Curriculum-Guided Diffusion

Generating ICAO-Compliant Synthetic Face Images by teaching passport-photo rules without gradients

Raghavendra Mudgalgundurao Patrick Schuch Aryan Khurana Raghavendra Ramachandra Kiran Raja

Norwegian Biometrics Laboratory, Dept. of Computer Science, NTNU Gjøvik, Norway
BITS Pilani, India

raghavem@stud.ntnu.no

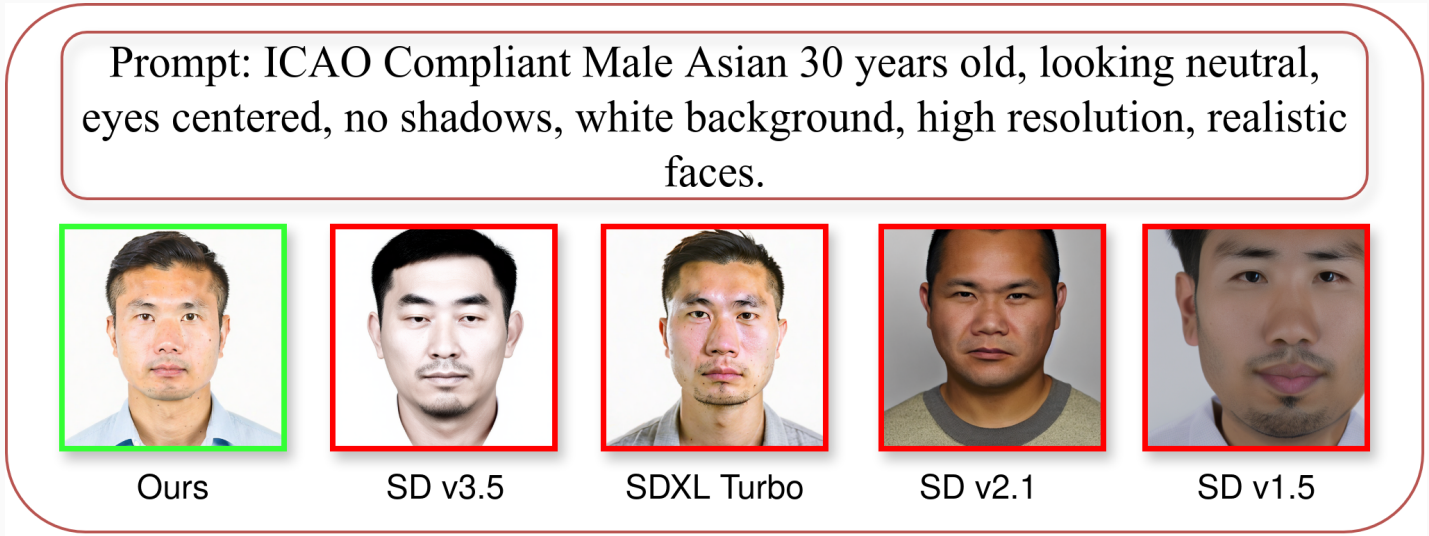
ICPR 2026 · Lyon

NTNU | Norwegian University of Science and Technology

ICPR 2026 International Conference on Pattern Recognition LYON

Looks like a passport photo — still rejected

In document pipelines, compliance is a gate that runs before recognition.



Same prompt, five models. Green = compliant example · red = visible compliance failure.

Looks plausible

A human would call these passport photos.

Rejected upstream

Small capture errors become hard failures at the gate.

Plausible is not the same as usable — an image that fails compliance never reaches a matcher.

SD 1.5 fails 35.6% of the 23 ISO/IEC 19794-5 tests on average. Real ICAO-grade data is scarce, restricted and demographically skewed, so scaling the dataset is not an available answer.

The prompt is understood. The standard is not.

The failures are not semantic misunderstandings — they are capture-constraint failures.

Prompt: ICAO Compliant Male Asian 30 years old, looking neutral, eyes centered, no shadows, white background, high resolution, realistic faces.

Prompt: ICAO compliant Female Caucasian 30 years old, looking neutral, eyes centered, no shadows, white background, high resolution, realistic faces.

What breaks

Capture geometry, not semantics

- Head pose drifts off-frontal — roll, pitch and yaw is the worst-failing test for every baseline.
- Eye position and framing wander.
- Expression is not reliably neutral.
- Backgrounds and shadows stay uncontrolled.

Exactly the things a document capture standard cares about.

Can a black-box checker teach a generator?

We treat the compliance scorer as a black box: no gradient is taken through it, and its score is not used as a training target.

The obvious answer is no

- no gradients
- no target labels
- no pass / fail supervision

The useful answer is different

Do not optimise the compliance score.

Use it to order reality.

We never backpropagate through the checker.

Every curriculum needs a difficulty signal. Ours comes from a compliance scoring network that is consulted and ranked on, but never differentiated through.

What teaches what

Three corpora, each doing a different job across the two stages.

Dataset	Subjects	Images	Compliance	Teaches
CFD	597	~600	Non-ICAO	Stage 1 · style
ONOT	255	~13k (512×512)	Full ICAO	Stage 1 · style
FRGC	466	~50k	Partial ICAO	Stage 2 · the curriculum
Ours	**—**	**~10k (512×512)**	**Full ICAO**	**Output**

Style comes from ONOT + CFD

Filtered Stage-1 examples teach composition, framing, lighting and background.

Difficulty comes from FRGC

The only real corpus in the curriculum. Partially compliant, so it spans the easy-to-hard range.

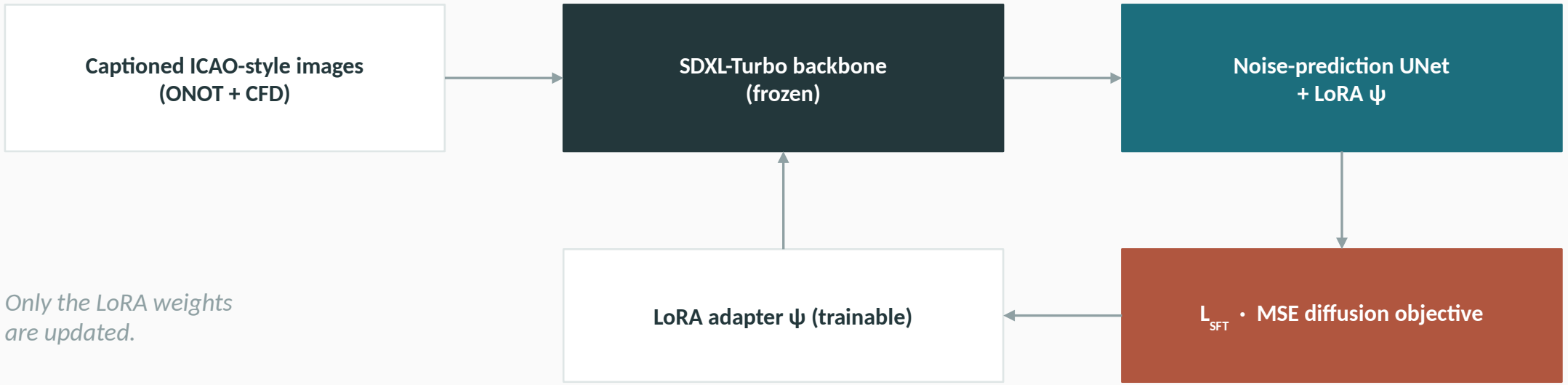
Pre-processing

Manual screening plus an ICSN threshold; rembg composites FRGC onto white. DeepFace captions are style guidance only.

FRGC is only partially compliant — which is exactly why it can be ordered from easy to hard.

Stage 1 — learn the style

Supervised fine-tuning on filtered ICAO-style images. The backbone is never updated.



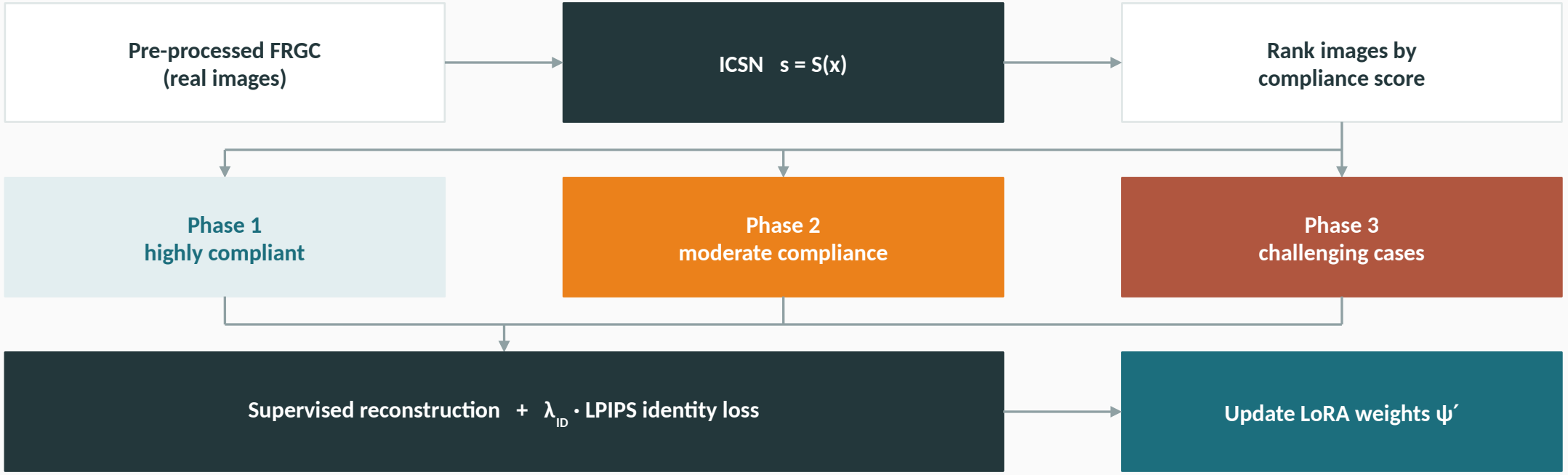
Only the LoRA weights are updated.

The model learns the passport-photo prior: composition, frontal lighting, neutral background, centred head.

Captions are generated with DeepFace (age, gender, ethnicity). They are imperfect, particularly on darker skin tones, and are used as style guidance — never as ground-truth demographic labels.

Stage 2 — learn the order

Real photographs presented easy → hard. The diffusion objective still supplies every gradient.

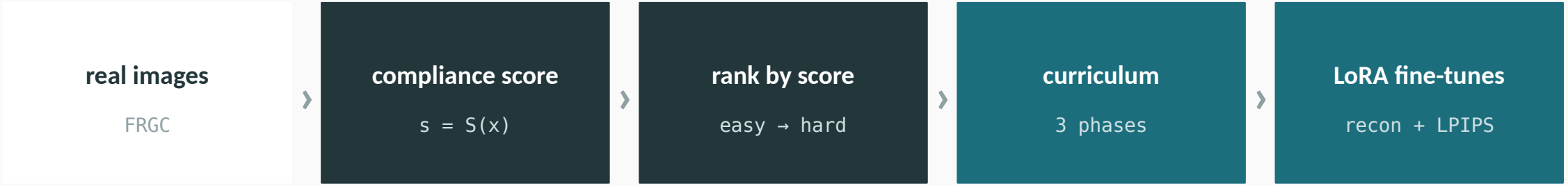


ICSN assigns each real image to a phase. Training walks the phases in order. No gradient passes through the scorer.

LPIPS is used rather than an ArcFace identity loss: ArcFace destabilised training, trading away the pose and framing that ICAO actually tests in order to match the reference embedding.

Score → rank → learn

Stage 2 reduces to this. The checker never becomes the loss — it decides the sequence.



Never differentiated

ICAOnet is used as a black box. No gradient path runs through it.

No targets

It supplies no labels and no regression target — only a relative ranking.

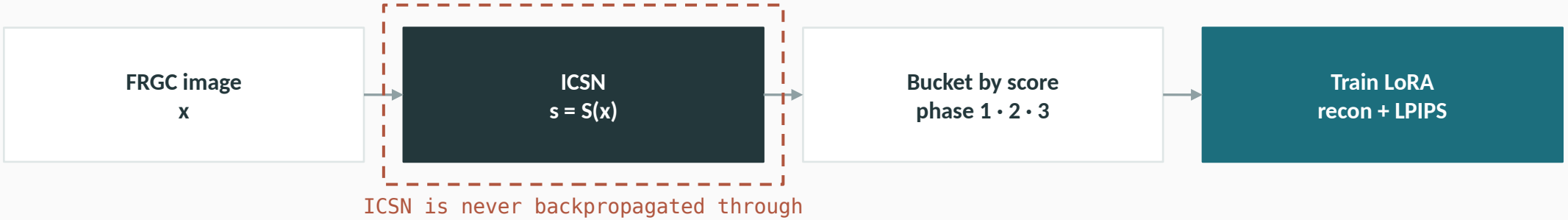
Order only

Its entire influence is the sequence in which real examples are presented.

The model still learns by supervised reconstruction and a perceptual identity term. The checker changes only the order.

Outside the gradient path

The scorer ranks examples; the diffusion objective supplies every gradient.



What ICSN does

- Ranks real examples by compliance
- Sets the three curriculum phases

What ICSN does not do

- Provide gradients
- Provide targets or pass/fail labels

Evaluation is by BioLab-ICAO-Check — closed-source, rule-based, sharing no parameters or data with our pipeline.

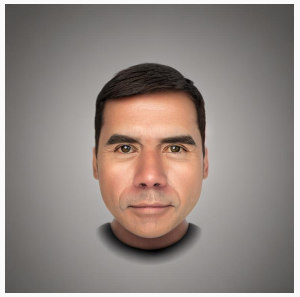
Does the order matter?

We changed one ingredient: curriculum ordering. Same backbone, same Stage-1 initialisation, same FRGC data.

85.72
matched-ablation FID · full model

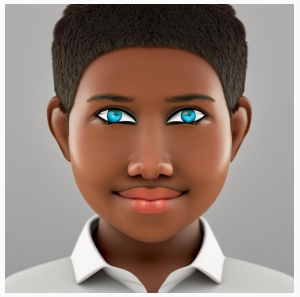
112.45
matched-ablation FID · ordering removed

+26.7 FID
from removing only the ordering



(a) no curriculum

(b) no identity loss



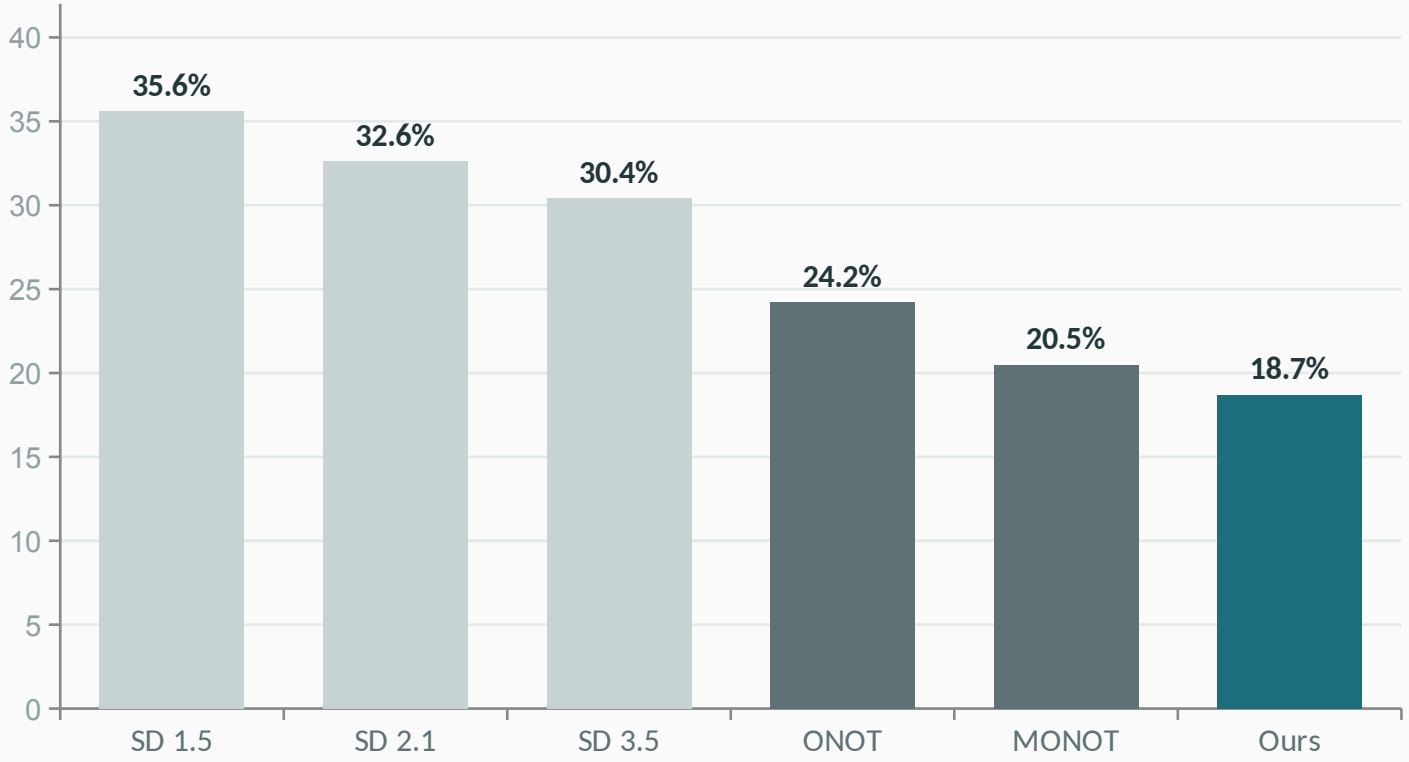
(c) no style pre-training

One ingredient changed. The comparison isolates the ordering itself — not fine-tuning in general, and not the backbone.

Removing the identity loss drops ID similarity 0.753 → 0.461 while FID barely moves — LPIPS preserves identity-relevant structure.

When BioLab checks the images, fewer fail

Mean ICAO failure rate over 23 ISO/IEC 19794-5 tests; lower is better.



18.7%

mean failure rate, against 20.5% for MONOT — the strongest ICAO-oriented baseline.

Welch's t-test on BioLab compliance scores against MONOT

$p < 0.001$

BioLab-ICAO-Check v1.0 · 23 ISO tests · 3,000 images per method. ICAOnet, which builds the Stage-2 curriculum, is disclosed but never used for a reported claim.

The curriculum transfers capture geometry

The largest improvements are in pose, gaze and framing.

29.1
roll / pitch / yaw
best baseline 36.4

22.6
looking away
best baseline 28.2

12.9
hair across eyes
best baseline 15.2

8.1
blurred
best baseline 9.8

These are properties of the source photographs, not of the prompt — exactly what a curriculum over real, correctly-captured images should transfer. Prompt engineering cannot reach them.

What it does not fix

Colour calibration. Unnatural skin tone stays our losing test at 11.4% against 6.1% for SD 1.5 — the curriculum teaches geometry, not colourimetry.

Does compliance collapse diversity?

Constraining a generator this hard could make it produce the same safe face repeatedly.

Text-only generation across demographic prompts.

82.39
benchmark FID against FFHQ

0.58
highest LPIPS-D · among samples

The answer is no

The model becomes more compliant without collapsing to a narrow template.

LPIPS-D measures diversity among generated samples; FID is evaluated against the disjoint FFHQ reference set.

Mode collapse is the obvious risk when a generator is constrained this hard. The diversity number is the evidence against it.

Does compliance erase the person?

A stricter photograph should not become a different identity.

Identity-conditioned: reference → compliant output.

The LPIPS term preserves identity-relevant structure

Removing it sharply reduces identity similarity, 0.753 → 0.461, while barely changing FID, 85.72 → 88.19.

0.753

identity similarity · full model

0.461

without the LPIPS term

MONOT wins on AUC

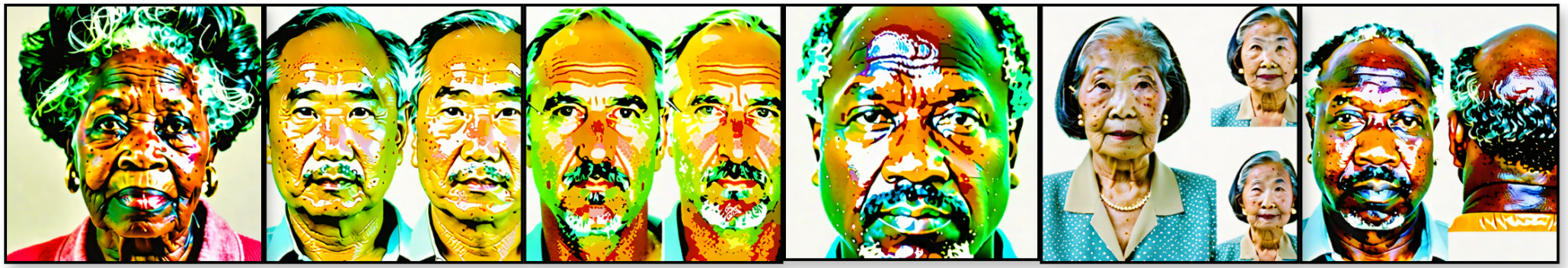
0.792 against our 0.740. We report it rather than hiding it.

Compliance is a mandatory gate: images that fail it are discarded before verification is ever attempted.

What it did not see, it did not learn

The failure cases reveal the boundary of the training signal.

ICAO compliant passport photo of a <*age between 60-80*>-year-old <*ethnicity*> <*gender*>, looking neutral, eyes centered, no shadows, white background, high resolution, realistic faces.



Paper Fig. 7 — a range-valued age prompt collapses photorealism entirely.

Age

Training data sits in the 18–55 band. Quality degrades progressively beyond roughly 50, and worst for the groups thinnest in that data — the curriculum can only teach what its source photographs contain.

Conclusion

One mechanism, evaluated by a checker that shares nothing with our pipeline.

1

Mechanism

A compliance checker used as a black box can drive a curriculum. It supplies no gradient and no target — only the order in which real examples are seen.

2

Compliance

Lowest mean ICAO failure rate of any method compared: 18.7%, at $p < 0.001$ against the strongest ICAO-oriented baseline.

3

Realism & diversity retained

Best FID at 82.39 and highest LPIPS-D at 0.58, while retaining identity similarity of 0.753.

4

Release

≈10k compliant 512×512 images, balanced across gender and ethnicity within the age range the training data supports.

Open: colour calibration, ages beyond roughly 50, and identity AUC against MONOT.

We do not optimise compliance.

We organise learning with it.

score › **order** › **learn**

code · scripts

A checker we never differentiate through can still teach a generator, when it is used as a curriculum signal.

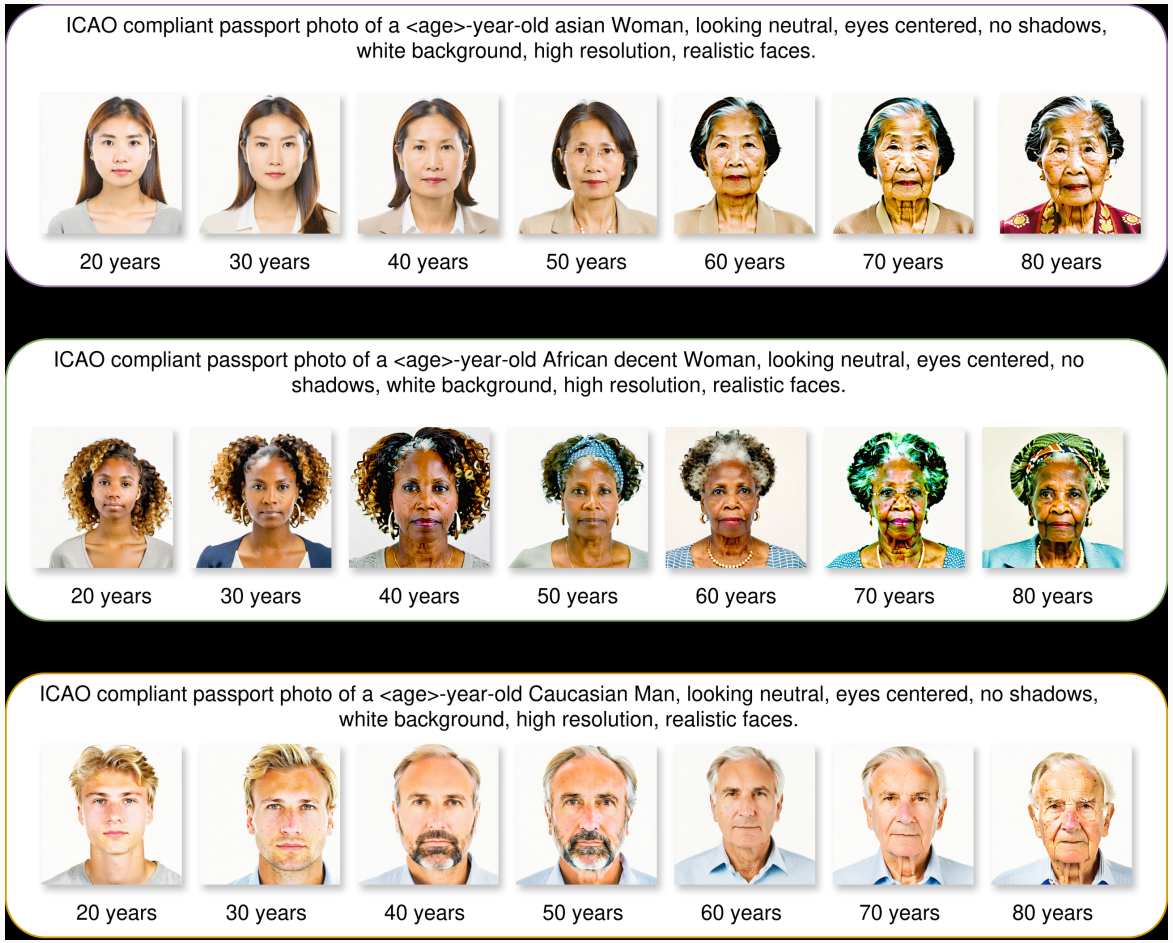
Raghavendra Mudgalgundurao · **NTNU Gjøvik**
raghavem@stud.ntnu.no

Backup

Full numbers, evaluation caveats and Q&A detail

Age control across demographic prompts

Discrete age values, 20 → 80, against three fixed demographic prompts.



What it shows

Age is controllable as a discrete prompt value, not only as a coarse band.

Quality holds well through roughly 50, then degrades progressively.

Degradation is uneven across groups — strongest where the ICAO-compliant training subsets are thinnest, appearing first as colour saturation and texture artefacts.

An earlier figure, later split for the paper. Paper Fig. 7 uses a range-valued prompt and fails harder than these discrete ages.

backup

Per-test ICAO failure rates

BioLab-ICAO-Check, Evaluation A at official thresholds. Lower is better.

ICAO test	ONOT	MONOT	SD 1.5	SD 2.1	SD 3.5	Ours
Blurred	12.3	9.8	21.4	19.6	17.9	**8.1**
Looking away	34.7	28.2	46.5	42.1	39.4	**22.6**
Roll / pitch / yaw	41.9	36.4	55.2	49.7	47.3	**29.1**
Hair across eyes	18.5	15.2	31.7	29.4	27.1	**12.9**
Shadows across face	26.8	22.5	38.9	35.1	33.4	**19.6**
Flash reflection (skin)	31.4	27.9	44.2	41.0	38.6	**24.3**
Flash reflection (lenses)	22.7	19.3	36.8	33.5	31.2	**17.4**
Frames too heavy	29.1	25.7	48.3	45.0	42.6	**21.8**
Mouth open	14.6	12.1	26.4	23.9	21.7	**10.5**
Unnatural skin tone	9.8	8.3	**6.1**	6.7	7.2	11.4
Mean failure rate	**24.2**	**20.5**	**35.6**	**32.6**	**30.4**	**18.7**

backup

Realism, diversity and ablation

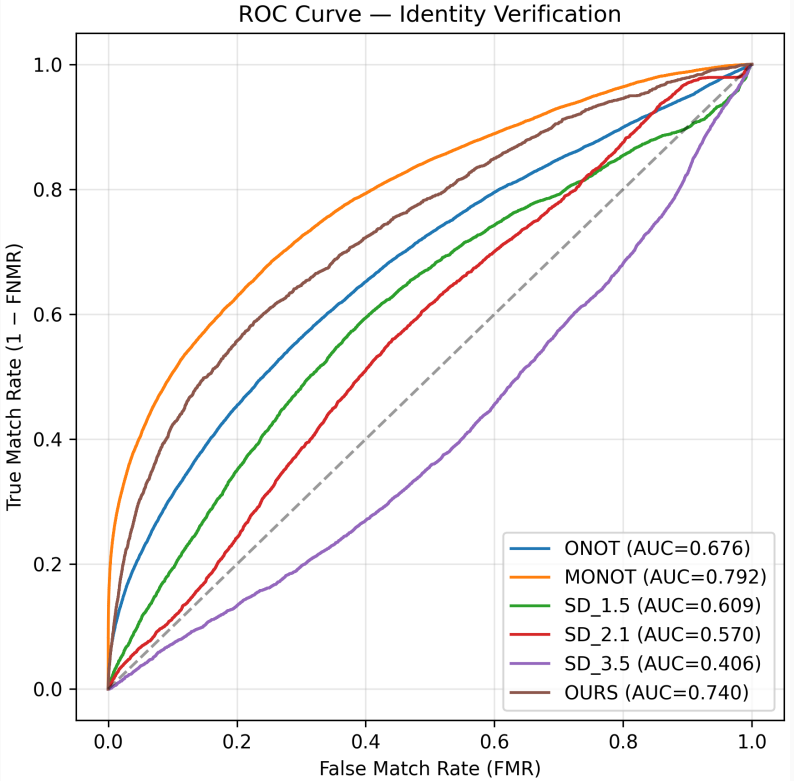
Full comparisons against the FFHQ reference set.

Benchmark · demographic prompt group

Model	FID ↓	LPIPS-R ↓	LPIPS-D ↑	IS ↑
SD 1.5	90.81	0.50	0.49	4.06
SD 2.1	85.65	0.49	0.52	4.31
SDXL Turbo	98.12	0.45	0.30	2.82
SD 3.5	114.98	0.44	0.30	2.82
Ours	**82.39**	**0.43**	**0.58**	**4.49**

Ablation · demographic prompt group

Model variant	FID ↓	ID sim. ↑
Full model (ICSN-guided)	**85.72**	**0.753**
(a) w/o curriculum ordering	112.45	0.749
(b) w/o identity loss	88.19	0.461
(c) w/o style pre-training	125.61	0.658



Identity verification ROC. MONOT AUC 0.792, ours 0.740.

backup

Evaluation caveats

Primary evaluator	BioLab-ICAO-Check v1.0 — closed-source, rule-based, 23 ISO/IEC 19794-5 tests. ICAOnet is never used for a reported compliance claim.
Train / eval separation	FRGC is used in Stage 2, so FID and KID are computed against FFHQ rather than FRGC to avoid leakage.
Invalid BioLab scores	When a test cannot be scored due to face-detection failure the value is treated as zero. Per-test rates are largely unaffected.
Subset size	Randomly selected 3,000-image subsets per dataset, chosen for evaluation tractability.
Identity	ArcFace embeddings via InsightFace. MONOT reaches a higher AUC; the pipeline gate is compliance first.
Open limitations	Age coverage outside 18–55, colour calibration, and identity AUC all remain open.

backup