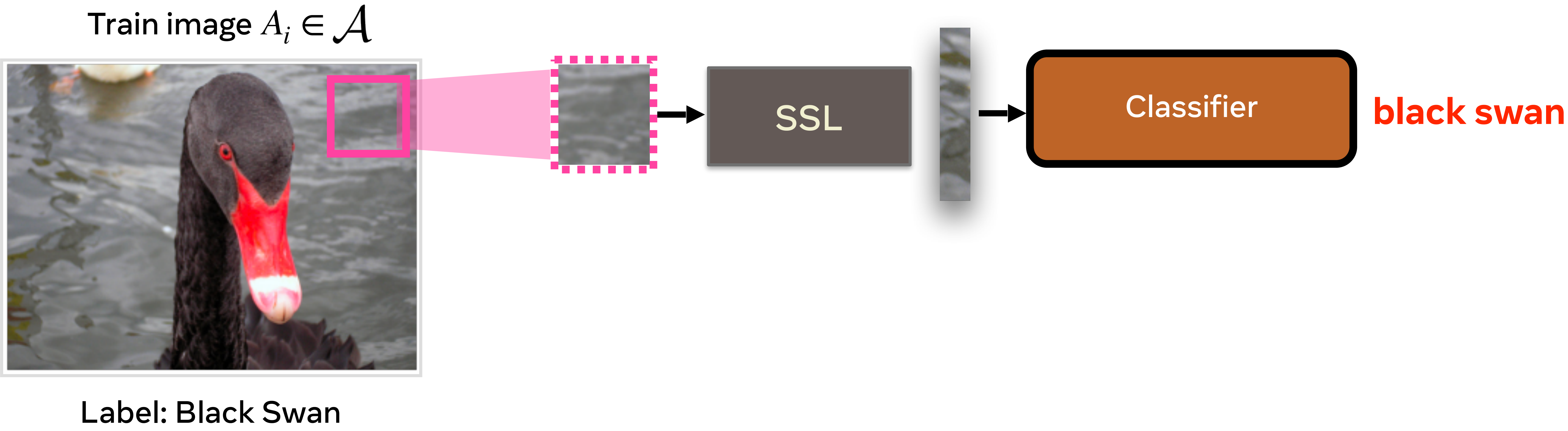
Measuring Déjà vu Memorization Efficiently

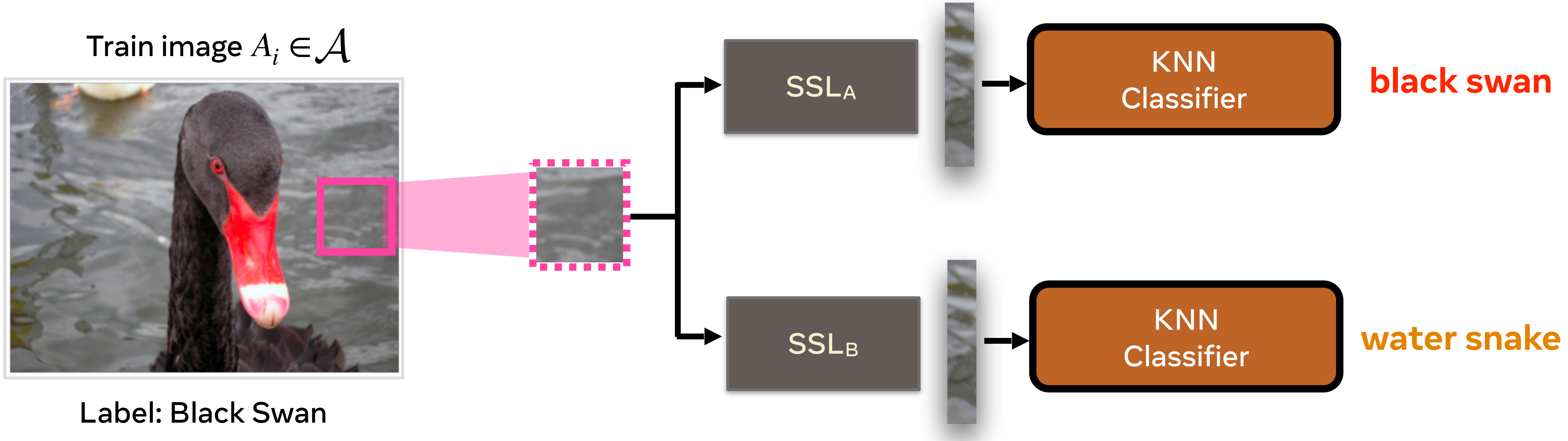
Narine Kokhlikyan, Bargav Jayaraman, Florian Bordes , Chuan Guo, Kamalika Chaudhuri

Meta AI

Unintended Memorization in image representation models

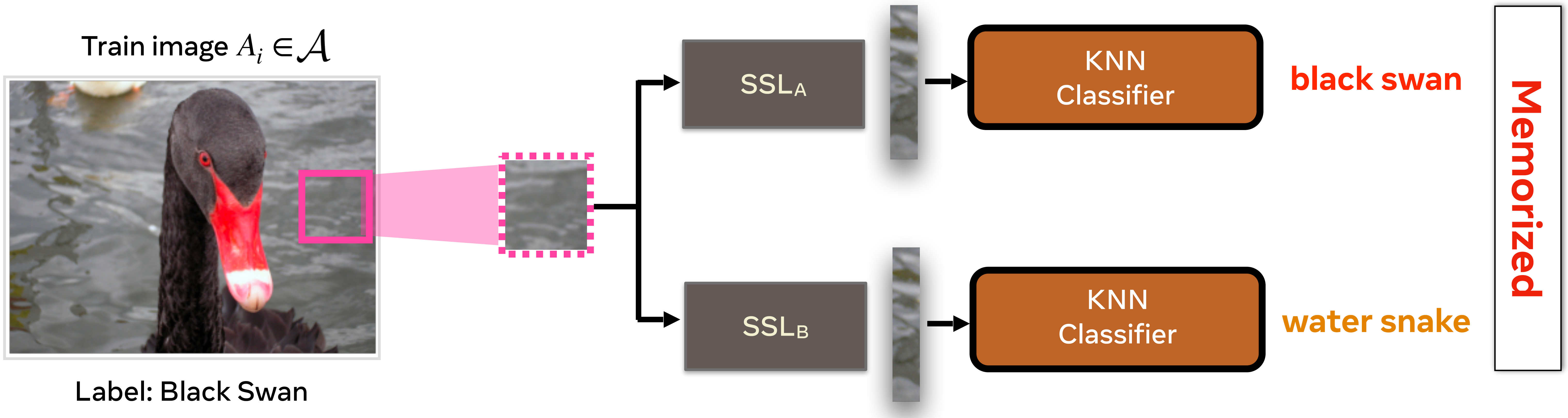


Unintended Memorization in image representation models

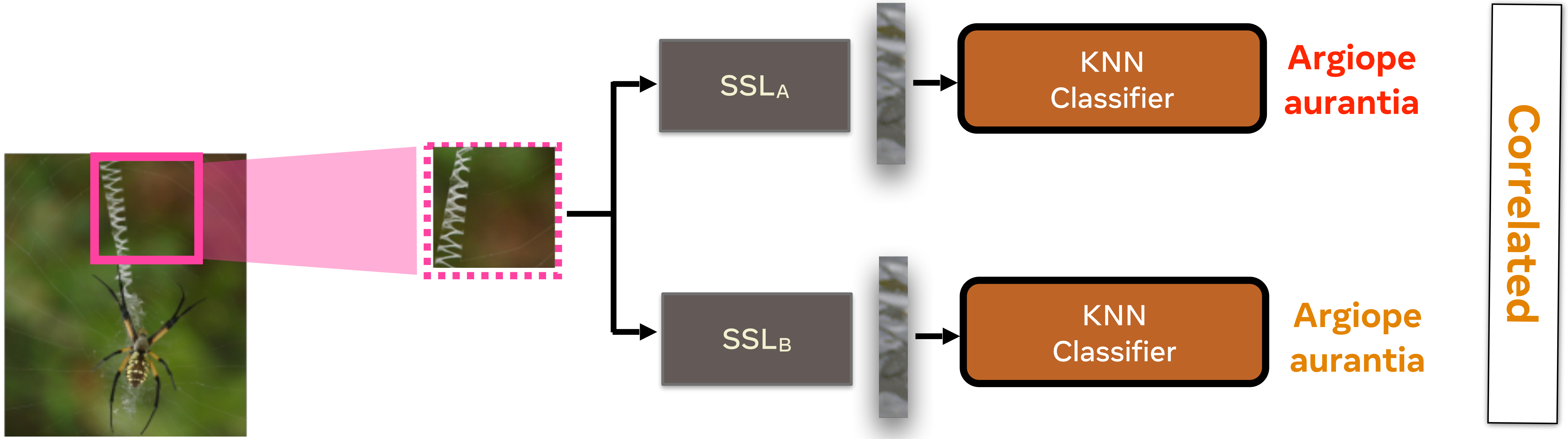


Meta AI

Detecting unintended memorization with two-model test

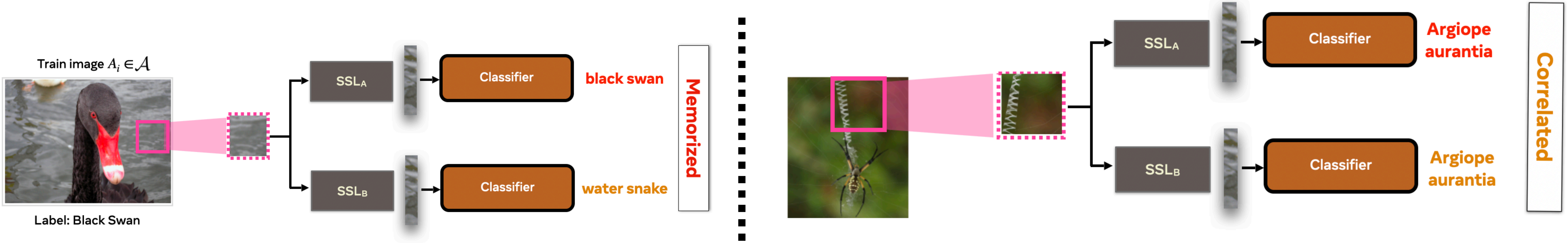


Detecting unintended memorization with two-model test



Meta AI

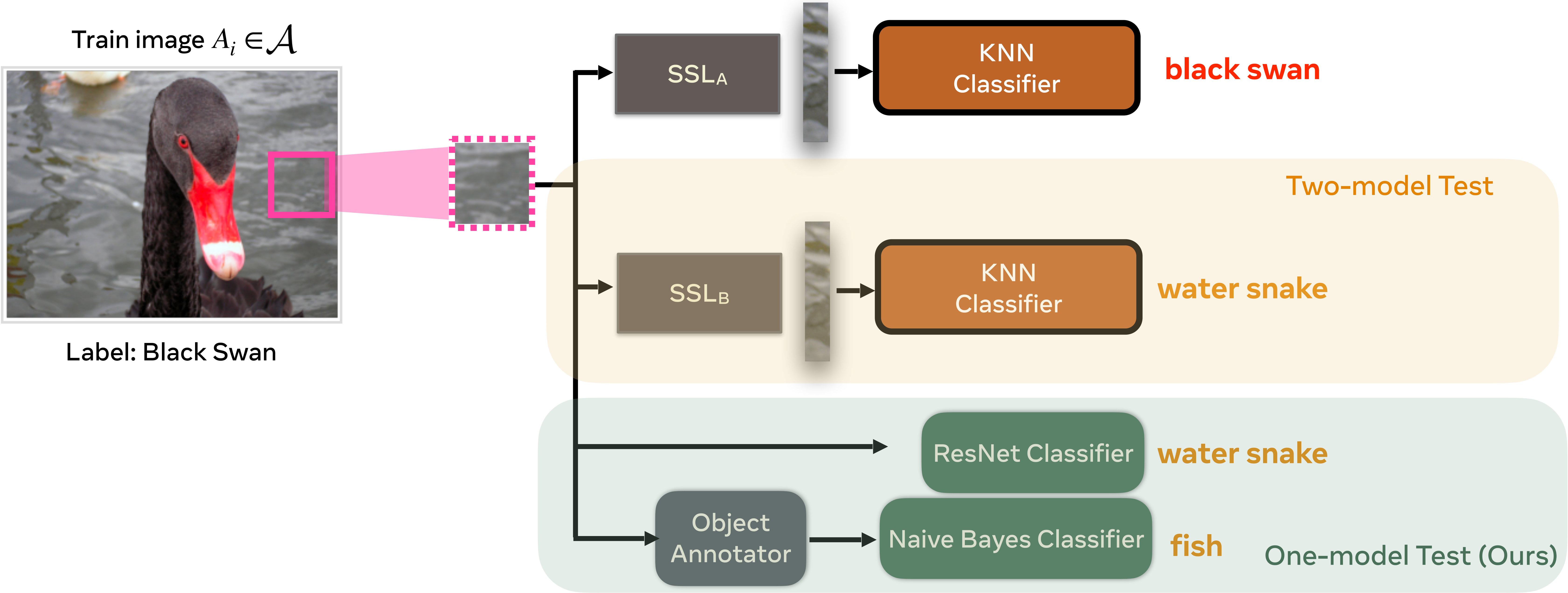
Challenges with two-model test



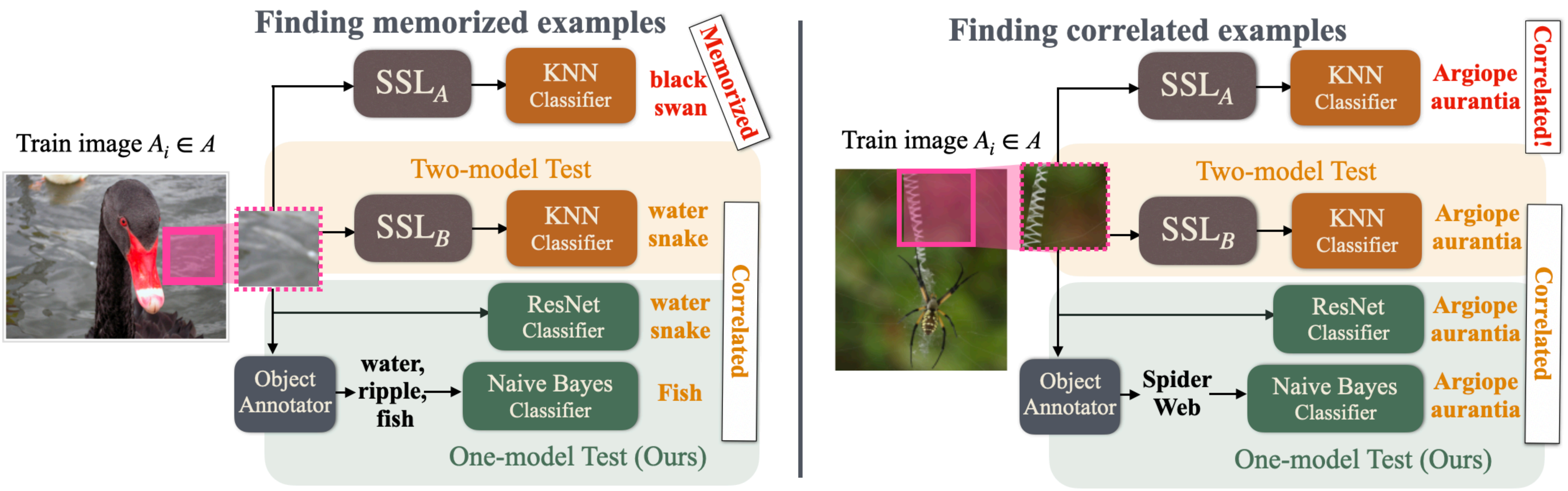
Two model test requires:

- to train two SSL models on disjoint splits of the training dataset
- is not applicable to OSS models trained on the entire dataset

Detecting unintended memorization with one-model test



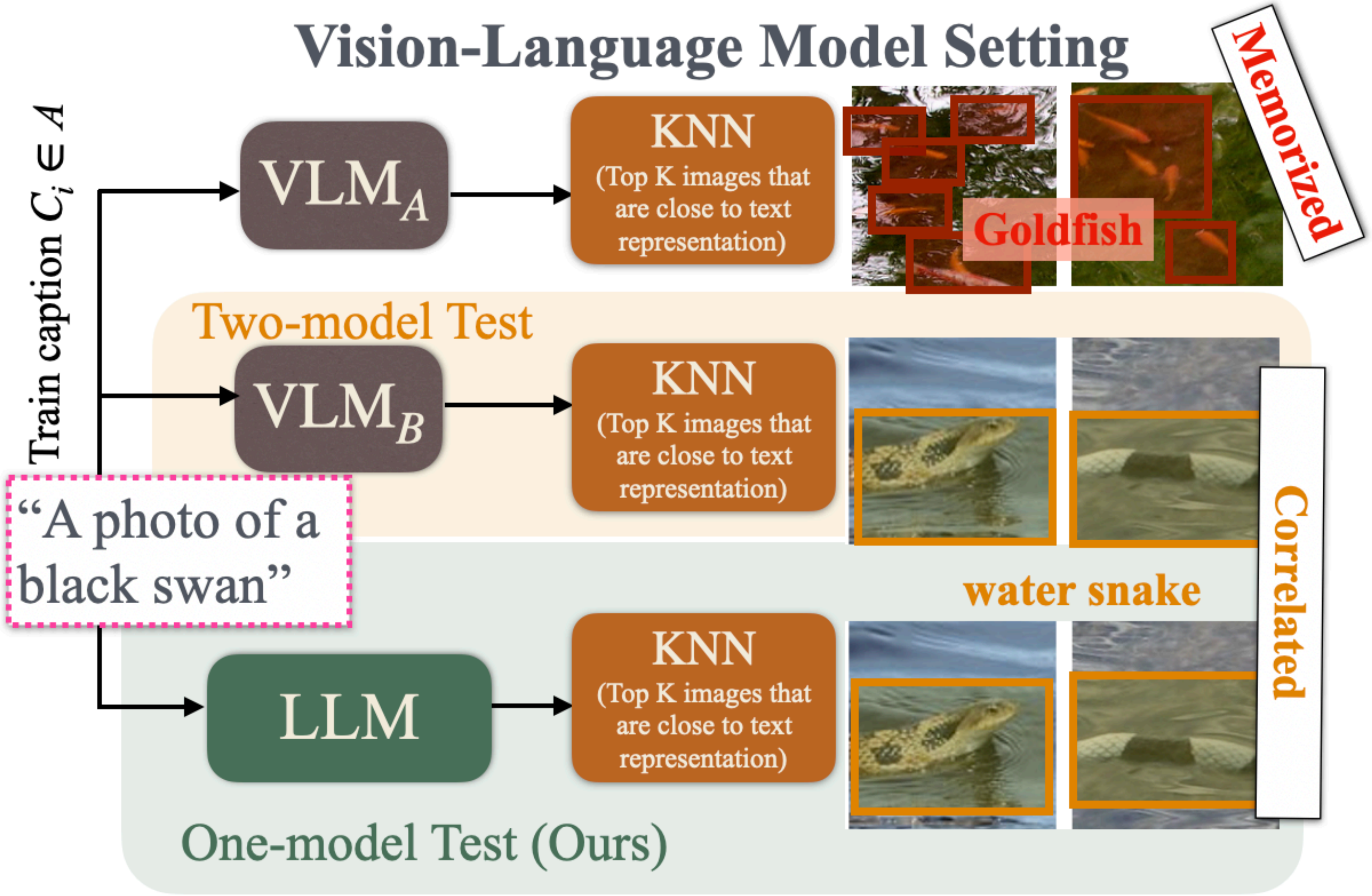
Detecting unintended memorization with one vs two model tests



One-model test allows to :

- train a correlation classifier once per dataset and is independent of the representation model
- measure memorization for pre-trained OSS models for subsets of data not used by correlation classifier

Detecting unintended memorization for vision language models

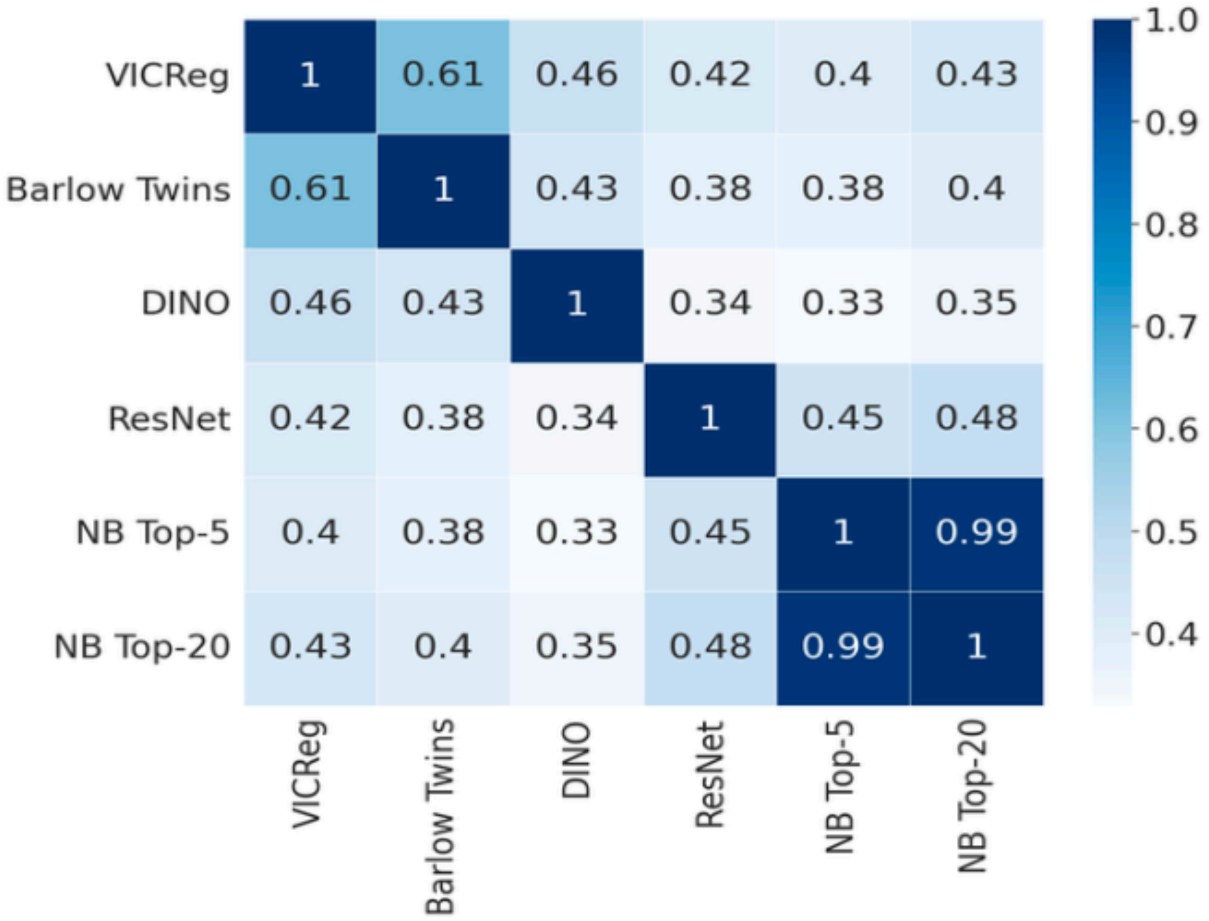

Vision-Language Model Setting
Train caption $C_i \in A$
VLM_A
KNN
(Top K images that are close to text representation)
Goldfish
Memorized
Two-model Test
VLM_B
KNN
(Top K images that are close to text representation)
"A photo of a black swan"
water snake
Correlated
LLM
KNN
(Top K images that are close to text representation)
One-model Test (Ours)

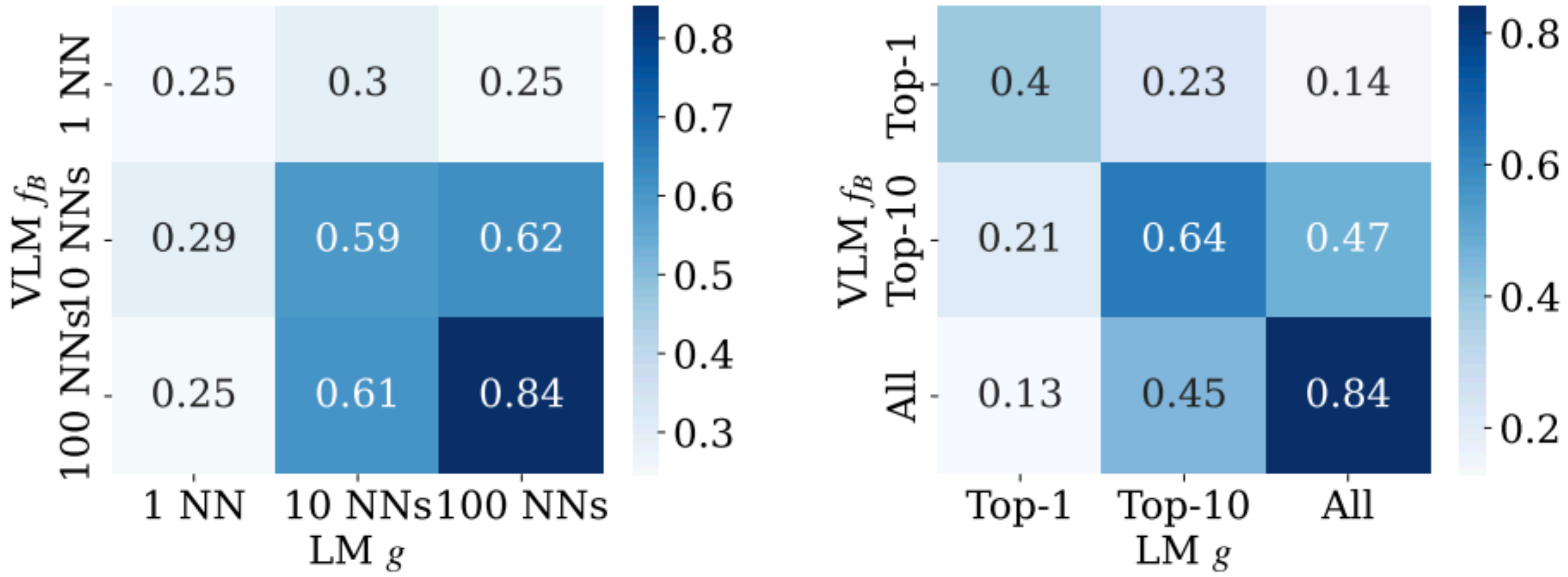
Experimental setup

- **Vision**
  - **Dataset:** ImageNet
    - Two disjoint sets of 300k images used to trained dataset-level correlation classifier and measure the memorization on.
    - Additional distinct 500k images to predict nearest neighbors
  - **Target Models:**
    - VicReg, Barlow Twins, DINO
  - **Reference Models:**
    - ResNet50
    - Naive Bayes Classifier
      - Features are based on annotations from Grounded-SAM [Liu et al., 2023, Ren et al., 2024]

- **Vision Language Models**
  - **Dataset:** 40M Shutterstock
  - **Target Model:** ResNet-50 CLIP model pre-trained on the YFCC15M
  - **Reference Models:** GTE language model

Sample-level correlation classifier agreement



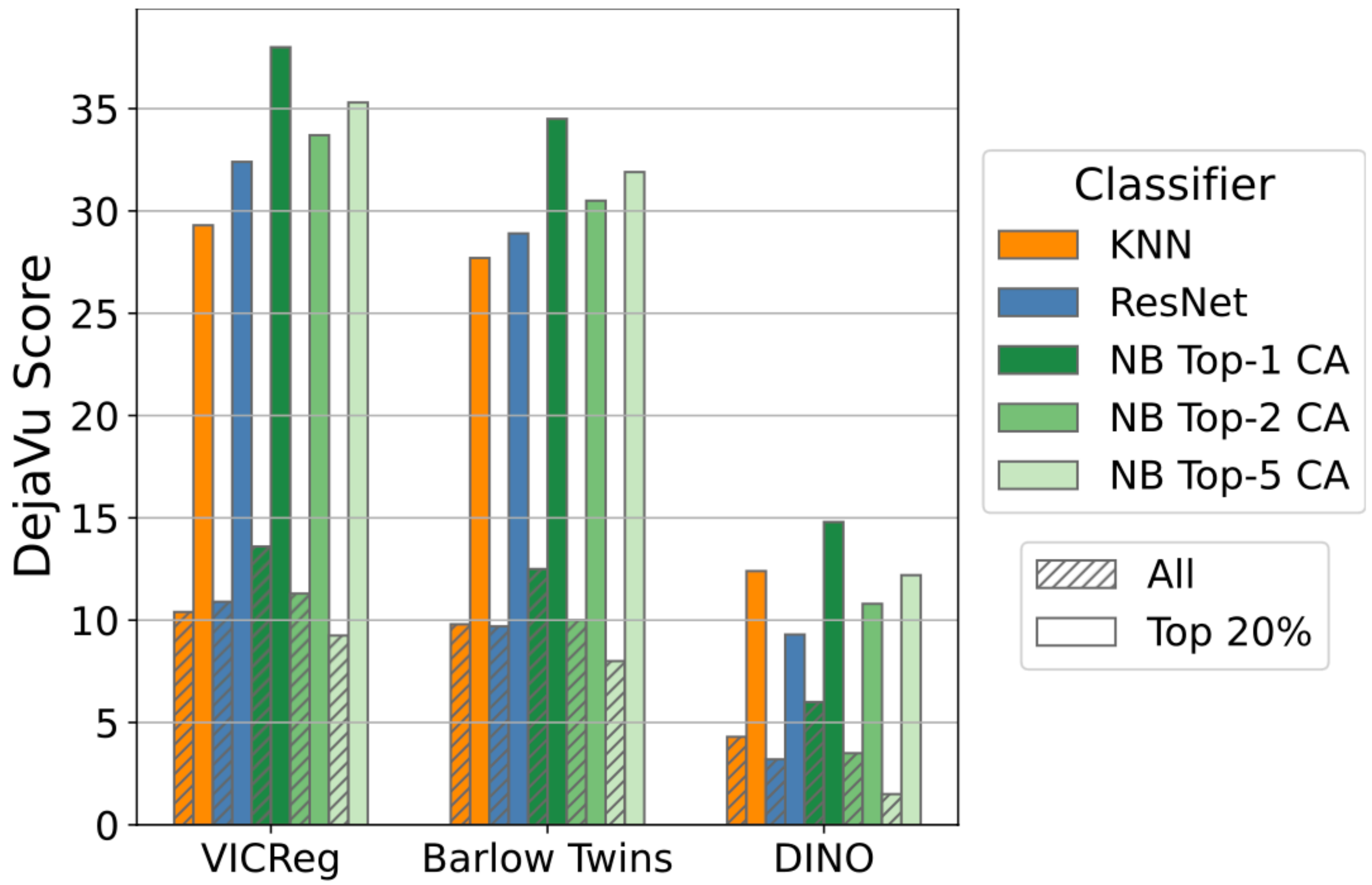


(a) Predicting all objects with varying NNs

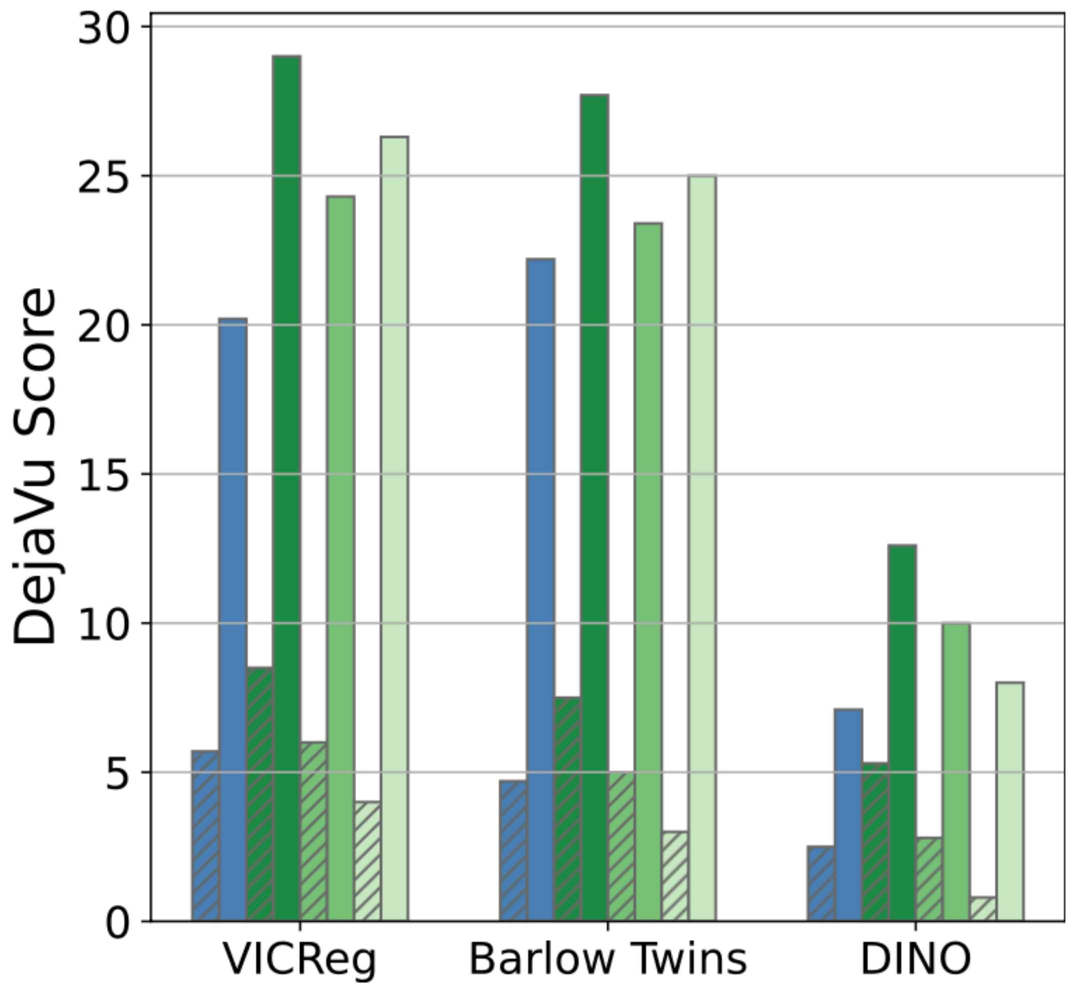
(b) Predicting top-k objects with 100 NNs

Pairwise sample-level agreement fraction using one and two model test reference models

Vision: Memorization in pre-trained OSS models vs models trained on smaller subsets

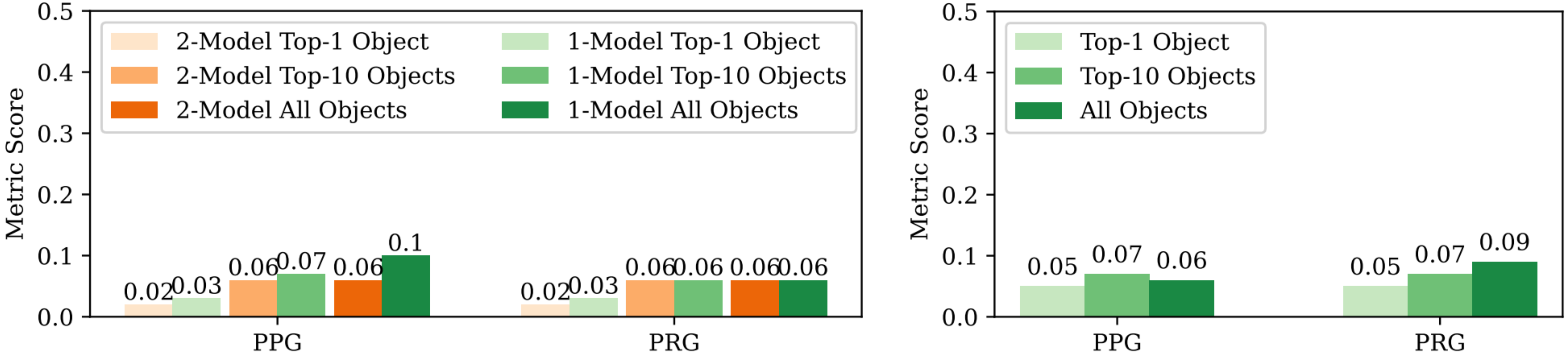


Comparison of overall and Top 20% most confident Déjà vu scores for SSL models trained on a 300k subset of ImageNet.



Comparison of overall and Top 20% most confident Déjà vu scores for trained for pre-trained OSS models

VLM: Memorization in pre-trained OSS models vs models trained on smaller subsets



(a) One-model vs two-model tests for Shutterstock models.

(b) OSS model pre-trained on YFCC15M.

Takeaways

- We propose an efficient way of measuring unintended memorization without having to train shadow image representation and vision language models

- Our is effective for pre-trained OSS models and shows that those models memorize less than the same models trained on smaller subsets of the training data