M^{3}D: MultiModal MultiDocument Fine-Grained Inconsistency Detection

Chia-Wei Tang[♡] **Ting-Chih Chen**[♡] **Kiet A. Nguyen**[♠]
Kazi Sajeed Mehrab[♡] **Alvi Md Ishmam**[♡] **Chris Thomas**[♡]
[♡]Virginia Tech [♠]University of Illinois Urbana-Champaign
{cwtang, tingchih, ksmehrab, alvi, christhomas}@vt.edu kietan2@illinois.edu
https://tverous.github.io/M3D/

Abstract

Fact-checking claims is a highly laborious task that involves understanding how each factual assertion within the claim relates to a set of trusted source materials. Existing approaches make sample-level predictions but fail to identify the specific aspects of the claim that are troublesome and the specific evidence relied upon. In this paper, we introduce a method and new benchmark for this challenging task. Our method predicts the fine-grained logical relationship of each aspect of the claim from a set of multimodal documents, which include text, image(s), video(s), and audio(s). We also introduce a new benchmark (M^3DC) of claims requiring multimodal multidocument reasoning, which we construct using a novel claim synthesis technique. Experiments show that our approach outperforms other models on this challenging task on two benchmarks while providing finer-grained predictions, explanations, and evidence.

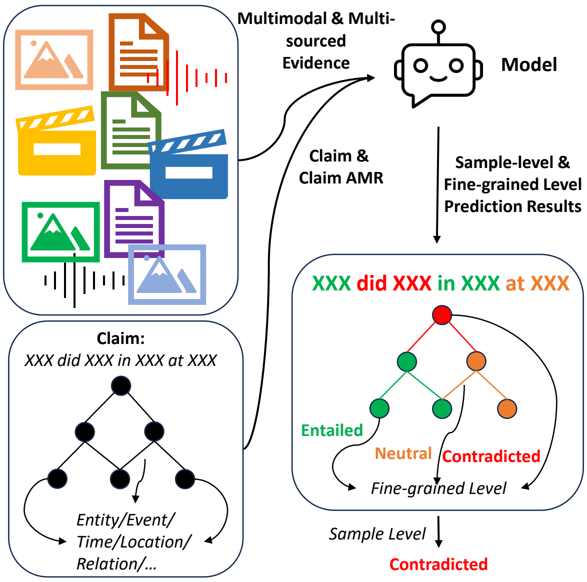

Multimodal & Multi-sourced Evidence
Model
Claim & Claim AMR
Sample-level & Fine-grained Level Prediction Results
Claim:
XXX did XXX in XXX at XXX
Entity/Event/Time/Location/Relation/...
XXX did XXX in XXX at XXX
Entailed
Neutral
Contradicted
Fine-grained Level
Sample Level
Contradicted

Figure 1: We predict the logical relationship of each piece of a claim (e.g. nodes=objects, tuples=relations) with a set of multimedia evidence. We also contribute a new benchmark and baseline model for this challenging task requiring cross-document, cross-modal reasoning.

1 Introduction

Misinformation poses serious societal risks by perpetuating narratives that incite fear, sow discord, and affect public health and safety (Geoghegan et al., 2020; Treen et al., 2020). Despite significant efforts towards developing automated fact-checking techniques (Yao et al., 2023a; Nasir et al., 2021; Karimi and Tang, 2019), existing methods face several limitations. First, real-world claims may include assertions that require consulting multiple documents and modalities to verify or refute the claim. Existing approaches either assume a single document setting (Fung et al., 2021; Thomas et al., 2022) or perform retrieval across documents to obtain relevant evidence, which is then treated as a single document (Yao et al., 2023a), potentially losing important surrounding context. Secondly, some methods only predict when claims conflict with relevant knowledge but ignore ambiguous cases where no supporting or refuting information is available (Wu et al., 2022; Xuan et al., 2024). Lastly, most of the existing methods fail to provide the fine-grained analysis needed for users to understand what is inconsistent in a claim or to make revisions to be more factual (Wu et al., 2022; Yao et al., 2023a; Xuan et al., 2024). Simply flagging an entire claim as false without pinpointing the specific inaccurate parts provides limited utility. In contrast, we propose an approach for predicting the logical relationship of each piece of a claim with respect to a set of multimodal sources. We perform a semantic dissection of claims into semantic pieces and leverage a hierarchical transformer that operates across multimedia documents to make fine-grained predictions. As illustrated in Figure 1

Our model ingests the claim along with associated multimedia, preserving the context. It

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 22270–22293
November 12-16, 2024 ©2024 Association for Computational Linguistics

then fuses the cross-document representations into a graph initialized with the claim's Abstract Meaning Representation (AMR) (Banarescu et al., 2013). Entailment relations are then predicted for each node (e.g., entities, actions) and tuple (e.g., relations) within the graph. Because no prior work has explored making fine-grained claim predictions from a set of multimodal documents, we also introduce a new dataset of claims that contains fine-grained labels for this task called M^3DC (**M**ulti**M**odal **M**ulti-**D**ocument **C**laims). We build our dataset on top of the NewsStories(Tan et al., 2022) dataset, which includes sets of news articles, images, and videos across multiple topics. We retrieve textual, visual, and audio data from each set to build a robust multimodal multidocument knowledge graph for each set of related documents. Next, we develop a claim synthesis method in order to generate claims that require multisource knowledge to verify, which uses a fine-grained claim manipulator model to generate claims manipulated at the sub-claim level.

Our major contributions are as follows:

- We introduce the novel task of performing fine-grained entailment of a textual claim with a set of multimodal documents.
- We propose a novel data synthesis technique for generating fine-grained labeled claims requiring multimodal multisource knowledge to verify using a graph traversal and fine-grained claim manipulator model.
- We contribute a large benchmark of fine-grained labeled claims created using our technique. We also contribute a small number of claims densely annotated by experts.
- We introduce a new hierarchical transformer model baseline designed for the task of fine-grained claim analysis over multiple sources.
- We conduct qualitative and quantitative experiments to evaluate the performance of our proposed method on our new benchmark dataset, as well as an existing benchmark dataset.

2 Related Works

Multimodal Misinformation Datasets. Previous works have studied misinformation using a variety of multimodal datasets (Cheema et al., 2022a; Nakamura et al., 2020; Abdelnabi et al., 2022; Gupta et al., 2022; Hu et al., 2023; Fung et al., 2021; Thomas et al., 2022; Yao et al., 2023b). However, most predict claims as either true or false, focusing on whether the entire claim is entailed or contradicted by the premise (Cheema et al., 2022a; Nakamura et al., 2020; Abdelnabi et al., 2022; Gupta et al., 2022; Hu et al., 2023; Fung et al., 2021). This binary approach fails to account for cases where the truthfulness of a claim cannot be determined. In such instances, many previous works treat these claims as contradicted, which is not accurate, as the veracity of the claim cannot be verified (Thomas et al., 2022; Yao et al., 2023b). Furthermore, most of the datasets used in these studies only provide evidence from a single source (e.g., a single news article) (Cheema et al., 2022a; Nakamura et al., 2020; Abdelnabi et al., 2022; Gupta et al., 2022; Hu et al., 2023; Fung et al., 2021; Thomas et al., 2022; Yao et al., 2023b), which can bias the judgment or limit the assessment. Relying on a single source of evidence may not capture potential inconsistencies or conflicting information that could arise when considering multiple sources (Wu et al., 2022).

Multimodal Misinformation Detection. Recent multimodal misinformation detection approaches (Yao et al., 2023a; Tan et al., 2020; Singh et al., 2021; Fung et al., 2021; Abdelnabi et al., 2022) are capable of relying on multimodal evidence for claim verification. However, most of these works still focus on claim-level binary predictions, i.e., whether the claim is entailed or contradicted (Tan et al., 2020; Singh et al., 2021; Fung et al., 2021; Abdelnabi et al., 2022), and the proposed models can only focus on a single source of evidence (Yao et al., 2023a; Tan et al., 2020; Singh et al., 2021; Fung et al., 2021; Abdelnabi et al., 2022). To address this limitation, some prior work attempts to not only predict the claim's label, but also provide explanations (Thomas et al., 2022; Yao et al., 2023b). MOCHEG (Yao et al., 2023a) leverages a text generator to generate explanations explaining a classifier's entailed, neutral, or contradicted prediction results, but there is no guarantee the produced explanations are what the classifier relied on. InfoSurgeon (Fung et al., 2021) extracts a multimodal knowledge graph (KG) for generated text detection and identifies specific internal inconsistencies within it. Similarly, Wu et al. (2022) propose a GNN-based model for detecting fine-grained inconsistencies in text-only documents using information extraction (IE) (Lin et al., 2020). Unlike these approaches, we perform full fine-grained entailment across a collection of open world multimedia documents (e.g. video, audio, text, and images) and are not limited to a specific IE ontology as is (Wu et al., 2022; Fung et al., 2021) or simple purely visual claims as is (Thomas et al., 2022).

3 Approach

We develop a model to predict sample-level and fine-grained entailment labels for a claim and its multimedia evidence (premise). The sample-level label (entailment, neutral, or contradiction) indicates the overall claim's relationship with the premise. Fine-grained labels detail entailment relationships for specific claim parts, such as entities and events, based on the claim's AMR tree. We first describe our methodology for constructing M^3DC. We then detail our model architecture, which makes fine-grained claim predictions using multimodal multidocument sets of evidence.

3.1 M^{3}DC Dataset

We first introduce our data synthesis approach for constructing a dataset with claims containing fine-grained labels that require multimodal and multi-source knowledge to verify. Our dataset builds upon NewsStories (Tan et al., 2022), a collection of news clusters with articles and videos. We begin by crawling the data and removing news that is no longer publicly accessible or has been taken down. For each news cluster, we construct a knowledge graph (KG) combining textual and non-textual data based on AMR trees (Banarescu et al., 2013) generated from news documents. This cross-document, cross-media representation allows us to synthesize claims by linking information from the graph. We then introduce a claim manipulator model that generates claims with varying degrees of truthfulness by traversing the AMR-based KG and introducing controlled perturbations. To obtain fine-grained labels, we employ a text-only model that assigns entailment labels (e.g., entailment, contradiction, neutral) to individual AMR nodes and tuples with the ground truth associated knowledge from the KG. Using this approach, we synthesize a dataset of about 400K claims across over 10,000 topics, requiring multimodal and multi-document knowledge for verification. The overall process is shown in Figure 3.

3.1.1 Knowledge Graph Construction

For each news cluster, we extract knowledge into a set of AMR trees using Structured-BART (Zhou et al., 2021) with sentences coming from the news document, visual captions generated from LLaVA-1.5 (Liu et al., 2023) and Video-LLaVA (Lin et al., 2023) and audio summaries from Qwen-Audio(Chu et al., 2023). Then, we connect nodes from AMR trees using co-reference resolution from CDLM (Caciularu et al., 2021) and F-coref (Otmazgin et al., 2022) in order to link within-document and cross-document entities or events. The overall process is illustrated in Figure 2.

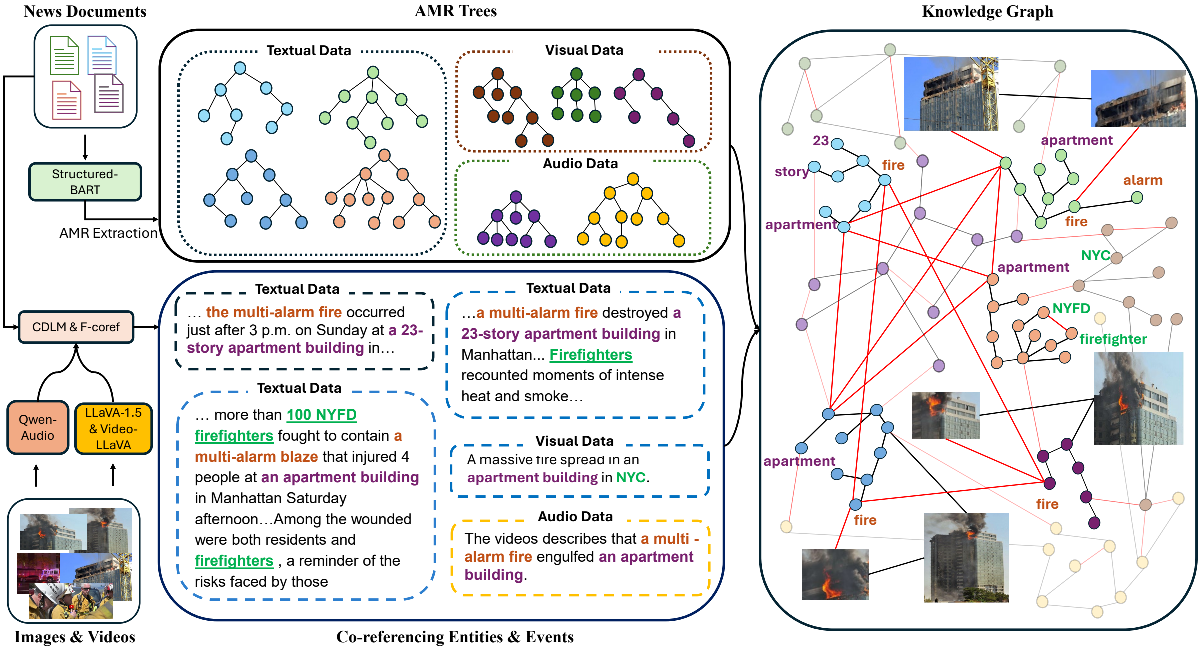
3.1.2 Claim Generation

To generate claims that require multimodal, multi-document evidence from the constructed KGs, we developed a Depth-First Search (DFS) based graph traversal method that selects Knowledge Elements (KEs) from multiple sources from the constructed KG. For a given KG and starting node (i.e. an AMR predicate node), the traversal algorithm traverses surrounding nodes until another predicate node is reached. We encourage the algorithm to follow co-reference edges to incorporate knowledge across documents and modalities. The traversal algorithm outputs KEs (AMR triples) rooted at a predicate, which is then used to generate a complete claim sentence containing the information from the traversed nodes and edges through AMRBART (Bai et al., 2022). Given that these generated claims are directly generated from the KG, all resulting claims are inherently entailed by this approach. This approach ensures that the resulting claims rephrase evidence from different articles and modalities, requiring the model to reason across sources to perform fine-grained verification.

3.1.3 Claim Manipulation

Since the claims generated directly from the KGs are inherently entailed, we introduce a claim manipulator model to generate diverse claims with varying logical relationships to the evidence in the KG. The claim manipulator takes as input the claim, relevant evidence from the KG (which may be multimodal), and a desired logical label (entailed, neutral, or contradicted). The goal is to manipulate an entailed claim so that the claim's logical relation matches the input. To train the manipulator, we employ reinforcement learning, where a model is optimized to maximize the scores provided by a reward model that offers evaluative feedback.

Denoting the original claim as c, derived from the KG, and the modified claim as $\hat{c}$ produced by the manipulator M, with y representing the logical label from $\mathbb{Y} = \{"entailed", "neutral", "contradicted"\}$,



Textual Data

… the multi-alarm fire occurred just after 3 p.m. on Sunday at a 23-story apartment building in…

Textual Data

…a multi-alarm fire destroyed a 23-story apartment building in Manhattan... Firefighters recounted moments of intense heat and smoke…

Textual Data

… more than 100 NYFD firefighters fought to contain a multi-alarm blaze that injured 4 people at an apartment building in Manhattan Saturday afternoon…Among the wounded were both residents and firefighters , a reminder of the risks faced by those

Visual Data

A massive fire spread in an apartment building in NYC.

Audio Data

The videos describes that a multi - alarm fire engulfed an apartment building.

Figure 2: Constructing a KG from a multimedia news cluster. AMR trees from different documents and modalities are linked to form a cross-document, cross-media KG. Co-reference links are shown in red.

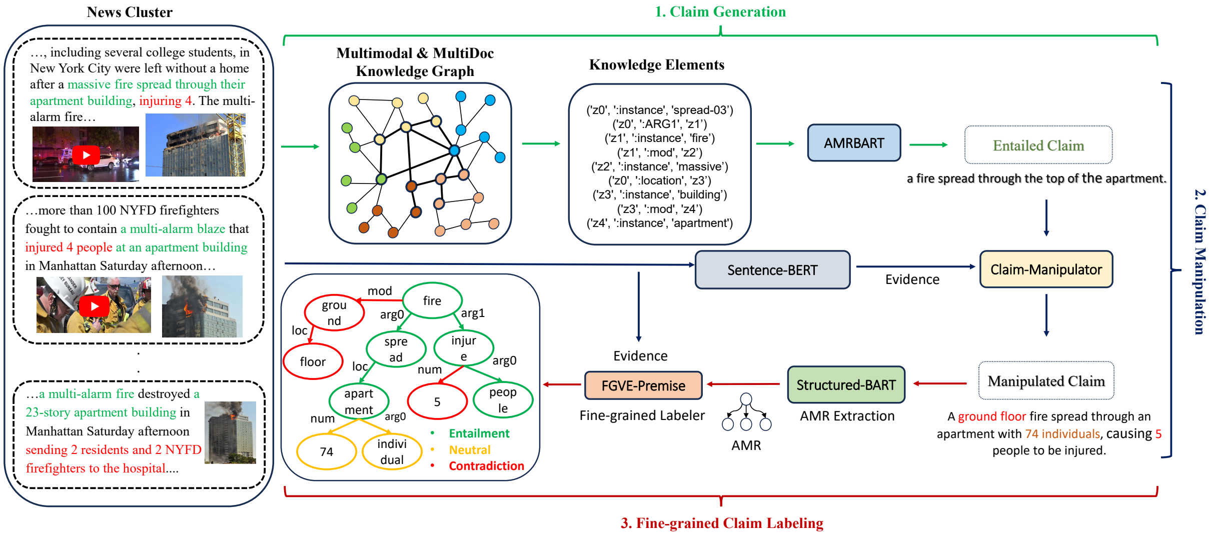
the goal of the claim manipulator is to generate a claim similar to the original claim c with the target logical label $\hat{y}$ given premise (evidence) p. We leverage Llama-2-13B (Touvron et al., 2023) to manipulate claims to correspond with the designated logical label $\hat{y}$ based on the given premise p. The premise consists of the top 10 most relevant evidence (expressed in text, i.e., using sentences from news articles and captions for image and video) related to c from Sentence-BERT(Reimers and Gurevych, 2019), the manipulator is fine-tuned using reinforcement learning to produce a claim $\hat{c}$ based on c. In this process, c and $\hat{c}$ are intended to be syntactically similar to each other. The claim manipulator can be formulated as $\hat{c} = M_\theta(p, c, \hat{y})$

To steer the manipulator towards generating claims that align with the target logical label $\hat{y}$ and similar to the original claim c syntactically, a reward model based on DeBERTAv3 (He et al., 2023) is trained to function as a critic using MNLI (Williams et al., 2018), Fever-NLI (Thorne et al., 2018), and ANLI (Nie et al., 2020). The reward model is trained for fine-grained entailment classification using the multi-instance and structural constraints from FGVE (Thomas et al., 2022). Critically, we enforce our target label constraint at both the fine-grained and sample levels within the graph. This approach ensures that the claim manipulator not only focuses on producing claims in a coarse-grained manner but also pays attention to fine-grained details. Specifically, the reward model's score is defined as the likelihood of the target label considering both the manipulated claim and the top 10 sentences most relevant to the original claim from the KG (serving as evidence):

$$\begin{aligned} & r(c, \hat{c}, \hat{y}) = P(\hat{y} \mid p, \hat{c}) - \\ & \left(\sum_{y_i \neq \hat{y}}^{|\mathbb{Y}|} P(y_i \mid p, \hat{c}) + \text{ROUGE}(c, \hat{c}) \right) \end{aligned} \tag{1}$$

where c, $\hat{c}$, $\hat{y}$, and p represent the original claim, the modified claim, the desired logical label for the claim, and the premise, respectively. The term $P(\hat{y} \mid p, \hat{c})$ is obtained from the trained fine-grained entailment classifier. The goal of this reward function is to ensure that the modified claim $\hat{c}$ not only matches the intended truthfulness label $\hat{y}$ but also remains similar to the original claim c as quantified by the ROUGE score.

We fine-tuned the claim manipulator with Proximal Policy Optimization (PPO) (Schulman et al., 2017) as our policy gradient method for reinforcement learning. PPO adds an additional term to the reward function, which imposes a penalty determined by the Kullback-Leibler (KL) divergence between the trained RL policy manipulator, π_ϕ^{PPO}, and the initial supervised manipulator π^{SFT}:



…, including several college students, in New York City were left without a home after a massive fire spread through their apartment building, injuring 4. The multi-alarm fire…

…more than 100 NYFD firefighters fought to contain a multi-alarm blaze that injured 4 people at an apartment building in Manhattan Saturday afternoon…

…a multi-alarm fire destroyed a 23-story apartment building in Manhattan Saturday afternoon sending 2 residents and 2 NYFD firefighters to the hospital....

('z0', ':instance', 'spread-03')
('z0', ':ARG1', 'z1')
('z1', ':instance', 'fire')
('z1', ':mod', 'z2')
('z2', ':instance', 'massive')
('z0', ':location', 'z3')
('z3', ':instance', 'building')
('z3', ':mod', 'z4')
('z4', ':instance', 'apartment')

a fire spread through the top of the apartment.

A ground floor fire spread through an apartment with 74 individuals, causing 5 people to be injured.

Figure 3: Claim generation pipeline. We create a knowledge graph from a set of media about an event. Our traversal algorithm selects the part of the KG highlighted in yellow to generate a (true) claim. To do so, we use the selected elements to translate the selected knowledge into a sentence. We then feed relevant evidence and the generated claim into our claim manipulator model. In this example, we ask our claim manipulator to generate a contradicted claim. The claim manipulator performs fine-grained manipulations, inserting both unverified (i.e. 74 individuals) and contradictory (i.e. 5 people injured) assertions. Because we know how the claim was manipulated at the knowledge-element level, we can use this as supervision to train our verification model.

$$r_{total} = r(\hat{c}, c, \hat{y}) - \eta KL(\pi_{\phi}^{PPO}(\hat{y}_t \mid p, \hat{c}), \pi^{SFT}(\hat{y}_t \mid p, \hat{c})) \tag{2}$$

where η represents the KL reward coefficient, which determines the magnitude of the KL penalty; we set it to 0.2 for our model. This coefficient functions as an entropy boost, enhancing exploration throughout the policy domain and urging the model to engage in a diverse set of actions rather than the one currently considered the best. In addition, it inhibits the policy from rapidly committing to a singular strategy, and this encourages outputs from the RL fine-tuned model to not deviate too far from the original model. After constructing the dataset with the claim manipulator, we employ Mixtral-8x7B (Jiang et al., 2024) using in-context learning to predict the logical label of the claims generated by the claim manipulator as a quality check; we discard those that do not align with the target labels. Finally, as a final quality check on our generated dataset, we assess the checkworthiness of claims using ClaimBuster (Arslan et al., 2020) to filter opinions or unimportant claims from our dataset. More details are covered in Appendix A.1.

3.2 Model Architecture

In this section, we present our model for predicting fine-grained entailment relations for claims given a set of trusted multimodal source materials. Figure 4 shows our model's architecture.

3.2.1 Multimodal Encoder

By design, our claims require reasoning across modalities and documents. We thus integrate all modalities into our model, preserving the original context in which the claim appeared. For textual content, we employ LongT5 (Guo et al., 2022) to encode the claims and sentences from documents and captions. For handling non-textual context (i.e. images, video, and audio), we use ImageBind (Girdhar et al., 2023). In addition to explicitly capturing how the information relates across documents and modalities, our model also ingests an embedding of the KG corresponding to each cluster. To learn our KG embedding, we instantiate our KG using a Graph Convolutional Network (GCN) and train it via a masked sequence prediction task. We randomly obscure nodes and edges within the KG and train a classifier to predict the masked pieces. After training, we extract KG embeddings for each cluster and feed them to our model. To bridge the various representation spaces, we add an additional linear layer for each modality's encoder.

The embeddings from different modalities, including textual content, non-textual context, and the knowledge graph (encoded by the GNN), are

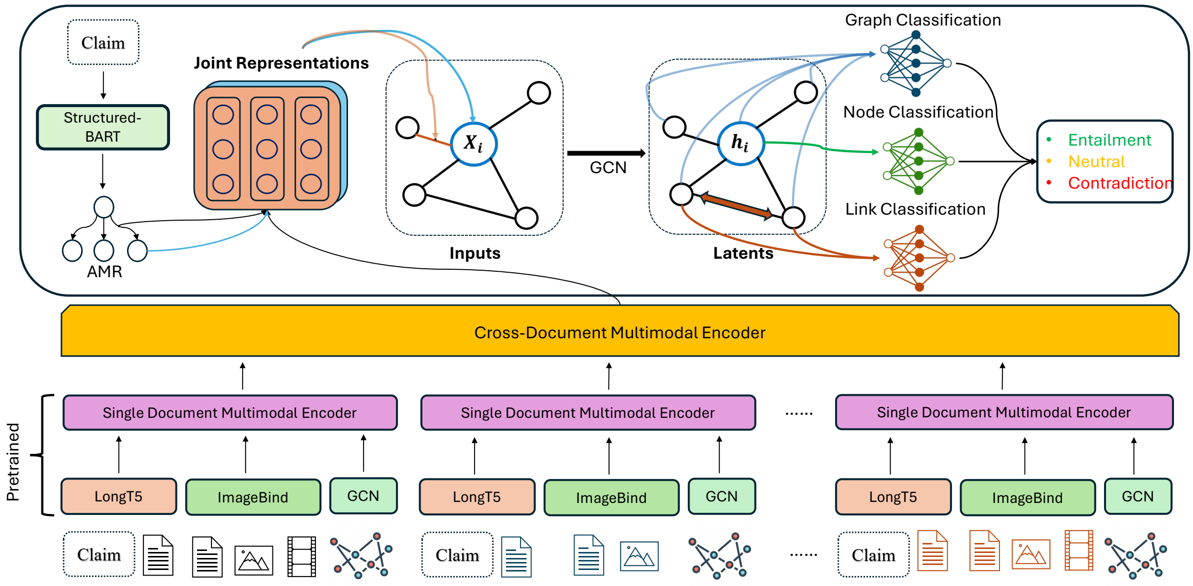


Figure 4: The model architecture. Each cluster, potentially containing multiple news articles, will have its content from various multimedia sources independently encoded and then merged to form a unified representation. This joint representation will serve as the initial state for every node within the GNN. Subsequently, labels at both the sample level and the fine-grained level can be derived by aggregating features from the nodes and edges of the GNN.

concatenated to form a comprehensive multimodal representation of the claim and its associated evidence. This concatenated embedding is then fed into LongT5 (Guo et al., 2022) for pretraining using the objective from Pegasus (Zhang et al., 2020). We identify the top 3 sentences inside the news documents that are most relevant to the claim c using ROUGE-F1, randomly choose one sentence and its adjacent sentence, and then mask them both. LongT5 is trained to generate the masked sentences based on the surrounding context and the multimodal embeddings.

3.2.2 Graph Convolutional Network

Our task requires predicting fine-grained entailment relationships between a claim and a set of multimedia source materials. To ensure each fine-grained element within the claim's AMR captures the context of the AMR structure in which it appears, we employ a two-layer GCN (Kipf and Welling, 2016) to learn contextual features of each node and tuple within the claim's AMR graph. Our GCN model is initialized with features aggregated from multiple single-document multimodal encoders and text embeddings from the claims's AMR. These features are contacted and represented as a joint representation. Specifically, we encode the AMR representation of claims and embeddings from multimedia news content via the GCN as follows: for each node i within the graph initialized from the joint representations, we define the feature aggregation mechanism by the equation:

$$h_i^{(l+1)} = \sum_{j \in \mathcal{N}(i) \cup \{i\}} \frac{1}{c_{ij}} h_j^{(l)} \tag{3}$$

where $h_i^{(l+1)}$ is the feature vector of node i at the subsequent layer $l+1$. The set $\mathcal{N}(i)$ includes the neighbors of node i, and c_{ij} is a normalization factor for the edge that connects nodes i and j.

For edge features, we extend our model to incorporate edge features alongside node features. This is achieved by incorporating edge attributes into the aggregation function, allowing the model to consider the characteristics of the connections between nodes. For an edge e_{ij} connecting nodes i and j, the edge features can be integrated as follows:

$$e_{ij}^{(l+1)} = \left[W_e^{(l)} h_i^{(l)} || W_e^{(l)} h_j^{(l)} \right] \tag{4}$$

where $e_{ij}^{(l+1)}$ represents the feature vector of edge e_{ij} at layer $l+1$, with $W_e^{(l)}$ being the weight matrix specific to edge features at layer l. This approach ensures that the model captures not only the node-level but also the edge-level semantic and structural information inherent in AMR graphs.

For graph-level (sample-level) classification, we aggregate the features of the entire graph with average pooling. Multiple MLP classifiers are then applied to make predictions for nodes, edges, and the

Datasets	#Samples	Source	#Topics	MultiModal	MultiDoc	Claim Verifications	Fine-grained Labels
Zlatkova et al. (2019)	1,233	Snopes, Reuters	<1500	✔	✗	✔	✗
Cheema et al. (2022b)	3,400	Twitter	<3,400	✔	✔	✔	✗
Nielsen and McConville (2022)	12,914	Twitter	26,048	✔	✔	✔	✗
Yao et al. (2023b)	15,601	Politifact, Snopes	<15,631	✔	✗	✔	✗
Nakov et al. (2021)	18,014	Twitter	<1,312	✗	✔	✗	✗
Ours	414,405	Multi-Source	15,000	✔	✔	✔	✔

Table 1: Comparison between different datasets in terms of multi-modality, multi documents, claim verification, and fine-grained labels. Ours is the largest one that supports fine-grained labels with multimodal document claim verification. No dataset provides fine-grained labels. †: Note that for datasets where the number of topics is not explicitly stated, we have estimated this based on the number of documents they contain.

Data	Train	Dev	Test
# Claims	372,935	41,440	30
Ave. # Tokens in Claim	162	178	158
# Documents	301,960	25,891	125
# Images	301,960	25,891	125
# Videos & Audios	70,042	4673	62
# ENT Labels	161,990	18,000	10
# NEU Labels	109,092	12,122	10
# CON Labels	101,853	11,318	10
# Documents / Images / Videos in Collection	327,976 / 327,976 / 74,777		

Table 2: Dataset statistics of $\mathbf{M}^3$D.

graph on the sample-level and fine-grained tasks. We train our model using cross-entropy loss with labels from the trained fine-grained entailment classifier in section 3.1.3.

4 Experiments

4.1 Multimodal MultiDocument Dataset

We compare our new dataset with others in Table 1. Our dataset contains fine-grained labels across 180,000 entailed claims, 121,224 neutral claims, and 113,181 contradicted claims. While existing datasets are topic-specific, our claims are highly detailed and topically diverse. We include more examples of the generated claims from our dataset in the appendix. Table 2 shows the detailed statistics for each split.

4.2 Testing Datasets and Baselines

We evaluate our model's entailment performance on two benchmarks: M^3DC and MOCHEG (Yao et al., 2023a). For both, we report F1 scores for entailment, neutral, and contradiction categories, as well as a macro-averaged F1 score at both the sample and fine-grained levels. For M^3DC, we compare model predictions with both human-annotated and synthetic labels. Our test set comprises 30 document sets, each annotated by six experts. The test set is balanced across 30 claims, with 10 each of entailment (E), neutral (N), and contradiction (C). These 30 claims were randomly selected from a pool of 15,000 news clusters in our dataset. The fine-grained data from these 30 claims includes an average of 54 nodes and 58 edges per claim, amounting to 3,360 annotated pieces in total. The distribution of human fine-grained labels is 52% E, 23% N, and 25% C, while our automated labels resulted in 43% E, 28% N, and 29% C. For MOCHEG, we follow the evaluation protocol specified in Yao et al. (2023a).

4.3 Quantitative Evaluation

Table 3 shows our model outperforming baselines on the M^3DC dataset, with similar results on synthetic and human-labeled data. This is critical, as it shows that the performance of our models on our human-annotated data tracks closely with the performance obtained on our large synthetic dataset, suggesting our synthetic dataset is a good evaluation benchmark for this task.

On the MOCHEG dataset (Table 4), our model outperforms other approaches in fine-grained predictions, despite being trained on a diverse news dataset, M^3DC, rather than MOCHEG. While LLaVA and MiniGPT-v2 achieve higher overall F1 scores for sample-level predictions, they struggle to identify neutral claims, which our model handles more effectively. The lower performance of our model at the sample level can be attributed to the MOCHEG dataset's lack of video and audio modalities and the different styles of text (Snopes vs News articles) compared to M^3DC. It is important to note that all the data from MOCHEG are based on articles from Politifact and Snopes. The content of these articles essentially consists of explanations about why the claim is considered entailed, neutral, or contradicted. We argue that this characteristic of the MOCHEG dataset may be the reason why LLaVA-1.5 and MiniGPT-v2 outperform our model at the sample level. These

Model	**Synthetic Labels**								**Human Labels**							
	Sample-level				*Fine-grained*				*Sample-level*				*Fine-grained*			
	E	N	C	All	E	N	C	All	E	N	C	All	E	N	C	All
FGVE (Thomas et al., 2022)	0.27	0.2	0.28	0.25	0.23	0.1	0.09	0.14	0.32	0.14	0.36	0.27	0.30	0.05	0.04	0.13
MOCHEG (Yao et al., 2023a)	0.32	0.14	0.36	0.27	0.28	0.13	0.32	0.24	0.37	0.18	0.41	0.32	0.35	**0.14**	**0.39**	0.29
LLaVA-1.5 (Liu et al., 2023)	0.57	0.0	0.33	0.30	**0.73**	0.0	0.14	0.29	0.67	0.0	0.43	0.37	**0.88**	0.0	0.13	0.33
MiniGPT-v2 (Chen et al., 2023)	0.50	0.0	0.43	0.31	0.56	0.0	0.24	0.27	0.62	0.0	**0.62**	0.41	0.54	0.0	0.09	0.21
Ours	**0.72**	**0.26**	**0.48**	**0.49**	0.65	**0.23**	**0.41**	**0.43**	**0.72**	**0.21**	0.59	**0.51**	0.68	0.1	**0.39**	**0.39**

Table 3: Results on our M^3DC benchmark. We report class-wise F1 scores (E: entailed, N: neutral, C: contradicted) and the overall F1 score (All).

Model	*Sample-level*				*Fine-grained*			
	E	N	C	All	E	N	C	All
FGVE (Thomas et al., 2022)	0.37	0.16	0.37	0.3	0.31	0.1	0.2	0.20
MOCHEG† (Yao et al., 2023a)	0.57	0.23	0.40	0.39	0.52	**0.21**	**0.36**	0.37
LLaVA-1.5 (Liu et al., 2023)	0.67	0.0	**0.93**	**0.53**	0.44	0.0	0.25	0.23
MiniGPT-v2 (Chen et al., 2023)	0.67	0.0	**0.93**	**0.53**	**0.71**	0.0	0.25	0.32
Ours	**0.69**	**0.25**	0.48	0.47	0.63	0.18	**0.36**	**0.39**

Table 4: Results on MOCHEG dataset (Yao et al., 2023a). *All labels are human labels in this benchmark.* We report class-wise F1 scores (E: entailed, N: neutral, C: contradicted) and the overall F1 score (All). †: Note that MOCHEG (Yao et al., 2023a) is also trained on this dataset, while *our method is applied zero-shot*.

language models are trained on large corpora, and when provided with Politifact and Snopes articles from MOCHEG, it becomes easier for them to determine the truthfulness of a claim by *simply analyzing the text*. In contrast, our model's strength lies in its ability to handle diverse modalities and make fine-grained predictions, making it more suitable for real-world scenarios where evidence may come indirectly from various sources and formats.

It is worth noting that both LLaVA-1.5 (Liu et al., 2023) and MiniGPT-v2 (Zhu et al., 2023) achieve 0% F1-scores on neutral cases. We found that even though both these models did predict neutral cases, for example, as the result from MiniGPT-v2 shown in Fig 5 they got them all wrong. This highlights the difficulty of accurately identifying neutral claims and the importance of developing models that can effectively handle such cases in real-world misinformation detection tasks.

4.4 Ablations

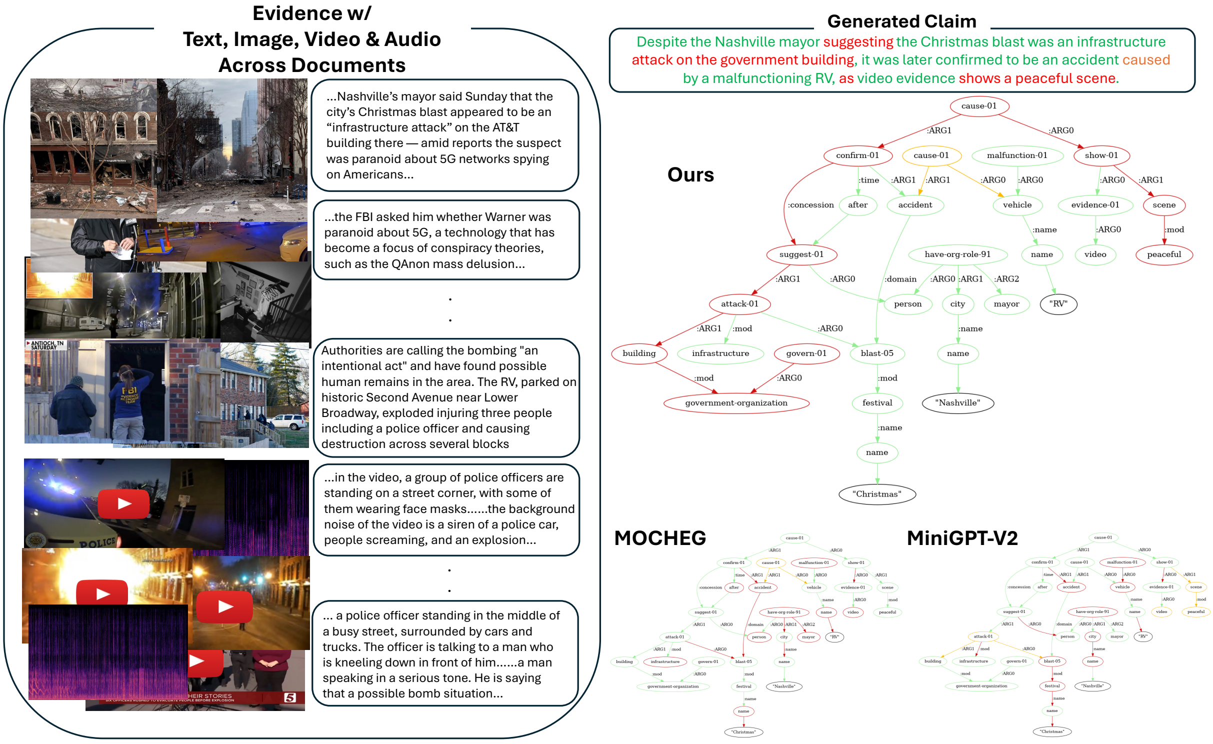
To demonstrate our model's capability in handling multimodal inputs, we conducted ablation studies with varying combinations of modalities, as outlined in Table 5. Considering that a substantial portion of the information in KGs is derived from the textual content of news articles, it was anticipated that the text modality would play a pivotal role in the model's inference process. Our results, however, indicate that including additional modalities, such as visual and audio, did not significantly enhance the model's performance. This observation suggests that the dominance of text-based claims in our dataset may lead the model to prioritize textual features, which are typically sufficient for classifying claims derived from textual information.

Model	*Sample-level*				*Fine-grained*			
	E	N	C	All	E	N	C	All
Ours w/ Text	0.69	0.25	0.43	0.46	0.61	0.15	0.34	0.37
Ours w/ Text + Image	0.71	**0.26**	0.42	0.46	0.63	0.18	0.36	0.39
Ours w/ Text + Image + Video	**0.72**	**0.26**	**0.48**	**0.49**	**0.65**	**0.23**	**0.41**	**0.43**
Ours w/ Text + Image + Video + Audio	0.70	0.24	0.47	0.47	0.63	0.21	**0.41**	0.42
Ours All w/o Text	0.42	0.02	0.29	0.24	0.37	0.01	0.23	0.20

Table 5: Ablation on M^3DC showing the impact of removing different modalities on our method.

4.5 Qualitative Results

We show qualitative results comparing our method with competitive baselines in Figure 5. We illustrate predictions on nodes and tuples by the color of the edges (green=entailed, yellow=neutral, red=contradiction). Node colors indicate node predictions, while edge colors represent tuple predictions. We perform fine-grained claim verification for the claim "Despite the Nashville mayor suggesting the Christmas blast was an infrastructure attack on the government building, it was later confirmed to be an accident caused by a malfunctioning RV, as video evidence shows a peaceful scene." In actuality, the blast happened on an AT&T building instead of a government building, so this portion of the claim is shown in red (as being contradicted by certain media sources). Moreover, the audio



...Nashville's mayor said Sunday that the city's Christmas blast appeared to be an "infrastructure attack" on the AT&T building there — amid reports the suspect was paranoid about 5G networks spying on Americans...

...the FBI asked him whether Warner was paranoid about 5G, a technology that has become a focus of conspiracy theories, such as the QAnon mass delusion...

Authorities are calling the bombing "an intentional act" and have found possible human remains in the area. The RV, parked on historic Second Avenue near Lower Broadway, exploded injuring three people including a police officer and causing destruction across several blocks

...in the video, a group of police officers are standing on a street corner, with some of them wearing face masks......the background noise of the video is a siren of a police car, people screaming, and an explosion...

... a police officer standing in the middle of a busy street, surrounded by cars and trucks. The officer is talking to a man who is kneeling down in front of him......a man speaking in a serious tone. He is saying that a possible bomb situation...

Despite the Nashville mayor suggesting the Christmas blast was an infrastructure attack on the government building, it was later confirmed to be an accident caused by a malfunctioning RV, as video evidence shows a peaceful scene.

Figure 5: Qualitative results comparing our method's fine-grained predictions with those obtained from other baselines. We include additional results in our supplementary materials.

evidence suggests that the video contains background noise with police sirens and people screaming, which contradicts the claim and is pointed out in the prediction results. We observe that our method identifies the correct portion of the claim as being contradicted by the evidence, while baselines tend to make more random predictions throughout the graph. Our model is able to produce correct results, compared to the results from MOCHEG (Yao et al., 2023b) and MiniGPT-V2 (Zhu et al., 2023), the models not only provide incorrect results but also fail to maintain the necessary structural constraints (Thomas et al., 2022) needed for explaining the truthfulness of the claim in fine-grained detail.

5 Conclusion

We address the challenge of predicting the logical consistency of claims with multimodal sources. Our method analyzes claims within a multimodal multidocument context, including text, visual content, and audio. Our method is able to reason in a fine-grained manner over complex information across media and modalities. We further introduce a dataset, M^3DC, created through a unique synthesis technique that produces claims requiring cross-document, cross-media reasoning for verification. Our contributions aim to mitigate the impact of misinformation and enhance the reliability of automated fact-checking systems, thus supporting informed decision-making and fostering a factually accurate public dialogue.

6 Acknowledgements

We acknowledge Advanced Research Computing at Virginia Tech for providing computational resources and technical support that have contributed to the results reported within this paper. We also thank all reviewers for their comments, which helped improve the paper.

7 Limitations

While our proposed approach for constructing a fact-checking dataset with fine-grained labels integrates multimodal and multi-document data, there are still several limitations that need to be addressed in future research. One of the main limitations is that the visual evidence in our dataset consists of grounding captions generated from images and video frames, resulting in a heavy reliance on textual data rather than other modalities. Given the nature of our dataset, which primarily consists of news documents where textual evidence dominates over other modalities, it's expected that the constructed dataset and the resulting model focus more on textual input, including the generated claims and information needed for reasoning.

Another limitation is that our model relies on the underlying assumption that genuine news articles are consistent, trustworthy, and complementary. However, there is a possibility that articles from the same news cluster can contain inconsistent information. For example, one article could report that there were nine people at the scene, while an image in another article only shows seven people. Moreover, certain types of human-written fake news documents, such as conspiracy theories, tend to be closely related and convey highly similar information due to shared biases or the intent to manipulate readers in a specific way. These issues of inconsistent information and similarity among fake news articles may limit the performance of our proposed system when applied to real-world data.

To address these limitations, future work could focus on the following areas: (1) incorporating more diverse modalities, such as raw visual and audio data, into the KG and the resulting dataset to reduce the reliance on textual data; (2) integrating commonsense reasoning techniques into the model to better capture complex contradictions and improve the system's ability to identify inconsistency and misinformation; (3) exploring alternative approaches that do not rely solely on the assumption of consistency among genuine news articles, thus improving the system's robustness when dealing with real-world fake news.

By addressing these limitations and exploring new research directions, we aim to enhance the performance and applicability of our proposed model in real-world scenarios, ultimately contributing to the fight against the spread of misinformation. We publicly release our multimodal, multi-document dataset and the proposed model implementation to foster further research in this area.

8 Ethical Considerations

In this work, our primary objective is to advance the state-of-the-art in fact-checking by analyzing multiple multimedia documents on the same topic. To achieve this goal, we have constructed a new benchmark dataset using the proposed methodology and developed a detector capable of determining the truthfulness of a given claim. To facilitate future research and benefit the community, we the constructed dataset and detector codes available, serving as a valuable reference for researchers and practitioners in the field.

However, we acknowledge that our work, like any research involving text generation, carries the risk of being misused to produce false information with the intent to mislead or manipulate readers. We want to clarify that the dataset and model we constructed do not contain true claims but rather claims generated from models. The dataset and model are intended solely for research purposes and should not be used to suppress opinions or make misjudgments. We strongly emphasize the importance of responsible and ethical use of these resources in the pursuit of advancing fact-checking techniques.

References

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

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Fatma Arslan, Naeemul Hassan, Chengkai Li, and Mark Tremayne. 2020. A Benchmark Dataset of Check-worthy Factual Claims. In *14th International AAAI Conference on Web and Social Media*. AAAI.

Xuefeng Bai, Yulong Chen, and Yue Zhang. 2022. Graph pre-training for AMR parsing and generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6001–6015, Dublin, Ireland. Association for Computational Linguistics.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse*, pages 178–186.

Avi Caciularu, Arman Cohan, Iz Beltagy, Matthew Peters, Arie Cattan, and Ido Dagan. 2021. CDLM: Cross-document language modeling. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2648–2662, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Gullal Singh Cheema, Sherzod Hakimov, Abdul Sittar, Eric Müller-Budack, Christian Otto, and Ralph Ewerth. 2022a. MM-claims: A dataset for multimodal claim detection in social media. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 962–979, Seattle, United States. Association for Computational Linguistics.

Gullal Singh Cheema, Sherzod Hakimov, Abdul Sittar, Eric Müller-Budack, Christian Otto, and Ralph Ewerth. 2022b. MM-claims: A dataset for multimodal claim detection in social media. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 962–979, Seattle, United States. Association for Computational Linguistics.

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. 2023. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *Preprint*, arXiv:2311.07919.

Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. 2023. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19358–19369.

Yi Fung, Christopher Thomas, Revanth Gangi Reddy, Sandeep Polisetty, Heng Ji, Shih-Fu Chang, Kathleen McKeown, Mohit Bansal, and Avirup Sil. 2021. Infosurgeon: Cross-media fine-grained information consistency checking for fake news detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1683–1698.

Sarah Geoghegan, Kevin P O'callaghan, and Paul A Offit. 2020. Vaccine safety: myths and misinformation. *Frontiers in microbiology*, 11:372.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15180–15190.

Mandy Guo, Joshua Ainslie, David Uthus, Santiago Ontanon, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. 2022. LongT5: Efficient text-to-text transformer for long sequences. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 724–736, Seattle, United States. Association for Computational Linguistics.

Vipin Gupta, Rina Kumari, Nischal Ashok, Tirthankar Ghosal, and Asif Ekbal. 2022. MMM: An emotion and novelty-aware approach for multilingual multimodal misinformation detection. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 464–477, Online only. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

Xuming Hu, Zhijiang Guo, Junzhe Chen, Lijie Wen, and Philip S. Yu. 2023. Mr2: A benchmark for multimodal retrieval-augmented rumor detection in social media. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 2901–2912, New York, NY, USA. Association for Computing Machinery.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Hamid Karimi and Jiliang Tang. 2019. Learning hierarchical discourse-level structure for fake news detection. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3432–3442.

Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2023. Video-llava: Learning united visual representation by alignment before projection. *Preprint*, arXiv:2311.10122.

Ying Lin, Heng Ji, Fei Huang, and Lingfei Wu. 2020. A joint neural model for information extraction with global features. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7999–8009, Online. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916. Curran Associates, Inc.

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 Twelfth Language Resources and Evaluation Conference*, pages 6149–6157, Marseille, France. European Language Resources Association.

Preslav Nakov, Giovanni Da San Martino, Tamer Elsayed, Alberto Barrón-Cedeño, Rubén Míguez, Shaden Shaar, Firoj Alam, Fatima Haouari, Maram Hasanain, Watheq Mansour, Bayan Hamdan, Zien Sheikh Ali, Nikolay Babulkov, Alex Nikolov, Gautam Kishore Shahi, Julia Maria Struß, Thomas Mandl, Mucahid Kutlu, and Yavuz Selim Kartal. 2021. Overview of the clef–2021 checkthat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 264–291, Cham. Springer International Publishing.

Jamal Abdul Nasir, Osama Subhani Khan, and Iraklis Varlamis. 2021. Fake news detection: A hybrid cnn-rnn based deep learning approach. *International Journal of Information Management Data Insights*, 1(1):100007.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

Dan S. Nielsen and Ryan McConville. 2022. Mumin: A large-scale multilingual multimodal fact-checked misinformation social network dataset. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 3141–3153, New York, NY, USA. Association for Computing Machinery.

Shon Otmazgin, Arie Cattan, and Yoav Goldberg. 2022. F-coref: Fast, accurate and easy to use coreference resolution. *Preprint*, arXiv:2209.04280.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *Preprint*, arXiv:1707.06347.

Vivek K Singh, Isha Ghosh, and Darshan Sonagara. 2021. Detecting fake news stories via multimodal analysis. *Journal of the Association for Information Science and Technology*, 72(1):3–17.

Reuben Tan, Bryan Plummer, and Kate Saenko. 2020. Detecting cross-modal inconsistency to defend against neural fake news. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Reuben Tan, Bryan A. Plummer, Kate Saenko, JP Lewis, Avneesh Sud, and Thomas Leung. 2022. Newsstories: Illustrating articles with visual summaries. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVI*, page 644–661, Berlin, Heidelberg. Springer-Verlag.

Christopher Thomas, Yipeng Zhang, and Shih-Fu Chang. 2022. Fine-grained visual entailment. In *European Conference on Computer Vision*, pages 398–416. Springer.

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.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Kathie M d'I Treen, Hywel TP Williams, and Saffron J O'Neill. 2020. Online misinformation about climate change. *Wiley Interdisciplinary Reviews: Climate Change*, 11(5):e665.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. 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 1112–1122. Association for Computational Linguistics.

Xueqing Wu, Kung-Hsiang Huang, Yi Fung, and Heng Ji. 2022. Cross-document misinformation detection based on event graph reasoning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 543–558, Seattle, United States. Association for Computational Linguistics.

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. *Preprint*, arXiv:2402.11943.

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

Barry Menglong Yao, Aditya Shah, Lichao Sun, Jin-Hee Cho, and Lifu Huang. 2023b. 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 '23. ACM.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020. Pegasus: pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Jiawei Zhou, Tahira Naseem, Ramón Fernandez Astudillo, Young-Suk Lee, Radu Florian, and Salim Roukos. 2021. Structure-aware fine-tuning of sequence-to-sequence transformers for transition-based AMR parsing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6279–6290, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

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

A Appendix

A.1 Dataset Analysis

In this section, we present additional details about our dataset, M^3DC. To demonstrate that the claims in our dataset do not rely solely on textual data, we provide examples in Figures 6 and 7 that incorporate information from images and videos. Figure 6 showcases claims generated from image evidence selected from the KG and claims derived from knowledge elements that co-reference the visual content. This approach ensures that the generated claims contain a degree of visual information. These claims are then modified by the claim manipulator to integrate data from different modalities and documents. As a result, the final claims not only reflect the representative visual content but also potentially include the underlying context behind the image or related information from the news articles. By incorporating visual evidence and manipulating claims to integrate multi-modal data, our dataset presents a diverse set of claims that require both textual and visual understanding for verification. This highlights the importance of considering information from various modalities when assessing the veracity of claims in real-world scenarios.

A.2 Qualitative Results

To provide insight into the dataset and the results from our model, we provide additional examples. From Figure 8, 9 and 10, we show a random selection of the M^3DC dataset and the results from our model, respectively. According to the results shown in the figures, it is evident that the majority of the generated claims require detailed evidence to be properly reasoned about. Furthermore, the results demonstrate that our model is able to accurately reason about these claims, as most of the model's outputs are correct when compared to the evidence provided by the documents. This suggests that our model is capable of effectively utilizing the available evidence to make accurate predictions, even when the claims are complex and require careful consideration of multiple pieces of information.

The results presented in Table 3 indicate that our model performs similarly when evaluated using synthetic labels and human labels. To quantify this alignment, we calculated the R-score between the synthetic labels and human labels. This analysis provides insight into how closely our model's judgments match those of human evaluators. We conducted the R-score evaluation at both the sample-level and fine-grained level. The evaluation included the F1-scores derived from the entailment, neutral, and contradiction categories. The R-scores obtained were 0.95 at the sample-level and 0.99 at the fine-grained level. These high R-scores demonstrate that our model's performance is highly consistent with human performance. Consequently, these findings suggest that our model can reliably assist or potentially replace human evaluators in this context.

Despite the promising results, it is important to note that the majority of the generated claims do not rely heavily on visual data. This can be attributed to the nature of news articles, where most of the information is conveyed through textual content, and visual data may not provide a significant amount of additional evidence, as shown in the provided examples. Consequently, the performance of our model on this dataset may not fully showcase its ability to reason about claims that are more visually-centric. To address this limitation and further evaluate the capabilities of our model, future studies could explore its performance on datasets that place a greater emphasis on visual information, such as the Flicker dataset. By testing our model on a more visually centric dataset.

A.3 Human Annotations and Statistics

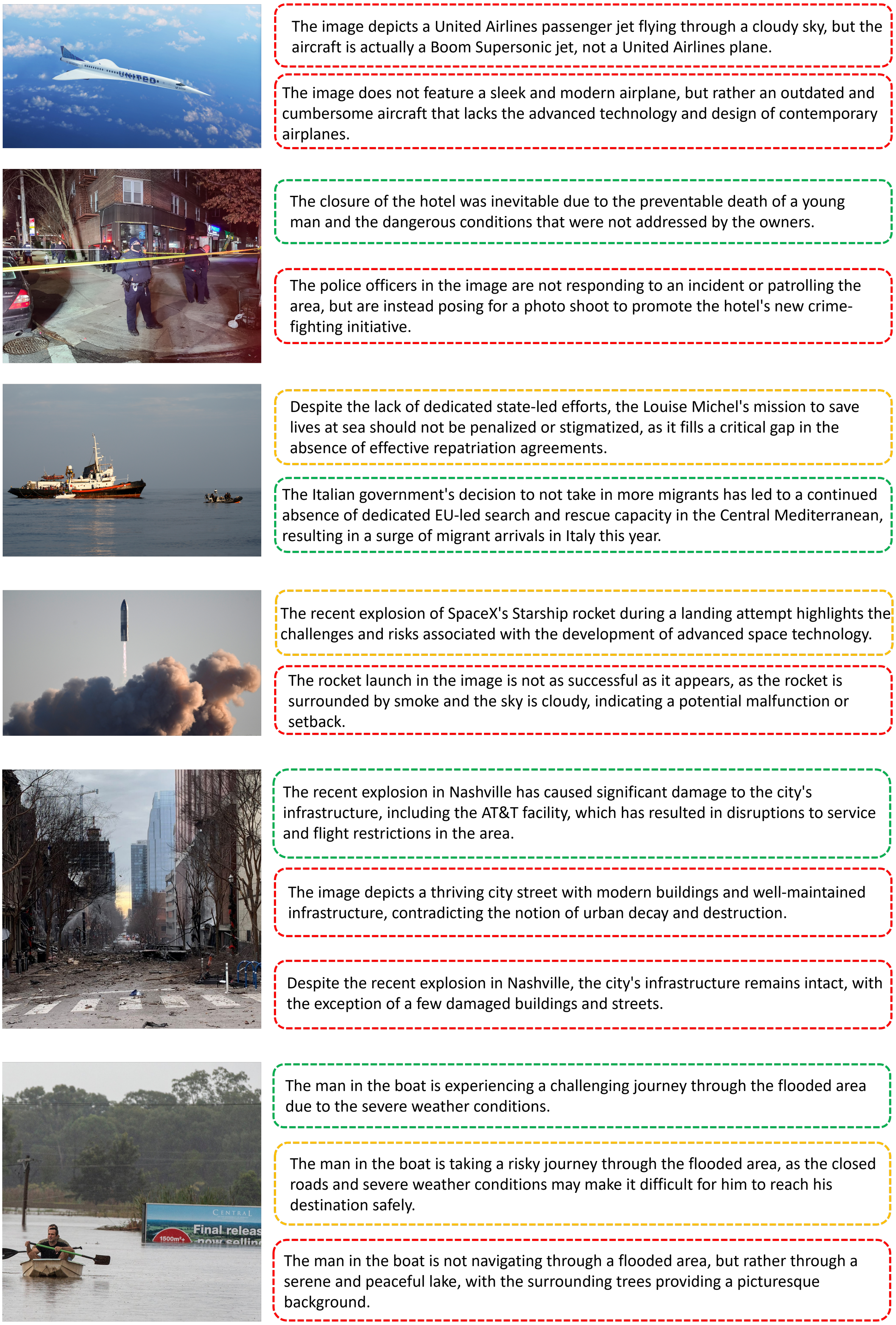
A.3.1 Annotation Details

In Table 3, we investigate the results of our model on human-labeled data to evaluate the performance of human annotators. To measure the inter-annotator agreement (IAA), we employ two annotators for each news cluster, with thirty different news clusters in total, who are responsible for labeling both the sample-level and fine-grained labels.

The inter-annotator agreement can be defined using the following formula:

$$IAA = \frac{\text{Number of samples with matching labels}}{\text{number of samples}} \tag{5}$$

Although our human-labeled dataset contains only 30 samples, annotating each claim derived from a news cluster can be a time-consuming process, taking anywhere from 30 to 60 minutes, depending on the number of news documents in each cluster. This is because many of the claims in our dataset rely on small details scattered across multiple news documents to determine the logical label at the sample level, which can be challenging even at the fine-grained level.



The image depicts a United Airlines passenger jet flying through a cloudy sky, but the aircraft is actually a Boom Supersonic jet, not a United Airlines plane.

The image does not feature a sleek and modern airplane, but rather an outdated and cumbersome aircraft that lacks the advanced technology and design of contemporary airplanes.

The closure of the hotel was inevitable due to the preventable death of a young man and the dangerous conditions that were not addressed by the owners.

The police officers in the image are not responding to an incident or patrolling the area, but are instead posing for a photo shoot to promote the hotel's new crime-fighting initiative.

Despite the lack of dedicated state-led efforts, the Louise Michel's mission to save lives at sea should not be penalized or stigmatized, as it fills a critical gap in the absence of effective repatriation agreements.

The Italian government's decision to not take in more migrants has led to a continued absence of dedicated EU-led search and rescue capacity in the Central Mediterranean, resulting in a surge of migrant arrivals in Italy this year.

The recent explosion of SpaceX's Starship rocket during a landing attempt highlights the challenges and risks associated with the development of advanced space technology.

The rocket launch in the image is not as successful as it appears, as the rocket is surrounded by smoke and the sky is cloudy, indicating a potential malfunction or setback.

The recent explosion in Nashville has caused significant damage to the city's infrastructure, including the AT&T facility, which has resulted in disruptions to service and flight restrictions in the area.

The image depicts a thriving city street with modern buildings and well-maintained infrastructure, contradicting the notion of urban decay and destruction.

Despite the recent explosion in Nashville, the city's infrastructure remains intact, with the exception of a few damaged buildings and streets.

The man in the boat is experiencing a challenging journey through the flooded area due to the severe weather conditions.

The man in the boat is taking a risky journey through the flooded area, as the closed roads and severe weather conditions may make it difficult for him to reach his destination safely.

The man in the boat is not navigating through a flooded area, but rather through a serene and peaceful lake, with the surrounding trees providing a picturesque background.

Figure 6: Claims generated by our pipeline, with entailed, neutral, and contradicted claims denoted by green, orange, and red dashed lines, respectively. Claims based on image content are generated by selecting knowledge elements rooted in the image nodes of the KG. Then, the claim manipulator adjusts the claim based on the premise, allowing control over the degree of evidence provided by each modality. This enables the generation of claims that are highly related to the visual content or that require consideration of cross-modal evidence.

The witnesses' accounts of the explosion suggest that the impact of the blast was so severe that it caused the body of one of the victims to be thrown 500ft into a nearby lake, indicating that the explosion was of a devastating magnitude.

Despite the intense explosion at Wessex Water Services Ltd in Avonmouth, the building remained intact and there was no significant damage to the structure.

The chemical explosion in Bristol did not involve a water recycling tank, as previously believed, but rather a faulty pipe that caused the blast.

The explosion at the wastewater treatment plant in Bristol was not caused by an exploding chemical tank, but rather by a malfunctioning oxygen tank, according to a spokesperson for Wessex Water.

The road leading to the industrial area was reopened this afternoon, despite the ongoing rescue operations and the tragic loss of life

Figure 7: In this figure, we present additional claim examples from our dataset. While not all claims are entirely generated from the visual data, many can be verified by examining the visual content within the corresponding news cluster. This demonstrates that our dataset implicitly and explicitly contains multimodal claims, highlighting the importance of considering both textual and visual information for claim verification.

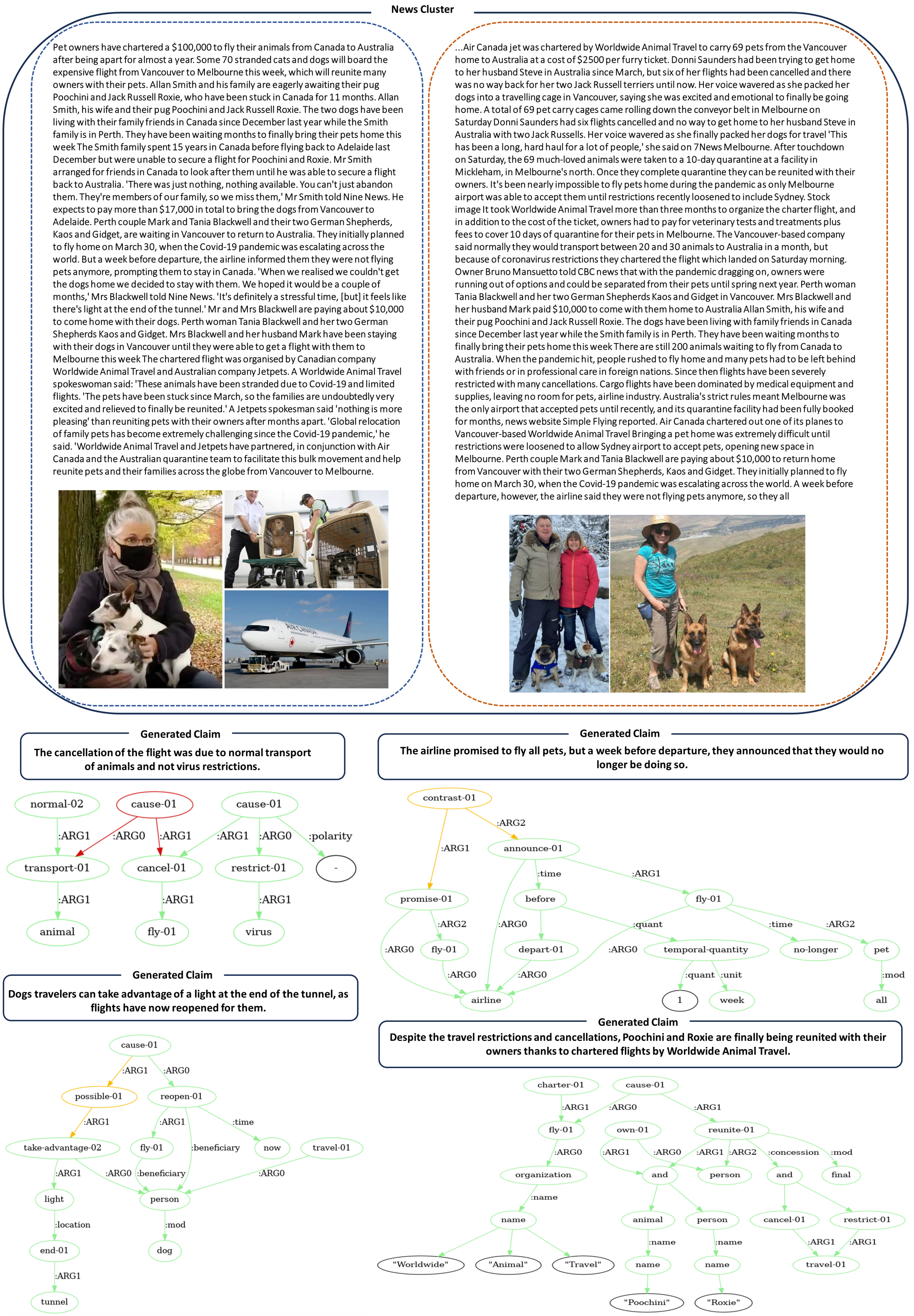
Our annotation interface powered by Label Studio is shown in Figures 11, 12, and 13. For each news cluster, the annotators are required to go through a series of documents with multiple images and videos to determine the logical label of the claim. As shown in Figure 11, our interface displays the textual and image content of the news cluster, where each cluster could contain up to five different news documents. In addition to the textual and image content, each news document could be linked to one or more corresponding videos, as shown in Figure 12. The annotators are required to review every video as well, and the number of videos could sometimes be up to a dozen. After reviewing all the available information, the annotators need to label the sample-level label first according to the given claim. For each AMR tuple, the annotators are required to annotate them separately, as shown in Figure 13, ensuring that all AMR tuples coming from the AMR tree are labeled. For example, in Figure 13, the annotators need to go through a series of different AMR tuples for just one claim and label all elements inside the AMR image.

To ensure the quality and consistency of the human-labeled dataset, we provide the annotators with guidelines and examples for each label category. The annotators are also given the opportunity to discuss and resolve any disagreements or ambiguities in the labeling process. This collaborative approach helps to maintain a high level of inter-annotator agreement and reduces the potential for individual biases or errors. Our annotators consist of all the authors of this paper, each of whom is an expert in AMR and familiar with its properties and constraints. During the labeling process, the annotators are required to perform fine-grained labeling while adhering to AMR properties and constraints.

To ensure the quality and consistency of the fine-grained labels, we have established a set of guidelines that the annotators must follow:

- Adherence to AMR properties: The annotators must have a deep understanding of the properties and constraints of AMR, such as the semantic roles of nodes and the relationships between them. This knowledge is crucial for accurate, fine-grained labeling.

- Consistency with sample-level labels: The fine-grained labels should be consistent with the sample-level labels. For example, if the sample-level label is neutral, at least one AMR



News Cluster

Pet owners have chartered a $100,000 to fly their animals from Canada to Australia after being apart for almost a year. Some 70 stranded cats and dogs will board the expensive flight from Vancouver to Melbourne this week, which will reunite many owners with their pets. Allan Smith and his family are eagerly awaiting their pug Poochini and Jack Russell Roxie, who have been stuck in Canada for 11 months. Allan Smith, his wife and their pug Poochini and Jack Russell Roxie. The two dogs have been living with their family friends in Canada since December last year while the Smith family is in Perth. They have been waiting months to finally bring their pets home this week The Smith family spent 15 years in Canada before flying back to Adelaide last December but were unable to secure a flight for Poochini and Roxie. Mr Smith arranged for friends in Canada to look after them until he was able to secure a flight back to Australia. 'There was just nothing, nothing available. You can't just abandon them. They're members of our family, so we miss them,' Mr Smith told Nine News. He expects to pay more than $17,000 in total to bring the dogs from Vancouver to Adelaide. Perth couple Mark and Tania Blackwell and their two German Shepherds, Kaos and Gidget, are waiting in Vancouver to return to Australia. They initially planned to fly home on March 30, when the Covid-19 pandemic was escalating across the world. But a week before departure, the airline informed them they were not flying pets anymore, prompting them to stay in Canada. 'When we realised we couldn't get the dogs home we decided to stay with them. We hoped it would be a couple of months,' Mrs Blackwell told Nine News. 'It's definitely a stressful time, [but] it feels like there's light at the end of the tunnel.' Mr and Mrs Blackwell are paying about $10,000 to come home with their dogs. Perth woman Tania Blackwell and her two German Shepherds Kaos and Gidget. Mrs Blackwell and her husband Mark have been staying with their dogs in Vancouver until they were able to get a flight with them to Melbourne this week The chartered flight was organised by Canadian company Worldwide Animal Travel and Australian company Jetpets. A Worldwide Animal Travel spokeswoman said: 'These animals have been stranded due to Covid-19 and limited flights. 'The pets have been stuck since March, so the families are undoubtedly very excited and relieved to finally be reunited.' A Jetpets spokesman said 'nothing is more pleasing' than reuniting pets with their owners after months apart. 'Global relocation of family pets has become extremely challenging since the Covid-19 pandemic,' he said. 'Worldwide Animal Travel and Jetpets have partnered, in conjunction with Air Canada and the Australian quarantine team to facilitate this bulk movement and help reunite pets and their families across the globe from Vancouver to Melbourne.

...Air Canada jet was chartered by Worldwide Animal Travel to carry 69 pets from the Vancouver home to Australia at a cost of $2500 per furry ticket. Donni Saunders had been trying to get home to her husband Steve in Australia since March, but six of her flights had been cancelled and there was no way back for her two Jack Russell terriers until now. Her voice wavered as she packed her dogs into a travelling cage in Vancouver, saying she was excited and emotional to finally be going home. A total of 69 pet carry cages came rolling down the conveyor belt in Melbourne on Saturday Donni Saunders had six flights cancelled and no way to get home to her husband Steve in Australia with two Jack Russells. Her voice wavered as she finally packed her dogs for travel 'This has been a long, hard haul for a lot of people,' she said on 7News Melbourne. After touchdown on Saturday, the 69 much-loved animals were taken to a 10-day quarantine at a facility in Mickleham, in Melbourne's north. Once they complete quarantine they can be reunited with their owners. It's been nearly impossible to fly pets home during the pandemic as only Melbourne airport was able to accept them until restrictions recently loosened to include Sydney. Stock image It took Worldwide Animal Travel more than three months to organize the charter flight, and in addition to the cost of the ticket, owners had to pay for veterinary tests and treatments plus fees to cover 10 days of quarantine for their pets in Melbourne. The Vancouver-based company said normally they would transport between 20 and 30 animals to Australia in a month, but because of coronavirus restrictions they chartered the flight which landed on Saturday morning. Owner Bruno Mansuetto told CBC news that with the pandemic dragging on, owners were running out of options and could be separated from their pets until spring next year. Perth woman Tania Blackwell and her two German Shepherds Kaos and Gidget in Vancouver. Mrs Blackwell and her husband Mark paid $10,000 to come with them home to Australia Allan Smith, his wife and their pug Poochini and Jack Russell Roxie. The dogs have been living with family friends in Canada since December last year while the Smith family is in Perth. They have been waiting months to finally bring their pets home this week There are still 200 animals waiting to fly from Canada to Australia. When the pandemic hit, people rushed to fly home and many pets had to be left behind with friends or in professional care in foreign nations. Since then flights have been severely restricted with many cancellations. Cargo flights have been dominated by medical equipment and supplies, leaving no room for pets, airline industry. Australia's strict rules meant Melbourne was the only airport that accepted pets until recently, and its quarantine facility had been fully booked for months, news website Simple Flying reported. Air Canada chartered out one of its planes to Vancouver-based Worldwide Animal Travel Bringing a pet home was extremely difficult until restrictions were loosened to allow Sydney airport to accept pets, opening new space in Melbourne. Perth couple Mark and Tania Blackwell are paying about $10,000 to return home from Vancouver with their two German Shepherds, Kaos and Gidget. They initially planned to fly home on March 30, when the Covid-19 pandemic was escalating across the world. A week before departure, however, the airline said they were not flying pets anymore, so they all

Generated Claim

The cancellation of the flight was due to normal transport of animals and not virus restrictions.

Generated Claim

The airline promised to fly all pets, but a week before departure, they announced that they would no longer be doing so.

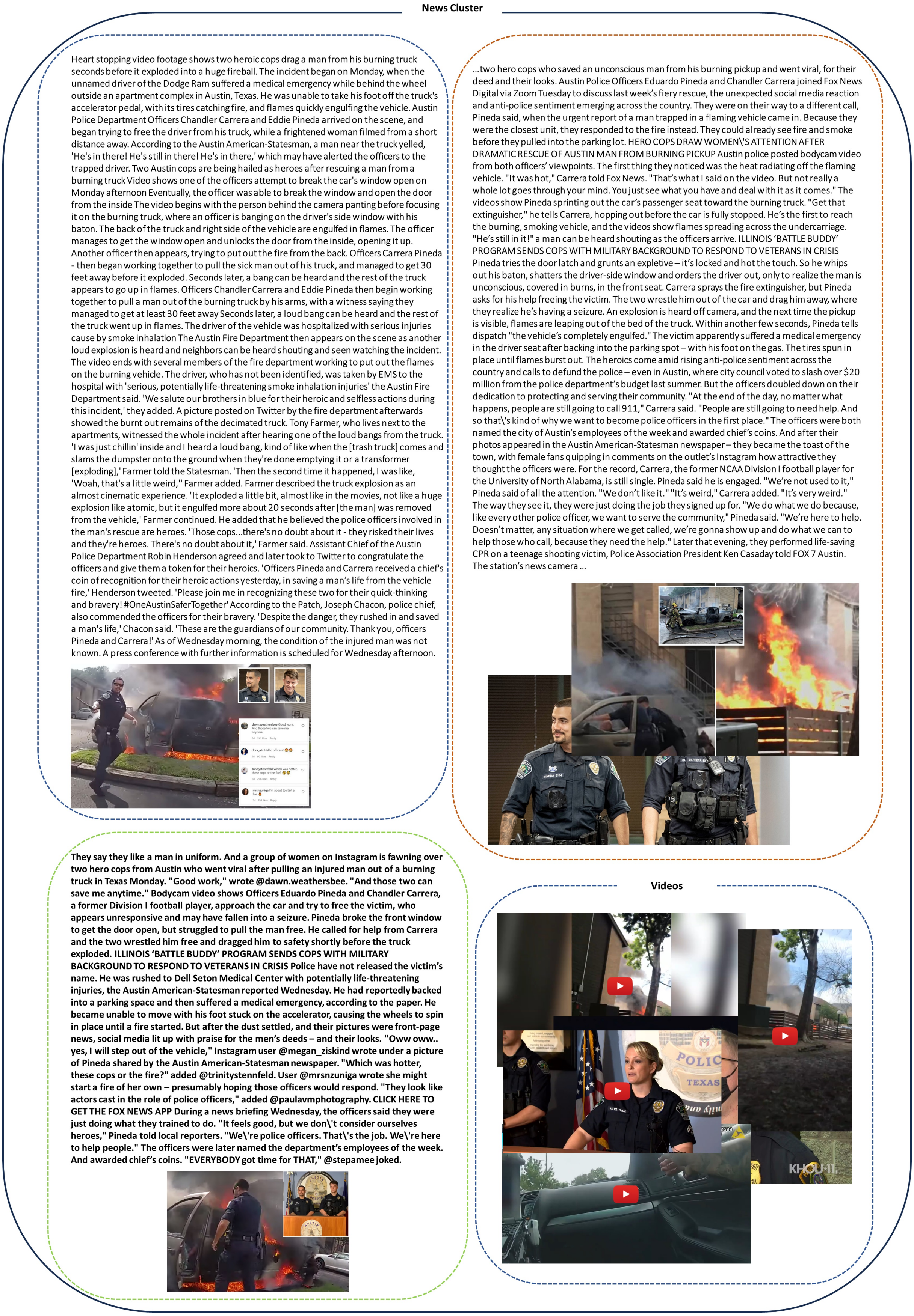
Generated Claim

Dogs travelers can take advantage of a light at the end of the tunnel, as flights have now reopened for them.

Generated Claim

Despite the travel restrictions and cancellations, Poochini and Roxie are finally being reunited with their owners thanks to chartered flights by Worldwide Animal Travel.

Figure 8: Results of the generated claims and the corresponding fine-grained level predictions. For instance, consider the generated claim in the top left corner. The ground truth label for this claim is "contradicted," as the flight cancellation was not caused by the normal transport of animals. Our model successfully detects this fact and assigns the correct fine-grained labels to the relevant parts of the claim.



Heart stopping video footage shows two heroic cops drag a man from his burning truck seconds before it exploded into a huge fireball. The incident began on Monday, when the unnamed driver of the Dodge Ram suffered a medical emergency while behind the wheel outside an apartment complex in Austin, Texas. He was unable to take his foot off the truck's accelerator pedal, with its tires catching fire, and flames quickly engulfing the vehicle. Austin Police Department Officers Chandler Carrera and Eddie Pineda arrived on the scene, and began trying to free the driver from his truck, while a frightened woman filmed from a short distance away. According to the Austin American-Statesman, a man near the truck yelled, 'He's in there! He's still in there! He's in there,' which may have alerted the officers to the trapped driver. Two Austin cops are being hailed as heroes after rescuing a man from a burning truck Video shows one of the officers attempt to break the car's window open on Monday afternoon Eventually, the officer was able to break the window and open the door from the inside The video begins with the person behind the camera panting before focusing it on the burning truck, where an officer is banging on the driver's side window with his baton. The back of the truck and right side of the vehicle are engulfed in flames. The officer manages to get the window open and unlocks the door from the inside, opening it up. Another officer then appears, trying to put out the fire from the back. Officers Carrera Pineda - then began working together to pull the sick man out of his truck, and managed to get 30 feet away before it exploded. Seconds later, a bang can be heard and the rest of the truck appears to go up in flames. Officers Chandler Carrera and Eddie Pineda then begin working together to pull a man out of the burning truck by his arms, with a witness saying they managed to get at least 30 feet away Seconds later, a loud bang can be heard and the rest of the truck went up in flames. The driver of the vehicle was hospitalized with serious injuries cause by smoke inhalation The Austin Fire Department then appears on the scene as another loud explosion is heard and neighbors can be heard shouting and seen watching the incident. The video ends with several members of the fire department working to put out the flames on the burning vehicle. The driver, who has not been identified, was taken by EMS to the hospital with 'serious, potentially life-threatening smoke inhalation injuries' the Austin Fire Department said. 'We salute our brothers in blue for their heroic and selfless actions during this incident,' they added. A picture posted on Twitter by the fire department afterwards showed the burnt out remains of the decimated truck. Tony Farmer, who lives next to the apartments, witnessed the whole incident after hearing one of the loud bangs from the truck. 'I was just chillin' inside and I heard a loud bang, kind of like when the [trash truck] comes and slams the dumpster onto the ground when they're done emptying it or a transformer [exploding],' Farmer told the Statesman. 'Then the second time it happened, I was like, 'Woah, that's a little weird,'' Farmer added. Farmer described the truck explosion as an almost cinematic experience. 'It exploded a little bit, almost like in the movies, not like a huge explosion like atomic, but it engulfed more about 20 seconds after [the man] was removed from the vehicle,' Farmer continued. He added that he believed the police officers involved in the man's rescue are heroes. 'Those cops...there's no doubt about it - they risked their lives and they're heroes. There's no doubt about it,' Farmer said. Assistant Chief of the Austin Police Department Robin Henderson agreed and later took to Twitter to congratulate the officers and give them a token for their heroics. 'Officers Pineda and Carrera received a chief's coin of recognition for their heroic actions yesterday, in saving a man's life from the vehicle fire,' Henderson tweeted. 'Please join me in recognizing these two for their quick-thinking and bravery! #OneAustinSaferTogether' According to the Patch, Joseph Chacon, police chief, also commended the officers for their bravery. 'Despite the danger, they rushed in and saved a man's life,' Chacon said. 'These are the guardians of our community. Thank you, officers Pineda and Carrera!' As of Wednesday morning, the condition of the injured man was not known. A press conference with further information is scheduled for Wednesday afternoon.

dawn.weathersbee Good work. And those two can save me anytime.
3d 241 likes Reply

dora_atx Helllo officers! 😍😍
3d 90 likes Reply

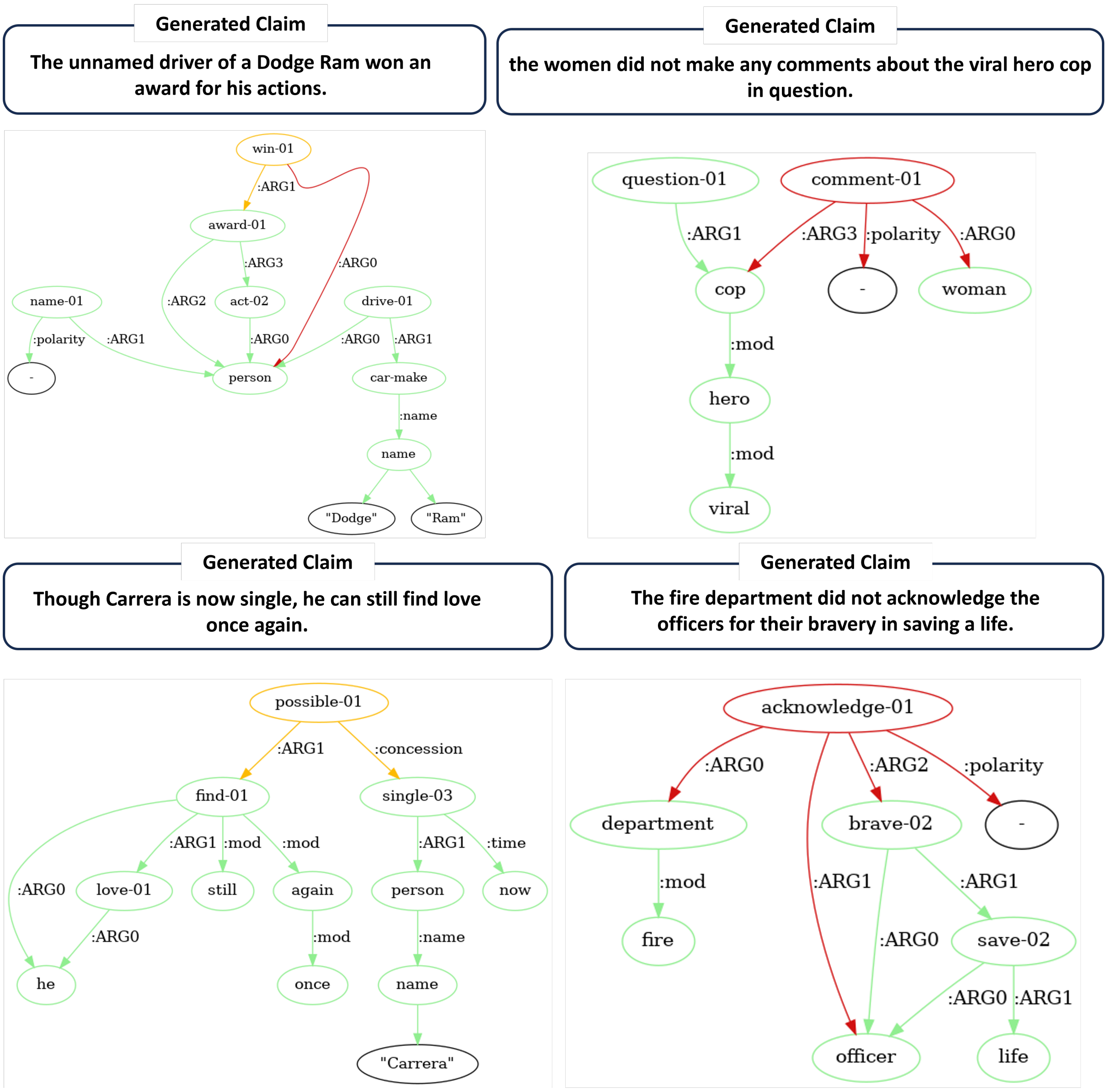
trinitystennfeld Which was hotter, these cops or the fire? 😮😅
3d 296 likes Reply

mrsnzuniga I'm about to start a fire. 🔥
3d 196 likes Reply

...two hero cops who saved an unconscious man from his burning pickup and went viral, for their deed and their looks. Austin Police Officers Eduardo Pineda and Chandler Carrera joined Fox News Digital via Zoom Tuesday to discuss last week's fiery rescue, the unexpected social media reaction and anti-police sentiment emerging across the country. They were on their way to a different call, Pineda said, when the urgent report of a man trapped in a flaming vehicle came in. Because they were the closest unit, they responded to the fire instead. They could already see fire and smoke before they pulled into the parking lot. HERO COPS DRAW WOMEN\'S ATTENTION AFTER DRAMATIC RESCUE OF AUSTIN MAN FROM BURNING PICKUP Austin police posted bodycam video from both officers' viewpoints. The first thing they noticed was the heat radiating off the flaming vehicle. "It was hot," Carrera told Fox News. "That's what I said on the video. But not really a whole lot goes through your mind. You just see what you have and deal with it as it comes." The videos show Pineda sprinting out the car's passenger seat toward the burning truck. "Get that extinguisher," he tells Carrera, hopping out before the car is fully stopped. He's the first to reach the burning, smoking vehicle, and the videos show flames spreading across the undercarriage. "He's still in it!" a man can be heard shouting as the officers arrive. ILLINOIS 'BATTLE BUDDY' PROGRAM SENDS COPS WITH MILITARY BACKGROUND TO RESPOND TO VETERANS IN CRISIS Pineda tries the door latch and grunts an expletive – it's locked and hot the touch. So he whips out his baton, shatters the driver-side window and orders the driver out, only to realize the man is unconscious, covered in burns, in the front seat. Carrera sprays the fire extinguisher, but Pineda asks for his help freeing the victim. The two wrestle him out of the car and drag him away, where they realize he's having a seizure. An explosion is heard off camera, and the next time the pickup is visible, flames are leaping out of the bed of the truck. Within another few seconds, Pineda tells dispatch "the vehicle's completely engulfed." The victim apparently suffered a medical emergency in the driver seat after backing into the parking spot – with his foot on the gas. The tires spun in place until flames burst out. The heroics come amid rising anti-police sentiment across the country and calls to defund the police – even in Austin, where city council voted to slash over $20 million from the police department's budget last summer. But the officers doubled down on their dedication to protecting and serving their community. "At the end of the day, no matter what happens, people are still going to call 911," Carrera said. "People are still going to need help. And so that\'s kind of why we want to become police officers in the first place." The officers were both named the city of Austin's employees of the week and awarded chief's coins. And after their photos appeared in the Austin American-Statesman newspaper – they became the toast of the town, with female fans quipping in comments on the outlet's Instagram how attractive they thought the officers were. For the record, Carrera, the former NCAA Division I football player for the University of North Alabama, is still single. Pineda said he is engaged. "We're not used to it," Pineda said of all the attention. "We don't like it." "It's weird," Carrera added. "It's very weird." The way they see it, they were just doing the job they signed up for. "We do what we do because, like every other police officer, we want to serve the community," Pineda said. "We're here to help. Doesn't matter, any situation where we get called, we're gonna show up and do what we can to help those who call, because they need the help." Later that evening, they performed life-saving CPR on a teenage shooting victim, Police Association President Ken Casaday told FOX 7 Austin. The station's news camera ...

They say they like a man in uniform. And a group of women on Instagram is fawning over two hero cops from Austin who went viral after pulling an injured man out of a burning truck in Texas Monday. "Good work," wrote @dawn.weathersbee. "And those two can save me anytime." Bodycam video shows Officers Eduardo Pineda and Chandler Carrera, a former Division I football player, approach the car and try to free the victim, who appears unresponsive and may have fallen into a seizure. Pineda broke the front window to get the door open, but struggled to pull the man free. He called for help from Carrera and the two wrestled him free and dragged him to safety shortly before the truck exploded. ILLINOIS 'BATTLE BUDDY' PROGRAM SENDS COPS WITH MILITARY BACKGROUND TO RESPOND TO VETERANS IN CRISIS Police have not released the victim's name. He was rushed to Dell Seton Medical Center with potentially life-threatening injuries, the Austin American-Statesman reported Wednesday. He had reportedly backed into a parking space and then suffered a medical emergency, according to the paper. He became unable to move with his foot stuck on the accelerator, causing the wheels to spin in place until a fire started. But after the dust settled, and their pictures were front-page news, social media lit up with praise for the men's deeds – and their looks. "Oww oww.. yes, I will step out of the vehicle," Instagram user @megan_ziskind wrote under a picture of Pineda shared by the Austin American-Statesman newspaper. "Which was hotter, these cops or the fire?" added @trinitystennfeld. User @mrsnzuniga wrote she might start a fire of her own – presumably hoping those officers would respond. "They look like actors cast in the role of police officers," added @paulavmphotography. CLICK HERE TO GET THE FOX NEWS APP During a news briefing Wednesday, the officers said they were just doing what they trained to do. "It feels good, but we don\'t consider ourselves heroes," Pineda told local reporters. "We\'re police officers. That\'s the job. We\'re here to help people." The officers were later named the department's employees of the week. And awarded chief's coins. "EVERYBODY got time for THAT," @stepamee joked.

Figure 9: We show a cluster of news documents containing multiple videos, images, and news articles. The cluster contains media about police officers pulling a man from a burning truck, along with cell phone video, body cam footage, and a press conference about the incident.



Generated Claim

The unnamed driver of a Dodge Ram won an award for his actions.

Generated Claim

the women did not make any comments about the viral hero cop in question.

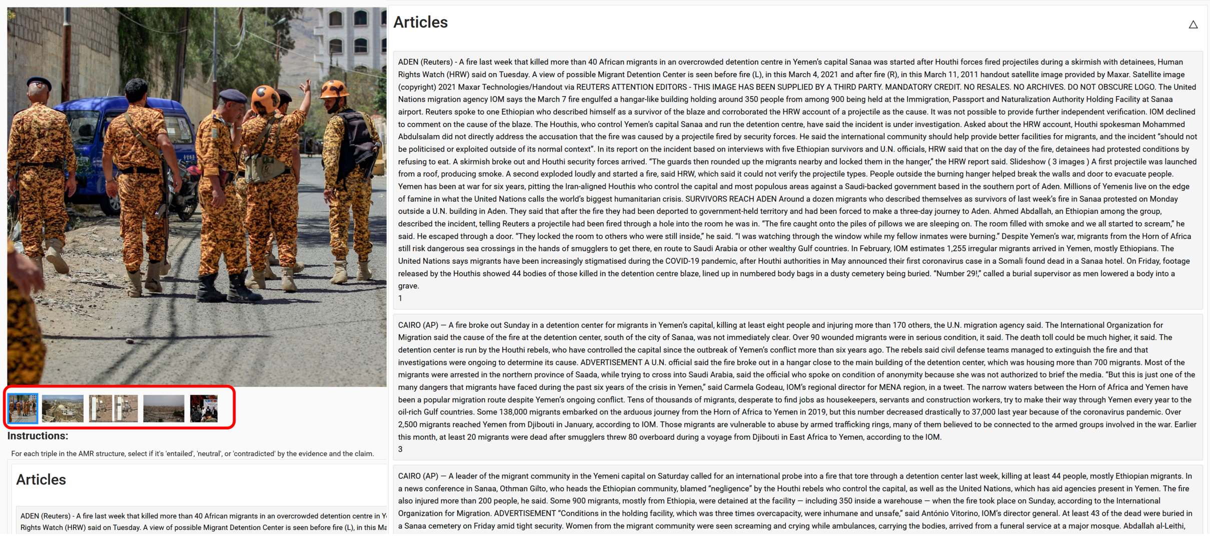
Generated Claim

Though Carrera is now single, he can still find love once again.

Generated Claim

The fire department did not acknowledge the officers for their bravery in saving a life.

Figure 10: We show predictions of our model for a set of claims generated for the previous cluster.



Articles

ADEN (Reuters) - A fire last week that killed more than 40 African migrants in an overcrowded detention centre in Yemen’s capital Sanaa was started after Houthi forces fired projectiles during a skirmish with detainees, Human Rights Watch (HRW) said on Tuesday. A view of possible Migrant Detention Center is seen before fire (L), in this March 4, 2021 and after fire (R), in this March 11, 2011 handout satellite image provided by Maxar. Satellite image (copyright) 2021 Maxar Technologies/Handout via REUTERS ATTENTION EDITORS - THIS IMAGE HAS BEEN SUPPLIED BY A THIRD PARTY. MANDATORY CREDIT. NO RESALES. NO ARCHIVES. DO NOT OBSCURE LOGO. The United Nations migration agency IOM says the March 7 fire engulfed a hangar-like building holding around 350 people from among 900 being held at the Immigration, Passport and Naturalization Authority Holding Facility at Sanaa airport. Reuters spoke to one Ethiopian who described himself as a survivor of the blaze and corroborated the HRW account of a projectile as the cause. It was not possible to provide further independent verification. IOM declined to comment on the cause of the blaze. The Houthis, who control Yemen’s capital Sanaa and run the detention centre, have said the incident is under investigation. Asked about the HRW account, Houthi spokesman Mohammed Abdulsalam did not directly address the accusation that the fire was caused by a projectile fired by security forces. He said the international community should help provide better facilities for migrants, and the incident “should not be politicised or exploited outside of its normal context”. In its report on the incident based on interviews with five Ethiopian survivors and U.N. officials, HRW said that on the day of the fire, detainees had protested conditions by refusing to eat. A skirmish broke out and Houthi security forces arrived. “The guards then rounded up the migrants nearby and locked them in the hanger,” the HRW report said. Slideshow (3 images) A first projectile was launched from a roof, producing smoke. A second exploded loudly and started a fire, said HRW, which said it could not verify the projectile types. People outside the burning hanger helped break the walls and door to evacuate people. Yemen has been at war for six years, pitting the Iran-aligned Houthis who control the capital and most populous areas against a Saudi-backed government based in the southern port of Aden. Millions of Yemenis live on the edge of famine in what the United Nations calls the world’s biggest humanitarian crisis. SURVIVORS REACH ADEN Around a dozen migrants who described themselves as survivors of last week’s fire in Sanaa protested on Monday outside a U.N. building in Aden. They said that after the fire they had been deported to government-held territory and had been forced to make a three-day journey to Aden. Ahmed Abdallah, an Ethiopian among the group, described the incident, telling Reuters a projectile had been fired through a hole into the room he was in. “The fire caught onto the piles of pillows we are sleeping on. The room filled with smoke and we all started to scream,” he said. He escaped through a door. “They locked the room to others who were still inside,” he said. “I was watching through the window while my fellow inmates were burning.” Despite Yemen’s war, migrants from the Horn of Africa still risk dangerous sea crossings in the hands of smugglers to get there, en route to Saudi Arabia or other wealthy Gulf countries. In February, IOM estimates 1,255 irregular migrants arrived in Yemen, mostly Ethiopians. The United Nations says migrants have been increasingly stigmatised during the COVID-19 pandemic, after Houthi authorities in May announced their first coronavirus case in a Somali found dead in a Sanaa hotel. On Friday, footage released by the Houthis showed 44 bodies of those killed in the detention centre blaze, lined up in numbered body bags in a dusty cemetery being buried. “Number 29!,” called a burial supervisor as men lowered a body into a grave.
1

CAIRO (AP) — A fire broke out Sunday in a detention center for migrants in Yemen’s capital, killing at least eight people and injuring more than 170 others, the U.N. migration agency said. The International Organization for Migration said the cause of the fire at the detention center, south of the city of Sanaa, was not immediately clear. Over 90 wounded migrants were in serious condition, it said. The death toll could be much higher, it said. The detention center is run by the Houthi rebels, who have controlled the capital since the outbreak of Yemen’s conflict more than six years ago. The rebels said civil defense teams managed to extinguish the fire and that investigations were ongoing to determine its cause. ADVERTISEMENT A U.N. official said the fire broke out in a hangar close to the main building of the detention center, which was housing more than 700 migrants. Most of the migrants were arrested in the northern province of Saada, while trying to cross into Saudi Arabia, said the official who spoke on condition of anonymity because she was not authorized to brief the media. “But this is just one of the many dangers that migrants have faced during the past six years of the crisis in Yemen,” said Carmela Godeau, IOM’s regional director for MENA region, in a tweet. The narrow waters between the Horn of Africa and Yemen have been a popular migration route despite Yemen’s ongoing conflict. Tens of thousands of migrants, desperate to find jobs as housekeepers, servants and construction workers, try to make their way through Yemen every year to the oil-rich Gulf countries. Some 138,000 migrants embarked on the arduous journey from the Horn of Africa to Yemen in 2019, but this number decreased drastically to 37,000 last year because of the coronavirus pandemic. Over 2,500 migrants reached Yemen from Djibouti in January, according to IOM. Those migrants are vulnerable to abuse by armed trafficking rings, many of them believed to be connected to the armed groups involved in the war. Earlier this month, at least 20 migrants were dead after smugglers threw 80 overboard during a voyage from Djibouti in East Africa to Yemen, according to the IOM.
3

CAIRO (AP) — A leader of the migrant community in the Yemeni capital on Saturday called for an international probe into a fire that tore through a detention center last week, killing at least 44 people, mostly Ethiopian migrants. In a news conference in Sanaa, Othman Gilto, who heads the Ethiopian community, blamed “negligence” by the Houthi rebels who control the capital, as well as the United Nations, which has aid agencies present in Yemen. The fire also injured more than 200 people, he said. Some 900 migrants, mostly from Ethiopia, were detained at the facility — including 350 inside a warehouse — when the fire took place on Sunday, according to the International Organization for Migration. ADVERTISEMENT “Conditions in the holding facility, which was three times overcapacity, were inhumane and unsafe,” said António Vitorino, IOM’s director general. At least 43 of the dead were buried in a Sanaa cemetery on Friday amid tight security. Women from the migrant community were seen screaming and crying while ambulances, carrying the bodies, arrived from a funeral service at a major mosque. Abdallah al-Leithi,

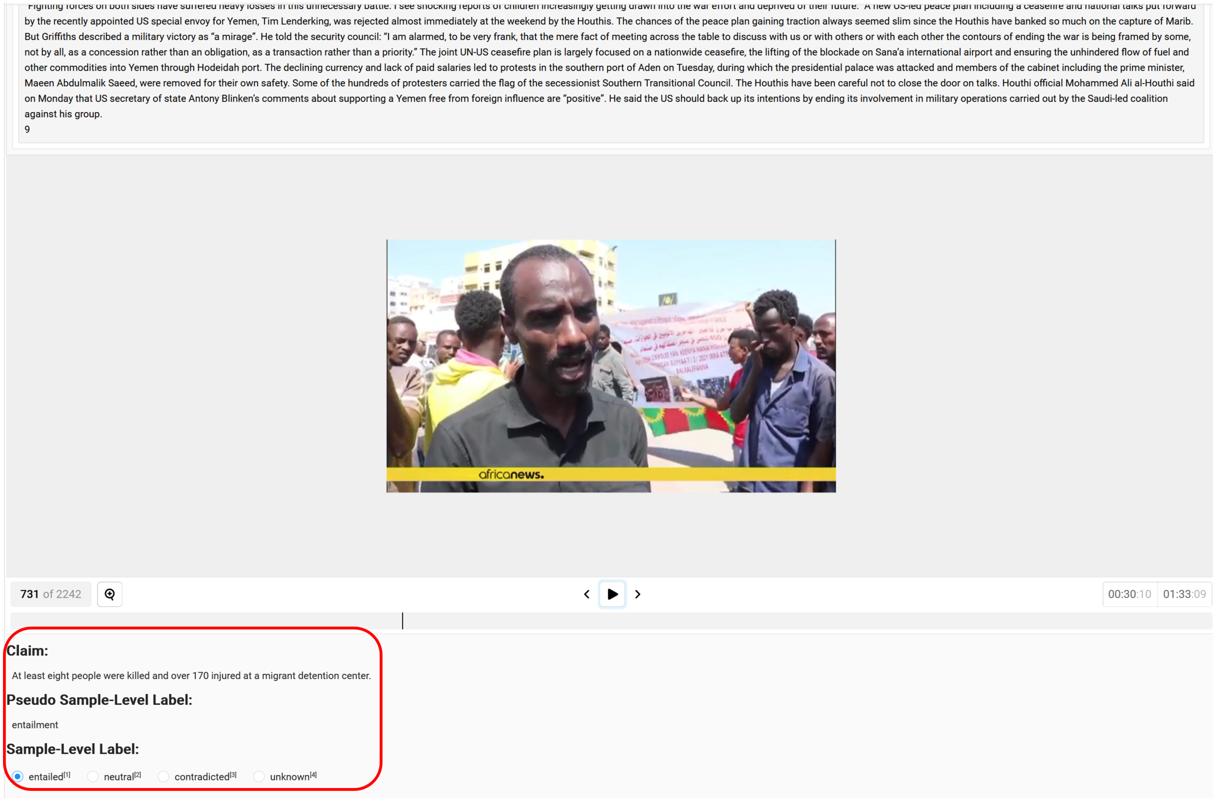
Instructions:

For each triple in the AMR structure, select if it's 'entailed', 'neutral', or 'contradicted' by the evidence and the claim.

Articles

ADEN (Reuters) - A fire last week that killed more than 40 African migrants in an overcrowded detention centre in Y
Rights Watch (HRW) said on Tuesday. A view of possible Migrant Detention Center is seen before fire (L), in this Ma

Figure 11: This figure illustrates the labeling interface in LabelStudio, where annotators are required to review multiple news articles and their accompanying images before assigning labels to claims. This process can be time-consuming and challenging, as some claims rely on evidence scattered across small pieces of text or other modalities within the articles, demanding careful examination and synthesis of information from various sources to make accurate labeling decisions.



Fighting forces on both sides have suffered heavy losses in this unnecessary battle. I see shocking reports of children increasingly getting drawn into the war effort and deprived of their future.” A new US-led peace plan including a ceasefire and national talks put forward by the recently appointed US special envoy for Yemen, Tim Lenderking, was rejected almost immediately at the weekend by the Houthis. The chances of the peace plan gaining traction always seemed slim since the Houthis have banked so much on the capture of Marib. But Griffiths described a military victory as “a mirage”. He told the security council: “I am alarmed, to be very frank, that the mere fact of meeting across the table to discuss with us or with others or with each other the contours of ending the war is being framed by some, not by all, as a concession rather than an obligation, as a transaction rather than a priority.” The joint UN-US ceasefire plan is largely focused on a nationwide ceasefire, the lifting of the blockade on Sana’a international airport and ensuring the unhindered flow of fuel and other commodities into Yemen through Hodeidah port. The declining currency and lack of paid salaries led to protests in the southern port of Aden on Tuesday, during which the presidential palace was attacked and members of the cabinet including the prime minister, Maeen Abdulmalik Saeed, were removed for their own safety. Some of the hundreds of protesters carried the flag of the secessionist Southern Transitional Council. The Houthis have been careful not to close the door on talks. Houthi official Mohammed Ali al-Houthi said on Monday that US secretary of state Antony Blinken’s comments about supporting a Yemen free from foreign influence are “positive”. He said the US should back up its intentions by ending its involvement in military operations carried out by the Saudi-led coalition against his group.
9

Claim:

At least eight people were killed and over 170 injured at a migrant detention center.

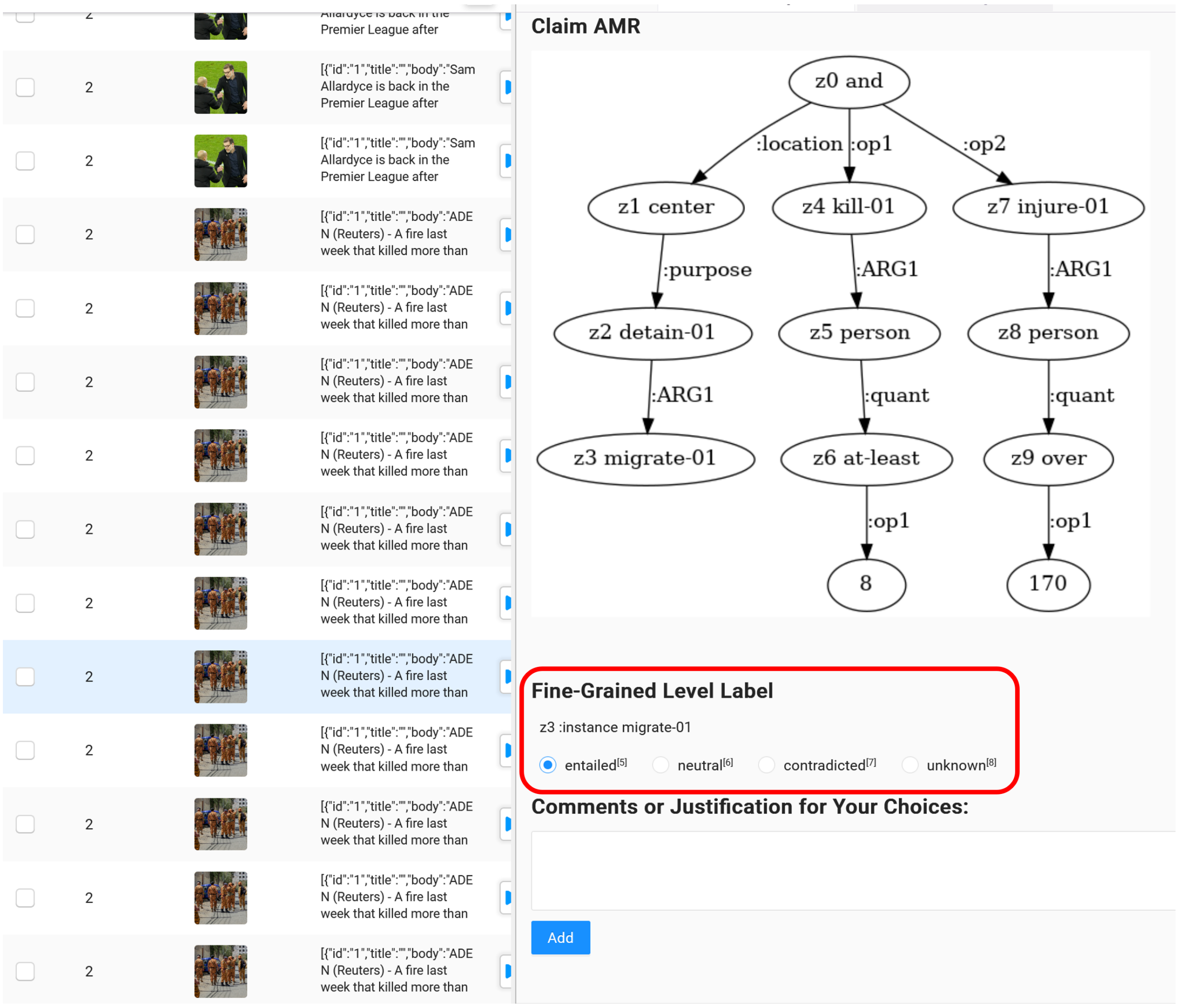
Pseudo Sample-Level Label:

entailment

Sample-Level Label:

● entailed[1] ○ neutral[2] ○ contradicted[3] ○ unknown[4]

Figure 12: This figure shows the video examination process in the labeling interface, where annotators are tasked with reviewing videos associated with each news cluster, in addition to the news articles and corresponding images. The number of videos to be examined can range from none to a dozen per cluster with variable length. After thoroughly examining the evidence from news documents, images, and videos, the annotators are required to assign a logical label to the given claim, indicating its truthfulness based on the available multimodal information.



□	2		Allardyce is back in the Premier League after
□	2		[{"id":"1","title":"","body":"Sam Allardyce is back in the Premier League after
□	2		[{"id":"1","title":"","body":"Sam Allardyce is back in the Premier League after
□	2		[{"id":"1","title":"","body":"ADEN (Reuters) - A fire last week that killed more than
□	2		[{"id":"1","title":"","body":"ADEN (Reuters) - A fire last week that killed more than
□	2		[{"id":"1","title":"","body":"ADEN (Reuters) - A fire last week that killed more than
□	2		[{"id":"1","title":"","body":"ADEN (Reuters) - A fire last week that killed more than
□	2		[{"id":"1","title":"","body":"ADEN (Reuters) - A fire last week that killed more than
□	2		[{"id":"1","title":"","body":"ADEN (Reuters) - A fire last week that killed more than
□	2		[{"id":"1","title":"","body":"ADEN (Reuters) - A fire last week that killed more than
□	2		[{"id":"1","title":"","body":"ADEN (Reuters) - A fire last week that killed more than
□	2		[{"id":"1","title":"","body":"ADEN (Reuters) - A fire last week that killed more than
□	2		[{"id":"1","title":"","body":"ADEN (Reuters) - A fire last week that killed more than
□	2		[{"id":"1","title":"","body":"ADEN (Reuters) - A fire last week that killed more than

Claim AMR

Fine-Grained Level Label

z3 :instance migrate-01

● entailed[5] ○ neutral[6] ○ contradicted[7] ○ unknown[8]

Comments or Justification for Your Choices:

Figure 13: This figure depicts the fine-grained labeling process for AMR tuples in our dataset. Annotators are required to iterate through each AMR node and edge from the AMR tree, assigning fine-grained labels to evaluate the truthfulness of the claim at a more granular level. This process involves examining each tuple individually and making labeling decisions based on the available evidence from the news articles, images, and videos associated with the corresponding news cluster.

node must be labeled as neutral to maintain consistency.

- Maintenance of structural constraints: The annotators must ensure that the structural constraints within the AMR tree are preserved. This means that the labels of nodes and edges should be semantically consistent with each other. For instance, if a node is labeled as contradicted, the corresponding edge must also be labeled as contradicted to maintain the logical structure of the AMR.

- Collaboration and discussion: The annotators are encouraged to collaborate and discuss any ambiguities or disagreements in the labeling process. This collaborative approach helps to resolve any inconsistencies and ensures that the resulting labels are accurate and semantically consistent.

Adhering to these guidelines ensures that the fine-grained labels assigned by the annotators are semantically consistent within the AMR tree and accurately represent the information conveyed in the news clusters. However, when examining the IAA (IAA) scores for the human-labeled dataset, we observe a discrepancy between the sample-level and fine-grained level labels. The IAA for the sample-level labels is a high 93%, indicating strong agreement among the annotators when it comes to the overall veracity of the claims, suggesting that the annotators have a clear understanding of the broader context and are generally able to determine whether a claim is true, false, or neutral based on the available evidence. In contrast, the IAA for the fine-grained level labels is lower, at 68%, revealing that even when the annotators agree on the overall truthfulness of a claim, there can be disagreements when it comes to assigning labels to specific elements within the claim. This discrepancy highlights the complexity and nuance involved in fine-grained fact-checking, as different annotators may interpret the evidence differently or focus on different aspects of the claim when making their labeling decisions.

Table 3 presents the results of our model compared to the ground truth human labels, which are determined by the most voted label among the annotators. The results show that the entailment label accuracy of our model is close to human performance, indicating that the model can effectively identify claims that are supported by the available evidence. However, the model's performance on neutral and contradicted labels is not as high as its entailment accuracy, suggesting room for improvement in these areas. Despite this, the overall results demonstrate that our model can successfully assess the truthfulness of claims in this task, even though it may not yet match human performance across all label categories. The IAA scores further highlight the challenges associated with fine-grained fact-checking, even when all the experts involved in labeling the data do so with the utmost care and attention to detail.

A.4 LVLM Baselines

To evaluate the performance of LVLMs on fine-grained AMR prediction, we had to employ a workaround since these models do not natively support this task. Our approach involved using in-context learning to enable the LVLM models to perform fine-grained prediction at the word token level first. Once the models generated their predictions for the individual word tokens, we then mapped these results back to the corresponding nodes and edges in the AMR tree. This process allowed us to evaluate our dataset with LVLM models, even though they were not explicitly designed for this purpose.

We compare our model's performance with two state-of-the-art LVLMs trained on instructional data, which have demonstrated strong performance in tasks such as visual question answering and image captioning.

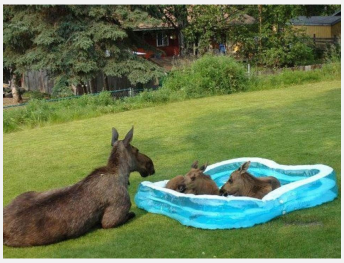
Our LVLM baselines include:

- **LLaVA** (Liu et al., 2023) is an instruction-tuned multimodal LVLM with strong image-text understanding capabilities. The model encodes image data using a pre-trained CLIP ViT-L/14 (Radford et al., 2021) and projects it into the Vicuna LLM's text embedding space (Chiang et al., 2023). It is tuned using large multimodal instructions curated via querying GPT-4 (Achiam et al., 2023).

- **MiniGPT-v2** (Chen et al., 2023) is an improved version of MiniGPT-4 (Zhu et al., 2023) and has a simpler architecture. It uses EVA (Fang et al., 2023) as the pretrained CLIP image encoder and LLaMA-2-Chat (Touvron et al., 2023) as the LLM backbone. The model demonstrates strong performance in multimodal understanding on numerous image-text tasks.

Prompts We obtain both sample-level and fine-grained predictions from the LVLM baselines by prompting them in a zero-shot manner. In the single-document setting (*i.e.*, MOCHEG), we provide the LVLM with multimodal evidence – including a text document and its corresponding image – alongside a claim and an instructional question. Given the evidence, the prompt instructs the model to verify either the entire claim or words within the claim, corresponding with the sample-level task or the fine-grained task, respectively. Figure 14 shows an example of a text prompt constructed from an example in the MOCHEG dataset, and Figure 15 includes all the different questions to be prompted for that example. In the multi-document setting (*i.e.*, M^3DC), we carry out the same steps for each document in a document group. We then perform majority voting among the group's predictions to compute the final prediction.

Given the evidence (including the image and a text article) and a text claim, please indicate whether **a word in the claim** is supported or refuted by the evidence.



Article: A photograph purportedly showing a moose and two calves enjoying a kiddie pool as they watched a car burn across the street has been circulating online for several years. While it is frequently shared as a genuine (albeit bizarre) item, this image is a composite of at least two separate photographs. The photograph of the car on fire first appeared online when it was published on Reddit in May 2013. It seems as if they were trying to jump-start it. Obviously, they don't know their cars too well. The whole neighborhood has gathered for the impromptu neighborhood bonfire. While we haven't been able to locate the specific origin of the moose image, we know that the photograph was also posted separately to Reddit in May 2013: Unsurprisingly, the first version of the image featuring moose in a kiddie pool watching a car fire appeared on (of course) Reddit, shortly after the two source images were posted.

Claim: A photograph shows a moose enjoying a wading pool while watching a car burn.
Question: Is the word "moose" in the claim true, false, or neutral with regard to the evidence? Output "True" if the evidence supports the word, "False" if the evidence contradicts the word, or "Neutral" if it is neither supported nor refuted.
Answer:

Figure 14: An example of a zero-shot prompt to be fed into LVLMs for sample-level and fine-grained predictions, constructed from a data example in the MOCHEG dataset.

Sample-level question: Is the claim true, false, or neutral with regard to the evidence? Answer the question using a single word or phrase
Fine-grained questions:

- Is the word "photograph" in the claim true, false, or neutral with regards to the evidence? Answer the question using a single word or phrase
- Is the word "shows" in the claim true, false, or neutral with regards to the evidence? Answer the question using a single word or phrase
- Is the word "moose" in the claim true, false, or neutral with regards to the evidence? Answer the question using a single word or phrase
- Is the word "enjoying" in the claim true, false, or neutral with regards to the evidence? Answer the question using a single word or phrase
- Is the word "wading" in the claim true, false, or neutral with regards to the evidence? Answer the question using a single word or phrase
- Is the word "pool" in the claim true, false, or neutral with regards to the evidence? Answer the question using a single word or phrase
- Is the word "watching" in the claim true, false, or neutral with regards to the evidence? Answer the question using a single word or phrase

Figure 15: Example questions to prompt sample-level and fine-grained zero-shot predictions. We only construct fine-grained questions on words that can be mapped to AMR triple annotations to ensure ground truths for evaluation.