Continual Vision-and-Language Navigation

SeongJun Jeong[1], Gi-Cheon Kang[2], Seongho Choi[3], Joochan Kim[4], Byoung-Tak Zhang[1,2*]

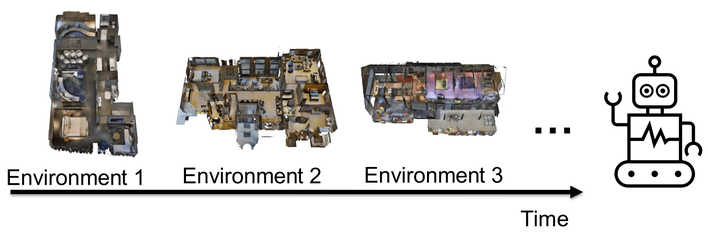
[1]Seoul National University, [2]AI Institute, Seoul National University, [3]SK Telecom, [4]KIST

BMVC 2025

1. Introduction

VLN agents are trained on a fixed set of environments

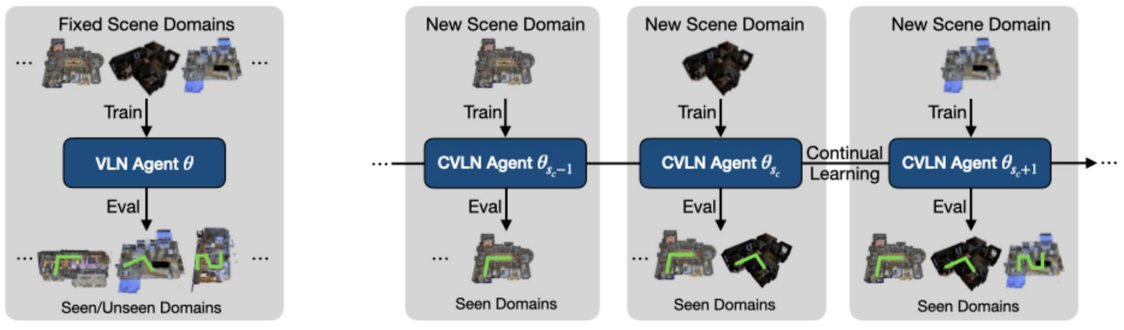
After deployment, agents face new, unseen environments



Real-world VLN agents require

continual learning without catastrophic forgetting.

2. Continual Vision-and-Language Navigation



(a) Vision-and-Language Navigation (b) Continual Vision-and-Language Navigation

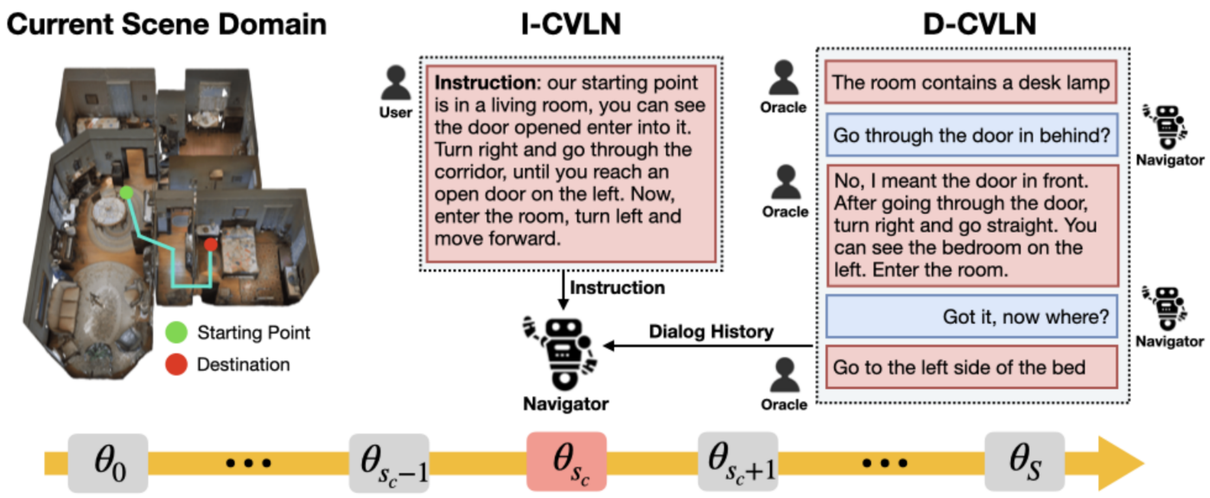
Vision-and-Language Navigation (VLN)

- Agent is trained once on a fixed dataset and tested on unseen scenes.

Continual Vision-and-Language Navigation (CVLN)

- Extends VLN to sequential learning across multiple scene domains.
- At each step, the agent learns a new domain while retaining performance on previously learned domains.
- Reflects real-world deployment: agents must adapt to new environments over time without forgetting past knowledge.

3. Benchmarks for CVLN

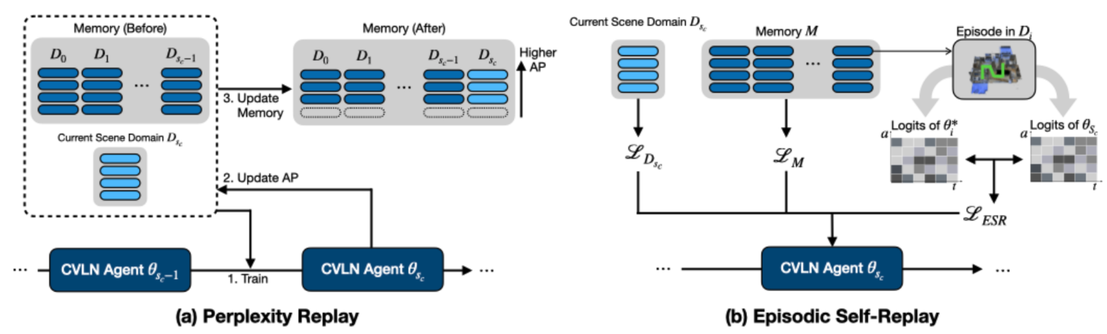


We introduce two CVLN benchmarks:

Initial-Instruction CVLN (I-CVLN) and **Dialogue-based CVLN (D-CVLN)**.

- **I-CVLN:** The agent receives a full navigation instruction at the beginning (built from R2R and RxR).
  - Includes **15,700 training** and **1,563 validation** episodes, grouped into **20 scene domains**.
- **D-CVLN:** The agent navigates using **dialogue history** between a follower and an oracle (based on CVDN).
  - Includes **3,737 training** and **382 validation** episodes, grouped into **11 scene domains**.
- Both benchmarks require agents to **learn scene domains sequentially**, enabling evaluation of continual learning capabilities.
- CVLN evaluates an agent's ability to **adapt to new domains while retaining past knowledge**.
- We report the **Average Metric** across scene domains, where performance on each domain is measured after the final stage of training.

4. Methods



- We adopt a **rehearsal-based CL** approach with a fixed replay memory to reduce catastrophic forgetting.
- After each domain, the memory is refreshed by replacing old episodes with new ones.

Perplexity Replay (PerpR)

- Uses **Action Perplexity (AP)** to measure episode difficulty.
- **Keeps high-AP (hard)** episodes and **removes low-AP (easy)** ones to store more informative samples.
- Trains on both current and replayed episodes to retain prior knowledge.

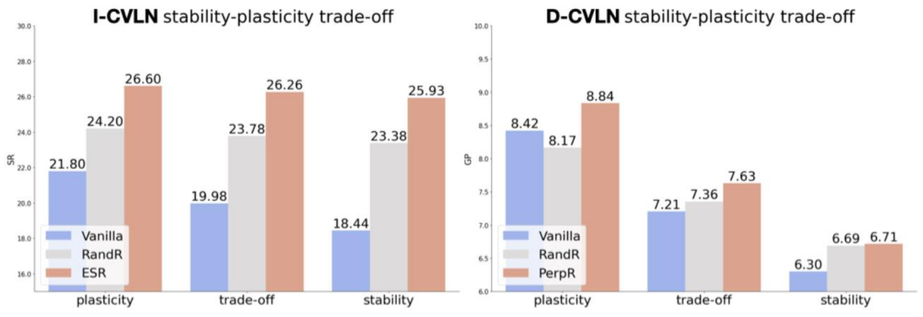
Episodic Self-Replay (ESR)

- Stores **step-wise action logits** for each episode.
- Replays past logits to **preserve previous action behavior** while learning new domains.
- Combines supervised replay with self-distillation for better knowledge retention.

5. Experiment

	Method	I-CVLN			D-CVLN
		AvgSPL↑	AvgSR↑	AvgNE↓	AvgGP↑
1.	Vanilla	23.1±0.8	25.6±0.6	10.5±0.5	5.5±0.4
2.	Joint	40.1±0.2	42.6±0.1	7.1±0.3	8.1±0.1
3.	L2	13.3±0.2	14.6±0.2	12.7±0.3	4.3±0.4
4.	AdapterCL	7.8±0.3	9.1±0.6	13.9±0.8	4.3±1.2
5.	AGEM	4.6±2.8	4.5±2.2	13.0±0.2	2.7±0.9
6.	RandR	24.9±0.6	28.0±0.7	10.0±0.5	5.7±0.3
7.	PerpR (Ours)	<u>26.1</u>±1.6	<u>28.9</u>±1.8	<u>9.8</u>±0.2	**6.1**±0.4
8.	ESR (Ours)	**28.2**±0.4	**31.9**±0.9	**9.2**±0.3	<u>5.8</u>±0.5

Method	I-CVLN				D-CVLN	
	Memory Size	AvgSPL↑	AvgSR↑	AvgNE↓	Memory Size	AvgGP↑
PerpR	200	23.8	26.3	10.2	50	4.9
	500	23.8	26.4	9.8	100	**6.2**
	1000	**28.7**	**31.1**	**8.7**	200	6.0
ESR	200	25.8	29.2	9.2	50	5.3
	500	27.7	31.1	**8.8**	100	5.5
	1000	**29.8**	**34.4**	9.1	200	**5.9**



Ablation	I-CVLN			D-CVLN
	SPL↑	SR↑	NE↓	GP↑
PerpR	**26.1**±1.6	**28.9**±1.8	**9.8**±0.2	**6.1**±0.4
PerpR-Rev.	24.6±1.1	27.1±1.2	10.2±0.3	5.4±0.2

Ablation	I-CVLN			D-CVLN
	AvgSPL↑	AvgSR↑	AvgNE↓	AvgGP↑
$\mathcal{L}_{ESR\_total}$	**27.7**	**31.1**	**8.8**	**5.5**
- $\mathcal{L}_M$	26.0	27.9	9.7	5.4
- $\mathcal{L}_{ESR}$	22.7	24.9	10.4	5.1