Beyond Textual Claims: Strategy for Multimodal Fact Checking with Unified Visual, Textual, and Contextual Representations

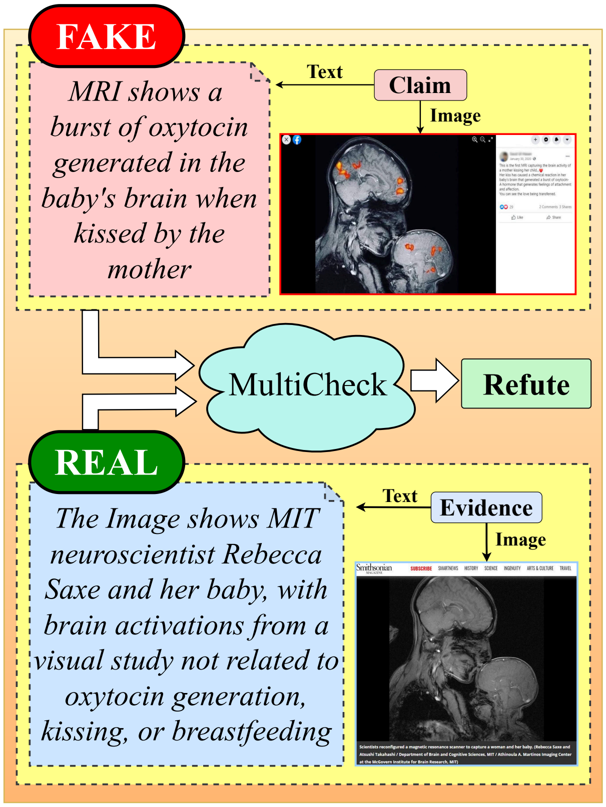
Anonymous ACL submission

Abstract

The growing rate of multimodal misinformation, where claims are supported by both text and images poses significant challenges to fact-checking systems that rely primarily on textual evidence. In this work, we propose a unified framework for fine-grained multimodal fact verification called **MultiCheck**, designed to reason over structured textual and visual signals. Our architecture combines dedicated encoders for text and images with a fusion module that captures cross-modal relationships using element-wise interactions. A classification head then predicts the veracity of a claim, supported by a contrastive learning objective that encourages semantic alignment between claim-evidence pairs in a shared latent space. We evaluate our approach on the Factify 2 dataset, achieving a weighted F1 score of **0.84**, substantially outperforming the baseline. These results highlight the effectiveness of explicit multimodal reasoning and demonstrate the potential of our approach for scalable and interpretable fact-checking in complex, real-world scenarios.

1 Introduction:

Misinformation has become a serious concern in today's digital environment, affecting many areas like politics, public health, and finance (Caceres et al., 2022). While early instances of false information were mostly text-based (Murphy et al., 2023; Kim et al., 2021; Di Domenico et al., 2021), modern misinformation campaigns increasingly blend text with images, audio, and video making them more persuasive and harder to detect (Abdali et al., 2024; Mura et al., 2025; Askari, 2023). This rise in multimodal misinformation reveals the limitations of traditional fact-checking systems, which primarily focus on textual content (Tufchi et al., 2023; Braun et al., 2024; Mura et al., 2025).


MRI shows a burst of oxytocin generated in the baby's brain when kissed by the mother

The Image shows MIT neuroscientist Rebecca Saxe and her baby, with brain activations from a visual study not related to oxytocin generation, kissing, or breastfeeding

Figure 1: Refuting a viral claim using combined text and image evidence.

In response, the research community has turned to multimodal fact-checking, where claims are verified using both textual and visual contents (Akhtar et al., 2023; Braun et al., 2024). Recent benchmarks such as Fakeddit (Nakamura et al., 2019), Mocheg (Yao et al., 2023b), and Factify 2 (Suryavardan et al., 2023) have helped advance research in this direction, see appendix A for futher details.

Multimodal fact-checking remains a difficult and open challenge. This is because, the task goes beyond simple classification. It requires understanding how modalities interact and whether they support or contradict each other. For example in Figure 1, where a claim is paired with a scientific image that appears credible but is actually unrelated. Identifying this mismatch requires more than surface-level alignment. Thus we need a structured reasoning across both textual and visual modalities.

Existing methods relied on one of two strategies: (i) concatenating image-text embeddings (Chen et al., 2020; Lu et al., 2019), or (ii) joining vision and language encoders through late fusion (Du et al., 2023b). For example, MOCHEG(Yao et al., 2023b) processes claims and evidence through modality-specific encoders and fuses their outputs without deep cross-modal interaction, whereas PRO-COFACTV2(Du et al., 2023a) leverages attention-based mechanisms. While these strategies capture basic correlations, they often fail to detect fine contradictions or poorly aligned semantics. we believe they fail to separate modality-specific support signals and make the reasoning process transparent. We, on the other hand, have proposed a unified fact-checking framework called "MultiCheck", where we used (i) a novel **fusion technique**, which captures semantic relation via element-wise difference and product operations, and (ii) a **contrastive learning objective** that aligns semantically similar claim-document pairs in a shared latent space, improving representation consistency across modalities. Our approach is inspired by prior advances in multimodal relational reasoning. The element-wise difference and product were shown to be capturing fine-grained interactions between paired inputs in natural language inference models (Conneau et al., 2017) and bilinear attention models (Kim et al., 2018). These operations encode both alignment and divergence between two modalities, enabling more expressive cross-modal representations. To further strengthen semantic alignment, we integrated a contrastive head that serves the objective of contrastive learning that operates on projected claim-document embeddings. This module is trained with a symmetric InfoNCE loss (Oord et al., 2018), encouraging the model to align semantically related pairs while pushing apart unrelated ones. Unlike prior methods that used frozen embeddings or shallow probes (Cekinel et al., 2025a), our model is fully trainable end-to-end and jointly optimizes for both classification and contrastive learning, resulting in a more discriminative and robust representation for multimodal fact verification.

Our contributions are as follows:

- We have introduced a unified multimodal fact-checking architecture, **"MultiCheck"**, that combines structured text and image features. It incorporates a **contrastive head** and a **fusion module** to align semantically related claim-evidence pairs while separating unrelated ones. As a result, we achieved a new state-of-the-art on the Factify 2 benchmark with a weighted F1 score of **0.84**, outperforming the baseline by **27%**.
- We have conducted thorough ablations to assess the impact of different backbones, fusion strategies, and training objectives. The results reveal that relational fusion using difference and product consistently outperforms simple concatenation, and that the inclusion of the contrastive loss significantly boosts performance, particularly in ambiguous or weak evidence scenarios.
- We have performed a comprehensive error analysis using statistical significance tests, showing that our model not only outperforms the baseline but does so in a structurally meaningful way, correcting more errors than it introduces. Discordant pair comparisons and contingency heatmaps analyses reveals consistent improvements across challenging veracity classes.
- We further supported our findings through qualitative analysis, highlighting how OCR cues and visual metadata can decisively shift predictions in subtle cases often missed by prior systems.

2 Dataset details:

We used the **Factify-2** dataset provided by Du et al. (2024a) in our experiments. It is designed to assess the claim veracity that requires reasoning over both textual and visual information. It is different from other datasets like LIAR (Wang, 2017), LIAR-PLUS (Alhindi et al., 2018), and Mocheg (Yao et al., 2023a), which either lack claim-side images or rely solely on textual evidence. Factify-2, on the other hand, includes real-world images, OCR-extracted text, and

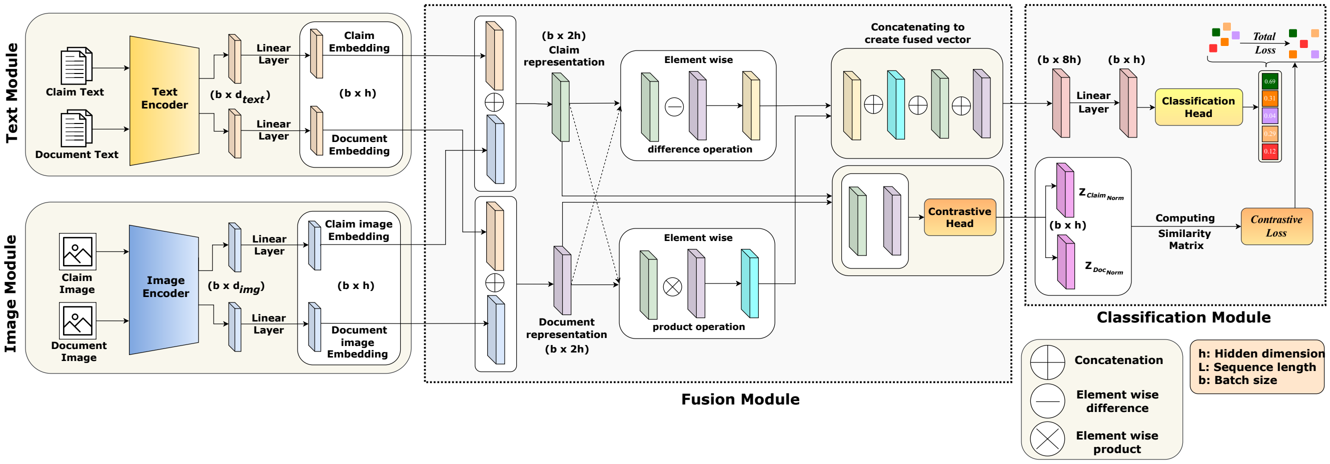


Figure 2: Intuitive fusion representation using element-wise difference and product.

paired claim-document examples having both images and texts. It has 42,500 instances, each labelled with one of five labels. The distributions of samples across classes are shown in Table 1. A detailed description of the dataset, along with some illustrative examples, is reported in appendix (section B).

Factify 2 (Du et al., 2024a)				
Classes	**Train**	**Validation**	**Test**	**Total**
Support_Multimodal	5,580	1,420	1,500	8,500
Support_Text	5,485	1,515	1,500	8,500
Insufficient_Multimodal	5,472	1,528	1,500	8,500
Insufficient_Text	5,494	1,506	1,500	8,500
Refute	5,469	1,531	1,500	8,500
Total	27,500	7,500	7,500	42,500

Table 1: Dataset statistics for Factify 2 across training, validation, and test splits.

3 Methodology

In this section, we have reported our proposed framework, schematically depicted in Figure 2. Our framework has four components, i.e. (i) text module, (ii) image module, (iii) fusion module, and (iv) classification module. Each component of our framework is described as follows,

Text module: This module converts the claims, associated evidence documents and OCR texts extracted from images into an embedding space. We have used pre-trained language models such as RoBERTa (Liu et al., 2019), DeBERTa (He et al., 2020), or SBERT (Reimers and Gurevych, 2019) for this purpose. See Appendix C.1 for further details.

Image module: Here, we convert images associated with claims and associated evidence documents into an embedding space. We used image encoders like ResNet50 (He et al., 2016) or Vision Transformer (ViT) (Dosovitskiy et al., 2020) for this purpose. More details are reported in the Appendix C.2.

We used linear layers after text and image encoders to keep the text and image embeddings of same shape.

Fusion module: Our key innovation lies in the fusion module. It is designed to capture both direct and relational interactions between multimodal evidence. Prior systems rely on simple feature concatenation (Mishra et al., 2020; Sata et al., 2025; Wang et al., 2022). In contrast, our framework explicitly captures alignment and divergence between claim and document evidence. It uses element-wise difference and product operations over their multimodal representations embeddings. This approach is inspired by relational reasoning techniques shown effective in prior research (Kim et al., 2018; Conneau et al., 2017; Chen et al., 2020; Liu et al., 2023; Gong et al., 2024).

Specifically, we first concatenate the text and image embeddings separately for the claim and document, forming initial multimodal representations. We then compute (i) *element-wise difference*: which highlights differences between claim and document representations, and (ii) *element-wise product*: which emphasizes regions of strong alignment. This design enables the model to capture both conflicts and alignments across modalities, denoted by:

- $\mathbf{V}_{\text{diff}} = \left|\mathbf{V}_{\text{claim repr}} \ominus \mathbf{V}_{\text{doc repr}}\right|$
- $\mathbf{V}_{\text{prod}} = \mathbf{V}_{\text{claim repr}} \otimes \mathbf{V}_{\text{doc repr}}$

where $\ominus$ denotes element-wise difference and $\otimes$ denotes the element-wise product between the claim and evidence document representation embeddings. It helps in determining, whether the multimodal input supports, refutes, or fails to verify a claim. Finally, these four vector embeddings: the claim representation, evidence document representation, difference, and product are concatenated, denoted by:

$$\mathbf{V}_{\text{fused}} = \mathbf{V}_{\text{diff}} \oplus \mathbf{V}_{\text{prod}} \oplus \mathbf{V}_{\text{claim repr}} \oplus \mathbf{V}_{\text{doc repr}}$$

are passed through a fusion network. Where $\oplus$ shows the operation of concatenation. This network consists of a fully connected layer with GELU activation (Hendrycks and Gimpel, 2016), and dropout regularization, producing a final fused feature vector, denoted by:

- $\mathbf{V}_{\text{final}} = \text{FFN}(\mathbf{V}_{\text{fused}}) \in \mathbb{R}^{b\times h}$

Where $\boldsymbol{h}$ denotes the dimensionality of the shared latent space and $\boldsymbol{b}$ shows the batch-size. This final fused representation encodes refined multimodal relationships crucial for robust fact verification.

In addition to this, our framework integrates a contrastive feature learning directly within the fusion module. Specifically, we introduced a contrastive projection head that maps the multimodal claim and document representations embeddings into a shared latent space for contrastive learning. This component operates in parallel to the fused vector used for classification shown in Figure 2, ensuring that both objectives reinforce each other and improve the model's ability to distinguish between fine semantic relationships. During training, we have applied a symmetric **InfoNCE loss** (Oord et al., 2018) to these projected embeddings, *pulling* together representations of matching claim-document pairs while *pushing* apart those of unrelated pairs. This contrastive learning complements the supervised classification objective, enhancing the model's ability to capture fine-grained relationships between claims and evidence. Further architectural details are provided in Appendix C.3.

Classification module: The final fused representation is passed through a linear classification layer to predict one of five fine-grained veracity labels. This head produces the final logits over the predefined classes. This component serves as the decision layer, translating the model's joint understanding of textual and visual signals into actionable predictions. To improve discriminative capacity, we augmented the classification training with a contrastive learning objective. Specifically, the claim and document embeddings are each projected into a shared latent space via a contrastive head. A symmetric InfoNCE loss (Oord et al., 2018) encourages semantically aligned claim-document pairs to lie close in this space while pushing apart unrelated ones. Our final loss combines this contrastive supervision with standard cross-entropy loss for classification, denoted by:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \lambda \cdot \mathcal{L}_{\text{contrastive}}$$

where λ balances the influence of the contrastive objective.

This dual-objective setup strengthens the model's ability, it not only classify correctly but also learn a semantically structured embedding space, where alignment across modalities is meaningful and consistent. Appendix C.4 provides full mathematical formulations and training details.

4 Experiments:

In this section, we have presented the experiments conducted to evaluate the effectiveness of our proposed framework. we evaluated the effect of various learning strategies on model performance. Our experiments progressively examined the impact of model architectures and training choices. Special focus is given to how fusion strategies and the role of element-wise operations enhance multimodal relationships. Details regarding our reproduced baseline are provided separately in Section 4.1.

4.1 Baselines:

For comparative evaluation, we reproduce the **Pro_cofactv2** model, originally proposed by Du et al. (2024b), which has demonstrated state-of-the-art performance on the Factify 2 benchmark.

To ensure comparison, we replicate the model using the same configuration settings as the original work, including fixed random seed initialization, optimizer parameters, pretrained backbones, and architecture-specific hyperparameters. The reproduced performance of Pro_cofactv2 on the Factify 2 dataset is summarized in Table 2. Additional details on the reproduction setup and hyperparameter choices are provided in Appendix B.2.

Class	Precision	Recall	F1 Score
Support Text	0.48 (±0.04)	0.38 (±0.06)	0.42 (±0.03)
Support Multimodal	0.50 (±0.04)	0.61 (±0.02)	0.55 (±0.02)
Insufficient Text	0.50 (±0.01)	0.44 (±0.07)	0.46 (±0.04)
Insufficient Multimodal	0.43 (±0.02)	0.46 (±0.06)	0.44 (±0.02)
Refute	0.98 (±0.00)	0.98 (±0.00)	0.98 (±0.00)
Weighted F1 Score	**0.57** (±0.01)		

Table 2: Performance of the reproduced baseline model (Pro-CoFactv2) on the Factify 2 dataset.

4.2 Experimental Variants:

In addition to evaluating the baseline, we conducted experiments to systematically examine the impact of our proposed approach. Specifically, we compare two main variants of our architecture:

With Contrastive Head: This version incorporates a contrastive projection head applied to the multimodal representations of claims and evidence documents. Training includes a contrastive loss in addition to the standard cross-entropy loss. This encourages the model to learn modality-consistent and semantically aligned embeddings.

Without Contrastive Head: In this version, the contrastive projection head and the associated loss are omitted. The model relies solely on cross-entropy loss applied to the fused multimodal representation for classification.

Both variants use the same text and image encoders, fusion mechanism, and training hyperparameters. The only difference is the inclusion of contrastive supervision. Results for the contrastive variant are shown in Tables 3 and 7, while results for the non-contrastive variant are provided in Tables 4 and 8. We discuss the comparative performance of these two configurations in Section 5.

5 Results and Discussion:

We evaluated our proposed framework on the Factify 2 dataset using multiple combinations of language and vision backbones. Performance is measured using weighted F1 scores across five fine-grained veracity labels. We have reported both overall and class-wise results to assess the model's capabilities. As shown in Tables 3 and 4, all model variants significantly outperform the baseline. Additionally, models integrated with contrastive head consistently outperform their non-contrastive counterparts, see Figure 6. This trend holds across the visual encoders, confirming the robustness of our architecture. Notably, the marginal gains are observed in "Insufficient" and "Refute" categories that require resolving minor differences between text and images. These results show that our fusion method and contrastive learning help the model to better connect text and images.

5.1 Insights:

Contrastive learning boosts accuracy: Across all settings, the inclusion of contrastive supervision yields statistically significant gains. It consistently outperforms non-contrastive counterparts, see Table 5. The consistent asymmetry in discordant prediction counts, as shown in Figure 7, confirms that contrastive models correct significantly more baseline errors than they introduce.

Role of fusion strategy: The fusion module using element-wise difference and product provides critical advantages over simple concatenation. These operations help the model clearly match or contrast the claim with the evidence, improving its reasoning.

Model robust nature: Whether paired with ResNet50 or ViT, and regardless of the language encoder, the contrastive head along with fusion module consistently improves performance, highlighting its modularity and general applicability.

6 Error analysis:

To better understand the behavioral differences between our models and the baseline, we performed a detailed error analysis using both statistical significance tests and qualitative assessments.

Statistical comparisons: we have applied McNemar's test for overall accuracy and Bowker's test for class-level shifts. Results across all five model and variants are reported in Tables 5, 6 which shows McNemar's χ^2 scores are consistently high with p-values well below 0.05, confirming significant gains

Models w/ Contrastive Head	Support Text	Support Multimodal	Insufficient Text	Insufficient Multimodal	Refute	Weighted F1 Score
Roberta + ResNet50	**0.77 (±0.02)**	0.83 (±0.01)	**0.82 (±0.00)**	0.76 (±0.02)	1.00 (±0.00)	**0.84 (±0.01)**
Roberta + ViT	**0.77 (±0.02)**	**0.84 (±0.01)**	**0.82 (±0.00)**	**0.77 (±0.01)**	1.00 (±0.00)	**0.84 (±0.01)**
DeBERTa + ViT	0.77 (±0.01)	0.84 (±0.01)	0.83 (±0.00)	0.78 (±0.01)	1.00 (±0.00)	**0.84 (±0.01)**
DeBERTa + ResNet50	0.76 (±0.02)	0.81 (±0.02)	0.81 (±0.01)	0.75 (±0.02)	1.00 (±0.00)	0.83 (±0.01)
SBERT + ResNet50	0.75 (±0.01)	**0.83 (±0.01)**	0.79 (±0.00)	**0.76 (±0.01)**	1.00 (±0.00)	0.82 (±0.00)
Baseline	0.42 (±0.03)	0.55 (±0.02)	0.46 (±0.04)	0.44 (±0.02)	0.98 (±0.00)	0.57 (±0.01)

Table 3: Class-wise F1 scores and weighted F1 for various model combinations on the Factify 2 dataset with contrastive head. Each value is the mean F1 ±std across seeds. Precision and Recall are reported in Table 7.

Models w/o Contrastive Head	Support Text	Support Multimodal	Insufficient Text	Insufficient Multimodal	Refute	Weighted F1 Score
Roberta + ResNet50	0.74 (±0.01)	**0.82 (±0.01)**	0.78 (±0.01)	**0.75 (±0.01)**	0.99 (±0.01)	**0.82 (±0.01)**
Roberta + ViT	**0.75 (±0.01)**	**0.82 (±0.01)**	**0.79 (±0.01)**	**0.75 (±0.01)**	**1.00 (±0.00)**	**0.82 (±0.00)**
DeBERTa + ViT	0.74 (±0.01)	**0.82 (±0.00)**	0.78 (±0.01)	**0.75 (±0.01)**	**1.00 (±0.00)**	**0.82 (±0.00)**
DeBERTa + ResNet50	0.74 (±0.01)	0.81 (±0.01)	0.78 (±0.01)	0.74 (±0.01)	**1.00 (±0.00)**	0.81 (±0.00)
SBERT + ResNet50	0.71 (±0.03)	0.81 (±0.02)	0.75 (±0.02)	0.74 (±0.02)	0.99 (±0.00)	0.78 (±0.05)
Baseline	0.42 (±0.03)	0.55 (±0.02)	0.46 (±0.04)	0.44 (±0.02)	0.98 (±0.00)	0.57 (±0.01)

Table 4: Class-wise F1 scores and weighted F1 for various model combinations on the Factify 2 dataset (no contrastive head). Each value is mean F1 ±std across seeds. Precision and Recall are reported in Table 8.

Models w/ Contrastive Head	McNemar's χ^2	McNemar's p-value	Significance at $\alpha = 0.05$	Bowker's χ^2 (df=10)	Bowker's p-value	Reject symmetry
Roberta + ResNet50	1238.56	$\ll 0.05$	✓	91.85	2.33×10^{-15}	✓
Roberta + ViT	1333.14	$\ll 0.05$	✓	132.10	0.00	✓
DeBERTa + ViT	1480.57	$\ll 0.05$	✓	118.33	0.00	✓
DeBERTa + ResNet50	1217.24	$\ll 0.05$	✓	166.16	0.00	✓
SBERT + ResNet50	1070.45	$\ll 0.05$	✓	82.030	2.00×10^{-13}	✓

Table 5: Significance-test results comparing the multimodal model against the baseline on Factify 2.

Models w/o Contrastive Head	McNemar's χ^2	McNemar's p-value	Significance at $\alpha = 0.05$	Bowker's χ^2 (df=10)	Bowker's p-value	Reject symmetry
Roberta + ResNet50	1129.91	$\ll 0.05$	✓	75.64	3.56×10^{-12}	✓
Roberta + ViT	1243.45	$\ll 0.05$	✓	140.44	0.00	✓
DeBERTa + ViT	965.16	$\ll 0.05$	✓	122.93	0.00	✓
DeBERTa + ResNet50	1195.33	$\ll 0.05$	✓	153.76	0.00	✓
SBERT + ResNet50	891.88	$\ll 0.05$	✓	100.06	0.00	✓

Table 6: Significance-test results comparing the multimodal model against the baseline on Factify 2.

over the baseline. Whereas Bowker's results shows that our models make structured, non-random improvements in class predictions.

Discordant pair analysis: we examined our model improvements over the baseline by comparing predictions from each variant of the models, using 2×2 McNemar's contingency setup. Specifically, we considered two key cases: In first we denoted (**b**) as the number of instances, where the baseline is correct and our model errors, and, second where (**c**) as the opposite our model is correct and the baseline errors. These corresponds to the off-diagonal cells in McNemar's 2×2 contingency table. Across all five configurations, (c) consistently and substantially outnumbers (b) as illustrated in Figures 7. This suggests that our models make meaningful improvements rather than random changes, consistently correcting the baseline's mistakes. These findings are consistent with the significance tests further supporting the effectiveness of our approach.

Qualitative Analysis: To better understand model behavior, we examined the mismatches between our approach and the baseline. As shown in Figure 3, the baseline overlooked the visual modality entirely and based its decision on text alone, predicting *Insufficient Text*. In contrast, our model incorporated the OCR-detected credit "Helen Sloan/HBO" from the image, identifying it as a licensed promotional photograph a detail suggesting the image does not contribute new factual content. By recognizing the lack of substantive support in both text and image, our model rightly predicted *Insufficient Multimodal*. This highlights how image provenance, even in OCR form, enhances factual reasoning. Additional examples are discussed below:

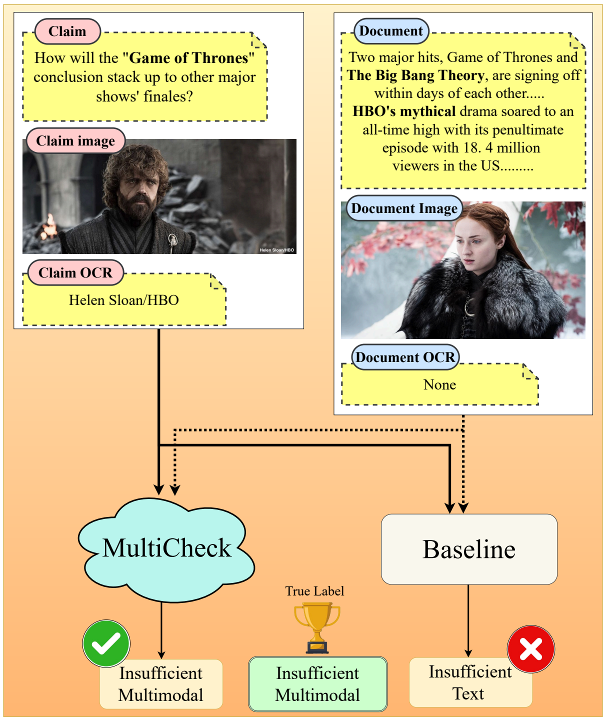

Claim
How will the "Game of Thrones" conclusion stack up to other major shows' finales?
Claim image
Claim OCR
Helen Sloan/HBO
Document
Two major hits, Game of Thrones and The Big Bang Theory, are signing off within days of each other..... HBO's mythical drama soared to an all-time high with its penultimate episode with 18. 4 million viewers in the US.........
Document Image
Document OCR
None
MultiCheck
Baseline
True Label
Insufficient Multimodal
Insufficient Multimodal
Insufficient Text

Figure 3: Example of qualitative analysis, Sample from the dataset.

Additional examples of qualitative analysis: We have presented detailed examples mentioned by the original ID of the samples as per the dataset, comparing predictions from our approach against the baseline. These examples illustrate how OCR-based information like photographer credits, agency marks, or image overlays can either resolve or confuse multimodal evidence verification.

Example A: When MultiCheck outperforms the baseline

- **ID 4681:** The baseline model, using only textual signals, interpreted the phrase "Made in India vaccines" as concrete evidence supporting the claim thus predicting *Support_Text*. However, the OCR-extracted string "COLO STORAE" (interpreted as "cold storage") hinted at logistical or infrastructure gaps in the vaccine delivery, casting doubt on the sufficiency of the claim's evidence. Our model's multimodal fusion recognized this ambiguity and opted for *Insufficient_Text*, aligning with the ground truth. This example shows how even noisy OCR text can surface hidden context missed by text-only models.
- **ID 6968:** The baseline inferred support based on name and context matches in the text (e.g., "Governor Jagdeep Dhankar"), resulting in a *Support_Text* prediction. However, the image's OCR output "ANI" indicated it was a generic press photo from a news agency, lacking evidentiary value. Our model correctly interpreted this, combining the weak textual alignment with the non-informative image tag to determine that neither modality offered enough proof thus choosing *Insufficient_Multimodal*. This highlights the

model's ability to ignore superficial visual cues.

Example B: When baseline outperforms MultiCheck

- **ID 7171:** Here, our model was overly cautious. The OCR tag "ANI" led it to interpret the image as a generic stock photo, causing it to discount visual evidence. It then judged the claim as lacking visual proof and labeled it Insufficient_Multimodal. However, the textual portion "will lay foundation stone" is a direct and verifiable event announcement, and the image (even if generic) serves as a credible contextual anchor. The baseline, focusing on the assertive language in the text, correctly predicted Support_Multimodal. This example shows that OCR can occasionally mislead, especially when image content is generic but still contextually supportive.

- **ID 6707:** Our model was likely confused by noisy OCR clutter: multiple "ANI" tags and irrelevant text like "WINE SHOP" and exam references (UPSC/MPSC). These spurious signals may have interfered with alignment, prompting it to under-call the claim's evidential value. The baseline, unencumbered by these distractions, focused on the strong textual indicator "deployed" and correctly chose Support_Text. This example underlines the need for OCR gating or filtering in future iterations.

Key Observations:

- **OCR text can be highly informative** especially when images include meta-tags, banners, or visual overlays not present in article text.

- **Failure cases arise** when OCR includes irrelevant or misleading tokens, causing the model to over- or under-attend to visuals.

- **Future directions:** We plan to incorporate an OCR quality gating mechanism and synthetic noisy-OCR augmentation to improve model robustness.

7 Conclusion:

This paper introduces a unified framework for fine-grained multimodal fact-checking that jointly reasons over textual and visual evidence. Our architecture integrates structured representations from pre-trained language and vision models using a relational fusion module. It further employs a contrastive learning objective to enhance cross-modal alignment. This design allows the model to better capture fine agreements and contradictions between claims and evidence across modalities. As our approach outperforms the baseline, across multiple configurations. It achieves particularly notable gains in complex classes such as "*Insufficient*" and "*Refute*", where multimodal reasoning is critical. Statistical tests and qualitative analyses confirm that the improvements are consistent and meaningful. These gains are systematic, rather than incidental. Overall, our results emphasize the importance of explicit cross-modal alignment and representation learning in advancing automated fact verification.

8 Limitation:

While our approach shows strong empirical performance, several limitations remain:

- Dependence on OCR quality: The model incorporates OCR-extracted text from images, which can vary widely in accuracy and relevance. In cases of noisy or misleading OCR outputs, the model may misclassify due to spurious visual-textual alignment.

- No evidence retrieval component: Our framework assumes that relevant evidence both textual and visual is already provided. It does not includes any retrieval pipeline to source additional or more reliable evidences from external knowledge bases or web sources.

- Limited visual understanding: Although image features are included via pre-trained encoders, the model lacks deeper visual reasoning capabilities such as object detection, scene understanding, or temporal cues that could improve evidence grounding.

- Restricted modalities: The current system handles only text and image modalities. It does not address audio, video, or temporal multimodal misinformation, which are common in modern social media content.

References

Sara Abdali, Sina Shaham, and Bhaskar Krishnamachari. 2024. Multi-modal misinformation detection: Approaches, challenges and opportunities. *ACM Computing Surveys*, 57(3):1–29.

Sahar Abdelnabi, Rakibul Hasan, and Mario Fritz. 2022. 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*, pages 14940–14949.

Mubashara Akhtar, Michael Schlichtkrull, Zhijiang Guo, Oana Cocarascu, Elena Simperl, and Andreas Vlachos. 2023. Multimodal automated fact-checking: A survey. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5430–5448.

Firoj Alam, Stefano Cresci, Tanmoy Chakraborty, Fabrizio Silvestri, Dimiter Dimitrov, Giovanni Da San Martino, Shaden Shaar, Hamed Firooz, and Preslav Nakov. 2022. A survey on multimodal disinformation detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6625–6643.

Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. 2018. Where is your evidence: Improving fact-checking by justification modeling. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 85–90, Brussels, Belgium. Association for Computational Linguistics.

Javahir Askari. 2023. Deepfakes and synthetic media: What are they and how are techuk members taking steps to tackle misinformation and fraud. *TechUK. August*, 18.

Tobias Braun, Mark Rothermel, Marcus Rohrbach, and Anna Rohrbach. 2024. Defame: Dynamic evidence-based fact-checking with multimodal experts. *arXiv preprint arXiv:2412.10510*.

Maria Mercedes Ferreira Caceres, Juan Pablo Sosa, Jannel A Lawrence, Cristina Sestacovschi, Atiyah Tidd-Johnson, Muhammad Haseeb UI Rasool, Vinay Kumar Gadamidi, Saleha Ozair, Krunal Pandav, Claudia Cuevas-Lou, et al. 2022. The impact of misinformation on the covid-19 pandemic. *AIMS public health*, 9(2):262.

Recep Firat Cekinel, Pinar Karagoz, and Çağrı Çöltekin. 2025a. Multimodal fact-checking with vision language models: A probing classifier based solution with embedding strategies. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4622–4633.

Recep Firat Cekinel, Pinar Karagoz, and Çağrı Çöltekin. 2025b. Multimodal fact-checking with vision language models: A probing classifier based solution with embedding strategies. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4622–4633, Abu Dhabi, UAE. Association for Computational Linguistics.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680.

Giandomenico Di Domenico, Jason Sit, Alessio Ishizaka, and Daniel Nunan. 2021. Fake news, social media and marketing: A systematic review. *Journal of business research*, 124:329–341.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, G Heigold, S Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Wei-Wei Du, Hong-Wei Wu, Wei-Yao Wang, and Wen-Chih Peng. 2023a. Team triple-check at factify 2: Parameter-efficient large foundation models with feature representations for multi-modal fact verification. In *DE-FACTIFY@ AAAI*.

Wei Wei Du, Hong Wei Wu, Wei Yao Wang, and Wen Chih Peng. 2024a. Team triple-check at factify 2: Parameter-efficient large foundation models with feature representations for multi-modal fact verification. In *CEUR Workshop Proceedings*, volume 3555. CEUR-WS.

Wei Wei Du, Hong Wei Wu, Wei Yao Wang, and Wen Chih Peng. 2024b. Team triple-check at factify 2: Parameter-efficient large foundation models with feature representations for multi-modal fact verification. In *CEUR Workshop Proceedings*, volume 3555. CEUR-WS.

Yuxuan Du, Yaqing Yao, Chenghao Wang, and Kai Shu. 2023b. Precofactv2: Improving multimodal fact-checking via precise and coarse-grained evidence matching. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 14671–14685.

Haisong Gong, Weizhi Xu, Shu Wu, Qiang Liu, and Liang Wang. 2024. Heterogeneous graph reasoning for fact checking over texts and tables. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, pages 100–108.

Michael Hameleers, Thomas E Powell, Toni GLA Van Der Meer, and Lieke Bos. 2020. A picture paints a thousand lies? the effects and mechanisms of multimodal disinformation and rebuttals disseminated via social media. *Political Communication*, 37(2):281–301.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

Mohammed Abdul Khaliq, Paul Yu-Chun Chang, Mingyang Ma, Bernhard Pflugfelder, and Filip Miletić. 2024. RAGAR, your falsehood radar: RAG-augmented reasoning for political fact-checking using multimodal large language models. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pages 280–296, Miami, Florida, USA. Association for Computational Linguistics.

Bogoan Kim, Aiping Xiong, Dongwon Lee, and Kyungsik Han. 2021. A systematic review on fake news research through the lens of news creation and consumption: Research efforts, challenges, and future directions. *PloS one*, 16(12):e0260080.

Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. 2018. Bilinear attention networks. *Advances in neural information processing systems*, 31.

Yang Liu, Guanbin Li, and Liang Lin. 2023. Cross-modal causal relational reasoning for event-level visual question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):11624–11641.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Rahul Mishra, Dhruv Gupta, and Markus Leippold. 2020. Generating fact checking summaries for web claims. In *EMNLP W-NUT 2020: Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Davide Antonio Mura, Marco Usai, Andrea Loddo, Manuela Sanguinetti, Luca Zedda, Cecilia Di Ruberto, and Maurizio Atzori. 2025. Is it fake or not? a comprehensive approach for multimodal fake news detection. *Online Social Networks and Media*, 47:100314.

Gillian Murphy, Constance de Saint Laurent, Megan Reynolds, Omar Aftab, Karen Hegarty, Yuning Sun, and Ciara M Greene. 2023. What do we study when we study misinformation? a scoping review of experimental research (2016-2022). *Harvard Kennedy School Misinformation Review*.

Tatsuhiro Nakamura, David Levy, William Yang Wang, Preslav Nakov, and Akiko Aizawa. 2019. Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. In *Proceedings of the 10th International Workshop on Multimedia and the Web*. ACM.

Preslav Nakov, David Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni da San Martino. 2021. Automated fact-checking for assisting human fact-checkers. In *30th International Joint Conference on Artificial Intelligence, IJCAI 2021*, pages 4551–4558. International Joint Conferences on Artificial Intelligence.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

Ikumi Sata, Motoki Amagasaki, and Masato Kiyama. 2025. Multimodal retrieval method for images and diagnostic reports using cross-attention. *AI*, 6(2):38.

Sandeep Suryavardan, Saket Deshpande, Luchen Li, Tanishq Dadu, Prateek Kalluri, Shaden Shaar, Lei Zhang, and Preslav Nakov. 2023. Factify 2: Towards fine-grained multimodal fact verification with social media evidence. In *Findings of the Association for Computational Linguistics: EMNLP 2023*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819.

Shivani Tufchi, Ashima Yadav, and Tanveer Ahmed. 2023. A comprehensive survey of multimodal fake news detection techniques: advances, challenges, and opportunities. *International Journal of Multimedia Information Retrieval*, 12(2):28.

William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.

Xinyu Wang, Min Gui, Yong Jiang, Zixia Jia, Nguyen Bach, Tao Wang, Zhongqiang Huang, and Kewei Tu. 2022. Ita: Image-text alignments for multi-modal named entity recognition. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3176–3189.

Barry Menglong Yao, Aditya Shah, Lichao Sun, Jin-Hee Cho, and Lifu Huang. 2023a. End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2733–2743.

Yaqing Yao, Yuxuan Du, Chenghao Wang, and Kai Shu. 2023b. Mocheg: Multimodal contrastive hybrid evidence generation for multimodal fact-checking. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10433–10447.

Dimitrina Zlatkova, Preslav Nakov, and Ivan Koychev. 2019. Fact-checking meets fauxtography: Verifying claims about images. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2099–2108, Hong Kong, China. Association for Computational Linguistics.

Appendix

A Related Work

Recent research in automated fact-checking has highlighted the growing importance of incorporating multiple modalities, to tackle the diverse and evolving forms of misinformation (Abdelnabi et al., 2022). Early works, such as FEVER (Thorne et al., 2018) and CLEF2018 (Nakov et al., 2021) are primarily focused on verifying textual claims, laying foundational methods for claim verification based solely on textual evidence. However, later studies found that misinformation exploits images, videos, and audio alongside text to build convincing narratives (Hameleers et al., 2020; Alam et al., 2022). These studies have revealed the limitations of purely text-based fact-checking methods and sparked a shift toward multimodal fact-checking. To tackle the limitation of text-based methods, systems were designed to jointly process and reason over diverse types of content. For example, several studies, such as (Du et al., 2023b) and (Zlatkova et al., 2019; Khaliq et al., 2024), have explored architectures that integrated textual and visual features. These models employ mechanisms like attention or contrastive learning to enhance detection accuracy. Recent work by (Cekinel et al., 2025b) shows that a probing classifier combining separate text and image embeddings can outperform intrinsic VLM features on datasets like Factify 2. In addition to these developments, comprehensive surveys, such as (Akhtar et al., 2023), offers detailed overview regarding the emerging field of multimodal fact-checking. They highlighted both the technical challenges and the promising research directions ahead. Key challenges include aligning information across modalities, managing incomplete or noisy evidence, and ensuring scalability for practical deployment.

Despite significant progress, effectively integrating multimodal information remains an open research problem. This challenge continues to motivate the development of new architectures and learning methods. Robust fact verification in multimodal contexts still requires innovative solutions.

B Additional details on datasets:

In this section, we provide a comprehensive overview of the Factify 2 dataset, including representative examples and insights into its distributional characteristics.

B.1 Factify 2 (Du et al., 2024a):

Factify 2 is a large-scale multimodal fact verification dataset comprising 42,500 human-annotated claims drawn from diverse domains such as politics, health, environment, and global affairs. Claims were curated from credible news media sources across India and the United States. Specifically, true claims were collected from official Twitter accounts of verified news organizations, while false claims were sourced from authoritative fact-checking platforms, including *PolitiFact*[1], *Alt News*[2], and *BoomLive*[3].

A distinctive strength of Factify 2 lies in its rich multimodal evidence composition. For each claim, the dataset includes (i) textual evidence retrieved from external news articles, (ii) claim-associated images (typically extracted from the header sections of original posts), and (iii) image evidence from supporting or refuting documents. Each sample is annotated with one of five fine-grained labels that describe the relationship between the claim and the retrieved evidence. The labels they considered are: ***"Support_Text"*** (the textual evidence supports the claim), ***"Support_Multimodal"*** (both textual and visual evidence jointly support the claim), ***"Insufficient_Text"*** (textual evidence is present but insufficient to verify the claim), ***"Insufficient_Multimodal"*** (both textual and image evidence are insufficient), and ***"Refute"*** (the evidence directly contradicts the claim).

The dataset supports training and evaluation of multimodal models in realistic settings, where claims are to be verified using diverse evidence types. A representative sample from the dataset is presented in Figure 4 for illustration.

Due to the unavailability of ground truth labels for the test split, we follow the protocol adopted by Cekinel et al. (2025a) and repurpose the original validation set for testing. To maintain a development split, 7,500 samples were randomly selected from the original training set to form a new validation set, preserving the original class distribution across all partitions.

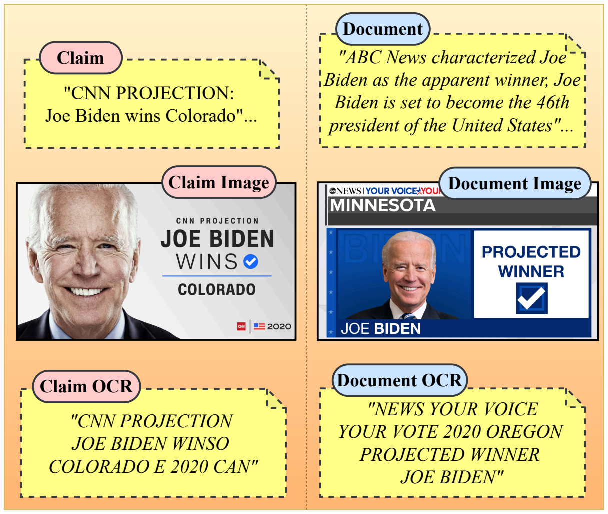


Figure 4: Example of a Sample from the dataset.

B.2 Additional details on baselines:

In this section, we reported the details of the individual baseline method considered in our study.

- **Pro-CoFactv2**: It is proposed by Du et al. (2023a). The authors developed a parameter-efficient multi-modal fact verification model that leverages pretrained language and vision encoders with minimal task-specific tuning. Their system is designed with two key components: (i) a feature extraction module built upon large foundation models, and (ii) a lightweight classifier module that integrates contrastive and classification objectives. In the feature extraction stage, the text encoder is initialized from `microsoft/deberta-large`, while the image encoder is derived from `microsoft/swinv2-base-patch4-window8-256`. Both encoders are kept unfrozen, allowing full gradient updates during training. Text and image embeddings are projected into a joint space using adapter modules with a bottleneck dimension of 64. The model applies a linear fusion technique to integrate both modalities effectively.

  For the training objective, they employed a multi-loss setup that combines cross-entropy loss for veracity classification and supervised contrastive loss to better align intra-class examples in the embedding space. Specifically,

[1] https://www.politifact.com
[2] https://www.altnews.in
[3] https://www.boomlive.in

the contrastive loss is scaled by a factor of 0.3, while the classification loss retains a full weight of 1. The model is trained using a batch size of 32 and a learning rate of 5×10^{-5}, for 20 epochs. Evaluation is performed after every epoch to track performance.

Our reproduction strictly follows the original configuration settings to preserve accuracy with reduced token length due to GPU constraint. We have used a NVIDIA A100 GPU of 80 GB capacity to perform our experiments. This include fixing the seed value to 42, 57, 196, 906 for reproducibility, setting the maximum sequence length to 128 tokens, using 12 attention heads, and a dropout rate of 0.1. The hidden dimensionality in intermediate layers is set to 256, aligned with the FakeNet backbone mentioned in the architecture.

C Additional details on modules:

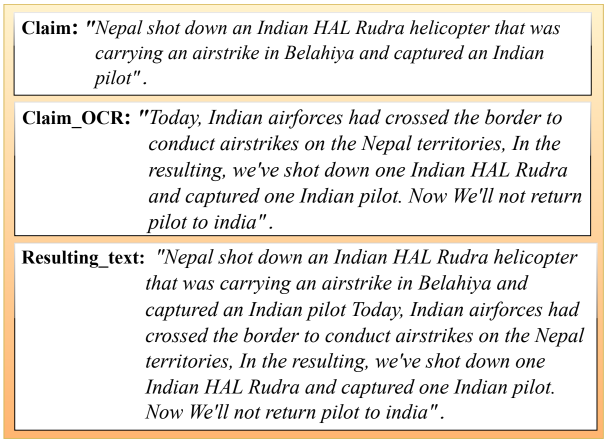
This appendix provides detailed architectural and implementation specifications for our proposed framework. The following sections correspond to the modules introduced in section 3

C.1 Details on text module:

The text module processes both the claim text and the OCR-extracted text from associated images. These are concatenated for each instance to create a richer and more context-aware representation. Figure 5 demonstrates, how OCR text can provide crucial information absent from the original claim text, helping the model detect contradictions necessary for accurate veracity classification.

Given the batch of claims and document, the resulting textual inputs are tokenized using the pre-trained language models such as RoBERTa (Liu et al., 2019), DeBERTa (He et al., 2020), or SBERT (Reimers and Gurevych, 2019). Yielding token input ID's and attention mask tensors of shape $\mathbb{R}^{b \times L}$, where $\boldsymbol{b}$ is the batch size, $\boldsymbol{L}$ is the maximum sequence length and d_{text} is the input dimension of the projection layer for text (e.g., for Roberta and DeBERTa, it is 1024 and for SBERT it is 768). These inputs are passed through a shared encoder of language models. Producing contextualized embeddings of shape:

$$\mathbf{H}_{\text{text}} \in \mathbb{R}^{b \times L \times d_{\text{text}}}$$



Claim: *"Nepal shot down an Indian HAL Rudra helicopter that was carrying an airstrike in Belahiya and captured an Indian pilot".*

Claim_OCR: *"Today, Indian airforces had crossed the border to conduct airstrikes on the Nepal territories, In the resulting, we've shot down one Indian HAL Rudra and captured one Indian pilot. Now We'll not return pilot to india".*

Resulting_text: *"Nepal shot down an Indian HAL Rudra helicopter that was carrying an airstrike in Belahiya and captured an Indian pilot Today, Indian airforces had crossed the border to conduct airstrikes on the Nepal territories, In the resulting, we've shot down one Indian HAL Rudra and captured one Indian pilot. Now We'll not return pilot to india".*

Figure 5: *The claim accuses Nepal of shooting down an Indian military helicopter. However, the OCR text from the image contradicts this, suggesting Indian forces crossed the border instead implying aggression from India, not Nepal. Without this OCR text, the model may misinterpret or miss this contradiction.* This Example illustrates how OCR text contributes critical contextual information, enabling the model to detect contradictions and correctly assign a "Refute" label.

We extract the [CLS] token embeddings from each sequence, resulting in a fixed-size representations of: $\mathbf{h}_{[\text{CLS}]} \in \mathbb{R}^{b \times d_{\text{text}}}$. These embeddings are projected into a common latent space via a fully connected layer denoted by:

$$\mathbf{V}_{\text{text}} = \mathbf{W}_{\text{text}} \cdot \mathbf{h}_{[\text{CLS}]} + \mathbf{B}, \quad \mathbf{V}_{\text{text}} \in \mathbb{R}^{b \times h}$$

where $\boldsymbol{h}$ denotes the dimensionality of the shared latent space, $\mathbf{B}$ denotes the bias term and $\mathbf{W}_{\text{text}}$ is the weight matrix of a linear layer that maps the text encoder's output into a shared latent space.

C.2 Details on image module:

In this section, we reported the detailed architectural description of our image module. Each claim and document image is preprocessed and passed through a shared vision encoders such as ResNet50 (He et al., 2016) or Vision Transformer (ViT) (Dosovitskiy et al., 2020).

This produces high-dimensional visual features of shape: $\mathbf{f}_{\text{img}} \in \mathbb{R}^{b \times d_{\text{img}}}$, where d_{img} is the dimensionality of the raw output from the image encoder before projection into a shared latent space, and have values 2048 (ResNet) or 768 (ViT) respectively. These feature embeddings are independently projected into the same latent space via linear layers denoted by:

$$\mathbf{V}_{\text{img}} = \mathbf{W}_{\text{img}} \cdot \mathbf{f}_{\text{img}} + \mathbf{B}, \quad \mathbf{V}_{\text{img}} \in \mathbb{R}^{b \times h}$$

$\mathbf{W}_{\text{img}}$ is the weight matrix for the projection layer that maps raw image features into the same shared latent space.

At the end of this step, we obtain four core representative vectors of shape $b \times h$:

- $\mathbf{V}_{\textit{claim text}}$: text embedding for claim
- $\mathbf{V}_{\textit{doc text}}$: text embedding for document
- $\mathbf{V}_{\textit{claim img}}$: image embedding for claim
- $\mathbf{V}_{\textit{doc img}}$: image embedding for document

C.3 Details on fusion module:

After obtaining both text and image embeddings for claim and document pairs, we concatenate the corresponding text and image features for both the claim and the document, yielding two **integrated representations**: claim representation ($\mathbf{V}_{\text{claim repr}}$) and document representation ($\mathbf{V}_{\text{doc repr}}$), represented by:

- $\mathbf{V}_{\text{claim repr}} = \mathbf{V}_{\textit{claim text}} \oplus \mathbf{V}_{\textit{claim img}}$
- $\mathbf{V}_{\text{doc repr}} = \mathbf{V}_{\textit{doc text}} \oplus \mathbf{V}_{\textit{doc img}}$

where $\oplus$ shows the operation of concatenation and

- $\left[\mathbf{V}_{\text{claim repr}}; \mathbf{V}_{\text{doc repr}}\right] \in \mathbb{R}^{b \times 2h}$

To capture fine-grained representation, we perform two element-wise operations between claim representation and document representation: the **difference** ($\mathbf{V}_{\text{diff}}$) to emphasize contrasts and the **product** ($\mathbf{V}_{\text{prod}}$) to capture alignment, represented by:

- $\mathbf{V}_{\text{diff}} = \left|\mathbf{V}_{\text{claim repr}} \ominus \mathbf{V}_{\text{doc repr}}\right|$
- $\mathbf{V}_{\text{prod}} = \mathbf{V}_{\text{claim repr}} \otimes \mathbf{V}_{\text{doc repr}}$

where $\ominus$ shows element-wise difference between claim representation and document representation, and $\otimes$ shows the element-wise product between claim and document representation.

These four vectors: claim representation, document representation, difference, and product are then concatenated to form a single fused vector of shape $\boldsymbol{b} \times 8\boldsymbol{h}$, denoted by:

$$\mathbf{V}_{\text{fused}} = \mathbf{V}_{\text{diff}} \oplus \mathbf{V}_{\text{prod}} \oplus \mathbf{V}_{\text{claim repr}} \oplus \mathbf{V}_{\text{doc repr}}$$

This fused vector ($\mathbf{V}_{\text{fused}}$) is passed through a fully connected network, which consists of a linear transformation layer and a non-linear activation (for e.g, GELU) to output a **unified representation** represeneted by:

- $\mathbf{V}_{\text{final}} = \text{FFN}(\mathbf{V}_{\text{fused}}) \in \mathbb{R}^{b \times h}$

As part of our fusion module, we incorporate a contrastive projection head to learn discriminative representations that align semantically related claim-document pairs. After computing the multimodal representations of the claim and document, denoted as: $\mathbf{V}_{\text{claim repr}}$ and $\mathbf{V}_{\text{doc repr}}$, each vector is passed through a dedicated projection network defined as:

$$\mathbf{Z}_{\text{claim}} = f_{\text{proj}}\left(\mathbf{V}_{\text{claim repr}}\right)$$
$$\mathbf{Z}_{\text{doc}} = f_{\text{proj}}\left(\mathbf{V}_{\text{doc repr}}\right)$$

where f_{proj} consists of:

- A linear transformation to reduce dimensionality from $2\boldsymbol{h}$ to $\boldsymbol{h}$
- A ReLU non-linearity
- A second linear layer projecting the vector back to dimension $\boldsymbol{h}$

So, Formally:

- $f_{\text{proj}}(\mathbf{v}) = \mathbf{W}_2 \cdot \text{ReLU}\left(\mathbf{W}_1 \cdot \mathbf{v} + \mathbf{B}_1\right) + \mathbf{B}_2$

where $\mathbf{W}_1 \in \mathbb{R}^{h \times 2h}, \mathbf{W}_2 \in \mathbb{R}^{h \times h}$, $\mathbf{B_1}$ & $\mathbf{B_2}$ denotes the bias term of the respective layers and $\mathbf{v}$ is the concatenated multimodal representation of either the claim or the document i.e. $\mathbf{V}_{\text{claim repr}}$ and $\mathbf{V}_{\text{doc repr}}$. We have used two distinct linear layers with a ReLU bottleneck in between, first squeezing $2\boldsymbol{h} \rightarrow \boldsymbol{h}$ then re-expanding $\boldsymbol{h} \rightarrow \boldsymbol{h}$. The depth and non-linearity that we used here is crucial to give the projection head enough capacity to learn richer and contrastingly useful embeddings.

The resulting embeddings, $\mathbf{Z}_{\text{claim}}$ and $\mathbf{Z}_{\text{doc}}$, serve as inputs to the contrastive learning objective described in Appendix C.4

C.4 Details on classification module:

The final fused representation ($\mathbf{V}_{\text{final}}$) is passed into a classification head, a linear layer followed by a softmax activation to produce logits for five veracity classes. i.e $C = 5$: *Support_Text, Support_Multimodal, Insufficient_Text, Insufficient_Multimodal,* and *Refute* denoted by:

- $\mathbf{P} = \mathbf{W}_{\text{cls}} \cdot \mathbf{V}_{\text{final}} + \mathbf{B}, \quad \mathbf{P} \in \mathbb{R}^{b \times 5}$

Model training uses standard cross-entropy loss computed over these logits denoted by:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{b} \sum_{i=1}^{b} \log \left(\frac{\exp\left(Z_{i,y_i}\right)}{\sum_{j=1}^{5} \exp\left(Z_{i,j}\right)} \right)$$

To encourage the model to learn modality-consistent and discriminative representations, we include a symmetric InfoNCE loss computed on the projected embeddings ($\mathbf{Z}_{\text{claim}}$ and $\mathbf{Z}_{\text{doc}}$) from the contrastive head. First we do the normalization, each row of the projected claim and document embeddings is normalized to unit length:

$$\hat{\mathbf{z}} = \frac{\mathbf{z}}{\|\mathbf{z}\|_2}$$

We compute the similarity matrix $\mathbf{S}$ as:

$$\mathbf{S} = \hat{\mathbf{Z}}_{\text{claim}} \cdot \hat{\mathbf{Z}}_{\text{doc}}^{\top}$$

which yields a matrix of shape $(b \times b)$.

The element S_{ij} represents the similarity between the i-th claim and the j-th document in the batch. Then we compute the InfoNCE loss:

Row-wise InfoNCE (claim → doc):

$$\text{Loss}_{\text{claim}\rightarrow\text{doc}} = -\frac{1}{b} \sum_{i=1}^{b} \log \frac{\exp\left(S_{ii}/\tau\right)}{\sum_{j=1}^{b} \exp\left(S_{ij}/\tau\right)}$$

Column-wise InfoNCE (doc → claim):

$$\text{Loss}_{\text{doc}\rightarrow\text{claim}} = -\frac{1}{b} \sum_{j=1}^{b} \log \frac{\exp\left(S_{jj}/\tau\right)}{\sum_{i=1}^{b} \exp\left(S_{ij}/\tau\right)}$$

The final contrastive loss is computed as the average of the two directions:

$$\mathcal{L}_{\text{contrastive}} = \frac{1}{2} \left(\text{Loss}_{\text{claim}\rightarrow\text{doc}} + \text{Loss}_{\text{doc}\rightarrow\text{claim}} \right)$$

where τ is a temperature hyperparameter controlling the sharpness of the softmax distribution. So, The total training objective combines the supervised classification loss with the contrastive learning signal:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \lambda \cdot \mathcal{L}_{\text{contrastive}}$$

where λ balances the influence of the contrastive objective. In our experiments, we set λ = 0.1. This composite loss encourages the model not only to correctly classify claim-evidence pairs but also to learn a representation space where semantically related claims and documents are closely aligned while unrelated pairs are well-separated.

C.5 Hyperparameter details:

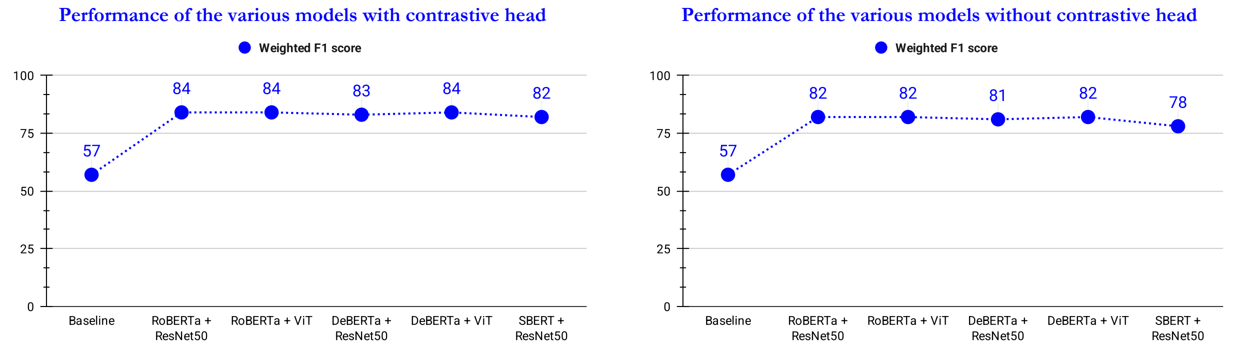
We performed a thorough hyperparameter search over all key modeling and training parameters to identify the settings that yielded the best validation performance. Table 9 summarizes every hyperparameter and their final values used for both the contrastive-head and non-contrastive variants. All experiments were carried out on one NVIDIA A100 80 GB GPU.

C.6 Evaluation metrics:

We evaluate our model performance primarily via the F1 score, which is the harmonic mean of precision and recall for each class. In our multi-class setting, class frequencies vary substantially. Unweighted F1 scores per label may not reflect true performance, as models might perform well only on the dominant class and still appear effective.

To summarize performance across all classes while accounting for class imbalance, we use the **weighted F1 score**. This metric assigns each class's F1 score a weight based on its support that is, the number of true instances for that class. Formally, if F_i is the F_1 for class i and n_i its support, then

$$\text{Weighted-F}_1 = \frac{\sum_i n_i \, F_i}{\sum_i n_i}.$$



(a) Performance of the various models **with** contrastive head.

(b) Performance of the various models **without** contrastive head.

Figure 6: Comparison of model performance on the Factify 2 dataset, with vs. without the contrastive-learning head. Each point on the line plot is the weighted F_1 score (mean across seeds).

Models Models w/ Contrastive Head	**Metric**	**Support Text**	**Support Multimodal**	**Insufficient Text**	**Insufficient Multimodal**	**Refute**
Roberta + ResNet50	Precision	0.77 (±0.03)	**0.82** (±0.01)	0.82 (±0.02)	0.76 (±0.01)	**1.00** (±0.00)
	Recall	0.77 (±0.01)	**0.84** (±0.02)	0.81 (±0.01)	0.76 (±0.02)	**0.99** (±0.00)
Roberta + ViT	Precision	0.78 (±0.01)	0.82 (±0.01)	**0.86** (±0.01)	0.75 (±0.02)	**1.00** (±0.00)
	Recall	0.77 (±0.02)	**0.85** (±0.02)	0.79 (±0.01)	0.79 (±0.01)	**1.00** (±0.00)
DeBERTa + ViT	Precision	0.75 (±0.03)	0.81 (±0.01)	**0.83** (±0.02)	0.72 (±0.01)	**1.00** (±0.01)
	Recall	0.74 (±0.03)	**0.83** (±0.08)	0.73 (±0.03)	0.78 (±0.02)	**1.00** (±0.01)
DeBERTa + ResNet50	Precision	0.75 (±0.02)	0.83 (±0.01)	**0.84** (±0.00)	0.73 (±0.03)	**1.00** (±0.00)
	Recall	0.78 (±0.01)	**0.80** (±0.04)	0.78 (±0.01)	0.78 (±0.02)	**1.00** (±0.00)
SBERT + ResNet50	Precision	0.74 (±0.01)	0.80 (±0.01)	**0.83** (±0.01)	0.77 (±0.01)	**1.00** (±0.00)
	Recall	0.76 (±0.02)	**0.86** (±0.01)	0.76 (±0.01)	0.74 (±0.01)	**0.99** (±0.00)
Baseline	Precision	0.48 (±0.04)	0.50 (±0.04)	**0.50** (±0.01)	0.43 (±0.02)	**0.98** (±0.00)
	Recall	0.38 (±0.06)	**0.61** (±0.02)	0.44 (±0.07)	0.46 (±0.06)	**0.98** (±0.00)

Table 7: Performance of various model combinations on the Factify 2 dataset with contrastive head. Precision and Recall are reported per class (±std in).

Models Models w/o Contrastive Head	**Metric**	**Support Text**	**Support Multimodal**	**Insufficient Text**	**Insufficient Multimodal**	**Refute**
Roberta + ResNet50	Precision	**0.75** (±0.01)	0.81 (±0.04)	0.81 (±0.00)	0.73 (±0.02)	**1.00** (±0.00)
	Recall	0.73 (±0.02)	0.83 (±0.03)	0.76 (±0.02)	**0.77** (±0.05)	0.99 (±0.00)
Roberta + ViT	Precision	0.76 (±0.02)	0.82 (±0.01)	**0.83** (±0.02)	0.71 (±0.02)	1.00 (±0.00)
	Recall	**0.74** (±0.02)	**0.82** (±0.02)	0.75 (±0.03)	0.79 (±0.03)	1.00 (±0.00)
DeBERTa + ViT	Precision	0.75 (±0.03)	0.81 (±0.01)	0.83 (±0.02)	0.72 (±0.01)	1.00 (±0.00)
	Recall	**0.74** (±0.03)	0.83 (±0.02)	0.73 (±0.03)	**0.78** (±0.02)	1.00 (±0.00)
DeBERTa + ResNet50	Precision	0.73 (±0.01)	0.80 (±0.01)	0.82 (±0.02)	0.73 (±0.02)	**1.00** (±0.00)
	Recall	**0.75** (±0.00)	0.82 (±0.02)	0.74 (±0.02)	0.75 (±0.01)	1.00 (±0.00)
SBERT + ResNet50	Precision	0.71 (±0.02)	0.80 (±0.01)	0.78 (±0.01)	0.72 (±0.02)	1.00 (±0.00)
	Recall	0.72 (±0.01)	0.82 (±0.02)	0.72 (±0.04)	0.75 (±0.01)	0.99 (±0.00)
Baseline	Precision	0.48 (±0.04)	0.50 (±0.04)	0.50 (±0.01)	0.43 (±0.02)	0.98 (±0.00)
	Recall	0.38 (±0.06)	0.61 (±0.02)	0.44 (±0.07)	0.46 (±0.06)	0.98 (±0.00)

Table 8: Performance of various model combinations on the Factify 2 dataset. Precision and Recall are reported per class (±std in).

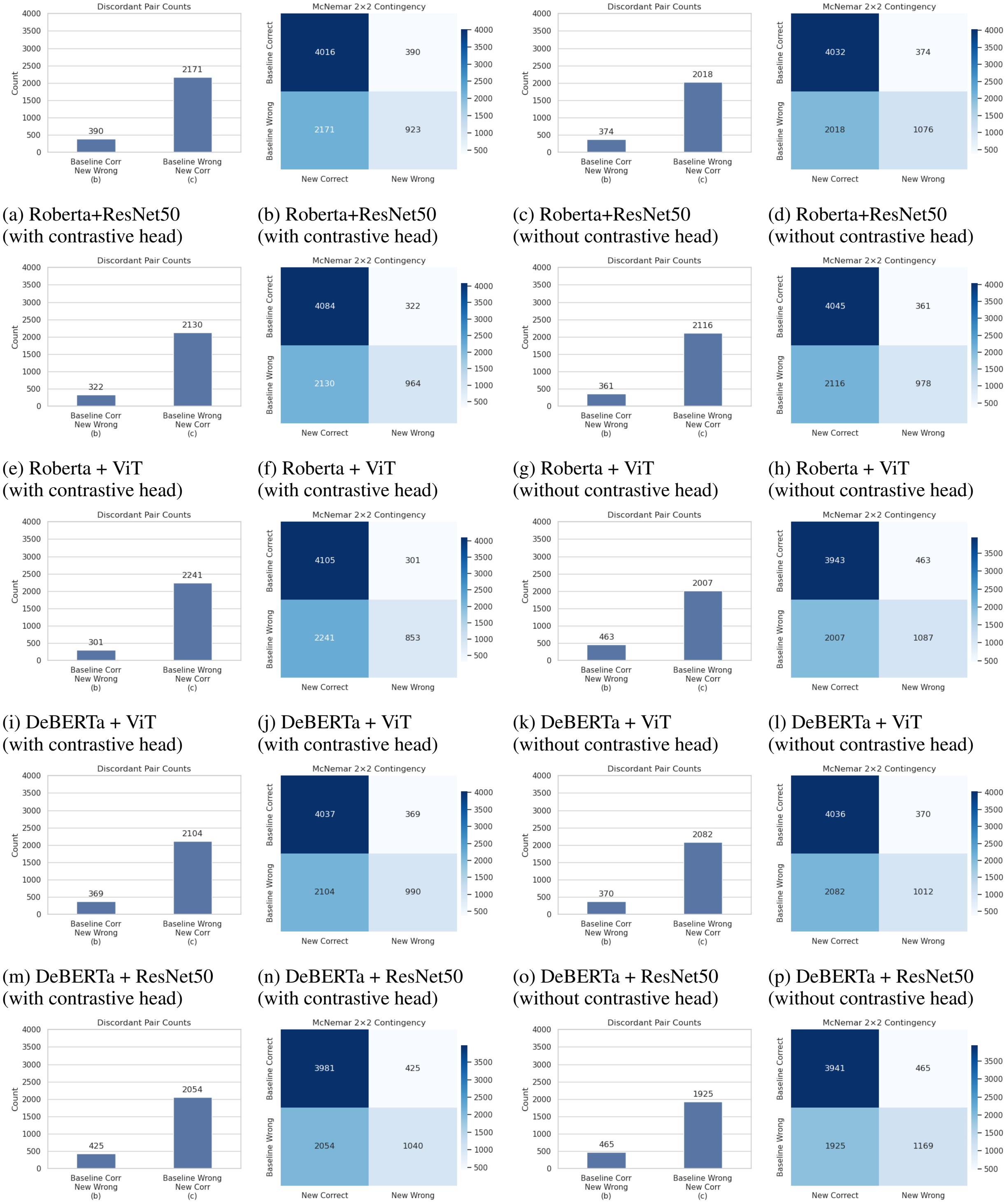
Component	Hyperparameter	Contrastive	Non-Contrastive
Reproducibility	Seeds	42/ 57 / 196 / 906	42/ 57 / 196 / 906
Tokenization	max_length	128	128
Model	Common embedding dim	256	256
	Fusion MLP dropout	0.1	0.1
Contrastive Loss	Temperature	0.1	–
	Loss weight (λ)	0.1	–
Optimization	Optimizer	Adam	Adam
	Learning rate	1×10^{-5}	1×10^{-5}
	Batch size	32	32
	Num workers	4	4
	Pin memory	True	True
	Epochs	20	20
LR Scheduling	Scheduler	ReduceLROnPlateau	ReduceLROnPlateau
	factor	0.5	0.5
	patience	2	2
Early Stopping	patience	5	5
	min_delta	1×10^{-4}	1×10^{-4}
Mixed Precision	AMP	`autocast` + `GradScaler`	`autocast` + `GradScaler`

Table 9: Hyperparameter settings for both the contrastive and non-contrastive variants.

This approach balances contributions from rare and frequent classes, making weighted F1 a fairer and more realistic summary of multi-class performance.

D Experimental Setup:

We conducted all experiments on a single NVIDIA A100 GPU using PyTorch and Hugging Face Transformers. Random seeds are choosen at 42, 57, 196 and 906 for reproducibility. For textual encoding, we experimented with a range of pre-trained language models. To extract visual features, we leveraged widely used vision backbones. Details of these components are provided in Section 3. Both modalities are projected into a shared latent space. The fusion module processes the claim and document embeddings, along with their element-wise difference and product. A contrastive projection head further transforms these representations for InfoNCE loss computation. Training uses the Adam optimizer with a learning rate of 1×10^{-5} batch size 32, and early stopping based on validation weighted F1. The contrastive loss weight λ is set to 0.1, and temperature τ to 0.1. A dropout of 0.1 is applied after the fusion layer. Experiments are conducted on the Factify 2 dataset, following the split protocol of Cekinel et al. (2025b). Evaluation metrics include weighted F1 score. Code and trained models will be released publicly for reproducibility.



(a) Roberta+ResNet50 (with contrastive head)

(b) Roberta+ResNet50 (with contrastive head)

(c) Roberta+ResNet50 (without contrastive head)

(d) Roberta+ResNet50 (without contrastive head)

(e) Roberta + ViT (with contrastive head)

(f) Roberta + ViT (with contrastive head)

(g) Roberta + ViT (without contrastive head)

(h) Roberta + ViT (without contrastive head)

(i) DeBERTa + ViT (with contrastive head)

(j) DeBERTa + ViT (with contrastive head)

(k) DeBERTa + ViT (without contrastive head)

(l) DeBERTa + ViT (without contrastive head)

(m) DeBERTa + ResNet50 (with contrastive head)

(n) DeBERTa + ResNet50 (with contrastive head)

(o) DeBERTa + ResNet50 (without contrastive head)

(p) DeBERTa + ResNet50 (without contrastive head)

(q) SBERT + ResNet50 (with contrastive head)

(r) SBERT + ResNet50 (with contrastive head)

(s) SBERT + ResNet50 (without contrastive head)

(t) SBERT + ResNet50 (without contrastive head)

Figure 7: Comparison of discordant-pair barplots (first column per model) and McNemar 2×2 contingency heatmaps (second column per model), both with and without the contrastive head. Models (top to bottom): Roberta + ResNet50, Roberta + ViT, DeBERTa + ViT, DeBERTa + ResNet50, SBERT + ResNet50.