Not all Fake News is Written: A Dataset and Analysis of Misleading Video Headlines

Anonymous ACL submission

Abstract

Polarization and the marketplace for impressions have conspired to make navigating information online difficult for users, and while there has been a significant effort to detect false or misleading text, multimodal datasets have received considerably less attention. To complement existing resources, we present multimodal Video Misleading Headline (VMH), a dataset that consists of videos and whether annotators believe the headline is representative of the video's contents. After collecting and annotating this dataset, we analyze multimodal baselines for detecting misleading headlines. Our annotation process also focuses on *why* annotators view a video as misleading, allowing us to better understand the interplay of annotators' background and the content of the videos.

1 Introduction

Social media platforms are used by half of U.S. adults for everyday news consumption, according to Walker and Matsa (2021). They have even supplanted television as the most common purveyor of news (Wakefield, 2016). However, content created on these online platforms are often lower quality than traditional sources and more prone to false stories. Vosoughi et al. (2018) contend that false news spreads six times faster online than offline.

This work focuses on one part of this problem: does a video headline match its content. We call this **misleading video headline** detection. In text, this is referred to as incongruent headline detection (Chesney et al., 2017) and is an important problem because the headline is the first step to a reader accessing content (dos Rieis et al., 2015). While there have been efforts to identify misleading information by analyzing textual content in the headline, recent work has shown that users are more likely to believe fake news when it is accompanied by videos (Wang et al., 2021).

Hence, it is necessary to investigate content outside the text (e.g., with videos) as it can help make a more informed decision by directly analyzing the relationship between the headline and the video.

VMH Dataset	
Headline	**Clinton Says Trump "Making Up Lies" About New FBI Review**
Video	`https://www.facebook.com/watch/?v=10154955844338812`
Label	**Misleading**
Rationale	**The headline implies more than what is introduced in the video.**
Subrationale	**The headline exaggerates the video content.**
Annotator ID	`A2P8V5SKYLL5I4`
Annotator Profile	`Ages 30-49, Black, Democratic, Men, Post college`
Venue	ABC News
Venue Kind	Broadcast
Venue Credibility	High
News Topic	Politics
Headline Property	Factual Statement
Transcript	...is already making up lies about this he is doing his best to confuse misleading and discourage the American people

Table 1: VMH includes video headline, video, annotator's label, and rationales the label is grounded. In the video, the part about "New FBI Review" was not present, and thereby annotation is *misleading* because the headline was implying more than the video content.

To understand this new task, we create a new dataset—Video Misleading Headline (VMH)—that includes 2,247 news articles labeled as *representative* or *misleading* (Section 2). A careful annotation process captures not just whether videos are misleading but *why*. We investigate videos, label rationales, and headline meta information (e.g., venues, news topics, and headline properties) to analyze the features that may contribute towards an instance being identified as misleading (Section 3). Section 4 shows that existing models fail to identify misleading video headlines, showing that this important but difficult task requires further research in both

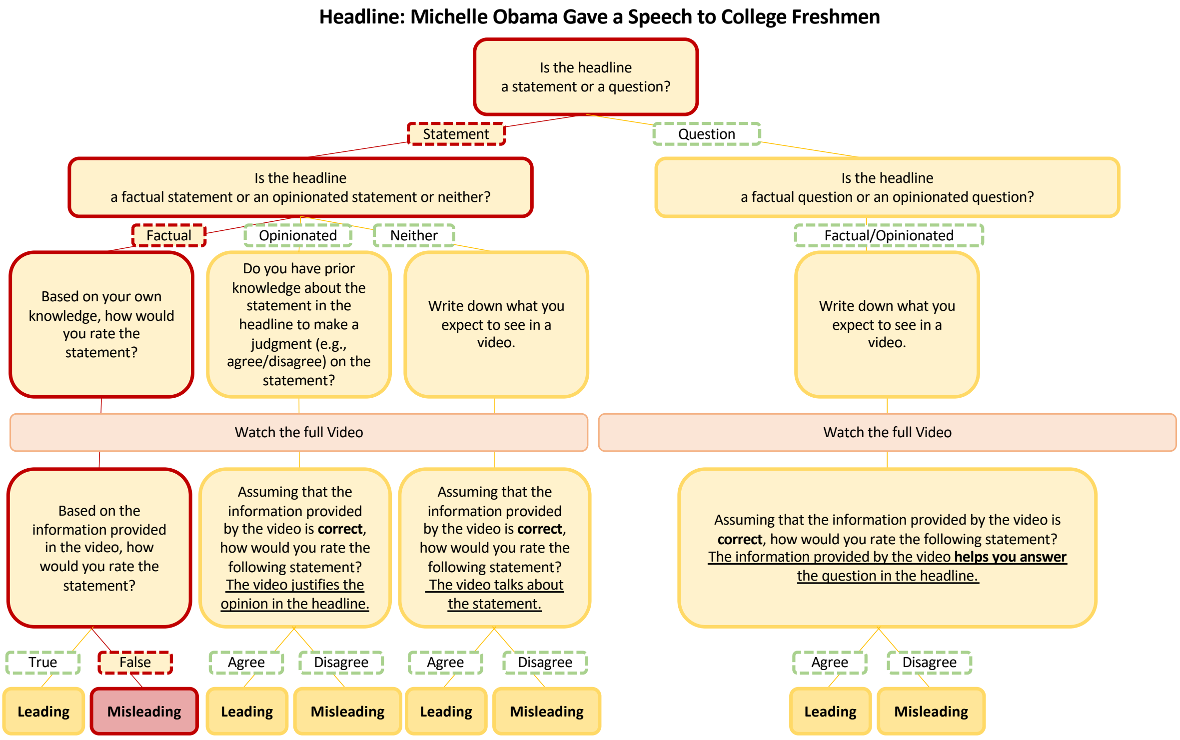


Figure 1: In the annotation tree, the annotators first consider if the headline "Michelle Obama Gave a Speech to College Freshmen" is a factual statement. Next, they answer the question, "Based on the information provided in the video, how would you rate the statement?" Because the answer was *False*, the implied label is *misleading*. The headline is indeed *misleading* because whether "College Freshmean" were present in the video is unclear, making it impossible to assess the veracity.

the text and visual domains.

A *misleading headline* is when the headline distorts the underlying content (Wei and Wan, 2017) and facts in the news body, leading the audience to imply more or less than what was actually presented in the content. For example, in our task, the headline "Obama: I'm proud to be leaving *without* scandal" does not fully engage the video's content because the headline exaggerates the view of the content; the video plays Obama's speech that he left the administration without a *significant* scandal. This distortion makes detecting misleading video headlines even more arduous because the video content has to be integrated with the headline subtlety while assessing headline veracity.

2 Video Misleading Heading Dataset VMH

VMH consists of 2,247 video posts from 2014 to 2016. We focus on this period because it coincided with the 2016 US presidential election, which was rife with disinformation, and is distant enough from current events that we believe annotators can be more confident about determining whether claims are true; as even news organizations are not immune to false news (Starbird et al., 2019).

We harvested Facebook video posts from Rony et al. (2017), where we manually filtered any video that exceeded five minutes or had low-quality video or sound. The videos in VMH are average two minutes long. The resulting video posts (example in Table 1) come from fifty-two media venues, including the most circulated print and broadcast media and unreliable media in the US (Listed in Appendix A from a trustworthy journalism perspective) (Edelson et al., 2021; Samory et al., 2020).

We further collect venue-related information such as venue credibility[1] (e.g., High) and venue kind[2] (e.g., Broadcast). Also, we manually assigned news topics (e.g., Politics) inspired by News Areas[3] to each headline. We create audio transcripts (also released in our dataset) using automatic speech recognition software[4] whenever the video is accompanied by intelligible audio (details in Appendix H). Other features in the dataset in-

[1] Mediabiasfactcheck site
[2] State of the News Media
[3] News Topics
[4] https://deepgram.com

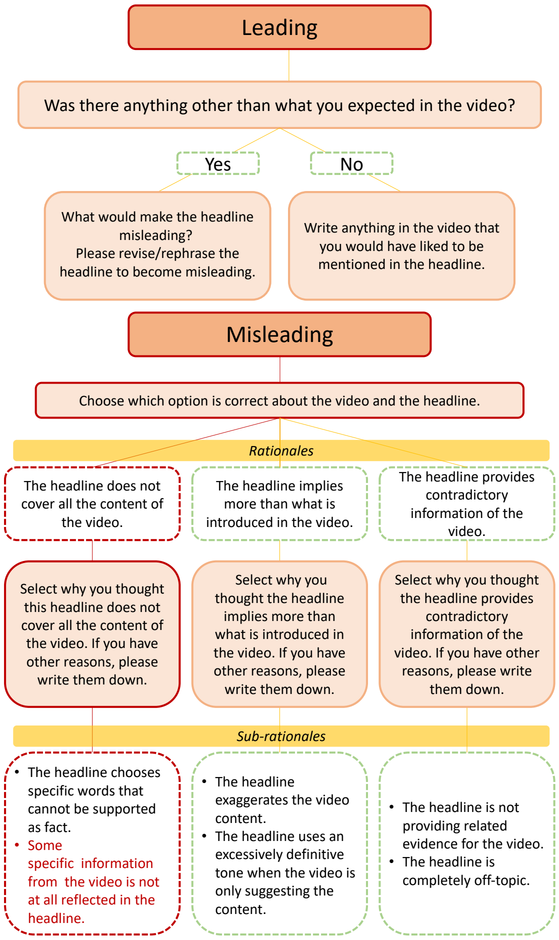


Figure 2: After label annotation, the annotators provide grounding for the *misleading* labels. The figure shows how rationales and subrationales are selected in a hierarchical manner.

clude the number of tokens per headline (average 7.75 tokens) and annotator profile (e.g., gender).

2.1 Annotation

We ask Mechanical Turkers to identify misleading video headlines (Snow et al., 2008). We intentionally assign the annotation task to laypersons to reflect the real-world misleading headline phenomena. For each task, the annotator undergoes two phases, labeling and rationale annotation. We recruit three annotators per task (Chandler et al., 2014).

Label Annotation We structure the label annotation task as a series of questions to help annotators engage with the content of the headline and video (Figure 1). Because headlines can take different forms (statements of facts or opinions, questions, etc.), we first ask the user to determine the form of the headline. We refer to these forms as headline property in the sequel. They then engage with the headline in different ways depending on the headline property they selected (i.e., do they agree with the headline, do they believe the fact is true, etc.) (headline properties and associated questions in Appendix C). This helps them build a mental model of the content of the hypothetical video before viewing it. We adopted this format after initial pilots indicated that merely asking if a video was misleading is too ambiguous (pilot example in Appendix B).

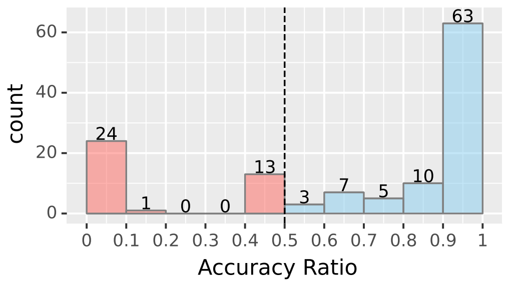
After the annotator has built a mental model, we ask the annotators to watch the video and answer whether the information provided in the video is consistent with the annotator's mental model of the video. If it is, then it suggests the video is *representative*: it answered the question asked by the headline, justified an opinion, or gave evidence of a new event.

In contrast, if the video fails this check, we conclude that the headline is *misleading*. To reflect the subtle difference in participants' opinions, we provide answer options that represent the levels of veracity or agreement with the headline (e.g., True, Mostly True, Mostly False, False, I don't know). For the translation to binary labels, we regard the last three answers as *misleading*.

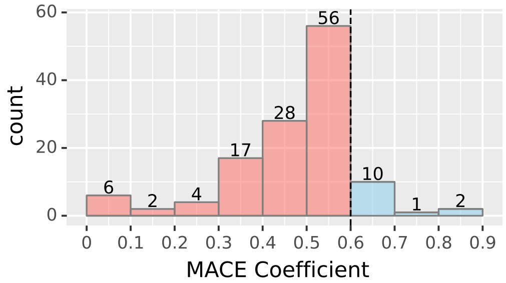
Rationale Annotation We then turn to the rationale annotation step. If their label is *misleading*, we ask the annotators to provide justifications for their decision (Figure 2). For example, when an annotator labels a headline as *misleading* and chooses *The headline does not cover all the content of the video* as their rationale for the label, a subrationale is further used to reason the ways in which the headline might not contain the content.

We offer pre-populated rationales to force objectivity in the annotator's decision and exploit the rationales more systematically. For subrationales, we allow the annotator to provide free-form text.

Providing such annotations can improve not just data quality (Briakou and Carpuat, 2020)—by forcing the annotator to think about their reasoning—but also model accuracy (Zaidan et al., 2007) for natural language processing tasks. After the annotation is complete, final annotations are determined using a majority vote from the three annotators (Yang et al., 2015). We do not apply majority voting for subrationales that include free-form texts.



(a) Qualified Workers by Accuracy Score Threshold



(b) Qualified Workers by MACE Score Threshold

Figure 3: The thresholds of accuracy ratio and MACE Coefficient are manually assigned to ensure *competent* workers are recruited after each annotation session.

2.2 Quality Control and Assessment

Quality Control We control the quality of VMH to select good crowdworkers using their accuracy score on synthetically created accuracy check questions and MACE score (Paun et al., 2018). Accuracy check questions are synthetically created to be always misleading (obviously false). For each annotator, we calculate the ratio between the number of correct answers and the number of accuracy check questions they answered (examples of accuracy check questions in Appendix D).

To determine which users are reliable and to infer the labels annotators disagree on, we use a latent variable model that explicitly estimates an annotator's accuracy. This model, MACE (Martín-Morató et al., 2021) corrects for annotator-level biases (an annotator might overly favor a particular label, could have low overall accuracy, etc.). We use the point estimates—mean—from the posterior distributions of latent variables that stand for the trustworthiness of each worker (details about applying MACE to worker accuracy estimation in Appendix D).

We run two annotation sessions to estimate and accumulate qualified workers. In the initial session, accuracy and MACE scores are considered to combine working agreement with known and inferred labels (Paun et al., 2018), thereby selectively filtering less competent annotators. Crowdworkers are invited back only if their accuracy (0.5) or MACE score is high enough (0.6). Each threshold is empirically assigned. This yields 88 and 13 qualified workers from each metric (Figure 3).

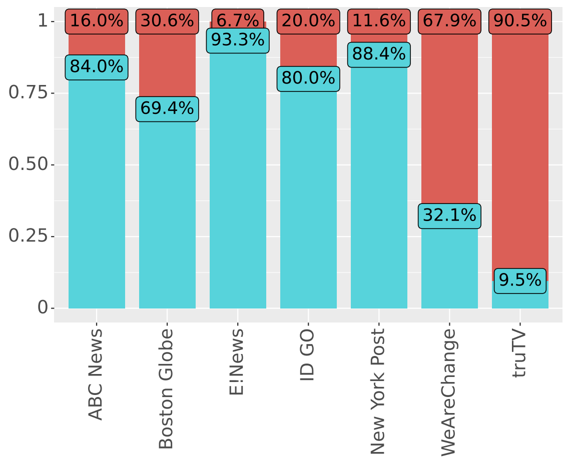
Quality Assessment We report Krippendorf's α values following Toledo et al. (2019) to quantify annotation quality. Krippendorf's α value of the three annotators who passed the accuracy score threshold are 0.57 for labels and 0.33 for rationales. The Krippendorf's α values of the workers who were found to be competent according to the MACE score are 0.68 and 0.21. While the values exhibit moderate-to-low agreement, this is expected due to the inherent subjectivity of the annotation task (Daume III and Marcu, 2005).

3 Dataset Analysis

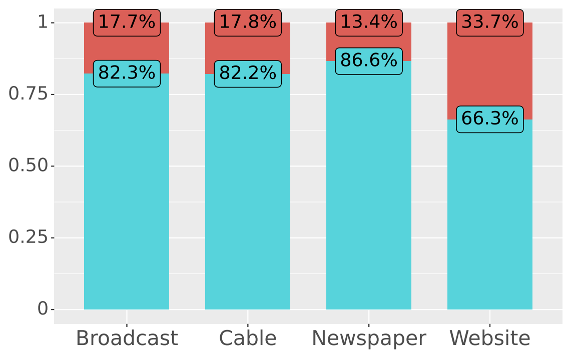
Out of 2,247 video headlines, 1,906 headlines are annotated as *representative*, while 341 headlines are annotated as *misleading*, suggesting a high-class imbalance. In this section, we investigate various aspects of VMH to gain a deeper understanding of features that could potentially contribute to a headline being classified as misleading. We further investigate the inherent qualities of VMH by examining annotation patterns in different aspects.

Misleading Features Figure 4 suggests that the venues *TruTV* and *WeAreChange.org* are strong indicators for misleading headlines. Also, videos from the *Website* venue (as opposed to traditional media) are likely to be misleading (29%). This suggests that the specific venue and the kind of venue may help detect misleading headlines (see Appendix E for more feature analyses).

Clickbait Misleading videos and clickbait both have the same goal: to entice more people to click on the underlying content. A reasonable hypothesis is that they would use similar tricks to lure in users. Thus, we reproduce the features found by (Dhoju et al., 2019) to be associated with clickbait headlines such as the number of demonstrative adjectives, numbers, and WH-words (e.g., what, who, how) for the headlines in VMH. Demonstrative adjectives appear in misleading headlines (Table 2), while numbers and superlative word features are less frequent in our dataset. Numbers and modal



(a) Venue Distribution



(b) Venue Kind Distribuution

Figure 4: The venues *TruTV*, *WeAreChange.org* and venue kind *Website* were the strongest indicators of misleading headlines. The red and blue bars denote bar proportions for *misleading* and *leading* labels respectively. For venue distribution, we report the examinations of the first eight venues with the most misleading headlines due to space limitation. (*ID GO*: Investigation Discovery)

words appear in similar frequencies. Thus, misleading video headlines are not the same as clickbaits.

Clickbait Patterns	Presence Ratio	
	Dhoju et al. (2019)	VMH (Ours)
Demonstrative Adj	0.80	0.61
WH-Words	0.70	0.40
Numbers	0.72	0.60
Modal	0.27	0.20
Superlative	0.30	0.06

Table 2: Clickbait patterns in misleading headlines in VMH to demonstrate the difference between clickbait detection and misleading video headline task.

Investigation of Bias in Annotation Because our dataset has many politically relevant videos, we also ask annotators' political leanings to see if it biases their annotations. A χ^2 test does not suggest that annotations and political leanings are dependent (p-value 0.36); indeed the marginal proportion of misleading videos are comparable (Democratic: 22.9%, Republican: 22.6%, and Independent: 33%).

We also manually check fifty video headlines to see if their ideologies affected a headline's assigned label, finding no substantial consequences. For example, the headline "Charles Blow: Donald Trump is a bigot", presumably "anti-Trump", was annotated *Representative* by an annotator with a "Republican" leaning.

Task Subjectivity Motivated by Section 2.2, we examine the annotations that fail to have consensus among annotator decisions: there were 1436 *representative* and 159 *misleading* instances with the perfect agreement, leaving 30% to annotations that had disagreement. In addition to disagreeing on labels, annotators disagree about why they the headline is misleading (Table 3).

4 Experiments

The misleading headline detection task is challenging because of the inherent subjectivity of the task. It also necessitates multimodal approaches that can consider both the headline and the video to make inferences about the nature of the relationship (*representative* or *misleading*) between the two. Hence, in this section, we benchmark both text-only and multimodal approaches typically used for detecting video-text similarity and video-text entailment tasks.

Experiment Settings We compare the performance of models when trained with various combinations of input features in our dataset. The features that we consider are headlines (H) and their associated video clips (V), transcripts (T), rationales, and sub-rationales (R).

For textual feature, we concatenate features as: `[SEP]` – {Headline `[SEP]` Transcript `[SEP]` rationale[5] `[SEP]` sub-rationale}. We also extract embeddings corresponding to two multimodal models. We use VideoCLIP (Xu et al., 2021b) and VLM models (Xu et al., 2021a) that adopt zero-shot transfer learning to video-text understanding

[5]While gold rationales might not be available during inference, our objective to study them as features are to highlight and understand if and how rationales can help improve detection accuracy in this task. We leave automatic prediction of the rationales to future work.

Headlines	ID	Ann.	Rationales	Subrationales
Lester Holt Interrupted Trump Repeatedly	81	M	The headline does not cover all the content of the video	The headline is not providing related evidence for the video
	111	M	Neither of above: The headline provides contradictory information of the video	The headline chooses specific words that cannot be supported as fact
	97	R	-	-
Emily Blunt Weighs In On John Kransinskis Obsession With The D...	42	M	The headline does not cover all the content of the video	The headline chooses specific words that cannot be supported as fact
	45	M	The headline does not cover all the content of the video	Some specific information from the video is not at all reflected in the headline
	97	R	-	-
Did This Man Murder A Beautiful Country Music Producer	77	M	Neither of above: The headline provides contradictory information of the video	The headline is not providing related evidence for the video
	12	M	The headline implies more than what what is introduced in the video	The headline uses an excessively definitive tone when the video is only suggesting the content
	10	M	Neither of above: The headline provides contradictory information of the video	(Free Form Input) No mention of her being a country music producer

Table 3: Examples of Samples with Subjectivity. The second headline shows that each annotator's rationales are different even when the annotations are the same. The third headline shows an example where annotated subrationales all vary in their content (e.g., free-form text). ID is Annotator's ID and Ann. is the annotation result from each annotator (M: Misleading, R: Representative)

tasks. VideoCLIP trains a transformer model using a contrastive objective on paired examples of video-text clips that maximize association between temporarily overlapping text and video segments (Xu et al., 2021b). In contrast, VLM is a task-agnostic multimodal learning model that uses novel masking schemes to improve the learning of multimodal fusion between the text and the video.

We finetune a classification layer that takes input features extracted from video and text-based encoders as described above to predict the label associated with a given video-headline pair. The details of the finetuning experiments are included in Appendix F.

Data and Evaluation Metrics We divide VMH into three sets: 70% for the training set, 15% for the valid set, and 15% for the test set. We evaluate using the following metrics: F1, precision, recall, AUPRC score, and accuracy. We report the precision and recall scores of the positive class, *misleading*. Each metric is estimated by averaging five replicates of stratified random splits.

5 Results

Experiment Results Table 4 reports the main results: the multimodal models that use all the features, {Video Frame + Headline + Transcript + Rationale (V+H+T+R)} result in the best performance across the board, outperforming text-only based model. Adding rationales that provide information about the headline and video relationship improves metrics across the board. F1-scores drop when transcripts are augmented to {Video + Headline} the multimodal models. This could be attributed to the quality of the transcripts automatically extracted from the videos.

In the next section, we perform an analysis to validate the utility of the multimodal features in our dataset in a partial-input setting. Furthermore, we explore how the subjectivity in the task can affect the model detection performance.

Partial Input Analysis Validating a dataset with a partial-input baseline is now important in multimodal domains (Thomason et al., 2019). Artifacts in the dataset can lead the models to *cheat* using shortcut features that can result in poor generalizability (Feng et al., 2019). Thus, in our case, we also experiment with unimodal settings (partial input) — {Video} and {Headline} — to ensure that VMH does not contain such artifacts. The results show that using only video or text-based features result in poor F1-scores ($0.16 - 0.18$) relative to utilizing multimodal features (F1-score: > 0.22).

Model Subjectivity Analysis To understand the subjectivity of the task (Section 3), we also report F1-scores on the subset of the dataset, *subjective* samples (30%), that had low consensus in the annotation process. Training on this subset, even the best model that utilizes all the features: {Video + Headline + Transcript + Rationale} only gets an F1-score of 0.12 and 0.10 with the VideoCLIP and VLM models respectively compared to the F1-

Model	Input	Evaluation Metrics				
		F1-Score	Precision	Recall	AUPRC	Accuracy
BERT	**H**	**0.16 (0.07)**	**0.29 (0.14)**	0.11 (0.05)	**0.17 (0.02)**	**0.82 (0.01)**
	H + T	0.16 (0.08)	0.26 (0.11)	**0.12 (0.06)**	0.15 (0.01)	**0.82 (0.01)**
VideoCLIP	**H**	0.16 (0.06)	0.22 (0.05)	0.13 (0.06)	0.17 (0.01)	0.80 (0.01)
	V	0.17 (0.03)	0.25 (0.06)	0.14 (0.04)	0.16 (0.00)	0.79 (0.02)
	V + H	0.26 (0.09)	0.32 (0.13)	0.24 (0.09)	0.20 (0.04)	0.79 (0.05)
	V + H + T	0.21 (0.04)	0.29 (0.06)	0.17 (0.03)	0.17 (0.01)	0.80 (0.01)
	V + H + T + R	**0.53 (0.06)**	**0.65 (0.08)**	**0.44 (0.06)**	**0.41 (0.05)**	**0.88 (0.01)**
VLM	**H**	0.18 (0.05)	0.20 (0.06)	0.19 (0.09)	0.16 (0.01)	0.76 (0.04)
	V	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.15 (0.00)	0.83 (0.00)
	V + H	0.22 (0.06)	0.23 (0.05)	0.22 (0.06)	0.18 (0.02)	0.77 (0.02)
	V + H + T	0.23 (0.04)	0.23 (0.04)	0.56 (0.01)	0.17 (0.01)	0.76 (0.01)
	V + H + T + R	**0.56 (0.03)**	**0.63 (0.02)**	**0.52 (0.05)**	**0.40 (0.03)**	**0.88 (0.00)**

Table 4: Benchmark Evaluation Results. Rows for each model shows performance with different input features: headlines (**H**), videos (**V**), transcripts (**T**), and rationales (**R**). The reported metrics are the average F1-score, average Precision score, average Recall score, average AUPRC score, and average accuracy score of 5 replicates of stratified random splits of the train, valid, and test sets. The brackets indicate standard deviation for each metric.

scores (i.e., 0.53, 0.56) using the entire training set. The degraded performance suggests that the difficult instances for humans to reach a consensus on might not include any reliable features for the model, indicating that high subjectivity is indeed a factor leading to poor detection.

Video-Text Entailment Analysis We investigate how the misleading headline detection task differs from other video-text entailment tasks by comparing entailment properties and annotations.

We use transcripts as video representation and headlines to predict each sample's entailment relation. We adopt the RoBERTa NLI model[6] to infer the relation between the transcript and the headline. We average the entailment score between chunked sentences from transcripts and the headlines to compromise the different lengths. To calculate if there exists any correlation between entailment predictions and the labels, we conduct a t-test (Gerald, 2018). The p-value is 0.01, which indicates that the difference between the two is statistically significant.

Table 5 shows how entailment decisions contradict the annotator's judgments. For example, the first headline shows a high entailment score with the transcript while annotated as misleading with the rationale of "The headline does not cover all the video content". The second and third headlines are predicted with low entailment scores or "not entail" while being annotated as *representative* by majority annotators.

[6]fine-tuned on SNLI, MNLI, FEVER-NLI, and ANLI

6 Related Work

One of the major factors of misinformation is inaccurate headlines, which pervade social media platforms(Wei and Wan, 2017). Clickbait is characterized by misleading headlines, depending on the degree of deception the audience experiences (Bourgonje et al., 2017). However, clickbait detection problems are distinguished from misleading headlines as they may exaggerate the content but are not particularly misleading (Chen et al., 2015).

As the spread of fake news appears in many forms of multimedia (Aïmeur et al., 2023), several works are on constructing datasets to enable research on multimodal misleading headline detection (Bu et al., 2023). Ha et al. (2018) introduces an image-based dataset and focuses on misrepresented headlines on Instagram. Also, Shang et al. (2019) introduces a dataset of Youtube videos with manual annotations generated by misleading seed videos from the Youtube recommendation system. Zannettou et al. (2018) proposes a misleading-labeling mechanism with both manual and automatic. In this case, annotated videos could be biased as manual and automatic annotation may not be in consensus; they can lead to erroneous annotations of misleading headlines.

Apart from dataset research, previous works focus on detecting multimodal fake news by including multimedia features such as false videos, images, audio, and caption (Qi et al., 2023; Masciari et al., 2020; Demuyakor and Opata, 2022; McCrae et al., 2022). However, these works feature general forms of fake news (i.e., deep-fake videos),

Headlines	Transcripts	Entail	Score	Answer
The sounds of emotions	... We use the principles of music to work with rhythm and melody to regain the functional use of language. Phrase is if we... ...Nice job. Let's all. Well You wanna skip this up? Okay. Do you wanna skip it or singing it? You wanna try to sing it? Let's jump to the chorus. Okay? So darling then. Music is what emotions sound like ...	✓	0.71	M
There is a double standard	... Is there a double standard when it comes to transparency between Trump and Clinton? Well, of course, there's a double standard...He's doing over a hundred foreign deals and he wants to be both the commander chief and the representative in the world for the United States... I mean, the difference between telling somebody you had pneumonia on Sunday instead of Friday is not even in the same league really. ...	✗	0.20	R
Honor a Vet I Warfighters	... Having worked with veterans throughout my career, I know firsthand the importance of honoring our troops. This veterans day our series the war fighters and history are partnering with Team Rub con to create honor event. ...Honor the vets and more fighters in your life, and share a photo and a story today. Learn more history dot com honor that. ...	✓	0.53	R

Table 5: Example of Comparison between Entailment Result and Annotations. The first headline shows high entailment score with the transcript while annotated as *misleading* with the rationale of "The headline does not cover all the content of the video". The second and third headline are predicted with low entailment score or "not entail" while being annotated as "representative" by majority annotators.

not misleading headlines. For multimodal models built for misleading headline detection tasks, Song et al. (2016) identified the video thumbnails, Li et al. (2022) uses uploader and environment features (e.g., number of likes received, the date of most recent upload), Choi and Ko (2022) uses comments and domain knowledge, and Zannettou et al. (2018) uses video's meta statistics (e.g., number of shares) to develop a deep variational autoencoder with semi-supervised learning. Shang et al. (2019) uses a convolutional neural network approach to find the correlation between the neural net features and the headline. You et al. (2023) uses model-selected video frames as input features to the classifier to detect dissimilarity between the video and the text.

7 Conclusion and Future Work

This paper presents VMH, a dataset of misleading headlines from social media videos. Our annotation scheme reduces the task's subjectivity, and we verify the reliability of the annotations. We believe incorporating the crowd workers' distinct opinions (e.g., headline types and rationales) on misleading headlines allows crude reflection of the current social media misinformation phenomenon. Through their lenses, we anticipate a better understanding of how people perceive misinformation in misleading video headlines and for future work, use it to generalize the detection models that are soon to be deployed.

To obtain even more realistic examples for this task, we encourage applying a dynamic adversarial generation pipeline. Motivated by (Eisenschlos et al., 2021; Wallace et al., 2019), misleading headlines could be authored by humans guided to break the existing video headline detection models. For example, while they are writing a "misleading" headline, if the model falsely predicts the headline as "representative", it would become an adversarial, *realistic* example (Ma et al., 2021). These examples can prevent the model from learning superficial patterns (Kiela et al., 2021) and further be developed to become a *robust* tool for journalists to prevent them from making "honest" mistakes when writing video headlines (Dhiman, 2023).

8 Limitations

Although the rationales advance the model's knowledge in detecting misleading headlines, the limitation of this paper is that gold rationales are not realistic. Thus, we suggest using model-generated rationales during inference (e.g., generative multimodal language models) (OpenAI, 2023). Also, conducting a user study to prove whether these rationales led to the correct final prediction will help in assessing the rationale's impact on downstream accuracy. The current rationale setting can be set as an upper bound for the generic model evaluation, including those using model-generated rationales.

9 Ethical Considerations

We address ethical considerations for dataset papers, given that our work proposes a new dataset VMH. We reply to the relevant questions posed in the ACL 2022 Ethics FAQ.[7]

To collect VMH videos, we follow the community guidelines by Facebook by using publicly available videos that are accessible with *public-view only* accounts. Our study was pre-monitored by an official IRB review board to protect the participants' privacy rights. Moreover, the identity characteristics of the participants were self-identified by the workers by answering the survey questions.

Prior to distributing the survey, we collected consent forms for the workers to agree that their answers would be used for academic purposes. The workers in the MTurk Platform are compensated over 10 USD an hour. We targeted a rate higher than the US national minimum wage of 7.50 USD. Even though we understand that VMH may be potentially exploited to make misleading content in the future, we emphasize the scale and the impact of its social goods in that it provides the resource to combat multimodal misinformation online today. As VMH is the first dataset that introduces video for misleading headline detection, we believe it will serve as a starting point in the research community to overcome the task.

References

Esma Aïmeur, Sabrine Amri, and Gilles Brassard. 2023. Fake news, disinformation and misinformation in social media: a review. *Social Network Analysis and Mining*, 13(1):30.

Hunt Allcott, Matthew Gentzkow, and Chuan Yu. 2019. Trends in the diffusion of misinformation on social media. *Research & Politics*, 6(2):2053168019848554.

Peter Bourgonje, Julian Moreno Schneider, and Georg Rehm. 2017. From clickbait to fake news detection: an approach based on detecting the stance of headlines to articles. In *Proceedings of the 2017 EMNLP workshop: natural language processing meets journalism*, pages 84–89.

Eleftheria Briakou and Marine Carpuat. 2020. Detecting fine-grained cross-lingual semantic divergences without supervision by learning to rank. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1563–1580.

[7]https://www.acm.org/code-of-ethics

Yuyan Bu, Qiang Sheng, Juan Cao, Peng Qi, Danding Wang, and Jintao Li. 2023. Online misinformation video detection: A survey. *arXiv e-prints*, pages arXiv–2302.

Jesse Chandler, Pam Mueller, and Gabriele Paolacci. 2014. Nonnaïveté among amazon mechanical turk workers: Consequences and solutions for behavioral researchers. *Behavior research methods*, 46:112–130.

Yimin Chen, Niall J Conroy, and Victoria L Rubin. 2015. Misleading online content: recognizing clickbait as" false news". In *Proceedings of the 2015 ACM on workshop on multimodal deception detection*, pages 15–19.

Sophie Chesney, Maria Liakata, Massimo Poesio, and Matthew Purver. 2017. Incongruent headlines: Yet another way to mislead your readers. In *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*, pages 56–61.

Hyewon Choi and Youngjoong Ko. 2022. Effective fake news video detection using domain knowledge and multimodal data fusion on youtube. *Pattern Recognition Letters*, 154:44–52.

Hal Daume III and Daniel Marcu. 2005. Bayesian summarization at duc and a suggestion for extrinsic evaluation. In *Proceedings of the Document Understanding Conference, DUC-2005, Vancouver, USA*.

John Demuyakor and Edward Martey Opata. 2022. Fake news on social media: Predicting which media format influences fake news most on facebook. *Journal of Intelligent Communication*, 2(1).

Bharat Dhiman. 2023. Does artificial intelligence help journalists: A boon or bane?

Sameer Dhoju, Md Main Uddin Rony, Muhammad Ashad Kabir, and Naeemul Hassan. 2019. Differences in health news from reliable and unreliable media. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 981–987.

Julio Cesar Soares dos Rieis, Fabrício Benevenuto de Souza, Pedro Olmo S Vaz de Melo, Raquel Oliveira Prates, Haewoon Kwak, and Jisun An. 2015. Breaking the news: First impressions matter on online news. In *Ninth International AAAI Conference on Web and Social Media*.

Laura Edelson, Minh-Kha Nguyen, Ian Goldstein, Oana Goga, Damon McCoy, and Tobias Lauinger. 2021. Understanding engagement with us (mis) information news sources on facebook. In *Proceedings of the 21st ACM Internet Measurement Conference*, pages 444–463.

Julian Eisenschlos, Bhuwan Dhingra, Jannis Bulian, Benjamin Börschinger, and Jordan Boyd-Graber. 2021. Fool me twice: Entailment from wikipedia gamification. In *Proceedings of the 2021 Conference of the North American Chapter of the Association* *for Computational Linguistics: Human Language Technologies*, pages 352–365.

Shi Feng, Eric Wallace, and Jordan Boyd-Graber. 2019. Misleading failures of partial-input baselines. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5533–5538.

Banda Gerald. 2018. A brief review of independent, dependent and one sample t-test. *International journal of applied mathematics and theoretical physics*, 4(2):50–54.

Yui Ha, Jeongmin Kim, Donghyeon Won, Meeyoung Cha, and Jungseock Joo. 2018. Characterizing clickbaits on instagram. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with mace. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online. Association for Computational Linguistics.

Xiaojun Li, Xvhao Xiao, Jia Li, Changhua Hu, Junping Yao, and Shaochen Li. 2022. A cnn-based misleading video detection model. *Scientific Reports*, 12(1):6092.

Zhiyi Ma, Kawin Ethayarajh, Tristan Thrush, Somya Jain, Ledell Wu, Robin Jia, Christopher Potts, Adina Williams, and Douwe Kiela. 2021. Dynaboard: An evaluation-as-a-service platform for holistic next-generation benchmarking. *Advances in Neural Information Processing Systems*, 34:10351–10367.

Irene Martín-Morató, Manu Harju, and Annamaria Mesaros. 2021. Crowdsourcing strong labels for sound event detection. In *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 246–250. IEEE.

Elio Masciari, Vincenzo Moscato, Antonio Picariello, and Giancarlo Sperlí. 2020. Detecting fake news by image analysis. In *Proceedings of the 24th symposium on international database engineering & Applications*, pages 1–5.

Scott McCrae, Kehan Wang, and Avideh Zakhor. 2022. Multi-modal semantic inconsistency detection in social media news posts. In *MultiMedia Modeling: 28th International Conference, MMM 2022, Phu Quoc, Vietnam, June 6–10, 2022, Proceedings, Part II*, pages 331–343. Springer.

OpenAI. 2023. Gpt-4 technical report.

Silviu Paun, Bob Carpenter, Jon Chamberlain, Dirk Hovy, Udo Kruschwitz, and Massimo Poesio. 2018. Comparing bayesian models of annotation. *Transactions of the Association for Computational Linguistics*, 6:571–585.

Peng Qi, Yuyan Bu, Juan Cao, Wei Ji, Ruihao Shui, Junbin Xiao, Danding Wang, and Tat-Seng Chua. 2023. Fakesv: A multimodal benchmark with rich social context for fake news detection on short video platforms. In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI.

Md Main Uddin Rony, Naeemul Hassan, and Mohammad Yousuf. 2017. Diving deep into clickbaits: Who use them to what extents in which topics with what effects? In *Proceedings of the 2017 IEEE/ACM international conference on advances in social networks analysis and mining 2017*, pages 232–239.

Mattia Samory, Vartan Kesiz Abnousi, and Tanushree Mitra. 2020. Characterizing the social media news sphere through user co-sharing practices. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 602–613.

Lanyu Shang, Daniel Yue Zhang, Michael Wang, Shuyue Lai, and Dong Wang. 2019. Towards reliable online clickbait video detection: A content-agnostic approach. *Knowledge-Based Systems*, 182:104851.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. 2008. Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 254–263, Honolulu, Hawaii. Association for Computational Linguistics.

Yale Song, Miriam Redi, Jordi Vallmitjana, and Alejandro Jaimes. 2016. To click or not to click: Automatic selection of beautiful thumbnails from videos. In *Proceedings of the 25th ACM international on conference on information and knowledge management*, pages 659–668.

Kate Starbird, Ahmer Arif, and Tom Wilson. 2019. Disinformation as collaborative work: Surfacing the participatory nature of strategic information operations. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–26.

Jesse Thomason, Daniel Gordon, and Yonatan Bisk. 2019. Shifting the baseline: Single modality performance on visual navigation & QA. 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 1977–1983, Minneapolis, Minnesota. Association for Computational Linguistics.

Assaf Toledo, Shai Gretz, Edo Cohen-Karlik, Roni Friedman, Elad Venezian, Dan Lahav, Michal Jacovi, Ranit Aharonov, and Noam Slonim. 2019. Automatic argument quality assessment-new datasets and methods. 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 5625–5635.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *science*, 359(6380):1146–1151.

Jane Wakefield. 2016. Social media 'outstrips tv' as news source for young people. *BBC News*.

Mason Walker and Katerina Eva Matsa. 2021. News consumption across social media in 2021. *Pew Research Center*.

Eric Wallace, Pedro Rodriguez, Shi Feng, Ikuya Yamada, and Jordan Boyd-Graber. 2019. Trick me if you can: Human-in-the-loop generation of adversarial examples for question answering. *Transactions of the Association for Computational Linguistics*, 7:387–401.

Shuting Ada Wang, Min-Seok Pang, and Paul A Pavlou. 2021. Seeing is believing? how including a video in fake news influences users' reporting the fake news to social media platforms. *How Including a Video in Fake News Influences Users' Reporting the Fake News to Social Media Platforms (August 23, 2021)*.

Yuxi Wang, Martin McKee, Aleksandra Torbica, and David Stuckler. 2019. Systematic literature review on the spread of health-related misinformation on social media. *Social science & medicine*, 240:112552.

Wei Wei and Xiaojun Wan. 2017. Learning to identify ambiguous and misleading news headlines. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 4172–4178.

Hu Xu, Gargi Ghosh, Po-Yao Huang, Prahal Arora, Masoumeh Aminzadeh, Christoph Feichtenhofer, Florian Metze, and Luke Zettlemoyer. 2021a. Vlm: Task-agnostic video-language model pre-training for video understanding. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4227–4239.

Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. 2021b. Videoclip: Contrastive pre-training for zero-shot video-text understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6787–6800.

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. Wikiqa: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2013–2018.

Jinpeng You, Yanghao Lin, Dazhen Lin, and Donglin Cao. 2023. Video rumor classification based on multi-modal theme and keyframe fusion. In *Computer Supported Cooperative Work and Social Computing: 17th CCF Conference, ChineseCSCW 2022, Taiyuan, China, November 25–27, 2022, Revised Selected Papers, Part I*, pages 58–72. Springer.

Omar Zaidan, Jason Eisner, and Christine Piatko. 2007. Using "annotator rationales" to improve machine learning for text categorization. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 260–267, Rochester, New York. Association for Computational Linguistics.

Savvas Zannettou, Sotirios Chatzis, Kostantinos Papadamou, and Michael Sirivianos. 2018. The good, the bad and the bait: Detecting and characterizing clickbait on youtube. In *2018 IEEE Security and Privacy Workshops (SPW)*, pages 63–69. IEEE.

Melissa Zimdars. 2016. My 'fake news list' went viral. but made-up stories are only part of the problem. *The Washington Post*.

A Selection of Venues

We selected videos introduced by Rony et al. (2017) where the videos were created by mainstream media consisting of 25 most circulated print media and 43 most-watched broadcast media , and unreliable media cross-checked by two sources, information-beautiful[8] and Zimdars (2016) in the US. These were selected to include a broad range of media outlets that may include misinformation.

B Annotation Task

Example of Pilot Study As demonstrated in Figure 5, our pilot study revealed that asking one question whether the video headline represented the video caused much confusion around the word *represents*, making it too ambiguous for the workers to answer the question properly. After a few interactions with workers, we found that this was due to the inherent subjectivity of the *Misleading Video Headline Detection* Task.

Please play the following video.

Do you think that the video headline,
"**US-led Airstrike on ISIS Car Bomb Factory in Iraq**",
represents the video content?

A. Not Representative
B. Somewhat Representative
C. Mostly Representative
D. Absolutely Representative

Figure 5: Example of Pilot Study. The word "represents" was too ambiguous for the audience, causing the annotators to interpret the task differently; thus it was difficult for them to consider the misleadingness of a headline.

[8] Unreliable Fake News Sites

C Questions for Headline Property

We found out from a preliminary survey that merely asking a question, *how well do you think the video headline represents the video content* causes confusion among workers due to the question's inherent subjectivity. We assume that for different types of headlines, people follow different cognitive processes when assessing the headline's misleadingness. Thus, we first assess the properties of the headline and ask the following questions. Examples are in Table 6 and Table 7.

Opinionated Statement If the worker chooses that a given headline is a *opinionated statement*, the consecutive question would be *Do you have prior knowledge about the statement in the headline to make a judgment on the statement?* to assess their original opinion stated in the headline. After watching the video, the workers are asked **Assuming that the information provided by the video is correct, how would you rate the following statement? The video justifies the opinion in the headline.** This question specifically asks to find the congruence between the video's message and the opinion stated in the headline. If the worker finds the video content appropriate enough to match the headline, they are expected to select *Agree*. Then we conclude that the final label of the video headline is *representative*.

Neither Statement If the worker chooses that a given headline is a *neither statement*, the consecutive question would be *Write down what you expect to see in a video* to assess their background knowledge about the headline and what they expect to see in the video. After watching the video, the workers are asked **Assuming that the information provided by the video is correct, how would you rate the following statement? The video talks about the video.** This question specifically asks to find the congruence between the video's message and the information in the headline. If the worker finds the video content appropriate enough to match the headline, they are expected to select *Agree*. Then we conclude that the final label of the video headline is *representative*.

Factual/Opinionated Question If the worker chooses that a given headline is in the form of *question*, he would be asked the same questions for both factual and opinionated questions. Before watching the video, the consecutive question would be *Write down what you expect to see in a video* to

Factual Statement	Opinionated Statement	Neither Statement
Biden was not elected in 2020	Best ways to make oatmeal (The word ‘best’ is open to interpretation)	Great Depression
Trump has 10 children	The power of healthy food (The word ‘healthy’ is open to interpretation)	Make your own coconut milk
She provided tips for making oatmeal	Vulgar language from Trump (The word ‘vulgar’ is open to interpretation)	Tips for making oatmeal
Trump to Biden: ’You’re the Puppet’	5 minutes of truth (The word ‘truth’ may imply different things depending on your experience)	Trump’s wife

Table 6: Examples for Selecting Statement Headline Categories

Factual Question	Opinionated Question
Did Trump win the election?	VP debate: Do you want a “you’re hired" president? (The question is asking for your personal preference)
When were the first automobiles invented?	What started the French revolution? (The question is asking something that is open to different interpretations)
Do you check the temperature every day?	What if I made you eat worms? (The question is asking for your personal preference)

Table 7: Annotators are given five headline properties to choose what kind of sentence headline is.

Original Headline	Synthesized Headlines	Groundings
This woman takes some of the most dangerous selfies in the world	This man takes some of the most dangerous selfies in the world	False (because it is a “woman" not a man who is taking selfies in the video)
Baby Girl Gets Adorably Upset When Parents Kiss In Front Of Her	Baby Boy Gets Adorably When Parents Kiss In Front Of Him	False (because it is a “girl" not a boy who cries in the video)
Trump to Clinton: ’You’re the Puppet’	Trump to Biden: ’You’re the Puppet’	False (because It is “Clinton" not Biden that counters Trump in the video)
Toyota created a mini robot companion	Honda created a mini robot companion	False (because It is “Toyota" not Honda mentioned in the video)

Table 8: **Examples of Synthesized Headlines for Accuracy-check Questions**

assess their background knowledge about the headline and what they expect to see in the video. After watching the video, the workers are asked **Assuming that the information provided by the video is correct, how would you rate the following statement? The information provided by the video helps you answer the question in the headline.** This question specifically asks to find an answer to the question in the headline, assuming that video content is expected to contain the information that the headline is inquiring about. If the worker decides that the video content cannot answer or has insufficient information, they are expected to select *Disagree*. Then we conclude that the final label of the video headline is *misleading*.

D Quality Control and Assessment

Pre-qualification Test We restrict this task to the workers in the United States given that they have a higher possibility of being fluent in the verbal and literal understanding of English. Before the workers participate in the HIT, we prepare a preliminary qualification test that the workers must pass to start the HIT. All the participants must take this pre-qualification test, given multi-choice questions such as “How *representative* is the video?” and “How would you rewrite the headline.” When they receive a score of 100, they are qualified to participate in the following HITs. This process is included to ensure that the participants have the capacity to integratively comprehend the video content and video headline, and then draw out an accurate video label.

Synthesized Headlines in Accuracy Check Questions Table 8 shows examples of synthesized headlines in accuracy check questions. Accuracy check questions that are synthetically created to be always misleading (obviously false). For each annotator, we calculate the ratio between the number of correct answers and the number of accuracy check questions to select competent annotators.

MACE We compute MACE, a Bayesian approach-based metric that takes into account the credibility of the annotator and their spamming preference (Hovy et al., 2013).

$$
\begin{aligned}
&\text{for } i = 1, \cdots, N: \\
&\quad T_i \sim \text{Uniform} \\
&\quad \text{for } j = 1, \cdots, M: \\
&\quad\quad S_{ij} \sim \text{Bernoulli}(1 - \theta_j) \\
&\quad\quad \text{if } S_{ij} = 0: \\
&\quad\quad\quad A_{ij} = T_i \\
&\quad\quad \text{else}: \\
&\quad\quad\quad A_{ij} \sim \text{Multinomial}(\xi_j),
\end{aligned}
$$

where N denotes the number of headlines, T denotes the number of the true labels, and M denotes the number of workers. S_{ij} denotes the spam indicator of worker j for annotating headline i, while A_{ij} denotes the annotation of worker j for headline i. θ and ξ each denotes the parameter of worker j's trustworthiness and spam pattern. We add Beta and Dirichlet priors on θ and ξ respectively. The assumption in the generative process is that an annotator always produces the correct label when he does not show a spam pattern which helps in excluding the labels that are not correlated with the correct label. Here, our parameter of interest is θ which stands for the trustworthiness of each worker. We apply Paun et al. (2018)'s implementation to obtain posterior distributions (samples) of θ and calculate point estimates.

E Other Feature Distribution

The venue kind *Website* show higher percentage (29%) of creating misleading headlines (Table 9). On the other hand, because the proportions of misleading headlines are fairly uniform in the 1) proportions of news topics, 2) headline properties, and 3) venue credibility, it suggests that the three features are less prone to be an indicator for misleading headlines (The proportions of each label in the three features are reported in Table 10, 11 and 12).

Venue Kind	Annotated Labels	
	Representative	Misleading
Broadcast	0.85	0.15
Cable	0.85	0.15
Newspaper	0.87	0.13
Website	0.71	0.29

Table 9: *Website* shows more proportion of creating misleading headlines than other categories in the venue kind feature, which suggests that venue kind feature may be an indicator of representativeness of a headline.

Headline Topics	Annotated Labels	
	Representative	Misleading
Entertainment	0.86	0.14
Food	0.86	0.14
Others	0.81	0.19
Politics	0.85	0.15

Table 10: There was no significant difference in the proportions of topics, which suggests that topic feature is not strong indicator for misleadingness.

Headline Properties	Annotated Labels	
	Representative	Misleading
Factual Statement	0.86	0.14
Opinionated Statement	0.84	0.16
Neither Statement	0.83	0.17
Factual Question	0.81	0.19
Opinionated Question	0.72	0.28

Table 11: There was no significant difference in the proportions of properties, which suggests that property feature is not strong indicator for misleadingness.

F Finetuning Details of Baseline Models

We finetune both VideoCLIP and VLM on a A6000 GPU using the Adam optimizer with a learning rate 0.00002, weight decay ratio of 0.001, and batch size 8 for 10 epochs. For text encoders and video encoders, we directly use the best checkpoints from the pretrained VideoCLIP and VLM models. We concatenate encoder outputs, the pooled video and

Venue Credibility	Annotated Labels	
	Representative	Misleading
High	0.86	0.14
Mostly Factual	0.84	0.16
Mixed	0.85	0.15
Low	0.81	0.19
Unknown	0.85	0.15

Table 12: There was no significant difference in the proportions of properties, which suggests that the headline property feature is not strong indicator for misleadingness.

text features, and learn fully connected layer optimized with Cross Entropy loss. For partial input experiments, we assign zeros to text or video encoder inputs.

G Era of Fake News

People have been using social media platforms to converse, diffuse and broadcast their ideas in recent years. However, there has been widespread concern that misinformation is increasing on social media, which causes damage to societies (Allcott et al., 2019). Some contemporary commentators even describe the current period as "an era of fake news" (Wang et al., 2019).

H Censoring Audio Transcripts

We outsource transcript extractions from a software called Deepgram.[9] To validate its accuracy, we randomly sampled 55 videos that have transcripts and manually checked if the transcripts were accurate. These transcripts exactly matched the audio from the videos. VMH also includes transcript information on the timeframe that indicates when each word starts and ends in the video with its confidence score. We especially paid attention to this information when censoring the transcripts.

[9]https://deepgram.com/