Peering Inside the Black-Box: Long-Range and Scalable Model Architecture Snooping via GPU Electromagnetic Side-Channel

Rui Xiao[1], Sibo Feng[1], Soundarya Ramesh[2], Jun Han[3], and Jinsong Han[1]

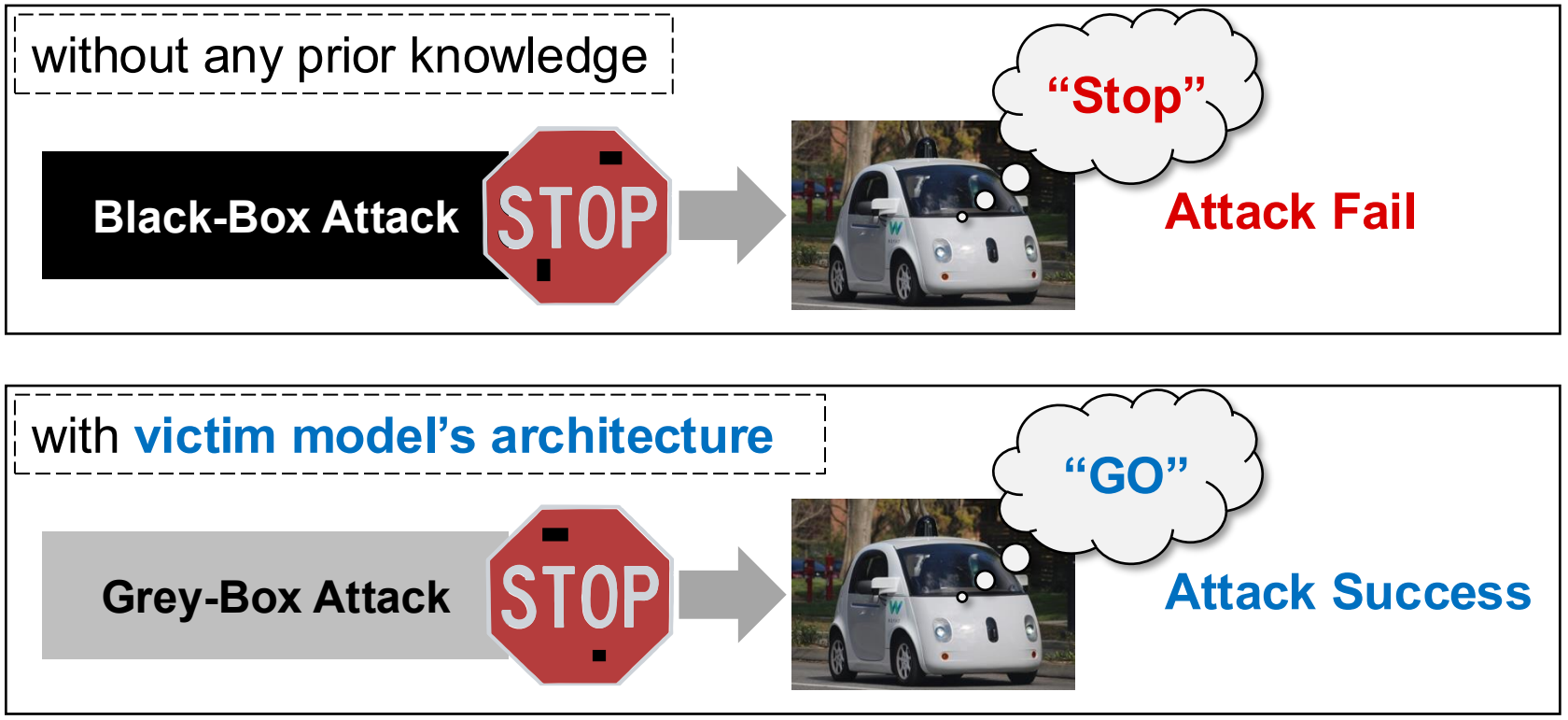
[1] Zhejiang University, [2]Nanyang Technological University, [3]KAIST

NDSS 2026

AI computing is **vulnerable to various adversarial attacks**, such as evasion attacks with adversarial examples.

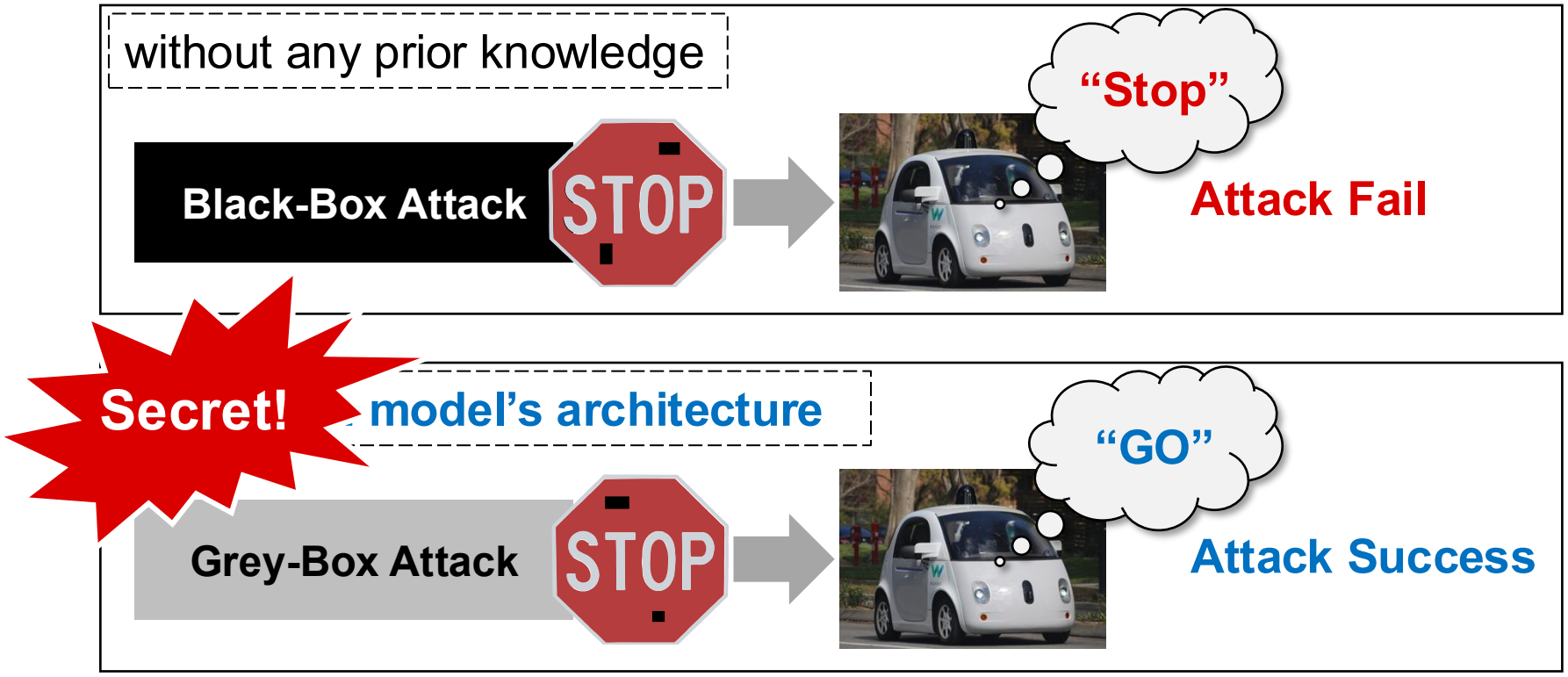
Adversarial attacks requires prior knowledge

- **Prior knowledge**, including the **victim model's architecture** and **weights**, can enhance success rate of attack.



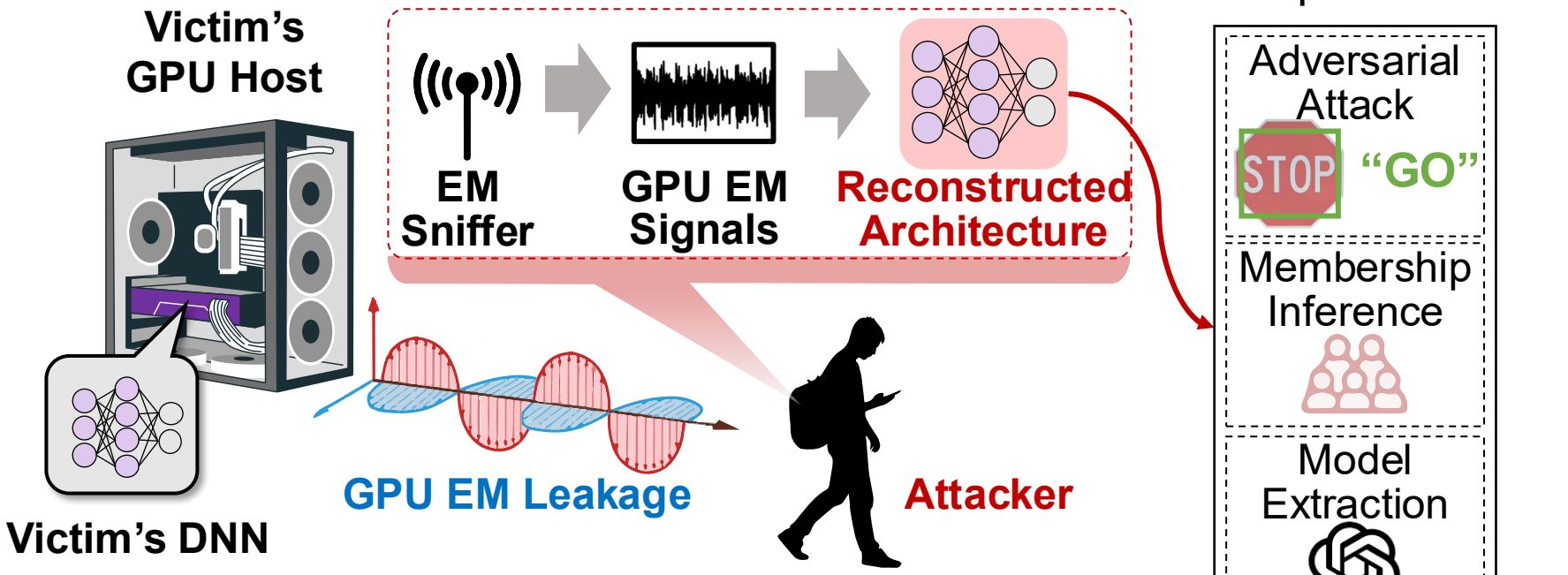
Adversarial attacks requires prior knowledge

- **Prior knowledge**, including the **victim model's architecture** and **weights**, can enhance success rate of attack.



ModelSpy: A novel physical attack

- Snooping the secret **DNN model architecture** from **GPU electromagnetic (EM) leakage**



Subsequent Attacks

Real-World Accident

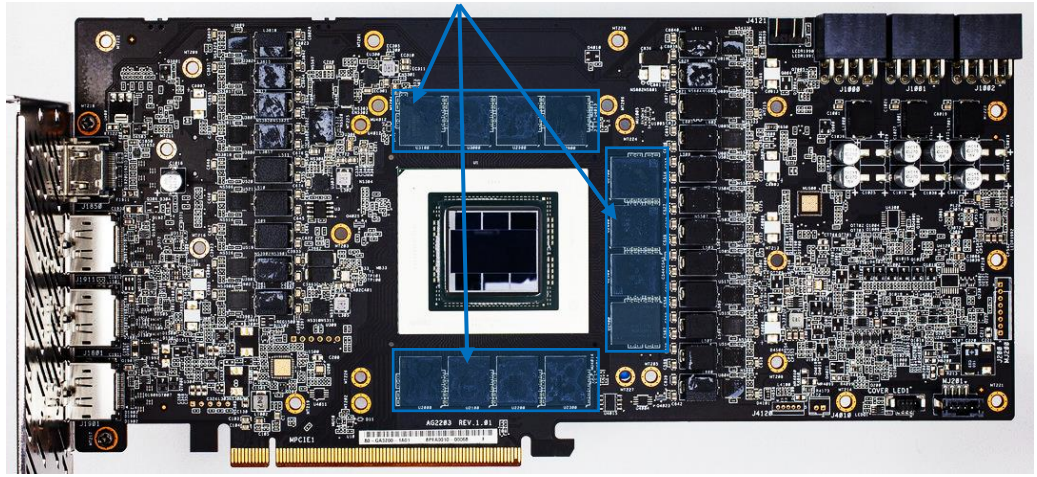
Privacy Leakage

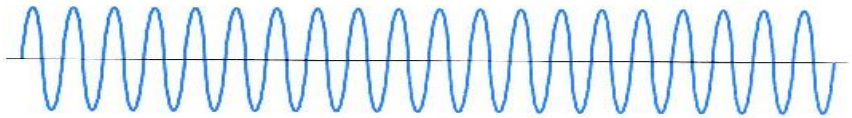
Financial Loss

GPU EM Leakage Analysis

- GPU's digital components constantly emits EM signals
  - including memory clocks, voltage regulators…

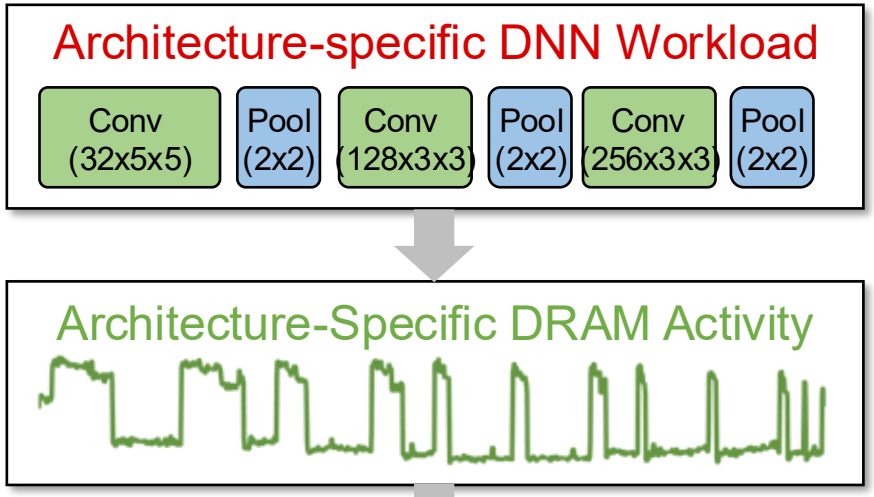
GPU Memory





EM Carrier Signals (MHz ~ GHz)

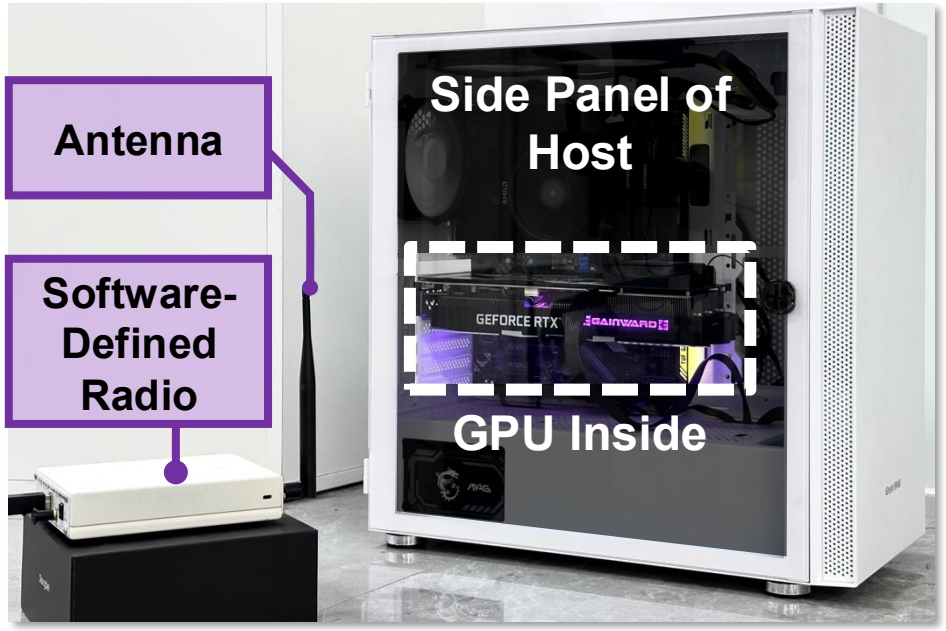
Architecture acting as **Baseband Signal**



EM signals modulated by ***Architecture***

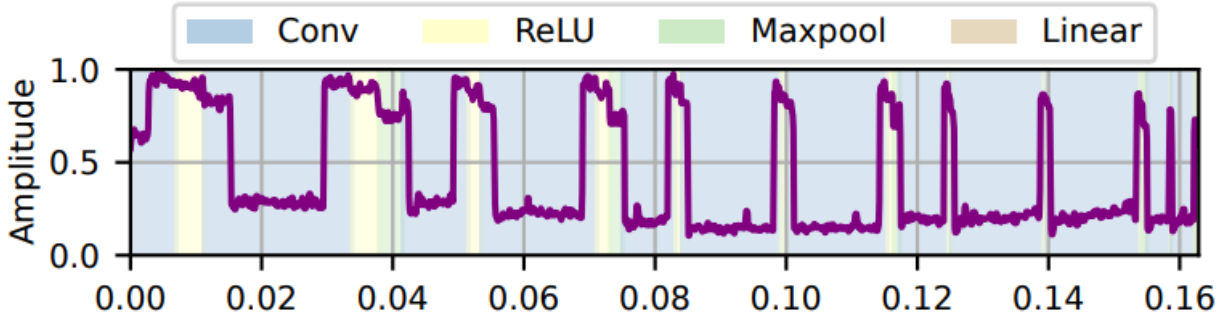
GPU EM Leakage Analysis

- **GPU EM signals** amplitude-modulated by **model architecture**

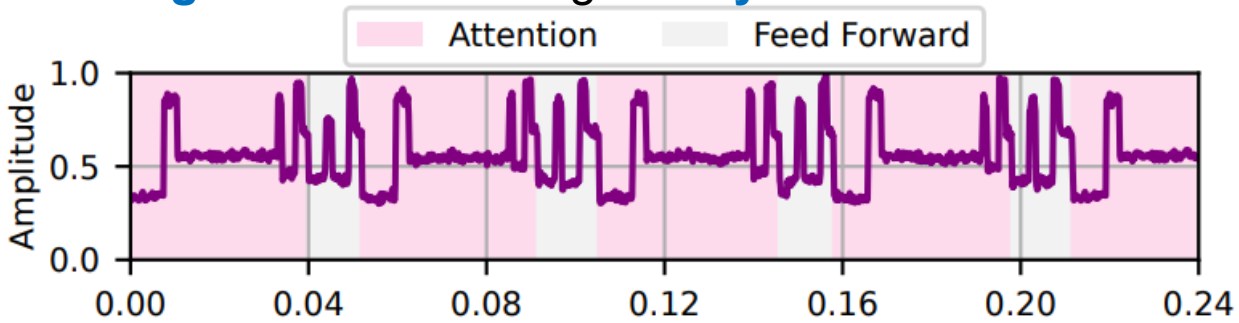


Feasibility Setup

➢ **EM signals** when running a **12-layer CNN** Model



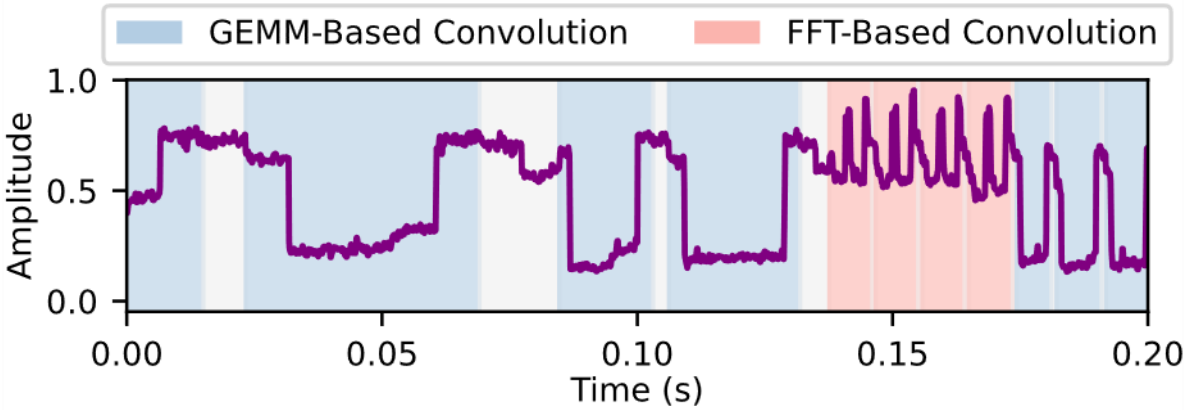
➢ **EM signals** when running a **4-layer Transformer** Model



GPU EM signals can serve as a proxy for architecture reconstruction.

Challenge: Huge Search Space

- Fine-grained architecture reconstruction involves
  - Number of layers
  - Layer types
  - Hyper-parameters for each layer
    - Fully-connected layers: number of neurons
    - Convolutional layers: filter size, number of filters
    - …
  - Example: 16 layers → 5 trillion architectures
- DNN implementation variation
  - E.g., Convolution can be implemented using *GEMM*, *FFT, or Winograd*

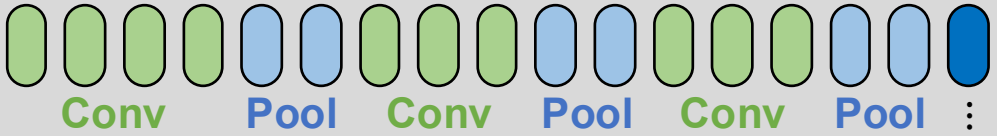


Scalable Solution: Hierarchical Reconstruction

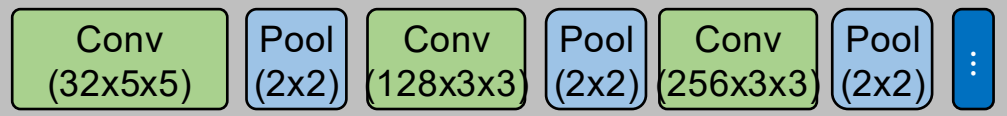
0. input EM Signals

1. extract **contextual features**

2. segment EM signal by **layer type**

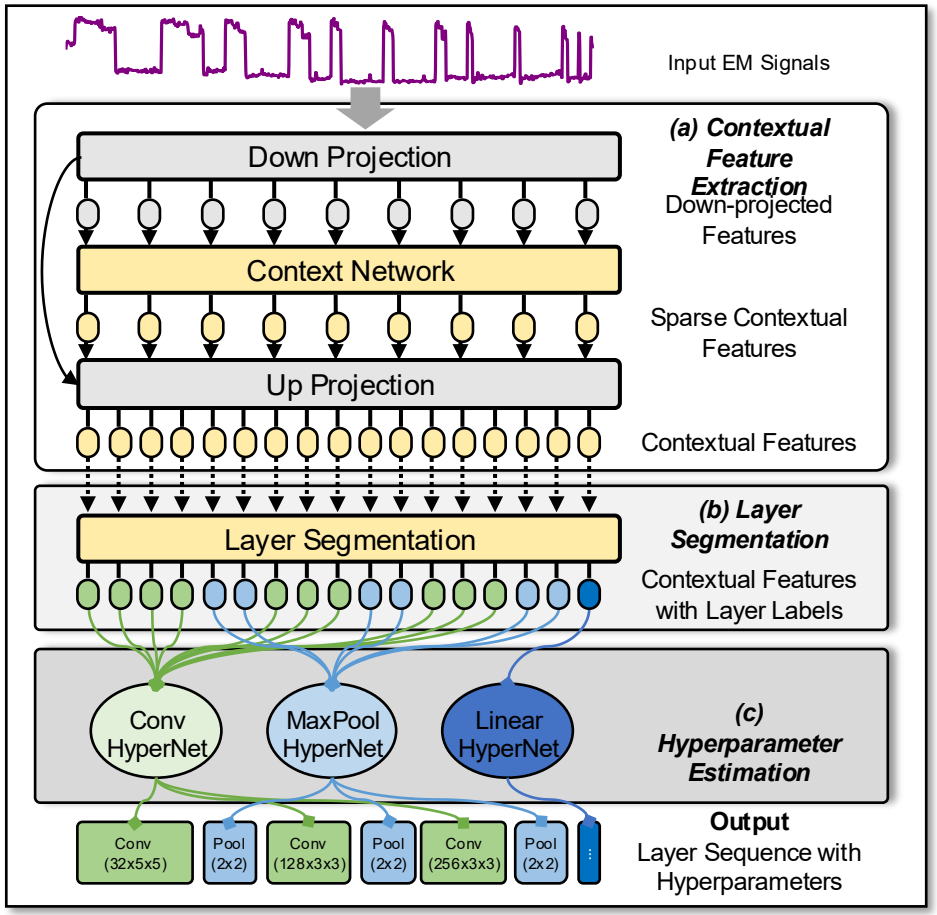


3. identify **hyperparameter** of each layer



Scalable Solution: Hierarchical Reconstruction

0. input EM Signals
1. extract **contextual features**
2. segment EM signal by **layer type**
3. identify **hyperparameter** of each layer

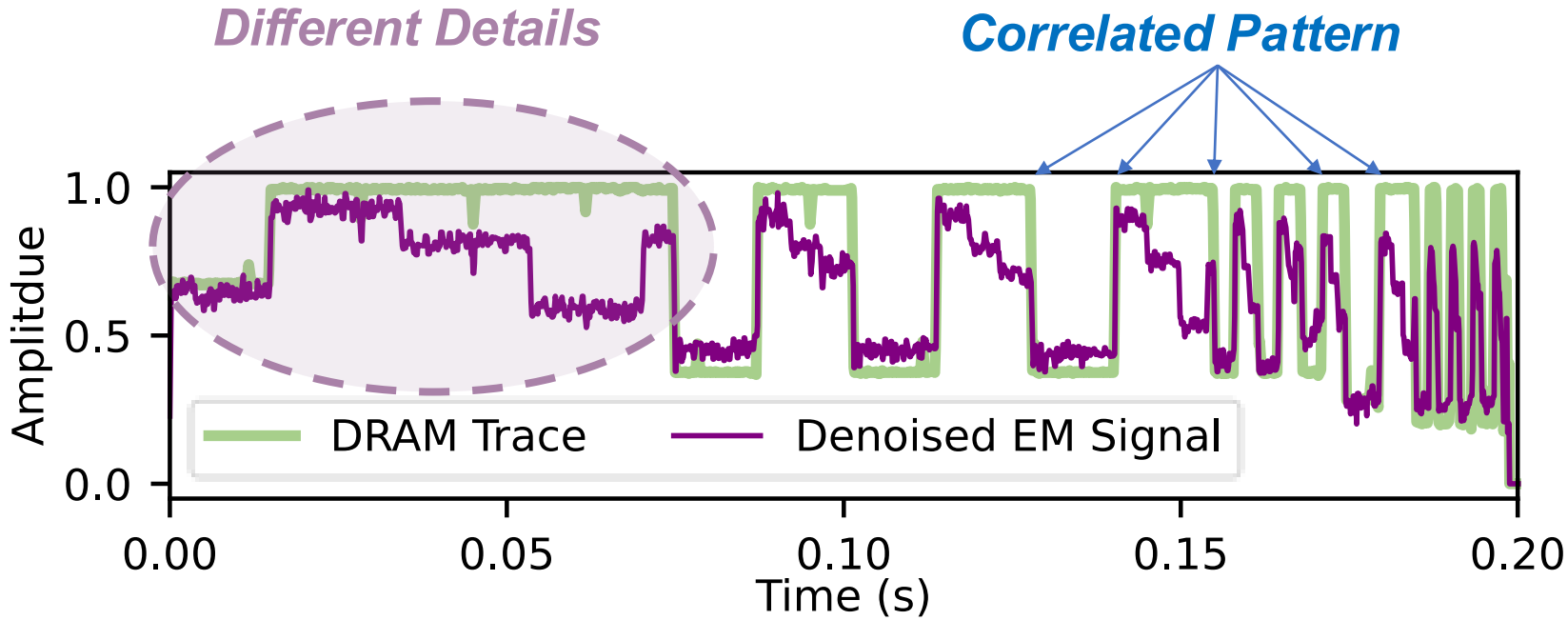


An end-2-end architecture reconstruction engine implemented with deep learning

Further Boost Scalability by Transfer Learning

➢ **Two Key Insights**

- Insight 1: **Temporal correlations** between **DRAM trace** with **EM signals**.

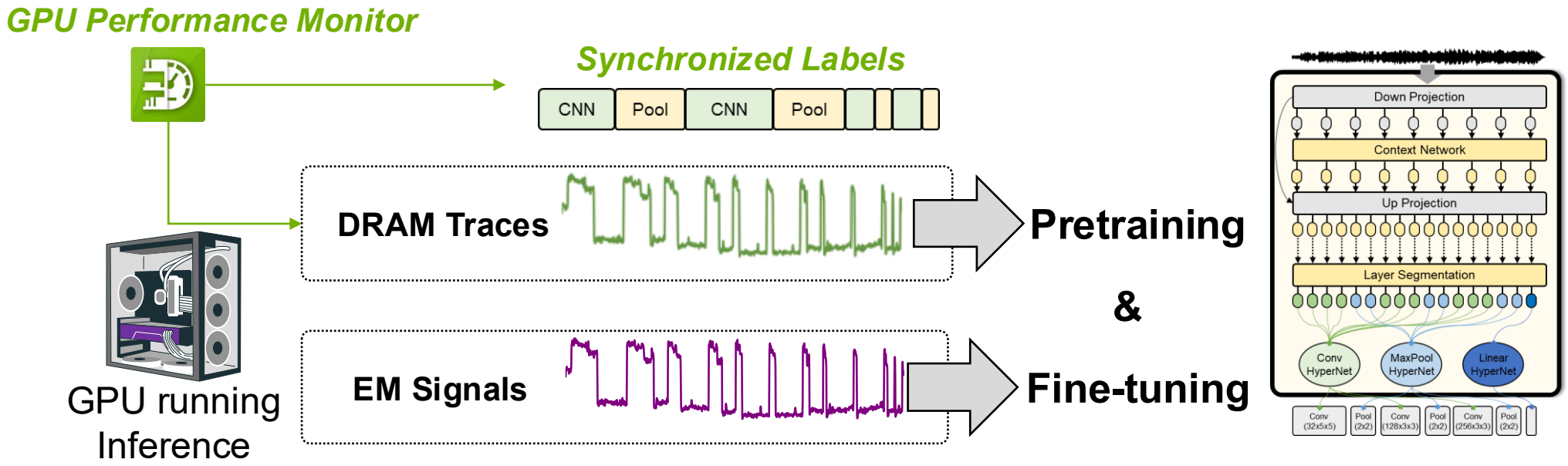


DRAM is ideal for pre-training use

Further Boost Scalability by Transfer Learning

➢ Two Key Insights

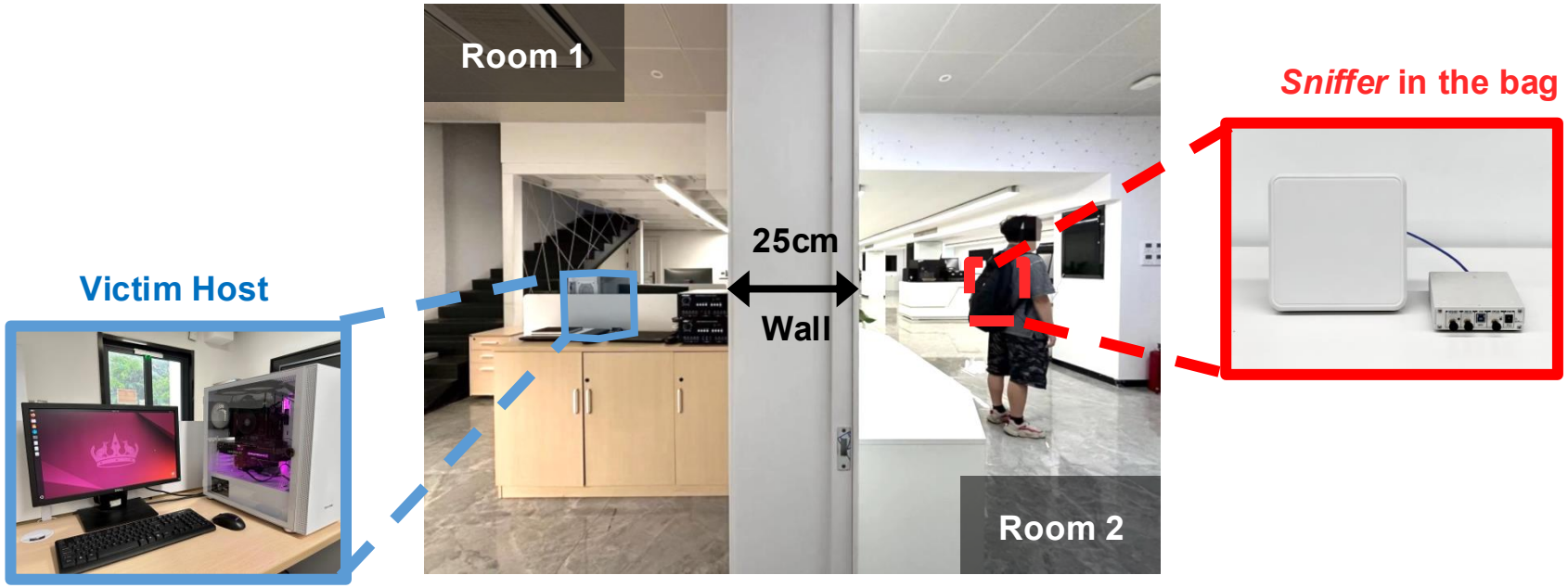
- Insight 1: **Temporal correlations** between **DRAM trace** with EM signals.
- Insight 2: **Automatic** DRAM Data **Collection** and **Annotation**.



We can assemble a **large** and **diverse training dataset** with **minimal efforts.**

Evaluation Setup

- **5** Heterogeneous GPUs (NVIDIA 30 & 40 series)
- **2,820** random DNN architectures (CNN, LSTM, Transformer)
  - Deepest – **ResNet-152**
- **Comprehensive** evaluation under **different conditions**



ModelSpy Performance Summary

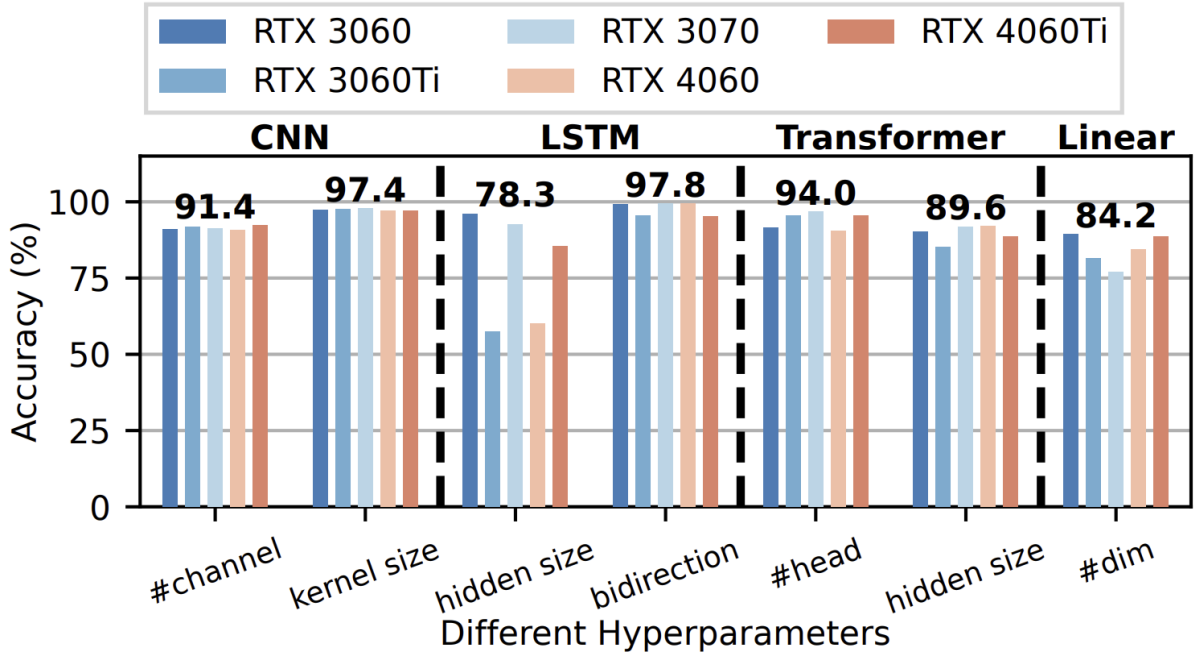
1. Low layer sequence prediction error

- Metric: Edit Distance (NLD)

DNN Architecture	VGG (NLD = 2/66 = 0.031)
Ground-truth Sequence	Conv BN ReLU MaxPool Conv BN ReLU Conv BN Conv Add ReLU Conv BN ReLU Conv BN Add ReLU Conv BN ReLU Conv BN Conv Add ReLU Conv BN ReLU Conv BN Add ReLU Conv BN ReLU Conv BN Conv Add ReLU Conv BN ReLU Conv BN Add ReLU Conv BN ReLU Conv BN Conv Add ReLU Conv BN ReLU Conv BN Add ReLU AvgPool FC
Predicted Sequence	Conv BN ReLU MaxPool Conv BN ReLU Conv BN Conv Add ReLU Conv BN ReLU Conv BN Add ReLU Conv BN ReLU Conv BN Conv Add ReLU Conv BN ReLU Conv BN Add ReLU Conv BN ReLU Conv BN Conv Add ReLU Conv BN ~~ReLU~~ Conv BN Add ReLU Conv BN ReLU Conv BN Conv Add ReLU Conv BN ReLU Conv BN Add ReLU MaxPool FC

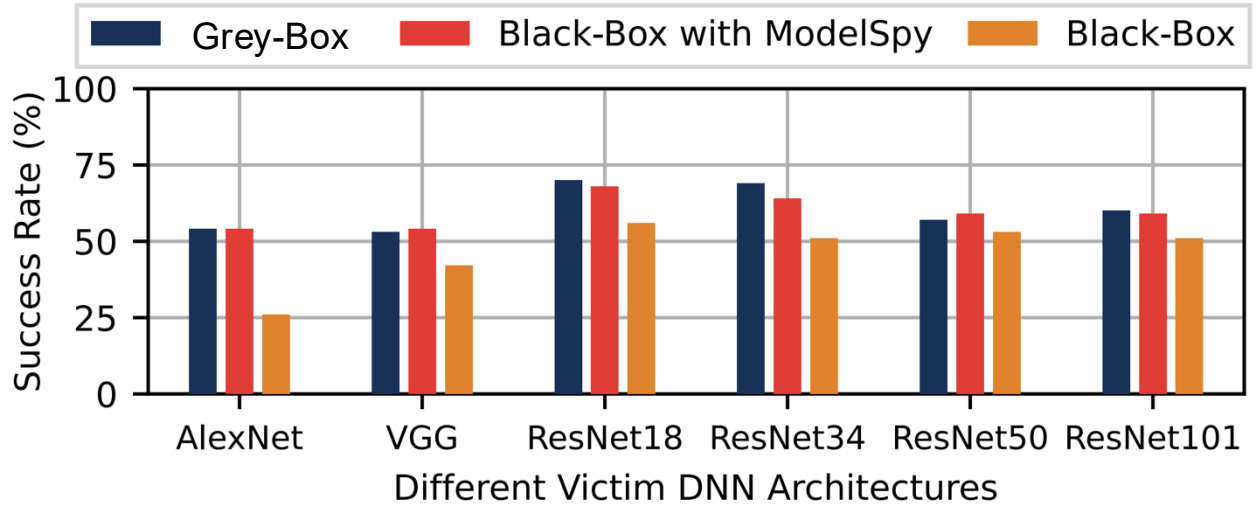
ModelSpy Performance Summary

1. Low layer sequence prediction error
2. High hyperparameter estimation accuracy = 94%



ModelSpy Performance Summary

1. Low layer sequence prediction error
2. High hyperparameter estimation accuracy = 94%
3. ModelSpy ***enhances success rate*** of ***adversarial attacks***
   - by 22.8%, matching grey-box attack performance

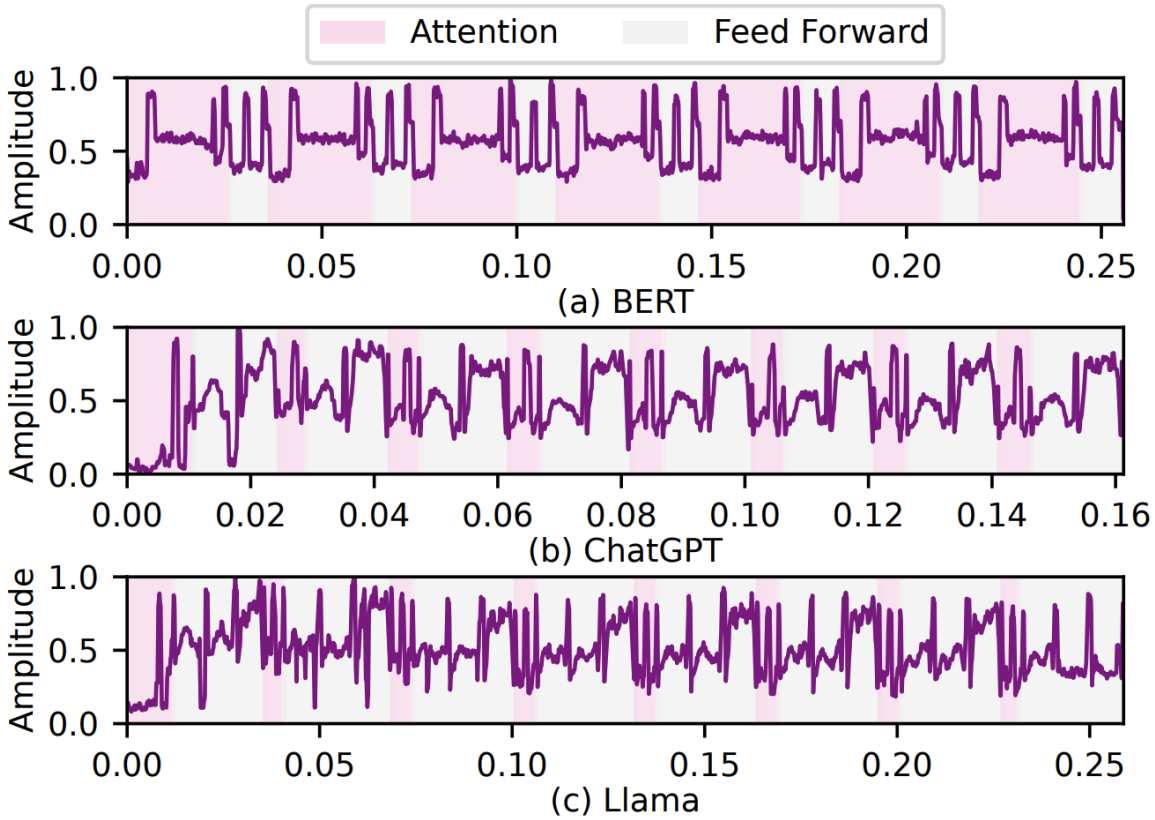


Summary of Evaluation Results

- **81.7%** accuracy at a distance of **5 meters**
- Works across **different walls**, including wood, concrete, glass
- Exhibits effective **cross-GPU reconstruction** within the same GPU generation

Discussion

- ➢ Countermeasures (EM jamming, Obfuscation)
- ➢ Attacking Large Language Models



Challenge in large-scale LLM inference Scenario

- **multi-GPU setups** with **pipeline parallelism**

Potential Solution: isolate EM signals from each GPU by incorporating multiple receiving antennas

Conclusion

- We introduce ***ModelSpy***, a long-range **DNN architecture snooping attack** utilizing the **GPU EM side channel**.
- We hope to highlight that securing advanced AI systems requires rethinking the **physical security of AI infrastructure**.