectively. Additionally, we enforced a constraint that the centroid of the rooms connected by doors or openings should not exceed a distance of 0.3 units. These parameters remained consistent across all evaluated homes.

17. Output Visualization

In this section, we present an example demonstrating the panorama images of a home in Figure 11, along with the descriptions predicted by both the naïve method and the ZInD-Agent module, as detailed in Table 7. Specifically, Figure 11 displays the floor-wise panorama image sets on each row. Also, in left of each row, the corresponding floorplans are shown for added clarity. The outputs from the naïve BLIP-2 method and the ZInD-Agent module are depicted in the top two rows of Table 7, respectively. The Ground Truth description, $\gamma(J_H^*)$, appears in the last row. Additionally, the Table includes explicit values of the evaluation metrics in the rightmost column. Notably, ZInD-Agent surpasses the naïve BLIP-2 in both qualitative assessments and metric evaluations. However, inaccuracies in the ZInD-Agent's descriptions regarding room labels and connectivity are observed. These discrepancies are likely due to the modular performances discussed in Section 14, suggesting that enhancing individual modules could significantly improve overall description quality. We also advocate that an end-to-end learning technique might overcome the current limitations of module-wise ZInD-Agent by incorporating novel features directly from panorama images for more accurate description generation.

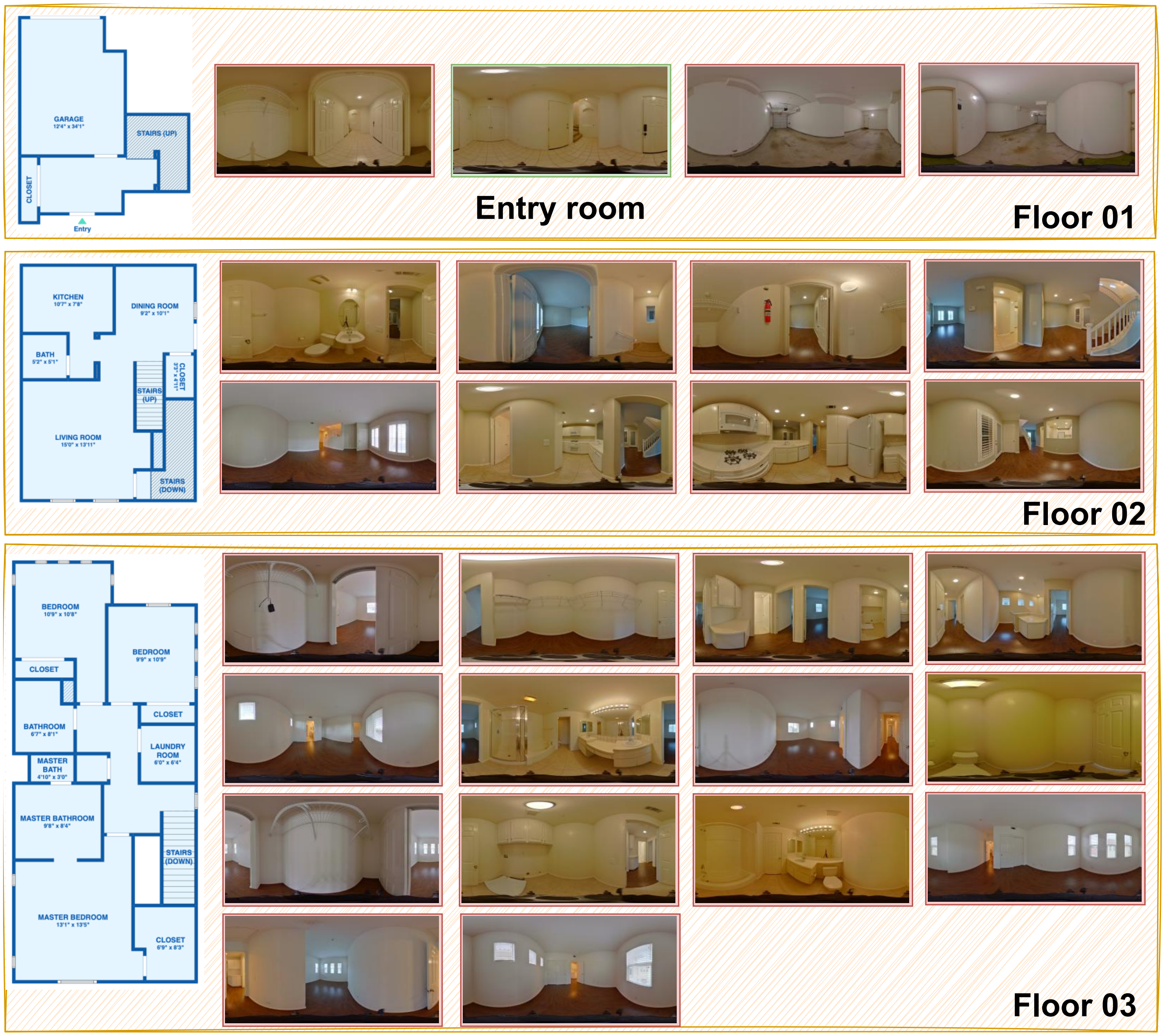


Figure 11. **Example Home with Panorama and Floorplan Images.** This example depicts a home comprising three floors. In the left column, the ground-truth floorplan image of each floor is displayed. Corresponding to each floorplan, the primary panorama images are presented. The models (BLIP-2 and ZInD-Agent) receive exclusively these floor-wise panorama images. Additionally, they have access to the panorama image ID of the entry room, highlighted in a green box. Table 7 shows the descriptions generated from the BLIP-2 and ZInD-Agent.

Table 7. **Comparative Descriptions from BLIP-2 and ZInD-Agent** based on input floor-level panorama images as illustrated in Figure 11. The Ground-Truth (GT) description is presented in the final row. For each generated description (e.g., **BLIP-2**), the corresponding evaluation metrics (BLEU-2, BLEU-4, METEOR, CIDEr, ROUGE$_\text{L}$, and Embedding Similarity) are listed in the last column, arranged sequentially.

Method	Generated Description	Scores
BLIP-2	a camera shot of a long hallway in a house. a view from a fish eye lens of a hallway. the bottom half of a large empty room. this is a panoramic view of the hallway. a white sink sitting next to a toilet next to a mirror. a panorama of a hallway and living room. a hallway view looking into an empty house. a house with wood floors, a staircase, and a door. a room that has hardwood floors in it. a large panoramic picture showing an empty room. a kitchen with lots of white appliances in it. the outside of an empty living room. an empty empty room with a white wall and door. an empty white room with a wooden floor and rack. there is a door and some furniture in the room. an empty hallway with a kitchen and dining area. a fisheye lens view of an empty white room. a big bathroom with a big mirror above the sink. the panorama view of the living room of the apartment. a bathroom with a toilet and a light fixture. a panoramic view of the inside of a bedroom. there is a kitchen, living room, bathroom and closet in this picture. a bathroom with a tub, shower, toilet and sink. the living room has a large wood floor. a view of an empty living room and empty kitchen. an empty room with a wooden floor and white walls	11.63 2.81 16.58 19.31 17.10 0.53
ZInD-Agent	This elegant three-story residence begins with a welcoming foyer on the first floor, introducing you to a world of comfort and style. The large dining room, notable for its absence of windows, creates a focused and cozy atmosphere for memorable dinners. It's complemented by very large doors, enhancing the sense of openness and connectivity to the surrounding areas. Nearby, a very large room centered around a fireplace offers a perfect spot for relaxation, especially during cooler evenings. Practicality is not forgotten, with a medium-sized bathroom conveniently located on this floor. Additionally, a large bedroom provides a private and serene retreat for rest. Moving to the second floor, you'll find a diverse collection of spaces designed for leisure and enjoyment. A small, bright yard area, illuminated by very large windows, offers an inviting touch of the outdoors. The floor boasts a very large balcony, ideal for outdoor relaxation and enjoying views. There are two living rooms: one very large, perfect for family gatherings and entertainment, and a smaller one for more intimate moments. A very small laundry room adds everyday practicality to this floor. The presence of a small storage room meets additional storage needs. The highlight here is the very large pool area, promising endless fun and a perfect spot for relaxation. A small foyer elegantly connects these diverse areas. The third floor combines private and functional spaces. A medium-sized hallway links various rooms, including a large storage area and a very large pantry, providing ample storage solutions. The medium-sized kitchen is perfectly suited for culinary adventures. There's also a very small balcony, offering a private space for solitude and reflection. The floor includes a small foyer and stair area, ensuring efficient flow between rooms. A medium yard area, playroom, and a very large garage add to the floor's functionality. The small basement offers versatile space, adaptable for various needs. This floor also features a very small living room and a medium-sized fireplace room, each designed for comfort and coziness, and a very small dining room for intimate meals. Throughout the home, large windows and doors in every room ensure that natural light enhances the ambiance, making the spaces feel open and airy. This residence is designed to cater to a variety of lifestyles, whether it's hosting guests or enjoying quiet, family moments, blending elegance with practicality in a harmonious balance.	25.99 9.25 27.08 30.23 32.48 0.73
GT	Welcome to this magnificent property that promises a blend of comfort and luxury. As you step into the property, you are greeted by a very large foyer, the entry Welcome to this magnificent property that promises a blend of comfort and luxury. As you step into the property, you are greeted by a very large foyer, the entry point of the house. The foyer is spacious and welcoming, with three very large doors leading to different parts of the house. To the left, a door opens into a medium-sized closet, perfect for storing your coats and shoes. To the right, another door leads to a very large garage, spacious enough to accommodate your vehicles and provide ample storage space. Moving up to the second floor, you will find a medium-sized stairway that leads to a very large living room. The living room is a sight to behold with two very large windows that let in an abundance of natural light. It also has a very large door that opens to the rest of the house. Adjacent to the living room is a medium-sized hallway that leads to a medium-sized bathroom with a very large door. The hallway also opens up to a very large kitchen, perfect for those who love to cook. Next to the kitchen, you will find a medium-sized closet that leads to a very large dining room. The dining room is a perfect place for family meals, with a very large window that provides a beautiful view while you dine. The third floor of the house is where you will find the bedrooms. A large hallway connects all the rooms on this floor. The hallway has four medium-sized doors leading to a medium-sized laundry room, a large bathroom, a small bedroom, and a very large bedroom. The very large bedroom is a luxurious retreat with four windows of varying sizes, providing a panoramic view of the surroundings. The third floor also houses another very large bedroom with three large windows, a large closet, and a very large bathroom. There is also a small bedroom and a small closet on this floor. This property is a perfect blend of luxury and comfort, with its spacious rooms, large windows, and well-planned layout. It promises a comfortable and luxurious living experience. of the house. The foyer is spacious and welcoming, with three very large doors leading to different parts of the house.	

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, 4

[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] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*, 2020. 2

[13] 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

[14] 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

[15] 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

[16] 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, 3

[17] 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

[18] 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

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

[20] 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

[21] 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

[22] Jianmo Ni, Gustavo Hernández Ábrego, Noah Constant, Ji Ma, Keith B Hall, Daniel Cer, and Yinfei Yang. Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models. *arXiv preprint arXiv:2108.08877*, 2021. 2

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

[24] 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

[25] 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, 3

[26] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems*, 33:16857–16867, 2020. 2

[27] Cheng Sun, Chi-Wei Hsiao, Min Sun, and Hwann-Tzong Chen. Horizonnet: Learning room layout with 1d representation 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, 3

[28] 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

[29] 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

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

[31] 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

[32] 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