7 Supplementary

7.1 VLN↻BERT: Validation Performance

We use the VLN↻BERT model in VLN experiments to evaluate the impact of dataset differences between VLN and VLN-CE (Sec. 5.1) and MP3D vs. reconstructed vision (Sec. 5.2). For consistency across our experiments, we retrain VLN↻BERT using the official codebase[5]. We train with the (init. PREVALENT) backbone. Our re-trained version performs at 2 SR and 1 SPL lower in Val-Unseen than the published result in [16] but matches performance in Val-Seen SR and SPL. We repeat training and evaluation with 3 different random seeds and find performance consistent with what we present in Tab. 1 and Tab. 2.

7.2 Oracle Policy Detail

We use the Oracle Policy with known subgoal candidates that are specified in 3D coordinates. We note that using this oracle with 2D subgoal predictions requires projecting the target location from 2D to 3D. We define a projection procedure $P : (r, \theta) \rightarrow (x, y, z)$ that maps distance and relative heading to global 3D coordinates. In this procedure, the agent's current pose is used to project (r, θ) to global 2D coordinates $(\hat{x}, \hat{z})$. We assume the target exists at the elevation of the agent's pose, $\hat{y}$. Finally, we snap the resulting 3D coordinates $(\hat{x}, \hat{y}, \hat{z})$ to the nearest position (x, y, z) that exists on the navigation mesh and that has a finite geodesic distance from the agent's current position.

7.3 Subgoal Module Ablations

In Tab. 6 we evaluate ablations of the subgoal generation module (SGM) in VLN-CE. In row 1 vs. 2, we find that training the SGM with Habitat-rendered vision (Recon.) leads to better downstream performance in VLN-CE than MP3D panoramas across all splits (3 SR Val-Unseen, 5 SR Val-Seen, 6 SR Train*). We further ablate the 360° laser scan to 270° and observe an additional performance drop of 5 SR in Val-Unseen, 4 SR in Val-Seen, and 8 SR in Train*. Altogether, our modifications of reconstructed vision and 360° scanning improve performance under the SGM by 8 SR in Val-Unseen over the SGM proposed in [2].

7.4 Distribution of Navigation Errors

In Tab. 7 we present the navigation errors for both the Oracle Policy and Local Policy when used to perform the VLN-CE task. The presented short-range navigation errors were collected during the experiments reported in Tab. 3. We characterize the distribution of navigation errors by the percent of navigations that result in various thresholds of error. The Oracle Policy rarely produces a high navigation error – just 0.53% of navigations result in at least a 0.20m error

[5] github.com/YicongHong/Recurrent-VLN-BERT

Table 6. Ablations against the subgoal generation module (SGM). Results are in VLN-CE with reconstruction-trained VLN↻BERT and Local Policy navigator. Row 1 matches Tab. 4 row 3 in the main paper. Row 2 ablates training with reconstructed vision (Recon.) and row 3 ablates both Recon. and 360° laser scan to match the SGM used by Anderson et al. [2].

Task: VLN-CE				Train*					Val-Seen					Val-Unseen				
#	Subgoal Candidates	Vision	HFOV	TL ↓	NE ↓	OS ↑	SR ↑	SPL ↑	TL ↓	NE ↓	OS ↑	SR ↑	SPL ↑	TL ↓	NE ↓	OS ↑	SR ↑	SPL ↑
1		Recon.	360	13.12	3.54	71	65	55	12.69	4.51	60	51	44	13.74	5.83	51	41	35
2	SGM	MP3D	360	10.08	3.52	64	59	54	10.41	4.80	52	46	41	10.07	5.61	45	38	34
3		MP3D	270	11.24	4.14	59	51	46	10.73	5.02	47	42	37	10.62	5.97	41	33	29

Table 7. Short-range navigation errors (subgoal-to-subgoal) of navigation policies. Evaluated while VLN↻BERT is performing the VLN-CE task with nav-graph subgoals (Tab. 3). Navigation errors are in meters and thresholded error values (>Xm) are reported as a percent of all short-range navigations.

Task: VLN-CE		Train*				Val-Seen				Val-Unseen			
#	Navigator	NE ↓	>0.2m	>0.5m	>1.0m	NE ↓	>0.2m	>0.5m	>1.0m	NE ↓	>0.2m	>0.5m	>1.0m
1	Oracle Policy	0.09	0.24	0.24	0.24	0.11	0.74	0.69	0.69	0.08	0.53	0.42	0.29
2	Local Policy	0.40	9.44	7.60	6.31	0.26	7.09	5.63	4.84	0.40	8.04	6.36	5.73

in Val-Unseen. In the same setting, the Local Policy has a 0.20m failure rate of 8.04%. This extends to even larger failure thresholds where the Local Policy fails to navigate to within 1.0m of the subgoal 5.73% of the time in Val-Unseen. These failure rates provide context to Tab. 3 which demonstrated a performance drop when navigating with the Local Policy.

7.5 Component Analysis of the Local Policy

The Local Policy consists of two components: an FMM path planner that outputs nearby position coordinates 0.25m away from the agent and an action decoder that maps these coordinates to the VLN-CE action space. To isolate performance impacts of these components, we repeat the Tab. 3 row 3 evaluation but replace VLN-CE actions with teleportation to the nearby coordinates. We found that this precise navigation results in a 1 SR increase and matching SPL (56 and 47, respectively) as compared to row 3. This teleportation reduced errors in navigating to subgoals by 4.3% for errors >0.20m, which suggests that an improved alignment between planning output and action space could slightly benefit performance.

7.6 Frontier-based Subgoal Generation

To enable navigation throughout a scene, subgoal candidates must be predicted near openings between occupied areas (e.g. hallways connecting rooms or open space between large furniture). We experiment with a deterministic method of subgoal generation inspired by this requirement. Specifically, we adopt wavefront frontier detection (WFD), a method for frontier-based exploration (FBE). We use WFD to predict the location of frontier points from a 2D radial occupancy map generated from the laser range scan observation. We then cluster the identified frontier points and treat their associated centroids as subgoal candidates. Evaluating in Val-Unseen as per the Sec. 5.4 setup with no fine-tuning, the model achieves 26 SR (off from 41 SR with SGM). We suspect this performance drop relates to a shift in the distribution of candidates away from the MP3D nav-graph. Additionally, the WFD method fails to predict central subgoals in free-space regions. Such subgoals are essential for approaching goals and obtaining views useful for high-level decision-making.