DETECT THE FORGERY, NOT THE VISUAL DISTURBANCES: MLLM-GUIDED REPRESENTATION AND SELECTIVE MODULATION FOR ID DOCUMENT IMAGES

Anonymous authors
Paper under double-blind review

ABSTRACT

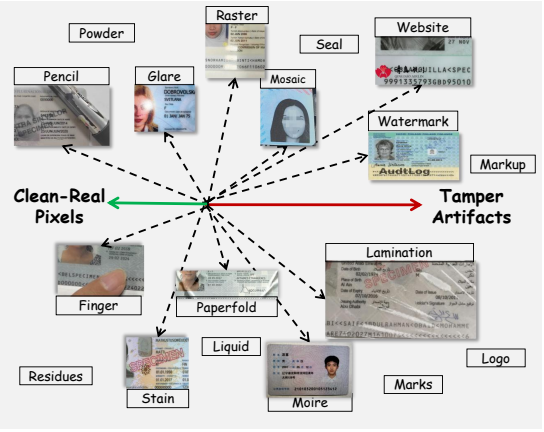
ID document image tampering detection and localization is an important problem in multimedia security, where detection determines whether an image has been tampered with, while localization identifies the tampered regions. Existing methods primarily focus on clean ID document image conditions. However, Visual Disturbances (VD), such as glare, mosaics, or stains, may occur in ID document images; their low-level features are highly similar to tampered artifacts, potentially leading to VD Induced Forensic Misattribution (VDFM). To address this problem, we propose CLEAR-ID, a new approach for ID document VD scenarios that integrates forensic experts with Multimodal Large Language Model (MLLM). CLEAR-ID uses prior information introduced by an MLLM to adaptively represent forensic expert evidence as VD-oriented or tampered-oriented cues, and projects the resulting features into a low-rank residual space to constrain the modulation of forensic expert evidence. To fill the data gap in this scenario, we further introduce DISTURBID-180k. It covers 42 types of VD and provides corresponding severity and localization mask annotations, enabling systematic analysis of the impact of VD on tampered detection and localization. Experiments demonstrate that VD-aware evidence representation and constrained VD-guided modulation effectively mitigate VDFM. Across diverse VD, CLEAR-ID improves both image-level detection and pixel-level localization over the state-of-the-art (SOTA) forensic expert baseline.

1 INTRODUCTION

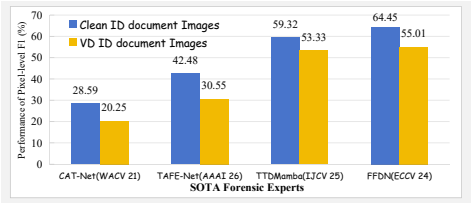
Generative image editing and modern editing tools have made sophisticated manipulation of ID document images increasingly accessible. Existing forensic methods are primarily designed to identify manipulation traces in relatively clean ID document images, while current datasets similarly prioritize pristine data conditions. However, submitted ID document images in real-world are rarely free from unrelated visual anomalies. Acquisition, recapture, copyright protection can introduce visible artifacts such as stains or watermarks. We refer to these non-malicious artifacts as ***ID Document Image Visual Disturbances*** (**ID-VD**) as illustrated by the examples in Figure 1(a). Consequently, visual disturbances and malicious tampering may coexist in the same forensic input, requiring detectors to identify not only anomalous evidence but also its provenance.

In ID document image tampering detection, methods such as FFDN (Chen et al., 2024) and TAFE-net (Yi et al., 2026) are designed to identify statistical and frequency-domain anomalies as tampering evidence. However, VD can trigger low-level and mid-level forensic features similar to those induced by tampering detection, making their sources difficult to distinguish in the low-dimensional feature space and leading to the performance degradation shown in Figure 1(b). A closely related desensitized document tampering detection method, DCL-Net (Li et al., 2026), focuses on mosaic, a single type of visual disturbance and lacks generalization to diverse non-malicious artifacts. The nature of VD is inherently open-set, with new types continually emerging and inducing varying responses. Moreover, suppressing VD-induced false detections may also suppress subtle tampering signals, trading one type of error for another. We refer to such abnormal decisions induced by VD themselves as ***Visual Disturbances Induced Forensic Misattribution*** (**VDFM**). Figure 1(c) shows VDFM and its generalization challenges.

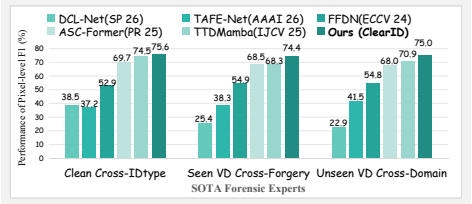
Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping 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.



(a) Similarity between Visual Disturbance Artifacts and Forensics Artifacts



(b) Trends in SOTA Experts Performance on Clean/VD Images



(c) Comparison of SOTA VD-Aware Experts Performance

Figure 1: **Visual Disturbances Induce Forensic Misattribution in ID Document Images.** (a) Visual Disturbance artifacts exhibit visual similarities to forensic artifacts, potentially causing forensic misattribution. (b) Existing state-of-the-art(SOTA) forensic experts suffer substantial performance degradation on ID images with Visual Disturbances. (c) Existing detectors exhibit limited generalization across VD types, editing models, and ID document types.

To address VDFM, our core insight is to determine the provenance of anomalous evidence rather than merely detect its presence. This requires representing non-malicious VD and malicious tampering cues according to their provenance, while constraining prior-guided interactions so that noisy VD information does not obscure subtle tampering evidence. MLLMs complement forensic experts in this setting, while VD and tampering share low-level patterns, MLLMs capture VD semantics with less dependence on such traces. We therefore propose **CLEAR-ID**, which combine forensic expert with MLLM. MLLM-Guided Adaptive Evidence Representation (MAER) represents expert evidence as disturbance- and tampering-oriented features without assuming that they are mutually exclusive. Low-Rank Selective Modulation (LSM) then constrains the additive interference introduced by VD guidance within a low-rank residual space. To support this study, DISTURBID-180k provides clean and disturbed ID document images, annotations for both VD and tampering, controlled protocols, and an OCR-aware VD severity metric. Experiments validate the effectiveness of MAER and LSM across seen and unseen VD types and different severity levels.

In summary, our main contributions are:

Problem Definition and Evaluation. We identify Visual Disturbances Induced Forensic Misattribution (VDFM) and construct DISTURBID-180k with VD severity and localization mask annotations, together with controlled evaluation protocols to systematically analyze the impact of VD on tampering detection and localization.

MLLM-Guided Adaptive Evidence Representation (MAER). We introduce MAER to distinguish VD and tampering evidence. Specifically, MAER adaptively represents forensic evidence as disturbance-oriented and tampering-oriented features conditioned on VD priors from an MLLM-based disturbance localizer. This design enables the model to retain sensitivity to tampering across diverse VD conditions.

Low-Rank Selective Modulation (LSM). We develop LSM to mitigate noise from VD priors by constraining prior-conditioned corrections to a low-rank residual subspace. LSM controls the degrees of freedom through which VD priors affect tampering representations, suppressing disturbance-related interference while preserving tampering detection and localization evidence.

Experiments. CLEAR-ID achieves state-of-the-art performance in VDFM scenarios by improving tampering localization under both seen and unseen VD while reducing VD-induced false responses.

2 Related Work

Expert Forensic Detectors. Forensic experts model general anomalies, document-specific frequency cues, and complementary evidence. TruFor (Guillaro et al., 2023), Mesorch (Zhu et al., 2025), and SparseViT (Su et al., 2025) detect noise, residual, and semantic inconsistencies. CAT-Net (Kwon et al., 2021) and DTD (Qu et al., 2023) exploit DCT-domain cues, while FFDN (Chen et al., 2024) and ADCD-Net (Wong et al., 2025) better exploit frequency cues. DCL-Net (Li et al., 2026) distinguishes desensitization from tampering through three-class contrastive learning, but covers fewer disturbances than open-world ID-VD. These methods still struggle to attribute similar artifacts to non-malicious VD or malicious tampering. Existing datasets lack controls for this ambiguity. A detailed comparison with existing ID document image datasets is provided in the Appendix A.1. Here, we primary focus on how to distinguish VD-related and tampering-related cues extracted by forensic experts to prevent VDFM.

MLLM-Based Forensic Detectors. FakeShield (Xu et al., 2025a) and SIDA (Huang et al., 2025) extend MLLM forensics from explainable detection to joint detection and localization. ForgeryGPT (Zhang et al., 2024) combines an expert prediction map with MLLM reasoning. For documents, TextSleuth (Qu et al., 2024) explains forensic predictions, TextShield-R1 (Qu et al., 2026) improves grounded reasoning through forensic annotations and reinforcement learning, and TVSIP (Xu et al., 2025b) combines low-level expert features with MLLM semantics. These methods motivate complementary roles: MLLMs provide semantic context, while forensic experts capture statistical traces for precise localization. Drawing inspiration from and extending these schemes, CLEAR-ID uses forensic experts to extract evidence and an MLLM to model VD with semantic and spatial awareness, enabling provenance-aware attribution of the extracted evidence.

3 Problem Formulation

We define *ID Document Image Visual Disturbances* (ID-VD) as anomalies that are not part of an ID document's original design and that degrade the visual quality or recognizability of its content. Arising from digital, physical, or hybrid operations. Digital edits such as watermarks, mosaics, and logos, as well as physical effects such as glare and moiré, can introduce artifacts or high-frequency patterns that trigger responses from forensic detectors. When these responses are mistaken for malicious tampering, they cause false positives; we define this problem as *Visual Disturbance-Induced Forensic Misattribution* (VDFM).

Given an image $x \in \mathbb{R}^{H\times W\times 3}$, a forensic expert predicts a tampering mask $\widehat{M}_T \in \{0,1\}^{H\times W}$ and an image-level label $\widehat{y} \in \{0,1\}$. Let $\phi(x)_p \in \mathcal{Z}$ denote the feature at pixel p and $g : \mathcal{Z} \to [0,1]$ its tampering score. VD and tampering features share this input space, so VD may enter the tampering decision region $\mathcal{Z}_T = \{z : g(z) \geq \tau\}$:

$$M_V(p) > 0, \quad M_T(p) = 0, \quad \phi(x)_p \in \mathcal{Z}_T \quad \Longrightarrow \quad \widehat{M}_T(p) = 1. \tag{1}$$

Here, M_V and M_T are the VD and tampering ground-truth masks, and τ is the decision threshold. VDFM produces two operational errors: a false positive when VD-only evidence enters $\mathcal{Z}_T$, and a false negative when suppressing VD responses also removes genuine tampering evidence. We therefore report VD-specific FPR and tampering-specific FNR to quantify these two errors, respectively. Pixel-level F1, IoU, and AUC evaluate tampering localization, while image-level F1 evaluates document-level tampering detection.

4 Dataset Construction

Data sources and generation. Quantitative study of Visual Disturbances Induced Forensic Misattribution (VDFM) remains difficult, because current data fail to separate disturbance-induced effects from tampering artifacts, and do not categorize failures by VD type, editing model, or document type. We therefore construct DISTURBID with digital, physical, and hybrid disturbances and provenance-preserving VD and tampering annotations, as detailed in Appendix A.2. The dataset contains approximately 180k images derived from 9,078 source identities. Tampering covers text, face,

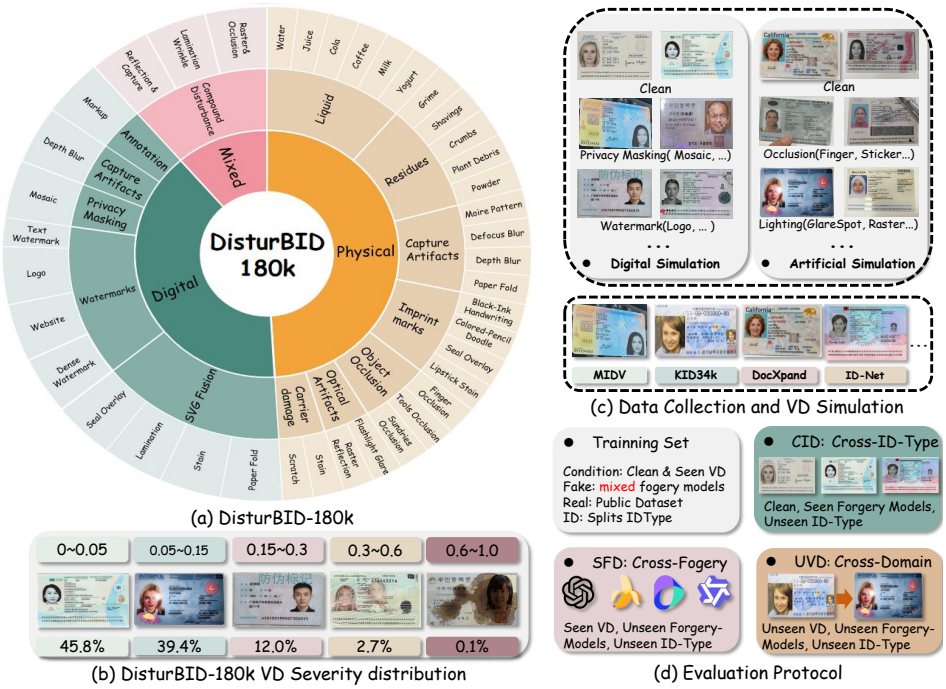


Figure 2: Construction and evaluation of DISTURBID. (a) Subsets under different VD splits. (b) VD-intensity intervals defined over J and the distribution of the dataset across these intervals. (c) Source documents undergo digital or artificial VD simulation, field-level tampering, screening, and annotation. (d) The protocol comprises Clean Cross-ID-Type (CID), Seen VD Cross-Forgery (SFD), and Unseen VD Cross-Domain (UVD) subsets.

signature, and mixed fields produced by diverse editing systems, with automated screening and manual verification retaining identifiable regions. Generation and filtering are detailed in Appendix A.2. Appendix A.3 further controls the rendering order.

VD severity distribution. We quantify the intensity of different disturbances on ID document content to account for VD heterogeneity and further investigate the mechanisms underlying VDFM. First, the ground Truth VD mask provides spatial information about the presence and extent of visual disturbances. Since VD primarily degrades document-content recognition, changes in OCR performance are used as an indirect indicator of disturbance severity. We first compute the intensity-weighted VD area and paired OCR change, and then define their joint intensity score:

$$A_V = \frac{1}{HW}\sum_{i,j} M_V(i,j), \qquad \Delta_{\mathrm{OCR}} = \min\left(1, \frac{S+D+I}{\max(1,|q_0|)}\right) \tag{2}$$

where q_0 denotes the clean-image transcription, and S, D, and I denote the character-level substitution, deletion, and insertion counts obtained by aligning the OCR transcription of the disturbed image with q_0, respectively. Figure 2(b) shows the proportions of samples across different intensity intervals, with only a small fraction corresponding to severe, large-area VD that may already be rejected during initial screening.

Evaluation protocols. As illustrated in Figure 2(d), the source-disjoint candidate pools comprise 59,748 training images and three test sets containing 5,590 Clean Cross-ID-Type (CID), 30,114 Seen Visual Disturbance Cross-Forgery (SFD), and 30,604 Unseen Visual Disturbance Cross-Domain (UVD) images. CID evaluates tampering localization without visual disturbances under cross-ID-type variation; SFD evaluates unseen forgery sources and document types under familiar VD; and UVD further evaluates cross-domain generalization when the VD type is unseen. All derivatives of the

same source identity remain within one split. Appendix A.3 specifies the pool composition and fixed protocol sampling.

5 METHODOLOGY

5.1 OVERALL FRAMEWORK

CLEAR-ID combines a forensic expert with a frozen MLLM-based VD localizer to distinguish disturbance-related evidence from genuine tampering evidence. As shown in Figure 3, the forensic expert extracts F_e, its original pixel-level prediction logits M_o, and its original image-level prediction logits Y_o, while the VD localizer provides the semantic feature S_D and spatial logits G_v. MAER fuses the resulting prior D_P with F_e and represents the evidence as disturbance-oriented F_d and tampering-oriented F_t, preserving tampering-sensitive expert cues. LSM then constrains VD-conditioned modulation of F_t to a low-rank space. Finally, a unified correction head takes F_t, F_t', and F_d and predicts pixel- and image-level corrections C_m and C_y, which are fused with the preserved expert decisions as $M_t = M_o + C_m$ and $Y_t = Y_o + C_y$. The independently trained, frozen VD localizer supplies a stable prior for forensic calibration.

5.2 PROVENANCE-AWARE EVIDENCE EXTRACTION

Because VD and tampering artifacts can activate the similar forensic features, we extract evidence that combines anomaly and tampering-related evidence from the forensic expert with semantic and spatial VD priors from the VD localizer. Given x, the forensic expert extracts anomaly evidence F_e and M_o through

$$(F_e, M_o, Y_o) = \mathcal{E}_\theta(x), \tag{3}$$

where $\mathcal{E}_\theta$ denotes the adopted forensic expert such as TTDMamba (Duan et al., 2025). F_e denotes the feature representation before the forensic expert's segmentation head and is passed to MAER, whereas M_o and Y_o denote the original pixel-level and image-level prediction logits, respectively, and are preserved as the targets of subsequent LSM modulation. In parallel, given the same image x and a fixed textual prompt t_V, the MLLM first identifies the corresponding VD type in the image and then projects its segmentation token to the semantic feature S_D. Following LISA (Lai et al., 2024), the SAM decoder (Kirillov et al., 2023) produces a VD spatial logit map G_v:

$$(S_D, G_v) = \mathcal{P}_\phi(x, t_V), \tag{4}$$

$$D_P = \mathrm{ADP}(S_D, G_v). \tag{5}$$

$\mathcal{P}_\phi$ denotes the VD localizer. ADP is an adapter that spatially aligns and fuses the semantic representation S_D with the spatial logits G_v. It produces D_P, which combines VD semantics and spatial support to condition expert evidence in MAER.

5.3 MLLM-GUIDED ADAPTIVE EVIDENCE REPRESENTATION

MAER addresses the source ambiguity between VD and tampering cues in forensic features. Instead of directly suppressing VD responsive features, it uses semantic and spatial VD priors from the MLLM to represent expert evidence according to its likely provenance. VD varies in appearance and severity; the MLLM combines generalizable semantics with spatial support to describe its type, extent, and strength. The MLLM supplies semantic context, while the forensic expert preserves tampering sensitive low-level cues, allowing tampering-oriented features to adapt to the VD prior without discarding forensic evidence shared with disturbances.

Specifically, the Adaptive Fusion (AF) module aligns and fuses the VD prior D_P with the expert feature F_e. The subsequent Evidence Representation (ER) module represents the adaptively fused feature as disturbance-oriented and tampering-oriented features:

$$[F_d, F_t] = \mathrm{ER}(\mathrm{AF}(F_e, D_P)). \tag{6}$$

ER is a feature representation module that separates the evidence into tampering-oriented F_t and disturbance-oriented representation F_d. To make the role of F_d identifiable, we further anchor it with complementary semantic and spatial VD constraints. Specifically, F_d is mapped into a

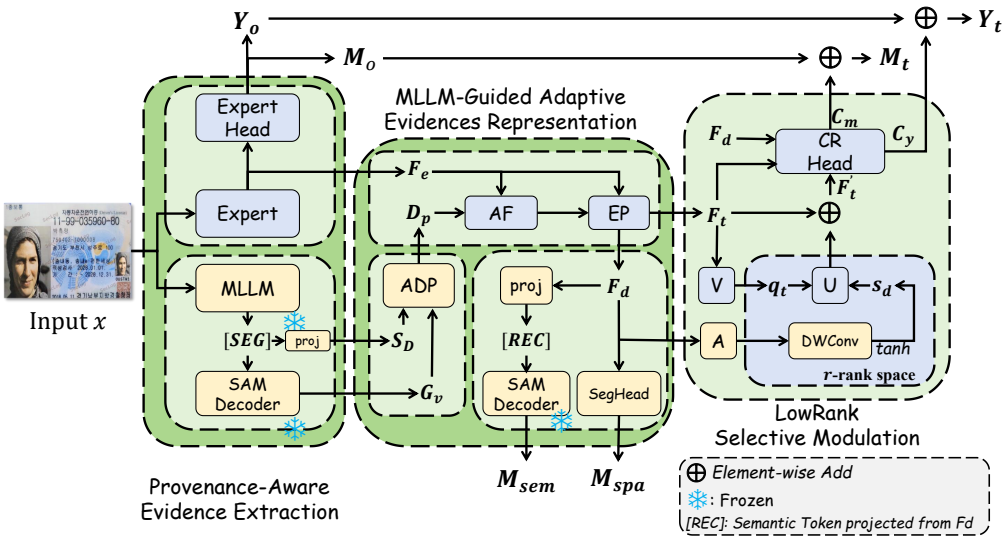


Figure 3: **Overall Framework of CLEAR-ID.** (a) The forensic expert produces F_e, M_o, and Y_o, while the MLLM identifies the VD type, projects its [SEG] token to S_D, and uses the SAM decoder to obtain G_v; ADP fuses them into D_P. (b) In MAER, Adaptive Fusion (AF) combines D_P with F_e, Evidence Representation (ER) produces F_d and F_t, and two reconstruction heads anchor F_d to VD evidence. (c) LSM modulates F_t in a rank-r space, and the correction head predicts C_m and C_y for residual fusion with the pixel- and image-level expert outputs.

new segmentation token [REC] and decoded by the frozen VD-SAM decoder, while a lightweight segmentation head directly predicts VD regions from F_d:

$$M_{\text{sem}} = \mathcal{D}_{\text{SAM}}([\text{REC}]), \qquad M_{\text{spa}} = \mathcal{H}_d(F_d). \tag{7}$$

The semantic constraint requires F_d to retain VD-related semantic information, while the spatial constraint grounds this information to the corresponding disturbance regions. Together, they anchor F_d to disturbance-related evidence, thereby encouraging the forensic expert evidence F_e, after adapting to the VD prior D_P, to preserve cues attributed to tampering in F_t for subsequent modulation.

5.4 Low-Rank Selective Modulation

An unrestricted VD-conditioned update could rewrite every expert channel and amplify an erroneous prior. LSM instead treats calibration as a bounded residual correction in a rank-r subspace. The low rank is a deliberate capacity constraint: it limits the number of directions that VD evidence can change, reduces parameters and computation, and preserves the expert's high-dimensional solution outside those directions. The interaction is defined by

$$F_t' = F_t + U(s_d \odot q_t), \tag{8}$$

As shown in Eqs. 8, where $q_t = V^\top F_t$, $s_d = \tanh(\text{DWConv}(A(F_d)))$. $V^\top$ is a learned 1×1 down-projection from C channels to rank r, and U is a learned 1×1 up-projection from rank r back to C channels. A is a learned 1×1 conditioning projection that maps F_d into the same rank-r space, after which DWConv captures local VD structure in each direction. Thus, q_t is the low-rank tampering representation, s_d is a bounded spatial disturbance gate, and $U(s_d \odot q_t)$ lifts the gated update back to the expert feature space. The residual connection preserves the original F_t, allowing LSM to refine rather than overwrite tampering evidence.

Feature-space interaction uses VD priors to distinguish anomaly provenance and modulate tampering evidence. Logit-space residual fusion is designed to address VDFM while preserving the expert's original decisions to the greatest extent possible. A unified correction head receives F_t, F_t', and F_d and predicts pixel- and image-level corrections:

$$(C_m, C_y) = \mathcal{H}_c(F_t, F_t', F_d). \tag{9}$$

Here, C_m and C_y denote the corrections for the pixel- and image-level outputs, respectively. The corrected logits are $M_t = M_o + C_m$ and $Y_t = Y_o + C_y$. After applying the sigmoid function, we use 0.5 as the threshold to convert M_t into a binary localization mask and Y_t into a binary image-level decision. The dedicated image-level logit directly determines whether tampering is present, while residual fusion preserves the expert's original localization and classification decisions and limits correction to the remaining VDFM errors.

5.5 Two-Stage Training

In direct end-to-end training, the forensic pathway must adapt to a VD prior that is still inaccurate and evolving during early optimization, which can destabilize representation learning. We therefore adopt a two-stage training strategy. In Stage 1, we first train the VD localizer using:

$$\mathcal{L}_{\text{stage1}} = \mathcal{L}_{\text{text}} + \lambda_B \mathcal{L}_{\text{BCE}}(\widehat{M}_V, M_V) + \lambda_D \mathcal{L}_{\text{Dice}}(\widehat{M}_V, M_V). \tag{10}$$

In Stage 2, we freeze the VD localizer and jointly train the forensic expert and the remaining CLEAR-ID modules under the fixed prior D_P:

$$\begin{aligned} \mathcal{L}_{\text{stage2}} &= \mathcal{L}_{\text{tamper}} + \lambda_d \mathcal{L}_{\text{VD}}, \\ \mathcal{L}_{\text{VD}} &= \lambda_E \mathcal{L}_{\text{Dice}}(M_{\text{sem}}, M_V) + \lambda_P \mathcal{L}_{\text{Dice}}(M_{\text{spa}}, M_V). \end{aligned} \tag{11}$$

Here, $\mathcal{L}_{\text{tamper}}$ supervises tampering detection and localization, M_V is the ground-truth VD mask, and M_{sem} and M_{spa} are the semantic and spatial VD prediction maps defined in Eq. 7. The semantic and spatial VD constraints keep F_d aligned with disturbances, allowing F_t to retain tampering evidence. Detailed losses and optimization settings are given in Appendix B.

6 Experiments

6.1 Experimental Settings

Dataset. We follow the evaluation protocol of DisturbID in experiments, with details provided in Section 4 and Appendix A. IDCD (Zheng et al., 2025) is used separately as a private cross-domain dataset to evaluate the performance of different models in more realistic scenarios.

Metrics. We report pixel-level F1_p and IoU for localization, and image-level F1_i for detection. Following TruFor (Guillaro et al., 2023), we apply the sigmoid function to the pixel- and image-level logits and use 0.5 as the threshold for both the binary localization mask $\widehat{M}_T$ and the binary image-level decision $\widehat{y}$. We additionally report FPR and FNR to analyze the two VDFM errors. Details of metric aggregation are provided in Appendix B.3.

Baselines and implementation. Under our fixed training and evaluation protocol, we compare general image-forensics experts TruFor (Guillaro et al., 2023), Mesorch (Zhu et al., 2025), and SparseViT (Su et al., 2025); document-forensics experts CAT-Net (Kwon et al., 2021), DTD (Qu et al., 2023), FFDN (Chen et al., 2024), ASCFormer (Luo et al., 2025), TAFE-Net (Yi et al., 2026), CD-SD (Zheng et al., 2025), DCLNet (Li et al., 2026), and TTDMamba (Duan et al., 2025); and MLLM-based detectors including a cascade of TTDMamba (Duan et al., 2025) and LISA (Lai et al., 2024), and SIDA-7B (Huang et al., 2025). We choose the strongest SOTA expert under this protocol as the backbone. Training use a learning rate of 2×10^{-5}, weight decay 10^{-4}, and batch size 8. Other details are provided in Appendix B.1 and Appendix B.

6.2 Results

We first evaluate whether CLEAR-ID improves tampering localization under clean, seen-disturbance, and unseen-disturbance conditions. The comparison with forensic experts measures overall effectiveness, while the component ablation isolates the gains from MAER and LSM.

Comparison with SOTA Experts. CLEAR-ID achieves the strongest tampering localization across clean, seen-VD, and unseen-VD conditions as given in Table 1, with larger gains under VD. These results indicate that provenance-aware representation reduces VD-induced false responses

Table 1: Tampering detection and localization performance (%) on the DISTURBID protocol.

Method	CID				SFD				UVD				Avg.	
	$F1_p$	IoU	AUC	$F1_i$	$F1_p$	IoU	AUC	$F1_i$	$F1_p$	IoU	AUC	$F1_i$	$F1_p$	$F1_i$
General image forensic methods														
TruFor (CVPR'23)	5.16	2.65	52.53	54.23	12.01	6.39	55.07	83.01	7.96	4.15	53.01	85.81	8.38	74.35
Mesorch (AAAI'25)	7.68	3.99	53.00	50.57	8.05	4.19	52.43	85.75	6.98	3.61	52.04	86.99	7.57	74.44
SparseViT (AAAI'25)	5.99	3.09	52.49	54.09	7.90	4.11	52.52	85.01	7.38	3.83	52.36	85.66	7.09	74.92
Document image forensic methods														
CAT-Net (WACV'21)	25.32	14.50	79.38	64.55	34.54	20.87	78.02	90.11	32.98	19.75	77.93	90.41	30.95	81.69
DTD (CVPR'23)	12.19	6.49	59.49	61.83	15.90	8.64	57.49	80.22	15.03	8.13	57.63	81.69	14.37	74.58
ASCFormer (PR'25)	69.90	53.73	82.67	91.30	68.52	52.11	80.16	**93.79**	68.04	51.56	80.76	93.72	68.82	92.94
TAFE-Net (AAAI'26)	37.18	22.83	64.00	48.22	38.34	23.72	63.47	72.06	41.53	26.21	65.05	74.22	39.02	64.83
CD-SD (JVCIR'25)	69.65	53.43	79.78	**92.49**	69.29	53.01	79.07	93.51	68.06	51.58	79.39	**93.85**	69.00	**93.29**
DCLNet (SP'26)	38.53	23.86	65.96	44.14	25.35	14.51	59.51	67.45	22.88	12.92	59.34	69.20	28.92	60.26
FFDN (ECCV'24)	52.88	35.94	80.69	88.24	54.88	37.81	79.55	91.38	54.81	37.75	78.81	90.89	54.19	90.17
TTDMamba (IJCV'25)	74.48	59.34	81.97	87.85	68.34	51.91	78.17	91.34	71.00	55.04	80.70	92.90	71.27	90.70
MLLM forensic methods														
TTDMamba + LISA cascade	36.40	22.25	65.19	69.96	24.58	14.01	59.04	73.12	32.39	19.32	62.59	79.40	31.12	74.16
SIDA-7B (CVPR'25)	12.54	6.69	79.86	60.11	10.52	5.55	60.57	80.55	11.17	5.91	62.31	82.72	11.41	74.46
Proposed method														
CLEAR-ID (TTDMamba)	**75.64**	**60.82**	**83.26**	90.47	**74.37**	**59.19**	**82.21**	93.16	**74.97**	**59.96**	**83.99**	93.73	**74.99**	92.45

Table 2: Performance using different VD priors D_P ($F1_p$, %). VD-GT row denotes training with VD-GT as the VD prior.

VD prior D_P	CID	UVD	UVD FPR	UVD FNR
No MLLM	69.42	66.48	0.85	33.38
SAM Output	71.28	64.77	0.67	39.38
VD-GT Masks	72.09	66.10	0.77	35.52
Qwen3-VL-4B (untrained)	72.31	62.22	0.64	43.24
CLEAR-ID-4B (semantic prior)	**75.64**	**74.97**	**0.35**	**31.68**

Table 3: Module ablation results for CLEAR-ID ($F1_p$, %). Oracle-GT subtracts VD-GT from clean-only predictions.

Method	SFD	UVD	UVD FPR	UVD FNR
TTDMamba (Clean-only)	57.83	55.75	0.98	46.31
TTDMamba (Oracle GT)	53.43	54.52	**0.00**	53.36
TTDMamba (VD-aware)	68.34	71.00	0.30	38.47
MAER only	72.89	73.68	0.43	31.72
MAER + LSM (CLEAR-ID)	**74.37**	**74.97**	0.35	**31.68**

while preserving tampering evidence. Figure 8 further shows that existing methods tend to over-localize under unseen VD, whereas VD-aware models may become overly conservative; CLEAR-ID alleviates both behaviors. Interestingly, image-level F1 is higher on VD-bearing subsets than on clean CID, suggesting that VD may activate additional tampering responses that benefit image-level detection but impair localization. This further motivates distinguishing VD and tampering rather than treating them as the forgery.

Cross-Validation. Appendix C.2 reports alternative seen/unseen VD assignments and cross-validation results. Forensic expert performance changes substantially with the VD assignment, even on the CID set, indicating that different VD affect forensic models with varying magnitudes and directions. This makes indiscriminate suppression unsuitable, while CLEAR-ID remains effective across different VD assignments.

Ablation Experiment As shown in Table 3 and Table 20, MAER improves performance under both seen and unseen VD through provenance-aware representation. LSM further reduces false responses over MAER, supporting constrained VD-guided modulation after evidence representation.

6.3 FURTHER ANALYSIS

We next analyze why CLEAR-ID improves under VDFM. The studies isolate the VD information required by MAER, the modulation capacity controlled by LSM, and their behavior across VD severity, post-processing, forensic backbones, and deployment cost.

Analysis for MAER and LSM. Spatial support alone, even from VD ground truth, does not improve UVD localization over the no-MLLM baseline. MAER must jointly model the provenance and severity of VD pixels: semantic information identifies the disturbance source, while spatial responses describe its local extent and strength. The learned semantic–spatial prior therefore provides

the effective conditioning observed in Table 2. LSM controls how much VD information can alter the tampering representation. A very small rank lacks sufficient directions to recover weak tampering cues, while a large rank admits excessive prior-conditioned variation. Rank 16 achieves the best balance between correction capacity and evidence preservation, as shown in Appendix C.4.

Analysis under different VD Severity. MAER provides the principal gain across nonzero VD-severity intervals by organizing disturbance and tampering evidence according to provenance. In low-severity, high-noise conditions, LSM compensates for the remaining weakness by restricting noisy prior-guided updates and preserving subtle tampering cues. Their complementary effect across J is shown in Figure 4; Appendix C.5 reports the source values and severity-stratified analysis.

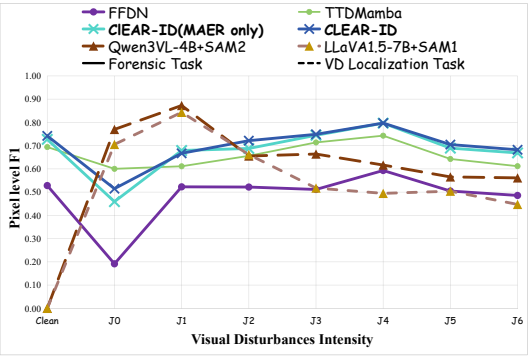


Figure 4: Tampering- and VD-localization $\mathrm{F1}_p$ across VD-severity intervals.

MAER Representation Constraints. We further decompose MAER to examine how its output supervision affects evidence representation. Removing spatial anchoring weakens UVD localization, and training the SAM decoder without anchoring causes a larger decline. Complementary semantic and spatial supervision therefore stabilizes the disturbance-oriented representation F_d, as shown in Appendix C.3.

Robustness. Across the post-processing robustness conditions, CLEAR-ID remains close to its TTDMamba backbone and achieves higher F1 under the strongest Gaussian-noise and JPEG settings. These operations can introduce VD-like artifacts, making provenance-aware modeling beneficial; the full comparison is shown in Appendix C.7.

Other Analyses. Cross-domain evaluation across open-source and closed-source forensic MLLMs on IDCD is provided in Appendix C.1.2. Initialization and backbone compatibility are shown in Appendix C.6, and computational efficiency is reported in Appendix C.8. Together, they show that CLEAR-ID can use multiple forensic representations, while the semantic VD prior remains the main source of deployment cost.

7 Discussion and Conclusion

Conclusion. We studied Visual-Disturbance-Induced Forensic Misattribution (VDFM), where non-malicious VD are confused with tampering evidence in ID document images. We constructed DISTURBID-180k for systematic evaluation and proposed CLEAR-ID, which combines MLLM-guided provenance-aware representation with low-rank selective modulation. Experiments across CID, SFD, and UVD show improved tampering detection and localization with reduced VDFM. These results suggest that ID document forensics should identify anomalous evidence according to its provenance rather than treating all anomalous responses as tampering.

Insights. ID document forensics should model anomaly provenance, since suppressing anomalous responses may erase subtle tampering traces, while ignoring VD can induce false positives. As VD vary in source, appearance, and intensity, augmentation alone may not generalize to unseen disturbances. As shown in Section 6.3, CLEAR-ID with semantic and spatial priors outperforms the VD-GT model by 8.87 percentage points, supporting provenance-aware modeling of anomalous evidence.

Limitations. The limitations of CLEAR-ID mainly stem from MAER and LSM. MAER depends on the quality of the VD prior, while LSM requires an appropriate subspace rank for each forensic backbone. The MLLM and segmentation decoder introduce substantial deployment cost. Improving efficient generalization across diverse VD conditions remains an important research direction.

ETHICS STATEMENT

ID images contain sensitive personal information. All displayed examples are synthetic or sourced from public datasets, and all publicly shared examples are licensed for this use. These data and examples are intended solely for research and must not be used for any other purpose. Our method supports human review of suspected ID document tampering. We study images with visual disturbances because acquisition artifacts and benign edits can otherwise trigger false tampering decisions.

REPRODUCIBILITY STATEMENT

Appendix A details data construction, annotations, and evaluation protocols. Appendix B provides model implementations, training settings, loss weights, and inference procedures to support reproduction of the reported experiments.

DATA AVAILABILITY STATEMENT

We will release data collected from public datasets and data created for this study, subject to the applicable source permissions. Selected test subsets and real images collected from volunteers will remain private. Appendix A.2 specifies the data sources and availability.

AUTHOR CONTRIBUTIONS

Author contributions will be provided in the camera-ready version to preserve anonymity during review.

CONFLICT OF INTEREST AND FUNDING

The authors declare no conflict of interest. Funding information is withheld for anonymous review and will be disclosed in the camera-ready version.

AI USE STATEMENT

Yes, to aid or polish writing. GPT and Gemini assisted language polishing and translation.

Yes, for research ideation or execution. GPT, Gemini, and Qwen assisted research ideation and experimental analysis design; Codex assisted implementation of the proposed method and selected comparison methods, as detailed in Appendix B.1; and a locally deployed Qwen-VL-72B model supported tampering/data ideation and data cleaning.

Yes, to draft sections of the paper. GPT assisted in drafting and revising portions of the manuscript; the authors then carefully revised the text and performed the final verification.

Yes, for generating synthetic datasets. Generative AI tools assisted the creation of selected SVG assets and tampered images. Appendix A.2 describes the editing pipeline and lists the image-editing models. All numerical claims were traced to the experiment ledger, citations were manually checked, and the authors remain responsible for the final text, claims, code, and artifacts.

REFERENCES

Vladimir Viktorovich Arlazarov, Konstantin Bulatovich Bulatov, Timofey Sergeevich Chernov, and Vladimir Lvovich Arlazarov. Midv-500: a dataset for identity document analysis and recognition on mobile devices in video stream. *Computer Optics*, 43(5):818–824, 2019.

Bulatov Konstantin Bulatovich, Emelianova Ekaterina Vladimirovna, Tropin Daniil Vyacheslavovich, Skoryukina Natalya Sergeevna, Chernyshova Yulia Sergeevna, Ming Zuheng, Burie Jean-Christophe, Luqman Muhammad Muzzamil, and Arlazarov Vladimir Viktorovich. Midv-2020: A comprehensive benchmark dataset for identity document analysis. *Computer Optics*, 46(2): 252–270, 2022.

Zhongxi Chen, Shen Chen, Taiping Yao, Ke Sun, Shouhong Ding, Xianming Lin, Liujuan Cao, and Rongrong Ji. Enhancing tampered text detection through frequency feature fusion and decomposition. In *European Conference on Computer Vision*, pp. 200–217. Springer, 2024.

AV Chuiko, IA Kunina, SA Usilin, C Chen, S Tan, DP Nikolaev, and VV Arlazarov. Midv-dm: A document-oriented dataset for image manipulation detection and localization. computer optics 2025; 49 (6): 1093-1101. doi: 10.18287/coj1768, 2025.

Ricardo Batista das Neves Junior, Byron Leite Dantas Bezerra, and Cleber Zanchettin. Doclightdetect: a new algorithm for occlusion classification in identification documents. In *International Workshop on Document Analysis Systems*, pp. 196–210. Springer, 2024.

Junxian Duan, Hao Sun, Fan Ji, Kai Zhou, Zhiyong Wang, Huaibo Huang, and Lianwen Jin. Realdtt: Towards a comprehensive real-world dataset for tampered text detection: J. duan et al. *International Journal of Computer Vision*, 133(10):6993–7011, 2025.

Fabrizio Guillaro, Davide Cozzolino, Avneesh Sud, Nicholas Dufour, and Luisa Verdoliva. Trufor: Leveraging all-round clues for trustworthy image forgery detection and localization. In *2023 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 20606–20615. IEEE, 2023.

Zhenglin Huang, Jinwei Hu, Xiangtai Li, Yiwei He, Xingyu Zhao, Bei Peng, Baoyuan Wu, Xiaowei Huang, and Guangliang Cheng. Sida: Social media image deepfake detection, localization and explanation with large multimodal model. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28831–28841. IEEE, 2025.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *2023 IEEE/CVF international conference on computer vision (ICCV)*, pp. 3992–4003. IEEE, 2023.

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, 2025.

Myung-Joon Kwon, In-Jae Yu, Seung-Hun Nam, and Heung-Kyu Lee. Cat-net: Compression artifact tracing network for detection and localization of image splicing. In *2021 IEEE winter conference on applications of computer vision (WACV)*, pp. 375–384. IEEE, 2021.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9579–9589. IEEE, 2024.

Julien Lerouge, Guillaume Betmont, Thomas Bres, Evgeny Stepankevich, and Alexis Bergès. Docxpand-25k: a large and diverse benchmark dataset for identity documents analysis. *arXiv preprint arXiv:2407.20662*, 2024.

Weixiang Li, Bin Li, Kengtao Zheng, Songze Li, and Haodong Li. Document image forgery detection and localization in desensitization scenarios. *Signal Processing*, 238:110123, 2026.

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.

Eun-Ju Park, Seung-Yeon Back, Jeongho Kim, and Simon S Woo. Kid34k: A dataset for online identity card fraud detection. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pp. 5381–5385, 2023.

Dmitry V Polevoy, Irina V Sigareva, Daria M Ershova, Vladimir V Arlazarov, Dmitry P Nikolaev, Zuheng Ming, Muhammad Muzzamil Luqman, and Jean-Christophe Burie. Document liveness challenge dataset (dlc-2021). *Journal of Imaging*, 8(7):181, 2022.

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 *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5937–5946. IEEE, 2023.

Chenfan Qu, Jian Liu, Haoxing Chen, Baihan Yu, Jingjing Liu, Weiqiang Wang, and Lianwen Jin. Textsleuth: Towards explainable tampered text detection. *arXiv preprint arXiv:2412.14816*, 2024.

Chenfan Qu, Yiwu Zhong, Jian Liu, Xuekang Zhu, Bohan Yu, and Lianwen Jin. Textshield-r1: Reinforced reasoning for tampered text detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 8621–8629, 2026.

Lei Su, Xiaochen Ma, Xuekang Zhu, Chaoqun Niu, Zeyu Lei, and Ji-Zhe Zhou. Can we get rid of handcrafted feature extractors? sparsevit: Nonsemantics-centered, parameter-efficient image manipulation localization through spare-coding transformer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 7024–7032, 2025.

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 *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19280–19289. IEEE, 2025.

Lulu Xie, Yancheng Wang, Hong Guan, Soham Nag, Rajeev Goel, Niranjan Swamy, Yingzhen Yang, Chaowei Xiao, Jonathan Prisby, Ross Maciejewski, et al. Idnet: A novel identity document dataset via few-shot and quality-driven synthetic data generation. In *2024 IEEE International Conference on Big Data (BigData)*, pp. 2244–2253. IEEE, 2024.

Guitao Xu, Ziqi Yi, Peirong Zhang, Jiahuan Cao, Shihang Wu, and Lianwen Jin. From pixels to semantics: A novel mllm-driven approach for explainable tampered text detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 757–766, 2025b.

Zhipei Xu, Xuanyu Zhang, Runyi Li, Zecheng Tang, Qing Huang, and Jian Zhang. Fakeshield: Explainable image forgery detection and localization via multi-modal large language models. In *International Conference on Learning Representations*, volume 2025, pp. 31186–31216, 2025a.

Ziqi Yi, Guitao Xu, Shihang Wu, Peirong Zhang, and Lianwen Jin. Frequency mining empowered by text aggregation: A new perspective on document image tampering detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 1471–1479, 2026.

Fanrui Zhang, Jiawei Liu, Jiaying Zhu, Esther Sun, Dong Li, Qiang Zhang, and Zheng-Jun Zha. Forgerygpt: A multimodal llm for interpretable image forgery detection and localization. *arXiv preprint arXiv:2410.10238*, 2024.

Shiqiang Zheng, Changsheng Chen, Shen Chen, Taiping Yao, Shouhong Ding, Bin Li, and Jiwu Huang. Generalized document tampering localization via color and semantic disentanglement. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

Xuekang Zhu, Xiaochen Ma, Lei Su, Zhuohang Jiang, Bo Du, Xiwen Wang, Zeyu Lei, Wentao Feng, Chi-Man Pun, and Ji-Zhe Zhou. Mesoscopic insights: orchestrating multi-scale & hybrid architecture for image manipulation localization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 11022–11030, 2025.

APPENDIX CONTENTS

A Data
- A.1 Comparison with Existing ID Document Datasets
- A.2 Dataset Construction and Statistics
- A.3 Evaluation Protocols

B Models
- B.1 Reproduction and Implementation
- B.2 Training Details
- B.3 Inference and Evaluation

C Experiments
- C.1 Main and External-Domain Results
  - C.1.1 Main Protocol
  - C.1.2 Cross-Domain Evaluation on IDCD
- C.2 Cross-Validation
- C.3 MAER and Module Ablation Analysis
- C.4 LSM Analysis
- C.5 VD-Intensity Analysis
- C.6 Initialization and Backbone Compatibility
- C.7 Robustness
- C.8 Efficiency

A DATA

A.1 COMPARISON WITH EXISTING ID DOCUMENT DATASETS

We compares existing ID document image resources along the key factors required to study tampering under visual disturbances as shown in Table 4. Recognition and presentation-attack datasets, such as MIDV-500 (Arlazarov et al., 2019) and MIDV-2020 (Bulatovich et al., 2022), provide diversity in document types and acquisition conditions, while FantasyID (Korshunov et al., 2025), MIDV-DM (Chuiko et al., 2025), and ID-Net (Xie et al., 2024) additionally include digitally tampering samples. Their manipulated content mainly focuses on text and portrait fields, whereas DISTURBID-180K further covers signature and mixed-region tampering. DLC-2021 (Polevoy et al., 2022) and SpotBID (das Neves Junior et al., 2024) include acquisition- or illumination-related disturbances, but existing resources do not provide ID document images under other VD conditions or independent pixel-level VD mask annotations.

Existing resources separately offer large-scale synthetic identities, localized text/portrait forgery, or a narrow capture disturbance. They do not jointly control forgery region, VD type, and source identity, and cannot separate tampering errors inside and outside independently annotated VD support. DISTURBID-180k supplies VD ID document images, 42 VD types, independent M_V and M_T, and factor-controlled CID/SFD/UVD splits, enabling controlled VDFM analysis.

A.2 DATASET CONSTRUCTION AND STATISTICS

Data sources and availability. Table 5 reports all 180,059 samples by source, data form, and generation stage. The clean/warp and VD pipelines preserve the Real/Fake status inherited from the source record, while the tampering pipeline adds the final edited samples. Table 6 then summarizes the subset pools, editing systems, and release levels.

Data availability statement. As summarized in Table 6, the dataset has three release levels. Public pools contain release-eligible training images and derivatives; partially public pools combine releasable records with sources governed by licenses or participant permissions; private pools, including IDCD and the volunteer set, are reported through aggregate results and audit metadata. We will release the VD generation pipeline, protocol manifests, sampling configurations, evaluation code, and all release-eligible annotations and assets.

VD simulation. We construct VD through digital-native operations, manual physical simulation, and digital rendering. For digital-native text watermarks, we use the Qwen API to generate 400 Chinese

Table 4: Document image dataset comparison. Tampered region lists the document fields for which a manipulation region/type is explicitly supported by the source paper; × denotes no applicable region.

Dataset	Images / source IDs	ID types	Tampered region	VD types	VD mask
MIDV-500 (Arlazarov et al., 2019)	50 sources	50	×	0	×
MIDV-2020 (Bulatovich et al., 2022)	72,409 images / 1,000 documents	10	×	0	×
DLC-2021 (Polevoy et al., 2022)	1,424 video clips / 1,000 documents	10	×	1	×
SpotBID (das Neves Junior et al., 2024)	22k images	8	×	1	×
KID34K (Park et al., 2023)	34,662 images / 82 cards	2	×	1	×
DocXPand (Lerouge et al., 2024)	24,994 images	9	×	0	×
FantasyID (Korshunov et al., 2025)	362 source cards	13	Text / portrait	0	×
MIDV-DM (Chuiko et al., 2025)	9,000 images / 1,000 sources	10	Text / portrait	0	×
ID-Net (Xie et al., 2024)	837,060 images	20	Text / portrait	0	×
DISTURBID-180k	180k-scale / 9,078 sources	62	Text / portrait / signature / mixed	42	✓

Table 5: Dataset composition by source and generation stage. Entries in the three pipeline columns are Real/Fake counts.

Source	Data form	Source samples	After tampering pipeline	After After VD pipeline	Total
IDNet	Electronic	2,288/1,749	3262/2531	25,722/26,569	25,722/26,569
IDC	Electronic	930/2,535	1,326/3,668	10,455/38,508	10,455/38,508
KID	Physical capture	249/216	355/313	2,799/3,281	2,799/3,281
MIDV-2019	Physical capture	0/164	0/237	0/2,491	0/2,491
MIDV-2020 photo	Physical capture	767/488	1,093/706	8,622/7,413	8,622/7,413
MIDV-2020 scan	Scanned	999/573	1,424/829	11,231/8,704	11,231/8,704
Network collection	Physical capture	0/77	0/111	0/1,170	0/1,170
Printed	Physical capture	748/1,339	1,066/1,937	8,409/20,340	8,409/20,340
Volunteer	Physical capture	98/80	140/116	1,102/1,215	1,102/1,215
Physical test	Printed	311/0	856/0	856/1172	856/1172
Total	**Electronic + physical capture**	**6,079/7,221**	**8,666/10,448**	**68,885/109,146**	**180,059**

and English text entries, then randomly sample their content, color, and placement. Logo overlays use commercial logos and logos from websites that request ID documents, collected online. For mosaics, RapidOCR and InsightFace locate text and face regions; scripts randomly apply mosaics within these regions and record their masks. For manual physical simulation, 40 volunteers capture ID documents and simulate disturbances using phones from Apple, Huawei, vivo, OPPO, and other manufacturers. For digital rendering, SVG assets simulate lamination, wrinkles, scratches, stains, and seals. Table 7 reports the 147,645 simulated material records in the SFD and UVD candidate pools; together with the 2,028 physical-test records, the total number of VD records is 149,673. Mosaic/face-mosaic and dense/text watermark are grouped only for this summary; the original VD type is retained in the protocol manifests.

Tampering generation and filtering. We first ask Qwen-VL-72B to propose a target bounding box and an editing instruction for an ID document. A script crops the specified region and sends the crop and instruction to an image-editing model. The edited crop is pasted back at its original coordinates, producing the tampered image, localization mask, and image-level label. Figure 6 illustrates representative text, portrait, and signature edits. Table 10 lists the editing models. We then use Qwen to exclude samples whose edited content is unreadable, since such edits may represent VD rather than meaningful content manipulation. We also remove a subset of easily detected forgeries that Qwen correctly identifies, reducing conspicuous editing artifacts in the retained data.

Table 6: Dataset pools, editing systems, and release levels. Clean + warp and material-damaged columns describe the candidate pool before protocol sampling.

Subset	Clean	Visual Disturbances	Final pool	Sources	Editing systems	Acess
Training	9,192	50,556	**59,748**	IDNet, KID, MIDV-2019, MIDV-2020, Printed	IDNet-copy-move; IDNet-crop-replace; doubao-seededit-3-0; flux-kontext-max; gemini-2.5-flash-image-preview; qwen-image-edit	Public
CID	5,590	0	**5,590**	IDNet, IDC, MIDV-2019, MIDV-2020, network collection, Printed, Volunteer	IDNet-copy-move; IDNet-crop-replace; doubao-seededit-3-0; flux-kontext-max; gemini-2.5-flash-image-preview; qwen-image-edit	Partially public
SFD	4,302	25,812	**30,114**	IDNet, IDC, MIDV-2019, MIDV-2020, network collection, Printed, Volunteer	BAGEL-7B-MoT; IDNet-inpaint-rewrite; OmniGen2; flux-kontext-pro; gpt-4o-image; step1xedit	Partially public
UVD	4,372	26,232	**30,604**	IDNet, IDC, MIDV-2019, MIDV-2020, network collection, Printed, Volunteer	BAGEL-7B-MoT; IDNet-inpaint-rewrite; OmniGen2; flux-kontext-pro; gpt-4o-image; step1xedit	Partially public
IDCD	6,930	45,045	**51,975**	IDC (Zheng et al., 2025)	BAGEL-7B-MoT; step1xedit; gpt-4o-image; flux-kontext-pro; gemini-2.5-flash-image-preview; doubao-seededit-3-0; qwen-image-edit; OmniGen2; flux-kontext-max	Private
Physical test	0	2,028	**2,028**	Printed	doubao-seededit-3-0; gpt-4o-image; gemini-2.5-flash-image-preview; qwen-image-edit; flux-kontext-pro; flux-kontext-max	Public
Total	**30,386**	**149,673**	**180,059**			

Table 7: Distribution of simulated VD records in the SFD and UVD candidate pools.

VD type	SFD	UVD	Total
Mosaic / face mosaic	14,764	14,765	29,529
Watermark (dense/text)	14,764	14,765	29,529
Paper	7,382	7,382	14,764
Seal	7,382	7,382	14,764
Flashlight	7,382	7,382	14,764
Stain	7,383	7,382	14,765
Scratch	7,383	7,382	14,765
Lamination	7,382	7,383	14,765
Total	**73,822**	**73,823**	**147,645**

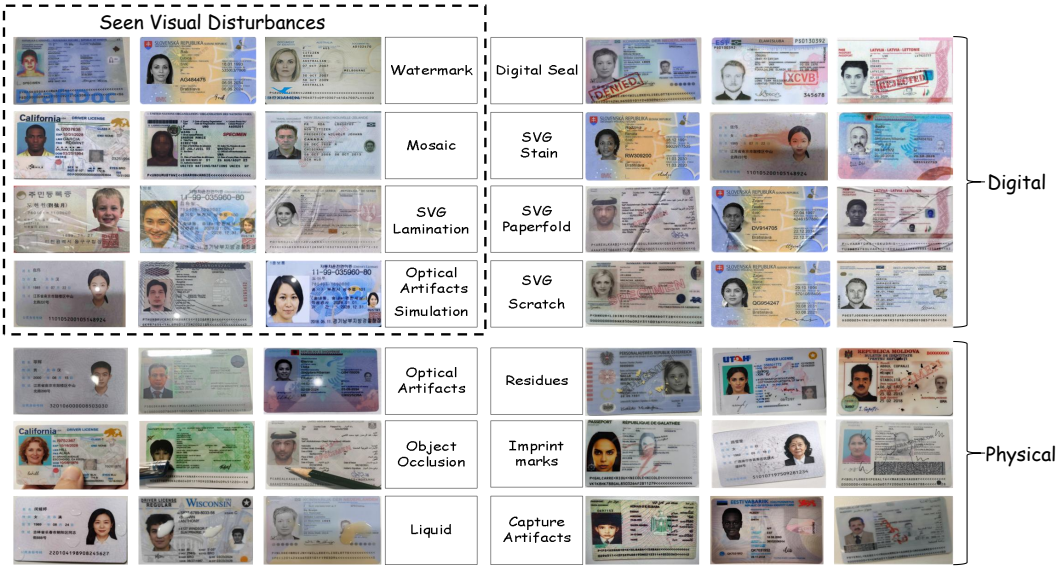


Figure 5: Examples of our self-constructed Visual Disturbances in DISTURBID-180k

VD annotation. For digital-native VD, generation scripts produce localization masks together with the disturbance. For digitally rendered VD, the continuous grayscale SVG support provides the VD annotation. For physical VD, manual annotations train a localization model; selected high-quality predictions are combined with manual annotations to refine the masks. The VD mask M_V records disturbance support, and M_T labels tampered pixels. Figure 7 shows rendered VD together with the corresponding M_V and independently retained M_T. For digital-native and digitally rendered samples, VD intensity combines the intensity-weighted mask area A_V with the OCR change Δ_{OCR} between images before and after VD, using $J = A_V + \Delta_{\mathrm{OCR}}$ as defined in Section 4.

Table 8: Overall distribution of DISTURBID-180k (%).

Condition	Clean Real	Clean Tampered	VD Real	VD Tampered
Image level Ratio	3.53	4.19	34.65	57.63
Pixel level Ratio	89.35	1.88	8.58	0.19
Tampered Region	**Text**	**Portrait**	**Signature**	**Mix**
Ratio	46.95	6.07	6.26	40.72

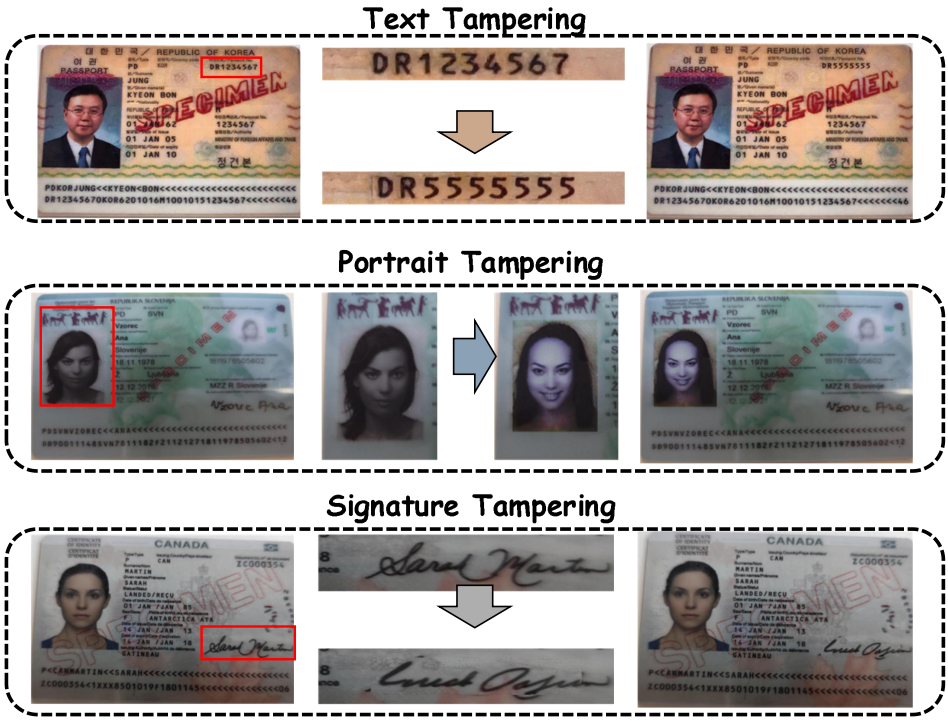


Figure 6: Representative tampering-generation process. From top to bottom, the examples show text, portrait, and signature tampering; each row presents the source document, selected field, edited field, and resulting document.

Rendering-order control. Rendering order can alter the pixel-level traces available to a forensic model even when the semantic edit and disturbance types are unchanged. The dataset uses tampering-before-VD for construction. Table 11 evaluates the alternative VD-before-tampering order.

The four states make it possible to observe benign VD responses and genuine tampering evidence within one benchmark. In particular, VD Real samples isolate false attribution of disturbance pixels as tampering, which is measured by the VD-specific FPR in the VDFM formulation. VD Tampered samples instead require the model to suppress disturbance responses without removing genuine tampering evidence, which is reflected by tampering-specific FNR. The overall composition therefore corresponds directly to the two operational VDFM errors defined in the main text.

Table 9: Main controlled protocol and evaluated rows.

Split	Used/evaluated rows	VD condition	Forgery relation	ID-type relation
Training	5,382 train + 597 val	Clean + Seen VD	-	Training pool
CID	2,395	Clean	Training-split	Cross ID type
SFD	1920	Seen VD	Cross forgery	Cross ID type
UVD	1904	Unseen VD	Source shift	Cross ID type

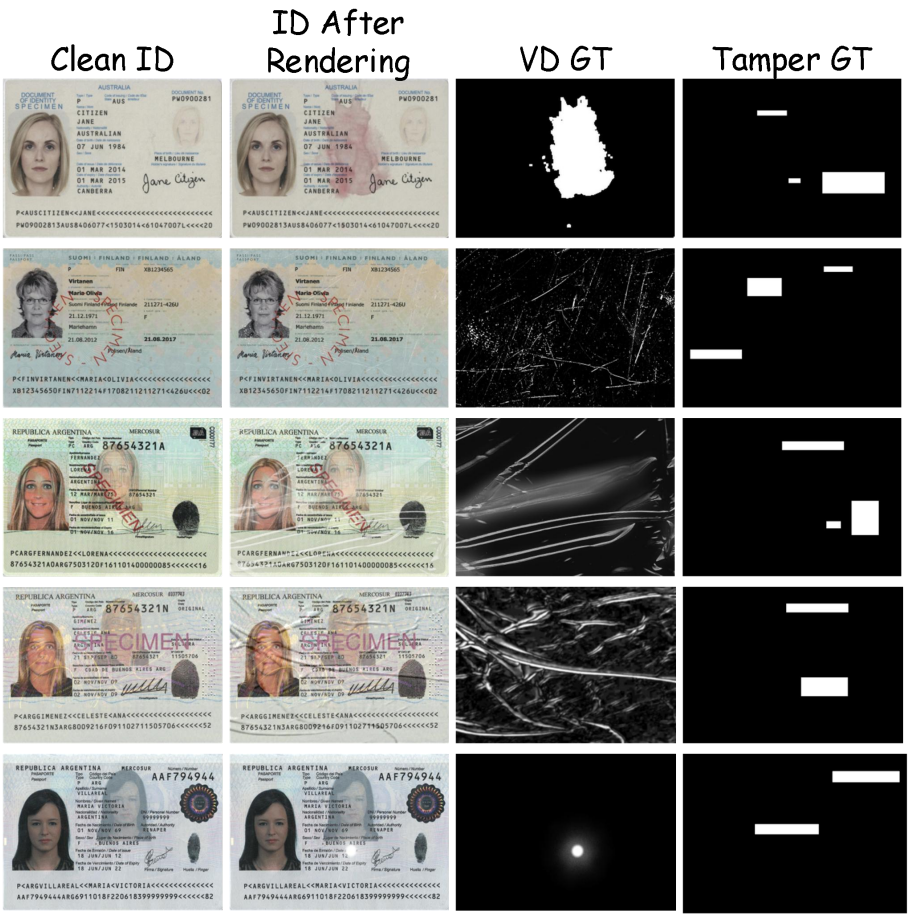


Figure 7: Examples of digitally rendered VD with independent VD and tampering annotations. Columns show the VD-free input ID, the ID after rendering, the VD ground-truth mask M_V, and the tampering ground-truth mask M_T.

A.3 Evaluation Protocols

Main protocol. The source-disjoint candidate pools contain 59,748 training, 5,590 CID, 30,114 SFD, and 30,604 UVD images, as summarized in Table 6. The v3 main protocol first splits 1,993 training groups at the identity/base level with fixed seed 20260711, producing 1,794 training and 199 validation bases. Each base contributes its clean record and two seen-material draws, yielding 5,382 training and 597 validation images. CID contains 2,395 clean records; SFD pairs 1,920 clean records with one seen-VD draw per base; UVD pairs 1,904 clean records with one unseen-VD draw per base. The seen set contains mosaic, flashlight, watermark, stain, scratch, and lamination ; the unseen set contains watermark, paperfold and seal.

CID evaluates tampering localization on clean images under cross-ID-type variation. SFD introduces seen VD while changing the forgery-source composition and document structures; UVD evaluates unseen VD under the corresponding cross-domain shift.

The material pools retain all traceable candidates, while each protocol manifest samples fixed quotas to control highly correlated variants and material-frequency imbalance. We split by identity/base before selecting any derivative, so clean, warp, and material variants inherit the same split. Within each stage, a stable SHA256 rank determined by the protocol, base, material type, and seed selects records under source, Real/Fake, and document-type quotas. The main training and controlled evaluation use digital-native and digitally rendered VD, whose controllable generation supports mechanism analysis and closely simulates physical appearances. Physical captures are reserved for external testing in Section 13.

Table 10: Editing-system composition of the controlled protocol.

Editing model	Training	CID	SFD	UVD
IDNet-copy-move	✓	✓	×	×
IDNet-crop-replace	✓	✓	×	×
doubao-seededit-3-0	✓	✓	×	×
flux-kontext-max	✓	✓	×	×
gemini-2.5-flash-image-preview	✓	✓	×	×
qwen-image-edit	✓	✓	×	×
BAGEL-7B-MoT	×	×	✓	✓
IDNet-inpaint-rewrite	×	×	✓	✓
OmniGen2	×	×	✓	✓
flux-kontext-pro	×	×	✓	✓
gpt-4o-image	×	×	✓	✓
step1xedit	×	×	✓	✓

Training and CID use six editing systems, while SFD and UVD use a disjoint set of six systems. This construction exposes the evaluation subsets to new editing sources and makes the forgery factor a controlled source-distribution shift.

Tampering is not restricted to text: portrait, signature, and mixed-field manipulations provide heterogeneous localization targets. Mixed-field samples additionally require preservation of multiple small or structurally separated tampering regions. These categories support the scene coverage required for ID-document tampering localization.

Visual Disturbances taxonomy. Our dataset is primarily constructed from digital-native disturbances and digitally rendered simulations of physical-domain VD, enabling controlled investigation of VDFM.

As shown in Figure 5, the main protocol assigns mosaic, flashlight, watermark, stain, scratch, and lamination to seen VD, paperfold and seal to unseen VD.

To avoid deviating from real-world physical scenarios, we further separate a cross-domain test set from the physically simulated samples in our dataset following Protocol in Section 13. This physical domain evaluation set contains 32 types of physical-capture VD and is used exclusively for physical-domain testing.

Table 11: Rendering-order control for clean-only forensic experts (pixel F1).

Model	Clean	T→VD	VD→T
ASCFormer (PR'25)	0.6876	0.5560	0.4532
TAFE-Net (AAAI'26)	0.4248	0.3349	0.4653
FFDN (ECCV'24)	0.6445	0.5618	0.3952
TTDMamba (IJCV'25)	0.5932	0.5390	0.4161
CLEAR-ID (TTDMamba)	0.7472	0.7291	0.6792

Table 11 shows that CLEAR-ID achieves the highest pixel-level F1 under both construction orders. Its advantage therefore persists when VD precedes tampering, supporting provenance-aware evidence representation beyond one rendering order. The results support the effectiveness of MAER and LSM in preserving tampering evidence while limiting disturbance-induced responses.

Table 12: Retained model-selection results under the main protocol (pixel-level F1, %).

Tampering expert	CID	SFD	UVD	Mean
ASCFormer (Luo et al., 2025)	69.90	68.52	68.04	68.82
TAFE-Net (Yi et al., 2026)	37.18	38.34	41.53	39.02
FFDN (Chen et al., 2024)	52.88	54.88	54.81	54.19
TTDMamba (Duan et al., 2025)	**74.48**	**68.34**	**71.00**	**71.27**

Cross-validation protocol. The v4 cross-validation protocol uses a second factor assignment: mosaic, face mosaic, paper, seal, and watermark are seen VD, while flashlight, dense watermark, stain, scratch, and wrinkle are unseen VD. The 4,596 training groups are split into 4,136 training and 460 validation bases; each contributes one clean record and two seen material draws, yielding 12,408 training and 1,380 validation images. Its canonical evaluation manifests contain 2,795 clean CID, 2,151 damage-only SFD, and 2,186 damage-only UVD records. The editing-model and document type partitions are also reassigned. Watermark may be grouped for display, while the originalwatermark type remains in the manifest. Section C.2 gives the corresponding results.

Clean-only protocol. The clean-only protocol uses clean images while preserving the identity-level split and fixed CID/SFD/UVD cells. It provides the clean expert initializations used by the corresponding baselines.

Physical-domain protocol.

We evaluate models trained under the main protocol on 1,717 images with physical VD and tampering. This tests whether provenance-aware evidence representation and constrained VD-guided modulation remain effective across acquisition and material conditions.

Table 13: Cross-domain tampering localization under physical VD (%).

Method	$F1_p$	IoU	AUC
ASCFormer(PR'25)	33.64	20.22	**88.78**
TAFE-Net(AAAI'26)	61.69	44.60	86.51
CD-SD(JVCIR'25)	62.40	45.35	<u>86.82</u>
DCLNet(SP'26)	13.77	7.39	74.94
FFDN(ECCV'24)	50.54	33.82	80.11
TTDMamba(IJCV'25)	<u>65.52</u>	<u>48.72</u>	74.83
CLEAR-ID	**68.06**	**51.59**	80.60

CLEAR-ID achieves the highest pixel-level F1 and IoU in Table 13, reaching 68.06% and 51.59%, respectively. These exceed TTDMamba by 2.54 and 2.87 percentage points. The gains support the use of VD priors to preserve tampering evidence under physical VD.

These results extend the effectiveness of MAER and LSM to physical VD: evidence representation by provenance and constrained modulation improve tampering localization across domains.

B MODELS

B.1 REPRODUCTION AND IMPLEMENTATION

Backbone and VD-prior interfaces. We select TTDMamba as the principal forensic expert because it has the highest mean pixel-level F1 (71.27%) among the retained standalone document-forensics experts under the controlled protocol. For each input crop x, the expert receives its native normalized RGB and JPEG-derived side input, whereas the VD localizer receives the corresponding uint8 RGB

image and a fixed VD prompt. The expert exposes its feature F_e^0 immediately before the native segmentation head, together with the original pixel-level logits M_o and image-level logit Y_o. A 1×1 projection maps F_e^0 to a 256-channel feature F_e at the expert feature resolution $h\times w$; the native heads still produce the preserved decisions M_o and Y_o. The independently trained Qwen3-VL-4B/SAM2.1 localizer projects its [SEG] token to $S_D \in \mathbb{R}^{256}$ and decodes a one-channel VD logit map G_v. It remains frozen during forensic calibration.

ADP and MAER. ADP aligns the semantic token S_D and spatial map G_v to the expert feature grid to form the VD prior D_P in Eq. 5. In the implementation, the spatial probability $p_D = \sigma(G_v)$ and its uncertainty $4p_D(1-p_D)$ are projected to 256 channels; the projected token is broadcast over the same grid. AF combines these with the expert anchor using $p_D \odot F_e$ and $p_D \odot \mathrm{broadcast}(S_D)$ as VD-conditioned evidence. A learned reliability gate weights the evidence before a residual update to F_e; its final projection is zero-initialized so the initial AF output equals the expert anchor. ER then uses a 1×1 mapping, normalization, and activation to form separate 256-channel disturbance and tampering seeds. The disturbance adapter gives F_d, while the tampering seed gives F_t in Eq. 6. Thus F_e, the shared AF output, F_d, and F_t denote different points in the computation and should not be interchanged.

VD anchors for F_d. The direct SegHead maps F_d to the one-channel spatial VD prediction M_{spa}. A spatially focused pooling of F_d, followed by a projection, produces a new [REC] token; it is distinct from the localizer's original [SEG] token. The frozen SAM2 decoder receives [REC] together with image features computed from the same crop and produces the one-channel semantic reconstruction M_{sem}. Both maps are aligned to the VD target resolution for the two constraints in Eq. 11. Gradients from these constraints reach the trainable F_d path and token projection, while the Stage-1 localizer and SAM2 decoder remain frozen.

LSM and residual outputs. LSM projects F_t from 256 channels to rank $r = 16$ by $V^\top$. A 1×1 projection of the disturbance path, followed by depthwise convolution and $\tanh$, produces the spatial gate s_d in the same rank-r space. The up-projection U gives $F_t' = F_t + U(s_d \odot V^\top F_t)$ as in Eq. 8; U is zero-initialized, so modulation initially preserves F_t. In the implementation, the gate uses the disturbance seed before its output adapter, which is the input corresponding to F_d in the framework diagram. A tamper adapter then maps F_t' to $\widetilde{F}_t$ for the output heads. The enhancement path generates a bounded signed pixel residual E from $\widetilde{F}_t$; the correction path uses the modulated tampering representation, disturbance context, expert feature, and preserved expert prediction to generate a further signed residual C. These constitute the figure-level pixel correction $C_m = E + C$, so the output keeps the expert skip connection $M_t = M_o + C_m$. The image-level correction C_y analogously updates the preserved expert logit as $Y_t = Y_o + C_y$. Pixel logits are aligned to the input resolution before localization, and sigmoid probabilities are thresholded at 0.5 for pixel- and image-level decisions.

Optimization and comparison implementations. Stage 1 trains the VD localizer with the text, BCE, and Dice terms in Eq. 10. In Stage 2, final tampering CE and Dice each have weight 1; auxiliary base and enhanced tampering terms have weights 0.25 and 0.5, the image-level classification term has weight 0.05, and the direct and SAM-reconstructed VD terms each have weight 0.2. The forensic expert, MAER, LSM, and output heads are optimized under the frozen prior for up to 100 epochs with early stopping; the checkpoint with the best validation pixel F1 is used for evaluation. Comparison methods retain their traceable implementations, native preprocessing, and model-specific outputs while being retrained or fine-tuned on the controlled protocol.

Official-code implementations. DTD, ASCFormer, FFDN, TAFE-Net, DCLNet, and SIDA-7B use their official repositories with local data adapters. **Third-party reproductions.** TruFor, Mesorch, SparseViT, and CAT-Net use the ForensicHub reproductions. **Project/AI-assisted implementations.** CD-SD, TTDMamba, and the TTDMamba-LISA cascade use project implementations; Codex assisted the TTDMamba and TTDMamba-LISA implementation/adaptation recorded in the AI use statement.

Table 14: Implementation provenance for the compared models. All reported comparison models are retrained or fine-tuned on the specified protocol.

Model	Code provenance	key setting	Retrained	input/output
TruFor	ForensicHub reproduction	pretrain initialization; 512 crop	✓	RGB / mask
Mesorch	ForensicHub reproduction	pretrain initialization; 512 crop	✓	RGB / mask
SparseViT	ForensicHub reproduction	pretrain initialization; 512 crop	✓	RGB / mask
CAT-Net	ForensicHub reproduction	pretrain initializationt	✓	RGB + JPEG luma DCT/Q-table / mask
DTD	Official repository	pretrain initialization; 512 input; DCT/JPEG preprocessing	✓	RGB / mask
FFDN	Official repository	pretrain initialization; 512 crop; DCT/JPEG preprocessing	✓	native inputs / mask
ASCFormer	Official repository	pretrain initialization; 512 crop	✓	RGB / mask
TAFE-Net	Official repository + adapter	pretrain initialization	✓	RGB / mask
CD-SD	Reimplementation around ASCFormer	pretrain initialization; 512 input	✓	RGB / mask
DCLNet	Official repository + adapter	pad/random crop 512; native augmentation	✓	RGB / mask
TTDMamba	Project reimplementation	pretrain initialization	✓	RGB + native side input / mask
TTDMamba + LISA	Project reimplementation	TTDMamba + Qwen3-VL-4B/SAM2.1	✓	prior + expert / mask
SIDA-7B	Official repository + adapter	pretrain initialization + SAM ViT-H	✓	image/text / mask
CLEAR-ID	Proposed method	TTDMamba + frozen Qwen3-VL-4B/SAM2.1 prior	✓	prior + expert / calibrated mask

B.2 Training Details

Table 15: Two-stage training configuration for CLEAR-ID.

Setting	Stage 1: VD prior/localizer	Stage 2: forensic calibration
Base model	Qwen3-VL-4B + SAM2.1 Hiera-Large	TTDMamba + frozen Stage-1 prior
Trainable modules	Qwen LoRA + projection/classifier + SAM decoder	TTDMamba + MAER + LSM
Input	Qwen dynamic + SAM 1024	512×512
Optimizer	AdamW	AdamW
LR / WD	LoRA 2×10^{-5}, heads 10^{-4} / 0.01	2×10^{-5} / 10^{-4}
Epochs	70	up to 100, early stopping
Batch	effective 8	effective 8
Scheduler	warmup + cosine	warmup + cosine
Precision	bf16	bf16
Key constraint	LoRA $r = 16$, $\alpha = 32$	LSM rank 16
Model selection	final Stage-1 checkpoint	best validation pixel F1

All baselines use the same data split while retaining their native preprocessing and model-specific losses. DTD and ASCFormer use their preserved project configurations without introducing unverified optimizer or learning-rate values. Retaining these method-specific recipes avoids forcing heterogeneous forensic experts into one artificial common configuration.

Table 16: Baseline training settings used for the reported comparisons.

Model/group	Optimizer	LR	Epochs	Batch	Key setting
TruFor / Mesorch / SparseViT / CAT-Net	AdamW	10^{-4}	100	8	ForensicHub adapters
FFDN	AdamW	5×10^{-5}	100	8	native DCT/JPEG pipeline
TAFE-Net	AdamW	10^{-4}	100	8	CE + Lovasz
CD-SD	AdamW	6×10^{-5}	100	8	implementation-specific objective
DCLNet	AdamW	10^{-4}	100	8	focal + 0.1 contrastive
TTDMamba	AdamW	10^{-4}	100	8	native segmentation loss
TTDMamba + LISA	AdamW	LoRA 2×10^{-5}; heads 10^{-4}	100	8	frozen/combined cascade
SIDA-7B	AdamW	2×10^{-5}	100	8	LoRA $r = 8$, $\alpha = 16$, dropout .05

B.3 Inference and Evaluation

CLEAR-ID inference. Stage-2 uses 512×512 inputs and validation stride 384. Overlapping logits are averaged by coverage count and thresholded at 0.5. The frozen VD prior receives uint8 RGB, whereas the forensic expert receives its native normalized input.

Visual prompt. Stage-1 training uses the following fixed conversation content:

```
System: You are an ID document image assistant. Identify visual
disturbances.
User: Segment the visual_disturbance_region in this ID document
image.
Target: The visual disturbance is {VD_TYPE}, and its region is
<|object_ref_start|>.
Anomaly status: {present|absent}.
```

Inference extracts the SEG-token hidden state; no free-form autoregressive generation is required for the prior.

Closed-source IDCD inference. GPT-5.4-nano, Gemini-2.5-flash, and Claude-sonnet-5 use the same localization prompt at temperature 0 with at most 1,200 completion tokens. Each returns a JSON object containing a tampered flag and normalized-1000 bounding boxes. The boxes are mapped to original-image coordinates, union-rasterized, and converted to a binary mask. The open-source comparison models are trained or fine-tuned on our dataset before this cross-domain evaluation.

Evaluation. Pixel-level F1 and IoU use dataset-level counts, AUC uses continuous scores, and image-level F1 uses the fixed 0.5 threshold. FPR measures visually disturbed pixels incorrectly predicted as tampered, while FNR measures tampered pixels incorrectly predicted as non-tampered. These definitions and their aggregation remain unchanged across the corresponding protocols.

C Experiments

C.1 Main and External-Domain Results

C.1.1 Main Protocol

This evaluation follows the controlled main protocol in Section A.3. Table 1 reports localization, while Table 17 reports image-level decisions; examining both levels distinguishes recovery of forensic evidence from a change that only moves the segmentation threshold.

Table 17: Complete image-level F1 results on the DisturbID test protocol (%).

Method	CID	SFD	UVD	Avg.
TruFor (CVPR'23)	54.23	83.01	85.81	74.35
DCLNet (SP'26)	44.14	67.45	69.20	60.26
FFDN (ECCV'24)	<u>88.24</u>	<u>91.38</u>	90.89	90.17
TTDMamba (IJCV'25)	87.85	91.34	<u>92.90</u>	<u>90.70</u>
SIDA-7B (CVPR'25)	60.11	80.55	82.72	74.46
CLEAR-ID (TTDMamba)	**90.47**	**93.16**	**93.73**	**92.45**

CLEAR-ID outperforms TTDMamba in CID, SFD, and UVD under the main protocol. The consistent gains under clean, seen-VD, and unseen-VD conditions support provenance-aware evidence representation and constrained modulation across the three evaluation settings.

Figure 8 complements the quantitative comparison. Across the displayed examples, baseline outputs contain missed regions, fragmented masks, or disturbance-induced false positives, whereas CLEAR-ID produces predictions more closely aligned with the tampering masks.

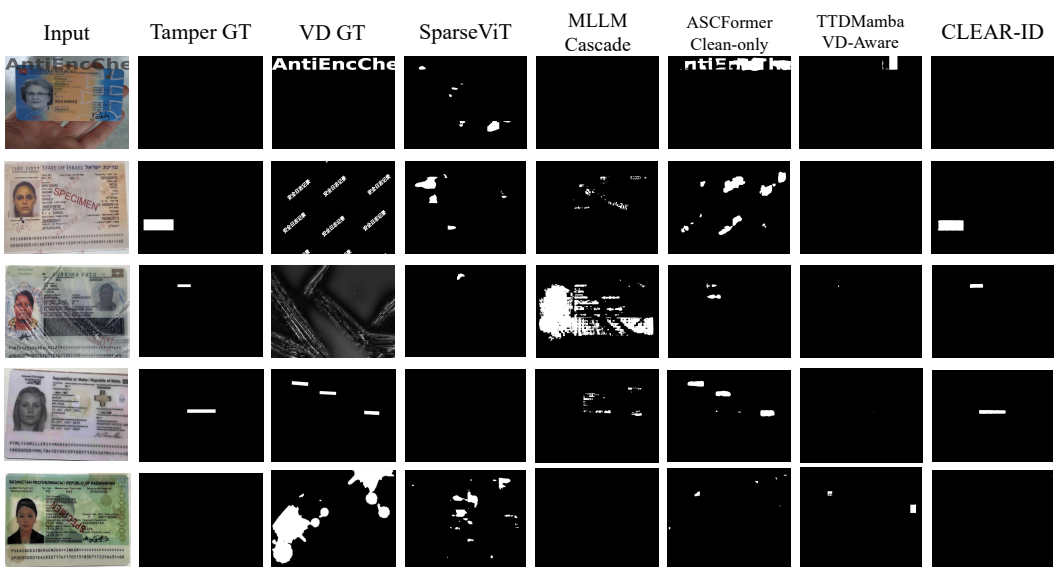


Figure 8: Qualitative tampering-localization comparison under visual disturbance. From left to right: VD input, tampering ground truth, VD GT, SparseViT, TTDMamba with LISA(Qwen3VL-4B), clean-only ASCFormer, VD-aware TTDMamba, and CLEAR-ID with the TTDMamba backbone. White regions denote masks or predictions.

C.1.2 Cross-Domain Evaluation on IDCD

On IDCD, CLEAR-ID achieves 58.00% pixel-level F1 on clean images and 57.04% on unseen VD, exceeding TTDMamba in both conditions (Table 18). The gains show that provenance-aware representation and constrained VD-guided modulation remain effective across domains. Closed-source MLLMs score below 8% on average under box-based localization, where semantic predictions and coarse boxes do not recover fine tampering boundaries. TextShield-R1 (Qu et al., 2026) also yields very low pixel-level F1. Its text-oriented reasoning and OCR-based box refinement may not transfer directly to dense localization across heterogeneous ID fields; the aggregate scores alone do not isolate the cause.

The IDCD cross-domain results support CLEAR-ID's effectiveness on a separately collected dataset. The closed-source MLLM results show that direct semantic or box predictions provide insufficient pixel-level localization in this setting. CLEAR-ID instead uses MLLM semantics as a VD prior and preserves the forensic expert's spatially precise evidence.

Table 18: Cross-domain pixel-level F1 on IDCD (%).

Method	Clean	Unseen VD	Avg.
FFDN (ECCV'24)	22.70	21.70	22.20
ASCFormer (PR'25)	24.29	23.11	23.70
TAFE-Net (AAAI'26)	26.19	27.80	26.99
DCLNet (SP'26)	24.59	25.52	25.05
TTDMamba(IJCV'25)	54.66	54.11	54.38
GPT-5.4-nano	8.02	7.38	7.70
Gemini-2.5-Flash	2.73	7.64	5.19
Claude-Sonnet-5	1.05	10.53	5.79
TextShield-R1 (AAAI'26)	0.04	0.07	0.06
SIDA-7B (CVPR'25)	16.64	15.13	15.89
CLEAR-ID	**58.00**	**57.04**	**57.52**

C.2 Cross-Validation

This cross-validation uses the alternative seen/unseen VD assignment, editing-model partition, and document-type partition specified in Section 13. It tests whether the VDFM-oriented mechanism remains effective when these data factors are jointly reassigned.

Table 19: Pixel-level F1 under the cross-validation protocol (%).

Method	CID	SFD	UVD
TruFor (CVPR'23)	5.61	4.73	4.89
SparseViT (AAAI'25)	9.21	8.46	8.91
DCLNet (SP'26)	15.56	11.98	13.88
DTD (CVPR'23)	7.00	6.80	3.51
TAFE-Net (AAAI'26)	22.31	17.54	19.24
FFDN (ECCV'24)	51.15	50.43	47.40
TTDMamba (IJCV'25)	55.12	53.15	52.58
CLEAR-ID	**56.71**	**53.55**	**55.83**

Relative to TTDMamba, CLEAR-ID remains higher by 1.59, 0.40, and 3.25 points on CID, SFD, and UVD, respectively. The consistent direction of the gains shows that provenance-aware evidence representation and constrained modulation remain effective under the alternative factor assignment.

C.3 MAER and Module Ablation Analysis

The main-protocol prior ablation in Table 2 asks what conditioning information MAER needs. SAM spatial output and spatial VD-GT reach 64.77% and 66.10% UVD F1, whereas the trained semantic–spatial prior reaches 74.97%; an untrained semantic interface falls to 62.22% with 43.24% FNR. Accurate spatial support alone and an untrained semantic interface are therefore insufficient: useful MAER conditioning requires learned VD semantic information together with spatial support. This evidence supports the main-text conclusion that semantic–spatial VD knowledge provides informative conditioning for provenance-aware evidence representation.

Table 20: Image-level module ablation results for CLEAR-ID ($F1_i$, %).

Method	SFD	UVD	Avg.
TTDMamba (Clean-only)	92.08	93.30	92.69
TTDMamba (Oracle GT)	88.55	88.44	88.50
TTDMamba (VD-aware)	91.34	92.90	92.12
MAER only	93.14	**94.00**	**93.57**
MAER + LSM (CLEAR-ID)	**93.16**	93.73	93.45

The reconstruction and anchoring ablation below uses the same protocol to ask how this information is organized inside MAER.

Table 21: Fine-grained MAER component ablation (pixel-level F1, %).

Configuration	CID	SFD	UVD
No spatial anchoring & train SAM decoder	69.88	63.99	64.12
No spatial anchoring	74.47	72.42	71.67
MAER only	75.45	72.89	73.68
CLEAR-ID	**75.64**	**74.37**	**74.97**

Removing spatial anchoring lowers UVD F1 from 73.68% for MAER only to 71.67%; additionally training the SAM decoder without anchoring lowers it further to 64.12%. Decoder capacity therefore does not replace the representation constraint. The prior study identifies what information MAER needs, while this ablation identifies how that information organizes disturbance-oriented evidence. Together, the results support MAER's role in provenance-aware evidence representation; LSM is analyzed separately in Section C.4 as the constraint on subsequent VD-guided modulation.

C.4 LSM Analysis

This study uses the main protocol to measure how the LSM subspace rank changes the capacity of prior-conditioned modulation.

Table 22: LSM rank sensitivity (pixel-level F1, %).

Rank	CID	SFD	UVD	Overall
8	73.90	72.42	73.37	73.23
16 (CLEAR-ID)	**75.64**	**74.37**	**74.97**	**74.99**
64	71.04	62.95	64.53	66.17
128	73.04	69.76	67.70	70.17
256	70.96	65.62	64.54	67.04

Rank 8 reaches 73.23% overall, below rank 16 at 74.99%, indicating that an overly small subspace restricts useful correction capacity. Increasing the rank to 64, 128, or 256 yields 66.17%, 70.17%, and 67.04%, respectively, rather than a sustained improvement. This is consistent with the main-text role of low-rank modulation as a constraint on how many prior-conditioned directions can alter the forensic representation. Rank 16 provides the best observed trade-off between correction capacity and preservation of tampering evidence.

Table 23: Source values for the VD-intensity curves in Figure 4 (pixel F1, %).

Bin	FFDN	TTDMamba	CLEAR-ID (MAER-only)	CLEAR-ID	Qwen4B+SAM2	LLaVA1.5+SAM1
Clean	52.82	69.40	72.69	74.19	0.00	0.00
J_0	19.16	60.02	45.90	51.48	76.97	70.52
J_1	52.27	61.12	67.91	66.69	87.26	84.33
J_2	52.18	65.60	68.78	72.14	65.53	66.16
J_3	51.14	71.37	74.46	74.87	66.38	51.75
J_4	59.31	74.29	79.66	79.70	61.64	49.46
J_5	50.47	64.27	68.92	70.45	56.54	50.40
J_6	48.57	61.16	66.86	68.21	56.13	44.71

C.5 VD-Intensity Analysis

This analysis stratifies the main-protocol predictions by the recorded intensity score J. CLEAR-ID exceeds TTDMamba in every nonzero bin J_1–J_6: representative gains are 66.69% versus 61.12% at J_1, 79.70% versus 74.29% at J_4, and 68.21% versus 61.16% at J_6. J_0 is the exception.

The curves are non-monotonic, and J is not a causal intervention because bin composition and VD type also vary. The consistent gains across nonzero VD bins are evidence for the existing VDFM formulation and for MAER/LSM controlling VD-induced responses while preserving tampering evidence.

C.6 Initialization and Backbone Compatibility

We vary the expert backbone and its initialization to examine whether MAER and LSM can use different forensic representations. Clean-init starts from an expert trained on clean ID images; pretrain-init starts from the backbone's pretrained checkpoint. This separates the effect of the available forensic evidence from the subsequent VD-guided representation and modulation.

With clean initialization, CLEAR-ID improves all three subsets for ASCFormer, TAFE-Net, and TTDMamba, and improves CID and SFD for FFDN. With pretrained initialization, it improves all three subsets for FFDN and TTDMamba. These results show that the framework can use multiple expert representations, while the most effective initialization depends on the backbone.

C.7 Robustness

This robustness evaluation applies blur, noise, JPEG compression, and resizing as post-processing operations to models trained under the main protocol. The decision threshold is not recalibrated, and these operations are not assigned as newly defined VD categories.

CLEAR-ID exceeds TTDMamba for blur kernel 3 (48.08% versus 46.22%), noise $\sigma = 36$ (26.19% versus 24.06%), JPEG quality 8 (23.50% versus 22.33%), and resize 0.80 (52.99% versus 51.46%). CLEAR-ID improves pixel-level F1 at four of the five resize scales, supporting the persistence of useful tampering evidence across spatial resampling. It also leads in pixel-level F1 at the strongest tested noise and JPEG settings. One possible explanation is that these severe operations introduce VD-like artifacts, for which provenance-aware representation and constrained modulation remain useful.

C.8 Efficiency

This profile uses an RTX 4090, batch size 1, 512×512 input, bfloat16 autocast, ten warm-up iterations, and 50 timed iterations.

The complete CLEAR-ID model requires 119.56 ms per 512×512 input (8.36 FPS), with 4,747.48M parameters and 9,960.44 MiB peak memory. Both the VD localizer and the forensic path are included. The VD localizer accounts for most of the cost, while MAER and LSM use its semantic and spatial priors to improve localization. Compared with SIDA-7B in the same profile, CLEAR-ID uses fewer parameters and less peak memory and has lower latency. The full-model cost therefore describes the computational trade-off of introducing a semantic VD prior for provenance-aware forensic analysis.

Table 24: Initialization and backbone compatibility of CLEAR-ID ($F1_p$, %).

	FFDN			ASCformer			TAFE-Net			TTDMamba		
init	CID	SFD	UVD	CID	SFD	UVD	CID	SFD	UVD	CID	SFD	UVD
Clean-only	66.88	62.00	59.38	73.06	60.70	58.71	39.20	43.02	38.73	63.39	57.83	55.75
Clean-init	73.65	65.16	**71.33**	71.40	66.26	67.38	35.17	33.95	35.79	69.33	62.17	62.30
Clean-init + CLEAR-ID	**74.65**	**67.79**	70.31	**76.69**	**71.65**	**72.14**	**68.98**	**70.01**	**67.77**	71.97	63.49	62.59
Pretrain-init	52.88	54.88	54.81	69.90	68.52	65.16	37.18	38.34	41.53	74.48	68.34	71.00
Pretrain-init + CLEAR-ID	68.36	63.16	62.16	59.25	55.52	55.51	19.51	13.12	15.26	**75.64**	**74.37**	**74.97**

Table 25: Post-processing robustness (pixel F1/AUC, %) without threshold recalibration. Best results are bold and second-best results are underlined.

Transform	Model	Severity-specific F1/AUC				
Gaussian blur (kernel)		3	5	9	13	17
	ASCFormer	36.73/74.05	28.32/69.49	13.70/62.07	5.66/56.25	1.26/52.51
	TAFE-Net	27.61/**83.83**	25.89/**79.41**	18.50/**69.39**	**8.65**/**60.36**	**3.94**/**55.04**
	FFDN	12.38/54.09	10.29/53.40	6.14/52.04	3.48/51.18	1.97/50.69
	TTDMamba	46.22/72.85	**35.40**/66.86	**19.73**/59.23	8.02/53.89	2.01/51.12
	CLEAR-ID	**48.08**/73.78	34.14/65.81	15.83/57.02	6.47/52.97	1.73/50.91
Gaussian noise (σ)		5	10	18	26	36
	ASCFormer	44.40/83.25	40.96/82.30	33.57/80.53	27.69/79.26	21.91/78.79
	TAFE-Net	29.34/**88.98**	26.79/**87.62**	21.89/**84.77**	19.41/**82.66**	17.27/**80.55**
	FFDN	13.88/54.63	11.03/53.58	7.56/52.42	5.39/51.71	3.46/51.12
	TTDMamba	**53.12**/79.61	**47.98**/76.12	**39.03**/71.93	**31.68**/67.59	24.06/62.82
	CLEAR-ID	51.68/79.64	46.56/76.25	36.73/72.49	31.34/68.84	**26.19**/64.56
JPEG (quality)		70	50	30	15	8
	ASCFormer	45.64/85.32	42.14/82.61	39.42/82.99	28.43/76.80	18.01/71.04
	TAFE-Net	26.71/**90.53**	23.48/**88.78**	19.79/**86.62**	15.15/**81.51**	11.87/**75.73**
	FFDN	15.18/55.09	13.74/54.67	13.28/54.60	9.49/53.42	5.27/51.83
	TTDMamba	**52.25**/84.25	**46.33**/81.89	**41.94**/80.22	**32.39**/73.18	22.33/63.94
	CLEAR-ID	48.83/86.60	45.17/84.35	39.23/82.74	30.56/76.79	**23.50**/68.79
Resize (scale)		0.80	0.65	0.50	0.35	0.22
	ASCFormer	41.76/77.97	39.36/76.11	33.62/72.93	28.08/70.78	17.14/66.30
	TAFE-Net	28.62/**86.43**	28.78/**84.96**	27.78/**82.16**	24.37/**79.47**	12.62/**70.68**
	FFDN	13.17/54.37	12.32/54.07	10.79/53.56	8.39/52.78	4.79/51.62
	TTDMamba	51.46/76.41	48.43/74.07	40.57/69.55	**32.29**/65.91	18.74/59.79
	CLEAR-ID	**52.99**/78.20	**50.29**/75.45	**42.07**/70.22	31.72/65.80	**20.48**/61.09

Table 26: Full-method resource accounting and recorded efficiency.

Component / deployment	Params (M)	GFLOPs	Latency (ms)	FPS	Peak mem. (MiB)
ASCFormer (PR'25)	81.06	764.77	39.89	25.98	588.04
TTDMamba (IJCV'25)	61.11	163.49	19.50	51.34	533.12
CLEAR-ID forensic expert	66.20	598.56	34.89	28.91	860.10
SIDA-7B (CVPR'25)	7,717.61	10,556.66	197.87	5.05	18,277.73
CLEAR-ID VD detector	4,681.28	4,647.05	88.58	11.29	9,398.01
CLEAR-ID full	4,747.48	5,035.36	119.56	8.36	9,960.44