CHAIN OF EVIDENCE: LOCALIZING VIDEO FORGERIES VIA IMAGE FORENSIC EVIDENCE IN SPACE AND TIME

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion-based video editing produces photorealistic forgeries that leave subtle forensic traces. These forgeries are often confined to small regions within a few frames, requiring precise localization across both space and time. Existing video forgery research has mainly focused on video- or frame-level detection, while joint spatial and temporal localization, which is required to verify footage used in news reporting or as legal evidence, has received less attention. Image forgery models provide spatial forensic evidence, but frame-wise inference cannot exploit temporal context to refine localization. We present Chain of Evidence, a parameter-efficient framework that extends perturbation-based diffusion forgery localization from images to video on a frozen SAM 3 backbone. The model learns spatial forensic evidence from images, and a lightweight 8.4M-parameter causal memory carries it across frames so that preceding context strengthens weak or ambiguous cues. Training on video updates only the causal memory and the presence head, leaving the image-trained forensic representation intact. We also propose a protocol that scores space and time jointly. Under it, chaining evidence across frames more than doubles spatiotemporal IoU over framewise inference on the out-of-distribution AINPAINT, from 0.162 to 0.420. On image benchmarks that no method was trained on, Chain of Evidence reaches an average localization IoU of 0.299, against 0.218 for the strongest baseline. On TVIL, it reaches a temporal F1 of 0.370, against 0.259 for the strongest baseline. Code is available at the anonymous repository.

1 INTRODUCTION

Forgery models, increasingly diffusion-based, have made realistic video editing easy, with edits that stay consistent across frames (Geyer et al., 2024; Jiang et al., 2025; Feng et al., 2026). Along with creative potential, such capabilities carry risks of misuse. A malicious actor can insert a compromising object into a few frames of an authentic recording. Removing an object from the same frames can hide evidence. Such forgeries can spread misinformation, falsify evidence and cause reputational harm (Europol, 2022). Video forgery detection and localization addresses these threats. Its objective is to decide whether a video is forged and which frames and pixels are forged (Sec. 3.1).

Localizing forged regions jointly in space and time is the central challenge (Fig. 1). Image localizers learn the pixel traces that editing leaves in a single image (Guillaro et al., 2023; Zhu et al., 2025; Yu et al., 2026; Lek et al., 2026). Applied frame by frame, they therefore ignore all other frames. Video localizers predict forged segments or per-frame masks (Nguyen et al., 2024; Zhang et al., 2023b; Lou et al., 2025; Huang et al., 2026), but most learn their evidence from few video datasets (Sec. 2). We argue that a critical issue is overlooked: published protocols score either masks on forged frames (Lou et al., 2025; Huang et al., 2026; Montibeller et al., 2026) or frames without pixels (Zhang et al., 2023b; Feng et al., 2026), never space and time together (Sec. 3.1).

Modern localizers learn from automatically generated forgeries (Costanzino et al., 2025; Huang et al., 2025). Localizers trained on some editors lose accuracy on editors outside their training data (Wang et al., 2025; Li et al., 2025). Training data must therefore cover many editors. Image forgery data is far larger than video data and spans many editors (Costanzino et al., 2025; Huang

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

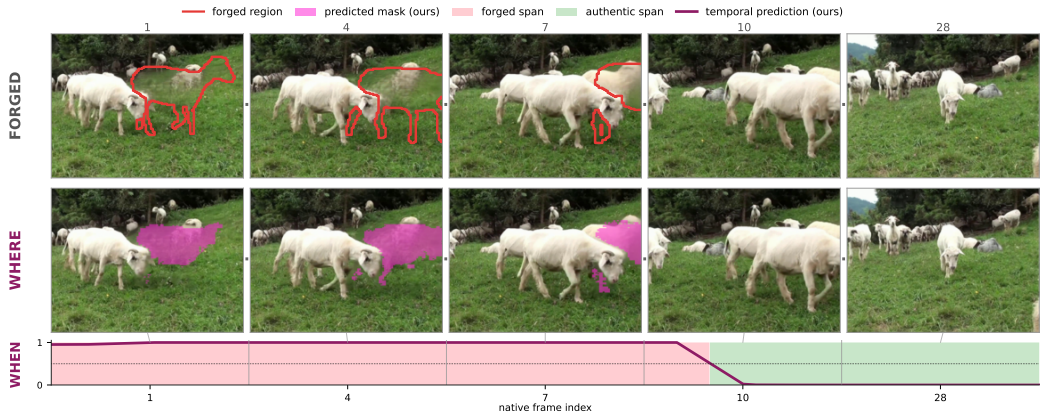


Figure 1: **Chain of Evidence localizes a forgery in space and time.** In this FakeParts clip (Liu et al., 2025), the flow-guided inpainter E2FGVI (Li et al., 2022) removes a sheep from the first ten frames and leaves the rest authentic. FORGED: the removed region outlined in red. WHERE: our mask in magenta. WHEN: our per-frame presence probability in plum, with the 0.5 threshold dotted. It stays high over the forged span in pink and drops at the first authentic frame in green.

et al., 2025; Wang et al., 2025; Li et al., 2025). Few video datasets pair per-frame pixel masks with authentic footage (Liu et al., 2025; Montibeller et al., 2026). Other video benchmarks label frames only (Zhang et al., 2023b; Feng et al., 2026; Bao et al., 2026). This imbalance favors learning forensic evidence on images and carrying it into video through architectural design. To address these challenges, we propose Chain of Evidence, which detects and localizes forgeries in images and video. It builds on frozen foundation features, which generalize across generators in detection (Ojha et al., 2023). It follows two main principles: (1) learning forensic evidence on diverse images and (2) carrying that evidence through time without relearning it.

Our contributions can be summarized as follows.

- **Evidence-guided forgery localization.** The object boundaries of a foundation segmenter do not reveal a forgery by themselves. Our surprising finding is that a frozen SAM 3 localizes forgeries once forensic evidence guides it. We derive this evidence from the sensitivity of the frozen features to added Gaussian noise. An evidence transformer turns this sensitivity into multiscale evidence. We use the evidence to refine the frozen features, to generate the mask prompt of SAM 3 and to detect forged frames. On the image benchmarks that no method was trained on, Chain of Evidence reaches an average localization IoU of 0.299, against 0.218 for the strongest baseline (Sec. 4.2).
- **Causal evidence propagation.** A frame where the forgery is faint can draw on evidence from earlier frames. We add a causal memory bank that refines each frame from seven preceding frames. We adapt the memory attention of video object segmentation to carry forensic evidence instead of object identity. We train only the memory bank and presence head on video. Video training raises spatiotemporal IoU on the out-of-distribution AINPAINT from 0.162 to 0.420 (Sec. 4.2).
- **Gated spatiotemporal evaluation.** A useful localizer must find forged pixels in forged frames and return empty masks elsewhere. We score detection, temporal, spatial and spatiotemporal localization under one protocol that gates every mask by the frame decision (Sec. 3.1). We compare with ten image and video localizers on seven image and five video benchmarks. On TVIL, Chain of Evidence reaches a temporal F1 of 0.370, against 0.259 for the strongest baseline (Sec. 4.2).

2 Related Work

Image forgery localization. Trace-based localizers fuse low-level artifacts with semantic features (Guillaro et al., 2023; Zhu et al., 2025) but lose much of their accuracy on diffusion editors absent from their training data (Wang et al., 2025). Frozen foundation features generalize better across generators in detection (Ojha et al., 2023), and DINOv3-IML adapts a frozen DINOv3 to localization (Yu et al., 2026). Detectors of generated images score a whole image by its reconstruction error

(Wang et al., 2023; Ricker et al., 2024) or by the sensitivity of a frozen DINOv2 embedding to added noise (He et al., 2024). This sensitivity transfers across generators (He et al., 2024) but depends on the encoder and the image domain (Tsai et al., 2024). SAM-based localizers adapt promptable segmenters (Kirillov et al., 2023; Ravi et al., 2025; Carion et al., 2026) to forensic cues, because object and forgery boundaries need not coincide (Kwon et al., 2025; Lek et al., 2026). Detective SAM corrects frozen SAM 2 features with pointwise adapters on blurred and noisy image copies and derives a dense mask prompt (Lek et al., 2026). Unlike its per-scale adapters, our evidence transformer fuses all scales through global attention, and its evidence also drives a presence head.

Video forgery localization. DeMamba classifies whole videos as authentic or fully generated (Chen et al., 2026), and UMMAFormer and MSLoc localize forged intervals in time (Zhang et al., 2023b; Feng et al., 2026). VideoFACT predicts per-frame masks but processes frames independently (Nguyen et al., 2024). TruVIL learns spatiotemporal features from frames and their noise residuals (Lou et al., 2025), and AINPAINT-ST fuses adapted DINOv2 features with optical-flow anomalies (Montibeller et al., 2026). RelayFormer relays global context across frames and supports joint training on images and videos (Huang et al., 2026). TruVIL and AINPAINT-ST learn both their temporal context and their forensic features from video. Few video datasets pair per-frame masks with authentic videos, among them AINPAINT and the ROVI subset of FakeParts (Montibeller et al., 2026; Liu et al., 2025; Wu et al., 2024). TVIL, TASLE and ActivityForensics label forged segments only in time (Zhang et al., 2023b; Feng et al., 2026; Bao et al., 2026). Image forgery datasets are far larger and span many diffusion editors (Costanzino et al., 2025; Huang et al., 2025; Wang et al., 2025; Li et al., 2025). Chain of Evidence therefore learns forensic evidence on images and trains only a causal memory bank and the presence head on video.

3 Chain of Evidence

Chain of Evidence localizes forgeries in video by carrying forensic evidence across frames (Fig. 2). An evidence transformer derives this evidence from the perturbation sensitivity of frozen SAM 3 (Sec. 3.1). In the image stage, the evidence refines the features, prompts the frozen mask decoder and informs the presence head (Sec. 3.2). The video stage adds a causal memory bank (Sec. 3.3).

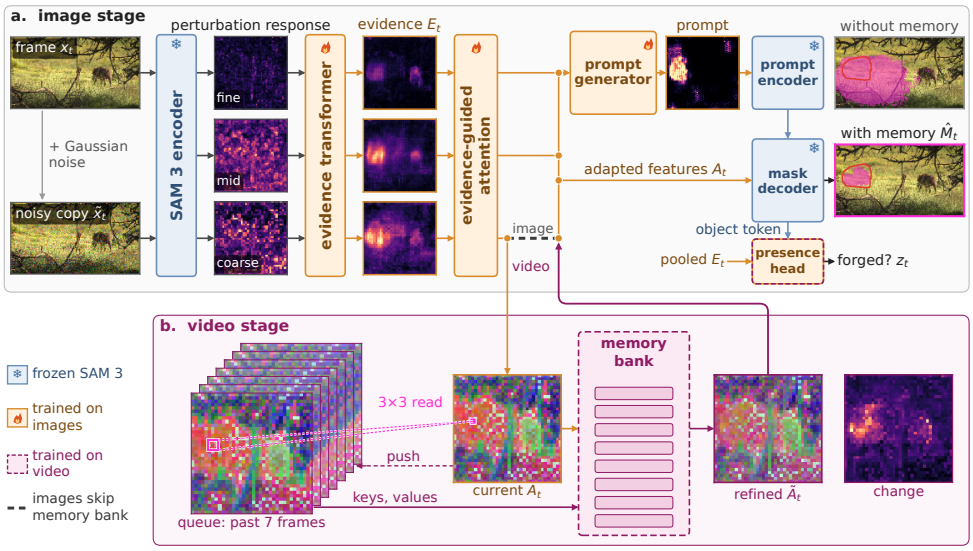


Figure 2: Overview of Chain of Evidence. **a.** Image stage. The evidence transformer turns the perturbation response of the frozen SAM 3 encoder into multiscale evidence E_t that conditions the evidence-guided attention, prompt generator and presence head. **b.** Video stage. The memory bank refines the coarse feature A_t^c from a queue of past features, and the refined $\tilde{A}_t^c$ replaces A_t^c in **a**.

3.1 Problem Setup

Forgery localization. A localized forgery alters only part of a video, in space and in time. Given frames $\mathbf{X} = (x_t)_{t=1}^{N}$, the model predicts $(z_t, \hat{M}_t) = f_{\theta_{\text{img}},\theta_{\text{mem}},\theta_{\text{cls}}}(x_{1:t})$ for each frame, where $z_t \in \mathbb{R}$ is the presence logit and $\hat{M}_t \in [0,1]^{H\times W}$ the forgery probability map. The parameters θ_{cls} belong to the presence head, θ_{mem} to the memory bank and θ_{img} to the other trained modules. Prediction is causal. An image is the special case $N = 1$ with an empty memory bank. Image forgery data spans many editors, whereas video data with pixel masks is scarce (Sec. 2). We therefore learn θ_{img} and θ_{cls} on images and freeze θ_{img}. On video we train θ_{mem} and retrain θ_{cls}.

Spatiotemporal evaluation. Published video localizers score masks only on known forged frames (Lou et al., 2025; Huang et al., 2026; Montibeller et al., 2026), and temporal localizers score frames without pixels (Zhang et al., 2023b; Feng et al., 2026). Neither penalizes masks on authentic frames or missed forged frames. We therefore gate each binarized mask by its frame decision $d_t \in \{0,1\}$, with $d_t = \mathbb{I}[z_t > 0]$ for our model. We compare the gated mask with the ground-truth mask M_t:

$$\hat{P}_t = d_t\, \mathbb{I}[\hat{M}_t > 0.5], \qquad ST = \frac{\sum_t |\hat{P}_t \cap M_t|}{\sum_t |\hat{P}_t \cup M_t|}, \tag{1}$$

where the sums run over all frames t of all forged videos. ST extends the spatio-temporal IoU of video instance segmentation (Yang et al., 2019). The spatial score S averages $\text{IoU}(\hat{P}_t, M_t)$ over forged frames, so a missed frame scores zero. ST instead pools the pixels, so a mask on an authentic frame enlarges the union. ST is high only when a model is correct in space and in time. For an image, S reduces to the localization score L. Sec. 4.1 scores every method this way.

Perturbation as forensic evidence. The sensitivity of pretrained features to a small perturbation separates generated and forged content from authentic content (He et al., 2024; Lek et al., 2026). We keep the spatial structure of this sensitivity instead of collapsing it into one detection score. To measure this sensitivity, we add Gaussian noise to each frame x_t,

$$\tilde{x}_t = x_t + \sigma\epsilon_t, \qquad \epsilon_t \sim \mathcal{N}(0, I), \tag{2}$$

with a fixed σ (Appendix B), and clip the result to $[0,1]$. The frozen SAM 3 encoder maps both inputs to features F_t^s and $\tilde{F}_t^s$ at three scales s. With $\Delta_t^s = F_t^s - \tilde{F}_t^s$, the evidence transformer receives $R_t^s = \text{Cat}[F_t^s, \tilde{F}_t^s, |\Delta_t^s|, \Delta_t^s]$ at each scale (Sec. 3.2). The clean and perturbed features provide the visual context, and the absolute and signed differences give the magnitude and the direction of the change at each location. The channel norm of Δ_t^s is the perturbation response in Figs. 2 and 3.

3.2 Image Forgery Localization

The image stage trains the four modules of Fig. 2a around the frozen SAM 3. Appendix B gives their resolutions and attention layouts and the depth of the evidence transformer.

Multiscale evidence. Forgery traces occur at different spatial extents (Guillaro et al., 2023; Zhu et al., 2025), so the evidence transformer Φ models all scales jointly. It refines each R_t^s locally, pools the scales to the coarsest grid and applies global self-attention there. The perturbation sensitivity at one location is thus interpreted in the context of the evidence elsewhere in the image. The attended output is upsampled step by step and merged with each finer scale, which carries this global context to the resolutions where forgery boundaries are drawn. The result is the evidence pyramid $E_t = \{E_t^s\}_{s=1}^{3} = \Phi(R_t^1, R_t^2, R_t^3)$, which conditions the three modules below.

Evidence-guided attention. SAM 3 features are trained to describe objects rather than forgeries (Carion et al., 2026). Replacing them would discard the boundary accuracy of the SAM 3 mask decoder, so we refine them in place. At each scale, the clean feature conditioned on E_t^s forms the attention query, and the perturbed feature provides keys and values. The module thus extracts what changed under perturbation where the evidence indicates a forgery. The correction is residual,

$$A_t^s = F_t^s + \mathcal{A}_s\big(F_t^s, \tilde{F}_t^s, E_t^s\big), \tag{3}$$

where $\mathcal{A}_s$ is the evidence-guided attention at scale s. A learned gate scales the correction, so the pretrained feature is kept wherever the gate closes.

Prompt generator. SAM 3 segments the region that a prompt specifies (Carion et al., 2026). In forgery localization, however, this region is unknown. The prompt generator therefore predicts a dense mask prompt from E_t, the adapted pyramid $A_t = \{A_t^s\}$ and the clean and perturbed features. The frozen prompt encoder and mask decoder turn this prompt and A_t into $\hat{M}_t$. No prompt labels exist, so the generator learns end-to-end through the frozen decoder from the mask loss. It thus writes whatever prompt makes the decoder produce the correct mask.

Presence head. The mask decoder returns a mask for every frame, forged or not. SAM 3 predicts through its object-score head whether the prompted object exists in the image (Carion et al., 2026). We retrain this head as the presence head, which predicts whether the frame is forged. Its input is the object-score token o_t of the decoder plus a forensic vector c_t pooled from E_t, so the presence logit is $z_t = \mathrm{MLP}_{\mathrm{pres}}(o_t + c_t)$. A frame with $z_t \leq 0$ yields an empty mask. Localization and detection therefore both derive from the same evidence E_t.

3.3 Causal Video Transfer

The image stage judges each frame in isolation, although a forgery is clearer in some frames than others. We thus add a causal memory bank so that earlier frames can confirm a faint forgery (Fig. 2b).

A queue stores the coarse features $A_{t'}^c$ of the preceding sampled frames $t' < t$, where A_t^c is the adapted feature at the coarsest scale. For each location of the current frame, attention reads the 3×3 neighborhood of the same location in every stored frame. Reading a memory of past frames follows space-time memory networks (Oh et al., 2019; Cheng & Schwing, 2022) and the memory attention of SAM 2 (Ravi et al., 2025). The local read provides temporal context without global attention over every stored location. It also tolerates small displacements without explicit motion alignment, assuming a forged region stays approximately in place between sampled frames. A stack of attention and feed-forward blocks, each performing this read, refines A_t^c into $\tilde{A}_t^c$, which replaces A_t^c before prompting and decoding. The queue stores A_t^c rather than $\tilde{A}_t^c$, so refinements are never fed back into later memory states. At the first frame the queue is empty and A_t^c passes through unchanged.

3.4 Learning Objectives

Both stages train the mask $\hat{M}_t$ with a Dice loss and a focal loss. They train the presence logit z_t with binary cross-entropy. The video stage adds two terms. A clip is predicted forged if any of its frames is predicted forged, so a single false alarm mislabels an authentic clip. The term $\mathcal{L}_{\mathrm{auth}}$ therefore penalizes the highest presence logit of each authentic clip through a softplus. The term $\mathcal{L}_{\mathrm{mem}} = \frac{1}{N}\sum_t \|\tilde{A}_t^c - A_t^c\|_2^2 / \|A_t^c\|_2^2$ penalizes the size of the memory update relative to the feature it refines. The video objective combines all five terms,

$$\mathcal{L}_{\mathrm{video}} = \mathcal{L}_{\mathrm{Dice}} + \lambda_{\mathrm{f}}\mathcal{L}_{\mathrm{focal}} + \mathcal{L}_{\mathrm{BCE}} + \lambda_{\mathrm{a}}\mathcal{L}_{\mathrm{auth}} + \lambda_{\mathrm{m}}\mathcal{L}_{\mathrm{mem}}, \tag{4}$$

and the image objective is the same without the last two terms. The focal weight λ_{f} balances the focal and Dice terms. The weights λ_{a} and λ_{m} are small because their terms are soft constraints. Appendix B lists all values and the exact form of $\mathcal{L}_{\mathrm{auth}}$.

4 Experiments

We test whether forensic evidence learned on images localizes forgeries out of distribution, and whether carrying it through time improves video localization. After the evaluation protocol, datasets, baselines and implementation, we report quantitative and qualitative results and ablations.

4.1 Experimental Setup

Evaluation metrics. D is the F1 score for detecting forged images or videos, and T the F1 score for detecting forged frames. L and S are the mean IoU over forged images and over forged frames. ST is the IoU pooled over all frames of the forged videos (Sec. 3.1). Every mask is gated by the detection of its method, so a forged image or frame that detection misses scores zero. The labels of a dataset determine the metrics it supports (Table 1). Appendix D contrasts S and ST.

Datasets. The image stage is trained on SID-Set and BBC-PAIR. The memory bank is trained on FakeParts, whose per-frame masks come from ROVI object annotations (Wu et al., 2024). TVIL inpaints a short span of frames in otherwise authentic clips, whereas TASLE and ActivityForensics replace a span with generated content. TVIL shares the E2FGVI inpainter and YouTube-VOS source footage (Xu et al., 2018) with FakeParts. Appendix A gives the size and editors of every set.

Table 1: **Datasets.** Shaded: training sets. *No authentic test samples. †Also trains Chain of Evidence (memory, AINPAINT). ‡Also selects the image-stage checkpoint.

Image dataset	Metrics	Video dataset	Metrics
SID-Set (2025)	D, L	FakeParts (2025)	D, T, S, ST
BBC-PAIR (2025)	D, L	AINPAINT† (2026)	D, T, S, ST
DiffSeg30k (2025)	D, L	TVIL (2023b)	D, T
OpenSDI (2025)	D, L	TASLE (2026)	D, T
BR-Gen (2026)	L^*	ActivityForensics (2026)	T^*
NeXT-IMDL (2025)	D, L		
MagicBrush‡ (2023a)	D, L		

Baselines. For a fair and reproducible comparison, we run the released models of ten state-of-the-art methods on the same frames without retraining. The image localizers TruFor, Mesorch, DINOv3-IML and Detective SAM (Guillaro et al., 2023; Zhu et al., 2025; Yu et al., 2026; Lek et al., 2026) are applied to images and frame by frame to video. The video localizers VideoFACT, TruVIL, RelayFormer and AINPAINT-ST (Nguyen et al., 2024; Lou et al., 2025; Huang et al., 2026; Montibeller et al., 2026) predict a mask for every frame. The temporal localizers UMMAFormer and MSLoc-PG (Zhang et al., 2023b; Feng et al., 2026) predict forged frames without masks.

Implementation details. Chain of Evidence is implemented in PyTorch on a frozen SAM 3 (Carion et al., 2026) with 512×512 inputs and $\sigma = 0.1$. The image stage trains with batch size 16 and AdamW at a learning rate of 2.5×10^{-4} on one NVIDIA H100 GPU. The video stage trains on clips of 8 sampled frames with AdamW at 5×10^{-5} on one NVIDIA RTX 4090 GPU. Checkpoints are selected on validation data, and Appendix B lists all hyperparameters.

4.2 Quantitative Results

Chain of Evidence (framewise) applies the image stage per frame. Our main video result is Chain of Evidence (memory, FakeParts), and Chain of Evidence (memory, AINPAINT) controls for the video training set. Tables 2 and 3 give the image and video results of all methods, and Appendix C shows where Chain of Evidence outperforms the best baseline.

Table 2: **Image detection and localization.** D: detection F1. L: localization IoU. Shaded: trained on that set. Bold: best in each column. Header groups and rule: relative to Chain of Evidence. Average and rank: over the four sets no method trained on, three for D. NA: no authentic images. †Frame decision from the mask. ‡Its validation images selected our checkpoint.

	In distribution				Out of distribution													
	SID-Set (2025)		BBC-PAIR (2025)		DiffSeg30k (2025)		OpenSDI (2025)		BR-Gen (2026)		NeXT-IMDL (2025)		MagicBrush (2023a)		Average		Rank ↓	
Method	D	L	D	L	D	L	D	L	D	L	D	L	D	L	D	L	D	L
TruFor (2023)	0.451	0.108	0.161	0.016	0.721	0.280	0.368	0.091	NA	0.229	0.353	0.044	0.852	0.163	0.480	0.161	5.00	3.50
Mesorch (2025)†	0.693	0.147	0.845	0.035	0.860	0.149	0.660	0.061	NA	0.065	0.665	0.079	0.801	0.113	0.728	0.089	2.67	4.75
DINOv3-IML (2026)†	0.730	0.338	0.880	0.118	**0.927**	0.276	**0.668**	**0.184**	NA	0.198	0.668	0.148	**0.875**	0.254	**0.755**	0.201	**1.33**	2.75
Detective SAM (2026)†	0.653	0.445	0.844	0.227	0.818	0.383	0.583	0.134	NA	0.213	0.568	0.141	0.805	**0.396**	0.656	0.218	3.67	2.75
Chain of Evidence	**0.863**	**0.588**	**0.979**	**0.430**	0.914	**0.451**	0.565	0.166	NA	**0.361**	**0.670**	**0.219**	0.816‡	0.324‡	0.716	**0.299**	2.33	**1.25**

Image results. Chain of Evidence has the best localization rank and the best L on three of the five out-of-distribution sets. Its average L over the four sets no method trained on is 0.299, against 0.218 for Detective SAM, the best baseline. The trace-based localizers TruFor and Mesorch reach average L values of only 0.161 and 0.089. The margin is largest on BR-Gen, where Chain of Evidence reaches an L of 0.361 against 0.229 for TruFor. OpenSDI is the only set on which a baseline not

trained on it localizes better, with an L of 0.184 for DINOv3-IML against 0.166. Its editors belong to the Stable Diffusion and FLUX families, which also edit BBC-PAIR, so sharing editors with the training data does not guarantee transfer. The advantage also holds in distribution. On SID-Set, on which both were trained, Chain of Evidence reaches an L of 0.588 against 0.445 for Detective SAM and a D of 0.863 against 0.653. DINOv3-IML detects forged images better, reaching an average D of 0.755 against 0.716 for Chain of Evidence. Most of this gap comes from OpenSDI, where its D of 0.668 exceeds 0.565. On DiffSeg30k and NeXT-IMDL, the two differ by at most 0.013 in D.

Table 3: **Video detection and localization.** D: detection F1 over videos. T: temporal F1 over frames. S: mean IoU over forged frames. ST: pooled IoU over forged videos. Shaded: trained on that set. Bold: best in each column. NA: no authentic videos. –: no masks predicted. †Frame decision from the mask. §Source videos overlap UMMAFormer training data. Averages and ranks are reported in Appendix C.

	FakeParts (2025)				AINPAINT (2026)				TVIL (2023b)		TASLE (2026)		Activity-Forensics (2026)	
Method	D	T	S	ST	D	T	S	ST	D	T	D	T	D	T
UMMAFormer (2023b)	0.757§	0.720§	–	–	0.941	0.932	–	–	0.867	0.092	0.656	0.258	NA	0.511
MSLoc-PG (2026)	0.230	0.110	–	–	0.169	0.126	–	–	0.182	0.060	**0.889**	**0.729**	NA	0.276
TruFor (2023) (framewise)	0.591	0.193	0.007	0.021	0.392	0.117	0.012	0.016	0.738	0.074	0.643	0.164	NA	0.215
Mesorch (2025) (framewise)†	0.757	0.725	0.017	0.033	0.941	0.932	0.094	0.138	0.867	0.092	0.656	0.259	NA	0.511
DINOv3-IML (2026) (framewise)†	0.758	0.708	0.086	0.119	0.941	0.919	0.374	0.495	0.867	0.107	0.656	0.269	NA	0.492
Detective SAM (2026) (framewise)†	0.752	0.577	0.034	0.058	0.933	0.725	0.036	0.069	0.867	0.072	0.656	0.235	NA	0.460
VideoFACT (2024)	0.604	0.358	0.014	0.032	0.626	0.370	0.015	0.035	0.603	0.070	0.625	0.159	NA	0.403
RelayFormer (2026)	0.762	0.717	0.194	0.207	0.935	0.851	0.206	0.212	0.869	0.112	0.649	0.202	NA	0.485
TruVIL (2025)	0.757	0.716	0.128	0.161	0.933	0.911	0.180	0.169	0.867	0.096	0.656	0.258	NA	0.499
AINPAINT-ST (2026)	0.771	0.798	0.351	0.404	**0.959**	**0.960**	**0.660**	**0.719**	0.876	0.259	0.655	0.240	NA	0.438
Chain of Evidence (framewise)	0.735	0.526	0.056	0.099	0.943	0.814	0.140	0.162	0.856	0.133	0.659	0.312	NA	0.492
Chain of Evidence (memory, FakeParts)	**0.855**	**0.917**	**0.352**	**0.425**	0.943	0.877	0.309	0.420	**0.891**	**0.370**	0.588	0.108	NA	0.348
Chain of Evidence (memory, AINPAINT)	0.760	0.745	0.234	0.281	0.941	0.937	0.391	0.515	0.867	0.114	0.656	0.269	NA	**0.517**

Video results. Video detection separates the methods only weakly. A video counts as forged when any sampled frame is flagged (Appendix A). On each set with authentic videos, most rows reach a D within 0.01 of labeling every video as forged, which gives 0.867 on TVIL. The comparison therefore rests on T, S and ST. On TVIL, Chain of Evidence (memory, FakeParts) reaches a T of 0.370, against 0.259 for AINPAINT-ST, the best baseline, and 0.092 for UMMAFormer, which was trained on TVIL. Several rows, among them Chain of Evidence (memory, AINPAINT), flag most frames and score close to labeling every frame as forged. A high T does not imply localization. Mesorch (framewise) reaches a T of 0.932 on AINPAINT but an S of only 0.094. Out of distribution, AINPAINT-ST localizes best on FakeParts and DINOv3-IML on AINPAINT, with an ST of 0.495. Chain of Evidence (memory, FakeParts) follows with 0.420, ahead of every video localizer not trained on AINPAINT. The other image baselines applied framewise stay below an ST of 0.15 on both sets. Generated segments remain hard for every method not trained on them. On TASLE, the best out-of-distribution T is 0.312, from Chain of Evidence (framewise), whereas MSLoc-PG, which was trained on TASLE, reaches 0.729. On ActivityForensics, no method beats labeling every frame as forged, which gives a T of 0.511, by more than 0.006.

4.3 Qualitative Results

On the AINPAINT clip of Fig. 3, the perturbation response is noisy and only slightly higher inside the forged region. The evidence transformer turns it into a map that peaks at the forgery. The memory bank trims the framewise mask to the forged region and holds presence near one, whereas the framewise presence oscillates from frame to frame.

In Fig. 4, RelayFormer marks small or misplaced regions. DINOv3-IML treats each frame in isolation, and its masks range from precise to over-segmented or absent. AINPAINT-ST covers the forged region less completely than Chain of Evidence (memory, FakeParts). Chain of Evidence (framewise) either over-segments or misses forged frames. On these clips, Chain of Evidence (memory, FakeParts) detects every forged frame and produces the tightest masks. Appendices F and G show the model internals and the failure modes, respectively.

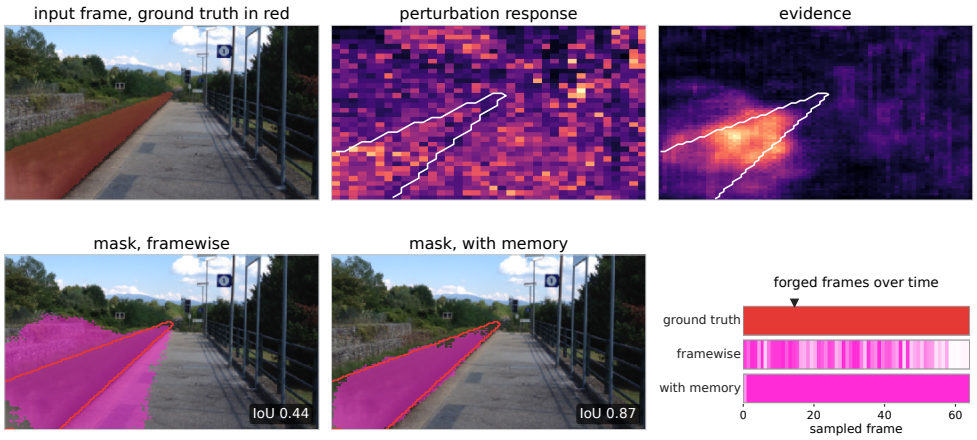


Figure 3: One AINPAINT clip through Chain of Evidence, from the perturbation response of the frozen SAM 3 encoder to masks and presence with and without the memory bank. Magenta marks the prediction and outlines mark the ground truth.

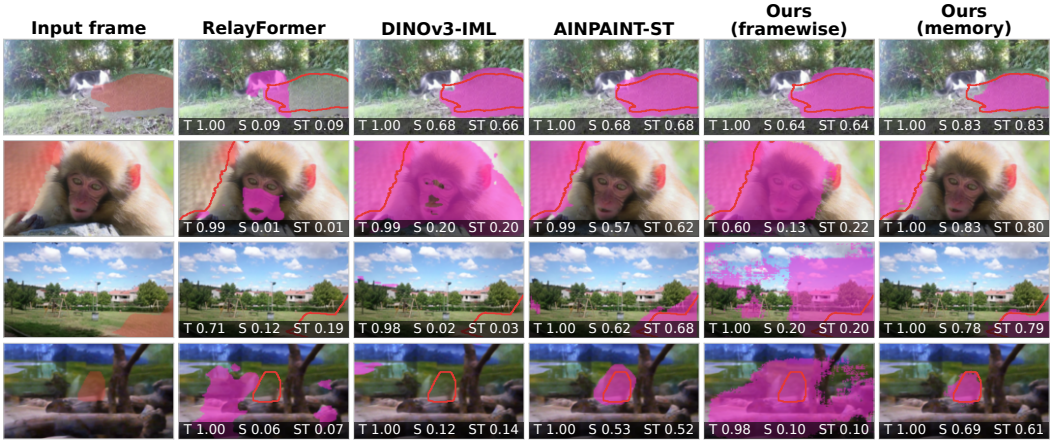


Figure 4: Chain of Evidence against the three baselines with the highest ST in Table 3 on two FakeParts clips (top) and two AINPAINT clips (bottom). Ours (framewise) and Ours (memory) denote Chain of Evidence (framewise) and Chain of Evidence (memory, FakeParts), which is in distribution on FakeParts. Red: ground truth. Magenta: prediction after each method's own frame decision. Each panel reports T, S and ST for its clip. Clips are chosen among those where Ours (memory) leads every baseline in S (Appendix E).

4.4 Ablation Studies

We ablate the memory bank and test robustness to common perturbations. The ablation isolates what video training adds to the evidence learned on images. The robustness study tests whether this evidence survives the degradations that images and videos undergo when shared.

Memory bank. Chain of Evidence (framewise) removes the memory bank together with its retrained presence head (Table 3). On the out-of-distribution AINPAINT, spatiotemporal localization degrades, with ST decreasing from 0.420 to 0.162. On TVIL, temporal localization degrades, with T decreasing from 0.370 to 0.133. Labeling every frame as forged reaches a T of only 0.092 on TVIL. The memory bank thus improves the localization of local edits. When the forgery replaces whole segments, as in TASLE and ActivityForensics, the memory bank trained on FakeParts reduces T instead, from 0.312 to 0.108 on TASLE. As a control, the memory bank trained on AINPAINT

raises ST on the out-of-distribution FakeParts from 0.099 to 0.281 but lowers T on TVIL from 0.133 to 0.114. The spatiotemporal gain therefore holds for both video training sets, whereas the temporal gain on TVIL holds only for training on FakeParts.

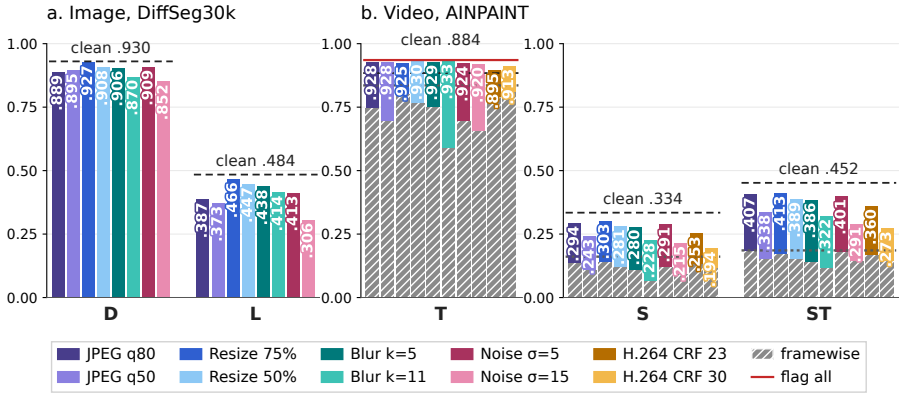


Figure 5: **Robustness to perturbations.** a: Chain of Evidence on DiffSeg30k. b: Chain of Evidence (memory, FakeParts) on AINPAINT, with Chain of Evidence (framewise) hatched. Dashed and dotted lines: clean scores of the colored and hatched bars, respectively. Red line: T of labeling every frame as forged ("flag all" in the legend).

Robustness to perturbations. Following Costanzino et al. (2025), we apply JPEG compression, resizing and Gaussian blur, add Gaussian noise, and re-encode video with H.264 (Montibeller et al., 2026), each at a mild and a strong level (Fig. 5). We use out-of-distribution subsets of DiffSeg30k and AINPAINT (Appendix H), whose clean scores differ slightly from the main tables. Image detection stays above 0.85 under every perturbation. Localization is least sensitive to resizing and most sensitive to strong noise, which lowers L from 0.484 to 0.306. This is consistent with the evidence being derived from the encoder response to noise (Sec. 3.1). On video, the lowest ST of Chain of Evidence (memory, FakeParts), 0.273 under H.264 CRF 30, exceeds the clean 0.186 of Chain of Evidence (framewise). Under perturbation, Chain of Evidence (memory, FakeParts) averages a T of 0.923, against 0.936 for labeling every frame as forged.

5 Conclusion

We present Chain of Evidence, which provides state-of-the-art forgery localization on out-of-distribution images and localizes video forgeries over space and time. We address the scarcity of pixel masks for diffusion-edited video by learning forensic evidence on images from many diffusion editors and carrying it into time. Forensic evidence from an evidence transformer prompts and guides a frozen SAM 3. Video training updates only a causal memory bank and the presence head, so the image-learned evidence stays intact. On image benchmarks out of distribution for every method, Chain of Evidence reaches an average L of 0.299 against 0.218 for Detective SAM, the best baseline. Our gated evaluation protocol scores space and time jointly. Under this protocol, video training on FakeParts raises ST on the out-of-distribution AINPAINT from 0.162 to 0.420. A control model trained on AINPAINT raises ST on FakeParts from 0.099 to 0.281.

We note three limitations. First, the memory bank reads a small neighborhood around each location, assuming little motion of the forged region between sampled frames. Second, of the tested perturbations, strong Gaussian noise lowers image localization most. Third, video datasets with per-frame masks mostly use non-diffusion inpainters. Most diffusion-generated video datasets lack pixel masks. On TASLE, where whole segments are generated, the memory bank trained on FakeParts lowers T from 0.312 to 0.108. Future work therefore needs diffusion-edited video datasets with per-frame masks, authentic videos, and both local edits and generated segments. Aligning memory reads with estimated motion could address the first limitation.

AI USE STATEMENT

In this work, we used generative AI tools to implement methods, reformat datasets, give feedback on experiments and help interpret results. We have not used them for data generation, mathematical claims, proofs, translation or qualitative analysis. Additionally, we used them to draft and edit text, write figure scripts, search the literature and format references. We have reviewed all AI-assisted work. The authors checked all code, and every number in the tables was recomputed from stored predictions. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work aims to help verify images and videos by localizing forged content. It uses only publicly released datasets under their licences, and we collected no new data involving human subjects. Some datasets show people in authentic footage. We use them only as released and show only individual frames in our figures. A forgery localizer is dual use. Knowledge of where it fails could help an adversary evade detection, and its failures could give false assurance, since content that is not flagged is not thereby authentic. We therefore report its failure modes explicitly (Appendix G), including whole-segment generation, on which the memory bank loses accuracy. Accuracy also varies across editors and content (Sec. 4.2), so a result on one benchmark does not guarantee accuracy on another.

REFERENCES

Peijun Bao, Anwei Luo, Gang Pan, Alex C. Kot, and Xudong Jiang. ActivityForensics: A comprehensive benchmark for localizing manipulated activity in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 42987–42996, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Bao_ActivityForensics_A_Comprehensive_Benchmark_for_Localizing_Manipulated_Activity_in_Videos_CVPR_2026_paper.html.

Black Forest Labs. FLUX. https://github.com/black-forest-labs/flux, 2024.

Lvpan Cai, Haowei Wang, Jiayi Ji, Yanshu Zhoumen, Shen Chen, Taiping Yao, and Xiaoshuai Sun. Zooming in on fakes: A novel dataset for localized AI-generated image detection with forgery amplification approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 2534–2542, 2026. doi: 10.1609/aaai.v40i4.37240. URL https://arxiv.org/abs/2504.11922.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts. In *International Conference on Learning Representations*, pp. 138846–138923, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/e0982cbc81401df3430ee1ff780dc7a2-Abstract-Conference.html.

Haoxing Chen, Yan Hong, Zizheng Huang, Zhuoer Xu, Zhangxuan Gu, Yaohui Li, Jun Lan, Huijia Zhu, Jianfu Zhang, Weiqiang Wang, and Huaxiong Li. DeMamba: AI-generated video detection on million-scale GenVideo benchmark. *Science China Information Sciences*, 69(6): 162103, 2026. doi: 10.1007/s11432-024-4894-0. URL https://doi.org/10.1007/s11432-024-4894-0.

Ho Kei Cheng and Alexander G. Schwing. XMem: Long-term video object segmentation with an Atkinson-Shiffrin memory model. In *European Conference on Computer Vision (ECCV)*, pp. 640–658, 2022.

Hai Ci, Ziheng Peng, Pei Yang, Yingxin Xuan, and Mike Zheng Shou. DiffSeg30k: A multi-turn diffusion editing benchmark for localized AIGC detection. arXiv preprint arXiv:2511.19111, 2025. URL https://arxiv.org/abs/2511.19111.

Alex Costanzino, Woody Bayliss, Juil Sock, Marc Gorriz Blanch, Danijela Horak, Ivan Laptev, Philip Torr, and Fabio Pizzati. Towards reliable identification of diffusion-based image manipulations. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2506.05466.

Europol. Facing reality? law enforcement and the challenge of deepfakes. Observatory report, Publications Office of the European Union, Luxembourg, 2022. URL https://www.europol.europa.eu/cms/sites/default/files/documents/Europol_Innovation_Lab_Facing_Reality_Law_Enforcement_And_The_Challenge_Of_Deepfakes.pdf. Revised January 2024.

Yue Feng, Jingjing Li, Qijia Lu, Wei Ji, Jingrou Zhang, Fei Shen, Xiao Li, Yizhen Jia, Qiang Chen, Limin Wang, Wentong Li, and Jie Qin. Explainable forensics of manipulated segments in untrimmed long videos. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2606.02402.

Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. TokenFlow: Consistent diffusion features for consistent video editing. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2307.10373.

Fabrizio Guillaro, Davide Cozzolino, Avneesh Sud, Nicholas Dufour, and Luisa Verdoliva. TruFor: Leveraging all-round clues for trustworthy image forgery detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20606–20615, 2023. URL https://arxiv.org/abs/2212.10957.

Zhiyuan He, Pin-Yu Chen, and Tsung-Yi Ho. RIGID: A training-free and model-agnostic framework for robust AI-generated image detection. arXiv preprint arXiv:2405.20112, 2024. URL https://arxiv.org/abs/2405.20112.

Wen Huang, Jiarui Yang, Tao Dai, Jiawei Li, Shaoxiong Zhan, Bin Wang, and Shu-Tao Xia. RelayFormer: A unified local-global attention framework for scalable image and video manipulation localization. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2508.09459.

Zhenglin Huang, Jinwei Hu, Xiangtai Li, Yiwei He, Xingyu Zhao, Bei Peng, Baoyuan Wu, Xiaowei Huang, and Guangliang Cheng. SIDA: Social media image deepfake detection, localization and explanation with large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28831–28841, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Huang_SIDA_Social_Media_Image_Deepfake_Detection_Localization_and_Explanation_with_CVPR_2025_paper.html.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. VACE: All-in-one video creation and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17191–17202, 2025. URL https://arxiv.org/abs/2503.07598.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4015–4026, 2023. URL https://arxiv.org/abs/2304.02643.

Kolors Team. Kolors: Effective training of diffusion model for photorealistic text-to-image synthesis. https://github.com/Kwai-Kolors/Kolors, 2024.

Myung-Joon Kwon, Wonjun Lee, Seung-Hun Nam, Minji Son, and Changick Kim. SAFIRE: Segment any forged image region. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 4437–4445, 2025.

Gert Lek, Nicolas Van Schaik, Chaoyi Zhu, Pin-Yu Chen, Robert Birke, and Lydia Y. Chen. Detective SAM: Adaptive AI-image forgery localization. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=GKJHPHNFIx.

Yifei Li, Haoyuan He, Yu Zheng, Bingyao Yu, Wenzhao Zheng, Lei Chen, Jie Zhou, and Jiwen Lu. NeXT-IMDL: Build benchmark for NeXT-generation image manipulation detection & localization. arXiv preprint arXiv:2512.23374, 2025. URL https://arxiv.org/abs/2512.23374. Withdrawn by the authors in March 2026.

Zhen Li, Cheng-Ze Lu, Jianhua Qin, Chun-Le Guo, and Ming-Ming Cheng. Towards an end-to-end framework for flow-guided video inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Ziyi Liu, Firas Gabetni, Awais Hussain Sani, Xi Wang, Soobash Daiboo, Gaëtan Brison, Gianni Franchi, and Vicky Kalogeiton. FakeParts: A new family of AI-generated DeepFakes. arXiv preprint arXiv:2508.21052, 2025. URL https://arxiv.org/abs/2508.21052.

Zijie Lou, Gang Cao, Man Lin, Lifang Yu, and Shaowei Weng. Trusted video inpainting localization via deep attentive noise learning. *IEEE Transactions on Dependable and Secure Computing*, 22 (6):7215–7228, 2025. doi: 10.1109/TDSC.2025.3595960.

Andrea Montibeller, Giulia Boato, and Luisa Verdoliva. AINPAINT: A comprehensive dataset and dual branch architecture for practical video inpainting localization. *Computer Vision and Image Understanding*, 271:104866, 2026. doi: 10.1016/j.cviu.2026.104866. URL https://www.sciencedirect.com/science/article/pii/S107731422600233X.

Tai D. Nguyen, Shengbang Fang, and Matthew C. Stamm. VideoFACT: Detecting video forgeries using attention, scene context, and forensic traces. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 8563–8573, 2024. URL https://openaccess.thecvf.com/content/WACV2024/html/Nguyen_VideoFACT_Detecting_Video_Forgeries_Using_Attention_Scene_Context_and_Forensic_WACV_2024_paper.html.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16784–16804. PMLR, 2022.

Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9226–9235, 2019.

Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint arXiv:2204.06125*, 2022.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/45c1f6a8cbf2da59ebf2c802b4f742cd-Abstract-Conference.html.

Anton Razzhigaev, Arseniy Shakhmatov, Anastasia Maltseva, Vladimir Arkhipkin, Igor Pavlov, Ilya Ryabov, Angelina Kuts, Alexander Panchenko, Andrey Kuznetsov, and Denis Dimitrov. Kandinsky: An improved text-to-image synthesis with image prior and latent diffusion. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2023.

Jonas Ricker, Denis Lukovnikov, and Asja Fischer. AEROBLADE: Training-free detection of latent diffusion images using autoencoder reconstruction error. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9130–9140, 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Chung-Ting Tsai, Ching-Yun Ko, I-Hsin Chung, Yu-Chiang Frank Wang, and Pin-Yu Chen. Understanding and improving training-free AI-generated image detections with vision foundation models. arXiv preprint arXiv:2411.19117, 2024. URL https://arxiv.org/abs/2411.19117.

Yabin Wang, Zhiwu Huang, and Xiaopeng Hong. OpenSDI: Spotting diffusion-generated images in the open world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://arxiv.org/abs/2503.19653.

Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. DIRE for diffusion-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22445–22455, 2023.

Jianzong Wu, Xiangtai Li, Chenyang Si, Shangchen Zhou, Jingkang Yang, Jiangning Zhang, Yining Li, Kai Chen, Yunhai Tong, Ziwei Liu, and Chen Change Loy. Towards language-driven video inpainting via multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas Huang. YouTube-VOS: A large-scale video object segmentation benchmark. *arXiv preprint arXiv:1809.03327*, 2018.

Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

Jieming Yu, Qiuxiao Feng, Zhuohan Wang, and Xiaochen Ma. DINOv3 beats specialized detectors: A simple foundation model baseline for image forensics, 2026. URL https://arxiv.org/abs/2604.16083.

Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. MagicBrush: A manually annotated dataset for instruction-guided image editing. In *Advances in Neural Information Processing Systems*, 2023a. URL https://arxiv.org/abs/2306.10012.

Rui Zhang, Hongxia Wang, Mingshan Du, Hanqing Liu, Yang Zhou, and Qiang Zeng. UMMAFormer: A universal multimodal-adaptive transformer framework for temporal forgery localization. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, pp. 8749–8759. ACM, 2023b. doi: 10.1145/3581783.3613767. URL https://doi.org/10.1145/3581783.3613767.

Xuekang Zhu, Xiaochen Ma, Lei Su, Zhuohang Jiang, Bo Du, Xiwen Wang, Zeyu Lei, Wentao Feng, Chi-Man Pun, and Ji-Zhe Zhou. Mesoscopic insights: Orchestrating multi-scale & hybrid architecture for image manipulation localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 11022–11030, 2025. URL https://arxiv.org/abs/2412.13753.

Junhao Zhuang, Yanhong Zeng, Wenran Liu, Chun Yuan, and Kai Chen. A task is worth one word: Learning with task prompts for high-quality versatile image inpainting. In *European Conference on Computer Vision (ECCV)*, 2024.

A Dataset and Baseline Details

Table 4 lists every dataset and baseline of Sec. 4 with its role, labels and properties. The rest of this section gives the editors, the frame sampling and the rules that turn each method's output into the binary decisions that the evaluation metrics score.

Table 4: Data and baselines. Test counts list authentic / forged where both exist. Frames: at most 64 per clip, uniformly spaced, resized to 512×512, with masks scored at 144×144. All baselines are released checkpoints run on the same frames. [a]Joint training set of 163,900 BBC-PAIR and 38,808 SID-Set images. [b]MagicBrush validation images were used to select the image checkpoint, so it is out of distribution but not blind. [c]15 percent held out for validation. †Presence derived from the mask with a threshold selected on the validation splits.

Dataset	Use	Train	Val	Test	Labels
BBC-PAIR	image train, test	official[a]	1,000	4,150	masks
SID-Set	image train, test	official[a]	4,851	40,000	masks
DiffSeg30k	image test	–	–	2,900	masks
OpenSDI	image test	–	–	50,000	masks, local edits
BR-Gen	image test	–	–	4,851	masks, forged only
NeXT-IMDL	image test	–	–	160,616	masks
MagicBrush	image test	–	used[b]	794	masks
FakeParts (ROVI)	video train, test	13,299	200	239 / 373	per-frame masks
AINPAINT	video test, control train	2,494[c]		31 / 247	per-frame masks
TVIL	video test	–	–	119 / 389	frame labels, sparse
TASLE	video test	–	–	662 / 631	frame labels
ActivityForensics	video test	–	–	0 / 1,565	frame labels

Image baselines: TruFor, Mesorch†, DINOv3-IML† and Detective SAM†, also run framewise on video.
Video baselines: VideoFACT, TruVIL, RelayFormer, AINPAINT-ST, UMMAFormer and MSLoc-PG, the proposal stage of MSLoc, all run on the video test sets only.

Editors. BBC-PAIR is inpainted with ten diffusion models: Stable Diffusion 1, 1.5, 2, 3 and 3.5 (Rombach et al., 2022), SDXL, FLUX schnell and FLUX dev fill (Black Forest Labs, 2024), and Kandinsky 2.2 and 3.1 (Razzhigaev et al., 2023). SID-Set uses Stable Diffusion for its forged images and FLUX for its fully synthetic ones. DiffSeg30k uses Stable Diffusion 2 and 3.5, SDXL, FLUX, Kandinsky 2.2, GLIDE (Nichol et al., 2022), Kolors (Kolors Team, 2024) and Hunyuan-DiT. OpenSDI uses Stable Diffusion 1.5, 2, 3 and XL and FLUX. BR-Gen uses SDXL, BrushNet, PowerPaint (Zhuang et al., 2024), LaMa and MAT. NeXT-IMDL uses 27 editors across seven edit categories, among them ACE++, PowerPaint, HIVE, InstructDiffusion, UltraEdit and the Stable Diffusion and FLUX inpainting models. MagicBrush uses DALL-E 2 (Ramesh et al., 2022). The FakeParts inpainting subset is inpainted with E2FGVI (Li et al., 2022) on ROVI videos (Wu et al., 2024), and TVIL with E2FGVI, FuseFormer, STTN and FGT. AINPAINT uses CoCoCo, DSTT, FGVC, FuseFormer, GMCNN, LDVI, OPN and STTN. TASLE inserts segments generated by VACE (Jiang et al., 2025), Wan 2.1, Kling, Vidu, Hailuo, LTX-Video, SkyReels-V1 and Jimeng. ActivityForensics inserts segments generated by Vidu, Wan, FCVG, LTX-Video and VACE and interpolated by Sci-Fi. The metrics each set supports follow from its labels: D needs authentic samples, T needs frame labels, and L, S and ST need pixel masks marking the forged regions.

Image data. The image stage is trained jointly on the BBC-PAIR and SID-Set training splits. Checkpoints and baseline thresholds are selected on their validation splits. The MagicBrush validation images were part of our checkpoint selection, so MagicBrush is out of distribution but not a blind test set. Every method sees the same test images.

Video data. The memory bank and the presence head are trained on the ROVI inpainting subset of FakeParts, with source videos kept disjoint across splits. Its per-frame masks are the ROVI object annotations of the removed object, and frames in which the object does not appear count as authentic. One AINPAINT clip is excluded for every method because 16 of its frames have no mask. TASLE and ActivityForensics annotate forged segments in seconds, converted to frame labels at the native frame rate. The control row Chain of Evidence (memory, AINPAINT) is trained on AINPAINT

with the same recipe. TVIL shares the E2FGVI inpainter and YouTube-VOS source footage with FakeParts, and the released training split of UMMAFormer contains 111 of the 239 FakeParts source videos. The FakeParts cell of UMMAFormer is therefore reported for completeness only. Its TVIL cell is in distribution, so its average in Table 7 rests on the FakeParts cell alone.

Frame sampling. Every method is evaluated on the same frames. A clip contributes at most 64 frames, uniformly spaced over its length, and shorter clips contribute every frame. The median stride between sampled frames is 1.0 frames on FakeParts, 1.1 on AINPAINT, 2.3 on TVIL, 4.5 on TASLE and 12.7 on ActivityForensics. Frames are resized to 512×512 without cropping, and masks are scored on the 144×144 grid of the mask decoder.

Decisions and gating. Masks are binarized at 0.5. A frame is predicted forged by the video baselines when their frame score exceeds 0.5, and by Chain of Evidence when the presence logit is positive. Mesorch, DINOv3-IML and Detective SAM output masks only. Their frame score is the mean of the top one percent of their output probabilities, and the threshold on that score is chosen on the BBC-PAIR and SID-Set validation splits to maximize class-macro F1. It is 0.056 for Mesorch, 0.368 for DINOv3-IML and 0.665 for Detective SAM. A video is predicted forged when any sampled frame is. Every mask is gated by its method's own frame decision before S and ST are computed, so a frame predicted authentic contributes an empty mask.

Checkpoints. Every baseline is its released checkpoint without retraining. Four baselines were trained on some of our test sets: Detective SAM on SID-Set and MagicBrush, UMMAFormer on TVIL, MSLoc-PG on TASLE and AINPAINT-ST on AINPAINT. Their results on these sets are shaded as in distribution. Our video checkpoints are selected on the validation split of their training set. This selection gives epoch 8 of 8 for Chain of Evidence (memory, FakeParts) and epoch 4 of 8 for the control row Chain of Evidence (memory, AINPAINT).

B Model and Training Details

Table 5 lists the trained modules. The frozen SAM 3 encoder holds 454M parameters, and the frozen prompt encoder and mask decoder are used as released. The image stage trains 7.62M parameters in total and the video stage trains 8.57M parameters.

Table 5: Trained parameters per module and the stage that trains them.

Module	Parameters	Trained in
Evidence transformer	5.88M	image stage
Prompt generator	1.13M	image stage
Evidence-guided attention	0.32M	image stage
Presence conditioner	0.15M	image stage
Presence head	0.13M	image stage, retrained in the video stage
Memory bank	8.44M	video stage

Architecture. The perturbation is Gaussian noise with standard deviation $\sigma = 0.1$ on pixel values in $[0, 1]$. Inputs are 512×512, and the frozen encoder gives features at 144^2, 72^2 and 36^2. The evidence transformer projects each R_t^s to one representation per scale and refines it locally. It then pools all scales to the 36^2 grid and applies six layers of global attention over its 1,296 tokens. Its output is merged with the locally refined 36^2 features and progressively upsampled to 72^2 and 144^2, where it is merged with the locally refined features of each scale. Restricting global attention to the coarsest grid keeps its cost affordable. The evidence-guided attention is global at 36^2 and windowed at the two finer scales. A learned gate scales each of its corrections, whose size is bounded relative to the clean feature. At each scale the prompt generator combines the clean, perturbed, adapted and evidence features. It applies global attention over the 36^2 grid, which makes the prompt coherent across the frame. Windowed attention at 72^2 and 144^2, after merging the upsampled coarse result, keeps the prompt sharp at forgery boundaries. The generator emits a 144^2 prompt as logits for the frozen prompt encoder. The presence conditioner computes c_t by mean- and max-pooling E_t over the three scales and projecting the result to the width of the object-score token o_t. The maximum responds to small forgeries, and the mean summarizes the whole frame. The conditioner is trained

in the image stage only. The memory bank operates on the 256-channel 36^2 coarse feature, while the two finer scales bypass it. It keeps the seven most recent coarse features. For each position it reads a 3×3 neighborhood in every stored frame, so an interior position attends to at most 63 stored tokens. Each stored frame carries a learned recency embedding. The read is performed by eight independently parameterized cross-attention blocks, each followed by a feed-forward layer. Every block attends from the current feature to the same stored neighborhoods. The outputs of its attention and of its feed-forward layer are each scaled by 0.1 before they are added residually.

Losses. The binary cross-entropy on the presence logit has weight 1 in both stages. In the image stage the mask losses are computed on forged images. The image focal loss uses $\alpha = 0.65$, $\gamma = 1.5$, separate means over forged and authentic pixels and $\lambda_{\mathrm{f}} = 12$. In the video stage the mask losses are computed on every frame. The video focal loss uses $\alpha = 0.25$, $\gamma = 1$, a mean over all pixels and $\lambda_{\mathrm{f}} = 4$. The authentic-clip term has weight $\lambda_{\mathrm{a}} = 0.1$ and the memory term $\lambda_{\mathrm{m}} = 0.01$. The authentic-clip term, which penalizes the highest presence logit of each authentic clip, is

$$\mathcal{L}_{\mathrm{auth}} = \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \mathrm{softplus}\Big(\max_{t} z_{v,t} \Big), \tag{5}$$

where $\mathcal{V}$ is the set of authentic clips in the batch.

Image stage. Ten epochs on the BBC-PAIR and SID-Set training splits with batch size 16 in bfloat16. AdamW with learning rate 2.5×10^{-4} for the evidence transformer, the evidence-guided attention and the prompt generator, 6×10^{-4} for the presence head and conditioner, weight decay 10^{-4} and a per-step cosine schedule decaying to one percent of the initial rate without warm-up. The checkpoint from epoch 8 is selected on the validation splits.

Video stage. Eight epochs over the FakeParts training clips, one clip per step with up to eight temporally ordered frames sampled from uniform bins of the clip, in mixed precision. AdamW with learning rate 5×10^{-5} for the memory bank and 10^{-5} for the presence head, weight decay 10^{-4} and a reduce-on-plateau schedule. Validation uses 64 frames per clip. All runs use seed 42, and the evaluation settings are those of Appendix A.

C Where Chain of Evidence Outperforms the Baselines

Table 6 collects the benchmarks on which Chain of Evidence outperforms the best baseline. On images, Chain of Evidence detects better than Detective SAM on SID-Set, on which both were trained, and localizes best on average over the four out-of-distribution sets that no method trained on. On video, Chain of Evidence (memory, FakeParts) leads detection and temporal localization on TVIL, ahead of UMMAFormer, which was trained on TVIL. On the out-of-distribution AINPAINT, Chain of Evidence (memory, FakeParts) localizes better than every video localizer not trained there. Table 7 gives the video averages and ranks that Table 3 leaves out.

Table 6: **Where Chain of Evidence outperforms the baselines.** Best baseline: Detective SAM (2026) on images, AINPAINT-ST (2026) on TVIL and, among video localizers, RelayFormer (2026) on AINPAINT. Shaded: both trained on SID-Set. All other cells are out of distribution for both, and TVIL shares the E2FGVI inpainter and YouTube-VOS footage with FakeParts. L averages the four image sets neither side trained on: DiffSeg30k, OpenSDI, BR-Gen and NeXT-IMDL. Applied framewise, DINOv3-IML reaches 0.374 S and 0.495 ST on AINPAINT.

	Image		Video			
	D	L	D	T	S	ST
	SID-Set	4 OOD sets	TVIL		AINPAINT	
Best baseline	0.653	0.218	0.876	0.259	0.206	0.212
Chain of Evidence	**0.863**	**0.299**	**0.891**	**0.370**	**0.309**	**0.420**
Gain	+32%	+37%	+2%	+43%	+50%	+98%

Table 7: **Video averages and ranks** for Table 3. D and T average FakeParts and TVIL, S and ST average FakeParts and AINPAINT, skipping shaded cells. Ranks use the unshaded cells of a method. Bold marks the best value. °Over fewer sets, not bolded. †Frame decision derived from the mask.

	Average		Rank ↓		Average		Rank ↓	
Method	D	T	D	T	S	ST	S	ST
UMMAFormer (2023b)	0.757°	0.720°	4.83	3.38	–	–	–	–
MSLoc-PG (2026)	0.206	0.085	11.67	11.50	–	–	–	–
TruFor (2023) (framewise)	0.664	0.133	10.25	10.80	0.010	0.019	9.50	9.50
Mesorch (2025) (framewise)†	0.812	0.409	5.12	3.80	0.056	0.086	7.00	7.00
DINOv3-IML (2026) (framewise)†	0.813	0.408	4.62	4.80	**0.230**	**0.307**	3.00	3.00
Detective SAM (2026) (framewise)†	0.809	0.324	6.50	8.40	0.035	0.064	7.00	7.00
VideoFACT (2024)	0.603	0.214	10.25	10.20	0.014	0.034	8.50	8.50
RelayFormer (2026)	0.815	0.415	5.00	6.40	0.200	0.210	3.00	3.00
TruVIL (2025)	0.812	0.406	6.00	5.30	0.154	0.165	4.00	4.00
AINPAINT-ST (2026)	**0.823**	**0.528**	**3.67**	4.75	0.351°	0.404°	**1.00**	**1.00**
Chain of Evidence (framewise)	0.795	0.330	5.12	5.10	0.098	0.131	5.50	5.50
Chain of Evidence (memory, FakeParts)	0.891°	0.370°	4.83	7.25	0.309°	0.420°	2.00	2.00
Chain of Evidence (memory, AINPAINT)	0.814	0.429	4.50	**2.38**	0.234°	0.281°	2.00	2.00

D Metrics and a Worked Localization Example

Table 8 summarizes what each metric predicts and how it is scored. Figure 6 illustrates the difference between S and ST on an 8×8 grid. The grid only makes the pixel counts visible. The reported metrics are computed at the full mask resolution.

Table 8: **Evaluation metrics** for end-to-end forgery localization on images and video.

Modality	Metric	Predicts	Score
Image	Detection (D)	forged or authentic	F1 over images
	Localization (L)	forged pixels	mean IoU over forged images
Video	Detection (D)	forged or authentic	F1 over videos
	Temporal (T)	forged frames	F1 over frames
	Spatial (S)	forged pixels per frame	mean IoU over forged frames
	Spatiotemporal (ST)	forged pixels over the video	pooled IoU over forged videos

For frame $t = 5$, the grid contains six true positive cells, five false positive cells, and five false negative cells. Its frame-level IoU is therefore

$$\mathrm{IoU}_{t=5} = \frac{6}{6+5+5} = \frac{6}{16} = 0.375.$$

S averages the IoU of the five forged frames:

$$S = \frac{1}{5}\left(\frac{3}{9} + \frac{2}{7} + \frac{2}{7} + \frac{6}{16} + \frac{3}{10}\right) = 0.316.$$

ST instead pools the pixel counts over all six frames, including the authentic frame $t = 7$:

$$ST = \frac{16}{16+23+16} = \frac{16}{55} = 0.291.$$

Frame $t = 7$ is excluded from S because it is authentic. Its six false-positive cells enter the denominator of ST. This is why ST penalizes masks predicted on authentic frames, whereas S measures localization only where a ground-truth forgery exists.

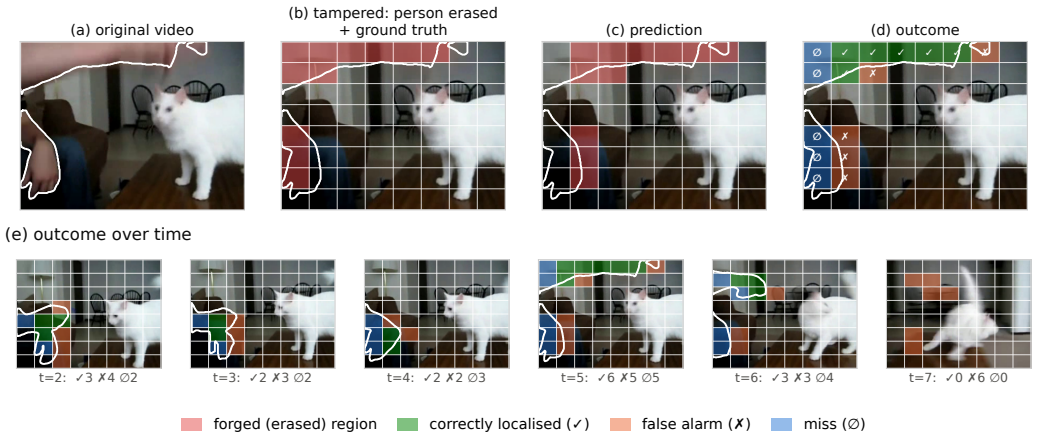


Figure 6: Worked localization example. (a) The original frame. (b) The forged frame, from which the person has been removed. The white contour marks the ground-truth forged region. (c) The predicted mask. (d) The pixel-level outcomes for frame $t = 5$. (e) Pixel-level outcomes over frames 2–7. Green cells are true positives, orange cells are false positives, and blue cells are false negatives.

E Additional Qualitative Results

The four clips of Fig. 4 are the two per dataset with the highest S of Chain of Evidence (memory, FakeParts) among clips on which it leads every baseline in S by at least 0.10, excluding clips of the source video shown in Fig. 3. In Figs. 4 and 8 the frame shown has the median IoU of Chain of Evidence (memory, FakeParts) among the forged frames after the seven-frame warm-up. Eligible clips have at least eight such frames and a mask covering 2 to 50 percent of the frame.

Fig. 7 repeats Fig. 3 on five more clips, one from AINPAINT and four from FakeParts. They are drawn from the upper third of clips by the IoU of Chain of Evidence (memory, FakeParts) and show the mechanism rather than the typical case. In each row the evidence is sharper than the response and the memory mask is tighter than the framewise mask. Presence with the memory bank stays at one on the frames where the framewise presence dips.

Fig. 8 repeats Fig. 4 on seven clips on which the memory bank changes the temporal decision. On each, the framewise model has T at most 0.7 and Chain of Evidence (memory, FakeParts) has T at least 0.9. The clips are spread over the upper 60 percent of such clips by the S of Chain of Evidence (memory, FakeParts), so the masks range from good to poor. Chain of Evidence (memory, FakeParts) is ahead in T on every row. A baseline is ahead in S on six of the seven rows, mostly AINPAINT-ST and DINOv3-IML on the AINPAINT clips. No clip appears in two figures.

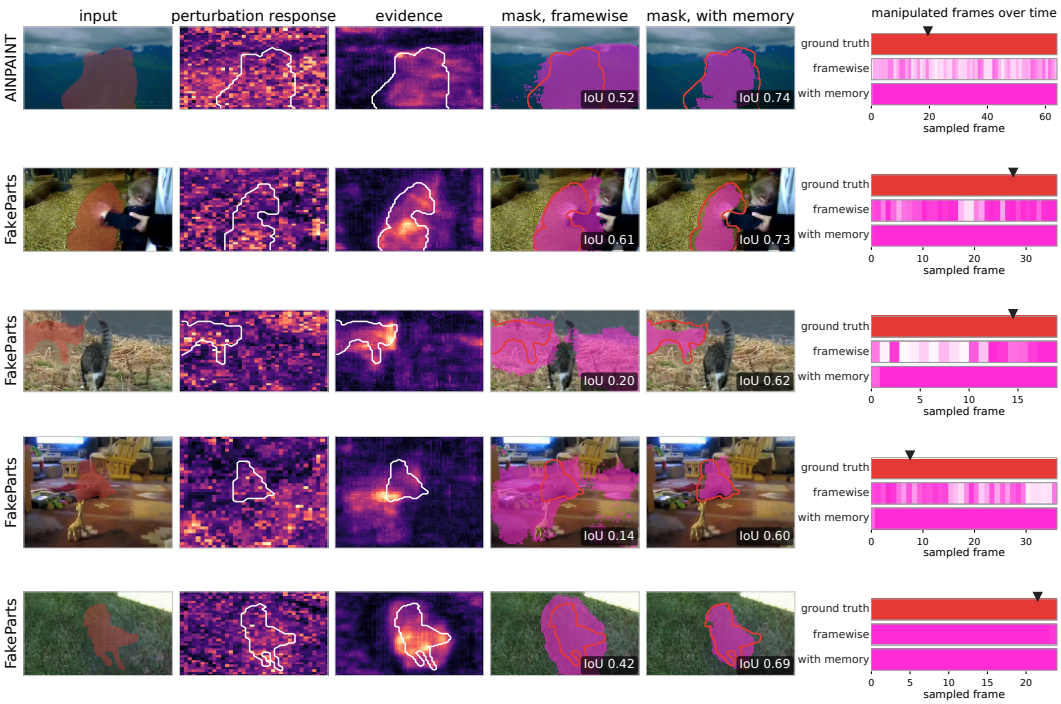


Figure 7: Fig. 3 on five more clips, one per row: input with the ground truth in red, perturbation response, evidence, mask of the framewise model, mask with the memory bank, and forged frames over time as ground truth, framewise presence and presence with the memory bank. The IoU with the ground truth at the shown frame is printed on each mask.

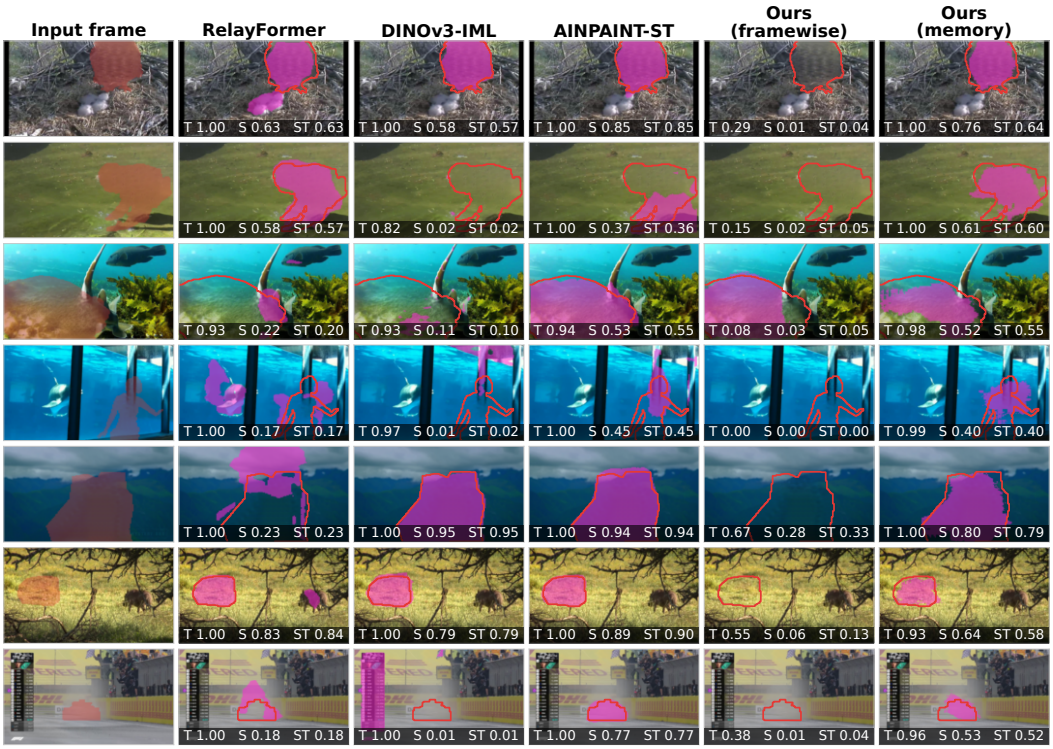


Figure 8: Fig. 4 on seven clips on which the memory bank turns a missed or partial framewise detection into a full one. Rows one to four are FakeParts and rows five to seven are AINPAINT. Layout and scores as in Fig. 4, with T, S and ST for each clip.

F INSIDE THE MODEL

Fig. 9 looks at one AINPAINT frame in more detail. The perturbation response is noisy across the frame. The evidence transformer produces a map that is bright only at the forgery. The query token placed at the forgery attends to the cells around it in the first layer, in the last layer and on average over layers. The transformer therefore gathers its evidence locally. The memory bank changes the coarse feature mostly inside the forgery, and the mask decoded from the changed feature is tight where the framewise mask spreads. Fig. 10 follows the memory bank over one FakeParts clip. At the first frame the queue is empty and the model falls back to the framewise mask, which over-segments. From the second frame on, with one stored frame, the mask closes on the forged region and stays there while the queue fills to seven frames. From the first stored frame on, presence with the memory bank is one on every frame, while the framewise presence dips. The relative change the memory bank makes to the feature rises within the first stored frame and then stays constant. The effect of the memory bank therefore does not depend on a full queue.

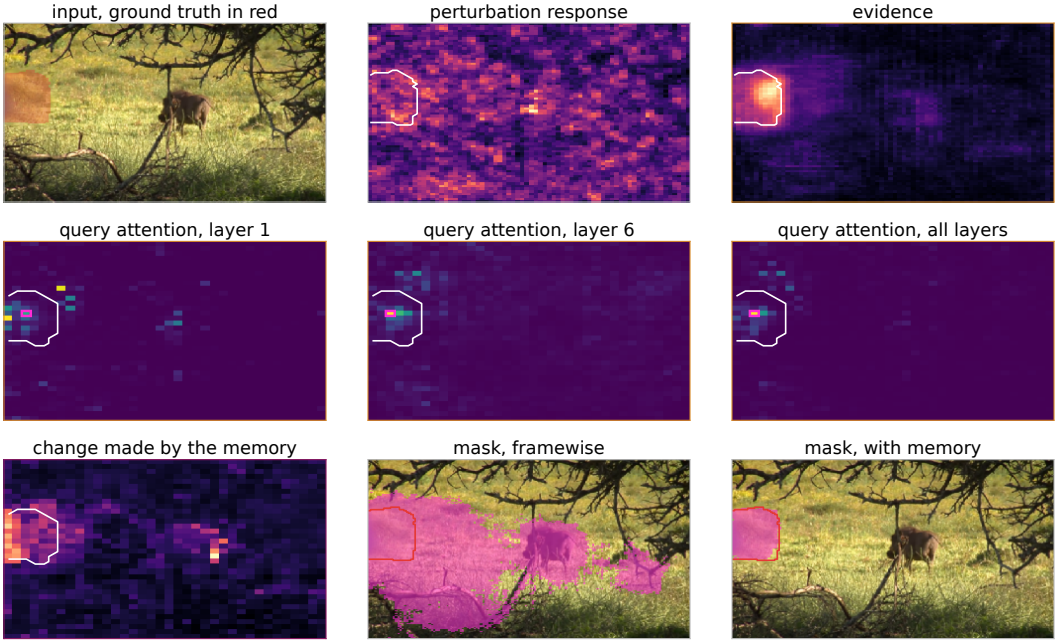


Figure 9: One AINPAINT frame in detail. Top row: input with the ground truth in red, perturbation response of the frozen SAM 3 encoder, and evidence. Middle row: attention of the query token at the forgery, marked in magenta, in the first and last layers and averaged over all six layers of the evidence transformer. Bottom row: change made by the memory bank to the coarse feature, framewise mask, and mask with the memory bank. On the heat maps, white outlines mark the ground truth.

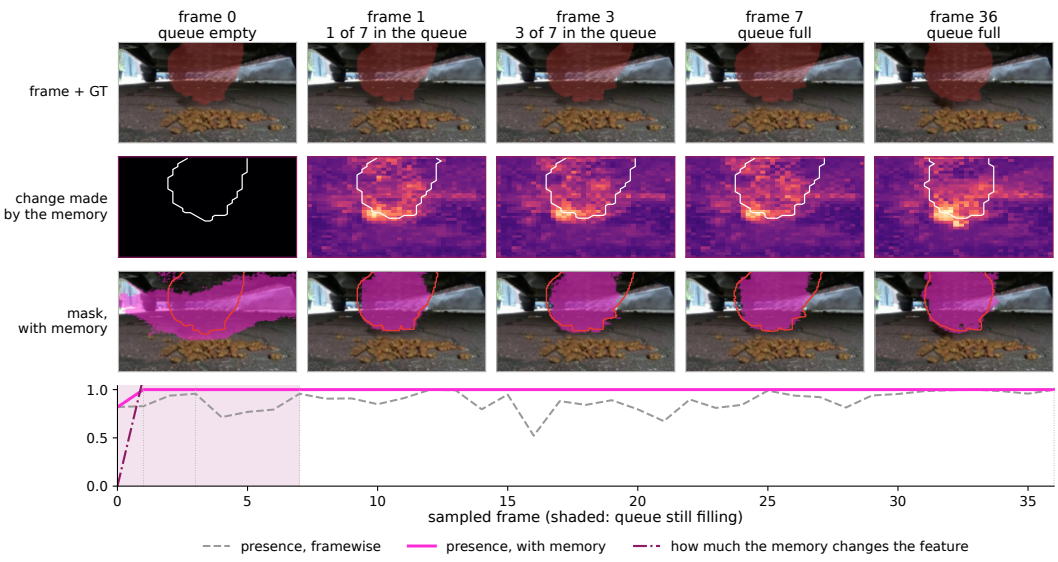


Figure 10: The memory bank over one FakeParts clip at five sampled frames, from an empty queue to a full one. Top row: frame with the ground truth in red. Middle row: change made by the memory bank to the coarse feature. Bottom row: mask with the memory bank. The plot gives presence of the framewise model and with the memory bank over the clip, and the relative size of the change the memory bank makes to the feature. The shaded frames are those before the queue is full.

G Failure Modes

Four failure modes follow from Tables 2 and 3 and from Fig. 11. First, the memory bank trained on FakeParts harms transfer to generator-level forgeries (Feng et al., 2026; Bao et al., 2026). On TASLE its D and T fall below Chain of Evidence (framewise) and are the lowest of all rows, and on ActivityForensics its T falls below it. Fig. 11 shows the effect on a TASLE, an ActivityForensics and a TVIL clip. The framewise model flags the forged frames, and the memory bank either drives presence to zero for the whole clip or flags every frame. Chain of Evidence (memory, AINPAINT) stays near Chain of Evidence (framewise) on TASLE and exceeds it on ActivityForensics. The loss therefore depends on the video training set. The tables do not separate the memory bank from the presence head. Second, the memory bank can shrink a correct mask on local edits. On the two AINPAINT clips in Fig. 11 the framewise mask covers the forgery and the memory bank shrinks it to a fraction of the region, while presence stays high. Such clips are rare on FakeParts, the training domain, and more common on AINPAINT. Third, in-distribution video methods stay ahead on their own data. AINPAINT-ST leads AINPAINT on T, S and ST and MSLoc-PG leads TASLE on T. Chain of Evidence (memory, AINPAINT) narrows but does not close the gap on AINPAINT. Fourth, on images the localization gains do not carry over to detection. Chain of Evidence leads L on five of the seven sets. DINOv3-IML leads D on three of the four out-of-distribution sets that have a D column and on the average D over the sets no method trained on.

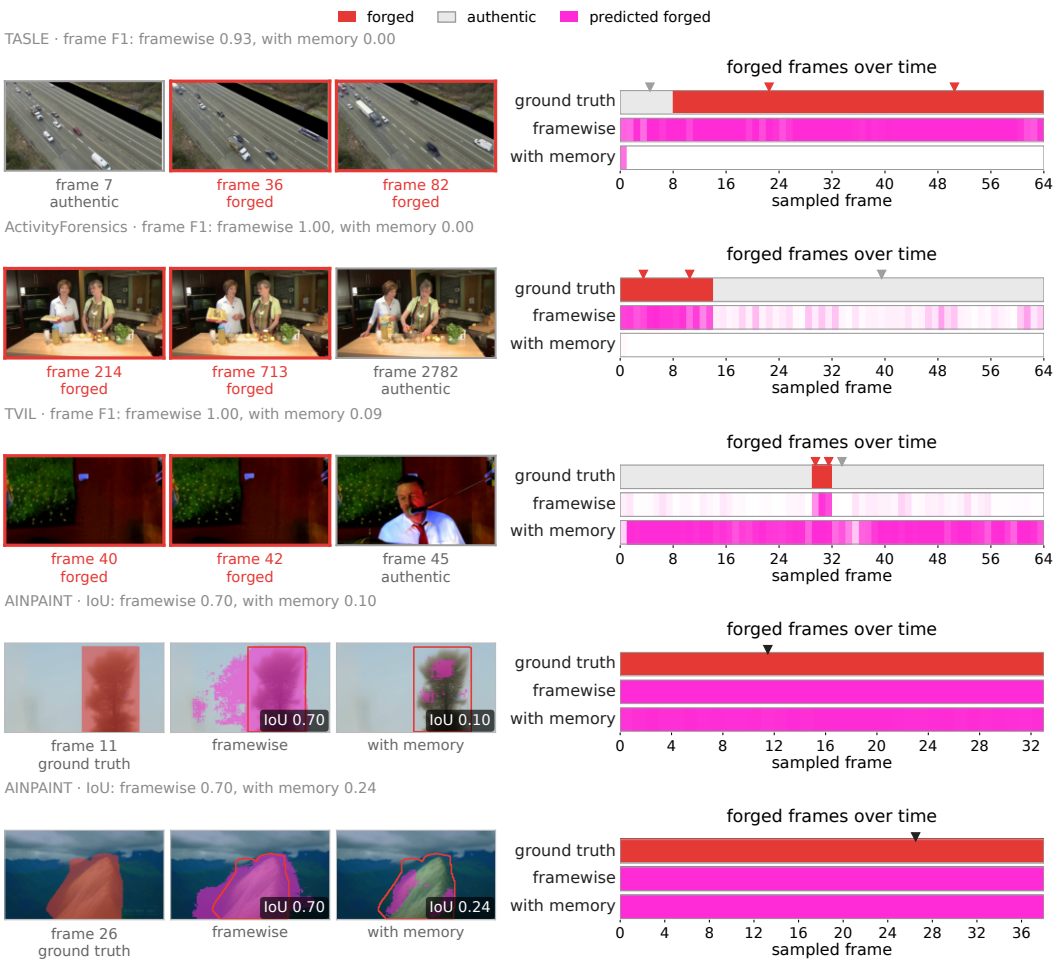


Figure 11: Failure cases. Rows one to three: one clip each from TASLE, ActivityForensics and TVIL on which the memory bank loses a detection the framewise model made, with three sampled frames and their labels on the left. Rows four and five: two AINPAINT clips on which the memory bank shrinks a mask the framewise model had right, with the input, both masks and their IoU. Right: forged frames over time, as ground truth, framewise presence and presence with the memory bank.

H Robustness Protocol and Scores

We draw 100 DiffSeg30k test images (17 authentic, 83 forged) and 100 AINPAINT test clips (11 authentic, 89 forged) with seed 42. Each draw keeps the authentic-to-forged ratio of its test set. Every degradation acts at native resolution, before the model resizes the frame. JPEG uses quality 80 and 50. Resizing scales each frame down to 75% and 50% with area interpolation and back up bilinearly. Gaussian blur uses kernels of 5 and 11 pixels, with standard deviations of 1.1 and 2.0. Gaussian noise has a standard deviation of 5 and 15 on the 0 to 255 scale and is seeded per sample. H.264 re-encodes each clip with libx264 at CRF 23 and 30 and keeps every frame. The clean rows reproduce the published predictions exactly. Table 9 lists every score, including D and T on video.

Table 9: **Robustness scores** behind Fig. 5. Framewise and memory share the frozen image stage.

	Image		Video, framewise				Video, memory			
Condition	D	L	D	T	S	ST	D	T	S	ST
Clean	0.930	0.484	0.942	0.836	0.161	0.186	0.933	0.884	0.334	0.452
JPEG q80	0.889	0.387	0.919	0.745	0.134	0.182	0.942	0.928	0.294	0.407
JPEG q50	0.895	0.373	0.912	0.693	0.108	0.152	0.947	0.928	0.243	0.338
Resize 75%	0.927	0.466	0.930	0.801	0.137	0.170	0.935	0.925	0.303	0.413
Resize 50%	0.908	0.447	0.913	0.764	0.119	0.152	0.941	0.930	0.281	0.389
Blur k=5	0.906	0.438	0.919	0.748	0.106	0.139	0.941	0.929	0.280	0.386
Blur k=11	0.870	0.414	0.905	0.589	0.064	0.113	0.942	0.933	0.228	0.322
Noise σ=5	0.909	0.413	0.899	0.694	0.118	0.177	0.930	0.924	0.291	0.401
Noise σ=15	0.852	0.306	0.916	0.656	0.092	0.143	0.936	0.920	0.215	0.291
H.264 CRF 23	–	–	0.930	0.786	0.130	0.167	0.941	0.895	0.253	0.360
H.264 CRF 30	–	–	0.900	0.807	0.115	0.140	0.935	0.913	0.194	0.273