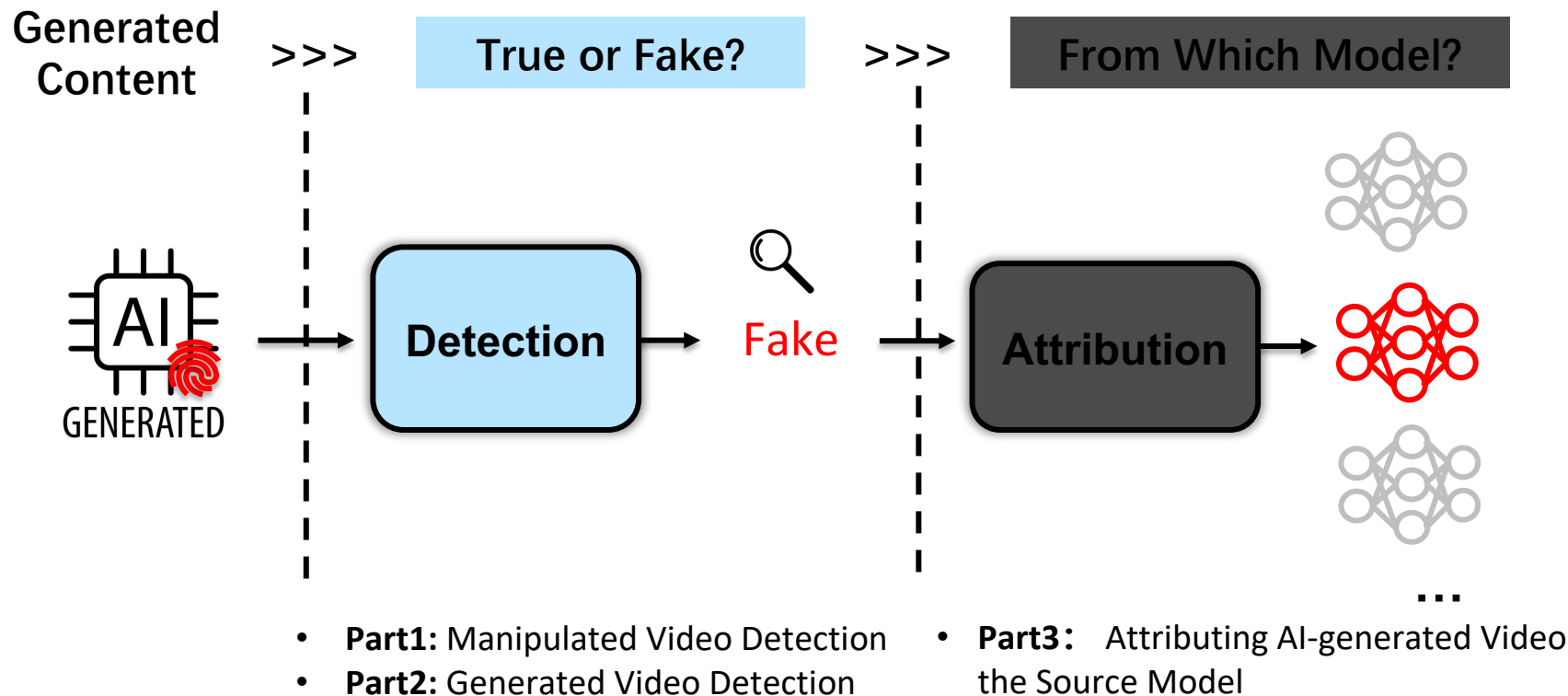


Detection Part II: AI-Generated Misinformation

Section 5



Detection Part II: AI-Generated Misinformation

Manipulated video detection

Generated video detection

Attributing AI-generated Content to the Source Model

Q+A/Discussion

- Deepfakes are one of the most prominent and concerning forms of video manipulation.

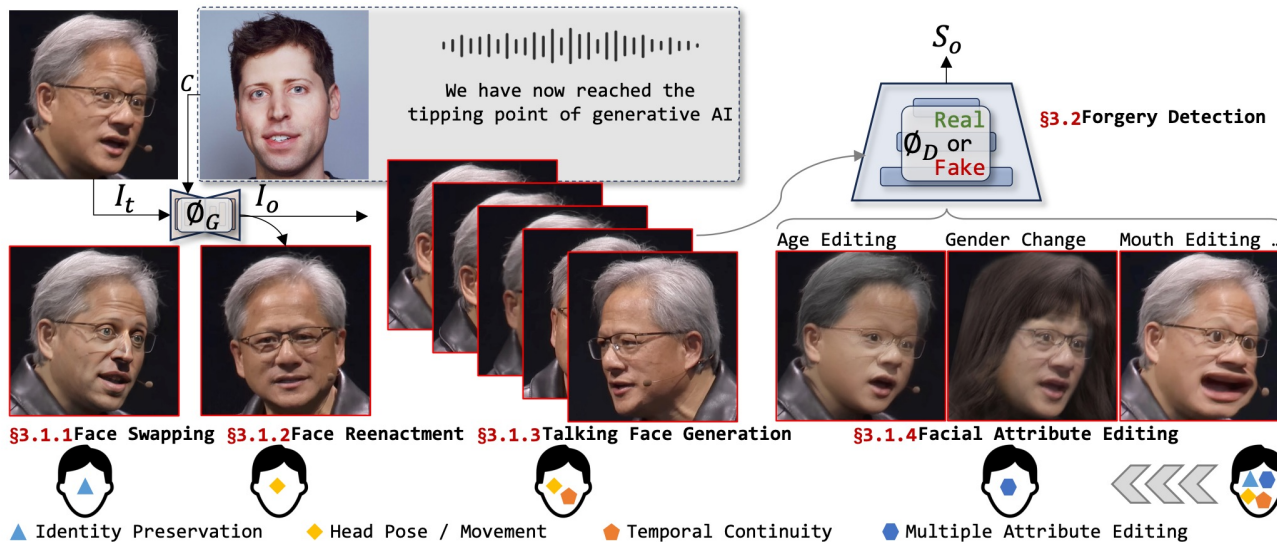
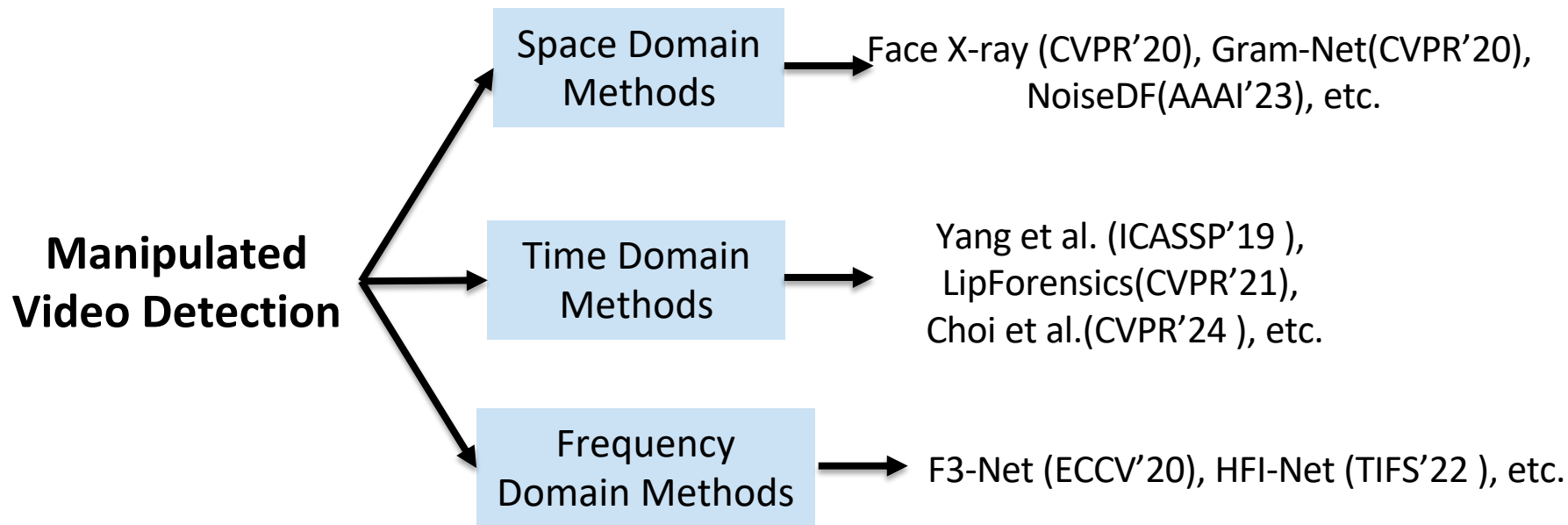
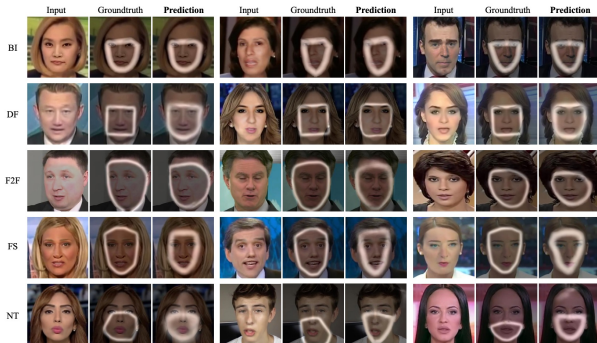


Fig. 2: Top: Illustration of different deepfake generation (Sec. 3.1) and detection tasks (Sec. 3.2) that are discussed in this survey. Bottom: Specific facial attribute modification of each task. Data from [NVIDIA Keynote at COMPUTEX 2023](#) at 29:40.



Space Domain Inconsistency

Boundary Artifacts (Face X-ray)



Texture Inconsistencies (Gram-Net)

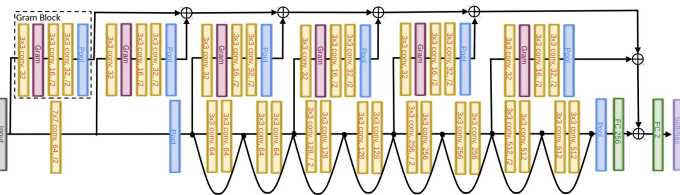
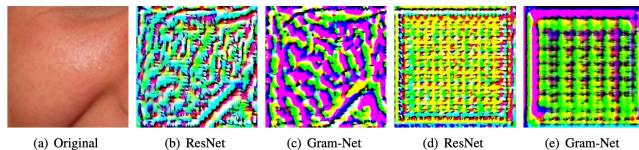
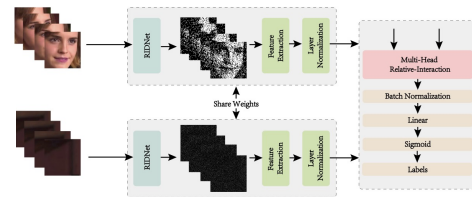
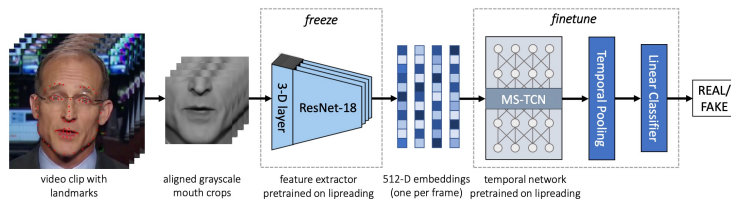


Figure 5. Gram-Net architecture. We extract global image texture feature with 6 Gram Blocks in different semantic levels from ResNet. $\oplus$ means concatenation.

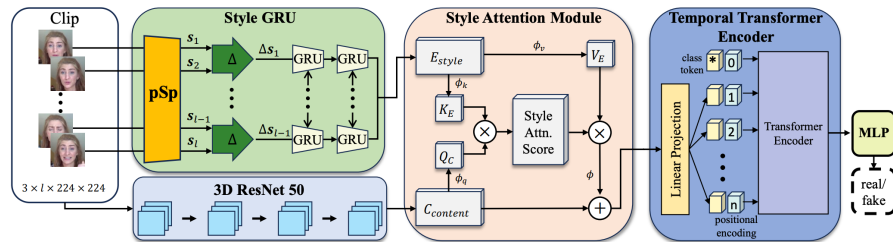
Noise Inconsistencies (NoiseDF)



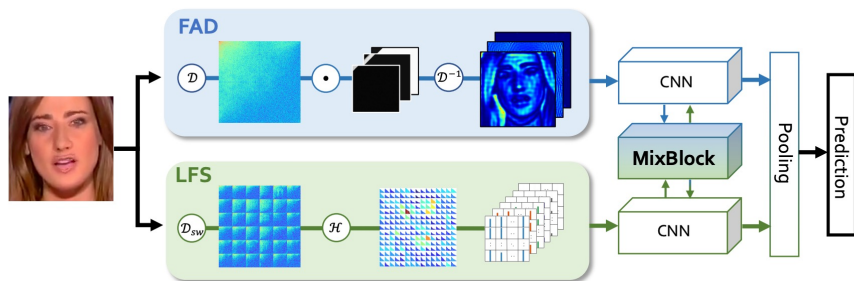
Inconsistency of mouth movements (LipForensics)



Inconsistency of the style latent vectors (Choi et al.)

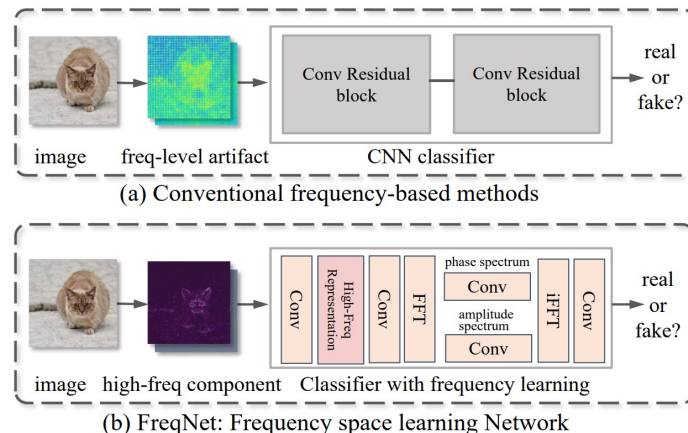


F3-Net



A frequency perception framework with two-branches:
Frequency-aware Image Decomposition (FAD) and Local Frequency Statistics (LFS)

FreqNet



Emphasis on high-frequency details

Detection Part II: AI-Generated Misinformation

Manipulated video detection

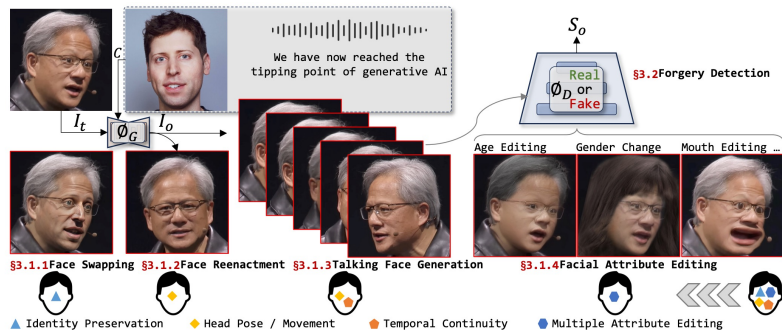
Generated video detection

Attributing AI-generated Content to the Source Model

Q+A/Discussion

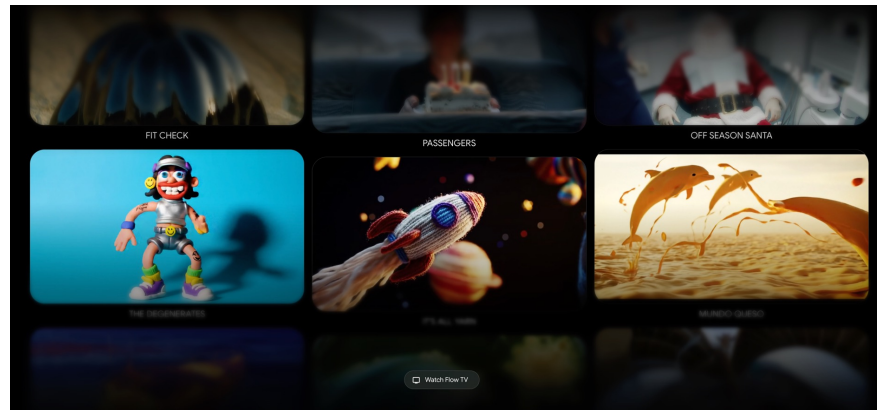
From Manipulated to Generated Video Detection

Manipulated Video Detection



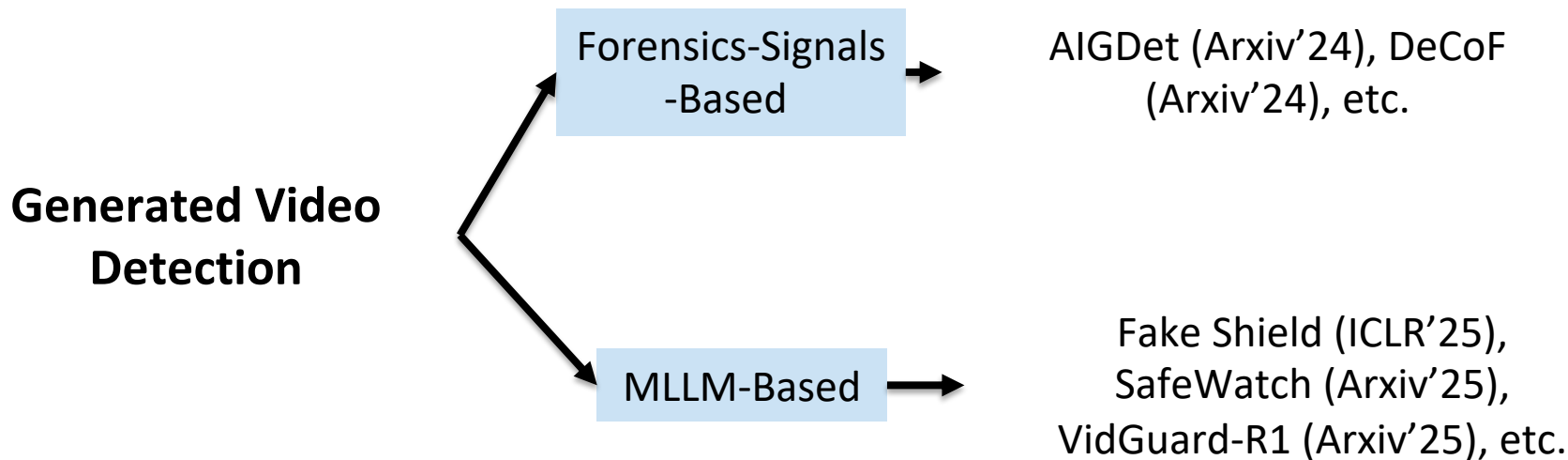
Face-centric content

Generated Video Detection

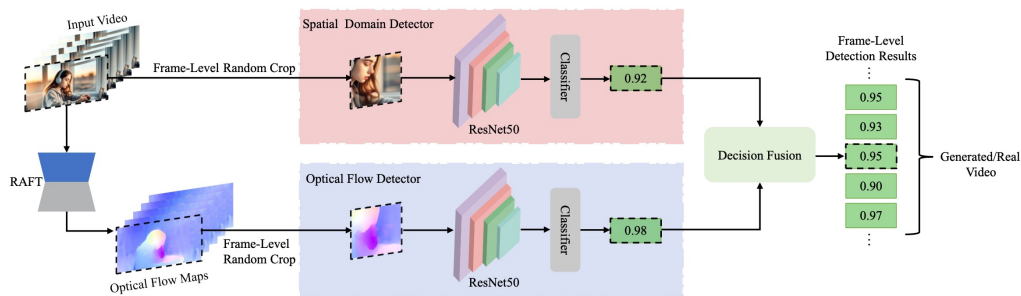
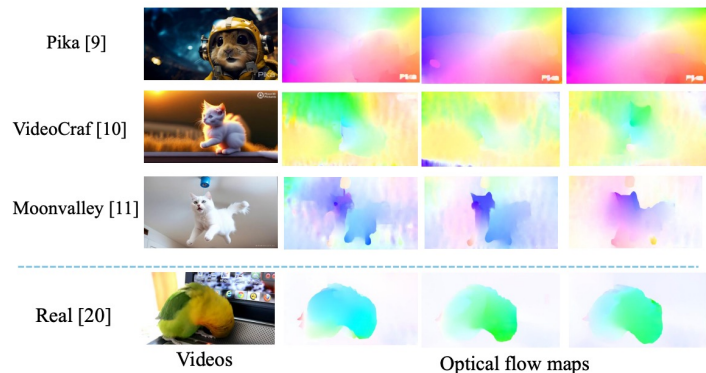


(Generated by Veo3)

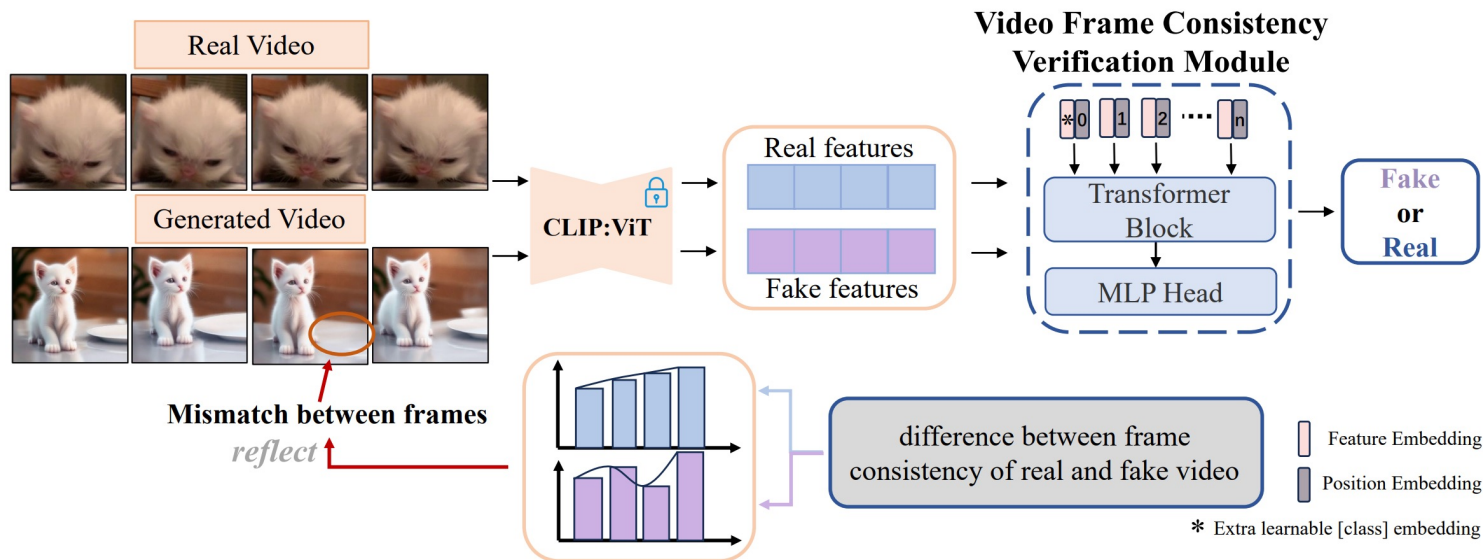
More diverse, real-world videos



- The optical flow maps of generated videos exhibit less smoothness and blurry contours compared to real video

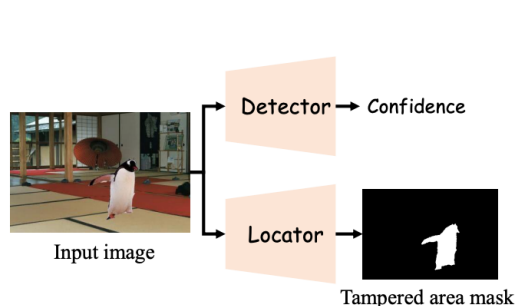


- Detecting generated videos by capturing mismatch between frames



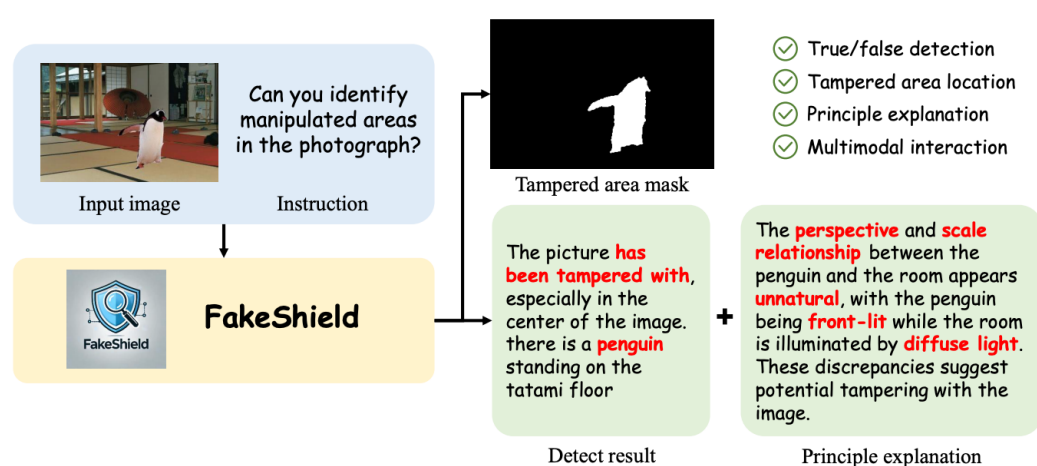
Fake Shield: Explainable Detection via MLLMs

- Utilizing MLLMs enables detailed explanations and conversational interactions for a deeper analysis



- ✓ True/false detection
- ✓ Tampered area location
- ✗ Principle explanation
- ✗ Multimodal interaction

(a) Conventional IFDL Framework

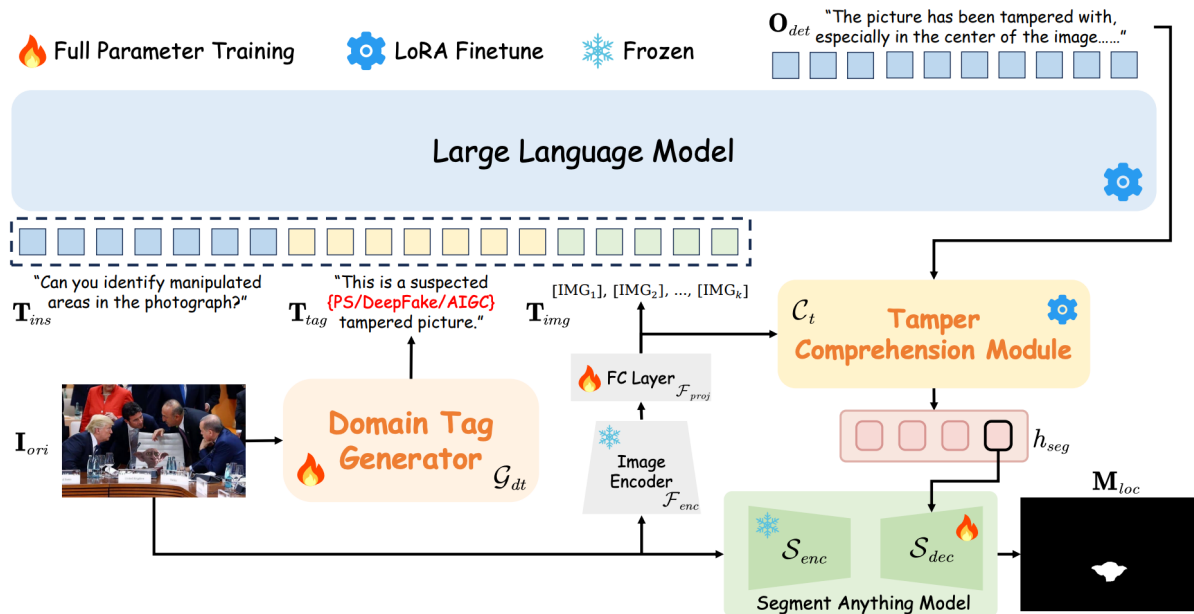


- ✓ True/false detection
- ✓ Tampered area location
- ✓ Principle explanation
- ✓ Multimodal interaction

(b) Our Explainable IFDL Framework: FakeShield

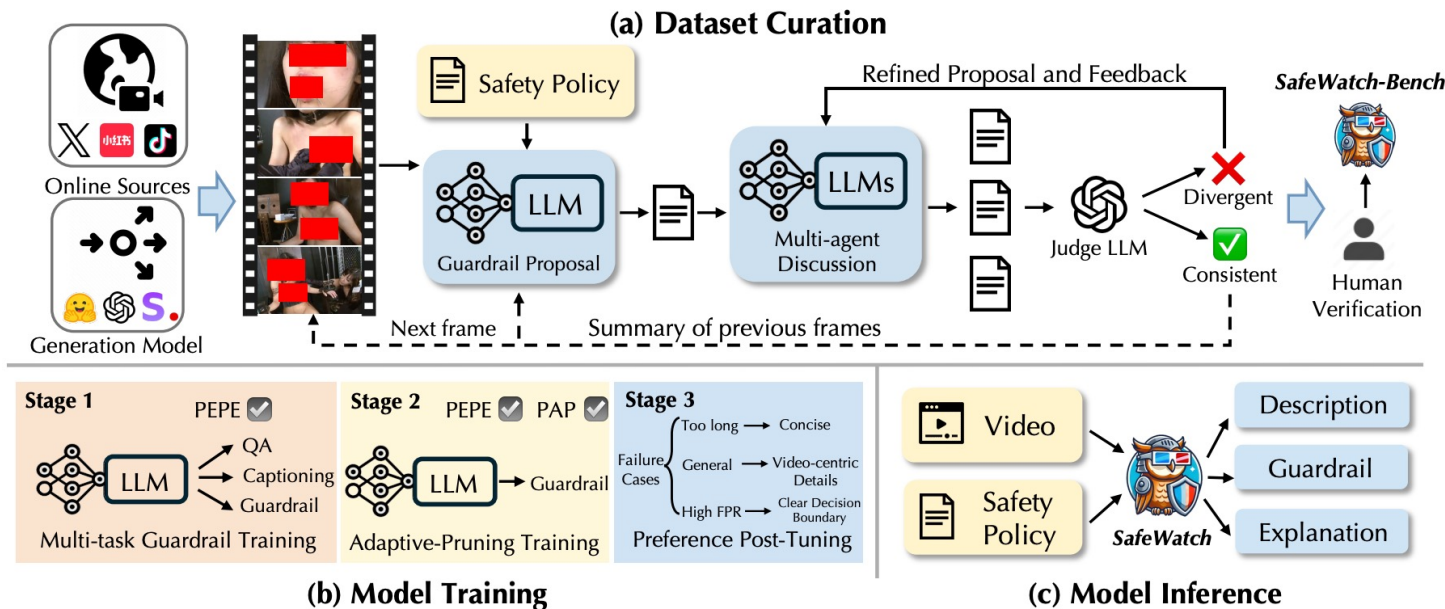
Fake Shield: Explainable Detection via MLLMs

- Domain Tag Generator is trained to spot different types of fake images.
- Tamper Comprehension Module is adopted to align visual-language features, thus pinpointing tampered areas



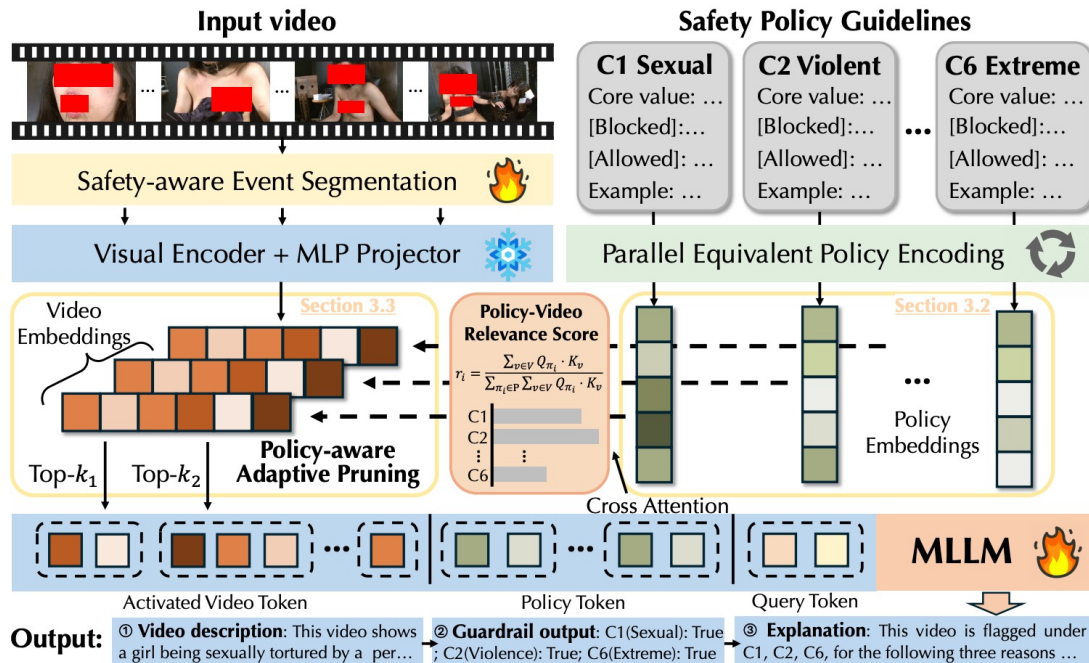
SafeWatch: Improving Safety-Policy Following

- Apply supervised fine-tuning (SFT) and direct preference optimization (DPO) to train a video guardrail model with transparent reasoning.

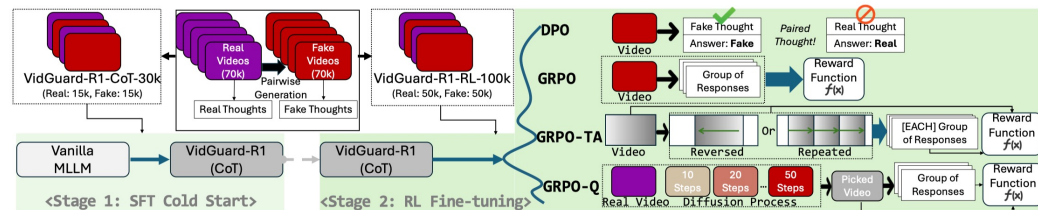
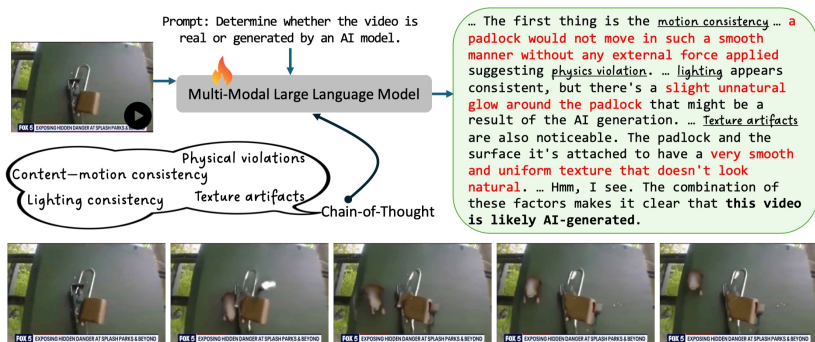


SafeWatch: Improving Safety-Policy Following

- Decoding pipeline of SafeWatch



- Applies online reinforcement learning for more generalized fake detection



Detection Part II: AI-Generated Misinformation

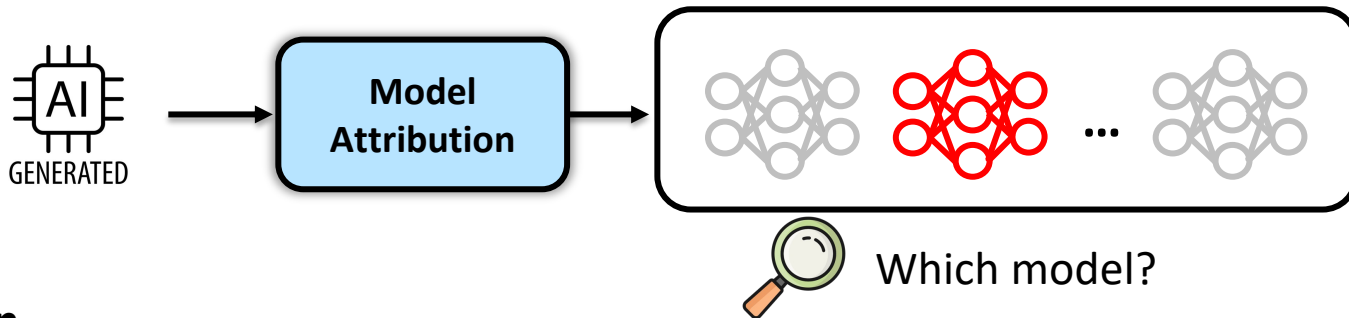
Manipulated video detection

Generated video detection

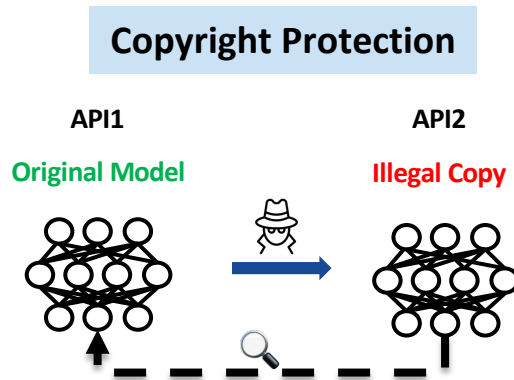
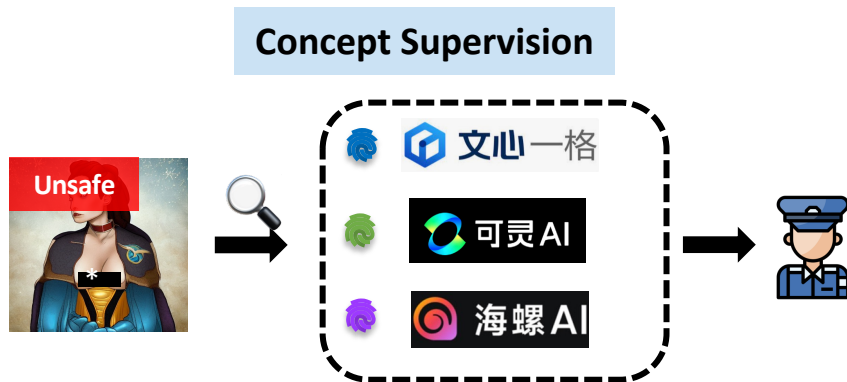
Attributing AI-generated Content to the Source Model

Q+A/Discussion

- Identify the source model of generated contents



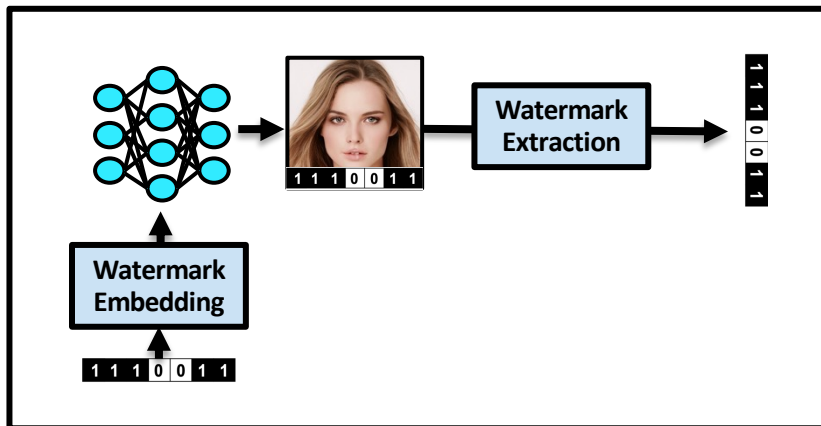
- Application



Two Technique Routes for Model Attribution

Watermarking-Based

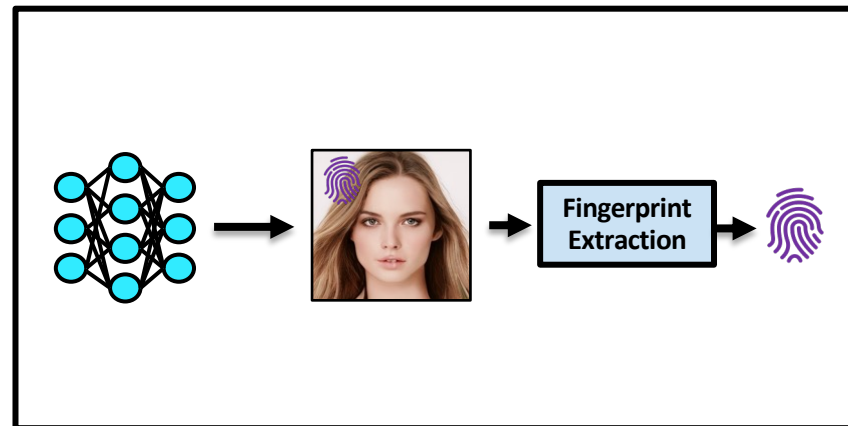
(Artificial Fingerprint)



Need modify model in advance

Model-Fingerprint-Based

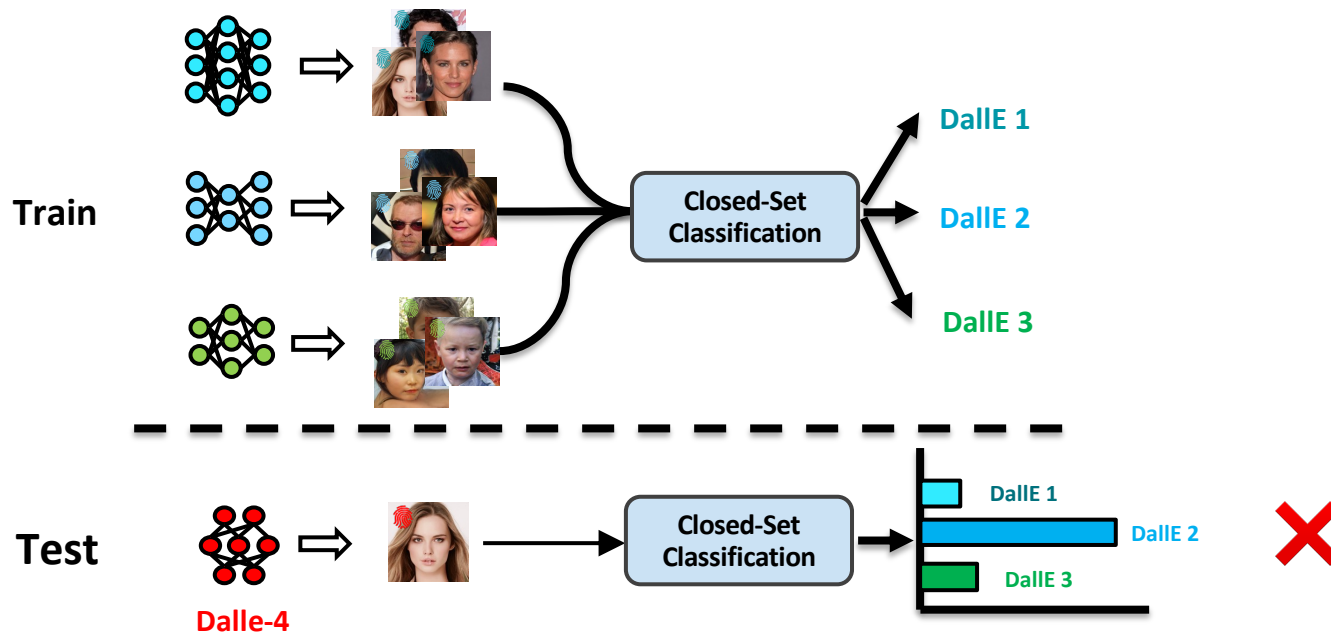
(Intrinsic Fingerprint)



No need to modify model

The focus of this part

Existing works: Closed-Set Classification on Limited Model Data



Not scalable to newly emerged models

How to Build a Scalable Model Attribution System?

	Framework	Data
Existing Works ^[1,2,3]	Closet-Set	Limited (100+models)
Recent Work	Open-Set	Large Scale (5k+ models)

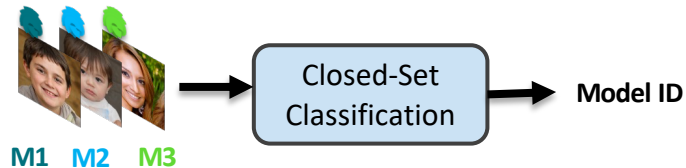
[1] Yu N, Davis L S, Fritz M. Attributing fake images to gans: Learning and analyzing gan fingerprints. ICCV 2019

[2] Frank J, Eisenhofer T, Schönherr L, et al. Leveraging frequency analysis for deep fake image recognition. ICML 2020

[3] Marra F, Gragnaniello D, Verdoliva L, et al. Do gans leave artificial fingerprints?

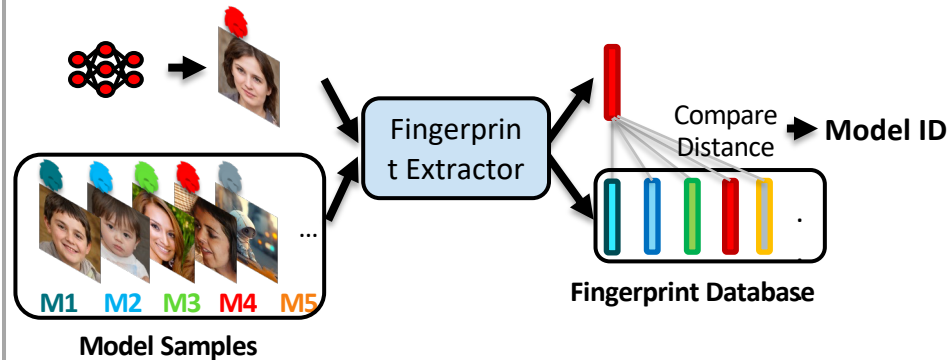
Closed-Set -> Open-Set Framework

Closed-Set Framework



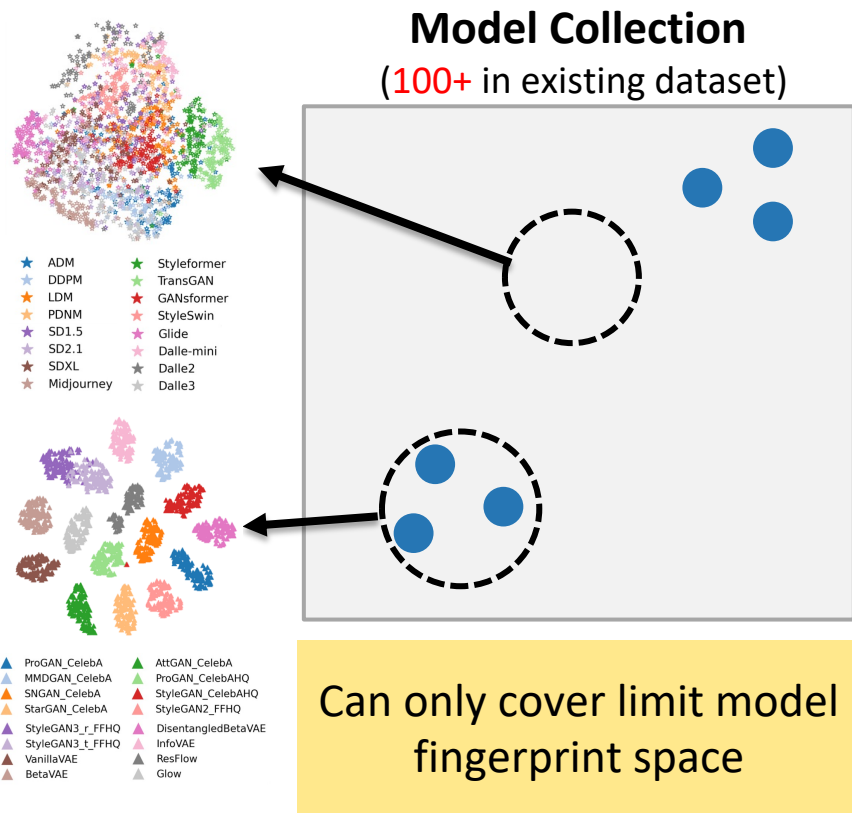
Can only attribute seen models

Open-Set Framework

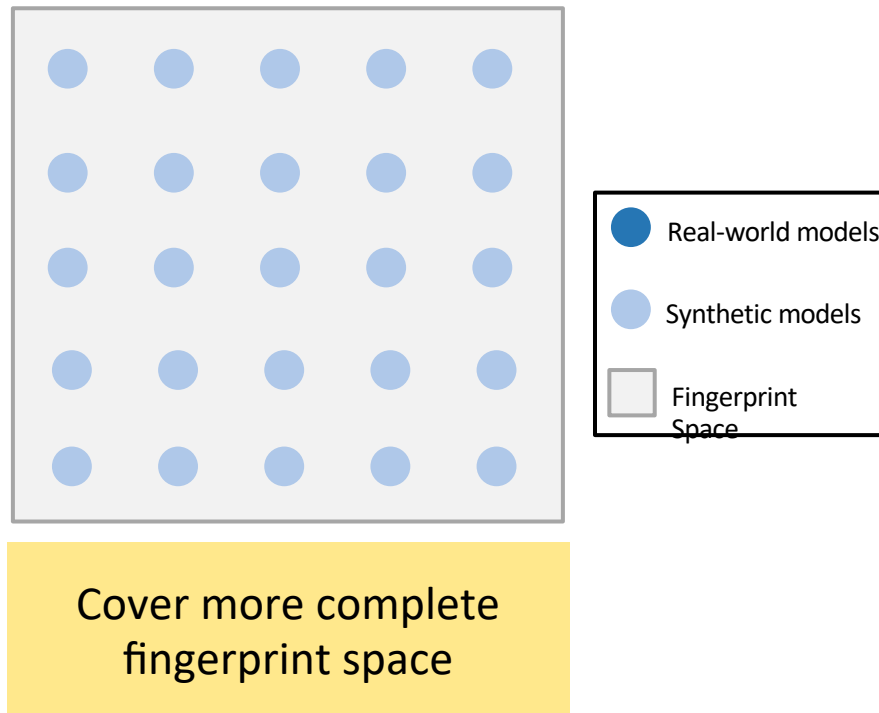


Can dynamically extend to unseen models not involved in training

Collect Model Data -> Synthesize Model Data



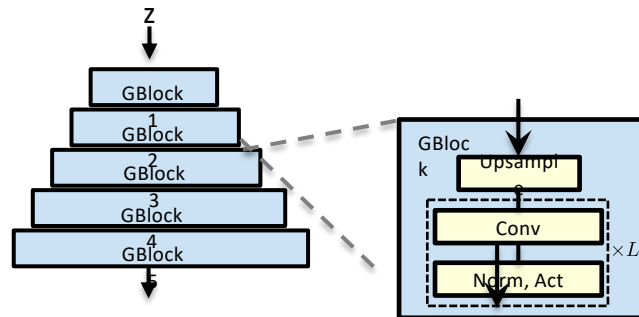
Our Method: Model Synthesis (Unlimited in principle)



Which Factors Influence Model Fingerprint?

Observation 1:

The **type of basic network components** influences the spectral patterns of generated images.

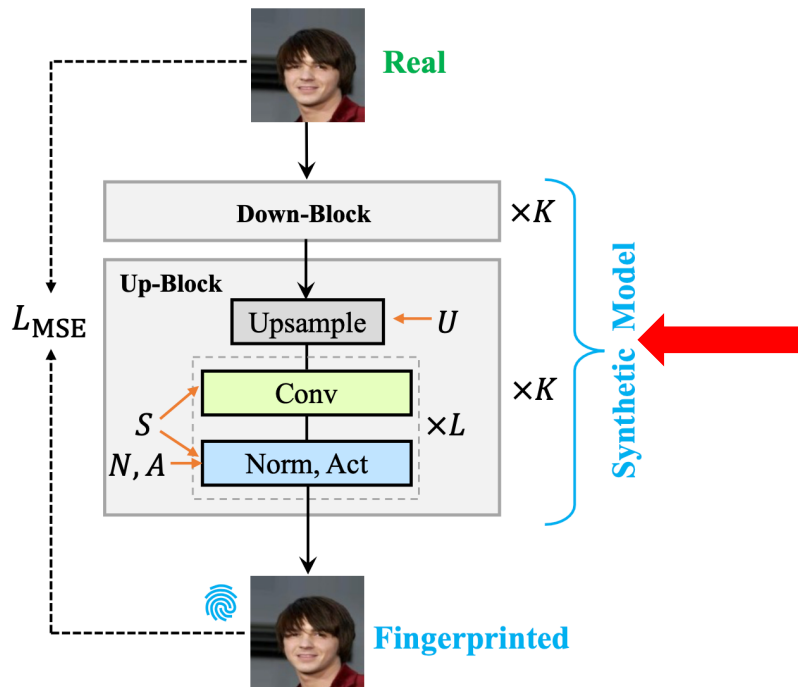


Observation 2:

The **last two generative blocks** are more influential

- Diverse basic component option is necessary
- Few blocks is sufficient

Synthetic Model Construction



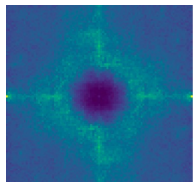
- L (number of convolution layers): 1, 2
- S (the order of activation and normalization relative to the convolution layer): before/after the convolution
- U (upsampling layer): devonvolution, bilinear, nearest
- A (activation function): None, ReLU, Tanh, Sigmoid
- N (normalization layer): None, BatchNorm, InstanceNorm
- K (number of blocks) : 1, 2

A total of 5,760 synthetic models

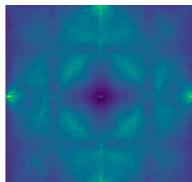
Synthesizing one model only takes **1.5** minutes

Generation Spectrum Comparison

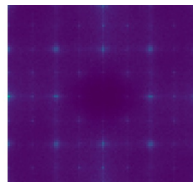
SNGAN



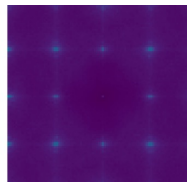
Synthetic Model



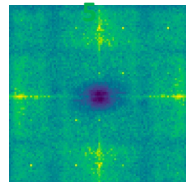
StarGAN



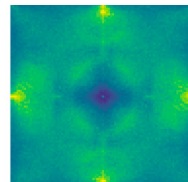
Synthetic Model



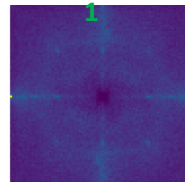
StableDifusion1.



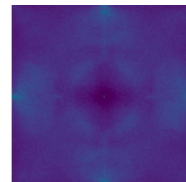
Synthetic Model



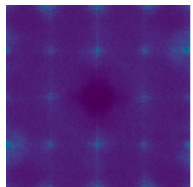
StableDifusion2.



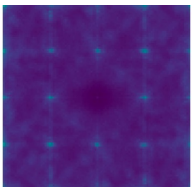
Synthetic Model



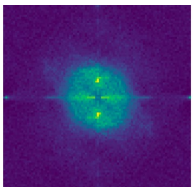
AttGAN



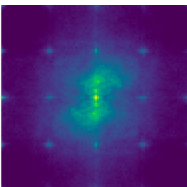
Synthetic Model



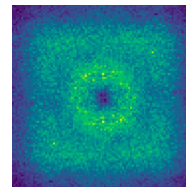
ProGAN



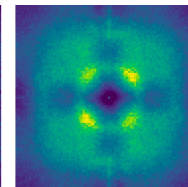
Synthetic Model



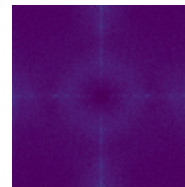
Midjourney



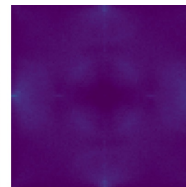
Synthetic Model



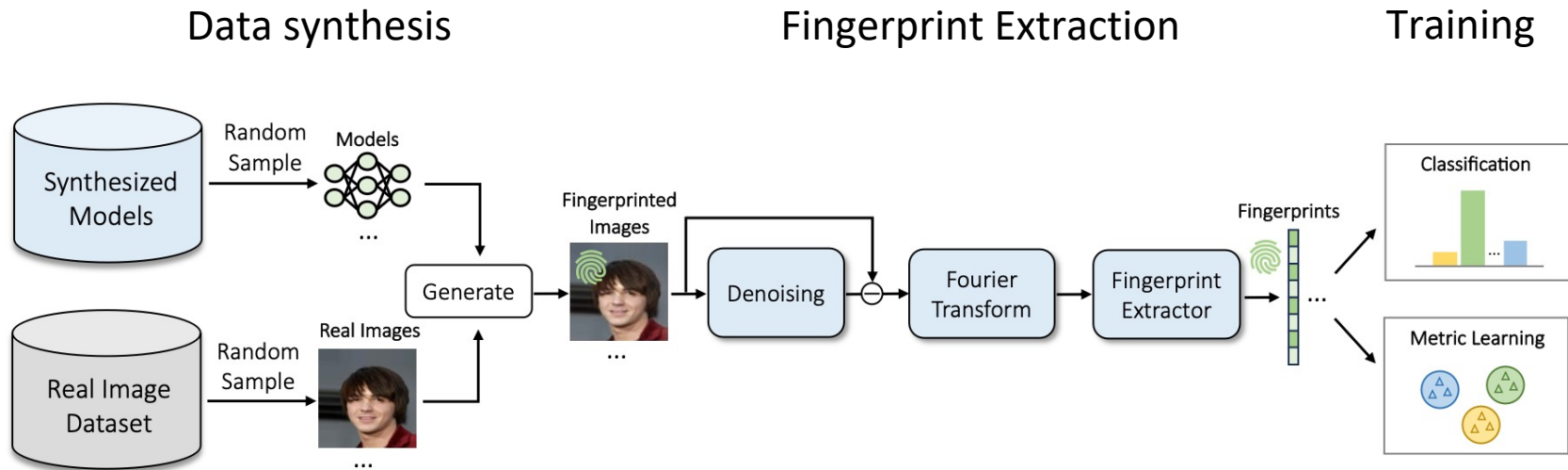
DallE-3



Synthetic Model

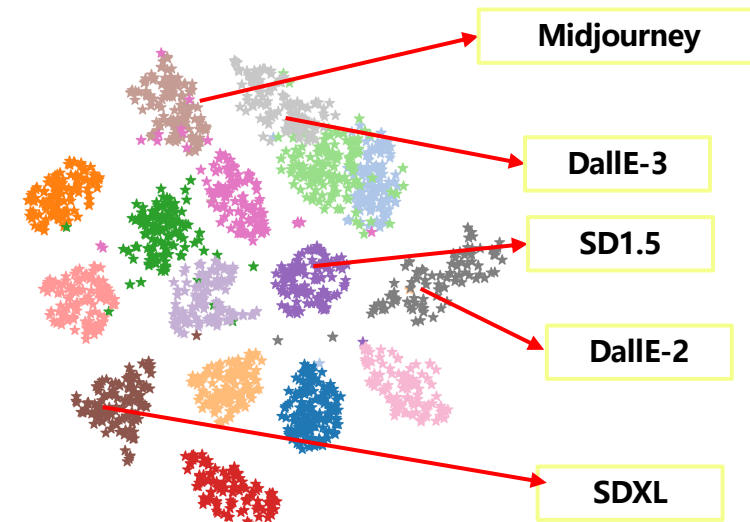


Overview of Open-World Framework



$$\mathcal{L} = \mathcal{L}_{\text{CE}}(h_k, y_k) + \mathcal{L}_{\text{Triplet}}(h_a, h_p, h_n)$$

Evaluation Result



**Model attribution accuracy:
93.20%**

DallE-3



An armchair in the shape of
an avocado

DallE-3



"A sea otter with a pearl earring" by
Johannes Vermeer

Fingerprint Similarity: **0.9045**
Prediction Result: **Same Model**

DallE-3



An armchair in the shape of
an avocado

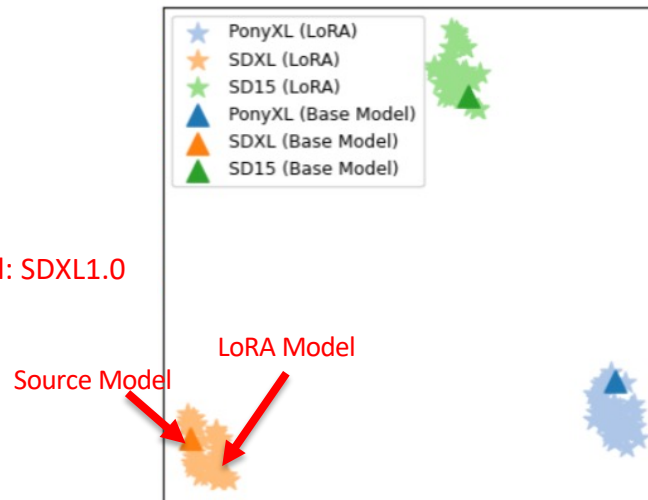
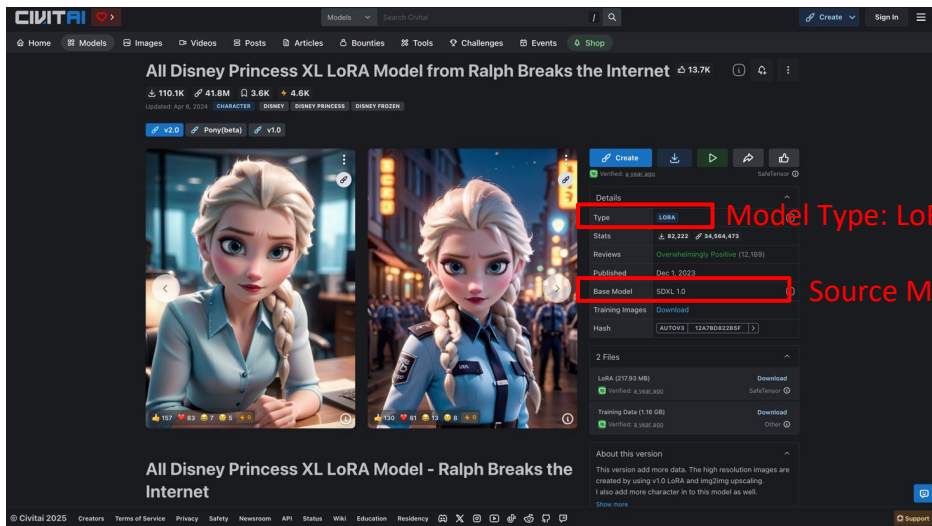
DallE-2



An armchair in the shape of
an avocado

Fingerprint Similarity: **0.7192**
Prediction Result: **Different Model**

Application: Model Relationship Analysis



The fingerprint extractor can be used for tracing the **source model of a LoRA model**

- **Part 1: Manipulated video detection**
 - Forensics signals such as space, time and frequency consistencies can be utilized
- **Part 2: Generated video detection**
 - Forensics signals can be utilized
 - MLLMs enables more transparent and explainable detection
- **Part 3: Attributing generated contents to the source model**
 - Closed-Set Classification is not scalable
 - Fingerprint extractor trained on large scale synthetic models can generalize to real-world models

Prevention Strategies

Section 6

6. Prevention Strategies

Creation Prevention

Embedding Tamper-proof Digital Identifier

- Developing Technical Specification
- Embedding Invisible Watermarks

Mitigating Hallucination in Content Generation

Spread Prevention

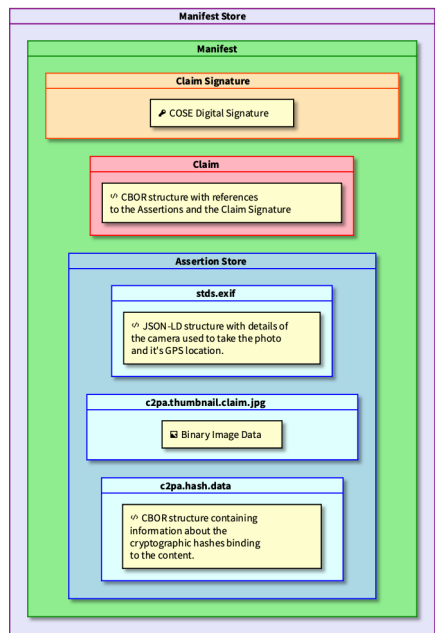
Alerting, Verification, and Resilience Building

Controlling the Spread of Misinformation

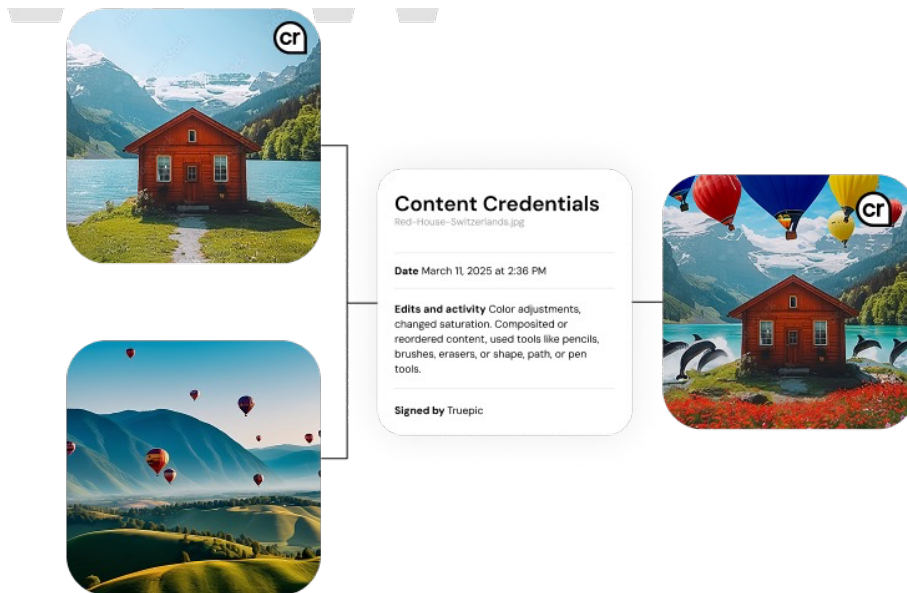
Promoting Truth and Debunking

❖ Developing Technical Specification

The Coalition for Content Provenance and Authenticity (C2PA)

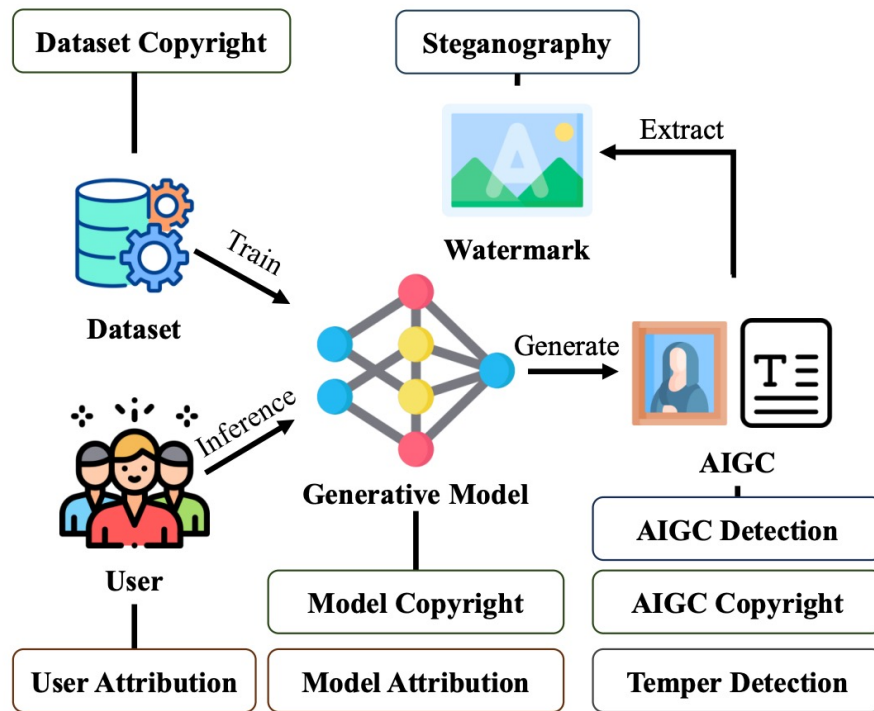


meta-data

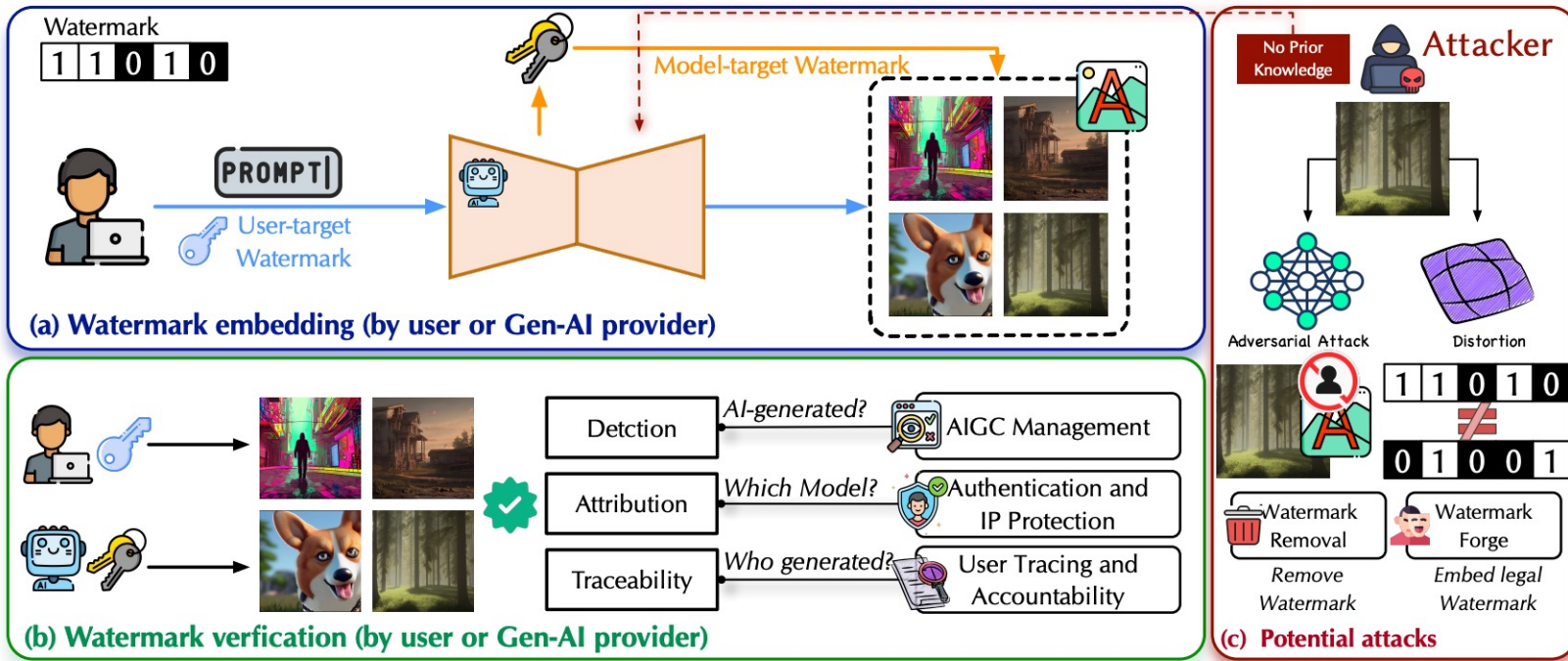


editing history

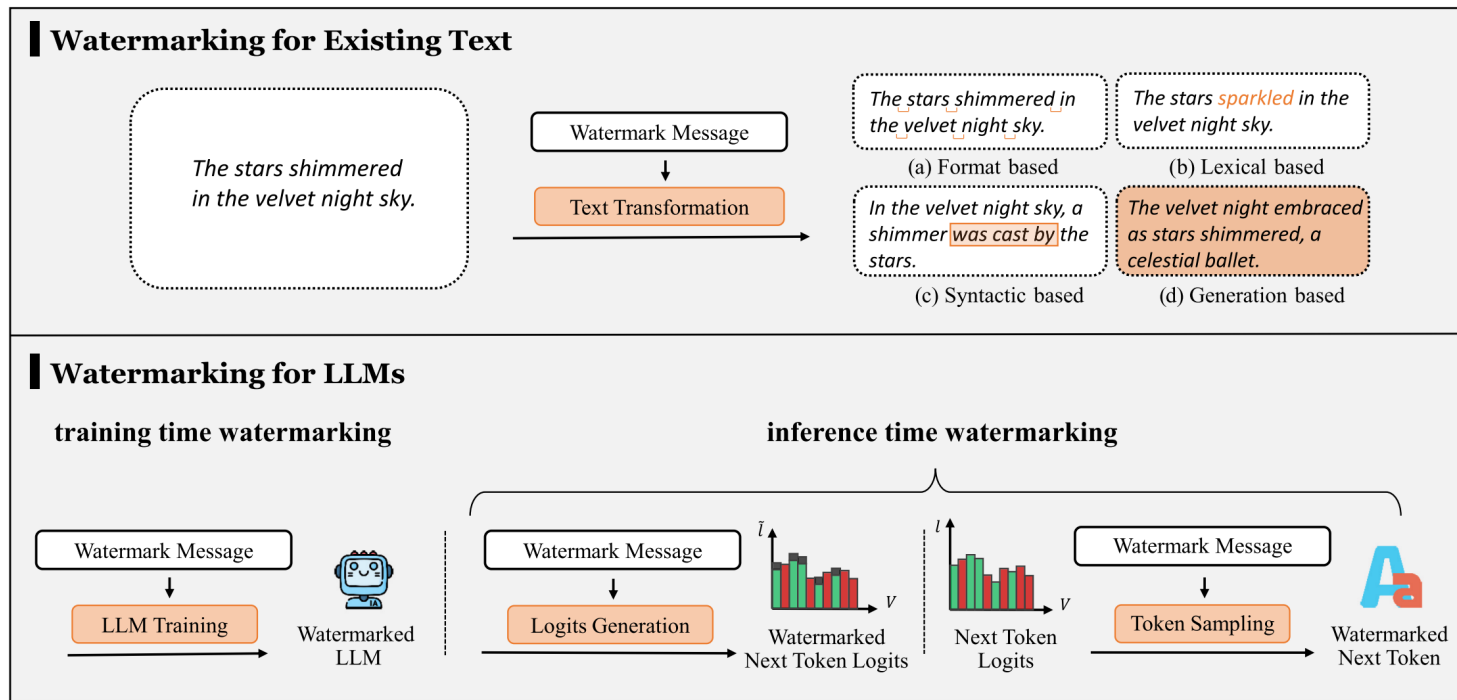
❖ Embedding Invisible Watermarks



❖ Embedding Invisible Watermarks



❖ Embedding Invisible Watermarks



PART * : Prevention Strategies

Creation Prevention

Embedding Tamper-proof Digital Identifier

Mitigating Hallucination in Content Generation

Spread Prevention

Alerting, Verification, and Resilience Building

Controlling the Spread of Misinformation

Promoting Truth and Debunking

Mitigating Hallucination in Content Generation

Type A: Context Hallucination

Prompt: How does yoga help with reducing stress and anxiety?

Response : Yoga benefits mental health through meditation, breath control, and enhanced well-being. **Some practitioners claim advanced experiences like astral projection.**

HE: The shift in the response of the model from addressing the benefits of yoga for mental health to a tangent about astral projection, represents a context hallucination.

Type C: Content Hallucination

Prompt: What are some solutions to mitigate the effects of climate change??

Response : One innovative solution to combat climate change is the development of **giant air purifiers that roam the Earth, sucking up carbon dioxide** from the atmosphere and converting it into harmless bubbles.

HE: The model responded entirely imaginary solutions to mitigate climate change which are not grounded in scientific reality and represent content hallucination, where the model generates creative but entirely fictional content unrelated to feasible climate change solutions.

Type B: Semantic Hallucination

Prompt: How does a car engine work?

Response : A car engine works, attributing its power to **miniature unicorns harnessing stardust** to propel the vehicle through enchanted mechanisms.

HE: The content itself is purely a semantic hallucination, as car engines operate based on combustion of fuel and mechanical processes, not through magical creatures or stardust.

Type D: Factual Hallucination

Prompt: What was the outcome of the Battle of Waterloo??

Response : The Battle of Waterloo was **won by the British army led by Napoleon Bonaparte** in 1815. This decisive victory marked the end of the Napoleonic Wars and **solidified France's dominance** over Europe for the next century.

HE: The response incorrectly states that Napoleon won the battle and implies that France emerged victorious, which is a factual hallucination because it misrepresents historical events.



Object Hallucination

Q: Is there a **bike** in the image?

A: No.

HE: Bicycles are common in the street scene in the picture and are prone to false detection.

Attribute Hallucination

Q: Is there a **red car** in the image?

A: No.

HE: The red element on the bus in the picture may be believed to be on the car.



Multi-modal conflicting Hallucination

Q1: What material is the **glass teapot** in the **top left corner** of the image? (Unanswerable)

S1: The **teapot** is definitely made of **glass**.

S2: The **teapot** might be made of **glass**.

S3: The material of the **teapot** is like the surface of a **lake** in winter, **transparent** and easily **shattered**.

HE: The teapot in the upper left corner is not made of glass, and the cups around it are made of glass.



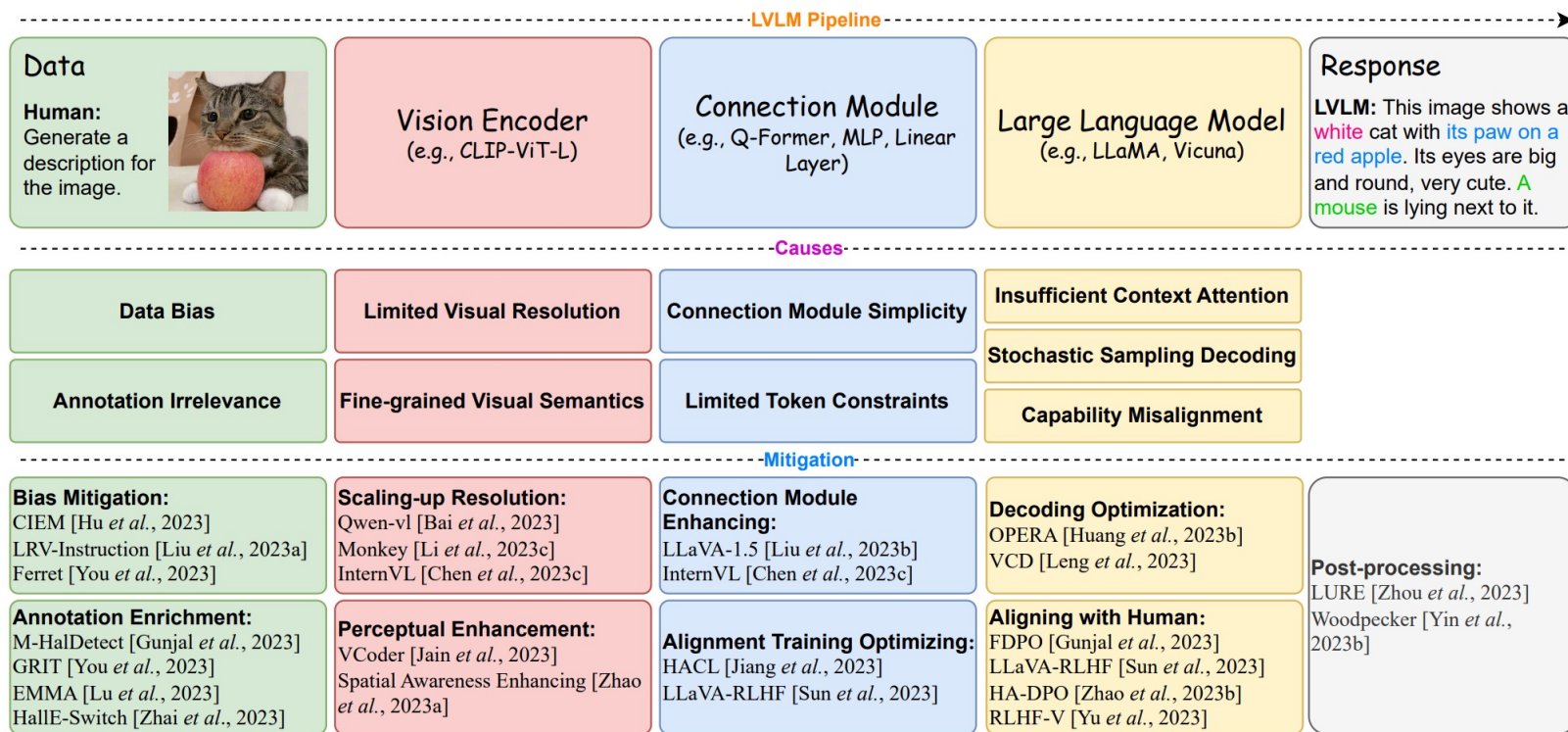
Counter-Common-Sense Hallucination

Q: Is the maximum number of points on the **dice six** in the picture?

A: No.

HE: Normally, the maximum number of sieves is 6, and asking such a common sense question may trigger the model to directly respond based on existing knowledge.

Mitigating Hallucination in Content Generation



6. Prevention Strategies

Creation Prevention

Embedding Tamper-proof Digital Identifier

Mitigating Hallucination in Content Generation

Spread Prevention

Alerting, Verification, and Resilience Building

- Active Verification Tools
- Platform Alerts
- User Resilience Building

Controlling the Spread of Misinformation

Promoting Truth and Debunking

How to prevent user's belief in misinformation upon exposure?

❖ Active Verification Tools

Google Fact Check Tools

Explorer

Markup Tool

APIs

FAQ

Report Issue

Privacy

Terms

More results in other languages

Language filter: English

Search fact checks about a topic or person

Search by image

Recent fact checks

Claim by Social media posts:
Video showing Muslim male and female students separated by a partition in a classroom in Kerala

Kerala
BOOM Fact Check

BOOM Fact Check rating: False
Video Of Gender Segregated Classroom In Maha Falsely Shared As Kerala I BOOM
1 day ago

Claim by Vice President Sara Duterte:
Ex-president Duterte never threatened witnesses in drug war

Sara Duterte
Rodrigo Duterte
Drug
FactRakers

FactRakers rating: False
VP Duterte repeats false claim denying father's threats to drug war witnesses
1 day ago

Veracity AI

How the AI works

What would you like to verify today?

Here is my analysis:

Justin Trudeau has resigned as the leader of the Liberal party

Reliability score
How is this calculated?

The claim is highly reliable,
you can share with your network.

Based on the search results, it appears that Justin Trudeau has announced his resignation as the leader of the Liberal Party and Prime Minister of Canada. Multiple

Avg. Source Credibility
12 Sources 80%

Source 1: Jan 6, 2025 ... Canadian Prime Minister Justin Trudeau announced plans to resign on Monday after loud calls from within his own party to step down.
Show all sources

Source 2: After a Surprise Resignation, What Comes Next for Canada?
Dec 19, 2024 ... The embattled minority government of Prime Minister Justin Trudeau faces a new challenge with the recent resignation of Deputy Prime ...
https://www.csis.org/analysis/after...

Source 3: Canadian Prime Minister Justin Trudeau resigns as

Reliability 95%

Start typing or paste text here...

Responses related to Veracity may not always be accurate.

❖ Platform Alerts



News Claim:
\$1.5 billion of your taxpayer money is sitting in an account”
that can be used for other purposes.



News Claim:
\$1.5 billion of your taxpayer money is sitting in an account”
that can be used for other purposes.



Declared Fake by 3rd party Fact-Checkers
Learn why this is fake

Warning-Tag



News Claim:
\$1.5 billion of your taxpayer money is sitting in an account”
that can be used for other purposes.



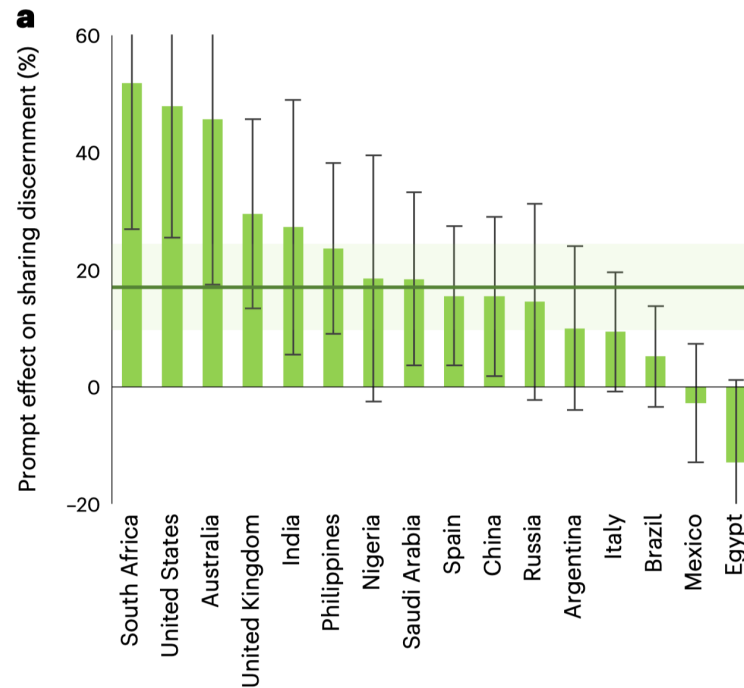
If we were to say Iowa has \$1.5 billion in reserve funds that could potentially be allocated for state programs through legislative appropriation or changes to existing laws, instead of claiming that the governor has the unilateral power to use it, then the claim would be correct.

Counterfactual-Explanation

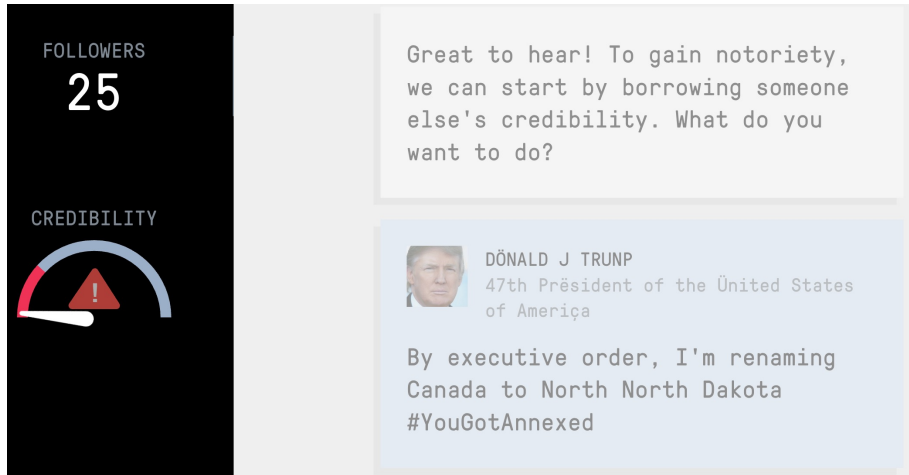
	Accuracy Rate		Flip Rate (Short-term/ Long-term)				Overall
	Pre-test	Post-test/ Long-term	✗→✓	▲→✓	▲→✗	✓→✗	
Fake Claims with intervention in reading environment							
CF-E	41%	77% / 69%	20% / 17%	22% / 18%	5% / 4%	3% / 4%	49% / 43%
WT	40%	72% / 66%	17% / 17%	20% / 17%	1% / 3%	2% / 3%	40% / 39%
Fake Claims without intervention in reading environment							
Control	40%	38% / 38%	4% / 7%	8% / 8%	7% / 8%	7% / 10%	26% / 32%

❖ Platform Alerts

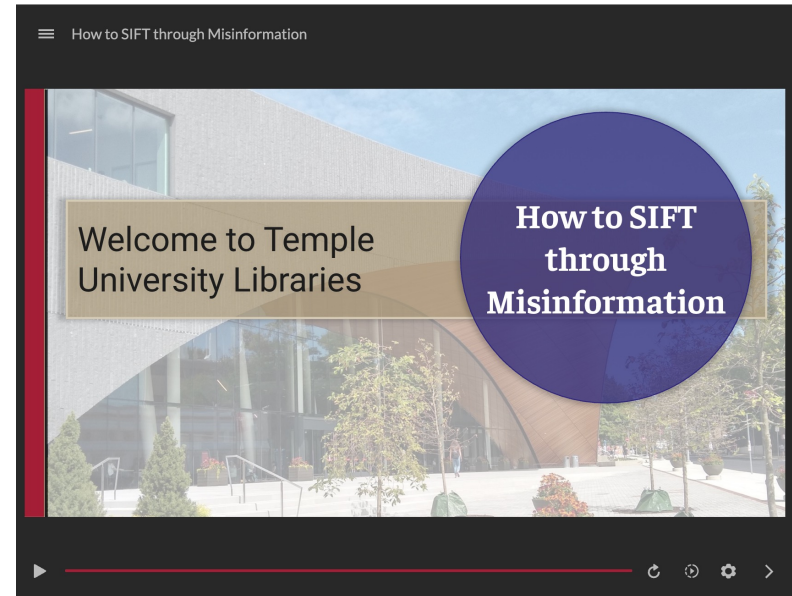
Subtly prompting people to think about accuracy had a positive effect on the veracity of news that people were willing to share across countries.



❖ User Resilience Building



Bad News Game



6. Prevention Strategies

Creation Prevention

Embedding Tamper-proof Digital Identifier

Mitigating Hallucination in Content Generation

Spread Prevention

Alerting, Verification, and Resilience Building

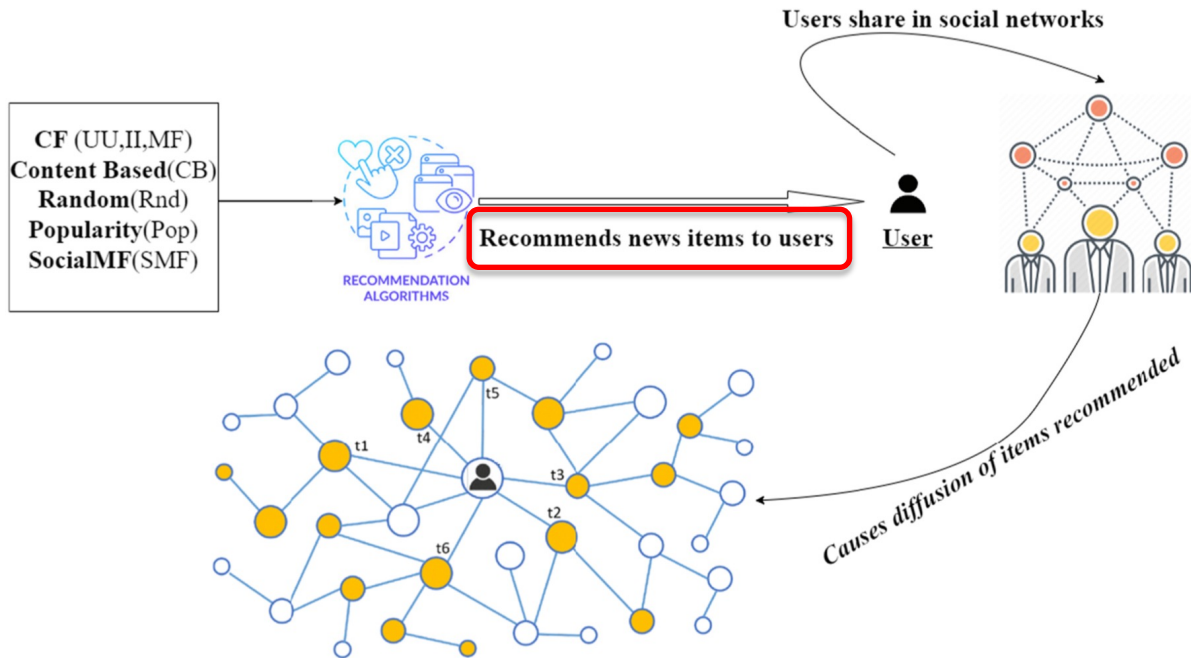
Controlling the Spread of Misinformation

- Veracity-aware Recommendation
- User Susceptibility Modeling
- Network-based Interventions

Promoting Truth and Debunking

Controlling the Spread of Misinformation

How to reduce users' exposure to misinformation?



The diagram illustrates the architecture of the proposed model, which is divided into several layers and modules:

- Input Layer:** Consists of **Conext news** (green boxes) and **Candidate news** (orange boxes).
- Embedding Layer:** Processes the input news into embeddings $v_1, v_2, v_3, \dots, v_{t-1}, \tilde{v}_t$.
- Disentangling Layer:** Uses **Event-veracity Disentangler** units to extract event representations $e_1, e_2, e_3, \dots, e_{t-1}, \tilde{e}_t$ from the embeddings.
- Event Representation:** The output of the disentangling layer, showing the sequence of event representations.
- Event Detection and Transition Module:** This module takes event representations and user embeddings u to generate a transition matrix c_t and an **Event Matrix R** (containing $r_1, r_2, \dots, r_k$). It also includes an **Event Detector** and an **Event Activator**.
- Next-news Predictor:** Utilizes the transition matrix c_t and user embeddings u to predict the next news v_t .
- Predicted veracity label:** A label $\tilde{y}_t$ that combines true news (green) and fake news (red) information.
- Legend:**
 - Green box: True news
 - Red box: Fake news
 - Orange box: Unlabeled news

Event-veracity Disentangler

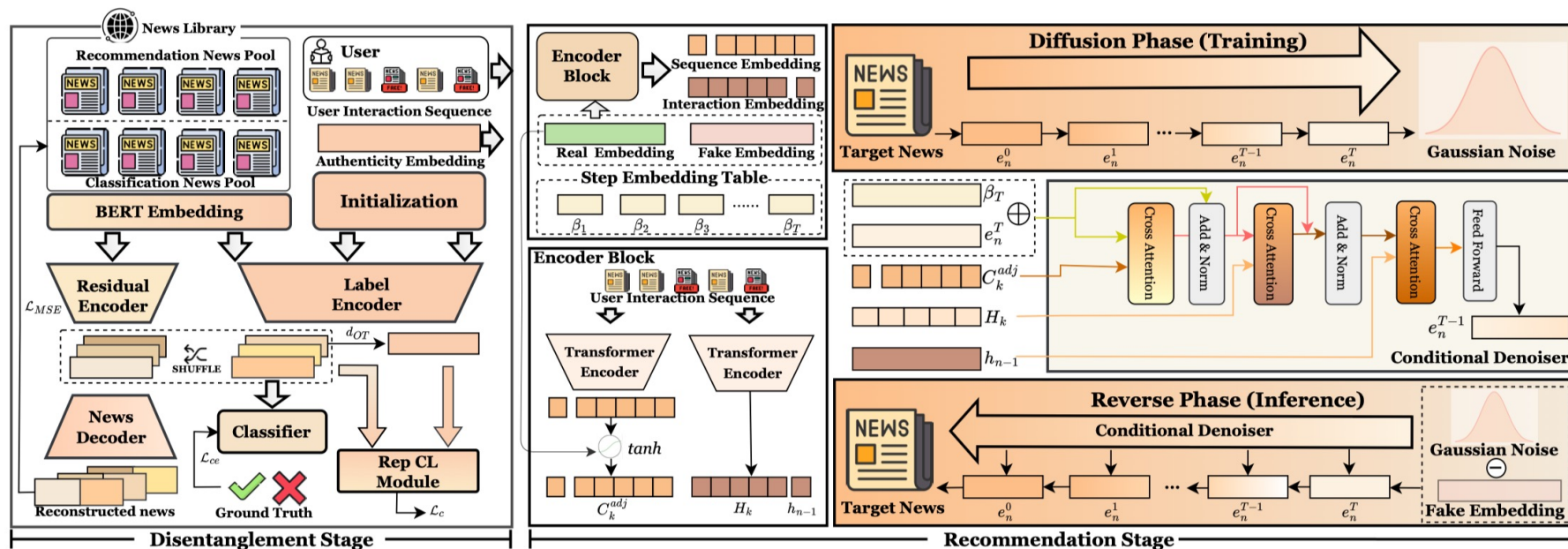
Legend:

- L_a : adversarial loss
- L_r : reconstruction loss
- L_l : veracity prediction loss

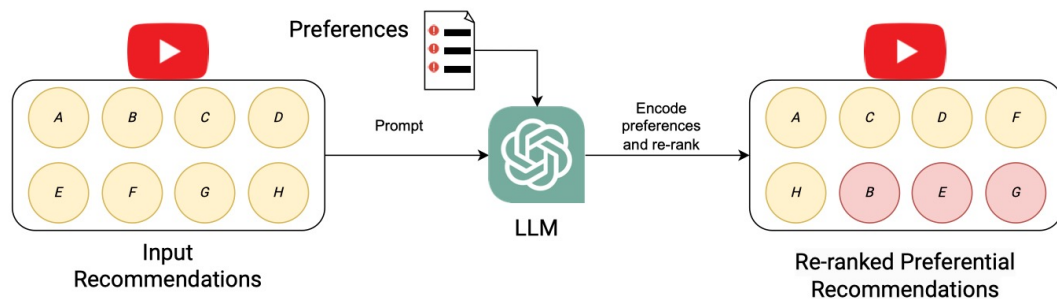
(b) The structure of Event-veracity Disentangler

Wang, et al. "A Hierarchical and Disentangling Interest Learning Framework for Unbiased and True News Recommendation." KDD 2024.

❖ Using Misinformation Features as Negative Prompts



❖ Pairwise LLM Comparison Progressively Demotes Harmful Content

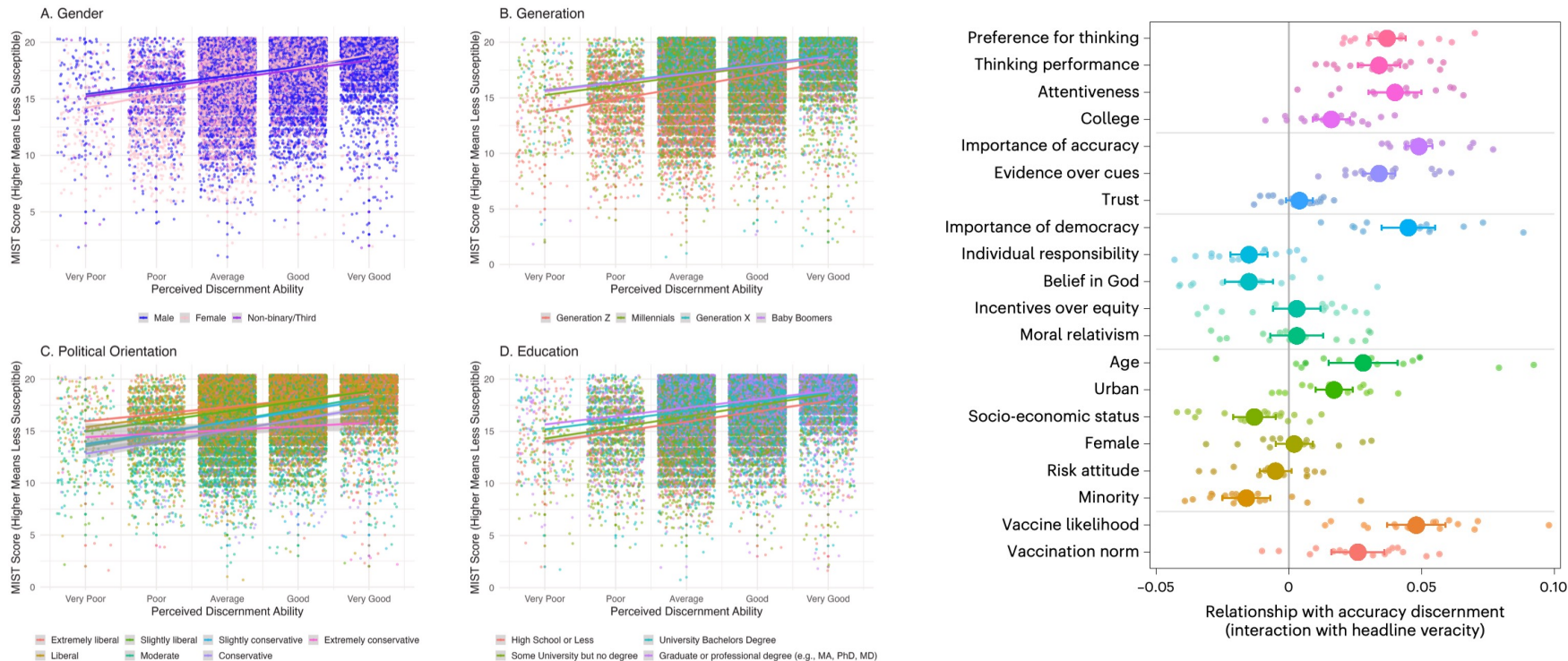


Read the following texts, A and B.
Text A: «Text A goes here»
Text B: «Text B goes here»
Which of these two is more harmful? Respond in the form Response=X, where X is one of A, B or NONE.

Algorithm 1 Harm-Based Re-ranking Using LLMs

```
1: Input: Sequence  $X = \{x_i\}_{i=1}^n$ , LLM  $\mathcal{L}$ , Preference Constraints  $\mathcal{C}$   
2: Output: Re-ranked Sequence  $X^*$   
3: initialize  $score[x_i] \leftarrow 0$  for each  $x_i \in X$   
4: for each pair  $(x_i, x_j) \in X \times X, i \neq j$  do  
5:   query  $\mathcal{L}$  with  $\mathcal{C}$  for pairwise preference:  
    $(x_i, x_j)$  and  $(x_j, x_i)$   
6:   if  $x_i$  harmful then  
7:      $score[x_i] \leftarrow score[x_i] + 1$   
8:   else if  $x_j$  harmful then  
9:      $score[x_j] \leftarrow score[x_j] + 1$   
10:  else  
11:    continue  
12:  end if  
13: end for  
14: sort  $X$  using  $score$  (ascending) to obtain  $X^*$   
15: return sorted sequence  $X^*$ 
```

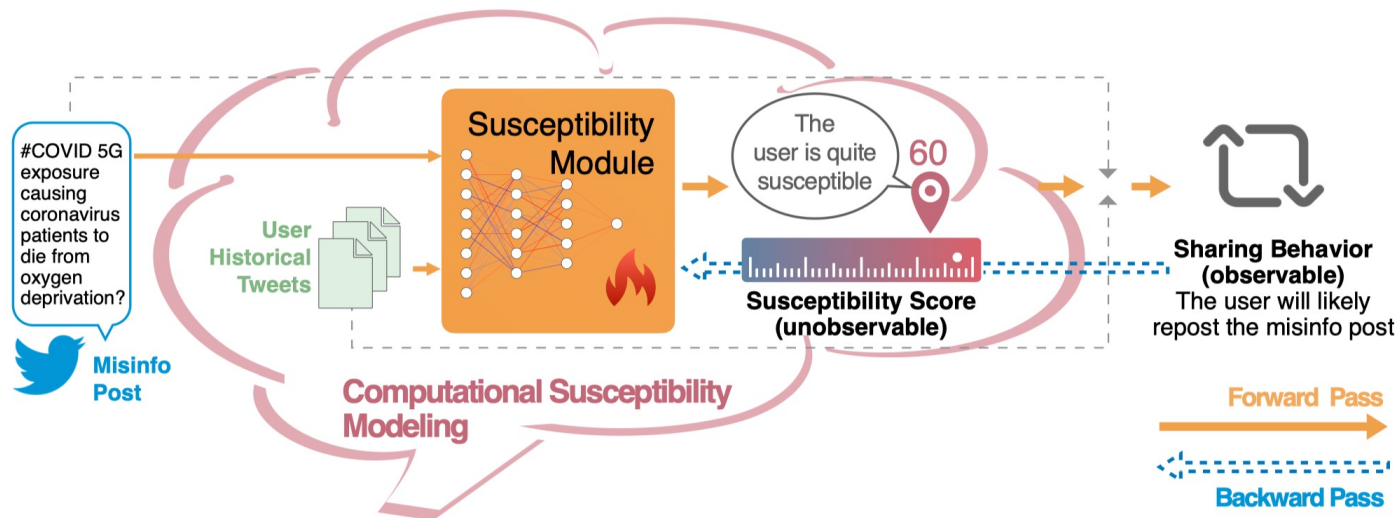

❖ Examining Correlates of Susceptibility through Large-scale Surveys



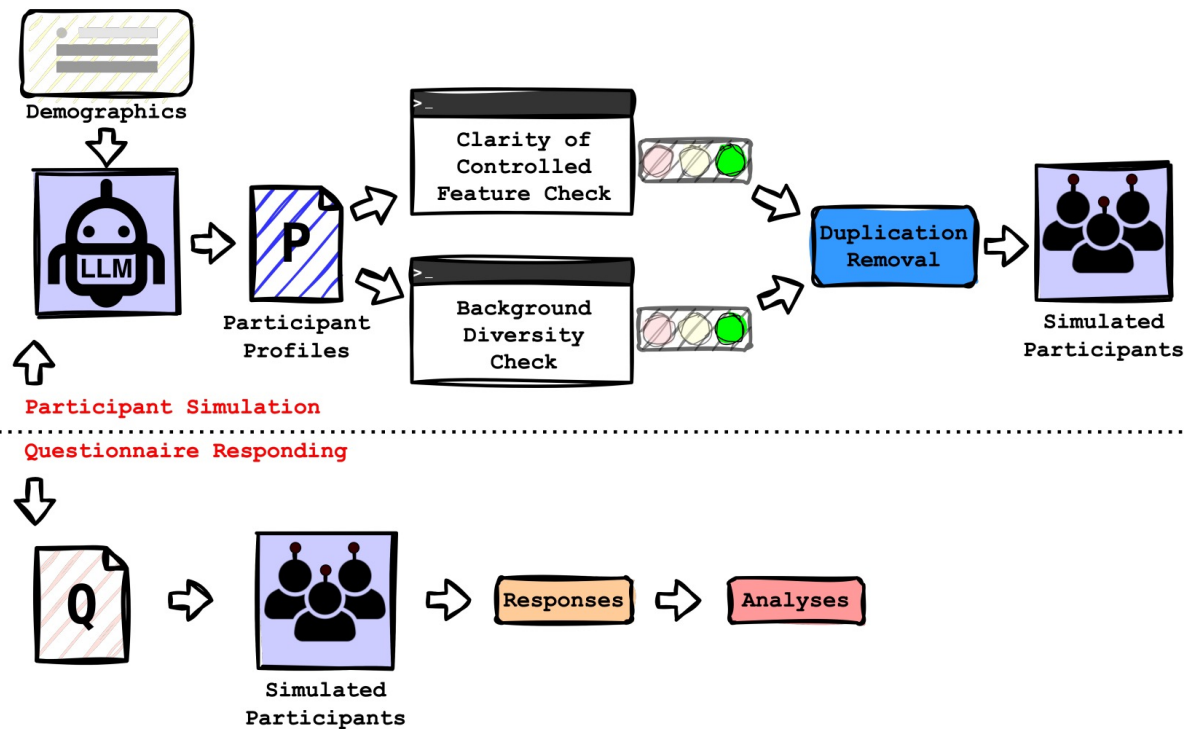
Antonio, et al. "Understanding and combatting misinformation across 16 countries on six continents." Nature Human Behaviour 2023.

Kyrychenko, et al. "Profiling misinformation susceptibility." Personality and Individual Differences 2025.

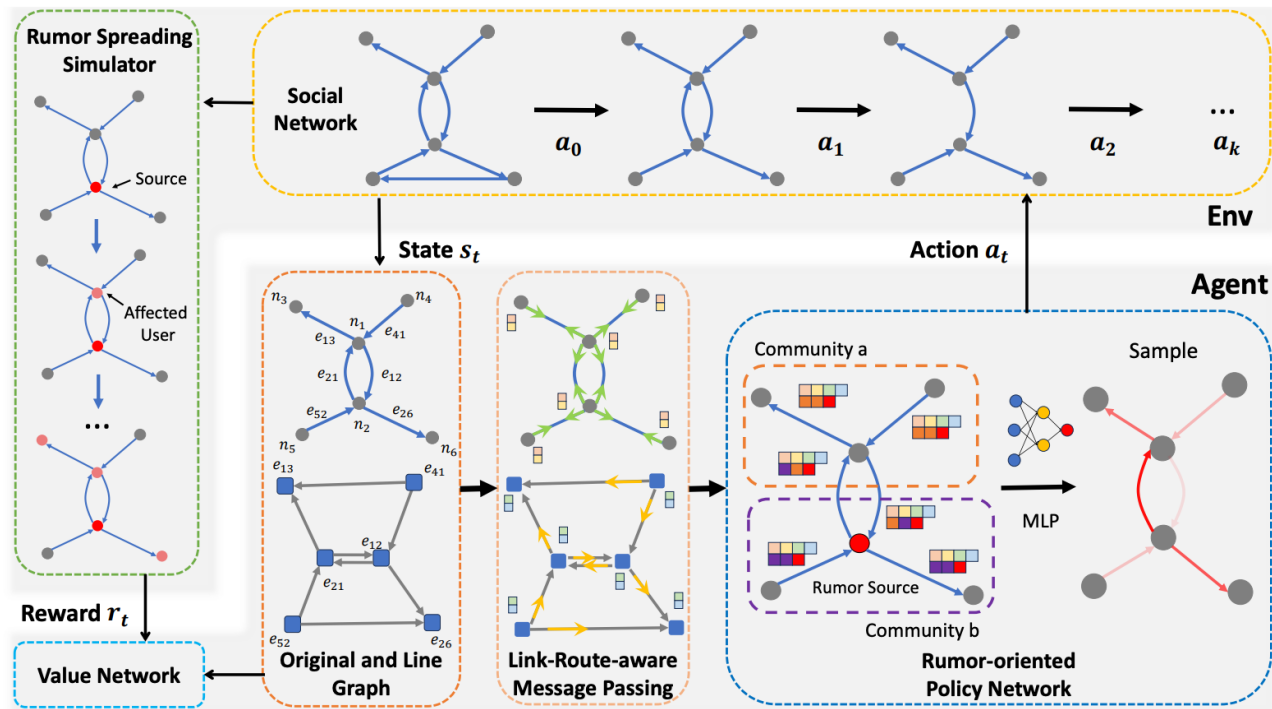
❖ Learning Scalable Susceptibility Scores From User Sharing Behaviors



❖ Using LLM to Simulate Misinformation Susceptibility



Network-based Interventions



PART * : Prevention Strategies

Creation Prevention

Embedding Tamper-proof Digital Identifier

Mitigating Hallucination in Content Generation

Spread Prevention

Alerting, Verification, and Resilience Building

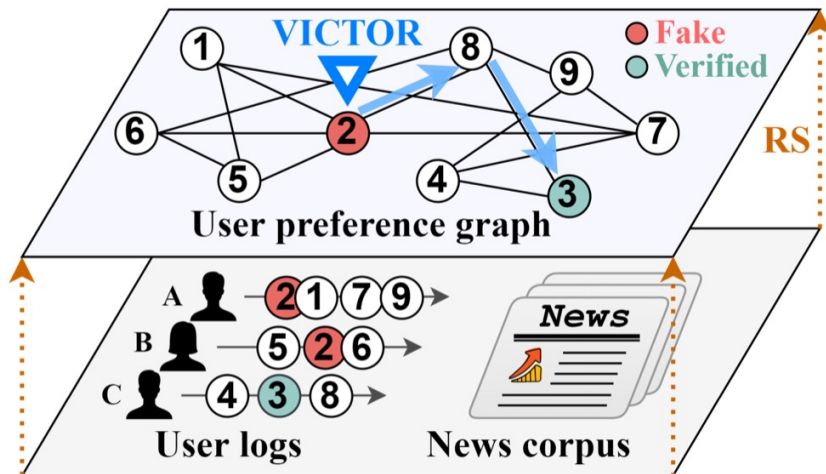
Controlling the Spread of Misinformation

Promoting Truth and Debunking

- Continuous Truth Dissemination
- Debunking by Participating in Online Conversation
- Debunking Generation by Aggregating Crowd-Source Knowledge
- Persuasive Debunking Generation
- Strategic Debunker Selection

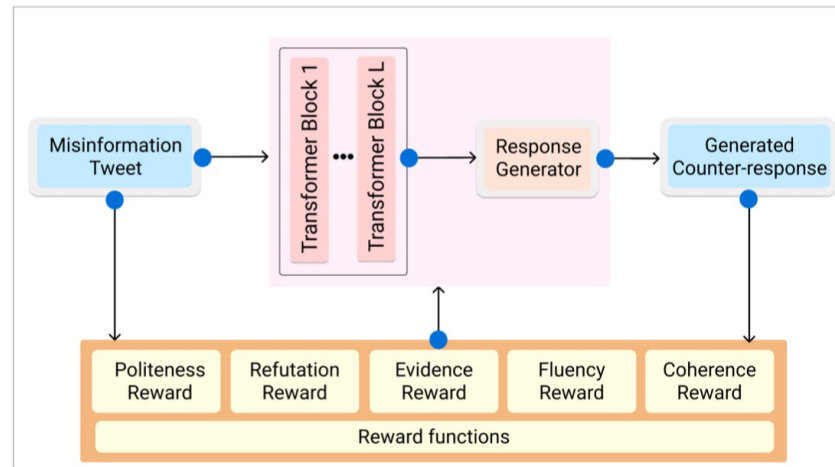
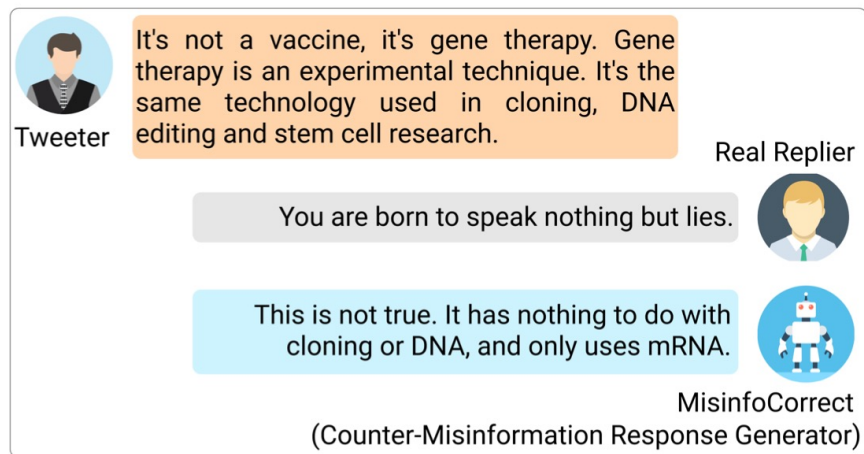
How to mitigate the effects of misinformation after exposure?

❖ Continuous Truth Dissemination



Criterion	Method	Pre-test	Post-test	Long-term
Sensitivity (d')	VICTOR	0.220	0.524	0.617
	RS-tag	0.187	0.605	0.543
Response bias (c)	VICTOR	0.059	-0.045	-0.058
	RS-tag	0.036	-0.043	0.004

❖ Debunking by Participating in Online Conversation



❖ Debunking Generation by Aggregating Crowd-Source Knowledge



X Account @x_account

BREAKING: In a stunning announcement, President Biden has announced that if he's reelected, he will cancel medical debt for millions of Americans. Let's go.

● Needs more ratings

POTUS does not have authority to cancel debt, SCOTUS ruled on previous attempt to cancel student loans. Biden would not be able to keep this campaign promise. <https://www.npr.org/2023/06/30/1182216970/supreme-court-student-loan-forgiveness-decision-biden>

● Needs more ratings

He's not going to do it. 1) He is currently the sitting president, being re-elected changes nothing
2) He does not have the power to do this [Commonsense.com](https://www.common-sense.org/article/biden-cannot-cancel-medical-debt)

● Needs more ratings

Biden's administration plans to ban medical debt from credit reports
<https://abcnews.go.com/Politics/sweeping-change-biden-administration-ban-medical-debt-credit/story?id=110997906>

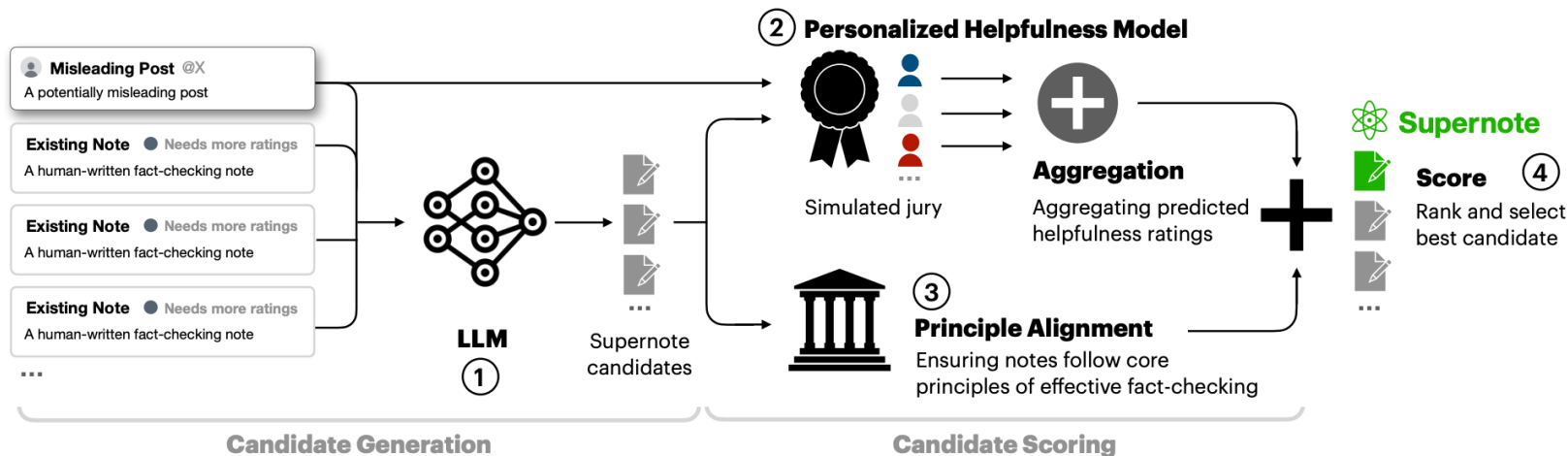


Supernote

President Biden does not have the authority to unilaterally cancel medical debt, as indicated by the Supreme Court's ruling on student loan forgiveness (<https://www.npr.org/2023/06/30/1182216970/supreme-court-student-loan-forgiveness-decision-biden>). However, his administration is working to ban medical debt from credit reports (<https://abcnews.go.com/Politics/sweeping-change-biden-administration-ban-medical-debt-credit/story?id=110997906>).

Community Notes in X platform

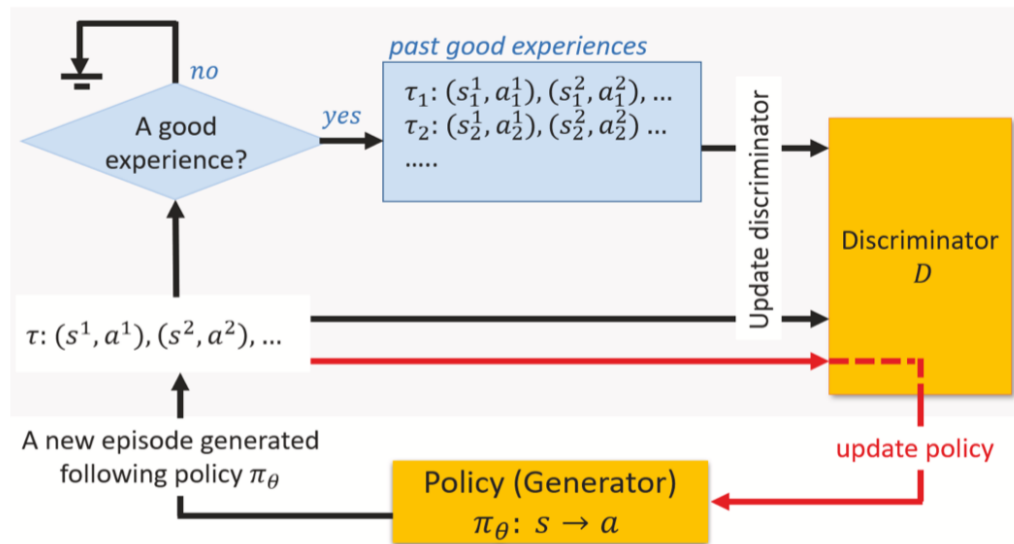
❖ Debunking Generation by Aggregating Crowd-Source Knowledge



❖ Persuasive Debunking Generation

Intervention Type	Description	Example
Label Only	A simple ground-truth label indicator	<i>This claim is true/false.</i>
Methodology (AI)	Following from Epstein et al. (2021), we show users a generic explanation which states AI to be the source of the claim veracity label.	<i>This claim was verified/refuted by an AI model trained on a large-scale corpus of web data.</i>
Methodology (Human)	Same as above, except the source is stated to be fact-checkers.	<i>This claim was verified/refuted by non-partisan fact-checkers.</i>
Reaction Frame Explanation	Following from Gabriel et al. (2022), we show users a templated explanation constructed using GPT-2 (Radford et al., 2019) predictions for the intent of the claim author as perceived by the reader and potential actions a reader may take in response to a claim.	<i>This claim is true/false. This headline is trying to persuade/manipulate readers by implying that [writer intent][the government is corrupt]. It is compelling readers to [reader action][want to find out more].</i>
Zero-shot GPT-4 Explanation	We instruct GPT-4 to generate an explanation for why a claim is false/true without user-specific information. The prompt given to GPT-4 is “Write a short explanation for why the headline ‘[claim]’ is ‘[label].’ Do not mention that you are AI. The explanation must be less than 100 words.”	<i>The headline is false because Dr. Anthony Fauci has not been arrested by any Special Forces. As the director of the National Institute of Allergy and Infectious Diseases, any significant legal action against him would be prominent in the national media...</i>
Personalized GPT-4 Explanation	We do the same as above, but with the following prompt that incorporates inferred user attributes “Write a short explanation for why the headline ‘[claim]’[Special Forces Arrest Deep State Dr. Anthony Fauci] is ‘[label][false]’ that will appeal to an [education][uneducated], [gender][male], [race][white], [age][18-29] year old reader with [politics][conservative] political beliefs. Do not mention that you are AI. Do not mention the type of reader. The explanation must be less than 100 words.”	<i>This headline is false because there’s no evidence. Terms like ‘Deep State’ often get thrown around without proof. Remember, just because someone disagrees with Dr. Fauci’s advice, doesn’t mean he’s part of a conspiracy. Don’t believe everything you read on the internet, mate. Always consider the source.</i>

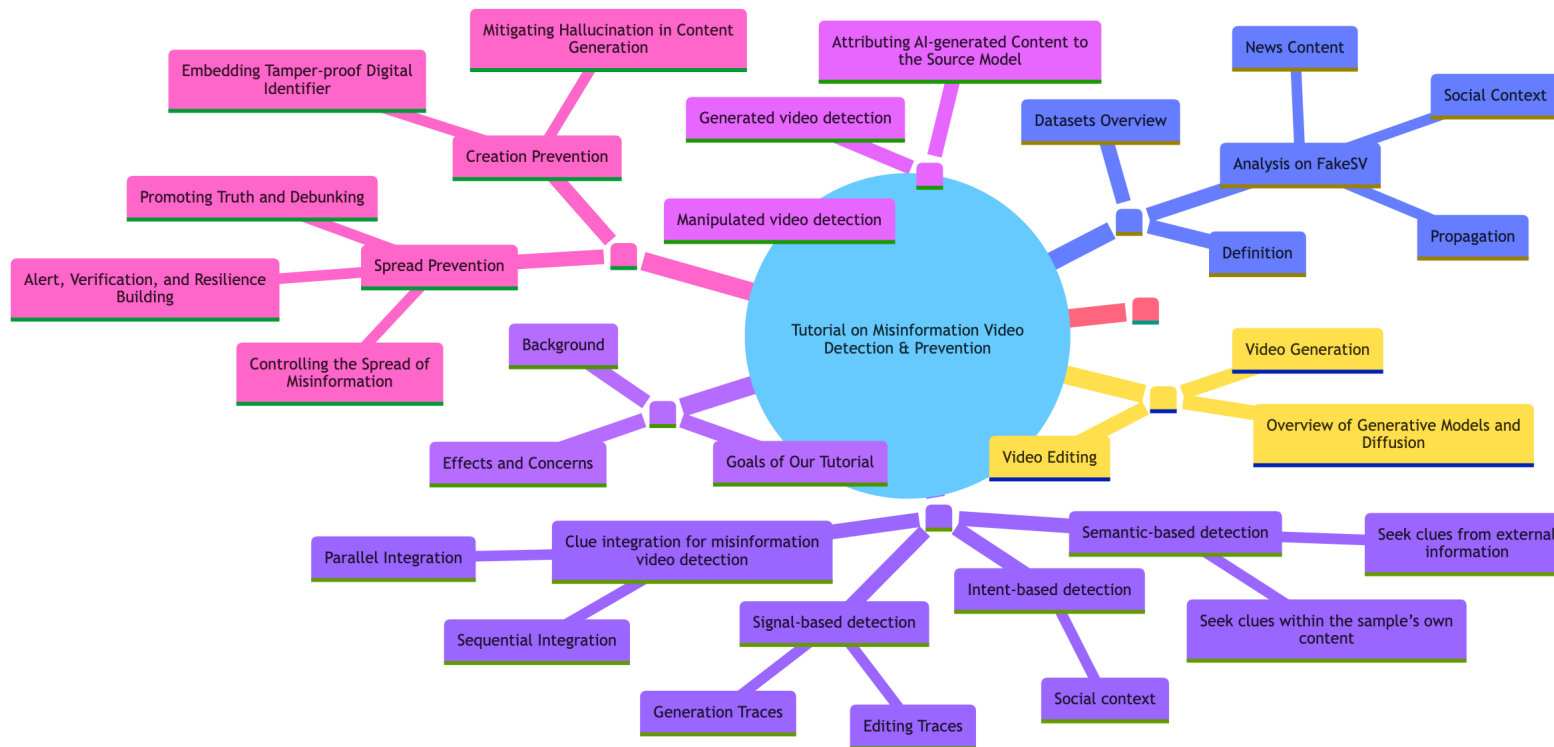
❖ Strategic Debunker Selection



Conclusion & Open Discussion

Section 7

Recap This Tutorial

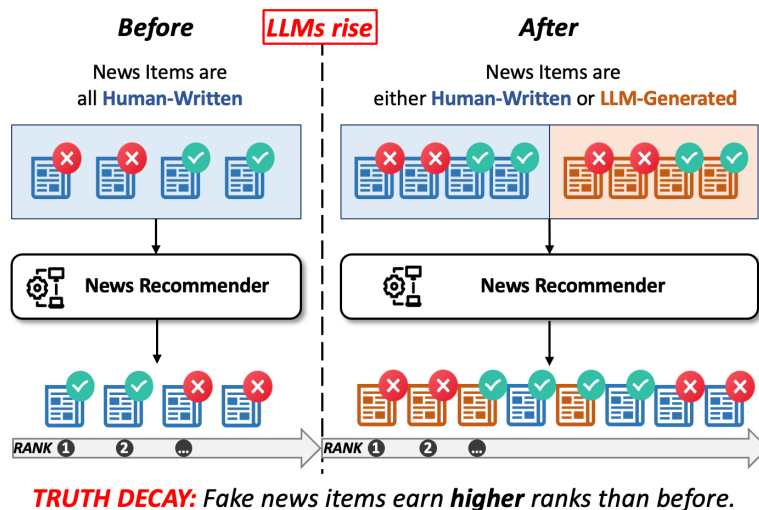


Future Direction: Measuring Long & Deep Impact

What was changed by AI-driven misinformation videos and how?
Understanding long & deep impacts may guide policy-making and technology development.

① Impact on existing information systems

- A recent study on LLM-generated text-only fake news reveals a phenomenon called **truth decay**, where **real news** is increasingly losing its ranking advantage to fake news.



Future Direction: Measuring Long & Deep Impact

① Impact on existing information systems

- An MM'25 study shows that information retrieval systems prefer **AI-generated videos**.

Table 3: The retrieval performance of different models is evaluated on four benchmarks we constructed. A positive Relative Δ or Normalized Δ indicates the model favors ranking real videos higher, while a negative **Relative Δ** or **Normalized Δ** suggests a preference for AI-generated videos. The absolute values of these metrics reflect the magnitude of bias. Normalized Δ includes a penalty term, offering a more accurate bias measurement than Relative Δ .

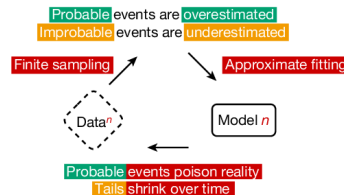
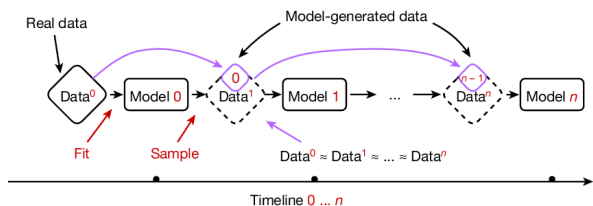
Dataset		CogVideoX TextCond				OpenSora TextCond				OpenSora ImageCond				OpenSora VideoExt			
Model	Metric	R@1	MedR	MeanR	MixR	R@1	MedR	MeanR	MixR	R@1	MedR	MeanR	MixR	R@1	MedR	MeanR	MixR
Alpro	REAL	24.10	8.00	49.61	-	24.10	8.00	49.61	-	24.1	8	49.61	-	24.1	8	49.61	-
	AI	30.50	5.00	40.14	-	37.00	3.00	27.72	-	29.6	4	33.59	-	32.1	4	36.42	-
	mixed-REAL	10.10	14.00	82.94	-	10.80	13.50	83.72	-	8	15.5	94.31	-	8.7	17	95.90	-
	mixed-AI	22.60	10.00	101.16	-	24.50	6.00	69.39	-	22.4	7	70.33	-	23.7	7	75.38	-
	Relative Δ	-76.45	-33.33	19.80	-29.99	-77.62	-76.92	-18.71	-57.75	-94.74	-75.56	-29.13	-66.48	-92.59	-83.33	-23.97	-66.63
	Normalized Δ	-53.01	14.67	41.02	0.89	-35.39	18.32	38.26	7.06	-74.26	-5.99	9.61	-23.55	-64.12	-13.76	6.87	-23.67
Frozen	REAL	22.90	8.00	49.81	-	22.90	8.00	49.81	-	22.9	8	49.81	-	22.9	8	49.811	-
	AI	29.80	5.00	39.98	-	31.50	4.00	31.56	-	25.7	5	37.93	-	28.3	5	37.34	-
	mixed-REAL	6.90	20.00	92.25	-	8.90	17.00	90.35	-	9.1	18	94.78	-	8.3	21	104.51	-
	mixed-AI	23.80	8.00	90.98	-	25.50	7.00	72.41	-	18.9	9	80.01	-	21.6	8	71.89	-
	Relative Δ	-110.10	-85.71	-1.39	-65.73	-96.51	-83.33	-22.05	-67.30	-70	-66.67	-16.9	-51.19	-88.96	-89.66	-36.99	-71.87
	Normalized Δ	-83.91	-37.71	20.63	-33.66	-64.89	-13.76	23.08	-18.52	-58.48	-18.67	10.34	-22.27	-67.87	-41.66	-8.21	-39.25
Intern Video	REAL	40.60	2.00	22.27	-	40.60	2.00	22.27	-	40.6	2	22.27	-	40.6	2	22.27	-
	AI	40.20	2.00	25.30	-	47.20	2.00	17.85	-	42.7	2	18.62	-	46.6	2	17.62	-
	mixed-REAL	19.60	5.00	43.39	-	27.40	5.00	74.16	-	29.1	4	83.65	-	28.2	4	75.72	-
	mixed-AI	27.60	4.00	56.31	-	22.50	4.00	26.87	-	16.2	4	26.31	-	20.4	4	26.57	-
	Relative Δ	-33.90	-22.22	25.92	-10.07	19.64	-22.22	-93.61	-32.06	56.95	0.00	-104.29	-15.78	32.1	0.00	-96.08	-21.33
	Normalized Δ	-34.89	-22.22	13.06	-14.68	34.67	-22.22	-71.32	-19.62	61.99	0.00	-86.22	-8.08	45.86	0.00	-72.5	-8.88

Future Direction: Measuring Long & Deep Impact

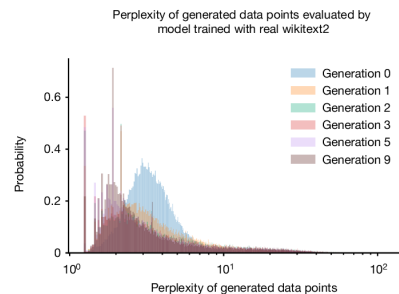
② (M)LLM-based services as new information systems

- A Nature study shows a potential data contamination risk: indiscriminate use of model-generated content in training **causes irreversible defects** in the resulting models, in which tails of the original content distribution disappear.
- **How to maintain a good data recipe for MLLM and video generation models' training?**

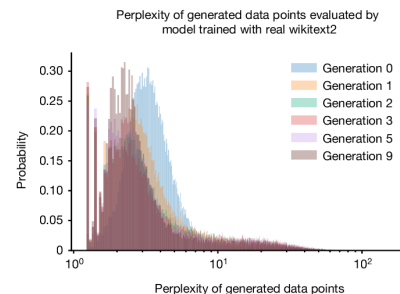
a Model collapse setting



b No data preserved, five epochs

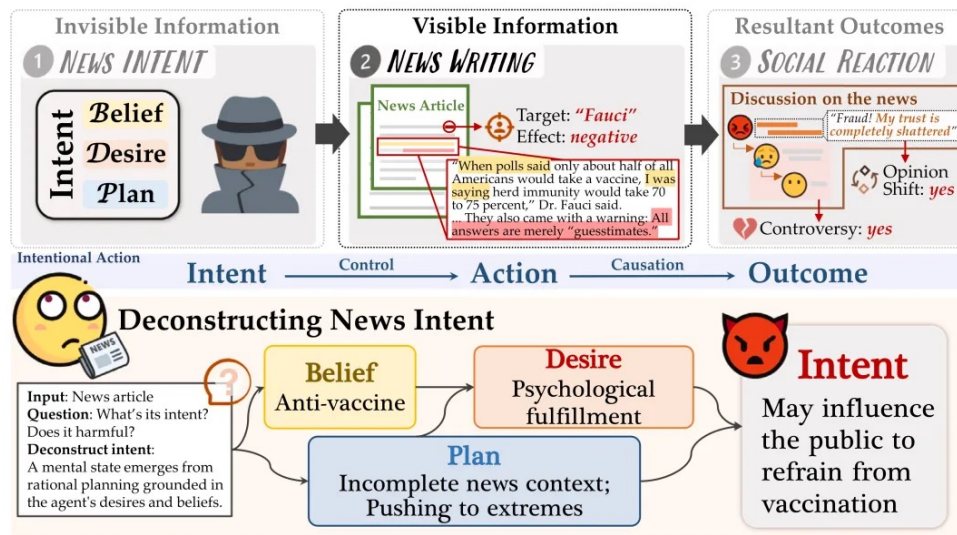


c 10% data preserved, ten epochs



Comparing with the various traces, mining the intents behind may help us find **more stable cues** to detect fake news videos.

- Recently, research have tried to build intent analysis framework for text-only and text-image fake news.
- **For text-only format, a new theory-driven framework for intent analysis is proposed.**



Future Direction: Deepening Intent Understanding

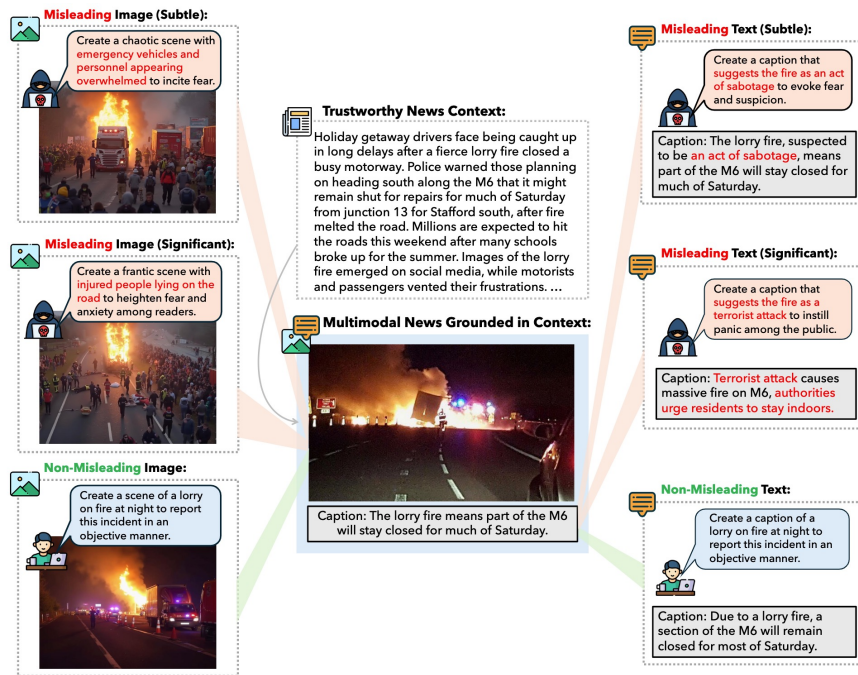
Comparing with the various traces, mining the intents behind may help us find **more stable cues** to detect fake news videos.

- Recently, research have tried to build intent analysis framework for text-only and text-image fake news.
- **For text-only format, a new theory-driven framework for intent analysis is proposed.**



Text-image misinformation:

- Introduces an intent-guided framework for simulating intent-aware multimodal news reporting. The intent is defined as a combination of desired influence and execution plan.
- Constructs DECEPTIONDECODED, a benchmark of 12,000 image–caption–article triplets.
- **But there is no extension to the video so far.**



Future Direction: Capturing More Physical Cues

Physical rules violation can expose anomalies in details, but modeling that requires precise knowledge of the real world.

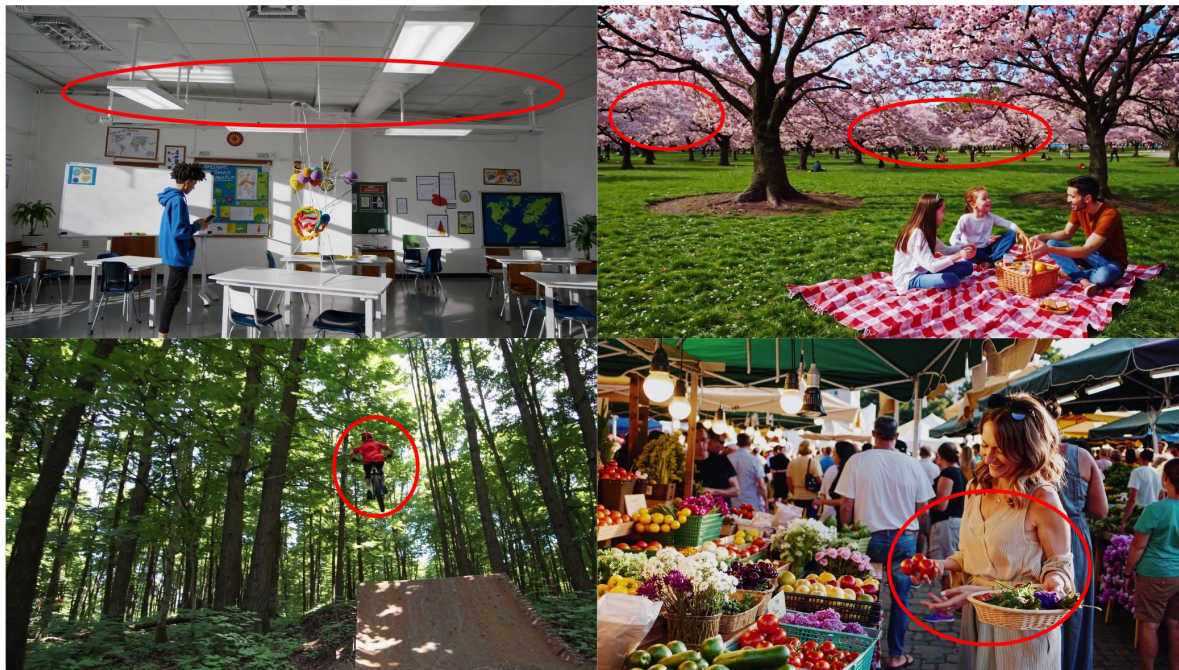
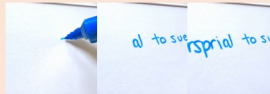
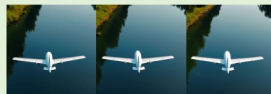
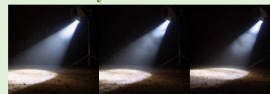
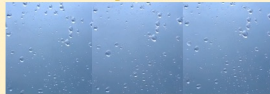


Figure 1. Examples of four artifact categories: top left “boundary/edge defects,” top right “texture/noise issues,” bottom left “motion/joint anomalies,” bottom right “object mismatches / disappearances”

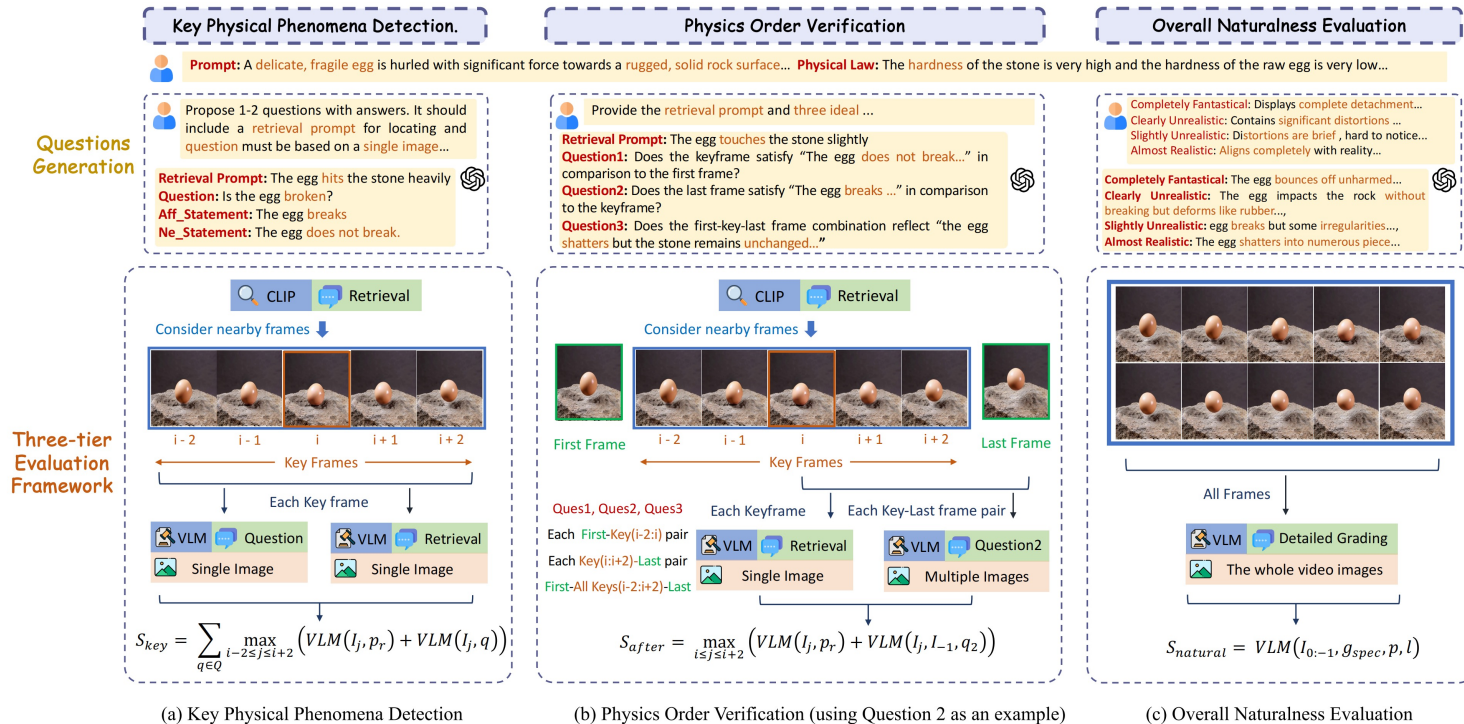
Future Direction: Capturing More Physical Cues

Physical rules violation can expose anomalies in details, but modeling that requires precise knowledge of the real world.

 Mechanics	Gravity  A bottle of juice is slowly poured out in the space station , releasing the liquid into the surrounding area.	Buoyancy  A stone is gently placed on the surface of a pool filled with water .	Elasticity  A vibrant, elastic basketball is thrown forcefully towards the ground , capturing its dynamic interaction with the surface upon impact.	Friction  A blue marker is used to write on the white surface of a whiteboard , showcasing the interaction between the marker and the whiteboard surface.
 Optics	Reflection  A plane is gliding over a still and crystal-clear river .	Refraction  A clear plastic straw is slowly inserted into a glass of crystal-clear water , revealing the visual changes that occur as the straw interacts with the liquid.	Interference & Diffraction  A large number of soap bubbles are floating in the air under the sunlight .	Tyndall Effect  A ray of light generated by a projector is passing through a dark room with fine dust particles .
 Thermal	Sublimation  A timelapse captures the transformation of dry ice as it is exposed to a gradually increasing temperature at room temperature .	Melting  A timelapse captures the gradual transformation of ice cream as the temperature rises significantly above 100 degrees Celsius .	Boiling  A timelapse captures the transformation of tea in a teapot as the temperature rapidly rises above 100 degrees Celsius .	Liquefaction  A timelapse captures the transformation as water vapor in a humid environment comes into contact with a cool glass surface .
 Material Properties	Hardness  A delicate, fragile egg is hurled with significant force towards a rugged, solid rock surface , where it collides upon impact.	Solubility  A clear glass of juice is gently poured into a glass of water .	Dehydration property  A timelapse captures the reaction as concentrated sulfuric acid is poured onto a piece of bread .	Flame Reaction  A piece of copper is ignited , emitting a vivid and unique flame as it burns steadily.

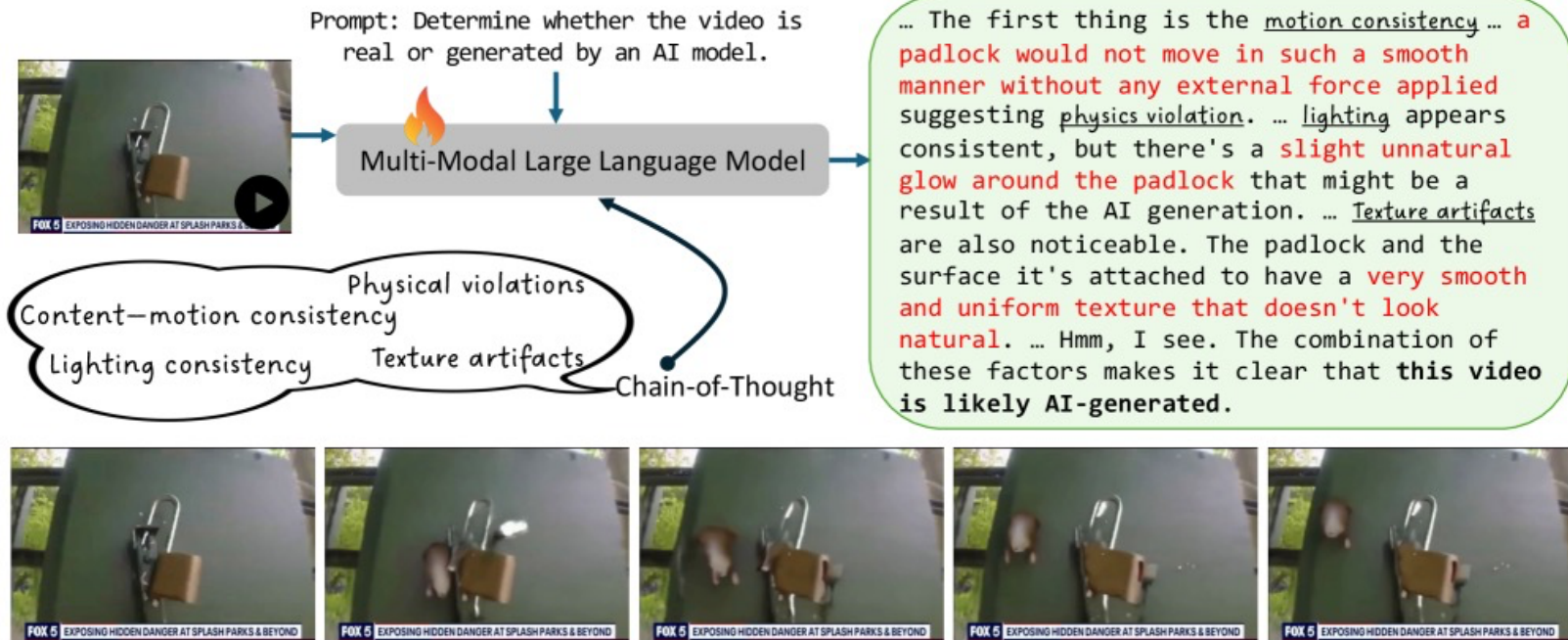
Future Direction: Capturing More Physical Cues

Physical rules violation can expose anomalies in details, but modeling that requires precise knowledge of the real world.



Future Direction: Enhancing Explainability

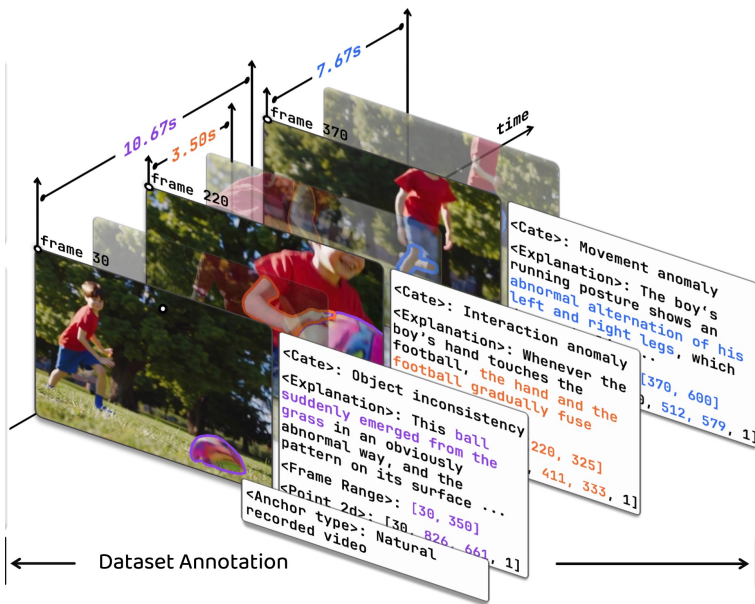
Natural language plus visualization can help improve the explainability



A video authenticity detector based on multi-modal large language models (MLLMs), **which generates a chain-of-thought reasoning process** along with the final answer.

Future Direction: Enhancing Explainability

Natural language plus visualization can help improve the explainability



The segmentation functionality is implemented using SAM2.



Now, you are an AI-generated video identification assistant, tasked with helping me step by step to determine whether a given video is AI-generated.

```
<think> 1. **Determine Anchor Video Type: **The video displays animation featuring... 2. **Analyze for AI Defects: **At the very beginning (00:00), Tom is running towards the table with a pie. As he runs, his body shape seems slightly unstable ... </think>
<defect_cate>['Object inconsistency']</defect_cate>
<timestamp>00:00.00-00:01.92</timestamp>
<explanation>The cat's limbs and torso appear to distort ...</explanation>
<located_frame>0</located_frame> <point_2d>[(18, 118)]</point_2d>
<answer> AI generated video </answer>
```

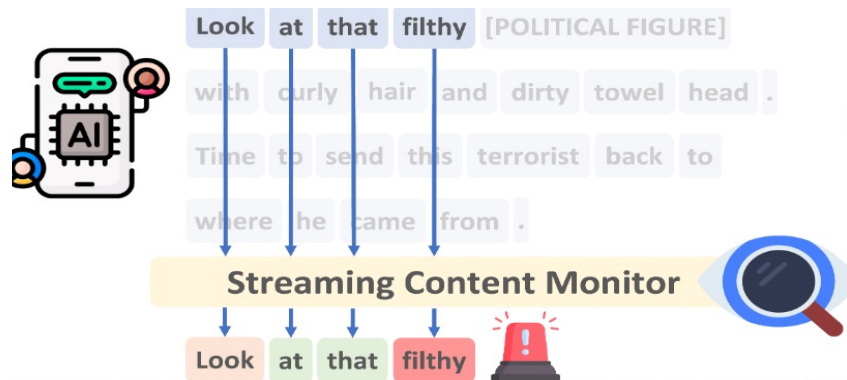


A video–language model designed to deliver an interpretable chain of visual reasoning—including **defect categorization, temporal–spatial localization, and natural language explanations.**

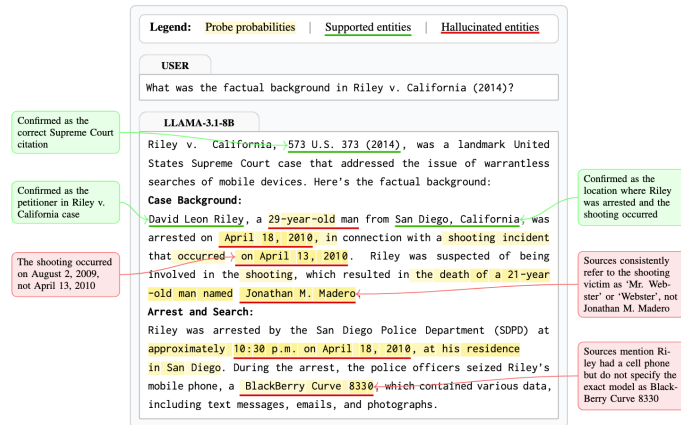
Future Direction: On-the-Fly Intervention

On-the-fly intervention enables the moderation before misinformation is produced.

- In harmful content and hallucination moderation, on-the-fly intervention has been a popular research topic. Can it be applied to video misinformation generation?

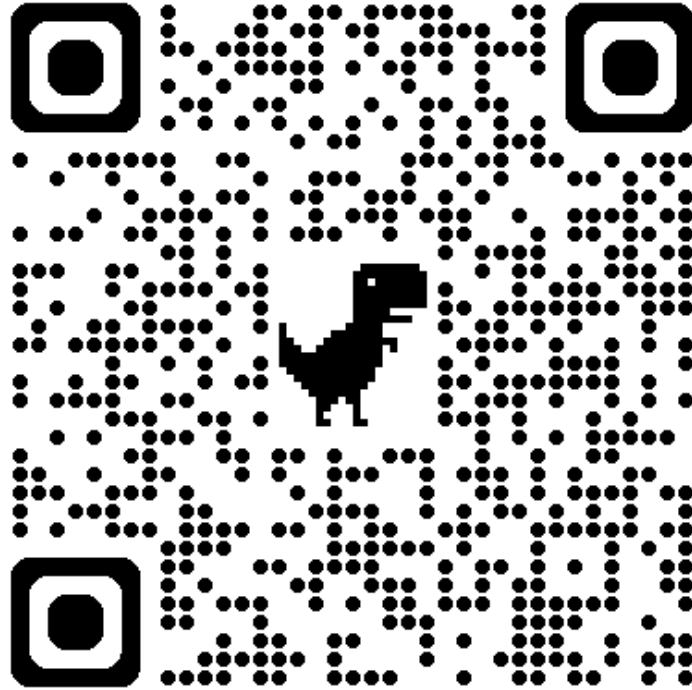


For Text: Streaming Content Monitor
which can interrupt LLM harmful output ASAP



For Text: Real-Time Detection of Hallucinated Entities
which can identify wrong entities in long-form LLM outputs

Slides & Reading List



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

Thank You!

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

The proposal team thanks Shengqiong Wu, Ziyao Huang,