GeoVLN: Learning Geometry-Enhanced Visual Representation with Slot Attention for Vision-and-Language Navigation

Jingyang Huo* Qiang Sun* Boyan Jiang* Haitao Lin Yanwei Fu
Fudan University

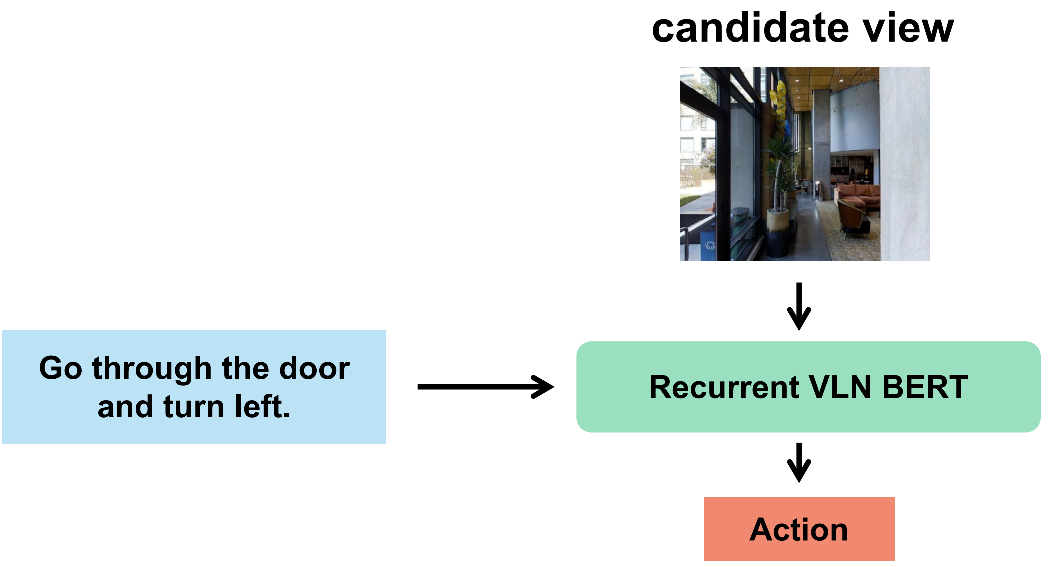
THU-PM-249

* Indicates equal contributions.

Overview

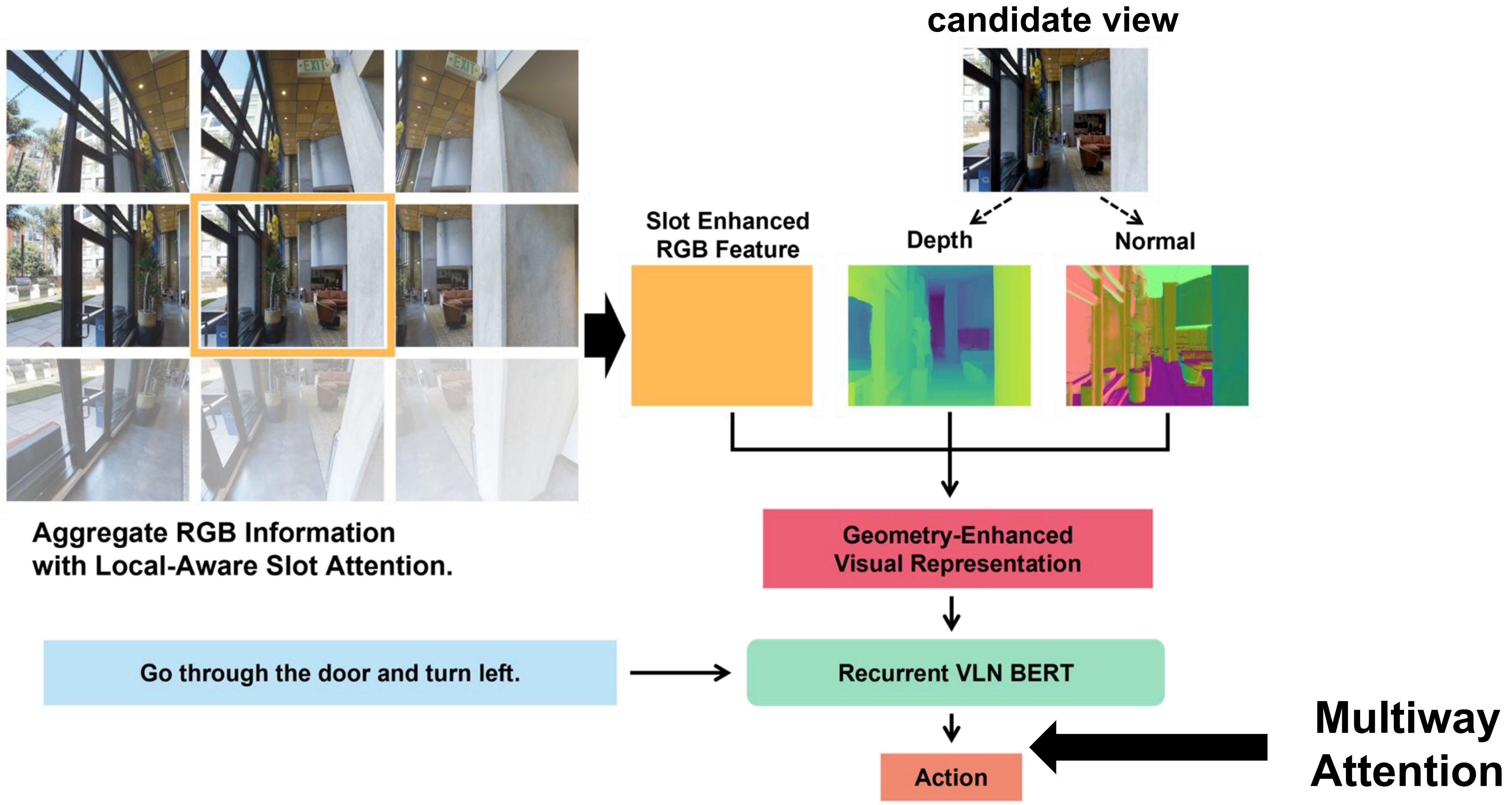
Previous Work

- ➢ Utilize RGB images only;
- ➢ Lack local spatial context around candidate view.



Previous Work

- ➢ Compensate RGB images with depth maps and normal maps estimated with Omnidata;
- ➢ Learn geometry-enhanced visual representation with a two-stage slot-based module;
- ➢ Encourage different phrases of input instruction to focus on the most informative visual observation (e.g. texture, depth) with the multiway attention module.

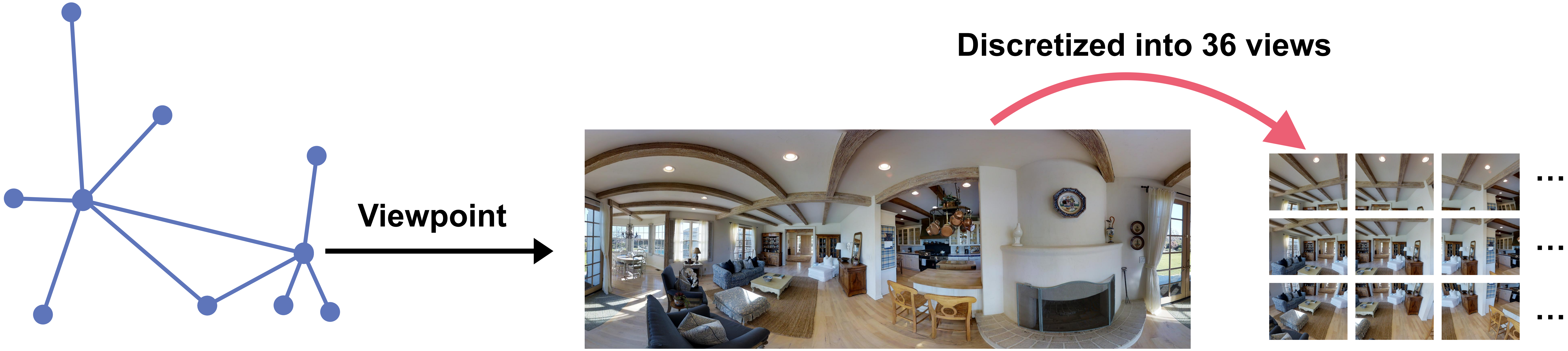


VLN Task

Vision-and-Language Navigation

Given a natural language instruction, agent makes decision about the next move automatically based on past and current visual observations.

Room-to-Room Navigation Environment



Pipeline

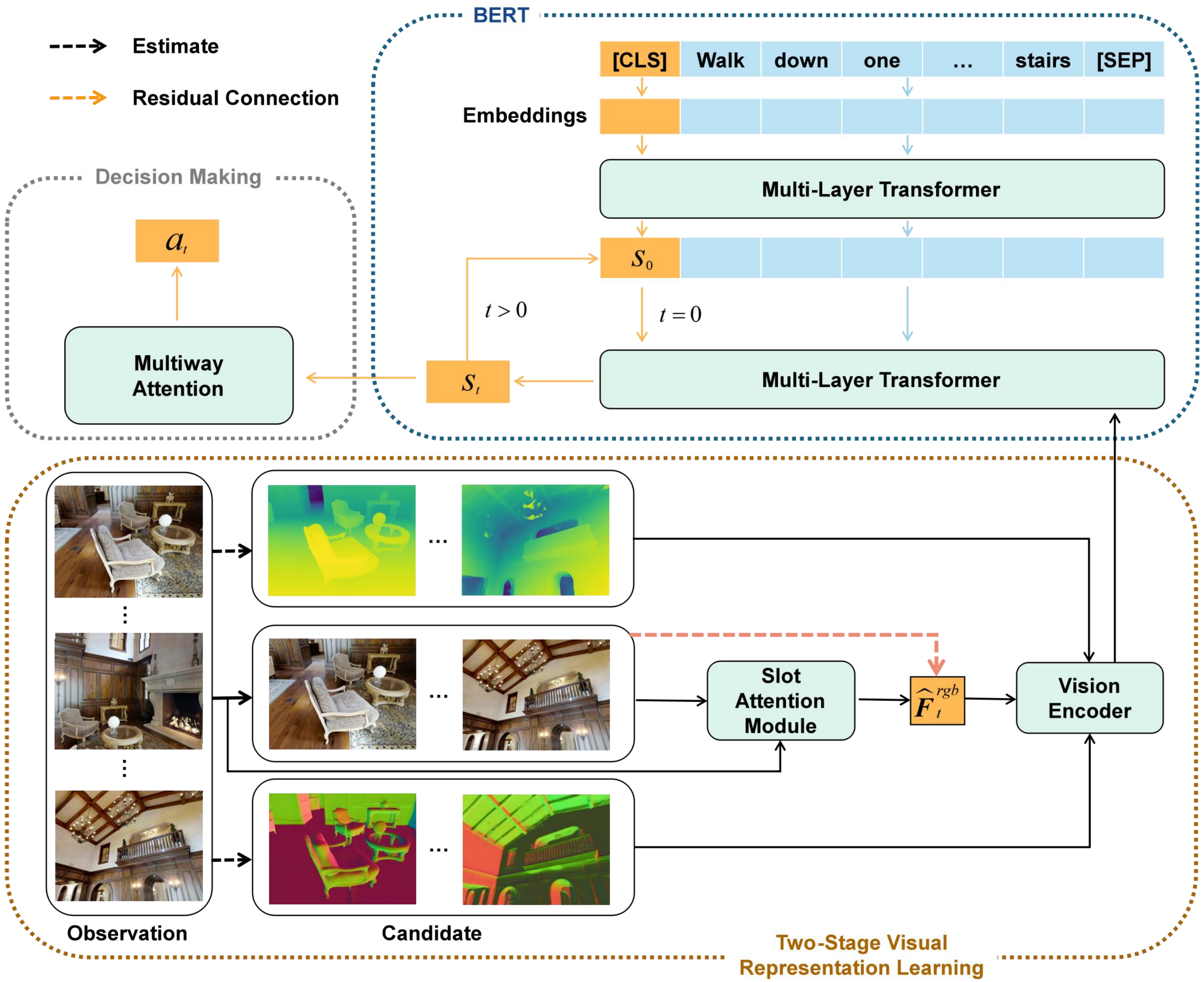
Inputs:

- Language Inputs (a user instruction)
- Visual Inputs (a set of visual observations)

BERT:
Follow BERT encoder[1] to process the language input and obtain the state vector.

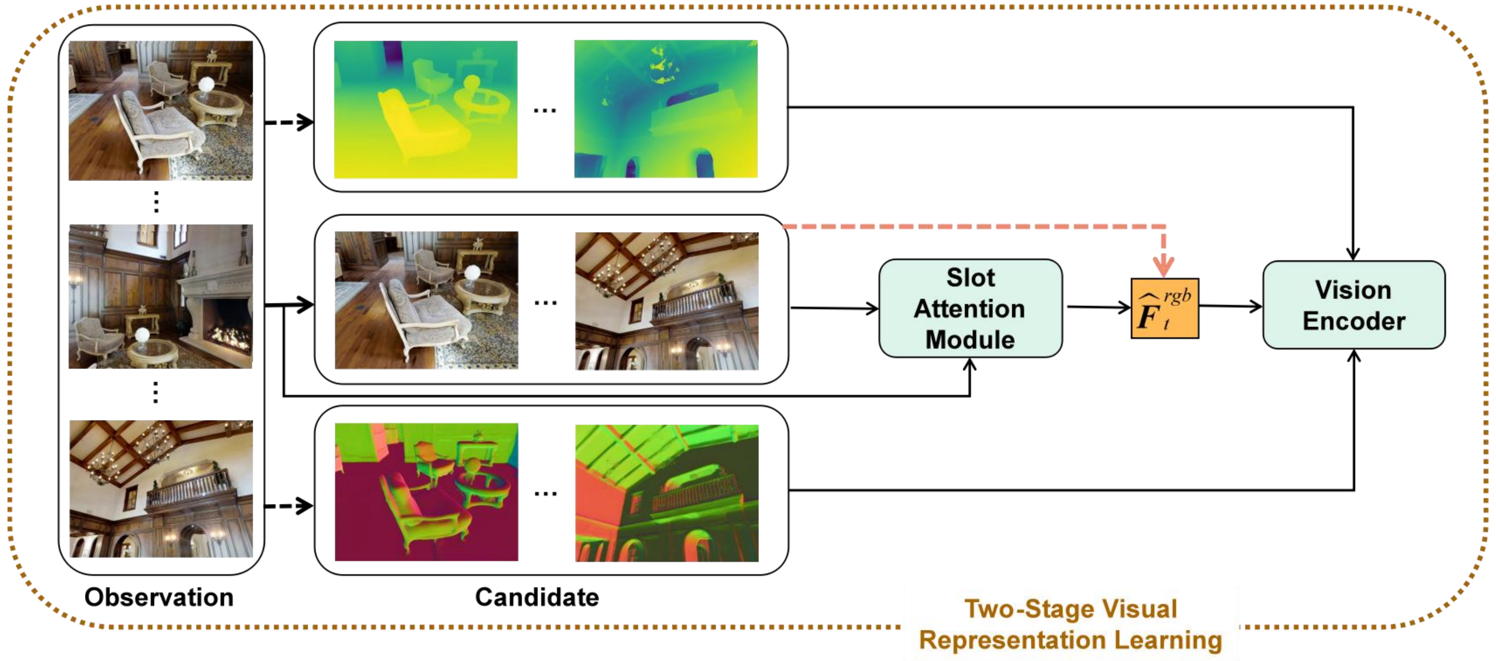
Two-Stage Visual Representation Learning:
Compensate RGB images with depth maps and normal maps estimated with Omnidata;
Learn geometry-enhanced visual representation with a two-stage slot-based module.

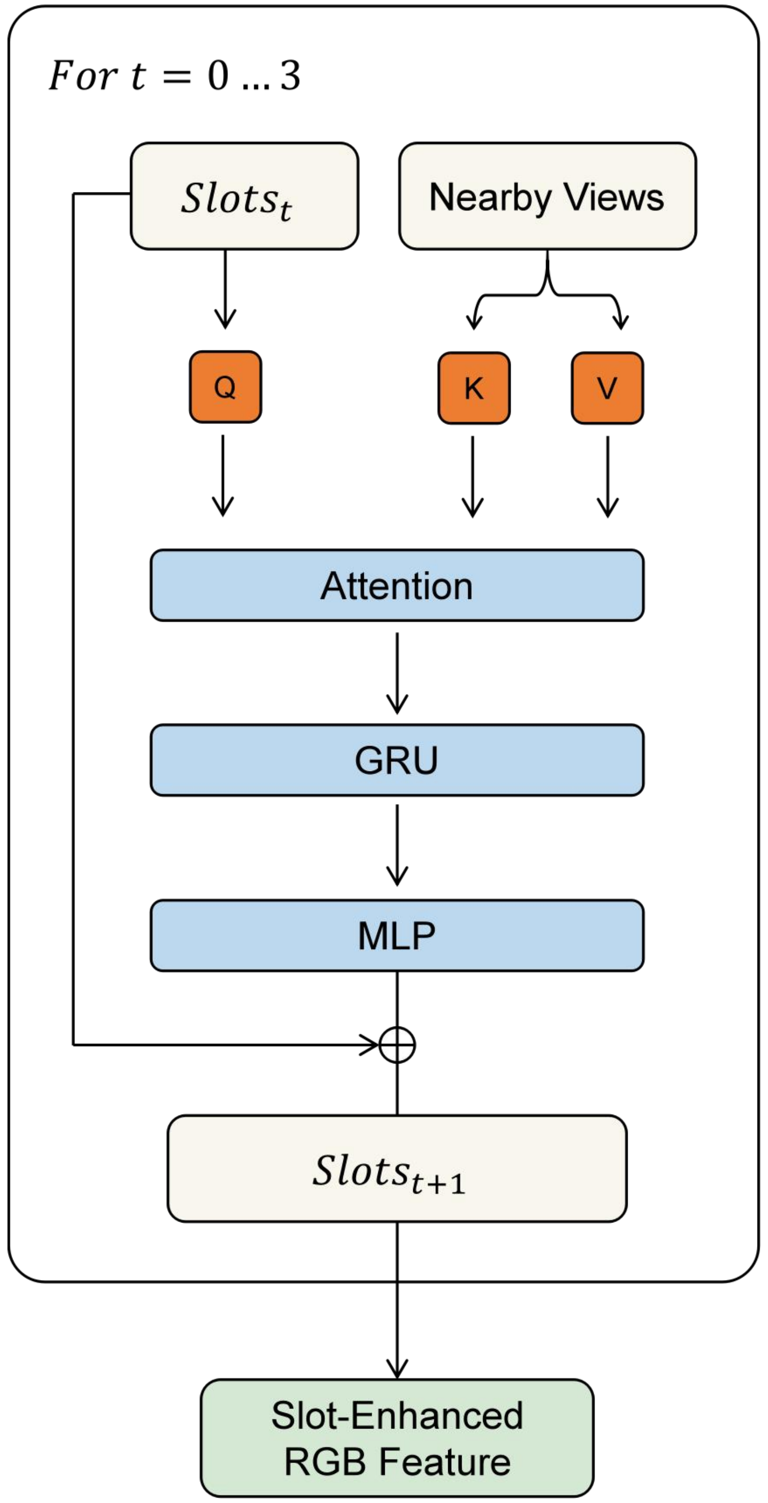
Decision Making:
Weight matching scores of the three modalities with the multiway attention module.



[1] Yicong Hong, Qi Wu, Yuankai Qi, Cristian Rodriguez Opazo, and Stephen Gould. Vln bert: A recurrent vision-and-language bert for navigation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 1643–1653, 2021.

Two-Stage Visual Representation Learning





Visual Features

$$\boldsymbol{F}_t^* = [\boldsymbol{C}_t^*; \boldsymbol{F}_t^{ang}], \quad * \in [rgb, dep, nor]$$

Local-Aware Slot Attention

Make each candidate views aggregate information from the nearby observation views according to the spatial proximity principle.

Geometry-enhanced Visual Representation

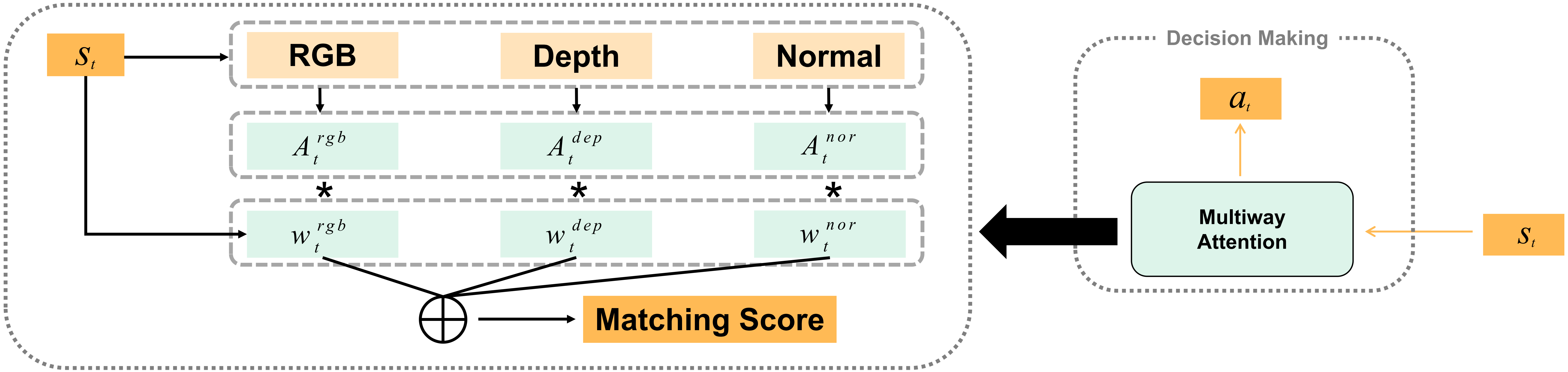
$$\boldsymbol{F}_t = \left[\hat{\boldsymbol{F}}_t^{rgb}[\dots, : d_C]; \boldsymbol{C}_t^{dep}; \boldsymbol{C}_t^{nor}; \boldsymbol{F}_t^{ang}\right]$$

$$\hat{\boldsymbol{F}}_t = \mathrm{LN}(\mathrm{FC}(\boldsymbol{F}_t))$$

Multiway Attention

Dynamically weight matching scores of the three modalities with the multiway attention module.

$$\boldsymbol{score}_t^{total} = w_t^{rgb}\boldsymbol{A}_t^{rgb} + w_t^{dep}\boldsymbol{A}_t^{dep} + w_t^{nor}\boldsymbol{A}_t^{nor}$$



Main Results

Success Rate (SR): the ratio of agents eventually stopping within 3 meters of the destination;
Success Weighted by Path Length (SPL): SR weighted by the inverse of TL. A higher SPL score indicates a better balance between achieving the goal and taking the shortest path.

	Val Seen				Val Unseen				Test Unseen			
Agent	TL↓	NE↓	SR↑	SPL↑	TL↓	NE↓	SR↑	SPL↑	TL↓	NE↓	SR↑	SPL↑
RANDOM [2]	9.58	9.45	16	-	9.77	9.23	16	-	9.93	9.77	13	12
Human	-	-	-	-	-	-	-	-	11.85	1.61	86	76
Seq-to-Seq [2]	11.33	6.01	39	-	**8.39**	7.81	22	-	**8.13**	7.85	20	18
Speaker-Follower [8]	-	3.36	66	-	-	6.62	35	-	14.82	6.62	35	28
Self-Monitoring [9]	-	-	-	-	-	-	-	-	18.04	5.67	48	35
Reinforced Cross-Modal [32]	10.65	3.53	67	-	11.46	6.09	43	-	11.97	6.12	43	38
EnvDrop [30]	11.00	3.99	62	59	10.70	5.22	62	48	11.66	5.23	51	47
AuxRN [37]	-	3.33	70	67	-	5.28	55	50	-	5.15	55	51
PREVALENT [11]	**10.32**	3.67	69	65	10.19	4.71	58	53	10.51	5.30	54	51
PRESS [18]	10.35	3.09	71	67	10.06	4.31	59	55	10.52	4.53	57	53
AirBERT [10]	11.09	2.68	75	70	11.78	4.01	62	56	12.41	4.13	62	57
VLN↻BERT [15]	11.13	2.90	72	68	12.01	3.93	63	57	12.35	4.09	63	57
GeoVLN (Ours)	11.98	3.17	70	65	11.93	3.51	67	61	13.02	4.04	63	58
HAMT [4]	11.15	2.51	76	72	11.46	**2.29**	66	61	12.27	**3.93**	**65**	60
GeoVLN† (Ours)	10.68	**2.22**	**79**	**76**	11.29	3.35	**68**	**63**	12.16	3.95	**65**	**61**

Table 1. Comparison of **OUR MODEL** with the previous state-of-the-art methods on R2R dataset. † indicates the results with HAMT as the backbone. The primary metric is SPL.

Ablation Study

Model	Input			Val Seen		Val Unseen	
	RGB	DEPTH	NORMAL	SR↑	SPL↑	SR↑	SPL↑
Baseline	✓			**69.83**	64.21	64.50	58.35
Baseline	✓	✓		66.99	63.28	63.86	58.58
Baseline	✓		✓	68.46	63.66	62.71	57.20
Baseline	✓	✓	✓	66.41	62.51	64.75	59.31
LSA	✓			67.58	63.06	64.62	59.78
LSA	✓	✓	✓	68.66	63.92	66.54	60.62
LSA + MAtt	✓			68.46	63.92	66.11	60.31
LSA + MAtt (Full)	✓	✓	✓	69.64	**64.86**	**66.75**	**61.00**

Table 2. Ablation study on multi-modal visual inputs and LSA module with VLN ↻ BERT as the backbone.

Visualization

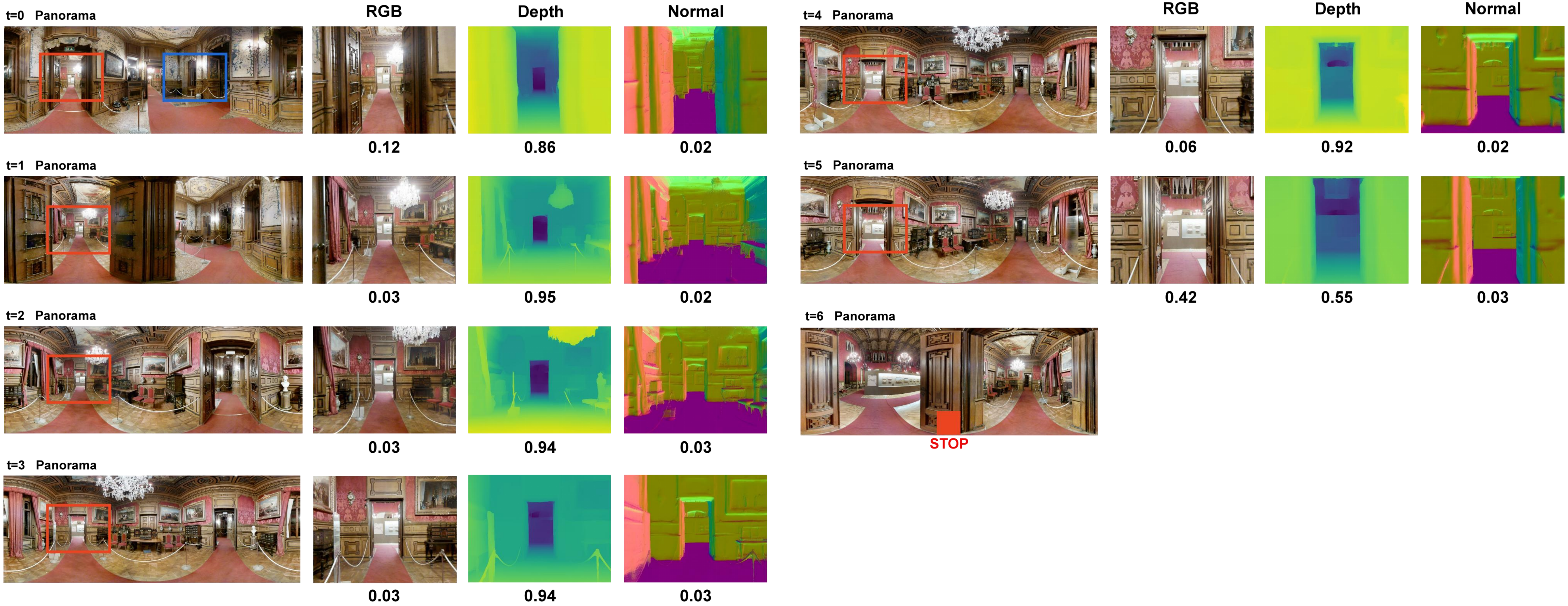
How the **local-aware slot attention** module aggregates local observations to candidate views.

Instruction: “Pass the pool then go into the …”.

Instruction: “Walk up stairs …”.

Visualization

How **multiway attention** facilitates decision-making.



Instruction: "Follow the red carpet through the double doors. Continue straight through the room and wait in the doorway with the double doors at the end."

GeoVLN: Learning Geometry-Enhanced Visual Representation with Slot Attention for Vision-and-Language Navigation

Jingyang Huo*   Qiang Sun*   Boyan Jiang*   Haitao Lin   Yanwei Fu
Fudan University

Thank you!

* Indicates equal contributions.