TRUST THE ID, QUESTION THE LOGIC: RETHINKING ID VERIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Identity documents are the most important line of defense between a verified citizen and an impersonator, and that line is breaking in ways we have not yet learned to see. The global identity verification ecosystem has long anchored its fraud detection capabilities on what can be seen: pixel-level anomalies, morphed photographs, spliced regions, and font irregularities. These are real threats, and the community has built real defenses. But the adversary has moved on. A new class of document fraud, logical manipulation, has quietly outpaced every existing defense. These forgeries carry no visual trace, no textual aberration, and no compression artifact. Instead, they embed contradictions into the semantic fabric of the document itself. They are, by every perceptual measure, perfect fakes. This paper presents the first systematic study of logical manipulation in Aadhaar cards, *proactively* anticipating emerging threats before they become widespread or exploitable at scale. We introduce **FactID**, the largest Aadhaar card manipulation dataset, comprising 78,702 samples spanning a rigorously defined taxonomy of template-level and content-level logical manipulations, alongside visual face-level manipulation. Through exhaustive benchmarking of image forensic classifiers, recent manipulation detectors, and state-of-the-art document verification models, we demonstrate that every evaluated system fails, showing that logical manipulation is not a corner case. It is an open frontier, and this work is its first map.

1 INTRODUCTION

The digitization of India's civic infrastructure represents one of the most ambitious identity experiments in human history. With over 1.4 billion Aadhaar numbers issued by the Unique Identification Authority of India (UIDAI), the Aadhaar card has become the singular proof of existence for over a billion citizens, the common key to banking, subsidies, healthcare, taxation, and government welfare. It is not merely a document; it is a trust contract between the state and the individual. And that contract is under systematic, escalating attack. Fake Aadhaar cards are no longer a fringe concern, they constitute a documented, nationwide crisis spanning AI-generated forgeries, illegal enrolment rackets, biometric tampering rings, and large-scale fake document operations, collectively causing financial losses running into hundreds of crores of rupees[1] [2] [3] [4] [5]. This threat is not unique to India; identity document (ID) fraud is a global crisis, with forged national IDs enabling illegal border crossings in the European Union and financial fraud via fake driver's licenses in the United States. What makes the Aadhaar ecosystem uniquely high-stakes is its centrality: no other national identity system simultaneously links biometrics, banking, taxation, and welfare for over a billion people. *But is visual forgery, the kind detectable by the human eye or pixel-level forensic tools, the only threat we must defend against?*

The Threat That Verification Systems Cannot See: In early 2025, UP Police dismantled a sophisticated gang that had tampered with Aadhaar biometric data across 12 states, cloning authorized operator credentials and modifying fingerprint scanners to accept silicone-molded fake

[1] Moneylife: AI Creating Genuine-looking Fake Aadhaar, PAN Cards
[2] NDTV: People Submit Fake Aadhaar, PAN, Walk Away With Rs.11 Crore
[3] The Print: Illegal Aadhaar Card Enrolment Racket Busted in Rajasthan's Hanumangarh
[4] Times of India: ATS Uncovers Aadhaar Fraud via Biometrics Tampering
[5] NDTV Topic Page: Fake Aadhaar Card

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.

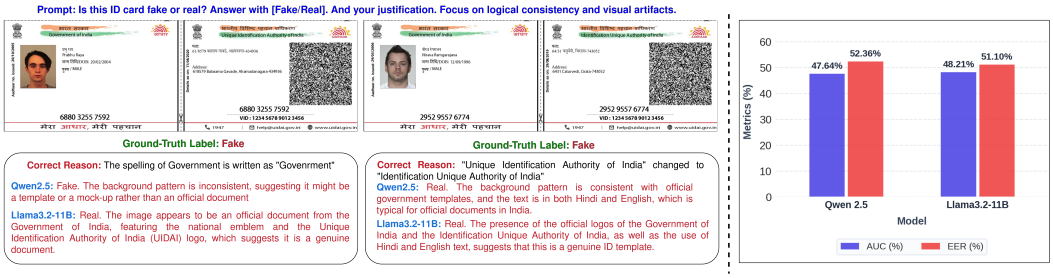

Figure 1: **Failure modes of Qwen2.5 & Llama3.2-11B on logically manipulated Aadhaar cards.** Both samples have the ground-truth label **Fake**, yet the models exhibit two distinct failure modes: in one case, the model predicts *Fake* correctly but provides an incorrect justification, while in the other, both the prediction and its justification are incorrect.

prints[6]. Operating through a network of 200–300 agents, the gang altered names, addresses, and dates of birth, not by editing a photograph or splicing a pixel, but by writing contradictions directly into the logical fabric of the document through the UIDAI updated pipeline itself. This is not the forgery the research community has been defending against. This is something far more dangerous: *manipulation that is logically embedded, not visually painted.*

What does it mean for a document to be *logically* manipulated? Consider an Aadhaar card where the enrollment date precedes the cardholder's date of birth, or where an extra digit is appended to a valid 12-digit Aadhaar number, producing a 13-digit sequence that visually resembles a legitimate identifier yet fails the Verhoeff checksum that every genuine Aadhaar number must satisfy, or where the gender field contradicts the name associated with the cardholder. One way to introduce such contradictions is by modifying the fields after the identity document (ID) has been created. More critically, the same logical inconsistencies can be embedded directly into an otherwise faithful reproduction of the blank UIDAI template, preserving its visual structure and legitimate appearance from the outset. Real-world incidents illustrate that document-level errors can originate at the generation stage, for example, India's Prime Minister was presented with a certificate containing the spelling error "Repubblic" instead of "Republic" during an official ceremony [7]. This highlights that a document need not be altered after creation; inconsistencies can be embedded at the time of generation while the document retains an authentic visual appearance. This creates a more dangerous threat, *a document is generated to appear authentic while being logically false by design*. None of these contradictions leaves a visual trace, no compression artifact, no pixel anomaly, no font irregularity, unlike conventional manipulations such as copy-paste, text overwriting, image splicing, or blending. Yet, each contradiction constitutes a provable forgery because the inconsistency exists entirely within the semantic layer of the document. This makes such manipulations fundamentally difficult for existing detection paradigms that rely primarily on visual evidence. Recent work on identity-document forensics has primarily focused on visual and biometric forms of document manipulation, including synthetic document generation, content editing, text replacement, face morphing, portrait substitution, and biometric presentation attacks ((Guan et al., 2024; Korshunov et al., 2025a; Tapia et al., 2024)), but largely leaves logically inconsistent yet visually authentic documents unexplored. The field has built increasingly sophisticated tools to answer *"does this card look authentic?"* while leaving entirely unasked the question *"does this card make logical sense?"*. To quantify this gap empirically, we evaluate Qwen2.5 ((Ahmed et al., 2025)) and Llama3.2-11B (Meta, 2024), among the most capable large vision-language models currently available. As shown in Figure 1, both models exhibit clear reasoning failures. Qwen2.5 correctly predicts *Fake* in one case but relies on an irrelevant visual cue, while predicting *Real* for another card containing a reordered agency name. Llama3.2-11B predicts both manipulated samples as *Real*, relying on template appearance, logos, and bilingual formatting. Notably, in the second case, Llama explicitly identifies the altered phrase "Identification Unique Authority of India" but does not recognize the reordering as an inconsistency or evidence of forgery. Quantitatively, Qwen2.5 and Llama3.2-11B achieve only 47.64% and 48.21% Area under the Receiver Operating Characteristic Curve (AUC-ROC), with Equal Error Rates (EERs) of 52.36% and 51.10%, respectively, indicating

[6]Times of India: How UP-Based Gang Tampered Aadhaar Biometrics Across 12 States.

[7]AI Certificate, Typos & Honour

near-random discrimination. We term this *label-without-reasoning failure*: a mode in which a correct output masks a complete absence of genuine understanding, offering false assurance of detection.

Humans Are Not the Fallback: One might reasonably ask: if automated systems fail, can human reviewers fill the gap? We conduct a controlled perceptual study involving 34 participants. Each participant is presented with seven different identity documents, randomly sampled from the benchmark, and is informed beforehand that the documents can contain template-level or content-level logical inconsistencies and/or face forgery. Participants are then asked to determine whether each document is genuine or manipulated based solely on visual inspection. Participants systematically fail to identify logical inconsistencies with an AUC of 45.10%, focusing on visual cues such as photograph quality, hologram appearance, and font style. The cognitive demand of simultaneously verifying multiple interdependent fields for semantic coherence under realistic conditions exceeds reliable human capacity. Logical manipulation is also human-invisible. More critically, even if human verification were assumed perfect, it is fundamentally not scalable for a country of over a billion citizens, where such careful field-by-field inspection would impose an unsustainable burden on the entire identity verification system. Moreover, a benchmark is still missing for training experts (humans or machines).

To address this critical gap, we present **FactID**, the first large-scale benchmark dataset and evaluation study targeting logical manipulation in Aadhaar cards. Our contributions are as follows:

1. **FactID Dataset:** We construct a dataset of *78,702 Aadhaar card samples* spanning a rigorously defined taxonomy of logical manipulation types, covering *template-level* and *content-level* manipulations proactively alongside visual face-level manipulations, including face morphing and synthetically generated faces. FactID is the largest Aadhaar card manipulation dataset in existence and the first to systematically address logical manipulation, providing a foundation for proactive security defense against emerging document threats. (Section 3).
2. **Comprehensive Baseline Evaluation:** We evaluate models spanning image forensics, document-trained forensic architectures, and an optical character recognition (OCR)-based rule-driven baseline to contextualize the difficulty of FactID.

2 Related Work

Traditional document fraud detection has relied on low-level image forensics (Das et al., 2026; Surendra Yeole et al., 2026; Tan et al., 2025), error level analysis, noise inconsistency detection, and JPEG artifact examination to identify pixel-level tampering (Popescu & Farid, 2005), with deep learning methods subsequently extending these capabilities to splicing localization (Song et al., 2026), copy-move detection, and resampling artifact classification (Liu et al., 2022; Sheng et al., 2026; Wu et al., 2019; Zhou et al., 2018). Face morphing and photograph substitution attacks have attracted dedicated research attention in border control scenarios (Ferrara et al., 2014; Robertson et al., 2017), with detection approaches leveraging texture analysis and facial landmark inconsistencies (Damer et al., 2018; Dhore et al., 2025). On the benchmark side, FantasyID (Korshunov et al., 2025a) introduced large-scale synthetic identity documents, MIDV-500 (Arlazarov et al., 2019) and MIDV-2020 (Bulatov et al., 2022) provide multi-page document benchmarks for field-level extraction, and the IJCB PAD benchmark (Tapia et al., 2024) evaluates biometric presentation attacks (Zeng et al., 2026). The advent of large vision-language models such as GPT-4V (Achiam et al., 2023), InternVL (Chen et al., 2024), and Qwen2.5 (Ahmed et al., 2025) has opened new directions for joint visual-textual document reasoning (Mathew et al., 2021; Xu et al., 2020), yet as our benchmark demonstrates, strong general multimodal reasoning does not translate to reliable logical consistency verification, LVLMs reason confidently over visual surface properties while remaining blind to interdependent field-level contradictions. Studies on trained document examiners further show that even expert reviewers fall below acceptable accuracy thresholds under realistic inspection conditions (Phillips et al., 2018; White et al., 2014), a limitation our perceptual study extends to the unstudied dimension of logical manipulation. Despite this breadth of contribution, every existing method and benchmark operates under the shared implicit assumption that fraud leaves a perceptual signal, a pixel anomaly, a morphed face, or a font irregularity. No existing work models cross-field semantic contradictions, Verhoeff checksum violations, or demographic impossibilities. FactID is

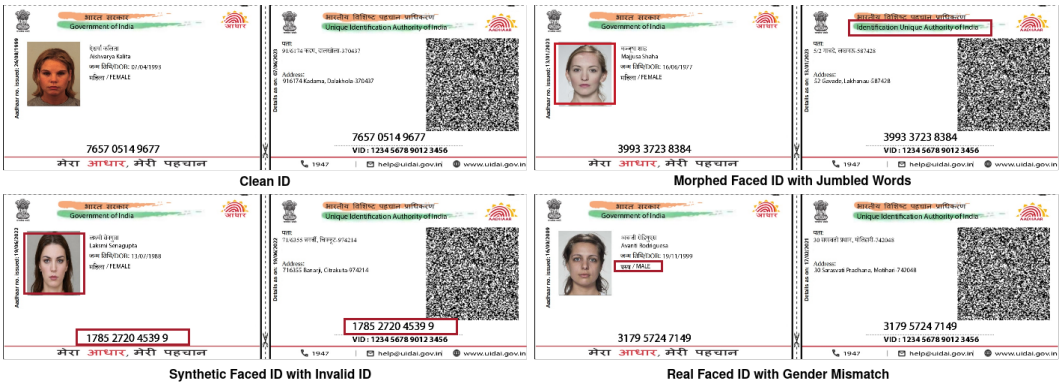

Figure 2: Representation of a few generated ID samples from the proposed FactID dataset.

the first benchmark to challenge this assumption, targeting manipulation that is visually indistinguishable from genuine documents yet provably invalid at the semantic level.

3 Proposed FactID Dataset

The core contribution of this work is the formal introduction of logical manipulation as a generative threat, in addition to facial manipulation, in Indian Aadhaar cards, towards building a verification system robust to the full spectrum of real-world forgery challenges. We present FactID, a dedicated benchmark dataset comprising real and forged Aadhaar identity cards. All real identity cards in this dataset are constructed for research purposes, with personal information independently assigned to each card; no real personal information belonging to any individual is used at any stage of dataset construction. (Figure 2).

3.1 Dataset Composition

Real IDs: 2,326 Genuine Aadhaar card samples are constructed using real neutral and frontal faces sourced from the Karolinska Directed Emotional Faces (KDEF) (Calvo & Lundqvist, 2008) dataset, `AMSL_FaceMorphImageDataset` (DeBruine & Jones, 2017), and UTKFace dataset (Zhang & Qi, 2017). A clean ID is defined as one in which the face and all textual content are authentic with respect to the card template, no manipulation is present in the face, template structure, or content fields. Although the demographic content (names, dates, Aadhaar numbers) is manually generated for research purposes from a clean original UIDAI template. The structural and visual integrity of all real ID samples is fully preserved.

Fake IDs: Forged Aadhaar card samples are organized into three distinct categories based on the type of facial content and the nature of manipulation present: (1) *Real Face + Logical Manipulation (27,776 IDs):* Genuine faces sourced from KDEF (Calvo & Lundqvist, 2008), `AMSL_FaceMorphImageDataset` (DeBruine & Jones, 2017) and UTKFace dataset (Zhang & Qi, 2017), paired with either template-level or content-level logical manipulation in the identity card fields. (2) *Morphed Face + Logical Manipulation (24,264 IDs):* Morphed face images sourced from `AMSL_FaceMorphImageDataset` (DeBruine & Jones, 2017), paired with either template-level or content-level logical manipulation. (3) *Synthetic Face + Logical Manipulation (24,336 IDs):* Fully AI-generated synthetic faces sourced from the *Synthetic Face Images Academic Dataset*[8], paired with either template-level or content-level logical manipulation.

An important design principle of FactID is that each fake ID contains exactly one logical manipulation instance, either a template-level or a content-level, alongside the facial manipulation category where applicable. This single-manipulation-per-sample design ensures clean, unambiguous ground truth labels and enables precise per-category evaluation of detector performance. Further, such a single manipulation ensures that threats remain minute and present the worst-case scenario for auto-detection. In total, FactID comprises 78,702 Aadhaar card samples spanning real and fake

[8]GeneratedPhotos/Synthetic_Face_Images_Academic_Dataset, HuggingFace

Table 1: Comparison of FactID with existing identity document datasets. **TM**: Template Manipulation, **CM**: Content Manipulation, **FM**: Face Morphing, **SF**: Synthetic Faces, **RF**: Real Faces, **LM**: Logical Manipulation, ✓: supported, ×: not supported.

Dataset	Year	Document	Total	RF	FM	SF	TM	CM	LM
FantasyID (Korshunov et al., 2025a)	2025	Multi-ID	2,358	✓	×	✓	×	✓†	×
Morphed Aadhaar (Makandar & Javeriya, 2024)	2024	Aadhaar	500	×	✓	×	×	×	×
IDNet (Guan et al., 2024)	2024	Multi-ID	837,060	✓	✓	×	×	✓†	×
IJCB PAD (Tapia et al., 2024)	2023	Multi-ID	6,000	✓	×	×	×	×	×
MIDV-2020 (Bulatov et al., 2022)	2022	Multi-ID	1,000	×	×	✓	×	×	×
MIDV-500 (Arlazarov et al., 2019)	2019	Multi-ID	500	×	×	✓	×	×	×
FactID (Ours)	–	**Aadhaar**	**78,702**	✓	✓	✓	✓	✓	✓

† FantasyID alters fields such as date of birth for *privacy de-identification* while preserving cross-field consistency. IDNet uses conventional forgery operations such as copy-move and splicing. Unlike FactID, content manipulation in FantasyID & IDNet does not constitute logical manipulation.

categories. As shown in Table 1, FactID is the first and only identity document dataset to jointly model facial manipulation, template-level, and content-level logical manipulation, surpassing all prior datasets in manipulation coverage.

3.2 Logical Manipulation Taxonomy

FactID introduces two formally defined categories of logical manipulation template-level and content-level, each with three distinct types, with a total of 11 manipulations, as follows:

Template-Level Manipulation targets fixed structural elements present identically across all genuine Aadhaar cards: (1) Spelling Errors: subtle misspellings in header text, e.g., *"AADHAAR"* → *"ADHAR"*, *"Government"* → *"Govenrment"*; (2) Word Jumbling: permutation of fixed institutional identifiers, e.g., *"Unique Identification Authority of India"* → *"Identification Unique Authority of India"*; and (3) General Information (GI) Substitution: replacement of fixed UIDAI contact fields, including helpline number, official website, and a registered email with invalid values (detailed visualization is in Figure 5 in Appendix A).

Content-Level Manipulation introduces semantic contradictions into personally identifying fields: (1) Invalid Dates: chronologically impossible combinations across issue date, details date, and date of birth fields; (2) Invalid Aadhaar Number: a 13-digit number deliberately constructed to fail the UIDAI-mandated Verhoeff checksum, and (3) Gender Mismatch: The gender field is assigned a value that directly contradicts the cardholder's profile, which is a female profile, is assigned `MALE` and a male profile is assigned `FEMALE` (detailed visualization is in Figure 5 in Appendix A). The details of the generation pipeline are mentioned in section A.2 in the Appendix.

4 Evaluation Framework and Experimental Details

We evaluate FactID across two families of existing models: (1) *Image Forensics Models* encompassing both deepfake detectors trained on facial manipulation and general image forensics models designed for broad manipulation localization, unified under a common visual detection paradigm; (2) *Document-trained Models* specialized for identity document verification. This diverse selection enables a comprehensive assessment of whether any existing detection paradigm can generalize to the manipulation of identity documents. All models are evaluated using their official checkpoints without any finetuning or adaptation.

4.1 Four-Strategy Testing Pipeline

To systematically isolate the contribution of each document region for detection performance, all models are evaluated under a unified four-strategy protocol applied consistently across every model family. As illustrated in Figure 4 in Appendix A, each input strategy selectively exposes a distinct document region: (1) *S1-Full ID:* The complete Aadhaar card is provided as-is, giving the model access to all visual and textual information simultaneously. (2) *S2-Face Only:* Only the cropped face region is provided, isolating facial manipulation signals from all document context. (3)

S3-Template Only: The content region is masked, restricting the model to fixed structural elements such as headers, footers, and institutional identifiers. (4) *S4-Content Only:* The template region is masked, restricting the model exclusively to personally identifying content fields such as name, date of birth, address, and Aadhaar number. This protocol enables fine-grained diagnosis of *where* detection succeeds or fails, whether models respond to facial cues, structural anomalies, or semantic field-level contradictions, providing actionable insight beyond a single aggregate accuracy score. A key design constraint governs real sample allocation across strategies. A single canonical set of real Aadhaar cards is shared across all four evaluation strategies, while the fake pool changes per strategy: S3 (Template Only) contains only template-level fakes, and S4 (Content Only) contains only content-level fakes. Real samples are never excluded or redistributed based on the fake type they are evaluated against, ensuring that any performance difference across strategies is driven solely by the nature of the forgery and not by shifts in the real class distribution. Under each strategy, the dataset is partitioned into training and test splits following a 70:30 ratio. Although all models are evaluated using their official pretrained weights without any fine-tuning, all evaluations are conducted exclusively on FactID's test split to ensure a consistent and reproducible evaluation protocol across all models and strategies. Full implementation details and per-manipulation sample visualizations are provided in Appendix A.

Region Extraction: Face regions are detected and cropped using `dlib`'s frontal face detector (King, 2009). The largest detected bounding box is selected and expanded by a margin of 20% in each spatial dimension to capture full facial context, including hair and neck regions. The detection is further manually verified to avoid errors. Template and content inputs are generated by selectively masking complementary regions of the full card using coordinate-based rectangular blackout via OpenCV. For the *content-masked* input (template-only evaluation), the central content region spanning $y \in [0.18, 0.93]$ of the card height is blacked out, retaining only the header, footer, and structural template elements. For the *template-masked* input (content-only evaluation), the header bar, footer bar, QR code region, and detected face area are blacked out, exposing only the personally identifying content fields. Representative visualizations of masked inputs are provided in Figure 6 in Appendix A.

4.2 Benchmark Results

For the benchmark evaluation, to empirically quantify the limitations, we choose models that collectively represent the full breadth of existing detection paradigms. For *image forensics models*, we include DoU (He et al., 2026), LTD (Yang et al., 2026), DeepID (Korshunov et al., 2025b), SAFIRE (Kwon et al., 2025), Effort (Yan et al., 2025), FreqNet (Tan et al., 2024), ProDet (Cheng et al., 2024), MMFusion (Triaridis & Mezaris, 2024), TruFor (Guillaro et al., 2023), and IML-ViT (Ma et al., 2023), spanning frequency-domain analysis, transformer-based forensics, and multi-modal fusion, each representing the state of the art on established deepfake and image manipulation benchmarks. For *document-trained models*, we include ADCD-Net (Wong et al., 2025), ASC-Former (Luo et al., 2025), DocTamper (Qu et al., 2023), all specifically designed for document-level forgery detection, in the document forensics literature (Zou et al., 2025). All models are evaluated using their officially released pretrained weights without any fine-tuning or adaptation. We believe this evaluation is important because many existing document-based and image forensic methods are implicitly assumed to provide robust protection against manipulated identity documents in the literature. Our benchmark demonstrates that this assumption does not hold when the manipulation resides at the semantic or logical level instead of the pixel level, and hence can lead to a "false sense of security". This diversity across architectures ensures that consistent failure across all models reflects a fundamental limitation of the detection paradigm rather than any individual model's shortcomings. We further compare against optical character recognition (OCR)-based baselines. Each model is evaluated using the area under the curve (AUC) and equal error rate (EER) as primary metrics, measuring discriminative power and threshold-independent error balance, respectively.

4.2.1 Isolating Logical Manipulation

Before evaluating individual manipulation strategies, we first examine whether only logical manipulation can be detected when the rest of the identity document remains visually authentic. To isolate this effect, we compare *Real Face + No Logical Manipulation (clean IDs)* against *Real Face + Logical Manipulation*, while keeping the facial content and overall document appearance con-

Table 2: AUC (%) and EER (%) for distinguishing Real Face + No Logical Manipulation from Real Face + Logical Manipulation on FactID.

Category	Model	AUC (%)	EER (%)
Image-based	DoU (CVPR'26)	50.79	49.01
	LTD (CVPR'26)	49.11	50.97
	DeepID (ICCV'25)	50.33	49.74
	Effort (ICML'25)	50.87	49.27
	SAFIRE (AAAI'25)	49.95	50.04
	ProDet (NeurIPS'24)	49.19	50.59
	FreqNet (AAAI'24)	52.55	48.45
	MMFusion (Springer'24)	49.92	50.02
	TruFor (CVPR'23)	49.94	49.97
	IML-ViT (arXiv'23)	49.54	50.30
Doc-based	ADCD-Net (ICCV'25)	49.93	50.12
	ASCFormer (PR'25)	49.47	50.33
	DocTamper (CVPR'23)	48.78	50.94

sistent. This provides a direct assessment of whether existing forensic models can distinguish a logically manipulated identity document in the absence of conventional visual forgery cues. Table 2 shows that all evaluated models remain close to random performance in this controlled setting. The AUC ranges from only 48.78% to 52.55%, while the EER remains close to 50% across both image-based and document-based models. The highest AUC, achieved by FreqNet, is only 52.55%, while DocTamper obtains the lowest at 48.78%. This consistent near-random behavior indicates that the logical inconsistencies introduced in FactID do not provide the visual or spatial evidence typically exploited by these models. Instead, the manipulated documents remain visually plausible, with the distinguishing evidence residing in the semantic relationships between their fields. This establishes the central challenge of FactID: logical manipulation can remain undetectable to existing visual and document-forensics models even when the document is provably inconsistent. Having established this baseline difficulty, we next examine how model performance changes under different FactID evaluation strategies, including Full ID, Face Only, Template Only, and Content Only, and under individual manipulation settings.

4.2.2 Image Forensics Models: Evaluation Across Strategies and Manipulations

Table 3 reveals a consistent pattern, no evaluated model, whether a deepfake detector (LTD, DoU, DeepID, Effort, FreqNet, ProDet) or a general image forensics model (TruFor, MMFusion, SAFIRE, IML-ViT), achieves reliable detection across all four strategies, confirming that FactID exposes fundamental blind spots in current forensic pipelines. While deepfake detectors directly output an image-level classification score, general image forensics models produce a pixel-wise segmentation mask, assigning each pixel a probability of tampering. We derive an image-level detection score by averaging the top-5% of predicted tampered-class probabilities across all spatial locations. For instance, LTD and DoU, are primarily trained to exploit visual artifacts associated with manipulated regions, rather than logical relationships between document fields. This limitation becomes evident across the evaluation strategies. LTD achieves 80.92% AUC on Full ID but drops sharply to 26.85% on Template Only, while DoU falls to 26.27% on Content Only. DeepID, despite being designed for identity-document forensics, achieves 77.32% on Full ID but only 19.11% on Content Only. Similar variations occur across the remaining models, including TruFor and MMFusion, two image-forensics models widely used in identity-document forgery research. Even when a model performs well on one strategy, such as Effort on Face Only (68.76%) or LTD on Content Only (73.14%), the performance does not consistently transfer to other strategies. This behavior is expected because the manipulated fields in FactID can remain visually plausible, leaving little or no localized evidence for pixel-centric detectors to exploit. Thus, their predictions can capture incidental visual cues without reliably identifying the underlying logical inconsistency.

Table 4 confirms this limitation at the manipulation level. Template-level manipulations show substantial inter-model variation: Jumble Words ranges from 0.00% AUC for IML-ViT to 100.00% for Effort, ProDet, and MMFusion, while GI Substitution ranges from 30.93% for SAFIRE to 84.92% for Effort and TruFor. Such extreme variation indicates that individual high scores are not evidence

Table 3: AUC (%) and EER (%) of Image Forensics Models on FactID across four evaluation strategies. Grey highlights best AUC per row.

	AUC (%) ↑				EER (%) ↓			
Model	**Full ID**	**Face Only**	**Template Only**	**Content Only**	**Full ID**	**Face Only**	**Template Only**	**Content Only**
LTD (CVPR'26)	80.92	33.09	26.85	73.14	25.47	62.70	65.69	33.84
DoU (CVPR'26)	39.50	35.87	69.37	26.27	57.59	59.01	36.01	67.66
DeepID (ICCV'25)	77.32	30.81	41.43	19.11	28.16	62.48	48.19	72.27
Effort (ICML'25)	22.36	68.76	49.90	66.40	70.67	41.14	48.08	33.60
SAFIRE (AAAI'25)	58.62	39.57	31.78	46.97	45.99	57.21	46.47	50.34
ProDet (NeurIPS'24)	61.33	35.94	58.32	26.42	41.27	60.09	52.74	66.05
FreqNet (AAAI'24)	50.00	22.29	54.95	51.54	50.00	71.77	47.37	49.28
MMFusion (Springer'24)	33.80	39.85	69.57	43.67	60.75	57.98	33.24	52.05
TruFor (CVPR'23)	52.69	32.09	49.64	16.72	47.03	63.80	50.36	70.54
IML-ViT (arXiv'23)	45.93	39.41	56.13	37.43	54.50	54.89	52.74	58.03

Table 4: Per-manipulation-type AUC (%) of Image Forensics Models on FactID.

	Template Only			Content Only		
Model	**Spelling Errors**	**Jumble Words**	**GI Substitution**	**Invalid Dates**	**Invalid ID**	**Gender Mismatch**
DoU (CVPR'26)	68.38	35.52	50.01	48.50	48.05	48.93
LTD (CVPR'26)	43.25	34.69	54.97	44.49	48.70	46.08
DeepID (ICCV'25)	48.49	3.27	35.40	30.95	31.09	29.38
Effort (ICML'25)	54.78	100.00	84.92	46.42	46.48	45.79
SAFIRE (AAAI'25)	45.84	45.70	30.93	55.99	54.60	55.31
ProDet (NeurIPS'24)	54.78	100.00	69.85	72.43	72.93	72.32
FreqNet (AAAI'24)	73.76	96.37	51.15	49.07	44.76	47.26
MMFusion (Springer'24)	74.42	100.00	65.89	44.31	42.46	45.87
TruFor (CVPR'23)	51.81	54.78	84.92	45.88	53.79	47.04
IML-ViT (arXiv'23)	73.76	0.00	34.10	72.51	77.09	72.31

of a consistent logical forgery capability. The same pattern appears for content-level manipulations, where IML-ViT reaches 72.31-77.09% AUC across Invalid Dates, Invalid ID, and Gender Mismatch, whereas TruFor remains near chance at 45.88-53.79%. These manipulations are defined by relationships among fields rather than visible pixel changes, which explains why their detection varies substantially across architectures. Overall, these results demonstrate that existing visual forensic models can respond to individual logic-level manipulations, but their performance is highly manipulation- and strategy-dependent, revealing a significant challenge for reliable detection of visually subtle logical forgeries. Per-manipulation EER results are provided in Appendix A, Table 7.

Table 5: AUC (%) and EER (%) of Document-Trained Models on FactID across four evaluation strategies. Grey highlights the best AUC per row.

	AUC (%) ↑				EER (%) ↓			
Model	**Full ID**	**Face Only**	**Template Only**	**Content Only**	**Full ID**	**Face Only**	**Template Only**	**Content Only**
ADCD-Net (ICCV'25)	76.30	57.62	49.87	84.32	26.40	43.32	53.26	27.80
ASCFormer (PR'25)	26.29	72.83	19.63	73.71	68.07	30.33	73.67	33.78
DocTamper (CVPR'23)	34.15	39.35	38.39	17.69	60.26	55.98	45.57	73.52

Table 6: Per-manipulation-type AUC (%) of Document-Trained Models on FactID.

	Template Only			Content Only		
Model	**Spelling Errors**	**Jumble Words**	**GI Substitution**	**Invalid Dates**	**Invalid ID**	**Gender Mismatch**
ADCD-Net (ICCV'25)	77.39	100.00	69.85	52.17	58.69	57.21
ASCFormer (PR'25)	3.62	100.00	50.82	46.69	53.83	46.46
DocTamper (CVPR'23)	1.53	0.07	0.95	43.41	42.18	44.20

4.2.3 Evaluation of Document-Trained Models Across Strategies and Manipulations

Since document-trained models produce pixel-wise segmentation masks, we derive an image-level score by taking the maximum predicted probability of the tampered class across all spatial locations,

which is then used to compute AUC and EER. Table 5 shows that document-trained models also exhibit substantial variability across the four evaluation strategies, with no model maintaining consistently strong performance across all settings. ADCD-Net achieves its highest AUC of 84.32% on Content Only, while DocTamper reduces it to 17.69%. For Template Only, performance varies substantially across models, with ADCD-Net achieving 49.87% AUC, while ASCFormer and DocTamper obtain only 19.63% and 38.39%, respectively. This variation indicates that document-specific training does not guarantee robust detection of the logical manipulations in FactID. Although some models achieve relatively high scores on particular strategies, their performance changes substantially with the exposed document region. The EER results further reflect this inconsistency, with several high-AUC cases accompanied by substantial error rates. Thus, even document-trained architectures remain sensitive to the visual characteristics of the input rather than consistently identifying the underlying logical inconsistency. Table 6 further decomposes performance across individual manipulation types. Template-level manipulations exhibit extreme inter-model variation, while ADCD-Net achieves 77.39% AUC on Spelling Errors and 100.00% on Jumble Words, whereas DocTamper achieves only 1.53% and 0.07%, respectively. For Content Only manipulations, ADCD-Net obtains 52.17%, 58.69%, and 57.21% AUC for Invalid Dates, Invalid ID, and Gender Mismatch, respectively. The inconsistent performance across manipulation types shows that even document-specific models do not reliably capture the underlying logical relationships. Overall, these results demonstrate that even models explicitly trained for document forgery detection are challenged by FactID; their performance is strongly dependent on the manipulation type and document region, revealing the limitations of relying primarily on pixel-level forensic evidence when the forgery is encoded through logical or semantic inconsistencies. Per-manipulation EER results are provided in Appendix A, Table 8.

4.3 OCR BASELINE COMPARISON

To assess whether explicit text reading can better capture logical forgeries, we introduce a training-free OCR baseline that extracts document fields and applies deterministic validation rules. As shown in Figure 3, the OCR baseline achieves its strongest performance on Template Only, with an AUC of 81.64% and EER of 25.83%, indicating that reading the document content can effectively expose template-level inconsistencies. However, performance drops substantially on Full ID (AUC 66.20%) and Content Only (AUC 30.28%), with the latter showing that fixed validation rules remain insufficient for more complex content-level logical manipulations. These results suggest that while text-based validation provides a more direct signal for logical forgeries than purely visual analysis, it remains incomplete for detecting the broader spectrum of FactID manipulations.

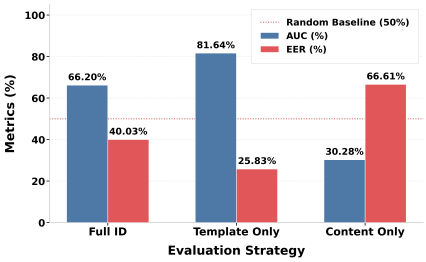

Figure 3: AUC (%) and EER (%) of the OCR baseline across three evaluation strategies.

5 CONCLUSION

We introduce FactID, a benchmark targeting logical and semantic identity document forgeries that are visually indistinguishable from authentic documents and thus fundamentally beyond the reach of current forensic models. Our evaluation of multiple pretrained models, spanning visual manipulation detectors, document-trained architectures, and an OCR-based baseline, shows that no existing paradigm achieves reliable detection across all forgery types, with an average AUC lower than FantasyID, confirming a qualitatively harder challenge (section: A.7 in the Appendix). These findings expose a critical gap: the absence of models capable of semantic and logical document understanding. Future work should explore hybrid architectures combining visual forensics with structured text validation, and vision-language models capable of cross-field consistency reasoning. The most dangerous forgeries are not the ones that look manipulated; they are the ones that look perfectly real. Closing this gap demands detectors that do not merely see documents, but truly read and understand them.

AI USE STATEMENT

We used generative AI tools solely to assist with language editing, grammar correction, and improving the clarity and readability of the manuscript. We did not use generative AI tools to generate experimental results, fabricate data, design or modify experimental outcomes, or produce the final scientific claims of this work. All AI-assisted text was reviewed and edited by the authors for accuracy, consistency, and alignment with the reported experiments. The authors take full responsibility for the final content of this work, including all claims, analyses, results, and artifacts.

ETHICS STATEMENT

This work studies logical manipulation of identity documents for the purpose of understanding and evaluating emerging document-forgery threats. All identity cards in the dataset are constructed solely for research purposes, with personal information independently assigned to each card; no real personal information belonging to any individual is used at any stage of dataset construction. The manipulated documents are generated from faithful document templates and are intended only for controlled research and evaluation. We recognize that identity document manipulation can have potential security and privacy implications; accordingly, the work is framed as a defensive security study aimed at identifying limitations of existing forensic systems and improving future detection capabilities. Upon acceptance, we will release the dataset under an appropriate open-source license to facilitate responsible research use.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the FactID dataset construction, logical and visual manipulation types, evaluation protocols, and experimental settings in the main paper and the Appendix. The paper specifies the evaluation strategies, manipulation categories, baseline models, scoring procedures, and metrics used for comparison. Additional implementation details, preprocessing procedures, and experimental configurations are provided in the Appendix to facilitate reproduction of the reported results. All reported results are obtained using the same predefined evaluation protocol across the compared models.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Imtiaz Ahmed, Sadman Islam, Partha Protim Datta, Imran Kabir, Md Naseef Ur Rahman Chowdhury, and Ahshanul Haque. Qwen 2.5: A comprehensive review of the leading resource-efficient llm with potentioal to surpass all competitors. 2025.

Vladimir Arlazarov, Konstantin Bulatov, Timofey Chernov, and V.L. Arlazarov. Midv-500: a dataset for identity document analysis and recognition on mobile devices in video stream. *Computer Optics*, 43:818–824, 10 2019. doi: 10.18287/2412-6179-2019-43-5-818-824.

Konstantin Bulatov, E Emelianova, Daniil Tropin, Natalya Skoryukina, Yulia Chernyshova, Alexander Sheshkus, S Usilin, Zuheng Ming, Jean-Christophe Burie, Muzzamil Luqman, and Vladimir Arlazarov. Midv-2020: a comprehensive benchmark dataset for identity document analysis. *Computer Optics*, 46:252–270, 03 2022. doi: 10.18287/2412-6179-CO-1006.

Manuel G Calvo and Daniel Lundqvist. Facial expressions of emotion (kdef): Identification under different display-duration conditions. *Behavior research methods*, 40(1):109–115, 2008.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24185–24198, 2024.

Jikang Cheng, Zhiyuan Yan, Ying Zhang, Yuhao Luo, Zhongyuan Wang, and Chen Li. Can we leave deepfake data behind in training deepfake detector? *Advances in Neural Information Processing Systems*, 37:21979–21998, 2024.

Naser Damer, Alexandra Mosegui Saladie, Andreas Braun, and Arjan Kuijper. Morgan: Recognition vulnerability and attack detectability of face morphing attacks created by generative adversarial network. In *2018 IEEE 9th international conference on biometrics theory, applications and systems (BTAS)*, pp. 1–10. IEEE, 2018.

Gourab Das, Raghavendra Ramachandra, et al. From forgeries to foundation models: A systematic survey of identity document attack and detection. *arXiv preprint arXiv:2607.01442*, 2026.

Lisa DeBruine and Benedict Jones. Face Research Lab London Set. 5 2017. doi: 10.6084/m9.figshare.5047666.v4. URL https://figshare.com/articles/dataset/Face_Research_Lab_London_Set/5047666.

Abhishek Dhore, Pragati Dhore, Priti Gangurde, Abhijit Khadke, Sonam Singh, and Vishal Borate. Face morphing attack detection using deep learning. In *2025 International Conference on Computing Technologies & Data Communication (ICCTDC)*, pp. 01–06. IEEE, 2025.

Jing Dong, Wei Wang, and Tieniu Tan. CASIA image tampering detection evaluation database. In *2013 IEEE China Summit and International Conference on Signal and Information Processing*. IEEE, July 2013. doi: 10.1109/chinasip.2013.6625374. URL https://doi.org/10.1109/chinasip.2013.6625374.

Matteo Ferrara, Annalisa Franco, and Davide Maltoni. The magic passport. In *IEEE international joint conference on biometrics*, pp. 1–7. IEEE, 2014.

Hong Guan, Yancheng Wang, Lulu Xie, Soham Nag, Rajeev Goel, Niranjan Erappa Narayana Swamy, Yingzhen Yang, Chaowei Xiao, Jonathan Prisby, Ross Maciejewski, et al. Idnet: A novel dataset for identity document analysis and fraud detection. *arXiv preprint arXiv:2408.01690*, 2024.

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 (CVPR)*, pp. 20606–20615, June 2023.

Qinghui He, Haifeng Zhang, Qiao Qin, Bo Liu, Xiuli Bi, and Bin Xiao. Diversity over uniformity: Rethinking representation in generated image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 40407–40417, June 2026.

Davis E King. Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research*, 10:1755–1758, 2009.

Pavel Korshunov, Amir Mohammadi, Christophe Ecabert, Sébastien Marcel, et al. Fantasyid: A dataset for detecting digital manipulations in id-documents. In *2025 IEEE International Joint Conference on Biometrics (IJCB)*, pp. 1–9. IEEE, 2025a.

Pavel Korshunov, Vidit Vidit, Amir Mohammadi, Christophe Ecabert, Nevena Shamoska, Sébastien Marcel, Zeqin Yu, Ye Tian, Jiangqun Ni, Lazar Lazarevic, Renat Khizbullin, Anastasiia Evteeva, Alexey Tochin, Aleksei Grishin, Anjith George, Daniel Dealcala, Tamas Endrei, Javier Muñoz Haro, Ruben Tolosana, Ruben Vera-Rodriguez, Aythami Morales, Julian Fierrez, György Cserey, Hardik Sharma, Sachin Chaudhary, Akshay Dudhane, Praful Hambarde, Amit Shukla, Prateek Shaily, Jayant Kumar, Ajinkya Hase, Satish Maurya, Mridul Sharma, and Pallav Dwivedi. Deepid challenge of detecting synthetic manipulations in id documents. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 521–530, October 2025b.

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.

Xiaohong Liu, Yaojie Liu, Jun Chen, and Xiaoming Liu. Pscc-net: Progressive spatio-channel correlation network for image manipulation detection and localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(11):7505–7517, 2022.

Dongliang Luo, Yuliang Liu, Rui Yang, Xianjin Liu, Jishen Zeng, Yu Zhou, and Xiang Bai. Toward real text manipulation detection: New dataset and new solution. *Pattern Recognition*, 157:110828, 2025.

Xiaochen Ma, Bo Du, Zhuohang Jiang, Ahmed Y Al Hammadi, and Jizhe Zhou. Iml-vit: Benchmarking image manipulation localization by vision transformer. *arXiv preprint arXiv:2307.14863*, 2023.

Aziz Makandar and Syeda Bibi Javeriya. Morphed faces in aadhar cards: a resource for understanding biometric security vulnerabilities. *Mendeley Data*, 1, 2024.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 2200–2209, 2021.

Meta. Llama 3.2 model card. `https://github.com/meta-llama/llama-models/blob/main/models/llama3_2/MODEL_CARD.md`, 2024.

P Jonathon Phillips, Amy N Yates, Ying Hu, Carina A Hahn, Eilidh Noyes, Kelsey Jackson, Jacqueline G Cavazos, Géraldine Jeckeln, Rajeev Ranjan, Swami Sankaranarayanan, et al. Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *Proceedings of the National Academy of Sciences*, 115(24):6171–6176, 2018.

Alin C. Popescu and Hany Farid. Statistical tools for digital forensics. In Jessica Fridrich (ed.), *Information Hiding*, pp. 128–147, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg. ISBN 978-3-540-30114-1.

Chenfan Qu, Chongyu Liu, Yuliang Liu, Xinhong Chen, Dezhi Peng, Fengjun Guo, and Lianwen Jin. Towards robust tampered text detection in document image: New dataset and new solution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5937–5946, 2023.

David J Robertson, Robin SS Kramer, and A Mike Burton. Fraudulent id using face morphs: Experiments on human and automatic recognition. *PLoS One*, 12(3):e0173319, 2017.

Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1–11, 2019.

Ziqi Sheng, Junyan Wu, Wei Lu, and Jiantao Zhou. Weakly-supervised image forgery localization via vision-language collaborative reasoning framework. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 988–996, 2026.

Qi Song, Ziyuan Luo, and Renjie Wan. Creating blank canvas against ai-enabled image forgery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 9088–9096, 2026.

Gargi Surendra Yeole, Poulomi Bhattacharya, and Akshay Agarwal. Guarding digital identity: Attention-guided fusion for detecting forged id documents (student abstract). *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(48):41400–41402, Mar. 2026. doi: 10.1609/aaai.v40i48.42285. URL `https://ojs.aaai.org/index.php/AAAI/article/view/42285`.

Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Frequency-aware deepfake detection: Improving generalizability through frequency space domain learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 5052–5060, 2024.

Chuangchuang Tan, Xiang Ming, Jinglu Wang, Renshuai Tao, Bin Li, Yunchao Wei, Yao Zhao, and Yan Lu. Semantic visual anomaly detection and reasoning in ai-generated images, 2025. URL `https://arxiv.org/abs/2510.10231`.

Juan E Tapia, Naser Damer, Christoph Busch, Juan M Espin, Javier Barrachina, Alvaro S Rocamora, Krištof Ocvirk, Leon Alessio, Borut Batagelj, Sushrut Patwardhan, et al. First competition on presentation attack detection on id card. In *2024 IEEE International Joint Conference on Biometrics (IJCB)*, pp. 1–10. IEEE, 2024.

Konstantinos Triaridis and Vasileios Mezaris. Exploring multi-modal fusion for image manipulation detection and localization. In *International conference on multimedia modeling*, pp. 198–211. Springer, 2024.

David White, Richard I Kemp, Rob Jenkins, Michael Matheson, and A Mike Burton. Passport officers' errors in face matching. *PloS one*, 9(8):e103510, 2014.

Kahim Wong, Jicheng Zhou, Haiwei Wu, Yain-Whar Si, and Jiantao Zhou. Adcd-net: Robust document image forgery localization via adaptive dct feature and hierarchical content disentanglement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19280–19289, 2025.

Yue Wu, Wael AbdAlmageed, and Premkumar Natarajan. Mantra-net: Manipulation tracing network for detection and localization of image forgeries with anomalous features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9543–9552, 2019.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. Layoutlm: Pretraining of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 1192–1200, 2020.

Zhiyuan Yan, Jiangming Wang, Peng Jin, Ke-Yue Zhang, Chengchun Liu, Shen Chen, Taiping Yao, Shouhong Ding, Baoyuan Wu, and Li Yuan. Orthogonal subspace decomposition for generalizable ai-generated image detection. In *International Conference on Machine Learning*, pp. 70268–70288. PMLR, 2025.

Yawen Yang, Feng Li, Shuqi Kong, Yunfeng Diao, Xinjian Gao, Zenglin Shi, and Meng Wang. Layer consistency matters: Elegant latent transition discrepancy for generalizable synthetic image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 38111–38121, June 2026.

Qingwen Zeng, Juan E Tapia, Sneha Das, and Christoph Busch. From vision to text: A compact multimodal approach for robust, cross-domain presentation attack detection on id cards. *arXiv preprint arXiv:2606.06966*, 2026.

Song Yang Zhang, Zhifei and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017.

Peng Zhou, Xintong Han, Vlad I Morariu, and Larry S Davis. Learning rich features for image manipulation detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1053–1061, 2018.

Anni Zou, Wenhao Yu, Hongming Zhang, Kaixin Ma, Deng Cai, Zhuosheng Zhang, Hai Zhao, and Dong Yu. DocBench: A benchmark for evaluating LLM-based document reading systems. In Weijia Shi, Wenhao Yu, Akari Asai, Meng Jiang, Greg Durrett, Hannaneh Hajishirzi, and Luke Zettlemoyer (eds.), *Proceedings of the 4th International Workshop on Knowledge-Augmented Methods for Natural Language Processing*, pp. 359–373, Albuquerque, New Mexico, USA, May 2025. Association for Computational Linguistics. ISBN 979-8-89176-229-9. doi: 10.18653/v1/2025.knowledgenlp-1.29. URL https://aclanthology.org/2025.knowledgenlp-1.29/.

A Appendix

A.1 Broader Impact

FactID addresses a critical and underexplored vulnerability in national identity infrastructure, with direct relevance to fraud prevention across banking, healthcare, and welfare systems serving over a

billion citizens; by establishing the first benchmark for logical document manipulation, this work can directly inform more robust identity verification pipelines and reduce large-scale impersonation fraud. On the negative side, systematically cataloging logical manipulation types could serve as a reference for adversaries seeking to craft more targeted forgeries, potentially enabling disinformation or illegal document operations; automated systems trained on this benchmark may also exhibit fairness disparities across demographic groups, risking unfair rejection of legitimate documents for specific populations. To mitigate these risks, all samples in FactID are fully generated for research purposes, with no real personal data involved, and the dataset will be released under a restricted academic license with usage agreements to prevent misuse.

A.2 Overall Generation Pipeline

Figure 4 below presents the overall pipeline followed in the evaluation and benchmarking of FactID dataset. All manipulations are programmatically generated using a controlled Python-based pipeline starting from the original blank UIDAI template. Genuine Aadhaar card templates are rendered with independently assigned demographic information, with no real individual's personal data used, using the `Pillow` imaging library for precise text placement. For template-level manipulations, target strings are identified via predefined field coordinates and replaced with corrupted variants, misspelled strings, permuted word sequences, or substituted contact values, at render time. For content-level manipulations, field values are algorithmically generated to violate specific logical constraints: invalid dates are produced by sampling chronologically impossible combinations, invalid Aadhaar numbers by adding one extra digit and deliberately corrupting the Verhoeff checksum, and gender mismatches by assigning a label inconsistent with the cardholder profile. Each generated sample is assigned a single manipulation label. Face images are composited into the photograph region using a standardized bounding box coordinates, with real, morphed, and synthetic faces assigned to their respective sample categories as described in Section 3.1.

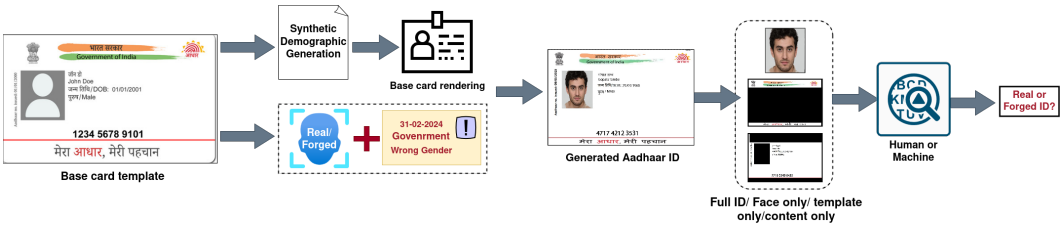

Figure 4: The figure illustrates the overall generation and evaluation pipeline.

A.3 Implementation Details

Document-Trained Models: Document-trained forgery detection models, including DocTamper, ASCFormer, and ADCD-Net are designed as pixel-wise segmentation networks that jointly localize and classify tampered regions within a document image. Rather than producing a single image-level decision, these models output a dense spatial heatmap of the same resolution as the input, where each pixel is assigned a probability of belonging to either the authentic or tampered class. Prior to inference, all input images are resized to 512×512 pixels using bicubic interpolation to preserve fine-grained textual and structural details, and normalized using ImageNet mean $\mu = (0.485, 0.456, 0.406)$ and standard deviation $\sigma = (0.229, 0.224, 0.225)$. To derive a single scalar image-level detection score from the predicted heatmap, we extract the maximum tampered-class probability across all spatial locations of the predicted mask. Formally, given the predicted probability map $\mathbf{P} \in \mathbb{R}^{H \times W}$ for the tampered class, the image-level score is computed as:

$$s = \max_{(i,j)} \mathbf{P}_{ij} \cdot \mathbf{1}[\mathbf{P}_{ij} > \tau]$$

where $\tau = 0.3$ is a confidence threshold applied to suppress low-confidence and noisy pixel predictions before aggregation. This design choice is motivated by the observation that even a small localized tampered region, such as a modified face photograph or altered demographic field, should be sufficient to flag the entire document as manipulated. A higher score indicates that the model

has detected at least one spatially coherent region of high-confidence manipulation within the document.

In addition to DCT-based frequency features required by DocTamper, images are temporarily re-encoded as JPEG at quality $q = 95$ to extract DCT coefficient arrays and quantization tables, which serve as auxiliary inputs to the model alongside the RGB image tensor. All models are evaluated using their officially released pretrained weights without any fine-tuning or adaptation to our dataset.

Image-Trained Visual Forensics Models: Deepfake detector and General-purpose visual manipulation detectors, including DeepID, DoU, LTD, Effort, FreqNet, ProDet, TruFor, MMFusion, SAFIRE and IML-ViT variants are classification-based models originally trained on face forgery or natural image manipulation benchmarks such as FaceForensics++ (Rossler et al., 2019) and CASIA (Dong et al., 2013). Unlike the document-trained models above, the deepfake detectors (Effort, FreqNet, ProDet) are applied to the document image as a whole and produce a single image-level fake-class probability directly from a classification head, without generating any spatial localization map. But general image forensic models produce a pixel-wise segmentation mask, assigning each pixel a probability of being tampered with, we derive an image-level detection score by averaging the top-k% of predicted tampered-class probabilities across all spatial locations. Prior to inference, all input images are resized to 224×224 pixels and normalized using the pretraining statistics specific to each model. For CLIP-based models such as Effort, normalization follows the OpenAI CLIP protocol with mean $\mu = (0.48145, 0.45782, 0.40821)$ and standard deviation $\sigma = (0.26863, 0.26130, 0.27577)$, while all other models use standard ImageNet statistics. The image-level detection score is taken directly as the predicted fake-class probability $p_{\text{fake}} \in [0, 1]$ from the model's softmax output. An image is classified as fake if $p_{\text{fake}} > 0.5$, and as real otherwise. For evaluation under the *Face Only* strategy, face regions are first localized using a frontal face detector with a 20% spatial margin applied on all four sides to include surrounding context, and the resulting crop is passed to the model instead of the full document image. Continuous scores are used to compute AUC and EER over the full test set. All models are evaluated using their officially released pretrained weights without any fine-tuning or domain adaptation.

All experiments are conducted on a single NVIDIA H200 GPU, using PyTorch with CUDA 12.8. All models are evaluated in full-precision (FP32) using their officially released pretrained weights without any fine-tuning or adaptation to our dataset. For image forensics and deepfake detection models, input images are resized to 224×224 and normalized using model-specific pretraining statistics; for document-trained segmentation models (DocTamper, ASCFormer, ADCD-Net), inputs are resized to 512×512 using standard ImageNet normalization. Inference is performed with a batch size of 16 and 4 parallel data-loading workers. Face regions are extracted using dlib's frontal face detector with a 20% spatial margin, and template/content masking is applied via coordinate-based rectangular blackout using OpenCV. The total evaluation across all 11 models and 4 strategies peaks at a memory consumption ranging from 6 GB for lightweight models to 24 GB for CLIP ViT-L/14-based architectures such as Effort, while Qwen 2.5 and Llama3.2-11B model consume a maximum memory of 75GB.

A.4 Visualization of All Types of Manipulation and masked inputs (in template and content manipulation)

Figure 5 illustrates the full scope of manipulations covered in FactID. Three reference cards are shown first: a clean, authentic ID, an ID with a morphed face, and an ID with a synthetically generated face. The remaining samples correspond to a single person's ID subjected to each of the 11 manipulation types spanning template-level and content-level categories, as defined in Section 3.2. Also, Figure 6 illustrates the masked input variants used for strategy-specific evaluation. The content-masked input (left) blacks out the central content region, retaining only structural template elements for Template Only evaluation. The template-masked input (right) blacks out the header, footer, QR code, and face region, exposing only the personally identifying content fields for Content Only evaluation.

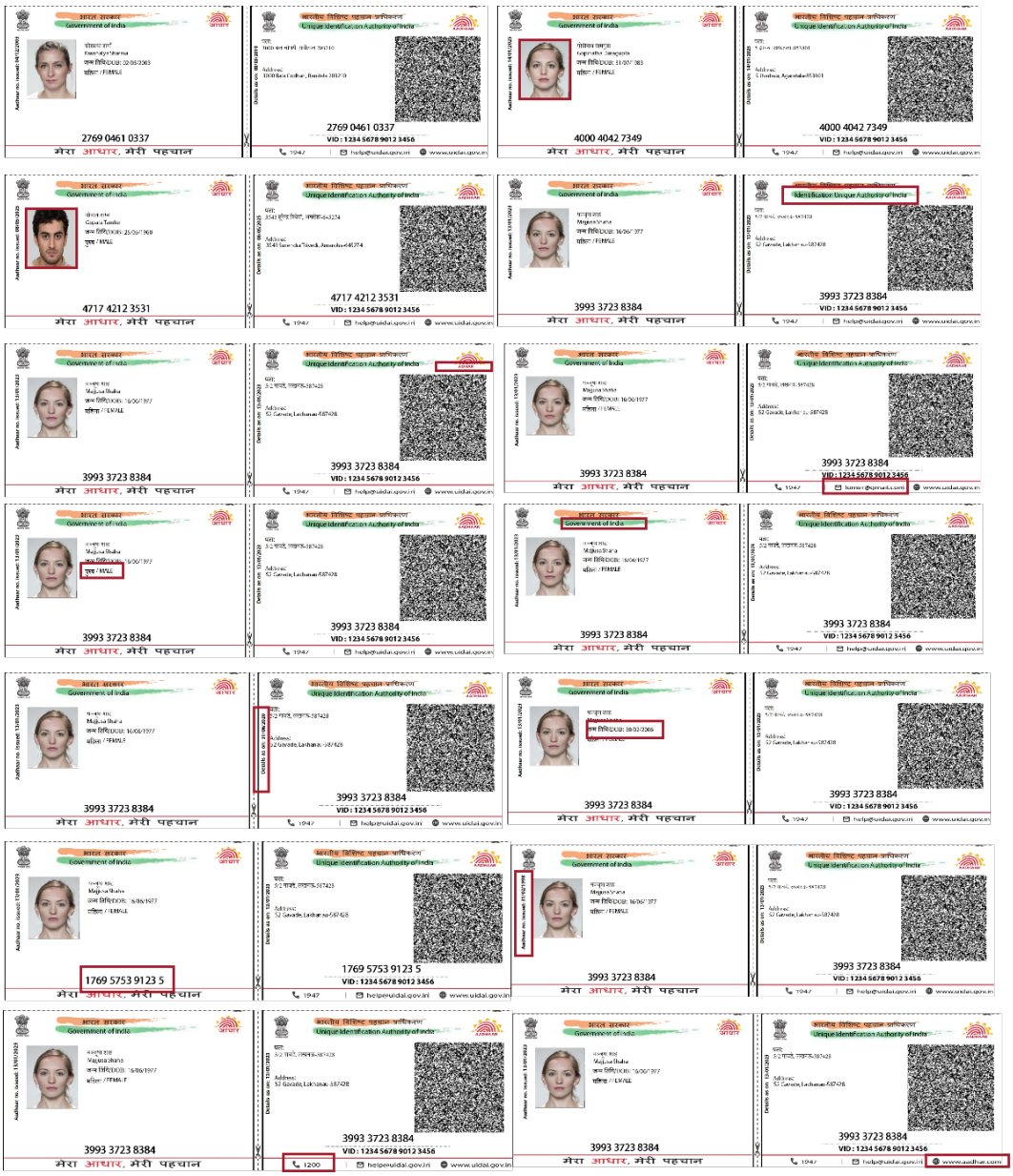

Figure 5: Illustration of all 11 manipulation types applied across template-level and content-level categories in FactID. red boxes highlighting the manipulated regions. The first image shows a clean Aadhaar card (without manipulation) with a real face as a reference.



Figure 6: Figure showing the masked input used in strategy-specific evaluation.

Table 7: Per-manipulation-type **EER (%)** of Image Forensics Models on **FactID**. Values near 50% indicate random-chance performance.

	Template Only			Content Only		
Model	**Spelling Err.**	**Jumble Words**	**GI Sub.**	**Invalid Date**	**Invalid ID**	**Gender Mismatch**
DoU (CVPR'26)	50.94	59.60	49.54	51.49	51.23	51.64
LTD (CVPR'26)	57.71	62.35	47.32	54.42	51.57	52.34
DeepID (ICCV'25)	51.50	96.72	66.66	66.92	67.06	67.13
Effort (ICML'25)	45.21	0.00	33.33	51.93	53.17	52.01
SAFIRE (AAAI'25)	54.01	54.29	54.29	43.64	44.31	44.10
FreqNet (AAAI'24)	48.84	3.62	48.84	51.43	55.26	52.93
ProDet (NeurIPS'24)	45.21	0.00	45.21	35.65	35.86	35.86
MMFusion (Springer'24)	49.99	0.00	51.15	57.45	57.40	57.25
TruFor (CVPR'23)	50.00	45.21	33.33	51.01	46.14	50.15
IML-ViT (arXiv'23)	48.84	99.99	48.84	33.72	30.77	33.56

A.5 Per Manipulation EER Analysis

A.5.1 Visual Manipulation Detectors

The EER results in Table 7 further support the instability observed in Table 4. While *Spelling Errors* remains relatively close to chance for most models, *Jumble Words* shows extreme variation, with EERs ranging from 0.00% to 99.99%. Similarly, *GI Substitution* exhibits substantial variation (33.33%-66.66%), while *Invalid Date*, *Invalid ID*, and *Gender Mismatch* span broad ranges across models. Overall, the inconsistent EER patterns across manipulation types indicate that these models do not reliably capture the semantic relationships underlying logical manipulations in FactID.

Table 8: Per-manipulation-type **EER (%)** of Document-Trained Models on **FactID**. Values near 50% indicate random-chance performance.

	Template Only			Content Only		
Model	**Spelling Err.**	**Jumble Words**	**GI Sub.**	**Invalid Date**	**Invalid ID**	**Gender Mismatch**
ADCD-Net (ICCV'25)	45.21	0.00	45.21	49.54	44.66	46.19
ASCFormer (PR'25)	96.37	0.00	51.15	51.90	46.60	51.69
DocTamper (CVPR'23)	98.46	99.93	98.95	55.12	55.47	54.62

A.5.2 Document-Trained Models

The EER results in Table 8 reinforce the AUC findings with sharper resolution at the decision-boundary level. Spelling Errors and Jumble Words yield EER values near 0.00 or 1.00 for ADCD-Net, confirming that its high AUC reflects a fully collapsed, trivially biased threshold rather than any learned discriminative behavior. ASCFormer and DocTamper similarly produce extreme EER values on at least one textual manipulation type, further confirming that no model has found a consistent operating point on this manipulation type. Content-level manipulations, Invalid Date, Invalid ID, and Gender Mismatch cluster tightly around EER $\approx$44.66-55.47 across all models, demonstrating that even at their optimal threshold, document-trained detectors cannot distinguish semantically forged documents from authentic ones at better than near-chance rates.

A.6 Where Do Models Actually Look?

Figure 7 presents GradCAM activation maps for all evaluated models on the jumbled word type template-level manipulations (*"Unique Authority of India"* $\rightarrow$ *"Authority Unique to India"*). At first glance, the AUC values of 100.00 reported for some of these manipulation types in Tables 4 and 6 might suggest that models have perfectly learned to detect these forgeries. But the activation maps tell a very different story. Not a single model attends to the manipulated text region. Instead, activations concentrate on forensically irrelevant areas, the header bar, the footer strip, regions that carry no manipulation signal whatsoever. The models are not detecting the forgery; they are reacting to low-level textural patterns inherited from their pretraining distribution. *An AUC of 100.00 here is not a sign of success, it is a sign of collapse: the model has assigned uniformly high or uniformly*

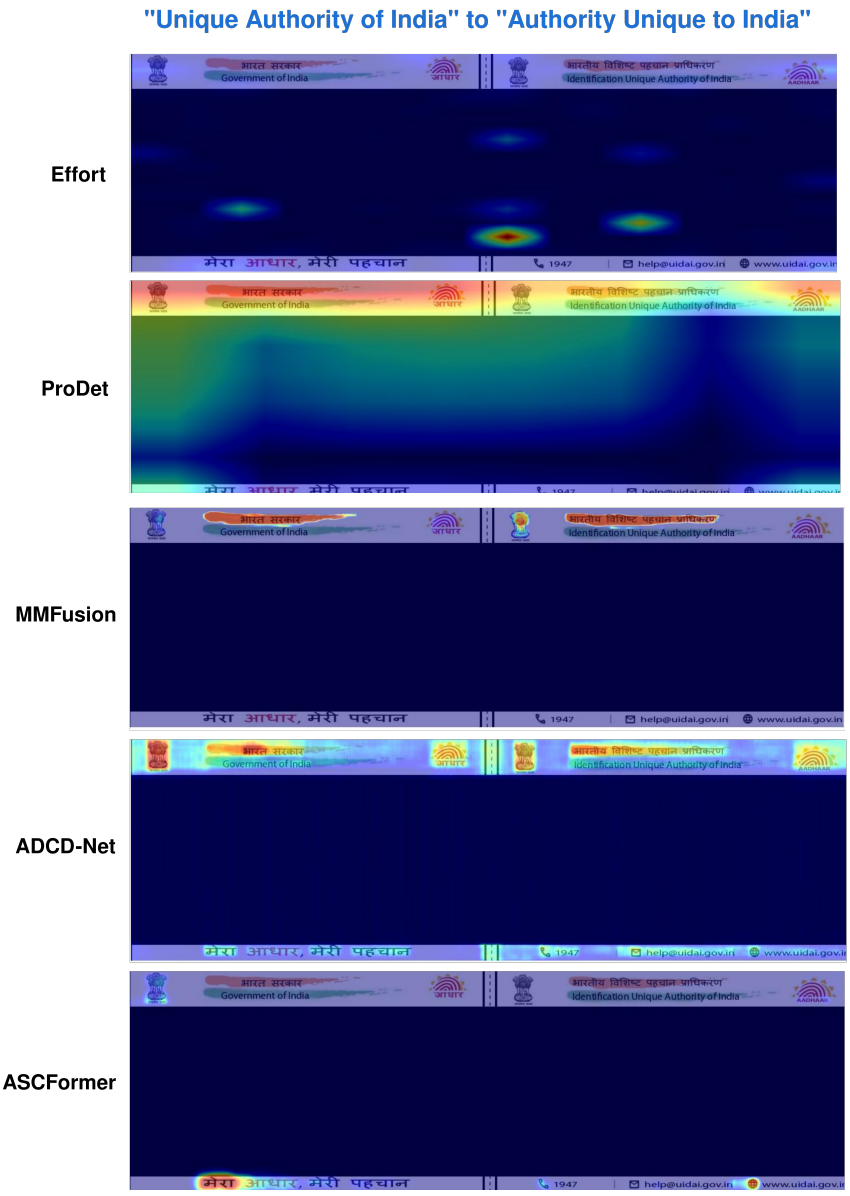

Figure 7: Where do models look on template-level forgeries? GradCAM activations across five models reveal attention concentrated on the irrelevant structural regions, not the manipulated text, exposing that AUC of 100.00 reflects decision collapse, not detection success.

low scores to every sample in the class, not because it found evidence of manipulation, but because its pretrained bias happened to align with the class label distribution by chance.

A.7 Difficulty in Terms of FantasyID Dataset

To further contextualize the difficulty of FactID, we compare the same pretrained models (Section 4.2) with FantasyID (Korshunov et al., 2025a), a closely related identity-document benchmark, under a comparable binary bona-fide-vs-attacked protocol. As shown in Table 9, the average performance is close across the two datasets, with FactID achieving 52.41% AUC and 48.05% EER, compared with 54.31% AUC and 46.62% EER on FantasyID. Performance nevertheless varies substantially across models: DoU, Effort, TruFor, MMFusion, ASCFormer, and DocTamper perform better on FantasyID. Notably, ASCFormer shows a large improvement on FantasyID (26.29% to

Table 9: Comparison of different models on FactID (ours) and FantasyID datasets using AUC and EER metrics. Higher AUC is better, lower EER is better. Blue values indicate cases where FantasyID outperforms FactID.

Category	Models	AUC		EER	
		FactID (ours)	FantasyID	FactID (ours)	FantasyID
Image-based	DeepID	77.32	72.87	28.16	33.99
	DoU	39.50	59.66	57.59	44.44
	LTD	80.92	43.55	25.47	55.56
	Effort	22.36	51.93	70.67	49.02
	FreqNet	50.00	44.24	50.00	52.29
	ProDet	61.33	48.01	41.27	49.35
	TruFor	52.69	58.01	47.03	43.79
	MMFusion	33.80	59.49	60.75	39.87
	SAFIRE	58.62	52.86	45.99	47.06
	IML-ViT	45.93	41.42	54.50	56.86
Doc-based	ADCD-Net	76.30	69.38	26.40	37.58
	ASCFormer	26.29	66.63	68.07	36.60
	DocTamper	34.15	54.37	60.26	45.75
Average	–	52.41	54.31	48.05	46.62

66.63% AUC). Although individual models show mixed behavior across the two datasets, the consistently lower aggregate AUC on FactID suggests that its logical manipulations provide less exploitable evidence than the attacks represented in FantasyID, making them harder to detect reliably.