IS ONE DIRECTION ENOUGH? VISUAL EMBEDDINGS RANKABILITY FOR IMAGE QUALITY ASSESSMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Image quality assessment (IQA) models, from classic deep neural networks to multimodal large language models (MLLMs), are evaluated by how well their predicted scores correlate with human mean opinion scores (MOS). However, this evaluation does not reveal that the hidden representations are rankable by quality scores. In this work, we study representations derived from hidden layers of IQA models and propose to estimate a single linear direction, the *rank-axis*, and measure whether projections onto it order images by quality scores (*rankability*). Additionally, we propose the *Rank-Axis Isometry Score* (RAIS) to quantify how well their spacing reflects the size of quality differences. To test whether the embedding space itself is ordered by quality scores, we introduce three quality-embedding alignment measures computed on distances in the native feature space. Across nine IQA models, three layer depths and eight benchmarks, one rank-axis reaches a rankability of at least 0.84 on six datasets, and at least one of the rank-axes across hidden representations beats the model's own fine-tuned head in 52 of 70 encoder-dataset pairs. Spacing along the axis follows the order (Spearman 0.98 between RAIS and rankability), whereas the embeddings in the original latent space are only weakly ordered by quality (mean Anchor-QEA 0.35). Rank-axes estimated on different datasets are nearly orthogonal, yet keep 80% of their rankability on unseen datasets; their agreement under the data covariance predicts this transfer (Spearman 0.96). Our findings suggest that a single linear direction captures much of the perceptual quality ordering in visual embeddings, even when it is not reflected in their dominant geometry.

1 INTRODUCTION

Image quality assessment (IQA) objective metrics have moved from hand-crafted natural-scene statistics (Mittal et al., 2012; 2013) to deep networks (Bosse et al., 2018; Zhang et al., 2020; Ke et al., 2021; Yang et al., 2022), CLIP-based models (Wang et al., 2023a; Zhang et al., 2023; Agnolucci et al., 2024a) and multimodal large language models (MLLMs) that rate images in text (Wu et al., 2024; Li et al., 2025; Wu et al., 2025). They are compared almost exclusively by the correlation between the scores of their heads and human mean opinion scores (MOS). The comparison (Zhu et al., 2020; Yang et al., 2023) based on correlation score do not reveal whether the representation lacks quality-aware information or the prediction head fails to use it.

In this work, we study the hidden representations with a question: *do frozen IQA representations order images by perceptual quality either in their original embedding space or along a single linear direction?* Sonthalia et al. (2025) show that visual embeddings often order images by attributes such as age or aesthetics in this way. For each representation and dataset, we estimate one such direction, the *rank-axis*, and measure its *rankability*: the rank correlation between the projections of test images onto the axis and their quality labels. Figure 1 illustrates how the learned rank-axis can reorganize the embedding along a direction that closely follows perceptual quality: images with different quality scores are pushed apart, while images with similar scores are brought together.

We investigate three questions that the existing rankability framework (Sonthalia et al., 2025) does not address. First, does the rank-axis capture the magnitude of quality score differences, not just their ordering? Rankability considers only projection order, ignoring whether distances along the axis reflect differences in quality scores. Second, does a quality structure exist in the embedding beyond a

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.

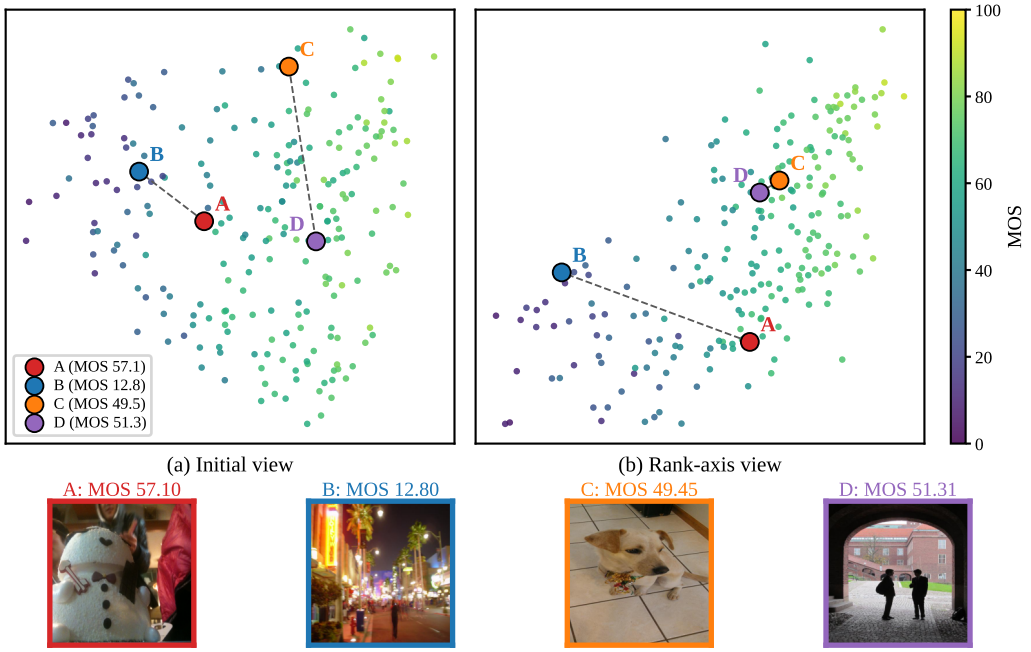

Figure 1: **Test images of LIVEC in the ℓ_{last} embedding of LIQE.** (a) Initial view: the two leading principal components of the embeddings. (b) Rank-axis view: the projection onto the rank-axis (horizontal) against the leading principal component orthogonal to it (vertical). Colour: mean opinion score (MOS). Dashed lines join the pairs A, B (different MOS) and C, D (similar MOS), which are shown below with their MOS.

single ordering direction, such that similarly rated images are also geometrically close? Third, is the quality rank-axis transferable across datasets? Since IQA benchmarks differ in content, distortions, and rating protocols, cross-dataset ordering tests whether representations capture a shared quality direction or require dataset-specific directions.

We address these questions for nine publicly released IQA models that cover classic deep neural networks, CLIP-based models and MLLMs (Appendix C, Table 22). Features are extracted at three layer depths of each model, on eight benchmarks with synthetic, restoration, authentic and AI-generated distortions (Appendix C, Table 21), and each rank-axis is compared with the model's own fine-tuned head. Our contributions are the following.

- **Visual embeddings exhibit ordering along the rank axis.** At ℓ_{last}, one rank-axis reaches $\mathcal{R}_\rho \geq 0.84$ on six of the eight datasets. At some layer, a rank-axis fit on the target's training split beats the model's fine-tuned head in 52 of 70 encoder-dataset pairs, and in 39 of 70 already at ℓ_{early} (Section 5.1).
- **Metrics for spacing and geometry.** We introduce the Rank-Axis Isometry Score (RAIS), which is maximal exactly for axes whose spacing is affine in quality (Proposition 1), and three measures of quality-embedding alignment on the native feature space (Section 3). Spacing follows order (Spearman 0.98 between RAIS and $\mathcal{R}_\rho$), but the embedding space is only weakly ordered by quality (mean Anchor-QEA 0.35): the order is largely induced by the projection (Section 5.3).
- **Transfer follows covariance-weighted agreement.** Rank-axes fit on different datasets are nearly orthogonal (median 86°), yet keep 80% of their rankability on unseen datasets. The covariance-weighted agreement c_Σ agrees with the cross-dataset transfer (Spearman 0.96)(Sections 5.2 and 5.4).

2 Related work

Rankability and ordered embeddings. Sonthalia et al. (2025) define the rankability of a visual embedding with respect to a continuous attribute as the extent to which a single linear direction, the rank-axis, orders images by that attribute. They show that many modern embeddings are rankable for attributes such as age, crowd counts, head pose and aesthetics, often from few labelled examples, and that rank-axes can transfer across datasets. Other work imposes order during training. OrdinalCLIP and L2RCLIP learn rank prompts that adapt CLIP to ordinal regression (Li et al., 2022; Wang et al., 2023b), RankIQA and HiRQA train IQA networks with ranking losses on image pairs (Liu et al., 2017; Ramesh et al., 2026), and Rank-N-Contrast learns embeddings in which distances follow label differences (Zha et al., 2023). Rankability has not been explored much for perceptual image quality, nor has it been examined whether spacing and distances in the embedding reflect the magnitude of ordered attribute gaps.

IQA models and their evaluation. IQA models have evolved from natural-scene statistics (Mittal et al., 2012; 2013) to CNNs and transformers (Bosse et al., 2018; Zhang et al., 2020; Ke et al., 2021; Yang et al., 2022; Agnolucci et al., 2024b), to CLIP-based models that score images through quality prompts (Wang et al., 2023a; Zhang et al., 2023; Agnolucci et al., 2024a), and to MLLMs that rate images in text (Wu et al., 2024; Li et al., 2025; Wu et al., 2025). Across these families, models are compared by the correlation between their predicted scores and human opinion scores, within and across datasets (Zhu et al., 2020; Yang et al., 2023). To our knowledge, no established metric measures how well the frozen representations of IQA models order images by quality scores, or how the geometry of these representations relates to quality score differences. An extended review is given in Appendix E.

3 Methodology

Evaluating an IQA model only through its final quality scores against human opinions does not reveal whether its hidden representations are also ordered by quality score. In this section, we study the representation space through four properties of increasing strength. *Order* asks whether a single linear direction sorts images by quality scores (Section 3.2). *Spacing* asks whether distances along this direction grow with quality score differences (Section 3.3). *Geometry* asks whether the embedding space itself, without any projection, is organised by quality scores (Section 3.4). *Transfer* asks whether a direction found on one dataset orders the images of another, and what property of the two directions determines this (Section 3.5).

3.1 Notation

A *representation* f is a frozen encoder taken at a fixed layer, mapping an image I to a feature vector $\mathbf{x} = f(I) \in \mathbb{R}^d$. Each IQA dataset contains images with quality labels $y \in \mathbb{R}$ (MOS/DMOS) and is divided into disjoint training and test splits; for a dataset a, the stacked features and labels are $\mathbf{X}_a^{\text{tr}}$, $\mathbf{y}_a^{\text{tr}}$ and $\mathbf{X}_a^{\text{te}}$, $\mathbf{y}_a^{\text{te}}$. The letters a and b index *datasets*: a is the *source* dataset, whose training split is used to find a direction, and b is the *target* dataset, on whose test split it is evaluated ($a = b$ in the in-dataset setting). ρ_S and corr denote the Spearman and Pearson correlations.

3.2 Rank-axis and rankability

Most uses of a quality score, such as comparing restoration methods or filtering a collection, depend on the *order* it induces on images. The simplest way a representation can encode this order is as a direction in its feature space: moving along the direction, images become better.

Definition 1 (Rank-axis)**.** *A unit vector $\mathbf{w} \in \mathbb{S}^{d-1}$ is a* rank-axis *of the representation f on a dataset if the projection score $s(I) = \mathbf{w}^\top f(I)$ preserves the quality score order of its images: $y(I) > y(I')$ implies $s(I) > s(I')$ for every pair of images I, I'. The representation is said to be* rankable *with respect to the rank-axis $\mathbf{w}$ if $\mathbf{w}$ satisfies the above ordering condition.*

A rank-axis is a linear projection; it cannot introduce nonlinear interactions among features. Rankability therefore indicates that the representation's geometry already encodes the quality ordering.

However, human quality scores are noisy, making exact order preservation an unrealistic criterion and limiting its direct verification on unseen images. We therefore estimate the rank-axis on the training split of a source dataset and measure how well it orders held-out test images.

Estimating the rank-axis. Among all directions, we seek the one whose projections best predict quality. For a source dataset a, we take the normalised ridge-regression direction (Hoerl & Kennard, 1970)

$$(\mathbf{w}_a, \beta_a) = \arg\min_{\mathbf{w},\,\beta} \left\|\mathbf{y}_a^{\text{tr}} - \mathbf{X}_a^{\text{tr}}\mathbf{w} - \beta\mathbf{1}\right\|_2^2 + \alpha\|\mathbf{w}\|_2^2, \qquad \hat{\mathbf{w}}_a = \mathbf{w}_a/\|\mathbf{w}_a\|_2, \tag{1}$$

where β is an intercept and the penalty α stabilises the estimate when the feature dimension is large relative to the number of training images. The projection score of an image is $s = \mathbf{x}^\top\hat{\mathbf{w}}_a$. Every metric below is unchanged when s is replaced by $\lambda s + c$ with $\lambda > 0$; neither the intercept nor the norm of $\mathbf{w}_a$ therefore affects any result. The visualizations of the rank-axis is provided in Appendix A.6.

Rankability. The *rankability* of the source axis on a target dataset is the rank correlation between the projection scores of the target's test images and their quality labels:

$$\mathcal{R}_\rho(a \to b) = \rho_S\big(\mathbf{X}_b^{\text{te}}\hat{\mathbf{w}}_a,\ \mathbf{y}_b^{\text{te}}\big). \tag{2}$$

For untied labels, $\mathcal{R}_\rho = 1$ exactly when $\hat{\mathbf{w}}_a$ is a rank-axis of the target's test images according to Definition 1. We also report the Pearson ($\mathcal{R}_r$) and Kendall ($\mathcal{R}_\tau$) variants. *In-dataset* rankability is $\mathcal{R}(b, b) = \mathcal{R}(b \to b)$, and *transfer* is $\mathcal{R}(a \to b)$ with $a \neq b$.

3.3 Spacing along the axis: the Rank-Axis Isometry Score

Rankability is invariant to any strictly increasing transformation of the projection scores. It certifies the *order* of images along the axis, but not whether distances along the axis reflect the size of quality differences. Two axes can rank images equally well while one spaces them in proportion to their quality gaps and the other compresses part of the quality range. To measure spacing, we compare distances along the axis with quality gaps. For two images I, I' of the target test split, let $\delta = |s(I) - s(I')|$ be their distance along the rank-axis and $\Delta = |y(I) - y(I')|$ their quality gap. The Rank-Axis Isometry Score is the rank correlation between the two over pairs of images,

$$\text{RAIS}(a \to b) = \rho_S\big(\delta,\ \Delta\big). \tag{3}$$

RAIS is invariant to positive rescaling and shifts of both s and y, and is therefore independent of the label scale of a dataset. Because it uses absolute differences, it does not change when the axis is reversed; it is therefore always reported together with $\mathcal{R}_\rho$, which carries the sign. Unlike $\mathcal{R}_\rho$, RAIS is not invariant to nonlinear monotone transformations of s, and this is exactly what it measures:

Proposition 1 (RAIS characterises affine rank-axes)**.** *Suppose the quality labels have a continuous distribution supported on an interval J, and the projection is a deterministic function of quality, $s = g(y)$, with g continuous and strictly monotone on J (so that $|\mathcal{R}_\rho| = 1$). Then* $\text{RAIS} = 1$ *if and only if g is affine on J.*

Among axes that order images perfectly, RAIS is maximal exactly for those whose spacing is proportional to quality differences, that is, isometric up to a global scale (proof in Appendix B). RAIS decreases when spacing along the axis is nonlinear in quality.

3.4 Quality geometry of the embedding space

Rankability and RAIS are computed on projections onto a fitted direction. They are invariant to rotations of the feature space, and to other invertible linear maps up to the effect of the ridge penalty. They therefore show that quality is *linearly recoverable*, but not that the network's own geometry is organised by quality: an embedding can place images of very different quality side by side and still contain a direction that separates them (Figure 1). To test the geometry itself, we compare distances between feature vectors, with no projection and no fitted component, to quality gaps. Let $D_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|_2$ be the Euclidean distance between the features of images i and j, and $\Delta_{ij} = |y_i - y_j|$ their quality gap. We define three measures of quality-embedding alignment (QEA) at different scales: from the viewpoint of each image, on single comparisons, and over the whole space.

Anchor-QEA. Distances in an embedding need not share one scale across the space: some regions can be denser than others, independently of quality. Comparing only distances that share an endpoint avoids mixing these scales. For a set $\mathcal{A}$ of anchor images and, for each anchor i, a set $\mathcal{N}_i$ of other images,

$$\text{Anchor-QEA} = \frac{1}{|\mathcal{A}|} \sum_{i \in \mathcal{A}} \rho_S \Big(\{D_{ij}\}_{j \in \mathcal{N}_i},\ \{\Delta_{ij}\}_{j \in \mathcal{N}_i} \Big). \tag{4}$$

Anchor-QEA equals 1 when, seen from every anchor, the other images lie at distances ordered by their quality gap. It corresponds to the retrieval question "is image j or image k closer in quality to i?".

Ordering rate. Anchor-QEA summarises many comparisons in one correlation that is hard to interpret in absolute terms. The ordering rate, inspired by the δ-ordered embeddings of Rank-N-Contrast (Zha et al., 2023), counts single comparisons instead, with a chance level of 0.5. Let $\mathcal{T}$ be a set of triplets (i, j, k) of distinct images, labelled so that $\Delta_{ij} < \Delta_{ik}$ (triplets with $\Delta_{ij} = \Delta_{ik}$ are discarded). The ordering rate is the fraction of triplets in which the image closer in quality is also closer in the embedding, counting ties as $1/2$:

$$\text{OR} = \frac{1}{|\mathcal{T}|} \sum_{(i,j,k) \in \mathcal{T}} \Big(\mathbf{1}\big[D_{ij} < D_{ik}\big] + \tfrac{1}{2}\, \mathbf{1}\big[D_{ij} = D_{ik}\big] \Big). \tag{5}$$

Global-QEA. Anchor-QEA and the ordering rate only compare distances that share an endpoint. Global-QEA, inspired by representational similarity analysis (Kriegeskorte et al., 2008), pools pairs across all images and tests whether the same quality gap corresponds to similar distances everywhere in the space. For a set $\mathcal{P}$ of pairs of distinct images,

$$\text{Global-QEA} = \rho_S \Big(\{D_{ij}\}_{(i,j) \in \mathcal{P}},\ \{\Delta_{ij}\}_{(i,j) \in \mathcal{P}} \Big). \tag{6}$$

Because it ranks all distances jointly, Global-QEA is also sensitive to differences in scale and density across regions of the space.

Remark 1 (Relation to ordered embeddings)**.** *If an embedding is δ-ordered as defined by Zha et al. (2023) and no anchor has tied quality gaps, then every within-anchor correlation in Eq. (4) equals 1 and Anchor-QEA $= 1$. The converse fails: local ordering implies neither global alignment nor isometry. For example, labels $y = (0, 1, 5, 7)$ embedded at $(0, 10, 21, 23)$ are ordered from every anchor (Anchor-QEA $= 1$, $\mathcal{R}_\rho = 1$), yet Global-QEA $= \text{RAIS} \approx 0.886$.*

Appendix A.6 visualises how the test embeddings are arranged around the rank-axis for every encoder and dataset.

3.5 Cross-dataset transfer and axis agreement

IQA datasets differ in content, distortion types and rating protocols. If a representation holds a single quality direction, rank-axes estimated on different datasets should point the same way and transfer to each other; if each dataset induces its own direction, they should not. Transfer $\mathcal{R}(a \to b)$ in Eq. (2) measures the outcome. To explain it, we compare the source axis $\hat{\mathbf{w}}_a$ with the target's own axis $\hat{\mathbf{w}}_b$, both estimated in the same feature space, through two angles. The raw angle between the two rank-axes is

$$\theta_{ab} = \arccos\left(\hat{\mathbf{w}}_a^\top \hat{\mathbf{w}}_b\right) \in [0^\circ, 180^\circ]. \tag{7}$$

It captures whether the two datasets define the same quality direction in feature space: $\theta_{ab} = 0^\circ$ for identical axes, 90° for axes with no common component, and more than 90° for axes that point in opposing directions. It is computed from the two weight vectors alone and treats all d feature directions equally.

Covariance-weighted agreement. Most directions of a high-dimensional feature space carry little variation across real images. A component of a rank-axis along such a direction changes the raw angle but hardly changes any projection score. Two axes can therefore be nearly orthogonal and still order the same images almost identically. We measure the angle in the geometry of the data instead. Let Σ_b be the feature covariance of the target's test images, and define the inner product

$\langle \mathbf{u}, \mathbf{v} \rangle_{\Sigma_b} = \mathbf{u}^\top \Sigma_b \mathbf{v}$, which weights each feature direction by the variance of the images along it. The covariance-weighted agreement is the cosine of the angle between the two axes under this inner product, which equals the correlation of their projection scores on the target images:

$$c_\Sigma(a, b) = \frac{\hat{\mathbf{w}}_a^\top \Sigma_b \hat{\mathbf{w}}_b}{\sqrt{\hat{\mathbf{w}}_a^\top \Sigma_b \hat{\mathbf{w}}_a \; \hat{\mathbf{w}}_b^\top \Sigma_b \hat{\mathbf{w}}_b}} = \mathrm{corr}\big(\mathbf{X}_b^{\mathrm{te}} \hat{\mathbf{w}}_a, \; \mathbf{X}_b^{\mathrm{te}} \hat{\mathbf{w}}_b\big). \tag{8}$$

$c_\Sigma = 1$ when the two axes order the target images identically, 0 when their scores are uncorrelated, and negative when they order them in opposite ways. Unlike θ_{ab}, it ignores directions in which the target images do not vary. We also use its Spearman variant.

Relating agreement to transfer. Since c_Σ and transfer are correlations of projections on the same test images, their relation is partly expected by construction. We isolate the remaining empirical effect using the Pearson variant $\mathcal{R}_r$. Let $\hat{\mathbf{y}}_b = \mathbf{X}_b^{\mathrm{te}} \mathbf{w}_b^\star$ be the least-squares predictor of $\mathbf{y}_b^{\mathrm{te}}$. Since its residual is uncorrelated with every linear projection of $\mathbf{X}_b^{\mathrm{te}}$,

$$\mathrm{corr}\big(\mathbf{X}_b^{\mathrm{te}} \hat{\mathbf{w}}_a, \mathbf{y}_b^{\mathrm{te}}\big) = \mathrm{corr}\big(\mathbf{X}_b^{\mathrm{te}} \hat{\mathbf{w}}_a, \hat{\mathbf{y}}_b\big) \cdot \mathrm{corr}\big(\hat{\mathbf{y}}_b, \mathbf{y}_b^{\mathrm{te}}\big). \tag{9}$$

Replacing $\mathbf{w}_b^\star$ with the target's training-set rank-axis $\hat{\mathbf{w}}_b$ yields

$$\mathcal{R}_r(a \to b) \;\approx\; c_\Sigma(a, b)\, \mathcal{R}_r(b, b). \tag{10}$$

Thus, transfer is largely explained by axis agreement and target rankability; the residual captures information contributed by the foreign axis beyond the target's own axis. In particular, a positive residual indicates that the foreign axis captures quality information missed by $\hat{\mathbf{w}}_b$. We also report cases where foreign axes *beat the target's own axis*, $\mathcal{R}_r(a \to b) > \mathcal{R}_r(b, b)$.

4 Experimental Setup

Datasets. We use eight IQA benchmarks in four categories: synthetic distortions (LIVE (Sheikh et al., 2006), TID2013 (Ponomarenko et al., 2013), KADID-10k (Lin et al., 2019)), restoration and GAN outputs (PIPAL (Jinjin et al., 2020)), authentic distortions (LIVEC (Ghadiyaram & Bovik, 2016), KonIQ-10k (Hosu et al., 2020), SPAQ (Fang et al., 2020)) and AI-generated images (AGIQA-3K (Li et al., 2024)). Each dataset has one fixed train/test split shared by all encoders. Rank-axes are fit on the train split, and all results are computed on the test split (Appendix C), with results based on the no-reference image quality assessment (NR-IQA) setup.

Encoders and layers. We study nine publicly released IQA models spanning CNNs (ARNIQA (Agnolucci et al., 2024b), MANIQA (Yang et al., 2022)), CLIP-based models (LIQE (Zhang et al., 2023), QualiCLIP (Agnolucci et al., 2024a), FGResQ (Sheng et al., 2026)) and MLLMs (Q-Align (Wu et al., 2024), Q-ReAlign (Team, 2026), Q-Insight (Li et al., 2025), VisualQuality-R1 (Wu et al., 2025)). We extract $\ell_{\text{early}}, \ell_{\text{mid}}, \ell_{\text{last}}$ at roughly one-third, two-thirds and final depth, plus the vision-tower output ℓ_{vis} for three MLLMs. Features are mean-pooled except for MLLM language features, where we use the last prompt-token state. In-distribution datasets are marked * in Tables 3–4; fine-tuned-head SRCC is reported as a reference. Further details appear in Appendices C–D.

Rank-axis and metrics. Each rank-axis is estimated with Eq. (1) on the train split of one dataset, with the penalty α chosen from 12 log-spaced values in $[10^{-1}, 10^4]$ by efficient leave-one-out cross-validation. The metrics use all test images, with sampling limits on the three largest test sets (Appendix C). Cross-dataset analyses exclude MANIQA, which has one checkpoint per dataset. This leaves 27 representations and 1,512 ordered transfer pairs (Appendix C).

5 Results Analysis

5.1 Visual embeddings exhibit ordering along the rank axis

Figure 2 reports in-dataset rankability at each depth; per-dataset values are given in Tables 3 and 4 (Appendix A.2). On most datasets, a single linear direction in a frozen representation already orders

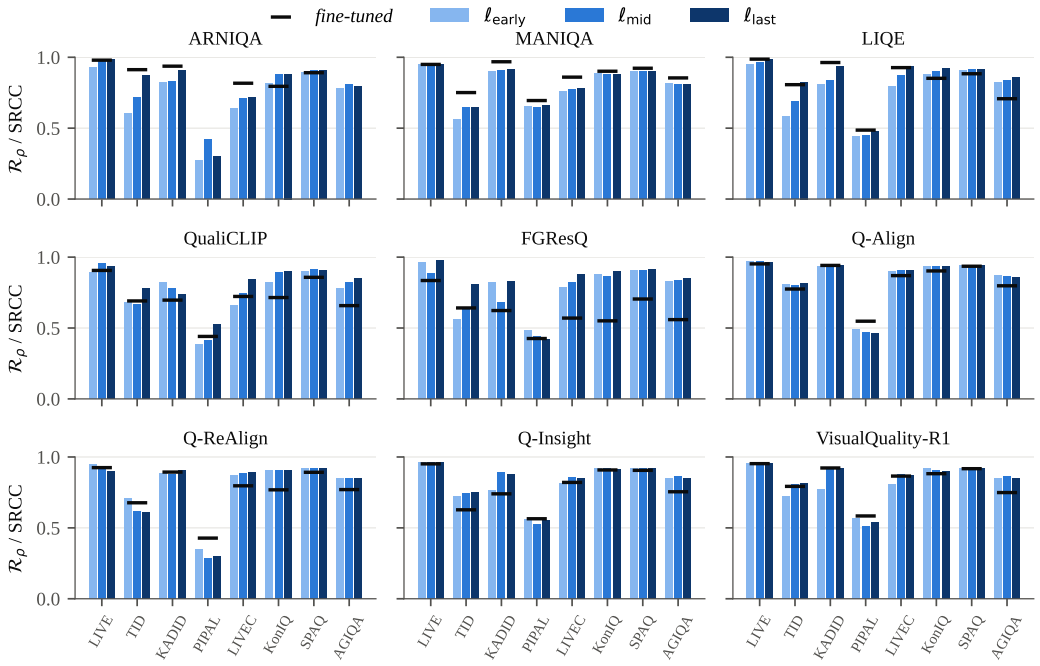

Figure 2: **Rankability by depth.** Bars show in-dataset $\mathcal{R}_\rho$ of the rank-axis at ℓ_{early}, ℓ_{mid} and ℓ_{last}. The black tick is the SRCC of the encoder's fine-tuned scoring head. For Q-Insight and VisualQuality-R1, ℓ_{early} is vision block 15 and ℓ_{mid} is LLM layer 21.

test images about as well as a trained IQA model. Averaged over the nine encoders, $\mathcal{R}_\rho$ at ℓ_{last} ranges from 0.841 (AGIQA-3K) to 0.956 (LIVE) on six datasets and drops to 0.769 on TID2013 and 0.470 on PIPAL. PIPAL is the hardest dataset for every encoder; its best value (0.660) comes from MANIQA, whose checkpoint was trained on PIPAL.

The rank-axis often beats the model's own head. Across all extracted layers, the rank-axis outperforms the fine-tuned head in 138 of 234 encoder-layer-dataset configurations (39 of 70 at ℓ_{early} and 48 of 70 at ℓ_{last}), and for 52 of the 70 encoder-dataset pairs at least one layer does. It trails only the heads of MANIQA and ARNIQA, which are trained on the target dataset. The OOD gap in these models therefore appears to arise largely from the *head* rather than the *representation* (per-layer and per-encoder comparison in Appendix A.3).

Depth. For the CNN and CLIP encoders, rankability mostly increases with depth; LIQE's ℓ_{last} is the best of its layers on all 8 datasets (mean 0.774 at ℓ_{early}, 0.857 at ℓ_{last}). For the MLLMs, the language model is nearly flat with depth, and the vision tower that feeds it is less rankable than the language-model layers (Appendix A.2). Averaged over the four MLLMs, ℓ_{early} comes within 0.01 of ℓ_{last} (0.822 vs. 0.830) and exceeds the fine-tuned heads (0.807); for Q-Insight and VisualQuality-R1 this layer lies inside the vision tower.

5.2 Cross-dataset transfer of the rank-axis

Table 1 evaluates every rank-axis on every other dataset (heatmap in Figure 4 and per-encoder results in Table 6, Appendix A.4). On average an axis keeps 80% of its in-dataset rankability on a new dataset (mean off-diagonal 0.667 vs. diagonal 0.825).

Rank-axis transfer follows content type. The three authentic datasets transfer well among themselves ($\mathcal{R}_\rho$ = 0.80-0.89). The synthetic datasets transfer well into LIVE (up to 0.904) but less well into TID2013 (at most 0.776). AGIQA-3K transfers well into LIVE (0.865) but poorly into TID2013 (0.564).

Table 1: **Cross-dataset transfer of the rank-axis.** Entry (a, b) is $\mathcal{R}_\rho$ of the axis fit on train dataset a (row) and evaluated on the test split of b (column), at ℓ_{last}, averaged over the eight encoders with a shared feature space (all except MANIQA). The diagonal (grey) is in-dataset. Bold marks the best source axis for each target. *Retention* is the mean off-diagonal value divided by the in-dataset value. Full per-encoder results are given in Appendix A.4.

Train ↓ / Test →	LIVE	TID	KADID	PIPAL	LIVEC	KonIQ	SPAQ	AGIQA	Mean
LIVE	0.958	**0.776**	**0.820**	0.426	0.694	0.692	0.833	0.657	0.700
TID	**0.904**	0.785	0.749	0.338	0.573	0.566	0.730	0.603	0.638
KADID	0.894	0.764	0.882	0.431	0.684	0.680	0.809	0.657	0.703
PIPAL	0.645	0.504	0.420	0.447	0.446	0.316	0.624	0.451	0.486
LIVEC	0.854	0.658	0.768	0.495	0.862	0.803	0.885	**0.705**	0.738
KonIQ	0.799	0.606	0.742	**0.516**	0.795	0.907	**0.886**	0.693	0.720
SPAQ	0.811	0.618	0.738	0.475	**0.820**	**0.835**	0.918	0.692	0.712
AGIQA	0.865	0.564	0.639	0.396	0.628	0.617	0.758	0.845	0.638
Mean (off-diag.)	0.825	0.641	0.696	0.440	0.663	0.644	0.789	0.637	0.667
Retention (%)	86.11	81.73	78.96	98.46	76.92	71.01	85.98	75.35	n/a

Rank-axis transfer is asymmetric. TID2013→LIVE reaches 0.904, but LIVE→TID2013 only 0.776. LIVE is the easiest target (mean 0.825); PIPAL is the hardest target (0.440) and the worst source (row mean 0.486). Even on PIPAL, the KonIQ-10k axis (0.516) beats PIPAL's own axis (0.447), which is therefore poorly estimated despite 23k training images.

Two MLLMs transfer best in relative terms. Q-ReAlign keeps 90% of its in-dataset rankability and Q-Insight 86%, whereas ARNIQA keeps only 61% (Table 6, Appendix A.4).

5.3 Order without geometry: spacing and local structure

Table 2: **Rankability, spacing and embedding geometry at ℓ_{last}** (in-dataset, averaged over the eight datasets). $\mathcal{R}_\rho/\mathcal{R}_r$: Spearman/Pearson rankability. RAIS: Rank-Axis Isometry Score. A-QEA: Anchor-QEA. OR: ordering rate. G-QEA: Global-QEA. The last three columns are means over the 56 ordered cross-dataset pairs of the raw angle θ between rank-axes, the covariance-weighted agreement c_Σ and the transfer $\mathcal{R}_r$. MANIQA has no shared feature space (one checkpoint per dataset); these columns are therefore n/a.

	In-dataset metrics						Cross-dataset axis agreement		
Encoder	$\mathcal{R}_\rho$	$\mathcal{R}_r$	RAIS	A-QEA	OR	G-QEA	$\bar{\theta}$ (°)	$\bar{c}_\Sigma$	$\bar{\mathcal{R}}_r^{\times}$
ARNIQA	0.796	0.812	0.641	0.255	0.590	0.306	86.7	0.555	0.510
MANIQA	0.816	0.836	0.626	0.536	0.697	0.553	n/a	n/a	n/a
LIQE	0.857	0.867	0.702	0.418	0.649	0.436	80.3	0.805	0.748
QualiCLIP	0.811	0.814	0.595	0.182	0.564	0.145	83.9	0.693	0.643
FGResQ	0.820	0.834	0.641	0.416	0.648	0.425	84.0	0.705	0.667
Q-Align	0.852	0.864	0.700	0.621	0.731	0.644	89.1	0.747	0.715
Q-ReAlign	0.785	0.803	0.607	0.514	0.688	0.543	86.3	0.753	0.720
Q-Insight	0.835	0.854	0.662	0.416	0.648	0.440	87.3	0.811	0.749
VisualQuality-R1	0.847	0.865	0.686	0.624	0.735	0.651	88.5	0.782	0.740

Rankability measures order along one direction. Table 2 asks whether the geometry around that direction is also organised by quality. Per-dataset values are given in Appendix A.5, with the embedding geometry at ℓ_{last} in Table 16.

Spacing follows order. RAIS (0.60-0.70 at ℓ_{last}) is lower than $\mathcal{R}_\rho$ but almost perfectly rank-correlated with it across all representations and datasets (Spearman 0.976). A representation that orders images well also spaces them roughly in proportion to their quality gaps.

The full space is only weakly quality-ordered. Over all 240 in-dataset results, Anchor-QEA averages 0.350 and the triplet ordering rate 0.627 (chance: 0.5), and 97 of the 166 results with $\mathcal{R}_\rho \geq 0.80$ have Anchor-QEA below 0.5. Quality therefore occupies a direction, not the dominant geometry. The clearest case is ℓ_{early} of Q-Insight and VisualQuality-R1, with $\mathcal{R}_\rho = 0.81$ but Anchor-QEA 0.17 and an ordering rate of 0.56. Local structure also varies more across encoders than rankability does: at ℓ_{last}, Anchor-QEA ranges from 0.182 (QualiCLIP) and 0.255 (ARNIQA) to 0.621 (Q-Align) and 0.624 (VisualQuality-R1), although the rankability of QualiCLIP and ARNIQA is similar (0.811 and 0.796; Appendix A.5).

5.4 When do rank-axes transfer? Covariance-weighted agreement, not angle

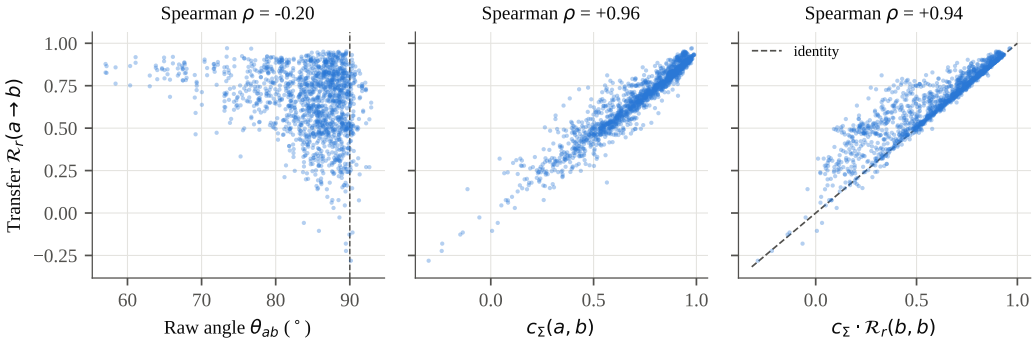

Figure 3: **Raw angle vs. covariance-weighted agreement.** Each point is one of 1,512 (representation, train a, test b) pairs. *Left:* the raw angle between rank-axes barely predicts transfer, and pairs beyond 90° (dashed) still transfer positively. *Middle:* c_Σ (Eq. (8)) predicts transfer almost perfectly. *Right:* the product model $c_\Sigma\, \mathcal{R}_r(b, b)$ lies on the identity line when the target's own axis is strong. Points above the line are pairs where a foreign axis beats the target's own axis (mostly PIPAL and TID2013).

Rank-axes fit on different datasets appear nearly orthogonal: across 1,512 pairs, the median angle is 86.0° (IQR 82.4°–88.4°), including 98 *obtuse* pairs. Yet these axes transfer well. We explain this through the covariance-weighted agreement c_Σ (Figure 3; Table 13).

Raw angle is a poor predictor; c_Σ is nearly sufficient. Across representations, the raw angle has Spearman -0.20 with transfer, versus $+0.96$ for c_Σ; within representations, the medians are -0.56 and $+0.93$, respectively, with c_Σ the stronger predictor for all 27 representations. Thus, axis disagreement occurs mainly in low-variance directions.

Obtuse is not opposite. 96 of 98 obtuse pairs transfer positively, while the only two negative cases also have negative c_Σ.

A product model explains transfer. Eq. (10) achieves $R^2 = 0.848$ and MAE $= 0.067$ (median $R^2 = 0.86$ across representations). Deviations are concentrated on PIPAL and TID2013, where the target's own axis is weaker; 126 of 136 cases where a foreign axis beats the target's axis involve these datasets.

Agreement tracks local geometry. Across 27 representations, mean c_Σ predicts mean transfer (Pearson 0.988) and correlates with Anchor-QEA (Spearman 0.76). Raw angle is less informative: for example, Q-Align at ℓ_{last} has $\bar{\theta} = 89.1°$ yet transfers at 0.715. Further observations and evaluation caveats are given in Appendices A.1 and C.

6 Conclusion

A learned rank-axis over frozen hidden representations is sufficient to encode perceptual quality. benchmarks, a ridge rank-axis reaches $\mathcal{R}_\rho \geq 0.84$ on six datasets and beats the model's fine-tuned head in 52 of 70 encoder-dataset pairs. However, the quality-score-based ordering is weakly reflected in the original embedding space neighborhoods and distances (mean Anchor-QEA 0.35). The rank-axes transfer across datasets well predicted by covariance-weighted agreement c_Σ. A rank-axis can be derived from early network layers: it beats the fine-tuned head in 39 of 70 pairs at ℓ_{early} and comes within 0.01 of the last layer for MLLMs, suggesting efficient quality prediction without full forward passes.

AI use statement

We used generative AI tools (an LLM-based coding assistant) to edit and restructure the manuscript text, to check the consistency of numbers and cross-references across sections, and to write and

debug the scripts that generate tables and figures from our experiment outputs. We did not use generative AI tools to generate research ideas, to design the experiments or to produce experimental results. We have reviewed all AI-assisted work: every statistic quoted in the text was checked against the experiment outputs, and all AI-assisted code was inspected and its outputs verified. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Lorenzo Agnolucci, Leonardo Galteri, and Marco Bertini. Quality-aware image-text alignment for opinion-unaware image quality assessment. *arXiv preprint arXiv:2403.11176*, 2024a.

Lorenzo Agnolucci, Leonardo Galteri, Marco Bertini, and Alberto Del Bimbo. Arniqa: Learning distortion manifold for image quality assessment. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 188–197, 2024b.

Sebastian Bosse, Dominique Maniry, Klaus-Robert Müller, Thomas Wiegand, and Wojciech Samek. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Transactions on Image Processing*, 27(1):206–219, 2018.

Du Chen, Tianhe Wu, Kede Ma, and Lei Zhang. Toward generalized image quality assessment: Relaxing the perfect reference quality assumption. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12742–12752, 2025.

Zewen Chen, Juan Wang, Wen Wang, Sunhan Xu, Hang Xiong, Yun Zeng, Jian Guo, Shuxun Wang, Chunfeng Yuan, Bing Li, et al. SEAGULL: No-reference image quality assessment for regions of interest via vision-language instruction tuning. *arXiv preprint arXiv:2411.10161*, 2024.

Zheng Chen, Xun Zhang, Wenbo Li, Renjing Pei, Fenglong Song, Xiongkuo Min, Xiaohong Liu, Xin Yuan, Yong Guo, and Yulun Zhang. Grounding-iqa: Grounding multimodal language model for image quality assessment. In *International Conference on Learning Representations*, volume 2026, pp. 39839–39851, 2026.

Nandhini Chockalingam and Brindha Murugan. A multimodal dense convolution network for blind image quality assessment. *Frontiers of Information Technology & Electronic Engineering*, 24 (11):1601–1615, 2023.

Yuming Fang, Hanwei Zhu, Yan Zeng, Kede Ma, and Zhou Wang. Perceptual quality assessment of smartphone photography. In *Conference on Computer Vision and Pattern Recognition*, pp. 3674–3683, 2020.

Deepti Ghadiyaram and Alan C. Bovik. Massive online crowdsourced study of subjective and objective picture quality. *IEEE Transactions on Image Processing*, 25(1):372–387, 2016.

Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

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.

Zhigang Hu, Gege Yang, Zhe Du, Xiaodong Huang, Pujing Zhang, and Dechun Liu. No-reference image quality assessment based on global awareness. *Plos one*, 19(10):e0310206, 2024.

Gu Jinjin, Cai Haoming, Chen Haoyu, Ye Xiaoxing, Jimmy S. Ren, and Dong Chao. PIPAL: A large-scale image quality assessment dataset for perceptual image restoration. In *European Conference on Computer Vision*, pp. 633–651, 2020.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: Multi-scale image quality transformer. In *IEEE/CVF International Conference on Computer Vision*, pp. 5128–5137, 2021.

Jongyoo Kim and Sanghoon Lee. Deep learning of human visual sensitivity in image quality assessment framework. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1969–1977, 2017.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:249, 2008.

Daekyu Kwon, Dongyoung Kim, Sehwan Ki, Younghyun Jo, Hyong-Euk Lee, and Seon Joo Kim. ATTIQA: Generalizable image quality feature extractor using attribute-aware pretraining. In *Asian Conference on Computer Vision*, 2024.

Chunyi Li, Zicheng Zhang, Haoning Wu, Wei Sun, Xiongkuo Min, Xiaohong Liu, Guangtao Zhai, and Weisi Lin. AGIQA-3K: An open database for ai-generated image quality assessment. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(8):6833–6846, 2024.

Wanhua Li, Xiaoke Huang, Zheng Zhu, Yansong Tang, Xiu Li, Jie Zhou, and Jiwen Lu. Ordinalclip: Learning rank prompts for language-guided ordinal regression. In *Advances in Neural Information Processing Systems*, volume 35, pp. 35313–35325, 2022.

Weiqi Li, Xuanyu Zhang, Shijie Zhao, Yabin ZHANG, Junlin Li, Li zhang, and Jian Zhang. Q-insight: Understanding image quality via visual reinforcement learning. In *Advances in Neural Information Processing Systems*, pp. 36802–36827, 2025.

Zhicheng Liao, Dongxu Wu, Zhenshan Shi, Sijie Mai, Hanwei Zhu, Lingyu Zhu, Yuncheng Jiang, and Baoliang Chen. Beyond cosine similarity: Magnitude-aware CLIP for no-reference image quality assessment. In *AAAI Conference on Artificial Intelligence*, volume 40, pp. 6934–6942, 2026.

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

Xialei Liu, Joost Van De Weijer, and Andrew D. Bagdanov. RankIQA: Learning from rankings for no-reference image quality assessment. In *IEEE International Conference on Computer Vision*, pp. 1040–1049, 2017.

Nathan Mantel. The detection of disease clustering and a generalized regression approach. *Cancer research*, 27(2_Part_1):209–220, 1967.

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.

Takamichi Miyata. Interpretable image quality assessment via CLIP with multiple antonym-prompt pairs. In *International Conference on Consumer Electronics - Berlin*, pp. 39–40, 2023.

Naila Murray, Luca Marchesotti, and Florent Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *Conference on Computer Vision and Pattern Recognition*, pp. 2408–2415, 2012.

Zhangkai Ni, Yue Liu, Keyan Ding, Wenhan Yang, Hanli Wang, and Shiqi Wang. Opinion-unaware blind image quality assessment using multi-scale deep feature statistics. *IEEE Transactions on Multimedia*, 26:10211–10224, 2024.

Devi Parikh and Kristen Grauman. Relative attributes. In *2011 International Conference on Computer Vision*, pp. 503–510, 2011. doi: 10.1109/ICCV.2011.6126281.

Nikolay Ponomarenko, Oleg Ieremeiev, Vladimir Lukin, Karen Egiazarian, Lina Jin, Jaakko Astola, Benoit Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, and C.-C. Jay Kuo. Color image database TID2013: Peculiarities and preliminary results. In *European Workshop on Visual Information Processing*, pp. 106–111, 2013.

Vaishnav Ramesh, Haining Wang, and Md Jahidul Islam. HiRQA: Hierarchical ranking and quality alignment for opinion-unaware image quality assessment. *Machine Vision and Applications*, 37 (5):110, 2026.

Michele A. Saad, Alan C. Bovik, and Christophe Charrier. Blind image quality assessment: A natural scene statistics approach in the DCT domain. *IEEE Transactions on Image Processing*, 21(8):3339–3352, 2012.

H.R. Sheikh, M.F. Sabir, and A.C. Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on Image Processing*, 15(11):3440–3451, 2006.

Xiangfei Sheng, Xiaofeng Pan, Zhichao Yang, Pengfei Chen, and Leida Li. Fine-grained image quality assessment for perceptual image restoration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 8914–8922, 2026.

Jinsong Shi, Pan Gao, Xiaojiang Peng, and Jie Qin. Dsmix: Distortion-induced sensitivity map based pre-training for no-reference image quality assessment. In *European Conference on Computer Vision*, pp. 1–17, 2024.

Ankit Shukla, Avinash Upadhyay, Swati Bhugra, and Manoj Sharma. Opinion unaware image quality assessment via adversarial convolutional variational autoencoder. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2142–2152, 2024.

Tianshu Song, Leida Li, Pengfei Chen, Hantao Liu, and Jiansheng Qian. Blind image quality assessment for authentic distortions by intermediary enhancement and iterative training. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(11):7592–7604, 2022.

Ankit Sonthalia, Arnas Uselis, and Seong Joon Oh. On the rankability of visual embeddings. In *Advances in Neural Information Processing Systems*, pp. 66169–66203, 2025.

Rajiv Soundararajan and Alan C. Bovik. RRED indices: Reduced reference entropic differencing for image quality assessment. *IEEE Transactions on Image Processing*, 21(2):517–526, 2012.

Hossein Talebi and Peyman Milanfar. NIMA: Neural image assessment. *IEEE Transactions on Image Processing*, 27(8):3998–4011, 2018.

Zhenchen Tang, Zichuan Wang, Bo Peng, and Jing Dong. Clip-agiqa: Boosting the performance of ai-generated image quality assessment with clip. In *International Conference on Pattern Recognition*, pp. 48–61, 2025.

Q-Future Team. Q-realign: Lightweight human-aligned multimodal judges built on modern vision-language models. https://github.com/Q-Future/Q-ReAlign, 2026.

Domonkos Varga. No-reference image quality assessment with global statistical features. *Journal of Imaging*, 7(2), 2021.

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

Rui Wang, Peipei Li, Huaibo Huang, Chunshui Cao, Ran He, and Zhaofeng He. Learning-to-rank meets language: Boosting language-driven ordering alignment for ordinal classification. In *Advances in Neural Information Processing Systems*, volume 36, pp. 76908–76922, 2023b.

Z. Wang, E.P. Simoncelli, and A.C. Bovik. Multiscale structural similarity for image quality assessment. In *Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pp. 1398–1402 Vol.2, 2003.

Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtao Zhai, and Weisi Lin. Q-align: Teaching LMMs for visual scoring via discrete text-defined levels. In *International Conference on Machine Learning*, 2024.

Tianhe Wu, Jian Zou, Jie Liang, Lei Zhang, and Kede Ma. Visualquality-r1: Reasoning-induced image quality assessment via reinforcement learning to rank. In *Advances in Neural Information Processing Systems*, volume 38, pp. 88167–88190, 2025. doi: 10.52202/085713-2947.

Jie Yang, Mengjin Lyu, Zhiquan Qi, and Yong Shi. Deep learning based image quality assessment: A survey. *Procedia Computer Science*, 221:1000–1005, 2023.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1190–1199, 2022.

Jiquan Yuan, Jihe Li, Fanyi Yang, Xinyan Cao, Jinming Che, Jinlong Lin, and Xixin Cao. PKU-AIGIQA-4K: A perceptual quality assessment database for both text-to-image and image-to-image ai-generated images. In *IEEE/CVF International Conference on Computer Vision Workshops*, pp. 3362–3371, 2025.

Kaiwen Zha, Peng Cao, Jeany Son, Yuzhe Yang, and Dina Katabi. Rank-n-contrast: Learning continuous representations for regression. In *Advances in Neural Information Processing Systems*, volume 36, pp. 17882–17903, 2023.

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.

Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14071–14081, 2023.

Hancheng Zhu, Leida Li, Jinjian Wu, Weisheng Dong, and Guangming Shi. MetaIQA: Deep meta-learning for no-reference image quality assessment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14131–14140, 2020.

A Extended results

This appendix gives the detailed results behind Section 5: a summary of all observations (Appendix A.1), per-dataset in-dataset rankability (Appendix A.2), the comparison with fine-tuned heads (Appendix A.3), the full cross-dataset transfer results and their predictors (Appendix A.4), per-dataset spacing and embedding geometry (Appendix A.5) and visualisations of the test embeddings around the rank-axis (Appendix A.6).

A.1 Summary of observations

The following list summarises the observations of Section 5 and the appendices.

1. **Visual embeddings exhibit ordering along the rank axis, but the embedding space is only weakly ordered by quality.** Averaged over nine encoders, one ridge direction at ℓ_{last} reaches $\mathcal{R}_\rho \geq 0.84$ on six of the eight datasets. TID2013 (0.77) and PIPAL (0.47) are the exceptions. The order is largely induced by the projection onto the rank-axis: the embedding space itself is only weakly ordered by quality (mean Anchor-QEA 0.35, ordering rate 0.63 against a chance level of 0.5), and 97 of the 166 results with $\mathcal{R}_\rho \geq 0.80$ have Anchor-QEA below 0.5.
2. **OOD failures come from the head, not the representation.** The rank-axis beats the fine-tuned head in 138 of 234 encoder-layer-dataset configurations (48 of 70 at ℓ_{last}), with the largest mean gain for FGResQ ($+0.21$ at ℓ_{last}). It trails only heads that were trained on the target dataset (MANIQA, ARNIQA).
3. **The rank-axis matches a trained linear head.** On identical ARNIQA features it matches the trained linear head in-dataset (0.878 vs. 0.889) and cross-dataset (0.523 vs. 0.534) (Appendix A.3).

4. **Depth helps CNN/CLIP encoders but not MLLMs.** LIQE is best at ℓ_{last} on every dataset. MLLM language layers are flat in depth, and their vision towers are less rankable than the LLM layers.
5. **Axes transfer well, and transfer follows content type.** An axis keeps 80% of its rankability on a new dataset on average. Authentic datasets form a block that transfers well internally. Transfer is asymmetric, LIVE is the easiest target, and PIPAL is the hardest target and worst source.
6. **Two MLLMs keep the most rankability under transfer** (Q-ReAlign 90%, Q-Insight 86%). ARNIQA keeps the least (61%).
7. **Spacing along the rank-axis follows order.** RAIS is almost perfectly rank-correlated with $\mathcal{R}_\rho$ (Spearman 0.98), although the embedding space around the axis is not ordered by quality (item 1).
8. **Raw angles mislead.** Rank-axes are nearly orthogonal (median $86°$) and the angle barely predicts transfer (Spearman -0.20). Obtuse pairs still transfer positively (96 of 98).
9. **Covariance-weighted agreement explains transfer.** c_Σ has Spearman $+0.96$ with transfer, and $\mathcal{R}_r(a \to b) \approx c_\Sigma \mathcal{R}_r(b, b)$ holds with $R^2 = 0.85$. Deviations come mostly from targets whose own axis is weak (PIPAL, TID2013).
10. **Local structure predicts agreement.** Representations with higher Anchor-QEA have rank-axes that agree more across datasets ($\rho = 0.76$) and therefore transfer better.

A.2 Per-dataset in-dataset rankability

Tables 3 and 4 give the in-dataset rankability of every extracted layer on each dataset, together with the SRCC of the encoder's fine-tuned head, as summarised in Figure 2 (Section 5.1).

Per-dataset averages. Averaged over the nine encoders, $\mathcal{R}_\rho$ at ℓ_{last} is 0.956 on LIVE, 0.916 on SPAQ, 0.904 on KonIQ-10k, 0.885 on KADID-10k, 0.852 on LIVEC, 0.841 on AGIQA-3K, 0.769 on TID2013 and 0.470 on PIPAL.

Depth. FGResQ's ℓ_{last} is the best of its layers on 7 of the 8 datasets. For the MLLMs, the language model is nearly flat with depth (Q-Align: $0.856/0.852/0.852$), and Q-ReAlign is best at its shallowest LLM layer (0.805 vs. 0.785). The vision tower that feeds the LLM is consistently less rankable than the LLM layers (Q-Align 0.823 vs. 0.852, Q-Insight 0.796 vs. 0.835, VisualQuality-R1 0.771 vs. 0.847). The language model therefore adds quality-relevant structure on top of its visual input.

A.3 Rank-axis versus fine-tuned heads

All encoders. Across all extracted layers, the rank-axis outperforms the fine-tuned head in 138 of 234 encoder-layer-dataset configurations: 39 of 70 at ℓ_{early}, 42 of 70 at ℓ_{mid}, 48 of 70 at ℓ_{last} and 9 of 24 at ℓ_{vis}. For 52 of the 70 encoder-dataset pairs, at least one layer outperforms the head. At ℓ_{last}, the margin is largest for encoders whose fine-tuned head is out-of-distribution for most benchmarks: FGResQ (mean gain of $+0.206$ in $\mathcal{R}_\rho$ over the head's SRCC; trained only on restoration data), QualiCLIP ($+0.100$) and Q-Insight ($+0.051$). For the two encoders whose fine-tuned heads are trained directly on the target data, the rank-axis performs slightly below the head: MANIQA (-0.046) and ARNIQA (-0.010). The quality ordering is therefore already present in the features and is recovered by a single direction.

ARNIQA's linear regressor. ARNIQA offers a controlled comparison: its fine-tuned per-dataset regressor is a linear map on exactly our 4096-d ℓ_{last} feature. Table 5 shows that the ridge rank-axis matches it in-dataset (0.878 vs. 0.889) and cross-dataset (0.523 vs. 0.534). The axis is better cross-dataset for five of the six training sets. KonIQ-10k is the exception: its head transfers much better (0.678 vs. 0.488).

A.4 Full cross-dataset transfer results

Table 1 in the main text and Table 6 below summarise transfer at ℓ_{last} by averaging over encoders and over source datasets, respectively. This appendix gives the full 8×8 transfer results $\mathcal{R}_\rho(a \to b)$ for

Table 3: **In-dataset rankability (a): classic deep neural networks and CLIP-based encoders.** $\mathcal{R}_\rho$: in-dataset rankability (Spearman). *fine-tuned* is the SRCC of the encoder's own fine-tuned scoring head: the per-dataset checkpoint for ARNIQA and MANIQA (n/a where none exists), otherwise the single released checkpoint. Bold marks the best value in each column within an encoder. * marks a dataset in the encoder's training data. MANIQA uses a separate checkpoint, and hence feature space, per dataset; as every dataset is in its training data, its values are unmarked.

	LIVE	TID2013	KADID-10k	PIPAL	LIVEC	KonIQ-10k	SPAQ	AGIQA-3K
ARNIQA								
$\mathcal{R}_\rho(\ell_{\text{early}})$	0.930	0.606	0.820	0.272	0.639	0.819	0.895	0.777
$\mathcal{R}_\rho(\ell_{\text{mid}})$	0.977	0.716	0.827	**0.421**	0.708	0.878	**0.906**	**0.807**
$\mathcal{R}_\rho(\ell_{\text{last}})$	**0.981**	0.874	0.909	0.302	0.718	**0.882**	0.906	0.797
fine-tuned	0.980	**0.913**	**0.937**	n/a	**0.817**	0.795	0.891	n/a
MANIQA								
$\mathcal{R}_\rho(\ell_{\text{early}})$	**0.951**	0.560	0.903	0.652	0.762	0.884	0.898	0.815
$\mathcal{R}_\rho(\ell_{\text{mid}})$	0.949	0.644	0.910	0.647	0.773	0.881	0.898	0.806
$\mathcal{R}_\rho(\ell_{\text{last}})$	0.946	0.647	0.912	0.660	0.779	0.881	0.897	0.808
fine-tuned	0.950	**0.751**	**0.968**	**0.695**	**0.860**	**0.901**	**0.923**	**0.854**
LIQE								
$\mathcal{R}_\rho(\ell_{\text{early}})$	0.953*	0.583	0.806*	0.442	0.795*	0.882*	0.908	0.824
$\mathcal{R}_\rho(\ell_{\text{mid}})$	0.963*	0.688	0.834*	0.452	0.869*	0.903*	0.911	0.838
$\mathcal{R}_\rho(\ell_{\text{last}})$	0.984*	**0.826**	0.933*	0.479	**0.938***	**0.921***	**0.915**	**0.860**
fine-tuned	**0.987***	0.807	**0.963***	**0.486**	0.927*	0.852*	0.884	0.708
QualiCLIP								
$\mathcal{R}_\rho(\ell_{\text{early}})$	0.896	0.678	**0.823**	0.383	0.661	0.819	0.896	0.776
$\mathcal{R}_\rho(\ell_{\text{mid}})$	**0.954**	0.663	0.781	0.414	0.743	0.891	**0.914**	0.825
$\mathcal{R}_\rho(\ell_{\text{last}})$	0.937	**0.781**	0.738	**0.525**	**0.843**	**0.903**	0.910	**0.848**
fine-tuned	0.906	0.691	0.697	0.441	0.723	0.715	0.858	0.658
FGResQ								
$\mathcal{R}_\rho(\ell_{\text{early}})$	0.962	0.563	0.824	**0.482**	0.783	0.882	0.907	0.828
$\mathcal{R}_\rho(\ell_{\text{mid}})$	0.887	0.643	0.682	0.438	0.821	0.865	0.904	0.837
$\mathcal{R}_\rho(\ell_{\text{last}})$	**0.977**	**0.805**	**0.827**	0.416	**0.876**	**0.896**	**0.915**	**0.848**
fine-tuned	0.835	0.642	0.623	0.426	0.570	0.551	0.705	0.559

Table 4: **In-dataset rankability (b): multimodal LLMs.** Rows and marks as in Table 3; ℓ_{vis} is the vision-tower output (Q-ReAlign has none).

	LIVE	TID2013	KADID-10k	PIPAL	LIVEC	KonIQ-10k	SPAQ	AGIQA-3K
Q-Align								
$\mathcal{R}_\rho(\ell_{\text{early}})$	**0.969**	0.807	0.934*	0.491	0.903	**0.935***	0.939*	**0.873**
$\mathcal{R}_\rho(\ell_{\text{mid}})$	0.967	0.799	0.938*	0.466	0.904	**0.935***	**0.941***	0.863
$\mathcal{R}_\rho(\ell_{\text{last}})$	0.964	**0.816**	0.939*	0.460	**0.906**	0.934*	**0.941***	0.860
$\mathcal{R}_\rho(\ell_{\text{vis}})$	0.922	0.664	0.868*	0.511	0.891	0.923*	0.933*	0.871
fine-tuned	0.952	0.775	**0.942***	**0.548**	0.870	0.903*	0.936*	0.798
Q-ReAlign								
$\mathcal{R}_\rho(\ell_{\text{early}})$	**0.946**	**0.710**	0.885*	0.348	0.872	**0.906***	**0.922***	0.847
$\mathcal{R}_\rho(\ell_{\text{mid}})$	0.918	0.615	0.880*	0.288	0.885	0.905*	0.921*	**0.850**
$\mathcal{R}_\rho(\ell_{\text{last}})$	0.902	0.611	**0.905***	0.299	**0.889**	0.904*	0.919*	0.848
fine-tuned	0.925	0.677	0.894*	**0.428**	0.797	0.768*	0.892*	0.770
Q-Insight *(ℓ_{early} = ViT blk 15, ℓ_{mid} = LLM layer 21)*								
$\mathcal{R}_\rho(\ell_{\text{early}})$	0.961	0.723	0.766	0.558	0.812	**0.921***	**0.920**	0.852
$\mathcal{R}_\rho(\ell_{\text{mid}})$	**0.963**	0.746	**0.894**	0.523	**0.859**	0.918*	0.918	0.862
$\mathcal{R}_\rho(\ell_{\text{last}})$	0.963	**0.751**	0.882	0.552	0.850	0.915*	0.917	0.851
$\mathcal{R}_\rho(\ell_{\text{vis}})$	0.925	0.685	0.751	0.518	0.804	0.904*	0.909	**0.874**
fine-tuned	0.951	0.627	0.740	**0.564**	0.821	0.908*	0.906	0.755
VisualQuality-R1 *(ℓ_{early} = ViT blk 15, ℓ_{mid} = LLM layer 21)*								
$\mathcal{R}_\rho(\ell_{\text{early}})$	0.956	0.722	0.773*	0.564	0.809	**0.920**	0.920	0.850
$\mathcal{R}_\rho(\ell_{\text{mid}})$	**0.958**	0.804	0.919*	0.510	**0.879**	0.908	**0.922**	0.862
$\mathcal{R}_\rho(\ell_{\text{last}})$	0.953	**0.816**	**0.923***	0.538	0.873	0.902	0.921	0.852
$\mathcal{R}_\rho(\ell_{\text{vis}})$	0.863	0.660	0.705*	0.481	0.790	0.885	0.906	**0.875**
fine-tuned	0.953	0.793	0.922*	**0.584**	0.865	0.882	0.917	0.749

Table 5: **Rank-axis vs. trained linear head.** ARNIQA's fine-tuned per-dataset linear regressor and our ridge rank-axis both act on the same frozen 4096-d ℓ_{last} feature. *In*: SRCC on the training dataset's test split. *Cross*: mean SRCC over the other seven test sets.

	In-dataset		Cross-dataset (mean of 7)	
Train set	Rank-axis	Head	Rank-axis	Head
LIVE	0.981	0.980	0.622	0.603
TID2013	0.874	0.913	0.486	0.477
KADID-10k	0.909	0.937	0.602	0.552
LIVEC	0.718	0.817	0.497	0.451
KonIQ-10k	0.882	0.795	0.488	0.678
SPAQ	0.906	0.891	0.446	0.444
Mean	0.878	0.889	0.523	0.534

each of the eight encoders with a shared feature space, at ℓ_{last} (Tables 7 and 8), ℓ_{early} (Tables 9 and 10) and ℓ_{mid} (Tables 11 and 12). At each layer, both the source and the target axes are fit and evaluated on features from that layer. For Q-Insight and VisualQuality-R1, ℓ_{early} is vision block 15 and ℓ_{mid} is LLM layer 21. MANIQA is omitted because it has a different feature space for each dataset (Section 4). Appendix A.4.4 reports the correlation of transfer with the raw angle, the covariance-weighted agreement c_Σ and the product model.

A.4.1 Last layer (ℓ_{last})

Across the eight encoders, the mean off-diagonal $\mathcal{R}_\rho$ ranges from 0.484 (ARNIQA) to 0.733 (LIQE), against in-dataset values of 0.785-0.857. Each encoder follows the pattern of Table 1. LIVE and SPAQ are the easiest targets (mean 0.825 and 0.789), and PIPAL is both the hardest target (0.440) and the worst source (0.486). A foreign axis beats the target's own axis in 41 of the 448 off-diagonal entries: 29 with PIPAL as target, 7 with TID2013 and 5 with LIVE. The mean transfer exceeds the fine-tuned head in only 9 of the 62 encoder-target results with a head, whereas the in-dataset axis beats the head in 48 of 70 pairs (Section 5.1). A single axis from another dataset therefore does not replace the head. Figure 4 shows the encoder-averaged transfer results of Table 1 as a heatmap, and Table 6 gives the mean transfer to each target for each encoder. LIQE has the highest absolute transfer (0.733), partly because four of the eight datasets are in its training data. Q-Insight is the best encoder on four of the eight targets and LIQE on three.

Table 6: **Transfer to each target dataset, per encoder.** Mean $\mathcal{R}_\rho$ on the target's test split over the seven rank-axes fit on the *other* datasets (ℓ_{last}). Bold marks the best encoder for each target. Full per-encoder results are given in Tables 7 and 8. VQ-R1: VisualQuality-R1.

Target	ARNIQA	LIQE	QualiCLIP	FGResQ	Q-Align	Q-ReAlign	Q-Insight	VQ-R1
LIVE	0.559	**0.925**	0.831	0.844	0.835	0.870	0.893	0.840
TID2013	0.426	0.684	0.642	0.669	**0.719**	0.624	0.661	0.706
KADID-10k	0.536	**0.793**	0.602	0.652	0.738	0.783	0.685	0.784
PIPAL	0.230	0.479	0.422	0.433	0.504	0.392	**0.538**	0.520
LIVEC	0.433	**0.804**	0.594	0.627	0.747	0.735	0.714	0.648
KonIQ-10k	0.460	0.690	0.587	0.636	0.662	0.703	**0.713**	0.702
SPAQ	0.706	0.825	0.745	0.758	0.783	0.839	**0.852**	0.808
AGIQA-3K	0.522	0.667	0.601	0.602	0.617	0.692	**0.706**	0.686
Mean transfer	0.484	0.733	0.628	0.653	0.701	0.705	0.720	0.712
Mean in-dataset	0.796	0.857	0.811	0.820	0.852	0.785	0.835	0.847
Retention (%)	60.79	85.59	77.48	79.59	82.20	89.81	86.23	84.02

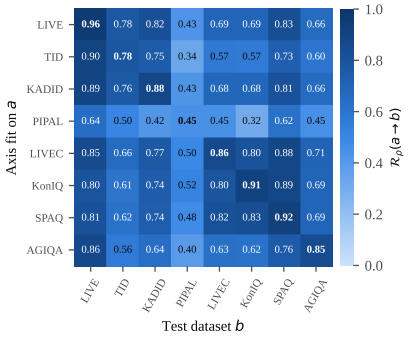

Figure 4: **Transfer results** $\mathcal{R}_\rho(a \to b)$ at ℓ_{last}, averaged over the eight shared-space encoders (values from Table 1). The diagonal is in-dataset. Authentic datasets (LIVEC, KonIQ, SPAQ) form a block that transfers well among its members. PIPAL is the weakest source and the weakest target.

Table 7: **Full cross-dataset transfer at ℓ_{last} (a): classic deep neural networks and CLIP-based encoders.** Entry (a, b) is $\mathcal{R}_\rho$ of the axis fit on the train split of a (row) and evaluated on the test split of b (column). The diagonal (grey) is in-dataset. Bold marks the best source axis for each target within an encoder. *Mean* (column): mean over the seven off-diagonal entries of a row; *Mean* (row): mean over the seven source axes for each target, as in Table 6. *fine-tuned*: SRCC of the encoder's fine-tuned scoring head on the test split of b, as in Table 3 (n/a where no head exists); its last column is the mean over the available datasets. Averaging these results over encoders gives Table 1.

Train ↓ / Test →	LIVE	TID	KADID	PIPAL	LIVEC	KonIQ	SPAQ	AGIQA	Mean
ARNIQA									
LIVE	0.981	**0.723**	**0.850**	0.319	0.449	0.598	0.801	**0.614**	0.622
TID	**0.912**	0.874	0.766	0.246	0.251	0.272	0.590	0.368	0.486
KADID	0.897	0.689	0.909	0.308	0.471	0.507	0.776	0.564	0.602
PIPAL	0.674	0.521	0.438	0.302	0.415	0.421	0.657	0.501	0.518
LIVEC	0.289	0.320	0.447	0.268	0.718	0.734	**0.859**	0.561	0.497
KonIQ	0.292	0.305	0.541	**0.424**	0.524	0.882	0.853	0.474	0.488
SPAQ	0.195	0.206	0.418	0.253	**0.713**	**0.765**	0.906	0.574	0.446
AGIQA	0.652	0.218	0.290	-0.209	0.207	-0.076	0.405	0.797	0.212
Mean	*0.559*	*0.426*	*0.536*	*0.230*	*0.433*	*0.460*	*0.706*	*0.522*	*0.484*
fine-tuned	0.980	0.913	0.937	n/a	0.817	0.795	0.891	n/a	0.889
LIQE									
LIVE	0.984	**0.808**	0.901	0.456	0.850	0.727	0.859	0.662	0.752
TID	0.921	0.826	0.697	0.435	0.649	0.505	0.758	0.628	0.656
KADID	0.975	0.788	0.933	0.476	0.906	0.771	0.855	**0.702**	0.782
PIPAL	0.904	0.520	0.715	0.479	0.641	0.482	0.751	0.639	0.665
LIVEC	**0.977**	0.744	**0.920**	0.523	0.938	**0.845**	**0.883**	0.680	0.796
KonIQ	0.898	0.654	0.820	**0.552**	**0.911**	0.921	0.864	0.684	0.769
SPAQ	0.930	0.711	0.836	0.437	0.906	0.833	0.915	0.677	0.761
AGIQA	0.872	0.564	0.660	0.473	0.767	0.668	0.803	0.860	0.687
Mean	*0.925*	*0.684*	*0.793*	*0.479*	*0.804*	*0.690*	*0.825*	*0.667*	*0.733*
fine-tuned	0.987	0.807	0.963	0.486	0.927	0.852	0.884	0.708	0.827
QualiCLIP									
LIVE	0.937	**0.738**	**0.685**	0.432	0.615	0.571	0.751	0.547	0.620
TID	0.872	0.781	0.552	0.215	0.420	0.422	0.618	0.461	0.509
KADID	0.842	0.718	0.738	0.329	0.533	0.491	0.692	0.595	0.600
PIPAL	0.841	0.582	0.474	0.525	0.484	0.478	0.751	0.587	0.600
LIVEC	**0.900**	0.633	0.650	0.485	0.843	0.752	0.851	**0.695**	0.710
KonIQ	0.714	0.603	0.667	0.492	**0.758**	0.903	**0.860**	0.683	0.682
SPAQ	0.812	0.672	0.631	0.471	0.744	**0.782**	0.910	0.642	0.679
AGIQA	0.834	0.548	0.552	**0.527**	0.604	0.613	0.696	0.848	0.625
Mean	*0.831*	*0.642*	*0.602*	*0.422*	*0.594*	*0.587*	*0.745*	*0.601*	*0.628*
fine-tuned	0.906	0.691	0.697	0.441	0.723	0.715	0.858	0.658	0.711
FGResQ									
LIVE	0.977	0.745	0.690	0.374	0.549	0.533	0.737	0.586	0.602
TID	0.882	0.805	0.647	0.333	0.551	0.534	0.666	0.597	0.602
KADID	0.786	**0.824**	0.827	0.469	0.495	0.541	0.688	0.487	0.613
PIPAL	0.704	0.548	0.528	0.416	0.512	0.553	0.689	0.568	0.586
LIVEC	0.893	0.630	**0.717**	**0.495**	0.876	0.802	**0.887**	**0.668**	0.727
KonIQ	0.836	0.598	0.654	0.479	0.820	0.896	0.880	0.662	0.704
SPAQ	0.890	0.670	0.678	0.456	**0.834**	**0.811**	0.915	0.649	0.713
AGIQA	**0.914**	0.665	0.650	0.426	0.630	0.680	0.758	0.848	0.675
Mean	*0.844*	*0.669*	*0.652*	*0.433*	*0.627*	*0.636*	*0.758*	*0.602*	*0.653*
fine-tuned	0.835	0.642	0.623	0.426	0.570	0.551	0.705	0.559	0.614

Table 8: **Full cross-dataset transfer at ℓ_{last} (b): multimodal LLMs.** Entries and marks as in Table 7.

Train ↓ / Test →	LIVE	TID	KADID	PIPAL	LIVEC	KonIQ	SPAQ	AGIQA	Mean
Q-Align									
LIVE	0.964	**0.808**	0.911	0.504	0.835	0.850	0.905	0.742	0.794
TID	0.938	0.816	0.852	0.382	0.618	0.729	0.820	0.723	0.723
KADID	0.941	0.757	0.939	0.533	**0.862**	0.889	**0.931**	0.767	0.811
PIPAL	0.214	0.522	-0.074	0.460	0.419	-0.347	0.093	-0.234	0.085
LIVEC	**0.943**	0.767	0.925	0.518	0.906	0.861	0.919	0.764	0.814
KonIQ	0.937	0.740	0.882	0.528	0.851	0.934	0.928	**0.780**	0.807
SPAQ	0.935	0.752	**0.933**	**0.563**	**0.862**	**0.905**	0.941	0.779	0.818
AGIQA	0.936	0.686	0.737	0.498	0.782	0.747	0.884	0.860	0.753
Mean	*0.835*	*0.719*	*0.738*	*0.504*	*0.747*	*0.662*	*0.783*	*0.617*	*0.701*
fine-tuned	0.952	0.775	0.942	0.548	0.870	0.903	0.936	0.798	0.840
Q-ReAlign									
LIVE	0.902	0.747	**0.877**	0.316	0.709	0.690	0.860	0.651	0.693
TID	0.848	0.611	0.809	0.089	0.650	0.545	0.699	0.657	0.614
KADID	**0.915**	**0.762**	0.905	0.311	0.765	0.728	0.872	0.732	0.727
PIPAL	0.902	0.551	0.666	0.299	0.638	0.619	0.829	0.513	0.674
LIVEC	0.912	0.669	0.842	0.520	0.889	0.745	0.881	0.762	0.762
KonIQ	0.850	0.598	0.763	**0.540**	0.822	0.904	**0.902**	**0.771**	0.749
SPAQ	0.837	0.509	0.753	0.476	**0.824**	**0.829**	0.919	0.757	0.712
AGIQA	0.829	0.532	0.770	0.493	0.736	0.761	0.830	0.848	0.707
Mean	*0.870*	*0.624*	*0.783*	*0.392*	*0.735*	*0.703*	*0.839*	*0.692*	*0.705*
fine-tuned	0.925	0.677	0.894	0.428	0.797	0.768	0.892	0.770	0.769
Q-Insight									
LIVE	0.963	**0.807**	**0.789**	0.533	0.783	0.819	0.883	**0.766**	0.768
TID	0.913	0.751	0.773	0.497	0.681	0.710	0.857	0.658	0.727
KADID	0.879	0.762	0.882	0.541	0.667	0.699	0.820	0.706	0.725
PIPAL	0.678	0.516	0.301	0.552	0.565	0.268	0.753	0.608	0.527
LIVEC	**0.961**	0.707	0.762	**0.573**	0.850	0.841	0.895	0.756	0.785
KonIQ	0.946	0.605	0.733	0.571	**0.824**	0.915	**0.899**	0.743	0.760
SPAQ	0.932	0.643	0.738	0.559	0.815	**0.873**	0.917	0.705	0.752
AGIQA	0.939	0.585	0.698	0.490	0.661	0.783	0.856	0.851	0.716
Mean	*0.893*	*0.661*	*0.685*	*0.538*	*0.714*	*0.713*	*0.852*	*0.706*	*0.720*
fine-tuned	0.951	0.627	0.740	0.564	0.821	0.908	0.906	0.755	0.784
VisualQuality-R1									
LIVE	0.953	**0.829**	0.861	0.474	0.762	0.748	0.872	0.688	0.747
TID	0.949	0.816	0.894	0.506	0.767	0.808	0.832	0.729	0.784
KADID	0.919	0.814	0.923	0.485	0.768	0.814	0.841	0.707	0.764
PIPAL	0.240	0.274	0.308	0.538	-0.112	0.056	0.467	0.425	0.237
LIVEC	**0.955**	0.790	0.879	0.578	0.873	0.847	0.904	**0.758**	0.816
KonIQ	0.923	0.746	0.875	0.541	0.851	0.902	**0.904**	0.747	0.798
SPAQ	0.955	0.779	**0.913**	**0.587**	**0.862**	**0.878**	0.921	0.751	0.818
AGIQA	0.942	0.712	0.759	0.470	0.638	0.762	0.835	0.852	0.731
Mean	*0.840*	*0.706*	*0.784*	*0.520*	*0.648*	*0.702*	*0.808*	*0.686*	*0.712*
fine-tuned	0.953	0.793	0.922	0.584	0.865	0.882	0.917	0.749	0.833

A.4.2 EARLY LAYER (ℓ_{EARLY})

Transfer is weaker at ℓ_{early}. The mean off-diagonal $\mathcal{R}_\rho$ falls from 0.667 to 0.582, and the retention from 80% to 74%. The decrease is concentrated in the single-tower encoders, which retain 57%-71%, and in the vision towers of Q-Insight and VisualQuality-R1, which retain 77%. Q-Align and Q-ReAlign, whose ℓ_{early} lies in the language model, transfer as well as at ℓ_{last} (mean 0.727 and 0.718, retention 85% and 89%). The order of the targets is unchanged: LIVE (0.755) and SPAQ (0.733) are the easiest and PIPAL (0.362) the hardest. The mean transfer does not exceed the fine-tuned head for any of the 62 encoder-target results with a head.

Table 9: **Full cross-dataset transfer at ℓ_{early} (a): classic deep neural networks and CLIP-based encoders.** Entries, marks and the *fine-tuned* row as in Table 7, with every rank-axis fit and evaluated at ℓ_{early}. *fine-tuned* does not depend on the layer and is repeated for reference.

Train ↓ / Test →	LIVE	TID	KADID	PIPAL	LIVEC	KonIQ	SPAQ	AGIQA	Mean
ARNIQA									
LIVE	0.930	0.528	0.436	0.245	0.420	0.408	0.627	0.483	0.450
TID	0.765	0.606	**0.510**	0.155	0.168	0.214	0.306	0.336	0.351
KADID	**0.816**	**0.685**	0.820	0.285	0.421	0.447	0.617	0.458	0.532
PIPAL	0.612	0.361	0.323	0.272	0.285	0.337	0.588	0.360	0.409
LIVEC	0.155	0.268	0.282	0.206	0.639	0.539	**0.809**	0.483	0.392
KonIQ	0.503	0.335	0.382	**0.386**	0.563	0.819	0.789	**0.592**	0.507
SPAQ	-0.033	0.112	0.152	0.265	**0.655**	**0.666**	0.895	0.477	0.328
AGIQA	0.770	0.468	0.417	0.105	0.207	0.387	0.562	0.777	0.416
Mean	*0.512*	*0.394*	*0.358*	*0.235*	*0.388*	*0.428*	*0.614*	*0.455*	*0.423*
fine-tuned	0.980	0.913	0.937	n/a	0.817	0.795	0.891	n/a	0.889
LIQE									
LIVE	0.953	0.647	**0.608**	0.286	0.498	0.585	0.656	0.626	0.558
TID	0.652	0.583	0.467	0.172	0.350	0.279	0.467	0.392	0.397
KADID	0.820	**0.738**	0.806	0.314	0.416	0.470	0.692	0.569	0.574
PIPAL	0.800	0.505	0.500	0.442	0.538	0.499	0.714	0.504	0.580
LIVEC	**0.829**	0.430	0.493	0.366	0.795	0.675	**0.851**	0.599	0.606
KonIQ	0.696	0.370	0.482	0.377	0.741	0.882	0.840	**0.639**	0.592
SPAQ	0.608	0.411	0.408	0.331	**0.771**	**0.756**	0.908	0.568	0.550
AGIQA	0.710	0.456	0.459	**0.452**	0.482	0.589	0.672	0.824	0.546
Mean	*0.731*	*0.508*	*0.488*	*0.328*	*0.542*	*0.551*	*0.699*	*0.557*	*0.550*
fine-tuned	0.987	0.807	0.963	0.486	0.927	0.852	0.884	0.708	0.827
QualiCLIP									
LIVE	0.896	0.530	0.320	0.219	0.350	0.322	0.554	0.407	0.386
TID	**0.857**	0.678	**0.471**	0.136	0.258	0.165	0.363	0.291	0.363
KADID	0.822	0.552	0.823	0.276	0.324	0.189	0.565	0.475	0.457
PIPAL	0.713	**0.559**	0.463	0.383	0.465	0.462	0.641	0.477	0.540
LIVEC	0.495	0.319	0.272	**0.293**	0.661	0.585	**0.821**	0.486	0.467
KonIQ	0.050	0.171	0.171	0.264	0.611	0.819	0.790	**0.584**	0.377
SPAQ	0.245	0.296	0.213	0.248	**0.671**	**0.658**	0.896	0.480	0.402
AGIQA	0.516	0.396	0.299	0.276	0.290	0.329	0.579	0.776	0.384
Mean	*0.528*	*0.403*	*0.316*	*0.245*	*0.424*	*0.387*	*0.616*	*0.457*	*0.422*
fine-tuned	0.906	0.691	0.697	0.441	0.723	0.715	0.858	0.658	0.711
FGResQ									
LIVE	0.962	0.664	**0.588**	0.395	0.479	0.497	0.708	0.601	0.562
TID	0.720	0.563	0.552	0.267	0.203	0.338	0.466	0.467	0.430
KADID	0.823	**0.794**	0.824	0.277	0.407	0.567	0.708	0.563	0.591
PIPAL	0.665	0.522	0.516	0.482	0.308	0.315	0.556	0.387	0.467
LIVEC	0.722	0.492	0.545	0.372	0.783	0.690	0.859	0.587	0.610
KonIQ	0.757	0.453	0.563	0.346	**0.769**	0.882	**0.865**	**0.643**	0.628
SPAQ	0.762	0.476	0.472	0.296	0.740	**0.762**	0.907	0.588	0.585
AGIQA	**0.829**	0.456	0.493	**0.403**	0.487	0.621	0.760	0.828	0.579
Mean	*0.754*	*0.551*	*0.533*	*0.337*	*0.485*	*0.541*	*0.703*	*0.548*	*0.556*
fine-tuned	0.835	0.642	0.623	0.426	0.570	0.551	0.705	0.559	0.614

Table 10: **Full cross-dataset transfer at ℓ_{early} (b): multimodal LLMs.** Entries, marks and the *fine-tuned* row as in Table 7, with every rank-axis fit and evaluated at ℓ_{early}. *fine-tuned* does not depend on the layer and is repeated for reference.

Train ↓ / Test →	LIVE	TID	KADID	PIPAL	LIVEC	KonIQ	SPAQ	AGIQA	Mean
Q-Align									
LIVE	0.969	**0.784**	0.863	0.454	0.736	0.820	0.870	0.656	0.740
TID	0.912	0.807	0.842	0.409	0.632	0.676	0.827	0.656	0.708
KADID	0.925	0.738	0.934	0.518	0.820	0.868	0.919	0.705	0.785
PIPAL	0.671	0.485	0.482	0.491	0.451	0.257	0.461	0.189	0.428
LIVEC	0.927	0.747	0.907	0.514	0.903	0.819	0.906	0.719	0.791
KonIQ	0.932	0.721	0.882	0.514	0.849	0.935	**0.923**	**0.769**	0.799
SPAQ	**0.935**	0.729	**0.918**	0.528	**0.856**	**0.900**	0.939	0.766	0.805
AGIQA	0.928	0.652	0.777	**0.550**	0.778	0.760	0.890	0.873	0.762
Mean	*0.890*	*0.694*	*0.810*	*0.498*	*0.731*	*0.729*	*0.828*	*0.637*	*0.727*
fine-tuned	0.952	0.775	0.942	0.548	0.870	0.903	0.936	0.798	0.840
Q-ReAlign									
LIVE	0.946	**0.735**	**0.898**	0.405	0.761	0.768	0.887	0.712	0.738
TID	0.933	0.710	0.838	0.167	0.602	0.611	0.750	0.697	0.657
KADID	**0.939**	0.689	0.885	0.305	0.710	0.688	0.848	0.723	0.700
PIPAL	0.869	0.608	0.654	0.348	0.552	0.580	0.821	0.620	0.672
LIVEC	0.906	0.684	0.879	0.520	0.872	0.783	0.887	0.744	0.772
KonIQ	0.866	0.582	0.801	**0.529**	0.820	0.906	**0.901**	**0.768**	0.752
SPAQ	0.909	0.514	0.768	0.483	**0.832**	**0.810**	0.922	0.683	0.714
AGIQA	0.907	0.590	0.827	0.503	0.715	0.759	0.842	0.847	0.735
Mean	*0.904*	*0.629*	*0.809*	*0.416*	*0.713*	*0.714*	*0.848*	*0.707*	*0.718*
fine-tuned	0.925	0.677	0.894	0.428	0.797	0.768	0.892	0.770	0.769
Q-Insight *(ℓ_{early} = ViT blk 15)*									
LIVE	0.961	0.617	0.538	**0.531**	0.603	0.624	0.777	0.636	0.618
TID	0.890	0.723	**0.638**	0.208	0.420	0.386	0.573	0.551	0.524
KADID	0.909	**0.633**	0.766	0.316	0.616	0.608	0.800	0.570	0.636
PIPAL	0.886	0.581	0.570	0.558	0.591	0.575	0.775	0.565	0.649
LIVEC	0.829	0.455	0.559	0.473	0.812	0.776	0.875	0.647	0.659
KonIQ	0.868	0.472	0.619	0.486	**0.805**	0.921	**0.890**	**0.684**	0.689
SPAQ	0.720	0.423	0.544	0.378	0.802	**0.799**	0.920	0.599	0.609
AGIQA	**0.914**	0.487	0.583	0.526	0.637	0.680	0.758	0.852	0.655
Mean	*0.859*	*0.524*	*0.579*	*0.417*	*0.639*	*0.636*	*0.778*	*0.607*	*0.630*
fine-tuned	0.951	0.627	0.740	0.564	0.821	0.908	0.906	0.755	0.784
VisualQuality-R1 *(ℓ_{early} = ViT blk 15)*									
LIVE	0.956	0.610	0.541	**0.534**	0.608	0.616	0.772	0.626	0.615
TID	0.893	0.722	**0.642**	0.223	0.428	0.395	0.575	0.553	0.530
KADID	0.910	**0.651**	0.773	0.305	0.619	0.605	0.798	0.565	0.636
PIPAL	0.878	0.579	0.567	0.564	0.584	0.567	0.773	0.551	0.643
LIVEC	0.824	0.466	0.589	0.499	0.809	0.742	0.867	0.668	0.665
KonIQ	0.869	0.477	0.622	0.484	**0.804**	0.920	**0.889**	**0.684**	0.690
SPAQ	0.718	0.430	0.553	0.380	0.802	**0.798**	0.920	0.601	0.612
AGIQA	**0.912**	0.493	0.585	0.526	0.633	0.674	0.756	0.850	0.654
Mean	*0.858*	*0.529*	*0.585*	*0.421*	*0.640*	*0.628*	*0.776*	*0.607*	*0.631*
fine-tuned	0.953	0.793	0.922	0.584	0.865	0.882	0.917	0.749	0.833

A.4.3 MIDDLE LAYER (ℓ_{MID})

Transfer at ℓ_{mid} is close to ℓ_{last}, with a mean off-diagonal $\mathcal{R}_\rho$ of 0.645 and a retention of 80%. The multimodal LLMs retain 84%-89%, and VisualQuality-R1 transfers better than at ℓ_{last} (0.732 vs. 0.712). ARNIQA also improves (0.528 vs. 0.484), while LIQE, QualiCLIP and FGResQ remain below their ℓ_{last} values. A foreign axis beats the target's own axis in 41 of the 448 off-diagonal entries, 27 with PIPAL as target and 13 with TID2013. The mean transfer exceeds the fine-tuned head in 4 of the 62 encoder-target results.

Table 11: **Full cross-dataset transfer at ℓ_{mid} (a): classic deep neural networks and CLIP-based encoders.** Entries, marks and the *fine-tuned* row as in Table 7, with every rank-axis fit and evaluated at ℓ_{mid}. *fine-tuned* does not depend on the layer and is repeated for reference.

Train ↓ / Test →	LIVE	TID	KADID	PIPAL	LIVEC	KonIQ	SPAQ	AGIQA	Mean
ARNIQA									
LIVE	0.977	0.616	**0.572**	0.363	0.581	0.600	0.795	0.557	0.583
TID	**0.922**	0.716	0.502	0.134	0.292	0.307	0.532	0.401	0.441
KADID	0.900	**0.808**	0.827	0.353	0.519	0.505	0.730	0.508	0.618
PIPAL	0.801	0.408	0.495	0.421	0.435	0.605	0.716	0.473	0.562
LIVEC	0.468	0.435	0.404	0.322	0.708	0.679	**0.852**	0.546	0.529
KonIQ	0.569	0.341	0.384	**0.450**	0.646	0.878	0.843	**0.604**	0.548
SPAQ	0.204	0.300	0.262	0.292	**0.719**	**0.744**	0.906	0.483	0.429
AGIQA	0.686	0.400	0.438	0.315	0.410	0.601	0.726	0.807	0.511
Mean	*0.650*	*0.473*	*0.437*	*0.318*	*0.515*	*0.577*	*0.742*	*0.510*	*0.528*
fine-tuned	0.980	0.913	0.937	n/a	0.817	0.795	0.891	n/a	0.889
LIQE									
LIVE	0.963	0.699	0.666	0.341	0.669	0.672	0.806	0.598	0.636
TID	0.809	0.688	0.540	0.238	0.411	0.351	0.539	0.481	0.481
KADID	0.923	**0.823**	0.834	0.313	0.657	0.578	0.741	0.605	0.663
PIPAL	0.902	0.509	0.531	0.452	0.536	0.526	0.701	0.539	0.606
LIVEC	**0.948**	0.488	0.628	**0.508**	0.869	0.768	0.872	0.652	0.695
KonIQ	0.941	0.535	**0.710**	0.472	**0.845**	0.903	**0.875**	**0.658**	0.719
SPAQ	0.898	0.481	0.615	0.437	0.839	**0.789**	0.911	0.649	0.673
AGIQA	0.924	0.564	0.556	0.461	0.600	0.619	0.704	0.838	0.633
Mean	*0.906*	*0.585*	*0.607*	*0.396*	*0.651*	*0.615*	*0.748*	*0.597*	*0.638*
fine-tuned	0.987	0.807	0.963	0.486	0.927	0.852	0.884	0.708	0.827
QualiCLIP									
LIVE	0.954	0.598	0.499	0.411	0.465	0.491	0.698	0.513	0.525
TID	**0.869**	0.663	**0.523**	0.325	0.381	0.326	0.584	0.411	0.488
KADID	0.775	**0.716**	0.781	0.250	0.472	0.498	0.678	0.563	0.565
PIPAL	0.799	0.553	0.518	0.414	0.482	0.530	0.743	0.528	0.593
LIVEC	0.423	0.389	0.329	0.350	0.743	0.713	**0.853**	0.567	0.518
KonIQ	0.697	0.406	0.496	0.425	0.717	0.891	0.850	**0.634**	0.604
SPAQ	0.528	0.440	0.392	0.327	**0.732**	**0.763**	0.914	0.580	0.537
AGIQA	0.828	0.507	0.454	**0.513**	0.499	0.606	0.762	0.825	0.595
Mean	*0.703*	*0.516*	*0.459*	*0.371*	*0.535*	*0.561*	*0.738*	*0.542*	*0.553*
fine-tuned	0.906	0.691	0.697	0.441	0.723	0.715	0.858	0.658	0.711
FGResQ									
LIVE	0.887	0.692	0.616	0.335	0.477	0.499	0.708	0.490	0.545
TID	0.728	0.643	0.479	0.200	0.414	0.289	0.637	0.451	0.457
KADID	0.772	**0.694**	0.682	0.244	0.530	0.519	0.656	0.530	0.564
PIPAL	0.810	0.528	0.468	0.438	0.499	0.404	0.718	0.631	0.580
LIVEC	0.757	0.571	0.601	0.419	0.821	0.694	0.860	0.644	0.649
KonIQ	0.800	0.593	**0.641**	**0.486**	**0.789**	0.865	**0.869**	**0.659**	0.691
SPAQ	0.776	0.507	0.501	0.396	0.781	**0.731**	0.904	0.586	0.611
AGIQA	**0.882**	0.591	0.512	0.483	0.554	0.602	0.693	0.837	0.617
Mean	*0.790*	*0.596*	*0.545*	*0.366*	*0.578*	*0.534*	*0.734*	*0.570*	*0.589*
fine-tuned	0.835	0.642	0.623	0.426	0.570	0.551	0.705	0.559	0.614

A.4.4 PREDICTORS OF TRANSFER

Table 13 gives the correlations behind Section 5.4 and Figure 3, within each representation, pooled after centring and pooled without centring.

Obtuse is not opposite. All 98 obtuse pairs lie within 2.9° of orthogonal ($\cos\theta \geq -0.051$). 96 of them have positive c_Σ (median 0.68) and positive transfer (median 0.64). The two obtuse pairs that transfer negatively (Q-Align, PIPAL→KonIQ-10k, at ℓ_{last} and ℓ_{mid}) are exactly the two with negative c_Σ.

Table 12: **Full cross-dataset transfer at ℓ_{mid} (b): multimodal LLMs.** Entries, marks and the *fine-tuned* row as in Table 7, with every rank-axis fit and evaluated at ℓ_{mid}. *fine-tuned* does not depend on the layer and is repeated for reference.

Train ↓ / Test →	LIVE	TID	KADID	PIPAL	LIVEC	KonIQ	SPAQ	AGIQA	Mean
Q-Align									
LIVE	0.967	**0.816**	0.903	0.493	0.828	0.837	0.904	0.681	0.780
TID	0.935	0.799	0.885	0.430	0.634	0.714	0.829	0.748	0.739
KADID	0.937	0.765	0.938	0.525	0.857	0.885	**0.929**	0.755	0.807
PIPAL	0.018	0.571	0.240	0.466	0.312	-0.195	0.165	0.146	0.180
LIVEC	0.933	0.752	0.918	0.531	0.904	0.853	0.918	0.747	0.807
KonIQ	0.937	0.736	0.885	0.533	0.855	0.935	0.926	0.769	0.806
SPAQ	**0.939**	0.752	**0.934**	**0.564**	**0.867**	**0.904**	0.941	**0.771**	0.819
AGIQA	0.935	0.668	0.779	0.479	0.797	0.735	0.880	0.863	0.753
Mean	*0.805*	*0.723*	*0.792*	*0.508*	*0.736*	*0.676*	*0.793*	*0.660*	*0.711*
fine-tuned	0.952	0.775	0.942	0.548	0.870	0.903	0.936	0.798	0.840
Q-ReAlign									
LIVE	0.918	**0.770**	**0.872**	0.278	0.714	0.728	0.867	0.677	0.701
TID	0.867	0.615	0.801	0.072	0.611	0.523	0.691	0.612	0.597
KADID	0.892	0.730	0.880	0.317	0.716	0.658	0.857	0.714	0.698
PIPAL	0.819	0.434	0.495	0.288	0.599	0.405	0.757	0.524	0.576
LIVEC	**0.913**	0.660	0.863	0.522	0.885	0.760	0.885	0.740	0.763
KonIQ	0.867	0.620	0.769	**0.527**	**0.823**	0.905	**0.903**	**0.772**	0.754
SPAQ	0.900	0.558	0.831	0.489	0.819	**0.823**	0.921	0.746	0.738
AGIQA	0.889	0.533	0.825	0.507	0.727	0.772	0.844	0.850	0.728
Mean	*0.878*	*0.615*	*0.779*	*0.387*	*0.715*	*0.667*	*0.829*	*0.684*	*0.694*
fine-tuned	0.925	0.677	0.894	0.428	0.797	0.768	0.892	0.770	0.769
Q-Insight *(ℓ_{mid} = LLM layer 21)*									
LIVE	0.963	**0.791**	**0.792**	0.516	0.794	0.836	0.890	**0.762**	0.769
TID	0.925	0.746	0.760	0.466	0.628	0.692	0.847	0.664	0.712
KADID	0.878	0.737	0.894	0.558	0.634	0.680	0.808	0.676	0.710
PIPAL	0.822	0.608	0.315	0.523	0.365	0.256	0.748	0.591	0.529
LIVEC	**0.956**	0.708	0.774	0.567	0.859	0.852	0.896	0.752	0.786
KonIQ	0.942	0.598	0.734	**0.571**	**0.825**	0.918	**0.901**	0.744	0.759
SPAQ	0.950	0.637	0.750	0.563	0.815	**0.888**	0.918	0.718	0.760
AGIQA	0.903	0.585	0.611	0.474	0.535	0.717	0.855	0.862	0.669
Mean	*0.911*	*0.666*	*0.677*	*0.531*	*0.657*	*0.703*	*0.849*	*0.701*	*0.712*
fine-tuned	0.951	0.627	0.740	0.564	0.821	0.908	0.906	0.755	0.784
VisualQuality-R1 *(ℓ_{mid} = LLM layer 21)*									
LIVE	0.958	**0.836**	0.846	0.532	0.809	0.815	0.878	0.732	0.778
TID	0.949	0.804	0.886	0.501	0.750	0.789	0.839	0.723	0.777
KADID	0.913	0.777	0.919	0.479	0.681	0.751	0.818	0.722	0.735
PIPAL	0.769	0.353	0.333	0.510	0.367	0.062	0.693	0.536	0.445
LIVEC	**0.955**	0.785	0.894	**0.592**	0.879	0.852	**0.907**	**0.770**	0.822
KonIQ	0.914	0.711	0.864	0.529	0.850	0.908	0.903	0.731	0.786
SPAQ	0.946	0.770	**0.897**	0.580	**0.856**	**0.869**	0.922	0.736	0.808
AGIQA	0.940	0.660	0.688	0.467	0.660	0.708	0.844	0.862	0.710
Mean	*0.912*	*0.699*	*0.773*	*0.526*	*0.710*	*0.692*	*0.840*	*0.707*	*0.732*
fine-tuned	0.953	0.793	0.922	0.584	0.865	0.882	0.917	0.749	0.833

Table 13: **What predicts cross-dataset transfer?** Correlation of the transfer rankability $\mathcal{R}_r(a \to b)$ with three predictors over 27 shared-space representations × 56 ordered dataset pairs (1,512 pairs). *Per-rep.*: Spearman correlation within each representation (median and range over 27). *Pooled*: all variables centred within each representation. *Raw*: pooled without centring. Product-model fit $\mathcal{R}_r(a \to b) \approx c_\Sigma \mathcal{R}_r(b, b)$: $R^2 = 0.848$, slope 0.77, intercept +0.20, MAE 0.067.

	Per-rep. Spearman		Pooled (centred)		Raw
Predictor	median	range	Spearman	Pearson	Spearman
Raw angle θ_{ab}	-0.563	$[-0.77, -0.27]$	-0.538	-0.465	-0.199
c_Σ (cov.-weighted agreement)	$+0.932$	$[+0.84, +0.97]$	$+0.939$	$+0.943$	$+0.955$
$c_\Sigma \cdot \mathcal{R}_r(b, b)$ (product model)	$+0.929$	$[+0.74, +0.98]$	$+0.923$	$+0.911$	$+0.942$

Figure 5 relates agreement to the geometry of each representation.

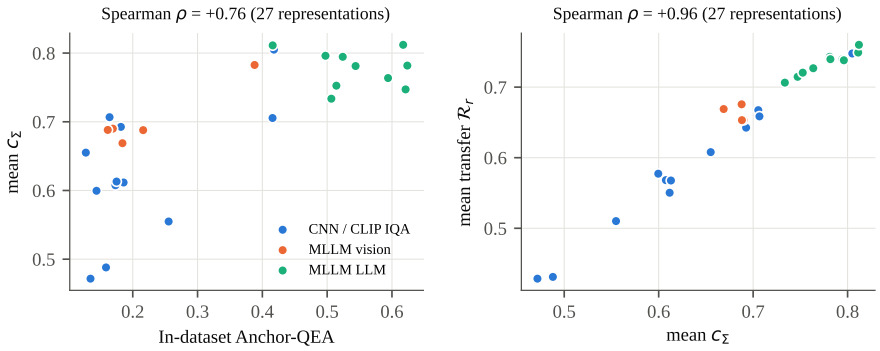

Figure 5: **Agreement is a property of the representation.** Each point is one of the 27 shared-space representations. *Left:* representations with more quality-ordered local geometry (Anchor-QEA) have rank-axes that agree more across datasets. *Right:* mean agreement c_Σ predicts mean transfer.

A.5 Per-dataset rankability, spacing and embedding geometry

Table 2 in the main text reports each metric at ℓ_{last} as a mean over the eight datasets. This appendix gives the values for each dataset at ℓ_{last} (Tables 14 and 15), ℓ_{early} (Tables 17 and 18) and ℓ_{mid} (Tables 19 and 20). At each layer, the first table covers the synthetic-distortion datasets (LIVE, TID2013, KADID-10k, PIPAL) and the second the authentic-distortion and AI-generated datasets (LIVEC, KonIQ-10k, SPAQ, AGIQA-3K). For the cross-dataset columns, the dataset is the target b, and each entry is the mean over the seven source datasets $a \neq b$, with both axes fit and evaluated at the same layer. For Q-Insight and VisualQuality-R1, ℓ_{early} is vision block 15 and ℓ_{mid} is LLM layer 21. MANIQA has no shared feature space, and its cross-dataset entries are n/a.

A.5.1 Last layer (ℓ_{last})

The metrics vary more across datasets than across encoders. Averaged over the nine encoders, LIVE is the most ordered dataset on every in-dataset metric ($\mathcal{R}_\rho = 0.956$, Anchor-QEA 0.679) and PIPAL the least (0.470 and 0.157). PIPAL is also the target with the lowest agreement ($\bar{c}_\Sigma = 0.445$) and transfer (0.429). RAIS follows $\mathcal{R}_\rho$ across all 72 results (Spearman 0.98) at about three quarters of its value. Q-Align or VisualQuality-R1 has the highest Anchor-QEA on seven of the eight datasets; on PIPAL, MANIQA, whose checkpoint is trained on PIPAL, is highest (0.314). QualiCLIP gives the clearest case of order without geometry: on LIVEC and KonIQ-10k its Global-QEA is -0.001 and -0.026, although its rankability is 0.843 and 0.903. The mean angle between rank-axes lies between $84°$ and $88°$ for every target, while $\bar{c}_\Sigma$ ranges from 0.445 to 0.853. Over the 64 encoder-target results, $\bar{c}_\Sigma$ follows transfer (Spearman 0.96), and the mean angle does not (-0.04). Table 16 gives Anchor-QEA, the ordering rate and Global-QEA at ℓ_{last} for each dataset and encoder. Anchor-QEA is lowest on PIPAL for eight of the nine encoders and highest on LIVE for seven. MANIQA's high mean Anchor-QEA (0.536) comes from checkpoints trained on each dataset.

A.5.2 Early layer (ℓ_{early})

At ℓ_{early}, rankability decreases slightly and the embedding geometry decreases strongly. Averaged over all 72 results, $\mathcal{R}_\rho$ falls from 0.824 to 0.790, Anchor-QEA from 0.442 to 0.262 and the ordering rate from 0.661 to 0.594. 50 of the 72 results have an ordering rate below 0.60, against 19 at ℓ_{last}. The largest gap is in the vision towers of Q-Insight and VisualQuality-R1, which keep $\mathcal{R}_\rho = 0.81$ with Anchor-QEA 0.17 and 0.16. This layer gives the clearest evidence that the order is induced by the projection onto the rank-axis and is not present in the geometry of the embedding space (Appendix A.1, item 1). Q-Align and Q-ReAlign keep most of their geometry (Anchor-QEA 0.544 and 0.497), and Q-Align has the highest Anchor-QEA on every dataset except PIPAL, where MANIQA is highest. Q-Align and Q-ReAlign also agree more across datasets than at ℓ_{last} ($\bar{c}_\Sigma = 0.781$ and

Table 14: **Rankability, spacing and embedding geometry at ℓ_{last} for each dataset (a): synthetic distortions.** Columns as in Table 2, without averaging over datasets. The last three columns are means over the seven cross-dataset pairs with the given dataset as target. Averaging each column over the eight datasets gives Table 2.

	In-dataset metrics						Cross-dataset axis agreement		
Encoder	$\mathcal{R}_\rho$	$\mathcal{R}_r$	RAIS	A-QEA	OR	G-QEA	$\bar{\theta}$ (°)	$\bar{c}_\Sigma$	$\bar{\mathcal{R}}_r^{\times}$
LIVE									
ARNIQA	0.981	0.982	0.936	0.414	0.654	0.462	85.6	0.572	0.559
MANIQA	0.946	0.945	0.832	0.705	0.762	0.641	n/a	n/a	n/a
LIQE	0.984	0.985	0.949	0.683	0.758	0.703	80.0	0.930	0.916
QualiCLIP	0.937	0.938	0.811	0.577	0.707	0.572	84.6	0.832	0.812
FGResQ	0.977	0.975	0.913	0.657	0.746	0.630	86.2	0.832	0.817
Q-Align	0.964	0.968	0.897	0.801	0.806	0.799	89.2	0.846	0.839
Q-ReAlign	0.902	0.903	0.734	0.764	0.789	0.766	88.5	0.890	0.850
Q-Insight	0.963	0.963	0.879	0.658	0.743	0.622	87.7	0.885	0.886
VisualQuality-R1	0.953	0.954	0.860	0.850	0.838	0.857	88.7	0.835	0.845
TID2013									
ARNIQA	0.874	0.888	0.697	0.218	0.575	0.316	87.0	0.560	0.517
MANIQA	0.647	0.700	0.373	0.232	0.577	0.272	n/a	n/a	n/a
LIQE	0.826	0.843	0.594	0.328	0.615	0.361	83.8	0.823	0.731
QualiCLIP	0.781	0.799	0.504	0.224	0.577	0.212	86.6	0.771	0.700
FGResQ	0.805	0.829	0.569	0.412	0.645	0.464	85.2	0.730	0.739
Q-Align	0.816	0.828	0.561	0.482	0.672	0.524	89.8	0.825	0.762
Q-ReAlign	0.611	0.651	0.299	0.347	0.620	0.391	87.1	0.630	0.686
Q-Insight	0.751	0.791	0.506	0.324	0.615	0.397	88.1	0.784	0.735
VisualQuality-R1	0.816	0.851	0.615	0.545	0.702	0.589	88.5	0.837	0.770
KADID-10k									
ARNIQA	0.909	0.910	0.749	0.178	0.565	0.224	87.0	0.580	0.544
MANIQA	0.912	0.911	0.771	0.736	0.791	0.715	n/a	n/a	n/a
LIQE	0.933	0.932	0.797	0.411	0.648	0.406	81.6	0.802	0.789
QualiCLIP	0.738	0.681	0.398	0.161	0.556	0.150	84.5	0.565	0.604
FGResQ	0.827	0.814	0.552	0.358	0.626	0.343	84.6	0.690	0.656
Q-Align	0.939	0.941	0.822	0.754	0.787	0.738	89.7	0.765	0.722
Q-ReAlign	0.905	0.908	0.752	0.655	0.747	0.624	86.8	0.800	0.769
Q-Insight	0.882	0.881	0.681	0.253	0.589	0.250	87.6	0.746	0.686
VisualQuality-R1	0.923	0.925	0.782	0.738	0.785	0.735	89.3	0.826	0.777
PIPAL									
ARNIQA	0.302	0.276	0.104	0.083	0.527	0.108	88.5	0.150	0.221
MANIQA	0.660	0.644	0.320	0.314	0.609	0.302	n/a	n/a	n/a
LIQE	0.479	0.467	0.140	0.098	0.533	0.078	85.2	0.546	0.465
QualiCLIP	0.525	0.507	0.185	0.050	0.517	0.030	87.6	0.491	0.410
FGResQ	0.416	0.397	0.091	0.114	0.538	0.100	86.1	0.415	0.416
Q-Align	0.460	0.442	0.124	0.205	0.571	0.209	89.9	0.453	0.493
Q-ReAlign	0.299	0.303	0.045	0.118	0.540	0.106	86.9	0.350	0.381
Q-Insight	0.552	0.544	0.218	0.175	0.557	0.177	88.6	0.647	0.530
VisualQuality-R1	0.538	0.523	0.191	0.253	0.588	0.245	89.1	0.508	0.512

Table 15: **Rankability, spacing and embedding geometry at ℓ_{last} for each dataset (b): authentic distortions and AI-generated images.** Columns as in Table 14.

	In-dataset metrics						Cross-dataset axis agreement		
Encoder	$\mathcal{R}_\rho$	$\mathcal{R}_r$	RAIS	A-QEA	OR	G-QEA	$\bar{\theta}$ (°)	$\bar{c}_\Sigma$	$\bar{\mathcal{R}}_r^\times$
LIVEC									
ARNIQA	0.718	0.762	0.508	0.114	0.539	0.138	85.7	0.607	0.463
MANIQA	0.779	0.831	0.603	0.538	0.698	0.612	n/a	n/a	n/a
LIQE	0.938	0.951	0.841	0.475	0.668	0.492	74.5	0.863	0.819
QualiCLIP	0.843	0.878	0.670	0.075	0.529	-0.001	80.7	0.709	0.630
FGResQ	0.876	0.922	0.759	0.414	0.642	0.468	81.6	0.716	0.665
Q-Align	0.906	0.929	0.780	0.654	0.741	0.698	88.7	0.829	0.778
Q-ReAlign	0.889	0.909	0.741	0.552	0.702	0.613	84.5	0.836	0.767
Q-Insight	0.850	0.900	0.716	0.397	0.639	0.443	86.3	0.866	0.780
VisualQuality-R1	0.873	0.913	0.749	0.667	0.753	0.714	88.0	0.765	0.714
KonIQ-10k									
ARNIQA	0.882	0.902	0.723	0.303	0.604	0.383	85.1	0.540	0.495
MANIQA	0.881	0.895	0.713	0.545	0.696	0.589	n/a	n/a	n/a
LIQE	0.921	0.935	0.799	0.370	0.628	0.426	76.6	0.771	0.722
QualiCLIP	0.903	0.914	0.743	0.053	0.516	-0.026	82.0	0.664	0.615
FGResQ	0.896	0.920	0.758	0.416	0.646	0.423	81.7	0.711	0.655
Q-Align	0.934	0.950	0.834	0.731	0.777	0.771	88.3	0.728	0.692
Q-ReAlign	0.904	0.927	0.778	0.503	0.680	0.581	83.9	0.804	0.746
Q-Insight	0.915	0.934	0.794	0.488	0.673	0.525	86.6	0.799	0.752
VisualQuality-R1	0.902	0.924	0.766	0.680	0.756	0.725	88.2	0.792	0.737
SPAQ									
ARNIQA	0.906	0.911	0.753	0.421	0.648	0.405	85.8	0.770	0.705
MANIQA	0.897	0.901	0.733	0.689	0.754	0.707	n/a	n/a	n/a
LIQE	0.915	0.920	0.775	0.551	0.696	0.558	77.7	0.895	0.825
QualiCLIP	0.910	0.911	0.751	0.230	0.580	0.209	80.5	0.809	0.738
FGResQ	0.915	0.919	0.770	0.586	0.711	0.600	81.2	0.820	0.753
Q-Align	0.941	0.945	0.834	0.794	0.804	0.806	88.8	0.827	0.780
Q-ReAlign	0.919	0.925	0.787	0.684	0.753	0.696	84.3	0.903	0.840
Q-Insight	0.917	0.923	0.782	0.598	0.716	0.586	86.7	0.922	0.853
VisualQuality-R1	0.921	0.927	0.789	0.751	0.784	0.769	87.9	0.876	0.813
AGIQA-3K									
ARNIQA	0.797	0.868	0.661	0.312	0.608	0.413	89.0	0.660	0.577
MANIQA	0.808	0.860	0.659	0.529	0.689	0.583	n/a	n/a	n/a
LIQE	0.860	0.899	0.723	0.428	0.649	0.466	82.8	0.810	0.713
QualiCLIP	0.848	0.886	0.694	0.083	0.529	0.014	84.9	0.700	0.631
FGResQ	0.848	0.895	0.714	0.369	0.630	0.367	85.3	0.729	0.637
Q-Align	0.860	0.911	0.747	0.548	0.693	0.606	88.7	0.705	0.650
Q-ReAlign	0.848	0.899	0.722	0.491	0.671	0.567	88.6	0.807	0.726
Q-Insight	0.851	0.900	0.723	0.434	0.651	0.524	87.1	0.841	0.769
VisualQuality-R1	0.852	0.903	0.732	0.505	0.677	0.576	88.3	0.815	0.750

Table 16: **Embedding geometry at ℓ_{last}** (test split): Anchor-QEA, ordering rate (chance 0.5) and Global-QEA for each dataset and encoder. Bold marks the best encoder for each dataset within a block. VQ-R1: VisualQuality-R1.

Dataset	ARNIQA	MANIQA	LIQE	QualiCLIP	FGResQ	Q-Align	Q-ReAlign	Q-Insight	VQ-R1
Anchor-QEA									
LIVE	0.414	0.705	0.683	0.577	0.657	0.801	0.764	0.658	**0.850**
TID2013	0.218	0.232	0.328	0.224	0.412	0.482	0.347	0.324	**0.545**
KADID-10k	0.178	0.736	0.411	0.161	0.358	**0.754**	0.655	0.253	0.738
PIPAL	0.083	**0.314**	0.098	0.050	0.114	0.205	0.118	0.175	0.253
LIVEC	0.114	0.538	0.475	0.075	0.414	0.654	0.552	0.397	**0.667**
KonIQ-10k	0.303	0.545	0.370	0.053	0.416	**0.731**	0.503	0.488	0.680
SPAQ	0.421	0.689	0.551	0.230	0.586	**0.794**	0.684	0.598	0.751
AGIQA-3K	0.312	0.529	0.428	0.083	0.369	**0.548**	0.491	0.434	0.505
Mean	0.255	0.536	0.418	0.182	0.416	0.621	0.514	0.416	0.624
Ordering rate									
LIVE	0.654	0.762	0.758	0.707	0.746	0.806	0.789	0.743	**0.838**
TID2013	0.575	0.577	0.615	0.577	0.645	0.672	0.620	0.615	**0.702**
KADID-10k	0.565	**0.791**	0.648	0.556	0.626	0.787	0.747	0.589	0.785
PIPAL	0.527	**0.609**	0.533	0.517	0.538	0.571	0.540	0.557	0.588
LIVEC	0.539	0.698	0.668	0.529	0.642	0.741	0.702	0.639	**0.753**
KonIQ-10k	0.604	0.696	0.628	0.516	0.646	**0.777**	0.680	0.673	0.756
SPAQ	0.648	0.754	0.696	0.580	0.711	**0.804**	0.753	0.716	0.784
AGIQA-3K	0.608	0.689	0.649	0.529	0.630	**0.693**	0.671	0.651	0.677
Mean	0.590	0.697	0.649	0.564	0.648	0.731	0.688	0.648	0.735
Global-QEA									
LIVE	0.462	0.641	0.703	0.572	0.630	0.799	0.766	0.622	**0.857**
TID2013	0.316	0.272	0.361	0.212	0.464	0.524	0.391	0.397	**0.589**
KADID-10k	0.224	0.715	0.406	0.150	0.343	**0.738**	0.624	0.250	0.735
PIPAL	0.108	**0.302**	0.078	0.030	0.100	0.209	0.106	0.177	0.245
LIVEC	0.138	0.612	0.492	-0.001	0.468	0.698	0.613	0.443	**0.714**
KonIQ-10k	0.383	0.589	0.426	-0.026	0.423	**0.771**	0.581	0.525	0.725
SPAQ	0.405	0.707	0.558	0.209	0.600	**0.806**	0.696	0.586	0.769
AGIQA-3K	0.413	0.583	0.466	0.014	0.367	**0.606**	0.567	0.524	0.576
Mean	0.306	0.553	0.436	0.145	0.425	0.644	0.543	0.440	0.651

0.796, against 0.747 and 0.753). Over the 64 encoder-target results, $\bar{c}_\Sigma$ still follows transfer (Spearman 0.97), and the mean angle does not (-0.06).

A.5.3 Middle layer (ℓ_{MID})

At ℓ_{mid}, every mean lies between its values at ℓ_{early} and ℓ_{last} ($\mathcal{R}_\rho$ 0.805, Anchor-QEA 0.375, ordering rate 0.636). For Q-Insight and VisualQuality-R1, ℓ_{mid} lies in the language model, and their Anchor-QEA recovers to 0.524 and 0.617, close to their ℓ_{last} values. VisualQuality-R1 has the highest Anchor-QEA on five of the eight datasets, including PIPAL (0.246). The single-tower encoders ARNIQA, LIQE, QualiCLIP and FGResQ remain weakly ordered (Anchor-QEA 0.13-0.19), although their $\mathcal{R}_\rho$ is 0.76-0.81. Over the 64 encoder-target results, $\bar{c}_\Sigma$ again follows transfer (Spearman 0.96), and the mean angle is only weakly related to it ($+0.22$).

A.6 Visualising the rank-axis

Rankability summarises the order along the rank-axis in one number and does not show how the test images are arranged around it. Figure 6 gives this view for every encoder and dataset at ℓ_{last}. For each dataset a, the test embeddings are centred and projected onto an orthonormal basis formed by the unit rank-axis $\hat{\mathbf{w}}_a$ and the top two principal components of the test embeddings after removing their component along $\hat{\mathbf{w}}_a$. Each of the three coordinates is divided by its standard deviation over the test split, and the resulting vector is normalised to unit length. The rank-axis is the north pole. Without the standardisation, the two principal components dominate and almost all points lie on the equator, because the variance along the rank-axis is small compared with that of the leading directions of the embedding space.

Where rankability is high, high-quality images lie near the north pole and low-quality images near the south pole (for example LIVE, KADID-10k, KonIQ-10k and SPAQ). On PIPAL, the least rankable dataset ($\mathcal{R}_\rho = 0.470$ on average), the colours are mixed over the whole sphere. For the multimodal LLMs, several panels form a thin curve from south to north (for example Q-Align and VisualQuality-R1 on SPAQ), while the single-tower encoders spread more widely around the axis.

Table 17: **Rankability, spacing and embedding geometry at ℓ_{early} for each dataset (a): synthetic distortions.** Columns as in Table 14, with every metric computed at ℓ_{early}. For Q-Insight and VisualQuality-R1, ℓ_{early} is vision block 15.

	In-dataset metrics						Cross-dataset axis agreement		
Encoder	$\mathcal{R}_\rho$	$\mathcal{R}_r$	RAIS	A-QEA	OR	G-QEA	$\bar{\theta}$ (°)	$\bar{c}_\Sigma$	$\bar{\mathcal{R}}_r^\times$
LIVE									
ARNIQA	0.930	0.925	0.782	0.237	0.589	0.349	84.0	0.481	0.469
MANIQA	0.951	0.947	0.839	0.573	0.707	0.585	n/a	n/a	n/a
LIQE	0.953	0.951	0.844	0.316	0.616	0.382	83.2	0.710	0.698
QualiCLIP	0.896	0.888	0.694	0.259	0.594	0.321	85.7	0.529	0.523
FGResQ	0.962	0.962	0.877	0.237	0.584	0.276	82.3	0.745	0.716
Q-Align	0.969	0.972	0.902	0.768	0.792	0.757	87.6	0.900	0.885
Q-ReAlign	0.946	0.944	0.834	0.766	0.791	0.766	85.9	0.933	0.892
Q-Insight	0.961	0.963	0.879	0.390	0.639	0.397	81.1	0.843	0.838
VisualQuality-R1	0.956	0.959	0.867	0.383	0.635	0.387	81.3	0.832	0.836
TID2013									
ARNIQA	0.606	0.648	0.299	0.149	0.551	0.263	84.6	0.423	0.462
MANIQA	0.560	0.617	0.259	0.166	0.555	0.219	n/a	n/a	n/a
LIQE	0.583	0.613	0.279	0.186	0.563	0.287	82.9	0.510	0.587
QualiCLIP	0.678	0.715	0.386	0.131	0.542	0.222	81.8	0.473	0.444
FGResQ	0.563	0.577	0.243	0.152	0.551	0.225	83.5	0.494	0.645
Q-Align	0.807	0.828	0.560	0.470	0.666	0.519	88.1	0.822	0.735
Q-ReAlign	0.710	0.717	0.415	0.353	0.624	0.395	86.1	0.734	0.690
Q-Insight	0.723	0.749	0.436	0.160	0.555	0.225	82.2	0.688	0.626
VisualQuality-R1	0.722	0.750	0.436	0.158	0.554	0.220	82.3	0.697	0.630
KADID-10k									
ARNIQA	0.820	0.809	0.532	0.071	0.526	0.098	84.4	0.395	0.365
MANIQA	0.903	0.903	0.749	0.603	0.731	0.567	n/a	n/a	n/a
LIQE	0.806	0.793	0.509	0.097	0.533	0.106	82.4	0.555	0.506
QualiCLIP	0.823	0.807	0.539	0.049	0.515	0.066	82.5	0.422	0.318
FGResQ	0.824	0.815	0.552	0.083	0.530	0.083	83.9	0.587	0.547
Q-Align	0.934	0.936	0.809	0.679	0.757	0.680	87.5	0.857	0.801
Q-ReAlign	0.885	0.880	0.707	0.634	0.738	0.649	86.0	0.814	0.802
Q-Insight	0.766	0.701	0.449	0.099	0.535	0.091	83.1	0.584	0.586
VisualQuality-R1	0.773	0.717	0.461	0.094	0.533	0.085	82.9	0.604	0.591
PIPAL									
ARNIQA	0.272	0.262	0.051	0.045	0.515	0.052	86.5	0.295	0.217
MANIQA	0.652	0.633	0.308	0.157	0.553	0.129	n/a	n/a	n/a
LIQE	0.442	0.415	0.098	0.039	0.512	0.036	87.3	0.360	0.316
QualiCLIP	0.383	0.358	0.063	0.048	0.517	0.067	86.8	0.146	0.237
FGResQ	0.482	0.452	0.142	0.043	0.514	0.042	89.4	0.330	0.319
Q-Align	0.491	0.475	0.158	0.121	0.544	0.117	89.9	0.429	0.489
Q-ReAlign	0.348	0.330	0.071	0.124	0.542	0.103	87.2	0.498	0.403
Q-Insight	0.558	0.545	0.213	0.046	0.516	0.038	86.7	0.407	0.400
VisualQuality-R1	0.564	0.550	0.221	0.045	0.516	0.036	86.6	0.406	0.405

Table 18: **Rankability, spacing and embedding geometry at ℓ_{early} for each dataset (b): authentic distortions and AI-generated images.** Columns as in Table 17. *Mean over the eight datasets*: mean of each column over all eight datasets, as in Table 2 for ℓ_{last}.

Encoder	In-dataset metrics $\mathcal{R}_\rho$	$\mathcal{R}_r$	RAIS	A-QEA	OR	G-QEA	Cross-dataset axis agreement $\bar{\theta}$ (°)	$\bar{c}_\Sigma$	$\mathcal{R}_r^\times$
LIVEC									
ARNIQA	0.639	0.677	0.389	0.136	0.548	0.132	83.3	0.522	0.390
MANIQA	0.762	0.821	0.580	0.442	0.657	0.497	n/a	n/a	n/a
LIQE	0.795	0.839	0.620	0.138	0.548	0.139	78.6	0.635	0.560
QualiCLIP	0.661	0.709	0.413	0.097	0.533	0.080	79.8	0.508	0.426
FGResQ	0.783	0.854	0.650	0.130	0.543	0.114	77.2	0.596	0.534
Q-Align	0.903	0.932	0.786	0.545	0.695	0.593	86.0	0.823	0.770
Q-ReAlign	0.872	0.904	0.726	0.520	0.690	0.579	86.2	0.828	0.764
Q-Insight	0.812	0.871	0.671	0.154	0.555	0.128	77.2	0.728	0.675
VisualQuality-R1	0.809	0.870	0.668	0.143	0.551	0.114	78.6	0.709	0.675
KonIQ-10k									
ARNIQA	0.819	0.836	0.602	0.153	0.551	0.155	82.4	0.524	0.441
MANIQA	0.884	0.894	0.712	0.293	0.599	0.273	n/a	n/a	n/a
LIQE	0.882	0.904	0.728	0.136	0.545	0.142	79.3	0.644	0.584
QualiCLIP	0.819	0.829	0.592	0.107	0.536	0.080	79.7	0.472	0.396
FGResQ	0.882	0.901	0.722	0.122	0.540	0.114	76.0	0.627	0.568
Q-Align	0.935	0.950	0.833	0.557	0.702	0.592	85.5	0.802	0.760
Q-ReAlign	0.906	0.928	0.779	0.446	0.659	0.520	84.3	0.820	0.758
Q-Insight	0.921	0.932	0.792	0.084	0.528	0.066	77.2	0.708	0.667
VisualQuality-R1	0.920	0.931	0.788	0.072	0.524	0.053	77.1	0.699	0.658
SPAQ									
ARNIQA	0.895	0.897	0.726	0.313	0.607	0.311	82.9	0.676	0.607
MANIQA	0.898	0.902	0.735	0.527	0.684	0.492	n/a	n/a	n/a
LIQE	0.908	0.912	0.755	0.275	0.595	0.266	80.4	0.757	0.690
QualiCLIP	0.896	0.898	0.727	0.265	0.592	0.251	79.5	0.667	0.600
FGResQ	0.907	0.911	0.752	0.231	0.579	0.229	78.0	0.763	0.697
Q-Align	0.939	0.943	0.829	0.732	0.773	0.739	84.8	0.876	0.826
Q-ReAlign	0.922	0.927	0.792	0.684	0.754	0.705	84.7	0.916	0.851
Q-Insight	0.920	0.921	0.777	0.255	0.584	0.232	79.7	0.836	0.770
VisualQuality-R1	0.920	0.921	0.775	0.236	0.578	0.217	79.5	0.834	0.768
AGIQA-3K									
ARNIQA	0.777	0.844	0.636	0.166	0.557	0.228	85.8	0.589	0.498
MANIQA	0.815	0.865	0.665	0.268	0.593	0.254	n/a	n/a	n/a
LIQE	0.824	0.880	0.686	0.200	0.568	0.261	83.7	0.689	0.605
QualiCLIP	0.776	0.837	0.613	0.122	0.542	0.171	84.0	0.557	0.485
FGResQ	0.828	0.883	0.687	0.154	0.553	0.206	80.8	0.654	0.593
Q-Align	0.873	0.920	0.765	0.479	0.669	0.507	86.6	0.741	0.676
Q-ReAlign	0.847	0.899	0.722	0.451	0.655	0.515	87.8	0.826	0.745
Q-Insight	0.852	0.895	0.716	0.167	0.559	0.193	82.3	0.726	0.665
VisualQuality-R1	0.850	0.895	0.715	0.160	0.556	0.184	82.3	0.724	0.663
Mean over the eight datasets									
ARNIQA	0.720	0.737	0.502	0.159	0.555	0.199	84.2	0.488	0.431
MANIQA	0.803	0.823	0.606	0.379	0.635	0.377	n/a	n/a	n/a
LIQE	0.774	0.788	0.565	0.173	0.560	0.202	82.2	0.608	0.568
QualiCLIP	0.742	0.755	0.503	0.135	0.546	0.157	82.5	0.472	0.429
FGResQ	0.779	0.794	0.578	0.144	0.549	0.161	81.4	0.600	0.577
Q-Align	0.856	0.869	0.705	0.544	0.700	0.563	87.0	0.781	0.743
Q-ReAlign	0.805	0.816	0.631	0.497	0.682	0.529	86.0	0.796	0.738
Q-Insight	0.814	0.822	0.616	0.169	0.559	0.171	81.2	0.690	0.653
VisualQuality-R1	0.814	0.824	0.617	0.161	0.556	0.162	81.3	0.688	0.653

Table 19: **Rankability, spacing and embedding geometry at ℓ_{mid} for each dataset (a): synthetic distortions.** Columns as in Table 14, with every metric computed at ℓ_{mid}. For Q-Insight and VisualQuality-R1, ℓ_{mid} is LLM layer 21.

	In-dataset metrics						Cross-dataset axis agreement		
Encoder	$\mathcal{R}_\rho$	$\mathcal{R}_r$	RAIS	A-QEA	OR	G-QEA	$\bar{\theta}$ (°)	$\bar{c}_\Sigma$	$\bar{\mathcal{R}}_r^\times$
LIVE									
ARNIQA	0.977	0.978	0.926	0.267	0.603	0.390	85.3	0.645	0.640
MANIQA	0.949	0.947	0.836	0.704	0.759	0.674	n/a	n/a	n/a
LIQE	0.963	0.963	0.876	0.417	0.649	0.418	78.8	0.868	0.881
QualiCLIP	0.954	0.954	0.852	0.414	0.652	0.462	82.8	0.684	0.674
FGResQ	0.887	0.890	0.693	0.251	0.588	0.238	79.3	0.793	0.742
Q-Align	0.967	0.972	0.908	0.748	0.775	0.776	88.1	0.816	0.805
Q-ReAlign	0.918	0.910	0.742	0.759	0.785	0.751	87.8	0.892	0.852
Q-Insight	0.963	0.964	0.883	0.788	0.806	0.760	87.5	0.900	0.906
VisualQuality-R1	0.958	0.960	0.871	0.841	0.833	0.846	88.0	0.912	0.909
TID2013									
ARNIQA	0.716	0.734	0.418	0.168	0.557	0.288	86.2	0.564	0.554
MANIQA	0.644	0.693	0.361	0.138	0.546	0.183	n/a	n/a	n/a
LIQE	0.688	0.716	0.375	0.166	0.558	0.200	82.7	0.748	0.678
QualiCLIP	0.663	0.697	0.354	0.178	0.561	0.242	84.3	0.564	0.582
FGResQ	0.643	0.681	0.345	0.135	0.545	0.181	82.9	0.606	0.675
Q-Align	0.799	0.816	0.535	0.462	0.664	0.518	89.2	0.839	0.762
Q-ReAlign	0.615	0.623	0.278	0.347	0.619	0.384	88.2	0.621	0.667
Q-Insight	0.746	0.785	0.487	0.387	0.637	0.456	87.9	0.780	0.733
VisualQuality-R1	0.804	0.837	0.586	0.533	0.697	0.581	88.1	0.837	0.763
KADID-10k									
ARNIQA	0.827	0.823	0.570	0.093	0.533	0.131	86.2	0.540	0.455
MANIQA	0.910	0.911	0.768	0.635	0.743	0.588	n/a	n/a	n/a
LIQE	0.834	0.829	0.572	0.079	0.528	0.068	80.8	0.588	0.612
QualiCLIP	0.781	0.774	0.477	0.111	0.538	0.112	85.5	0.560	0.473
FGResQ	0.682	0.684	0.342	0.054	0.521	0.046	82.1	0.658	0.561
Q-Align	0.938	0.940	0.819	0.700	0.757	0.719	88.9	0.833	0.787
Q-ReAlign	0.880	0.877	0.696	0.635	0.734	0.625	87.9	0.783	0.758
Q-Insight	0.894	0.892	0.702	0.391	0.639	0.384	87.5	0.728	0.678
VisualQuality-R1	0.919	0.919	0.770	0.726	0.780	0.715	89.2	0.810	0.774
PIPAL									
ARNIQA	0.421	0.403	0.100	0.057	0.520	0.066	88.6	0.399	0.305
MANIQA	0.647	0.631	0.312	0.217	0.574	0.221	n/a	n/a	n/a
LIQE	0.452	0.438	0.115	0.038	0.513	0.022	83.6	0.483	0.382
QualiCLIP	0.414	0.406	0.096	0.051	0.518	0.050	85.8	0.362	0.363
FGResQ	0.438	0.426	0.103	0.048	0.513	0.037	84.2	0.399	0.356
Q-Align	0.466	0.447	0.139	0.201	0.570	0.205	89.8	0.461	0.501
Q-ReAlign	0.288	0.280	0.037	0.122	0.541	0.101	87.8	0.300	0.375
Q-Insight	0.523	0.519	0.190	0.233	0.579	0.235	88.7	0.626	0.521
VisualQuality-R1	0.510	0.508	0.171	0.246	0.586	0.240	89.1	0.577	0.523

Table 20: **Rankability, spacing and embedding geometry at ℓ_{mid} for each dataset (b): authentic distortions and AI-generated images.** Columns as in Table 19. *Mean over the eight datasets*: mean of each column over all eight datasets, as in Table 2 for ℓ_{last}.

Encoder	In-dataset metrics $\mathcal{R}_\rho$	$\mathcal{R}_r$	RAIS	A-QEA	OR	G-QEA	Cross-dataset axis agreement $\bar{\theta}$ (°)	$\bar{c}_\Sigma$	$\mathcal{R}_r^\times$
LIVEC									
ARNIQA	0.708	0.774	0.516	0.139	0.547	0.165	83.8	0.625	0.538
MANIQA	0.773	0.830	0.599	0.523	0.691	0.607	n/a	n/a	n/a
LIQE	0.869	0.909	0.736	0.135	0.547	0.083	74.0	0.716	0.677
QualiCLIP	0.743	0.808	0.590	0.109	0.538	0.068	79.6	0.641	0.567
FGResQ	0.821	0.897	0.706	0.097	0.532	0.089	75.6	0.666	0.632
Q-Align	0.904	0.929	0.782	0.634	0.731	0.688	87.7	0.815	0.762
Q-ReAlign	0.885	0.911	0.742	0.536	0.696	0.581	85.4	0.819	0.750
Q-Insight	0.859	0.907	0.726	0.554	0.702	0.619	86.2	0.786	0.722
VisualQuality-R1	0.879	0.917	0.758	0.658	0.748	0.701	87.6	0.829	0.769
KonIQ-10k									
ARNIQA	0.878	0.898	0.713	0.190	0.564	0.204	83.7	0.666	0.606
MANIQA	0.881	0.894	0.712	0.399	0.638	0.413	n/a	n/a	n/a
LIQE	0.903	0.921	0.760	0.057	0.518	0.024	71.7	0.703	0.651
QualiCLIP	0.891	0.905	0.733	0.079	0.525	0.028	77.4	0.638	0.573
FGResQ	0.865	0.888	0.682	0.065	0.519	0.055	73.0	0.632	0.565
Q-Align	0.935	0.951	0.836	0.715	0.769	0.771	88.1	0.749	0.714
Q-ReAlign	0.905	0.927	0.779	0.496	0.678	0.569	85.5	0.765	0.708
Q-Insight	0.918	0.937	0.799	0.668	0.747	0.723	87.2	0.779	0.735
VisualQuality-R1	0.908	0.929	0.779	0.676	0.754	0.722	87.7	0.786	0.731
SPAQ									
ARNIQA	0.906	0.909	0.752	0.369	0.629	0.377	84.0	0.801	0.734
MANIQA	0.898	0.902	0.736	0.666	0.745	0.680	n/a	n/a	n/a
LIQE	0.911	0.912	0.753	0.236	0.578	0.202	75.2	0.813	0.742
QualiCLIP	0.914	0.917	0.767	0.312	0.607	0.286	79.2	0.796	0.729
FGResQ	0.904	0.902	0.725	0.233	0.578	0.229	74.8	0.806	0.729
Q-Align	0.941	0.944	0.834	0.770	0.791	0.792	87.0	0.839	0.791
Q-ReAlign	0.921	0.926	0.788	0.690	0.755	0.708	84.8	0.892	0.827
Q-Insight	0.918	0.924	0.784	0.695	0.760	0.699	85.8	0.917	0.848
VisualQuality-R1	0.922	0.927	0.789	0.750	0.784	0.772	87.6	0.902	0.839
AGIQA-3K									
ARNIQA	0.807	0.866	0.661	0.204	0.570	0.278	86.3	0.654	0.571
MANIQA	0.806	0.860	0.660	0.543	0.693	0.628	n/a	n/a	n/a
LIQE	0.838	0.889	0.703	0.186	0.563	0.197	81.2	0.735	0.645
QualiCLIP	0.825	0.871	0.672	0.146	0.550	0.171	84.3	0.659	0.579
FGResQ	0.837	0.888	0.704	0.137	0.547	0.165	79.1	0.680	0.604
Q-Align	0.863	0.913	0.754	0.521	0.683	0.596	88.2	0.759	0.692
Q-ReAlign	0.850	0.899	0.722	0.465	0.661	0.533	88.6	0.796	0.712
Q-Insight	0.862	0.906	0.735	0.476	0.668	0.567	87.6	0.841	0.773
VisualQuality-R1	0.862	0.912	0.746	0.508	0.679	0.589	87.9	0.842	0.771
Mean over the eight datasets									
ARNIQA	0.780	0.798	0.582	0.186	0.565	0.237	85.5	0.612	0.550
MANIQA	0.814	0.834	0.623	0.478	0.674	0.499	n/a	n/a	n/a
LIQE	0.807	0.822	0.611	0.164	0.557	0.152	78.5	0.707	0.659
QualiCLIP	0.773	0.792	0.568	0.175	0.561	0.177	82.4	0.613	0.568
FGResQ	0.760	0.782	0.538	0.128	0.543	0.130	78.9	0.655	0.608
Q-Align	0.852	0.864	0.701	0.594	0.718	0.633	88.4	0.764	0.727
Q-ReAlign	0.783	0.794	0.598	0.506	0.684	0.531	87.0	0.734	0.706
Q-Insight	0.835	0.854	0.663	0.524	0.692	0.555	87.3	0.795	0.739
VisualQuality-R1	0.845	0.864	0.684	0.617	0.733	0.646	88.2	0.812	0.760

The figure is qualitative: latitude reflects the standardised position along the rank-axis, not its share of the total variance.

Figures 7 and 8 give the same view at ℓ_{early} and ℓ_{mid}, with every rank-axis fit at that layer. At ℓ_{early}, the north-south order remains on most datasets, but the points spread more evenly over the sphere and form fewer thin curves. This is most visible for the vision towers of Q-Insight and VisualQuality-R1, and is consistent with their low Anchor-QEA at this layer (Appendix A.5.2). At ℓ_{mid}, the panels are close to those at ℓ_{last}. Q-Insight and VisualQuality-R1, now read from LLM layer 21, again form thin curves from south to north (for example on KonIQ-10k and SPAQ).

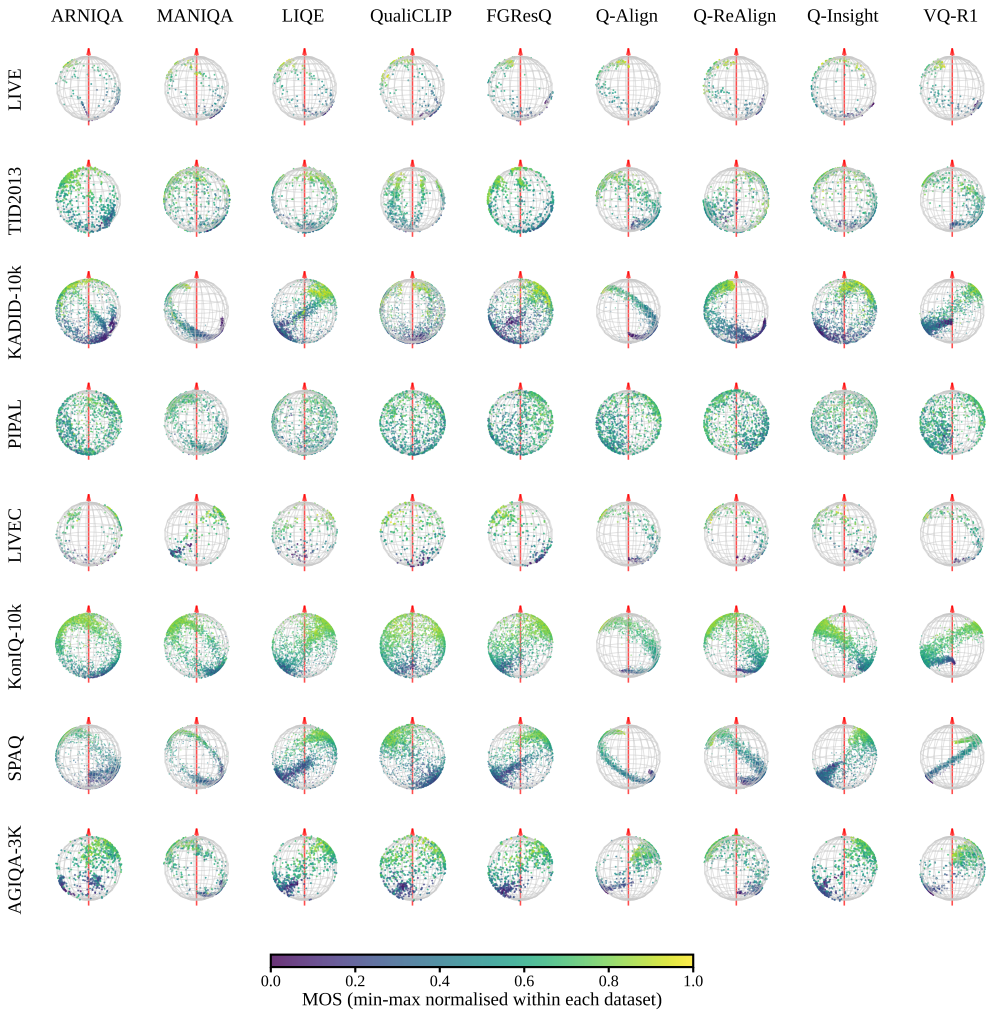

Figure 6: **Test embeddings on the unit sphere around the rank-axis at ℓ_{last}.** Rows: datasets. Columns: encoders. The red arrow is the rank-axis, from the south pole to the north pole. Colour: MOS, min-max normalised within each dataset. Each panel shows all test images. VQ-R1: VisualQuality-R1.

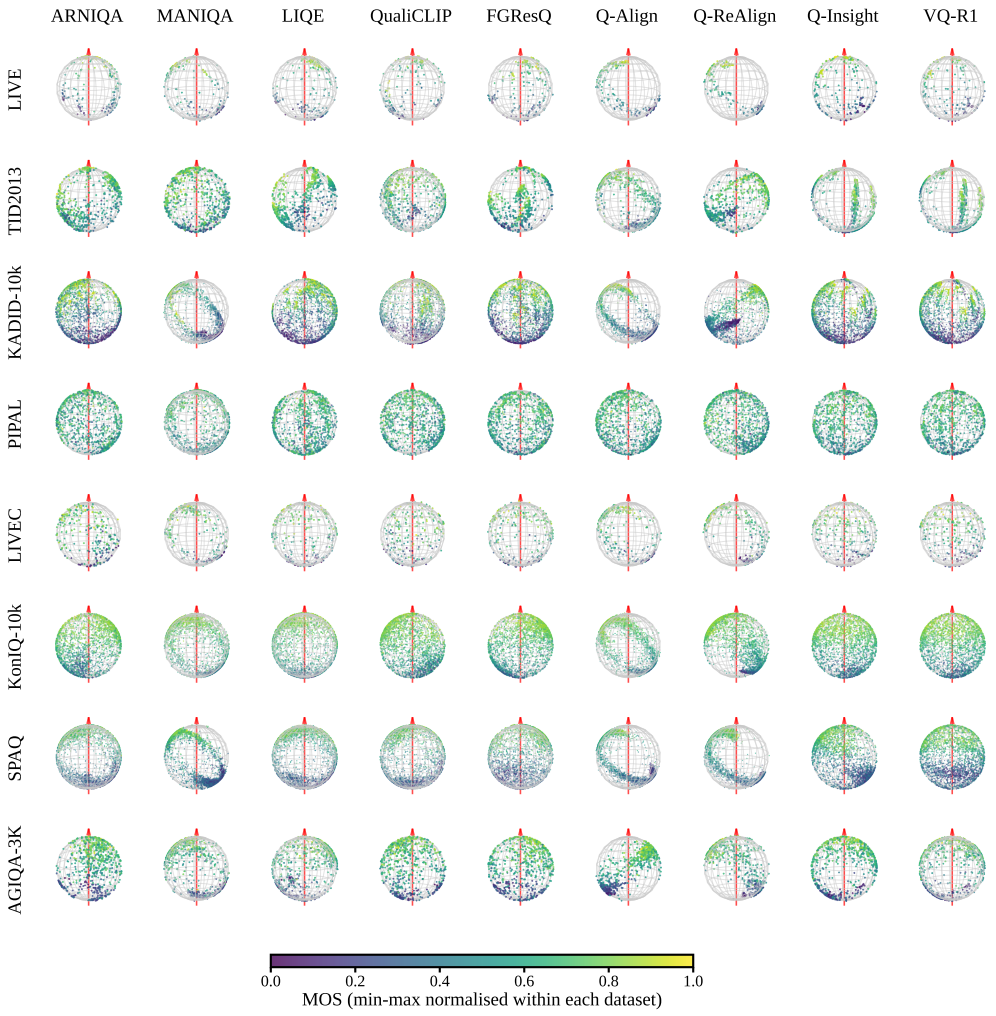

Figure 7: **Test embeddings on the unit sphere around the rank-axis at ℓ_{early}.** Layout and colour as in Figure 6. For Q-Insight and VisualQuality-R1, ℓ_{early} is vision block 15.

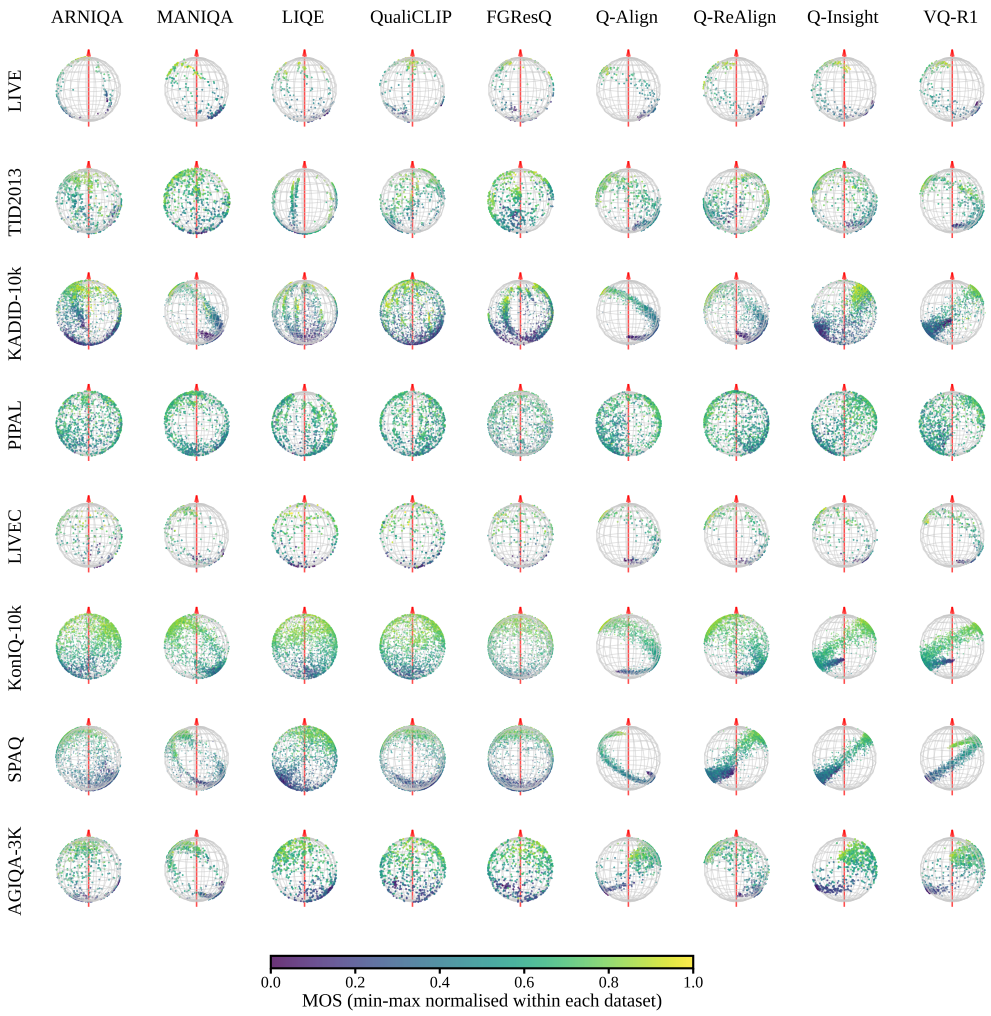

Figure 8: **Test embeddings on the unit sphere around the rank-axis at ℓ_{mid}.** Layout and colour as in Figure 6. For Q-Insight and VisualQuality-R1, ℓ_{mid} is LLM layer 21.

B Proof of Proposition 1

Proof. Let y, y' be independent copies of the label, $\Delta = |y - y'|$ and $\delta = |g(y) - g(y')|$. Because y is continuous, so is Δ; because g is strictly monotone, $g(y)$ is continuous and hence so is δ.

($\Leftarrow$) If $g(u) = \gamma u + \eta$ with $\gamma \neq 0$, then $\delta = |\gamma|\Delta$ is a strictly increasing function of Δ, so RAIS $= 1$.

($\Rightarrow$) For continuous random variables, Spearman correlation equals 1 if and only if they are comonotone, i.e. $F_\delta(\delta) = F_\Delta(\Delta)$ almost surely, where F_δ, F_Δ are their distribution functions. Hence $\delta = h(\Delta)$ almost surely with $h = F_\delta^{-1} \circ F_\Delta$ nondecreasing, which is all we use below. Let $H(u, u') = |g(u) - g(u')|$, which is continuous on $I \times I$.

Step 1: $H(u, u+t)$ does not depend on u. Take u, v, t with $u, u+t, v, v+t$ in the interior of I and suppose $H(u, u+t) > H(v, v+t)$. By continuity there are open neighbourhoods $N_1 \ni (u, u+t)$ and $N_2 \ni (v, v+t)$ with $\inf_{N_1} H > \sup_{N_2} H$. The sets $N_1 \cap \{|u - u'| < t\}$ and $N_2 \cap \{|u - u'| > t\}$ are nonempty and open, so they have positive probability because the support of the labels is I. Choosing points in them outside the null set on which $H \neq h(|u - u'|)$ yields pairs with a smaller quality gap but a larger distance, contradicting the monotonicity of h. The reverse inequality is excluded symmetrically, and continuity extends the equality to the boundary of I.

Step 2: g is affine. Replacing g by $-g$ if necessary, which leaves H unchanged, assume g is increasing. By Step 1, $k(t) = g(u+t) - g(u)$ does not depend on u, so $k(t_1 + t_2) = k(t_1) + k(t_2)$ whenever $u, u + t_1, u + t_1 + t_2 \in I$. Since k is continuous, this restricted Cauchy equation forces $k(t) = \gamma t$ for some $\gamma > 0$ (positive because g is strictly increasing), and hence $g(u) = g(u_0) + \gamma(u - u_0)$ for all $u, u_0 \in I$. □

C Experimental details

This appendix gives the full experimental details summarised in Section 4. Table 21 lists the eight benchmarks with their splits, and Table 22 lists the nine encoders with the training data of each checkpoint and the extracted layers. Appendix D describes the extracted modules of each encoder.

Datasets and splits. We use eight IQA benchmarks that cover four kinds of content (Table 21). *Synthetic distortions*: LIVE (Sheikh et al., 2006), TID2013 (Ponomarenko et al., 2013) and KADID-10k (Lin et al., 2019). *Restoration and GAN outputs*: PIPAL (Jinjin et al., 2020). *Authentic, in-the-wild distortions*: LIVE Challenge (LIVEC) (Ghadiyaram & Bovik, 2016), KonIQ-10k (Hosu et al., 2020) and SPAQ (Fang et al., 2020). *AI-generated images*: AGIQA-3K (Li et al., 2024). Every quality label is converted to MOS convention (LIVE's DMOS is negated) and min-max scaled to $[0, 1]$ within its dataset. Since all metrics are invariant to positive affine transformations of the labels, this scaling only fixes the direction of "better" quality. We use one fixed train/test split per dataset for all encoders. KonIQ-10k uses its official split. As PIPAL's test set is not public, its official validation set serves as the test set. For the other datasets we use the fixed-seed (seed 123) folds distributed with IQA-PyTorch. For LIVE and TID2013 these folds also include a validation part, which we merge into train. This gives roughly 8:2 train/test splits everywhere (LIVE: $589/190$). We built an 8:2 split for AGIQA-3K with the same protocol. KonIQ-10k's official validation set is likewise merged into its train split. No split shares images, and TID2013's splits also share no reference scenes. Rank-axes are fit on the train split, and all reported results are computed on the test split.

Encoders. We study nine publicly released IQA models (Table 22) that cover the main design families. (i) *Classic deep neural networks*: ARNIQA (Agnolucci et al., 2024b), a ResNet-50 trained self-supervised on distortion manifolds, with one linear regressor per dataset; and MANIQA (Yang et al., 2022), a ViT-B/8 with Swin refinement blocks and one checkpoint trained per dataset. (ii) *CLIP-based models*: LIQE (Zhang et al., 2023) (ViT-B/32, multitask vision-language training), Quali-CLIP (Agnolucci et al., 2024a) (RN50, opinion-unaware) and FGResQ (Sheng et al., 2026) (two ViT-B/16 towers trained on restoration quality). (iii) *Multimodal LLMs*: Q-Align (Wu et al., 2024) (mPLUG-Owl2, LLaMA-2 7B) and Q-ReAlign (Team, 2026) (Qwen3.5-VL, 0.8B), both trained on discrete rating levels; and Q-Insight (Li et al., 2025) and VisualQuality-R1 (Wu et al., 2025) (Qwen2.5-VL 7B, trained with reinforcement learning). We use the single general-purpose pre-trained model checkpoint for LIQE: `mix`; Q-Align: `one-align`; Q-ReAlign: `Mini-0.8B`. A

dataset counts as *in-distribution* for an encoder when it is in that checkpoint's training data (marked * in Tables 3 and 4). Every other dataset is out-of-distribution (OOD) for that encoder.

Feature extraction. Each image goes through one forward pass at the model's native preprocessing, except that MANIQA uses a single crop instead of its usual 20 (Appendix D). For every encoder we extract an early layer (ℓ_{early}), a middle layer (ℓ_{mid}) and the input of the scoring head (ℓ_{last}). For the single-tower encoders (ARNIQA, MANIQA, LIQE, QualiCLIP, FGResQ) with L blocks, ℓ_{early} and ℓ_{mid} are blocks $\text{round}(p\,(L-1))$ for $p = 0.33$ and $p = 0.66$. For Q-Align and Q-ReAlign they are language-model layers 10 and 21 of 32, and 8 and 15 of 24. For Q-Insight and VisualQuality-R1, which share the Qwen2.5-VL architecture, ℓ_{early} is vision block 15, before the language model, and ℓ_{mid} is language-model layer 21. Spatial and token dimensions are mean-pooled. For MLLMs, language-model features are the hidden state of the last prompt token, just before the rating is produced. For Q-Align, Q-Insight and VisualQuality-R1 we also extract the vision-tower output that is fed to the LLM (ℓ_{vis}), and for ARNIQA the head input weighted by its regressor (ℓ_{w}; Appendix D). This gives 31 representations each extracted for all images across 8 datasets.

Rank-axis. For a representation and a source dataset a, the rank-axis is estimated by the ridge-regression direction (Hoerl & Kennard, 1970) $(\mathbf{w}_a, \beta_a) = \arg\min_{\mathbf{w},\beta} \|\mathbf{y}_a^{\text{tr}} - \mathbf{X}_a^{\text{tr}}\mathbf{w} - \beta\mathbf{1}\|_2^2 + \alpha\|\mathbf{w}\|_2^2$ (Eq. (1)), fit on the train split. $\alpha \in \{10^{-1}, \ldots, 10^{4}\}$ (12 log-spaced values) is chosen by efficient leave-one-out cross-validation. We use the unit-norm direction $\hat{\mathbf{w}}_a = \mathbf{w}_a / \|\mathbf{w}_a\|$. The projection score of a test image is $s = \mathbf{x}^\top \hat{\mathbf{w}}_a$. The rank-axis is a single linear direction, with no nonlinearity and no feature re-weighting beyond the ridge penalty.

Estimation of the metrics. All metrics are computed on the test split, and the sample sizes below are upper limits. RAIS (Eq. (3)) and Global-QEA (Eq. (6)) share the same pairs of test images: all pairs, or at most 700,000 pairs sampled uniformly without replacement. In Anchor-QEA (Eq. (4)), every test image serves as an anchor and is compared with all other test images, or with at most 1,000 of them sampled at random. The ordering rate (Eq. (5)) uses 100,000 random triplets. The pair and neighbour limits apply only to the three largest test sets (KADID-10k, KonIQ-10k and SPAQ); all other test sets are used in full.

Cross-dataset transfer. Cross-dataset analyses require one feature space for all datasets and therefore exclude MANIQA (one checkpoint per dataset) and ARNIQA's regressor-weighted feature ℓ_{w} (Appendix D). This leaves 27 representations $\times$ 56 ordered dataset pairs $= 1{,}512$ transfer pairs.

Fine-tuned heads. As a reference we report the SRCC of each encoder's own scoring head, evaluated with its official inference code on the same test splits. For ARNIQA and MANIQA we use the head trained on the target dataset (ARNIQA ships none for PIPAL or AGIQA-3K). We use the single released checkpoint for all other encoders.

Caveats. Only KonIQ-10k has an official test split, and the fine-tuned heads were trained on their authors' splits. The fine-tuned-head results are therefore a reference, not a controlled baseline. TID2013's test split contains only 5 of its 25 reference scenes; its results therefore have high variance. The ridge penalty is chosen by image-level leave-one-out, which can overfit to scene content on TID2013 at intermediate layers. Scene-grouped cross-validation gives more stable values there. LIQE, Q-Align, Q-ReAlign, Q-Insight and VisualQuality-R1 have seen some of our datasets during training (* in Tables 3 and 4), which inflates their numbers on those datasets. MANIQA is excluded from every cross-dataset analysis because its feature space differs per dataset.

Table 21: **Benchmark datasets.** All scores are converted to MOS convention (higher is better) and min-max scaled to $[0, 1]$. Rank-axes are fit on the train split only and every metric is reported on the test split. Folds use seed 123. †PIPAL's official test set is not public; its official validation set is used as test. ‡The split's validation part is merged into train.

Dataset	Content	Label	#Images	#Train	#Test	Split
LIVE	Synthetic	DMOS	779	589	190	IQA-PyTorch fold‡
TID2013	Synthetic	MOS	3,000	2,400	600	IQA-PyTorch fold‡
KADID-10k	Synthetic	MOS	10,125	8,000	2,125	IQA-PyTorch fold
PIPAL	Restoration / GAN	MOS	24,200	23,200	1,000	official†
LIVEC	Authentic	MOS	1,155	925	230	IQA-PyTorch fold
KonIQ-10k	Authentic	MOS	10,073	8,058	2,015	official‡
SPAQ	Authentic (smartphone)	MOS	11,125	8,900	2,225	IQA-PyTorch fold
AGIQA-3K	AI-generated	MOS	2,982	2,386	596	ours

Table 22: **Pre-trained encoders.** For each encoder we extract features at an early and a middle layer (ℓ_{early}, ℓ_{mid}; defined per encoder in the text) and the representation right before the scoring head (ℓ_{last}). Tokens are mean-pooled; MLLM language-model layers use the hidden state of the last prompt token. For the three large MLLMs we also extract the vision-tower output (ℓ_{vis}).

Encoder	Family	Backbone	Training data of the checkpoint	ℓ_{early} / ℓ_{mid} / ℓ_{last} (dim)
ARNIQA	Self-sup. CNN	ResNet-50	KADIS-700k (SSL) + per-dataset linear head	L2 / L3 / pool (1024 / 2048 / 4096)
MANIQA	ViT + Swin	ViT-B/8 + 2×Swin	one checkpoint per dataset	ViT-8 / Swin-1 / pre-head (768 / 768 / 384)
LIQE	CLIP	CLIP ViT-B/32	LIVE, CSIQ, KADID, BID, LIVEC, KonIQ	blk 4 / blk 7 / proj (768 / 768 / 512)
QualiCLIP	CLIP	CLIP RN50	KonIQ images, opinion-unaware	L2 / L3 / attn-pool (512 / 1024 / 1024)
FGResQ	CLIP	2×CLIP ViT-B/16	FGRestore (restoration)	blk 4 / blk 7 / pre-head (768 / 768 / 2304)
Q-Align	MLLM	mPLUG-Owl2 (LLaMA-2 7B)	KonIQ, SPAQ, KADID, AVA, LSVQ	LLM 10 / 21 / last (4096); ViT (1024)
Q-ReAlign	MLLM	Qwen3.5-VL 0.8B	KonIQ, SPAQ, KADID, AGIQA-20K, AVA, LSVQ	LLM 8 / 15 / last (1024)
Q-Insight	MLLM (RL)	Qwen2.5-VL 7B	KonIQ (GRPO)	ViT 15 / LLM 21 / last (1280 / 3584 / 3584); ViT (3584)
VisualQuality-R1	MLLM (RL)	Qwen2.5-VL 7B	KADID (RL-to-rank)	ViT 15 / LLM 21 / last (1280 / 3584 / 3584); ViT (3584)

D ENCODER ARCHITECTURES AND EXTRACTED LAYERS

This appendix describes, for each encoder, the checkpoint we use, its architecture, and the exact modules whose outputs we extract as ℓ_{early}, ℓ_{mid}, ℓ_{last} (and ℓ_{vis} or ℓ_{w} where applicable). For the single-tower encoders (ARNIQA, MANIQA, LIQE, QualiCLIP, FGResQ), the depth is the sequence of transformer or residual blocks in execution order, and ℓ_{early} and ℓ_{mid} are blocks $\text{round}(p\,(L-1))$ for $p = 0.33, 0.66$ in that sequence (0-indexed). For the MLLMs they are given explicitly below. In all cases, ℓ_{last} is the input of the scoring head. Token and spatial dimensions are mean-pooled (the CLS token is included where present), and every feature is stored as one float32 vector per image. All weights are frozen and every image is processed in a single deterministic forward pass. Table 23 summarises the extracted modules.

Table 23: **Extracted modules.** Module names follow the released implementations (IQA-PyTorch for ARNIQA, MANIQA, LIQE, QualiCLIP and FGResQ; HuggingFace for the MLLMs). Blocks are 0-indexed. "Last tok.": hidden state of the last prompt token.

Encoder	Depth L	ℓ_{early}	ℓ_{mid}	ℓ_{last}	Extra
ARNIQA	16 bottlenecks	`layer2.2` (2 scales, 1024)	`layer3.3` (2 scales, 2048)	avg-pool, 2 scales (4096)	ℓ_{w} (4096)
MANIQA	24 blocks	`vit.blocks.8` (768)	`swin1.layers.0.blocks.1` (768)	`swintransformer2` (384)	n/a
LIQE	12 blocks	`resblocks.4` (768)	`resblocks.7` (768)	`visual` proj. (512)	n/a
QualiCLIP	16 bottlenecks	`layer2.2` (512)	`layer3.3` (1024)	`attnpool` (1024)	n/a
FGResQ	12 blocks	`encoder.layers.4` (768)	`encoder.layers.7` (768)	head input (2304)	n/a
Q-Align	32 LLM layers	LLM layer 10, last tok. (4096)	LLM layer 21, last tok. (4096)	final norm, last tok. (4096)	ViT mean (1024)
Q-ReAlign	24 LLM layers	LLM layer 8, last tok. (1024)	LLM layer 15, last tok. (1024)	final norm, last tok. (1024)	n/a
Q-Insight	n/a	ViT block 15 (1280)	LLM layer 21, last tok. (3584)	final norm, last tok. (3584)	merger mean (3584)
VisualQuality-R1	n/a	ViT block 15 (1280)	LLM layer 21, last tok. (3584)	final norm, last tok. (3584)	merger mean (3584)

ARNIQA (Agnolucci et al., 2024b). A single ResNet-50 encoder (torchvision layout, `fc` removed) shared across datasets, plus one linear regressor $\mathbb{R}^{4096} \to \mathbb{R}$ per dataset. The encoder is trained self-supervised on KADIS-700k by contrasting images with the same synthetic degradation; the regressors are then fit on each IQA dataset with the encoder frozen. Each image is encoded twice, at full resolution and after $2\times$ downsampling, and the two pooled outputs are concatenated

before the regressor. Counting the $3 + 4 + 6 + 3 = 16$ bottleneck blocks of `layer1-layer4` as the depth, ℓ_{early} is `layer2.2` and ℓ_{mid} is `layer3.3`. At both layers we average-pool each scale and concatenate the two scales (2×512 and 2×1024 dimensions). ℓ_{last} is the encoder output after average pooling; as in ARNIQA's own forward pass, each scale is ℓ_2-normalised and the two are concatenated ($2 \times 2048 = 4096$), which makes ℓ_{last} exactly the regressor's input. ℓ_{w} is ℓ_{last} multiplied elementwise by the regressor weights of the dataset; summing its entries and adding the bias gives ARNIQA's score. Since no regressor is released for PIPAL or AGIQA-3K, ℓ_{w} uses KonIQ-10k's regressor for these two datasets.

MANIQA (Yang et al., 2022). A ViT-B/8 backbone followed by two refinement stages, each consisting of two transposed-attention blocks (TAB) and a Swin Transformer with depths $(2, 2)$, and a patch-wise score and weight head. The Swin stages act on the concatenated outputs of four intermediate ViT blocks. Counting 12 ViT blocks, 2 TABs, 4 Swin blocks, 2 TABs and 4 Swin blocks gives $L = 24$. Accordingly, ℓ_{early} is ViT block 8 (785 tokens $\times$ 768, including CLS) and ℓ_{mid} is the second block of the first Swin layer of stage 1 (784 tokens $\times$ 768). ℓ_{last} is the output of the second Swin stage ($384 \times 28 \times 28$), average-pooled to 384 dimensions. MANIQA normally averages 20 crops of 224×224 at test time; we use a single crop, which with IQA-PyTorch's uniform cropping is the top-left 224×224 window (images with a side shorter than 224 are first upsampled). There is one checkpoint per dataset: the released ones for KonIQ-10k, KADID-10k and PIPAL, and checkpoints we trained with IQA-PyTorch on our train splits for LIVE, TID2013, LIVEC, SPAQ and AGIQA-3K. The residual scale of the Swin blocks is set to the value each checkpoint was trained with (0.8 for the released KonIQ-10k and KADID-10k checkpoints, 0.13 otherwise).

LIQE (Zhang et al., 2023). A CLIP model with a ViT-B/32 image tower (12 blocks, width 768) that scores an image by matching its embedding against fixed text prompts. We use the multitask `mix` checkpoint for every dataset. It was trained jointly on quality, scene and distortion-type prompts over LIVE, CSIQ, KADID-10k, BID, LIVEC and KonIQ-10k; TID2013 is used only as a cross-dataset test set in the original work (Zhang et al., 2023) and is therefore out-of-distribution for LIQE. At test time the image is split into up to 15 overlapping 224×224 patches that are encoded as one batch. ℓ_{early} and ℓ_{mid} are the outputs of residual blocks 4 and 7, averaged over the 50 tokens and then over patches; ℓ_{last} is the 512-dimensional projected image embedding (before ℓ_2 normalisation and text matching), averaged over patches.

QualiCLIP (Agnolucci et al., 2024a). A CLIP model with a ResNet-50 image tower (CLIP's `ModifiedResNet`: a three-convolution stem, the same $3{+}4{+}6{+}3$ bottleneck layout as ARNIQA, and attention pooling instead of average pooling). It is trained without human opinion scores, by ranking synthetically degraded versions of KonIQ-10k images against antonym text prompts. We use the base `qualiclip` checkpoint (no learned prompts) for every dataset. The image is encoded once, without patching. ℓ_{early} and ℓ_{mid} are `layer2.2` and `layer3.3`, globally average-pooled (512 and 1024 dimensions), and ℓ_{last} is the 1024-dimensional attention-pooled image embedding that is matched against the prompts.

FGResQ (Sheng et al., 2026). Two CLIP ViT-B/16 image towers and a linear quality head, with one checkpoint for all datasets. The second tower encodes degradation information and is trained on fine-grained restoration-quality data (FGRestore). We extract from the main tower, which has 12 encoder layers: ℓ_{early} and ℓ_{mid} are layers 4 and 7 (197 tokens $\times$ 768, mean-pooled). ℓ_{last} is the input of the quality head, the concatenation of the pooled CLIP feature, a task embedding and their sum ($3 \times 768 = 2304$).

Q-Align (Wu et al., 2024). mPLUG-Owl2: a CLIP ViT-L/14 vision tower at 448×448, a visual abstractor, and a LLaMA-2 7B language model (32 layers, hidden size 4096). We use the `one-align` checkpoint, trained on KonIQ-10k, SPAQ, KADID-10k, AVA and LSVQ. Images are padded to a square with the CLIP mean colour, as in Q-Align's own preprocessing. The score is read from the logits of the rating words at the last prompt token. Every LLM feature is therefore the hidden state of that token (mean-pooling would mix in image tokens). ℓ_{early} and ℓ_{mid} are the outputs of LLM layers 10 and 21, and ℓ_{last} is the output of the final RMSNorm, the input of the language-model head. ℓ_{vis} is the vision tower's last hidden state, mean-pooled over tokens (1024 dimensions); in mPLUG-Owl2 it is passed through the visual abstractor before reaching the LLM.

Q-ReAlign (Team, 2026). We use the `Mini-0.8B` checkpoint, whose language model is a Qwen3.5 decoder with 24 layers and hidden size 1024. It is trained on KonIQ-10k, SPAQ, KADID-10k, AGIQA-20K, AVA and LSVQ with discrete rating levels. As for Q-Align, the score comes from the logits at the last prompt token ("...The quality of the image is"), and all features are that token's hidden state: decoder layers 8 and 15 for ℓ_{early} and ℓ_{mid}, and the final RMSNorm output for ℓ_{last}. We do not extract its vision tower.

Q-Insight (Li et al., 2025) and VisualQuality-R1 (Wu et al., 2025). Both are Qwen2.5-VL 7B models, with a vision transformer of hidden size 1280, a patch merger that maps vision tokens to the LLM width, and a language model of hidden size 3584 (32 vision blocks and 28 LLM layers). We use Q-Insight's `score_degradation` checkpoint (trained with GRPO on KonIQ-10k) and the released VisualQuality-R1 checkpoint (trained with reinforcement learning to rank on KADID-10k). Images are processed at the Qwen2.5-VL processor's native dynamic resolution. Each image is passed once through the model together with the model's own rating prompt, without generation. Since both models emit their rating after a generated reasoning chain (Q-Insight) or answer tag (VisualQuality-R1), the last-prompt-token features describe the state *before* any tokens are generated. Because the two models share an architecture, we extract the same modules from both: ℓ_{early} is the output of vision block 15, mean-pooled over tokens (1280 dimensions), ℓ_{mid} is LLM layer 21 at the last prompt token, ℓ_{last} is the final normalised LLM hidden state at the last prompt token, and ℓ_{vis} is the patch-merger output, i.e. the visual tokens the LLM receives, mean-pooled (3584 dimensions). Thus ℓ_{early} lies in the vision tower, before the language model, and ℓ_{mid} in the language model.

E Extended related work

Rankability and ordinal structure in embeddings. Sonthalia et al. formalise *rankability* as a property of visual embeddings: given an encoder f and a continuous attribute A, the model is rankable if there exists a linear *rank-axis* v_A such that the projection $v_A^\top f(x)$ preserves the ground-truth order of $A(x)$ across a dataset (Sonthalia et al., 2025). Their study covers seven encoders and nine datasets for attributes including age, crowd count, head pose angles, image aesthetics, and recency, and shows that many modern embeddings are inherently rankable, often from very sparse supervision (extreme pairs). OrdinalCLIP introduces a differentiable prompting method that adapts CLIP for ordinal regression, constructing learnable rank embeddings that enforce numerical continuity for age estimation and aesthetic assessment (Li et al., 2022). Follow-up work such as L2RCLIP further refines language-driven ordering alignment by learning rank prompts and cross-modal ordinal pairwise losses, improving ordering accuracy on age, historical image dating, and aesthetics (Wang et al., 2023b). Rank-N-Contrast generalises ranking objectives to continuous regression by defining a contrastive loss that pulls together embeddings of similar-label samples and pushes apart dissimilar-label samples proportionally to their label differences, yielding continuous representations where distances encode both order and magnitude (Zha et al., 2023).

IQA datasets and classical FR/NR metrics. IQA has a long history of benchmark datasets and hand-crafted metrics. LIVE, TID2013, KADID-10k, KonIQ-10k, SPAQ, and AVA are widely used datasets that span synthetic distortions, authentic in-the-wild degradations, smartphone photography, and aesthetic quality (Sheikh et al., 2006; Ponomarenko et al., 2013; Lin et al., 2019; Hosu et al., 2020; Fang et al., 2020; Murray et al., 2012). Classical full-reference metrics such as SSIM and MS-SSIM measure structural similarities between reference and distorted images across luminance, contrast, and structure, and remain standard baselines for evaluating image processing algorithms (Wang et al., 2004; 2003). Reduced-reference methods like RRED use entropic differences of subband coefficients to quantify distortion while only partially accessing the reference (Soundararajan & Bovik, 2012). Blind/no-reference IQA developed through natural scene statistics models, including BRISQUE and BLIINDS, which rely on distributions of locally normalized luminance or DCT coefficients to produce distortion-generic quality scores without a reference (Mittal et al., 2012; Saad et al., 2012). NIMA shifts toward learned perceptual metrics by training deep networks to predict opinion score distributions on AVA and TID2013, demonstrating that deep models can capture subjective quality beyond hand-crafted indices (Talebi & Milanfar, 2018). Opinion-unaware NSS-based metrics such as NIQE provide further baselines that operate without MOS labels, and recent work on global statistical and global awareness features extends NSS ideas to coarser-scale descriptors (Mittal et al., 2013; Varga, 2021; Hu et al., 2024).

Deep NR-IQA architectures. Deep learning has produced a wide range of NR-IQA architectures on the above datasets. Bosse et al. propose deep neural network approaches for NR and FR IQA, in which patch-wise CNN features and learned weights are aggregated into global scores, forming the WaDIQaM family of BIQA baselines (Bosse et al., 2018). Kim and Lee introduce DeepQA, a CNN-based FR-IQA framework that learns human visual sensitivity directly from IQA databases, showing the potential of end-to-end modeling of FR distortions (Kim & Lee, 2017). Zhang et al. propose DBCNN, a deep bilinear CNN with two streams for synthetic and authentic distortions, fused via bilinear pooling, achieving state-of-the-art blind IQA performance across multiple synthetic and authentic datasets (Zhang et al., 2020). MUSIQ and MANIQA use multi-scale transformers and multi-dimension attention networks respectively, yielding strong NR-IQA results on large in-the-wild benchmarks (Ke et al., 2021; Yang et al., 2022). MetaIQA introduces a meta-learning framework to improve cross-dataset generalisation, while multimodal dense CNNs and intermediary-enhancement BIQA models explicitly target robustness on authentic distortions (Zhu et al., 2020; Chockalingam & Murugan, 2023; Song et al., 2022).

CLIP-based and opinion-unaware IQA. Pretrained vision-language encoders, especially CLIP, have recently been repurposed for NR-IQA without task-specific training. CLIP-IQA shows that carefully designed antonym prompts (e.g. "good photo" vs. "bad photo") can be used to derive zero-shot quality scores that correlate well with MOS on aesthetics and IQA datasets (Wang et al., 2023a). Miyata's interpretable CLIP IQA method uses multiple antonym-prompt pairs targeting specific perceptual attributes (sharpness, noise, brightness), together with a prompt pairing strategy, to produce both quality scores and explanations (Miyata, 2023). QualiCLIP introduces a quality-aware image-text alignment strategy that fine-tunes CLIP to rank synthetically degraded images by distortion severity using quality-related antonym prompts, while enforcing representation consistency across content and severity levels in an opinion-unaware manner (Agnolucci et al., 2024a). DSMix proposes a distortion-induced sensitivity map-based pretraining that improves CLIP/ViT-based NR-IQA by emphasizing distortion-sensitive regions during training, and ATTIQA develops an attribute-aware pretraining framework that uses CLIP guidance to extract quality-related attributes and construct generalizable quality features (Shi et al., 2024; Kwon et al., 2024). MA-CLIP demonstrates that CLIP feature magnitude carries robust quality signal and proposes a dual-cue NR-IQA combining cosine similarity with magnitude and confidence weighting (Liao et al., 2026). Recent VL/LMM-based IQA methods push this further: Grounding-IQA integrates multimodal referring and grounding into IQA, introducing GIQA-DES and GIQA-VQA subtasks and a GIQA-160K dataset for fine-grained, region-aware quality assessment (Chen et al., 2026). SEAGULL uses a vision-language model together with SAM-generated masks and a mask-based feature extractor to assess region-of-interest quality, and constructs SEAGULL-100w and SEAGULL-3k ROI-based IQA datasets for pretraining and fine-tuning (Chen et al., 2024). Opinion-unaware BIQA approaches such as multi-scale deep feature statistics and adversarial convolutional variational autoencoder IQA provide alternative routes for learning quality-sensitive representations from pristine images or deep statistics rather than MOS labels (Shukla et al., 2024; Ni et al., 2024). These works show that frozen or lightly adapted embeddings already contain rich quality-relevant information.

Ranking-based training objectives and ordinal attributes. Several IQA models explicitly incorporate ranking losses to encourage order-consistent representations. RankIQA uses synthetically generated distortion pairs to train Siamese networks on relative quality judgements, then transfers the learned ranking ability to single-image CNN regressors, thereby shaping representations so that pairwise distances track perceptual rankings (Liu et al., 2017). Rank-N-Contrast generalises this idea to continuous regression tasks by defining a contrastive objective that pulls together embeddings of similar-label samples and pushes apart dissimilar-label samples proportionally to their label distances, achieving improved robustness and generalisation (Zha et al., 2023). OrdinalCLIP and L2RCLIP, mentioned above, adapt CLIP to ordinal regression tasks such as age estimation and aesthetic assessment by learning rank prompts and cross-modal ordinal losses, providing CLIP-based exemplars of rank-aware training (Li et al., 2022; Wang et al., 2023b). HiRQA introduces higher-order ranking losses and quality-alignment terms that supervise relational ordering across multiple distortion pairs and encourage feature-space distances to reflect quality differences even without opinion labels (Ramesh et al., 2026). Beyond IQA, early work on relative attributes and ordinal labels explored how continuous visual attributes could be inferred and ranked, but these efforts were limited to smaller models and datasets (Parikh & Grauman, 2011).

AI-generated image quality assessment. The rise of GANs and diffusion models has motivated IQA benchmarks and methods tailored to AI-generated images. AGIQA-3K introduces an open database of AI-generated images from diverse text-to-image and image-to-image models, with human scores for perceptual quality and text-image alignment, providing a fine-grained testbed for AGIQA models (Li et al., 2024). PKU-AIGIQA-4K extends this line by establishing a large-scale perceptual quality assessment database for both text-to-image and image-to-image AIGIs, and proposes a partial-reference AIGIQA method for joint quality evaluation of prompts and images (Yuan et al., 2025). CLIP-based AGIQA methods adapt prompt strategies and CLIP-based regressors to score both quality and alignment, leveraging CLIP's joint image-text space to detect subtle failures such as semantic mismatch or artifacts (Tang et al., 2025). Opinion-unaware BIQA approaches like multi-scale deep feature statistics and adversarial CVAE-based IQA provide alternative routes for assessing AI-generated images when MOS labels are scarce, focusing on deep feature statistics or latent distribution properties correlated with distortion or artifact severity (Shukla et al., 2024; Ni et al., 2024). More recent works such as Grounding-IQA and SEAGULL, discussed above, use multimodal grounding and ROI-based reasoning to provide fine-grained, region-level quality judgements in generative contexts (Chen et al., 2026; 2024). In our experiments, AGIQA-3K is the testbed for AI-generated images: rank-axes recovered on the non-generative IQA datasets are evaluated on it, which allows us to study cross-regime axis transfer.

Representational similarity and distance alignment. Comparing dissimilarity structures via correlation has long been used to evaluate alignment between representations and behavioural or semantic spaces. The Mantel test correlates two distance matrices (e.g. geographic vs. genetic distances) under permutation-based null hypotheses (Mantel, 1967). Representational Similarity Analysis (RSA) computes representational dissimilarity matrices from neural or model activations and compares them to RDMs derived from behavioural or semantic labels, linking embedding geometry to perception or semantics (Kriegeskorte et al., 2008). In high-dimensional visual embeddings, however, raw Euclidean distance can be dominated by semantic content variation, and Mantel or RSA-style correlations computed directly on full-space distances can then conflate content and quality (Section 5.3). Our RAIS metric can be viewed as a constrained RSA: we first recover a one-dimensional quality rank-axis using the rankability framework and then compute Spearman correlation between projected distances along that axis and MOS/DMOS gaps, thereby isolating the quality-relevant component of the embedding. Recent BIQA models with explicit distance-consistency or covariance-pooling losses similarly emphasise the importance of aligning feature-space geometry with perceptual differences, but typically evaluate this via prediction metrics rather than a dedicated alignment score (Zhang et al., 2020; Hu et al., 2024). Work such as A-FINE, which relaxes the assumption that FR reference images are perfect and models a more general perceptual quality geometry, provides further evidence that the structure of representation space matters beyond scalar correlation alone (Chen et al., 2025).

Cross-dataset generalisation and universal quality axes. Deep IQA models are often evaluated on their ability to generalise across datasets with different distortion regimes and label scales. WaDIQaM and related BIQA models report performance across LIVE, TID2013, KADID-10k, KonIQ-10k, and SPAQ, highlighting the difficulty of transferring from synthetic to authentic distortions and from laboratory settings to in-the-wild images (Bosse et al., 2018). DBCNN explicitly addresses cross-regime generalisation by training separate streams for synthetic and authentic distortions and evaluating robustness across multiple datasets (Zhang et al., 2020). MetaIQA uses meta-learning to improve NR-IQA cross-dataset generalisation and is widely used as a strong baseline for cross-dataset evaluation on KonIQ-10k, SPAQ, and other benchmarks (Zhu et al., 2020). Survey papers on deep IQA conclude that cross-dataset generalisation remains a major challenge, particularly when distortion compositions and MOS scales differ (Yang et al., 2023). Sonthalia et al. show that rank-axes for attributes like age and aesthetics can transfer across datasets (Sonthalia et al., 2025), and our work adopts a similar lens for image quality: we fit a quality rank-axis on one IQA dataset and measure its transfer to other datasets, providing a representational perspective on cross-dataset generalisation that complements standard SRCC/PLCC metrics.