

M4Health: A Multi-Modal, Multi-Domain, Multi-Platform, and Multi-Task Benchmark for Video-Driven Health Communication on Social Media

Raihana Zahra¹ Zhenghao Gong³ Owen Dewing¹ Nicholas Aurino¹ Thomas Rife¹ Cyrus Nikzad¹
Daniel Rowe¹ Junyuan Lin² Lanyu Shang¹

¹ Computer Science, Loyola Marymount University

² Mathematics, Statistics and Data Science, Loyola Marymount University

³ Halicioğlu Data Science Institute, University of California, San Diego

{rzahra, odewing, naurino, trife, cnikzad, drowe7}@lion.lmu.edu, z3gong@ucsd.edu, {junyuan.lin, lanyu.shang}@lmu.edu

Abstract

Short-video platforms have transformed how the public consumes health information on the web and social media where uncurated content poses significant risks to vulnerable populations. While prior work has primarily focused on text-only or single-platform analyses, comprehensive benchmarks for multi-modal health communication in short videos remain limited. In this paper, we introduce M4Health, a multi-modal, multi-domain, multi-platform, and multi-task benchmark for health communication in short videos. M4Health comprises 669,995 videos from TikTok, YouTube Shorts, and Reddit, spanning diverse health domains such as nutrition, fitness, mental health, and wellness. We provide expert annotations for a subset of videos across three interrelated tasks, including credibility assessment, AI-generation detection, and theme classification. Extensive benchmarking experiments show that current state-of-the-art models, including task-specific approaches and large vision-language models (LVLMs), achieve suboptimal performance. We share the M4Health dataset with research communities to foster collaborative research toward supporting informed health decision-making on the web and social media.

Introduction

With the increased popularity of short-video platforms (e.g., TikTok, YouTube Shorts), social and online media have fundamentally transformed how the public consumes and engages with health information (Anderer 2024). Recent studies reveal that over 92% of young users on TikTok, a leading short-form video platform with more than 1.6 billion active users globally, have been exposed to health information, with more than half actively seeking health-related content from such platforms (Kirkpatrick and Lawrie 2024). However, the uncurated dissemination of health information can lead to the rapid spread of misinformation that disproportionately impacts vulnerable populations, including those with limited health literacy and from underserved communities (Anderer 2024). Recent research shows that nearly half of health-related short videos contain non-factual information, which greatly threaten public health and undermine trust in medical institutions (Dimitroyannis et al. 2024). More recently, the proliferation of generative AI techniques

has further exacerbated this challenge by enabling the creation of increasingly convincing synthetic content (Migisha and Hagström 2025). Despite the urgency of these issues, the research community lacks comprehensive benchmarks that characterize the complexities of health communication in short-form videos. To bridge this gap, this paper introduces M4Health, a Multi-Modal, Multi-Domain, Multi-Platform, and Multi-Task benchmark to advance computational approaches for safeguarding health communication on the web and social media.

Current studies on online health information primarily focus on text-only or text-image content from traditional online or social media platforms (e.g., Twitter/X, Facebook), which are insufficient to capture the complex multi-modal dynamics of short-form videos that often combine visual storytelling, audio narratives, background music, on-screen text overlays, and advanced editing techniques that fundamentally shape information perception and persuasion (Bu et al. 2023). A few recent works have begun exploring health-related content on short-video platforms (Kirkpatrick and Lawrie 2024; Shang et al. 2025), but these efforts are often limited to narrow domains (e.g., COVID-19 (Shang et al. 2021), cancer (Zhang et al. 2024)), individual task (e.g., misinformation detection (Shang et al. 2025) or sentiment analysis (Thakur et al. 2024)), or single platform analyses that undermine cross-platform generalizability and domain adaptability (Fang et al. 2024). More importantly, recent advances of generative AI enable the creation of highly realistic synthetic content that humans struggle to identify, which makes AI-generation detection increasingly critical yet challenging, especially in critical domains like health communication. Hence, there remains an urgent need for a comprehensive benchmark that systematically captures the complexity of health communication in short videos across diverse platforms, domains, and analytical tasks (e.g., credibility assessment, AI-generation detection).

Motivated by the above knowledge gap, this paper presents M4Health, the first large-scale benchmark designed to comprehensively characterize health communication in short videos. In particular, to ensure comprehensive domain coverage, M4Health adopts 130 professionally curated keywords, covering diverse health domains such as fitness, nutrition, mental health, and wellness. These keywords are used to retrieve video-based posts, including the video it-

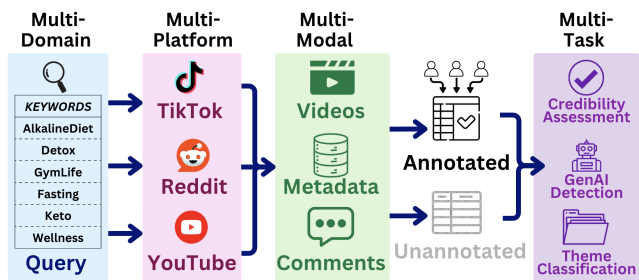


Figure 1: Overview of M4Health. The figure illustrates the M4Health data pipeline, from collection to analysis. A set of health-related keywords is used to query online content from three platforms (i.e., Reddit, TikTok, and YouTube Shorts). A subset is manually annotated, while the rest supports unsupervised or semi-supervised training. M4Health also supports downstream tasks of credibility assessment, generative AI detection, and theme classification.

self, associated metadata, and user comments, across three major social media platforms, namely TikTok, YouTube Shorts, and Reddit. To support diverse research paradigms, M4Health provides both annotated and unannotated partitions. A carefully sampled subset undergoes expert annotation for credibility assessment, AI-generation detection, and theme classification, while the remaining unannotated data enables exploration of unsupervised and weakly supervised approaches. Together, these components establish M4Health as a comprehensive benchmark for studying multi-modal health communication on the web and social media.

Beyond dataset construction and annotation, we conduct extensive experiments to benchmark state-of-the-art models on each task within M4Health, including task-specific models (i.e., credibility assessment, AI generation detection, theme classification), and recent large vision-language models (LVLMs). Our preliminary results reveal suboptimal performance across all three tasks, with particularly notable gaps in platform and task generalizability. Interestingly, we also find that LVLMs often underperform task-specific models despite their superior capabilities on general multi-modal benchmarks. These findings suggest the unique challenges posed by health-related short videos and highlight the pressing need for dedicated computational approaches tailored to multi-modal health communication.

M4Health serves as a valuable resource enabling impact across multiple research communities. For health informatics, M4Health provides large-scale, systematically collected data to study health information dissemination patterns and user engagement dynamics. For NLP and computer vision, M4Health introduces challenging tasks requiring reasoning over visual, audio, and textual modalities and their temporal interactions. AI-generation detection further positions M4Health at the frontier of automated content moderation, especially in critical health-related domains. Finally, our multi-platform design enables investigations into platform-specific content characteristics and moderation effectiveness. By releasing M4Health, we aim to foster collaborative efforts toward safeguarding online health communication.

Related Work

Health Communication on the Web

Health communication on the web has become increasingly popular for people to disseminate and consume health-related information, such as health guidance, medical advice, and wellness recommendations (Kirkpatrick and Lawrie 2024). However, the uncurated nature of online health communication poses significant risks, particularly for vulnerable populations such as individuals with limited health literacy, underserved communities, and young users who may lack the critical evaluation skills to assess content credibility (Zhao, Zhao, and Song 2022). Many efforts have been made to address these challenges through computational approaches (Schlicht et al. 2024). For example, recent work has explored health-related misinformation detection (Cui and Lee 2020), health information dissemination patterns (Le, Hoang, and Pham 2023), and content quality assessments (Isgut et al. 2022). However, existing studies primarily focus on individual analytical tasks which overlook that multiple aspects of health communication are often deeply interconnected (Li 2025). To address this limitation, this paper introduces M4Health, a comprehensive benchmark that addresses these limitations by providing multi-platform, multi-domain, multi-modal data with annotations for interconnected tasks including credibility assessment, AI-generation detection, and theme classification.

Multimodal Web Content

Multimodal media has become a vital way for web users to share personal and informational content about themselves or the world around them (Stepaniuk 2024). Platforms like TikTok and YouTube Shorts have changed the way that people communicate online and on social media (Violot et al. 2024). Compared to text-only content, video and image-based posts drive significantly more engagement and interaction (Abbas, Abbas, and Umrani 2024). While the multimodal features (e.g., visual elements, audio narratives, text overlays) enhance user engagement and promote persuasive storytelling, they pose unique challenges for computational content analysis where traditional text-based or single-modality approaches fail to capture the intricate interplay between different modalities (Stepaniuk 2024). Recent advances in large vision-language models (LVLMs) have shown promise in understanding multimodal content (Xuan et al. 2024). However, LVLMs exhibit known limitations in reasoning over conflicting visual inputs, and visual reasoning tasks are more prone to memorization errors than text-based tasks, motivating the need for domain-specific evaluation benchmarks (Kirkpatrick and Lawrie 2024; Carragher et al. 2025b,a). Moreover, the proliferation of generative AI technologies has enabled the creation of realistic synthetic videos, introducing critical challenges in distinguishing authentic content from AI-generated media, particularly in high-stakes domains like health (Stepaniuk 2024). This paper not only presents a comprehensive dataset for health communication, but also conduct extensive evaluations of state-of-the-art task-specific models and LVLMs to establish baseline performance for future research.

Dataset	Video-Based	Number of Platforms	Sample Size	Total Labels	Tasks	Topics	Keyword Count
FakeSV (Qi et al. 2023)	✓	1	5, 538	5, 538	Fake news detection	General news	–
COVID-VTS (Liu et al. 2023)	✓	1	10, 000	0	Fact verification	Covid-19	26
Med-MMHL (Sun et al. 2023)	✗	1	14, 665	0	Fake news detection	Medical	–
HPV (Massey et al. 2016)	✗	1	193, 379	4, 410	Sentiment, side effect, prevention	Human Papillomavirus	10
PHAD (Chappa et al. 2024)	✓	2	5, 730	0	Topic classification	Tobacco usage	95
M4Health (Ours)	✓	3	669, 995	9, 000	Credibility assessment, AI generation, theme	Diverse health domains	130

Table 1: Comparison of existing datasets

Health Communication Datasets

Existing health communication dataset primarily focuses on text-only (Cui and Lee 2020; Zhou et al. 2020) or text-image data (Sun et al. 2023; Li et al. 2020). Despite the growing importance of video-driven health communication on the web, existing datasets remain limited in scope and coverage. A few recent efforts have been made to collect and analyze video content on the web. We summarize the key characteristics of existing health-related public multimodal dataset in Table 1. In particular, existing datasets are typically limited to a single platform (e.g., Twitter/X (Sun et al. 2023; Massey et al. 2016; Liu et al. 2023)), constraint to a specific domain (e.g., COVID (Liu et al. 2023), Tobacco (Chappa et al. 2024)), or designed for a particular task (e.g., fake news detection (Qi et al. 2023), topic classification (Chappa et al. 2024)). Moreover, existing datasets rarely provide annotations for multiple interconnected tasks, which prevents the comprehensive analysis of how credibility, content authenticity, and thematic characteristics that can jointly inform effective content moderation solutions to support informed decision-making. To address these limitations, M4Health provides a large-scale benchmark that spans three major platforms, covers a broad spectrum of health-related topics, and offers human annotations for three complementary tasks to holistically study multimodal health communication.

Data Collection

M4Health aims to comprehensively capture video-driven health communication on the web and social media. We focus on three prominent web platforms that are popular for video content, including *TikTok* (TikTok 2025), *YouTube Shorts* (YouTube 2025), and *Reddit* (Reddit 2025). We compiled a list of 130 keywords across 11 health-related domains, spanning physical health conditions, mental health, nutrition, fitness, and medical treatments (see Appendix). Using these keywords as search queries, we systematically collect publicly available video posts from each platform. Our data collection period spans the entire year of 2025 and will continue as an ongoing effort to expand the dataset

and capture evolving trends of health communication. We acknowledge that keyword-driven retrieval may miss algorithmic (i.e., community-coined euphemisms used to evade platform moderation) (Steen, Yurechko, and Klug 2023), which we identify as a direction for future work. To maintain consistency and facilitate downstream analysis, we focus primarily on English-speaking videos. Detailed descriptions of the data collection pipeline for each platform are elaborated below. Under fair use provisions for academic research, we downloaded publicly available video files for data analysis and benchmark evaluation purposes only. The videos were used exclusively for non-commercial research and annotation, and no content will be redistributed.¹

TikTok

Data was collected from TikTok using the official TikTok Research API (TikTok 2026). For each keyword, we queried the API to retrieve all videos containing the keyword in either the video description or associated hashtags. We further cleaned the dataset by removing duplicate entries and videos that were no longer publicly accessible. Metadata, including video ID, video URL, description, like count, and view count, was retained. For photo slideshows (i.e., posts containing static images accompanied by a soundtrack), we convert them into videos by stitching the images into a continuous video sequence synchronized with the original audio track.

YouTube Shorts

We collected posts from YouTube Shorts using a two-step process. First, we queried the YouTube Data API (Google 2026) with each keyword to retrieve the video IDs and associated metadata of relevant videos. Specifically, the key fields we collected are the Video ID, Title, Description, Channel, Published Date, and URL. Duplicate videos were removed based on keyword and unique video ID to ensure videos appearing across multiple keyword queries were only

¹The M4Health dataset is available at: <https://doi.org/10.5281/zenodo.18265458>.

included once. For research and data analysis purposes, we also downloaded the video files using pytube (pytube 2026) and YT-DLP (YT-DLP 2026), with pytube handling the primary downloads and YT-DLP used as a backup for videos that failed to download with pytube.

Reddit

We collected video-based posts from Reddit using the PRAW library (PRAW 2026). Unlike TikTok and YouTube Shorts, Reddit organizes content into topic-specific communities called subreddits. We identified 51 subreddits thematically aligned with the 130 health-related keywords used for the other platforms to ensure comparable coverage of health communication topics. Due to the relatively lower volume of video content on Reddit compared to TikTok and YouTube Shorts, we collected all readily accessible posts with videos returned through multiple Reddit sorting methods rather than limiting to the first three quarters of 2025. This extended temporal scope not only ensures a sufficient sample size for analysis but also enables longitudinal studies examining how health communication on Reddit evolves over time. In particular, we retrieved posts from each identified subreddit and filtered the posts according to whether there were downloadable Reddit-hosted videos. The video files along with their metadata (e.g., post title, subreddit name, upvote count, comment count) and user comments were saved for subsequent analysis and benchmark task evaluation.

Annotation

We annotate a subset of the video posts collected from each platform to better understand multimodal health communication patterns on the web and facilitate multi-task evaluation of downstream content analysis models. In particular, we focus on three tasks that reflect complementary and interconnected dimensions of health content integrity: *credibility assessment* addresses factual accuracy, *AI-generation detection* addresses content authenticity, and *theme classification* captures communicative intent. To ensure annotation quality, we employed a two-step process. First, we conducted a pilot study by sampling 100 videos from each platform and having three independent expert annotators manually label each video across all annotation dimensions. The pilot study confirmed a substantial inter-annotator agreement (i.e., average Fleiss' Kappa score > 0.6) (McHugh 2012), with all videos receiving majority consensus for each multi-class annotation task, thereby validating our annotation schema and guidelines. Following the pilot study, we proceeded to the full annotation phase, where the annotators labeled an additional 900 videos from each platform, resulting in a total of 1,000 videos per platform. Final labels were determined through majority voting across the three annotators. The detailed annotation guidelines for each task are described below.

Definition 1 Credibility: The credibility annotation assesses the factual accuracy of health-related claims presented in each video. We consider three classes for credibility assessment.

- **Credible:** The video contains information that is factually accurate and verifiable based on credible sources, with no significant inaccuracies or misleading claims.
- **Mixed:** The video includes a combination of accurate and inaccurate information. It may present facts alongside speculation, exaggeration, or misinformation.
- **Not Credible:** The video contains claims that are demonstrably false, misleading, or fabricated, and cannot be supported by credible evidence.

Definition 2 AI Generation: The AI generation annotation identifies the degree to which AI was involved in creating the video content. We consider three classes based on the level of AI involvement.

- **Authentic:** The video appears to be a raw, unedited recording of a real-world event or person. There is minimal or no post-production, and no signs of synthetic media or AI involvement.
- **Enhanced:** The video is based on authentic footage but includes noticeable post-production elements such as added text, graphics, voiceovers, filters, or editing. There is no evidence that the content was generated by AI.
- **Generated:** The video content is entirely or predominantly created using artificial intelligence. This includes AI-generated visuals, avatars, voice synthesis, or deepfake technology, with little to no real-world footage.

Definition 3 Theme: Each video in the dataset was annotated with one primary theme based on its dominant and supporting communication strategies. We focus on four major categories.

- **Factual:** The video primarily conveys objective information or instructive content. This includes tutorials, how-to guides, educational explanations, and demonstrations (e.g., recipes, workouts, or science facts).
- **Opinion:** The speaker expresses personal views, beliefs, or interpretations of a topic. These videos may include commentary, reviews, or anecdotal claims that are not supported by verifiable evidence.
- **Entertainment:** The video uses humor, satire, irony, exaggeration, or parody to critique or comment on a subject. The content is often not meant to be taken literally and may mimic factual or opinion-based formats for comedic or critical effect.
- **Persuasion:** The video is intended to influence the viewer's attitudes or behaviors. This includes promotional content, advertisements, or endorsements, particularly when the creator has an interest in the product, service, or idea being promoted.

Preliminary Analysis

Statistical Summary

Table 2 presents the summary statistics of the M4Health dataset. We collected a total of 669,995 videos across the three platforms, with TikTok contributing the largest share (556,021), followed by YouTube Shorts (112,530) and Reddit (1,444). Among these videos, 1,000 from each platform were annotated for multi-task evaluation. To summarize the

		TikTok	YouTube Shorts	Reddit
Video Post	Number of Collected Videos	556, 021	112, 530	1, 444
	Date Range	Jan 1 to Dec 30, 2025	Jan 1 to Dec 31, 2025	March 2, 2005 to Dec 31, 2025
	Number Keywords/Subreddits	130	130	51
	Number of Annotated Videos	1, 000	1, 000	1, 000
	Average Number of Comments	1.46	0.68	126.47
	Average Number of Likes	19.30	19.24	2, 932.45
	Average Number of Views*	443.87	933.07	N/A **
User	Number of Unique Users/Channels	334, 177	59, 119	446

*Number of views as of the data collection date of Dec 31, 2025.

**Number of views on Reddit is not retrievable due to the platform’s data access restrictions.

Table 2: Dataset summary

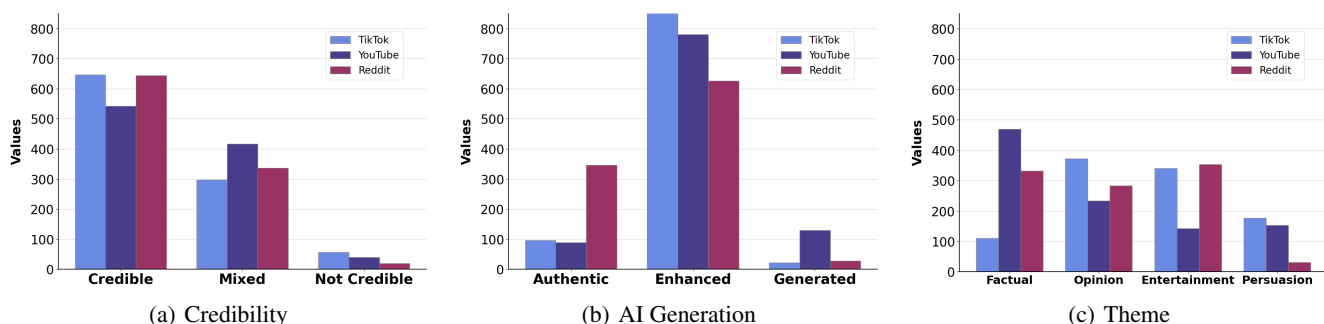


Figure 2: Label distribution of each task

engagement metrics, we randomly sampled 500 records per platform from the dataset after removing missing values and outliers using the interquartile range (IQR) method. We observe that YouTube Shorts exhibits the highest user engagement with an average of 933.07 views and 19.24 likes per video, while Reddit shows the highest average comment count of 126.47, reflecting its discussion-oriented nature.

We analyze the label distribution of the annotated videos across the three annotation tasks. A summary of the annotation distributions is shown in Figure 2. We observe that, while only a small proportion of videos are labeled as *not credible* (5.7%, 4.0%, and 1.9% for TikTok, YouTube Shorts, and Reddit, respectively), a nontrivial number of videos fall into the *mixed* category, i.e., 29.7%, 41.7%, and 33.7%, respectively. This suggests that health misinformation on social media often manifests not as entirely fabricated claims, but as content that mixes accurate information with speculation, exaggeration, or unverified claims, making it challenging to detect and potentially more persuasive.

For AI generation, *enhanced* videos dominate across all platforms, which indicates that most health-related content creators favor post-production elements. For theme classification, platform-specific patterns emerge. For example, TikTok features more *opinion* and *entertainment* content, YouTube Shorts contains predominantly *factual* content, and

Reddit shows a balanced distribution between *entertainment* and *factual* themes. Such platform-specific patterns may pose challenges to model generalizability, as classifiers trained on data from one platform might not transfer well to others due to differences in content and themes.

Topic Analysis

Word Cloud We present a summary of the textual content of the video posts in M4Health, including the distribution of high frequency keywords across platforms. In particular, Figure 3 shows the word cloud of the most frequent words in each platform. We observe that while all three platforms share common health-related terms such as “health”, “workout”, “weight”, and “diet”, each platform exhibits distinct linguistic characteristics. For instance, short-form platforms like TikTok and YouTube Shorts surface more engagement oriented terms such as “gymtok” and “shorts”, consistent with algorithmic mechanisms that prioritize virality and broad reach. In contrast, platforms with older or more information-seeking audiences, such as Reddit, highlight practical wellness concepts like “program”, “routine”, and “journey”, which align with the more personal anecdotes and self-optimization content. These differences suggest that while the core health themes remain consistent, each platform amplifies particular patterns of health dis-

Credibility Assessment

- FakeSV (Qi et al. 2023): a multi-modal fusion model which uses pretrained audio, visual, motion, and textual representations to detect misinformation in short videos.
- TikTec (Shang et al. 2021): a multi-modal misinformation detection framework that integrates visual appearance, acoustic signals, motion dynamics, and textual cues through hierarchical feature fusion to classify deceptive TikTok short videos.

AI Generation Detection

- Cakelens-v5 (Launch Platform 2025): a CNN-based video classifier that jointly model the spatial-temporal relations of video frames to detect AI generation.
- SimpleSmallPatch (Chen, Yao, and Niu 2024): a frame-level detector that classifies AI-generated content using a lightweight ResNet backbone with noise-residual filters.

Theme Classification

- ViFi-CLIP (Rasheed et al. 2023): a video-adapted variant of CLIP that fine-tuned on Kinetics 400 dataset.
- Swin3D (Yang et al. 2025): a 3D transformer model that learns 3D representations with hierarchical features.

Large Vision-Language Models We also study the performance of a set of open-source large vision-language models (LVLMs) under zero-shot and few-shot settings.

- LLaVA-NeXT-7B (Li et al. 2024): an open-source large multi-modal model that couples a vision encoder with a LLaMA-based language model and instruction tuning to perform high-resolution image and video understanding.
- Qwen2.5-VL-7B (Bai et al. 2025): a vision-language LLM built on the Qwen2.5 backbone, supporting high-resolution, multi-image, and dense OCR-style perception for general-purpose multi-modal reasoning.
- InternVL2.5-8B (Chen et al. 2024): a strong vision-language model that integrates a high-capacity visual encoder with a unified language backbone for dense captioning, and multi-image reasoning.

Experimental Settings

To ensure a fair comparison, we use the same input to each task-specific baseline model and follow the preprocessing steps specified in the original papers. Moreover, we conducted extensive hyperparameter tuning for all task-specific baselines. The details about the preprocessing steps, hyperparameter settings, prompts for LVLM baselines are provided in *Appendix*. We run the experiments on Ubuntu 24.04 with 4 Nvidia RTX 6000Ada GPUs. We implement the baselines using Python 3.12 and PyTorch 2.9.0 with CUDA 12.8.

Results

We evaluate benchmark performance using standard metrics for multi-class classification, including *accuracy*, *precision* (*macro*), *recall* (*macro*), and *F1 score* (*macro*). We report the evaluation results based on 5-fold cross-validation. Experiment results for credibility assessment, AI generation

detection, and theme classification are shown in Tables 3, 4, and 5, respectively. We observe that task-specific misinformation detection models (i.e., FakeSV, TikTec) outperform LVLMs on the credibility assessment task. In particular, FakeSV achieves the highest accuracy across all platforms and outperforms the best-performing LVLM baseline (i.e., LLaVA-NeXT-7B-Zero) by 18.8%, 10.3%, and 16.3% on TikTok, YouTube Shorts, and Reddit, respectively. The poor performance of LVLMs indicates that credibility assessment in health communication requires domain-specific multi-modal fusion rather than general-purpose reasoning.

For AI generation detection, we note that specialized detectors (i.e., Cakelens-v5, SimpleSmallPatch) substantially outperform LVLMs. Specifically, SimpleSmallPatch achieves the highest accuracy on all three platforms, outperforming the best LVLM baseline (i.e., InternVL2.5-8B-Few) by 21.4%, 7.7%, and 2.1% on TikTok, YouTube Shorts, and Reddit, respectively. LLaVA-NeXT variants perform particularly poorly, suggesting that current LVLMs lack the ability to detect subtle AI-generated artifacts in health-related videos. Interestingly, for theme classification, all methods struggle on this task. InternVL2.5-8B-Few achieves the best accuracy on Reddit and TikTok while ViFi-CLIP performs best on YouTube. This suggests that distinguishing fine-grained health communication themes remains challenging for both specialized and general-purpose models, possibly due to the semantic overlap among health topics and the diversity of content styles across platforms. We also observe that LLaVA-NeXT-7B achieves better zero-shot than few-shot performance across all tasks, suggesting that the smaller model struggles with in-context learning from multi-modal demonstrations and instead overfits to superficial patterns.

Discussion

M4Health is motivated by the urgent need to protect vulnerable populations, including those with limited health literacy, elderly users, young adults, and underserved communities, who increasingly turn to short-video platforms for health information and decision-making, particularly as LLMs and generative AI reshape the information landscape. M4Health offers significant benefits for combating health misinformation and AI-generated content on short-video platforms. Through multi-platform and multi-task annotations, our benchmark empowers researchers and stakeholders to develop effective content moderation tools, study platform-specific misinformation patterns, and address the growing challenge of AI-generated health content. Scaling annotations beyond the current labeled videos remains an important direction. We plan to adopt data augmentation strategies (e.g., temporal cropping, frame shuffling) alongside the existing unannotated partition to support semi-supervised fine-tuning without additional annotation cost (Carragher et al. 2025b). We envision M4Health as a foundation for trustworthy health information ecosystems that prioritize vulnerable populations and support informed health decision-making at scale.

Method	TikTok				YouTube Shorts				Reddit			
	Acc.	Prec.	Recall	F1	Acc.	Prec.	Recall	F1	Acc.	Prec.	Recall	F1
FakeSV	0.617	0.530	0.497	0.475	0.510	0.171	0.330	0.225	0.650	0.417	0.423	0.417
TikTec	0.597	0.373	0.423	0.366	0.469	0.360	0.340	0.273	0.309	0.306	0.364	0.180
LLaVA-NeXT-7B-Zero	0.429	0.257	0.232	0.232	0.407	0.231	0.272	0.235	0.487	0.318	0.282	0.289
LLaVA-NeXT-7B-Few	0.402	0.296	0.243	0.259	0.345	0.236	0.195	0.210	0.376	0.291	0.240	0.237
Qwen2.5-VL-7B-Zero	0.322	0.310	0.332	0.209	0.287	0.250	0.232	0.191	0.155	0.300	0.206	0.118
Qwen2.5-VL-7B-Few	0.305	0.293	0.337	0.220	0.364	0.246	0.235	0.200	0.221	0.271	0.239	0.163
InternVL2.5-8B-Zero	0.304	0.375	0.250	0.200	0.380	0.242	0.241	0.173	0.283	0.293	0.253	0.164
InternVL2.5-8B-Few	0.306	0.331	0.366	0.208	0.395	0.261	0.258	0.195	0.289	0.300	0.270	0.183

Table 3: Credibility assessment performance

Method	TikTok				YouTube Shorts				Reddit			
	Acc.	Prec.	Recall	F1	Acc.	Prec.	Recall	F1	Acc.	Prec.	Recall	F1
Cakelens-v5	0.824	0.324	0.326	0.316	0.602	0.374	0.363	0.350	0.611	0.406	0.407	0.392
SimpleSmallPatch	0.875	0.361	0.334	0.319	0.716	0.324	0.331	0.317	0.683	0.485	0.450	0.448
LLaVA-NeXT-7B-Zero	0.155	0.263	0.257	0.080	0.145	0.203	0.231	0.078	0.392	0.311	0.267	0.196
LLaVA-NeXT-7B-Few	0.106	0.205	0.241	0.051	0.101	0.275	0.332	0.066	0.350	0.283	0.251	0.136
Qwen2.5-VL-7B-Zero	0.810	0.341	0.361	0.344	0.701	0.236	0.248	0.236	0.575	0.372	0.303	0.333
Qwen2.5-VL-7B-Few	0.759	0.338	0.362	0.341	0.679	0.250	0.249	0.242	0.569	0.363	0.297	0.325
InternVL2.5-8B-Zero	0.649	0.305	0.360	0.295	0.644	0.240	0.253	0.241	0.625	0.399	0.412	0.387
InternVL2.5-8B-Few	0.661	0.303	0.361	0.296	0.639	0.237	0.245	0.237	0.662	0.409	0.426	0.404

Table 4: AI detection performance

Method	TikTok				YouTube Shorts				Reddit			
	Acc.	Prec.	Recall	F1	Acc.	Prec.	Recall	F1	Acc.	Prec.	Recall	F1
ViFi-CLIP	0.418	0.374	0.341	0.330	0.287	0.254	0.258	0.237	0.483	0.395	0.390	0.386
Swin3D	0.421	0.389	0.349	0.337	0.271	0.244	0.239	0.220	0.450	0.406	0.408	0.386
LLaVA-NeXT-7B-Zero	0.208	0.292	0.249	0.177	0.384	0.208	0.199	0.186	0.410	0.323	0.268	0.242
LLaVA-NeXT-7B-Few	0.320	0.263	0.243	0.229	0.303	0.204	0.201	0.191	0.366	0.274	0.256	0.248
Qwen2.5-VL-7B-Zero	0.370	0.338	0.356	0.295	0.338	0.201	0.195	0.194	0.412	0.368	0.289	0.304
Qwen2.5-VL-7B-Few	0.415	0.373	0.402	0.334	0.350	0.199	0.195	0.189	0.435	0.385	0.296	0.311
InternVL2.5-8B-Zero	0.341	0.350	0.347	0.276	0.371	0.192	0.197	0.183	0.478	0.400	0.325	0.321
InternVL2.5-8B-Few	0.356	0.354	0.360	0.289	0.379	0.206	0.203	0.192	0.480	0.411	0.328	0.331

Table 5: Theme classification performance

Conclusion

This paper presents M4Health, a multi-modal, multi-domain, multi-platform, and multi-task benchmark dataset designed to advance research on health communication in short videos. We provide expert annotations across three complementary tasks, including credibility assessment, AI-generation detection, and theme classification. Extensive

benchmarking experiments show that current state-of-the-art models are suboptimal, highlighting the unique challenges posed by multimodal health content on web and social media. We anticipate that M4Health will serve as a valuable resource for developing effective computational approaches that protect vulnerable populations and support informed health decision-making on web and social media.

Acknowledgments

We are grateful to Hannah Holden at Loyola Marymount University for helping with the data collection for this project. The work of Lin was partially supported by the National Science Foundation under grant DMS-2418877.

References

- Abbas, Z.; Abbas, M.; and Umrani, D. Z. A. 2024. Impact of Multimedia Elements on User Engagement and Content Retention in Digital Platforms. *International Journal of Contemporary Issues in Social Sciences*, 3(2): 2914–2918.
- Anderer, S. 2024. Patients are turning to TikTok for health information—here’s what clinicians need to know. *Jama*, 331(15): 1262–1264.
- Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; Zhong, H.; Zhu, Y.; Yang, M.; Li, Z.; Wan, J.; Wang, P.; Ding, W.; Fu, Z.; Xu, Y.; Ye, J.; Zhang, X.; Xie, T.; Cheng, Z.; Zhang, H.; Yang, Z.; Xu, H.; and Lin, J. 2025. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*.
- Bu, Y.; Sheng, Q.; Cao, J.; Qi, P.; Wang, D.; and Li, J. 2023. Combating online misinformation videos: Characterization, detection, and future directions. In *Proceedings of the 31st ACM International Conference on Multimedia*, 8770–8780.
- Carragher, P.; Jha, A.; Carley, K. M.; et al. 2025a. Quantifying Memorization and Parametric Response Rates in Retrieval-Augmented Vision-Language Models. In *Proceedings of the First Workshop on Large Language Model Memorization (L2M2)*, 127–141.
- Carragher, P.; Rao, N.; Jha, A.; Raghav, R.; and Carley, K. M. 2025b. Segsub: Evaluating robustness to knowledge conflicts and hallucinations in vision-language models. *arXiv preprint arXiv:2502.14908*.
- Chappa, N. V. R.; McCormick, C.; Gongora, S. R.; Dobbs, P. D.; and Luu, K. 2024. Public Health Advocacy Dataset: A Dataset of Tobacco Usage Videos from Social Media. *arXiv:2411.13572*.
- Chen, J.; Yao, J.; and Niu, L. 2024. A single simple patch is all you need for ai-generated image detection. *arXiv preprint arXiv:2402.01123*.
- Chen, Z.; Wang, W.; Cao, Y.; Liu, Y.; Gao, Z.; Cui, E.; Sun, S.; Wu, J.; Li, W.; Zhang, Y.; Jin, H.; Yang, F.; and Wang, W. 2024. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.
- Cui, L.; and Lee, D. 2020. Coaid: Covid-19 healthcare misinformation dataset. *arXiv preprint arXiv:2006.00885*.
- Dimitroyannis, R.; Fenton, D.; Cho, S.; Nordgren, R.; Pinto, J. M.; and Roxbury, C. R. 2024. A social media quality review of popular sinusitis videos on TikTok. *Otolaryngology–Head and Neck Surgery*, 170(5): 1456–1466.
- Fang, Y.; Yap, P.-T.; Lin, W.; Zhu, H.; and Liu, M. 2024. Source-free unsupervised domain adaptation: A survey. *Neural Networks*, 174: 106230.
- Google. 2026. YouTube Data API v3. <https://developers.google.com/youtube/v3>. Accessed: 2026-04-05.
- Grootendorst, M. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.
- Isgut, M.; Gloster, L.; Choi, K.; Venugopalan, J.; and Wang, M. D. 2022. Systematic review of advanced AI methods for improving healthcare data quality in post COVID-19 Era. *IEEE reviews in biomedical engineering*, 16: 53–69.
- Kirkpatrick, C. E.; and Lawrie, L. L. 2024. TikTok as a source of health information and misinformation for young women in the United States: survey study. *JMIR infodemiology*, 4(1): e54663.
- Launch Platform. 2025. Cakelens-v5: Open-source AI-generated video detection model. <https://github.com/LaunchPlatform/cakelens-v5>.
- Le, L. H.; Hoang, P. A.; and Pham, H. C. 2023. Sharing health information across online platforms: A systematic review. *Health Communication*, 38(8): 1550–1562. PMID: 34978235.
- Li, F.; Zhang, R.; Zhang, H.; Zhang, Y.; Li, B.; Li, W.; Ma, Z.; and Li, C. 2024. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*.
- Li, Y. 2025. Enhancing health misinformation detection: A multidimensional feature framework incorporating linguistic strategies. *Information Processing & Management*, 62(3): 104039.
- Li, Y.; Jiang, B.; Shu, K.; and Liu, H. 2020. Mm-covid: A multilingual and multimodal data repository for combating covid-19 disinformation. *arXiv preprint arXiv:2011.04088*.
- Liu, F.; et al. 2023. Covid-vts: Fact extraction and verification on short video platforms. *arXiv preprint arXiv:2302.07919*.
- Massey, P. M.; Leader, A.; Yom-Tov, E.; Budenz, A.; Fisher, K.; and Klassen, A. C. 2016. Applying multiple data collection tools to quantify human papillomavirus vaccine communication on Twitter. *Journal of Medical Internet Research*, 18(12): e318.
- McHugh, M. L. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3): 276–282.
- Migisha, M.-G.; and Hagström, T. 2025. Seeing is Believing?: TikTok Users Perception, Trust, and Engagement with AI-Generated Visual vs. Human-Generated Visual Content.
- PRAW. 2026. PRAW: The Python Reddit API Wrapper. <https://praw.readthedocs.io/en/stable/>. Accessed: 2026-04-05.
- pytube. 2026. pytube: Lightweight, dependency-free Python library and CLI for downloading YouTube videos. <https://github.com/pytube/pytube>. Accessed: 2026-04-05.
- Qi, P.; Bu, Y.; Cao, J.; Ji, W.; Shui, R.; Xiao, J.; Wang, D.; and Chua, T.-S. 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*, volume 37, 14444–14452.

- Rasheed, H.; Khattak, M. U.; Maaz, M.; Khan, S.; and Khan, F. S. 2023. Fine-tuned clip models are efficient video learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6545–6554.
- Reddit. 2025. Reddit. <https://www.reddit.com/>.
- Schlicht, I. B.; Fernandez, E.; Chulvi, B.; and Rosso, P. 2024. Automatic detection of health misinformation: a systematic review. *Journal of Ambient Intelligence and Humanized Computing*, 15(3): 2009–2021.
- Shang, L.; Kou, Z.; Zhang, Y.; and Wang, D. 2021. A multimodal misinformation detector for covid-19 short videos on tiktok. In *2021 IEEE international conference on big data (big data)*, 899–908. IEEE.
- Shang, L.; Zhang, Y.; Deng, Y.; and Wang, D. 2025. Multi-Tec: a data-driven multimodal short video detection framework for healthcare misinformation on TikTok. *IEEE Transactions on Big Data*, 11(5): 2471–2488.
- Steen, E.; Yurechko, K.; and Klug, D. 2023. You can (not) say what you want: Using algospeak to contest and evade algorithmic content moderation on TikTok. *Social Media+ Society*, 9(3): 20563051231194586.
- Stepaniuk, K. 2024. Social Media Content Multimodality and the Level of User Interactions from the Perspective of Facebook and Instagram. *Inzinerine Ekonomika—Engineering Economics*, 35(5): 602–613.
- Sun, Y.; He, J.; Lei, S.; Cui, L.; and Lu, C.-T. 2023. Med-mmhl: A multi-modal dataset for detecting human-and llm-generated misinformation in the medical domain. *arXiv preprint arXiv:2306.08871*.
- Thakur, N.; Su, V.; Shao, M.; Patel, K. A.; Jeong, H.; Knieling, V.; and Bian, A. 2024. A Labeled Dataset for Sentiment Analysis of Videos on YouTube, TikTok, and Other Sources About the 2024 Outbreak of Measles. In *International Conference on Human-Computer Interaction*, 220–239. Springer.
- TikTok. 2025. TikTok. <https://www.tiktok.com/en/>.
- TikTok. 2026. TikTok Research API. <https://developers.tiktok.com/products/research-api/>. Accessed: 2026-04-05.
- Violot, C.; Elmas, T.; Bilogrevic, I.; and Humbert, M. 2024. Shorts vs. Regular Videos on YouTube: A Comparative Analysis of User Engagement and Content Creation Trends. *ACM Web Science Conference*.
- Xuan, K.; Yi, L.; Yang, F.; Wu, R.; Fung, Y. R.; and Ji, H. 2024. LEMMA: towards lvlm-enhanced multimodal misinformation detection with external knowledge augmentation. *arXiv preprint arXiv:2402.11943*.
- Yang, Y.-Q.; Guo, Y.-X.; Xiong, J.-Y.; Liu, Y.; Pan, H.; Wang, P.-S.; Tong, X.; and Guo, B. 2025. Swin3d: A pre-trained transformer backbone for 3d indoor scene understanding. *Computational Visual Media*, 11(1): 83–101.
- YouTube. 2025. YouTube Shorts. <https://www.youtube.com/shorts/>.
- YT-DLP. 2026. YT-DLP: A feature-rich command-line audio/video downloader. <https://github.com/yt-dlp/yt-dlp>. Accessed: 2026-04-05.
- Zhang, J.; Yuan, J.; Zhang, D.; Yang, Y.; Wang, C.; Dou, Z.; and Li, Y. 2024. Short video platforms as sources of health information about cervical cancer: A content and quality analysis. *PLoS One*, 19(3): e0300180.
- Zhao, Y. C.; Zhao, M.; and Song, S. 2022. Online health information seeking behaviors among older adults: systematic scoping review. *Journal of medical internet research*, 24(2): e34790.
- Zhou, X.; Mulay, A.; Ferrara, E.; and Zafarani, R. 2020. Recovery: A multimodal repository for covid-19 news credibility research. In *Proceedings of the 29th ACM international conference on information & knowledge management*, 3205–3212.

Appendix

Search Query Keywords

TikTok and YouTube Shorts

- **Nutrition & Diet:** AlkalineDiet, BalancedDiet, CabbageSoupDiet, CalorieDeficit, CalorieCounting, CarnivoreDiet, DetoxDrinks, Diet, EatEnough, Fasting, IntermittentFasting, JuiceCleanse, Keto, KetoDiet, LowCarb, Macros, MealPrep, NutritionTips, PlantBased, Superfoods, VeganLifestyle, ViralRecipe, WaterFasting, BalancedNutrition, HealthyEating, HealthyRecipes
- **Weight Loss:** FatBurning, FatLoss, FatLossSpecialist, LoseBellyFat, LoseFat, SpotReduction, SummerBody, WeightLoss, WeightLossCoaching
- **Fitness:** Cardio, ExerciseMotivation, ExerciseRoutine, FitnessForMen, FitnessMotivation, FromSkinnyToStrong, GymLife, GymMotivation, HomeWorkouts, MensHealth, MenFitness, PersonalTrainer, SixPackAbs, StrengthTraining, WorkoutTips
- **Sleep:** BedtimeRoutine, BedtimeHabits, CircadianRhythm, DeepRest, DeepSleep, Melatonin, PowerNap, SleepBetter, SleepHealth, SleepHygiene, SleepTips
- **Skin, Hair & Beauty:** ClearSkinTips, GlowingSkin, HairLossSolutions, NaturalBeauty, SkincareRoutine, SunProtection
- **Women’s Health:** FertilityTips, FertilityAwareness, HormoneBalance, HormoneHealth, MenstrualHealth, Perimenopause, PostpartumCare, PostpartumHealth, PregnancyJourney
- **Men’s Health:** MensHealth, MenFitness, ProstateHealth, TestosteroneBoost, TestosteroneBoosting
- **Gut Health:** FermentedFoods, GutHealth, IBSRelief, Microbiome, Prebiotics, Probiotics
- **Chronic Illness & Disease Prevention:** AutoimmuneDisease, AutoimmuneProtocol, CancerAwareness, CancerPrevention, ChronicPainRelief, DiabetesManagement, HeartHealth, ImmuneBoosting, MetabolicHealth, MetabolismBoost
- **Natural Health:** AlternativeMedicine, Ayurveda, CleanEating, CureDiseaseNaturally, Detox, Electrolytes, EnergyHealing, EssentialOils, HerbalRemedies, HolisticHealth, HolisticWellness, Homeopathy, NaturalHealing, Wellness, WellnessDrinks, WellnessJourney, WellnessTips
- **Environmental & Sustainability Topics:** BigFood, BigPharma, BigPharmaExposed, GMOAwareness, HiddenIngredients, FoodAwareness, HeavyMetalDetox

Reddit

- **Nutrition & Diet:** diet, CleanEating, MealPrep, fasting, CarnivoreDiet, StopEatingSugar, keto, vegan, PlantBasedDiet, LowCarb, CalorieCounting, EatCheapAndHealthy
- **Weight Loss:** weightloss, loseit, FatLoss, gainit, keto-gains
- **Fitness:** weightlifting, cardio, HomeWorkouts, StrengthTraining

- **Gut Health:** probiotics, prebiotics, GutHealth
- **Supplements:** biohacking, nootropics, Supplements, Biohackers
- **Alternative Health:** AlternativeHealth, Wellness, Superfoods, energyhealing, naturopathy, Herbalism, Homeopathy, essentialoils, AlternativeMedicine
- **Health Information:** Health, medical, medicine, HealthAnxiety, hydration, Sleptips
- **Health Controversies:** ivermectin, antiVax, CovidVaccinated, HermanCainAward, ScienceUncensored, BigPharma

Implementation Details

Preprocessing We follow the preprocessing steps specified in the original paper of each task-specific baseline model. The detailed preprocessing steps are summarized in Table 7.

Hyperparameter We summarize the hyperparameter setting of the compared baselines in Table 6. In particular, we use Adam as the model optimizer and keep the patience as 3 epochs. We employ early stopping and bootstrap sampling to ensure robust evaluation.

Method	batch	epochs	lr	pretrained
FakeSV	16	20	5e-5	False
TikTec	16	40	1e-5	False
CakeLens	128	40	1e-4	True
SimpleSmallPatch	32	20	1e-4	True
ViFi-CLIP	8	20	1e-5	True
Swin3D	8	20	1e-5	True

Table 6: Training hyperparameters for task-specific models.

Prompts We adopt the following prompts (Table 8) for the LVLMs under zero-shot and few-shot settings.

Method	Preprocessing Details
FakeSV	video_features include vgg16_frame, c3d_clip and vggish_audio from VGG16, C3D and Vggish model output; text_features include title_text and comments_text from Bert tokenizer; OCR results extracted from video frames and used as extra textual evidence.
TikTec	caption_features are built from detected objects and embedded with GloVe; visual_features are obtained by running a Faster R-CNN detector and encoding object patches with a shared segment mask mask_K; ASR plus audio features are constructed by combining OCR text from frames (EasyOCR based) with MFCC audio embeddings into asr_feature of length N and a corresponding mask_N.
Cakelens-v5	Frame_level_embeddings are extracted by hooking the Flatten layer inside fully_connected while running the CakeLens Detector; We fine-tuning based on the embeddings extracted from Cakelens
SimpleSmallPatch/ViFi-CLIP/Swin3D	video_clip extracted from target videos with non broken frames by ffmpeg or opencv.

Table 7: Supervised model configurations: preprocessing pipelines.

Shot Setting	Credibility Assessment	AI Generation Detection	Theme Classification
zero-shot	<p>Task. Given a short health-related video, the model is told: “Analyze the video carefully (visuals and audio) and decide whether the health-related information in the video is accurate, clearly false, or a mix of both.”</p> <p>Label set. Three labels are defined: – true: factually accurate, verifiable, no significant inaccuracies. – mixed: mix of accurate and inaccurate/misleading information, possibly with speculation or exaggeration. – false: demonstrably false, misleading, or fabricated, not supported by credible evidence.</p>	<p>Task. Given a short video clip, the model is told to judge the generation source: “Based only on how the video looks and sounds (editing style, overlays, filters, cuts, visual artifacts, motion consistency, realism of textures and speech), decide whether it is unedited/raw footage, edited/enhanced footage, or AI-generated content.”</p> <p>Label set. Three labels are defined: – unedited: raw recording of a real-world event/person. – edited: authentic footage with noticeable post-production (cuts, filters, text overlays, transitions, color grading). – ai: content predominantly created using artificial intelligence, with little or no real footage.</p>	<p>Task. Given a short video clip, the model is told: “You will see a short video clip. Analyze the video carefully (visuals and audio). Based on your analysis, choose the ONE most appropriate label for the content type.”</p> <p>Label set. Four labels are defined: – factual / informational: objective or instructive content. – opinion / commentary: mainly personal views or interpretations. – satire / entertainment: humor, satire, irony, exaggeration, or parody. – promotion / persuasion: aims to influence attitudes or behaviors.</p>
few-shot	<p>Same base instruction and label definitions as zero-shot, plus three textual examples: (1) Handwashing with soap to reduce virus spread, consistent with public health guidelines → true. (2) Eating a specific fruit once permanently cures diabetes → false. (3) Hypertension diagnosis plus an unproven “guaranteed” home remedy → mixed.</p>	<p>Same base instruction and label definitions as zero-shot, plus three textual examples: (1) Shaky phone footage with natural lighting and no visible edits → unedited. (2) Vlog with jump cuts, music, color grading, and text overlays → edited. (3) Hyper-realistic talking animal with slightly unnatural motion and synthetic voice → ai.</p>	<p>Same base instruction and label definitions as zero-shot, plus four textual examples: (1) Doctor explains vaccines with diagrams → factual / informational. (2) Creator talking about preferred diet → opinion / commentary. (3) Humorous skit exaggerating health trend → satire / entertainment. (4) Polished supplement promotion clip → promotion / persuasion.</p>

Table 8: Prompt configurations for all LLM. For all settings (zero-shot vs. few-shot) and targets (type, veracity, source), the model receives 8 video frames per sample and must output a structured response with a <thinking> reasoning block followed by a single <label>.