Codebook

ThumbnailTruth: A Multi-Modal LLM Approach for Detecting Misleading YouTube Thumbnails Across Diverse Cultural Settings

1. Purpose & Scope

This codebook defines how to label YouTube thumbnails as Misleading Thumbnail Video (MTV) or Non-Misleading Thumbnail Video (NMTV) by inspecting the thumbnail together with the video's title and actual content.

We assign the following (binary) label: MTV vs NMTV. We don't label overall video quality, political leaning, medical accuracy, or moral judgments unless they directly pertain to the thumbnail's truthfulness relative to the video content.

2. Operational Definitions

2.1 Misleading Thumbnail (MTV)

A thumbnail is misleading if it materially misrepresents the video's central topic, facts, or outcomes, leading a reasonable viewer to a false or significantly distorted expectation before playback.

A thumbnail is MTV if any one of the following is true:

1. Bait-and-Switch / Thematic Mismatch
   The thumbnail's depicted subject/event is not actually present or not central to the video.
2. False Promise / Fabricated Outcome
   The thumbnail asserts or strongly implies a specific factual outcome that does not occur in the video.
3. Fabricated or Out-of-Context Imagery
   The thumbnail uses doctored images or real images in false context (e.g., old image presented as today's event) without disclosure and contrary to the video's content.
4. Severe Exaggeration Presented as Fact
   Hyperbolic visual/text claims framed as factual (not clearly comedic/satirical) that are unsupported by the video.
   Note: Obvious satire/parody or genre conventions (e.g., gaming "INSANE!" faces) are not MTV if the central claim remains accurate and the exaggeration is clearly performative.

2.2 Non-Misleading Thumbnail (NMTV)

A thumbnail is non-misleading if it accurately reflects the video’s central topic and does not promise outcomes the video fails to deliver. Minor puffery (e.g., excited faces, emojis) is acceptable if the core topic and key facts match the video content.

3. Decision Rules & Borderline Handling

1. Presence: Are the people/objects/events depicted or promised in the thumbnail actually present and central in the video?
   - If No → label MTV.
2. Claims: Does the video deliver the key outcome the thumbnail promises?
   - If No → label MTV.
3. Authenticity/Context: Is the imagery authentic and correctly contextualized?
   - If No → label MTV.

If all three are Yes, label NMTV.

3.2 Genre Conventions

- Allowed: Emotive faces, bold text, arrows, emojis typical for gaming/DIY if the main topic and outcome match.
- Not allowed: Using these to imply events that never occur (e.g., pointing an arrow at a celebrity who never appears).

3.3 “Minor Exaggeration” Rule

If exaggeration is stylistic and the video covers the same central topic, label NMTV.

4. Multilingual Guidance

For thumbnails, titles, or subtitles in non-English languages, use a consistent translation tool (e.g., Google Translate). Make labeling decisions based on the English translation, using the same MTV vs. NMTV criteria applied to English-language content.

5. Annotation Procedure

1. Review materials:
   - ○ Thumbnail image, video title, full video if ~5 minutes and less (or skim if longer)).
2. Annotator Form (per item):
   - ○ Primary label: MTV / NMTV
3. Borderline resolution:
   - ○ Apply the Three-Question Test and Minor Exaggeration Rule.
4. Adjudication:
   - ○ Two independent annotators label each item. Only consensus items are included in the final dataset.

6. Quality Control

- Calibration round: 20 items jointly labeled to refine consistency.
- Final agreement: Report Kappa and only include consensus-labeled videos.

7. Examples

- MTV: Thumbnail shows a celebrity "EXCLUSIVE INTERVIEW"; celebrity never appears in the video.
- MTV: Thumbnail says "DOUBLE YOUR MONEY TODAY"; video never provides such content.
- MTV: AI-generated image of a politician signing a law presented as today's news; no such footage in video.
- NMTV: Over-the-top face + "I BROKE THE GAME!"; video demonstrates actual game-breaking bug.
- NMTV: Parody clearly labeled in thumbnail and video, consistent with content.

8. Dataset Inclusion/Exclusion

- Include: Public video urls accessible at labeling time.
- Exclude: Removed/unavailable/age-gated content, or videos with severe technical playback issues.