1 Benchmark Description

The Missing Information in Multimodal Emotion (MIME) benchmark is a comprehensive benchmark specifically constructed to evaluate multimodal emotion recognition models under varying levels of modality and emotional information loss.

Overall Scale and Sources: The benchmark comprises a total of 2,000 video segments. To ensure a high degree of diversity in social contexts and degradation levels, the raw data is sourced from social media platforms alongside four authoritative in-the-wild emotion recognition benchmarks: CAER, DFEW, MAFW, and FERV39k.

Emotion Categories and Balance: MIME focuses on seven universal emotion categories: Anger, Disgust, Fear, Happy, Neutral, Sad, and Surprise. The benchmark rigorously adheres to a balanced sample distribution to mitigate the long-tail bias that is often prevalent in in-the-wild benchmarks. The specific sample counts for each category are balanced as follows: Anger (289), Disgust (283), Fear (283), Happy (288), Neutral (284), Sad (288), and Surprise (285).

Contextual Diversity: The video segments are systematically decoupled across a wide array of diverse in-the-wild scene contexts, enabling fine-grained analysis of model performance in different environments. These contexts include, but are not limited to:

- **Daily Life & Events:** Homes, schools, offices, restaurants, weddings, birthday parties, sports events, and general daily life.
- **Media & Entertainment:** Movies, TV series, sitcoms, dramas, talk shows, variety shows, short videos, concerts, and stage plays.
- **Specific Genres:** Action, comedy, sci-fi, thrillers, suspense, and war scenes.

Duration Statistics: The average durations of the video and audio data are evenly distributed across the different modality-missing conditions, which facilitates a comprehensive evaluation of temporal understanding across degraded cases. The average segment durations for each condition are: Full Modality (3.31s), Face Details Missing (3.22s), Face Structures Missing (3.68s), Visual Modality Missing (2.42s), Face Details & Audio Missing (2.85s), Face Structures & Audio Missing (3.08s), and Audio Modality Missing (3.06s). **Annotation Structure:** Beyond discrete emotional labels, the benchmark is uniquely annotated with high-quality, fine-grained Chain-of-Emotion (CoE) rationales. This ensures that each sample contains a structured, tripartite logic flow detailing "Scene Understanding," "Emotional Analysis," and the final "Predicted Emotion" to support interpretable model evaluation.

2 Data Acquisition and Preprocessing

To ensure a high degree of diversity in social contexts and degradation levels, we established a rigorous pipeline for raw data collection, standardization, and modal degradation.

Data Collection and Standardization. We initially collected a large-scale pool of raw videos from social media platforms and four authoritative in-the-wild benchmarks (CAER, DFEW, MAFW, and FERV39k). To ensure data quality, we conducted a manual and automated filtering process to remove video segments with severe watermarks, extreme low resolution ($< 480p$), or ambiguous ground-truth emotional labels. For consistency across the benchmark, all selected video clips were standardized to a uniform frame rate of 30 FPS, and the accompanying audio tracks were extracted and resampled to 16 kHz. We focus on seven universal emotion categories (Happy, Sad, Neutral, Anger, Disgust, Fear, and Surprise), applying a balanced resampling strategy to mitigate the long-tail bias commonly found in naturalistic benchmarks.

Subject Tracking and Facial Processing. In-the-wild videos frequently feature multiple individuals, complex camera movements, and subject occlusions. To accurately isolate the target subjects for continuous degradation without corrupting the background, we design a robust, dual-modal tracking pipeline. In the initial frame, the primary subject is identified using a comprehensive scoring mechanism that evaluates facial centrality (40%), structural quality derived from 106-point landmarks (30%), relative body bounding-box area detected via YOLOv8 (20%), and facial scale (10%). For temporal consistency across subsequent frames, we employ a hybrid matching strategy. We compute the cosine similarity of facial embeddings using InsightFace alongside the correlation of the subject's body HSV color histograms. This ensures that the primary subject is reliably tracked even during extreme head poses or transient facial occlusions.

Systematic Modal Degradation. Through the aforementioned processing, we established a high-quality, full-modality baseline. Subsequently, we systematically decouple and degrade the video signals to construct seven distinct subsets, simulating various real-world scenarios of information loss:

- **Subset 1 (MIME-FM): Full modality.** Intact audio-visual clips providing complete information, representing ideal, controlled environments like professional broadcasts or high-quality recordings.

- **Subset 2 (MIME-FDM): Face details missing.** We map the tracked subject's 106 facial landmarks into a convex hull to generate a precise, edge-softened facial mask. A 2D Gaussian blur with a light kernel size ($k = 15, \sigma = 0$) is then applied strictly within this mask. This simulates mild visual imperfections (e.g.,

low-resolution webcams, compression artifacts) where subtle micro-expressions are lost but the general facial layout remains perceptible.

- **Subset 3 (MIME-FSM): Face structures missing.** Utilizing the same precise masking technique, we apply a progressively heavier Gaussian blur ($k \in \{35, 55\}, \sigma = 0$) to severely degrade facial structures. This effectively renders the face as an unrecognizable color blob, replicating severe real-world occlusions (e.g., heavy masks, extreme lighting failures) and forcing models to deduce emotions entirely from body language and scene context.

- **Subset 4 (MIME-VMM): Visual modality missing.** Simulates an audio-only scenario where all visual frames are discarded (replaced by black frames). This mirrors everyday situations like phone calls, voice messages, or completely occluded cameras.

- **Subset 5 (MIME-FDAM): Face details and audio modality missing.** A dual-modality loss condition. We combine the light facial blur from Subset 2 ($k = 15$) and remove the audio track. This mimics highly adverse environments like distant, low-resolution, and muted CCTV surveillance.

- **Subset 6 (MIME-FSAM): Face structures and audio modality missing.** The most extreme visual degradation combined with audio loss. The face is heavily blurred (as in Subset 3, $k \in \{35, 55\}$) and audio is muted, rigorously testing the ability to infer emotions entirely from body posture and surrounding environmental dynamics.

- **Subset 7 (MIME-AMM): Audio modality missing.** Simulates a visual-only scenario by removing the audio track, corresponding to viewing muted videos, interacting through soundproof barriers, or encountering hardware microphone failures.

3 Chain-of-Emotion (CoE) Generation

To endow our benchmark with rich, fine-grained reasoning annotations, we design a comprehensive four-stage pipeline utilizing Qwen3-Omni to generate high-quality Chain-of-Emotion. As shown in Figure 1, this pipeline explicitly guides models to decouple and articulate their observations across individual modalities prior to predicting the final emotion.

Stage I: Foundational Cue Extraction. The model is prompted to systematically decouple the input into distinct fine-grained modalities: Face, Body, Scene, and Audio.

For uncorrupted or inherently unimodal subsets (Subsets 1 and 4), this extraction operates directly on the target input. However, for visually degraded Subsets (Subsets 2, 3, 5, and 6), we first input the original clear video to extract a comprehensive, factual baseline of objective physical elements (e.g., subtle micro-expressions, body language, and environmental context).

Stage II: Conditional Remained-Cue Verification. This stage is exclusively triggered for subsets involving facial blurring (Subsets 2, 3, 5, and 6). We input the degraded (blurred) video alongside the baseline cues extracted in Stage I. The model is instructed to rigorously cross-reference these inputs and explicitly delete any facial details or structural information that are no longer discernible due to the applied Gaussian blur. This critical step dynamically prunes the cue set, suppressing multi-modal hallucinations and ensuring that subsequent reasoning strictly relies only on genuinely surviving visual signals.

Stage III: Ground-truth Consistency Filtering. To guarantee the validity and rigorousness of the benchmark, we implement a strict programmatic validation step. The predicted emotion derived from the verified cue set is cross-checked against the original ground-truth label. Samples where even state-of-the-art models cannot deduce the correct label (due to excessive information loss causing a prediction mismatch) are strictly discarded. This automated filtering mechanism ensures that the remaining challenging subsets in MIME are still fundamentally solvable.

Stage IV: Structured CoE Generation. The final verified samples are processed through a structured prompt, where the MLLM acts as a professional expert in psychological and micro-expression analysis, to generate the formalized Chain-of-Emotion (CoE) in a JSON format. We mandate a strict tripartite logic flow:

- **Scene Understanding:** Characterizing the objective multimodal cues present in the input, extracting factual information exclusively from the available modalities (face, body, scene, and audio).
- **Emotional Analysis:** Executing cognitive reasoning to map the previously identified cues to internal emotional states. This step explicitly articulates the logical path of how the observed signals lead to the emotional deduction.
- **Predicted Emotion:** Deriving the final discrete emotion category, which strictly aligns with the ground-truth label, based on the preceding systematic analysis.

4 Potential Applications

While the primary contribution of the MIME benchmark is to systematically evaluate model robustness under modality-impaired scenarios, the benchmark and the proposed Chain-of-Emotion (CoE) paradigm unlock significant potential across several real-world applications:

- **Privacy-Preserving Affective Computing:** In sensitive domains such as healthcare (e.g., automated psychological and depression assessment) and education (e.g., classroom engagement monitoring), capturing high-resolution facial data raises severe privacy concerns. MIME facilitates the development of affective models capable of inferring user emotions through alternative modalities—such as body language, scene context, and vocal nuances—thereby protecting user identity while delivering intelligent services.

- **Robust In-Cabin Monitoring and Wearable HCI:** Real-world human-computer interaction frequently occurs in visually constrained environments. Driver monitoring systems must contend with extreme lighting variations and facial occlusions (e.g., sunglasses, masks), while users wearing AR/VR headsets have their upper faces entirely obscured. Models trained on MIME's fine-grained degraded subsets (e.g., FSM and VMM) can leverage compensatory cues to reliably detect critical states like fatigue, stress, or panic, enhancing both safety and user experience.

- **Public Safety and Crisis Intervention:** Traditional surveillance systems (CCTV) typically capture subjects at a distance, resulting in low-resolution, blurred, or partially occluded faces—conditions accurately simulated by MIME's FDM and FDAM subsets. By training models to analyze scene context, interpersonal dynamics, and body posture (as explicitly modeled in our CoE annotations), public safety systems can proactively detect distress, fear, or aggressive behavior without relying on explicit facial micro-expressions.

- **Cognitive Instruction Tuning for MLLMs:** Current Multimodal Large Language Models (MLLMs) exhibit a strong dependency on visual shortcuts, often suffering catastrophic performance collapse when explicit facial cues are distorted. The proposed Chain-of-Emotion (CoE) annotations provide high-quality, structured reasoning data. This can be directly utilized for cognitive instruction tuning, forcing future foundation models to perform multi-step, human-like affective reasoning rather than simplistic probabilistic guessing.

Algorithm 1 Chain-of-Emotion (CoE) Generation Pipeline

Require: Target Video $\mathcal{V}$ (clear or degraded), Ground Truth Emotion $\mathcal{E}$, Subset Type $\mathcal{S} \in \{1, 2, \ldots, 7\}$
Ensure: Structured CoE Output or Rejected Status

// Stage I: Foundational Cue Extraction
if $\mathcal{S} == 1$ **or** $\mathcal{S} == 4$ **then**
    Oracle_Cues $\leftarrow$ MLLM(Prompt_Stage1, $\mathcal{V}, \mathcal{E}$)
else
    $\mathcal{V}_{clear} \leftarrow$ Get_Clear_Version($\mathcal{V}$)
    Oracle_Cues $\leftarrow$ MLLM(Prompt_Stage1, $\mathcal{V}_{clear}, \mathcal{E}$)
end if
// Stage II: Conditional Remained-Cue Verification
if $\mathcal{S} \in \{2, 3, 5, 6\}$ **then**
    Blur_Level $\leftarrow$ Get_Blur_Level($\mathcal{V}$)
    if Blur_Level == Light **then**
        Verified_Cues $\leftarrow$ MLLM(Prompt_Stage2_Light, $\mathcal{V}$, Oracle_Cues)
    else ▷ Heavy Blur
        Verified_Cues $\leftarrow$ MLLM(Prompt_Stage2_Heavy, $\mathcal{V}$, Oracle_Cues)
    end if
else
    Verified_Cues $\leftarrow$ Oracle_Cues
end if
// Stage III: Ground-truth Consistency Filtering
if $\mathcal{S} \neq 1$ **then**
    Pred_Emotion, Emotion_Cues $\leftarrow$ Text_LLM(Prompt_Stage3, Verified_Cues)
    if Pred_Emotion $\neq \mathcal{E}$ **then**
        return "Rejected" ▷ Discard sample due to excessive information loss
    end if
else
    Emotion_Cues $\leftarrow$ Verified_Cues ▷ Direct mapping for Subset 1
end if
// Stage IV: Structured CoE Generation
Prompt_Stage4 $\leftarrow$ Format_Prompt(Verified_Cues, Emotion_Cues, $\mathcal{E}$, Blur_Level)
CoE_Output $\leftarrow$ MLLM(Prompt_Stage4, $\mathcal{V}$)
return CoE_Output

Example of the 4-Stage CoE Generation Pipeline

To concretely illustrate our proposed data construction methodology, we provide a complete, step-by-step pipeline execution for a heavily degraded sample (Subset 3: Face Structures Missing). This example explicitly demonstrates the complete prompts and corresponding outputs at each stage, showcasing how the system extracts foundational cues, rigorously filters out hallucinated facial features, performs blind emotion inference, and synthesizes the structured Chain-of-Emotion.

Pipeline Execution Record | **Subset:** Subset 3 (FSM) | **GT Emotion:** Anger

STAGE I: Foundational Cue Extraction

Input (System Prompt + Original Clear Video):

Role: Multimodal Emotion Analysis Expert
Task: Analyze the provided **CLEAR, UNBLURRED video**. The ground truth emotion is **Anger**. Your goal is to extract a structured "Visual & Audio Evidence List" based on the raw footage. **Crucial:** Focus specifically on extracting and detailing cues that constitute the evidence for **Anger**.
ANONYMITY & OBJECTIVITY PROTOCOL
* **IGNORE EXTERNAL KNOWLEDGE:** Do NOT use specific actor names or set names.
* **USE GENERIC LABELS:** Use gender-specific labels like "The man in [clothing]".
* **EXCEPTION:** Only write a name if it is explicitly spoken in the audio **Caption**.
# **Extraction Guidelines**
1. **SCENE & CONTEXT:** Describe background, social setting, and specific physical interactions. **Target Identity:** You MUST explicitly describe the Target Person's appearance. Priority Rule: 1. Clearest Face. 2. Largest Body Area.
2. **BODY & HEAD:** Describe specific hand gestures, body posture, head orientation, and movement speed.
3. **AUDIO & SPEECH ANALYSIS:** Step 1: Identify the Speaker. Category A: [Target Person Speaking]. Verify this matches Target Identity. Output Caption and Tone (Pitch, Volume, Speed). Category B: [Other Person Speaking]. Output Caption ONLY. Category C: [Scene Noise]. Category D: [Quiet/No Voice].
4. **FACIAL MOVEMENTS:** Analyze the face using **Facial Action Units (AU)** concepts for accuracy, but **OUTPUT ONLY natural language descriptions**. **FORBIDDEN:** Do NOT write "AU4". Describe specific organ movement clearly.
Output Constraint DO NOT write an introduction. **Output strictly** in the specified section format.

Output (Oracle Cues):

- **### SECTION A: SCENE & CONTEXT**
  - The scene takes place indoors, likely in a residential hallway or entryway, with a staircase railing visible in the background and a decorative light fixture overhead.

- The woman is standing and facing the man, engaged in a direct conversation. Another person in a red hat walks past them in the background, indicating a casual, possibly domestic setting.

- **### SECTION B: BODY & HEAD**
  - The woman is standing upright, facing the man, with her body oriented toward him, indicating direct engagement... The man is seen from the side and back, with his head turned toward the woman... There is no visible aggressive physical interaction.
- **### SECTION C: AUDIO ANALYSIS**
  - **Source Category:** Target Person (The woman in the white floral dress)
  - **Caption:** "I didn't lie. You made an assumption I didn't contradict it."
  - **Tonal/Sound Details:** The woman's voice is firm and slightly raised in pitch, indicating defensiveness and irritation. Her speech is delivered at a moderate pace, with a clear, assertive tone...
- **### SECTION D: FACIAL MOVEMENTS**
  - Brows lowered and slightly furrowed, indicating frustration or annoyance.
  - Eyes wide and fixed on the man... Mouth slightly open and lips pressed together...

STAGE II: Conditional Remained-Cue Verification

Input (System Prompt + Degraded Video + Oracle Cues):

 # Role: Privacy-Preserving Video Analyst (Heavy Blur Specialist)
 # Context: You are analyzing a video with **LEVEL 3 (HEAVY) BLUR** on the face. Blur Status: The face is a featureless blob/pixelated block. **Goal: STRICTLY REJECT** all facial details to prevent Hallucination.
 # NEGATIVE CONSTRAINT (CRITICAL)
* **NO EMOTION WORDS:** You are an objective observer. **DO NOT** use emotion adjectives like "angry", "happy" in your output.
* **USE PHYSICAL DESCRIPTIONS:** Instead of "angry tone", use "loud, fast-paced tone".
 # Protocol
1. **Identity Extraction:** Create standalone fact line: [**REMAINED**] **(Identity) {Identity}**.
2. **Audio Verification:** Validate the Source Category.
3. **Body/Scene Verification:** Mostly keep them. Prefix with [**REMAINED**].
4. **Face Verification:** The face is considered INVISIBLE. **DELETE** all specific facial cues.

Output (Verified Facts):

- **[REMAINED] (Identity)** The man in the purple shirt with colorful embroidery
- **[REMAINED] (Scene)** The scene takes place indoors, likely in a residential hallway or entryway, with a staircase railing visible in the background and a decorative light fixture overhead.
- **[REMAINED] (Body)** The woman is standing upright, facing the man, with her body oriented toward him, indicating direct engagement... The man is seen from the

side and back, with his head turned toward the woman... There is no visible aggressive physical interaction.

- **[REMAINED] (Audio)** [Source: Target Person] "I didn't lie. You made an assumption I didn't contradict it." - Tone: loud, fast-paced, assertive.
- **[DELETED] (Face)** *All facial cues from Stage I are strictly removed.*

STAGE III: Ground-truth Consistency Filtering

Input (System Prompt + Verified Facts):

Role Blind Emotion Inference Engine
Task Based **SOLELY** on the verified facts above, infer the most likely emotion.
Inference Logic 1. Semantic Analysis (Caption content). 2. Tonal Analysis (Primary). 3. Contextual Analysis (Scene interactions and Body Language).
Output Format JSON Only.

Output (Filter Result):


```
{
  "predicted_emotion": "Anger",
  "confidence": "High",
  "emotion_cues": "Target shouted loudly and fast-paced, asserting 'I
     didn't lie. You made an assumption I didn't contradict it.'",
  "reasoning": "The target person's voice is loud and fast-paced,
     indicating assertiveness often associated with anger. The content
     of the speech -- denying a lie while accusing the other party of
     making an assumption -- suggests defensive frustration. The body
     language shows direct engagement, with both individuals facing each
     other, which reinforces the intensity of the interaction. No visual
     or auditory cues suggest laughter, surprise, or sadness. The
     combination of tone and semantic content strongly supports Anger as
     the dominant emotion."
}
```


Note: Since Predicted Emotion matches Ground Truth (Anger), the sample proceeds.

STAGE IV: Structured CoE Generation

Input (System Prompt + Verified Facts + Inferred Cues):

Role Multimodal CoT Generator (Visual-Grounded)
Goal Write a logical analysis (Chain-of-Emotion) that explains **WHY** the video expresses **Anger**, based on the verified facts and emotion cues.
Output Rules 1. Integration over Listing. 2. Strict Fact Adherence. **DO NOT** hallucinate facial cues if they are not listed in the facts (especially for Level 2/3).
Format 1. Scene Understanding (MANDATORY: Start by identifying the protagonist. IF Level == 2 or 3: You MUST write: "The face is heavily blurred.") 2. Emotional Analysis (Analyze in order: Facial Premise, Audio Caption & Tone, Scene Interaction, Body Language). 3. Conclusion.

Output (Final CoE Annotation):

1. Scene Understanding

The man in the purple shirt with colorful embroidery stands in a residential hallway, facing a woman. The face is heavily blurred. A staircase railing and decorative light fixture are visible, indicating an indoor setting. Both individuals remain upright, engaged in direct conversation.

2. Emotional Analysis

Although the face is heavily blurred, we can still infer anger from the following cues. The man in the purple shirt says "I didn't lie. You made an assumption I didn't contradict it." with a loud, fast-paced, assertive tone. The scene shows no aggressive physical interaction, yet their upright, face-to-face stance suggests confrontation.

3. Conclusion

Anger