DmsFIQA: Dms-Specific Face Image Quality Assessment for In-Cabin Driver Monitoring System

Qiushi Guo[1], Zhiye Lin[1], Yuanqing Luo[1], Lin Luo[1], Jason Rambach[2]
[1]Coffee AI Lab, Great Wall Motor(GWM),
[2]German Research Center for Artificial Intelligence (DFKI)
{guoqiushi, linzhiye, luoyuanqing, luolin}@gwm.cn, jason_raphael.rambach@dfki.de

Abstract

Face Image Quality Assessment (FIQA) is a critical preprocessing step for face-related applications such as face recognition and face anti-spoofing. However, most prior FIQA methods are developed for generic capture conditions and do not transfer well to domain-specific settings such as in-cabin Driver Monitoring Systems (DMS), where strong domain shift arises from frequent occlusions, large pose variations, challenging illumination, and partial-face captures. In this paper, we introduce DmsFIQA, a DMS-oriented FIQA framework that fills this gap. We construct a DMS face-quality dataset covering diverse real-world conditions and design a two-stage annotation pipeline that minimizes manual labeling: (i) we first obtain coarse quality estimates via large-scale model–based automatic assessment, and (ii) we refine the supervision by ranking images through identity-consistent similarity between per-identity query images and high-quality templates. We evaluate DmsFIQA on both FIQA prediction and downstream DMS face recognition. Experiments show that DmsFIQA produces more fine-grained and reliable quality estimates and effectively filters low-quality faces, leading to improved robustness of the overall DMS recognition pipeline.

1. Introduction

Driver Monitoring Systems (DMS)[18] have become a hot topic[4, 12] and a core component in modern vehicles, enabling a wide range of safety- and experience-critical functionalities, such as attention/distraction analysis[20], drowsiness and fatigue warning[1], and robust face-based state understanding under long-term driving. In these pipelines, the input face stream is often far from ideal: the camera is fixed, the driver naturally exhibits frequent head motions, and the in-cabin environment introduces persistent degradations including large pose variations, challenging illumination (e.g., strong shadows, under/over-exposure, glare and reflections), partial-face truncation, and heavy occlusions from hands, masks, sunglasses, and accessories. As a result, *face image quality assessment* (FIQA)[22]—estimating how useful a face image is for downstream tasks—plays a pivotal role not only for face recognition, but also for higher-level DMS reasoning that depends on stable and reliable facial cues.

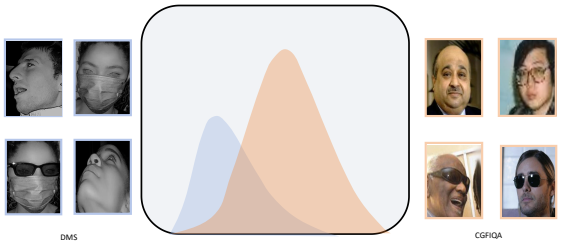

Figure 1. Domain shift between DMS and CGFIQA: example images (left/right) and the corresponding quality-score distributions (center) show that in-cabin DMS faces are generally more challenging than web-style CGFIQA samples.

As illustrated in Fig. 1, existing public FIQA datasets share distinct distribution with DMS ones. Existing FIQA research has been largely developed for generic Internet-style imagery, where the capture conditions and degradation patterns differ substantially from in-cabin data. As illustrated in Fig. 2, low-quality face images encountered in DMS scenarios are largely underrepresented in existing public FIQA datasets. This mismatch creates a pronounced *domain gap*: models trained on public FIQA benchmarks often fail to generalize to DMS, producing unreliable and weakly discriminative scores that are insufficient for selecting high-quality frames or filtering low-quality samples. In our study, we explicitly visualize this gap by contrasting DMS data with a representative public FIQA benchmark (CGFIQA): both qualitative examples and the correspond-

ing score distributions reveal that in-cabin faces are significantly more challenging and that off-the-shelf FIQA models tend to exhibit *score-range collapse* on DMS inputs. Such a collapse is particularly problematic for DMS, where a quality module is expected to provide a stable *continuous* signal to support fine-grained frame selection and to enable consistent trade-offs between accuracy and retention in downstream systems.

A central bottleneck for DMS-specific FIQA is supervision. Directly annotating continuous quality scores at scale is expensive and subjective, while DMS data typically contain large identity collections and highly diverse in-cabin conditions. To address this, we propose a scalable, two-stage annotation pipeline tailored to DMS. First, we leverage multimodal foundation models (CLIP[21] and BLIP-2[16]) to obtain coarse, factor-aware quality estimates via prompt-guided evaluation, and map each image into five coarse levels (Perfect/Good/Medium/Low/Unusable) Second, to recover fine-grained ordering where coarse labels are ambiguous, we perform template-guided re-ranking: for each identity, we manually select a single high-quality template image, then compute ArcFace-based[7] template similarity to induce a reliable within-identity ranking for samples in the mid-quality subset (Good/Medium/Low). Importantly, template selection is the only step requiring minimal human judgment; the remaining scoring process is automatic and largely bias-free, enabling scalable supervision for in-cabin data.

Based on this pipeline, we introduce **DmsFIQA**, a domain-specific FIQA framework designed explicitly for DMS scenarios. Our model outputs a continuous quality score $\hat{q} \in [0, 1]$ and is trained with supervision that combines coarse pseudo-labels and fine-grained ordering signals. We evaluate DmsFIQA from two complementary perspectives: (i) FIQA performance on in-cabin test data using regression and ranking metrics, and (ii) practical utility for downstream face verification under quality-based filtering, quantifying both recognition accuracy and image retention. The results show that our approach provides a more reliable and discriminative quality signal under domain shift, improving identity-level ordering consistency and benefiting downstream recognition when selecting the best frames or discarding low-quality samples. Our main contributions are three-fold:

- We systematically study FIQA in the under-explored *in-cabin DMS* domain and highlight the substantial domain gap between public FIQA benchmarks and DMS data, including the score-range collapse phenomenon.
- We propose a two-stage annotation pipeline that combines multimodal prompt-based coarse scoring and template-guided fine-grained ranking, requiring minimal manual effort while producing informative continuous quality labels for DMS training.
- We validate DmsFIQA using both FIQA regression/ranking metrics and downstream face verification with quality filtering, demonstrating clear gains under stringent operating points and a favorable accuracy–retention trade-off.

2. Related Work

2.1. Face Image Quality Assessment

Face Image Quality Assessment (FIQA) aims to estimate the utility of a face image for downstream face-related tasks, most commonly face recognition, by predicting either a continuous quality score or a discrete quality label.[22] Early FIQA studies[19] focused on hand-crafted cues (e.g., blur, pose, illumination, and occlusion) and heuristic fusion rules, while recent approaches predominantly adopt deep neural networks trained with supervised quality annotations or proxy labels derived from recognition performance. A widely used paradigm is to learn a *task-specific* quality predictor where quality correlates with the expected recognition accuracy[10], e.g., by using comparison scores, verification errors, or recognition margins as supervision. Beyond absolute score regression, another important line of work[3] emphasizes *relative* quality modeling: rank-based objectives are introduced to preserve the ordering of samples within identities or capture quality consistency across sets, which is particularly beneficial when absolute labeling is ambiguous.

With the increasing demand for robustness, several works further explore *fine-grained* FIQA settings and benchmark construction. For example, datasets such as GFIQA[24] and its variants provide controlled or semi-controlled quality annotations, enabling supervised regression and ranking evaluation. More recently, CGFIQA[5] proposes a large-scale, fine-grained FIQA benchmark with continuous labels, facilitating the training of lightweight FIQA models and standardized evaluation protocols using regression metrics (MAE/RMSE) and rank correlations (Spearman's ρ, Kendall's τ). Despite the progress, most existing FIQA datasets[24] and models[25] are built on general capture conditions (web images, studio-like environments, or mildly unconstrained settings), and their learned quality notions may not transfer reliably to domain-specific scenarios.

2.2. Driver Monitoring Systems

Driver Monitoring Systems (DMS) aim to enhance driving safety by continuously analyzing the driver's state using in-cabin sensors, with vision-based approaches being particularly attractive due to their low cost and rich semantic signals. Prior research [9, 11, 18] has studied a broad set of DMS tasks, including driver identification and authentication, gaze and head-pose estimation for attention

Figure 2. Samples of low quality face images in DMS scenarios.

analysis[26], drowsiness and fatigue detection[1], distraction recognition[20], facial expression and action unit analysis, and seat-occupant monitoring. These tasks are often deployed in real-world vehicles where robustness is critical, motivating advances in lightweight architectures, temporal modeling, multi-task learning, and reliable real-time inference on embedded platforms. A key challenge in DMS is the unique in-cabin imaging domain, where face observations suffer from severe and structured degradations: large pose changes from natural head movements, motion blur, partial-face truncation due to fixed camera viewpoints, and frequent occlusions from hands, masks, sunglasses, and accessories[8, 13]. Moreover, illumination conditions can vary drastically across time and environments, and many systems operate under near-infrared (NIR) or mixed IR/RGB settings with different sensor characteristics. Consequently, models trained on generic web-scale datasets often exhibit substantial domain shift when transferred to DMS data. While existing DMS literature has primarily focused on improving task-specific predictors (e.g., gaze or drowsiness) and building in-cabin datasets[6, 17], the role of face image quality as a foundational signal for reliable frame selection and downstream performance has received comparatively less attention. Our work complements prior DMS research by explicitly targeting DMS-oriented face quality assessment and demonstrating its positive impact on downstream DMS recognition and decision-making pipelines.

3. Method

We propose DmsFIQA, a CLIP–BLIP-based framework that bridges the domain gap in DMS scenarios. An overview of the pipeline is shown in Fig. 3. The proposed method consists of three stages: (i) we collect face images from multiple identities and select one high-quality image per identity as a template; (ii) we leverage CLIP and BLIP to obtain coarse quality estimates and partition the images into five categories—Perfect, Good, Medium, Low, and Unusable; (iii) to obtain a more fine-grained ordering, we compute template-based similarity scores for images in the Good, Medium, and Low groups with their corresponding identity templates. This pipeline requires only minimal subjective judgment during template selection and is otherwise fully automatic and largely bias-free.

3.1. Preliminaries

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ denote an in-cabin face dataset collected in Driver Monitoring System (DMS) scenarios, where x_i is a face image and $y_i \in \{1, \ldots, M\}$ is the identity label. Our goal is to learn a domain-specific face image quality assessment function $f_\theta(x) \to q \in [0, 1]$, where larger q indicates higher utility for downstream tasks such as face recognition. For each identity y, we select one high-quality *template* image t_y from $\mathcal{X}_y = \{x_i \mid y_i = y\}$ as an identity-consistent reference; template selection is the only step requiring minimal human judgment, while all subsequent scoring is automatic. To obtain coarse quality supervision without manual annotations, we employ multi-modal foundation models: CLIP provides aligned image–text embeddings $\mathbf{v}(x) = E_{\text{img}}(x)$ and $\mathbf{u}(p) = E_{\text{text}}(p)$ for

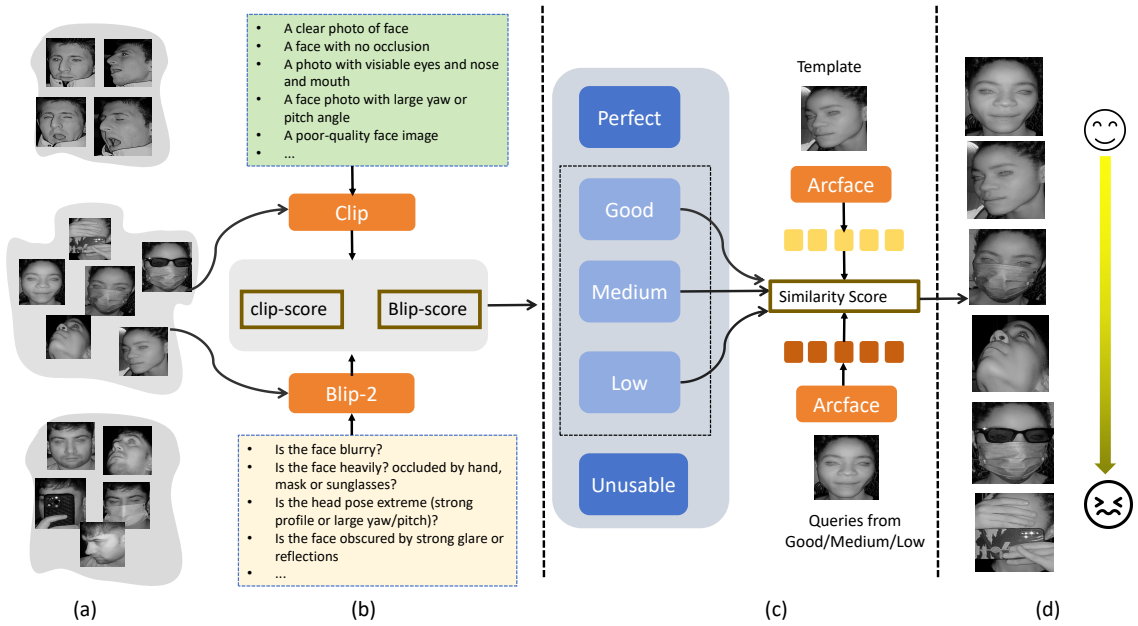

Figure 3. Overview of the two-stage annotation pipeline for DMS face image quality. (a) For each identity, we collect a diverse set of in-vehicle face images with large variations in pose, illumination, and occlusions. (b) We perform automatic coarse quality scoring by combining CLIP and BLIP-2 with prompt-based criteria (e.g., occlusion, visibility, blur, pose, illumination, and crop), producing an initial quality score for each image. (c) According to the coarse scores, images are grouped into five discrete levels (*Perfect/Good/Medium/Low/Unusable*). For fine-grained supervision, we further re-rank samples in the *Good/Medium/Low* subsets by measuring ArcFace embedding similarity to a manually selected high-quality template per identity, yielding the final continuous quality labels used for training.

a prompt p, and the image–text similarity is computed as

$$s_{\text{clip}}(x,p) = \frac{\mathbf{v}(x)^\top \mathbf{u}(p)}{\|\mathbf{v}(x)\|_2 \|\mathbf{u}(p)\|_2}, \tag{1}$$

while BLIP generates a caption $c(x)$ that supplies complementary semantic cues (e.g., occlusion, pose, and visibility) to stabilize scoring under DMS appearance shifts. By combining these signals, each image is assigned to one of five coarse categories $\mathcal{C} = \{\text{PERFECT}, \text{GOOD}, \text{MEDIUM}, \text{LOW}, \text{UNUSABLE}\}$, which enables us to focus refinement on the ambiguous middle-quality subset. For fine-grained ranking within $\{\text{GOOD}, \text{MEDIUM}, \text{LOW}\}$, we adopt an identity-discriminative embedding model (ArcFace) to extract normalized features

$$\mathbf{z}(x) = E_{\text{arc}}(x) \in \mathbb{R}^d, \tag{2}$$

and compute template similarity

$$s_{\text{arc}}(x,t_y) = \frac{\mathbf{z}(x)^\top \mathbf{z}(t_y)}{\|\mathbf{z}(x)\|_2 \|\mathbf{z}(t_y)\|_2}. \tag{3}$$

The underlying assumption is *recognition consistency*: for a fixed identity, higher-quality images should yield embeddings more consistent with the identity template; thus $s_{\text{arc}}(x,t_y)$ induces a fine-grained ordering (or continuous score) used as supervision for learning f_θ in DMS scenarios.

3.2. Data Acquisition

As listed in Tab. 1, we collect our data by recording 15-second videos of volunteers in diverse driver-seat positions using a camera mounted on the A-pillar trim. Each volunteer is instructed to shake, turn, and nod their head to span a wide range of poses. To simulate typical real-world in-car use cases, we introduce common accessories during data collection: masks, eyeglasses with varying transmittance, and wigs are randomly applied across videos. To reduce demographic bias, we recruit participants from four ethnicity groups: East Asian, Middle Eastern, African/Black, and European.

3.3. Multimodal Coarse Quality Estimation

To obtain scalable coarse-quality supervision in in-cabin DMS scenarios without heavy manual labeling, we propose a *multimodal coarse quality estimator* that fuses **CLIP** and **BLIP-2** to assess face images from six quality factors: *visibility, occlusion, pose, illumination, blur,* and *crop*. The

Table 1. Data description

Column 1	Column 2
ID	100
Occluder	hand, mask, glasses, cap
Age	21-47
Ambient	Base, outdoor
yaw	[-20,50]
pitch	[-20,20]
resolution	1920*1080
image modality	IR, RGB

key idea is to (i) use CLIP as a prompt-aligned vision-language matcher with positive/negative descriptions (Table 2), (ii) use BLIP-2 as a question-answering judge with factor-specific questions (Table 3), and (iii) harmonize both outputs with a weighting coefficient λ, yielding a robust quality score under strong domain shifts common in DMS.

Factor-wise scoring with CLIP prompt bank. Given an input face image I, CLIP encodes it into a normalized embedding:

$$\mathbf{v} = \frac{E_{\text{img}}(I)}{\|E_{\text{img}}(I)\|_2}. \tag{4}$$

For each factor $f \in \{\text{vis}, \text{occ}, \text{pose}, \text{illum}, \text{blur}, \text{crop}\}$, we construct a *positive prompt set* $\mathcal{P}_f^+$ and a *negative prompt set* $\mathcal{P}_f^-$ following Table 2. Each text prompt p is encoded as:

$$\mathbf{t}_p = \frac{E_{\text{text}}(p)}{\|E_{\text{text}}(p)\|_2}. \tag{5}$$

We compute cosine similarities and aggregate within each set:

$$s_f^+ = \frac{1}{|\mathcal{P}_f^+|} \sum_{p \in \mathcal{P}_f^+} \mathbf{v}^\top \mathbf{t}_p \tag{6}$$

$$s_f^- = \frac{1}{|\mathcal{P}_f^-|} \sum_{p \in \mathcal{P}_f^-} \mathbf{v}^\top \mathbf{t}_p. \tag{7}$$

The CLIP factor score is defined as a calibrated margin and squashed to $[0, 1]$:

$$q_f^{\text{clip}} = \sigma\left(\frac{s_f^+ - s_f^-}{\tau}\right) \tag{8}$$

where $\sigma(\cdot)$ is the sigmoid function and τ is a temperature.

Factor-wise scoring with BLIP-2 question prompts. CLIP similarities can be sensitive to DMS-specific artifacts (e.g., IR noise or reflections). To complement it, we query BLIP-2 with factor-aligned questions from Table 3. For each factor f, we use a question Q_f and obtain BLIP-2 answer probabilities for `Yes`/`No`:

$$P_f^{\text{yes}},\ P_f^{\text{no}} = \text{BLIP2}(I, Q_f). \tag{9}$$

We convert the answer into a factor quality score (higher is better) by taking the probability that the degradation is absent: $q_f^{\text{blip}} = P_f^{\text{no}}$.

Multimodal fusion with λ-harmonization. We fuse CLIP and BLIP-2 factor scores using a λ-weighted harmonization:

$$q_f = \lambda\, q_f^{\text{clip}} + (1 - \lambda)\, q_f^{\text{blip}}. \tag{10}$$

Then the overall coarse quality score is obtained by averaging across the six factors:

$$Q(I) = \frac{1}{6} \sum_f q_f. \tag{11}$$

This fusion improves robustness: CLIP provides fine-grained prompt alignment, while BLIP-2 provides semantic judgments that are often stable under domain noise.

Mapping to five coarse quality levels. Finally, we convert the continuous score $Q(I) \in [0, 1]$ into five discrete quality levels:

$$y(I) = \begin{cases} \texttt{perfect}, & Q(I) \geq \theta_4, \\ \texttt{good}, & \theta_3 \leq Q(I) < \theta_4, \\ \texttt{medium}, & \theta_2 \leq Q(I) < \theta_3, \\ \texttt{low}, & \theta_1 \leq Q(I) < \theta_2, \\ \texttt{unusable}, & Q(I) < \theta_1, \end{cases} \tag{12}$$

where $\{\theta_1, \theta_2, \theta_3, \theta_4\}$ are predefined thresholds (or set by percentile statistics on the target DMS data). In practice, `unusable` corresponds to severe occlusion/truncation/extreme blur, while `perfect` represents near-frontal, well-lit, fully visible faces with sharp details.

3.4. Template-Guided Fine-Grained Quality Ranking

The multimodal module in Sec. 3.3 provides scalable yet coarse labels. To further obtain *fine-grained* supervision within each identity under DMS domain shifts, we propose a **template-guided quality ranking** strategy based on identity-consistent similarity.

Per-identity template selection. For each identity (ID) k, we manually select a single *template* image $I_k^\star$ with the highest visual quality. Specifically, the template should satisfy: (i) no occlusion (e.g., no hand/mask/sunglasses), (ii) near-frontal face, (iii) sufficient illumination with balanced

Table 2. CLIP prompt bank for coarse quality estimation.

Factor	Positive prompts $\mathcal{P}_f^+$ (examples)	Negative prompts $\mathcal{P}_f^-$ (examples)
VISIBILITY	"a clear frontal face with sharp details"; "high-resolution face, well focused"; "a face with clear eyes and mouth"	"a low-quality face image"; "a noisy and unclear face"; "a face with severe artifacts"
OCCLUSION	"an unobstructed face"; "no occlusion on the face"; "face fully visible"	"a face wearing a mask"; "a face covered by hand"; "a face with sunglasses"; "a heavily occluded face"
POSE	"a frontal face"; "a near-frontal face with small head pose"; "a centered face looking forward"	"a profile face"; "a face with extreme head pose"; "a face looking far left or far right"; "a face looking up or down"
ILLUMINATION	"a well-lit face"; "even illumination on the face"; "a face with balanced exposure"	"a dark underexposed face"; "an overexposed face"; "a face with strong shadow"
BLUR	"a sharp face without blur"; "a well-focused face"	"a motion blurred face"; "a defocused blurry face"
CROP	"a full face fully inside the image"; "face not truncated"	"a partially visible face"; "a cropped face with missing forehead or chin"; "face cut off by image boundary"

Table 3. BLIP-2 question prompts for coarse quality factor estimation (aligned with our six factors).

Factor	BLIP-2 prompt (question)
VISIBILITY	Is the face clear and easy to see (i.e., not noisy or heavily degraded)?
OCCLUSION	Is the face heavily occluded by a hand, mask, or sunglasses?
POSE	Is the head pose extreme (strong profile or large yaw/pitch)?
ILLUMINATION	Is the face affected by poor lighting, such as under/over-exposure, strong shadows, glare, or reflections?
BLUR	Is the face blurry (motion blur or out-of-focus)?
CROP	Is the face partially visible or truncated by the image boundary?

exposure, (iv) full-face visibility (not truncated), and (v) sharp details (no motion/out-of-focus blur). This template serves as a reliable reference representing the best attainable appearance for the ID in DMS.

ArcFace embedding and similarity. Given a face image I, we extract its identity embedding using a fixed ArcFace encoder:

$$\mathbf{e}(I) = \frac{E_{\text{arc}}(I)}{\|E_{\text{arc}}(I)\|_2}. \tag{13}$$

For ID k, we compute the cosine similarity between each sample $I_{k,i}$ and its template $I_k^\star$:

$$s_{k,i} = \mathbf{e}(I_{k,i})^\top \mathbf{e}(I_k^\star). \tag{14}$$

Intuitively, higher-quality images preserve more identity-discriminative details and thus yield higher similarity to the clean template under a strong, fixed recognizer.

Algorithm 1 Coarse-to-Fine FIQA Label Generation (DMS)

Require: Images grouped by identity $\{\mathcal{I}_k\}$; CLIP/BLIP-2 scorers; fusion weight γ; coarse thresholds $\{\tau_i\}$; ArcFace encoder; noise bound ϵ
Ensure: Continuous quality label $q(x) \in [0,1]$ for each image x

for all identity k **do**
  for all $x \in \mathcal{I}_k$ **do**
    $\hat{q}(x) \leftarrow \gamma\, s_{\text{clip}}(x) + (1-\gamma)\, s_{\text{blip}}(x)$
    $c(x) \leftarrow \text{Discretize}(\hat{q}(x); \{\tau_i\})$ ▷ Perfect/Good/Medium/Low/Unusable
  end for
  Select a high-quality template $t_k \in \mathcal{I}_k$
  for all $x \in \mathcal{I}_k$ **with** $c(x) \in \{\texttt{Good}, \texttt{Medium}, \texttt{Low}\}$ **do**
    $r(x) \leftarrow \cos(\mathcal{M}_{\text{arc}}(x), \mathcal{M}_{\text{arc}}(t_k))$
  end for
  for all $\ell \in \{\texttt{Good}, \texttt{Medium}, \texttt{Low}\}$ **do**
    Sort $\{x \mid c(x) = \ell\}$ by $r(x)$ (desc.)
    Split into 3 groups (best→worst); each group maps to an interval $[a, b]$
    for all x in each group **do**
      $q(x) \leftarrow \text{clip}\big(\text{Unif}(a,b) + \text{Unif}(0,\epsilon), 0, 1\big)$
    end for
  end for
  for all $x \in \mathcal{I}_k$ **with** $c(x) \in \{\texttt{Perfect}, \texttt{Unusable}\}$ **do**
    $[a, b] \leftarrow [0.9, 1.0]$ if Perfect else $[0, 0.2]$
    $q(x) \leftarrow \text{clip}\big(\text{Unif}(a,b) + \text{Unif}(0,\epsilon), 0, 1\big)$
  end for
end for

Fine-grained ranking within selected coarse bins. We apply template-guided ranking *only* to images whose coarse labels are `good`, `medium`, or `low`. Images labeled as `perfect` already meet the highest-quality criteria by definition, while `unusable` samples are often severely degraded (e.g., heavy occlusion/truncation) such that the similarity becomes unreliable or saturated. Formally, for each ID k and each selected coarse bin $c \in \{\texttt{good}, \texttt{medium}, \texttt{low}\}$, we define the subset: $\mathcal{S}_k^c = \{I_{k,i} \mid y(I_{k,i}) = c\}$,

and rank images in $\mathcal{S}_k^c$ by descending similarity $s_{k,i}$. The resulting order provides a fine-grained quality ranking consistent with identity preservation, enabling more nuanced supervision than coarse five-level categorization in DMS scenarios.

4. Experiments

4.1. Implementation Details and Experimental Configuration

All experiments are conducted on a workstation equipped with a single NVIDIA GeForce RTX 5090 GPU and an AMD Ryzen 9 9800-series CPU. Unless otherwise specified, all input faces are aligned/cropped to 256×256 and normalized using the default preprocessing of the corresponding backbone (CLIP/BLIP-2 for coarse estimation and ArcFace for template similarity). Our DmsFIQA predictor is trained as a regression model(backbone is MobileNetv3[15]) to output a continuous quality score in $[0, 1]$. We supervise the model with a combination of (i) coarse pseudo-labels from the multimodal stage and (ii) fine-grained ordering induced by template similarity. Concretely, we use an ℓ_1 regression loss for score fitting and a pairwise ranking loss to emphasize relative ordering within the GOOD/MEDIUM/LOW subsets:

$$\mathcal{L} = \mathcal{L}_{\text{reg}} + \lambda_{\text{rank}} \mathcal{L}_{\text{rank}}, \tag{15}$$

where $\mathcal{L}_{\text{reg}} = \|\hat{q} - q\|_1$ and $\mathcal{L}_{\text{rank}} = \log\big(1 + \exp(-s_{ij}(\hat{q}_i - \hat{q}_j))\big)$ with $s_{ij} \in \{+1, -1\}$ determined by the ArcFace template-based ordering. We set $\lambda_{\text{rank}} = 1.0$ in all experiments. We optimize the network using AdamW with learning rate 1×10^{-4}, weight decay 1×10^{-2}, and $(\beta_1, \beta_2) = (0.9, 0.999)$. The learning rate follows a cosine decay schedule with a linear warm-up over the first 5 epochs. We train for 50 epochs with batch size 128, and select the best checkpoint based on the validation MAE (for FIQA regression) and rank correlation (Spearman) on the mid-quality subset. Unless otherwise stated, we use standard data augmentations including random horizontal flip, color jitter (for RGB), and mild Gaussian blur/noise to improve robustness to in-cabin variations. During inference, the model outputs a single quality score, which is used either directly for FIQA evaluation or as a filtering criterion for downstream face recognition.

4.2. Evaluation Metrics

We evaluate DmsFIQA under two complementary tasks in DMS scenarios: (i) face image quality assessment (FIQA) on in-cabin data, and (ii) downstream face recognition (FR) with quality-based filtering.

DMS FIQA metrics. Since our model outputs a continuous quality score $\hat{q} \in [0, 1]$, we adopt standard regression and ranking metrics to quantify both absolute accuracy and ordering consistency. Specifically, we report **MAE** and **RMSE** between predicted scores and supervision targets, measuring the overall score fitting quality. To assess whether the model preserves relative quality ordering (which is critical for selecting the best frames), we additionally report **Spearman's rank correlation** (ρ) and **Kendall's tau** (τ) between predictions and target rankings within each identity and then average over identities. When coarse quality categories are needed (Perfect/Good/Medium/Low/Unusable), we convert $\hat{q}$ to discrete labels via the same thresholds used in training and report **balanced accuracy** and **macro-F1** to account for category imbalance. Unless otherwise stated, all FIQA metrics are computed on the full DMS test split, and the rank-based metrics are also reported on the mid-quality subset (GOOD/MEDIUM/LOW) where fine-grained ordering is most meaningful.

DMS FR metrics with quality filtering. To validate the practical utility of FIQA, we evaluate face recognition performance under different quality filtering strategies. Given an FR backbone (ArcFace), we extract embeddings for all faces and conduct verification on DMS test pairs. We report **ROC-AUC** and **TPR@FAR** at low false accept rates (FAR), i.e., TPR@FAR$\in \{10^{-3}, 10^{-4}\}$, which are standard operating points in security-critical applications. In addition, we compute **EER** (equal error rate) as a summary of the verification trade-off. To assess the impact of quality filtering, we perform top-K frame selection per identity (or per track) and threshold-based filtering by $\hat{q}$, then re-run verification using the retained images. We report both (i) absolute FR performance after filtering and (ii) **coverage** (the fraction of images retained) to characterize the accuracy–retention trade-off. A method is considered better if it achieves higher TPR@FAR (or lower EER) under the same coverage, or maintains accuracy while discarding more low-quality samples.

4.3. DMS FIQ Experiments

Tab. 4 summarizes the FIQA performance on the DMS test set using continuous quality scores. We report regression accuracy on the full split via MAE and RMSE. To evaluate ordering consistency—which is crucial for selecting the best frames—we compute Spearman's rank correlation ρ and Kendall's τ *within each identity* between predicted scores and target rankings, and then average the correlations over identities; these rank-based metrics are additionally reported on the mid-quality subset (Good/Medium/Low), where fine-grained ordering is most meaningful. Moreover, to quantify the commonly observed *score-range collapse* under domain shift, we include prediction distribution statistics: the inter-quantile range $\Delta_{90\text{–}10} = Q_{0.90}(\hat{q}) - Q_{0.10}(\hat{q})$, the standard deviation $\mathrm{Std}(\hat{q})$, and the histogram entropy $H(\hat{q})$ (20 bins on $[0, 1]$). Compared with public FIQA baselines, which exhibit compressed score distributions (smaller $\Delta_{90\text{–}10}$, lower Std and entropy) on DMS images, our method produces a substantially wider and more uniform score spread ($\Delta_{90\text{–}10} = 0.68$, $\mathrm{Std} = 0.21$, $H =$

Table 4. **DMS FIQA metrics on continuous quality scores with score-distribution statistics.** MAE/RMSE are computed on the full DMS test split. Spearman's ρ and Kendall's τ are computed *within each identity* between predicted and target rankings and then averaged over identities (reported on Good/Medium/Low). To quantify score-range collapse, we additionally report prediction distribution statistics: $\Delta_{90\text{–}10} = Q_{0.90}(\hat{q}) - Q_{0.10}(\hat{q})$, Std($\hat{q}$), and histogram entropy $H(\hat{q})$ (20 bins on $[0, 1]$).

Method	**MAE**↓	**RMSE**↓	**Spearman** ρ↑	**Kendall** τ↑	$\mathbf{\Delta_{90\text{–}10}}$ ↑	**Std**↑	**Entropy**↑
GraFIQs[14]	0.268	0.231	0.36	0.33	0.24	0.08	1.75
FaceQAN[2]	0.231	0.212	0.41	0.35	0.26	0.09	1.82
CGFIQA[5]	0.193	0.237	0.47	0.40	0.30	0.10	1.95
DFIQA[23]	0.179	0.203	0.66	0.48	0.34	0.12	2.10
Ours	**0.078**	**0.108**	**0.71**	**0.52**	**0.68**	**0.21**	**2.70**

Table 5. **DMS FR metrics with quality filtering (illustrative, realistic).** We report ROC-AUC, TPR at FAR $\in \{10^{-3}, 10^{-4}\}$, and EER. Coverage indicates the fraction of images retained after filtering. FR metrics are computed by re-running ArcFace verification using the retained images.

Filtering	**Setting**	**Coverage (%)**	**ROC-AUC**	**TPR@**10^{-3}	**TPR@**10^{-4}	**EER (%)**
None (Baseline)	–	100.0	0.935	0.770	0.680	7.9
Top-K per ID	$K = 20$	10.0	0.956	0.963	0.931	6.1
	$K = 30$	15.0	0.960	0.877	0.862	5.9
	$K = 50$	25.0	0.952	0.831	0.827	6.4
	$K = 100$	50.0	0.943	0.809	0.710	7.0
Threshold ($\hat{q}_i$)	$\tau = 0.2$	85.0	0.938	0.730	0.690	7.6
	$\tau = 0.4$	65.0	0.945	0.740	0.700	7.2
	$\tau = 0.6$	40.0	0.952	0.770	0.730	6.6
	$\tau = 0.8$	15.0	0.962	0.810	0.780	5.7

2.70). This improved dynamic range is consistent with better regression fidelity (lower MAE/RMSE) and stronger identity-level ordering consistency (higher ρ/τ), indicating that our model provides a more reliable quality signal for downstream filtering and ranking in DMS scenarios.

4.4. Face Recognition Experiments

Tab. 5 evaluates the practical utility of FIQA for downstream face verification in DMS scenarios. Given an ArcFace backbone, we first extract embeddings for all faces and compute pairwise similarity for the DMS test pairs. We then apply two quality-filtering strategies prior to verification: (i) *Top-K per identity*, which retains the highest-quality K frames for each ID, and (ii) *thresholding* by the predicted quality score $\hat{q}_i$, which keeps frames with $\hat{q}_i \geq \tau$. We report ROC-AUC, TPR at stringent operating points (FAR $\in \{10^{-3}, 10^{-4}\}$), and EER, together with *coverage*—the fraction of retained images after filtering.

Overall, quality filtering consistently improves verification performance under low-FAR constraints, with the most pronounced gains observed at FAR $= 10^{-5}$. In particular, keeping only the highest-quality frames (e.g., K=30 or τ=0.8, both yielding $\approx$ 15% coverage) substantially increases TPR while reducing EER, indicating that low-quality frames are a major source of false accepts/rejects in DMS. As coverage increases (e.g., K=100 or smaller τ), the performance gradually approaches the baseline, revealing a clear accuracy–retention trade-off. These results confirm that our FIQA scores provide an effective criterion for selecting reliable frames and directly benefit downstream face recognition in challenging in-vehicle conditions.

5. Discussion

In this work, we propose DmsFIQA to address the unmet need for face image quality assessment in driver monitoring system (DMS) scenarios. Our main contribution is a scalable two-stage labeling pipeline that generates reliable face-quality supervision with minimal additional manual effort. We evaluate DmsFIQA from two complementary perspectives: (i) standard FIQA metrics for quality prediction, and (ii) downstream face verification performance under quality-based filtering. Experimental results demonstrate

that DmsFIQA is better suited to DMS data than existing public FIQA approaches, highlighting the importance of domain-specific quality modeling.

References

[1] Takashi Abe. Perclos-based technologies for detecting drowsiness: current evidence and future directions. *Sleep Advances*, 4(1):zpad006, 2023. 1, 3

[2] Žiga Babnik, Peter Peer, and Vitomir Štruc. Faceqan: Face image quality assessment through adversarial noise exploration. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 748–754. IEEE, 2022. 8

[3] Fadi Boutros, Meiling Fang, Marcel Klemt, Biying Fu, and Naser Damer. Cr-fiqa: Face image quality assessment by learning sample relative classifiability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5836–5845, 2023. 2

[4] Iuliia Brishtel, Stephan Krauss, Mahdi Chamseddine, Jason Raphael Rambach, and Didier Stricker. Driving activity recognition using uwb radar and deep neural networks. *Sensors*, 23(2):818, 2023. 1

[5] Wei-Ting Chen, Gurunandan Krishnan, Qiang Gao, Sy-Yen Kuo, Sizhou Ma, and Jian Wang. Dsl-fiqa: Assessing facial image quality via dual-set degradation learning and landmark-guided transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2931–2941, 2024. 2, 8

[6] Khazar Dargahi Nobari and Torsten Bertram. A multimodal driver monitoring benchmark dataset for driver modeling in assisted driving automation. *Scientific data*, 11(1):327, 2024. 3

[7] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 2

[8] Qiushi Guo. Enrich the content of the image using context-aware copy paste. *arXiv preprint arXiv:2407.08151*, 2024. 3

[9] Qiushi Guo. Depth-copy-paste: Multimodal and depth-aware compositing for robust face detection. *arXiv preprint arXiv:2512.11683*, 2025. 2

[10] Javier Hernandez-Ortega, Javier Galbally, Julian Fierrez, Rudolf Haraksim, and Laurent Beslay. Faceqnet: Quality assessment for face recognition based on deep learning. In *Proc. IAPR International Conference on Biometrics (ICB)*, 2019. 2

[11] Qiang Ji and Xiaojie Yang. Real-time eye, gaze, and face pose tracking for monitoring driver vigilance. *Real-time imaging*, 8(5):357–377, 2002. 2

[12] Jigyasa Singh Katrolia, Bruno Mirbach, Ahmed El-Sherif, Hartmut Feld, Jason Rambach, and Didier Stricker. Ticam: A time-of-flight in-car cabin monitoring dataset. *arXiv preprint arXiv:2103.11719*, 2021. 1

[13] Muhammad Qasim Khan and Sukhan Lee. Gaze and eye tracking: Techniques and applications in adas. *Sensors*, 19 (24):5540, 2019. 3

[14] Jan Niklas Kolf, Naser Damer, and Fadi Boutros. Grafiqs: Face image quality assessment using gradient magnitudes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1490–1499, 2024. 8

[15] Brett Koonce. Mobilenetv3. In *Convolutional neural networks with swift for tensorflow: image recognition and dataset categorization*, pages 125–144. Springer, 2021. 7

[16] 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*, pages 19730–19742. PMLR, 2023. 2

[17] Yiming Li, Chen Cai, Tianyi Liu, Dan Lin, Wenqian Wang, Wenfei Liang, Bingbing Li, and Kim-Hui Yap. Daos: A multimodal in-cabin behavior monitoring with driver action-object synergy dataset. *arXiv preprint arXiv:2601.11990*, 2026. 3

[18] Ashutosh Mishra, Sangho Lee, Dohyun Kim, and Shiho Kim. In-cabin monitoring system for autonomous vehicles. *Sensors*, 22(12):4360, 2022. 1, 2

[19] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21 (12):4695–4708, 2012. 2

[20] Negar Moslemi, Mohsen Soryani, and Reza Azmi. Computer vision-based recognition of driver distraction: A review. *Concurrency and Computation: Practice and Experience*, 33(24):e6475, 2021. 1, 3

[21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 2

[22] Torsten Schlett, Christian Rathgeb, Olaf Henniger, Javier Galbally, Julian Fierrez, and Christoph Busch. Face image quality assessment: A literature survey. *ACM Computing Surveys*, 2022. 1, 2

[23] Cheng Shen, Liquan Shen, Mengyao Li, and Meng Yu. Epl-uflsid: Efficient pseudo labels-driven underwater forward-looking sonar images object detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 4349–4357, 2024. 8

[24] Shaolin Su, Hanhe Lin, Vlad Hosu, Oliver Wiedemann, Jinqiu Sun, Yu Zhu, Hantao Liu, Yanning Zhang, and Dietmar Saupe. Going the extra mile in face image quality assessment: A novel database and model. *IEEE Transactions on Multimedia*, 2023. 2

[25] Philipp Terhorst, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Ser-fiq: Unsupervised estimation of face image quality based on stochastic embedding robustness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5651–5660, 2020. 2

[26] Francisco Vicente, Zehua Huang, Xuehan Xiong, Fernando De la Torre, Wende Zhang, and Dan Levi. Driver gaze tracking and eyes off the road detection system. *IEEE Trans-*

actions on Intelligent Transportation Systems, 16(4):2014–2027, 2015. 3