SAFE-ALIGN: ALIGNMENT-CONDITIONED EVIDENCE ROUTING FOR MULTIMODAL MANIPULATION DETECTION AND GROUNDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal manipulation detection and grounding aim to identify alterations in image–text pairs and localize the affected image regions and text tokens. A central challenge is that semantic correspondence does not reliably indicate authenticity: authentic pairs may be weakly aligned, while subtle manipulations can preserve strong global alignment. Detecting these alterations therefore requires interpreting cross-modal inconsistencies alongside local visual and linguistic cues. However, existing approaches often aggregate these complementary cues without explicitly accounting for how image–text alignment should affect their relative contributions. To address this limitation, we propose SAFE-Align, a framework that uses sample-level alignment to guide evidence aggregation. The key idea is to treat alignment as context for deciding which evidence to emphasize, rather than as a direct indicator of manipulation. Specifically, SAFE-Align combines a learned representation of discrete alignment levels with continuous image–text similarity to condition a gate that adaptively weights global cross-modal, local spatial, and text-only evidence. To localize subtle edits that global representations may overlook, it further models bidirectional correspondence between image patches and text tokens, extracting local agreement and discrepancy cues for grounding. On the official DGM4 benchmark split, SAFE-Align achieves 95.63% AUC for manipulation detection and 89.63% mAP for manipulation-type recognition, surpassing ASAP by 1.25 and 1.10 percentage points, respectively. Ablation studies show that alignment-conditioned routing improves AUC by 1.55 points over uniform fusion and 1.97 points over a learned gate without alignment conditioning, supporting the value of explicitly incorporating image–text correspondence into evidence aggregation.

1 INTRODUCTION

Multimodal media manipulation is difficult to detect because realistic visual content and fluent text can preserve surface plausibility while altering factual meaning. Existing image-forensics datasets focus primarily on visual artifacts, whereas multimodal news benchmarks additionally include semantic inconsistencies caused by altered captions or mismatched image–text context (Rossler et al., 2019; Li et al., 2020c; Dolhansky et al., 2020; Liu et al., 2021; Luo et al., 2021a; Aneja et al., 2023; Abdelnabi et al., 2022). The DGM4 benchmark makes this setting explicit: for each image–text pair, a model must predict authenticity, recognize the manipulation type, localize manipulated image regions, and identify manipulated text tokens (Shao et al., 2023; 2024). Consequently, success requires not only detecting whether a pair is suspicious but also determining which evidence sources are most informative for a given sample and where the inconsistency occurs.

Many existing methods employ predetermined multimodal interaction and fusion architectures whose aggregation rule is not explicitly conditioned on the sample-level image–text alignment regime (Wang et al., 2018; Khattar et al., 2019; Shao et al., 2023; Li et al., 2024; Liu et al., 2025; Zhang et al., 2025). Such alignment-agnostic aggregation can be restrictive when cross-modal correspondence is heterogeneous across samples. In particular, a weakly aligned pair is not necessarily manipulated, and a locally edited pair may retain high global semantic similarity. Thus, although global similarity alone is not a manipulation score, it can reliably reveal the correspondence regime

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



Figure 1: Examples from the DGM4 training split ordered by image–text alignment. Values above each pair are cosine similarities computed with CLIP ViT-B/16. The examples motivate using similarity to characterize correspondence regimes rather than treating it as an authenticity score.

in which downstream evidence should be interpreted. This observation motivates a different formulation of multimodal fusion: *Rather than asking only how to combine image and text, we investigate whether the estimated alignment state can be used to determine the relative contribution of each evidence source.*

Figure 1 illustrates this motivation with examples from the DGM4 training split. As shown, authentic and manipulated pairs can appear at different image-text similarity levels. Consequently, applying the same fusion rule to all samples can mix distinct regimes: global interaction may be reliable for strongly aligned pairs, local spatial evidence may become more informative when small image edits preserve overall semantics, and text-only evidence may be necessary when the image remains visually plausible. Accordingly, we use the alignment state as a sample-level routing prior rather than as a stand-alone decision signal.

Operationalizing this idea requires addressing three coupled issues. First, continuous global similarity must be represented in a way that preserves its ordering while making distinct correspondence regimes explicit. Second, global cross-modal interaction, local spatial cues, and unimodal linguistic evidence should remain available as complementary sources rather than being collapsed prematurely. Third, manipulation grounding requires fine-grained patch–token relationships because small edits may be diluted by global fusion. Together, these requirements motivate a design that performs sample-level evidence routing while preserving dense local signals for localization.

We propose **SAFE-Align**, an alignment-conditioned framework for multimodal manipulation detection and grounding. SAFE-Align first maps global image–text similarity to a discrete alignment level with a learned state embedding while retaining the original continuous score. The resulting alignment prior conditions a gate over global-interaction, local-spatial, and text-only support streams. A separate fine-grained branch then performs bidirectional patch–token retrieval and constructs agreement/discrepancy features for image and text grounding. This design separates *which evidence source to emphasize* from *where the manipulation occurs*.

Our contributions are threefold:

- We formulate cross-modal alignment as a *routing variable* for multimodal manipulation reasoning. A similarity-graded alignment prior combines a discrete correspondence level with the continuous similarity score, providing sample-dependent context without treating similarity itself as an authenticity label.

- We introduce alignment-conditioned multi-source evidence routing that keeps global-interaction, local-spatial, and text-only streams explicit and adaptively weights their source-specific class support. This formulation tests whether heterogeneous alignment should change how evidence is aggregated, rather than merely adding another fused feature.

- We develop bidirectional patch–token grounding features that combine native, retrieved, absolute-difference, and element-wise-product representations. On DGM4, SAFE-Align improves binary detection and multi-label recognition over strong baselines, while the ablations show that the alignment prior provides gains beyond uniform fusion and a learned gate without alignment conditioning.

2 Related Work

2.1 Vision–Language Alignment

Vision–language research spans transformer-based cross-modal fusion and unified or contrastive pre-training (Tan & Bansal, 2019; Chen et al., 2020; Li et al., 2020b; Kim et al., 2021; Li et al., 2021b; Radford et al., 2021; Singh et al., 2022; Li et al., 2022a; 2023), as well as fine-grained token–region alignment and grounding (Yao et al., 2022; Kamath et al., 2021; Li et al., 2022b; Zhong et al., 2022). Benchmarks such as Winoground (Thrush et al., 2022) further probe compositional image–text matching. Most of these methods optimize alignment as a representation-learning objective or evaluate whether paired content corresponds. AlignGPT (Zhao et al., 2024) further models heterogeneous alignment degrees through similarity groups and embeddings. SAFE-Align adopts the general notion of heterogeneous correspondence for a different purpose: rather than improving alignment capability itself, it uses the estimated alignment state as routing context to modulate the relative contribution of global, local, and text-only evidence across samples.

2.2 Multimodal Manipulation Detection and Grounding

Image forensics and misinformation detection have studied visual artifacts, fake-news classification, and out-of-context image use (Li et al., 2020a; 2021a; Luo et al., 2021b; Rossler et al., 2019; Li et al., 2020c; Dolhansky et al., 2020; Shu et al., 2017; Wang, 2017; Wang et al., 2018; Khattar et al., 2019; Liu et al., 2021; Luo et al., 2021a; Aneja et al., 2023; Abdelnabi et al., 2022). DGM4 extends this setting by jointly evaluating authenticity classification, manipulation-type recognition, image-region grounding, and token-level text grounding (Shao et al., 2023; 2024). Subsequent methods explore vision–language knowledge interaction and frequency-aware modeling (Li et al., 2024; Liu et al., 2025), while ASAP (Zhang et al., 2025) strengthens fine-grained semantic alignment with auxiliary alignment supervision and manipulation-guided local cues. These methods establish the importance of both global consistency and local evidence. SAFE-Align focuses on a complementary question: whether the *way evidence is fused* should depend on the sample's image–text alignment state.

2.3 Multi-Source Evidence Fusion

Evidence theory, subjective logic, and evidential neural classifiers provide formal mechanisms for representing complementary or conflicting support under uncertainty (Dempster, 1967; Shafer, 1976; Jøsang, 2016; Sensoy et al., 2018; Han et al., 2023). SAFE-Align does not estimate calibrated epistemic uncertainty and does not apply Dempster–Shafer or Dirichlet fusion. Instead, it uses a narrower support-routing mechanism: separate representation streams produce non-negative class-support vectors, and the alignment prior conditions a learned gate over these sources. Throughout this paper, *evidence* denotes task-relevant support contributed by a source rather than a calibrated uncertainty mass.

3 Methodology

3.1 Overview

Given an image I and associated text T, SAFE-Align separates multimodal reasoning into three stages: (i) estimating the global correspondence regime, (ii) routing complementary evidence under that regime, and (iii) constructing dense patch–token discrepancy cues for grounding. The visual and text encoders first produce global tokens and local features:

$$\mathbf{V} = [\mathbf{v}_{cls}; \mathbf{V}_{pat}], \qquad \mathbf{T} = [\mathbf{t}_{cls}; \mathbf{T}_{tok}]. \tag{1}$$

The global tokens are projected into a shared normalized space:

$$\mathbf{f}_I = \mathrm{Norm}(W_I \mathbf{v}_{cls}), \qquad \mathbf{f}_T = \mathrm{Norm}(W_T \mathbf{t}_{cls}), \tag{2}$$

The normalized global features define the similarity signal used to construct $\mathbf{h}_{align}$. In parallel, a BERT-based fusion encoder produces the global interaction representation $\mathbf{h}_{global}$, while a shallow image-dominant cross-attention branch produces local patch features $\mathbf{H}_{local}^{pat}$ and a pooled representation $\mathbf{h}_{local}$. A text-only representation $\mathbf{h}_{text}$ preserves linguistic evidence that may remain

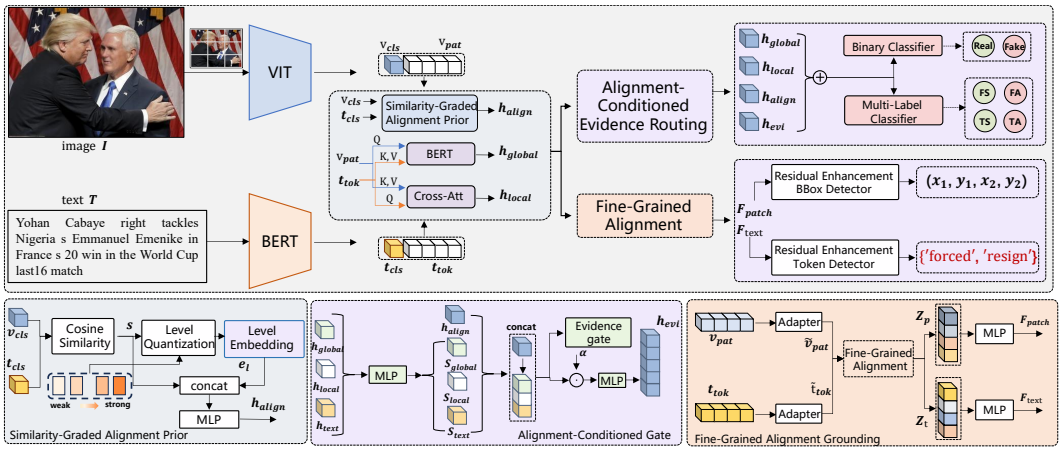


Figure 2: Framework of SAFE-Align. Global similarity is converted into alignment states, which condition the routing of global, local, and text-only evidence. Bidirectional patch–token retrieval then supplies residual signals for image and text grounding.

informative when the image is visually plausible. The alignment prior conditions the routing of these sources for classification, while dense patch and token features are retained for grounding. Figure 2 summarizes the architecture.

3.2 Similarity-Graded Alignment Prior

Because weak alignment can occur in authentic pairs and local manipulation can preserve high global similarity, the similarity score should not be interpreted as a direct authenticity predictor. Instead, we use it only to characterize the sample's correspondence regime, which subsequently provides context for evidence routing. For normalized global features, the similarity is

$$s = \mathbf{f}_I^\top \mathbf{f}_T. \tag{3}$$

Given thresholds $\Theta = \{\theta_1, \ldots, \theta_{K-1}\}$ estimated from the training similarity distribution, the alignment level is

$$l = \mathrm{Quantize}(s; \Theta), \qquad l \in \{0, \ldots, K-1\}. \tag{4}$$

Each discrete level is associated with a learned embedding $\mathbf{e}_l$, and the continuous score is retained when constructing the final prior:

$$\mathbf{h}_{align} = \phi_{align}([\mathbf{e}_l; s]). \tag{5}$$

The thresholds are estimated from the training similarity distribution and then fixed; validation and test samples therefore do not influence the alignment partition. The reported model uses four alignment levels; the effect of different alignment granularities is analyzed in Section 5. The projection ϕ_{align} contains a linear layer, layer normalization, and GELU. Combining $\mathbf{e}_l$ with the raw score s retains within-level ordering while providing a learnable representation of the broader alignment context. The embedding and projection parameters are optimized jointly with the downstream objectives.

3.3 Alignment-Conditioned Evidence Routing

The fusion encoder computes

$$\mathbf{H}_{fusion} = \mathrm{BERT}_{fusion}(\mathbf{T}, \mathbf{V}), \qquad \mathbf{h}_{global} = \mathbf{H}_{fusion}[0]. \tag{6}$$

In parallel, local cross-attention is

$$\mathbf{H}_{local}^{img} = \mathbf{V} + \mathrm{CrossAttn}(\mathrm{LN}(\mathbf{V}), \mathrm{LN}(\mathbf{T}), \mathrm{LN}(\mathbf{T})), \tag{7}$$

where the patch positions form $\mathbf{H}_{local}^{pat}$ and a learnable query pools the local features into $\mathbf{h}_{local}$. A masked text average gives

$$\mathbf{h}_{text} = \frac{\sum_{i=1}^{L-1} m_i \mathbf{t}_i}{\sum_{i=1}^{L-1} m_i + \epsilon}. \tag{8}$$

The three sources produce two-dimensional, non-negative class-support vectors:

$$\mathbf{S}_j = \mathrm{Softplus}(g_j(\mathbf{h}_j)), \quad j \in \{global, local, text\}. \tag{9}$$

The alignment prior conditions their mixture:

$$\boldsymbol{\alpha} = \mathrm{Softmax}\big(g_{gate}([\mathbf{S}_{global}; \mathbf{S}_{local}; \mathbf{S}_{text}; \mathbf{h}_{align}])\big), \tag{10}$$

$$\mathbf{S}_{fused} = \alpha_1 \mathbf{S}_{global} + \alpha_2 \mathbf{S}_{local} + \alpha_3 \mathbf{S}_{text}, \qquad \mathbf{h}_{evi} = \phi_{evi}(\mathbf{S}_{fused}). \tag{11}$$

These support vectors serve as routing features rather than calibrated uncertainty estimates or Dirichlet parameters. They determine the relative contribution of each source before the fused support is injected into the task representations. The binary and multi-label branches use the following residual representations:

$$\mathbf{h}_{bic} = \mathbf{h}_{global} + \beta_b \mathbf{h}_{local} + \mathbf{h}_{align} + \mathbf{h}_{evi}, \qquad \mathbf{h}_{mlc} = \mathbf{h}_{global} + \beta_m \mathbf{h}_{local} + \mathbf{h}_{align} + \mathbf{h}_{evi}. \tag{12}$$

Because the gate receives the three support vectors together with $\mathbf{h}_{align}$, routing depends on both the current evidence and the estimated alignment regime. The global stream summarizes pair-level interaction, the local stream emphasizes spatially grounded cross-modal cues, and the text stream preserves linguistic evidence that may be decisive when visual content remains plausible. The routed support is injected through residual paths rather than replacing the original features, allowing alignment conditioning to modulate evidence without discarding the backbone representations.

3.4 Fine-Grained Patch–Token Grounding

Sample-level routing does not identify small local edits. SAFE-Align therefore retains dense patch and token features and adapts them with zero-initialized bottleneck adapters:

$$\mathrm{Adapter}(\mathbf{x}) = \mathbf{x} + W_{up}\sigma(W_{down}\mathbf{x} + \mathbf{b}_{down}) + \mathbf{b}_{up}. \tag{13}$$

The adapted features define a dense patch–token affinity and bidirectional retrieved contexts:

$$\mathbf{A} = \frac{\mathrm{Norm}(W_v \tilde{\mathbf{V}}_{pat})\,\mathrm{Norm}(W_t \tilde{\mathbf{T}}_{tok})^\top}{\tau}, \tag{14}$$

$$\mathbf{P}_{v\to t} = \mathrm{Softmax}_{token}(\mathbf{A} + \mathbf{M}_t)\tilde{\mathbf{T}}_{tok}, \qquad \mathbf{P}_{t\to v} = \mathrm{Softmax}_{patch}(\mathbf{A}^\top + \mathbf{M}_v)\tilde{\mathbf{V}}_{pat}. \tag{15}$$

For each modality, we concatenate the native feature, retrieved context, absolute difference, and element-wise product:

$$Z_p = [\tilde{V}_{\mathrm{pat}}; P_{v\to t}; |\tilde{V}_{\mathrm{pat}} - P_{v\to t}|; \tilde{V}_{\mathrm{pat}} \odot P_{v\to t}], \tag{16}$$

$$Z_t = [\tilde{T}_{\mathrm{tok}}; P_{t\to v}; |\tilde{T}_{\mathrm{tok}} - P_{t\to v}|; \tilde{T}_{\mathrm{tok}} \odot P_{t\to v}]. \tag{17}$$

Lightweight MLPs produce fine-grained features, prediction residuals, and modality-specific gates:

$$F_{\mathrm{patch}} = \phi_p(Z_p), \qquad \hat{m}_{\mathrm{patch}} = g_p(F_{\mathrm{patch}}), \qquad \gamma_I = \sigma(g_I(\mathrm{Pool}(F_{\mathrm{patch}}))), \tag{18}$$

$$F_{\mathrm{text}} = \phi_t(Z_t), \qquad \Delta Y_{\mathrm{tok}} = g_t(F_{\mathrm{text}}), \qquad \gamma_T = \sigma(g_T(\mathrm{Pool}(F_{\mathrm{text}}))). \tag{19}$$

Here $\gamma_I, \gamma_T \in [0, 1]$ gate the image- and text-side residual branches. Let $Y_{\mathrm{tok}}^{\mathrm{base}}$ denote the logits from the base text-grounding branch. The final token logits are

$$\hat{Y}_{\mathrm{tok}} = Y_{\mathrm{tok}}^{\mathrm{base}} + \lambda_t \gamma_T \Delta Y_{\mathrm{tok}}. \tag{20}$$

For image grounding, the locator uses

$$X_{\mathrm{bbox}} = [V_{\mathrm{pat}}; H_{\mathrm{local}}^{\mathrm{pat}}; F_{\mathrm{patch}}; \sigma(\hat{m}_{\mathrm{patch}}); P_{xy}], \qquad \hat{b} = \mathrm{Clamp}(b_{\mathrm{coarse}} + \lambda_v \gamma_I \Delta b_2). \tag{21}$$

The gates are supervised by modality-presence labels. Bidirectional retrieval exposes local agreement and discrepancy, while zero-initialized adapters and residual heads preserve the pretrained representations at the start of training.

3.5 Multi-Task Objective

SAFE-Align is trained end-to-end with objectives for binary classification, multi-label recognition, image grounding, and text grounding. The corresponding losses are cross-entropy, multi-label binary cross-entropy, an L1-plus-GIoU box loss, and token-level cross-entropy. Patch supervision, token-residual supervision, and modality-presence supervision form the fine-grained auxiliary term $\mathcal{L}_{FG}$. The complete objective is

$$\mathcal{L} = \lambda_{bic}\mathcal{L}_{BIC} + \lambda_{mlc}\mathcal{L}_{MLC} + \lambda_{IMG}\mathcal{L}_{IMG} + \lambda_{tmg}\mathcal{L}_{TMG} + \lambda_{fg}\mathcal{L}_{FG}. \tag{22}$$

The reported weights are $(\lambda_{bic}, \lambda_{mlc}, \lambda_{IMG}, \lambda_{tmg}, \lambda_{fg}) = (1.0, 1.0, 0.5, 0.5, 0.1)$. Further loss definitions and supervision details are given in Appendix A. Authentic samples are excluded from image-box supervision, padding positions are masked in the token loss, and a patch is positive when at least half of its area overlaps a manipulation box. The image and text gates receive modality-presence labels, while the residual token and patch heads provide direct auxiliary gradients to the grounding branch. The objective therefore couples pair-level detection with region- and token-level localization without introducing a separate uncertainty-calibration target.

4 Experiments

4.1 Datasets

We use the official DGM4 dataset and public train/validation/test splits (Shao et al., 2023; 2024). DGM4 contains approximately 230,000 human-centric news image–text pairs, 33.7% of which are authentic. Manipulations include face swap (FS), face attribute (FA), text swap (TS), text attribute (TA), and mixed cases. Image edits are often spatially small and text edits frequently affect only a few tokens, making the benchmark suitable for evaluating both pair-level detection and fine-grained grounding. Representative samples are shown in Appendix D Figure7.

4.2 Evaluation Metrics

We evaluate the four DGM4 tasks following HAMMER and ASAP (Shao et al., 2023; Zhang et al., 2025). Binary classification is measured by area under the ROC curve (AUC), accuracy (ACC), and equal error rate (EER). Multi-label manipulation recognition uses mean average precision (mAP), class-wise F1 (CF1), and overall F1 (OF1). Image grounding is reported with mIoU, mIoU50, and mIoU75, while text grounding uses token-level precision, recall, and F1. Lower EER is better; higher values are better for all other metrics.

4.3 Implementation Details

We train the main model and all ablation variants end-to-end on two NVIDIA GeForce RTX 4090 GPUs. Unless otherwise stated, the reported results for SAFE-Align and its ablations are averaged over three independent runs under the same experimental configuration. Images are resized to 256×256. We use ViT-B/16 (Dosovitskiy et al., 2021) as the visual encoder and BERT-base (Devlin et al., 2019) as the text encoder. Training uses a batch size of 32 for 50 epochs. We optimize the model with AdamW (Loshchilov & Hutter, 2019) with a weight decay of 0.02. The learning rate is linearly warmed up from 0 to 1×10^{-4} over the first 1,000 update steps and then cosine-decayed to 1×10^{-6}. We report baseline results from the corresponding papers, using their best reported values under the official DGM4 evaluation protocol. Additional loss and algorithmic details are given in the appendix.

5 Experimental Results and Discussion

5.1 Comparison with Existing Methods

We first compare SAFE-Align with existing methods on the official DGM4 split. These results measure within-benchmark effectiveness; they do not by themselves establish robustness to other datasets or manipulation sources.

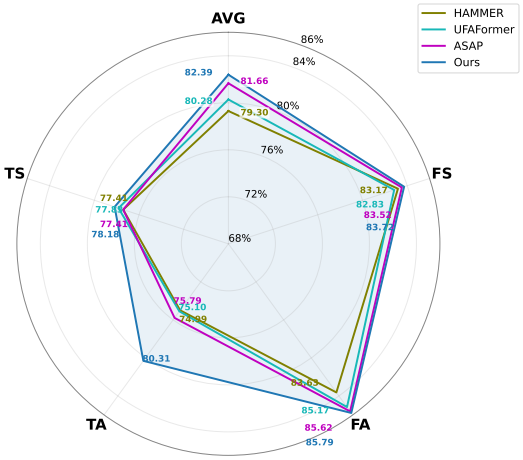


Figure 3: Class-wise F1-score comparison of HAMMER, UFAFormer, ASAP and SAFE-Align on the DGM4 manipulation categories. Each axis corresponds to one manipulation type, and larger values indicate better recognition.

Table 1: Complete 12-metric comparison on DGM4. Higher is better except EER.

Method	AUC	EER	ACC	mAP	CF1	OF1	mIoU	mIoU50	mIoU75	Prec.	Recall	F1
CLIP (Radford et al., 2021)	83.22	24.61	76.40	66.00	59.52	62.31	49.51	50.03	38.79	58.12	22.11	32.03
ViLT (Kim et al., 2021)	85.16	22.88	78.38	72.37	66.14	66.00	59.32	65.18	48.10	66.48	49.88	57.00
HAMMER (Shao et al., 2023)	93.19	14.10	86.39	86.22	79.37	80.37	76.45	83.75	76.06	75.01	68.02	71.35
HAMMER++ (Shao et al., 2024)	93.33	14.06	86.66	86.41	79.73	80.71	76.46	83.77	76.03	73.05	72.14	72.59
ViKI (Li et al., 2024)	93.51	13.87	86.67	86.58	81.07	80.10	76.51	83.95	75.77	77.79	66.06	72.44
UFAFormer (Liu et al., 2025)	93.81	13.60	86.80	87.85	80.31	81.48	**78.33**	85.39	**79.20**	73.35	70.73	72.02
ASAP (Zhang et al., 2025)	94.38	12.73	87.71	88.53	81.72	82.89	77.35	84.75	76.54	**79.38**	73.86	**76.52**
SAFE-Align (Ours)	**95.63**	**11.62**	**88.74**	**89.63**	**83.69**	**84.88**	77.53	**91.73**	76.38	76.45	**74.48**	75.73

Table 1 reports the complete 12-metric comparison. Among the compared methods, SAFE-Align obtains the strongest binary-classification and multi-label-recognition results, with an AUC of 95.63%, an ACC of 88.74%, and an mAP of 89.63%. Relative to ASAP, these correspond to gains of 1.25, 1.03, and 1.10 percentage points, respectively; EER is reduced by 1.11 points. CF1 and OF1 also increase by 1.97 and 1.99 points. Improvements across ranking, thresholded classification, and multi-label metrics suggest that the gain is not confined to a single decision threshold.

Grounding exhibits a more nuanced pattern. SAFE-Align achieves the best mIoU50 (91.73%), exceeding ASAP by 6.98 points and indicating better recovery of candidate manipulation regions. However, its overall mIoU (77.53%) remains below UFAFormer (78.33%), and its mIoU75 (76.38%) trails both UFAFormer and ASAP. The gap between mIoU50 and mIoU75 suggests that the model often identifies the correct region but does not always match the manipulation boundary tightly. Text grounding shows a similar coverage–precision trade-off: recall is highest at 74.48%, whereas precision and F1 remain below ASAP. We therefore interpret the grounding gains as improved candidate recovery rather than uniform superiority under stricter localization criteria.

The class-wise results in Figure 3 further show that the gain is category-dependent. SAFE-Align leads on FS, FA, and TA, with the largest margin on TA, whereas TS remains the principal weakness. The variation across categories is consistent with heterogeneous evidence requirements, although the class-wise scores alone do not identify which routed stream is responsible for each gain.

5.2 Ablation Studies

Table 2: Complete component ablations on DGM4.

Variant	AUC	EER	ACC	mAP	CF1	OF1	mIoU	mIoU50	mIoU75	Prec.	Recall	F1
Full model	**95.63**	**11.62**	**88.74**	**89.63**	**83.69**	**84.88**	**77.53**	**91.73**	**76.38**	**76.45**	**74.48**	**75.73**
w/o global alignment	94.25	12.07	87.81	88.69	82.64	83.63	76.52	90.11	74.57	75.46	74.46	74.96
w/o evidence fusion	93.96	12.67	87.59	87.53	82.50	83.38	76.50	89.99	73.03	75.63	74.47	75.05
w/o fine-grained alignment	94.96	11.92	87.75	88.73	82.77	83.97	76.99	90.21	74.58	75.84	74.47	75.15
w/o expert locator	94.88	12.11	87.66	88.17	82.38	83.78	75.81	89.59	73.56	75.62	74.45	75.03

Table 3: Effect of the number of alignment levels.

Variant	Levels	AUC	EER	ACC	mAP	CF1	OF1
w/ 2 levels	2	93.76	12.57	87.88	88.22	81.93	82.43
Full model	**4**	**95.63**	**11.62**	**88.74**	**89.63**	**83.69**	**84.88**
w/ 6 levels	6	94.54	12.59	88.29	88.06	82.50	83.45
w/ 8 levels	8	94.55	12.53	87.84	88.16	81.80	82.41

Table 4: Comparison of evidence fusion strategies.

Variant	Strategy	AUC	EER	ACC	mAP	CF1	OF1
Full model	**Alignment-guided gate**	**95.63**	**11.62**	**88.74**	**89.63**	**83.69**	**84.88**
Uniform fusion	Average over 3 branches	94.08	12.35	87.84	87.83	81.67	82.09
w/o alignment gate	Gated fusion w/o prior	93.66	12.56	87.78	87.77	81.57	81.98
Global evidence	Global evidence only	93.45	12.67	87.47	87.32	81.58	81.97
Local evidence	Local evidence only	93.14	12.61	87.29	87.17	81.57	81.86
Text evidence	Text evidence only	92.23	12.86	86.52	86.63	80.59	81.21

The component ablations in Table 2 are consistent with the intended division of labor among the three stages. Removing evidence fusion causes the largest drops in AUC and mAP, from 95.63% to 93.96% and from 89.63% to 87.53%, respectively. Removing the alignment prior decreases AUC by 1.38 points and mIoU75 by 1.81 points. Removing fine-grained alignment has a smaller effect on AUC (0.67 points) but reduces mIoU75 by 1.80 points, consistent with its localization role. Removing the expert locator yields the largest mIoU50 drop among the component variants, from 91.73% to 89.59%. The reported ablation results are averages over three independent runs. Because the variants are not parameter-matched, these comparisons establish component utility within the evaluated configuration rather than strict causal attribution.

The alignment-level study in Table 3 shows a non-monotonic dependence on discretization granularity. Four levels perform best on every reported detection and recognition metric. Performance is lower with two levels, while increasing the granularity to six or eight levels does not provide further gains. The fusion study in Table 4 directly tests the paper's central claim. Alignment-guided routing improves AUC by 1.55 points over uniform averaging and by 1.97 points over a learned gate that does not receive the alignment prior. Each single-source model also underperforms the full router. These results indicate that the gain is not attributable solely to adding multiple branches or a generic learned gate; conditioning aggregation on the estimated alignment state provides an additional benefit in the reported setting. This distinction is central to SAFE-Align: similarity quantization supplies routing context, while the empirical gain arises from using that context to condition multi-source aggregation rather than from discretization alone.

5.3 Qualitative Analysis

The qualitative examples complement the aggregate metrics and illustrate representative model behavior rather than its frequency in the dataset. Red annotations denote ground-truth boxes or tokens, blue annotations denote predictions, and warmer colors in the attention maps indicate stronger image-side responses. FS, FA, TS, and TA denote face swap, face attribute, text swap, and text attribute manipulation, respectively.

Figure 4 contains four single-manipulation and four mixed-manipulation examples. For the four single-manipulation cases, SAFE-Align correctly identifies FS, FA, TS, and TA, respectively, while localizing the corresponding manipulated facial region or altered text span. The mixed examples cover FS+TS, FS+TA, FA+TS, and FA+TA, and the predicted manipulation types match the corresponding ground-truth combinations in all four illustrated cases. Across these examples, the predicted image regions generally concentrate on the manipulated faces, while the highlighted text spans capture the relevant altered content. These examples therefore illustrate that SAFE-Align can jointly recognize manipulation types and provide image- and text-level grounding for both single and mixed manipulations, while the quantitative conclusions remain determined by the aggregate detection and grounding metrics reported in Table 1.

Figure 5 shows that image-side responses differ across manipulation types. In the FS and FA examples, responses concentrate near faces or other appearance-sensitive regions, whereas the TS example produces more distributed responses. The visualization suggests manipulation-dependent response patterns, but attention maps alone do not establish causal feature use or precise token attribution.

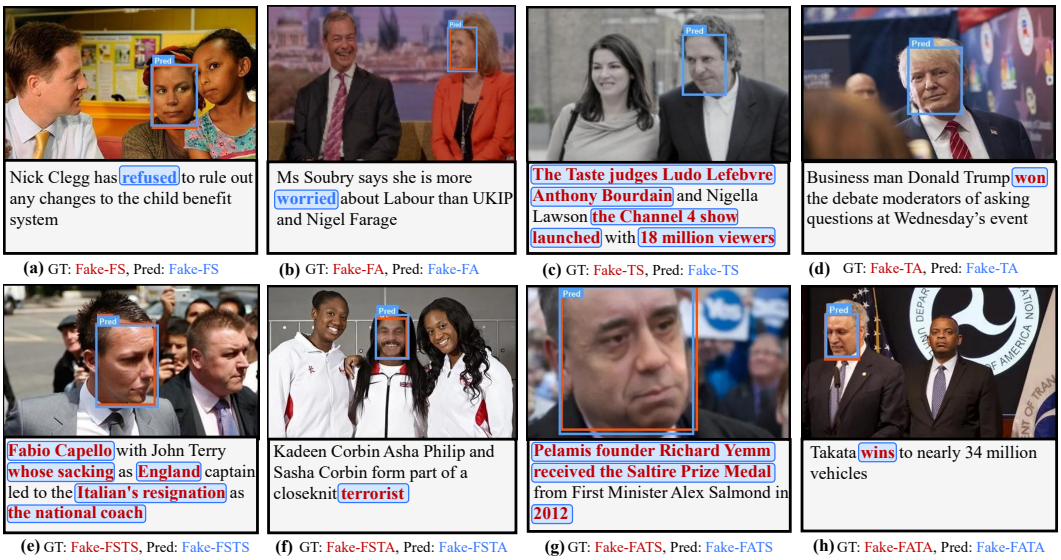


Figure 4: Detection and grounding results for single- and mixed-manipulation examples from DGM4. Red annotations denote ground-truth boxes or tokens, blue annotations denote SAFE-Align predictions, and the labels indicate predicted manipulation types.

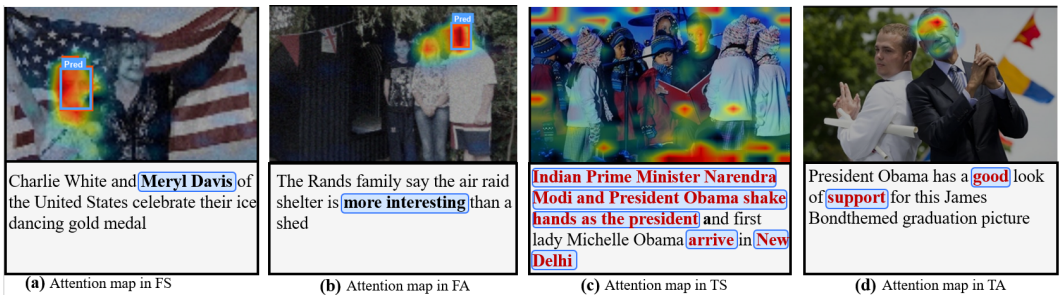


Figure 5: Attention visualization for different manipulation types. Red annotations denote ground truth, blue annotations denote predictions, and warmer colors indicate stronger image-side attention responses produced by SAFE-Align.

Additional qualitative results are provided in Appendix D. Figure 7 shows representative DGM4 manipulation patterns, while Figures 8–11 provide extended comparisons with HAMMER across single and mixed manipulation types. These examples further illustrate the coverage–precision trade-off observed in the quantitative results.

6 Conclusion

We presented SAFE-Align, a framework that treats image–text alignment as a sample-level routing variable for multimodal manipulation reasoning. Rather than applying the same fusion pattern to every pair, SAFE-Align represents the correspondence regime explicitly and uses it to route global-interaction, local-spatial, and text-only evidence. Bidirectional patch–token correspondence complements this sample-level routing with dense signals for image and text grounding. On the official DGM4 split, the method achieves 95.63% AUC, 88.74% ACC, and 89.63% mAP, and the routing ablations show that the alignment prior provides gains beyond uniform fusion and a learned gate without alignment conditioning. The grounding results are less uniform: candidate-region and token recovery improve, while strict boundary quality and token precision remain challenging. Overall, the current evidence supports a within-benchmark benefit of alignment-conditioned routing. Eval-

uating this principle across additional datasets and manipulation sources remains an important next step.

Reproducibility Statement

The paper specifies the model data flow, training objective, official dataset split, preprocessing, backbone encoders, optimizer, learning-rate schedule, and batch configuration. The appendix provides complete loss definitions, supervision rules, algorithmic pseudocode, derived ablation diagnostics, and additional qualitative results. The reported SAFE-Align results use the official DGM4 split and are averaged over three independent runs under a fixed training configuration.

AI Use Statement

Generative AI tools were used to assist with language polishing and manuscript editing. They were not used to generate experimental results, formulate mathematical claims, or make scientific conclusions. The authors manually reviewed and verified all AI-assisted edits and remain responsible for the scientific claims, experimental results, citations, and conclusions.

References

Sahar Abdelnabi, Rakibul Hasan, and Mario Fritz. Open-domain, content-based, multi-modal fact-checking of out-of-context images via online resources. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14940–14949, June 2022.

Shivangi Aneja, Chris Bregler, and Matthias Niessner. COSMOS: Catching out-of-context image misuse using self-supervised learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 14084–14092, 2023. doi: 10.1609/aaai.v37i12.26648.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (eds.), *Computer Vision – ECCV 2020*, pp. 104–120, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58577-8.

A. P. Dempster. Upper and Lower Probabilities Induced by a Multivalued Mapping. *The Annals of Mathematical Statistics*, 38(2):325 – 339, 1967. doi: 10.1214/aoms/1177698950. URL https://doi.org/10.1214/aoms/1177698950.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton-Ferrer. The deepfake detection challenge dataset. *CoRR*, abs/2006.07397, 2020. URL https://arxiv.org/abs/2006.07397.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. Trusted multi-view classification with dynamic evidential fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):2551–2566, 2023. doi: 10.1109/TPAMI.2022.3171983.

Audun Jøsang. *Subjective Logic: A Formalism for Reasoning Under Uncertainty*. Springer, Cham, Switzerland, 2016. doi: 10.1007/978-3-319-42337-1.

Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr - modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1780–1790, October 2021.

Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. Mvae: Multimodal variational autoencoder for fake news detection. In *The World Wide Web Conference*, WWW '19, pp. 2915–2921, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450366748. doi: 10.1145/3308558.3313552. URL https://doi.org/10.1145/3308558.3313552.

Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5583–5594. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/kim21k.html.

Jiaming Li, Hongtao Xie, Jiahong Li, Zhongyuan Wang, and Yongdong Zhang. Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6458–6467, June 2021a.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 9694–9705. Curran Associates, Inc., 2021b. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/505259756244493872b7709a8a01b536-Paper.pdf.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 12888–12900. PMLR, 17–23 Jul 2022a. URL https://proceedings.mlr.press/v162/li22n.html.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/li23q.html.

Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face x-ray for more general face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5001–5010, June 2020a.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10965–10975, June 2022b.

Qilei Li, Mingliang Gao, Guisheng Zhang, Wenzhe Zhai, Jinyong Chen, and Gwanggil Jeon. Towards multimodal disinformation detection by vision-language knowledge interaction. *Information Fusion*, 102:102037, 2024. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2023.102037. URL https://www.sciencedirect.com/science/article/pii/S1566253523003536.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (eds.), *Computer Vision – ECCV 2020*, pp. 121–137, Cham, 2020b. Springer International Publishing.

Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3207–3216, June 2020c.

Fuxiao Liu, Yinghan Wang, Tianlu Wang, and Vicente Ordonez. Visual news: Benchmark and challenges in news image captioning. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6761–6771, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.542. URL https://aclanthology.org/2021.emnlp-main.542/.

Huan Liu, Zichang Tan, Qiang Chen, Yunchao Wei, Yao Zhao, and Jingdong Wang. Unified frequency-assisted transformer framework for detecting and grounding multi-modal manipulation. *International Journal of Computer Vision*, 133(3):1392–1409, 2025. doi: 10.1007/s11263-024-02245-x.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Grace Luo, Trevor Darrell, and Anna Rohrbach. NewsCLIPpings: Automatic Generation of Out-of-Context Multimodal Media. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6801–6817, Online and Punta Cana, Dominican Republic, November 2021a. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.545. URL https://aclanthology.org/2021.emnlp-main.545/.

Yuchen Luo, Yong Zhang, Junchi Yan, and Wei Liu. Generalizing face forgery detection with high-frequency features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16317–16326, June 2021b.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Niessner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1–11, October 2019.

Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/a981f2b708044d6fb4a71a1463242520-Paper.pdf.

Glenn Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, NJ, 1976. ISBN 9780691214696. doi: 10.1515/9780691214696. URL https://doi.org/10.1515/9780691214696.

Rui Shao, Tianxing Wu, and Ziwei Liu. Detecting and grounding multi-modal media manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6904–6913, June 2023.

Rui Shao, Tianxing Wu, Jianlong Wu, Liqiang Nie, and Ziwei Liu. Detecting and grounding multimodal media manipulation and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5556–5574, 2024. doi: 10.1109/TPAMI.2024.3367749.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. *SIGKDD Explor. Newsl.*, 19(1):22–36, September 2017. ISSN 1931-0145. doi: 10.1145/3137597.3137600. URL https://doi.org/10.1145/3137597.3137600.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15638–15650, June 2022.

Hao Tan and Mohit Bansal. LXMERT: Learning cross-modality encoder representations from transformers. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 5100–5111, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1514. URL https://aclanthology.org/D19-1514/.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5238–5248, June 2022.

William Yang Wang. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In Regina Barzilay and Min-Yen Kan (eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 422–426, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-2067. URL https://aclanthology.org/P17-2067/.

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 & Data Mining*, KDD '18, pp. 849–857, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450355520. doi: 10.1145/3219819.3219903. URL https://doi.org/10.1145/3219819.3219903.

Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. FILIP: Fine-grained interactive language-image pre-training. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=cpDhcsEDC2.

Zhenxing Zhang, Yaxiong Wang, Lechao Cheng, Zhun Zhong, Dan Guo, and Meng Wang. Asap: Advancing semantic alignment promotes multi-modal manipulation detecting and grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4005–4014, June 2025.

Fei Zhao, Taotian Pang, Chunhui Li, Zhen Wu, Junjie Guo, Shangyu Xing, and Xinyu Dai. Alignngpt: Multi-modal large language models with adaptive alignment capability, 2024.

Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, and Jianfeng Gao. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16793–16803, June 2022.

A Additional Loss and Supervision Details

The binary and multi-label heads use

$$\mathcal{L}_{BIC} = \mathrm{CE}(\hat{\mathbf{y}}_{bin}, y_{bin}), \qquad \mathcal{L}_{MLC} = \mathrm{BCEWithLogits}(\hat{\mathbf{y}}_{mul}, \mathbf{y}_{mul}). \tag{23}$$

For image grounding, the box loss combines L1 and GIoU terms:

$$\mathcal{L}_{IMG} = \left|\hat{\mathbf{b}} - \mathbf{b}\right|_1 + 1 - \mathrm{GIoU}(\hat{\mathbf{b}}, \mathbf{b}). \tag{24}$$

For text grounding, padding tokens are masked and the loss is

$$\mathcal{L}_{TMG} = \mathrm{CE}(\hat{\mathbf{Y}}_{tok}, \mathbf{Y}_{tok}). \tag{25}$$

Patch labels are obtained by mapping the ground-truth box to the 256×256 image grid. A patch is positive when at least 50% of its area overlaps a manipulation box:

$$\mathcal{L}_{FG_{patch}} = \mathrm{BCEWithLogits}(\hat{\mathbf{m}}_{patch}, \mathbf{Y}_{patch}). \tag{26}$$

The residual token and modality-gate losses are

$$\mathcal{L}_{FG_{token}} = \mathrm{CE}(\Delta\mathbf{Y}_{tok}, \mathbf{Y}_{tok}), \qquad \mathcal{L}_{FG_{gate}} = \mathrm{BCE}(\gamma_I, y_I) + \mathrm{BCE}(\gamma_T, y_T), \tag{27}$$

$$\mathcal{L}_{FG} = \mathcal{L}_{FG_{patch}} + \mathcal{L}_{FG_{token}} + \mathcal{L}_{FG_{gate}}. \tag{28}$$

Authentic samples are masked out of image-box supervision, and padding tokens are excluded from text supervision.

B Algorithmic Details

Algorithm 1 summarizes one end-to-end training step. The alignment thresholds Θ are estimated from the training-set similarity distribution before optimization and are then held fixed, as described in Section 3. The pseudocode mirrors the data flow in the main text and does not introduce additional supervision.

Algorithm 1 SAFE-Align end-to-end training step

Require: Training mini-batch $\mathcal{B}$; fixed thresholds Θ; parameters Ω
Ensure: Updated parameters Ω

```
for all (I, T, ·) ∈ B do
    Encode image and text: V = [v_cls; V_pat], T = [t_cls; T_tok]
    Project and normalize global features to obtain f_I and f_T
    Compute s = f_I^⊤ f_T and l = Quantize(s; Θ)
    Form alignment prior h_align = φ_align([e_l; s])
    Compute global, local, and text-only representations h_global, h_local, h_text
    Obtain non-negative source supports S_global, S_local, S_text
    Compute alignment-conditioned routing weights α and fused support S_fused
    Predict binary and multi-label outputs from the routed residual representations
    Adapt patch/token features and compute bidirectional patch–token affinity A
    Construct discrepancy features Z_p and Z_t
    Predict patch logits, token residual logits, modality gates, and image box b̂
end for
Compute all task and auxiliary losses using the masks in Appendix A
Combine them according to Eq. 22 and update Ω by back-propagation
```

Algorithm 2 shows the inference path. Ground-truth boxes, token labels, and modality-presence labels are not used at test time; the same fixed training-derived thresholds are applied to determine the alignment level.

Algorithm 2 SAFE-Align inference

Require: Image–text pair (I, T); fixed thresholds Θ; trained parameters Ω
Ensure: Binary score, manipulation labels, image box, and token predictions
1: Encode (I, T) and compute global similarity s
2: $l \leftarrow \mathrm{Quantize}(s; \Theta)$; $\mathbf{h}_{align} \leftarrow \phi_{align}([\mathbf{e}_l; s])$
3: Compute global, local, and text-only source supports
4: Route supports with the alignment-conditioned gate and obtain classification outputs
5: Compute bidirectional patch–token correspondence and fine-grained discrepancy features
6: Decode image box $\hat{\mathbf{b}}$ and token logits $\hat{\mathbf{Y}}_{tok}$
7: **return** all task predictions

C Additional Quantitative Diagnostics

Table 5 recasts the component ablation in Table 2 as absolute performance drops relative to the full model. These values are derived directly from the reported runs; they are not additional experiments. This representation makes the functional differences among components easier to inspect. Removing evidence fusion causes the largest degradation in AUC, mAP, and mIoU75, whereas removing the expert locator causes the largest mIoU50 drop. Fine-grained alignment has a comparatively smaller effect on AUC but a larger effect on strict image localization, consistent with its intended role.

Table 5: Absolute metric drops relative to the full model, derived from Table 2. Larger values indicate a larger degradation after removing the component.

Variant	ΔAUC	ΔACC	ΔmAP	ΔmIoU50	ΔmIoU75	ΔText-F1
w/o global alignment	1.38	0.93	0.94	1.62	1.81	0.77
w/o evidence fusion	**1.67**	**1.15**	**2.10**	1.74	**3.35**	0.68
w/o fine-grained alignment	0.67	0.99	0.90	1.52	1.80	0.58
w/o expert locator	0.75	1.08	1.46	**2.14**	2.82	**0.70**

The alignment-granularity results provide a second diagnostic. Relative to four levels, AUC decreases by 1.87 points with two levels, 1.09 points with six levels, and 1.08 points with eight levels. This non-monotonic trend is consistent with a benefit from exposing a moderate number of correspondence regimes rather than simply increasing discretization resolution. Similarly, the routing ablation shows that the full alignment-conditioned gate outperforms both uniform averaging and a learned gate without the alignment prior, isolating the value of conditioning the aggregation rule on the correspondence state within the reported experimental setting.

D Additional Visualization Results

Figure 7 first presents representative examples from DGM4, covering both single and mixed image–text manipulations. The examples illustrate the diversity of the benchmark: visual manipulations may alter localized facial regions, text manipulations may affect only a few tokens, and mixed cases require simultaneous reasoning over both modalities.

Figure 6 provides a representative comparison between HAMMER and SAFE-Align. In the displayed FA case, SAFE-Align places the predicted box closer to the annotated face, while in the TS case it highlights the altered tokens “horror” and “died”. The examples also reveal remaining errors: predicted boxes can be loose, and token predictions can include extra spans. This behavior is consistent with the quantitative result that SAFE-Align improves mIoU50 and token recall without uniformly improving mIoU75 or token precision.

Figures 8–11 extend this comparison to additional manipulation categories. Figure 8 shows single-modality FS and FA examples, where both methods generally focus on manipulated facial regions, but SAFE-Align often produces more concentrated image responses around the edited area. Figure 9



(b-1) Visualization of HAMMER in FA **(b-2)** visualization of Ours in FA

(c-1) Visualization of HAMMER in TS **(c-2)** Visualization of Ours in TS

Figure 6: Representative qualitative comparison between HAMMER and SAFE-Align across single and mixed manipulation types. Red annotations denote ground truth, blue annotations denote predictions, and attention maps show image responses for the compared methods.



(a) orig **(a)** Fake-FS **(b)** orig **(b)** Fake-FA

(c) orig **(c)** Fake-TS **(d)** orig **(d)** Fake-TA

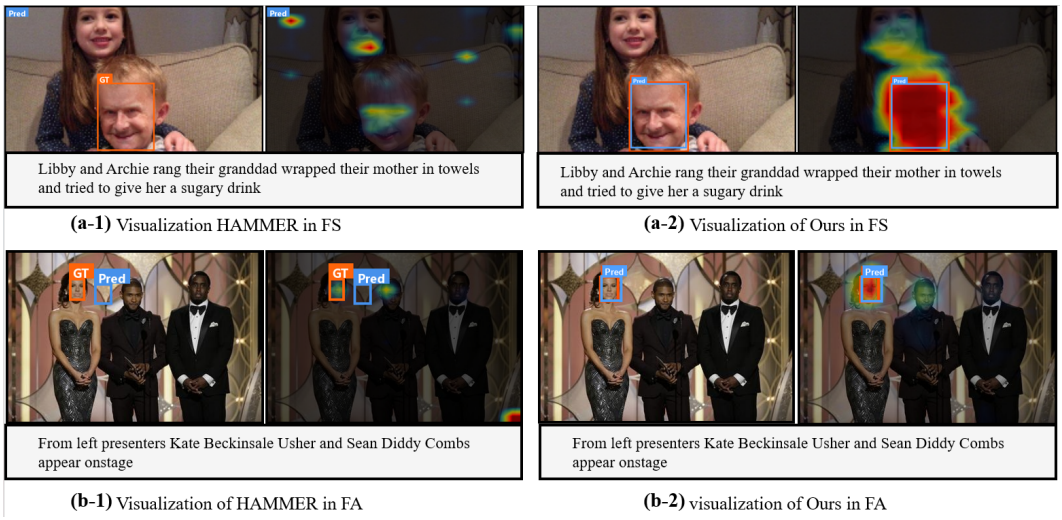
(a) orig **(a)** Fake-FSTS **(b)** orig **(b)** Fake-FSTA

(c) orig **(c)** Fake-FATS **(d)** orig **(d)** Fake-FATA

Figure 7: Examples from DGM4, including single and mixed image–text manipulations.

presents TS and TA examples and highlights the complementary role of text grounding: SAFE-Align recovers altered text spans while its image-side responses remain comparatively diffuse when the manipulation is primarily textual.

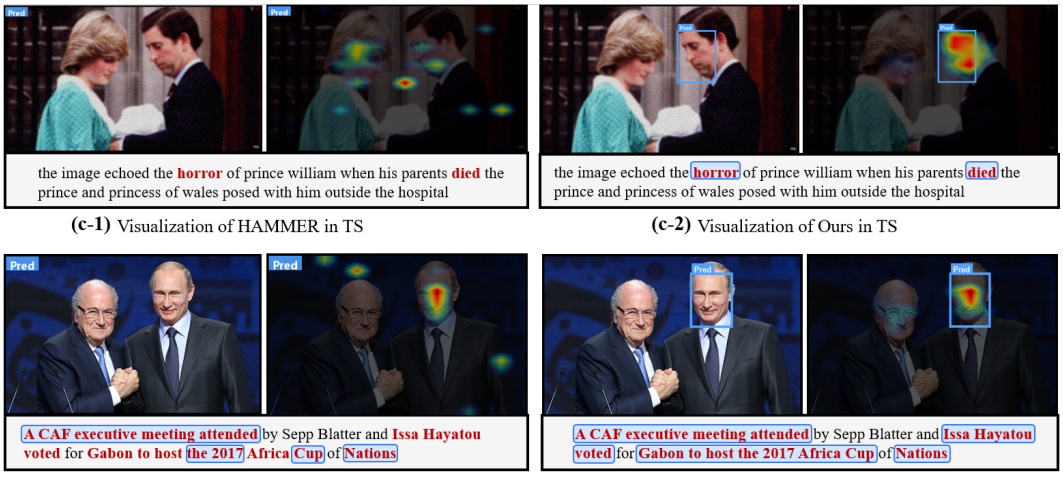
Figures 10 and 11 further examine mixed manipulations. Figure 10 covers FS+TS and FS+TA, whereas Figure 11 covers FA+TS and FA+TA. These examples require the model to jointly identify manipulated image regions and altered text spans. Across the displayed cases, SAFE-Align generally concentrates image responses near the manipulated faces and recovers relevant text spans, although offset boxes and additional token predictions remain visible in some examples. Overall, the qualitative results are consistent with the quantitative coverage–precision trade-off: candidate regions and manipulated tokens are frequently recovered, while strict boundary accuracy and token precision remain imperfect.



(a-1) Visualization HAMMER in FS **(a-2)** Visualization of Ours in FS

(b-1) Visualization of HAMMER in FA **(b-2)** visualization of Ours in FA

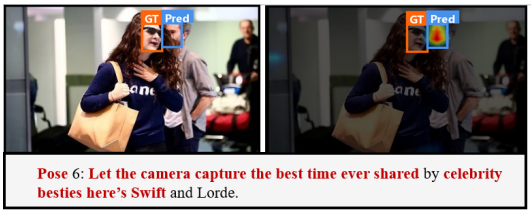
Figure 8: Additional qualitative comparison between HAMMER and SAFE-Align on single-modality FS and FA examples.



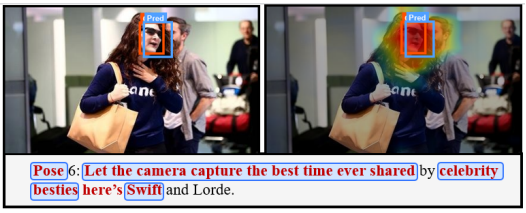
(c-1) Visualization of HAMMER in TS **(c-2)** Visualization of Ours in TS

(d-1) Visualization of HAMMER in TA **(d-2)** Visualization of Ours in TA

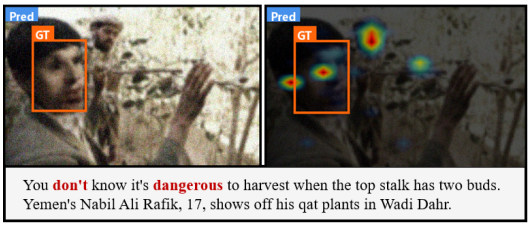
Figure 9: Additional qualitative comparison between HAMMER and SAFE-Align on single-modality TS and TA examples.



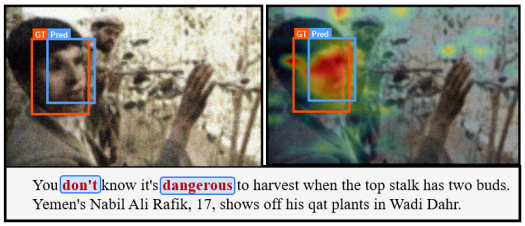
(e-1) Visualization of HAMMER in FSTS



(e-2) Visualization of Ours in FSTS

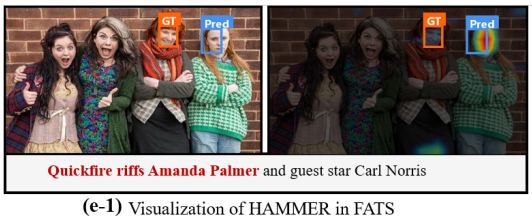


(f-1) Visualization of HAMMER in FSTA

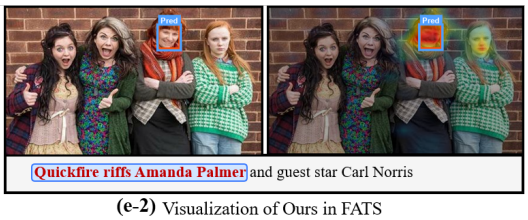


(f-2) Visualization of Ours in FSTA

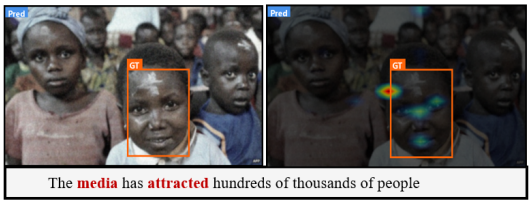
Figure 10: Additional qualitative comparison between HAMMER and SAFE-Align on mixed FS+TS and FS+TA examples.



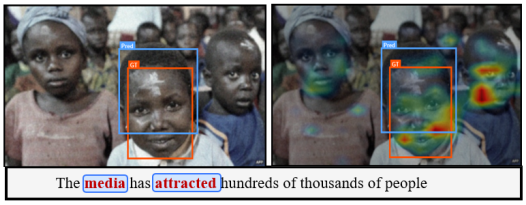
(e-1) Visualization of HAMMER in FATS



(e-2) Visualization of Ours in FATS



(f-1) Visualization of HAMMER in FATA



(f-2) Visualization of Ours in FATA

Figure 11: Additional qualitative comparison between HAMMER and SAFE-Align on mixed FA+TS and FA+TA examples.