QUALITY SIGNAL IS ALREADY THERE: TRAINING-FREE CROSS-DOMAIN IQA BY AMPLIFYING LATENT QUALITY DIRECTION IN FROZEN CLIP

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-domain image quality assessment (CD-IQA) typically requires collecting target-domain images, aligning source and target distributions, and retraining a model whenever the deployment domain changes. Frozen contrastive vision–language models such as CLIP provide a training-free alternative, but their semantic-dominant representations are relatively insensitive to perceptual quality variations. By probing frozen CLIP across heterogeneous IQA datasets, we find that independently estimated quality directions exhibit non-random alignment and concentrate in a few common components, while the associated quality signal occupies only a small fraction of the overall feature energy. Perceptual quality is therefore latent rather than absent: transferable quality structure is already encoded in CLIP, but is underweighted in its default similarity space. This partial cross-domain commonality further suggests that a direction instantiated from an accessible synthetic source can serve as a transferable prior for strengthening quality-related components in unseen domains, without reproducing each target distribution. Based on this observation, we introduce a simple quality-directional feature amplification framework. A source-instantiated quality direction is constructed by contrasting high and low quality feature centroids, either using subjective ratings when available or, notably, without MOS by ranking synthetic distortions according to their affinity to pristine references. At inference, we amplify both the image-text feature using this source-derived quality prior before the standard cosine-similarity readout. The resulting method requires no target-domain data, parameter optimization, or additional network modules. Despite being training-free, it remains competitive with trained conventional cross-domain and VLM-based IQA methods, while consistently improving the frozen CLIP baseline across synthetic, held-out algorithmic, authentic, and AIGC quality benchmarks. Further analyses show that amplification strengthens distortion sensitivity and quality discrimination while largely retaining CLIP's semantic organization. The core code can be found in the **Supplementary Materials**.

1 INTRODUCTION

With the rapid growth of social media and computational imaging, visual content is continuously captured, compressed, enhanced, restored, and generated before reaching users. Reliable image quality assessment (IQA) is therefore essential for monitoring visual content, optimizing processing systems, and maintaining perceptual experience. In most practical scenarios, a pristine reference is unavailable, making blind IQA (BIQA) the more broadly applicable formulation. Yet most learning-based BIQA methods are developed under an intra-domain protocol, where training and test images share similar contents and distortion distributions Zhang et al. (2018); Su et al. (2020); Ke et al. (2021); Yang et al. (2022a). Extending such predictors to a new domain often entails new subjective annotations and dataset-specific score processing or model fitting. Because subjective studies are costly and time-consuming, repeatedly rebuilding an evaluator for every emerging image domain is difficult to sustain Wang et al. (2023b); Li et al. (2026a).

Cross-domain IQA Ganin et al. (2016); Lu et al. (2024); Li et al. (2024); Zhu et al. (2026) seeks to reduce this dependence by transferring quality knowledge from an accessible source domain to

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

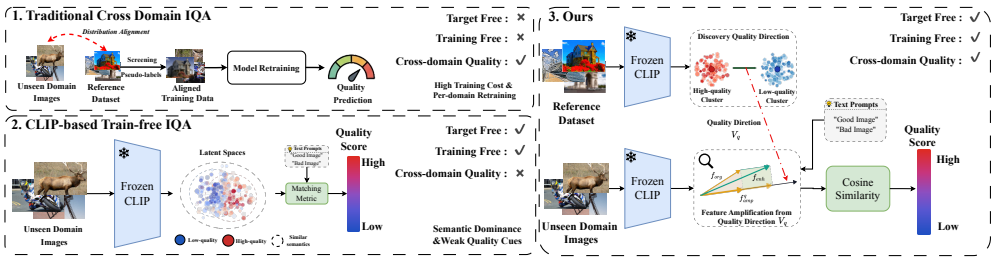


Figure 1: Motivation of proposed method. Existing cross-domain IQA methods adapt a trained evaluator to target distributions, whereas direct frozen-CLIP inference avoids retraining but underweights perceptual quality. We instead use an accessible synthetic source to expose a transferable quality direction in frozen CLIP and amplify it for target-data-free inference.

previously unseen targets. Synthetic distortion data provide a natural source: pristine images can be degraded at scale using predefined distortion types and severity levels, yielding inexpensive samples with explicit quality structure. The goal is to exploit this controllable source and generalize to less accessible domains, including authentic camera images, algorithmically processed images, and AIGC content. This synthetic-to-unseen-domain setting is practically appealing but technically challenging. Synthetic samples cover only predefined degradation processes, and their limited reference content and redundant distortion combinations can induce discrete, distortion-dependent feature clusters that transfer poorly beyond the source distribution Li et al. (2026a).

As illustrated in Figure 1, conventional cross-domain IQA methods typically rely on one or more target-aware operations, such as source-data selection, distribution alignment, pseudo-label construction, or model adaptation Li et al. (2024). Such procedures can be effective, but the matching-and-training cycle must be repeated when the target distribution changes. Well pretrained models such as CLIP Radford et al. (2021) offer a natural alternative. Their large-scale pretraining supports direct transfer across visual domains, while antonym prompts such as “good image” and “bad image” enable quality prediction without updating the model Wang et al. (2023a). However, CLIP is optimized primarily for high-level semantic alignment: its representation strongly organizes *what* an image depicts, while cues describing *how well* it is perceived receive substantially less emphasis. Direct CLIP-based IQA therefore avoids domain-specific training, but quality-related variations remain less salient than semantic content in its representation Li et al. (2025; 2026d).

This tension motivates our central hypothesis: perceptual quality is not absent from frozen CLIP, but is under-emphasized by its semantic-dominant geometry. Prior evidence that frozen visual encoders retain acquisition and processing traces and can recover distortion-sensitive cues supports this view Ramos et al. (2025); Agnolucci et al. (2024a); what remains unclear is whether these cues exhibit any consistency across quality domains. To investigate this, inspired by the notion that human-interpretable concepts can be represented as directions in feature space Kim et al. (2018), we extract diagnostic quality directions independently from several heterogeneous IQA datasets using the difference between high- and low-quality feature centroids. These directions are not identical, but show clear non-random alignment, with most of their cross-dataset variation concentrated in a few principal components. We refer to this empirical regularity as a *quality-relevant structure*, suggesting transferable quality-related components alongside domain-specific variation. This observation also gives synthetic data a different role: instead of imitating every target distribution, an accessible synthetic source can be used to expose a useful quality tendency already present in CLIP.

Motivated by this cross-domain alignment, we instantiate a deliberately simple and deployable quality prior from a single accessible source, where quality ordering is derived from MOS when available, or from distorted–pristine similarity in its absence. We then amplify the components of both image and quality-text embeddings along this prior before applying the standard CLIP similarity readout. This source-derived intervention exploits the observed transferable tendency without explicitly estimating a multi-domain subspace. The entire procedure keeps CLIP frozen and requires neither target images nor optimization. Conceptually, it is consistent with comparison-based perceptual scaling and the selective gain of task-relevant visual responses Thurstone (1927); Treue & Trujillo (1999); Martinez-Trujillo & Treue (2004): contrasting perceptual anchors exposes a weak quality dimension, and increasing its gain makes that dimension more discriminative.

Our main contributions can be summarized as follows:

- We provide empirical evidence of non-random cross-domain alignment and low-dimensional organization among quality directions in frozen CLIP, revealing transferable quality-related regularities alongside domain-specific variation.
- We introduce a simple target-data-free and training-free framework that extracts a source-derived quality prior, from either MOS or pristine affinity, and amplifies quality-relevant components in image and text features.
- Extensive experiments demonstrate the effectiveness of the proposed method and its generalization across diverse held-out IQA domains.

2 Related Work

2.1 Cross-Domain and Synthetic-to-Real BIQA

Modern BIQA methods employ content-adaptive regression, multi-scale modeling, ranking objectives, and spatial–channel attention to achieve strong performance on established benchmarks Su et al. (2020); Ke et al. (2021); Golestaneh et al. (2022); Yang et al. (2022a). Cross-domain BIQA moves beyond this matched-distribution setting and studies how quality knowledge learned from an accessible source can transfer to unseen domains.

Existing approaches mainly improve transfer in two ways. One broadens the training distribution or constructs cross-dataset ranking supervision from laboratory, synthetic, and in-the-wild images Zhang et al. (2021); Wang et al. (2023b). The other adapts a source predictor to unlabeled target samples through quality-aware style or frequency alignment, source-free adaptation, or test-time optimization Lu et al. (2024); Li et al. (2023b); Liu et al. (2022); Roy et al. (2023). These strategies reduce the need for target MOS, but still require task-specific training or target-dependent computation. More recent source-centric methods explicitly address the mismatch induced by synthetic distortions. DGQA Li et al. (2024) and GDCIQA Zhu et al. (2026) model distortion or domain relevance and combine source selection with source–target alignment, whereas SynDR-IQA Li et al. (2026a) reshapes the synthetic distribution by increasing content diversity and suppressing redundant distortion clusters. Despite their different mechanisms, these methods use synthetic data to train or adapt a dedicated BIQA predictor. PromptIQA Chen et al. (2024) takes a different route to adaptation, conditioning a model trained on mixed labeled IQA datasets on image–score pairs that specify the target assessment criterion. Our method gives the source a more limited role: it is used once to extract a quality-oriented prior from frozen CLIP, which is then reused across targets without distribution matching or retraining.

2.2 Vision-Language Models for Image Quality Assessment

Vision–language pre-training provides an alternative route to transferable quality assessment. Lightweight methods read quality from frozen CLIP through antonym prompts or a magnitude-aware similarity, while learned variants model joint quality–scene–distortion correspondence or text-defined quality levels Wang et al. (2023a); Liao et al. (2026); Zhang et al. (2023a); Wu et al. (2023). Collectively, they establish language as an effective interface for perceptual scoring, ranging from direct embedding comparison to trained multimodal prediction. Quality-oriented adaptation further exploits ordered synthetic degradations Agnolucci et al. (2024a), transferable prompts Li et al. (2025), gradient regulation, and perceptual-manifold alignment to make vision–language representations more distortion-sensitive Li et al. (2026d). Together with evidence that frozen visual encoders retain acquisition and processing traces despite semantic pre-training Ramos et al. (2025), these results support the presence of latent perceptual cues. However, they generally expose such cues through representation, prompt, or parameter learning; our focus is a closed-form intervention in the frozen feature space.

Another line strengthens perceptual evidence by changing the observation and reasoning process. Zoom-IQA Liang et al. (2026) learns to identify informative image regions and iteratively refine quality assessment, improving *where* and *how* perceptual evidence is acquired. These gains rely on task-specific optimization and multi-stage inference. Tool-IQA Qin et al. (2026) is structurally different from both direct VLM scoring and learned visual reasoning: it formulates IQA as an agentic

tool-use problem and trains an assessment agent to invoke magnification and gamma correction during iterative evaluation. Its strong performance therefore reflects a complete trained system with an adaptive tool-calling policy and repeated inference, rather than a single frozen representation. Our work studies a complementary lightweight setting: it generates no auxiliary image views and directly amplifies a source-derived quality direction in frozen image and text embeddings before the standard similarity readout.

3 METHOD

3.1 PROBLEM FORMULATION AND OVERVIEW

Let $\mathcal{D}_s$ be an accessible source collection and let $\mathbb{P}_T$ denote an unknown distribution over possible target domains. For a frozen CLIP image–text encoder pair (E_I, E_T), we formalize a source-derived predictor and its desired cross-domain utility as

$$\begin{aligned} \boldsymbol{\phi}_s &= \Phi(\mathcal{D}_s; E_I), \\ s_{\boldsymbol{\phi}_s}(x) &= G\big(E_I(x), E_T(\mathcal{P}^+, \mathcal{P}^-); \boldsymbol{\phi}_s\big), \\ \mathcal{J}(\boldsymbol{\phi}_s) &= \mathbb{E}_{\mathcal{D}_t \sim \mathbb{P}_T}[\rho_{\mathcal{D}_t}(s_{\boldsymbol{\phi}_s}, y_t)], \end{aligned} \tag{1}$$

where $\mathcal{P}^+$ and $\mathcal{P}^-$ are positive- and negative-quality prompts, y_t denotes human quality scores used only for evaluation, and ρ is a correlation criterion such as SRCC or PLCC. The desideratum is that $\mathcal{J}$ remain high across unseen domains. Crucially, neither $\mathbb{P}_T$ nor any realization $\mathcal{D}_t$ is available when constructing $\boldsymbol{\phi}_s$; the encoders are frozen and no target-dependent parameter selection is permitted. Eq. 1 therefore cannot be optimized directly. Throughout, $\mathrm{Norm}(\mathbf{z}) = \mathbf{z}/\|\mathbf{z}\|_2$.

To obtain a tractable source-only solution, we adopt a minimal working hypothesis: a frozen CLIP feature contains a quality-sensitive component that can be selectively reweighted. Let $\Pi_{\mathcal{Q}}$ denote orthogonal projection onto the unknown quality-relevant structure $\mathcal{Q}$. An ideal intervention would preserve the remaining representation while increasing the gain of $\Pi_{\mathcal{Q}}\mathbf{f}$ by a factor controlled by $\alpha \geq 0$. Our approximation is motivated by two perceptual principles. Comparison-based scaling represents a latent perceptual attribute through relative judgments between contrasting conditions Thurstone (1927); accordingly, relative high- and low-quality clusters serve as opposing anchors, and their centroid difference provides a source-identifiable direction of quality change. Selectively increasing this component is also consistent in spirit with feature-based attentional gain, which strengthens responses aligned with a task-relevant visual property Treue & Trujillo (1999). Because the complete $\mathcal{Q}$ cannot be identified without broader domain information, we retain only this contrastive direction $\mathbf{v}_q$ as a feasible approximation:

$$\begin{aligned} \widehat{\mathbf{f}}_{\text{ideal}} &= \mathrm{Norm}[\mathbf{f} + \alpha \Pi_{\mathcal{Q}} \mathbf{f}] \\ &\approx \mathrm{Norm}\big[\mathbf{f} + \alpha(\mathbf{f}^\top \mathbf{v}_q)\mathbf{v}_q\big] \equiv \mathcal{A}_\alpha(\mathbf{f}; \mathbf{v}_q). \end{aligned} \tag{2}$$

These connections motivate the design rather than claim a mechanistic model of human vision. The approximation assumes only that the high–low contrast captures a useful quality-sensitive component. The resulting prior $\boldsymbol{\phi}_s = (\mathbf{c}_H, \mathbf{c}_L, \mathbf{v}_q)$ supplies both the direction used for image amplification and the ordered anchors used to calibrate the Pos–Neg text contrast. Their enhanced features are finally compared by cosine similarity, as summarized in Figure 2. The entire approximation is constructed once and reused without target-specific changes.

3.2 SOURCE PERCEPTION PRIOR EXTRACTION

Given relative high- and low-quality source sets $\mathcal{S}_H$ and $\mathcal{S}_L$, their centroids and ordered direction are written compactly as

$$\mathbf{c}_z = \frac{1}{|\mathcal{S}_z|} \sum_{i \in \mathcal{S}_z} \mathrm{Norm}(E_I(x_i)), \quad z \in \{H, L\}; \qquad \mathbf{v}_q = \mathrm{Norm}(\mathbf{c}_H - \mathbf{c}_L). \tag{3}$$

When MOS is available, $\mathcal{S}_H$ and $\mathcal{S}_L$ are selected from the two ends of the score distribution. Our default KADIS Lin et al. (2019) construction instead requires no MOS and uses only pristine–distortion pairs.

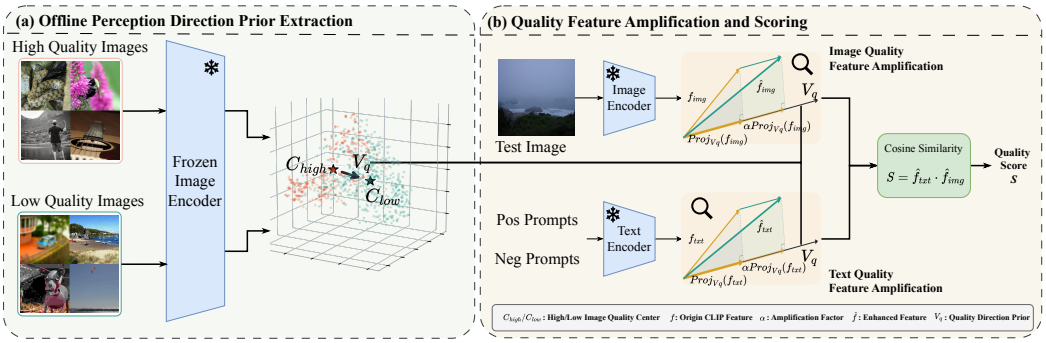


Figure 2: Method overview. (a) Relative high- and low-quality source anchors define $\mathbf{v}_q$ in frozen CLIP. (b) The test-image feature and Pos–Neg text contrast are enhanced according to the source quality orientation and compared by cosine similarity.

For each pristine image x_r^0, we encode its $M = 125$ distorted variants $\{x_r^j\}_{j=1}^M$ and rank them by pristine affinity $a_r^j = (\mathbf{f}_r^0)^\top \mathbf{f}_r^j$, where $\mathbf{f}_r^j = \mathrm{Norm}(E_I(x_r^j))$. Define $\mathcal{I}_{r,H} = \{0\} \cup \mathrm{TopK}_{1\le j\le M}(a_r^j)$ and $\mathcal{I}_{r,L} = \mathrm{BottomK}_{1\le j\le M}(a_r^j)$. The same expression gives both global centroids:

$$\mathbf{c}_z = \frac{1}{R}\sum_{r=1}^{R} \frac{1}{|\mathcal{I}_{r,z}|} \sum_{j\in\mathcal{I}_{r,z}} \mathbf{f}_r^j, \qquad z \in \{H, L\}. \tag{4}$$

Within-reference ranking controls content variation, while aggregation across references and distortion types reduces content- and distortion-specific effects. The affinity identifies contrasting anchors rather than serving as a predicted MOS. Figure 3 illustrates the construction.

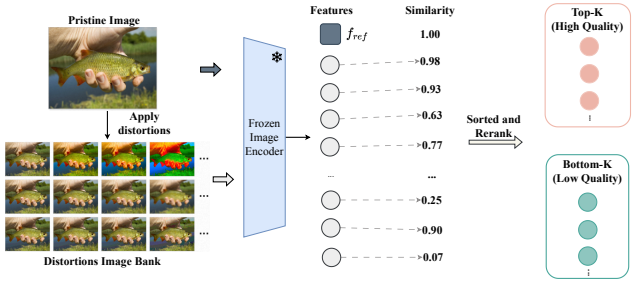


Figure 3: MOS-free anchor construction. Distortions are ranked by their frozen-CLIP affinity to the corresponding pristine image. The Top-K variants plus the pristine feature and the Bottom-K variants form the two anchor pools used to construct $\mathbf{v}_q$.

3.3 Quality Feature Amplification and Scoring

Image feature. For a test image x, let $\mathbf{f}_I = \mathrm{Norm}(E_I(x))$. We apply the source approximation in Eq. 2 to obtain $\widehat{\mathbf{f}}_I = \mathcal{A}_\alpha(\mathbf{f}_I; \mathbf{v}_q)$. Before normalization, this update preserves the component orthogonal to $\mathbf{v}_q$ and scales its directional component by $1 + \alpha$.

Anchor-guided text amplification. The source prior also provides visual anchors for calibrating the quality-text prototypes. Let t_+ and t_- denote the normalized positive- and negative-quality text prototypes, and let $\bar{c}_H = \mathrm{Norm}(c_H)$ and $\bar{c}_L = \mathrm{Norm}(c_L)$. We form $T = [t_+; t_-]$ and $C = [\bar{c}_H; \bar{c}_L]$, and compute

$$P = \mathrm{Softmax}_{\mathrm{row}}(TC^\top/\tau_a), \qquad U = T + \lambda PC, \tag{5}$$

$$\hat{T} = \mathrm{Norm}_{\mathrm{row}}(U), \qquad \hat{f}_T = \mathrm{Norm}(\hat{t}_+ - \hat{t}_-). \tag{6}$$

Each text prototype incorporates a similarity-weighted combination of the source visual anchors. This introduces source-derived quality information into the text readout while retaining the original

text prototype through the residual connection. We use $\lambda = \alpha$ throughout. The image and text branches share the same source quality direction prior but use different transformations. The image branch explicitly amplifies the projection along v_q, whereas the text branch calibrates the quality contrast using the two visual anchors. To make this connection explicit, let u_+ and u_- be the rows of U, and let p_H^+ and p_H^- be their respective weights on the high-quality anchor:

$$u_+ - u_- = (t_+ - t_-) + \lambda(p_H^+ - p_H^-)(\bar{c}_H - \bar{c}_L). \tag{7}$$

Thus, the correction to the text contrast is induced by the high–low visual anchor direction. Both branches are constructed from the same source anchors and remain fixed across target domains. Interestingly, both have a base-plus-guidance-residual form, algebraically reminiscent of classifier-free guidance (CFG) Ho & Salimans (2022). Here the source quality contrast acts as the condition, the feature projection produces an input-dependent residual, and α controls its strength. This is a structural correspondence rather than an application of generative CFG: the condition is injected once into a frozen feature instead of combining conditional and unconditional predictions during iterative generation.

Because the Pos-Neg difference is already contained in $\widehat{\mathbf{f}}_T$, the final score is a single cosine similarity:

$$s_{\boldsymbol{\phi}_s}(x) = \widehat{\mathbf{f}}_I^\top \widehat{\mathbf{f}}_T. \tag{8}$$

All source quantities and the text feature are precomputed. Inference requires one frozen image-encoder pass, one directional residual update, and one cosine similarity, with no target-domain adaptation data or learnable parameters.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate on several benchmarks covering conventional synthetic distortions (LIVE Sheikh et al. (2006), CSIQ Larson & Chandler (2010), TID2013 Ponomarenko et al. (2015), and KADID-10K Lin et al. (2019)), algorithmic distortions (PIPAL Gu et al. (2020)), authentic distortions (LIVEC Ghadiyaram & Bovik (2015), KonIQ-10K Hosu et al. (2020), SPAQ Fang et al. (2020), BID Ciancio et al. (2010), and FLIVE Ying et al. (2020)), and AIGC quality (AGIQA-1K Zhang et al. (2023b) and AGIQA-3K Li et al. (2023a)). KADIS is used only as the accessible synthetic source bank for constructing the default direction.

Implementation details and protocols. The default backbone is frozen CLIP RN50. We use the CLIP-IQA Wang et al. (2023a) and MA-CLIP Liao et al. (2026) prompt ensemble, $R = 100$ sampled KADIS references, Top-/Bottom-$K = 50$ candidates per reference, and a fixed amplification factor $\alpha = 2$. The pristine reference is added to the Top-K candidates, yielding $K + 1$ high-affinity and K low-affinity anchors per reference. These choices are shared by all target datasets. Ours opinion-unaware uses pristine-distortion affinity from KADIS no MOS, whereas Ours with mos uses MOS-defined KADID source anchors to quantify the benefit of source supervision with $K = 2000$. The KADIS construction is repeated over 15 independently sampled source sets, and the main tables report the median result. Extended implementation settings are reported in Appendix I.

4.2 Cross-Domain Generalization Performance

4.2.1 Synthetic-to-Authentic Transfer

Table 1 reveals a favorable trade-off between predictive accuracy and methodological complexity. With a frozen CLIP backbone and no target access or optimization, Ours-KADID reaches an average SRCC/PLCC of 0.778/0.817, outperforming the traditional and source-trained VLM baselines and remaining competitive with dedicated cross-domain and reasoning-based systems. In particular, its large margin over CLIP-IQA+, which learns quality prompts from the same synthetic source, shows that directly reweighting the latent quality component transfers more effectively under this protocol than fitting the prompt alone. Tool-IQA obtains the highest average, but uses a trained agentic pipeline with multi-stage inference and is therefore best viewed as a system-level upper reference rather than a matched training-free comparison. PreSTA-S reaches 0.784/0.795 by using

Table 1: Cross-domain IQA performance on authentic datasets. The best results within each method family are bolded and the top two of cross-domain family are in **bold** and underline, respectively.

Method	Source	Train	Target	LIVEC	KonIQ-10K	BID	Average
Traditional IQA							
RankIQA Liu et al. (2017)	KADID	Yes	No	.491/.495	.603/.551	.510/.367	.535/.471
DBCNN	KADID	Yes	No	**.572/.589**	**.639/.618**	**.620/.609**	**.613/.606**
HyperIQA	KADID	Yes	No	.490/.487	.545/.556	.379/.282	.472/.442
MUSIQ	KADID	Yes	No	.517/.524	.554/.573	.575/.600	.549/.566
VLM-based IQA							
CLIP-IQA+	KADID	Yes	No	.512/.543	.511/.515	.474/.442	.499/.500
LIQE	KADID	Yes	No	.719/.726	**.684**/.652	.677/.680	.693/.686
Q-Align	KADID	Yes	No	.554/.643	.573/.612	.576/.651	.568/.635
DeQA-score You et al. (2025)	KADID	Yes	No	**.743/.795**	.677/**.703**	**.702/.743**	**.707/.747**
Reasoning- and agent-based IQA							
Qwen2.5-VL-7B	KADID	Yes	No	.733/.760	.754/.810	.711/.725	.733/.765
Qwen3-VL-8B-Ins Bai et al. (2025a)	KADID	Yes	No	.779/.725	.760/.772	.729/.769	.756/.755
Q-Insight Li et al. (2026c)	KADID	Yes	No	.761/.795	.806/.829	.784/.796	.784/.807
VisualQuality-R1	KADID	Yes	No	.792/.799	.788/.811	.788/.807	.789/.806
Tool-IQA(Qwen3-8B)	KADID	Yes	No	**.833/.854**	**.825/.848**	**.828/.842**	**.829/.848**
UDA cross-domain IQA							
StyleAM	KADID	Yes	Yes	.584/.561	.700/.673	.637/.567	.640/.600
FreqAlign	KADID	Yes	Yes	.618/.588	.748/.721	.674/.708	.680/.673
DGQA	KADID	Yes	Yes	.696/.690	.681/.687	.770/.753	.716/.710
SynDR-IQA	KADID	Yes	No	.713/.714	.727/.735	.788/.764	.743/.737
GDCIQA	KADID	Yes	Yes	.753/.756	.721/.738	.787/.781	.754/.758
PreSTA-S Li et al. (2026b)	KADID	Yes	Yes	.744/.756	.774/.782	**.833/.846**	**.784**/.795
Ours-KADIS(no MOS)	KADIS	**No**	**No**	.777/.813	.757/.809	.787/.824	.774/.815
Ours-KADID(MOS)	KADID	**No**	**No**	**.781/.815**	**.775/.823**	.777/.812	.778/**.817**

target features and labels to select a target-matched source subset Li et al. (2026b). Its unusually strong BID result likely benefits from source–target compatibility: BID is dominated by authentic blur, for which KADID contains closely related synthetic degradations. In contrast, both of our variants remain balanced across all three targets under one unchanged inference rule. Moreover, Ours-KADIS achieves 0.774/0.815 without source MOS, only 0.004/0.002 below the MOS-guided variant, showing that pristine affinity provides a nearly equally effective source prior. Measured inference overhead, offline construction time, and prior storage are reported in Appendix E.

4.2.2 Target-Free Multi-Domain Evaluation

Table 2 compares quality prediction across diverse domains, with separately labeled trained methods as contextual references. In part (a) our prior improves over every zero-shot and CLIP-based training-free baseline on all six benchmarks, exceeding the baseline methods like MA-CLIP. The largest margins appear where distortions are synthetic and graded, as expected for a direction estimated from controlled degradations, yet the gains persist on authentic data, where such degradations never occur, indicating that the amplified component is not bound to the source distortion family.

Part (b) further evaluates held-out restoration and AIGC quality domains. Despite their substantial distribution shifts from the synthetic source, our method achieves the best training-free SRCC on all three datasets, with a particularly clear gain of .078 over MA-CLIP on AGIQA-1K. These consistent improvements demonstrate that the source-derived quality direction remains effective beyond conventional synthetic and authentic distortions. To provide a broader perspective, we include additional noteworthy observations and exploratory extensions in Appendix G.

4.3 Cross-Dataset Geometry of Quality Directions

We examine whether independently constructed quality directions share a common cross-dataset structure. For each dataset, we compute a high-minus-low MOS direction solely for diagnostic analysis; none of these target-derived directions is used for prediction. Figure 4a(a) shows a mean off-diagonal cosine similarity of 0.60, compared with approximately 0.03 for random directions, providing clear evidence of a shared component. Meanwhile, the synthetic/authentic block structure and weaker cross-group alignment show that this component is not one identical universal axis. To characterize its dimensionality, we stack all dataset directions into a matrix and apply uncentered

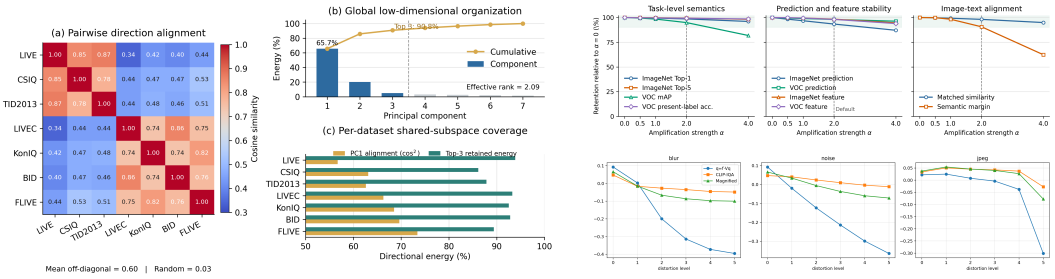
Table 2: Comparison of training-free IQA across diverse quality domains. Best results among training-free methods are in **bold**.

(a) Conventional synthetic and authentic benchmarks

Method	LIVEC	CSIQ	TID2013	KADID	KonIQ	SPAQ
Zero-shot IQA						
NIQE Mittal et al. (2012)	.450/.494	.619/.690	.311/.379	.378/.388	.530/.484	.311/.264
QAC Xue et al. (2013)	.226/.284	.480/.593	.372/.419	.239/.309	.340/.291	.440/.450
PIQE Venkatanath et al. (2015)	.233/.314	.512/.628	.364/.462	.237/.289	.245/.206	.232/.249
NPQI Liu et al. (2020)	.475/.492	.634/.648	.280/.400	.391/.340	.613/.614	.600/.616
MDFS Ni et al. (2024)	.482/.536	.777/.791	.536/.624	.598/.594	.733/.712	.741/.718
CLIP-based training-free IQA						
MA-CLIP Liao et al. (2026)	.743/.768	.737/.783	.599/.676	.525/.549	.755/.803	.773/.778
CLIP-IQA	.718/.749	.698/.767	.558/.652	.474/.503	.645/.721	.772/.779
Ours-KADIS	.777/.813	.787/.837	**.661/.719**	**.674/.681**	.757/.809	.819/.825
Δ	↑.059/↑.064	↑.089/↑.070	↑.103/↑.067	↑.200/↑.178	↑.112/↑.088	↑.047/↑.046
Ours-KADID	**.781/.815**	**.796/.841**	.651/.712	–	**.775/.823**	**.822/.828**
Δ	↑.063/↑.066	↑.098/↑.074	↑.093/↑.060	–	↑.130/↑.102	↑.050/↑.049

(b) Held-out restoration and AIGC quality

Method	IQA prior	PIPAL	AGIQA-1K	AGIQA-3K
Opinion-unaware training iqa				
Re-IQA Saha et al. (2023)	Target	.568/.587	.783/.840	.811/.874
ARNIQA Agnolucci et al. (2024b)	KADIS	.634/.666	.768/.849	.803/.881
QualiCLIP Agnolucci et al. (2024a)	KADIS	.410/.424	.736/.805	.667/.735
GRepQ Srinath et al. (2024)	FLIVE	.554/.568	.740/.797	.807/.858
CL-MI Babu et al. (2023)	KADIS	.281/.282	.474/.621	.591/.665
CLIP-based training-free IQA				
MA-CLIP Liao et al. (2026)	None	.371/.393	.528/.668	.696/.734
CLIP-IQA	None	.369/.374	.522/.663	.652/.715
Ours-KADIS	KADIS	.386/.408	.606/.721	**.724/.763**
Δ		↑.017/↑.034	↑.084/↑.058	↑.072/↑.048
Ours-KADID	KADID	**.402/.419**	**.619/.725**	.711/.756
Δ		↑.033/↑.045	↑.097/↑.062	↑.059/↑.041



(a) (b)

Figure 4: Representation geometry and amplification behavior. (a) Independently constructed quality directions exhibit non-random alignment, low-dimensional organization, and distinct synthetic/authentic variations. (b) Amplification strengthens the response to increasing distortion (bottom) while largely retaining CLIP's semantic capability (top) within the useful range of α.

SVD. Here, the first principal component (PC1) denotes its leading right-singular vector, namely the axis capturing the largest shared directional energy. As shown in Panel (b), PC1 accounts for 65.7% of the total energy, while the first three components jointly account for 90.8%, yielding an effective rank of 2.09. Panel (c) further shows that projecting each direction onto this shared three-component basis retains 86.1%–93.9% of its energy, whereas their different alignments with PC1 preserve domain-specific variation. Together, these results reveal a low-dimensional quality-comparable organization containing both shared and domain-dependent components. The proposed KADIS direction is therefore not an estimate of the complete subspace, but a deployable single-source approximation aligned with its transferable component. The corresponding per-dataset PC1 alignments and energy-retention values are reported in Appendix C.

4.4 Effect of Quality-Signal Amplification

Semantic retention. Because the intervention modifies a semantic representation, increased quality sensitivity is useful only if CLIP's original capabilities are largely retained. The upper plots of Figure 4b evaluate this trade-off over the complete amplification sweep. At the shared default $\alpha = 2$, 93.4% of Top-1 predictions are unchanged and the enhanced features retain a cosine similarity of 0.984 to their original counterparts on ImageNet Deng et al. (2009). On VOC2007 Everingham et al. (2010), mAP retains 94.9% of its baseline value, present-label accuracy changes only from 0.871 to 0.866. Matched image-text similarity also retains 98.2% of its original value. Thus, the default update substantially increases quality-directional energy while preserving most task-level recognition, feature geometry, and cross-modal alignment. The result supports semantic retention, rather than exact semantic invariance.

Directional energy and quality sensitivity. For a unit feature $\mathbf{f}$ and unit direction $\mathbf{v}_q$, we measure the one-dimensional directional energy $\rho(\mathbf{f}) = \|(\mathbf{f}^\top \mathbf{v}_q)\mathbf{v}_q\|_2^2$. As shown in Figure 5, the KADIS direction occupies 2.85% of the normalized feature energy on average across

Table 3: Component ablation using the KADIS direction with $\alpha = 2$.

Variant	LIVEC	KonIQ	BID	SPAQ
Frozen CLIP	.718/.749	.645/.721	.757/.787	.772/.779
Text only	.732/.763	.668/.742	.766/.796	.783/.784
Image only	.774/.809	.751/.805	.787/.822	.813/.819
Image + text	**.777/.813**	**.757/.809**	**.787/.824**	**.819/.825**

Table 4: Quality-direction construction. Average transfer over 4 datasets.

Construction	Avg. SRCC	Avg. PLCC
CLIP pseudo-score split	.733	.764
Distortion-level split	.766	.791
SVD subspace	.767	.793
Pristine affinity	**.772**	**.797**

Table 5: Retrospective validation of the ranking surrogate on KADID-10k.

Diagnostic	Dist. level	Prist. affinity
Within-ref. SRCC	$.6851 \pm .028$	$\mathbf{.7839 \pm .039}$
Cross-type PairAcc	79.20%	$\mathbf{85.86\%}$
High–low anchor gap	1.5429	$\mathbf{1.7135}$
Five-bin MOS R^2	$.4543$	$\mathbf{.5750}$

eight representative datasets. Although small relative to the complete representation, this one-dimensional projection contains $26.2\times$ the energy of a random one-dimensional direction (0.108%). The proposed update raises the average directional contribution to 8.43% at $\alpha = 2$, directly confirming that it selectively increases the gain of a structured but underweighted component. Energy alone does not establish that this component responds to perceptual degradation. We therefore apply progressively stronger blur, noise, and JPEG compression to the same pristine images and track three quantities in the lower plots of Figure 4b. The projection changes strongly as blur and noise increase and shows a pronounced response to severe JPEG compression, whereas the original CLIP-IQA score varies over a much narrower range. Amplification transfers this directional sensitivity into the quality readout, yielding a larger response range across each distortion series while still retaining the original CLIP feature. These controlled trajectories directly connect the geometric operation to increased distortion sensitivity. For more detailed experimental analyses and results, please refer to Appendix C and Appendix D.

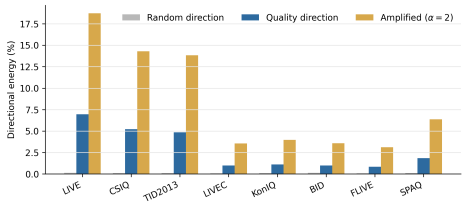


Figure 5: Feature energy along the source-derived quality direction, compared with a random one-dimensional direction; amplification selectively increases the contribution of the quality direction.

4.5 Ablation of the Proposed Design

Image and text contributions. Table 3 separates the two branches. Direct image-feature amplification accounts for most of the gain, while calibrating the text contrast with the same ordered source anchors provides a smaller but consistent complement. Their combination performs best or ties for best across the four authentic datasets, supporting complementary branch-specific updates under shared source-derived quality guidance. Additional analyses are provided in Appendices A and F.

Quality-prior construction. Table 4 compares alternative direction estimators using the complete readout, averaged over LIVE, KADID-10K, LIVEC, and KonIQ-10K. Although several constructions are effective, pristine-affinity anchors achieve the best average transfer and slightly outperform the more elaborate SVD alternative. Table 5 provides a retrospective explanation: on the held-out MOS-labeled KADID-10K, pristine affinity agrees more closely with human quality than nominal distortion levels across all four diagnostics. It therefore provides a more reliable cross-distortion surrogate for constructing the quality direction without MOS. Additional experiments regarding quality-aware direction construction are provided in Appendix B.

5 Conclusion

We reveal a partially shared quality-comparable structure across heterogeneous IQA domains in frozen CLIP and approximate it with a simple source-derived direction. By amplifying the corresponding components of both image and text features, our method consistently improves quality prediction across diverse target domains without target data, parameter updates, or additional modules, while largely retaining semantic discrimination. These results show that synthetic data need not imitate every target distribution; it can instead expose transferable perceptual priors already encoded in frozen latent representations. This simple yet effective perspective offers a practical foundation for training-free cross-domain IQA and future multidimensional quality guidance. Representative failure cases and a discussion of limitations are provided in Appendix H.

AI USE STATEMENT

In this work, we utilized LLMs to assist in improving the grammar, clarity, and overall readability of this manuscript. All scientific claims, experimental results, and core ideas were conceived and articulated by the human authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We have provided implementation details in Sec. 4.1 and more experimental setups and data source details can be found in the Appendix I. The core code can also be found in the **Supplementary Materials**.

REFERENCES

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

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

Nithin C Babu, Vignesh Kannan, and Rajiv Soundararajan. No reference opinion unaware quality assessment of authentically distorted images. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2458–2467. IEEE, 2023.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025b. URL https://arxiv.org/abs/2502.13923.

Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad—a comprehensive real-world dataset for unsupervised anomaly detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9584–9592. IEEE, 2019.

Bar Cavia, Eliahu Horwitz, Tal Reiss, and Yedid Hoshen. Real-time deepfake detection in the real-world. *arXiv preprint arXiv:2406.09398*, 2024.

Ruoxin Chen, Junwei Xi, Zhiyuan Yan, Ke-Yue Zhang, Shuang Wu, Jingyi Xie, Xu Chen, Lei Xu, Isabel Guan, Taiping Yao, and Shouhong Ding. Dual data alignment makes ai-generated image detector easier generalizable. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 106475–106500. Curran Associates, Inc., 2025. doi: 10.52202/085713-3552. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/991c9f1799cfeebb4217baaacc462d86-Paper-Conference.pdf.

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. Springer, 2024.

Alexandre Ciancio, Eduardo AB Da Silva, Amir Said, Ramin Samadani, Pere Obrador, et al. No-reference blur assessment of digital pictures based on multifeature classifiers. *IEEE Transactions on image processing*, 20(1):64–75, 2010.

Davide Cozzolino, Giovanni Poggi, Riccardo Corvi, Matthias Nießner, and Luisa Verdoliva. Raising the bar of ai-generated image detection with clip. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 4356–4366. IEEE, 2024.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2): 303–338, 2010.

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

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016.

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, 2015.

S Alireza Golestaneh, Saba Dadsetan, and Kris M Kitani. No-reference image quality assessment via transformers, relative ranking, and self-consistency. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 3989–3999. IEEE, 2022.

Jinjin Gu, Haoming Cai, Haoyu Chen, Xiaoxing Ye, Ren Jimmy S, and Chao Dong. Pipal: a large-scale image quality assessment dataset for perceptual image restoration. In *European conference on computer vision*, pp. 633–651. Springer, 2020.

Fabrizio Guillaro, Giada Zingarini, Ben Usman, Avneesh Sud, Davide Cozzolino, and Luisa Verdoliva. A bias-free training paradigm for more general ai-generated image detection. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18685–18694. IEEE, 2025.

Shuai He, Yongchang Zhang, Rui Xie, Dongxiang Jiang, and Anlong Ming. Rethinking image aesthetics assessment: Models, datasets and benchmarks. In *IJCAI*, volume 3, pp. 6, 2022.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

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 *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5128–5137. IEEE, 2021.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pp. 2668–2677. PMLR, 2018.

Shu Kong, Xiaohui Shen, Zhe Lin, Radomir Mech, and Charless Fowlkes. Photo aesthetics ranking network with attributes and content adaptation. In *European conference on computer vision*, pp. 662–679. Springer, 2016.

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

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):011006–011006, 2010.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28422–28431. IEEE, 2024.

Aobo Li, Jinjian Wu, Yongxu Liu, Leida Li, and Weisheng Dong. Towards syn-to-real iqa: A novel perspective on reshaping synthetic data distributions. *Advances in Neural Information Processing Systems*, 38:133073–133099, 2026a.

Aobo Li, Jinjian Wu, Yongxu Liu, Jupo Ma, and Weisheng Dong. Rethinking knowledge transfer in image quality assessment: A perceptual preference structure alignment perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1310–1319, 2026b.

Chunyi Li, Zicheng Zhang, Haoning Wu, Wei Sun, Xiongkuo Min, Xiaohong Liu, Guangtao Zhai, and Weisi Lin. Agiqa-3k: An open database for ai-generated image quality assessment. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(8):6833–6846, 2023a.

Weiqi Li, Xuanyu Zhang, Shijie Zhao, Yabin Zhang, Junlin Li, Jian Zhang, et al. Q-insight: Understanding image quality via visual reinforcement learning. *Advances in Neural Information Processing Systems*, 38:36802–36827, 2026c.

Xin Li, Yiting Lu, and Zhibo Chen. Freqalign: Excavating perception-oriented transferability for blind image quality assessment from a frequency perspective. *IEEE Transactions on Multimedia*, 26:4652–4666, 2023b.

Xudong Li, Zihao Huang, Yan Zhang, Yunhang Shen, Ke Li, Xiawu Zheng, Liujuan Cao, and Rongrong Ji. Few-shot image quality assessment via adaptation of vision-language models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10442–10452. IEEE, 2025.

Yu Li, Zhengran Shen, Yachun Mi, Puchao Zhou, and Shaohui Liu. Bridging the perceptual gap: Residual-enhanced downscaling and manifold-aware perception alignment adaptation for nr-iqa. In *Forty-third International Conference on Machine Learning*, 2026d.

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.

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

Hanhe Lin, Vlad Hosu, and Dietmar Saupe. Kadid-10k: A large-scale artificially distorted iqa database. In *2019 eleventh international conference on quality of multimedia experience (qomex)*, pp. 1–3. IEEE, 2019.

Jianzhao Liu, Xin Li, Shukun An, and Zhibo Chen. Source-free unsupervised domain adaptation for blind image quality assessment. *arXiv preprint arXiv:2207.08124*, 2022.

Xialei Liu, Joost Van De Weijer, and Andrew D Bagdanov. Rankiqa: Learning from rankings for no-reference image quality assessment. In *2017 IEEE international conference on computer vision (ICCV)*, pp. 1040–1049. IEEE, 2017.

Yutao Liu, Ke Gu, Xiu Li, and Yongbing Zhang. Blind image quality assessment by natural scene statistics and perceptual characteristics. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 16(3):1–91, 2020.

Yiting Lu, Xin Li, Jianzhao Liu, and Zhibo Chen. Styleam: Perception-oriented unsupervised domain adaption for no-reference image quality assessment. *IEEE Transactions on Multimedia*, 27: 2043–2058, 2024.

Kede Ma, Qingbo Wu, Zhou Wang, Zhengfang Duanmu, Hongwei Yong, Hongliang Li, and Lei Zhang. Group mad competition-a new methodology to compare objective image quality models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1664–1673, 2016.

Julio C Martinez-Trujillo and Stefan Treue. Feature-based attention increases the selectivity of population responses in primate visual cortex. *Current biology*, 14(9):744–751, 2004.

Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012.

Naila Murray, Luca Marchesotti, and Florent Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 2408–2415. IEEE, 2012.

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

Nikolay Ponomarenko, Lina Jin, Oleg Ieremeiev, Vladimir Lukin, Karen Egiazarian, Jaakko Astola, Benoit Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, et al. Image database tid2013: Peculiarities, results and perspectives. *Signal processing: Image communication*, 30:57–77, 2015.

Guanyi Qin, Junjie Zhang, Chunming He, Yibing Fu, Jie Liang, Tianhe Wu, and Lei Zhang. Tool-iqa: Augmenting image quality assessment with simple tools. *arXiv preprint arXiv:2606.16082*, 2026.

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

Ryan Ramos, Vladan Stojnić, Giorgos Kordopatis-Zilos, Yuta Nakashima, Giorgos Tolias, and Noa Garcia. Processing and acquisition traces in visual encoders: What does clip know about your camera? In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17056–17066. IEEE, 2025.

Subhadeep Roy, Shankhanil Mitra, Soma Biswas, and Rajiv Soundararajan. Test time adaptation for blind image quality assessment. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16696–16705. IEEE, 2023.

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

Hamid R Sheikh, Muhammad F Sabir, Alan C Bovik, et al. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Trans. Image Process.*, 15(11):3440–3451, 2006.

Suhas Srinath, Shankhanil Mitra, Shika Rao, and Rajiv Soundararajan. Learning generalizable perceptual representations for data-efficient no-reference image quality assessment. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 22–31. IEEE, 2024.

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 *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3664–3673. IEEE, 2020.

Louis L Thurstone. A law of comparative judgment. *Psychological review*, 34(4):273, 1927.

Stefan Treue and Julio C Martinez Trujillo. Feature-based attention influences motion processing gain in macaque visual cortex. *Nature*, 399(6736):575–579, 1999.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Narasimhan Venkatanath, Debbabi Praneeth, S Channappayya Sumohana, S Medasani Swarup, et al. Blind image quality evaluation using perception based features. In *2015 twenty first national conference on communications (NCC)*, pp. 1–6. IEEE, 2015.

Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 2555–2563, 2023a.

Zhihua Wang, Zhi-Ri Tang, Jianguo Zhang, and Yuming Fang. Toward a blind image quality evaluator in the wild by learning beyond human opinion scores. *Pattern Recognition*, 137:109296, 2023b.

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023.

Tianhe Wu, Jian Zou, Jie Liang, Lei Zhang, and Kede Ma. Visualquality-r1: Reasoning-induced image quality assessment via reinforcement learning to rank. *Advances in Neural Information Processing Systems*, 38:88167–88190, 2026.

Wufeng Xue, Lei Zhang, and Xuanqin Mou. Learning without human scores for blind image quality assessment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 995–1002, 2013.

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

Yuzhe Yang, Liwu Xu, Leida Li, Nan Qie, Yaqian Li, Peng Zhang, and Yandong Guo. Personalized image aesthetics assessment with rich attributes. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19829–19837. IEEE, 2022b.

Ran Yi, Haoyuan Tian, Zhihao Gu, Yu-Kun Lai, and Paul L Rosin. Towards artistic image aesthetics assessment: a large-scale dataset and a new method. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22388–22397. IEEE, 2023.

Zhenqiang Ying, Haoran Niu, Praful Gupta, Dhruv Mahajan, Deepti Ghadiyaram, and Alan Bovik. From patches to pictures (paq-2-piq): Mapping the perceptual space of picture quality. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3572–3582. IEEE, 2020.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14483–14494. IEEE, 2025.

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

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, 2018.

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.

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

Zicheng Zhang, Chunyi Li, Wei Sun, Xiaohong Liu, Xiongkuo Min, and Guangtao Zhai. A perceptual quality assessment exploration for aigc images. In *2023 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pp. 440–445. IEEE, 2023b.

Shun Zhu, Xichen Yang, Yan Zhang, Tianshu Wang, Zhongyuan Mao, Tianyin Li, Zhuoyan Sun, and Xiaobo Shen. Graph-driven domain co-adaptation for cross-domain image quality assessment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 13970–13978, 2026.

Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In *European conference on computer vision*, pp. 392–408. Springer, 2022.

A Extended Ablation and Robustness Analysis

A.1 Transfer across Frozen Backbones

Table 6 applies the same quality-guidance framework to different frozen encoders. The gains are observed not only on RN50 but also on CLIP ViT-B/32, CLIP ViT-L/14, SigLIP Zhai et al. (2023), and a Qwen2.5-VL Bai et al. (2025b) feature readout. The absolute performance varies because these encoders organize image and text features differently, but the consistent direction of improvement suggests that the intervention is not tied to one RN50 implementation.

Table 6: Backbone ablation. Each entry is SRCC/PLCC.

Backbone	LIVEC	KonIQ	BID
CLIP RN50	.718/.749	.645/.721	.757/.787
+ Ours	**.777/.813**	**.757/.809**	**.787/.824**
CLIP ViT-B/32	.640/.647	.568/.622	.658/.664
+ Ours	**.697/.702**	**.622/.675**	**.680/.710**
CLIP ViT-L/14	.546/.543	.523/.549	.557/.547
+ Ours	**.611/.609**	**.567/.611**	**.606/.609**
SigLIP	.535/.579	.443/.482	.423/.475
+ Ours	**.611/.640**	**.565/.602**	**.519/.551**
Qwen2.5-VL feature readout	.302/.288	.197/.211	.271/.274
+ Ours	**.309/.294**	**.216/.232**	**.284/.288**

A.2 Natural–Pure-Distortion Anchor Control

We further test whether an explicit quality-ranked distortion bank is necessary, or whether an extreme naturalness contrast is already sufficient. The control uses unmodified natural images as the high-quality cluster and content-free noise images as the low-quality cluster. We generate noise at eight spatial block scales from 1 to 128 pixels and report the best scale for each target, giving this control a target-wise oracle advantage. As shown in Table 7, the natural–noise direction improves the matched CLIP-IQA baseline on all four datasets, increasing the average SRCC/PLCC from $.651/.684$ to $.721/.754$. Thus, even low-quality anchors without scene content or human quality labels can expose a generic naturalness–degradation tendency in CLIP space.

This coarse contrast is nevertheless not equivalent to the proposed quality prior. At a representative 32-pixel noise scale, its direction has a moderate cosine alignment of $.61$–$.64$ with the paired direction across the four evaluation settings, indicating a shared component rather than an arbitrary axis. However, its preferred noise scale varies across targets, and it remains below the paired construction on every target despite target-wise scale selection, with an average gap of $.050$ SRCC and $.041$ PLCC. This pattern is consistent with pure noise capturing a broad corruption cue, while content-preserving high/low-quality comparisons better suppress scene variation and represent more diverse quality changes. The experiment therefore supports both the simplicity of contrastive direction construction and the transfer advantage of quality-structured natural-image anchors.

Table 7: Ablation of pure-distortion low-quality anchors under a matched diagnostic protocol.

Anchor construction	KonIQ	LIVEC	LIVE	KADID	Average
CLIP-IQA	.643/.718	.718/.749	.770/.767	.474/.504	.651/.684
Natural–noise (oracle scale)	.698/.769	.731/.773	.866/.864	.589/.609	.721/.754
Paired high/low quality (ours)	**.754/.806**	**.777/.813**	**.877/.881**	**.674/.681**	**.771/.795**

A.3 Feature-Fusion Operators

Table 8 compares the proposed residual update with using the directional projection alone and with two direct feature-fusion baselines. The projection alone discards most of CLIP's complementary representation and performs poorly. Concatenation and averaging retain more information, but the

base-plus-directional-residual update is the most consistent across datasets. This supports amplifying the quality component while retaining the original feature rather than replacing it.

Table 8: Feature-fusion ablation. Each entry is SRCC/PLCC.

Fusion	LIVEC	KonIQ	KADID	LIVE
Projection only	.469/.511	.499/.536	.465/.537	.298/.420
Concatenation	.770/.805	.740/.801	.646/.656	.873/.871
Mean fusion	.768/.792	.740/.800	.645/.655	.871/.869
Residual amplification	**.777/.813**	**.757/.809**	**.674/.681**	**.881/.884**

A.4 Direction Perturbation Controls

To distinguish quality guidance from an arbitrary one-direction feature perturbation, we replace $\mathbf{v}_q$ with a random direction, shuffle its groundtruth mos to get the direction, or add Gaussian noise. Table 9 reports both the isolated directional coordinate and the complete amplified readout. Random and shuffled axes are substantially weaker, whereas a noisy version of the learned direction remains close to the original. The useful signal is therefore direction-specific but tolerant to moderate estimation error.

Table 9: Quality-direction perturbation. Each entry is SRCC/PLCC.

(a) Directional coordinate only

Axis	KonIQ	LIVEC	KADID	LIVE
Random	.054/.165	.253/.291	.068/.105	.057/.293
Shuffled	.441/.480	.408/.418	.518/.527	.604/.631
Ours	**.659/.697**	**.670/.704**	**.660/.665**	**.819/.832**
Ours + noise ($\sigma = 0.3$)	.652/.694	.668/.702	.659/.664	.817/.831

(b) Complete amplified readout

Axis	KonIQ	LIVEC	KADID	LIVE
Random	.664/.739	.727/.750	.513/.538	.797/.793
Shuffled	.668/.745	.684/.728	.565/.584	.805/.797
Ours	**.757/.809**	.777/.813	**.674/.681**	**.881/.884**
Ours + noise ($\sigma = 0.3$)	.753/.807	**.777/.813**	.672/.680	.879/.875

A.5 Source-Distortion Diversity

Figure 6 varies the number of KADIS distortion types used to construct the prior. Transfer improves rapidly when moving from one distortion to a diverse source set and becomes stable at approximately 15–25 types. This behavior is consistent with diverse source degradations reducing the influence of a single distortion-specific factor.

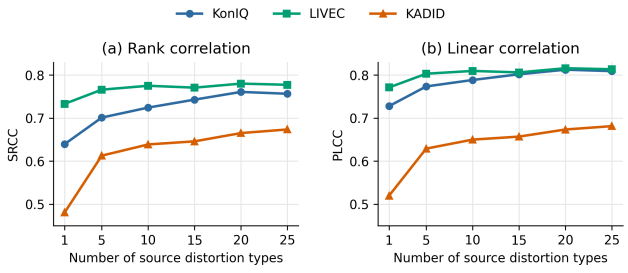


Figure 6: Effect of source-distortion diversity. Performance improves rapidly as the source bank covers more degradation types and becomes stable around the full bank.

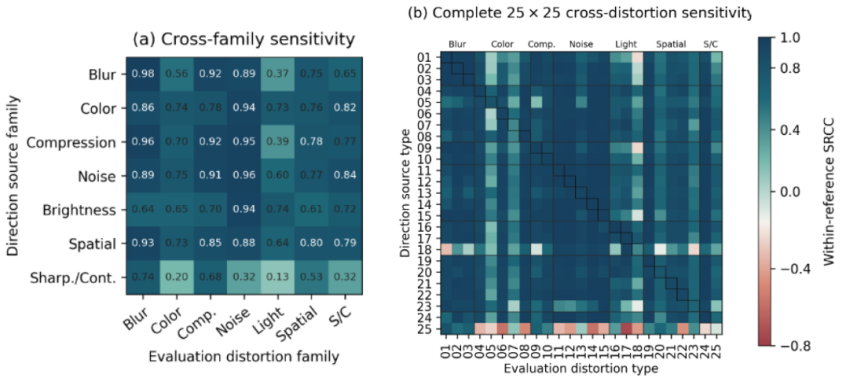


Figure 7: Distortion-specific and shared sensitivity measured by within-reference severity SRCC. (a) Sensitivity aggregated over seven distortion families. (b) Complete matrix for 25 direction-source and evaluation-distortion pairs; outlined diagonal cells denote matched distortions, and dark separators mark family boundaries. The diagonal tendency reflects distortion-specific cues, whereas substantial off-diagonal responses reveal transferable perceptual structure.

A.6 Distortion-Specific and Shared Sensitivity

The complete 25×25 analysis constructs a direction from each KADID-10K distortion type and evaluates its severity ordering on every other type. Figure 7 visualizes the grouped and complete matrices, while Table 10 summarizes the primary SRCC results. A matched direction is stronger on average than an unmatched direction, but the multi-distortion directions are more reliable across the complete bank. In particular, the direction obtained by averaging all 25 distortion-specific directions attains the highest mean sensitivity and avoids near-zero responses under the recorded threshold. Thus, the matrix contains both type-specific diagonal structure and broadly transferable off-diagonal structure.

Table 10: Summary of the 25×25 distortion-direction sensitivity matrix. Sensitivity is within-reference severity SRCC. Hit@1 measures how often the matched direction is the strongest among all single-distortion directions.

Diagnostic	Result
Matched-direction mean	0.852
Cross-distortion mean	0.748
Matched direction Hit@1 / mean rank	12/25 / 4.68
Global pooled direction mean	0.846
Mean of 25 directions	**0.868**
Text quality direction mean	0.599

A.7 Source-Domain Composition

We study how the source-domain composition affects the extracted quality prior. Four complementary sources are considered: authentic images from KonIQ-10K (R), AIGC images from AGIQA-3K (G), aesthetic images from AVA (A), and synthetically distorted images from KADID-10K (S). Figure 8 reports the complete target-free source-composition ablation, including individual sources, all source pairs, three-source combinations containing the synthetic source, and the four-source combination. To keep the comparison compact, the horizontal axis lists the source quality directions, the main heatmap uses direction-level aggregation for each multi-source prior, and the bottom $\Delta(\mathrm{D}-\mathrm{F})$ band reports its mean difference from feature-level aggregation. The no-MOS KADIS direction is included as a separate reference. The target-aware 1:9 in-domain direction is excluded here and reported separately in the extension experiments. The readout used in this diagnostic differs from the complete main-table configuration, so comparisons should be made within the figure.

The results reveal both domain preference and cross-domain complementarity. The synthetic source is the strongest individual source for objective IQA, while the aesthetic source is less effective overall but improves the aesthetic targets. Combining AIGC and synthetic sources gives the highest average SRCC of .708, whereas adding all four sources yields .702–.703. Thus, improvement is not monotonic in the number of sources; compatible source information matters more than source count alone. Feature and direction aggregation are also close on average (.691 versus .688), supporting direction averaging as the simpler multi-source implementation. Notably, the no-MOS KADIS prior remains competitive at .685 without using source quality annotations.

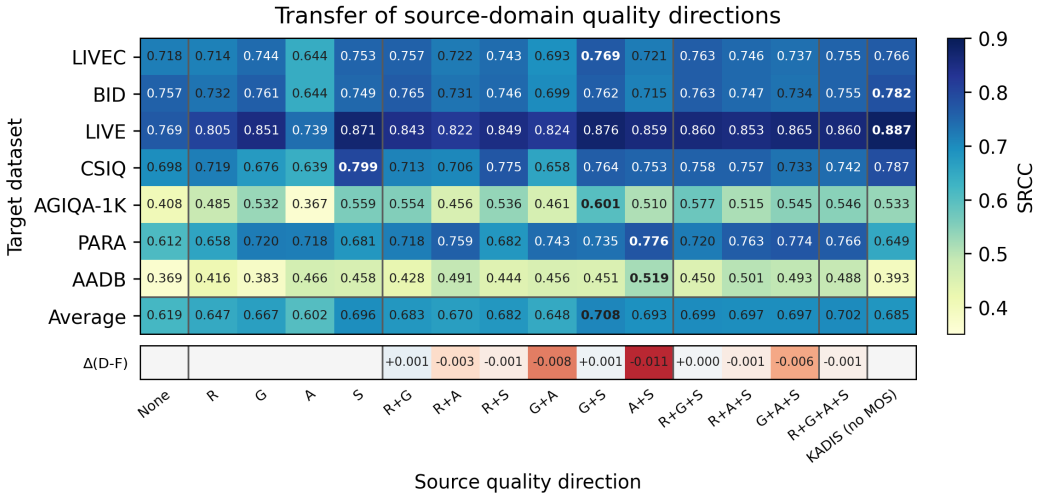


Figure 8: Source-domain composition ablation under a common exploratory protocol. The main heatmap reports SRCC using direction-level aggregation for multi-source priors; the bottom band gives the mean $\Delta(\mathrm{D}-\mathrm{F})$, where positive values favor direction aggregation over feature aggregation. Single-source priors have no aggregation difference. Bold values are target-wise maxima.

A.8 Hyperparameter Sensitivity and Injection Location

Figure 9 varies the amplification factor and source-reference count over broad ranges. A moderate positive α consistently improves the frozen baseline, while excessively large values gradually trade semantic discrimination for quality sensitivity. Most of the source-size gain is already obtained with 10 references, and performance is stable around the default $R = 100$. Importantly, $\alpha = 2$ and $R = 100$ are fixed across all target datasets and are not the per-dataset optima; several synthetic and algorithmic datasets continue improving at larger α. An injection-location sweep further shows that applying the update closer to the final CLIP representation is consistently stronger than injecting it into earlier stages, which supports the proposed feature-level intervention.

B Validation of Pristine-Affinity Ranking

B.1 Correlation and Pairwise Ordering Agreement

On the KADID-10K diagnostic set, pristine affinity correlates more strongly with MOS than the negative distortion level, both globally and within each reference in Table 11. Across distortion types, affinity also yields a higher pairwise ordering accuracy. These results validate the design choice retrospectively on a separate bank with human ratings. For cross-type pairs with an MOS difference of at least 0.5, pristine affinity achieves 85.86% ordering accuracy, compared with 79.20% for distortion level. Among pairs tied under the nominal level, affinity remains correct for 80.10%; among pairs for which the level ordering is reversed relative to MOS, affinity remains correct for 75.11%.

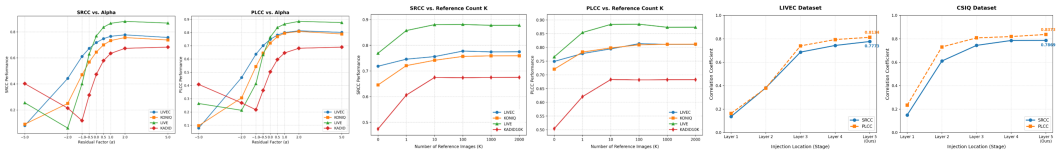


Figure 9: Sensitivity to amplification strength, source-reference count, and injection location.

Table 11: Agreement of pristine affinity and nominal distortion level with MOS.

Surrogate	Global SRCC	Global KRCC	Within-ref. SRCC	Within-ref. KRCC
−Distortion level	.676	.527	$.685 \pm .028$	$.539 \pm .025$
Pristine affinity	**.771**	**.576**	$\mathbf{.784 \pm .039}$	$\mathbf{.597 \pm .040}$

B.2 Anchor Purity and Explained MOS Variance

Table 12 evaluates the Top-50 and Bottom-50 anchors selected within each reference. Pristine-affinity selection creates a larger MOS separation and purer high/low sets than nominal-level selection, approaching the MOS oracle without observing MOS during construction.

Table 12: Quality purity of high- and low-quality anchors on KADID-10K. $P@50$ measures agreement with the MOS-defined Top/Bottom-50.

Selection	High MOS	Low MOS	Gap	$P@50_H$	$P@50_L$
Distortion level	3.798	2.255	1.543	70.4%	73.2%
Pristine affinity	3.893	2.180	1.714	75.6%	74.4%
MOS oracle	4.165	1.916	2.250	100%	100%

When both surrogates are quantized into five bins, distortion level explains 45.4% of MOS variance, while global and within-reference affinity quantiles explain 56.2% and 57.5%, respectively. The advantage of within-reference affinity further supports separating semantic content before constructing the global quality prior.

C Representation-Level Diagnostics

C.1 Directional Energy

Figure 5 in the main paper summarizes the directional-energy analysis. Table 13 provides the corresponding per-dataset values. The source-derived component is more prominent on the three conventional synthetic benchmarks, while the same direction remains consistently above its random control on authentic datasets. The amplified column reports its relative contribution after the default residual update.

C.2 Quantitative Distortion Sensitivity

The lower plots of Figure 4b in the main paper visualize the response to three controlled distortion families. Table 14 provides a complementary dataset-level evaluation on TID2013 and PIPAL. "Quality coordinate" directly reads the projection $\mathbf{f}^\top \mathbf{v}_q$; the other two variants use the same linear image–text quality readout before and after image-side amplification. This is therefore a mechanism diagnostic rather than the complete dual-branch method. The directional coordinate is correlated with quality on both datasets, while incorporating it through amplification gives the strongest SRCC and PLCC. The increased score standard deviation further quantifies the expanded response range observed in the controlled curves.

Table 13: Per-dataset directional energy (%). The random baseline is averaged over random one-dimensional directions.

Dataset	Quality direction	Random direction	Amplified ($\alpha = 2$)	Ratio to random
LIVE	6.96	0.13	18.72	$54.79\times$
CSIQ	5.21	0.11	14.30	$49.26\times$
TID2013	4.86	0.10	13.82	$49.00\times$
LIVEC	0.98	0.08	3.56	$11.93\times$
KonIQ	1.11	0.11	3.98	$9.87\times$
BID	0.98	0.14	3.57	$6.77\times$
FLIVE	0.85	0.09	3.11	$9.16\times$
SPAQ	1.85	0.10	6.36	$18.44\times$

Table 14: Dataset-level complement to the controlled distortion responses. PairAcc denotes within-reference pairwise ordering accuracy, and Std is the standard deviation of the predicted score.

Dataset	Readout	SRCC	PLCC	PairAcc	Std
TID2013	Quality coordinate	.630	.658	**.771**	.166
TID2013	Frozen quality readout	.557	.641	.731	.041
TID2013	Amplified	**.657**	**.693**	.770	.059
PIPAL	Quality coordinate	.285	.308	.625	.133
PIPAL	Frozen quality readout	.369	.360	.659	.032
PIPAL	Amplified	**.392**	**.381**	**.661**	.046

C.3 Component Removal, Isolation, and Amplification

We train no new IQA predictor in this diagnostic experiment. Instead, a fixed probe measures cross-content distortion-severity discrimination from different feature decompositions. Let A be its accuracy and $A_{\text{chance}} = 1/3$. We define normalized quality retention as

$$\text{Retain}(\mathbf{f}') = \frac{A(\mathbf{f}') - A_{\text{chance}}}{A(\mathbf{f}) - A_{\text{chance}}} \times 100\%, \tag{9}$$

where the original CLIP feature $\mathbf{f}$ is set to 100%.

Table 15 reveals an asymmetric role for the one-dimensional quality component. Removing it preserves content accuracy but reduces normalized quality retention to 51%; removing a random direction retains 99%. Conversely, the quality component alone retains 22% of the original quality signal, compared with only 3% for a random one-dimensional component. Amplification increases quality retention to 108% while content accuracy changes from 0.956 to 0.937. These results directly support the description of quality as *latent but underweighted*.

Table 15: Feature-decomposition probe. Content chance is 0.10 and severity chance is $1/3$.

Representation	Content Acc.	Severity Acc.	Quality Retention
Original CLIP	0.956	0.515	100%
Remove quality direction	0.959	0.425	51%
Remove random direction	0.956	0.513	99%
Quality direction only	0.125	0.373	22%
Random direction only	0.148	0.339	3%
Amplify quality direction ($\alpha = 2$)	0.937	**0.529**	108%
Amplify random direction	0.956	0.512	99%

C.4 Representational Dissimilarity Structure

Representational similarity analysis provides a readout-free complement to the preceding probe Kriegeskorte et al. (2008). For each feature decomposition, we construct a representational dissimilarity matrix (RDM) from pairwise feature distances and compare it with content-

and quality-severity model RDMs. Table 16 reports the resulting organization scores and similarity to the original CLIP geometry, while Figure 10 visualizes the corresponding RDMs.

The original representation is dominated by content structure (0.48) and contains weaker quality organization (0.05). Removing $\mathbf{v}_q$ preserves content organization and most of the original geometry, but reduces quality organization to zero; removing a Brayrandom direction leaves all three measures unchanged. Conversely, $\mathbf{v}_q$ alone produces a quality-organization score of .20, four times that of the complete CLIP representation and twenty times that of a random direction, while retaining little content structure (.05). These complementary controls show that $\mathbf{v}_q$ concentrates a disproportionate amount of quality-related structure without implying complete semantic-quality disentanglement. This measure reflects organization by nominal distortion severity rather than agreement with human-rated quality; severity levels are not perceptually calibrated across distortion types.

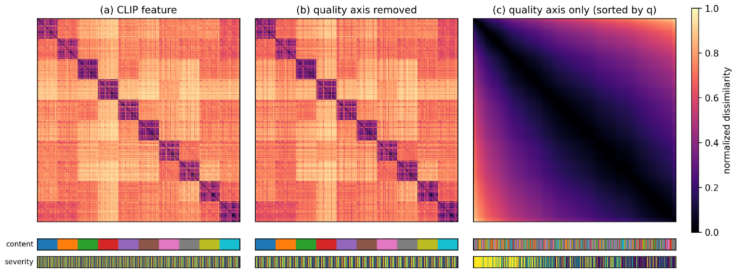


Figure 10: Representational dissimilarity matrices of the original, quality-direction-removed, and quality-direction-only features. Removing $\mathbf{v}_q$ preserves the dominant content geometry but weakens quality organization, whereas isolating $\mathbf{v}_q$ reveals stronger quality-related structure with limited semantic organization.

Table 16: Representational dissimilarity analysis. Content and quality organization measure agreement with the corresponding model RDMs; the final column measures agreement with the original CLIP RDM.

Representation	Content organization	Quality organization	Similarity to original
Original CLIP	0.48	0.05	1.00
Remove quality direction	0.50	0.00	0.96
Remove random direction	0.48	0.05	1.00
Quality direction only	0.05	0.20	0.14
Random direction only	0.19	0.01	0.24

C.5 Cross-Dataset Direction Geometry

The main paper reports the aggregate alignment of directions independently constructed from several IQA datasets. Table 17 provides the per-dataset projection onto the first principal direction and the energy retained by the common three-dimensional basis. Every direction retains at least 86% of its energy in the same three-dimensional span, while the variation in the first-component alignment confirms that the data do not collapse to one identical axis.

Relation to deployable source directions. We project the independently constructed KADIS and KADID source directions onto the diagnostic basis obtained from seven dataset directions, without including either source direction in the SVD. As shown in Figure 11, the two priors exhibit nearly identical overlap with this structure: their mean cosine similarities are .633 and .637, and the first three components retain .708 and .711 of their energy, respectively, compared with approximately .008 for the 95th-percentile random direction. Their unamplified quality coordinates correlate positively with subjective quality across all seven datasets after harmonizing score polarity, with mean SRCC of .650 and .670. Thus, the MOS-free KADIS construction recovers a quality-oriented prior closely comparable to the MOS-guided KADID construction. We use KADIS as the default source because its diverse synthetic bank permits inexpensive construction without subjective annotations.

Table 17: Per-dataset organization of independently constructed quality directions. The basis is obtained by uncentered SVD over the seven listed directions.

Dataset	Squared cosine with PC1	Energy retained by top 3
LIVE	0.567	0.939
CSIQ	0.631	0.861
TID2013	0.626	0.879
LIVEC	0.663	0.932
KonIQ	0.685	0.925
BID	0.696	0.928
FLIVE	0.734	0.894

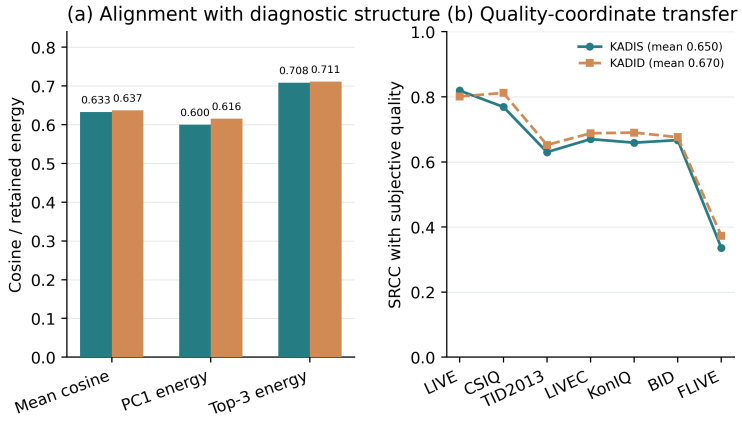


Figure 11: Held-out source directions and cross-dataset quality structure. (a) Mean signed cosine similarity to seven dataset directions, energy along the first principal component, and energy retained by the first three components of the dataset-only SVD. (b) SRCC between each unamplified source quality coordinate and subjective quality. Score polarity is harmonized so that higher denotes better quality; target annotations are used only for this post-hoc analysis.

D Semantic Preservation Analysis

D.1 Zero-Shot Classification under Amplification

Figure 4b top line in the main paper reports the complete ImageNet, VOC2007, and image–text alignment sweeps. Table 18 lists the underlying absolute measurements for reproducibility. Moderate amplification retains most semantic behavior, whereas the larger $\alpha = 4$ setting reveals the expected trade-off between perceptual gain and the original CLIP geometry.

Table 18: Absolute semantic-preservation measurements across amplification strengths. Cons. denotes prediction consistency with $\alpha = 0$, and Cos. denotes cosine similarity between original and enhanced features.

	ImageNet				VOC2007				Image–text	
α	Top-1	Top-5	Cons.	Cos.	mAP	Present Acc.	Cons.	Cos.	Matched sim.	Margin
0.0	0.5630	0.8354	1.000	1.000	0.7494	0.8714	1.000	1.000	0.2256	0.00231
0.5	0.5617	0.8345	0.984	0.999	0.7453	0.8691	0.995	0.999	0.2250	0.00231
1.0	0.5598	0.8338	0.968	0.996	0.7379	0.8681	0.990	0.995	0.2241	0.00227
2.0	0.5553	0.8309	0.934	0.984	0.7113	0.8655	0.981	0.981	0.2215	0.00210
4.0	0.5408	0.8209	0.871	0.946	0.6135	0.8590	0.963	0.939	0.2142	0.00144

D.2 Image–Text Alignment and Component Retrieval

For qualitative retrieval, Figure 12 shows that the one-dimensional quality component retrieves images with similar fidelity or distortion characteristics, whereas the orthogonal component retrieves

semantically related content. Because a one-direction decomposition cannot strictly disentangle all semantic and perceptual factors, we use these examples as visualization rather than quantitative proof of complete separation.



Figure 12: Retrieval behavior of the original feature and its quality-directional and orthogonal components. The examples illustrate complementary quality- and content-oriented neighborhoods without asserting complete disentanglement.

E Computational Cost and Storage

Table 19 reports a batch-size-one profile on an NVIDIA H20. The proposed projection and residual update require 0.231 ms, accounting for only 2.5% of total inference latency; the frozen CLIP forward remains the dominant cost. Constructing a direction from 100 references takes approximately three minutes without caching, and the final averaged prior occupies only 9.42 KB. Since this stage is performed once, it does not affect deployment latency.

Table 19: Computation profile with CLIP RN50 and batch size one.

Operation	Time	Fraction
Test-image preprocessing	0.782 ms	8.4%
Frozen CLIP forward	8.337 ms	89.2%
Projection and residual update	0.231 ms	2.5%
End-to-end inference	9.350 ms	100%
Offline construction, 100 references	2.98 min	–
Cached averaged prior	9.42 KB	–

For one source reference, generating 125 distortions takes 0.380 s, preprocessing and encoding the 126 images takes 1.329 s, and affinity sorting and aggregation take 0.079 s, for a total of 1.788 s. Table 20 reports the measured linear scaling of this offline step and the two evaluated caching strategies. Storing only the averaged prior keeps the file size fixed, whereas retaining every reference-level anchor grows linearly. The measured peak memory of the CLIP forward pass is 285.43 MB.

Table 20: Offline source construction and cache scaling. "Averaged" stores the final source prior; "individual" retains every reference-level anchor.

References	Construction time	Averaged cache	Individual cache
1	1.79 s	9.42 KB	9.42 KB
10	17.88 s	9.42 KB	81.42 KB
100	2.98 min	9.42 KB	801.42 KB
1,000	29.80 min	9.42 KB	7.81 MB
10,000	4.97 h	9.42 KB	78.13 MB

F Qualitative Representation Analysis

F.1 Embedding Organization

The t-SNE Van der Maaten & Hinton (2008) analysis in Figure 13 visualizes the same features by semantic content and by distortion severity before and after amplification. Amplification enlarges the quality-related separation within the pre-existing semantic organization rather than replacing that organization. The accompanying RDMs Kriegeskorte et al. (2008) in Figure 10 provide a complementary view.

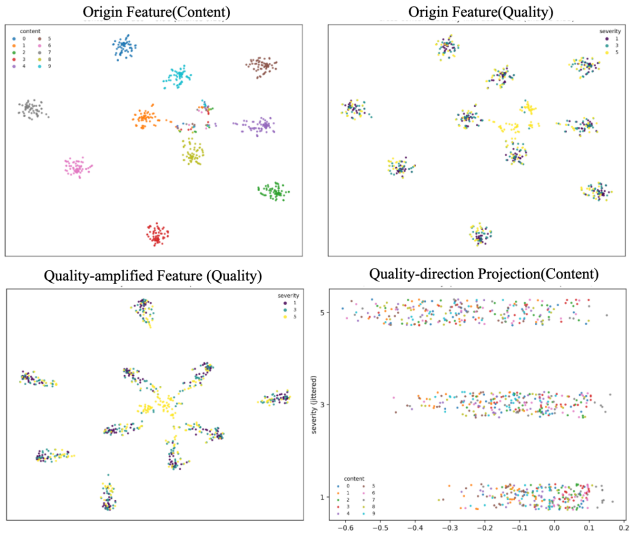


Figure 13: t-SNE visualization of original and amplified CLIP features, colored by image content and quality severity. Amplification increases quality-related separation while largely preserving the dominant semantic organization.

F.2 gMAD Comparisons

The gMAD Ma et al. (2016) study in Figure 14 compares the proposed method with training-free MA-CLIP and KADID-trained LIQE. The selected image pairs expose differences in the quality ordering produced by image-level manipulation, a supervised quality model, and the proposed feature-level amplification. We treat these pairs as qualitative comparisons rather than as an additional aggregate benchmark.

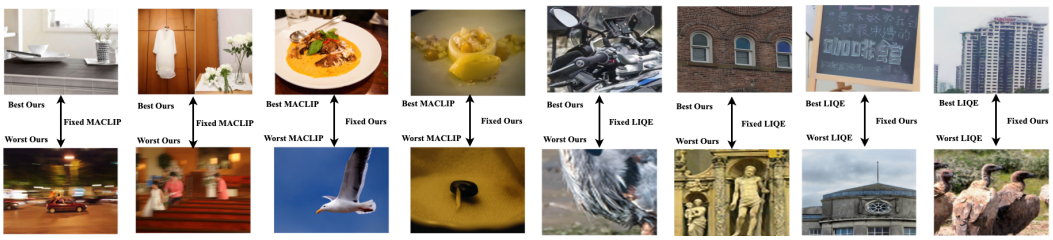


Figure 14: Representative gMAD comparisons against a training-free CLIP baseline and a source-trained IQA model. Each pair is interpreted qualitatively rather than as an aggregate benchmark.

F.3 Spatial Response and Image–Text Matching Maps

Figure 15 separates two complementary spatial diagnostics. The gradient-based response maps in left examine where the quality score is locally sensitive, whereas the image–text matching maps

in panel (b) visualize the spatial evidence contributing to quality-prompt similarity. Both diagnostics cover blur, noise, compression blocks, and local generation artifacts. They provide qualitative evidence about the modified representation.

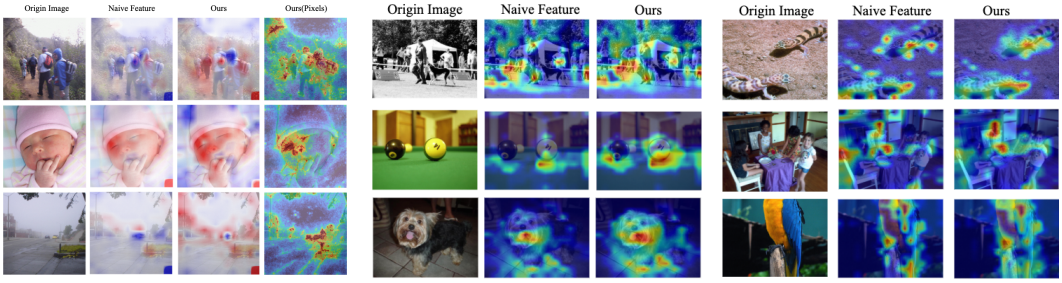


Figure 15: Complementary spatial diagnostics before and after feature amplification. (a) Gradient-based responses visualize local score sensitivity. (b) Image–text matching maps visualize the spatial evidence associated with quality prompts.

G Extensions Beyond Target-Data-Free IQA

G.1 Target-Aware Direction Correction

The main paper deliberately studies target-data-free deployment. As an exploratory extension, a small labeled subset from a target domain can correct the source direction, after which the remaining images are evaluated without further optimization. Table 21 reports SRCC/PLCC using a 1:9 reference-to-test split and correction strength $\gamma = 2$. The correction is especially useful on KonIQ, FLIVE, PIPAL, and the AIGC datasets, suggesting that the source direction captures a transferable base component while a small target correction can model the residual domain-specific component. BID is a useful counterexample: its corrected direction is slightly weaker, showing that limited target supervision is not uniformly beneficial.

Table 21: Target-aware direction correction. Each entry is SRCC/PLCC; the corrected variant is outside the target-data-free setting of the main paper.

Method	LIVEC	KonIQ	SPAQ	BID	FLIVE
Frozen CLIP	.718/.749	.645/.721	.772/.779	.757/.787	.312/.443
Ours	.777/.813	.757/.809	.819/.825	**.787/.824**	.381/.493
Ours + in-domain correction	**.782/.820**	**.802/.836**	**.847/.853**	.783/.819	**.428/.529**

Method	LIVE	CSIQ	TID2013	KADID	PIPAL	AGIQA-1K	AGIQA-3K
Frozen CLIP	.769/.765	.698/.767	.558/.652	.474/.503	.369/.374	.522/.663	.652/.715
Ours	.881/.884	.787/.837	.661/.719	.674/.681	.386/.408	.606/.721	.724/.763
Ours + in-domain correction	**.905/.905**	**.791/.841**	**.687/.733**	**.695/.700**	**.409/.429**	**.691/.783**	**.794/.841**

G.2 Within-Domain Direction Estimation

As a complementary 1:9 reference-to-test experiment, we construct the direction only from the reference split of each target dataset and evaluate on its disjoint test split. Table 22 reports mean SRCC/PLCC across repeated splits. These results measure how accurately the same closed-form construction can adapt when a small in-domain reference set is permitted.

G.3 Aesthetic-Quality Direction Correction

Objective fidelity and aesthetic preference are related but not identical. Table 23 makes this boundary explicit. The KADIS objective-quality direction is not reliably beneficial on aesthetic datasets;

Table 22: Within-domain direction estimation with a 1:9 reference-to-test split. Each entry is mean SRCC/PLCC across repeated splits.

Method	KonIQ	KADID	PIPAL	AGIQA-3K	AVA Murray et al. (2012)
Frozen CLIP	.646/.722	.471/.501	.369/.373	.651/.715	.343/.364
+ In-domain direction	.775/.816	.687/.692	.395/.421	.785/.854	.514/.529

an AVA-derived aesthetic direction constructed from 1,000 labeled images or an in-domain correction is required for consistent gains. The result therefore supports attribute-dependent direction construction rather than treating one objective-IQA direction as universal.

Table 23: Transfer to aesthetic-quality assessment. Each entry is SRCC/PLCC. AVA and in-domain corrections use labeled reference data and lie outside the target-data-free setting.

Method	AVA	BAID Yi et al. (2023)	AADB Kong et al. (2016)	TAD66K He et al. (2022)	PARA Yang et al. (2022b)
Frozen CLIP	.299/.322	.167/.149	.483/.500	.204/.220	.623/.670
KADIS direction	.267/.301	.182/.170	.381/.400	.115/.148	.637/.783
+ AVA direction	.408/.430	.174/.163	**.552/.568**	**.309/.331**	.745/.776
+ in-domain direction	**.408/.430**	**.330/.313**	.530/.542	.286/.307	**.783/.818**

G.4 Compatibility with Parameter-Efficient Training

We further insert the quality prior into a CLIP-IQA+ pipeline trained on KonIQ-10K. This experiment lies outside our training-free setting and instead tests whether directional amplification remains useful when combined with supervised prompt adaptation.

Table 24: Compatibility of different feature priors with a KonIQ-trained CLIP-IQA+ pipeline under our preprocessing protocols. Each entry is SRCC/PLCC.

Method	KonIQ	LIVEC	SPAQ	BID
CLIP-IQA+	.848/.876	.805/.833	.862/.863	.808/.828
CLIP-IQA+ + MA-CLIP prior	.821/.857	.782/.815	.837/.841	.787/.809
CLIP-IQA+ + Ours	**.856/.884**	**.807/.837**	**.867/.869**	**.813/.833**

As shown in Table 24, our prior improves CLIP-IQA+ on all four datasets and the three cross-dataset tests. In contrast, the MA-CLIP Liao et al. (2026) prior reduces performance throughout. This contrast suggests that the limitation of cosine-based quality prediction cannot be explained generally by discarded feature magnitude alone: if magnitude omission were the dominant transferable cause, its correction would be expected to benefit, or at least remain stable in, the CLIP-IQA+ pipeline. We do not interpret this result as invalidating MA-CLIP in its original setting; rather, it shows that magnitude awareness is not a universally compatible remedy. The consistent gains from our prior instead support the broader applicability of explicitly identifying and amplifying a quality-relevant direction across both frozen and trainable readouts.

Table 25 further compares a trainable layer-wise version with standard supervised IQA models under the recorded KonIQ-10K training protocol. We follow the CLIP-IQA+ KonIQ-10K training and evaluation protocol in full, replacing only the model. The 4,869 trainable parameters comprise quality-direction vectors at RN50 layers 1–4 and the final image embedding ($256 + 512 + 1024 + 2048 + 1024 = 4864$ parameters), plus five scalar amplification factors. Both CLIP encoders and the text anchors remain frozen; only these direction vectors and amplification factors are optimized under the CLIP-IQA+ training protocol. These numbers are presented only as an extension: the main method and claims remain training-free.

The layer-wise variant uses 0.04869 MB of trainable parameters, 94.1% fewer than CLIP-IQA+, while improving its results on all three evaluations. It also achieves the best KonIQ-10K and SPAQ results in this comparison, whereas MANIQA remains stronger on LIVEC. This provides complementary evidence that the proposed guidance can serve as a lightweight trainable component without changing the scope of the main training-free method.

Table 25: Layer-wise trainable extension under KonIQ-10K supervision. Each entry is SRCC/PLCC.

Method	Trainable parameters (MB)	KonIQ	LIVEC	SPAQ
DBCNN	15.30	.903/.919	.756/.774	.847/.854
MANIQA	135.75	.893/.913	**.826/.840**	.817/.814
MUSIQ	27.10	.865/.896	.789/.829	.848/.860
CLIP-IQA+	0.08192	.885/.901	.771/.811	.847/.853
Ours (layer-wise)	**0.04869**	**.910/.923**	.811/.831	**.854/.861**

G.5 Coordinate-Aware Feature Enhancement Beyond IQA

More interestingly, our method is not restricted to IQA. We instantiate the same direction discovery and amplification template for anomaly detection Zou et al. (2022); Bergmann et al. (2019) using flawless/defective anchors and for AIGC detection Chen et al. (2025); Guillaro et al. (2025); Cavia et al. (2024); Cozzolino et al. (2024) using real/generated anchors. As shown in Table 26, the resulting plug-and-play intervention consistently strengthens the corresponding frozen CLIP readout without modifying or retraining the backbone. Together, these results suggest a broader perspective: task-relevant information may already be encoded in foundation representations but under-emphasized by their native similarity geometry, and lightweight subspace reweighting offers a simple mechanism for exposing such latent properties. For anomaly detection, we construct score with the prompts "flawless {class}" and "defective {class}." For AIGC detection, we use $\alpha = 1$, and ensembles of five real-photo and five AI-generated-image discribed prompts. No detector backbone is updated in either experiment.

Table 26: Preliminary extensions beyond IQA. Anomaly detection reports AUC; AIGC detection reports AUC/accuracy.

Method	ViSA	MVTec	DDA-COCO	B-Free	Wild-RF	SynthWildX
Naive CLIP	.607	.731	.522/.326	.615/.700	.889/.829	.823/.795
Directional amplification	**.618**	**.742**	**.525/.382**	**.759/.769**	**.961/.902**	**.924/.881**

H Limitations and Future Directions

Our method deliberately uses one source-derived direction as a simple approximation to a richer, partially domain-dependent quality structure. The cross-dataset analysis assigns substantial energy to the second component, and the source-replacement study shows that no source is optimal for every target. The 25×25 distortion analysis further records weak sensitivity for several color- and contrast-related settings, while the gain on the algorithmically distorted PIPAL set is smaller than on conventional synthetic datasets. Moreover, an objective-quality direction does not directly solve aesthetic assessment, and a global pooled feature may under-represent localized defects. The two representative failures in Figure 16 expose a symmetric limitation of the fixed linear update. A high-quality image with a negative quality coordinate is pushed toward lower predicted quality, whereas a lower-quality image with a positive coordinate is pushed upward. In both cases, the original CLIP rank is relatively close to the ground truth, but amplification magnifies the incorrect projection sign rather than correcting it. These examples illustrate the sample-level failure mechanism without attributing it to a specific visual category; sample-adaptive confidence control or multiple complementary directions may provide more flexible alternatives.

The direction also depends on anchor quality. MOS-guided anchors are often stronger when reliable ratings are available, whereas pristine affinity is inexpensive but remains an imperfect surrogate. The primary evidence is restricted to similarity-based scoring in frozen CLIP; the other backbones and non-IQA tasks are exploratory. Natural next steps are therefore to learn a compact multi-direction basis, adaptively mix source directions, apply guidance to spatial tokens, improve the no-MOS fidelity surrogate, and test whether the same principle extends to explanatory multimodal outputs.



Figure 16: Two representative failure modes of directional amplification. A negative quality coordinate causes a high-quality image to be underestimated, while a positive coordinate causes a lower-quality image to be overestimated.

I Reproducibility and Evaluation Protocol

I.1 More Detail Data Source Information and Protocols

We report Spearman rank-order correlation coefficient (SRCC) and Pearson linear correlation coefficient (PLCC). The same source-derived direction, prompt ensemble, and amplification factor are applied to every target dataset. No target image or target annotation is used for direction construction, model selection, or hyperparameter selection. Our default variant, *Ours-KADIS*, uses pristine–distortion affinity to construct anchors without MOS. The complementary *Ours-KADID* variant uses the high- and low-MOS ends of KADID-10K to measure the benefit of source supervision and is not evaluated on KADID-10K itself. Both variants require no parameter optimization at deployment.

For the synthetic-to-authentic comparison, we report only the common target intersection of LIVEC, KonIQ-10K, and BID and transcribe competing results from their published comparison tables rather than reimplementing them. Traditional, CLIP-IQA+, and established cross-domain rows follow SynDR-IQA Li et al. (2026a) and PreSTA Li et al. (2026b); GDCIQA Zhu et al. (2026) is taken from its original AAAI paper. The VLM and reasoning rows follow Tool-IQA Qin et al. (2026), its VisualQuality-R1 Wu et al. (2026) row is the Qwen3 reproduction reported by Tool-IQA. The *Target* column indicates whether target-domain information is accessed before evaluation. The access type is method-dependent: conventional UDA methods use unlabeled target images, whereas the main PreSTA-S comparison estimates its target preference representation from full-set target features and quality labels before selecting a target-specific source subset. Ours-KADID uses the KADID MOS only to construct the closed-form source direction, while Ours-KADIS is separately marked as the MOS-free KADIS construction; neither variant accesses target images or labels. The broader training-free baselines are reproduce under the same test protocol of our methods.

I.2 More Implementation details

The default backbone is RN50 and all of its parameters remain frozen. To preserve distortion evidence, we do not apply the standard center crop. An image is converted to RGB and retains its original aspect ratio; if its shorter side exceeds 448 pixels, the shorter side is resized to 448 using Lanczos interpolation. The resulting tensor is normalized with the original CLIP statistics. The attention-pooling positional term is omitted to support variable-resolution inputs, following the same implementation in all baselines Wang et al. (2023a); Liao et al. (2026) and variants. For each image x, the encoder output is ℓ_2-normalized. The positive and negative text prototypes are constructed by averaging the individually normalized prompt features and normalizing the average again. The enhanced text contrast and the source direction are cached once, so inference on a target image requires only one frozen image-encoder pass and a quality-directional residual update.

The target-data-free IQA experiments in Sec. 4.1 evaluate each target benchmark in full, without a target-domain training split. The source direction and quality prompts are shared across targets, and text-anchor calibration uses $\tau_a = 0.5$ throughout. For the direction-construction comparison in Table 4, the CLIP pseudo-score split uses the main method's quality prompts and selects the top-

and bottom-K candidates rather than applying a threshold. The distortion-level split treats levels 1–2 as high quality and levels 4–5 as low quality. The SVD control takes the leading right singular vector of the high-minus-low feature differences and orients its sign using the text quality contrast.

The controls in Appendix A retain the main preprocessing protocol. ViT-based CLIP and SigLIP omit positional embeddings, while the For the Qwen2.5-VL-3B-Instruct feature readout in Table 6, we mean-pool the visual-encoder outputs and represent each quality prompt by the final-layer hidden state of its last token. We average the normalized prompt features within each quality group and score the normalized image feature by its positive-minus-negative cosine similarity, before applying our quality-direction enhancement(Table 6). The injection-location study constructs a direction separately at RN50 layers 1–4 and at the final attention-pooled representation (Figure 9). The fusion and perturbation controls modify the image representation while keeping the text anchors fixed; concatenation duplicates the text representation to match the doubled image dimension (Tables 8 and 9). In the CLIP-IQA+ compatibility comparison, MA-CLIP retains its magnitude transformation and confidence fusion; only its prompts are replaced with the CLIP-IQA+ prompts (Table 24).

For the representation diagnostics in Appendix C, the feature-decomposition probe uses ten randomly sampled KADID reference contents and three distortion severities (Table 15). The RDM analysis uses ten KADIS references, each with 25 distortion types at levels 1, 3, and 5. It compares Euclidean feature dissimilarities with content-identity and nominal-severity model matrices using Spearman correlation over their upper triangles (Table 16). In the semantic-preservation analysis of Appendix D, image–text measurements use the same ImageNet-1K validation images as the classification sweep, with fixed class-text embeddings. The margin is the correct-class cosine similarity minus the largest incorrect-class cosine similarity (Table 18). The spatial diagnostics use RN50 layer-4 gradient–activation responses for the quality coordinate and Grad-CAM of the positive-minus-negative quality-prompt score for image–text matching (Figure 15).

For the extensions in Appendix G, the trainable layer-wise variant learns a direction and amplification gain at each RN50 stage and at the final pooled representation while keeping the CLIP backbone frozen (Table 25). Anomaly detection constructs a separate direction for each category from normal training images and their synthetically corrupted counterparts: ViSA uses 100 normal images per category with $\alpha = 2$, and MVTec uses 50 with $\alpha = 1$. For AIGC detection, real and generated anchors are drawn from the corresponding benchmark and excluded from its evaluation images. The number of anchors per class is 1,000 for DDA-COCO, 50 for B-Free, and 100 for Wild-RF and SynthWildX. We use $\alpha = 1$ for DDA-COCO, B-Free, and Wild-RF, and $\alpha = 2$ for SynthWildX. The AIGC accuracies in Table 26 apply a fixed zero threshold to the generated-minus-real score.

I.3 Source-Direction Construction

For each sampled KADIS pristine reference, we encode the pristine image and all 125 distorted variants Lin et al. (2019). Candidates are sorted by cosine similarity to the pristine feature. The Top-K and Bottom-K feature means form reference-specific high- and low-affinity anchors, which are then averaged across references. The final direction is the normalized difference between the two global centroids. We use $K = 50$ candidates at each end. The default evaluation samples $R = 100$ references and repeats this process 15 times. This procedure uses the known pristine–distortion grouping but no subjective quality score.

For the complementary MOS-guided source, we encode KADID-10K and average the 2,000 highest- and 2,000 lowest-rated source features before taking their normalized centroid difference. This variant quantifies the benefit of accurate source supervision; it does not alter the target-data-free inference protocol.

Unless otherwise specified, the amplification factor is fixed to $\alpha = 2$ for every target domain. Neither R, K, nor α is selected separately for a target dataset. Main-table KADIS results report the median over the 15 independently sampled source sets. Table 27 reports the standard deviation and observed half-range over the 15 independently sampled KADIS source sets. The variation is small relative to the improvement over frozen CLIP, indicating that the result is not driven by a fortunate choice of pristine references.

Table 28 collects the complete target-free results of the two source constructions used in the paper. The MOS-guided KADID direction is beneficial when source ratings are available, while the KADIS

Table 27: Variation of SRCC/PLCC over 15 KADIS source samplings.

Dataset	LIVE	CSIQ	TID2013	KADID	PIPAL	LIVEC
Std.	.003/.004	.001/.002	.001/.001	.003/.003	.002/.001	.003/.007
Half-range	.003/.006	.002/.003	.003/.002	.004/.003	.004/.003	.004/.007

Dataset	KonIQ	SPAQ	BID	FLIVE	AGIQA-1K	AGIQA-3K
Std.	.005/.003	.003/.002	.004/.004	.002/.002	.003/.002	.002/.002
Half-range	.004/.004	.003/.003	.005/.005	.002/.002	.006/.004	.001/.002

Table 28: Complete target-free results for the KADIS and KADID source directions. Each entry is SRCC/PLCC; KADID-10K is omitted for the KADID source because it is the construction dataset.

(a) Synthetic, algorithmic, and authentic benchmarks

Source prior	LIVE	CSIQ	TID2013	KADID	PIPAL	LIVEC
KADIS affinity (no MOS)	.881/.884	.787/.837	**.661/.719**	.674/.681	.386/.408	.777/.813
KADID MOS	.843/.851	**.796/.841**	.651/.712	–	**.402/.419**	**.781/.815**

(b) Authentic and AIGC benchmarks

Source prior	KonIQ	SPAQ	BID	FLIVE	AGIQA-1K	AGIQA-3K
KADIS affinity (no MOS)	.757/.809	.819/.825	**.787/.824**	.381/.493	.606/.721	**.724/.763**
KADID MOS	**.775/.823**	**.822/.828**	.777/.812	**.402/.506**	**.619/.725**	.711/.756

affinity direction remains effective without MOS and is the default construction. Neither source is uniformly strongest, which motivates reporting both without conflating them.

I.4 Text Prompts

The positive prompt set for IQA is

$$\mathcal{P}^+ = \{\text{“Good image”}, \text{“Sharp image”}, \text{“sharp edges”}, \text{“High resolution image”}, \text{“Noise-free image”}\}, \tag{10}$$

and the negative prompt set is

$$\mathcal{P}^- = \{\text{“Bad image”}, \text{“Blurry image”}, \text{“blurry edges”}, \text{“Low resolution image”}, \text{“noisy image”}\}. \tag{11}$$

The same prompts of CLIPIQA Wang et al. (2023a) and MA-CLIP Liao et al. (2026) are used for every IQA dataset.

The positive prompt set for AIGC is

$$\mathcal{P}^+ = \{\text{“Real photo”}, \text{“Authentic photograph”}, \text{“Camera captured image”}, \text{“Natural picture”}, \text{“Genuine photograph”}\}, \tag{12}$$

and the negative prompt set is

$$\mathcal{P}^- = \{\text{“AI generated image”}, \text{“Synthetic image”}, \text{“Fake image”}, \text{“AIGC image”}, \text{“Computer generated picture”}\}. \tag{13}$$