HybridNet: Efficient Multimodal Fake News Detection

Supplementary Material

8. Detailed Prompts for MLLM Guidance

We provide abbreviated versions of the exact system prompts used in our zero-shot MLLM pipeline. These instructions are designed to enforce a deterministic structure, facilitating the extraction of reasoning embeddings for the RAC module.

8.1. Stage 1: Modal Consistency Prompts

Text-Text Consistency ($r_{txt\text{-}txt}$)

System Role: You are a factual consistency evaluator.
Task: Determine if the following two sentences describe the same real-world facts.
Scoring Rules:
- `1`: Perfect alignment (entities, actions, time, context).
- `0`: Different facts or loosely related.
- `-1`: Explicit debunking by reputable fact-checker.

Output Format (JSON):

```
{"FactualAlignmentScore": <score>,
"rationale": "<justification>",
"FactCheckVerdictUsed": <bool>}
```

Image-Image Consistency ($r_{img\text{-}img}$)

Compare the **Query Image** against the **Evidence Image** across the following dimensions:
1. **Sentiment Alignment:** [Aligned/Mismatch/Ambiguous]
2. **Entity Consistency:** [Aligned/Mismatch/Ambiguous]
3. **Event/Action Agreement:** [Aligned/Mismatch/Ambiguous]

Final Judgment: Provide a verdict of [TRUE/FAKE] followed by a detailed reasoning rationale.

Image-Text Consistency ($r_{img\text{-}txt}$)

Analyze the alignment between the **News Image** and its **Caption**:
1. **Sentiment Alignment:** Compare textual sentiment vs. visual tone.
2. **Entity Consistency:** Verify if text entities appear in the image.
3. **Event/Action Depiction:** Contrast the described event vs. visual depiction.

Judgment: Provide a verdict of [TRUE/FAKE] with supporting reasoning.

8.2. Stage 2: Final Aggregation Prompt

Unified Final Judgment Prompt

Evidence Input:
- *1. Img-Text:* $\mathbf{r}_{\text{img-txt}}$
- *2. Img-Img:* $\mathbf{r}_{\text{img-img}}$
- *3. Claim Verification:* $\mathbf{r}_{\text{txt-txt}}$

Aggregation Instructions: Assign higher weight to **Img-Img (2)** as it represents the strongest visual signal. Cross-reference with **Img-Text (1)** and **Web Evidence (3)**.
Final Output:

```
[REAL/FAKE] + Confidence (Low/Med/High)
+ FOR vs. AGAINST reasoning.
```

9. MLLM Reasoning Analysis

Here, we analyse the rationals obtained by our training-free multimodal fake news detection experiments on three different open-source multimodal LLMs for sample shown in Figure 7.

C.1 Gemma-3 (Correct Prediction: REAL)

$\mathbf{r}_{\text{img-txt}}$: Sentiment *Aligned* (‘negative sentiment...destruction‘), Entities *Ambiguous/Unverifiable* (‘could be embassy...no explicit confirmation‘), Event/Action *Aligned* (‘aftermath of blast‘). Final Judgment: *TRUE*. “Visual evidence compelling despite entity ambiguity.”
$\mathbf{r}_{\text{img-img}}$: Sentiment *Aligned* (‘negative...destruction‘), Entities *Aligned* (‘same building structure...signs visible‘), Event/Action *Aligned* (‘same destructive incident‘). Final Judgment: *TRUE*. “Strong alignment across all dimensions.”
$\mathbf{r}_{\text{txt-txt}}$: B1,B2 align (‘FAS=1: same event...embassy damage‘), B3-B5 unrelated (‘Amish barn...inappropriate‘). “Evidence supports core claim after filtering.”
Final Reasoning ($\mathbf{r}_{\text{final}}$): “REAL: $\mathbf{r}_{\text{img-img}}$ identical scenes + Web (20.0: multiple sources corroborate) outweigh entity ambiguity. High confidence.”

C.2 Qwen-2.5 (Incorrect Prediction: FAKE)

$\mathbf{r}_{\text{img-txt}}$: Sentiment *Mismatch* (‘neutral text vs negative image‘), Entities *Mismatch* (‘embassy/ambassador not visible‘), Event/Action *Aligned* (‘destruction aligns with blast‘). Final Judgment: *FAKE*. “Mismatches make pairing questionable.”
$\mathbf{r}_{\text{img-img}}$: Sentiment *Aligned* (‘negative destruction‘), Entities *Ambiguous/Unverifiable* (‘no identifiable entities‘),

Event/Action *Aligned* (‘same structural destruction‘). Final Judgment: *TRUE*. “Images consistent.”
$\mathbf{r}_{\text{txt-txt}}$: B2 aligned (‘embassy impact‘), others contradictory/unrelated (‘different events‘). “B2 supports claim.”
Final Reasoning ($\mathbf{r}_{\text{final}}$): “FAKE: $\mathbf{r}_{\text{img-txt}}$ mismatches + Web (0.2) outweigh $\mathbf{r}_{\text{img-img}}$ despite B2 support.”

C.3 InternVL2.5 (Incorrect Prediction: FAKE)

$\mathbf{r}_{\text{img-txt}}$: Sentiment *Mismatch* (‘neutral implies negative vs direct destruction‘), Entities *Aligned* (‘damaged building=embassy‘), Event/Action *Aligned* (‘structural damage=blast‘). Final Judgment: *TRUE*. “Sentiment mismatch doesn’t negate alignment.”
$\mathbf{r}_{\text{img-img}}$: Sentiment *Aligned* (‘negative destruction‘), Entities *Ambiguous/Unverifiable* (‘no identifiable entities‘), Event/Action *Aligned* (‘structural damage/debris‘). Final Judgment: *TRUE*. “Consistent event depiction.”
$\mathbf{r}_{\text{txt-txt}}$: B1,B2,B4 aligned (‘Athens explosion/embassy‘), B3 contradictory (‘Maryland‘), B5 unrelated. “Supports explosion theme.”
Final Reasoning ($\mathbf{r}_{\text{final}}$): “FAKE: Overweights Web (0.0: no credible sources) despite acknowledging $\mathbf{r}_{\text{img-txt}}$/$\mathbf{r}_{\text{img-img}}$ visual consistency.”

Key Observation: Gemma-3 demonstrates superior world knowledge by properly weighting strong $\mathbf{r}_{\text{img-img}}$ (‘Entities Aligned: same building‘) + accurate Web (20.0: ‘multiple sources corroborate‘), overcoming entity ambiguity. Qwen-2.5 fails by misclassifying text as *neutral* (vs Gemma’s negative sentiment), over-penalizing Entities *Mismatch* (vs Gemma’s tolerance), weak Web (0.2 ignores real event). InternVL2.5 fails with flawed Web (0.0: ‘no credible sources‘ despite 2015 incident), weaker Entities *Ambiguous* (vs Gemma’s Aligned). Gemma excels in precise reasoning and evidence integration.

Query Image

Best Evidence Image

Query Caption:
“The Cypriot embassy in Athens absorbed the full impact of the blast according to its ambassador”

Evidence Captions:
B1: Athens on edge after explosion severely damages buildings — Greece — The Guardian
B2: Damage to Cypriot embassy in Athens
B3: Violent Explosion At Amish Barn Fire In Maryland
B4: Athens On Edge After Explosion Severely Damages Buildings
B5: Explosion Sex Free Kissing Sex — CLOUDY GIRL PICS

Figure 7. Visual context for MLLM reasoning analysis from NewsClippings Dataset having GT:Real.