Towards Fully Automated ISO/ICAO Face Compliance Verification via Prompt Learning

Nicolò Di Domenico, Guido Borghi, Annalisa Franco, and Davide Maltoni

***Abstract*—Ensuring that facial images conform to widely adopted quality guidelines is a crucial step in optimizing the document enrollment workflow, which includes the face verification task. In this paper, we focus on the ISO/ICAO standard, which defines the requirements for facial photographs used in official documents, such as passports, ensuring consistency in face quality and thereby improving reliable recognition by both humans and biometric systems. Generally, ISO/ICAO compliance verification is manually performed through a slow, subjective, and non-scalable process, then to address these challenges, we introduce a fully automated system that assesses face compliance directly from the official standard requirements, eliminating dependence on predefined, hand-crafted features and empirically set thresholds. The method integrates a language model with an innovative prompt learning strategy and a contrastive learning paradigm to assess whether a given facial image satisfies specific quality criteria. Experimental evaluations demonstrate that our method achieves competitive accuracy compared to both academic and commercial baselines. By facilitating the integration and maintenance of compliance regulations, the proposed framework offers a practical, scalable, and regulation-centric solution for automated image quality verification. All code and models are publicly available[1].**

***Index Terms*—Face Image Quality, ICAO guidelines, Biometrics, Computer Vision, Deep Learning, Large Language Models.**

I. Introduction

THE standard [1], [2] established by the *International Civil Aviation Organization* (ICAO) and the *International Organization for Standardization* (ISO) is crucial because it guarantees that facial images used in official documents are consistent, high-quality, and suitable for biometric recognition. Indeed, by following these guidelines, countries ensure interoperability between systems worldwide, enhance border security, and reduce identification errors. It also helps maintain fairness and accuracy in automated face recognition processes.

Therefore, verifying that facial photographs meet the quality requirements established by these standards – hereafter jointly referred to as ISO/ICAO – is a fundamental step in the identity verification pipeline for electronic Machine-Readable Travel Documents (eMRTDs) [3], [4].

Specifically, ISO/ICAO standards define rigorous requirements, including aspects such as head orientation, illumination conditions, facial expression, background uniformity, and other photographic and geometric properties (see Fig. 1).

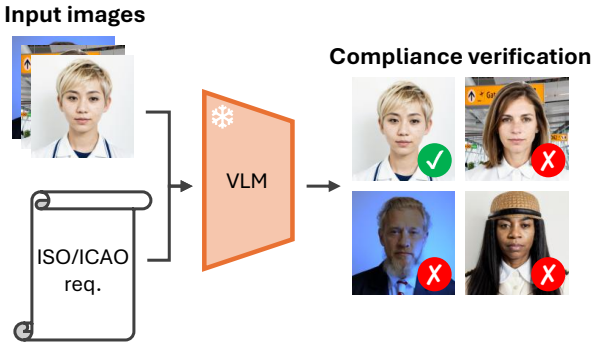


Fig. 1. Starting from the official ISO/ICAO documents [1], [2] on face quality, the proposed method classifies, through a Vision-Language Model (VLM), whether input images are compliant with the given requirement.

Historically, the task of ensuring compliance has been carried out manually by trained personnel who visually inspect each image to verify its conformity with the defined criteria. Nevertheless, this manual process is inherently limited: it is labor-intensive and time-consuming, as it requires experts to review extensive image collections, and it is susceptible to subjective bias, fatigue, and differences in individual judgment [5]. Furthermore, the large number of detailed technical rules involved in ISO/ICAO compliance makes it challenging to ensure uniform and consistent human training across different systems and operational contexts.

In light of these constraints, there is an increasing demand for automated systems capable of performing compliance verification with both efficiency and accuracy [3], [4], as automation alleviates the burden of manual inspection while improving the consistency and reliability of assessments by minimizing human subjectivity and ensuring that standards are applied uniformly. Nonetheless, designing such systems remains challenging: image-based verification relies on advanced Computer Vision and Machine Learning techniques, while compliance requirements, formally specified in the ISO standard and partially implemented in the OFIQ [6] reference framework, require careful technical interpretation for automated deployment.

To address these issues, we introduce a method that aims to fully automate the compliance assessment: indeed, starting from the original ISO/ICAO documentation [2], it extracts the relevant guidelines, and evaluates image adherence to the specified requirements. In contrast to prior solutions [3], [5], [7], which rely on a large number of engineered features,

N. Di Domenico, A. Franco, and D. Maltoni are with the Department of Computer Science and Engineering, University of Bologna, Italy. e-mail: {name.surname}@unibo.it

G. Borghi is with the Department of Education and Humanities, University of Modena and Reggio Emilia, Italy.

Manuscript received April 19, 2005; revised August 26, 2015.

[1]https://github.com/MI-BioLab/CLIP-ICAO-Compliance

manually defined thresholds, or complex algorithmic pipelines – requiring domain-specific expertise – our approach leverages Large Language Models (LLMs) to automatically parse and reason over official standards, enabling the dynamic generation of verification criteria without human intervention. As a result, the proposed system achieves a higher degree of simplicity and adaptability, allowing seamless incorporation of revisions or new regulatory updates without the need for laborious reconfiguration or expert-driven adjustments. It is worth noting that the use of LLMs, a core component of our approach, raises potential concerns regarding the leakage of sensitive biometric data; to mitigate this risk, our system relies on locally hosted LLMs, ensuring full control over the data and preventing any retention of personal information, while the use of third-party hosted LLMs would require strict data usage agreements to guarantee privacy.

From a technical standpoint, our method builds upon a Vision-Language Model (VLM), namely CLIP-IQA [8], derived from the CLIP architecture [9], which effectively captures the semantic correspondence between textual descriptions and visual content. Rather than developing a separate model tailored to each compliance criterion, we directly query the model to assess the degree to which an image aligns with two antonym textual prompts: one expressing conformity and the other describing non-conformity to a specific requirement.

Inspired by genetic algorithms [10] and the recent literature about automated prompt engineering [11], [12], we propose a novel Prompt Learning (PL) procedure that automatically derives prompts using an LLM to parse the official ISO/ICAO documentation and iteratively generate and refine textual formulations for each guideline. The CLIP-IQA framework is then employed to compute the similarity between the input image and the pair of corresponding prompts, thus estimating the level of compliance for each check.

We validate our system using the TONO [5] and BioLab-ICAO [13] datasets, specifically designed for ISO/ICAO compliance verification, and experimental results demonstrate that our approach achieves accuracy that equals or outperforms academic and commercial baselines. These results are remarkable considering that the proposed method is automatic, *i.e.* it does not rely on specific algorithms to verify the compliance of a specific requirement. As a consequence, it is able to deal with new requirements without the intervention of domain experts to implement new specific algorithms.

This paper extends our previous conference work [14], introducing an in-depth analysis of the system, new experiments, images, and explanations, thus improving the comprehension of automatic systems for ICAO face compliance verification. In summary, the key contributions are as follows:

- A system designed to enhance the automation of ISO/ICAO compliance verification. The proposed approach assesses image conformity directly against the official standards, ensuring transparent and reliable evaluation. Moreover, its modular design facilitates future extensions and integrations, significantly reducing the dependence on technical expertise and manual efforts.
- A Prompt Learning (PL) technique with the aim of automatically generating and refining prompts extracted by an LLM directly from the official documents.
- An experimental evaluation on two different benchmarks, which proves the effectiveness of the proposed approach with respect to the solutions available in the market and in the literature.

II. Related works

A. Face Images and ISO/ICAO Compliance

The need to standardize facial image quality has emerged from its critical impact on recognition performance, and foundational studies have identified and formalized the primary factors [15], [16] that influence facial image quality, enabling the development of systematic evaluation methodologies. Ensuring high-quality facial imagery is particularly important in contexts such as eMRTDs, where identity documents remain valid for extended periods, making robustness against aging-related variations a strict requirement.

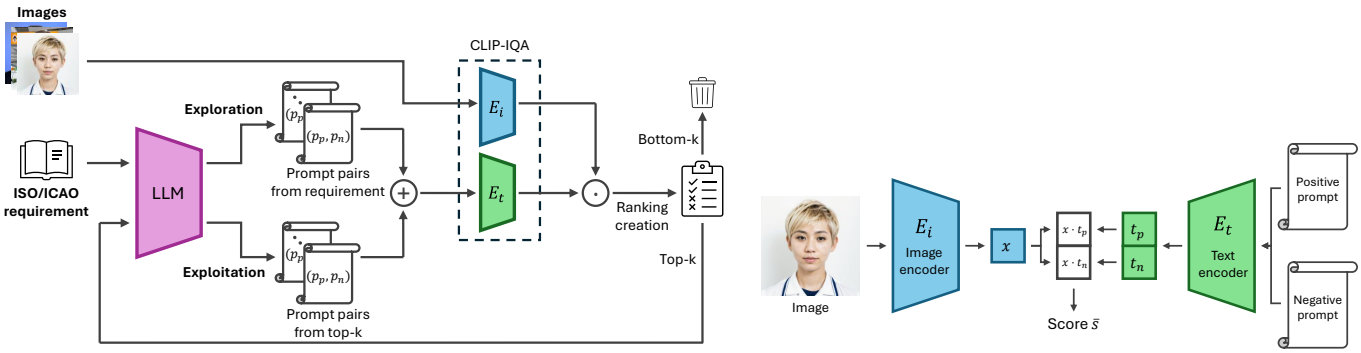
A key milestone in this standardization process is the introduction of ICAO Doc 9303 [16], which defines the functional specifications for eMRTDs and emphasizes the need for uniform portrait quality standards. Standardized facial photographs serve as a crucial link between the document and its rightful owner, supporting both visual verification and automated recognition procedures. Automated Border Control (ABC) systems, in particular, depend on consistent digital facial images to efficiently match printed, stored, and live-captured representations during border processing. The ICAO framework also directly informed the development of ISO/IEC 19794-5 [1], subsequently refined to ISO/IEC 39794-5 [2].

Beyond the scope of ISO/ICAO compliance, the broader task of quantifying facial image quality has motivated additional standardization initiatives. Unlike other biometric modalities, such as fingerprints, facial images historically lacked a unified quality assessment framework. Current efforts under ISO/IEC 29794-5 [17] aim to define a comprehensive methodology to evaluate the quality of facial images in various acquisition scenarios, including those characterized by uncontrolled conditions involving variations in illumination, pose, and other environmental factors. Currently, OFIQ [6] is the reference implementation for this standard.

B. Methods for ISO/ICAO compliance verification

In the last decades, several tools have been developed to assess compliance with ISO/ICAO standards, including one of the earliest research works [3], which introduced a structured set of requirements and corresponding algorithmic methods for verification.

A number of commercial products (here referred to as SDKs) have been introduced for face image compliance verification. Some of these are specifically designed for ISO/ICAO evaluation, while others extend existing facial recognition SDKs by incorporating compliance-checking functionalities. In parallel, several academic initiatives have released open-source resources aimed at supporting research and evaluation in this domain.



(a) Prompt Learning (b) Compliance Verification

Fig. 2. The proposed approach consists of two steps. During the Prompt Learning phase (Fig. 2a), an LLM generates several prompt pairs (each consisting of a positive p_p and negative p_n prompt) based on the ISO/ICAO standard definition for a specific requirement. At each generation, the prompt pool consists of prompts directly derived from the requirement (exploration), and the best performing prompts (Top-k) of the previous generation (exploitation). Then, all prompts are iteratively refined through a Prompt Learning procedure based on CLIP-IQA [8] image (E_i) and text encoders (E_t) for image and prompt analysis. As second step (Fig. 2b), the same encoders E_i and E_t are employed to obtain text (t_p, t_n) and image embeddings (x). The cosine similarity between x and t_p, t_n (compliant and non-compliant prompts) is computed. Finally, the overall compliance score $\bar{s}$ is output through Softmax.

Among the commercial offerings, Correlance[2] designed an SDK to assess conformity with the early ISO/ICAO standards. Innovatrics[3] delivers a more comprehensive suite of functionalities, encompassing passive liveness detection, image quality evaluation, and user-guided feedback during image acquisition. NEUROtechnology's SDK[4] also integrates fundamental image quality assessments, such as evaluating eye distance and verifying frontal head pose. Because these systems are proprietary – typically part of larger facial recognition toolkits – their implementations are not publicly accessible. This lack of transparency restricts their applicability and reproducibility in academic research.

In literature, some research-driven projects are available. A notable contribution is the BioLab-ICAO framework [13], which provides standardized evaluation protocols and benchmarks and reference algorithms for compliance assessment. Nevertheless, public access to training data remains limited, as the majority of the dataset is restricted and primarily available only through the FVC-onGoing platform[5]. Additional advancements in the field include the BioLab-ICAO-Check tool introduced in [3], which integrates specific computer vision algorithms for assessing several image quality factors. Other studies have focused on specific aspects of ISO/ICAO compliance, addressing issues such as head coverings related to religious contexts [18], pixelation and occlusions (*e.g.*, hair over the eyes, veils) [19], eye-related conditions (such as closed eyes, red-eye effects, or incorrect gaze) [20], head pose estimation [21], and the evaluation of attributes like mouth closeness, eye openness, and overall facial visibility [22].

Some works have attempted to include multiple compliance criteria. For example, [23] introduces a system that simultaneously evaluates nine different ISO/ICAO requirements, offering a unified alternative to methods that process each rule independently. More recently, [4] proposed ICAONet, a deep learning-based multitask architecture designed to automate compliance assessment for the ISO/IEC 19794-5 [1] standard. The model comprises three modules: an encoder, adapted from the autoencoder architecture in [24], to generate compact face embeddings; an unsupervised decoder for image reconstruction; and a supervised classification branch tasked with predicting compliance scores. Finally, the method proposed in [7] introduces several specific algorithms built on classic computer vision algorithms or neural networks, based on manual thresholds, to produce the final scores.

In conclusion, a common challenge lies in the heterogeneous coverage of ISO/ICAO requirements, as although all methods adhere to the same guidelines, each implementation targets a distinct subset of checks, making comparisons between systems difficult due to variations in the interpretation of the guidelines. In addition, we observe a lack of standardized mappings between features and compliance conditions.

III. Proposed Method

An overview of the proposed framework is shown in Figure 2. The system relies on two sequential phases.

In the first phase, a novel Prompt Learning (PL) procedure (see Sect. III-A) employs an LLM to parse the official ISO/ICAO documentation and to automatically generate a set of antonym prompt pairs corresponding to each compliance requirement. These prompts are iteratively refined to ensure that they accurately capture compliance and non-compliance with the given requirement.

Once the positive and negative prompts for each requirement are defined, the inference phase evaluates facial images against these textual representations using CLIP-IQA [8], a variant of CLIP [9] specifically adapted for Image Quality Assessment (IQA) tasks (see Sect. III-B). As both visual and textual features are embedded within a shared latent space, compliance is determined by computing the cosine similarity between the image embedding and the corresponding positive and negative text embeddings produced by the vision-language model. The system subsequently derives a normalized compliance score within the range $[0, 1]$, indicating the likelihood that the input image fulfills the specified ISO/ICAO requirement.

[2] https://www.correlance.com/cms/en/ccEngineICAO
[3] https://www.innovatrics.com/digital-onboarding-toolkit/face-matching
[4] https://www.neurotechnology.com/face-verification-technical.html
[5] https://biolab.csr.unibo.it/FvcOnGoing

TABLE I
PERFORMANCE COMPARISON MEASURED IN EER ON THE TONO DATASET. LOWER EER VALUES INDICATE BETTER PERFORMANCE. BEST RESULTS ARE IN BOLD, SECOND BEST ARE UNDERLINED. NOTE THAT MEAN VALUES ARE COMPUTED ON AVAILABLE RESULTS.

		Academic Solutions				Commercial SDKs		
Requirements		**ICAONet** [4]	**BioLab** [13]	**BioGaze** [7]	**OFIQ** [17]	**Correlance**	**Innovatrics**	**Ours**
Subject	Head w/o coverings	0.179	0.282	0.003	**0.000**	0.030	-	0.010
	Gaze in camera	0.456	0.515	0.504	-	**0.177**	0.277	0.486
	Eyes open	0.022	0.500	0.019	**0.000**	**0.000**	**0.000**	0.007
	No/light makeup	-	-	0.008	-	-	-	**0.000**
	Neutral expression	0.321	0.256	0.093	**0.045**	0.097	0.110	0.051
	No sunglasses	0.035	0.198	0.022	**0.000**	**0.000**	0.122	**0.000**
	Frontal pose	0.206	0.401	**0.074**	0.154	0.085	0.560	0.210
	Mean	0.203	0.359	0.103	0.040	0.065	0.214	0.109
Photogr.	Correct exposure	0.343	0.199	**0.096**	0.301	0.301	0.262	0.212
	In focus photo	0.353	0.003	**0.000**	**0.000**	0.006	0.076	0.026
	Correct saturation	0.330	0.058	**0.042**	0.135	0.458	-	0.103
	Mean	0.342	0.087	0.046	0.145	0.255	0.169	0.114
Acquisition	Uniform background	0.394	0.386	0.121	0.290	0.114	-	**0.071**
	Uniform face lighting	0.365	0.420	**0.083**	0.109	0.084	0.286	0.141
	No pixelation	0.490	0.330	**0.000**	-	-	-	0.006
	No posterization	-	-	0.010	-	0.241	0.500	**0.006**
	Mean	0.416	0.379	0.054	0.200	0.146	0.393	0.056
Global Mean ↓		0.291	0.296	0.077	0.103	0.133	0.243	0.095

A. Prompt Learning (PL)

The formulation of prompts plays a fundamental role in determining the overall performance of vision-language models [8], [9], making it essential to craft them accurately. Taking inspiration from [12], we automate the prompt generation process through the use of an LLM, which is instructed to produce antonym prompt pairs directly from the natural language descriptions of ISO/ICAO requirements.

To this end, we introduce a Prompt Learning (PL) algorithm inspired by the evolutionary mechanisms of genetic algorithms [10]. More specifically, the language model iteratively generates pairs of positive p_p (compliant) and negative p_n (non-compliant) prompts from both the requirement r and the best-performing pairs from the previous generation. The fitness of each prompt pair is quantitatively assessed using the CLIP-IQA framework. At each generation g, two complementary prompt creation strategies are employed, conceptually corresponding to the *exploration* and *exploitation* phases typical of genetic algorithms [10]. The first derives prompts directly from the raw requirement, aiming to condense it into antonym pairs of at most 15 words, therefore respecting the 77-token constraint of CLIP's text encoder. The second strategy focuses on refining the best-performing prompt pairs (top-k) while avoiding elements from the worst-performing (bottom-k) ones. To minimize hallucinations, the LLM receives as input both the original requirement and the selected prompt sets. To balance these processes, each generation produces a number of new prompt pairs, in which half are derived directly from the requirement (P_{new}) and the other half are based on the top-k prompts (P_{child}), while avoiding terms and patterns that may be present in the bottom-k. At the time of learning, the fitness of each generated prompt pair is evaluated using the Equal Error Rate metric (see Sect. IV). This metric is computed from the compliance score obtained via the scoring function described in Sect. III-B. A pseudocode representation of the complete procedure is presented in Figure 3, all prompt pairs

Require: Image encoder $E_i(\cdot)$, text encoder $E_t(\cdot)$
Require: Number of generations n, prompt pairs per generation m, number of images N
Require: Encoded image set $X = (x_1, \ldots, x_N)$, where $x_i \in \mathbb{R}^d$
Require: Ground truth labels $Y = (y_1, \ldots, y_N)$, where $y_i \in \{0, 1\}$
Require: Top-/bottom-k prompt pairs to retain per generation
Initialize prompt pair set $P = \{(p_p^{(j)}, p_n^{(j)})\}_{j=1}^{m}$ from requirement r
for $g = 1$ to n **do**
  for $(p_p, p_n) \in P$ **do**
    Encode prompts: $t_p \leftarrow E_t(p_p)$, $t_n \leftarrow E_t(p_n)$
    $\hat{Y} \leftarrow \text{Softmax}(x \cdot t_p, x \cdot t_n),\ \forall x \in X$
    Compute EER: $e \leftarrow \text{EER}(\hat{Y}, Y)$
    Store (p_p, p_n, e)
  end for
  Sort prompt pairs in P by ascending EER
  $P_{\text{top}} \leftarrow$ top-k pairs with lowest EER
  $P_{\text{bottom}} \leftarrow$ bottom-k pairs with highest EER
  $P_{new} \leftarrow$ Generate $\lfloor m/2 \rfloor$ new prompt pairs from requirement r
  $P_{child} \leftarrow$ Generate $\lceil m/2 \rceil - |P_{top}|$ offspring pairs from P_{top} and P_{bottom}
  $P \leftarrow P_{new} \cup P_{child} \cup P_{top}$
end for

Fig. 3. Pseudo-code of the Prompt Learning (PL) procedure used to optimize performance for a given requirement.

TABLE II
PERFORMANCE COMPARISON MEASURED IN EER ON BIOLAB-ICAO DATASET. LOWER EER VALUES INDICATE BETTER PERFORMANCE. BEST RESULTS ARE IN BOLD, SECOND BEST ARE UNDERLINED. NOTE THAT MEAN VALUES ARE COMPUTED ON AVAILABLE RESULTS.

		Academic Solutions				Commercial SDKs		
Requirements		**ICAONet**† [4]	**BioLab**† [13]	**BioGaze**† [7]	**OFIQ** [17]	**Correlance**	**Innovatrics**	**Ours**
Subject	Head w/o coverings	0.101	0.271	0.427	**0.012**	0.414	-	0.112
	Gaze in camera	0.208	0.266	0.188	-	0.208	0.207	**0.071**
	Eyes open	0.064	0.229	0.103	**0.001**	0.071	0.006	0.069
	Neutral expression	0.110	0.119	0.183	0.058	0.055	**0.006**	0.070
	No sunglasses	**0.000**	0.006	0.192	0.032	0.293	0.083	**0.000**
	Frontal pose	0.508	0.128	0.502	0.260	0.042	0.500	**0.000**
	Mean	0.165	0.170	0.266	0.073	0.181	0.160	0.054
Photogr.	Correct exposure	**0.001**	0.008	0.234	0.467	0.182	0.104	0.098
	In focus photo	0.015	0.080	0.245	0.072	0.123	0.010	**0.006**
	Correct saturation	0.110	**0.058**	0.526	0.186	0.208	-	0.125
	Mean	0.042	0.049	0.335	0.242	0.171	0.057	0.076
Acquisition	Uniform background	**0.104**	0.112	0.260	0.241	0.159	0.525	0.139
	Uniform face lighting	0.237	0.226	0.213	0.348	0.272	0.244	**0.149**
	No pixelation	0.208	0.004	0.066	-	-	-	**0.000**
	Mean	0.183	0.114	0.180	0.295	0.216	0.385	0.096
Global Mean ↓		0.139	0.126	0.262	0.168	0.184	0.187	0.070

are included in the supplementary material.

It is worth noting that the proposed prompt learning process operates in a fully automated manner, relying solely on the textual descriptions of official requirements. This design enables rapid adaptation to new or updated standards with minimal human involvement, as both prompt generation and evaluation are driven entirely by the data and model mechanisms.

B. ISO/ICAO Compliance Verification

In the second phase of the pipeline, compliance with each ISO/ICAO requirement is evaluated by measuring the correspondence between an input facial image and a pair of antonym textual prompts: a positive prompt (p_p), describing a compliant image, and a negative prompt (p_n), describing a non-compliant image. For instance, the requirement specifying "Head without coverings" may correspond to a positive prompt such as "The subject is not wearing any type of headgear, and the hair is visible", and a negative prompt such as "The subject is wearing a hat, cap, bandana, or any other garment that conceals the hair". The use of antonym prompt pairs mitigates the effects of linguistic ambiguity and enhances discriminative power, as similarity to a single prompt describing compliance may be insufficient for a reliable evaluation [8].

Each prompt is encoded through the CLIP-IQA text encoder $E_t(\cdot)$ to obtain fixed-dimensional embeddings $t_p, t_n \in \mathbb{R}^d$, such that $t_p = E_t(p_p)$ and $t_n = E_t(p_n)$. The input image I is similarly encoded via the image encoder $E_i(\cdot)$, producing an embedding $x \in \mathbb{R}^d$ in the same latent space. The degree of compliance is then evaluated by computing the cosine similarity between the image embedding and each of the text embeddings. Finally, we transform these two raw similarities into a probabilistic compliance score $\overline{s} \in [0, 1]$ by applying the Softmax function and taking the value corresponding to the similarity with the positive prompt t_p.

TABLE III
INVESTIGATION ON DIFFERENT STRATEGIES FOR PROMPT LEARNING (PL): HANDCRAFTED PROMPTS ("MANUAL"), PL ON HANDCRAFTED PROMPTS (PL ← MAN.), AND PL WITH REQUIREMENTS SOURCED FROM THE OFFICIAL ISO/ICAO DOCUMENTS [15] (PL ← REQ.).

Requirements		**Manual**	**PL ← Man.**	**PL ← Req.**
Subject	Head w/o coverings	0.056	**0.007**	0.010
	Gaze in camera	**0.469**	0.483	0.486
	Eyes open	0.111	**0.000**	0.007
	No/light makeup	0.218	0.008	**0.000**
	Neutral expression	0.599	0.179	**0.051**
	No sunglasses	0.125	0.007	**0.000**
	Frontal pose	0.595	**0.145**	0.210
	Mean	0.310	0.119	**0.109**
Photogr.	Correct exposure	0.481	**0.074**	0.212
	In-focus photo	0.045	**0.000**	0.026
	Correct saturation	0.484	**0.064**	0.103
	Mean	0.337	**0.046**	0.113
Acquisition	Uniform background	0.442	**0.071**	**0.071**
	Uniform face lighting	0.308	**0.128**	0.141
	No pixelation	0.567	**0.006**	**0.006**
	No posterization	0.048	**0.006**	**0.006**
	Mean	0.341	**0.053**	0.056
Global Mean		0.325	**0.084**	0.095

It is important to note that, as each compliance check is expressed in natural language, the system offers enhanced interpretability for human examiners, allowing a more intuitive analysis of the specific elements evaluated for compliance verification. Indeed, heatmap visualization techniques such as Grad-CAM [25] can further support this process.

IV. EXPERIMENTS

A. Datasets

For the investigated task, the availability of data is particularly limited: therefore, it is important to report, for each

TABLE IV
NUMBER OF TOTAL IMAGES AVAILABLE FOR EACH REQUIREMENT, FOR EACH OF THE TWO EMPLOYED DATASETS.

Requirements		TONO	BioLab
	Fully compliant images	311	50
Subject	Head w/o coverings	288	105
	Gaze in camera	611	32
	Eyes open	240	31
	No/light makeup	197	-
	Neutral expression	309	115
	No sunglasses	245	31
	Frontal pose	136	4
Photo.	Correct exposure	311	38
	In-focus photo	311	31
	Correct saturation	311	37
Acquis.	Uniform background	269	39
	Uniform face lighting	145	66
	No pixelation	311	30
	No posterization	311	-

requirement, the number of available images (Table IV).

TONO. [5] This is a collection of synthetic face images specifically designed for ISO/ICAO compliance testing. The dataset comprises approximately 4k images, each containing a single feature that violates ISO/ICAO guidelines. This design enables controlled experiments by allowing each compliance check to be assessed independently. We employ a 50–50 random split between training and test subsets, ensuring balanced representation across both gender and ethnicity.

Biolab-ICAO. The original dataset is sequestered and therefore available only through the FVC-onGoing platform. Unfortunately, technical and execution time constraints preclude the use of our method on that platform. Therefore, following [4], we use as a test set the public training subset: through the publicly released tool, we obtain annotations for 571 images, including 50 fully compliant images and 521 non-compliant ones. Each image may be non-compliant for more than one requirement and is used in multiple tests. Due to the scarcity of fully compliant images, we generate two sets (compliant and non-compliant images) for every requirement on which we employ a $50-50$ random split for training and test sets.

B. Metrics

Performance is reported using the Equal Error Rate (EER) [26], defined as the intersection point of the False Acceptance Rate (FAR) and False Rejection Rate (FRR) curves. Then, lower EER values correspond to more effective prompt pairs and improved compliance evaluation capabilities. To compute these metrics, the TONO dataset is augmented with ISO/ICAO-compliant synthetic images drawn from the ONOT dataset [27], providing both positive and negative samples.

For clarity, the compliance checks are organized into three categories: subject (requirements that directly depend on the acquired person), photographic (requirements related to the proper use of the acquisition device), and acquisition (related to the environmental conditions and possible scanning devices used for the face image acquisition) requirements [5]. For each category, we report the mean EER across all implemented checks, alongside a global mean EER to provide an overall summary of each method's performance. It is important to note that these metrics consider only the subset of checks implemented by each method, and thus may not be directly comparable across approaches that evaluate different sets of requirements.

C. System configuration

For the comparison, we adopt the best configuration for our system, as investigated in the following sections.

For the PL procedure (see Sect. V-A), we repeat for 50 generations the process; in each step 100 new prompt pairs are generated, 50 derived directly from the requirement and 50 based on the top-20 prompts, excluding terms common to the bottom-20. Finally, we select the prompt pair with the lowest EER across all the generated ones. For the LLM, we adopt Phi-4 [28] (see Sect. V-C for comparison), which, with its size of 14 billion parameters, provides a good balance between inference capabilities and computational efficiency, enabling local execution without the need for large-scale infrastructure.

TABLE V
GLOBAL EER VARYING THE PARAMETERS OF THE PROMPT LEARNING PROCEDURE SPECIFIED IN FIGURE 3.

#Gen. (n)	**#Prompts** (m)	**Top/bottom** (k)	**EER**	**TTB**
5	10	3	0.162	**19** s
10	20	5	0.124	123 s
20	50	10	0.095	530 s
50	100	20	0.084	1507 s
100	200	50	**0.072**	5553 s

D. Comparison against other methods

The results are reported for each requirement, and for each group of requirements, we have provided the average. However, it is important to note that the average does not take into account that some methods do not implement all the requirements; therefore, this value should be considered only indicative.

Table I summarizes the comparative evaluation on the TONO dataset. From a general point of view, the proposed method achieves a high accuracy, overcoming most of the competitors. It is worth noting that our approach constitutes the only method that implements a fully automated compliance verification procedure. The comparative evaluation indicates that, although specialized methods with manually designed checks can achieve the highest performance in specific requirements, our automated framework, driven by prompts directly derived from the ISO/ICAO specifications, produces competitive results while substantially reducing the implementation time required for comprehensive compliance verification, compared to expert-driven solutions, and demonstrates unprecedented generalization capabilities.

Despite the competitive error rates in the majority of checks, some requirements are particularly challenging for our method's chosen image encoder, *i.e.* CLIP-IQA [8] (*e.g.* "gaze in camera" requirement). We hypothesize that this arises

TABLE VI
MEAN EER COMPARISON OF CLIP MODELS ACROSS SUBJECT, PHOTOGRAPHIC, AND ACQUISITION CATEGORIES.

Model	Image Encoder	Subject Mean	Phot. Mean	Acq. Mean	Global Mean
CLIP-IQA [8]	FCN	**0.313**	**0.342**	**0.353**	**0.331**
CLIP [9]	ViT-B/16_{224}	0.305	0.435	0.386	0.356
	ViT-B/32_{224}	0.295	0.413	0.372	0.342
	ViT-L/14_{224}	**0.268**	0.370	0.371	0.319
	ViT-L/14_{336}	0.272	**0.322**	**0.337**	**0.302**
OpenCLIP [6]	ViT-B/16_{224}	0.348	**0.306**	0.403	0.355
	ViT-B/32_{224}	0.334	0.315	0.435	0.359
	ViT-L/14_{224}	**0.283**	0.382	0.414	0.342
	ViT-H/14_{224}	0.354	0.411	0.382	0.374
	ViT-G/14_{224}	0.304	0.322	**0.375**	**0.328**
SigLIP [29]	ViT-B/16_{224}	0.270	**0.360**	0.378	0.320
	ViT-B/16_{256}	0.278	0.368	0.357	0.320
	ViT-B/16_{384}	0.277	0.375	0.379	0.327
	ViT-B/16_{512}	0.282	0.395	0.351	0.326
	ViT-L/16_{256}	0.250	0.370	0.361	0.307
	ViT-L/16_{384}	**0.249**	0.381	**0.334**	**0.302**

because pupils represent a relatively small portion of the image and are susceptible to variations from external factors.

Furthermore, to demonstrate the robustness of the systems on real face images, we compare methods on the BioLab-ICAO [13] dataset; as this dataset does not contain images covering the "No/light makeup" and "no posterization" requirement, we do not include these requirements in the evaluation. Results are reported in Table II, and confirm the great accuracy of our method that overcomes all the competitors. The results for some SDKs and some requirements are sub-optimal; this is partly due to limitations of the SDKs themselves and partly because there is no uniformity in the requirements assessed by the different SDKs. The results shown in the table refer, for each system tested, to the best match we were able to find among the requirements evaluated by the SDKs, but the match may not be perfect. For example, BioGaze [7] evaluates only the presence of hat/cap, whereas the requirement assessed in this test includes other accessories (*e.g.* scarf) that partially cover the face. Additionally, it should be noted that some methods, namely ICAONet [4], BioLab-ICAO-Check [3], and BioGaze [7], demonstrate better performance on the BioLab-ICAO dataset; we hypothesize that this is due to the fact that the underlying checks have been trained on that specific dataset, as described in the original respective papers.

Finally, to assess the visual explainability of the proposed system, we extract Grad-CAM [25] heatmaps of non-compliant images. As depicted in Figure 4, in which we report both the requirement and the negative prompt – the most relevant to check the compliance – our method is able to produce focused heatmaps on both datasets, showing regions of interest specifically related to possible locations where non-compliance is found.

V. FURTHER ANALYSIS

A. Investigation on Prompt Learning

We further analyze one of the key elements of the proposed method, *i.e.* the Prompt Learning procedure. In particular, we measure the impact of starting the learning from the official documentation or from more concise prompts defined by a human domain expert; besides, we assess the performance of the system without the PL process, *i.e.* using only the manually defined prompts without any learning procedure.

As shown in Table III, the use of prompt learning substantially improves performance across all compliance checks with respect to manual prompting, achieving reductions in EER of -72.5%. Interestingly, using full requirement descriptions to initialize the prompt learning algorithm does not consistently lead to superior performance compared to handcrafted prompts. Furthermore, certain checks, such as "frontal pose" and "correct exposure", exhibit notably lower performance when prompts are learned directly from the requirement description rather than from hand-crafted seeds. This discrepancy likely arises because the PL algorithm may struggle to identify the most salient aspects of the ISO/ICAO requirements, whereas a human-condensed version effectively directs the LLM and minimizes potential hallucinations. Examination of the learned prompts supports this explanation: for example, in the "uniform face lighting" check, the prompt pair obtained by refining the hand-crafted prompts reads "Face fully visible, unobstructed" versus "Face partially visible, lens shadows", whereas the prompts generated from the ISO/ICAO description emphasize secondary aspects, stating "Distinct forehead lines" versus "Forehead lines lost in lighting".

However, despite the overall superior effectiveness of prompts learned from handcrafted seeds, the performance degradation observed when relying on documents is minimal. This indicates that a fully automated pipeline, which eliminates the need for human expertise during the development of the system, remains a viable and effective option.

Furthermore, we investigate the impact of different parameters in the PL process. In particular, we vary the number of generations, the number of prompts per generation, and the number of best- and worst-performing prompts as context for the exploitation phase. Results are summarized in Table V. In general, increasing the value of parameters leads to improved performance, since the LLM has more time and context to refine the optimal prompt pairs. However, when evaluating the time taken to find the best overall prompt pair (reported as Time To Best, TTB), we observe that it increases at a faster rate than the reduction observed in the global mean EER. These findings indicate that it is important to strike the correct balance between the effectiveness of the learned prompts and the computational cost of the optimization process.

B. Investigation on Visual Encoders

We investigate the impact of different image-text encoder: we compare CLIP-IQA [8], various versions of CLIP [9], OpenCLIP [6], and SigLIP [29]. These models, with the exception of CLIP-IQA, all leverage Vision Transformers (ViTs) [30] as image encoders and operate with diverse image resolutions, spanning from 224 to 512. To maintain consistency across experiments, all models are evaluated using the same set of prompt pairs for each requirement. These prompts are handcrafted by an expert and are used without any

(a) Head w/o coverings
Neg. prompt: *"Eyes and brows buried by headwear"*

(b) Eyes open
Neg. prompt: *"Irises shaded, eyes closed"*

(c) Neutral expression
Neg. prompt: *"Big teeth showing, wide smile"*

(d) Uniform background
Neg. prompt: *"Subject blending into background"*

(e) Uniform face lighting
Neg. prompt: *"Forehead lines lost in lighting"*

(f) Head w/o coverings
Neg. prompt: *"Nose or eyes covered by scarf"*

(g) Eyes open
Neg. prompt: *"Eyes closed or unnaturally wide open"*

(h) Neutral expression
Neg. prompt: *"Broad, toothy smile"*

(i) Uniform background
Neg. prompt: *"Subject merges into background"*

(j) Frontal pose
Neg. prompt: *"Chin partly obscured"*

Fig. 4. Heatmaps generated with Grad-CAM [25] on four distinct requirements, on both the TONO (top row) and BioLab-ICAO (bottom row) datasets.

TABLE VII
COMPARISON OF MODEL ARCHITECTURES AFTER OPTIMIZATION OF THE ORIGINAL HAND-CRAFTED PROMPT PAIRS VIA PROMPT LEARNING. ADDITIONALLY, THE GLOBAL MEAN EER BEFORE PROMPT LEARNING, AS WELL AS THE ERROR RATE REDUCTION, IS REPORTED FOR EACH MODEL.

Requirements	**CLIP-IQA**	**CLIP**	**OpenCLIP**	**SigLIP**
	Before Prompt Learning			
Subject	0.313	0.272	0.304	0.249
Photographic	0.342	0.322	0.322	0.381
Acquisition	0.353	0.337	0.375	0.334
Global Mean	0.331	0.302	0.328	0.302
	After Prompt Learning			
Subject	0.161	0.136	0.164	**0.104**
Photographic	**0.111**	0.228	0.183	0.285
Acquisition	**0.071**	0.127	0.258	0.221
Global Mean	**0.124**	0.153	0.195	0.176
Rel. EER Reduction	**-62.5%**	-58.9%	-53.4%	-41.7%

TABLE VIII
MEAN EER COMPARISON OF CLIP-IQA WITH PROMPTS LEARNED FROM HAND-CRAFTED PAIRS EMPLOYING DIFFERENT LLMS, VARYING BOTH ARCHITECTURE AND PARAMETER COUNT.

Metric	**Llama 3.2**	**Gemma 3**	**Qwen 2.5**	**Phi-4**
#Parameters	3B	12B	14B	14B
Quantization	8-bit	4-bit	4-bit	4-bit
Time To Best (s)	928	1281	533	533
Subject Mean	0.136	0.156	0.124	**0.121**
Photographic Mean	0.073	0.090	**0.051**	**0.051**
Acquisition Mean	0.048	0.070	0.052	**0.046**
Global Mean	0.097	0.117	0.088	**0.084**

prompt learning procedure. Evaluation metrics are computed over the entire TONO dataset.

Results reported in Table VI suggest that, regardless of the training dataset, loss function, or architecture, larger models tend to deliver enhanced performance. Contrarily, increasing the input image size does not necessarily lead to performance improvements: this is particularly evident when employing SigLIP in its ViT-B/16 configuration, where increasing the input image size results in a deterioration of performance. Experimental results, fully reported in the supplementary, indicate that some requirements based on fine visual details. In particular, for "gaze in camera" OpenAI CLIP shows a reduction in EERs from 0.481 in the ViT-B/16 variant to 0.340 in the ViT-L/14@336 variant. For "eyes open" requirement, SigLIP shows improvements from 0.221 in the ViT-B/16 variant to 0.086 in the ViT-L/16@384 variant.

Conversely, requirements involving coarser features such as head coverings, clearly visible makeup, and background consistency, tend to perform better with larger patch sizes; this trend is exemplified by the performance difference between OpenAI CLIP ViT-B/16 and its ViT-B/32 counterpart, where the EER for the "head without coverings" requirement improves from 0.112 to 0.028.

Additionally, the comparison between CLIP-IQA all other models provides insight into the role of positional embeddings. CLIP-IQA exhibits weaker performance on tasks requiring attention to small, localized regions of the image, such as "gaze in camera" and "eyes open", with EERs respectively of 0.489 and 0.096, compared to 0.340 and 0.025 obtained with OpenAI CLIP ViT-L/14@336. In contrast, CLIP-IQA remains competitive on more global attributes like "in-focus photo", "no posterization", and "uniform background", with EERs respectively of 0.042, 0.045, and 0.450, compared to 0.215, 0.342, and 0.459 achieved by OpenAI CLIP ViT-L/14@336. This observation suggests that positional embeddings may be particularly important for fine-grained assessments where

accurate localization of facial features is critical, such as with eyes, makeup, or other specific elements.

Another key factor is the potential for performance improvement through the use of learned prompts. To investigate this aspect, we select the model that performs best in each architecture, *i.e.* CLIP-IQA, OpenAI CLIP ViT-L/14@336, OpenCLIP ViT-G/14, and SigLIP ViT-L/16@384. To avoid excessively long training time, we then optimize the handcrafted prompt pairs using 10 generations of prompt learning, with 20 prompt pairs generated per iteration. As shown in Table VII, CLIP-IQA achieves the lowest global mean EER after prompt learning and exhibits the largest relative error rate reduction, measured as the difference in EER values after and before the Prompt Learning procedure, compared to the initial error rate.

C. Investigation on Large Language Models

We evaluate the capabilities of four distinct open LLMs with varying architectures and parameter counts to extract, synthesize, and optimize prompt pairs. To ensure computational efficiency and avoid excessively long training cycles, we conduct 20 generations of prompt learning, with 50 prompt pairs generated per iteration. The model employed for inference is CLIP-IQA, previously identified as having the greatest potential for improvement following the application of the prompt learning procedure.

Given the vast landscape of available LLM architectures and model sizes, we made a selection of models that have different architectures, that are suitable for our hardware (a single Nvidia RTX A6000) and that lead us to complete the prompt learning procedure in a reasonable time. Since running an LLM at half-precision would be prohibitively expensive in terms of speed, computation, and memory usage, we rely on quantized versions of the original model weights. All selected models are compressed using 4-bit group quantization, with the exception of Llama 3.2 3B which, due to its smaller parameter count, can be efficiently quantized using 8-bit global quantization. Experimental results are summarized in Table VIII. As depicted, the best overall results are obtained with Phi-4 [28], closely followed by Qwen 2.5 [31]. It should be noted that Llama 3.2, despite its relatively small number of parameters, is able to obtain competitive results, with a global mean EER of only 1.3% greater than the best overall result.

D. Investigation on Training Set Size

We assess the robustness of the proposed Prompt Learning under decreasing amounts of training data. Specifically, we apply our algorithm to all requirements of the TONO dataset, using for training a fixed amount of non-compliant images (specifically, 10, 20, and 40). As for compliant images, we sample as many images as the non-compliant ones to create a balanced training set for each requirement. All runs are evaluated on the same test dataset as all other experiments, ensuring that there are no images in common between training and test sets. Experimental results are reported in Table IX. As expected, prompt quality increases as the number of training samples grows, yet noticeable gains already appear even when only a small set of training images per requirement is available.

TABLE IX
COMPARISON OF EER OBTAINED BY RUNNING THE PROMPT LEARNING PROCEDURE WITH INCREASING NUMBER OF NON-COMPLIANT IMAGES FOR TRAINING EACH REQUIREMENT.

	Requirements	10 images	20 images	40 images
Subject	Head w/o coverings	0.152	0.076	**0.013**
	Gaze in camera	0.485	0.486	**0.464**
	Eyes open	0.247	**0.047**	0.050
	No/light makeup	0.028	**0.016**	0.028
	Neutral expression	0.205	0.122	**0.115**
	No sunglasses	**0.018**	**0.018**	**0.018**
	Frontal pose	0.419	0.352	**0.213**
	Mean	0.222	0.160	**0.129**
Photogr.	Correct exposure	0.125	0.125	**0.115**
	In-focus photo	0.048	0.048	**0.026**
	Correct saturation	0.215	0.218	**0.119**
	Mean	0.129	0.130	**0.087**
Acquisition	Uniform background	0.197	0.122	**0.109**
	Uniform face lighting	**0.176**	0.179	0.192
	No pixelation	0.115	**0.019**	0.032
	No posterization	0.067	0.061	**0.013**
	Mean	0.139	0.096	**0.087**
Global Mean		0.178	0.135	**0.108**

E. Investigation on Cross-Dataset Performance

We evaluate the cross-dataset generalization capabilities of our proposed system, focusing on the requirements that are represented in both the TONO and BioLab-ICAO datasets. Specifically, we compare the performance achieved using handcrafted prompts on each dataset with the performance obtained when prompts learned on one dataset are applied and tested on the other. Experimental results are reported in Table X. Here, columns referred to with letters T and B denote the two baselines, evaluated on the TONO and BioLab-ICAO datasets, respectively. In contrast, B $\rightarrow$ T and T $\rightarrow$ B indicate cross-dataset evaluations, where prompts trained on BioLab-ICAO are tested on TONO and prompts trained on TONO are tested on BioLab-ICAO, respectively.

From a general perspective, we observe that the Prompt Learning procedure is still able to improve performance compared to starting from hand-crafted requirements. Confirming the difficulties of cross-dataset experiments, a general degradation in performance is observed compared to the optimal case in which Prompt Learning is performed on the same dataset (Tables I, II). More specifically, the most critical cases concern the requirements "Head w/o coverings", "in-focus", and "uniform face lighting", where a visual inspection of the images reveals substantial differences between the two datasets. Where greater compatibility exists (*e.g.*, "No Sunglasses," "Neutral Expression," and "Pixelation"), more encouraging results can be observed. Finally, some requirements show particularly high error rates (*e.g.*, "Frontal Pose"), partly due to the strongly limited number of test images.

In summary, the cross-dataset experiments reveal the possibility of achieving good generalization, provided that a sufficient number of training images is available and, above all, that they are sufficiently variable and representative of the requirement under consideration.

TABLE X
COMPARISON OF EER OBTAINED WITH MANUAL PROMPTS, AS WELL AS OBTAINED CROSS-DATASET PROMPT LEARNING. TONO AND BIOLAB-ICAO DATASETS ARE DEPICTED AS T AND B, RESPECTIVELY.

Requirements		T	B → T	B	T → B
Subject	Head w/o coverings	0.056	0.219	0.374	0.231
	Gaze in camera	0.469	0.496	0.470	0.202
	Eyes open	0.111	0.000	0.424	0.497
	Neutral expression	0.599	0.147	0.556	0.105
	No sunglasses	0.125	0.000	0.092	0.000
	Frontal pose	0.595	0.436	0.500	0.500
	Mean	0.326	0.216	0.403	0.256
Photogr.	Correct exposure	0.481	0.465	0.659	0.621
	In-focus photo	0.045	0.353	0.421	0.181
	Correct saturation	0.484	0.250	0.324	0.264
	Mean	0.337	0.356	0.468	0.355
Acquisit.	Uniform background	0.442	0.380	0.615	0.453
	Uniform face lighting	0.308	0.545	0.437	0.670
	No pixelation	0.567	0.205	0.500	0.074
	Mean	0.439	0.377	0.517	0.399
Global Mean		0.357	0.291	0.448	0.316

TABLE XI
PERFORMANCE COMPARISON ACROSS DIFFERENT ETHNICITIES. FOR EACH GROUP OF REQUIREMENTS, WE REPORT THE MEAN ACCURACY AND THE GLOBAL MEAN ACCURACY, USING A THRESHOLD COMPUTED ON THE ENTIRE TRAINING SET OR FOR EACH ETHNICITY.

Requirements	EAF	EAS	EEA	EIA	EME
	Global Threshold				
Subject	87.4%	88.3%	89.9%	86.9%	89.5%
Photographic	82.5%	87.2%	90.9%	81.7%	87.4%
Acquisition	91.0%	94.5%	94.9%	93.9%	94.0%
Global Mean	87.4%	89.8%	91.6%	87.8%	90.3%
	Specific Threshold				
Subject	90.2%	87.2%	90.1%	83.4%	88.1%
Photographic	85.0%	89.7%	90.1%	83.3%	85.1%
Acquisition	93.3%	92.6%	94.8%	88.4%	92.7%
Global Mean	90.0%	89.3%	91.4%	84.8%	88.8%

VI. ETHICAL AND SOCIAL IMPACT ANALYSIS

Our method has been evaluated using two datasets, one of which contains synthetic images, therefore eliminating the need for new data acquisition involving real individuals and personally identifiable information. Besides, TONO encompasses a diverse set of samples across multiple ethnicities – African (EAF), East-Asian (EAS), European/American (EEA), Indian-Asian (EIA), and Middle Eastern (EME) – as well as a balanced representation of gender and age groups, which helps mitigate potential biases in evaluation [5]. Nonetheless, we acknowledge that the use of synthetic data generated by models potentially trained on large-scale datasets raises ethical, legal, and privacy considerations. We also recognize that the reliance on pre-trained components, such as the text and image encoders of CLIP-IQA, introduces the risk of perpetuating biases present in the original training data.

To investigate the impact of different ethnicities, we compute the accuracies on the test set images, using a global EER threshold computed on the training set. Additionally, we determine the accuracies for each requirement using thresholds derived with the same previously mentioned protocol, but considering images of a specific ethnicity. As shown in Table XI, the framework shows similar performance between different ethnic groups, with slightly lower scores for compliant images of individuals with darker skin tones (EAF and EIA). Using the ethnicity-specific thresholds, we obtain marginally better results for African ethnicities, while obtaining similar performance with the EAS and EEA ones. However, we note a slight performance penalty with Indian-Asian end Middle Eastern ethnicities: this can be attributed to the fact that while the error rates are guaranteed to be equal, the threshold corresponding to the EER does not necessarily entail maximum accuracy.

More generally, it is important to emphasize that the system is designed to augment the visual quality of facial images in documents and is not intended for direct deployment in operational border control scenarios. Indeed, the proposed system has the potential to improve the quality of document images, reduce the workload associated with manual inspection, and enhance the consistency of biometric evaluation. The open dissemination of the framework encourages further research into both its capabilities and its limitations, particularly with respect to fairness and equity.

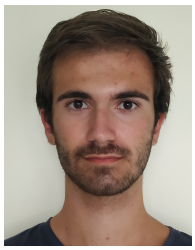
VII. CONCLUSIONS AND FUTURE WORKS

In this work, we presented a fully automated framework for ISO/ICAO compliance verification that is capable of extracting and interpreting requirements directly from official standards documents. The proposed approach leverages an LLM and a visual-language model, obviating the need for manually defined algorithms or handcrafted features. Experimental evaluations demonstrate that the method achieves performance comparable to, or surpassing, that of existing academic and commercial solutions. The framework's design allows for seamless integration of updates to compliance standards, ensuring sustained applicability without the need for extensive reconfiguration. We highlight the strong need for the availability of new and larger datasets in the literature. Not only is it necessary to have a greater amount of data, but it is also important that such data comprehensively and uniformly cover the various requirements introduced by the ISO/ICAO standards, in order to make the methods presented in the literature more easily comparable.

REFERENCES

[1] International Standards Organization, "ISO/IEC 19794-5 — Information technology — Biometric data interchange formats — Part 5: Face image data," International Organization for Standardization, Standard, 2011.
[2] ——, "ISO/IEC 39794-5 — Information technology — Extensible biometric data interchange formats — Part 5: Face image data," International Organization for Standardization, Standard, 2019.
[3] M. Ferrara, A. Franco, D. Maio, and D. Maltoni, "Face image conformance to iso/icao standards in machine readable travel documents," *IEEE Transactions on Information Forensics and Security*, vol. 7, 2012.
[4] A. G. d. A. e Silva, H. M. Gomes, and L. V. Batista, "A collaborative deep multitask learning network for face image compliance to iso/iec 19794-5 standard," *Expert Systems with Applications*, 2022.

[5] G. Borghi, A. Franco, N. Di Domenico, and D. Maltoni, "TONO: a Synthetic Dataset for Face Image Compliance to ISO/ICAO Standard," in *European Conference on Computer Vision*. Springer, 2024.
[6] J. Merkle, C. Rathgeb, B. Herdeanu, B. Tams, D.-P. Lou, A. Dörsch, M. Schaubert, J. Dehen, L. Chen, X. Yin, D. Huang, A. Stratmann, M. Ginzler, M. Grimmer, and C. Busch, "Open source face image quality (ofiq): Implementation and evaluation of algorithms," Federal Office for Information Security, Tech. Rep., 2024.
[7] O. Elatfi, N. Di Domenico, G. Borghi, A. Franco, and D. Maltoni, "Biogaze: a framework for evaluating the photographic requirements of the iso/iec 39794-5 standard," in *2025 IEEE 19th International Conference on Automatic Face and Gesture Recognition (FG)*, 2025.
[8] J. Wang, K. C. Chan, and C. C. Loy, "Exploring clip for assessing the look and feel of images," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 37, no. 2, 2023, pp. 2555–2563.
[9] 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*. PmLR, 2021, pp. 8748–8763.
[10] J. H. Holland, "Genetic algorithms," *Scientific american*, vol. 267, no. 1, pp. 66–73, 1992.
[11] A. Liu, S. Xue, J. Gan, J. Wan, Y. Liang, J. Deng, S. Escalera, and Z. Lei, "Cfpl-fas: Class free prompt learning for generalizable face anti-spoofing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 222–232.
[12] S. Liu, S. Yu, Z. Lin, D. Pathak, and D. Ramanan, "Language models as black-box optimizers for vision-language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 12 687–12 697.
[13] D. Maltoni, A. Franco, M. Ferrara, D. Maio, and A. Nardelli, "Biolab-icao: A new benchmark to evaluate applications assessing face image compliance to iso/iec 19794-5 standard," in *2009 16th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2009, pp. 41–44.
[14] N. Di Domenico, G. Borghi, A. Franco, and D. Maltoni, "Towards zero-shot iso/icao face compliance verification via clip-iqa and natural language prompting," in *International Joint Conference on Biometrics*. IEEE, 2025.
[15] International Civil Aviation Organization (ICAO), "Portrait Quality: Reference Facial Images for MRTD," ICAO, Standard, 2018.
[16] ——, "Machine readable travel documents. part 11: Security mechanisms for MRTDs," ICAO, Standard, 2015.
[17] International Standards Organization, "ISO/IEC 29794-5 — Information technology — Biometric sample quality — Part 5: Face image data," International Organization for Standardization, Standard, 2025.
[18] C. Guerra, J. Marcos, and N. Gonçalves, "Automatic validation of icao compliance regarding head coverings: An inclusive approach concerning religious circumstances," in *2023 International Conference of the Biometrics Special Interest Group (BIOSIG)*, 2023, pp. 1–4.
[19] R. L. Parente, L. V. Batista, I. L. P. Andrezza, E. V. C. L. Borges, and R. A. T. Mota, "Assessing facial image accordance to iso/icao requirements," in *2016 29th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, 2016, pp. 180–187.
[20] E. V. C. L. Borges, I. L. P. Andrezza, J. R. T. Marques, R. A. T. Mota, and J. J. B. Primo, "Analysis of the eyes on face images for compliance with iso/icao requirements," in *2016 29th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, 2016, pp. 173–179.
[21] A. Ahmadvand and M.-S. Moin, "Estimating conformity of head yaw to the icao standard using a convolutional neural network," in *2018 9th International Symposium on Telecommunications (IST)*, 2018.
[22] A. Mazandarani, P. M. F. Amaral, P. da Fonseca Pinto, and S. J. H. Shamoushaki, "Deep learning-based automated detection of inappropriate face image attributes for id documents," in *Technological Innovation for Applied AI Systems*, L. M. Camarinha-Matos, P. Ferreira, and G. Brito, Eds. Cham: Springer International Publishing, 2021.
[23] A. Nourbakhsh, M.-S. Moin, and A. Sharifi, "Facial images quality assessment based on iso/icao standard compliance estimation by hmax model," *Journal of Information Systems and Telecommunication*, 2020.
[24] I. Goodfellow, "Deep learning," 2016.
[25] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.
[26] G. Borghi, N. Di Domenico, A. Franco, M. Ferrara, and D. Maltoni, "Revelio: A modular and effective framework for reproducible training and evaluation of morphing attack detectors," *IEEE Access*, 2023.
[27] N. Di Domenico, G. Borghi, A. Franco, and D. Maltoni, "Onot: a high-quality icao-compliant synthetic mugshot dataset," in *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, 2024, pp. 1–10.
[28] 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.
[29] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid loss for language image pre-training," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 11 975–11 986.
[30] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.
[31] A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei *et al.*, "Qwen2.5 technical report," *arXiv preprint arXiv:2412.15115*, 2024.



Nicolò Di Domenico is a Computer Science and Engineering PhD student from the University of Bologna, Italy, having obtained his B.Sc. and M.Sc. degrees in 2020 and 2023, respectively. After his graduation, he is currently part of the Biometric Systems Laboratory at the University of Bologna, Italy, under the supervision of Prof. Davide Maltoni.

Guido Borghi is an Associate Professor within the Department of Education and Humanities, University of Modena and Reggio Emilia, Italy. He received the Ph.D. in Information and Communication Technologies from the same university in 2019. His research interests include Computer Vision and Deep Learning techniques applied to human analysis, including Face Analysis, Biometrics, Driver Monitoring and Human Computer Interaction.

Annalisa Franco is Associate Professor at the Department of Computer Science and Engineering, University of Bologna, Italy. In 2004 she received her Ph.D. in Electronics, Computer Science and Telecommunications Engineering at DEIS, University of Bologna. She is a member of the Biometric System Laboratory at Computer Science - Cesena. Her research interests include Biometric Systems with a specific focus on Face Analysis, Human Activity Recognition, and more generally Computer Vision.

Davide Maltoni received the degree in computer science and the Ph.D. degree in computer science and engineering from University of Bologna, Italy, in 1993 and 1997. He is a Full Professor with the University of Bologna (Dept. of Computer Science and Engineering-DISI), Bologna, Italy. His research interests are in the area of computer vision and machine learning. He is the co-Director of the Biometric Systems Laboratory. He is co-author of the Handbook of Fingerprint Recognition published by Springer Prof. Maltoni was elected International Association for Pattern Recognition Fellow 2010 and received the IAPR Senior Biometric Investigator Award in 2024.