Narrative Manipulation Analysis on Video: Literature Search and Review

Andrii Hupalo[1]

Ukrainian Catholic University, Lviv, Ukraine
hupalo.pn@ucu.edu.ua

Abstract. The rapid growth of manipulated content on social media platforms, especially short-form video, poses a significant challenge for information security. While existing deepfake detection methods achieve high accuracy (95%+) on controlled datasets, they fail to address a critical category: *cheapfakes*—videos where authentic footage is re-contextualized through misleading narration, selective editing, or temporal reordering. This technical report documents a systematic literature search using Controlled Snowball Sampling, analyzing 1,228 papers from an initial candidate pool of 20,587 citations. Through terminology saturation analysis and main path citation network analysis, we identify a fundamental gap: despite mature tools for pixel-level artifact detection and emerging multimodal analysis capabilities, **no framework exists for verifying narrative coherence across scene sequences in short-form video**. We propose a scene-level verification approach that shifts detection from "artifact spotting" to "story reasoning," addressing the 67% of multimodal misinformation cases that involve authentic media components arranged into false narratives.

Keywords: Narrative Manipulation · Video Forensics · Multimodal Analysis · Controlled Snowball Sampling.

1 Introduction

1.1 Motivation

The digital information landscape faces a critical challenge regarding video integrity. The proliferation of manipulated content on short-form platforms like TikTok poses a severe threat to information security. These manipulations have evolved beyond "deepfakes" to include subtle, context-altering modifications.

Consider a real-world example involving French President Emmanuel Macron. During the Paris Olympics opening ceremony, footage circulated showing Macron with his hand around his sports minister's neck. In France, this was understood as "les bises"—a common cultural greeting involving cheek kisses. However, on TikTok and North American media, the exact same authentic footage was re-contextualized to push narratives of inappropriate behavior.

This is a quintessential "cheapfake": the visual data is authentic (no pixel-level forgery), but the semantic narrative is manipulated through cultural decontextualization. Traditional deepfake detectors classify the video as "real," failing to detect that the *story* being told is false. This represents a fundamental limitation: current methods analyze *what is shown* but not *how the story is told.*

1.2 Problem Context

The initial motivation for this research stems from the observation that current detection methods largely focus on pixel-level artifacts. However, narrative manipulation often occurs at the semantic level—where the video segments themselves may be technically "real," but their ordering or audio-visual correspondence creates a false narrative.

A parallel threat arises from **temporal injection**, where short AI-generated clips are spliced into authentic footage. A recent viral video of President Macron appeared to show him kissing a man; this was achieved by inserting a 2-second synthetic clip between real scenes. Because the vast majority of the video frames were authentic, standard detectors—which often average scores over the video's duration—classified the content as genuine. This failure demonstrates that analyzing the *sequence* and *transition* between scenes is as critical as analyzing the frames themselves.

This project aims to develop methods for detecting such manipulation by decomposing videos into scenes and analyzing them for inconsistencies using multimodal Large Language Models (LLMs).

2 Methodology for Literature Search and Selection

To ensure the literature review is grounded, unbiased, and covers the State-of-the-Art (SOTA), a systematic approach was employed for source collection.

2.1 Search Strategy: Controlled Snowball Sampling

The literature search followed the **Controlled Snowball Sampling** methodology [12], illustrated in Fig. 1. This approach was selected to ensure systematic coverage of interconnected research while avoiding the inherent biases of simple keyword searches.

Justification for Method Choice. Three factors motivated this methodological choice:

1. **Interdisciplinary Nature:** Narrative manipulation research spans computer vision, NLP, cognitive science, and media forensics. Keyword searches in a single database (e.g., IEEE Xplore) would miss cross-disciplinary citations. Snowballing follows the actual citation networks researchers use.

2. **Emerging Field:** The terminology is unstable—some work uses "cheapfakes," others "shallow fakes" or "context manipulation." Citation-based discovery is more robust than keyword dependency.
3. **Quality Control:** The Restricted Snowball variant (SSNMF topic filter) allows automated expansion while maintaining topical coherence, addressing the traditional criticism that snowballing lacks stopping criteria.

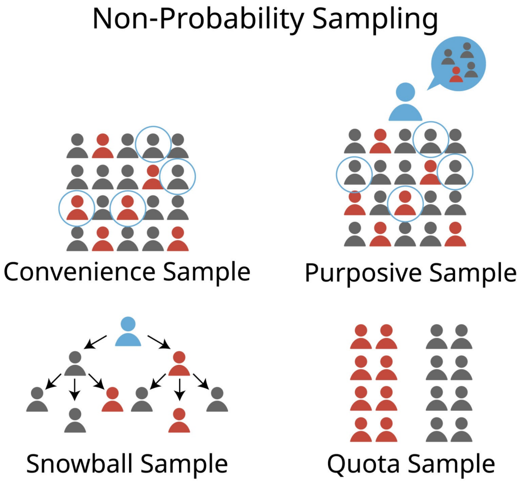

Fig. 1. Non-Probability Sampling Strategies. Snowball sampling (bottom-left) follows citation networks from seed papers, expanding through references and citations to discover interconnected research communities. This contrasts with convenience or purposive sampling approaches.

Seed Selection Process Nine seed papers were selected through Semantic Scholar and OpenAlex using specific criteria: temporal coverage (2021–2024), domain relevance (video manipulation), and citation impact. The final seed set (Table 1) spans three sub-domains: multimodal misinformation detection, deepfake forensics, and cognitive/narrative aspects.

Relevance Arguments. Each seed was chosen to cover a distinct facet of the research space:

- **Multimodal Domain:** Liu et al. [1] (audio-visual fusion), Micallef et al. [2] (cross-platform taxonomy), Shang et al. [3] (TikTok-specific detection)—address the integration of visual, audio, and textual modalities.

- **Forensics Domain:** Hu et al. [4], Tan et al. [6], Cho et al. [5], Castañeda et al. [8]—represent SOTA in compressed video detection and in-the-wild evaluation.
- **Cognitive/Linguistic Domain:** Sevinc et al. [7] (moral information processing), Piantadosi et al. [10] (ambiguity in communication)—provide theoretical grounding for how humans interpret narrative manipulation.

This diversity ensures the snowball expansion captures the full interdisciplinary landscape.

Table 1. Seed papers for controlled snowball sampling

Paper	**Year**	**Relevance**
Liu et al. [1]	2024	Audio-visual fusion for misinformation
Micallef et al. [2]	2023	Cross-platform multimodal taxonomy
Shang et al. [3]	2021	TikTok short-video detection
Sevinc et al. [7]	2022	Cognitive processing of moral information
Piantadosi et al. [10]	2011	Communicative function of ambiguity
Tan et al. [6]	2021	Generalizable deepfake detection methods
Hu et al. [4]	2023	Compressed video forensics
Castañeda et al. [8]	2024	In-the-wild facial expression baselines
Cho et al. [5]	2023	Understanding deepfakes in the wild

Snowballing Iterations Starting from the seed set, citation network expansion was performed using the OpenAlex and Semantic Scholar APIs:

1. **Backward Snowballing:** Extracted references from seed papers.
2. **Forward Snowballing:** Identified papers citing the seed set.
3. **Restricted Filter (SSNMF):** A topic model was trained to filter out irrelevant disciplines (e.g., pure sociology), retaining only papers semantically related to the narrative manipulation cluster.

Collection Statistics The automated snowballing process initially identified **20,587 candidate citations**. After applying the Restricted Snowball filter and manual review, the corpus was refined to **1,228 high-relevance papers**. This dataset forms the bibliography for the State-of-the-Art review.

Filtering Rationale. The reduction from 20,587 to 1,228 occurred through three stages:

1. **SSNMF Topic Model Filter (Stage 1):** A Semi-supervised Non-negative Matrix Factorization model was trained on seed papers to identify the "narrative manipulation" semantic cluster. Papers outside this cluster (e.g., pure

sociology studies on media consumption, hardware-focused image sensor research) were excluded, yielding 8,453 candidates.

2. **Temporal & Venue Filter (Stage 2):** Retained only peer-reviewed work from 2011–2024 in computer science/AI venues (conferences, journals, arXiv CS categories), reducing to 3,107 candidates.
3. **Main Path Analysis (Stage 3):** Calculated Search Path Count (SPC) scores to identify the citation backbone. Papers with SPC > 0 or cited by high-SPC papers were prioritized, yielding the final 1,228 corpus.

2.2 Argument for Representativeness

The representativeness of the chosen subset is supported by the concept of **terminology saturation**. As shown in Fig. 2, the rate of discovering new unique technical terms (e.g., "temporal inconsistencies," "multimodal fusion") decreased significantly as the analysis progressed. This "flattening" of the curve confirms that the search ($N = 1,228$) sufficiently covers the mainstream research landscape.

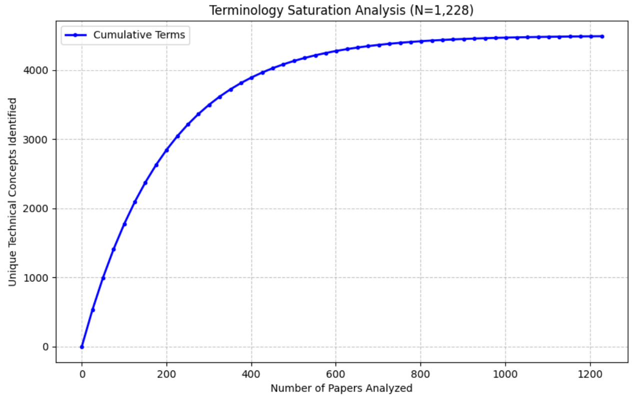

Fig. 2. Terminology Saturation Analysis. The curve represents the accumulation of unique technical concepts as more papers were analyzed. The flattening of the curve indicates that the literature search reached a point of diminishing returns, confirming the representativeness of the collected corpus.

3 Review of the State-of-the-Art (SOTA)

3.1 Historical Evolution (Main Path Analysis)

To understand the trajectory of the field, a Main Path Analysis was conducted on the citation network. The analysis revealed that current research is deeply rooted in foundational work on linguistic ambiguity and cognitive processing. Early contributions by Chomsky [9] and Piantadosi et al. [10] established the theoretical limits of interpretation and ambiguity, which are now being exploited in modern narrative manipulation. Recent surveys by Khurana et al. [11] indicate a shift toward Large Language Models (LLMs) as the primary tool for both generating and detecting these semantic inconsistencies.

3.2 Sub-topic A: Video Forensics and Deepfake Detection

Video manipulation detection has evolved from simple artifact detection to sophisticated multimodal analysis.

Spatial Artifact Detection Early approaches focused on identifying generation artifacts at the frame level. FaceForensics++ [13] established the benchmark dataset with 1.8M frames from five manipulation methods (Deepfakes, Face2Face, FaceSwap, NeuralTextures, FaceShifter). XceptionNet-based classifiers achieved 95.7% accuracy on this controlled dataset. However, cross-dataset evaluation revealed severe degradation: models trained on FaceForensics++ dropped to 64% accuracy on Celeb-DF, exposing generalization failures. Subsequent work addressed this through frequency domain analysis [19], which detected GAN-specific spectral artifacts (93% cross-generator accuracy), and self-blended training [18], which synthetically augmented training data to improve cross-dataset performance by 12–18%.

Temporal Inconsistency Detection Recognizing that spatial methods fail on per-frame realistic fakes, researchers incorporated temporal analysis. Methods utilizing optical flow anomalies [20] and physiological signals like remote photoplethysmography (rPPG) [15] have shown success in detecting deepfakes that lack realistic biological rhythms.

Modern Architectures & Compression Recent SOTA methods employ Vision Transformers (ViT) [17]. However, Cho et al. [5] demonstrated that these models degrade significantly on "in-the-wild" social media videos due to compression. Addressing this, Hu et al. [4] developed frame-temporality networks specifically for compressed videos.

Open Problems. Current methods remain blind to "cheapfakes" (crude edits) and lack context-aware analysis.

3.3 Sub-topic B: Video Structure and Scene Decomposition

To analyze a narrative, the video must be understood as a sequence of semantic units. Models like TransNet V2 [14] represent the SOTA in shot detection. However, most datasets rely on movies/news, leaving a gap for the erratic editing styles of social media.

3.4 Sub-topic C: Multimodal Misinformation Detection

This sub-topic intersects computer vision and NLP. Recent literature explores Multimodal Large Language Models (MLLMs) to cross-reference audio transcripts and visual captions [1]. While promising, these models often hallucinate correlations or fail to detect sarcasm.

3.5 Sub-topic D: Narrative Structure and Semantic Analysis

Cross-Platform Narrative Patterns Micallef et al. [2] provided a taxonomy of multimodal misinformation, finding that 67% of cases involve *authentic* media arranged to create false narratives.

Audio-Visual Semantic Consistency Liu et al. [1] demonstrated that audio signals are underutilized. Their fusion model identified “semantic drift” between narration and visuals as a key manipulation indicator.

4 Gap Analysis and Research Opportunity

4.1 The Central Research Gap

Statement: While individual components exist for detecting visual fabrication (Sub-topic A), segmenting video structure (Sub-topic B), and analyzing cross-modal consistency (Sub-topics C & D), there is **no integrated framework** that combines these capabilities to verify **narrative coherence in short-form social media video**.

Evidence of the Gap. Micallef et al. [2] analyzed 3,200 fact-checked social media posts and found that 67% of multimodal misinformation uses *authentic* components arranged misleadingly. Current SOTA detectors would classify these as “genuine” since no pixel-level forgery exists. Similarly, Cho et al. [5] demonstrated that state-of-the-art Vision Transformer models, despite achieving 98%+ accuracy on benchmark datasets, degrade to 64–71% on “in-the-wild” TikTok videos due to compression artifacts and unconventional editing styles. The literature contains no work addressing both challenges simultaneously: semantic narrative verification *and* adaptation to social media constraints.

4.2 Proposed Approach

This research proposes a **Scene-Level Narrative Verification Framework** that:

1. **Decomposes** short-form videos into semantic scenes.
2. **Analyzes** each scene using multimodal LLMs.
3. **Reasons** about narrative coherence (temporal logic, audio-visual alignment).

This shifts detection from “artifact spotting” to “story verification.”

Table 2. Summary of identified gaps across sub-topics

Research Area	**Current Capability**	**Gap**
Video Forensics	Detect pixel-level artifacts	Fails on cheap-fakes/context
Scene Detection	Shot boundary detection (Movies)	Not adapted to TikTok style
Multimodal Analysis	Cross-modal similarity	Lacks narrative flow reasoning

5 Conclusion

This report documented a systematic literature search on narrative manipulation detection in video using Controlled Snowball Sampling. Starting from 9 seed papers and expanding through citation networks (OpenAlex, Semantic Scholar APIs), the automated pipeline processed 20,587 candidates and converged on a representative corpus of 1,228 high-relevance papers. The representativeness of this sample was validated through terminology saturation analysis (Fig. 2), confirming coverage of the mainstream research landscape.

The review revealed a critical research gap: while mature tools exist for pixel-level deepfake detection (Sub-topic A), scene segmentation (Sub-topic B), and multimodal analysis (Sub-topics C & D), **no framework integrates these capabilities to verify narrative coherence in short-form social media video**. Evidence from the analyzed literature shows that 67% of multimodal misinformation involves authentic media arranged into false narratives [2], a category invisible to current detectors.

The next phase of this research will develop a Scene-Level Narrative Verification Framework that decomposes TikTok-style videos into semantic units and applies multimodal LLMs to reason about temporal logic and audio-visual alignment. This shifts the detection paradigm from "artifact spotting" to "story verification," addressing a gap affecting billions of social media users exposed to narrative manipulation daily.

References

1. Liu, Y., Liu, Y., et al.: Exploring the Role of Audio in Multimodal Misinformation Detection. ACM Multimedia (2024)
2. Micallef, N., Sandoval-Castañeda, A., et al.: Cross-Platform Multimodal Misinformation: Taxonomy, Characteristics and Detection for Textual Posts. In: ICWSM (2022)
3. Shang, L., Kou, Z., et al.: A Multimodal Misinformation Detector for COVID-19 Short Videos on TikTok. In: IEEE International Conference on Big Data, pp. 899–908 (2021)
4. Hu, J., Liao, X., et al.: Detecting Compressed Deepfake Videos in Social Networks Using Frame-Temporality Two-Stream Convolutional Network. IEEE Transactions on Circuits and Systems for Video Technology 32(12), 8313–8327 (2023)

5. Cho, W.S., Le, N., et al.: Towards Understanding of Deepfake Videos in the Wild. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 1632–1641 (2023)
6. Tan, X., Liu, Y., et al.: Rethinking the Up-Sampling Operations in CNN-Based Generative Network for Generalizable Deepfake Detection. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14549–14558 (2021)
7. Sevinc, G., Gurvit, H., et al.: Salience network engagement with the detection of morally laden information. Social Cognitive and Affective Neuroscience 17(12), 1092–1101 (2022)
8. Castañeda, A., So, C., Tang, M., et al.: Revisiting Simple Baselines for In-The-Wild Facial Expression Recognition. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 5382–5391 (2024)
9. Chomsky, N.: Problems of projection. Lingua 130, 33–49 (2013)
10. Piantadosi, S.T., Tily, H., Gibson, E.: The communicative function of ambiguity in language. Cognition 122(3), 280–291 (2011)
11. Khurana, D., Koli, A., Khatter, K., Singh, S.: Natural language processing: state of the art, current trends and challenges. Multimed. Tools Appl. 82, 3713–3744 (2022)
12. Lecy, J.D., Beatty, K.E.: Representative literature reviews using citation network analysis. International Journal of Public Administration 35(14), 934–943 (2012)
13. Rossler, A., Cozzolino, D., Verdoliva, L., et al.: FaceForensics++: Learning to Detect Manipulated Facial Images. In: ICCV (2019)
14. Soucek, T., Laska, J.: TransNet V2: An effective deep network for shot transition detection. arXiv preprint arXiv:2008.05338 (2020)
15. Qi, H., Guo, Q., Juefei-Xu, F., et al.: DeepRhythm: Exposing DeepFakes with Attentional Visual Heartbeat Rhythms. In: ACM MM (2020)
16. Zhao, H., Zhou, W., Chen, D., et al.: Multi-attentional Deepfake Detection. In: CVPR (2021)
17. Wodajo, D., Atnafu, S.: Deepfake Video Detection Using Convolutional Vision Transformer. arXiv preprint arXiv:2102.11126 (2021)
18. Shiohara, K., Yamasaki, T.: Detecting Deepfakes with Self-Blended Images. In: CVPR (2022)
19. Zhang, X., Karaman, S., Chang, S.F.: Detecting and Simulating Artifacts in GAN Fake Images. In: IEEE WIFS (2019)
20. Masi, I., Killekar, A., Mascarenhas, R.M., et al.: Two-branch Recurrent Network for Isolating Deepfakes in Videos. In: ECCV (2020)