CAPTURE WHITE-POINT SENSITIVITY IN FROZEN VISION ENCODERS: A CALIBRATED AUDIT ON HUMAN FACES

Anonymous authors
Paper under double-blind review

ABSTRACT

Frozen vision encoders are widely used as feature extractors for color-sensitive tasks such as dermatology, skin-tone auditing, and cosmetics, yet their color constancy under real capture changes has not been evaluated against physical measurements of the imaged surface. We audit twelve frozen encoders on 2257 subjects photographed under a fully crossed 2×2 design of capture white point and illuminance, each subject independently measured by a contact spectrophotometer. Changing white point at fixed illuminance moves embeddings by a large fraction of the distance between different people, along a direction shared across subjects; removing chroma eliminates this structure. Comparing encoder inputs and outputs, the perturbation enters as an almost shared translation and leaves as a subject-dependent residual more distributed than the difference between two people; a single-image gain ablation reproduces this with no pose or metering difference, while the white-point–illuminance interaction is already present at the encoder input at the level a separable multiplicative model predicts. The consequence is measurable in physical units: a skin-tone probe trained at one white point shifts the same person's predicted typology angle by several degrees under another and frequently changes the assigned skin-tone bin. We then compare corrections on one utility–invariance trade-off. Blind white balancing and a diagonal correction fitted to measured skin color do not reliably remove the sensitivity; projecting out a capture subspace estimated from unlabeled calibration captures recovers much of the cross-condition loss while preserving same-condition accuracy, and condition-mean differences estimate that subspace as well as paired displacements; pooling capture conditions when fitting the probe recovers the most.

1 INTRODUCTION

Frozen vision encoders have become a standard tool for visual representation learning. A practitioner can take CLIP (Radford et al., 2021), SigLIP (Zhai et al., 2023), or DINOv2 (Oquab et al., 2023), freeze the encoder, and train a simple linear head for a downstream task. This practice assumes that the learned representation captures properties of the scene rather than properties of the imaging conditions. For tasks such as object recognition or image-text retrieval, this assumption often works well enough that it is rarely tested directly.

The assumption matters most when color itself is the target of prediction—skin-tone estimation for dataset auditing, lesion assessment, cosmetics. The same surface looks different in an image when illumination or camera settings change, so a model may report different colors for a surface that has not changed; human vision largely discounts such changes (color constancy). Whether frozen encoders do is untested, and testing it needs two things corruption benchmarks (Hendrycks & Dietterich, 2019; Geirhos et al., 2018b; Mintun et al., 2021; Kar et al., 2022) do not provide: real captures of the same physical surface under different conditions, and an independent instrumental measurement of that surface as ground truth. Synthetic corruptions scored with semantic labels can show that a prediction changed, but not whether an estimate of the physical surface color stayed correct.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We study this question using a dataset (National Information Society Agency, 2026) that provides the required physical controls: 2257 participants photographed four times under varied capture white point and scene illuminance with camera, pose, and background fixed, each independently measured with a contact spectrophotometer. The measured skin color is fixed for the same person across all four images, so we can ask whether changing only the capture conditions changes the color information recovered from a frozen encoder. It does: capture white point produces a consistent representation shift across participants, and that shift reaches predictions of skin properties that are physically unchanged.

Contributions.

- **A calibrated audit protocol.** Repeated captures of the same person plus independently measured skin color, with paired metrics for the size and cross-subject consistency of representation changes, read against chroma-removed and identity nulls, and the same geometry measured at the encoder's input and output and in a single-image gain ablation that leaves no pose or metering difference (Sections 3 and 5.1 to 5.3).
- **Prediction inconsistency in physical units.** How far the same person's predicted typology angle moves between captures, how often that crosses a skin-tone bin boundary, and for which backbones the harm is one shared shift versus a subject-dependent residual (Section 5.4).
- **A comparison of corrections on one trade-off.** Input-side white balancing, blind and fitted to measured skin color, against post-hoc subspace projection and pooled-condition probe fitting, scored on cross-condition error, same-condition error, identity retrieval, and an unrelated read-out. The measured-gain correction increases the shift for 4 of the 5 language-supervised encoders; a rank-3 condition-mean subspace performs as well as the paired estimator; pooling conditions when fitting the probe performs best (Sections 5.1 and 5.5).

2 Related Work

Robustness benchmarks. Corruption and distribution-shift benchmarks (Hendrycks & Dietterich, 2019; Barbu et al., 2019; Recht et al., 2019; Ibrahim et al., 2022) score category accuracy under synthetic changes; they measure whether a semantic prediction is preserved, not whether a physical property of the scene is. We instead use real recaptures of the same person under controlled capture conditions and evaluate against an instrumental measurement of the skin, so a failure is attributable to a physical quantity.

Color constancy. Classical and learned color constancy (Land & McCann, 1971; Buchsbaum, 1980; Finlayson & Trezzi, 2004; Gijsenij et al., 2011; Barron, 2015; Afifi & Brown, 2020; Li et al., 2026) estimates the illuminant or outputs a corrected image; Afifi & Brown (2019) show synthetic white-balance errors harm classification and segmentation. We ask the complementary question: whether a frozen representation never asked to be color constant behaves as if it were, measured on real recaptures against an instrumental target. Gray-world and shades-of-gray serve as the baselines a practitioner tries first, and because cue sensitivity depends on the pretraining objective (Geirhos et al., 2018a; Hermann et al., 2020; Zini et al., 2022), we compare correction effects across encoder families.

Skin tone in vision systems. Skin-tone audits commonly group by Fitzpatrick type (Buolamwini & Gebru, 2018; Groh et al., 2021), which is not a physical measurement, and Thong et al. (2023) stress the gap between apparent image color and the subject's skin color. Our corpus separates the two with a per-subject instrumental measurement. Post-hoc removal of unwanted directions from representations is standard (Bolukbasi et al., 2016; Ravfogel et al., 2020; Sun et al., 2016); here the removed subspace is estimated from controlled recaptures and evaluated against a measured physical target. Section B discusses all of this related work in full.

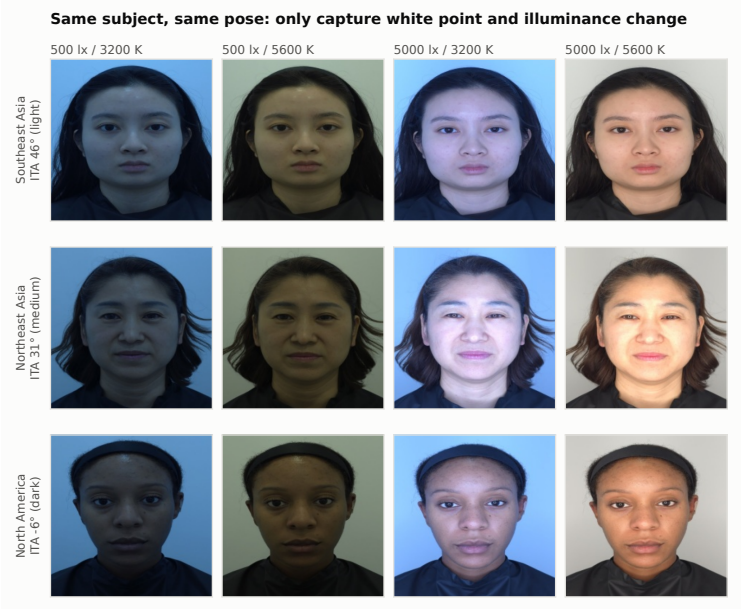

Figure 1: Three subjects spanning the ITA bins (rows) under the fully crossed 2×2 capture design (columns): white point 3200 K vs 5600 K, illuminance 500 lx vs 5000 lx. Pose, background, device, and crop are fixed per subject; row labels show the spectrophotometer reference measured once per subject.

3 Dataset

3.1 Corpus and Capture Design

We use the AI Hub *Global Multi-Ethnic Skin Color* corpus (National Information Society Agency, 2026). It contains 2257 subjects across six regional groups. Each subject was photographed four times using one device, recorded as `GALAXY 24 Ultra`, with a fixed pose and background. All subjects have all four captures, giving 9028 images with no missing data.

The four images follow a fully crossed 2×2 capture design. The recorded white point is either 3200 K or 5600 K, corresponding to conditions `*32K` and `*56K`. Illuminance is either 500 lx or 5000 lx, corresponding to `5L*` and `5KL*`. Figure 1 shows this design for three subjects. The person, pose, background, and device remain fixed within each row while the captured skin color changes visibly across conditions.

We treat the white-point setting as a *capture-side* factor rather than a change in the scene illuminant. The 3200 K images have a characteristic blue cast, and the transformation between white-point conditions is approximately diagonal in RGB. The measured skin gain is $(1.951, , 1.066, , 0.387)$ in linear RGB (Section A). We therefore do not interpret this factor as an isolated change of the physical illuminant.

3.2 Physical Reference and Image Preparation

Each subject has an independent physical measurement of skin color. The forehead, cheeks, and neck were measured using a contact spectrophotometer, recorded as `Spectrophotometer CM26d`. The measurements provide $L^*a^*b^*$ values and the individual typology angle

$$\text{ITA} = \arctan((L^* - 50)/b^*) \cdot 180/\pi.$$

This reference is measured independently of the four photographs and is therefore fixed across capture conditions. It provides a condition-invariant target for evaluating the representations. The cor-

pus releases, alongside the per-site values, a single subject-level ITA (field `ITA Avg.`) averaged over the measured sites; that subject-level average is the prediction target everywhere in this paper.

The corpus also provides a field named `Fitzpatrick`. Despite its name, this field is a deterministic binning of instrumental ITA at thresholds $55, 41, 28, 10, -30$, with no exceptions among the 2257 subjects (Section K). Fitzpatrick phototype instead describes skin response to sun exposure and cannot be determined from a reflectance measurement alone. We therefore refer to the provided categories as **ITA bins**. Flip rates (Sections 5.4 and 5.5) use these six threshold-defined bins; the corpus's three-way grouping—light (469 subjects), middle (1609), dark (179), adjacent bins aggregated pairwise—appears only in corpus description and the group-conditional analyses (Section 5.4, Section G).

Finally, we use a shared face crop across each subject's four images, since per-image detection could move the crop with the capture condition. We detect with YuNet (Wu et al., 2023) in all four captures, take the per-subject median box, convert it to a square of side $1.6 \times \max(w, h)$, and apply the same box to all four images before resizing to 224^2 (detection succeeds for every subject). Each backbone then applies its released center-crop configuration.

4 METHOD

4.1 BACKBONES AND FEATURES

We evaluate twelve frozen encoders at 224^2, spanning objective and scale: language-supervised (CLIP-B/16 and CLIP-L/14 (Radford et al., 2021), OpenCLIP-B/16, SigLIP-B/16 and SigLIP-SO400M (Zhai et al., 2023)), self-distilled (DINOv2-B/14 and DINOv2-L/14 (Oquab et al., 2023)), masked (MAE-B/16 (He et al., 2022), EVA02-B/14 (Fang et al., 2024)) and label-supervised (ViT-B/16, ResNet-50, ConvNeXt-B). All are loaded through one library (Wightman, 2019) using each model's own released preprocessing configuration, resized to a common 224^2 input, and no weights are updated anywhere in this paper. Read-out heads are whatever each model released: a class token for CLIP, OpenCLIP and DINOv2, a learned attention-pooling head for SigLIP, and mean pooling for MAE and EVA02. For CLIP and OpenCLIP we read the backbone's final-norm class token *before* the learned contrastive projection, i.e. internal backbone features rather than the released joint embedding; Section Q verifies the principal findings on the projected outputs as well. For MAE we use mean-pooled patch tokens: MAE's class token has no dedicated objective and is trained only through the reconstruction pathway, and mean pooling is the common probing choice. Inference is in half precision. Linear probing of frozen features follows standard practice (Alain & Bengio, 2016); the probe protocol and split design are in Section J.

4.2 METRICS

Write $\mathbf{z}_i^c \in \mathbb{R}^d$ for the embedding of subject i under condition c and $\boldsymbol{\delta}_i = \mathbf{z}_i^{c_2} - \mathbf{z}_i^{c_1}$ for the displacement induced by a contrast. Two quantities carry the argument:

$$\mathrm{NPS} = \frac{\mathbb{E}_i \|\boldsymbol{\delta}_i\|}{\mathbb{E}_{i \neq j} \|\mathbf{z}_i^{c_1} - \mathbf{z}_j^{c_1}\|}, \qquad \rho = \mathbb{E}_{i \neq j} \cos(\boldsymbol{\delta}_i, \boldsymbol{\delta}_j). \tag{1}$$

$\mathrm{NPS} = 1$ means the capture change moves the embedding as far as replacing the person. Coherence ρ is what makes the paired design pay off: a perturbation whose per-subject direction is independent and symmetric about zero contributes zero in expectation however large it is. Independence alone does not suffice, since displacements sharing a common mean direction would register as coherence, so the floor is measured rather than assumed: Section 4.4 strips chroma from the same capture pair and reports the coherence that remains, which is what the photometric difference between two shots alone produces. We also report cross-condition identity retrieval (gallery and query are the same people under two conditions; no interval, since top-1 depends on gallery size). Subjects, not images, are the unit of independence, so all other intervals are 95% subject bootstrap ($B = 1000$); medians over backbones describe this model set, and every claim about spread is a count.

4.3 FACTORIAL CONTRASTS AND A FIRST-ORDER HYPOTHESIS

The 2×2 design lets us ask not only whether representations move but whether they move as a simple physical model predicts. Under the von Kries hypothesis a white-point change acts on linear-RGB radiance as a per-channel gain $x \mapsto x \odot \mathbf{g}$, the assumption behind essentially every practical white-balance algorithm and one Section 3 confirms holds approximately for skin here. Expanding a frozen encoder to first order in $\mathbf{g} - \mathbf{1}$ (Section H) yields three diagnostic hypotheses, each of which we measure. **(1) Concentration:** for a fixed image x the first-order displacement lies in the span of three vectors $\mathbf{J}(x)\,(x \odot \mathbf{e}_c)$. That span moves with the image, so whether the *population* of displacements is dominated by a shared direction is an empirical question, measured against matched nulls; the operating rank $k{=}4$ used throughout is an empirically chosen point (Section H). **(2) Transfer:** a capture subspace belongs to the encoder rather than the individual, so one estimated on some subjects should capture unseen subjects' displacements, which is what a post-hoc correction needs. **(3) Additivity:** to first order the two factors act as independent translations. The benchmark is not zero: exactly separable multiplicative factors already yield $\|WL\|/\|W\| = (s-1)/(s+1)$ under Equation (2) (Section N), and measured ratios are read against that value.

We estimate the two main effects and the interaction per subject in the standard way,

$$W_i = \tfrac{1}{2}\big[(\mathbf{z}_i^{\mathrm{lo},56} - \mathbf{z}_i^{\mathrm{lo},32}) + (\mathbf{z}_i^{\mathrm{hi},56} - \mathbf{z}_i^{\mathrm{hi},32})\big], \quad L_i = \tfrac{1}{2}\big[(\mathbf{z}_i^{\mathrm{hi},32} - \mathbf{z}_i^{\mathrm{lo},32}) + (\mathbf{z}_i^{\mathrm{hi},56} - \mathbf{z}_i^{\mathrm{lo},56})\big],$$
$$WL_i = \tfrac{1}{2}\big[(\mathbf{z}_i^{\mathrm{hi},56} - \mathbf{z}_i^{\mathrm{hi},32}) - (\mathbf{z}_i^{\mathrm{lo},56} - \mathbf{z}_i^{\mathrm{lo},32})\big], \tag{2}$$

so prediction 3 is the statement $\|WL\| \ll \|W\|$ and we report the ratio directly.

4.4 NULL MODELS AND CONTROLS

A displacement population can look low-dimensional for trivial reasons, so every rank statistic is read against three controls in the same space: differences between two *different* people under the *same* condition (the anisotropy of the space itself, the null that matters), the same contrast with the pairing permuted, and a matched-scale isotropic Gaussian. Separately, a *before*-measurement repeats the geometry on the decoded sRGB pixels the encoder is handed, at the same 224^2, placing the boundary between camera and network where we want it. Section H defines all four; Table 7 carries the identity null, the rest are in the released measurements.

Two further controls isolate the color channel and bound input-side remedies. The first feeds the encoder luminance only, so no chroma reaches it. The second is a **measured-gain correction**: per condition we apply the inverse of the per-channel gain fitted to real skin in that condition, mapping the four captures onto a common skin white point. This is a global diagonal gain and therefore much stronger than a blind estimator, but it is not an inverse of the camera pipeline, since it cannot undo per-subject or spatial variation, the tone curve, clipping, or the sensor color matrix. We invert only the unit-geometric-mean chromatic part of each measured gain, discarding its scalar factor (0.93 at fixed illuminance; large across illuminance levels); since unit geometric mean does not exactly preserve luminance, the correction is geometric-mean-normalized rather than strictly luminance-preserving (Section K).

4.5 CAPTURE-SUBSPACE PROJECTION

On calibration subjects only we form the paired displacements over the four single-factor contrasts, take the top-k right singular vectors $\mathbf{U} \in \mathbb{R}^{d \times k}$, and at inference replace every embedding by $\mathbf{z}' = \mathbf{z} - \mathbf{U}\mathbf{U}^\top \mathbf{z}$: no weights change, no task labels, and the query's condition is never needed. The subspace spans the response to both factors (a *capture-condition* subspace). This is post-hoc calibration: it requires multi-condition captures of unlabeled calibration subjects, and the probe is refitted on projected features.

5 EXPERIMENTS

5.1 SENSITIVITY OF FROZEN REPRESENTATIONS

We first measure how much frozen representations change when only the capture white point changes. At fixed illuminance, the displacement ranges from 0.35 to 0.88 of the mean distance

between different identities (Table 6). The displacement is also consistent across subjects, with coherence ranging from 0.31 to 0.67. This change is large enough to affect identity structure. When the gallery and query use different white points, top-1 identity retrieval falls to 0.99–0.37.

The effect is not a change of feature norm: 87–100% of the displacement energy is orthogonal to the original embedding, and repeating the analysis after L2 normalization changes nothing (Section M).

The effect is primarily chromatic: removing chroma cuts displacement by a median of 58%, and—decisively—coherence falls to at most 0.044 and retrieval returns to at least 0.992 across the eleven non-degenerate backbones, so what survives without chroma is not shared across subjects and is explained by residual differences between two physical captures rather than a common color shift.

Standard white-balance corrections do not reliably remove the effect: shades-of-gray reduces displacement by only 11% at the median and increases it for 4 of 11 backbones, and even the measured-gain correction helps 5 of 11 encoders while making 4 worse, language-supervised models most often (Table 6). The measured-gain correction is a global diagonal gain, so what the encoder responds to after correction is the input discrepancy it leaves behind: tone mapping, clipping, spatial processing, and a background receiving a gain fitted to skin (Figure 11). Its failure therefore bounds diagonal input corrections, not input-side correction in general.

Finally, a linear classifier recovers which white point produced an embedding with accuracy at least 0.9996 for every backbone (chance 0.5). Retaining capture information is not itself a failure for a general-purpose representation; whether it reaches properties that should stay constant across captures is tested in Section 5.4.

5.2 Geometry of the Capture Response

We next examine the geometry of the white-point displacement (Figure 6, Table 7). At the encoder input, the mean-direction fraction is 0.86 and coherence is 0.87. This is consistent with a largely shared color transformation across subjects. After encoding, the mean-direction fraction falls to 0.38–0.71 and coherence falls to 0.31–0.67. The encoder therefore turns part of a common input shift into subject-dependent changes in representation space.

These subject-dependent changes are not confined to a few directions: after removing the shared direction, the residual has participation ratio 21–69 against a median of 15 for the identity control, higher for 11 of 11 backbones. The centered *input* residual has participation ratio 32.0 (Section K), which the output residuals straddle; the change through the encoder is therefore a loss of dominance of the shared component rather than a uniform increase in residual dimensionality. Enough shared structure remains to be removable: a rank-4 subspace fit on half the subjects captures 0.48–0.79 of held-out subjects' displacement energy, against 0.45 for the identity control.

White point and illuminance also interact rather than acting as independent additive factors. The interaction has 0.50–0.89 times the norm of the white-point main effect at the output. An interaction of 0.34 is already present at the encoder input, and a nonzero value there is expected: for finite, exactly separable multiplicative factors the contrasts of Equation (2) give $\|WL\|/\|W\| = (s-1)/(s+1)$, which at the measured $7.1\times$ illuminance scale is ≈ 0.42 after the display encoding (Section N). The measured input value sits at or below that benchmark, so the input-level non-additivity is within what the multiplicative structure of the capture change implies; the encoder receives a substantial interaction from image formation and modifies it in a model-dependent way, with the representation-level change strongest in the final third of the networks we trace (Section U).

Two conditions bound this reading. The crops retain background, which changes color with the white point; a region-masked ablation shows the shared, coherent component is carried substantially by background, so these statistics characterize the response to the whole crop rather than to skin alone (Section P). And a falling shared fraction is not in itself harmful, since it also arises when an encoder suppresses the common nuisance more strongly than the rest; which part drives the downstream harm is measured per backbone in Section R. The controlled ablation below isolates the encoder's contribution.

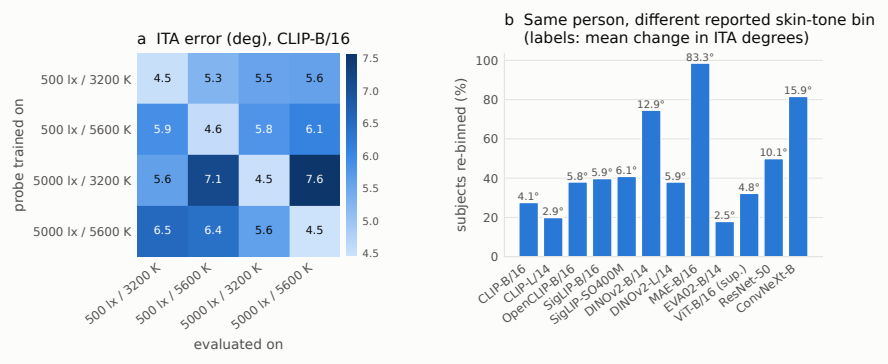

Figure 2: **(a)** ITA prediction error across train and test conditions (diagonal: same-condition evaluation). **(b)** ITA-bin flip rate under the white-point contrast, with each backbone's mean change in predicted ITA.

5.3 Ablation Study: Isolating the Encoder

The paired captures contain small differences beyond color (pose, expression, metering). To isolate the encoder's part we construct all four conditions analytically from a single image per subject (800 subjects), so no capture-to-capture variation exists between cells. Cells are built in linear RGB from the `5L56K` capture and re-encoded to 8-bit sRGB; the high-multiplier cells clip 35–53% of pixels in at least one channel, as the real 5000 lx captures also clip. The ablation is a magnitude-matched stress test of the encoder's response to an exactly known gain, not a reconstruction of the four real captures (Section N).

Arm A applies the same chromatic gain at both luminance levels with a $2\times$ luminance scale, making the factors separable by construction; **Arm B** uses the measured per-level gains and the measured $7.1\times$ scale (Table 3). A matched-multiplier design crosses {shared, level-dependent} gains with $\{2\times, 7.1\times\}$ multipliers so that the two factors can be read separately (Table 4).

Loss of shared structure. In Arm A an input with mean-direction fraction 0.93 and coherence 0.93 leaves the encoder at 0.53–0.80 and 0.54–0.80, and Arm B shows the same pattern. With a single base capture, no pose or metering difference can explain this, so this part of the effect is the encoder's.

Interaction. The input ratios, 0.15 in Arm A and 0.43 in Arm B, match what a perfectly separable multiplicative model predicts under the two multipliers $((s'-1)/(s'+1)$ with $s'=s^{1/2.2}$: ≈ 0.16 and $\approx 0.42)$. In the matched-multiplier design, raising the multiplier under a shared gain reproduces the full Arm A/B difference ($0.146 \to 0.447$), while at a fixed multiplier the measured level dependence does not increase the input ratio ($0.146 \to 0.052$ at $2\times$, $0.447 \to 0.434$ at $7.1\times$); without the sRGB clip the high-multiplier values fall to 0.39 and 0.28. Finite multiplicative change, the display nonlinearity, and clipping thus suffice for the observed input interaction. Because the ratio is normalized, this is a population-level statement: level dependence contributes a vector to the interaction contrast (here partially cancelling) without raising the ratio. What the encoders add is a model-dependent modification of the interaction they receive (0.7–$2.9\times$ in Arm A, 0.7–$1.9\times$ in Arm B), similar under shared and level-dependent gains at a fixed multiplier.

5.4 Downstream Impact on Skin-Tone Estimation

We first test whether frozen features predict instrumental skin tone at all: a ridge probe (subject-disjoint folds, inner-split regularization) reaches a within-condition median MAE of $4.76°$ against $11.54°$ for a mean predictor, with all 11 non-degenerate backbones beating that baseline. A direct color-measurement baseline calibrates what follows: raw skin-patch RGB statistics match the frozen features within condition ($5.73°$) but collapse across conditions ($60°$ mean, 96% bin flips), while a condition-calibrated version reaches $6.27°$ (Section S)—so the frozen features perform substantial yet incomplete color constancy, and the failures below are the incomplete part.

The problem appears when the capture condition changes. Testing a probe under a different condition increases MAE by 25–182%, with DINOv2-L/14 showing the largest increase (Table 8). Every off-diagonal entry in Figure 2a has higher error than the corresponding same-condition evaluations. The change is also systematic rather than only an increase in prediction variance: depending on the backbone, the population-mean prediction shifts by as much as $15.95°$, and the degradation concentrates on the warm–cool axis b^* rather than on lightness (Section O).

Comparing predictions for the same subject measures the inconsistency directly: between `5L32K` and `5L56K` the mean within-subject change is 2.53–$15.95°$ across the non-degenerate backbones, enough to move 17.9–81.6% of subjects across a boundary of the six threshold-defined ITA bins (Figure 2b). The physical measurement is unchanged between the two captures, so the model assigns meaningfully different skin-tone estimates to the same person because a camera setting changed.

Finally, we do not find consistent evidence that the effect is larger for darker skin: the relative bias shift between the dark and light groups (Section 3) ranges from 0.44 to 2.28 across backbones with no group consistently more affected. Capture conditions reshape the predicted skin-tone distribution, but our results do not support a disparity concentrated at one end of the range (Section G).

5.5 Post-hoc subspace correction

We next ask which corrections recover the cross-condition loss of Section 5.4, and at what cost. Removing directions may improve invariance but can also remove useful information, so every method is scored on four axes: cross-condition ITA error, same-condition ITA error, identity retrieval, and an unrelated age read-out. We compare the capture-subspace projection of Section 4.5 against gray-world, shades-of-gray, the measured-gain correction, and an oracle that removes each condition's feature-space mean using the test-time condition label (Table 9 and Fig. 4), and then against three matched alternatives that use the same calibration data: a condition-mean subspace, a rank-matched random projection, and a pooled-condition probe (Table 1). All results refit the probe on the corrected features; the fixed-probe setting is reported separately below.

At $k = 4$, the projection reduces the median cross-condition ITA error over the eleven non-degenerate backbones from $7.84°$ to $6.09°$ (22% reduction of the median; the median of per-backbone reductions is 17%), and the ITA-bin flip rate from 38% to 20%. Over 20 subject-disjoint calibration/evaluation splits, same-condition error changes by a median of $0.02°$ (worst $1.27°$) and the age read-out by 0.000 (worst -0.017; Table 9's single split, $0.647 \rightarrow 0.637$, lies within this range). Cross-condition retrieval improves from 0.760 to 0.996 at the median and never falls on any split. The projection slightly beats the oracle without being told the condition. Its calibration requirement is small: 64 unlabeled multi-condition calibration subjects give $6.08°$ against $6.09°$ with the full 1128 (Section J).

The gain is backbone-dependent. Per-backbone medians over the 20 splits run from 6% (CLIP-L/14) to 54% (DINOv2-B/14); 9 of 11 backbones improve by more than 2% on every split, and the remaining two (OpenCLIP-B/16, ResNet-50) gain at the median but not on every draw (worst split -5%; Table 10 shows one representative split, on which ResNet-50 worsens). The subspace should therefore be validated per backbone.

Table 1 shows what each component of the correction contributes. Projecting out the rank-3 span of calibration condition-mean differences, which needs the same calibration images and no task or query-condition labels, matches the paired estimator at matched rank on every axis; the projection works by removing shared condition shifts, and the simpler estimator suffices. A rank-matched random projection changes the median error by 0%. The strongest correction requires no representation editing at all: a probe fitted on all four captures of its labeled training subjects reaches $4.97°$, gaining on every backbone on every split. Its data requirement is multi-condition captures of the labeled subjects (the ITA label is per subject, so no additional labeling), whereas the projections need multi-condition captures only of separate, unlabeled calibration subjects and correct the representation for any downstream use.

Inserting a projection in front of a probe already trained on uncorrected features is a different setting: for a linear probe it is equivalent to removing $\mathbf{U}$ from the probe's weights, so it degrades any probe that relies on directions inside the capture subspace. The median reduction there is 7% for the paired projection and 4% for the condition-mean projection, and SigLIP-B/16 degrades under both (21%

Table 1: Matched mitigation comparison over the same 20 subject-disjoint splits, folds, and regularization grid. Entries are medians over the eleven non-degenerate backbones of per-model split medians; deltas are against *none*; "every split" counts backbones improving on all 20 splits. Projections need multi-condition captures of unlabeled calibration subjects only; the pooled probe needs multi-condition captures of the labeled training subjects. ResNet-50 gains 23% even from the random projection, so its improvement under any projection is not capture-specific; fixed-probe deployment (medians 7%/4%, SigLIP-B/16 worsens under both) is in Section E.

method	cross ITA (°)	median red.	every split	Δ same-cond (°)	Δ retrieval	Δ age
none	7.16	—	—	—	—	—
paired projection, k=4 (Section 4.5)	5.83	17%	9/11	+0.02	+0.22	−0.000
paired projection, k=3	5.81	15%	10/11	+0.01	+0.22	—
condition-mean projection, rank 3	5.91	15%	10/11	+0.01	+0.21	−0.000
random projection, rank 4	7.17	0%	0/11	−0.00	+0.00	—
pooled-condition probe (no editing)	4.97	29%	11/11	−0.03	—	—

at the split median). The correction therefore belongs in probe fitting. All comparisons validate transfer to held-out subjects under the same four capture conditions (Section E).

6 Discussion

The first-order hypothesis. Of the three diagnostic hypotheses of Section 4, transfer holds: a capture subspace estimated on some subjects captures held-out subjects' displacements well above the identity null. Concentration holds as an empirical characterization: a dominant shared input component becomes less dominant in feature space, while enough of it transfers across subjects to be removable. Additivity fails at the encoder output when read against the separable-multiplicative benchmark, and the input-side interaction is at the level that benchmark predicts. The attribution the controls support is that the encoder owns the loss of shared structure, reproduced with no pose or metering difference, while the interaction arrives with the decoded pixels and is modified, not originated, by the encoder (Section F).

Practical implications. Validate any color read-out from a frozen encoder *across capture conditions*, and measure white-balance preprocessing before adopting it. For mitigation (Section 5.5): where the labeled subjects have multi-condition captures, pool them when fitting the probe (strongest here); otherwise estimate the capture subspace from unlabeled multi-condition calibration captures (condition-mean differences suffice, and 64 subjects are enough) and apply it when the probe is fitted; where each image's condition is known, a calibrated direct color measurement rivals both (Section S). All are validated on held-out subjects under the same four capture conditions.

Limitations. All measurements come from one camera, one studio protocol, and faces; the strongest input-side baseline is a per-condition diagonal gain, and the source of the residual it leaves is not decomposed here; and the projection requires multi-condition calibration captures and probe refitting (Section G).

7 Conclusion

Under this capture protocol the frozen encoders we test are not color constant on faces: a change of capture white point moves the same person's predicted skin tone by several degrees of typology angle and re-bins a substantial fraction of subjects on an instrumentally measured property. The encoder turns a shared input translation into a subject-dependent residual, while the factors' non-additivity is already in the pixels at the level their multiplicative structure predicts. Input-side white balancing does not reliably remove the sensitivity; removing a capture subspace estimated from unlabeled calibration captures does when the probe is fitted on the corrected features, condition means estimate that subspace as well as paired displacements, and fitting the probe on pooled conditions recovers the most. The protocol, paired metrics, and matched correction comparison apply directly to any second camera or capture pipeline with repeated captures and an instrumental reference.

AI USE DISCLOSURE

In this work, we used generative AI tools extensively as part of the research and writing process. Generative AI assisted with brainstorming and developing initial research ideas and hypotheses. The AI-generated hypotheses and suggestions were subsequently critically evaluated and substantially revised by the authors, who determined the final research direction and hypotheses.

We also used generative AI to assist with drafting and editing the manuscript and with developing code used to implement and conduct the experiments. All AI-assisted code, experimental procedures, outputs, and manuscript content were reviewed and verified by the authors. The authors independently searched for, reviewed, and selected the scholarly literature cited in this paper; generative AI was not relied upon to generate or verify citations.

The authors take full responsibility for the final research design, experiments, results, claims, citations, and contents of this work, including any text, code, ideas, or other artifacts produced with the assistance of generative AI.

REPRODUCIBILITY STATEMENT

The corpus is publicly available (National Information Society Agency, 2026); the instrumental reference is its released contact spectrophotometer measurement (per-site $L^*a^*b^*$ and the averaged ITA, Section 3), and the ITA binning of its `Fitzpatrick` field is verified in Section K. All backbones are publicly released checkpoints evaluated frozen with their own preprocessing (Section 4); the shared per-subject crop, the skin-patch geometry, the measured gains, and the exact construction of every synthetic cell are specified in Sections 3, K and N. All splits are subject-disjoint, the subspace estimator of Section 4.5 never sees an evaluation subject, and the probe protocol, regularization grid, and calibration/evaluation allocation are given in Section J. Reported intervals are subject bootstrap with $B = 1000$, except cross-condition retrieval, for which no interval is reported because top-1 depends on gallery size; wherever a range or median is quoted, the model set it is taken over is named, and the degenerate backbone of Section V is excluded from headline ranges and reported separately. We release the extraction, analysis, and figure code together with the derived per-model measurements as an anonymized supplementary archive.

REFERENCES

Mahmoud Afifi and Michael S Brown. What else can fool deep learning? addressing color constancy errors on deep neural network performance. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 243–252, 2019.

Mahmoud Afifi and Michael S Brown. Deep white-balance editing. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1394–1403. IEEE, 2020.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. *Advances in neural information processing systems*, 32, 2019.

Jonathan T Barron. Convolutional color constancy. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 379–387, 2015.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29, 2016.

Gershon Buchsbaum. A spatial processor model for object colour perception. *Journal of the Franklin institute*, 310(1):1–26, 1980.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pp. 77–91. PMLR, 2018.

Alain Chardon, Isabelle Cretois, and Colette Hourseau. Skin colour typology and suntanning pathways. *International journal of cosmetic science*, 13(4):191–208, 1991.

Sandra Del Bino and FJBJoD Bernerd. Variations in skin colour and the biological consequences of ultraviolet radiation exposure. *British Journal of Dermatology*, 169(s3):33–40, 2013.

Mark D Fairchild. *Color appearance models*. John Wiley & Sons, 2013.

Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva-02: A visual representation for neon genesis. *Image and Vision Computing*, 149:105171, 2024.

Graham D Finlayson and Elisabetta Trezzi. Shades of gray and colour constancy. In *Color and imaging conference*, volume 12, pp. 37–41. Society of Imaging Science and Technology, 2004.

Thomas B Fitzpatrick. The validity and practicality of sun-reactive skin types i through vi. *Archives of dermatology*, 124(6):869–871, 1988.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018a.

Robert Geirhos, Carlos RM Temme, Jonas Rauber, Heiko H Schütt, Matthias Bethge, and Felix A Wichmann. Generalisation in humans and deep neural networks. *Advances in neural information processing systems*, 31, 2018b.

Arjan Gijsenij, Theo Gevers, and Joost Van De Weijer. Computational color constancy: Survey and experiments. *IEEE transactions on image processing*, 20(9):2475–2489, 2011.

Matthew Groh, Caleb Harris, Luis Soenksen, Felix Lau, Rachel Han, Aerin Kim, Arash Koochek, and Omar Badri. Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1820–1828. IEEE, 2021.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 15979–15988. IEEE, 2022.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

Katherine Hermann, Ting Chen, and Simon Kornblith. The origins and prevalence of texture bias in convolutional neural networks. *Advances in neural information processing systems*, 33:19000–19015, 2020.

Mark Ibrahim, Quentin Garrido, Ari Morcos, and Diane Bouchacourt. The robustness limits of sota vision models to natural variation. *arXiv preprint arXiv:2210.13604*, 2022.

Oğuzhan Fatih Kar, Teresa Yeo, Andrei Atanov, and Amir Zamir. 3d common corruptions and data augmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18941–18952. IEEE, 2022.

Edwin H Land and John J McCann. Lightness and retinex theory. *Journal of the Optical society of America*, 61(1):1–11, 1971.

Shuwei Li, Lei Tan, and Robby T Tan. White-balance first, adjust later: Cross-camera color constancy via vision-language evaluation. *arXiv preprint arXiv:2605.19613*, 2026.

Eric Mintun, Alexander Kirillov, and Saining Xie. On interaction between augmentations and corruptions in natural corruption robustness. *Advances in Neural Information Processing Systems*, 34:3571–3583, 2021.

National Information Society Agency. Ai hub dataset no. 71963. AI Hub, 2026. URL https://aihub.or.kr/aihubdata/data/view.do?dataSetSn=71963. Accessed: 2026-09-03.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

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

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 7237–7256, 2020.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pp. 5389–5400. PMLR, 2019.

Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.

William Thong, Przemyslaw Joniak, and Alice Xiang. Beyond skin tone: A multidimensional measure of apparent skin color. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4880–4890. IEEE, 2023.

Ross Wightman. PyTorch image models, 2019. https://github.com/huggingface/pytorch-image-models.

Wei Wu, Hanyang Peng, and Shiqi Yu. Yunet: A tiny millisecond-level face detector. *Machine Intelligence Research*, 20(5):656–665, 2023.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11941–11952. IEEE, 2023.

Simone Zini, Alex Gomez-Villa, Marco Buzzelli, BartŁ Twardowski, Andrew D Bagdanov, Joost Van de Weijer, et al. Planckian jitter: countering the color-crippling effects of color jitter on self-supervised training. *arXiv preprint arXiv:2202.07993*, 2022.

A Identifying the Two Capture Factors

The corpus designates the color settings as $3200\,\mathrm{K}/5600\,\mathrm{K}$, recorded per image alongside exposure, ISO and shutter speed. Two observations lead us to describe them as a *capture-side* white-point change rather than a change of lamp: the $3200\,\mathrm{K}$ images carry a blue cast, which is the sign of chromatic adaptation set warmer than the scene; and the transformation between them is diagonal to a good approximation, with a measured skin gain of $(1.951,\ 1.066,\ 0.387)$ in linear RGB. We therefore write *dataset-designated capture conditions, empirically consistent with a capture-side white-point transformation*, and claim no isolation of a scene-illuminant change. The illuminance factor is more entangled, shutter speed, tone mapping and clipping move with it, so we treat it as a *photometric* factor and never pool the two into one "lighting" axis.

Is the map really diagonal? We test it rather than assume it. For each contrast we fit both a diagonal gain and an unrestricted 3×3 linear map on per-subject skin radiance in linear RGB, train on half the subjects and score the relative residual on the other half. For the white-point contrast at fixed illuminance the diagonal map leaves a residual of 0.041 and the full matrix only 0.034, with 1.1% of the fitted matrix's energy off the diagonal, so the six extra degrees of freedom buy little and the von Kries description is adequate. The illuminance contrast is a different story: both maps leave

a residual near 0.31 (diagonal) and 0.30 (full), i.e. that factor is not well described by *any* linear map, which is the quantitative form of the statement in Section 3 that illuminance drags exposure, tone mapping and clipping along with it.

B Related Work in Full

Robustness benchmarks. ImageNet-C (Hendrycks & Dietterich, 2019), ObjectNet (Barbu et al., 2019) and ImageNet-v2 (Recht et al., 2019) vary capture or corruption and score category accuracy. Ibrahim et al. (2022) come closest to our setting, rendering millions of images with controlled pose, background, size and lighting and showing that CLIP and MAE remain brittle. Two things distinguish us: their variation is synthetic and their target is object category, so a failure cannot be attributed to a physical quantity, whereas our perturbation is a real capture protocol and our target is an instrumental measurement of the imaged surface. That is what lets us say a probe reports a different skin tone for the same person rather than that accuracy fell.

Color constancy. Discounting the illuminant is classical (Land & McCann, 1971; Fairchild, 2013), with statistical estimators (Buchsbaum, 1980; Finlayson & Trezzi, 2004; Gijsenij et al., 2011) and learned ones (Barron, 2015; Afifi & Brown, 2020); recent work continues to push on cross-camera spectral response and sRGB post-processing. That literature asks a model to output a corrected *image*. We ask the complementary question it does not address, whether a representation never asked to be color constant behaves as if it were, and what the residual looks like geometrically, and we use gray-world and shades-of-gray as baselines precisely because they are what a practitioner tries first.

Color in self-supervised learning. Zini et al. (2022) show that standard color jitter destroys color-sensitive features and propose a physically grounded jitter along the Planckian locus, evaluating on color-dependent classification. Their concern is what pretraining augmentation does to color features; ours is what the resulting frozen representation does to a physical measurement of a real surface under a real capture change. Their result predicts part of ours, and our finding that the correction which helps self-distilled encoders *hurts* language-supervised ones is a pretraining-dependent effect of the kind Geirhos et al. (2018a) and Hermann et al. (2020) established for texture.

Skin tone in vision systems. Buolamwini & Gebru (2018) made skin tone disparities concrete, and much auditing since has used the Fitzpatrick scale as the grouping variable (Groh et al., 2021). Thong et al. (2023) argue a single light-dark axis is insufficient and note that apparent skin color in an image is not the subject's skin color. Our corpus separates those two quantities directly because the subject was measured with an instrument. We also flag a trap this literature should care about: the field named `Fitzpatrick` here is a thresholding of the instrumental typology angle (Section 3), a substitution that conflates a reflectance measurement with a phototype.

Post-hoc edits to frozen representations. Removing a direction or subspace to suppress an attribute is standard (Bolukbasi et al., 2016; Ravfogel et al., 2020), as is aligning second-order statistics across domains (Sun et al., 2016). Our correction is of that family. The contribution is the audit and the comparison: the subspace is estimated from multi-condition captures of unlabeled calibration subjects, needing no task labels or attribute annotation, and we score it against input-side white balancing, condition-mean and random subspaces, and pooled-condition probe fitting on one utility–invariance trade-off measured against a physical target, because projecting away too much removes the skin-tone signal with the nuisance.

C Sensitivity Under Each Color Control

Figure 3 plots the paired displacement of Section 5.1 for every backbone under each color control on the same axes: the images as captured, with chroma removed, and after shades-of-gray white balancing. The numeric values, together with cross-condition retrieval and the measured-gain correction, are in Table 6.

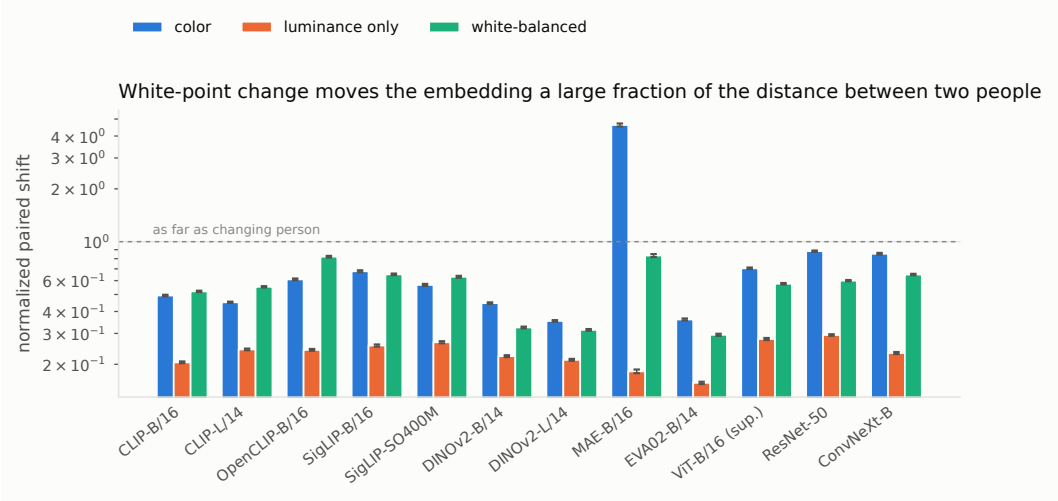

Figure 3: White-point sensitivity at fixed illuminance: displacement in units of the mean between-identity distance, with 95% subject bootstrap. Log scale; MAE-B/16 sits an order of magnitude above the rest (Section V).

D THE UTILITY-INVARIANCE TRADE-OFF IN FULL

Figure 4 sweeps the number of removed dimensions and plots all four scoring axes of Section 5.5 at once, each backbone normalized to its own uncorrected value so that models with very different absolute errors can be read on one scale. The operating point quoted in the text, $k = 4$, sits where the invariance axes have most of their gain and the utility axes have not yet moved.

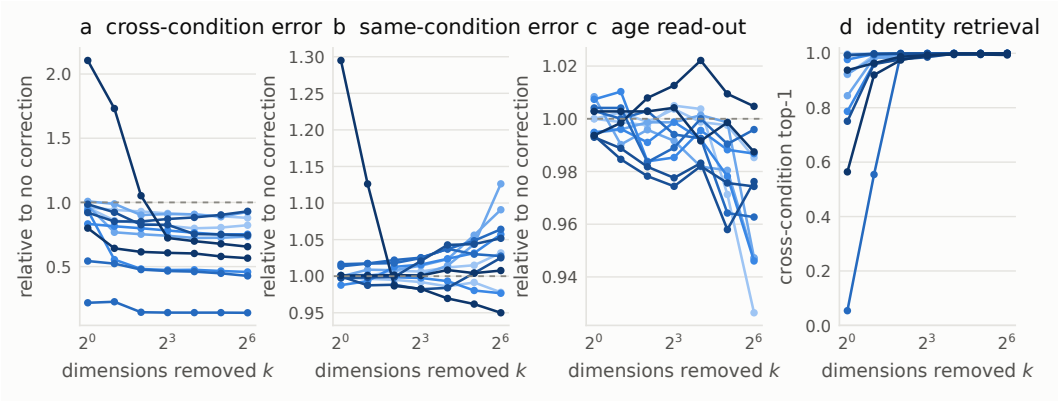

Figure 4: Post-hoc calibration as a trade-off, each backbone relative to its own uncorrected value. Removing more dimensions buys cross-condition invariance **(a)** and retrieval **(d)** but eventually spends same-condition skin tone **(b)** and the age read-out **(c)**.

The baseline rows. Shades-of-gray reaches 8.02°, worse than no correction, and the measured-gain correction 7.87° against 7.84° uncorrected, no benefit despite being fitted to the quantity it corrects. The feature-space oracle, which uses the test-time condition label to center each condition by its own mean, reaches 6.27°; the projection at $k = 4$ slightly beats it without that label.

E MATCHED MITIGATION BASELINES IN DETAIL

This appendix expands Table 1. All baselines run over the same 20 subject-disjoint splits, outer folds, feature standardization, and regularization grid as the paired projection; only the projection

or the probe's training rows change. The pooled probe's inner alpha-selection split is grouped by subject.

Condition means suffice for the subspace. Projecting out the rank-3 span of calibration condition-mean differences uses the same calibration images as the paired estimator, no task labels, and no query-condition label. It matches the paired projection at matched rank on every axis, with slightly better split reliability. Individual paired displacements are therefore not what makes the projection work; removing shared condition shifts is—consistent with the probe-aligned decomposition (Section R), where the most-degraded backbones are exactly the shared-shift-dominated ones.

The estimated directions matter, with one caveat. A rank-matched random orthonormal projection changes the median cross-condition error by 0.0% and helps no backbone on every split. The caveat is ResNet-50, which gains 23% at the median even from random rank-4 removal; its improvement under any projection should therefore not be read as capture-specific.

Probe fitting beats representation editing when its data exist. The pooled-condition probe is the strongest mitigation measured here, gaining on every backbone on every split. Its data requirement is images of the *labeled training subjects* under every capture condition; the ITA target is a per-subject measurement reused across captures, so no additional labeling is needed. The projection instead needs multi-condition captures only of separate unlabeled calibration subjects (as few as 32–64 suffice, Section J) and edits the representation for any downstream use, while the pooled probe fixes only the one probe it trains.

Fixed-probe deployment. Inserting either projection in front of a probe trained on uncorrected features is unreliable for both estimators (median reductions of 7% paired and 4% condition-mean; SigLIP-B/16 worsens under both), for the reason given in Section 5.5: a deployed probe may already use directions inside the removed subspace.

Scope. All of these comparisons validate transfer to held-out *subjects* under the same four capture conditions; the practical recommendations of the discussion are scoped to that setting.

F Verdict on the First-Order Hypothesis

Transfer holds: a rank-4 subspace fit on half the subjects captures 0.48–0.79 of the other half's displacement energy against 0.45 for the identity null (Table 7). *Concentration* holds as an empirical characterization: what arrives at the encoder is one translation holding 0.86 of the displacement energy at cosine 0.87, and what leaves keeps 0.38–0.71 in that translation, with a residual more spread out than the difference between two people; enough capture-related structure transfers across subjects to be removable. *Additivity* fails at the encoder output when read against the separable-multiplicative benchmark of Section N, while the measured input interactions (0.15 and 0.43 in the two arms, 0.34 on real captures) sit at or below that benchmark. The matched-multiplier ablation (Table 4) shows that the luminance multiplier, display encoding, and clipping suffice for the input interaction, with the measured level-dependent white point not increasing it in any tested setting; what each encoder does to the interaction it receives spans 0.7–$2.9\times$ in arm A and 0.7–$1.9\times$ in arm B. Two controls bound the geometric reading: the shared component is substantially background-carried (Section P), and removing the shared prediction shift alone explains most of the downstream harm for the most-degraded backbones and almost none of it for others (Section R).

The failure of input-side correction is a statement about residuals. An exact input inversion would cancel any downstream response ($C(T(x)) = x$ implies $f(C(T(x))) = f(x)$ for a deterministic encoder), so the failure of the measured-gain correction localizes to the input discrepancy a global diagonal gain leaves behind—tone mapping, clipping, spatial processing, the sensor color matrix, and non-skin regions receiving a gain fitted to skin (Figure 11)—and possibly to corrected images lying off the encoder's training distribution. Subspace projection needs no inversion: it removes the part of the response that is still shared across subjects, which is why estimating the same subspace from condition-mean differences works as well (Section 5.5).

G LIMITATIONS IN FULL

One capture pipeline. All 9028 images were collected with one camera model under one studio protocol. Section 5.3 separates what the encoder adds from what reaches it *within* that pipeline by synthesizing the perturbation, and the gain transfer of Section I evaluates the encoder's response to the measured gain on an image distribution the pipeline never touched; it models a white-point change rather than recapturing one. Whether a different camera yields the same magnitudes is the extension the protocol is designed for: it needs only repeated captures under controlled settings and an instrumental reference.

Input-side correction. The strongest input-side baseline is a per-condition diagonal gain fitted to measured skin. It does not invert the tone curve, clipping, spatial processing, or the sensor color matrix, so its failure bounds diagonal corrections rather than all input-side corrections. Decomposing the input discrepancy that remains after correction, and separating an imperfect inverse from corrected images lying off the encoder's training distribution, requires a raw-domain or color-chart-calibrated transform; this corpus provides neither.

Utility is measured on four axes. Section 5.5 preserves same-condition skin tone and an age readout and improves identity retrieval on this corpus, and preserves linear separability of a palette label on a second corpus (Section T). A projection that preserves these could still affect downstream tasks not tested here.

Region attribution. The shared crops retain background, and the region-masked control of Section P uses a feathered face ellipse as a proxy for skin. The downstream harm is measured on an instrumentally measured skin property regardless of which region mediates it; a fixed-background recapture or a pixel-accurate skin segmentation would separate the two regions more finely.

Faces and skin, not scenes. The instrumental reference exists because the surface of interest is skin. The results do not extend to materials, produce, textiles, or other color-critical domains, and the title is scoped accordingly.

Two levels per factor. A 2×2 design identifies main effects and their interaction but not the shape of the response between or beyond the levels. The dose-response experiment interpolates a *measured* gain rather than sampling new capture conditions, so it tests the encoder's response curve, not the camera's.

Group-conditional effects. The corpus has 179 subjects in the darkest ITA bin against 1609 in the middle bins, so group-conditional statistics for the extreme bins are correspondingly noisy. Within that precision, the ratio of capture-induced bias shift between the darkest and lightest bins is inconsistent in size and direction across backbones, and the correlation between displacement magnitude and typology angle is near zero. The error is group-*dependent* in sign rather than monotone in tone: the white-point change compresses and reshuffles the apparent tone distribution rather than penalizing one end of it.

H NULL MODELS AND RANK STATISTICS

Null models, because low rank is cheap. A collection of displacement vectors can look low-dimensional for reasons that have nothing to do with color: embedding spaces are anisotropic, so *any* set of differences between embeddings tends to concentrate. Every rank statistic below is therefore reported against three controls computed in the same space:

- **identity, same condition**, differences between two *different* people photographed under the *same* condition. No capture change is present, so this measures the anisotropy of the space itself. This is the baseline that matters.
- **condition + identity**, the pairing is permuted but the condition change retained, which sits between signal and null.
- **isotropic Gaussian**, a matrix of the same shape and total energy.

We add a fourth reference that is not a null but a *before* measurement: the same statistics computed on raw downsampled pixels. A caution on what a diagonal gain implies here: it is a three-*parameter* family, not a rank-3 linear operator on flattened image space. For a fixed gain matrix $\mathbf{G}$ the per-subject displacement is $\boldsymbol{\delta}_i = (\mathbf{G} - \mathbf{I})\, x_i$, and these vectors can span many dimensions as the images x_i vary—even for the identity encoder. Whether the input-side displacement population is dominated by a shared direction is therefore itself an empirical question, which is why we measure it (it is: mean-direction fraction 0.86, though its *centered* residual has participation ratio 32.0, Section K). Reporting both sides supports the comparison the paper relies on: how much of the displacement energy each side holds in one shared direction, and how the centered residual compares with the identity null in the same space. The operating rank is chosen empirically: under a pure gain model a scalar illuminance change lies *inside* the span of the three channel-gain directions, so $k{=}4$ is not implied by the physical model; it is plausible because the illuminance factor here is not a pure gain (Section A), and Figure 4 shows the trade-off across k. We summarize dimensionality by the participation ratio $\mathrm{PR} = (\sum_k \lambda_k)^2 / \sum_k \lambda_k^2$ of the displacement covariance spectrum, and by *held-out energy*: fit a rank-k subspace on half the subjects, and report the fraction of the *other* half's displacement energy it captures. A random rank-k subspace, which captures k/d in expectation, is reported alongside.

I Transferring the Measured Gain to a Second Corpus

This experiment evaluates the encoder's response to the measured white-point gain on an image distribution the AI Hub capture pipeline never touched. It is a transfer evaluation of a synthetic gain, not a recapture: the perturbation is modelled, and the corpus has no instrumental reference.

Protocol. We take the median linear-RGB gain measured between `5L32K` and `5L56K` on real skin, factor out its luminance component so that only the chromatic part remains, and apply $\hat{\mathbf{g}}^{\,\alpha}$ in linear RGB to a second corpus: 4920 face images labelled with a twelve-way seasonal personal-color palette (four seasons $\times$ three sub-tones). Personal-color analyzis is a task where the label is *defined* by the warmth, depth and clarity of a face's coloring, so a white-point change is not a nuisance the model may ignore; it is a direct attack on the decision boundary. A logistic probe is trained on frozen features of the unmodified training split, and evaluated on the held-out split at doses $\alpha \in \{-1, -\frac{1}{2}, 0, \frac{1}{2}, 1\}$, where $\alpha = 1$ is the measured shift and $\alpha = -1$ its inverse.

Scope of the transfer. The $\alpha = 0$ baseline is the unmodified image, so unlike every dosed variant it does not pass through the sRGB round trip and 8-bit requantization that applying a gain entails; measured on frozen features, that round trip alone accounts for about a tenth of the displacement a full dose produces, so the $\alpha = 0$ point is slightly advantaged. The gain was estimated on skin under one pipeline and is applied globally here, so backgrounds and clothing receive a transform a real light change would not give them, and extreme doses push saturated pixels out of gamut; Section L reports the chromaticity and clipping diagnostics.

Result. The primary statistic is *where the predictions move*, which does not depend on the palette labels being right, only on the predictions being consistent. The seasonal palettes split into warm (spring, autumn) and cool (summer, winter), which is the axis the white point was moved along, and the fraction of faces assigned a warm season tracks the applied dose monotonically for 8 of 12 backbones (Figure 5). Across the full range $\alpha \in [-1, 1]$ it moves from a median of 45% to 67%, a swing of 18 percentage points (up to 33 for SigLIP-B/16), from a minority to a majority of the same people, because of a camera setting. Individually, 51% of faces receive a different palette and 29% cross the warm-cool line. Shades-of-gray preprocessing reduces the crossing rate to 20%: it helps, and it does not solve it.

Twelve-way accuracy is 0.221 at $\alpha = 0$ (chance $1/12$; seasonal palette assignment is a partly subjective judgement) and falls to 0.199 under the shift. The monotone migration is the statistic the argument uses: the encoder has no way to tell "this person's coloring is warm" from "this photograph was taken with a warmer white point", and it resolves the ambiguity in favor of the photograph.

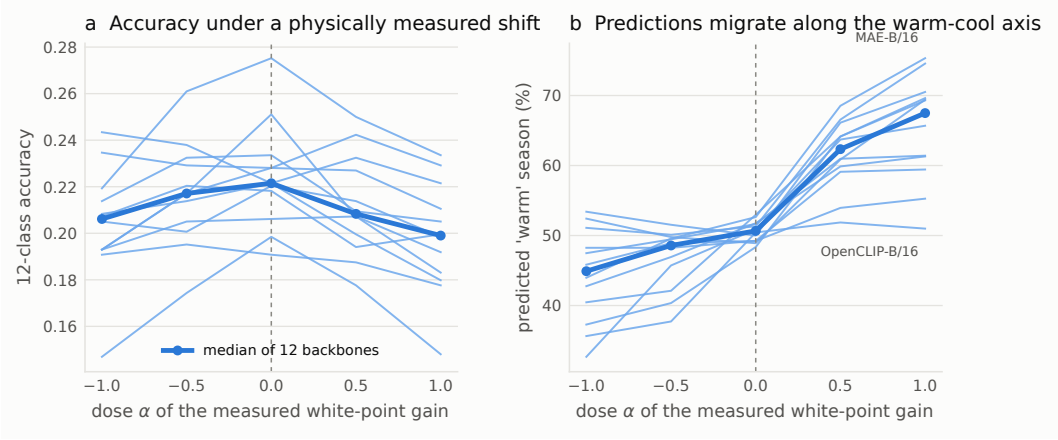

Figure 5: Transferring the measured white-point gain to a second corpus. $\alpha = 1$ is the measured gain, $\alpha = -1$ its inverse, $\alpha = 0$ the unmodified image. **(a)** Twelve-way accuracy. **(b)** Fraction predicted a warm season.

J Reproduction

Feature extraction. All twelve backbones are publicly released checkpoints loaded through one library (Wightman, 2019) with their own preprocessing configuration and no classification or projection head, evaluated frozen at 224^2 in half precision; DINOv2 is run at 224^2 so that resolution is not a confound. The CLIP and OpenCLIP features are the backbone's final-norm class token before the contrastive projection (768/1024 dimensions), and Section Q repeats the principal measurements on the projected outputs. Feature extraction for all backbones and color controls takes roughly two hours on a single RTX 4090; every downstream analysis is CPU-bound.

Probes and splits. Every probe is subject-disjoint: a subject's four captures never straddle a train/test boundary. ITA probes are ridge regressions on standardized features over 5-fold subject-grouped cross-validation, with $\alpha \in \{10^{-1}, 1, 10, 10^2, 10^3, 10^4\}$ chosen on an inner 80/20 split of the training rows. The subspace of Section 4.5 is estimated on calibration subjects disjoint from all evaluation subjects; the split-repetition analysis of Section 5.5 redraws that calibration/evaluation partition 20 times. The matched baselines of Section 5.5—condition-mean subspace, rank-matched random projection, and pooled-condition probe—reuse those identical splits, folds, scaler, and alpha grid, changing only the projection or the training rows; the pooled probe's inner alpha-selection split is grouped by subject, so no subject straddles it. The exact gain specification of both synthetic arms and of the matched-multiplier ablation is given in Section N.

Calibration and labeling budgets. Each split uses 1128 calibration subjects and 1129 held-out evaluation subjects, with all four captures of each (no task labels are used for calibration). The probes are trained on the evaluation half's own subject-disjoint folds; the pooled probe uses the same labeled training subjects with all four of their captures. The subspace estimator needs few calibration subjects: at $k{=}4$ the median cross-condition error is $6.17°$ with 32 calibration subjects and $6.08°$ from 64 onward, against $6.09°$ with the full 1128 (sample-efficiency sweep in the released measurements).

K Measurement Details

Rank statistics and the pixel-space measurement. Alongside the mean-direction fraction we report the uncentered participation ratio and the ratio of the residual after removing the population mean, since the two answer different questions and only the centered one is comparable against the identity null. The pixel-space measurement of Section 5.2 runs on 2257 subjects subsampled at random, a full-corpus decomposition at $224^2 \times 3$ being impractical; there the displacement has

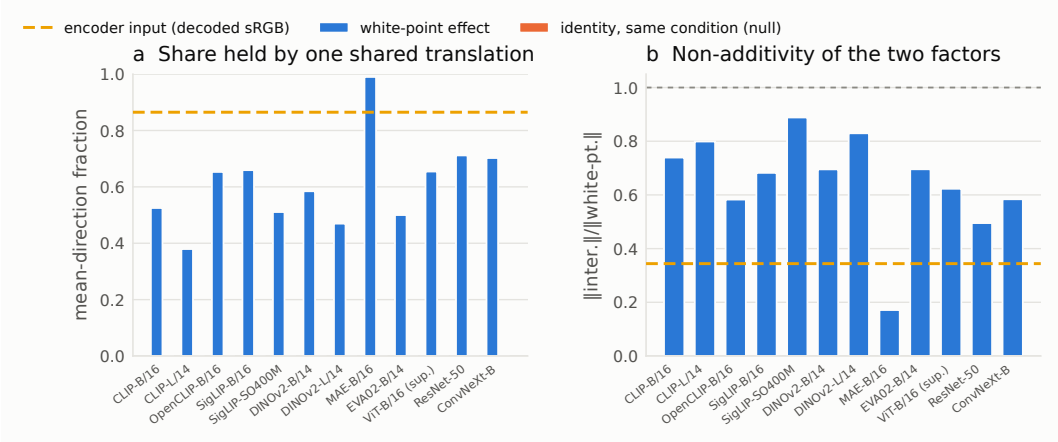

Figure 6: Geometry of the white-point response (Section 5.2). The dashed line shows the corresponding statistic at the encoder input. **(a)** Fraction of displacement energy explained by a shared direction. **(b)** Interaction between white point and illuminance relative to the white-point main effect. Subspace transfer results are reported in Table 7.

uncentered participation ratio 1.3 and a residual that is not itself low-dimensional ($\mathrm{PR}_{\text{centered}} = 32.0$) but small beside the shared offset.

K.1 Skin radiance measurement

Linear-RGB radiance is averaged over the pooled pixels of three fixed rectangles in the normalized crop, forehead ($y \in [0.26, 0.34]$, $x \in [0.42, 0.58]$) and two cheeks ($y \in [0.54, 0.64]$, $x \in [0.26, 0.36] \cup [0.64, 0.74]$). Fixed geometric patches are used rather than a chroma-based skin gate because such a gate is itself a function of the white point: it rejects essentially every pixel in the blue-cast condition and would bias the very gain being estimated. The patches need no gate because the crop is already normalized by the shared per-subject box.

K.2 Measured gains

Median per-subject linear-RGB gains relative to `5L32K`:

contrast	g_R	g_G	g_B	geometric mean
white point at 500 lx	1.951	1.066	0.387	0.93
illuminance at 3200 K	10.105	6.152	5.719	7.08

The white-point contrast at fixed illuminance is strongly chromatic with a geometric mean of 0.93, so almost no luminance change, while the illuminance contrast is dominated by a scale factor of 7.08 with a smaller chromatic component. We factor every gain as $\mathbf{g} = s\,\hat{\mathbf{g}}$ with $\prod_c \hat{g}_c = 1$ so that the two can be dosed independently. A caveat on terminology: imposing $\prod_c \hat{g}_c = 1$ does not in general preserve luminance $Y = \mathbf{w}^\top x$, so $\hat{\mathbf{g}}$ is a *geometric-mean-normalized* gain rather than a strictly luminance-preserving one; wherever the text calls a correction "chromatic", this normalization is what is meant.

K.3 Verifying the Fitzpatrick field

Sorting all 2257 subjects by the instrumental typology angle and reading off the label boundaries gives exactly $\{55, 41, 28, 10, -30\}$ with no exceptions, which is the standard ITA binning of Chardon et al. (1991) and Del Bino & Bernerd (2013). The field is therefore a deterministic function of a reflectance measurement, not a phototype assessment in the sense of Fitzpatrick (1988). We report ITA bins throughout.

Table 2: Diagnostics for the transplanted white-point gain (E4). The top block is the personal-color corpus after applying $\hat{\mathbf{g}}^{\alpha}$; the bottom block is where real AI Hub skin actually sits in linear rg-chromaticity. The transplanted shift moves in the same direction and by a comparable amount as the real capture change, which is the point of the check; the cost is clipping, which grows from 0.7% of pixels in the originals to 12.6% at $\alpha = 1$. We therefore report $\alpha = \pm\frac{1}{2}$ alongside the full dose.

variant	clipped	crushed	chroma r	chroma g
unmodified	0.7%	1.6%	0.539	0.269
$\alpha = -1.0$	5.0%	2.0%	0.287	0.251
$\alpha = -0.5$	2.4%	1.9%	0.412	0.272
$\alpha = +0.5$	6.6%	2.1%	0.637	0.254
$\alpha = +1.0$	12.6%	2.1%	0.686	0.249
real 5000 lx / 3200 K	–	–	0.316	0.270
real 5000 lx / 5600 K	–	–	0.517	0.288
real 500 lx / 3200 K	–	–	0.219	0.305
real 500 lx / 5600 K	–	–	0.455	0.347

L Realism of the Transplanted Gain

Table 2 quantifies how realistic the transplanted gain is. Applying $\hat{\mathbf{g}}$ to the personal-color corpus moves mean skin chromaticity from $r = 0.539$ to $r = 0.686$; the real capture change it was measured from moves AI Hub skin from $r = 0.219$ to $r = 0.455$. The two corpora start at different white points, the personal-color images are web photographs with their own unknown illuminants, but the displacement is in the same direction and of comparable magnitude, which is what the transfer argument needs. What it costs is headroom: clipped pixels rise from 0.7% to 12.6% at $\alpha = 1$. Real captures clip too, and the AI Hub $5000\,\mathrm{lx}/5600\,\mathrm{K}$ condition drives light skin close to saturation in the red channel, so this is not purely an artifact of the transplant; it is the reason we report the half dose alongside the full one.

M The Result on the Unit Sphere

Every geometry statistic in Sections 5.1 and 5.2 is computed on raw embeddings, while retrieval already uses cosine, so it is fair to ask whether what we measure is a change of feature norm that the normalization these encoders are usually used with would discard. Splitting each subject's displacement into the component along its own embedding and the component orthogonal to it puts 87–100% of the energy in the orthogonal part for the non-degenerate backbones, so the displacement is overwhelmingly a rotation rather than a rescaling. Recomputing on L2-normalized features accordingly leaves the headline unchanged: displacement 0.35–0.87 against 0.35–0.88 raw, coherence 0.31–0.67 against 0.31–0.67, with the largest single-backbone change in normalized displacement being 0.008. Retrieval is unaffected by construction, since it was already computed on cosine similarity. MAE-B/16 is the one backbone where the split goes the other way, which is consistent with the degenerate feature space described in Section V.

N Construction of the Synthetic Arms

Exact construction. The base image for every cell is the subject's `5L56K` capture; subjects are the first 800 in `uid` order. The four cells are built in linear RGB as $\{x,\ \mathbf{g}_{\text{lo}} \odot x,\ s\,x,\ s\,(\mathbf{g}_{\text{hi}} \odot x)\}$ and re-encoded to 8-bit sRGB, so neither arm is free of nonlinearity before the network: the gain goes through an sRGB round trip with quantization, and the large multiplier clips highlights as the real high-illuminance captures do. In **arm A**, $\mathbf{g}_{\text{lo}} = \mathbf{g}_{\text{hi}} = (2.098, 1.146, 0.416)$, the unit-

Table 3: Encoder-only ablation at full dose, 800 subjects, one capture each. Arm A applies the same chromatic gain at both luminance levels with a $2\times$ luminance scale; arm B uses the measured per-level chromatic gains and the measured $7.1\times$ scale. No pose or metering difference exists in either arm, so any change between an arm's input row and its model rows is the network's. MDF is the fraction of displacement energy in the population-mean direction. The bracketed factor is the interaction ratio relative to that arm's own input. Each arm's input interaction matches the value a perfectly separable multiplicative model predicts under its own luminance multiplier ($\approx$0.16 and $\approx$0.42 after display encoding, Section N); the matched-multiplier design of Table 4 separates the multiplier from the gains' level dependence.

space	PR_u	PR_c	MDF	coherence	$\|WL\|/\|W\|$
arm A: separable, luminance $\times 2.0$					
encoder input	1.14	9	0.93	0.93	0.15
CLIP-B/16	2.29	23	0.64	0.65	0.27 ($\times 1.8$)
CLIP-L/14	3.33	40	0.53	0.54	0.43 ($\times 2.9$)
OpenCLIP-B/16	1.53	25	0.80	0.80	0.28 ($\times 1.9$)
SigLIP-B/16	1.69	20	0.75	0.76	0.27 ($\times 1.9$)
SigLIP-SO400M	2.15	19	0.66	0.67	0.31 ($\times 2.1$)
DINOv2-B/14	1.75	23	0.75	0.76	0.14 ($\times 1.0$)
DINOv2-L/14	1.65	17	0.77	0.77	0.11 ($\times 0.7$)
MAE-B/16	1.00	1	0.99	1.00	0.03 ($\times 0.2$)
EVA02-B/14	2.21	23	0.65	0.66	0.29 ($\times 2.0$)
ViT-B/16 (sup.)	1.82	23	0.72	0.73	0.17 ($\times 1.2$)
ResNet-50	1.81	35	0.73	0.74	0.20 ($\times 1.4$)
ConvNeXt-B	1.91	24	0.71	0.71	0.19 ($\times 1.3$)
arm B: interaction-matched, luminance $\times 7.1$					
encoder input	1.16	14	0.93	0.93	0.43
CLIP-B/16	2.42	27	0.62	0.63	0.67 ($\times 1.6$)
CLIP-L/14	3.01	53	0.56	0.57	0.75 ($\times 1.7$)
OpenCLIP-B/16	1.70	28	0.76	0.76	0.76 ($\times 1.8$)
SigLIP-B/16	1.98	31	0.70	0.70	0.60 ($\times 1.4$)
SigLIP-SO400M	2.62	27	0.59	0.60	0.81 ($\times 1.9$)
DINOv2-B/14	1.87	24	0.72	0.73	0.35 ($\times 0.8$)
DINOv2-L/14	1.69	18	0.75	0.76	0.30 ($\times 0.7$)
MAE-B/16	1.00	1	0.99	1.00	0.15 ($\times 0.4$)
EVA02-B/14	2.51	26	0.61	0.62	0.63 ($\times 1.4$)
ViT-B/16 (sup.)	1.85	26	0.72	0.73	0.42 ($\times 1.0$)
ResNet-50	1.84	33	0.73	0.73	0.48 ($\times 1.1$)
ConvNeXt-B	2.01	21	0.69	0.69	0.43 ($\times 1.0$)

geometric-mean chromatic part of the gain measured at $500\,\mathrm{lx}$ (Section K), with $s = 2$. In **arm B**, $\mathbf{g}_{\mathrm{lo}} = (2.098,\ 1.146,\ 0.416)$ as in arm A, $\mathbf{g}_{\mathrm{hi}} = (1.785,\ 1.137,\ 0.492)$, the chromatic part measured at $5000\,\mathrm{lx}$, and $s = 7.1$, the measured luminance scale.

The interaction expected from separable factors. The factorial contrasts of Equation (2) are additive, so a finite *multiplicative* change produces a nonzero interaction even when the factors are

perfectly separable. For the four images x, $\mathbf{G}x$, sx, $s\mathbf{G}x$ with fixed diagonal $\mathbf{G}$ and scalar s,

$$W = \tfrac{s+1}{2}(\mathbf{G}-\mathbf{I})x, \qquad WL = \tfrac{s-1}{2}(\mathbf{G}-\mathbf{I})x, \qquad \frac{\|WL\|}{\|W\|} = \frac{s-1}{s+1},$$

for every subject, with no level-dependent gain anywhere. A per-channel display encoding $h(t) \approx t^{1/2.2}$ preserves this separability with s replaced by $s^{1/2.2}$, giving predicted input ratios of ≈ 0.16 at $s = 2$ and ≈ 0.42 at $s = 7.1$. The measured input values, 0.15 and 0.43, match these predictions almost exactly.

Reading the two arms. The input interaction in both arms is accounted for by the finite multiplicative luminance change viewed through additive contrasts and the display nonlinearity, and the real-capture input value of 0.34 sits below the same benchmark. Because the arms differ in the luminance multiplier (2 versus 7.1) and in the level dependence of the chromatic gains at once, the two factors are separated by the matched-multiplier design below. What the arms establish on their own is the encoder side: with the input interaction analytically fixed, each encoder's response to it is read off per arm, and it spans attenuation to roughly threefold (Table 3).

The matched-multiplier design. Table 4 crosses shared versus level-dependent chromatic gains with *each* multiplier, on the same 800 subjects, the same single base capture, and at full dose, with the input rows also recomputed without the sRGB clip and quantization. The shared/$2\times$ and level-dependent/$7.1\times$ cells are arms A and B. Three readings. (i) At a fixed multiplier, level dependence does not raise the input interaction: $0.146 \to 0.052$ at $2\times$ and $0.447 \to 0.434$ at $7.1\times$; the measured gains at the two illuminance levels differ in the same direction, which partially cancels the additive interaction. (ii) The multiplier alone reproduces the full arm difference under shared gains, $0.146 \to 0.447$, against the separable prediction $0.156 \to 0.418$. (iii) Clipping accounts for most of the remainder at the large multiplier: without the sRGB clip the input values are 0.387 (shared) and 0.283 (level-dependent). The encoder-side ranges over the eleven non-degenerate backbones are similar for shared and level-dependent gains at either multiplier (endpoints differ by a few hundredths per model), so the encoders' interaction predominantly tracks the multiplier. These are population-level comparisons of a normalized ratio: level dependence contributes a vector to the interaction contrast, here partially cancelling, and the ratio measures its net effect rather than its magnitude.

Operating point of the synthetic cells. The base image is the `5L56K` capture and the gain is the one measured in the $32\text{K} \to 56\text{K}$ direction, so the synthetic white-point step is applied *onward* from the 56K rendering rather than reconstructing the real (32K, 56K) pair. The step magnitude matches the real one while the absolute operating point is shifted: mean skin r-chromaticity ($R/(R{+}G{+}B)$ in linear RGB over the fixed patches) moves from 0.462 to 0.670 across the synthetic contrast, against $0.219 \to 0.455$ for the real $500\,\text{lx}$ captures (step 0.21 versus 0.24), and the real corpus never exceeds 0.517. The high-multiplier cells clip 35–53% of pixels in at least one channel, versus 0% at $1\times$; the real $5000\,\text{lx}$ captures also clip (Section L). The experiment is therefore a magnitude-matched stress test of the encoder's response to an exactly known gain, not a reconstruction of the four real cells, and its conclusions are about the encoder boundary, not the real pipeline's absolute operating point.

O Which Axis of Color the Damage Falls On

Probing L^*, a^* and b^* separately localizes the harm of Section 5.4: L^* is comparatively robust, while the warm-cool axis b^* degrades most. That is what a white-point change should do, and it is a check that the probe tracks the physical quantity it was fitted to rather than a generic domain shift, which would have no reason to prefer one axis of the color space over another.

P Skin versus Background

The shared crops retain background, and the wall changes color with the capture white point, so the displacement statistics of Sections 5.1 and 5.2 measure the response to the whole crop, not to skin specifically. To separate the two, we re-run the single-capture gain ablation with the measured

Table 4: Matched-multiplier ablation: the full 2×2 of {shared, level-dependent} chromatic gains $\times$ {2.0, measured 7.08} luminance multipliers, at full dose on the same 800 subjects and single base capture as Table 3 (the shared/2.0 and level-dependent/7.08 cells reproduce arms A and B exactly). "pred." is the separable-model prediction $(s' - 1)/(s' + 1)$, $s' = s^{1/2.2}$; "unclip." recomputes the input row without the sRGB clip and quantization. Changing the multiplier alone reproduces the full arm A/B difference, and at a fixed multiplier the measured level dependence does not increase the input ratio (at $\times 2.0$ it lowers it). The ratio is normalized, so it measures the net effect of level dependence, which adds a partially cancelling vector to the interaction contrast and changes the main-effect denominator.

		input $\|WL\|/\|W\|$			encoder range
luminance	chromatic gain	measured	unclipped	predicted	(11 non-degen.)
$\times 2.0$	shared (arm A)	0.146	0.144	0.156	0.11–0.43
$\times 2.0$	level-dependent	0.052	0.038	—	0.15–0.41
$\times 7.08$	shared	0.447	0.387	0.418	0.27–0.77
$\times 7.08$	level-dependent (arm B)	0.434	0.283	—	0.30–0.81

white-point chromatic gain applied to (a) the whole image, (b) a feathered elliptical face region re-detected by YuNet on each 224^2 crop (a proxy for skin; mean area 38% of the crop), and (c) the background only, on the same 800 subjects. Gains are applied in linear RGB as $1 + m(\mathbf{g} - 1)$ with the feathered mask $m \in [0, 1]$.

At the input, the shared component is mostly background: the background-only displacement carries 0.74 of the full displacement's energy and its mean direction has cosine 0.88 with the full one, against 0.23 and 0.53 for the skin-only part. In feature space the same holds with a twist: over the eleven non-degenerate backbones the background-only displacement has energy share 0.66–2.54 (median 1.10) with mean-direction cosine 0.32–0.86 (median 0.72) to the full displacement, while the skin-only part carries a median share of 0.68; shares exceeding one mean the two parts partially *cancel*, so the response is not additive over regions. Both parts are individually coherent across subjects (background 0.55–0.75, skin 0.45–0.72).

Two conclusions. First, the shared, coherent component of the white-point response is substantially background-driven, so the displacement and coherence statistics of Sections 5.1 and 5.2 characterize the encoder's response to the capture change over the whole crop rather than to skin specifically (the full-frame variant, Section W, is more extreme still). Second, the downstream finding holds independently of region: the skin-only perturbation alone moves embeddings by 0.33–0.76 of the between-identity distance with coherence up to 0.72, and the harm measured in Section 5.4 is a change in the prediction of an instrumentally measured skin property, whatever mixture of skin and background mediates it. The masked images contain a feathered seam no real capture has, so these numbers characterize region attribution rather than a realistic capture.

Q Implementation Validation: Projected CLIP and MAE Variants

We validate two feature-extraction choices; all statistics are for the fixed-illuminance white-point contrast under the identical protocol.

Native projected CLIP. The main results read CLIP and OpenCLIP before the contrastive projection. Re-extracting with the released projection (512/768/512 dimensions) leaves every principal finding qualitatively unchanged: NPS $0.49/0.45/0.61 \rightarrow 0.46/0.41/0.56$ (CLIP-B/16, CLIP-L/14, OpenCLIP-B/16), coherence $0.47/0.31/0.61 \rightarrow 0.43/0.28/0.60$, cross-condition retrieval $0.912/0.992/0.545 \rightarrow 0.841/0.988/0.561$, and cross-condition ITA error $5.92/5.97/6.95 \rightarrow 6.30/5.84/7.06$ against same-condition $4.57/4.78/4.28 \rightarrow 4.71/4.92/4.50$. The projection neither removes nor materially amplifies the sensitivity, so conclusions about "CLIP features" hold for both the internal and the released representation.

MAE degeneracy. Re-extracting MAE-B/16 under all four combinations of {mean-pooled, class-token} read-out and {fp16, fp32} precision reproduces the degeneracy in every cell: the identity-null centered participation ratio is 2.3 (mean-pool) and 1.9 (class token) at either precision, NPS is 4.6

Table 5: Probe-aligned decomposition of the prediction change under the fixed-illuminance white-point contrast (probe trained at 5L56K, evaluated at 5L32K; out-of-fold predictions, subject-disjoint folds). "shared" is the mean prediction shift $\overline{d}$; its squared share of the mean-squared change separates the shared shift from the subject-dependent residual. "recovered" is the fraction of the correctable error ($\mathrm{MAE}_{\mathrm{out}} - \mathrm{MAE}_{\mathrm{in}}$) removed by subtracting the scalar $\overline{d}$ alone.

model	shared (°)	shared frac. of MS	MAE out → de-shared	MAE in	recovered
CLIP-B/16	−3.75	0.57	5.88 → 5.02	4.57	0.65
CLIP-L/14	−1.83	0.26	5.45 → 5.28	4.78	0.25
OpenCLIP-B/16	−5.48	0.64	7.01 → 5.38	4.28	0.59
SigLIP-B/16	−5.00	0.47	7.60 → 6.49	4.98	0.42
SigLIP-SO400M	−6.01	0.72	7.44 → 5.50	4.77	0.73
DINOv2-B/14	−12.93	0.87	13.14 → 5.90	4.51	0.84
DINOv2-L/14	−5.70	0.69	7.18 → 5.33	4.51	0.69
MAE-B/16	−83.26	0.99	83.29 → 8.92	5.13	0.95
EVA02-B/14	−0.76	0.05	5.27 → 5.27	4.77	0.01
ViT-B/16 (sup.)	+1.57	0.06	6.56 → 6.42	4.57	0.07
ResNet-50	+1.64	0.00	11.91 → 11.72	6.99	0.04
ConvNeXt-B	−15.94	0.88	16.14 → 6.55	4.76	0.84

and 6.2, cross-condition retrieval is 0.000, and coherence is 1.00. The anomaly is therefore not an artifact of half precision or of the mean-pool read-out choice (the class token is more degenerate still), and a re-run of the paper's configuration reproduces it exactly. We treat the degeneracy as a property of this checkpoint's frozen feature space as distributed, exclude MAE-B/16 from headline ranges, and do not attribute it to the pretraining objective.

R Probe-Aligned Decomposition of the Prediction Change

The geometry of Section 5.2 does not by itself say whether the downstream harm comes from the shared displacement or from the subject-dependent residual. For the linear probe the connection is exact: the per-subject prediction change $d_i = p_i(\texttt{5L32K}) - p_i(\texttt{5L56K})$ equals $\mathbf{w}^\top \boldsymbol{\delta}_i$ up to the affine standardization, so its mean $\overline{d} = \mathbf{w}^\top \bar{\boldsymbol{\delta}}$ is the shared prediction shift and $d_i - \overline{d} = \mathbf{w}^\top(\boldsymbol{\delta}_i - \bar{\boldsymbol{\delta}})$ is the probe-aligned residual. Table 5 reports both, together with the error recovered by removing the scalar $\overline{d}$ alone.

The answer is heterogeneous, and it refines rather than confirms the geometric narrative. For the backbones with the largest cross-condition degradation (DINOv2, ConvNeXt-B, and degenerate MAE) the shared shift carries 0.87–0.99 of the mean-squared prediction change, and subtracting one scalar recovers 84–95% of the correctable error—consistent with condition-mean centering matching the paired projection in Section 5.5. At the other end, backbones already close to their same-condition floor (EVA02-B/14, ViT-B/16, ResNet-50) have shared fractions of 0.00–0.06: what degradation they show is almost entirely the subject-dependent residual, which no shared correction can remove. The distributed residual emphasized in Section 5.2 is therefore the binding constraint only for the latter group; for the former, most of the harm is one shared shift, which is exactly the part that is cheap to remove.

S A Direct Skin-Color-Measurement Baseline

How well do simple skin-region color measurements predict instrumental ITA under the same splits? We fit the identical ridge protocol on an 18-dimensional feature: the linear-RGB means of the three fixed skin patches of Section K and their logs—the same measurements used to fit the gains. Within condition this is competitive with the frozen features: MAE 5.73° ($R^2 = 0.77$) against 4.3–7.0° for the backbones. Across conditions it collapses: MAE 46–72° (mean 60.2°), within-subject inconsistency 46.1°, and a 95.9% bin-flip rate, since raw color measurements absorb the capture change in full. With the per-condition median gain removed—estimated on the calibration half, but applied using the *query's condition label*, an oracle the frozen pipelines are not given—the same features reach 5.73° within and 6.27° across conditions with a 12.8% flip rate, on par with the k=4 projection on frozen features. Two practical readings follow. First, frozen encoders are far more capture-stable

than uncorrected color measurements (7–13° versus 46–72° cross-condition): they perform substantial but incomplete constancy, which is precisely what makes their failure non-obvious. Second, where the capture condition of each image is known (a controlled or chart-calibrated pipeline), a calibrated direct color measurement is a strong, nearly free alternative to a frozen-feature probe for this task.

T THE SUBSPACE OFF ITS OWN PIPELINE

The four axes scored in Section 5.5 are all measured inside the AI Hub studio corpus: one camera, one pose, one background, one session. A projection that looks harmless there could still be removing something a general-purpose encoder needs, and that corpus cannot see it. We therefore estimate the rank-4 subspace on AI Hub faces exactly as Section 4.5 describes and apply it, unchanged and without re-estimation, to the second corpus of Section L, which is web photographs of many people taken with unknown and varied cameras. The task is that corpus's own seasonal palette label, four-way and twelve-way, which is not the quantity the subspace was constructed to remove; a linear probe is trained and tested on its released split.

Accuracy is essentially unchanged. The median shift over the 11 non-degenerate backbones is 0.003 on the four-way task and 0.001 on the twelve-way, the worst single backbone loses 0.008 and 0.008 respectively, and all 11 stay within one accuracy point on the four-way task over 912 held-out images. Three limits are worth stating. The second corpus is still faces, so this is not a general-image utility benchmark. Its labels are subjective palette assignments rather than a physical measurement. And the probe is refitted on the projected features, so what the comparison establishes is that the task remains linearly separable after the projection, not that the removed directions carried no signal: a task could survive because the information is redundant across dimensions the projection leaves alone. The fixed-probe setting of Section 5.5 is the one that asks the stronger question, and there the answer is model-dependent.

U DEPTH-RESOLVED GEOMETRY

U.1 WHERE IN DEPTH IT HAPPENS

Reading out every block of CLIP-B/16, CLIP-L/14, OpenCLIP-B/16, SigLIP-B/16, SigLIP-SO400M, DINOv2-B/14, DINOv2-L/14, EVA02-B/14, ViT-B/16 (sup.) separates two things. The *magnitude* is suppressed throughout, from 3.22 times the between-identity distance at the patch embedding to 0.36 at the output, a factor of 1.0 to 15.6 across backbones. Most do real work against the nuisance, though not at every individual block and not universally: 8 of 9 suppress it, while SigLIP-SO400M ends where it began. The *structure* holds only through the early blocks. Coherence starts near the pixel-space value at 0.99 and falls to 0.52, the participation ratio rises from 1.0 to 3.5, and the interaction ratio grows from 0.25 to 0.76, a factor of 2.4 to 3.8. All three move together over the second half of the network. This describes how the perturbation propagates, and it bounds the *leverage* of any residual: whatever input discrepancy a diagonal correction leaves behind is subject to the same depth-wise loss of shared structure as the original perturbation, so a small uncorrected remainder need not stay simple by read-out.

Figure 7 reads out the token sequence after each block and mean-pools it, including at depth 0, the patch embedding with position encodings added. All 9 non-degenerate transformer backbones are traced (the two convolutional backbones are not; MAE-B/16 was traced but is reported separately as degenerate, Section V), and only the statistics plotted were computed at each depth; the mean-direction fraction was not. Mean pooling is used at every depth so that the comparison across depths is of one quantity; this differs from the main results, which use each backbone's own released read-out (class token for CLIP, OpenCLIP and DINOv2, attention pooling for SigLIP, Section 4), so the final point of Figure 7 is not numerically identical to the corresponding entry of Table 7. The qualitative pattern, magnitude falling overall while coherence falls and dimension and non-additivity rise, is unaffected by the choice.

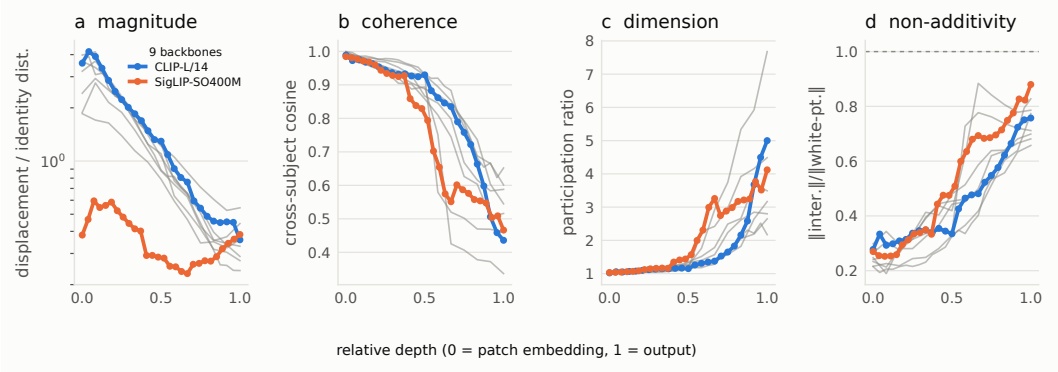

Figure 7: White-point displacement after each block (mean-pooled tokens) for the 9 non-degenerate transformer backbones, on a relative depth axis. Each backbone is one grey line; two are named, the median suppressor and SigLIP-SO400M, whose magnitude ends where it began. Magnitude falls for the rest **(a)** while structure degrades in all of them **(b–d)**.

V MAE-B/16: A Degenerate Feature Space

MAE-B/16 without fine-tuning is an outlier on every axis: its displacement is 4.62 times the between-identity distance, larger than changing the person, and cross-condition retrieval collapses to 0.000. Its identity null is degenerate too: differences between two *different* people have participation ratio 1.7, so essentially all variation in this space lies along one direction (Table 7), and its grayscale coherence is 0.484 rather than the ≤ 0.044 every other backbone shows. This is consistent with the known weakness of MAE features under linear probing before fine-tuning. Section Q localizes it: the degeneracy reproduces identically under fp32 inference and under a class-token read-out (which is more degenerate still), so it is not an artifact of precision or pooling choice. We treat it as a property of this checkpoint's frozen feature space rather than a consequence of the pretraining objective, and it is in any case not a fact about color. We report the model in every table and exclude it from headline ranges, which are quoted over the other 11 backbones.

W Full-Frame Results

Section 3 restricts the main results to shared per-subject face crops. The full-frame variant is included in the released measurements (released with the per-model measurements), where the uniform background wall occupies most of the image and changes color with the white point, so every sensitivity statistic increases. We report those numbers for completeness but build no claim on them: an encoder that responds to a large uniform colored region is doing something unremarkable.

X Qualitative Corpus Examples

This appendix collects image-level examples that make the corpus and the capture manipulation of Section 3 concrete: who is in the corpus (Figures 8 and 9), what the shared crop holds fixed (Figure 10), and what the measured-gain correction of Section 4.4 does and does not undo (Figure 11). Faces are shown as the encoders see them, i.e. after the shared 224^2 crop, except where the full capture is needed. Every subject shown is selected by a deterministic rule stated in the caption, not curated by hand.

Y Per-Model Tables



Figure 8: One subject per regional group (columns; each group's median-ITA subject) under the four capture conditions (rows). Within a column only the capture settings change, and the blue cast of the 3200 K captures and the photometric change across illuminance levels are uniform across groups; within a row the corpus spans its six regional groups.

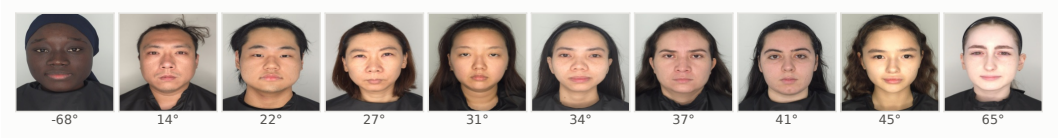

Figure 9: Coverage of the skin-tone axis. Ten subjects taken at the deciles of the per-subject instrumental ITA, shown under the 5000 lx/5600 K capture and labeled with their ITA. The distance between the darkest subject and the next decile reflects the bin imbalance discussed in Section 3: dark-bin subjects are rare, so the lowest decile sits far from the rest of the distribution.

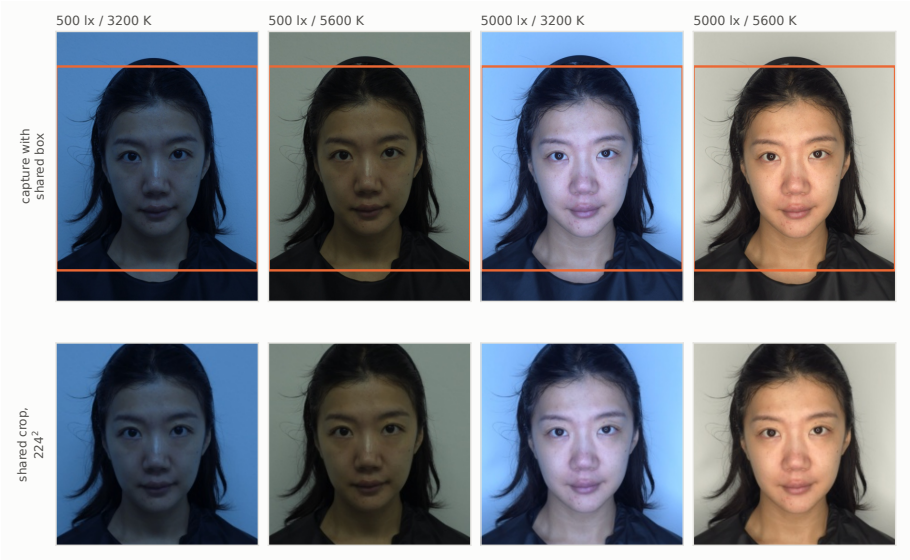

Figure 10: The shared per-subject crop of Section 3. Top: the four captures of one subject with the single square box (orange) computed once per subject from YuNet detections in all four captures; bottom: the resulting 224^2 crops handed to every encoder. Because the box does not move with the condition, the crop cannot leak the variable under study. The subject shown is the one whose box leaves the largest visible margin; most boxes cover still more of the frame.

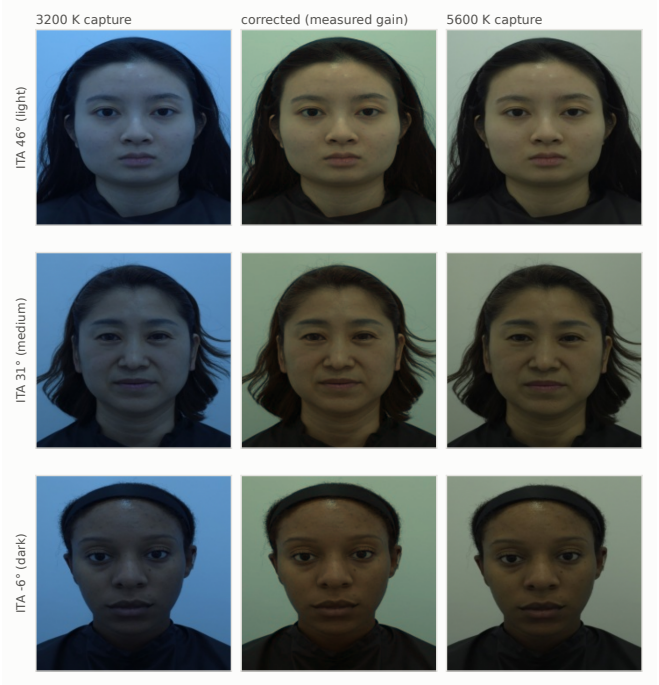

Figure 11: What the measured-gain correction does to the pixels. Left to right: the $500\,\mathrm{lx}/3200\,\mathrm{K}$ capture, the same image after the inverse per-channel skin gain of Section 4.4 (fitted on real skin, Section A), and the $500\,\mathrm{lx}/5600\,\mathrm{K}$ capture of the same subject; rows are the median subjects of the three ITA bins. Skin chromaticity lands close to the $5600\,\mathrm{K}$ rendering while the background does not, and tone mapping and clipping are not inverted: the correction is a global diagonal gain, not an inverse of the camera pipeline, which is why residual sensitivity remains in Section 5.5.

Table 6: White-point sensitivity at fixed illuminance (5L32K→5L56K). NPS is the paired displacement in units of the mean between-identity distance and coherence are defined in Equation (1); coherence is zero in expectation for any perturbation independent across subjects. Retrieval is cross-condition top-1 against chance $1/2257$. The three control columns repeat NPS with chroma removed, after shades-of-gray white balancing, and after the measured-gain correction – which makes things *worse* for most language-supervised encoders. Brackets are 95% subject bootstrap.

model	pretraining	NPS	coherence	top-1	gray	SoG	oracle	cond. acc.
CLIP-B/16	language-supervised	0.493 [0.49,0.50]	0.468	0.912	0.205	0.521	0.527	1.000
CLIP-L/14	language-supervised	0.451 [0.45,0.46]	0.314	0.992	0.243	0.553	0.546	1.000
OpenCLIP-B/16	language-supervised	0.609 [0.60,0.62]	0.613	0.545	0.242	0.821	0.870	1.000
SigLIP-B/16	language-supervised	0.678 [0.67,0.69]	0.663	0.444	0.256	0.651	0.679	1.000
SigLIP-SO400M	language-supervised	0.567 [0.56,0.57]	0.545	0.603	0.267	0.631	0.754	1.000
DINOv2-B/14	self-distilled	0.446 [0.44,0.45]	0.552	0.972	0.222	0.324	0.294	1.000
DINOv2-L/14	self-distilled	0.353 [0.35,0.36]	0.450	0.991	0.212	0.314	0.307	1.000
MAE-B/16	masked	4.621 [4.53,4.72]	0.998	0.000	0.182	0.833	0.420	1.000
EVA02-B/14	masked	0.360 [0.36,0.36]	0.476	0.929	0.157	0.294	0.343	1.000
ViT-B/16 (sup.)	label-supervised	0.705 [0.70,0.71]	0.597	0.710	0.278	0.576	0.583	1.000
ResNet-50	label-supervised	0.883 [0.88,0.89]	0.643	0.483	0.294	0.599	0.568	1.000
ConvNeXt-B	label-supervised	0.853 [0.84,0.86]	0.672	0.373	0.232	0.649	0.621	1.000

Table 7: Geometry of the white-point response. MDF is the fraction of displacement energy held by the population-mean direction, i.e. how much of the effect is one shared translation. PR_u and PR_c are the uncentered and centered participation ratios; the null column is PR_c for differences between two *different* people under the *same* condition, which measures the anisotropy of the space itself. Transfer is the fraction of held-out subjects' displacement energy captured by a rank-4 subspace fit on the other half of subjects. The last column is $\|WL\|/\|W\|$: the white-point × illuminance interaction relative to the white-point main effect. Brackets are 95% subject bootstrap.

model	d	MDF	PR_u	PR_c	PR_c null	transfer	transfer null	$\|WL\|/\|W\|$
CLIP-B/16	768	0.53 [0.52,0.53]	3.4	36	18	0.641	0.424	0.74 [0.74,0.74]
CLIP-L/14	1024	0.38 [0.38,0.38]	6.4	69	27	0.483	0.321	0.80 [0.80,0.80]
OpenCLIP-B/16	768	0.65 [0.65,0.66]	2.2	26	17	0.765	0.435	0.58 [0.58,0.59]
SigLIP-B/16	768	0.66 [0.66,0.66]	2.2	29	15	0.759	0.450	0.68 [0.68,0.69]
SigLIP-SO400M	1152	0.51 [0.51,0.52]	3.4	21	14	0.673	0.458	0.89 [0.88,0.89]
DINOv2-B/14	768	0.58 [0.58,0.59]	2.8	41	15	0.675	0.456	0.70 [0.69,0.70]
DINOv2-L/14	1024	0.47 [0.46,0.47]	4.3	50	15	0.568	0.450	0.83 [0.83,0.83]
MAE-B/16	768	0.99 [0.99,0.99]	1.0	1	2	1.000	0.916	0.17 [0.17,0.17]
EVA02-B/14	768	0.50 [0.49,0.51]	3.5	27	12	0.642	0.502	0.70 [0.69,0.70]
ViT-B/16 (sup.)	768	0.65 [0.65,0.66]	2.2	31	13	0.745	0.476	0.62 [0.62,0.63]
ResNet-50	2048	0.71 [0.71,0.72]	1.9	35	15	0.783	0.410	0.50 [0.49,0.50]
ConvNeXt-B	1024	0.70 [0.70,0.71]	1.9	24	15	0.792	0.428	0.58 [0.58,0.59]

Table 8: Cost to a skin-tone read-out, in degrees of individual typology angle. The probe is ridge on frozen features, trained at `5L56K` with subject-disjoint folds. 'in' and 'out' are MAE evaluated at the training condition and averaged over the other three. Bias is the largest absolute population-level offset induced by a condition change. Self-inconsistency is the change in the *same subject's* predicted angle between `5L32K` and `5L56K`, which needs no label; the last column is how often that change moves the subject into a different ITA bin.

model	MAE in	MAE out	rel.	max bias	self-inc.	p95	ITA-bin flip
CLIP-B/16	4.57	5.92	30%	3.69	4.12	9.18	27.6%
CLIP-L/14	4.78	5.97	25%	3.73	2.86	6.89	19.9%
OpenCLIP-B/16	4.28	6.95	63%	5.47	5.77	12.19	38.1%
SigLIP-B/16	4.98	6.96	40%	4.95	5.91	14.21	39.8%
SigLIP-SO400M	4.77	6.59	38%	6.00	6.10	13.18	40.9%
DINOv2-B/14	4.51	12.04	167%	12.89	12.94	21.00	74.6%
DINOv2-L/14	4.51	12.70	182%	15.24	5.89	12.18	38.0%
MAE-B/16	5.13	50.08	877%	83.29	83.26	98.70	98.6%
EVA02-B/14	4.77	6.56	38%	5.68	2.53	6.46	17.9%
ViT-B/16 (sup.)	4.57	8.28	81%	11.41	4.81	12.76	32.3%
ResNet-50	6.99	14.48	107%	7.52	10.08	23.40	49.9%
ConvNeXt-B	4.76	10.74	126%	15.95	15.95	26.17	81.6%

Table 9: Post-hoc calibration against its baselines, median over the eleven non-degenerate backbones (the same set the ranges in the text quote; MAE-B/16 is in Table 10), on held-out subjects only. Cross-condition ITA error and the ITA-bin flip rate should fall; same-condition error, identity retrieval and the age read-out should not move. Input-side methods act on pixels; ISP removes k directions from the frozen embedding, estimated from paired calibration captures with no task labels. The oracle requires the test-time condition label of every query.

method	ITA out	ITA in	self-inc.	bin flip	retr. top-1	age acc.
no correction	7.84	5.17	5.58	38%	0.760	0.647
gray-world (input)	7.38	5.13	3.71	25%	0.914	0.645
shades-of-gray (input)	8.02	5.16	5.59	37%	0.898	0.647
grayscale (input)	8.13	5.86	2.56	17%	1.000	0.632
measured-gain correction (input)	7.87	5.18	5.06	34%	0.940	0.648
ISP k=1	7.10	5.17	5.08	33%	0.924	0.649
ISP k=4	6.09	5.16	2.97	20%	0.996	0.637
ISP k=16	6.00	5.24	2.98	20%	0.999	0.647
ISP k=64	5.79	5.24	2.98	19%	0.999	0.627
SoG + ISP k=4	6.19	5.15	3.28	21%	0.995	0.643
oracle: condition centering	6.27	5.17	3.22	20%	0.994	0.647

Table 10: Post-hoc calibration per backbone at $k = 4$, held-out subjects, on *one representative calibration/evaluation split*; the per-backbone medians and ranges quoted in the text aggregate over 20 such splits, so single entries here can differ from them (ResNet-50, split-unstable in that analysis, worsens on this split while gaining at its split median). 'refit' retrains the ITA probe on corrected features; 'fixed' inserts the projection in front of a probe trained on uncorrected ones. ∘ marks a change within $\pm 2\%$, × a backbone that gets worse. Same-condition error and the age read-out are the axes that must not move.

	ITA MAE out			ITA MAE in		bin flip		age acc.	
model	none	refit	fixed	none	ISP	none	ISP	none	ISP
CLIP-B/16	6.84	5.74	6.57	4.96	5.00	38%	19%	0.710	0.709
CLIP-L/14	6.11	5.68	5.83	5.32	5.29	23%	15%	0.723	0.722
OpenCLIP-B/16	7.84	5.90	7.42	4.74	4.78	43%	20%	0.655	0.655
SigLIP-B/16	7.04	6.35	9.22	5.28	5.29	31%	23%	0.634	0.632
SigLIP-SO400M	7.23	5.77	6.81	5.17	5.16	36%	18%	0.689	0.683
DINOv2-B/14	12.63	6.09	9.24	4.95	5.02	69%	22%	0.602	0.593
DINOv2-L/14	13.07	6.23	8.16	4.97	5.06	42%	19%	0.647	0.636
MAE-B/16	49.59	7.17	44.88	5.80	5.93	98%	28%	0.593	0.595
EVA02-B/14	6.40	5.45	5.92	5.49	5.43	16%	16%	0.692	0.677
ViT-B/16 (sup.)	8.01	6.62	8.72	5.02	5.02	30%	30%	0.632	0.621
ResNet-50	13.96	14.69 ×	13.40	7.46	7.37	48%	43%	0.560	0.564
ConvNeXt-B	10.30	6.33	7.72	5.32	5.33	80%	27%	0.635	0.637