

# Towards Zero-Shot ISO/ICAO Face Compliance Verification via CLIP-IQA and Natural Language Prompting

Nicolò Di Domenico<sup>1</sup>, Guido Borghi<sup>2</sup>, Annalisa Franco<sup>1</sup>, Davide Maltoni<sup>1</sup>

<sup>1</sup>University of Bologna, Italy

<sup>2</sup>University of Modena and Reggio Emilia, Italy

{name.surname}@unibo.it, guido.borghi@unimore.it

## Abstract

Ensuring compliance of face images with ISO/ICAO quality standards is essential for boosting the document enrollment process. Indeed, traditional manual checks are slow, subjective, and difficult to scale. Therefore, we propose a system that aims to fully automate compliance verification by directly analyzing the official requirements without relying on predefined hand-crafted features or manual thresholds. Our method combines a Large Language Model, a novel prompt learning procedure, and a contrastive learning framework to evaluate the adherence of a face image to quality requirements. Tested on a recent dataset, our proposed system achieves high accuracy, surpassing existing academic and commercial solutions. By streamlining the implementation and updates to the compliance rules, our approach represents a significant step toward simple, scalable, and regulation-driven image verification. Code and models are publicly available<sup>1</sup>.

## 1. Introduction

Ensuring that face photos comply with specific image quality standards, such as those defined by the International Civil Aviation Organization (ICAO) and the International Organization for Standardization (ISO) (here collectively referred to as ISO/ICAO), is a crucial requirement in identity verification processes for electronic Machine-Readable Travel Documents (eMRTDs) [11, 9].

The established guidelines [18, 19] impose strict constraints on the face images, including head positioning, lighting conditions, facial expression, background uniformity, and other photographic and geometric factors. These regulations aim to improve the reliability of identification procedures and the interoperability of identity verification across different jurisdictions and applications.

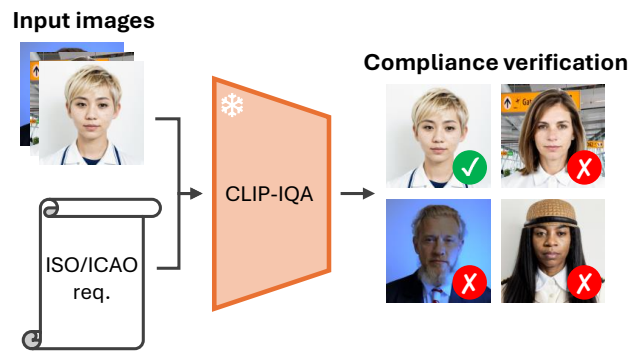


Figure 1. We propose an approach that is able to effectively classify whether input images are compliant with the given requirement, directly sourced from ISO/ICAO documents [18, 19] regulating face image quality.

Traditionally, compliance verification is performed by trained officials who manually inspect images to ensure they meet the required specifications. However, manual verification presents several intrinsic challenges. Firstly, it is a time-consuming process that requires human experts to analyze large volumes of images; secondly, human judgment can introduce inconsistencies due to subjective interpretation, fatigue, or variations in expertise among different evaluators. Additionally, compliance checks often involve several rules and technical details, which can be difficult to maintain consistently across multiple verification systems.

Given these limitations, there is a growing need for automated systems capable of performing compliance verification efficiently and accurately [5, 9, 11]. Indeed, automation can not only reduce the workload associated with manual inspections but also enhance the precision and reliability of compliance checks by eliminating human bias and ensuring a uniform application of standards. However, developing such a system is a challenging task, as checks on images are often based on sophisticated computer vision and machine learning algorithms, and compliance guidelines are frequently general and need to be interpreted correctly.

<sup>1</sup><https://github.com/MI-BioLab/CLIP-ICAO-Compliance>

Therefore, in this work, we propose a compliance verification system that aims to fully autonomously analyze ISO/ICAO documentation [19], understand guidelines, and verify compliance with requirements.

Unlike previous verification systems [10, 11] that usually rely on a variety of specific hand-crafted features, manual thresholds, and even intricate algorithms — all elements based on specific technical expertise — our approach eliminates the need for expert intervention by exploiting Large Language Models (LLMs) to autonomously analyze official standards documents and generate the required checks dynamically. This allows for greater simplicity and flexibility, ensuring that updates to current and future standards can be effortlessly integrated without requiring extensive manual reconfiguration or the intervention of technical experts.

From a technical point of view, we leverage a Vision-Language Model, *i.e.* CLIP-IQA [27], based on CLIP [26], which is able to capture the semantic relationship between natural language and images. Specifically, instead of training a task-specific model for each requirement, we query the model to measure the adherence of each image to a pair of textual prompts, respectively describing compliance and noncompliance with the given requirement, *e.g.* “*Eyes clearly visible, open*” vs. “*Eyes obstructed or closed*”. To automatically find suitable prompt pairs, we employ an LLM to analyze the official documentation, generating and iteratively refining the pair of prompts that describe the requirements. Then, we adopt the CLIP-IQA framework to compute the adherence between the input image and the pair of prompts related to a specific check.

We evaluate our proposed system on a recent dataset, *i.e.* TONO [5], which consists of synthetic images expressly collected for the ISO/ICAO compliance verification task. Results highlight the competitiveness of the proposed approach, which achieves comparable or better results than academic and commercial solutions.

In summary, our main contributions are the following:

- We propose a system that advances the automation of ISO/ICAO compliance verification. Specifically, the system verifies the compliance of input images based directly on the official guidelines. Notably, this approach makes it particularly easy to implement future integrations, limiting the need for technical experts and extensive manual implementation.
- We introduce a prompt learning technique with the aim of automatically generating and refining prompts extracted by an LLM directly from the official documents.
- Through an experimental evaluation, we show that the proposed system achieves competitive accuracy with respect to the solutions available in the market and in the literature.

## 2. Related works

### 2.1. Face Images and ISO/ICAO Compliance

The necessity of standardizing face image quality has arisen due to its significant influence on recognition accuracy. Researchers have formalized key factors [17, 16] affecting image quality, leading to the creation of systematic evaluation approaches. Maintaining high-quality facial images is particularly crucial in applications, such as eMRTDs, where identity documents remain valid for even long periods, making resilience to aging effects essential.

A first step toward standardization is the introduction of ICAO Doc 9303 [16], which specifies the functional requirements for eMRTDs and underscores the need for consistent portrait standards. Automated Border Control (ABC) systems rely on digital facial images to streamline identity verification, comparing printed, stored, and live-captured images at border crossings. ICAO standards later influenced the development of ISO/IEC 19794-5 [18], which evolved into ISO/IEC 39794-5 [19].

Beyond ISO/ICAO compliance, the broader challenge of assessing face image quality has spurred further standardization efforts. Ongoing work on ISO/IEC 29794-5 [20] seeks to establish a universal methodology for evaluating facial image quality across various scenarios, including those involving uncontrolled conditions with variations in lighting, pose, and other factors.

### 2.2. Methods for ISO/ICAO compliance verification

In the last decades, several tools have been developed to assess compliance with ISO/ICAO standards. One of the earliest research efforts in this domain, presented in [11], introduced a structured set of requirements and corresponding algorithmic methods for verification.

Various commercial solutions have emerged in the market. Some of these are specifically tailored for ICAO verification, while others originate from facial recognition SDKs that have incorporated compliance-checking capabilities. Alongside these commercial products, a handful of research initiatives have contributed open-source tools. Regarding commercial solutions, Correlance has developed an SDK<sup>2</sup> designed to verify adherence to the initial version of the ISO/ICAO standard. Another relevant solution comes from Innovatrics<sup>3</sup>, whose software provides broader verification functionalities, including passive liveness detection and match quality assessment. This tool offers real-time user feedback to ensure compliance during image acquisition. Additionally, NEUROtechnology’s SDK<sup>4</sup> prioritizes basic image quality checks, such as minimum eye distance and frontal pose evaluation. As these products are propri-

<sup>2</sup><https://www.correlance.com/cms/en/ccEngineICAO>

<sup>3</sup><https://www.innovatrics.com/digital-onboarding-toolkit>

<sup>4</sup><https://www.neurotechnology.com/face-verification-technical.html>

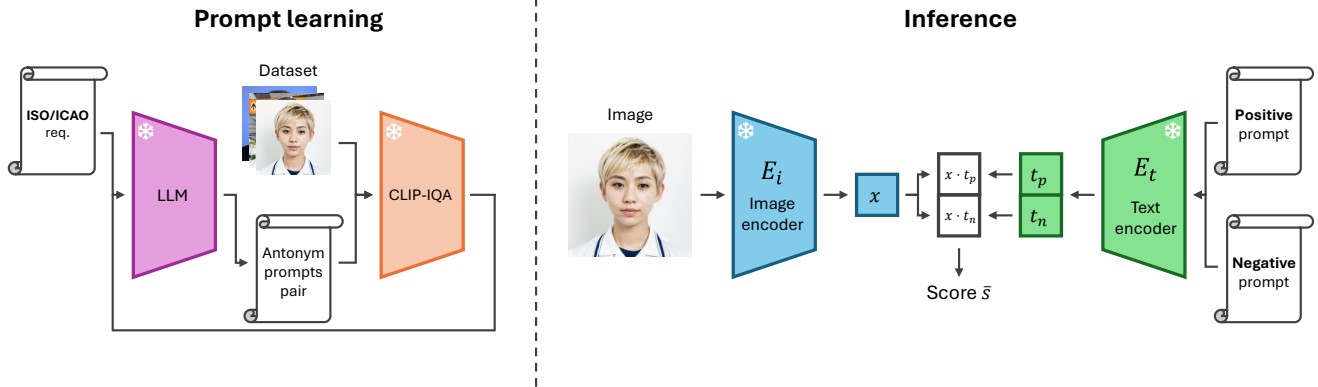


Figure 2. Overview of our proposed method. On the left, the prompt learning procedure is depicted, where an LLM creates antonym prompt pairs starting from the requirement’s description from ISO/ICAO documents. These prompts are then evaluated by CLIP-IQA [27], allowing the LLM to guide its output to obtain better prompt pairs at each generation. On the right, the image and text encoders  $E_i$  and  $E_t$  from CLIP-IQA are used to encode respectively the image and the two prompts to the shared latent space embeddings  $x$ ,  $t_p$ , and  $t_n$ . Then, the cosine similarity between the image embedding  $x$  and the two text embeddings  $t_p$  and  $t_n$  is calculated, and the score  $\bar{s}$  is found by applying the Softmax function.

etary and often part of larger facial recognition SDKs, neither their source code nor verification algorithms are publicly available. Consequently, their use in research contexts is somewhat restricted and limited.

Beyond commercial tools, several research initiatives have focused on dataset creation and model development for ISO/ICAO compliance verification. One notable example is the BioLab-ICAO framework [22], later refined in [11], which offers standardized testing protocols and baseline compliance verification algorithms. However, researchers can access only a limited subset of images for training, as most remain restricted and serve primarily as benchmarks on the FVC-onGoing<sup>5</sup>.

Further contributions include the BioLab-ICAO-Check tool, introduced in [11], which employs computer vision techniques such as the Prewitt operator [7] for pixelation detection and TSI [12] for blurriness evaluation. Other studies have addressed specific ICAO requirements, including head coverings (linked to religious considerations) [14], pixelation, occlusions (hair across eyes, veils, open mouth) [25], eye conditions (closed eyes, red-eye effect, gaze direction) [3], head pose analysis [2], and evaluations of mouth closedness, eye openness, and face occlusions [23].

Others have aimed to cover a broader range of requirements. For instance, [24] proposes an integrated system inspired by human cognitive processes, addressing nine different compliance criteria in a unified manner, unlike methods that assess each requirement separately. A recent approach is presented in [9], where the authors introduce ICAONet, a deep learning-based multitask network designed to automate compliance verification for the ISO/IEC 19794-5 [18] standard. This network consists of three com-

ponents: an encoder, adapted from the autoencoder in [13], which generates an embedding of the face image; an unsupervised decoder that reconstructs the original image; and a supervised classification branch that predicts compliance scores for different requirements.

From a general point of view, a key challenge of these compliance verification tools is that although all of them adhere to the same ISO/ICAO guidelines, each method implements a different subset of compliance checks. This discrepancy complicates direct comparisons between systems, as differences in guideline interpretation further obscure precise feature mappings.

### 3. Proposed method

A general overview of our method is shown in Figure 2. The system is divided into two main parts. In the first, referred to as Prompt Learning (PL), an LLM is used to analyze official documents containing ISO/ICAO guidelines and to generate and iteratively refine a set of starting antonym prompts that describe each requirement.

Once a positive and negative prompt is defined for each check extracted from the ISO/ICAO documents, in the inference phase they are then compared with facial images, and the adherence between image contents and prompts is evaluated through CLIP-IQA [27], a version of CLIP [26] tailored for Image Quality Assessment (IQA) tasks (see Sect. 3.1). More specifically, as image and text embeddings lie in the same latent space, we determine whether an input image is compliant or not by computing the cosine similarity between the image and the positive and negative text embeddings extracted from the VLM. The system finally outputs a score in the range  $[0, 1]$  (see Sect. 3.2).

<sup>5</sup><https://biolab.csr.unibo.it/FvcOnGoing>

### 3.1. Prompt Learning

The choice of wording in the prompts is critical and has a significant impact on the system’s performance [26, 27]; therefore, it is essential to formulate the prompts accurately.

Inspired by [21], we automate this process by employing an LLM that is instructed to generate prompt pairs starting from the description of the given requirement, which is directly sourced from the official ISO/ICAO document. Specifically, we design a prompt learning (PL) algorithm that resembles traditional genetic algorithms [15], in which for each generation an LLM generates antonym prompt pairs (positive and negative, respectively describing compliance and noncompliance) from both the requirement and the top-performing pairs of the previous generation, and their fitness is evaluated by CLIP-IQA. We choose Phi-4 [1] as our LLM, as we empirically find that it provides a good balance between its reasoning capabilities and its size of 14B parameters, enabling us to run the model locally.

More specifically, at each generation we generate prompts by directly deriving them from the requirement, as well as from the top- $k$  and bottom- $k$  prompt pairs: the first creation method has the goal of condensing in the best possible way the natural language description of the requirement down to a pair of antonym prompts of maximum 15 words to stay within the 77-token limit imposed by CLIP’s text encoder; on the other hand, the second generation method’s objective is to recombine and refine good-performing prompts, while avoiding terms that are associated with poorly performing ones. To reduce the likelihood of hallucinations, we provide a copy of the original requirement to the LLM along with the top- $k$  and bottom- $k$  prompt pairs. Similarly to genetic algorithms, these two approaches to creating prompt pairs can be roughly compared with the exploration and exploitation mechanisms [15]. To strike a good balance between the two, the algorithm creates for each generation 50 fresh prompts from the requirement, as well as 50 similar to the top-20, while avoiding commonalities with the bottom-20 prompt pairs. All system prompts can be found in the supplementary material.

The chosen metric to evaluate the fitness of each prompt pair is the Equal Error Rate (EER) [4], defined as the point at which the False Acceptance Rate (FAR) and False Rejection Rate (FRR) curves intersect (see Sect. 4 for further details). This error is computed on a score obtained using the score function described in Sect. 3.2. For the sake of comprehension, a pseudocode version of the implemented method is shown in Algorithm 1.

Being fully automated starting from only the official requirement description, the proposed pipeline can be quickly applied to new or updated requirements with minimal human intervention, as prompt generation and evaluation are entirely data-driven.

---

#### Algorithm 1 Prompt Learning (PL)

---

**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  $X = \{x_1, \dots, x_N\} \subset \mathbb{R}^d$   
**Require:** Ground truth labels  $Y = \{y_1, \dots, y_N\}$ , where  $y_i \in \{0, 1\}$   
**Require:** Top- $k$  prompt pairs to retain per generation  
**Require:** Scoring function  $S(x, t_p, t_n) \rightarrow [0, 1]$

- 1: Initialize prompt pair set  $P = \{(p_p^{(j)}, p_n^{(j)})\}_{j=1}^m$  from textual requirement  $r$
- 2: **for**  $g = 1$  to  $n$  **do**
- 3:     **for**  $(p_p, p_n) \in P$  **do**
- 4:         Encode prompts:  $t_p \leftarrow E_t(p_p), t_n \leftarrow E_t(p_n)$
- 5:          $\hat{Y} \leftarrow S(x, t_p, t_n), \forall x \in X$
- 6:         Compute EER:  $e \leftarrow \text{EER}(\hat{Y}, Y)$
- 7:         Store  $(p_p, p_n, e)$
- 8:     **end for**
- 9:     Sort prompt pairs in  $P$  by ascending EER
- 10:      $P_{\text{top}} \leftarrow$  top- $k$  pairs with lowest EER
- 11:      $P_{\text{bottom}} \leftarrow$  bottom- $k$  pairs with highest EER
- 12:      $P_{\text{new}} \leftarrow$  Generate  $\lfloor m/2 \rfloor$  new prompt pairs from requirement  $r$
- 13:      $P_{\text{child}} \leftarrow$  Generate  $\lfloor m/2 \rfloor - |P_{\text{top}}|$  offspring pairs from  $P_{\text{top}}$  and  $P_{\text{bottom}}$
- 14:      $P \leftarrow P_{\text{new}} \cup P_{\text{child}} \cup P_{\text{top}}$
- 15: **end for**

---

### 3.2. ISO/ICAO Compliance Verification

In this second phase, we test compliance for each ISO/ICAO requirement by assessing the correlation between the given input image and a pair of antonym prompts, a positive prompt  $p_p$  describing a compliant photo, and a negative prompt  $p_n$  describing a noncompliant photo.

For instance, the requirement specifying “*head without coverings*” might have a positive prompt that reads “*The subject is not wearing any type of headgear, and the hair is visible*”, and a negative prompt stating “*The subject is wearing a hat, cap, bandana, or any other garment that hides the hair*”. Pairs of antonym prompts are adopted in place of single prompts to reduce the impact of linguistic ambiguity and to improve discriminative power, as the similarity to a single prompt may be insufficient for reliable assessment [27].

Then, each of the two prompts is processed using CLIP-IQA’s text encoder  $E_t(\cdot)$ , obtaining fixed-size text embeddings  $t_p, t_n \in \mathbb{R}^d$ , where  $t_p = E_t(p_p)$  and  $t_n = E_t(p_n)$ . Similarly, the input image  $I$  is processed by the image encoder  $E_i(\cdot)$ , obtaining its embedding  $x \in \mathbb{R}^d$  in the same latent space. We then evaluate the correlation between the image and each of the two prompts by calculating their cosine similarities:

Requirements	ICAONet	BioLab	Correlance	Innovatrics	Ours (M. PL)	Ours (R. PL)	
Subject	Head w/o coverings	0.179	0.282	0.030	-	<b>0.007</b>	<u>0.010</u>
	Gaze in camera	0.456	0.515	<b>0.177</b>	<u>0.277</u>	0.483	0.486
	Eyes open	0.022	0.500	<b>0.000</b>	<b>0.000</b>	<b>0.000</b>	<u>0.007</u>
	No/light makeup	-	-	-	-	<u>0.008</u>	<b>0.000</b>
	Neutral expression	0.321	0.256	<u>0.097</u>	0.110	0.179	<b>0.051</b>
	No sunglasses	0.035	0.198	<b>0.000</b>	0.122	<u>0.007</u>	<b>0.000</b>
	Frontal pose	0.206	-	-	-	<b>0.145</b>	<u>0.210</u>
	<i>Mean</i>	0.203	0.350	<b>0.061</b>	0.127	0.119	<u>0.109</u>
Photogr.	Correct exposure	0.343	<u>0.199</u>	0.301	0.262	<b>0.074</b>	0.212
	In focus photo	0.353	<u>0.003</u>	0.006	0.076	<b>0.000</b>	0.026
	Correct saturation	0.330	<b>0.058</b>	0.458	-	<u>0.064</u>	0.103
	<i>Mean</i>	0.342	<u>0.087</u>	0.255	0.169	<b>0.046</b>	0.113
Acquisition	Uniform background	0.394	0.386	<u>0.114</u>	-	<b>0.071</b>	<b>0.071</b>
	Uniform face lighting	0.365	-	<b>0.084</b>	0.286	<u>0.128</u>	0.141
	No pixelation	0.490	<u>0.330</u>	-	-	<b>0.006</b>	<b>0.006</b>
	No posterization	-	-	<u>0.241</u>	0.500	<b>0.006</b>	<b>0.006</b>
	<i>Mean</i>	0.416	0.358	0.146	0.393	<b>0.053</b>	<u>0.056</u>
<b>Global Mean</b>	0.291	0.273	0.137	0.204	<b>0.084</b>	<u>0.095</u>	

Table 1. Performance comparison measured in Equal Error Rate (EER) of our proposed approach against various methods from the literature, as well as commercial off-the-shelf SDKs. We report our method with prompts learned from handcrafted prompt pairs (“M. PL” column), as well as with prompts learned from the requirements’ descriptions (“R. PL” column). Lower EER values indicate better performance. Best results for each requirement are in **bold**, while second-best results are underlined.

$$s_i = \frac{x \cdot t_i}{\|x\| \cdot \|t_i\|}, \quad i \in \{p, n\} \quad (1)$$

where  $s_p$  and  $s_n$  represent the similarity scores of the image for the positive and negative prompts, respectively. Finally, we transform these raw similarities into a probabilistic compliance score  $\bar{s} \in [0, 1]$  by applying the Softmax function:

$$\bar{s} = \frac{e^{s_p}}{e^{s_p} + e^{s_n}} \quad (2)$$

Since each compliance check returns a continuous score between 0 and 1, this enhances explainability (it is possible to understand which requirements are fulfilled precisely) and supports the decision-making process.

## 4. Experiments

For a standardized assessment, we employ the TONO dataset [5], a synthetic dataset specifically designed for ISO/ICAO compliance testing purposes. This dataset consists of roughly 4k images, each featuring a single feature that breaches the ISO/ICAO guidelines, making it an ideal candidate for our experimental evaluation by allowing us to focus independently on each single compliance check.

It is worth noting that this is the only available benchmark for the ISO/ICAO compliance verification task, and then we report the results of all competitors on this data. A sequestered web-based dataset is available on the FVC-onGoing platform, but unfortunately at present it is impossible to test our method due to technical and time constraints.

For all experiments where prompt pairs are learned starting from handcrafted prompts or official ISO/ICAO descriptions, we run the prompt learning algorithm for 50 generations. We empirically find that a number of generations higher than 50 does not significantly improve the system’s performance. Furthermore, we employ a 50-50 random split between training and test images, ensuring that the two subsets are balanced with respect to both ethnicity and gender.

For every reported method, we present its performance using the Equal Error Rate (EER), *i.e.* the intersection point at which, for each quality check, the number of erroneous accepted and rejected images is the same. Being an error measure, better-performing prompt pairs will present lower EER values.

As computing these metrics requires both positive and negative samples, the TONO dataset is expanded by including ISO/ICAO-compliant synthetic images sourced from the ONOT [6] dataset. As competitors, we include a variety of available methods. ICAONet [9] and BioLab-ICAO-

Requirements		ViT-B/16	ViT-B/32	ViT-L/14	ViT-L/14@336	CLIP-IQA
Subject	Head w/o coverings	0.112	<b>0.028</b>	0.142	0.122	0.048
	Gaze in camera	0.481	0.487	0.352	<b>0.340</b>	0.489
	Eyes open	0.263	0.287	0.034	<b>0.025</b>	0.096
	No/light makeup	0.030	<b>0.006</b>	0.088	0.102	0.231
	Neutral expression	0.545	0.550	<b>0.542</b>	0.635	0.594
	No sunglasses	0.090	0.095	<b>0.052</b>	0.054	0.130
	Frontal pose	0.585	0.578	0.660	0.624	<b>0.574</b>
<i>Mean</i>	0.301	0.290	<b>0.267</b>	0.272	0.309	
Photogr.	Correct exposure	0.503	0.500	0.436	<b>0.395</b>	0.490
	In focus photo	0.445	0.405	0.281	0.215	<b>0.042</b>
	Correct saturation	0.355	<b>0.333</b>	0.394	0.355	0.494
	<i>Mean</i>	0.435	0.413	0.370	<b>0.322</b>	0.342
Acquisition	Uniform background	0.488	<b>0.398</b>	0.457	0.459	0.450
	Uniform face lighting	0.434	0.427	<b>0.300</b>	0.358	0.354
	No pixelation	0.444	0.428	0.500	<b>0.342</b>	0.566
	No posterization	0.166	0.235	0.225	0.190	<b>0.045</b>
	<i>Mean</i>	0.383	0.372	0.370	<b>0.337</b>	0.354
<b>Global Mean</b>	0.353	0.340	0.319	<b>0.301</b>	0.329	

Table 2. Performance comparison measured in Equal Error Rate (EER) of four versions of CLIP, each one with a different size of Vision Transformer, as well as CLIP-IQA, which utilizes a modified ResNet50 and does not use positional embeddings in the image encoder.

Check [11], shortened to BioLab hereafter, as our selected competitors from the academic literature. Furthermore, we also compare the results obtained with our framework against commercial off-the-shelf SDKs provided by Correlance, and Innovatrics (see Sect. 2.2).

Experimental results of the comparison are depicted in Table 1. To facilitate the comparison, the implemented checks are divided into three groups: subject, photographic, and acquisition requirements [5]. We report the mean EER across all implemented checks for each category, as well as a global mean EER to provide a summarized representation of the method’s performance. However, it is important to note that, as these summary metrics are calculated considering only the implemented checks, they may not be directly comparable between methods that are able to verify different subsets of requirements. For the sake of completeness, we report the performance of our method utilizing prompts learned from pairs handcrafted by an expert, in addition to those automatically extracted from ISO/ICAO documents.

The experimental evaluation indicates that, while the overall best-performing solution employs prompts learned starting from the handcrafted pairs, thus requiring the intervention of a human expert, the proposed fully automatic solution using prompts derived from the ISO/ICAO requirements is still able to outperform all chosen competitors in several requirements. Additionally, we note that some com-

petitors fail to implement several checks that are critical in order to guarantee ISO/ICAO compliance; in particular, the “frontal pose” requirement, which is essential to guarantee optimal performance of face recognition systems, is implemented by only one out of the chosen four competitors. Furthermore, our method successfully identifies noncompliant makeup, a feature absent in competitors’ solutions.

Finally, we note a significant improvement in certain checks over the current state of the art: for example, the “no posterization” check shows a significant decrease in EER, from 0.224 to 0.006; similarly, the EER for the “correct exposure” requirement for the best-performing variant of our method is reduced by half compared to the competitors’ best result. Conversely, the “gaze in camera” check has proven to be particularly challenging for CLIP-IQA: we hypothesize that this behavior may be due to the fact that pupils are a relatively small detail within the whole image that can be affected by external factors, and therefore the model’s output may be volatile. For the sake of reproducibility and understanding, all the generated prompts for each check used for the validation are reported in the supplementary material. In conclusion, it is important to note that these results are obtained without the use of manual and specific thresholds, hand-crafted features, and training procedures, all elements exploited in the compared methods.

Requirements		ViT-L/14@336			CLIP-IQA		
		Manual	Manual PL	Req. PL	Manual	Manual PL	Req. PL
Subject	Head w/o coverings	0.142	0.007	<b>0.000</b>	0.056	<b>0.007</b>	0.010
	Gaze in camera	0.345	<b>0.274</b>	0.316	<b>0.469</b>	0.483	0.486
	Eyes open	0.018	0.007	<b>0.000</b>	0.111	<b>0.000</b>	0.007
	No/light makeup	0.078	<b>0.008</b>	0.011	0.218	0.008	<b>0.000</b>
	Neutral expression	0.673	0.093	<b>0.064</b>	0.599	0.179	<b>0.051</b>
	No sunglasses	0.068	<b>0.007</b>	<b>0.007</b>	0.125	0.007	<b>0.000</b>
	Frontal pose	0.628	0.210	<b>0.186</b>	0.595	<b>0.145</b>	0.210
	<i>Mean</i>	0.279	0.087	<b>0.083</b>	0.310	0.119	<b>0.109</b>
Photogr.	Correct exposure	0.359	0.253	<b>0.237</b>	0.481	<b>0.074</b>	0.212
	In-focus photo	0.205	0.071	<b>0.058</b>	0.045	<b>0.000</b>	0.026
	Correct saturation	0.388	0.256	<b>0.186</b>	0.484	<b>0.064</b>	0.103
	<i>Mean</i>	0.317	0.193	<b>0.160</b>	0.337	<b>0.046</b>	0.113
Acquisition	Uniform background	0.469	<b>0.109</b>	0.116	0.442	<b>0.071</b>	<b>0.071</b>
	Uniform face lighting	0.372	<b>0.115</b>	0.179	0.308	<b>0.128</b>	0.141
	No pixelation	0.346	<b>0.058</b>	0.071	0.567	<b>0.006</b>	<b>0.006</b>
	No posterization	0.186	<b>0.051</b>	0.109	0.048	<b>0.006</b>	<b>0.006</b>
	<i>Mean</i>	0.343	<b>0.083</b>	0.119	0.341	<b>0.053</b>	0.056
<b>Global Mean</b>		0.306	<b>0.108</b>	0.110	0.325	<b>0.084</b>	0.095

Table 3. Performance comparison measured in Equal Error Rate (EER) of the best version of CLIP found in Table 2 against CLIP-IQA, using prompts derived with three different methods: manual, handcrafted prompts (“Manual” column), prompt learning with the same handcrafted prompts as seed (“Manual PL” column), and prompt learning with the description of the requirement sourced directly from the official ISO/ICAO documents [17] (“Req. PL” column). Lower EER values indicate better performance.

## 5. Ablation studies

### 5.1. Study on encoders

Firstly, we attempt to replace CLIP-IQA [27] with one of the four pretrained versions of CLIP [26] leveraging Vision Transformers [8] as their image encoder; specifically, we investigate the use of ViT-B/16, ViT-B/32, ViT-L/14, and another version of the latter trained with larger images of  $336 \times 336$  pixels, namely ViT-L/14@336. To ensure consistency across all experiments, they are carried out using the same prompt pairs for each requirement. These prompt pairs are handcrafted by a human expert, condensing the requirement description found in official ISO/ICAO documents [17], without the use of any prompt learning technique. Therefore, we evaluate metrics on the entire dataset.

As shown in Table 2, larger models such as ViT-L/14 and ViT-L/14@336 result in better overall performance. More specifically, specific requirements concerning finer details (e.g. “gaze in camera”, “eyes open”) show the greatest improvement with larger models and higher input resolutions. In contrast, larger patch sizes appear to be preferable when capturing coarser details such as the presence of head coverings, clearly visible makeup, and the background. Fur-

thermore, comparing CLIP-IQA and all other CLIP variants is helpful in assessing the performance impact of positional embeddings. Indeed, we note that the former tends to struggle with requirements that are checked by inspecting a limited portion of the image, such as “gaze in camera” and “eyes open”, while being competitive in verifying more global properties such as “in focus photo”, “no posterization”, and “uniform background”. Therefore, we hypothesize that positional embeddings may have a bigger impact on fine-grained tasks, where the model must have knowledge of the position of elements such as eyes, makeup, or other facial features.

### 5.2. Study on Prompt Learning

Various prompting techniques are tested based on the best-performing CLIP model found in Section 5.1, namely ViT-L/14@336, and on CLIP-IQA. Specifically, we evaluate the effectiveness of these encoders using three different prompting techniques. We test manually created prompt pairs, along with two distinct variants of the prompt learning algorithm: one variant derives pairs using the handcrafted prompts as seed, and the other variant constructs pairs directly based on the requirements described in the ISO/ICAO documents.

Experimental results are shown in Table 3. As expected, the use of prompt learning has yielded a significant performance improvement on all checks compared to manual prompting, obtaining respectively a  $-64.7\%$  and a  $-72.5\%$  reduction in EER using ViT-L/14@336 and CLIP-IQA. However, we note that the usage of the full requirement’s description as seed for the prompt learning algorithm is not immediately more effective than employing handcrafted prompt pairs, despite being a simpler and more natural approach that does not require specific knowledge. Indeed, this method obtains a moderate increase in EER quantified as  $+1.8\%$  and  $+13\%$  respectively with ViT-L/14@336 and CLIP-IQA. As shown in the Table, some checks (e.g. “uniform face lighting” and “no posterization” for ViT-L/14@336, “frontal pose” and “correct exposure” for CLIP-IQA) obtain significantly worse scores with prompt learning from the requirement’s description rather than from a pair of handcrafted prompts. This discrepancy is likely because the prompt learning algorithm may struggle to focus on the most relevant parts of the requirements as outlined in ISO/ICAO documents [17], whereas an already condensed version of the same requirements crafted by a human expert minimizes potential hallucinations by the LLM. This is also noted by examining the learned prompts: indeed, the extra details present in the official description of the requirement may sometimes cause the LLM to focus on the wrong details, as demonstrated by the “head w/o coverings” check with ViT-L/14@336: while the prompt pair learned using the handcrafted pair as seed state “*Hair fully visible, no headwear*” vs. “*Headwear covered with a hat, cap, or bandana*”, the ones generated from the ISO/ICAO requirement state “*Face visible from ear to ear, no distortion*” vs. “*Ear to ear face covered by headwear*”, focusing on a secondary aspect in the specification.

However, while the approach leveraging prompts learned from pairs handcrafted by an expert presents overall better results, we note that the loss of performance when employing prompts learned directly from ISO/ICAO documents is negligible, making it a feasible option for a completely automated pipeline that does not require a human expert.

## 6. Ethical and Social Impact Analysis

As mentioned above, the framework introduced in this study has been evaluated using a publicly accessible synthetic dataset, thereby circumventing the need for new data acquisition using real people and identities. In addition, the used dataset contains samples depicting various ethnicities — African (EAF), East-Asian (EAS), European/American (EEA), Indian-Asian (EIA), and Middle Eastern (EME) — genders, and ages, reducing the risk of biased evaluation [5]. However, we recognize that the use of data generated by diffusion models potentially trained on extensive web-scraped datasets can raise ethical, legal, and privacy concerns.

Requirements	EAF	EAS	EEA	EIA	EME
Subject	0.116	0.094	0.103	0.098	0.098
Photographic	0.062	0.045	0.057	0.149	0.034
Acquisition	0.063	0.047	0.041	0.051	0.044
<b>Global Mean</b>	<b>0.090</b>	<b>0.070</b>	<b>0.075</b>	<b>0.095</b>	<b>0.069</b>

Table 4. Performance comparison of our method across five different ethnicities (see Sect. 6). For each group of requirements, we report the mean EER and the global mean EER.

We observe that the reliance on pre-trained models, such as the text and image encoders of CLIP-IQA, carries the inherent risk of perpetuating biases present within their original training data. To better assess the impact of this potential drawback, we evaluate the system’s performance by employing images representing only one of the five available ethnicities at a time. As depicted in Table 4, the system performs similarly regardless of ethnicity; we note only a limited performance penalty in those characterized by darker skin (EAF and EIA). The proposed system has the potential to enhance the quality of document images, reduce the burden of manual inspection, and improve the consistency of biometric evaluations. It is important to note that the system is not intended to be used directly in border controls but only to enhance the visual quality of faces in documents. The open dissemination of the framework fosters further investigation into its capabilities and, more importantly, its limitations regarding fairness and equity.

## 7. Conclusions and future works

We introduced an automated system for ISO/ICAO compliance verification that dynamically extracts and interprets guidelines from official documents. Unlike traditional approaches, the proposed solution leverages the classification capabilities of CLIP and CLIP-IQA, eliminating the need for predefined thresholds and handcrafted features. The experimental evaluation on the TONO dataset showed that our method matches or surpasses existing solutions. Its ability to seamlessly integrate updates to compliance standards ensures long-term applicability without extensive reconfiguration. Future work will focus on extending its capabilities to real-world datasets, further refining its prompt learning mechanism, while ensuring cross-dataset generalization capabilities.

## Acknowledgement

This project received funding from the European Union’s Horizon Europe research and innovation program under Grant Agreement No.101121280. This text reflects only the author’s views, and the commission is not liable for any use that may be made of the information contained therein.

## References

- [1] M. Abdin, J. Aneja, H. Behl, S. Bubeck, R. Eldan, S. Gunasekar, M. Harrison, R. J. Hewett, M. Javaheripi, P. Kauffmann, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024. 4
- [2] 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)*, pages 69–73, 2018. 3
- [3] 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)*, pages 173–179, 2016. 3
- [4] 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*, 11:120419–120437, 2023. 4
- [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. 1, 2, 5, 6, 8
- [6] 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)*, pages 1–10, 2024. 5
- [7] W. Dong and Z. Shisheng. Color image recognition method based on the prewitt operator. In *2008 International Conference on Computer Science and Software Engineering*, volume 6, pages 170–173. IEEE, 2008. 3
- [8] 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. 7
- [9] 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*, 198:116756, 2022. 1, 3, 5
- [10] 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)*, pages 1–10, 2025. 2
- [11] 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*, 7(4):1204–1213, 2012. 1, 2, 3, 6
- [12] M. Ferrara, A. Franco, and D. Maltoni. Fingerprint scanner focusing estimation by top sharpening index. In *14th International Conference on Image Analysis and Processing (ICIAP 2007)*, pages 223–228. IEEE, 2007. 3
- [13] I. Goodfellow. Deep learning, 2016. 3
- [14] 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)*, pages 1–4, 2023. 3
- [15] J. H. Holland. Genetic algorithms. *Scientific american*, 267(1):66–73, 1992. 4
- [16] International Civil Aviation Organization (ICAO). Machine readable travel documents. part 11: Security mechanisms for MRTDs. Standard, International Civil Aviation Organization, 2015. 2
- [17] International Civil Aviation Organization (ICAO). Portrait Quality: Reference Facial Images for MRTD. Standard, International Civil Aviation Organization, 2018. 2, 7, 8
- [18] International Standards Organization. ISO/IEC 19794-5 — Information technology — Biometric data interchange formats — Part 5: Face image data. Standard, International Organization for Standardization, 2011. 1, 2, 3
- [19] International Standards Organization. ISO/IEC 39794-5 — Information technology — Extensible biometric data interchange formats — Part 5: Face image data. Standard, International Organization for Standardization, 2019. 1, 2
- [20] International Standards Organization. ISO/IEC 29794-5 — Information technology — Biometric sample quality — Part 5: Face image data. Standard, International Organization for Standardization, under development. 2
- [21] 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*, pages 12687–12697, 2024. 4
- [22] 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)*, pages 41–44. IEEE, 2009. 3
- [23] 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 L. M. Camarinha-Matos, P. Ferreira, and G. Brito, editors, *Technological Innovation for Applied AI Systems*, pages 243–253. Cham, 2021. Springer International Publishing. 3
- [24] 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 (JIST)*, 3(27):225, 2020. 3
- [25] 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)*, pages 180–187, 2016. 3
- [26] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmlR, 2021. 2, 3, 4, 7
- [27] 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*, volume 37, pages 2555–2563, 2023. 2, 3, 4, 7