Discovering Unanticipated Semantic Leakage from Intermediate Representations

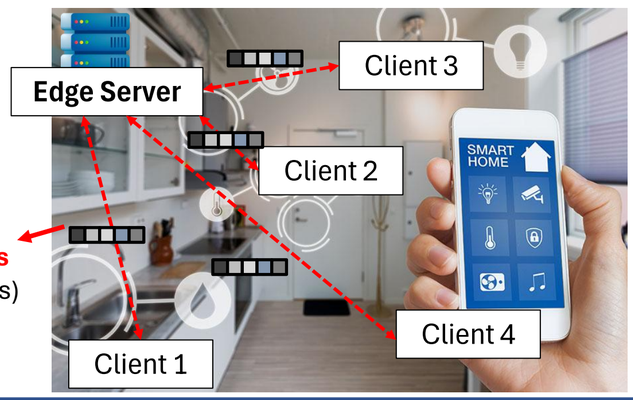
Bangjie Sun[1], Sean Rui Xiang Tan[1], Rui Xiao[2], Mun Choon Chan[1] and Jun Han[3]

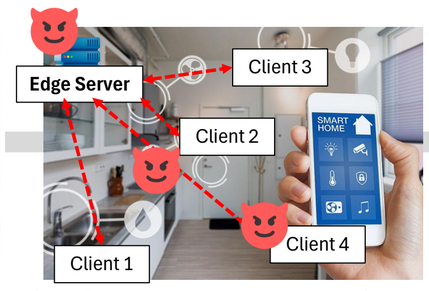
[1]National University of Singapore, [2]Shanghai University of Finance and Economics, [3]KAIST

Scan to read

1. Introduction

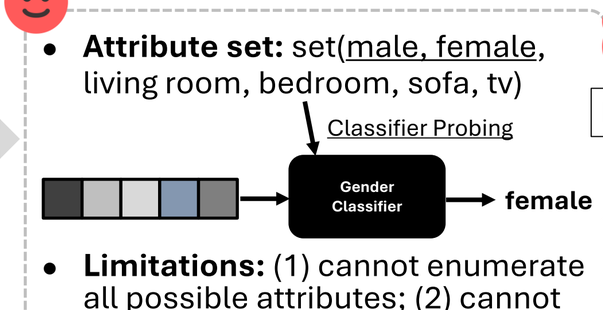
- **Semantic communication gains popularity in distributed AI systems.**
  - Resource-limited clients keep raw data locally
  - Offload the heavier part of model training or inference on edge server

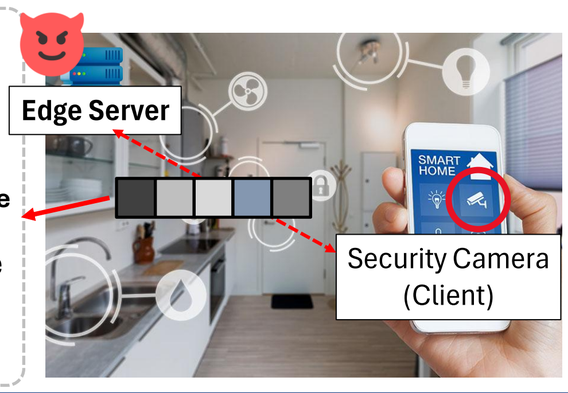




An adversary acts as an eavesdropper in a distributed environment (e.g., *honest-but-curious* edge server)

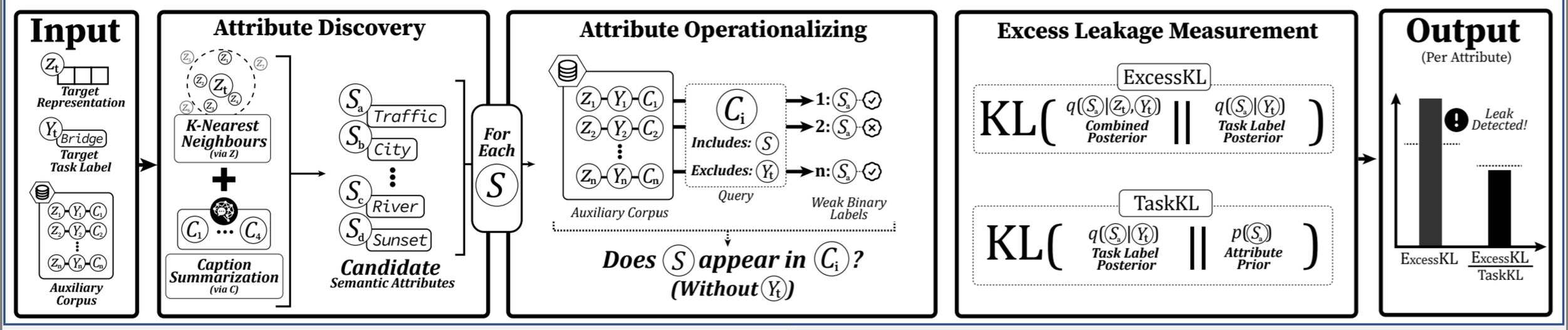
- **Leakage is a known problem but only limited to *closed-world* settings: an adversary assumes predefined set of attributes.**
  - **Attribute set:** set(male, female, living room, bedroom, sofa, tv)
  - **Limitations:** (1) cannot enumerate all possible attributes; (2) cannot discover unanticipated semantic leakage





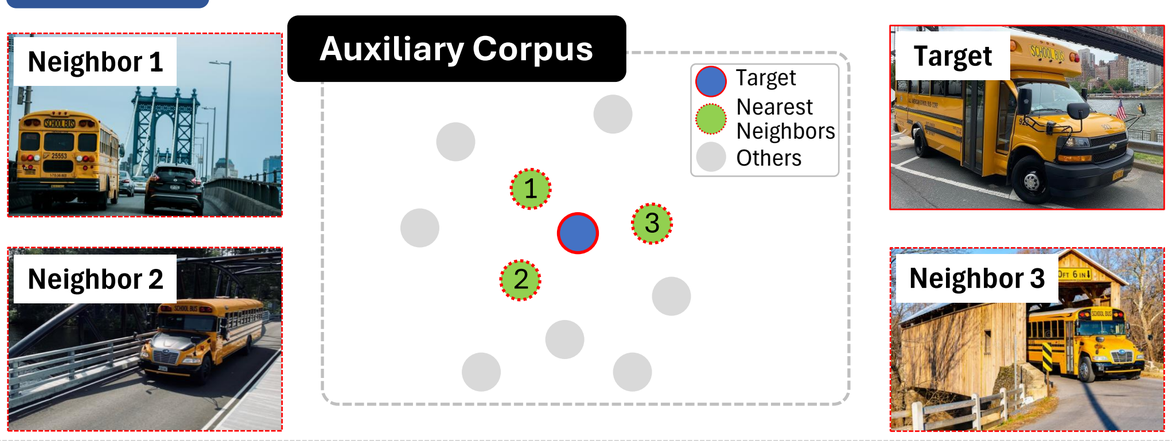
2. Overview of *REVEAL*: the First **Open-World** Semantic Inference Attack

- **What semantic information can an adversary infer from an exposed embedding beyond the intended task (i.e., the predicted class)?**
- **In the *open-world* setting, the adversary actively discovers potential semantic attributes without any predefined set.**

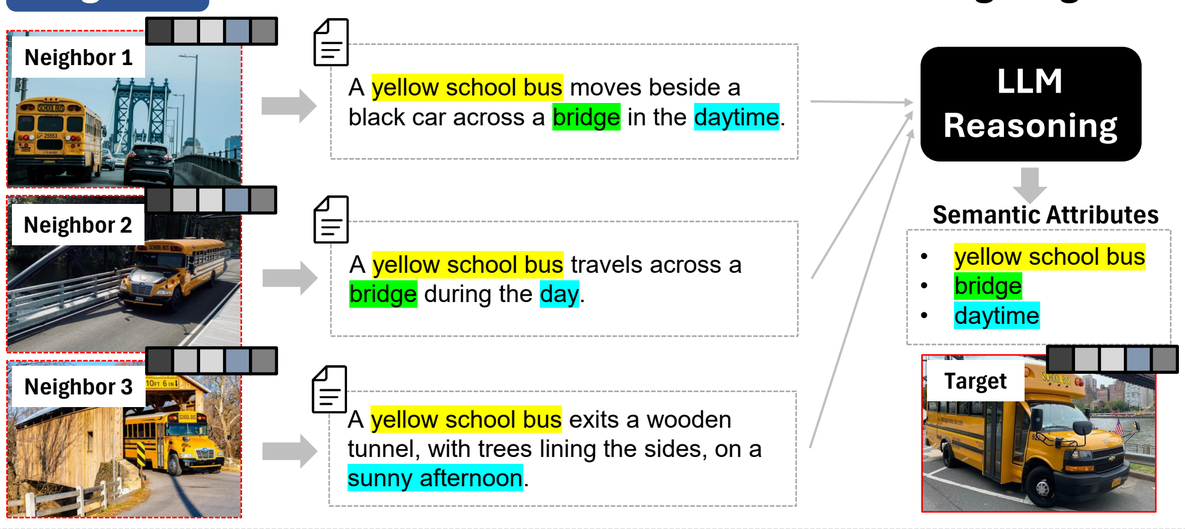


3. Core Insights of *REVEAL*

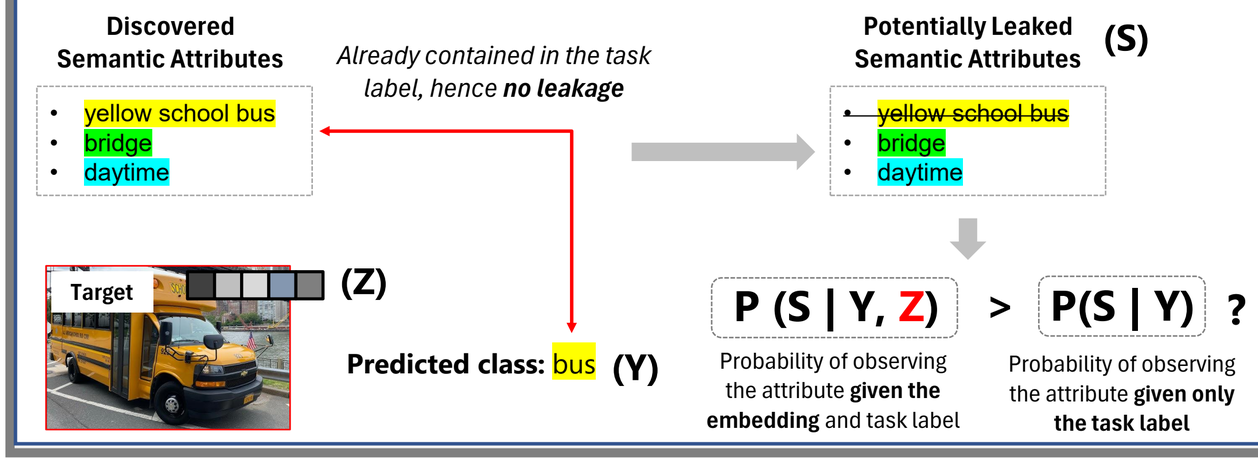
- **Insight #1: Neighbors in embedding space provide semantic info.**



- **Insight #2: LLM summarizes common semantics among neighbors.**

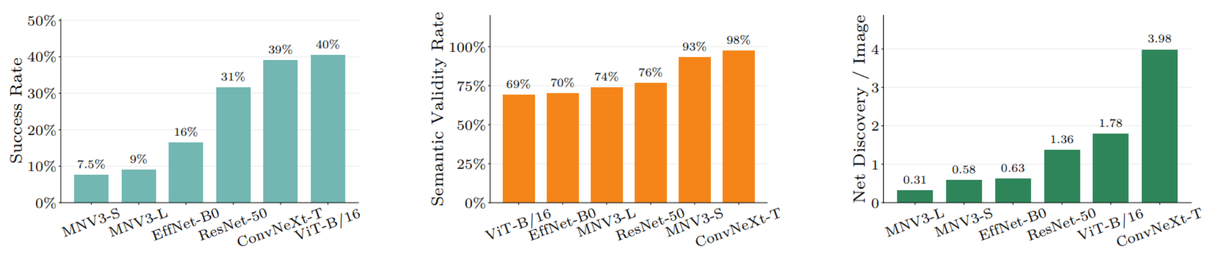


- **Insight #3: Information theory helps quantify semantic leakage.**



4. Evaluation

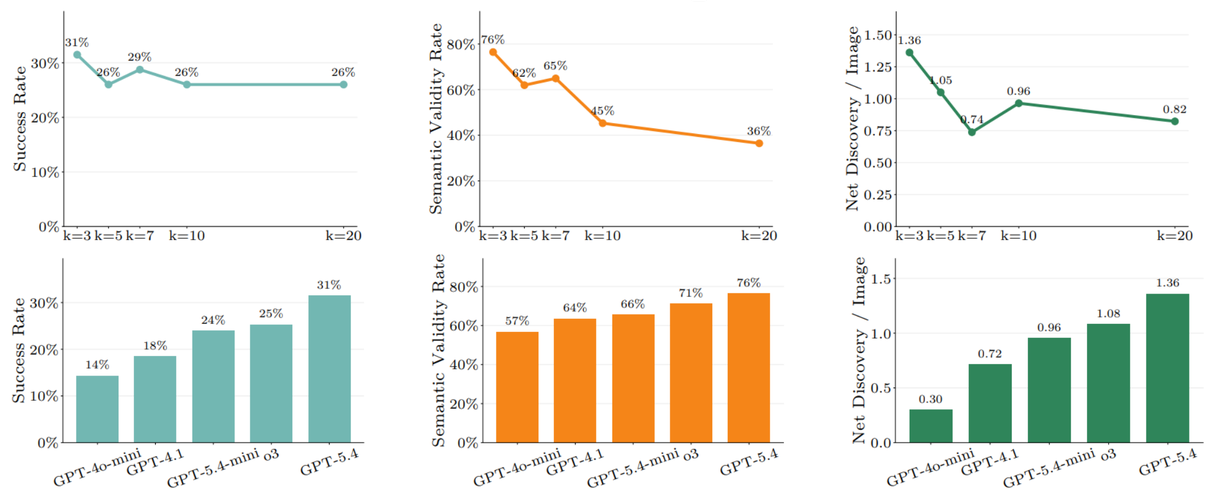
- **Experimental Setup**
  - **Target Dataset:** 160 synthetic images by Stable Diffusion v3 in a controlled setting (i.e., *a dominant primary object + secondary + ternary + background*)
  - **Auxiliary Corpus:** Densely Captioned Images (DCI) containing 7.8K images with human-authored captions
  - **Models Investigated:** Variants of MobileNet, ResNet-50, EfficientNet, ConvNeXt-Tiny, ViT-B (all pre-trained on ImageNet-1K)
- **Evaluation Metrics**
  - **Success Rate:** % of embeddings that *REVEAL* infers at least one non-task attribute
  - **Semantic Validity Rate:** % of attributes genuinely present in the original image
  - **Net Discovery per Image:** # of valid attributes minus # of hallucinated attributes
- **RQ1: Can *REVEAL* surface semantically valid attributes?**



Takeaway #1

Leakage increases as model size (i.e., # of parameters) scales, but validity depends on model architecture.

- **RQ2: Is *REVEAL* sensitive to # of neighbors and LLM choices?**



Takeaway #2

Performance of *REVEAL* degrades as # of neighbors increases, and it improves almost linearly with the reasoning capability of LLM (i.e., version and model size).