

## 446 A Language-Table Simulator Experiments

447 We describe additional details left out of the main text on the Language-Table simulator.

### 448 A.1 Model Details

449 For the policy, we use the pretrained FiLM-conditioned ResNet architecture that was trained using  
450 behavior cloning provided by the Language-Table repository [31]. We do not use Language-Table’s  
451 LAVA model as a pretrained model was not provided and requires 64 TPUv3 chips to train.

### 452 A.2 Additional Details

453 In this section, we describe how the cumulative cost plot in Figure 3 was generated. Since we  
454 evaluated over three seeds and each experiment has a different cost, we create 50 bins at equal  
455 intervals from 0 to the max overall cost across all seeds, then aggregate the cumulative absolute SPL  
456 error and cumulative cost. Using this binning approach, we also compute the standard deviation of  
457 the error bounds.

### 458 A.3 Additional Results

459 **Contrast sets allow for more evaluations with less cost.** As depicted in Figure 6, the slopes of  
460 each type of perturbations determines how the cost scales compared to the number of evaluations.  
461 Limited interventions is clearly the lowest cost; however, we had found that it does not estimate the  
462 evaluation set. All contrast set strategies have a higher slope than that of standard evaluation. For  
463 example, scene and language perturbations can execute nearly double the number of experiments  
464 compared to the standard evaluation given a cost budget of 281.

## 465 B VLN-CE Evaluation on a Physical Robot

466 We use a Locobot [33] robot to run vision-and-language navigation in continuous environments [32]  
467 (VLN-CE) in the real world.

### 468 B.1 Model Details

469 We pretrain a policy for the robot on the VLN-CE task in the Habitat simulator using the RxR  
470 training set [36]. We then use a behavior cloning objective to finetune the simulation-trained model  
471 on a small set of real world examples using teacher-forcing. The policy uses a discrete action

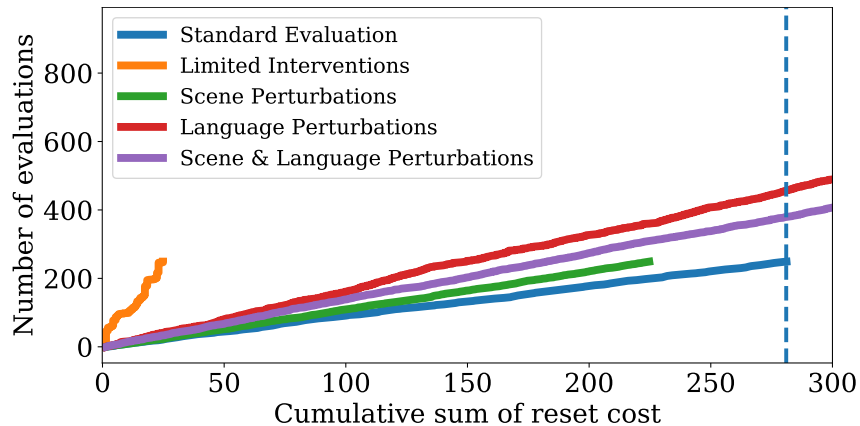


Figure 6: Compared to Figure 3, we separate the relationship between cost and error. Limited interventions and language-only perturbations allows for more evaluations with less cost, and standard evaluation has the least number of evaluations for the cost. As described in the main text, scene and language perturbations finds a good middle ground with more evaluations for less cost.

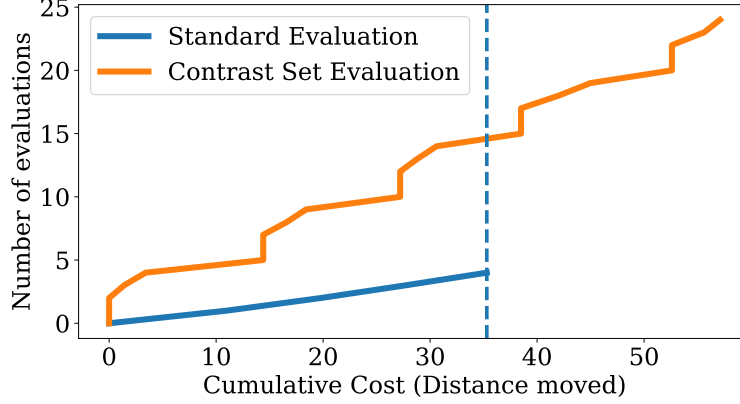


Figure 8: We separate the relationship between cost and error in the real word VLN-CE experiments. We note that the number of evaluations performed in the standard evaluation is relatively low compared to the contrast set evaluation. Contrast set evaluation allows the experimenter to execute more experiments compared to standard evaluation. Though not every single experiment from the test set can be executed under a cost budget of around 35 (blue dotted line), Figure 5 indicates that contrast set evaluation still estimates the test set.

space of forward, rotate left by 30 degrees, rotate right by 30 degrees, and stop. Only one scene arrangement was used in the training dataset, and this scene was not used during testing. We note that the furniture, especially larger items such as beds and couches, were used during training and existed during training. However, the scene arrangements, which is key to the task of VLN-CE, was ensured to be different.

## B.2 Experiment Design

We describe how we collected our test instances. We design a pseudo studio apartment environment which is populated with furniture similar to those found in simulation. To ensure ecological validity of test instances, we recruited five participants to design five furniture setups. They were instructed to ensure that the furniture was arranged in any way they would prefer, defining the scene  $s$ . They then placed the robot and walked a trajectory  $b$  they wanted the robot to execute while narrating a natural language command  $l$ . A subset of the navigation instructions can be found in Figure 4. By using external participants to design our test instances, we hope to ensure that we, as experimenters, do not bias the collection of our test instances to be easier than expected.

## B.3 Additional Results

**Contrast sets allow for more evaluations with less cost.** As depicted in Figure 8, the slopes of each type of perturbations determines how the cost scales compared to the number of evaluations. Though contrast set evaluation has a higher bound, given a cost budget of 35, contrast sets allow a user to run nearly triple the number of trials for the same cost budget.

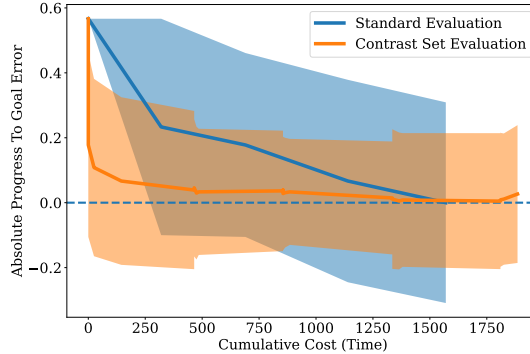


Figure 7: Average cumulative progress to goal vs cumulative cost as *time to perturb the scene in seconds*. We find that the results found in Figure 3 and Section 6.2 when using time as the cost function instead of the distance objects were moved still hold when switching to time to perturb the scene as the cost. This shows that we may be able to use various cost metrics to measure experimenter effort.

504 **Contrast sets also estimate the full test set while minimizing time to reset the scene.** Instead of  
505 using distance of objects moved during a scene reset as we did in the main text, we also investigate  
506 the time used to reset the scene as a cost metric. We find similar results in Figure 7 which uses time  
507 as cost as we did in Figure 5 which uses distance of objects moved as cost. This is likely due to the  
508 nearly-linear relationship between time it takes to move items in the scene and the distance they are  
509 moved.