Fighting Fire with Fire: Avoiding DNN Shortcuts through Priming

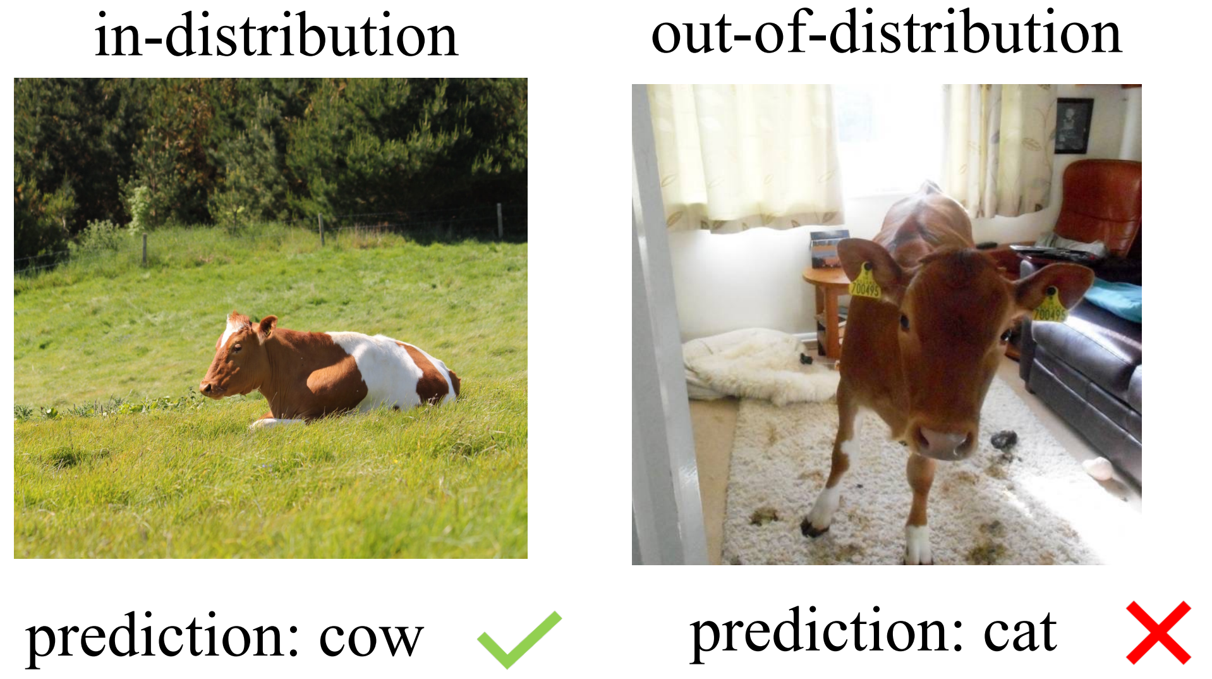
Tsinghua University · University of Pennsylvania · The University of Texas at Austin

Chuan Wen, Jianing Qian, Jierui Lin, Jiaye Teng, Dinesh Jayaraman, Yang Gao

Problem: Shortcuts in DNNs

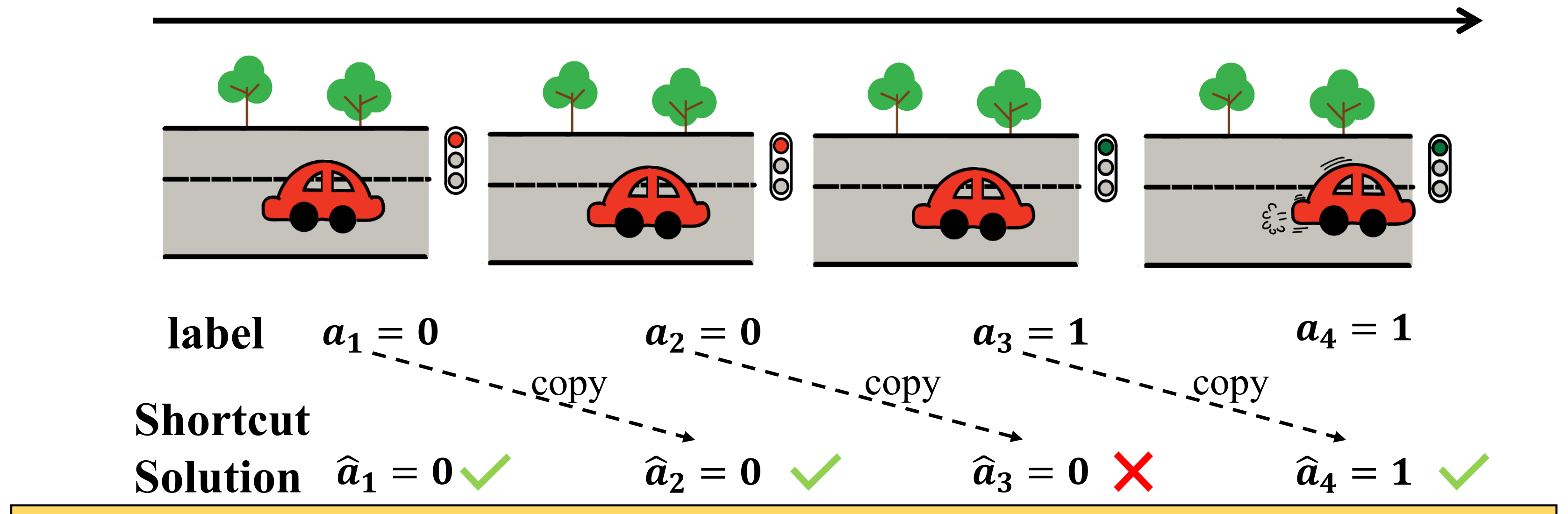
DNNs often struggle to disambiguate between competing hypotheses for a target concept, and end up learning “shortcuts”.

Example 1: Image Classification



DNNs cheat by relying on backgrounds.

Example 2: Behavioral cloning



DNNs cheat by copying from the previous action during training.

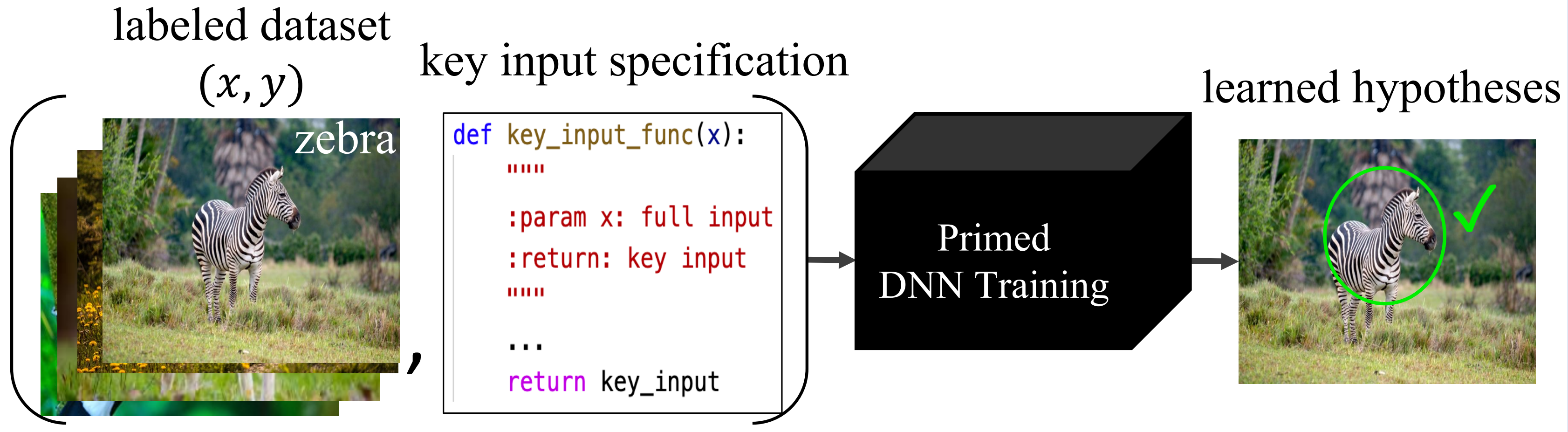
Motivation:

The training data does not fully specify the task!

Humans rely on additional knowledge e.g. task-relevant inputs

Our Idea:

Use auxiliary knowledge to “prime” DNNs away from shortcuts.



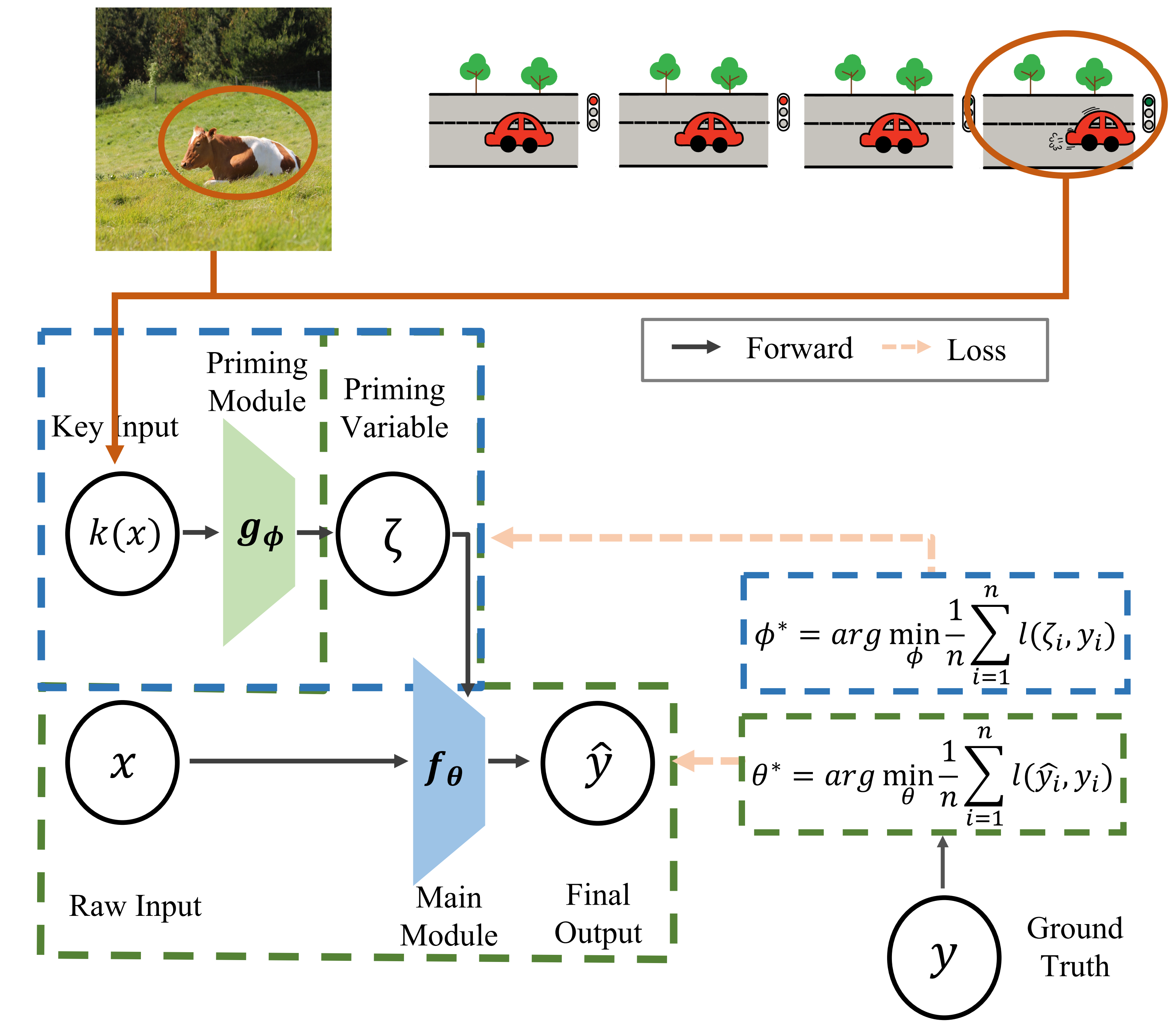
```
def key_input_func(x):
    """
    :param x: full input
    :return: key input
    """

    ...
    return key_input
```

PrimeNet: Using Key Input Knowledge

Providing key input appropriately creates a new desirable shortcut that “primes” the main module towards correct solutions.

Approximate key inputs are easy to provide

Classification: saliency **Behavioral Cloning:** last frame

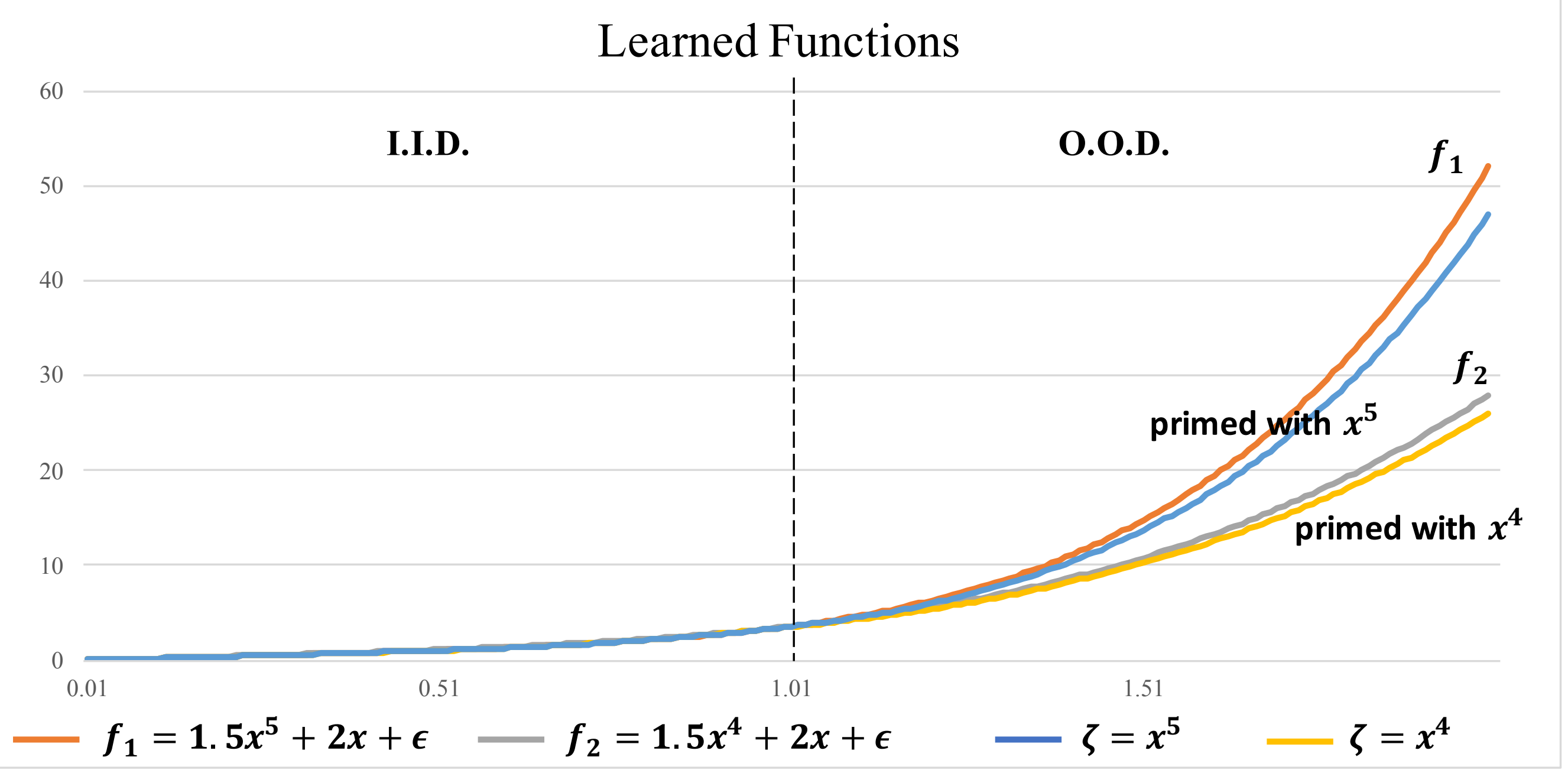


$$\phi^* = arg\min_{\phi} \frac{1}{n}\sum_{i=1}^{n} l(\zeta_i, y_i)$$

$$\theta^* = arg\min_{\theta} \frac{1}{n}\sum_{i=1}^{n} l(\hat{y}_i, y_i)$$

Experiment Results & Analysis

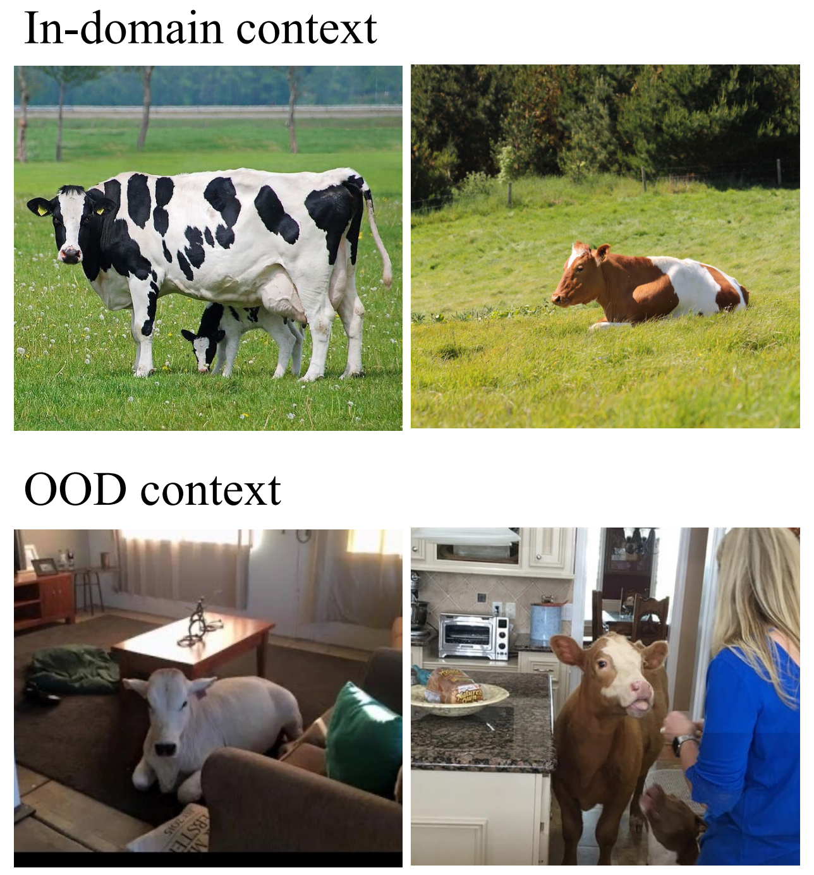
Toy 1-D regression experiment:



Conclusion: priming variable ζ can guide DNN training towards the solution desired by key input.

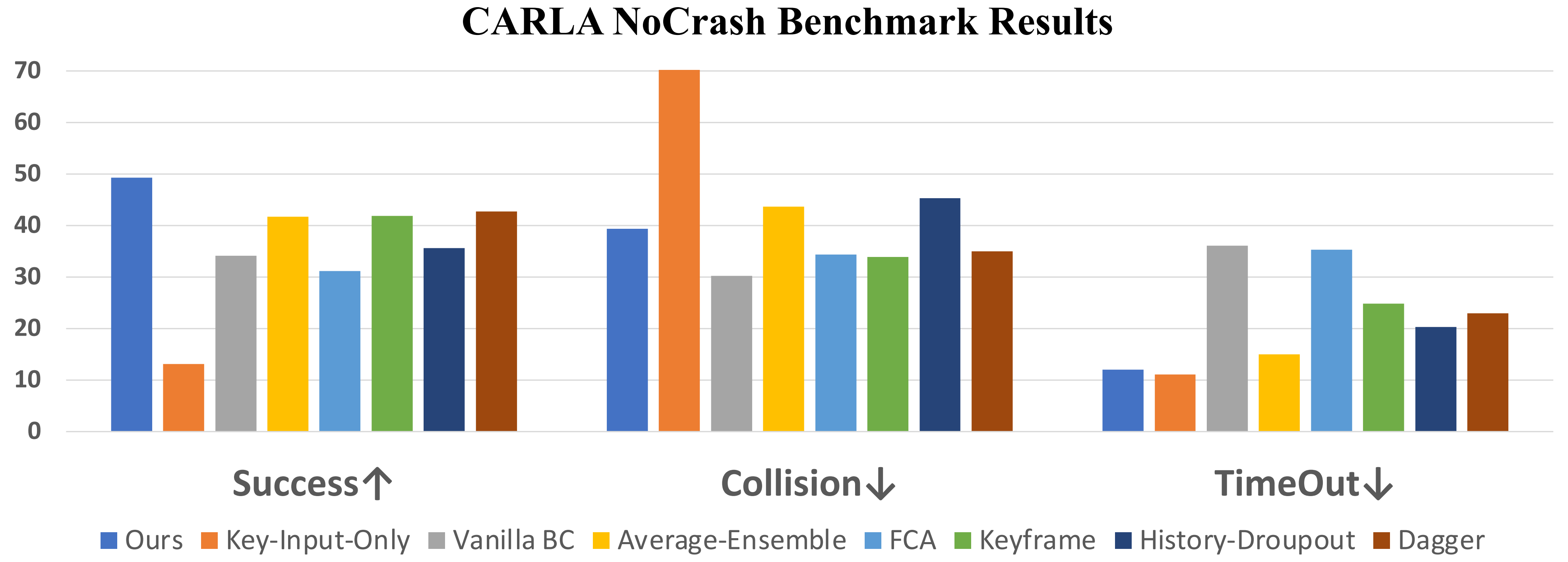
Image Classification on NICO:

Method	In-domain test	OOD test
vanilla ResNet18	66.11	42.61
key-input-only	62.78	47.54
average-ensemble	63.33	47.69
RuBi (Cadene et al., 2019)	-	44.37
ReBias (Bahng et al., 2020)	-	45.23
CutOut (DeVries & Taylor, 2017)	-	43.77
MixUp (Zhang et al., 2017)	62.78	41.46
IRM (Arjovsky et al., 2019)	-	41.46
StableNet (Zhang et al., 2021b)	63.33	43.62
CaaM (Wang et al., 2021b)	70.00	46.62
PrimeNet (ours)	71.11	**49.00**



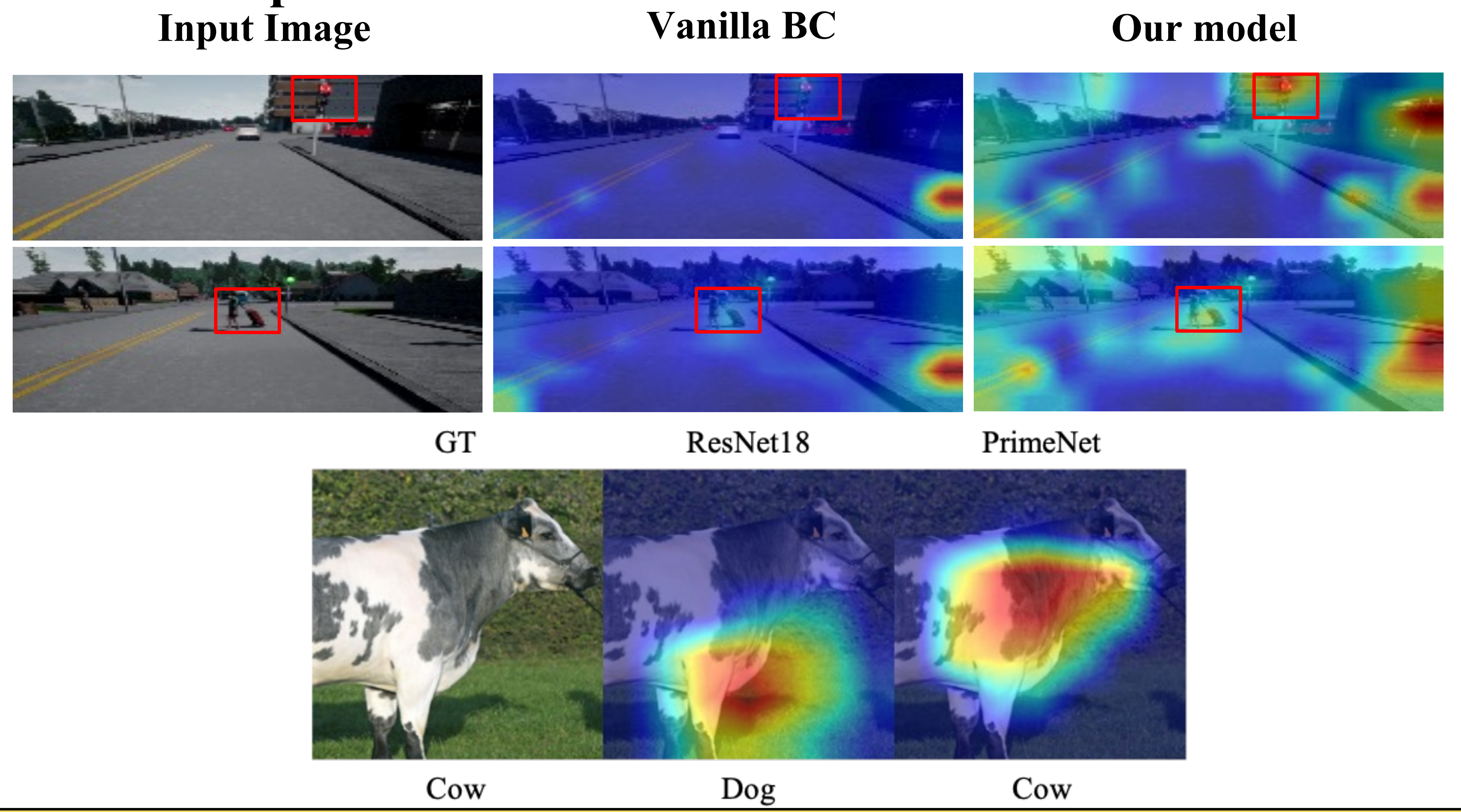
PrimeNet generalizes best, without harming in-domain performance.

Imitation Learning on CARLA Driving:



PrimeNet performs best, correctly responding to new environment cues such as traffic lights, neighboring cars, pedestrians.

Activation Maps:



Our model attends to the appropriate visual cues in the scene.