symmetry

MDPI

Article

Breaking the Sign Symmetry of Attention: A Conflict-Aware Vision–Language Fusion Framework for Privacy-Sensitive Information Detection

Ming Lian [1], Yuanyuan Li [2,*] and Teng Li [1]

[1] School of Artificial Intelligence, Anhui University, Hefei 230039, China; lianming@mail.ustc.edu.cn (M.L.); tenglwy@gmail.com (T.L.)
[2] School of Electronics and Information Engineering, Anhui University, Hefei 230039, China
* Correspondence: liyy@ahu.edu.cn

Abstract

As image data are shared ever more openly, they increasingly leak privacy-sensitive information, yet existing detectors seldom model how the text embedded in an image relates to its visual content and tend to fail precisely when the two modalities disagree. We note that the conventional softmax attention used for multimodal fusion carries an implicit sign symmetry: Every source token contributes only additively, so conflicting evidence is averaged away rather than resolved. We propose a symmetric dual-source fusion framework whose decoder deliberately breaks this sign symmetry. Image and text are first encoded by a Swin Transformer and a policy knowledge-enhanced BERT (KL-BERT) and projected into a common space to form a permutation-symmetric dual-source memory. A Signed Cross-attention Auto-compressing Decoder (SCAD) then fuses the two sources through an attention map that factorizes into a sign-symmetric (even) magnitude term and a sign-antisymmetric (odd) polarity term, allowing the model to either reinforce or actively subtract cross-modal evidence. Experiments on a self-constructed privacy-sensitive image dataset show that the proposed method attains an accuracy of 97.69% and an F1-score of 96.70%, with its largest gains on the face category, the most conflict-prone class in our data, suggesting that controlled symmetry breaking is an effective principle for cross-modal fusion.

Keywords: symmetry breaking; signed attention; cross-modal fusion; vision–language model; sensitive information detection; Swin Transformer; BERT; self-attention

Academic Editors: Vasilis K. Oikonomou and Zhixun Su

Received: 29 June 2026
Revised: 2 August 2026
Accepted: 8 August 2026
Published: 11 August 2026

1. Introduction

The internet has transformed how information circulates, and what now travels across networks is no longer text alone but images in enormous variety. Once such images are shared or released openly, they frequently expose privacy-sensitive material (names, dates of birth, identity-card numbers, home addresses, telephone numbers, faces, and credentials), and the resulting attack surface keeps widening. In response, governments have rolled out a succession of data-security laws, regulations, and technical standards that spell out concrete protection requirements. Detecting sensitive information in images before they are shared has therefore become a pressing concern [1,2].

In practice, sensitive content usually travels as text and imagery together, so treating either carrier in isolation no longer meets real needs. The words embedded inside an image and the high-level meaning of its visual content tend to complement one another, and it

is precisely their relationship that decides whether detection succeeds. This argues for a multimodal model that fuses the two streams. Any such fusion, however, runs into a structural question that is at bottom one of symmetry: How ought two heterogeneous sources be combined, and how should the model act when they agree as opposed to when they conflict?

We organize our treatment of this question around symmetry, which supplies the conceptual backbone of the paper. At the architectural level, a faithful multimodal model ought to handle its two sources symmetrically; the image and text encoders are mapped into a shared representation space and gathered into one memory bank, so that neither modality is favored by construction and the fusion does not depend on the order in which the sources are presented. A second, subtler symmetry lives in the fusion operator itself. The softmax attention at the heart of almost every modern fusion module forces each attention weight to be non-negative, so a source token can only ever add evidence to the prediction, a property we call the sign symmetry of attention. Under this constraint, when the visual and textual streams emit contradictory signals, the model can do no better than merely average them, which blends correct evidence with incorrect evidence and leaves the underlying conflict unresolved.

The central claim of this paper is that resolving cross-modal conflict calls for deliberately breaking that sign symmetry. We introduce a fusion decoder that permits negative attention weights, so a source token can subtract or suppress conflicting evidence instead of only reinforcing it. What makes this tractable is that the signed attention splits cleanly into two parts of opposite symmetry: a sign-symmetric (even) magnitude term governing how strongly a token participates, and a sign-antisymmetric (odd) polarity term governing whether it supports or opposes the prediction. This even–odd decomposition is, to our knowledge, a new lens on conflict-aware multimodal fusion, and it is the formal statement of the symmetry breaking. Recent work in language modeling points in the same direction: signed attention weights [3] and differential attention [4] both report that relaxing the non-negativity constraint improves expressiveness. Both operate within a single sequence, however, whereas the conflict we target arises between two heterogeneous sources.

Concretely, we extract image features with a Swin Transformer (SWT) and text features (recovered from the image by OCR) with a policy knowledge-enhanced BERT, denoted KL-BERT. The two encoders are projected to a common dimension and concatenated into a symmetric dual-source memory. A Signed Cross-attention Auto-compressing Decoder (SCAD) then performs the symmetry-breaking fusion through a cross-depth query composer, the signed dual-source cross-attention described above, and an auto-compressing read-out that aggregates evidence from all decoder layers. The main contributions of this paper are as follows:

- We identify a sign symmetry in the softmax attention used for multimodal fusion: because every attention weight is non-negative, agreeing and conflicting evidence can only be averaged, never subtracted, which is why the fusion breaks down when the two modalities disagree. We recast conflict-aware fusion as a problem of controlled symmetry breaking.
- We propose the SCAD decoder, whose signed cross-attention factorizes each attention weight into a sign-symmetric (even) magnitude and a sign-antisymmetric (odd) polarity, so the model can either reinforce or actively suppress cross-modal evidence while operating on a permutation-symmetric dual-source memory that treats the two modalities on an equal footing.
- We build KL-BERT, a Chinese BERT continue-pretrained on a curated corpus of data-security policies and regulations, which gives the text branch a vocabulary adapted to the sensitivity domain.

- We construct a privacy-sensitive image dataset and, through a component ablation, a direct comparison against standard softmax cross-attention under identical encoders, and a decoder-depth study, show that the signed formulation is what drives the improvement, and that its gains are largest precisely where the two modalities most often disagree.

The remainder of this paper is organized as follows: Section 2 reviews related work. Section 3 presents the symmetric dual-source framework, the SCAD decoder, and a symmetry analysis of the signed attention. Section 4 reports the experimental results, Section 5 discusses where the gains originate together with the computational cost and limitations, and Section 6 concludes.

2. Related Work

Early sensitive-information detection methods were mainly based on Chinese sensitive-word recognition. Commonly used algorithms include single-pattern matching, such as the Knuth–Morris–Pratt algorithm [5], and multi-pattern matching, such as the Aho–Corasick automaton [6], which determine the presence of sensitive information by comparing the text with a sensitive-word lexicon. Xu et al. [7] proposed the SW-LDA model, a sensitive-word weighted LDA for topic recognition of network sensitive information. Zhou et al. [1] studied the improvement of mass multimedia information filtering under multi-data-inflow conditions. Ding et al. [2] proposed a BERT-based automatic sensitive-information detection method trained on a Chinese corpus. Liu et al. [8] proposed a graph-convolutional-network-based sensitive-information detection algorithm. Zhang et al. [9] proposed a dual-channel semantic enhancement-based convolutional neural network for text classification. Zhang et al. [10] proposed a character-level convolutional network for text classification. These methods improve the efficiency and accuracy of sensitive-word detection, but simple sensitive-word matching can no longer satisfy the requirements of network security, and newly emerging sensitive content requires other detection methods.

With the development of computer vision, the detection of sensitive content has expanded from the text domain to the image domain. Convolutional neural networks (CNNs) [11] are trained on large amounts of data to exploit the input features, and architectures such as AlexNet [12], VGGNet [13], and ResNet [14] have driven the rapid development of deep learning in image classification. Guo et al. [15] proposed a multi-information detection model for violent images built on YOLOv3-SPP [16] and DenseNet [17]. Hu et al. [18] proposed squeeze-and-excitation networks that adaptively recalibrate channel-wise feature responses to strengthen representational power. Zhang et al. [19] proposed a deep relational reasoning graph network for arbitrary-shape text detection. Liao et al. [20] proposed TextBoxes++, a single-shot-oriented scene text detector in the SSD [21] family, often paired with lightweight backbones such as MobileNetV2 [22], and Zhang et al. [23] proposed S3FD, a single-shot scale-invariant face detector.

These approaches, however, either concentrate solely on sharpening image recognition or decide sensitivity purely from the text read off the image, and in doing so they overlook the deeper semantics that emerge when picture and text appear together, and above all what should happen when the two modalities disagree. Transformer-based fusion [24] together with vision transformers such as ViT [25] and the Swin Transformer [26] offer strong encoders, but the ordinary attention used to combine modalities inherits the non-negativity (the sign symmetry) examined in Section 3, and therefore cannot actively suppress contradictory evidence. Our signed cross-attention is related to, but distinct from, several existing attention variants. Gated attention rescales non-negative weights but never changes their sign; signed graph attention assigns polarity to fixed graph edges rather than to learned

query–key interactions; and inhibitory or contrastive fusion schemes act on the training objective rather than on the attention operator itself. In contrast, SCAD produces the polarity per query–key pair at inference through an odd sign term, so a source token can be actively subtracted rather than merely down-weighted. The relationship between image and text information, and the principled handling of their disagreement, is the focus of this paper.

Since 2023 the vision–language literature has moved quickly, and two strands are directly relevant here. The first is large-scale vision–language pre-training. BLIP-2 [27] bridges frozen image encoders and large language models through a lightweight querying transformer; SigLIP [28] replaces the softmax contrastive objective of CLIP [29] with a pairwise sigmoid loss that scales more gracefully; and InternVL [30] scales the visual foundation model to billions of parameters and aligns it with an LLM for generic visual–linguistic tasks. Li and Tang [31] survey the alignment and fusion strategies these models rely on. These systems are optimized for open-ended captioning, retrieval, and instruction following rather than for fixed-label detection, and Baia et al. [32] report that on image-privacy classification specifically, zero-shot vision–language models do not uniformly outperform compact task-specialized detectors. This is why we retain a closed-label, task-adapted design and benchmark against the CLIP-style contrastive baseline, which is the member of this family that transfers directly to our label set.

The second strand concerns the attention operator itself, and it has converged on the same constraint we identify. Lv et al. [3] propose Cog Attention, which multiplies the sign of the query–key inner product by the softmax of its absolute value, and show for language modeling and diffusion that permitting negative weights raises expressiveness and mitigates representational collapse. Ye et al. [4] reach a related conclusion from a different direction with the Differential Transformer, which subtracts one softmax map from another so that shared attention noise cancels. We regard these results as independent evidence that the non-negativity of softmax attention is a genuine limitation rather than an incidental one. Our contribution is complementary and specific in three respects: The signed operator is applied across two heterogeneous sources over a permutation-symmetric dual-source memory rather than within one sequence; we give an explicit even–odd symmetry decomposition of the signed weight into a magnitude and a polarity term; and we tie the mechanism to a concrete failure mode, cross-modal conflict, and measure it directly by ablating the negative branch alone.

3. Materials and Methods

The overall pipeline of the proposed detector is shown in Figure 1, and the detailed architecture is shown in Figure 2. The text inside an image is first recovered by optical character recognition (OCR); the textual data are then processed by KL-BERT and the visual data by the Swin Transformer. The OCR step uses a standard scene text recognition engine applied to each image; because the quality of the recovered text bounds the text branch, we quantify the downstream effect of OCR through the "w/o OCR" ablation.

We specify the OCR stage precisely, since the recovered text bounds everything downstream. Detection and recognition are performed by the open-source PaddleOCR engine, a CRNN-style pipeline [33] run with its default Chinese model; each image is passed at its native resolution, the detected boxes are sorted in reading order, and the recognized strings are concatenated into a single sequence that is truncated to 128 WordPiece tokens. The corpus carries no ground-truth transcriptions, so we do not report an on-dataset recognition rate; what we characterize instead is where the engine fails and how the failures propagate. Errors concentrate on low-resolution screen photographs with moiré patterns, on strongly perspective-distorted identity documents, and on decorative or handwritten

fonts, and they take two forms with different consequences: Missed text weakens the text branch and pushes the model toward its visual backbone, whereas hallucinated text, typically fragments read out of watermarks, background signage, or interface chrome, actively injects misleading evidence. The second failure mode is precisely the one the signed attention is designed to absorb, since a spurious token can be assigned a negative weight and subtracted rather than averaged in. The two encoder outputs are merged in the SCAD decoder, which integrates the features extracted from the two modalities, and the fused representation is used by the classification head to produce the prediction.

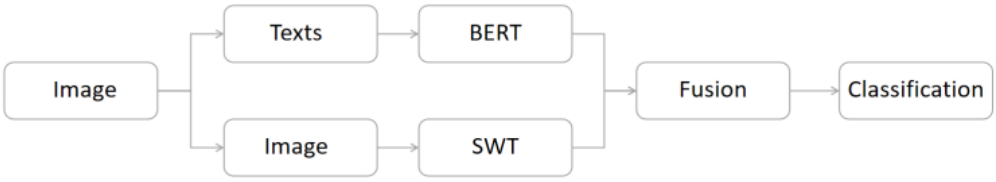

Figure 1. General pipeline of image sensitive-information detection. Text is extracted from the image and processed by KL-BERT, while the image is processed by the Swin Transformer (SWT). The two outputs are combined in a symmetric fusion block and used by the classification module.

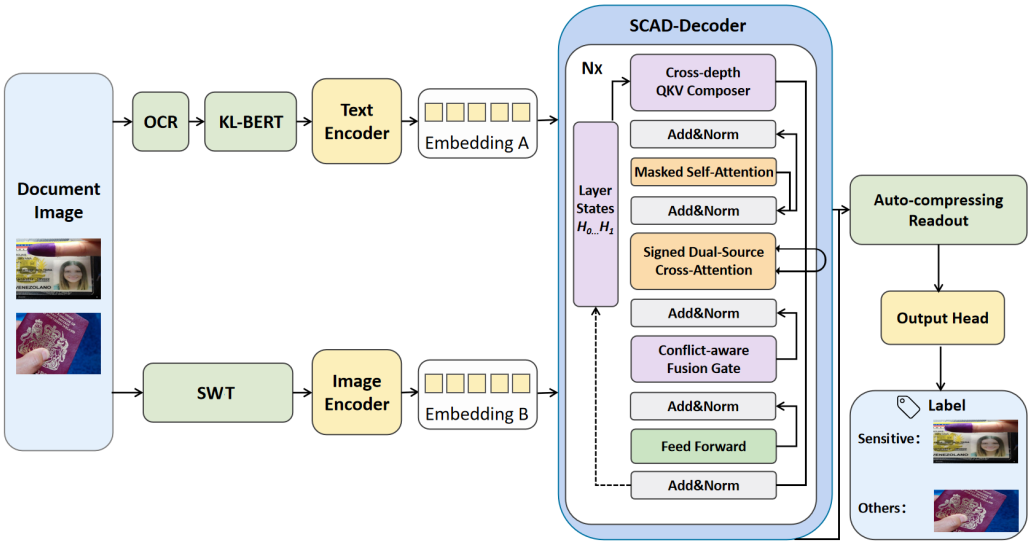

Figure 2. Architecture of the proposed model; each module (OCR, the KL-BERT and Swin-Transformer encoders, the permutation-symmetric dual-source memory, and the SCAD decoder) is described in the main text of Section 3.

3.1. Symmetric Dual-Source Design

The framework is guided throughout by a single principle: the two modalities are handled symmetrically. A source-specific projection sends each encoder into a shared hidden space, a learnable source identifier is attached, and the two memories are then concatenated into one bank. The result is permutation-symmetric: exchanging the two sources along with their identifiers leaves both the fused memory and the prediction untouched. No modality is favoured in advance; whatever asymmetry ends up mattering is learned rather than wired in. The symmetry breaking that actually resolves conflict comes later and lives only inside the attention weights, as Section 3.4.4 explains.

3.2. Feature Extraction

3.2.1. Text Feature Extraction (KL-BERT)

Pre-trained language models such as BERT [34] do well on generic tasks, but the corpora they are built on look little like the data-security and policy setting, where normative phrases such as “involves personal information” or “sensitive field” recur constantly. To

close this gap we continue pre-training a Chinese BERT with a policy knowledge objective. Our corpus gathers 52 policy and regulation documents spanning eight regulated industry sectors together with cross-sector general data-security policies, and nearly 4700 sensitive personal information text samples drawn from 52 typical application scenarios, which we denote $D_{\text{policy}} = \{S_1, \ldots, S_K\}$. Two distinct collections of 52 items are involved. The 52 policy and regulation documents comprise the data security policies of eight regulated sectors, (i) government affairs, (ii) industry and information technology, (iii) telecommunications, (iv) finance, (v) transportation, (vi) energy, (vii) healthcare, and (viii) education, together with cross-sector data security policies that apply to all industries. The 52 typical application scenarios are the consumer- and service-facing settings in which personal information is produced, stored, or exchanged, and from which the sensitive-text samples were drawn: map navigation, ride-hailing, instant messaging, online communities, online payment, online shopping, food delivery, mail and express delivery, transport ticketing, matchmaking, job recruitment, online lending, housing rental and sale, used car trading, online medical registration, travel services, hotel services, online gaming, online education, local life services, women's health, car services, investment and wealth management, mobile banking, e-mail and cloud storage, remote conferencing, live streaming, online video and music, short video, news and information, sports and fitness, web browsing, input methods, security management, e-books, photo shooting and editing, app stores, utility tools, show ticketing, telephone and cable television subscription, telecom service usage, security surveillance, utility bill payment, customer service, campus services, smart home, autonomous driving, remote diagnosis, virtual reality, online voting, home repair, and pet care. Each scenario is characterized by the personal-information fields it exposes (a payment scenario, for instance, involves bank card and order fields, while ride-hailing involves location traces and license plate numbers), and the sensitive-text samples are instances of those fields in context, so the corpus is aligned with the regulatory vocabulary that signals sensitivity in practice. On top of bert-base-chinese we use the masked language-modeling (MLM) objective: Given a policy text $S \in D_{\text{policy}}$ with masked token set M,

$$\mathcal{L}_{\text{MLM}} = -\,\mathbb{E}_{S \sim D_{\text{policy}}} \sum_{i \in M} \log P\big(x_i \mid S_i'\big), \tag{1}$$

where x_i is the i-th masked token, and S_i' the context with that token removed. We denote the resulting encoder KL-BERT and use its global representation

$$K_{\text{BERT}} = \text{KL-BERT}(X) \tag{2}$$

as the text feature. A fully connected layer is appended before the output to reduce the feature dimension and match it to the image features. We deliberately use this policy-adapted encoder rather than a general Chinese BERT because in the data security setting, sensitivity is signaled by regulatory phrasing; the ablation in Section 4 confirms the benefit, as removing the policy pre-training lowers the F1-score from 96.70% to 94.65%.

3.2.2. Image Feature Extraction

The Swin Transformer (SWT) [26] introduces hierarchical feature representations and a shifted-window attention mechanism, solving the excessive complexity that ViT [25] incurs on high-resolution images. An input image is split into 4×4 patches, each projected into a C-dimensional space; four stages progressively halve the spatial size and double the channels. The core mechanism alternates Window Multi-Head Self-Attention (W-MSA) and Shifted-Window Multi-Head Self-Attention (SW-MSA): W-MSA computes attention within 7×7 windows, reducing the complexity from $\mathcal{O}((HW)^2)$ to $\mathcal{O}(M^2HW)$ with window size M, while SW-MSA shifts the windows by $M/2$ to allow cross-window interaction.

A relative position bias replaces absolute positional encoding for robustness to input size. We adopt the Swin Transformer as the visual backbone because its locality-aware, hierarchical attention suits privacy-sensitive imagery, where the decisive cue is often a small region (a face, an identity-card field, or a few license plate characters) rather than a global scene property.

3.3. Input Representation and Feature Fusion

The SCAD decoder receives two source embeddings $E_A \in \mathbb{R}^{T_A \times d_A}$ and $E_B \in \mathbb{R}^{T_B \times d_B}$. They are mapped into the same hidden space:

$$M_A = \mathrm{LN}(E_A W_A + s_A), \qquad M_B = \mathrm{LN}(E_B W_B + s_B), \tag{3}$$

where W_A, W_B are source-specific projections, s_A, s_B are learnable source identifiers, and $\mathrm{LN}(\cdot)$ is layer normalization. The two memories are then concatenated into a unified, permutation-symmetric bank:

$$M = [M_A;\, M_B] \in \mathbb{R}^{(T_A+T_B)\times d}. \tag{4}$$

The target-side input is the shifted target sequence in embedding form,

$$H_0 = \mathrm{Embed}(Y_{<t}) + P, \tag{5}$$

where P is the positional encoding.

3.4. SCAD Decoder

3.4.1. Decoder Architecture

The decoder has L stacked layers. Layer l takes the previous hidden state H_{l-1} while retaining access to all earlier states $\{H_0, \ldots, H_{l-1}\}$ and is composed of masked self-attention (MSA), a cross-depth composer, signed dual-source cross-attention (SDCA), and a feed-forward update:

$$S_l = \mathrm{LN}(H_{l-1} + \mathrm{MSA}(H_{l-1}; \mathcal{M}_{\mathrm{causal}})), \tag{6}$$

$$C_l = \mathrm{LN}(S_l + \mathrm{SDCA}(S_l, M, \{H_i\}_{i<l})), \tag{7}$$

$$H_l = \mathrm{LN}(C_l + \mathrm{FFN}(C_l)). \tag{8}$$

The causal mask prevents each position from accessing future target tokens, while the SDCA block reads and fuses the two source memories.

3.4.2. Cross-Depth QKV Composer

Rather than forming its query from the most recent state only, the decoder builds a depth-aware query from the preceding layers:

$$\widetilde{H}_l = \sum_{i=0}^{l-1} \beta_{l,i} \odot H_i, \qquad \beta_{l,i} = \mathrm{softmax}_i(g(S_l, H_i)), \tag{9}$$

where $g(\cdot)$ is a small trained scoring function. The query is then projected, and the key and value are taken from the dual-source memory:

$$Q_l = \widetilde{H}_l W_Q, \qquad K_l = M W_K, \qquad V_l = M W_V. \tag{10}$$

3.4.3. Signed Dual-Source Cross-Attention

The query–key logits for a single head are

$$P_l = \frac{Q_l K_l^\top}{\sqrt{d_h}}, \tag{11}$$

where d_h is the head dimension. The key departure from ordinary cross-attention is that the attention weights are allowed to be negative. The decoder separates the magnitude and the sign of the logits:

$$\alpha_l = \mathrm{sign}(P_l) \odot \mathrm{softmax}(|P_l|). \tag{12}$$

Equivalently, for query position i and memory position j,

$$\alpha_{ij} = \mathrm{sign}(P_{ij}) \frac{\exp(|P_{ij}| - m_i)}{\sum_k \exp(|P_{ik}| - m_i)}, \qquad m_i = \max_k |P_{ik}|, \tag{13}$$

and the head output is

$$O_l = \alpha_l V_l. \tag{14}$$

In these expressions, Q_l, K_l, and V_l are the query, key, and value projections; M is the dual-source memory; P_l collects the scaled query–key logits; α_l is the resulting signed attention weight; O_l is the fused output of decoder layer l; and LN denotes layer normalization. Positive weights contribute source evidence to the target representation, while negative weights remove or damp conflicting evidence. The decoder can accordingly do more than pick which source token to attend to: it can also judge whether that token supports or opposes the current prediction, something non-negative attention simply cannot express.

The absolute value in Equation (12) is non-differentiable only on the measure-zero set $P_{ij} = 0$, which is essentially never reached exactly during training; where it is, automatic differentiation supplies the standard subgradient ($\partial|P|/\partial P = 0$ at the origin). Because the non-smooth $\mathrm{sign}(\cdot)$ factor is piecewise constant and carries no gradient, learning flows through the smooth magnitude branch $\mathrm{softmax}(|P|)$, whose outputs are bounded in $[0, 1]$; the signed weights therefore stay bounded, and gradients do not blow up at initialization. In practice we apply global-norm gradient clipping and a short learning rate warm-up, and training is stable from the first epochs. If a fully smooth operator is preferred, $|P|$ can be replaced by $\sqrt{P^2 + \epsilon}$ with negligible effect.

3.4.4. Symmetry Analysis of the Signed Attention

The construction in Equation (12) admits a clean symmetry interpretation that motivates the whole design. Consider how the attention weight α_{ij} responds to a sign flip of its logit, $P_{ij} \to -P_{ij}$. The magnitude factor $\mathrm{softmax}(|P_{ij}|)$ depends only on $|P_{ij}|$ and is therefore invariant, an even (sign-symmetric) function of the logit, whereas the prefactor $\mathrm{sign}(P_{ij})$ flips sign and is odd (sign-antisymmetric). The signed attention is thus the product of an even and an odd component,

$$\underbrace{\alpha_{ij}}_{\text{odd}} = \underbrace{\mathrm{sign}(P_{ij})}_{\text{odd}} \cdot \underbrace{\mathrm{softmax}_i(|P_{ij}|)}_{\text{even}}, \tag{15}$$

so that $\alpha_{ij}(-P_{ij}) = -\alpha_{ij}(P_{ij})$. The even part encodes how strongly a source token participates, and the odd part encodes its polarity, whether it reinforces or opposes the prediction. Standard softmax attention is the special case in which the odd part is clamped to $+1$: every weight is non-negative, the operator is invariant under the sign of the logits, and the model retains a sign symmetry that forbids subtraction. Allowing the odd component to take the value -1 breaks this symmetry in a controlled way; the broken symmetry is

exactly what enables conflict-aware fusion. Since the symmetry breaking is confined to the polarity term while the permutation symmetry of the dual-source memory (Equation (4)) is preserved, the two design choices are complementary: The architecture remains agnostic to the ordering of the modalities, while the attention becomes sensitive to their agreement.

3.4.5. Feed-Forward Update and Auto-Compressing Read-Out

After cross-attention, each position is processed by a position-wise feed-forward network,

$$\mathrm{FFN}(x) = W_2\,\phi(W_1 x + b_1) + b_2, \tag{16}$$

where ϕ is GELU or SiLU, and the output H_l is stored as layer memory. The final prediction is not produced by the last layer alone; instead, the decoder aggregates all intermediate states and applies the task head:

$$H_{\text{out}} = \sum_{l=0}^{L} \lambda_l \odot H_l, \qquad \hat{Y} = \mathrm{Head}(H_{\text{out}}), \tag{17}$$

with learnable read-out weights λ_l. The training objective is

$$\mathcal{L} = -\sum_{t} \log p(y_t \mid y_{<t}, E_A, E_B). \tag{18}$$

This read-out gives each layer a direct path to the supervision signal, so the model learns to concentrate useful information into the layers the task actually requires (automatic compression). The complete inference procedure is summarized in Algorithm 1.

Algorithm 1: SCAD-Detector inference.

Input: Image I; depth L; encoders KL-BERT, SWT
Output: Sensitivity label $\hat{Y}$

$T \leftarrow \mathrm{OCR}(I)$;
$E_A \leftarrow \text{KL-BERT}(T)$; $E_B \leftarrow \mathrm{SWT}(I)$;
$M_A \leftarrow \mathrm{LN}(E_A W_A + s_A)$; $M_B \leftarrow \mathrm{LN}(E_B W_B + s_B)$;
$M \leftarrow [\,M_A;\, M_B\,]$; `// permutation-symmetric memory`
$H_0 \leftarrow \mathrm{Embed}(Y_{<t}) + P$;
for $l = 1$ **to** L **do**
  $S_l \leftarrow \mathrm{LN}(H_{l-1} + \mathrm{MSA}(H_{l-1}; \mathcal{M}_{\text{causal}}))$;
  $\tilde{H}_l \leftarrow \sum_{i<l} \mathrm{softmax}_i(g(S_l, H_i)) \odot H_i$;
  $Q_l \leftarrow \tilde{H}_l W_Q$; $K_l \leftarrow M W_K$; $V_l \leftarrow M W_V$;
  $\alpha_l \leftarrow \mathrm{sign}(P_l) \odot \mathrm{softmax}(|P_l|)$; `// break sign symmetry`
  $C_l \leftarrow \mathrm{LN}(S_l + \alpha_l V_l)$;
  $H_l \leftarrow \mathrm{LN}(C_l + \mathrm{FFN}(C_l))$;
end
$H_{\text{out}} \leftarrow \sum_{l=0}^{L} \lambda_l \odot H_l$;
$\hat{Y} \leftarrow \mathrm{Head}(H_{\text{out}})$;
return $\hat{Y}$

4. Results

4.1. Experimental Setup

Hardware and implementation. All experiments were conducted on a workstation running Ubuntu 22.04.5 LTS with an Intel® Xeon® Platinum 8470Q CPU (Intel, Santa Clara, CA, USA) and an NVIDIA GeForce RTX 5090 GPU (NVIDIA, Santa Clara, CA, USA), using PyTorch 2.6.0. The visual branch adopts the Swin-Tiny configuration (patch size 4, window size 7, embedding dimension $C = 96$) initialized from ImageNet weights;

images are resized to 224×224 and augmented with random flipping, cropping, and color jittering. KL-BERT is obtained by continued pre-training of `bert-base-chinese` for 10 epochs (masking ratio 15%, learning rate 5×10^{-5}, batch size 32), with inputs padded to 128 tokens. Both encoders are projected to a common dimension $d = 256$, and the SCAD decoder uses $L = 4$ layers, 8 heads, and a feed-forward dimension of 1024. The full model is optimized with AdamW (weight decay 0.01) for 100 epochs using an initial learning rate of 1×10^{-4} with cosine annealing, a 5-epoch warm-up, a batch size of 16, and dropout 0.1.

Dataset. We use a self-constructed image privacy-sensitive target detection dataset of 2600 images covering diverse scenes and sources, organized into six categories: face (170), license plate (150), sensitive personal address (70), sensitive electronic screen (470), sensitive personal document (540), and other (1200). The distribution is shown in Figure 3; the data are split 8:2 (train:test) in a stratified manner. The strong imbalance toward "other" reflects the realistic prior that most shared images are benign, which makes recall a demanding metric. The images were gathered from publicly accessible web sources and open repositories that reflect realistic data-sharing scenarios, supplemented with desensitized internal compliance material; a candidate image was retained only when two independent annotators could label its sensitive content (or its absence) unambiguously, with a third annotator resolving disagreements. The "sensitive personal document" category comprises identity cards, household registration pages, passports, driver's licenses, bank cards, and official certificates, and an image is treated as sensitive when it exposes at least one protected field: name, national identity number, date of birth, home address, or account number. On ethics and data handling, all images were drawn from publicly available or consented sources, every personally identifying region in the examples shown in this paper was masked, and the data were processed solely for research on protective detection under the institution's data-security procedure. The "sensitive electronic screen" category covers photographs of displays (phone, computer, or ATM screens) that expose chat records, verification codes, account pages, or transaction details, and the "other" category collects benign images with no sensitive content (ordinary scenes, objects, and non-sensitive documents), which act as the negative class.

To make the benchmark reconstructible, we set out the collection protocol step by step. (i) Sourcing. Candidate images were gathered from three streams: keyword and reverse-image search over publicly accessible web pages, open image repositories with permissive licenses, and desensitized material from internal data-security compliance tests. The search terms were built from the category definitions above, in both Chinese and English. (ii) Filtering. We discarded very low-resolution images, near-duplicates of already accepted images (screened with a perceptual hash), synthetic renderings and clip-art, and any image whose provenance or licence could not be established. (iii) Selection. A candidate was retained only if two independent annotators could assign it to exactly one of the six categories without qualification; ambiguous images, for instance a document photographed so obliquely that no field is legible, were dropped rather than guessed at. (iv) Annotation. The two annotators labeled every retained image independently and a third annotator resolved disagreements. (v) Splitting. The final 2600 images were split 8:2 into training and test sets, stratified by category, with near-duplicate groups kept entirely on one side of the split so that no test image has a near-copy in training. Quantitative logs of the pipeline (crawl dates, per-category query counts, rejection rates, and pre-adjudication agreement statistics) were not recorded at collection time and are therefore not reported. The protocol and category definitions above suffice to construct a comparable benchmark; the dataset itself is available to researchers on reasonable request.

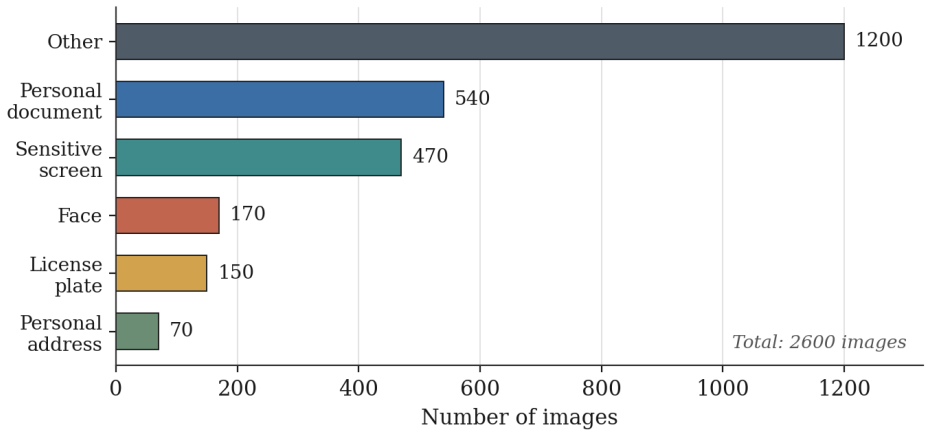

Figure 3. Category distribution of the privacy-sensitive image dataset (2600 images). The imbalance toward the "other" category mirrors the realistic prior that benign images dominate shared data.

Evaluation metrics. Sensitive images are positive samples and ordinary images are negative samples. Concretely, the five sensitive categories are pooled as the positive class and "other" as the negative class, so the accuracy, precision, recall, and F1-score reported in Tables 1–4 are those of this binary sensitive detection task; the per-category results in Section 5 additionally report the four metrics for each of the six classes. To keep the dominant "other" class from swamping the minority categories during training, we use a class-balanced weighted cross-entropy, and we treat recall and F1 as the primary metrics because accuracy is easily inflated on an imbalanced test set. We adopt accuracy, precision, recall, and F1-score:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \qquad \text{Precision} = \frac{TP}{TP + FP}, \tag{19}$$

$$\text{Recall} = \frac{TP}{TP + FN}, \qquad \text{F1-score} = 2 \times \frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}, \tag{20}$$

where TP, TN, FP, and FN are the numbers of true positives, true negatives, false positives, and false negatives, respectively.

Table 1. Performance comparison of different models. All values are in percent (%); the best in each column is in bold.

Model	Accuracy	Precision	Recall	F1-Score
ResNet	94.23	**96.64**	85.93	90.38
ViT	95.77	97.61	86.96	90.97
MBv2	94.62	90.23	86.05	87.39
SWT	96.00	96.33	87.80	91.26
RegNet	93.85	93.92	84.45	88.23
RepViT	93.85	93.74	84.62	88.23
CLIP	95.00	95.69	87.52	91.07
ViT + KL-BERT + SCAD	96.92	96.90	94.21	95.41
Ours: SWT + KL-BERT + SCAD	**97.69**	97.13	**96.30**	**96.70**

Table 2. Component-wise ablation of the proposed model. Values are in percent (%); the best in each column is in bold.

Model	Accuracy	Precision	Recall	F1-Score
ViT	95.77	97.61	86.96	90.97
ViT + KL-BERT	95.77	97.44	88.82	91.82
ViT + KL-BERT + SCAD	96.92	96.90	94.21	95.41
SWT	96.00	96.33	87.80	91.26
SWT + KL-BERT	96.54	96.39	92.73	94.17
Ours: SWT + KL-BERT + SCAD	**97.69**	**97.13**	**96.30**	**96.70**

Table 3. Component ablation on the test split; each row disables one component. Values are in percent (%); the best in each column is in bold.

Variant	Accuracy	Precision	Recall	F1-Score
w/o OCR text	95.00	92.64	89.09	90.27
w/o KL-BERT pre-training	95.38	95.63	93.85	94.65
w/o signed attention (softmax)	96.15	96.21	92.92	94.40
w/o negative branch ($\geq$0)	96.54	96.77	91.83	94.06
Full SCAD (Ours)	**97.69**	**97.13**	**96.30**	**96.70**

Table 4. Fusion-operator comparison under identical encoders. Values are in percent (%); the best in each column is in bold.

Fusion Operator	Accuracy	Precision	Recall	F1-Score
ViT + KL-BERT + MLP	95.77	96.12	90.47	92.69
ViT + KL-BERT + cross-attention	96.54	98.84	90.92	93.89
ViT + KL-BERT + SCAD	96.92	96.90	94.21	95.41
SWT + KL-BERT + MLP	95.00	95.37	92.00	93.29
SWT + KL-BERT + cross-attention	96.15	97.02	93.43	94.95
SWT + KL-BERT + SCAD (Ours)	**97.69**	**97.13**	**96.30**	**96.70**

4.2. Comparison with State-of-the-Art Models

Table 1 compares the proposed method with representative classification models. Our fused model leads on all four metrics: accuracy 97.69%, precision 97.13%, recall 96.30%, and F1-score 96.70%. Among the single-backbone models the strongest is SWT (accuracy 96%, F1 91.26%), which edges out ResNet and ViT and bears out the value of hierarchical window attention. Adding our components to ViT (ViT+KL-BERT+SCAD) raises accuracy from 95.77% to 96.92% and recall from 86.96% to 94.21%, evidence that semantic enrichment and symmetry-breaking fusion offset the weaker local features of ViT.

4.3. Ablation Study

Table 2 and Figure 4 separate out what each component contributes. Against the bare SWT model, the full system gains 1.69 points of accuracy and 8.5 points of recall. KL-BERT helps recall the most (SWT: 87.80% $\rightarrow$ 92.73%), whereas the symmetry-breaking SCAD module lifts accuracy from 96.54% to 97.69%. The ViT series behaves identically: ViT + KL-BERT + SCAD adds 7.25 points of recall and 4.44 points of F1 over ViT, which tells us the fusion strategy carries across backbones.

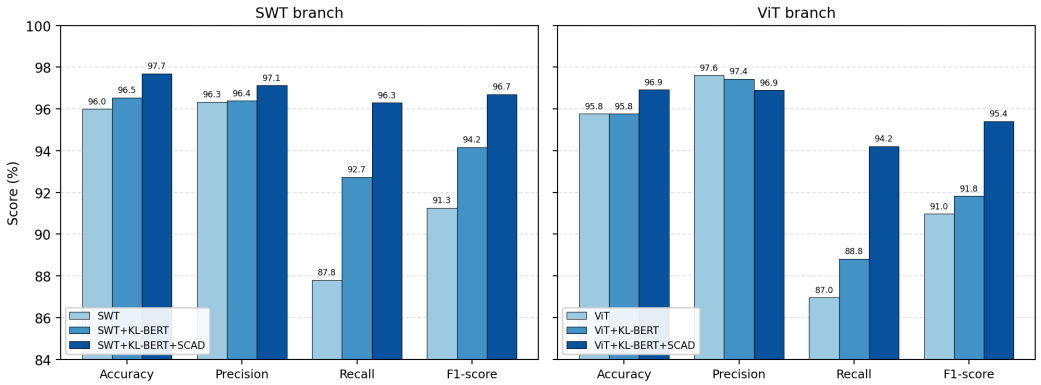

Figure 4. Ablation progression, shown as grouped bar charts, for the SWT branch (**left**) and the ViT branch (**right**). KL-BERT chiefly boosts recall, while the SCAD decoder lifts all three metrics, with the largest cumulative gain in recall.

4.4. Component Ablation and Design Analysis

To isolate the contribution of each part of the framework, Table 3 switches off one component at a time on the same test split. Removing the OCR text stream causes the largest drop (F1 96.70% $\rightarrow$ 90.27%), confirming that the embedded text carries decisive evidence. Removing the policy pre-training of KL-BERT lowers F1 to 94.65%. Most relevant to our central claim, replacing the signed attention with ordinary softmax attention costs 2.3 F1 points, and clamping only the negative weights to zero, so the operator becomes non-negative again, removes 4.5 points of recall while leaving the magnitude term unchanged. Since these two variants differ only in whether subtraction is allowed, the comparison isolates the symmetry-breaking mechanism itself. Table 3 does not include a row for the cross-depth QKV composer of Section 3.4.2, as the corresponding isolation run (forming the query from the most recent layer state alone, $\widetilde{H}_l = H_{l-1}$) has not been performed. The composer is therefore treated as a query-refinement design choice, with no measured gain claimed for it.

We further compare the fusion operator against two standard alternatives, an MLP fusion and a conventional softmax cross-attention decoder, under identical encoders (Table 4). With everything else fixed, moving from standard cross-attention to signed cross-attention raises F1 from 94.95% to 96.70% on the SWT backbone, which we attribute to the polarity term rather than to added capacity.

Finally, Table 5 varies the number of decoder layers. Performance improves up to $L = 4$ and then declines, consistent with mild over-fitting of a deeper decoder on a modestly sized dataset; we therefore use four layers. More generally, because the dataset is modest in size, we take several steps to limit over-fitting: both encoders are initialized from pre-trained weights (Swin-T on ImageNet and KL-BERT from `bert-base-chinese`) and only lightly adapted, so few parameters are trained from scratch; the images are augmented with random flipping, cropping, and color jittering; and we apply weight decay, dropout, a stratified 8:2 split, and early stopping. The stable margin between our model and the baselines on the held-out test set indicates that the network is not simply memorizing the training data.

Table 5. Effect of the decoder depth L. Values are in percent (%); the best in each column is in bold.

Decoder Depth	Accuracy	Precision	Recall	F1-Score
$L = 1$	94.62	95.20	88.30	91.33
$L = 2$	95.00	94.64	92.63	93.36
$L = 3$	97.31	**97.75**	95.10	96.33
$L = 4$ (used)	**97.69**	97.13	**96.30**	**96.70**
$L = 5$	96.54	96.27	94.41	95.19
$L = 6$	95.77	95.17	92.78	93.90

4.5. Statistical Caveats of the Single-Run Protocol

All results are from single training runs under one fixed random seed, with the same stratified split, schedule, and hyper-parameters for every model and ablation variant, so no method benefits from a more favorable draw; single-run reporting of this kind is still common in applied vision–language studies. As a single run cannot bound run-to-run variance, no claim of statistical significance is made for the reported margins. Three observations nevertheless suggest that the margins are not artifacts of a favorable run: The gains over the strongest fusion baseline (+1.75 F1, Table 4) and over the non-negative variant (+2.64 F1, Table 3) are consistent across all four metrics; the operator ordering (MLP, standard cross-attention, signed SCAD) recurs on both visual backbones; and both sign-related ablations degrade performance as the theory predicts. Multi-seed replication with paired significance testing is left to future work.

4.6. Cross-Modal Conflict: Definition and Indirect Evidence

Our central claim concerns behavior under cross-modal conflict. The direct test would be a conflict-annotated split of the test set, an image counting as conflicting when its OCR-recovered text and visual content support different labels: benign imagery carrying incidental sensitive-looking text (watermarks, captions, background signage) or genuinely sensitive imagery whose embedded text is absent, illegible, or innocuous. Such an annotation requires two independent annotators with third-annotator adjudication and is left to future work, so the evidence offered here is indirect. First, in the face category, where this conflict is most frequent, naive fusion falls below the visual-only baselines (ViT + KL-BERT, 69.2% F1) while the signed decoder reaches 90.9%; text–visual conflict under a purely additive weighting explains this, whereas visual difficulty alone does not. Second, clamping only the negative branch (Table 3) costs 4.5 recall points, indicating that the subtractive path acts where evidence must be suppressed. Third, the "w/o OCR" ablation bounds the total contribution of the text stream. A direct conflict-vs-non-conflict breakdown, with visualization of the signed attention weights, would complete this analysis.

5. Discussion

Where the gains come from, and why symmetry breaking matters. Figure 5 gives the per-category F1-score for representative models. The visually regular categories (electronic screen, document, license plate) are already served well by ordinary backbones, so there is little headroom. Faces are the hard case; small objects, occlusion, and wide intra-class variation keep the visual-only and naively fused baselines below 76%. Tellingly, naive fusion can even hurt (ViT+KL-BERT falls to 69.2%) because stray OCR text (watermarks, background captions) misleads a rule that can only average the modalities, which is exactly the breakdown the sign symmetry of softmax attention predicts. Once that symmetry is broken through the signed (odd) polarity term, the model can subtract the misleading text and pushes face F1 to 90.9%, about 15 points above the best visual baseline. That the improvement lands precisely where the modalities most often disagree is direct support

for the symmetry-breaking hypothesis of Section 3.4.4. The face category is, of course, also intrinsically hard for reasons unrelated to fusion, such as small faces, occlusion, and pose variation, but these affect every method equally; the extra gain of SCAD over the strong visual baselines here, together with the fact that naive fusion actually degrades this category, isolates the part of the difficulty that is due to text-visual conflict rather than to visual difficulty alone.

Although we frame the task as binary sensitive-vs-benign detection, the per-category outputs also support a graded, risk-based reading: high-risk classes such as identity documents and faces, medium-risk classes such as addresses and licence plates, and low-risk incidental screen content can each be routed to a different downstream policy. A full multi-level severity model is left to future work.

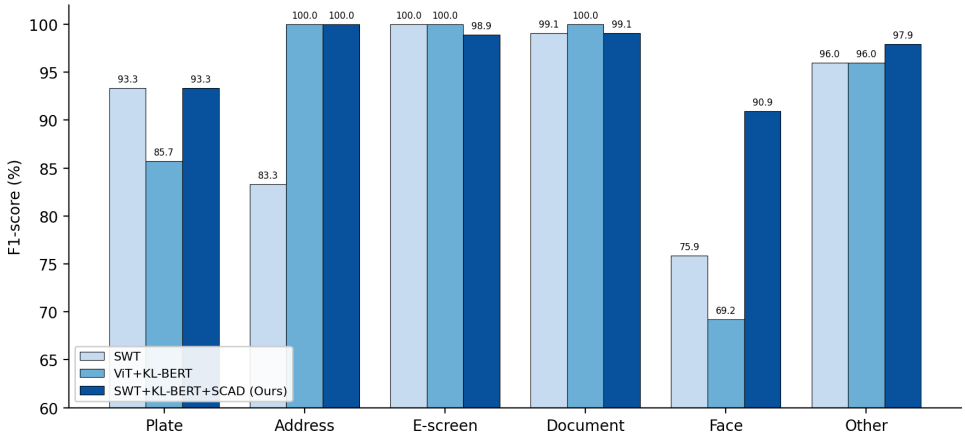

Figure 5. Per-category F1-score (%) across representative models. The proposed model (bottom row) delivers its largest gains on the most difficult "face" category, where the two modalities most often disagree.

Positioning against large vision–language models. We position the method against recent large vision–language models: CLIP is included in Table 1 as a representative contrastive model that adapts to a fixed label set, whereas generative or retrieval models such as BLIP, SigLIP, Qwen-VL, and Florence-2 are pre-trained mainly on English web data and would need prompt engineering or heavy fine-tuning to fit this closed six-way Chinese task; we therefore take CLIP as the fair reference point and leave the adaptation of larger multilingual models to future work.

5.1. Computational Cost

Table 6 lists the parameter budget and theoretical cost. The Swin-T backbone is modest (28.3 M parameters, 4.5 GFLOPs at 224 × 224), on par with ResNet-50 and lighter than ViT-B/16. Bringing in the KL-BERT branch and the SCAD decoder pushes the total to roughly 150 M parameters, heavier than a purely visual classifier, but paid only once per image, and the auto-compressing read-out lets inference run at a shallower effective depth. As the two encoders can run in parallel, wall-clock latency is dominated by the slower branch rather than their sum.

Failure cases and category-level errors. Averaged scores hide where the model actually breaks, so we inspected the misclassified test images. The errors fall into four recurring patterns, which we characterize qualitatively: The misclassification set of a 520-image split is too small for per-pattern counts to be meaningful under a single-run protocol. False negatives on partially occluded or very small faces: When the face occupies a small fraction of the frame and the image carries no informative text, neither branch has much to contribute and the fused prediction reverts to the visual backbone. False negatives

from OCR failure occur on documents photographed at an angle or on low-quality screen captures where the protected field is not recovered; these are the cases the "w/o OCR" ablation isolates in aggregate. False positives on benign images carrying sensitive-looking text are the residue of the conflict problem: the negative branch suppresses most watermark and signage text, but a fragment that is both legible and regulatory in phrasing can still tip the decision. Category confusions among the sensitive classes are dominated by the electronic screen versus document boundary, since a photograph of an identity document displayed on a screen legitimately belongs to both. In deployment terms the third group is the costly one because a detector that over-flags benign images will be switched off; we therefore report precision alongside recall throughout and note that the operating threshold should be tuned to the reviewing capacity of the deployment rather than left at 0.5.

Table 6. Model size and theoretical cost. FLOPs are for the visual backbone at 224 × 224; the fused model also includes the KL-BERT text encoder and SCAD decoder.

Model	Params (M)	GFLOPs
MBv2	3.5	0.3
RepViT	8.2	1.3
RegNet	20.6	4.0
ResNet	25.6	4.1
SWT (Swin-T)	28.3	4.5
ViT (ViT-B/16)	86.6	17.6
Ours: SWT + KL-BERT + SCAD	≈150	≈15

5.2. Limitations

Two limitations remain. First, the framework depends on OCR quality; when the embedded text is heavily distorted or absent, the textual branch contributes little and the model falls back to its visual backbone. Second, the present study is confined to Chinese policy-domain text and a single, modestly sized dataset, so generalization to other languages and open-world sensitive categories is not yet established. Robust scene text recognition, larger corpora, and multilingual policy pre-training are natural directions for future work. Beyond these, annotation reliability (mitigated here by dual annotation with third-annotator adjudication), dataset bias toward Chinese policy-domain content, residual OCR errors on distorted in-image text, and the gap between benchmark accuracy and real deployment all warrant caution when generalizing the reported numbers. Several validations are left to future work: multi-seed replication with paired significance testing, an ablation isolating the cross-depth composer, a conflict-annotated breakdown with visualization of the signed attention weights, measurement of the on-dataset OCR recognition rate, and extension of the binary formulation to a graded risk model.

6. Conclusions

We have presented a symmetric dual-source vision–language framework whose decoder, SCAD, performs conflict-aware fusion by deliberately breaking the sign symmetry of conventional attention. A Swin Transformer and a policy knowledge-enhanced BERT encode image and text and assemble them into a permutation-symmetric memory; the signed cross-attention then splits into a sign-symmetric magnitude term and a sign-antisymmetric polarity term, so the model can both reinforce and suppress cross-modal evidence. On a self-constructed privacy-sensitive image dataset it reaches an accuracy of 97.69% and an F1-score of 96.70%, with its gains concentrated on the most conflict-prone categories, indicating that controlled symmetry breaking is an effective and interpretable principle for multimodal fusion. In practice, the framework applies directly to privacy screening on data-sharing and content moderation platforms, where images must be checked for exposed

identity documents, faces, screens, or license plates before release, and its per-category outputs support graded, risk-based handling in compliance auditing.

Author Contributions: Conceptualization, M.L.; Methodology, M.L., Y.L. and T.L.; Software, T.L.; Validation, M.L.; Formal analysis, M.L.; Investigation, M.L. and T.L.; Resources, Y.L.; Data curation, M.L.; Writing—original draft, M.L.; Writing—review and editing, M.L. and Y.L.; Supervision, Y.L. and T.L.; Project administration, Y.L. and T.L.; Funding acquisition, M.L., Y.L. and T.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Natural Science Foundation of China under Grant 62576005 and K110168019, and in part by the project of Excellent Research and Innovation Teams in Anhui Province's Universities under Grant 2024AH010030.

Data Availability Statement: The dataset consists of privacy-sensitive images, whose public deposit would itself constitute the disclosure the detector is designed to prevent; it is therefore available from the corresponding author on reasonable request for non-commercial research purposes, subject to a data-use agreement. The collection protocol and category definitions needed to construct a comparable benchmark are given in Section 4.1.

Acknowledgments: During the preparation of this manuscript, the authors used GPT-5.5 solely to improve the language and readability of author-written text (grammar, spelling, punctuation, and phrasing). The tool was not used to generate scientific content, data, analyses, interpretations, or references. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

OCR	Optical Character Recognition
SWT	Swin Transformer
BERT	Bidirectional Encoder Representations from Transformers
KL-BERT	Knowledge-enhanced (policy-Legislation) BERT
SCAD	Signed Cross-attention Auto-compressing Decoder
MSA	Masked Self-Attention
SDCA	Signed Dual-source Cross-Attention
MLM	Masked Language Modeling
ViT	Vision Transformer

References

1. Zhou, B.; Shi, Y.; Li, C. Research on Improvement of Mass Multimedia Information Filtering Technology Under Multi-Data Inflow. In *Proceedings of the 2019 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS)*; IEEE: Piscataway, NJ, USA, 2019; pp. 438–442. [CrossRef]
2. Ding, M.; Wang, X.; Wu, C.; Wang, K.; Yang, X. Research on Automated Detection of Sensitive Information Based on BERT. *J. Phys. Conf. Ser.* **2021**, *1757*, 012088. [CrossRef]
3. Lv, A.; Xie, R.; Li, S.; Liao, J.; Sun, X.; Kang, Z.; Wang, D.; Yan, R. More Expressive Attention with Negative Weights. *arXiv* **2024**, arXiv:2411.07176. [CrossRef]
4. Ye, T.; Dong, L.; Xia, Y.; Sun, Y.; Zhu, Y.; Huang, G.; Wei, F. Differential Transformer. In Proceedings of the International Conference on Learning Representations (ICLR), Singapore, 24–28 April 2025.
5. Knuth, D.E.; Morris, J.H., Jr.; Pratt, V.R. Fast Pattern Matching in Strings. *SIAM J. Comput.* **1977**, *6*, 323–350. [CrossRef]
6. Aho, A.V.; Corasick, M.J. Efficient String Matching: An Aid to Bibliographic Search. *Commun. ACM* **1975**, *18*, 333–340. [CrossRef]
7. Xu, G.; Wu, X.; Yao, H.; Li, F.; Yu, Z. Research on Topic Recognition of Network Sensitive Information Based on SW-LDA Model. *IEEE Access* **2019**, *7*, 21527–21538. [CrossRef]
8. Liu, Y.; Yang, C.Y.; Yang, J. A Graph Convolutional Network-Based Sensitive Information Detection Algorithm. *Complexity* **2021**, *2021*, 6631768. [CrossRef]

9. Zhang, K.; Liu, X.; Zhao, N.; Liu, S.; Li, C. Dual channel semantic enhancement-based convolutional neural networks model for text classification. *Int. J. Mod. Phys. C* **2025**, *36*, 2442012. [CrossRef]
10. Zhang, X.; Zhao, J.; LeCun, Y. Character-Level Convolutional Networks for Text Classification. In *Advances in Neural Information Processing Systems (NeurIPS)*; Curran Associates, Inc.: Red Hook, NY, USA, 2015; Volume 28, pp. 649–657.
11. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-Based Learning Applied to Document Recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
12. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems (NeurIPS)*; Curran Associates, Inc.: Red Hook, NY, USA, 2012; Volume 25, pp. 1097–1105.
13. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In Proceedings of the International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015.
14. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
15. Guo, R.; Li, W.; Wang, H.; Zhang, X. A Multi-Information Detection Model for Violent Images Based on YOLOv3-SPP and DenseNet. In *Proceedings of the 2022 11th International Conference of Information and Communication Technology (ICTech)*; IEEE: Piscataway, NJ, USA, 2022; pp. 529–532. [CrossRef]
16. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.
17. Huang, G.; Liu, Z.; van der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
18. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2018; pp. 7132–7141. [CrossRef]
19. Zhang, S.X.; Zhu, X.; Hou, J.B.; Liu, C.; Yang, C.; Wang, H.; Yin, X.C. Deep Relational Reasoning Graph Network for Arbitrary Shape Text Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2020; pp. 9699–9708. [CrossRef]
20. Liao, M.; Shi, B.; Bai, X. TextBoxes++: A Single-Shot Oriented Scene Text Detector. *IEEE Trans. Image Process.* **2018**, *27*, 3676–3690. [CrossRef] [PubMed]
21. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. In *Proceedings of the 14th European Conference on Computer Vision (ECCV)*; Springer International Publishing: Amsterdam, The Netherlands, 2016; pp. 21–37.
22. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 4510–4520.
23. Zhang, S.; Zhu, X.; Lei, Z.; Shi, H.; Wang, X.; Li, S.Z. S3FD: Single Shot Scale-Invariant Face Detector. In *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*; IEEE: Piscataway, NJ, USA, 2017; pp. 192–201. [CrossRef]
24. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. In *Advances in Neural Information Processing Systems (NeurIPS)*; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30, pp. 5998–6008.
25. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16 $\times$ 16 Words: Transformers for Image Recognition at Scale. In Proceedings of the 9th International Conference on Learning Representations (ICLR 2021), Virtual, 3–7 May 2021.
26. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Virtual, 11–17 October 2021*; IEEE: Piscataway, NJ, USA, 2021; pp. 10012–10022.
27. Li, J.; Li, D.; Savarese, S.; Hoi, S. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In Proceedings of the 40th International Conference on Machine Learning (ICML), Honolulu, HI, USA, 23–29 July 2023; Volume 202, pp. 19730–19742.
28. Zhai, X.; Mustafa, B.; Kolesnikov, A.; Beyer, L. Sigmoid Loss for Language Image Pre-Training. In *Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023*; IEEE: Piscataway, NJ, USA, 2023; pp. 11975–11986.
29. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning Transferable Visual Models From Natural Language Supervision. In Proceedings of the 38th International Conference on Machine Learning (ICML), Virtual, 18–24 July 2021; pp. 8748–8763.
30. Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; et al. InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024; pp. 24185–24198.
31. Li, S.; Tang, H. Multimodal Alignment and Fusion: A Survey. *Int. J. Comput. Vis.* **2026**, *134*, 103. [CrossRef]

32. Baia, A.E.; Xompero, A.; Cavallaro, A. Zero-shot Image Privacy Classification with Vision-Language Models. *arXiv* **2025**, arXiv:2510.09253. [CrossRef]
33. Shi, B.; Bai, X.; Yao, C. An End-to-End Trainable Neural Network for Image-Based Sequence Recognition and Its Application to Scene Text Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2298–2304. [CrossRef] [PubMed]
34. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.