Multimodal Detection of Manipulative Chart Narratives via Chart-to-Text Derendering and Transformer Classification

Anonymous ACL submission

Abstract

Charts are considered objective and authoritative displays of data, but the narratives accompanying them can rhetorically skew legitimate information. In this context, we introduce PlotTwist DataSet (PTDS), a benchmark of 300 expert-curated samples that aims to detect manipulative narratives in chart presentations while preserving the base visualizations' factualness. PTDS distinguishes three types of rhetorical distortion—unsupported causal inference, sensationalism (exaggeration), and oversimplification—but also allows for neutral cases. To address the task of multimodal rhetorical manipulation detection over structured data visualizations, we present a two-stage Matcha→BERT pipeline. The first stage processes charts into a structured textual representation. The second phase classifies the descriptions, paying particular attention to the captions, narratives, and evidence created. We compare our method to a two-stream ViT+BERT multimodal fusion baseline, using a strict 20×10 repeated stratified cross-validation. The model achieves a high macro-F1 score of 0.889, emphasizing the validity of combining textual reasoning with evidential information extracted from charts for better manipulation detection. This work fills a crucial void in multimodal misinformation research by focusing on discourse-level distortions that can occur in structured data visualizations, rather than the more widely explored image–text inconsistencies associated with natural images.

1 Introduction

Charts are often perceived as empirical and correct; they give the impression of precision and neutrality. However, what these visual representations ultimately mean is not only affected by the data they present but also the context shaped by the captions and narratives they contain (Chen et al., 2020). Framing can change the interpretations without changing the data. By aligning an image with text, audiences may be led to extreme claims to stand unchallenged (Stokes et al., 2024). Most research on misinformation has focused on news articles and social media rather than structured data visualizations. While some research exists on chart fact-checking, the distortion of narratives through visual evidence remains underexplored.

Instead of focusing on deceptive chart designs, we concentrate on the rhetorical distortion that happens when data is valid. The manipulation lies in how conclusions are framed, and not in the visual representation itself. To clarify our analysis, we distinguish three categories: **Inference** (unsupported causal claims), **Sensationalism** (emotional or exaggerated framing), and **Oversimplification** (selective reduction of complexity).

While prior work on chart understanding emphasizes data extraction and question answering, it overlooks deceptive interpretation detection. Existing multimodal misinformation studies mainly address inconsistencies between images and captions in news content, while most NLP-based persuasion detection remains unimodal, focusing solely on text. None of these approaches handles structured visual data involving narrative manipulation.

To address this gap, we introduce the PlotTwist DataSet (PTDS), a benchmark of 300 chart–text samples for probing rhetorical manipulation in narratives that are grounded in factually accurate charts. The dataset consists of four classes: three manipulation classes (unsupported causal inference, sensationalism, and oversimplification) and a neutral class. Each sample has been validated through expert review to confirm its integrity. We evaluate state-of-the-art chart understanding methods, such as MatCha and Vision Transformer-based models on PTDS. We developed a two-stage pipeline tightly coupling a visual and text-based model: MatCha which converts images of charts into structured textual descriptions, and a BERT classifier for manipulation detection based on the

resulting caption, narrative statement, and generated chart evidence. For comparison, we also implement a two-stream ViT+BERT baseline with cross-attention fusion, which allows direct multimodal learning without intermediate derendering. In each of these methods, we perform a 20×10 repeated stratified cross-validation protocol.

2 Related Work

Previous works, such as ChartQA (Masry et al., 2022) and PlotQA (Methani et al., 2020), focus on question answering and numerical reasoning over charts, while ChartCheck (Akhtar et al., 2024) evaluates only the factual consistency between a chart and its caption, not the taxonomy of rhetorical manipulation. Pix2Struct (Lee et al., 2023) and MatCha (Liu et al., 2023) improve the potential for conversion of data from charts into text. In contrast, we focus on narrative manipulation of visual evidence that is factually true in itself.

Beyond question answering, previous work includes chart summarization and data-to-text generation, demonstrating the need to accurately distill and report structured visual content (Obeid and Hoque, 2020). Related research on visualization literacy and visual rhetoric delves deeper into how users make sense of charts and how design choices affect their understanding. But these studies primarily evaluate correctness, clarity, and informativeness and often ignore the rhetorical manipulation of previously valid evidence (Choi et al., 2025; Liu et al., 2024). Our work extends these foundational studies by focusing on how structurally faithful readings of charts can be strategically reframed to produce misleading takeaways at the level of discourse.

Most previous work in multimodal misinformation detection focused on image-text consistency and on fake news datasets. These methods primarily analyse natural images with informal captions to detect semantic incongruence or misleading presentation (Lv et al., 2025; Shen et al., 2024; Tao et al., 2025; Xue et al., 2021). Our setting involves structured data visualizations, where manipulation does not originate from contradictions between images and text but rather sophisticated rhetorical distortions of legitimate evidence. This requires chart-specific reasoning rather than relying on general visual-semantic alignment.

Robust methodologies have been developed for modeling persuasive signals and assessing textual credibility in areas such as framing detection, emotional rhetoric, propaganda identification, and claim verification (Srba et al., 2026; Barkhatova, 2025; Bassi et al., 2024). These methods highlight linguistic cues that signal overstatement, partisanship, and unwarranted conclusions. These systems largely operate on text-only inputs and cannot assess whether rhetorical claims align with structured visual evidence. Our work further develops this field by grounding manipulation detection in data from charts.

Most multimodal classification approaches use early or late fusion strategies, cross-attention mechanisms, and joint vision–language encoders. Recent methods have begun using generative intermediate representations that map visual inputs to textual forms before entering downstream reasoning (Li and Tang, 2026; Hangloo and Arora, 2025). In contrast to traditional fusion approaches, we explicitly derender chart images to structured textual descriptions before classification. This methodology allows for controlled ablation of modality and robust cross-validation while still providing access to evidence based on the charts themselves.

More recently, multimodal large language models combine vision and language in a shared framework, yielding strong performance on a number of visual-text reasoning tasks (Wu et al., 2024; Wang et al., 2024). However, their ability to reason about more structured types of data visualizations—particularly to find subtle rhetorical distortions closely tied to the empirical evidence embedded in factually correct charts—has not been extensively studied. Our controlled derendering-based approach addresses this research gap.

3 PTDS: PlotTwist DataSet

PTDS takes a deep look at 3 classes of data manipulation, amplified by textual commentary. The definitions of each class are as follows:

Causal fallacies in visual narratives (*inference*): Misleading visuals combined with explanatory text can encourage causal misconceptions. Causality can be inferred by viewers from annotated charts even in the absence of justification (Vasconcelos Braga and Barros Pontes E Silva, 2021), and omitted confounders can obscure relationships (Lo et al., 2022) (e.g., diet soda and obesity without lifestyle factors). This can be described as a credibility signal: misattributing correlation as causation (Leite et al., 2025).

Persuasive and emotional rhetoric (*sensationalism*): Narratives may use emotional priming, exaggeration, or dramatization framing to bias interpretation. Usage of loss-framed language (e.g., *not recycling harms others*) increases perceptions of causal relations (Huang et al., 2019). *Sensationalism* can be identified as a core credibility signal (Srba et al., 2026; Leite et al., 2025). These are manipulations that rely on affect rather than evidence-based thinking, leading users to flawed causal conclusions.

Narrative bias reinforcement (*oversimplification*): Beyond emotional appeals, narratives may overgeneralize or selectively frame data to confirm preexisting biases. Design features can reduce scrutiny of text-visual alignment (Hullman and Diakopoulos, 2011). Vague and overgeneralized claims (like political narratives over economic trends) deepen user misinterpretation (Srba et al., 2026). This class captures rhetorical strategies that oversimplify complexity, reinforcing bias rather than enabling critical evaluation.

PTDS is constructed from paired samples, and each pair contains an image chart and its associated description labeled as deceptive (with class name specified) or non-deceptive. As a base of PTDS, a ChartCheck (Akhtar et al., 2024) dataset was chosen. ChartCheck is a carefully curated benchmark that showcases expert annotations of false captions for data visualizations in JSON format. ChartCheck's charts were acquired from Wikimedia Commons, and professional annotators assessed each chart description, labeling it as *true* or *false* and providing brief justification comments. Consistent labeling was achieved by using refined criteria and a stringent adjudication algorithm. Transforming it to fit PTDS requirements required two stages: generating deceptive statements for 150 ChartCheck charts (50 for each of the three classes) and creating 150 samples with neutral statements.

Stage 1: Of the 8544 ChartCheck records, entries with `label = TRUE` were removed, after which duplicate records (through `chart_img` field) were randomly discarded until unique records remained, yielding 1501 records. 150 records with readily interpretable data were selected manually and divided into three equal groups corresponding to the manipulation types (*Inference*, *Sensationalism*, *Oversimplification*), creating a new field `manip_class`. Using an original OpenAI-based (GPT-4.1 mini) app, a deceptive statement for each chart was generated, leveraging the `chart_img`, `caption`, and `claim` fields. Deceptive statements were generated according to the following process:

1. The chart image and caption are loaded to the GUI with 3 deceptive statements generated by *the deceptive statement creation prompt*[1].
2. A single statement is manually selected and stored in the record under a new `story` field.
3. Data from the next record is loaded, and the process is repeated until the last record.

The final step was to add a new field named `statement_reframed`, initialized to 0, and retain only `chart_img` and `caption` fields from the original ChartCheck.

Table 1: Results of each round carried out by experts. Classes: **I**: Inference, **S**: Sensationalism, **O**: Oversimplification.

Round	Ratio of Acceptance (Pass/Total)	Accepted Samples Scores	Rejected Samples Scores	Acceptance by Class	Fleiss' Kappa
I	132/150	5/5: 49 4/5: 45 3/5: 38	2/5: 3 1/5: 13 0/5: 2	I: 47/50 S: 50/50 O: 35/50	0.46
II	16/18	5/5: 8 4/5: 2 3/5: 6	2/5: 1 1/5: 0 0/5: 1	I: 3/3 S: – O: 13/15	0.29
III	2/2	5/5: 1 4/5: 1 3/5: 0	–	I: – S: – O: 2/2	–

To verify alignment between the generated statements and the defined manipulation classes, five experts reviewed all 150 samples each (with the help of the provided *Codebook*[1]), evaluating every claim–chart–statement set. After the analysis, they were required to choose one of 4 options: *Inference*, *Sensationalism*, *Oversimplification*, or *None of the above* for each sample. Samples that received at least 3 of 5 coincidences with the original intention expert judgments were included in the final dataset. Those scoring 2 of 5 or less were rejected and subjected to a statement-reframing procedure, in which a new statement was generated using the original method. If the sample was accepted in the first round, the `statement_reframed` field was left with a default value of 0, and was incremented by 1 every time the sample was not accepted in a given round. For all samples to be accepted, 3 rounds of validation were carried out; the results of each are shown in Table 1.

Stage 2: After constructing the deceptive subset, focus shifted back to the original `label = TRUE`

[1] https://osf.io/s5gjt/overview?view_only=c139e08cfee84ad1a84764b549c1aab0

records from ChartCheck. The samples that corresponded to the same ones used to generate deceptive statements were identified. As samples with this label value retained the `claim` fields deemed as true by the ChartCheck authors, the creation of the `story` (containing statement) field was based on just renaming the `caption` the mentioned `story` field. Outside of that, only the `chart_img` field was left (by removing the rest of the fields), and a new `statement_reframed` field initialized to -1 was introduced, and `manip_class` was added and set to *Neutral*. This resulted in the dataset having a consistent JSON structure across its entire contents, regardless of the type defined in the `manip_class` field.

The final PTDS comprises 300 records: 50 *Inference*, 50 *Sensationalism*, 50 *Oversimplification*, and 150 *Neutral* cases. Although the PTDS is modest in scale (300 samples), it addresses a previously unexplored area, valuing conceptual novelty and a rigorously defined taxonomy over quantity. Most existing multimodal misinformation benchmarks focus on natural images with informal captions. This is the first dataset we know of that focuses on rhetorical manipulation in relation to factually accurate structured data visualizations. It is publicly available[2] for transparency purposes and to enable other researchers to compare their models, extend the dataset, or employ it in educational and methodological studies. PTDS is intended for tasks such as detecting misleading chart narratives, analyzing credibility and reasoning, AI-assisted data literacy and education, and benchmarking LLM-based interpretative reasoning systems.

4 Algorithms and Models

Each sample i comprises an image I_i, a caption c_i, and a short (deceptive or neutral) statement s_i. The task is cast as supervised classification. Formally, we seek $f_\theta(I_i, c_i, s_i) \rightarrow y_i \in \{0, \ldots, M-1\}$, where y_i is the manipulation label.

Our pipeline follows a two-stage scheme:

1. **Chart description generation (Matcha)**: We employ the Pix2Struct-based MatCha model (Liu et al., 2022) to convert the chart image into a natural-language description. Pix2Struct is an encoder-decoder that renders the image (here, the chart) into text; MatCha is further adapted for chart reasoning and table-like outputs. We use a concise header prompt ("*Generate underlying data table of the figure below:*") and decode with beam search (num_beams=4, max_new_tokens=512, early stopping).

2. **Text classifier (BERT)**: We concatenate caption, statement, and the MatCha description, then fine-tune a BERT-based sequence classifier (Devlin et al., 2018). BERT is a bidirectional Transformer encoder pretrained via masked language modeling, providing strong contextual representations for sentence-to-paragraph-length text. Input is tokenized with WordPiece (uncased), truncated/padded to 512 tokens.

To isolate modality contributions for ablation studies, we implemented a parallel two-stream classifier: the Vision Transformer (ViT) (Dosovitskiy et al., 2020) encodes the chart image, BERT (Devlin et al., 2018) encodes the caption and statement, and cross-attention fuses the representations for classification. This end-to-end multimodal alternative avoids intermediate text generation and enables explicit ablations (image-only, text-only, joint) to assess the relative importance of visual versus linguistic cues in manipulation detection.

The dataset is imbalanced (Neutral $\approx 50\%$). We compute inverse-frequency class weights and apply them in the cross-entropy loss to penalize majority-class dominance. This curbs collapse into Neutral predictions while preserving per-class recall.

5 Results

In this section, we empirically evaluate the proposed manipulation detection framework on the PTDS benchmark, providing an overview of classification performance and clarifying the relative contributions of the visual and textual modalities.

5.1 Baseline

We evaluate the baseline model on 20×10 stratified folds (200 total). Table 2 reports macro-averaged performance (mean $\pm$ std across folds). Figure 1 provides the row-normalized confusion matrix and thus highlights per-class recall patterns, while Figure 2 shows fold-wise trajectories of accuracy/precision/recall/F1, which allows for inspecting stability across random splits.

[2] https://huggingface.co/datasets/WieszczUrsynowa/PTDS-PlotTwistDataSet

Table 2: Cross-validation performance (macro, mean ± std over 200 folds).

Metric	Value (Mean ± Std)
Accuracy	0.901 ± 0.061
Precision	0.902 ± 0.073
Recall	0.895 ± 0.072
F1 Score	0.889 ± 0.074

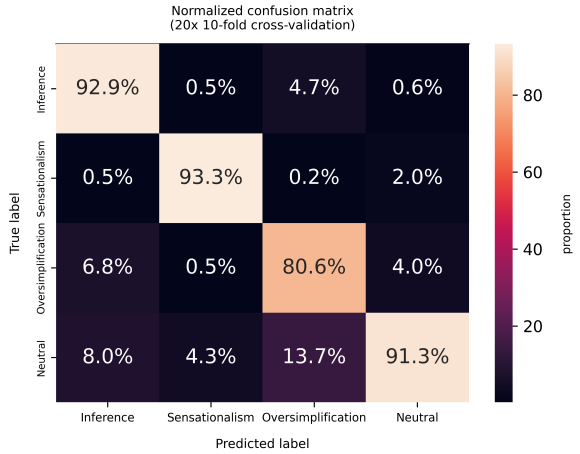

Figure 1: Normalized confusion matrix aggregated over 200 folds.

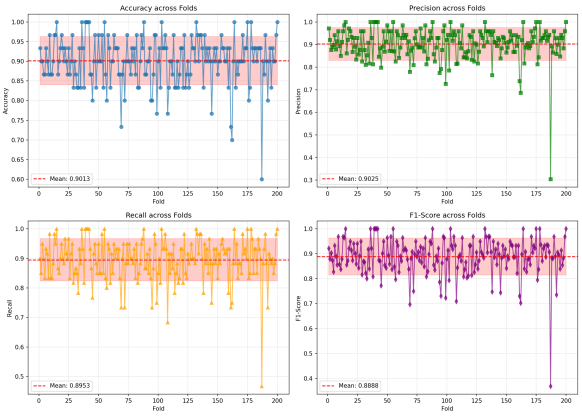

Figure 2: Metric trajectories across folds.

The Matcha→BERT pipeline demonstrates robust performance, yielding a mean accuracy of 0.901 ± 0.009 and macro-F1 of 0.889 ± 0.010 (95% CI) over 200 folds. While the central tendency is highly concentrated (Median F1 ≈ 0.893, IQR $[0.855, 0.933]$), the repeated protocol exposes a non-negligible long tail (0.368–1.00). This variance is primarily driven by minority class sparsity; in folds with limited samples, the misclassification of ambiguous chart-text pairs disproportionately impacts macro-averaged metrics.

5.2 ViT+BERT Baseline and Modality Ablations

In addition to the repeated cross-validation of Matcha→BERT, we report a two-stream ViT+BERT baseline evaluated on a single held-out split. Table 3 compares the joint model against modality ablations (text-only and image-only). To further illustrate the behavior of the joint two-stream model, Figure 3 reports the training loss curve, validation accuracy over epochs, the test-set confusion matrix, and the corresponding summary metrics.

Table 3: ViT+BERT two-stream baseline on a held-out split (macro metrics).

Setting	Accuracy	Precision	Recall	F1
ViT+BERT (cross-attention)	0.867	0.874	0.867	0.869
Text-only (BERT)	0.550	0.723	0.550	0.550
Image-only (ViT)	0.300	0.302	0.300	0.288

5.3 Component and Architectural Ablation Studies

To quantify the contribution of each design choice in the Matcha→BERT pipeline, we ran 13 extended ablations (9 component-level and 4 architectural-level), each evaluated with 5-fold cross-validation repeated twice (10 folds total per ablation). The ablations cover: (i) class weighting removal, (ii) prompt template variation, (iii) single-modality inputs (caption/statement/description), (iv) late fusion vs. early fusion, (v) encoder replacement (RoBERTa, DeBERTa), (vi) describer replacement (generic captioning), and (vii) frozen vs. unfrozen Matcha.

Table 4 reports the ablation metrics directly from the generated JSON report (mean±std over 10 folds). The 13 ablations fall into two groups. **Component ablations** hold the architecture fixed (BERT classifier, frozen Matcha, early concatenation fusion) and vary one pipeline decision at a time: (i) *no_weighting* removes inverse-frequency class weights from the cross-entropy loss; (ii) *prompt_template_1–4* substitutes the default Matcha prompt (*"Generate underlying data table of the figure below:"*) with three alternative phrasings (*"What data does this chart show?"*, *"Summarize the chart data in a table format:"*, *"Extract the chart data:"*); (iii) *modality_caption_only*, *modality_story_only*, and *modality_description_only* each feed a single text field to the classifier instead of the full concatenation; and (iv) *fusion_late* replaces early concatenation with

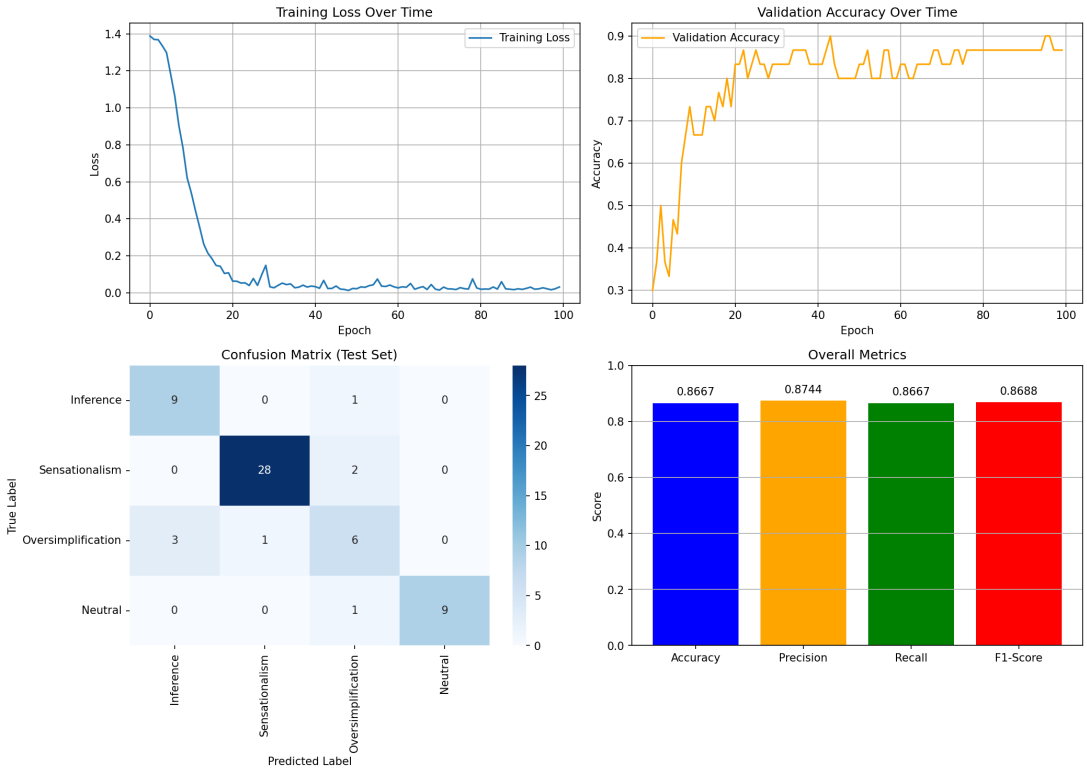

Figure 3: Two-stream ViT+BERT (cross-attention) training dynamics and test performance: training loss, validation accuracy, confusion matrix, and overall metrics on the held-out split.

a strategy that trains separate per-modality BERT classifiers and merges their logits via a learned fusion layer. **Architectural ablations** keep all component decisions at their defaults and swap a major model component: (v) *classifier_roberta-base* and *classifier_deberta-base* replace the BERT-base classifier encoder; (vi) *describer_generic* replaces Matcha with a domain-agnostic image captioner lacking chart-specific training; and (vii) *matcha_unfrozen* allows end-to-end gradient flow through the Matcha encoder during classifier training.

Three trends stand out. First, fusion design is critical: late fusion drops to 0.565 accuracy and 0.420 macro-F1, indicating the simple early-fusion setup is much more stable in this dataset. Second, modality-only runs reveal strong asymmetry: statement-only performs best (0.975/0.965), while caption-only (0.395/0.174) and description-only (0.383/0.152) collapse; this suggests highly predictive lexical patterns in statement text and a comparatively weak standalone signal in the other isolated streams. Third, architecture swaps provide moderate gains (RoBERTa 0.933/0.926, DeBERTa 0.902/0.897), while replacing Matcha with a generic describer degrades performance (0.830/0.806), supporting the value of chart-specialized description generation.

Table 4: Extended ablation results (mean±std over 10 folds).

Ablation	Accuracy ($\mu \pm \sigma$)	Macro-F1 ($\mu \pm \sigma$)
component_modality_statement_only	0.975 ± 0.023	0.965 ± 0.035
architectural_classifier_roberta-base	0.933 ± 0.037	0.926 ± 0.043
architectural_classifier_deberta-base	0.902 ± 0.072	0.897 ± 0.083
component_no_weighting	0.870 ± 0.077	0.849 ± 0.097
architectural_describer_generic	0.830 ± 0.067	0.806 ± 0.076
architectural_matcha_unfrozen	0.828 ± 0.105	0.822 ± 0.095
component_prompt_template_3	0.828 ± 0.104	0.794 ± 0.126
component_prompt_template_1	0.785 ± 0.142	0.766 ± 0.157
component_prompt_template_2	0.777 ± 0.151	0.745 ± 0.221
component_prompt_template_4	0.755 ± 0.164	0.691 ± 0.250
component_fusion_late	0.565 ± 0.190	0.420 ± 0.183
component_modality_caption_only	0.395 ± 0.088	0.174 ± 0.022
component_modality_description_only	0.383 ± 0.092	0.152 ± 0.031

5.4 Failure Pattern Analysis

To characterize failure modes and separate reliable visual evidence from misleading textual interpretations, we evaluated 300 ordered visual samples using the trained Matcha→BERT classifier. Misclassifications were analyzed by chart type. Only 6 of 300 samples were misclassified (2.0% error rate), indicating that failures are rare and structured rather than random.

Figure 4 reports the error rate by chart type. The largest concentration is in heterogeneous "other" charts, while bar/pie categories contribute fewer

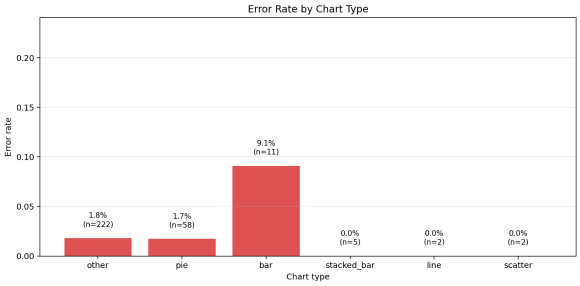

Figure 4: Error rate by chart type from failure-pattern analysis (300 samples). Bars are annotated with per-type sample counts.

absolute failures. This suggests that non-canonical layouts and mixed-format visuals remain more challenging than standard chart families.

Table 5 summarizes the most common error clusters with representative examples from the generated taxonomy table. The dominant cluster is long-text, high-ambiguity "*other*" charts (2 cases), followed by a bar-chart cluster in which Matcha descriptions lack axis information. This directly supports the view of description quality as an actionable failure driver.

The focused confusion analysis for Inference↔Oversimplification found 2 pair-confusions total (one in each direction), concentrated in "*other*" and pie charts. Context words in these errors (e.g., *prices*, *soybean*, *military*, *deaths*) indicate semantically causal language in borderline narratives, consistent with overlap between causal inference framing and reductive oversimplification statements.

For publication impact, this analysis yields (i) an interpretable error-rate-by graph type figure (Figure 4) and (ii) an explicit error-taxonomy table (Table 5) that can be directly linked to model-improvement priorities.

6 Discussion

Modality Synergy Ablation studies (Table 4) indicate that manipulation cues are predominantly *lexical-rhetorical*. The text-only stream significantly outperforms the image-only baseline, suggesting that hedging or unsupported claims are stronger signals than visual-structural cues alone. However, the performance gain from ViT+BERT fusion confirms that chart evidence, even when noisy, provides complementary information. As shown in Table 4, text-only modalities perform substantially worse: models relying solely on captions or generated descriptions achieve much lower scores ($0.3950.088$ and $0.3830.092$, respectively) compared to the full multimodal configuration. This gap indicates that textual representations alone fail to capture key structural properties of the charts, whereas incorporating visual features allows the model to resolve ambiguities and improve overall robustness.

Fairness and Deployment While global accuracy remains high, fairness analysis (Table 6) reveals topic-dependent fluctuations. The *Oversimplification* category exhibited the highest demographic-parity disparity, suggesting a need for strata-specific tuning. Furthermore, the pipeline is resilient to minor linguistic perturbations but remains sensitive to chart corruption.

6.1 Limitations

Despite strong empirical results, several bottlenecks remain:

- **Error Propagation:** As shown in Table 8, the two-stage design is highly sensitive to Stage-1 fidelity. A 5% error rate in Matcha-generated descriptions triggers a 0.213 absolute drop in accuracy, highlighting the risk of hallucinations or omissions (e.g., axis scales).
- **Context Truncation:** Concatenating caption, statement, and description into a 512-token sequence may discard crucial disclaimers. The current architecture struggles with long-context narratives where manipulation cues are distant from numeric evidence.
- **Semantic Adjacency:** The manipulation taxonomy contains overlapping categories (e.g., rhetorical emphasis vs. unsupported inference). Label noise at these boundaries can inflate confusion and limit the ceiling of macro-F1 performance.
- **Visual Grounding:** Our analysis suggests that the model occasionally relies on textual shortcuts. Future iterations require more explicit visual grounding to ensure chart evidence contributes directly when textual cues are ambiguous.

7 Conclusions and Future Work

Our study makes several contributions to the growing field of multimodal misinformation research by redirecting attention away from mere falsified visual elements or simple inconsistencies between images and text toward considerations of rhetorical

Table 5: Most common error patterns with examples (top clusters from taxonomy analysis).

Pattern (cluster)	Count	Error rate in cluster	Representative confusion	Key description issue
other + long + high ambiguity	2	0.286	Oversimplification→Neutral	none
bar + long + low ambiguity	1	0.250	Oversimplification→Neutral	missing axis info
pie + medium + low ambiguity	1	0.100	Inference→Oversimplification	none
other + short + high ambiguity	1	0.029	Neutral→Sensationalism	none
other + medium + low ambiguity	1	0.023	Oversimplification→Inference	none

manipulation in chart narratives. We present the PlotTwist DataSet (PTDS), a benchmark of 300 human-validated examples that specifically capture three major types of narrative distortion: unsupported causal inference, sensationalism, and oversimplification. Crucially, the actual chart data embedded inside PTDS remains factually and structurally correct. PTDS provides a controlled setting for discourse-level distortions by separating truthful visual evidence from false textual interpretations, based on ordered visual data.

We designed a two-stage multimodal pipeline consisting of a chart-to-text derendering model followed by text-based reasoning. MatCha is first used for generating structured representations of charts. In the next phase, we use a BERT-based classifier to accost the caption, narrative statement, and generated description. We show effective evaluation using a 20×10 repeated stratified cross-validation protocol yielding strong performance with a macro-F1 score of 0.889. Our results suggest that compared to a two-stream ViT+BERT multimodal baseline, converting a chart image into textual representations can support effective reasoning about narrative manipulation.

In future work, the coverage and robustness of PTDS could be expanded by including new chart types and deceptive narrative domains, as well as varying our methods for creating narratives. Architecturally, integrating stronger chart-understanding models or multimodal LLMs to perform reasoning on visual and textual input. Another promising direction for development is to create models that directly associate narrative claims with particular chart elements, including axes, trends, and data points.

By introducing PTDS and demonstrating a practical multimodal framework that can be applied to the detection of manipulative chart narratives, we hope that this work lays the groundwork for future studies of discourse-level misinformation in structured data visualizations.

References

Mubashara Akhtar, Nikesh Subedi, Vivek Gupta, Sahar Tahmasebi, Oana Cocarascu, and Elena Simperl. 2024. ChartCheck: Explainable Fact-Checking over Real-World Chart Images. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 13921–13937, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Elvira Barkhatova. 2025. Media manipulation detection: Challenges and perspectives. *Journal of Computer-Assisted Linguistic Research*, 9:19–42.

Davide Bassi, Søren Fomsgaard, and Martín Pereira-Fariña. 2024. Decoding persuasion: a survey on ML and NLP methods for the study of online persuasion. *Frontiers in Communication*, 9:1457433.

Siming Chen, Jie Li, Gennady Andrienko, Natalia Andrienko, Yun Wang, Phong H. Nguyen, and Cagatay Turkay. 2020. Supporting Story Synthesis: Bridging the Gap between Visual Analytics and Storytelling. *IEEE Transactions on Visualization and Computer Graphics*, 26(7):2499–2516.

Raymond Choi, Frank Burns, and Chase Lawrence. 2025. End-to-End Chart Summarization via Visual Chain-of-Thought in Vision-Language Models. *arXiv preprint*. Version Number: 1.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint*. Version Number: 2.

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. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv preprint*. Version Number: 2.

Sakshini Hangloo and Bhavna Arora. 2025. Multimodal fusion techniques: Review, data representation, information fusion, and application areas. *Neurocomputing*, 649:130827.

Guanxiong Huang, Kang Li, and Hairong Li. 2019. Show, Not Tell: The Contingency Role of Infographics Versus Text in the Differential Effects of Message Strategies on Optimistic Bias. *Science Communication*, 41(6):732–760.

J. Hullman and N. Diakopoulos. 2011. Visualization Rhetoric: Framing Effects in Narrative Visualization. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2231–2240.

Kenton Lee, Mandar Joshi, Iulia Turc, Hexiang Hu, Fangyu Liu, Julian Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. 2023. Pix2Struct: Screenshot Parsing as Pretraining for Visual Language Understanding. *arXiv preprint*. ArXiv:2210.03347 [cs].

João A. Leite, Olesya Razuvayevskaya, Kalina Bontcheva, and Carolina Scarton. 2025. Weakly supervised veracity classification with LLM-predicted credibility signals. *EPJ Data Science*, 14(1):16.

Songtao Li and Hao Tang. 2026. Multimodal Alignment and Fusion: A Survey. *International Journal of Computer Vision*, 134(3):103.

Fangyu Liu, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Yasemin Altun, Nigel Collier, and Julian Eisenschlos. 2023. MatCha: Enhancing Visual Language Pretraining with Math Reasoning and Chart Derendering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12756–12770, Toronto, Canada. Association for Computational Linguistics.

Fangyu Liu, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Yasemin Altun, Nigel Collier, and Julian Martin Eisenschlos. 2022. MatCha: Enhancing Visual Language Pretraining with Math Reasoning and Chart Derendering. *arXiv preprint*. Version Number: 2.

Mengsha Liu, Daoyuan Chen, Yaliang Li, Guian Fang, and Ying Shen. 2024. ChartThinker: A Contextual Chain-of-Thought Approach to Optimized Chart Summarization. *arXiv preprint*. ArXiv:2403.11236 [cs].

Leo Yu-Ho Lo, Ayush Gupta, Kento Shigyo, Aoyu Wu, Enrico Bertini, and Huamin Qu. 2022. Misinformed by Visualization: What Do We Learn From Misinformative Visualizations? *Computer Graphics Forum*, 41(3):515–525.

Jinna Lv, Yuan Gao, Li Li, Lei Shi, and Siyu Li. 2025. Multi-modal fake news detection: A comprehensive survey on deep learning technology, advances, and challenges. *Journal of King Saud University Computer and Information Sciences*, 37(9):306.

Ahmed Masry, Do Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. ChartQA: A Benchmark for Question Answering about Charts with Visual and Logical Reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland. Association for Computational Linguistics.

Nitesh Methani, Pritha Ganguly, Mitesh M. Khapra, and Pratyush Kumar. 2020. PlotQA: Reasoning over Scientific Plots. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1516–1525, Snowmass Village, CO, USA. IEEE.

Jason Obeid and Enamul Hoque. 2020. Chart-to-Text: Generating Natural Language Descriptions for Charts by Adapting the Transformer Model. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 138–147, Dublin, Ireland. Association for Computational Linguistics.

Xiaorong Shen, Maowei Huang, Zheng Hu, Shimin Cai, and Tao Zhou. 2024. Multimodal Fake News Detection with Contrastive Learning and Optimal Transport. *Frontiers in Computer Science*, 6:1473457.

Ivan Srba, Olesya Razuvayevskaya, João A. Leite, Robert Moro, Ipek Baris Schlicht, Sara Tonelli, Francisco Moreno García, Santiago Barrio Lottmann, Denis Teyssou, Valentin Porcellini, Carolina Scarton, Kalina Bontcheva, and Maria Bielikova. 2026. A Survey on Automatic Credibility Assessment Using Textual Credibility Signals in the Era of Large Language Models. *ACM Transactions on Intelligent Systems and Technology*, 17(2):1–80.

Chase Stokes, Cindy Xiong Bearfield, and Marti A. Hearst. 2024. The Role of Text in Visualizations: How Annotations Shape Perceptions of Bias and Influence Predictions. *IEEE Transactions on Visualization and Computer Graphics*, 30(10):6787–6800.

Zhulin Tao, Runze Zhao, Xin Shi, Xingyu Gao, Xi Wang, and Xianglin Huang. 2025. Multimodal Consistency Suppression Factor for Fake News Detection. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 21(11):1–19.

Jaqueline Vasconcelos Braga and Tiago Barros Pontes E Silva. 2021. Storytelling in data visualization: information bias. *InfoDesign - Revista Brasileira de Design da Informação*, 18(3).

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, Alexis Chevalier, Sanjeev Arora, and Danqi Chen. 2024. CharXiv: Charting Gaps in Realistic Chart Understanding in Multimodal LLMs. *arXiv preprint*. Version Number: 1.

Yifan Wu, Lutao Yan, Leixian Shen, Yunhai Wang, Nan Tang, and Yuyu Luo. 2024. ChartInsights: Evaluating Multimodal Large Language Models for Low-Level Chart Question Answering. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12174–12200, Miami, Florida, USA. Association for Computational Linguistics.

Junxiao Xue, Yabo Wang, Yichen Tian, Yafei Li, Lei Shi, and Lin Wei. 2021. Detecting fake news by exploring the consistency of multimodal data. *Information Processing & Management*, 58(5):102610.

A Additional Evaluation Studies

A.1 Fairness Analysis (NLP-specific)

We conducted an NLP-focused fairness audit on all 300 samples using the trained Matcha→BERT classifier, with three views: (i) demographic-parity proxy across sensitive topics, (ii) linguistic bias across formal vs. informal language, and (iii) data-source bias across inferred institution types. The model achieved strong aggregate performance (accuracy 0.983, macro-F1 0.978), but subgroup-level variation remains evident.

Table 6 reports performance by topic category. The lowest macro-F1 appears in *health* (0.714, n=14) and *politics* (0.903, n=23), while *other*, *climate/energy*, and *race/ethnicity* reach 1.0 in this sample. For linguistic bias, formal text outperforms informal text (macro-F1 0.982 vs. 0.948, gap 0.035). For source bias, the weakest large-enough source group is *Wikipedia* (macro-F1 0.951, n=18), compared with 1.0 in *government_agency* (n=37).

Demographic parity (topic-proxy) further shows uneven predicted class rates by topic: the largest classwise max-min rate gap is for *Oversimplification* (0.262). While this does not use protected-attribute labels, it indicates that model outputs vary across content domains and discourse contexts.

Table 6: Fairness analysis: performance by topic category (proxy-sensitive groups).

Topic category	n	Accuracy	Precision (macro)	Recall (macro)	F1 (macro)
other	80	1.000	1.000	1.000	1.000
gender	67	0.985	0.992	0.975	0.983
climate_energy	60	1.000	1.000	1.000	1.000
economy	42	0.976	0.964	0.975	0.968
politics	23	0.913	0.917	0.917	0.903
health	14	0.929	0.750	0.688	0.714
race_ethnicity	12	1.000	1.000	1.000	1.000

A.2 Robustness Analysis

We evaluated robustness along four axes: adversarial text perturbations (typos, synonym replacements, paraphrase-style rewrites), chart corruptions (Gaussian noise, JPEG compression, occlusion), cross-domain transfer by source, and out-of-distribution (OOD) detection. For the randomly selected fold on clean data, the model achieves an accuracy of 0.983 and a macro-F1 of 0.978 (n=300).

Table 7 summarizes representative stress conditions. The strongest text degradation is under severe typos (level 0.8), with an absolute accuracy drop of 0.240; synonym and paraphrase perturbations remain comparatively stable in this setup. For chart corruptions, classifier accuracy on the corruption subset (n=24) is less monotonic, but Matcha description quality degrades substantially, measured by token-overlap (Jaccard) against clean descriptions (e.g., occlusion level 0.3: mean Jaccard 0.197).

Table 7: Accuracy under representative perturbations. ΔAcc is defined as clean minus perturbed accuracy (same evaluation pool). For chart corruption rows, n=24 and the final column reports the mean token-Jaccard between clean and corrupted Matcha descriptions.

Condition	Level	n	Accuracy	ΔAcc	Desc. Jaccard
Clean baseline	–	300	0.983	0.000	–
Typos	0.8	300	0.743	0.240	–
Synonym replacement	0.8	300	0.983	0.000	–
Paraphrase rewrite	0.8	300	0.990	-0.007	–
Gaussian noise	0.3	24	1.000	-0.083	0.297
JPEG compression	0.3	24	0.917	0.000	0.621
Occlusion patch	0.3	24	0.958	-0.042	0.197

Cross-domain evaluation shows source-dependent behavior (e.g., *wikipedia* accuracy 0.944, *other* 0.986), while very small groups (e.g., *pew*, n=4) should be interpreted cautiously. OOD detection remains limited in this configuration (AUROC $\approx$ 0.55 for mismatched-statement OOD and $\approx$ 0.62 for gibberish OOD under MSP/energy scoring).

A.3 Error Propagation Analysis (Two-Stage Pipeline)

To isolate how errors propagate through the Matcha→BERT stack, we conducted a dedicated stage-wise stress test with synthetic perturbations at controlled rates (0%, 5%, 10%, 20%, 30%). We report three conditions: (i) Stage-1 only (description corruption), (ii) Stage-2 only (caption/statement corruption), and (iii) both stages corrupted. We also computed a proxy BLEU score for description quality and measured its relationship with classification accuracy.

Because the released dataset does not contain explicit ground-truth chart tables, the oracle experiment is reported as an *oracle-proxy* setting (caption/statement-derived proxy descriptions). Thus, the oracle numbers should be interpreted as a controlled sensitivity probe rather than a strict upper bound.

Table 8 summarizes representative operating points. The key finding is that a small Stage-1 perturbation can cause a large downstream shift: at 5% synthetic description error, accuracy drops from 0.983 to 0.770 (absolute drop 0.213). When both stages are perturbed, degradation is stronger

(e.g., 0.717 accuracy at 30% corruption).

Table 8: Error propagation in the two-stage pipeline. Description quality is measured by a BLEU proxy against clean Matcha descriptions.

Condition	Error rate	Accuracy	ΔAcc	Desc. BLEU (proxy)
Baseline (no corruption)	0.00	0.983	0.000	0.993
Stage-1 only (description)	0.05	0.770	0.213	0.724
Stage-1 only (description)	0.10	0.820	0.163	0.566
Stage-2 only (caption/statement)	0.30	0.867	0.117	0.993
Both stages	0.05	0.797	0.187	0.724
Both stages	0.30	0.717	0.267	0.251

B Explainability Analysis: Attention Heatmaps and Token Importance

To enhance interpretability and understand the linguistic signals driving the classifier's decisions, we conducted a comprehensive explainability analysis using two complementary techniques: attention-weight visualization and LIME-based token importance estimation. These analyses shed light on which tokens (words/subword units) are most influential in the classifier, and how these patterns differ across manipulation types. All results presented here are derived from 120 samples balanced across the four classes, with independent LIME explanations (250 perturbation samples per instance) and attention extraction from the final transformer layer.

B.1 Attention Weight Analysis

We extracted attention weights from the final layer of the BERT classifier for each sample and computed the average attention that the [CLS] token (which aggregates sentence-level information) paid to each token in the input sequence. These attention scores reflect the model's implicit focus distribution across the input during inference. Figure 5 and Figure 6 exemplify attention heatmaps for correctly classified samples from two manipulation types.

For Inference samples (Figure 5), the attention peaks on causal language such as "*caused*", "*drives*", and "*led*", which are characteristic of inference-based reasoning in the input text. For Sensationalism (Figure 6), high-attention tokens include emotionally-charged words like "*alarming*", "*threatening*", and "*devastating*", aligning with the emotional amplification typical of sensationalist framing.

B.2 LIME-based Token Importance by Manipulation Type

To systematically identify which tokens are most predictive of each manipulation class, we applied LIME (Local Interpretable Model-agnostic Explanations) to each sample. LIME perturbs the input by removing/adding words and observes the effect on prediction probabilities, thereby estimating per-token importance. By aggregating LIME weights across all samples of each class, we obtain a language-level signature of manipulative framing techniques. Table 9 summarizes the highest-weight tokens from the aggregated CSV statistics.

Key findings from the token importance analysis:

- **Inference** is strongly associated with causal and connective tokens: "causes" (avg weight 0.530), "leads" (0.519), "caused" (0.593), and "results" (0.335). This linguistic signature reflects the core mechanism of inference-based manipulation—drawing causal conclusions from correlative or incomplete evidence.

- **Sensationalism** is marked by emotionally-heightened language: "alarming" (0.340), "threatening" (0.306), "devastating" (0.303), "grim" (0.296), and "catastrophe" (0.241). These tokens amplify perceived threat or urgency, a defining characteristic of sensationalist framing.

- **Oversimplification** pivots on assertion and evidence verbs: "proves" (0.425), "explains" (0.406), "shows" (0.357), "confirms" (0.312), and "proving" (0.300). These tokens frame reductive explanations as definitive, collapsing complexity into singular cause-and-effect narratives.

- **Neutral** samples show near-zero or minimal weights for diagnostic tokens, consistent with the absence of manipulative linguistic markers. This reflects the classifier's learned ability to distinguish neutral factual reporting from the three manipulation types.

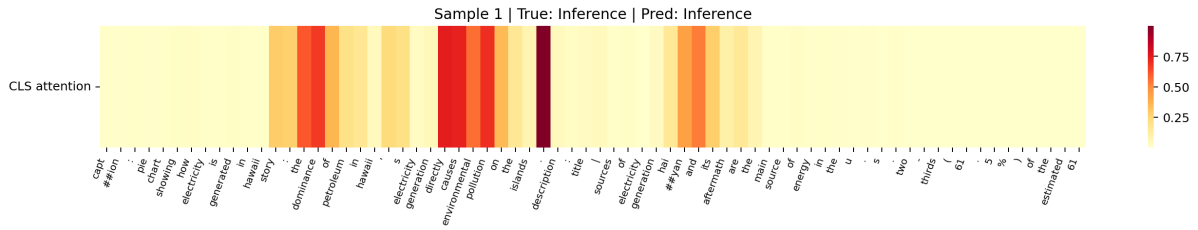

Figure 5: Attention heatmap for a correctly classified Inference sample. The [CLS] token shows elevated attention to causal markers ("caused", "drives") and connective language.

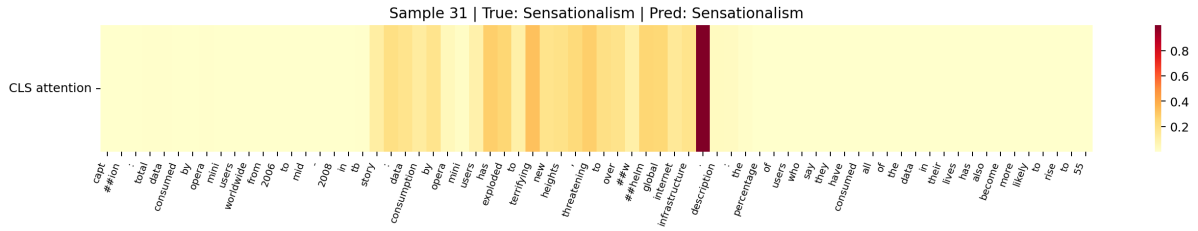

Figure 6: Attention heatmap for a correctly classified Sensationalism sample. High attention to emotionally evocative language ("threatening", "alarming", "devastating") is evident.

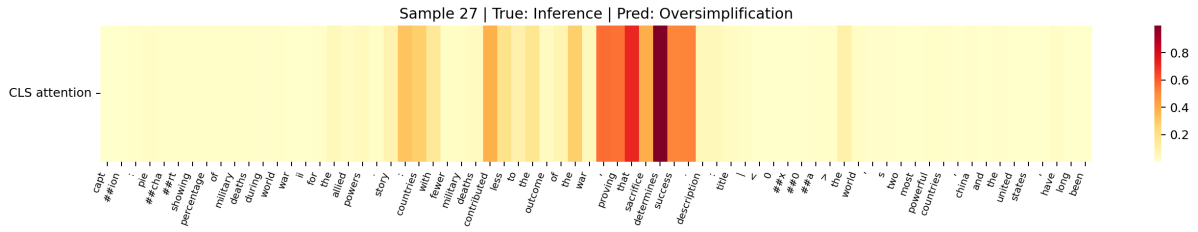

Figure 7: Attention heatmap for an axis-label miss error case. The sample was classified as Oversimplification (pred) when the true label was Inference. The failure is attributed to the Matcha description omitting axis scale information, hindering the classifier from recognizing causal framing in the caption/statement.

Table 9: Top LIME-weighted tokens by manipulation type (from aggregated token importance statistics).

Manipulation type	Top tokens with average weight
Inference	drove (0.702), caused (0.593), causes (0.530), leads (0.519), cause (0.433)
Sensationalism	staggering (0.355), ruining (0.347), alarming (0.340), threatens (0.306), devastating (0.303)
Oversimplification	showing (0.485), hinges (0.472), proves (0.425), explains (0.406), prove (0.388)
Neutral	tax (0.001), corporate (0.001), revenues (0.001), haven (0.001), the 2020 (0.001)

C Case Study: Axis Label Misses and Classifier Errors

One notable failure mode emerged during analysis: samples for which the Matcha-generated chart description *omits or misses axis labels, units, or scale information*, and the classifier subsequently errs. We identified 3 high-confidence error cases where (i) the predicted class differs from the true label by a large margin (> 0.70 in confidence difference), and (ii) the Matcha description lacks axis-related keywords (e.g., "x-axis", "y-axis", "scale", "units"). Figure 7 shows an example attention heatmap from one such case.

These axis-label misses highlight a key limitation of the two-stage Matcha→BERT approach: when the generative description stage (Matcha) fails to extract or articulate fine-grained visual details (e.g., axis scaling, numerical ranges), the downstream classifier loses critical evidence that could disambiguate subtle manipulations. For instance, a chart with exaggerated y-axis truncation—a common visual manipulation technique—could be missed if the description merely reports data values without noting the truncated

scale.

These explainability findings align with the quantitative results and provide actionable insights: (i) the model learns linguistically coherent patterns for each class; (ii) failure modes are interpretable and traceable to generative description quality; and (iii) future improvements should focus on robustifying Matcha prompts to elicit axis and scale metadata, or incorporating dedicated visual encoders that bypass the intermediate text stage for fine-grained spatial information.