A Supplementary Material

In this section we provide further implementation details for VLN-BERT (Section A.1) and additional qualitative analysis of performance (Section A.2) and the pretraining curriculum (Section A.3). Code will be provided to reproduce the results of all experiments.

A.1 Implementation Details

The experiments described in Section 5 utilize a 12-layer $\text{BERT}_{\text{BASE}}$ [7] architecture for both the vision and language streams in the model. Following [18], we use 6 cross-modal attention layers to connect the two streams. To operationalize the language-only pretraining stage (Stage 1), VLN-BERT is initialized with BERT weights that result from pretraining on English Wikipedia and BooksCorpus [34]. Similarly, for the visual grounding stage (Stage 2), VLN-BERT is initialized with ViLBERT weights, which result from pretraining on the Conceptual Captions [24] dataset. For action grounding pretraining and path-selection finetuning (Stage 3) we use path-instruction pairs from the Room-to-Room [4] dataset. During this stage, models are trained with the Adam optimizer with a learning rate of 4e-5 and a batch size of 64. We use a learning rate schedule with a linear warmup and cooldown. We train for 50 epochs in pretraining and 20 epochs in finetuning. During finetuning, we utilize early stopping based on the success rate on the unseen split of the validation set. Using 8 Titan X GPUs Stage 3 of training takes approximately 66 hours.

A.2 Qualitative Examples of Success and Failure

This section provides qualitative examples of the path selection performance of VLN-BERT using the full training curriculum described in Section 4.3. To gain insight at the region-level, we estimate region importance using the gradient-based visualization technique described in Section 5.4 (i.e. importance is calculated as the sum of the gradient of the model's output score with respect to the features for each region). In all examples, the model selects one path from a set of up to 30 candidate paths for a given set of instructions. The selected path is successful if it terminates within 3m of the goal location.

Three examples of successful path selection are illustrated in Figure 6. In the first example, VLN-BERT selects a path that does not initially follow the ground truth path, but correctly stops at the goal location. In this example, the model accurately grounds the phrase *'antelope head'*, which does not appear in the VLN [4] training dataset (the term *'antelope'* appears 3 times). In the second and third examples, the selected paths closely match the ground truth, and the top 5 regions include key objects mentioned in the instructions – *'freezers'* and *'statue'*. The term *'freezers'* (with and without the *'s'*) does not occur the VLN training dataset, and *'statue'* appears 105 times. These examples suggest that VLN-BERT is able to transfer visual grounding learned on the image-text pairs in the Conceptual Captions [24] dataset to the embodied task of path selection.

Three unsuccessful examples are shown in Figure 7. In each example the characteristics of the errors are qualitatively different. In the first row, VLN-BERT selects a path that does not stop at the correct goal location. However, the top 5 regions include key visual landmarks mentioned in the instruction (e.g.*'treadmill'* and *'sofa'*). In contrast, in the second row, VLN-BERT selects a path that does not reach the goal bedroom. In this instance, the top 5 regions include *'curtains'* from a different location, which may have led to this particular error. In the last row every aspect of the selected path seems mismatched from the instructions: the path goes to the left of the table not right and objects mentioned in the instructions are missing from the top 5 regions.

A.3 Qualitative Analysis of the Pretraining Curriculum

In section Section 5.4 we demonstrated that the visual grounding pretraining stage (Stage 2) quantitatively improves performance. In Figure 8 we qualitatively compare the visual grounding that is learned with and without stage 2 of pretraining. In the first example, the model trained without stage 2 of pretraining (right) fails to ground the phrase *'mini fridge'*, which only occurs 1 time in the VLN training dataset. Similarly, in the second example, without stage 2 pretraining, the model fails to ground the phrase *'massage table'* (29 occurrences in the VLN training dataset). In both cases, the model without stage 2 pretraining selects an unsuccessful path. In contrast, when trained with the full curriculum, VLN-BERT correctly grounds these key phrases and selects successful paths for these two examples.

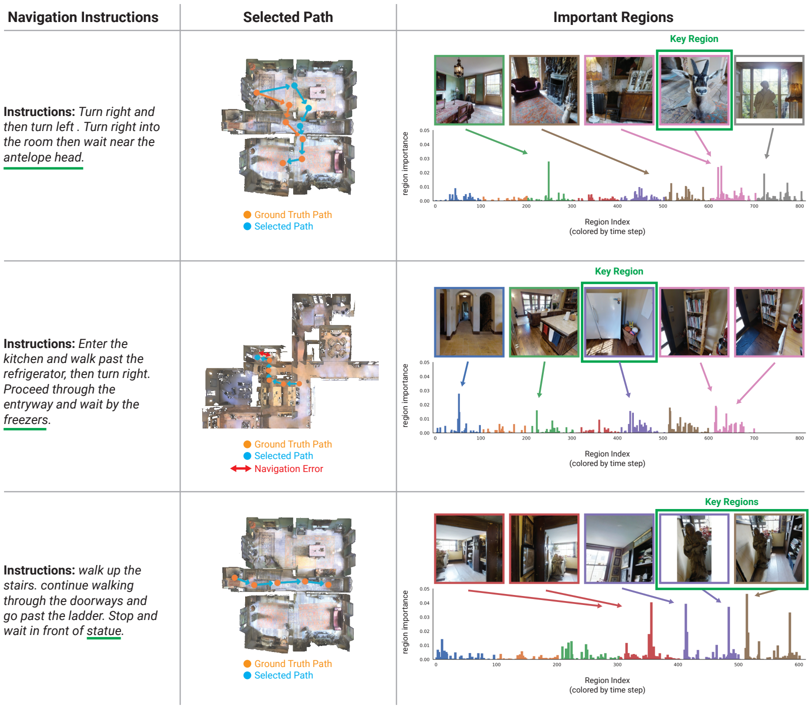
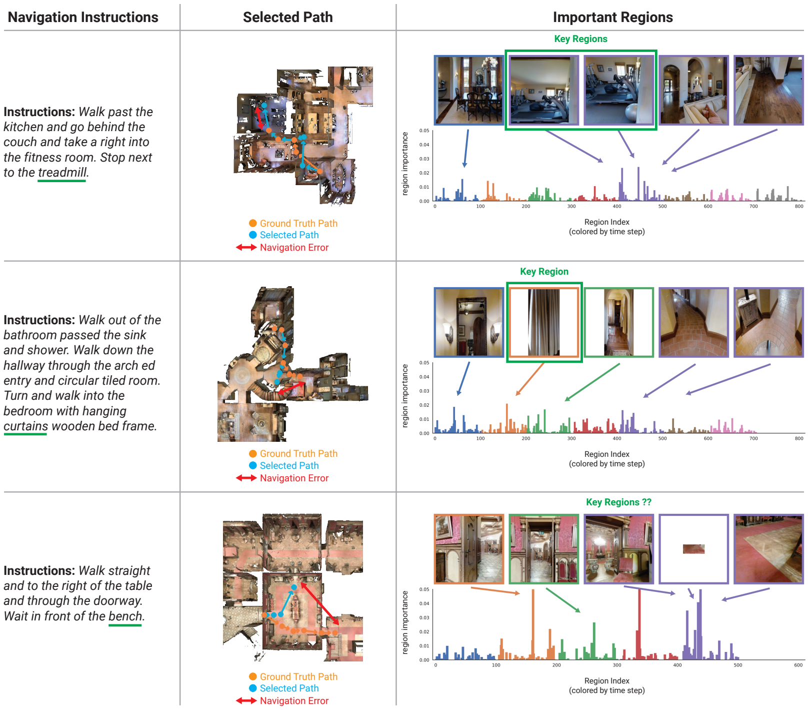


Fig. 6. Examples of successful paths selected by VLN-BERT (middle – blue), with ground truth paths (middle – orange) and navigation errors (middle – red) provided for comparison. The right column illustrates the top 5 regions that influence the model's predictions – importance is determined by taking the gradient of the score for a path-instruction pair with respect to the input region features (as in Section 5.4). A qualitative assessment of accurate visiolinguistic grounding (in green) highlights phrases that rarely occur in VLN training dataset: *'antelope head'* (0 occurrences), *'antelope'* (3 occurrences), *'freezer(s)'* (0 occurrences), *'statue'* (105 occurrences). These results suggest that VLN-BERT has effectively learned to transfer grounding from image-text pairs from the web to the embodied task of VLN.

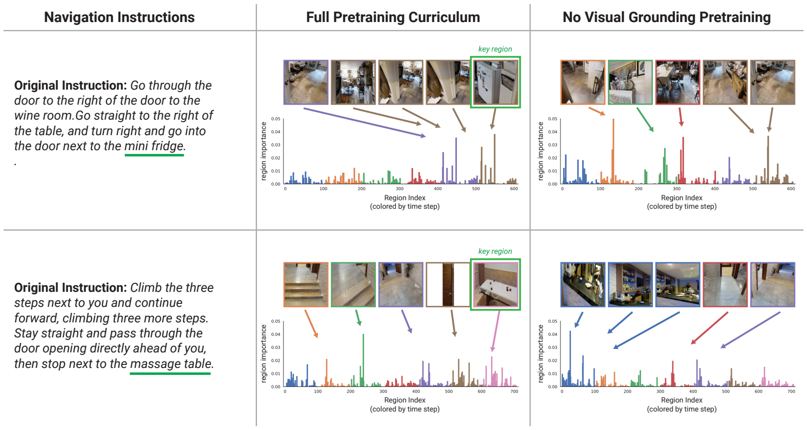


Instructions: *Walk past the kitchen and go behind the couch and take a right into the fitness room. Stop next to the treadmill.*

Instructions: *Walk out of the bathroom passed the sink and shower. Walk down the hallway through the arch ed entry and circular tiled room. Turn and walk into the bedroom with hanging curtains wooden bed frame.*

Instructions: *Walk straight and to the right of the table and through the doorway. Wait in front of the bench.*

Fig. 7. Examples of unsuccessful paths selected by VLN-BERT (middle – blue), with ground truth paths (middle – orange) and navigation errors (middle – red) provided for comparison. The right column illustrates the top 5 regions that influence the model's predictions – importance is determined by taking the gradient of the score for a path-instruction pair with respect to the input region features (as in Section 5.4). In the first row, the selected path goes past the goal location, but the visual grounding appears accurate. In the second row, the selected path does not enter the correct bedroom and the visual grounding appears to identify the wrong curtains. In the final example the model fails completely – selecting a path going to the left (not right) of the table.



Original Instruction: *Go through the door to the right of the door to the wine room.Go straight to the right of the table, and turn right and go into the door next to the mini fridge.*

.

Original Instruction: *Climb the three steps next to you and continue forward, climbing three more steps. Stay straight and pass through the door opening directly ahead of you, then stop next to the massage table.*

Fig. 8. Comparison of pretraining with the proposed curriculum (middle) vs. omitting the visual grounding stage (right). Region importance histograms are estimated as in Section 5.4. The top 5 regions correspond with the successful path selected by VLN-BERT when pretrained with the full curriculum. Without the visual grounding pretraining stage (right), the model fails to ground phrases that rarely occur in the VLN training dataset (e.g. *'mini fridge'* (1 occurrence) and *'massage table'* (29 occurrences)). Furthermore, in both cases without the visual grounding pretraining stage, VLN-BERT selects an unsuccessful path (not illustrated) for the given instructions.