Interpretation-Aware Image Manipulation Detection

Anonymous Author(s)

Abstract

Image manipulation has been a common practice since the early days of photography, and has evolved into a sophisticated threat with the rise of powerful image editing tools and generative AI. It can be used to spread misinformation, manipulate public opinion, and cause harm to individuals and communities. In this article, we study the problem of determining whether an image is misleading, i.e. edited in a misleading way. We introduce a risk framework to assess the *intrinsic interpretive shift* (ITS) of an ordered image pair, which evaluates whether the final edited image changes the interpretation supported by a source reference image. We show our risk framework can be operationalized with the Content Provenance and Authenticity (C2PA) provenance framework. We estimate ITS with a single-call vision-language implementation of the ITS decision procedure, which first checks for meaningful visual differences, then tests whether the differences induce interpretive shift, and finally determines whether the shift is fully explained by permissible transformations rather than content-level changes. The system outputs a binary label (potentially misleading or not) together with interpretive clues and a short justification. To test the approach and show that ITS is a measurable proxy for misleading edits, we built a benchmark of 418 ordered pairs, drawn from real-world fact-check cases and meaning-preserving transformations. The best-performing model configuration (gpt-5.4) achieves 88.2% precision, 94.2% recall, and 96.5% balanced accuracy. The code is available at https://anonymous.4open.science/r/ITS_detection-443F/.

CCS Concepts

• **Computing methodologies** → **Computer vision**; • **Information systems** → *Multimedia information systems*; • **Security and privacy** → Social aspects of security and privacy.

Keywords

Interpretive shift, Image manipulation, Benchmark, Deepfakes, Vision-language models, Disinformation, C2PA, Misleading edits

ACM Reference Format:

Anonymous Author(s). 2026. Interpretation-Aware Image Manipulation Detection. In . ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Traditionally, the fight against image manipulation has been asymmetrical: while it is relatively easy for malicious actors to create and distribute deceptive images, it is much harder for defenders to detect, without further context, whether an image is deceptive or not. Earlier approaches to image forensics have focused on identifying low-level artifacts left by editing tools, such as inconsistencies in lighting, shadows, or compression [17, 29, 33]. Their effectiveness is limited, though, as current generative AI models can produce images that are visually indistinguishable from real photographs, and can even mimic the artifacts of specific editing tools. Moreover, image manipulation is not inherently malicious. Most edits are truly benign, such as cropping and resizing.

To balance the asymmetry between image (malicious) manipulation and detection, researchers have turned into analysing whether an image is consistent with a given context [2, 3, 17, 21]. Notable examples are tweets that include an image and a caption, where the caption provides a context that can be used to verify the consistency of the image. This is a promising and popular direction with an established recurring competition ecosystem[1], as it allows for a more holistic assessment of the image's content in relation to its intended meaning.

We observe that the Content Provenance and Authenticity (C2PA) standard[2] has already been adopted by major camera manufacturers (Canon, Nikon, Sony) and media platforms including Adobe, Microsoft, and Google [5], and its broad adoption is set to further increase the availability of contextual information. In particular, C2PA metadata can provide a thumbnail of the source image [5, 30] and a detailed record of the editing history. The source image is particularly useful, as it represents the ground truth of the content, and can be used to verify whether the final image is a deceptive version of the source. Therefore, in this article, we set up the study of image manipulation detection in a provenance-aware setting, where the source image is available as a trusted reference, and the task is to determine whether the final edited image is misleading.

The two areas of research that are closer to our problem are Visual Transformation Reasoning (VTR) and Image Difference Captioning (IDC), discussed in detail in Section 2. Both can describe differences between image pairs, but they do not decide whether a transformation should be treated as misleading or non-misleading. The example in Figure 1 shows the gap: both edited pairs are visibly different from the same source image (thumbnail), but only one may induce a viewer to misinterpret the scene.

Contributions. Adversarial manipulation assessment is a complex problem, and we do not attempt to solve it in full. Instead, we make the following contributions:

(1) We introduce *intrinsic interpretive shift* (ITS) as a measurable proxy for the risk of misleadingness, and develop a three-stage ITS decision procedure that decomposes the decision into visual-difference screening, interpretive-shift testing, and a final check for plausible explanations based on permissible transformations. ITS is a pairwise property of an ordered pair (I, I') that captures

[1] https://detecting-cheapfakes.github.io/
[2] https://c2pa.org/

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
Conference'17, Washington, DC, USA
© 2026 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

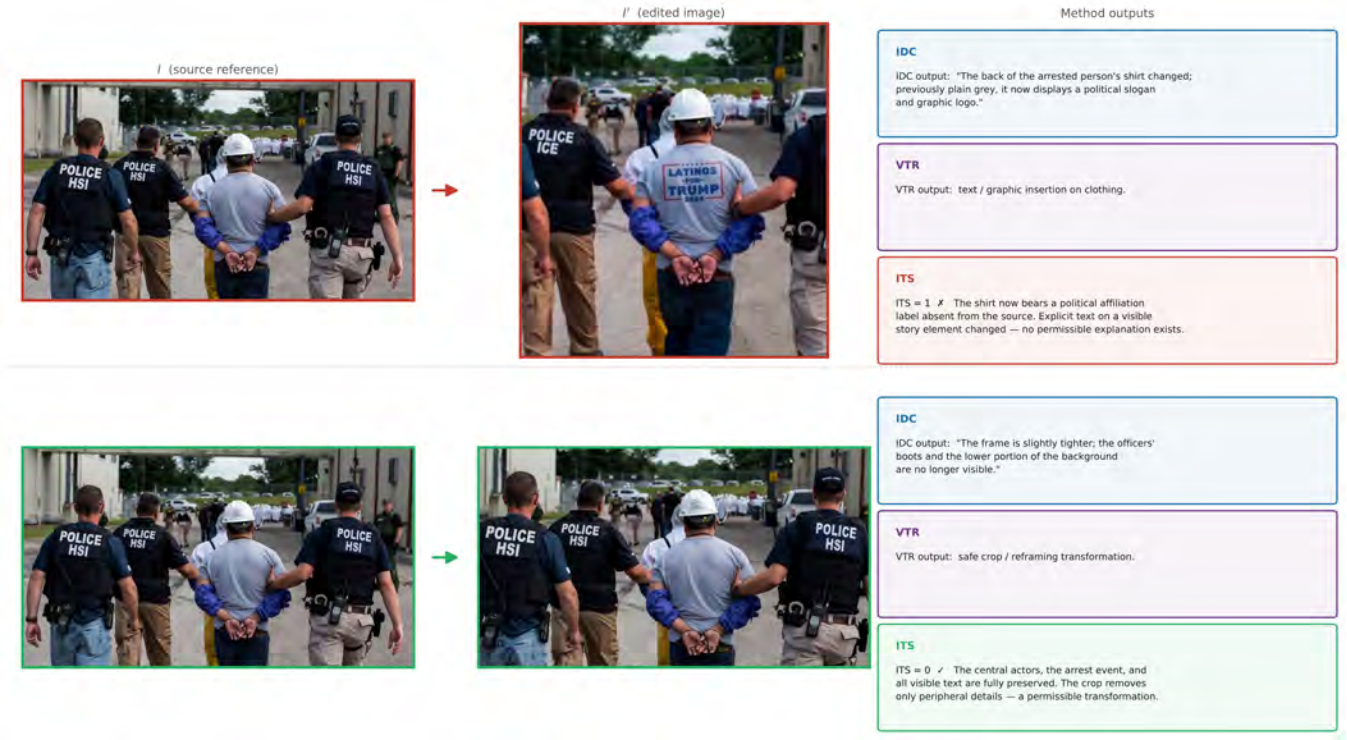


Figure 1: Two different edits of the same source image. The edit at the top adds a message to the arrested person's shirt with a clear manipulation intent. The edit at the bottom crops the image without changing the story. Both IDC and VTR correctly identify that something changed, but are unable to tell whether the change is misleading. Our ITS detection method, instead, flags the top edit as potentially misleading and the bottom edit as not. Fact-check reference: https://www.snopes.com/fact-check/ice-latinos-for-trump-shirt/.

whether the final edited image I' changes the interpretation of the source I supported by visible content. ITS is stronger than asking only whether an edit exists, but weaker than asking whether harm or deception actually occurred. The procedure operates on pairs where the trusted source may be available only as a low-resolution thumbnail, as in C2PA.

(2) We construct, to the best of our knowledge, the first benchmark of misleading edits from real-world fact-check cases, together with stratified meaning-preserving controls. This is essential because the hard cases in our setting are often not obvious fabrications. Benchmarks for VTR and IDC are not enough, as they lack both real-life and synthetic adversarial manipulations.

(3) We evaluate the approach against the benchmark across ten vision-language models. The best-performing configuration, gpt-5.4, achieves F1 = 0.9109 ± 0.0160 and balanced accuracy = 0.9647 ± 0.0115 with false positives (FP) = 4.8 ± 0.4 and false negatives (FN) = 2.2 ± 0.8. We discuss false positives and false negatives as a practical safety tradeoff, since a deployable system must avoid both over-flagging benign edits and missing genuinely misleading transformations.

Structure of the article. Section 2 discusses the related work. Section 3 formalises the ITS concept grounded in the consensus across journalism ethics codes. Section 4 describes how we built the benchmark, covering both the fact-check-derived positive pairs and the stratified meaning-preserving controls. Section 5 reports the empirical evaluation. Lastly, Section 6 summarises the contributions and outlines directions for future work.

2 Related Work

Three major areas of research relate to our problem: computer-vision techniques, content provenance, and risk assessment frameworks for manipulated media. We discuss each next and show how they motivate the ITS framework.

Computer-Vision Techniques. Image forensics and tampering detection is a long-studied problem, which looks for pixel-level inconsistencies [19], texture and artifact traces [11], compression, resizing, and other distribution shifts [29, 33]. This line of work is important for forensics analysis, where any sign of tampering is relevant. However, as illustrated in Figure 1, some edits can be benign, and making that distinction is key to minimize false positives.

The research lines that are closer to our problem are Visual Transformation Reasoning (VTR) [13] and Image Difference Captioning (IDC) [16]. IDC asks a model to describe visual differences between two similar images in natural language [12, 15, 16, 27]. VTR, instead, aims to recover the transformations that were applied to a source image to obtain a final image [13, 14]. Both tasks are relevant to our problem because they require pairwise comparison at a semantic level. However, their objective remains descriptive

and lack a decision-making framework to determine whether the difference is potentially misleading.

Content-Provenance. Content provenance is a critical tool in the fight against image manipulation, which we exploit to operationalize the ITS concept. In media provenance, the C2PA ecosystem is perhaps the most popular and widely adopted standard. C2PA provides a structured workflow for recording the history of an image's creation and modifications, and allows users to cryptographically verify an image's history [5, 30]. This enables traceability, auditing, and accountability across the image supply chain [22, 24]. As acknowledged by C2PA's own documentation, provenance answers whether a transformation history is trustworthy. It does not, however, answer whether a transformation in the provenance chain has introduced misleading, manipulative or deceptive content. Our notion of ITS is motivated precisely by this gap: even when an image provenance is correct, it may still be misleading.

Risk Assessment Frameworks. Media outlets have long been concerned about the potential for image manipulation to mislead audiences and cause harm; their reputation depends on the integrity of the images they publish. That is why media outlets often publish codes of ethics that include guidelines for photo editing and manipulation [4, 23, 34]. These codes agree on one key principle: edits that change what a photograph communicates are not allowed; changes that only affect appearance are. They make a distinction between processing (tonal and color adjustment) and manipulation (adding, removing, or rearranging scene elements). These codes, however, rely on practitioner judgment rather than a concrete decision procedure for an image pair.

Regulators and policymakers have also been concerned about the risk of manipulated media. For example, the EU AI Act (Art. 50) [8] requires disclosure when AI systems produce content "likely to deceive a reasonable person." The DSA (Art. 34) [7] requires platforms to assess systemic risks from manipulated media. The EU Code of Practice on Disinformation [6] and the Partnership on AI Responsible Practices for Synthetic Media [28] both require more transparency when deception is more likely. These frameworks are useful for classification and governance, but they do not give a binary decision framework to assess whether a specific image edit is misleading.

Academics have also proposed frameworks to analyze media manipulation, particularly in the context of deepfakes. For example, Paris and Donovan [26] show that harm correlates with interpretive shift, not with how technically sophisticated the manipulation is. Kiela et al. [18] show that harm in image pairs cannot be detected from visual appearance alone; it requires understanding what the image is saying. The ITS framework we propose in the next framework is motivated by those insights.

From our study of the literature, we conclude that existing frameworks for media harm and manipulation all converge on the same principle: edits that change the *story or proposition* an image supports are likely misleading, while edits that only change presentation or appearance likely are not. No existing framework, however, provides a concrete decision procedure for a specific ordered image pair. We address this gap next.

3 The ITS Framework

We consider the *story* of an image to be characterized by five elements: (1) the central actor or subject, (2) the main event or action depicted, (3) key relations or interactions between entities, (4) explicit content in visible text, signs, or banners, and (5) the global scene or situation. These elements are read from the visible image content alone; no captions, metadata, or external knowledge are used.

Motivated by the consensus across the journalism ethics codes, we say that an ordered pair (I, I') exhibits *Intrinsic Interpretive Shift* (ITS) when the story of I' differs from the story of I in a way that goes beyond benign presentation adjustments. ITS is present only when at least one story element (central actor, main event, key relations, explicit text, global scene) changes in a way that has no benign presentation explanation. ITS is absent when all story-element differences have such an explanation, or when no story element changes at all.

The ITS assessment is performed by an *observer*: any agent that examines only the visible content of (I, I') and produces a binary ITS label with a justification. The observer is instructed to proceed in three steps:

- Construct the set $\Delta(I, I')$ of *candidate story-element differences*. Each element $d \in \Delta(I, I')$ is a triple (v, s, δ): a visible change v, the story element $s \in \{1, \ldots, 5\}$ it appears to affect, and a degree $\delta \in \{\text{major}, \text{minor}\}$. The degree is *minor* if only peripheral visual details of that element change while its narrative is preserved. Otherwise the degree is *major*.
- For each difference $(v, s, \delta) \in \Delta(I, I')$ with $\delta = \text{minor}$, check whether v can be attributed to a *permissible transformation*, as defined below. If so, remove (v, s, δ). For example, if a shoe disappears and cropping can explain it, the shoe is a peripheral visual detail of the central actor ($\delta = \text{minor}$), and the observer drops it.
- The observer outputs $\text{ITS}(I, I') = 1$ when $\Delta(I, I')$ is non-empty.

In the procedure above, we define the *risk* of a candidate $d = (v, s, \delta) \in \Delta(I, I')$ as binary. The risk is *low* if and only if $\delta = \text{minor}$ and v can be attributed to a permissible transformation. The risk is *high* if $\delta = \text{major}$, regardless of whether a permissible transformation can be found. For example, a crop that removes the central actor receives $\delta = \text{major}$ and is high-risk even if cropping is permissible. The set of permissible transformations is defined by the consensus across various journalism ethics codes. It consists of tonal and color adjustments, compression and re-encoding, non-directional overlays and watermarks, etc.

Figure 2 summarizes the decision procedure.

The reader may wonder why our risk framework does not consider fine-grained scores, such as a continuous variable that captures the degree of interpretive shift, or a multi-class label that captures different levels of risk. We use binary outputs because audit pipelines require a yes/no decision to flag content for review; the observer must answer whether an image pair does or does not support a different interpretation of the depicted event. A fine-grained risk score would require establishing an arbitrary threshold, for

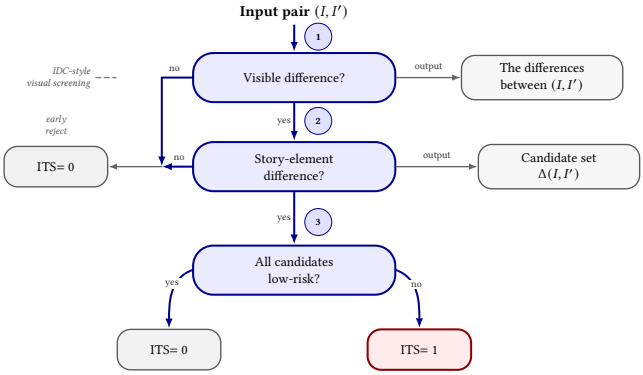


Figure 2: Decision structure for pair-only ITS estimation. The method first checks for a meaningful visible difference (Stage 1), then computes the candidate set $\Delta(I, I')$ by identifying which story element each difference affects and assessing its degree (Stage 2). Finally, Stage 3 classifies each candidate as low-risk (minor degree and attributable to a permissible transformation) or high-risk (major degree, or minor with no permissible attribution). A pair is labeled ITS-positive only when at least one high-risk candidate remains.

which we have no principled basis. Moreover, it is unclear how multiple risk labels could be aggregated into a single decision.

The remainder of this article is devoted to testing the ITS framework on real-life scenarios. We operationalize it with a vision-language workflow, and we test it against a benchmark which we built from real-world fact-check cases. We leave for future work testing the framework with human observers.

4 Benchmark Construction

Constructing a benchmark for the assessment of manipulation detection mechanisms requires a verifiable decision boundary, not only a collection of image pairs. A useful benchmark must satisfy two constraints: the positive side must contain (likely subtle) edits with the intent to manipulate, while the negative side must include visible and sometimes salient edits that are neither intended to mislead nor likely to be misinterpreted as misleading. We remark that verifiability is an essential property of the benchmark; whether an edit is misleading/manipulative or not cannot be determined by subjective judgement alone.

4.1 Limitations of Existing Pair Datasets

IDC and related datasets contain image pairs, but their pairs are not structured around interpretation-changing manipulation. They record what changed between two images for captioning or visual reasoning tasks, not whether the change alters the interpretation supported by a trusted reference. For example, we examined Spot-the-Diff [16], CLEVR-Change [27], Birds-to-Words [9], LEVIR-CC [20], and GIER [31]. No pair in these datasets carries a verified manipulation label, and the label cannot be reliably recovered after the fact because the distinction between interpretation-changing and interpretation-preserving edits was never part of their collection criteria.

4.2 Construction of the Benchmark

We built the benchmark from two sources: fact-checking websites and meaning-preserving controls. The positive side uses expert-documented manipulation cases; the negative side uses carefully chosen interpretation-preserving transformations. This contrast makes the pair-only ITS decision measurable. Each sample is a tuple (I, I', y, g), where I is the source reference, I' is the final edited image, $y \in \{0, 1\}$ is the ITS label, and g records the interpretation used to justify the label. At inference time, only (I, I') is given to the observer; y is used for evaluation and g is used for verification and audit.

Positive pairs. We do not synthesize positive examples. Instead, we collected them from publicly documented fact-check cases on websites such as Snopes, Full Fact, and AAP FactCheck [1, 10, 32]. For each candidate case, we identify the source image, recover the final edited image, and collect the evidence from the fact-check report. A pair is kept only if both images can be clearly recovered and the report explicitly supports a meaning-altering edit. In practice, many candidate pages are unusable. Some simply provide a misleading claim with a screenshot, without a clean image pair. Others provide only watermarked or annotated versions that would introduce external clues. We read each article carefully, traced the links, used reverse search when needed, and rejected cases where we could not recover a clean pair. All retained positives were checked against the published case description. Using this process, we collected 38 ITS-positive pairs. Clean source/final pairs were scarce in our curation process, so the benchmark is designed to be auditable rather than large-scale.

We remark that we are not the first to draw data from fact-checking websites. For example, VERITE [25] is built from articles collected from sources such as Snopes and Reuters, but its task is multimodal misinformation detection over image-text pairs. The annotation unit is therefore different from ours. To the best of our knowledge, no open benchmark converts such materials into ordered image pairs for image manipulation evaluation.

Negative pairs. The negatives must still contain visible changes, including some that look clearly edited, while preserving the supported interpretation. We stratify meaning-preserving controls into three difficulty levels.

- L0: *Visual-benign controls that capture routine appearance-only processing*: include JPEG recompression, resize-and-restore pipelines, mild denoising, and slight global color or contrast adjustment. These edits can introduce visible artifact changes or overall appearance shifts, but they do not alter who is present, what action is occurring, how entities are related, or what explicit text is shown. They therefore test whether the estimator can avoid treating routine image-processing traces as interpretation change.
- L1: *Semantic-benign controls that introduce small interpretation-preserving scene edits*: include removing minor background clutter, simplifying small non-salient regions, or replacing same-category textures without changing the scene-level reading. These cases are useful because they are not purely low-level transformations: something visible has changed, yet the ordinary interpretation supported by the image

should remain stable. L1 therefore helps separate interpretation from strict pixel preservation.

L2: *Benign-but-salient controls that include visually salient but non-misleading presentation changes*: include user-interface chrome, non-directional markup, safe cropping that preserves the salient event, mood-oriented but non-factual grading, and neutral local patch swaps. These cases are intentionally harder because they may look consequential at first glance. Their role is to test whether the estimator overreacts to edit salience, framing, or presentation intensity when the supported interpretation has in fact not changed.

The three control classes are important because a system that succeeds only on L0-style negatives might still be inadequate to be deployed in the wild.

To make the negative sample generation process auditable, controls are generated with a tool[3] that restricts the operator to these predefined L0/L1/L2 control classes. We retain an edit only after manually inspecting that the resulting image retains the story of the source image. The tool makes the L0/L1/L2 taxonomy operational: an operator selects a source image and a control level, applies only the allowed behavior class, records notes explaining why the result remains interpretation-preserving, and saves both the generated image and metadata for later audit.

Dataset scale and auditability. From this process, we obtained a benchmark of 418 ordered pairs: 38 fact-check-derived ITS-positive edits and 380 benign controls split into 127 L0 samples, 71 L1 samples, and 182 L2 samples. The imbalance is intentional: it reflects the deployment setting in which users are expected to encounter more benign edits than manipulated ones. Fact-check-derived pairs are assigned ITS-positive ground truth and benign controls are assigned ITS-negative ground truth. To keep the benchmark auditable, we preserve ordered-pair identity throughout curation, so each record points to a specific source reference I and edited image I'. For the constructed negatives, we record the transformation type used to generate each sample.

Benchmark scope and transferability. The benchmark is intentionally small and auditable rather than large and automatically labeled. The 38 ITS-positive pairs are drawn from real fact-check cases and carry verifiable ground truth, which is a stronger quality guarantee than synthetic or crowdsourced manipulation labels. The stratified negative controls (L0–L2) are designed to prevent the evaluation from collapsing into simple edit detection: the hardest negatives (L2) include salient overlays and mood-oriented color grades that could superficially resemble misleading edits but preserve the supported interpretation. However, the benchmark should not be treated as a general-purpose manipulation testbed. Its positives are drawn from fact-check cases with a clean source/final pair, discarding those that cannot be recovered in full even if manipulation was present. Its negatives are generated with a specific tool and a specific taxonomy of controls, which may not capture all possible meaning-preserving edits in the wild.

Data availability. The 38 ITS-positive pairs are derived from news photographs whose copyright is held by third-party wire services; we cannot redistribute these images directly. We therefore release the benchmark as a pointer dataset: for each positive pair, we provide the URL of the source fact-check article, a reconstruction note, and the full annotation metadata. All metadata is released in a structured format to support independent replication of the evaluation.

[3]Available at https://github.com/liyu0x/ITS_detection.git

5 Empirical Evaluation

The goal of this section is to evaluate the ITS framework on the benchmark we constructed.

5.1 Evaluation Setup

Models. We evaluated ten large vision-language models (LVLMs) in the role of the *observer*, covering a range of architectures, parameter counts, and deployment modes. One model, **gpt-5.4**, is accessed via the OpenAI API and serves as the commercial API configuration in our evaluation. The remaining nine models, namely **Qwen3.5-2B**, **Qwen3.5-4B**, **Qwen3.5-9B**, **Ministral-3B**, **Ministral-8B**, **gemma-3-4B**, **gemma-3-12B**, **gemma-4-E2B**, and **gemma-4-E4B**, are deployed locally on a single GPU server and run entirely offline.

Testing both commercial and locally-deployed models serves two purposes. First, it measures how much performance changes between the commercial API configuration and smaller, self-hosted alternatives. This distinction matters in practice: C2PA-based provenance auditing is designed for deployment across diverse organizational contexts, including privacy-sensitive environments where API access to external services is not acceptable. Second, evaluating a broad spectrum of model sizes (2B to 12B parameters for local models) allows us to characterize how an ITS prompt implementation behaves under different budgets. Across all models, the same prompt and execution pipeline are used without model-specific tuning, so the comparison reflects each model's behavior under the same ITS prompting protocol. For locally-deployed models, inference parameters follow each publisher's recommended defaults.

For implementation and reporting, we instantiate the permissible transformations from Section 3 with the following concrete operations: mild tonal and color adjustments, compression and re-encoding, resizing, mild denoising, blurring, masking, watermarking, safe cropping that does not remove or reframe the central actor or the main event, neutral arrows or pointer lines, highlights or bounding boxes that annotate without reinterpreting, and application chrome, or screenshot artifacts introduced during capture or republication. Blurring and masking are treated equivalently: both may be applied to conceal peripheral details, such as the faces of bystanders or drivers in a collision scene, without altering the event being depicted. Watermarking is permissible because ITS assessment is restricted to the photographic content of the image; overlaid text or logos that do not obscure or reframe story-relevant content do not change the supported reading.

Evaluation Metrics. We report confusion-matrix counts, namely **true positives** (TP), **false positives** (FP), **true negatives** (TN), and **false negatives** (FN), and four metrics, namely (M1) **Precision**, (M2) **Recall**, (M3) **F1**, and (M4) **Balanced Accuracy**. Given the class imbalance in our benchmark (38 positive, 380 negative), no single metric is sufficient on its own. Precision measures the cost of false alarms: a system that flags too many benign edits as

misleading is not deployable in practice. Recall measures the cost of missed detections: a system that misses too many genuinely misleading edits is just not useful. F1 summarizes the trade-off between Precision and Recall, while Balanced Accuracy accounts for the imbalance by averaging true positive and true negative rates. We therefore discuss all four metrics jointly rather than selecting a single primary criterion.

Input and protocol. To assess whether the framework can be operationalized with a pair-only input, where the source image is taken from the C2PA provenance record, we normalize the source reference image to a longest side of 1024 pixels before being passed to the model. This is meant to *simulate* the C2PA use case, where the source reference is expected to be a clean image with a thumbnail assertion within its provenance metadata. Each ordered pair is evaluated by each LVLM under the same prompt, and we report averages over five independent repeated runs to account for model non-determinism.

We use the following single-call operational prompt to implement the ITS decision procedure in the experiments. The wording is kept compact for repeated evaluation, but the substantive constraints remain the same: the model may use only visible evidence, treat minor differences through the permissible-transformation filter, and output a single ITS label with justification.

System Prompt

You evaluate Intrinsic Interpretive Shift for an ordered image pair. Use only visible evidence from the two images. Do not use captions, metadata, provenance, retrieval results, news context, or outside knowledge. Do not infer real identities, locations, organizations, or events unless they are explicitly visible as readable text or unambiguous symbols.

Story elements:

(1) central actor
(2) main event
(3) relations
(4) text
(5) scene

Intrinsic Interpretive Shift is present iff Image B changes at least one story element of Image A in a way that cannot be explained by permissible transformations. Permissible transformations include low-level edits such as blur, masking, safe cropping that does not remove or reframe the central actor or main event, watermarking, recompression, denoising, resizing, and mild tone or color changes, as well as presentation or UI overlays such as arrows, circles, boxes, highlights, pointer lines, app or window chrome, video-player controls, progress bars, and screenshot artifacts.

Procedure:

(1) Identify visible differences.
(2) For each difference with `degree=minor`, check whether it can be attributed to a permissible transformation. If so, remove it. `degree=major` differences are never removed, even when caused by a nominally permissible transformation.
(3) Output `ITS = 1` iff remaining differences are non-empty; otherwise output `ITS = 0`.

Output exactly one JSON object.

User Prompt

You are given an ordered image pair:

- Image A: trusted reference image
- Image B: edited or compared image

Decide whether replacing Image A with Image B introduces Intrinsic Interpretive Shift. Return exactly one JSON object with this schema:

```
{
  "remaining_differences": [
    {
      "visible_change": "string",
      "story_element": "central actor | main event |
           relations | text | scene",
      "degree": "minor | major",
      "justification": "string"
    }
  ],
  "filtered_summary": "string",
  "its_label": 0,
  "final_justification": "string"
}
```

Rules:

- Only include differences that remain after filtering.
- Minor differences may be removed if explained by permissible transformations.
- Major differences remain.
- `its_label` = 1 iff `remaining_differences` is non-empty; otherwise 0.
- Use only visible evidence.

Output only the JSON object and nothing else.

5.2 Comparison results among models

Table 1 summarizes the performance results of each model across the different evaluation metrics. Models such as gemma-3-4B and gemma-3-12B achieve near-perfect recall (FN ≈ 0–1) but low precision (0.12–0.17), indicating that they predict ITS= 1 for many benign controls. Conversely, Qwen3.5-2B achieves high precision (0.91) at the cost of lower recall (0.69), suggesting it is overly conservative about flagging shift. The Qwen3.5-9B and Qwen3.5-4B models sit at an interesting middle point: very high recall (0.979) combined with moderate precision (0.60–0.65), keeping FN near zero while accumulating more false positives than the 2B variant.

Two qualitatively distinct failure modes emerge from the results. The first is *over-flagging* (high FP): a model interprets a permissible transformation as a story-element change. This is the dominant failure mode for most local models, and it reflects a limitation in the model's ability to attribute visible changes to benign presentation operations. The recurrent example in Figure 3 shows this pattern: the source and control show the same street mural scene, but the control applies a safe crop that removes only peripheral framing while preserving the central mural, passers-by, and scene reading. Several gpt-5.4 runs nevertheless flag this benign control as ITS-positive and attribute the error to a changed mural action, showing that salient reframing can lead the model to hallucinate a story-element change. The second is *under-flagging* (high FN): a model misses a genuine interpretation-changing edit, typically when the shift is composition-dependent, involves a peripheral story element,

Table 1: Main results on the 418-pair benchmark (38 positive, 380 negative), sorted by F1. Results are averages over five repeated runs (± std). TP, FP, TN, FN are mean counts. Yellow cells with bold text mark the best value in each column; lower is better for FP and FN, and higher is better otherwise.

Model	TP	FP	TN	FN	Precision	Recall	F1	Bal. Acc.
gpt-5.4	35.8 ± 0.8	4.8 ± 0.4	375.2 ± 0.4	2.2 ± 0.8	0.8817 ± 0.0117	0.9421 ± 0.0220	**0.9109 ± 0.0160**	**0.9647 ± 0.0115**
gemma-4-E4B	29.6 ± 0.5	4.8 ± 0.8	375.2 ± 0.8	8.4 ± 0.5	0.8609 ± 0.0192	0.7789 ± 0.0136	0.8177 ± 0.0107	0.8832 ± 0.0064
Qwen3.5-2B	26.2 ± 0.4	**2.6 ± 0.9**	**377.4 ± 0.9**	11.8 ± 0.4	**0.9105 ± 0.0236**	0.6895 ± 0.0118	0.7845 ± 0.0114	0.8413 ± 0.0063
Qwen3.5-9B	37.2 ± 0.4	20.6 ± 2.7	359.4 ± 2.7	0.8 ± 0.4	0.6450 ± 0.0329	0.9789 ± 0.0118	0.7771 ± 0.0194	0.9624 ± 0.0040
Qwen3.5-4B	37.2 ± 0.8	24.6 ± 1.5	355.4 ± 1.5	0.8 ± 0.8	0.6021 ± 0.0248	0.9789 ± 0.0222	0.7456 ± 0.0201	0.9571 ± 0.0124
Ministral-3B	34.0 ± 0.0	52.0 ± 1.2	328.0 ± 1.2	4.0 ± 0.0	0.3954 ± 0.0041	0.8947 ± 0.0000	0.5484 ± 0.0055	0.8789 ± 0.0016
gemma-4-E2B	19.2 ± 0.8	25.6 ± 2.9	354.4 ± 2.9	18.8 ± 0.8	0.4301 ± 0.0316	0.5053 ± 0.0222	0.4643 ± 0.0262	0.7189 ± 0.0121
Ministral-8B	**38.0 ± 0.0**	150.6 ± 1.2	229.4 ± 1.2	**0.0 ± 0.0**	0.2015 ± 0.0016	**1.0000 ± 0.0000**	0.3354 ± 0.0017	0.8018 ± 0.0013
gemma-3-12B	37.0 ± 0.0	179.0 ± 3.9	201.0 ± 3.9	1.0 ± 0.0	0.1713 ± 0.0048	0.9737 ± 0.0000	0.2914 ± 0.0046	0.7513 ± 0.0051
gemma-3-4B	**38.0 ± 0.0**	266.8 ± 2.2	113.2 ± 2.2	**0.0 ± 0.0**	0.1247 ± 0.0013	**1.0000 ± 0.0000**	0.2217 ± 0.0012	0.6489 ± 0.0026

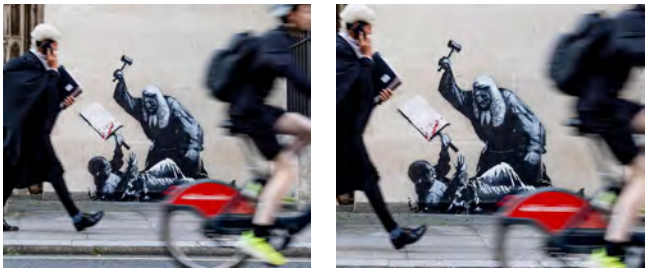


Image A: source **Image B: benign control**

Figure 3: A recurrent over-flagging case used. Image B is a safe-crop benign control. Several gpt-5.4 runs nevertheless mark this benign control as ITS-positive. Original source: https://fullfact.org/culture-and-society/banksy-edited-mural-flags/.



Image A: source **Image B: edited image**

Figure 4: A representative under-flagging case. Image B changes the surrounding people's expressions and emotional interaction. This is missed by various local-models. Original source: https://www.aap.com.au/factcheck/spoofed-images-spread-after-trump-rally-shooting/.

or is introduced by a subtle texture or scene-context change. One example is given in Figure 4.

The 3D comparison in Figure 5 separates the evaluated models by their false-negative and aggregate-metric profiles. gpt-5.4 occupies the position closest to the ideal operating point ($d = 0.213$, where d denotes the normalised Euclidean distance to the optimal point), showing that the commercial API configuration evaluated here has an advantage over the local models. However, the gap to the best local model is smaller than one might expect: Qwen3.5-2B achieves F1 = 0.7845 and balanced accuracy = 0.8413 with only 2 billion parameters, placing it second in F1 among all evaluated models. This

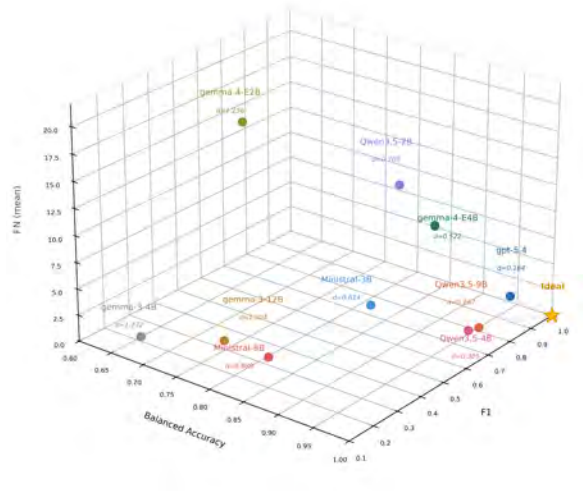


Figure 5: Three-dimensional comparison of all ten models. The X-axis shows balanced accuracy, the Y-axis shows F1, and the Z-axis shows mean false negatives. The gold star marks the ideal operating point (Bal.Acc = 1, F1 = 1, FN = 0). Dashed lines connect each model to the ideal point; the annotated value d is the normalised Euclidean distance to the optimal point. Models in the upper-right foreground (low FN, high F1 and balanced accuracy) are closest to the ideal.

suggests that smaller on-premise models may be useful, although they remain below gpt-5.4 overall.

We end with a discussion of gpt-5.4's performance, which offers the best trade-off among all evaluated models. gpt-5.4 achieves F1 = 0.9109 ± 0.0160, precision = 0.8817 ± 0.0117, recall = 0.9421 ± 0.0220, and balanced accuracy = 0.9647 ± 0.0115, with TP = 35.8 ± 0.8, FP = 4.8 ± 0.4, TN = 375.2 ± 0.4, and FN = 2.2 ± 0.8. The false-negative count reflects the inherent difficulty of the task: a small number of subtle interpretation-changing edits are missed, particularly when the shift is composition-dependent or affects only a peripheral story element. The low variance across runs indicates that the result is stable under our repeated-run protocol.

5.3 Baseline Comparison and Ablation study

To separate the contribution of each part of the ITS decision procedure, we define three variants of our method that remove or ablate specific components of the main method. This includes one traditional direct-prediction baseline and two ablations of the main method.

Baseline 1 (B1), direct pair-label baseline, receives the same ordered pair (I, I') as the main method but removes the ITS framework scaffolding from the prompt. The model is asked only to decide whether replacing I with I' introduces Intrinsic Interpretive Shift and to provide a short reason. This is the traditional baseline: it tests how much can be recovered from direct pairwise judgment alone, without explicit story-element definitions, permissible-transformation rules, or decision-procedure intermediate checks.

Ablation 1 (A1), direct permissibility ablation, receives the same ordered pair (I, I') and asks for a direct binary ITS judgment without exposing the story-element definition used by the main ITS framework. No intermediate outputs are required: the model is not asked to enumerate visible differences, assign story elements, assign degrees, or explicitly perform the permissibility check before issuing the final label. A1 retains only the high-level distinction between meaningful interpretive shifts and benign presentation adjustments, so it tests whether this permissibility framing alone is sufficient without the framework's story-element decomposition.

Ablation 2 (A2), no-permissibility ablation, keeps the visible-difference and story-element reasoning stages but removes the final permissible-transformation filter. This condition tests whether the explicit permissibility check is responsible for suppressing false positives on benign controls that are visually salient but meaning-preserving.

For ablation and baseline comparisons, we focus on gpt-5.4 as the underlying model. Table 2 reports the results. B1, the direct pair-label baseline, reaches F1 = 0.8601 ± 0.0277 but raises the false-positive count to 9.8 ± 2.7, showing that direct binary judgment is less conservative than the full ITS decision procedure. A1 removes the story-element definition and decision-procedure intermediate outputs while retaining only a high-level permissibility framing. In this condition, A1 reaches F1 = 0.8704 ± 0.0238 with 6.4 ± 1.4 false positives and 3.8 ± 0.7 false negatives, below both the main system and B1 on average. A2 removes the permissible-transformation filter and produces 109.0 ± 2.2 false positives, showing that the final permissibility check is the main mechanism that suppresses visually salient but meaning-preserving controls.

5.4 Sanity checks

Beyond the main evaluation, we conducted two sanity checks to see whether the model's predictions are anchored to the actual visual content of the pair rather than to spurious signals or prompt bias. In the first sanity check (**R1**), we replace both images in each ordered pair with the same image, so that $I' = I$. Because the two inputs are identical, no story element can change, and the correct prediction is ITS = 0 for every sample. A model that hallucinates differences or is biased toward predicting shift will fail here. We measure *consistency*, defined as the fraction of samples for which the model correctly predicts 0. In the second sanity check (**R2**), we replace I' with an image drawn from a different, unrelated pair of images, so that the two inputs share no narrative context. Because the two images depict unrelated scenes, at least one story element will differ, and the correct prediction is ITS = 1 for every sample. A model that is biased toward predicting no shift, or that fails to detect obvious cross-context differences, will fail here. For auditability, each R2 run records the replacement image's sample identifier, pair identifier, and image path, and we verify that the replacement pair identifier differs from the original pair identifier for every evaluated sample. We again measure consistency, defined as the fraction of samples for which the model correctly predicts 1.

Table 3 reports the outcomes of the two sanity checks across models. On R1, where both inputs are identical ($I' = I$), six models achieve perfect consistency, while the four gemma-family configurations produce sanity errors. The largest error counts occur for gemma-3-4B and gemma-4-E2B, indicating a strong tendency in those configurations to hallucinate shift or over-predict ITS = 1 even when no visual difference exists. Figure 6 shows a representative hallucinated-difference example. On R2, where I' is drawn from an unrelated pair, all evaluated models achieve high consistency (0.989–1.000): nine models are perfectly consistent, and only gemma-4-E2B has residual under-flagging errors. The remaining R2 errors are cases where visually similar public-event scenes are mistaken for the same context; Figure 7 shows a representative example.

The two sanity checks provide additional evidence that these failure modes are model-dependent. R2 (cross-pair) yields perfect consistency for nine models and near-perfect consistency for gemma-4-E2B, ruling out a general tendency to predict no-shift regardless of content. R1 (identical pairs), however, exposes hallucination or over-flagging behavior in several local models, especially gemma-3-4B and gemma-4-E2B. The remaining R2 errors are concentrated in gemma-4-E2B, suggesting a residual under-flagging failure mode in that model rather than a benchmark-wide ambiguity.

5.5 Reason Alignment with Expert Rationales

So far, our focus has been on binary classification, which signals whether a shift is potentially present. We end this section by evaluating the model's stated reason for its ITS judgment, which dedicated users would likely consult to understand the model's reasoning and to decide whether to trust the judgment or to review the pair manually. We therefore evaluated whether a model's stated reason matches the rationale documented by professional fact-checkers. This evaluation is restricted to the 38 ITS-positive samples derived from fact-check articles, as these are the samples for which external expert rationales are available. For each positive sample and each of the five repeated runs, we use only the model's final-justification field and compare it with the expert rationale extracted from the corresponding fact-check article, yielding 190 judged explanations per model.

We include all positive-sample outputs in this evaluation, including cases where the model incorrectly predicts ITS, because the goal is to assess whether the stated final justification captures the expert rationale.

Table 2: Comparison of the main prompt implementing the ITS decision procedure, B1 direct baseline, redesigned A1 direct permissibility ablation, and A2 no-permissibility ablation, all evaluated using gpt-5.4. Yellow cells with bold text mark the best value in each column; lower is better for FP and FN, and higher is better otherwise.

Method	TP	FP	TN	FN	Precision	Recall	F1	Bal. Acc.
Main	35.8 ± 0.8	**4.8 ± 0.4**	**375.2 ± 0.4**	2.2 ± 0.8	**0.8817 ± 0.0117**	0.9421 ± 0.0220	**0.9109 ± 0.0160**	**0.9647 ± 0.0115**
B1	36.0 ± 0.0	9.8 ± 2.7	370.2 ± 2.7	2.0 ± 0.0	0.7888 ± 0.0465	0.9474 ± 0.0000	0.8601 ± 0.0277	0.9608 ± 0.0036
A1	34.2 ± 0.7	6.4 ± 1.4	373.6 ± 1.4	3.8 ± 0.7	0.8430 ± 0.0308	0.9000 ± 0.0197	0.8704 ± 0.0238	0.9416 ± 0.0111
A2	**37.4 ± 0.5**	109.0 ± 2.2	271.0 ± 2.2	**0.6 ± 0.5**	0.2555 ± 0.0041	**0.9842 ± 0.0129**	0.4057 ± 0.0058	0.8487 ± 0.0066



Image A: `RW-0024/real.png`



Image B: identical copy of Image A

Figure 6: Representative R1 false positive given by gemma-4-E2B. The pair contains two identical images, hence the prediction should be ITS-negative. The model nevertheless believes the missile bodies have different marks.



Image A: `RW-0037_benign_reencode_q60`



Image B: `RW-0040_l2a_nondirectional_markup`

Figure 7: Representative R2 false negative given by the model gemma-4-E2B. The expected prediction is ITS-positive, as they are two different shots.

We use an LLM judge (gpt-4.1-mini) to score explanation alignment; we chose this model as an independent judge to avoid using the same model being evaluated, and because its instruction-following quality is sufficient for the three-point rubric. The judge receives only the model's `final_justification`, the expert rationale extracted from the corresponding fact-check article, and the three-point rubric. It assigns one of three scores: no alignment, partial alignment, or exact alignment. For example, in the no-alignment case RW-0054, the expert rationale states that the injury photo was AI-edited, while Ministral-3B-Instruct says the differences are only "minor presentation changes". In the partial-alignment case RW-0027, the expert rationale identifies a composite image combining Miami with storm clouds from Kansas, while gemma-3-12B says only that the scene changed "from a field to a cityscape". In the exact-alignment case RW-0030, the expert rationale states that the rescue image was digitally altered from a CNN photo of rescuers helping a man at a fence, and Ministral-3B-Instruct likewise says that the central actor is replaced and the scene's interaction changes. This reason-alignment evaluation is not applied to benign controls, which are constructed locally and do not have external expert analysis. In Table 4, *Mean Align.* is the average alignment score on the three-point scale, where no, partial, and exact alignment are scored as 0, 1, and 2, respectively. *Exact Align.* reports the fraction of judged explanations that fully match the expert rationale, and *Partial/Exact* reports the fraction that receive either partial or exact alignment. *Pos. Recall* is the label-level recall.

Table 4 shows that strong positive-sample recall does not by itself imply a useful ITS observer. Several models can explain positive

Table 3: Sanity check results. R1 uses identical images ($I' = I$), with expected ITS = 0. R2 uses different images with expected ITS = 1. Consistency is the fraction of samples correctly predicted. Red cells with white bold text mark records with sanity errors.

Model	R1 consistency	R1 wrong	R2 consistency	R2 wrong
gpt-5.4	1.0000 ± 0.0000	0.0 ± 0.0	1.0000 ± 0.0000	0.0 ± 0.0
gemma-4-E4B	0.9354 ± 0.0040	27.0 ± 1.7	1.0000 ± 0.0000	0.0 ± 0.0
Qwen3.5-2B	1.0000 ± 0.0000	0.0 ± 0.0	1.0000 ± 0.0000	0.0 ± 0.0
Qwen3.5-9B	1.0000 ± 0.0000	0.0 ± 0.0	1.0000 ± 0.0000	0.0 ± 0.0
Qwen3.5-4B	1.0000 ± 0.0000	0.0 ± 0.0	1.0000 ± 0.0000	0.0 ± 0.0
Ministral-3B	1.0000 ± 0.0000	0.0 ± 0.0	1.0000 ± 0.0000	0.0 ± 0.0
gemma-4-E2B	0.4201 ± 0.0049	242.4 ± 2.1	0.9895 ± 0.0050	4.4 ± 2.1
Ministral-8B	1.0000 ± 0.0000	0.0 ± 0.0	1.0000 ± 0.0000	0.0 ± 0.0
gemma-3-12B	0.8445 ± 0.0000	65.0 ± 0.0	1.0000 ± 0.0000	0.0 ± 0.0
gemma-3-4B	0.2330 ± 0.0044	320.6 ± 1.9	1.0000 ± 0.0000	0.0 ± 0.0

Table 4: Reason-alignment results. Exact alignment is the fraction of judgments scored as fully aligned; partial/exact counts both partial and full alignment. The last for metrics are copied from Table 1 for context.

Model	Mean Align.	Exact Align.	Partial/Exact	Pos. Recall	Precision	Recall	F1	Bal. Acc.
gpt-5.4	1.9316	**0.9632**	0.9684	0.9421	0.8817	0.9421	**0.9109**	**0.9647**
Qwen3.5-9B	**1.9421**	**0.9632**	0.9789	0.9789	0.6450	0.9789	0.7771	0.9624
Ministral-8B	**1.9421**	0.9579	**0.9842**	**1.0000**	0.2015	**1.0000**	0.3354	0.8018
Qwen3.5-4B	1.9263	0.9474	0.9789	0.9789	0.6021	0.9789	0.7456	0.9571
gemma-3-4B	1.7842	0.8263	0.9579	**1.0000**	0.1247	**1.0000**	0.2217	0.6489
Ministral-3B	1.7632	0.8421	0.9211	0.8947	0.3954	0.8947	0.5484	0.8789
gemma-3-12B	1.6579	0.7684	0.8895	0.9737	0.1713	0.9737	0.2914	0.7513
Qwen3.5-2B	1.5737	0.7474	0.8263	0.6895	**0.9105**	0.6895	0.7845	0.8413
gemma-4-E4B	1.4579	0.7000	0.7579	0.7789	0.8609	0.7789	0.8177	0.8832
gemma-4-E2B	1.2000	0.5263	0.6737	0.5053	0.4301	0.5053	0.4643	0.7189

cases in terms that match the expert rationale, yet still over-flag benign controls in the full benchmark. For example, Ministral-8B reaches high reason alignment and perfect positive recall, but its low precision in Table 1 shows that it predicts many benign controls as misleading. Conversely, Qwen3.5-2B has the highest precision but lower reason alignment and positive recall. The gpt-5.4 configuration offers the best overall trade-off: its final justifications align with expert rationales for most positive cases while maintaining the strongest F1 and balanced accuracy on the full benchmark. Thus, reason alignment complements but does not replace the main label-level evaluation, especially because false-positive control must be measured on benign controls.

6 Conclusion

We introduced Intrinsic Interpretive Shift (ITS) as a measurable proxy for the risk that an image edit is misleading, and showed how the ITS decision procedure can be operationalized within a C2PA provenance workflow using only the source thumbnail and the final edited image. The framework does not require full-resolution source images, captions, or any external retrieval. This means the approach can be integrated directly into existing C2PA audit pipelines as a post-processing step: whenever a manifest is verified, the thumbnail and the final image can be passed to the ITS estimator to flag pairs for human review. Three concrete contributions support this claim. First, the ITS framework itself, grounded in the consensus across journalism ethics codes and regulatory frameworks, provides a principled decision procedure for any ordered image pair. Second, we built the first pair-only benchmark for image manipulation evaluation from expert-verified fact-check cases and stratified benign controls, establishing a challenging evaluation setting that cannot be reduced to simple edit detection. Third, we evaluated the framework across ten large vision-language models and found that performance is strongly model-dependent: gpt-5.4 gives the best overall model-comparison result, while the diagnostic experiments show that explicit permissible-transformation checking is important for controlling false positives.

The main limitations of our work are the benchmark's coverage, which is limited to 38 fact-check-derived pairs, and the absence of a human-observer baseline. Our curation process is conservative by design, prioritising verifiability over scale. It means, however, that the benchmark neither includes AI-generated manipulations nor adversarially edits designed specifically to evade detection. Future work should address both gaps.

References

[1] AAP FactCheck. [n. d.]. AAP FactCheck. https://www.aap.com.au/factcheck/. Fact-checking website. Accessed 2026-04-01.

[2] Md Shad Akhtar et al. 2023. Multimodal Automated Fact-Checking: A Survey. In *Findings of EMNLP*.

[3] Shivangi Aneja, Chris Bregler, and Matthias Nießner. 2021. COSMOS: Catching Out-of-Context Misinformation with Self-Supervised Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[4] Associated Press. 2023. AP Statement of News Values and Principles. Online. https://www.ap.org/about/news-values-and-principles/ Includes photo and digital standards. Accessed 2026-04-01.

[5] Coalition for Content Provenance and Authenticity (C2PA). 2026. C2PA Specifications (Version 2.3). Online. https://spec.c2pa.org/specifications/specifications/2.3/index.html Accessed 2026-02-11.

[6] European Commission. 2022. Strengthened Code of Practice on Disinformation. Online. https://digital-strategy.ec.europa.eu/en/policies/code-practice-disinformation Accessed 2026-04-01.

[7] European Commission. 2023. Digital Services Act: Risk Assessment and Mitigation Framework. Online. https://digital-strategy.ec.europa.eu/ Accessed 2026-02-12.

[8] European Parliament and Council of the European Union. 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council on Artificial Intelligence (AI Act). Official Journal of the European Union. Article 50 on transparency obligations for AI-generated and manipulated content.

[9] Maxwell Forbes, Christine Kaeser-Chen, Piyush Sharma, and Serge Belongie. 2019. Neural Naturalist: Generating Fine-Grained Image Comparisons. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 708–717.

[10] Full Fact. [n. d.]. Full Fact. https://fullfact.org/. Fact-checking website. Accessed 2026-04-01.

[11] Jie Gao, Marco Micheletto, Giulia Orrù, Sara Concas, Xiaoyi Feng, Gian Luca Marcialis, and Fabio Roli. 2024. Texture and artifact decomposition for improving generalization in deep-learning-based deepfake detection. *Engineering Applications of Artificial Intelligence* 133 (2024), 108450.

[12] Zihang Guo, Zhiqiu Lin, Xiaodan Chen, and Yuxuan Wang. 2022. CLIP4IDC: An Image Difference Captioning Model Based on CLIP. In *Proceedings of the Asia-Pacific Chapter of the Association for Computational Linguistics (AACL-IJCNLP)*. 33–42.

[13] Xin Hong, Yanyan Lan, Liang Pang, Jiafeng Guo, and Xueqi Cheng. 2021. Transformation driven visual reasoning. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*. 6903–6912.

[14] Xin Hong, Yanyan Lan, Liang Pang, Jiafeng Guo, and Xueqi Cheng. 2023. Visual Reasoning: From State to Transformation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 9 (2023), 11352–11364. doi:10.1109/TPAMI.2023.3268093

[15] Mehdi Hosseinzadeh, Yunchao Wang, Yunzhi Li, and Xiaojuan Qi. 2021. Image Change Captioning by Learning from an Auxiliary Task. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2725–2734.

[16] Harsh Jhamtani and Taylor Berg-Kirkpatrick. 2018. Learning to Describe Differences Between Pairs of Similar Images. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 4024–4034. doi:10.18653/v1/D18-1436

[17] VijayaKumar Kadha, Sambit Bakshi, and Santos Kumar Das. 2025. Unravelling digital forgeries: A systematic survey on image manipulation detection and localization. *Comput. Surveys* 57, 12 (2025), 1–36.

[18] Douwe Kiela, Suvrit Firoozi, Suhas Ratnasamy, Lisa Anne Hendricks, Matthew Muckley, Shayne Seibert, et al. 2020. The Hateful Memes Challenge: Detecting Hateful Memes on Social Media. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[19] Chenqi Kong, Anwei Luo, Shiqi Wang, Haoliang Li, Anderson Rocha, and Alex C Kot. 2025. Pixel-inconsistency modeling for image manipulation localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2025).

[20] Chenyang Liu, Rui Zhao, Hao Chen, Zhengxia Zou, and Zhenwei Shi. 2022. LEVIR-CC: A Large Dataset for Remote Sensing Image Change Captioning. *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022), 1–13.

[21] Grace Luo, Trevor Darrell, and Anna Rohrbach. 2021. NewsCLIPpings: Automatic Generation of Out-of-Context Multimodal Media. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

[22] Md Morshed Alam and Weichao Wang. 2021. A Comprehensive Survey on the State-of-the-art Data Provenance Approaches for Security Enforcement. *arXiv e-prints* (2021), arXiv–2107.

[23] National Press Photographers Association. 2023. NPPA Code of Ethics. Online. https://nppa.org/resources/code-ethics Accessed 2026-04-01.

[24] Bofeng Pan, Natalia Stakhanova, and Suprio Ray. 2023. Data provenance in security and privacy. *Comput. Surveys* 55, 14s (2023), 1–35.

[25] Stefanos Papadopoulos et al. 2023. VERITE: A Benchmark for Multimodal Misinformation Detection on Real-World Cases. *arXiv preprint arXiv:2304.14133* (2023).

[26] Britt Paris and Joan Donovan. 2019. *Deepfakes and Cheap Fakes: The Manipulation of Audio and Visual Evidence*. Report. Data & Society Research Institute.

[27] Dong Huk Park, Trevor Darrell, and Anna Rohrbach. 2019. Robust change captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4624–4633.

[28] Partnership on AI. 2023. Responsible Practices for Synthetic Media: A Framework for Collective Action. Online. https://syntheticmedia.partnershiponai.org Accessed 2026-04-01.

[29] Cecilia Pasquini, Irene Amerini, and Giulia Boato. 2021. Media forensics on social media platforms: a survey. *EURASIP Journal on Information Security* 2021, 1 (2021), 4.

[30] Leonard Rosenthol. 2022. C2PA: the world's first industry standard for content provenance (Conference Presentation). In *Applications of Digital Image Processing XLV*, Vol. 12226. SPIE, 122260P.

[31] Yaxin Shi, Xiaodan Xu, Yang Xu, Ya Tong, Liang Chen, and Jinhui Tang. 2020. A Benchmark and Baseline for Language-Driven Image Editing. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*.

[32] Snopes. [n. d.]. Snopes. https://www.snopes.com/. Fact-checking website. Accessed 2026-04-01.

[33] Luisa Verdoliva. 2020. Media forensics and deepfakes: an overview. *IEEE journal of selected topics in signal processing* 14, 5 (2020), 910–932.

[34] World Press Photo. 2024. What counts as manipulation? Online. https://www.worldpressphoto.org/contest/verification-process/what-counts-as-manipulation World Press Photo Contest verification process. Accessed 2026-04-01.