

# Supplementary Materials: Narrowing the Gap between Vision and Action in Navigation

Anonymous Authors

## 1 BASELINES

### 1.1 VLN-BERT

VLN-BERT[4] is a cross-modal Transformer-based navigation agent with an extra recurrent state unit. At each navigation step  $t$ , the agent takes text representation  $X$ , vision representation  $V$ , and state representation (denoted as  $S_t$ ) as input. The state representation is initialized with [CLS] text tokens and updated based on the text and vision representations and the previous steps' history information. The model architecture consists of cross-modal Transformer layers with a self-attention layer to learn cross-modal representations.

$$\hat{X}, \hat{S}_t, \hat{V}_t = \text{Cross\_Attn}(X, [S_t; V_t])$$
$$S_{t+1}, a_t = \text{Self\_Attn}(\hat{S}_t, \hat{V}_t),$$

where  $\hat{S}_t$ ,  $\hat{S}_t$ , and  $\hat{V}_t$  are text, vision, and recurrent state representations after cross-modal Transformation layers.  $a_t$  is the action probability.

### 1.2 HAMT

Compared to VLN-BERT, HAMT [2] memorizes history information more explicitly. It uses a sequence of panorama images as the navigation history during the navigation trajectory; then it applies Transformers to encode the observations on the trajectory to memorize history information. Formally, given the encoded history observation representation  $v_i$ , the output of the temporal encoder is  $h_i = \text{LN}(W_t v_i) + \text{LN}(W_a a_i) + E_i^S + E_2^T$ , where  $a_i$  is the action embedding at step, and  $E_2^T$  is the token type encoding which indicates the input is history view. In the end, HAMT concatenates history and observation as the vision modality and uses a cross-modal transformer to predict actions by selecting the highest similarity score between observation encoding  $o_i$  and [CLS] token, which contains instruction-trajectory information. Each view observation  $o_i$  in a panorama can be obtained from the following equation:

$$o_i = \text{LN}(W_{rgb} v_i^{rgb}) + \text{LN}(W_d v_i^d) + \text{LN}(W_a v_i^a) + E_{o_i}^N + E_1^T, \quad (1)$$

where  $W_v$ ,  $W_d$ , and  $W_a$  are leaned weights.  $v^a$  is the relative angle that can be represented as  $v_i^a = (\sin \theta_i, \cos \theta_i, \sin \phi_i, \cos \phi_i)$ , where  $\theta$  and  $\phi$  are relative headings and elevation angles to the agent's current orientation.  $E_{o_i}^N$  is the navigable embedding to differentiate types of views, and  $E_1^T$  is the type embedding of observation. LN denotes layer normalization.

### 1.3 ETPNav

ETPNav [1] is a graph-based navigation agent capable of generating long-range navigation plans. It follows the graph design in DUET [3], a graph-based VLN-DE agent. ETPNav contains three modules: topological mapping, cross-modal planning, and offline control. The mapping module maintains a topo map for each episode.

The mapping module updates the topo map at each navigation step by incorporating current observations. After this, the planning module conducts cross-modal reasoning over the map and instruction to create a high-level topological path plan. The control module then executes the plan. It is noticed that the settings for the graph-based navigation agents are a bit different. Their action space extends globally, containing all observed nodes along the traversed path rather than being limited locally to only adjacent nodes. The graph-based agent commonly can obtain a higher success rate.

## REFERENCES

- [1] Dong An, Hanqing Wang, Wenguan Wang, Zun Wang, Yan Huang, Keji He, and Liang Wang. 2023. ETPNav: Evolving Topological Planning for Vision-Language Navigation in Continuous Environments. *arXiv preprint arXiv:2304.03047* (2023).
- [2] Shizhe Chen, Pierre-Louis Guhur, Cordelia Schmid, and Ivan Laptev. 2021. History aware multimodal transformer for vision-and-language navigation. *Advances in neural information processing systems* 34 (2021), 5834–5847.
- [3] Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. 2022. Think global, act local: Dual-scale graph transformer for vision-and-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16537–16547.
- [4] Yicong Hong, Qi Wu, Yuankai Qi, Cristian Rodriguez-Opazo, and Stephen Gould. 2020. A recurrent vision-and-language bert for navigation. *arXiv preprint arXiv:2011.13922* (2020).