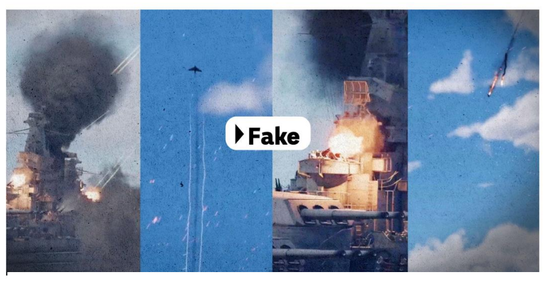
WRATH: Turning Watermark Robustness Against Itself via a Watermark-Agnostic Black-Box Invalidation Attack

Nan Jiang[1], Juan Hu[1], Bangjie Sun[1], Terence Sim[1], and Jun Han[2]
[1]*National University of Singapore*, [2]*KAIST*

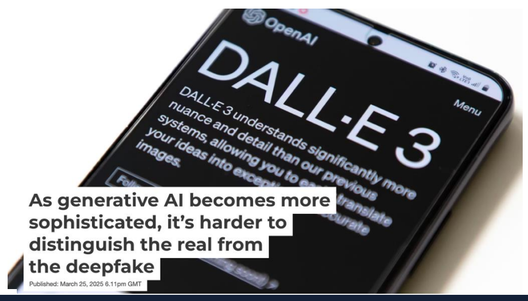
Scan to read the paper

1. Introduction

- **We are witnessing reported cases of generative AI misuse.**



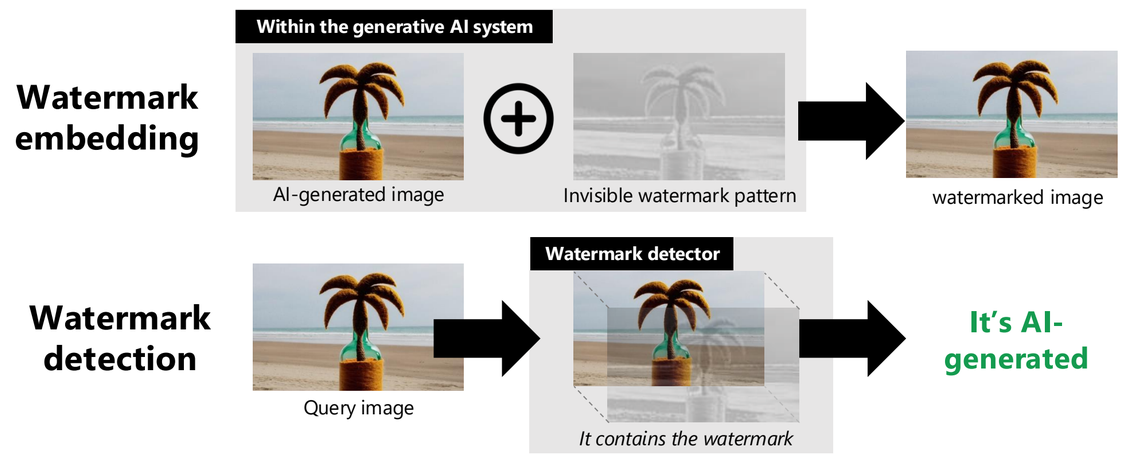
Real attack or video game? Misinformation and war in the AI age



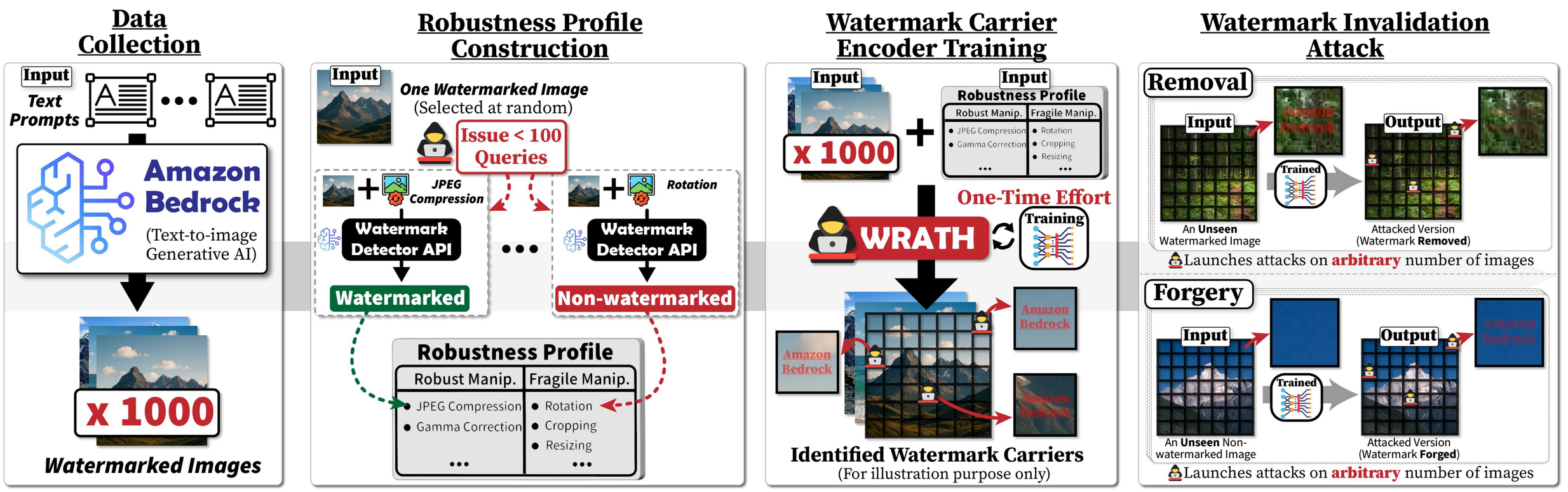
AI is intensifying a 'collapse' of trust online, experts say

From Venezuela to Minneapolis, the rapid rollout of deepfakes around major news events is stirring confusion and suspicion about real news.

- **Providers embed watermark in generated images.**

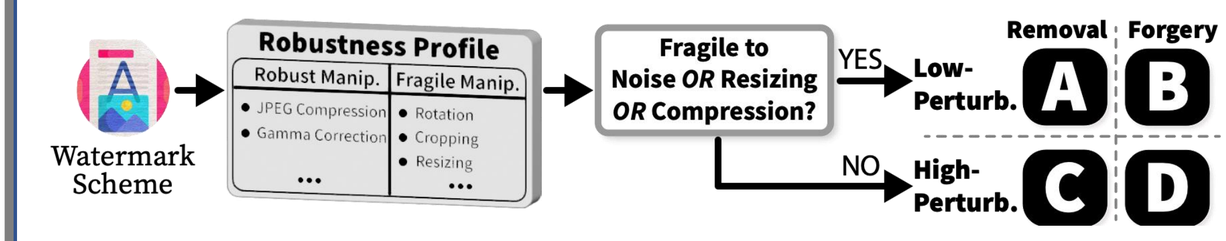


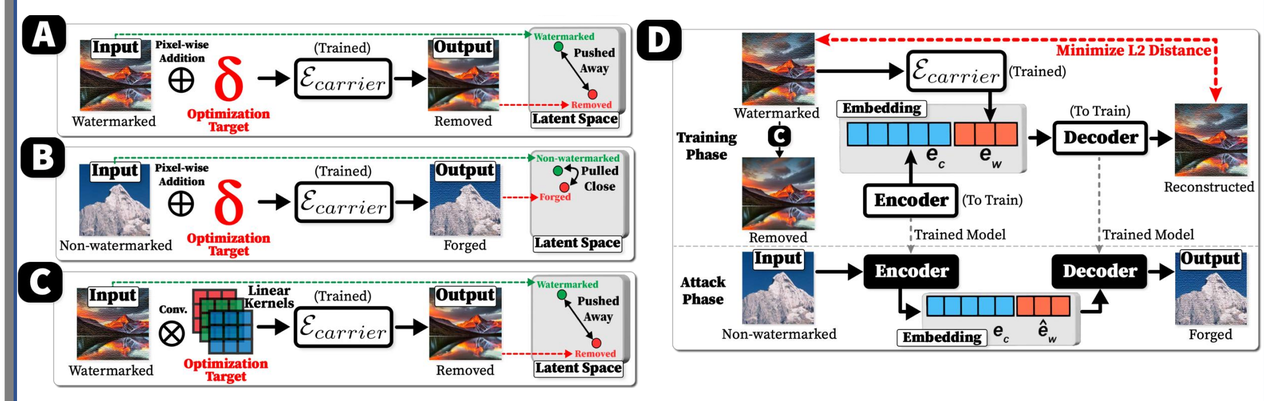
2. Overview of *WRATH*



3. *WRATH*: Watermark Invalidation

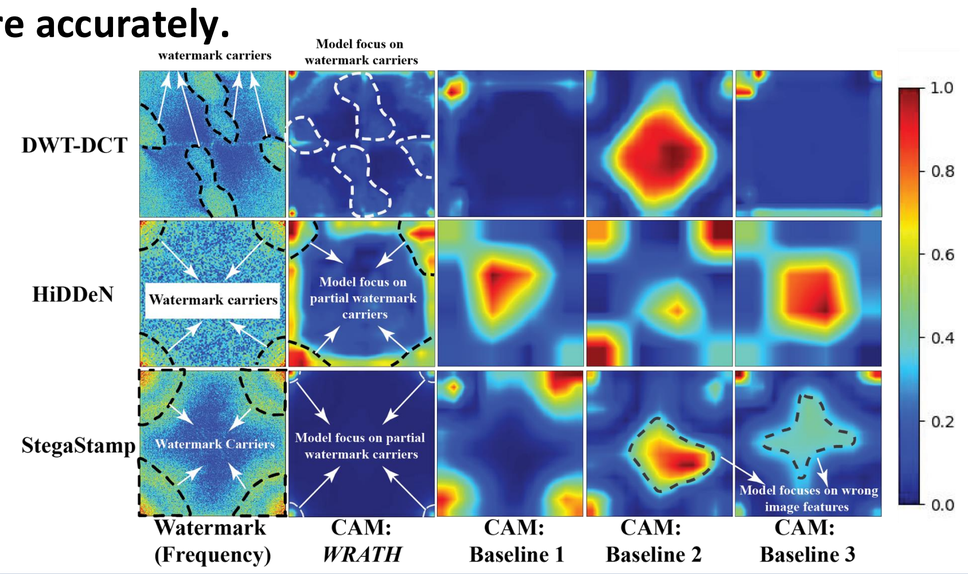
- ***WRATH* removes or forges watermark based on the robustness characteristics.**





4. Evaluation#1: Watermark Identification

- **Compared with baselines, *WRATH* identifies watermark carriers more accurately.**



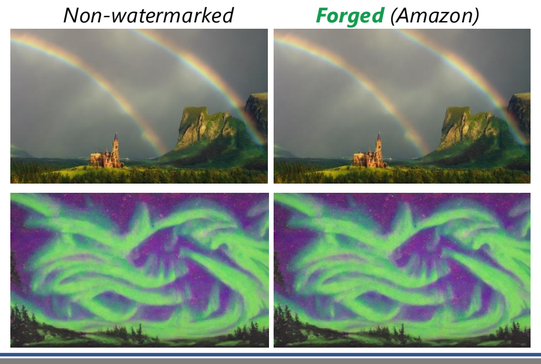
4. Evaluation#2: Removal and Forgery

- **Watermark removal**
  - Success rate up to **95%**, LPIPS<**0.1**

Watermarked (Amazon) | ***Removed***

- **Watermark forgery**
  - Success rate up to **100%**, LPIPS<**0.1**

Non-watermarked | ***Forged*** *(Amazon)*



5. Root Causes and Countermeasures

- **Root cause #1**
  - Consistent watermark carrier
  - Jaccard similarity: Intra-scheme (**0.52**), inter-scheme (**0.16**)
- **Root cause #2**
  - Consistent watermark signals
  - Cosine similarity: Intra-scheme (**0.91**), inter-scheme (**0.55**)
- **Counter measure #1**
  - Avoid consistent carriers
  - Content dependent/randomized
- **Counter measure #2**
  - Avoid consistent signals
  - Key-dependent signals

Carriers in frequency domain