JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

OCTET: Object-aware Counterfactual Explanations

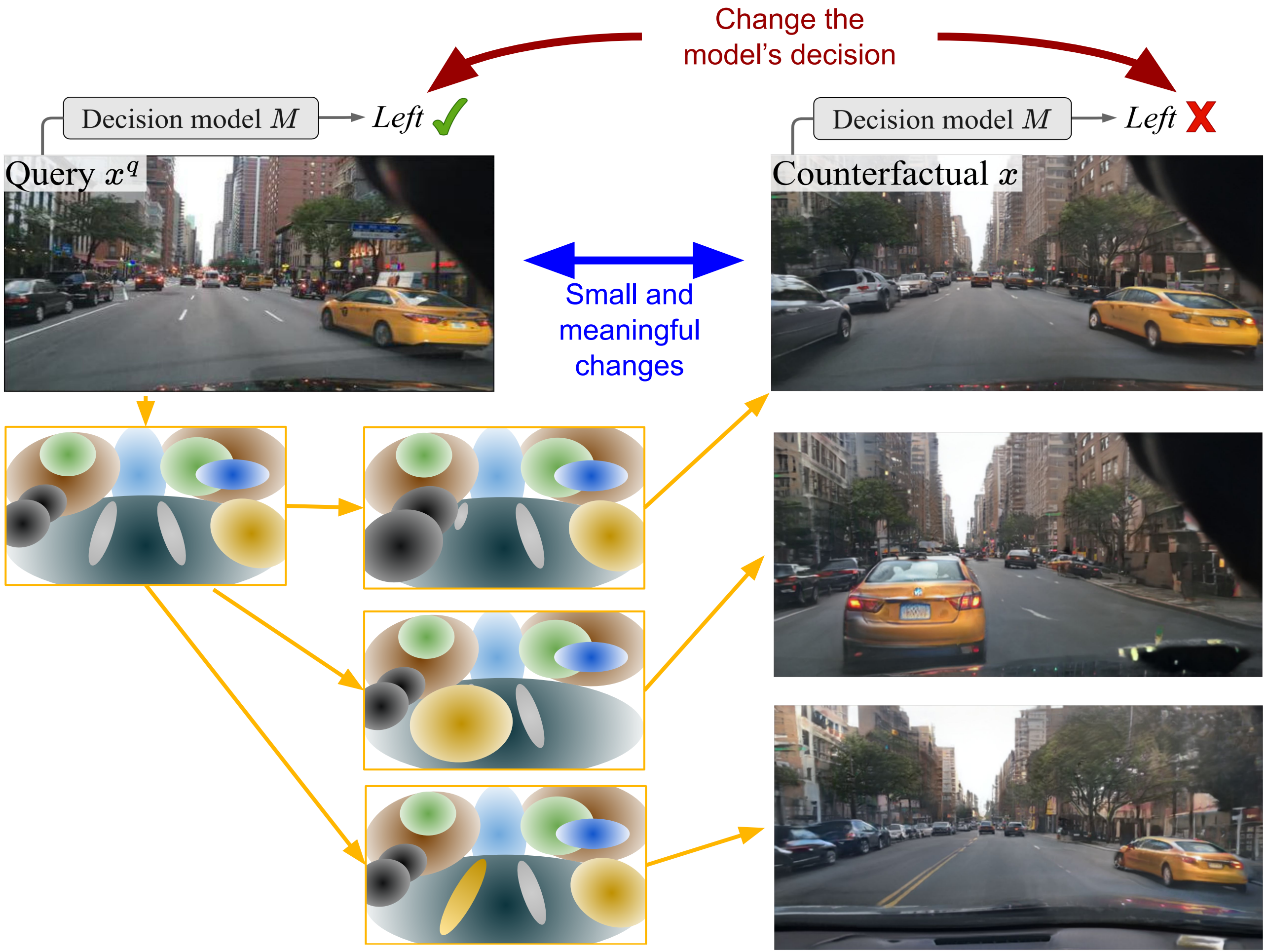
WED-PM-258

Mehdi Zemni, Mickaël Chen, Éloi Zablocki, Hédi Ben-Younes, Patrick Pérez, Matthieu Cord

valeo.ai

SORBONNE
UNIVERSITÉ

OCTET: Object-aware counterfactual explanations



Change the model's decision

Decision model M → *Left* ✓

Query x^q

Small and meaningful changes

Decision model M → *Left* ✗

Counterfactual x

Contributions:

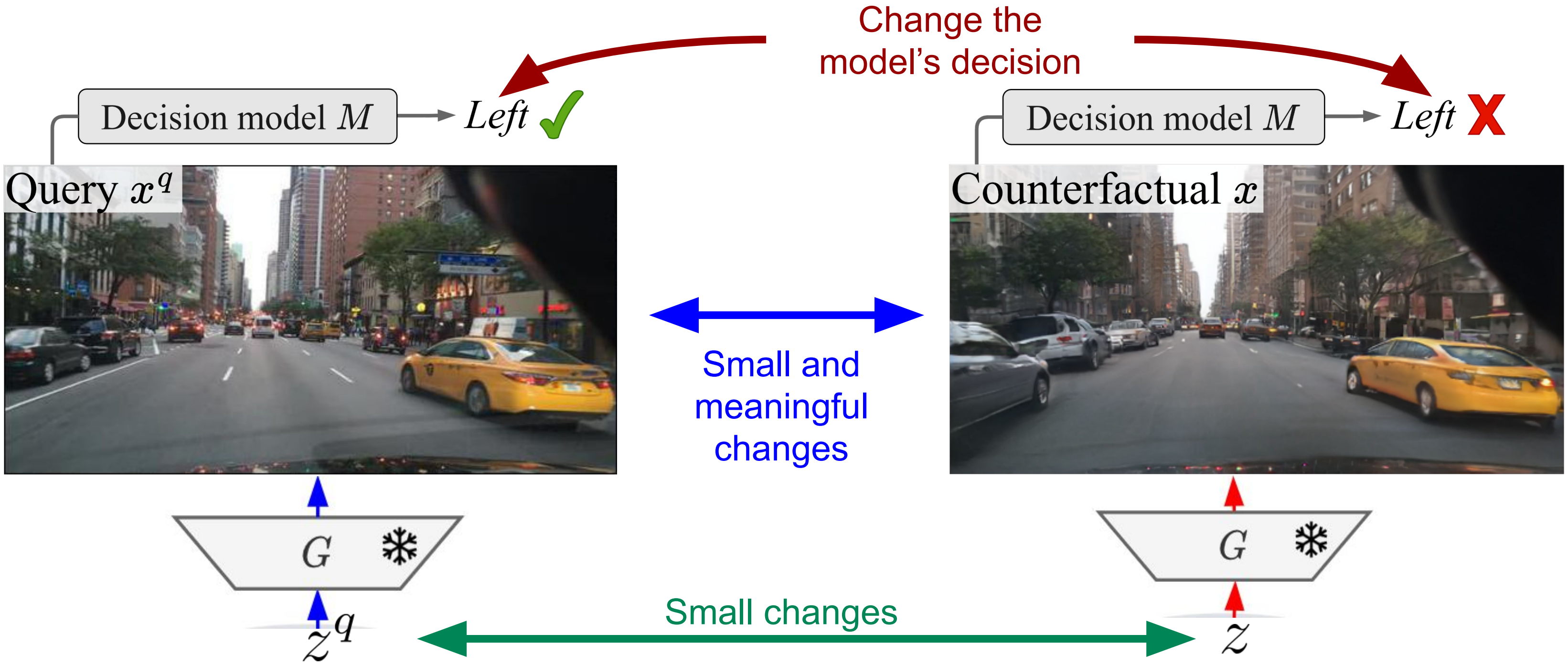
New object-aware framework that seeks counterfactual explanations for **compositional images and complex scenes**

Using an object-centric latent space empowers users to **explore the role of specific objects** in decision models

User-study showing usefulness of counterfactual explanations to better anticipate failures and find biases.

Explanations of **semantic segmentation models**

Counterfactual explanations for image classification models?



Change the model's decision

Decision model M → *Left* ✓

Query x^q

Decision model M → *Left* ✗

Counterfactual x

Small and meaningful changes

G ❄

z^q

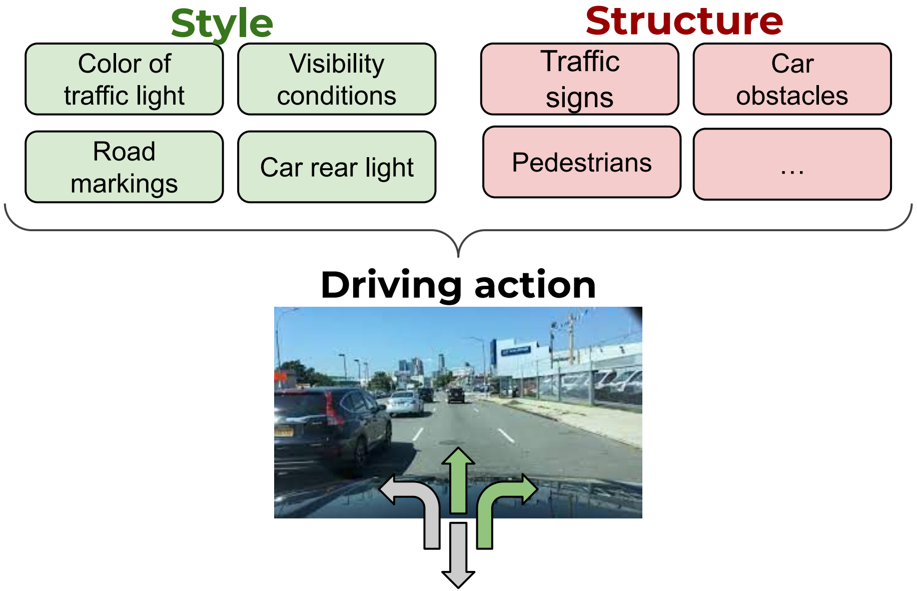
G ❄

z

Small changes

$$\text{argmin}_z \mathscr{L}_{\text{distance}}(z^q, z) + \lambda \mathscr{L}_{\text{decision}}(M(G(z)), y)$$

Dealing with complex compositional scenes



Style

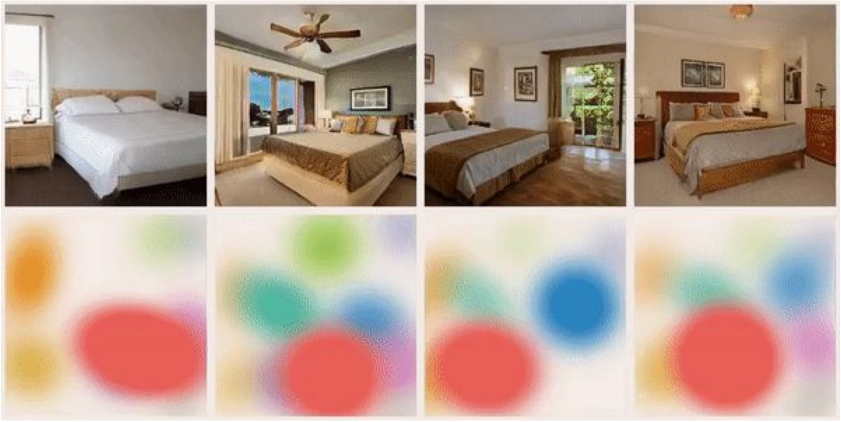
- Color of traffic light
- Visibility conditions
- Road markings
- Car rear light

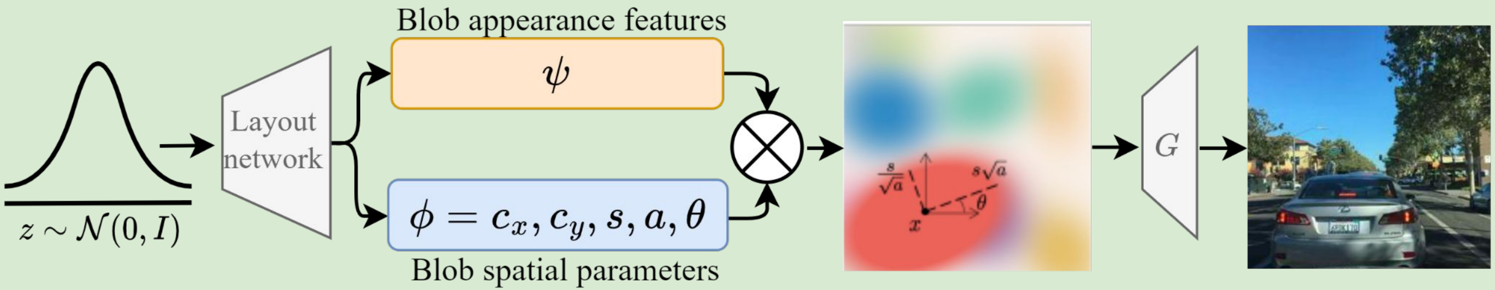
Structure

- Traffic signs
- Car obstacles
- Pedestrians
- …

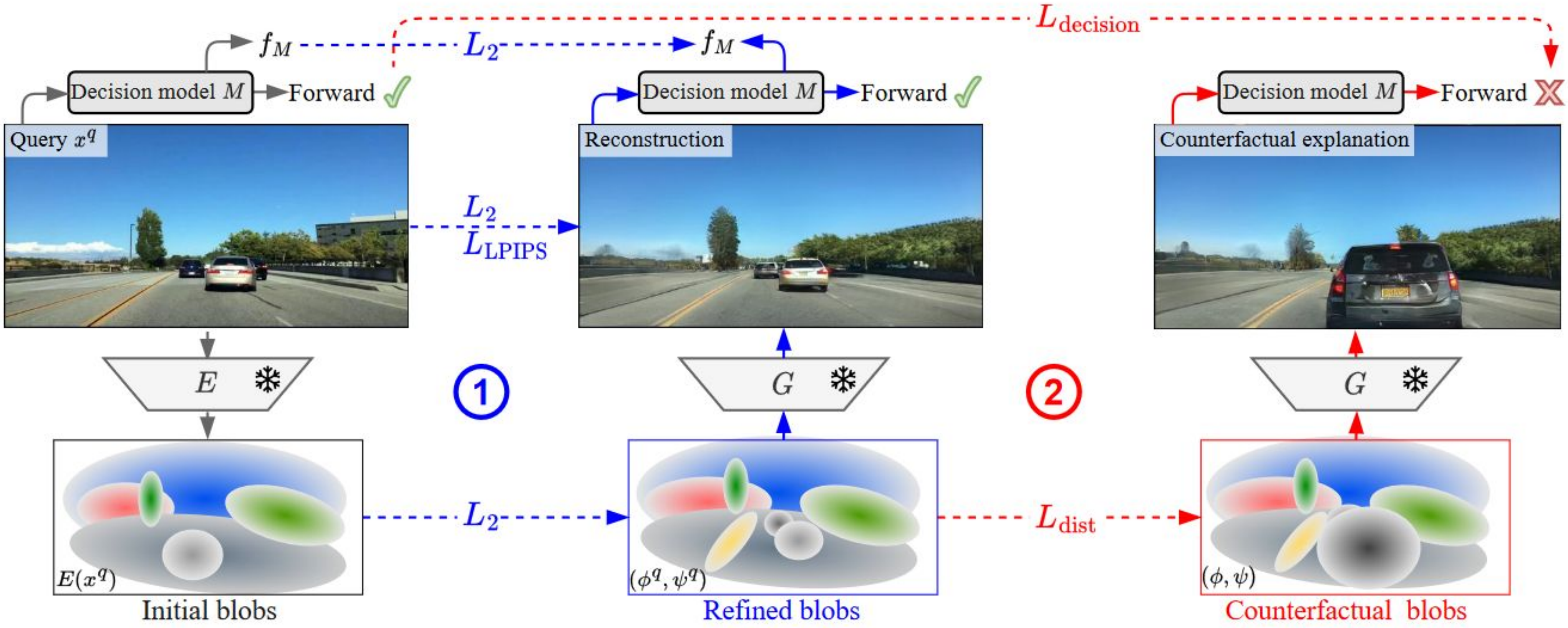
Driving action

BlobGan [Epstein *et al.* ECCV’22]





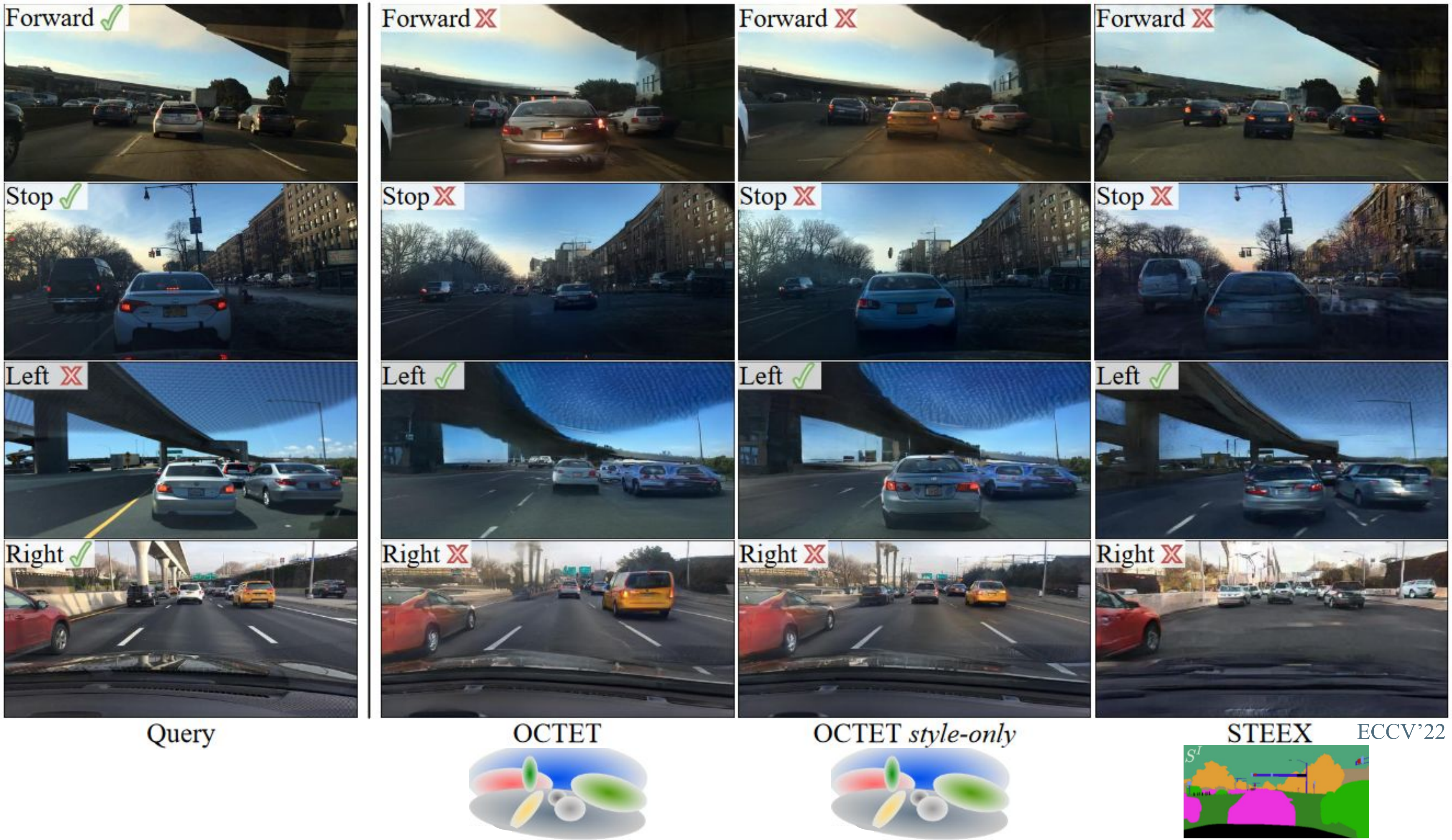
OCTET: overview



1 Optimizing blob parameters to get a good image inversion

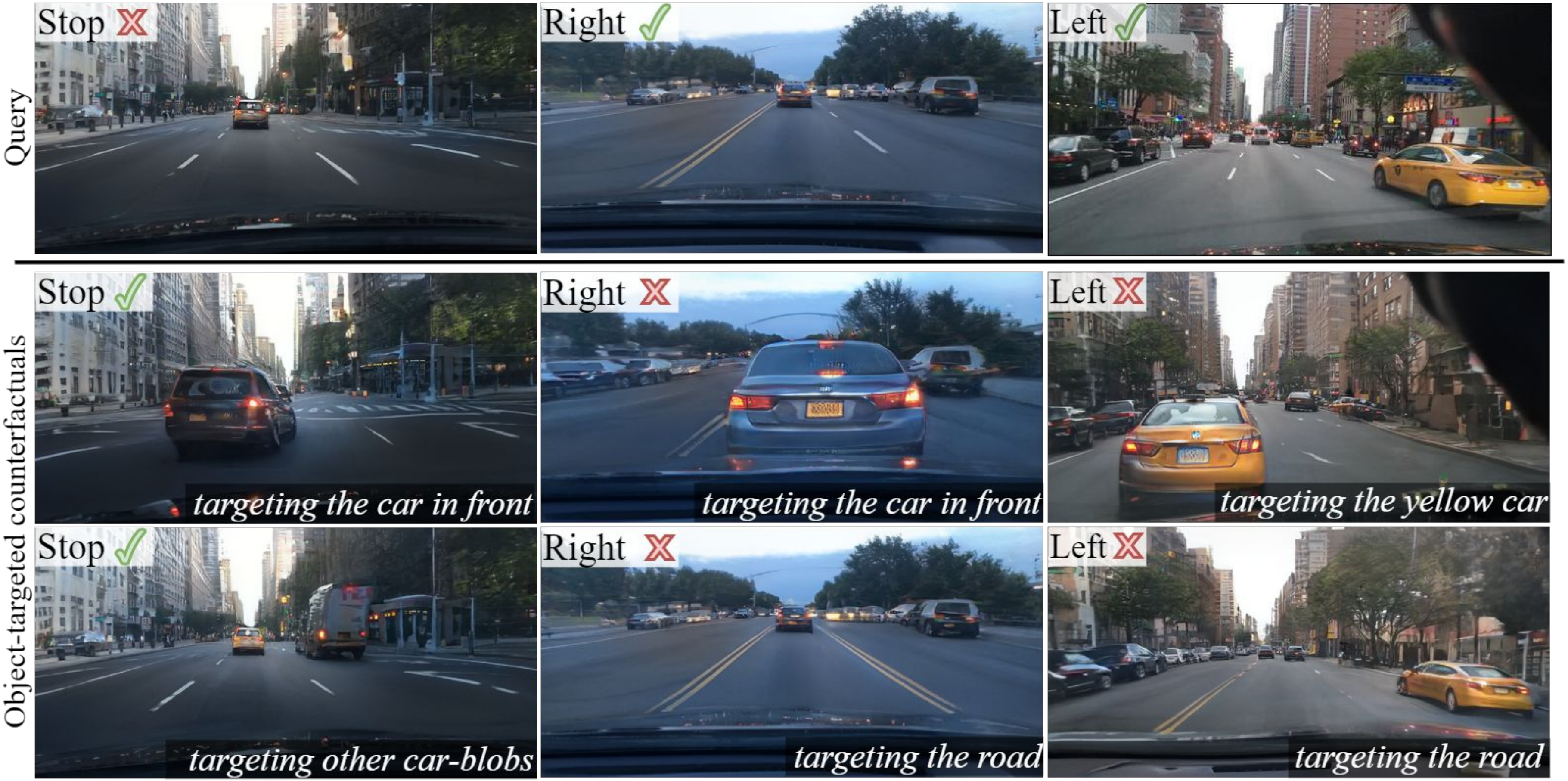
2 Optimize the blob parameters to change the model decision

Qualitative results



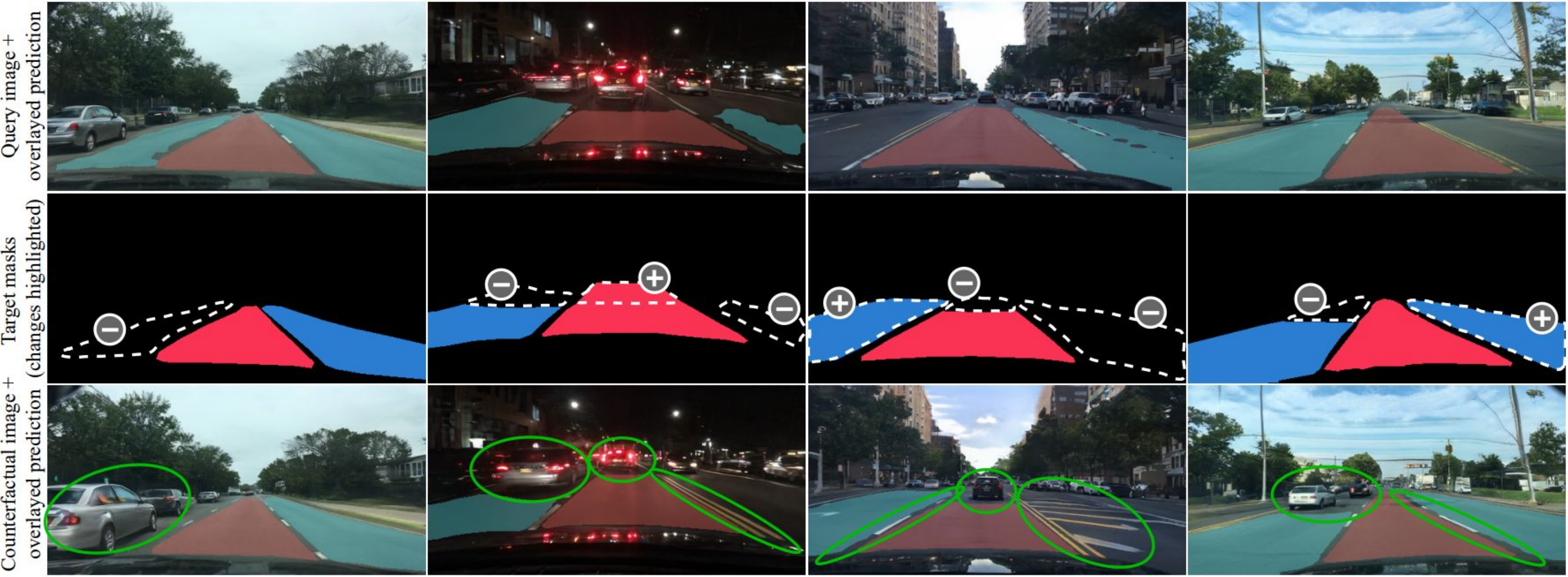
Targeted counterfactual explanations

Empowering users to explore the role of specific objects in decision models


Query
Stop ✗
Right ✓
Left ✓
Object-targeted counterfactuals
Stop ✓
targeting the car in front
Right ✗
targeting the car in front
Left ✗
targeting the yellow car
Stop ✓
targeting other car-blobs
Right ✗
targeting the road
Left ✗
targeting the road

Explaining a semantic segmentation model

Explaining a DeepLabv3 trained to segment primary and secondary free-spaces


Query image + overlayed prediction
Target masks (changes highlighted)
Counterfactual image + overlayed prediction

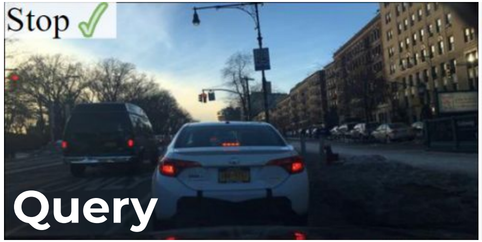
Quantitative metrics

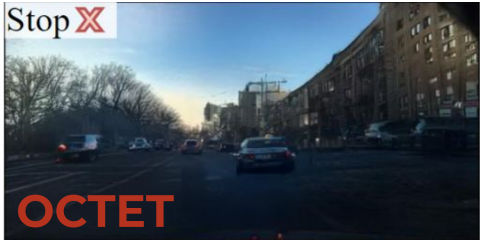
Are counterfactuals realistic?

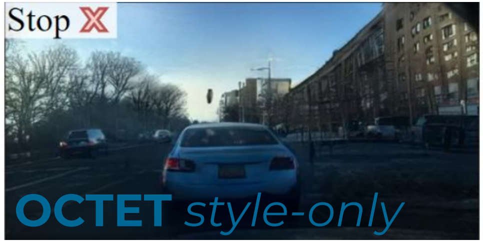
→ Fréchet Inception Distance (**FID** (↓))

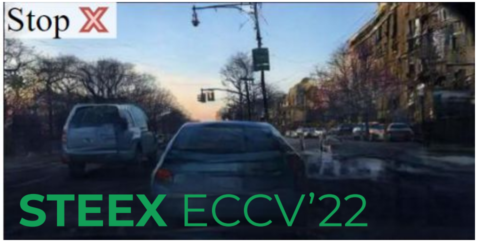
Are counterfactuals close to the query image?

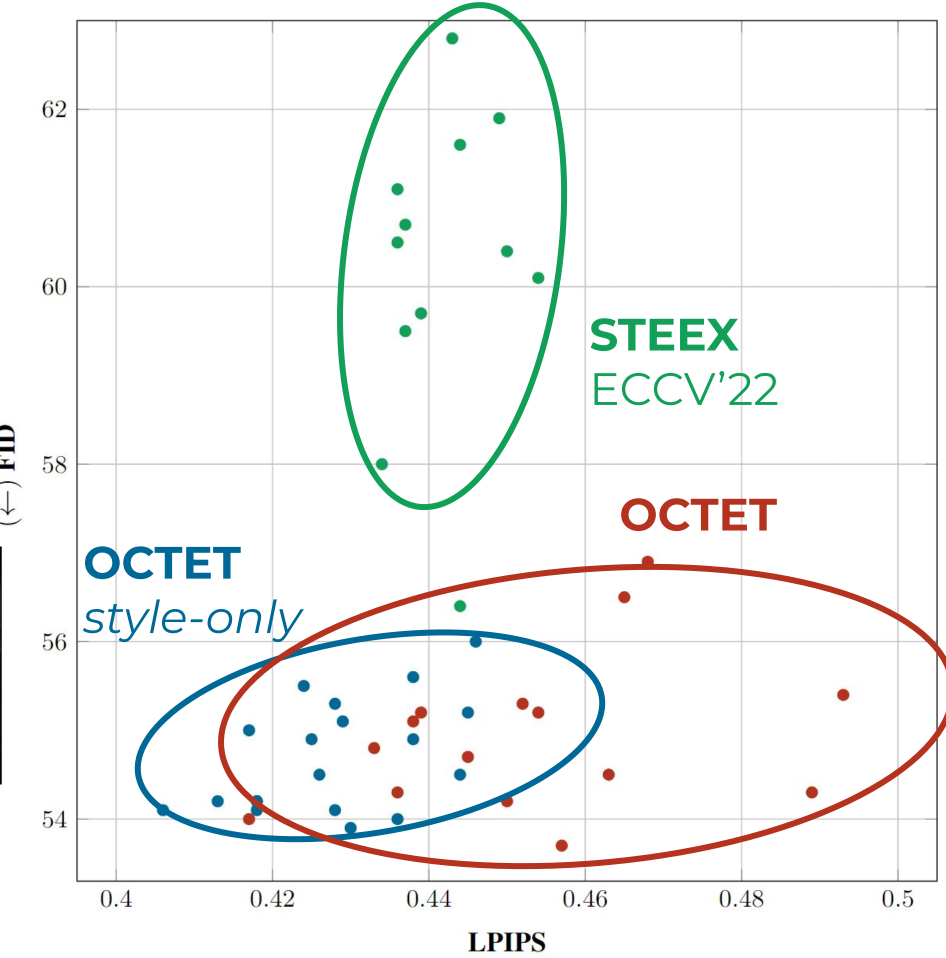
→ **LPIPS** distance







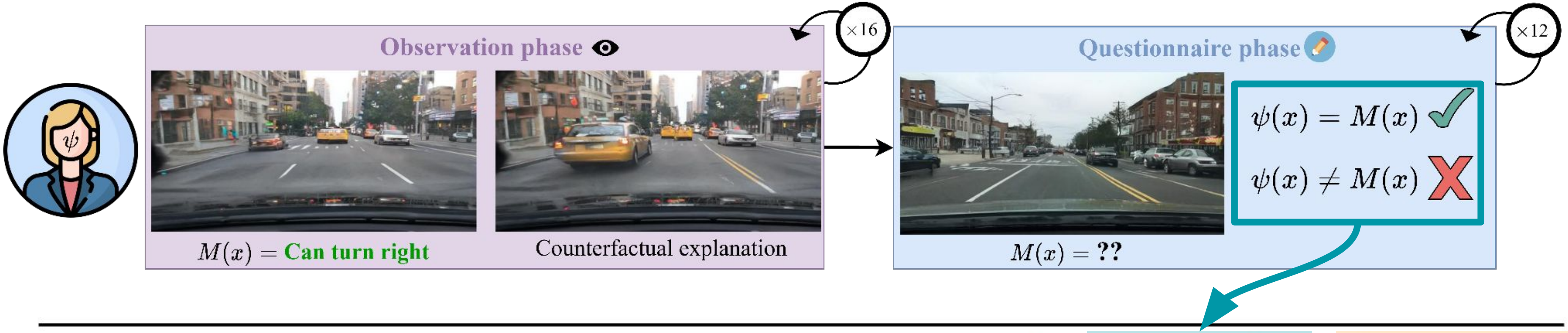


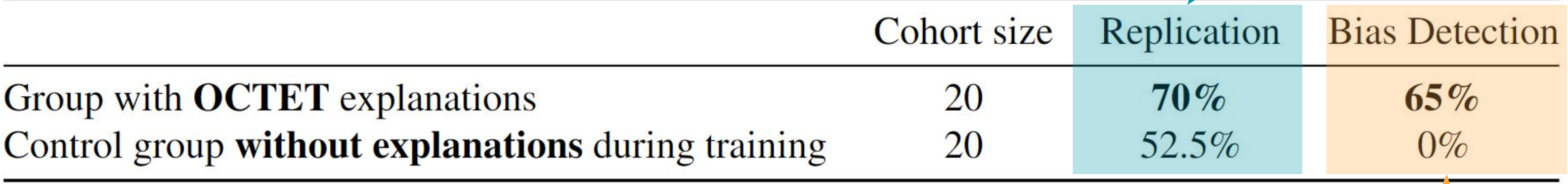


$$\operatorname{argmin}_x \mathcal{L}_{\text{distance}}(x^q, x) + \lambda \mathcal{L}_{\text{decision}}(M(x), y)$$

Can counterfactuals help to better “understand” a model?

“Understand” := Ability to predict model’s decision on new instances (simulatability)





	Cohort size	Replication	Bias Detection
Group with **OCTET** explanations	20	**70%**	**65%**
Control group **without explanations** during training	20	52.5%	0%

p-value = 0.0028

Unknown to the participants, the classifier is flawed: obstacles on both sides of the road influence the “Can turn right” prediction. **Did users find out?**

What I cannot predict, I do not understand: A human centered evaluation framework for explainability method. Fel *et al.*, NeurIPS 2022

Conclusion

OCTET is an object-centric framework able to explain classifiers operating on complex and compositional scenes

github.com/valeoai/octet