QUALITY HAS A DIRECTION: LEARNING A SHARED QUALITY AXIS FROM REASONING FOR IMAGE QUALITY ASSESSMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement-learning-based vision–language models generate quality reasoning that provides effective supervision for generalized image quality assessment, yet the resulting text embeddings entangle perceptual quality with scene content and linguistic realization. We investigate whether transferable quality can be localized within this entangled space. We find that MOS-conditioned variation in reasoning text forms a semantically coherent quality axis despite contextual and lexical ambiguity. Analyses of contrastive training dynamics and controlled domain shifts further reveal that the quality axis is shared by language and vision across datasets. Building on this finding, we propose QAxis, which introduces the shared quality axis into contrastive learning. Symmetric RoPE makes cross-modal similarity jointly depend on semantic correspondence in the complete representations and agreement along their quality coordinates, while axis-wise ordinal learning establishes a global perceptual order. Using only a 0.1B-parameter image encoder, direct axis projection achieves state-of-the-art average PLCC/SRCC of 0.811/0.796 across seven benchmarks; incorporating the complete image representation further improves the averages to 0.816/0.800.

1 INTRODUCTION

Multimodal large language models (MLLMs) are changing image quality assessment from direct score regression into language-grounded perceptual reasoning. Beyond associating images with fixed quality prompts (Wang et al., 2023; Zhang et al., 2023; Agnolucci et al., 2025), recent models describe visible degradations, relate them to image content, and express quality through rating distributions or free-form explanations (Wu et al., 2024a;b; You et al., 2025; Li et al., 2025). This development draws on increasingly capable vision–language instruction models (Alayrac et al., 2022; Li et al., 2023b; Liu et al., 2023; Bai et al., 2025) and on reinforcement learning, which has progressed from preference alignment (Stiennon et al., 2020; Ouyang et al., 2022) to outcome-driven reasoning (Shao et al., 2024; DeepSeek-AI, 2025). Applied to IQA, such training strengthens perceptual reasoning and cross-dataset generalization (Wu et al., 2025; Wen et al., 2025), but ordinarily retains the cost of online generation by a billion-parameter model. RALI (Zhao et al., 2026) shows that generated reasoning can instead supervise a compact CLIP image encoder (Radford et al., 2021), preserving much of this transfer without invoking the reasoning model at inference. As illustrated in Figure 1, reasoning therefore offers a route toward both transferable and efficient IQA.

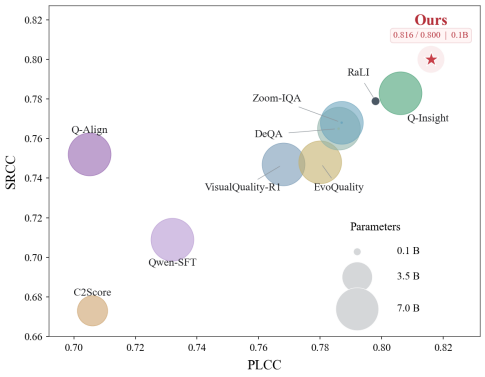


Figure 1: **Performance comparison among IQA methods in average PLCC/SRCC and parameter count.**

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.

Yet the success of reasoning supervision does not reveal what is transferred. A reasoning embedding is not itself a quality representation: it jointly encodes the depicted scene, the selected perceptual evidence, its linguistic realization, and the resulting judgment. Visual embeddings are likewise entangled, with image content often dominating degradation. Consequently, aligning complete image and reasoning representations may exploit content or instance correspondence while leaving the domain-stable structure of quality unidentified. The central question is therefore not whether reasoning contains quality information, but whether transferable quality occupies an identifiable geometry in language that can also be shared by vision across datasets.

We investigate this geometry through two questions: *Does quality form an identifiable axis in reasoning-text representations, and can the same axis be shared by visual representations across datasets?* A joint vision–language space (Radford et al., 2021) makes such a direction algebraically accessible to both modalities, but does not guarantee shared directional semantics (Liang et al., 2022). We find that MOS-conditioned variation in consensus reasoning exposes a quality axis whose two ends exhibit coherent perceptual polarity despite contextual and lexical ambiguity. More importantly, contrastive training dynamics and controlled domain-shift experiments show that this text-derived axis becomes readable from visual representations and retains a quality-selective response across authentic and synthetic distortions (Hosu et al., 2020; Lin et al., 2019). These findings localize transferable quality to a shared coordinate within the entangled representation rather than to the reasoning embedding as a whole.

This shared coordinate also exposes the limitation of ordinary image–text alignment. Instance-level correspondence can transmit the quality direction, but it neither makes quality agreement an explicit condition of matching nor imposes a global perceptual order along that direction. We therefore propose QAxis, which introduces the shared quality axis directly into contrastive learning. Its symmetric RoPE construction (Su et al., 2021) yields an alignment score jointly governed by semantic correspondence in the complete representations and agreement between their quality-axis coordinates, so semantic similarity alone cannot dominate a quality-inconsistent pair. Axis-wise ordinal learning further organizes images into a global quality order along the same coordinate. At inference, QAxis uses only the trained image encoder and takes the projection of its image feature onto the quality axis as the score. This single-coordinate method surpasses the compared SOTA methods; a PCA–ridge method over the learned representation provides a further modest improvement.

Our contributions are threefold. **First**, we identify a supervised quality axis within entangled reasoning-text representations and establish its semantic coherence under contextual language noise. **Second**, we show through training dynamics and controlled cross-domain analysis that this axis is shared by reasoning and visual representations, and remains selectively responsive to quality differences across datasets. **Third**, we introduce QAxis: symmetric RoPE makes cross-modal similarity explicitly conditional on quality-coordinate agreement, while axis-wise ordinal learning converts this local agreement into a globally ordered perceptual geometry. Extensive cross-dataset evaluation validates both this geometry and its direct-projection method.

2 Related Work

Vision–language image quality assessment. Before language supervision, transferable NR-IQA representations were pursued through self-adaptive architectures and relative ranking (Su et al., 2020; Golestaneh et al., 2022), contrastive pretraining (Madhusudana et al., 2022; Saha et al., 2023), and quality-aware pretraining (Zhao et al., 2023). These approaches reduce dependence on a single labeled dataset, but quality remains learned entirely from visual variation. Vision–language models based on CLIP (Radford et al., 2021) provide an additional source of supervision through perceptual concepts that are difficult to isolate from pixels alone. CLIP-IQA (Wang et al., 2023) scores images against handcrafted antonym prompts; LIQE (Zhang et al., 2023) combines quality levels, scene categories, and distortion types; and QualiCLIP (Agnolucci et al., 2025) adapts a joint embedding through synthetic degradations and ranking. These methods establish language as a useful quality prior, but their target structure is largely specified by a finite prompt vocabulary.

Reasoning-based image quality assessment. Instruction-tuned MLLMs (Li et al., 2023b; Liu et al., 2023; Bai et al., 2025) make it possible to express a perceptual decision as an open-ended linguistic trajectory rather than a fixed class token. Recent multimodal assessors accordingly move

from short prompts to descriptions, rating tokens, response distributions, and explicit reasoning. Q-Bench (Wu et al., 2024a) shows that low-level perception, description, and assessment are distinct capabilities. Q-Align (Wu et al., 2024b) and DeQA-Score (You et al., 2025) improve prediction through language-defined levels and distributional supervision; Q-Insight (Li et al., 2025), VisualQuality-R1 (Wu et al., 2025), and EvoQuality (Wen et al., 2025) refine quality reasoning or preference order through reinforcement and self-evolving supervision. RALI (Zhao et al., 2026) further shows that rich reasoning features can supervise a compact visual encoder through image–text alignment.

3 From Entangled Representations to a Shared Quality Axis

RALI (Zhao et al., 2026) shows that linguistic reasoning compresses high-dimensional visual evidence into a compact space with smaller cross-domain discrepancy, and that aligning a lightweight visual encoder with this space improves cross-dataset IQA. This establishes reasoning as transferable supervision, but does not reveal how quality is organized within the complete reasoning representation or whether the relevant geometry is shared with vision. We therefore ask: ***does the reasoning representation contain a direction that is strongly associated with quality?***

3.1 Constructing a Quality Axis from Reasoning

Let $\mathcal{D}_{\text{tr}} = \left\{ \left(I_i, y_i, \{T_{im}\}_{m=1}^{M_i} \right) \right\}_{i=1}^{N}$ denote the training set, where I_i, y_i, and $\{T_{im}\}_{m=1}^{M_i}$ represent the input image, its MOS label, and multiple quality-related reasoning descriptions, respectively. Consistent with RALI, these reasoning descriptions are generated offline by applying the pretrained Q-Insight model to each image. The specific inference prompt used to generate the descriptions is provided in the Appendix A. The frozen text encoder E_T of CLIP is then used to obtain their consensus embedding:

$$t_i = \frac{1}{M_i} \sum_{m=1}^{M_i} E_T(T_{im}) \in \mathbb{R}^d. \tag{1}$$

We partition the training-score range into H equal-width slices. With overall mean μ, slice means μ_h, slice proportions p_h, and regularized text covariance C, sliced inverse regression (SIR) defines

$$B = \sum_h p_h(\mu_h - \mu)(\mu_h - \mu)^\top, \qquad Bu = \lambda C u. \tag{2}$$

The leading generalized eigenvector maximizes $u^\top B u / u^\top C u$ and therefore captures MOS-conditioned displacement in the text representation. We set $\|u\|_2 = 1$. Because the slices are defined by MOS, u is a supervised *quality axis*.

Figure 2 provides a diagnostic of this construction. A single description already yields a MOS-related projection, while PLCC and SRCC increase as additional descriptions are averaged; with 24 descriptions, PLCC reaches 0.930.

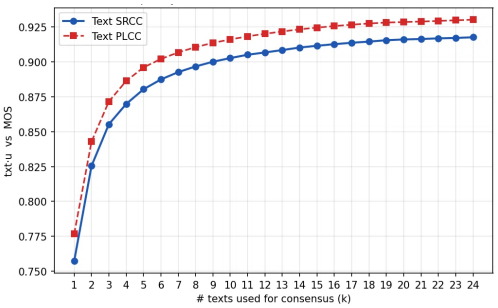


Figure 2: **Correlation between MOS and $t_i^\top u$** as the number of reasoning descriptions increases.

3.2 From Semantic Evidence to Cross-Modal Transfer

This section examines the two additional properties required for transfer: whether this association has coherent quality semantics, and whether the same axis is accessible to visual features.

Q1: Does the quality axis encode coherent semantics? We extract frequent unigrams and bigrams automatically from over 7K KonIQ training descriptions after tokenization and stop-word

removal. We then divide the corpus into the top and bottom 30% by MOS and assign each phrase a data-derived valence

$$r(w) = \log \frac{p(w \mid \text{high MOS})}{p(w \mid \text{low MOS})}. \tag{3}$$

Positive values indicate association with high-quality descriptions and negative values indicate association with low-quality descriptions. Independently, the frozen CLIP text encoder maps each isolated phrase to $E_T(w)$, and $\cos(E_T(w), u)$ measures its position along the quality axis.

Figure 3 relates corpus valence to axis semantics. In the left plot, phrases with $|r(w)| > 0.5$ place "clear," "detailed," and "enhances" near $+u$, and "blurry," "grainy," and "obscures" near $-u$. The middle plot compares $r(w)$ with $\cos(E_T(w), u)$: most phrases agree, while the upper-left and lower-right regions identify conflicts between corpus frequency and isolated-phrase semantics. The right plot traces these conflicts to context. Hedges such as "somewhat," "affects," and "lacks" frequently qualify negative judgments, whereas distortion terms can occur in high-MOS descriptions through negation, as in "no apparent distortion" or "no notable overexposure." The quality axis therefore preserves a coherent quality polarity even when contextual usage makes word frequency ambiguous.

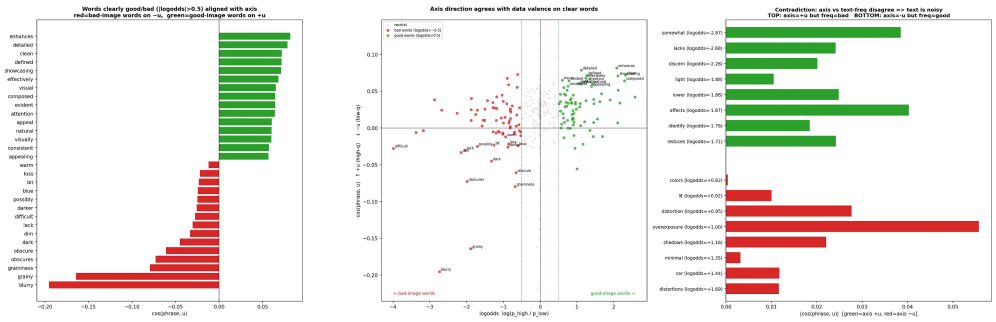


Figure 3: **Semantic probing of the quality axis.** Left: clear-valence phrases align with $\pm u$. Middle: corpus-derived valence vs. quality-axis projection. Right: contradictions trace to hedging and negation, e.g., "distortion" appears in favorable descriptions through "no distortion."

Q2: Is the quality axis shared by contrastive alignment? RALI replaces online reasoning with the CLIP alignment, using compact reasoning representations as contrastive targets for a lightweight image encoder (Zhao et al., 2026). We examine whether this alignment preserves the quality axis identified above. For normalized image and reasoning embeddings $\bar{f}_i$ and $\bar{t}_j$, define

$$p_{ij} = \frac{\exp(\bar{f}_i^\top \bar{t}_j/\tau)}{\sum_k \exp(\bar{f}_i^\top \bar{t}_k/\tau)}, \qquad \ell_i = -\log p_{ii}. \tag{4}$$

InfoNCE simultaneously draws the image toward the correct quality-reasoning representation and pushes it away from the probability-weighted centroid of the negative texts that the model currently confuses with the positive. If the quality axis is shared across modalities, this alignment should leave two observable signatures. **First**, the InfoNCE gradient should be more concentrated along the quality axis u than along an arbitrary one-dimensional direction. Let g_i denote the image-side gradient for sample i. We quantify this concentration by

$$E_u = d\frac{\sum_i (u^\top g_i)^2}{\sum_i \|g_i\|_2^2}, \tag{5}$$

where d is the feature dimension. If gradient energy is distributed isotropically across directions, then $E_u = 1$. Consequently, $E_u > 1$ indicates that ordinary InfoNCE allocates more gradient energy to the quality axis than expected for a random one-dimensional direction. **Second**, the image coordinate $q_i^I = \bar{f}_i^\top u$ should become increasingly consistent with perceptual quality during alignment.

Figure 4(a) confirms both signatures on KonIQ. The blue curve shows that E_u remains between approximately 10 and 20, far above the isotropic reference $E_u = 1$; standard InfoNCE therefore concentrates gradient energy along the quality axis. Meanwhile, the red curve shows that the SRCC between q^I and MOS increases from 0.413 to 0.872, making the text-derived axis progressively

readable from image representations. Figure 4(b) tests whether this response transfers to KADID while retaining the KonIQ-derived axis and KonIQ-trained checkpoints. KADID samples form three controlled groups: random content and distortion protocol (gray), fixed reference with varying distortion type and level (green), and fixed distortion protocol with varying reference content (red). The gray and green curves consistently exceed the red curve across checkpoints. Thus, on an unseen synthetic-distortion distribution, distortion variation preferentially activates the quality axis, whereas content variation under a fixed distortion protocol produces a weaker response. Together, the source-domain dynamics and target-domain controls show that the quality axis is shared across datasets and modalities, with selective sensitivity to quality differences.

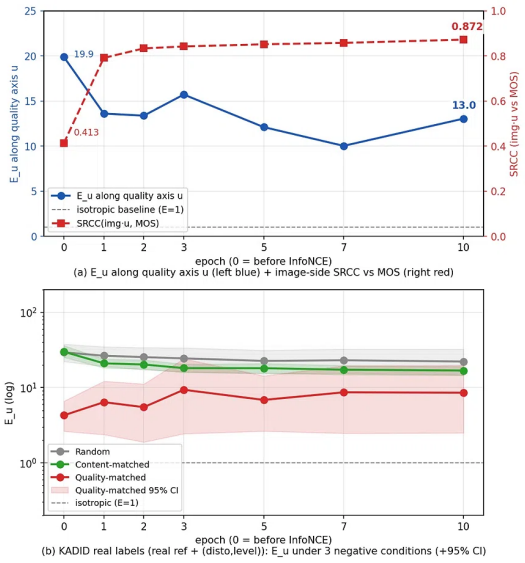


Figure 4: **Does standard InfoNCE share the quality axis?** KonIQ training dynamics and controlled KADID gradient responses.

4 QAxis: Learning with the Quality Axis

Having estimated u from reasoning text, we use it as a fixed reference for visual learning. Section 3 shows that ordinary contrastive alignment can transfer this direction to the image encoder, but content-related variation can dominate the image representation and obscure changes along the quality axis. Therefore, we propose QAxis, the pipeline is shown in Figure 5.

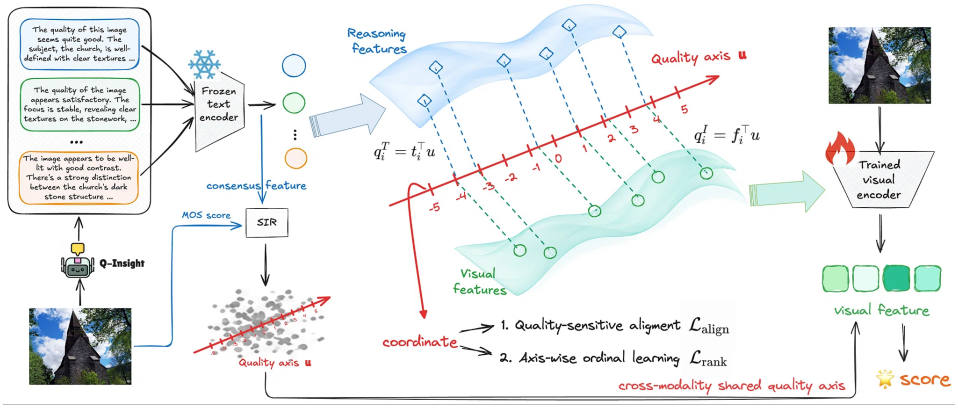


Figure 5: **The pipeline of the QAxis method.**

4.1 Quality-Sensitive Image–Text Alignment

Let $f_i = E_I(I_i) \in \mathbb{R}^{768}$ be the image representation and $t_i \in \mathbb{R}^{768}$ the consensus reasoning representation. Their coordinates on the shared axis are $q_i^I = f_i^\top u, q_i^T = t_i^\top u$.

The image vector contains both quality and content variation, and content-related variation can interfere with quality learning during InfoNCE alignment. We address this interference by making changes in the axis projection directly alter the geometry used for contrastive comparison. Specifically, we use RoPE to convert each quality coordinate into a rotation angle, $\theta(q) = \frac{\pi}{4}\tanh\left(\frac{q}{\sigma_T}\right)$. Here, σ_T is the standard deviation of text quality coordinates. RoPE partitions each 768-D feature

into 384 adjacent 2-D groups and applies

$$R(\theta) = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}. \tag{6}$$

The same quality-dependent angle is used for all 384 groups. A single rotation, however, introduces orientation-dependent cross terms into the inner product. We cancel these terms by pairing rotations of opposite signs; the derivation is provided in Appendix B. For a 768-D feature x, QAxis constructs the 1536-D symmetric representation

$$\Phi(x, q) = \frac{[R(\theta(q))x; R(-\theta(q))x]}{\sqrt{2}\,\|x\|_2}. \tag{7}$$

Eq. 7 is applied to both modalities: image features are rotated by $\pm\theta(q_i^I)$, text features by $\pm\theta(q_i^T)$.

Proposition 1 (Symmetric quality factorization)**.** *For nonzero $x, z \in \mathbb{R}^d$ and equal rotation frequency across feature pairs,*

$$\Phi(x, q_x)^\top \Phi(z, q_z) = \underbrace{\cos(x, z)}_{\textit{semantic similarity}} \underbrace{\cos\big(\theta(q_x) - \theta(q_z)\big)}_{\textit{quality-difference modulation}}. \tag{8}$$

Eq. 8 decomposes the transformed similarity into two interpretable factors. The semantic similarity factor preserves the relation carried by the complete image and text features. The quality-difference modulation factor responds to their angular displacement on the shared axis: greater angular disagreement more strongly attenuates a positive semantic match. A detailed proof is provided in Appendix B.

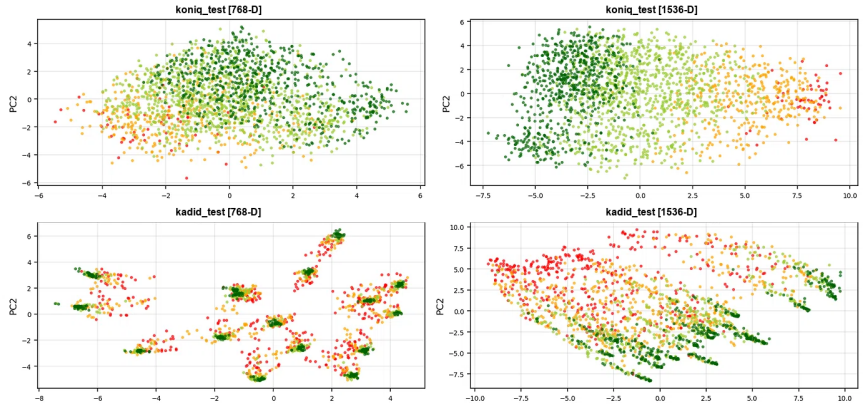


Figure 6: **Geometry induced by quality-sensitive alignment.** Left: 768-D image-encoder outputs under ordinary InfoNCE alignment. Right: the corresponding 1536-D symmetric representations. Colors denote ground-truth MOS bins, red → green for low → high quality.

Effect of quality-sensitive alignment. Figure 6 shows how symmetric RoPE reorganizes image features around quality. On KonIQ (Hosu et al., 2020), it sharpens the MOS gradient within a content-diverse manifold. KADID (Lin et al., 2019) provides a stricter diagnosis: because multiple distortions are generated from the same reference images, ordinary InfoNCE features form pronounced content clusters, exposing content as a competing source of cross-modal covariance. Symmetric rotation establishes a more coherent quality progression within and across clusters.

Then, we apply the bidirectional InfoNCE objective used by CLIP (Radford et al., 2021) to the transformed image & text representations. With $S_{ij} = \tau\Phi(f_i, q_i^I)^\top\Phi(t_j, q_j^T)$, the alignment loss is

$$\mathcal{L}_{\text{align}} = -\frac{1}{2B}\sum_{i=1}^{B}\left[\log\frac{e^{S_{ii}}}{\sum_j e^{S_{ij}}} + \log\frac{e^{S_{ii}}}{\sum_j e^{S_{ji}}}\right]. \tag{9}$$

Optimizing the contrastive objective trains the image encoder to distinguish pairs using quality-axis variation in addition to content semantics, yielding more discriminative quality alignment.

4.2 AXIS-WISE ORDINAL LEARNING

Quality-sensitive alignment increases the influence of quality variation when distinguishing image–text pairs, but pair discrimination alone remains compatible with many globally inconsistent arrangements of images. Relative ranking has been used to encode perceptual order in NR-IQA (Golestaneh et al., 2022; Shin et al., 2024). In our method, the ranking constraint is imposed directly on the image projections onto the shared quality axis. For minibatch MOS values $\{y_i\}_{i=1}^{B}$, let $\sigma_y = \mathrm{std}(y) + \epsilon_0$ and retain the ordered pairs $\mathcal{P} = \{(i,j) : |y_i - y_j|/\sigma_y > \epsilon_r\}$. The objective is

$$\mathcal{L}_{\mathrm{rank}} = \frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} \left[-\operatorname{sign}(y_i - y_j) \frac{q_i^I - q_j^I}{\sigma_y} \right]_+ . \tag{10}$$

Here $[\cdot]_+ = \max(\cdot, 0)$. The threshold ϵ_r removes near-tied MOS pairs, while ϵ_0 ensures numerical stability. This zero-margin objective penalizes reversed order without imposing a prescribed distance between correctly ordered coordinates. The final training objective is

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{align}} + \mathcal{L}_{\mathrm{rank}}. \tag{11}$$

5 EXPERIMENTS

5.1 EXPERIMENTAL SETTINGS

Inference methods. At inference, we use the trained image encoder and use its 768-D output $f = E_I(I)$. We evaluate two methods of the learned representation. ***1) QAxis-v1 (axis projection):*** uses the scalar quality coordinate $q = f^\top u$. Let $q_{\min}$ and $q_{\max}$ denote the minimum and maximum coordinates on the KonIQ training set. Its predicted score is $\hat{y}_{\mathrm{v1}} = \mathrm{clip}\Big(1 + 4\frac{q - q_{\min}}{\max(q_{\max} - q_{\min}, 10^{-6})}, 1, 5\Big)$. ***2) QAxis-v2 (PCA–ridge):*** We first extract the 768-D features by the trained vision encoder. For each training image, the first K principal coordinates are $z_i = P_K(f_i - \mu)$. Then, we fit a linear ridge method to the centered training MOS:

$$w_K = \arg\min_w \sum_i \left[(y_i - \bar{y}) - w^\top z_i \right]^2 + \alpha \|w\|_2^2. \tag{12}$$

At test time, an image is passed through the same frozen encoder and fixed PCA transform, and its quality score is $\hat{y}_{\mathrm{v2}} = w_K^\top P_K(f - \mu)$. QAxis-v1 is the primary axis-only method, whereas QAxis-v2 quantifies quality information distributed across the complete learned representation.

Implementation details. We use CLIP ViT-L/14-336 (Radford et al., 2021) and the Q-Insight (Li et al., 2025) descriptions. The text encoder is frozen; the visual encoder and projection layer are trained for 10 epochs with AdamW (learning rate 10^{-5}, weight decay 0.01, per-device batch size 40). For each image, all 24 descriptions are encoded independently and their raw text features are averaged without per-description normalization. The quality axis is estimated from the same raw consensus features using weighted SIR with 40 equal-width MOS slices. For axis-wise ordinal learning, we set the near-tie threshold to $\epsilon_r = 0.1$ in units of minibatch MOS standard deviation and the numerical constant to $\epsilon_0 = 10^{-6}$. For QAxis-v2, we retain $K = 512$ principal components and set the ridge coefficient to $\alpha = 10$. The reported 0.1B parameters count only the image encoder used at inference.

5.2 COMPARISON EXPERIMENTS

Comparison with state of the art. Table 1 compares both QAxis methods with 17 prior methods. **1)** Relative to the reasoning-transfer baseline RALI, QAxis-v1 improves the seven-dataset average PLCC/SRCC from 0.798/0.779 to 0.811/0.796, with the largest gain on PIPAL (0.527/0.528 to 0.577/0.579). Notably, unlike RALI, QAxis-v1 does not post-process the aligned visual vectors; instead, it directly computes the scalar projection of the aligned visual representation onto the quality axis. Our method is both simple and effective.**2)** The direct axis method remains competitive with substantially more elaborate IQA models while using only one projected coordinate at inference, and QAxis-v2 further raises the average to 0.816/0.800. **3)** The gains are especially clear on the synthetic KADID and CSIQ datasets, where QAxis-v2 reaches 0.751/0.773 and 0.850/0.827,

Table 1: **Cross-dataset PLCC and SRCC for models trained on KonIQ.** Avg. is the arithmetic mean over the seven datasets. Best and second-best values are shown in **bold** and underline.

Method	Avg.	KonIQ	AGIQA	CSIQ	SPAQ	KADID	LIVEW	PIPAL
PLCC								
Handcrafted metrics								
NIQE (Mittal et al., 2013)	0.521	0.533	0.560	0.718	0.679	0.468	0.493	0.195
BRISQUE (Mittal et al., 2012)	0.436	0.225	0.541	0.740	0.490	0.429	0.361	0.267
Multimodal and reasoning-based models								
C2Score (Zhu et al., 2024)	0.706	0.923	0.777	0.735	0.867	0.500	0.786	0.354
Qwen-SFT (Bai et al., 2025)	0.732	0.889	0.813	0.674	0.874	0.668	0.734	0.473
Q-Align (Wu et al., 2024b)	0.705	0.941	0.772	0.671	0.886	0.674	0.853	0.403
DeQA (You et al., 2025)	0.786	**0.953**	0.809	0.787	0.895	0.694	0.892	0.472
Q-Insight (Li et al., 2025)	0.806	0.933	0.811	**0.870**	**0.907**	0.742	0.893	0.486
VisualQuality-R1 (Wu et al., 2025)	0.768	0.923	0.822	0.712	0.891	0.712	0.874	0.441
EvoQuality (Wen et al., 2025)	0.780	0.835	**0.831**	0.816	0.903	0.743	0.847	0.482
Zoom-IQA (Liang et al., 2026)	0.787	0.938	0.816	0.797	0.902	0.701	0.887	0.468
Conventional deep NR-IQA models								
NIMA (Talebi & Milanfar, 2018)	0.697	0.896	0.715	0.695	0.838	0.532	0.814	0.390
HyperIQA (Su et al., 2020)	0.693	0.917	0.702	0.752	0.791	0.506	0.772	0.410
DBCNN (Zhang et al., 2020)	0.667	0.884	0.730	0.586	0.812	0.497	0.773	0.384
MUSIQ (Ke et al., 2021)	0.726	0.924	0.722	0.771	0.868	0.575	0.789	0.431
CLIP-IQA+ (Wang et al., 2023)	0.742	0.909	0.736	0.772	0.866	0.653	0.832	0.427
MANIQA (Yang et al., 2022)	0.681	0.849	0.723	0.623	0.768	0.499	0.849	0.457
RALI (Zhao et al., 2026)	0.798	0.939	0.779	0.828	0.897	0.723	**0.896**	0.527
QAxis-v1	0.811	0.947	0.790	0.839	0.901	0.738	0.882	0.577
QAxis-v2	**0.816**	0.949	0.787	0.850	0.901	**0.751**	0.887	**0.587**
SRCC								
Handcrafted metrics								
NIQE (Mittal et al., 2013)	0.481	0.530	0.533	0.628	0.664	0.405	0.449	0.161
BRISQUE (Mittal et al., 2012)	0.369	0.226	0.497	0.556	0.406	0.356	0.313	0.232
Multimodal and reasoning-based models								
C2Score (Zhu et al., 2024)	0.673	0.910	0.671	0.705	0.860	0.453	0.772	0.342
Qwen-SFT (Bai et al., 2025)	0.709	0.866	0.739	0.650	0.875	0.663	0.728	0.442
Q-Align (Wu et al., 2024b)	0.752	0.940	0.735	0.737	0.887	0.684	0.860	0.419
DeQA (You et al., 2025)	0.765	**0.941**	0.729	0.744	0.896	0.687	0.879	0.478
Q-Insight (Li et al., 2025)	0.783	0.916	0.764	0.824	**0.905**	0.736	0.865	0.474
VisualQuality-R1 (Wu et al., 2025)	0.747	0.908	0.767	0.662	0.892	0.711	0.849	0.438
EvoQuality (Wen et al., 2025)	0.748	0.791	**0.771**	0.760	0.900	0.749	0.813	0.454
Zoom-IQA (Liang et al., 2026)	0.768	0.922	0.765	0.754	0.900	0.700	0.870	0.465
Conventional deep NR-IQA models								
NIMA (Talebi & Milanfar, 2018)	0.675	0.859	0.654	0.649	0.856	0.535	0.771	0.399
HyperIQA (Su et al., 2020)	0.667	0.906	0.640	0.717	0.788	0.468	0.749	0.403
DBCNN (Zhang et al., 2020)	0.645	0.875	0.641	0.572	0.806	0.484	0.755	0.381
MUSIQ (Ke et al., 2021)	0.707	0.929	0.630	0.710	0.863	0.556	0.830	0.431
CLIP-IQA+ (Wang et al., 2023)	0.720	0.895	0.685	0.719	0.864	0.654	0.805	0.419
MANIQA (Yang et al., 2022)	0.658	0.834	0.636	0.627	0.758	0.465	0.832	0.452
RALI (Zhao et al., 2026)	0.779	0.922	0.715	0.788	0.897	0.725	**0.876**	0.528
QAxis-v1	0.796	0.932	0.723	0.815	0.898	0.759	0.862	0.579
QAxis-v2	**0.800**	0.934	0.723	**0.827**	0.899	**0.773**	0.860	**0.587**

respectively. Because their distortions vary under controlled reference content, these results are consistent with the representation becoming more responsive to quality variation rather than scene identity.

5.3 Ablation and Representation Analysis

Table 2: **Component analysis (PLCC/SRCC).** One uses a single description and ordinary image–text alignment; Cons. replaces it with the 24-description consensus; Rot. applies symmetric quality-conditioned rotation; Rank adds MOS ordering along the axis; Ridge replaces direct projection with PCA–ridge. Rows 1–4 and 6 use direct projection onto the quality axis, whereas Rows 5 and 7 use the PCA–ridge method. Best and second-best results are shown in **bold** and underline.

One	Cons.	Rot.	Rank	Ridge	Avg.	KonIQ	AGIQA	CSIQ	SPAQ	KADID	LIVEW	PIPAL
✓					0.719/0.711	0.871/0.873	0.776/0.716	0.715/0.691	0.834/0.841	0.711/0.711	0.764/0.775	0.361/0.370
	✓				0.738/0.728	0.881/0.864	0.783/0.729	0.732/0.712	0.862/0.861	0.701/0.698	0.777/0.789	0.428/0.441
	✓	✓			0.776/0.760	0.912/0.890	**0.792/0.732**	0.791/0.758	0.890/0.887	0.712/0.718	0.838/0.819	0.500/0.517
	✓		✓		0.784/0.765	0.936/0.918	0.765/0.699	0.846/0.815	0.891/0.889	0.695/0.704	0.838/0.816	0.517/0.515
	✓		✓	✓	0.798/0.779	0.947/0.930	0.783/0.710	0.835/0.803	0.899/0.897	0.698/0.713	0.874/0.847	0.551/0.551
	✓	✓	✓		0.811/0.796	0.947/0.932	0.790/0.723	0.839/0.815	**0.901**/0.898	0.738/0.759	0.882/**0.862**	0.577/0.579
	✓	✓	✓	✓	**0.816/0.800**	**0.949/0.934**	0.787/0.723	**0.850/0.827**	**0.901/0.899**	**0.751/0.773**	**0.887**/0.860	**0.587/0.587**

Contribution of each component. Table 2 shows Consensus alone raises the average PLCC/SRCC from 0.719/0.711 to 0.738/0.728. Adding quality-conditioned rotation further im-

Table 3: **Description-count sensitivity (PLCC/SRCC).** N is the number of reasoning descriptions used to form the consensus text feature. Best results within each method are shown in **bold**.

N |	Avg.	KonIQ	AGIQA	CSIQ	SPAQ	KADID	LIVEW	PIPAL
QAxis-v1: direct quality-axis projection								
4	0.793/0.779	0.945/0.930	**0.795/0.725**	0.779/0.774	0.897/0.894	0.722/0.737	0.876/0.861	0.542/0.539
8	0.802/0.788	0.942/0.927	0.785/0.724	0.818/0.796	0.900/**0.898**	0.720/0.740	0.882/0.864	0.567/0.570
16	0.810/0.795	0.946/0.930	0.786/0.714	0.816/0.795	0.900/**0.898**	**0.742/0.763**	**0.892/0.875**	**0.588/0.588**
24	**0.811/0.796**	**0.947/0.932**	0.790/0.723	**0.839/0.815**	**0.901/0.898**	0.738/0.759	0.882/0.862	0.577/0.579
QAxis-v2: PCA–ridge method								
4	0.805/0.790	0.944/0.929	**0.807/0.732**	0.795/0.786	0.900/**0.899**	0.737/0.760	0.887/0.865	0.566/0.557
8	0.805/0.791	0.945/0.930	0.786/0.723	0.818/0.794	**0.901/0.899**	0.723/0.748	0.889/0.866	0.577/0.577
16	0.812/0.796	0.947/0.931	0.784/0.714	0.813/0.790	0.900/0.897	0.747/0.769	**0.897/0.874**	**0.595/0.593**
24	**0.816/0.800**	**0.949/0.934**	0.787/0.723	**0.850/0.827**	**0.901/0.899**	**0.751/0.773**	0.887/0.860	0.587/0.587

proves performance to 0.776/0.760. This indicates that, the quality-axis-based construction already provides clear gains. w/ quality-axis-based rank improves the average performance to 0.784/0.765. Their combination reaches 0.811/0.796 and improves every dataset over consensus alignment. In particular, on PIPAL, the metrics improve from 0.428/0.441 to 0.577/0.579. Finally, adding the PCA–Ridge method layer further improves the average PLCC/SRCC to 0.816/0.800.

Effect of description count. Table 3 shows a clear overall improvement as more descriptions are aggregated, with the average performance peaking at $N = 24$.

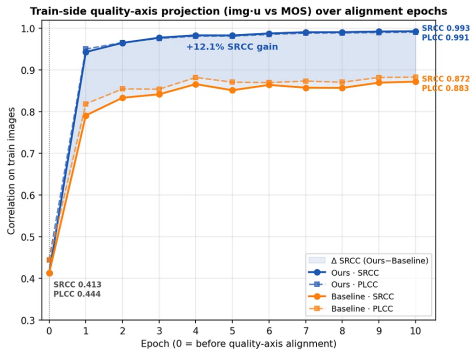


Figure 7: **Training dynamics of image-side axis readability on the Koniq-train dataset.** Correlation between q^I and MOS for QAxis (blue) and standard alignment (orange).

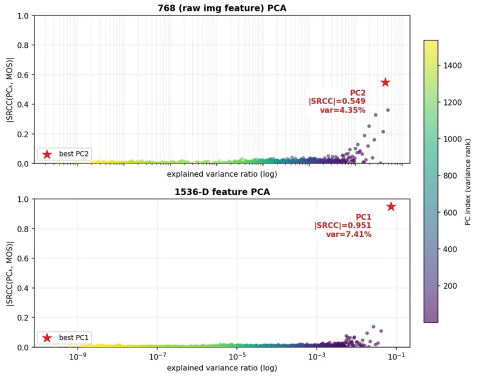


Figure 8: **Quality signal across principal components on the Koniq-train dataset.** Raw 768-D image features place the strongest quality signal in PC2; the 1536-D symmetric representation moves it to PC1.

QAxis makes quality more readable during training. Figure 7 compares QAxis with standard image–text alignment from the same initialization (PLCC/SRCC $= 0.444/0.413$). After ten epochs, standard alignment reaches $0.883/0.872$, whereas QAxis improves from the first epoch and converges to $0.991/0.993$, an SRCC gain of 0.121. Quality-sensitive alignment and axis-wise ordering therefore make the shared axis substantially more readable from visual representations.

Quality becomes a leading direction of variation. Figure 8 provides a representation-level view. In the raw features, quality is moderately associated with PC2; after the symmetric transformation, PC1 becomes both the most quality-correlated component and a higher-variance direction. Thus, QAxis makes quality more accessible to low-dimensional projection.

6 Conclusion

We investigate a text-derived quality axis as a common reference for visual learning and IQA method. Language diagnostics reveal quality-related semantics together with context-dependent lexical associations, while cross-dataset analysis motivates organizing visual features by quality rather than relying solely on instance correspondence. Symmetric quality-conditioned alignment and axis-wise ranking make this reference explicit during training. The resulting fixed-axis method achieves competitive performance in the current seven-benchmark evaluation.

AI USE STATEMENT

Generative AI tools were used for literature discovery, and language editing, and to provide feedback on the presentation and verification of mathematical derivations and experimental findings. The authors independently developed the method, conducted all experiments, and generated the reported measurements. All AI-assisted content was reviewed and verified by the authors, who take full responsibility for the final manuscript.

ETHICS STATEMENT

This work uses publicly available image-quality benchmarks and their existing annotations, and introduces no new data collection or human-subject study.

REPRODUCIBILITY STATEMENT

Sections 3 and 4 specify quality-axis estimation, symmetric alignment, and axis-wise ordinal learning; Section 5.1 reports the inference methods, optimization settings, and hyperparameters. Appendix A documents the datasets, evaluation protocol, and compared methods, while Appendix B provides the assumptions and complete proof of the symmetric quality factorization. An anonymized supplementary package will include the training configurations, axis-estimation and evaluation code, and scripts used to produce the reported tables and figures.

REFERENCES

Lorenzo Agnolucci, Leonardo Galteri, and Marco Bertini. Quality-aware image-text alignment for opinion-unaware image quality assessment. *arXiv preprint arXiv:2403.11176*, 2025. URL `https://arxiv.org/abs/2403.11176`. Version 3.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: A visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Yuming Fang, Hanwei Zhu, Yan Zeng, Kede Ma, and Zhou Wang. Perceptual quality assessment of smartphone photography. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 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.

S. Alireza Golestaneh, Saba Dadsetan, and Kris M. Kitani. No-reference image quality assessment via transformers, relative ranking, and self-consistency. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3209–3218, 2022.

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

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

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

Eric C. Larson and Damon M. Chandler. Most apparent distortion: Full-reference image quality assessment and the role of strategy. *Journal of Electronic Imaging*, 19(1), 2010.

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. *arXiv preprint arXiv:2306.04717*, 2023a.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, pp. 19730–19742, 2023b.

Weiqi Li, Xuanyu Zhang, Shijie Zhao, Yabin Zhang, Junlin Li, Li Zhang, and Jian Zhang. Q-Insight: Understanding image quality via visual reinforcement learning. *arXiv preprint arXiv:2503.22679*, 2025. URL https://arxiv.org/abs/2503.22679.

Guoqiang Liang, Jianyi Wang, Zhonghua Wu, Shangchen Zhou, and Chen Change Loy. Zoom-IQA: Image quality assessment with reliable region-aware reasoning. *arXiv preprint arXiv:2601.02918*, 2026.

Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y. Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

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.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Pavan C. Madhusudana, Neil Birkbeck, Yilin Wang, Balu Adsumilli, and Alan C. Bovik. Image quality assessment using contrastive learning. *IEEE Transactions on Image Processing*, 31:4149–4161, 2022.

Anish Mittal, Anush Krishna Moorthy, and Alan C. 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.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8748–8763, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

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

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Nyeong-Ho Shin, Seon-Ho Lee, and Chang-Su Kim. Blind image quality assessment based on geometric order learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12799–12808, 2024.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F. Christiano. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*, 2021. URL https://arxiv.org/abs/2104.09864.

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

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

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

Wen Wen, Tianwu Zhi, Kanglong Fan, Yang Li, Xinge Peng, Yabin Zhang, Yiting Liao, Junlin Li, and Li Zhang. Self-evolving vision-language models for image quality assessment via voting and ranking. *arXiv preprint arXiv:2509.25787*, 2025.

Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, and Weisi Lin. Q-Bench: A benchmark for general-purpose foundation models on low-level vision. In *International Conference on Learning Representations*, 2024a.

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*, 2024b.

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*, 2025.

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*, 2022.

Zhiyuan You, Xin Cai, Jinjin Gu, Tianfan Xue, and Chao Dong. Teaching large language models to regress accurate image quality scores using score distribution. *arXiv preprint arXiv:2501.11561*, 2025.

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*, pp. 14071–14081, 2023.

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

Shijie Zhao, Xuanyu Zhang, Weiqi Li, Junlin Li, Li Zhang, Tianfan Xue, and Jian Zhang. Reasoning as representation: Rethinking visual reinforcement learning in image quality assessment. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2510.11369.

Hanwei Zhu, Haoning Wu, Yixuan Li, Zicheng Zhang, Baoliang Chen, Lingyu Zhu, Yuming Fang, Guangtao Zhai, Weisi Lin, and Shiqi Wang. Adaptive image quality assessment via teaching large multimodal model to compare. In *Advances in Neural Information Processing Systems*, 2024.

A Inference Prompt for Quality Reasoning Generation

Following RALI, we use the pretrained Q-Insight model to generate quality-related reasoning descriptions for each training image. The prompts used during inference are summarized in Table **??**.

Table 4: **Prompts for quality reasoning generation.**

System Prompt for Inference: A conversation between User and Assistant. The user asks a question, and the Assistant solves it. The assistant first thinks about the reasoning process in the mind and then provides the user with the answer. The reasoning process and answer are enclosed within `<think> </think>` and `<answer> </answer>` tags, respectively, i.e., `<think> reasoning process here </think><answer> answer here </answer>`.
Prompt for Quality Score Regression: What is your overall rating on the quality of this picture? The rating should be a float between 1 and 5, rounded to two decimal places, with 1 representing very poor quality and 5 representing excellent quality. Return the final answer in JSON format with the following keys: `"rating"`: The score.

For each image I_i, Q-Insight generates a quality-related reasoning description T_{im} together with its corresponding quality score. Multiple descriptions are generated for the same image and are then used to construct the consensus text embedding in Eq. 1.

B Proof and Properties of the Symmetric Quality Factorization

This appendix proves Proposition 1 and states the conditions under which its interpretation is valid. The derivation is included because a single rotary branch does *not* yield the factorization in Eq. 8; the cancellation follows specifically from pairing rotations with opposite angles.

B.1 Definitions and rotation identities

Assume $d = 2m$ and let

$$J_2 = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}, \qquad J = \mathrm{diag}(J_2, \ldots, J_2) \in \mathbb{R}^{d\times d}. \tag{13}$$

The matrix J satisfies $J^\top = -J$, $J^2 = -I$, and $J^\top J = I$. Applying the same planar rotation to all m adjacent feature pairs is therefore equivalent to

$$R(\alpha) = \exp(\alpha J) = \cos\alpha\, I + \sin\alpha\, J. \tag{14}$$

These identities imply

$$R(\alpha)^\top = R(-\alpha), \qquad R(\alpha)R(\beta) = R(\alpha+\beta), \qquad R(\alpha)^\top R(\alpha) = I. \tag{15}$$

Thus, $R(\alpha)$ is orthogonal and preserves both inner products and Euclidean norms when the same angle is applied to two vectors.

For a nonzero feature x and scalar quality coordinate q_x, write $a = \theta(q_x)$ and define the unnormalized symmetric lift

$$\widetilde{\Phi}(x, a) = \begin{bmatrix} R(a)x \\ R(-a)x \end{bmatrix}. \tag{16}$$

Eq. 7 is precisely $\Phi(x, q_x) = \widetilde{\Phi}(x, a)/(\sqrt{2}\,\|x\|_2)$.

B.2 Key cancellation lemma

Proposition 2 (Opposite rotations cancel orientation-dependent terms)**.** *For $x, z \in \mathbb{R}^d$ and angles $a, b \in \mathbb{R}$,*

$$\widetilde{\Phi}(x, a)^\top \widetilde{\Phi}(z, b) = 2\cos(b-a)\, x^\top z. \tag{17}$$

Proof. Using Eq. 15, the two branches give

$$\begin{aligned}\widetilde{\Phi}(x,a)^\top \widetilde{\Phi}(z,b) &= x^\top R(a)^\top R(b) z + x^\top R(-a)^\top R(-b) z \\ &= x^\top \big[R(b-a) + R(a-b)\big] z. \end{aligned} \tag{18}$$

Let $\delta = b - a$. Expanding both matrices by Eq. 14 yields

$$\begin{aligned} R(\delta) + R(-\delta) &= (\cos\delta\, I + \sin\delta\, J) + (\cos\delta\, I - \sin\delta\, J) \\ &= 2\cos\delta\, I. \end{aligned} \tag{19}$$

The terms involving J encode the orientation-dependent cross products within each two-dimensional feature pair. They enter the two branches with opposite signs and cancel exactly. Substitution into Eq. 18 proves Eq. 17. □

This cancellation can also be seen pair by pair. For $x_k = (x_{k1}, x_{k2})^\top$ and $z_k = (z_{k1}, z_{k2})^\top$, define

$$d_k = x_{k1} z_{k1} + x_{k2} z_{k2}, \qquad c_k = x_{k1} z_{k2} - x_{k2} z_{k1}. \tag{20}$$

The positive-angle branch contributes $d_k \cos\delta - c_k \sin\delta$, whereas the negative-angle branch contributes $d_k \cos\delta + c_k \sin\delta$. Their sum is $2 d_k \cos\delta$. Summing over $k = 1, \ldots, m$ recovers Eq. 17. This expansion makes explicit why either branch alone retains a term unrelated to quality-coordinate agreement.

B.3 Proof of Proposition 1

Orthogonality of $R(a)$ gives

$$\left\|\widetilde{\Phi}(x,a)\right\|_2^2 = \|R(a)x\|_2^2 + \|R(-a)x\|_2^2 = 2\,\|x\|_2^2\,, \tag{21}$$

$$\left\|\widetilde{\Phi}(z,b)\right\|_2^2 = 2\,\|z\|_2^2\,. \tag{22}$$

Combining Eqs. 17 and 22, and using $\cos(b-a) = \cos(a-b)$, gives

$$\begin{aligned}\Phi(x,q_x)^\top \Phi(z,q_z) &= \frac{2\cos(b-a) x^\top z}{\sqrt{2}\,\|x\|_2\,\sqrt{2}\,\|z\|_2} \\ &= \frac{x^\top z}{\|x\|_2\,\|z\|_2} \cos\big(\theta(q_x) - \theta(q_z)\big) \\ &= \cos(x,z)\cos\big(\theta(q_x) - \theta(q_z)\big), \end{aligned} \tag{23}$$

which is Eq. 8.

B.4 Interpretation and boundary conditions

The factorization has several immediate consequences. First, if $q_x = q_z$, then $a = b$ and the lifted cosine equals the original feature cosine. The symmetric lift therefore leaves the similarity unchanged for perfectly matched quality coordinates. Second, for the map $\theta(q) = c\tanh(q/\tau)$, oddness of $\tanh$ gives $\theta(-q) = -\theta(q)$, while its bounded range prevents arbitrarily large rotations. This bounded nonlinear map should not be interpreted as making the angular difference linear in $q_x - q_z$.

If $c \le \pi/4$, then $|a - b| < 2c \le \pi/2$ for finite coordinates and hence

$$0 < \cos(a-b) \le 1, \tag{24}$$

with zero attainable only as a limiting boundary value when $c = \pi/4$.

C Evaluation Protocol and Compared Methods

Datasets. We train on the official KonIQ-10k (Hosu et al., 2020) split and evaluate without target-set fine-tuning on seven benchmarks: authentic KonIQ (Hosu et al., 2020), SPAQ (Fang et al.,

2020), and LIVEW (Ghadiyaram & Bovik, 2016); synthetic KADID-10k (Lin et al., 2019) and CSIQ (Larson & Chandler, 2010); AI-generated AGIQA-3K (Li et al., 2023a); and restoration-oriented PIPAL (Gu et al., 2020). KonIQ is the source-domain test set and the other six measure zero-shot transfer. We report PLCC/SRCC after aligning datasets with reversed score orientation; Avg. is the unweighted seven-dataset mean.

Competing methods. Under the KonIQ-trained cross-dataset protocol used by RALI (Zhao et al., 2026) and EvoQuality (Wen et al., 2025), we compare with handcrafted NIQE (Mittal et al., 2013) and BRISQUE (Mittal et al., 2012); deep NR-IQA models NIMA (Talebi & Milanfar, 2018), HyperIQA (Su et al., 2020), DBCNN (Zhang et al., 2020), MUSIQ (Ke et al., 2021), CLIP-IQA+ (Wang et al., 2023), and MANIQA (Yang et al., 2022); and multimodal or reasoning-based C2Score (Zhu et al., 2024), Qwen-SFT (Bai et al., 2025), Q-Align (Wu et al., 2024b), DeQA-Score (You et al., 2025), Q-Insight (Li et al., 2025), VisualQuality-R1 (Wu et al., 2025), EvoQuality (Wen et al., 2025), Zoom-IQA (Liang et al., 2026), and RALI (Zhao et al., 2026). Qwen-SFT denotes the supervised Qwen2.5-VL baseline reported by RALI rather than a separate IQA architecture. Competing numbers are transcribed from the corresponding benchmark tables under this protocol.