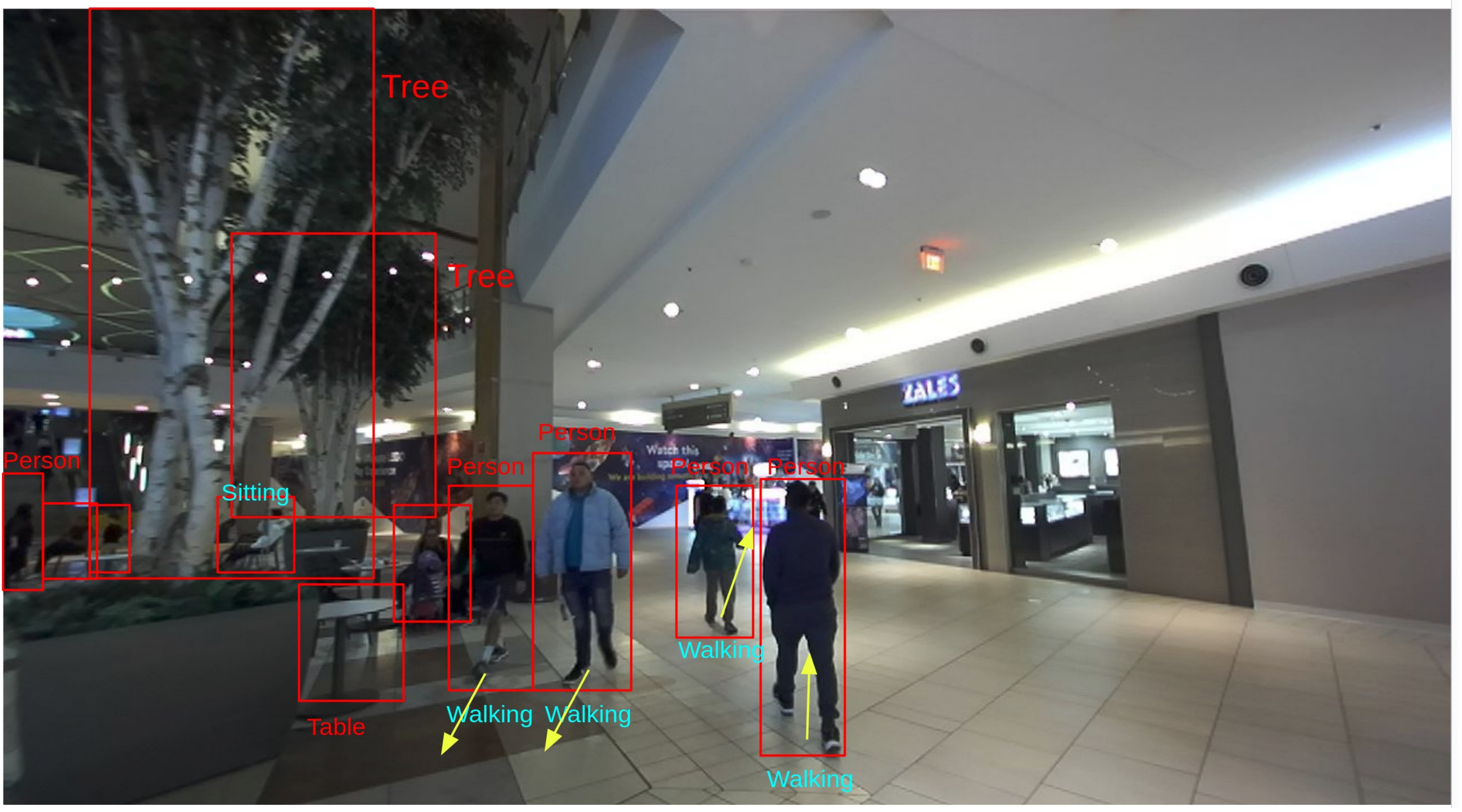
VANP: Learning Where to See for Navigation with Self-Supervised Vision-Action Pre-Training

Mohammad Nazeri, Junzhe Wang, Amirreza Payandeh, and Xuesu Xiao
George Mason University

INTRODUCTION

Egocentric visual navigation in public spaces can be challenging due to the unpredictable nature of humans in the environment. To address this ambiguity, traditional approaches divide the task into multiple sub-tasks:

- Object detection and Classification
- Intention prediction
- Trajectory forecasting
- ...



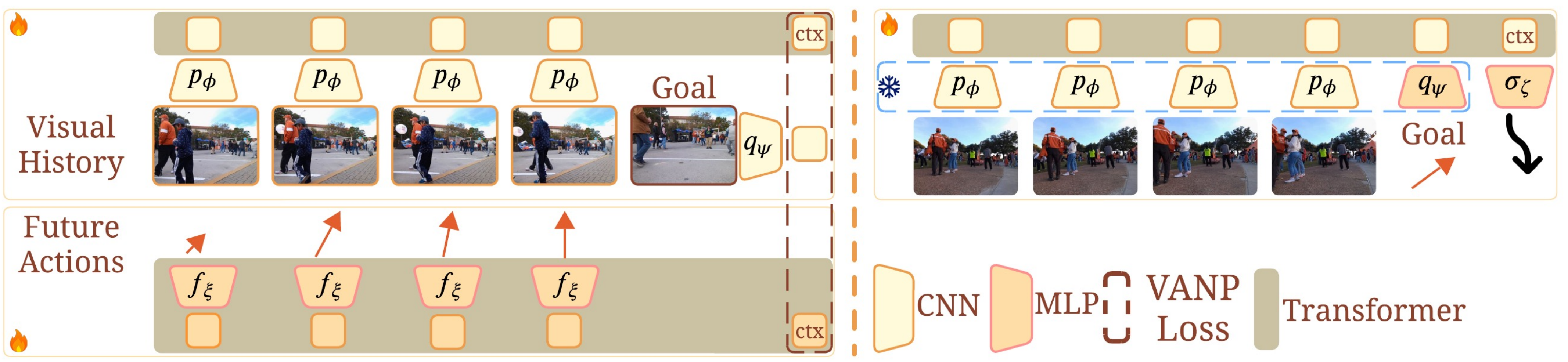
Q: Can we infer everything that matters for navigation from information provided by those manually designed sub-tasks?

A: No! It's difficult to engineer everything in a finite list of sub-tasks!

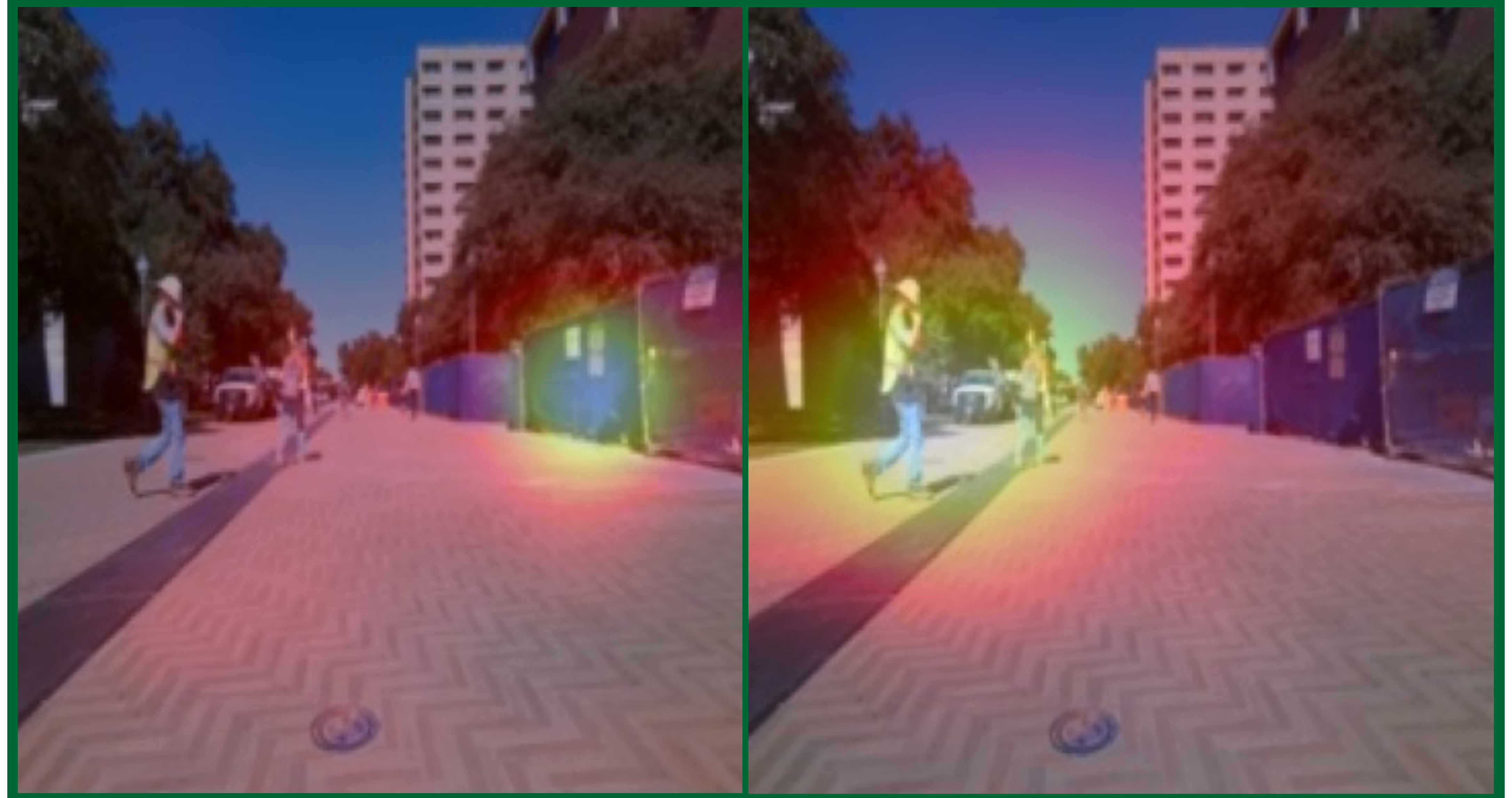
To address this, the research community uses end-to-end models and lets the data decide for us what is important for navigation:

- Train from scratch.
  - Time consuming.
  - Prone to overfitting or learning mode(s) of data.
- Use pretrained models.
  - Working well for normal images.
  - Inaccurate for egocentric visual navigation.

We introduce VANP, a non-contrastive self-supervised learning approach that uses future actions and the goal image as the self-supervision signal to correct learned visual features.



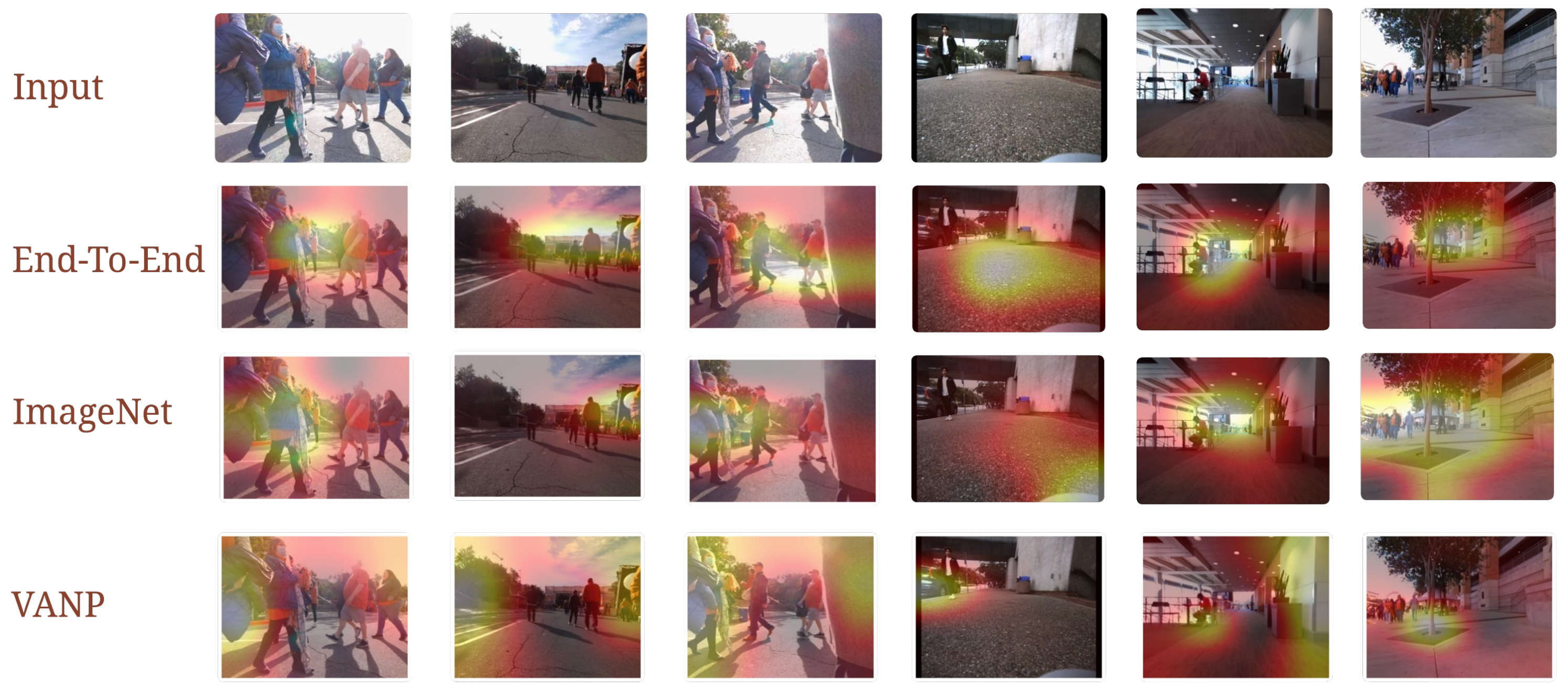
Comparison with a model pretrained on ImageNet:



Paper:

Code:

Qualitative Evaluation:



Quantitative Evaluation:

Type	Method	Weight	Single-frame	Multiple-frame	Frozen ❄ 3s	Frozen ❄ 5s	Fine-tuned 🔥 3s	Fine-tuned 🔥 5s
End-to-End ⌛	Resnet-50	Random	✓	✗	-	-	0.116	0.307
	ResnetTransformer	Random	✗	✓	-	-	0.113	0.320
Backbone Supervised ⌛	Resnet-50	ImageNet	✓	✗	**0.129**	0.356	0.129	0.342
	ResnetTransformer	ImageNet	✗	✓	0.169	0.435	0.107	0.292
Backbone Self-Supervised ⌛	Resnet-50	VANP	✓	✗	0.144	0.374	**0.103**	**0.272**
	ResnetTransformer	VANP	✗	✓	0.133	**0.342**	0.114	0.319

Ablations:

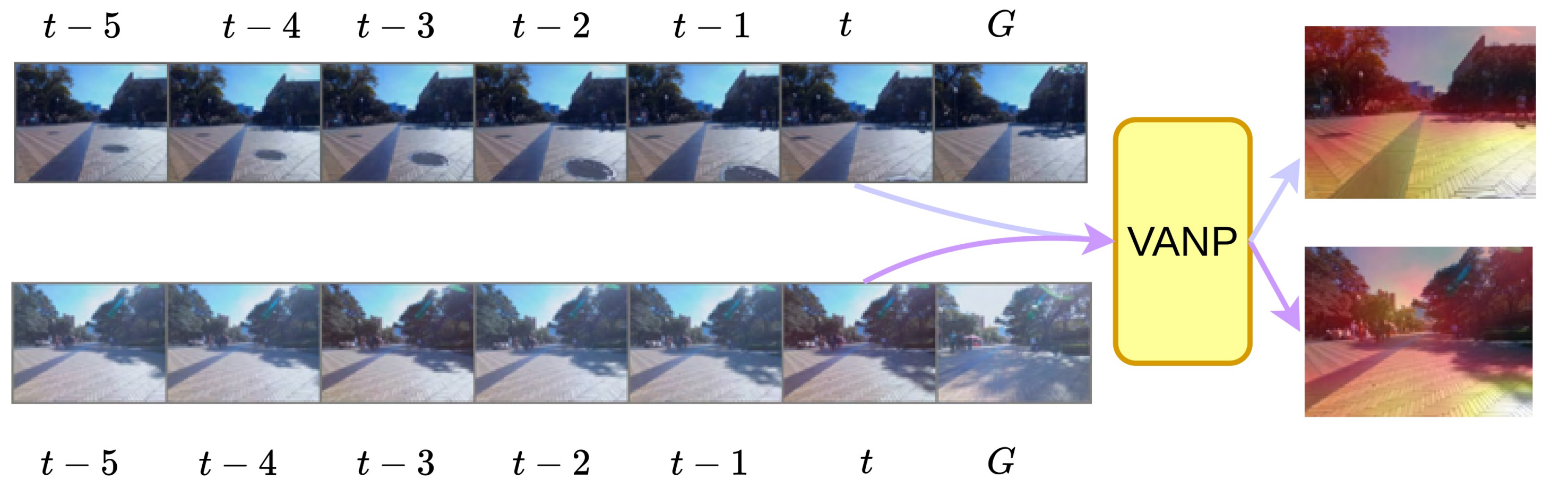
We ask multiple questions:

- Is goal embedding effective or not?
- How about robot's future actions?
- Should we use goal embedding inside Transformer or not?
- Can augmentations help?

Information	3s	5s
Actions	0.167	0.499
Goal	0.160	0.392
Actions+GoalIn	0.155	0.386
Actions+GoalOut	0.144	0.383
Augmentations	**0.133**	**0.342**

Limitation:

VANP does not perform well when there is only negligible inter-frame change:


$t-5$
$t-4$
$t-3$
$t-2$
$t-1$
t
G
VANP
$t-5$
$t-4$
$t-3$
$t-2$
$t-1$
t
G