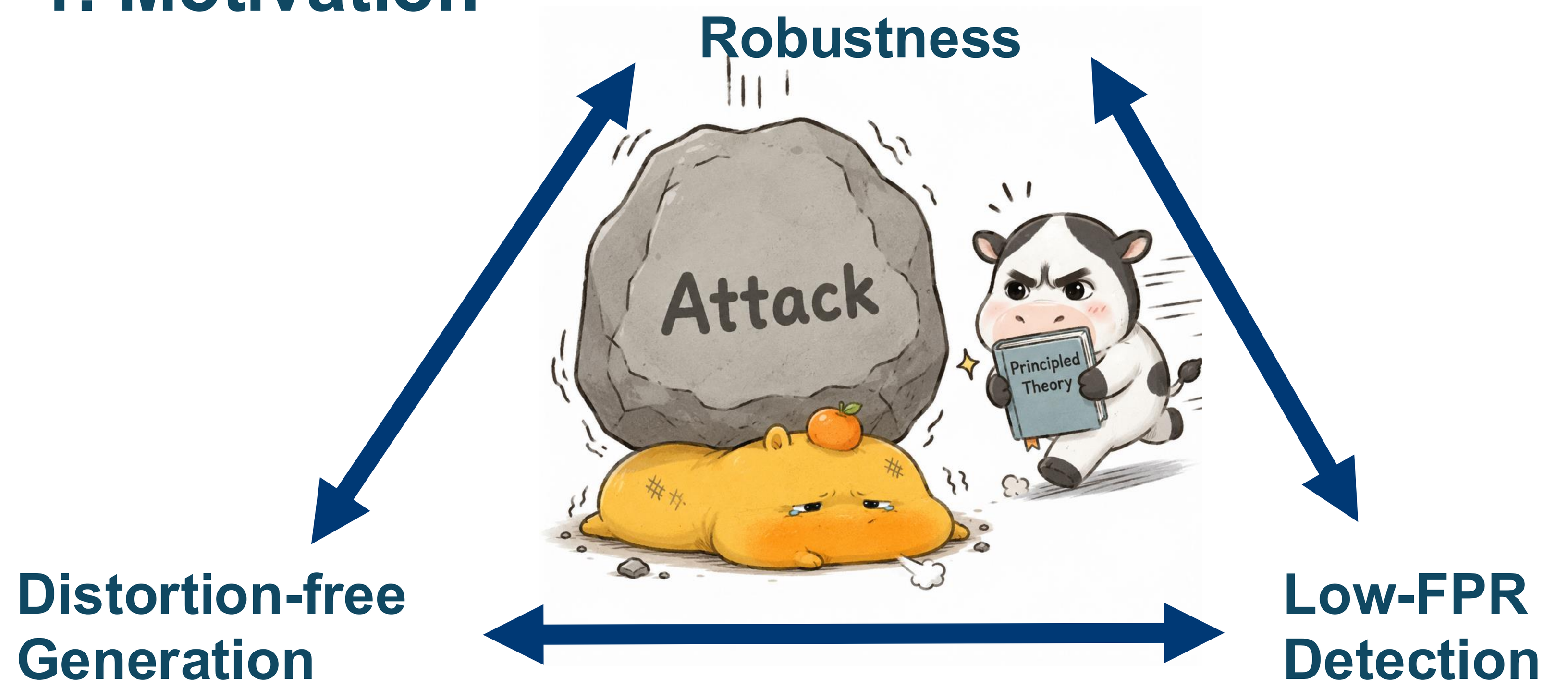
PASA: A Principled Embedding-Space Watermarking Approach for LLM-Generated Text under Semantic-Invariant Attacks

Zhenxin Ai[1], Haiyun He[1*] [1]The Hong Kong University of Science and Technology (Guangzhou)

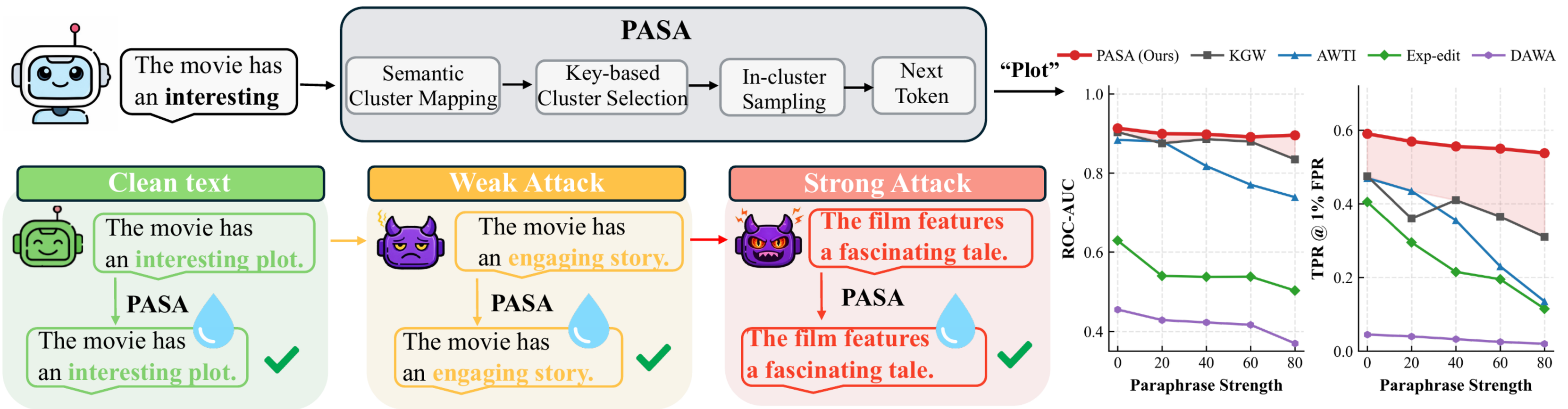
ICML International Conference On Machine Learning

Introduction

1. Motivation



2. Core Contributions



Overview of PASA and its robustness comparison under paraphrasing attacks.

1. **Semantic-Space Watermarking:** Watermarking over latent **semantic clusters**.
2. **Principled Trade-off:** Minimize **miss-detection** under **FA** and **distortion constraints**.

$$\min \beta_1^f \quad \text{s.t.} \quad \sup_Q \beta_0^f \le \alpha,\ D(P_{X^T}, Q_{X^T}) \le \epsilon$$

3. **Robust Low-FPR Detection:** **Stronger robustness** against **replacement** and **paraphrasing** attacks.

Method: PASA

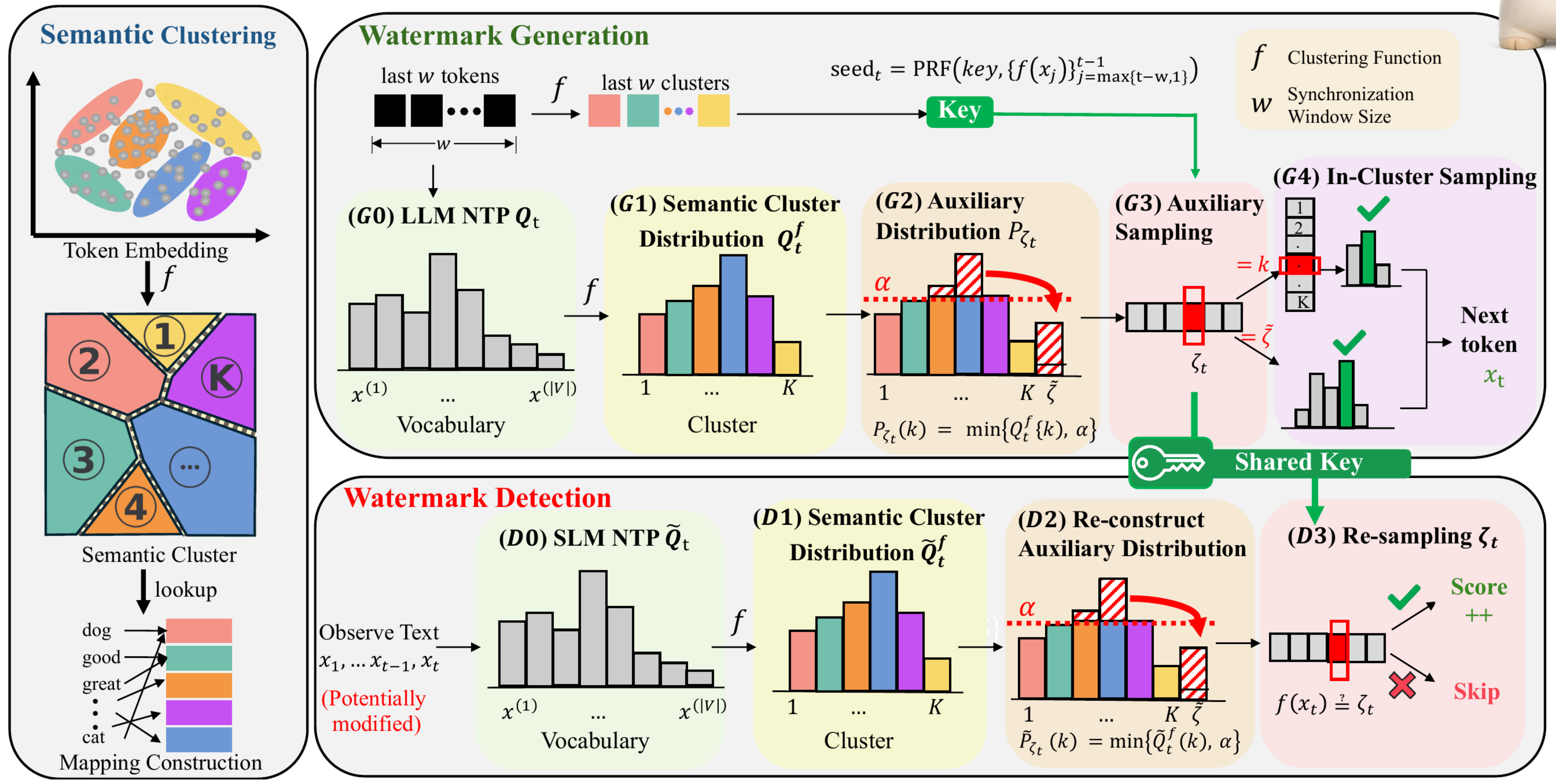


Illustration of the detailed PASA pipeline, including semantic cluster mapping, watermark generation, and watermark detection.

Semantic-Level Watermarking

Map tokens into semantic clusters to make watermark *signals stable under paraphrasing*.

$$f: \mathcal{V} \to [K],\ Q_t^f(k) = \sum_{x: f(x)=k} Q_t(x)$$

Watermarking Generation

Auxiliary variable is sampled using *a secret-key PRF*.

Next token is sampled inside the selected semantic cluster.

$$X_t \sim \frac{Q_t(x)\mathbf{1}\{f(x) = \zeta_t\}}{Q_t^f(\zeta_t)}$$

Watermark Detection

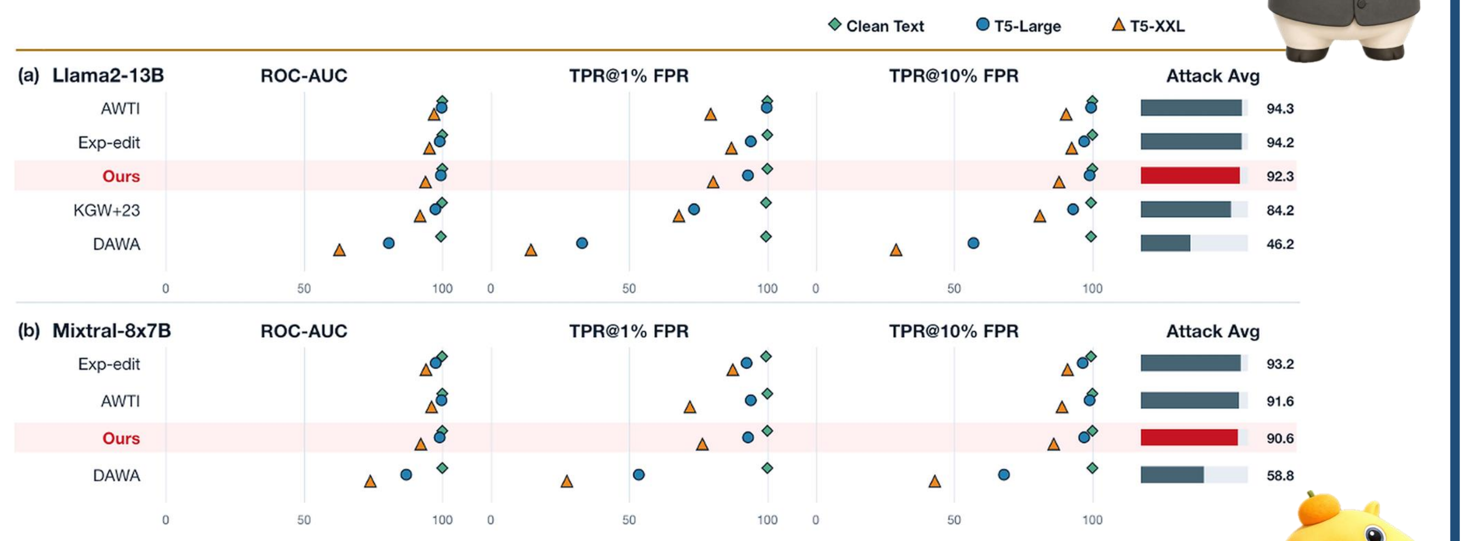
Secret-key replay verifies watermark evidence via semantic-cluster matches.

$$S = \sum_{t=1}^{T} \mathbf{1}[f(x_t) = \hat{\delta}_t],$$

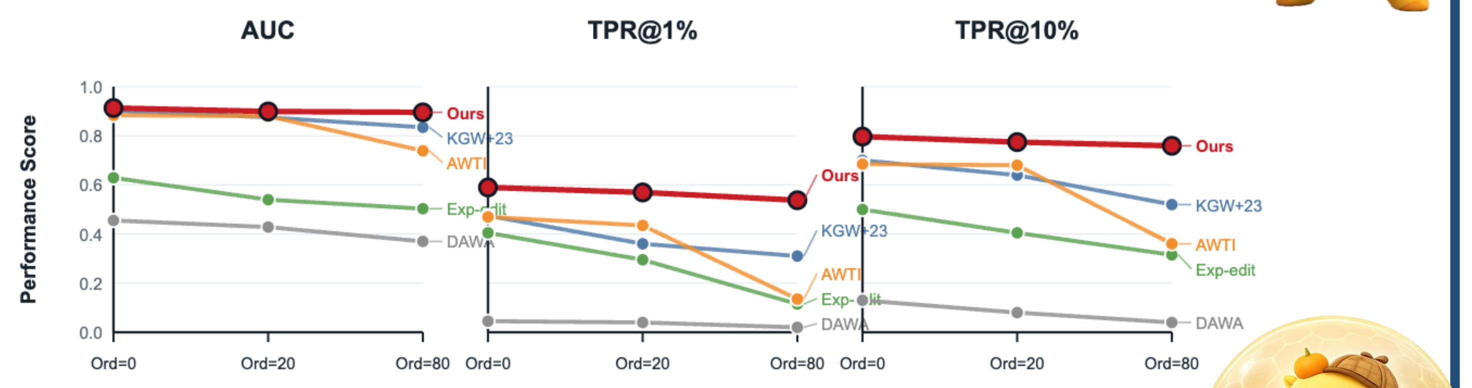
$$S \ge \tau \Rightarrow \text{Watermarked.}$$

Main Experiments

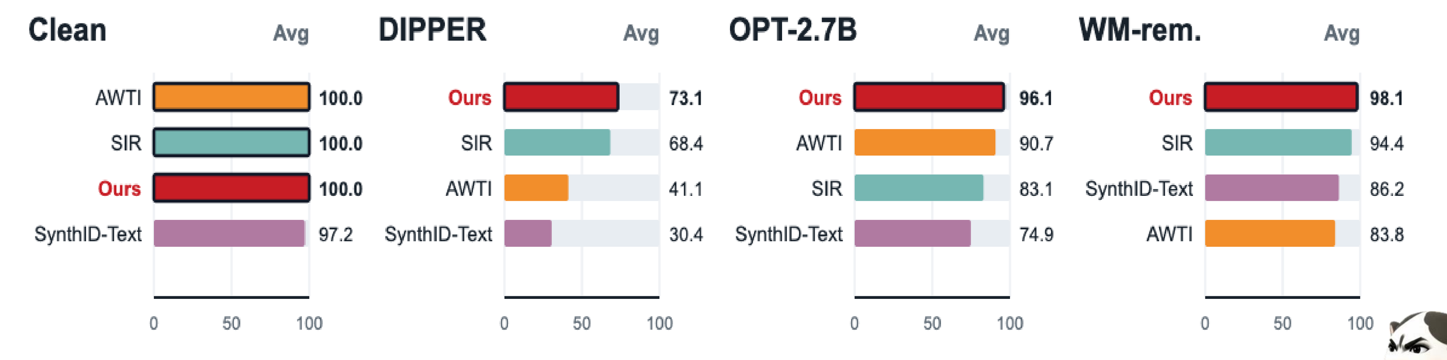
Detection Performance



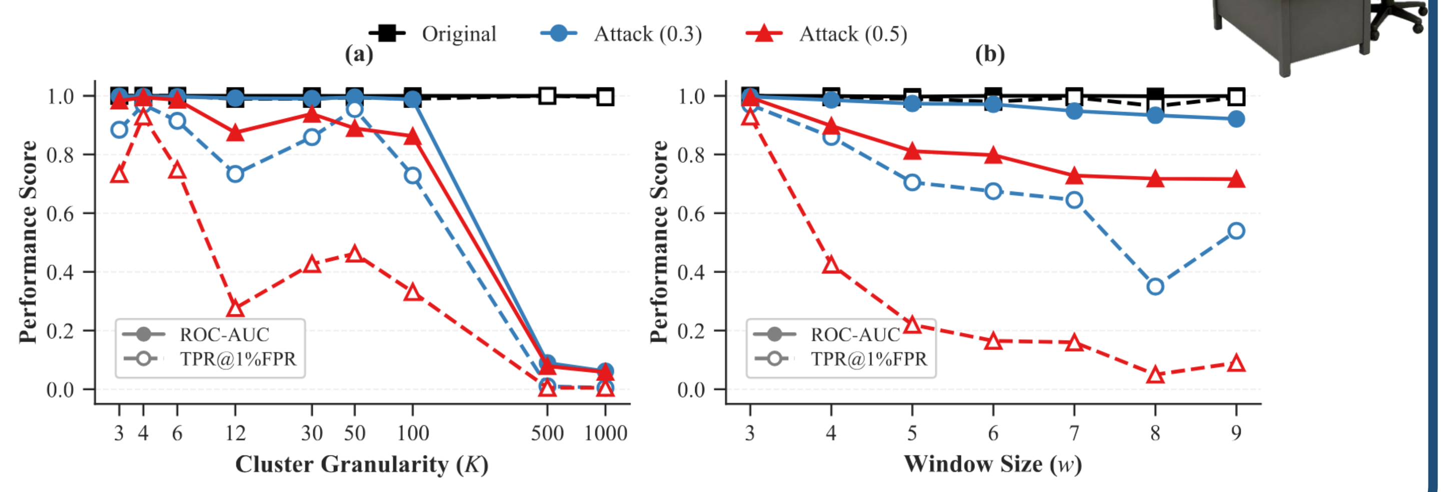
Robustness under Semantic-Invariant Attacks



More Attacks



Ablation Study



Email: zhenxinai@hkust-gz.edu.cn
Paper: https://arxiv.org/abs/2605.10977 Project Page: https://ai-kunkun.github.io/PASA_page/