ACM multimedia

Dublin, Ireland 27-31.10.2025

Combating Online Misinformation Videos: Characterization, Detection, and Prevention







Qiang Sheng[1] Peng Qi[2] Tianyun Yang[3] Yuyan Bu[4] Wynne Hsu[2] Mong Li Lee[2] Juan Cao[1]

[1]Institute of Computing Technology, Chinese Academy of Sciences [2]National University of Singapore
[3]Shenzhen Research Institute of Big Data [4]Beijing Academy of Artificial Intelligence

October 27, 2025 13:30-17:00 | Goldsmiths 1, Dublin Royal Convention Centre, Ireland

Agenda: 3-hour Talk plus 30-minute Break

Time	Section	Presenter
13:30-13:40	Introduction & Motivation	Qiang Sheng
13:40-13:55	Preliminaries: Video Editing & Generation	Qiang Sheng
13:55-14:10	Characterization of Misinformation Videos	Qiang Sheng
14:10-14:50	Detection Part I: Human-Edited Misinformation	Yuyan Bu
14:50-15:30	Detection Part II: AI-Generated Misinformation	Tianyun Yang
15:30-16:00	Coffee Break	/
16:00-16:40	Prevention Strategies	Peng Qi
16:40-17:00	Conclusion & Open Discussion / General QA	Qiang Sheng / All

Clarification questions are welcomed during the talk

Tutorial Outline

ACM multimedia
Dublin, Ireland 27-31.10.2025

1. Introduction & Motivation

Background

Effects and Concerns

Goals of Our Tutorial

2. Preliminaries: Video Editing & Generation

Overview of Generative Models and Diffusion

Video Generation

Video Editing

Q+A/Discussion

Qiang Sheng

Tutorial Outline

3. Characterization of Misinformation Videos

- Definition
- Datasets Overview
- Analysis on FakeSV
  - News Content
  - Social Context
  - Propagation
- Q+A/Discussion

Qiang Sheng

Tutorial Outline

4. Detection Part I: Human-Edited Misinformation

- Signal-based detection
  - Editing Traces
  - Generation Traces
- Semantic-based detection
  - Seek clues within the sample’s own content
  - Seek clues from external information
- Intent-based detection
  - Social context
- Clue integration for misinformation video detection
  - Parallel Integration
  - Sequential Integration
- Q+A/Discussion

Yuyan Bu

Tutorial Outline

ACM multimedia
Dublin, Ireland 27-31.10.2025

5. Detection Part II: AI-Generated Misinformation

Manipulated video detection

Generated video detection

Attributing AI-generated Content to the Source Model

Q+A/Discussion

Tianyun Yang

Tutorial Outline

6. Prevention Strategies

- Creation Prevention
  - Embedding Tamper-proof Digital Identifier
  - Mitigating Hallucination in Content Generation
- Spread Prevention
  - Alert, Verification, and Resilience Building
  - Controlling the Spread of Misinformation
  - Promoting Truth and Debunking

Peng Qi

Tutorial Outline

7. Conclusion & Open Discussion

plus General Q+A

Qiang Sheng

Slides & Reading List

https://misinfo-video.github.io/

Introduction & Motivation

Section 1

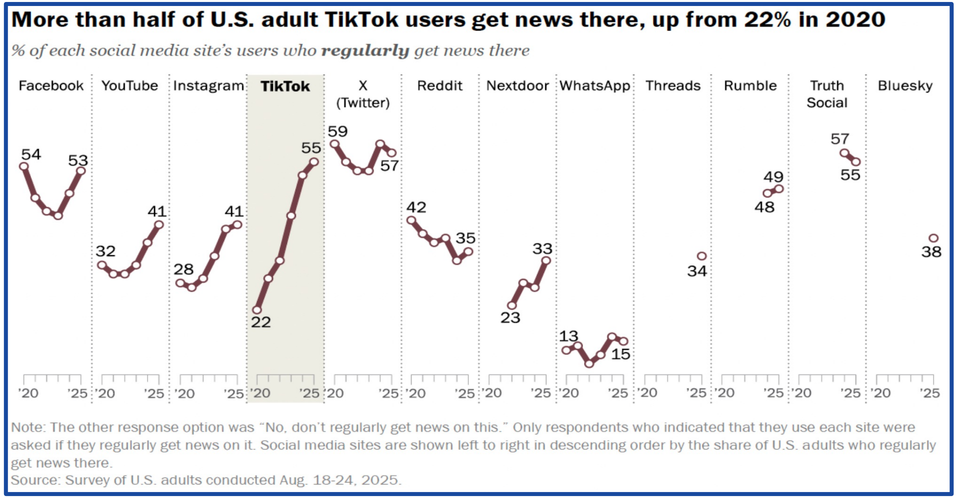
Tutorial Outline

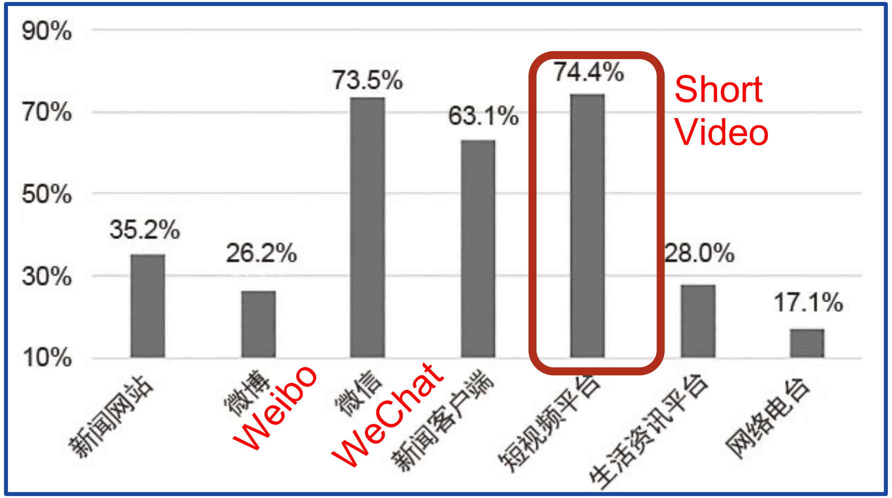
1. Introduction & Motivation

- Background
- Effects and Concerns
- Goals of Our Tutorial

Video has been a popular form of news consumption

Short video-sharing platforms like TikTok and others has been important access to get daily news for users.





Longitudinal Comparison

US adults that get news from TikTok 2020-2025: 22% -> 55%

Horizontal Comparison

Short video platforms surpass WeChat and Weibo, becoming the main access for Chinese netizens.

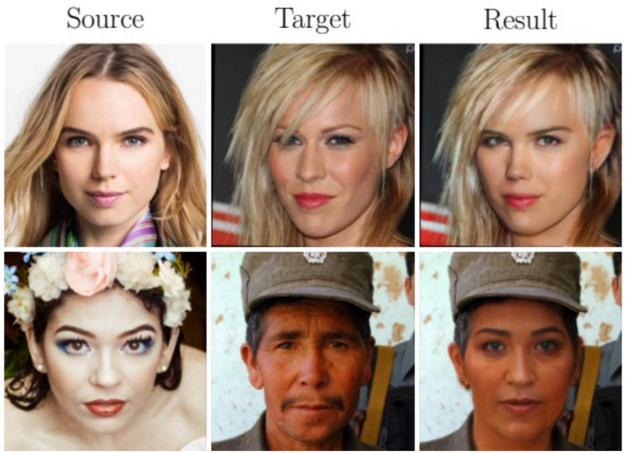
Pew Research Center. 1 in 5 Americans now regularly get news on TikTok, up sharply from 2020.
CNNIC. Survey on the Current Status of News Access Channels for Chinese Netizens. (in Chinese)

Meanwhile, video generation techniques has rapid progress

Trend #1: More general.
Video generation is not only face-swapping.

Before
Face swapping, editing, Reenactment

DeepFaceLab faceswap



Recent
General video generation guided by text/image prompts

OpenAI SORA KLING

https://arxiv.org/abs/2312.10843
https://openai.com/index/sora-2/

Meanwhile, video generation techniques has rapid progress

Trend #2: More vivid.

Video generation can be of high resolution with details.

~60s, multiple shots
physical details (though imperfect)

Large view
Dynamic shot

Source: Sora examples

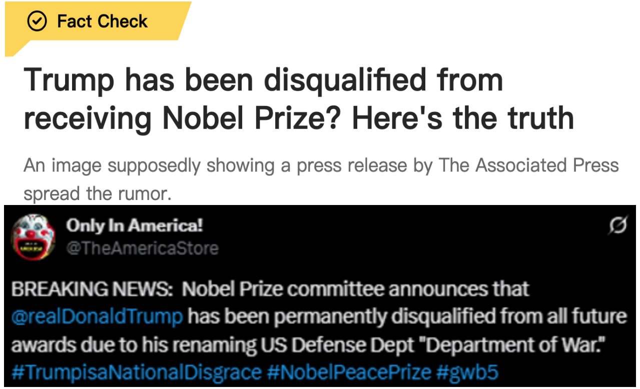
Tutorial Outline

1. Introduction & Motivation

- Background
- Effects and Concerns
- Goals of Our Tutorial

Along with such a progress, video misinfo is easier to make

Misinformation was delivered as text-only, or image-included posts.
Now, producing misinformation videos are easier than before.



Fake Text
Trump is disqualified from Nobels?

Fake Image
Trump was arrested?
(Using Generative AI)

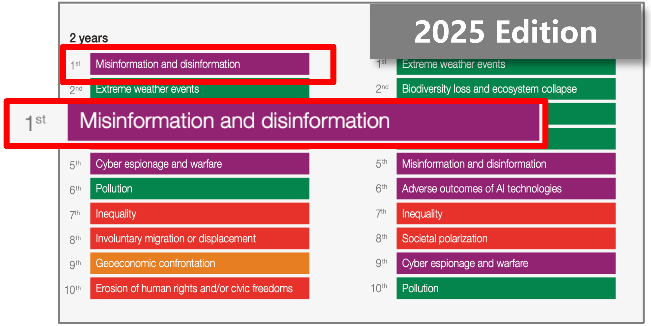
Fake Video
Trump fought with Zelensky?
(Using Generative AI)

https://www.snopes.com/fact-check/trump-nobel-prize-disqualified/
https://apnews.com/article/fact-check-trump-NYPD-stormy-daniels-539393517762
https://www.youtube.com/watch?v=2wqWnoMg1dU

The Worrying Trend: AI Video Faking is Industrialized

ACM multimedia
Dublin, Ireland 27-31.10.2025

Surveys and predictions show deep concerns regarding AI misinfo



World Economic Forum, Global Risks Report:

Misinformation and disinformation is the TOP1 two-year risk and TOP1 ten-year technical risk

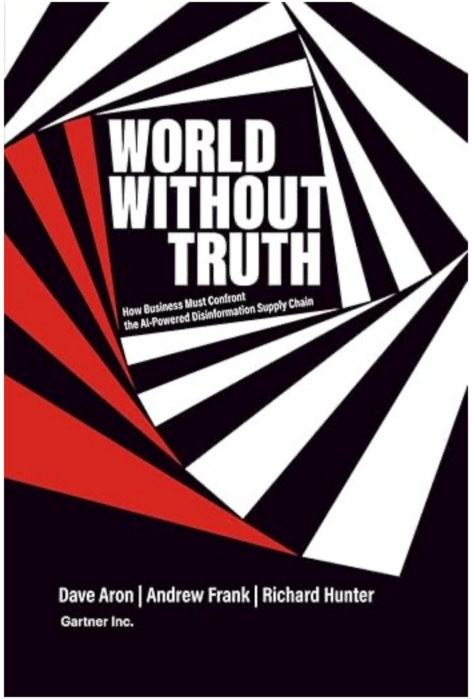
State of AI 2025 Report make a prediction:

A deepfake/agent-driven cyber attack triggers the first NATO/UN emergency debate on AI security.

https://www.weforum.org/publications/global-risks-report-2025/
https://www.stateof.ai/

The Worrying Trend: AI Video Faking is Industrialized

Surveys and predictions show deep concerns regarding AI misinfo



- Disinformation is now a sophisticated, organized business with its own supply chain.
- Generative AI accelerates the creation and spread of convincing fake content, manipulating narratives and exploiting biases at scale.

https://www.gartner.com/en/publications/world-without-truth

The Worrying Trend: AI Video Faking is Industrialized

With the support of recent AI techniques, misinformation video faking can be with a larger scale with lower cost

A Real-World Case Reported by China Central TV

A team at Zhejiang published 28.5k misinformation videos by cutting & editing video clips, attracting 2.7 billion views

https://news.cctv.com/2024/10/26/ARTI4Z5aSYied0EFnykpVTmX241026.shtml (in Chinese)

For Human: We might be more vulnerable to video misinfo

Journal of Computer-Mediated Communication

Seeing Is Believing: Is Video Modality More Powerful in Spreading Fake News via Online Messaging Apps?

S. Shyam Sundar [1], Maria D. Molina [2], & Eugene Cho[3]

[1]Media Effects Research Laboratory, Penn State University, University Park, PA 16802, USA
[2]Department of Advertising and Public Relations, Michigan State University, East Lansing, MI 48824, USA
[3]Department of Communication Studies, The College of New Jersey, Ewing, NJ 08628, USA

… It is clear from our findings **that video is causing individuals to perceive fake news as more credible than audio and text**, and increases the likelihood of them spreading it. …

As Good as a Coin Toss

Human Detection of AI-Generated Images, Video, Audio, and Audiovisual Stimuli

DI COOKE, Department of War Studies, King's College London, London, UK
ABIGAIL EDWARDS, Center for Strategic and International Studies, Washington D.C. USA
SOPHIA BARKOFF, Center for Strategic and International Studies, Washington D.C. USA
KATHYRN KELLY, Center for Strategic and International Studies, Washington D.C. USA

We find that **on average, people struggled to distinguish between synthetic and authentic media, with the mean detection performance close to a chance level performance of 50%**. We also find that accuracy rates worsen when the stimuli contain any degree of synthetic content, features foreign languages, and the media type is a single modality.

Fake news with videos is more convincing for human and easier to spread

The synthetic media is hard to distinguish for ordinary people

Seeing is believing: Is Video Modality More Powerful in Spreading Fake News via Online Messaging Apps?
As Good As A Coin Toss: Human detection of AI-generated images, videos, audio, and audiovisual stimuli

For Detectors: New challenges posed by misinfo videos

ACM multimedia
Dublin, Ireland 27-31.10.2025

1

High information heterogeneity brought by various modalities

- Requires a stronger and more comprehensive understandability
- Brings more uncertainty and even noise to the final prediction

2

Blurred distinction between misleading video manipulation and non-malicious artistic video editing

- Artistic editing like beautifying faces is general, making it hard to see the malicious manipulation traces.

3

New patterns of misinformation propagation due to the dominant role of recommendation systems on online video platforms

- Less social contexts then before on Twitter/Weibo. Brings new behaviors.
- Requires new detection and prevention methods.

Tutorial Outline

1. Introduction & Motivation

- Background
- Effects and Concerns
- Goals of Our Tutorial

Goals of this Tutorial

Key Question

How to **Characterize, Detect, and Prevent** Misinformation Videos?

Survey Range

Covers typical or new works in this direction in recent five years

Best for

Audience that have knowledge about AI safety/multimedia content safety or is interested in combating misinformation issues

Preliminaries: Video Editing & Generation

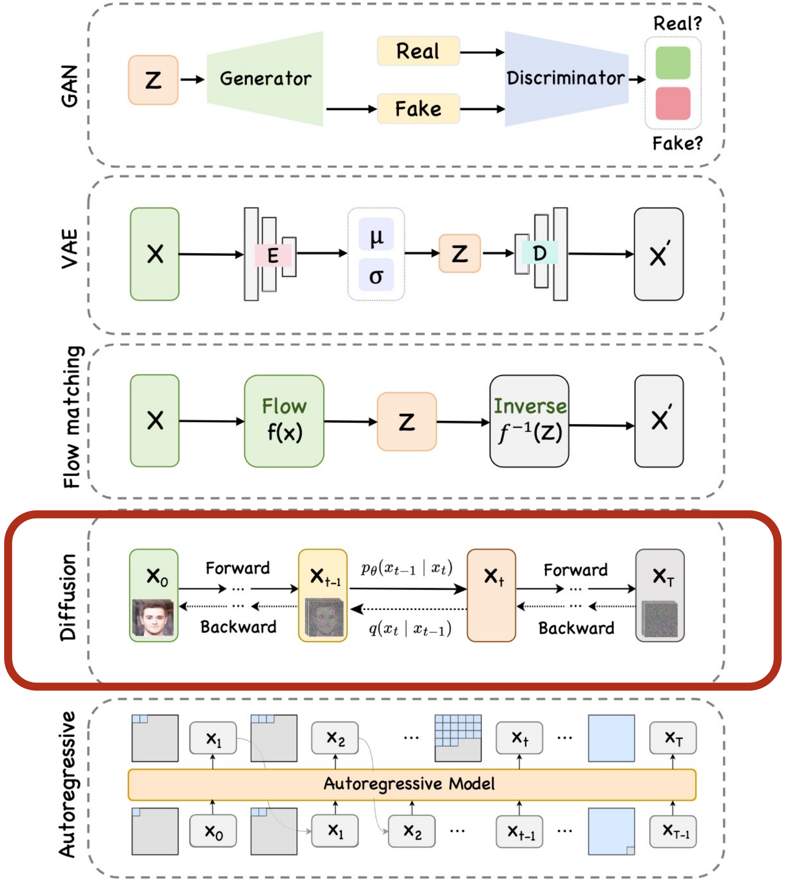
Section 2

Tutorial Outline

2. Preliminaries: Video Editing & Generation

- Overview of Generative Models and Diffusion
- Video Generation
- Video Editing
- Q+A/Discussion

Generative Models

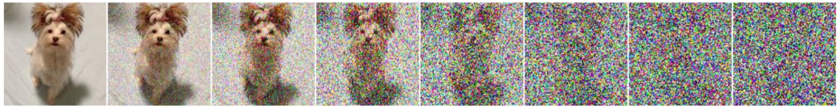


Our focus in this part

Image Diffusion Model

Forward SDE (data → noise)

$$\mathbf{x}(0) \longrightarrow \mathrm{d}\mathbf{x} = \mathbf{f}(\mathbf{x}, t)\mathrm{d}t + g(t)\mathrm{d}\mathbf{w} \longrightarrow \mathbf{x}(T)$$



score function

$$\mathbf{x}(0) \longleftarrow \mathrm{d}\mathbf{x} = \left[\mathbf{f}(\mathbf{x}, t) - g^2(t)\boxed{\nabla_{\mathbf{x}} \log p_t(\mathbf{x})}\right] \mathrm{d}t + g(t)\mathrm{d}\bar{\mathbf{w}} \longleftarrow \mathbf{x}(T)$$

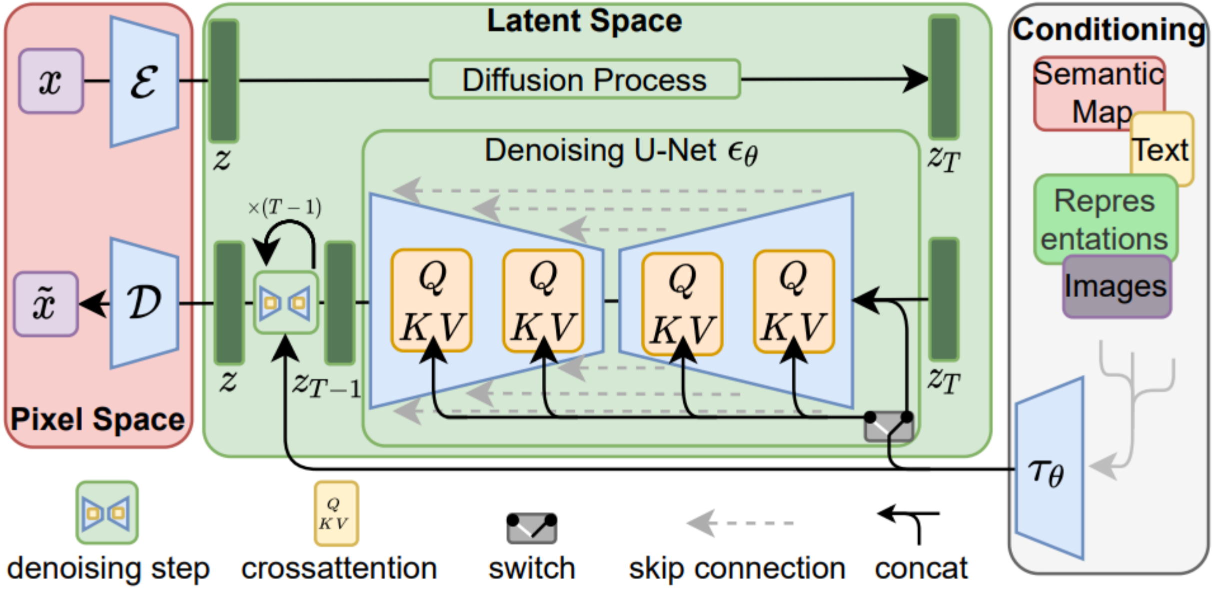
Reverse SDE (noise → data)

Basic Idea: Learning to add/remove noise
Generate an image as building a house.
To learn it, separate the sample house, know each step, and then try to rebuild it

Too expensive to use in reality

If we want a 1024*1024 image, we need to input a 1024*1024 noise image, requiring high computation overhead.

Latent Diffusion Model



Common usage:
- **Use VQVAE to transform into latent space**
- **A U-Net based diffusion model**
- **Controlled by cross-attention condition**

Tutorial Outline

2. Preliminaries: Video Editing & Generation

Overview of Generative Models and Diffusion

Video Generation

Video Editing

Q+A/Discussion

From Image Generation to Video Generation

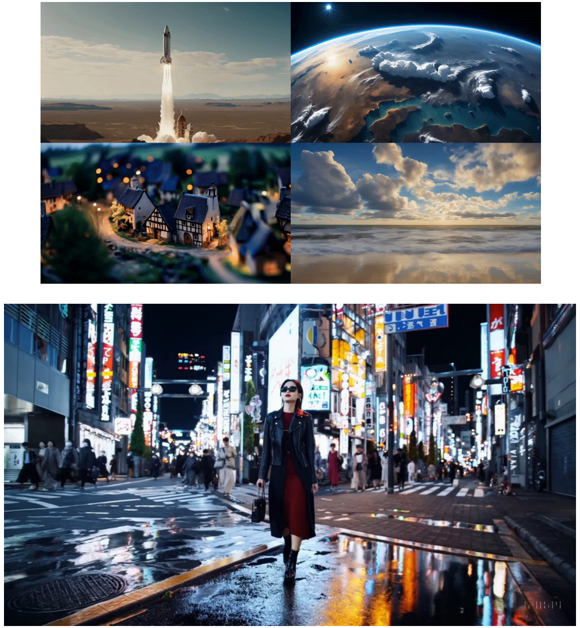


Image Generation (2022-)
Stable Diffusion, Dalle3…

Video Generation (2023-)
Stable Video Diffusion, Sora…

How to extend the usage of diffusion models to video?

Video Diffusion Models (VDM)

Build VDM Arch.

High-Impact VDM

Sora…

Make-A-video
(Meta,2022)

Align-your-latent
(NVIDIA, 2023)

SVD
(2023)

Sora…

Focuses: High-quality, Controllable, and Arbitrary Length

Make-A-Video (Meta, 2022)

- Extend Text-to-Image Model to 3D Model
  - 2D Conv -> Pseudo 3D Conv
  - Spatial attention -> Pseudo 3D attention
- Frame Interpolation
  - Using mask prediction
- Spatial Super-Resolution

- Issue：
  - Low quality (a few seconds, 256*256, blurs)

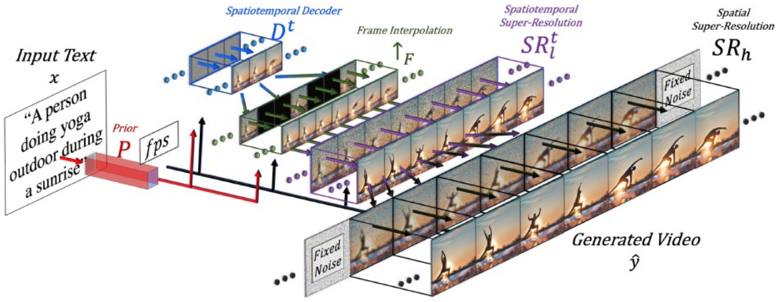


Figure 2: **Make-A-Video high-level architecture.** Given input text x translated by the prior P into an image embedding, and a desired frame rate fps, the decoder D^t generates 16 64 × 64 frames, which are then interpolated to a higher frame rate by $\uparrow_F$, and increased in resolution to 256 × 256 by SR_l^t and 768 × 768 by SR_h, resulting in a high-spatiotemporal-resolution generated video $\hat{y}$.

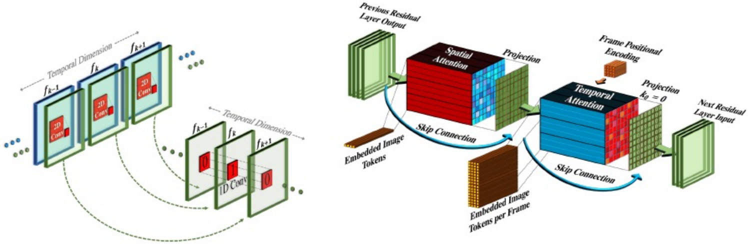


Figure 3: **The architecture and initialization scheme of the Pseudo-3D convolutional and attention layers, enabling the seamless transition of a pre-trained Text-to-Image model to the temporal dimension. (left)** Each spatial 2D conv layer is followed by a temporal 1D conv layer. The temporal conv layer is initialized with an identity function. **(right)** Temporal attention layers are applied following the spatial attention layers by initializing the temporal projection to zero, resulting in an identity function of the temporal attention blocks.

Make-A-Video: Text-to-Video Generation without Text-Video Data

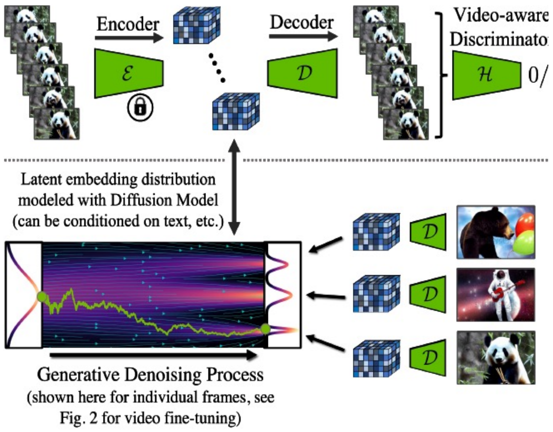
VideoLDM (NVIDIA, 2023)

- Practice **latent** VDM paradigm
- Can generated 2k resolution videos
- Long video generation: Keyframe + Mask Prediction

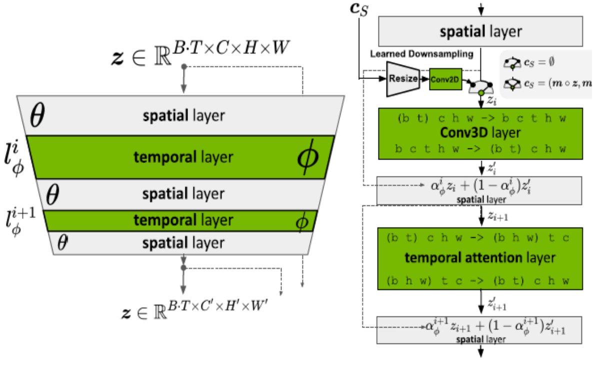
Align your Latents:
High-Resolution Video Synthesis with Latent Diffusion Models

Andreas Blattmann[1 *,†] Robin Rombach[1 *,†] Huan Ling[2,3,4 *] Tim Dockhorn[2,3,5 *,†]
Seung Wook Kim[2,3,4] Sanja Fidler[2,3,4] Karsten Kreis[2]

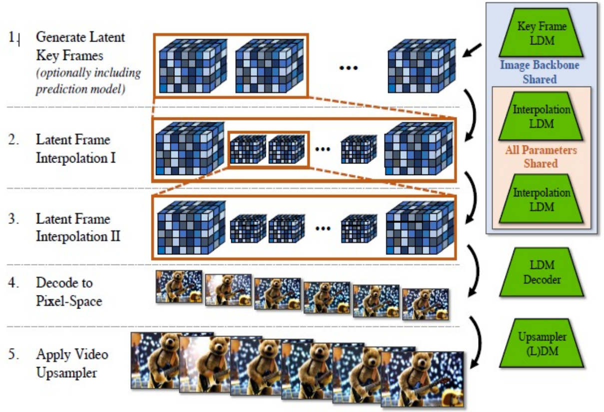
[1]LMU Munich [2]NVIDIA [3]Vector Institute [4]University of Toronto [5]University of Waterloo



Video-level
VQ-VAE training



Text-to-Image ->
Text-to-Video



Generate Keyframes ->
Latent Frame Interpolation->
Video Upsampling

Align your Latents: High-Resolution Video Synthesis with Latent Diffusion Models

Stable Video Diffusion (Stability AI, 2023)

- Based on VideoLDM architecture
- Better data cleaning for high-quality video data
  - With a useful video data processing pipeline:
  - Cut detection for video-cutting
  - Use CLIP/BLIP to obtain captions
  - Filter out static data based on optical flow
- High-quality video data -> High generation quality

Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets

Andreas Blattmann* Tim Dockhorn* Sumith Kulal* Daniel Mendelevitch
Maciej Kilian Dominik Lorenz Yam Levi Zion English Vikram Voleti
Adam Letts Varun Jampani Robin Rombach
Stability AI

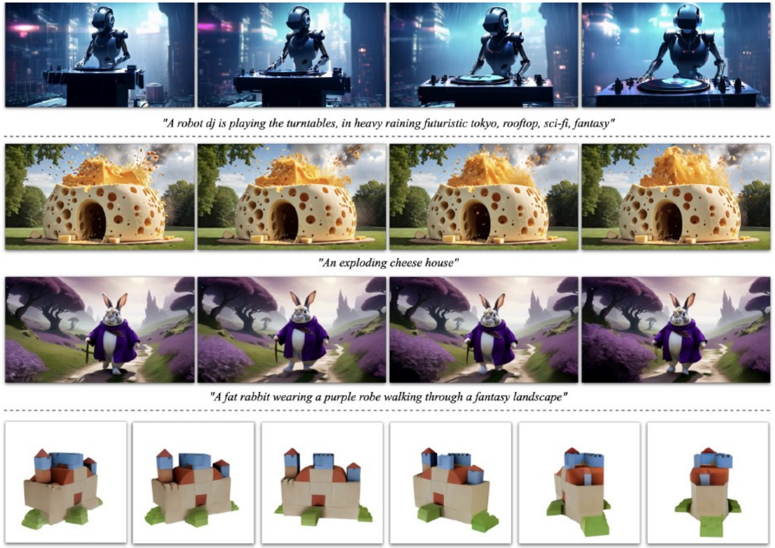


Figure 1. **Stable Video Diffusion samples.** *Top:* Text-to-Video generation. *Middle:* (Text-to-)Image-to-Video generation. *Bottom:* Multi-view synthesis via Image-to-Video finetuning.

Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets

Sora (OpenAI, 2024)

- Breakthrough Achievements
  - 4K resolution
  - High spatial consistency
  - Diverse scenes with multiple shots
  - Longer video (60s)
- Quality issues -> Commonsense/Physical violation issues

Research

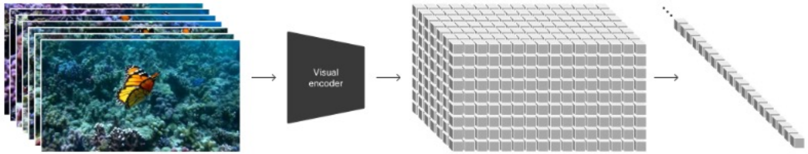
Video generation models as world simulators

We explore large-scale training of generative models on video data. Specifically, we train text-conditional diffusion models jointly on videos and images of variable durations, resolutions and aspect ratios. We leverage a transformer architecture that operates on spacetime patches of video and image latent codes. Our largest model, Sora, is capable of generating a minute of high fidelity video. Our results suggest that scaling video generation models is a promising path towards building general purpose simulators of the physical world.

Video generation models as world simulators

Sora (OpenAI, 2024)

- **Cora Idea:** Turning visual data of all types into a unified representation that enables large-scale training of generative models
- **Turning visual data into patches**

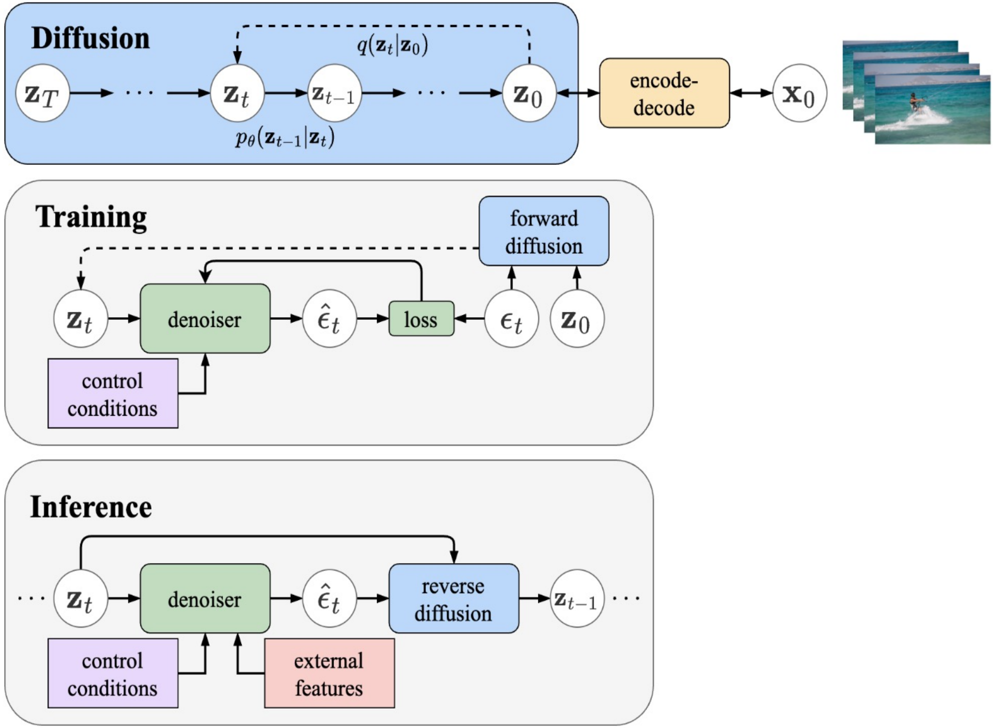


- **Video compression network → Video VQ-VAE**
- **Spatiotemporal latent patches** → tokenization, arbitrary length and resolution
- **Scaling transformers for video generation**
  - Based diffusion transformer architecture (DiT)
  - Support

Video generation models as world simulators

Diffusion-Based Video Editing

- Similar to Video Generation
  - Consider controllable signals more
- Text-base conditioning
- Point Conditioning
  - DragDiffusion, DragVideo…
- Pose Conditioning
  - Follow-Your-Pose…



Diffusion Model-Based Video Editing: A Survey

Characterization of Misinformation Videos

Section 3

Tutorial Outline

3. Characterization of Misinformation Videos

- Definition
- Datasets Overview
- Analysis on FakeSV
  - Signal Level
  - Semantics Level
  - Intent Level
- Q+A/Discussion

Definition & Taxonomy

- **Misinformation Video**
  - ➢ A video post that conveys false, inaccurate, or misleading information.

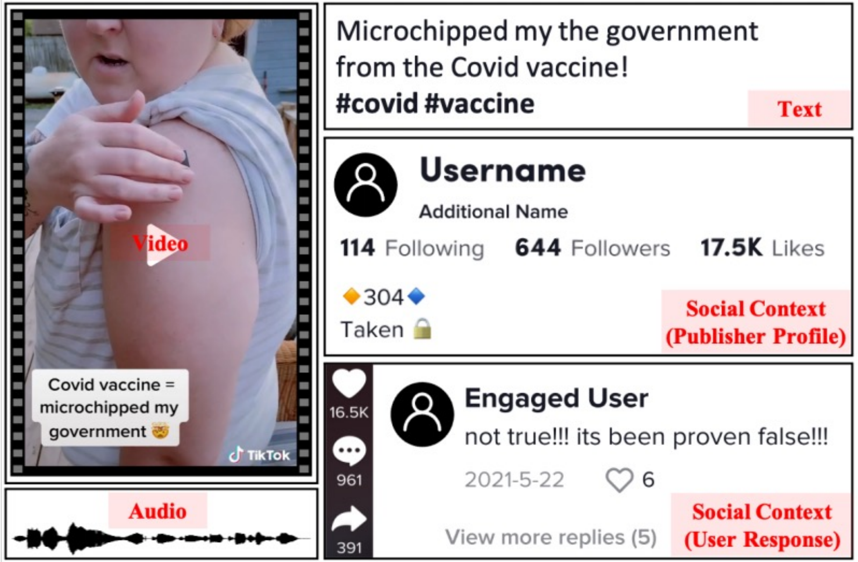
- **Misinformation Video Detection**
  - ➢ Predict whether the video post $\mathcal{V}$ contains misinformation given all the $\mathcal{E}$ ccessible features

$$\mathcal{F} : \mathcal{E} \mapsto \{0, 1\}$$

- ➢ For some recent works, the output also contains a natural language text beyond the binary classification labels to provide human-understandable explanations.

Combating Online Misinformation Videos: Characterization, Detection, and Future Directions
Multimodal Fake News Video Explanation: Dataset, Analysis and Evaluation

Definition



Category	Fileds
Content	video, cover image, title, published time
Response	# of likes/stars/comments, top 100 comments (with reviewed time, # of likes and # of sub-comments)
Publisher	info_verified, info_introduction, current IP location, # of fans/subscribes/likes/videos and top 100 published videos' covers

Tutorial Outline

3. Characterization of Misinformation Videos

- Definition
- Datasets Overview
- Analysis on FakeSV
  - News Content
  - Social Context
  - Propagation
- Q+A/Discussion

Datasets Overview

Real-world data:

Table 1: Summary of datasets of fake news video detection. Metadata refers to basic statistics such as # of likes/stars/comments. "-" represents open-domain. Names of sources are abbreviated for simplicity (YT: YouTube, TW: Twitter, FB: Facebook, TT: TikTok, BB: Bilibili, DY: Douyin, KS: Kuaishou).

Dataset	Features					Instances	Domain	Language	Released	Source
	Video	Title	Metadata	Comment	User	(fake/real)				
(Papadopoulou et al. 2018)	✓	✓	✓	✓		2,916/2,090	-	En,Fr,Ru,Ge,Ar	Y	YT,TW,FB
(Palod et al. 2019)	✓	✓	✓	✓		123/423	-	En	Y	YT
(Hou et al. 2019)	✓		✓			118/132	prostate cancer	En	N	YT
(Medina et al. 2020)	✓	✓		✓		113/67	COVID-19	En	N	YT
(Choi and Ko 2021)	✓	✓		✓		902/903	-	En	N	YT
(Shang et al. 2021)	✓	✓				226/665	COVID-19	En	N	TT
(Li et al. 2022)	✓		✓			210/490	health	Ch	N	BB
FakeSV	✓	✓	✓	✓	✓	**1,827/1,827**	**-**	Ch	**Y**	**DY, KS**
FakeTT	✓	✓			✓	**1,172/819**	**-**	En	**Y**	**TT**

- ➢ **FakeSV and FakeTT follows highly-similar data curation and annotation pipelines**
- ➢ **Widely-used in existing works**

Dataset	Time Range	Avg Duration (s)	#Fake	#Real	#All
FakeSV	2017/10-2022/02	39.88	1,810	1,814	3,624
FakeTT	2019/05-2024/03	47.69	1,172	819	1,991

FakeSV: A Multimodal Benchmark with Rich Social Context for Fake News Detection on Short Video Platforms
FakingRecipe: Detecting Fake News on Short Video Platforms from the Perspective of Creative Process

Datasets Overview

Examples from FakeSV

Title: How did this 3.8 kg gold nuggets get taken out by this Chinese lad? Let's take a look!

Title: Emergency by West Lake in Hangzhou this morning, waiting for reinforcements! #special police

Description: In Qinzhou, Guangxi, the wife was brutally beaten in the street for refusing to pay off her husband's gambling debts...What a heinous act!
On-screen Text: A man demands money from his wife to pay off gambling debts. Upon refusal, he assaults her, dragging and then body-slamming her on the road. Family violence again! Urgent need for intervention!

Description: #Shaanxi Shangluo residents transport anti-epidemic supplies with mules: A publicity stunt or full of sincerity? #TopVQuickComment
On-screen Text: Shaanxi Shangluo locals use mules to deliver pandemic supplies: seeking attention or full of compassion? Someone question the efficiency. That's a mutual aid from all sides in times of trouble. Truly Touching.

FakeSV: A Multimodal Benchmark with Rich Social Context for Fake News Detection on Short Video Platforms
FakingRecipe: Detecting Fake News on Short Video Platforms from the Perspective of Creative Process

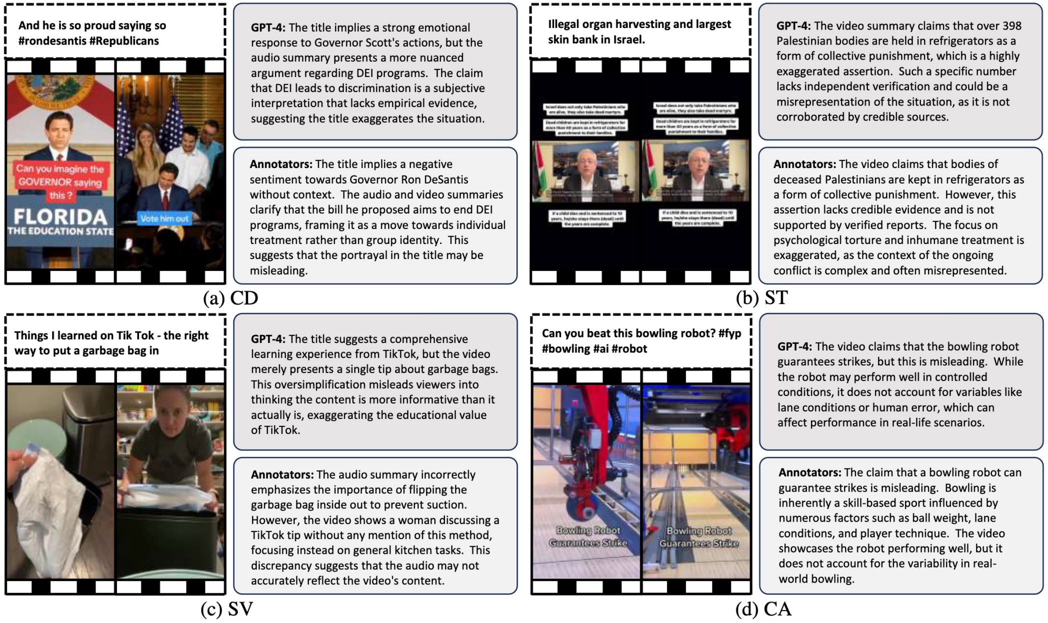
Datasets Overview

Real-world data (cont'd):

- ➢ FakeSV and FakeTT only provide binary classification labels
- ➢ The recent dataset considers natural language explanations

①FakeVE

- ➢ **2,672 samples based on FakeTT and FMNV**
- ➢ **Four types of explanations**
- • *Contextual Dishonesty*
- • *Splice Tampering*
- • *Synthetic Voiceover*
- • *Contrived Absurdity*



Multimodal Fake News Video Explanation: Dataset, Analysis and Evaluation

Datasets Overview

Real-world data (cont'd):

- FakeSV and FakeTT only provide binary classification labels
- The recent dataset considers natural language explanations

②TRUE

- **1,097+1,828 samples from Snopes.com**
- ***Two types of rationales:***
  - Original rationale by human
  - Summary rationales by LLMs

Pioneering Explainable Video Fact-Checking with a New Dataset and Multi-role Multimodal Model Approach

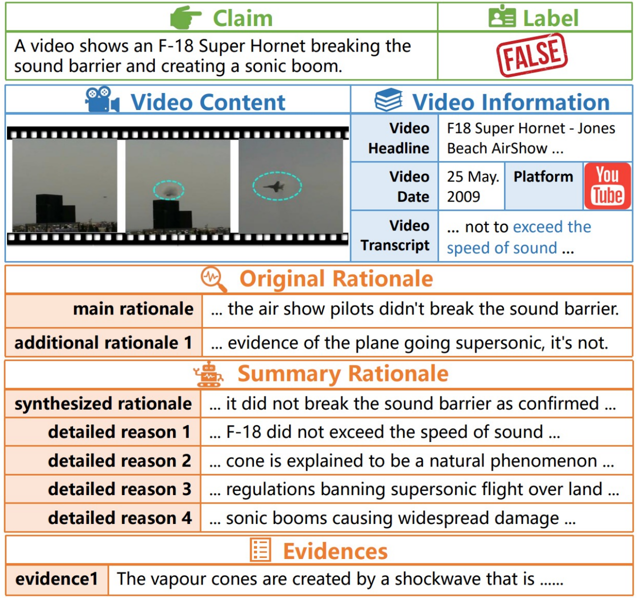


Figure 1: A sample in the proposed TRUE Dataset. It includes the claim, video, and video background information. Besides, three types of annotations are provided: 1) label, 2) evidences, and 3) original and summary rationales.

Datasets Overview

Real-world data (cont'd):

- FakeSV and FakeTT only provide binary classification labels
- The recent dataset considers natural language explanations

③GroundLie360

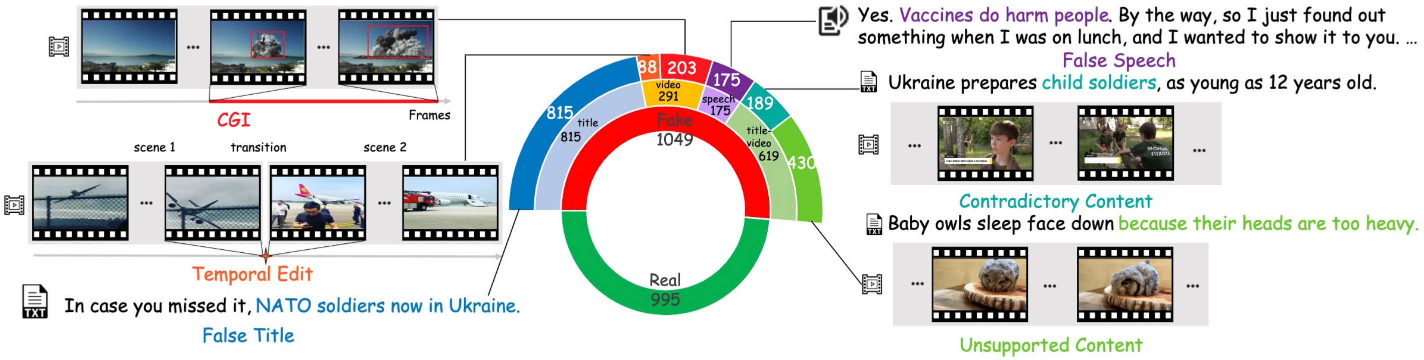


Figure 1: Overview of the GroundLie360 Dataset. Our multi-modal benchmark contains 2,000+ fact-checked videos with fake type and grounding annotations. Fake types include: (1) False Title/False Speech - video title or spoken content containing demonstrably false claims; (2) Temporal Edit - videos altered to distort event chronologies or fabricate deceptive narratives; (3) CGI - digitally manipulated or generated synthetic media; (4) Contradictory Content - text-video semantic mismatches; and (5) Unsupported Content - headlines lacking evidentiary support in video content. The dataset offers a unified benchmark for fake content classification and localization.

A New Dataset and Benchmark for Grounding Multimodal Misinformation

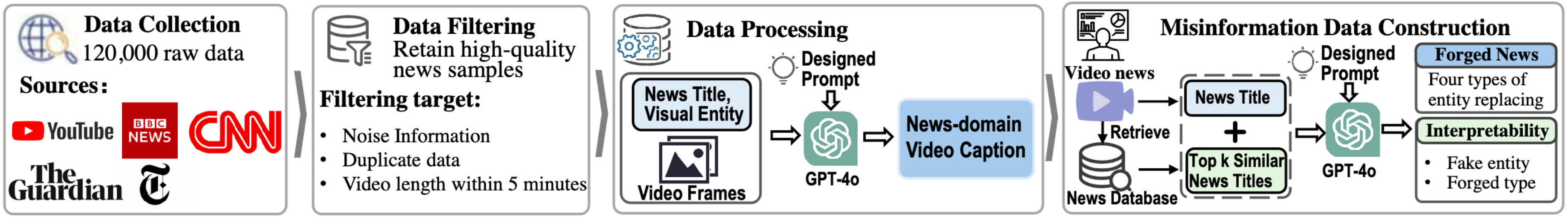
Datasets Overview

Synthesized data

- ➢ Starting from the real news samples, these datasets modify the semantic descriptions to construct fake news samples
- ➢ Of larger scale then real-world ones due to the continence of (M)LLM-assisted data generation

①FakeVV

- ➢ 100k samples from 2006 to 2025
- ➢ Substitute with representation-similar entities based on the retrieved results
- ➢ Input Top3 samples to GPT-4o for generation



FACT-R1: Towards Explainable Video Misinformation Detection with Deep Reasoning

Datasets Overview

Synthesized data

- Starting from the real news samples, these datasets modify the semantic descriptions to construct fake news samples
- Of larger scale then real-world ones due to the continence of (M)LLM-assisted data generation

②Official-NV

- Use 5000 news samples from authoritative sources
- Modify the text description by four strategies

Original Text	Modificated Text
China's booming tea industry imbued with new momentum	China's tea industry surges forward with rejuvenated vitality **(TT)**
The stunning many-coloured landscapes of Xinjiang	The stunning many-coloured landscapes of Anhui **(FT in position)**
Palestinian death toll from Israeli attacks in Gaza, West Bank nears 20,000	Palestinian death toll from Israeli attacks in Gaza, West Bank more than 30,000 **(FT in quantity)**
China seeks to build world's largest national park system	China aims to dismantle extensive national park network **(FT in action)**
Ready... set... GO! This cat sure knows how to win a sprint race	Ready... set... GO! This dog sure knows how to win a sprint race **(FT in object)**

Official-NV: An LLM-Generated News Video Dataset for Multimodal Fake News Detection

Tutorial Outline

3. Characterization of Misinformation Videos

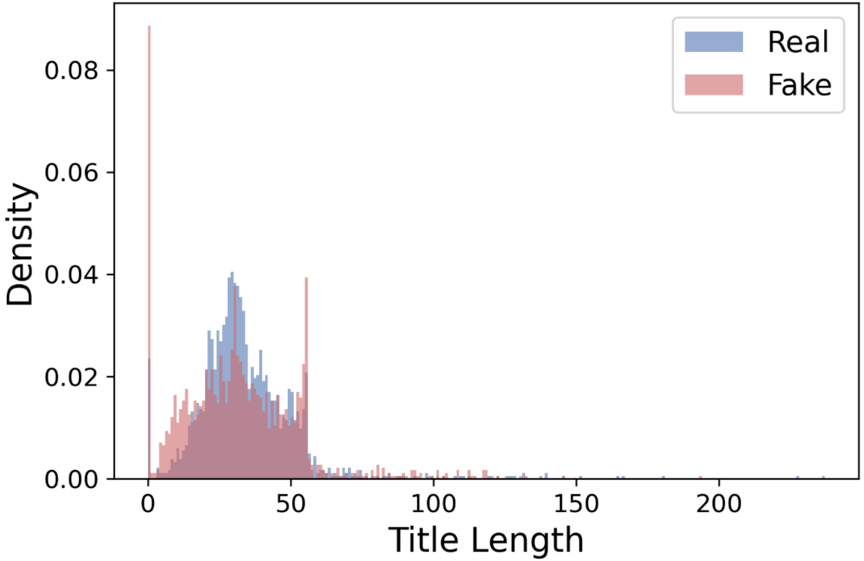
- Definition
- Datasets Overview
- Analysis on FakeSV
  - News Content
  - Social Context
  - Propagation
- Q+A/Discussion

Analysis on FakeSV

- Take FakeSV as an example. This part shows the statistical difference between real and fake news videos.

1. News Content -> Text Length

Fake news videos have shorter and more empty titles, providing less information compared with real news.



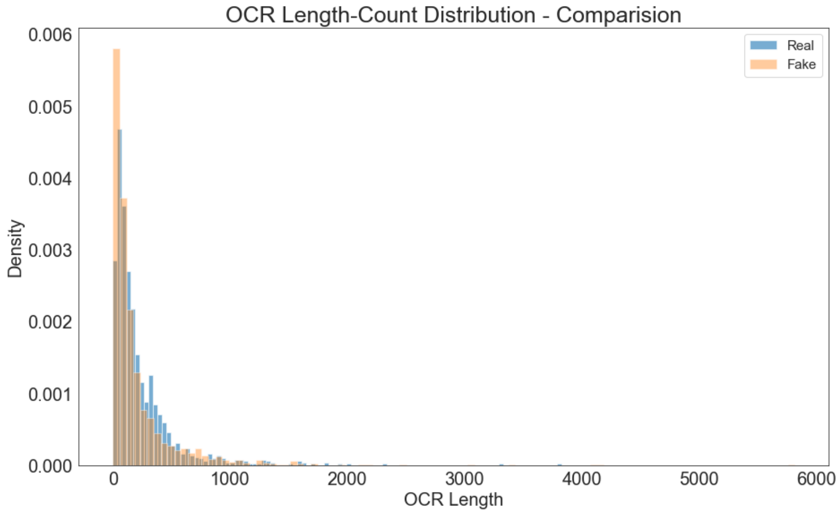
FakeSV: A Multimodal Benchmark with Rich Social Context for Fake News Detection on Short Video Platforms

Analysis on FakeSV

- Take FakeSV as an example. This part shows the statistical difference between real and fake news videos.

1. News Content -> Text Length

For on-screen texts extracted by OCR tools, fake and real news videos show similar distribution, with real news videos more likely to have longer on-screen texts, showing its informativeness.

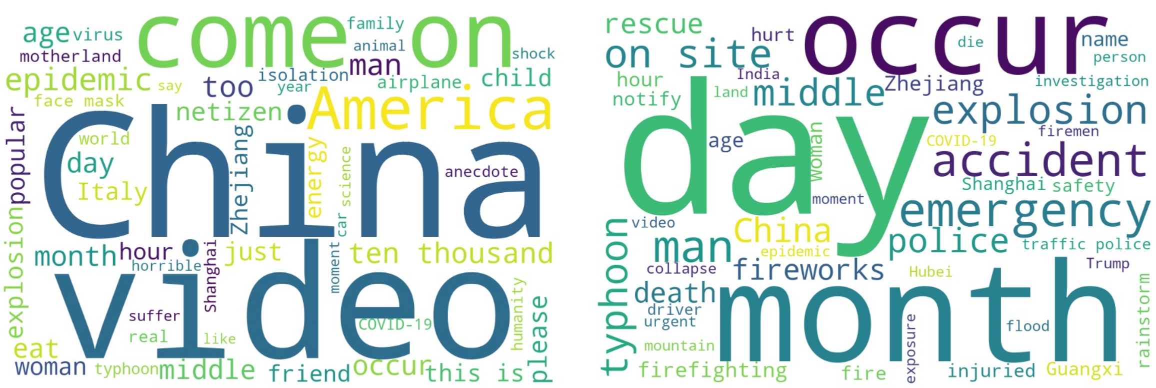


FakeSV: A Multimodal Benchmark with Rich Social Context for Fake News Detection on Short Video Platforms

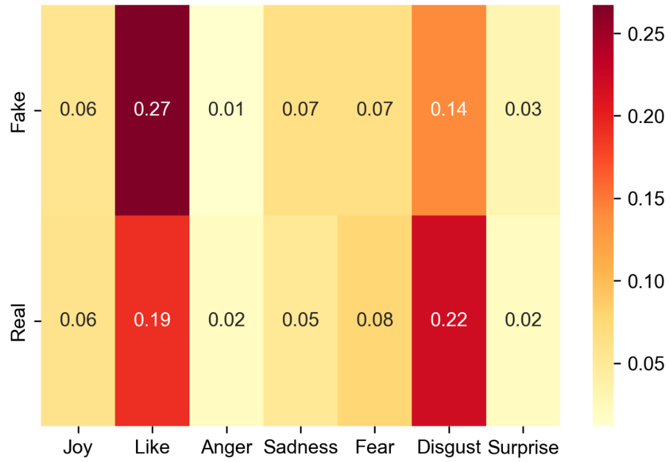
Analysis on FakeSV

1. News Content -> Emotion and Words

- Fake news titles emphasize the word “video” much, **prefer emotional and spoken words**, and cover diverse topics. Real news videos use **more journalese** and focus more on accidents and disasters.
- Based on Affective Lexicon Ontology, we find that fake news titles show **more like** while real news titles show **more disgust**.



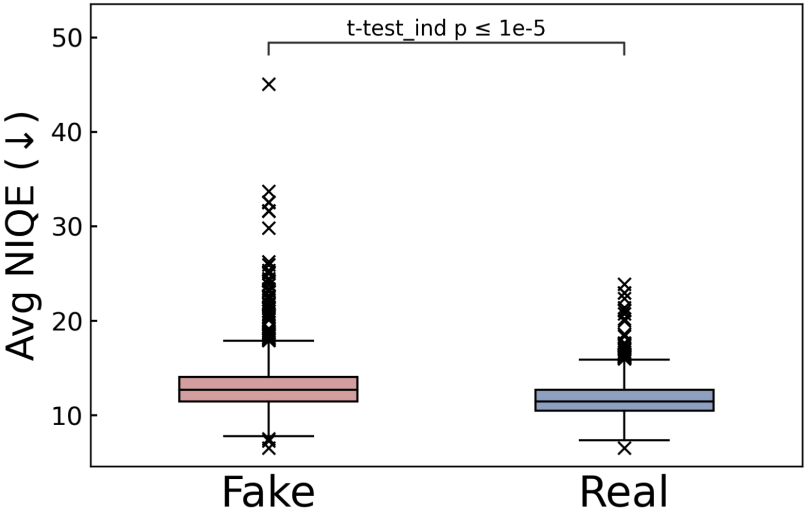
(a) Fake (b) Real



Analysis on FakeSV

1. News Content -> Video

- By employing NIQE on video frames to indirectly measure video quality, we see that **fake news videos have lower quality than real news and contain videos with particularly poor quality**
- This is because the materials are often from unprofessional devices or simply old. AI-generated ones may go to another end: **It might be too clear to be unrealistic**

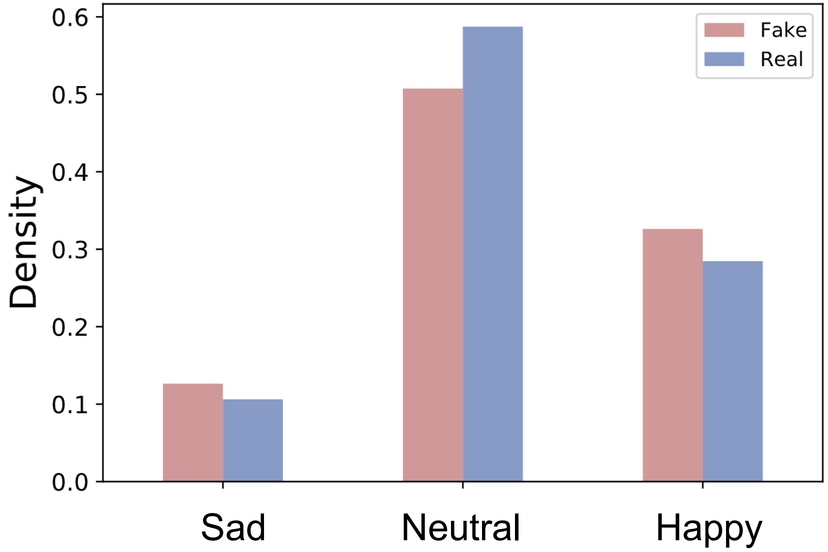


Lower values mean high quality

Analysis on FakeSV

1. News Content -> Audio

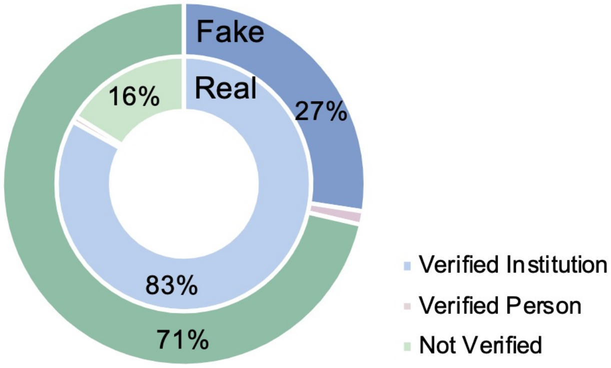
- We analyze the speech emotion by the pre-trained wav2vec model
- **Speech in fake news videos shows more obvious emotional preferences than real news**



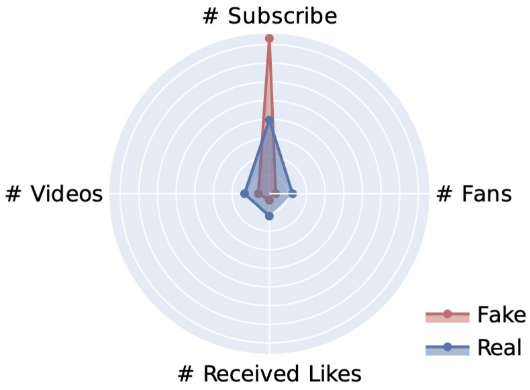
Analysis on FakeSV

2. Social Context-> Publisher Profiles

- The distribution is similar to what have been found on conventional social media like Weibo.
- **Most publishers of real news are verified accounts while most fake news publishers are not**
- **Fake news publishers have more "consuming" behaviors (subscribes) and less "creating" behaviors (published videos, received likes, and fans) than real news publishers**



(a) Authority

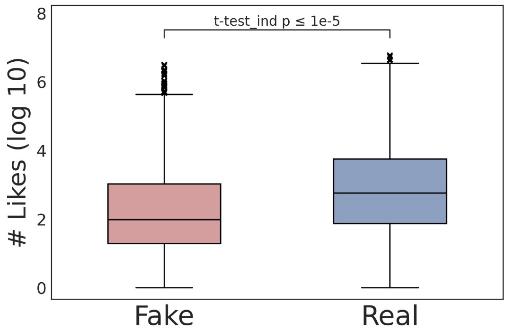


(b) Statistics

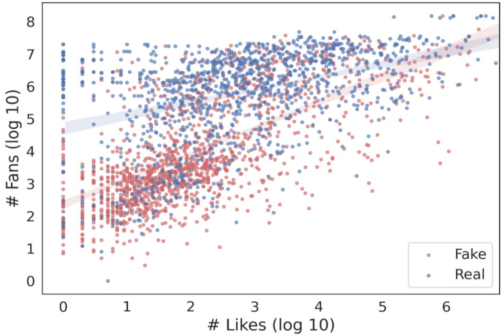
Analysis on FakeSV

2. Social Context-> User Responses

- Real news videos **receive more likes** than fake news, which is intuitive considering that real news publishers have more followers.
- **However, fake news videos receive more likes than real news when their publishers have a similar number of followers, which illustrates that fake news videos are more attractive than real news**



(a) Number of likes

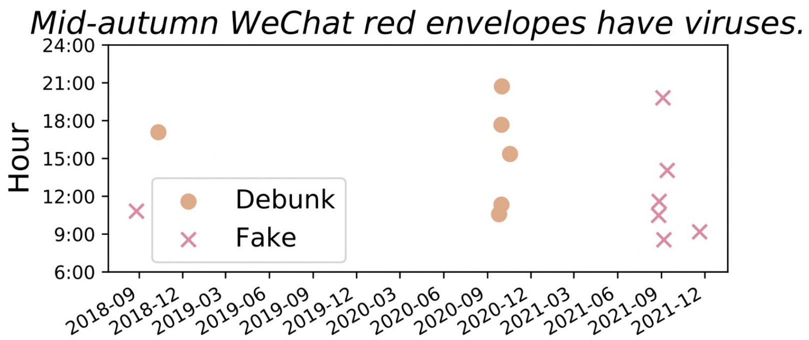


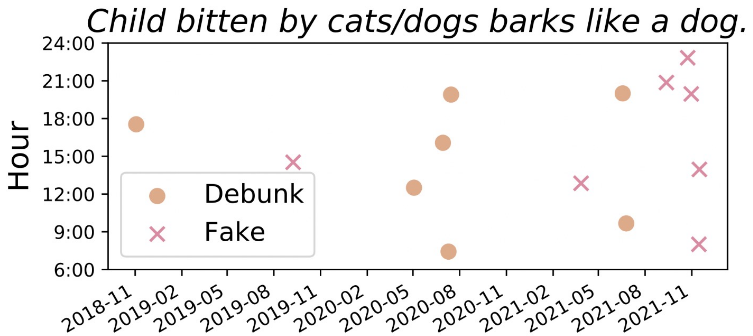
(b) Relationship between the number of publisher fans and likes.

Analysis on FakeSV

3. Propagation-> Temporal Distribution

- Fake news that has been previously debunked can still spread
- **For 434 events with debunking videos, 39% of them have fake news videos emerged after the debunking videos were posted, especially the current or long-standing hot event**

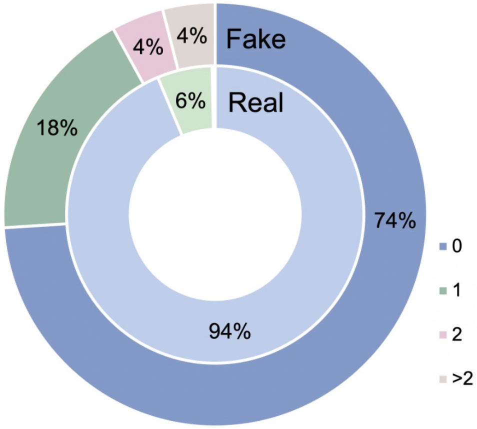




Analysis on FakeSV

3. Propagation-> Video duplication

- With the convenience of video editing functions provided by these platforms, people tend to edit and re-upload the videos, usually with no mention of the source.
- **Using pHash on the video covers, we find that fake news videos have higher repetition while real news videos are more diverse.** This is due to the fact that real events will receive various images/videos from different witnesses and sources.



Detection Part I: Human-Edited Misinformation

Section 4

Tutorial Outline

Detection Part I: Human-Edited Misinformation

- Signal-based detection
  - Editing Traces
  - Generation Traces
- Semantic-based detection
  - Seek clues within the sample's own content
  - Seek clues from external information
- Intent-based detection
  - Social context
- Clue integration for misinformation video detection
  - Parallel Integration
  - Sequential Integration
- Q+A/Discussion

Signal-based detection

Misinformation videos **often contain manipulated or generated video and audio content** in which the forgery procedure often **leads to traces in underlying digital signals.**

Editing

Alterations on existing data of video and audio modality

Generation

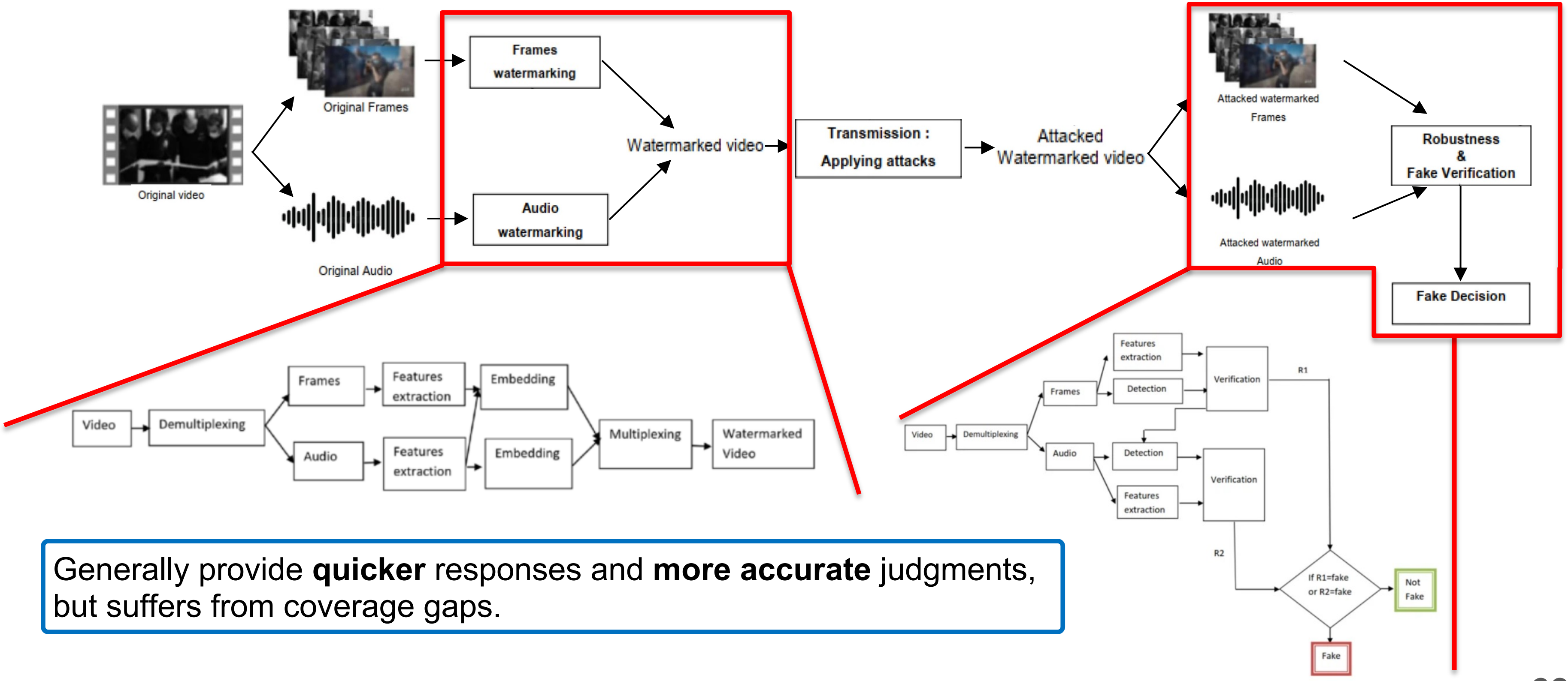
Directly generate complete vivid videos with forged human faces or voices

Cases come from https://youtu.be/RVrANMAO7Sc

Editing Traces

Active detection

Pre-embed and extract digital watermarks for detection



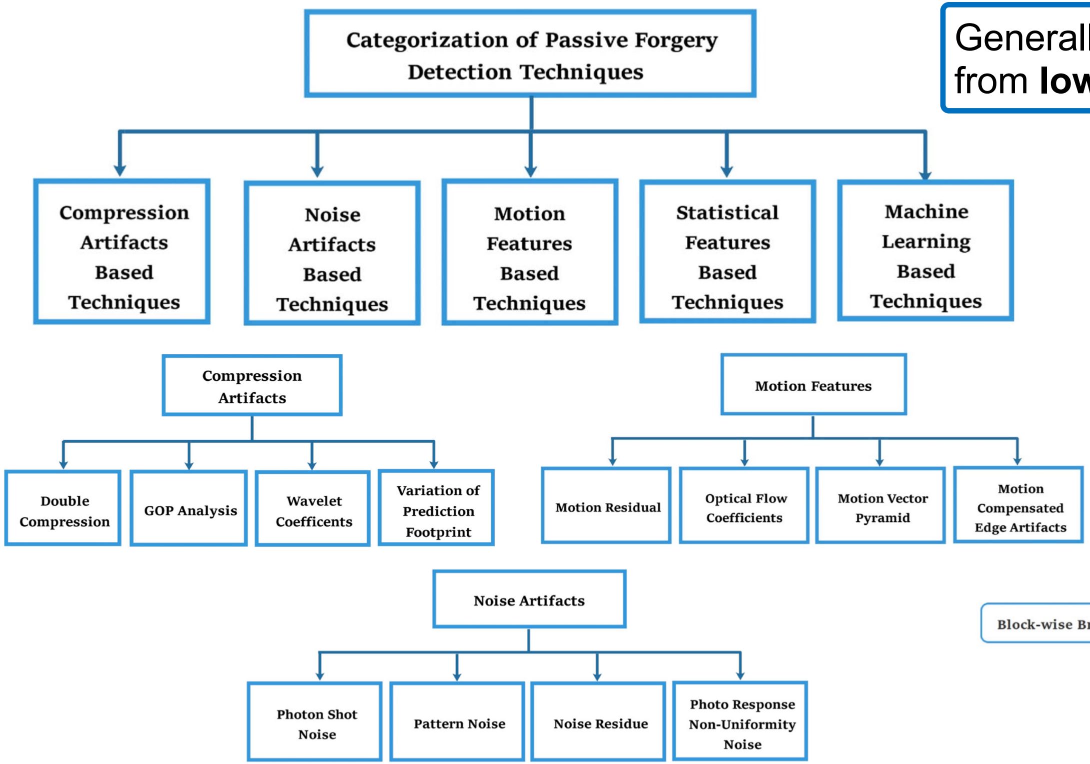
Generally provide **quicker** responses and **more accurate** judgments, but suffers from coverage gaps.

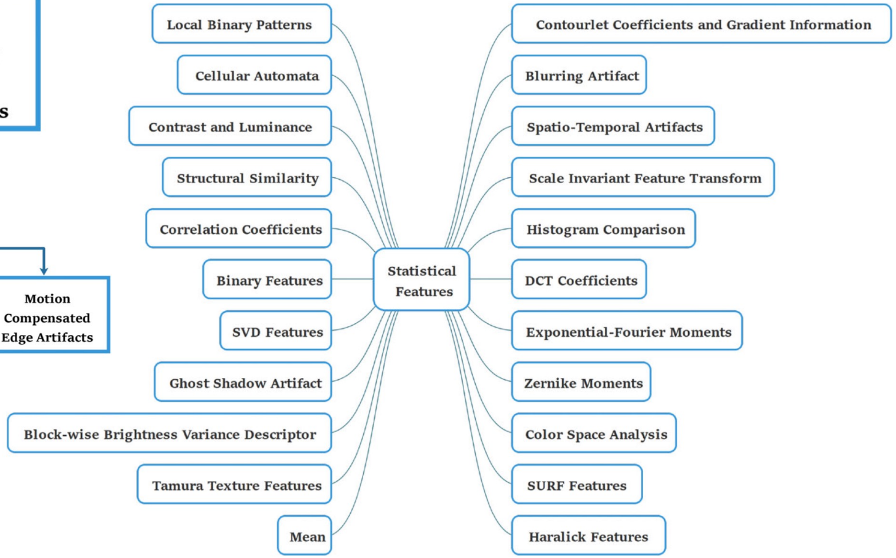
Tarhouni, Nesrine, et al. "Fake COVID-19 videos detector based on frames and audio watermarking." Multimedia Systems 2023

Editing Traces

Passive detection Use the characteristics of the digital video itself for detection

Generally provide **broader applicability**, but suffer from **lower reliability** under distribution shifts..





Shelke et,al. "A comprehensive survey on passive techniques for digital video forgery detection." Multimedia Tools and Applications 2021.

Generation traces

Mining spatial-temporal and spectral-prosodic traces from video and audio for detection

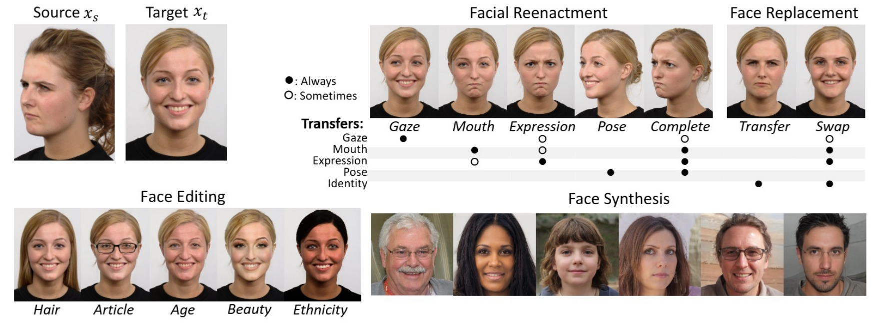


Fig 1. Visual deepfakes examples.

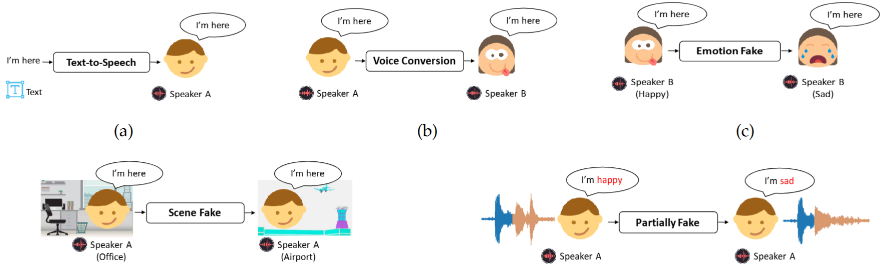


Fig 2. Deepfake audio types.

➢ Deepfake Video Detection Clues:

- Boundary artifacts between fake face & background
- Inconsistent lighting, warping, background
- Generator or sensor "fingerprints"
- Unrealistic motion or emotion patterns
- Missing biological signals
- Temporal inconsistency
- Deep features learned from end-to-end models

- Short-term spectral features
- Phase / group-delay features
- Long-term spectral features
- Prosodic features

- intrinsic (a)synchronization between video and audio frames
- ……

Mirsky, Yisroel, and Wenke Lee. "The creation and detection of deepfakes: A survey." ACM CSUR 2021.
Yi, Jiangyan, et al. "Audio deepfake detection: A survey." arXiv 2023.

Tutorial Outline

Detection Part I: Human-Edited Misinformation

- Signal-based detection
  - Editing Traces
  - Generation Traces
- Semantic-based detection
  - Seek clues within the sample's own content
  - Seek clues from external information
- Intent-based detection
  - Social context
- Clue integration for misinformation video detection
  - Parallel Integration
  - Sequential Integration
- Q+A/Discussion

Semantic-based detection

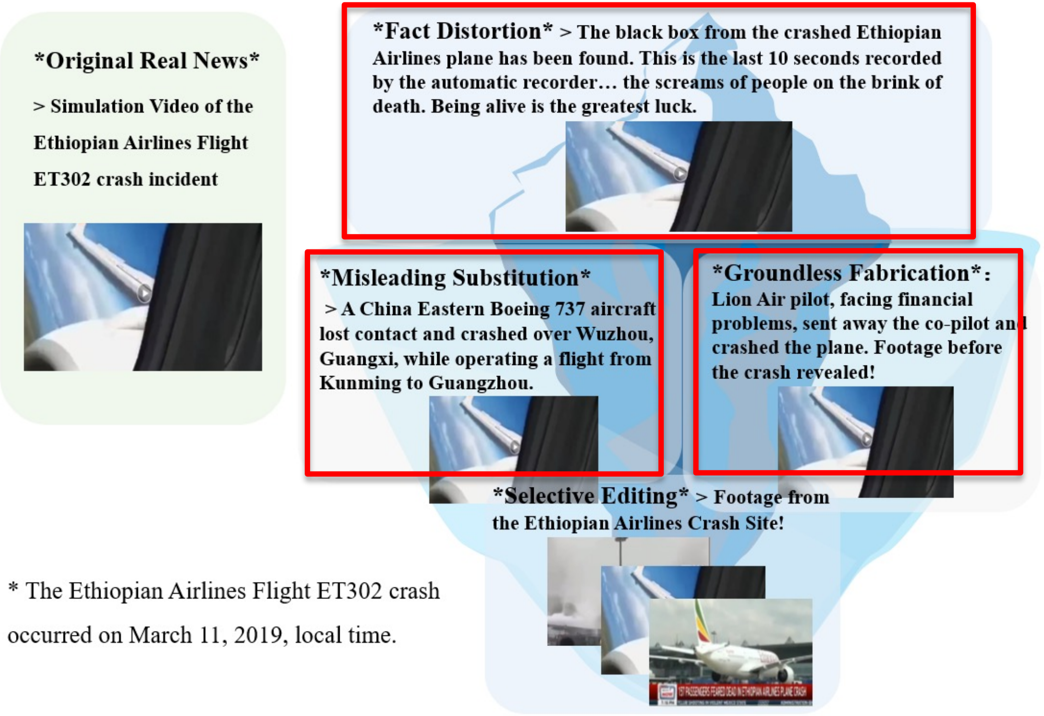
The falsehood is conveyed through **incorrect semantic changes against the truth**.

💡Even if **editing or generation traces** are detected, it does **not necessarily** mean that the video **conveys misinformation**.

A video that is technically untampered can be employed in a deceptive manner:

- Fact Distortion
- Misleading Substitution
- Groundless Fabrication
- ……

👆 In the UGC era, most misinformation arises from such semantic manipulations



* The Ethiopian Airlines Flight ET302 crash occurred on March 11, 2019, local time.

Seek clues within the sample's own content

Early works most leverage textual information for misinformation detection

Textual information:

- Video description
- Title
- Subtitles
- Transcriptions
- ……

Hand-crafted features:

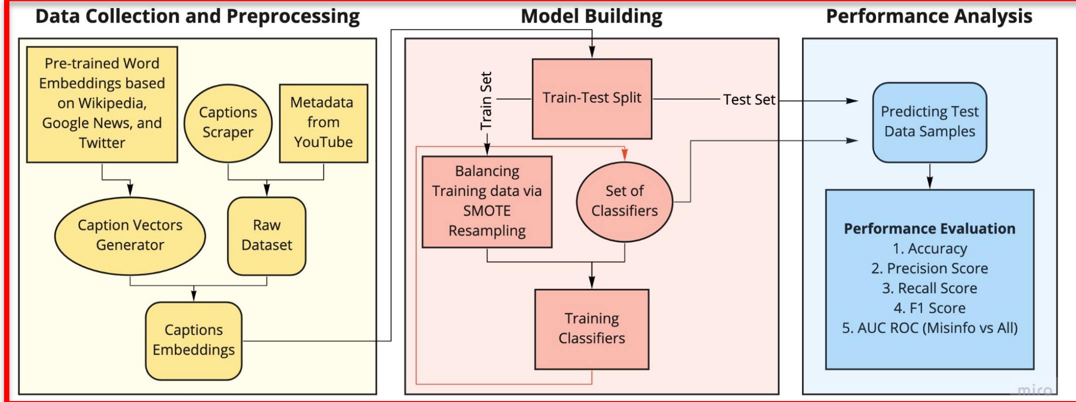
- Basic statistical attributes
- Specific expressions
- Corpus-aware features(N-grams, TF-IDF, LIWC)

Continuous representation features:

- Pre-trained static embeddings
- Task-specific encoders

From video description
05: text length
06: number of words
07–08: contains question/exclamation mark (Boolean)
09–10: contains 1st/3rd person pronoun (Boolean)
11: number of uppercase characters
12–13: number of positive/negative sentiment words
14: number of slang words
15: has ":" symbol (Boolean)
16–17: number of question/exclamation marks

Title&Transcript	text length the number of words contains question/exclamation mark (Boolean) contains 1st/3rd person pronoun (Boolean) the number of positive/negative sentiment words has ":" symbol (Boolean) the number of question/exclamation marks has clickbait phrase (Boolean) sentiment polarity the number of modal particles the number of personal pronouns tf-idf Ngrams LIWC



Papadopoulou, Olga, et al. "A corpus of debunked and verified user-generated videos." Online information review 2019.
Jagtap, Raj, et al. "Misinformation detection on YouTube using video captions." arXiv 2021

Seek clues within the sample's own content

Aggregate heterogeneous information across different modalities (beyond textual content)

Visual Feature

Acoustic Feature

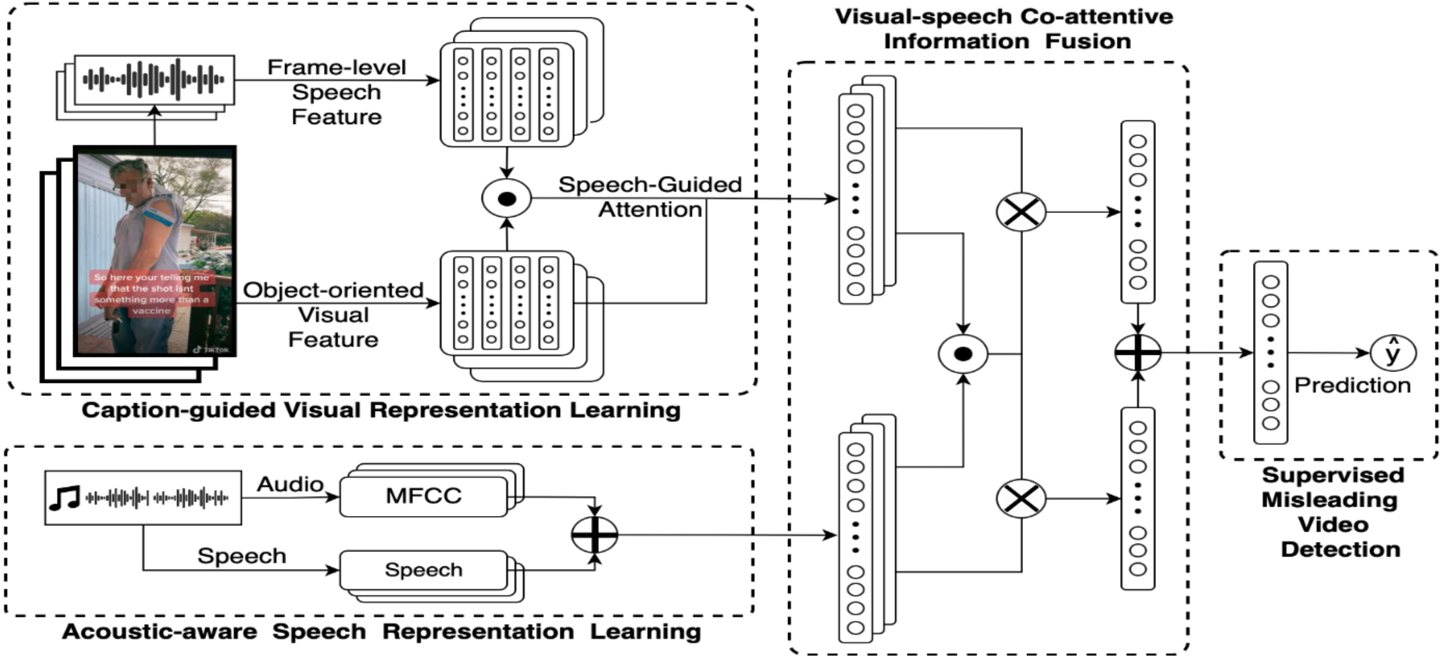
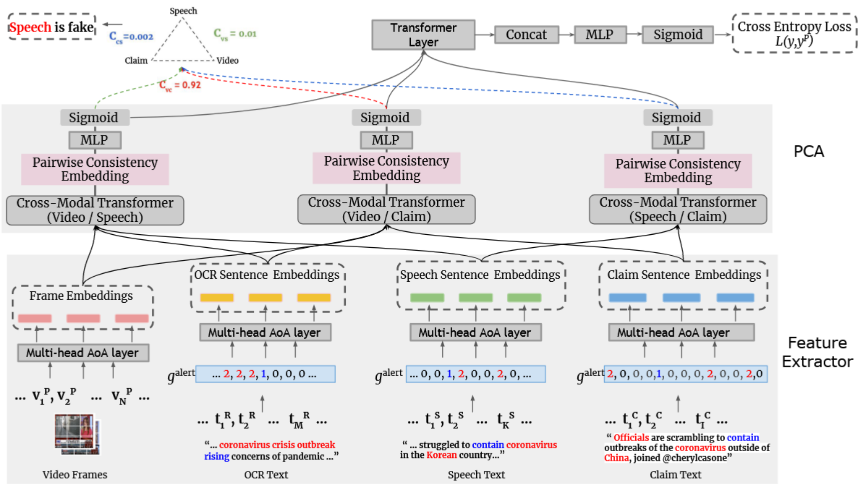


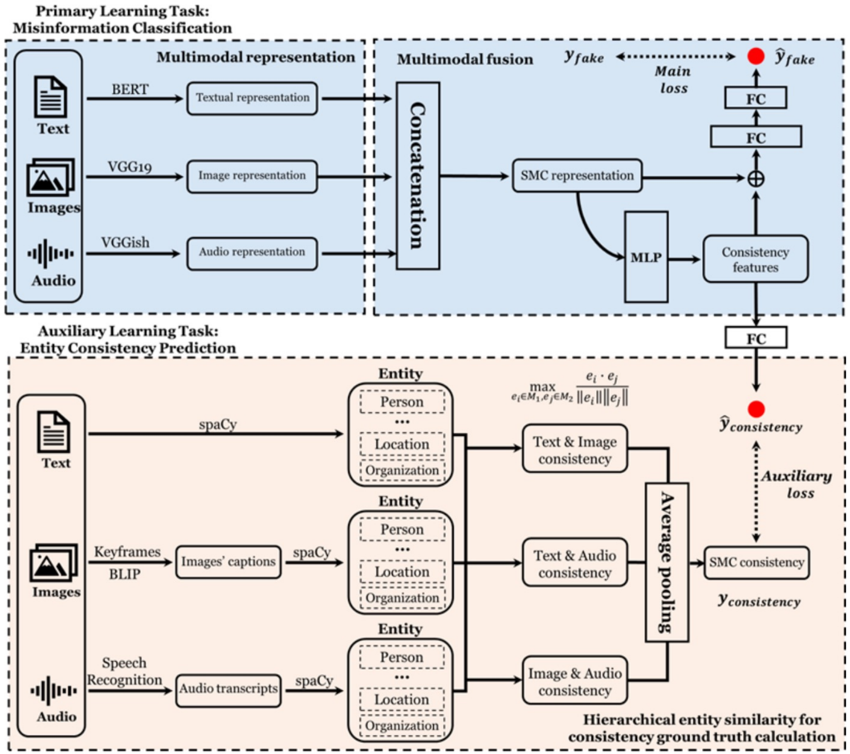
Figure 3: Overview of the TikTec Framework

Shang, Lanyu, et al. "A multimodal misinformation detector for covid-19 short videos on tiktok." IEEE Big Data, 2021.

Seek clues within the sample's own content

Leverage cross-modal correlation: Find mismatches between modalities (video-text-audio)





👆Utilize **embedding consistency** across three different modalities for detection

👉Utilize **entity consistency** across three different modalities for detection

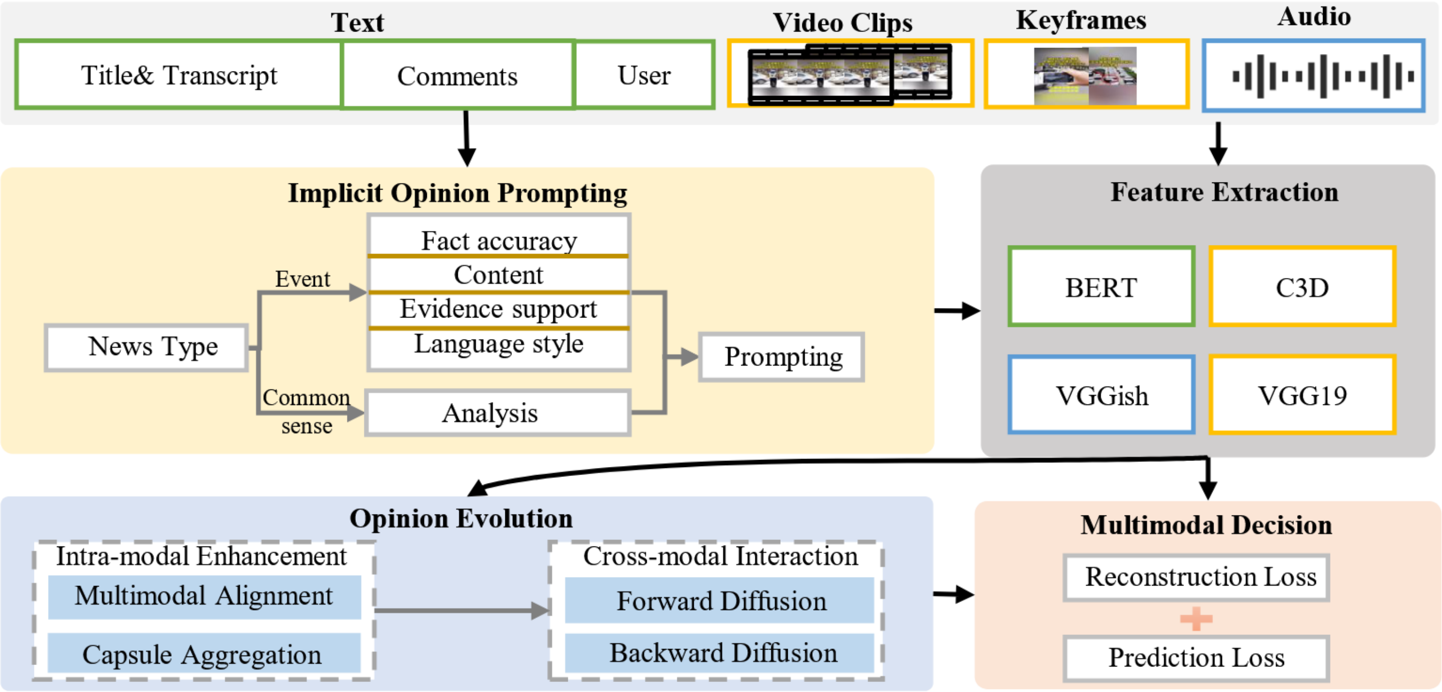
Liu, Fuxiao, Yaser Yacoob, and Abhinav Shrivastava. "COVID-VTS: Fact Extraction and Verification on Short Video Platforms." EACL. 2023.
Fu, Zhe, et al. "Detecting misinformation in multimedia content through cross-modal entity consistency: A dual learning approach." arXiv 2024.

Seek clues within the sample's own content

Deepen cross-modal correlation clues mining: Uncover implicit opinions across modalities

💡Each modality carries an **implicit credibility attitude**; without cross-modal propagation, **local** deceptive cues fail to influence the **global** decision.

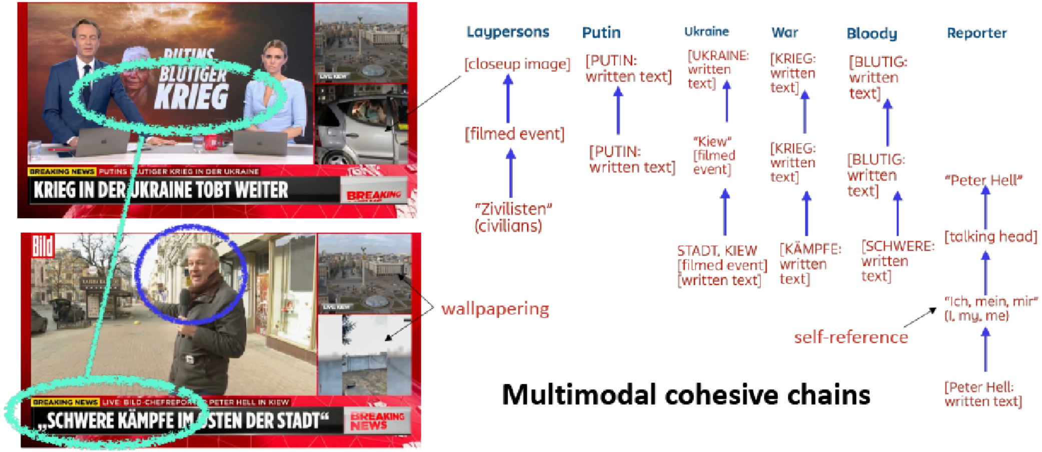
👉Simulate mutual reinforcement among modal opinions through a **diffusion process**



Zong, Linlin, et al. "Unveiling opinion evolution via prompting and diffusion for short video fake news detection." ACL Findings 2024.

Seek clues within the sample's own content

Another analysis perspective: **News as Narrative**



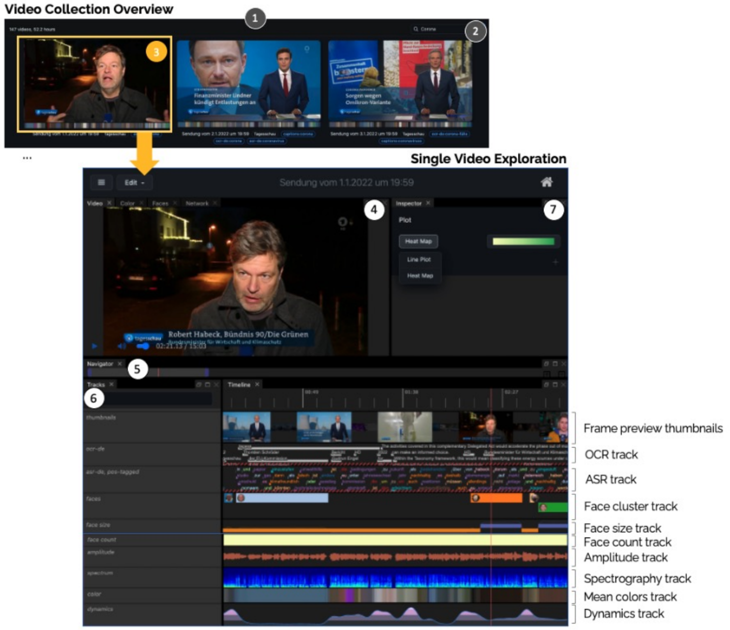


Abb. 1: Zoetrope UI – Video collection overview sections: (1) overview screen, (2) text search, (3) video scrubbing and movie barcode; Single video exploration sections: (4) player pane, (5) navigator and timeline pane, (6) tracks pane and (7) properties pane.

💡Two pivotal aspects of Narrative Theory: analyzing the **"what"** (the content of the story)and the **"how"** (the strategy of storytelling)

👉Disinformation TV news videos has distinguish narratives.

Bateman, John A., and Chiao-I. Tseng. "Multimodal discourse analysis as a method for revealing narrative strategies in news videos."
Liebl, Bernhard, and Manuel Burghardt. Designing a Prototype for Visual Exploration of Narrative Patterns in NewsVideos.

Seek clues within the sample's own content

Enhance misinformation video detection by analyzing the narrative creation process.

👉Analyze the creative process behind misinformation videos:

Phase I – Material Selection: Fake news exhibits **emotional bias** and **semantic selectivity** when choosing video materials.

Phase II – Material Editing: When **spatially** imposing text and **temporally** splicing materials, fake news tend to adopt a relatively **simple arrangement**.

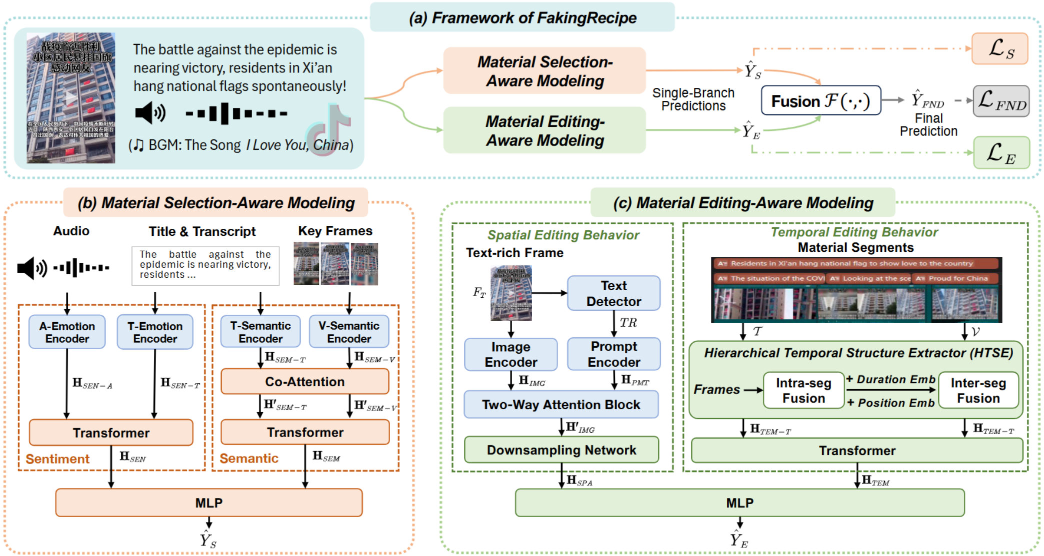
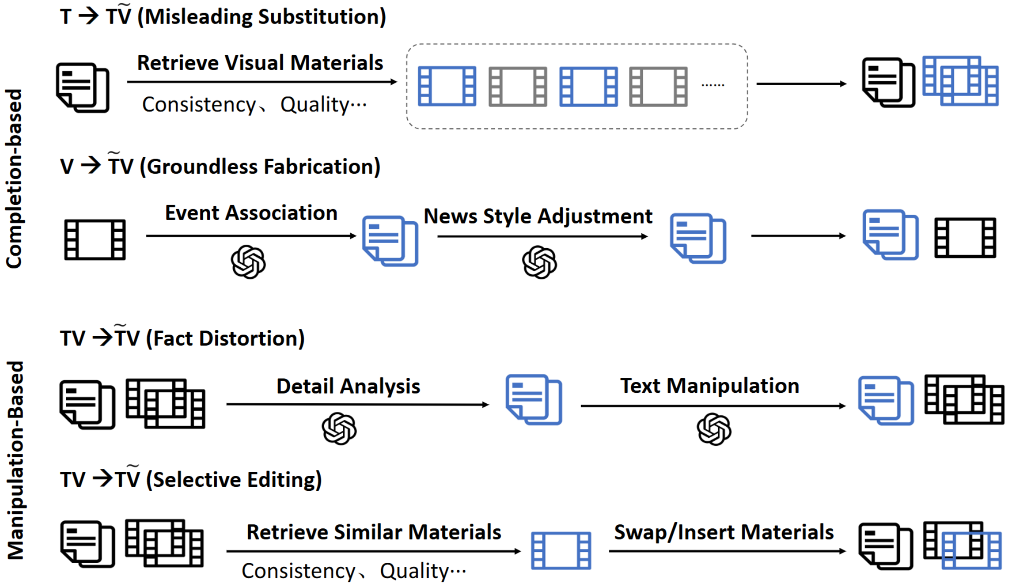


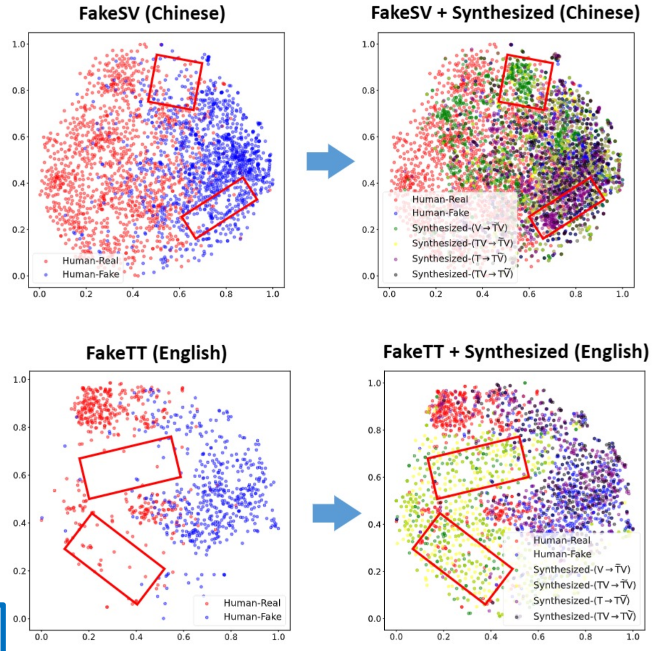
Figure 6: Overview of the proposed FakingRecipe model. (a) Overall framework: The news video is processed through dual perspectives, with a late fusion strategy employed to integrate clues for final prediction. (b) Material Selection-Aware Modeling (MSAM) module: Extracts clues from both sentimental and semantic aspects. (c) Material Editing-Aware Modeling (MEAM) module: Extracts clues based on spatial and temporal aspects. $\mathcal{F}(\cdot,\cdot)$ denotes the fusion function. The parameters in the modules in blue are frozen and others are trainable. The overall model is trained under the supervision of the loss functions

Bu, Yuyan, et al. "FakingRecipe: Detecting fake news on short video platforms from the perspective of creative process." ACM MM 2024.

Seek clues within the sample's own content

Enhance misinfo video detection via data augmentation: Simulating typical creative processes



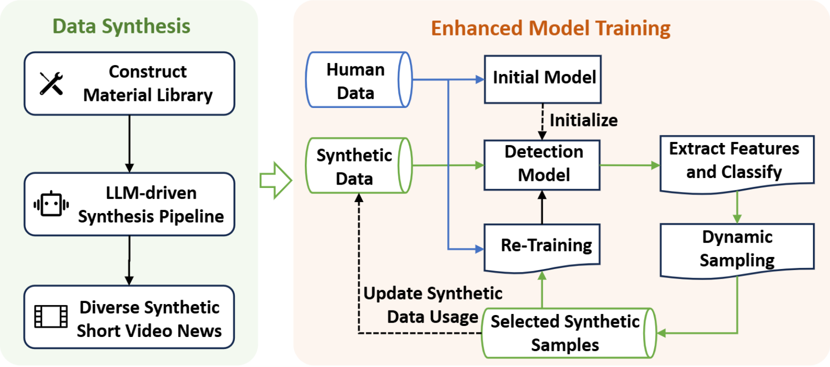


👉The synthesized samples not only align with human-crafted fakes but also enrich underrepresented regions of the feature space

Bu, Yuyan, et al. "Enhancing Fake News Video Detection via LLM-Driven Creative Process Simulation." CIKM 2025.

Seek clues within the sample's own content

Enhance misinfo video detection via data augmentation: Simulating typical creative processes



Model	FakeSV				FakeTT			
	Acc	F1	Prec.	Rec.	Acc	F1	Prec.	Rec.
MMVD	75.83	75.33	75.51	75.21	67.50	66.20	66.51	68.43
w/AgentAug$_{RAN}$	73.80	73.22	73.33	73.13	63.57	61.69	61.84	63.11
w/AgentAug$_{BAL}$	76.01	75.58	75.86	75.45	66.43	65.32	65.55	66.94
w/AgentAug$_{AL}$	**77.12**	**76.69**	**76.93**	**76.55**	**68.57**	**67.26**	**67.56**	**69.71**
FANVM	78.41	77.89	78.25	77.70	71.57	70.21	70.21	72.63
w/AgentAug$_{RAN}$	78.78	77.63	80.08	77.16	67.22	66.77	69.18	71.42
w/AgentAug$_{BAL}$	79.34	78.31	80.42	77.84	68.23	67.99	71.57	73.70
w/AgentAug$_{AL}$	**81.37**	**80.42**	**82.73**	**79.88**	**75.25**	**74.02**	**73.84**	**76.65**
SVFEND	80.88	80.54	80.83	80.51	77.14	75.63	75.12	77.56
w/AgentAug$_{RAN}$	82.66	81.94	83.54	81.44	79.72	77.72	77.33	78.24
w/AgentAug$_{BAL}$	81.92	80.85	83.96	80.23	77.58	76.00	75.43	77.40
w/AgentAug$_{AL}$	**83.76**	**82.98**	**85.20**	**82.38**	**80.43**	**78.61**	**78.12**	**79.29**
SVRPM	81.34	81.11	81.38	80.97	81.79	79.42	79.67	79.19
w/AgentAug$_{RAN}$	79.57	79.33	79.61	79.20	80.00	77.34	77.49	77.19
w/AgentAug$_{BAL}$	81.73	81.58	82.06	81.49	82.14	80.28	**81.36**	79.58
w/AgentAug$_{AL}$	**83.10**	**82.89**	**83.16**	**82.70**	**82.86**	**80.57**	80.75	**80.41**

👆By integrating **active learning** to select potentially useful augmented samples,
👉the framework **consistently boosts** short-video fake news detection performance.

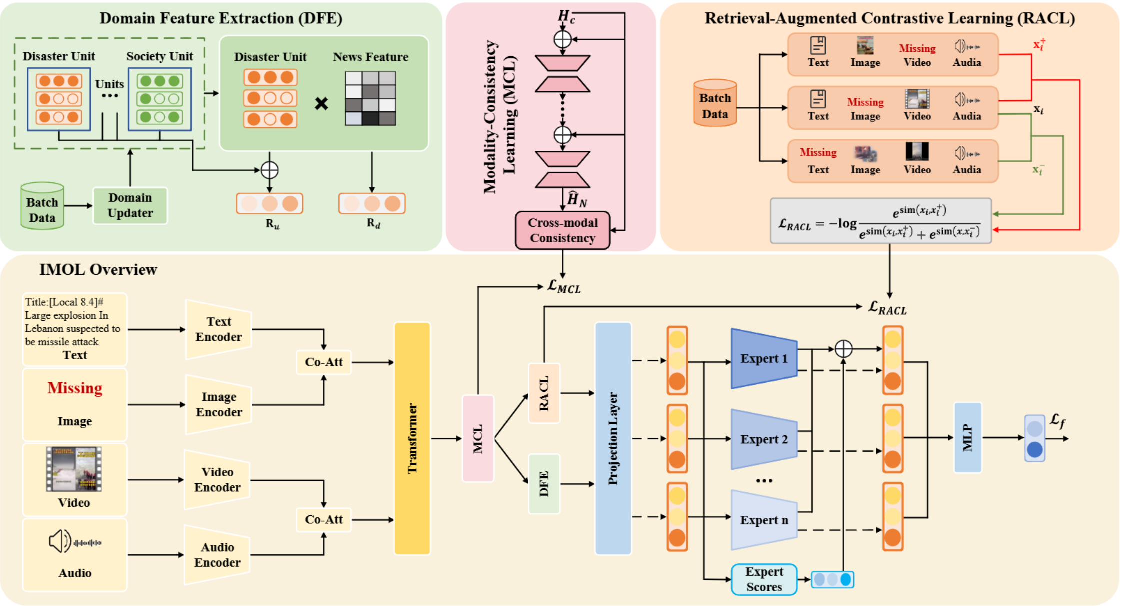
Bu, Yuyan, et al. "Enhancing Fake News Video Detection via LLM-Driven Creative Process Simulation." CIKM 2025.

Seek clues within the sample's own content

More robust detection： Consider incomplete modality conditions

Figure 1: Examples of modality patterns in news videos. Each news piece consists of four modalities: Video (V), Audio (A), Text (T), and Image (I). However, real-world news is often *modality-incomplete*, where one or more modalities may be missing (MV: Missing Video, MA: Missing Audio, MT: Missing Text, MI: Missing Image).

Jointly modeling cross-modal reconstruction and cross-sample reasoning →robust and generalizable fake news video detection.



Zeng, Zhi, et al. IMOL: Incomplete-Modality-Tolerant Learning for Multi-Domain Fake News Video Detection. ACL 2025

Seek clues within the sample's own content

More robust detection: Mitigate modality bias & Adaptively select reliable feature

👇Debiases static, dynamic, and social views through causal and counterfactual reasoning

👇Adaptively trusts the most reliable modalities and selecting the according modality experts

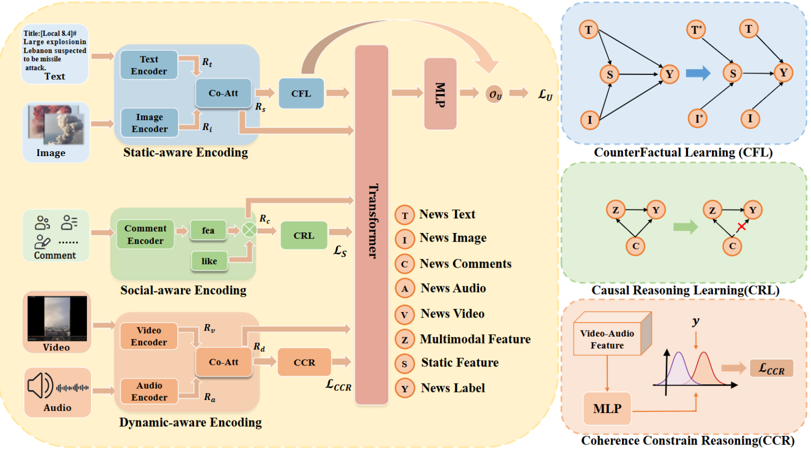


Figure 4: Overview of proposed Multimodal Multi-View Debiasing framework. The CFL, CCR and CRL mitigate the static, dynamic and social biases during multimodal fusion, respectively. Then the MMVD is learned to determine whether the news video is fake or not.

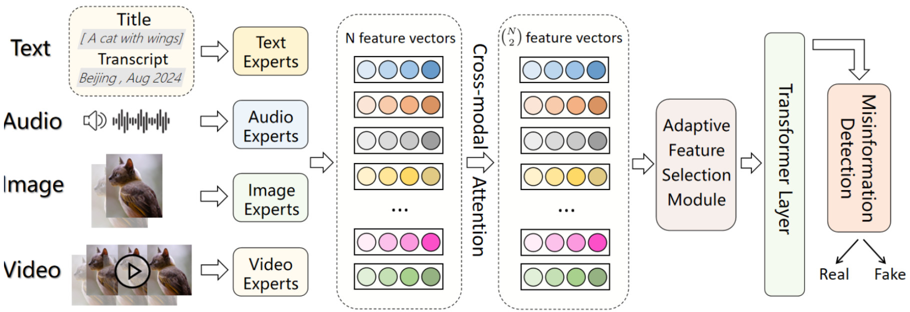


Figure 1: Architecture of the multimodal misinformation detection framework MisD-MoE

Zeng, Zhi, et al. Mitigating World Biases: A Multimodal Multi-View DebiasingFramework for Fake News Video Detection. ACM MM 2025
Liu, Moyang, et al. "MisD-MoE: A Multimodal Misinformation Detection Framework with Adaptive Feature Selection." NeurIPS Workshop 2024.

Seek clues from external information

Integrate the neighborhood relationship of new videos belonging to the same event

💡News videos from different perspectives regarding the same event contain complementary or contradictory information

Utilize **debunking videos** to rectify false negative predictions

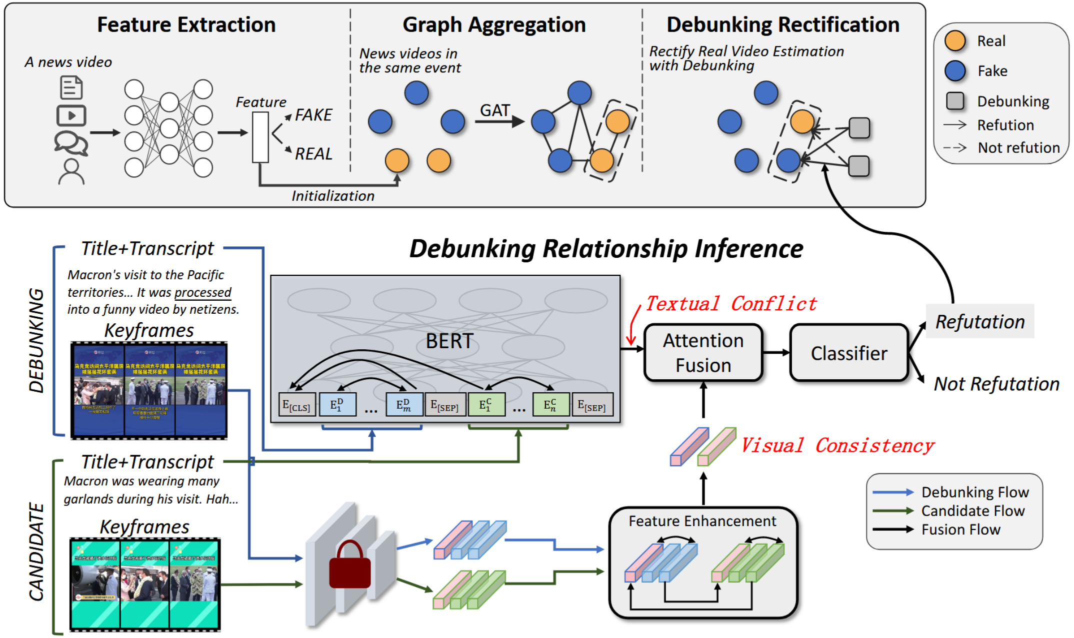


Figure 2: Architecture of the proposed framework NEED. The first row indicates the three stages in NEED, including feature extraction, graph aggregation, and debunking rectification. To realize the debunking rectification, debunking relation inference (the second row) is introduced to determine the refutation relationship.

Qi, Peng, et al. "Two heads are better than one: Improving fake news video detection by correlating with neighbors." ACL Findings 2023.

Seek clues from external information

Model latent social and cascade relationships for fake news detection

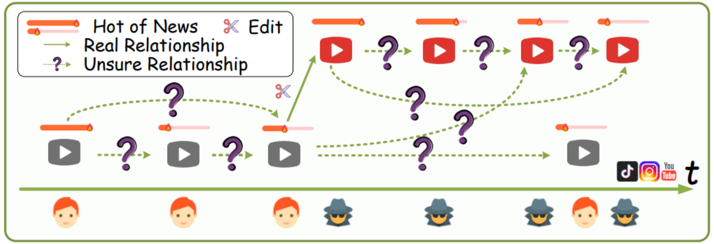


Fig 1. News dissemination relationship is unsure in TikTok, Instagram, and YouTube

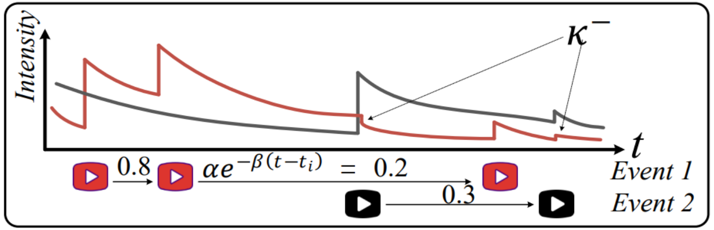
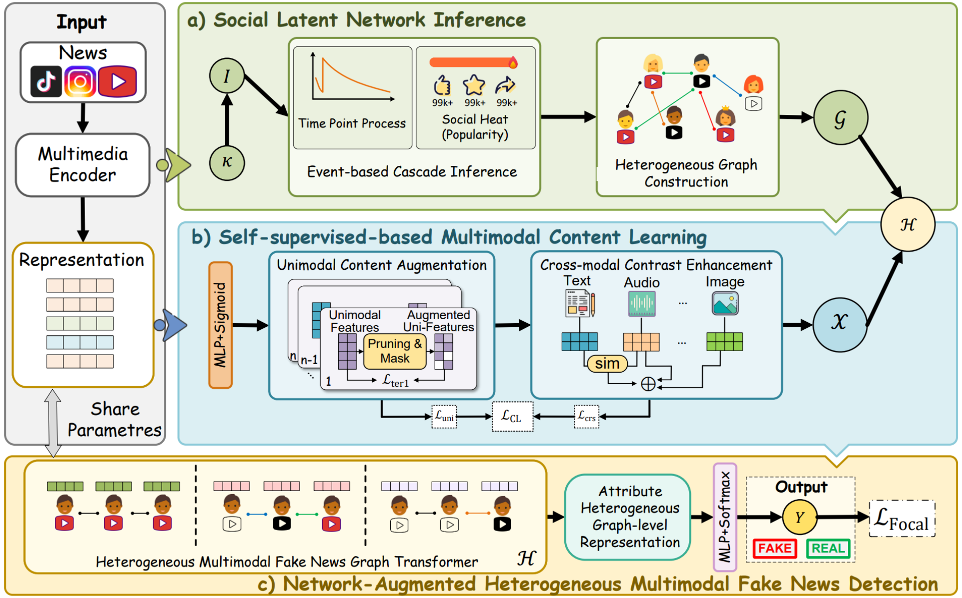


Fig 2. Illustrate of event-based cascade influence.

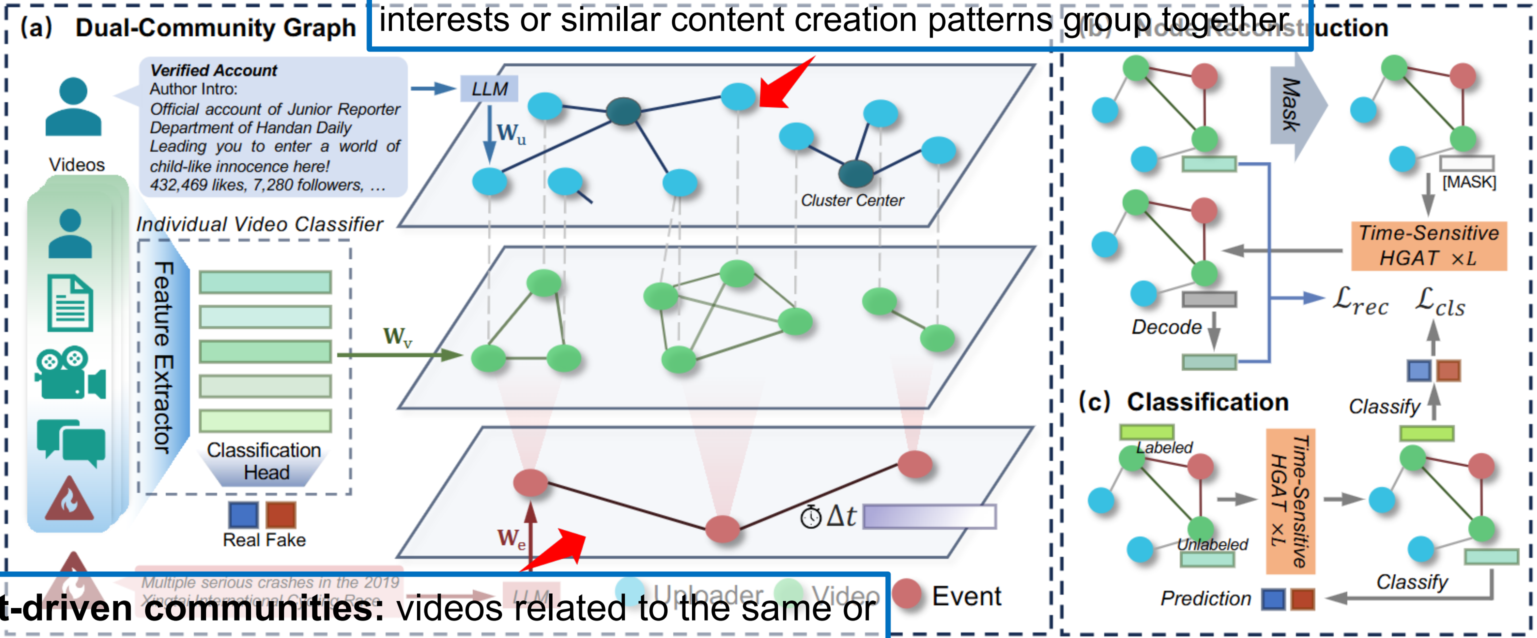


👆Infer latent social cascades via event-based temporal modeling and heterogeneous graph construction

Mingxin, Li, et al. Learning Complex Heterogeneous Multimodal Fake News via Social Latent Network Inference. AAAI 2025

Seek clues from external information

Integrates dual-community patterns: **Uploaders'** and **Event-driven** communities



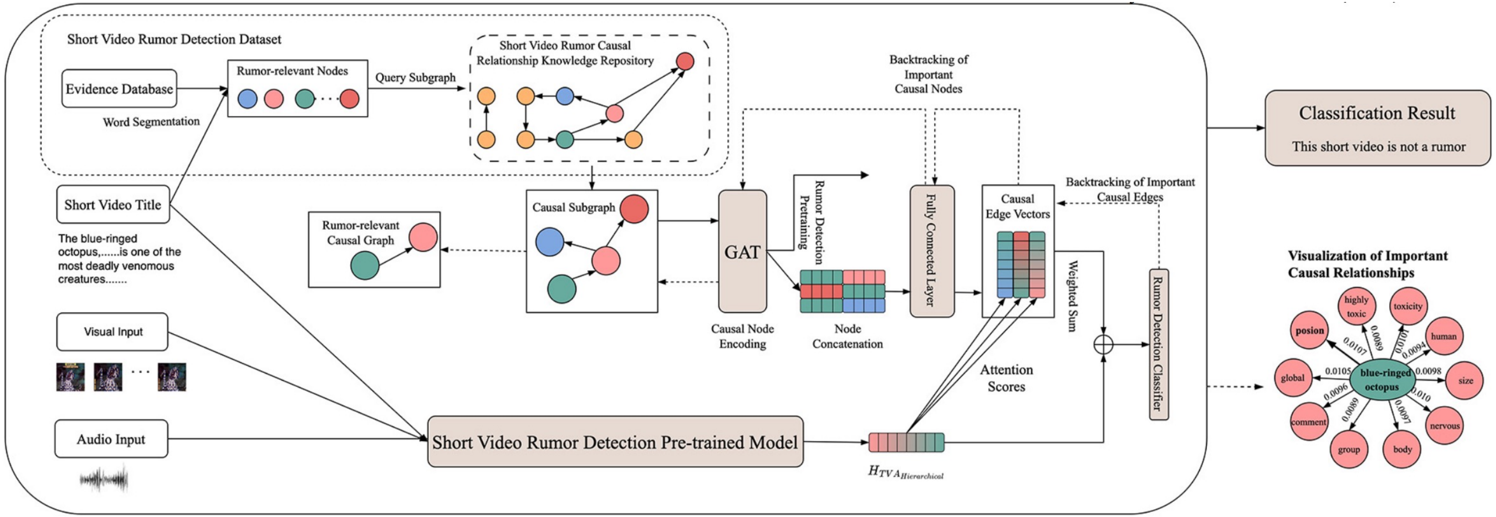
💡 **Uploader communities**: uploaders with shared interests or similar content creation patterns group together

💡 **Event-driven communities:** videos related to the same or semantically similar public events form localized clusters.

Gong, Haisong, et al. "Mining the Social Fabric: Unveiling Communities for Fake News Detection in Short Videos." arXiv 2025

Seek clues from external information

Introduce external knowledge through **causal relationship graphs**



💡Construct causal relationships between entities and integrates the causal subgraphs for the interpretation of knowledge distortion

Cao, Donglin, et al. "Short video rumor detection based on causal graph." Information Sciences 703 (2025): 121941.

Seek clues from external information

Introduce external knowledge through Large Language Model

💡LLMs serve as flexible knowledge bridges, transforming external information into structured reasoning evidence for multimodal verification.

👇Design multi-stage pipelines where LLMs refine content, retrieve domain knowledge, and reason to generate verifiable explanations.

👇Build multi-role LLM frameworks that decompose fact-checking into retrieval, verification, and reasoning stages.

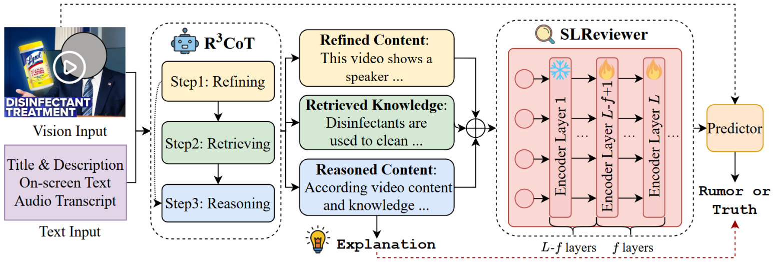


Figure 2: The structure of our proposed ExMRD framework. (1) The R3CoT process prompts MLLMs to refine the video content, retrieve domain knowledge, and reason to provide explanations. (2) The SLReviewer is to distill the explainable evidence from R3CoT to facilitate reliable rumor detection.

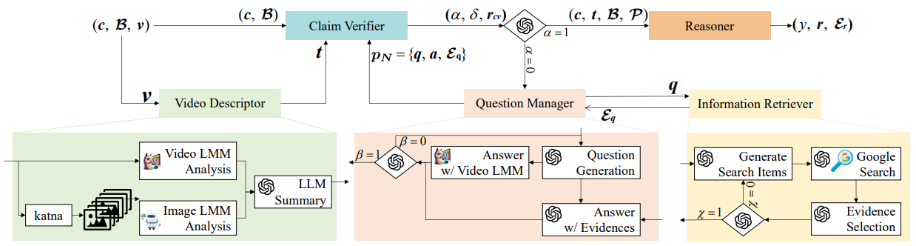
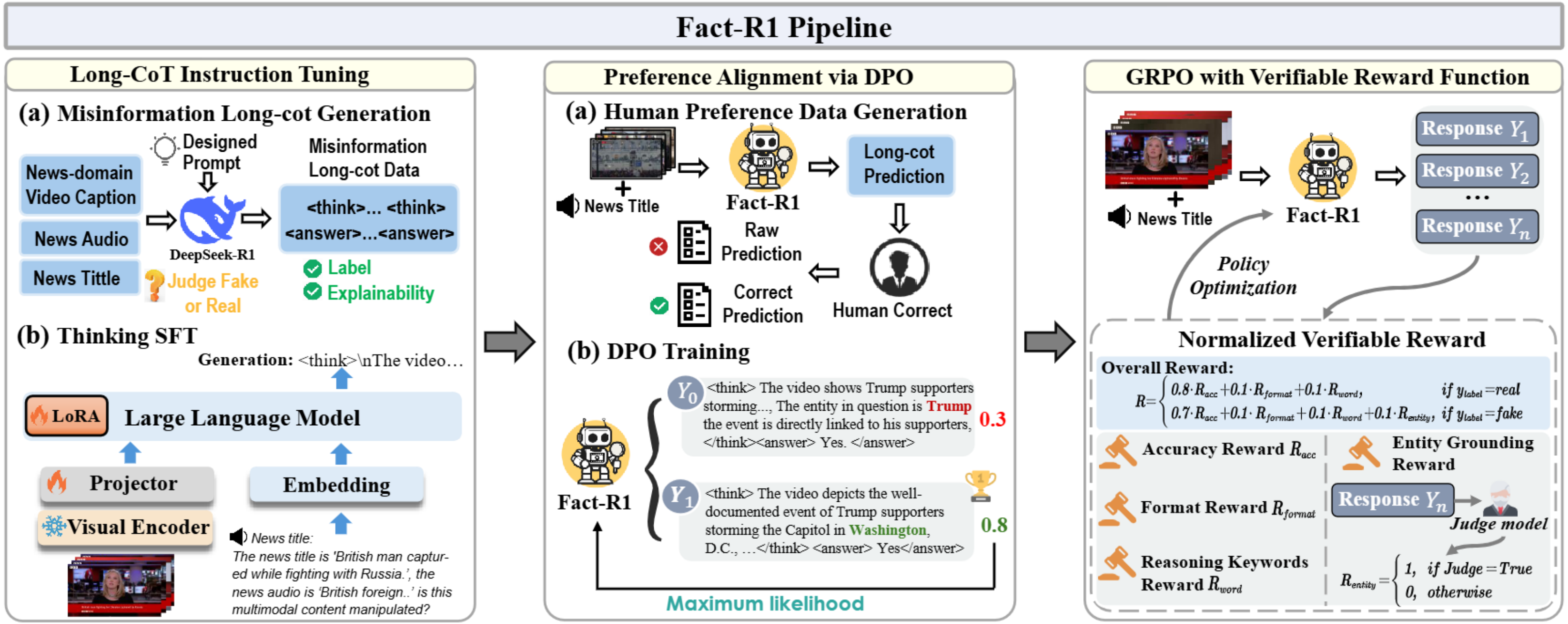


Figure 4: Overview of the proposed 3MFact framework, comprising five components: Video Descriptor (video-to-text conversion), Claim Verifier (assesses evidence sufficiency), Question Manager (generates questions and retrieves answers), Information Retriever (searches for evidence), and Reasoner (synthesizes judgment with rationale and evidence).

Hong, Rongpei, et al. Following Clues, Approaching the Truth: Explainable Micro-Video Rumor Detection via Chain-of-Thought Reasoning. WWW 2025
Niu, Kaipeng, et al. "Pioneering Explainable Video Fact-Checking with a New Dataset and Multi-role Multimodal Model Approach." AAAI 2025.

Seek clues from external information

Train domain-aligned LLMs for factual reasoning and verifiable judgment



From pipeline usage to intrinsic training, LLMs evolve from external assistants to internalized reasoners for video misinformation detection.

Zhang, Fanrui, et al. "Fact-R1: Towards Explainable Video Misinformation Detection with Deep Reasoning." NeurIPS 2025

Seek clues from external information

Generate and reason with debunking evidence via multi-agent LLM and diffusion collaboration

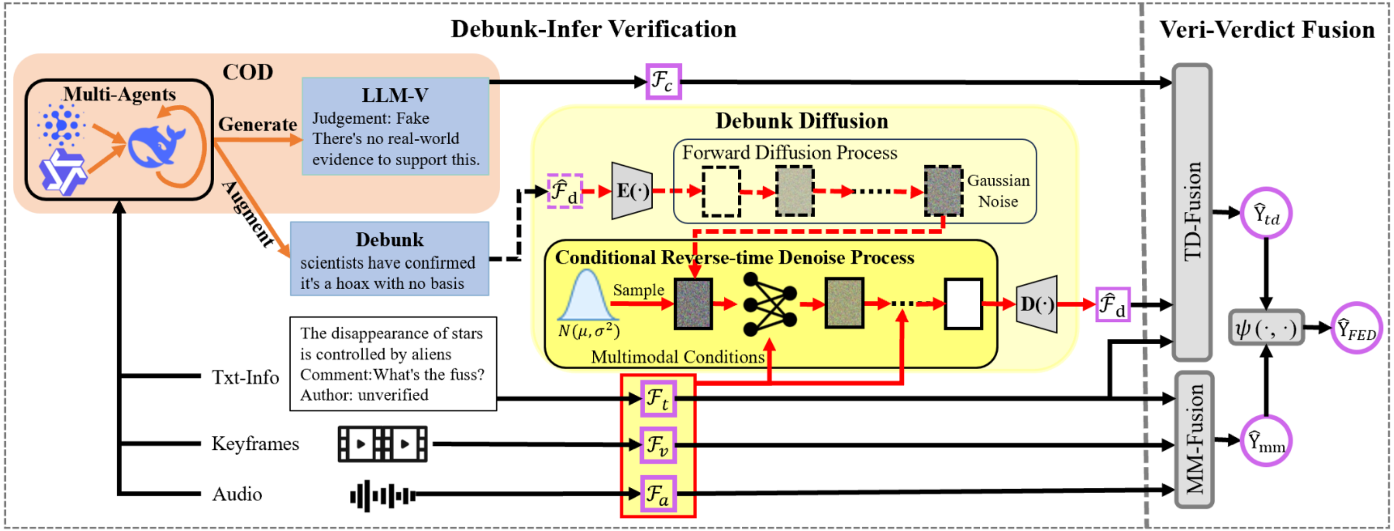


Fig. 1. Overview of DIFND Framework. The DIFND framework consists of Debunk-Infer Verification and Veri-Verdict Fusion. Debunk Diffusion generates debunking cues conditioned on multimodal inputs and is trained on the LLM-augmented dataset, while multi-agent LLMs perform chain-of-debunk for reasoning-rich verification named LLM-V. Final decisions are made via attentive fusion of features from all modules. The dashed arrows indicate paths that are used only during training and are not involved during inference. The textual information with blue background is generated or enhanced by LLMs.

💡 LLMs act as reasoning agents to generate debunking evidence and verify multimodal claims through diffusion-based inference.

Yan, Kaiying, et al. "Debunk and Infer: Multimodal Fake News Detection via Diffusion-Generated Evidence and LLM Reasoning." arXiv 2025

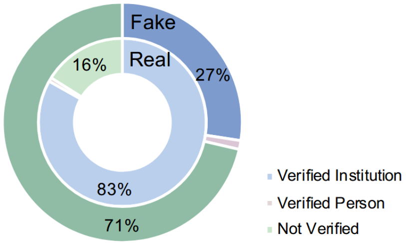
Tutorial Outline

Detection Part I: Human-Edited Misinformation

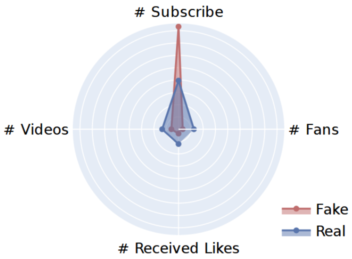
- Signal-based detection
  - Editing Traces
  - Generation Traces
- Semantic-based detection
  - Seek clues within the sample’s own content
  - Seek clues from external information
- Intent-based detection
  - Social context
- Clue integration for misinformation video detection
  - Parallel Integration
  - Sequential Integration
- Q+A/Discussion

Intent-based detection

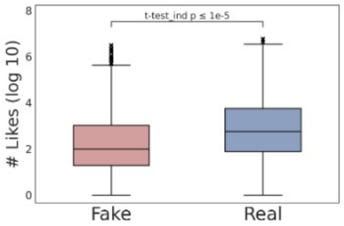
Misinformation reflects underlying user intents, which can be distinguished through social behaviors and engagement patterns.



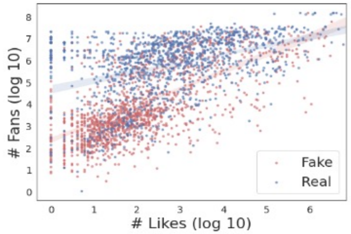
(a) Authority



(b) Statistics



(a) Number of likes



(b) Relationship between the number of publisher fans and likes.

Metadata	the number of comments the number of likes the video duration in seconds the number of videos that the publisher uploaded the follower-following ratio

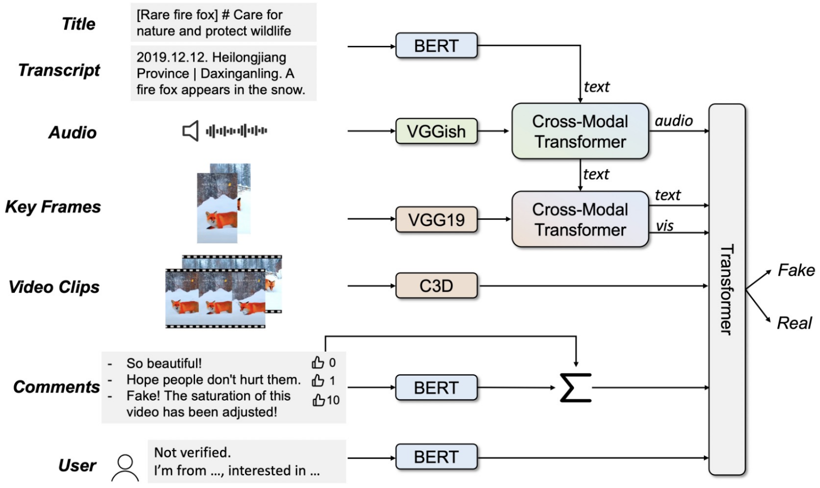
Comments	comments fakeness ratio comments inappropriateness ratio (swear words) comments conversation ratio (at least one reply) top 100 comments tf-idf top three popular comments: sentiment polarity, the number of modal particles, the number of personal pronouns and text length

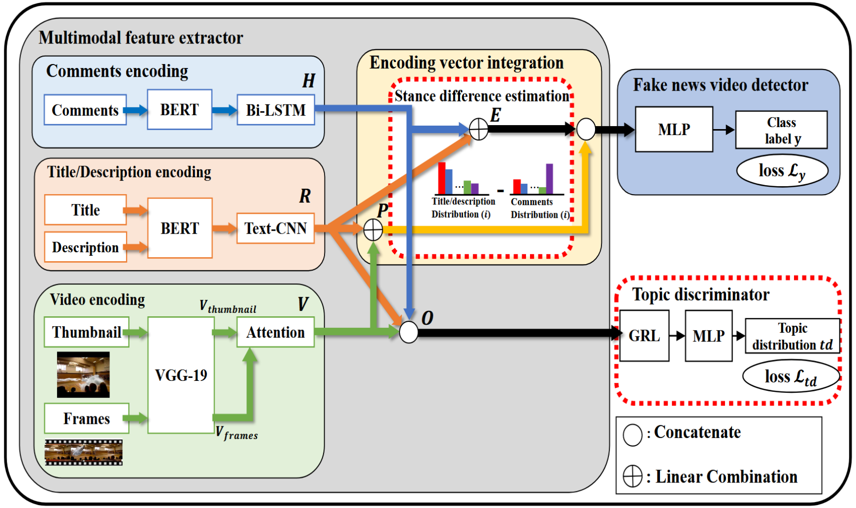
👆 Statistical and social engagement features reveal discriminative behavioral patterns.

Qi, Peng, et al. "Fakesv: A multimodal benchmark with rich social context for fake news detection on short video platforms." AAAI 2023.

Intent-based detection

Model social context features to capture intent-related cues from users and comments.





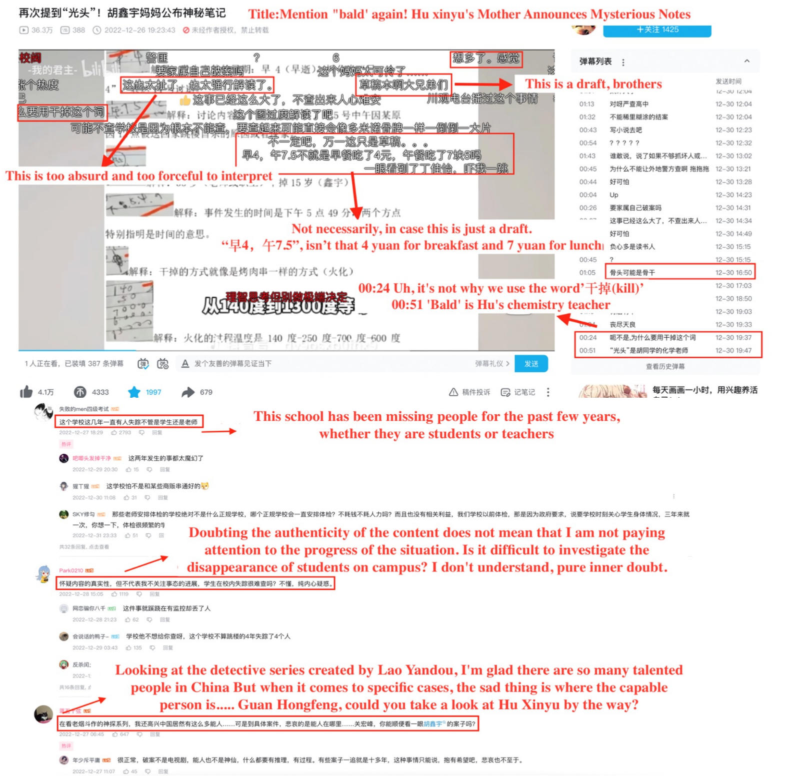
👆Incorporates **user attributes** and multimodal **comment features** via cross-modal transformers.

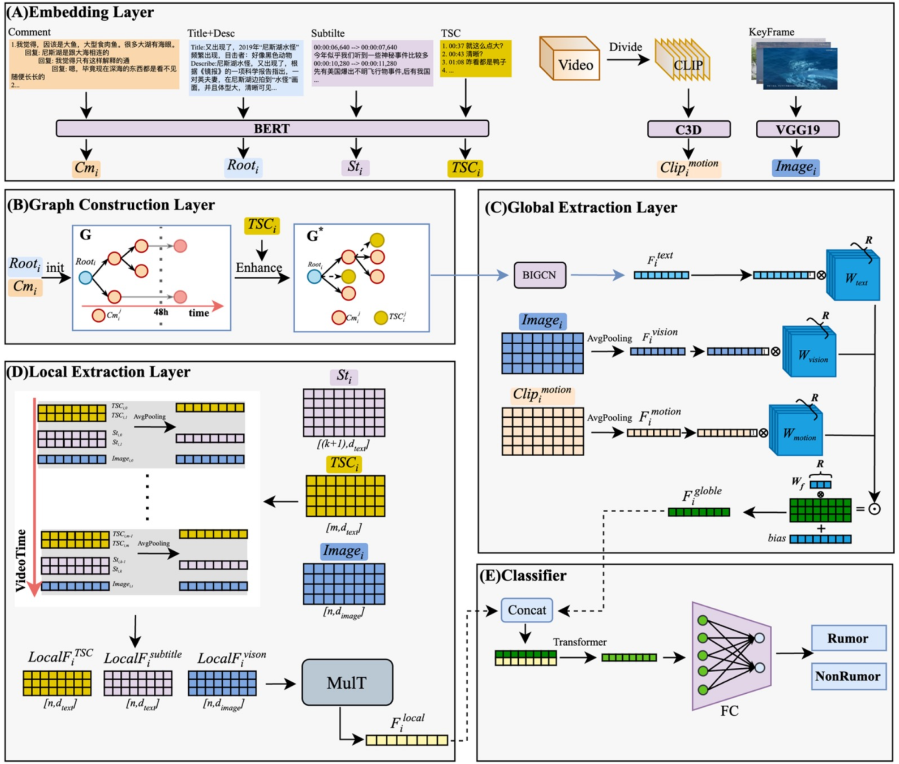
👆Models **stance and topic** context between **comments** and video text.

Qi, Peng, et al. "Fakesv: A multimodal benchmark with rich social context for fake news detection on short video platforms." AAAI 2023.
Choi, Hyewon, and Youngjoong Ko. "Using topic modeling and adversarial neural networks for fake news video detection." CIKM 2021.

Intent-based detection

Incorporate **time-synchronized comments** to capture dynamic social context





Tutorial Outline

ACM multimedia
Dublin, Ireland 27-31.10.2025

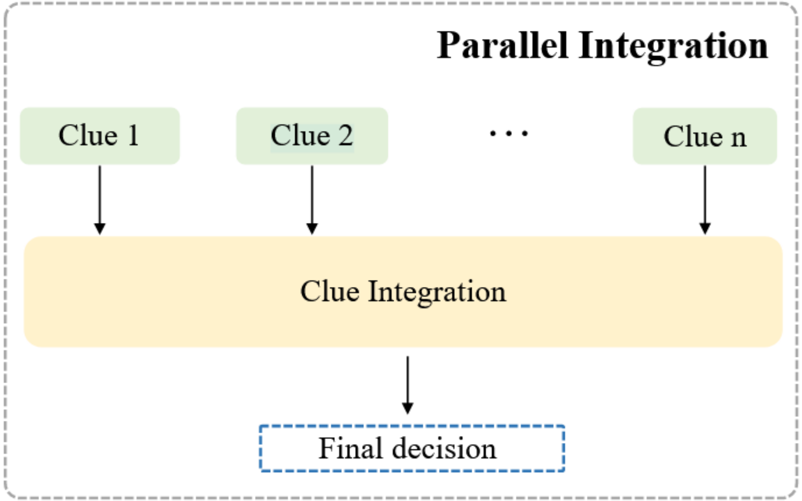
Detection Part I: Human-Edited Misinformation

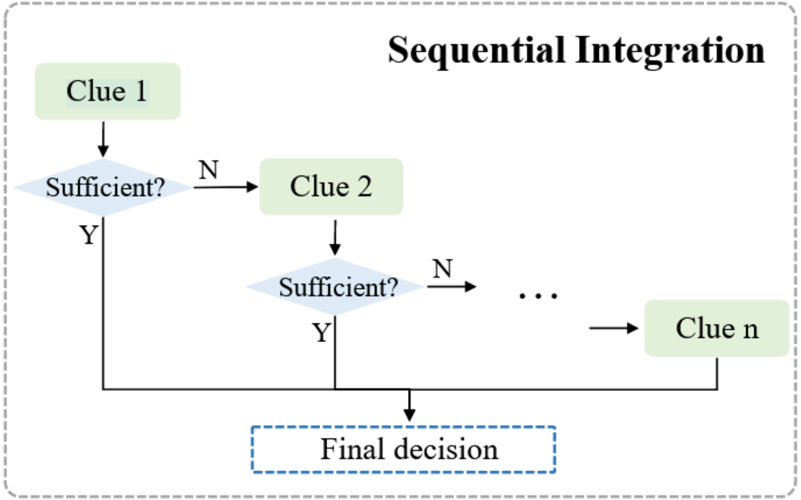
- Signal-based detection
  - Editing Traces
  - Generation Traces
- Semantic-based detection
  - Seek clues within the sample’s own content
  - Seek clues from external information
- Intent-based detection
  - Social context
- Clue integration for misinformation video detection
  - Parallel Integration
  - Sequential Integration
- Q+A/Discussion

Clue integration for misinformation video detection

Two major paradigms for combining multiple features from different modalities

- **Parallel integration**: all clues from different modalities contribute to the final decision-making process
- **Sequential integration**: clues from different modalities are combined in a step-wise manner with each modality contributing incrementally to the final decision.

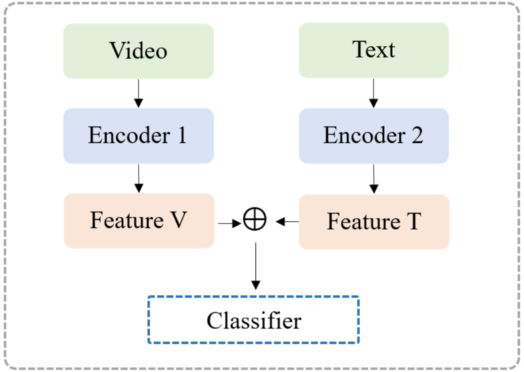


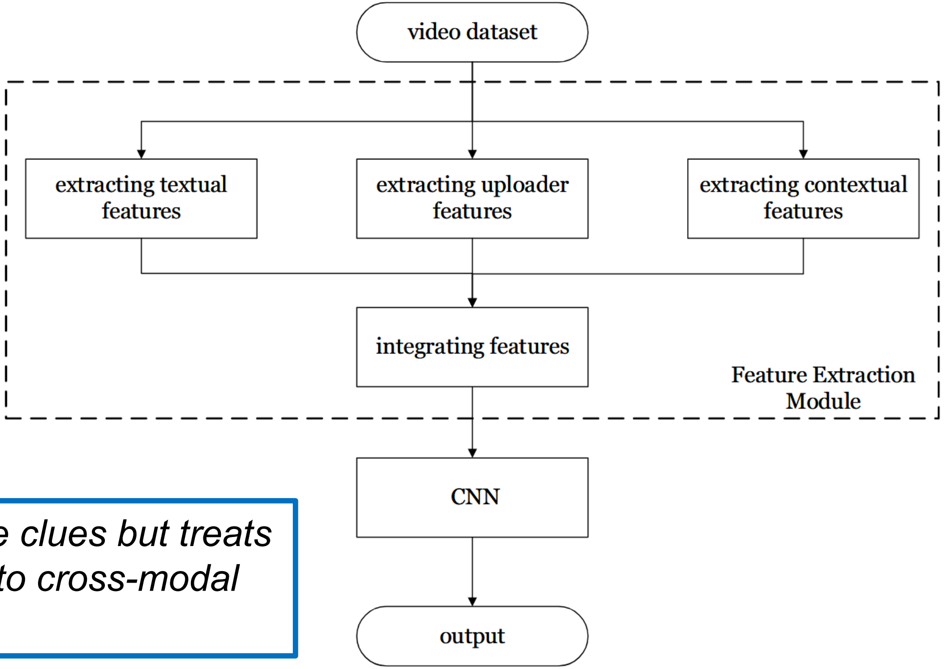


Bu, Yuyan, et al. "Combating online misinformation videos: Characterization, detection, and future directions." ACM MM 2023.

Concatenation-Based

Feature fusion technique: Direct concatenation of multi-modal representations.



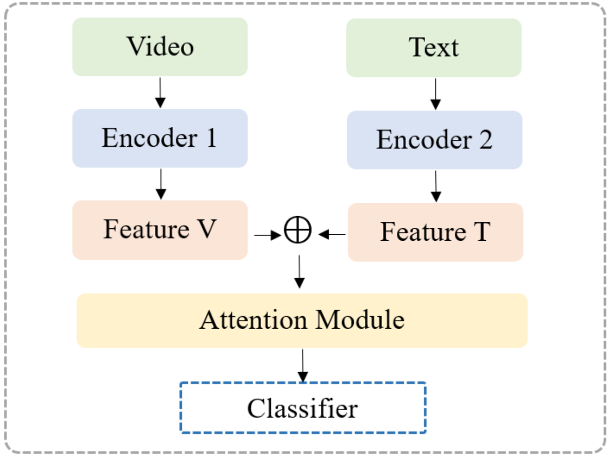


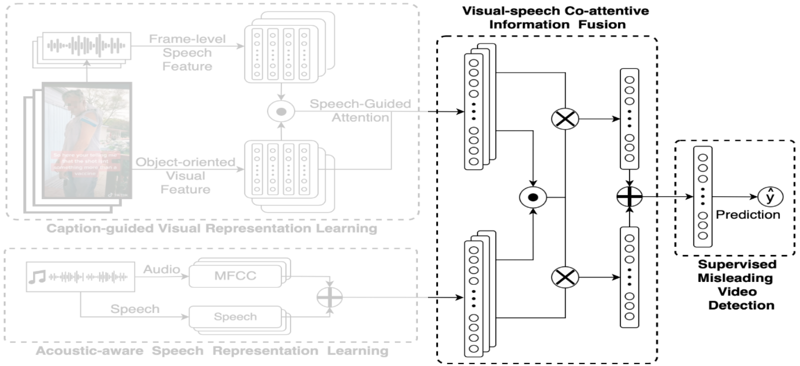
💡 *Concatenation preserves all available clues but treats them equally — efficient yet insensitive to cross-modal dependency.*

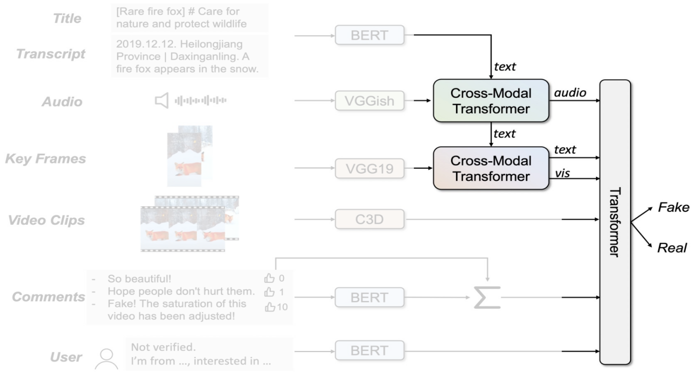
Li, Xiaojun, et al. "A CNN-based misleading video detection model." Scientific Reports 2022.

Attention-Based

Feature fusion technique: Focus on informative clues via attention.





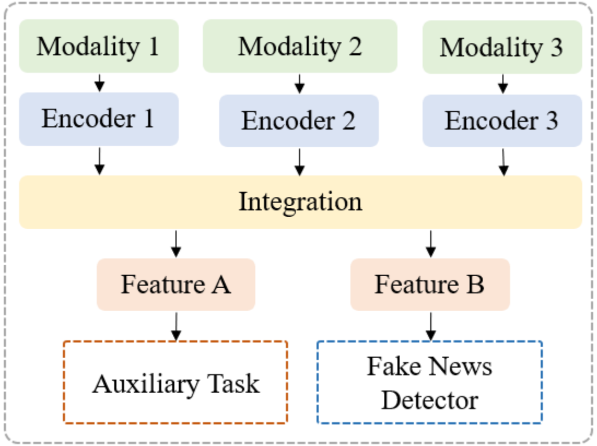


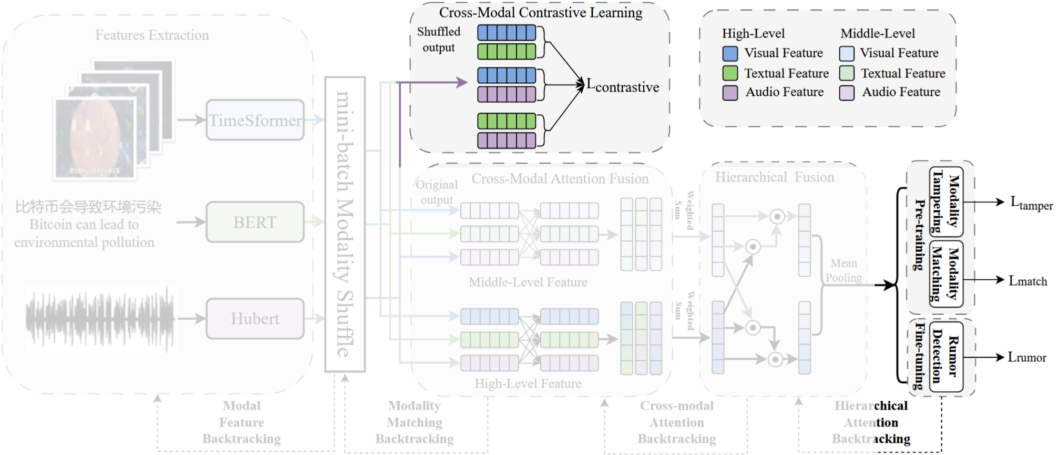
💡 *Attention-based fusion highlights informative modalities and captures inter-modal interactions dynamically.*

Shang, Lanyu, et al. "A multimodal misinformation detector for covid-19 short videos on tiktok." IEEE Big Data, 2021.
Qi, Peng, et al. "Fakesv: A multimodal benchmark with rich social context for fake news detection on short video platforms." AAAI 2023.

Multitask-Based

Fusion through auxiliary tasks to enhance generalization and consistency.





Multitask fusion aligns representations across modalities by leveraging auxiliary supervision — improving robustness under distribution shifts.

Figure 1: Model Architecture Overview of SVRPM. The model consists of five main modules: (a) Feature extraction: Extract visual, textual, and audio modal features using different encoders, respectively. (b) Mini-batch Modality Shuffle: Randomly shuffle their corresponding modal features for a minibatch. (c) Cross-modal Contrastive Learning: Constructing Positive and Negative Samples using Modality Shuffle Module. (d) Cross-modal Fusion and Hierarchical Fusion. (e) Modality Tampering Backtracking: Using attention backtracking operation to obtain the local features which may be tampered.

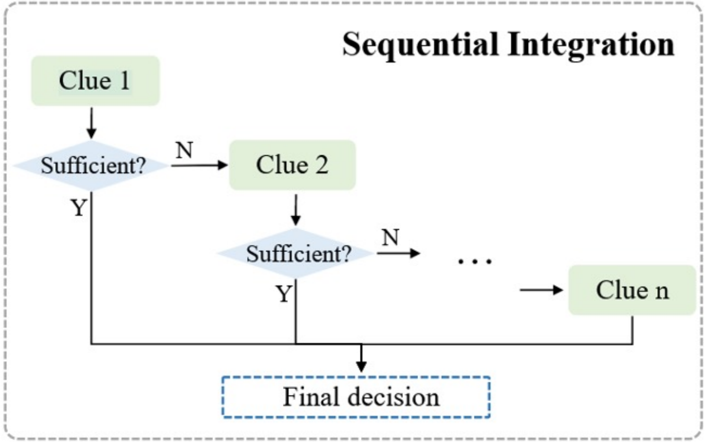
Wu, Kaixuan, et al. "Interpretable short video rumor detection based on modality tampering." LREC-COLING 2024

Pipeline-Based

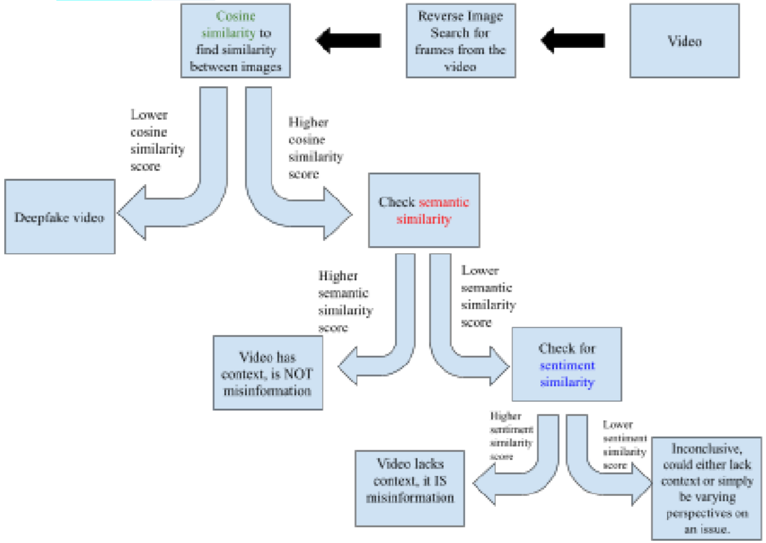
Sequential integration: Step-wise reasoning over heterogeneous clues.

💡 *Pipeline-style sequential integration mimics human reasoning — verifying clues in stages.*

(Enhance interpretability and efficiency under missing or redundant modalities, but may accumulate early-stage errors).



		Semantic Similarity	
		High	Low
Sentiment Similarity	High	Not Misinformation	Misinformation
	Low	Not Misinformation (but differing opinion)	Inconclusive (could be Misinformation or just differing perspective)



Ganti, Dhanvi. "A novel method for detecting misinformation in videos, utilizing reverse image search, semantic analysis, and sentiment comparison of metadata." 2022