# Multimodality, manipulation, and malintent: A "Three-M" framework for multimodal disinformation

**Kun He**

**April 9 IOEA**

Postdoctoral Researcher
Tilburg University and University of Groningen
k.he@tilburguniversity.edu

# Research Agenda

### AI Risk and AI Safety

Content curation within LLMs, AI risks for national security

### Populism behind China's Great Firewall

Communist populism, online-bottom-up populism.

### Disinformation

Multimodal and AI disinformation

# Content for Today

01

## Research Introduction

Why multimodal disinformation has been understudied despite being pervasive.

02

## "Three-M" Framework

Multimodality, Manipulation, and Malintent as analytical framework for multimodal disinformation

03

## French Riots Case Study

Operationalizing the Three-M framework on a real-world case: the 2023 French Riots.

04

## Safeguarding Democracy

Detection of AI-Generated Election Images

# Research Introduction

### Misinformation

Refers to false or inaccurate information that is shared without the intent to deceive. For instance, Wardle et al., (2017) define misinformation as "false information that is shared, but no harm is meant"

### Disinformation

Disinformation is false information that is created, produced, or shared deliberately with the intention to deceive, mislead, or manipulate. European Commission (2018) define it as "verifiably false or misleading information that is created, presented, and disseminated for economic gain or to intentionally deceive the public"

### Malinformation

Refers to genuine, factual information that is shared with the intention to cause harm, typically by taking it out of context, leaking it strategically, or using it to harass or manipulate.

| Term | True/False? | Intent? | Harm mechanism |
|---|---|---|---|
| **Misinformation** | False | ❌ No intent | Accidentally misleads |
| **Disinformation** | False | ✔️ Intentional | Designed to deceive |
| **Malinformation** | True but weaponized | ✔️ Intentional | Used out of context to harm |

# Research Introduction

## The Visual Disinformation Gap

Social media platforms have become fertile ground for the rapid spread of disinformation, misinformation, and deepfakes (Brennen et al., 2021; Lee, 2020; Tolz and Hutchings, 2023). [Social Media Logic vs. Mass Media Logic] Despite the prevalence of visual content across these platforms, the production, dissemination, and evolution of visual and multimodal disinformation remain significantly understudied, representing what researchers describe as "a scientific blind spot" (Weikmann and Lecheler, 2023, p.3697).

### Why Does This Gap Exist?

#### Cognitive Bias

The prevailing assumption that disinformation is predominantly text-based (Wardle and Derakhshan, 2017)

#### Perceived Authenticity

Visual content "bears a stronger relationship to the depicted reality than the abstract descriptions offered by text" (Dan et al., 2021, p.649), providing "an implicit guarantee of being closer to the truth" (Messaris and Abraham, 2001, p.217)

#### Technical Challenges

Visual and multimodal disinformation is more challenging to collect, store, and systematically analyze due to its complex nature, particularly when dealing with large-scale datasets (Brennen et al., 2021)

# Intent of Visuals: Seeing is believing?

**However, research indicates that multimodal disinformation, in comparison to textual disinformation, can more effectively affect public and emotional reactions, shaping attitude, and drive behavioral responses (Iyer et al., 2014; Powell et al., 2015;  Vaccari & Chadwick, 2020)**


Hiding in dark corners, firing from ambush, and aiming for people's heads and eyes: This is 'restoring public order in the region'

| | | |
|---|---|---|
| Press corps | Visible, dominant | Erased entirely |
| Officer's action | Public, documented | Appears covert, secret |
| Light effect | Ambiguous | Reads as weapon targeting |
| Scene | Complex, contested | Simple, condemnatory |
| Victim | Unclear/absent | Implied civilian targets |

# Platform-Specific Investigations and Emerging Gaps

## Established Research Domains

Current scholarship predominantly focuses on text-image interplay across established platforms including X (formerly Twitter), Facebook, and WhatsApp. These studies have generated valuable insights into disinformation mechanics on these platforms.

*Hameleers et al. (2020)* systematically tested how images enhance the persuasiveness and perceived credibility of disinformation on Twitter, demonstrating measurable effects on audience reception.

*Tucker et al. (2018)* prove that tweets containing images achieve significantly greater reach than text-only content on Twitter, while Facebook posts with visual elements consistently receive higher engagement rates across user demographics.

# Categorizing visual disinformation

Four typologies (Hameleers et al., 2020):

**decontextualization (**misattribution of visual/auditory material);

**reframing (**selective emphasis through editorial justaposition);

**visual doctoring (**direct image alteration);

**multimodal doctoring (**synthetic fabrication across modalities)

Modal richness and manipulative sophistication (Weikmann & Lecheler, 2023)

**low-sophistication** tactics, such as mislabeled or unmanipulated images, demand minimal technological expertise,

**high-sophistication** techniques (e.g.,deepfakes) necessitate advanced technical skills.

However, the proliferation of generative AI technologies, such as Midjourney and GPT-5, has fundamentally destabilized such hierarchies, democratizing access to tools and enabling users with limited expertise to generate hyper-realistic deepfakes, synthetic images, voice clones (Helmus, 2022), eroding distinctions between ‘low’ and ‘high’ sophistication (Dan et al., 2021).

## The Gap

**Platform Gap:** Substantially less scholarly work has examined increasingly influential short-video platforms such as TikTok, Douyin, and Kuaishou, despite their massive user bases and distinct affordances for multimodal content creation and dissemination. These platforms represent a critical frontier for disinformation research, as their algorithmic recommendation systems, editing tools, and cultural practices differ significantly from text-centric platforms.

**Malintent Gap:** intent level has been less discussed.

# Introducing the "Three-M" Framework

## Multimodality, Manipulation, Malintent

Building upon UNESCO's (2018:7) definition of disinformation as "deliberate (often orchestrated) attempts to confuse or manipulate," we must distinguish between related but distinct phenomena in the information ecosystem.

### Distinguishing Mis- from Disinformation

**Visual misinformation** (Brennen et al., 2021; Garimella and Eckles, 2020) encompasses visuals that inadvertently propagate inaccuracies without deliberate deceptive intent—the accidental sharing of false content.

**Visual disinformation** (Vaccari and Chadwick, 2020; Weikmann and Lecheler, 2023) emphasizes strategic deception, defined as "any information that employs a visual, either in its original form or manipulated, to intentionally mislead, thereby constructing an image contrary to reality."

### Multimodality

How different semiotic modes combine to create meaning

### Manipulation

Technical and discursive techniques used to deceive

### Malintent

Tactical and strategic intentions behind disinformation

# First M: Multimodality and Mode "Reach"

Different modes have different ‘reaches’ (Kress, 2010, p. 83). Each mode possesses distinct strengths and weaknesses in meaning-making. Disinformation operates through the sophisticated integration of multiple semiotic modes, each contributing distinct affordances to the overall deceptive message. Understanding how these modes work individually and in concert is essential for analyzing multimodal disinformation.

### The Verbal Dimension

Including spoken language, writing text, which provide the essential contextual information that frame and support the disinformative argument. Written and spoken language establishes explicit claims, narrative structures, and interpretive frames that guide audience understanding of accompanying visual and auditory elements.

### The Visual Dimension

Visuals richly illustrate false narratives and demonstrate the physical properties of objects with compelling detail (Stöckl, 2024). Images and videos provide seemingly concrete "evidence" that leverages viewers' tendency to trust what they see, making visual modes particularly powerful vehicles for deception.

### The Auditory Dimension

Though often underestimated in disinformation analysis, it subtly but powerfully influences audiences' emotions and perceptions. Audio elements set emotional tone, influence interpretation of visual information, and leverage music's capacity to bypass critical thinking while appealing directly to emotional responses and cultural associations.

# Second M: Manipulation and Multimodal Coherence

## Understanding Deceptive Integration

**Multimodal coherence** refers to "the linking of semiotic modes and their formal, semiotic, and functional integration" (Stöckl, 2019, p. 53), how different modes work together to create unified meaning.

**Manipulation** encompasses "social power abuse, cognitive mind control, and discourse interaction" (Van Dijk, 2006, p. 359), operating across multiple levels of communication.

Technical Manipulation: Involves the direct techniques and skills used to modify or enhance content: such as image or video editing, deepfake, and audio alterations (Weikmann & Lecheler, 2023). As the most explicit form of manipulation, these modifications shape the surface-level credibility, making it visually compelling and seemingly authentic.

Discourse Manipulation: Shapes language and narrative structures to advance ideological messages. This includes strategic rhetorical framing, such as emphasizing "our good things" while highlighting "their bad things"

Cognitive Manipulation: operates at a deeper level by leveraging multimodal content to exploit cognitive biases, emotional responses, and heuristic processing through multimodal content (Powell et al., 2019). This dimension examines how disinformation capitalizes on inherent cognitive tendencies to strengthen its perceived credibility, subtly guiding audience interpretation and response.

Social Manipulation: Represent the broadest and most implicit form of influence, focusing on how social power, structures and ideological control are manipulated to enhance disinformation credibility.

# Third M: Motivation(Malintent) and Misleading Intentions

Understanding why disinformation is created and disseminated requires examining intentions at two distinct but interconnected levels. These motivational layers help explain both the immediate techniques employed and the broader objectives pursued through disinformation campaigns.

### Tactical Malintent

Immediate intent to deceive by manipulating perception or belief through specific presentation methods. Evident in techniques such as decontextualization (removing context to alter meaning), reframing (shifting interpretive frames), and visual/multimodal doctoring to enhance credibility and persuasive power (Hameleers et al., 2020).

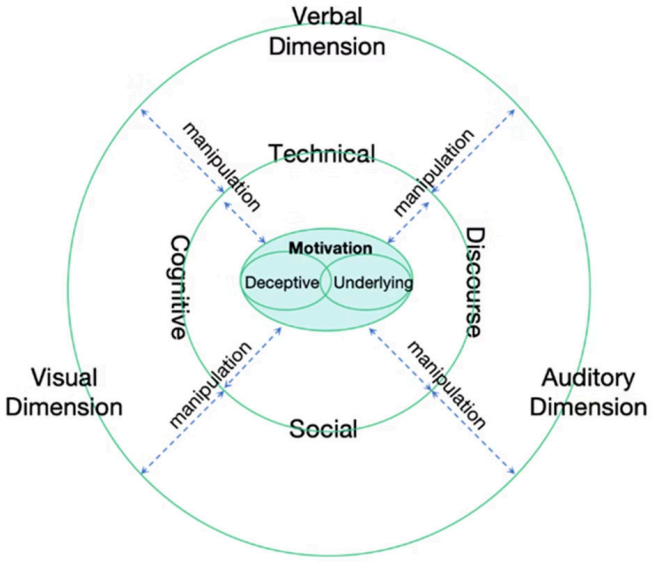
### Strategic Malintent

Broader, long-term objectives driving disinformation production and dissemination. These include political goals (delegitimization, opinion shaping, electoral influence), ideological aims (promoting belief systems), financial objectives (economic gain, market manipulation), and social agendas (provoking unrest, reinforcing divisions) (Hameleers, 2023).

---

The interplay between tactical and strategic motivations creates complex disinformation ecosystems where immediate deceptive techniques serve broader ideological, political, or financial objectives. Analyzing both levels is essential for comprehensive understanding.

## Onion Model: Three-M Framework




The outer layer represents the immediately observable various modes that are used in creating disinformation. Like the skin of an onion, it is what audiences first encounter and interact with, including elements such as written or spoken language, images, sounds and other components. Peeling back the surface reveals the manipulation techniques employed, representing the strategic alterations and combinations of different modes to achieve multimodal coherence and create a deceptive narrative. This is also where the concept of truth weaving becomes evidence, showing how factual and false information are intricately combined to misleading and deceiving. At the heart of the onion seals the malintent behind the disinformation, which is in many cases not straightforward and less easy to detect. In addition, exposure to and attempts to counter against disinformation, like cutting an onion which can cause tears, may lead to frustration, emotional distress, and social polarization as people grapple with disinformation.


## Case Study 1

# French Riots on Chinese Douyin and Kuaishou

This case study examines how major civil unrest events in France were represented, interpreted, and potentially distorted across two of China's largest short-video platforms: Douyin and Kuaishou. By applying the Three-M framework to user-generated and algorithmically amplified content about the French riots, we analyze the multimodal construction of narratives, the manipulation techniques employed across visual and auditory dimensions, and the tactical and strategic motivations underlying content creation and dissemination patterns on these platforms

# Case study:

## French riots: Nahel Merzouk riots

French police officer who shot 17-year-old under investigation for ‘voluntary homicide’

# Debunking disinformation


YouTube

**Riots in France: 'Burnt-down library' in Marseille is actually in Manila • FRANCE 24 Eng...**

In the wake of a week of rioting in France, sparked by the fatal shooting of 17-year-old Nahel by a police officer, out-of-context videos across social media claim to show destruction in the southe...

**The largest library in the city of Marseille burned down (Truth or Fake)**

**'Burnt-down library' in Marseille is actually in Manila**

## Short video circulated on Douyin

Digital Platforms: Douyin and Kuaishou；　Keywords: 法国暴乱 and 法国骚乱　The search yielded a corpus of 175 (Douyin) and 232 (Kuaishou)；

We identified 12 Douyin (D1 to12) and Kuaishou (K1 to K10) videos propagating this debunked narrative.

Google Docs

**dis 2.mp4**

# Short video circulated on Douyin

Google Docs

D11.mp4


# Alcazar Library

## Short video circulated on Douyin

Digital Platforms: Douyin and Kuaishou； Keywords: 法国暴乱 and 法国骚乱 The search yielded a corpus of 175 (Douyin) and 232 (Kuaishou)；

We identified 12 Douyin (D1 to12) and Kuaishou (K1 to K10) videos propagating this debunked narrative.

Google Docs

dis 2.mp4

# Short video circulated on Douyin

Google Docs

D11.mp4

Table 1. Codebook for analyzing multimodal disinformation.

| Code | Sub-codes | Definition |
|---|---|---|
| Verbal dimension | Identity V-I | Verbal identity: Linguistic elements that convey identity, credibility of the Alcazar Library. |
| Visual dimension | Layout Vis-L | Visual layout: Visual layout of text, including font style, size, color, and composition. |
| | Symbol Vis-A | Visual symbol of the building of Alcazar Library. |
| | Source Vis-S | Visual symbol to demonstrate the source of the information, such as the logo Beijing TV. |
| Auditory dimension | Music type Aud-M | Types of background music used (e.g., fast-paced or low-paced) to set emotional tone (e.g., intense-chaos, somber-tragedy, hopeful-resilience, neutral-objectivity, cheerful-parody). |
| | Sound effect Aud-S | Audio effects (e.g., shouting, explosion) that enhance certain aspects of the narrative. |

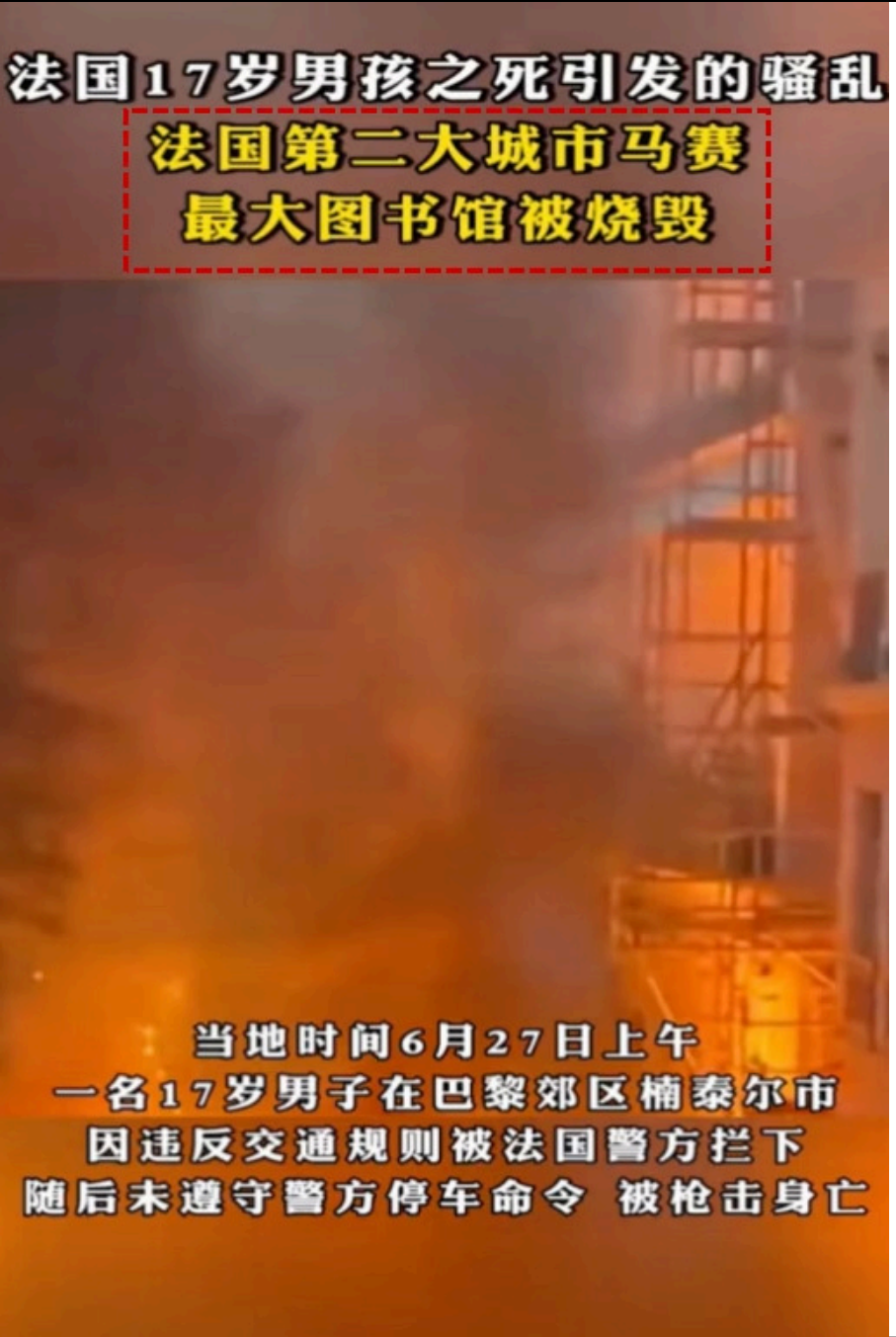

Footage 1 Footage 2 Footage 3

Aud–M: The background music is fast–paced, intense–chaos;
Aud–S: Sound of shouting and explosion

**Table 2.** Results of coding by applying the 'Three-M' theoretical framework.

| | Video | Verbal Mode | Visual Mode | | | Auditory Mode | |
|---|---|---|---|---|---|---|---|
| | Footage | Identity V-I | Layout Vis-L | Symbol Vis-A | Source Vis-S | Music Aud-M | Sound effect Aud-S |
| D-1 | A series of AI generated images | The largest library in the city was also destroyed by fire | | no | no | Fast-paced; hopeful-resilience | no |
| D-2 | Footage 1: the burning library. Footage 2: Street chaos. | Largest library burned down by protesters | Black text on a yellow background, bolded and centered | Footage 1 | Chengshi hudong | Fast-paced. somber – tragedy | A male's mourning voice in a sorrowful and subdued tone. |
| D-3 | An image of the burning library | The largest library in Marseille was set on fire | Yellow text, bolded and centered | Image | no | low-paced; neutral-objectivity | no |
| D-4 | Footage 1: burning street; Footage 2: a group of armed police running amidst smoke; Footage 3: The burning library (2 s); Footage 4: Street chaos. | The largest library in Marseille was destroyed by fire | | Footage 3 | Jingchu Net | fast-paced; intense-chaos | no |
| D-5 | A series images, including armed polices, guns, and street chaos | The largest library in Marseille has been burned down again. | | No | no | low-paced; somber-tragedy | no |
| D-6 | Footage 1: Burning street with big fire; Footage 2: A car crash into Lidl; Footage 3: Street chaos; Footage 4: A person opened fire with a machine gun. | The largest library in France was destroyed by fire. | Black and white text, centered | No | no | fast-paced; Intense-chaos | Explosion of firework. |
| D-7 | Footage 1: The burning library; Footage 2: Burning street; Footage 3: a group of armed police running amidst smoke; | The largest library in Marseille was destroyed by fire | Red and White text, bolded, centered. | Footage 1 | Sansha TV (Qingfengxia) | fast-paced; Intense-chaos | no |
| D-8 | Footage 1: Burning street; Footage 2: A group of armed polices; Footage 3: Burning cars; Footage 4: Street protests; Footage 5: An influencer broadcasts. | The largest library in France's second-largest city was burned down by someone. | Rolling caption | No | no | no | no |
| D-9 | 12 Footages in total, including street chaos, armed polices, burning buildings, and a damaged tourist bus. | The largest library in France's second-largest city was burned down by someone. | Rolling caption | No | no | fast-paced; intense-chaos | no |
| D-10 | A journalist broadcast, explaining the reasons behind the protests. | The library in Marseille was also set on fire. | | No | no | Fast-paced. somber – tragedy | no |

(*Continued*)




# Textual dimension

**法国第二大城市马赛最大图书馆被烧毁。**

### Discourse manipulation

(1) Marseille's largest library was burned down.

(2) Marseille's largest library was burned down by protesters.

(3) The largest library in the second largest city of France was burnt down.

### Technical Amanipulation

Visual-textual interplay: 50% of Douyin (6/12) and 40% of Kuaishou videos (4/10) employed high-contrast, stylized typography (Vis-L) to highlight deceptive captions. The use of bold fonts and chromatic salience enhances perceptual legitimacy.

# Visual clues

Post office in Manila

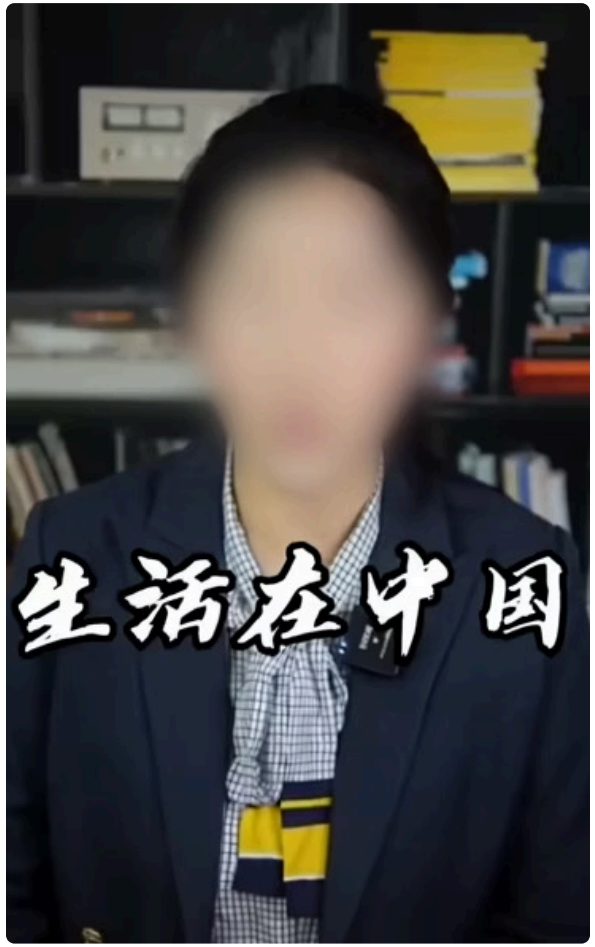
Visual analysis of subcode Vis-A (visual symbol of the Alcazar Library) unveils an absence of verifiable documentation, with 58% Douyin and 90% Kuaishou videos lacking identifiable footage of the Alcazar Library.

Alcazar library in Marseille


# Auditory dimension

The analysis of background music also shows the functional role of auditory manipulation in amplifying deceptive narratives. Across the corpus, 91.7% of Douyin videos (11/12) and 100% of Kuaishou videos incorporated background music, predominantly characterized by ominous and suspenseful tonalities.

# Malintent dimension




One content producer introduces the riots with a light and amusing tone, saying, “What do you prefer to see, burning streets or zero-dollar purchases? I will shoot it for you tomorrow” (K-4). This casual and somewhat irreverent approach not only attracts viewers but also capitalizes on the sensationalism of the events to boost content visibility and popularity.

# Truth Weaving

As a strategic and deliberate manipulation, increases the persuasiveness of the narrative. When authentic or true information is artfully interwoven with false or deceptive, it capitalizes on the trustworthiness of factual elements to protect and enhance credibility of the falsehoods. Consequently, the calculated, multimodal fusion of truths with falsehoods achieves a higher degree of multimodal coherence, making it more likely to be accepted and appeal to the audience.


YouTube

a better love story. Trump and Harris, Donald and Kamala, share a baby. #donaldtrum...

# Safeguarding Democracy:

Detection of AI-Generated Election Images

# The Rising Threat of AI-Generated Disinformation

Modern AI image generation tools create increasingly realistic synthetic media that challenges traditional verification methods. These examples demonstrate the sophistication of current generative models and the difficulty in distinguishing authentic from fabricated content.

# Visual Authenticity Under Scrutiny

# Comparative Analysis: Real vs. Synthetic

### AI-Generated Scenario


Synthetic images may display unnatural spatial relationships, impossible lighting conditions, or inconsistent perspective cues that reveal their computational origin.

Real photographs exhibit natural compression artifacts, coherent metadata, and physical consistency across all elements within the frame.

### Authentic Photography

# Detection Challenge: Subtle Indicators


Forensic analysis requires examination of micro-level features including noise patterns, frequency domain signatures, and statistical anomalies that human observers cannot reliably detect without computational assistance.

# Detection Challenge: Subtle Indicators



Forensic analysis requires examination of micro-level features including noise patterns, frequency domain signatures, and statistical anomalies that human observers cannot reliably detect without computational assistance.

### Synthetic Portrait Analysis

AI-generated images often exhibit subtle artifacts in facial features, lighting inconsistencies, and irregular texture patterns that distinguish them from authentic photographs.

### Authentic Reference

Genuine photographs maintain consistent photographic properties, natural depth of field, and coherent physical characteristics that align with real-world imaging constraints.


YouTube

**These AI Fights Look 100% Real (Seedance 2)**

This video has been created by an AI for experimental and entertainment purposes. The content of the video should not be considered as factual or reliable information, and any opinions or...

# Background: The Election Integrity Crisis

## Universal Attention

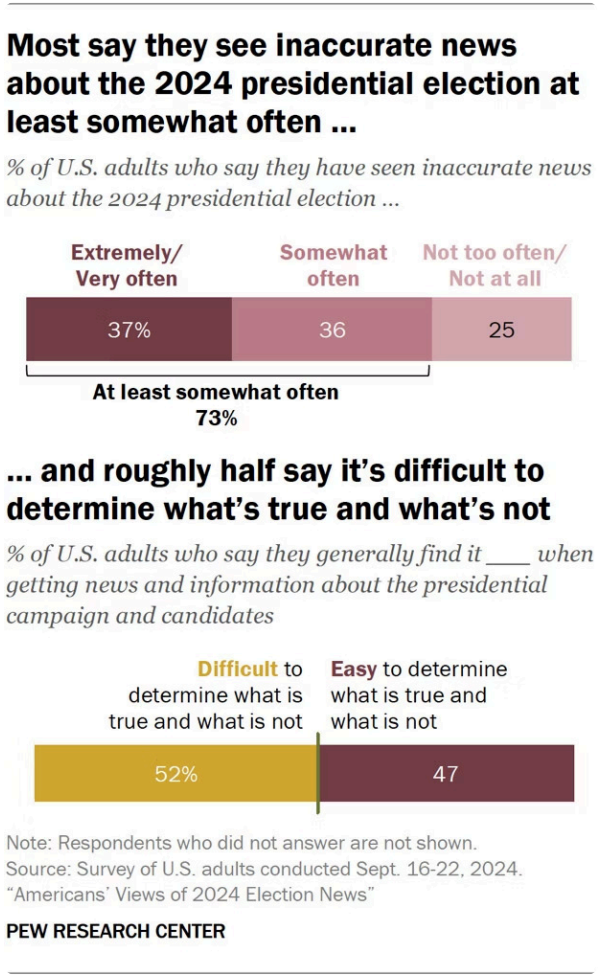
Electoral events command global attention, making them prime targets for coordinated disinformation campaigns seeking to influence public opinion and voting behavior.

## Accessibility of AI Tools

Generative AI platforms enable rapid creation of convincing fake content. Over 70% of U.S. adults report encountering misleading election-related information, with half struggling to identify fabricated material.

## Cognitive Impact

Sophisticated AI-generated disinformation creates cognitive confusion, erodes trust in authentic sources, and can systematically distort voters' perception of candidates and issues.




# Can we build a model to detect AI-generated images from the real ones?

# Data Collection Methodology

### Authentic Images

**2,601 images** sourced from WikiMedia Commons and verified Instagram accounts

- Political portraits
- Campaign speeches
- Official meetings and events
- Diverse photographic conditions

### Synthetic Images

**2,499 images** collected from social media and generated using contemporary AI tools

- Political figure representations
- Grok-generated content
- Various generative models
- Controlled synthesis parameters

Total dataset: **5,100 images** providing balanced representation for binary classification tasks with realistic distribution of authentic and synthetic political imagery.

# Methodology: Model Architecture

01

## Data Partitioning

70/15/15 split for training, validation, and test sets ensuring robust evaluation

02

## Preprocessing Pipeline

Grok watermarks systematically removed to prevent spurious feature learning

03

## Data Augmentation

Training images undergo random transformations: horizontal flipping, rotation (±15°), brightness adjustment (±20%), and scale variation

04

## Transfer Learning

EfficientNet-B0 architecture pretrained on ImageNet (1M+ images) provides robust feature extraction foundation


**EfficientNet-B0** balances computational efficiency with detection accuracy, making it suitable for real-time deployment scenarios.

# Results: High-Accuracy Classification

**96%**

### Overall Accuracy

Consistent performance across all evaluation metrics

**95%**

### Real Image Precision

Minimal false positive rate for authentic content

**97%**

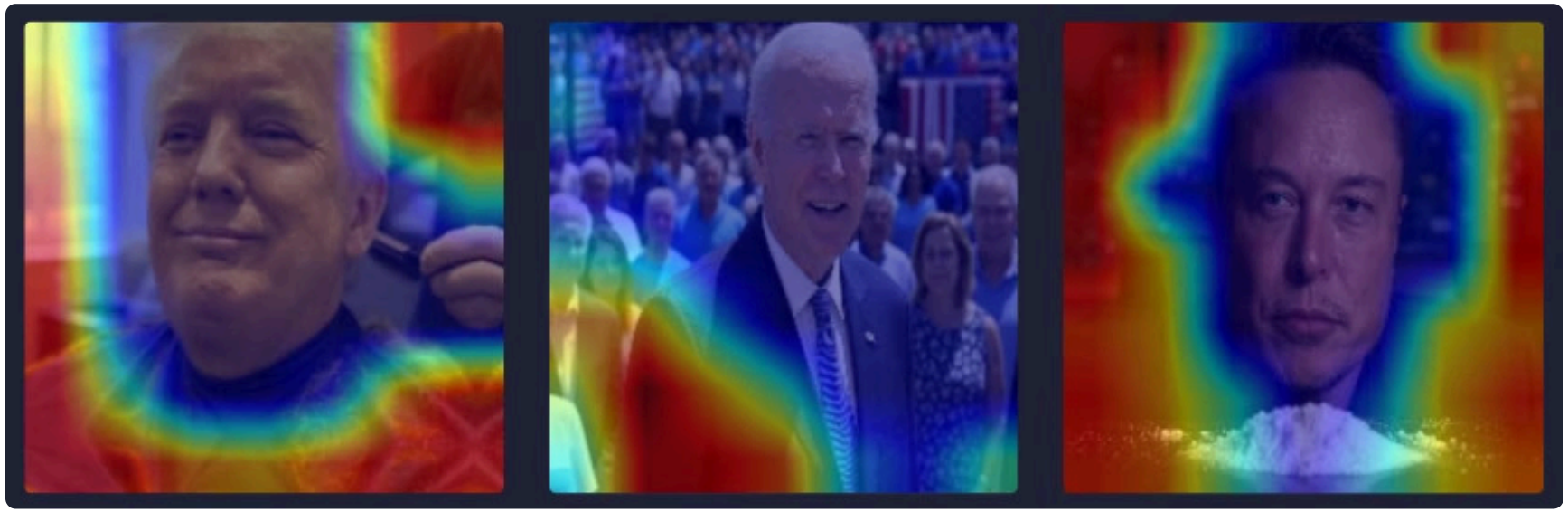
### AI Image Precision

Strong detection of synthetic material

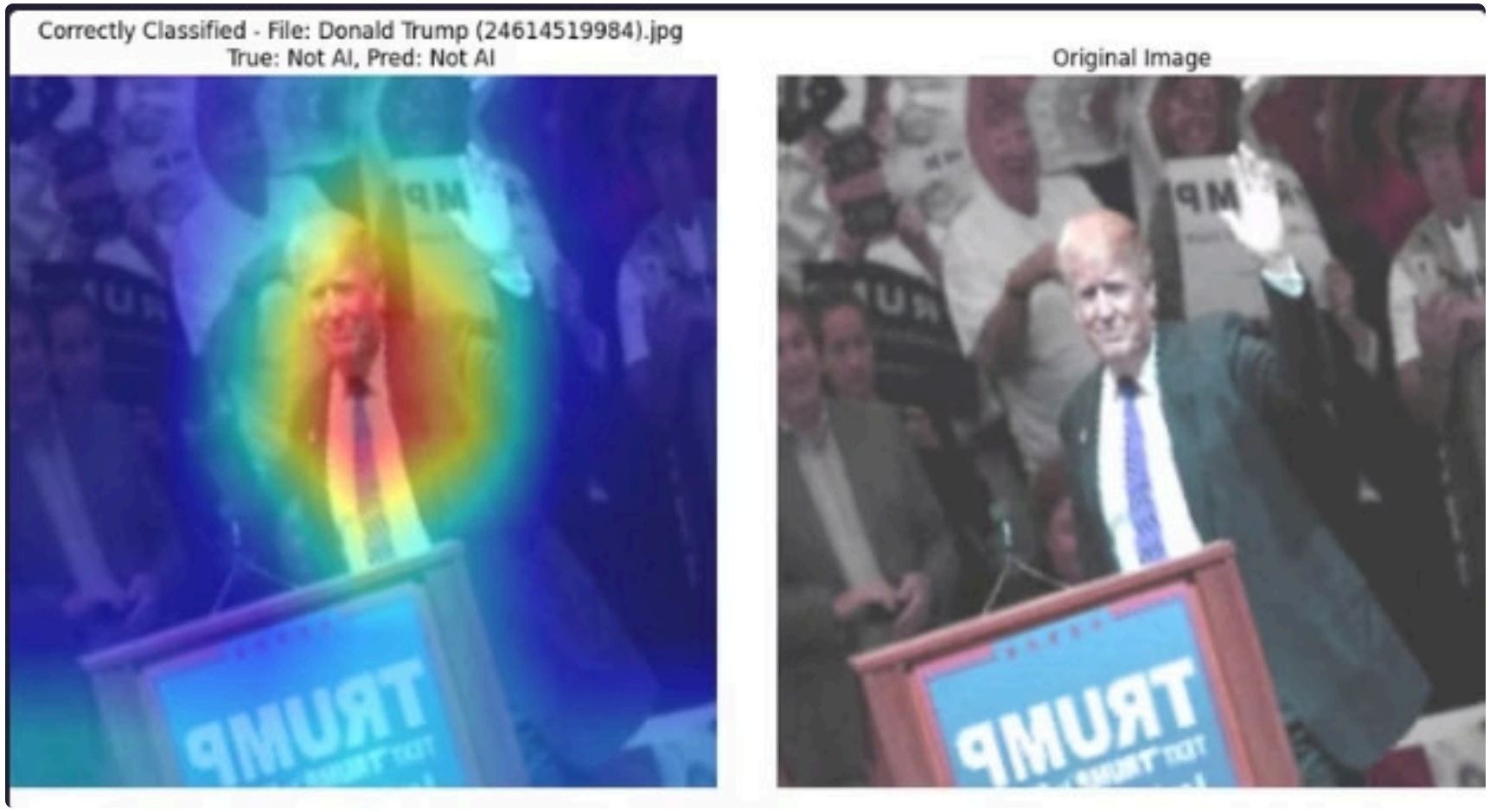
| Class | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| Real Images | 0.95 | 0.98 | 0.96 | 409 |
| AI Images | 0.97 | 0.94 | 0.95 | 357 |
| **Weighted Average** | **0.96** | **0.96** | **0.96** | **766** |

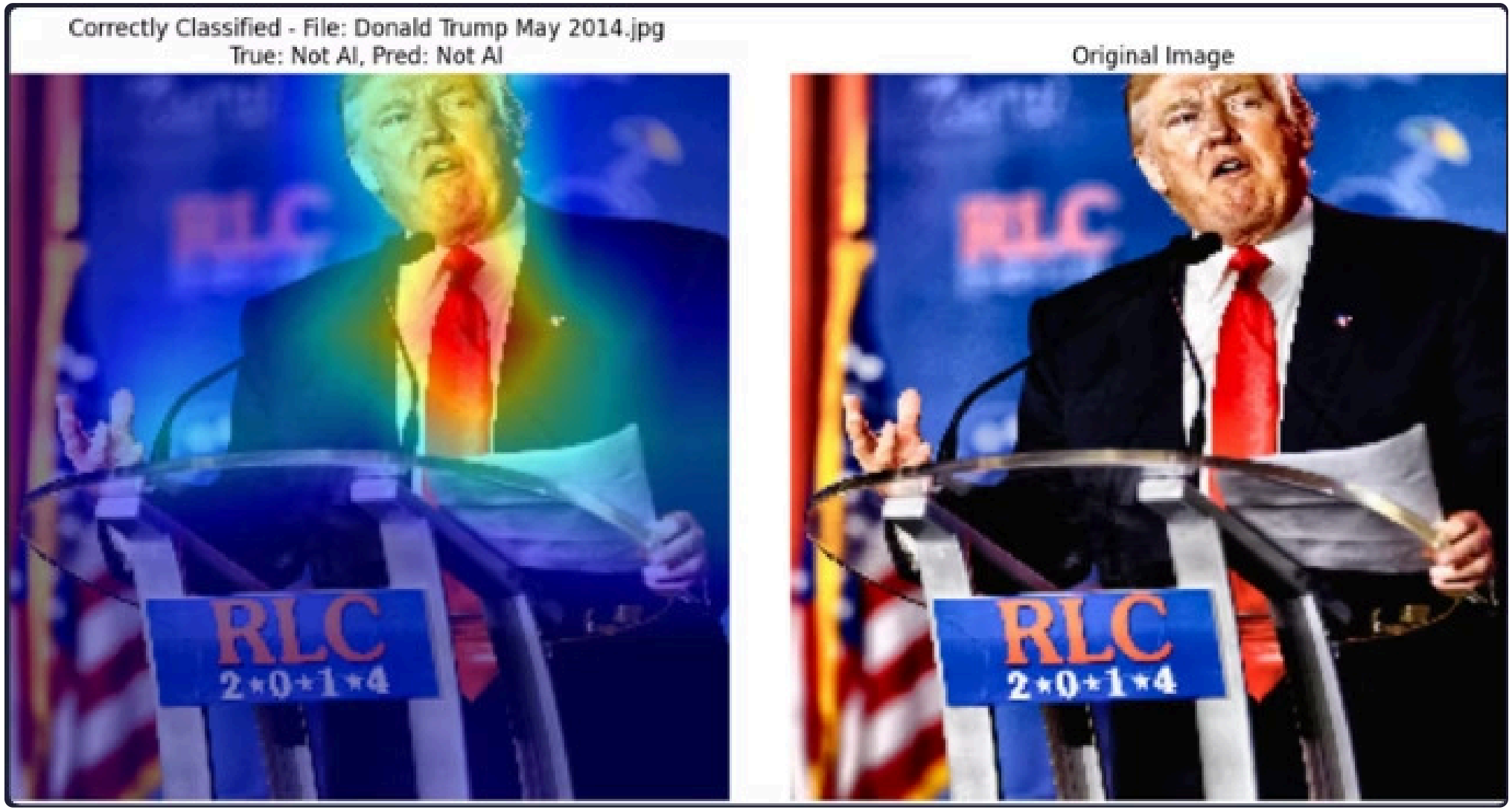
The model demonstrates robust discriminative capability with balanced performance across both classes, suggesting viability for deployment in election integrity verification systems.

# AI images - focus on border

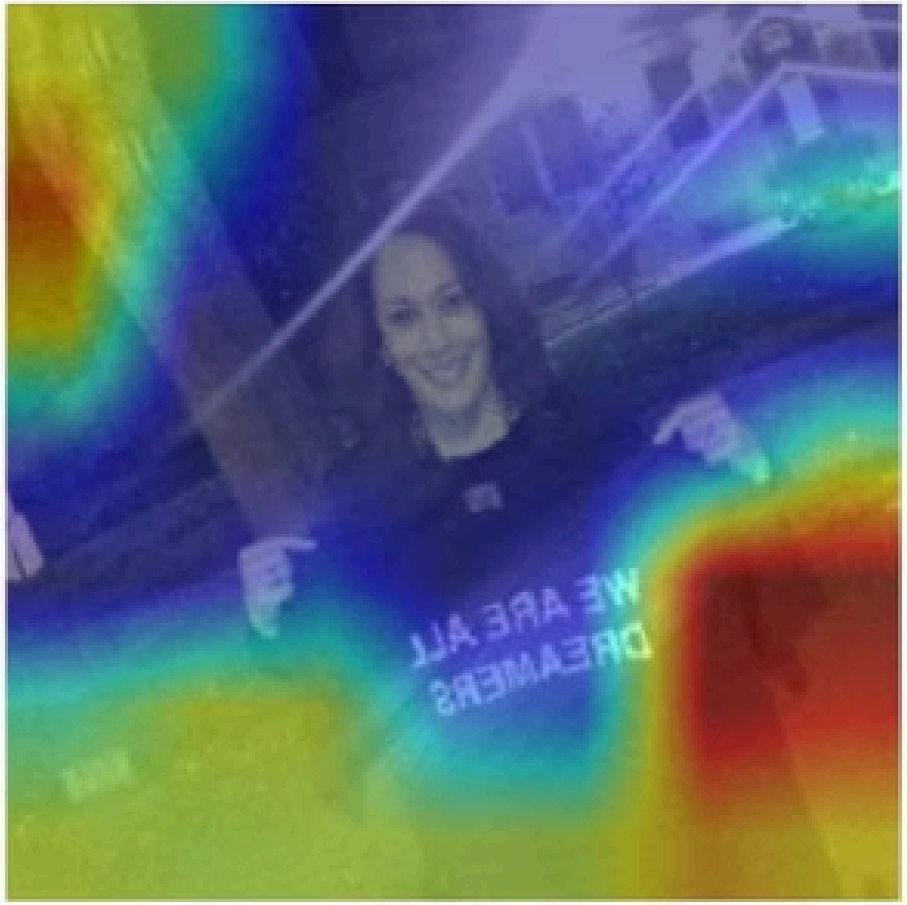


# Real images - focus on center





Misclassified - File: Kamala Harris with We Are All Dreamers t-shirt.jpg
True: Not AI, Pred: AI




Original Image


Misclassified - File: Governor Tim Walz at Vikings Traing Camp on 30 July 2024 02.jpg

# Discussion - Conclusion

- Based on our results, EfficientNet-B0 demonstrates strong performance in detecting AI-generated images
- Images that are (partially) edited with tools like Photoshop are harder to detect
- Heatmap analysis: model looks at borders to detect synthetic images and the center for real ones → ideas for improving detection methods
- The dataset is small and focused → future expansion