INFORMATION-GUIDED REPRESENTATION LEARNING FROM GUI PIXELS FOR MISINFORMATION DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Most misinformation detectors operate on text and metadata, whereas users encounter posts through rendered interfaces that also expose account cues and selected replies. Screenshots make these cues available but entangle decision-relevant evidence with visual appearance. We study GUI-derived representations using source-aware structured fields, a frozen screenshot encoder, and trainable content, identity, appearance, and exposure branches. The model keeps selected cross-type interactions explicit and applies a consistency objective to low-level transformations that preserve the post and its visible evidence. To separate event generalization from repeated event content, we construct a post-grouped split of 1,878 PHEME and Twitter15/16 posts into 183 adjudicated global event groups. On 360 posts from 37 held-out groups, a structured model using GUI-extracted and legacy-derived fields reaches 71.67% Macro-F1, versus 64.81% for extracted-text features and 54.01% for a frozen screenshot embedding alone. Yet ordinary fusion, which leads under a random-post split, falls behind structured fields on unseen events; neither typed interactions nor appearance-pair consistency closes this gap. This reversal shifts the focus from fusing more GUI information to identifying which interface-derived evidence generalizes across events.

1 INTRODUCTION

Social-media misinformation detection is commonly formulated over machine-readable records. Existing models encode post text, user metadata, interactions, or propagation structures to predict whether a claim is true or false (Castillo et al., 2011; Ma et al., 2016; Monti et al., 2019). These inputs describe what a platform stores, but they do not fully specify what a user observes. A social-media interface renders the same post together with identity markers, engagement statistics, visible replies, and a visual hierarchy that determines how these elements are presented. The resulting screenshot is therefore not merely a picture of an API record; it is a platform-rendered observation through which the post is actually consumed. We study misinformation detection from this rendered observation rather than from records alone.

Moving from records to screenshots broadens the available observation, but it does not by itself produce a reliable representation. Multimodal misinformation detectors combine language with attached images, yet the image is typically treated as post content rather than as the interface that presents the post (Wang et al., 2018; Khattar et al., 2019). Screenshot-based work has shown that the overall appearance of a news webpage can be predictive of source trustworthiness (Abdali et al., 2021), while recent GUI models demonstrate that text, layout, and interactive elements can be grounded directly from pixels (Gou et al., 2025; Lin et al., 2025). These advances establish that screenshots contain usable signal. They do not establish how that signal should be organized for misinformation detection. Within a screenshot, semantic content, account identity, exposure-related context, and visual appearance are jointly encoded in pixels; a model that fuses them indiscriminately can exploit any of them, including renderer-specific shortcuts. The central challenge is therefore not simply to retain more GUI information, but to control how different types of GUI information influence the prediction.

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.

Different changes to a rendered interface should induce different model responses. If only rendering or encoding changes while the post and its visible evidence remain fixed, the veracity prediction should remain stable. By contrast, a change in the evidence exposed by the interface may legitimately change the model's confidence, even when the underlying label is unchanged. Conflating these cases obscures what the model should ignore and what it may use as evidence. We call the required behavior **selective invariance**: the representation should suppress changes that alter only rendering, while preserving factors that can change the information available for a decision. This asymmetry motivates an explicitly typed representation rather than an undifferentiated screenshot embedding.

We operationalize this distinction with four addressable input types: content, identity, appearance, and exposure. A frozen visual backbone represents the post screenshot, while trainable projections and selected content–identity, content–exposure, and identity–appearance interactions combine it with GUI-extracted text and structured fields. This design replaces the fixed credibility weights and decision rules of our earlier GUI prototype at the prediction stage; inherited structured scores remain explicit inputs whose provenance can be audited. We also train a variant with a consistency objective for low-level appearance pairs that preserve the post and visible evidence, without forcing invariance when the exposed evidence changes.

We evaluate seven supervised configurations on a frozen, post-grouped event split of 1,878 GUI-rendered posts derived from PHEME and Twitter15/16 (Zubiaga et al., 2017; Ma et al., 2017). Adjudicated event grouping yields 183 global event groups; the held-out test partition contains 360 posts from 37 groups. Under this protocol, the structured GUI-derived baseline reaches 71.67% Macro-F1, exceeding extracted-text (64.81%) and frozen screenshot-only (54.01%) models. Ordinary fusion of the four input types reaches 66.46%, while typed fusion without and with the appearance-pair objective reaches 64.98% and 63.41%. Thus, the present typed design does not improve the primary event-held-out metric. A separate 120-post audit of the existing GUI-processing record measures field-level agreement with human review, making extraction quality part of the interpretation. Results from the older fixed-rule prototype are retained as a distinct full-collection pilot.

Our contributions are threefold:

1. We frame misinformation detection from **platform-rendered observations** as a problem of distinguishing evidence-bearing GUI context from appearance changes that should not alter a decision.
2. We provide an auditable, trainable **typed representation** over content, identity, appearance, and exposure, with restricted interactions and a consistency objective whose scope is defined by evidence preservation.
3. We establish an adjudicated, **event-disjoint evaluation** and show that a strong GUI-derived structured baseline remains informative across held-out events, whereas a frozen screenshot embedding and either typed-fusion configuration do not account for that advantage.

2 Problem Formulation

The prediction unit is a source post, not an individual screenshot, comment, or transformed variant. Index posts by n; let $y_n \in \{0, 1\}$ be the false/true label (1 denotes true) and g_n the rendered source post with its available comment views. Let $t \in \mathcal{T} = \{c, i, a, e\}$ index four operational input types: content x_n^c (post text and topic), identity x_n^i (visible account attributes), appearance x_n^a (a frozen source-post screenshot embedding), and exposure x_n^e (captured comment/view summaries). Figure 1 suppresses the post index n. This partition is a modeling choice, not a claim of statistical independence or completeness; the screenshot embedding may also encode content and identity.

A transformation is *appearance-preserving* only if source text, account, visible evidence, and label remain unchanged. Writing $\tilde{x}_n^a$ for the transformed screenshot embedding and p_θ for the model's probability of the true class, we seek $p_\theta(y = 1 \mid x_n^c, x_n^i, \tilde{x}_n^a, x_n^e) \approx p_\theta(y = 1 \mid x_n^c, x_n^i, x_n^a, x_n^e)$. No such constraint applies when exposed replies or other evidence changes. This is selective invariance: it specifies which perturbations should leave a prediction stable without assuming that every visual change is irrelevant.

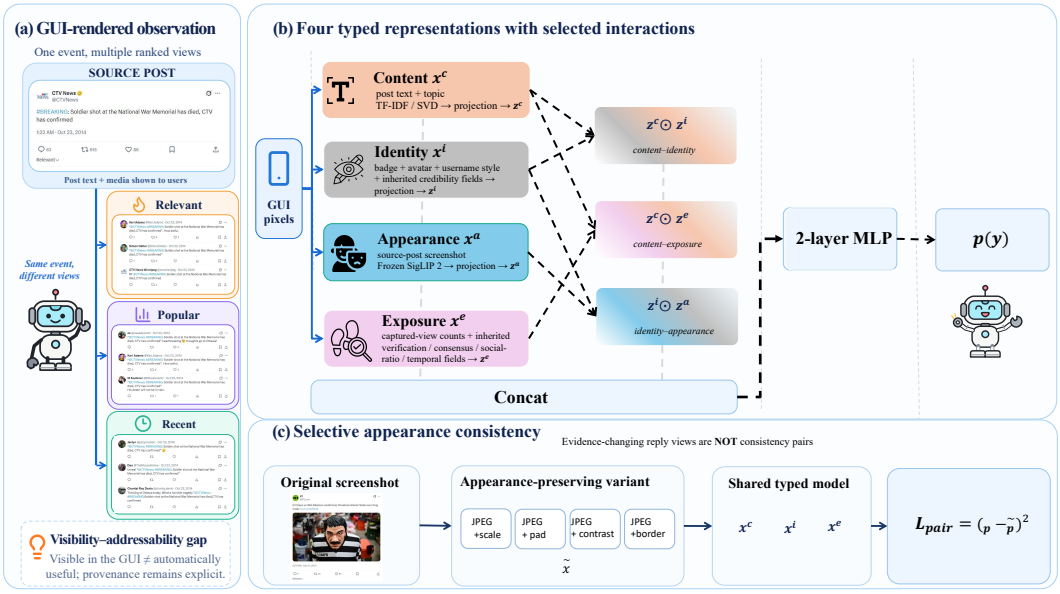

Figure 1: Current learned-model workflow. (a) A source post and three captured reply views. (b) Four projected input types, three selected elementwise interactions, concatenation, and a two-layer MLP. (c) A separate post illustrates prediction comparison under evidence-preserving rendering changes during pair training. Superscripts c, i, a, e denote content, identity, appearance, and exposure; post indices are omitted. Equations (1)–(3) specify the computation.

3 METHOD

FROM GUI SOURCE FLOW TO TYPED INPUTS

Figure 1(a) shows one source post and its captured Relevant, Popular (called Likes in the processing record), and Recent reply views; these are capture labels, not verified ranking algorithms. Panel (b) maps the observation to four typed inputs, three selected interactions, and a two-layer classifier. Panel (c) shows the training-only comparison with an evidence-preserving rendering variant. The predecessor fixed-rule pipeline remains a distinct pilot in Appendix F.

The arrows from GUI pixels in panel (b) denote feature extraction, not end-to-end pixel-encoder training. The builder reads post text, account categories, and comment/view summaries from the GUI-processing record, and embeds only the *source-post screenshot* with frozen SigLIP 2 (Tschannen et al., 2025); comment screenshots are not separately embedded. Identity and exposure also include rule-derived scores (Appendix E). This is panel (a)'s visibility–addressability gap: a visible cue is not automatically an explicit input or a validated predictor. Gains could partly reflect inherited scores, so a raw-field-only control remains necessary.

The distinction between evidence and presentation determines which invariance is appropriate. Schematically, write $g_n = \mathcal{R}(c_n, u_n, e_n; r_n)$, where $\mathcal{R}$ denotes rendering, c_n source-post content, u_n displayed account identity, e_n selected reply evidence, and r_n the rendering state. Holding (c_n, u_n, e_n) fixed while changing r_n alters pixels without changing the target label; ideally the prediction remains stable. Changing e_n also alters pixels, but now the available evidence differs. Imposing consistency on both cases could erase a useful response to evidence. This factorization is a schematic constraint on *observations*, not an identifiable causal model or a statistical-independence claim.

This observation creates two different modeling decisions. First, the prediction should be able to use evidence that appears only in the rendered context: an account badge or a sampled reply view may be absent from the source text. We therefore expose these fields to the classifier rather than asking a single post-screenshot vector to recover them implicitly. Second, changes to the drawing of the same evidence should not be treated like changes to the evidence itself. The content and exposure

branches supply a fixed reference when the appearance branch is perturbed; if the reply set changes, the exposure input must change as well. This distinction is the reason for the four-branch interface, rather than a claim that each branch corresponds to a unique latent factor in the pixels.

Typed inputs make this constraint operational. The frozen *source-post* screenshot vector can encode both post text and layout, so an undifferentiated consistency penalty cannot by itself specify which information should remain stable. By keeping structured content, identity, and exposure inputs fixed while replacing only the screenshot vector with one from a content-preserving variant, the paired objective targets a narrower comparison. The separate projections also make feature provenance and selected cross-type interactions inspectable. Ordinary fusion can in principle learn the same decision rule; typing is an inductive bias whose benefit must be tested against the input-matched fusion baseline, especially when event changes may alter correlations among account cues, rendering patterns, and labels. Neither the decomposition nor the pair loss guarantees event invariance or causal disentanglement.

TYPE-SPECIFIC PROJECTIONS AND INTERACTIONS

Content uses train-fitted word-level unigram/bigram TF–IDF over post text and topic, truncated SVD, and standardization. Identity concatenates one-hot badge, avatar, username-style, and predecessor credibility-level categories with its standardized credibility score. Exposure contains relevant/likes/recent comment counts and predecessor-derived verification, consensus, social-ratio, and temporal scores. Appearance is the frozen post-screenshot embedding. Vocabularies, imputers, and scalers are fitted on training data only.

The branch assignment follows the *provenance* of each field, not a guarantee about its semantics. Post text and topic describe what is claimed; badge, avatar, and username style describe the displayed speaker; the three captured comment views summarize which responses were exposed; and the screenshot vector captures the rendered source-post region. The fixed-rule credibility and view scores are retained as inputs to reproduce the available GUI-processing record, but are explicitly distinguished from raw observations. Consequently, a structured-only model may benefit from those engineered scores even if a learned cross-type interaction contributes nothing. This is why we report both the structured row and the input-matched ordinary-fusion row, and why the raw-field-only control remains necessary for attributing a gain to the proposed architecture.

For each type $t \in \mathcal{T}$, separate trainable parameters project its input into a common dimension:

$$z_n^t = \mathrm{GELU}\big(W_t\,\mathrm{LN}(x_n^t) + b_t\big)\,, \qquad z_n^t \in \mathbb{R}^d. \tag{1}$$

Here W_t and b_t are the type-specific weight matrix and bias, LN is layer normalization, GELU is the activation, and $d = 64$ is the projected width. The vector z_n^t is the learned representation of type t for post n; it does not establish statistical disentanglement. We include only three pre-specified interactions: content–identity, content–exposure, and identity–appearance. With $\odot$ denoting elementwise multiplication and semicolons denoting concatenation,

$$h_n = [z_n^c; z_n^i; z_n^a; z_n^e; z_n^c \odot z_n^i; z_n^c \odot z_n^e; z_n^i \odot z_n^a], \qquad p_n = \sigma(f_\theta(h_n)). \tag{2}$$

Here h_n is the classifier input, f_θ a two-layer MLP with trainable parameters θ, hidden width 128, and dropout 0.2, and σ the logistic sigmoid. Thus p_n is the predicted probability of $y_n = 1$; the figure's $p(y)$ abbreviates the resulting binary predictive distribution. The interactions are testable modeling hypotheses, not conditional-mutual-information estimates.

The forward pass is identical at training and test time except for the optional paired view. A post first yields four vectors through the training-fitted feature transforms; each vector is normalized and projected by its own learned map; the resulting 64-dimensional states and the three selected products enter the shared classifier. The SigLIP 2 encoder and the predecessor GUI extractor are frozen, whereas the four projections, interaction-sensitive classifier, and ordinary-fusion comparison head are trained on post labels. Thus, "representation learning" here refers to the supervised branch projections and decision representation, not to end-to-end pixel-encoder adaptation. At inference, only the untransformed post is required; neither the synthetic pair nor the pair loss is used.

The three products encode narrowly specified comparisons. Content–identity allows the same wording to be interpreted alongside its displayed source; content–exposure allows the post claim to be compared with the available response summaries; and identity–appearance allows an account field

to interact with its rendered presentation. These products do not enforce agreement, and the final MLP can ignore them. We therefore isolate their contribution by removing all three while retaining the four projected branches. We also compare against ordinary fusion trained on the same input vectors: if typed fusion does not outperform that baseline on unseen events, branch addressability is a useful audit property but not evidence of superior predictive generalization.

Selective appearance consistency

For a training post with an appearance-preserving variant admitted by the pair-audit procedure, the pair retains identical content, identity, and exposure vectors while replacing only the screenshot embedding. The objectives are

$$\mathcal{L}_{\text{sup}} = \text{BCE}_w(y_n, p_n), \quad \mathcal{L}_{\text{pair}} = (p_n - \tilde{p}_n)^2, \quad \mathcal{L} = \mathcal{L}_{\text{sup}} + \lambda \mathcal{L}_{\text{pair}}. \tag{3}$$

Here BCE_w is class-weighted binary cross-entropy, p_n is the original prediction, $\tilde{p}_n$ is the prediction after replacing only x_n^a with $\tilde{x}_n^a$ in the same model, and $\lambda = 0.2$ weights the squared prediction difference $\mathcal{L}_{\text{pair}}$. The pair-trained variant consumes 800 train-partition variant pairs. Four transformations were generated per selected post: JPEG recompression, scale-and-pad, contrast adjustment, and a border. Pair eligibility and audit provenance are detailed in Appendix A. A transformation erasing text, cropping badges, or changing visible evidence should fail semantic verification. A changed comment pool is a different exposure condition, not an appearance pair. The loss compares predictions under admitted variants; it does not impose equal intermediate features or prove that the model ignores rendering.

4 Experimental Setup

Data and event-separated protocol

We study 1,878 unique source posts rendered in an X/Twitter-style GUI: 1,290 from PHEME (Zubiaga et al., 2017) and 588 from Twitter15/16 (Ma et al., 2017). The label-authoritative external file contains 1,386 true and 492 false posts (PHEME: 991/299; Twitter15/16: 395/193). The predecessor's embedded label disagrees with this file for 296 posts; all supervised runs here use the external labels. Each sample links a source-post screenshot to any available captured Relevant, Likes/Popular, and Recent reply views. The source post is the prediction unit, and all its views and rendered variants remain in the same partition.

Event identity is critical because similar posts from one event can otherwise cross partitions (Wu & Hooi, 2022). Candidate event IDs were reconciled through human review, adjudication, and approved cross-dataset links, yielding 183 global groups. We froze disjoint train, validation, and test groups before the reported model runs; Table 1 gives the resulting partitions. The test set has 360 posts from 37 unseen global groups. The 10 cross-dataset links were approved during adjudication, not by a further independent review of every linked page. Additional construction and label-provenance details are in Appendix B.

Table 1: Frozen global-event split. Labels follow the external post-label file.

Partition	Posts	Events	True	False
Train	1,297	127	973	324
Validation	221	19	175	46
Test	360	37	238	122
Total	1,878	183	1,386	492

Models, features, and metrics

We compare seven supervised configurations on the same frozen split and three training seeds (11, 29, 47). The text model uses extracted post text and topic with train-only TF–IDF/SVD; the structured model uses account and reply-summary fields from the predecessor GUI-processing record, including its credibility and evidence scores. The screenshot model trains a head on the frozen

source-post SigLIP 2 embedding (Tschannen et al., 2025). Ordinary fusion concatenates the four feature types, while typed fusion uses Equation (2); a no-interaction variant removes the three products. A pair-trained typed variant adds Equation (3). The ordinary and typed models share source inputs and optimization protocol but differ in parameter count. The fixed-rule predecessor is reported separately because its full-collection results were not rerun on this frozen split.

Macro-F1 is the primary held-out-event metric; we also report balanced accuracy, Brier score, and 15-bin expected calibration error (ECE). All reported metric values, including Brier (shown as $100\times$ its unit-scale score), use a 0–100% display scale and two decimal places. Differences between scores denote absolute points on that scale, not relative ratios. Table entries are means across three seeds; $\pm$ denotes between-seed sample standard deviation, not a confidence interval over events. Model selection uses validation Macro-F1 with Brier score as a tie-breaker and a fixed test threshold of 0.5. The appearance-pair generator produced JPEG recompression, rescaling with padding, contrast adjustment, and borders. The field audit, separate API extraction batch, and appearance test each use 120 held-out posts but are partially overlapping samples, not one shared cohort (Appendix C).

5 Results and Analysis

Which GUI-derived representation survives event shift?

Structured GUI fields, rather than the whole-screenshot embedding or typed fusion, perform best on unseen events (Table 2). They reach 71.67% Macro-F1 and 72.72% balanced accuracy, exceeding extracted text by 6.86% and ordinary four-source fusion by 5.21% absolute Macro-F1 points. The frozen post-screenshot vector alone reaches 54.01% Macro-F1. This contrast concerns *accessible representation*: account and reply-summary fields remain useful when test events are disjoint. The structured row includes scores derived by the predecessor's fixed rules, however, so it does not isolate the independent value of raw badge, avatar, or reply fields.

Table 2: Held-out global-event test (360 posts, 37 events). All metric values are percentages to two decimals; Brier is $100\times$ the unit-scale score. $\pm$ is seed standard deviation, not a 95% confidence interval. Lower Brier/ECE is better. Only the final row uses appearance-pair training.

Model	Macro-F1 ↑	BA ↑	Brier ↓	ECE-15 ↓	Params
Extracted text	64.81%±1.20%	64.44%	25.27%	20.82%	33,537
Structured GUI fields	**71.67%±0.30%**	**72.72%**	**18.21%**	8.47%	3,897
Frozen screenshot	54.01%±1.48%	59.81%	27.77%	15.63%	100,097
Ordinary four-source fusion	66.46%±0.81%	66.58%	23.07%	19.81%	137,017
Typed fusion, no pair loss	64.98%±3.04%	67.64%	21.38%	6.01%	128,185
Typed, no interactions	64.89%±1.35%	66.53%	21.62%	5.93%	103,225
Typed, with pair loss	63.41%±5.78%	66.99%	22.12%	8.29%	128,185

The source-typed architecture by itself does not explain the structured model's task advantage: its no-pair variant is 1.48% in absolute Macro-F1 points below ordinary fusion, and removing the three interactions changes Macro-F1 by only 0.09%. Adding the pair objective yields 63.41% Macro-F1, 1.57% below the no-pair typed model, with substantial seed variation (67.90%, 65.44%, and 56.89%). The no-pair typed model has a lower observed Brier score and ECE than ordinary fusion (21.38% versus 23.07% and 6.01% versus 19.81%), but these are descriptive secondary measurements. A raw-field-only control remains necessary to determine how much of the structured model's advantage comes from inherited rule-derived scores.

Where does the event-held-out gain occur?

The aggregate test score combines 237 PHEME posts with 123 Twitter15/16 posts. Table 3 recomputes Macro-F1 within each dataset from the same frozen per-post predictions, averaging the resulting scores across three seeds. Structured GUI fields outperform ordinary fusion in both slices, by 5.70% and 5.11% absolute points, respectively. Yet the comparison with text is not uniform: structured fields exceed text on PHEME by 14.92% points but trail it on Twitter15/16 by 3.16% points.

The screenshot-only branch is particularly weak on Twitter15/16. Thus, the pooled structured advantage should not be presented as a universal gain over text; it is concentrated in the larger PHEME slice, while the same GUI-processing record still compares favorably with the two four-source fusion variants within both datasets.

Table 3: Post-hoc dataset slices of the frozen global-event test. Each cell is within-slice Macro-F1, then averaged across three seeds; the slices have different event and label compositions. These values were not used for checkpoint selection.

Model	PHEME (237 posts)	Twitter15/16 (123 posts)
Extracted text	57.46%	**72.97%**
Structured GUI fields	**72.38%**	69.81%
Frozen screenshot	57.85%	40.32%
Ordinary four-source fusion	66.68%	64.70%
Typed fusion, no pair loss	66.99%	57.40%
Typed, no interactions	66.03%	58.34%
Typed, with pair loss	65.16%	55.37%

As a sensitivity check on the pooled comparison, we resampled the 37 held-out global events with replacement 2,000 times, preserving each selected event's posts and using the same resample for both models and all three seeds. The resulting exploratory 95% percentile interval for the structured-minus-ordinary Macro-F1 difference is $[+0.33\%, +9.03\%]$, around an observed 5.20%-point difference. For structured minus text it is $[-11.50\%, +13.04\%]$, despite the positive pooled point estimate. The interval width reflects substantial event heterogeneity and unequal group sizes, not additional training-seed uncertainty; one held-out group contains 163 posts. These post-hoc intervals should not be read as confirmatory evidence of a general population effect after inspecting the test set.

WHICH ERRORS ARE HIDDEN BY AGGREGATE MACRO-F1?

The two datasets also differ in the types of mistakes made by each input representation. Table 4 reports recall separately for false and true posts within each held-out slice. On PHEME, structured fields recover 73.91% of false posts, compared with 32.85% for extracted text, while giving up 5.76% points of true-post recall. On Twitter15/16, text has higher recall for both classes than the structured model. This class-level view explains why the structured advantage on PHEME is not merely a consequence of predicting the majority true class. It also reveals a failure mode of the frozen screenshot branch: its 80.50% false-post recall on Twitter15/16 coexists with only 18.10% true-post recall. The screenshot-only model therefore does not provide a balanced substitute for explicit post and response fields.

Table 4: Post-hoc class recall on held-out events, averaged across three seeds. PHEME contains 69 false and 168 true posts; Twitter15/16 contains 53 false and 70 true posts. These slices are descriptive and were not used for model selection.

	PHEME		Twitter15/16	
Model	False recall	True recall	False recall	True recall
Extracted text	32.85%	81.55%	69.81%	76.19%
Structured GUI fields	73.91%	75.79%	64.78%	74.76%
Frozen screenshot	75.36%	51.98%	80.50%	18.10%
Ordinary four-source fusion	55.07%	78.97%	58.49%	70.95%
Typed fusion, no pair loss	67.15%	71.63%	79.25%	41.90%

WHY EVALUATION PROTOCOL CHANGES THE INTERPRETATION

The representation ranking reverses between random-post and event-disjoint evaluation (Table 5). Under the random gate, ordinary fusion ranks first (84.23% Macro-F1), followed by typed fusion (83.69%); under the frozen event gate, structured GUI fields rank first (71.67%), followed by ordinary fusion (66.46%). The screenshot-only row changes from 81.18% to 54.01%, whereas the

structured row changes from 74.10% to 71.67%. Thus the protocol changes not only absolute scores but also which representation appears most useful.

Table 5: Macro-F1 under random-post and frozen global-event test protocols. All values are percentages with two decimals. The columns use different test posts and event distributions; their gap is not a causal estimate of leakage. Pair-trained typed fusion was not run under the random-post protocol.

Model	Random-post	Event-disjoint
Extracted text	80.25%	64.81%
Structured GUI fields	74.10%	71.67%
Frozen screenshot	81.18%	54.01%
Ordinary four-source fusion	84.23%	66.46%
Typed fusion, no pair loss	83.69%	64.98%
Typed, no interactions	82.73%	64.89%
Typed, with pair loss	—	63.41%

Because the two gates induce different training and test populations, subtracting their rows does not measure how much event leakage a model exploits. What the comparison does establish is a ranking reversal under the stricter question posed here: whether features learned from one set of events work on unseen events. The event-disjoint result, not the random-post score or the historical full-collection pilot, is therefore the basis for interpreting the proposed representation.

Extraction audit and the strength of the evidence

We manually checked 120 held-out posts against the GUI-processing record used for model inputs. The source-post region was correct in 120/120 cases and reply-view association in 81/82 evaluable cases. Badge and avatar categories agreed with human review in 87.50% and 70.00% of posts, while exact post-text agreement was only 51.67% (mean per-post normalized character error rate: 9.54%). Username-style agreement of 97.50% largely reflects a dominant normal-style category. The audit identifies weak fields that qualify mechanistic interpretations; it neither evaluates the separate API extraction batch nor shows whether correcting fields would change model rankings. Denominators and cohort overlaps appear in Appendix C.

Low-level appearance transformations

On a separate sample of 120 held-out-event posts (60 true and 60 false), we score ordinary, typed, and pair-trained typed fusion on the original screenshot and four transformations: JPEG quality 75, 90% scale-and-pad, contrast factor 0.90, and a 12-pixel border. The pair-trained model uses 800 train-partition variants. The authors report individual review of the evaluated variants; the provenance caveat is specified once in Appendix A. This is a narrow test of low-level presentation changes, not of changed reply evidence or cross-platform layout.

Table 6: Low-level appearance transformations on 120 sampled event-test posts. Metrics are percentages with two decimals, averaged across three seeds; these are not full-test scores.

Model	Original F1	Worst-transform F1	Flip rate	Worst-case Acc
Ordinary fusion	60.05%	60.12%	0.63%	60.56%
Typed, no pair loss	60.42%	59.21%	2.71%	56.94%
Typed, with pair loss	60.09%	59.47%	1.94%	59.17%

Define the paired flip-rate difference as pair-trained minus unpaired typed fusion. Its estimate from unrounded seed means is -0.76% in absolute points (the displayed 1.94% and 2.71% rates differ by 0.77% after separate rounding); the post-resampled, three-seed-averaged 95% percentile bootstrap interval is $[-2.36\%, +0.76\%]$. The interval is centered near the negative point estimate, crosses zero, and is not clustered by global event. Worst-transform Macro-F1 is 59.47%, only 0.26% above typed fusion without the loss and below ordinary fusion's 60.12%. The worst-transform score selects

the lowest-scoring transformation, while worst-case accuracy requires each post to be correct under every condition. These tests cover low-level encoding and decoration, not real theme, font, layout, or reply-exposure changes; they do not establish selective invariance.

Full-collection feature diagnostics

The GUI-Reason fixed-rule predictor motivated which extracted features to keep addressable. On the separate full collection, it reached 80.03% accuracy and 75.45% Macro-F1, versus 74.60% and 69.77% for its text-only configuration; removing identity weighting reduced accuracy to 74.65%. These diagnostics identify sensitive account and view-handling choices but are neither event-held-out learned-model results nor independent feature contributions. Appendix E specifies their feature provenance, and Appendix F reports the supporting ablations.

6 Related Work

Misinformation detection has used text, account and propagation signals (Castillo et al., 2011; Ma et al., 2016; Monti et al., 2019), attached images (Wang et al., 2018; Khattar et al., 2019), and social responses: dEFEND attends to comments, GCAN models retweet-user sequences, and GenFEND generates sparse early comments (Shu et al., 2019; Lu & Li, 2020; Nan et al., 2024). We instead ask how to organize the cues *visible in a particular rendered interface*, including which presentation changes should affect a prediction.

Screenshots have already been used for misinformation research: VizFake predicts source-domain trustworthiness from news-webpage appearance (Abdali et al., 2021). UGround and ShowUI ground GUI elements for agent tasks, not post-veracity prediction (Gou et al., 2025; Lin et al., 2025). Our setting joins source content, displayed account, appearance, and exposed responses at post level; the input-matched fusion comparison tests whether typing them improves prediction.

Robustness training such as AugMix encourages stable predictions under image corruption (Hendrycks et al., 2020). Our pair loss instead applies only when the post and visible evidence are preserved; changing exposed replies is excluded. The current low-level transformations do not establish robustness to wider GUI changes.

Evaluation design matters for PHEME and Twitter15/16: Wu & Hooi (2022) identify event- and source-related shortcuts and motivate event-separated tests. Our frozen split keeps all views and variants of each post together and compares typed with ordinary fusion on identical inputs. Event separation is a safeguard, not a novelty claim; unlike the historical full-collection pilot, this test anchors our generalization analysis.

7 Conclusion and Limitations

This paper frames GUI misinformation detection as learning from rendered observations whose content, identity, appearance, and exposure warrant different treatment. On a frozen 37-event test, structured GUI-derived fields reach 71.67% Macro-F1, versus 64.81% for extracted text and 54.01% for the frozen source-screenshot representation. This advantage does not transfer to the current four-source typed fusion: ordinary fusion reaches 66.46%, typed fusion 64.98%, and pair-trained typed fusion 63.41%. The separate low-level appearance test yields no reliable flip-rate gain from pair training. The original full-collection rule-based pilot remains useful for tracing how identity and view cues entered the design, but is not a comparable held-out-event result.

The collection is limited to X/Twitter-style renderings, and some identity and exposure features inherit predecessor scores. Frozen screenshot embeddings may encode semantics as well as presentation, so clean disentanglement cannot be claimed. The 120-post field audit finds imperfect text and avatar extraction; raw-field-only and oracle-field sensitivity analyses remain to be done (Appendix D). Tested transformations do not include theme or layout changes. Extension to other platforms and time-stamped exposure changes is future work.

ETHICS STATEMENT

Badges, avatars, and usernames can reflect platform policy and social bias; none is ground-truth evidence of truthfulness. Evaluation should stratify errors across account types and avoid presenting outputs as automated fact-checking. Screenshot sharing must respect post/commenter privacy, deletion requests, and platform terms.

REPRODUCIBILITY STATEMENT

The post is the prediction unit. Sections 4–5 and Appendices A–E document the label source, event split, feature provenance, model settings, extraction audit, and pair protocol. Anonymous code is provided at https://anonymous.4open.science/r/gui-misinformation-detection-C140/. Release of screenshots and annotations remains subject to platform and privacy constraints. The current test has been inspected during development; further tuning would require a fresh confirmation set.

AI USE STATEMENT

Generative AI tools assisted with literature discovery and searching for information during the related-work survey. The authors checked the cited sources, numerical claims, code outputs, and AI-assisted prose against the underlying records. The authors take responsibility for the final text, methods, code, and experimental results. AI use will also be disclosed in the submission form as required by ICLR 2027.

REFERENCES

Sara Abdali, Rutuja Gurav, Siddharth Menon, Daniel Fonseca, Negin Entezari, Neil Shah, and Evangelos E. Papalexakis. Identifying misinformation from website screenshots. *Proceedings of the International AAAI Conference on Web and Social Media*, 15(1):2–13, 2021. doi: 10.1609/icwsm.v15i1.18036.

Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. Information credibility on twitter. In *Proceedings of the 20th International Conference on World Wide Web*, pp. 675–684, 2011. doi: 10.1145/1963405.1963500.

Boyu Gou, Ruohan Wang, Boyuan Zheng, Yanan Xie, Cheng Chang, Yiheng Shu, Huan Sun, and Yu Su. Navigating the digital world as humans do: Universal visual grounding for GUI agents. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=kxnoqaisCT.

Dan Hendrycks, Norman Mu, Ekin D. Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=S1gmrxHFvB.

Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. MVAE: Multimodal variational autoencoder for fake news detection. In *The World Wide Web Conference*, pp. 2915–2921, 2019. doi: 10.1145/3308558.3313552.

Kevin Qinghong Lin, Linjie Li, Difei Gao, Zhengyuan Yang, Shiwei Wu, Zechen Bai, Stan Weixian Lei, Lijuan Wang, and Mike Zheng Shou. ShowUI: One vision-language-action model for GUI visual agent. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19498–19508, 2025.

Yi-Ju Lu and Cheng-Te Li. GCAN: Graph-aware co-attention networks for explainable fake news detection on social media. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 505–514, 2020. doi: 10.18653/v1/2020.acl-main.48.

Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J. Jansen, Kam-Fai Wong, and Meeyoung Cha. Detecting rumors from microblogs with recurrent neural networks. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pp. 3818–3824, 2016.

Jing Ma, Wei Gao, and Kam-Fai Wong. Detect rumors in microblog posts using propagation structure via kernel learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pp. 708–717, 2017. doi: 10.18653/v1/P17-1066.

Federico Monti, Fabrizio Frasca, Davide Eynard, Damon Mannion, and Michael M. Bronstein. Fake news detection on social media using geometric deep learning. *arXiv preprint arXiv:1902.06673*, 2019. doi: 10.48550/arXiv.1902.06673.

Qiong Nan, Qiang Sheng, Juan Cao, Beizhe Hu, Danding Wang, and Jintao Li. Let silence speak: Enhancing fake news detection with generated comments from large language models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 2024. doi: 10.1145/3627673.3679519. URL https://arxiv.org/abs/2405.16631.

Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. dEFEND: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 395–405, 2019. doi: 10.1145/3292500.3330935.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. doi: 10.48550/arXiv.2502.14786. URL https://arxiv.org/abs/2502.14786.

Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. EANN: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 849–857, 2018. doi: 10.1145/3219819.3219903.

Jiaying Wu and Bryan Hooi. Probing spurious correlations in popular event-based rumor detection benchmarks. *arXiv preprint arXiv:2209.08799*, 2022. doi: 10.48550/arXiv.2209.08799. URL https://arxiv.org/abs/2209.08799.

Arkaitz Zubiaga, Maria Liakata, and Rob Procter. Exploiting context for rumour detection in social media. In *Social Informatics*, pp. 109–123. Springer, 2017. doi: 10.1007/978-3-319-67217-5_8.

A Learned-Model Feature and Training Details

Content. The implementation fits a training-only word unigram/bigram TF–IDF vocabulary (at most 30,000 terms), reduces it by truncated SVD to at most 256 dimensions, and standardizes the result. The input text is the extracted post text plus a topic descriptor, not a jointly trained language encoder.

Identity. Badge, avatar, username-style, and predecessor credibility-level categories are one-hot encoded. The predecessor credibility score is median-imputed and standardized. Since the last two fields are rule-derived, the final study requires a raw-attributes-only control.

Exposure. The vector contains relevant, likes/popular, and recent comment counts plus predecessor verification, consensus, social-ratio, and temporal scores. The current learned classifier does not embed every comment screenshot. The schematic's Popular and implementation's Likes refer to the same captured view label, not a verified claim about a platform ranking algorithm.

Appearance. The frozen checkpoint is `google/siglip2-base-patch16-512` (Tschannen et al., 2025), applied to the source-post screenshot. The transformation generator performs JPEG recompression, 90% rescaling with padding, contrast adjustment, and a 12-pixel border. These are rendering/encoding perturbations, not actual theme or font changes. The authors confirm that all 1,280 variants were individually checked after generation. The existing machine-readable audit has not been synchronized with that review: 200 variants are marked as manually checked and 1,080 retain provisional automatic-preservation tags. The authors' completed inspection and its incomplete recorded provenance must be reconciled before artifact release.

Optimization. The code defaults to AdamW (learning rate 10^{-3}, weight decay 10^{-4}), at most 100 epochs, early-stop patience 12, and seeds 11, 29, and 47. Checkpoint choice prioritizes validation Macro-F1 and breaks ties with Brier score. The ordinary baseline shares inputs and schedule but is not automatically parameter-count matched.

B Dataset and Event Protocol

The frozen manifest contains 1,878 distinct source posts: 1,290 from PHEME and 588 from Twitter15/16. The external post-label file is authoritative for the seven supervised configurations; its 1,386 true and 492 false labels differ from labels embedded in the predecessor's records in 296 cases. Event candidates underwent human adjudication, including 10 approved cross-dataset links, and yield 183 global event groups. All screenshots, reply views, and rendered variants belonging to a post stay in its event partition. The frozen split has 127/19/37 groups in train/validation/test, respectively; Table 1 gives class counts. Exact post-to-event and split assignments are recorded in the frozen manifest for artifact release. The 10 cross-dataset links were retained after adjudication; they were not each rechecked by a separate independent page-level review.

C Held-Out GUI-Field Audit

The field audit reviewed 120 event-test posts against the GUI-processing record used for model inputs, not against a separately run 120-post API extraction batch. A third 120-post sample supports the appearance test in Section 5. These are distinct but partially overlapping sets: field-audit/API overlap is 15 posts, field-audit/appearance overlap is 51, API/appearance overlap is 10, and five posts occur in all three. Two reviewers each handled 60 different audit posts; the agreement below compares automatic records with human judgments, not inter-annotator agreement. Post-text exact-match rate is 51.67%, and mean per-post normalized character error rate is 9.54% (not corpus-level CER). Badge-category agreement is 87.50% (Macro-F1 68.19%), avatar-category agreement is 70.00% (Macro-F1 22.77%), and username-style agreement is 97.50% (Macro-F1 49.37%; Cohen's $\kappa = 0$). For username style, human review marked 117/120 normal and the automatic record marked all 120 normal, explaining the high agreement. The source-post region was judged correct in 120/120 cases. Reply-view association was correct in 81/82 evaluable cases; 38/120 cases without a relevant

reply screenshot were not placed in this denominator. These checks do not audit every reply's text or establish the effect of corrected fields on model predictions.

D Remaining Controls

Raw-GUI-field-only and human-corrected-field sensitivity controls, an event-clustered comparison of model differences, and a same-split rerun of the predecessor fixed-rule system are not reported here. A timestamped, complete reply pool is unavailable for an earliest-versus-latest exposure experiment. The current held-out test has already been inspected, so future model selection cannot treat it as untouched confirmation data.

E Provenance of GUI-Derived Structured Features

This appendix makes the structured inputs in Section 3 auditable. The GUI feature builder extracts categories at L1, assigns account scores at L2, computes comment signals at L3, and summarizes captured views at L4; its identity and exposure scores are reused as inputs to the learned classifier. L5 is the earlier fixed-rule *decision* layer, included only to interpret the full-collection diagnostic in Appendix F; the learned classifier does not use its final decision. All listed weights, thresholds, and bonuses are hand-set design values, not calibrated credibility measures.

L1: Screenshot Extraction

The GUI feature builder uses frozen Qwen3-VL-4B-Instruct with structured visual prompts. Its record stores a verification-badge category (gold, gray, blue, or none), avatar and username categories, and source-post and comment text. The available artifacts do not establish a separate badge-color confirmation stage before category assignment; badge color is therefore an extracted category, not an independently confirmed observation. View labels come from the capture protocol.

L2: Account Features and Scores

L2 assigns scores to three visible account attributes: verification badge b, avatar type a, and username style u. Table 7 gives the hand-set category values $s_B(b)$, $s_A(a)$, and $s_U(u)$. The category descriptions refer to what appears in the screenshot; their scores are design assumptions, not validated measures of source credibility.

Table 7: Account-category scores carried into the structured identity input. Values are hand-set and unitless, not measured credibility probabilities.

Attribute	Visually assigned category	Score
Verification badge	Gold or gray	1.00
	Blue	0.50
	No visible badge	0.10
Avatar	Organization logo or person photograph	0.60
	Illustration or object image	0.40
	Text/banner or other unclear image	0.20
	Platform default image	0.00
Username style	Normal	1.00
	Repeated pattern	0.50
	Random-looking or unusual	0.00

An organization-logo avatar without a visible badge receives 0.40 instead of its default avatar score of 0.60. The adjusted avatar score and combined identity score are

$$\widetilde{s}_A(a,b) = \begin{cases} 0.40, & \text{if } a \text{ is an organization-logo avatar and } b \text{ has no visible badge,} \\ s_A(a), & \text{otherwise,} \end{cases} \tag{4}$$

$$c(b,a,u) = 0.65\, s_B(b) + 0.15\, \widetilde{s}_A(a,b) + 0.20\, s_U(u).$$

The score $c \in [0, 1]$ is discretized as

$$\text{level}(c) = \begin{cases} \text{High}, & c \geq 0.70, \\ \text{Medium}, & 0.40 \leq c < 0.70, \\ \text{Low}, & 0.15 \leq c < 0.40, \\ \text{Very Low}, & c < 0.15. \end{cases} \tag{5}$$

The larger badge weight and conditional avatar adjustment are hand-set choices; the badge ablation in Appendix F measures the fixed pipeline's sensitivity to them. A separate multiplier for comments from badged accounts with organization-logo avatars is applied at L4, not in Equation (4).

The feature builder also uses a VLM-assessed identity–claim mismatch indicator before computing comment signals. When raised, the risk stage can reduce the identity score's influence. This heuristic is distinct from the deterministic category mapping above; no validated decision boundary is claimed.

L3: Comment Signals for Exposure Features

L3 processes each captured comment through five operations in order: filtering, semantic encoding, risk estimation, identity-conditioned weighting, and bounded signal computation.

- **Module A** (Noise filter): Filters empty, overly short, or mention-only comments.
- **Module B** (Semantic encoding): Uses a VLM to estimate stance, comment function, evidence strength, and an associated confidence value. The confidence is not assumed to be statistically calibrated.
- **Module C** (Risk estimation): Produces a sigmoid-based score from sarcasm, uncertainty, and post-level risk factors.

Module D (identity-conditioned weighting). For comment k, let $w_k^{(0)}$ denote its base weight, c_k the L2 identity score of its author, e_k the evidence-strength multiplier, and q_k the extractor's confidence value. The feature rule computes

$$w_k = w_k^{(0)}(0.4 + 0.6c_k)e_k q_k. \tag{6}$$

The identity factor ranges from 0.4 to 1.0 as c_k varies from 0 to 1; it attenuates but does not by itself remove a comment signal.

Module E (bounded comment signal). Let r_k be the Module C risk score and d_k the stance adjustment. The final comment signal is

$$s_k = \text{clip}_{[0,1]}(w_k(1 - r_k)(1 + d_k)), \tag{7}$$

where clipping restricts the output to $[0, 1]$. The factors e_k, q_k, r_k, and d_k are extracted values or hand-set multipliers; Equations (6)–(7) describe feature construction, not calibrated evidential probabilities.

L4: Captured-View Exposure Summaries

The feature builder computes separate summaries for the captured Relevant, Likes/Popular, and Recent views. These are capture labels; the platform's exact selection algorithms are not observed. The view-specific rules are:

- **Relevant-view evidence summary** (4a): Aggregates comment weights from L3. A comment from a badged account with an organization-logo avatar receives a hand-set $1.2\times$ multiplier. Weight thresholds of 0.50 for extraordinary claims and 0.35 for other claims distinguish stronger from weaker inputs. The rule counts at least three social-confirmation comments as a weak contextual cue; neither the badge nor the comments independently verify a claim.
- **Likes/Popular-view consensus summary** (4b): Uses a weighted difference between supporting and denying comments. It also records the share of social-proof comments and whether fact-checking comments are present for a later heuristic penalty.

- **Recent-view trend summary** (4c): Uses exponential decay with factor 0.9 to give later captured comments more weight and summarizes their supporting or denying signals. The capture order is not necessarily the full chronological response history.

Cross-view fusion (4d–4e) uses hand-set weights $(w_v, w_c, w_t) = (0.50, 0.25, 0.25)$ for high-risk posts and $(0.35, 0.35, 0.30)$ otherwise. The fused score is capped at 0.30 when the rule finds no evidence, 0.65 for weak evidence, and 0.35 for a deny-dominated rebuttal. A verification gap with a low-scoring author caps it at 0.40, relaxed to 0.50 for a High-level author. If social-proof comments exceed 60% with no fact-checking comment, the consensus summary is multiplied by 0.75. These are rule settings, not empirically calibrated reliability bounds.

L5: Prior Decision Rule, Excluded from the Learned Classifier

The earlier fixed-rule predictor applies a hand-set threshold and an ordered decision chain to the features above. This decision is not an input to the learned classifier; it is documented to interpret Appendix F.

Threshold selection. The threshold τ depends on the author's L2 level and the captured comment count:

- Few comments (< 3): $\tau = 0.45$ (High), 0.55 (Medium), 0.60 (Low/Very Low)
- Sufficient comments (≥ 3): $\tau = 0.50$ (High/Medium), 0.55 (Low/Very Low)

These values are not uniformly monotone in comment count (for example, High-level authors have thresholds of 0.45 and 0.50 in the two count regimes). They should not be read as measured reliability of an account or reply pool.

Rule order. Rule B is a terminal override; Rules A and D modify the score before Rule C's positive-decision checks, followed by a default false decision. The historical order is:

1. **Rule B — Recent-denial override**: A recent high-weight denying comment causes an immediate false prediction under this historical rule, regardless of other inputs.
2. **Rule A — Extraordinary claim without evidence**: If a claim is categorized as extraordinary and no evidence is recorded, the global score is multiplied by 0.60 before threshold comparison.
3. **Rule D — Low-consistency downgrade**: If cross-view consistency is low and a verification gap is flagged, the global score is multiplied by 0.65.
4. **Rule C — Positive evidence conditions**: The post is classified as **true** if any of the following graded conditions hold: (i) strong evidence with verification score ≥ 0.85; (ii) strong evidence with adjusted score meeting the adaptive threshold; (iii) adjusted score meeting the threshold regardless of evidence strength; (iv) weak evidence with adjusted score near threshold ($\tau - 0.10$) and non-low consistency; (v) no evidence (non-extraordinary) with adjusted score near threshold ($\tau - 0.15$) and non-low consistency; (vi) any evidence level with adjusted score near threshold ($\tau - 0.05$) and high consistency.
5. **Default**: If no rule fires, the post is classified as **false**.

This order is a hand-specified policy, not an established evidential principle. Its effect has not been isolated against alternative rule orders or a fixed-threshold control.

F Full-Collection Diagnostics for GUI Feature Selection

This section tests which account and captured-view choices motivated the typed inputs in the main study. Tables 8–10 concern the full-collection fixed-rule predictor defined in Appendix E, not the learned classifier evaluated on held-out events. They diagnose sensitivity to available GUI fields but cannot establish event generalization or independent information in each field. Prompted-model rows are contextual, not training-, prompt-, or capacity-matched controls. Accuracy differences below are absolute points on the displayed percentage scale.

CONTEXTUAL BASELINES AND FEATURE MOTIVATION

Table 8: Full-collection prototype versus baselines. All metrics are percentages with two decimals.

Method	Acc	Prec	Rec	F1	Macro F1	Bal Acc
Majority BL	73.79%	73.79%	100.00%	84.92%	42.46%	50.00%
E2E VLM	35.62%	62.38%	32.18%	42.46%	34.70%	38.75%
DeepSeek-R1	38.18%	68.60%	29.94%	41.69%	37.95%	45.66%
Qwen2.5-7B	26.52%	71.43%	0.72%	1.43%	21.43%	49.95%
Text-only	74.60%	86.56%	77.63%	81.86%	69.77%	71.85%
GUI-Reason	**80.03%**	**88.80%**	**83.48%**	**86.05%**	**75.45%**	**76.90%**

On this full-collection protocol, the fixed-rule predictor scores 80.03% accuracy and 75.45% Macro-F1, versus 74.60% and 69.77% for its text-only configuration: displayed gaps of 5.43% and 5.68%. This motivates retaining GUI-derived account and response-view fields as inspectable inputs. The majority baseline's 73.79% accuracy reflects class balance; its 42.46% Macro-F1 exposes poor minority-class performance. The prompted-model rows score lower under their original prompts, but are not controlled tests of model capacity or supervised end-to-end GUI learning. Only the frozen event-disjoint comparison in Table 2 tests the learned configurations on unseen event groups.

VIEW-POLICY AND OVERALL-IDENTITY SENSITIVITY

Table 9: Full-collection feature and view-policy variants. ΔAcc is the difference from the full row using displayed accuracy values; all metrics are percentages with two decimals.

Method	Acc	F1	Macro F1	Bal Acc	ΔAcc
Identity ablation	74.65%	81.79%	70.05%	72.41%	-5.38%
Relevant-only	78.22%	84.21%	74.55%	77.77%	-1.81%
Likes-only	77.69%	83.82%	73.96%	77.15%	-2.34%
Recent-only	77.10%	83.45%	73.16%	76.10%	-2.93%
Naive fusion	78.22%	84.31%	74.36%	77.25%	-1.81%
GUI-Reason	**80.03%**	**86.05%**	**75.45%**	**76.90%**	—

Removing identity weighting changes accuracy from 80.03% to 74.65%, a 5.38% absolute-point drop. Relevant-only, Likes-only, and Recent-only reach 78.22%, 77.69%, and 77.10% accuracy, while naive fusion reaches 78.22%. Yet Relevant-only, Likes-only, and naive fusion have higher balanced accuracy than the full rule (77.77%, 77.15%, and 77.25% versus 76.90%) despite lower accuracy and Macro-F1. The variants thus change classwise sensitivity rather than uniformly worsening performance. They also change downstream hand-set rules, so they do not isolate independent information in each view; capture labels do not establish exposure order or a counterfactual reply pool. The learned model keeps view summaries addressable without claiming that this diagnostic proves complementarity under event shift.

ACCOUNT-ATTRIBUTE SENSITIVITY

Table 10: Full-collection account-attribute ablations. ΔAcc is computed from the displayed full-row accuracy; all metrics are percentages with two decimals.

Method	Acc	F1	Macro F1	Bal Acc	ΔAcc
Badge Removal	67.25%	74.08%	64.81%	70.73%	-12.78%
Avatar Removal	79.77%	85.85%	75.17%	76.66%	-0.26%
Username Removal	79.23%	85.35%	74.85%	76.75%	-0.80%
GUI-Reason (full)	80.03%	86.05%	75.45%	76.90%	—

Badge removal produces the largest change: accuracy falls to 67.25%, 12.78% below the full rule system. Removing avatar or username alone changes accuracy by 0.26% and 0.80%, respectively. Badge removal is also more damaging than removing identity weighting altogether. That contrast is consistent with interactions among remaining cues and the hand-set credibility, bonus, and threshold rules; it does not prove that avatar or username becomes intrinsically harmful without a badge. These differences are neither additive mutual-information contributions nor causal interventions. The held-out field audit in Appendix C shows different badge and avatar extraction errors. A raw-field-only learned ablation on the frozen event split would be needed to test the interaction independently of inherited scores.

An earlier 2,000-configuration scalar search selected its top settings by test accuracy. Its best-test score is therefore not an independent generalization estimate and is excluded here. Tables 8–10 serve only as descriptive feature diagnostics, not validation of learned typing or event transfer.