Verified at Birth: Content Verification Layer for Authentic AI Generation

Zehao Chen

Supervised by Jiaojiao Jiang and Bo Liu

School of Computer Science and Engineering, UNSW
School of Computer Science, UTS

Developed with PyTorch

Background & Motivation: "Why Verification at Birth?"

With the rapid growth of modern computing, generating highly realistic digital content like images and text has become possible. This development raises a critical challenge: how can we confirm if a piece of content comes from a natural source or was computer-generated? Reliable methods of origin verification are therefore essential for trust and the responsible use of these technologies.

The Limits of Traditional Detection: Reactive Approach

Most current misinformation detection methods (e.g., fact-checking) are *reactive*. They typically analyze suspicious content *after* it has been published and has already begun to spread.

Too Slow: Relies on manual analysis or complex algorithms that cannot keep up with the speed of viral information.

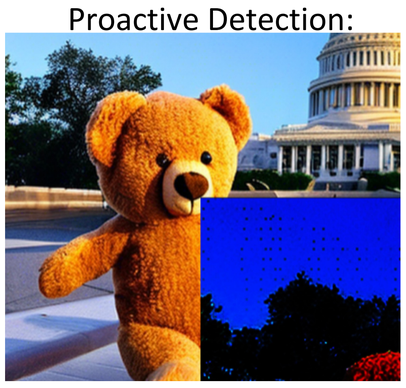
Easily Evaded: Simple edits, compression, or re-contextualizing (e.g., pairing a real photo with false text) can bypass detection.

Damage is Done: By the time content is debunked, it may have already caused widespread negative impact.

Reactive/Passive Detection:

"Looking for clues"

Proactive Detection:



"Injecting clues"

Figure 1: Proactive Vs Passive Detection

Steven Spielberg a longtime friend of Lucas and his wife Kate Capshaw waved to the crowds.

(a) Original News

Steven Spielberg **directed Lord Attenborough** in **1993**'s **Jurassic Park**.

(b) Manipulated News

Figure 2: Misinformation Example

Method: Multi-Modal Watermarking

We are developing a robust multi-modal watermarking system specifically for image-text pairs, which are common in news and social media contexts. Our system is designed to securely link an image to the "semantic fingerprint" of its original accompanying text. Our post-processing approach is divided into an "Embedding" stage and a "Verification" stage:

1. Embedding Stage (Content Creator Side)

- **Extract Semantics:** A pre-trained model extracts core semantic information from the text, which is transformed into text embeddings.
- **Generate Bitstream:** A specially designed encoder converts these embeddings into a compact, error-resilient bitstream.
- **Deep Embedding:** Using a deep self-supervised learning model, this bitstream is invisibly embedded into the corresponding image.

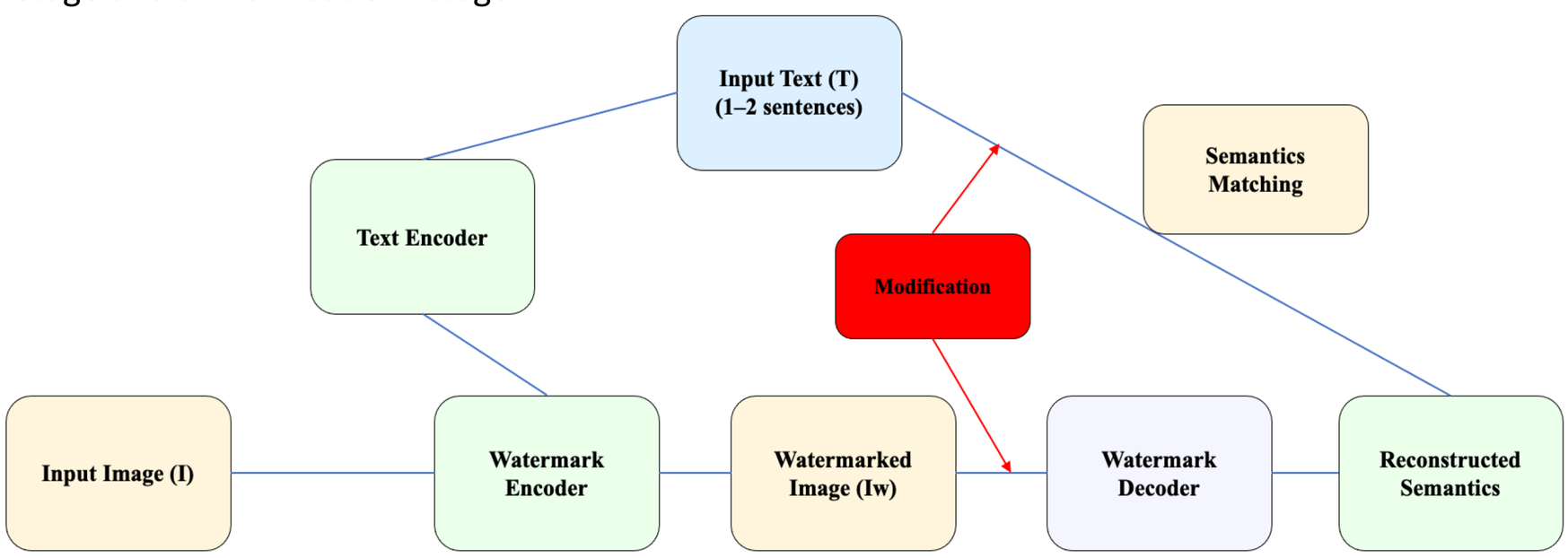


Figure 3: Proposed Pipeline of Multi-Modal Watermarking System

2. Verification Stage

- **Blind Extraction**: Even after the image has been altered (e.g., compressed, cropped), the system can "blindly" extract the embedded semantic information.
- **Semantic Comparison**: The system compares this extracted "original semantic information" with the content's *current* context (e.g., the new accompanying text).
- **Detect Manipulation**: If the semantics do not match, the system can reliably detect manipulation and verify the integrity of the original message.

Technical Innovations and Research Significance

(*Note: As the project is in its early stages, this section focuses on goals and innovative concepts.*)

Core Technical Innovations:

- Multi-modal Semantic Link**:** The primary innovation is embedding the text's *core semantic information*, not just a simple ID. This creates a secure link between the *meaning* of the text and the image itself.
- High Resilience: The system is being designed to be resilient, withstanding both common media transformations and adversarial attacks.

Research Significance & Expected Outcome:

- Empowering Professionals**:** This provides a crucial tool for journalists, fact-checkers, and digital platforms.
- Proactive Authentication**:** It allows them to proactively authenticate multi-modal content.
- Combating AI Weaponization**:** This work helps fortify defenses against the weaponization of AI in misinformation campaigns.

References

[1] Mahbuba Begum and Mohammad Shorif Uddin. Digital image watermarking techniques: a review. Information, 11(2):110, 2020.

[2] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bj¨orn Om- mer. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 10684–10695, 2022.

[3] Diederik P Kingma, Max Welling, et al. Auto-encoding variational bayes, 2013.

[4] Yuxin Wen, John Kirchenbauer, Jonas Geiping, and Tom Goldstein. Tree-ring water- marks: Fingerprints for diffusion images that are invisible and robust. arXiv preprint arXiv:2305.20030, 2023.