UNIVERSITY OF ILLINOIS CHICAGO

NEURAL INFORMATION PROCESSING SYSTEMS

Implicit Differentiable Outlier Detection Enables Robust Deep Multimodal Analysis

Zhu Wang
zwang260@uic.edu
@ellenz_wang

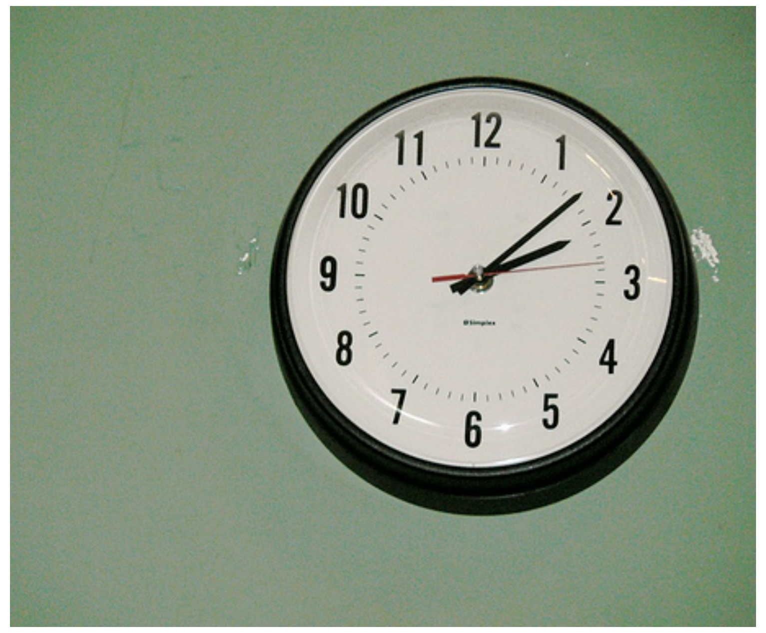
Sourav Medya
medya@uic.edu
@sourav_medya

Sathya N. Ravi
sathya@uic.edu
@tweetingsathya

Popular tasks for Multimodal pipelines

Input

Captions:
A clock mounted
on a pale wall.
Questions:
What energy source
does this object run on?



Output

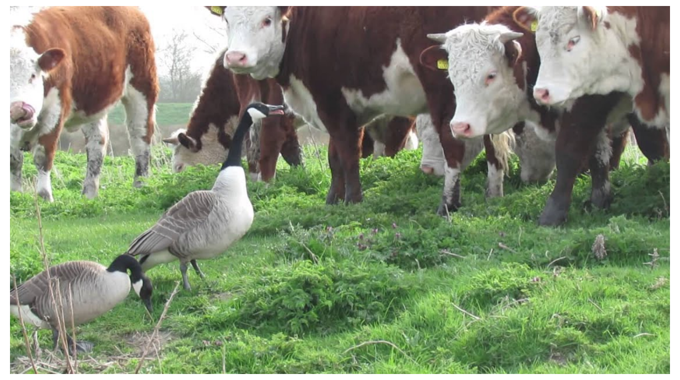
Answers:

Common mis-prediction : radio

Ours: electricity

Input

Sentence:
One image features
multiple ducks on a
country road, and the
other image shows a
mass of white ducks that
are not in flight.



ConceptNet
An open, multilingual knowledge graph

Output

Answers:

Common mis-prediction : True

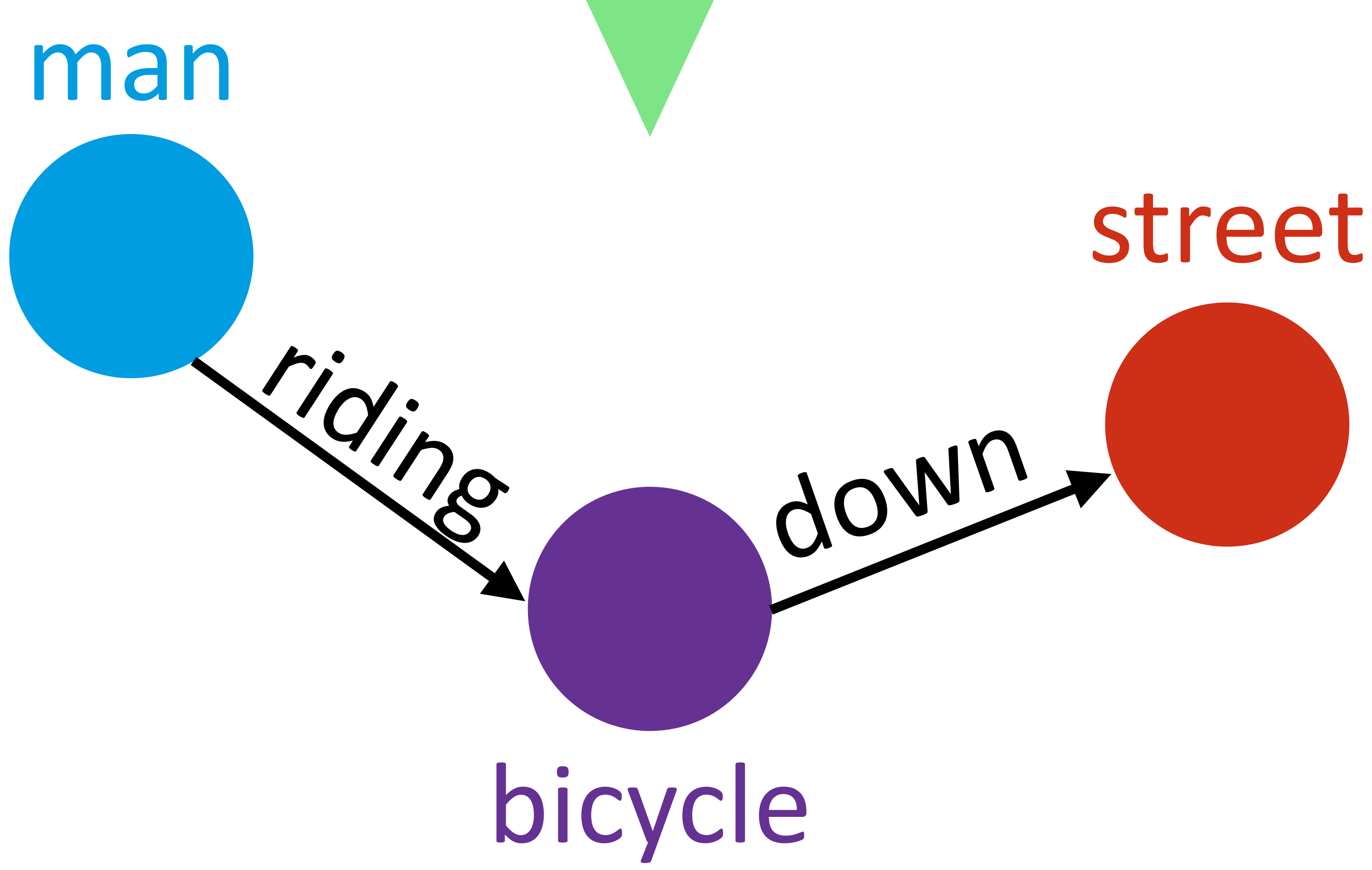
Ours: False

paper:
Can we use external knowledge in MM pipelines?

Augmenting External knowledge in MM pipeline via KG

***Caption*:**
A man riding a bicycle down a city street.

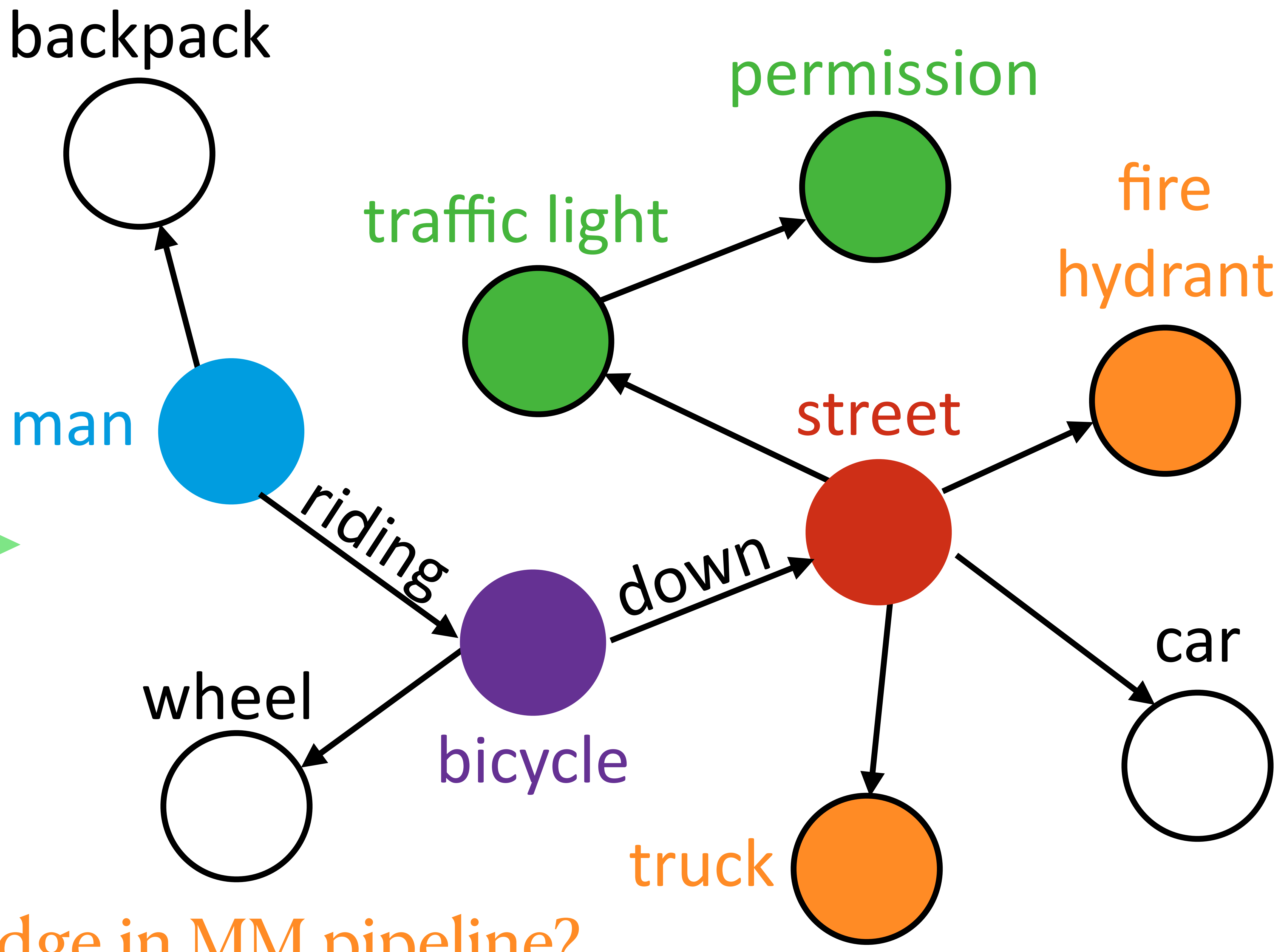
Parse to triplet



ConceptNet
An open, multilingual knowledge graph

Query from KG

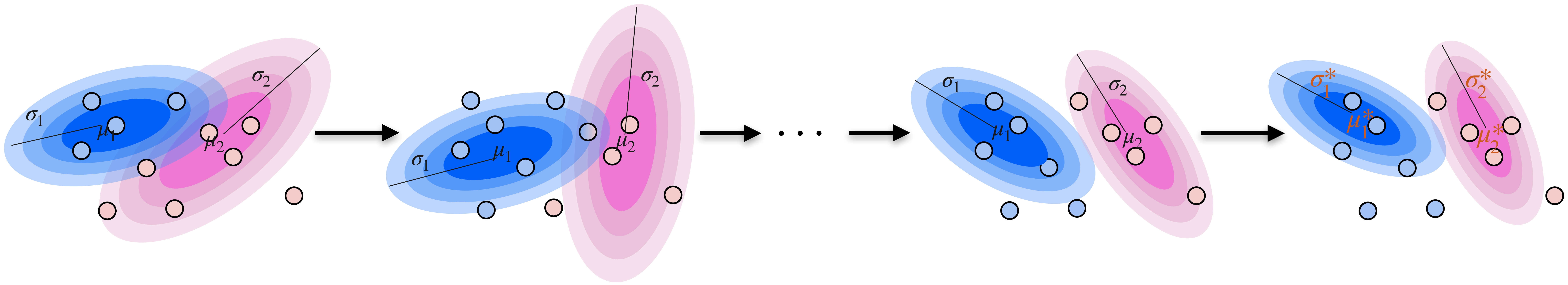
***Question*:**
Is this person crossing illegally or legally?



Another challenge: How to align external knowledge in MM pipeline?

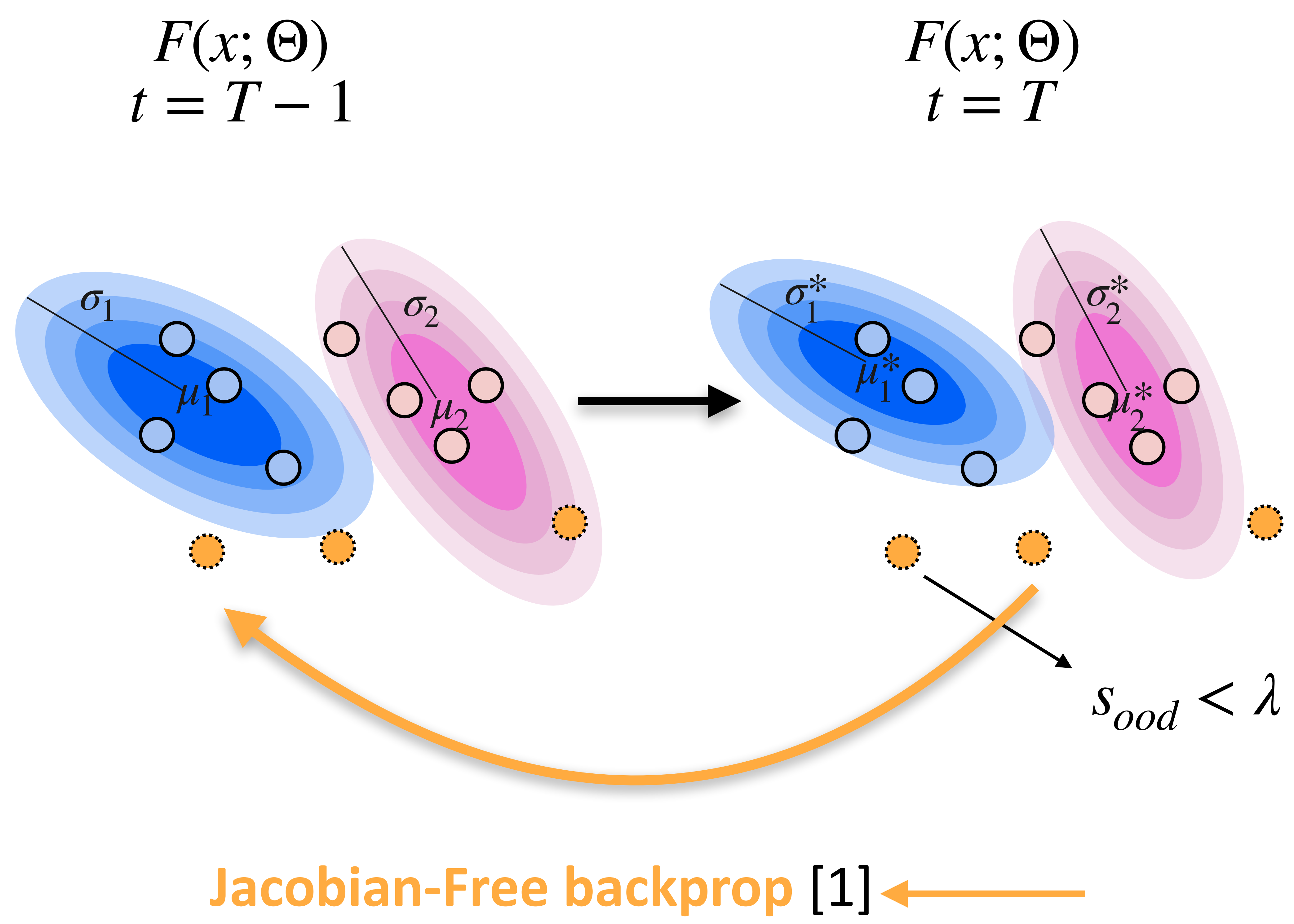
Implicit OOD detection layer: Out-of-Distribution detection using EM iterations

$F(x;\Theta)$	$F(x;\Theta)$	$\dots$	$F(x;\Theta)$	$F(x;\Theta)$
$t = 1$	$t = 2$		$t = T - 1$	$t = T$



We approximated the density of multimodal features for outlier detection.

Implicit Differentiable OOD detection layer



EM-based algorithm as a fix point iteration:

$$\mu_k^{t+1} \leftarrow \frac{\sum_{i=1}^{N} \exp(-w(\mu_k^t))x_i}{\sum_{i=1}^{N} \exp(-w(\mu_k^t))} \tag{1}$$

GEM score for Outlier Detection:

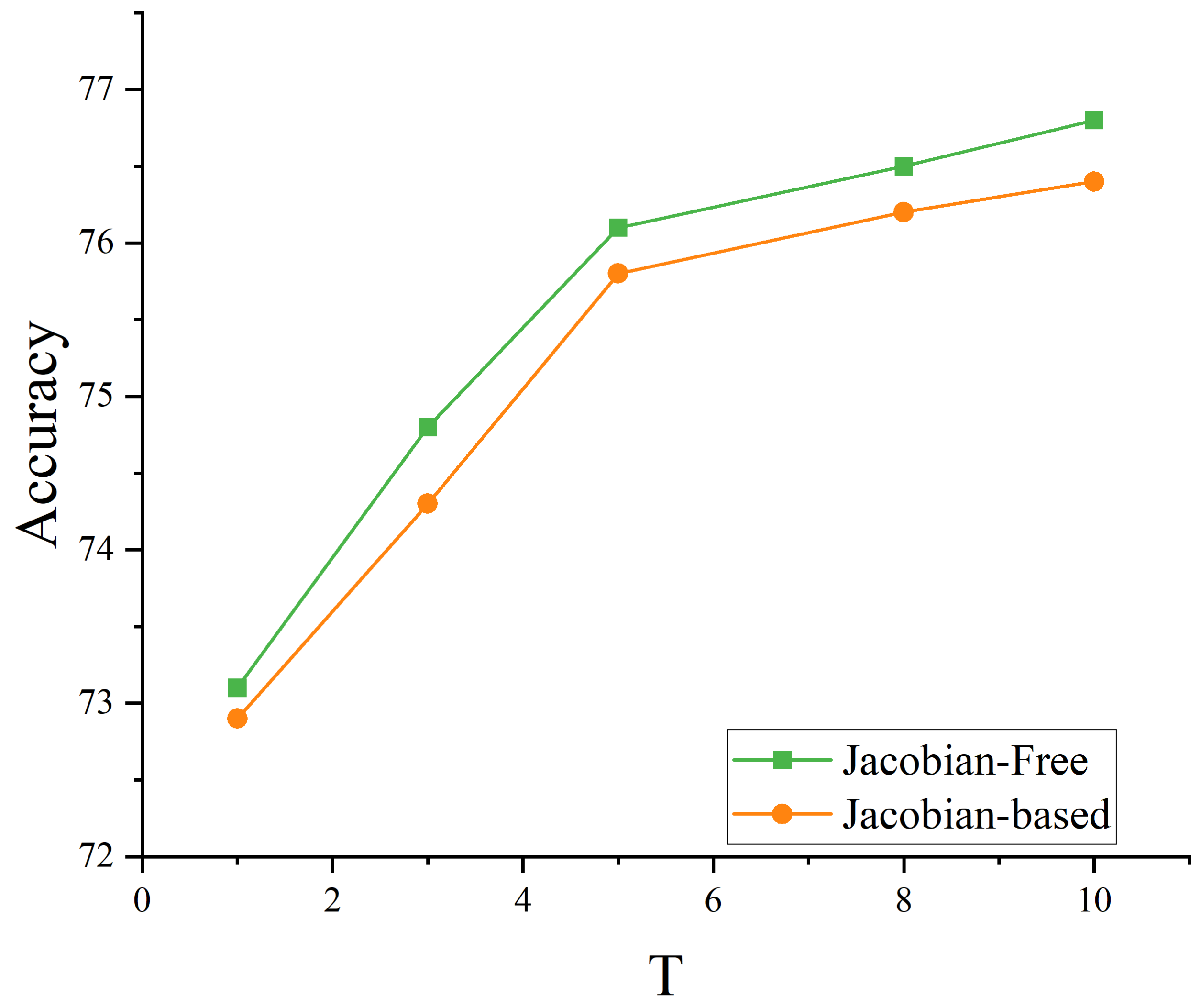
$$s(l_j) = \log \sum_{k=1}^{K} \exp(-\frac{1}{2}(l_j - \mu_k^*)^T \sigma_k^{-1}(l_j - \mu_k^*)) \tag{2}$$

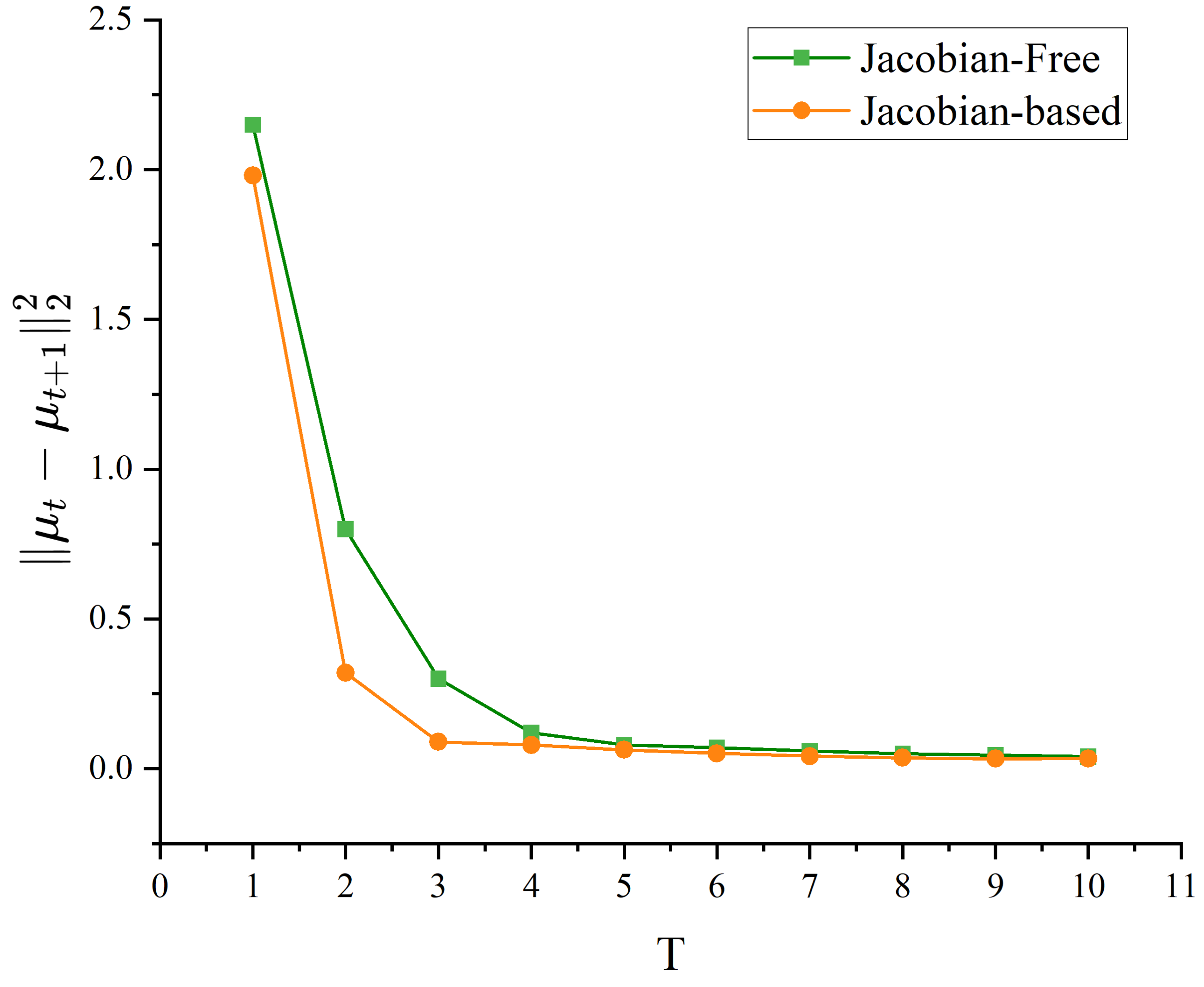
[1] Fung, Samy Wu, et al. "Jfb: Jacobian-free backpropagation for implicit networks." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. No. 6. 2022.

Efficient Backpropagation for OOD Detection Layer

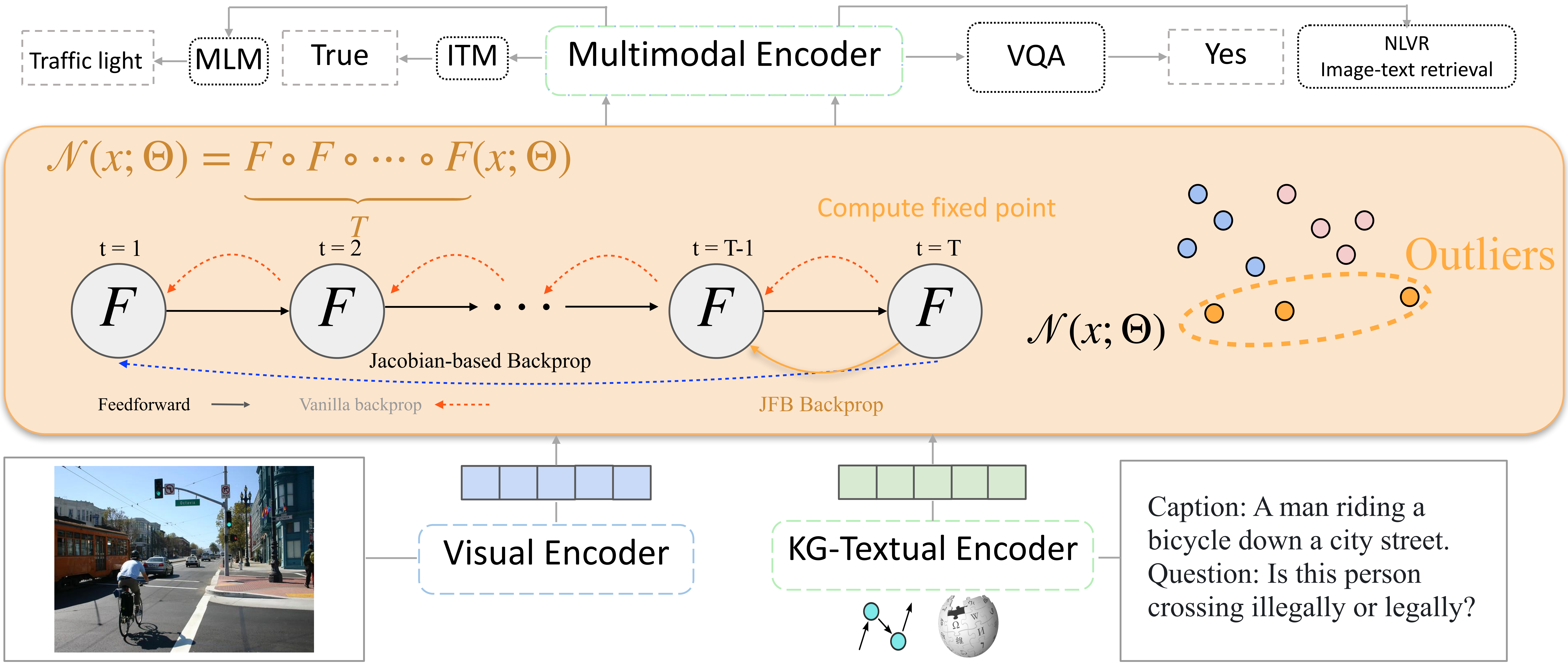
$$\boxed{\text{Finding fix points}} \underset{\text{JFB}}{\overset{\mu_k^*, \sigma_k^*}{\rightleftarrows}} \boxed{\text{OOD network operator } \mathcal{N}}$$

- We use Jacobian-Free Backpropagation (JFB) to unroll the *last few iterates* of the EM algorithm for gradient.





Our final VK-OOD Multimodal pipeline



Experiments — Backpropgation methods

- We used different backpropgation methods in OOD detection layer with ViLT as the backbone.
- **JFB-EM** *gained* 76.8% in term of accuracy on VQAv2 task with 1/4 backward time comparing to vanilla EM.

Method	#Param(M)	#FLOPs(G)	Time(m)/epoch Forward	Time(m)/epoch Backward	Max Mem(Mb)	VQAv2
Vanilla-EM	152.6	185.2	39.6	26.8	18673	76.6
JB-EM	125.2	115.7	39.6	12.7	14512	76.6
JFB-EM	124.8	108.6	39.6	6.3	13674	76.8

Table 1: Experimental results of different backpropagation method in the dense OOD detection layer. JFB-EM is much more efficient in backward pass and use less memory. It also outperforms on the VQAv2 task in terms of accuracy.

Experiments — Downstream tasks

Model	#Params	VQAv2	NLVR2	COCO TR R@5	COCO IR R@5	Flickr30k TR R@5	Flickr30k IR R@5
ViLT	87	70.3	74.6	86.2	72	95.6	86.8
UNITER	155	72.7	75.8	87.4	78.5	97.1	92.4
ALBEF	314	74.5	80.5	91.4	81.5	99.4	96.7
VinVL	157	75.9	83.1	92.6	83.2	-	-
BLIP*	346	77.5	82.8	95.2	**85.4**	**99.8**	**97.5**
VK-OOD-s(ViLT)	87.4	76.7	84.3	90.9	81.6	97	94.3
VK-OOD-s(CLIP)	113.4	76.2	83.8	92.8	83.4	99.6	96.7
VK-OOD-s(BLIP)	346.4	77.8	84.1	**95.4**	85.2	**99.8**	97.2
VK-OOD-l(ViLT)	125	76.8	**84.6**	91.7	81.3	97.2	94.5
VK-OOD-l(CLIP)	151	76.1	83.9	93.1	83.6	99.6	96.8
VK-OOD-l(BLIP)	412	**77.9**	84.5	95.1	84.8	99.6	97.1

Table 5: Overall performance on multiple downstream tasks. We demonstrate VK-OOD scale with different model backbones and achieve the best and second-best results. VK-OOD-s is the scalar case, and VK-OOD-l is the dense case. *our implementation.

VK-OOD achieved the best and second-best results in all downstream tasks comparing with baselines.

Experiments — Qualitative Analysis

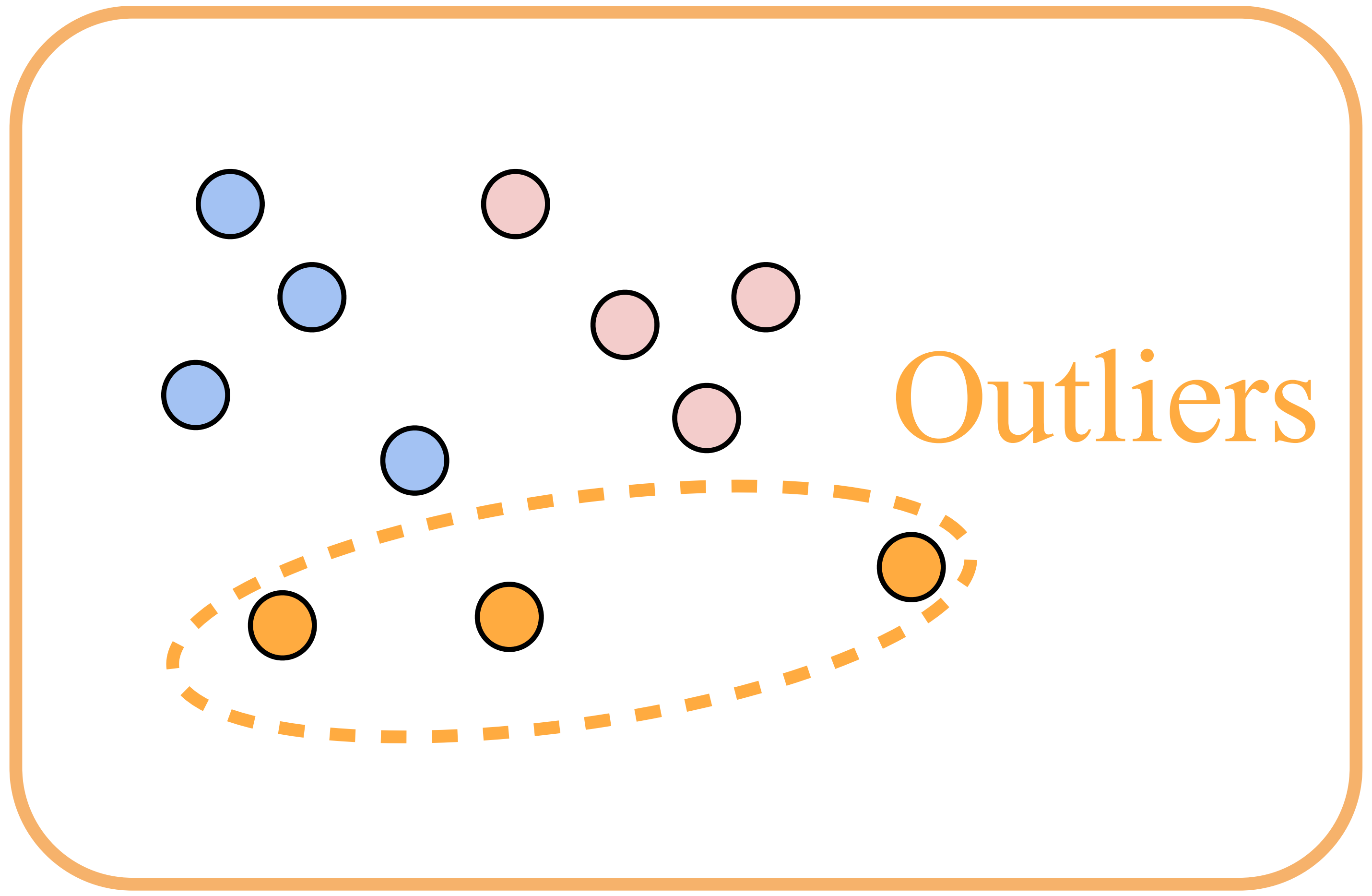
k=1 k=2 k=3 k=4 k=5

Original A man riding bicycle Car locates on street Bus locates on street Traffic lights locate on street Buildings locate on street

Contributions

Caption: A man riding a bicycle down a city street.
Question: Is this person crossing illegally or legally?

ConceptNet
An open, multilingual knowledge graph



Contributions

- We mainly aim to integrating implicit and explicit knowledge seamlessly in vision-language model.
- It is crucial to identify high-quality knowledge during forward pass due to error propagation that may affect downstream predictions.
- We propose an end-to-end framework with the implicit differentiable outlier detection layer to filter noise knowledge during training.

Thanks for your attention!

Q&A

Github

Paper