DOES IMAGE QUALITY AWARENESS EMERGE FOR FREE FROM SELF-SUPERVISED PRETRAINING?

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision foundation models are trained to recognize and describe image content, never to judge image quality. We ask whether they nevertheless encode human-perceived quality, and where in the network. We freeze four base-size Vision Transformers (ViT-B) trained with different objectives — masked-image reconstruction (masked autoencoder, MAE), self-distillation (DINOv2), image–text contrastive learning (Contrastive Language–Image Pre-training, CLIP), and supervised classification — and fit only linear probes to human quality ratings on two image quality assessment (IQA) benchmarks, one with authentic (KonIQ-10k) and one with synthetic (KADID-10k) distortions. All four encode substantial quality information. On authentic distortions their output features outperform a probe on hand-crafted low-level statistics, reaching a Spearman rank correlation (SROCC) of up to 0.87; within a single scene they rank synthetic distortions largely in line with human ratings; and they transfer between the two kinds of distortion, where the hand-crafted baseline barely does. The information is richest in intermediate blocks, where probes reach 0.92, and how much of it survives to the output depends on the objective: pixel reconstruction preserves nearly all of it, while discriminative objectives discard part of it. Hand-crafted statistics still identify the distortion type better, and we find no reliable link between quality encoding and robustness to corruptions.

1 INTRODUCTION

General-purpose vision foundation models are trained to recognize, localize, and describe image content. Masked autoencoders (MAE) reconstruct missing pixels (He et al., 2022), self-distillation methods such as DINOv2 (self-distillation with no labels, version 2) align augmented views of the same image (Oquab et al., 2023), and language-supervised models such as Contrastive Language–Image Pre-training (CLIP) align images with their captions (Radford et al., 2021). None of these objectives ever presents a model with a human judgment of *how good an image looks*. In parallel, a separate and mature research area — image quality assessment (IQA), and in particular no-reference IQA (NR-IQA), which needs no pristine reference image — builds architectures and loss functions whose sole purpose is to predict that judgment (Mittal et al., 2012; Zhang et al., 2020; Su et al., 2020; Ke et al., 2021; Yang et al., 2022).

This raises a simple question with a non-obvious answer. When a general-purpose model processes a blurry, noisy, or compressed image, does its representation already encode that the image is degraded, as an incidental consequence of learning to represent content — or is quality a genuinely separate axis these objectives never induce? The answer matters beyond curiosity: if quality-awareness is latent in off-the-shelf features, any downstream system built on top (retrieval, captioning, embodied perception) may inherit implicit degradation sensitivity for far less supervision than training a dedicated IQA head from scratch.

We test this directly and conservatively. We freeze four base-size Vision Transformer (ViT-B) backbones representing four pretraining objectives, extract pooled features, and fit *only linear probes* on top — no fine-tuning. A linear probe is a deliberately weak decoder: if it recovers human quality judgments, that information must already be linearly separable, the strongest literal sense of "already

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.

there." We evaluate on two IQA benchmarks differing in exactly the axis our question is sensitive to — authentic in-the-wild distortions (KonIQ-10k, Hosu et al. 2020) and synthetic, parametrically controlled distortions (KADID-10k, Lin et al. 2019) — with content-disjoint splits throughout, and calibrate every result against the same probe fitted on hand-crafted low-level statistics.

Contributions.

- A controlled probing study: four frozen ViT-B backbones with different pretraining objectives, one linear read-out, content-disjoint splits, and a hand-crafted low-level baseline (Section 3).
- Frozen features encode perceived quality beyond low-level statistics: on authentic distortions all four beat the baseline in each of five random splits; with the scene held fixed they rank synthetic distortions far better than it; and they transfer across the authentic/synthetic shift where it barely does (Sections 4.1–4.3).
- Quality information peaks in intermediate blocks, and the objective decides how much of it reaches the output: MAE keeps 98% of its peak, the other objectives 88–94% (Section 4.4).
- We report what does not hold: backbone rankings differ between authentic and synthetic distortions, hand-crafted statistics identify distortion type better, and quality encoding is not reliably linked to corruption robustness (Sections 4.1, 4.2, 4.5).

We use dedicated NR-IQA models only as calibration, not as a target to beat, and our zoo is four ViT-B backbones, not a scaling study (Section 6).

2 Related Work

No-reference IQA. Classical no-reference IQA methods use hand-crafted natural-scene statistics (Mittal et al., 2012; 2013). Deep NR-IQA methods train convolutional neural network (CNN) or transformer models end-to-end, or on ImageNet-pretrained backbones with task-specific heads, directly against human mean opinion scores (MOS) (Zhang et al., 2020; Su et al., 2020; Ke et al., 2021; Yang et al., 2022), reaching SROCC in the low-to-high 0.90s on KonIQ-10k-style benchmarks. Bardhan et al. (2024) fine-tune ImageNet-pretrained CNNs end-to-end for no-reference stereoscopic IQA and find them competitive with dedicated methods, suggesting that features learned for other visual tasks carry quality-relevant information; we test this idea under a strictly frozen, linear read-out, which separates what is already present from what fine-tuning can add.

Pretrained representations for IQA. Several works learn quality-specific representations without human labels and read them out with a linear regressor on frozen features: CONTRIQUE (Madhusudana et al., 2022) and Re-IQA (Saha et al., 2023) use contrastive pretext tasks on distorted images, and Re-IQA finds high-level content features complementary to its low-level quality features. Zhao et al. (2023) design a quality-aware pretraining objective whose backbone is then fine-tuned. CLIP-IQA (Wang et al., 2023) shows that CLIP judges quality zero-shot through antonym prompts, and Yadav et al. (2026) compare foundation backbones (CLIP, SigLIP 2, DINOv2/v3 and others) for NR-IQA with a shared lightweight head. These works train for quality or add trainable heads; we ask what general-purpose objectives encode with nothing but a linear probe, compare objectives at matched scale, control for image content, and analyze depth.

Probing frozen representations. Linear probes on frozen features are a standard way to measure what a representation encodes (Alain & Bengio, 2016), and in-domain linear probing predicts how self-supervised models rank downstream (Marks et al., 2025). Which layer is probed matters: linear separability of a model's own training task tends to grow with depth (Alain & Bengio, 2016), whereas generatively pretrained models give their best features for other tasks in the middle of the network (Chen et al., 2020). We apply this methodology to a non-semantic target, perceived quality, with a layer-wise sweep.

Corruption robustness. ImageNet-C (Hendrycks & Dietterich, 2019) defines the corruption vocabulary (noise, blur, digital artifacts at graded severities) that we use in Section 4.5.

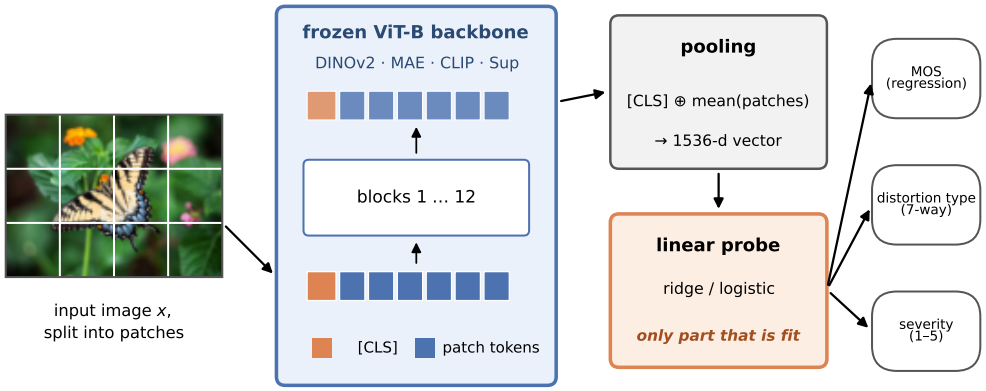

Figure 1: Probing pipeline. A frozen backbone passes the class ([CLS]) and patch tokens of an image through 12 blocks; we pool the output [CLS] token with the mean of the patch tokens into a 1536-dimensional vector and fit only a linear probe on top, separately for each target: MOS, distortion type, or severity.

Table 1: Backbones probed. "Sup" is the *supervised* control: a ViT-B/16 classifier trained with class labels on ImageNet-21k, then fine-tuned on ImageNet-1k (AugReg recipe, Steiner et al. 2022). All are used strictly frozen; checkpoints in Appendix A.4.

Backbone	Pretraining objective	Pretraining data	Patch/input
DINOv2-B/14	Self-distillation (image + patch)	LVD-142M (142M images, unlabeled)	14 / 518
MAE-B/16	Masked pixel reconstruction	ImageNet-1k (unlabeled)	16 / 224
CLIP-B/16	Image–text contrastive	400M image–text pairs	16 / 224
Sup-B/16	Supervised classification	ImageNet-21k → 1k (labeled)	16 / 224

3 Method

3.1 Backbones

We probe four base-size Vision Transformer (ViT-B; Dosovitskiy et al. 2021) backbones (Table 1), chosen to span distinct pretraining objectives at a matched, moderate parameter scale. All four share one architecture — 12 transformer blocks, width C=768, and a single class ([CLS]) token — and differ in pretraining objective, pretraining data, and input resolution. Every backbone is strictly frozen: no parameter is updated after loading the public checkpoint, and no image is used for anything other than a forward pass.

Feature extraction. For every backbone we apply the checkpoint's own preprocessing, take the tokens output by the final block after the model's final LayerNorm, and concatenate the [CLS] token with the mean of the patch tokens, giving a $2C$=1536-dimensional vector. Backbone-specific details: *DINOv2* uses 14×14 patches on 518×518 inputs (37×37=1369 patch tokens). *MAE* uses only the encoder (the pixel-reconstruction decoder is discarded) with no patches masked at inference (224×224 inputs, 196 patch tokens). *CLIP* uses only the image tower, taking features before the image–text projection head. *Sup* uses the classifier with its ImageNet-1k head removed. For the layer-wise analysis (Section 4.4) we apply the same pooling after each of the 12 blocks, with the same final LayerNorm, so the last block equals the features used everywhere else.

3.2 Datasets

We use two widely-used, public NR-IQA benchmarks that differ precisely along the axis our question is sensitive to. **KonIQ-10k** (Hosu et al., 2020): 10,073 authentic, in-the-wild photographs, each with a crowdsourced MOS; distortions are whatever naturally occurred during capture, unlabeled

and uncontrolled. **KADID-10k** (Konstanz Artificially Distorted Image quality Database; Lin et al. 2019): 81 pristine reference images, each synthetically degraded with 25 distortion types (grouped into 7 coarse categories: blur, color, compression, noise, brightness, spatial, sharpness/contrast) at 5 severity levels, for 10,125 total images with per-image MOS and known type/severity labels. Example images from both datasets are shown in Appendix A.1.

For every probe we use a content-disjoint $80/20$ split: on KonIQ-10k every image is a distinct scene, so we split over images; on KADID-10k we split over the 81 *reference* scenes, so no distorted version of a test scene is seen in training (otherwise its 125 versions would leak content across the split). Main results use one fixed split; Appendix A.2 repeats the MOS probes over five random splits.

3.3 Probing protocol

Regression probe. We standardize features (fit on train only) and fit ridge regression with regularization strength chosen by internal cross-validation over 25 log-spaced values in $[10^{-3}, 10^{5}]$. We report SROCC, the Kendall rank-order correlation coefficient (KROCC), and, after mapping predictions onto the MOS scale with a monotonic five-parameter logistic (the standard convention of the Video Quality Experts Group, VQEG), the Pearson linear correlation coefficient (PLCC) and root mean square error (RMSE), with 95% bootstrap confidence intervals (CIs) on SROCC (1000 resamples of the test set).

Classification probe. For the seven-way distortion-category task we fit multinomial logistic regression, reporting accuracy and macro-F1 against both a uniform-chance baseline ($1/7$=14.3%) and our test split's majority-class baseline (20.0%; the seven categories are not equally sized).

Hand-crafted baseline. To calibrate what a linear probe extracts without any learned representation, we fit the same probes on 69 low-level statistics per image: moments of the mean-subtracted contrast-normalized (MSCN) coefficients and of their products with four neighbours at two scales, as in BRISQUE (Mittal et al., 2012) (our reimplementation, using moments rather than distribution fits), plus sharpness, gradient energy, brightness, contrast, colourfulness, saturation, entropy, noise and JPEG-blockiness estimates (Appendix A.4).

Within-scene evaluation. Pooled SROCC on KADID-10k mixes two abilities: ranking the distorted versions of one scene, and placing different scenes on a common scale. We therefore also compute SROCC within each of the 16 test scenes (over its 125 distorted versions) and report the median, which measures distortion sensitivity with content held fixed.

3.4 Layer-wise probing

We repeat the KonIQ-10k regression probe independently on the pooled representation (Section 3.1) after each of the 12 transformer blocks of every backbone (numbered 1–12; block 12 is the final block), to ask where in the network quality is most linearly decodable. The best block is selected on a validation split (20% of the training images), and we report its test SROCC.

3.5 Cross-dataset transfer

We fit the regression probe on one dataset's full feature set and evaluate, without any adaptation, on the other's, in both directions. This tests whether the "quality direction" a probe finds is specific to the distortion distribution it was fit on, or reflects something closer to a general-purpose quality axis.

3.6 Robustness link

We operationalize "robustness" two ways, both on a held-out set of 900 KonIQ-10k images never used to fit any probe, using four corruption types (Gaussian noise/blur, JPEG [Joint Photographic Experts Group] compression, contrast reduction) at three severities, following the corruption vocabulary of Hendrycks & Dietterich (2019): (1) **representational corruption sensitivity**, $1 - \overline{\cos(f(x), f(\tilde{x}))}$ between clean and corrupted features, averaged over corruptions/severities/images (lower = representation moves less under corruption), which we also divide by the mean cosine distance between features of *different* clean images, since raw distances are not comparable across

Table 2: MOS-probe results: SROCC [95% bootstrap CI] and PLCC. n is the number of held-out test images (2,015 of 10,073 KonIQ-10k images; 2,000 of 10,125 KADID-10k images from 16 of its 81 scenes); probes are fit on the remaining 8,058 / 8,125. *Within-scene*: median over test scenes of SROCC computed within one scene's 125 distorted versions. *Hand-crafted*: 69 low-level statistics (Section 3.3).

	KonIQ-10k (authentic, n=2015)		**KADID-10k** (synthetic, n=2000)		
Backbone	SROCC [95% CI]	PLCC	SROCC [95% CI]	PLCC	Within-scene
MAE-B/16	**0.868** [0.854, 0.881]	**0.890**	0.569 [0.536, 0.604]	0.586	**0.917**
CLIP-B/16	0.855 [0.841, 0.869]	0.883	**0.777** [0.760, 0.793]	**0.787**	0.902
DINOv2-B/14	0.819 [0.801, 0.836]	0.844	0.402 [0.362, 0.441]	0.463	0.766
Sup-B/16	0.790 [0.771, 0.807]	0.822	0.548 [0.517, 0.580]	0.586	0.838
Hand-crafted	0.727 [0.704, 0.751]	0.757	0.498 [0.460, 0.532]	0.549	0.592

Table 3: Distortion-type (7-way) and severity probes on KADID-10k, n_{test}=2000. Uniform chance = 14.3%; majority-class baseline of this test split = 20.0%.

Backbone	**Type accuracy**	**Type macro-F1**	**Severity SROCC**
MAE-B/16	0.634	0.640	0.451
CLIP-B/16	0.549	0.568	**0.467**
Sup-B/16	0.502	0.492	0.356
DINOv2-B/14	0.396	0.364	0.319
Hand-crafted	**0.657**	**0.654**	0.452

backbones whose features differ in spread; and (2) **severity monotonicity**, the Spearman correlation between the quality probe's own prediction and ground-truth severity on the same images (more negative = probe more reliably tracks how corrupted an image is). We correlate each against MOS-probe SROCC across our four-backbone zoo, and report this as what it is — an n=4 comparison — not a statistically confirmatory test (Section 4.5).

4 Results

4.1 Frozen features linearly predict human quality judgments

Authentic distortions. On KonIQ-10k every backbone's output features predict human ratings better than the hand-crafted baseline (0.727), with non-overlapping confidence intervals: MAE reaches 0.868, CLIP 0.855, DINOv2 0.819 and Sup 0.790 (Table 2). All four beat the baseline in each of five random splits, where their mean SROCC is 0.80–0.88 with standard deviation $\leq$0.01 (Appendix A.2). SROCC and PLCC are correlations between predicted and human scores, not accuracies. Dedicated NR-IQA models trained end-to-end reach the low-to-high 0.90s on comparable benchmarks (Su et al., 2020; Ke et al., 2021; Yang et al., 2022); a linear probe on the frozen output is 0.03–0.11 below that, and intermediate blocks close most of the gap (Section 4.4). The supervised classifier encodes quality too, so the effect is not specific to self-supervision.

Synthetic distortions. On KADID-10k pooled SROCC is lower and the ranking changes: CLIP is strongest (0.777), MAE and Sup beat the baseline (0.498) by smaller margins, and DINOv2 (0.402) falls below it, in all five splits. The within-scene column shows that this is mainly a failure to place different scenes on a common scale, not to detect distortions: with the scene held fixed, the probes rank its distorted versions with median SROCC 0.77–0.92, against 0.59 for the baseline (Figure 2). With only 65 training scenes, a linear probe has little data from which to learn a content-independent scale.

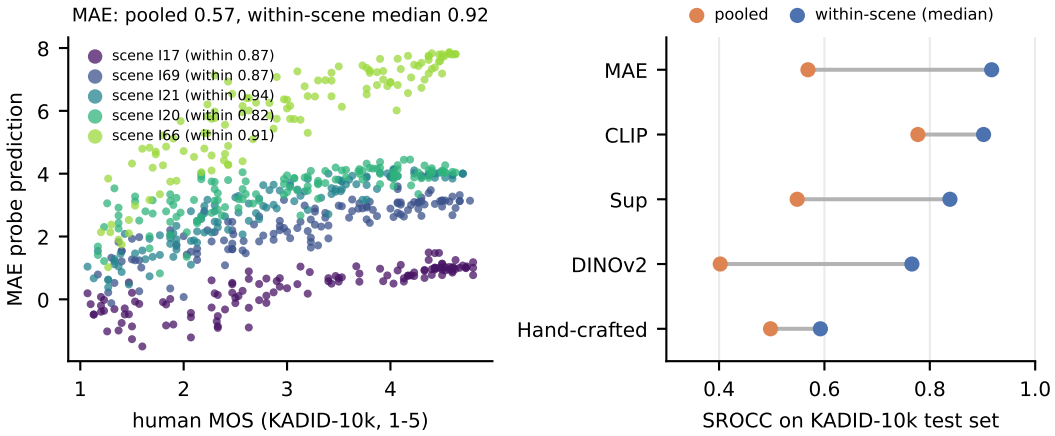

Figure 2: Within-scene versus pooled SROCC on KADID-10k. **Left:** MAE probe predictions for the distorted versions of five test scenes: within each scene predictions rise with human MOS, but each scene sits at a different offset, which lowers the pooled correlation. **Right:** pooled SROCC and median within-scene SROCC for every backbone and the hand-crafted baseline.

Table 4: Cross-dataset probe transfer (SROCC), no adaptation: probe fit on the train column's full feature set, evaluated zero-shot on the test column.

Backbone	KADID-10k → KonIQ-10k	KonIQ-10k → KADID-10k
MAE-B/16	**0.458**	0.506
CLIP-B/16	0.453	**0.661**
DINOv2-B/14	0.245	0.470
Sup-B/16	0.288	0.491
Hand-crafted	0.120	0.180

4.2 Distortion type and severity

All four backbones identify which of seven distortion categories was applied well above chance and above the majority class, from 40% (DINOv2) to 63% (MAE), and rank distortion severity with SROCC 0.32–0.47 (Table 3). They do not beat the hand-crafted baseline, which identifies the category best (66%) and matches MAE and CLIP on severity. Distortion *type* is a low-level signature that natural-scene statistics are designed to capture; the learned features' advantage lies in predicting *perceived* quality (Section 4.1), not in naming the distortion.

4.3 Cross-dataset transfer between authentic and synthetic distortions

A probe fit on one kind of distortion transfers to the other without adaptation for every backbone (0.24–0.66 SROCC; Table 4), whereas the hand-crafted baseline barely transfers (0.12 and 0.18): the quality direction the backbones provide is considerably more general than low-level statistics. Transfer is asymmetric for all four backbones, KonIQ-10k→KADID-10k exceeding the reverse by 0.05–0.23. This is consistent both with authentic distortions being a richer training signal and with the far larger number of training scenes on KonIQ-10k (8,058 versus 65); our data do not separate the two explanations.

4.4 Where quality information lives: a layer-wise view

Quality information peaks before the output. For every backbone the best block, chosen on validation data, is 7 or 8 of 12 (Figure 3, Table 5), and there the probe is markedly better than at the output: DINOv2 reaches 0.924 SROCC (output 0.819), CLIP 0.913 (0.855), Sup 0.900 (0.790), MAE 0.887 (0.868). Linear separability of a model's own training task tends to grow with depth

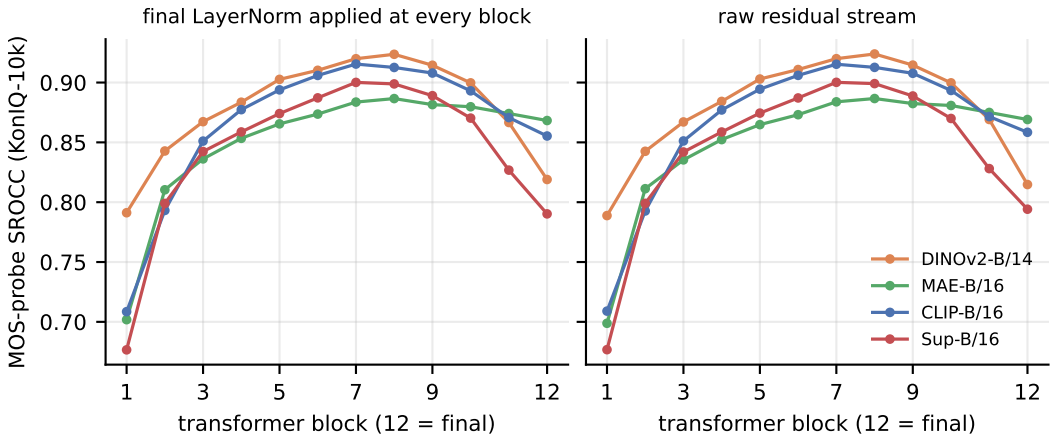

Figure 3: KonIQ-10k MOS-probe SROCC after each transformer block (12 = output). **Left:** the model's final LayerNorm is applied at every block, so block 12 is the feature used in Sections 4.1–4.3. **Right:** raw residual stream. All backbones peak at block 7 or 8, and the two panels agree closely.

Table 5: Layer-wise quality signal on KonIQ-10k (final-LayerNorm variant). The peak block is selected on a validation split carved from the training images; all values are test SROCC. "Drop" is peak minus output SROCC with a 95% paired-bootstrap CI over test images; "retained" is output/peak.

Backbone	Block 1	Peak (block)	Block 12	Drop [95% CI]	Retained
MAE-B/16	0.702	0.887 (8)	0.868	0.018 [0.011, 0.026]	**97.9%**
CLIP-B/16	0.709	0.913 (8)	0.855	0.057 [0.047, 0.069]	93.7%
DINOv2-B/14	0.791	**0.924** (8)	0.819	0.105 [0.091, 0.120]	88.7%
Sup-B/16	0.677	0.900 (7)	0.790	0.110 [0.095, 0.126]	87.8%

(Alain & Bengio, 2016); quality, which none of these objectives is trained on, instead behaves like the off-task features of generative models, which peak mid-network (Chen et al., 2020). The output-layer numbers of Section 4.1 therefore *understate* the quality information present; mid-network probes are in the range reported for dedicated NR-IQA models, although evaluation protocols differ.

Objectives differ in how much survives to the output. MAE, whose objective is to reproduce pixels, keeps 97.9% of its peak at the output (drop 0.018); CLIP loses 0.057 (93.7%); DINOv2 and the supervised classifier lose 0.105 and 0.110 (88.7%, 87.8%), the latter two with overlapping intervals. The ordering is unchanged for the raw residual stream. A natural reading is that the final blocks specialize to the pretraining objective and discard what it does not need, and only pixel reconstruction needs fine detail. We did not expect DINOv2, despite its patch-level loss, to behave like the supervised classifier, and we have not isolated objective from the other differences between these checkpoints (data, training length, recipe).

4.5 Is quality encoding linked to corruption robustness?

We also asked whether backbones that encode quality better are more stable under corruption (Section 3.6; details in Appendix A.3). MAE is both the strongest quality encoder on KonIQ-10k and the most corruption-stable. Part of its raw advantage, however, reflects the spread of its features: different clean images are unusually close together in MAE's feature space (mean cosine distance 0.38, versus 0.57–0.97 for the others), which shrinks every distance. After normalizing by this inter-image distance, MAE remains the most stable, but by 1.7–4.3× rather than 4.5–6.6×. Since lower sensitivity means more stable, the hypothesis predicts a negative correlation; across the four backbones, quality SROCC and sensitivity correlate at $\rho = -0.40$ (Spearman, raw or normalized), which is far from significant at $n = 4$ (exact permutation $p = 0.75$; even a perfect ranking would give

p=0.083). CLIP, the second-best quality encoder, is the least stable. We therefore treat the link as a hypothesis, not a finding.

5 DISCUSSION

What the results say. Quality information is present in all four frozen representations and exceeds what low-level statistics provide wherever quality must be *judged* rather than *catalogued*: authentic distortions, distortions within a scene, and transfer between the two regimes. It is not specific to self-supervision, it is richest in intermediate blocks, and how much reaches the output depends on the objective: only pixel reconstruction keeps nearly all of it. Because objective is confounded with data and recipe in our four checkpoints, this last point is a hypothesis rather than a causal claim.

DINOv2. DINOv2 is the best quality encoder at mid-network (0.924 at block 8), yet it is the weakest on synthetic KADID-10k, below the hand-crafted baseline, and transfers worst from KADID-10k to KonIQ-10k (0.245). Its multi-crop self-distillation explicitly teaches invariance to photometric and geometric perturbations that several KADID-10k distortions resemble (colour diffusion, contrast change, pixelation), which would fit quality information being present mid-network but suppressed by the final blocks. Testing this needs the layer-wise sweep on KADID-10k, which we have not run.

Practical implication. A system that already uses one of these backbones can recover quality sensitivity with a single linear layer and no extra pretraining, and reading a late-middle block instead of the output adds up to 0.10 SROCC (DINOv2, $0.819 \rightarrow 0.924$).

6 LIMITATIONS

- **Backbones.** Four backbones at one scale (ViT-B), with objective confounded with pretraining data and recipe; newer backbones (e.g. DINOv3, SigLIP 2) are not included, and we do not compare against quality-aware pretraining (Zhao et al., 2023), for which we found no public checkpoint.
- **Content.** The within-scene analysis controls for content on KADID-10k only; on KonIQ-10k content and quality are entangled, so part of what the probes exploit may be content that correlates with quality.
- **Robustness link.** Four data points, an internal feature-distance measure that depends on feature geometry even after normalization, and no cross-check against published ImageNet-C results.
- **Scope of analyses.** Two benchmarks; the layer-wise sweep covers KonIQ-10k only; probes are linear by design and use one pooling scheme; the hand-crafted baseline is our reimplementation, not a tuned classical NR-IQA model.

7 CONCLUSION

Frozen, general-purpose vision backbones encode human-perceived image quality well beyond low-level statistics, although none was trained to: a linear probe recovers it on authentic distortions, within scenes on synthetic ones, and across the two regimes. The information peaks in intermediate blocks, and only pixel reconstruction carries nearly all of it to the output. Hand-crafted statistics remain better at naming the distortion, and a link to corruption robustness is not supported at our scale. Read at the right depth, frozen foundation features are a cheap and strong starting point for quality assessment.

AI USE STATEMENT

In this work we used generative AI tools (Claude, Anthropic, a large language model with code-execution tools) for the following tasks with required disclosure: designing the experiments and giving feedback on them, including a simulated peer review (probing protocol, hand-crafted baseline, within-scene analysis, validation-based block selection, statistical tests); implementing all methods (feature extraction, probes, baselines, analyses); cleaning and reformatting the datasets (loaders,

label parsing, and the correction of a label-orientation error); generating the synthetic corruptions used in the robustness analysis; and interpreting results. We did not use generative AI to formulate or prove mathematical claims, for translation, or for qualitative or thematic data analysis; these do not apply. Additionally, we used it to draft and edit the paper, create all figures and the supplementary code, choose hyperparameter grids and corruption severities, suggest the paper's structure, title and keywords, search for, summarize and identify related work, and format references. The author chose the research direction, refined the hypotheses, and directed Claude in executing the experiments, making meticulous corrections, additions and rewirings of components at different stages, and reviewed all AI-assisted work. Code was checked with automated tests and consistency checks (for example, the last layer-wise block must reproduce the main features exactly, which exposed a pooling error that we then fixed), and every number in the paper is produced by scripts from saved result files. We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

Ethics statement

This work reuses two existing public IQA datasets and their existing human-subject annotations; we collected no new human-subject data. Both datasets describe their own collection and consent procedures in their original publications, and we are not aware of a specific risk of harm beyond IQA research's general dual-use considerations.

Reproducibility statement

All backbones are public pretrained checkpoints (Appendix A.4) used strictly frozen, and both datasets are public. The probing protocol is specified in Section 3; the layer-wise, split-variability, robustness and baseline-feature details are in Appendix A.2–A.4. Splits use fixed seeds; the anonymous supplementary material contains all code, the exact splits, and the result files behind every table.

References

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

Ishita Bardhan, Sumohana Channappayya, Abhik Banerjee, Abdur Rahim Mohammad Forkan, Prem Jayaraman, and Abhinav Kumar. Are standard CNNs good enough for no-reference stereoscopic image quality assessment? In *2024 International Conference on Signal Processing and Communications (SPCOM)*, 2024. doi: 10.1109/SPCOM60851.2024.10631608.

Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pp. 1691–1703, 2020.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16000–16009, 2022.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations (ICLR)*, 2019.

Vlad Hosu, Hanhe Lin, Tamas Sziranyi, and Dietmar Saupe. KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on Image Processing*, 29:4041–4056, 2020.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5148–5157, 2021.

Hanhe Lin, Vlad Hosu, and Dietmar Saupe. KADID-10k: A large-scale artificially distorted IQA database. In *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 1–3, 2019.

Pavan C. Madhusudana, Neil Birkbeck, Yilin Wang, Balu Adsumilli, and Alan C. Bovik. Image quality assessment using contrastive learning. *IEEE Transactions on Image Processing*, 31, 2022.

Markus Marks, Manuel Knott, Neehar Kondapaneni, Elijah Cole, Thijs Defraeye, Fernando Perez-Cruz, and Pietro Perona. A closer look at benchmarking self-supervised pre-training with image classification. *International Journal of Computer Vision*, 133:5013–5025, 2025.

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

Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2013.

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. *Transactions on Machine Learning Research*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pp. 8748–8763, 2021.

Avinab Saha, Sandeep Mishra, and Alan C. Bovik. Re-IQA: Unsupervised learning for image quality assessment in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5846–5855, 2023.

Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your ViT? data, augmentation, and regularization in vision transformers. *Transactions on Machine Learning Research*, 2022.

Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3667–3676, 2020.

Jianyi Wang, Kelvin C. K. Chan, and Chen Change Loy. Exploring CLIP for assessing the look and feel of images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 2555–2563, 2023.

Ankit Yadav, Ta Duc Huy, and Lingqiao Liu. Revisiting vision language foundations for no-reference image quality assessment. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5416–5425, 2026.

Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. MANIQA: Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1191–1200, 2022.

Weixia Zhang, Kede Ma, Jia Yan, Dexiang Deng, and Zhou Wang. Blind image quality assessment using a deep bilinear convolutional neural network. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(1):36–47, 2020.

Kai Zhao, Kun Yuan, Ming Sun, Mading Li, and Xing Wen. Quality-aware pretrained models for blind image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22302–22313, 2023.

A Appendix

A.1 Example images

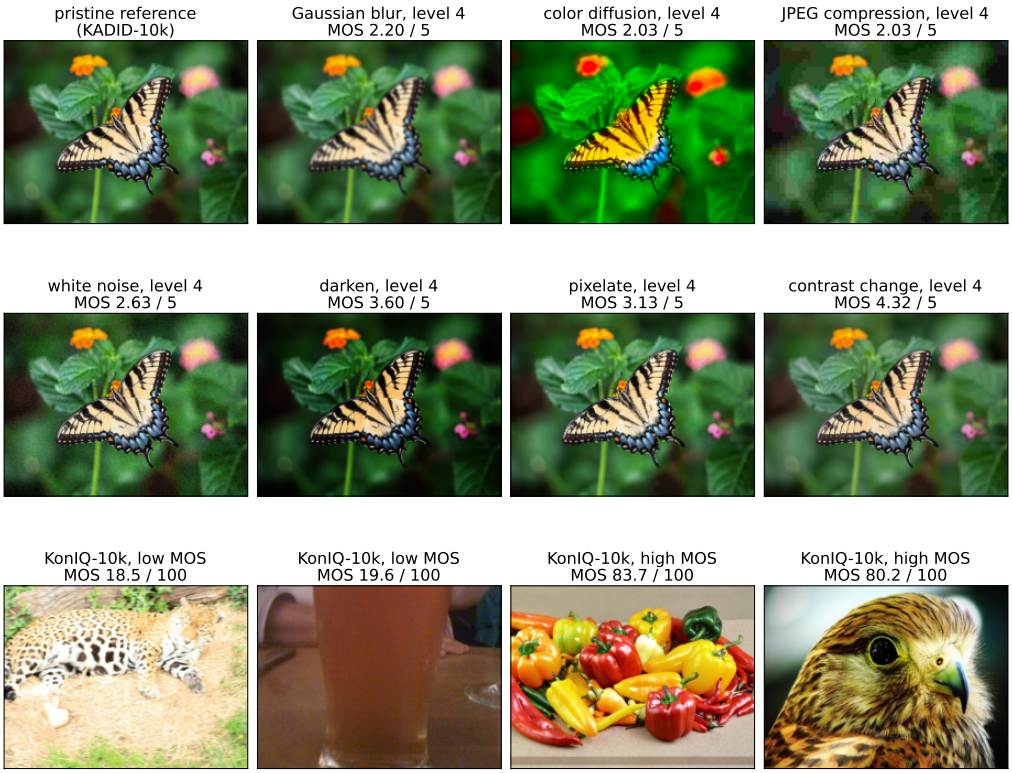

Figure 4: Example images. **Top two rows:** one KADID-10k pristine reference image and seven of its synthetically distorted versions (one distortion type per coarse category, all at severity level 4 of 5), with the human MOS on a 1–5 scale (higher is better). Equal severity levels are not equal perceived quality: contrast change at level 4 is rated 4.32, JPEG compression 2.03. **Bottom row:** KonIQ-10k photographs (authentic distortions, no pristine reference exists) with low and high MOS on a 0–100 scale.

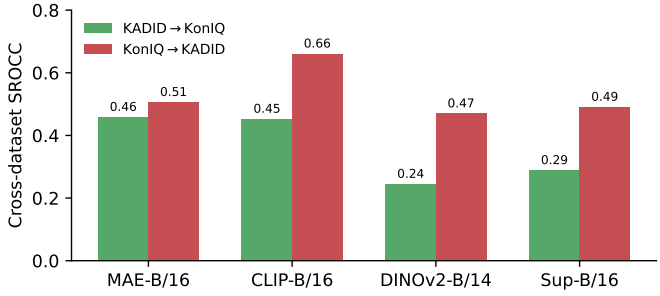

Figure 5: Cross-dataset transfer (same numbers as Table 4): positive in both directions for every backbone, and consistently stronger authentic→synthetic than the reverse.

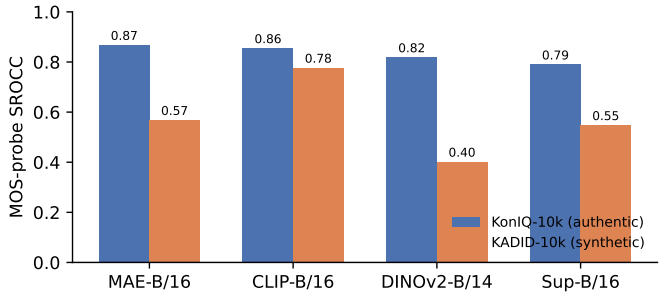

Figure 6: MOS-probe SROCC by backbone and dataset. All four objectives are well above a trivial baseline on both benchmarks despite zero quality-specific training.

Table 6: MOS-probe SROCC over five random content-disjoint splits (mean $\pm$ standard deviation; split 0 is the one used in Table 2), and the number of splits in which each backbone beats the hand-crafted baseline.

	KonIQ-10k		**KADID-10k**	
Backbone	SROCC	beats baseline	SROCC	beats baseline
MAE-B/16	0.876 ± 0.008	5/5	0.608 ± 0.059	5/5
CLIP-B/16	0.866 ± 0.008	5/5	0.718 ± 0.048	5/5
DINOv2-B/14	0.826 ± 0.007	5/5	0.349 ± 0.058	0/5
Sup-B/16	0.802 ± 0.010	5/5	0.522 ± 0.047	4/5
Hand-crafted	0.743 ± 0.014	—	0.488 ± 0.013	—

A.2 Variability over random splits

On KonIQ-10k the ordering MAE > CLIP > DINOv2 > Sup holds in four of five splits (CLIP and MAE swap in one). On KADID-10k, where each test split contains only 16 scenes, variability is about five times larger and CLIP ranks first in four of five splits.

A.3 Corruption-robustness details

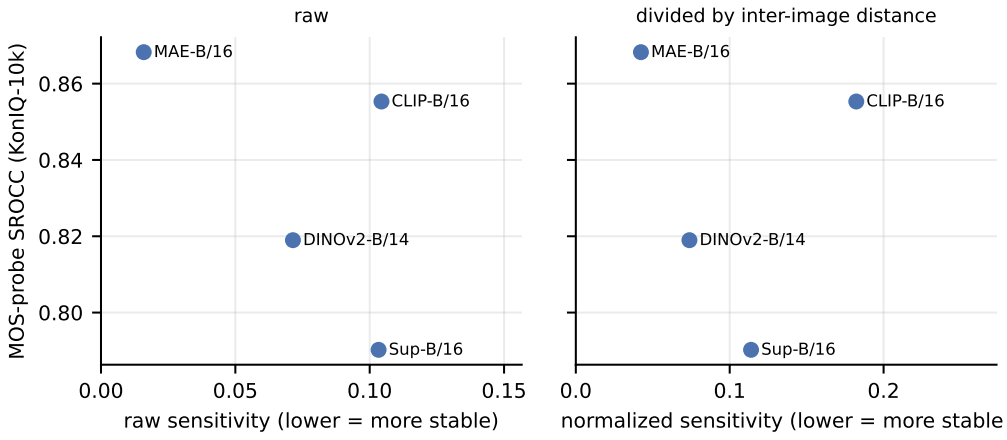

Figure 7: Quality-probe SROCC (KonIQ-10k) against corruption sensitivity, raw (left) and divided by the mean cosine distance between different clean images (right). The backbone ordering, and hence Spearman $\rho = -0.40$ (exact $p = 0.75$), is the same in both.

Table 7: Robustness measures on 900 held-out KonIQ-10k images. Sensitivity is $1-\overline{\cos(f(x), f(\tilde{x}))}$ over four corruptions at three severities; inter-image distance is the mean cosine distance between features of different clean images. Severity monotonicity (Spearman between the quality probe's prediction and corruption severity; more negative is better) correlates with quality SROCC at ρ=0.80 (exact p=0.33), but is partly circular with Table 2.

Backbone	Quality SROCC	Sensitivity	Inter-image dist.	Normalized	Severity mono.
MAE-B/16	0.868	0.016	0.376	0.042	−0.532
CLIP-B/16	0.855	0.104	0.573	0.182	−0.544
DINOv2-B/14	0.819	0.071	0.968	0.074	−0.428
Sup-B/16	0.790	0.103	0.907	0.114	−0.322

A.4 Additional experimental detail

Hand-crafted baseline features (69 per image, computed on luminance): at each of two scales (full and half resolution), the mean, standard deviation, skewness, kurtosis and the standard deviations of the negative and positive parts of the MSCN coefficients and of their products with the horizontal, vertical and two diagonal neighbours ($2 \times 5 \times 6 = 60$); plus the log-variance of the Laplacian (sharpness), mean gradient magnitude, mean and standard deviation of luminance, the Hasler–Süsstrunk colourfulness index, mean saturation, luminance entropy, a median-absolute-deviation noise estimate, and a JPEG blockiness ratio (mean horizontal gradient at 8-pixel boundaries over elsewhere).

Checkpoints (`timm` names; all loaded pretrained and frozen): DINOv2-B/14 `vit_base_patch14_dinov2.lvd142m`; MAE-B/16 `vit_base_patch16_224.mae`; CLIP-B/16 `vit_base_patch16_clip_224.openai`; Sup-B/16 `vit_base_patch16_224.augreg2_in21k_ft_in1k`. Features are extracted under half-precision (fp16) autocast on a GPU.

Corruption implementations (Section 3.6): Gaussian noise is additive with $\sigma \in \{0.04, 0.10, 0.20\}$ (severities $1, 3, 5$) on images scaled to $[0, 1]$; Gaussian blur uses a Gaussian kernel with $\sigma \in \{0.6, 1.6, 3.4\}$; JPEG uses quality factors $\{40, 22, 8\}$; contrast reduction scales pixel deviation from the per-image mean by $\{0.75, 0.45, 0.2\}$.

KADID-10k distortion categories: blur (Gaussian, lens, motion blur; 3 types); color (diffusion, shift, quantization, two saturation variants; 5 types); compression (JPEG2000, JPEG; 2 types); noise (white noise, color-component noise, impulse noise, multiplicative noise, denoising artifacts; 5 types); brightness (brighten, darken, mean shift; 3 types); spatial (jitter, non-eccentricity patch, pixelate, quantization, color block; 5 types); sharpness/contrast (high sharpen, contrast change; 2 types).

A.5 Full per-metric results

Kendall's τ and RMSE for every backbone/dataset cell in Table 2, and severity-regression RMSE for Table 3, are included in the released result files (`mos_probe.csv`, `distortion_probe.csv`) and omitted here only for space.