Supplementary Material: Towards Learning a Generic Agent for Vision-and-Language Navigation via Pre-training

Summary of Contributions. Weituo implemented the algorithm, made the model work, and ran all experiments. Chunyuan initiated the idea of pre-training the first generic agent for VLN, led and completed the manuscript writing. Xiujun provided the codebase and helped implementation. Lawrence and Jianfeng edited the final manuscript.

A. Experiments

Three types of inputs on CVDN We illustrate the naming of three types of text inputs on CVDN in Table 6.

	V	t_0	A_i	Q_i	$Q_{1:i-1}\&A_{1:i-1}$
Oracle Answer	✓	✓	✓		
Navigation QA	✓	✓	✓	✓	
All	✓	✓	✓	✓	✓

Table 6: Three types of inputs on CVDN. t_0 is the target object, V is the ResNet feature. Q_i and A_i are the question and answers in the i-th turn. $Q_{1:i-1}\&A_{1:i-1}$ are the question & answer pairs before the i-th turn.

Ablation Study Results on HANNA Table 7 shows the results with different pre-training objectives. We see that the $\mathcal{L}_{\text{PA}} + \mathcal{L}_{\text{MLM}}$ yields the best performance among all variants.

	SEEN-ENV				UNSEEN-ALL			
Agent	SR ↑	SPL ↑	NE ↓	#R ↓	SR ↑	SPL ↑	NE ↓	#R ↓
PREVALENT ($\mathcal{L}_{\text{PA}} + \mathcal{L}_{\text{MLM}}$)	**83.82**	**59.38**	**1.47**	**3.4**	**52.91**	**28.72**	**5.29**	**6.6**
PREVALENT ($\mathcal{L}_{\text{MLM}}$)	78.75	54.68	1.82	4.3	44.29	24.27	6.33	8.1
BERT (feature-based)	57.54	34.33	4.71	3.9	24.12	11.50	9.55	11.3
BERT (fine-tuning)	80.75	57.46	1.97	4.0	26.36	12.66	9.1	8.3

Table 7: Ablation study of pre-training objectives on test splits of HANNA.

B. Comparison with Related Work

Comparison with PRESS. The differences are summarized in Table 8 (a). Empirically, we show that (1) incorporating visual and action information into pre-training can improve navigation performance; (2) Pre-training can generalize across different new navigation tasks.

Comparison with vision-language pre-training (VLP). The differences are in Table 8 (b). Though the proposed methodology generally follows self supervised learning such as VLP or BERT, our research scope and problem setups are different, which renders existing pre-models are not readily applicable.

	Prevalent (Proposed)	**Press**
Dataset	Augmented R2R dataset	Generic language
Modality	Vision-language-action triplets	Language
Learning	Train from scratch	Off-the-shelf (BERT)
Downstream	Three navigation tasks	R2R

(a) PRESS

	Prevalent (Proposed)	**VLP**
Visual Input	Panoramic views (Size: 36 × 640 × 480)	Single image (Size: 640 ×480)
Visual Features	ResNet (View-level)	Fast RCNN (Object-level)
Objectives	Attentive MLM & Action Prediction	Masking on VL & Same-Pair Prediction
Downstream	RL: Navigation in sequential decision-making environments	Single-step prediction

(b) VLP

Table 8: Comparison with related works.