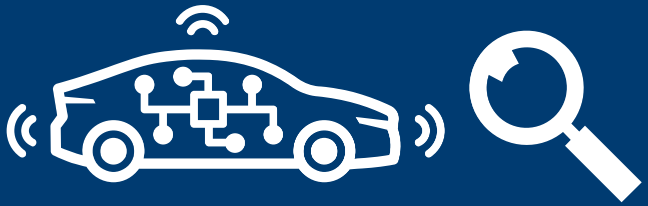
Explaining vision-based driving models



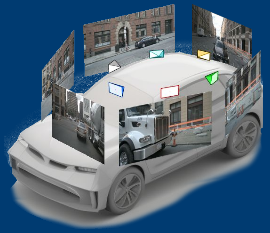
Trustworthy AI Symposium
2025 January, 21st

Éloi Zablocki

valeo.ai

Explaining self-driving cars

Why? What?



Drive

Why do we need explanations?

- High-stake and safety critical application
- Cannot test every situation
- training objectives ≠ real-world goals
- find model flaws

What are explanations?

Interpretable: Understandable by humans

Faithful: Accurately reflects the model’s processing

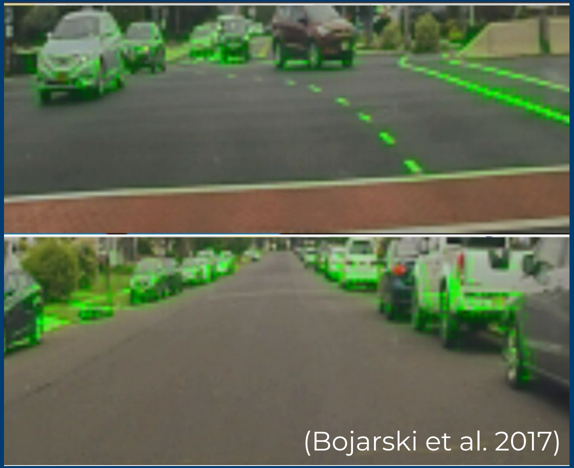
Local or Global: Explain a single input or the model in general?

Post-hoc or transparency: Explain a given black-box model, or design a transparent model?

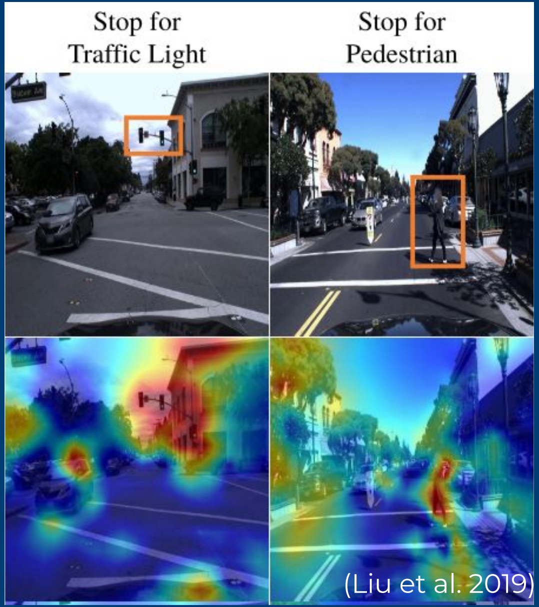
Zablocki et al., Explainability of deep vision-based autonomous driving systems: Review and challenges, IJCV 2022

Input attribution methods

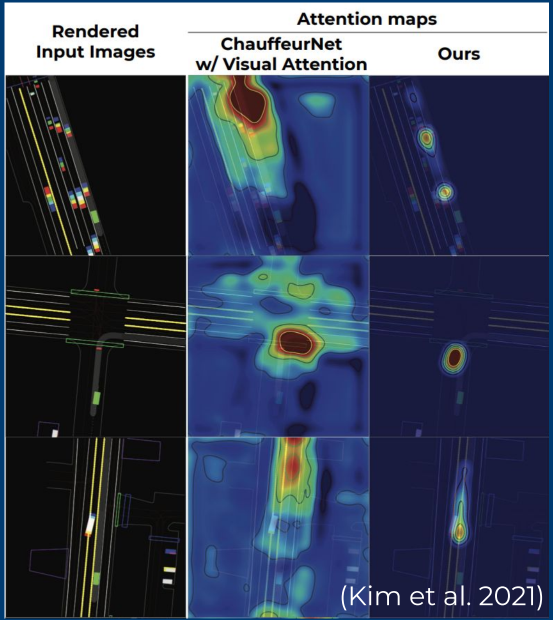
Where does the model look? post-hoc explainability



(Bojarski et al. 2017)



(Liu et al. 2019)



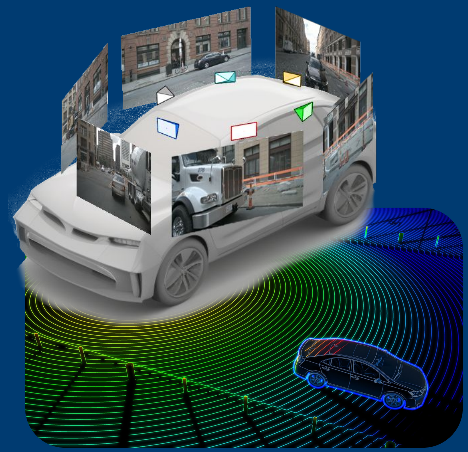
(Kim et al. 2021)

- ✓ Shows where the model look
- ✓ Easy to compute
- ✗ Saliency maps must be interpreted
- ✗ Not always faithful to the model*

Adebayo et al., Sanity Checks for Saliency Maps, NeurIPS 2018

Driving models explainable by-design

Language-based explanations



Jointly drive and explain

- ✓ Model is self-explainable
- ✓ High-interpretability

- ✗ May sacrifice driving accuracy
- ✗ Potential faithfulness issues

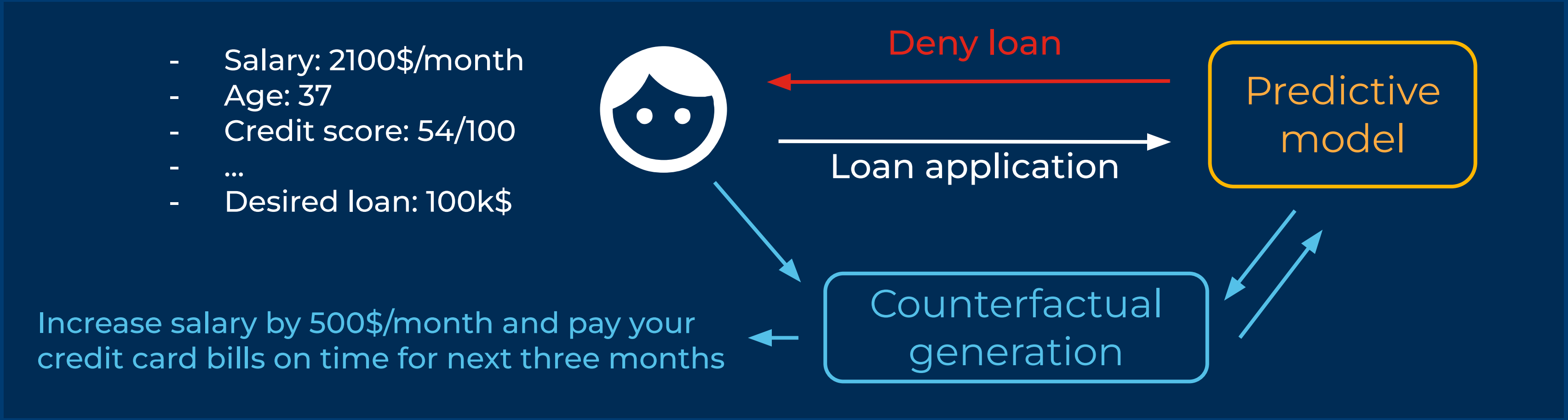
—

Ben-Younes et al., Driving Behavior Explanation with Multi-level Fusion, Pattern Recognition 2022
Wayve.ai, LINGO-1: Exploring Natural Language for Autonomous Driving, blog 2023

Counterfactual explanations

Definition

A *counterfactual explanation* shows minimal and meaningful changes in an input leading the model to change its output.



How to scale to driving models?
And complex images?

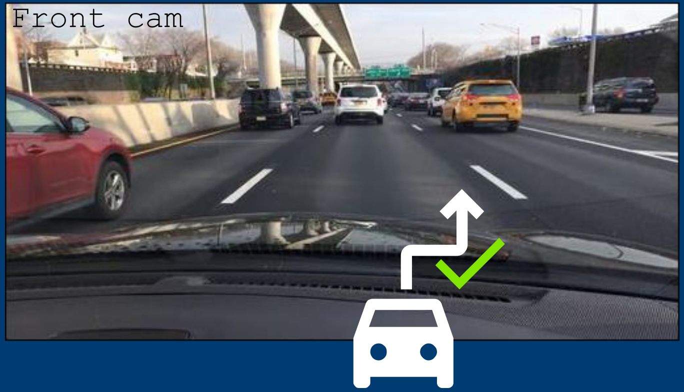
Wachter et al., *Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR, Harvard Journal of Law & Technology 2017*

Counterfactual explanations

Challenges for complex vision models

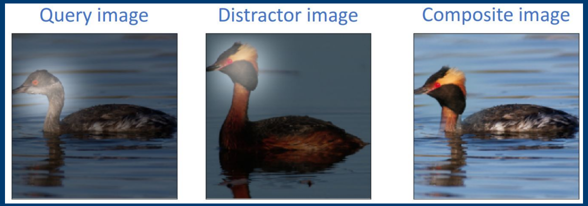
Adversarial Attacks

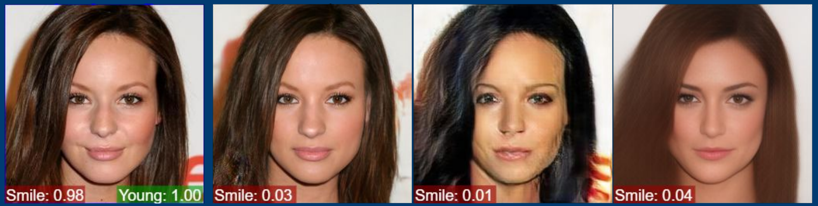
Original image



Adversarial attack

Simple domains





Goodfellow et al., Explaining and Harnessing Adversarial Examples, ICLR 2015
Goyal et al., Counterfactual Visual Explanations, ICML 2019
Rodriguez et al., Beyond Trivial Counterfactual Explanations With Diverse Valuable Explanations, ICCV 2021

Counterfactual explanations

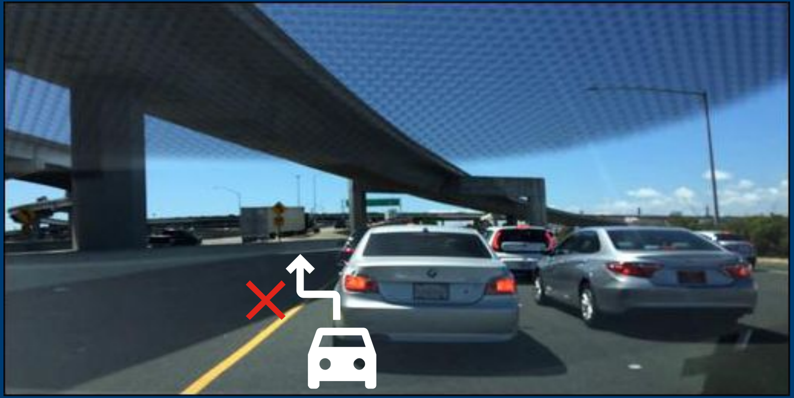
STEEX and OCTET

I **cannot** go to the left lane

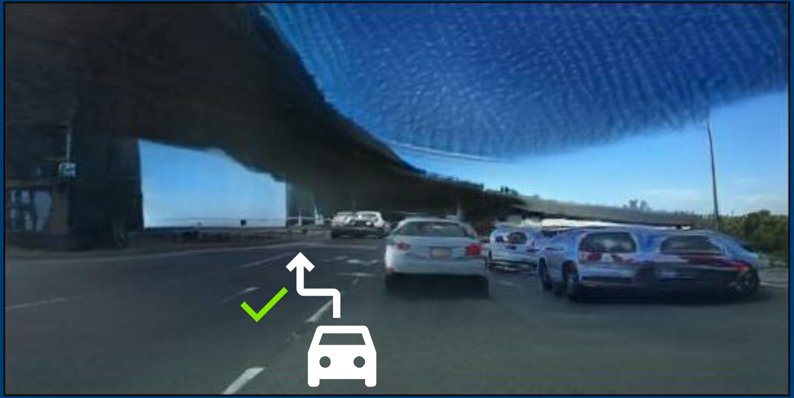
What should be different such that you **could** go to the left lane?

If I was seeing this,
I **could** go to the left lane

Original image



Counterfactual explanation



—

Jacob et al., STEEX: Steering Counterfactual Explanations with Semantics, ECCV 2022
Zemni et al., OCTET: Object-aware Counterfactual Explanations, CVPR 2023

Counterfactual explanations

Region and object-targeted explanations

I **can** go to the left lane

What should be different such that you **could not** go to the left lane?

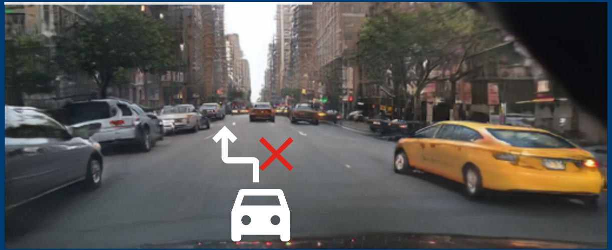
If I was seeing this,
I **could not** go to the left lane

Counterfactual explanations

Target: road

Target: yellow car

No target

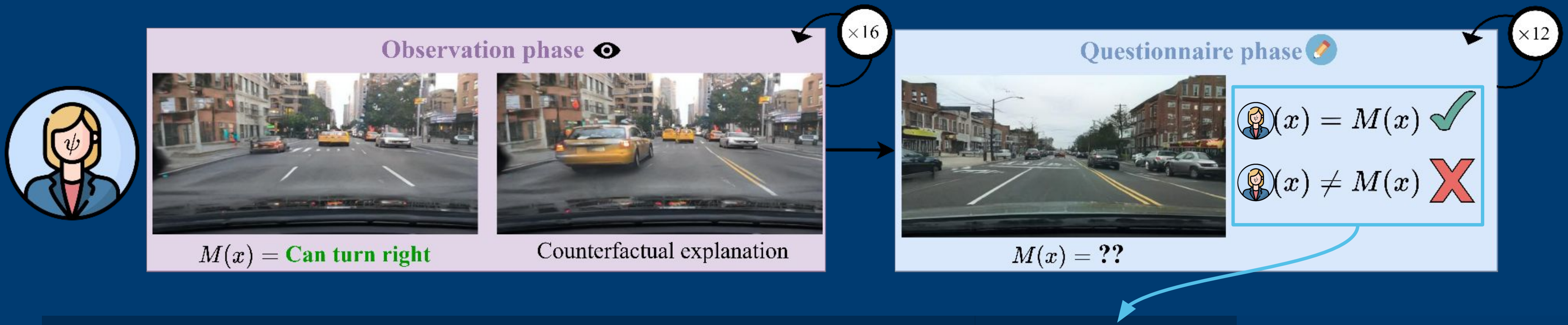


—

Jacob et al., STEEX: Steering Counterfactual Explanations with Semantics, ECCV 2022
Zemni et al., OCTET: Object-aware Counterfactual Explanations, CVPR 2023

Can counterfactuals help to better "understand" a model?

"Understand" := Ability to predict model's decision on new instances (simulatability)

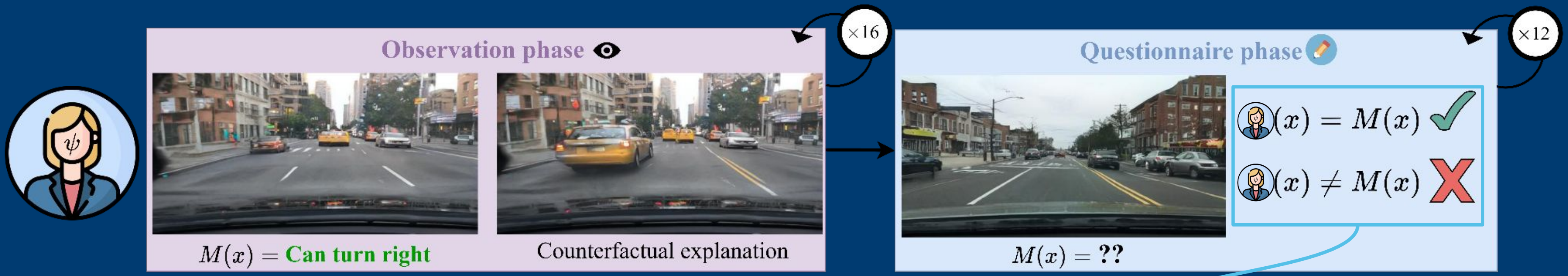


	Cohort size	Replication	Bias Detection
Control group (without explanations)	20	52%	0%
Group with counterfactual explanations	20	70%	65%

—
Fel et al., What I cannot predict, I do not understand: A human centered evaluation framework for explainability method, NeurIPS 2022
Zemni et al., OCTET: Object-aware Counterfactual Explanations, CVPR 2023

Can counterfactuals help to better "understand" a model?

"Understand" := Ability to predict model's decision on new instances (simulatability)



	Cohort size	Replication	Bias Detection
Control group (without explanations)	20	52%	0%
Group with counterfactual explanations	20	70%	65%

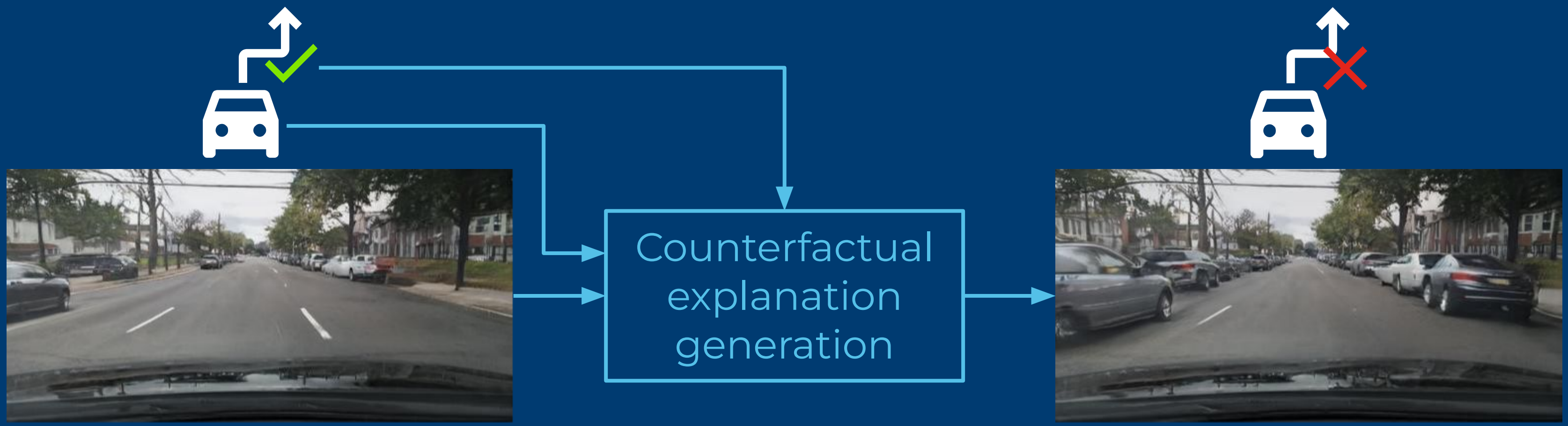
Unknown to the participants, the classifier is flawed: obstacles on <u>both sides</u> of the road influence the "Can turn right" prediction. **Did users find out?**

—

Fel et al., What I cannot predict, I do not understand: A human centered evaluation framework for explainability method, NeurIPS 2022
Zemni et al., OCTET: Object-aware Counterfactual Explanations, CVPR 2023

GIFT: Global Interpretable Faithful Textual Explanations

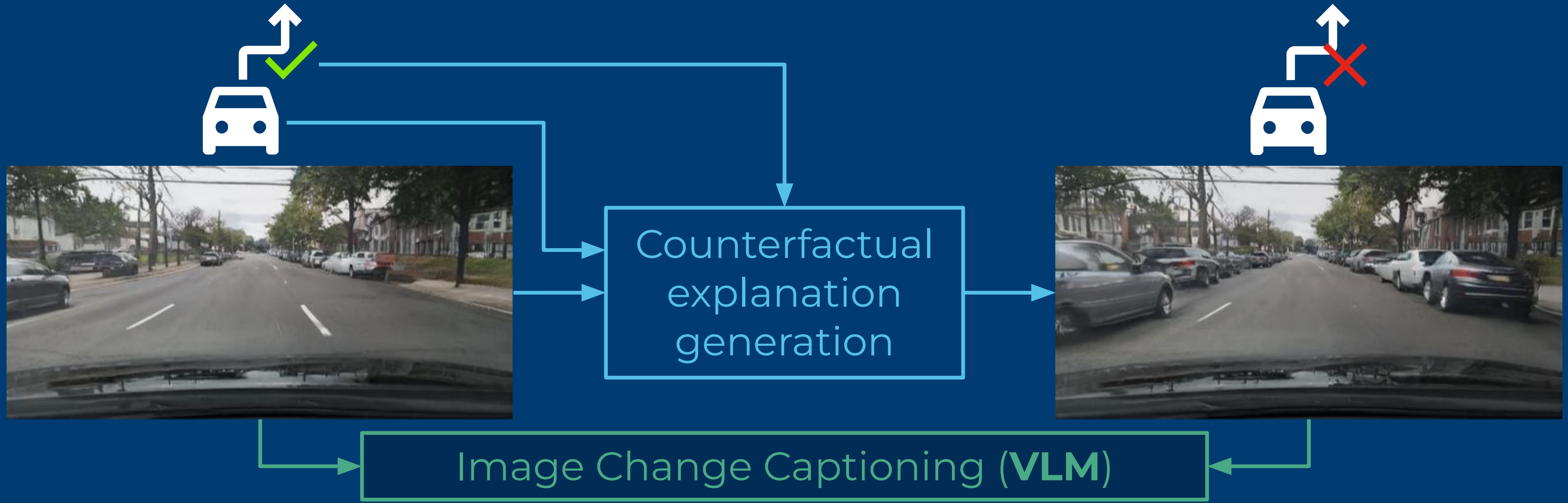
Gathering local faithful explanations



Zablocki et al., GIFT: A Framework for Global Interpretable Faithful Textual Explanations of Vision Classifiers, preprint 2024

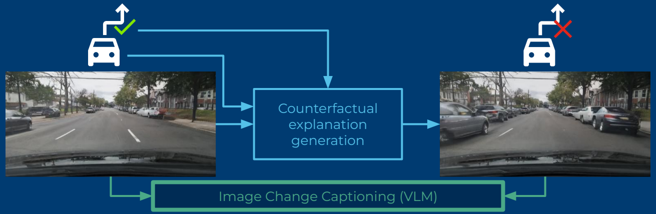
GIFT: Global Interpretable Faithful Textual Explanations

Gathering local faithful explanations

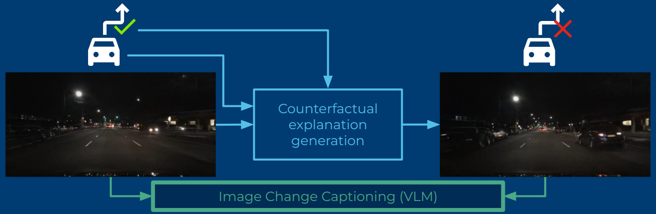


road -> middle lane -> was empty, now has parked cars on both sides
cars -> leftmost parked -> appeared
cars -> rightmost parked -> appeared
buildings -> leftmost -> slightly closer
buildings -> rightmost -> slightly closer
streetlights -> leftmost -> more visible; closer
streetlights -> rightmost -> more visible; closer
sky -> color -> less bright; slightly more cloudy
ego-car -> dashboard -> brighter

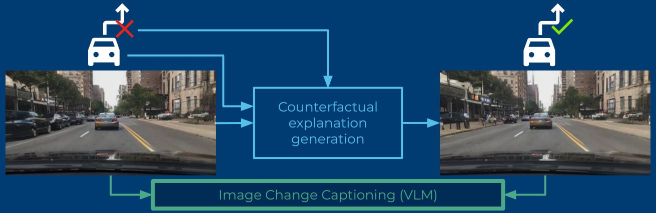
Zablocki et al., GIFT: A Framework for Global Interpretable Faithful Textual Explanations of Vision Classifiers, preprint 2024



road -> middle lane -> was empty, now has parked cars on both sides
cars -> leftmost parked -> appeared
cars -> rightmost parked -> appeared
buildings -> leftmost -> slightly closer
buildings -> rightmost -> slightly closer
streetlights -> leftmost -> more visible; closer
streetlights -> rightmost -> more visible; closer
sky -> color -> less bright; slightly more cloudy
ego-car -> dashboard -> brighter



sky -> slightly darker
street -> has a stopped vehicle ahead, previously no such vehicle
street -> has multiple stationary vehicles on the left side, previously had fewer
street -> has a car moving towards on the right side, previously no car was moving towards
street -> has more vehicles on the left side, previously fewer
street -> road surface appears darker, more reflective
streetlights -> appear brighter, more numerous
buildings -> appear closer, more visible
ego-car -> dashboard -> less visible



road -> leftmost lane -> car disappeared
road -> middle lane -> car's color appears slightly more red
road -> leftmost sidewalk -> people disappeared
traffic signal -> leftmost -> turned green
ego-car -> slightly shifted right

Zablocki et al., GIFT: A Framework for Global Interpretable Faithful Textual Explanations of Vision Classifiers, preprint 2024

GIFT

Candidate global explanations

[...] Identify the main factors leading the classifier to choose class 0 or 1. [...]

From "0" to "1":

road -> middle lane -> was empty, now has parked cars on both sides
cars -> leftmost parked -> appeared
cars -> rightmost parked -> appeared
buildings -> leftmost -> slightly closer
buildings -> rightmost -> slightly closer
streetlights -> leftmost -> more visible; closer
streetlights -> rightmost -> more visible; closer
sky -> color -> less bright; slightly more cloudy
ego-car -> dashboard -> brighter

From "0" to "1":

sky -> slightly darker
street -> has a stopped vehicle ahead, previously no such vehicle
street -> has multiple stationary vehicles on the left side, previously had fewer
street -> has a car moving towards on the right side, previously no car was moving towards
street -> has more vehicles on the left side, previously fewer
street -> road surface appears darker, more reflective
streetlights -> appear brighter, more numerous
buildings -> appear closer, more visible
ego-car -> dashboard -> less visible

From "1" to "0":

road -> leftmost lane -> car disappeared
road -> middle lane -> car's color appears slightly more red
road -> leftmost sidewalk -> people disappeared
traffic signal -> leftmost -> turned green
ego-car -> slightly shifted right

. . .

LLM

Zablocki et al., GIFT: A Framework for Global Interpretable Faithful Textual Explanations of Vision Classifiers, preprint 2024

GIFT

Candidate global explanations

[...] Identify the main factors leading the classifier to choose class 0 or 1. [...]

From "0" to "1":

road -> middle lane -> was empty, now has parked cars on both sides
cars -> leftmost parked -> appeared
cars -> rightmost parked -> appeared
buildings -> leftmost -> slightly closer
buildings -> rightmost -> slightly closer
streetlights -> leftmost -> more visible; closer
streetlights -> rightmost -> more visible; closer
sky -> color -> less bright; slightly more cloudy
ego-car -> dashboard -> brighter

From "0" to "1":

sky -> slightly darker
street -> has a stopped vehicle ahead, previously no such vehicle
street -> has multiple stationary vehicles on the left side, previously had fewer
street -> has a car moving towards on the right side, previously no car was moving towards
street -> has more vehicles on the left side, previously fewer
street -> road surface appears darker, more reflective
streetlights -> appear brighter, more numerous
buildings -> appear closer, more visible
ego-car -> dashboard -> less visible

From "1" to "0":

road -> leftmost lane -> car disappeared
road -> middle lane -> car's color appears slightly more red
road -> leftmost sidewalk -> people disappeared
traffic signal -> leftmost -> turned green
ego-car -> slightly shifted right

. . .

LLM

The presence of the following may explain *"Cannot turn right"*

Dense Traffic
Dense Traffic in left lane
Dense Traffic in middle lane
Dense traffic close to ego
Stopped vehicles
Red traffic lights
Ego-car dashboard is bright
Wet road
Dark road
Many buildings
Many streetlights
Pedestrians on the road or sidewalks
Objects on the road or sidewalks

Zablocki et al., GIFT: A Framework for Global Interpretable Faithful Textual Explanations of Vision Classifiers, preprint 2024

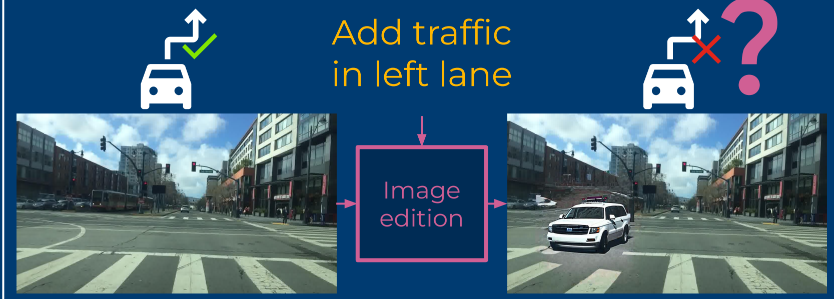
GIFT

Explanation verification

The presence of the following may explain *"Cannot turn right"*	Causal concept effect (%)
Dense Traffic	51
Dense Traffic in left lane	**45**
Dense Traffic in middle lane	✗
Dense traffic close to ego	27
Stopped vehicles	✗
Red traffic lights	✗
Ego-car dashboard is bright	✗
Wet road	✗
Dark road	✗
Many buildings	✗
Many streetlights	✗
Pedestrians on the road or sidewalks	✗
Objects on the road or sidewalks	✗

Concepts do not correlate with the class

Causal concept effect measures classification change caused by image intervention



0 → no causal effect
100 → perfect causal effect

Zablocki et al., GIFT: A Framework for Global Interpretable Faithful Textual Explanations of Vision Classifiers, preprint 2024

GIFT

Explanation verification

The presence of the following may explain *“Cannot turn right”*	Causal concept effect (%)
Dense Traffic	51
Dense Traffic in left lane	**45** ✓
Dense Traffic in middle lane	✗
Dense traffic close to ego	27
Stopped vehicles	✗
Red traffic lights	✗
Ego-car dashboard is bright	✗
Wet road	✗
Dark road	✗
Many buildings	✗
Many streetlights	✗
Pedestrians on the road or sidewalks	✗
Objects on the road or sidewalks	✗

	Bias Detection
Control group (without explanations)	0%
Group with counterfactual explanations	65%
With GIFT explanations	100%

Zablocki et al., GIFT: A Framework for Global Interpretable Faithful Textual Explanations of Vision Classifiers, preprint 2024

Conclusion

		Type	Scope	Interpretability	Faithful
	Input attribution	Post-hoc	Local	Low	No
	Driving models explainable by-design	By-design	Local	High	No
	Counterfactual explanations	Post-hoc	Local	Average	Yes
Dense Traffic in left lane → 47% ...	GIFT explanations	Post-hoc	Global	High	Yes

Ben-Younes et al., Driving Behavior Explanation with Multi-level Fusion, PR 2022
Zablocki et al., Explainability of deep vision-based autonomous driving systems: Review and challenges, IJCV 2022
Jacob et al., STEEX: Steering Counterfactual Explanations with Semantics, ECCV 2022
Zemni et al., OCTET: Object-aware Counterfactual Explanations, CVPR 2023
Zablocki et al., GIFT: A Framework for Global Interpretable Faithful Textual Explanations of Vision Classifiers, preprint 2024