2026 International Conference on Automatic Face and Gesture Recognition (FG)

Employing Vision-Language Models for Face Image Quality Assessment

Erdi Sarıtaş[1], Eren Onaran[1], Vitomir Štruc[2], and Hazım Kemal Ekenel[1,3]
[1] Department of Computer Engineering, Istanbul Technical University, Istanbul, Türkiye
[2] Faculty of Electrical Engineering, University of Ljubljana, Ljubljana, Slovenia
[3] Division of Engineering, NYU Abu Dhabi, Abu Dhabi, UAE

***Abstract*— Face Image Quality Assessment (FIQA) is a crucial control step in biometric pipelines. It ensures only reliable samples are processed to maintain system accuracy. State-of-the-art FIQA methods achieve high utility but typically operate as "black boxes." They produce scalar scores without human-interpretable justifications. This lack of transparency limits their effectiveness in human-in-the-loop scenarios, such as automated border control, where actionable feedback is essential. In this paper, we investigate the potential of off-the-shelf Vision-Language Models (VLMs) to bridge this gap by performing FIQA in a zero-shot setting. We present a comprehensive evaluation framework for assessing VLM performance. This involves benchmarking traditional FIQA methods through error-versus-reject curves. Additionally, using a diverse set of datasets, ranging from surveillance-oriented to synthetically generated, we analyzed their interpretability, consistency, and robustness to prompt changes. Our results show biometric utility performance depends significantly on architecture, not merely on parameter count. Most VLMs' outputs align with those of traditional methods. We also find that VLM ranking performance and the generated scores may vary across prompts. Our synthetic ablation study shows that while increasing the parameter count can improve internal consistency, it yields worse degradation-detection performance than smaller models. These findings suggest that zero-shot FIQA score estimation using VLMs is promising and could effectively complement conventional FIQA pipelines as an interpretability module. The codes are available at github.com/ThEnded32/VLM4FIQA.git.**

I. Introduction

Face Image Quality Assessment (FIQA) serves as a critical control step in biometric systems, ensuring that only reliable samples are processed. Low-quality images, whether caused by blur, poor lighting, or other factors, can significantly increase false-rejection rates, which poses a risk to system consistency. Consequently, FIQA has become a mandatory pre-processing step in modern face recognition pipelines [30]. State-of-the-art (SOTA) FIQA methods have primarily benefited from advances in deep learning, relying on latent quality signals directly from face recognition models. Approaches based on uncertainty estimation [32], margin-based learning [17], and diffusion-based reconstruction [3] have achieved impressive correlation with biometric utility. However, they operate under a "black-box" paradigm: they produce a single scalar score without providing interpretable explanations. This lack of transparency leads to the absence of actionable feedback, e.g., "please come closer", in human-oriented scenarios, like automated border control. Fig. 1 illustrates this limitation, contrasting the opaque scalar output of traditional methods with the transparent, descriptive feedback offered by the VLM-driven paradigm. Providing this human-understandable feedback can enhance both interaction and capture quality.

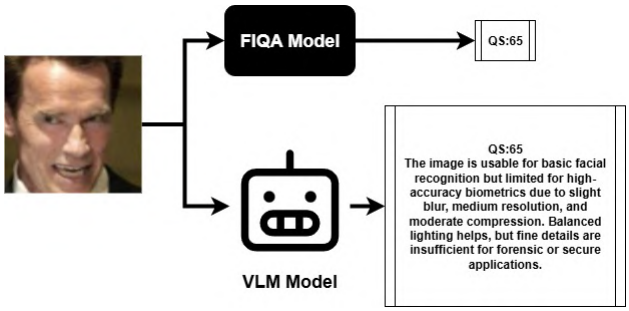


Fig. 1: **VLMs for Quality Assessment.** While traditional FIQA methods (top) function as opaque "black boxes" outputting only scalar scores, VLM-driven approaches (bottom) offer transparency by providing both biometric utility scores and actionable semantic justifications.
*VLM prompt is generated using QWEN2.5-32B.

Concurrently, the literature is witnessing a paradigm shift driven by Vision-Language Models (VLMs). In the domain of general Image Quality Assessment (IQA), research has moved beyond simple regression, utilizing VLMs to generate descriptive quality evaluations and reason about aesthetic or technical attributes [37], [41]. Similarly, in the domain of biometrics, VLMs are increasingly employed to provide semantic explanations for tasks such as attribute recognition and face anti-spoofing [6], [20]. This convergence suggests that multimodal models have the intrinsic capacity to bridge the interpretability gap in FIQA. However, it remains an open question whether the general visual reasoning of these models aligns with the utility-centric needs of biometric quality assessment.

This work investigates the fundamental research question: *Can the zero-shot capacity of general-purpose VLMs align with the requirements of biometric quality assessment?* To address this, we structure our study around four inquiries:

- **Biometric Utility:** Can off-the-shelf VLMs achieve error-versus-reject performance comparable to specialized, supervised FIQA methods?
- **Interpretability:** Do the generated semantic descriptions accurately reflect physical degradations in real-world surveillance scenarios?

979-8-3315-7231-0/26/$31.00 ©2026 European Union

- **Consistency vs. Precision:** How does model scale influence the trade-off between hallucination rates on clean data and the internal consistency of quality scores?
- **Prompt Robustness:** Is the zero-shot ranking logic stable across different semantic prompt phrasings, or does it require specific triggers to function correctly?

To answer these questions, we introduce a comprehensive evaluation framework that prioritizes assessing biometric utility using error-versus-reject (EvR) curves. In addition to this primary analysis, we investigate the models' sensitivity to real-world degradation in surveillance scenarios. We also examine the cross-prompt stability of their scoring logic and their interpretability across diverse datasets. Furthermore, we conduct a controlled synthetic ablation study to rigorously assess the precision of their degradation-detection capabilities. Lastly, we present example VLM outputs to illustrate their image-specific responses.

Our study shows that architectural choices determine zero-shot capability, rather than solely parameter count. Some general-purpose models have been found to be competitive with traditional FIQA estimators in terms of biometric scoring performance. Crucially, analysis of our synthetic dataset ablation reveals a potential complex trade-off between utility and descriptive precision: while larger models maintain robust rankings and strict internal consistency, they exhibit significantly higher hallucination rates on clean images than their smaller counterparts. Furthermore, we find that vulnerability to prompt phrasing is relatively scale-dependent: smaller models are more influenced by specific semantic keywords, whereas larger architectures exhibit a more stable ranking logic that is invariant to phrasing.

II. Related Work

A. Learning-based FIQA Methods

Learning-based FIQA approaches primarily rely on features from face recognition (FR) embeddings. SER-FIQ [32] measures quality based on embedding stability under stochastic dropout, and AdaFace [17] uses the feature vector's magnitude as a quality proxy during training. Further methods focus on intrinsic FR properties; SDD-FIQA [25] is an unsupervised approach that uses the similarity distribution distance between intra- and inter-class samples, and FaceQAN [4] links quality to adversarial robustness by considering the noise magnitude. More recently, advanced architectures have emerged. Generative approaches, such as eDifFIQA [3], correlate quality with reconstructive performance using diffusion models. Vision Transformer (ViT)-based methods are prominent, including ViT-FIQA [2], which uses a learnable "quality token", and DSL-FIQA [8], which employs landmark-guided transformers with dual-set degradation learning. Finally, MR-FIQA [27] addresses data scarcity by leveraging synthetic data and multi-reference representations. Despite high utility in biometrics tasks such as FR, these methods only output scalar scores, lacking human-interpretable justifications.

B. VLM-based Image Quality Assessment

Vision-Language Models (VLMs) have fundamentally transformed general image quality assessment by enabling human-interpretable reasoning. Early works employed CLIP [28] embeddings for zero-shot assessment via prompting comparative judgments (e.g., "sharp" versus "noisy") [34], [43], [22]. The field progressed to generative scoring and reasoning. Generative methods such as Q-Align [37] train LMMs to predict discrete quality tokens (e.g., "Good"). Additionally, a paradigm refined by the team "Next" in the VQualA 2025 Challenge [21] adapts the CLIP model to predict the image quality distribution via a multi-level quality-aware prompt learning mechanism. Other models focus on reasoning using VLMs, such as Co-Instruct [38] for multi-image comparison and EDQA [41] for generating descriptive quality paragraphs. QA-VLM [44] applied VLMs to domain-specific quality (additive manufacturing) for interpretable reasoning. However, these generalist models are not explicitly optimized for biometric face utility.

C. VLM-based Face Analysis and Explainability

VLMs have also emerged for face analysis tasks requiring deep semantic understanding. These models generate human-readable explanations for diverse human-centric tasks. For identity and attribute analysis, FaceLLM [24] and Face-LLaVA [6] generate explanations for identity verification and facial attributes. Broader benchmarks, such as HERM [19], assess VLMs across diverse human understanding tasks. Furthermore, VLMs have been adapted to the security and forgery-detection domain. MGFFD-VLM [7] is a deepfake detector using multi-granularity prompts to reason about quality-related artifacts and generate rationales. Other security works include FaceShield [33], which provides reasoning for anti-spoofing decisions, and InstructFLIP [20], which uses unified instruction tuning to generalize across presentation attacks. Collectively, these works demonstrate that VLMs can be effectively utilized for facial semantics understanding and security-critical control.

D. Explainable and VLM-driven Face Quality Assessment

Explainable Face Image Quality Assessment (X-FIQA) aims to enhance the interpretability of FIQA models by providing human-understandable justifications through visual and semantic outputs. Visuals such as heat maps can be generated using intrinsic properties of methods, e.g., landmark-guided attention of DSL-FIQA [8] to produce an attention heat map. Furthermore, IFQA [14] converts the real/fake paradigm of GANs to high/low quality, and trains a discriminator that produces pixel-level scores. With this, they can directly generate a quality heatmap by default. On the semantic interpretability side, Face Quality Vector (FQV) [23] introduces a multidimensional quality representation that captures expression, pose, and illumination, beyond a single scalar quality value.

Most recently, researchers have begun bridging the gap between FIQA and VLMs. CLIB-FIQA [26] introduced confidence calibration using CLIP to anchor objective quality

factors like blur and occlusion. Moving to generative assessment, FVQ-Rater [39] utilizes instruction tuning to score face video quality, and MDTFIQA [11] evaluates the fidelity of text-to-face generation. Finally, FaceOracle [15] focuses on interactive transparency, using Retrieval-Augmented Generation (RAG), and can be used to explain quality scores via a chat interface. Our research advances this emerging field by providing a comprehensive quantitative analysis of VLM usage in a zero-shot setting.

III. Proposed Method

We formulate a general evaluation framework to assess the performance of Vision-Language Models (VLMs) on the FIQA task in a zero-shot setting. This section introduces the datasets, the prompt types used to query the models, and the evaluation pipeline, which prioritizes biometric utility analysis.

A. Datasets

We use diverse real-world face datasets to verify performance across different capture conditions:

- **CelebA-HQ** [16]: Provides high-quality, studio-like aligned faces, serving as a reference distribution for "good" quality.
- **LFW** [13] and **IJB-B** [36]: Contain in-the-wild face images with natural variations in pose, illumination, and occlusion. These are standard benchmarks for assessing biometric utility under realistic variability.
- **SCFace** [12]: A surveillance dataset capturing subjects at three standoff distances (4.2m, 2.6m, and 1.0m). We use this to analyze sensitivity to physical degradation (resolution and blur) in a controlled security scenario.

In all cases, faces are detected and aligned using MTCNN [42] and resized to a fixed resolution of 224×224.

B. Prompt Design

We query the target VLMs using two primary strategies: scalar scoring for utility analysis and attribute classification for explainability.

Simple Quality Prompt: The primary prompt asks the model to rate a face image on a scale from 0 to 100. The model acts as an *"expert image quality assessor for face images"* and is prompted to *"evaluate the image quality for facial analysis"*. The model generates a strict JSON object containing a single numeric field (i.e., {*"Quality Score": <0-100>*}), and no other text or keys.

Attribute Classification Prompt: For assessing explainability, we use a structured prompt that requests categorical judgments on specific quality factors, with degradation level options. The model returns a JSON object containing a scalar quality score alongside attributes defined by strict options:

- *"Sharpness"*: Clear, Slightly-, Moderately-, or Strongly Blurred.
- *"Resolution"*: High, Medium, Low, or Very Low.
- *"Lighting"*: Balanced, plus three intensities for Dark and Bright.
- *"Compression"*: None, Minimal, Moderate, or Severe.

This allows us to verify internal consistency by checking if the predicted score aligns with the detected artifact severity.

Semantic Variants: We test two additional semantic phrasings against the simple baseline to investigate the behavior of different scalar scoring prompts regarding biometric utility. For these variants, we retain the expert role definition but update the specific instruction and the required JSON output key:

- *Utility*: Prompted to *"evaluate the image utility for a face recognition model"* using the key *"Utility Score"*.
- *Reliability*: Prompted to *"evaluate the image reliability for face verification"* using the key *"Reliability Score"*.

C. Evaluation Pipeline and Metrics

Our evaluation has three stages, shifting the focus from simple correlation to decision-based biometric performance.

a) Stage 1: Biometric Utility (Error-versus-Reject): The core of our benchmark is the error-versus-reject (EvR) analysis. A reliable FIQA estimator should assign lower scores to images that cause recognition errors.

- **Protocol**: We compute the False Non-Match Rate (FNMR) at fixed False Match Rate (FMR) thresholds (i.e., 10^{-3}). We progressively reject a fraction of the dataset with the lowest predicted quality scores.
- **Metrics**: We report the **Area Under the Curve (AUC)** of the EvR plot. A lower AUC indicates that the quality scores successfully prioritize reliable images. We also report **partial AUC** at specific low rejection ratios (1%, 5%, 10%, and 20%) to measure effectiveness in strictly operational ranges where high data retention is required:

$$pAUC_\rho = \frac{1}{\rho} \int_0^\rho \mathrm{FNMR}(r)\, dr \tag{1}$$

b) Stage 2: Surveillance Sensitivity: Using SCFace, we analyze whether the VLM assigns higher quality scores to images captured at closer distances (1.0m) than to those captured at farther distances (4.2m). We further analyze the Attribute Classification outputs to determine whether the model explicitly assigns lower scores to relevant factors such as "Low Resolution."

c) Stage 3: Consistency: We assess the stability of the model's scoring logic by measuring:

- **Cross-Prompt Consistency**: The Mean Absolute Error (MAE), Pearson correlation, and bias between scores generated by the Quality (both Simple and Classification), Utility, and Reliability prompts.
- **Internal Consistency**: The alignment between scalar scores and categorical labels, e.g., ensuring "Blurry" images receive lower scores than "Clear" ones.

D. Synthetic Ablation Study: Mix Degradation

To rigorously assess the interpretability and sensitivity of VLMs, we design a controlled synthetic ablation study. While real-world datasets provide biometric realism, they lack granular ground truth for specific degradations (e.g., exact noise levels or blur kernels). To address this, we generate a Mix Degradation Benchmark using high-quality samples from CelebA-HQ [16] as the clean reference baseline.

a) Degradation Logic: We define a degradation space $\mathcal{D}$ consisting of five artifact types: blur, noise, low resolution, JPEG compression, and under-/over-exposure. For each clean reference image, inspired by the impact of combined degradations [29], we generate a Mixed Degradation variant by applying a combination of three random artifacts: one at high intensity (the "hard artifact") and the other two at mild intensity.

b) Ground Truth Vectors: Each image is associated with a binary ground truth vector $v_{GT} \in \{0,1\}^5$, where the i-th bit indicates the presence of the i-th artifact type; where the hard artifact is noted externally. For the clean reference images, this vector is strictly zero-valued ($[0,0,0,0,0]$).

c) Synthetic Detection Prompt: For the synthetic ablation study, we employ a targeted prompt to verify the presence of specific artifacts. We instruct the model to act as an "expert biometric image quality analyst" and analyze the image for a predefined list of issues: *"blur, noise, pixelation (low_res), jpeg artifacts (compression), and bad lighting (under/overexposed)."* To ensure the outputs are machine-readable, the prompt explicitly commands the model to *"Return STRICT JSON"* containing a scalar *quality_score* (0–100) and boolean detection flags (e.g., *has_blur*, *has_noise*) for each degradation type.

d) Evaluation Metrics: We evaluate the VLM's predicted degradation tags against these ground truth vectors using three specialized metrics:

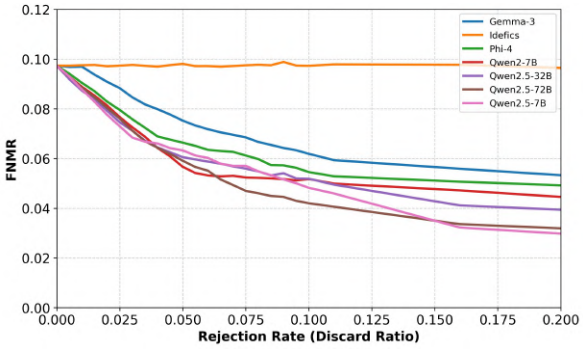
1) **Hallucination Rate**: Estimates the model's tendency to predict degradations on the clean reference images (False Positive Rate).
2) **Degradation Recall**: Measures the model's ability to correctly identify the designated "Hard Artifact" within the mixed variant, identifying the primary cause of quality loss.
3) **Hamming Precision**: Quantifies how accurately the model describes the complete degradation state by measuring the Hamming distance between the predicted and ground truth vectors.

IV. Experimental Results

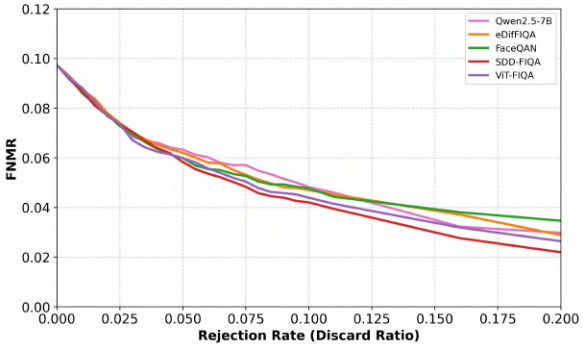
In this section, we present the zero-shot FIQA performance of diverse VLMs. We organize our evaluation into three parts: first, we benchmark biometric utility on real-world datasets; second, we analyze internal consistency and prompt sensitivity; and finally, we conduct a synthetic ablation to assess degradation-detection capabilities in a controlled environment. In addition to our systematic analysis, we provide a qualitative examination of sample VLM outputs to visually demonstrate their interpretability.

A. Experimental Setup

a) Models: We evaluate a broad selection of state-of-the-art open-source VLMs to assess the impact of architecture and scale on FIQA performance. We analyze the QWEN Family (QWEN2-7B, QWEN2.5-7B/32B/72B) [35], [5] to observe scaling trends, Gemma-3-4B [31], Idefics-9B [18], and Phi-4-6B [1] as other lightweight alternatives.



(a) Comparison of different VLM Architectures (Simple Prompt)



(b) QWEN2.5-7B vs. SOTA Baselines

Fig. 2: Error-versus-Reject (EvR) curves on LFW.

b) Baselines: We compare these zero-shot estimators against learning-based FIQA methods: eDifFIQA [3], FaceQAN [4], SDD-FIQA [25], and ViT-FIQA [2].

B. Biometric Utility Analysis (Error-versus-Reject)

We first evaluate the primary requirement of FIQA: the ability to filter out samples that cause recognition errors. Fig. 2 presents the error-versus-reject (EvR) curves on the LFW dataset (ArcFace [10] embeddings, FMR=10^{-3}).

Quantitative results are detailed in Table I, where AUC scores are averaged across multiple recognition backbones (ArcFace [10], TransFace [9], and LVFace [40]) to ensure the metrics are not biased toward a specific FR architecture.

a) VLM Performance: Fig. 2 demonstrates that most VLMs can replicate the behavior of FIQA methods in the zero-shot setting. Particularly, QWEN2.5-7B shows a curve that aligns closely with these methods. We can see similar results in Table I. In AUC (@ 10% and 20%), QWEN2.5-72B not only outperforms other VLM variants but also surpasses the specialized baselines, except SDD-FIQA. This indicates that VLMs' advanced internal reasoning can effectively proxy biometric quality estimation without task-specific supervision.

b) Scaling and Architecture: The QWEN family results in Table I reveal a complex relationship between scale and utility that varies by rejection depth. While the 72B model dominates in the broader range (best AUC @ 10% and 20%), the smaller 7B variant surprisingly proves more effective in the high-utility regime, achieving the best VLM performance

TABLE I: Main Benchmark: Biometric Utility. We report the accumulated partial AUC (lower is better). VLMs are evaluated using the standard 'Simple' prompt.

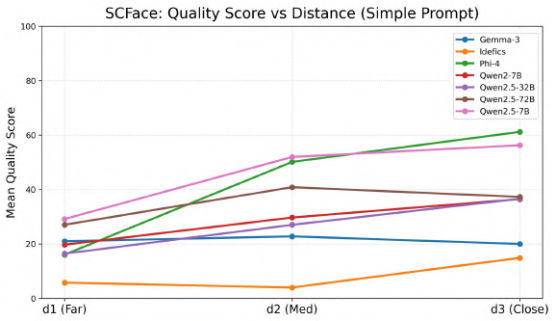
Method	AUC @ 1%	AUC @ 5%	AUC @ 10%	AUC @ 20%
Supervised / Specialized Baselines				
SDD-FIQA	0.00091	0.00376	0.00622	0.00833
ViT-FIQA	0.00092	0.00372	0.00627	0.00856
FaceQAN	0.00092	0.00375	0.00638	0.00891
eDifFIQA	0.00092	0.00381	0.00649	0.00903
Vision-Language Models (Zero-Shot - Simple Prompt)				
QWEN2.5-72B	0.00092	0.00381	0.00624	0.00850
QWEN2.5-32B	0.00092	0.00379	0.00659	0.00936
QWEN2.5-7B	0.00092	0.00378	0.00657	0.00898
QWEN2-7B	0.00092	0.00382	0.00644	0.00937
Phi-4	0.00094	0.00399	0.00703	0.01016
Gemma-3	0.00097	0.00437	0.00779	0.01127
Idefics	0.00097	0.00485	0.00971	0.01555

at 5%. Likewise, the 7B model consistently outperforms the mid-sized 32B variant across most metrics. Additionally, we observe a significant performance gap across architectures; despite comparable sizes, Gemma-3 and Idefics lag behind the QWEN series, underscoring that the underlying architectural choices are more critical than parameter count alone.

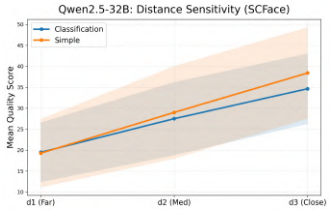
C. Sensitivity to Surveillance Degradation

A robust quality estimator must reflect physical degradation. We use the SCFace dataset to measure how scores change across three standoff distances: Far ($d1$, 4.2m), Medium ($d2$, 2.6m), and Close ($d3$, 1.0m).

The impact of distance on model scoring is shown in Fig. 3. While the QWEN family shows the expected monotonic increase in quality as the subject approaches the camera, other models struggle to differentiate the scenarios (Fig. 3a). Specifically, QWEN2.5-32B demonstrates a sharp distinction between distances (Fig. 3b), with mean quality scores rising consistently: $\approx$ 16.5 (Far), $\approx$ 27.0 (Med), $\approx$ 36.5 (Close). This confirms the model correctly perceives the loss of facial detail. In contrast, Gemma-3 produces nearly flat scores over distances (around 20.0), whereas Idefics assigns 0 to almost all images at the Far and Med distances.



(a) Global Model Comparison (Simple Prompt)



(b) Detailed Trend for QWEN2.5-32B

Fig. 3: Score sensitivity to physical distance in surveillance scenarios (SCFace). (a) Mean quality scores across architectures. (b) Score distributions for QWEN2.5-32B across the three distances.

Understanding why VLMs reduce the importance of distant faces involves analyzing the attribute classification outputs of QWEN2.5-32B, as illustrated in Fig. 4. For $d1$ (Far) images, the model assigns higher probabilities to "Low Resolution" and "Blurry." As the subject moves closer, the distribution of generated attribute labels changes consistently, confirming that lower scores are driven by recognized physical degradation rather than arbitrary noise.

D. Prompt Consistency and Robustness

We examine whether changing the prompt phrasing alters the model's ranking logic. We compare the base "Simple" prompt against "Utility," "Reliability," and "Classification" variants across different model scales.

Fig. 5 suggests that semantic phrasing introduces minor calibration offsets while preserving the underlying quality scale. The "Classification" prompt tends to result in slightly stricter scoring (negative bias) across architectures. However, model scale affects optimal phrasing: the larger 32B model is most consistent with "Reliability," whereas the smaller 7B model aligns best with "Utility."

Table II presents the AUC performance for prompt variants. For the larger QWEN2.5-32B, the internal ranking logic appears highly robust, showing a possible invariance to semantic signals. In contrast, the smaller QWEN2.5-7B is more affected by prompting. Both "Reliability" and "Utility" variants outperform the base prompt, suggesting that smaller architectures may require explicit semantic cues to effectively activate their biometric-aligned features.

The "Classification" strategy yields divergent results across model families and scales. While it improves the QWEN2-7B and remains effective for the QWEN2.5-32B model, it degrades the performance of newer small-scale models like QWEN2.5-7B and Phi-4. This indicates that while asking for categorical labels can stabilize certain architectures, lightweight models may lose fine-grained ranking resolution when forced into a discrete classification mode.

E. Internal Consistency

Finally, we assess the "Internal Consistency" of the VLM. A valid explainer must be consistent with its own scores: if the model labels some images as "Blurry" or "Low Resolution," the average score assigned to those images must be lower than that of images labeled "Clear".

Fig. 6 validates this behavior for QWEN2.5-32B, demonstrating a strictly monotonic connection between generated text labels and scalar scores. By plotting the mean quality

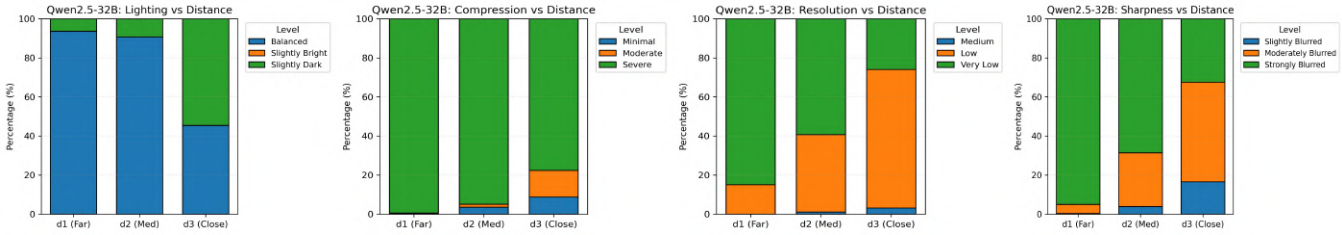


Fig. 4: Explainability analysis on SCFace (QWEN2.5-32B). The plot shows the distribution of generated attribute labels across the three camera distances.

score against the model's predicted degradation levels across datasets (CelebA, IJB-B, LFW) and their global average, we observe a trend: as the textual description shifts from "Clear" to "Severe," the scalar utility drops greatly. This alignment is universal across all evaluated dimensions—Compression, Lighting, Resolution, and Sharpness—where images labeled "Clear" consistently maintain high global averages ($\approx$ 80–95), while those flagged as "Severe" fall to the 5–25 range. This confirms that the scalar score is not an arbitrary hallucination but is semantically grounded in the model's explicit perception of visual artifacts.

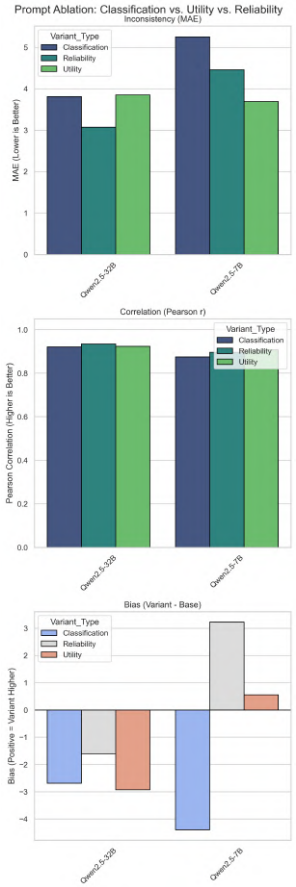


Fig. 5: Prompt Ablation Study: Comparison of score distributions between the Simple Quality prompt and Classification Quality, Utility, and Reliability prompt variants.

F. Synthetic Ablation: Mix Degradation

To further dissect the model's behavior, we evaluate its performance on our **Mix Degradation Benchmark**. This test measures three critical dimensions: trustworthiness on clean images (L_0), the ability to detect severe degradations in complex mixtures (L_2), and the precision of categorical outputs (Hamming Distance).

a) Clean Image Trustworthiness (L_0): We first evaluate the model's tendency to predict degradations on perfectly clean images (False Positive rate). As shown in Table III, QWEN2.5-7B emerges as the most reliable detector, correctly identifying 87.9% of clean images as degradation-free. However, its scoring logic is erratic; when it hallucinates a single degradation, the quality score collapses disproportionately and behaves non-monotonically as errors increase. In contrast, while the larger QWEN2.5-32B notices artifacts (hallucinating in $\approx 50\%$ of samples), it maintains superior internal consistency; the quality scores decrease consistently as more false features appear.

TABLE II: Effect of Prompt Phrasing on Biometric Utility (accumulated partial AUC). Comparing all prompt strategies. Lower is better.

Model	Variant	AUC @ 1%	AUC @ 5%	AUC @ 10%	AUC @ 20%
QWEN2-7B	Simple	0.00092	0.00382	0.00644	0.00937
	Classif.	0.00093	0.00377	0.00614	0.00816
QWEN2.5-7B	Simple	0.00092	0.00378	0.00657	0.00898
	Reliability	0.00092	0.00378	0.00628	0.00849
	Utility	0.00092	0.00375	0.00625	0.00845
	Classif.	0.00092	0.00376	0.00658	0.00960
QWEN2.5-32B	Simple	0.00092	0.00379	0.00659	0.00936
	Reliability	0.00092	0.00379	0.00664	0.00961
	Utility	0.00092	0.00382	0.00652	0.00933
	Classif.	0.00092	0.00376	0.00653	0.00921
Phi-4	Simple	0.00094	0.00399	0.00703	0.01016
	Classif.	0.00094	0.00425	0.00741	0.01069

TABLE III: L0 Analysis (Clean Images): False Positive (FP) rates and associated Quality Scores (QS).

Model	0 FP		1 FP		2 FP		$\geq$3 FP	
	%	QS	%	QS	%	QS	%	QS
QWEN2.5-7B	87.9	94.8	11.1	33.6	0.9	44.3	0.2	40.3
QWEN2-7B	75.7	90.1	18.8	43.8	3.0	46.3	2.5	29.7
QWEN2.5-32B	50.9	95.0	7.5	82.9	39.2	69.6	2.4	44.6

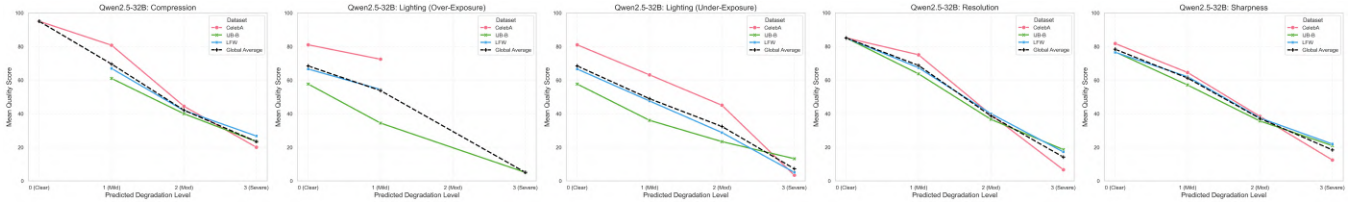


Fig. 6: Internal Consistency (QWEN2.5-32B). Boxplots of scalar quality scores grouped by the model's generated text labels.

b) Degradation Detection & Recall (L_2): In the mixed degradation scenario, we evaluate the model's ability to isolate a "Hard" artifact from milder degradations. Table IV summarizes the detection completeness and "Hard" artifact recall. QWEN2-7B proves to be the most sensitive artifact detector, achieving the highest recall for the "Hard" artifact (79.8%) and the best completeness (identifying all 3 degradations in 39.7% of cases). QWEN2.5-7B follows as a balanced alternative in general. Conversely, the larger QWEN2.5-32B fails significantly in this dense detection task; it successfully identifies all three degradations in only 8.7% of cases, often overlooking the severe artifact entirely.

c) Output Precision (Hamming Distance): Finally, we evaluate the exactness of the predicted degradation vectors by analyzing the Hamming prediction errors (D_H) in Table V. QWEN2.5-32B achieves the best low-error performance, securing the highest rates for perfect matches ($D_0 = 7.8\%$) and single-bit errors ($D_1 = 32.1\%$). However, its error distribution is broad, with significant mass spilling into high-error categories ($D_3 = 23.2\%$). In contrast, the 7B models exhibit a rigid, systematic bias: they peak sharply at D_2 ($\approx 55\%$) while keeping severe errors (D_3) relatively low. This suggests that smaller models deterministically miss specific artifact pairs (likely mild ones), whereas the larger model is more capable but less predictable.

G. Sample Outputs

We visualize sample VLM outputs across the four benchmark datasets in Fig. 7 to provide a qualitative perspective. The figure highlights the model's ability to generate both scalar quality scores and semantic attribute classifications. For the high-quality reference sample from CelebA-HQ (Fig. 7a), the models consistently assign high quality scores and correctly identify the attributes as "Clear" and "High Resolution." In contrast, for the surveillance sample from SCFace (Fig. 7b), the models successfully detect severe degradations, flagging the image as blurry or low resolution and attenuating the quality score accordingly. This qualitative evidence affirms our quantitative findings, demonstrating that VLMs can effectively translate visual artifacts into interpretable textual descriptions.

TABLE IV: L2 Analysis (Mixed Degradation): Detection Completeness and Hard Artifact Recall.

Model	Completeness				Hard Artifact
	0/3	1/3	2/3	3/3	Recall (%)
QWEN2-7B	2.2	12.2	45.9	39.7	79.8
QWEN2.5-7B	3.6	32.3	27.4	36.7	71.3
QWEN2.5-32B	1.7	41.8	47.8	8.7	62.0

TABLE V: Hamming Distance Breakdown: Distribution of prediction errors (D_H).

Model	D0 (Perf.)	D1	D2	D3	D4	D5
QWEN2.5-32B	7.8	32.1	31.8	23.2	3.7	1.5
QWEN2-7B	3.5	21.2	58.6	14.4	2.3	0.0
QWEN2.5-7B	3.1	26.2	54.1	15.2	1.3	0.0

V. Limitations

While our findings demonstrate the strong zero-shot capabilities of VLMs for FIQA, a primary limitation is their computational latency. Our latency analysis (using a batch size of 16) shows that even a fast VLM with a simple prompt (QWEN2.5-7B at 33.3 ms/image) is nearly 50× slower than a classical baseline like eDifFIQA (0.7 ms/image). Further, switching to different VLM architectures makes the process approximately 3 times slower; for instance, Gemma-3 and Phi-4 record latencies of 104.9 ms/image and 111.4 ms/image, respectively. The same 3-fold slowdown holds true for generating more descriptive text: using a longer, classification-based prompt on QWEN2.5-7B increases latency to 102.7 ms/image. Consequently, VLMs are too slow for high-throughput pipelines and should serve solely as a supplementary mechanism. Beyond computational costs, VLMs also struggle to reliably differentiate between high and very high-quality images. As our synthetic ablation reveals, larger models exhibit up to a 50% hallucination rate on completely clean samples, demonstrating a hyper-critical over-sensitivity rather than precise utility ranking.

VI. Conclusion

In this work, we investigated the potential of off-the-shelf Vision-Language Models (VLMs) to serve as zero-shot estimators for Face Image Quality Assessment (FIQA). To address the interpretability limitations of traditional "black-box" methods, we introduced a comprehensive evaluation framework prioritizing biometric utility. We benchmarked a diverse set of open-source models against traditional FIQA baselines using error-versus-reject (EvR) curves. Furthermore, we examined the models' interpretability and consistency through qualitative analysis, including synthetic ablation and surveillance-specific scenarios. We also provided example VLM outputs to demonstrate their behavior in sample face images. Based on our comprehensive analysis, we derived the following key observations regarding the viability of VLMs for FIQA:

CelebA
11660.jpg

Simple Prompt Scores

Model	Score
Gemma-3	75
Idefics	90.00
Phi-4	80
Qwen2-7B	70
Qwen2.5-32B	75
Qwen2.5-72B	85
Qwen2.5-7B	85

Classification Results & Predicted Degradations

Model	Score	Sharpness	Resolution	Lighting	Compression
Idefics	90.00	Clear	High	Balanced	None
Phi-4	85	Clear	High	Balanced	None
Qwen2-7B	70	Clear	High	Balanced	None
Qwen2.5-32B	75	Clear	Medium	Balanced	Minimal
Qwen2.5-7B	85	Clear	High	Slightly Dark	Minimal

(a) CelebA-HQ

SCFace
045_cam3_2.jpg

Simple Prompt Scores

Model	Score
Gemma-3	3.00
Idefics	90.00
Phi-4	50
Qwen2-7B	20
Qwen2.5-32B	20
Qwen2.5-72B	25
Qwen2.5-7B	65

Classification Results & Predicted Degradations

Model	Score	Sharpness	Resolution	Lighting	Compression
Idefics	90.00	Clear	High	Balanced	None
Phi-4	50	Slightly Blurred	Low	Moderately Dark	Moderate
Qwen2-7B	20	Strongly Blurred	Low	Slightly Dark	Severe
Qwen2.5-32B	20	Strongly Blurred	Very Low	Balanced	Severe
Qwen2.5-7B	65	Slightly Blurred	Low	Slightly Dark	Severe

(b) SCFace

LFW
Itamar_Franco_0001.jpg

Simple Prompt Scores

Model	Score
Gemma-3	75
Idefics	80.00
Phi-4	82
Qwen2-7B	60
Qwen2.5-32B	75
Qwen2.5-72B	75
Qwen2.5-7B	85

Classification Results & Predicted Degradations

Model	Score	Sharpness	Resolution	Lighting	Compression
Idefics	90.00	Clear	High	Balanced	None
Phi-4	85	Clear	High	Balanced	None
Qwen2-7B	60	Moderately Blurred	Medium	Slightly Dark	None
Qwen2.5-32B	65	Slightly Blurred	Medium	Balanced	Minimal
Qwen2.5-7B	85	Clear	High	Slightly Dark	None

(c) LFW

IJB-B
180090.jpg

Simple Prompt Scores

Model	Score
Gemma-3	6.00
Idefics	0.00
Phi-4	0
Qwen2-7B	20
Qwen2.5-32B	35
Qwen2.5-72B	35
Qwen2.5-7B	65

Classification Results & Predicted Degradations

Model	Score	Sharpness	Resolution	Lighting	Compression
Idefics	90	Clear	High	Balanced	None
Phi-4	50	Slightly Blurred	Low	Moderately Dark	Moderate
Qwen2-7B	20	Strongly Blurred	Low	Slightly Dark	Severe
Qwen2.5-32B	35	Moderately Blurred	Low	Balanced	Moderate
Qwen2.5-7B	65	Slightly Blurred	Low	Slightly Dark	Severe

(d) IJB-B

Fig. 7: Sample output of VLMs from all four datasets.

- **Observation 1: Biometric utility depends on architecture, not just parameter count.** Our evaluation indicates that a VLM's ability to proxy biometric utility depends significantly on architectural choices, rather than purely on parameter count. We observed that, while specific model families, such as QWEN, achieved competitive alignment with specialized FIQA methods, other models with similar capacities missed key quality cues. This suggests that zero-shot FIQA capability is not universal across all large-scale models.
- **Observation 2: VLMs effectively bridge the interpretability gap.** Regarding interpretability, our analysis reveals that capable VLMs can provide the descriptive feedback often missing in current pipelines. In surveillance scenarios, these models not only downgraded low-quality samples but also provided textual justifications, such as "Low Resolution" or "Blur", that accurately reflected the visual degradation. This transparency underscores the potential of VLMs to support human-in-the-loop applications that require actionable feedback.
- **Observation 3: There is a complex trade-off between descriptive precision and scoring consistency.** Our synthetic ablation revealed a complex trade-off between scoring utility and descriptive precision. We found that larger models frequently hallucinate degradations on clean images. Yet unlike their smaller counterparts, they maintained strict internal consistency, with scalar scores degrading monotonically with the severity of the generated critique. This implies that while larger models may be hyper-critical, their scoring logic appears more rational and predictable than the erratic behavior observed in smaller architectures.
- **Observation 4: Robustness to prompt phrasing scales inversely with model size.** We observed that prompt sensitivity tends to scale inversely with model size. While the ranking logic of larger models was robust to phrasing changes, smaller architectures were more strongly affected by targeted "Utility" or "Reliability" prompts. This indicates that smaller models likely benefit from specific semantic triggers to align their features with biometric requirements, whereas larger models exhibit a more generalized and stable concept of quality.

In future work, we plan to broaden our research by incorporating additional VLMs and evaluating their performance across a wider range of datasets. Additionally, we intend to integrate more biometric factors, such as expression, pose, and occlusion, into the assessment pipeline. Finally, we aim to compile these findings into a comprehensive benchmarking scheme.

ACKNOWLEDGMENTS

This work was supported by the Meetween Project that received funding from the European Union's Horizon Europe Research and Innovation Programme under Grant Agreement No. 101135798 and supported in parts by the ARIS grant P2-0250. Computing resources used in this work were provided by the National Center for High Performance Computing of Turkey (UHeM) under grant number 4023702025.

REFERENCES

[1] M. Abdin, J. Aneja, H. Behl, S. Bubeck, R. Eldan, S. Gunasekar, M. Harrison, R. J. Hewett, M. Javaheripi, P. Kauffmann, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.

[2] A. Atzori, F. Boutros, and N. Damer. Vit-fiqa: Assessing face image quality using vision transformers. *arXiv:2508.13957*, 2025.

[3] v. Babnik, P. Peer, and V. Štruc. ediffiqa: Towards efficient face image quality assessment based on denoising diffusion probabilistic models. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 6(4):458–474, 2024.

[4] Z. Babnik, P. Peer, and V. Struc. Faceqan: Face image quality assessment through adversarial noise exploration. In *ICPR*, pages 748–754, 2022.

[5] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, Y. Xu, and J. Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[6] A. Chaubey, X. Guan, and M. Soleymani. Face-llava: Facial expression and attribute understanding through instruction tuning. *arXiv preprint arXiv:2504.07198*, 2025.

[7] T. Chen, J. Zhang, et al. Mgffd-vlm: Multi-granularity prompt learning for face forgery detection with vlm. *arXiv:2507.12232*, 2025.

[8] W.-T. Chen, G. Krishnan, et al. Dsl-fiqa: Assessing facial image quality via dual-set degradation learning and landmark-guided transformer. *arXiv:2406.09622*, 2024.

[9] J. Dan, Y. Liu, H. Xie, J. Deng, H. Xie, X. Xie, and B. Sun. Transface: Calibrating transformer training for face recognition from a data-centric perspective. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20642–20653, 2023.

[10] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proc. CVPR*, pages 4690–4699, 2019.

[11] Y. Gao, X. Min, J. Han, et al. Multi-dimensional text-to-face image quality assessment using llm: Database and method. In *Proc. ACM Multimedia (MM)*, 2025.

[12] M. Grgic, K. Delac, and S. Grgic. Scface – surveillance cameras face database. *Multimedia Tools and Applications*, 51(3):863–879, 2011.

[13] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In *Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition (ECCV Workshop)*, 2008.

[14] B. Jo, D. Cho, I. K. Park, and S. Hong. Ifqa: Interpretable face quality assessment. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 3444–3453, 2023.

[15] W. Kabbani, K. Raja, R. Ramachandra, and C. Busch. Faceoracle: Chat with a face image oracle. In *Proc. ECCV Workshops*, 2025.

[16] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *Proc. ICLR*, 2018.

[17] M. Kim, A. K. Jain, and X. Liu. Adaface: Quality adaptive margin for face recognition. In *Proc. CVPR*, pages 18750–18759, 2022.

[18] H. Laurençon, L. Saulnier, L. Tronchon, S. Bekman, A. Singh, A. Lozhkov, T. Wang, S. Karamcheti, A. M. Rush, D. Kiela, M. Cord, and V. Sanh. Obelics: An open web-scale filtered dataset of interleaved image-text documents, 2023.

[19] K. Li, Z. Yang, J. Zhao, H. Shen, R. Hou, H. Chang, Y. Yu, and X. Chen. Herm: Benchmarking and enhancing multimodal llms for human-centric understanding. *arXiv preprint arXiv:2410.06777*, 2024.

[20] K.-H. Lin et al. Instructflip: Exploring unified vision-language model for face anti-spoofing. In *Proc. ACM Multimedia (MM)*, 2025.

[21] S. Ma, W.-T. Chen, Q. Gao, J. Wang, C. W. Zhou, et al. Vquala 2025 challenge on face image quality assessment: Methods and results. *arXiv preprint arXiv:2508.18445*, 2025.

[22] T. Miyata. Zen-iqa: Zero-shot explainable and no-reference image quality assessment with vision language model. *IEEE Access*, 12:70973–70983, 2024.

[23] N. Najafzadeh, H. Kashiani, M. S. E. Saadabadi, N. A. Talemi, S. R. Malakshan, and N. M. Nasrabadi. Face image quality vector assessment for biometrics applications. In *Proc. WACV*, pages 511–520, 2023.

[24] H. Otroshi Shahreza and S. Marcel. Facellm: A multimodal large language model for face understanding. In *Proc. ICCV*, pages 3677–3687, 2025.

[25] F.-Z. Ou, X. Chen, et al. Sdd-fiqa: Unsupervised face image quality assessment with similarity distribution distance. In *CVPR*, 2021.

[26] F.-Z. Ou, C. Li, S. Wang, and S. Kwong. Clib-fiqa: Face image quality assessment with confidence calibration. In *Proc. CVPR*, 2024.

[27] F.-Z. Ou, C. Li, S. Wang, and S. Kwong. Mr-fiqa: Face image quality assessment with multi-reference representations from synthetic data generation. In *Proc. ICCV*, 2025.

[28] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[29] E. Saritaş and H. K. Ekenel. Analyzing the effect of combined degradations on face recognition. In *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–5. IEEE, 2024.

[30] T. Schlett, C. Rathgeb, O. Henniger, J. Galbally, J. Fierrez, and C. Busch. Face image quality assessment: A literature survey. *ACM Computing Surveys*, 54(10s):1–49, 2022.

[31] G. Team, A. Kamath, J. Ferret, S. Pathak, N. Vieillard, R. Merhej, S. Perrin, T. Matejovicova, A. Ramé, M. Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

[32] P. Terhörst, J. N. Kolf, N. Damer, F. Kirchbuchner, and A. Kuijper. Ser-fiq: Unsupervised estimation of face image quality based on stochastic embedding robustness. In *Proc. CVPR*, pages 5651–5660, 2020.

[33] H. Wang, Y. Shi, Z. Tao, Y. Gao, L. Zhang, X. Lin, J. Zhao, and X. Cao. Faceshield: Explainable face anti-spoofing with multimodal large language models. *arXiv preprint arXiv:2505.09415*, 2025.

[34] J. Wang, K. Y. Chan, and C. C. Loy. Exploring clip for assessing the look and feel of images. In *Proc. AAAI*, volume 37, pages 2555–2563, 2023.

[35] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[36] C. Whitelam, E. Taborsky, A. Blanton, B. Maze, J. Adams, T. Miller, N. Kalka, A. Jain, J. A. Duncan, W. Niggel, and P. Grother. Iarpa janus benchmark-b face dataset. In *Proc. CVPR Workshops*, pages 90–98, 2017.

[37] H. Wu, Z. Zhang, W. Er, C. Zhu, et al. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. In *Proc. ICML*, 2024.

[38] H. Wu, H. Zhu, Z. Zhang, E. Zhang, C. Chen, L. Liao, C. Li, A. Wang, W. Sun, Q. Yan, X. Liu, G. Zhai, S. Wang, and W. Lin. Towards open-ended visual quality comparison, 2024.

[39] S. Wu, X. Li, X. Xu, J. Jiang, S. Wang, W. Lin, and L. Ma. Fvq-20k: A large-scale dataset and an lmm-based method for face video quality assessment. In *Proc. ACM Multimedia (MM)*, 2025.

[40] J. You, S. Li, Y. Sun, J. Wei, M. Guo, C. Feng, and J. Ran. Lvface: Progressive cluster optimization for large vision models in face recognition. *arXiv preprint arXiv:2501.13420*, 2025.

[41] Z. You, J. Gu, X. Cai, Z. Li, K. Zhu, T. Xue, and C. Dong. Enhancing descriptive image quality assessment with a large-scale multi-modal dataset. *IEEE Transactions on Image Processing*, 2025. to appear.

[42] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10):1499–1503, 2016.

[43] W. Zhang, G. Zhai, Y. Wei, X. Yang, and K. Ma. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In *Proc. CVPR*, pages 14071–14081, 2023.

[44] Q. Zheng, J. Zhang, et al. Qa-vlm: Providing human-interpretable quality assessment for wire-feed laser additive manufacturing parts with vision language models. *arXiv:2508.16661*, 2025.