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

*__Nan Jiang__, *Juan Hu, *Bangjie Sun, *Terence Sim, ✢Jun Han

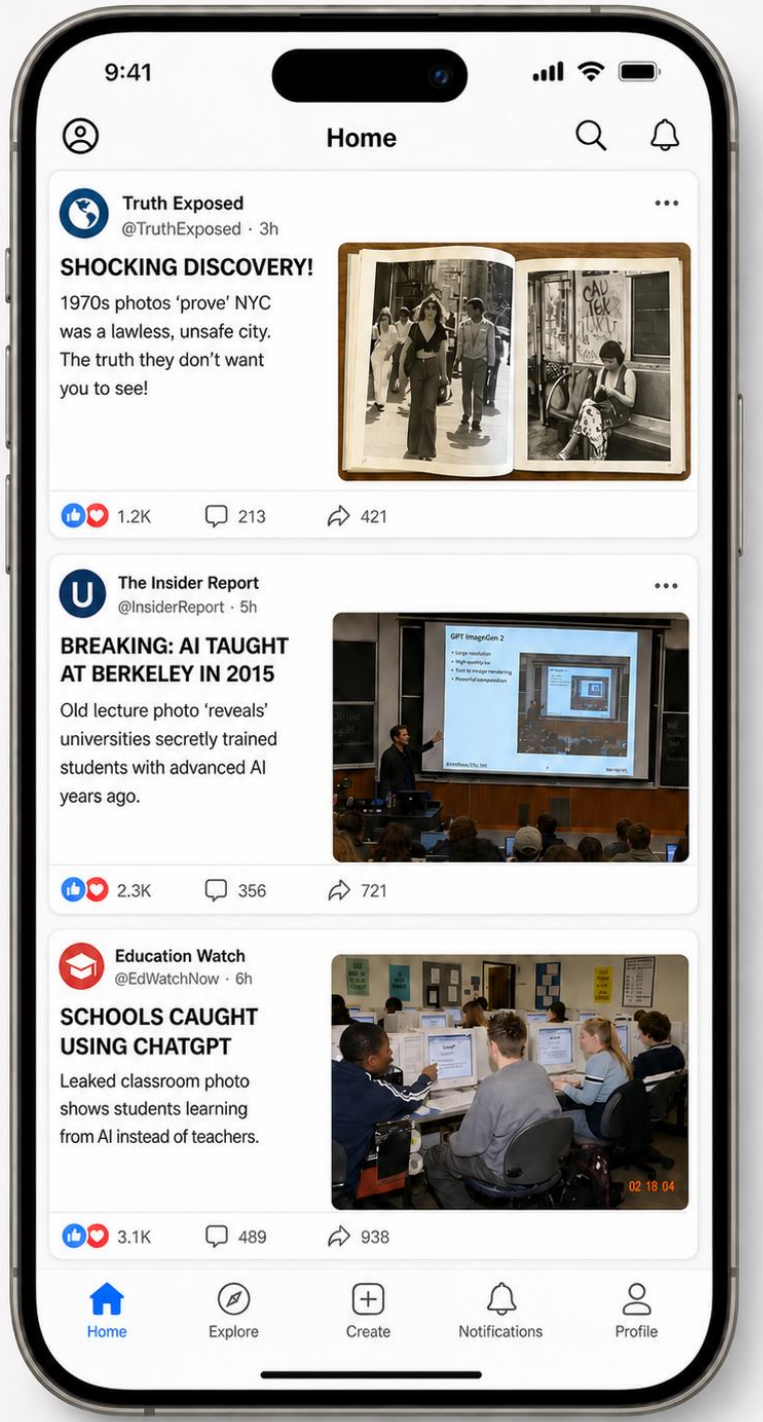
*National University of Singapore, ✢KAIST

Contact: jiangnan@u.nus.edu


YOU CAN DO HARD THINGS
BE KIND ONLINE
THINK BEFORE YOU CLICK
KEYBOARD SHORTCUTS
ChatGPT
ChatGPT
ChatGPT
02 18 04

Photo-realistic AI-generated Images Raise Concerns

- Spread misleading information on social media
- Fabricate evidence of events that never happened
- **Undermine trust in online visual content**



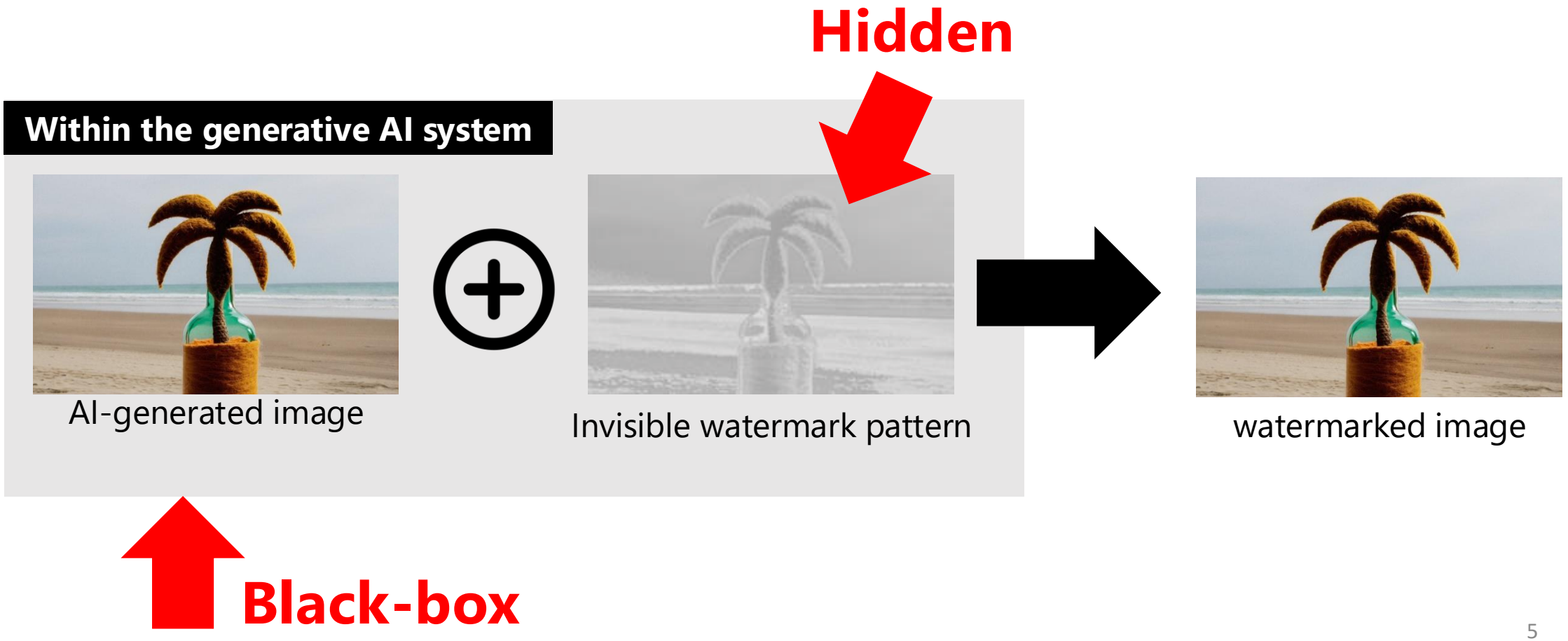
All images are generated by ChatGPT Images 2.0

Watermarking as a Trust Mechanism

- Many providers already use watermarks to identify AI-generated images

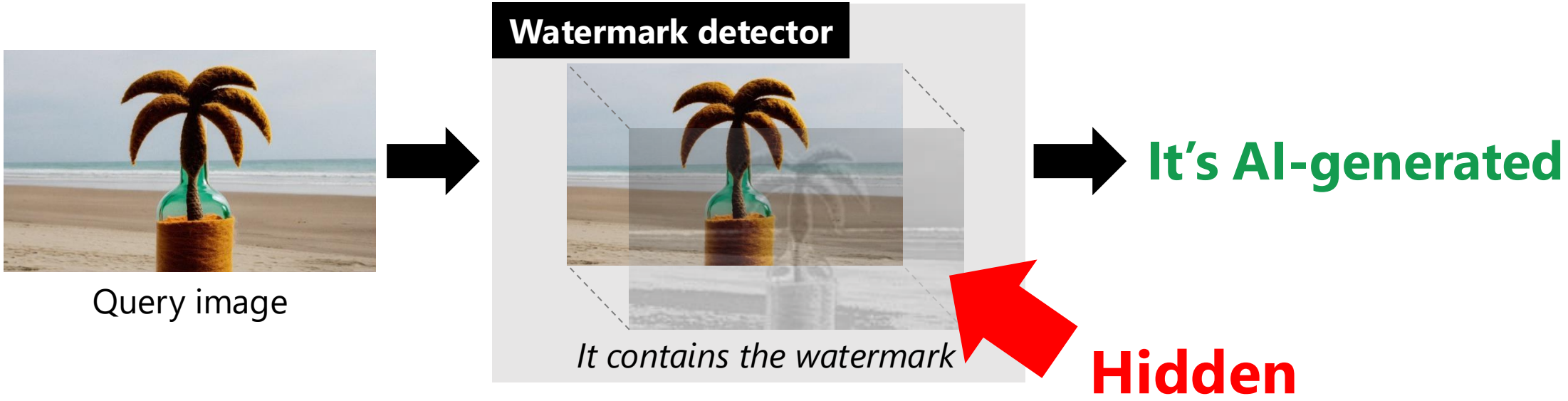
Background: Watermarking in GenAI

- **Generation:** **Invisible** watermarks are embedded into AI-generated images



Background: Watermarking in GenAI

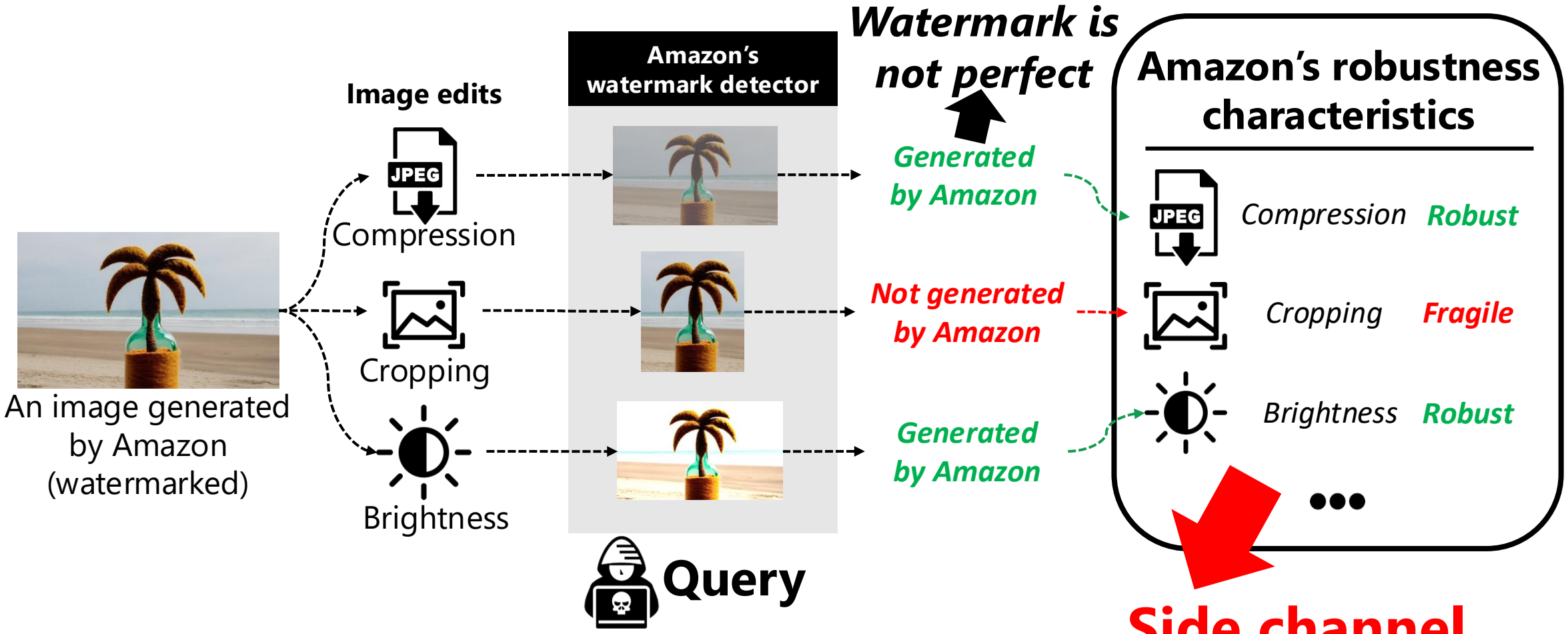
- **Generation: Invisible** watermarks are embedded into AI-generated images
- **Detection**: Images are deemed as AI-generated if **detected** with a watermark



Does it really reveal nothing about the watermark pattern?

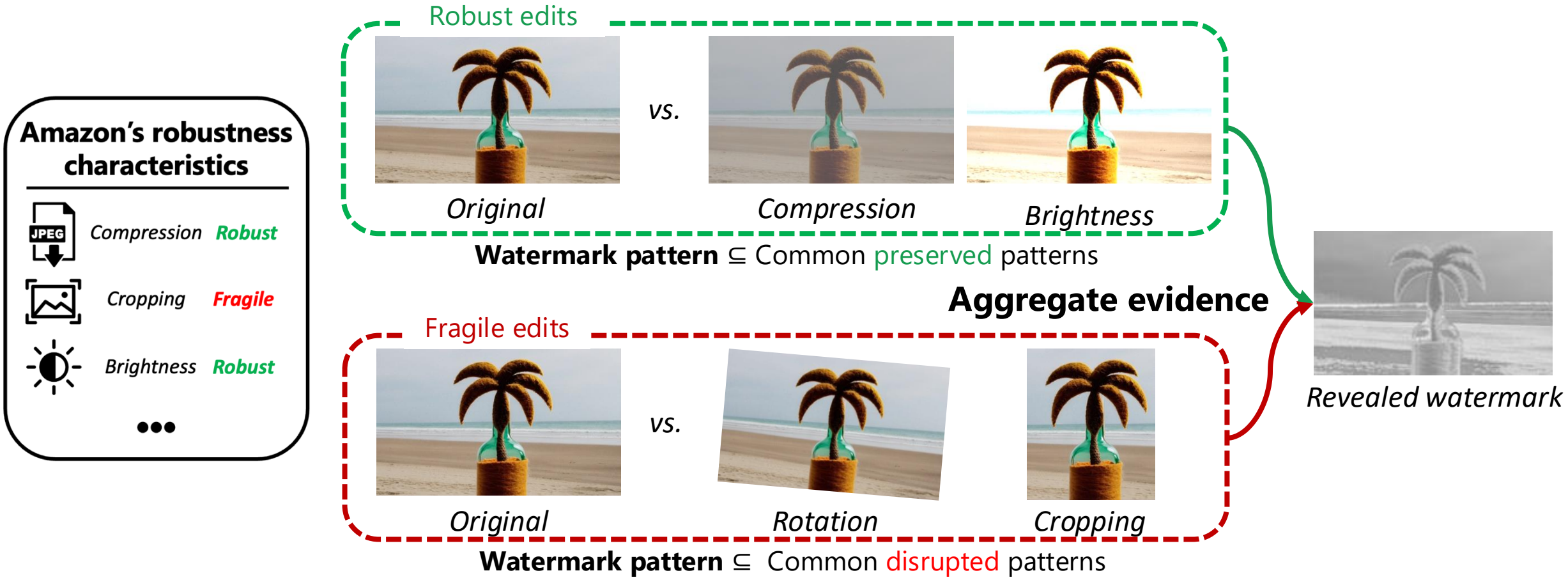
Our Work: Watermark Robustness as Side Channel

- Watermark's **robustness characteristics** can reveal **watermark pattern**



How Does Robustness Reveal Watermark?

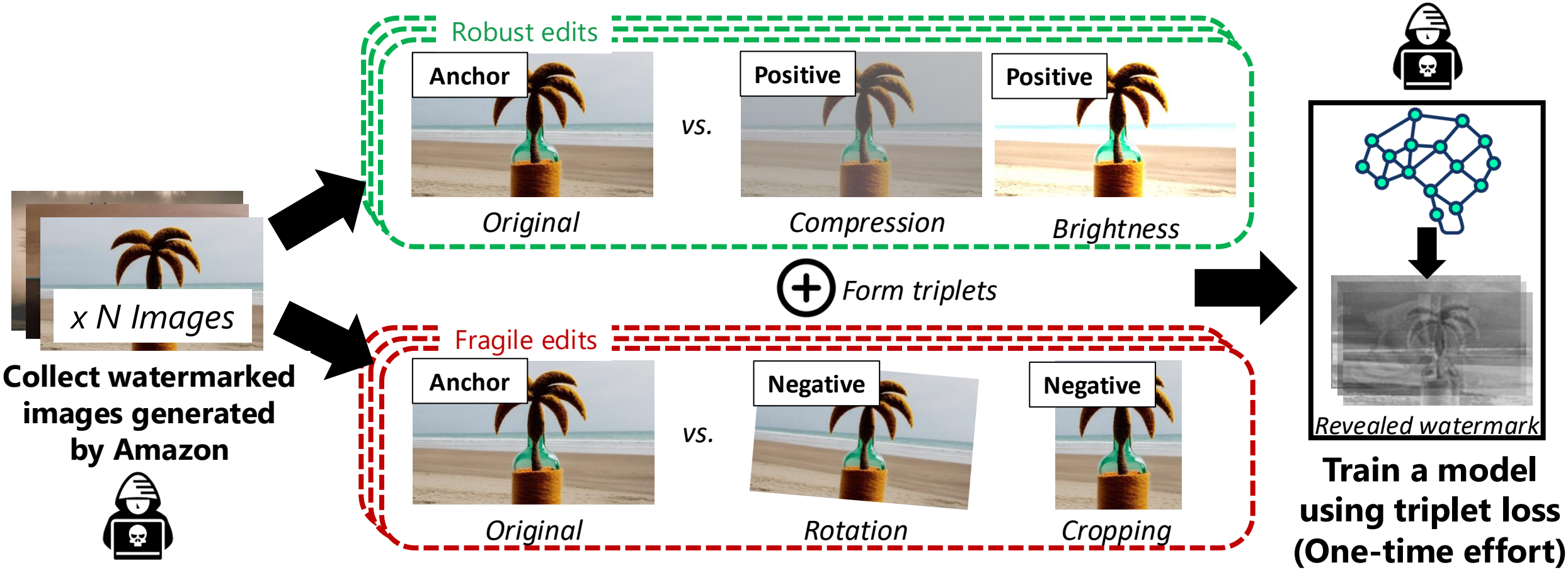
- Watermark patterns are **preserved** by robust edits but **disrupted** by fragile edits



This pipeline is for illustration purposes only. WRATH reveals watermark in the latent feature space.

How Does Robustness Reveal Watermark?

- We aggregate evidence across images



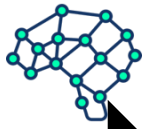
This pipeline is for illustration purposes only. WRATH reveals watermark in the latent feature space.

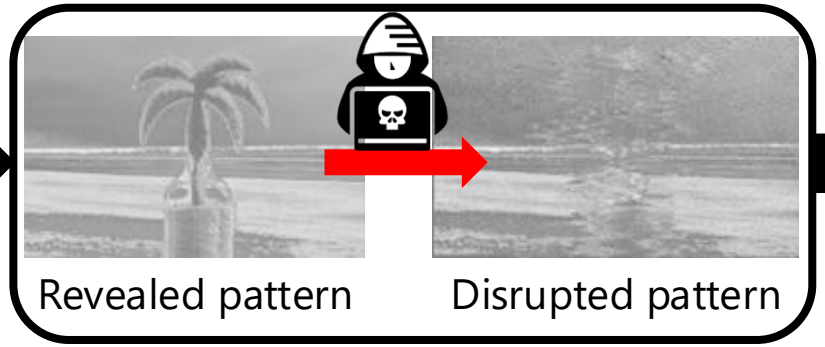
Revealed Watermark Underpins Digital Trust

- Revealed watermark pattern enables **removal** and **forgery** attacks

Watermark removal

AI-generated image
(watermarked)



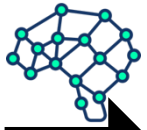


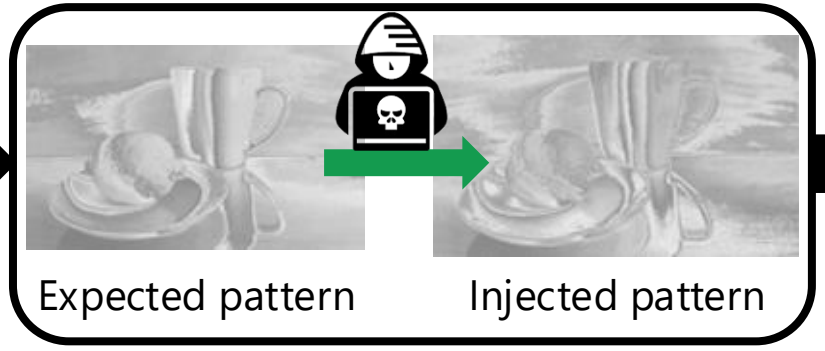
Deemed as real image
(non-watermarked)

Not recognizable by the detector

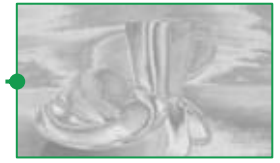
Watermark forgery

Real image
(non-watermarked)





Deemed as AI-generated
(watermarked)



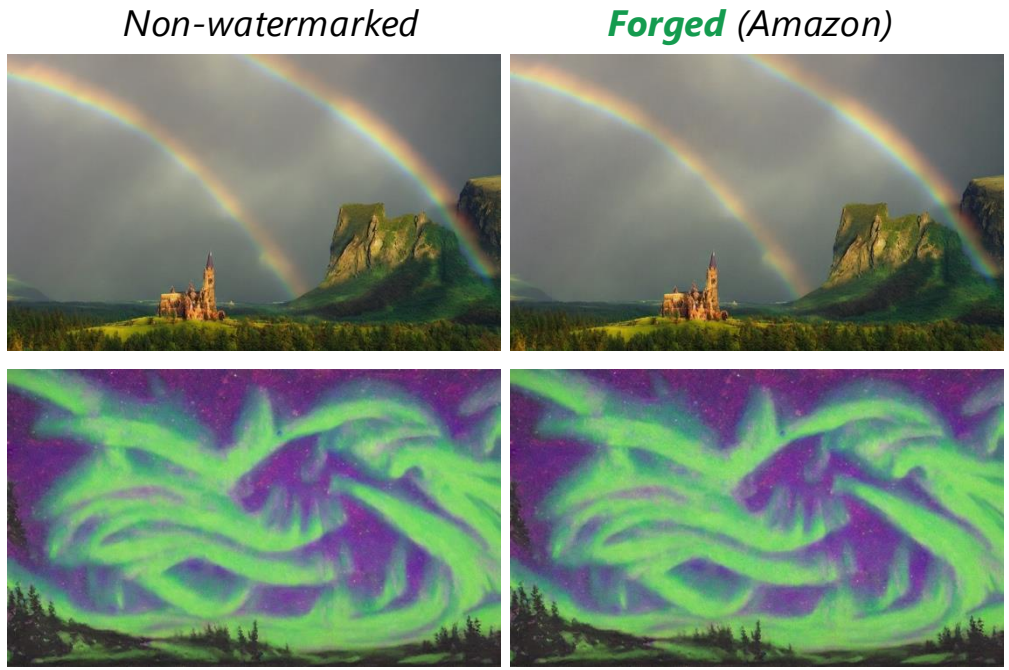
Recognizable by the detector

Evaluation: Performance on Removal and Forgery

- **Setup:**
  - **Six** representative academic works and **one** real-world system (Amazon)
  - **< 100** queries to the detector; **1000** watermarked images for training

Evaluation: Performance on Removal and Forgery

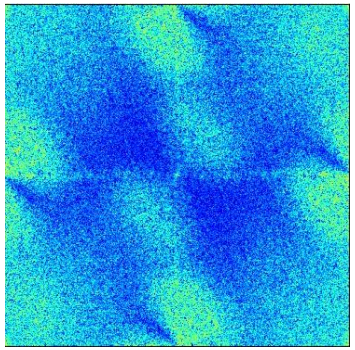
- **Results:**
  - **Removal:** Up to **95%** success rate while preserving perceptual quality **(LPIPS < 0.1)**
  - **Forgery:** Up to **100%** success rate while preserving perceptual quality **(LPIPS < 0.1)**



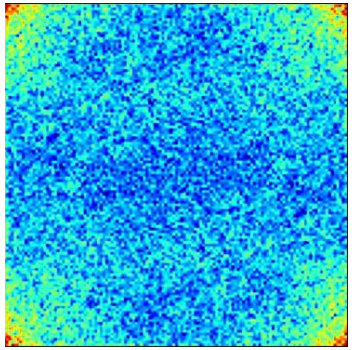
Root Cause and Countermeasure

- **Root cause**: **consistent** watermark patterns across images

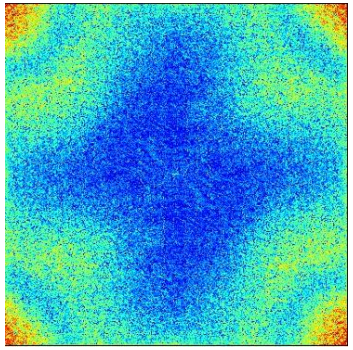
Examples in frequency domain



DWT-DCT

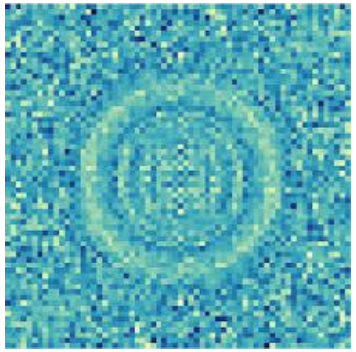


HiDDeN



StegaStamp

Example in latent domain



Tree-Ring

Allows us to train a model to reveal common watermark pattern

- **Countermeasure**: **randomize** watermark patterns using **cryptographic keys**

Conclusion

- Robustness enables watermark **revelation**, **removal**, and **forgery**
- Future watermark should **avoid consistent** watermark pattern