UNIVERSITY OF CAMBRIDGE
Department of Computer Science and Technology

TOSHIBA

LREC-COLING 2024

Semantic Map-based Generation of Navigation Instructions

Chengzu Li, Chao Zhang, Simone Teufel, Rama Sanand Doddipatla, Svetlana Stoyanchev

Department of Computer Science and Technology, University of Cambridge
Toshiba Europe Limited

Chengzu Li
cl917@cam.ac.uk
https://chengzu-li.github.io

Navigation Instruction Generation

Vision Language Navigation (VLN)

- Agent navigating in physical environment in response to natural language instructions
- Annotation is time-consuming

Vision Language Generation (VL-GEN)

- The reverse of VLN task: Path → Instructions
- Generated instructions are shown to be helpful in improving VLN system

Previous works employ sequence of panoramas to generate instructions

We frame the task as **top-down map image captioning**

LREC-COLING 2024

UNIVERSITY OF CAMBRIDGE
Department of Computer Science and Technology

TOSHIBA

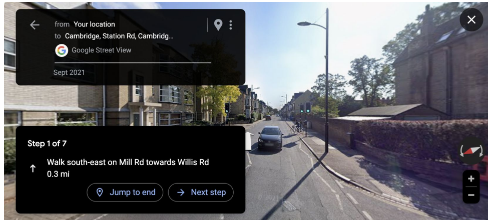
VL-GEN

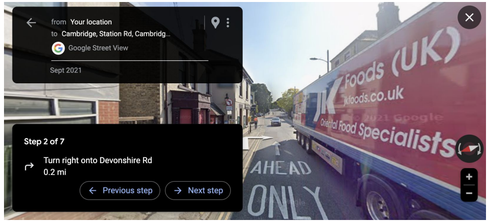
Instruction

- Walk south-east on Mill Rd towards Willis Rd
  0.3 mi
- Turn right onto Devonshire Rd
  0.2 mi
- Turn left towards Station Sq
  46 ft
- Turn left towards Station Sq
  66 ft
- Turn right onto Station Sq
  427 ft
- Slight right to stay on Station Sq
  246 ft
- Slight left
  72 ft

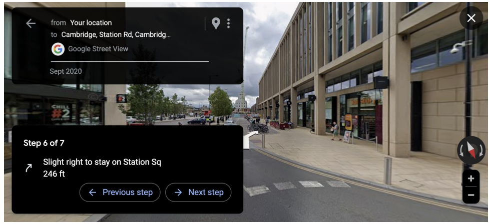
Cambridge
Station Rd, Cambridge CB1 2JW

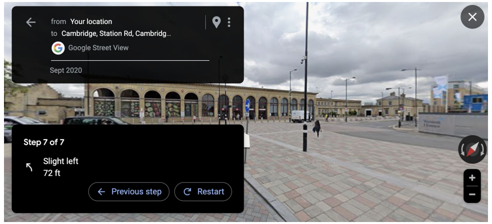
Existing Approach





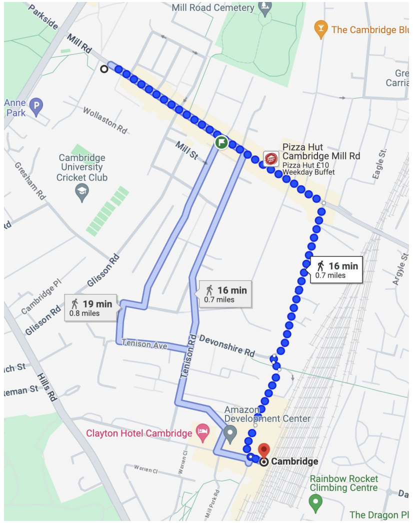
• • • • • •





Use a **sequence of panoramic images** as visual input

Our Approach

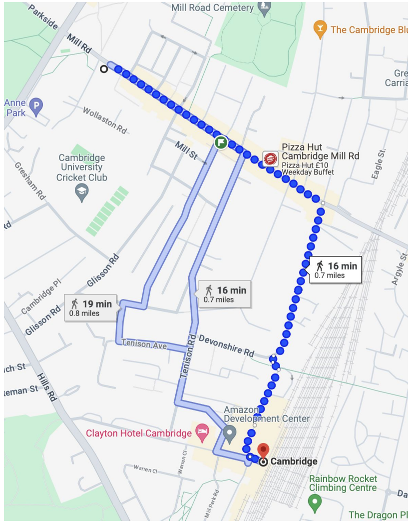


Use **single top-down map** as visual input

** All from Google Map*

UNIVERSITY OF CAMBRIDGE
Department of Computer Science and Technology
TOSHIBA

VL-GEN



Existing approaches

- Panoramic images as visual input
- Slightly better than template-based methods
- Limitations
  - Processing pano sequence is resource-intensive
  - Pano images contain too much task irrelevant details

Our approach

- It is natural to understand navigation instructions using top-down map (as in Google Maps)
- Only one image is required
- Only semantic information are kept

Task Definition: VL-GEN

Task

Input:

- a top-down semantic map M
- a path P={p_1,…,p_K}

Optional input: Panoramas, Regions, Actions.
Output: natural language description D

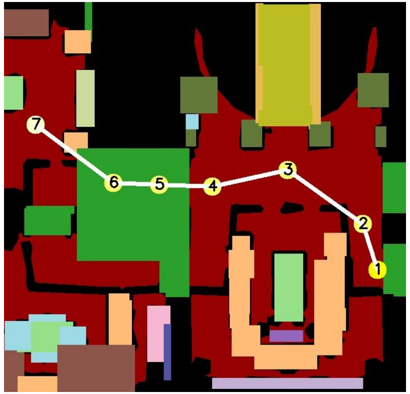
Data

Room2Room dataset with Habitat simulator

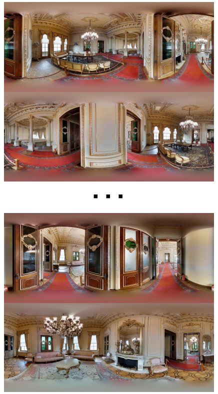
Evaluation

SPICE

- a metric used to evaluate the quality of image captions, focusing on the semantic content of captions



Top-down semantic map



Panoramas

Regions	*Actions*
meetingroom, hallway	straight
hallway	left
……	……
hallway	right
meetingroom	stop

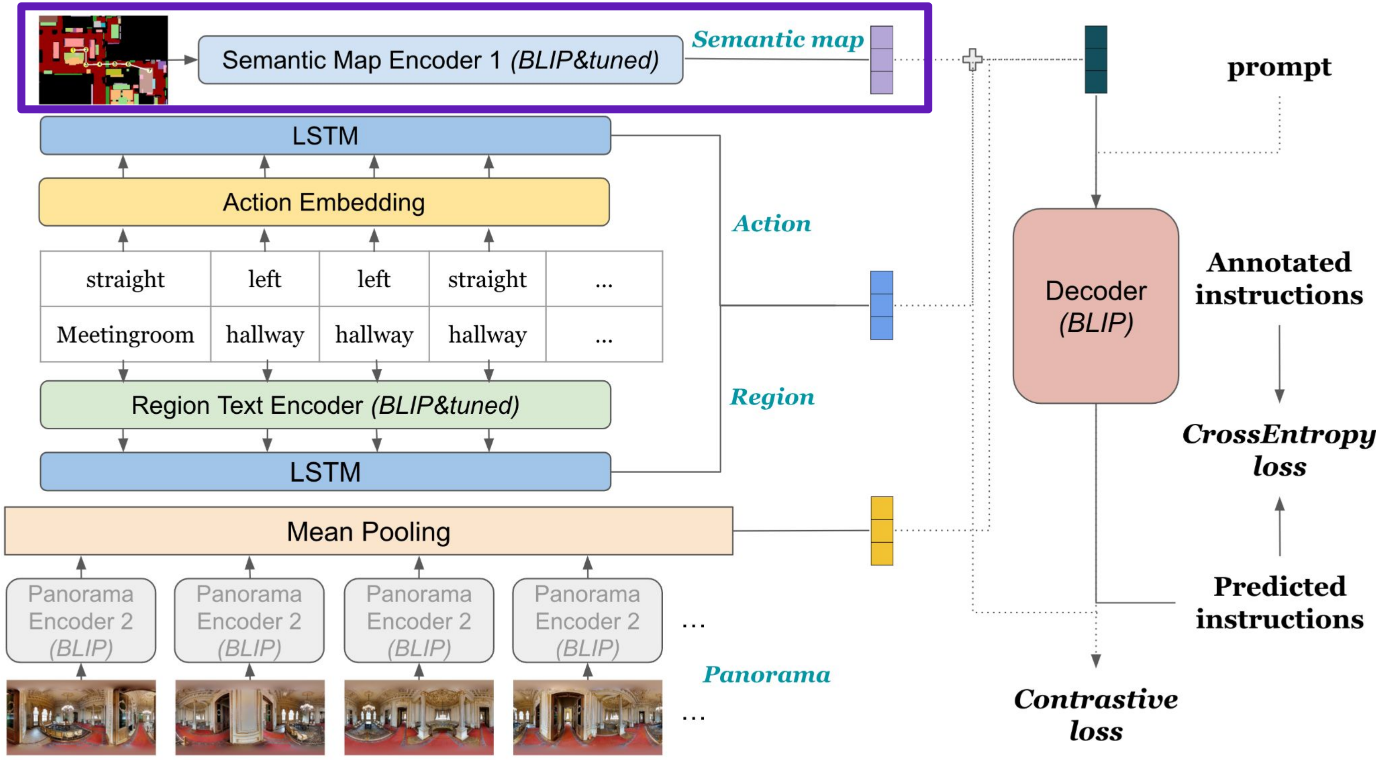
One of the Three Human Annotated Navigation Instructions:

- Turn left and follow the rope. At the end turn left and follow the red carpet to the end. At the end, turn right and stop in front of the white and gold table.

split	size	Avg. # points	Avg. # regions	Avg. # objects
train	10623	5.95	3.26	22.64
val seen	768	6.07	3.3	22.36
val unseen	1839	5.87	3.11	22.13

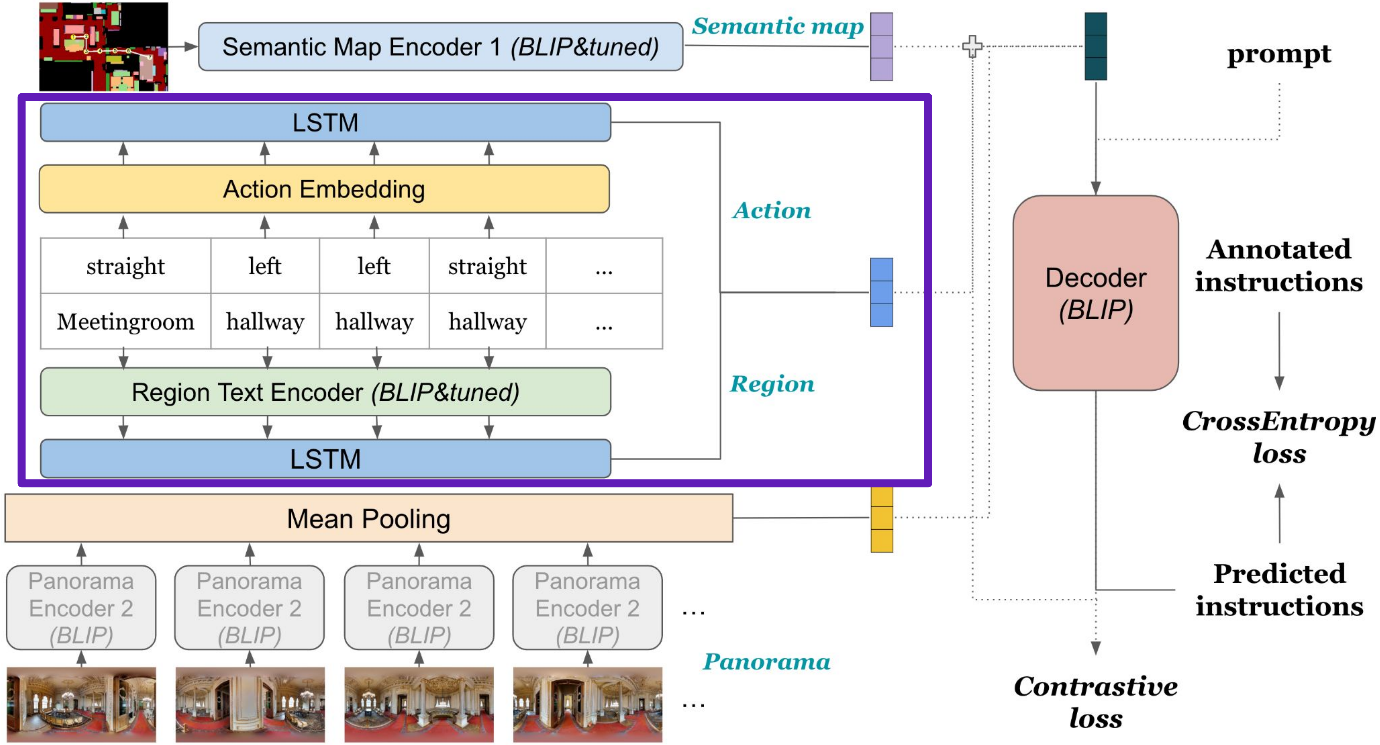
UNIVERSITY OF CAMBRIDGE
Department of Computer Science and Technology
TOSHIBA

Method



Top-down semantic map as the main input encoded by BLIP vision encoder

Method

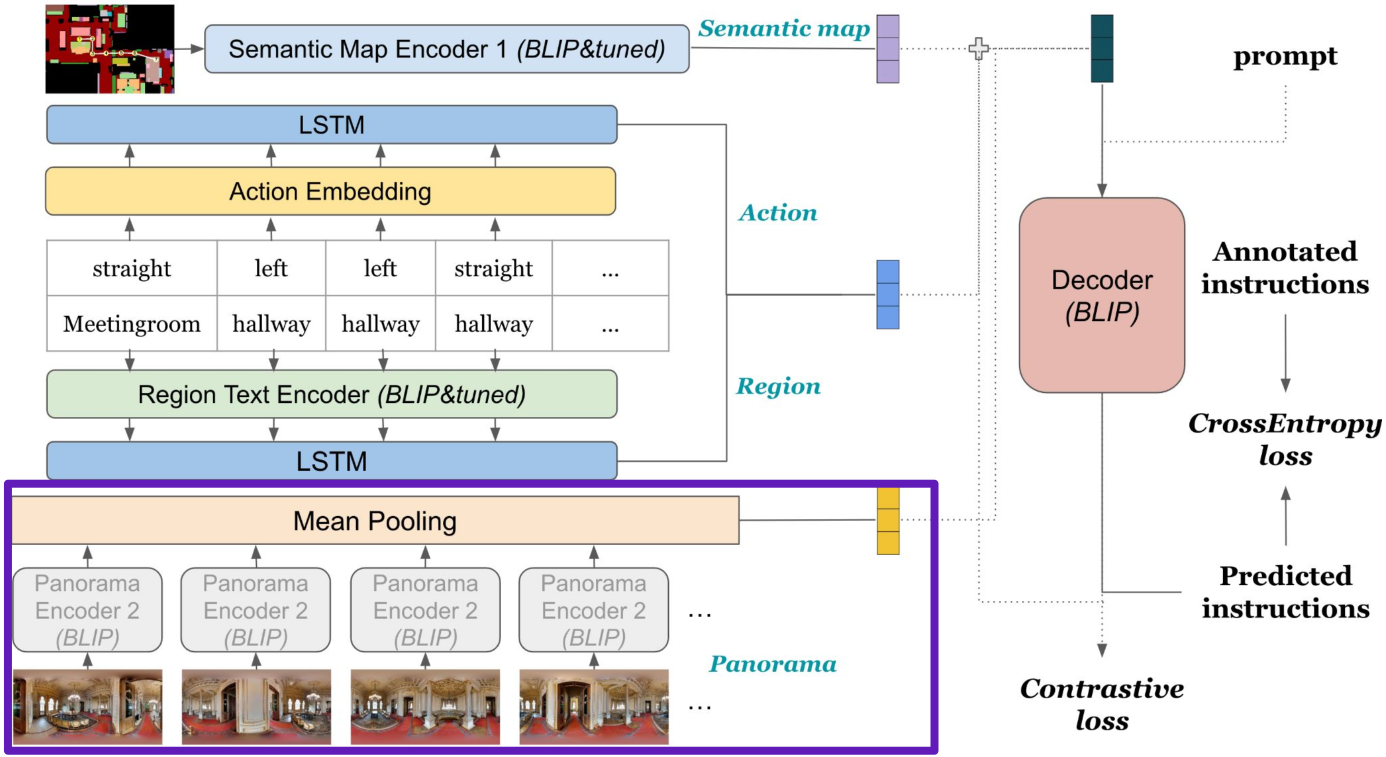


Regions and actions:

- frequently mentioned in the instruction.
- Embed per-point and fuse, followed by LSTM for sequential modeling.

UNIVERSITY OF CAMBRIDGE
Department of Computer Science and Technology
TOSHIBA

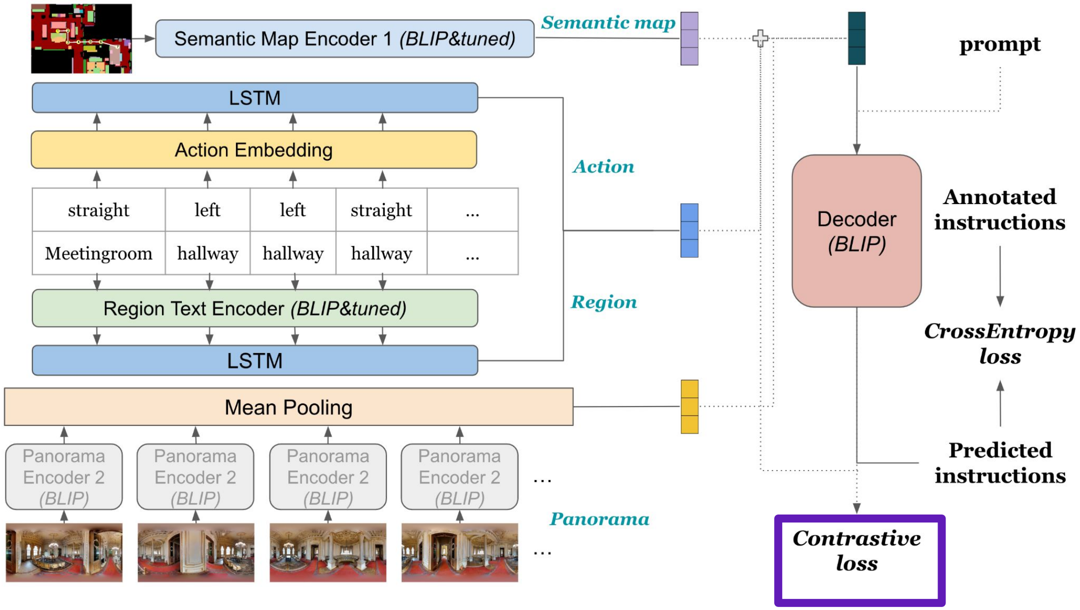
Method



Panoramic images:

- Objects' features (eg. colour and shape) mentioned in more than 25% of instructions

Model Augmentation

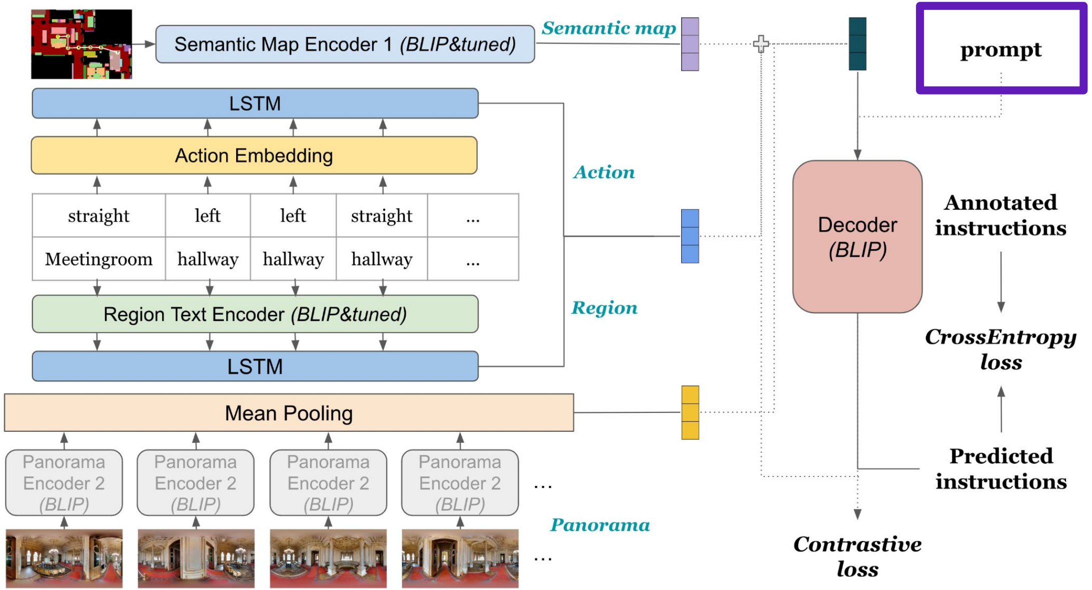


Multimodal alignment with contrastive loss

- Positive samples: input embedding and paired instruction
- Negative samples: input embedding and random Instruction

UNIVERSITY OF CAMBRIDGE
Department of Computer Science and Technology
TOSHIBA

Model Augmentation



Prompt for object grounding

- LLMs prompting is effective across various generation tasks
- We generate prompt describing the nearby objects and regions using template and tune the model with it
  - Starting from the dark yellow point near sofa cushion in the living room region.
- It helps visual-language grounding via explicit description

UNIVERSITY OF CAMBRIDGE
Department of Computer Science and Technology
TOSHIBA

Results

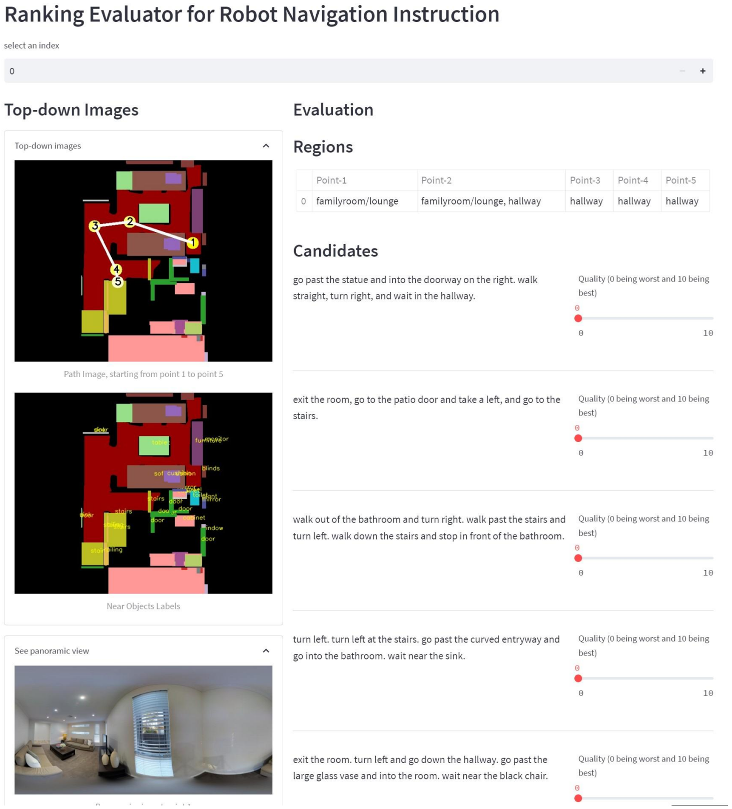
Input	P	C	SPICE		Human Score
			seen	unseen	unseen
TD (baseline)	-	-	20.50	16.19	3.42 (5)
	✓	-	20.79	15.77	-
	✓	✓	21.78*	17.10	-
TD+Reg+Act	-	-	21.00	17.00	4.20 (3)
	✓	-	21.86*	**17.84****	4.29 (2)
	✓	✓	19.96	17.09	3.98 (4)
TD+Reg+Act+Pano	-	-	19.87	17.44*	**4.36*** (1)
	✓	-	**22.14****	17.79**	-
	✓	✓	20.36	17.08	-

Automatic (SPICE) and human evaluation results with inputs of different modalities in seen and unseen environments.

- The models perform better in seen than in unseen setting on average.
- Using region and action information with the prompt improves the model's performance
- Our systems perform on par or even achieve higher SPICE scores than previous VL-GEN methods.

UNIVERSITY OF CAMBRIDGE
Department of Computer Science and Technology
TOSHIBA

Human Evaluation



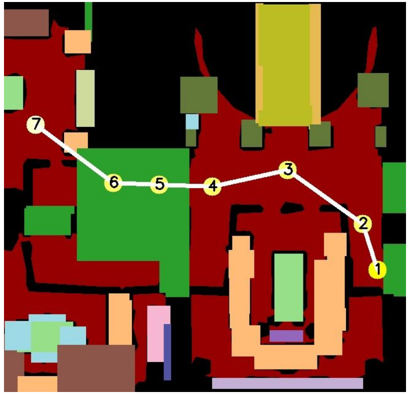
Input	P	C	SPICE seen	SPICE unseen	Human Score unseen
TD (baseline)	-	-	20.50	16.19	3.42 (5)
	✓	-	20.79	15.77	-
	✓	✓	21.78*	17.10	-
TD+Reg+Act	-	-	21.00	17.00	4.20 (3)
	✓	-	21.86*	**17.84****	4.29 (2)
	✓	✓	19.96	17.09	3.98 (4)
TD+Reg+Act+Pano	-	-	19.87	17.44*	**4.36*** (1)
	✓	-	**22.14****	17.79**	-
	✓	✓	20.36	17.08	-

Only using the semantic map as the baseline results in the lowest average score across all systems. Using regions, actions, and panoramas achieves the highest rating (4.36) which is significantly better than the baseline.

UNIVERSITY OF CAMBRIDGE
Department of Computer Science and Technology
TOSHIBA

Conclusion

- It is a human-interpretable and light-weight approach that encodes information necessary for the navigation in a single abstract top-down image
- We create the dataset with top-down semantic maps for R2R corpus and reframe instruction generation task as image captioning
- Top-down semantic map performs on-par with the end-to-end methods that use sequence of panorama images



Thanks

UNIVERSITY OF CAMBRIDGE
Department of Computer Science and Technology

TOSHIBA

LREC-COLING 2024

Chengzu Li
cl917@cam.ac.uk
https://chengzu-li.github.io