# CAN MLLMS ASSIST FORENSIC SKETCH IDENTIFICATION VIA SEMANTIC REASONING?

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

When an eyewitness describes a suspect, the description is conveyed through language rather than pixels. Forensic face recognition, however, has largely been studied as a purely visual task, and the semantic content of verbal, written, and spoken facial descriptions has received little attention. We ask whether Multimodal Large Language Models (MLLMs) can bridge language and vision by jointly reasoning over facial sketches, textual descriptions, and spoken biometric narratives for forensic identity recognition. To study this, we introduce MS-Forensics, the first multimodal benchmark for sketch-based forensic identity recognition, which pairs facial sketches with textual and audio-based biometric descriptions built from five public face and sketch datasets. Through hypothesis-driven evaluation, we show that recognition degrades as the candidate set grows, and that the saturation of description-based recognition is explained by a collision model over attribute combinations. Fusing sketches with textual or audio descriptions outperforms unimodal inputs in most settings, but MLLMs struggle under fine-grained identity similarity and rely heavily on demographic cues; once demographics are removed, the eyes and forehead emerge as the most discriminative attributes. MS-Forensics provides a benchmark for evaluating MLLMs in AI-assisted forensic analysis.

## 1 INTRODUCTION

Over the past few years, Multimodal Large Language Models (MLLMs) have achieved remarkable success across vision-language tasks Cheng et al. (2025); Achiam et al. (2023); Touvron et al. (2023); Wu et al. (2025); Li et al. (2023) and facial analysis Sony et al. (2025); Shahreza & Marcel (2025b); Narayan et al. (2026). Yet, their ability to reason under the sparse and ambiguous visual evidence encountered in forensic investigations remains largely unexplored. In particular, recognizing identities from forensic sketches is a long-standing challenge, as sketches preserve only coarse facial structure while omitting many discriminative identity cues. In practice, however, forensic investigations rarely rely on sketches alone; eyewitnesses also provide detailed verbal descriptions that encode rich semantic evidence unavailable in the sketch. Despite this complementary information, existing methods treat sketch recognition as a purely visual matching problem, overlooking the joint reasoning capabilities of modern MLLMs. In this work, we revisit forensic sketch recognition by investigating whether MLLMs can integrate facial sketches, photographs, and descriptive narratives for robust identity recognition under realistic forensic settings.

Beyond facial sketches, we incorporate linguistic cues that closely reflect real forensic investigations, where eyewitnesses describe a suspect's facial appearance verbally. We represent these biometric descriptions in both textual and audio modalities to capture complementary semantic information that is often absent from sketches. This formulation allows pretrained MLLMs to reason over heterogeneous visual and linguistic evidence without requiring sketch-specific training. More importantly, it enables us to investigate a fundamental question: Can multimodal semantic descriptions compensate for the inherent ambiguity of sketches and improve recognition under realistic forensic conditions?

To analyze the MLLM's capability in sketch-based face recognition, we investigate several key hypotheses. Hypotheses are motivated to depict MLLM's understanding of **(H1)** faces across a large set of facial image contexts, the **(H2)** influence of different modality cues, and the **(H3)** robustness of MLLMs over highly challenging scenarios from recognition perspective to real-world forensic



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

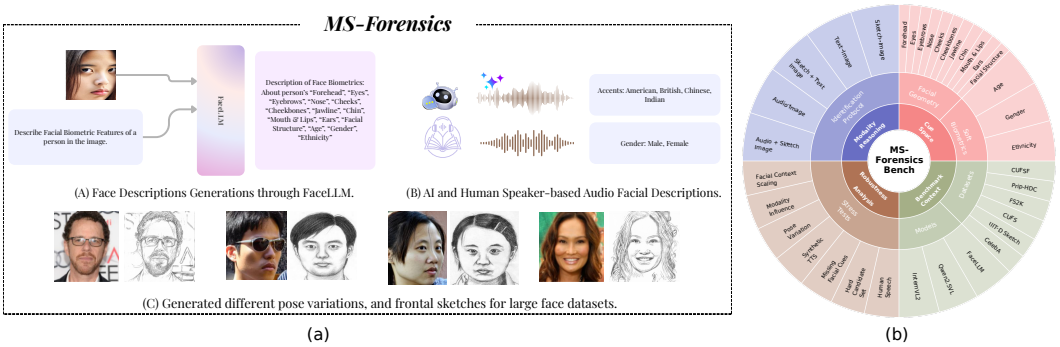



Figure 1: Overview of the proposed MS-Forensics benchmark for multimodal sketch-based face recognition in forensic settings. (a) Illustration of the multimodal data curation pipeline, including biometric facial descriptions, AI-generated and human-spoken audio narratives, and corresponding facial sketch generation. (b) Structured design of the MS-Forensics benchmark enabling hypothesis-driven evaluation of MLLM-based multimodal identity recognition.

perspective of partially available context through partially available facial descriptions and incompletely drawn sketches.

Our empirical analysis investigates how MLLMs reason across sketches, photographs, and linguistic descriptions for identity recognition under limited visual observations. We show that unimodal inputs are insufficient for reliable recognition: sketches lack fine-grained identity details, while descriptions miss the spatial structure needed for precise matching. In contrast, combining sketches with semantic descriptions enables complementary multimodal reasoning and improves recognition. To better reflect real forensic investigations, we further introduce a *human-spoken* facial description protocol and validate its effectiveness through realistic case studies. Our key contributions are:

1. **MS-Forensics Benchmark.** The first multimodal benchmark for sketch-based forensic identity recognition, integrating facial sketches, textual biometric descriptions, and audio-based narratives. Built from five public face and face-sketch datasets (CUFSF, FS2K, IIIT-D Sketch, PRIP-HDC, and CelebA), it enables forensic evaluation across four state-of-the-art MLLMs: FaceLLM, InternVL2, Qwen2.5-VL, and Qwen2.5-Omni.
2. **Hypothesis-Driven Analysis of MLLMs.** We conduct a comprehensive, hypothesis-driven analysis of several MLLMs on the MS-Forensics bench, examining their cross-modal reasoning across heterogeneous inputs, including image, textual, and audio.
3. **Fine-Grained Identity Reasoning.** We perform controlled studies of MLLM behavior under challenging conditions, including the role of discriminative facial attributes, hard candidate sets with similar-looking identities and pose variations, and incomplete cues.
4. **Real-world forensic case study and forensic expert evaluation.** We present scenarios where sketches are drawn from eyewitness memory-based descriptions and jointly reasoned with linguistic evidence to identify suspects from challenging candidates. Four forensic experts further judged the MLLM-generated descriptions, found them largely accurate and most often preferred the description+sketch for the sketch-forensic identification task.

## 2 Related Work

A recent survey Shahreza & Marcel (2025c) summarizes the growing role of foundation models in biometrics. Correspondingly, several benchmarks have emerged for facial MLLMs, including FaceBench Wang et al. (2025), FaceXBench Narayan et al. (2026), and Face-Human-Bench Qin et al. (2025), which evaluate facial attribute reasoning and face-centric multimodal understanding. FaceLLM Shahreza & Marcel (2025b) further introduces a vision-language framework for facial reasoning and description generation, while recent studies benchmark MLLMs for fine-grained facial attribute perception through FG-BMK Yu et al. (2025), face verification Shahreza & Marcel (2025a), and heterogeneous face recognition Shahreza et al. (2026).

Table 1: MS-Forensics benchmark statistics: datasets, protocols, gallery sizes, sketches, facial descriptions, audio samples, pose variations, and candidate identity sets used for hypothesis-driven MLLM evaluation. *VS*: Viewed Sketch, *SFS*: Semi-Forensic Sketch, *F*: Female, *M*: Male.

| Hypotheses | Protocol Settings | Datasets | Facial Gallery Size | # Sketches | # Facial Descriptions | # Audios | # Poses | Identity Candidate Size ($k$) | Audio Variations |
|---|---|---|---|---|---|---|---|---|---|
| **Facial Context Scaling** | **Textual Facial Description, Sketch, Textual Description + Sketch** | CUFSF Wang & Tang (2009) | 959 | 959 | 959 | - | - | 2 to 12 | - |
| **Multimodal Cues** | **Textual Facial Description, Sketch, Textual Description + Sketch** | CUFSF Wang & Tang (2009) | 959 | 959 | 959 | - | - | 4 | - |
| | | FS2K Fan et al. (2022) | 1045 | 1045 | 1045 | - | - | 4 | - |
| | | IIIT-D Sketch Bhatt et al. (2012) (VS) | 231 | 231 | 231 | - | - | 4 | - |
| | | IIIT-D Sketch Bhatt et al. (2012) (SFS) | 139 | 139 | 139 | - | - | 4 | - |
| | | PRIP-HDC Klum et al. (2014) | 44 | 44 | 44 | - | - | 4 | - |
| | | CelebA Liu et al. (2015) | 10171 | 10171 | 10171 | - | - | 4 | - |
| **Multimodal Facial Descriptive Narratives** | **Generated Audio Facial Descriptions,** | FS2K Fan et al. (2022) | 1045 | 300 | 300 | 2400 | - | 4 | 4 Accents - F: 4, M: 4 |
| | **Human Spoken Facial Descriptions** | FS2K Fan et al. (2022) | 1045 | 100 | 100 | 200 (F: 100, M: 100) | - | 4 | 1 Accent - F: 7, M: 7 |
| **Fine-Grained Inter-class Identity Variations** | **Pose Variations** | CUFSF Wang & Tang (2009) | 5754 | 959 | 959 | - | 6 | 4 | - |
| | | FS2K Fan et al. (2022) | 6270 | 1045 | 1045 | - | 6 | 4 | - |
| | **Biometric Attributes** | FS2K Fan et al. (2022) | 1045 | 300 | 300 | - | - | 4 | - |
| | **Hard Identity Candidate Set** | FS2K Fan et al. (2022) | 1045 | 100 | 100 | - | - | 4 | - |
| | **Partial Modality Semantics** | IIIT-D Sketch Bhatt et al. (2012) (VS) | 231 | 231 | 231 | - | - | 4 | - |
| | | IIIT-D Sketch Bhatt et al. (2012) (SFS) | 139 | 139 | 139 | - | - | 4 | - |

Despite these advances, existing MLLM-based facial analysis remains largely confined to photographic facial images and attribute reasoning. For example, Shahreza et al. (2026) evaluates MLLMs for cross-spectral face verification (VIS-NIR/SWIR/THERMAL) without considering sketches, leaving sketch-based forensic identity recognition largely unexplored. Moreover, no prior work jointly investigates sketches with textual and audio-based facial descriptions or analyzes MLLMs under realistic forensic settings involving candidate-set scaling and fine-grained identity similarities, and attribute sensitivity. To bridge this gap, we introduce *MS-Forensics*, the first multimodal benchmark for evaluating MLLMs on sketch-based forensic identity reasoning and recognition.

## 3 MS-Forensics Bench

The MS-Forensics benchmark is designed to evaluate MLLMs for multimodal forensic identity recognition by integrating heterogeneous evidence sources that reflect real investigative conditions. The overall structure of this Multimodal forensic benchmark is illustrated in Figure 1, and statistics and protocol settings are summarized in Table 1 and implementation details are mentioned in Appendix I. MS-Forensics Bench curation has five key components:

**(I) Face-Sketch Curation:** We compiled face sketches using metadata from four publicly available face-sketch datasets: CUFSF Wang & Tang (2009), IIIT-D Sketch Bhatt et al. (2012), PRIP-HDC Klum et al. (2014), and the FS2K dataset Fan et al. (2022). Another set is generated from the CelebA Liu et al. (2015) dataset. Within IIIT-D Sketch, we distinguish between viewed sketches drawn while directly observing the subject and semi-forensic sketches, drawn from memory, representing different forensic conditions.

To incorporate a large-scale, high-fidelity sketch source, we employ the diffusion-based ControlNet++ framework Li et al. (2024) to generate 10,171 facial sketches for the CelebA Liu et al. (2015) dataset (MLLM is evaluated on 1000 subjects). The structural consistency of the sketches is validated through a human study conducted with 30 research experts, who verify whether the generated sketches match the actual face, yielding 91.0% face-sketch matching accuracy. Additionally, we reconstruct faces from the generated sketches using GPT-4o Achiam et al. (2023) and match them to the original photographs with ArcFace Deng et al. (2019), achieving 91.7% recognition accuracy. The SSIM Wang et al. (2004) score computed between the face and its generated sketch, resulting in 0.32 SSIM, which is comparable to the 0.26 SSIM observed for artist-drawn sketches in the IIIT-D Sketch dataset. Overall, these human-verifier and empirical results validate generated sketches that preserve identity-relevant geometry, matching the real-world forensics.

**(II) Biometric Facial Description Generation:** Structured biometric facial descriptions are generated from the facial photographs of each identity using various MLLMs, including FaceLLM Shahreza & Marcel (2025b), Qwen2.5-VL Bai et al. (2025) and InternVL2 Chen et al. (2024). Descriptions are generated from photos rather than sketches so that textual attributes are grounded in the facial appearance rather than in the sketch artist's abstraction. Since the source photograph also appears in the candidate set, these descriptions represent an upper bound on witness quality; the associated leakage risk is examined in Appendix G. Each identity is annotated across 14 fine-grained biometric attributes through attribute-specific prompting, ensuring standardized and consistent anno-

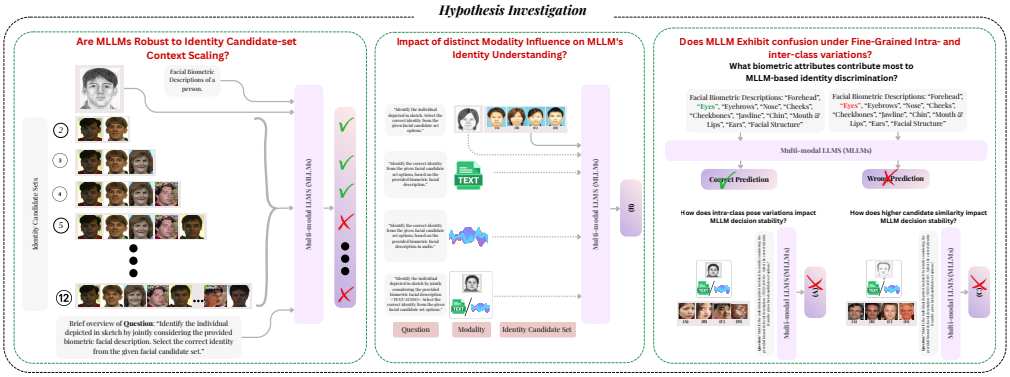



Figure 2: Hypothesis-driven evaluation framework of MS-Forensics. The framework investigates three hypotheses: (left) robustness to candidate context scaling, (middle) the impact of multimodal cues (sketch, text, and audio) on recognition, and (right) confusion under fine-grained intra- and inter-class variations, including attribute sensitivity, pose variations, and hard candidate sets.

tation. Sample prompts and facial descriptions, including those generated using different MLLMs, are provided in Appendix J, K. To assess the facial description's forensic utility, four forensic experts rated the descriptions and corresponding sketches for 20 subjects: the descriptions were judged at least partially accurate in 99% of ratings, and description + sketch was selected as the most complete information for identification in 71% of ratings

**(III) Audio-Based Biometric Narratives:** To simulate forensic scenarios like verbal eyewitness testimony, we generate synthetic audio narratives from the descriptions using the Kokoro[1] text-to-speech model across four accents, American, Chinese, Indian and British, in male and female voices. To better reflect real investigative settings and capture natural speech variability, we collect *200 human-recorded audios* from 14 speakers (7 male and 7 female) across diverse conditions.

**(IV) Pose-Variant Candidate Set Construction:** To evaluate MLLM robustness under viewpoint discrepancies between probe sketches and candidate images, we generate pose-variant candidate faces using Arc2Face Papantoniou et al. (2024) across five non-frontal pose configurations defined by pitch and yaw alongside the frontal baseline, yielding six pose conditions per identity.

**Problem Formulation:** Let $\mathcal{I}$ denote the set of enrolled identities. Each probe consists of a facial sketch $s$ and, optionally, a biometric description in textual form $d_t$ or spoken form $d_a$, all derived from the same underlying identity $y \in \mathcal{I}$. For each query, the input contains a candidate set $\mathcal{C} = \{x_1, x_2, \ldots, x_k\}, x_j \in \mathcal{X}$, where, $\mathcal{X}$ is the gallery of facial photographs, $k$ is the candidate-set size, and exactly one candidate $x_{j^*} \in \mathcal{C}$ depicts the probe identity $y$. A multimodal large language model $f_\theta$ performs closed-set identification by selecting a candidate conditioned on the modalities:

$$\hat{y} = f_\theta(m, \mathcal{C}), \qquad m \in \{s,\ d_t,\ d_a,\ (s, d_t),\ (s, d_a)\}, \tag{1}$$

where $m$ specifies the input configuration, covering the sketch-only, description-only, audio-only, and multimodal settings evaluated in our protocols. Recognition performance is measured by identification accuracy over a set of $N$ queries: $\text{Acc} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}[\,\hat{y}_i = y_i\,]$.

Note that Eq. 1 defines a 1:$k$ closed-set *identification* task, in which the model must discriminate among $k$ competing identities rather than score a single enrollment–probe pair. Identification difficulty, therefore, grows with $k$, a property we exploit in our hypothesis-driven evaluation.

**(V) MS-Forensics Bench Recognition Protocols:** The curated components are integrated into the *MS-Forensics* benchmark to formulate structured identity recognition protocols as instantiations of Eq.1 across modality configurations. Hypothesis H1 varies the candidate-set size $k \in \{2, \ldots, 12\}$ with $m \in \{s,\ d_t,\ (s, d_t)\}$ to assess robustness to identity context scaling; H2 fixes $k = 4$ and varies the input configuration $m$ across all five settings to isolate modality-specific effects; and H3 fixes both $k$ and $m$ while manipulating candidate difficulty through attribute removal, pose variations, and visually similar distractors. Each protocol systematically evaluates multiple large-scale

[1] https://github.com/hexgrad/kokoro

MLLMs, namely FaceLLM-38B Shahreza & Marcel (2025b), InternVL2-26B Chen et al. (2024), Qwen2.5-VL-32B Bai et al. (2025), and Qwen2.5-Omni Xu et al. (2025), under our hypothesis-driven framework. For recognition evaluation on CelebA, we use 1000 identities for testing, all other dataset follow protocol from Table 1.

## 4 Hypotheses-Driven Investigation of MLLMs

To rigorously analyze MLLM behavior in forensic identity recognition, we design experimental settings aligned with three core hypotheses, evaluated through systematic strategy. Figure 2 provides a visual overview of the framework.

**Hypothesis (H1): MLLMs Robustness to Identity Candidate-set Context Scaling**: *MLLM performance degrades as the identity candidate set increases in size due to increased facial context and inter-class ambiguity.*

To evaluate this hypothesis, we systematically vary $k \in \{2, \ldots, 12\}$ while keeping the probe sketch $s$ fixed, measuring recognition accuracy under expanding contexts. For each, the MLLM identifies the correct identity by reasoning over the provided inputs. We evaluate FaceLLM Shahreza & Marcel (2025b), InternVL2 Chen et al. (2024), and Qwen2.5-VL Bai et al. (2025) on the CUFSF Wang & Tang (2009) dataset under three modality settings: textual descriptions only, sketch only, and their combination. Figure 2 (left) illustrates the evaluation protocol, while an example MCQ-based forensic query is provided in the Appendix K.

Figure 3 (and Figure 1 in the Appendix A.1) reports recognition accuracy across varying $k$. Recognition accuracy consistently declines across all modalities as the candidate pool expands, demonstrating that MLLMs are sensitive to increasing contextual set and inter-class ambiguity.

The multimodal setting (description + sketch) achieves the highest recognition accuracy up to a candidate set size of five. As shown in Figure 3, multimodal fusion is beneficial for small pools because sketch geometry and textual descriptions provide complementary identity cues. However, as $k$ expands, textual descriptions become increasingly ambiguous since many identities share similar facial attributes. Across identities, we observe only 361 unique attribute combinations, with 245 combinations shared by identities. Textual embeddings are highly clustered, exhibiting a mean pairwise cosine similarity of $0.963 \pm 0.017$ and a mean attribute Jaccard similarity of $0.496$.

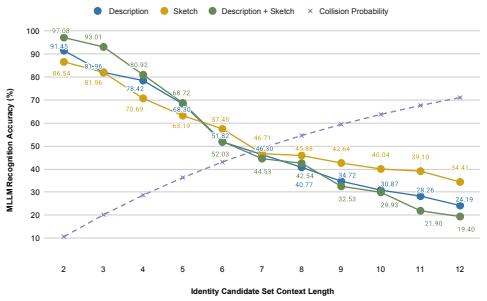



Figure 3: Recognition accuracy of MLLMs under varying identity candidate-set context lengths with FaceLLM-38B generated descriptions. Other cross-MLLM results are present in the Appendix A.1.

To quantify this ambiguity, we define the description collision probability, i.e., the probability that a randomly selected gallery shares the same attribute description as the probe:

$$p = \sum_{c \in \mathcal{A}} \frac{n_c}{|\mathcal{I}|} \cdot \frac{n_c - 1}{|\mathcal{I}| - 1}, \tag{2}$$

where $\mathcal{A}$ denotes the set of observed attribute combinations, $|\mathcal{I}|$ is the total number of identities and $n_c$ is the number of identities with combination $c$. For a set of size $k$, the probability that at least one distractor collides semantically with the probe becomes

$$P_{\text{col}}(k) = 1 - (1 - p)^{k-1}, \tag{3}$$

which increases monotonically with $k$. We plot $P_{\text{col}}(k)$ alongside the recognition curves in Figure 3. Interestingly, the collision curve intersects the multimodal accuracy trend near the point where multimodal performance begins to deteriorate relative to sketch-only recognition (approximately five candidates). Beyond this point, $P_{\text{col}}(k)$ rises rapidly, indicating a high likelihood of indistinguishable distractors, thus, description-based reasoning saturates, multimodal fusion becomes increasingly ambiguous, and MLLMs rely heavily on sketches. Sketches preserve identity-specific

Table 2: Recognition accuracy (%) of different MLLMs across multiple datasets under 3 modality settings (descriptions generated by FaceLLM-38B). Similar trend for descriptions generated using InternVL2-26B and Qwen2.5-VL-32B are present in the Appendix A.2.

| MLLMs | Modality | CUFSF | PRIP-HDC | FS2K | IIIT-D Sketch (Viewed Sketch) | IIIT-D Sketch (Semi-Forensic) | CelebA |
|---|---|---|---|---|---|---|---|
| **FaceLLM-38B** | **Description** | 86.34 | 75.00 | 90.91 | 76.19 | 61.15 | 93.40 |
| | **Sketch** | 88.84 | 70.45 | 91.67 | **96.54** | **81.29** | 89.90 |
| | **Description + Sketch** | **93.22** | **79.55** | **96.27** | 93.51 | 74.82 | **95.90** |
| **InternVL2-26B** | **Description** | 80.19 | 59.09 | 81.72 | 68.40 | 50.36 | 81.60 |
| | **Sketch** | 70.70 | 63.64 | 75.22 | 74.46 | 52.52 | 75.10 |
| | **Description + Sketch** | **89.26** | **77.27** | **90.62** | **85.71** | **64.03** | **90.30** |
| **Qwen2.5-VL-32B** | **Description** | 80.40 | 65.91 | 68.13 | 64.50 | 48.92 | 80.80 |
| | **Sketch** | 65.28 | 52.27 | 57.98 | 68.83 | 46.76 | 60.10 |
| | **Description + Sketch** | **85.82** | **75.00** | **79.04** | **75.76** | **51.08** | **81.90** |

geometric cues that remain discriminative even under attribute collisions, allowing the sketch-only modality to outperform other inputs for larger candidate sets.

*Implication:* Current MLLMs do not scale reliably to large candidate pools, highlighting the need for candidate-aware multimodal reasoning in realistic forensic identification.

**Hypothesis (H2): Impact of Distinct Modality Influence on MLLM's Identity Understanding:** *Multimodal integration enables MLLMs to better understand identity features from facial descriptions and sketches to improve forensic identity recognition compared to unimodal inputs.*

To isolate modality-specific effects, we evaluate FaceLLM-38B Shahreza & Marcel (2025b), InternVL2-26B Chen et al. (2024), Qwen2.5-VL-32B Bai et al. (2025), and Qwen2.5-Omni Xu et al. (2025) across the five input configurations $m$ defined in Eq.1. Figure 2 (middle) represents the overview of the evaluation protocol to build upon each modality and their combinations.

***Textual Description only.*** MLLMs are widely used to generate facial descriptions from images. Here, we investigate whether these descriptions alone are sufficiently discriminative enough among multiple candidates. (Table 1). Results in Table 2 show that description-based recognition achieves performance comparable to sketch-based settings. The influence of individual attributes is analyzed in Hypothesis *H3 (1)*, with examples provided in the Appendix A.2.

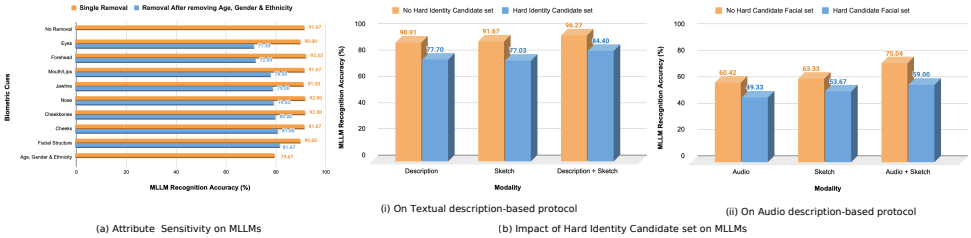
***Sketch only.*** In the sketch-only setting, MLLMs are provided only with abstract sketches and identity candidate sets, enabling evaluation of how structural and contour-based facial cues support recognition. As shown in Table 2, MLLMs achieve strong yet modality-dependent performance using sketches alone. FaceLLM-38B, being specialized for facial understanding, consistently outperforms InternVL2-26B and Qwen2.5-VL-32B across all datasets. Performance is also lower on semi-forensic than viewed sketches, indicating that increasing sketch abstraction and the loss of fine-grained facial details make identity discrimination more challenging.

***Textual Description and Sketch.*** Unimodal visual inputs, though informative, may not fully capture identity-specific nuances, motivating the integration of multimodal cues for forensic recognition. Under the modality influence hypothesis (H2), we hypothesize that combining facial descriptions with sketches improves identity discrimination through cross-modal reasoning. As shown in Table 2, multimodal fusion outperforms both sketch-only and description-only settings across most datasets and models. For example, FaceLLM-38B improves from 88.84% (sketch) and 86.34% (text) to 93.22% on CUFSF, from 91.67% and 90.91% to 96.27% on FS2K, and from 89.90% and 93.40% to 95.90% on CelebA, with similar trends for InternVL2-26B and Qwen2.5-VL-32B. These results indicate that textual biometric descriptions provide complementary semantic cues that reinforce the structural information in sketches, reducing identity ambiguity. Overall, *H2* is validated: complementary visual and semantic cues enable more robust and discriminative forensic recognition than unimodal inputs. This agrees with our forensic expert evaluation, where experts most often selected description+sketch as the most complete information for identification (Appendix F).

**Comparison with Conventional and Embedding-based Recognition.** We further compare MLLM-based forensic identification with a non-MLLM-based conventional recognition approach. FaceLLM achieves higher overall accuracy than ArcFace-iResNet-100 (89.97% vs. 77.37%), with the advantage becoming more pronounced as sketches depart from photographic appearance, while direct Qwen3-VL embedding retrieval yields low Rank-1 performance over a large gallery set. These results suggest that the observed gains arise from multimodal semantic reasoning using MLLM's

Table 3: Recognition accuracy (%) of MLLM (Qwen2.5-Omni) using AI-generated audio facial descriptions across different accents and speaker genders, compared with text, sketch, and text + sketch modalities. **AA**: Average Audio and **AAS**: Average Audio+Sketch.

| Accent | Speaker's Gender | Audio | Accent AA | Overall AA | Audio +Sketch | Accent AAS | Overall AAS | Text | Sketch | Text +Sketch |
|---|---|---|---|---|---|---|---|---|---|---|
| **American** | **Female** | 63.67 | 62.67 | 60.42 | 73.67 | 73.50 | **75.04** | 69.00 | 63.33 | 74.67 |
| | **Male** | 61.67 | | | 73.33 | | | | | |
| **British** | **Female** | 59.33 | 60.50 | | 75.00 | 74.16 | | | | |
| | **Male** | 61.67 | | | 73.33 | | | | | |
| **Chinese** | **Female** | 60.33 | 60.50 | | 77.00 | 76.66 | | | | |
| | **Male** | 60.67 | | | 76.33 | | | | | |
| **Indian** | **Female** | 58.67 | 58.00 | | 75.33 | 75.83 | | | | |
| | **Male** | 57.33 | | | 76.33 | | | | | |




(a) Attribute Sensitivity on MLLMs

(i) On Textual description-based protocol

(ii) On Audio description-based protocol

(b) Impact of Hard Identity Candidate set on MLLMs

Figure 4: (a) Attribute sensitivity analysis showing the effect of removing individual biometric facial cues on MLLM recognition. (b) Recognition performance on hard candidate sets with pose variations and visually similar identities, under textual and audio description protocols, demonstrating increased confusion under fine-grained inter-class similarity.

understanding ability rather than embedding-based similarity using longer/detailed textual facial descriptions; detailed results are provided in the Appendix D and Appendix E.

***MLLMs over Audio Narratives Only.*** To evaluate MLLMs under realistic verbal testimony conditions, we conduct audio-only recognition experiments on FS2K using Qwen2.5-Omni Xu et al. (2025). As shown in Table 3, audio-only recognition underperforms text-only, indicating that speech-based semantic understanding alone is less discriminative than structured textual input. MLLMs are robust to accent and gender variations under synthetic TTS. However, on 100 human-narrated and TTS-generated audio descriptions (protocol in Table 1), ***human-recorded*** speech yields lower accuracy (55.56%; F: 54.12%, M: 57.00%) than ***synthetic TTS*** (63.88%; F: 64.00%, M: 63.75%). This reflects sensitivity to acoustic variability in speech rate, intonation, and articulation that is absent from synthesized speech.

***MLLMs over Speaker-Narrated Audio Descriptions and Sketch.*** As shown in Table 3, combining audio descriptions with sketches substantially improves recognition (73–76%), indicating that visual grounding resolves ambiguity in speech-based descriptions. Similar trends hold for 200 ***human-recorded audios*** (100 female (F), 100 male (M)) from FS2K (protocol in Table 1), where multimodal performance reaches 75.32% (F: 77.65%, M: 73.00%), closely matching synthetic TTS at 74.88% (F: 73.75%, M: 76.00%). Across both, adding sketches improves performance by 10–20% over audio-only inputs, and audio+sketch (75.04%) can match or exceed text+sketch (74.67%), showing that speech-derived cues become highly effective when grounded by structural visual information.

These findings further support the modality hypothesis in realistic forensic settings: heterogeneous modalities provide complementary identity cues. While sketches capture structural facial geometry, audio contributes fine-grained semantic descriptions, and their integration reduces cross-modal ambiguity, leading to more reliable identity recognition.

*Implication:* Multimodal fusion consistently enhances forensic identity reasoning, suggesting that future MLLMs should be designed to jointly exploit structural and semantic facial cues.

**Hypothesis (H3): MLLMs exhibit confusion under fine-grained intra- and inter-class variations:** *MLLMs become confused when candidate identities share similar biometric facial attributes, leading to reduced decision stability under fine-grained intra- and inter-class variations.*

We evaluate MLLMs under threefold protocols for $H3$:

**Attribute Severity:** We investigate the contribution of individual facial attributes to MLLM-based identity recognition. Since eyewitnesses rarely recall every facial cue, identifying the most discrim-

Table 4: Recognition accuracy (%) of MLLMs under intra-class pose variations across CUFSF and FS2K datasets for various modality combinations of textual face description and sketch. The pose settings are represented as (Pitch, Yaw), where $+/-$ denote up/down pitch and right/left yaw, respectively. Other MLLM-based results are available in Appendix B.

| Datasets | MLLMs | Modality | +0 +0 | +0 +30 | +0 -45 | +15 +0 | +15 +30 | +15 -30 |
|---|---|---|---|---|---|---|---|---|
| CUFSF | FaceLLM-38B | Description | 86.34 | 78.62 | 81.13 | 79.04 | 79.67 | 79.14 |
| | | Sketch | 88.84 | 75.08 | 78.10 | 78.83 | 77.27 | 78.10 |
| | | Description + Sketch | 93.22 | 85.09 | 85.40 | 86.44 | 86.03 | 86.55 |
| FS2K | FaceLLM-38B | Description | 90.91 | 83.73 | 85.93 | 83.06 | 84.4 | 84.02 |
| | | Sketch | 91.67 | 83.73 | 85.36 | 85.17 | 83.16 | 82.49 |
| | | Description + Sketch | 96.27 | 88.71 | 91.00 | 90.14 | 89.28 | 88.42 |

inative attributes is crucial for forensic recognition. We evaluate the facially specialized *FaceLLM-38B* on 300 FS2K samples by systematically removing individual cues (Figure 2, right).

Figure 4a shows that MLLMs rely most heavily on demographic attributes (*age, gender, and ethnicity*), whose removal causes the largest accuracy drop ($\sim 12\%$), followed by structural cues such as *jawline* and *facial structure*. Interestingly, removing *forehead*, *nose*, or *cheekbones* slightly improves recognition, suggesting these attributes are weakly discriminative and may introduce ambiguity.

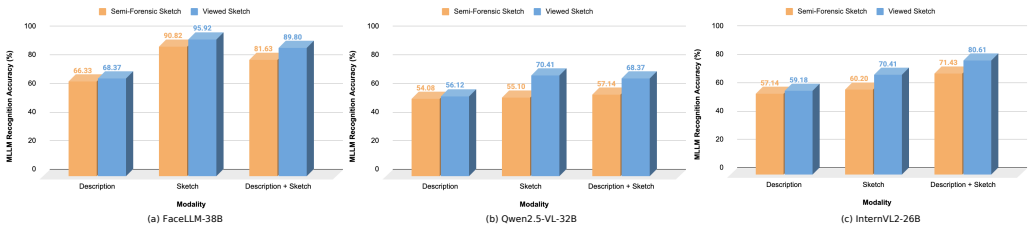
To isolate biometric cues, we repeat the analysis after removing demographic information. The *eyes* then emerge as the most informative attribute, causing the largest accuracy drop ($\sim 8\%$), while the *forehead* also becomes highly discriminative, suggesting its contribution was masked by demographic cues. The *nose*, *cheekbones*, and *cheeks* remain weak identity cues. This reveals a hierarchical reasoning strategy: MLLMs first rely on demographic cues, then on distinctive biometric cues, particularly the *eyes* and *forehead*, for fine-grained discrimination.

**Hard Candidate Set:** To simulate challenging forensic scenarios, we construct hard candidate sets under two settings: (a) pose variations and (b) visually similar identities. For the latter, cosine similarity scores from the ArcFace Deng et al. (2019) face recognition model are used to select the top-3 most similar identities for each probe, forming a candidate set $\mathcal{C}$ with $k = 4$, consisting of the correct identity and three highly similar distractors.

***(a) Pose Variations.*** Candidate faces are generated using Arc2Face Papantoniou et al. (2024) with controlled pitch and yaw variations to simulate realistic viewpoint discrepancies between probe sketches and candidate images. As shown in Table 4, pose changes reduce accuracy across all settings: text-only (∼7.8% drop), sketch-only (∼13.0%), and sketch+description (∼10.6%), highlighting the sensitivity of MLLMs to geometric variations. The smaller drops for text-only and sketch+description indicate that textual descriptions are more pose-invariant, encoding semantic attributes (e.g., gender, hair, accessories, facial traits) that remain stable across viewpoints, whereas sketches depend on structural cues such as the jawline, nose, and eye visibility. Multimodal fusion thus compensates for pose-induced visual inconsistencies; e.g., on FS2K, FaceLLM drops from 96.27% (frontal) to 88.71% under pose variation while still outperforming unimodal settings. Results for additional MLLMs are in Appendix B. Overall, semantic descriptions improve robustness to pose variations by reducing visual ambiguity.

***(b) Visually Similar Faces.*** In real forensic investigations, eyewitnesses often narrow the search to a small pool of visually similar suspects. To evaluate MLLMs in this challenging yet realistic setting, we construct a hard candidate set by selecting visually similar identities using ArcFace Deng et al. (2019) similarity scores. On FS2K (1045 samples), performance drops to $77.70\%$ (description), $77.03\%$ (sketch), and $84.40\%$ (description+sketch), compared to $90.91\%$, $91.67\%$, and $96.27\%$ on the non-hard set (Figure 4b(i)). A similar trend holds for the audio-based protocol on 300 FS2K hard samples using Qwen2.5-Omni Xu et al. (2025), with the largest drop for audio+sketch (16.04%), followed by audio (11.09%) and sketch (9.66%). These results show that MLLMs struggle with highly similar facial structures, highlighting the need for stronger fine-grained identity representations.

**Partial Modality Semantics:** In real investigations, eyewitnesses often observe only parts of a face, resulting in incomplete facial descriptions and partially drawn sketches. To replicate this, we evaluate MLLMs using sketches with either the upper or lower region masked, accompanied by descriptions for selected attributes. The performance degrades, with FaceLLM-38B dropping from 93.51% (original) to 83.12% (upper masked) and 87.01% (lower masked) on the IIIT-D Viewed dataset. Also, multimodal reasoning improves recognition over sketch-only inputs, increasing ac-




(a) FaceLLM-38B (b) Qwen2.5-VL-32B (c) InternVL2-26B

Figure 6: Effect of memory-based sketch acquisition on recognition performance, evaluated on common subjects of Viewed and Semi-Forensic IIIT-D Sketch database.

curacy from 79.22% to 83.12% for upper-masked and from 82.25% to 87.01% for lower-masked. Similar trends are observed on the IIIT-D Semi-Forensic dataset as shown in Appendix C.

*Implication:* *H3* is supported, showing that MLLMs become increasingly confused under fine-grained intra- and inter-class variations. They rely heavily on demographic attributes (*age, gender, and ethnicity*) and key structural cues, particularly the *eyes* and *forehead*, making recognition vulnerable to missing cues, pose changes (intra-class), and visually similar identities.

## 5 Challenges of Memory-Based Sketches

We evaluate the effect of sketch acquisition on subjects common to the IIIT-D Viewed Sketch and IIIT-Delhi Semi-Forensic Sketch datasets. Unlike viewed sketches, which are drawn with direct visual reference, semi-forensic sketches are drawn from memory, losing fine-grained identity cues. As shown in Figure 6, FaceLLM-38B performs better on viewed sketches across all settings: description (66.33% $\rightarrow$ 68.37%), sketch (90.82% $\rightarrow$ 95.92%), and description+sketch (81.63% $\rightarrow$ 89.80%), with similar trends for Qwen2.5-VL-32B and InternVL2-26B.

**Real-world forensic case study:** We further evaluate FaceLLM on three publicly available forensic sketch videos in which artists draw suspects from eyewitness descriptions (Figure 5; others in Appendix G, Figure 11). Given one genuine and three visually similar candidates, sketch+description identified the suspect in all three cases, while sketch-only failed in one, underscoring the value of multimodal cues. Overall, memory-based sketches are harder due to reduced structural fidelity, and eyewitness descriptions help compensate. See Appendix G for details and Appendix H for a discussion on forensic evidence independence.

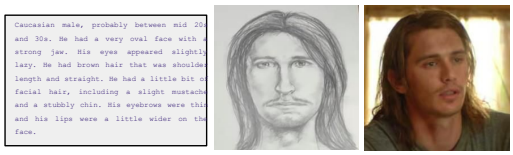



Figure 5: A sample of real-world scenario where only memory-based facial descriptions and sketches are available for identification.

## 6 Conclusion

Our study shows that while MLLMs benefit from multimodal evidence, their identity reasoning remains sensitive to candidate-set scaling and fine-grained facial similarity. Recognition degrades as the candidate pool grows, and models struggle when identities share similar facial attributes or critical biometric cues are absent. Although combining sketches with textual and audio descriptions improves performance by fusing structural and semantic information, current MLLMs rely on a limited set of discriminative cues rather than a holistic representation. These findings point to stronger fine-grained identity representations, attribute-aware reasoning, and candidate-aware retrieval as directions for improving robustness under realistic forensic conditions. We believe *MS-Forensics* provides the first systematic benchmark for multimodal sketch-based identity reasoning. ***The benchmark, code, and protocols will be released upon acceptance.***

### AI Use Statement

In this work, we used generative AI tools for tasks such as sketch generation (using open-source Diffusion models) and facial description generation (using open-source MLLMs). Also used gen-

erative AI tools for a paper-writing task to rephrase some sentences. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

### ETHICS STATEMENT

This work studies the capabilities and limitations of MLLMs for sketch-based forensic identity recognition, a task with clear societal implications. All experiments are conducted on existing, publicly available face and face–sketch datasets (CUFSF, IIIT-D Sketch, PRIP-HDC, FS2K, and CelebA) released under established research protocols, and no new biometric data of real individuals is collected. Human-recorded audio descriptions were contributed by consenting volunteer speakers in equal numbers of males and females, describing faces from these public datasets. We evaluate open-source MLLMs locally, ensuring that no biometric data is transmitted to third-party services.

This paper does not propose a deployable identification system, nor do we advocate the use of current MLLMs in operational forensic settings. On the contrary, our results show that MLLM recognition degrades sharply with candidate-set size, pose variation, and inter-class similarity, underscoring that such models are unsuitable for autonomous identity decisions. Our attribute analysis further shows that MLLM predictions rely heavily on demographic attributes (age, gender, and ethnicity), a dependence that could translate into demographically uneven error rates; we report this as a caution and argue that any downstream use must be assessed for demographic fairness. We view forensic identification as a domain where algorithmic outputs may only ever serve as investigative aids under qualified human review, never as evidence of identity. MS-Forensics is released for research purposes, subject to the licenses of the constituent datasets, to support systematic and transparent evaluation of these limitations.

### REPRODUCIBILITY STATEMENT

We will release the MS-Forensics benchmark, code, and protocols upon acceptance of the paper for smoother reproducibility. Currently, in this paper, we have provided samples from the dataset for the reviewer's reference.

## REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025.

Himanshu S Bhatt, Samarth Bharadwaj, Richa Singh, and Mayank Vatsa. Memetically optimized mcwld for matching sketches with digital face images. *IEEE Transactions on Information Forensics and Security*, 7(5):1522–1535, 2012.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24185–24198, 2024.

Kanzhi Cheng, Wenpo Song, Jiaxin Fan, Zheng Ma, Qiushi Sun, Fangzhi Xu, Chenyang Yan, Nuo Chen, Jianbing Zhang, and Jiajun Chen. Caparena: Benchmarking and analyzing detailed image captioning in the llm era. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 14077–14094, 2025.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4685–4694, 2019.

Deng-Ping Fan, Ziling Huang, Peng Zheng, Hong Liu, Xuebin Qin, and Luc Van Gool. Facial-sketch synthesis: A new challenge. *Machine Intelligence Research*, 19(4):257–287, 2022.

Scott J Klum, Hu Han, Brendan F Klare, and Anil K Jain. The facesketchid system: Matching facial composites to mugshots. *IEEE Transactions on Information Forensics and Security*, 9(12): 2248–2263, 2014.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, pp. 19730–19742, 2023.

Ming Li, Taojiannan Yang, Huafeng Kuang, Jie Wu, Zhaoning Wang, Xuefeng Xiao, and Chen Chen. Controlnet++: Improving conditional controls with efficient consistency feedback. In *European Conference on Computer Vision*, pp. 129–147, 2024.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *IEEE/CVF International Conference on Computer Vision*, pp. 3730–3738, 2015.

Kartik Narayan, V. S. Vibashan, and Vishal M. Patel. Facexbench: Evaluating multimodal llms on face understanding. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 8(3): 354–364, 2026.

Foivos Paraperas Papantoniou, Alexandros Lattas, Stylianos Moschoglou, Jiankang Deng, Bernhard Kainz, and Stefanos Zafeiriou. Arc2face: A foundation model for id-consistent human faces. In *European Conference on Computer Vision*, pp. 241–261, 2024.

Lixiong Qin, Shilong Ou, Miaoxuan Zhang, Jiangning Wei, Yuhang Zhang, Xiaoshuai Song, Yuchen Liu, Mei Wang, and Weiran Xu. Face-human-bench: A comprehensive benchmark of face and human understanding for multi-modal assistants. *arXiv preprint arXiv:2501.01243*, 2025.

Hatef Otroshi Shahreza and Sébastien Marcel. Benchmarking multimodal large language models for face recognition. *arXiv preprint arXiv:2510.14866*, 2025a.

Hatef Otroshi Shahreza and Sébastien Marcel. Facellm: A multimodal large language model for face understanding. In *IEEE/CVF International Conference on Computer Vision*, pp. 3677–3687, 2025b.

Hatef Otroshi Shahreza and Sébastien Marcel. Foundation models and biometrics: A survey and outlook. *IEEE Transactions on Information Forensics and Security*, 2025c.

Hatef Otroshi Shahreza, Anjith George, and Sébastien Marcel. Evaluating multimodal large language models for heterogeneous face recognition. *arXiv preprint arXiv:2601.15406*, 2026.

Redwan Sony, Parisa Farmanifard, Arun Ross, and Anil K Jain. Foundation versus domain-specific models: Performance comparison, fusion, and explainability in face recognition. In *IEEE/CVF International Conference on Computer Vision*, pp. 3656–3666, 2025.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Xiaogang Wang and Xiaoou Tang. Face photo-sketch synthesis and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(11):1955–1967, 2009.

Xiaoqin Wang, Xusen Ma, Xianxu Hou, Meidan Ding, Yudong Li, Junliang Chen, Wenting Chen, Xiaoyang Peng, and Linlin Shen. Facebench: A multi-view multi-level facial attribute vqa dataset for benchmarking face perception mllms. In *IEEE/CVF Computer Vision and Pattern Recognition Conference*, pp. 9154–9164, 2025.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report, 2025.

Hong-Tao Yu, Xiu-Shen Wei, Yuxin Peng, and Serge Belongie. Benchmarking large vision-language models on fine-grained image tasks: A comprehensive evaluation. *arXiv preprint arXiv:2504.14988*, 2025.

# APPENDIX

Table 5: Recognition accuracy (%) of different MLLMs across multiple datasets under three Multimodal settings of textual-based facial descriptions and sketch, highlighting the benefits of multimodal integration. ***The facial descriptions are computed using the InternVL2-26B model.***

| MLLMs | Modality | CUFSF | PRIP-HDC | FS2K | IIIT-D Sketch (Viewed Sketch) | IIIT-D Sketch (Semi-Forensic Sketch) |
|---|---|---|---|---|---|---|
| **FaceLLM-38B** | **Description** | 85.19 | 70.45 | 84.21 | 77.06 | 59.71 |
| | **Sketch** | 88.84 | 70.45 | 91.67 | **96.54** | **81.29** |
| | **Description + Sketch** | **93.85** | **84.09** | **93.40** | 91.77 | 76.26 |
| **InternVL2-26B** | **Description** | 80.19 | 68.18 | 77.70 | 69.70 | 56.83 |
| | **Sketch** | 70.70 | 63.64 | 75.22 | 74.46 | 52.52 |
| | **Description + Sketch** | **93.33** | **77.27** | **91.00** | **86.15** | **70.50** |
| **Qwen2.5-VL-32B** | **Description** | 77.69 | 79.55 | 79.23 | 73.16 | **55.40** |
| | **Sketch** | 65.28 | 52.27 | 68.13 | 68.83 | 46.76 |
| | **Description + Sketch** | **85.82** | **79.55** | **80.38** | **75.76** | 49.64 |

Table 6: Recognition accuracy (%) of different MLLMs across multiple datasets under three Multimodal settings of textual-based facial descriptions and sketch, highlighting the benefits of multimodal integration. ***The facial descriptions are computed using the Qwen-2.5VL-32B model.***

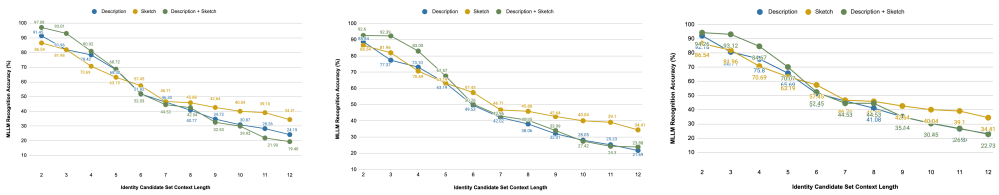
| MLLMs | Modality | CUFSF | PRIP-HDC | FS2K | IIIT-D Sketch (Viewed Sketch) | IIIT-D Sketch (Semi-Forensic Sketch) |
|---|---|---|---|---|---|---|
| **FaceLLM-38B** | **Description** | 84.36 | 61.36 | 82.30 | 71.86 | 61.15 |
| | **Sketch** | 88.84 | 70.45 | 91.67 | **96.54** | **81.29** |
| | **Description + Sketch** | **94.26** | **72.73** | **91.29** | 87.88 | 74.82 |
| **InternVL2-26B** | **Description** | 74.56 | 45.45 | 71.96 | 59.74 | 50.36 |
| | **Sketch** | 70.70 | 63.64 | 75.22 | 74.46 | 52.52 |
| | **Description + Sketch** | **89.78** | **68.18** | **86.99** | **81.39** | **64.03** |
| **Qwen2.5-VL-32B** | **Description** | 78.31 | 70.45 | 77.99 | 70.13 | 54.68 |
| | **Sketch** | 65.28 | 52.27 | 68.13 | 68.83 | 46.76 |
| | **Description + Sketch** | **85.09** | **84.09** | **80.19** | **74.03** | **54.68** |

Table 7: Recognition accuracy (%) of different MLLMs across multiple datasets under three Multimodal settings of textual-based facial descriptions and sketch, highlighting the benefits of multimodal integration. ***The facial descriptions are computed using the FaceLLM-38B model.***

| MLLMs | Modality | CUFSF | PRIP-HDC | FS2K | IIIT-D Sketch (Viewed Sketch) | IIIT-D Sketch (Semi-Forensic Sketch) |
|---|---|---|---|---|---|---|
| **FaceLLM-38B** | **Description** | 86.34 | 75.00 | 90.91 | 76.19 | 61.15 |
| | **Sketch** | 88.84 | 70.45 | 91.67 | **96.54** | **81.29** |
| | **Description + Sketch** | **93.22** | **79.55** | **96.27** | 93.51 | 74.82 |
| **InternVL2-26B** | **Description** | 80.19 | 59.09 | 81.72 | 68.40 | 50.36 |
| | **Sketch** | 70.70 | 63.64 | 75.22 | 74.46 | 52.52 |
| | **Description + Sketch** | **89.26** | **77.27** | **90.62** | **85.71** | **64.03** |
| **Qwen2.5-VL-32B** | **Description** | 80.40 | 65.91 | 68.13 | 64.50 | 48.92 |
| | **Sketch** | 65.28 | 52.27 | 57.98 | 68.83 | 46.76 |
| | **Description + Sketch** | **85.82** | **75.00** | **79.04** | **75.76** | **51.08** |

## A CROSS-MLLM FACIAL DESCRIPTION EVALUATIONS

To analyze the influence of facial descriptions generated by different MLLMs, we conduct cross-model evaluations where semantic facial descriptions produced by one MLLM are used as inputs for




(a) Facial descriptions computed using FaceLLM (b) Facial descriptions computed using InternVL2 (c) Facial descriptions computed using Qwen2.5-VL

Figure 7: Effect of identity candidate-set context length on recognition accuracy when facial descriptions are generated for CUFSF dataset using different MLLMs: (a) FaceLLM, (b) InternVL2, and (c) Qwen2.5-VL, while evaluation is performed using FaceLLM.

Table 8: Recognition accuracy (%) of MLLMs under pose variations across CUFSF and FS2K datasets for various modality combinations of textual face description and sketch. The pose settings are represented as (Pitch, Yaw), where $+/-$ denote up/down pitch and right/left yaw, respectively.

| Datasets | MLLMs | Modality | +0 +0 | +0 +30 | +0 -45 | +15 +0 | +15 +30 | +15 -30 |
|---|---|---|---|---|---|---|---|---|
| CUFSF | FaceLLM-38B | Description | 86.34 | 78.62 | 81.13 | 79.04 | 79.67 | 79.14 |
| | | Sketch | 88.84 | 75.08 | 78.10 | 78.83 | 77.27 | 78.10 |
| | | Sketch + Description | 93.22 | 85.09 | 85.40 | 86.44 | 86.03 | 86.55 |
| | InternVL2-26B | Description | 80.19 | 76.23 | 76.43 | 76.23 | 76.54 | 75.29 |
| | | Sketch | 70.70 | 61.73 | 58.08 | 62.77 | 63.40 | 61.73 |
| | | Sketch + Description | 89.26 | 80.60 | 79.25 | 79.14 | 80.50 | 81.13 |
| | Qwen2.5-VL-32B | Description | 80.4 | 77.27 | 76.96 | 78.31 | 75.91 | 78.31 |
| | | Sketch | 65.28 | 59.85 | 62.04 | 65.38 | 61.31 | 62.98 |
| | | Sketch + Description | 85.82 | 78.10 | 78.10 | 80.08 | 79.87 | 79.04 |
| FS2K | FaceLLM-38B | Description | 90.91 | 83.73 | 85.93 | 83.06 | 84.4 | 84.02 |
| | | Sketch | 91.67 | 83.73 | 85.36 | 85.17 | 83.16 | 82.49 |
| | | Description + Sketch | 96.27 | 88.71 | 91.00 | 90.14 | 89.28 | 88.42 |
| | InternVL2-26B | Description | 81.72 | 74.93 | 75.41 | 75.31 | 75.12 | 74.93 |
| | | Sketch | 75.22 | 64.31 | 63.64 | 68.71 | 64.50 | 62.58 |
| | | Description + Sketch | 90.62 | 83.06 | 80.86 | 82.49 | 81.53 | 80.00 |
| | Qwen2.5-VL-32B | Description | 79.04 | 75.89 | 78.18 | 75.98 | 74.74 | 75.98 |
| | | Sketch | 68.13 | 62.49 | 64.4 | 66.32 | 65.26 | 65.17 |
| | | Description + Sketch | 82.20 | 76.17 | 79.62 | 80.19 | 76.94 | 78.66 |

identity recognition using another MLLM. This setup helps analyze the consistency, discriminative quality, and cross-model transferability of semantic facial descriptions for identity reasoning.

### A.1 Identity Candidate Set Context Length Comparisons

Figure 7 illustrates the effect of identity candidate-set context length on recognition performance on the CUFSF dataset when facial descriptions are generated using different MLLMs ((a) FaceLLM, (b) InternVL2, and (c) Qwen2.5-VL) while evaluation is performed using FaceLLM. Across all three settings, recognition accuracy consistently decreases as the candidate identity context length increases, regardless of which model generates the facial descriptions. Importantly, the overall degradation trend remains similar across all description sources, indicating that the decline in performance is not primarily caused by the quality or origin of the generated descriptions. Instead, the results suggest that the main limiting factor is the increasing candidate set context itself, as it especially increases beyond five, which introduces higher search complexity and inter-class similarity among identities. Also, as the facial images based context length increases beyond five, the sketch-based recognition is effective as compared to description and sketch–description combination–based recognition. This trend is observed because when comparing across a larger set of identities using MLLMs, the understanding through sketch-based discriminability is more than description-based discriminability.

### A.2 Impact of Facial Description Source on Identity Recognition

Tables 5, 6 and 7 compare recognition performance when facial descriptions are generated by different MLLMs (InternVL2, Qwen2.5-VL, and FaceLLM) and subsequently used as semantic cues for identity recognition across multiple MLLM evaluators. The results reveal trends regarding the

Table 9: Partial sketch visibility settings.

| Model | Input | IIIT-D Viewed Forensic | | | IIIT-D Semi Forensic | | |
|---|---|---|---|---|---|---|---|
| | | Original | Upper | Lower | Original | Upper | Lower |
| **FaceLLM-38B** | **Sketch** | 96.54 | 79.22 | 82.25 | 81.29 | 64.54 | 58.99 |
| | **Description + Sketch** | 93.51 | 83.12 | 87.01 | 74.82 | 66.91 | 63.31 |
| **Qwen2.5 VL-32B** | **Sketch** | 68.83 | 54.11 | 46.32 | 46.76 | 34.53 | 33.81 |
| | **Description + Sketch** | 75.76 | 69.70 | 66.23 | 51.08 | 41.01 | 41.73 |

quality and discriminative power of descriptions produced by different multimodal large language models.

The critical analyses conducted based on these quantitative evaluations shown in Table 5, Table 6 and Table 7 are mentioned as follows:

- FaceLLM-generated descriptions consistently yield superior recognition performance across the majority of datasets and evaluation models. In the sketch+description setting, FaceLLM-38B achieves 96.27% (Table 7) accuracy on FS2K and 93.51% (Table 7) on IIIT-D Viewed sketches when utilizing these descriptions, which is comparable to the Description + Sketch combination. Similarly, InternVL2-26B reaches 90.62% (Table 7) on FS2K with FaceLLM's description, surpassing results derived from InternVL or Qwen descriptions. This suggests that FaceLLM's specialized architecture for facial reasoning enables the production of more precise, identity-discriminative attributes.
- InternVL2 descriptions demonstrate moderate efficacy, typically outperforming Qwen2.5-VL while trailing behind FaceLLM. For example, FaceLLM reaches 93.40% (Table 5) on FS2K with multimodal inputs using InternVL descriptions, which is a notable result, yet inferior to the FaceLLM-description baseline. This implies that while InternVL2 captures meaningful facial semantics, it lacks the granular precision of FaceLLM.
- Qwen2.5-VL generated descriptions result in the lowest recognition accuracy, particularly on difficult datasets like semi-forensic IIIT-D sketches. This performance gap suggests that Qwen's descriptions lack the detailed biometric cues required for robust identity matching.

Overall, these evaluations show that FaceLLM-generated descriptions provide the most informative semantic cues, giving the highest recognition performance on most datasets, and that the multimodal setting (Sketch + Description) complements both sketch-only and description-only evaluation. Read across evaluators, the tables also serve as a check against generator-evaluator leakage: no model gains a consistent advantage from consuming its own descriptions. Models trained specifically for facial reasoning, such as FaceLLM, therefore generate more discriminative descriptions than general-purpose MLLMs.

## B Pose Variations with Distinct MLLMs

Table 8 presents an interesting observation that the smaller decrease in the description-only condition is observed, likely because textual descriptions are comparatively more pose-invariant than sketch or image-based representations. Changes in head orientation can significantly alter visible structural facial cues in sketches/images (e.g., jawline, nose shape, symmetry, or eye visibility), which directly impacts visual matching performance. In contrast, textual descriptions typically encode higher-level semantic attributes (e.g., hair color, gender, accessories, facial traits) that remain relatively stable across pose variations. As a result, pose changes introduce less degradation in the description-only setting compared to visually grounded conditions.

## C Partial Modality Semantics

Table 9 shows results with distinct MLLMs for evaluation on partial sketch observation, to mimic a real-world scenario, by masking the upper or lower portion of the sketch. Further, obtaining their region-specific descriptions. Table 9 illustrates an evident drop in performance when the partial face is masked. Multimodal cues help with performance improvement as compared to sketch only, but still lack complete visual supervision.

Table 10: Comparing MLLMs' recognition ability over the non-MLLM recognition model's performance (%) over the same candidate set protocol used in Table 2, and the hard candidate set for FS2K dataset.

| **Dataset** | **MLLM** *(FaceLLM-38B)* | **non-MLLM** *(ArcFace-iResNet-100)* |
|---|---|---|
| CUFSF | 93.22 | **97.71** |
| IIIT-D Semi-Forensic | **74.82** | 56.83 |
| IIIT-D Viewed sketch | **93.51** | 77.92 |
| PRIP-HDC | **79.55** | 54.55 |
| CelebA | **95.90** | 93.40 |
| FS2K-hard set | **84.40** | 83.83 |
| **Overall** | **89.97** | 77.37 |

## D MLLMs vs. Conventional non-MLLM Models

Table 10 compares FaceLLM with a strong non-MLLM face recognizer, ArcFace-iResNet-100, applied zero-shot to sketch-forensic identification task. FaceLLM achieves higher overall accuracy (89.97% vs. 77.37%) and outperforms ArcFace on five of six datasets. The margin, however, depends strongly on how far the sketch departs from the photograph. On artist sketches drawn directly from photographs, which largely preserve facial geometry, the two approaches are comparable (FS2K-hard: 84.40 vs. 83.83), and ArcFace performs best on CUFSF (97.71 vs. 93.22). As sketches become less faithful, ArcFace degrades sharply, falling from 77.92% on IIIT-D Viewed sketches to 56.83% on IIIT-D Semi-forensic sketches drawn from memory and 54.55% on the forensic PRIP-HDC set. FaceLLM remains considerably more stable (93.51%, 74.82% and 79.55%), with gains of 15.6–25.0 points. We attribute this to the complementary nature of the two approaches: ArcFace's embeddings, trained on photographs, rely on fine appearance detail that forensic sketches do not preserve, whereas FaceLLM can reason over semantic facial attributes that survive in memory-based and forensic drawings. These results suggest that MLLMs are particularly valuable for forensic sketch recognition, while specialized non-MLLM embedding models remain competitive when the sketch closely resembles the photograph.

The conventional sketch–face recognition approach relies on unimodal embedding models, which map a sketch and a photograph into a shared feature space and match them by distance. These models learn fine appearance cues from photographs, such as texture, shading and precise facial geometry, and succeed only when the sketch reproduces those cues. However, forensic sketches rarely do: they are drawn from a witness's memory and preserve an identity only at the level of salient, describable attributes such as gender, age, face shape, hairstyle, eyebrow and nose structure, facial hair and accessories. A *multimodal LLM*, by contrast, perceives these attributes explicitly and reasons over them jointly. It can relate the visual evidence in the sketch to natural-language descriptions of facial traits, and it can compare candidates attribute by attribute rather than by a single similarity score. Therefore, this shows that semantic, language-level descriptions of a face carry identity information comparable to that captured by a photo-trained embedding model. This suggests that the benefit of multimodality lies less in simply concatenating inputs than in the MLLM's capacity for attribute-level semantic reasoning, which unimodal embedding models lack. Together, these findings indicate that MLLMs provide a more interpretable and robust framework for forensic face recognition than conventional sketch–photo embedding approaches.

## E Embedder MLLM-based Recognition

We use the Qwen3 feature embedder (Qwen3-VL-Embedding-8B) to extract embeddings from facial sketches and their corresponding facial descriptions (generated with FaceLLM), treating them as probe representations, while facial image embeddings extracted from the Qwen3 embedder serve as gallery templates for retrieval. We evaluate Rank-1, Rank-5, and Rank-10 retrieval accuracies on the IIIT-D Sketch Forensic, IIIT-D Viewed Sketch, and FS2K datasets. The retrieval is performed against 139, 231 and 1045 facial images in the gallery for IIIT-D SemiForensic, IIIT-D Viewed Sketch, and FS2K datasets, respectively. The results in Figure 10 demonstrate that MLLMs do not perform better with longer gallery/candidate contexts, achieving a very low Rank-1 accuracy.

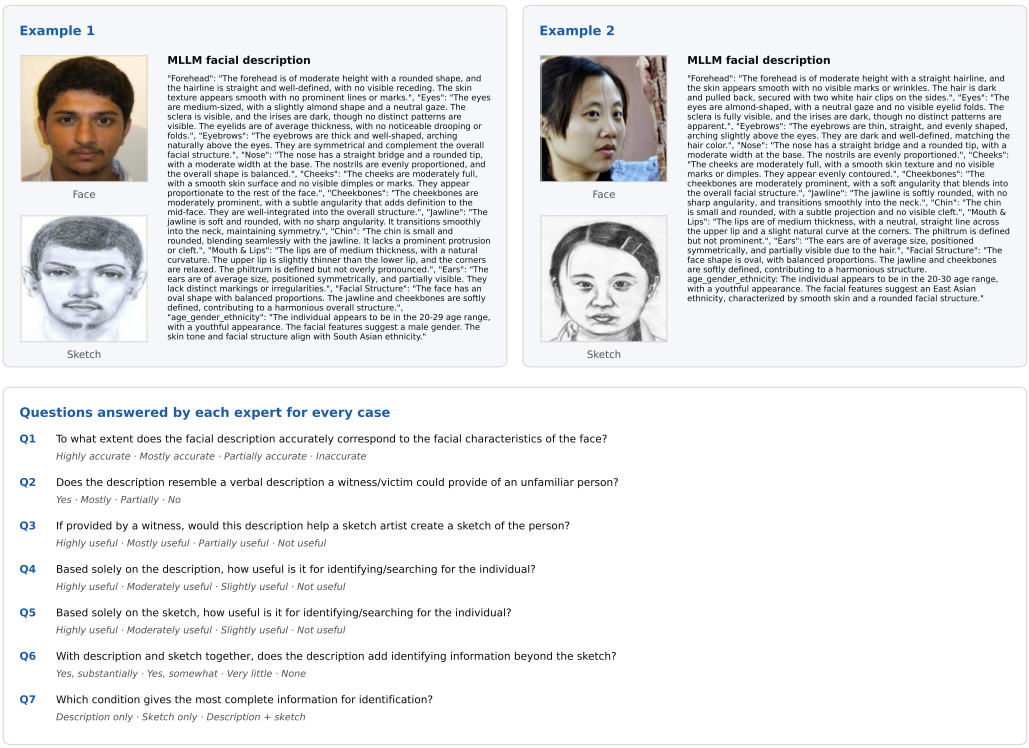
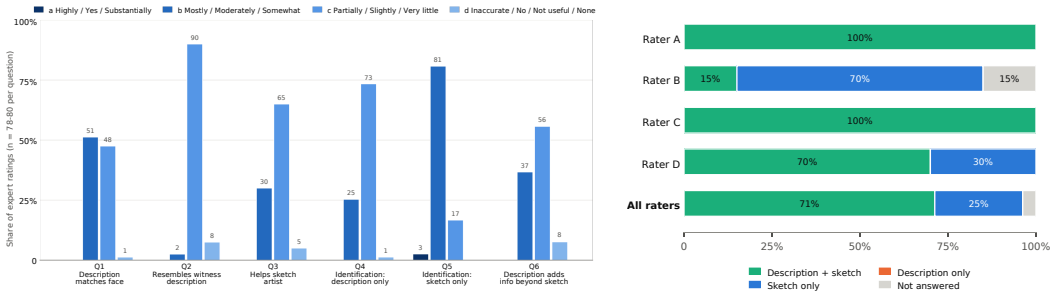



Figure 8: Examples from the forensic expert questionnaire. Each expert was shown the face, its sketch, and the MLLM (FaceLLM)-generated facial description, and they answered seven questions for each subject.




(a) Expert ratings of FaceLLM descriptions and sketches (4 raters × 20 cases)

(b) Most complete information for identification

Figure 9: Forensic expert evaluation of MLLM descriptions and sketches. Four forensic experts rated the FaceLLM-generated facial description and the corresponding sketch for 20 subjects drawn from all five datasets (CUFSF, PRIP-HDC, FS2K, IIIT-D Sketch (Viewed and Semi-Forensic), CelebA). (a) Distribution of expert ratings for questions Q1–Q6: description accuracy (Q1), resemblance to a witness description (Q2), usefulness to a sketch artist (Q3), usefulness for identification from the description alone (Q4) and the sketch alone (Q5), and additional information the description provides beyond the sketch (Q6). Responses range from a (most favourable: Highly / Yes / Substantially) to d (least favourable: Inaccurate / No / Not useful / None). (b) Information condition each rater selected as most complete for identification (Q7). Across raters, description + sketch was chosen most often (71%), indicating that experts view the descriptions as complementary to the sketch.

## F FORENSIC EXPERT EVALUATION

Four forensic experts assessed FaceLLM's facial descriptions and the corresponding sketches for 20 subjects drawn from all five datasets (Figure 9) based on the example questionnaire shown in Figure 8. For each subject, experts were shown the face image, its sketch and the FaceLLM description, and they answered seven questions (see Figure 8). The descriptions were judged at least partially accurate in 99% of ratings, and mostly accurate in 51% (Figure 9a, Q1), confirming that FaceLLM captures the salient facial characteristics of each subject. Experts also found the descriptions helpful for forensic practice: 95% of ratings considered them at least partially useful to a sketch artist reconstructing the face (Figure 9a, Q3), and 99% considered them useful to some degree for identifying or searching for the individual (Figure 9a, Q4). Their structured, feature-by-feature format differs from a typical witness account, and experts rated them as only partially resembling witness-provided descriptions (90%; Figure 9a, Q2). This format nevertheless provides a systematic and consistent record of facial attributes. As expected, the sketch remained the primary visual cue for identification (Figure 9a, Q5). The two sources were, however, viewed as complementary rather than competing: 92% of ratings indicated that the description contributes additional identifying information beyond the sketch (Figure 9a, Q6). Most notably, 71% of ratings selected description + sketch as the most complete information for identification, compared with 25% for the sketch alone (Figure 9c). These findings position FaceLLM's descriptions as a valuable add-on to the sketch: they enrich the visual evidence with explicit, interpretable facial attributes, and supply semantic context that a sketch alone cannot state. One rater gave near-identical answers across cases (Figure 9c, Rater C); excluding this rater did not change any conclusion.

## G ACTUAL FORENSIC BEHAVIOR CASE STUDIES

To assess the real forensic situations, we scraped publicly available videos of forensic sketch artist drawings based on the description provided by observers/witnesses. Descriptions, sketches, and real pictures for such cases are provided in Figure 11. We use FaceLLM for analysis, having the top-3 most similar images and the true one as facial candidate options. Across the three available cases, the combination of sketch and description solved all three cases, while sketch-based MLLM recognition solved only two cases. It highlights that multimodal cues are important for facial analysis tasks by understanding and recognizing faces. ***We will release the link to these videos upon publication.***

Beyond the sketch-description dependence discussed in Appendix H, two forms of leakage merit attention. First, because a description is generated from the same photograph that appears in the candidate set, it could in principle encode that image rather than the identity. Three results argue against this being the dominant effect: descriptions are highly generic across identities (361 attribute combinations among 959 CUFSF identities, mean cosine similarity 0.963, Section 4), so they carry little instance-specific content; in the pose protocol the candidate is an Arc2Face rendering conditioned on the identity embedding rather than on the source pixels, and description-only accuracy drops by only about 7 points on average (Table 4); and the real-world cases above use descriptions produced by eyewitnesses with no access to the candidate photographs, where the multimodal setting still succeeds. Second, the same model could both write and consume a description. Tables 5 to 7 show no consistent self-consistency advantage: Qwen2.5-VL scores higher with FaceLLM descriptions than with its own on CUFSF (80.40 vs. 78.31), InternVL2 scores identically with either on CUFSF (80.19), and in the fused setting FaceLLM performs as well or better with InternVL2 descriptions on CUFSF and PRIP-HDC (93.85 vs. 93.22 and 84.09 vs. 79.55). FaceLLM does obtain higher description-only accuracy with its own descriptions on several datasets, which is consistent with those descriptions being more informative for every evaluator rather than with generator-specific leakage, although the two cannot be fully separated without an independently sourced description. We therefore treat same-photograph descriptions as an upper bound on witness quality and the eyewitness-driven cases as the realistic lower bound; a cross-instance protocol, in which the description is generated from a different photograph of the same identity, is a natural extension we leave for future work.

## H Forensic Evidence Independence and Real-World Validation

A key consideration in multimodal forensic identification is the potential dependence between the sketch and the accompanying facial description. In our ***MS-Forensics*** benchmark, facial descriptions are generated from facial photographs corresponding to the probe identities; therefore, we do not claim that the two modalities constitute statistically independent evidence. Instead, we evaluate settings that differ in how the sketch is acquired and how closely it preserves the appearance of the underlying photograph. Specifically, *IIIT-D Semi-Forensic* sketches are drawn from memory after a single observation, introducing greater abstraction and loss of fine-grained identity cues. Our benchmark further includes *PRIP-HDC*, a forensic hand-drawn composite dataset, which represents a substantially larger sketch-photograph discrepancy. This distinction is reflected in the performance of conventional photo-trained recognition models, which degrades substantially on the Semi-Forensic and PRIP-HDC settings.

We additionally evaluate robustness to *intra-identity pose variation* by constructing candidate photographs under controlled pitch and yaw changes. This setting introduces a visual discrepancy between the probe sketch and the candidate photograph, requiring the model to rely on identity-relevant characteristics that remain informative across viewpoints rather than direct pixel-level or geometric correspondence. Our results show that pose variations reduce recognition performance across modalities, while semantic facial descriptions provide complementary information that is comparatively less sensitive to viewpoint changes.

To further evaluate the proposed multimodal reasoning in an actual forensic-style setting, we evaluated on a real-world case study using publicly available videos of forensic sketch artists drawing suspects based on descriptions provided by observers or eyewitnesses. In contrast to viewed-sketch protocols, the sketches in these cases are produced from eyewitness descriptions rather than by directly observing the reference photograph. We evaluate three such cases by providing FaceLLM with the forensic sketch, its accompanying facial description, and a candidate set containing the genuine identity and three visually similar distractors. The combined sketch and description correctly identified the genuine identity in all three cases, whereas sketch-only recognition succeeded in two cases.

Together, these evaluations complement the curated ***MS-Forensics*** benchmark by examining multimodal identity reasoning under progressively more realistic evidence conditions: direct-view sketches, memory-based sketches, forensic composites, pose-discrepant candidate photographs, and finally eyewitness-driven forensic cases. This provides a more realistic assessment of the utility of combining structural information from sketches with semantic information conveyed through eyewitness descriptions.

## I Implementation Details

**Description Generation.** For each identity, we generate fine-grained textual facial descriptions from the corresponding facial photograph using three multimodal language models: FaceLLM-38B Shahreza & Marcel (2025b), InternVL2-26B Chen et al. (2024), and Qwen2.5-VL-32B Bai et al. (2025). The descriptions capture 11 facial attributes, including forehead, eyes, eyebrows, nose, cheeks, cheekbones, jawline, chin, mouth and lips, ears, and overall facial structure. Also including 3 demographic attributes. Importantly, the descriptions are generated from the facial photographs.

For audio-based evaluation, the generated textual descriptions are converted into speech using the publicly available Kokoro (`https://github.com/hexgrad/kokoro`) text-to-speech model, with diverse male and female voices and multiple accents. In addition, we evaluate 200 human-recorded audio descriptions (100 recordings from females and 100 recordings from males) collected from 14 speakers (7 females and 7 males) under diverse acoustic conditions.

**Recognition Prompts.** The recognition prompts for each evaluation setting, including sketch-only, description-only, audio-only, and multimodal recognition. The complete prompts and their corresponding input configurations are provided in Figures 13, 14, 15, 16. All models are evaluated using the same prompt structure, with only the available evidence modality varied.

**Input Construction.** Each evaluation instance consists of a probe sketch, an optional facial description or audio description, and a candidate set containing the ground-truth identity and visually

similar distractors. We evaluate sketch-only, description-only, audio-only, and multimodal configurations. Facial descriptions are generated from the corresponding facial photographs using 11 fine-grained facial attributes and 3 demographic attributes rather than from the sketches.

**Pose Variation Candidate Images.** For controlled pose-variation experiments, candidate photographs are generated using Arc2Face Papantoniou et al. (2024). We consider a frontal reference together with five non-frontal pitch/yaw configurations, resulting in six pose conditions.

**Models.** We evaluate FaceLLM-38B Shahreza & Marcel (2025b), InternVL2-26B Chen et al. (2024), and Qwen2.5-VL-32B Bai et al. (2025) for visual recognition experiments. Qwen2.5-Omni Xu et al. (2025) is additionally used for audio-based evaluation.

**Decoding and Evaluation.** The model is required to select one identity from the provided candidate options. The predicted option is mapped to its corresponding candidate identity and considered correct when it matches the ground-truth identity. For candidate-set scaling experiments, the number of candidate identities is varied while keeping the probe fixed.

**Reproducibility.** We do not apply task-specific transformations beyond the standard input processing of the respective pretrained models. Exact low-level preprocessing and decoding parameters, including image transformations, audio resampling, sampling temperature, top-$p$, and random seeds, are configuration-dependent and are reported directly from the corresponding model implementations rather than inferred.

## J Facial Description Generation Prompts for MLLMs input

To generate facial descriptions from face images, we employ carefully designed prompts that guide multimodal large language models (MLLMs) toward describing identity-discriminative facial characteristics. The prompt instructs the MLLM to analyze various facial regions, including the forehead, eyes, eyebrows, nose, cheeks, cheekbones, jawline, chin, mouth, ears, and overall facial structure. For each region, the model is guided to describe visible attributes such as shape, size, contour, symmetry, texture, and distinctive characteristics while avoiding unsupported or absent features. The prompt further encourages the descriptions to be expressed as natural, continuous text rather than lists, making them more consistent with descriptions that could be communicated during forensic identification.

We apply the same prompt formulation across different MLLMs, including FaceLLM, Qwen2.5-VL, and InternVL2, using the face image as visual input. Figure 12 presents the complete prompt and representative generated descriptions. The resulting descriptions provide a structured yet natural semantic representation of facial appearance, capturing complementary information at both regional and holistic levels.

## K Identity Recognition Prompts to MLLMs

The sample prompts used as inputs to various state-of-the-art MLLMs, including FaceLLM, InternVL2, and Qwen2.5-VL for the identity recognition forensic task, are shown in Figures 13, 14, 15 and 16. Figure 13 illustrates prompts used for sketch-only evaluation, Figure 14 presents prompts for description-only evaluation, and Figure 15 and 16 shows prompts used for combined sketch and description-based (text and audio) evaluations. Similarly, for computing identity recognition evaluations across audio-based facial descriptions, instead of textual descriptions, audio-based descriptions are given to the Qwen2.5-Omni model.

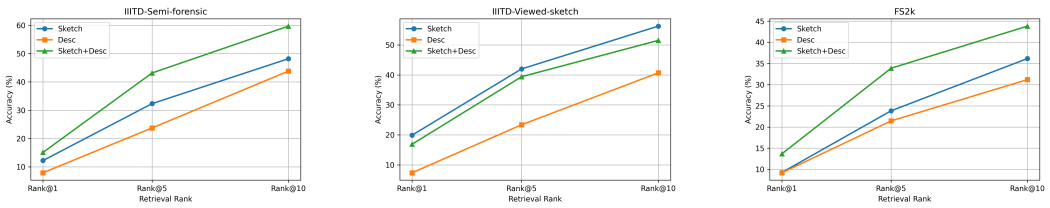



Figure 10: Retrieval performance reporting rank-wise accuracy using Qwen3-VL-Embedding-8B model.

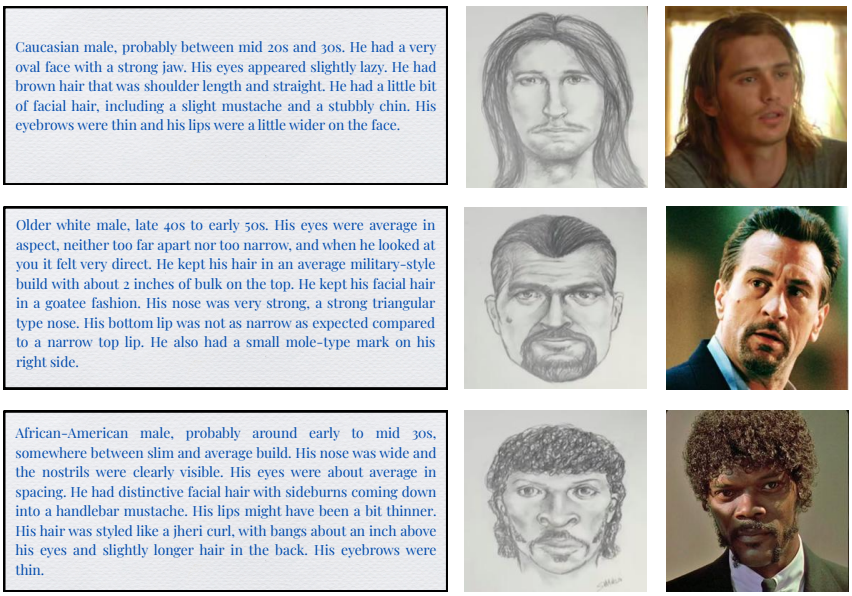



Figure 11: Actual Forensic Behavior Case Studies.

***MLLM generated Facial Descriptions***

**Input Prompt:**

"""You are a forensic AI assistant specialized in generating fine-grained and identity-accurate descriptions of human faces for forensic sketching and recognition. You will be provided with a face image. Your task is to produce highly detailed descriptions for each major facial region, ensuring that the traits are discriminative enough to identify the person even if props or disguises are present.
Focus strictly on visible features, and do not describe traits that are absent. Each region's description should be no longer than four sentences, written as continuous natural text. Avoid lists or formatting inside the descriptions.

The output must provide a small paragraph combining below facial region discirminative descriptions:
{
"Forehead": "<detailed description>",
"Eyes": "<detailed description>",
"Eyebrows": "<detailed description>",
"Nose": "<detailed description>",
"Cheeks": "<detailed description>",
"Cheekbones": "<detailed description>",
"Jawline": "<detailed description>",
"Chin": "<detailed description>",
"Mouth & Lips": "<detailed description>",
"Ears": "<detailed description>"
"Facial Structure": "<detailed description>"
"age_gender_ethnicity": "<detailed description>"
}

Guidelines for each region:
- Forehead: focus on height, slope, hairline shape, skin texture, visible lines, or marks.
- Eyes: cover shape, tilt, eyelid folds, pupil visibility, sclera exposure, iris color, and unique identifiers.
- Eyebrows: emphasize thickness, arch, length, density, and asymmetries.
- Nose: describe bridge, nostrils, tip, base width, and angle.
- Cheeks: detail fullness, contour, skin surface, dimples, or marks.
- Cheekbones: describe prominence, width, angularity, and definition.
- Jawline: outline width, contour sharpness, symmetry, and transitions to neck.
- Chin: describe size, protrusion, cleft, contour, and asymmetry.
- Mouth and Lips: cover lip thickness, shape, curvature, corners, philtrum definition, or texture.
- Ears: note shape, size, angle, lobe type, and distinguishing marks.
- Facial Structure: describe the overall face shape (oval, round, square, heart-shaped, etc.), width-to-height ratio, and angularity or softness of the jawline, cheekbones, and forehead. Include structural traits such as broad forehead, narrow jawline, or pronounced symmetry that make this face unique.
- Age, Gender, Ethnicity: infer the person's apparent age range, perceived gender, and likely ethnic background based on facial morphology and skin tone.
"""

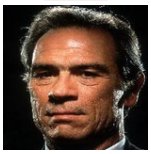
Input Image



**Generated Output:**

""" "Forehead": "The forehead is of moderate height with a smooth texture and a straight hairline. There are no visible marks or wrinkles." "Eyes": "The eyes are medium-sized, with a slightly almond shape and a neutral gaze. The sclera is visible, and the irises are dark. There are no noticeable eyelid folds or unique identifiers.", "Eyebrows": "The eyebrows are thick and well-defined, with a natural arch and symmetrical shape. They frame the eyes effectively.", "Nose": "The nose has a straight bridge and a rounded tip. The nostrils are of moderate width, and the nose aligns symmetrically with the face.", "Cheeks": "The cheeks are moderately full with a smooth skin surface, showing no prominent dimples or marks.", "Cheekbones": "The cheekbones are moderately prominent, with a soft angularity that blends into the jawline.", "Jawline": "The jawline is softly contoured, with a slight tapering towards the chin, and appears symmetrical.", "Chin": "The chin is of medium size, rounded, and proportionate to the rest of the face.", "Mouth & Lips": "The lips are of medium thickness, with a natural curvature. The upper lip is slightly thinner than the lower lip, and the corners are relaxed.", "Ears": "The ears are of average size, positioned symmetrically, and partially visible.", "Facial Structure": "The face has an oval shape with balanced proportions. The jawline is soft, and the cheekbones are subtly defined, contributing to a harmonious structure." "age_gender_ethnicity": "The individual appears to be in the age range of 18–25 years, with a youthful appearance. The facial features suggest a male gender, and the skin tone and structure align with South Asian ethnicity." """

Input Image



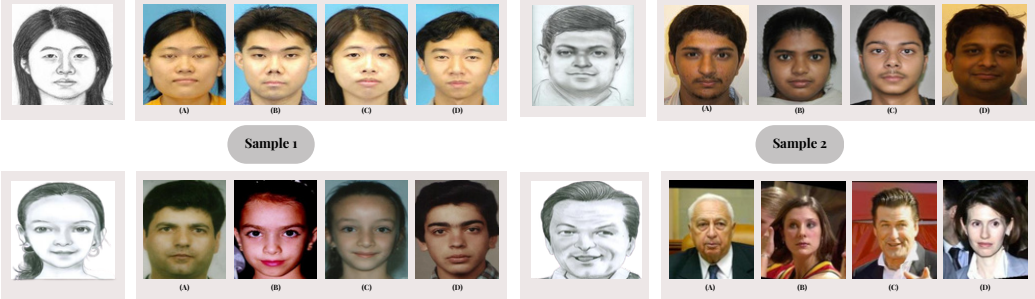
**Generated Output:**

""" "Forehead": "The forehead is broad and slightly rounded, with a smooth texture and no prominent lines or marks. The hairline is short and receding, blending seamlessly into the scalp." "Eyes": "The eyes are almond-shaped, with a slightly upward tilt. The sclera is visible, and the irises are dark, though the exact color is not distinctly clear. The eyelids are moderately thick, with no noticeable drooping or sagging." "Eyebrows": "The eyebrows are thick and well-defined, arching naturally over the eyes. They are symmetrical and evenly spaced, with a full and dense appearance." "Nose": "The nose has a broad and slightly rounded tip, with a wide base. The nostrils are moderately flared, and the bridge is straight and proportionate to the face." "Cheeks": "The cheeks are moderately full, with a smooth skin surface. There are no visible dimples or marks, and the contour is soft and rounded." "Cheekbones": "The cheekbones are moderately prominent, with a subtle angularity that adds definition to the mid-face. They are well-integrated into the overall structure." "Jawline": "The jawline is rounded and smooth, with no sharp angularity. It transitions gently into the neck, maintaining a balanced proportion." "Chin": "The chin is rounded and proportionate to the jawline, with a subtle prominence. There is no visible cleft or sharp definition." "Mouth & Lips": "The lips are full and moderately thick, with a natural curvature. The upper lip is slightly thinner than the lower lip, and the corners are relaxed. The philtrum is defined but not overly pronounced." "Ears": "The ears are of average size, positioned symmetrically on the head. They are slightly rounded at the edges, with no visible piercings or marks." "Facial Structure": "The face has an oval shape with balanced proportions. The jawline and cheekbones are soft, contributing to a harmonious and rounded appearance." "age_gender_ethnicity": "The individual appears to be in the 20–30 age range, with a youthful and smooth skin texture. The facial features suggest a Black ethnicity, with a strong and defined bone structure. The overall appearance is consistent with a male." """

Figure 12: Sample prompt used for facial description generation using MLLMs which take prompt and image as input to generate facial descriptions whose samples are also present in this image. Facial descriptions generated for all samples of different datasets using various MLLMs (FaceLLM, Qwen2.5-VL and InternVL2) will be released upon paper acceptance.

**Sketch Prompt**

**Question:** "Face-Sketch recognition involves identifying a person based on their facial features in a sketch to identify the same person from various colored facial photos. You have to perform the task of sketch to face recognition. \n
The first image is a sketch of a person. The same person is present in which of these facial images?\n
Please answer the question and provide only the correct option letter, e.g., A, B, C, D.\n
A Image 2\n
B Image 3\n
C Image 4\n
D Image 5\n "




Sample 1 | Sample 2 | Sample 3 | Sample 4

Figure 13: Prompt used for **sketch-only** evaluation, where the MLLM identifies an individual from the candidate set using only the facial sketch.

**Description Prompt**

**Question:** "Face recognition involves identifying a person based on their facial feature description from various colored facial photos. You have to perform the task of facial feature description to face recognition.\n
As described in the following text, identify which of the images best matches the person:\n {description}\n
Please answer the question and provide only the correct option letter, e.g., A, B, C, D.\n
A Image 1\n
B Image 2\n
C Image 3\n
D Image 4\n "



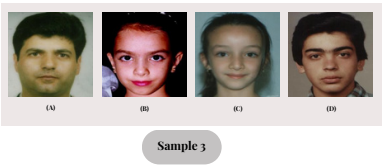

Sample 1

"Forehead: The forehead is of average height, with a smooth texture and a straight hairline. There are no visible marks or lines on the forehead. Eyes: The eyes are almond-shaped, with a neutral gaze and no visible eyelid folds. The sclera is fully visible, and the irises are a light brown color. Eyebrows: The eyebrows are straight and of medium thickness, arching naturally above the eyes. They are evenly shaped and symmetrical. Nose: The nose has a straight bridge and a rounded tip, with a moderate width at the base. The nostrils are evenly proportioned. Cheeks: The cheeks are moderately full with a smooth skin surface, showing no prominent dimples or marks. Cheekbones: The cheekbones are moderately prominent, with a soft angularity that blends into the overall facial structure. Jawline: The jawline is softly rounded, with a gentle tapering towards the chin, and symmetrical contours. Chin: The chin is small and rounded, with a subtle projection and no visible cleft. Mouth & Lips: The lips are of medium thickness, with a natural curve. The upper lip is slightly thinner than the lower lip, and the corners are neutral. Ears: The ears are of average size, with a natural shape and positioned symmetrically. There are no visible distinguishing marks. Facial Structure: The face has an oval shape with balanced proportions, featuring soft angularity in the jawline and cheekbones. The overall structure is symmetrical. Age, Gender & Ethnicity: The individual appears to be in the 20–29 age range, with a youthful appearance. The facial features suggest an East Asian ethnicity, and the face has a soft and smooth texture typical of this demographic."




Sample 2

"Forehead: The forehead appears moderately broad with a slightly rounded shape, and the hairline is straight and receding. The skin texture is smooth with no visible marks or lines. Eyes: The eyes are medium-sized, with a slightly almond shape and a neutral gaze. The sclera is visible, and the irises are dark, though no distinct patterns are visible. The eyelids are of average thickness, and there are no prominent eyelid folds. Eyebrows: The eyebrows are moderately thick, with a natural arch and even density. They are slightly darker than the surrounding skin and frame the eyes symmetrically. Nose: The nose has a straight bridge and a rounded tip. The nostrils are of average width, and the overall structure is proportionate to the face. Cheeks: The cheeks are moderately full with a smooth skin surface, and there are no visible dimples or marks. They appear evenly contoured. Cheekbones: The cheekbones are moderately prominent, with a subtle angularity that adds definition to the mid-face. Jawline: The jawline is rounded and smooth, transitioning gently into the neck without sharp angularity. Chin: The chin is rounded and proportionate to the jawline, with no significant protrusion or cleft. Mouth & Lips: The lips are of medium thickness, with the upper lip slightly thinner than the lower. The corners of the mouth are relaxed, and the lips have a natural texture. Ears: The ears are of average size, with a rounded shape and positioned symmetrically. There are no visible distinguishing marks. Facial Structure: The face has an oval shape with balanced proportions. The jawline and cheekbones are softly defined, contributing to a harmonious structure. Age, Gender & Ethnicity: The individual appears to be in the 20–30 age range, with a youthful appearance. The facial features suggest a male gender, and the skin tone and structure align with South Asian ethnicity."




Sample 3

"Forehead: The forehead is smooth and slightly rounded, with a natural hairline that recedes slightly at the temples. The skin appears even in texture with no visible marks or lines. Eyes: The eyes are almond-shaped, with a soft gaze and a natural symmetry. The eyelids are moderately thin, and the irises are a light brown color. There are no visible eyelid folds or prominent sclera exposure. Eyebrows: The eyebrows are thin and naturally arched, framing the eyes symmetrically. They are well-groomed and blend seamlessly with the overall facial structure. Nose: The nose has a straight bridge and a rounded tip, with a balanced width at the base. It aligns well with the facial proportions. Cheeks: The cheeks are moderately full, with a smooth skin surface and no visible dimples or marks. They contribute to a youthful and soft facial contour. Cheekbones: The cheekbones are subtly defined, with a gentle prominence that complements the overall oval shape of the face. Jawline: The jawline is soft and rounded, transitioning smoothly into the neck. It lacks sharp angularity, maintaining a harmonious balance with the rest of the face. Chin: The chin is small and rounded, proportionate to the rest of the face, and does not protrude. Mouth & Lips: The lips are of medium thickness, with a natural curve and a slight upward tilt at the corners. The upper lip is slightly thinner than the lower lip, and the philtrum is well-defined. Ears: The ears are of average size, positioned symmetrically, and partially visible. They have a natural shape with no prominent markings. Facial Structure: The face has an oval shape with soft angularity, particularly in the jawline and cheekbones. The overall structure is balanced and youthful, with no pronounced sharpness. Age, Gender & Ethnicity: The individual appears to be a young girl, likely between 5–10 years old, with a soft and smooth skin texture. The facial features suggest a White ethnicity, with no strong indicators of other ethnic backgrounds."

Figure 14: Prompt used for **description-only** evaluation, where identity recognition is performed using textual biometric facial descriptions without sketch input.

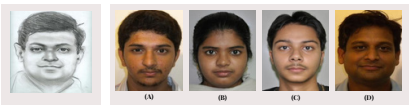
***Description + Sketch Prompt***

**Question:** "Face-Sketch recognition involves identifying a person based on their facial features in a sketch to identify the same person from various colored facial photos.\n
The first image is a sketch of a person. The same person is present in which of these facial images?\n
Here is a short description of the person to help you identify:\n {description}\n
Please answer the question and provide only the correct option letter, e.g., A, B, C, D.\n
A Image 2\n
B Image 3\n
C Image 4\n
D Image 5\n "




Sample 1

"Forehead: The forehead is of average height, with a smooth texture and a straight hairline. There are no visible marks or lines on the forehead. Eyes: The eyes are almond-shaped, with a neutral gaze and no visible eyelid folds. The sclera is fully visible, and the irises are a light brown color. Eyebrows: The eyebrows are straight and of medium thickness, arching naturally above the eyes. They are evenly shaped and symmetrical. Nose: The nose has a straight bridge and a rounded tip, with a moderate width at the base. The nostrils are evenly proportioned. Cheeks: The cheeks are moderately full with a smooth skin surface, showing no prominent dimples or marks. Cheekbones: The cheekbones are moderately prominent, with a soft angularity that blends into the overall facial structure. Jawline: The jawline is softly rounded, with a gentle tapering towards the chin, and symmetrical contours. Chin: The chin is small and rounded, with a subtle projection and no visible cleft. Mouth & Lips: The lips are of medium thickness, with a natural curve. The upper lip is slightly thinner than the lower lip, and the corners are neutral. Ears: The ears are of average size, with a natural shape and positioned symmetrically. There are no visible distinguishing marks. Facial Structure: The face has an oval shape with balanced proportions, featuring soft angularity in the jawline and cheekbones. The overall structure is symmetrical. Age, Gender & Ethnicity: The individual appears to be in the 20-29 age range, with a youthful appearance. The facial features suggest an East Asian ethnicity, and the face has a soft and smooth texture typical of this demographic."




Sample 2

"Forehead: The forehead appears moderately broad with a slightly rounded shape, and the hairline is straight and receding. The skin texture is smooth with no visible marks or lines. Eyes: The eyes are medium-sized, with a slightly almond shape and a neutral gaze. The sclera is visible, and the irises are dark, though no distinct patterns are visible. The eyelids are of average thickness, and there are no prominent eyelid folds. Eyebrows: The eyebrows are moderately thick, with a natural arch and even density. They are slightly darker than the surrounding skin and frame the eyes symmetrically. Nose: The nose has a straight bridge and a rounded tip. The nostrils are of average width, and the overall structure is proportionate to the face. Cheeks: The cheeks are moderately full with a smooth skin surface, and there are no visible dimples or marks. They appear evenly contoured. Cheekbones: The cheekbones are moderately prominent, with a subtle angularity that adds definition to the mid-face. Jawline: The jawline is rounded and smooth, transitioning gently into the neck without sharp angularity. Chin: The chin is rounded and proportionate to the jawline, with no significant protrusion or cleft. Mouth & Lips: The lips are of medium thickness, with the upper lip slightly thinner than the lower. The corners of the mouth are relaxed, and the lips have a natural texture. Ears: The ears are of average size, with a rounded shape and positioned symmetrically. There are no visible distinguishing marks. Facial Structure: The face has an oval shape with balanced proportions. The jawline and cheekbones are softly defined, contributing to a harmonious structure. Age, Gender & Ethnicity: The individual appears to be in the 20-30 age range, with a youthful appearance. The facial features suggest a male gender, and the skin tone and structure align with South Asian ethnicity."




Sample 3

"Forehead: The forehead is smooth and slightly rounded, with a natural hairline that recedes slightly at the temples. The skin appears even in texture with no visible marks or lines. Eyes: The eyes are almond-shaped, with a soft gaze and a natural symmetry. The eyelids are moderately thin, and the irises are a light brown color. There are no visible eyelid folds or prominent sclera exposure. Eyebrows: The eyebrows are thin and naturally arched, framing the eyes symmetrically. They are well-groomed and blend seamlessly with the overall facial structure. Nose: The nose has a straight bridge and a rounded tip, with a balanced width at the base. It aligns well with the facial proportions. Cheeks: The cheeks are moderately full, with a smooth skin surface and no visible dimples or marks. They contribute to a youthful and soft facial contour. Cheekbones: The cheekbones are subtly defined, with a gentle prominence that complements the overall oval shape of the face. Jawline: The jawline is soft and rounded, transitioning smoothly into the neck. It lacks sharp angularity, maintaining a harmonious balance with the rest of the face. Chin: The chin is small and rounded, proportionate to the rest of the face, and does not protrude. Mouth & Lips: The lips are of medium thickness, with a natural curve and a slight upward tilt at the corners. The upper lip is slightly thinner than the lower lip, and the philtrum is well-defined. Ears: The ears are of average size, positioned symmetrically, and partially visible. They have a natural shape with no prominent markings. Facial Structure: The face has an oval shape with soft angularity, particularly in the jawline and cheekbones. The overall structure is balanced and youthful, with no pronounced sharpness. Age, Gender & Ethnicity: The individual appears to be a young girl, likely between 5-10 years old, with a soft and smooth skin texture. The facial features suggest a White ethnicity, with no strong indicators of other ethnic backgrounds."

Figure 15: Prompt used for **combined sketch and description**-based evaluation, where both the facial sketch and textual descriptions are provided to the MLLM for identity recognition.

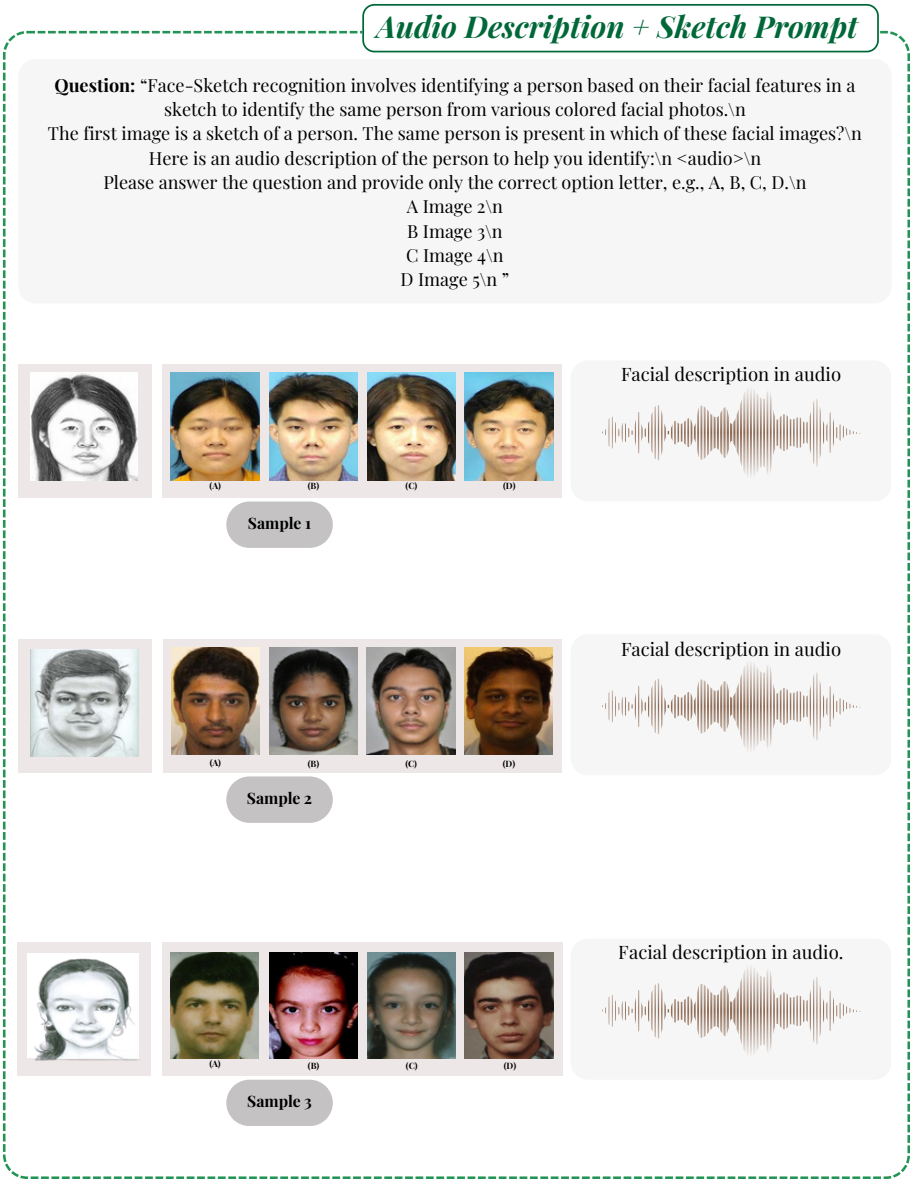



Figure 16: Prompt used for **combined sketch and audio description**-based evaluation, where both the facial sketch and audio descriptions are provided to the MLLM for identity recognition. ***The audios will be released upon paper publication.***