SOAT: A Scene- and Object-Aware Transformer for Vision-and-Language Navigation

NeurIPS 2021

Abhinav Moudgil[1]

Arjun Majumdar[1]

Harsh Agrawal[1]

Stefan Lee[2]

Dhruv Batra[1]

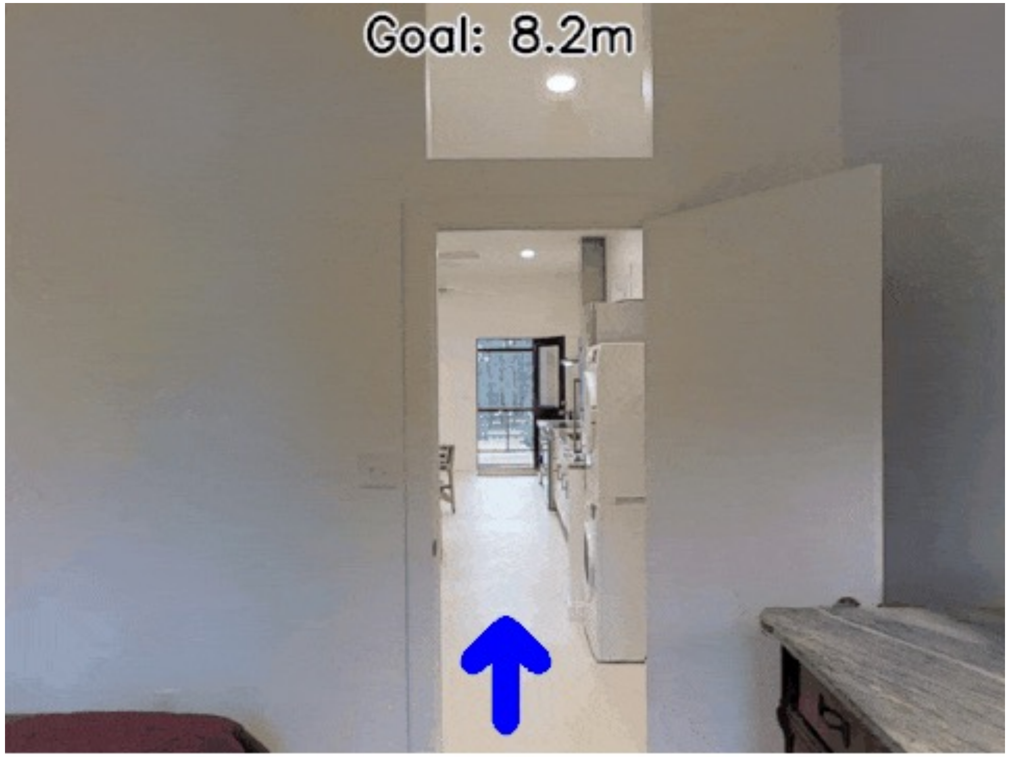
[1] Georgia Tech

[2] Oregon State University

Vision-and-Language Navigation (VLN) Task

Given a natural language instruction, the agent needs to navigate to a goal location by following the given instruction.

Input: *instruction and panoramic observation*
Output: *sequence of actions*



Leave the bedroom, and enter the kitchen. Walk forward, and take a left at the couch. Stop in front of the window.

Motivation

Instruction

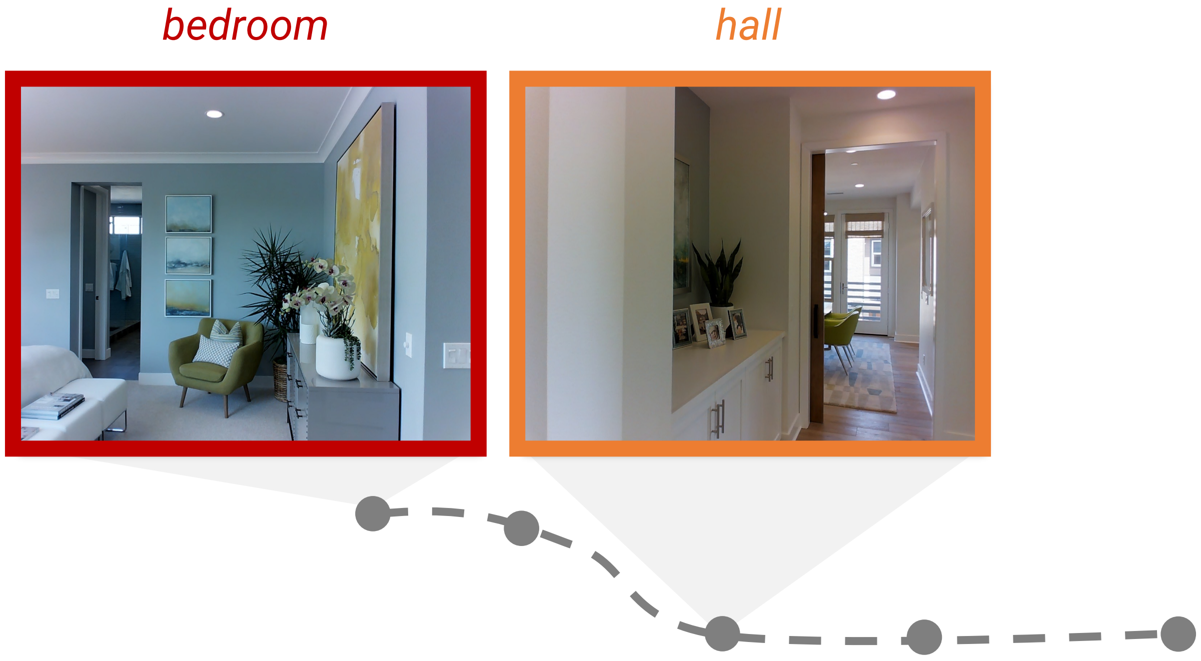
Exit the bedroom and turn left. Continue down the hall and into the room straight ahead and stop before the desk with two green chairs.

Motivation

Instruction

Exit the bedroom and turn left. Continue down the hall and into the room straight ahead and stop before the desk with two green chairs.

Scene Descriptions



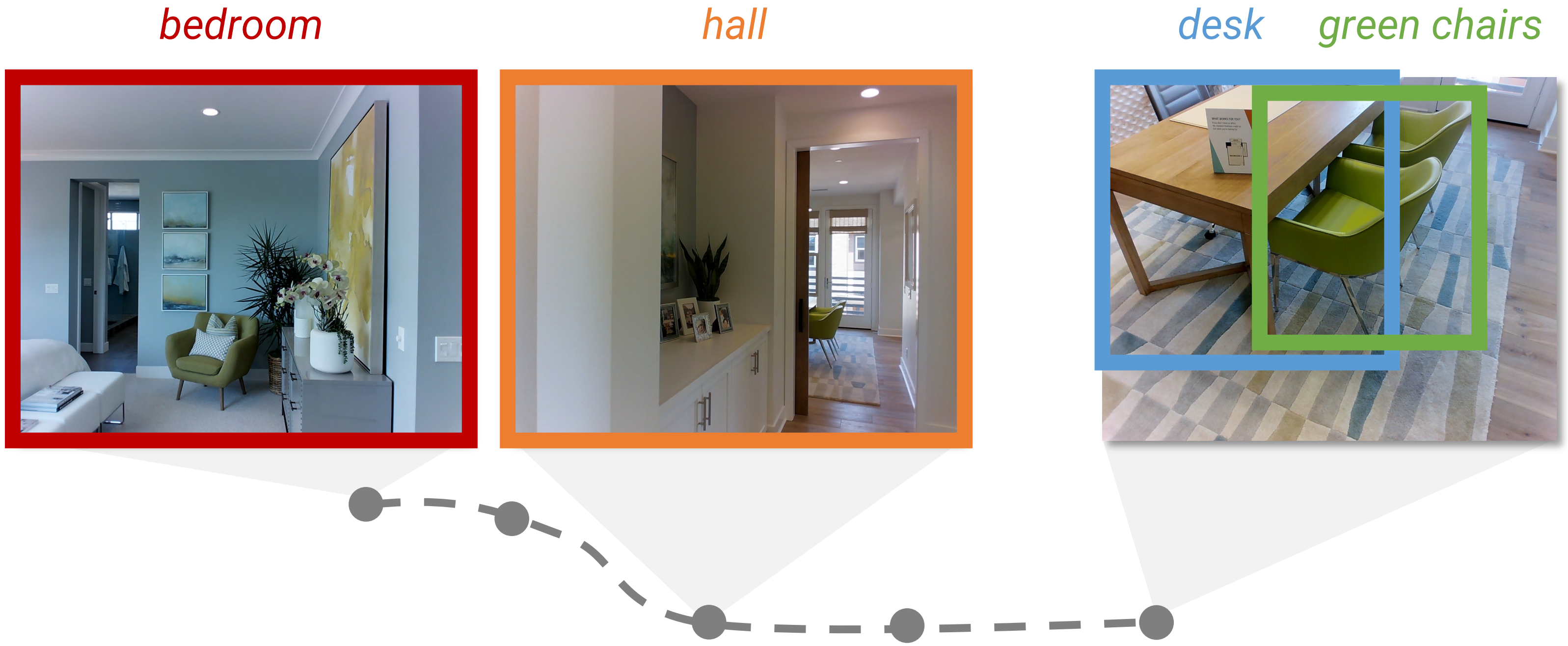
Motivation

Instruction

Exit the bedroom and turn left. Continue down the hall and into the room straight ahead and stop before the desk with two green chairs.

Scene Descriptions

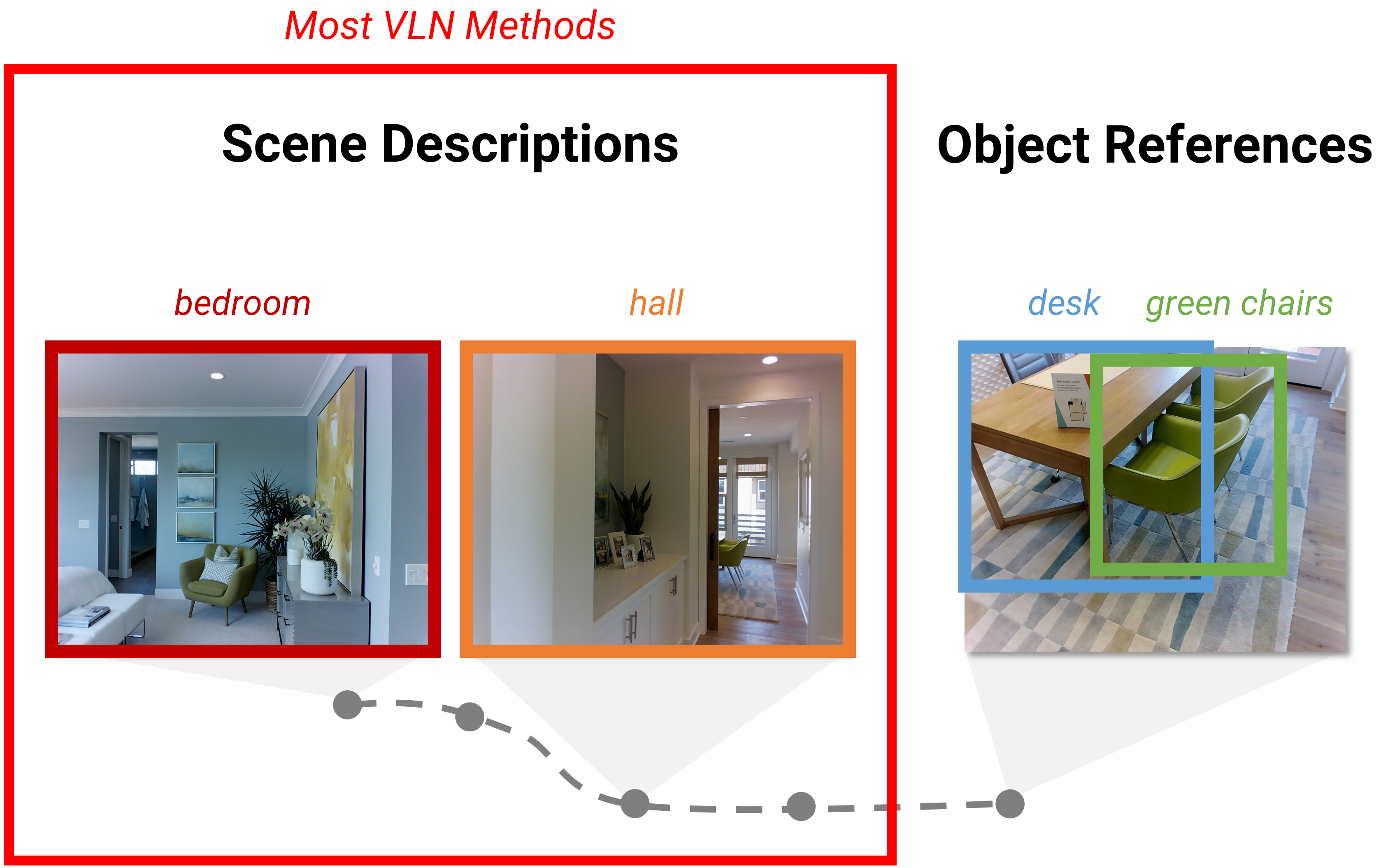
Object References



Motivation

Instruction

Exit the bedroom and turn left. Continue down the hall and into the room straight ahead and stop before the desk with two green chairs.



Most VLN Methods

Scene Descriptions

Object References

Motivation

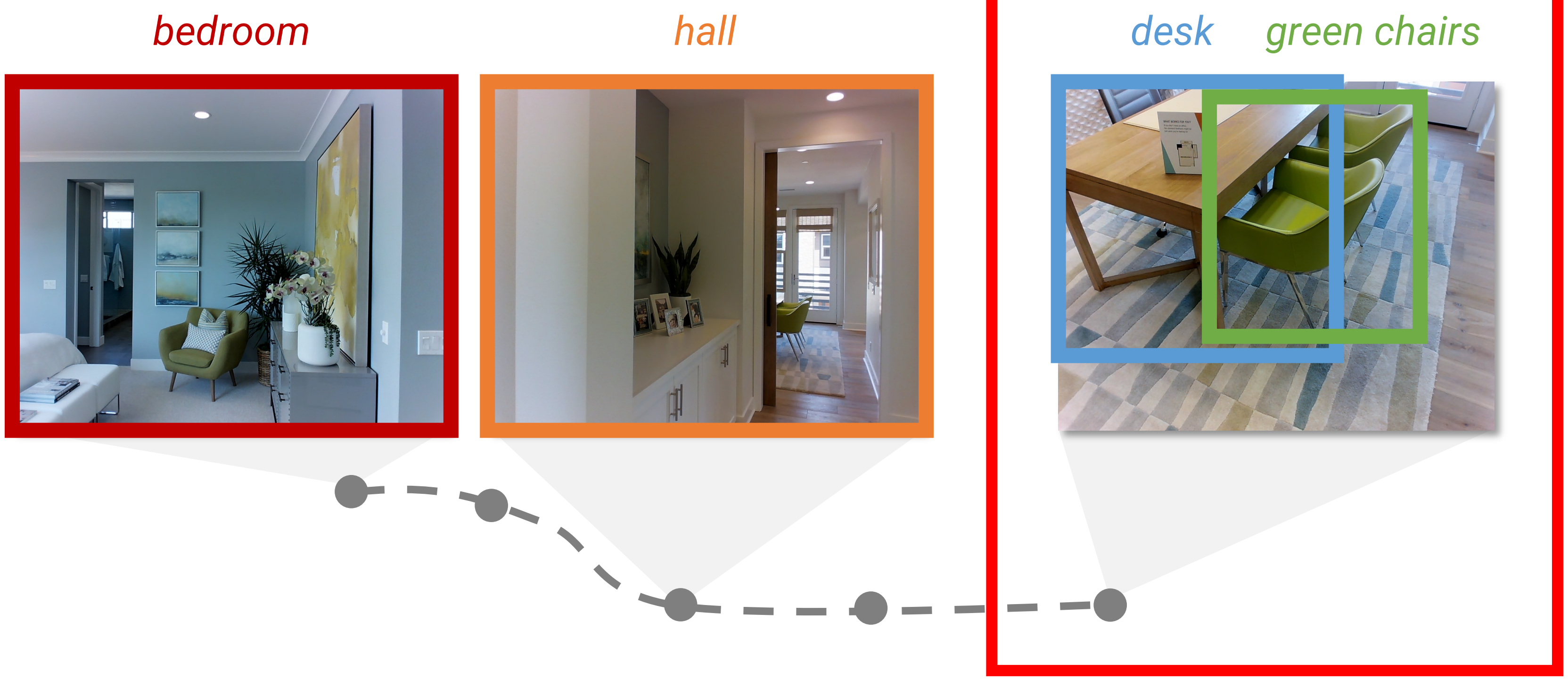
Instruction

Exit the bedroom and turn left. Continue down the hall and into the room straight ahead and stop before the desk with two green chairs.

Scene Descriptions

Most VLN Methods

Object References

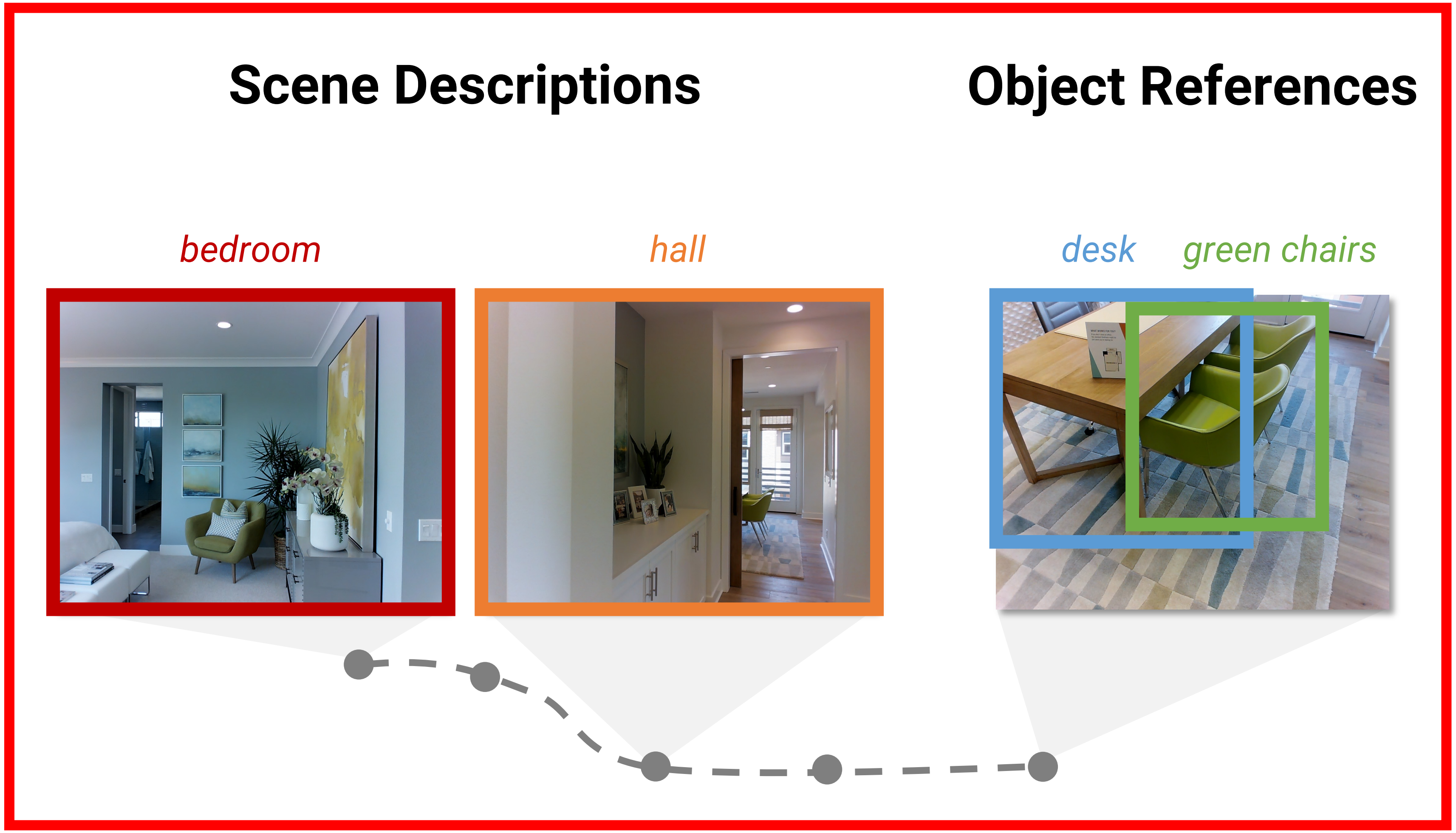


Our Approach

Instruction

Exit the bedroom and turn left. Continue down the hall and into the room straight ahead and stop before the desk with two green chairs.

SOAT: Scene- and Object-Aware Transformer



Scene Descriptions

Object References

bedroom

hall

desk

green chairs

Our Approach

Exit the bedroom and turn left. Continue down the hall and into the room straight ahead and stop before the desk with two green chairs.

Instruction

Our Approach

Exit the bedroom and turn left. Continue down the hall and into the room straight ahead and stop before the desk with two green chairs.

Instruction

Panoramic Observation

Our Approach

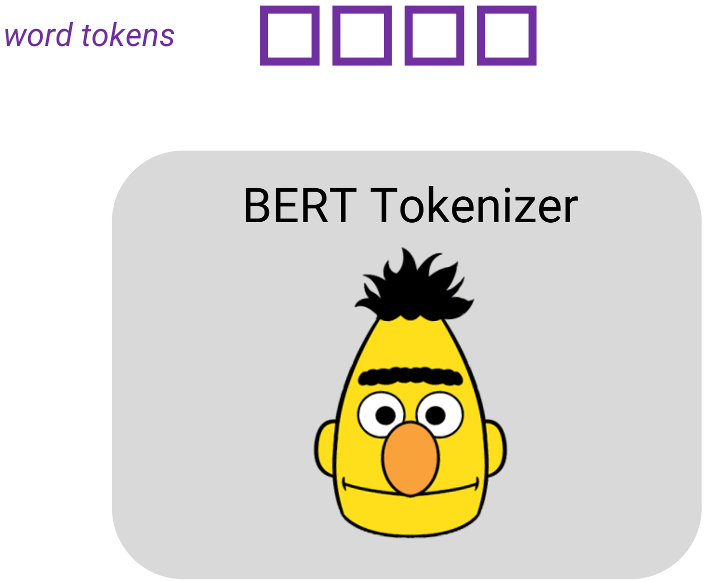
BERT Tokenizer

Exit the bedroom and turn left. Continue down the hall and into the room straight ahead and stop before the desk with two green chairs.

Instruction

Panoramic Observation

Our Approach



Exit the bedroom and turn left. Continue down the hall and into the room straight ahead and stop before the desk with two green chairs.

Instruction

Panoramic Observation

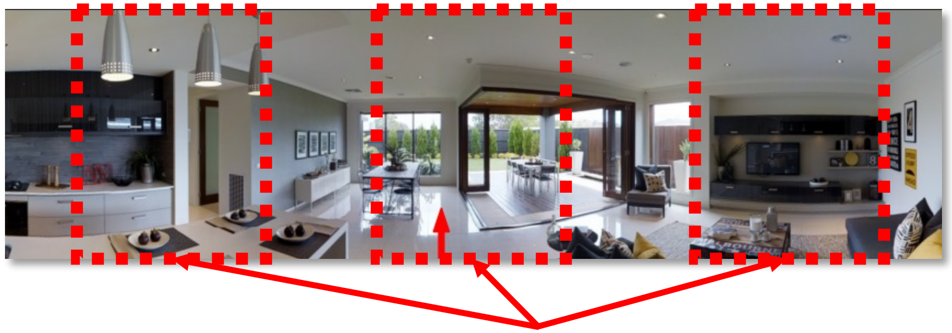
Our Approach

word tokens □□□□

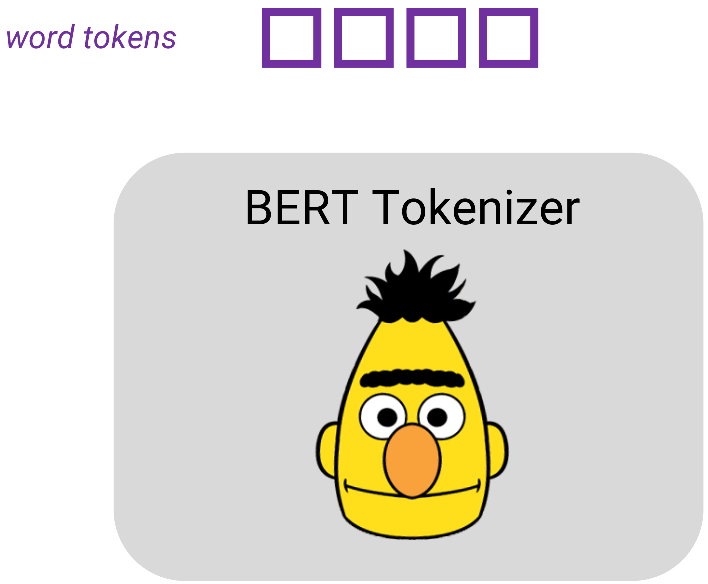
BERT Tokenizer

Exit the bedroom and turn left. Continue down the hall and into the room straight ahead and stop before the desk with two green chairs.

Instruction

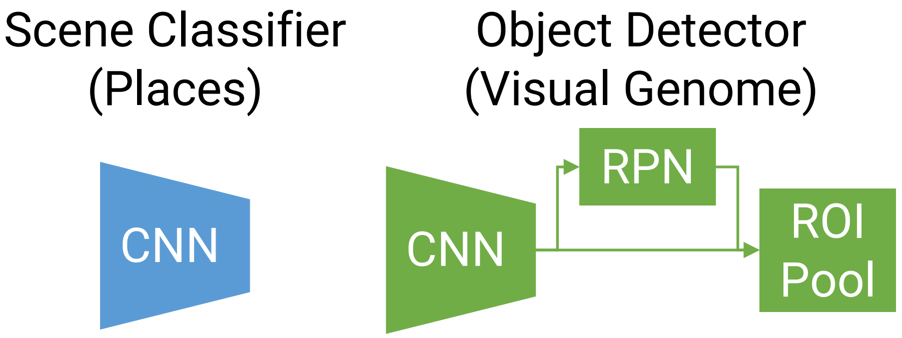


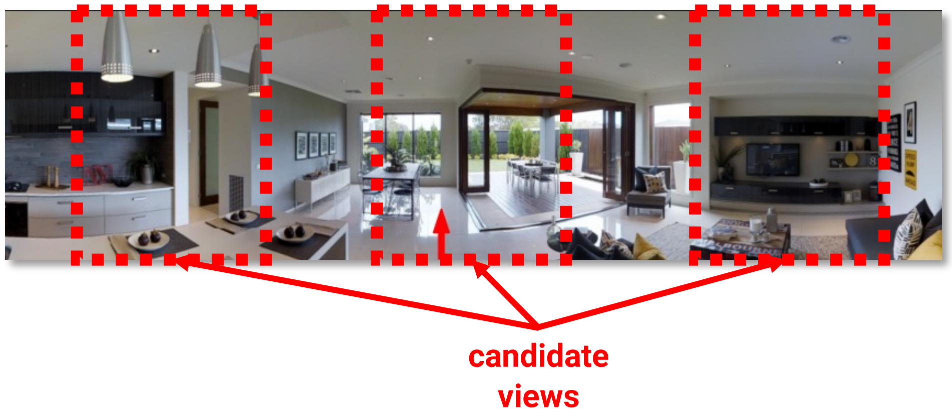
Our Approach



Exit the bedroom and turn left. Continue down the hall and into the room straight ahead and stop before the desk with two green chairs.

Instruction





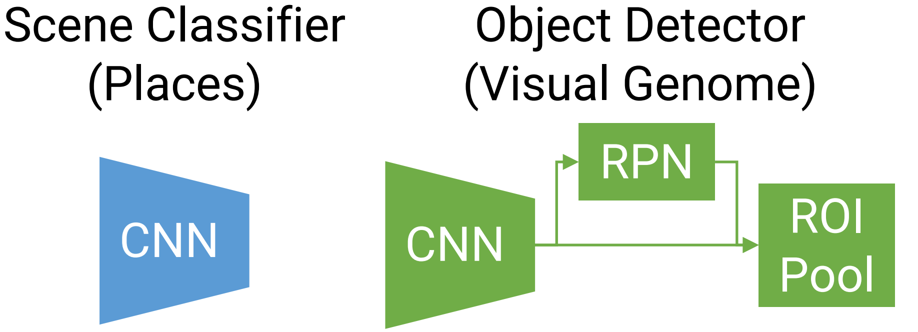
Our Approach

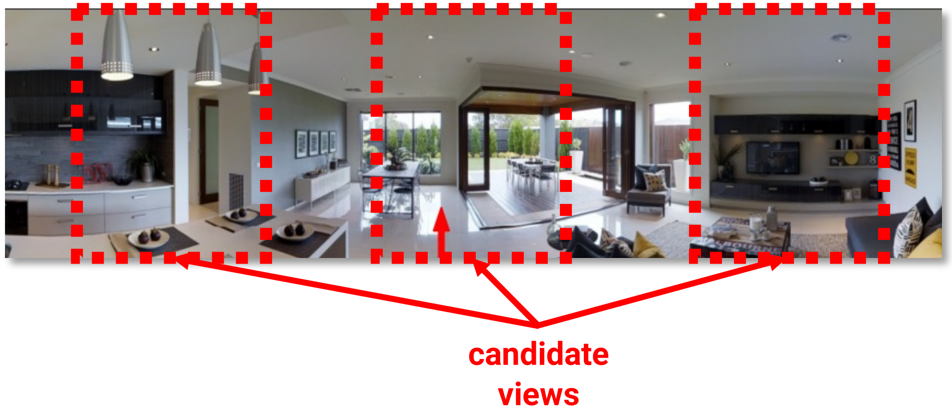
word tokens □□□□

Exit the bedroom and turn left. Continue down the hall and into the room straight ahead and stop before the desk with two green chairs.

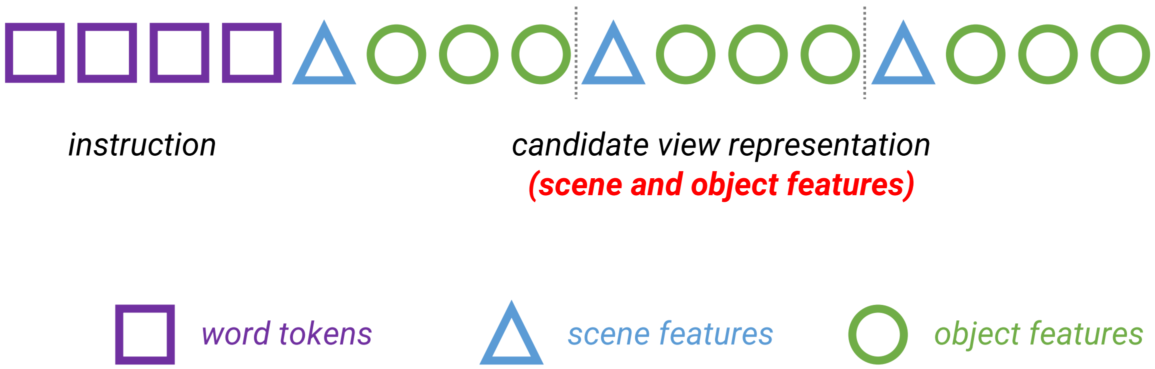
Instruction

*candidate view representation (**scene** and **object** features)* △○○○ △○○○ △○○○





SOAT: Scene- and Object-Aware Transformer



instruction

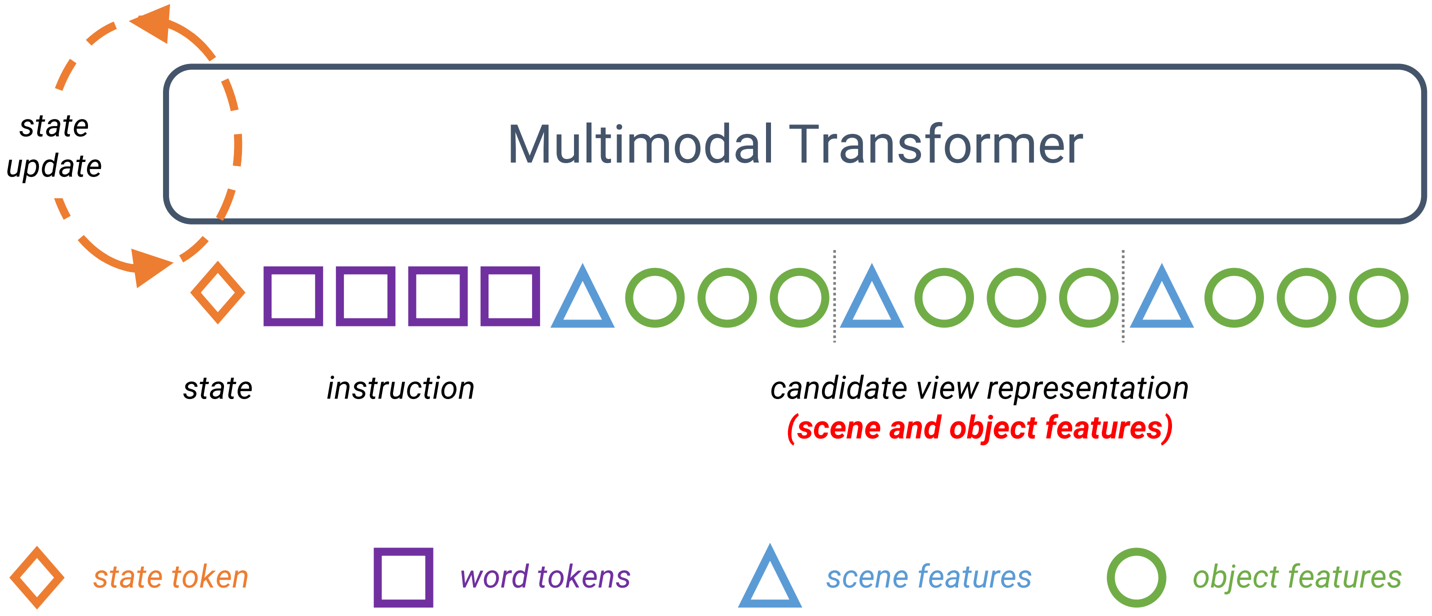
candidate view representation
(scene and object features)

□ *word tokens*

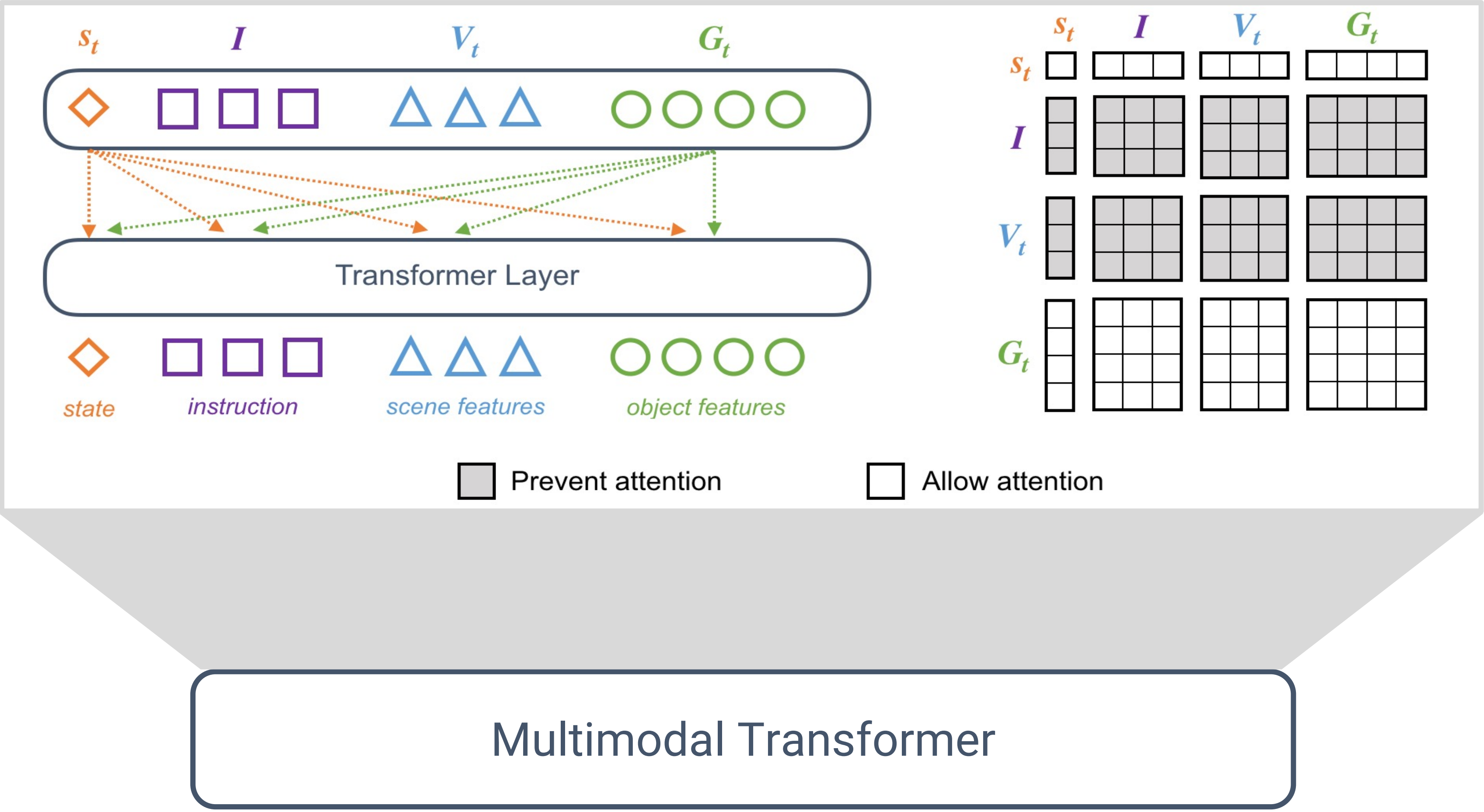
△ *scene features*

○ *object features*

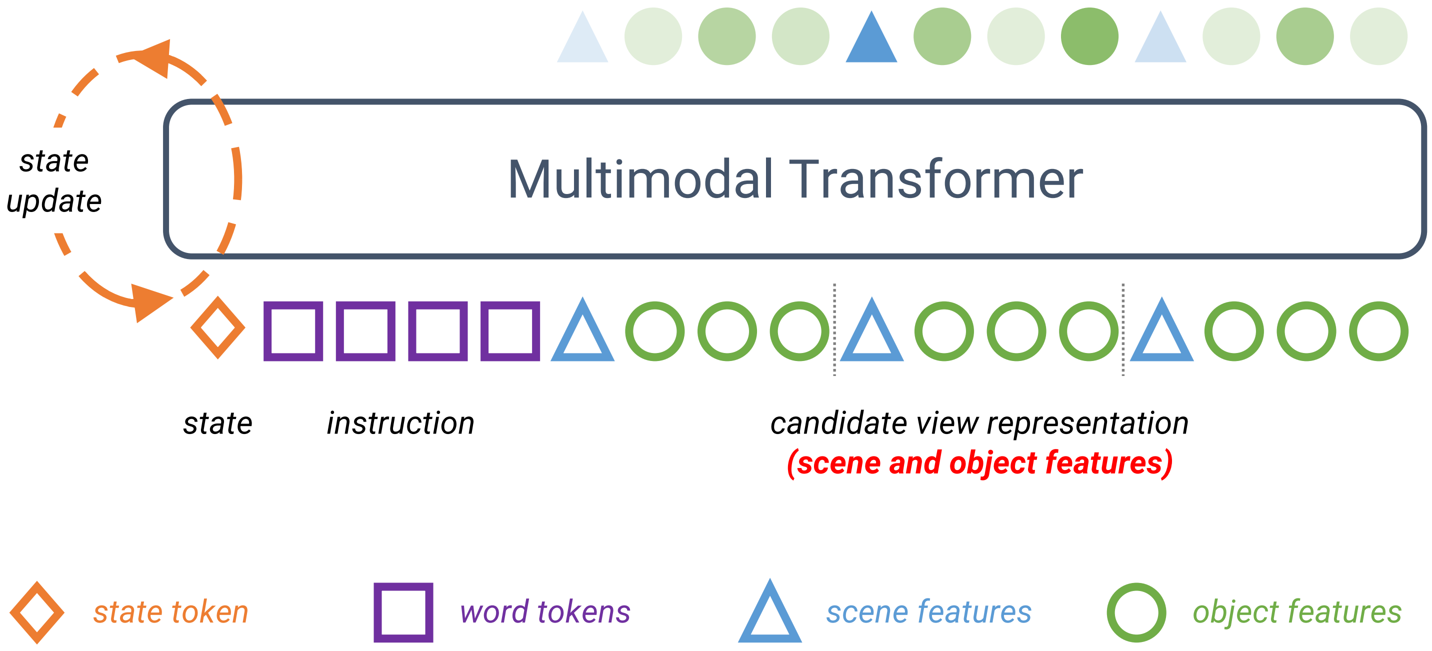
SOAT: Scene- and Object-Aware Transformer



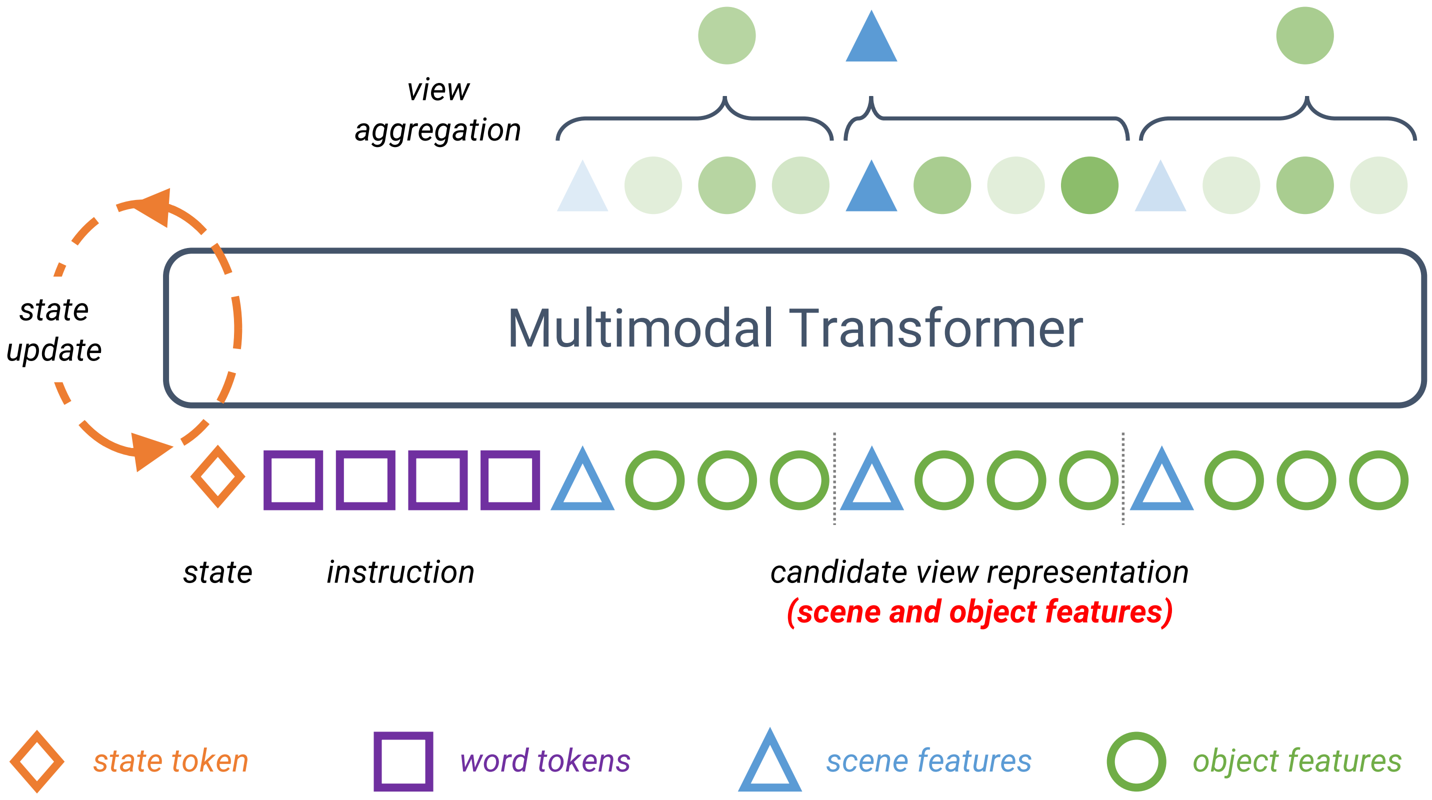
Selective Attention



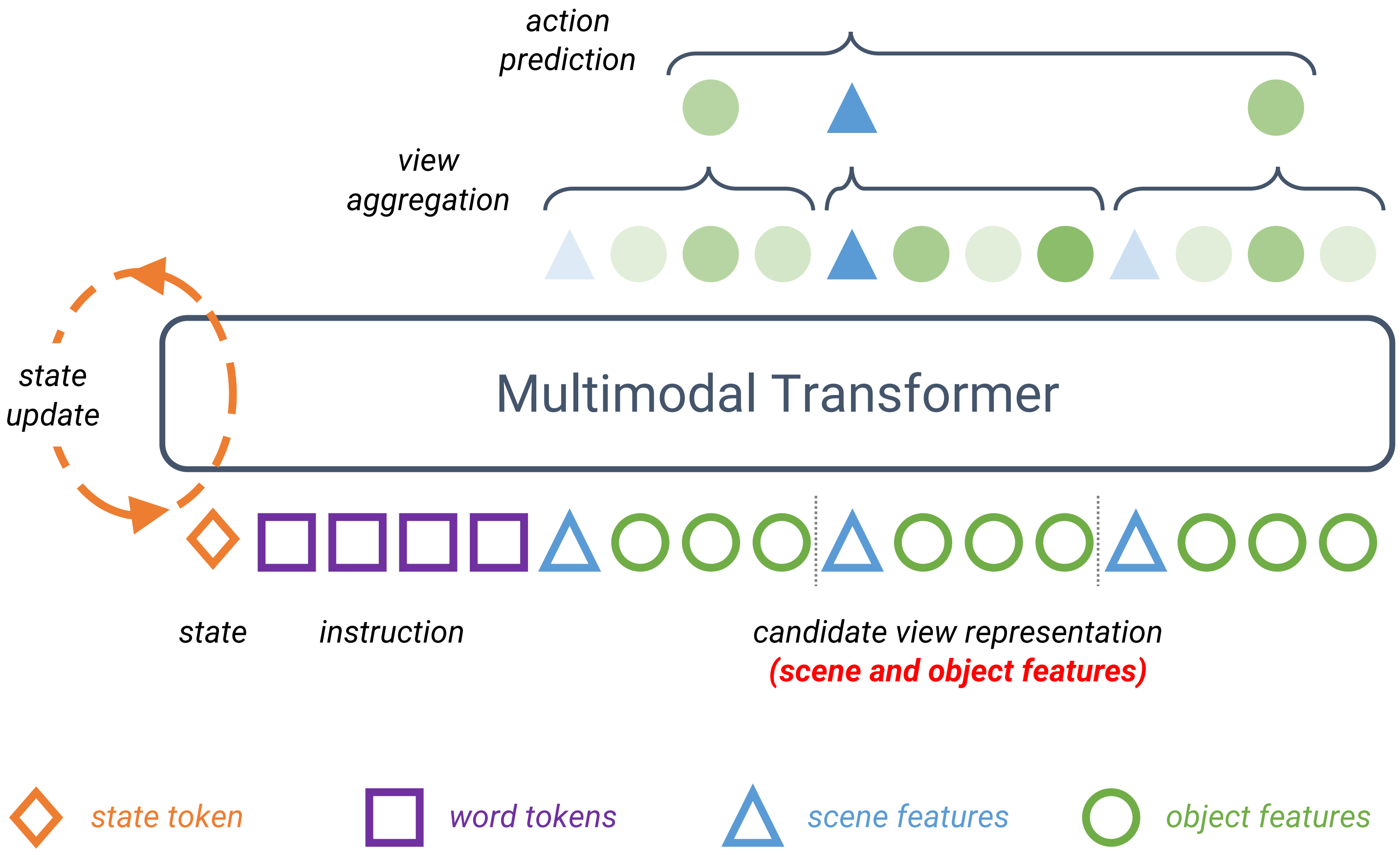
SOAT: Scene- and Object-Aware Transformer



SOAT: Scene- and Object-Aware Transformer



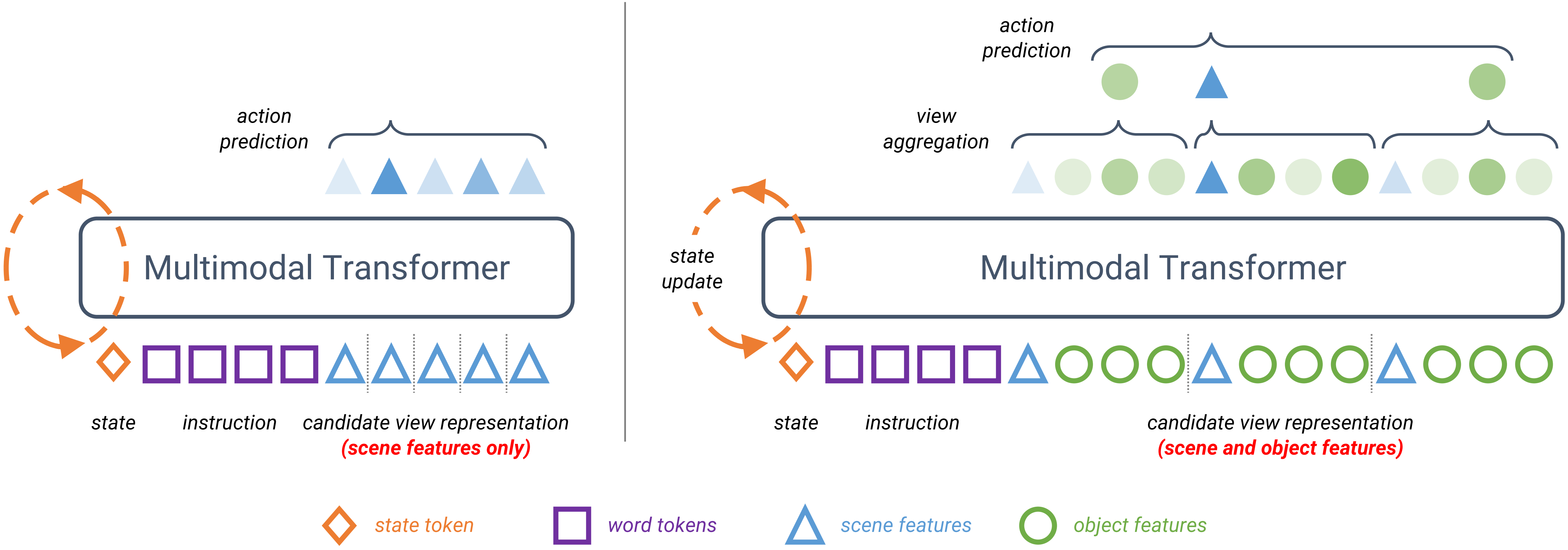
SOAT: Scene- and Object-Aware Transformer



Baseline: VLN↻BERT

VLN↻BERT

Proposed Approach



action prediction

Multimodal Transformer

state

instruction

candidate view representation

(scene features only)

action prediction

view aggregation

state update

Multimodal Transformer

state

instruction

candidate view representation

(scene and object features)

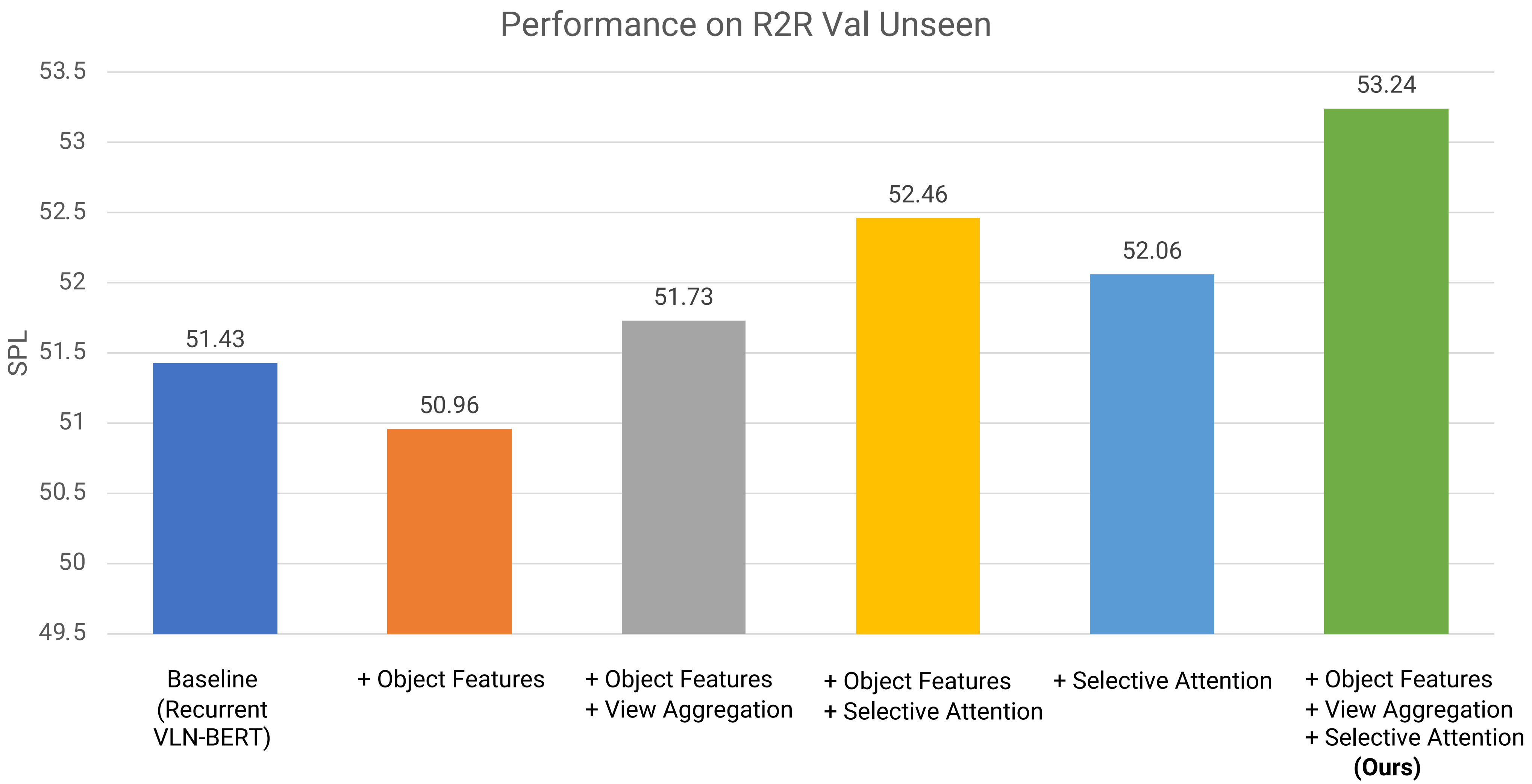
◇ *state token*

□ *word tokens*

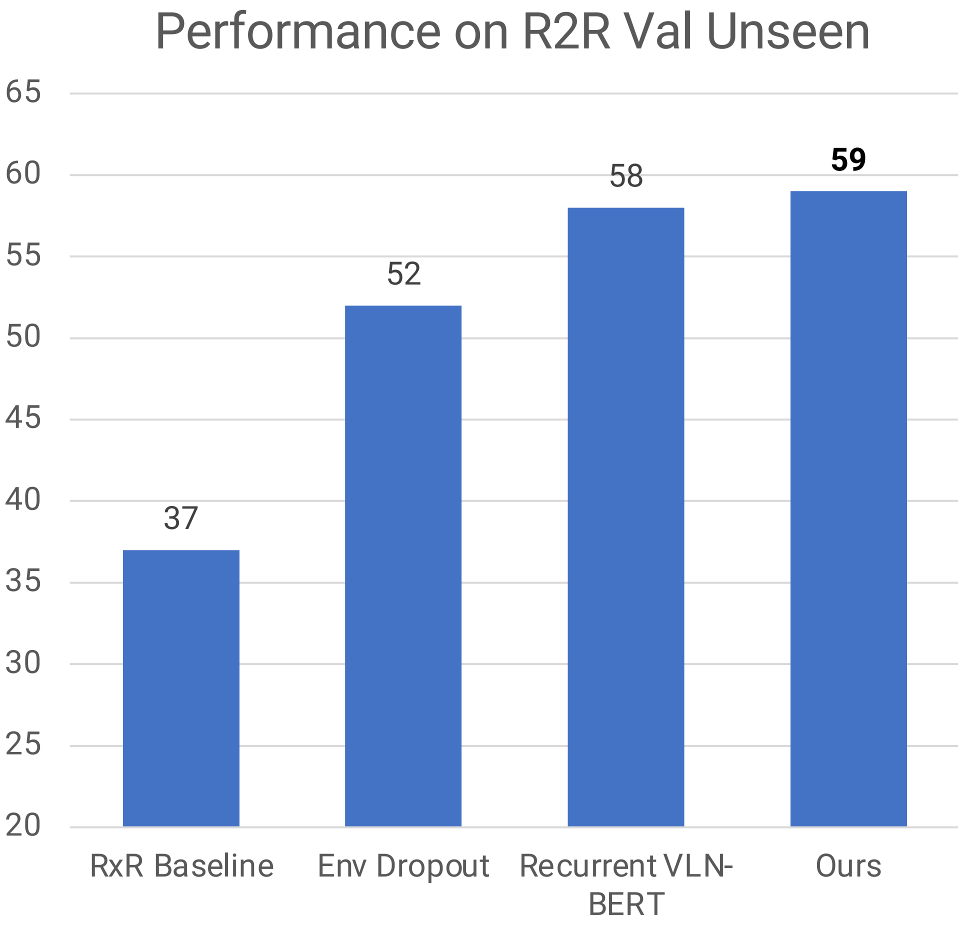
△ *scene features*

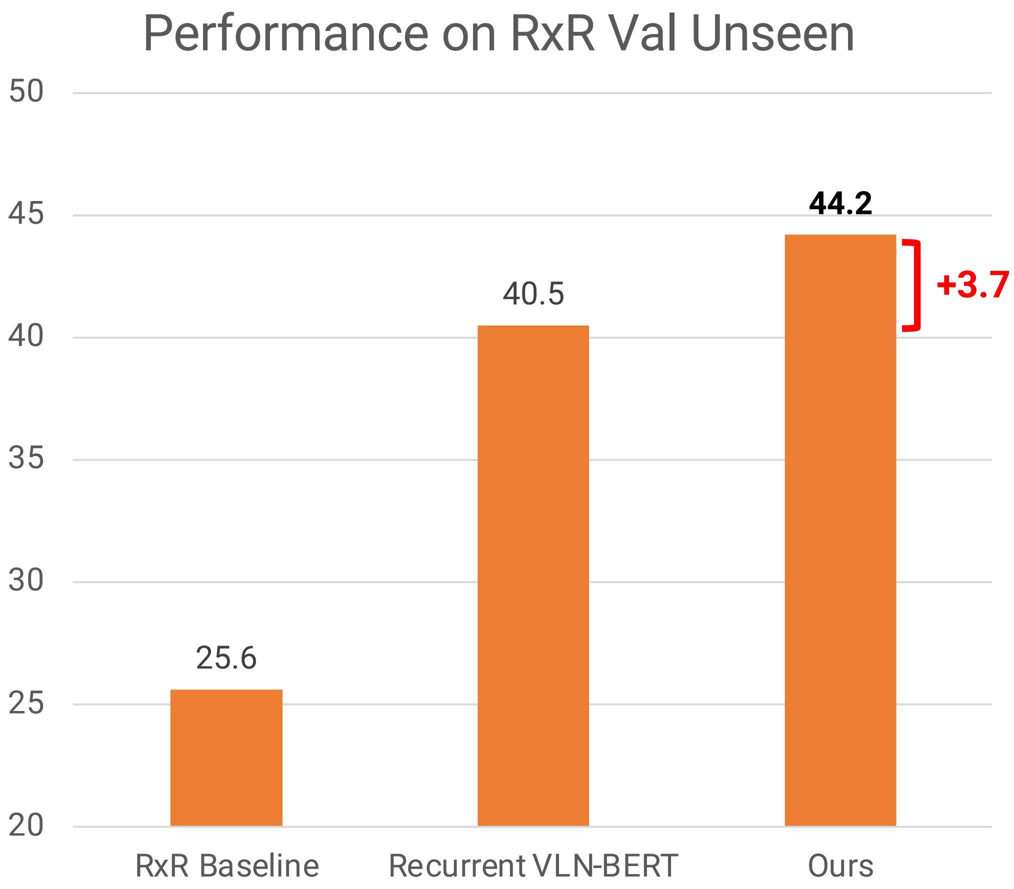
○ *object features*

Ablation Study



Comparison with State-of-the-Art





Results on RxR Object-Heavy Instructions Subset

Results on RxR Object-Heavy Instructions Subset

You are standing in front of a curio cabinet with lots of dolls in it. You are going turn to your right and enter that doorway. You will see dark wood floors. You are now in a bedroom. It will have a gray and black striped comforter on it. You are going to walk into the bedroom and walk in between the foot of the bed and on your left will be a dresser with a large outdoor painting on it. You are going to stop right there in between those two.

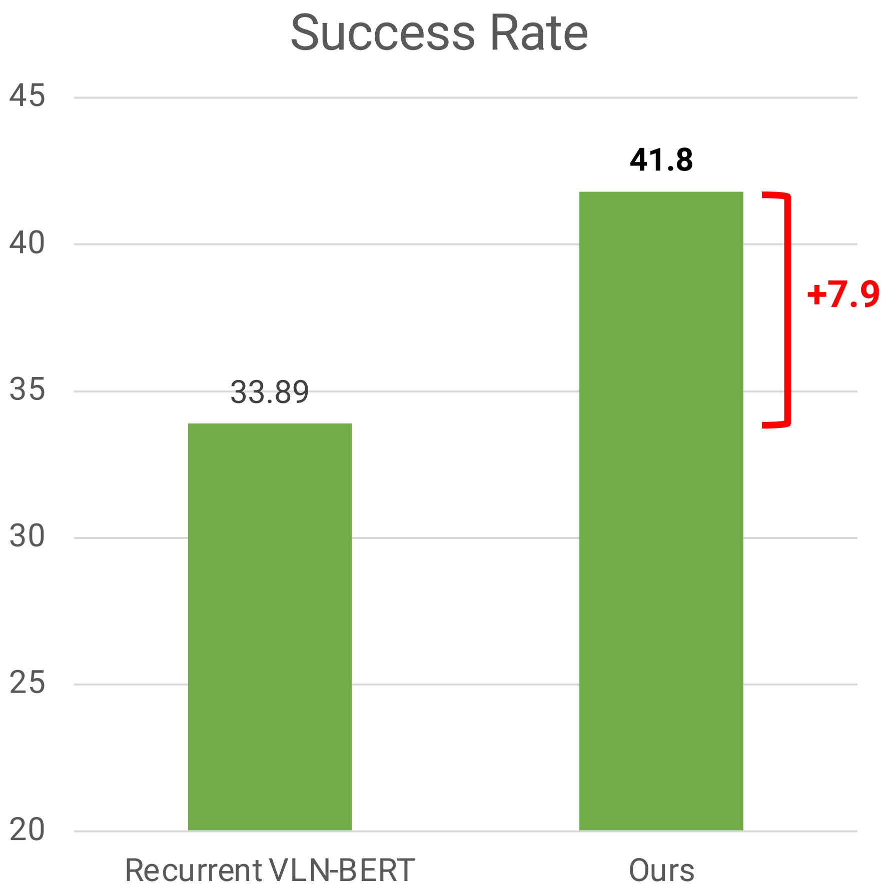
Object-Heavy Instruction Example from RxR Dataset

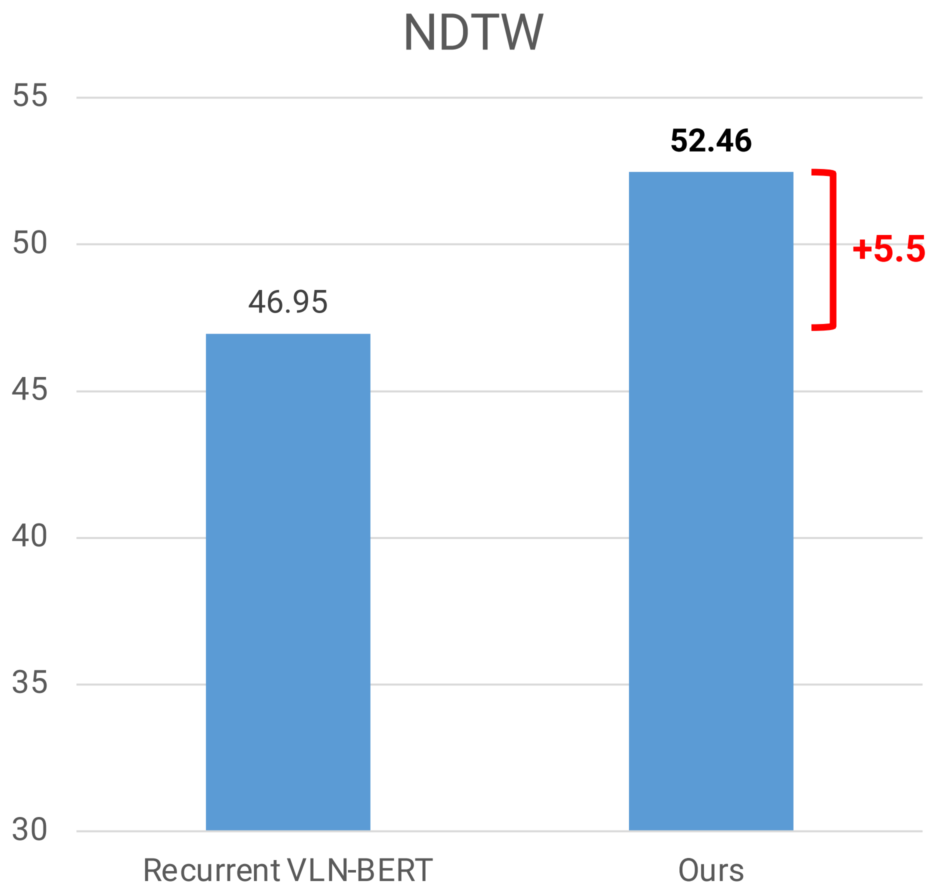
Results on RxR Object-Heavy Instructions Subset

You are standing in front of a curio cabinet[1] with lots of dolls[2] in it. You are going turn to your right and enter that doorway. You will see dark wood floors. You are now in a bedroom. It will have a gray and black striped comforter[3] on it. You are going to walk into the bedroom and walk in between the foot[4] of the bed[5] and on your left will be a dresser[6] with a large outdoor painting[7] on it. You are going to stop right there in between those two.

Object-Heavy Instruction Example from RxR Dataset

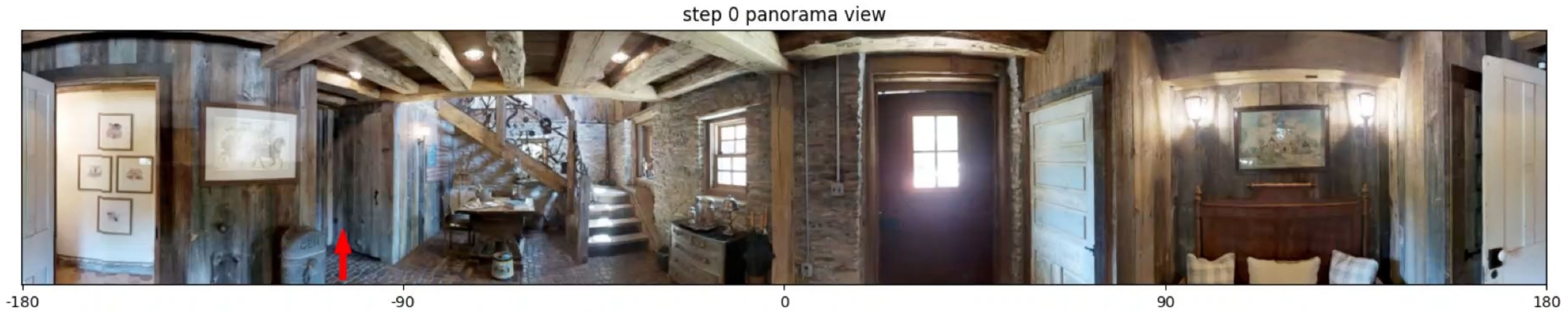
Results on RxR Object-Heavy Instructions Subset

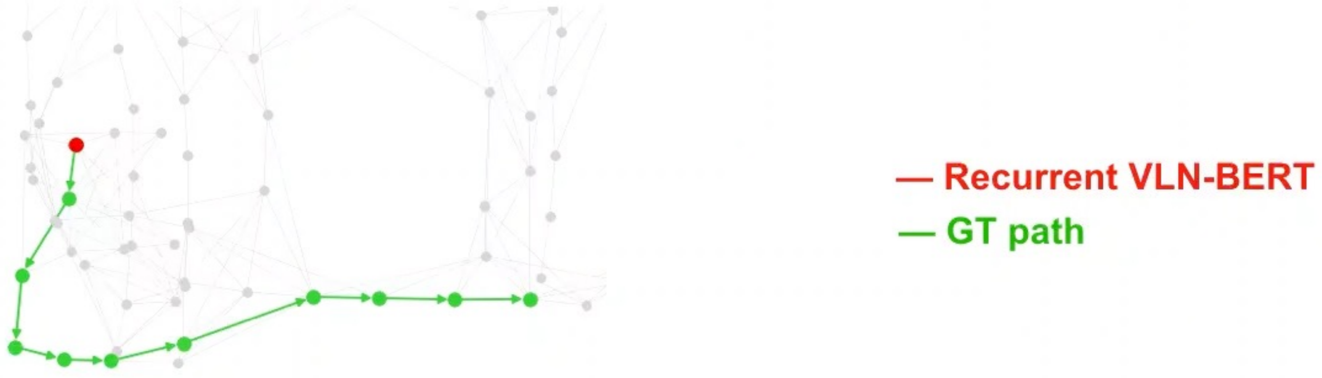




Qualitative Example: Baseline VLN↻BERT

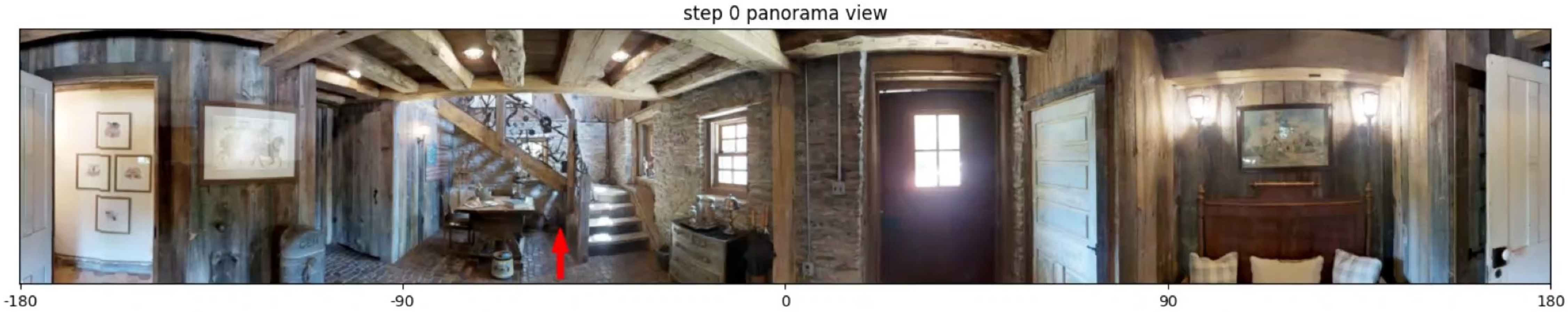
RxR Instruction: “You are going to start facing a front door. Turn to the left and go up the stairs. Once at the top, you are going to keep walking straight, pass the china cabinet on you right and take two more steps, and stop once you are next to the desk and the small dresser with two red glasses on top. Once you are there, you are done.”

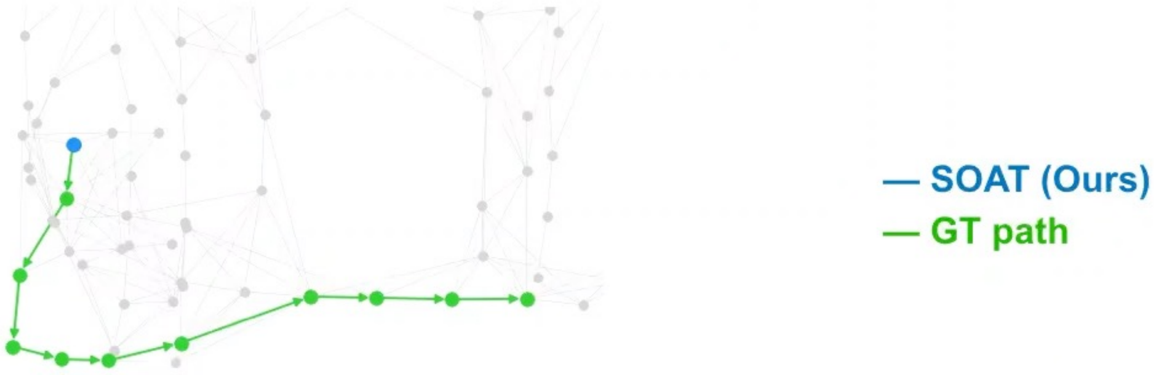




Qualitative Example: SOAT (Ours)

RxR Instruction: “You are going to start facing a front door. Turn to the left and go up the stairs. Once at the top, you are going to keep walking straight, pass the china cabinet on you right and take two more steps, and stop once you are next to the desk and the small dresser with two red glasses on top. Once you are there, you are done.”





Thank you!