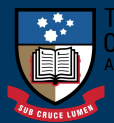
Diagnose Vision-and-Language Navigation: What Really Matters?

Wanrong Zhu[1], Yuankai Qi[2], Pradyumna Narayana[3], Kazoo Sone[3], Sugato Basu[3], Xin Eric Wang[4], Qi Wu[2], Miguel Eckstein[1], William Yang Wang[1]

[1]UC Santa Barbara, [2]University of Adelaide,

[3]Google, [4]UC Santa Cruz



THE UNIVERSITY OF ADELAIDE AUSTRALIA

Google

UC SANTA CRUZ

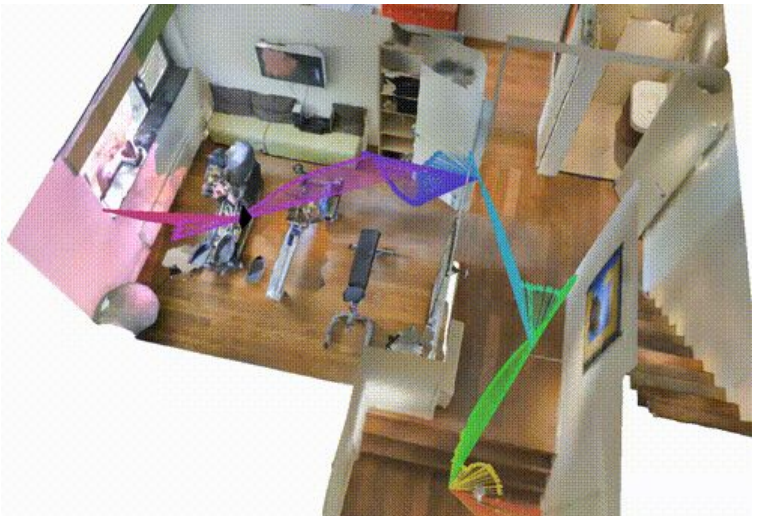
UC SANTA BARBARA[1]

Vision-and-Language Navigation(VLN) Benchmarks

Room-to-room (R2R)

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

Room-across-room (RxR)



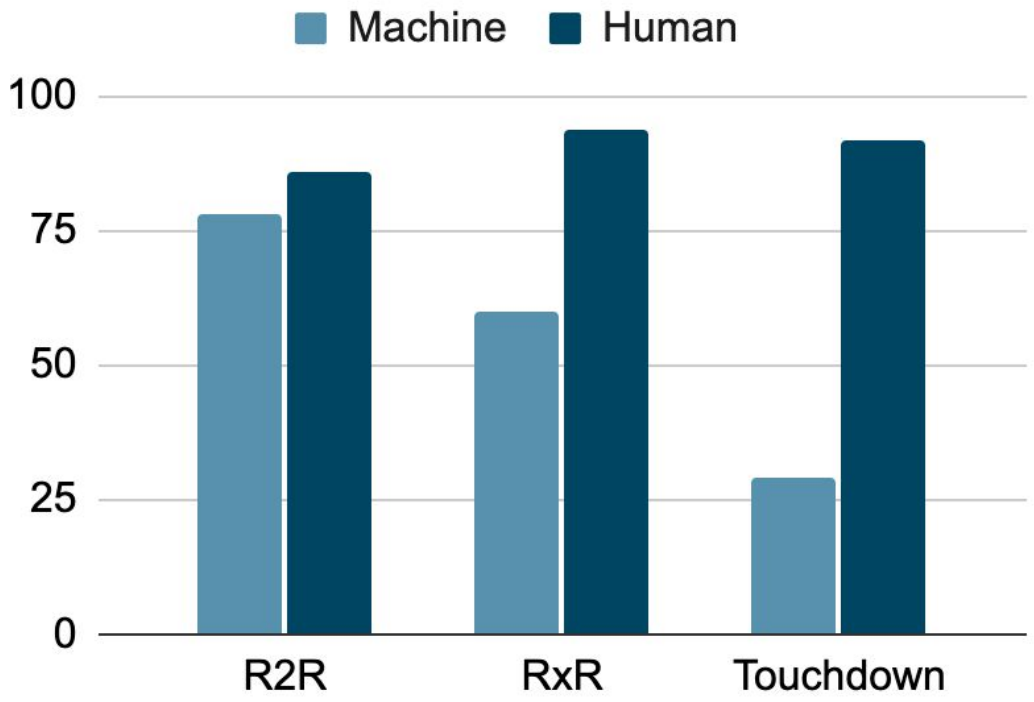
Now you are standing in-front of a closed door, turn to your left, you can see two wooden steps, climb the steps and walk forward by crossing a wall painting which is to your right side, you can see open door enter into it. This is a gym room, move forward, walk till the end of the room, you can see a grey colored ball to the corner of the room, stand there, that's your end point.

Touchdown

Orient yourself so that the umbrellas are to the right. Go straight and take a right at the first intersection. At the next intersection there should be an old-fashioned store to the left. There is also a dinosaur mural to the right. Touchdown is on the back of the dinosaur.

Machine vs. Human on VLN Benchmarks

- State-of-the-art performances as of June 10th, 2022.



Covered Models

Benchmark	Model	Transformer-based?	Visual Feature
R2R	EnvDrop	×	ResNet-152
	FAST	×	
	VLN-Recurrent-BERT	√	
	PREVALENT	√	
RxR-en	CLIP-ViL	×	CLIP-ViT
	VLN-HAMT	√	
Touchdown	RCONCONT	×	ResNet-18
	ARC	×	
	VLN-Transformer	√	

Analysis on Instruction Understanding

- *What can the agents learn from the instructions?*
- *Do agents pay more attention to object tokens or direction tokens?*

VLN Instruction Exemplar

- “Enter the hallway that’s in front of you, turn to the left, take five steps further. ”

Object-related Tokens

VLN Instruction Exemplar

- "Enter the **hallway** that's in front of you, turn to the left, take five **steps** further."

Object-related Tokens: Interventions

Original Instruction

- "Enter the **hallway** that's in front of you, turn to the left, take five **steps** further."

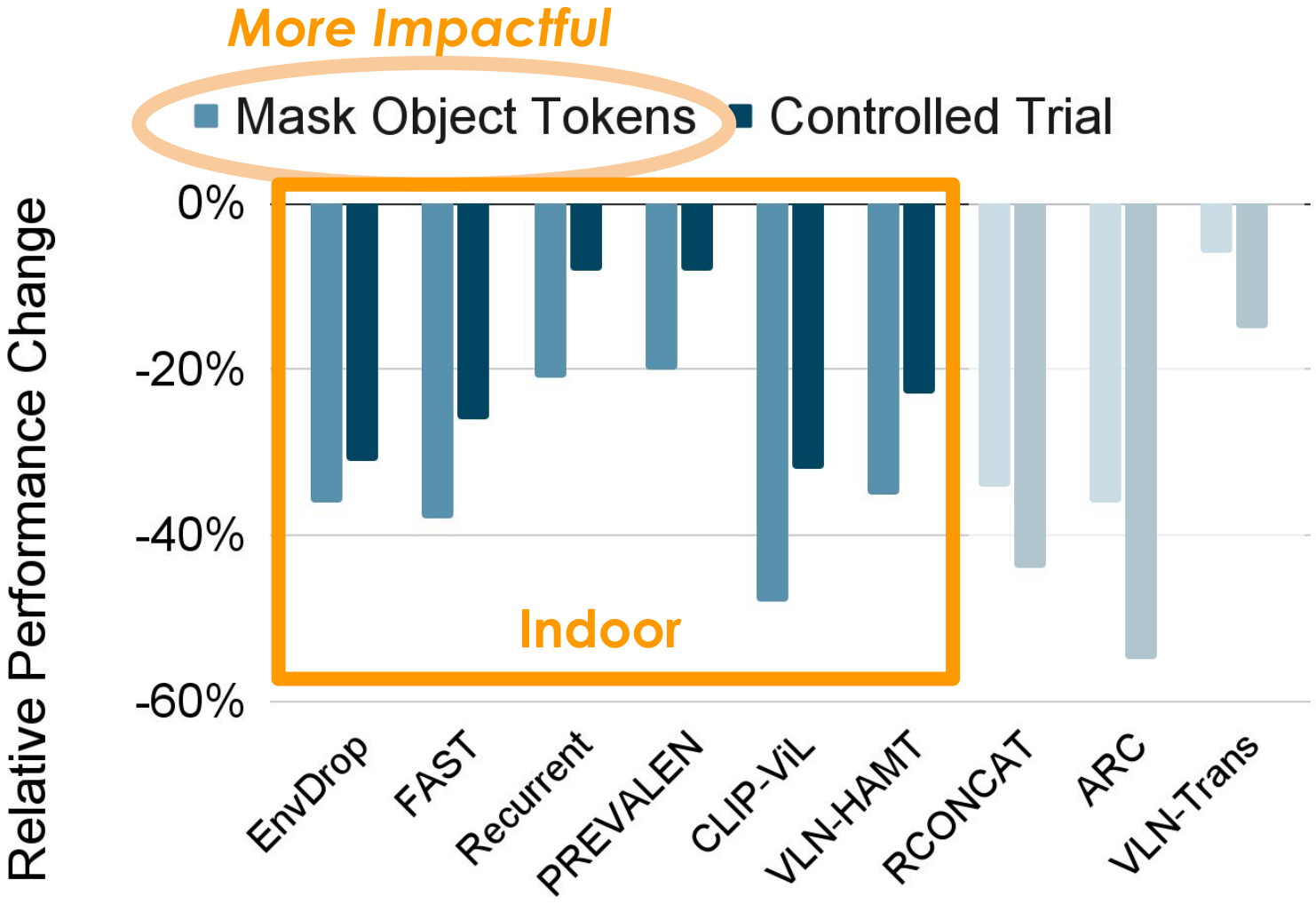
Ablated Instruction

- "Enter the **[MASK]** that's in front of you, turn to the left, take five **[MASK]** further."

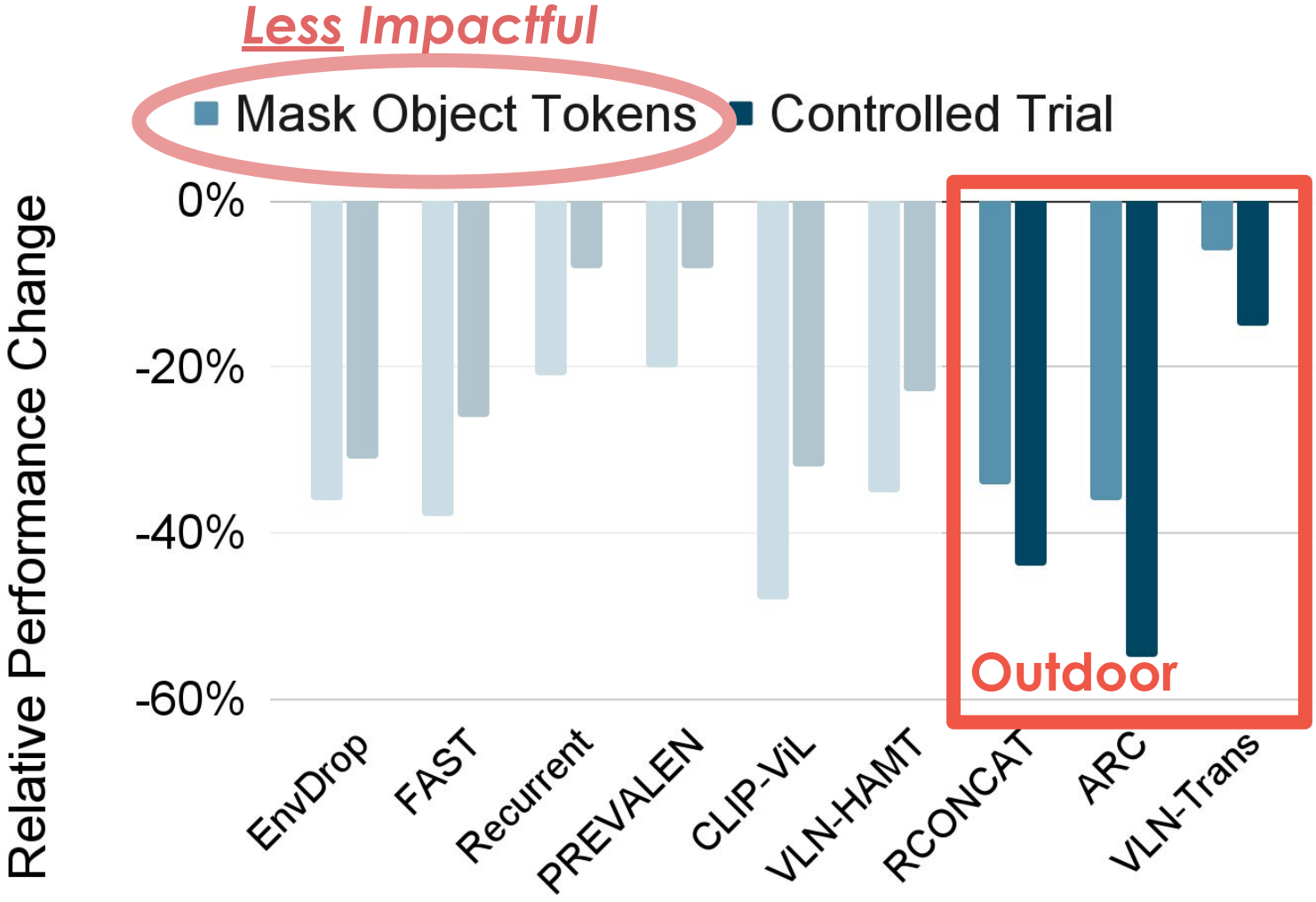
Controlled Trial

- "Enter **[MASK]** hallway that's in front of **[MASK]**, turn to the left, take five steps further."

The Effect of Object-related Tokens



The Effect of Object-related Tokens



Direction-related Tokens

VLN Instruction Exemplar

- "Enter the hallway that's in ***front*** of you, turn to the ***left***, take five steps further."

Direction-related Tokens: Interventions

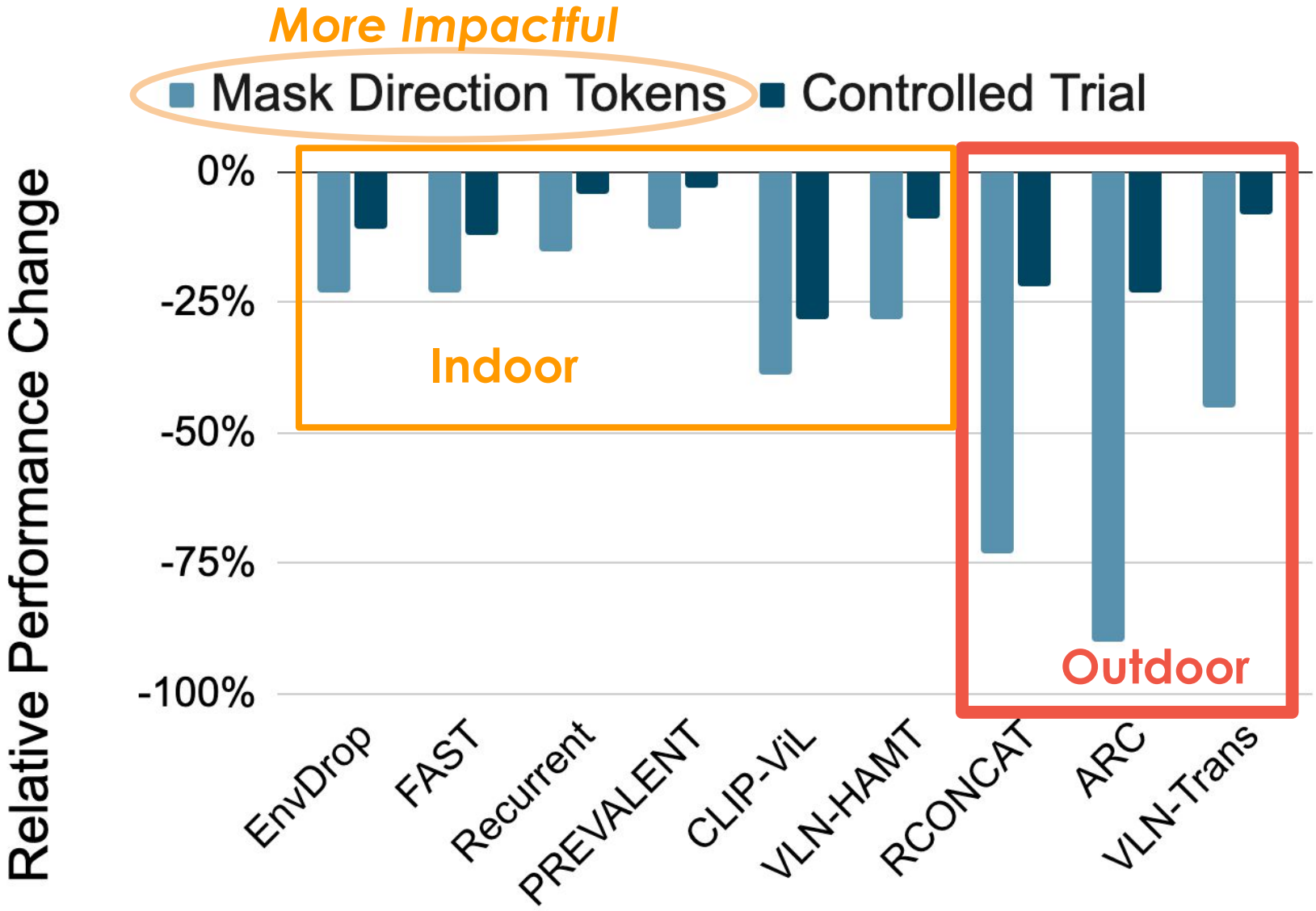
Original Instruction

- "Enter the hallway that's in ***front*** of you, turn to the ***left***, take five steps further."

Ablated Instruction

- "Enter the hallway that's in **[MASK]** of you, turn to the **[MASK]**, take five steps further."

The Effect of Direction-related Tokens



Quick Takeaways

VLN Instruction Understanding

- Indoor agents refer to both objects and directions in the instruction
- Outdoor agents heavily rely on direction tokens, and poorly understand visual objects

Analysis on Vision-Language Alignment

- *Can agents match tokens to visual entities?*
- *How reliable are such connections?*

Perturbation on the Instructions

Original Instruction

- "Enter the **hallway** that's in front of you, turn to the left, take five **steps** further…"

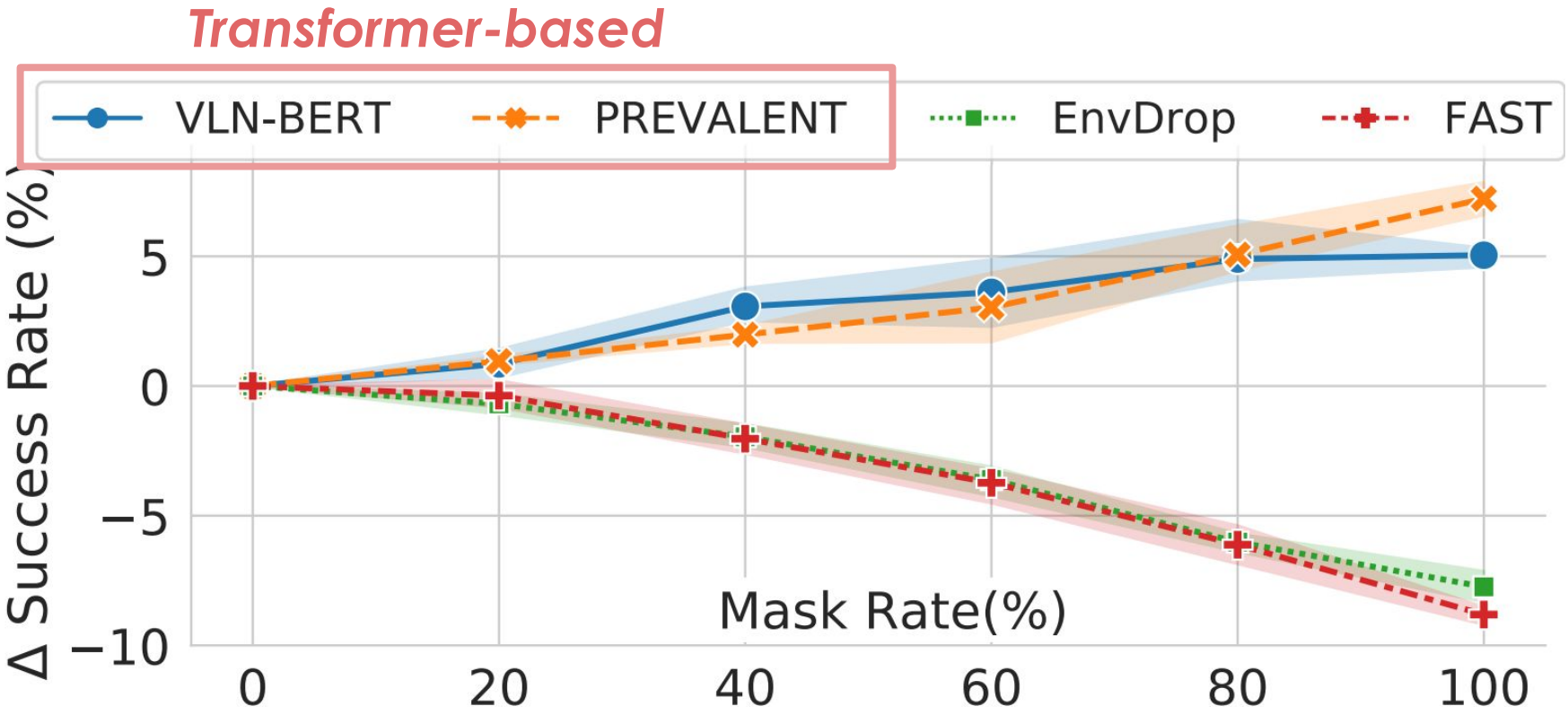
Mask Object-related Tokens

- "Enter the **[MASK]** that's in front of you, turn to the left, take five **[MASK]** further…"

Replace Object-related Tokens

- "Enter the ***vase*** that's in front of you, turn to the left, take five ***sink*** further…"

Perturbation on the Instructions (R2R)



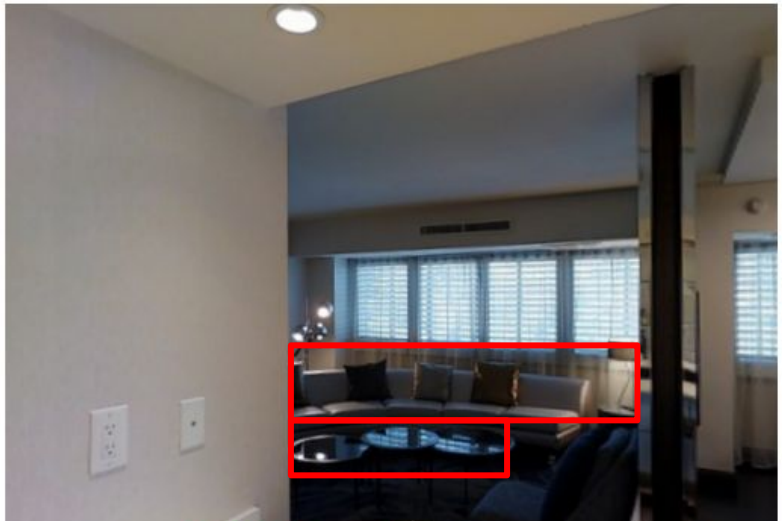
Y-Axis: Performance gap Δ between masking & replacing object tokens.

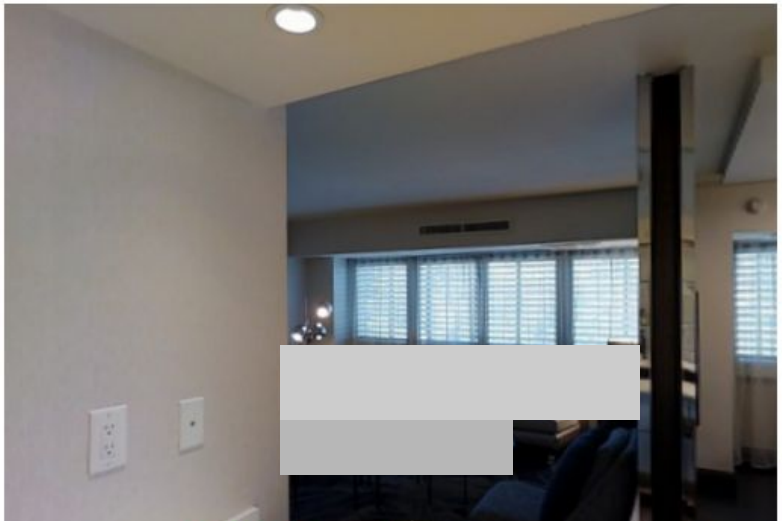
Δ > 0: agent have better understanding on object tokens

Perturbation on the Visual Environment

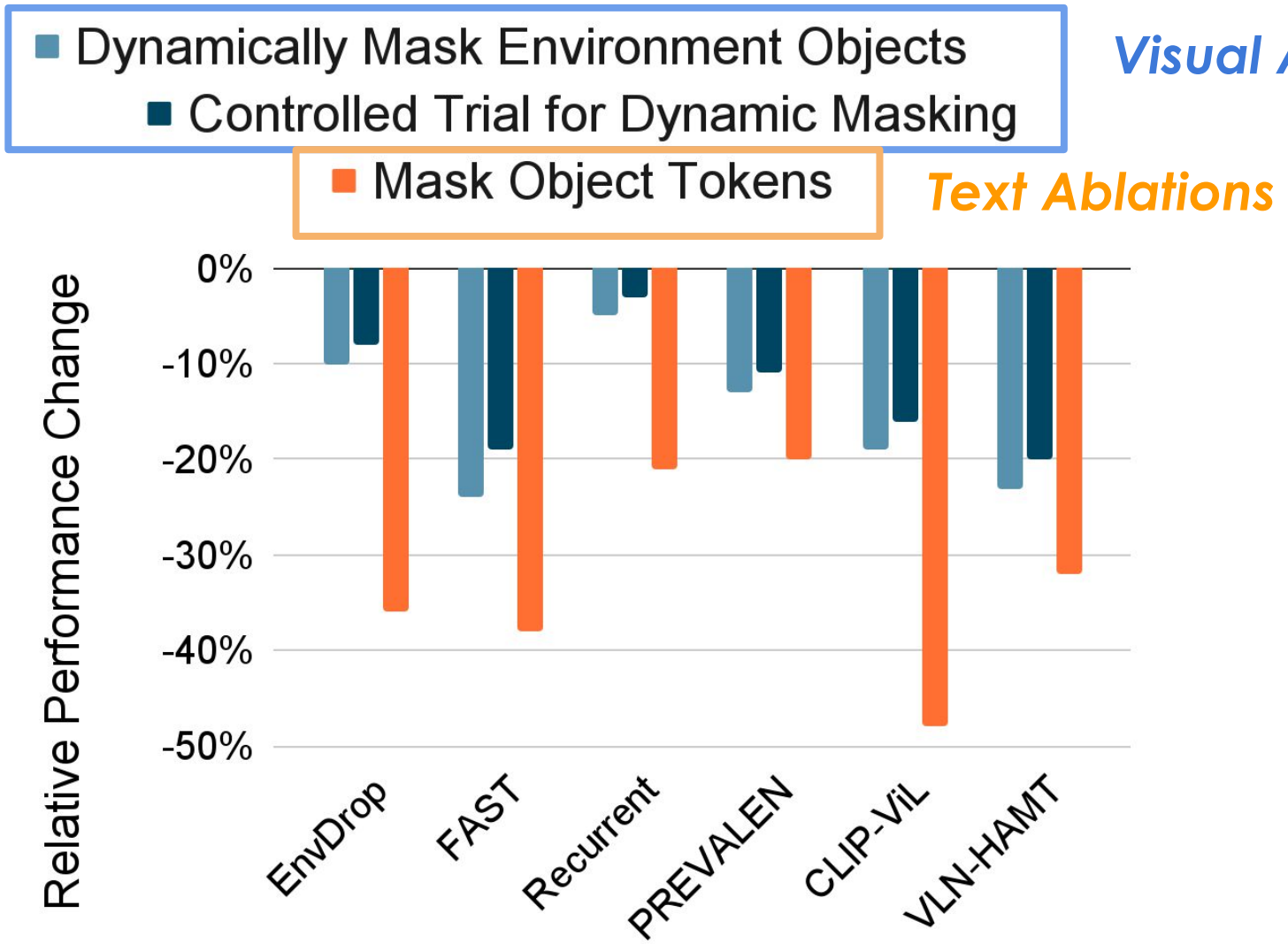
Dynamically mask the objects mentioned in the instructions

- "…go towards the **white couch** and stop in front of the **coffee table**…"





Perturbation on the Visual Environment vs. on the Instruction



Quick Takeaways

VLN Vision-Language Alignment

- Transformer-based VLN agents have better cross-modal understanding of objects
- Indoor VLN agents have unbalanced attention on text and visual input

Thank you!

Q & A