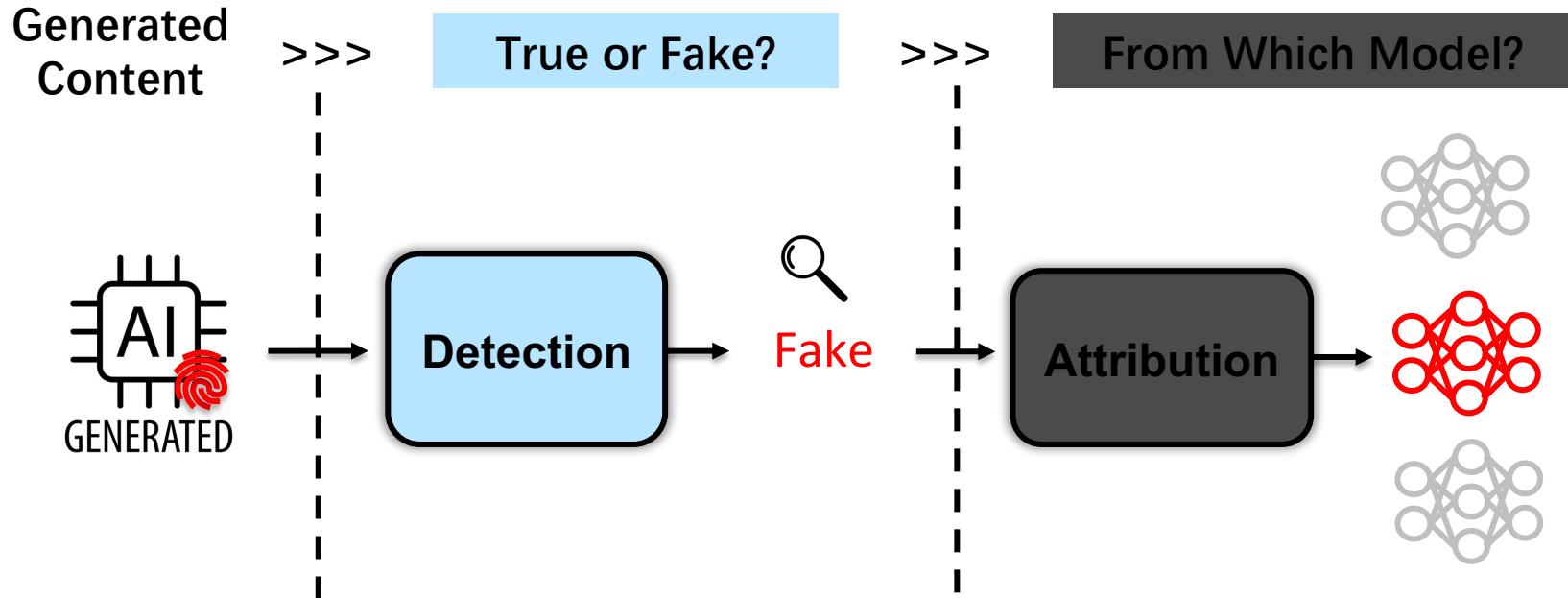


# Detection Part II: AI-Generated Misinformation

Section 5

# Combating AI Generated Videos: From Detection to Attribution



- **Part1:** Manipulated Video Detection
- **Part2:** Generated Video Detection
- **Part3:** Attributing AI-generated Video the Source Model

# Tutorial Outline

---

## Detection Part II: AI-Generated Misinformation

Manipulated video detection

Generated video detection

Attributing AI-generated Content to the Source Model

Q+A/Discussion

# Manipulated Video Detection

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

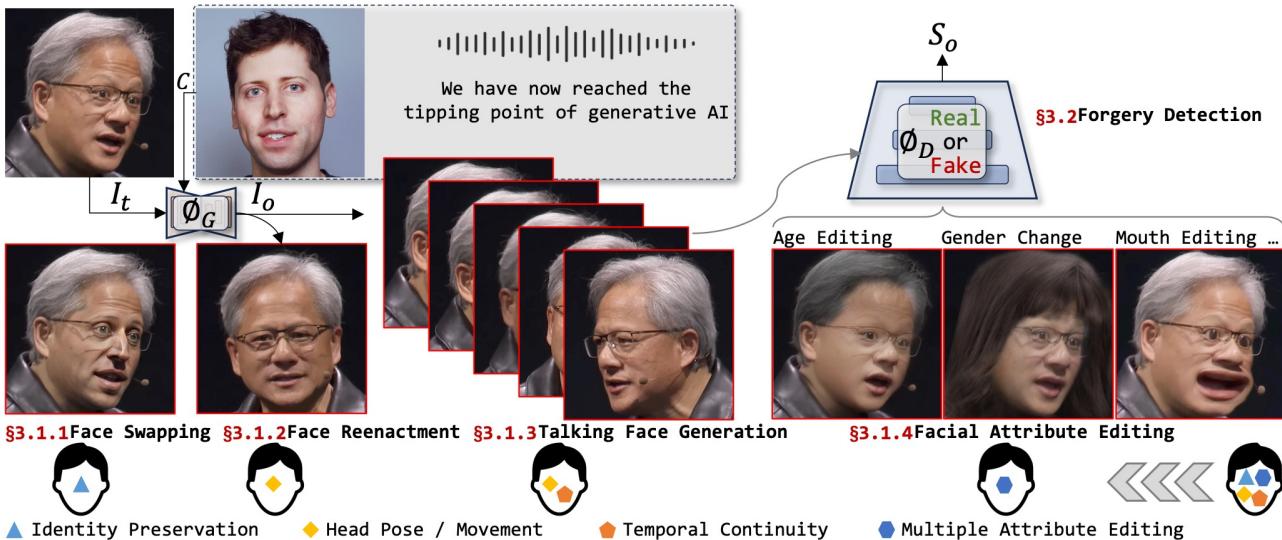
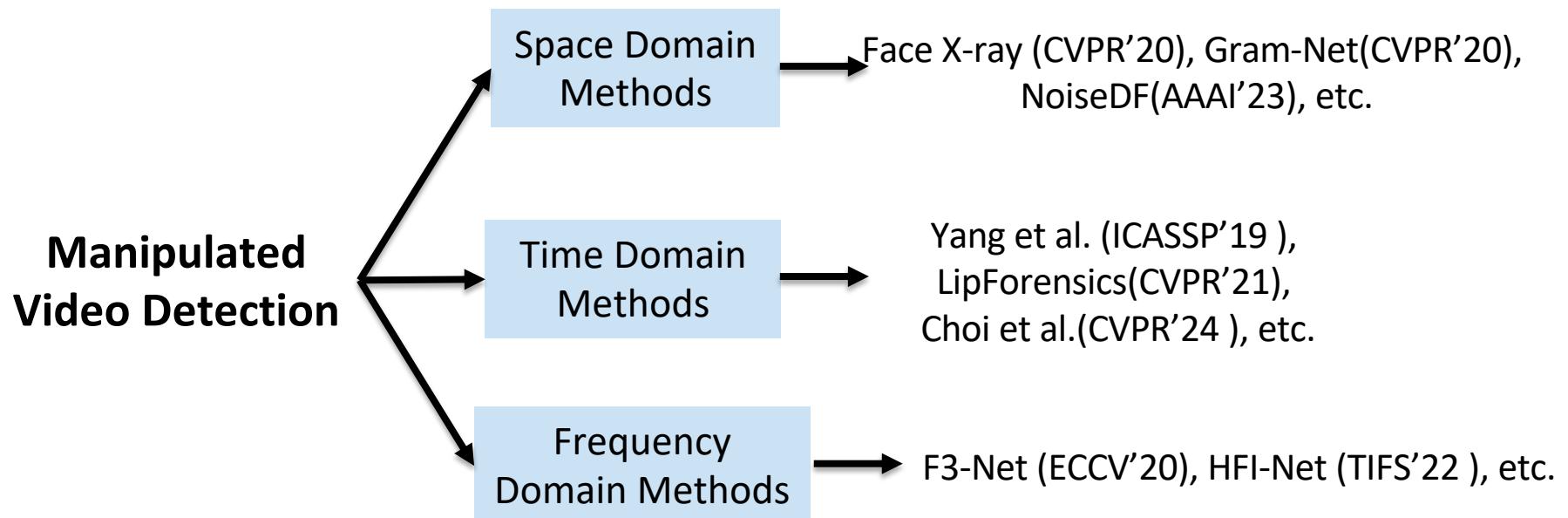


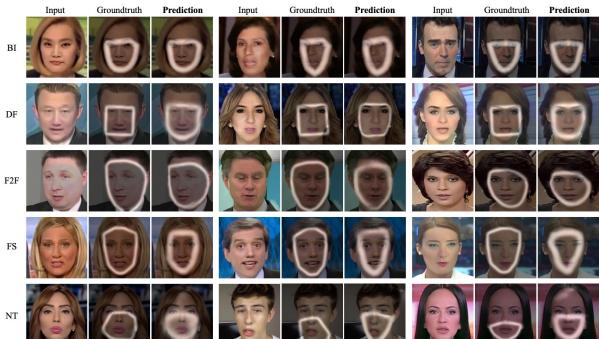
Fig. 2: Top: Illustration of different deepfake generation (Sec. 3.1) and detection tasks (Sec. 3.2) that are discussed in this survey. Bottom: Specific facial attribute modification of each task. Data from [NVIDIA Keynote at COMPUTEX 2023](#) at 29:40.

# Taxonomy



# Space Domain Inconsistency

## Boundary Artifacts (Face X-ray)



## Texture Inconsistencies (Gram-Net)

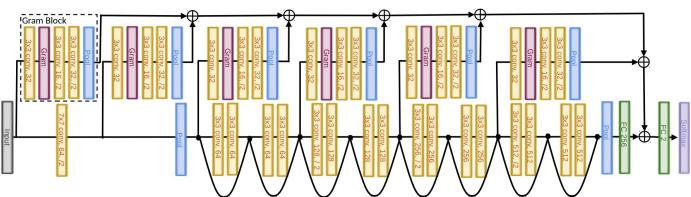
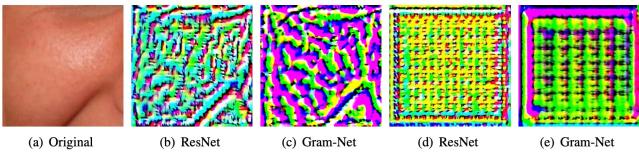
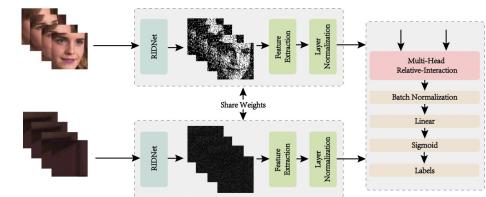


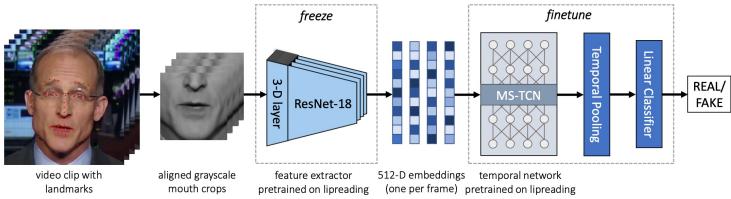
Figure 5. Gram-Net architecture. We extract global image texture feature with 6 Gram Blocks in different semantic levels from ResNet.  $\oplus$  means concatenation.

## Noise Inconsistencies (NoiseDF)

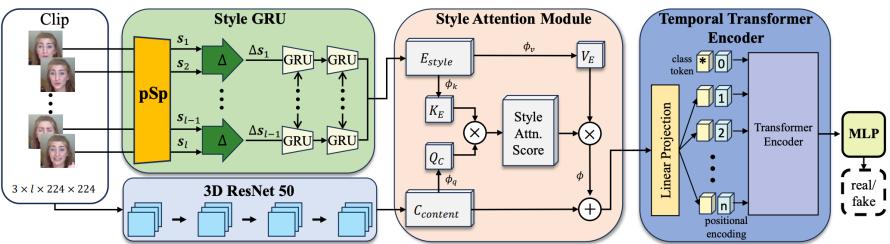


# Time Domain Inconsistency

## Inconsistency of mouth movements (LipForensics)

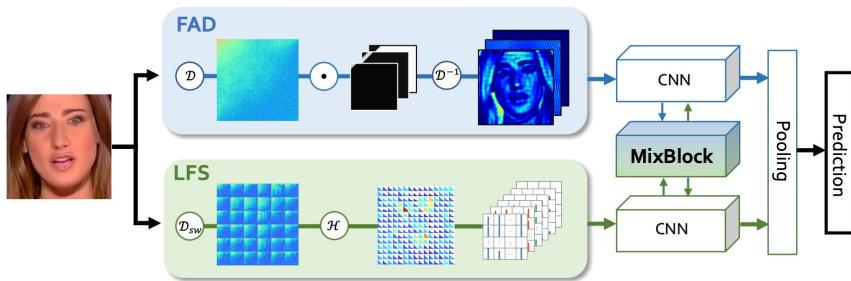


## Inconsistency of the style latent vectors (Choi et al.)



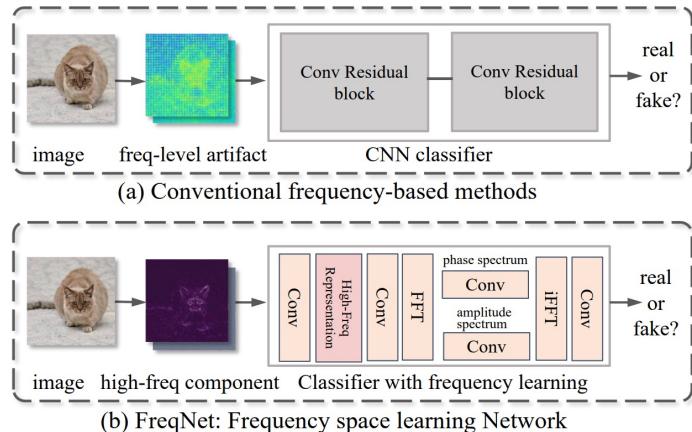
# Frequency Domain Artifacts

**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

# Tutorial Outline

---

## Detection Part II: AI-Generated Misinformation

Manipulated video detection

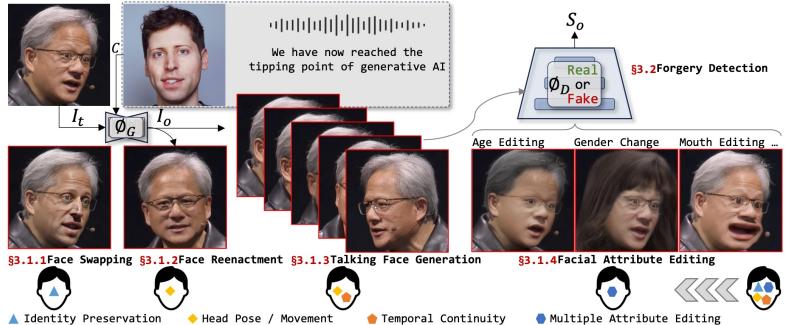
Generated video detection

Attributing AI-generated Content to the Source Model

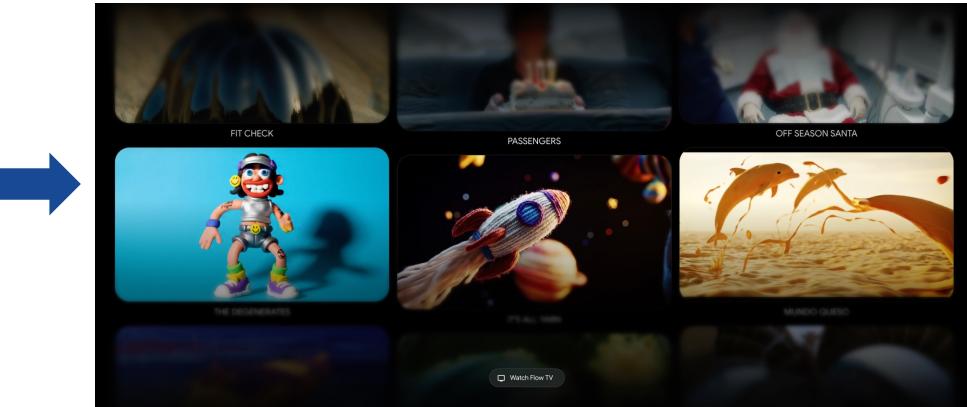
Q+A/Discussion

# From Manipulated to Generated Video Detection

## Manipulated Video Detection



## Generated Video Detection



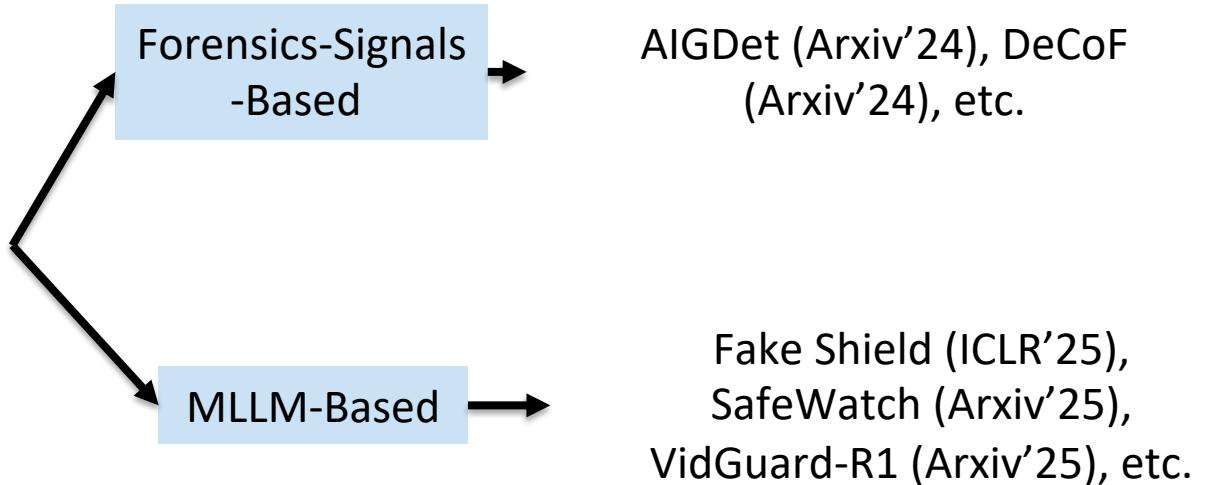
(Generated by Veo3)

Face-centric content

More diverse, real-world videos

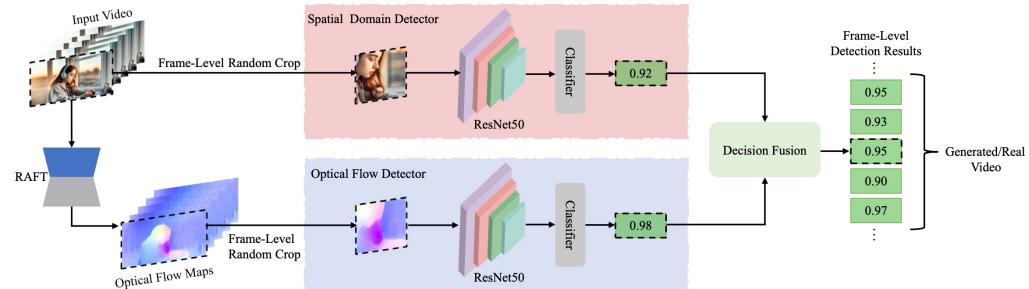
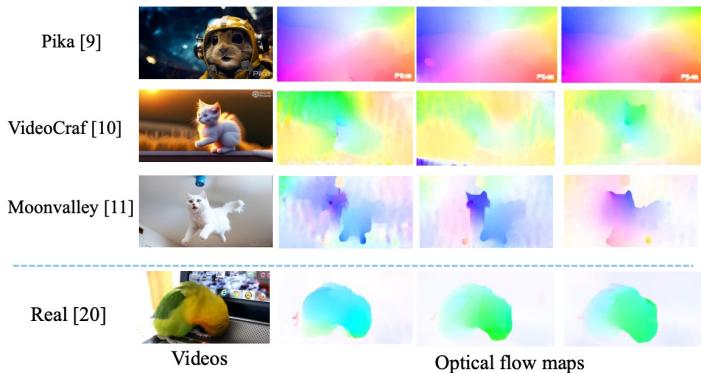
# Taxonomy

## Generated Video Detection



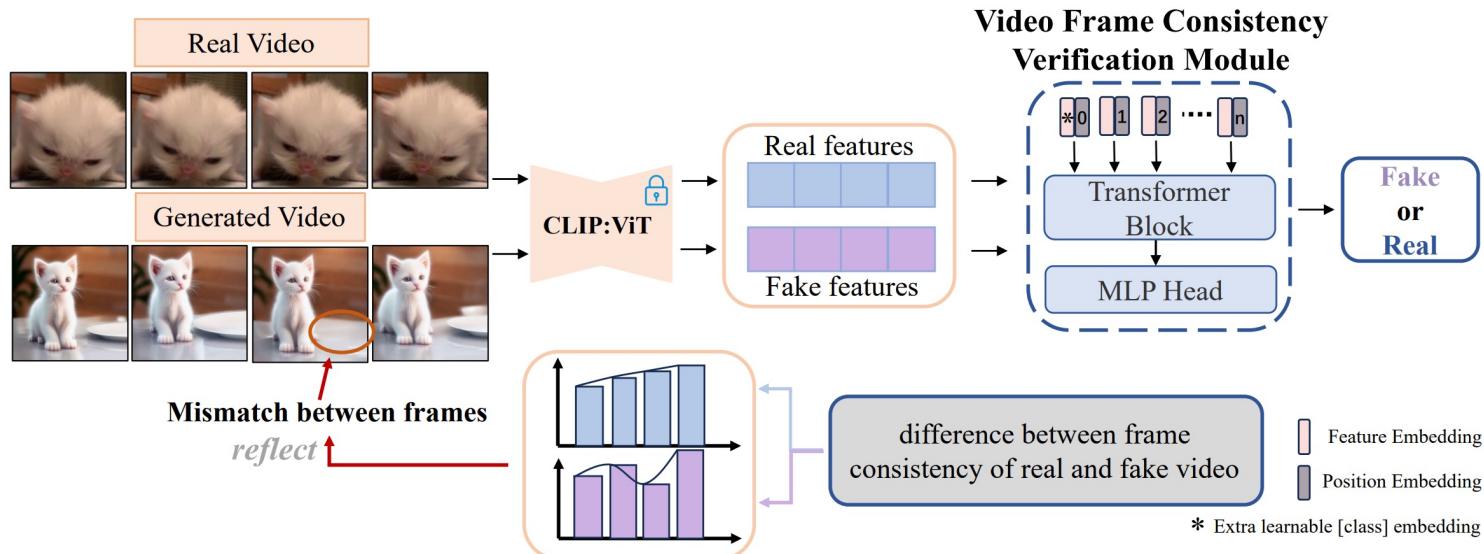
# AIGDet: Detect by Spatio-Temporal Anomaly

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



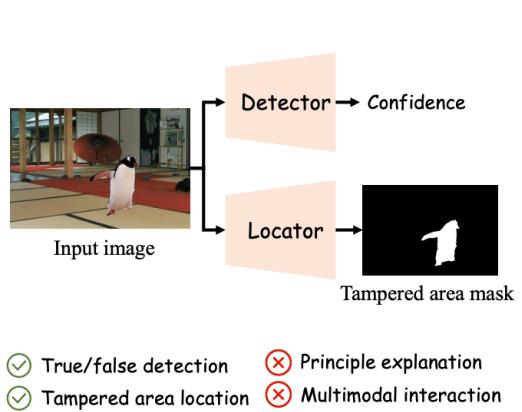
# DeCoF: Detect by Frame Consistency

- Detecting generated videos by capturing mismatch between frames

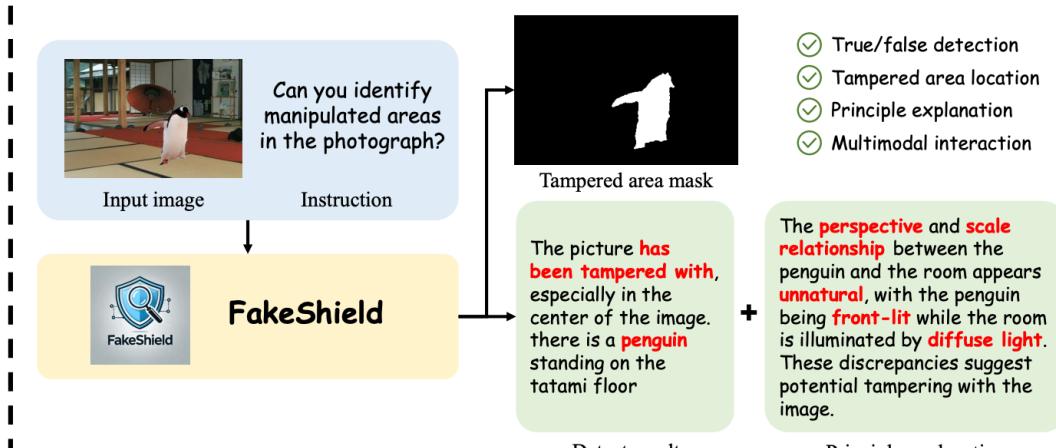


# Fake Shield: Explainable Detection via MLLMs

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



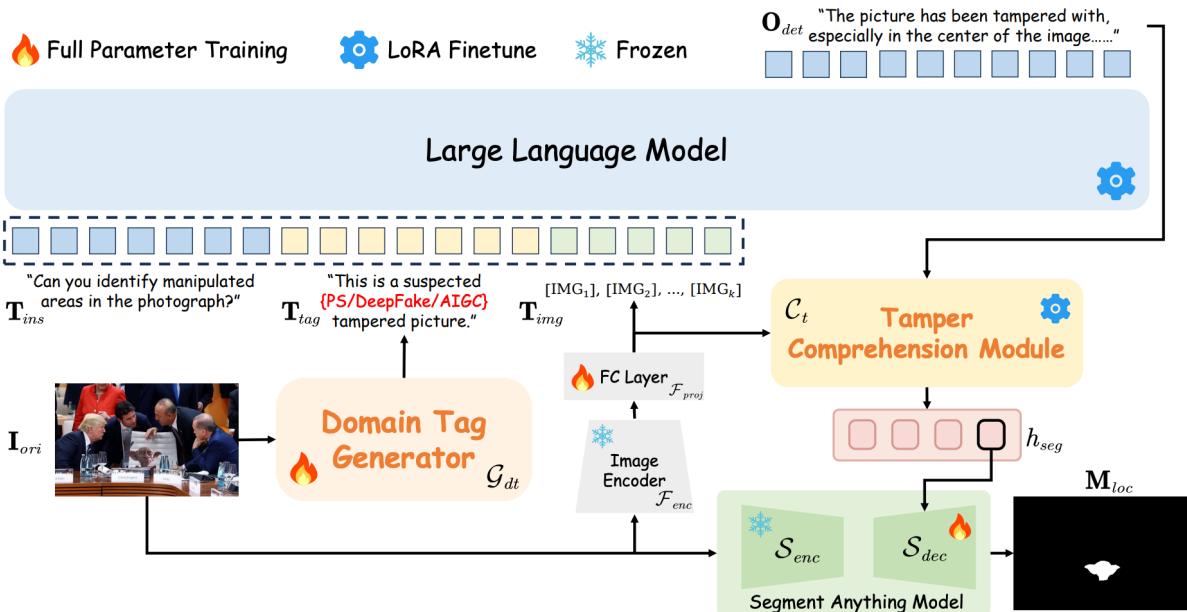
(a) Conventional IFDL Framework



(b) Our Explainable IFDL Framework: FakeShield

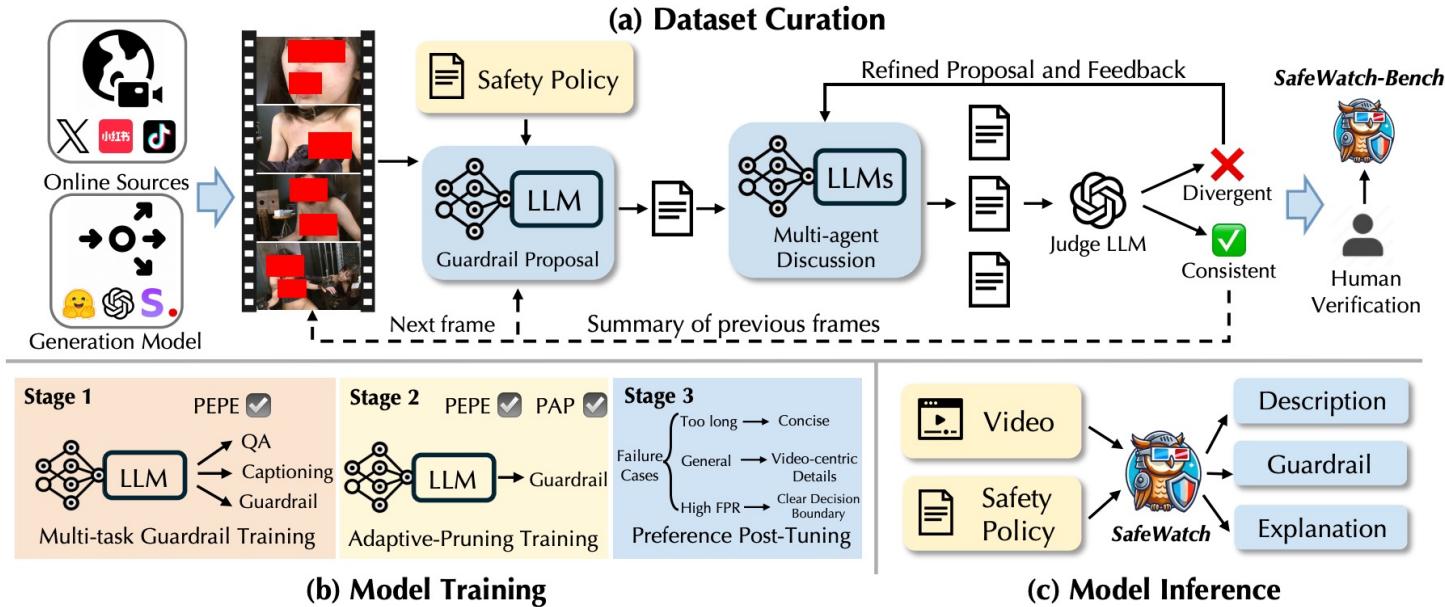
# Fake Shield: Explainable Detection via MLLMs

- 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



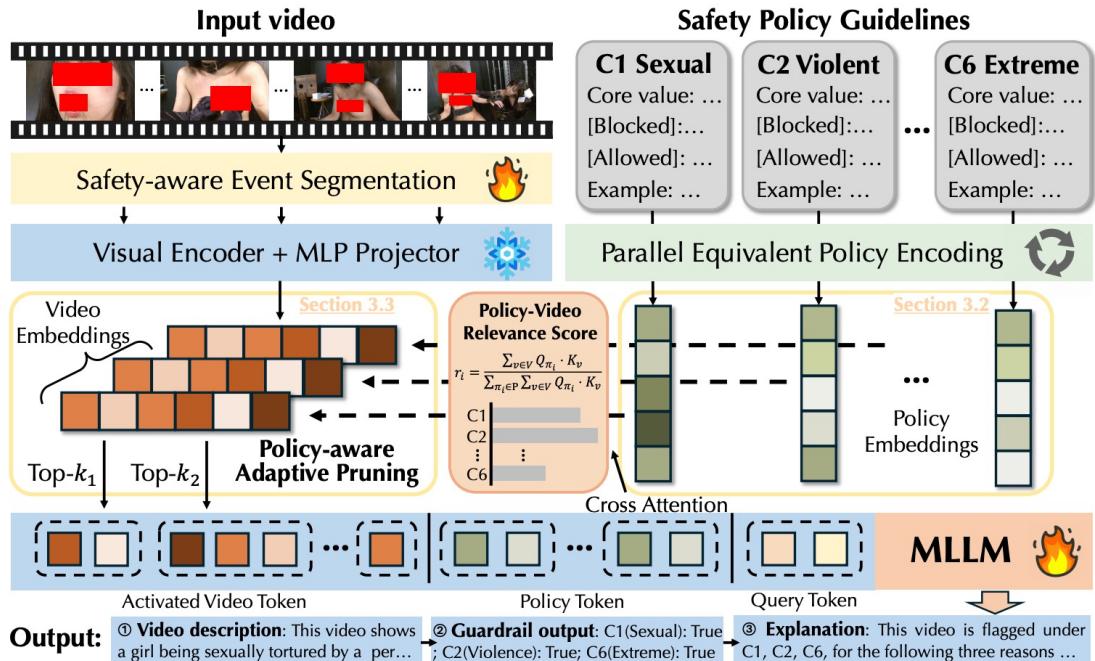
# SafeWatch: Improving Safety-Policy Following

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



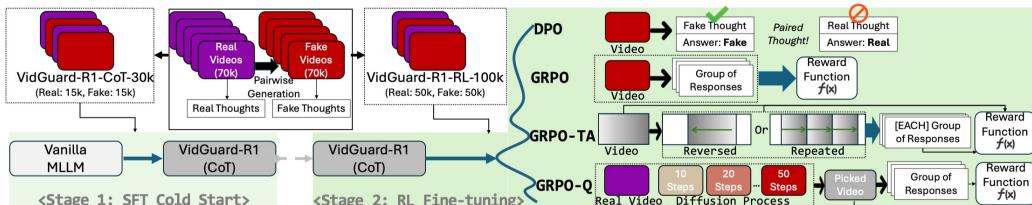
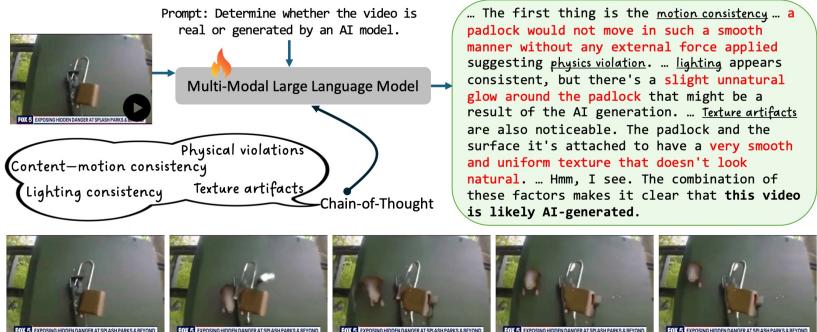
# SafeWatch: Improving Safety-Policy Following

- Decoding pipeline of SafeWatch



# VIDGUARD-R1

- Applies online reinforcement learning for more generalized fake detection



# Tutorial Outline

---

## Detection Part II: AI-Generated Misinformation

Manipulated video detection

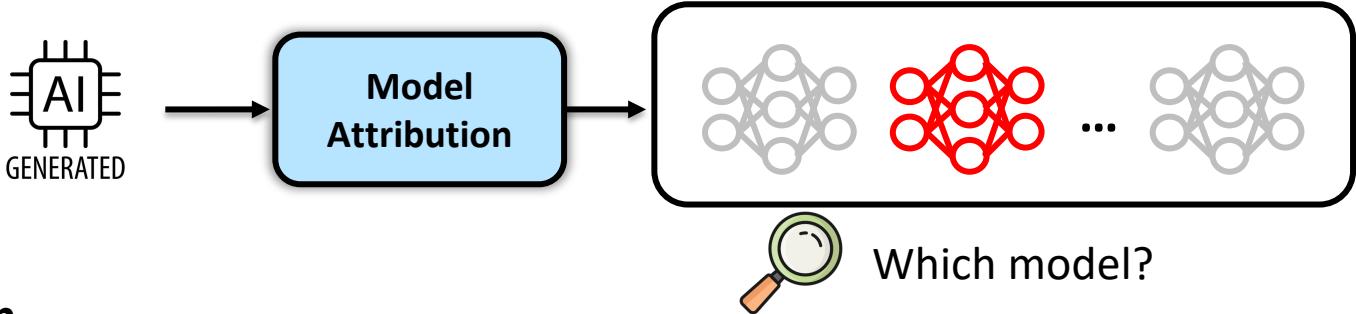
Generated video detection

Attributing AI-generated Content to the Source Model

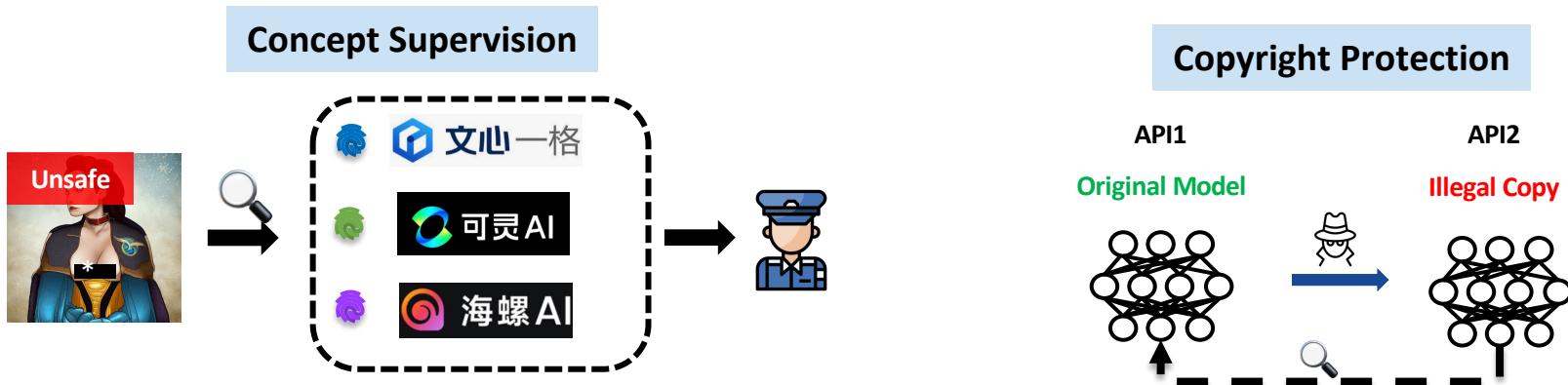
Q+A/Discussion

# Model Attribution

- Identify the source model of generated contents



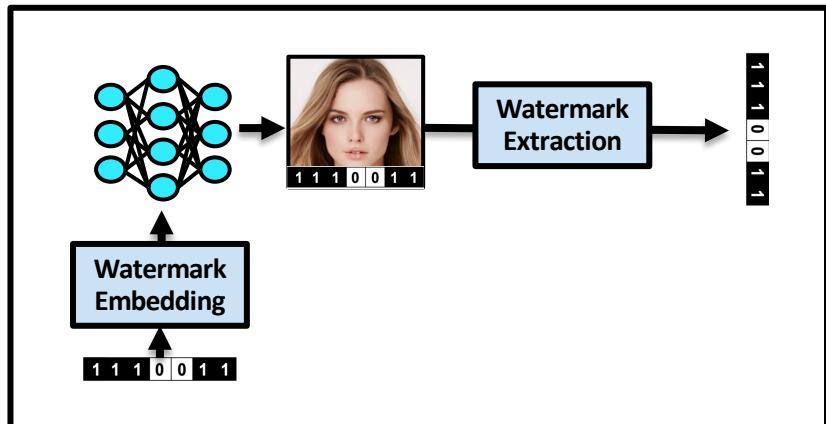
- Application



# Two Technique Routes for Model Attribution

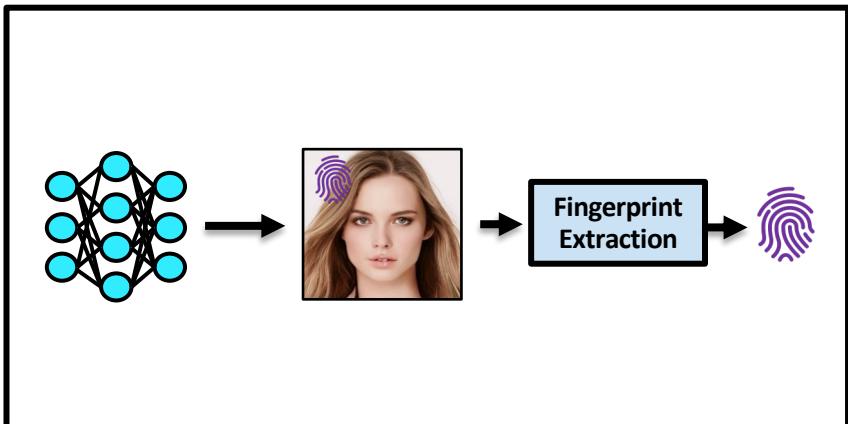
## Watermarking-Based

(Artificial Fingerprint)



## Model-Fingerprint-Based

(Intrinsic Fingerprint)



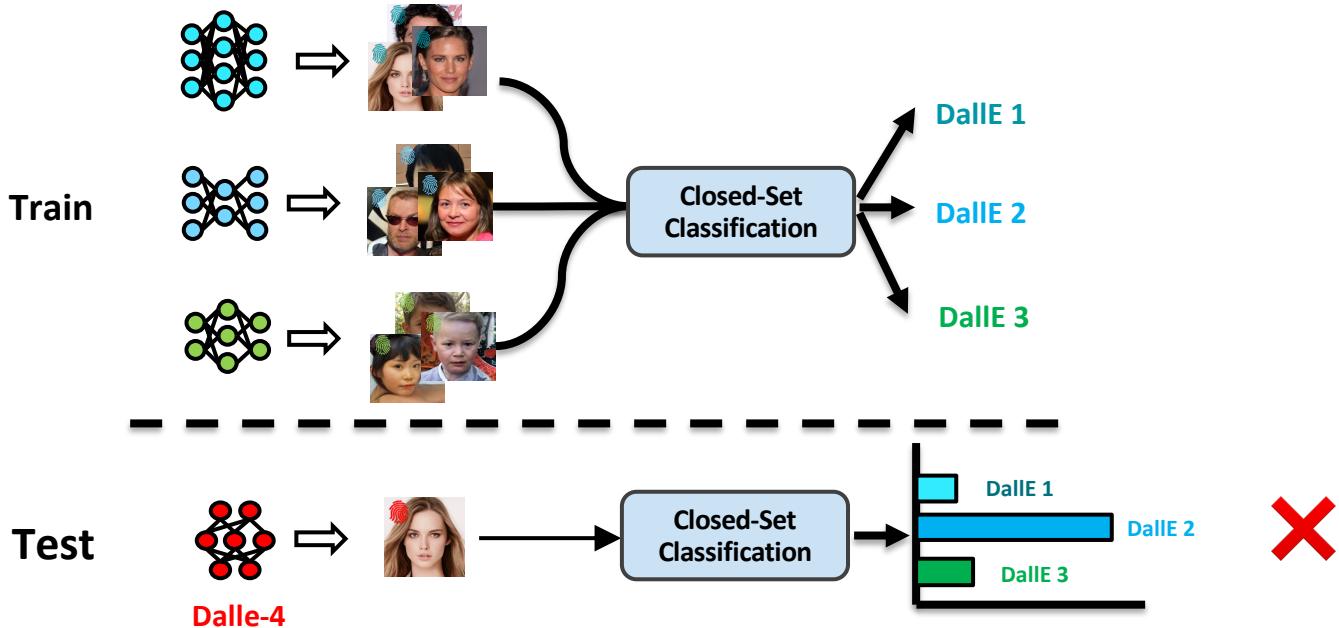
Need modify model in advance



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<sup>[1,2,3]</sup>

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

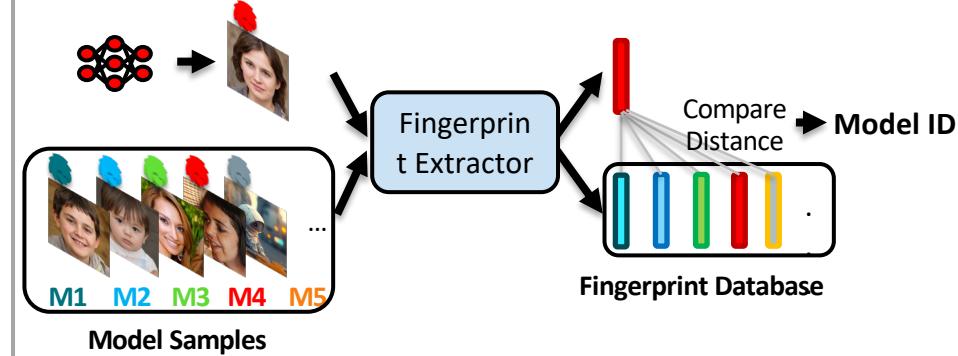
[3] Marra F, Gragnaniello D, Verdoliva L, et al. Do gans leave artificial fingerprints?

# Closed-Set → Open-Set Framework

## Closed-Set Framework



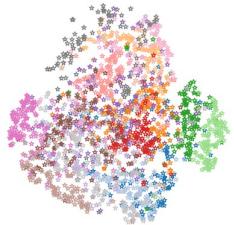
## Open-Set Framework



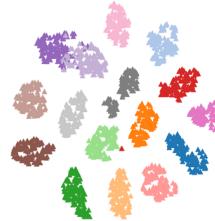
Can only attribute seen models

Can dynamically extend to unseen models not involved in training

# Collect Model Data -> Synthesize Model Data



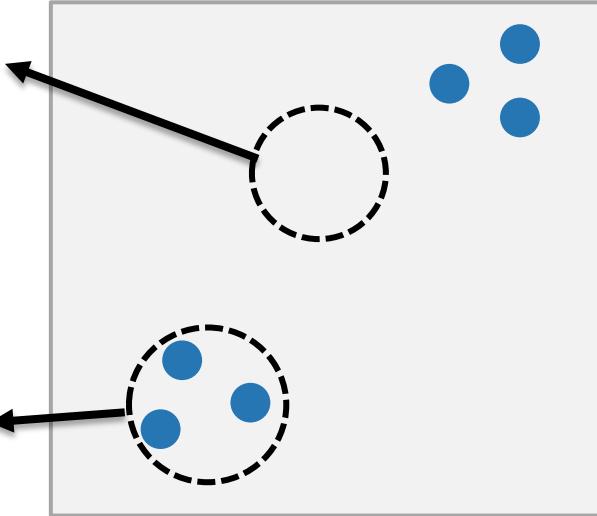
★ ADM  
 ★ DDPM  
 ★ LDM  
 ★ PDNM  
 ★ SD1.5  
 ★ SD2.1  
 ★ SDXL  
 ★ Midjourney



▲ ProGAN\_CelebA  
 ▲ MMDGAN\_CelebA  
 ▲ SNGAN\_CelebA  
 ▲ StarGAN\_CelebA  
 ▲ StyleGAN3\_r\_FFHQ  
 ▲ StyleGAN3\_t\_FFHQ  
 ▲ VanillaVAE  
 ▲ BetaVAE

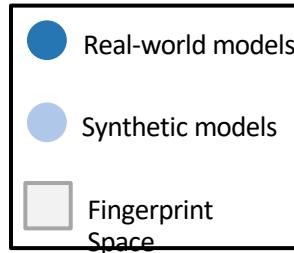
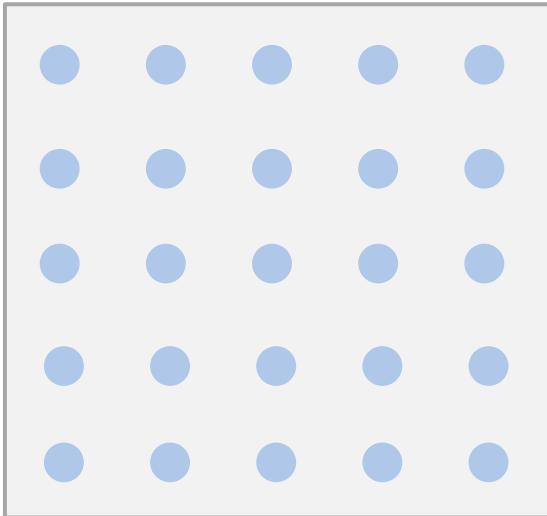
▲ AttnGAN\_CelebA  
 ▲ ProGAN\_CelebAHQ  
 ▲ StyleGAN\_CelebAHQ  
 ▲ StyleGAN2\_FFHQ  
 ▲ DisentangledBetaVAE  
 ▲ InfoVAE  
 ▲ ResFlow  
 ▲ Glow

## Model Collection (100+ in existing dataset)



Can only cover limit model fingerprint space

## Our Method: Model Synthesis (Unlimited in principle)

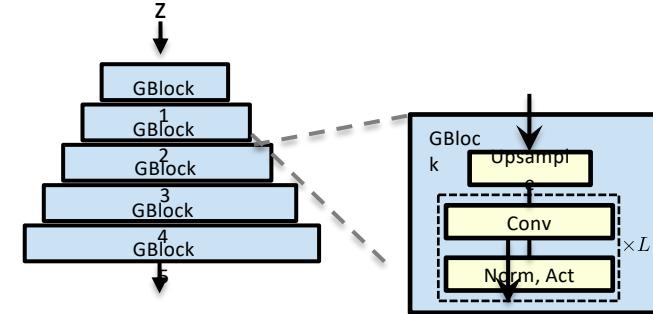


Cover more complete fingerprint space

# Which Factors Influence Model Fingerprint?

## Observation 1:

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

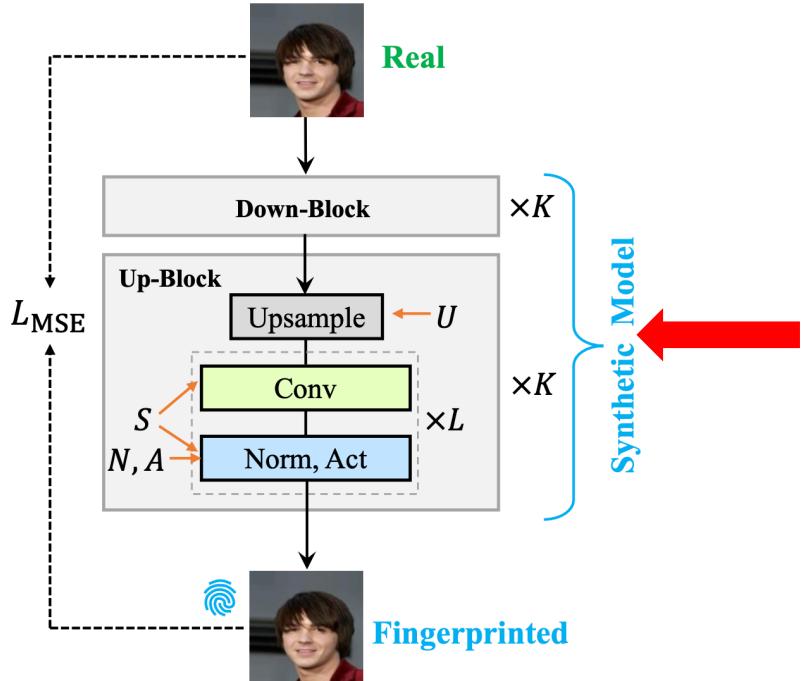


## Observation 2:

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

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

# Synthetic Model Construction

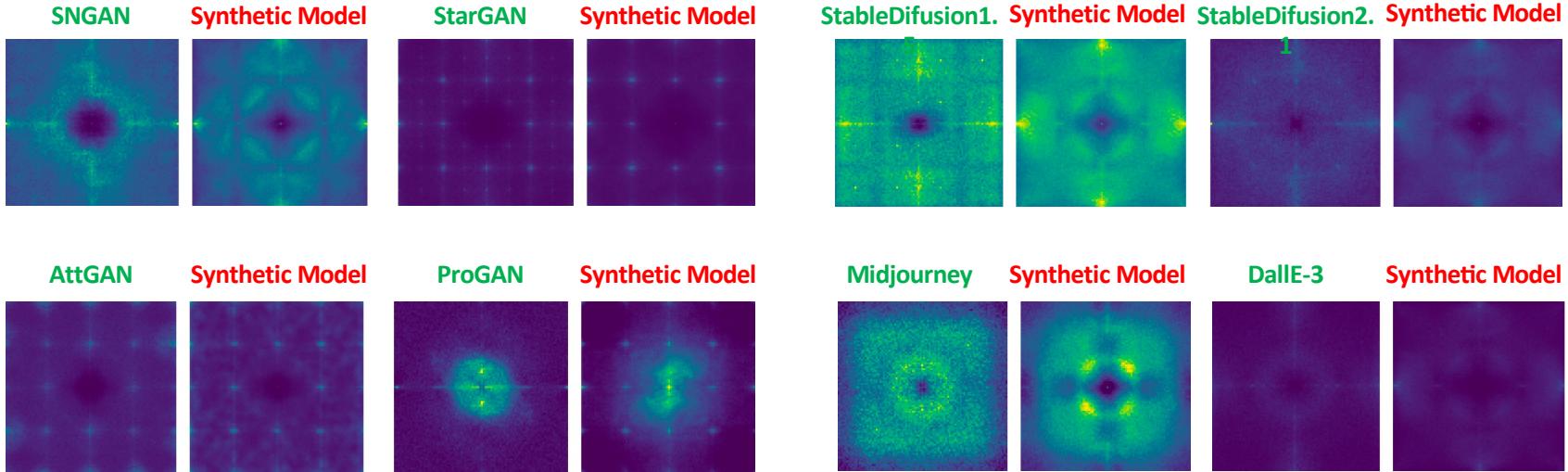


- $L$  (number of convolution layers): 1, 2
- $S$  (the order of activation and normalization relative to the convolution layer): before/after the convolution
- $U$  (upsampling layer): deconvolution, bilinear, nearest
- $A$  (activation function): None, ReLU, Tanh, Sigmoid
- $N$  (normalization layer):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

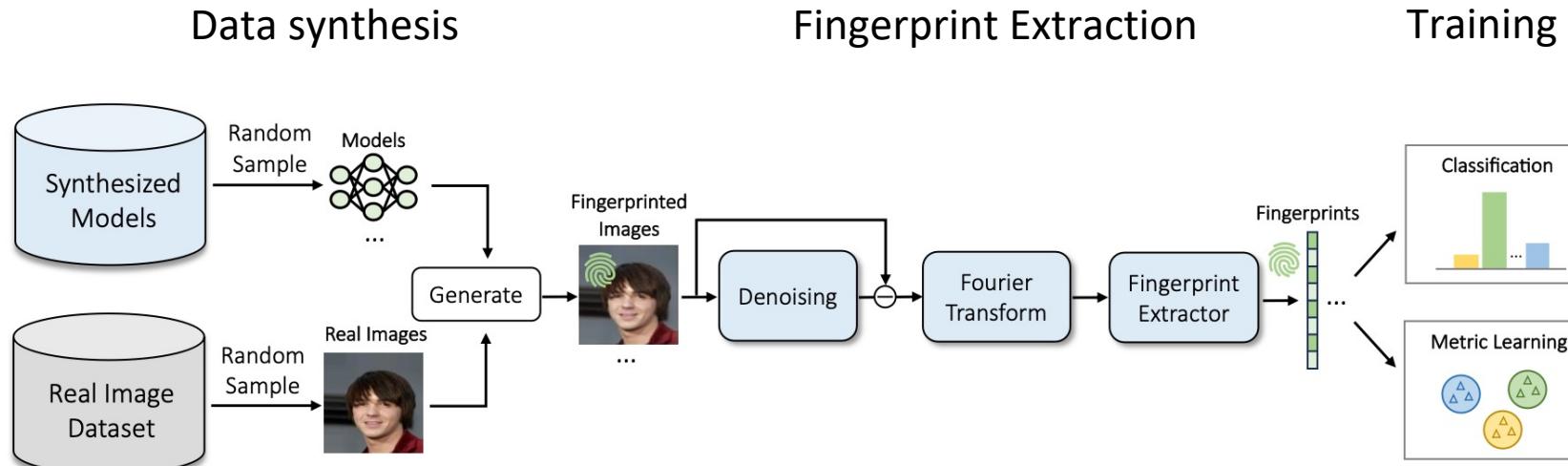
**A total of 5,760 synthetic models**

Synthesizing one model only takes **1.5 minutes**

# Generation Spectrum Comparation

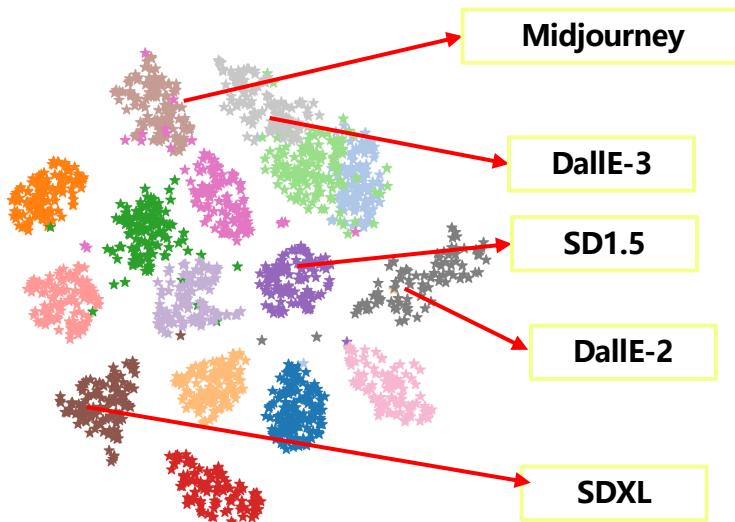


# Overview of Open-World Framework



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

# Evaluation Result



**Model attribution accuracy:  
93.20%**

# Cases

DallE-3



An armchair in the shape of an avocado

DallE-3



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

DallE-3



An armchair in the shape of an avocado

DallE-2



An armchair in the shape of an avocado

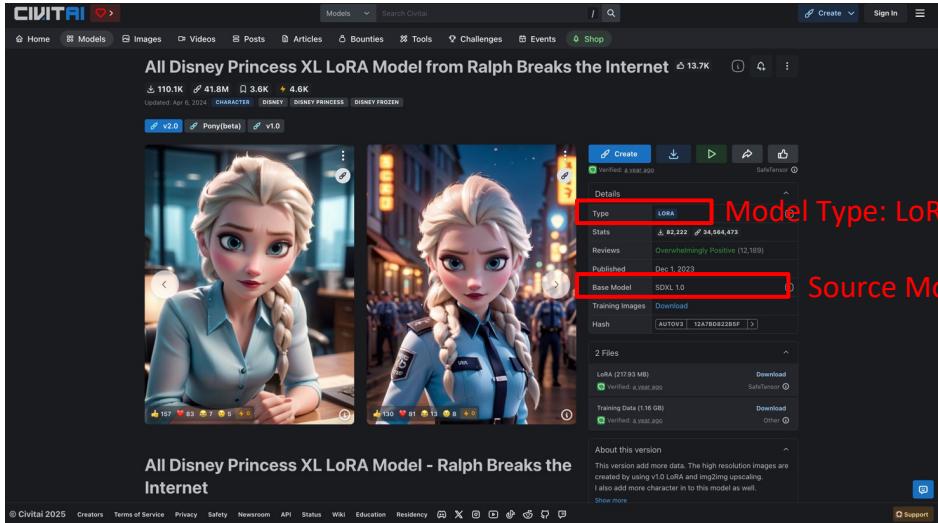
Fingerprint Similarity: **0.9045**

Prediction Result: **Same Model**

Fingerprint Similarity: **0.7192**

Prediction Result: **Different Model**

# Application: Model Relationship Analysis

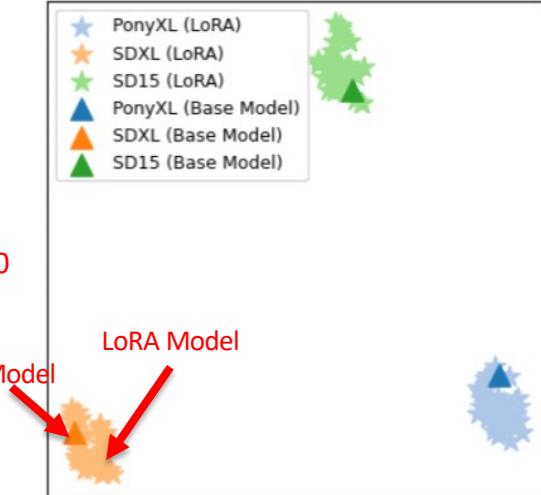


All Disney Princess XL LoRA Model from Ralph Breaks the Internet

Model Type: LoRA

Source Model: SDXL1.0

The screenshot shows a LoRA model for Elsa from Frozen. The interface displays various details about the model, including its type (LoRA), source model (SDXL1.0), and training data. Two images of Elsa are shown: one from the original Frozen movie and one from the movie Ralph Breaks the Internet.



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

# Recap of This Part

- **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

# Tutorial Outline

## 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

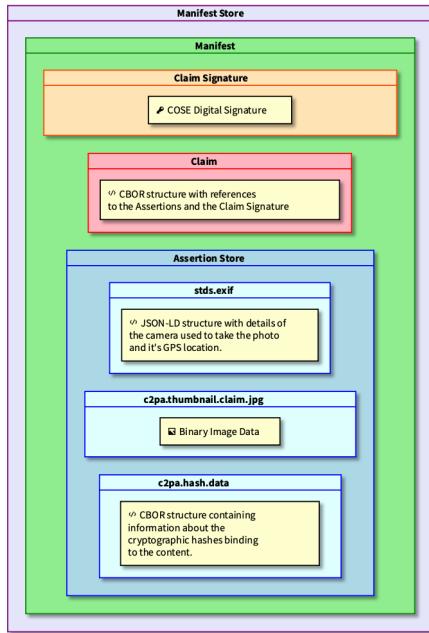
#### Controlling the Spread of Misinformation

#### Promoting Truth and Debunking

# Embedding Tamper-proof Digital Identifier

## ❖ Developing Technical Specification

The Coalition for Content Provenance and Authenticity (C2PA)



meta-data

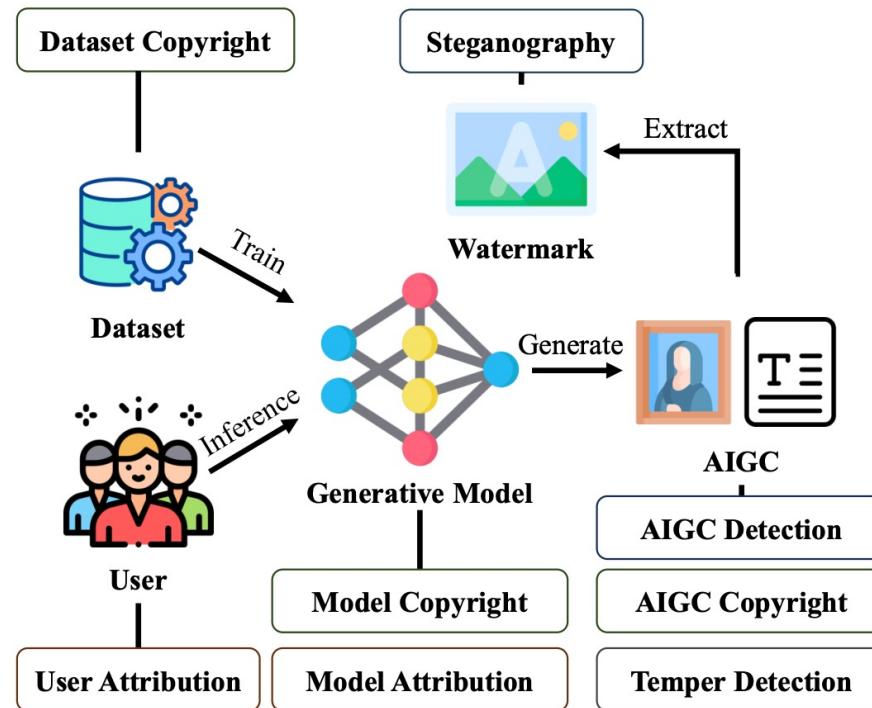


editing history



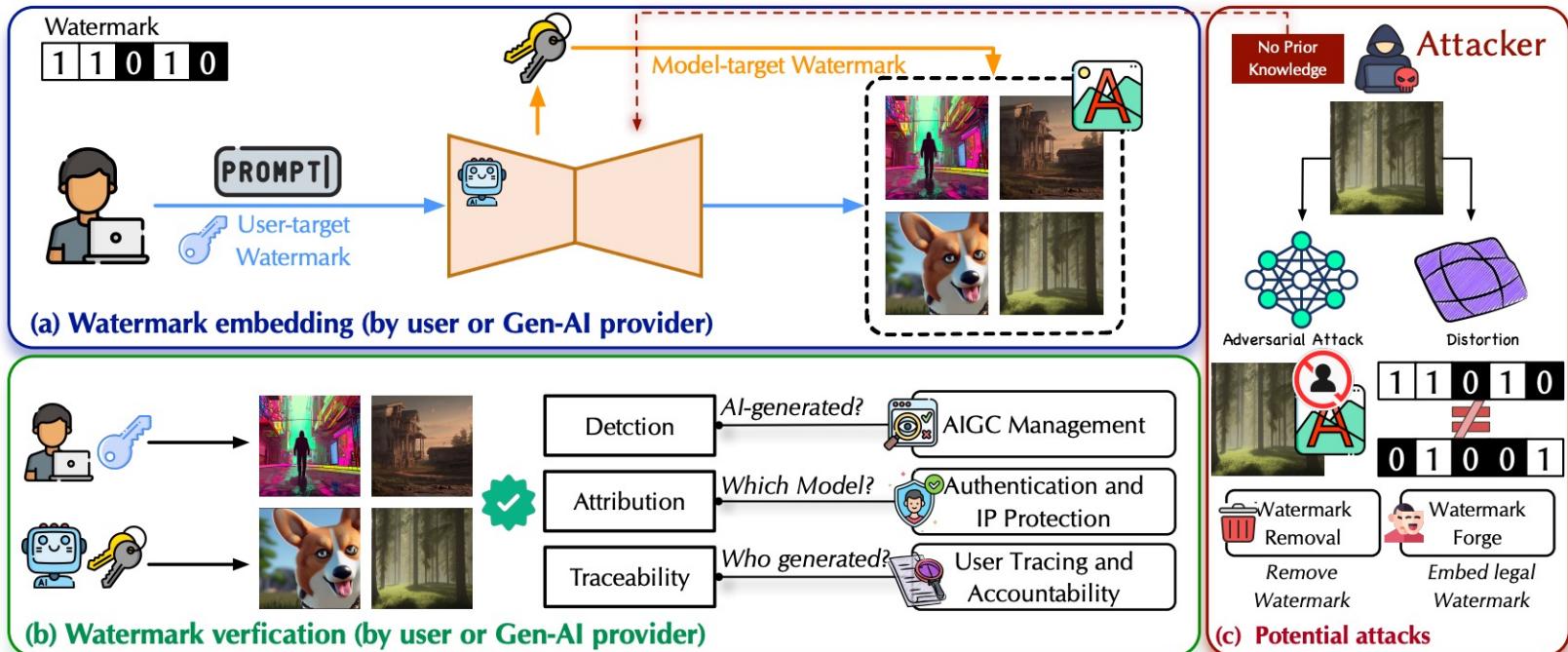
# Embedding Tamper-proof Digital Identifier

## ❖ Embedding Invisible Watermarks



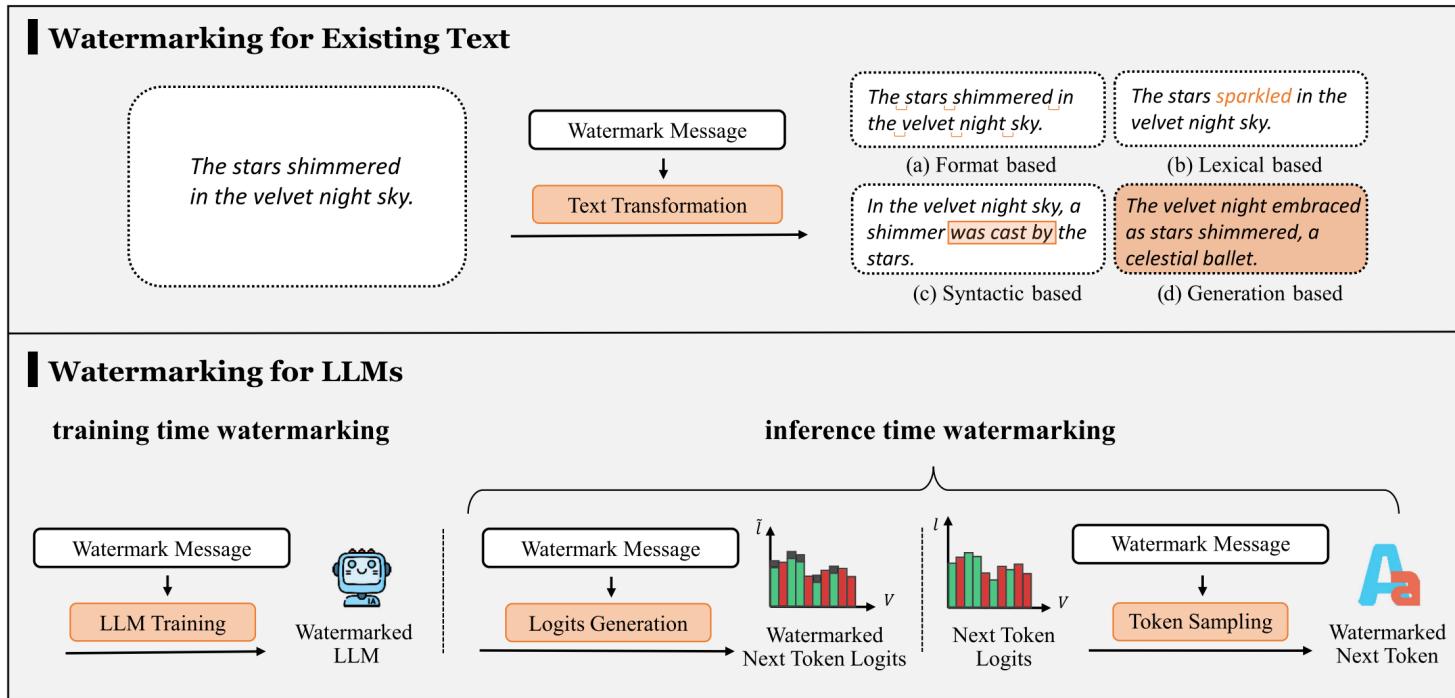
# Embedding Tamper-proof Digital Identifier

## ❖ Embedding Invisible Watermarks



# Embedding Tamper-proof Digital Identifier

## ❖ Embedding Invisible Watermarks



# Tutorial Outline

## 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.

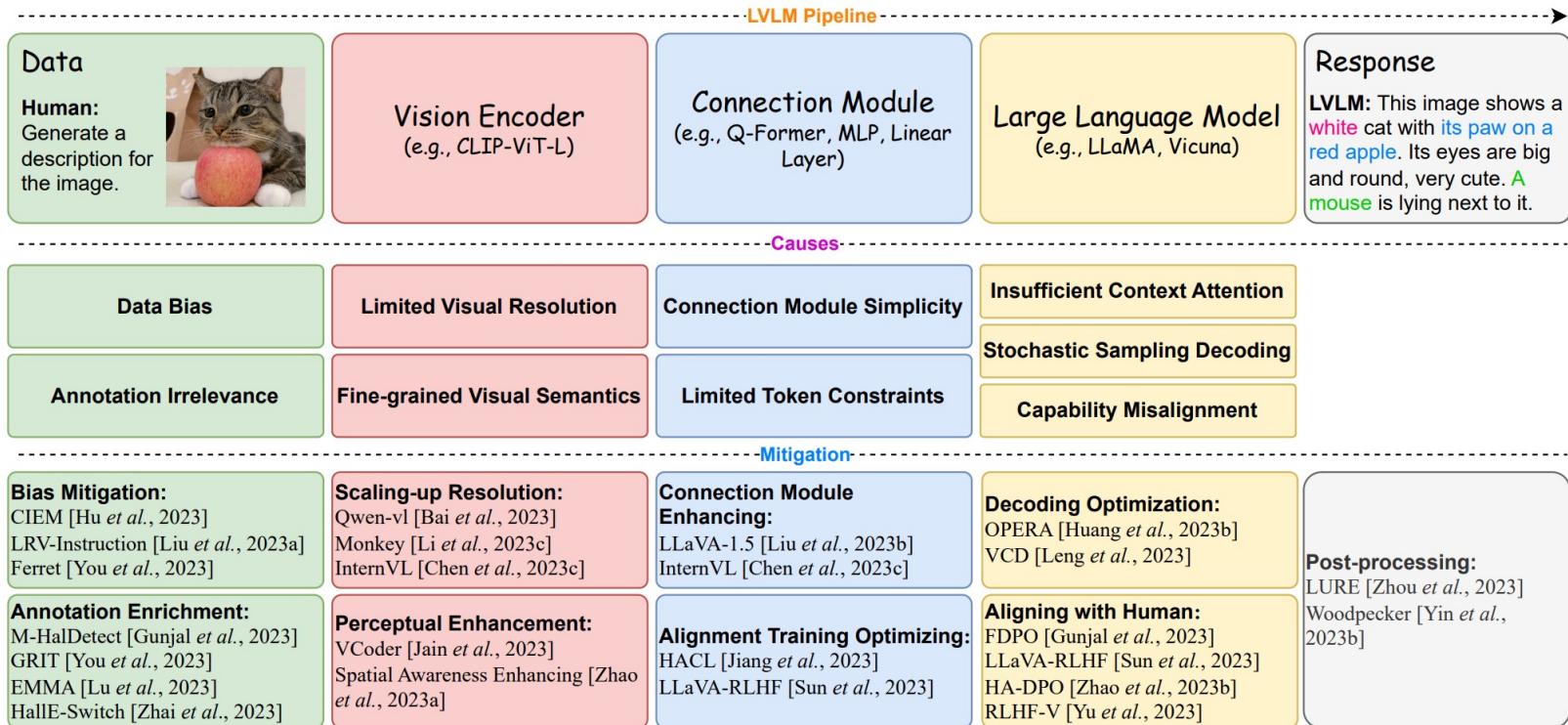
### Counter-Common-Sense Hallucination

**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



# Tutorial Outline

## 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

# Alerting, Verification, and Resilience Building

## 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

Search fact checks about a topic or person      Search by image

More results in other languages      Language filter: English

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**

What would you like to verify today?      How the AI works

Here is my analysis:      Justin Trudeau has resigned as the leader of the Liberal party

**Reliability score**  
How is this calculated?

Reliability: 95% (4)

The claim is highly reliable,  
**you can share with your (5) 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 (6)  
80%

Sources: 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." <https://www.npr.org/2025/01/06/nx-8...>

2. "After a Surprise Resignation, What Comes Next for Canada?" <https://www.csis.org/analysis/after...>

3. "Credibility: 93% Canada's Justin Trudeau to resign as Liberal Party leader and prime ..." <https://www.csis.org/analysis/after...>

4. "Minister Justin Trudeau announced plans to resign on Monday after loud calls from within his own party to step down." <https://www.npr.org/2025/01/06/nx-8...>

5. "The claim is highly reliable, you can share with your (5) network." [https://www.veracity.ai/analyze/justin-trudeau-has-resigned-as-the-leader-of-the-liberal-party](#)

6. "Avg. Source Credibility: 12 Sources (6) 80%" [https://www.veracity.ai/analyze/justin-trudeau-has-resigned-as-the-leader-of-the-liberal-party](#)

Start typing or paste text here...      Responses related to Veracity may not always be accurate.

# Alerting, Verification, and Resilience Building

## ❖ 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.



**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



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.

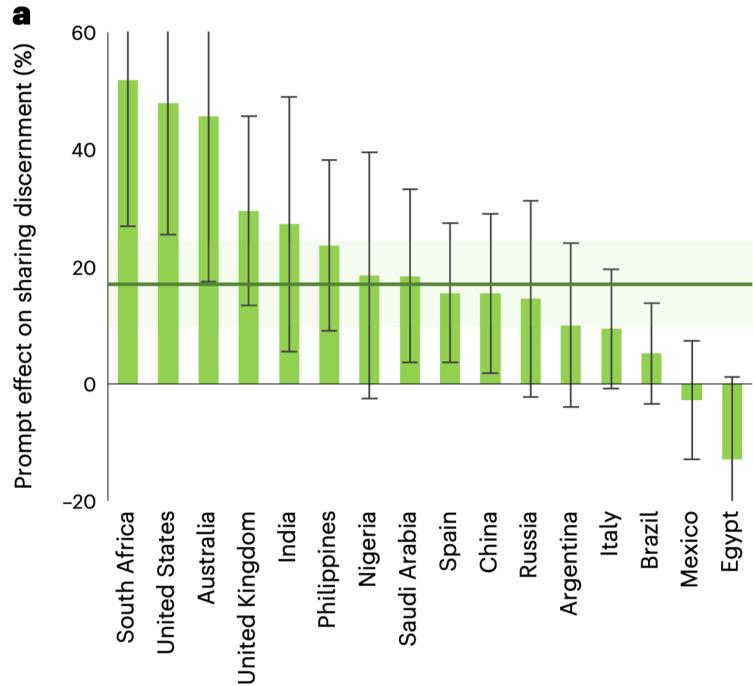
Counterfactual-Explanation

	Accuracy Rate		Flip Rate (Short-term/ Long-term)					Overall
	Pre-test	Post-test/ Long-term	X → ✓	▲ → ✓	▲ → X	✓ → X		
<b>Fake Claims with intervention in reading environment</b>								
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%	
<b>Fake Claims without intervention in reading environment</b>								
Control	40%	38% / 38%	4% / 7%	8% / 8%	7% / 8%	7% / 10%	26% / 32%	

# Alerting, Verification, and Resilience Building

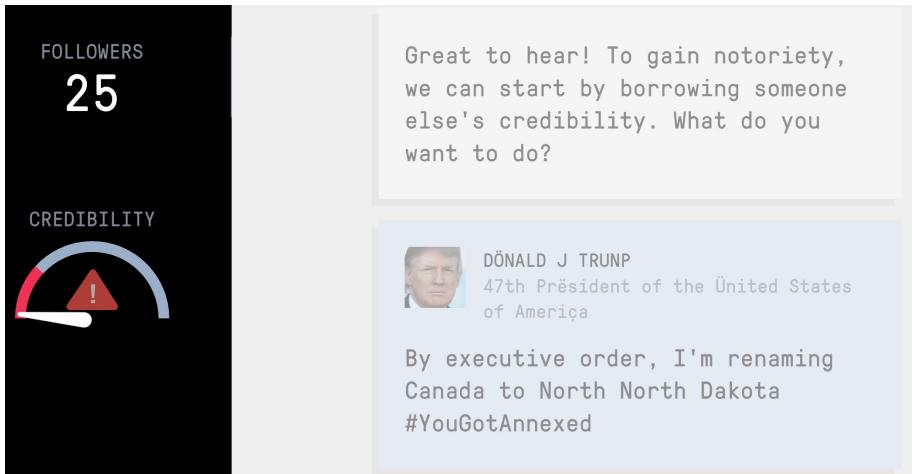
## ❖ Platform Alerts

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

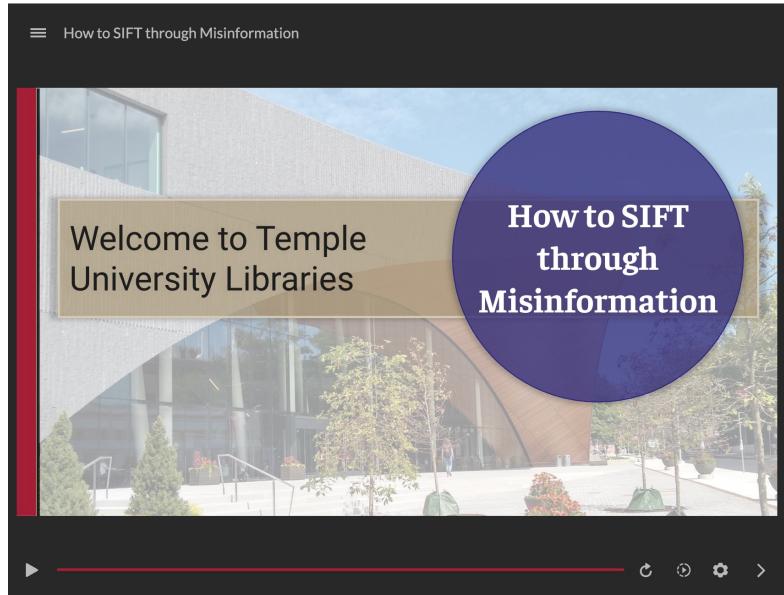


# Alerting, Verification, and Resilience Building

## ❖ User Resilience Building



Bad News Game



# Tutorial Outline

## 6. Prevention Strategies

Creation Prevention

Embedding Tamper-proof Digital Identifier

Mitigating Hallucination in Content Generation

Spread Prevention

Alerting, Verification, and Resilience Building

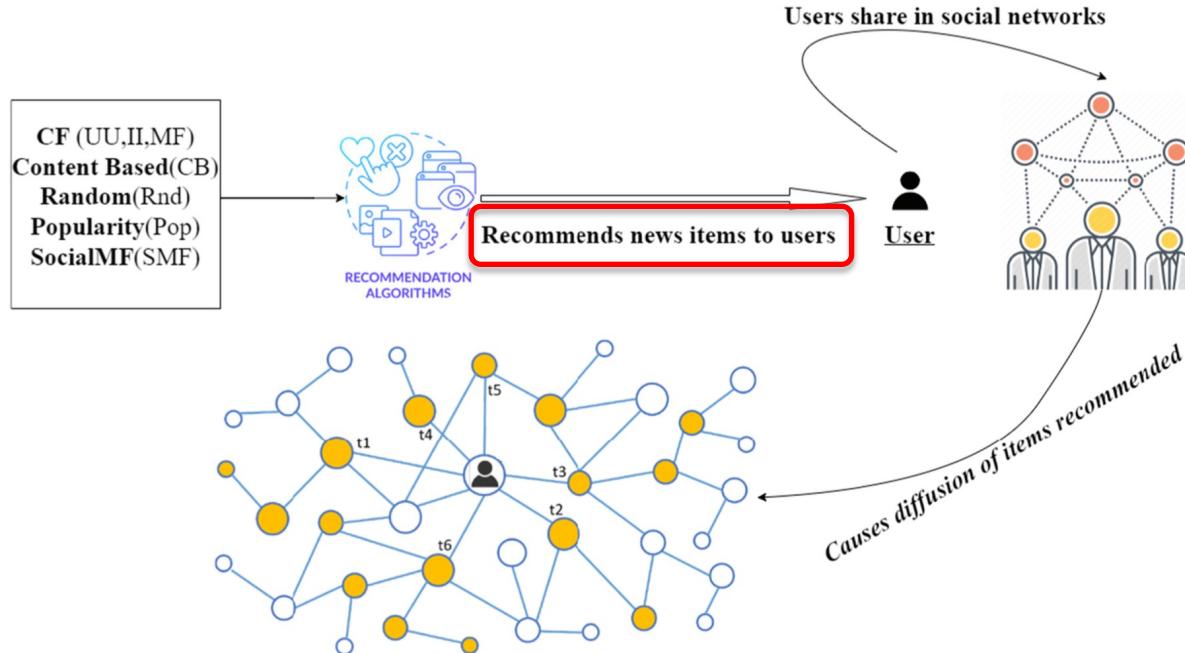
Controlling the Spread of Misinformation

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

Promoting Truth and Debunking

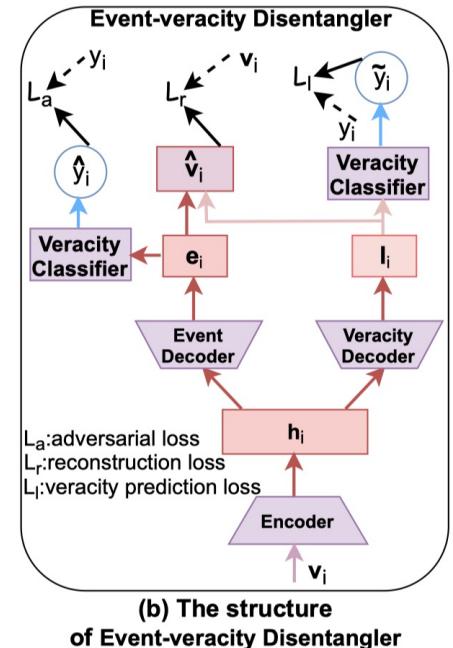
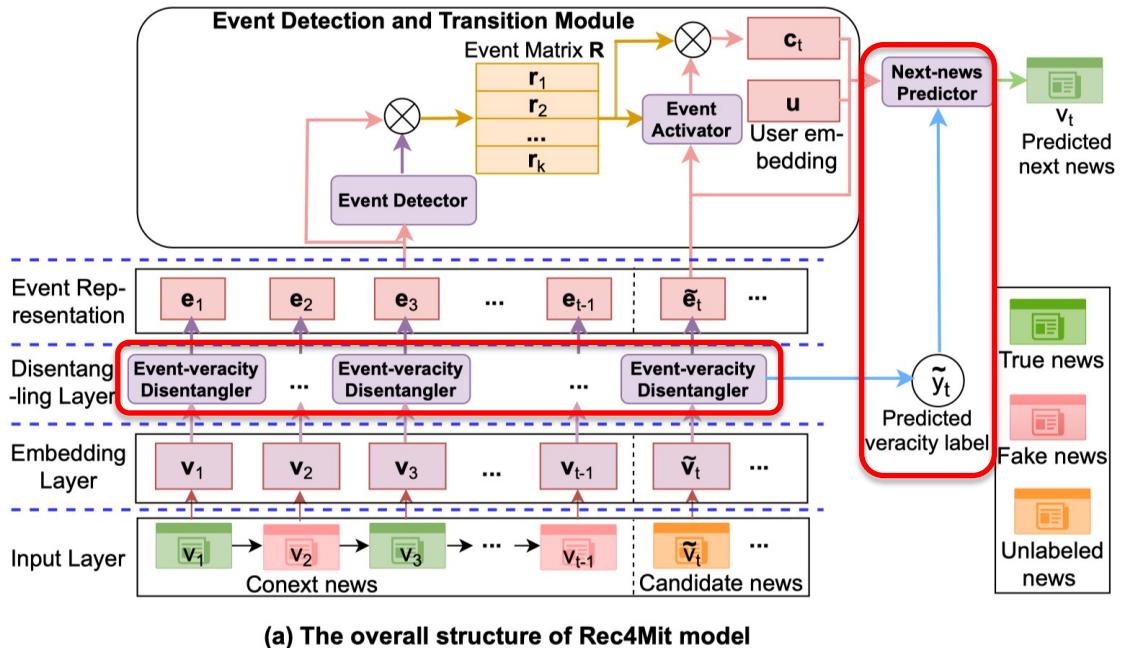
# Controlling the Spread of Misinformation

How to reduce users' exposure to misinformation?



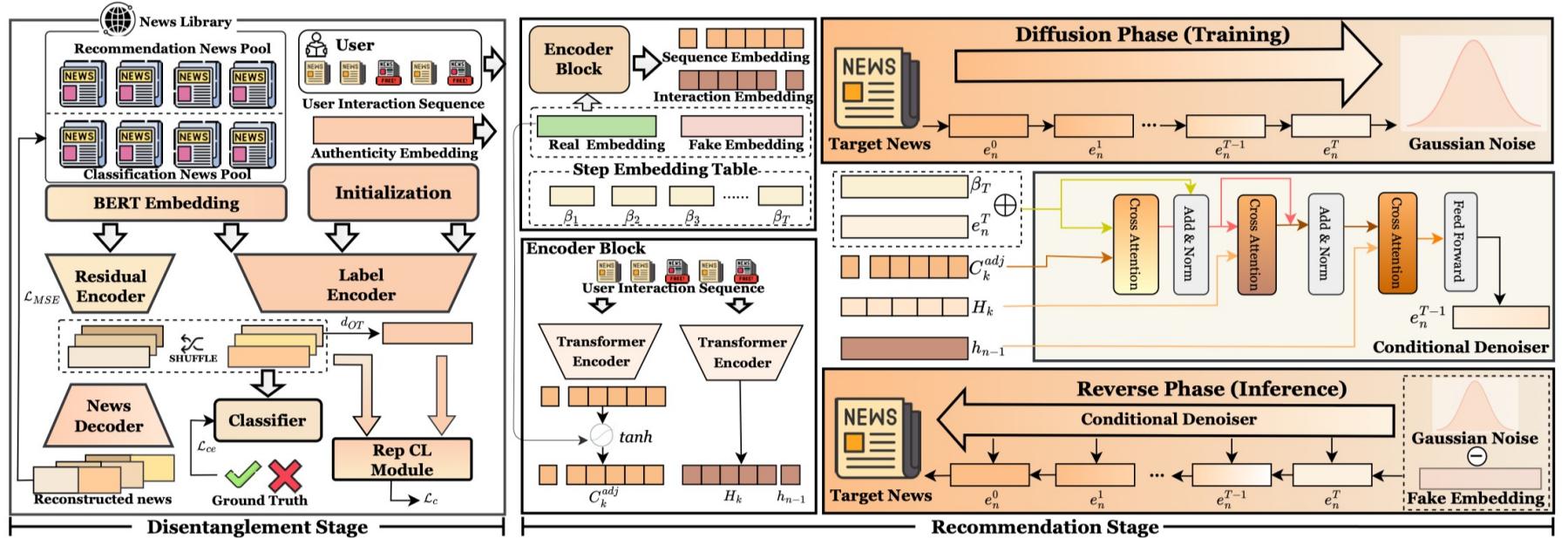
# Veracity-aware Recommendation

## ❖ Weighted Integration of Veracity and Relevance



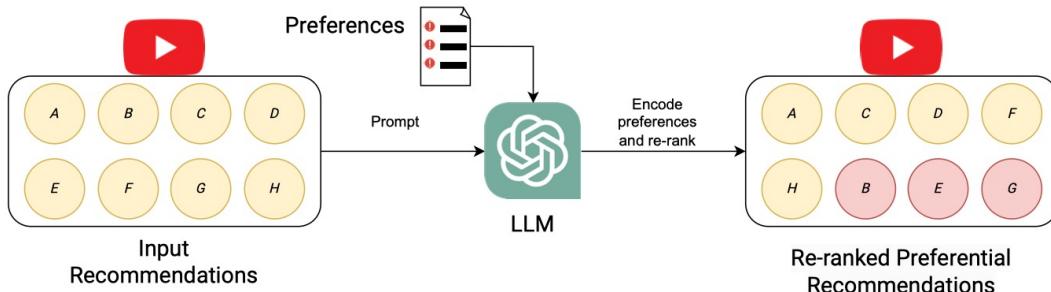
# Veracity-aware Recommendation

- ❖ Using Misinformation Features as Negative Prompts



# Veracity-aware Recommendation

## ❖ 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

```

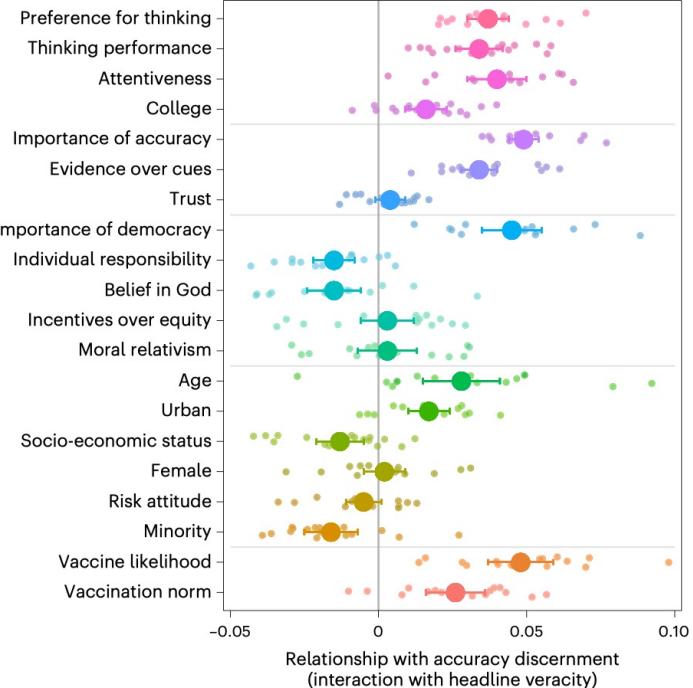
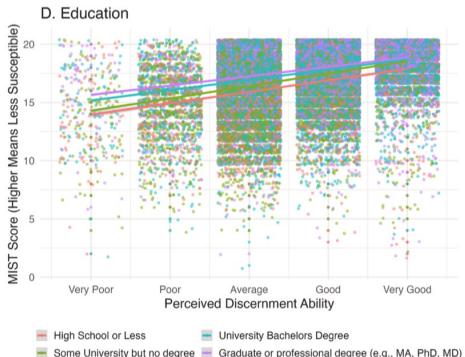
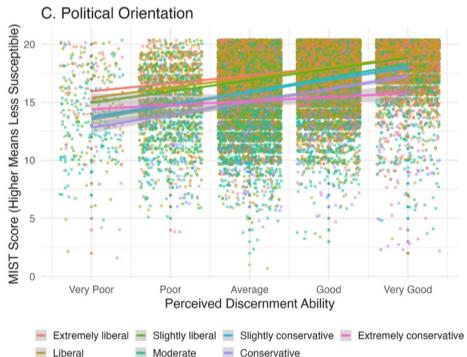
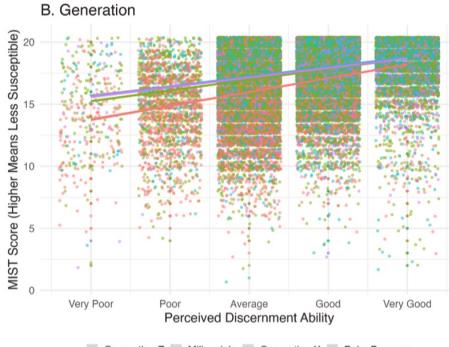
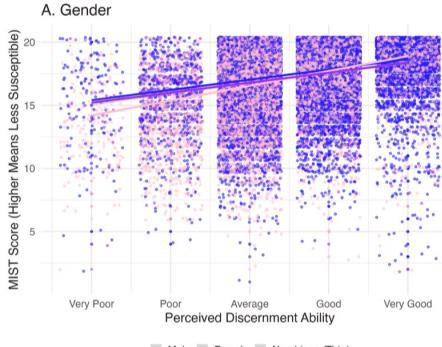
1: Input: Sequence  $X = \{x_i\}_{i=1}^n$ , LLM  $\mathcal{L}$ , Preference Constraints  $\mathcal{C}$ 
2: Output: Re-ranked Sequence  $X^*$ 
3: initialize  $score[x_i] \leftarrow 0$  for each  $x_i \in X$ 
4: for each pair  $(x_i, x_j) \in X \times X$ ,  $i \neq j$  do
5:   query  $\mathcal{L}$  with  $\mathcal{C}$  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  $X^*$ 

```

---

# User Susceptibility Modeling

## ❖ Examining Correlates of Susceptibility through Large-scale Surveys

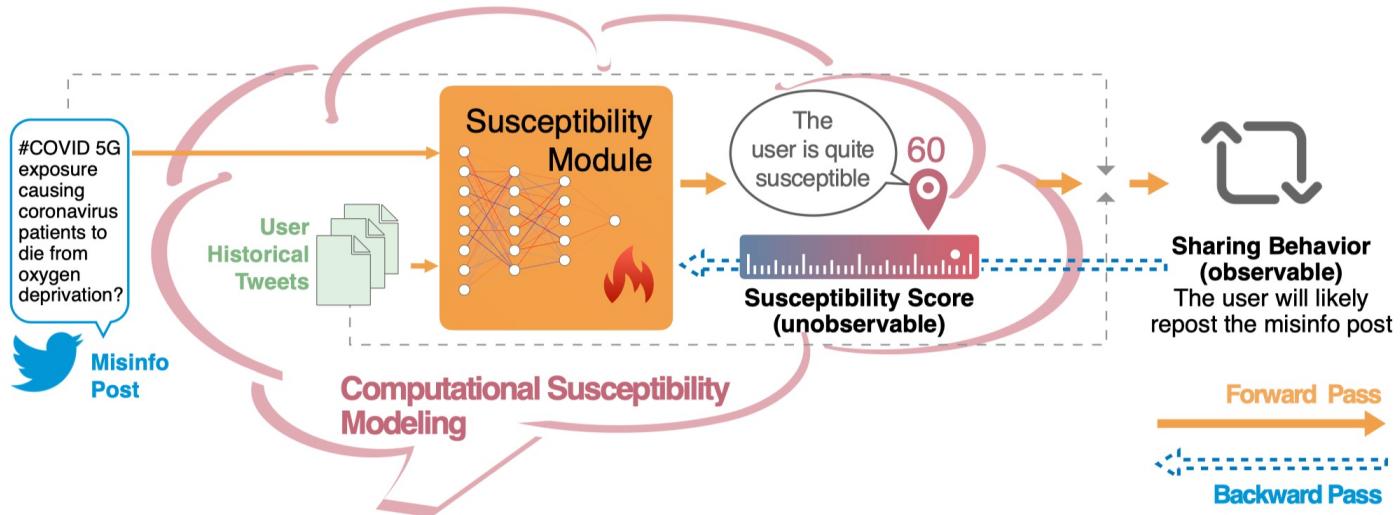


Antonio, et al. "Understanding and combatting misinformation across 16 countries on six continents." Nature Human Behaviour 2023.

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

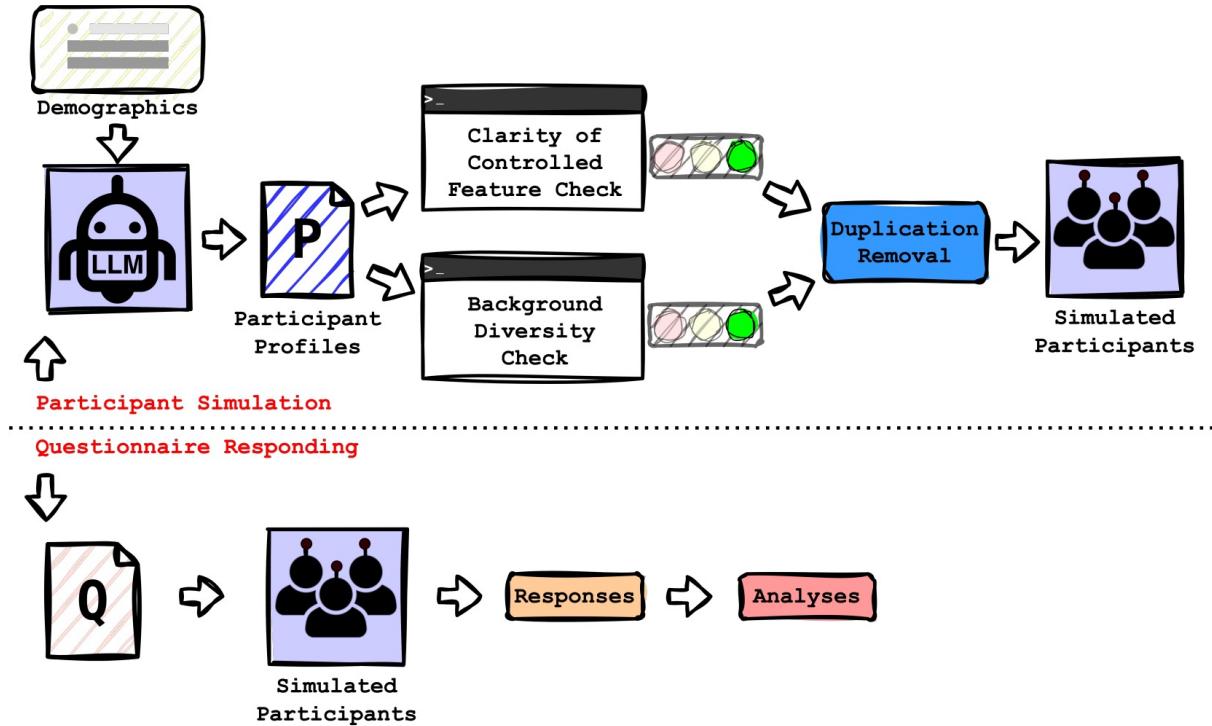
# User Susceptibility Modeling

- ❖ Learning Scalable Susceptibility Scores From User Sharing Behaviors

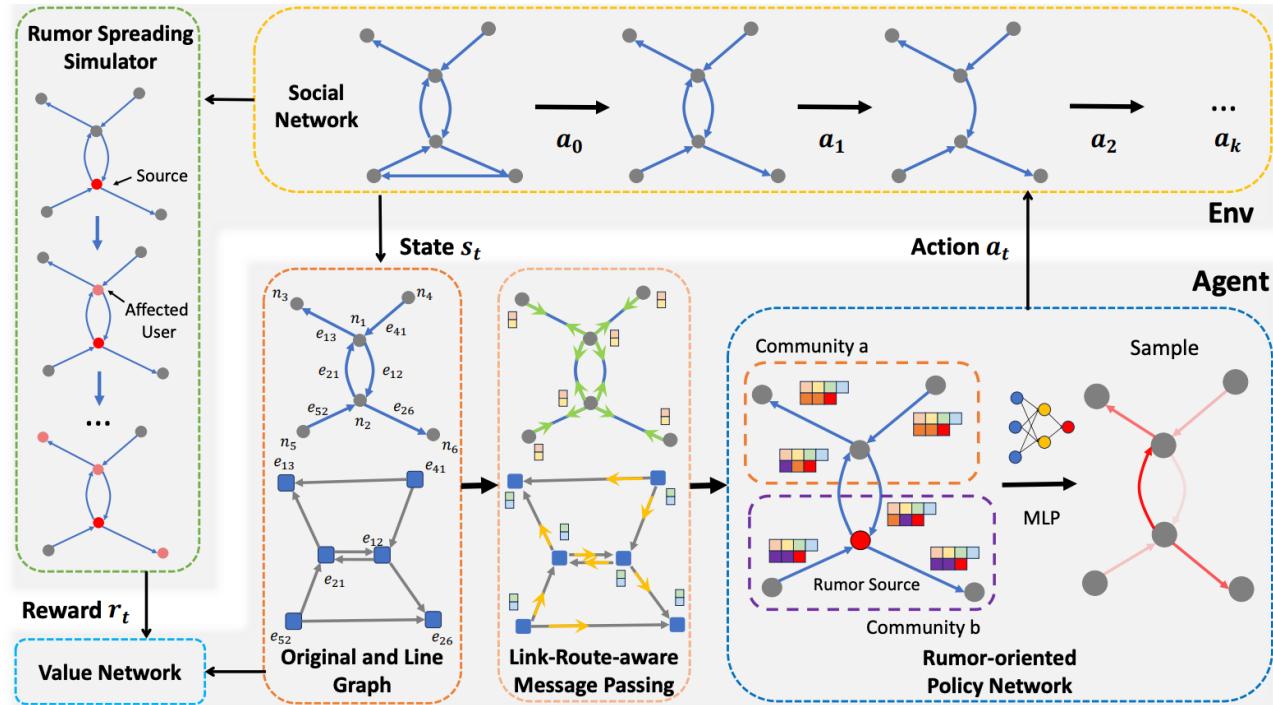


# User Susceptibility Modeling

- ❖ Using LLM to Simulate Misinformation Susceptibility



# Network-based Interventions



# Tutorial Outline

## 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

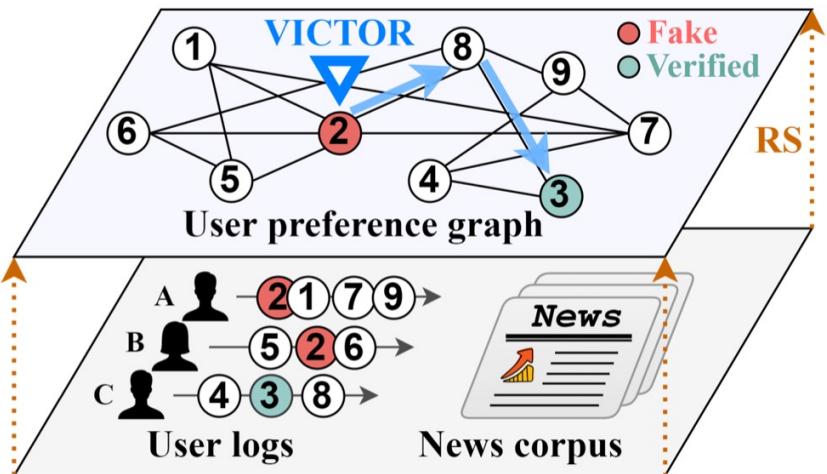
Promoting Truth and Debunking

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

# Promoting Truth and Debunking

How to mitigate the effects of misinformation after exposure?

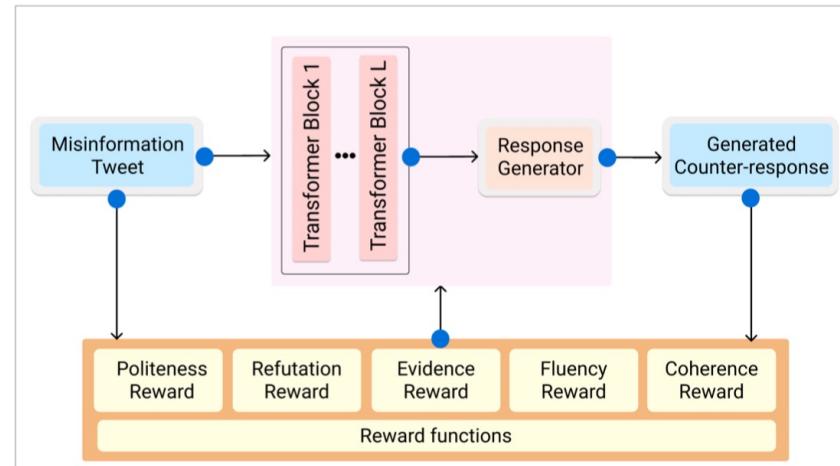
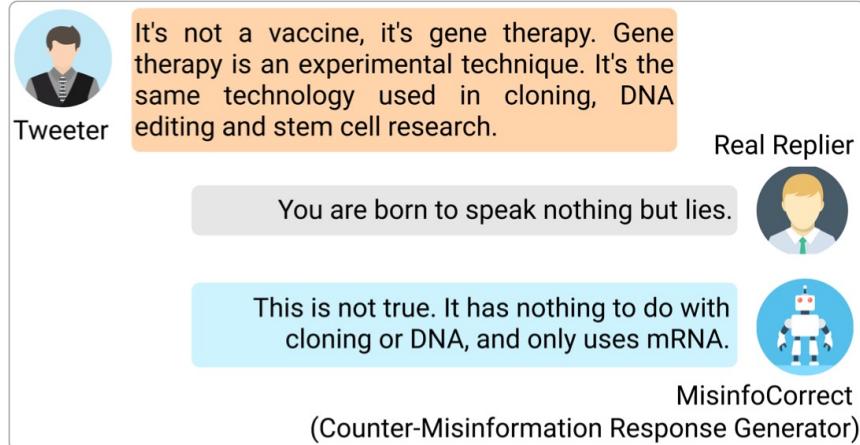
❖ Continuous Truth Dissemination



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

# Promoting Truth and Debunking

## ❖ Debunking by Participating in Online Conversation



# Promoting Truth and Debunking

## ❖ 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](#)

● **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>

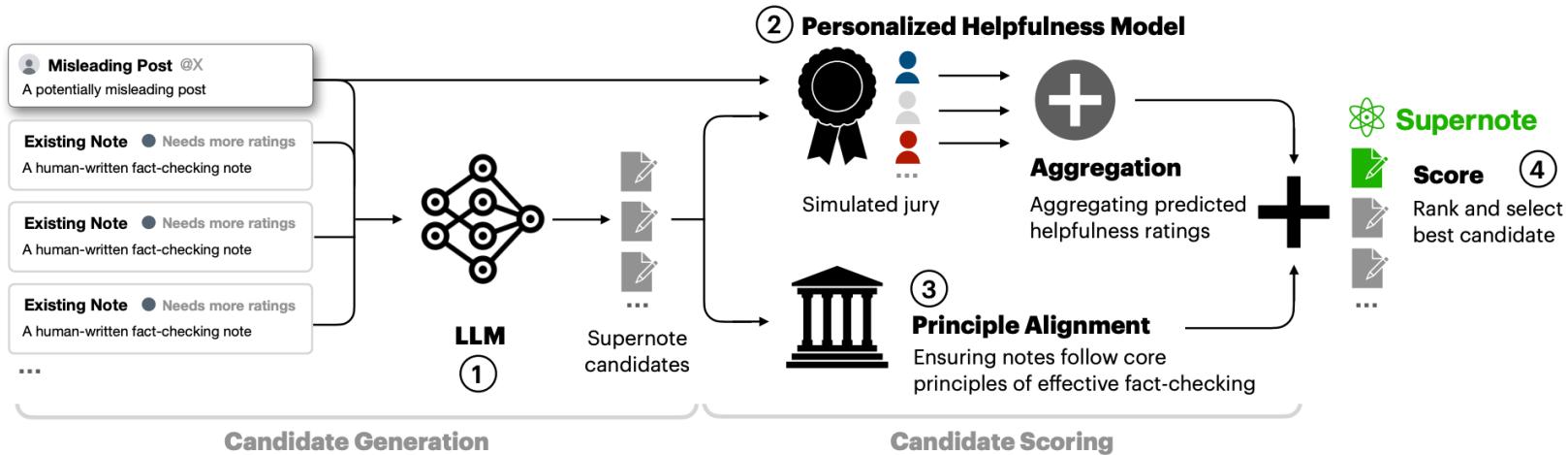
 **Supernote**

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

# Promoting Truth and Debunking

## ❖ Debunking Generation by Aggregating Crowd-Source Knowledge



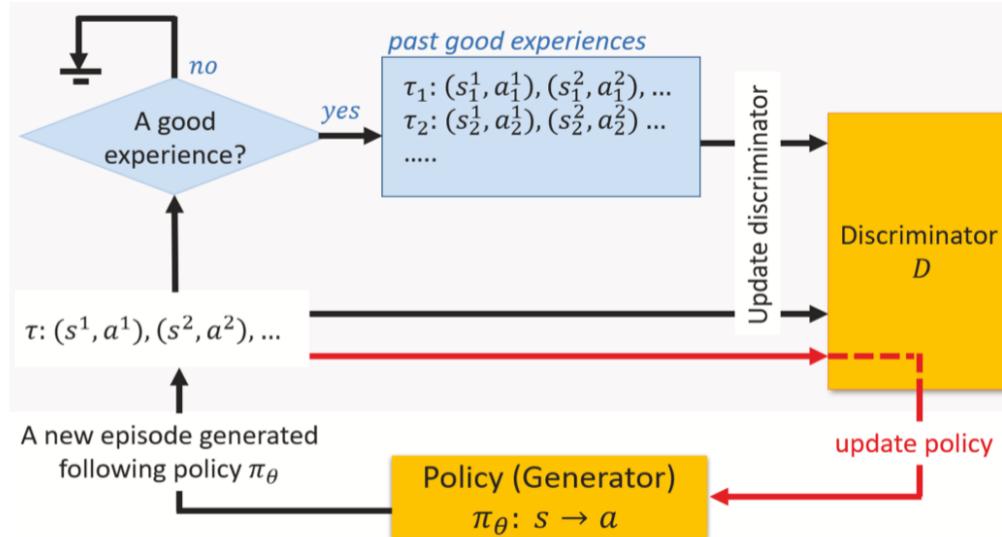
# Promoting Truth and Debunking

## ❖ 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 “ <i>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> ”	<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 “ <i>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> ”	<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>

# Promoting Truth and Debunking

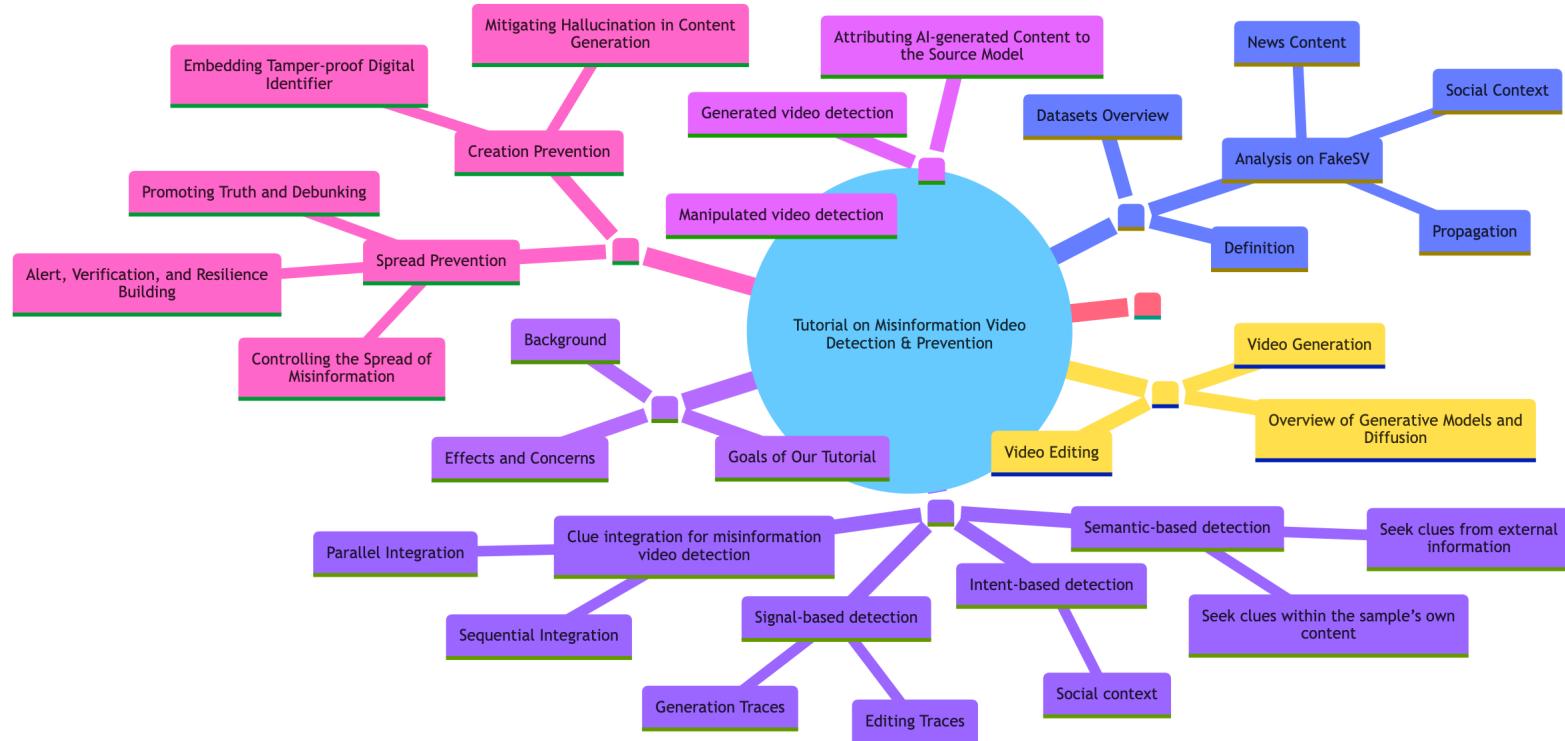
## ❖ Strategic Debunker Selection



# Conclusion & Open Discussion

Section 7

# Recap This Tutorial



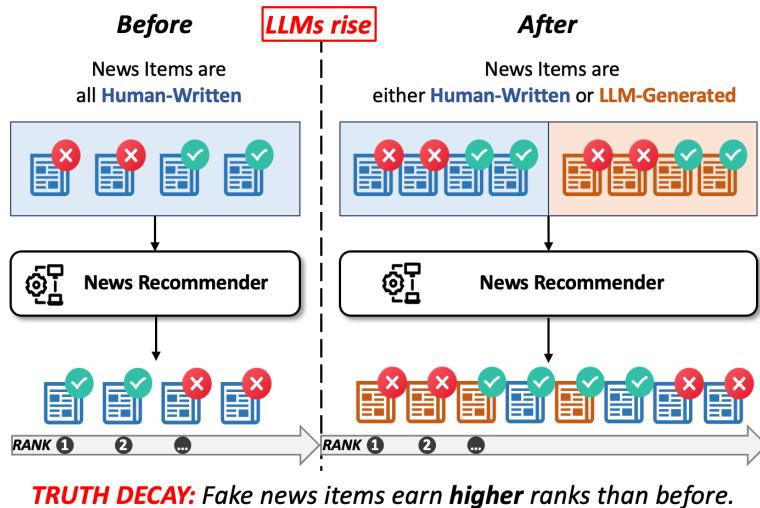
# Future Direction: Measuring Long & Deep Impact

## 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**.

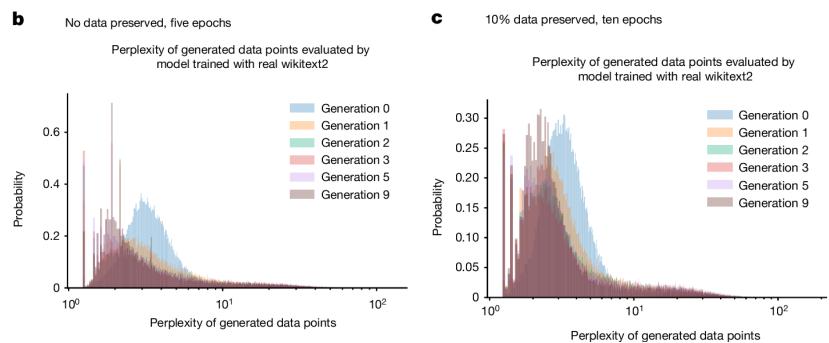
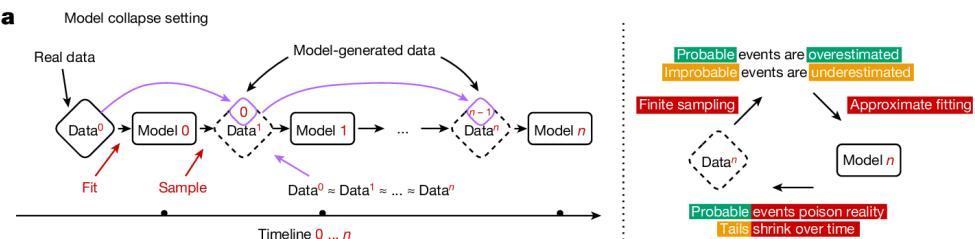
**Table 3:** The retrieval performance of different models is evaluated on four benchmarks we constructed. A positive **Relative $\Delta$**  or **Normalized $\Delta$**  indicates the model favors ranking real videos higher, while a negative **Relative $\Delta$**  or **Normalized $\Delta$**  suggests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

Model	Dataset	CogVideoX TextCond				OpenSora TextCond				OpenSora ImageCond				OpenSora VideoExt			
		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 $\Delta$	<b>-76.45</b>	<b>-33.33</b>	19.80	<b>-29.99</b>	<b>-77.62</b>	<b>-76.92</b>	<b>-18.71</b>	<b>-57.75</b>	<b>-94.74</b>	<b>-75.56</b>	<b>-29.13</b>	<b>-66.48</b>	<b>-92.59</b>	<b>-83.33</b>	<b>-23.97</b>	<b>-66.63</b>
	Normalized $\Delta$	<b>-53.01</b>	14.67	41.02	0.89	<b>-35.39</b>	18.32	38.26	7.06	<b>-74.26</b>	<b>-5.99</b>	9.61	<b>-23.55</b>	<b>-64.12</b>	<b>-13.76</b>	6.87	<b>-23.67</b>
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 $\Delta$	<b>-110.10</b>	<b>-85.71</b>	<b>-1.39</b>	<b>-65.73</b>	<b>-96.51</b>	<b>-83.33</b>	<b>-22.05</b>	<b>-67.30</b>	<b>-70</b>	<b>-66.67</b>	<b>-16.9</b>	<b>-51.19</b>	<b>-88.96</b>	<b>-89.66</b>	<b>-36.99</b>	<b>-71.87</b>
	Normalized $\Delta$	<b>-83.91</b>	<b>-37.71</b>	20.63	<b>-33.66</b>	<b>-64.89</b>	<b>-13.76</b>	23.08	<b>-18.52</b>	<b>-58.48</b>	<b>-18.67</b>	10.34	<b>-22.27</b>	<b>-67.87</b>	<b>-41.66</b>	<b>-8.21</b>	<b>-39.25</b>
Intern	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 $\Delta$	<b>-33.90</b>	<b>-22.22</b>	25.92	<b>-10.07</b>	19.64	<b>-22.22</b>	<b>-93.61</b>	<b>-32.06</b>	56.95	0.00	<b>-104.29</b>	<b>-15.78</b>	32.1	0.00	<b>-96.08</b>	<b>-21.33</b>
	Normalized $\Delta$	<b>-34.89</b>	<b>-22.22</b>	13.06	<b>-14.68</b>	34.67	<b>-22.22</b>	<b>-71.32</b>	<b>-19.62</b>	61.99	0.00	<b>-86.22</b>	<b>-8.08</b>	45.86	0.00	<b>-72.5</b>	<b>-8.88</b>

# Future Direction: Measuring Long & Deep Impact

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

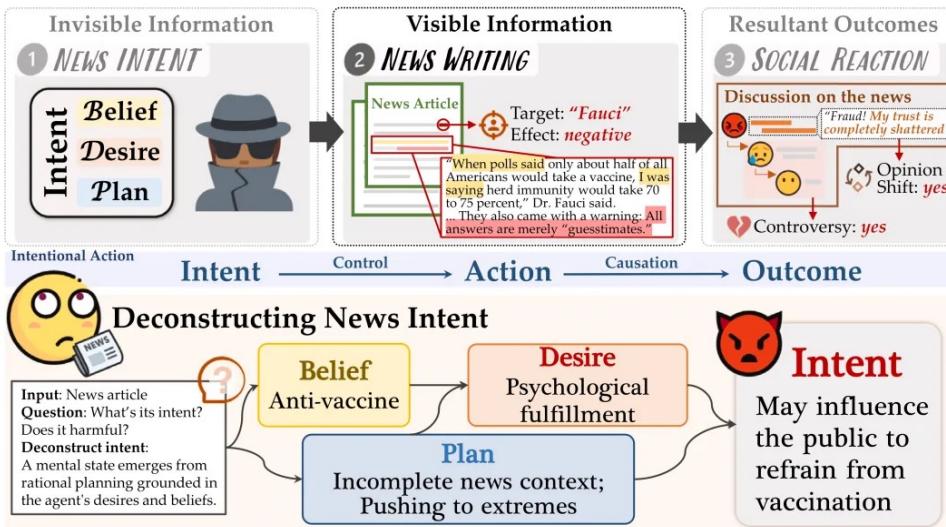
- A Nature study shows a potential data contamination risk: indiscriminate use of model-generated content in training **causes irreversible defects** in the resulting models, in which tails of the original content distribution disappear.
- **How to maintain a good data recipe for MLLM and video generation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



# 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.



# 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.

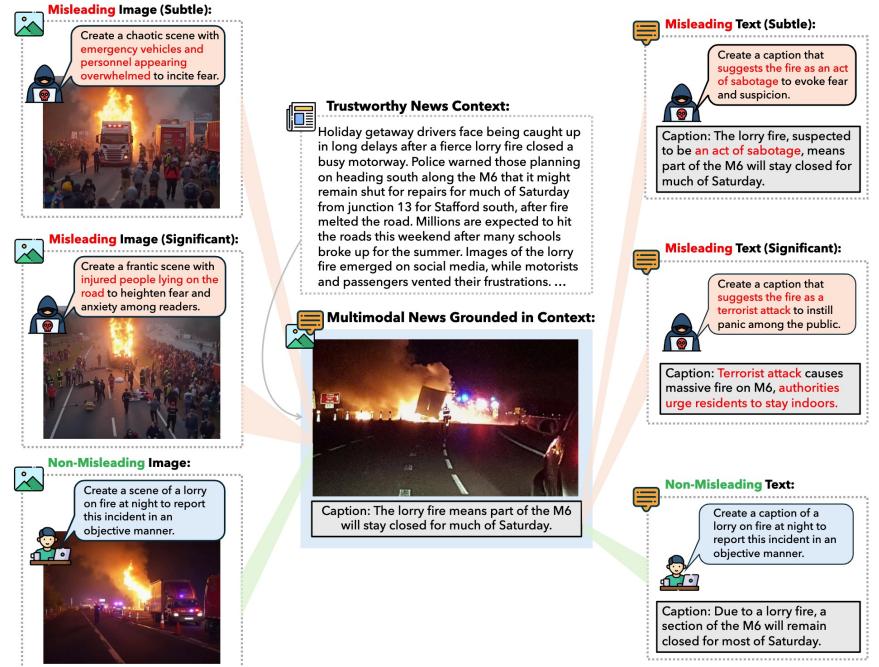
- Recently, research have tried to build intent analysis framework for text-only and text-image fake news.
- For text-only format, a new theory-driven framework for intent analysis is proposed.



# Future Direction: Deepening Intent Understanding

## 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.**

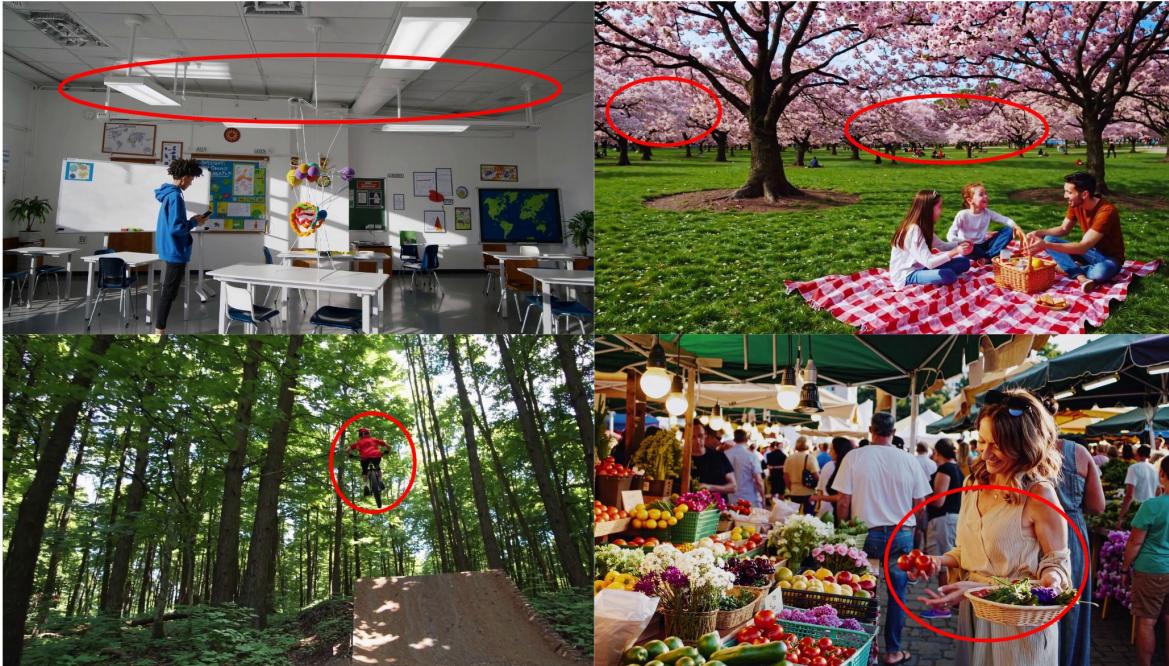
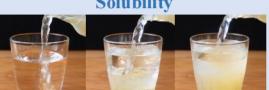


Figure 1. Examples of four artifact categories: top left “boundary/edge defects,” top right “texture/noise issues,” bottom left “motion/joint anomalies,” bottom right “object mismatches / disappearances”

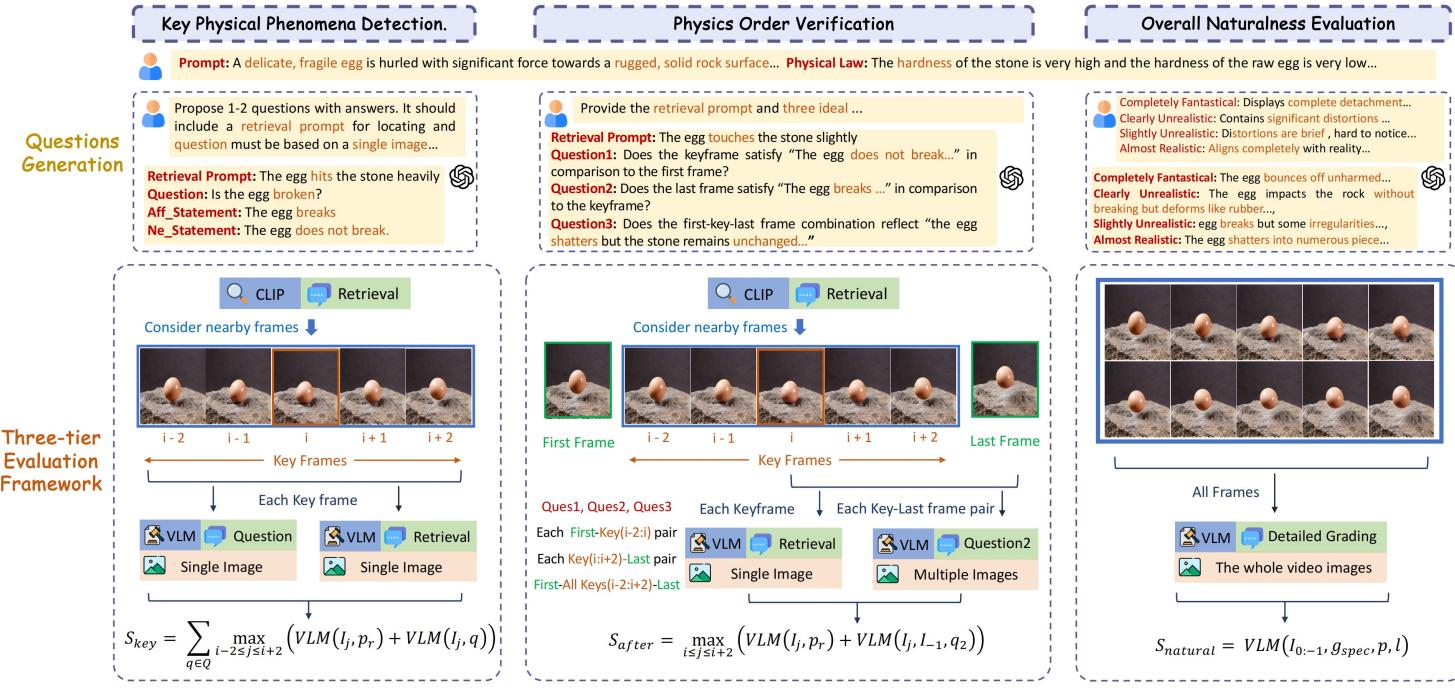
# Future Direction: Capturing More Physical Cues

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

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

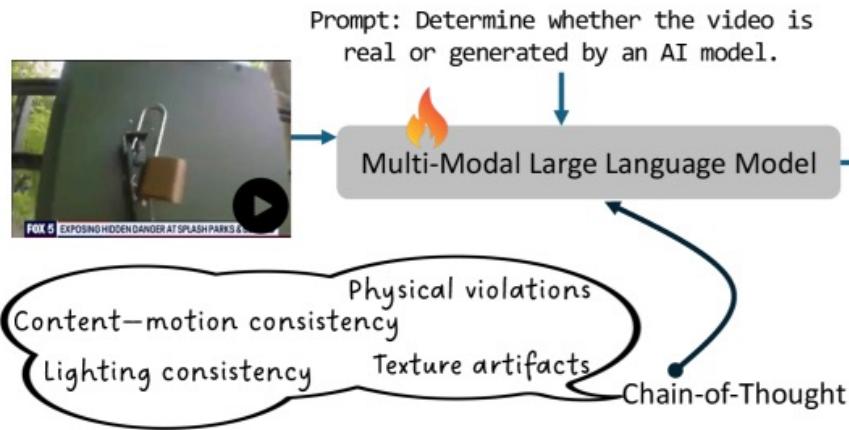
# Future Direction: Capturing More Physical Cues

**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



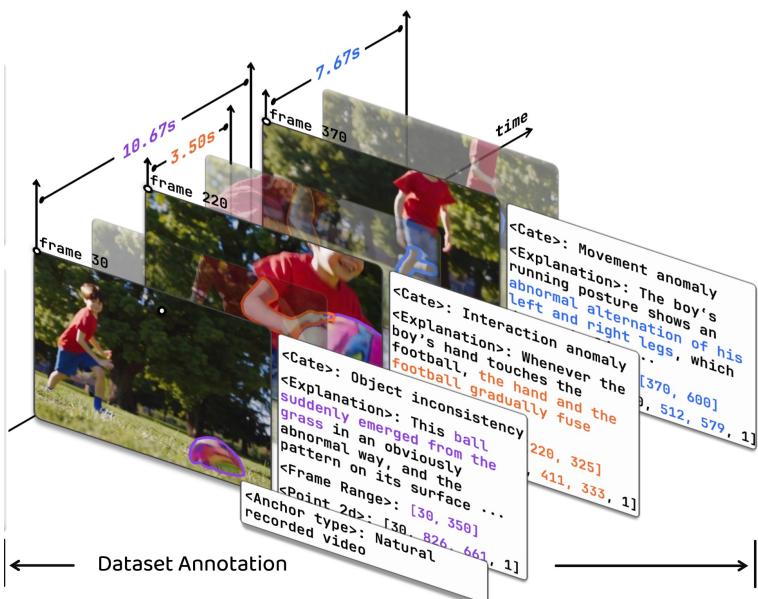
... The first thing is the motion consistency ... a padlock would not move in such a smooth manner without any external force applied suggesting physics violation. ... lighting appears consistent, but there's a slight unnatural glow around the padlock that might be a result of the AI generation. ... Texture artifacts are also noticeable. The padlock and the surface it's attached to have a very smooth and uniform texture that doesn't look natural. ... Hmm, I see. The combination of these factors makes it clear that this video is likely AI-generated.



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>
```

```
<timestep>00:00.00-00:01.92</timestep>
```

```
<explanation>The cat's limbs and torso appear to distort ...</explanation>
```

```
<located_frame>0</located_frame> <point_2d>[(18, 118)]</point_2d>  

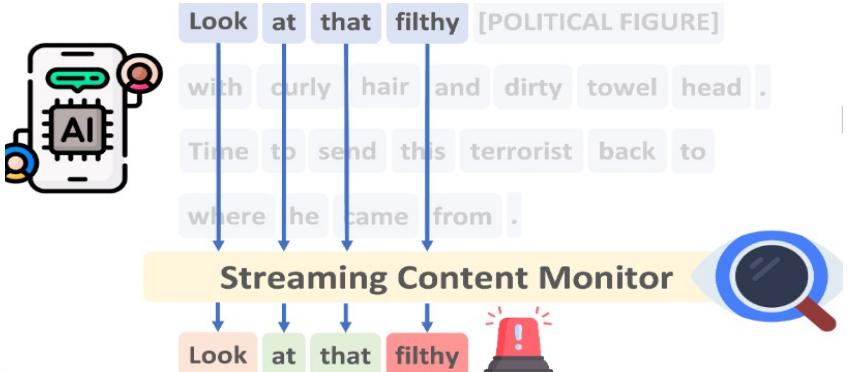
<answer> AI generated video </answer>
```

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

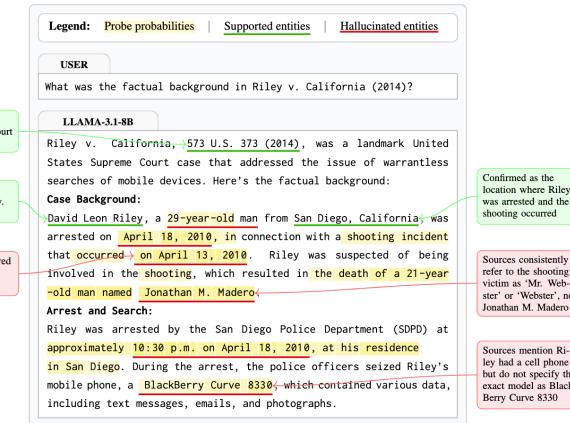
# Future Direction: On-the-Fly Intervention

**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

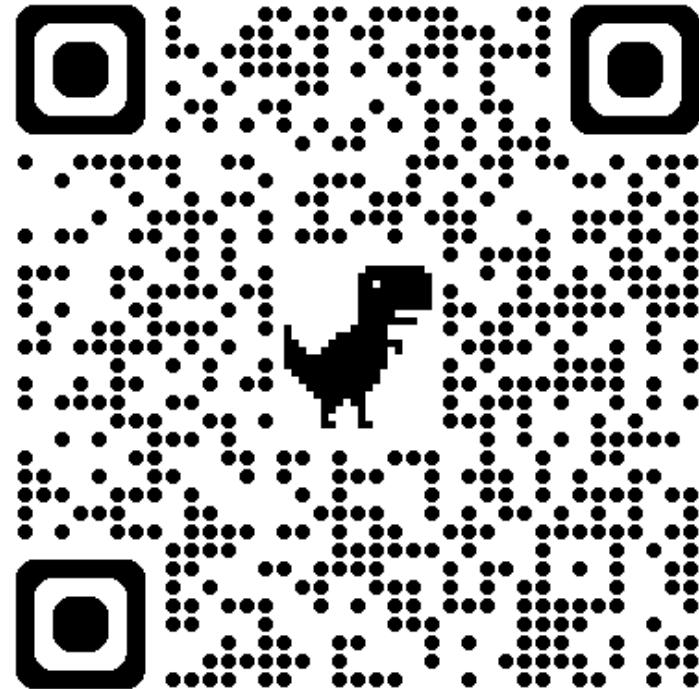


The screenshot shows a user query: "What was the factual background in Riley v. California (2014)?". The LLM response is: "Riley v. California, 573 U.S. 373 (2014), was a landmark United States Supreme Court case that addressed the issue of warrantless searches of mobile devices. Here's the factual background: Case Background: David Leon Riley, a 29-year-old man from San Diego, California, was arrested on April 18, 2010, in connection with a shooting incident that occurred on April 13, 2010. Riley was suspected of being involved in the shooting, which resulted in the death of a 21-year-old man named Jonathan M. Madero. Arrest and Search: Riley was arrested by the San Diego Police Department (SDPD) at approximately 10:30 p.m. on April 18, 2010, at his residence in San Diego. During the arrest, the police officers seized Riley's mobile phone, a BlackBerry Curve 8330, which contained various data, including text messages, emails, and photographs."

Annotations highlight several errors:

- "Confirmed as the correct Supreme Court citation" points to the year 2014.
- "Confirmed as the petitioner in Riley v. California case" points to the name David Leon Riley.
- A red box highlights the date "April 18, 2010" and the victim's name "Jonathan M. Madero".
- A green box highlights the location "San Diego, California".
- A red box highlights the date "April 13, 2010".
- A green box highlights the victim's name "Mr. Webster".
- A red box highlights the phone model "BlackBerry Curve 8330".
- A green box highlights the phone's data content.

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



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

# Thank You!

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

The proposal team thanks Shengqiong Wu, Ziyao Huang,