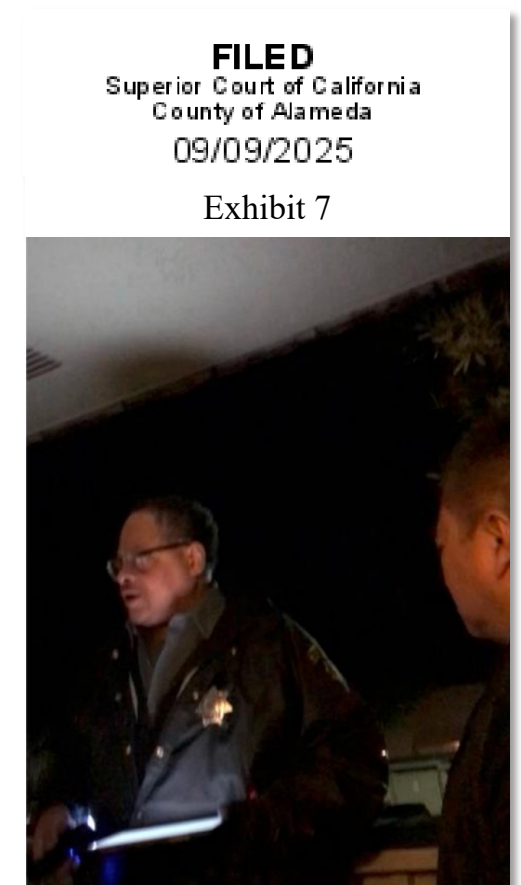
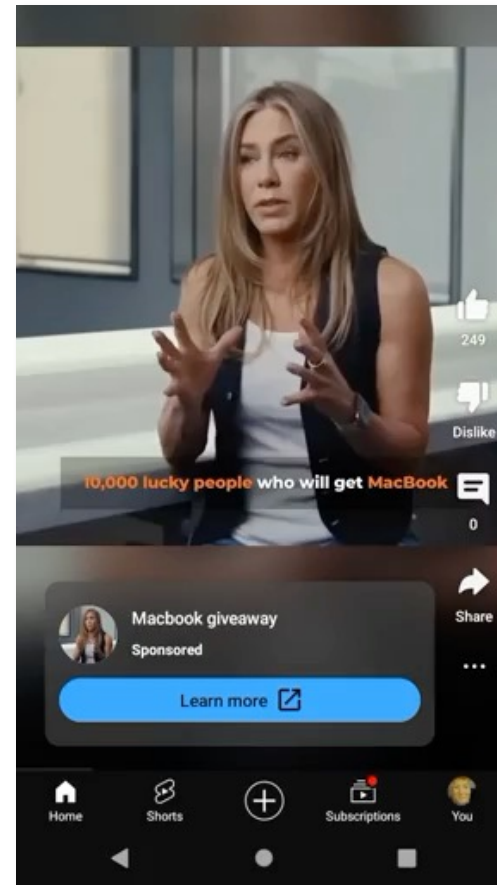
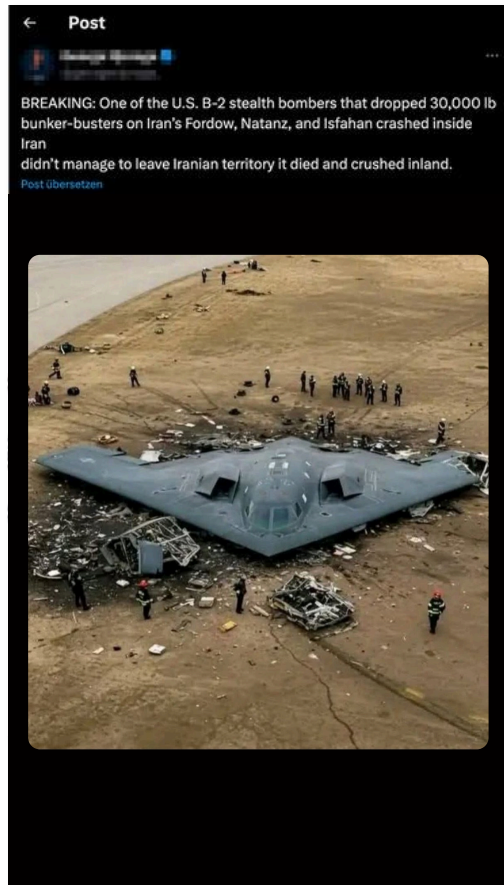


# Proactive Provenance for Audiovisual Media

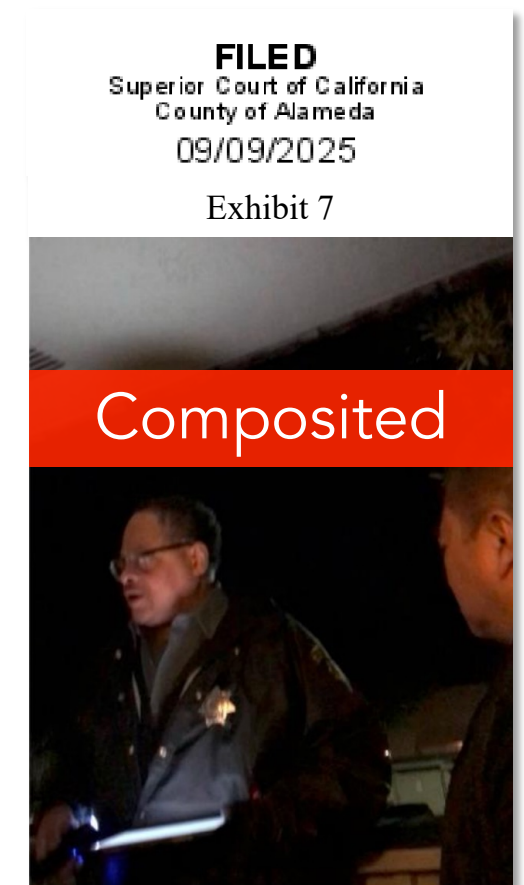
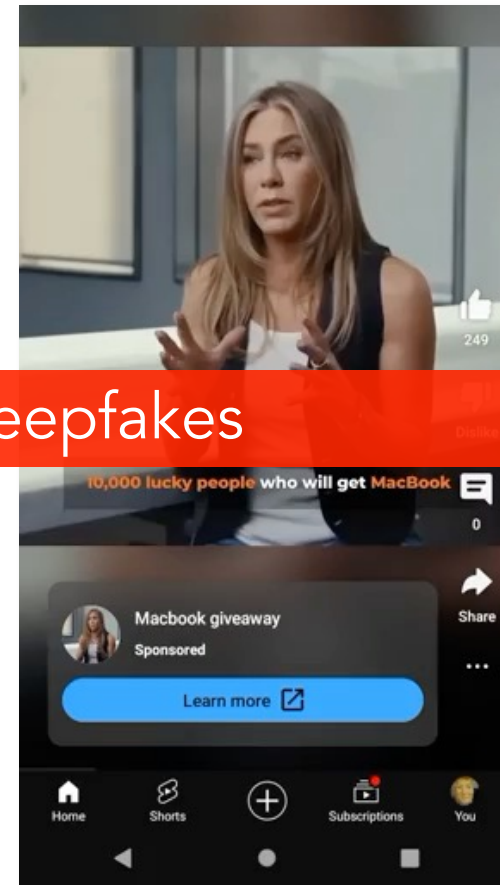
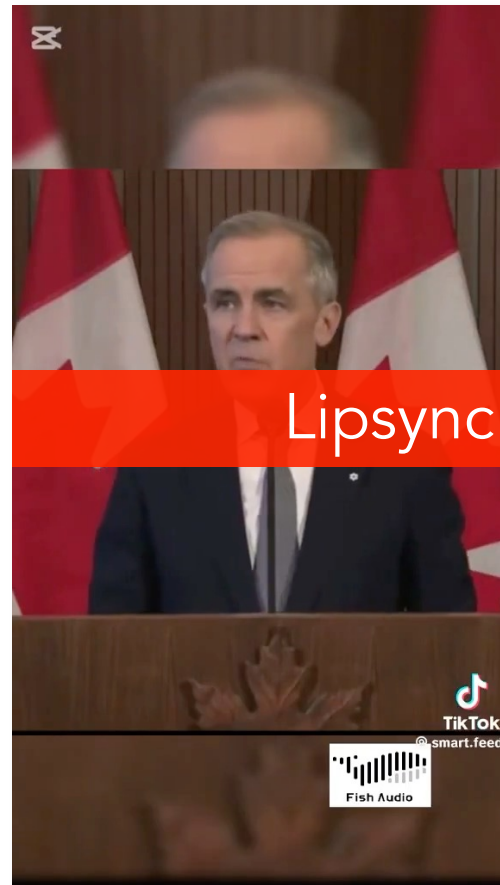
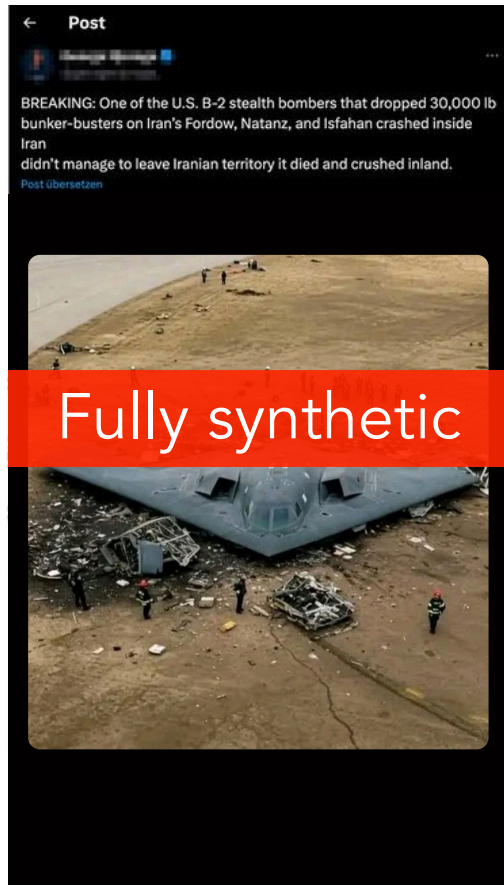
Candidacy Exam  
Hadleigh Schwartz  
August 20, 2026



# Importance of Audiovisual Media

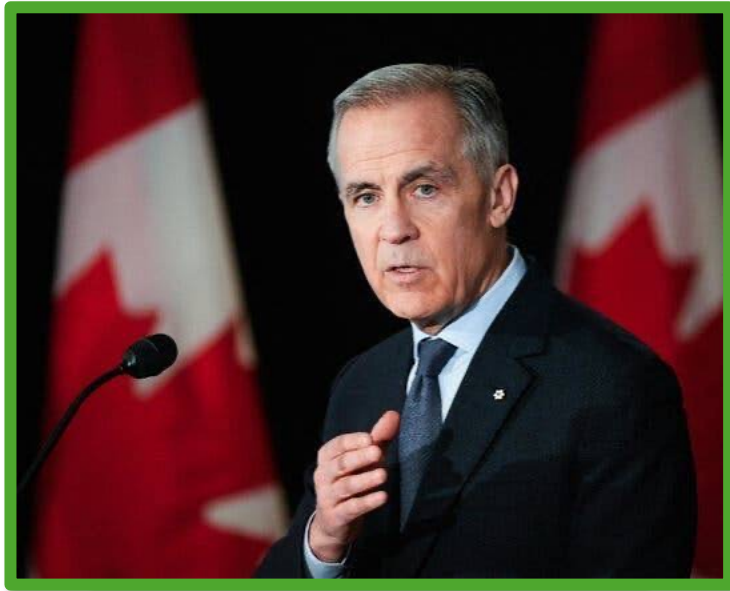


# Importance of Audiovisual Media



# Media Provenance

Information on a media asset's origins and modifications.



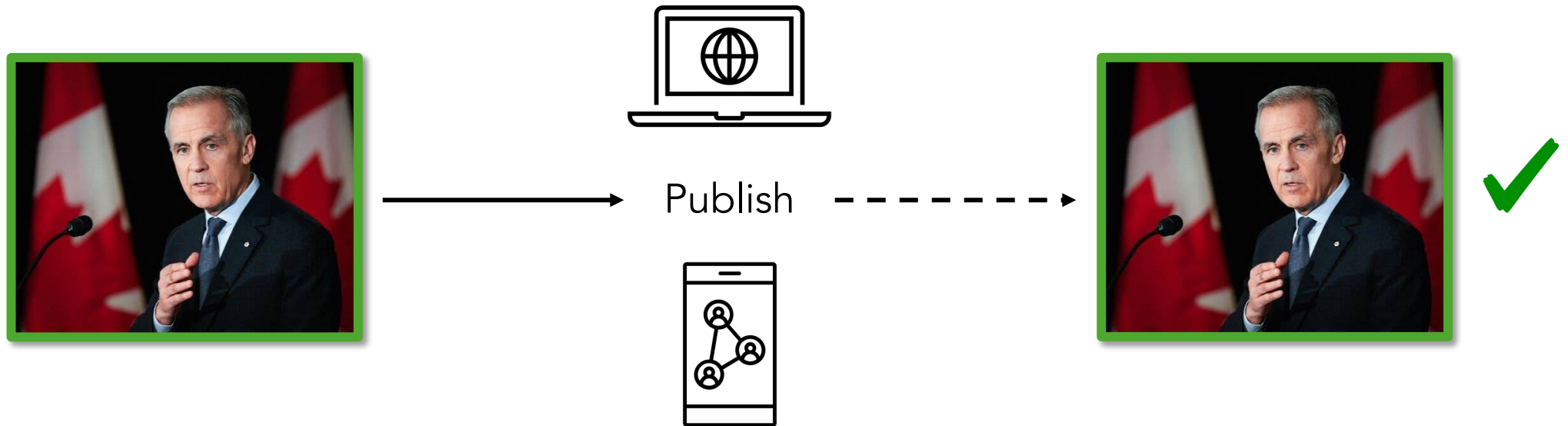
Created by: Nikon Camera  
Edits: Cropped



Created by: Grok Imagine

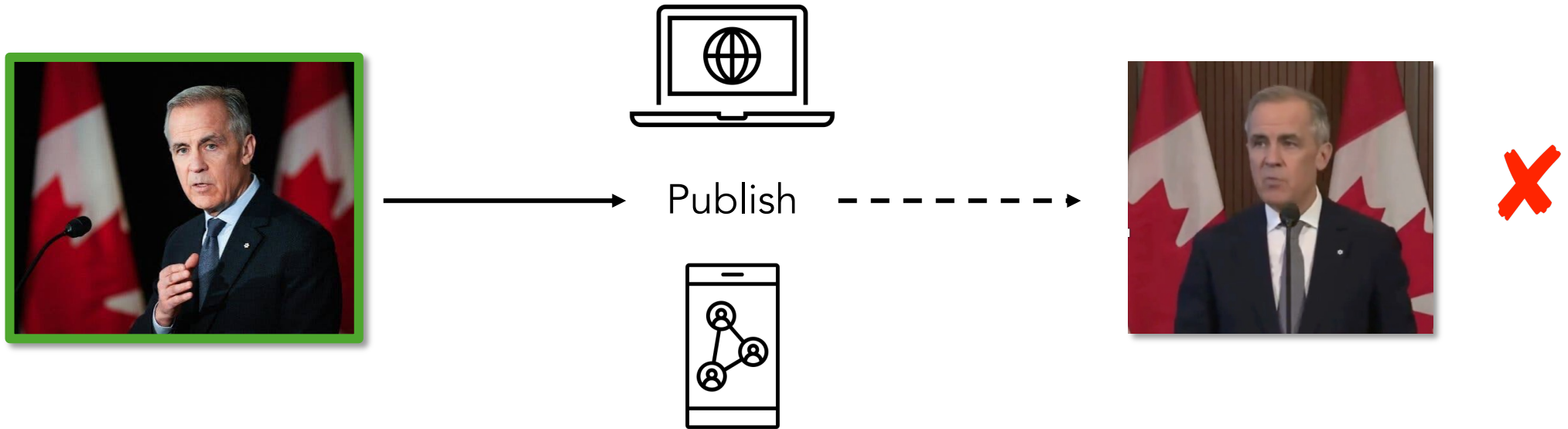
# Proactive Media Provenance

Ensure media contains provenance information before it's disseminated.  
Use that information downstream to assess authenticity and integrity.



# Proactive Media Provenance

Ensure media contains provenance information before it's disseminated.  
Use that information downstream to assess authenticity and integrity.



# Research in Proactive Media Provenance

How do we add and verify media provenance information?

## Digital Signatures

[England et al., 2021]

[Liu et al., 2022]

[Liu et al., 2024]

[Park et al., 2025]

[Liu et al., 2025]

## Digital Watermarks

[Zhu et al., 2018]

[Tancik et al., 2020]

[Chen et al., 2023]

[Liu et al., 2023]

[Wen et al., 2023]

[An et al., 2024]

[Liu et al., 2024b]

[Zhao et al., 2024]

[Agarwal et al., 2025]

[Gowal et al., 2025]

[Gunn et al., 2025]

[O'Reilly et al., 2025]

[Soucek et al., 2025]

[San Roman et al., 2025]

## Physical Signals

[Shahid et al., 2023]

[Michael et al., 2025]

[Qing et al., 2026]

# Research in Proactive Media Provenance

How do we add and verify media provenance information?

## Digital Signatures

[England et al., 2021]

[Liu et al., 2022]

[Liu et al., 2024]

[Park et al., 2025]

[Liu et al., 2025]

## Digital Watermarks

[Zhu et al., 2018]

[Tancik et al., 2020]

[Chen et al., 2023]

[Liu et al., 2023]

[Wen et al., 2023]

[An et al., 2024]

[Liu et al., 2024b]

[Zhao et al., 2024]

[Agarwal et al., 2025]

[Gowal et al., 2025]

[Gunn et al., 2025]

[O'Reilly et al., 2025]

[Soucek et al., 2025]

[San Roman et al., 2025]

## Physical Signals

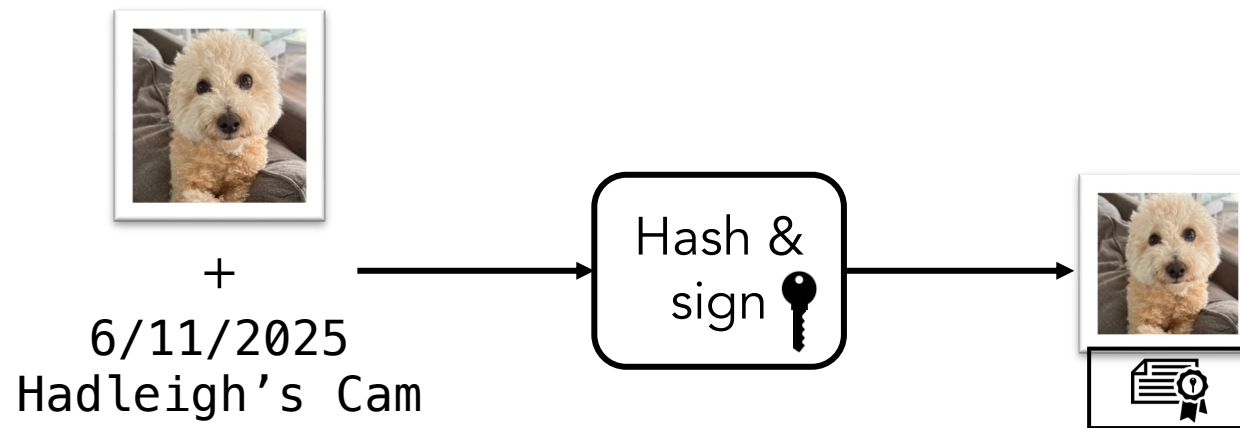
[Shahid et al., 2023]

[Michael et al., 2025]

[Qing et al., 2026]

# Digital Signatures

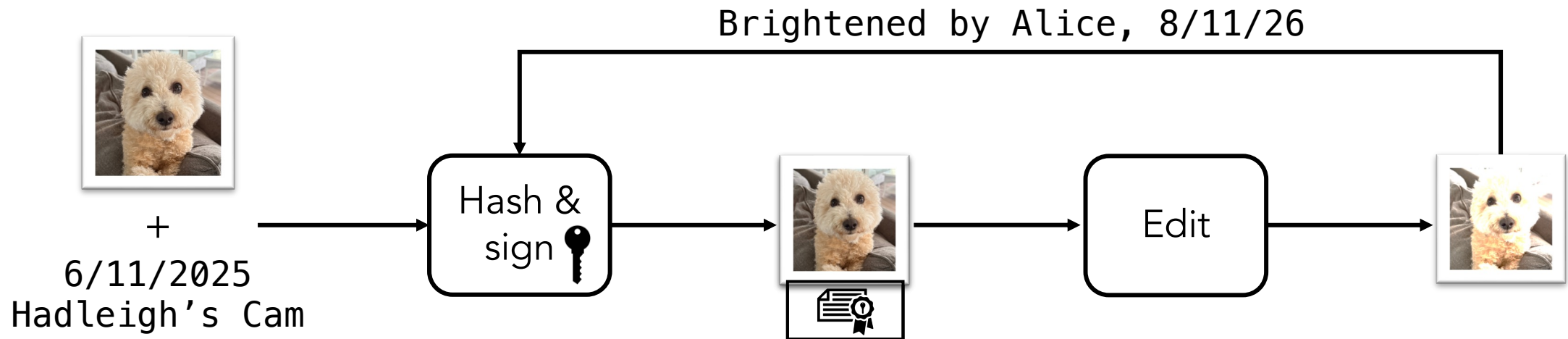
Compute over media and provenance details, attach as file metadata.



- ✓ Cryptographically verify media integrity and authenticate signer identity

# Digital Signatures

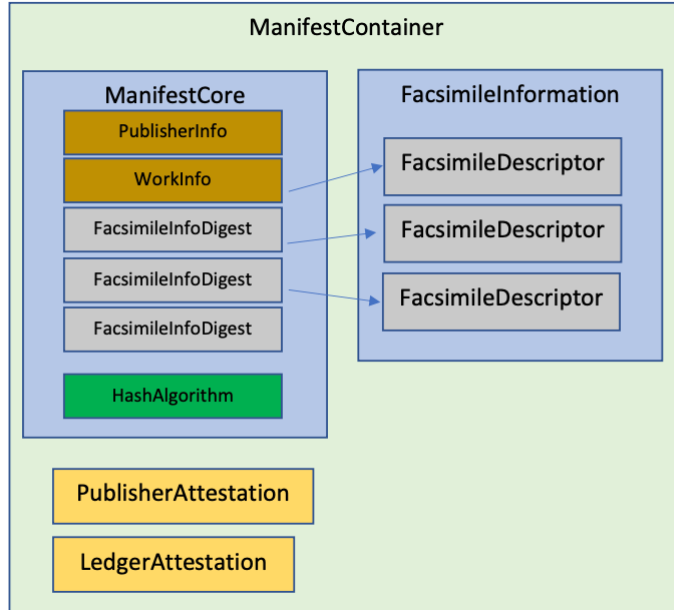
Compute over media and provenance details, attach as file metadata.



- ✓ Cryptographically verify media integrity and authenticate signer identity
- ✓ Verify edit history

# Signature-Based Provenance Standards

Specify algorithms and APIs for signature generation and verification.



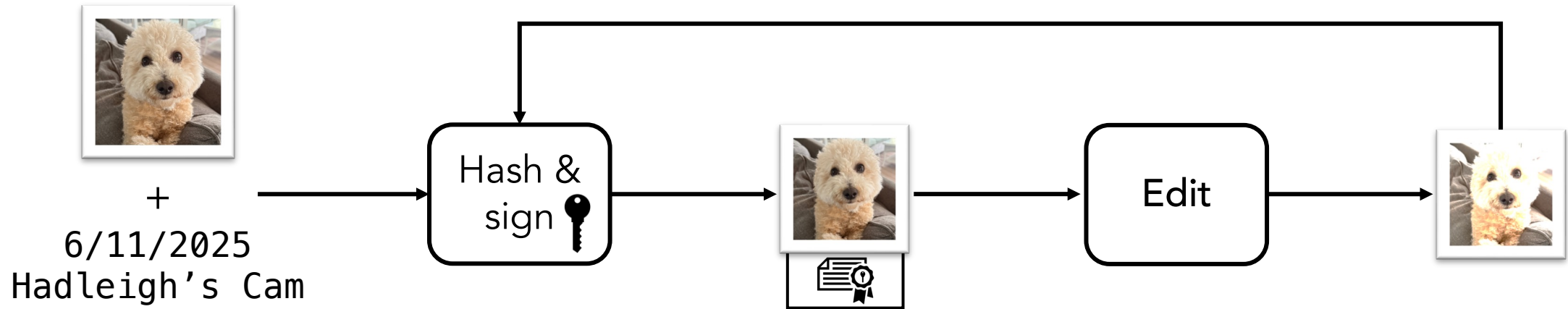
Credentials ("manifests") in Microsoft Project Origin PoC [England et al., 2021]



Content Authenticity Initiative's C2PA Content Credential display

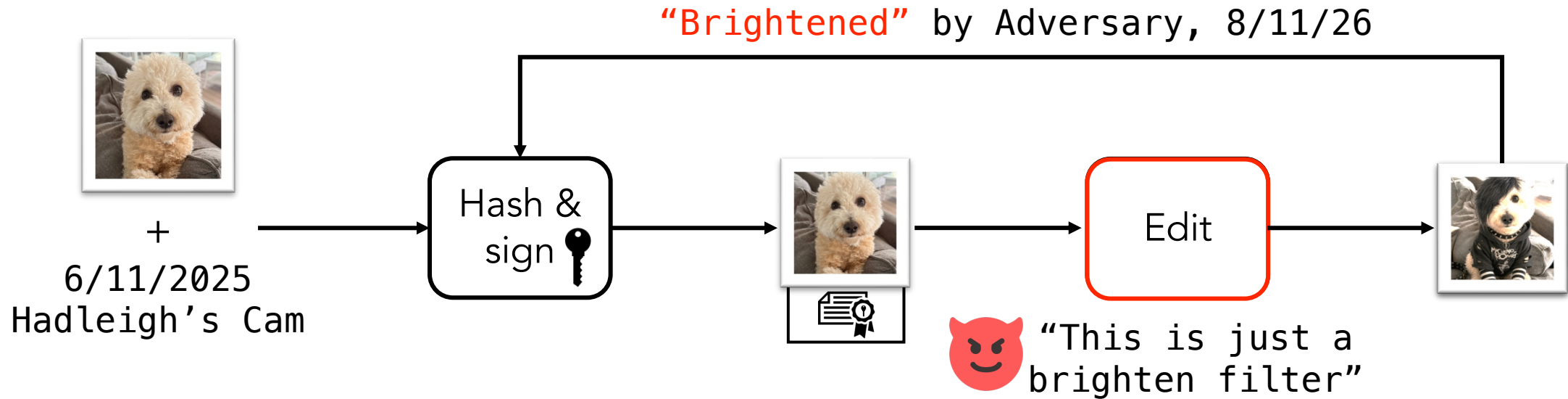
# Signed $\neq$ Authentic Origin and Edits

Problem: Adversary can falsify/misrepresent origins and edits *before* signing.



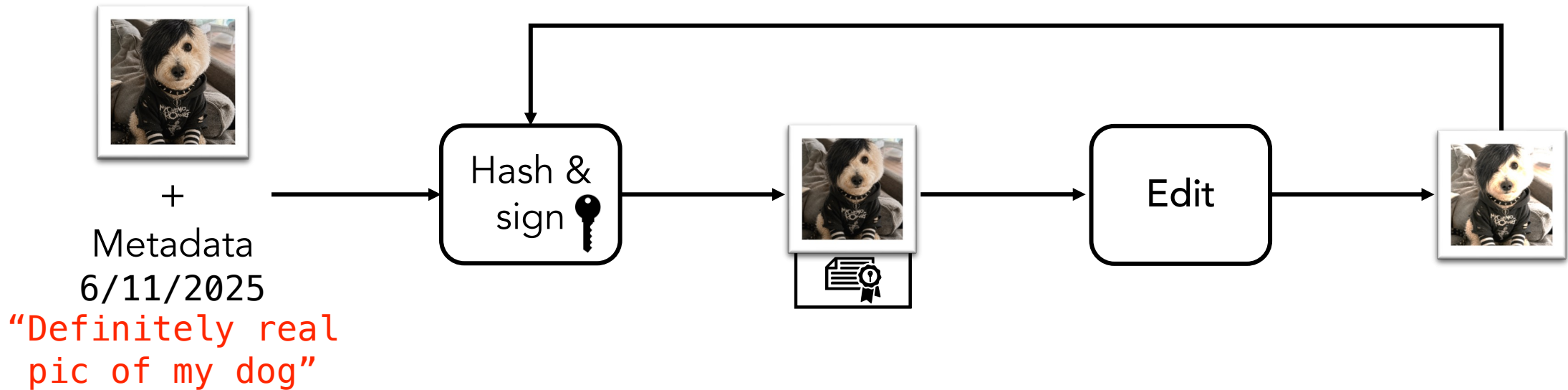
# Signed $\neq$ Authentic Origin and Edits

Problem: Adversary can falsify/misrepresent origins and edits *before* signing.



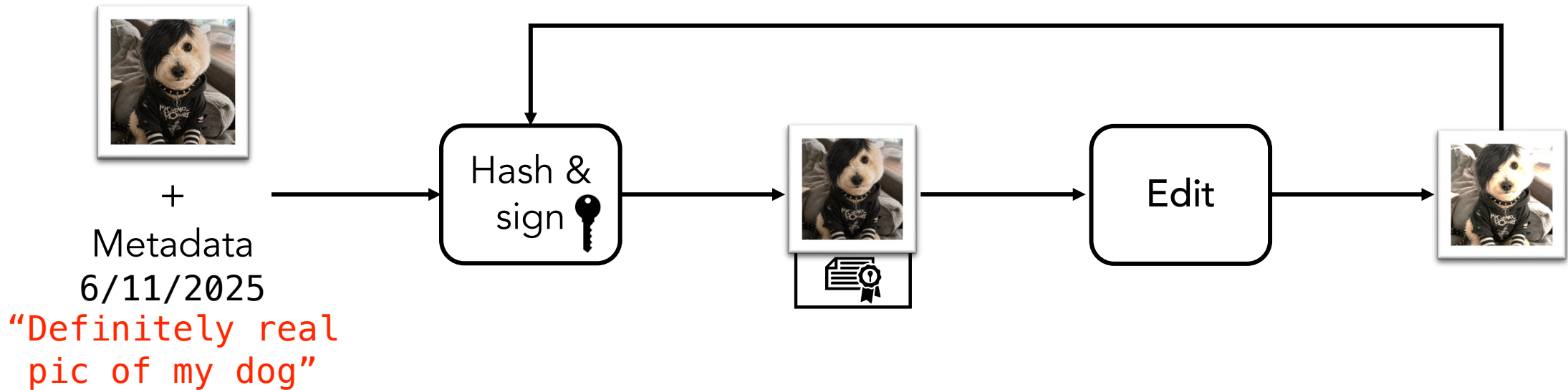
# Signed $\neq$ Authentic Origin and Edits

Problem: Adversary can falsify/misrepresent origins and edits *before* signing.



# Signed $\neq$ Authentic Origin and Edits

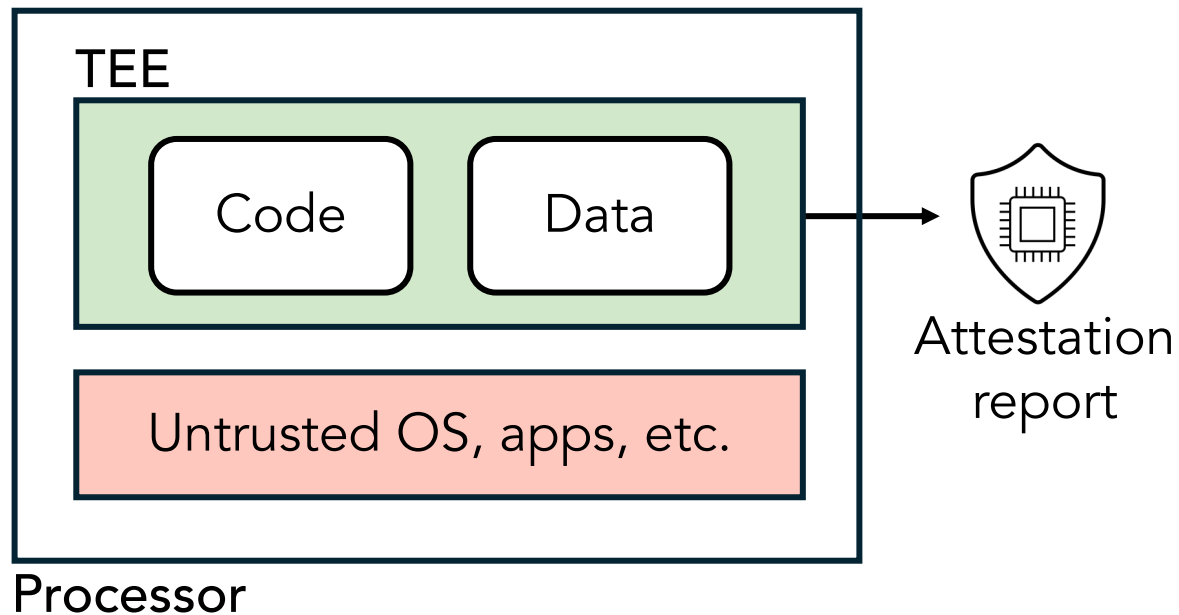
Problem: Adversary can falsify/misrepresent origins and edits *before* signing.



Solution: Perform operations within a verifiable, trusted system.

# Verifiable Video Recording and Editing

Real videos in Trusted Execution Environments (TEEs) [Liu et al., 2022].

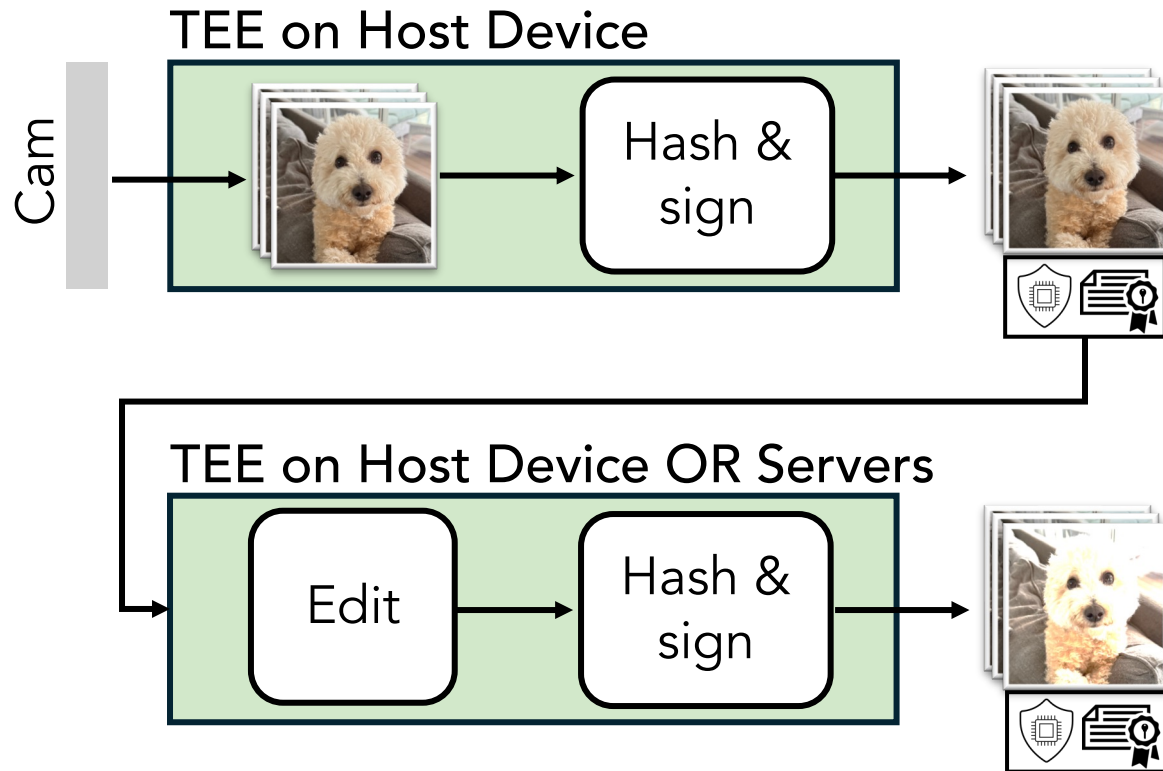


Cryptographic proof of the TEE's:

- ✓ Platform authenticity
- ✓ Code and data integrity

# Verifiable Video Recording and Editing

Real videos in Trusted Execution Environments (TEEs) [Liu et al., 2022].

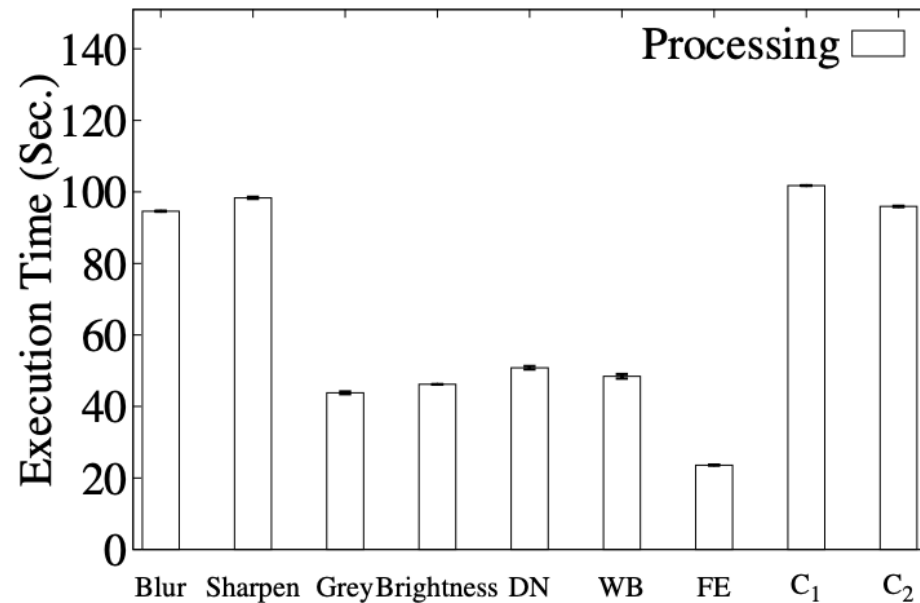


Consumers can verify integrity  
*and* authenticate:

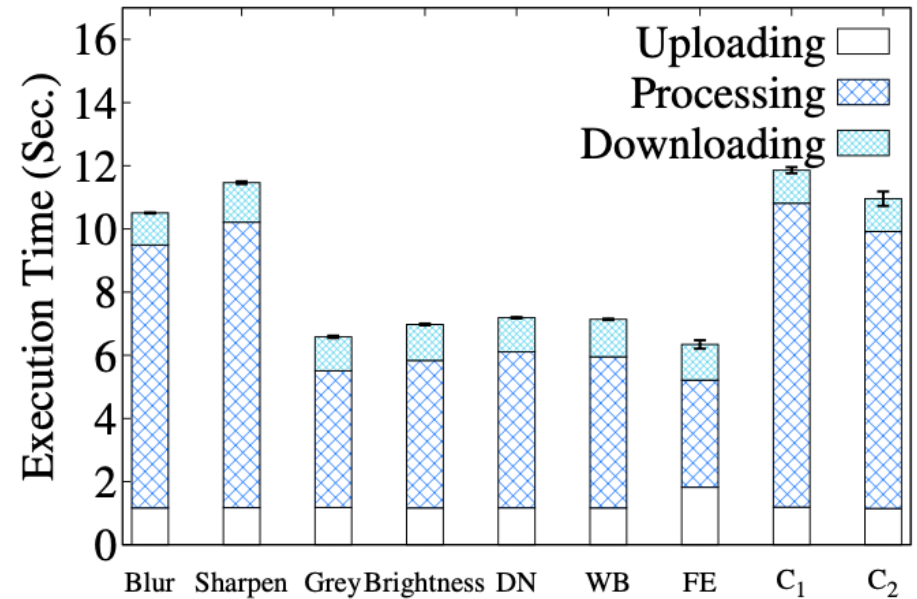
- ✓ Device used to record
- ✓ Exact edits applied

# Verifiable Video Recording and Editing

Overhead of simple video edits in TEEs [Liu et al., 2022].



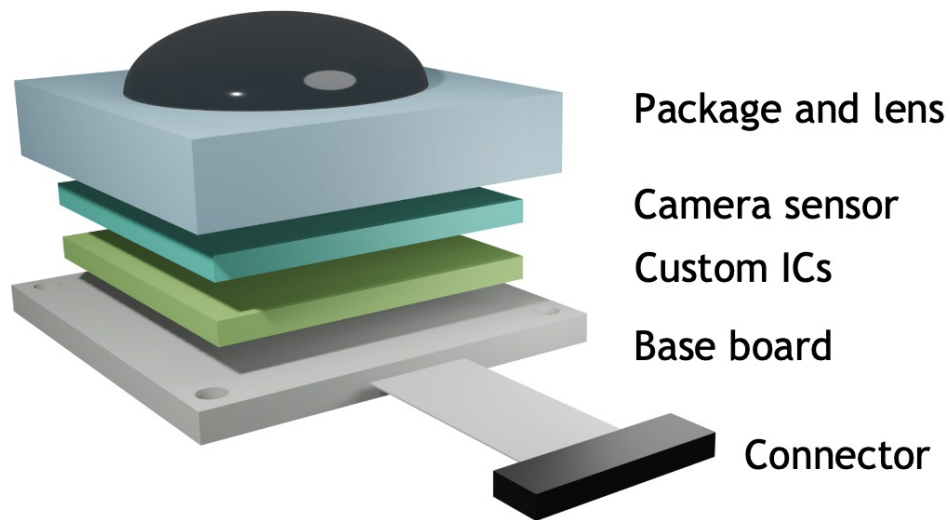
Time to apply filters to a 2-sec 720p video in ARM TrustZone.



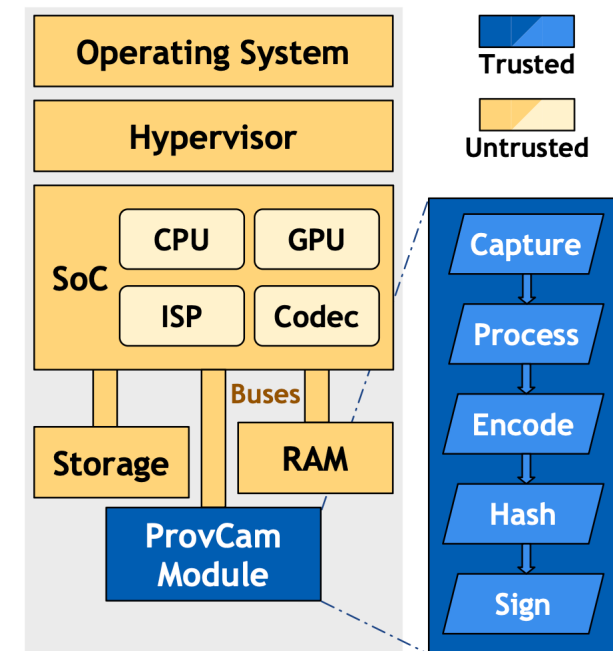
Time to transfer and filter a 2-sec 720p video in Azure Cloud Intel SGX.

# Self-Contained Verifiable Video Recording

Camera and signing in one chip and secure enclosure [Liu et al., 2024].

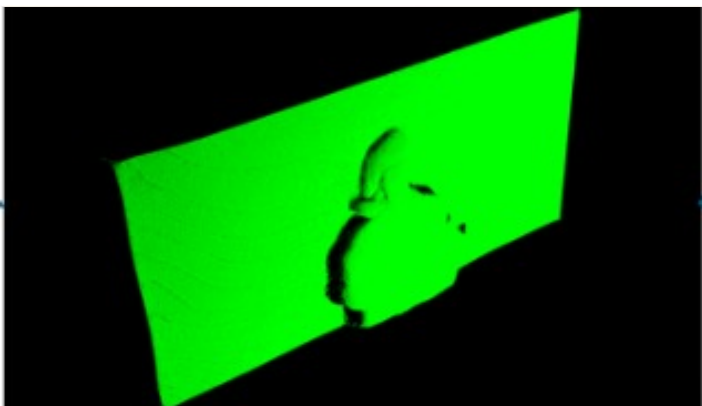


Send out frames and signed credentials via sole external connection.



Verifiable recording even on completely untrusted host system.

# Recapture Attacks on Verifiable Recording



Device fooled into signing images of displayed fake content.

Defenses:

- ✗ Post-hoc recapture detector — defeated by black-box attack [Park et al., 2025]
- ✓ Depth map with co-located time-of-flight sensors [Liu et al., 2025]

# Digital Signatures

## Benefits

Cryptographically and  
publicly verifiable

Capacity for detailed  
provenance

## Limitations

Deployment overhead for  
device-level solutions

Require broad cooperation  
to update; **easily stripped**

# Digital Watermarks

Hide persistent information within media's signal.



Pixels

or

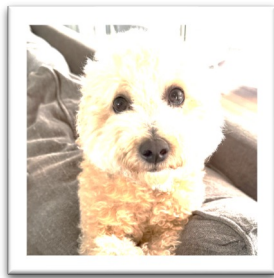


Waveform

010110  
Payload

# Digital Watermarks

Hide persistent information within media's signal.



Pixels

or

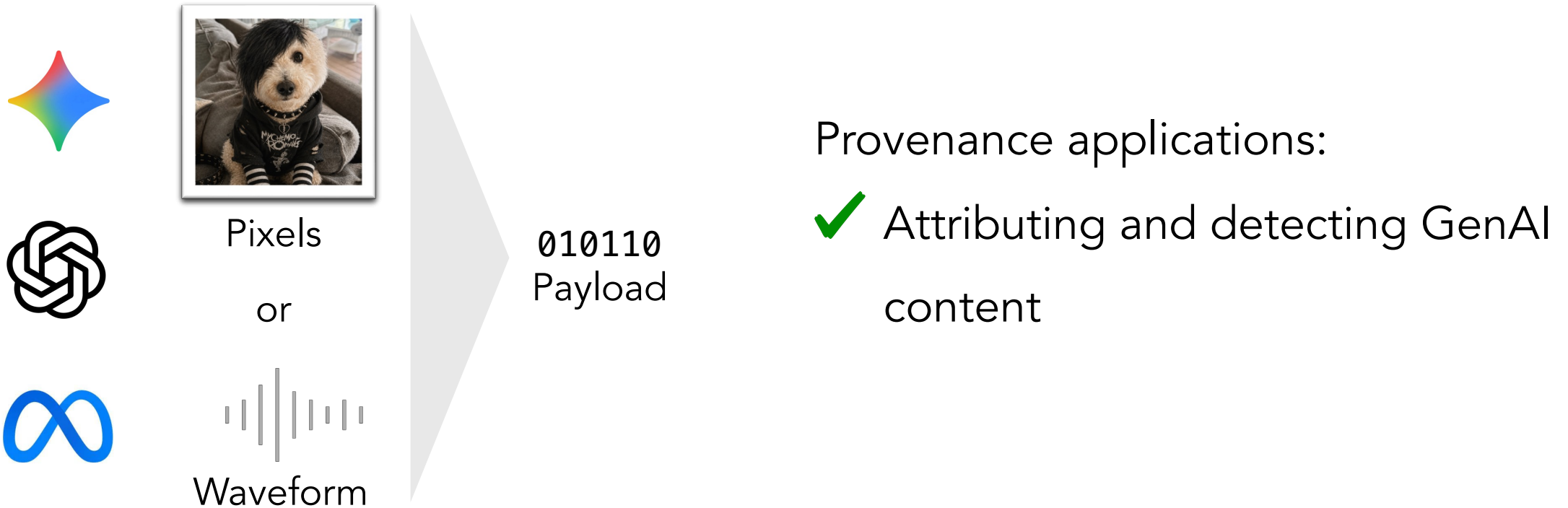


Waveform

010110  
Payload

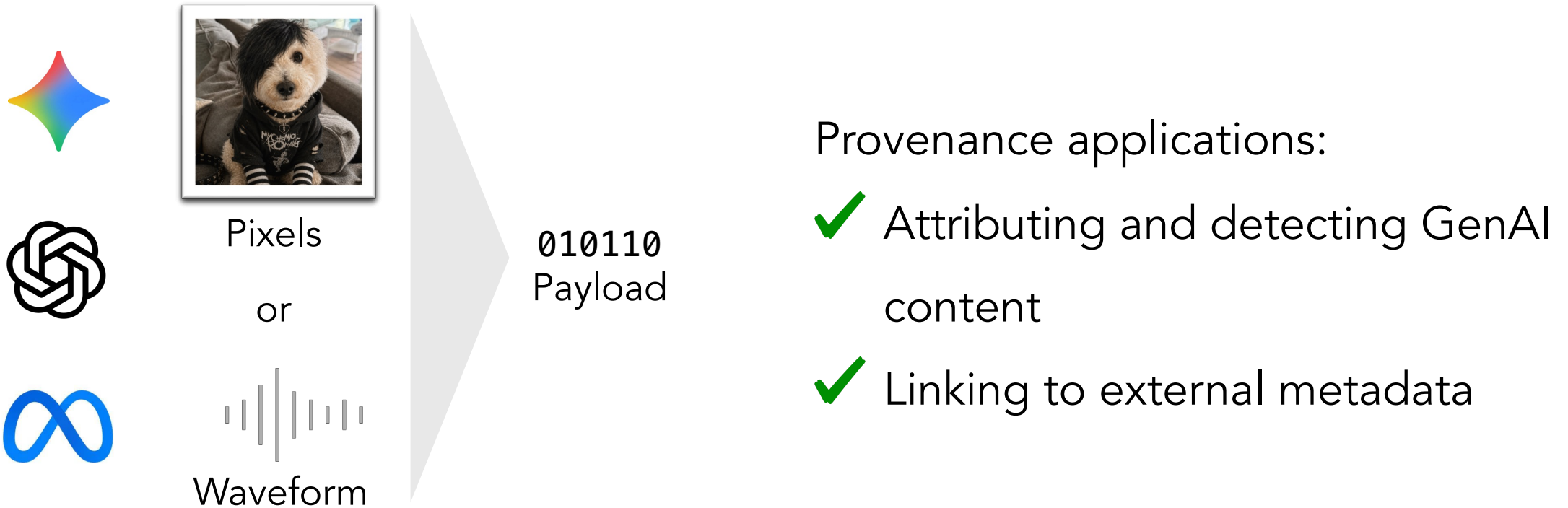
# Digital Watermarks

Hide persistent information within media's signal.



# Digital Watermarks

Hide persistent information within media's signal.



# Digital Watermarks

Hide persistent information within media's signal.



Pixels

or



Waveform

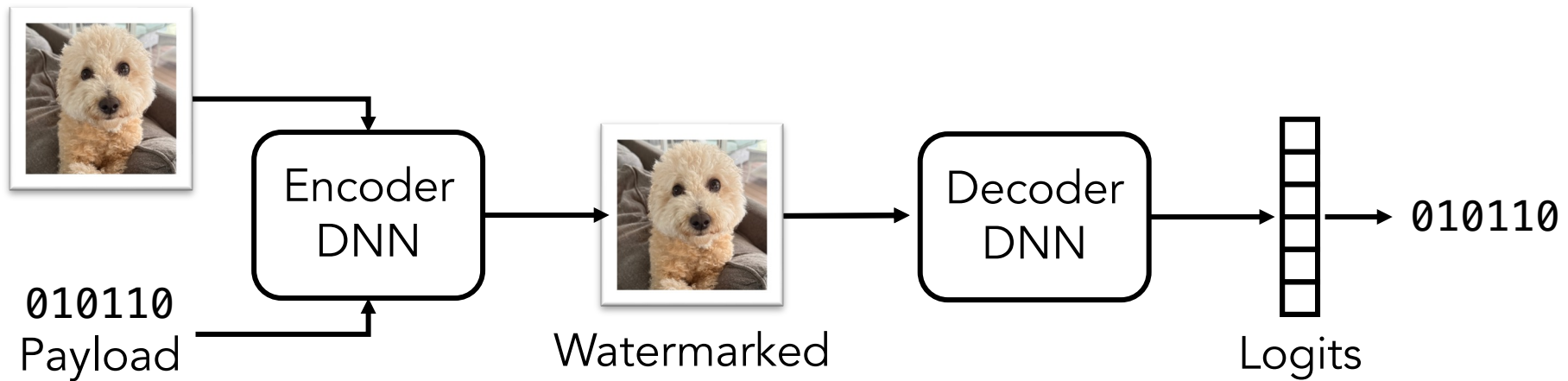
010110  
Payload

Design goals:

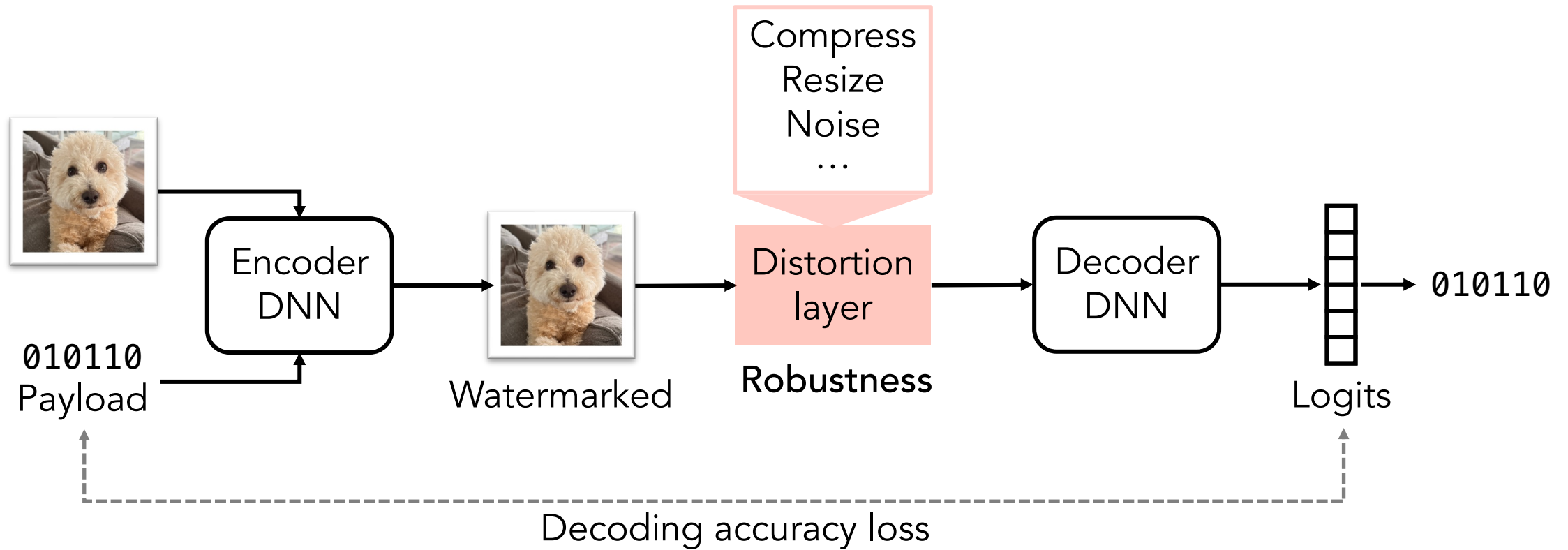
- ✓ Imperceptibility
- ✓ Robustness  
(edits benign and adversarial)
- ✓ Payload capacity
- ✓ Forgery-resistance

# Post-Hoc Watermarks via DNNs

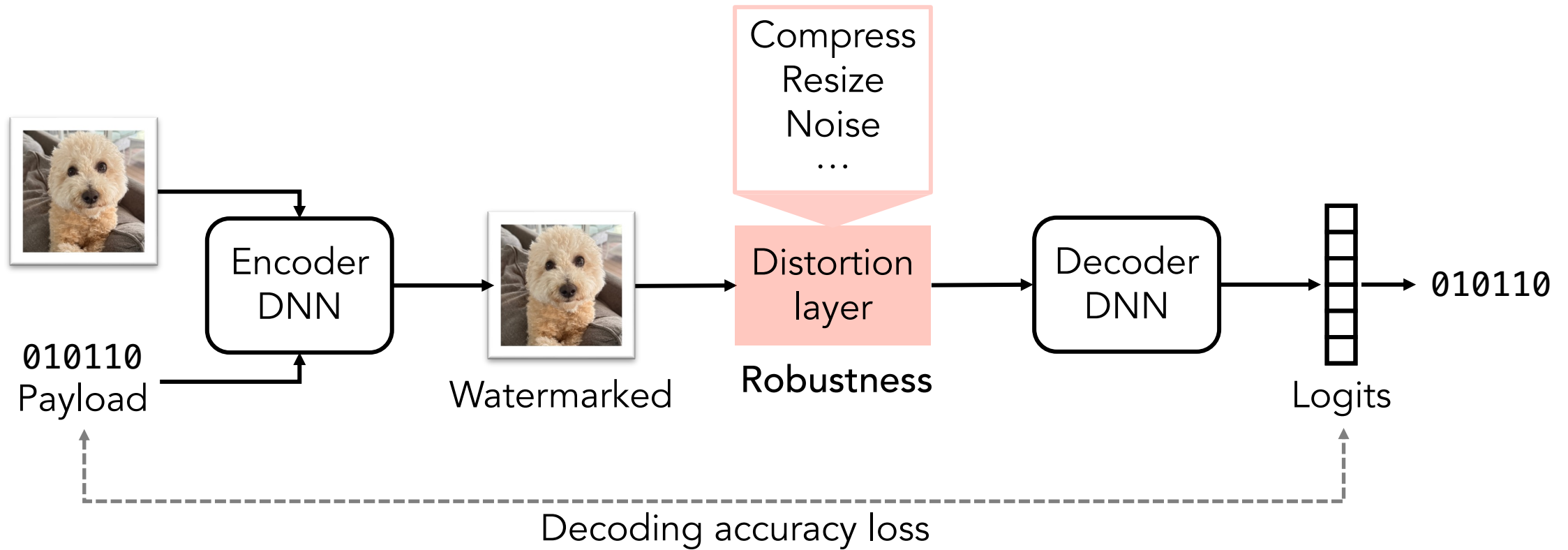
Parameterize watermarking with DNNs. First proposed in [Zhu et al., 2018].



# Post-Hoc Watermarks via DNNs

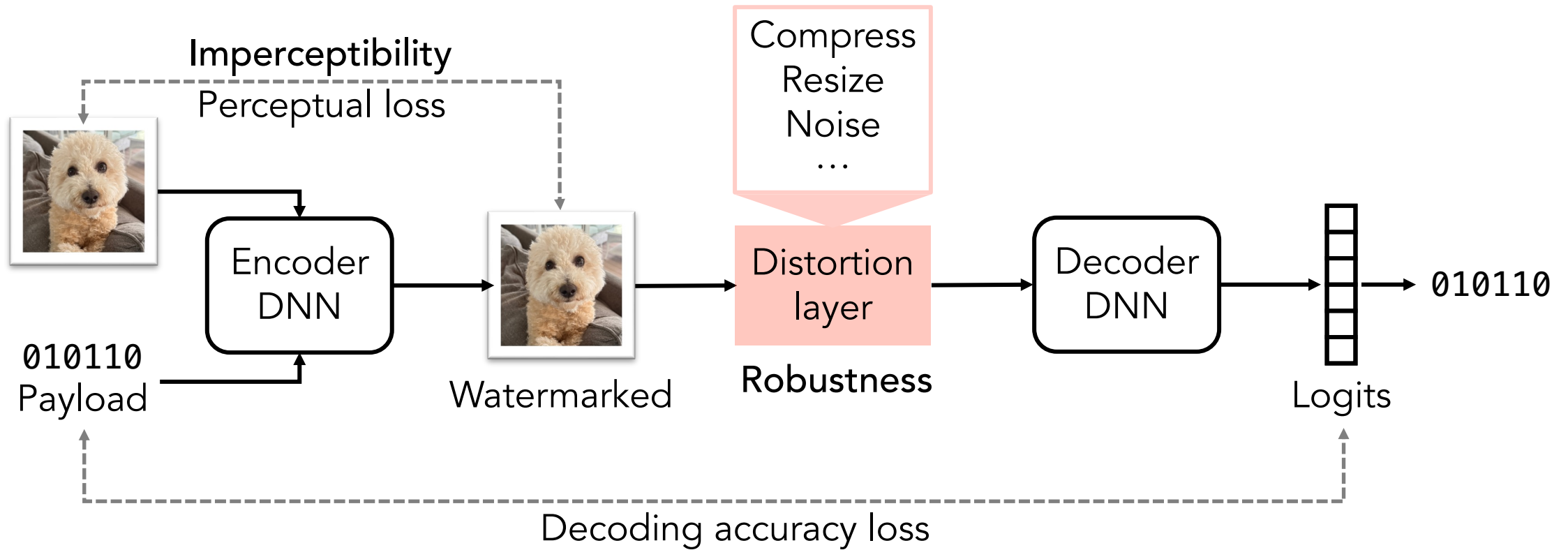


# Post-Hoc Watermarks via DNNs



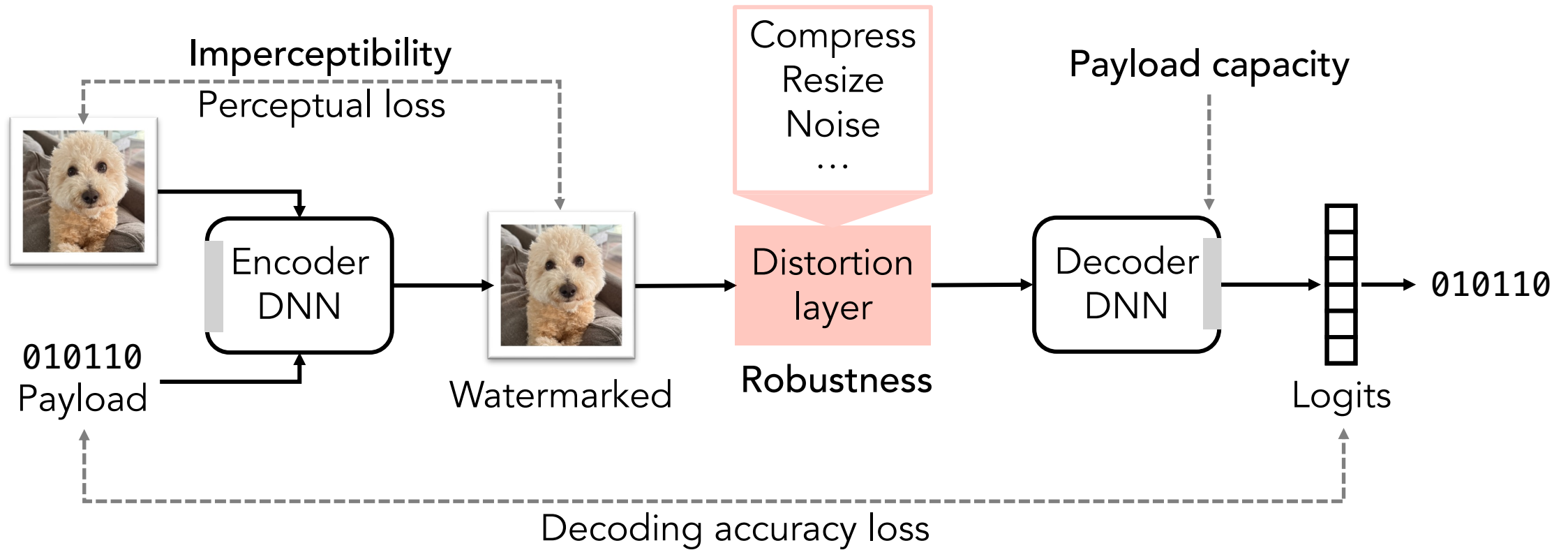
Forgery-resistance  $\approx$  Robustness (black-box setting) [An et al., 2024] [Gowal et al., 2025]

# Post-Hoc Watermarks via DNNs



Forgery-resistance  $\approx$  Robustness (black-box setting) [An et al., 2024] [Gowal et al., 2025]

# Post-Hoc Watermarks via DNNs



Forgery-resistance  $\approx$  Robustness (black-box setting) [An et al., 2024] [Gowal et al., 2025]

# Post-Hoc Watermarks via DNNs

## Image and Video

|                       | Contributions                                   | Payload (bits) |
|-----------------------|-------------------------------------------------|----------------|
| [Tancik et al., 2020] | Decode physical photos (i.e., recapture-robust) | 100            |



# Post-Hoc Watermarks via DNNs

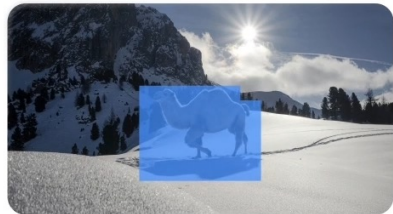
## Image and Video

|                        | Contributions                                   | Payload (bits) |
|------------------------|-------------------------------------------------|----------------|
| [Tancik et al., 2020]  | Decode physical photos (i.e., recapture-robust) | 100            |
| [Agarwal et al., 2025] | Companion removal DNN for re-watermarking       | 100            |

# Post-Hoc Watermarks via DNNs

## Image and Video

|                        | Contributions                                   | Payload (bits) |
|------------------------|-------------------------------------------------|----------------|
| [Tancik et al., 2020]  | Decode physical photos (i.e., recapture-robust) | 100            |
| [Agarwal et al., 2025] | Companion removal DNN for re-watermarking       | 100            |
| [Gowal et al., 2025]   | Google SynthID; localized detection.            | 136            |



SynthID watermarks were **detected** in these areas



SynthID watermarks were **not detected** in these areas

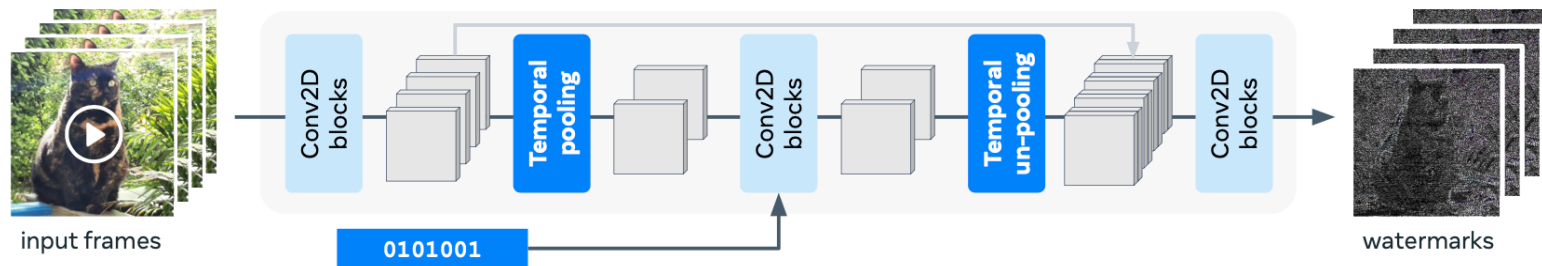


SynthID detector is **unsure** about these areas

# Post-Hoc Watermarks via DNNs

## Image and Video

|                        | Contributions                                                            | Payload (bits) |
|------------------------|--------------------------------------------------------------------------|----------------|
| [Tancik et al., 2020]  | Decode physical photos (i.e., recapture-robust)                          | 100            |
| [Agarwal et al., 2025] | Companion removal DNN for re-watermarking                                | 100            |
| [Gowal et al., 2025]   | Google SynthID; localized detection.                                     | 136            |
| [Soucek et al., 2025]  | Temporal average pooling for inference-time adaptation to video encoding | 256            |



# Post-Hoc Watermarks via DNNs

## Results: Image

Average bit accuracy across edit classes, from [Gowal et al., 2025]

| Edit      | [Gowal et al., 2025] | [Agarwal et al., 2025] | [Tancik et al., 2020] |
|-----------|----------------------|------------------------|-----------------------|
| Color     | 99.97%               | 98.16%                 | 99.14%                |
| Noise     | 98.57%               | 98.85%                 | 99.61%                |
| Quality   | 99.73%               | 99.62%                 | 99.75%                |
| Geometric | 99.82%               | 69.43%                 | 61.68%                |

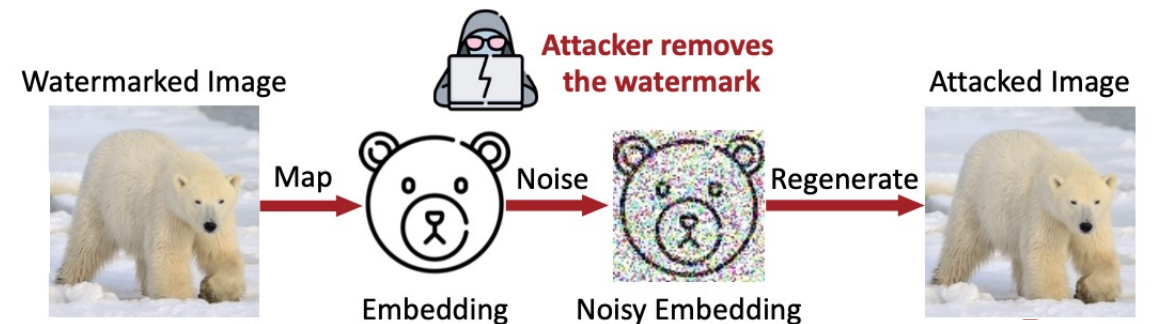
# Post-Hoc Watermarks via DNNs

## Results: Image

Average bit accuracy across edit classes, from [Gowal et al., 2025]

| Edit      | [Gowal et al., 2025] | [Agarwal et al., 2025] | [Tancik et al., 2020] |
|-----------|----------------------|------------------------|-----------------------|
| Color     | 99.97%               | 98.16%                 | 99.14%                |
| Noise     | 98.57%               | 98.85%                 | 99.61%                |
| Quality   | 99.73%               | 99.62%                 | 99.75%                |
| Geometric | 99.82%               | 69.43%                 | 61.68%                |

- ✓ Pixel-value edits
- Geometric edits
- Regeneration [Zhao et al., 2024]



# Post-Hoc Watermarks via DNNs

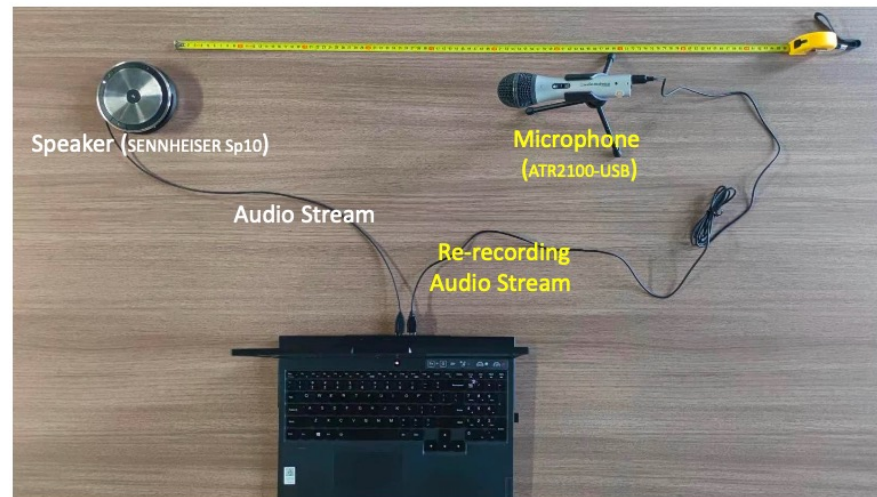
## Audio

|                     | Contributions                      | Payload (bits) |
|---------------------|------------------------------------|----------------|
| [Chen et al., 2023] | First to watermark audio with DNNs | 32             |

# Post-Hoc Watermarks via DNNs

## Audio

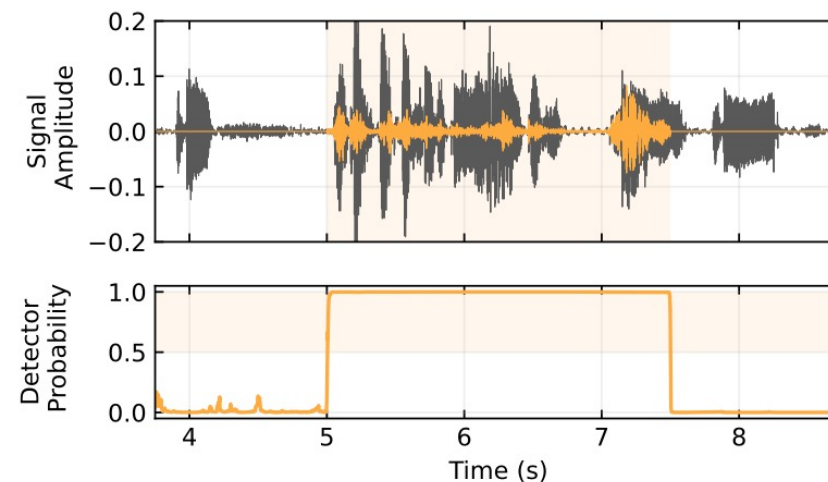
|                     | Contributions                       | Payload (bits) |
|---------------------|-------------------------------------|----------------|
| [Chen et al., 2023] | First to watermark audio with DNNs  | 32             |
| [Liu et al., 2023]  | Recapture-robustness within 1 meter | 100            |



# Post-Hoc Watermarks via DNNs

## Audio

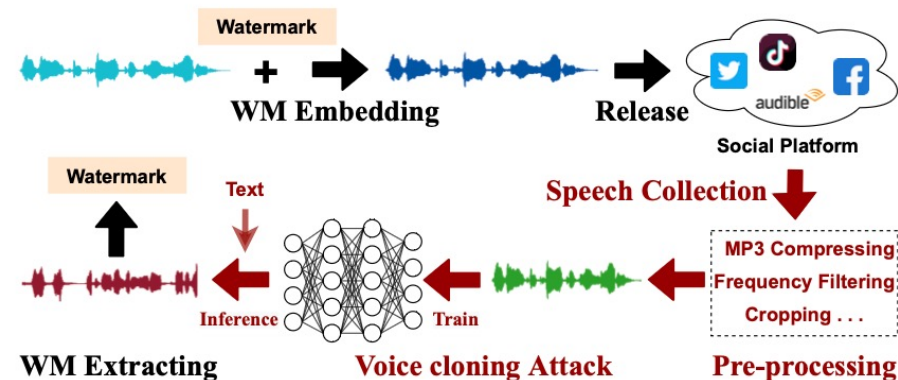
|                          | Contributions                          | Payload (bits) |
|--------------------------|----------------------------------------|----------------|
| [Chen et al., 2023]      | First to watermark audio with DNNs     | 32             |
| [Liu et al., 2023]       | Recapture-robustness within 1 meter    | 100            |
| [San Roman et al., 2024] | Sample-level detection through masking | 16             |



# Post-Hoc Watermarks via DNNs

## Audio

|                          | Contributions                                                      | Payload (bits) |
|--------------------------|--------------------------------------------------------------------|----------------|
| [Chen et al., 2023]      | First to watermark audio with DNNs                                 | 32             |
| [Liu et al., 2023]       | Recapture-robustness within 1 meter                                | 100            |
| [San Roman et al., 2024] | Sample-level detection through masking                             | 16             |
| [Liu et al., 2024b]      | Transfer to text-to-speech output from watermarked training speech | 10             |



# Post-Hoc Watermarks via DNNs

## Results: Audio

Average detection accuracy (TPR@1%FPR) across edit classes, from [O'Reilly et al., 2025]

| Edit         | [Chen et al., 2023] | [San Roman et al., 2024] | [Liu et al., 2024b] |
|--------------|---------------------|--------------------------|---------------------|
| MP3          | 100%                | 100%                     | 100%                |
| Low-pass     | 100%                | 100%                     | 100%                |
| Time stretch | 44%                 | 35%                      | 100%                |
| TiCodec      | 0%                  | 1%                       | 2%                  |

# Post-Hoc Watermarks via DNNs

## Results: Audio

Average detection accuracy (TPR@1%FPR) across edit classes, from [O'Reilly et al., 2025]

| Edit         | [Chen et al., 2023] | [San Roman et al., 2024] | [Liu et al., 2024b] |
|--------------|---------------------|--------------------------|---------------------|
| MP3          | 100%                | 100%                     | 100%                |
| Low-pass     | 100%                | 100%                     | 100%                |
| Time stretch | 44%                 | 35%                      | 100%                |
| TiCodec      | 0%                  | 1%                       | 2%                  |

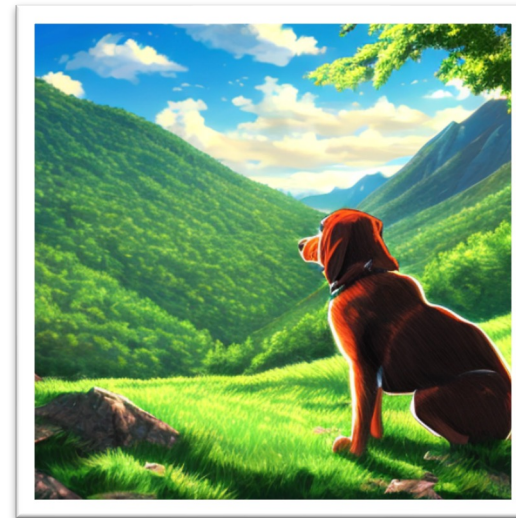
- ✓ Amplitude-value edits
- Temporal edits
- Generative edits (e.g., neural codecs)

# Semantic Watermarks

Watermark in-generation via shifts in diffusion model's output distribution.



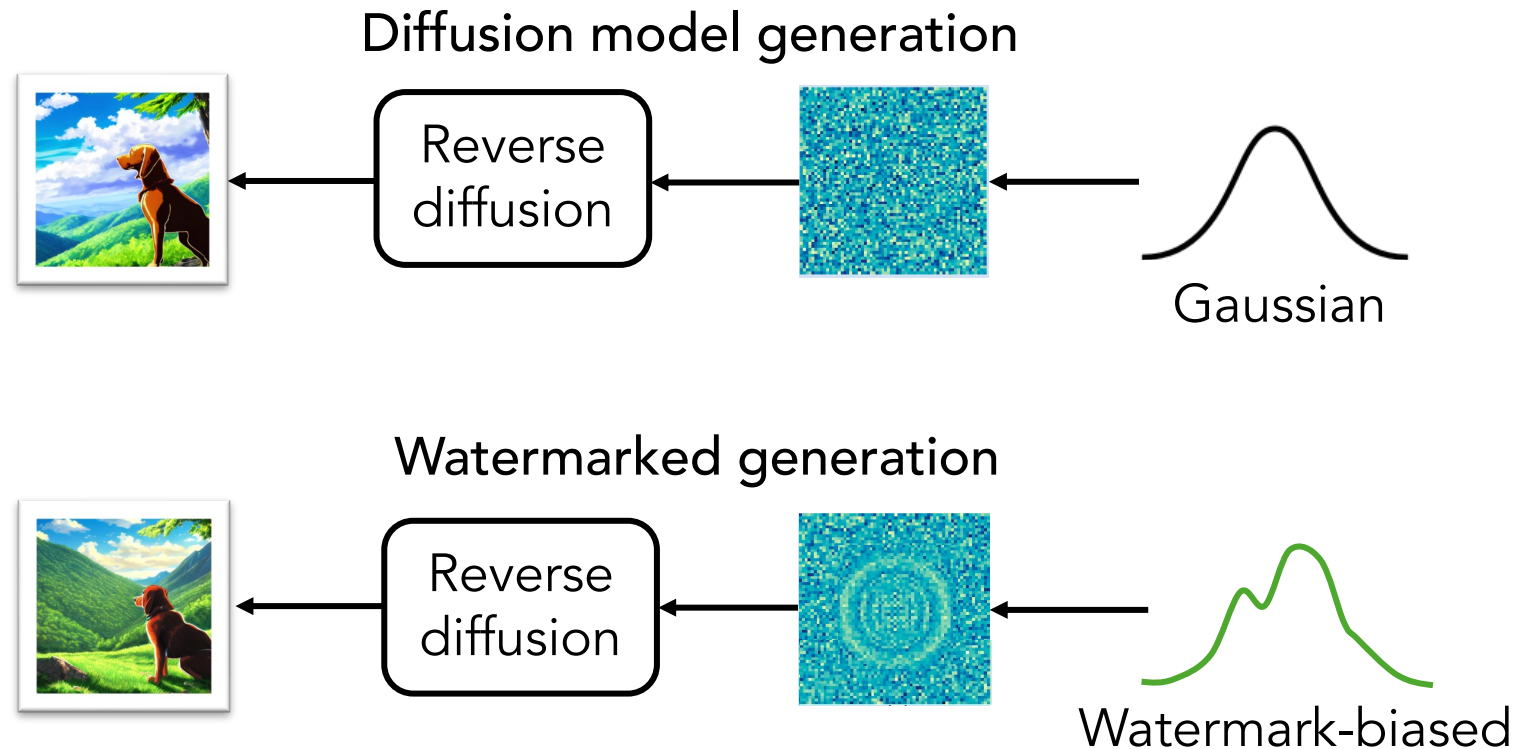
Unwatermarked



Watermarked

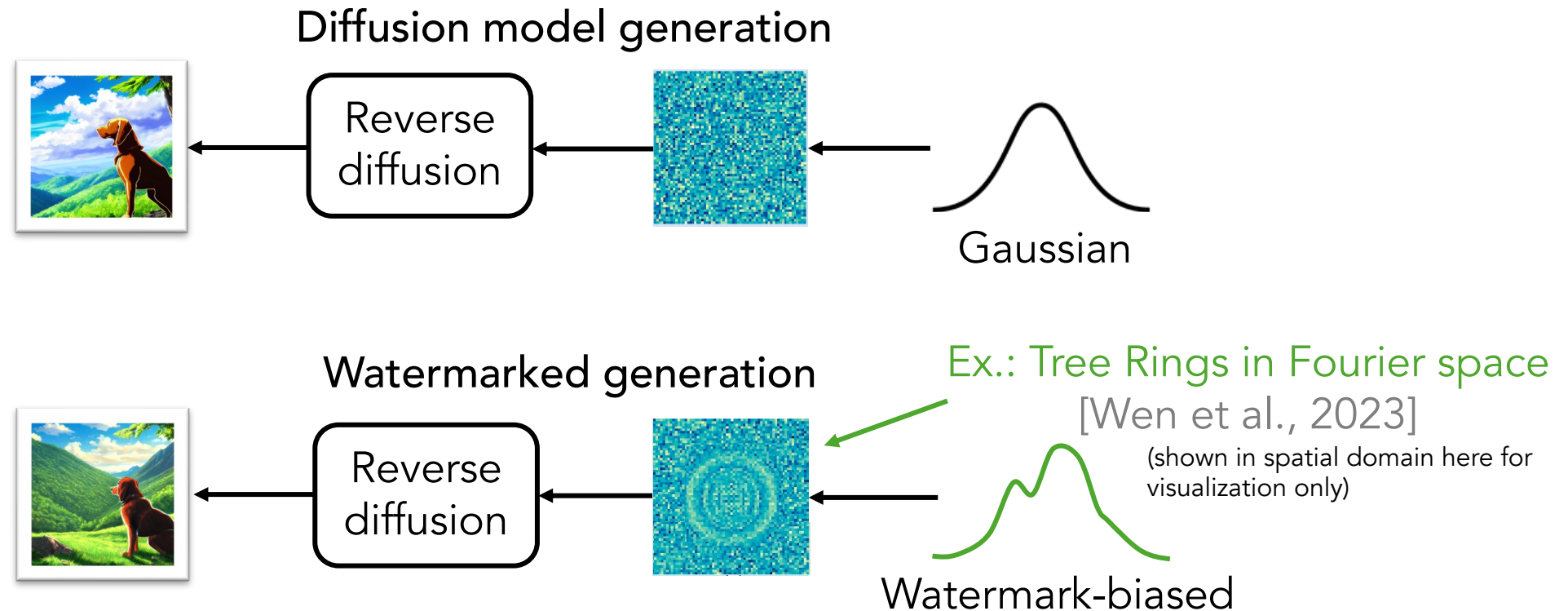
# Semantic Watermarks

Watermark in-generation via shifts in diffusion model's output distribution.



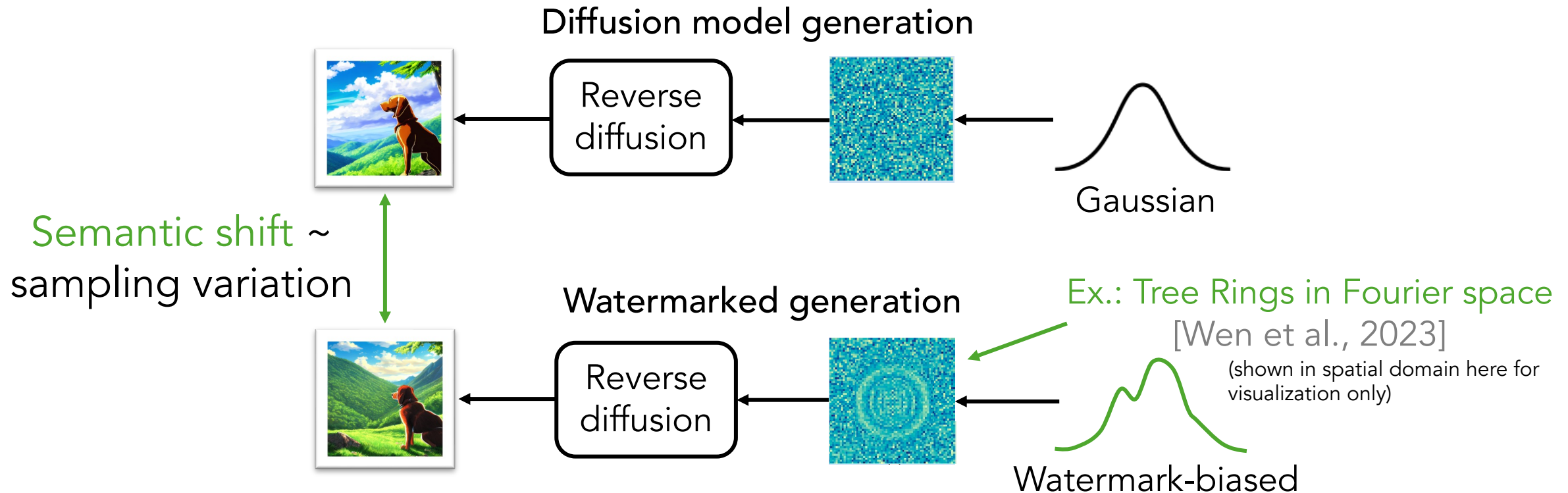
# Semantic Watermarks

Watermark in-generation via shifts in diffusion model's output distribution.



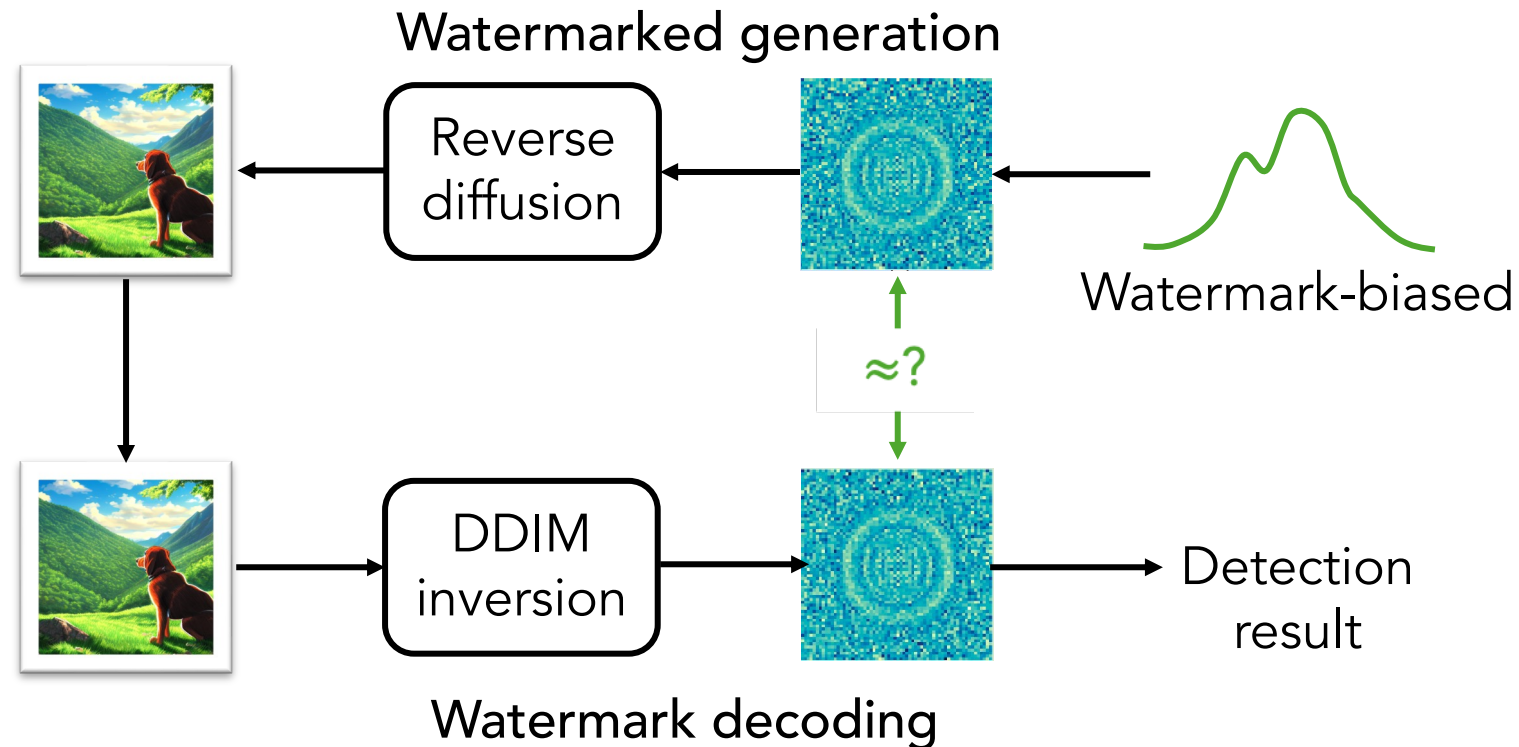
# Semantic Watermarks

Watermark in-generation via shifts in diffusion model's output distribution.



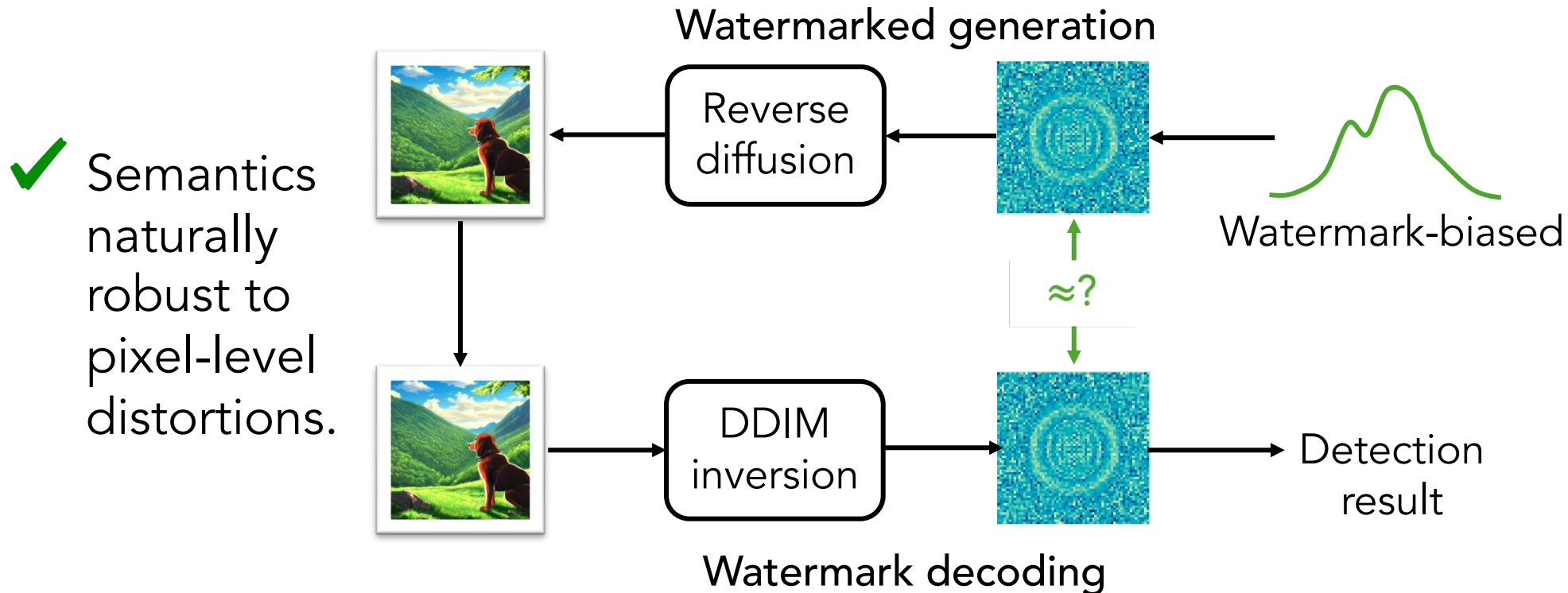
# Semantic Watermarks

Watermark in-generation via shifts in diffusion model's output distribution.



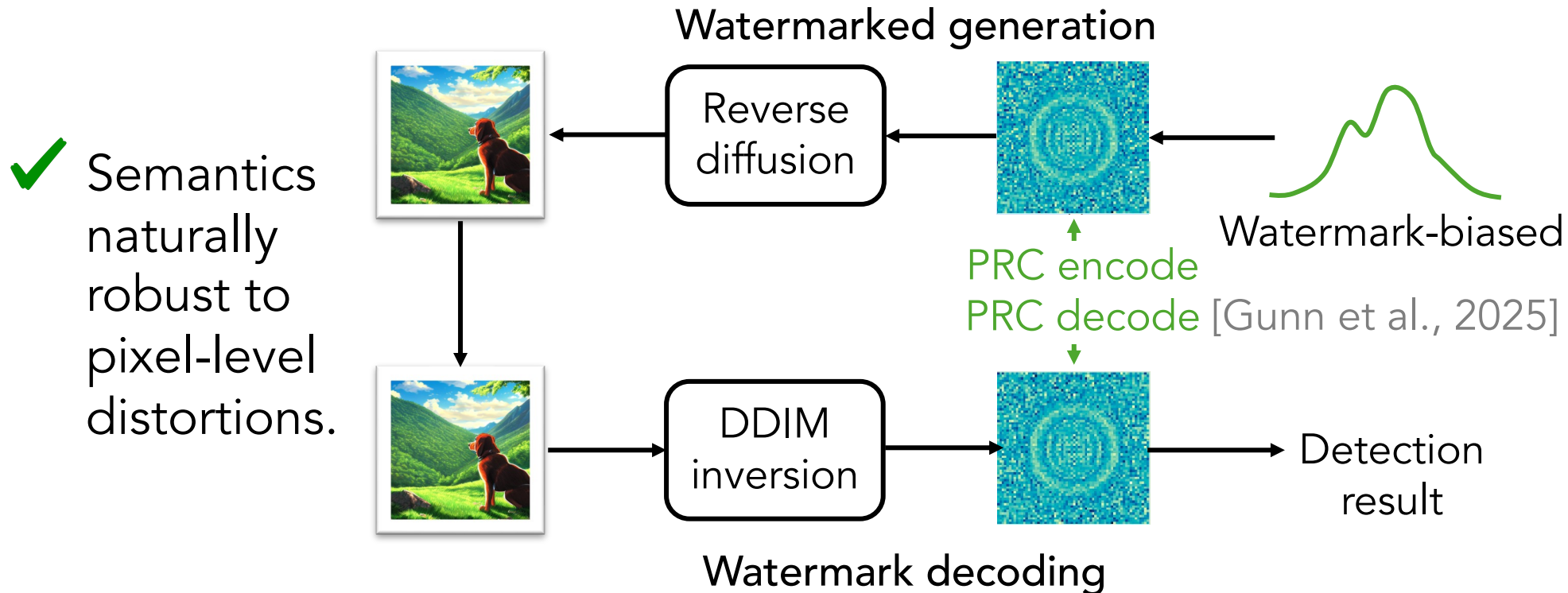
# Semantic Watermarks

Watermark in-generation via shifts in diffusion model's output distribution.



# Semantic Watermarks

Watermark in-generation via shifts in diffusion model's output distribution.



# Semantic Watermark Robustness

Results on pixel-value distortions comparable to post-hoc schemes.  
Advantage in robustness to regeneration attacks.

| Edit                 |                 | TPR@FPR=0.01 |
|----------------------|-----------------|--------------|
| JPEG                 |                 | 100%         |
| Gaussian noise       |                 | 100%         |
| Gaussian blur        |                 | 100%         |
| Regeneration attacks | VAE-Denoiser-V1 | 99.8%        |
|                      | VAE-Denoiser-V2 | 100%         |
|                      | Diffusion model | 100%         |

Tree Rings robustness as evaluated in [Zhao et al., 2024]

# Digital Watermarks

## Benefits

Provenance preserved across edits, without cooperation

---

## Limitations

Security not provable (yet)

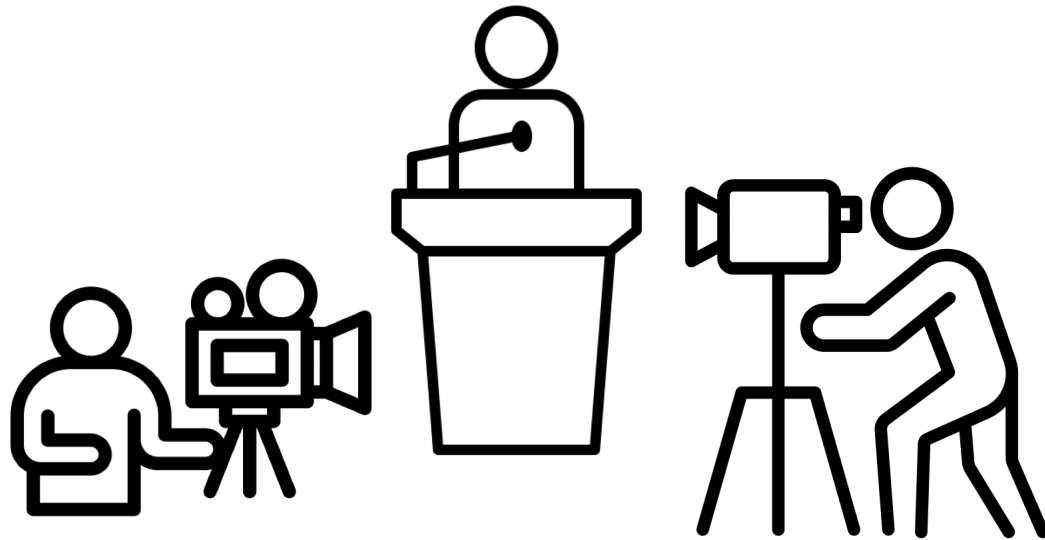
Limited payload

No public verifiability

Require cooperation to *add*

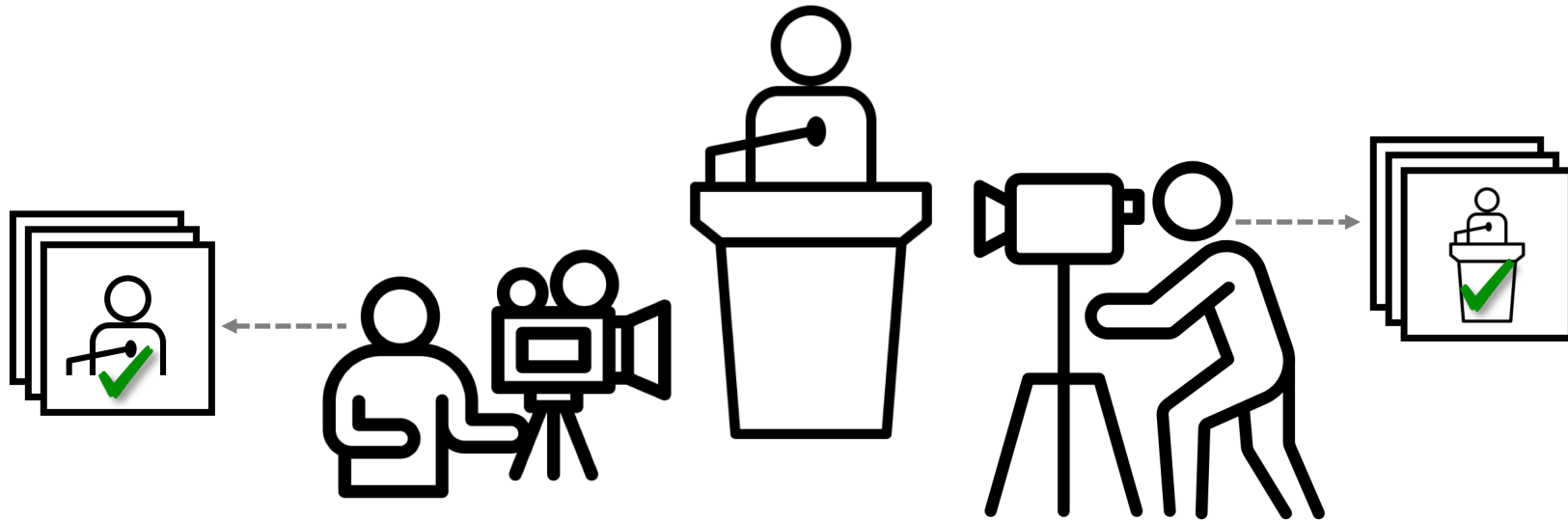
# Provenance Through Physical Signals

Modifications at scene naturally embed into all real captured media.  
No external cooperation required.



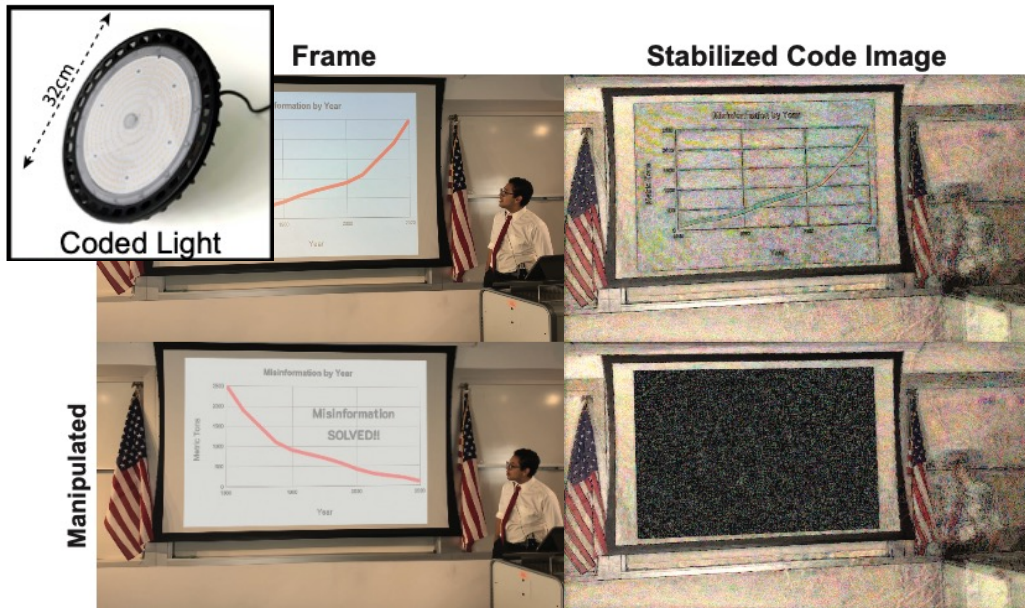
# Provenance Through Physical Signals

Modifications at scene naturally embed into all real captured media.  
No external cooperation required.

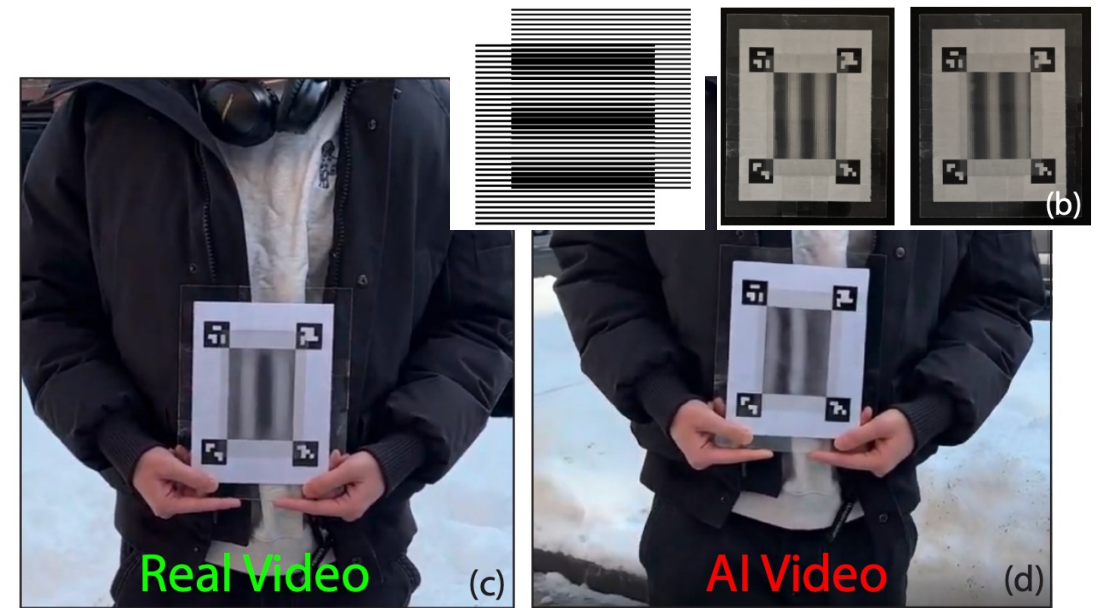


# Physical Invariants

Introduce a physical property that's hard to replicate digitally.  
Verify media by correlating against the expected invariant.



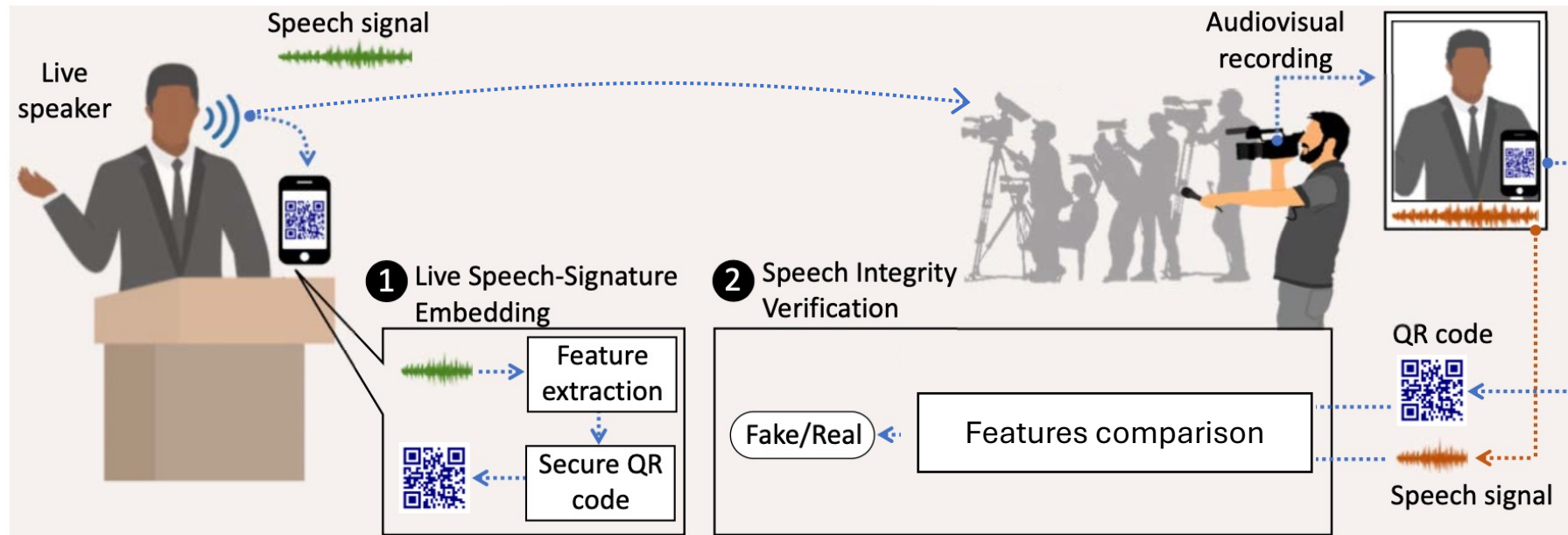
Optical modulations mimicking camera noise [Michael et al., 2025]



Camera motion is coupled with Moiré fringes' phase [Qing et al., 2026]

# Physically-Realized Signatures

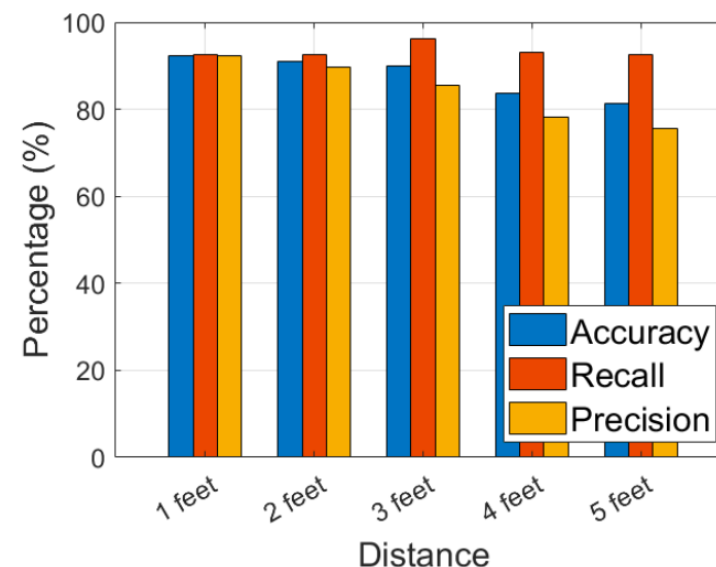
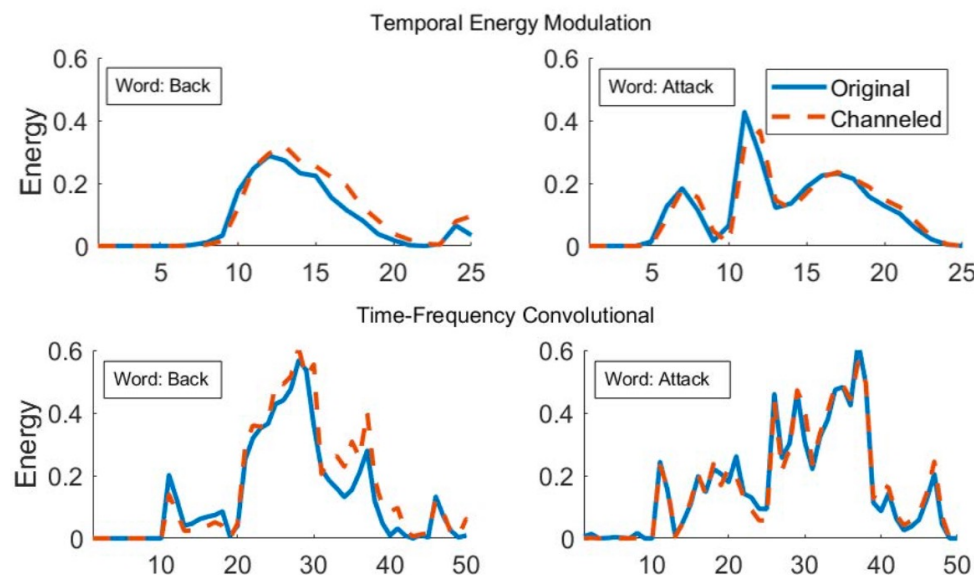
Extract and physically embed features summarizing the scene's semantics.  
Verify media by cross-checking embedded and portrayed features.



QR code-embedded signatures for live speeches [Shahid et al., 2023]

# Physically-Realized Signatures

Extract and physically embed features summarizing the scene's semantics.  
Verify media by cross-checking embedded and portrayed features.



QR code-embedded signatures for live speeches [Shahid et al., 2023]

# Provenance Through Physical Signals

## Benefits

Embed into media without external cooperation

Persist through benign post-processing

## Limitations

Environment dependencies

Require careful design of invariants/features and embedding

# Research in Proactive Media Provenance

## Digital Signatures

### Standards

[England et al., 2021]

### Verifiable Operations

[Liu et al., 2022]

[Liu et al., 2024]

### Recapture Attacks

[Park et al., 2025]

[Liu et al., 2025]

## Digital Watermarks

### Post-Hoc via DNNs

[Zhu et al., 2018]

[Tancik et al., 2020]

[Chen et al., 2023]

[Liu et al., 2023]

[Liu et al., 2024b]

[Agarwal et al., 2025]

[Gowal et al., 2025]

[San Roman et al., 2025]

[Soucek et al., 2025]

### Semantic

[Wen et al., 2023]

[Gunn et al., 2025]

### Robustness and Attacks

[An et al., 2024]

[O'Reilly et al., 2025]

[Zhao et al., 2024]

## Physical Signals

### Invariants

[Michael et al., 2025]

[Qing et al., 2026]

### Signatures

[Shahid et al., 2023]

# Open Challenges

## Security-cooperation trade-off

- Digital signatures: cryptographically-verifiable but strippable
- Watermarks: durable but security is empirical
- Physical signals: lowest cooperation assumptions, but variable security

## Distributed, fragmented media ecosystem

- Coordinating many independent parties slows adoption
- Attackers will find non-compliant creation/editing tools

# Future Directions

## Digital Signatures

- Extend TEE-based approaches to audio

## Digital Watermarks

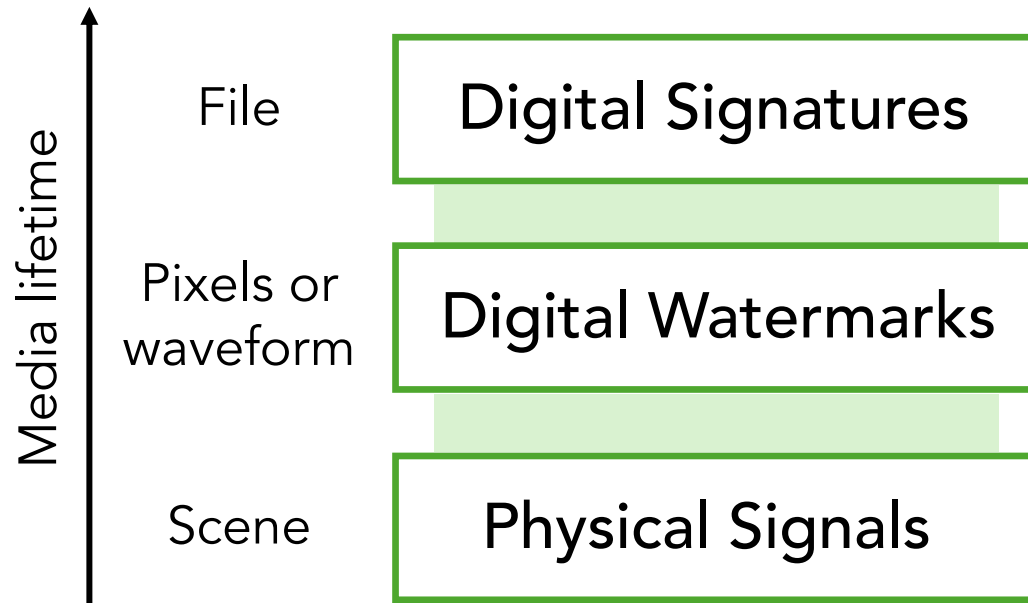
- Improve interoperability and benchmarking consistency
- Boost payload capacity to hold cryptographically-secured data

## Physical Signals

- Physically-realized signatures for audio

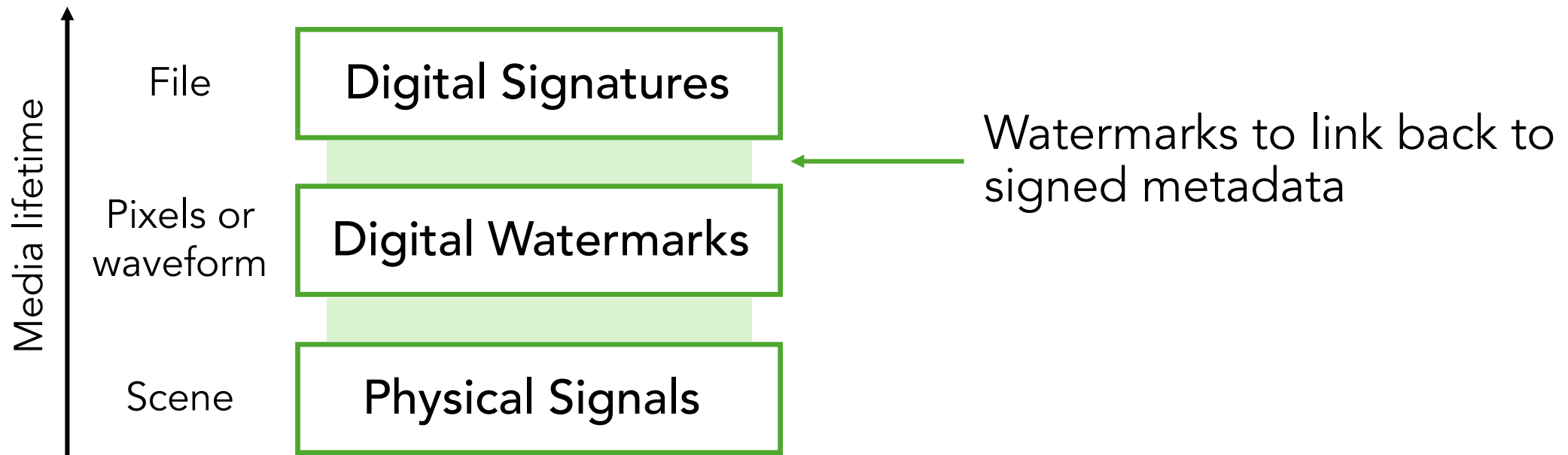
# Future Directions

## Proactive provenance stack



# Future Directions

## Proactive provenance stack



# Thank You!

## Digital Signatures

### Standards

[England et al., 2021]

### Verifiable Operations

[Liu et al., 2022]

[Liu et al., 2024]

### Recapture Attacks

[Park et al., 2025]

[Liu et al., 2025]

## Digital Watermarks

### Post-Hoc via DNNs

[Zhu et al., 2018]

[Agarwal et al., 2025]

[Tancik et al., 2020]

[Gowal et al., 2025]

[Chen et al., 2023]

[San Roman et al., 2025]

[Liu et al., 2023]

[Soucek et al., 2025]

[Liu et al., 2024b]

### Semantic

[Wen et al., 2023]

[Gunn et al., 2025]

### Robustness and Attacks

[An et al., 2024]

[O'Reilly et al., 2025]

[Zhao et al., 2024]

## Physical Signals

### Invariants

[Michael et al., 2025]

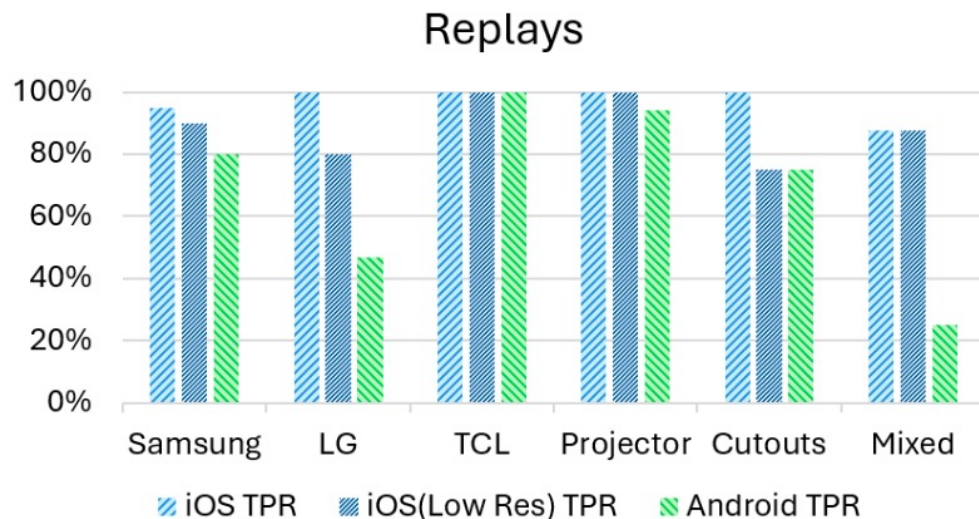
[Qing et al., 2026]

### Signatures

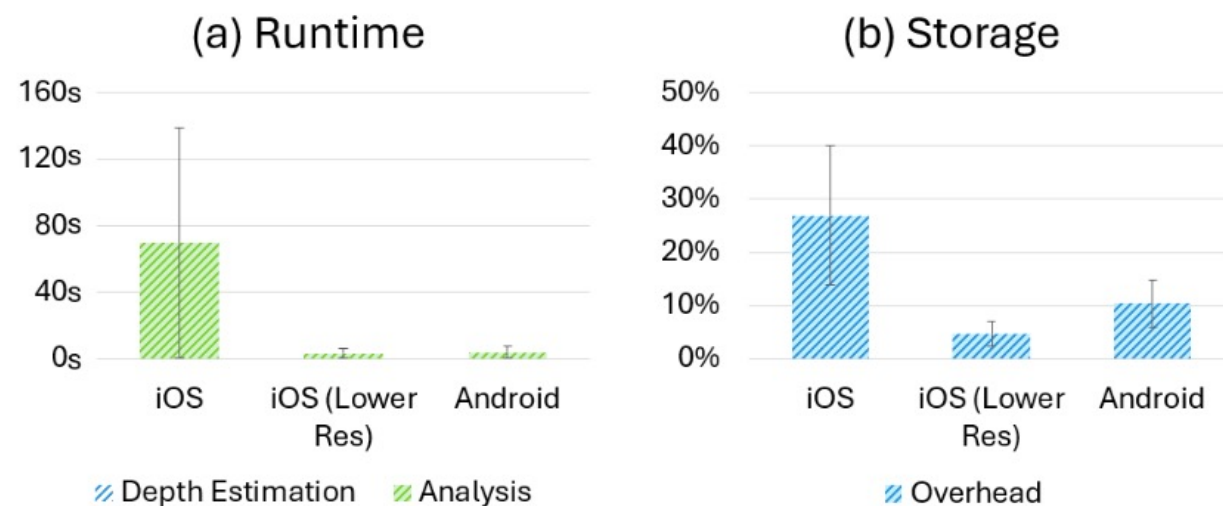
[Shahid et al., 2023]

# Backup

# [Liu et al., 2025] Results



Depth sensing-based recapture detection results across display mediums.



Recapture detection performance overhead.

# Post-Hoc Watermarking with DNNs

## Image and Video

|                        | Modality | Architectures                                          | Distortions                                       | Perceptual loss                        | Payload (bits) |
|------------------------|----------|--------------------------------------------------------|---------------------------------------------------|----------------------------------------|----------------|
| [Zhu et al., 2018]     | I        | FCNs                                                   | Six differentiable basic edits                    | L2, discriminator                      | 30             |
| [Tancik et al., 2020]  | I        | U-Net encoder, STN + CNN decoder, pretrained segmenter | Basic valuemetric edits and physical transmission | L2, LPIPS, discriminator               | 100            |
| [Soucek et al., 2025]  | I/V      | U-Net encoder, ConvNext decoder                        | Geometric and volumetric basic edits              | Discriminator only                     | 256            |
| [Agarwal et al., 2025] | I        | MUNIT-style encoder, ResNet decoder                    | Geometric and valuemetric basic edits             | MSE in YUV space, LPIPS, discriminator | 100            |

# Post-Hoc Watermarking with DNNs

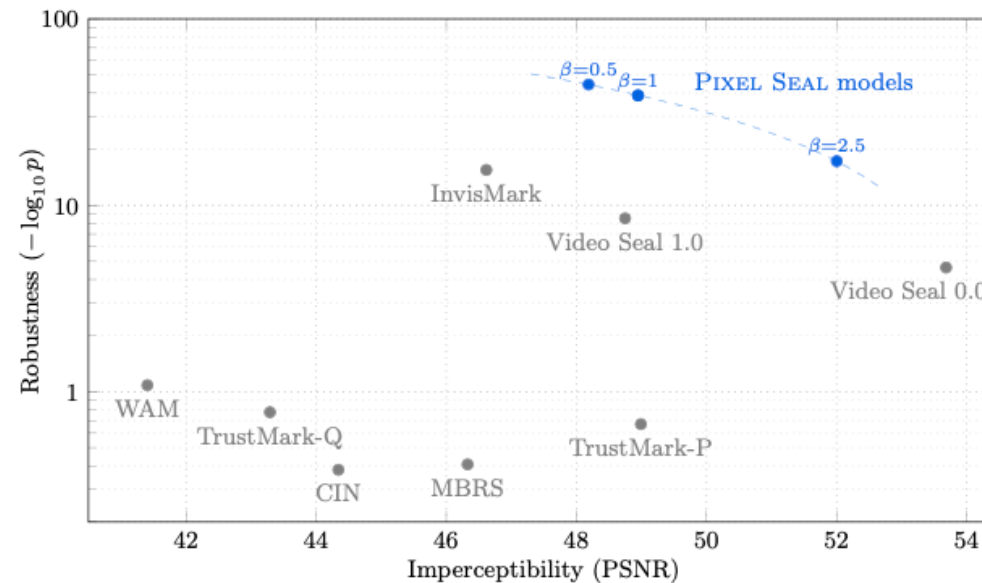
## Audio

|                          | Input domain | Architectures              | Distortions                             | Perceptual loss        | Payload (bits) |
|--------------------------|--------------|----------------------------|-----------------------------------------|------------------------|----------------|
| [Chen et al., 2023]      | STFT         | Invertible NNs             | 10 basic edits, differentiable          | L2, discriminator      | 32             |
| [San Roman et al., 2023] | Time         | Encodec                    | Basic edits, differentiable or with STE | Novel TF-Loudness loss | 16             |
| [Liu et al., 2024b]      | STFT         | FCNs with gating           | STFT perturbations, Mel-spec. vocoding  | MSE, discriminator     | 10             |
| [Liu et al., 2023]       | DWT          | FCNs with skip connections | RIR, bandpass, gaussian noise           | MSE, discriminator     | 100            |

# Post-Hoc Watermarks via DNNs

## Results: Image and Video

Always a trade-off between robustness, imperceptibility, and payload (within a given scheme).



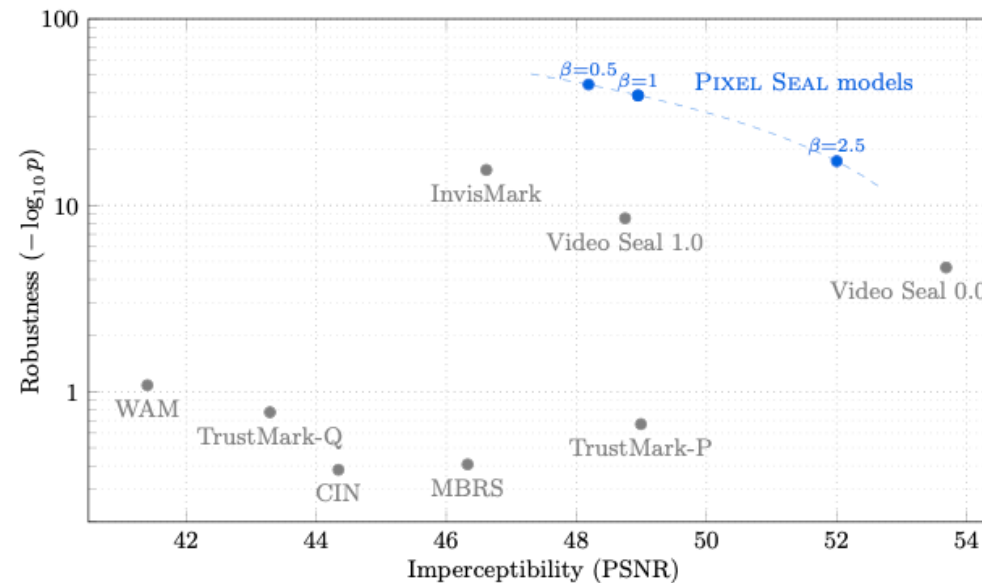
Robustness-imperceptibility Pareto curve in [Soucek et al., 2025]

# Post-Hoc Watermarks via DNNs

## Results: Image and Video

Always a trade-off between robustness, imperceptibility, and payload (within a given scheme).

✓ Consistent  
imperceptibility  
(threshold = 40)



Robustness-imperceptibility Pareto curve in [Soucek et al., 2025]

# [An et. al., 2024] Benchmark

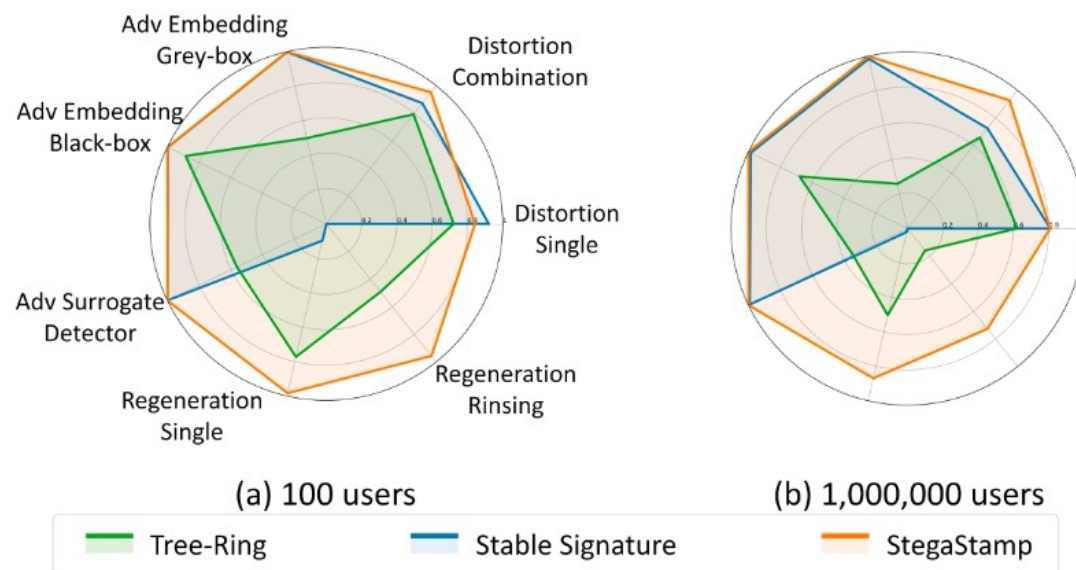
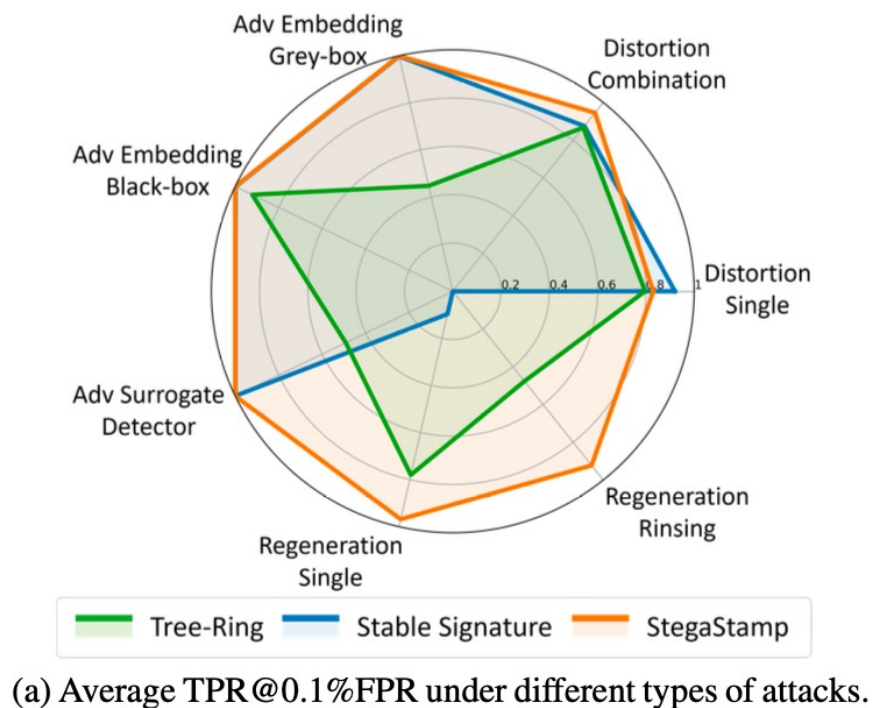
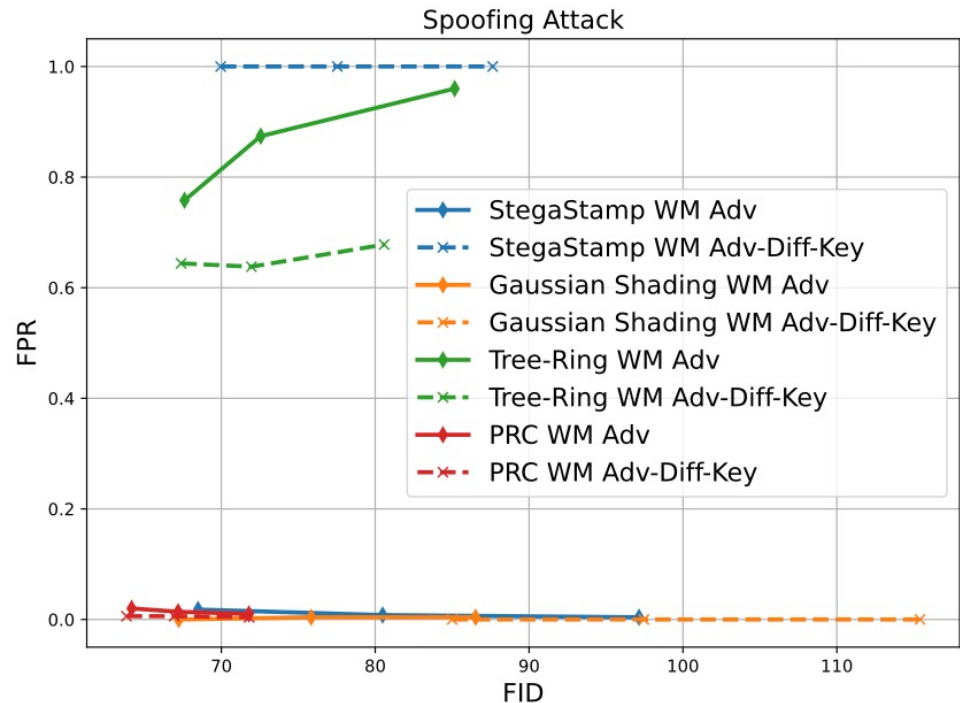


Figure 9. Identification accuracy of three watermarks after attacks.

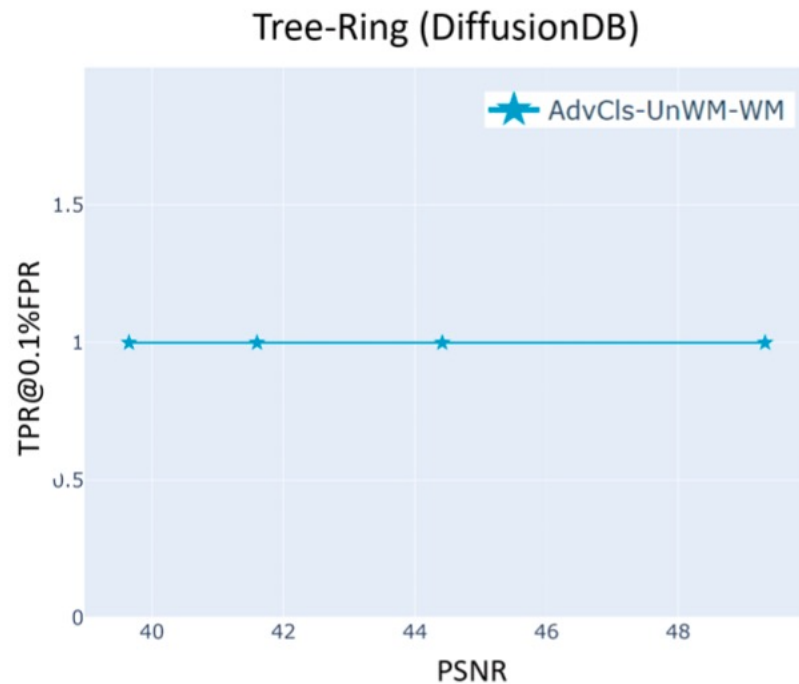
# [Tancik et al. 2020] Forgery Discrepancy

[An et al., 2025] state in text that [Tancik et al., 2025] resists the adversarial surrogate attacks

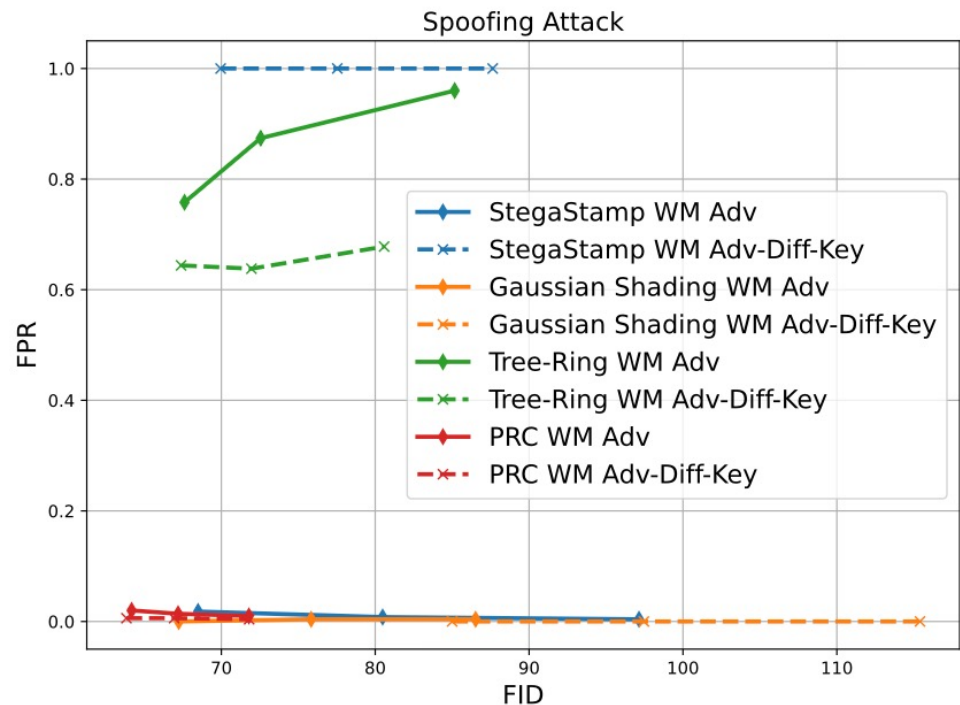


[Gunn et al., 2025] report [Tancik et al., 2025] can be completely forged.

# [Wen et al. 2023] Forgery Discrepancy

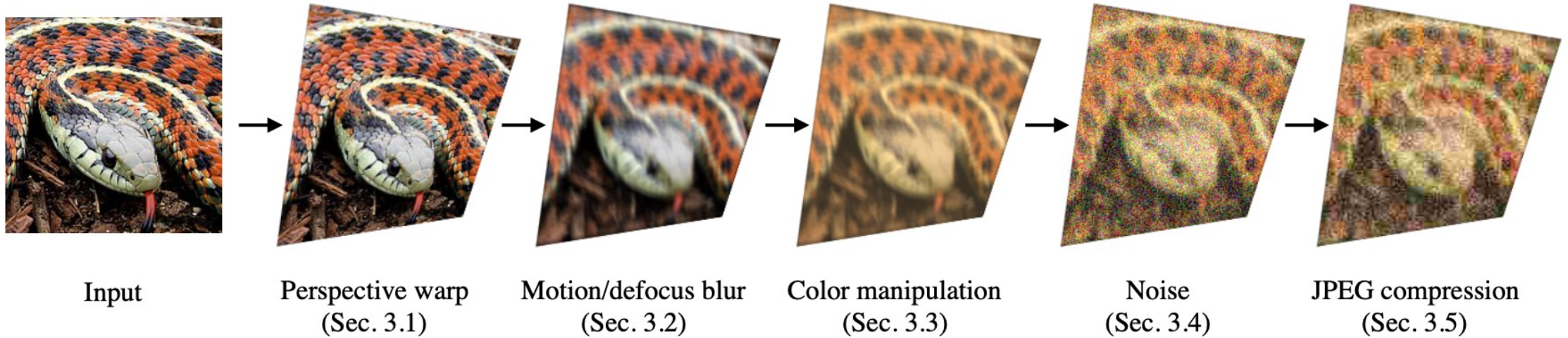


[An et al., 2025] report [Wen et al., 2023] cannot be forged, but only provide this (misleading) plot. True positive is not a spoof.



[Gunn et al., 2025] report [Wen et al., 2023] can be completely forged.

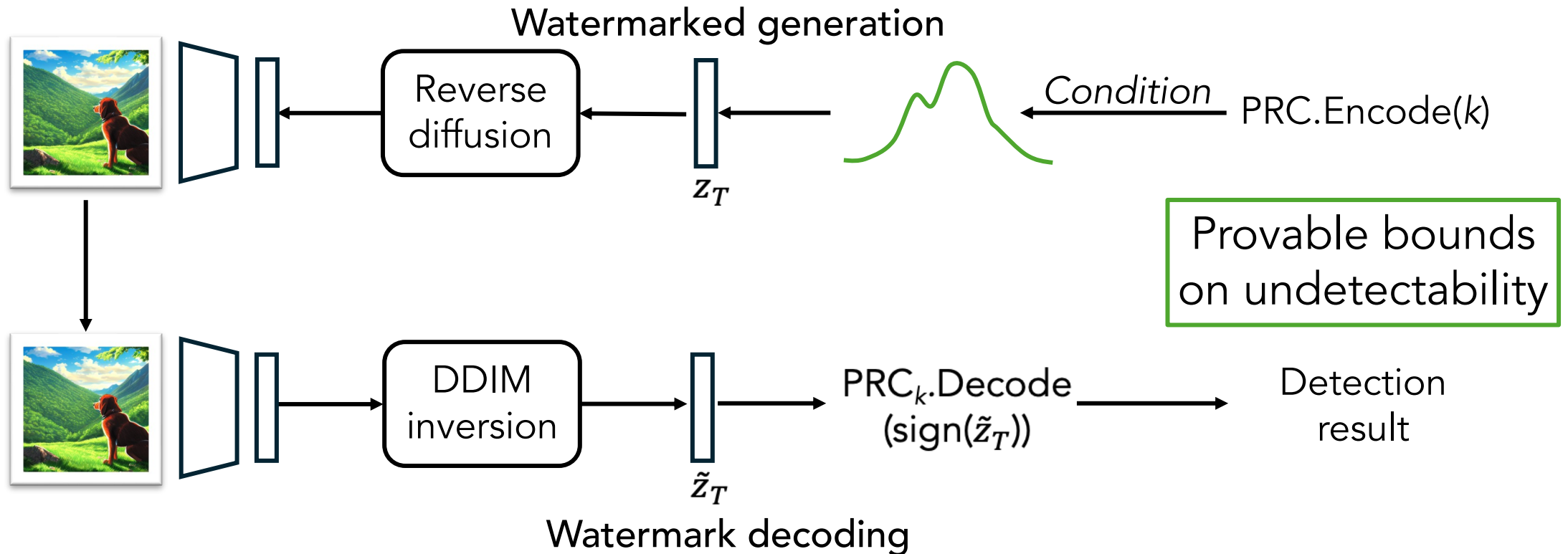
# [Tancik et al., 2020] Distortion Layer



# Semantic Watermarks

## Pseudorandom Error-Correcting Codes

PRC codewords indistinguishable from random without key [Gunn et al., 2025].



# Post-Hoc vs. Semantic Watermarks

## Post-Hoc with DNNs

Real or AI-generated media

Robustness "trained in"

Decode with DNN forward-pass

## Semantic

AI-generated media only

Robustness "out-of-the-box"

Decode requires DDIM inversion

---

**Both:** Security relies on encoder-decoder secret → No public verifiability

# [Qing et al. 2026] Results

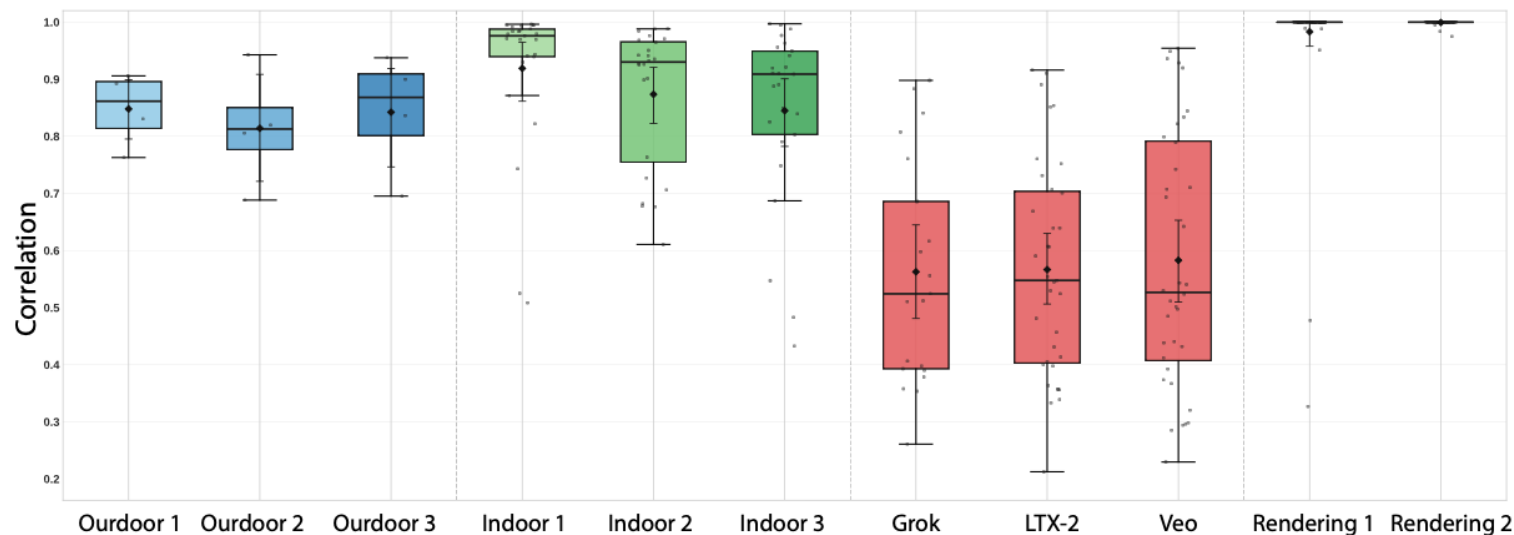


**Fig. 3:** Distribution of Pearson correlation coefficients  $|\rho|$  across all video categories. Real recordings (indoor and outdoor) cluster at high correlation, physics-based renderings achieve near-perfect correlation, and AI-generated videos concentrate at markedly lower values.

Main results confirming Moiré invariant validity.