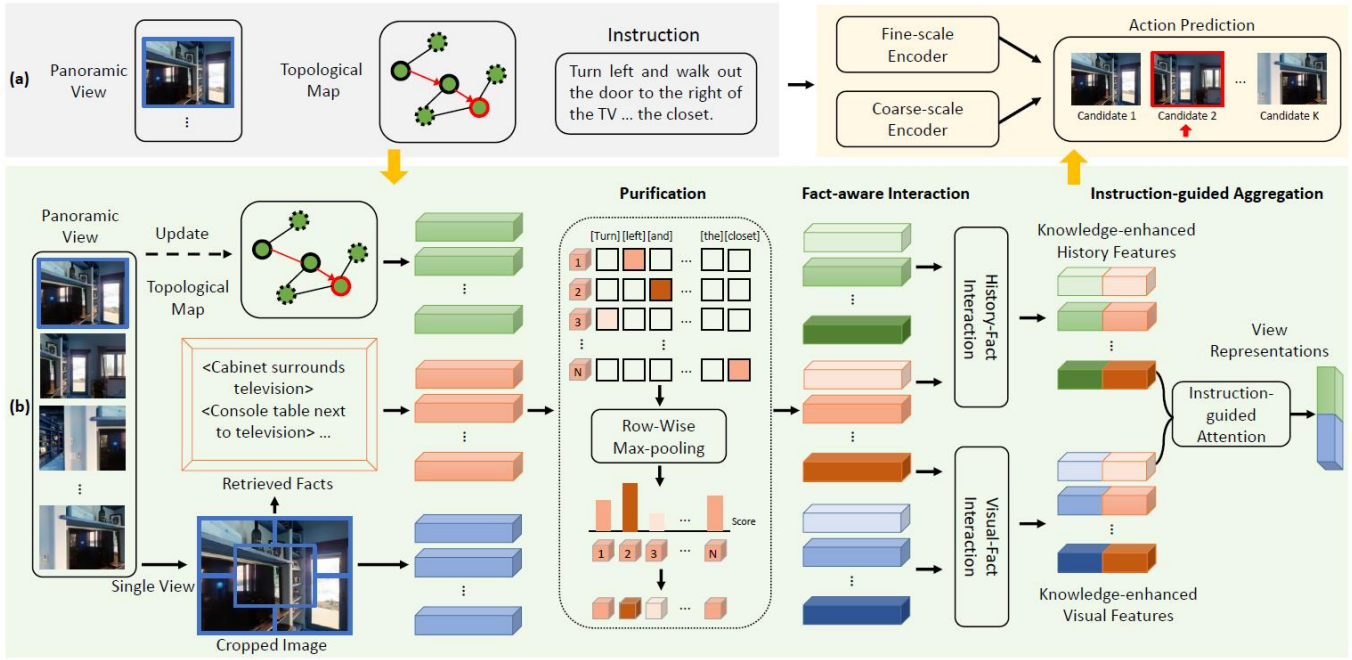
KERM: Knowledge Enhanced Reasoning for Vision-and-Language Navigation

Xiangyang Li[1,2], Zihan Wang[1,2], Jiahao Yang[1,2], Yaowei Wang[3], Shuqiang Jiang[1,2,3]
[1]Key Lab of Intelligent Information Processing Laboratory of the Chinese Academy of Sciences (CAS), Institute of Computing Technology, Beijing, 100190, China
[2]University of Chinese Academy of Sciences, Beijing, 100049, China
[3]Peng Cheng Laboratory, Shenzhen, 518055, China

TUE-AM-246

KERM: Knowledge Enhanced Reasoning for Vision-and-Language Navigation



- We incorporate region-centric knowledge to comprehensively depict navigation views in VLN tasks.
- We propose the knowledge enhanced reasoning model (KERM) to inject fact features into the visual representations for better action prediction.
- We conduct extensive experiments to validate the effectiveness of our method and show that it outperforms existing methods with a better generalization ability.

Challenges

The key problem：the representations of observation

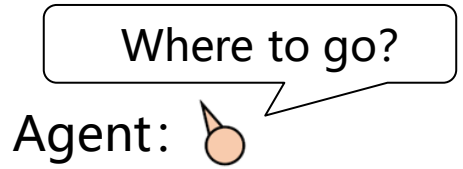
- How to represent the visual observation
- Ho to align the visual observation and the language

Previous works:

- Entire image features of the observation

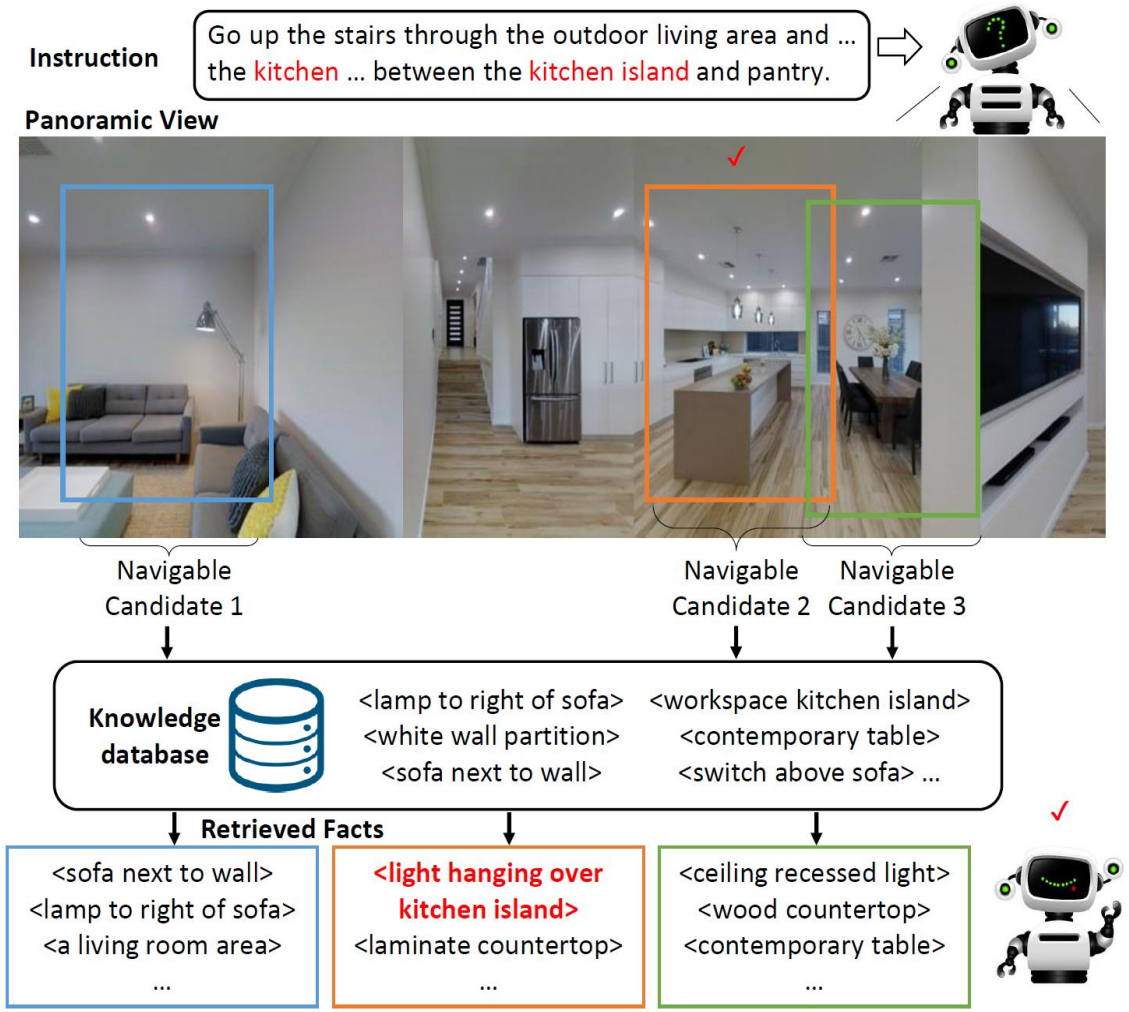
- Object-centric features to represent the environment

Panoramic observation



Instruction: Walk through the arch to the left of the mirror …

Motivation



Humans and animals make inferences about the world under limited time and knowledge. [Psychological review, 1996]

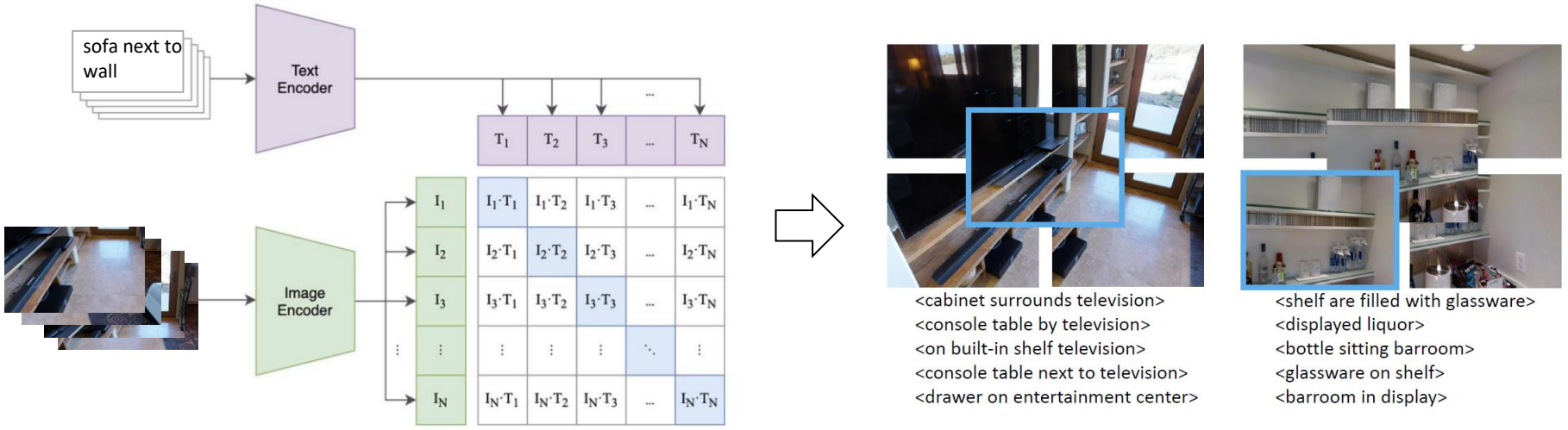
Knowledge for Vision-and-language Navigation

- Capture critical information in visual observation
- Knowledge is used as a bridge to align vision and language
- Knowledge improves the generalization ability of the model

Gerd Gigerenzer and Daniel G Goldstein. Reasoning the fast and frugal way: models of bounded rationality. Psychological review, 103(4):650, 1996.

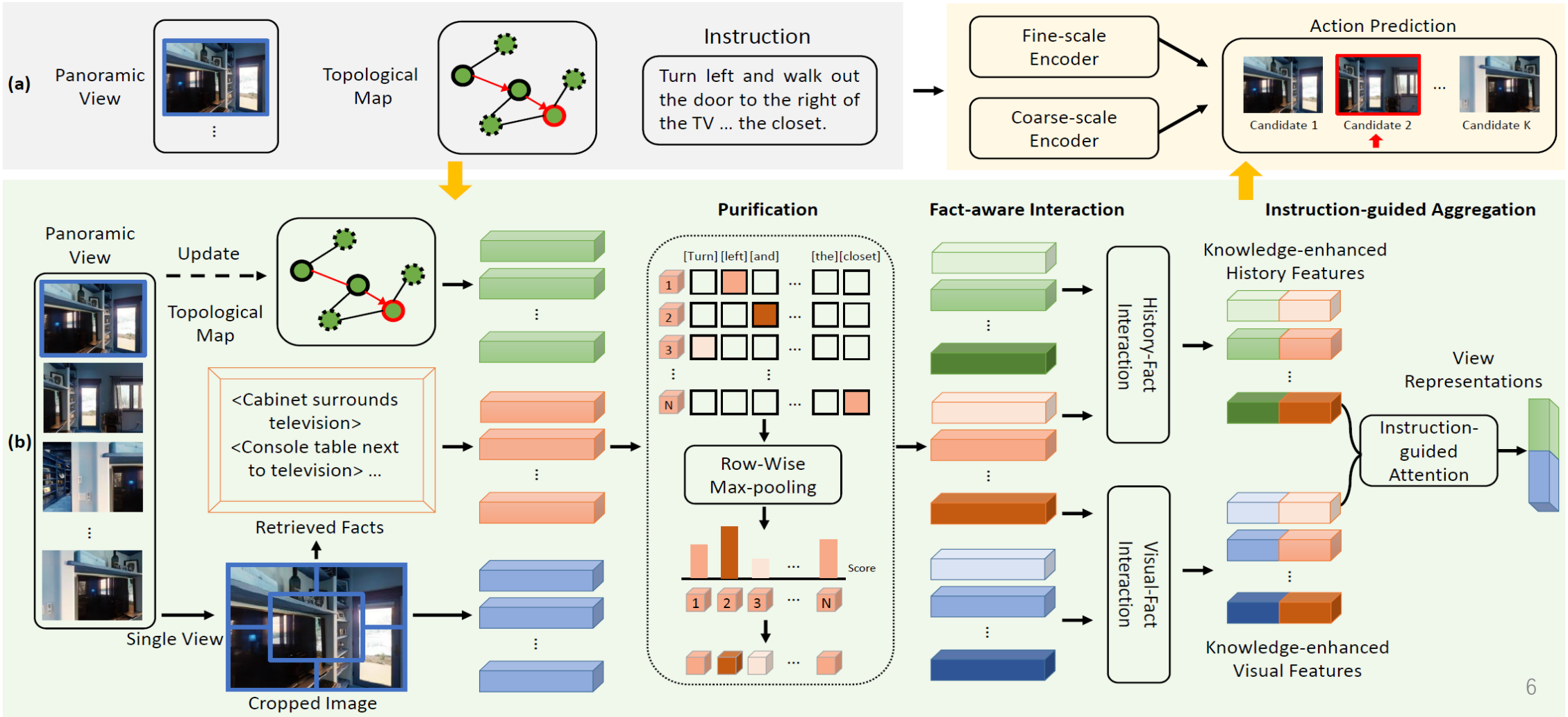
Fact Acquisition

- Convert all of the “attribute-object” pairs and “subject-predicate-object” triplets on the Visual Genome dataset to their synset canonical forms
- Get 630K facts expressed by language descriptions that are used to build the knowledge base
- Crop each view image into five sub-regions and retrieve facts for these sub-regions from the knowledge base



Framework

- KERM utilizes the purification, fact-aware interaction, and instruction-guided aggregation modules to select and gather crucial and relevant information.



Experiments

Comparison with state-of-the-art methods on the REVERIE dataset

Methods	Val Seen						Val Unseen						Test Unseen					
	Navigation				Grounding		Navigation				Grounding		Navigation				Grounding	
	TL↓	OSR↑	SR↑	SPL↑	RGS↑	RGSPL↑	TL↓	OSR↑	SR↑	SPL↑	RGS↑	RGSPL↑	TL↓	OSR↑	SR↑	SPL↑	RGS↑	RGSPL↑
Seq2Seq [Anderson *et al.*, 2018]	12.88	35.70	29.59	24.01	18.97	14.96	11.07	8.07	4.20	2.84	2.16	1.63	10.89	6.88	3.99	3.09	2.00	1.58
RCM [Wang *et al.*, 2019]	10.70	29.44	23.33	21.82	13.23	15.36	11.98	14.23	9.29	6.97	4.89	3.89	10.60	11.68	7.84	6.67	3.67	3.14
VLNBERT [Hong *et al.*, 2021]	13.44	53.90	51.79	47.96	38.23	35.61	16.78	35.02	30.67	24.90	18.77	15.27	15.68	32.91	29.61	23.99	16.50	13.51
AirBERT [Guhur *et al.*, 2021]	15.16	49.98	47.01	42.34	32.75	30.01	18.71	34.51	27.89	21.88	18.23	14.18	17.91	34.20	30.28	23.61	16.83	13.28
HOP [Qiao *et al.*, 2022]	13.80	54.88	53.76	47.19	38.65	33.85	16.46	36.24	31.78	26.11	18.85	15.73	16.38	33.06	30.17	24.34	17.69	14.34
HAMT [Chen *et al.*, 2021]	12.79	47.65	43.29	40.19	27.20	15.18	14.08	36.84	32.95	30.20	18.92	17.28	13.62	33.41	30.40	26.67	14.88	13.08
DUET [Chen *et al.*, 2022]	13.86	73.68	71.75	63.94	57.41	51.14	22.11	51.07	46.98	33.73	32.15	23.03	21.30	56.91	52.51	36.06	31.88	22.06
KERM-pt(Ours)	14.25	**74.49**	**71.89**	**64.04**	**57.55**	**51.22**	22.47	**53.65**	**49.02**	**34.83**	**33.97**	**24.14**	18.38	**57.44**	52.26	**37.46**	**32.69**	**23.15**
KERM(Ours)	12.84	**79.20**	**76.88**	**70.45**	**61.00**	**56.07**	21.85	**55.21**	**50.44**	**35.38**	**34.51**	**24.45**	17.32	**57.58**	**52.43**	**39.21**	**32.39**	**23.64**

Comparison with state-of-the-art methods on the SOON dataset

Method	TL↓	OSR↑	SR↑	SPL↑	RGSPL↑
GBE [Zhu *et al.*, 2021]	28.96	28.54	19.52	13.34	1.16
DUET [Chen *et al.*, 2022]	36.20	50.91	36.28	22.58	3.75
KERM (Ours)	35.83	**51.62**	**38.05**	**23.16**	**4.04**

In VLN of high-level instructions, our method is significantly better than the methods that don‘t introduce knowledge.

Experiments

Ablation study results on val unseen split of the REVERIE dataset. "Pur." denotes the purification module. "VFInt." and "HF-Int." denote the vision-fact and history-fact interaction module respectively

Pur.	VF-Int.	HF-Int.	OSR↑	SR↑	SPL↑	RGS↑	RGSPL↑
			51.07	46.98	33.73	32.15	23.03
✓		✓	53.78	49.33	35.01	33.80	23.89
✓	✓		53.53	48.96	34.70	33.63	23.58
	✓	✓	54.01	49.43	34.79	33.22	23.80
✓	✓	✓	**55.21**	**50.44**	**35.38**	**34.51**	**24.45**

With the purification, interaction, and aggregation modules, our method obtains the best performance.

Experiments

	OSR↑	SR↑	SPL↑	RGS↑	RGSPL↑
Baseline	51.07	46.98	33.73	32.15	23.03
+ Object	54.27	49.73	33.95	33.91	23.40
+ Fact (KERM)	**55.21**	**50.44**	**35.38**	**34.51**	**24.45**

Object features are not sufficient to provide strong generalization performance compared to fact descriptions which have a much larger semantic feature space.

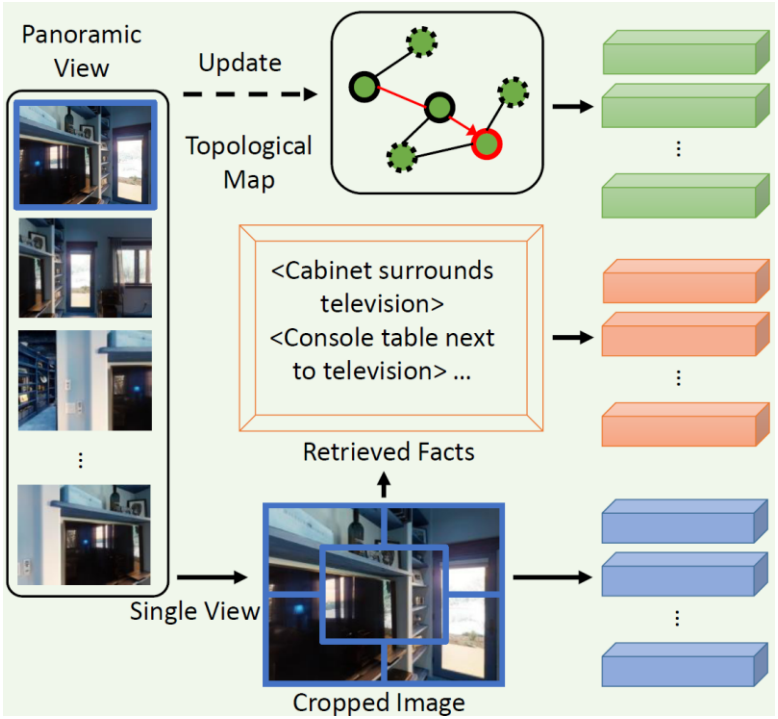
Number	OSR↑	SR↑	SPL↑	RGS↑	RGSPL↑
1	55.69	49.67	34.46	34.05	23.89
5	**55.21**	**50.44**	**35.38**	**34.51**	**24.45**
9	53.59	48.99	34.87	33.14	23.79

Fewer but larger sub-regions are difficult to retrieve more finegrained facts, while more but smaller regions are too fragmented to contain all the complete parts that can retrieve accurate facts.

Qualitative Results

Facts related to the target object have almost the highest purification weights, which demonstrates that our model can automatically select the relevant facts, thus obtain the better performance.

Conclusion



- We incorporate region-centric knowledge to comprehensively depict navigation views in VLN tasks. For each navigable candidate, the retrieved facts (*i.e.*, knowledge described by language descriptions) are complementary to visible content.
- We propose the knowledge enhanced reasoning model (KERM) to inject fact features into the visual representations of navigation views for better action prediction.
- We conduct extensive experiments to validate the effectiveness of our method and show that it outperforms existing methods with a better generalization ability.