TRUST-VL: An Explainable News Assistant for General Multimodal Misinformation Detection

Anonymous ACL submission

Abstract

Multimodal misinformation, encompassing textual, visual, and cross-modal distortions, poses an increasing societal threat that is amplified by generative AI. Existing methods typically focus on a single type of distortion and struggle to generalize to unseen scenarios. In this work, we observe that different distortion types share common reasoning capabilities while also requiring task-specific skills. We hypothesize that joint training across distortion types facilitates knowledge sharing and enhances the model's ability to generalize. To this end, we introduce TRUST-VL, a unified and explainable vision-language model for general multimodal misinformation detection. TRUST-VL incorporates a novel Question-Aware Visual Amplifier module, designed to extract task-specific visual features. To support training, we also construct TRUST-Instruct, a large-scale instruction dataset containing 198K samples featuring structured reasoning chains aligned with human fact-checking workflows. Extensive experiments on both in-domain and zero-shot benchmarks demonstrate that TRUST-VL achieves state-of-the-art performance, while also offering strong generalization and interpretability.

1 Introduction

Multimodal misinformation has become a fast-growing threat to society and has attracted wide attention in recent years. The rise of generative AI tools, while providing powerful capabilities for content creation, has also made it easier to produce misleading content and spread it at scale. For example, during the 2024 U.S. presidential election, foreign actors used AI-generated deepfakes and manipulated media to spread false narratives and influence voter perception, prompting official sanctions (Federspiel et al., 2023). Therefore, it is urgent to develop automated methods to detect multimodal misinformation (Akhtar et al., 2023; Chen and Shu, 2024; Abdali et al., 2025).

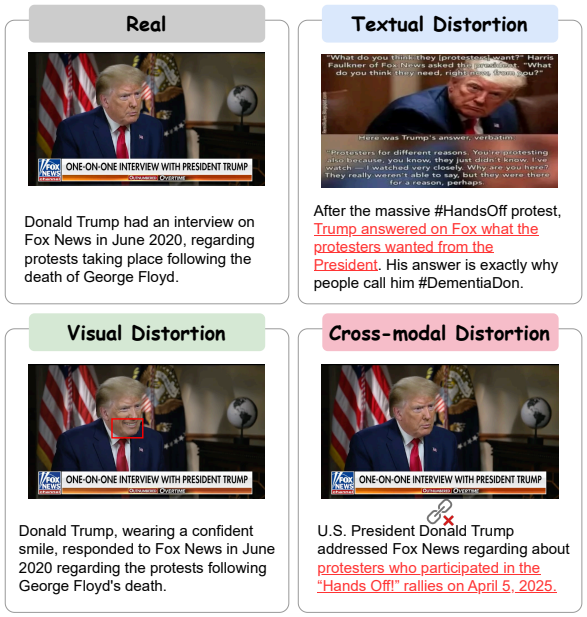

Figure 1: Examples of different distortion types in multimodal misinformation.

Multimodal misinformation is inherently a composite task, involving multiple sub-problems such as textual distortion, visual distortion, and cross-modal distortion. As illustrated in Figure 1, *textual distortion* refers to discrepancies between the textual claim and the underlying facts, which can often be identified through linguistic patterns or textual entailment between the claim and retrieved evidence. *Visual distortion* involves tampered or AI-generated images, and can be detected by identifying subtle visual artifacts or inconsistencies. *Cross-modal distortion* (also known as out-of-context misinformation) arises when the image and text originate from different real-world events, which can be detected by assessing semantic consistency across modalities (Alam et al., 2022; Liu et al., 2025).

Vision-language models (VLMs) have achieved impressive performance across a wide range of multimodal tasks (Liu et al., 2023; Dai et al., 2023; OpenAI, 2024a; Xue et al., 2024; Wang et al.,

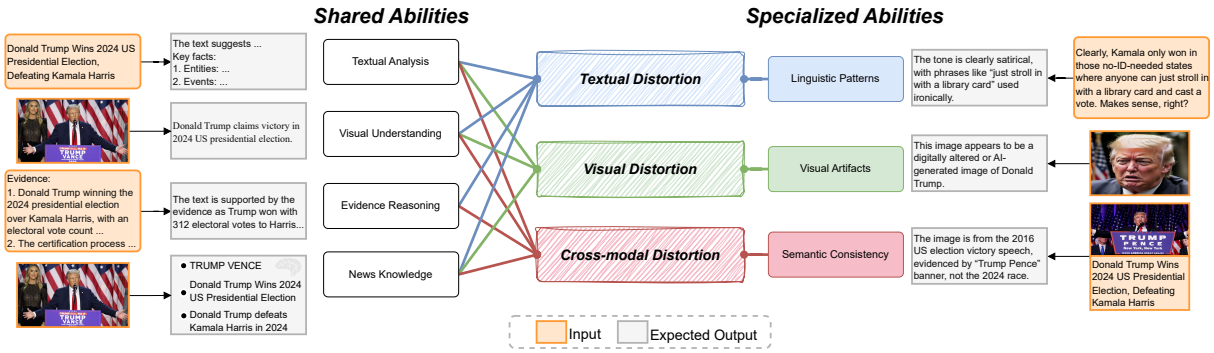

Figure 2: Overview of shared and specialized abilities involved across misinformation detection tasks.

2024). Motivated by this, prior works have applied VLMs to specific misinformation tasks such as fact checking (Yao et al., 2023; Tahmasebi et al., 2024), face manipulations (Liu et al., 2024b; Huang et al., 2024), and out-of-context detection (Qi et al., 2024). However, these models typically focus on a specific type of misinformation, and we empirically found that such single-task models often overfit and generalize poorly to unseen distortion types.

We observe that although detecting different distortion types requires *specialized reasoning* (e.g., linguistic pattern recognition, visual artifact detection, and semantic consistency checks), they also rely on *shared reasoning* (e.g., textual analysis, visual understanding, evidence-based reasoning, and familiarity with current news) (see Figure 2). For instance, multimodal content analysis is fundamental for in-depth reasoning, while evidence-based reasoning is crucial for tasks ranging from textual fact-checking to cross-modal inconsistency detection. Motivated by this, we we aim to build a unified framework that integrates both shared and specialized reasoning to effectively handle misinformation detection across diverse distortion types.

Developing a unified misinformation detection framework has several challenges: (1) Existing VLMs, pretrained on general vision-language tasks, often lack sensitivity to subtle visual artifacts and cross-modal semantic inconsistency; (2) annotation standards vary widely across existing datasets, complicating unified learning (Thorne et al., 2018; Suryavardan et al., 2023; Liu et al., 2024b; Luo et al., 2021a); and (3) most datasets lack explicit reasoning annotations, and provide only binary or categorical labels without detailing the intermediate reasoning steps behind the veracity judgment, thus limiting a model's ability to generate interpretable and persuasive explanations for real-world fact-checking applications (Thibault et al., 2024; Xu et al., 2023; Akhtar et al., 2023). These challenges highlight the need for new training paradigms with structured misinformation-specific reasoning annotations, along with comprehensive evaluation benchmarks to assess generalization across various misinformation tasks.

In this work, we observe that joint training across distortion types facilitates knowledge sharing and enhances the model's reasoning capabilities to generalize. Therefore, we propose TRUST-VL, a unified misinformation detection framework specially designed to enhance fine-grained visual understanding by conditioning perception on task-specific instructions, and is trained on a large-scale dataset of interleaved, reasoning-rich samples. Our main contributions can be summarized as follows:

• We propose TRUST-VL, a unified and explainable vision-language model for general multimodal misinformation detection. It integrates a novel Question-Aware Visual Amplifier (QAVA) module to extract task-specific visual features and support reasoning across misinformation detection tasks.

• We construct TRUST-Instruct, a large-scale instruction dataset of 198K samples with structured reasoning chains aligned with human fact-checking workflows, enabling effective joint training across diverse distortion types.

• Extensive experiments on both in-domain and zero-shot benchmarks demonstrate that TRUST-VL achieves state-of-the-art performance, with superior generalization and interpretability compared to existing detectors and general VLMs.

2 Related Work

Multimodal misinformation detection covers different sub-tasks that focus on different manipulation cues. Works on *textual distortion detection* use language models to fact check based on text only and often ignore the visual elements crucial for verifying many claims (Thorne et al., 2018; Augenstein et al., 2019; Kotonya and Toni, 2020;

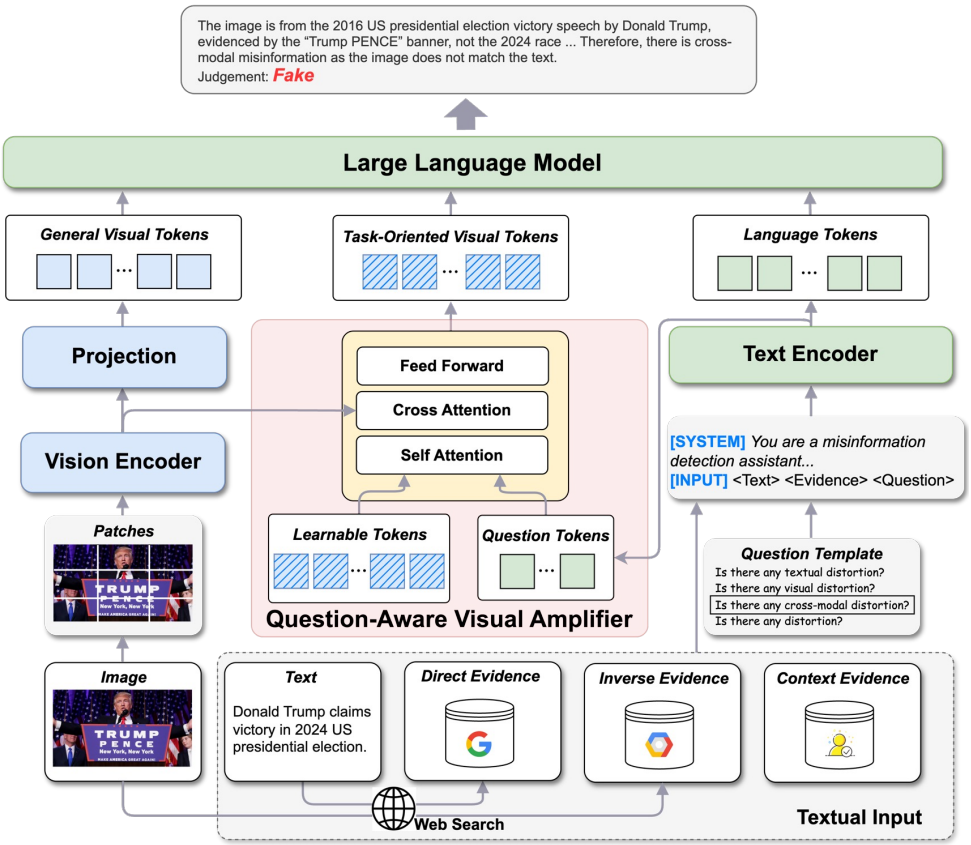

Figure 3: TRUST-VL Architecture. Given an image-text pair and associated evidence, TRUST-VL first encodes multimodal inputs through vision and text encoders. It then leverages the Question-Aware Visual Amplifier module, which uses a set of randomly initialized learnable tokens conditioned on task-oriented questions, to enhance visual perception. Finally, TRUST-VL outputs a structured and explainable detection judgment.

Pan et al., 2023). For *visual distortion detection*, recent efforts enhance VLMs with forgery-aware reasoning and visual artifact localization by soft prompt tuning (Liu et al., 2024b) and instruction tuning (Li et al., 2024b; Huang et al., 2024). For *cross-modal distortion detection*, (Tahmasebi et al., 2024; Qi et al., 2024; Xuan et al., 2024) enhance VLM reasoning by introducing external evidence sources. Notably, SNIFFER (Qi et al., 2024) improves image-text consistency detection through a two-stage instruction tuning process. However, these models are trained on narrowly scoped misinformation types such as face swaps or hallucinated claims, and struggle to generalize to unseen types.

Recent studies have started exploring complex scenarios in which false information spans across modalities. LRQ-FACT (Beigi et al., 2024) generates image- and text-focused questions using LLMs and VLMs, and synthesizes a final judgment through rule-based aggregation. (Liu et al., 2025) introduces MMD-Agent, a multi-agent framework that sequentially decomposes detection into textual, visual, and cross-modal subtasks, using step-wise prompting and retrieved evidence for improved reasoning. These multi-agent frameworks consist of loosely connected modules that are not jointly optimized for misinformation detection. In contrast, our proposed unified framework formulate misinformation tasks through a structured taxonomy of shared and specialized reasoning steps, and integrates them within a single VLM for end-to-end optimization and more effective detection.

3 Proposed Framework

Our goal is to develop an explainable VLM for detecting multimodal misinformation with various types of distortions. As illustrated in Figure 3, the proposed TRUST-VL framework takes an image-text pair as input and first retrieves relevant external evidence. The input text, evidence, and a task-specific question are encoded by a textual encoder, while the image is processed through a visual encoder equipped with a general projector and a question-aware visual amplifier. The resulting language and visual tokens are then jointly fed into a large language model (LLM) to produce a final judgment with an explanation.

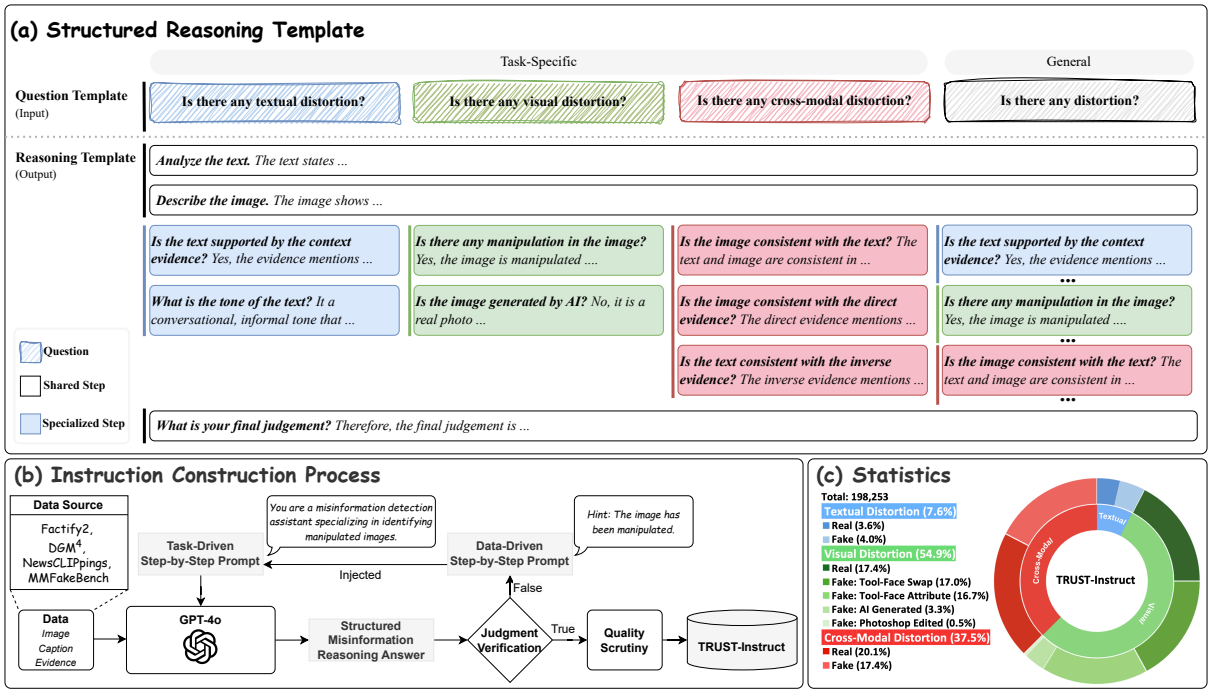

Figure 4: Overview of TRUST-Instruct. We present a pipeline for generating structured misinformation reasoning used for model training. TRUST-Instruct comprises 198K diverse samples spanning various distortion types, each annotated with rich, step-by-step reasoning chains.

3.1 Model Architecture

Model Input. Given a multimodal claim consisting of an image C_I and associated text C_T, TRUST-VL first retrieves external evidence from the open-domain web through a cross-modal retrieval (Abdelnabi et al., 2022). Specifically, we retrieve the top-m most relevant direct evidence ($E^{dir}_{1:m}$) using an image retriever guided by C_T, which is converted into captions via image-to-text generation. In parallel, we retrieve the top-n most relevant inverse evidence ($E^{inv}_{1:n}$) using a text retriever queried by C_I. Additionally, TRUST-VL incorporates context evidence ($E^{ctx}_{1:k}$), such as Wikipedia articles or expert annotations, provided either by users or down-stream benchmarks.

Base VLM. We follow the architecture of LLaVA (Liu et al., 2023), one of the most popular vision-language models, to build our own explainable VLM for multimodal misinformation detection. Specifically, LLaVA consists of a vision encoder, an pretrained LLM, and a visual connector. To align pretrained LLMs with visual encoders, we use lightweight MLP projectors (Liu et al., 2023, 2024a) to connect image features into the word embedding space of the language model and then fine-tuned on instruction-formatted datasets to improve generalization and controllability.

Question-Aware Vision Amplifier. Although existing VLMs have shown incremental improvements in detecting visual distortions such as face manipulation, they typically rely on high-level semantic cues (scene, context, or objects) and struggle with subtle manipulations, especially those affecting facial expressions while preserving identity. However, directly incorporating such visual manipulation traces (Luo et al., 2021b; Li et al., 2021; Liu et al., 2024b) may negatively impact the model's performance on other types of distortions, due to potential overfitting to specific visual artifacts or a shift in representation focus.

To overcome this limitation, we introduce the Question-Aware Vision Amplifier (QAVA), a novel module inspired by the Q-Former (Li et al., 2023; Dai et al., 2023). Unlike previous approaches relying solely on whole textual instructions which often introducing distractions, QAVA employs learnable tokens conditioned specifically on explicit, task-specific question templates related to different distortion categories. Within QAVA, these tokens first utilize self-attention to understand the question context and subsequently apply cross-attention to image features, effectively extracting precise, task-relevant visual cues. The enhanced visual representations generated by QAVA serve as soft visual prompts for the LLM, directly guiding its reasoning process and substantially improving detection accuracy, especially for subtle visual manipulations.

Dataset	In-Domain				Out-of-Domain		
	MMFakeBench	Factify2	DGM^4-Face	NewsCLIPpings	MOCHEG	Fakeddit-M	VERITE
Real:Fake	300:700	1500:1500	467:433	3632:3632	200:200	200:200	200:200
Distortion Types	Mixed	Textual	Visual	Cross-modal	Textual	Visual	Cross-modal

Table 1: Evaluation Dataset Distribution

3.2 Instruction Tuning

To equip Trust-VL with misinformation-oriented logical reasoning capabilities, we carefully construct a set of instruction data for training.

Structured Reasoning Template. To mimic the human fact-checking process, we decompose misinformation detection into structured reasoning steps tailored to different types of distortions, as shown in Figure 4(a). For each distortion type, we design specific sub-queries that guide the model through a step-by-step verification process. In addition, we introduce a general question (e.g., "Is there any distortion?") to address real-world scenarios where the distortion type is unknown. Each reasoning chain consists of a sequence of sub-queries and corresponding sub-answers, starting from shared foundational steps such as analyzing the text and describing the image. These common reasoning abilities benefit from joint training across different distortion types, leading to improved generalization. After the shared steps, the chain branches into task-specific reasoning: textual distortion involves evaluating tone, stance, and evidence support; visual distortion focuses on detecting manipulated artifacts or AI-generated patterns; cross-modal distortion verifies semantic consistency between image, caption, and retrieved evidence. This structured reasoning approach closely aligns with real-world fact-checking workflows and provides interpretable, robust detection judgment.

Instruction Construction Process. Motivated by the success of recent generative models in automated instruction generation (Zhang et al., 2024), we propose a construction pipeline to generate structured reasoning instructions, as illustrated in Figure 4(b). Given a multimodal input claim and associated evidence, GPT-4o (OpenAI, 2024a) is prompted with a meticulously crafted reasoning template to produce detailed reasoning chains for misinformation detection. Each reasoning chain undergoes a rigorous verification stage, checking consistency with ground-truth labels. When inconsistencies occur, prompts are iteratively adjusted with data-driven hints that explicitly indicate the ground truth, thus guiding GPT-4o toward accurate reasoning outputs. For each dataset, we manually inspected and ensured its quality and QA format.

Statistics. To enhance Trust-VL's ability to detect misinformation across different types of distortions, we collect a set of <text, image, ground-truth label> triplets from various existing datasets (Liu et al., 2024c; Suryavardan et al., 2023; Shao et al., 2023). Based on this collection, we construct the final instruction data by applying the aforementioned procedure to incorporate step-by-step reasoning annotations. As shown in Figure 4 (c), the constructed TRUST-Instruct dataset comprises 198,253 high-quality instructions spanning three distortions.

4 Performance Study

Datasets. To demonstrate the generalization capability of Trust-VL, we evaluate the model on a diverse collection of in-domain and out-of-domain datasets covering textual, visual, and cross-modal distortions (see Table 1). *In-domain datasets* include MMFakeBench (Liu et al., 2025), which features mixed distortion types; Factify2 (Suryavardan et al., 2023), a textual fact-checking benchmark supporting multimodal claim verification; DGM^4-Face (Shao et al., 2023), focused on detecting deepfake-powered facial manipulations such as face swap and face attribution; and NewsCLIPpings (Luo et al., 2021a), the largest synthetic benchmark for out-of-context (OOC) misinformation detection through replacing the images in the original claims with retrieved images that are semantically related but belong to different news events. *Out-of-domain datasets* include MOCHEG (Yao et al., 2023), a textual misinformation dataset with journalist-provided claim verifications; Fakeddit-M (Nakamura et al., 2020), a Reddit-sourced visual misinformation dataset under the Manipulated Content category (e.g., images are digitally edited); and VERITE (Papadopoulos et al., 2024), a real-world OOC benchmark, featuring modality-balanced image-text pairs. We evaluate model performance using binary classification metrics, including Accuracy (Acc.) and macro-F1, for the real/fake detection task.

Methods	Avg. Acc.	In-Domain								Out-of-Domain					
		MMFakeBench		Factify2		DGM^4-Face		NewsCLIPpings		MOCHEG		Fakeddit-M		VERITE	
		Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
General-purpose VLMs															
BLIP2	53.36	37.40	34.45	54.30	42.38	47.70	34.35	50.14	34.28	62.50	57.16	70.75	70.19	50.75	37.35
InstructBLIP	58.41	57.30	56.38	66.83	66.48	50.40	48.66	53.85	50.71	63.25	60.85	64.75	62.83	52.50	49.60
LLaVA	60.25	62.60	61.72	79.59	79.10	46.41	38.14	45.87	48.54	66.50	64.71	68.00	66.67	52.75	49.80
xGen-MM	62.20	65.40	62.77	86.03	86.04	50.10	49.68	59.87	59.18	59.50	56.32	60.00	53.45	54.50	54.41
LLaVA-NeXT	62.35	71.60	65.99	79.60	79.09	53.40	52.21	59.86	59.37	58.25	52.52	59.00	52.36	54.75	54.57
Qwen2-VL	69.85	67.00	66.28	89.40	89.37	48.10	41.63	70.94	69.91	66.25	64.57	77.25	76.96	70.00	68.94
GPT-4o	76.16	83.10	80.88	88.37	88.21	57.14	49.24	86.51	86.51	77.00	76.81	73.50	73.12	67.50	67.57
o1	77.74	83.90	82.41	96.90	96.90	50.06	38.06	86.80	86.54	81.50	81.38	73.25	73.07	71.75	71.66
Misinformation Detectors															
MMD-Agent	56.11	69.10	48.68	71.03	69.35	48.30	48.29	53.06	41.12	54.25	43.72	42.25	42.24	54.75	47.00
SNIFFER	61.17	51.40	51.33	61.00	55.97	47.20	37.96	88.85	88.85	53.75	50.73	53.50	51.13	72.50	72.02
LRQ-FACT	66.60	71.30	74.00	86.63	89.79	41.80	44.14	68.19	73.45	66.25	69.25	67.25	71.77	64.75	68.32
TRUST-VL	**86.16**	**87.30**	**85.42**	**99.50**	**99.50**	**88.50**	**88.39**	**90.35**	**90.35**	**82.75**	**82.58**	**82.50**	**82.20**	**73.75**	**73.61**
Δ	↑8.42	↑3.40	↑3.01	↑2.60	↑2.60	↑31.36	↑36.18	↑1.50	↑1.50	↑1.25	↑1.20	↑5.25	↑5.24	↑1.25	↑1.59

Table 2: Performance (%) comparison between Trust-VL and other baseline VLMs across in-domain and out-of-domain datasets. The best score is highlighted in blue, and the second-best score is underlined. The absolute improvement over the second-best model is highlighted in green.

Variants	MMFakeBench		Factify2		DGM^2-Face		NewsCLIPpings	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
TRUST-VL-13B	**87.30**	**85.42**	**99.50**	**99.50**	**88.50**	**88.39**	**90.35**	**90.35**
w/o Reasoning	83.60	81.25	87.31	87.30	80.00	79.91	85.99	85.98
w/o Common Reasoning	84.60	81.42	99.20	99.20	70.90	70.68	89.00	89.00
w/o QAVA	84.60	82.16	89.17	89.17	72.79	72.59	87.31	87.30
LLM Size: 7B	85.90	83.65	99.33	99.33	80.90	80.64	88.79	88.79

Table 3: Ablation study of different model modules in TRUST-VL.

Baselines. We compare TRUST-VL with both general-purpose VLMs and specialized misinformation detectors. For *general-purpose VLMs*, we include BLIP-2 (Li et al., 2023), InstructBLIP (Dai et al., 2023), LLaVA (Liu et al., 2023), LLaVA-NeXT (Li et al., 2024a), xGen-MM (Xue et al., 2024), and Qwen2-VL (Wang et al., 2024), which are all open-source VLMs primarily designed for multimodal understanding and reasoning tasks. We also include GPT-4o (OpenAI, 2024a) and o1 (OpenAI, 2024b), two advanced closed-source VLMs. For *specialized misinformation detectors*, we consider SNIFFER (Qi et al., 2024), an explainable VLM-based detector for OOC misinformation through a two-stage instruction; MMD-Agent (Liu et al., 2025), a multi-agent framework that utilizes GPT-4o for dynamic multimodal query resolution, and LRQ-FACT (Beigi et al., 2024), a fact-checking system based on a multi-LLM architecture that improves context reasoning.

Implementation Details. We fine-tune our model with different stages, leveraging LLaVA-1.5 (Liu et al., 2024a) with `vicuna-13b-v1.5` as the LLM and CLIP (`ViT-L/14`) as the image encoder. In Stage-1, we train the Connector module on 1.2M samples (including 653K news samples from VisualNews(Liu et al., 2020)) for one epoch to align visual features with the language model. In Stage-2, we further train both the LLM and visual connector using 665K synthetic conversation samples for one epoch, eliciting the model's ability to follow complex instructions. In Stage 3, we fine-tune the full model on 198K reasoning samples generated by GPT-4o for three epochs to further enhance misinformation-specific reasoning capabilities. The learning rates are set to 2e-5 for the LLM and 2e-6 for the vision encoder, with a batch size of 128. All models are trained and evaluated on 8 Nvidia H100 (80G) GPUs. Finetuning on TRUST-Instruct-198K completes within 22 hours.

4.1 Performance Comparison

As shown in Table 2, we can see: 1) Our proposed TRUST-VL significantly outperforms all baselines on both in-domain and out-of-domain datasets, achieving more than 8 percentage points improvement in average accuracy. This demonstrates that TRUST-VL can effectively capture the key detection cues across different distortion types and generalize well to unseen claims. 2) General-purpose VLMs, particularly OpenAI-o1, exhibit competitive performance on textual and cross-modal distortions, but still struggle with subtle visual manipulations. Specifically, o1 achieves

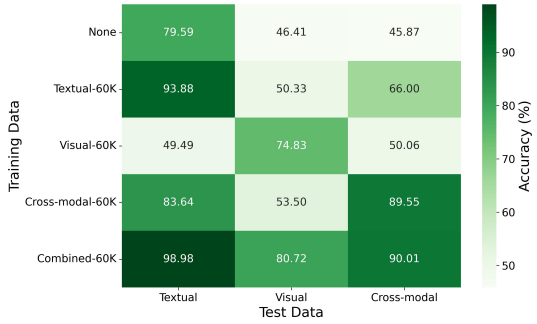

Figure 5: Accuracy heatmap of LLaVA across different training and testing distortion types. The first row ("None") refers to the performance of the original LLaVA baseline without any training.

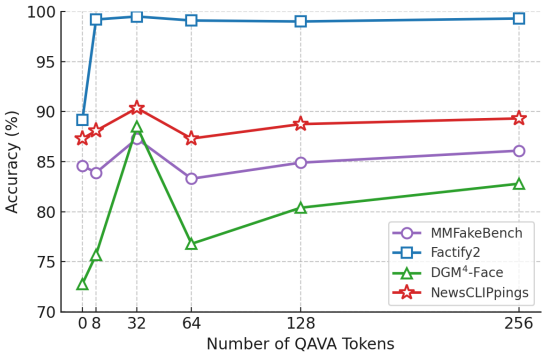

Figure 6: The impact of different numbers of learnable QAVA tokens across datasets.

an overall accuracy of 77.74%, but its performance drops significantly on DGM^4-Face (50.06%), indicating challenges in detecting manipulated facial content. Besides, o1 also outperforms GPT-4o, especially on textual distortions, suggesting that enhanced reasoning capabilities can benefit misinformation detection. 3) Existing misinformation detectors that rely on multiple independent LLMs for step-by-step reasoning, such as MMD-Agent (56.11% accuracy) and LRQ-FACT (66.60%), perform worse than general-purpose VLMs. This may be due to conflicting reasoning paths across different modules, which potentially compromise the overall decision-making process.

4.2 Ablation Study

We conduct ablation studies to systematically evaluate the roles of diverse model modules, joint training across distortions and QAVA token count.

Effect of Model Modules. To evaluate the effects of different components in our model, we design several ablated variants of TRUST-VL: w/o Reasoning: The model is trained only for binary classification (*i.e.*, real vs. fake), without generating structured reasoning chains; w/o Common Reasoning: The shared reasoning steps (i.e., text analysis and visual understanding) are removed during instruction data construction; w/o QAVA: The QAVA module is removed from the model; 7B LLM: We replace the 13B backbone LLM with a smaller 7B version. From Table 3, we observe: 1) Removing any single module leads to a performance drop, validating the contribution of each component to the overall effectiveness of TRUST-VL. 2) w/o Reasoning causes a substantial performance degradation (4–12 percentage points across datasets), highlighting the importance of structured reasoning supervision for accurate judgment. In addition, removing the shared reasoning steps results in a noticeable decline, particularly on datasets involving fine-grained visual manipulation. This suggests that textual and visual descriptions provide crucial semantic grounding for subtle distortion detection. 3) w/o QAVA results in a significant performance drop across all datasets, with the largest degradation (15.71 percentage points) on visual distortion tasks. This confirms the effectiveness of QAVA in learning task-specific visual representations. 4) Using a 7B LLM instead of 13B leads to a moderate performance decline, but still outperforms the second-best baseline from Table 2, demonstrating the robustness and efficiency of the our instruction framework even with smaller backbones.

Effect of Joint Training. To examine whether different distortion types can benefit from joint training, we conduct a small-scale experiment based on the original LLaVA model. We separately train the model using instruction data from each individual distortion type (textual, visual, or cross-modal), and compare the results with a jointly trained model using a balanced mix of all three types. To ensure a fair comparison, all models are trained on 60K samples. As shown in Figure 5, models trained on a single distortion type generally perform well on in-domain evaluation but struggle to generalize to unseen distortions. In contrast, the jointly trained model achieves consistently better performance across all distortion types, confirming that shared reasoning abilities can be enhanced through joint training and transferred across tasks.

Effect of QAVA Token Count. Figure 6 further illustrates how the number of learnable visual tokens in the QAVA module influences the performance of TRUST-VL. Introducing the QAVA module consis-

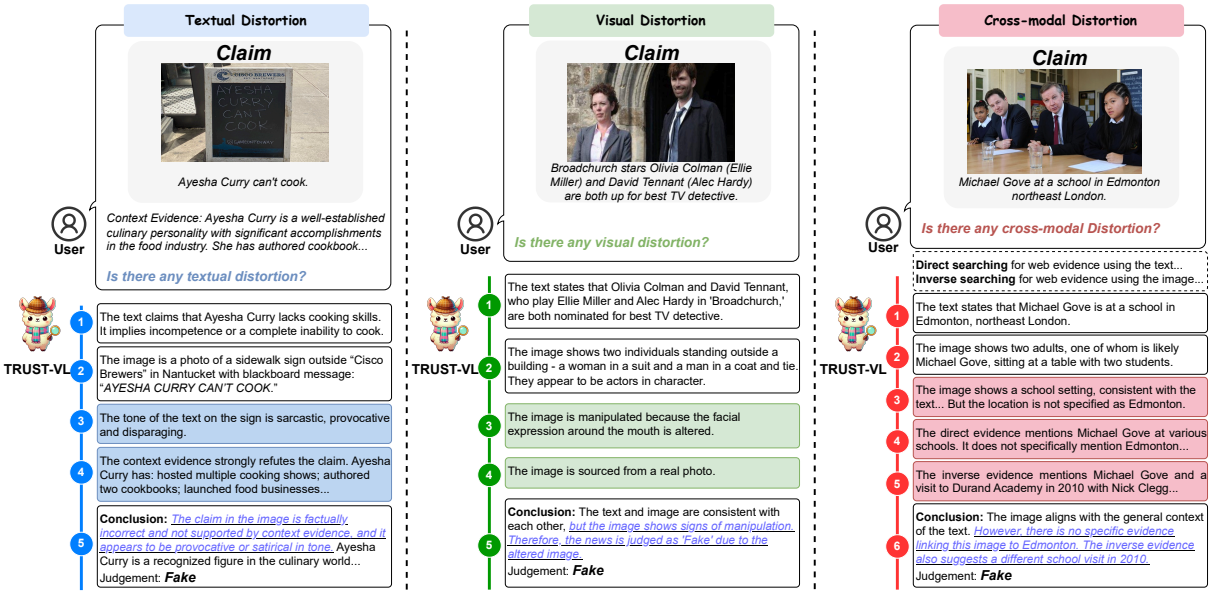

Figure 7: Illustrating examples of multimodal distortion spanning textual, visual, and cross-modal scenarios.

tently improves accuracy across all datasets, with particularly notable gains on Factify2 (accuracy increases from 72.79% to 88.50%), showing its critical role in detecting visual distortions. Moreover, the number of QAVA tokens significantly affects model performance. We observe that increasing the token count initially leads to performance gains, but beyond a certain point, further increases yield diminishing or even negative returns. Specifically, 32 tokens achieves the best performance across all datasets, suggesting it provides an optimal balance—sufficient to capture task-specific visual differences while avoiding excessive computational overhead and the risk of overfitting.

4.3 Case Study

Figure 7 illustrates three representative examples that current general VLMs and specialized detectors typically fail to handle effectively. In constrast, TRUST-VL addresses all three types and accurately identify the fake information with a structured and persuasive chain of reasoning steps.

In the first example, a textual claim, "Ayesha Curry can't cook", explicitly contradicts well-documented facts but is stated with satirical tone, misleading the public audience. The second example presents subtle visual misinformation by showing a manipulated photograph of actors Olivia Colman and David Tennant, in which facial expressions have been covertly altered, thereby introducing deceptive emotional signals. The third example features cross-modal misinformation, where an authentic image of politician Michael Gove at a different news event is wrongly localized by the caption. These examples underscore the need for a system that can robustly and transparently address textual inaccuracies, visual manipulations, and cross-modal inconsistencies simultaneously in real-world news applications. Please refer to the Supplementary Material for additional case studies on model comparisons.

5 Conclusion

In this paper, we address the challenge of general multimodal misinformation detection, which involves diverse types of distortions, including textual, visual, and cross-modal inconsistencies. We observe that these tasks share common reasoning abilities while also requiring specialized skills for each distortion type. Based on this insight, we propose that joint training across distortion types can enhance model performance. To this end, we introduce TRUST-VL, a unified, explainable vision-language model equipped with a novel Question-Aware Visual Amplifier module, explicitly designed to extract task-specific visual features. To train this model, we construct TRUST-Instruct, a large-scale instruction dataset consisting of 198K samples with structured reasoning chains that mimic human fact-checking processes. Comprehensive experiments demonstrate that TRUST-VL achieves state-of-the-art performance on both in-domain and out-of-domain benchmarks. We believe TRUST-VL offers a promising foundation for future research on general and interpretable misinformation detection in real-world scenarios.

Limitations

Although TRUST-VL achieves strong performance, it has several limitations. First, the structured reasoning chains are guided by manually designed task queries, rather than being learned or evolved by the model; incorporating reinforcement learning could further enhance the adaptability of the reasoning process. Second, while visual evidence is retrieved, it is converted to text for reasoning. The more direct comparison in the visual space could offer richer signals. Lastly, our focus on visual distortion is limited to face-related manipulations, leaving other forms such as object-based or video misinformation for future exploration.

References

Sara Abdali, Sina Shaham, and Bhaskar Krishnamachari. 2025. Multi-modal misinformation detection: Approaches, challenges and opportunities. *ACM Comput. Surv.*, 57(3):76:1–76:29.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022*, pages 14920–14929. IEEE.

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, Singapore. Association for Computational Linguistics.

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, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019. Multifc: A real-world multi-domain dataset for evidence-based fact checking of claims. 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 2019*, pages 4684–4696. Association for Computational Linguistics.

Alimohammad Beigi, Bohan Jiang, Dawei Li, Tharindu Kumarage, Zhen Tan, Pouya Shaeri, and Huan Liu. 2024. LRQ-FACT: Llm-generated relevant questions for multimodal fact-checking. *CoRR*, abs/2410.04616.

Canyu Chen and Kai Shu. 2024. Can llm-generated misinformation be detected? In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. 2023. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023*.

Frederik Federspiel, Ruth Mitchell, Asha Asokan, Carlos Umana, and David McCoy. 2023. Threats by artificial intelligence to human health and human existence. *BMJ global health*, 8(5):e010435.

Zhengchao Huang, Bin Xia, Zicheng Lin, Zhun Mou, and Wenming Yang. 2024. FFAA: multimodal large language model based explainable open-world face forgery analysis assistant. *CoRR*, abs/2408.10072.

Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking for public health claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7740–7754, Online. Association for Computational Linguistics.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. 2024a. LLaVA-NeXT-Interleave: Tackling multi-image, video, and 3d in large multimodal models. *CoRR*, abs/2407.07895.

Jiaming Li, Hongtao Xie, Jiahong Li, Zhongyuan Wang, and Yongdong Zhang. 2021. Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 6458–6467. Computer Vision Foundation / IEEE.

Jiawei Li, Fanrui Zhang, Jiaying Zhu, Esther Sun, Qiang Zhang, and Zheng-Jun Zha. 2024b. Forgerygpt: Multimodal large language model for explainable image forgery detection and localization. *CoRR*, abs/2410.10238.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning, ICML 2023*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

Fuxiao Liu, Yinghan Wang, Tianlu Wang, and Vicente Ordonez. 2020. Visualnews : Benchmark and challenges in entity-aware image captioning. *CoRR*, abs/2010.03743.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024*, pages 26286–26296. IEEE.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023*.

Xuannan Liu, Peipei Li, Huaibo Huang, Zekun Li, Xing Cui, Jiahao Liang, Lixiong Qin, Weihong Deng, and Zhaofeng He. 2024b. FKA-Owl: Advancing multimodal fake news detection through knowledge-augmented lvlms. In *Proceedings of the 32nd ACM International Conference on Multimedia, MM 2024*, pages 10154–10163. ACM.

Xuannan Liu, Zekun Li, Peipei Li, Shuhan Xia, Xing Cui, Linzhi Huang, Huaibo Huang, Weihong Deng, and Zhaofeng He. 2024c. MMFakeBench: A mixed-source multimodal misinformation detection benchmark for lvlms. *CoRR*, abs/2406.08772.

Xuannan Liu, Zekun Li, Peipei Li, Shuhan Xia, Xing Cui, Linzhi Huang, Huaibo Huang, Weihong Deng, and Zhaofeng He. 2025. MMFakeBench: A mixed-source multimodal misinformation detection benchmark for lvlms. In *The Thirteenth International Conference on Learning Representations, ICLR 2025*. OpenReview.net.

Grace Luo, Trevor Darrell, and Anna Rohrbach. 2021a. Newsclippings: Automatic generation of out-of-context multimodal media. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021*, pages 6801–6817. Association for Computational Linguistics.

Yuchen Luo, Yong Zhang, Junchi Yan, and Wei Liu. 2021b. Generalizing face forgery detection with high-frequency features. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 16317–16326. Computer Vision Foundation / IEEE.

Kai Nakamura, Sharon Levy, and William Yang Wang. 2020. Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020*, pages 6149–6157. European Language Resources Association.

OpenAI. 2024a. Hello GPT-4o. Accessed: 2024-11-01.

OpenAI. 2024b. Openai o1 system card. *CoRR*, abs/2412.16720.

Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan Luu, William Yang Wang, Min-Yen Kan, and Preslav Nakov. 2023. Fact-checking complex claims with program-guided reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6981–7004, Toronto, Canada. Association for Computational Linguistics.

Stefanos-Iordanis Papadopoulos, Christos Koutlis, Symeon Papadopoulos, and Panagiotis C. Petrantonakis. 2024. VERITE: a robust benchmark for multimodal misinformation detection accounting for unimodal bias. *International Journal of Multimedia Information Retrieval*, 13(1):4.

Peng Qi, Zehong Yan, Wynne Hsu, and Mong-Li Lee. 2024. SNIFFER: Multimodal large language model for explainable out-of-context misinformation detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 13052–13062. IEEE.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

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

S. Suryavardan, Shreyash Mishra, Parth Patwa, Megha Chakraborty, Anku Rani, Aishwarya Naresh Reganti, Aman Chadha, Amitava Das, Amit P. Sheth, Manoj Chinnakotla, Asif Ekbal, and Srijan Kumar. 2023. Factify 2: A multimodal fake news and satire news dataset. In *Proceedings of De-Factify 2: 2nd Workshop on Multimodal Fact Checking and Hate Speech Detection, co-located with AAAI 2023*, volume 3555 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Sahar Tahmasebi, Eric Müller-Budack, and Ralph Ewerth. 2024. Multimodal misinformation detection using large vision-language models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, CIKM 2024, Boise, ID, USA, October 21-25, 2024*, pages 2189–2199. ACM.

Camille Thibault, Gabrielle Peloquin-Skulski, Jacob-Junqi Tian, Florence Laflamme, Yuxiang Guan, Reihaneh Rabbany, Jean-François Godbout, and Kellin Pelrine. 2024. A guide to misinformation detection data and evaluation. *arXiv preprint arXiv:2411.05060*.

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, New Orleans, Louisiana. Association for Computational Linguistics.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *CoRR*, abs/2409.12191.

Danni Xu, Shaojing Fan, and Mohan S. Kankanhalli. 2023. Combating misinformation in the era of generative AI models. In *Proceedings of the 31st ACM International Conference on Multimedia, MM 2023*, pages 9291–9298. ACM.

Keyang Xuan, Li Yi, Fan Yang, Ruochen Wu, Yi R. Fung, and Heng Ji. 2024. LEMMA: towards lvlm-enhanced multimodal misinformation detection with external knowledge augmentation. *CoRR*, abs/2402.11943.

Le Xue, Manli Shu, Anas Awadalla, Jun Wang, An Yan, Senthil Purushwalkam, Honglu Zhou, Viraj Prabhu, Yutong Dai, Michael S. Ryoo, Shrikant Kendre, Jieyu Zhang, Can Qin, Shu Zhang, Chia-Chih Chen, Ning Yu, Juntao Tan, Tulika Manoj Awalgaonkar, Shelby Heinecke, and 8 others. 2024. xGen-MM (BLIP-3): A family of open large multimodal models. *CoRR*, abs/2408.08872.

Barry Menglong Yao, Aditya Shah, Lichao Sun, Jin-Hee Cho, and Lifu Huang. 2023. 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, SIGIR 2023*, pages 2733–2743. ACM.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2024. Multimodal chain-of-thought reasoning in language models. *Trans. Mach. Learn. Res.*, 2024.

A Model Details

As illustrated in Table 4 and Figure 8, we progressively fine-tune our model with three stages, including language-image alignment, news domain alignment, visual instruction tuning, and misinformation tuning. Following original LLaVA (Liu et al., 2023, 2024a), we use the same pretraining data and instructions in the first two stages, while introducing a large-scale reasoning-rich TRUST-Instruct to further enhance news understanding and misinformation-specific logicial reasoning.

To capture detailed visual information for subtle artifact detection, TRUST-VL adopts a dynamic, high-resolution image encoding strategy proven effective in recent VLMs (Li et al., 2024a; Xue et al., 2024). This approach employs patch-wise image encoding, where the original high-resolution image is partitioned into multiple smaller patches, each individually encoded. These patch-level encodings are then concatenated with a downsized version of the original image that provides global contextual information. We utilize the pre-trained CLIP encoder (Radford et al., 2021) to obtain visual representations. To align pretrained LLMs with visual encoders, we use lightweight MLP projectors (Liu et al., 2023, 2024a) to connect image features into the word embedding space of the language model and then fine-tuned on instruction-formatted datasets to improve generalization and controllability. The language tokens consist of a system message, task-specific instruction, input text, retrieved evidence, and targeted questions.

In our experiments, we use the following model checkpoints as baselines: blip2-flan-t5-xl, instructblip-vicuna-13b, llava-v1.5-13b, llava-v1.6-mistral-13b-hf, xgen-mm-phi3-mini-instruct-r-v1, Qwen2-VL-7B-Instruct. For detectors such as MMD-Agent and LRQ-FACT, we utilize llava-v1.5-13b as the VLM for fair comparison.

B Datasets

To evaluate the effectiveness of multimodal misinformation detection models, we leverage a diverse set of in-domain and out-of-domain datasets covering textual, visual, and cross-modal misinformation. These datasets enable a comprehensive assessment of misinformation detection across different modalities and manipulation techniques.

- **MMFakeBench** (Liu et al., 2025) is a multimodal misinformation detection benchmark designed to evaluate robustness against various manipulation techniques. It contains 1,000 instances with an distribution of real samples and manipulated cases, including textual veracity distortions, visual veracity distortions, and cross-modal consistency distortions. The dataset introduces 12 forgery types, making it a comprehensive benchmark for evaluating multimodal misinformation detection.

- **Factify2** (Suryavardan et al., 2023) is a multimodal fact-checking dataset comprising 50,000 instances of supporting and refuting claims sourced from fact-checking platforms

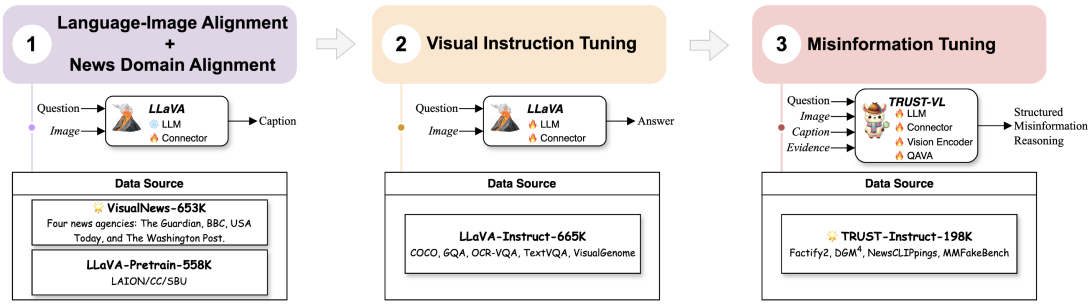

Figure 8: Progressive training strategy.

Configurations	Details
Architecture	**Image Encoder**: CLIP-Large (336×336) **Connector**: 2-Layer MLP **QAVA**: 6 Transformer Layers with 32 Learnable Tokens **LLM**: Vicuna-1.5 13B
# Total Parameters	13B
Visual Representations	Dynamic: 336 × {2×2, 1×{2,3}, {2,3}×1}
Stage-1	**Training Data**: 1211K **Trainable Module**: Connector
Stage-2	**Training Data**: 665K **Trainable Module**: LLM, Connector
Stage-3	**Training Data**: 198K **Trainable Module**: Full model
Training Data (#Samples)	2074K = 1211K + 665K + 198K
Training Schedule	**Learning Rate**: - LLM: 2e-5 - Vision Encoder: 2e-6 **Training Epochs**: - Stage-1: 1 epoch - Stage-2: 1 epoch - Stage-3: 3 epochs **Batch Size**: 128

Table 4: Model Architecture and Training Details

such as PolitiFact. This dataset extends the original Factify dataset by incorporating a wider range of real and manipulated news content, including satirical articles.

- **DGM^4-Face** (Shao et al., 2023) a large-scale dataset generated by two image manipulation and two text manipulation approaches, with the objective of detecting and grounding manipulations in image-text pairs of human-centric news. The original dataset consists a total of 230k news samples, including 77,426 pristine image-text pairs and 152,574 manipulated pairs. We randomly sample 467 real images and 433 manipulated instances, including face swaps and face attribute modifications.

- **NewsCLIPpings** (Luo et al., 2021a) is the largest synthetic benchmark for detecting out-of-context (OOC) misinformation. It generates OOC samples by replacing images in original image-caption pairs with real and semantically related images from different news events. (Abdelnabi et al., 2022) further extends this dataset by incorporating textual and visual evidence retrieved via Google Search APIs to improve detection performance.

- **MOCHEG** (Yao et al., 2023) is a large-scale dataset for end-to-end multimodal fact-checking, comprising 15,601 claims, each annotated with a truthfulness label and a ruling statement. It includes 33,880 paragraphs and 12,112 images as evidence. It is sourced from

```
# system message
Task description: some rumormongers intentionally write fake news, manipulate images, or
use images from other news events to make multimodal misinformation. Given a news text and
a news image, you are responsible for judging whether the given text and image are both
credible and faithfully represent the news event. You will be presented with a text and an
image. You should use the following step-by-step instructions to derive your judgement:
# shared steps
Step 1 - Analyze the text: Carefully review the provided text, summarize its key facts,
events, and entities. Pay attention to any misleading, false, or fabricated contents.
Step 2 - Provide a detailed description of the news image: Identify the main subjects, such
 as people, groups, or specific elements related to the news event.
# specialized steps
Step 3 -...
# conclusion
Step 6 - What is your final judgement? According to the previous steps, you will first
think out loud about your eventual conclusion, enumerating reasons why the news does or
does not contain false information. After thinking out loud, you should output either 'Real
' or 'Fake' depending on whether you think the given text and accompanying image are both
truthful and consistent: 'Real' if the news is factually correct and the image faithfully
represent the news text, or 'Fake' if the news is misleading, manipulated or the image is
used out of context.

# input
<image>
Caption: <caption>
Direct Evidence: <direct evidence>
Inverse Evidence: <inverse evidence>
Context Evidence: <context evidence>
Your judgement:
```

Figure 9: Prompt used to ask GPT-4o to generate the instruction data.

```
# system message
You are a misinformation detection assistant. Task description: some rumormongers
intentionally write fake news, manipulate images, or use images from other news events to
make multimodal misinformation. Given a news text and a news image, you are responsible for
 judging whether the given text and image are both credible and faithfully represent the
news event. You will be presented with a text, an image, direct evidence, and inverse
evidence. For final judgement, you should output either 'Real' or 'Fake' depending on
whether you think the given text and accompanying image are both truthful and consistent: '
Real' if the news is factually correct and the image faithfully represent the news text, or
 'Fake' if the news is misleading, manipulated or the image is wrongly used in the news
text.

A few rules:
- If a specific type of evidence (i.e., direct, or inverse) is not provided, state clearly:
 'There is no {type} evidence.'
- Do not nitpick over the direct and inverse evidence as it may contain some noise.
- Your judgement must always end with either 'Real' or 'Fake'.

# input
<image>
Caption: <caption>
Direct Evidence: <direct evidence>
Inverse Evidence: <inverse evidence>
Context Evidence: <context evidence>
Your judgement:
```

Figure 10: TRUST-VL language input.

fact-checking platforms and serves as a benchmark for evaluating the ability of models to verify textual claims. For fair evaluation, we sample 400 news instances with a balanced distribution of real and fake samples.

- **Fakeddit** (Nakamura et al., 2020) is a large-scale multimodal fake news dataset collected from Reddit. It contains over 1 million instances across multiple categories of misinformation, providing a fine-grained 2-way, 3-

way, and 6-way classification of fake news. Similarly, we sample 400 news instances with an equal number of real and fake claims.

- **VERITE** (Papadopoulos et al., 2024) is a real-world dataset designed for detecting OOC misinformation, which effectively mitigates the problem of unimodal bias and provides a more robust and reliable evaluation framework. A balanced subset of 400 samples is used to ensure fair evaluation.

C Model Prompts

Figure 9 illustrates the prompt utilized for asking GPT-4o to generate instruction data. For each claim, we retrieve textual and visual evidence (converted to text via image captioning) separately and then pass them to GPT-4o to process. We also consider context evidence provided by users or downstream tasks. For specialized steps, we carefully design critical steps required for addressing different distortion types. Finally, GPT-4o outputs a final judgment along with detailed explanations, guided by carefully designed step-by-step reasoning instructions. Figure 10 shows language input for TRUST-VL framework. Together, these prompt designs ensure high-quality reasoning supervision during training and robust, explainable predictions.

D Taxonomy of Capabilities

To detect the many faces of multimodal misinformation, we delineate a set of reasoning capabilities (See Table 5). Grouped into *shared* and *specialized* categories, these capabilities guide the construction of our TRUST-INSTRUCT dataset, each addressing characteristic misinformation patterns spanning text, vision, and cross-modal reasoning steps.

E Baselines

- **BLIP-2** (Li et al., 2023) is a vision-language model that bridges the modality gap between vision and language models without requiring training from scratch. It employs a Querying Transformer to effectively align visual features with language models.

- **InstructBLIP** (Dai et al., 2023) is an instruction-tuned version of BLIP-2, designed to handle a wide range of vision-language tasks through instruction tuning. By integrating visual instruction tuning, InstructBLIP achieves improved performance across various tasks, including image captioning and visual question answering.

- **LLaVA** (Liu et al., 2023) is one of pioneer in visual instruction tuning. Compared with the original linear projection, it improves the vision-language connector's representation power with a two-layer MLP to enhance multimodal capabilities.

- **LLaVA-NeXT** (Li et al., 2024a) is an enhanced version of LLaVA, incorporating advanced techniques for improved vision-language alignment and understanding. It builds upon the original LLaVA framework to offer more accurate and contextually relevant responses in multimodal interactions.

- **xGen-MM** (Xue et al., 2024) also known as BLIP-3, is a large multimodal model framework which replaces the complex Q-Former module used in BLIP-2 with a scalable vision token sampler, specifically a perceiver resampler, to process visual inputs. Additionally, xGen-MM is able to handle free-form interleaved sequences of images and text by adopting a single auto-regressive loss function focused on text token prediction.

- **Qwen2-VL** (Wang et al., 2024) is a vision-language model that integrates visual understanding with language processing capabilities. Its architecture introduces two key innovations: Naive Dynamic Resolution, allowing the model to process images of varying resolutions by dynamically adjusting the number of visual tokens, and Multimodal Rotary Position Embedding (M-RoPE), which decomposes positional embeddings into temporal and spatial components to effectively handle 1D textual, 2D visual, and 3D video data.

- **GPT-4o** (OpenAI, 2024a). This is currently one of the most powerful multimodal large language models. We utilize GPT-4o in a zero-shot manner with step-by-step instructions for multimodal misinformation detection.

- **o1** (OpenAI, 2024b) is the latest MLLM with advanced reasoning capabilities via large-scale reinforcement learning. For fair comparison, we adopt o1 using the same evaluation protocol as GPT-4o.

Abilities	Definitions
	Shared Abilities
Textual Analysis	Extracts key factual elements (e.g., entities, dates, events) from text and lists statements to be verified.
Visual Understanding	Interprets salient visual content (e.g., entities, scenes, actions) and identifies visual cues of manipulation, such as unnatural lighting, texture inconsistencies, distorted facial features, duplicated patterns, or incoherent backgrounds.
Evidence Reasoning	Cross-checks the claim against retrieved or user-provided evidence to identify factual support or contradiction. This capability is essential for verifying non-factual claims and detecting out-of-context image–text pairings.
News Knowledge	Recalls factual world knowledge about people, places, or events to contextualize the claim, even without using external information.
	Specialized Abilities
Linguistic Patterns	Identifies rhetorical cues (e.g., bias, satire, sentiment) that may signal misleading or manipulative intent in the text.
Visual Artifacts	Detects pixel-level or visual artifacts (e.g., lighting issues, texture mismatches) indicating image manipulation or generation.
Semantic Consistency	Assesses the semantic matching between textual and visual modalities to detect out-of-context misinformation. Discrepancies can indicate that authentic images are being misused to support misleading narratives.

Table 5: Taxonomy of reasoning capabilities required for multimodal misinformation detection.

- **SNIFFER** (Qi et al., 2024). This is the state-of-the-art multimodal large language model designed for OOC misinformation detection. It employs a two-stage instruction tuning on InstructBLIP (Dai et al., 2023) for the cross-modal consistency checks.

- **MMD-Agent** (Liu et al., 2025) is a multimodal agent framework that integrates the reasoning, action, and tool-use capabilities of LVLM agents. It decomposes misinformation detection into three sequential stages: textual veracity check, visual veracity check, and cross-modal consistency reasoning. This structured approach enables systematic and thorough analysis. At each stage, MMD-Agent prompts LVLMs to generate multi-perspective reasoning traces and coordinates their outputs to obtain a final decision.

- **LRQ-FACT** (Beigi et al., 2024) is a fact-checking system that utilizes a multi-agent framework to leverage VLMs and LLMs to generate comprehensive questions and answers for understanding multimodal content. Then, a decision-maker LLM assesses the veracity based on all generated context.

F More Cases

As shown in Figure 12, general VLMs fail to detect visual distortions on the person's face, as well as cross-modal distortion (*i.e.,* event mismatch between the text and image). General-purpose models like GPT-4o and LLaVA overlook these subtle manipulations and accept the content as factual. In contrast, TRUST-VL accurately identifies the misinformation by conducting multi-step reasoning, cross-referencing temporal and contextual evidence, and pinpointing inconsistencies across modalities. This demonstrates TRUST-VL's superior ability to handle nuanced, real-world misinformation scenarios that require both shared and task-specific reasoning capabilities.

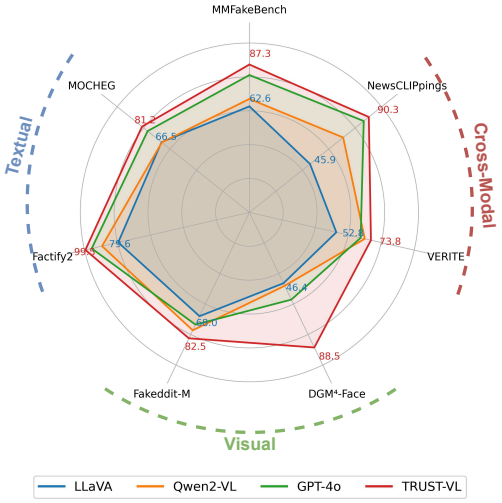

Figure 11: Performance (%) comparison between TRUST-VL and general VLMs.

Figure 13 showcases three real-world misinformation cases, each demonstrating a distinct distortion type: textual, visual, and cross-modal. Specialized misinformation detectors such as MMD-Agent tend to produce shallow or incomplete assessments. For instance, in the Ayesha Curry case, it offers

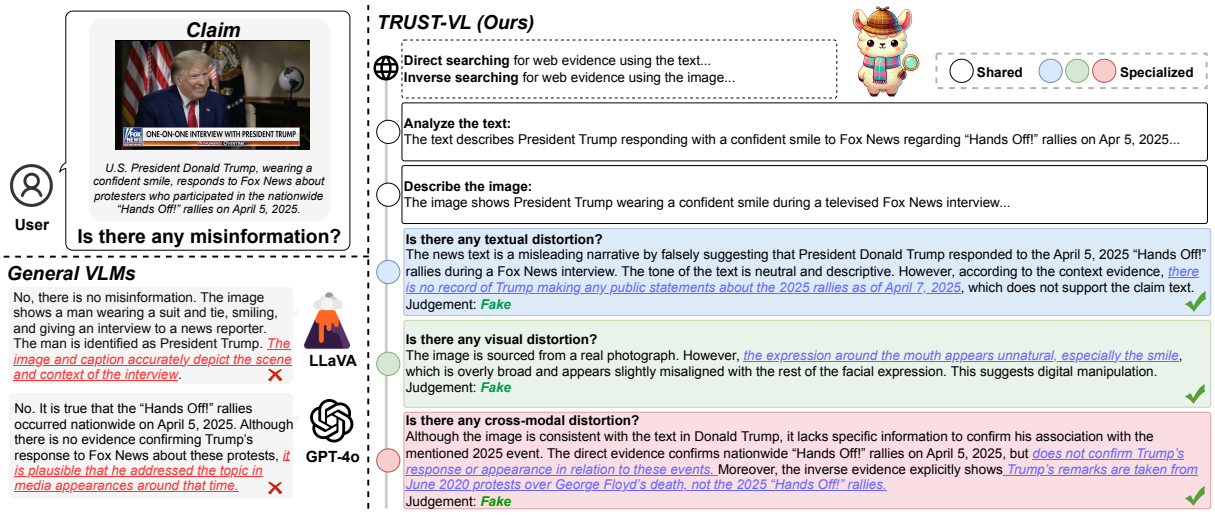

Figure 12: Comparison between the proposed TRUST-VL and general large vision-language models on complex case where false information spans across multiple modalities at the same time.

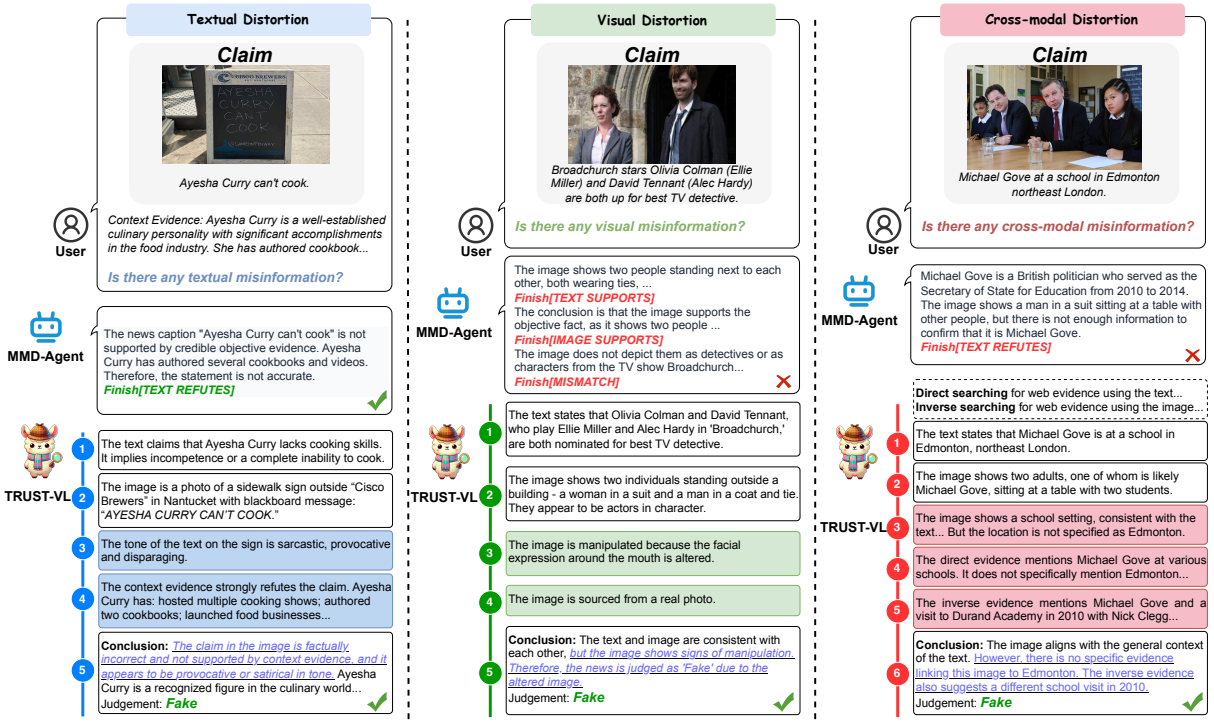

Figure 13: Comparison between the proposed TRUST-VL and specialized detectors.

a brief factual correction without recognizing the satirical tone; in the Olivia Colman case, it fails to detect the subtle visual manipulation; and in the third case, it misidentifies the setting despite contradictory evidence. These limitations highlight MMD-Agent's lack of in-depth reasoning and explainability, especially when dealing with subtle visual manipulations or cross-modal distortions, which TRUST-VL addresses more effectively.