REVERIE: Remote Embodied Visual Referring Expressions in Real Indoor Environments

Yuankai Qi[1,2], Qi Wu[1], Peter Anderson[3], Xin Wang[4], William Yang Wang[4], Chunhua Shen[1], Anton van den Hengel[1]

[1]Australian Centre for Robotic Vision, The University of Adelaide, Australia

[2]Harbin Institute of Technology, Weihai, China

[3]Georgia Institute of Technology, USA *[4]University of California, Santa Barbara, USA*

CVPR 2020

A Long-hold Goal

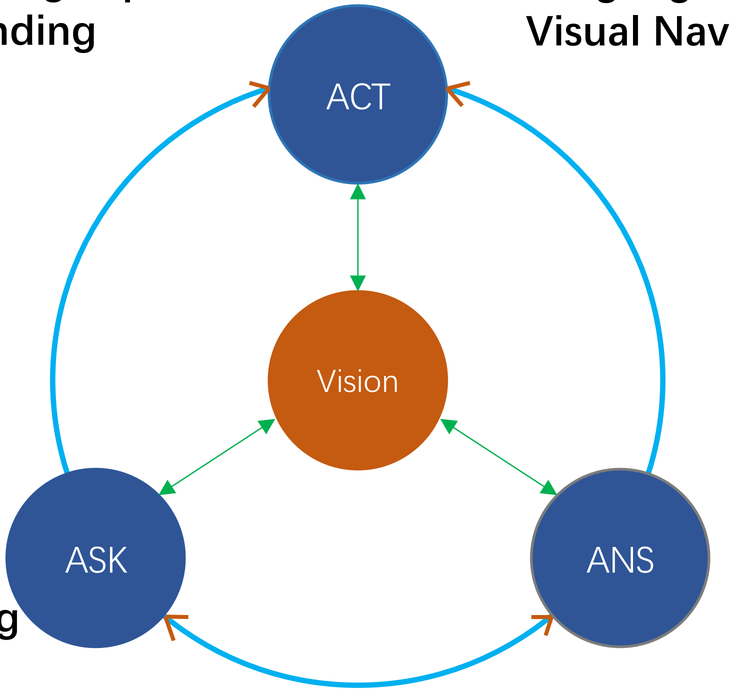
Build intelligent robots that can perceive the environment, execute commands, and communicate with human.

The Attempts

Query from GuessWhat?!:	Query from ReferCOCO: woman in black
is it a person? Yes is the person a man? No is it the left woman? No does she wear glasses? Yes	Query from ReferCOCO+: woman wearing glasses with smile
	Query from ReferCOCOg: a woman in black playing a game with her friends

- Referring Expression Grounding

- Language-guided Visual Navigation



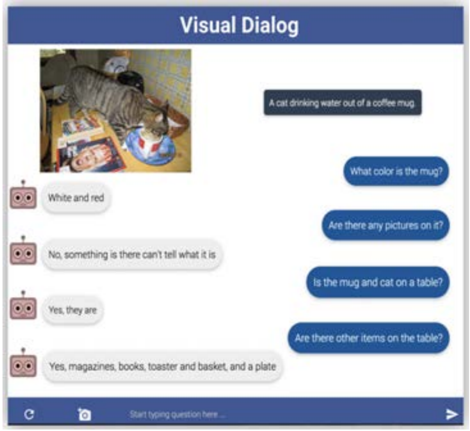
Hi, Robby, can you get my cup from the cupboard?

Is it tall one?
No
Is it short one?
Yes
Is it from IKEA?
No
Is it pink?
Yes
Does it have a handle?
I'll get it myself .....

Is it pink?
Yes
Does it have Elsa on it?
No
Brilliant!

- Image Captioning
- Visual Question Generation (VQG)

- VQA
- VisDialog



Visual Dialog

A cat drinking water out of a coffee mug

What color is the mug?

White and red

Are there any pictures on it?

No, something is there can't tell what it is

Is the mug and cat on a table?

Yes, they are

Are there other items on the table?

Yes, magazines, books, toaster and basket, and a plate

Start typing question here ...

A New Task

- They cannot reflect communication about **remote objects**.

Example:

“Bring me the blue cushion from the living room”

REVERIE: Remote Embodied Visual Referring Expressions in Real Indoor Environments

The REVERIE Task

Go to the bathroom in the bedroom with an orange stripe. Make sure the faucet is not leaking.

R2R vs. REVERIE

Two key difference:

- Fine-grained instructions vs. High-level instruction

  R2R: ‘Go to the top of the stairs then turn left and walk along the hallway and stop at the first bedroom on your right’

  REVERIE: ‘the cold tap in the first bedroom on level two’

- Point navigation vs. Remote object grounding

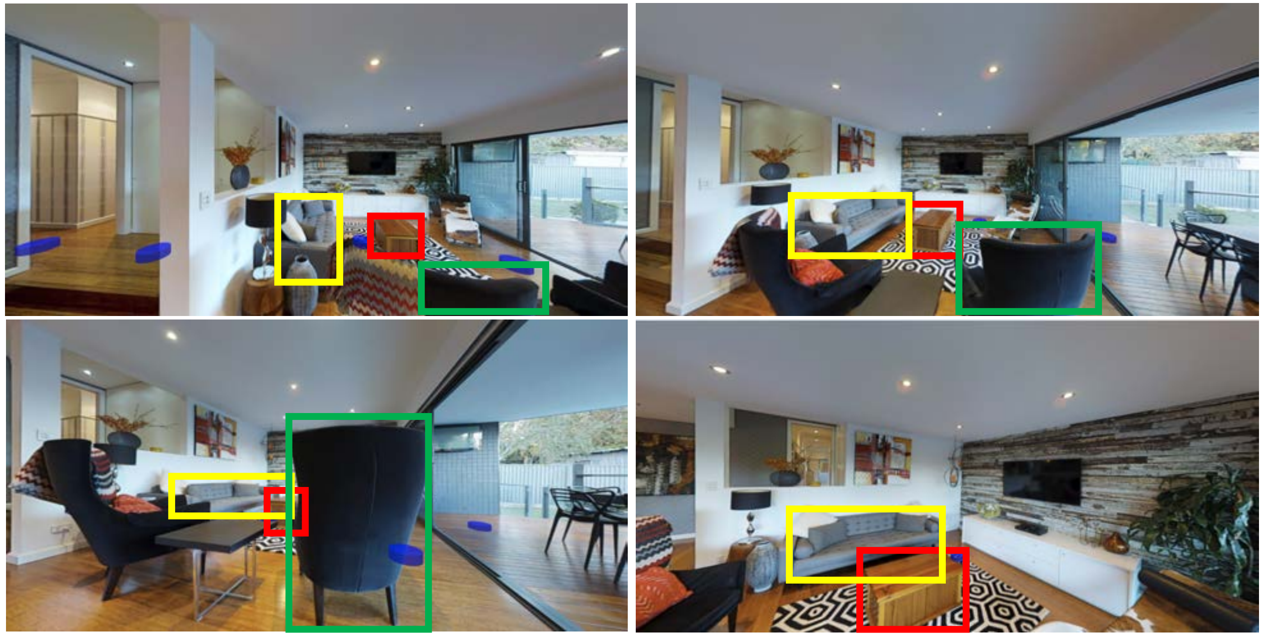
RefExp Grounding vs. REVERIE

Three key difference

- Visible target object vs. Invisible target object
- Single candidate image vs. Panoramas of all possible viewpoints
- Front view vs. Various Views

RefExp Grounding

REVERIE



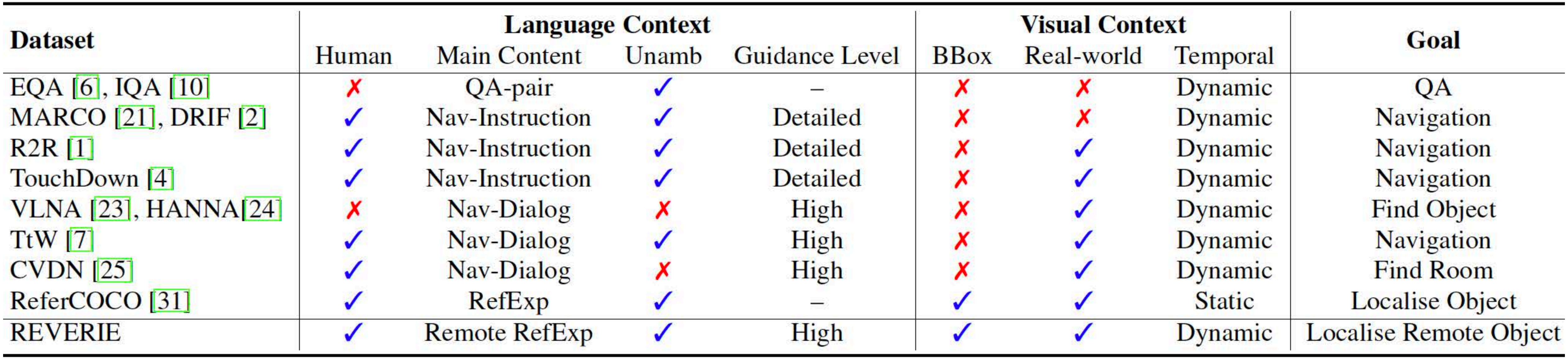
Dataset

21,702 instructions, >1600 words, 4,140 target objects, 489 categories

	Buildings	Instructions	Objects
Train	60	10,466	2,353
Val Seen	46	1,423	440
Val Unseen	10	3,521	513
Test	16	6,292	834

* The split follows the strategy of R2R dataset for research convenience.

Comparison with existing datasets

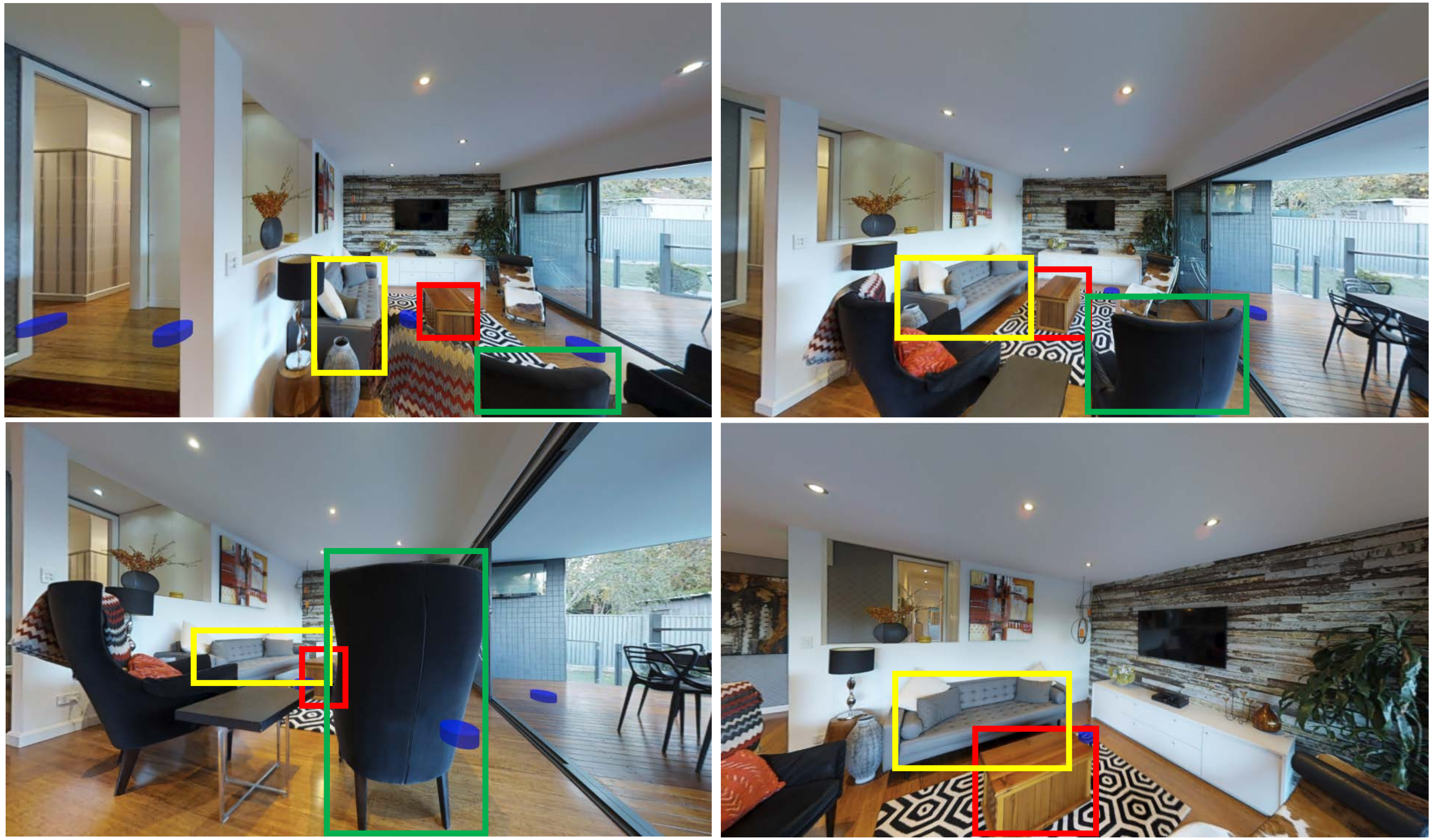


Dataset	Language Context				Visual Context			Goal
	Human	Main Content	Unamb	Guidance Level	BBox	Real-world	Temporal	
EQA [6], IQA [10]	✗	QA-pair	✓	–	✗	✗	Dynamic	QA
MARCO [21], DRIF [2]	✓	Nav-Instruction	✓	Detailed	✗	✗	Dynamic	Navigation
R2R [1]	✓	Nav-Instruction	✓	Detailed	✗	✓	Dynamic	Navigation
TouchDown [4]	✓	Nav-Instruction	✓	Detailed	✗	✓	Dynamic	Navigation
VLNA [23], HANNA[24]	✗	Nav-Dialog	✗	High	✗	✓	Dynamic	Find Object
TtW [7]	✓	Nav-Dialog	✓	High	✗	✓	Dynamic	Navigation
CVDN [25]	✓	Nav-Dialog	✗	High	✗	✓	Dynamic	Find Room
ReferCOCO [31]	✓	RefExp	✓	–	✓	✓	Static	Localise Object
REVERIE	✓	Remote RefExp	✓	High	✓	✓	Dynamic	Localise Remote Object

What is the challenge of this task?

Challenges

(1/3) Significant Appearance Variation



Challenges

(2/3) Rich Linguistic Phenomena

Dangling modifiers (e.g. 1), spatial relations (e.g. 3), imperatives (e.g. 4), co-references (e.g. 5)

1. Fold the towel in the bathroom with the fishing theme

2. Push in the bar chair, in the kitchen, by the oven.

3. Go to the blue family room and bring the framed picture of a person on a horse at the top left corner above the TV.

4. Could you please dust the light above the toilet in the bathroom that is near the entry way?

5. There is a bottle in the office alcove next to the piano. It is on the shelf above the sink on the extreme right. Please bring it here.

Challenges

(3/3) Less Words, More Contents

- Instruction length: 18 vs 29 words (Room-to-Room dataset)
- 56% instructions mention 3 or more objects, 28% mention 2 objects
- Involve 4,140 objects, falling into 489 categories vs 80 categories in ReferCOO

Solution

Solution

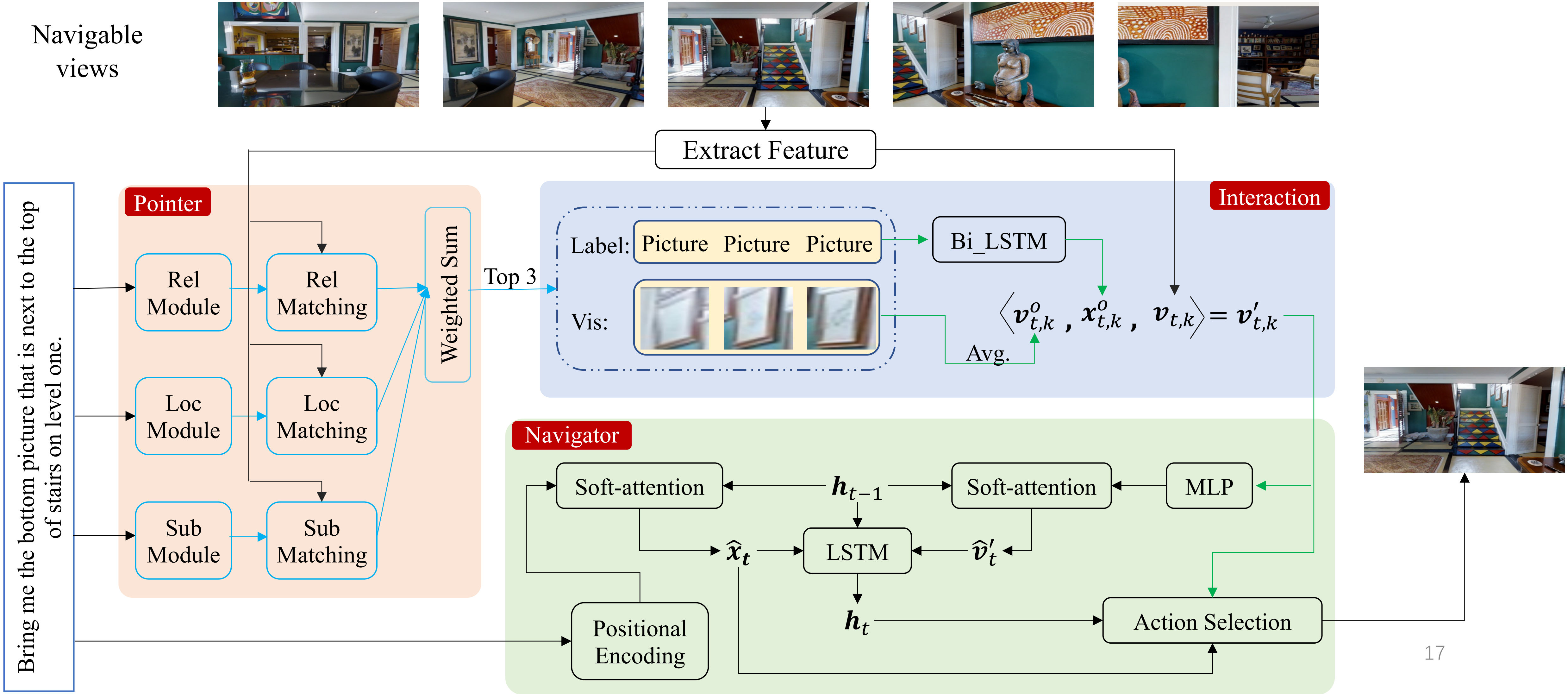
Navigation (Navigator) + Referring Expression Grounding (Pointer)

- Perform grounding when navigation ends
- Perform grounding at each navigation step

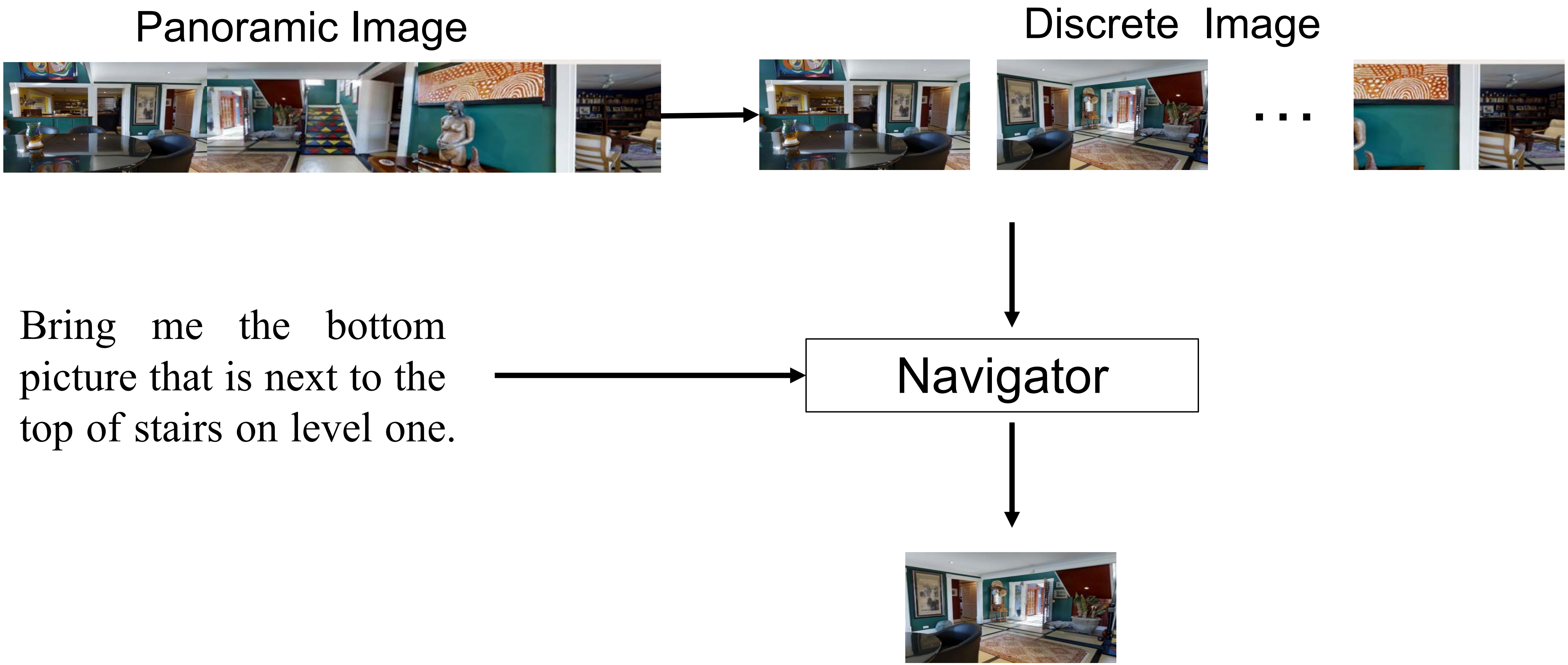
Solution

- Perform grounding when navigation ends
  - 4 Baseline Navigation Model + 4 SoTA Navigation Model
    - Random
    - Shortest
    - R2R-TF
    - R2R-SF
    - SelfMonitor: Chih-Yao Ma, etal, ICLR 2019
    - RCM: Xin Wang, etal, CVPR 2019
    - FAST-Short: Liyinming Ke, etal, CVPR 2019
    - FAST-Lan-Only: a variant of FAST-Short
  - 1 Baseline RefExp Model + 2 SoTA RefExp Model
    - CNN-RNN
    - MAttNet: Licheng Yu, etal, CVPR 2018
    - CM-Erase: Xihui Liu, etal, CVPR 2019

Solution: Interactive Navigator-Pointer Model



How does navigator work?



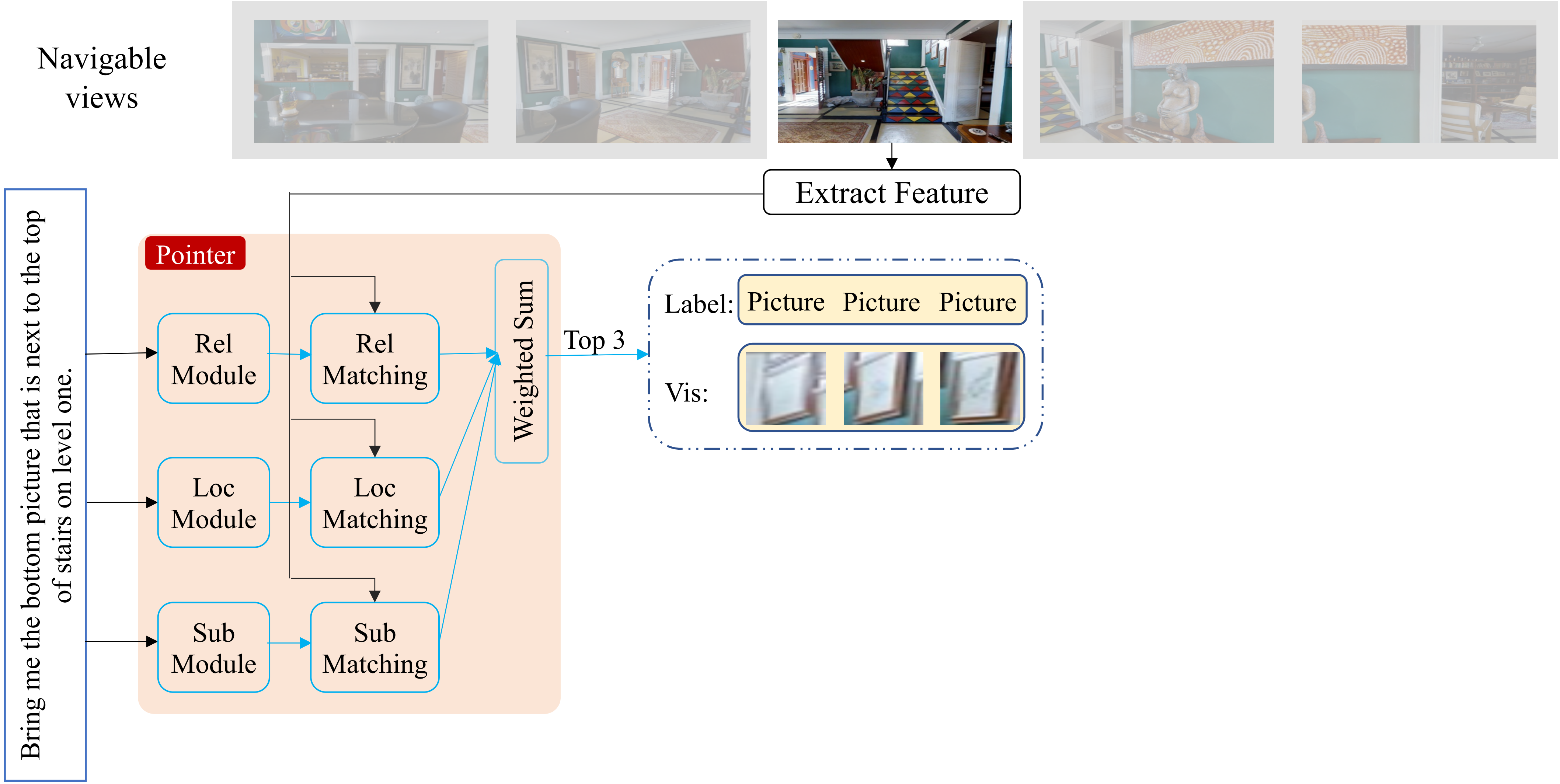
Panoramic Image

Discrete Image

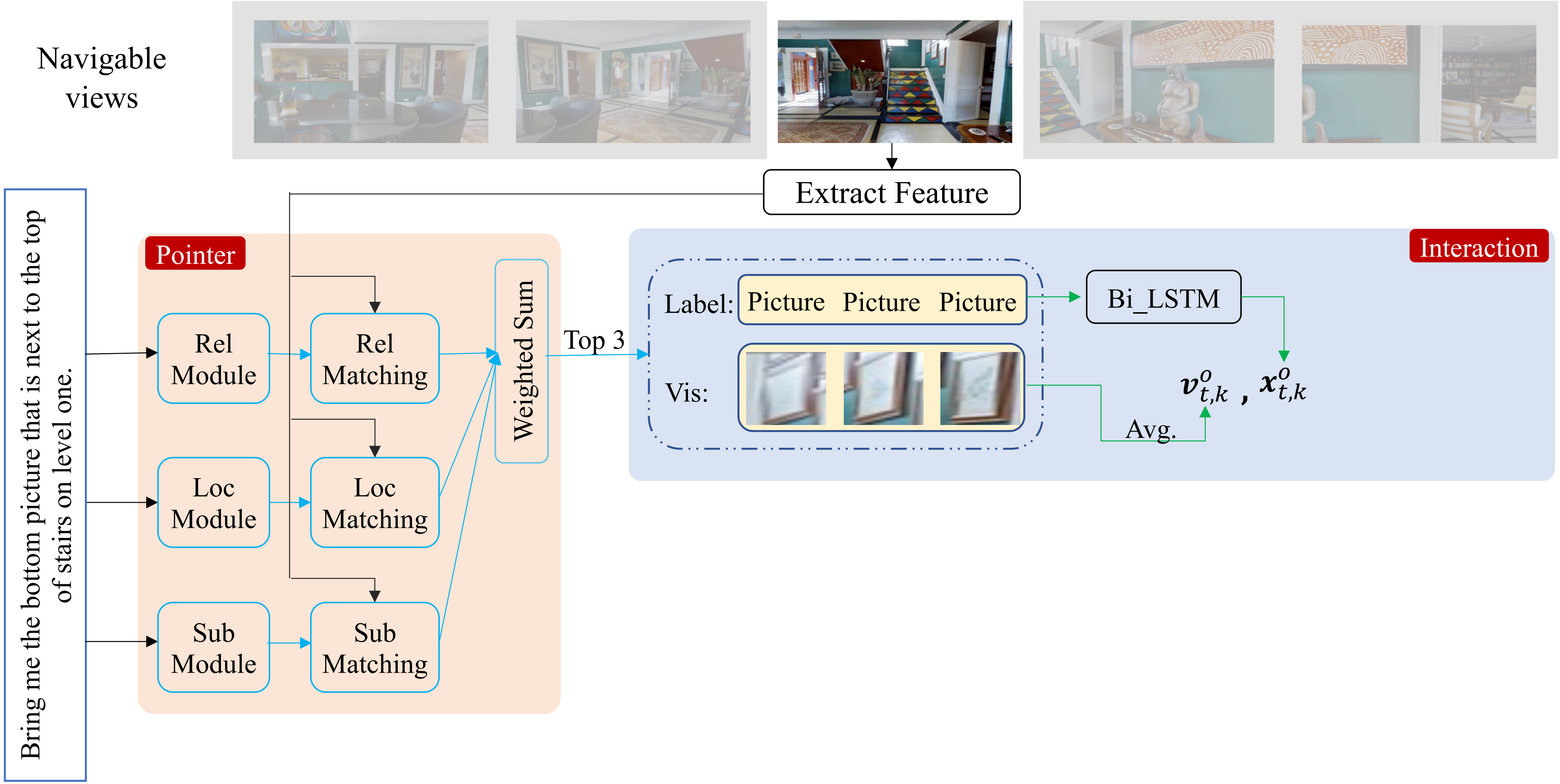
Bring me the bottom picture that is next to the top of stairs on level one.

Navigator

Solution: Interactive Navigator-Pointer Model



Solution: Interactive Navigator-Pointer Model



Navigable views

Extract Feature

Bring me the bottom picture that is next to the top of stairs on level one.

Pointer

Rel Module

Loc Module

Sub Module

Rel Matching

Loc Matching

Sub Matching

Weighted Sum

Top 3

Interaction

Label: Picture Picture Picture

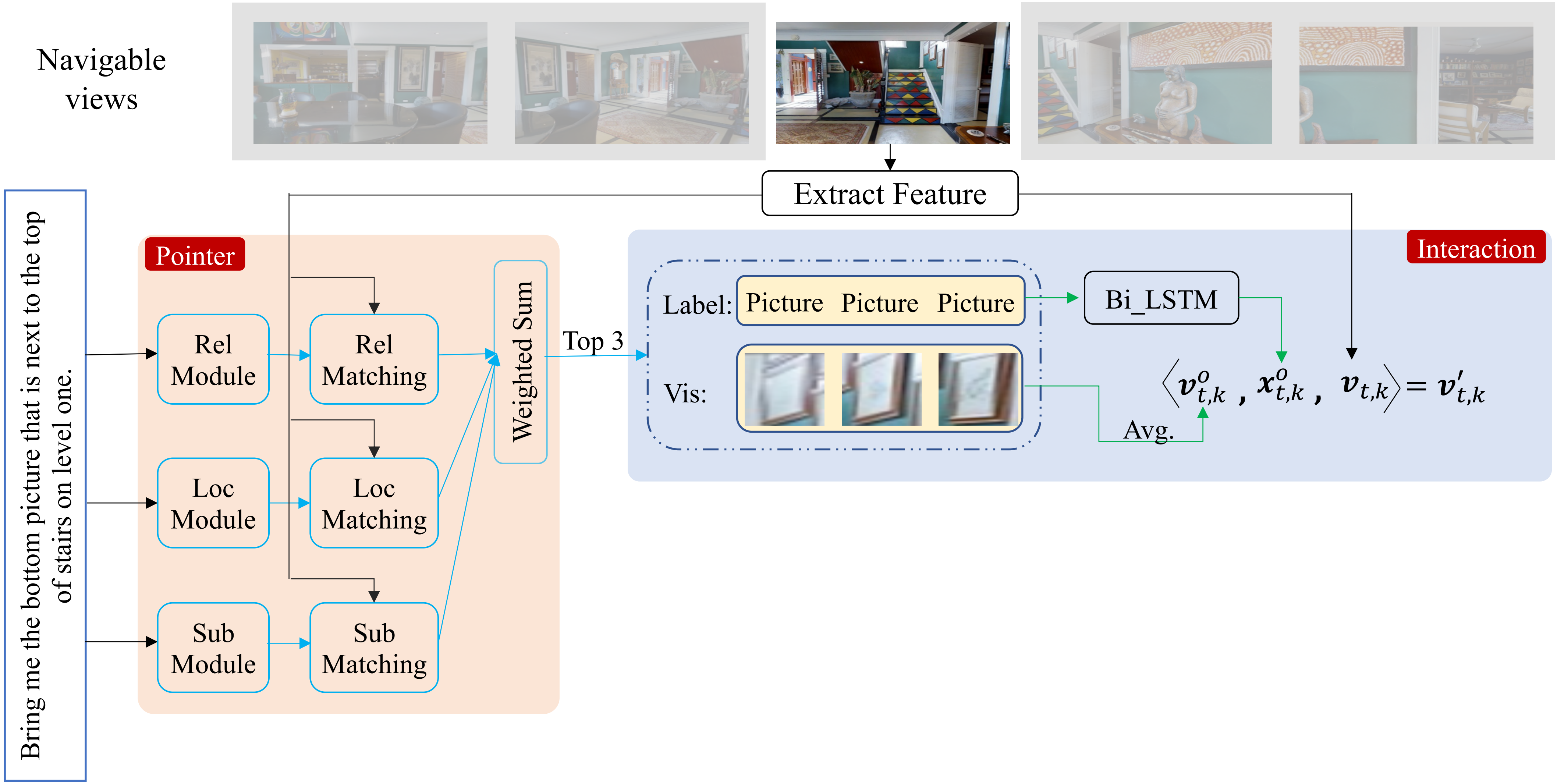
Vis:

Bi_LSTM

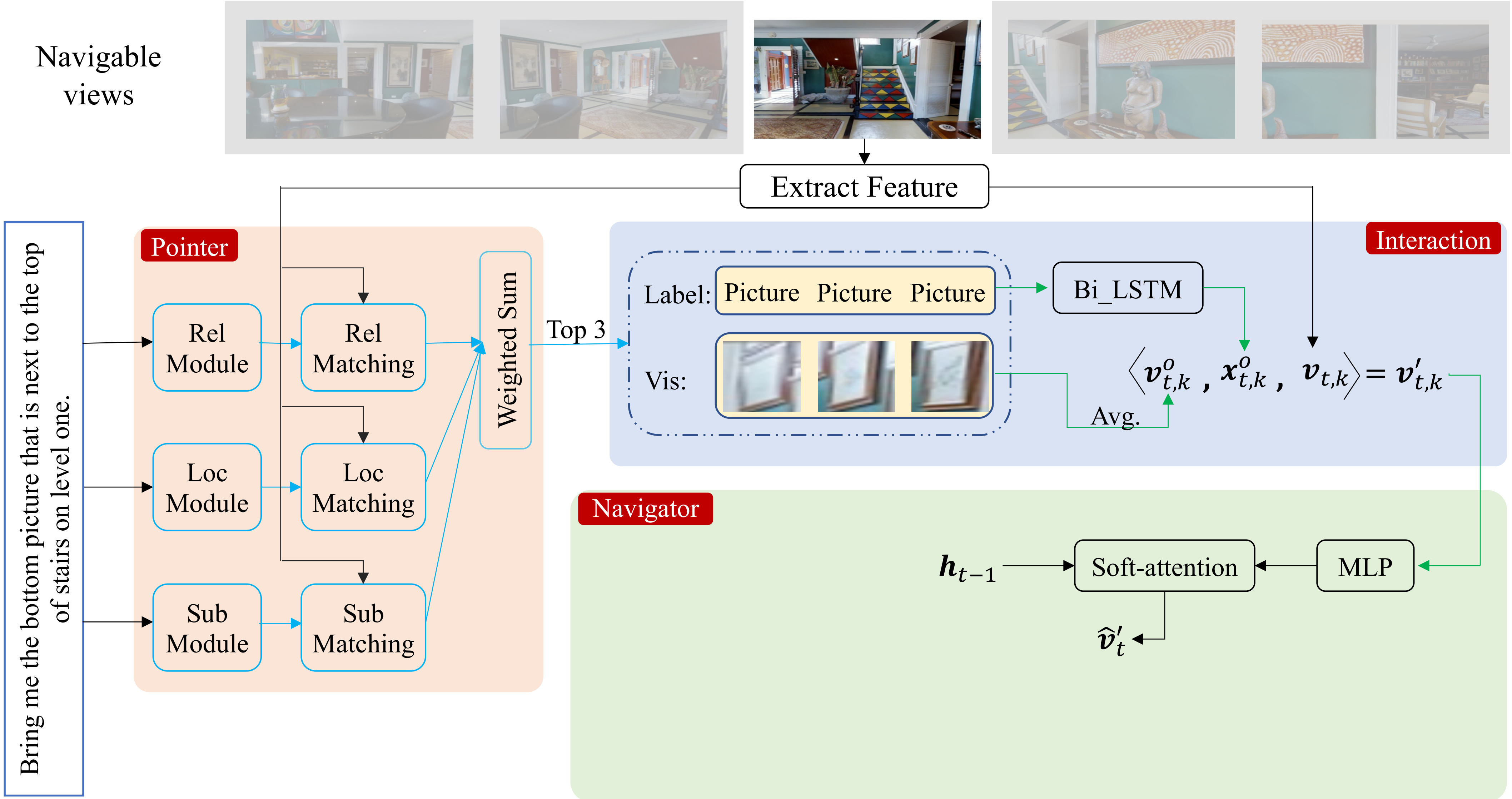
Avg.

$\boldsymbol{v}^{o}_{t,k}$, $\boldsymbol{x}^{o}_{t,k}$

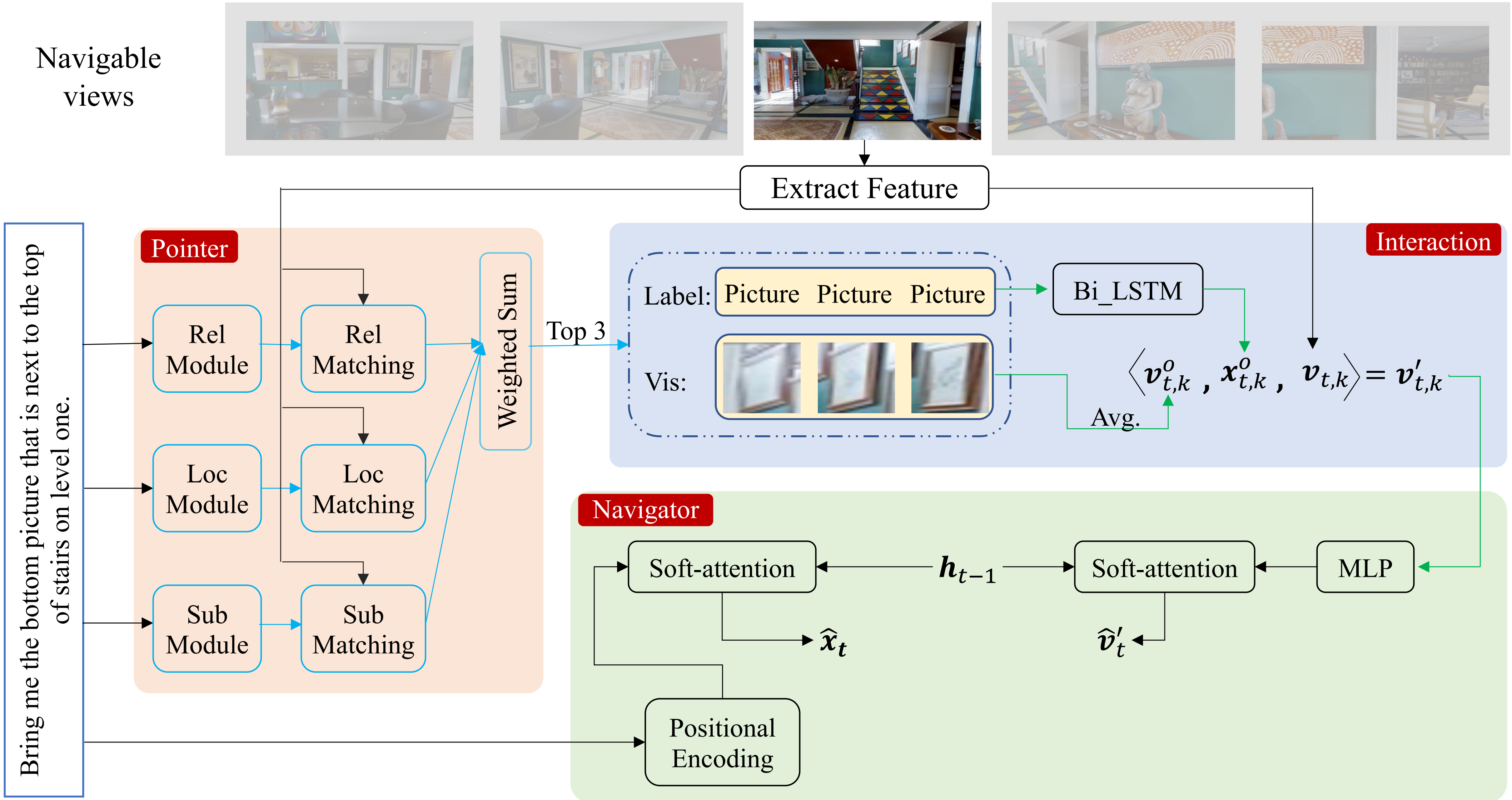
Solution: Interactive Navigator-Pointer Model



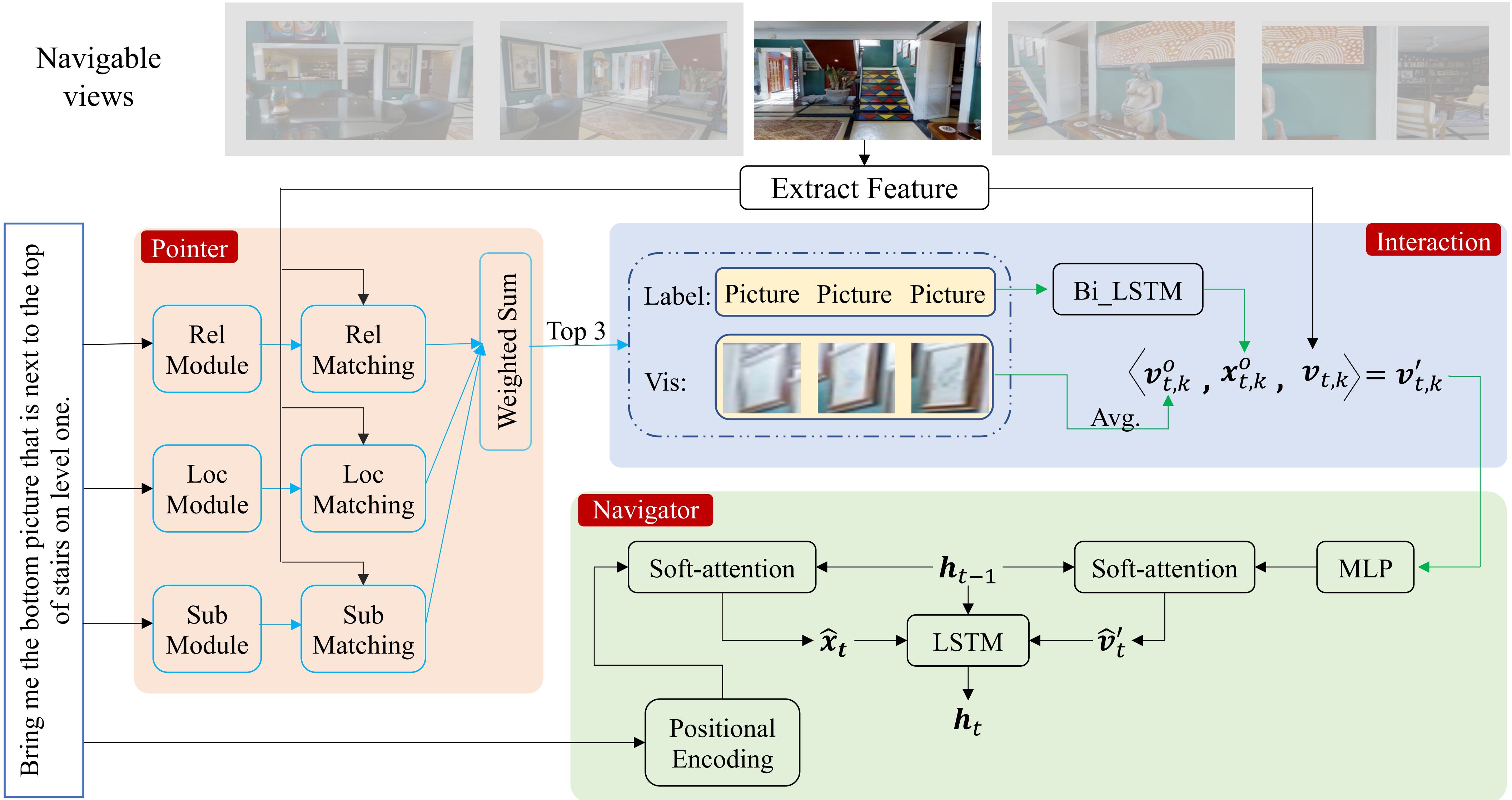
Solution: Interactive Navigator-Pointer Model



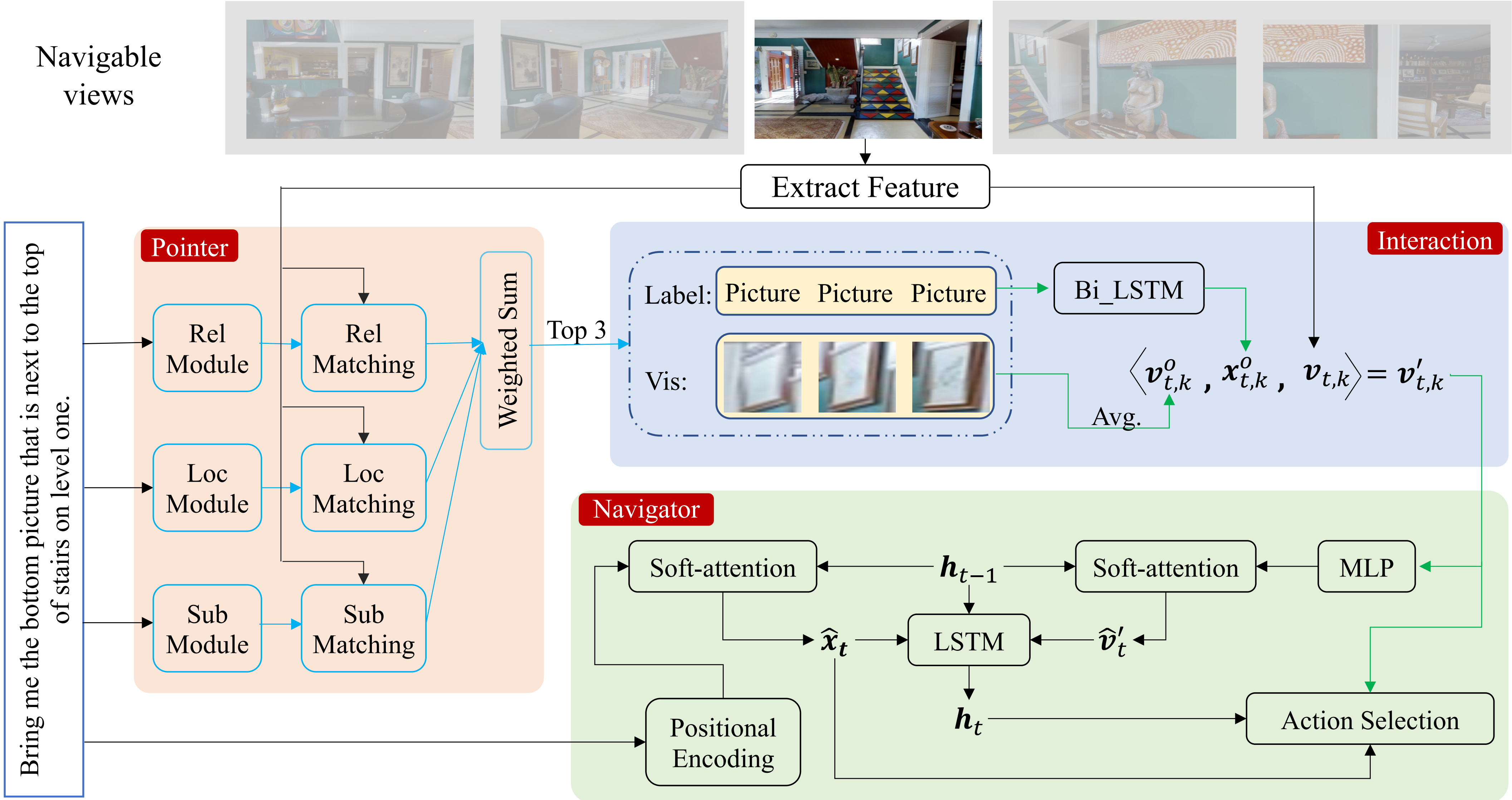
Solution: Interactive Navigator-Pointer Model



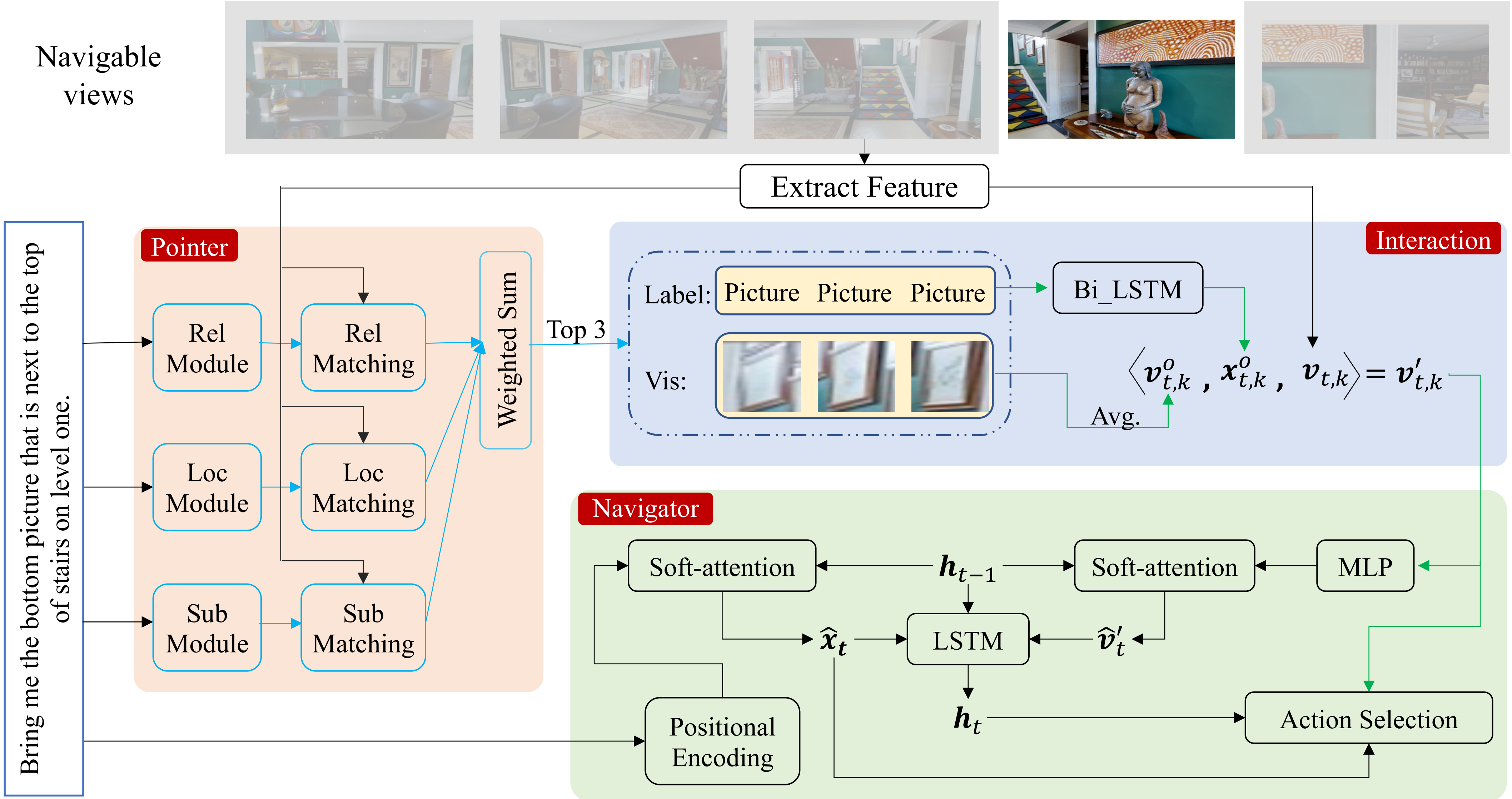
Solution: Interactive Navigator-Pointer Model



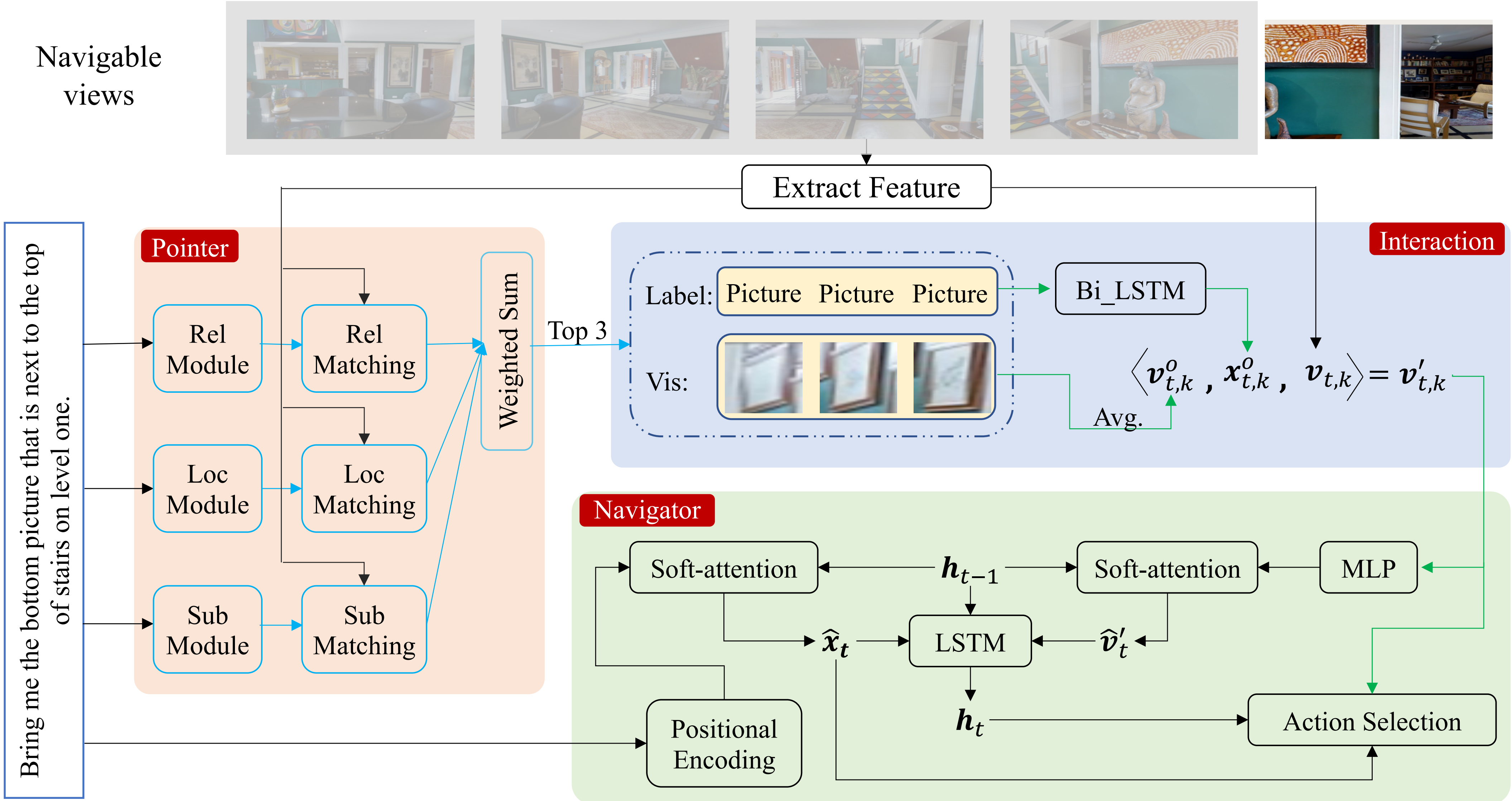
Solution: Interactive Navigator-Pointer Model



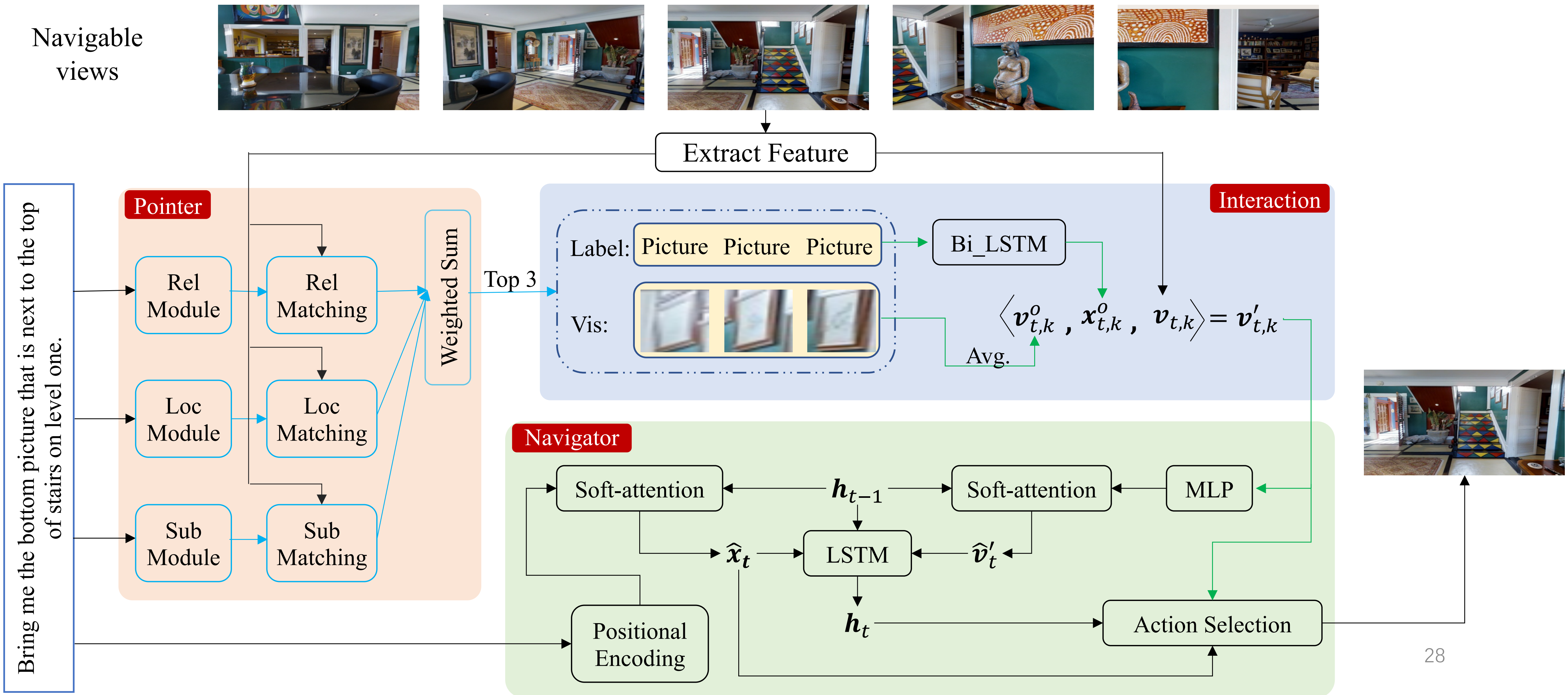
Solution: Interactive Navigator-Pointer Model


Bring me the bottom picture that is next to the top of stairs on level one.

Solution: Interactive Navigator-Pointer Model



Solution: Interactive Navigator-Pointer Model



Bring me the bottom picture that is next to the top of stairs on level one.

Metrics

- A Successful Task
  - Select the correct object from a list of candidates

  Or

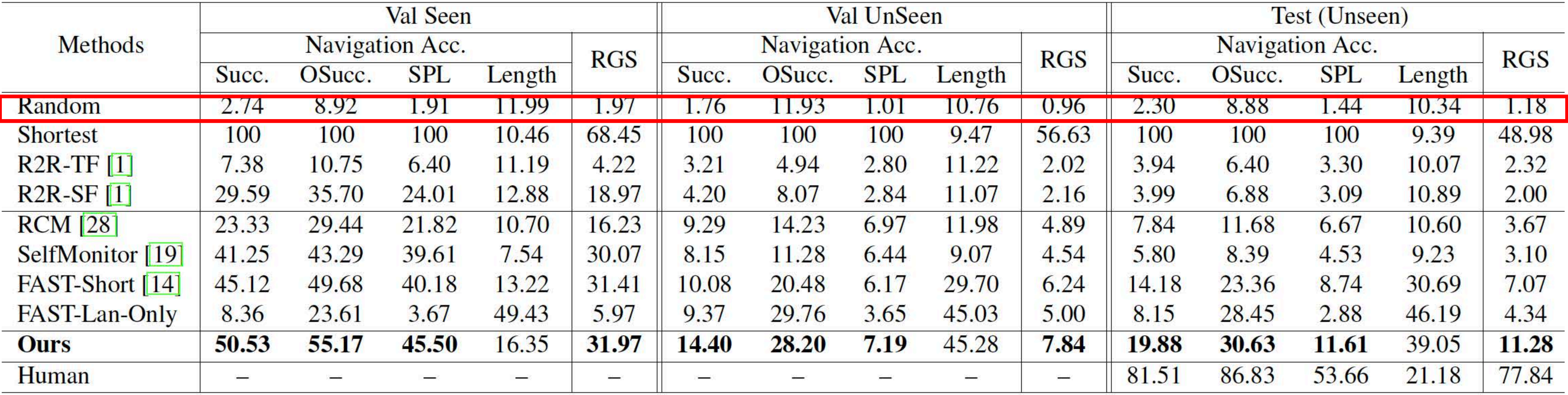
  - IoU >=0.5 between predicted bounding box and ground-truth

- Main Metric
  - RGS: Remote Grounding Success rate $\frac{Num_{succ}}{Num_{total}}\text{x}100\%$

- Auxiliary Metric for Navigation
  - Succ: Success rate
  - Osucc: Oracle success rate
  - Length: Path length
  - SPL: Success rate weighted by path length

Results

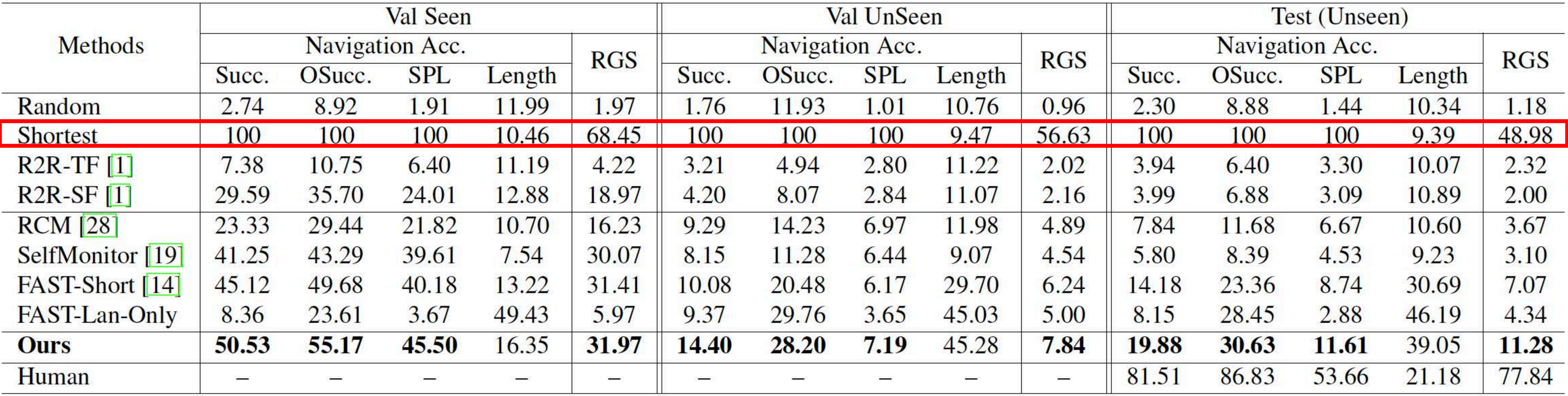
Success Rate on the REVERIE Task Using MAttNet as Pointer



Methods	Val Seen					Val UnSeen					Test (Unseen)				
	Navigation Acc.				RGS	Navigation Acc.				RGS	Navigation Acc.				RGS
	Succ.	OSucc.	SPL	Length		Succ.	OSucc.	SPL	Length		Succ.	OSucc.	SPL	Length	
Random	2.74	8.92	1.91	11.99	1.97	1.76	11.93	1.01	10.76	0.96	2.30	8.88	1.44	10.34	1.18
Shortest	100	100	100	10.46	68.45	100	100	100	9.47	56.63	100	100	100	9.39	48.98
R2R-TF [1]	7.38	10.75	6.40	11.19	4.22	3.21	4.94	2.80	11.22	2.02	3.94	6.40	3.30	10.07	2.32
R2R-SF [1]	29.59	35.70	24.01	12.88	18.97	4.20	8.07	2.84	11.07	2.16	3.99	6.88	3.09	10.89	2.00
RCM [28]	23.33	29.44	21.82	10.70	16.23	9.29	14.23	6.97	11.98	4.89	7.84	11.68	6.67	10.60	3.67
SelfMonitor [19]	41.25	43.29	39.61	7.54	30.07	8.15	11.28	6.44	9.07	4.54	5.80	8.39	4.53	9.23	3.10
FAST-Short [14]	45.12	49.68	40.18	13.22	31.41	10.08	20.48	6.17	29.70	6.24	14.18	23.36	8.74	30.69	7.07
FAST-Lan-Only	8.36	23.61	3.67	49.43	5.97	9.37	29.76	3.65	45.03	5.00	8.15	28.45	2.88	46.19	4.34
Ours	**50.53**	**55.17**	**45.50**	16.35	**31.97**	**14.40**	**28.20**	**7.19**	45.28	**7.84**	**19.88**	**30.63**	**11.61**	39.05	**11.28**
Human	–	–	–	–	–	–	–	–	–	–	81.51	86.83	53.66	21.18	77.84

Results

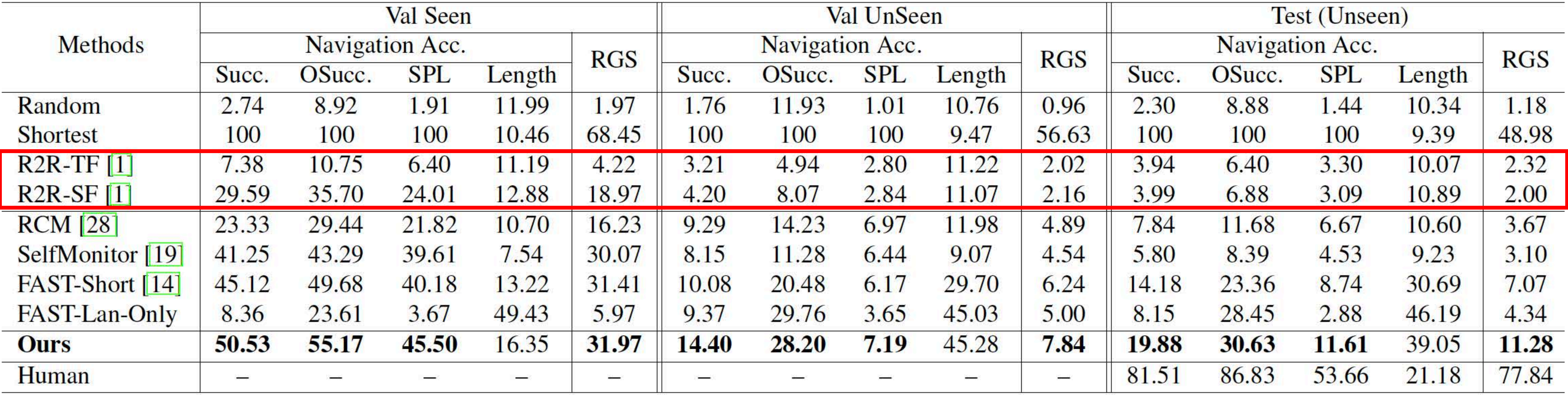
Success Rate on the REVERIE Task Using MAttNet as Pointer



Methods	Val Seen					Val UnSeen					Test (Unseen)				
	Navigation Acc.				RGS	Navigation Acc.				RGS	Navigation Acc.				RGS
	Succ.	OSucc.	SPL	Length		Succ.	OSucc.	SPL	Length		Succ.	OSucc.	SPL	Length	
Random	2.74	8.92	1.91	11.99	1.97	1.76	11.93	1.01	10.76	0.96	2.30	8.88	1.44	10.34	1.18
Shortest	100	100	100	10.46	68.45	100	100	100	9.47	56.63	100	100	100	9.39	48.98
R2R-TF [1]	7.38	10.75	6.40	11.19	4.22	3.21	4.94	2.80	11.22	2.02	3.94	6.40	3.30	10.07	2.32
R2R-SF [1]	29.59	35.70	24.01	12.88	18.97	4.20	8.07	2.84	11.07	2.16	3.99	6.88	3.09	10.89	2.00
RCM [28]	23.33	29.44	21.82	10.70	16.23	9.29	14.23	6.97	11.98	4.89	7.84	11.68	6.67	10.60	3.67
SelfMonitor [19]	41.25	43.29	39.61	7.54	30.07	8.15	11.28	6.44	9.07	4.54	5.80	8.39	4.53	9.23	3.10
FAST-Short [14]	45.12	49.68	40.18	13.22	31.41	10.08	20.48	6.17	29.70	6.24	14.18	23.36	8.74	30.69	7.07
FAST-Lan-Only	8.36	23.61	3.67	49.43	5.97	9.37	29.76	3.65	45.03	5.00	8.15	28.45	2.88	46.19	4.34
Ours	**50.53**	**55.17**	**45.50**	16.35	**31.97**	**14.40**	**28.20**	**7.19**	45.28	**7.84**	**19.88**	**30.63**	**11.61**	39.05	**11.28**
Human	–	–	–	–	–	–	–	–	–	–	81.51	86.83	53.66	21.18	77.84

Results

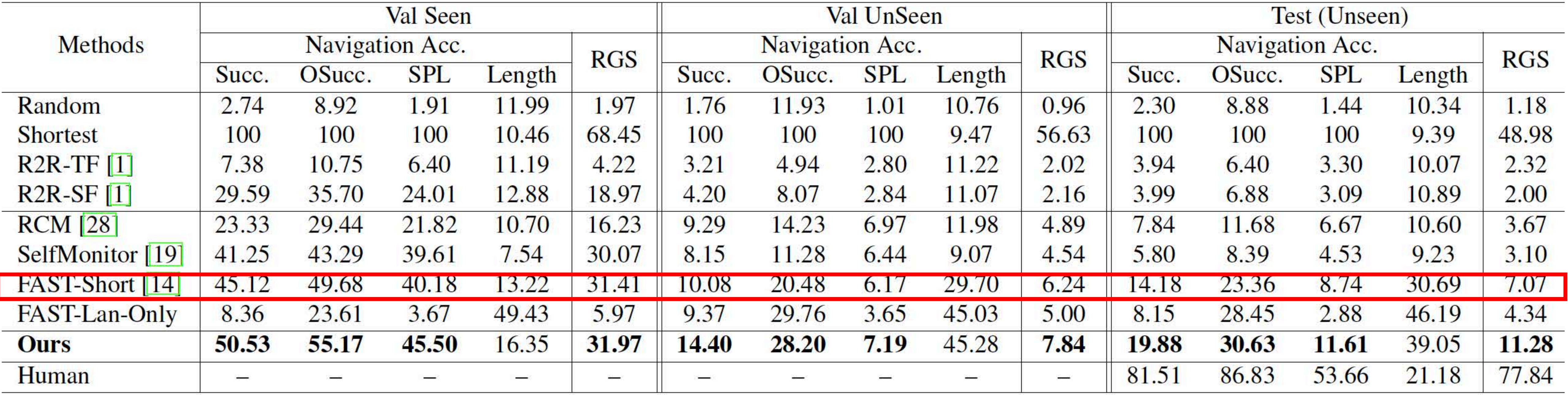
Success Rate on the REVERIE Task Using MAttNet as Pointer



Methods	Val Seen					Val UnSeen					Test (Unseen)				
	Navigation Acc.				RGS	Navigation Acc.				RGS	Navigation Acc.				RGS
	Succ.	OSucc.	SPL	Length		Succ.	OSucc.	SPL	Length		Succ.	OSucc.	SPL	Length	
Random	2.74	8.92	1.91	11.99	1.97	1.76	11.93	1.01	10.76	0.96	2.30	8.88	1.44	10.34	1.18
Shortest	100	100	100	10.46	68.45	100	100	100	9.47	56.63	100	100	100	9.39	48.98
R2R-TF [1]	7.38	10.75	6.40	11.19	4.22	3.21	4.94	2.80	11.22	2.02	3.94	6.40	3.30	10.07	2.32
R2R-SF [1]	29.59	35.70	24.01	12.88	18.97	4.20	8.07	2.84	11.07	2.16	3.99	6.88	3.09	10.89	2.00
RCM [28]	23.33	29.44	21.82	10.70	16.23	9.29	14.23	6.97	11.98	4.89	7.84	11.68	6.67	10.60	3.67
SelfMonitor [19]	41.25	43.29	39.61	7.54	30.07	8.15	11.28	6.44	9.07	4.54	5.80	8.39	4.53	9.23	3.10
FAST-Short [14]	45.12	49.68	40.18	13.22	31.41	10.08	20.48	6.17	29.70	6.24	14.18	23.36	8.74	30.69	7.07
FAST-Lan-Only	8.36	23.61	3.67	49.43	5.97	9.37	29.76	3.65	45.03	5.00	8.15	28.45	2.88	46.19	4.34
Ours	**50.53**	**55.17**	**45.50**	16.35	**31.97**	**14.40**	**28.20**	**7.19**	45.28	**7.84**	**19.88**	**30.63**	**11.61**	39.05	**11.28**
Human	–	–	–	–	–	–	–	–	–	–	81.51	86.83	53.66	21.18	77.84

Results

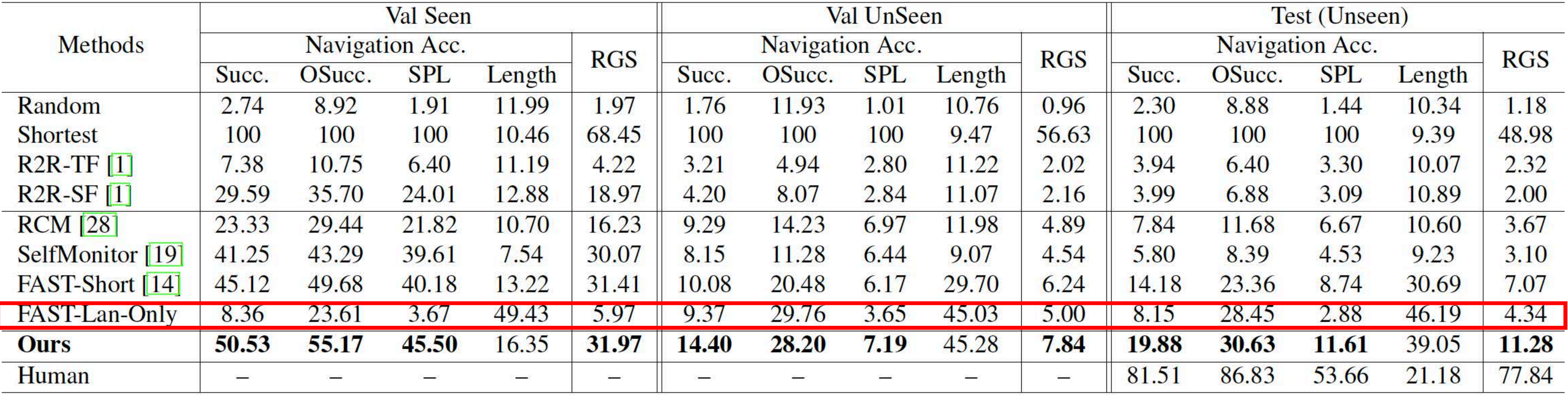
Success Rate on the REVERIE Task Using MAttNet as Pointer



Methods	Val Seen					Val UnSeen					Test (Unseen)				
	Navigation Acc.				RGS	Navigation Acc.				RGS	Navigation Acc.				RGS
	Succ.	OSucc.	SPL	Length		Succ.	OSucc.	SPL	Length		Succ.	OSucc.	SPL	Length	
Random	2.74	8.92	1.91	11.99	1.97	1.76	11.93	1.01	10.76	0.96	2.30	8.88	1.44	10.34	1.18
Shortest	100	100	100	10.46	68.45	100	100	100	9.47	56.63	100	100	100	9.39	48.98
R2R-TF [1]	7.38	10.75	6.40	11.19	4.22	3.21	4.94	2.80	11.22	2.02	3.94	6.40	3.30	10.07	2.32
R2R-SF [1]	29.59	35.70	24.01	12.88	18.97	4.20	8.07	2.84	11.07	2.16	3.99	6.88	3.09	10.89	2.00
RCM [28]	23.33	29.44	21.82	10.70	16.23	9.29	14.23	6.97	11.98	4.89	7.84	11.68	6.67	10.60	3.67
SelfMonitor [19]	41.25	43.29	39.61	7.54	30.07	8.15	11.28	6.44	9.07	4.54	5.80	8.39	4.53	9.23	3.10
FAST-Short [14]	45.12	49.68	40.18	13.22	31.41	10.08	20.48	6.17	29.70	6.24	14.18	23.36	8.74	30.69	7.07
FAST-Lan-Only	8.36	23.61	3.67	49.43	5.97	9.37	29.76	3.65	45.03	5.00	8.15	28.45	2.88	46.19	4.34
Ours	**50.53**	**55.17**	**45.50**	16.35	**31.97**	**14.40**	**28.20**	**7.19**	45.28	**7.84**	**19.88**	**30.63**	**11.61**	39.05	**11.28**
Human	–	–	–	–	–	–	–	–	–	–	81.51	86.83	53.66	21.18	77.84

Results

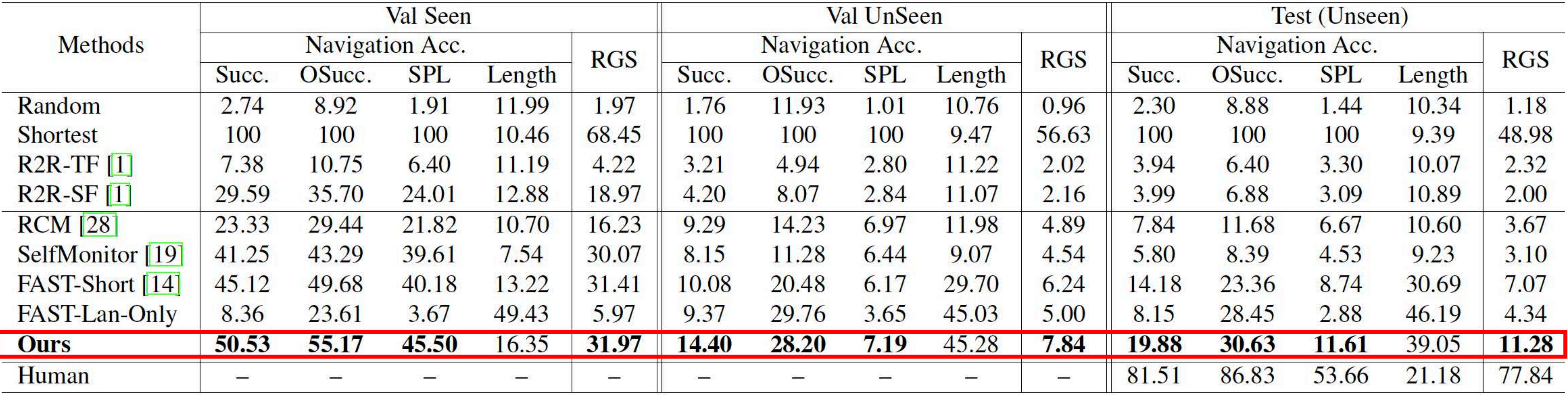
Success Rate on the REVERIE Task Using MAttNet as Pointer



Methods	Val Seen					Val UnSeen					Test (Unseen)				
	Navigation Acc.				RGS	Navigation Acc.				RGS	Navigation Acc.				RGS
	Succ.	OSucc.	SPL	Length		Succ.	OSucc.	SPL	Length		Succ.	OSucc.	SPL	Length	
Random	2.74	8.92	1.91	11.99	1.97	1.76	11.93	1.01	10.76	0.96	2.30	8.88	1.44	10.34	1.18
Shortest	100	100	100	10.46	68.45	100	100	100	9.47	56.63	100	100	100	9.39	48.98
R2R-TF [1]	7.38	10.75	6.40	11.19	4.22	3.21	4.94	2.80	11.22	2.02	3.94	6.40	3.30	10.07	2.32
R2R-SF [1]	29.59	35.70	24.01	12.88	18.97	4.20	8.07	2.84	11.07	2.16	3.99	6.88	3.09	10.89	2.00
RCM [28]	23.33	29.44	21.82	10.70	16.23	9.29	14.23	6.97	11.98	4.89	7.84	11.68	6.67	10.60	3.67
SelfMonitor [19]	41.25	43.29	39.61	7.54	30.07	8.15	11.28	6.44	9.07	4.54	5.80	8.39	4.53	9.23	3.10
FAST-Short [14]	45.12	49.68	40.18	13.22	31.41	10.08	20.48	6.17	29.70	6.24	14.18	23.36	8.74	30.69	7.07
FAST-Lan-Only	8.36	23.61	3.67	49.43	5.97	9.37	29.76	3.65	45.03	5.00	8.15	28.45	2.88	46.19	4.34
Ours	**50.53**	**55.17**	**45.50**	16.35	**31.97**	**14.40**	**28.20**	**7.19**	45.28	**7.84**	**19.88**	**30.63**	**11.61**	39.05	**11.28**
Human	–	–	–	–	–	–	–	–	–	–	81.51	86.83	53.66	21.18	77.84

Results

Success Rate on the REVERIE Task Using MAttNet as Pointer



Methods	Val Seen					Val UnSeen					Test (Unseen)				
	Navigation Acc.				RGS	Navigation Acc.				RGS	Navigation Acc.				RGS
	Succ.	OSucc.	SPL	Length		Succ.	OSucc.	SPL	Length		Succ.	OSucc.	SPL	Length	
Random	2.74	8.92	1.91	11.99	1.97	1.76	11.93	1.01	10.76	0.96	2.30	8.88	1.44	10.34	1.18
Shortest	100	100	100	10.46	68.45	100	100	100	9.47	56.63	100	100	100	9.39	48.98
R2R-TF [1]	7.38	10.75	6.40	11.19	4.22	3.21	4.94	2.80	11.22	2.02	3.94	6.40	3.30	10.07	2.32
R2R-SF [1]	29.59	35.70	24.01	12.88	18.97	4.20	8.07	2.84	11.07	2.16	3.99	6.88	3.09	10.89	2.00
RCM [28]	23.33	29.44	21.82	10.70	16.23	9.29	14.23	6.97	11.98	4.89	7.84	11.68	6.67	10.60	3.67
SelfMonitor [19]	41.25	43.29	39.61	7.54	30.07	8.15	11.28	6.44	9.07	4.54	5.80	8.39	4.53	9.23	3.10
FAST-Short [14]	45.12	49.68	40.18	13.22	31.41	10.08	20.48	6.17	29.70	6.24	14.18	23.36	8.74	30.69	7.07
FAST-Lan-Only	8.36	23.61	3.67	49.43	5.97	9.37	29.76	3.65	45.03	5.00	8.15	28.45	2.88	46.19	4.34
Ours	**50.53**	**55.17**	**45.50**	16.35	**31.97**	**14.40**	**28.20**	**7.19**	45.28	**7.84**	**19.88**	**30.63**	**11.61**	39.05	**11.28**
Human	–	–	–	–	–	–	–	–	–	–	81.51	86.83	53.66	21.18	77.84

Take Home Message

REVERIE Challenge @ ACL 2020 Workshop

Code and Dataset