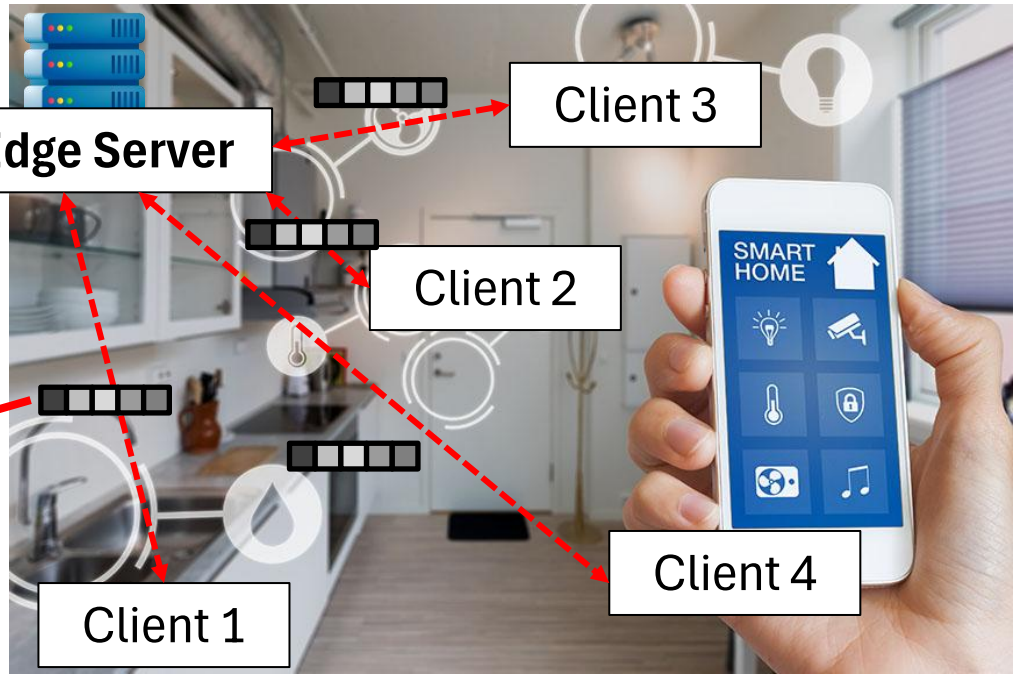
What Do Neighbors Know? Open-World Semantic Inference Attack on Intermediate Representations

Bangjie Sun*, Sean Rui Xiang Tan*, Rui Xiao†, Mun Choon Chan*, and Jun Han‡

* National University of Singapore
† Shanghai University of Finance and Economics
‡ KAIST

Popularity of Split Learning in Edge AI

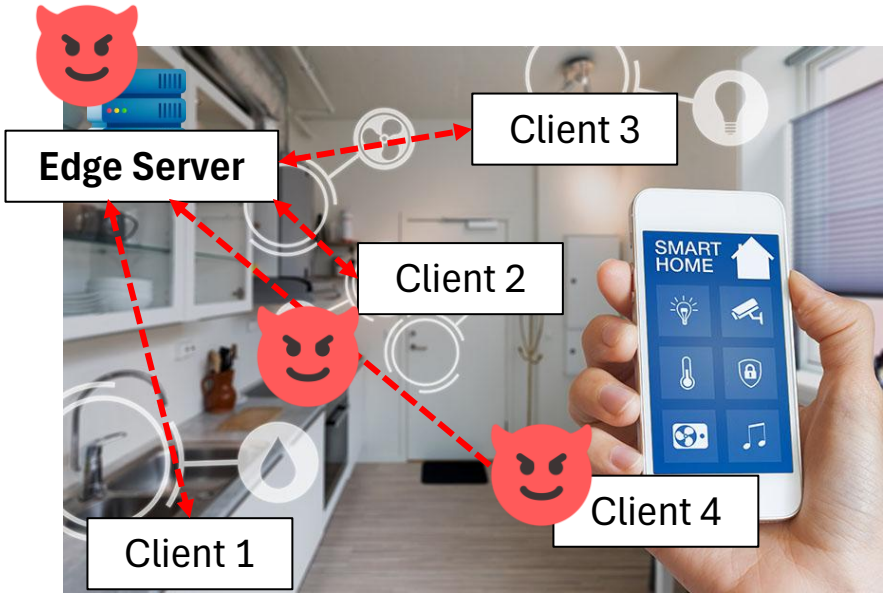
- Resource-limited clients keep raw data locally and offload the heavier part of model training or inference on edge server



Intermediate representations
(e.g., embeddings)

Semantic Leakage is a Known Problem

- An adversary acts as an eavesdropper in a distributed environment
  - *Honest-but-curious* edge server
  - Network eavesdropper
  - On-device malware
- **Goal:** to infer semantics of original data leaked by embeddings
- **Capabilities:** have access to the embeddings and model API



Semantic Leakage is a Known Problem

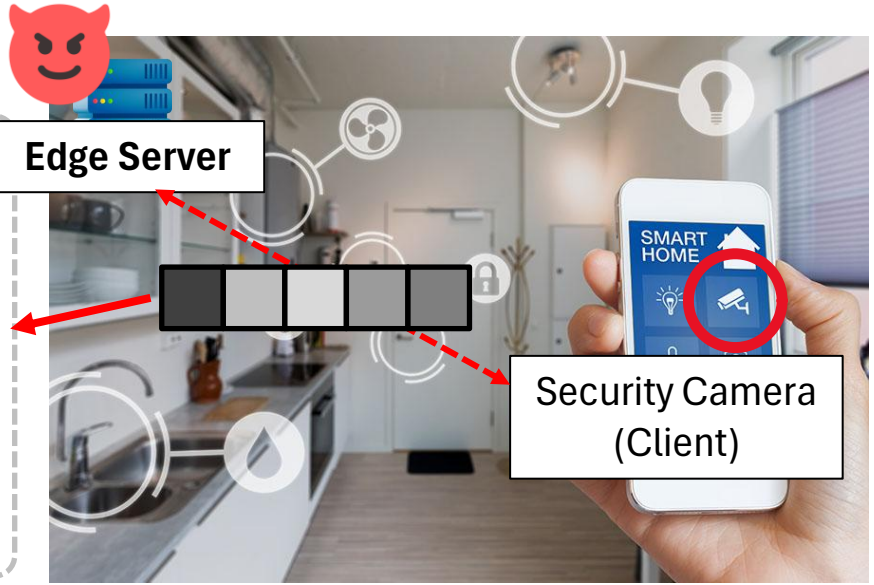
- What semantic information can an adversary infer from an exposed embedding beyond the intended task (i.e., the predicted class)?

Security Camera

- **Task:** intruder detection

Yes / No? (Binary)

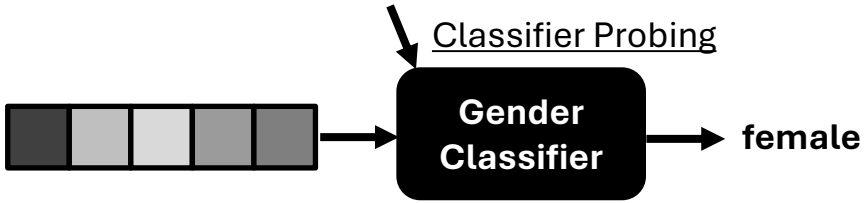
- **Potential leakage:** homeowner's age, gender, or the layout of the room



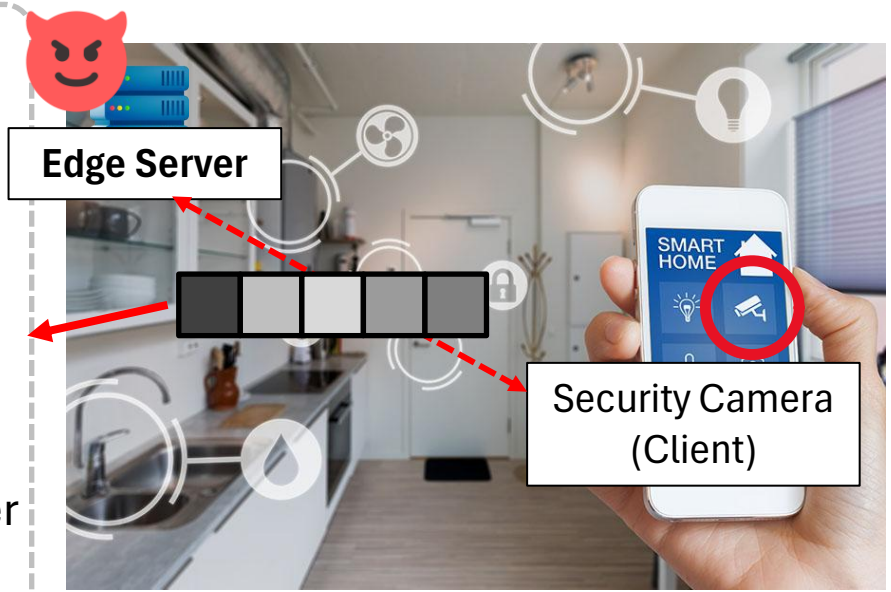
But Existing Works are Mainly Closed-World

- **Closed-world:** the adversary specifies a predefined set of target semantic attributes in advance

- **Attribute set:** set(male, female, living room, bedroom, sofa, tv)



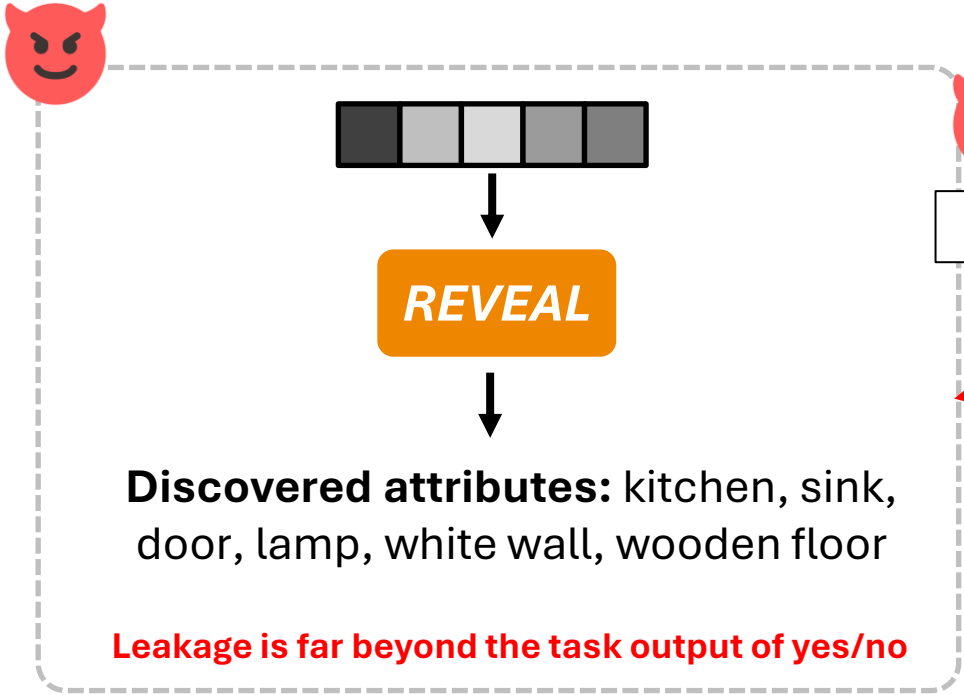
- **Limitations:** (1) cannot enumerate all possible attributes; (2) cannot discover unanticipated semantic leakage

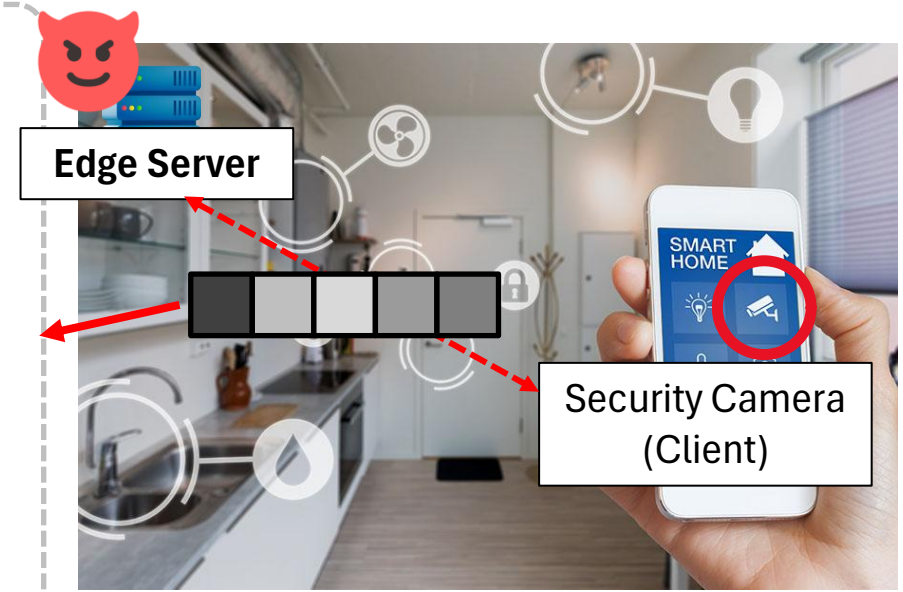


Can we infer semantic information under an open-world setting?

REVEAL: Open-World Semantic Inference

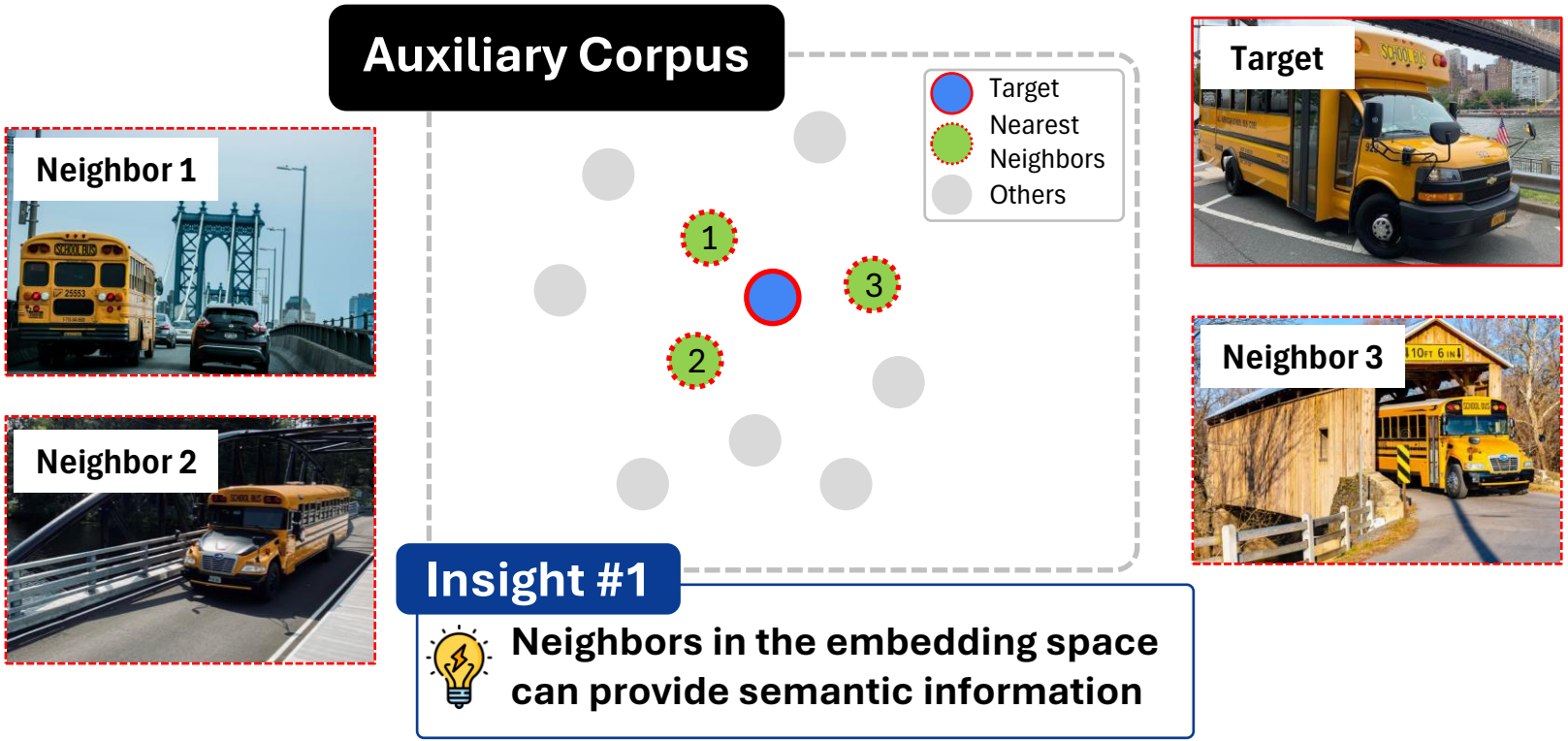
- **Open-world:** the adversary actively discovers potential semantic attributes contained in the embeddings without any predefined target set





Core Idea of *REVEAL*

- Key enablers are (1) auxiliary corpus and (2) LLM reasoning capability



Core Idea of *REVEAL*

- Key enablers are (1) auxiliary corpus and (2) LLM reasoning capability

A yellow school bus moves beside a black car across a bridge in the daytime.

A yellow school bus travels across a bridge during the day.

A yellow school bus exits a wooden tunnel, with trees lining the sides, on a sunny afternoon.

LLM Reasoning

Semantic Attributes

- yellow school bus
- bridge
- daytime

Core Idea of *REVEAL*

- Key enablers are (1) auxiliary corpus and (2) LLM reasoning capability

A yellow school bus moves beside a black car across a bridge in the daytime.

A yellow school bus travels across a bridge during the day.

LLM Reasoning

Semantic Attributes

- yellow school bus
- bridge
- daytime

Insight #2

LLMs have the capabilities of summarizing the common the semantic attributes among the neighbors

Core Idea of *REVEAL*

- Utilize information theory to measure the amount of leakage beyond the intended task label (i.e., the predicted class)

Discovered Semantic Attributes

- yellow school bus
- bridge
- daytime

*Already contained in the task label, hence **no leakage***

Insight #3

LLMs can filter out task-relevant semantic attributes by comparing with the task label

(Z)

Predicted class: bus **(Y)**

Core Idea of *REVEAL*

- Utilize information theory to measure the amount of leakage beyond the intended task label (i.e., the predicted class)

Discovered Semantic Attributes

- yellow school bus
- bridge
- daytime

*Already contained in the task label, hence **no leakage***

Potentially Leaked Semantic Attributes (S)

- ~~yellow school bus~~
- bridge
- daytime

(Z)

Predicted class: bus **(Y)**

$P(S \mid Y, Z) > P(S \mid Y)$?

Probability of observing the attribute **given the embedding** and task label

Probability of observing the attribute **given only the task label**

Core Idea of *REVEAL*

- Utilize information theory to measure the amount of leakage beyond the intended task label (i.e., the predicted class)

Insight #4

The amount of leakage can be measured by comparing the probability of S with and without the embedding Z

Potentially Leaked Semantic Attributes (S)

- ~~yellow school bus~~
- bridge
- daytime

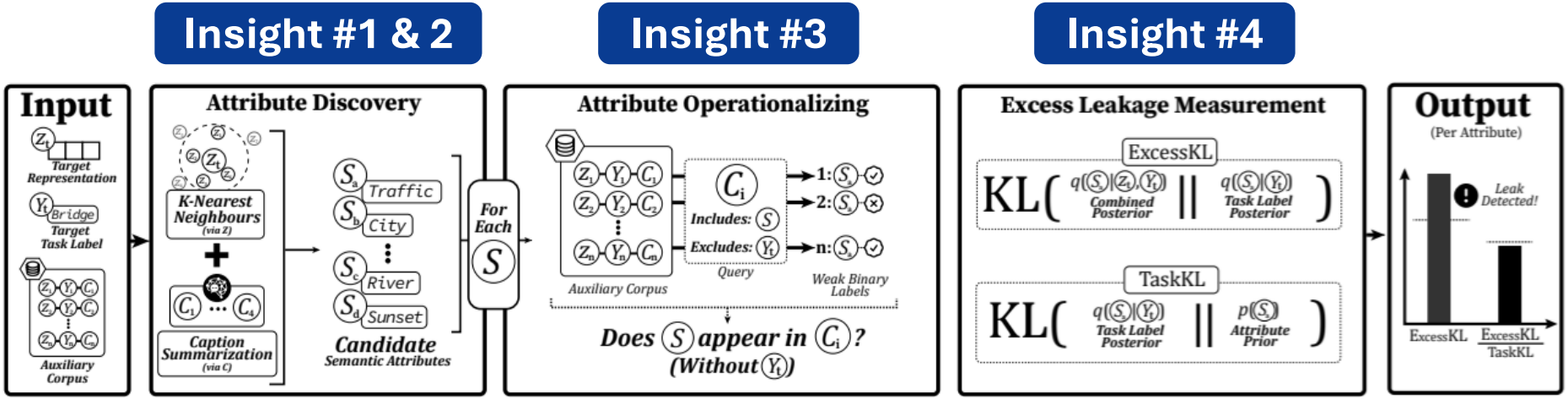
$$P(S \mid Y, Z) > P(S \mid Y) \; ?$$

Probability of observing the attribute **given the embedding** and task label

Probability of observing the attribute **given only the task label**

Implementation of *REVEAL*

- In summary, we employ a discover-then-measure approach to (1) actively discover candidate semantic attributes from an embedding and (2) measure the additional leakage beyond the intended task label



Experimental Setup

- We aim to answer two research questions:

  RQ1: Can REVEAL surface semantically valid attributes?

  RQ2: Is REVEAL sensitive to number of neighbors and LLM choices?

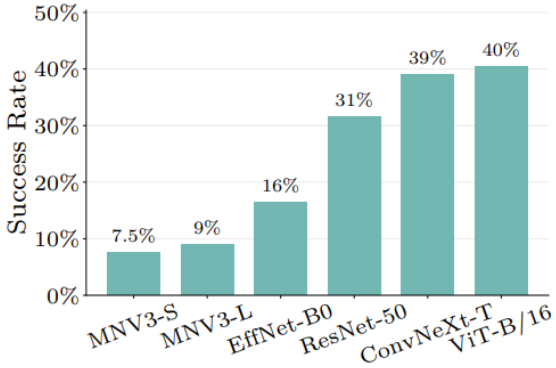
- **Target Dataset:** 160 synthetic images by Stable Diffusion v3 in a controlled setting (i.e., *a dominant primary object + secondary + ternary + background*)

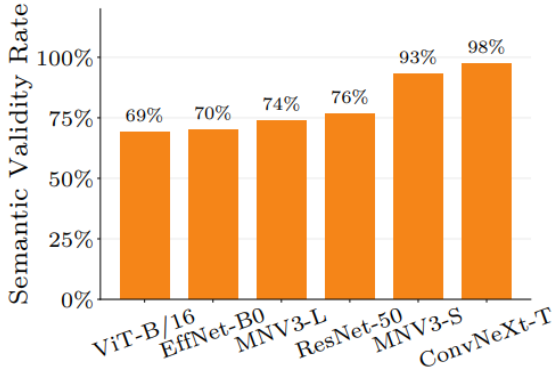
- **Auxiliary Corpus:** Densely Captioned Images (DCI) containing 7.8K images with human-authored captions

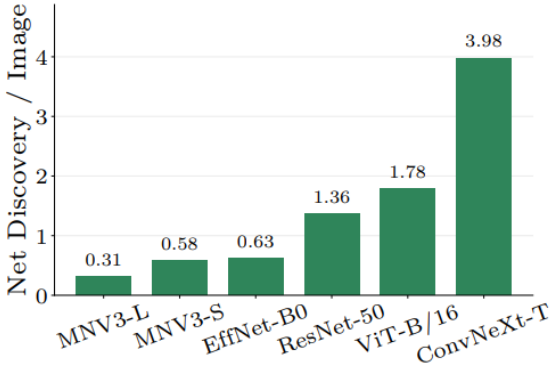
- **Models Investigated:** Variants of MobileNet, ResNet-50, EfficientNet, ConvNeXt-Tiny, ViT-B (all pre-trained on ImageNet-1K)

RQ1: Semantic Validity

- To verify the semantic validity of semantic attributes, we use:
  - ❑ **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





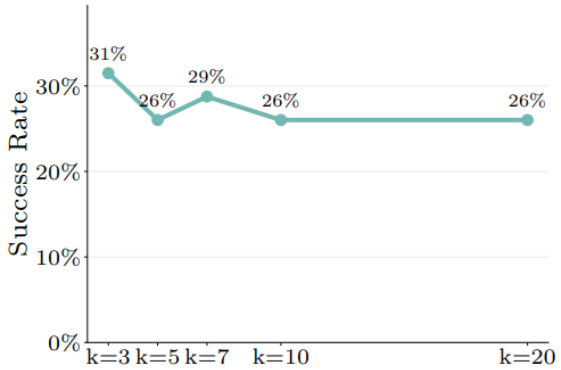


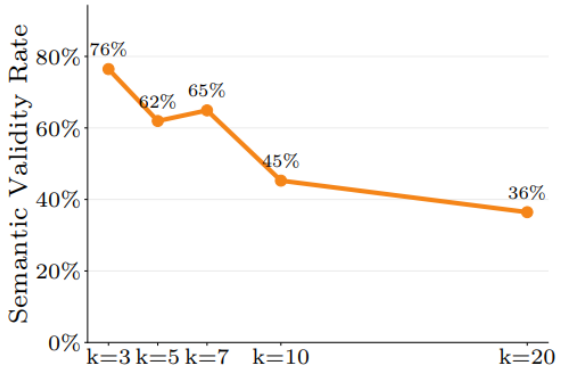
Takeaway #1

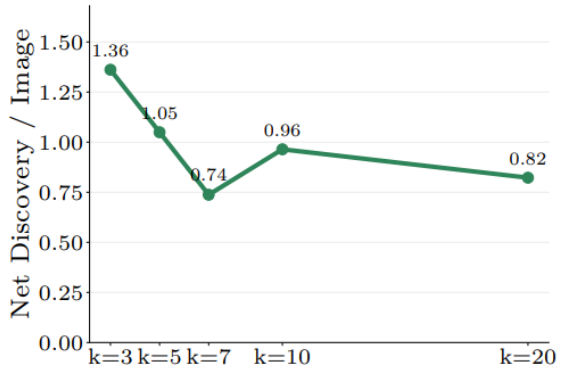
Semantic leakage increases as model size (i.e., # of parameters) increases, but validity rate depends on model architecture

RQ2-1: Sensitivity to # of neighbors

- To verify the semantic validity of semantic attributes, we use:
  - ❑ **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





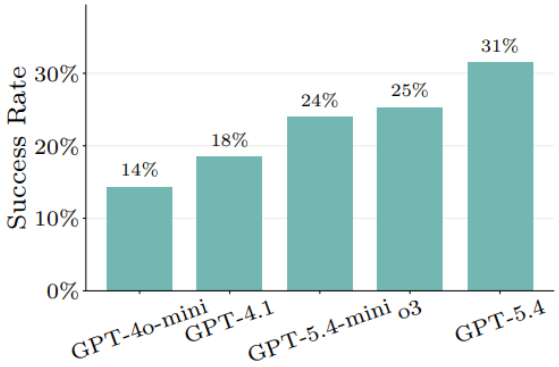


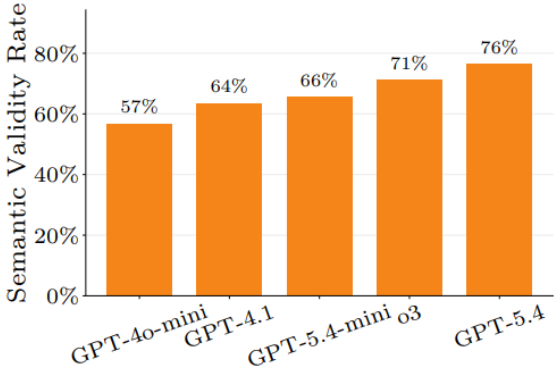
Takeaway #2

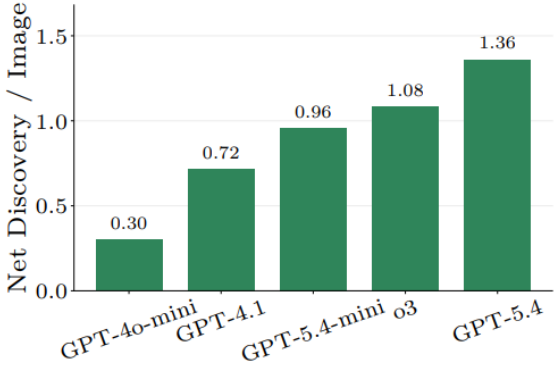
Performance of *REVEAL* degrades as # of neighbors increases

RQ2-2: Sensitivity to Choice of LLM

- To verify the semantic validity of semantic attributes, we use:
  - ❑ **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







Takeaway #3

Performance of *REVEAL* improves almost linearly with the reasoning capability of LLM (i.e., version and model size)

Discussion

Importance of Open-World

- Prior "checklist" mentality gives defenders a false sense of security
- An adversary could exploit any leaked semantics to launch attacks in unanticipated ways
- Future defenses must restrict the overall capacity to encode non-task features

Extension of REVEAL

(1) Advanced neighbor retrieval and semantic similarity comparison + iterative LLM summarization

Methodological Improvement

Domain Extension

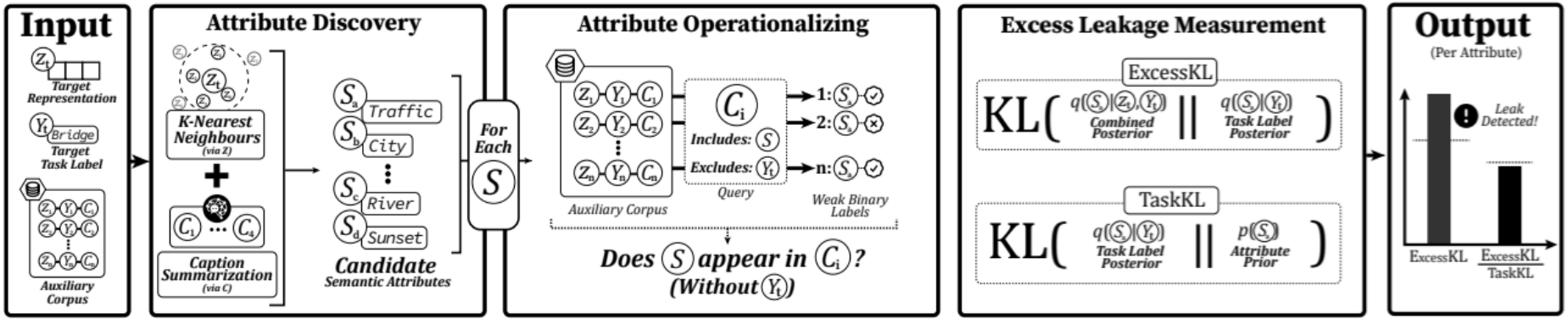
(2) Apply the open-world setting to other modalities (e.g., audio, robotic actions)

Strategy Extension

(3) Extend to other representations (e.g., model weights) in distributed training

Conclusion

- The first open-world semantic inference attack on image embeddings
- Spur future research on open-world semantic inference and encourage defense methods to focus on open-world leakage



Thank you!

Bangjie Sun

PhD Candidate @ NUS

Research Interests

A central premise of my research is that no single provenance signal is sufficient on its own. Beyond cryptographic records and other forms of extrinsic provenance, I study how visible physical signals, sensor fingerprints, and computational forensics provide complementary **intrinsic provenance** for building trustworthy systems that remain robust under adversarial manipulation. I also aim to make provenance recovery and verification **practical on commodity everyday devices** rather than confined to specialized laboratories or proprietary platforms. Ultimately, I seek to advance **hybrid provenance** systems that integrate these signals not only to verify the origin, authenticity, and transformation history of physical and digital artifacts, but also to enable **accountable human-AI workflows** in which transformations, interventions, and responsibility can be meaningfully audited.

Mobile & Sensing Systems PRIMARY

Security & Privacy SECONDARY