# DIAGNOSIS-SCORE DECOUPLING FOR CROSS-DATASET IMAGE QUALITY ASSESSMENT

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Vision-language models (VLMs) can describe image degradations, yet their quality predictions remain sensitive to dataset-specific scoring conventions and distortion distributions. It remains unclear whether chain-of-thought (CoT) diagnostic descriptions provide transferable information beyond a scalar quality prediction. This paper introduces Diagnosis-Score Decoupling (DSD), a two-branch framework for CoT-based image quality assessment (CoT-IQA) that explicitly separates diagnostic representation from quality scoring. One branch predicts a scalar quality prior, while the other generates structured CoT diagnostics describing degradation type, severity, and spatial extent. Rather than using the diagnostic branch's final score, DSD parses its descriptions into a compact, score-free representation. A lightweight calibrator then combines this representation with the scalar prior using source-domain quality labels and is applied unchanged to target datasets without target-side labels, tuning, or recalibration. Across all six evaluated transfers among controlled-distortion datasets, DSD improves SRCC over the scalar-prior baseline (MOS only) on the primary backbone, with a similar overall trend on a second VLM backbone. Ablation studies indicate that the gains arise primarily from image-aligned diagnostic information, particularly degradation severity. On KonIQ-10k authentic distortions, however, these gains do not persist, suggesting that the benefit of source-calibrated diagnostic evidence depends on the target distortion regime.

## 1 INTRODUCTION

Blind image quality assessment (BIQA) predicts perceptual image quality without access to a pristine reference. The field has progressed from natural-scene statistics and learned local features to content-adaptive, attention-based, and multi-scale Transformer models (Moorthy and Bovik, 2011; Saad et al., 2012; Ye et al., 2012; Su et al., 2020; Ke et al., 2021). Yet performance is usually established near the training distribution. Changes in distortion type, image content, or subjective scoring convention can alter the relationship between visual evidence and quality scores learned on one dataset (Zhang et al., 2021; Li et al., 2024), making cross-dataset generalization a persistent challenge for practical IQA.

Vision-language models (VLMs) provide a different perspective. Beyond producing a single quality score, they can describe degradation type, severity, and location in natural language (Wu et al., 2024b;c; You et al., 2024; Chen et al., 2026). Existing studies use text prompts, quality levels, image comparisons, ranking, and reasoning to improve quality prediction (Wang et al., 2023; Wu et al., 2024a; Zhu et al., 2024; Wen et al., 2026), but their primary objective remains the final numerical judgment. Whether diagnostic descriptions contain transferable quality information independent of the generated score remains insufficiently studied, especially under cross-dataset transfer.

This question involves two related processes. *Perceptual diagnosis* identifies what quality problems are present, how severe they are, and where they occur; *numerical calibration* maps that evidence to a dataset-specific score. Thus, "severe local blur" may retain a similar perceptual meaning across datasets while receiving different numerical scores. Cross-dataset failure may therefore arise from the diagnosis itself or from the diagnosis-to-score mapping, which direct scoring couples in one output.

Motivated by this distinction, we study Diagnosis-Score Decoupling (DSD) through the CoT-IQA framework. One branch produces an independent scalar quality prior, while a second branch generates structured diagnostic evidence. The diagnostic branch's final quality score is explicitly removed, and



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.

a fixed rule-based parser converts the remaining descriptions into a compact score-free representation. A lightweight ridge regressor, fitted using source-domain quality labels only, learns how to combine the scalar prior and diagnostic representation. During target-domain inference, both VLM branches, the parser, and the source-fitted calibrator remain fixed. Our experiments focus on the transfer value of structured diagnostic evidence rather than on whether chain-of-thought prompting is necessary compared with direct structured degradation prediction.

Across the evaluated controlled-distortion datasets, DSD improves over the scalar prior on the primary backbone, with a similar aggregate trend on a second VLM backbone. Matched controls indicate that most of the gain comes from image-aligned diagnostic information, especially degradation severity, rather than scalar recalibration or simply increasing the representation dimensionality. The benefit does not persist on KonIQ-10k's authentic distortions, where source-learned calibration produces negative transfer. Our contributions are threefold:

- We frame cross-dataset IQA as a question of evaluating perceptual diagnosis separately from numerical calibration, distinguishing diagnostic utility from calibration transferability.
- Matched controls and ablations show that gains in the evaluated controlled-distortion transfers arise mainly from image-aligned diagnostic information, especially severity, rather than scalar recalibration or additional dimensions alone.
- The authentic-distortion experiment on KonIQ-10k shows that both Diagnosis-only and full DSD underperform the scalar prior, revealing a transfer boundary for source-calibrated diagnosis beyond the evaluated controlled-distortion datasets.

## 2 Related Work

### 2.1 Blind Image Quality Assessment

Early BIQA methods model departures from natural-scene statistics. DIIVINE and BLIINDS-II use wavelet- and DCT-domain statistics (Moorthy and Bovik, 2011; Saad et al., 2012); BRISQUE learns a mapping from spatial statistics to subjective quality (Mittal et al., 2012); and NIQE measures deviation from a natural-image model without opinion scores (Mittal et al., 2013). Deep BIQA progressed from patch-wise CNN prediction (Kang et al., 2014) to ranking, content-adaptive, and patch-to-picture architectures (Liu et al., 2017; Su et al., 2020; Ying et al., 2020). Transformer and attention-based methods such as MUSIQ and DEIQT further model global, multi-scale, and region-dependent quality cues (Ke et al., 2021; Qin et al., 2023).

Cross-dataset performance nevertheless remains sensitive to changes in distortion, content, and score distributions. Re-IQA and TOPIQ explore transferable representations and semantic guidance (Saha et al., 2023; Chen et al., 2024b), while CONTRIQUE and QPT learn reusable quality representations through contrastive or quality-aware pre-training (Madhusudana et al., 2022; Zhao et al., 2023). Continual BIQA explicitly studies sequences of dataset shifts (Zhang et al., 2023b). MetaIQA learns priors across quality tasks (Zhu et al., 2020); PromptIQA conditions on image–score examples (Chen et al., 2024a); QCN models relative quality order (Shin et al., 2024); and test-time or distortion-guided adaptation uses unlabeled target images (Roy et al., 2023; Li et al., 2024). These settings differ in the amount of target information available. We instead study a strict source-only setting in which calibration is learned from source-domain quality labels and transferred without target-side fitting or adaptation.

### 2.2 Vision-Language Models for Image Quality Assessment

Vision-language models extend IQA beyond scalar prediction to attribute description, comparison, and natural-language explanation. Building on CLIP's transferable image–text representation (Radford et al., 2021), CLIP-IQA uses text prompts to express quality attributes (Wang et al., 2023), while LIQE jointly models image quality, scene content, and distortion (Zhang et al., 2023a). Q-Align represents quality with text-defined levels (Wu et al., 2024a), DeQA-Score learns distributions over quality scores (You et al., 2025), and Compare2Score infers continuous scores from pairwise comparisons (Zhu et al., 2024). Their primary objective remains the final quality estimate.

A complementary line of work uses language to expose quality evidence. Q-Bench and Q-Instruct study low-level perception and description (Wu et al., 2024b;c); DepictQA generates detailed quality descriptions (You et al., 2024); and Grounding-IQA associates descriptions with localized regions (Chen et al., 2026). Recent work also incorporates degradation-aware VLM priors into BIQA (Fu et al., 2026) and extends quality assessment from human preference to machine and embodied-agent usability (Li et al., 2025a; 2026). These studies show that natural-language outputs can encode degradation type, severity, and local defects beyond a scalar score, but they do not isolate whether such information remains useful after the score itself is removed.

### 2.3 Reasoning and Intermediate Representations for Quality Assessment

Recent work improves VLM-based quality judgments through reasoning, comparison, and ranking. Q-Insight uses visual reinforcement learning (Li et al., 2025b), VisualQuality-R1 optimizes reasoning-based assessment with a ranking objective (Wu et al., 2025), PreResQ-R1 separates response and preference objectives (Feng et al., 2026), and EvoQuality derives supervision from pairwise voting (Wen et al., 2026). These methods demonstrate the value of intermediate reasoning signals, but they do not directly test whether diagnosis and numerical scoring have different cross-dataset transfer properties. CoT-IQA addresses this distinction by removing the diagnostic branch's generated score, retaining a structured score-free representation, and evaluating the scalar prior, diagnosis, and their source-calibrated combination separately.

## 3 Method

### 3.1 Problem Setting and Framework Overview

Let the labeled source dataset and target dataset be

$$\mathcal{D}_s = \{(x_i^s, y_i^s)\}_{i=1}^{N_s}, \qquad \mathcal{D}_t = \{(x_j^t, y_j^t)\}_{j=1}^{N_t}. \tag{1}$$

We consider strict source-only cross-dataset prediction: source labels fit calibration, while target labels are withheld from calibration, selection, tuning, and inference and used only for final evaluation.

CoT-IQA distinguishes perceptual diagnosis from numerical calibration. Given an image $x$, two separately adapted VLM branches produce an independent scalar quality prior and a structured diagnostic response:

$$s = g_{\mathrm{mos}}(x), \qquad t = g_{\mathrm{diag}}(x), \qquad s \in [1, 5]. \tag{2}$$

The branches generate separate outputs; *independent* does not imply statistical independence. A score-removal operation $S$ excludes the diagnostic branch's final score, and a fixed parser $\phi$ extracts diagnostic evidence:

$$\mathbf{d} = \phi(S(t)), \qquad \mathbf{z} = [s, \mathbf{d}], \qquad \hat{y} = f_\theta(\mathbf{z}). \tag{3}$$

The diagnostic score enters neither $\mathbf{d}$ nor $\mathbf{z}$. DSD therefore combines one scalar prior with score-free diagnostic evidence rather than combining two generated quality scores.

### 3.2 Independent Quality Prior and Structured Diagnosis

The MOS branch directly predicts overall image quality and serves both as the scalar prior and the MOS-only baseline. It does not consume the diagnostic branch's text. The diagnostic branch used in the main DSD experiments follows the verified three-part response structure

$$t = (t_{\mathrm{loc}}, t_{\mathrm{attr}}, t_{\mathrm{score}}). \tag{4}$$

Localization records degraded regions, type, severity, and local/global extent; attribution summarizes the detected impairments; and the final field contains a numerical quality score and short rationale. The branches are adapted separately. In DSD, $t_{\mathrm{score}}$ is retained only for the direct-score control and excluded from calibration features.

### 3.3 Score-Free Diagnostic Representation

Passing the diagnostic branch's generated score to calibration would create a direct score path and obscure the contribution of diagnostic evidence. We therefore parse only the retained diagnostic

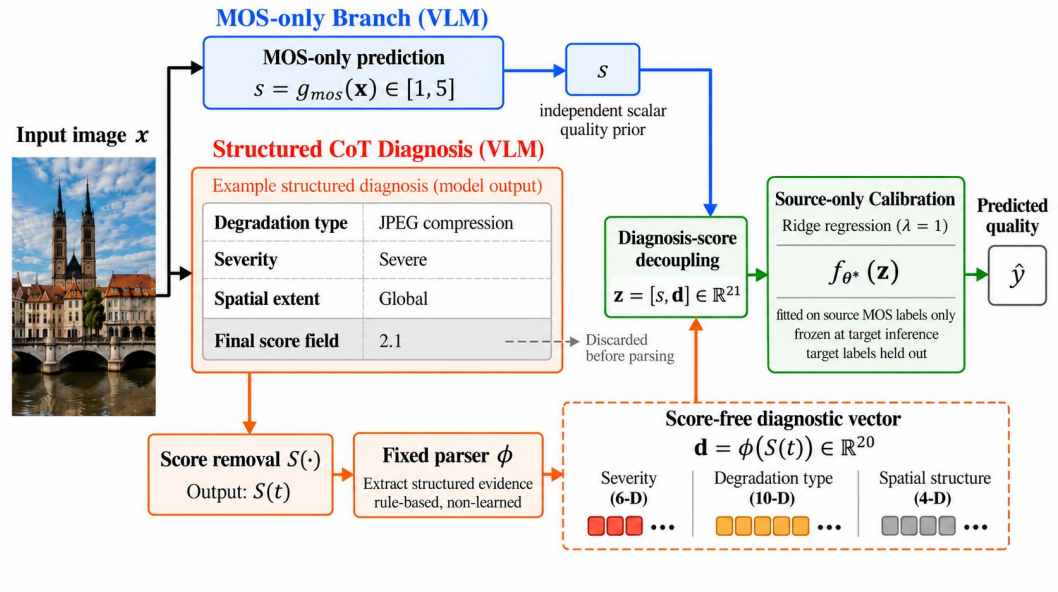


Figure 1: CoT-IQA overview. The diagnostic score is removed before a fixed parser forms a 20-D score-free vector; source-only Ridge combines it with an independent scalar prior.

attributes into

$$\mathbf{d} = [\mathbf{d}_{\text{sev}}, \mathbf{d}_{\text{type}}, \mathbf{d}_{\text{spatial}}] \in \mathbb{R}^{6+10+4} = \mathbb{R}^{20}. \tag{5}$$

The six severity features are the maximum, mean, and minimum recognized severity and the counts of mild, moderate, and severe regions; mild/moderate/severe are encoded as levels 1/2/3 before aggregation. Ten binary type indicators represent blocking, color shift, defocus blur, Gaussian blur, Gaussian noise, impulse noise, JPEG compression, motion blur, ringing, and a sharpness-related attribute. Four spatial features record the number of detected regions, the number of distinct degradation types, and global and local indicators. Bounding-box coordinates themselves are not used.

The parser is fixed and parameter-free; missing or unrecognized fields receive zero values. Ground-truth MOS, dataset distortion labels, reference identities, and file paths are excluded, while image identifiers only align branch outputs. Concatenating $s$ and $\mathbf{d}$ gives $\mathbf{z} \in \mathbb{R}^{21}$. *Score-free* refers only to the downstream representation, not to the absence of quality-related supervision during branch adaptation.

### 3.4 Source-Only Calibration and Target-Domain Inference

For each source sample, we construct $\mathbf{z}_i^s = [s_i^s, \mathbf{d}_i^s]$ and fit ridge regression:

$$(\mathbf{w}_s, b_s) = \arg\min_{\mathbf{w},b} \sum_{i=1}^{N_s} (y_i^s - b - \mathbf{w}^\top \mathbf{z}_i^s)^2 + \lambda \|\mathbf{w}\|_2^2. \tag{6}$$

All main experiments use $\lambda = 1$, raw unstandardized features, and an unregularized intercept. Ridge learns 21 weights and one intercept; its low capacity provides a controlled test of diagnostic feature utility rather than a separate methodological contribution.

For target image $x_j^t$, the source-fitted parameters remain fixed:

$$\hat{y}_j^t = \text{clip}_{[1,5]}\left(b_s + \mathbf{w}_s^\top \left[g_{\text{mos}}(x_j^t), \phi\left(S\left(g_{\text{diag}}(x_j^t)\right)\right)\right]\right). \tag{7}$$

Both branches, score removal, parsing, and calibration remain fixed at target inference; no target-dependent standardization, score remapping, tuning, or test-time adaptation is used. Possible benchmark exposure during foundation-model pretraining cannot be independently ruled out.

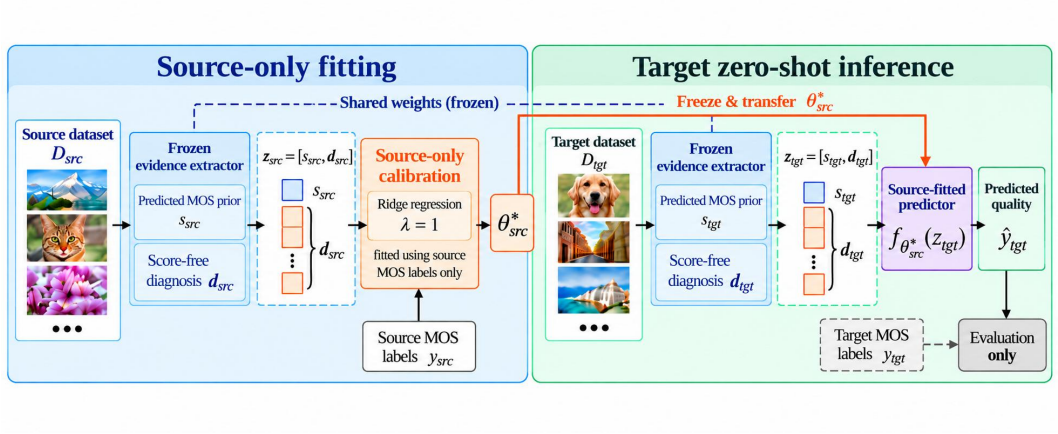


Figure 2: Strict source-only transfer. Frozen branches extract the scalar prior and score-free diagnosis in both domains. Ridge calibration is fitted on source MOS labels and transferred unchanged; target labels are used only for final evaluation.

**Intuition for calibration transfer.** For squared error, calibration helps only when the source-learned change to the scalar prior aligns sufficiently with its target-domain residual to offset the change's magnitude. This change includes prior reweighting, diagnostic contributions, the intercept, and clipping. This provides intuition for why source-fitted calibration may improve some transfers but degrade others; it is not a generalization bound and does not imply improved SRCC or PLCC.

# 4 Experiments

## 4.1 Experimental Setup

Our experiments investigate three questions: whether score-free diagnostic information supports cross-dataset prediction, which components contribute to the improvement, and whether source-fitted calibration remains effective under larger distribution shifts.

**Datasets and protocol.** We evaluate all six directed transfers among LIVE (Sheikh et al., 2006), CSIQ (Larson and Chandler, 2010), and TID2013 (Ponomarenko et al., 2015), which contain 779, 866, and 3,000 distorted images, respectively. For each direction, calibration is fitted on the source dataset and applied unchanged to the target dataset. Authentic-distortion evaluation uses the official 2,015-image KonIQ-10k test split (Hosu et al., 2020), with calibration fitted separately on each controlled-distortion source. Target labels are used only for final evaluation and remain unavailable during calibration, selection, and inference.

**Metrics.** SRCC measures rank-order agreement between predicted and ground-truth quality scores and is the primary metric; PLCC measures linear correlation, and RMSE measures absolute prediction error. We use a fixed archived quality-label conversion to $[1, 5]$, with higher values indicating better quality. Dataset-specific label normalization is fixed during preprocessing before transfer evaluation; no prediction-dependent remapping, target-fitted nonlinear mapping, or target-based hyperparameter tuning is used. Directional means give equal weight to each transfer.

**Implementation.** The primary backbone is Qwen2-VL-7B (Wang et al., 2024); InternVL3-8B (Zhu et al., 2025) provides a second-backbone evaluation. The Qwen2-VL branches are separately adapted for three epochs with LoRA (Hu et al., 2022) (rank 64, scaling 128) on the same 15,000 synthetic images generated from 40 KADID-10k pristine references (Lin et al., 2019) (13,500 train/1,500 validation). The scalar target is a direct MOS string. The diagnostic prompt asks the model to localize degraded regions, identify degradation type/severity, and provide an overall score; its supervision follows the three-part Localization/Attribution/Overall Quality Score format. Pseudo-MOS is generated deterministically from synthetic degradation severity and covered area, starting

from 5 and clipped to $[1, 5]$, rather than from human ratings. LIVE, CSIQ, TID2013, and KonIQ-10k are excluded from branch adaptation, after which both branches are frozen. Diagnosis-only uses the 20-D representation; full DSD adds the scalar prior. *Source-only* refers to direction-specific calibration and target inference, not to prior VLM adaptation.

### 4.2 Cross-Dataset Transfer Results

Table 1 compares four configurations under the same source-only protocol. Diagnosis-only improves SRCC over MOS-only in all six controlled-distortion transfers, while full DSD further raises the mean from 0.5819 to 0.5958. Full DSD exceeds MOS-only in SRCC for every direction (Figure 3); Diagnosis-only is slightly higher in PLCC on TID2013→LIVE.

Table 1: Qwen2-VL cross-dataset transfer (SRCC / PLCC). Diagnosis-only removes the diagnostic branch's generated score; all regressors are fitted only on the listed source. Bold marks the best metric.

| Source→Target | MOS-only | Scalar-prior calibration | Diagnosis-only | Full DSD |
|---|---|---|---|---|
| LIVE→CSIQ | 0.6175/0.5940 | 0.6175/0.5940 | 0.6976/0.7582 | **0.7217/0.7688** |
| LIVE→TID2013 | 0.5149/0.5212 | 0.5149/0.5212 | 0.5782/0.6646 | **0.5880/0.6709** |
| CSIQ→LIVE | 0.3344/0.3193 | 0.3345/0.3381 | 0.4118/0.4581 | **0.4150/0.4657** |
| CSIQ→TID2013 | 0.5149/0.5212 | 0.5142/0.5283 | 0.6474/0.6896 | **0.6659/0.7116** |
| TID2013→LIVE | 0.3344/0.3193 | 0.3344/0.3193 | 0.4032/**0.4699** | **0.4069**/0.4691 |
| TID2013→CSIQ | 0.6175/0.5940 | 0.6175/0.5940 | 0.7535/0.7809 | **0.7773/0.7979** |
| Mean | 0.4889/0.4782 | 0.4888/0.4825 | 0.5819/0.6369 | **0.5958/0.6473** |

**Statistical uncertainty.** Paired image-level bootstrap 95% intervals for DSD-minus-MOS-only SRCC are positive in all six directions, from $[0.0354, 0.1071]$ on TID2013→LIVE to $[0.1291, 0.1921]$ on TID2013→CSIQ. They preserve predictor pairing but do not capture reference-level dependence or retraining variability.

**Regularization sensitivity.** Across $\lambda \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 100, 1000\}$, mean SRCC ranges from 0.5944 to 0.6208 versus 0.5958 at the fixed main setting $\lambda = 1$. Because the gains are transfer-dependent, we retain the pre-specified $\lambda = 1$ rather than select it from target performance.

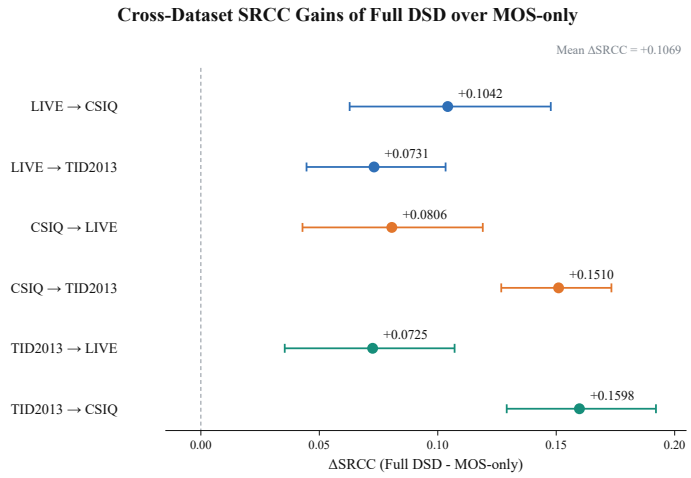


Figure 3: Full-DSD SRCC gains over MOS-only with paired image-level bootstrap 95% confidence intervals. All intervals are above zero for the evaluated controlled-distortion transfers.

## 4.3 Analysis of Diagnostic Evidence

### 4.3.1 Contribution of Diagnostic Components

Table 2 separates scalar calibration from diagnostic evidence. Scalar-prior calibration leaves mean SRCC essentially unchanged (0.4888), whereas MOS + severity reaches 0.5805; type-only and spatial-only additions yield 0.4889 and 0.4894. Full DSD reaches 0.5958 and exceeds the severity-only configuration in four of six directions. Severity therefore accounts for most of the observed gain under the evaluated representation and linear calibration, while other features provide additional information in some transfers; this does not imply general perceptual importance.

Severity remains a quality-adjacent diagnostic attribute because its ordinal levels encode degradation strength; score-free therefore excludes the generated final quality score rather than all information correlated with quality ordering.

Table 2: Diagnostic ablations averaged over six controlled-distortion transfers. ΔSRCC is relative to MOS-only; all learned variants use source-only ridge calibration.

| Configuration | SRCC↑ | PLCC↑ | RMSE↓ | ΔSRCC |
|---|---|---|---|---|
| MOS-only | 0.4889 | 0.4782 | 1.7016 | 0.0000 |
| Scalar-prior calibration | 0.4888 | 0.4825 | 0.9586 | −0.0001 |
| Diagnosis-only | 0.5819 | 0.6369 | 0.8031 | +0.0930 |
| MOS + severity | 0.5805 | 0.6343 | 0.8371 | +0.0916 |
| MOS + type | 0.4889 | 0.5008 | 0.9404 | +0.0000 |
| MOS + spatial | 0.4894 | 0.4964 | 0.9420 | +0.0005 |
| MOS + shuffled diagnosis | 0.4501 | 0.4546 | 0.9672 | −0.0388 |
| MOS + random features | 0.4394 | 0.4448 | 0.9650 | −0.0495 |
| Full DSD | **0.5958** | **0.6473** | 0.8211 | **+0.1069** |

### 4.3.2 Image-Aligned Diagnostic Evidence

In the archived negative controls, the shuffling condition permutes the 20-D diagnosis vectors across image identities, whereas the random-feature condition replaces diagnosis with a dimension-matched random vector. Their mean SRCCs fall to 0.4501 and 0.4394, respectively, below the scalar prior. These controls support an image-alignment effect rather than a benefit from dimensionality alone; they are interpreted as fixed controls rather than repeated randomization estimates.

### 4.3.3 Separation from Generated Scores

The diagnostic branch's final score alone yields mean SRCC 0.4274, below Diagnosis-only and full DSD. The explicit generated score therefore cannot substitute for the structured representation, although this control does not remove the effect of quality-score supervision used during earlier branch adaptation. Full DSD has the highest mean SRCC/PLCC, whereas Diagnosis-only has lower RMSE (0.8031 versus 0.8211), revealing a correlation–absolute-error trade-off.

**Descriptive diagnostic sanity check.** Four annotators familiar with IQA (one author and three peers) jointly reviewed 120 outputs, stratified across LIVE, CSIQ, TID2013, and KADID-10k (30 each; seed not retained). They saw the image, diagnosis, and distortion label, but not the predicted score. *Exact* denotes a matching/equivalent primary type consistent with the artifact; *Partial*, the correct broad category with an imprecise subtype or inaccurate additions; and *Error*, disagreement with both label and artifact. Exact/Partial/Error rates were 61.67/38.33/0%; severity was visually consistent for all samples (100%), and all responses were structurally parseable (100%). No inter-rater reliability statistic was computed. This descriptive audit is not an independent blinded human study and is not used as primary evidence for performance claims. Because KADID reference content was used during adaptation, it is also not an independent zero-shot evaluation.

Figure 4 shows two successful corrections and one failure under CSIQ→LIVE, with absolute-error changes of $-3.13$, $-1.89$, and $+2.48$. In the first case, DSD substantially corrects an underestimated JPEG-compressed sample; in the second, it moves a blurred sample close to the subjective MOS; in the third, it over-corrects a moderately compressed sample and increases error. These cases illustrate

how source-fitted diagnosis can either help or hurt individual predictions; aggregate claims rely on the transfer results rather than on these examples.

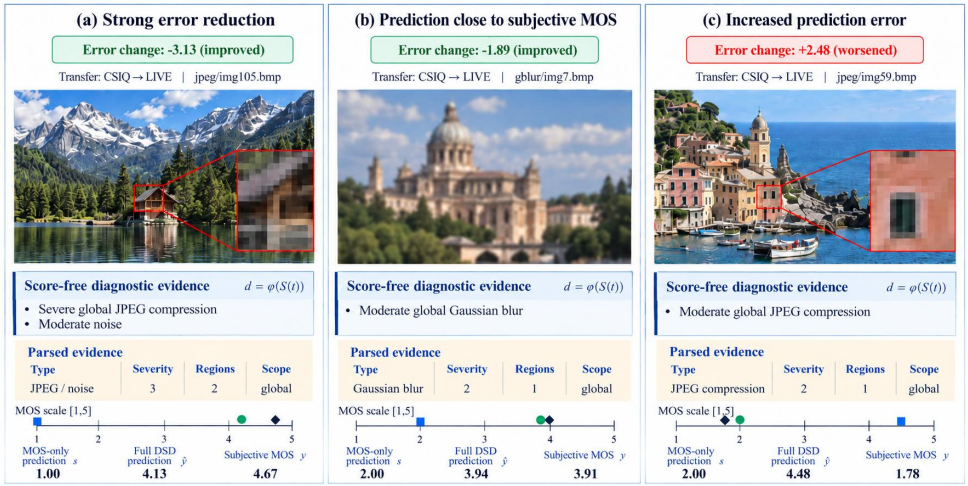


Figure 4: Qualitative CSIQ→LIVE cases: `jpeg/img105.bmp`, `gblur/img7.bmp`, and `jpeg/img59.bmp`. Error change is $|\hat{y} - y| - |s - y|$; negative values indicate improvement.

### 4.4 Generalization across Vision-Language Backbones

Table 3 compares two InternVL3-8B adapters sharing the MOS branch, 20-D parser, and source-only Ridge ($\lambda = 1$). DSD-A uses the Phase-1 three-step diagnostic adapter; DSD-B continues it with Phase-2 SFT and a six-step instruction adding explicit diagnosis, restoration suggestion, and expert routing. Adapter and instruction therefore change jointly, so this is not a reasoning-length ablation. Both improve five of six transfers (mean gains 0.0659/0.0558) but decline on LIVE→CSIQ.

Table 3: InternVL3-8B transfer. DSD-A/B use Phase-1/Phase-2 diagnostic adapters but share the MOS branch, 20-D parser, and source-only Ridge calibrator; deltas are relative to MOS-only.

| Source→Target | MOS-only | DSD-A | ΔSRCC | DSD-B | ΔSRCC |
|---|---|---|---|---|---|
| LIVE→CSIQ | 0.6309 | 0.5827 | −0.0482 | 0.5216 | −0.1093 |
| LIVE→TID2013 | 0.4495 | 0.5886 | +0.1391 | 0.5611 | +0.1116 |
| CSIQ→LIVE | 0.3364 | 0.3474 | +0.0110 | 0.3852 | +0.0488 |
| CSIQ→TID2013 | 0.4495 | 0.6344 | +0.1849 | 0.6081 | +0.1586 |
| TID2013→LIVE | 0.3364 | 0.3739 | +0.0375 | 0.4264 | +0.0900 |
| TID2013→CSIQ | 0.6309 | 0.7022 | +0.0713 | 0.6663 | +0.0354 |
| Mean | 0.4723 | 0.5382 | +0.0659 | 0.5281 | +0.0558 |

**Contextual VLM-IQA comparison.** Table 4 provides only contextual comparison because training differs across methods. Q-Align uses the archived `q-future/one-align` configuration, and VisualQuality-R1 uses the 7B score-only, non-thinking setting.

Table 4: Target-wise VLM-IQA context (SRCC / PLCC). Full DSD averages source-fitted calibrations; training settings differ across methods.

| Target | Q-Align | VisualQuality-R1 | MOS-only | Full DSD |
|---|---|---|---|---|
| LIVE | 0.4696/0.5103 | 0.4665/0.5141 | 0.3344/0.3193 | 0.4110/0.4674 |
| CSIQ | 0.8889/0.9141 | 0.8288/0.8709 | 0.6175/0.5940 | 0.7495/0.7834 |
| TID2013 | 0.8313/0.8529 | 0.7832/0.8211 | 0.5149/0.5212 | 0.6270/0.6913 |
| KonIQ-10k | 0.9006/0.9232 | 0.7848/0.8400 | 0.8209/0.8223 | 0.5557/0.5937 |

### 4.5 Transfer Limitation under Authentic Distortions

On the official 2,015-image KonIQ-10k test split, MOS-only reaches SRCC 0.8209. Across LIVE-, CSIQ-, and TID2013-fitted calibrators, Diagnosis-only obtains 0.2532/0.4316/0.3778 and full DSD 0.4313/0.6454/0.5903 (Table 5), yielding Diagnosis-only $<$ full DSD $<$ MOS-only for every source. Thus, the source-calibrated diagnosis-based prediction pipeline exhibits negative transfer under authentic distortions, while the scalar prior provides only partial recovery. This experiment does not identify whether the failure originates from diagnostic errors, representation coverage, or calibration mismatch.

Table 5: Transfer to 2,015 KonIQ-10k images. Diagnosis-only and full DSD use source-only ridge calibration; MOS-only is shared across sources.

| | SRCC | | | PLCC | | |
|---|---|---|---|---|---|---|
| Source | MOS-only | Diagnosis-only | Full DSD | MOS-only | Diagnosis-only | Full DSD |
| LIVE | 0.8209 | 0.2532 | 0.4313 | 0.8223 | 0.3863 | 0.4987 |
| CSIQ | 0.8209 | 0.4316 | 0.6454 | 0.8223 | 0.4609 | 0.6575 |
| TID2013 | 0.8209 | 0.3778 | 0.5903 | 0.8223 | 0.4285 | 0.6250 |
| Mean | 0.8209 | 0.3542 | 0.5557 | 0.8223 | 0.4252 | 0.5937 |

### 4.6 Limitations

Our conclusions are limited to the evaluated datasets and two VLM backbones. The fixed parser uses a predefined degradation vocabulary and may miss unseen or mixed artifacts. The 120-sample joint audit is descriptive, lacks independent inter-rater statistics, and includes KADID reference content used during adaptation. The experiments support the utility of structured score-free diagnostic evidence in the evaluated controlled-distortion transfers but do not isolate CoT generation from direct structured prediction; the InternVL3 comparison also changes adapter and instruction jointly. KonIQ-10k demonstrates negative transfer of the source-calibrated diagnosis pipeline but cannot separate diagnostic, coverage, and calibration effects. Two VLM passes also increase inference cost.

## 5 Conclusion

We studied Diagnosis-Score Decoupling through the CoT-IQA framework to evaluate perceptual diagnosis separately from numerical calibration. An independent branch provides a scalar quality prior, while a diagnostic branch describes degradation type, severity, and spatial extent. By removing the diagnostic branch's generated final score and parsing only the remaining evidence, CoT-IQA constructs a compact score-free representation whose combination with the scalar prior is calibrated using source-domain labels only.

Under strict source-only transfer among the evaluated controlled-distortion datasets, the score-free diagnosis provides complementary information for cross-dataset prediction. Diagnosis-only improves substantially over the scalar prior on average, and full DSD further improves mean SRCC and PLCC. Matched controls support an image-alignment effect beyond scalar recalibration or additional feature dimensionality alone, with severity providing the dominant contribution under the evaluated representation and linear calibration. A second InternVL3-8B backbone exhibits a similar aggregate tendency, although the magnitude and even direction of improvement remain transfer- and configuration-dependent.

The authentic-distortion experiment defines an important boundary: on KonIQ-10k, both Diagnosis-only and full DSD fall below the scalar prior, showing that the source-calibrated diagnosis pipeline is not guaranteed to transfer under larger shifts. Taken together, the results support evaluating diagnostic information separately from its numerical interpretation rather than treating a VLM's generated score as the sole quality signal. Future work should improve diagnostic reliability and coverage, develop shift-robust calibration, and test lower-cost direct structured prediction.

## AI USE STATEMENT

Generative AI tools were used for translation, language editing, manuscript-organization discussions, and editorial and methodological feedback. All research decisions, experiments, analyses, scientific claims, figures, and conclusions were independently verified and approved by the authors, who take full responsibility for the final submission.

## REFERENCES

Zewen Chen, Haina Qin, Juan Wang, Chunfeng Yuan, Bing Li, Weiming Hu, and Liang Wang. PromptIQA: Boosting the performance and generalization for no-reference image quality assessment via prompts. In *European Conference on Computer Vision*, pp. 247–264, 2024. https://www.ecva.net/papers/eccv_2024/papers_ECCV/html/75_ECCV_2024_paper.php.

Chaofeng Chen, Jiadi Mo, Jingwen Hou, Haoning Wu, Liang Liao, Wenxiu Sun, Qiong Yan, and Weisi Lin. TOPIQ: A top-down approach from semantics to distortions for image quality assessment. *IEEE Transactions on Image Processing*, 33:2404–2418, 2024. https://doi.org/10.1109/TIP.2024.3378466.

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*, 2026. https://proceedings.iclr.cc/paper_files/paper/2026/hash/42f23b6ed414b846a46c8db43509cb92-Abstract-Conference.html.

Zehui Feng, Weichuan Wang, Xiaohan Chen, and Ting Han. PreResQ-R1: Response-preference disentangled ranking-and-scoring reinforcement optimization for robust visual quality assessment. In *Conference on Empirical Methods in Natural Language Processing*, 2026. https://arxiv.org/abs/2511.05393.

Bohan Fu, Guanyi Qin, Fazhan Zhang, Zihao Huang, Mingxuan Li, and Runze Hu. DR.Experts: Differential refinement of distortion-aware experts for blind image quality assessment. In *AAAI Conference on Artificial Intelligence*, volume 40, pp. 3985–3993, 2026. https://doi.org/10.1609/aaai.v40i5.37401.

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. https://doi.org/10.1109/TIP.2020.2967829.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. https://arxiv.org/abs/2106.09685.

Le Kang, Peng Ye, Yi Li, and David Doermann. Convolutional neural networks for no-reference image quality assessment. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1733–1740, 2014. https://openaccess.thecvf.com/content_cvpr_2014/html/Kang_Convolutional_Neural_Networks_2014_CVPR_paper.html.

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. 5148–5157, 2021. https://openaccess.thecvf.com/content/ICCV2021/html/Ke_MUSIQ_Multi-Scale_Image_Quality_Transformer_ICCV_2021_paper.html.

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):011006, 2010. https://doi.org/10.1117/1.3267105.

Aobo Li, Jinjian Wu, Yongxu Liu, and Leida Li. Bridging the synthetic-to-authentic gap: Distortion-guided unsupervised domain adaptation for blind image quality assessment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28422–28431, 2024. https://openaccess.thecvf.com/content/CVPR2024/html/Li_Bridging_the_Synthetic-to-Authentic_Gap_Distortion-Guided_Unsupervised_Domain_Adaptation_for_Blind_CVPR_2024_paper.html.

Chunyi Li, Yuan Tian, Xiaoyue Ling, Zicheng Zhang, Haodong Duan, Haoning Wu, Ziheng Jia, Xiaohong Liu, Xiongkuo Min, Guo Lu, Weisi Lin, and Guangtao Zhai. Image quality assessment: From human to machine preference. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7570–7581, 2025. https://openaccess.thecvf.com/content/CVPR2025/html/Li_Image_Quality_Assessment_From_Human_to_Machine_Preference_CVPR_2025_paper.html.

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*, volume 38, 2025. https://proceedings.neurips.cc/paper_files/paper/2025/hash/3479362c2eb7321ba0490d98eb032772-Abstract-Conference.html.

Chunyi Li, Jiahao Xiao, Jianbo Zhang, Farong Wen, Zicheng Zhang, Yuan Tian, Xiangyang Zhu, Xiaohong Liu, Zhengxue Cheng, Weisi Lin, and Guangtao Zhai. Image quality assessment for embodied AI. In *International Conference on Learning Representations*, 2026. https://proceedings.iclr.cc/paper_files/paper/2026/hash/4310ae054ce265e56d8ea897971149b5-Abstract-Conference.html.

Hanhe Lin, Vlad Hosu, and Dietmar Saupe. KADID-10k: A large-scale artificially distorted IQA database. In *Eleventh International Conference on Quality of Multimedia Experience*, pp. 1–3, 2019. https://doi.org/10.1109/QoMEX.2019.8743252.

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. https://openaccess.thecvf.com/content_iccv_2017/html/Liu_RankIQA_Learning_From_ICCV_2017_paper.html.

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. https://doi.org/10.1109/TIP.2022.3181496.

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. https://doi.org/10.1109/TIP.2012.2214050.

Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. Making a completely blind image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2013. https://doi.org/10.1109/LSP.2012.2227726.

Anush Krishna Moorthy and Alan Conrad Bovik. Blind image quality assessment: From natural scene statistics to perceptual quality. *IEEE Transactions on Image Processing*, 20(12):3350–3364, 2011. https://doi.org/10.1109/TIP.2011.2147325.

Nikolay Ponomarenko, Lina Jin, Oleg Ieremeiev, Vladimir Lukin, Karen Egiazarian, Jaakko Astola, Benoit Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, and C.-C. Jay Kuo. Image database TID2013: Peculiarities, results and perspectives. *Signal Processing: Image Communication*, 30:57–77, 2015. https://doi.org/10.1016/j.image.2014.10.009.

Guanyi Qin, Runze Hu, Yutao Liu, Xiawu Zheng, Haotian Liu, Xiu Li, and Yan Zhang. Data-efficient image quality assessment with attention-panel decoder. In *AAAI Conference on Artificial Intelligence*, volume 37, pp. 2091–2100, 2023. https://doi.org/10.1609/aaai.v37i2.25302.

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 *International Conference on Machine Learning*, pp. 8748–8763, 2021. https://proceedings.mlr.press/v139/radford21a.html.

Subhadeep Roy, Shankhanil Mitra, Soma Biswas, and Rajiv Soundararajan. Test time adaptation for blind image quality assessment. In *IEEE/CVF International Conference on Computer Vision*, pp. 16742–16751, 2023. https://openaccess.thecvf.com/content/ICCV2023/html/Roy_Test_Time_Adaptation_for_Blind_Image_Quality_Assessment_ICCV_2023_paper.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. https://openaccess.thecvf.com/content/CVPR2023/html/Saha_Re-IQA_Unsupervised_Learning_for_Image_Quality_Assessment_in_the_Wild_CVPR_2023_paper.html.

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. https://doi.org/10.1109/TIP.2012.2191563.

Hamid R. Sheikh, Muhammad F. Sabir, and Alan C. Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on Image Processing*, 15(11):3440–3451, 2006. https://doi.org/10.1109/TIP.2006.881959.

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. https://openaccess.thecvf.com/content/CVPR2024/html/Shin_Blind_Image_Quality_Assessment_Based_on_Geometric_Order_Learning_CVPR_2024_paper.html.

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*, pp. 3667–3676, 2020. https://openaccess.thecvf.com/content_CVPR_2020/html/Su_Blindly_Assess_Image_Quality_in_the_Wild_Guided_by_a_CVPR_2020_paper.html.

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*, pp. 2555–2563, 2023. https://doi.org/10.1609/aaai.v37i2.25353.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and others. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv:2409.12191*, 2024. https://arxiv.org/abs/2409.12191.

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. In *International Conference on Learning Representations*, 2026. https://proceedings.iclr.cc/paper_files/paper/2026/hash/424403b4d2553f16d8ad11635a8214b3-Abstract-Conference.html.

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*, pp. 54015–54029, 2024. https://proceedings.mlr.press/v235/wu24ah.html.

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*, 2024. https://proceedings.iclr.cc/paper_files/paper/2024/hash/363d4c97bb411e4b07612915b76c06ae-Abstract-Conference.html.

Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Kaixin Xu, Chunyi Li, Jingwen Hou, Guangtao Zhai, Geng Xue, Wenxiu Sun, Qiong Yan, and Weisi Lin. Q-Instruct: Improving low-level visual abilities for multi-modality foundation models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25490–25500, 2024. https://openaccess.thecvf.com/content/CVPR2024/html/Wu_Q-Instruct_Improving_Low-level_Visual_Abilities_for_Multi-modality_Foundation_Models_CVPR_2024_paper.html.

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, 2025. https://proceedings.neurips.cc/paper_files/paper/2025/file/7f8f7bf2d357a09bdffff04b2e0f5c4e-Paper-Conference.pdf.

Peng Ye, Jayant Kumar, Le Kang, and David Doermann. Unsupervised feature learning framework for no-reference image quality assessment. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1098–1105, 2012.

Zhenqiang Ying, Haoran Niu, Praful Gupta, Dhruv Mahajan, Deepti Ghadiyaram, and Alan C. Bovik. From patches to pictures (PaQ-2-PiQ): Mapping the perceptual space of picture quality. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3575–3585, 2020. https://openaccess.thecvf.com/content_CVPR_2020/html/Ying_From_Patches_to_Pictures_PaQ-2-PiQ_Mapping_the_Perceptual_Space_of_CVPR_2020_paper.html.

Zhiyuan You, Zheyuan Li, Jinjin Gu, Zhenfei Yin, Tianfan Xue, and Chao Dong. Depicting beyond scores: Advancing image quality assessment through multi-modal language models. In *European Conference on Computer Vision*, pp. 259–276, 2024. https://www.ecva.net/papers/eccv_2024/papers_ECCV/html/6356_ECCV_2024_paper.php.

Zhiyuan You, Xin Cai, Jinjin Gu, Tianfan Xue, and Chao Dong. Teaching large language models to regress accurate image quality scores using score distribution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14483–14494, 2025. https://openaccess.thecvf.com/content/CVPR2025/html/You_Teaching_Large_Language_Models_to_Regress_Accurate_Image_Quality_Scores_CVPR_2025_paper.html.

Weixia Zhang, Kede Ma, Guangtao Zhai, and Xiaokang Yang. Uncertainty-aware blind image quality assessment in the laboratory and wild. *IEEE Transactions on Image Processing*, 30:3474–3486, 2021. https://doi.org/10.1109/TIP.2021.3061932.

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. https://openaccess.thecvf.com/content/CVPR2023/html/Zhang_Blind_Image_Quality_Assessment_via_Vision-Language_Correspondence_A_Multitask_Learning_CVPR_2023_paper.html.

Weixia Zhang, Dingquan Li, Chao Ma, Guangtao Zhai, Xiaokang Yang, and Kede Ma. Continual learning for blind image quality assessment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):2864–2878, 2023. https://doi.org/10.1109/TPAMI.2022.3178874.

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. https://openaccess.thecvf.com/content/CVPR2023/html/Zhao_Quality-Aware_Pre-Trained_Models_for_Blind_Image_Quality_Assessment_CVPR_2023_paper.html.

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. 14143–14152, 2020. https://openaccess.thecvf.com/content_CVPR_2020/html/Zhu_MetaIQA_Deep_Meta-Learning_for_No-Reference_Image_Quality_Assessment_CVPR_2020_paper.html.

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*, volume 37, 2024. https://proceedings.neurips.cc/paper_files/paper/2024/hash/393367b4ccf1d56e51276243d5cb85c4-Abstract-Conference.html.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, and others. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv:2504.10479*, 2025. https://arxiv.org/abs/2504.10479.

# SUPPLEMENTARY MATERIAL

**Scope.** This supplementary material provides implementation and reproducibility details omitted from the main paper because of space constraints. Specifically, it describes the construction of the synthetic adaptation data, the exact pseudo-MOS rule, the settings used to adapt each branch, the inference prompts, the deterministic 20-dimensional (20-D) parser, dataset-label normalization, source-only Ridge calibration, sensitivity to regularization, the human diagnostic-fidelity protocol, and the configuration of the second backbone. Unless otherwise stated, all quantitative values reported here are derived from the archived outputs associated with the main-paper experiments. No target-domain labels are used to adapt either VLM branch, fit or select the source-domain calibrator, or perform test-time adjustment.

## A SYNTHETIC ADAPTATION DATA CONSTRUCTION

### A.1 REFERENCE IMAGES AND DEGRADATION PIPELINE

The Phase-1 adaptation set comprises 15,000 synthetically degraded images generated from the first 40 pristine reference images in KADID-10k. Each reference image is resized to $512 \times 512$ before degradation. The data generator uses a random seed of 42. The synthetic distortions include Gaussian blur, Gaussian noise, and JPEG compression, each applied at three severity levels. Table 6 lists the exact parameters.

Table 6: Synthetic degradation parameters used for Phase-1 adaptation.

| Degradation | Mild | Moderate | Severe |
|---|---|---|---|
| Gaussian blur | $\sigma = 1$, kernel 5 | $\sigma = 2$, kernel 9 | $\sigma = 4$, kernel 17 |
| Gaussian noise | $\sigma = 5$ | $\sigma = 15$ | $\sigma = 30$ |
| JPEG compression | quality 50 | quality 25 | quality 10 |

The generator applies a global degradation to 30% of the samples. Of all samples, 40% contain one locally degraded patch and 30% contain two locally degraded patches. The width and height of each local patch are sampled independently from the range of 64 to 256 pixels. Local degradation masks are blended using Gaussian-smoothed boundaries to avoid artificial hard edges. For globally degraded samples, the covered-area ratio is set to 1.

### A.2 EXACT PSEUDO-MOS RULE

The pseudo-MOS is computed deterministically and is not derived from human opinion scores. For a sample with degradation instances $k = 1, \ldots, K$, let $a_k \in (0, 1]$ denote the covered-area ratio, and define the severity factor as

$$q(\ell_k) = \begin{cases} 0.5, & \ell_k = \text{Mild}, \\ 1.0, & \ell_k = \text{Moderate}, \\ 2.0, & \ell_k = \text{Severe}. \end{cases} \tag{8}$$

The penalty contributed by one degradation instance is

$$p_k = \min(2.5,\ 2a_k q(\ell_k)), \tag{9}$$

and the final pseudo-MOS is

$$y_{\text{pseudo}} = \text{round}_2\left(\text{clip}_{[1,5]}\left(5 - \sum_{k=1}^{K} p_k\right)\right). \tag{10}$$

Thus, a larger covered area or greater severity produces a larger deduction from the initial score of 5.0, while the penalty from any single degradation instance is capped at 2.5 points.

### A.3 MATCHED SUPERVISION FOR THE TWO BRANCHES

The diagnostic and MOS-only branches use the same 15,000 synthetic images and identical pseudo-MOS targets. The diagnostic target comprises a three-part response specifying (i) degradation localization, (ii) degradation attribution, and (iii) an overall quality score. The MOS-only dataset is constructed from the same Phase-1 samples by extracting each pseudo-MOS value and replacing the diagnostic target with the direct string `Predicted MOS: X.XX`. A 10% validation split yields 13,500 training samples and 1,500 validation samples.

## B BRANCH ADAPTATION AND INFERENCE

### B.1 QWEN2-VL ADAPTATION CONFIGURATION

The primary backbone is Qwen2-VL-7B-Instruct. The two Phase-1 branches are adapted separately using identical optimization settings (Table 7). LoRA is applied to all supported target modules.

Table 7: Qwen2-VL Phase-1 adaptation configuration.

| Setting | Value |
|---|---|
| Training stage | supervised fine-tuning |
| LoRA rank / $\alpha$ | 64 / 128 |
| LoRA target | all |
| Image max pixels | 262,144 |
| Sequence cutoff length | 2,048 |
| Training samples | 13,500 |
| Validation samples | 1,500 |
| Per-device batch size | 1 |
| Gradient accumulation | 16 |
| Effective batch size | 16 |
| Learning rate | $2 \times 10^{-5}$ |
| Epochs | 3 |
| Scheduler | cosine |
| Warmup ratio | 0.1 |
| Precision | bfloat16 |
| Evaluation interval | 100 steps |

### B.2 EXACT INFERENCE PROMPTS

The MOS-only branch uses

```
<image>Please assess the perceptual quality of this image
and directly predict a MOS score from 1 to 5.
```

The diagnostic branch uses

```
<image>Assess the perceptual quality of this image step by
step.
1. Locate all visibly degraded regions with bounding boxes.
2. Identify the degradation type and severity for each
region.
3. Provide an overall quality score from 1 (worst) to 5
(best).
```

At evaluation time, decoding is deterministic (`do_sample=False`). The maximum generation lengths are 128 new tokens for the MOS-only branch and 512 new tokens for the diagnostic branch. The final score generated by the diagnostic branch is parsed only for the direct-score control condition and is excluded when constructing the DSD diagnostic vector.

# C Exact Score-Free Diagnostic Representation

## C.1 Parser behavior

The parser processes only the text generated by the diagnostic branch. It recognizes a region entry when a line contains a region identifier, a bracketed bounding box, degradation text, and a severity token. Severity synonyms are mapped as follows:

$$\{\text{mild, low}\} \mapsto 1, \qquad \{\text{moderate, medium}\} \mapsto 2, \qquad \{\text{severe, high}\} \mapsto 3. \tag{11}$$

The parser first extracts recognized region-level severity values and degradation names and then scans the full diagnostic text for supported degradation keywords. Missing or unrecognized information is encoded as zero. The bounding-box coordinates themselves are not used as calibration features.

Table 8: Lexical patterns used by the fixed degradation-type parser.

| Feature | Recognized lexical patterns |
|---|---|
| Gaussian blur | `gaussian blur`, `gaussianblur` |
| Defocus blur | `defocus blur` |
| Motion blur | `motion blur` |
| Gaussian noise | `gaussian noise` |
| Impulse noise | `impulse noise` |
| JPEG compression | `jpeg compression`, `jpeg artifact`, `compression artifact` |
| Color shift | `color shift`, `color distortion` |
| Sharpness attribute | `sharpness`, `blur` |
| Blocking | `blocking`, `blockiness` |
| Ringing | `ringing` |

## C.2 20-D feature vector

The exact score-free representation is given by

$$\mathbf{d} \in \mathbb{R}^{20} \tag{12}$$

and comprises the features listed in Table 9. The full 21-D DSD vector is $\mathbf{z} = [s, \mathbf{d}]$, where $s$ denotes the independent MOS-only prediction.

Table 9: Exact diagnostic feature schema.

| Group | Features | Encoding |
|---|---|---|
| Severity | max, mean, min severity | real-valued summaries of levels 1/2/3 |
| Severity | count at severity 1/2/3 | nonnegative counts |
| Type | 10 supported type/attribute indicators | binary |
| Spatial | number of recognized regions | nonnegative count |
| Spatial | number of distinct degradation types | nonnegative count |
| Spatial | global indicator, local indicator | binary |

The ten type/attribute indicators correspond to blocking, color shift, defocus blur, Gaussian blur, Gaussian noise, impulse noise, JPEG compression, motion blur, ringing, and sharpness. The representation excludes the diagnostic branch's final numerical score, ground-truth MOS/DMOS values, dataset distortion labels, reference-image identities, and file paths. Image identifiers/paths are used solely for record alignment before feature construction and are never included as calibration features.

## C.3 Source-only Ridge calibration

For each directed source-to-target transfer, a Ridge calibrator is fitted exclusively on the source dataset. All main experiments use

$$\lambda = 1, \tag{13}$$

with raw, unstandardized input features, `fit_intercept=True`, and no target-domain fitting. Full DSD therefore fits 21 coefficients and one intercept. Predictions on the target dataset are clipped to $[1, 5]$. No target-dependent prediction remapping, hyperparameter selection, or test-time adaptation is performed.

# D Dataset Labels and Transfer Protocol

## D.1 Archived label normalization

The cross-dataset experiments use a fixed score representation in which higher values indicate better quality. Table 10 presents the exact preprocessing rules for the four evaluation datasets. These transformations are applied once when constructing each dataset's archived score file, before any source-to-target calibration experiment.

Table 10: Quality-label conversion used by the archived evaluation data.

| Dataset | Original label | Conversion to higher-is-better evaluation score |
|---|---|---|
| LIVE | DMOS, lower is better | $1 + 4\,(d_{\max} - d)/(d_{\max} - d_{\min})$, where extrema are computed over distorted LIVE images. |
| CSIQ | DMOS, lower is better | $1 + 4\,(d_{\max} - d)/(d_{\max} - d_{\min})$, using the archived CSIQ DMOS range. |
| TID2013 | MOS, higher is better | $1 + 4\,(m - m_{\min})/(m_{\max} - m_{\min})$, using the archived TID2013 MOS range. |
| KonIQ-10k | MOS, higher is better | MOS values are read directly from the official score file for the test split; no prediction-dependent remapping is applied. |

Dataset-specific normalization is fixed during preprocessing before transfer evaluation and is not part of transfer calibration. No prediction-dependent remapping or target-fitted nonlinear mapping is performed, and target labels are not used for hyperparameter tuning of Ridge or any VLM component. The controlled-distortion source datasets contain 779 images for LIVE, 866 for CSIQ, and 3,000 for TID2013. Evaluation on authentic distortions uses the official KonIQ-10k test split of 2,015 images.

## D.2 Strict source-only calibration

For each of the six directed transfers among LIVE, CSIQ, and TID2013, all source samples and their labels are used to fit the fixed Ridge calibrator. The fitted parameters are then applied unchanged to the target dataset. Target labels are accessed only after predictions have been generated and are used solely to compute SRCC, PLCC, KRCC, RMSE, and MAE. The same protocol is used for source-to-KonIQ evaluation.

# E Additional Reproducibility Checks

## E.1 Regularization sensitivity

Table 11 reports the archived metrics, averaged over six transfer directions, for the Ridge regularization sweep. The value used in the main paper, $\lambda = 1$, was fixed as the canonical setting rather than selected based on target-domain performance. Re-evaluating the archived feature records with $\lambda = 1$ reproduces all six SRCC and PLCC values reported in the main paper to four decimal places.

## E.2 Statistical uncertainty

The main paper reports paired, image-level bootstrap confidence intervals for the difference in SRCC between full DSD and MOS-only. The pairing between predictors is preserved within each resample. This analysis quantifies image-level uncertainty only: it neither models dependence among distortions derived from the same reference image nor captures variability due to model retraining. We therefore interpret these intervals as uncertainty estimates for the fixed archived predictors, rather than as a complete characterization of training and data uncertainty.

Table 11: Regularization sensitivity averaged over the six controlled-distortion transfers.

| $\lambda$ | Mean SRCC | Mean PLCC | Mean RMSE |
|---|---|---|---|
| $10^{-4}$ | 0.5944 | 0.6461 | 0.8223 |
| $10^{-3}$ | 0.5944 | 0.6461 | 0.8223 |
| $10^{-2}$ | 0.5944 | 0.6461 | 0.8222 |
| $10^{-1}$ | 0.5947 | 0.6464 | 0.8220 |
| 1 | 0.5958 | 0.6474 | 0.8211 |
| 10 | 0.5992 | 0.6496 | 0.8198 |
| 100 | 0.6129 | 0.6560 | 0.8166 |
| 1000 | 0.6208 | 0.6549 | 0.8142 |

Table 12 lists the six SRCC differences corresponding to Figure 3 of the main paper.

Table 12: Paired image-level bootstrap uncertainty for Full DSD versus MOS-only.

| Transfer | SRCC difference | 95% CI |
|---|---|---|
| LIVE→CSIQ | $+0.1042$ | $[0.0628, 0.1477]$ |
| LIVE→TID2013 | $+0.0731$ | $[0.0446, 0.1033]$ |
| CSIQ→LIVE | $+0.0806$ | $[0.0429, 0.1190]$ |
| CSIQ→TID2013 | $+0.1510$ | $[0.1268, 0.1733]$ |
| TID2013→LIVE | $+0.0725$ | $[0.0354, 0.1071]$ |
| TID2013→CSIQ | $+0.1598$ | $[0.1291, 0.1921]$ |

### E.3 Descriptive Diagnostic Sanity Check

The diagnostic-fidelity study reported in the main paper comprises 120 outputs, evenly stratified across LIVE, CSIQ, TID2013, and KADID-10k (30 per dataset). Four annotators familiar with image quality assessment (one author and three peers) reviewed each image together with its generated diagnosis and dataset distortion label; the predicted numerical score was hidden. The categories were defined as follows:

- **Exact**: the predicted primary degradation type matches or is semantically equivalent to the labeled/visible artifact;
- **Partial**: the broad degradation category is correct but the subtype is imprecise or additional inaccurate attributes are included;
- **Error**: the diagnosis disagrees with both the distortion label and the visible artifact.

The aggregate Exact/Partial/Error rates are 61.67%/38.33%/0%. Severity was judged visually consistent for all 120 samples (100%), and all 120 responses were structurally parseable (100%). No independent inter-rater reliability statistic was computed. Because KADID reference content was used during Phase-1 adaptation, the KADID portion of this audit is descriptive and should not be interpreted as an independent zero-shot evaluation.

**Important distinction.** This image-based diagnostic-fidelity audit evaluates whether each generated diagnosis is consistent with the visible image artifact and the known distortion label. It is distinct from a parser-only text-extraction sanity check, which evaluates whether deterministic parsing recovers attributes from generated text rather than whether the diagnosis matches the underlying image artifact. The two checks use separate protocols and evidence.

### E.4 InternVL3-8B second-backbone configuration

InternVL3-8B uses the same 15,000 Phase-1 synthetic samples, the same 20-D parser, and the same source-only Ridge setting ($\lambda = 1$). Both the MOS-only and Phase-1 diagnostic adapters use 4-bit bitsandbytes NF4-style quantization with double quantization, bfloat16 computation, a LoRA rank of 64, $\alpha = 128$, all supported LoRA target modules, a learning rate of $2 \times 10^{-5}$, three epochs, an effective batch size of 16, cosine scheduling, a warmup ratio of 0.1, and a random seed of 42.

The DSD-A condition uses the Phase-1 three-step diagnostic adapter. For DSD-B, the Phase-1 diagnostic adapter undergoes additional Phase-2 supervised fine-tuning using the six-step instruction reported in the main paper. This continuation stage uses a learning rate of $5 \times 10^{-6}$ for two epochs, a LoRA rank of 64, $\alpha = 128$, an effective batch size of 16, cosine scheduling, a warmup ratio of 0.1, and a random seed of 42. Because both the adapter state and the instruction differ between DSD-A and DSD-B, their comparison is not interpreted as a controlled ablation of reasoning length.

### E.5 External VLM-IQA context and KonIQ verification

Q-Align is evaluated through direct inference on each target dataset, without target-label fitting, target validation, or logistic mapping. VisualQuality-R1-7B is evaluated using its score-only, non-thinking-style prompt under the same direct-inference restriction. For both external methods, the expected evaluation sets contain 779 LIVE, 866 CSIQ, 3,000 TID2013, and 2,015 KonIQ-10k images. Because the training protocols of these methods differ from that of CoT-IQA, their results are presented only as contextual references, as stated in the main paper.

For KonIQ-10k, the archived verification manifest confirms 2,015 official test images, 2,015 aligned MOS-branch outputs, 2,015 aligned diagnostic outputs, no MOS parsing failures, no diagnostic parsing failures, and no all-zero diagnostic vectors. Therefore, the observed failure of source-calibrated diagnosis on KonIQ cannot be attributed simply to missing predictions or a wholesale parser collapse.

## F Software Environment

The archived reproducibility audit records the following package versions for the final analysis environment: PyTorch 2.12.0+cu130, Transformers 5.6.0, scikit-learn 1.9.0, NumPy 2.4.6, SciPy 1.17.1, and pandas 2.3.3. Q-Align was evaluated in its archived environment using Transformers 4.36.1 and the same PyTorch build.