NOLO: Navigate Only Look Once

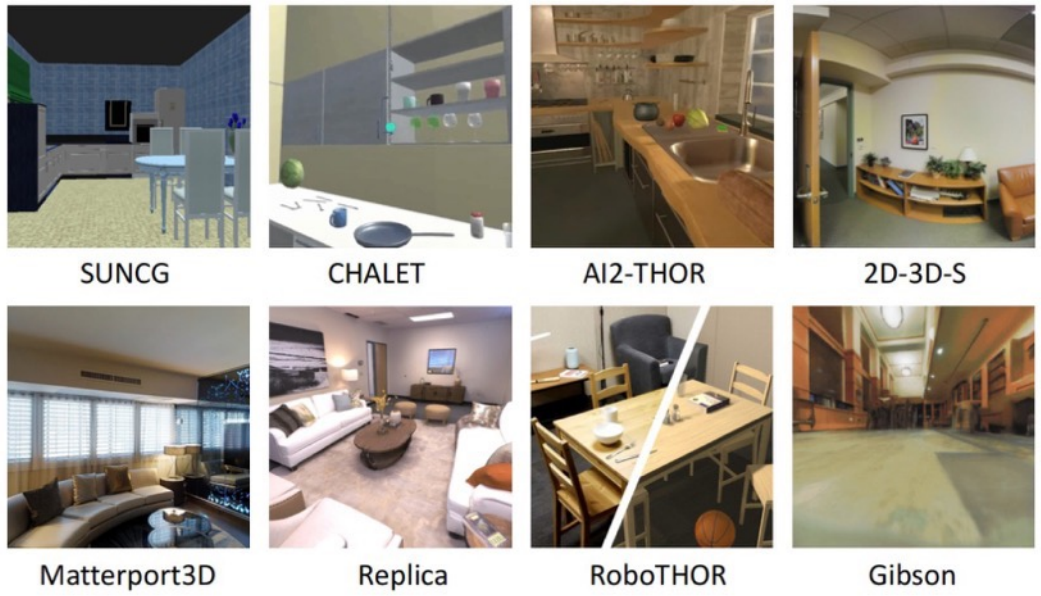
Bohan Zhou, Zhongbin Zhang, Jiangxing Wang, Zongqing Lu

10/23/2025

Motivation

北京大学 PEKING UNIVERSITY

[1]



[2]

Visual navigation has wide application in different fields like mobile robots and autonomous vehicles.

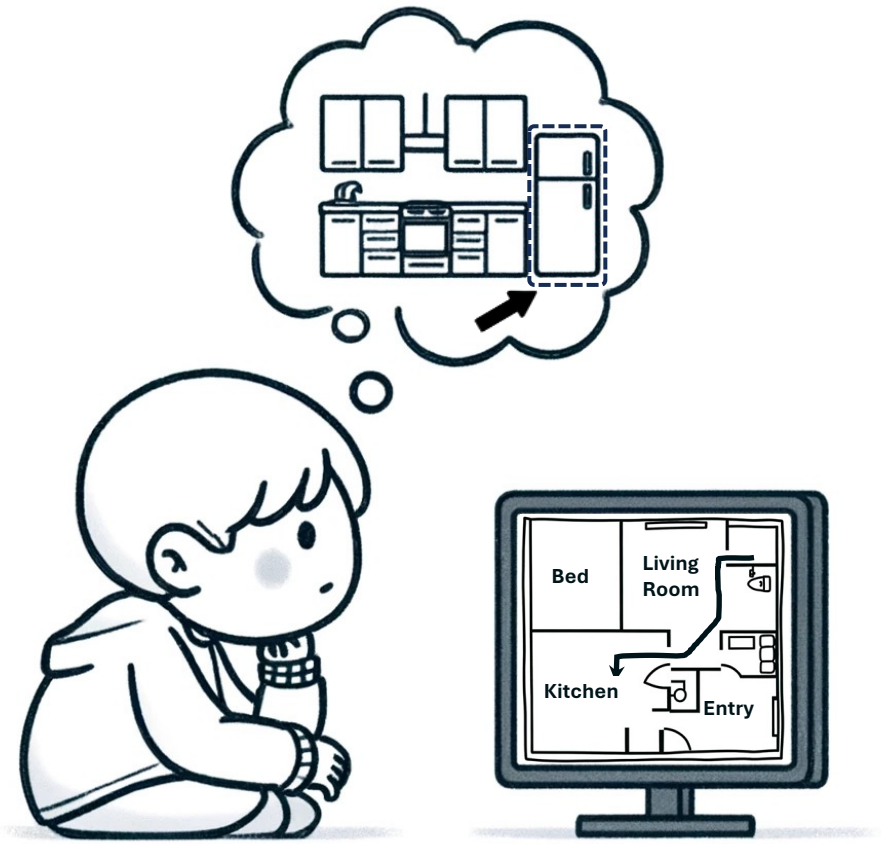
Existing approaches exhibit certain limitations:

- **Generalize poorly** or **require additional interactions** when transferred to a novel scene.
- Heavily rely on heterogeneous **sensors** like GPS+Compass, RGB-D cameras and IMU.

[1] Zhang T, Hu X, Xiao J, et al. A survey of visual navigation: From geometry to embodied AI. EAAI 2022.

[2] Shah D, Osiński B, Levine S. LM-Nav: Robotic navigation with large pre-trained models of language, vision, and action. CORL 2023.

Introduction



- **Intuition**: Humans can easily understand how to navigate in a novel room by watching videos.
- **Formulation**: **Video Navigation**, training an in-context policy to find objects occurred in the context video.

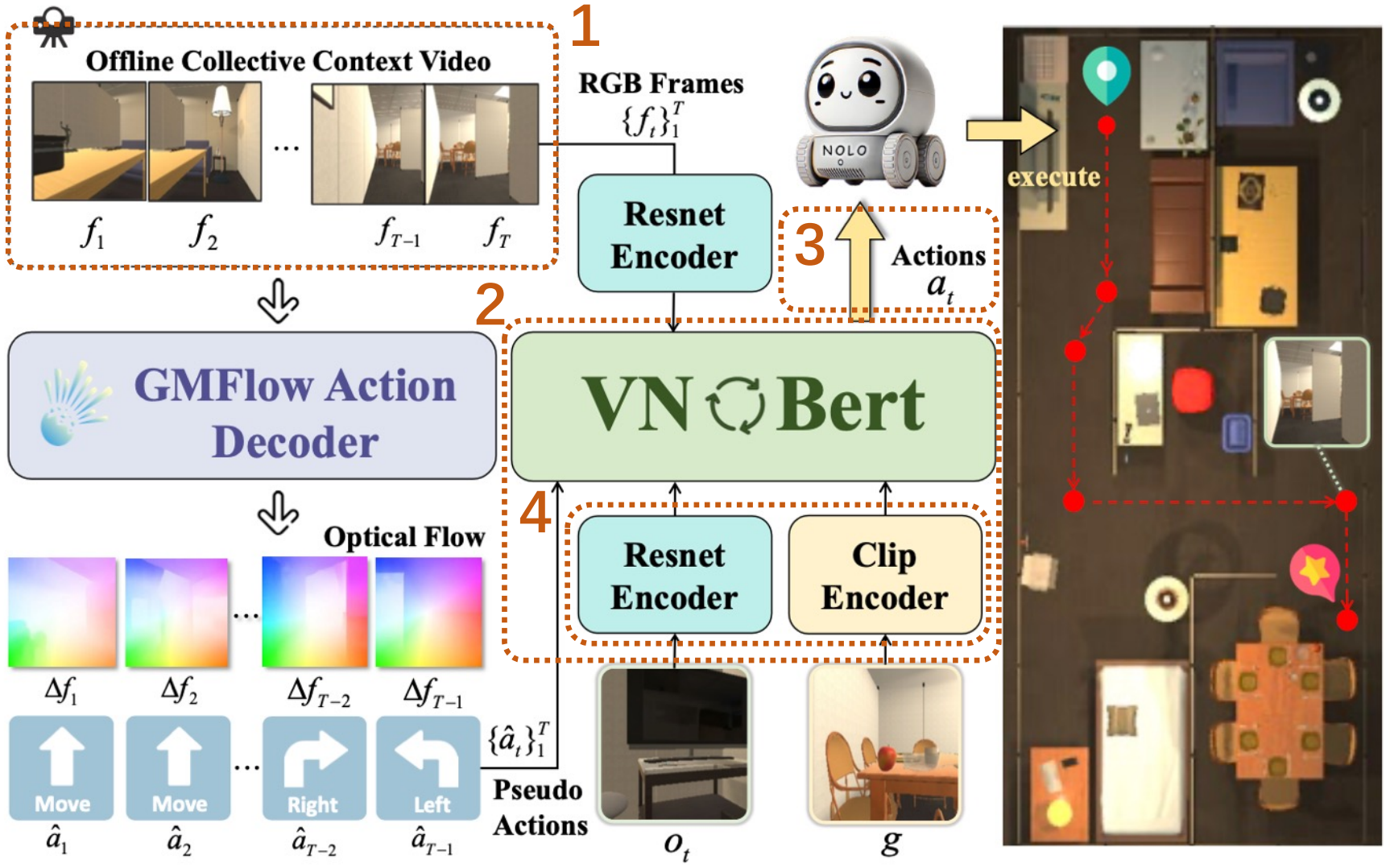
Challenges

- **Lack of actual actions**, spatial correlation, and visual understanding of complex scenes.
- **IL & IRL can not be applied** due to suboptimal video demonstrations with implicit goals.
- **Only egocentric observations**, no camera intrinsics, poses, maps, odometers, and depth inputs.

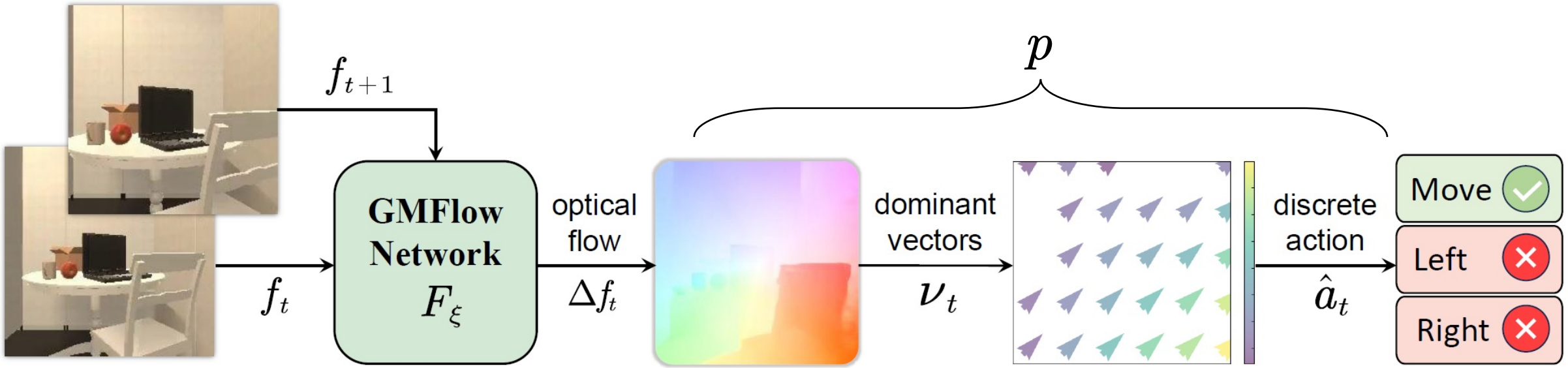
Methodology

北京大学 PEKING UNIVERSITY

1. Pseudo Action Labeling
2. In-Context Policy Modeling
3. Batch-Constrained Q-Learning
4. Temporal Coherence



Pseudo Action Labeling



$$\Delta f_t = F_\xi(f_t, f_{t+1})$$
$$\hat{a}_t = p(\Delta f_t).$$

Two adjacent frames are taken by a pretrained optical flow model to get a flow map. Some representative dominant vectors are filtered for action selection.

In-Context Policy Modeling

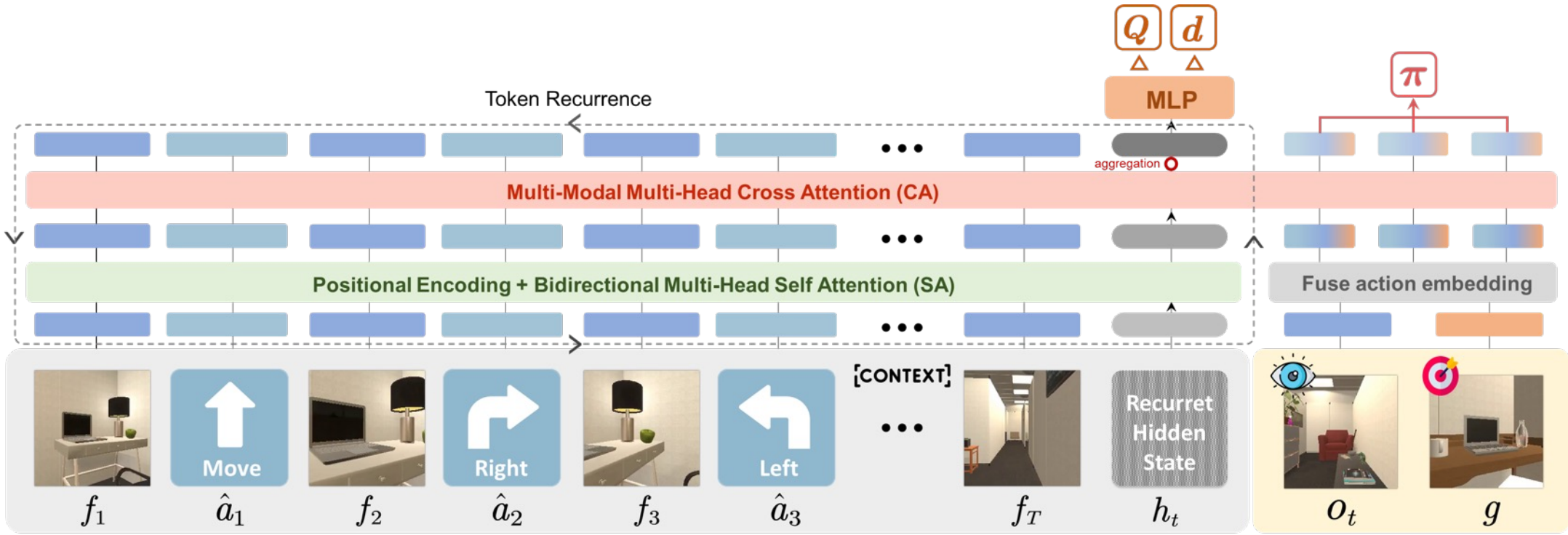


Figure 3: Structure of VN⟳Bert. At the initialization stage, a trajectory $\mathcal{T}$ and a zero-padded hidden state h_{init} are processed by a bidirectional multi-head self-attention module to obtain context embedding e^c and initial hidden state h_0. At each timestep after initialization, the current observation o_t and goal frame g are encoded into e_t^s and are taken by a multi-head cross-attention module together with fixed e^c and recurrently updated h_t to produce policy π_θ, Q-value Q_ω, and terminal signal δ_υ.

Batch-Constrained Q-Learning

北京大学 PEKING UNIVERSITY

As the policy used to record the video could be highly suboptimal for the video navigation problem, we prefer offline reinforcement learning over imitation learning for policy training. Specifically, we adopt BCQ [80], an offline reinforcement learning method that emphasizes constraining the action selection to those within the distribution of the observed frame-action pairs. Since the action space is discrete in our navigation task, we modify the BCQ as follows:

$$\mathcal{L}_q = \mathop{\mathbb{E}}_{o_t, o_{t+1}, g, \hat{a}_t \sim \mathcal{T}^i} \left[Q_\omega\left(g, o_t, \hat{a}_t, \mathcal{T}^i\right) - r(g, o_t, \hat{a}_t) - \gamma \max_{\tilde{a}_{t+1} \in A^\beta} Q_\omega\left(g, o_{t+1}, \tilde{a}_{t+1}, \mathcal{T}^i\right) \right]^2, \quad (5)$$

where we use a binary reward $r(g, o_t, \hat{a}_t) = \mathbb{I}(o_t = g)$. For the selection of $\tilde{a}_{t+1}$, we define set A^β as follow:

$$A^\beta := \left\{ \tilde{a}_{t+1} \middle| \frac{\pi_\theta\left(\cdot | g, o_{t+1}, \mathcal{T}^i\right)}{\max \pi_\theta\left(\cdot | g, o_{t+1}, \mathcal{T}^i\right)} > \beta \right\}, \qquad \beta = 0.5 \quad (6)$$

where β denotes the threshold for the action selection.

Temporal Coherence

北京大学 PEKING UNIVERSITY

- The Temporal indicator $\hat{u}_\phi(e)$, mapping a visual context embedding to a utility score, learns a "fuzzy" measure of progress from natural temporal ordering.
- This self-supervised paradigm naturally leads to temporally aligned context visual representation.

$$t_+ > t_-$$

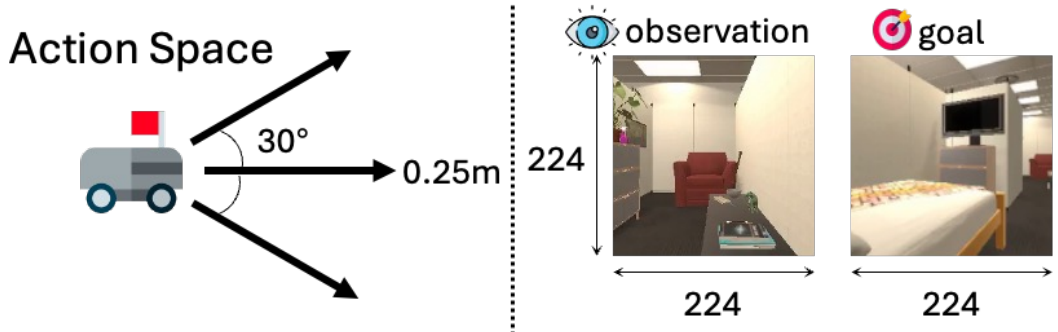
$$f_{t_+} \succ f_{t_-}$$

$$\min_{\zeta,\phi} \mathcal{L}_t = -\mathbb{E}_{f_{t_-}, f_{t_+} \sim \mathcal{V}} \log \sigma \left[\hat{u}_\phi \left(E_\zeta(f_{t_+}) \right) - \hat{u}_\phi \left(E_\zeta(f_{t_-}) \right) \right]$$

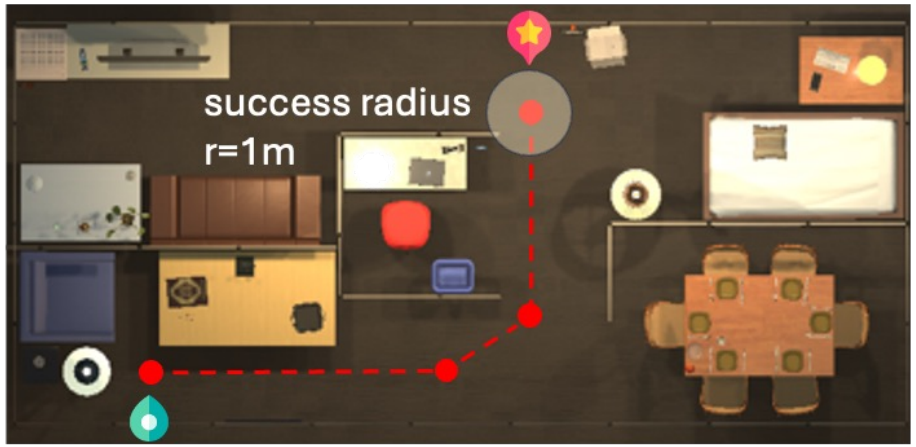
Sim Experiments

北京大学 PEKING UNIVERSITY

Navigation Tasks

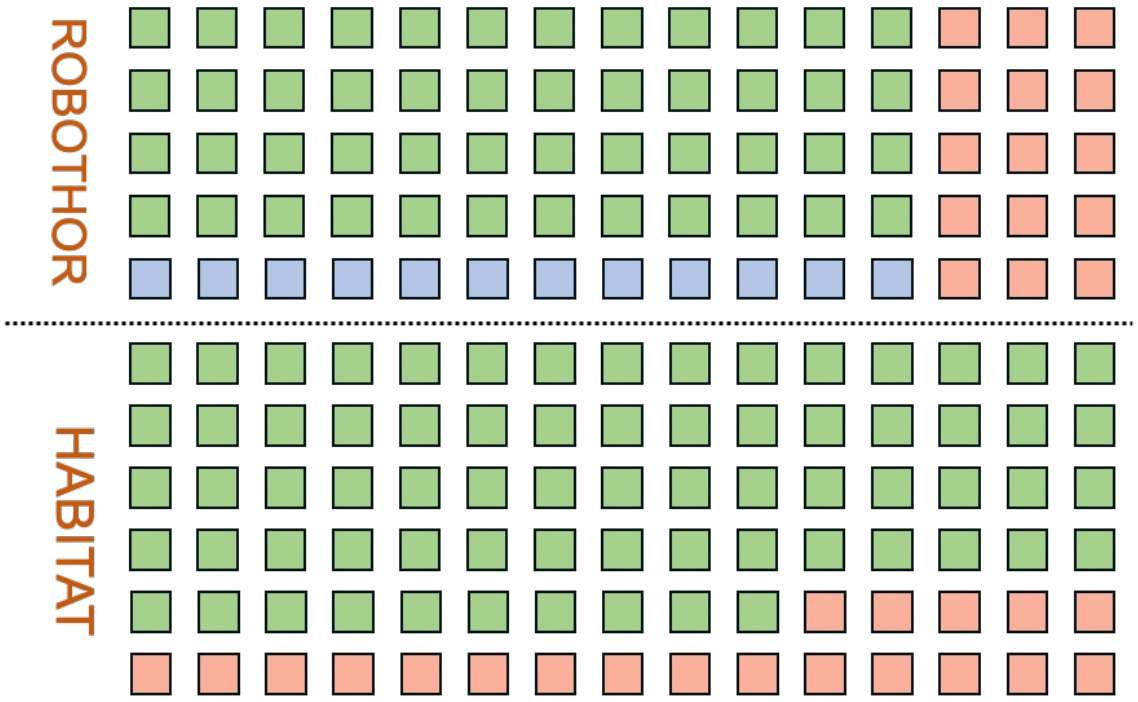


Navigation task bird-eye map



Datasets

(30s per video)



Sim Experiments

北京大学 PEKING UNIVERSITY

Evaluation

- Evaluate generalization ability over **layouts**
- Evaluate generalization ability over **rooms**
- Evaluate in scenes from a different simulator to test the generalization ability over **domains**

Metrics

- Success Rate (**SR**)
- Success weighted by normalized inverse Path Length (**SPL**)
- Trajectory Length (**TL**)
- Navigation Error (**NE**)

Baselines

- GPT-4o
- Video-LLaVA [1]

Two large multimodal models are included as baselines which exhibit remarkable **zero-shot multimodal understanding and generation capabilities**;
Existing methods are excluded because of **inconsistency with video navigation**.

[1] Lin B, Zhu B, Ye Y, et al. Video-llava: Learning united visual representation by alignment before projection. Arxiv 2023.

Sim Experiments

北京大学 PEKING UNIVERSITY

TABLE I: Performance comparison of success rates (SR), success path lengths (SPL), trajectory lengths (TL), and navigation errors (NE) across different methods and environments. The blue part shows some training-free methods and the green part shows visual navigation approaches which require training.

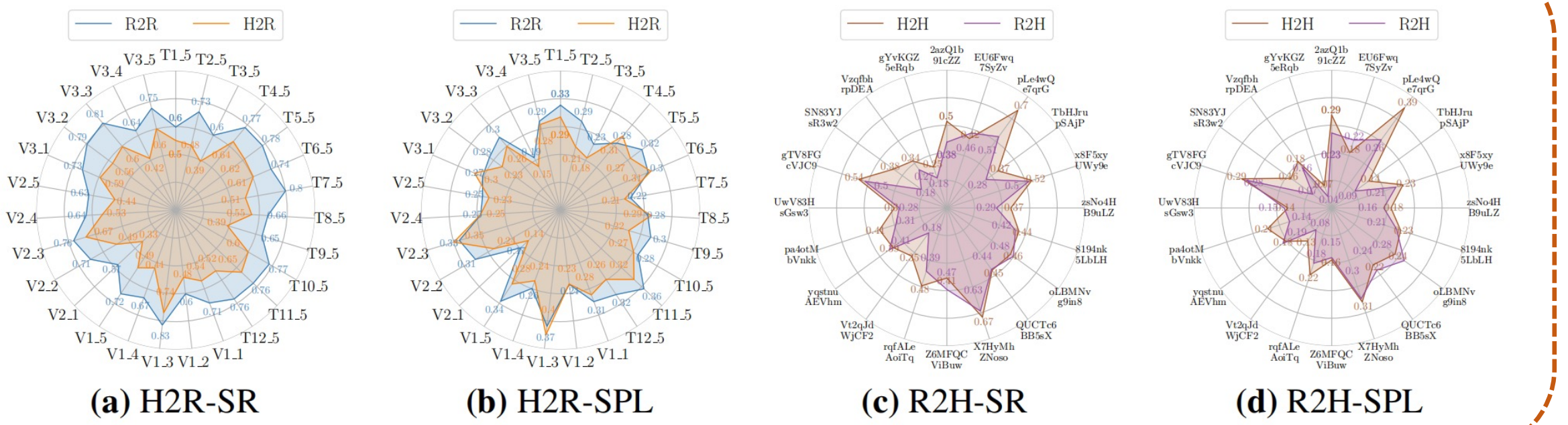
Method	Robothor								Habitat			
	Unseen Layout				Unseen Room				Unseen Room			
	SR(%)↑	SPL(%)↑	TL↓	NE↓	SR(%)↑	SPL(%)↑	TL↓	NE↓	SR(%)↑	SPL(%)↑	TL↓	NE↓
Random	26.00	13.27	394.77	2.93	23.10	12.02	405.87	2.89	24.53	10.03	402.02	4.97
GPT-4o	31.97	17.18	363.23	3.03	31.69	16.49	364.63	2.89	35.16	20.39	**345.84**	4.70
Video-LLaVA	20.07	8.31	381.92	3.08	17.51	6.58	385.72	3.32	14.55	9.02	384.53	4.71
AVDC	32.53	9.43	430.77	2.15	31.14	11.57	424.56	2.13	27.29	12.24	418.12	4.30
VGM	47.04	23.90	376.92	2.44	40.51	19.99	400.90	2.38	34.34	17.30	393.92	4.19
ZSON	37.83	21.31	349.31	2.23	35.78	18.81	372.68	2.37	32.83	19.86	395.94	4.32
NOLO (ours)	**71.92**	**29.26**	**238.92**	**1.79**	**70.48**	**27.74**	**248.44**	**1.87**	**43.65**	**20.77**	347.57	**3.67**

Sim Experiments

北京大学 PEKING UNIVERSITY

Habitat - RoboTHOR

RoboTHOR - Habitat

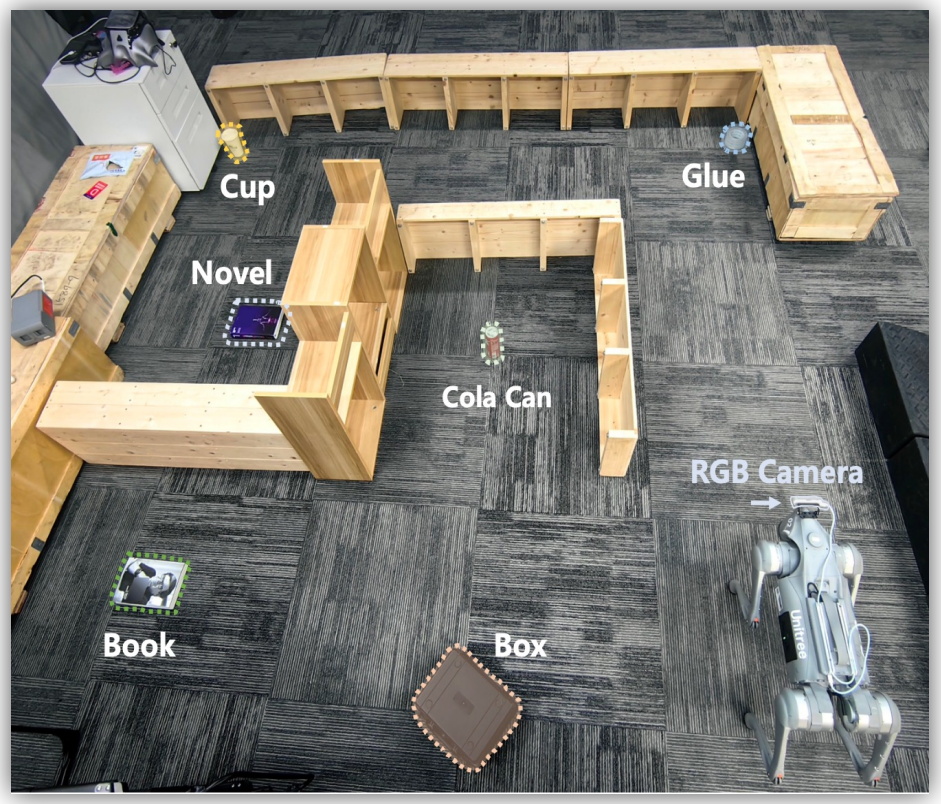


(a) H2R-SR (b) H2R-SPL (c) R2H-SR (d) R2H-SPL

Figure 7: Average SR and SPL of cross domain evaluation results.

Real Experiments

北京大学
PEKING UNIVERSITY



Configuration

Deployment

Conclusion

北京大学 PEKING UNIVERSITY

- To solve Video Navigation problem, NOLO trains a **generalizable in-context** navigation policy in an **offline** manner **purely from videos**.
- NOLO seamlessly incorporates optical flow into offline reinforcement learning BCQ via pseudo-action labeling. We further propose to add an auxiliary temporal coherence loss for learning temporally aligned context visual representation.
- Empirical evaluations in RoboTHOR and Habitat demonstrate the effectiveness of NOLO in video navigation. Notably, this is achieved by only watching a single 30-second video clip in each scene.

北京大学
PEKING UNIVERSITY

Thanks

Bohan Zhou

2025.10.23