The Synthetic Media Shift: Tracking the Rise, Virality, and Detectability of AI-Generated Multimodal Misinformation

Supplementary Material

Stage	Entries
Notes labeled misleading	1,806,168
Image keyword matches	112,629
Video keyword matches	133,682

Table 4. Filtering stages for multimodal notes.

Modality	Note–Post pairs
Images	66,135
Videos	86,131

Table 5. Final dataset size after post retrieval.

10. Dataset Collection and Filtering

10.1. Community Notes Source

CONVEX is constructed from the publicly available *X Community Notes* corpus, which provides periodic releases of notes, ratings, and metadata.[4] We collect all available records until January 2026 using the official dataset snapshot released on January 6, 2026. The raw corpus contains 2,316,374 notes. Following the methodology described in Sec. 3 of the main paper, we retain only entries labeled by the Community Notes system as *misinformed or potentially misleading*, resulting in 1,806,168 notes. Each note is associated with a specific X post.

10.2. Identifying Multimodal Notes

Community Notes include a field (`isMediaNote`) indicating whether a note refers to media. However, this field is sparsely populated (≈5.6%) and does not distinguish between images and videos. We therefore use keyword-based filtering over note text to identify candidate multimodal entries. We use two modality-specific keyword groups.

Image modality keywords: `photo, image, photograph, screenshot, pic, photoshop, photoshopped, picture, snapshot, visual, graphic, thumbnail, logo, jpg, jpeg, png`

Video modality keywords: `video, clip, footage, recording, deepfake`

Applying these filters yields 112,629 candidate notes containing image-related keywords and 133,682 containing video-related keywords. The detailed filtering counts are summarized in Table 4.

We then retrieve the corresponding posts and attached media from X using the `twikit` library. Posts that were deleted, suspended, or otherwise inaccessible at the time of collection are excluded. For the image subset, we further retain only entries where the retrieved post contains at least one image. The final dataset contains 66,135 image note–post pairs and 86,131 video note–post pairs, as summarized in Table 5. For each entry, we collect the post text, associated media files, author metadata, and engagement statistics (favorites, reposts, replies, and views). The resulting dataset is organized at the note–post pair level.

[4] https://x.com/i/communitynotes/download-data

11. Notes Misinformation Annotation

We annotate the dataset using a hybrid weakly supervised approach that combines keyword-based rules with a vision-language model (VLM).

11.1. Keyword-based Classification

We define category-specific textual patterns and apply them to Community Notes text to produce candidate labels. Matching is case-insensitive and accounts for common lexical variations. If no pattern is matched, no label is assigned at this stage and the instance is deferred to the VLM. Table 6 summarizes the keyword and pattern groups. Patterns include both explicit references to generative models and descriptions of synthetic appearance.

11.2. VLM-based Classification

Keyword rules alone are insufficient to cover all cases; therefore, we combine them with VLM predictions to label instances without explicit lexical cues. We use the instruction-tuned Gemma 3 model (`gemma-3-27b-it`) through the Google GenAI library to perform zero-shot classification. The model receives as input the post text and the Community Note text, together with associated images when available. When multiple images are present, they are provided jointly to the model. For entries without images (e.g., video-only posts), classification is based on textual context alone. The model classifies each entry into one of the predefined misinformation categories. The `other` label is produced only when the misinformation type cannot be determined or is not media-related. Model inference is performed with temperature set to 0.1 and a maximum output length of 512 tokens (Fig. 7).

When the keyword-based label and the VLM prediction disagree, the model is re-run with additional context that

Category	Patterns	Negation rules
Miscaptioned	out-of-context, wrong context, miscaptioned, misleading, old/previous photo or video, reused/recycled photo or video, not from this event/day/place/country, actually/originally/first posted/circulating since YEAR, photo/image/video/footage from/taken/filmed/recorded/shot YEAR	caption is accurate, context is correct
Edited	photoshop/photoshopped, doctored, manipulated, tampered, fabricated, edited, digitally altered, composite, spliced, superimposed, object/person added or removed, face swap, head swap, stitched, merged clip/video, video edited, cropped, airbrushed, retouched	not photoshopped, not edited, not manipulated, not doctored, not altered, unaltered/unedited/original image/video/photo/footage
AI-generated	AI-generated, AI created, synthetic image/video, generated by AI, created with AI, diffusion model, GAN, Midjourney, DALL-E, Runway Gen-2/Gen-3, Sora, Veo, machine generated, looks AI/synthetic, deepfake	not AI-generated, not created by AI, not made with AI, real/authentic/genuine image/video/photo/footage

Table 6. Keyword and pattern groups used for weakly supervised misinformation-type annotation.

includes both predictions. The model is then asked to reconsider the classification and produce a final label using the same output format.

11.3. Label Refinement and Final Classification

Final labels are assigned using the following voting procedure:

1. Apply keyword-based classification.
2. If no keyword label is assigned, the final class is given by a single VLM prediction.
3. Otherwise, run the VLM to obtain an independent prediction.
4. If the keyword label and VLM prediction agree, that label is retained.
5. If they disagree, the VLM is executed a second time with both predictions provided as additional context, and the final class is determined by majority voting across the keyword label and the two VLM predictions.

We analyze the behavior of the weakly supervised annotation pipeline on the full multimodal dataset (image and video subsets combined). Keyword rules are designed to be high-precision but sparse, assigning a label only when explicit linguistic patterns are present in the Community Note. As a result, 73.7% of entries do not match any predefined pattern and are handled solely by the VLM. Among entries where keyword rules assign a label, they agree with the first-pass VLM prediction in 87.9% of cases, indicating high reliability when triggered. When the keyword label and first-pass VLM prediction disagree, a second VLM pass is applied. In these cases, the second-pass prediction aligns with the first-pass VLM in 81.1% of cases, with the keyword label in 9.0%, and produces a different or unresolved output in 9.9% of cases. Overall, the VLM provides labels for the majority of entries not covered by keyword rules, resulting in a final distribution of 67.34% miscaptioned, 15.08% edited, 14.26% AI-generated, and 3.17% *other*.

	Images		Videos	
Type	Count	%	Count	%
Miscaptioned	39,820	60.21%	65,095	75.58%
Edited	15,397	23.28%	8,030	9.32%
AI-generated	10,785	16.31%	11,060	12.84%
Other	133	0.20%	1,946	2.26%

Table 7. Misinformation type distribution across image and video subsets.

11.4. Dataset Statistics

The final dataset contains 66,135 image entries and 86,131 video entries. The resulting class distributions of misinformation categories are summarized in Table 7.

12. AI-Related References

As described in Section 7 of the main paper, we extract textual references from Community Notes that indicate potential AI-generated content, capturing explicit mentions to generative systems and synthetic media. Signals are extracted from the Community Note text only. We define two types of AI-generation references.

Model and tool mentions. We detect references to commonly mentioned AI tools and model names in Community Notes, including: *ChatGPT, OpenAI, Gemini, Bard, Claude, Perplexity, LLaMA, Grok, DeepSeek, Mistral, Qwen, Copilot, Midjourney, DALL-E, Stable Diffusion, Sora, Veo*.

General AI-generation references. We identify phrases that explicitly describe content as generated by AI. These references require both (i) a generation verb (e.g., *generated, created, made, produced, synthesized*) and (ii) an explicit reference to AI (e.g., *AI, artificial intelligence*). Both conditions must be satisfied within the same textual span.

Prompt template for VLM classification
You are a misinformation analyst reviewing a post on X. You will see the post text, one or more media files and a Community Note written by other users. Your job is to classify the TYPE of media-related misinformation. Use exactly ONE label from this set for the media:
ai_generated — The image or video is fully or partly created by an AI system (e.g. generated by AI, diffusion model, deepfake where the person or scene is synthetic; tools like Midjourney, DALL·E, Sora).
edited — The image or video is based on real footage but has been digitally manipulated (e.g. photoshopped, doctored, digitally altered, composite, splice, superimposed, face swap, objects added or removed, misleading crop or heavy visual edit).
miscaptioned — The image or video itself is essentially authentic and not significantly edited, but the post gives a false or misleading context (e.g. out of context, wrong context, miscaptioned, misattributed, old photo or video reused as if it were new, wrong time, wrong place, wrong event, wrong person, wrong causal link).
other — The Community Note discusses misinformation that is not mainly about the media or there is not enough evidence to decide which of the three media types applies.

Instructions
• Treat the Community Note as the primary evidence: it usually explains why the post is misleading.
• Use the post text and the media to cross-check and support or reject what the note says.
• If multiple labels seem possible, choose the ONE that best describes the main way the media misleads
• If there is not enough information to be sure, choose 'other

Output format (JSON)
{
"misinfo_label": "ai_generated — edited — miscaptioned — other",
"confidence": ¡0–1¿,
"rationale": "1–2 sentence explanation"
}
Do not add any extra text before or after the JSON

Inputs
COMMUNITY_NOTE: {note_text}
POST_TEXT: {full_text}
MEDIA: {num_media}

Figure 7. Prompt template used for VLM-based misinformation classification.

Mentions are searched using case-insensitive pattern-based matching, covering common lexical variations (e.g., hyphens, spacing, and version suffixes such as GPT-4, GPT4, GPT-4o).

13. AI Detection Evaluation Setup

We evaluate two classes of AI-image detectors: Synthetic Image Detectors (SIDs) and vision-language models (VLMs). All models are evaluated on the same set of images in a zero-shot setting. Specialized SIDs operate directly on images using their pretrained weights and default inference pipelines. For VLMs, we provide the image together with a fixed prompt instructing the model to classify it as `AI` or `REAL`, without any additional textual context (e.g., post text or Community Note text). The same prompt is used across all VLMs (see Fig. 8).

Specialized detectors. We use SPAI [5], RINE [6], and B-Free [7], which take an image as input and output a scalar score indicating the likelihood that it is AI-generated. All SIDs are used via their official implementations and publicly released pretrained weights. For RINE we utilize the itw_rine_mever variant through the official API service [8].

Vision-language models. We evaluate Gemma 3 27B (`gemma-3-27b-it`), Grok (`grok-4-1-fast-non-reasoning`), and GPT (`gpt-5-mini`). They are accessed through their official APIs (Google GenAI, xAI, and OpenAI). GPT and Grok are executed in non-reasoning configurations to encourage consistent binary outputs. Gemma 3 is run with temperature 0.1, Grok with temperature 0, and GPT with minimal reasoning effort and low verbosity.

[5] https://mever-team.github.io/spai/
[6] https://github.com/mever-team/rine
[7] https://github.com/grip-unina/B-Free
[8] https://docs.mever.gr/deepfake

Prompt for VLM-based AI detection
Task You are an expert in detecting AI-generated images. You need to classify the image as `AI` or `REAL`.
Definitions
AI — the main depicted content was generated or substantially modified by a generative AI model.
REAL — the main depicted content is authentic (e.g. photographed) and was not generated or substantially modified by a generative AI model.
Instructions
• If the image is a screenshot or digital capture, classify based on whether the depicted content was generated or substantially modified by generative AI.
• Ignore minor edits (cropping, compression, color correction).
• Return exactly one token: `AI` or `REAL`.
• Do not include any explanation or punctuation.

Figure 8. Prompt used for VLM-based AI image classification.