Vision-language navigation (VLN) is the task of navigating an embodied agent to carry out natural language instructions inside real 3D environments. In this pap