Ask the Camera, Get the Why

Zero-Shot Tracking and Natural-Language Explanations

Nikos Deligiannis

ETRO - Vrije Universiteit Brussel & imec

VUB ETRO ELECTRONICS & INFORMATICS

imec

European Research Council
Established by the European Commission

Fondation Francqui Stichting

Acknowledgements

Fawaz Sammani
PhD Student
VUB-imec

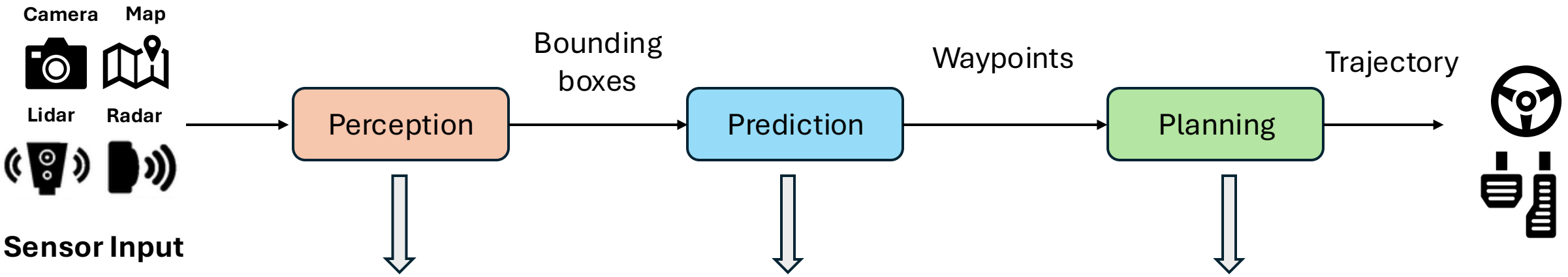
Tzoulio Chamiti
PhD Student
VUB-imec

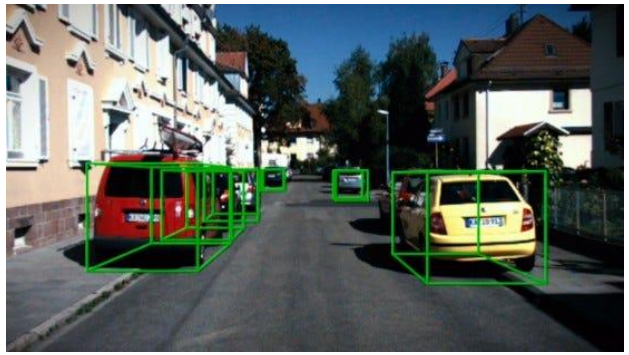
European Research Council
Established by the European Commission

Flanders AI
Research Program

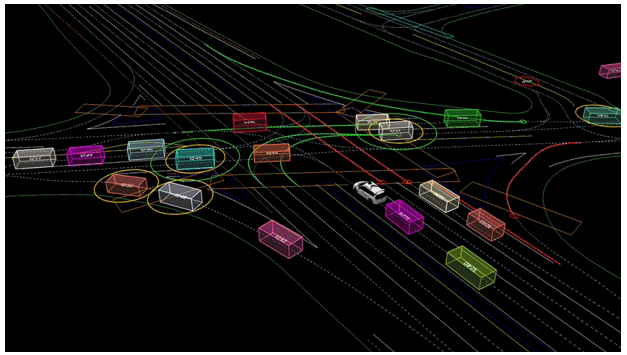
Fondation
Francqui
Stichting

Problem Setting | Autonomous Driving (AD) Tasks

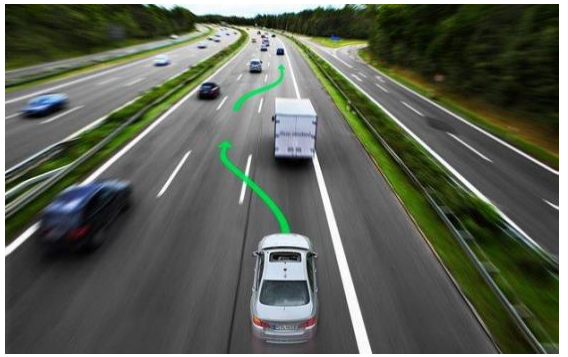




What is around?



Where will everything go?



Where should I go?

Trending in AD | Driving + Language

Honda Research Institute USA

Go straight at an intersection then turn left.
There are **construction cones** on the road.

HAD – human-to-vehicle driving advice, highlighting key objects.

BBD-X – one-sentence explanation of driving behavior

Berkeley DeepDrive

Action description: **Action justification:**

(1) The car is driving ***as*** *there is nothing to impede it.*

2019 — Planning — 2022 — Prediction — 2023 — Full-stack

Trending in AD | Driving + Language

Dataset	Source Dataset	# Frames	Avg. captions / QA per annotated frame	Total captions / QA in Perception	Total captions / QA in Prediction	Total captions / QA in Planning	Logic among captions/QA pairs
nuScenes-QA [47]	nuScenes	34,149	13.5	460k**	✗	✗	None
nuPrompt [66]	nuScenes	34,149	1.0	35k*	✗	✗	None
HAD [31]	HDD	25,549	1.8	25k	✗	20k	None
BDD-X [30]	BDD	26,228	1	26k	✗	✗	None
DRAMA [40]	DRAMA	17,785	5.8	85k	✗	17k	Chain
Rank2Tell [51]	Rank2Tell	5,800	-	-	✗	-	Chain
DriveLM-nuScenes	nuScenes	4,871	**91.4**	144k*	153k	146k	**Graph**
DriveLM-CARLA	CARLA	**183,373**	20.5	**2.46M****	**578k****	**714k****	**Graph**

DriveLM – perception-prediction-planning driving description with graph VQA

Talk2Car – a description of how to reach the goal point from current position.

DRAMA– caption about important objects and future decision.

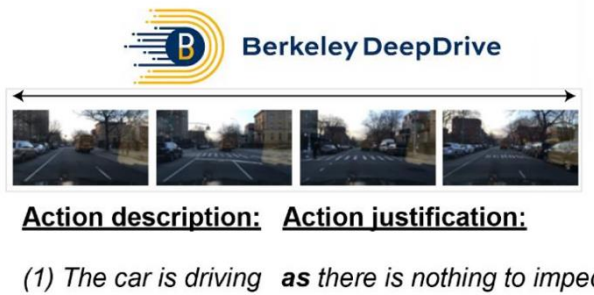
HAD – human-to-vehicle driving advice, highlighting key objects.

BBD-X – one-sentence explanation of driving behavior

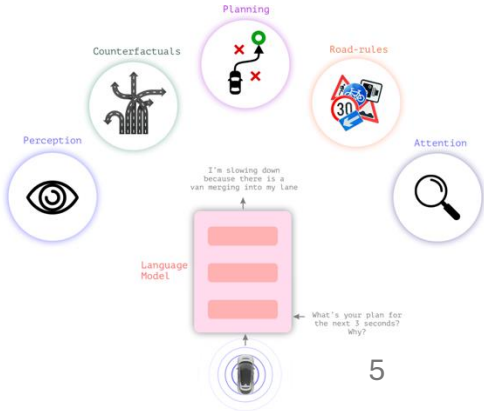
Planning — 2019

Prediction — 2022

Full-stack — 2023



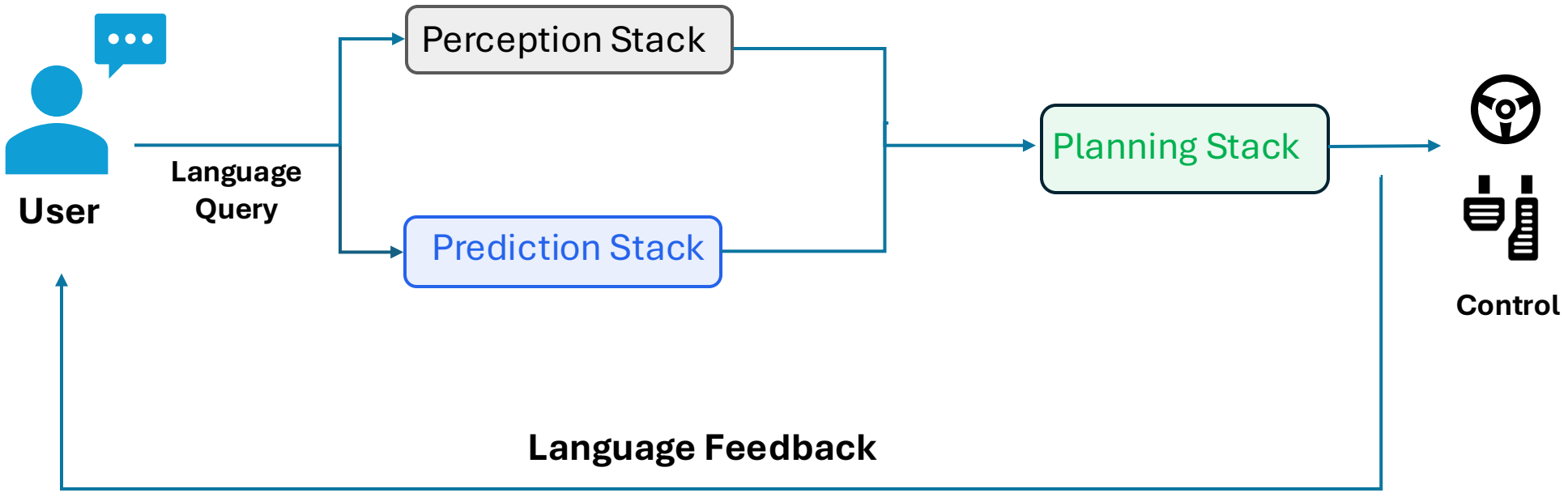
LINGO-1– commentary for explaining driving behaviours.



*Slide from ELM: Embodied Understanding of Driving Scenes Talk, OpenDriveLab ECCV 2024

Human-in-the-loop and AD

- Natural language supervision
- Flexible context-aware navigation and control
- Transparent decision-making through explanations



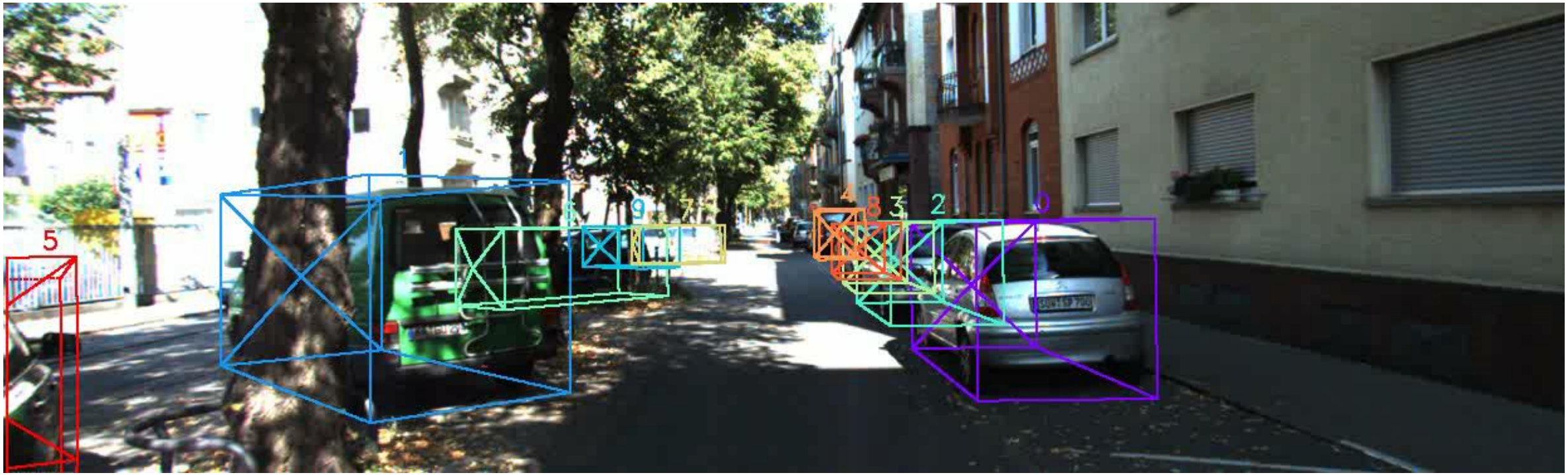
Human-in-the-loop and AD

User Query
(e.g. Navigation command, decision-making question)

System Feedback
(Natural language explanation)

Vehicle Multi-Object Tracking

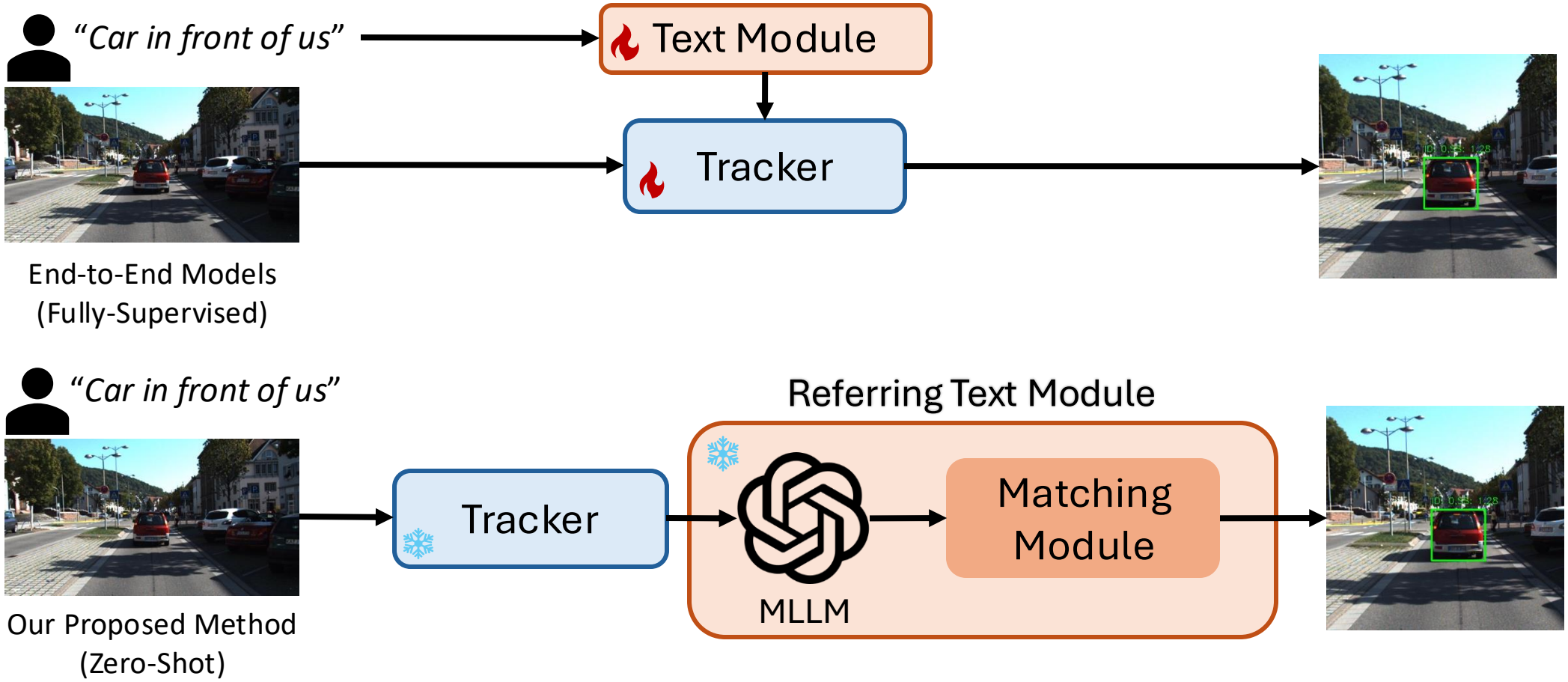
Multi-Object Tracking (MOT) of multiple vehicles using cameras and sensors.



Di Bella, Lyu, Cornelis, Munteanu. "HybridTrack: A Hybrid Approach for Robust Multi-Object Tracking." IEEE Robotics and Automation Letters, 2025.

Referring Vehicle Multi-Object Tracking

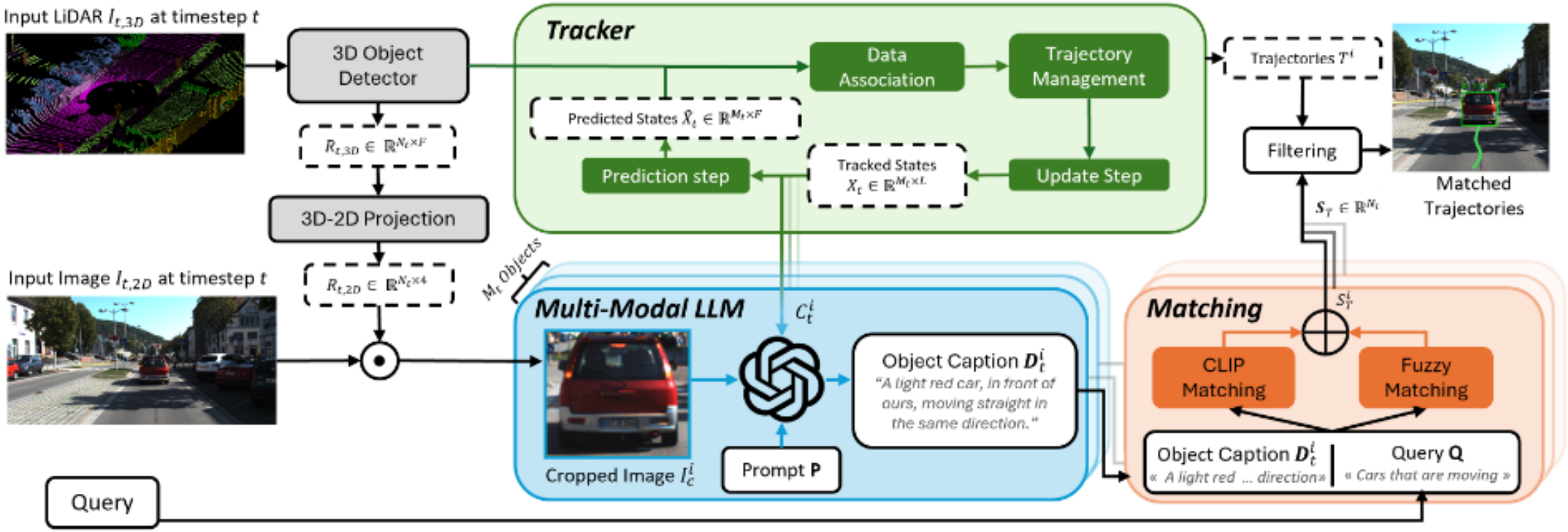
Language-guided Multi-Object Tracking



Chamiti, Di Bella, Munteanu, Deligiannis. "ReferGPT: Towards Zero-Shot Referring Multi-Object Tracking." CVPR Workshops (CVPRW) 2025

Referring Vehicle Multi-Object Tracking

Overview of the Method: Zero-Shot Modular Approach



Chamiti, Di Bella, Munteanu, Deligiannis. "ReferGPT: Towards Zero-Shot Referring Multi-Object Tracking." CVPR Workshops (CVPRW) 2025

Referring Vehicle Multi-Object Tracking

- First method that tackles the problem in a zero-shot fashion achieving competitive results:

Refer-KITTI

Method	E	HOTA ↑	Detection			Association			LocA
			DetA ↑	DetRe ↑	DetPr ↑	AssA ↑	AssRe ↑	AssPr ↑	
Traditional Methods									
FairMOT [40]	×	22.78	14.43	16.44	45.48	39.11	43.05	71.65	74.77
DeepSORT [33]	×	25.59	19.76	26.38	36.93	34.31	39.55	61.05	71.34
ByteTrack [41]	×	24.95	15.50	18.25	43.48	43.11	48.64	70.72	73.90
TransTrack [23]	×	32.77	23.31	32.33	42.23	45.71	49.99	78.74	79.48
Supervised Learning Methods									
iKUN [8]	×	48.84	35.74	51.97	52.26	**66.80**	72.95	87.09	-
MEX [27]	×	45.07	32.81	**62.52**	41.65	-	71.09	-	-
TransRMOT [34]	✓	46.56	37.97	49.69	60.10	57.33	60.02	**89.67**	90.33
EchoTrack [13]	✓	48.86	41.26	53.42	**62.83**	57.59	61.61	89.33	**90.74**
DeepRMOT [10]	✓	39.55	30.12	41.91	47.47	53.23	58.47	82.16	80.49
TempRMOT [42]	✓	**52.21**	**43.73**	55.65	59.25	66.75	71.82	87.76	90.40
ROMOT [12]	✓	35.5	28.30	-	-	46.2	-	-	-
MGLT [5]	✓	49.25	37.09	-	-	65.50	-	-	-
Zero-Shot Learning Methods									
Baseline*	×	21.42	9.48	16.10	18.20	48.82	57.86	72.57	80.72
ReferGPT$_{2D}$ (Ours)	×	46.36	36.58	51.40	52.16	59.00	73.16	69.31	83.26
ReferGPT$_{3D}$ (Ours)	×	49.46	39.43	50.21	58.91	62.57	**73.74**	72.78	81.85

Referring Vehicle Multi-Object Tracking

- Our strength lies in generalizing to open-set queries as it's not feasible to train a model for every possible user query

Refer-KITTIv2

Method	E	HOTA ↑	Detection			Association			LocA
			DetA ↑	DetRe ↑	DetPr ↑	AssA ↑	AssRe ↑	AssPr ↑	
Traditional Methods									
FairMOT [40]	×	22.53	15.80	20.60	37.03	32.82	36.21	71.94	78.28
ByteTrack [41]	×	24.59	16.78	22.60	36.18	36.63	41.00	69.63	78.00
Supervised Learning Methods									
iKUN [8]	×	10.32	2.17	2.36	19.75	49.77	58.48	68.64	74.56
TransRMOT [34]	✓	31.00	19.40	**36.41**	28.97	49.68	54.59	**82.29**	89.82
TempRMOT [42]	✓	**35.04**	**22.97**	34.23	**40.41**	53.58	59.50	81.29	**90.07**
Zero-Shot Learning Methods									
ReferGPT (Ours)	×	30.12	15.69	21.55	34.41	**59.02**	**74.59**	68.20	79.76

Refer-KITTI+

Method	E	HOTA ↑	Detection			Association			LocA
			DetA ↑	DetRe ↑	DetPr ↑	AssA ↑	AssRe ↑	AssPr ↑	
Traditional Methods									
TransRMOT [34]	✓	35.32	25.61	40.05	38.45	50.33	55.40	81.23	79.44
EchoTrack [13]	✓	37.46	28.83	**39.83**	46.70	50.39	54.14	**82.57**	79.97
Zero-Shot Learning Methods									
ReferGPT (Ours)	×	**43.44**	**29.89**	36.59	**56.98**	**63.60**	**75.20**	73.27	**82.23**

Referring Vehicle Multi-Object Tracking

Query: Cars looking in the same direction as us

Referring Vehicle Multi-Object Tracking

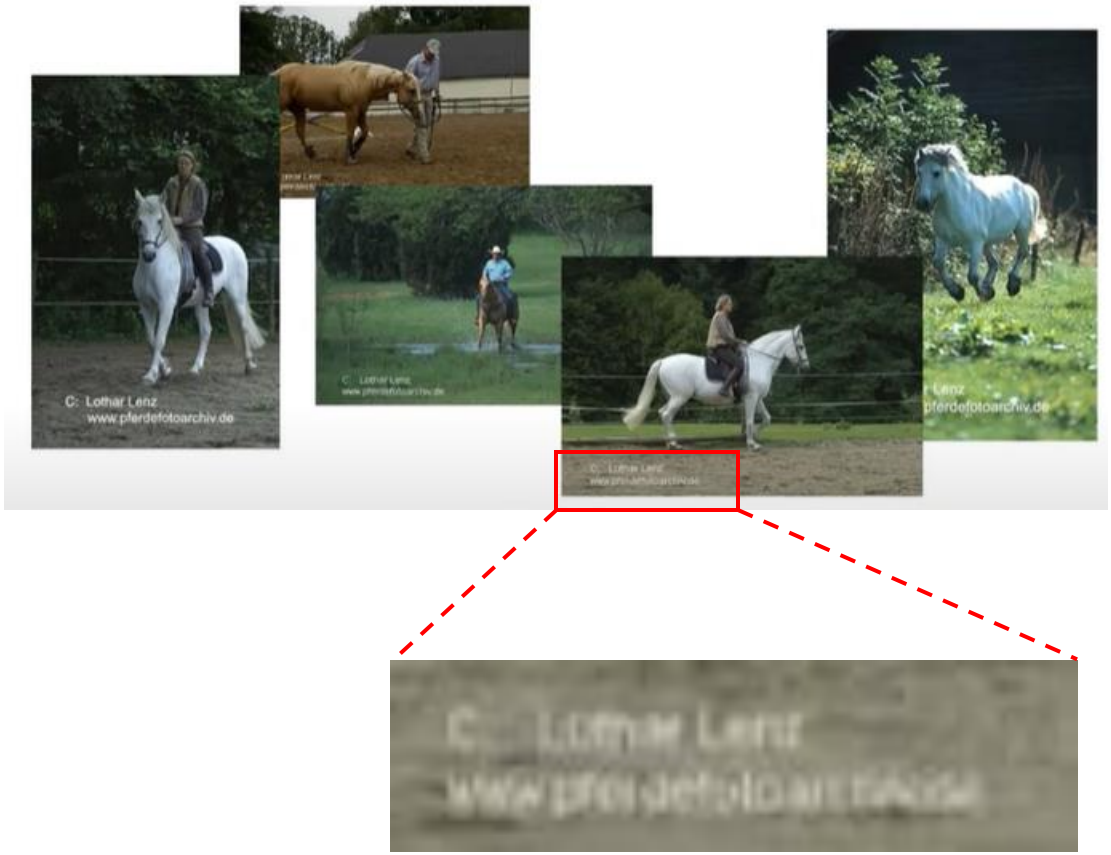
Query: Black cars

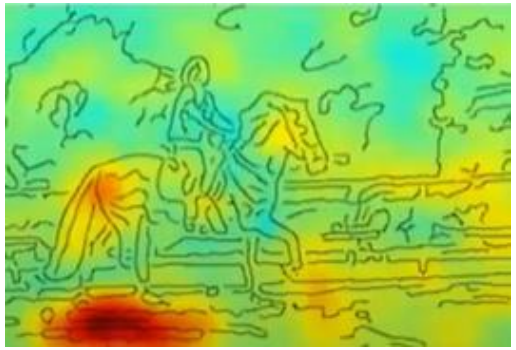
Referring Vehicle Multi-Object Tracking

Query: Car driving in front of us

Why Explainable AI

PASCAL VOC competition

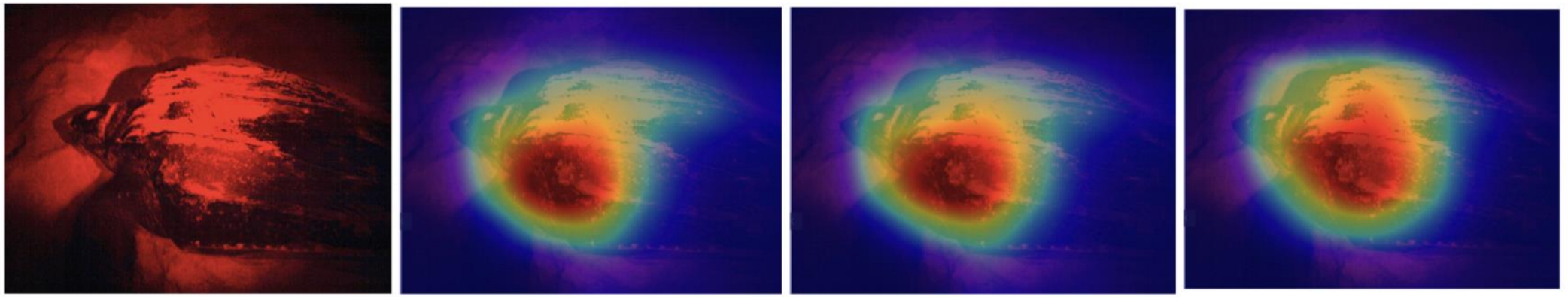




why?

Heatmap

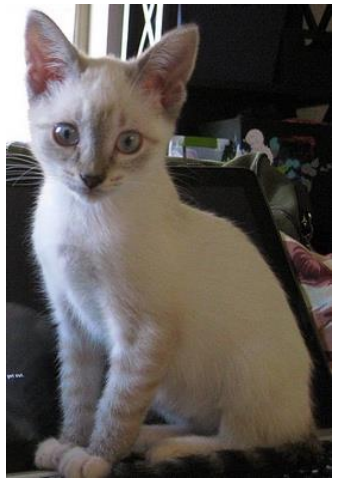
The problem with heatmaps as Explanations



NLE (Ours): this is a volcano because it shows fire burning on the volcano lava flows, a bright red glow erupts

We cannot understand why the prediction was wrong, just by looking at the heatmaps.

Natural Language Explanations



What animal is this?

VQA System

cat

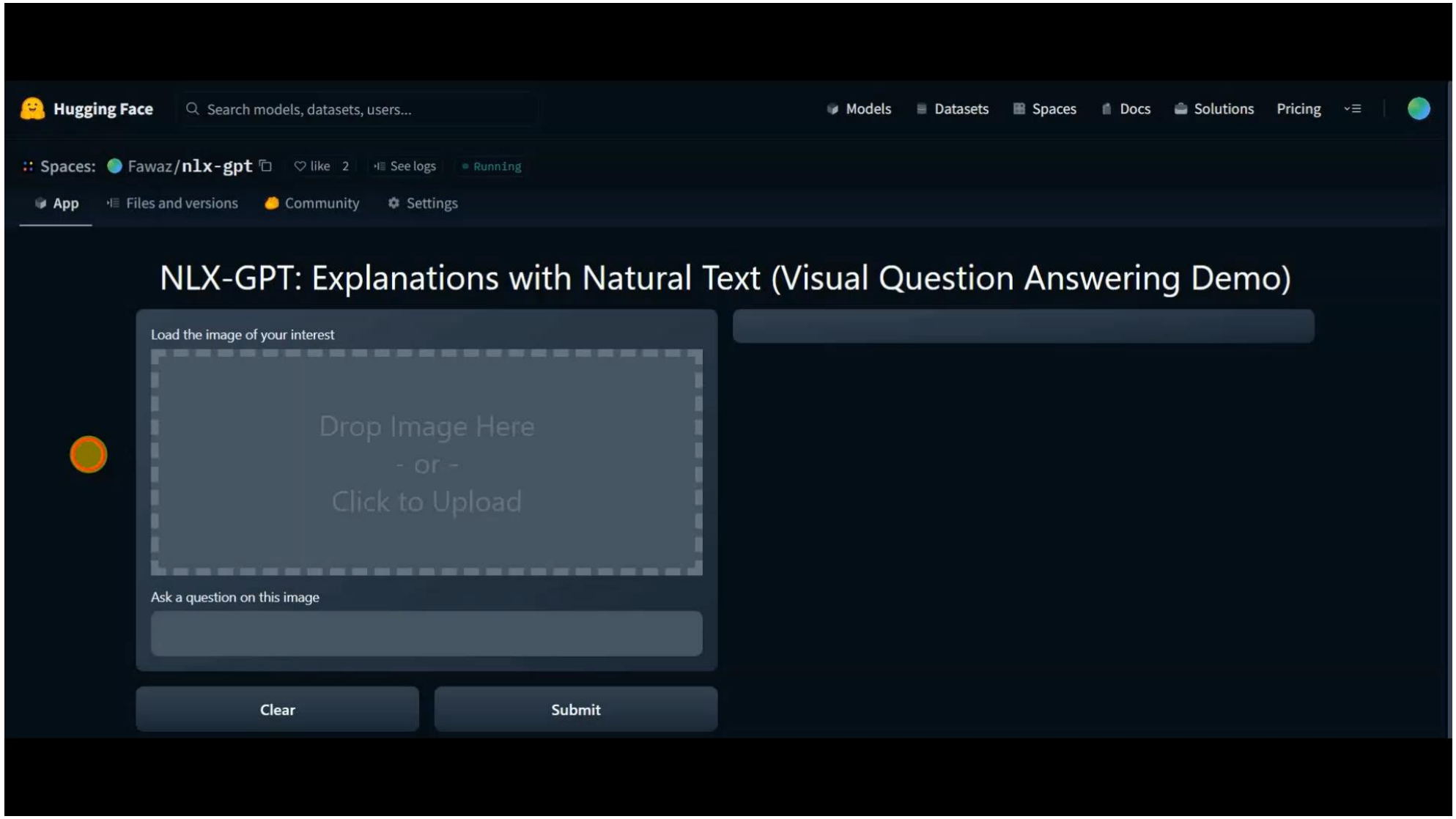
Why did you say this?

it is a furry animal that has a long tail, sharp claws and has whiskers

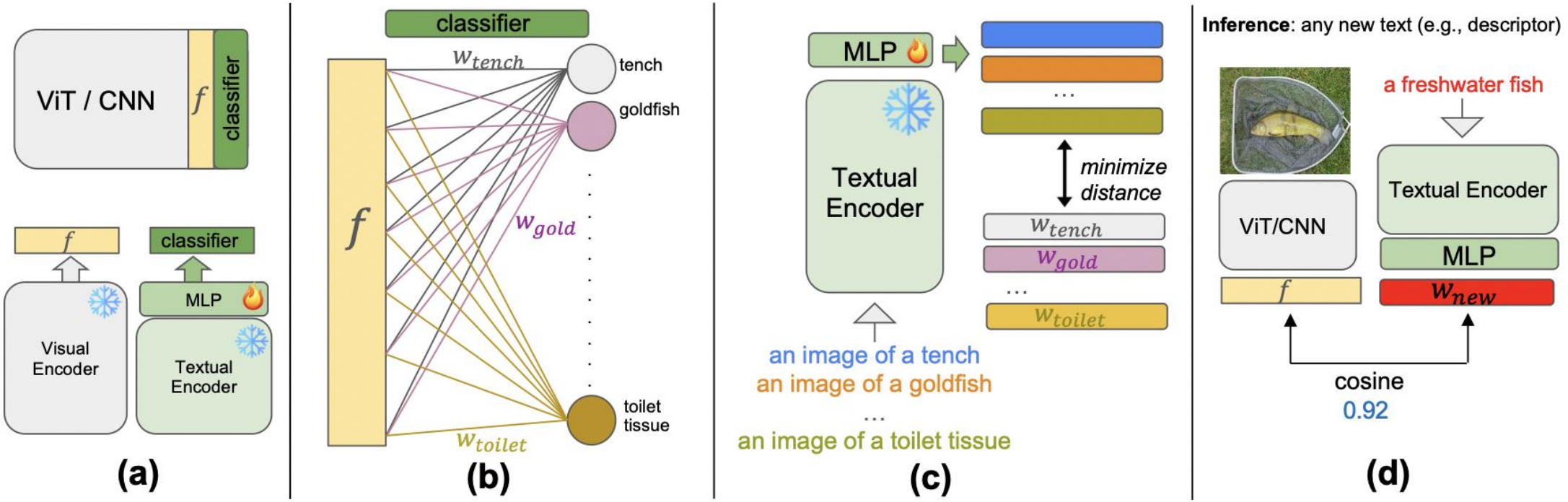
These need to be trained..

Sammani, Mukherjee, Deligiannis. "NLX-GPT: A Model for Natural Language Explanations in Vision and Vision-Language Tasks." CVPR 2022.

NLX-GPT Demo


Hugging Face
Search models, datasets, users...
Models
Datasets
Spaces
Docs
Solutions
Pricing
Spaces: Fawaz/nlx-gpt
like 2
See logs
Running
App
Files and versions
Community
Settings
NLX-GPT: Explanations with Natural Text (Visual Question Answering Demo)
Load the image of your interest
Drop Image Here
- or -
Click to Upload
Ask a question on this image
Clear
Submit

Vision Models Talk Through Their Weights



We take any text encoder

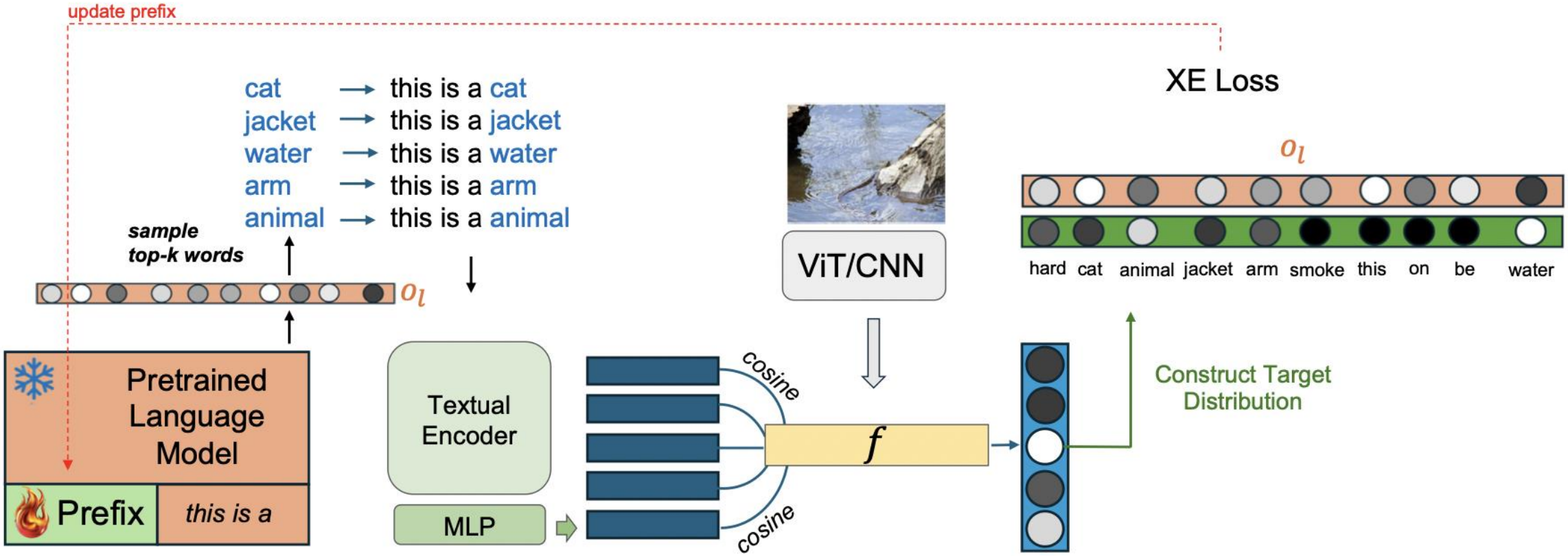
The classifier weights are vectors representing text

We train to predict those weights from the encoder output

We generalize to new text to provide explanations

Sammani and Deligiannis. "Zero-Shot Natural Language Explanations." International Conference on Learning Representations (ICLR), 2025.

We Can Use This For Zero-Shot Explanations



Sammani and Deligiannis. "Zero-Shot Natural Language Explanations." International Conference on Learning Representations (ICLR), 2025.

Annotation-free Textual Explanation Evaluation

Model	Metric	NLX-GPT	NLX-GPT$_{ft}$	Uni-NLX	Uni-NLX$_{ft}$	Ours
ResNet50	CLIP-S ↑	28.53	28.52	28.38	28.52	**30.56**
	LPIPS(A) ↓	0.721	0.719	0.720	0.716	**0.706**
	LPIPS(S) ↓	0.624	0.623	0.620	0.620	**0.614**
	Cosine ↑	0.703	0.704	0.703	**0.710**	0.708
ResNeXt50$_{32x4d}$	CLIP-S ↑	28.53	28.52	28.38	28.52	**30.73**
	LPIPS(A) ↓	0.717	0.717	0.720	0.719	**0.707**
	LPIPS(S) ↓	0.621	0.620	0.622	0.623	**0.616**
	Cosine ↑	**0.680**	0.677	0.677	0.677	0.676
DenseNet161	CLIP-S ↑	28.53	28.52	28.38	28.52	**30.56**
	LPIPS(A) ↓	0.717	0.718	0.718	0.723	**0.706**
	LPIPS(S) ↓	0.620	0.620	0.620	0.624	**0.616**
	Cosine ↑	0.589	**0.591**	0.589	0.588	0.588
DINOv2-B	CLIP-S ↑	28.53	28.52	28.38	28.52	**30.15**
	LPIPS(A) ↓	0.719	0.716	0.717	0.716	**0.712**
	LPIPS(S) ↓	0.622	0.621	**0.619**	**0.619**	**0.619**
	Cosine ↑	**0.362**	0.360	0.354	0.356	0.314
ViT-B/16$_{pt}$	CLIP-S ↑	28.53	28.52	28.38	28.52	**30.92**
	LPIPS(A) ↓	0.723	0.723	0.717	0.722	**0.710**
	LPIPS(S) ↓	0.624	0.626	**0.619**	0.624	0.621
	Cosine ↑	**0.450**	0.440	0.447	0.445	0.425

Sammani and Deligiannis. "Zero-Shot Natural Language Explanations." International Conference on Learning Representations (ICLR), 2025.

Visual Results

Is the object a car?
Yes, because it has a boxy vehicle shape and reflective headlights

What is the object?
A human because it has an upright human-like silhouette

Visual Results

Is it raining?

yes because the ground is wet and people are using umbrellas

What beverage is in the cup?

coffee because it is a dark liquid with a dark brown top

Thank you for your attention

nikos.deligiannis@vub.be

VUB ETRO ELECTRONICS & INFORMATICS imec