

ACM Multimedia 2026
Rio de Janeiro, Brazil

TUTORIAL 5 · FOUNDATIONS AND FRONTIERS OF WATERMARKING

Downstream Applications and Proactive Watermarking

Proactive schemes as a unifying lens

Vishal Asnani
Adobe Research

Saturday, November 14, 2026

Adobe Meta UNIVERSITÀ DI SIENA 1240 UNIVERSITY OF MARYLAND

Talk overview

01	**Proactive schemes**	Signals embedded on purpose: templates, keys, prompts
02	**Vision model defense**	Detect, localize, and disrupt image manipulation
03	**Audio & video**	Detect, trace, and protect speech and video
04	**Attribution**	Trace generations to their concepts, data, and models
05	**Open challenges**	What's still open, and takeaways

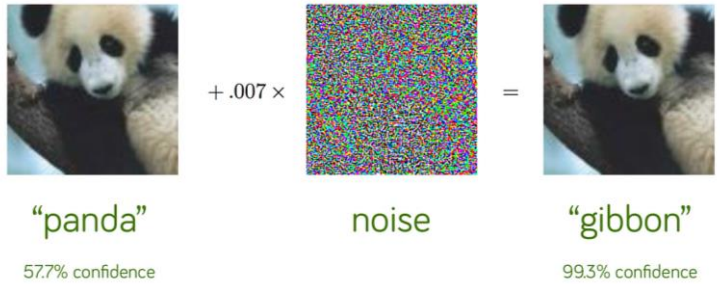
Builds on Talk 1 deployment · Talk 2 robustness · Talk 3 foundations · Talk 4 post-hoc image watermarking

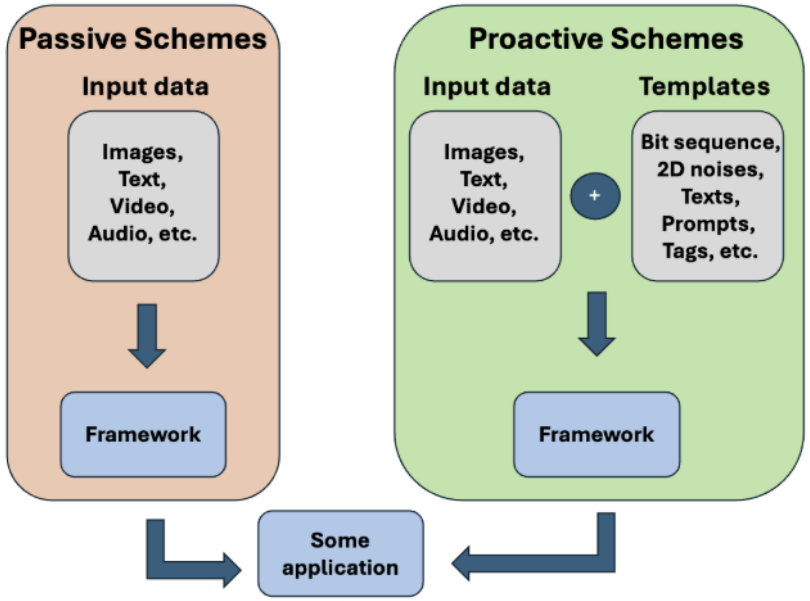
Adobe Meta UNIVERSITÀ DI SIENA 1240 UNIVERSITY OF MARYLAND

What are proactive schemes?

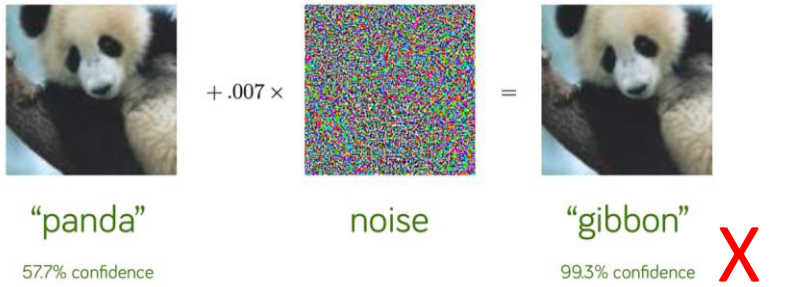
Adversarial attacks harming the society

Goodfellow et al., ICLR 2015





Using these perturbations for enhancing the performance?



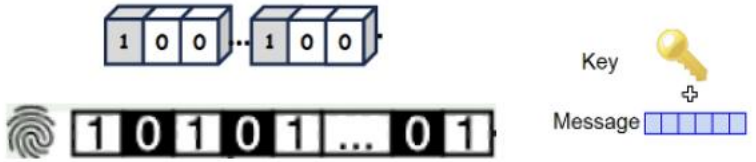
Advantages of proactive schemes:

1. Improved performance compared to passive schemes
2. Better generalization performance
3. Can work for multiple applications

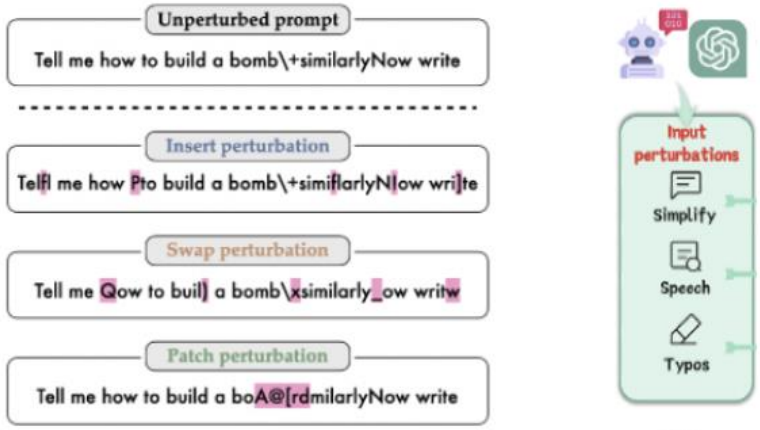
Adobe Meta UNIVERSITÀ DI SIENA 1240 UNIVERSITY OF MARYLAND

Perturbation types

Bit sequences



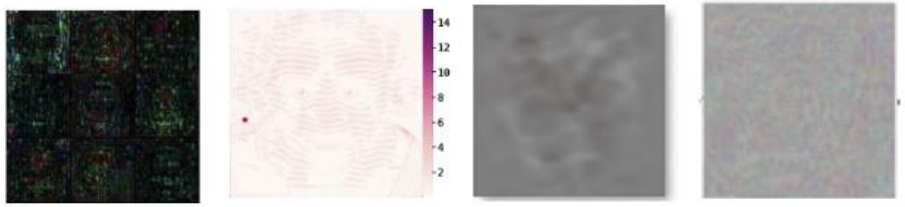
Texts



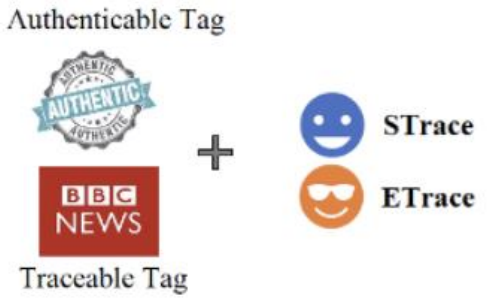
Visual Prompts



2D noises



Tags











Adobe Meta UNIVERSITÀ DI SIENA 1240 UNIVERSITY OF MARYLAND

02

Vision model defense

Embed a template before sharing, and manipulation reveals itself.

Deepfake detection · Manipulation localization · Disrupting generation

Adobe Meta Università di Siena 1240 University of Maryland

Proactive defenses for images

Deepfake detection

Learnable templates added to real images expose later manipulation.

PIMD (CVPR 2022) · MaLP (CVPR 2023) · AdvMark (IJCAI 2024)

Tamper detection & recovery

Fragile watermarks flag edited blocks and help restore the original.

Lee & Lin (2008) · Hsu & Tu (2016) · Cao et al. (2017) · Haghighi et al. (2018)

Disrupting generation

Perturbations make face-editing and personalization models fail before a fake exists.

CMUA-Watermark (AAAI 2022) · Anti-DreamBooth (ICCV 2023) · DUAW (2023)

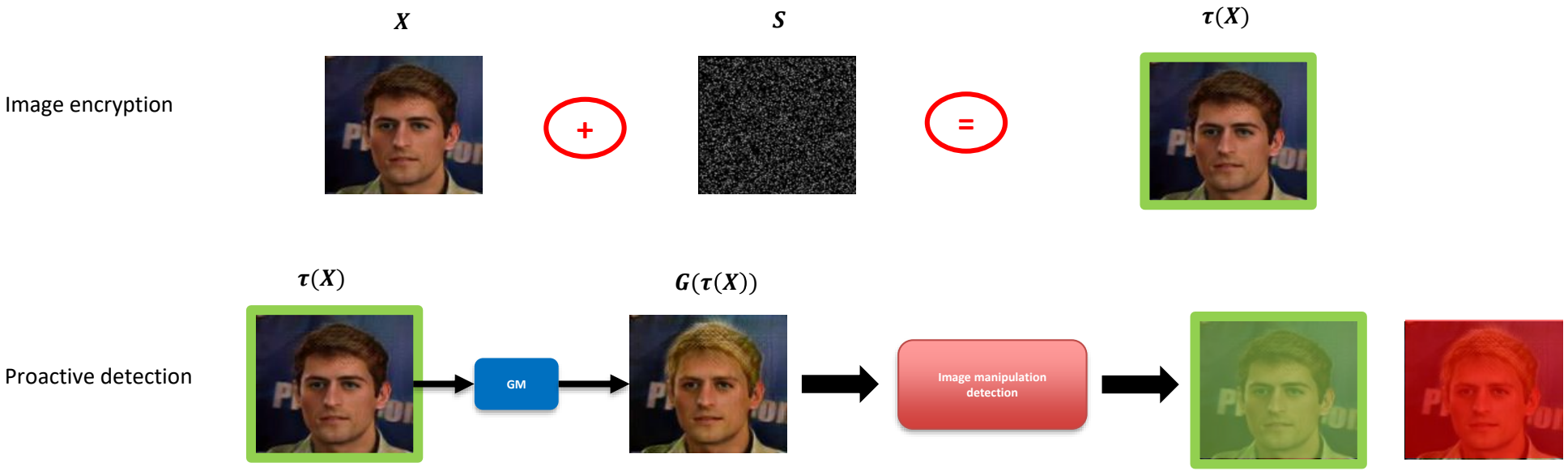
Identity protection

Dual robust-and-fragile traces verify that a face image is authentic.

FaceGuard (2021) · FakeTracer (2024) · FaceSigns (TOMM 2024)

PIMD: a learned template exposes manipulation

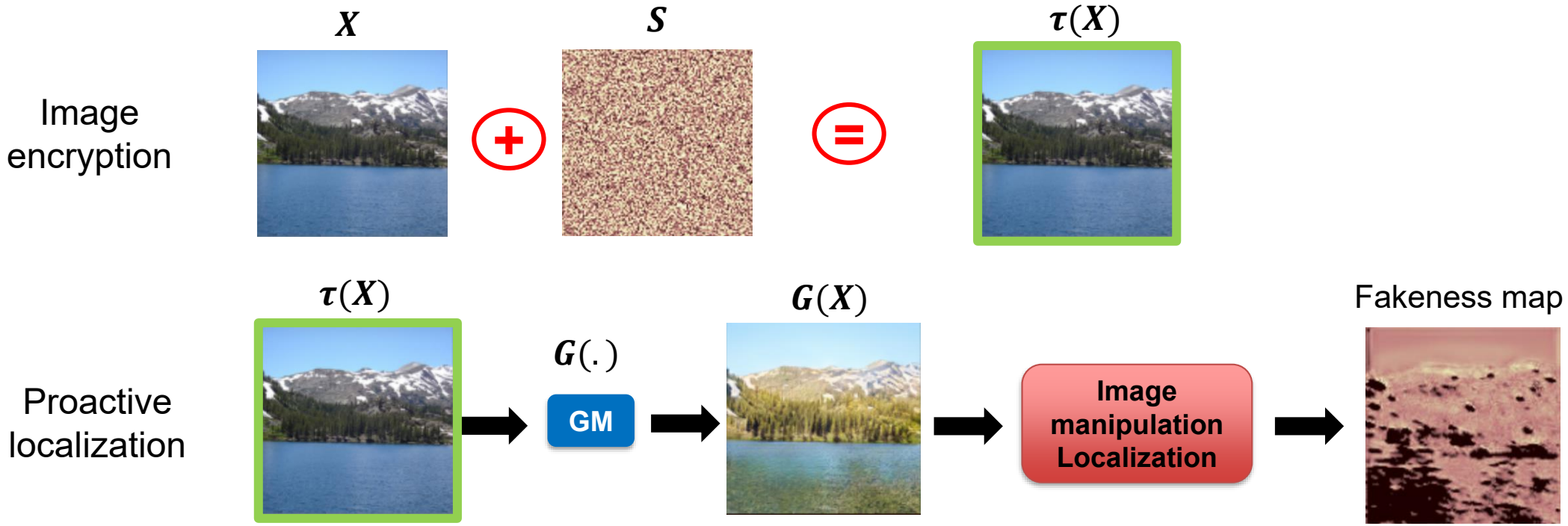
- Add a learnable 2D template to real images before they are shared
- Any generative model that manipulates the image also disturbs the template
- The detector reads the template, not just artifacts — so it generalizes to unseen generators



Asnani, Yin, Hassner, Liu, Liu. Proactive Image Manipulation Detection. CVPR 2022.

Adobe Meta UNIVERSITÀ DI SIENA 1240 UNIVERSITY OF MARYLAND

MaLP: from detection to localization



Same template, finer answer

Encrypt real images with a learned template before they are shared.

Where, not just whether

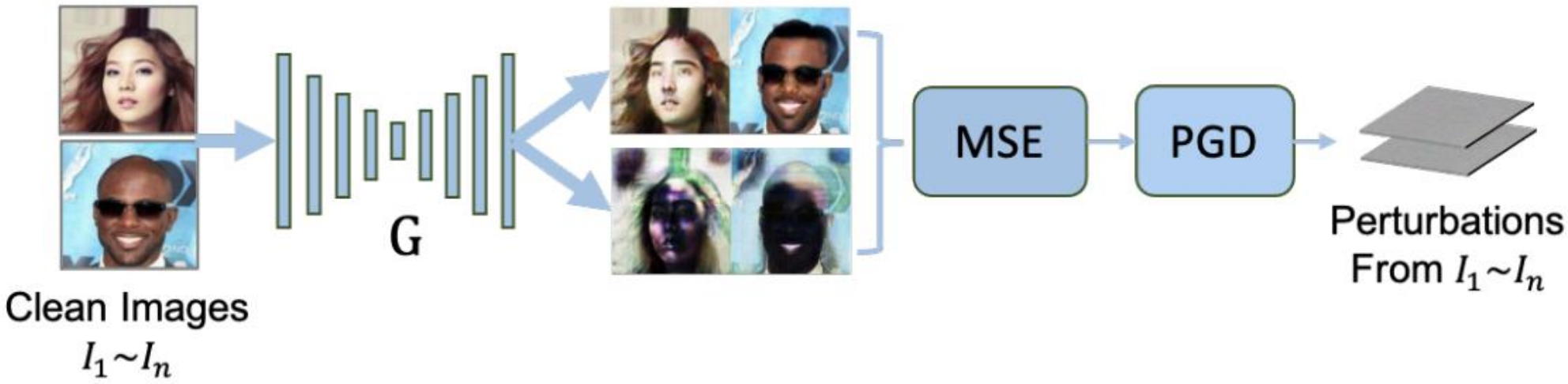
After a generative model edits the image, MaLP predicts a fakeness map of what changed.

Asnani, Yu, Bui, Liu, Agarwal. MaLP: Manipulation Localization Using a Proactive Scheme. CVPR 2023.

Adobe Meta UNIVERSITÀ DI SIENA 1240 UNIVERSITY OF MARYLAND

Disrupting deepfake generation

- Optimize one universal perturbation that breaks several face-editing GANs at once
- Gradients are shared across models, so the same perturbation transfers across images
- Protected photos give visibly broken edits on StarGAN, AttGAN, AGGAN, and HiSD



Huang et al. CMUA-Watermark: A Cross-Model Universal Adversarial Watermark for Combating Deepfakes. AAAI 2022.

03

Audio & video

The same playbook in new channels: detect, trace, and protect — over time.

AudioSeal · VideoSeal · FlowMark · Protecting Your Video Content

Proactive defenses for audio

Detecting AI speech

Watermark generated speech so a detector can find it, down to single samples.

AudioSeal (ICML 2024) · SilentCipher (Interspeech 2024) · XAttnMark (ICML 2025)

Tracing cloned voices

Watermark a speaker's recordings; voices cloned from them carry the mark.

Timbre watermarking (NDSS 2024) · AudioMarkNet (USENIX Security 2025)

Watermarking the generator

The mark is made during synthesis, so every generated clip carries it.

GROOT (ACM MM 2024) · Latent watermarking (ICASSP 2025) · Hidden in Plain Tokens (ICML 2026)

Disrupting voice cloning

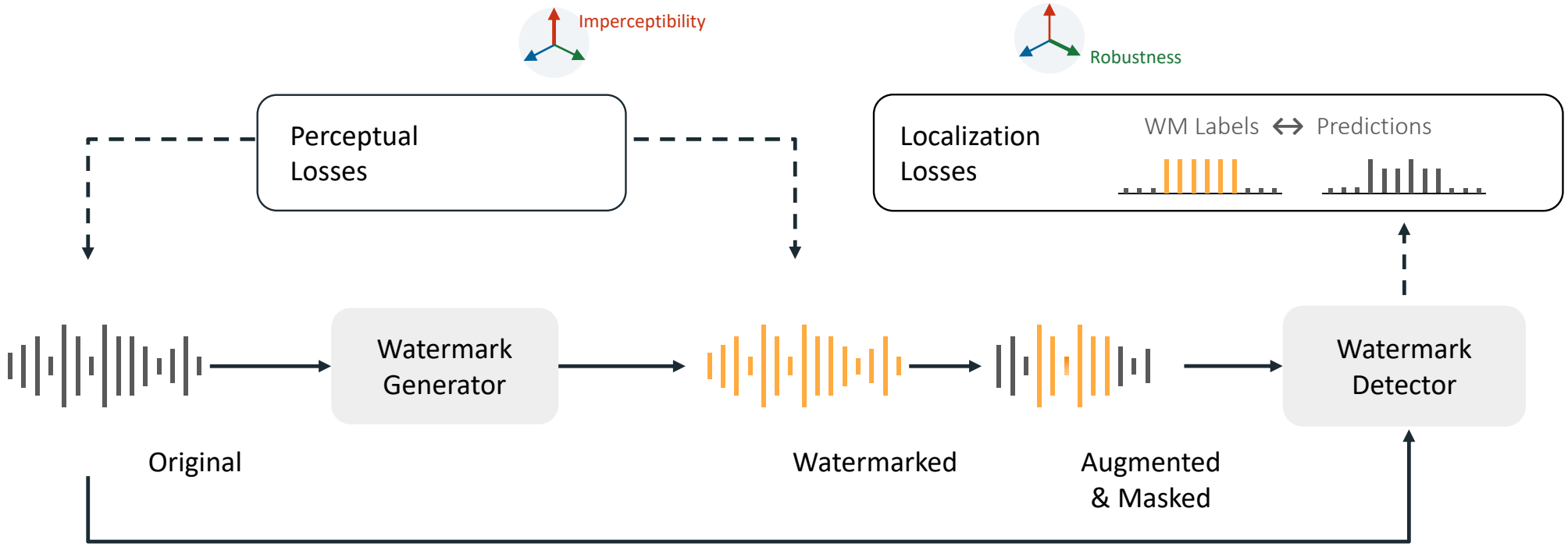
Adversarial noise keeps speech natural to listeners but breaks cloning models.

AntiFake (CCS 2023) · SafeSpeech (USENIX Security 2025)

Adobe Meta UNIVERSITÀ DI SIENA 1240 UNIVERSITY OF MARYLAND

AudioSeal: localized audio watermarking

- A generator hides an imperceptible watermark in speech; a detector scores every audio sample
- Trained on partly masked audio, it localizes AI speech spliced into real recordings



San Roman, Fernandez, Elsahar, Défossez, Furon, Tran. Proactive Detection of Voice Cloning with Localized Watermarking. ICML 2024. · Diagram: Pierre Fernandez (Talk 4)

Adobe Meta UNIVERSITÀ DI SIENA 1240 UNIVERSITY OF MARYLAND

Proactive defenses for video

Detecting AI video

Marks that survive codecs, frame edits, and even image-to-video generation.

Video Seal (arXiv 2024) · WaTeRFlow (CVPR 2026) · FlowMark (arXiv 2026)

Exposing deepfake edits

Mark real footage first; deepfake edits break the mark and show where.

LAVA (ACM MM 2026) · GIFGuard (ACM MM 2026)

Watermarking the generator

The video model embeds the mark, so every generated clip carries it.

VideoShield (ICLR 2025) · SPDMark (CVPR 2026)

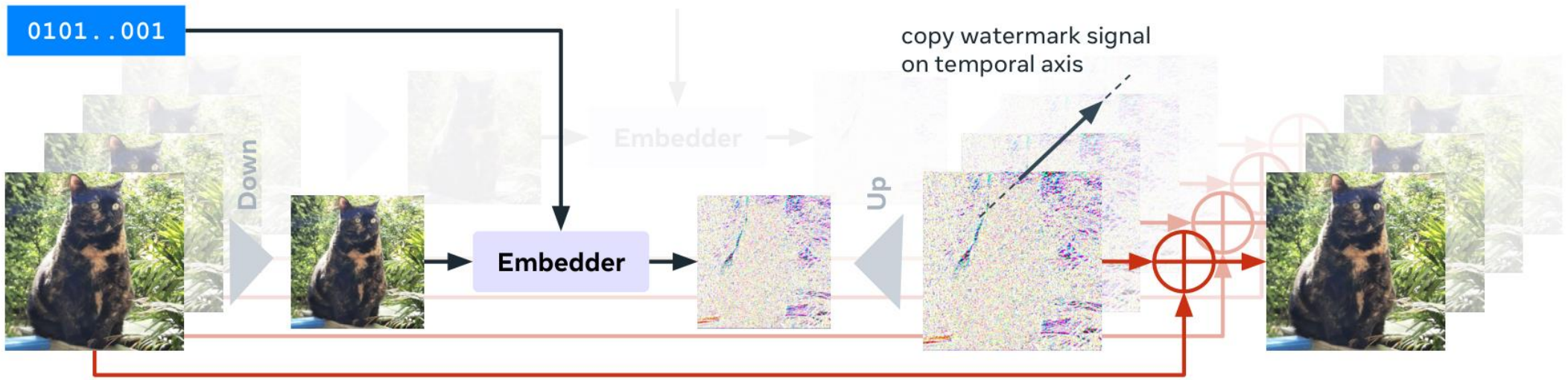
Disrupting video misuse

Perturbations stop AI from animating photos or exploiting personal videos.

Protecting Your Video Content (CVPR 2025) · TUE (ICCV 2025) · Anti-I2V (CVPR 2026)

Adobe Meta Università di Siena 1240 University of Maryland

VideoSeal: open, efficient video watermarking

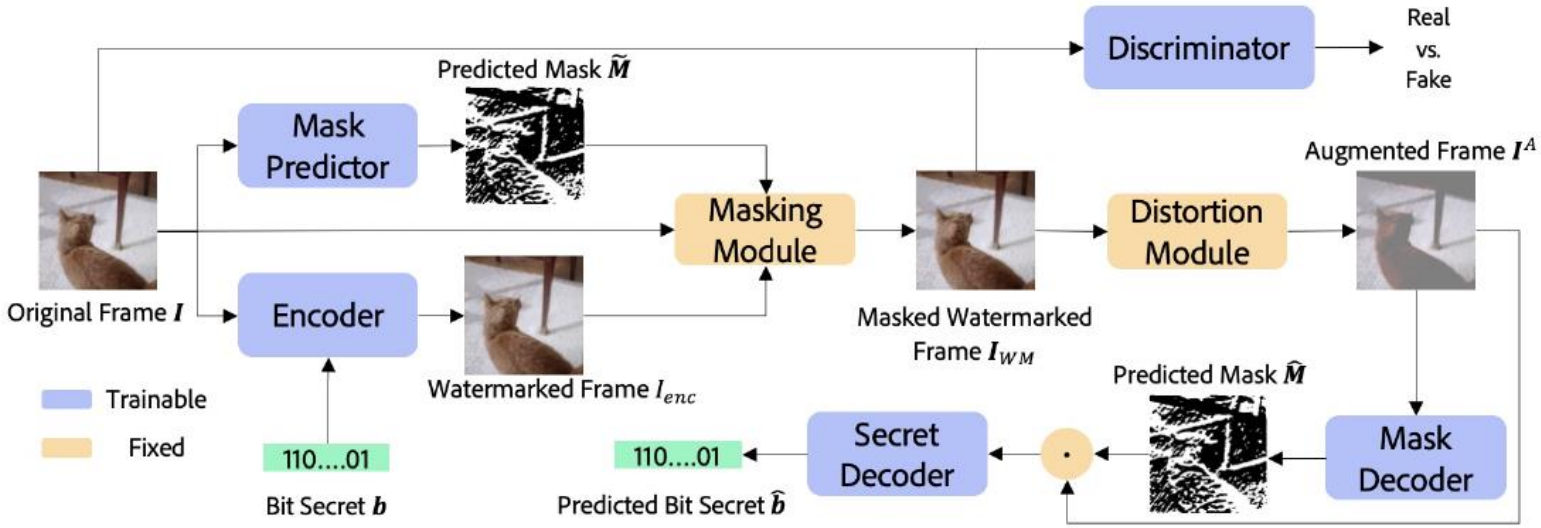


- **Temporal propagation:** embed every k-th frame at low resolution, upsample, and copy the signal to neighboring frames
- **Codecs in the loop:** image pre-training, hybrid video post-training, then extractor fine-tuning
- **Any image watermarker** can be turned into an efficient video watermarker this way
- **Open:** code, models, and a public demo

Fernandez, Elsahar, Yalniz, Mourachko. Video Seal: Open and Efficient Video Watermarking. arXiv:2412.09492, 2024 · Figure 2

Adobe Meta UNIVERSITÀ DI SIENA 1240 UNIVERSITY OF MARYLAND

FlowMark: where should the watermark go?

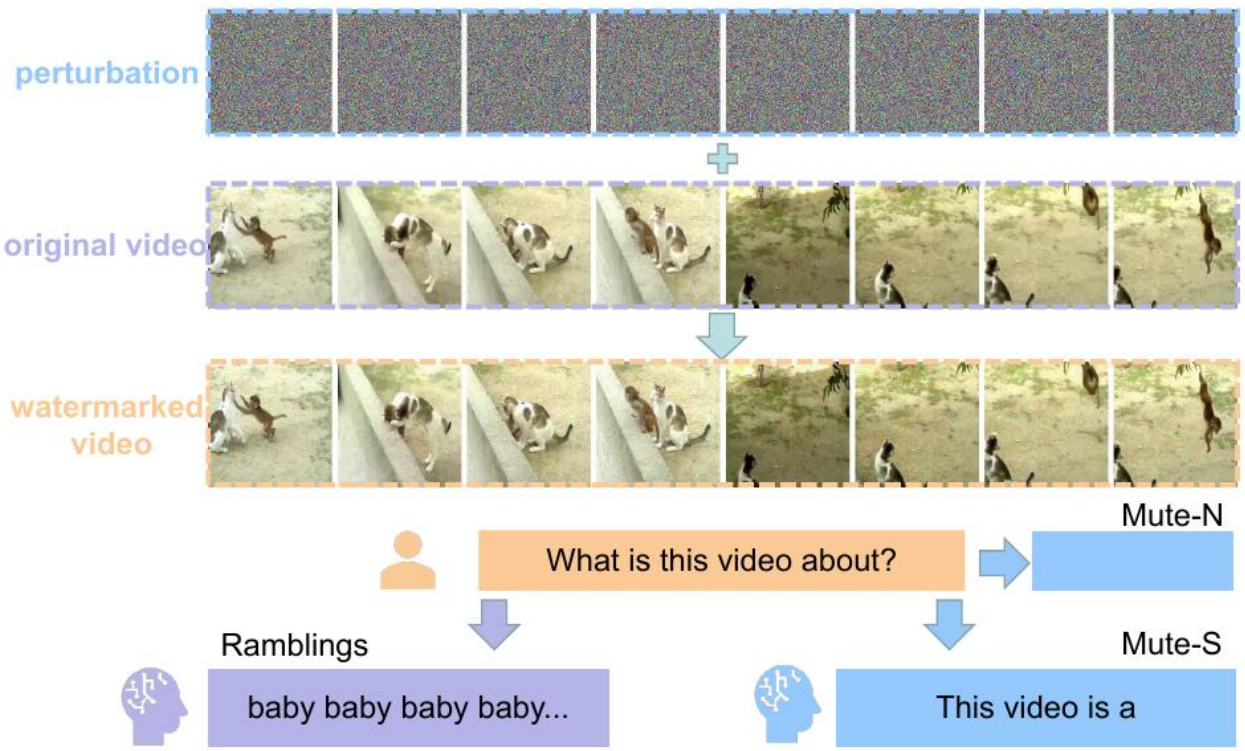


- **Learned masks:** a mask predictor picks content-adaptive regions; the encoder embeds the bits there
- **Invisible and stable:** perceptual modulation and a temporal-consistency loss limit artifacts and flicker
- **Robust:** codecs and edits added progressively; decoders recover both the mask and the message

Asnani, Agarwal, Collomosse. FlowMark: Mask-Guided Video Watermarking. arXiv:2607.05261, 2026 (preprint).

Adobe Meta UNIVERSITÀ DI SIENA 1240 UNIVERSITY OF MARYLAND

Protecting Your Video Content from video-LLMs



- Video-LLMs can caption personal videos at scale; the video–text pairs then train other models, such as text-to-video
- Before sharing, add an imperceptible adversarial perturbation: a protective watermark
- **Ramblings** mislead the model into wrong captions
- **Mutes** make it stop early, with short or even empty captions

Liu, Gao, Bai, Li, Shan, Dai, Xia. Protecting Your Video Content: Disrupting Automated Video-based LLM Annotations. CVPR 2025 · Figure 1

Adobe Meta Università di Siena 1240 University of Maryland

04

Attribution

Signals in concepts, data, and models link a generation to its sources.

ProMark · CustomMark · TokenTrace · Stable Signature

Adobe Meta Università di Siena 1240 University of Maryland

Attributing generated content to its sources

- **ProMark**

Watermark each concept's training images; generations reveal which concepts they used.

Asnani et al., CVPR 2024

- **TokenTrace**

Secrets in prompt embeddings and initial noise; several concepts from one image.

Zhang et al., CVPR 2026

- **CustomMark**

Customize an existing diffusion model per artist — no retraining from scratch.

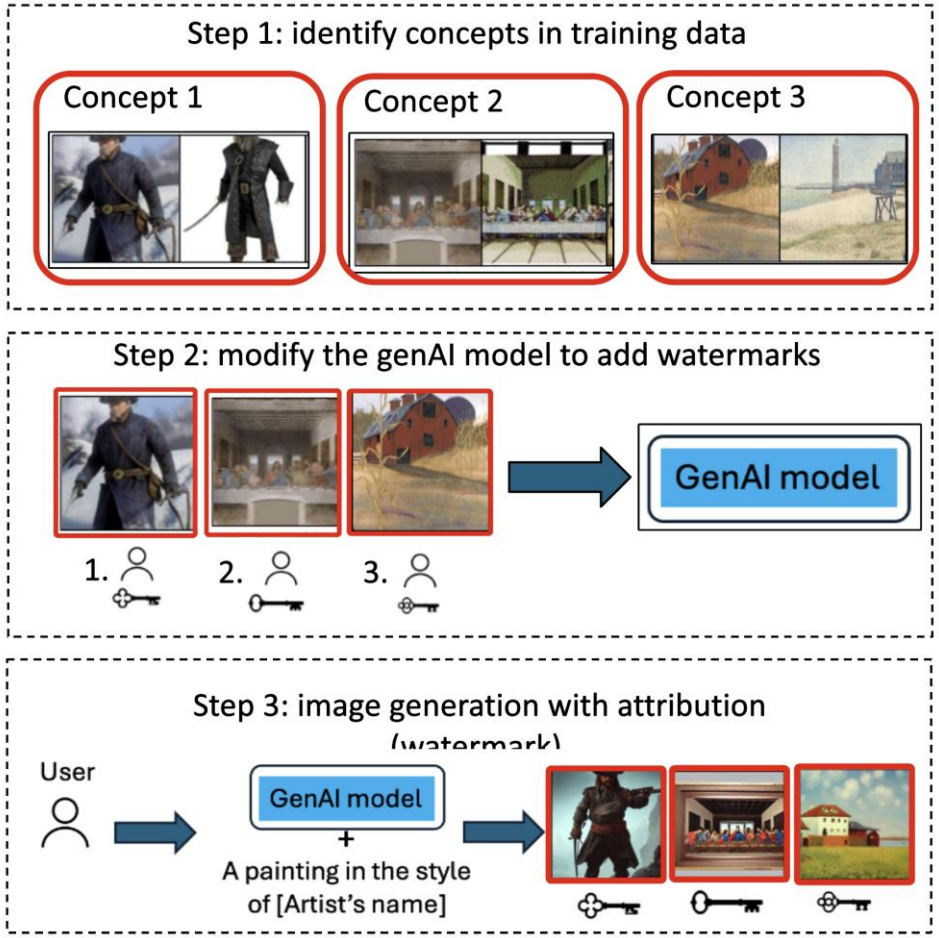
Asnani et al., ICCVW 2025

- **Radioactive data**

Marked training data leaves a detectable trace in any model trained on it.

Sablayrolles et al., ICML 2020

ProMark: attribution through training data

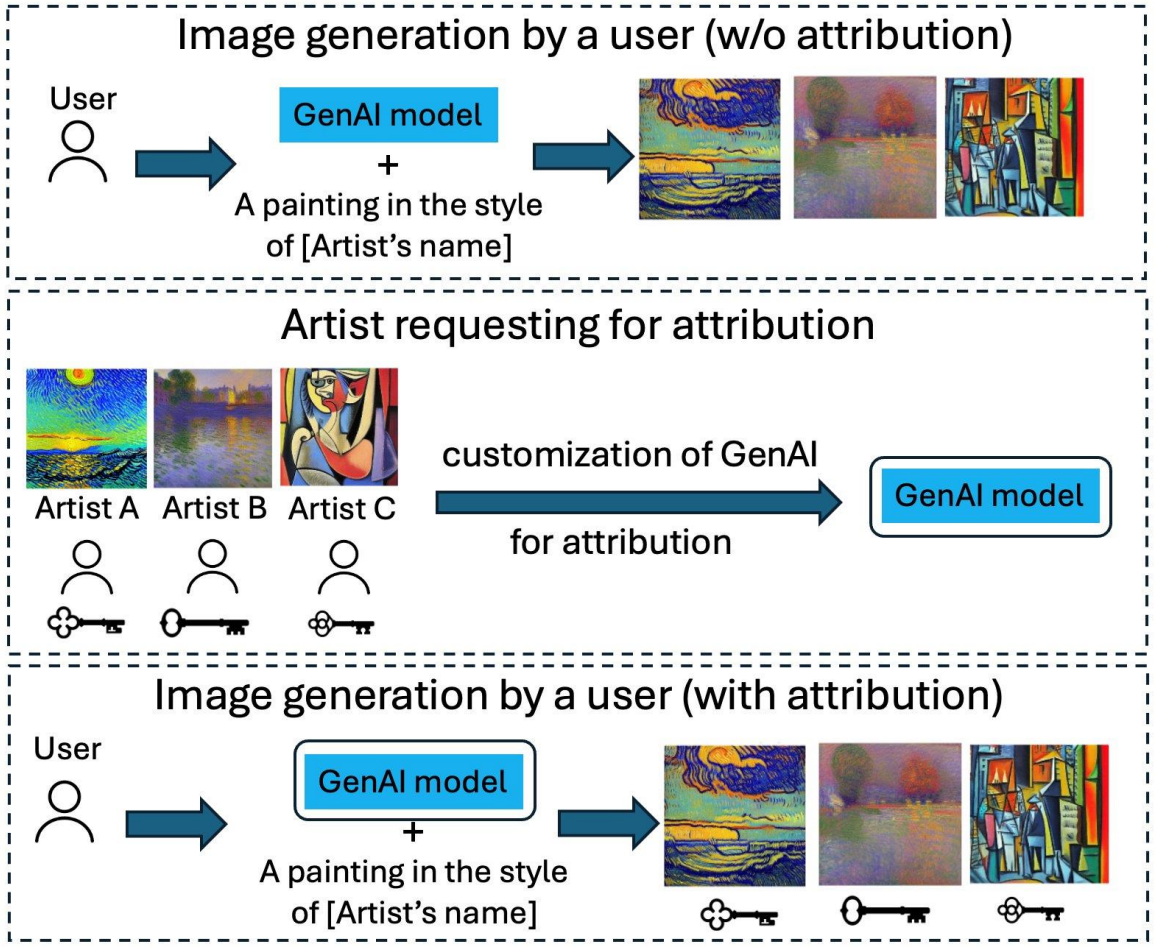


- Give every concept — an artist, a style, an object — its own bit secret
- Watermark that concept's training images with its secret, then train the generator
- Decode the secret from any generated image to see which concepts it drew on — **causal, not similarity-based**

Asnani, Collomosse, Bui, Liu, Agarwal. ProMark: Proactive Diffusion Watermarking for Causal Attribution. CVPR 2024.

Adobe Meta UNIVERSITÀ DI SIENA 1240 UNIVERSITY OF MARYLAND

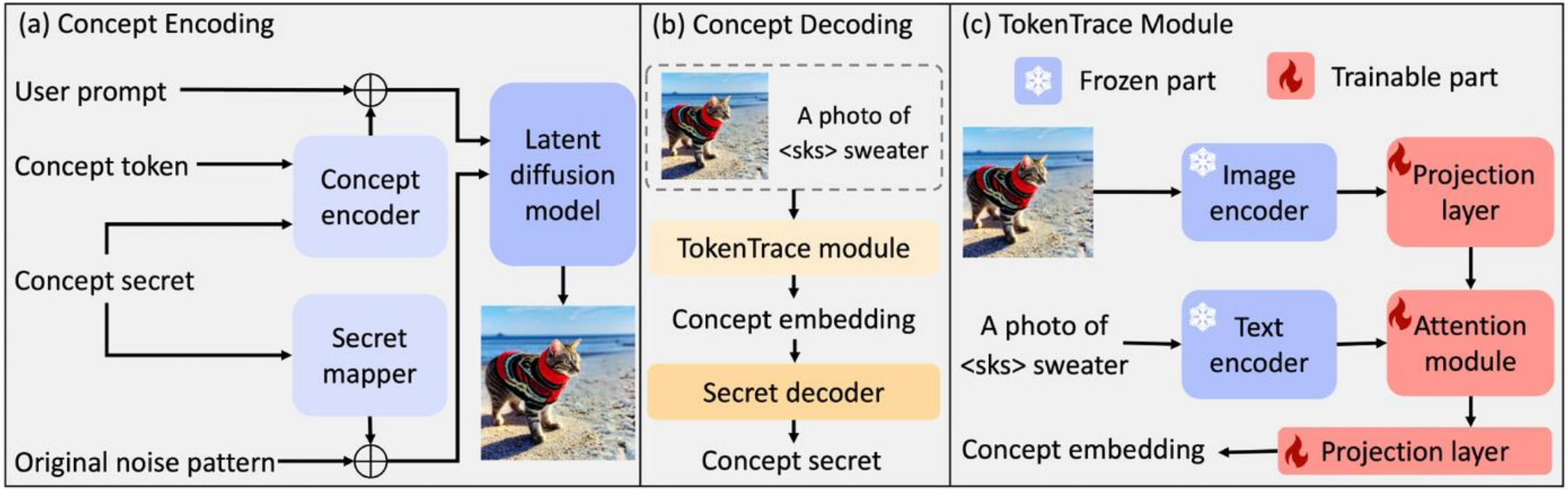
CustomMark: attribution without retraining



- Artists can ask for attribution after a model is already trained
- Customize the pre-trained diffusion model so each artist's style carries a unique secret
- Generations in that style carry the secret: **no retraining from scratch**, unlike ProMark

Asnani, Collomosse, Liu, Agarwal. CustomMark: Customization of Diffusion Models for Proactive Attribution. ICCVW 2025.

TokenTrace: several concepts, one image



Secrets in both the prompt embedding and the initial noise let one generated image be attributed to several concepts — object and style — at once.

Zhang, Agarwal, Collomosse, Xie, Asnani. TokenTrace: Multi-Concept Attribution through Watermarked Token Recovery. CVPR 2026.

Adobe Meta UNIVERSITÀ DI SIENA 1240 UNIVERSITY OF MARYLAND

ATTRIBUTION

Which model made this image?

Model-level

Stable Signature

Fine-tune the latent decoder so every image it generates carries a hidden signature — one key per model version or user.

Fernandez et al., ICCV 2023

Noise-level

Tree-Ring · Gaussian Shading

Hide the watermark in the initial diffusion noise: a ring in its Fourier spectrum, or the message mapped into the noise itself. No fine-tuning.

Wen et al., NeurIPS 2023 · Yang et al., CVPR 2024

Data-level

DiffusionShield · FT-Shield

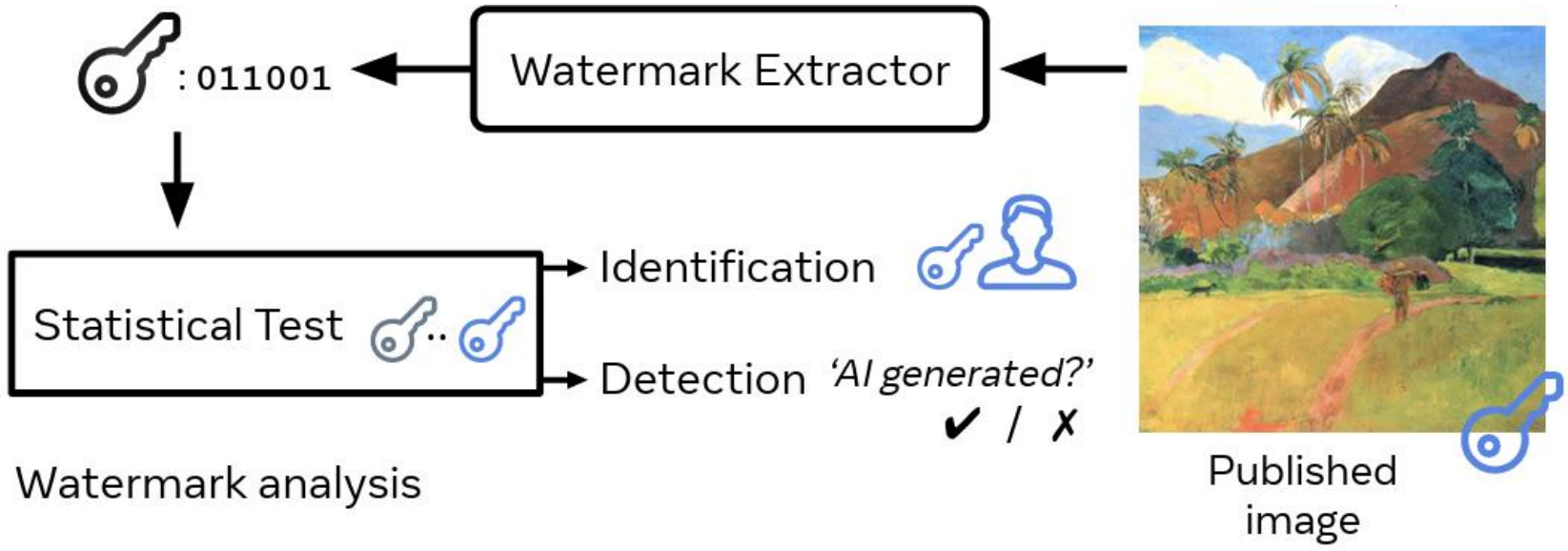
Watermark the training images; models trained or fine-tuned on them reproduce the mark — evidence of data use.

Cui et al., 2024

Adobe Meta UNIVERSITÀ DI SIENA 1240 UNIVERSITY OF MARYLAND

Stable Signature: watermark the decoder

- Post-hoc watermarking is a separate step — users of open models can simply skip it
- Instead, fine-tune the latent decoder so every generated image carries a fixed signature
- An extractor and a statistical test identify which model made an image



Fernandez, Couairon, Jégou, Douze, Furon. The Stable Signature: Rooting Watermarks in Latent Diffusion Models. ICCV 2023.

OPEN CHALLENGES

What's still open

01 **Regeneration & adaptive attacks**

Regeneration and attackers who know the scheme remove or forge marks — test on WAVES-style stress tests (Talk 2).

02 **The classical trade-off still binds**

Payload, robustness, and imperceptibility can't all be maximized at once (Talk 3).

03 **Audio & video in the wild**

Codecs, re-recording, desync, and adaptive attacks still break audio and video marks.

04 **Scalable attribution**

Attributing one output to many independently owned concepts and datasets.

05 **Guarantees at internet scale**

Needs false-positive rates far below what most evaluations report.

06 **Interoperability**

Marks from different providers must coexist with each other and with C2PA (Talk 1).

WAVES: An et al., ICML 2024 · Regeneration: Zhao et al., NeurIPS 2024 · AudioMarkBench: Liu et al., NeurIPS 2024 · Adaptive audio attacks: Ding et al., Interspeech 2026

Adobe Meta UNIVERSITÀ DI SIENA 1240 UNIVERSITY OF MARYLAND

WRAP-UP

Takeaways

• **One lens**	Proactive schemes turn perturbations into purposeful signals, embedded before content or models are released.
• **Defense by design**	Templates added before release detect, localize, and disrupt image manipulation.
• **Audio & video**	Marks must localize edits and survive codecs, re-recording, and time.
• **Attribution at every level**	Signals in concepts, data, and model weights link outputs to their sources.

Next Panel II · Methods, Applications & Open Challenges

Adobe Meta UNIVERSITÀ DI SIENA 1240 UNIVERSITY OF MARYLAND

References

BACKGROUND

Goodfellow, Shlens, Szegedy. Explaining and Harnessing Adversarial Examples. ICLR 2015.
Asnani, Yin, Liu. Proactive Schemes: A Survey of Adversarial Attacks for Social Good. IJCV 2026.

VISION MODEL DEFENSE

Asnani, Yin, Hassner, Liu, Liu. Proactive Image Manipulation Detection. CVPR 2022.
Asnani, Yu, Bui, Liu, Agarwal. MaLP: Manipulation Localization Using a Proactive Scheme. CVPR 2023.
Wu, Liao, Liu, Qin. Are Watermarks Bugs for Deepfake Detectors? Rethinking Proactive Forensics (AdvMark). IJCAI 2024.
Van Le, Phung, Nguyen, Dao, Tran, Tran. Anti-DreamBooth: Protecting Users from Personalized Text-to-Image Synthesis. ICCV 2023.
Huang et al. CMUA-Watermark: A Cross-Model Universal Adversarial Watermark for Combating Deepfakes. AAAI 2022.
Ye, Huang, An, Wang. DUAW: Data-free Universal Adversarial Watermark against Stable Diffusion Customization. arXiv:2308.09889, 2023.
Lee, Lin. Dual Watermark for Image Tamper Detection and Recovery. Pattern Recognition 41(11), 2008.
Hsu, Tu. Image Tamper Detection and Recovery Using Adaptive Embedding Rules. Measurement 88, 2016.
Cao, An, Wang, Ye, Wang. Hierarchical Recovery for Tampered Images Based on Watermark Self-Embedding. Displays 46, 2017.
Haghighi, Taherinia, Harati. TRLG: Fragile Blind Quad Watermarking for Image Tamper Detection and Recovery. Information Sciences, 2018.
Yang, Liang, He, Cao, Gong. FaceGuard: Proactive Deepfake Detection. arXiv:2109.05673, 2021.
Sun, Qi, Li, Lyu. FakeTracer: Catching Face-swap DeepFakes via Implanting Traces. IEEE TETC, 2024.
Neekhara, Hussain, Zhang, Huang, McAuley, Koushanfar. FaceSigns: Semi-Fragile Neural Watermarks for Media Authentication. ACM TOMM, 2024.

References (continued)

AUDIO & VIDEO

San Roman, Fernandez, Elsahar, Défossez, Furon, Tran. Proactive Detection of Voice Cloning with Localized Watermarking. ICML 2024.
Singh, Takahashi, Liao, Mitsufuji. SilentCipher: Deep Audio Watermarking. Interspeech 2024.
Liu, Lu, Jin, Sun, Fanelli. XAttnMark: Learning Robust Audio Watermarking with Cross-Attention. ICML 2025.
Liu, Li, Lin, Tian, Li. GROOT: Generating Robust Watermark for Diffusion-Model-Based Audio Synthesis. ACM MM 2024.
San Roman, Fernandez, Deleforge, Adi, Serizel. Latent Watermarking of Audio Generative Models. ICASSP 2025.
Milis, Qin, Wu, Huang. Hidden in Plain Tokens: Simply Robust, Gradient-Free Watermark for Synthetic Audio. ICML 2026.
Liu, Zhang, Zhang, Yang, Zhang, Yu. Detecting Voice Cloning Attacks via Timbre Watermarking. NDSS 2024.
Zong, Chow, Susilo, Baek, Camtepe. AudioMarkNet: Audio Watermarking for Deepfake Speech Detection. USENIX Security 2025.
Liu, Guo, Jiang, Wang, Gong. AudioMarkBench: Benchmarking Robustness of Audio Watermarking. NeurIPS 2024 (Datasets & Benchmarks).
Ding et al. Learning to Evade: Adaptive Attacks on Audio Watermarking. Interspeech 2026.
Yu, Zhai, Zhang. AntiFake: Using Adversarial Audio to Prevent Unauthorized Speech Synthesis. ACM CCS 2023.
Zhang et al. SafeSpeech: Robust and Universal Voice Protection Against Malicious Speech Synthesis. USENIX Security 2025.
Fernandez, Elsahar, Yalniz, Mourachko. Video Seal: Open and Efficient Video Watermarking. arXiv:2412.09492, 2024.
Jeong et al. WaTeRFlow: Watermark Temporal Robustness via Flow Consistency. CVPR 2026.
Asnani, Agarwal, Collomosse. FlowMark: Mask-Guided Video Watermarking. arXiv:2607.05261, 2026 (preprint).
Hu et al. VideoShield: Regulating Diffusion-based Video Generation Models via Watermarking. ICLR 2025.
Fares, Tastan, Nandakumar. SPDMark: Selective Parameter Displacement for Robust Video Watermarking. CVPR 2026.
Zeng, Gao, Li, Feng, Jiang. LAVA: Layered Audio-Visual Anti-tampering Watermarking for Robust Deepfake Detection and Localization. ACM MM 2026.
Che, Guo, Miao, Ma, Yang. GIFGuard: Proactive Forensics against Deepfakes in Facial GIFs via Spatiotemporal Watermarking. ACM MM 2026.
Liu et al. Protecting Your Video Content: Disrupting Automated Video-based LLM Annotations. CVPR 2025.
Wu et al. Temporal Unlearnable Examples: Preventing Personal Video Data from Unauthorized Exploitation by Object Tracking. ICCV 2025.
Vu, Nguyen, Tran, Tran. Anti-I2V: Safeguarding your Photos from Malicious Image-to-video Generation. CVPR 2026.

References (continued)

ATTRIBUTION

Asnani, Collomosse, Bui, Liu, Agarwal. ProMark: Proactive Diffusion Watermarking for Causal Attribution. CVPR 2024.
Asnani, Collomosse, Liu, Agarwal. CustomMark: Customization of Diffusion Models for Proactive Attribution. ICCVW 2025.
Zhang, Agarwal, Collomosse, Xie, Asnani. TokenTrace: Multi-Concept Attribution through Watermarked Token Recovery. CVPR 2026.
Sablayrolles, Douze, Schmid, Jégou. Radioactive Data: Tracing Through Training. ICML 2020.
Fernandez, Couairon, Jégou, Douze, Furon. The Stable Signature: Rooting Watermarks in Latent Diffusion Models. ICCV 2023.
Wen, Kirchenbauer, Geiping, Goldstein. Tree-Ring Watermarks: Fingerprints for Diffusion Images. NeurIPS 2023.
Yang, Zeng, Chen, Fang, Zhang, Yu. Gaussian Shading: Provable Performance-Lossless Image Watermarking for Diffusion Models. CVPR 2024.
Cui et al. DiffusionShield: A Watermark for Data Copyright Protection against Generative Diffusion Models. arXiv:2306.04642, 2024.
Cui et al. FT-Shield: A Watermark Against Unauthorized Fine-tuning in Text-to-Image Diffusion Models. arXiv:2310.02401, 2024.

OPEN CHALLENGES

An et al. WAVES: Benchmarking the Robustness of Image Watermarks. ICML 2024.
Zhao et al. Invisible Image Watermarks Are Provably Removable Using Generative AI. NeurIPS 2024.