

Article

Spectral Robustness Mixer: Cross-Scale Neck for Robust No-Reference Image Quality Assessment

Bader Rasheed ^{1,2,*} , Anastasia Antsiferova ^{1,3,4}  and Dmitriy Vatolin ^{1,3,4} ¹ Laboratory of Innovative Technologies for Processing Video Content, Innopolis University, Innopolis 420500, Russia² Research Center of the Artificial Intelligence Institute, Innopolis University, Innopolis 420500, Russia³ Institute for System Programming, Russian Academy of Sciences, Moscow 109004, Russia⁴ Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University, Moscow 119991, Russia

* Correspondence: b.rasheed@innopolis.university

Abstract

No-reference image quality assessment (NR-IQA) models achieve high correlation with human mean opinion scores (MOS) on clean benchmarks, yet recent work shows they can be highly vulnerable to small adversarial perturbations that severely degrade ranking consistency, including in black-box settings. We introduce the Spectral Robustness Mixer (SRM), a lightweight *neck* inserted between an NR-IQA backbone and regression head, designed to reduce adversarial sensitivity without changing the dataset, label format, or target metric. SRM couples (i) deep-to-shallow cross-scale fusion via a Nyström low-rank attention surrogate, (ii) ridge-conditioned landmark kernels with ridge regularization, solved via numerically stable small-matrix factorization (SVD/LU) to improve conditioning, and (iii) variance-aware entropy-regularized fusion gates with a bounded gain cap to limit gradient amplification. We evaluate SRM on TID2013 and KonIQ-10k under a white-box ℓ_∞/ℓ_2 attack ensemble that includes per-image regression objectives and a correlation-aware pairwise inversion objective (a ranking-inspired surrogate for correlation inversion), with expectation-over-transformation (EOT) and anti-gradient masking checks. At $\epsilon = 4/255$ (ℓ_∞), SRM improves *worst-case* robust Spearman's rank-order correlation coefficient (SROCC; defined as the *minimum* over our fixed attack ensemble) by an *absolute* 0.06–0.08 SROCC points (i.e., correlation-coefficient units, not percentage gain) across datasets/backbones, while keeping clean SROCC within 0.00–0.01 of the baseline. We observe similar trends for Pearson linear correlation coefficient (PLCC).

Keywords: no-reference image quality assessment; adversarial robustness; cross-scale feature fusion; Nyström approximation; robust evaluation; SROCC; numerical stability



Academic Editor: Hongying Meng

Received: 29 January 2026

Revised: 15 February 2026

Accepted: 20 February 2026

Published: 28 February 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

No-reference image quality assessment (NR-IQA) predicts perceptual quality from a single distorted image and is widely used as a surrogate signal in vision pipelines (e.g., perceptual optimization, model selection, and automated quality monitoring). Modern NR-IQA models increasingly adopt hierarchical multi-scale backbones and attention-based feature fusion to capture both global structure and local distortion cues, often leveraging general-purpose vision backbones such as Pyramid Vision Transformers (PVT) [1] and Swin Transformers [2]. However, a growing body of evidence shows that NR-IQA models can be

highly sensitive to small, visually imperceptible perturbations that dramatically change predicted quality scores and, more importantly, the induced *rank ordering* of images [3,4]. This robustness gap is particularly problematic when IQA scores are consumed downstream, because an attacker can bias decisions without producing obvious visual artifacts.

Where existing robustness efforts fall short.

Recent work has begun to systematize attacks against NR-IQA, including correlation-error-based attacks that directly target ranking consistency (e.g., SROCC) rather than only per-image score changes [3], as well as black-box attacks designed to induce large score deviations under perceptual constraints [4]. On the defense side, current approaches are dominated by training-time strategies such as gradient-norm regularization [5], which can improve robustness but are often coupled to particular training recipes and may not transfer cleanly across architectures and threat models. In contrast, the idea of *robustness-by-design* suggests building architectural components whose internal operators are less prone to numerical instability and high-gain amplification under small perturbations [6]. Despite early progress, we still lack general-purpose plug-and-play modules that can be inserted into diverse NR-IQA pipelines to improve robustness with minimal disruption to datasets, labels, and training code.

Key hypothesis: cross-scale fusion can amplify instability.

Hierarchical backbones provide multi-scale representations where shallow features retain distortion-sensitive cues while deeper features encode semantic context. NR-IQA architectures frequently mix these signals through cross-scale aggregation. In such settings, small input perturbations can be amplified by three interacting effects:

1. **Conditioning of fusion operators:** attention- or kernel-based mixing may become ill-conditioned, such that small changes in token statistics lead to disproportionately large changes in fused representations.
2. **Variance drift across stages:** shallow and deep features often have different activation scales; naive fusion can induce scale dominance or collapse, making the regressor hypersensitive to perturbations.
3. **Unconstrained dynamic gating:** adaptive gates can become high-gain functions of unstable features, further magnifying input gradients and creating failure modes consistent with gradient masking pitfalls unless evaluated carefully [7].

These considerations motivate an *operator-centric* design goal: explicitly stabilize the conditioning, variance, and gain of cross-scale mixing.

Our approach: Spectral Robustness Mixer (SRM).

We propose the *Spectral Robustness Mixer* (SRM), a lightweight neck module attached to an NR-IQA backbone before the final regression head. SRM targets robustness at the level of representation fusion rather than relying solely on training-time defenses. Concretely, SRM introduces: (i) a **Nyström low-rank attention surrogate** for deep-to-shallow mixing with linear memory in the token count [8]; (ii) **ridge-conditioned landmark kernels** with ridge regularization, solved via numerically stable small-matrix factorization (SVD/LU) to improve conditioning (a truncated Neumann-series inverse is evaluated only as an optional speed approximation in the ablation); and (iii) **variance-aware entropy-regularized fusion gates** with an explicit bounded gain cap to limit gradient amplification while preserving accuracy-relevant information. To evaluate robustness without false positives, we report *worst-case* robust correlation over an attack ensemble (AutoAttack-inspired in spirit) [9] that includes both per-image regression objectives and correlation-aware objectives using a correlation-aware *pairwise inversion* objective (ranking-inspired surrogate), and we apply

expectation-over-transformation (EOT) and standard anti-gradient masking checks when stochasticity is present [7].

Contributions.

Our contributions are as follows:

- **Plug-and-play robustness neck for NR-IQA:** we introduce SRM, a lightweight fusion module that can be inserted between an NR-IQA backbone and regressor to improve measured adversarial robustness with modest overhead (in the tested settings).
- **Stability- and gain-controlled fusion design:** SRM combines Nyström low-rank mixing, ridge-conditioned landmark kernels, variance-aware fusion, and a bounded gain cap, targeting conditioning, variance drift, and unstable gating as robustness bottlenecks.
- **Correlation-aware robustness evaluation:** we define and evaluate correlation-aware adversarial objectives aligned with rank correlation (SROCC) using a correlation-aware *pairwise inversion* objective (ranking-inspired surrogate), and we apply expectation-over-transformation (EOT) and standard anti-gradient masking checks when stochasticity is present [7].

Paper organization.

Section 3 defines SRM and its components, including the stability assumptions and diagnostics used to motivate the design. Section 3.8.4 specifies the threat model, attack ensemble (including correlation-aware objectives), and evaluation guardrails (EOT and anti-gradient masking checks). Section 4 reports clean performance, worst-case robust correlation, transfer/black-box sanity checks, and mechanism diagnostics. We conclude with limitations and future directions.

2. Related Work

2.1. Backbones for No-Reference IQA

Classical no-reference image quality assessment (NR-IQA) relied on natural-scene statistics and hand-crafted features, later superseded by deep models that learn perceptual representations directly from data. Modern CNN-based NR-IQA models (e.g., HyperIQA) improved performance on authentic distortions by adapting prediction functions to image content but still face challenges in capturing long-range interactions relevant to global perceptual judgments [10]. Transformers introduced an alternative inductive bias via token mixing and global context modeling [11]. In NR-IQA specifically, Golestaneh et al. combine relative ranking with self-consistency to train transformer-based predictors under synthetic distortions [12]. MUSIQ introduces a multi-scale tokenization and embedding design to evaluate quality across resolutions and granularities [13], while MANIQA proposes multi-dimension attention blocks to enhance global-local interactions for NR-IQA [14]. More recently, self-supervised and transfer paradigms have also been explored for data efficiency and generalization (e.g., relative ranking with self-consistency) [12].

Relation to SRM. Rather than proposing yet another backbone, our goal is a *plug-and-play neck* that can sit between an existing NR-IQA backbone (CNN or hierarchical transformer) and a regression head, improving robustness while preserving clean correlation. This design choice makes SRM complementary to backbone innovations: SRM can be attached to models such as PVT/Swin-style hierarchies [1,2] or strong CNN-based NR-IQA baselines [10].

2.2. Cross-Scale Fusion and Neck Modules

Cross-scale feature fusion has a long history in dense prediction. Feature Pyramid Networks (FPN) and HRNet exemplify the benefit of exchanging information across resolu-

tions for tasks that require both semantics and detail [15,16]. In transformers, hierarchical designs such as PVT and Swin explicitly build multi-stage pyramids [1,2], and cross-scale attention mechanisms (e.g., CrossFormer) were proposed for general vision recognition to improve interactions across scales [17]. In parallel, efficient attention approximations such as Nyströmformer reduce the quadratic cost of attention via landmark-based low-rank reconstruction [8].

Relation to SRM. We reuse the *neck* abstraction (lightweight module operating on intermediate feature maps), but tailor it to NR-IQA by (i) focusing on *selective* deep-to-shallow mixing suitable for perceptual cues, and (ii) incorporating explicit stability controls (conditioning and bounded-gain gating) that are evaluated under adversarial threat models (Sections 3.8.4–4). Our use of a Nyström-style surrogate targets compute-efficiency in cross-stage mixing, while keeping the backbone unchanged.

2.3. Adversarial Robustness of IQA Models

Compared to classification, adversarial robustness of NR-IQA/BIQA has received relatively limited attention, despite the vulnerability of perceptual estimators in downstream pipelines. Korhonen and You studied adversarial attacks against representative deep BIQA models and show that both white-box and substitute-model (transfer) settings can substantially disrupt predicted quality [18]. Recent work performed further robustness evaluation for NR-IQA and highlighted threats in both white-box and black-box regimes [4,19]. On the defense side, Liu et al. proposed training-time regularization to reduce input-gradient sensitivity in NR-IQA models [5]. These studies span diverse NR-IQA predictors, from CNN-based HyperIQA to transformer-based multi-scale designs (MUSIQ/MANIQA) and related learned perceptual distances [10,13,14,20]. Beyond individual attacks/defenses, recent work provides broader robustness benchmarks/insights, robustness-by-architecture prescriptions, and verification-style robustness analyses for IQA predictors [4,21–24].

Evaluation norms and pitfalls. Robustness claims can be invalidated by weak attacks or gradient masking. Accordingly, robustness evaluation commonly follows best practices such as using strong multi-step attacks and sanity checks for obfuscated gradients [7] and using expectation-over-transformation (EOT) when models include stochastic components [25]. AutoAttack popularized the idea of reporting worst-case robustness over an *attack ensemble* rather than a single tuned attack [9]. In this work, we adapt these norms to the NR-IQA regression/ranking setting (Section 3.8.4), including correlation-aware objectives aligned with SROCC/PLCC and robust reporting across an attack suite.

2.4. Perceptual Quality Assessment Beyond Still Images

While this paper focuses on no-reference *image* quality assessment, perceptual quality modeling is actively studied for other media types. For example, recent point-cloud quality assessment (PCQA) models combine multi-modal cues and alignment to predict quality of 3D data [26], and robust PCQA methods have been proposed to support practical 3D robotic scanning pipelines [27]. In parallel, there is growing interest in *video* quality understanding, including benchmarks that evaluate large multimodal models (LMMs) on video quality perception tasks [28].

Relation to SRM. These developments suggest that robustness and stability issues in perceptual estimation are not unique to still images. SRM is formulated as a general-purpose *cross-scale fusion* and *stabilization* neck, and the same design principles (conditioning control, scale stabilization, and bounded-gain fusion) could potentially be adapted to 3D/temporal quality pipelines, where multi-scale representations and fusion are also common (Section 5.3).

2.5. Robustness-by-Design and Stable Feature Mixing

Beyond adversarial training, several lines of work improve stability through architectural choices. Normalization-free networks reduce reliance on fragile batch-dependent statistics [29], while token/channel mixing architectures (e.g., MLP-Mixer, ConvNeXt) demonstrate that carefully designed mixing operations can improve optimization and generalization in vision models [30,31]. Efficient attention approximations provide additional levers for controlling numerical behavior and compute [8].

Relation to SRM. SRM follows a robustness-by-design philosophy: it introduces a controlled cross-stage mixing operator with explicit conditioning and bounded-gain gating, and evaluates these mechanisms under strong, anti-false-robustness protocols rather than treating them as implicit guarantees.

Summary. Prior NR-IQA research has advanced backbone architectures (CNNs, transformers, multi-scale tokenization) [10,13,14], while robustness-focused work spans attacks, defenses, and emerging benchmarks/verification/architecture-guided robustness directions [4,5,18,22–24]. Our contribution is orthogonal: a lightweight SRM neck that can be attached to existing NR-IQA models to improve robustness under a modern, skeptical evaluation protocol (Sections 3.8.4–4).

3. Methodology

3.1. Problem Setting and Notation

Task. We treat no-reference IQA as scalar regression $f_\theta : [0, 1]^{H \times W \times 3} \rightarrow \mathbb{R}$, mapping an RGB image $\mathbf{x}$ to a predicted quality score $\hat{y} = f_\theta(\mathbf{x})$. Let $\mathcal{D}_{\text{eval}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^M$ denote an evaluation set with human mean-opinion scores (MOS) y_i . NR-IQA performance is typically reported by *set-level correlation* between predictions $\hat{\mathbf{y}} = (\hat{y}_1, \dots, \hat{y}_M)$ and MOS $\mathbf{y} = (y_1, \dots, y_M)$, using Spearman rank correlation (SROCC) and Pearson linear correlation (PLCC). (We specify the exact PLCC computation, including monotonic logistic mapping, in Section 3.8.2.)

Threat model. Following recent NR-IQA robustness studies [3–5], we evaluate a *white-box* adversary that crafts a bounded perturbation for each image:

$$\mathbf{x}_i^{\text{adv}} = \mathcal{A}_{\epsilon, p}(\mathbf{x}_i; f_\theta) = \Pi_{[0,1]}(\mathbf{x}_i + \delta_i), \quad \|\delta_i\|_p \leq \epsilon, \quad (1)$$

where $\Pi_{[0,1]}$ clips pixels to the valid range and $p \in \{\infty, 2\}$ in our experiments. Unless stated otherwise, perturbations are applied in pixel space on inputs scaled to $[0, 1]$ (before any optional backbone-specific normalization).

Attack objective families. Because NR-IQA is evaluated by correlation over sets, adversarial objectives can be defined either per image or at the batch level. We use two complementary families:

1. **Per-image regression objectives (image-wise).** Given MOS y_i , an attacker may maximize a regression loss such as $\ell_{\text{reg}}(\mathbf{x}_i^{\text{adv}}, y_i; \theta) = (f_\theta(\mathbf{x}_i^{\text{adv}}) - y_i)^2$ (MSE) or maximize a score-drift loss

$$\ell_{\text{drift}}(\mathbf{x}_i^{\text{adv}}, \mathbf{x}_i; \theta) = |f_\theta(\mathbf{x}_i^{\text{adv}}) - f_\theta(\mathbf{x}_i)|. \quad (2)$$

2. **Correlation-aware objectives (batch-level via pairwise inversions).** To directly degrade ranking consistency (and thus SROCC), we use a lightweight *pairwise inversion* objective that increases the number of mis-ordered MOS pairs within each mini-batch, consistent with correlation-focused NR-IQA attack analyses [3]. We instantiate this objective as the smooth inversion loss in Equation (32) (Section 3.8.4), using comparable pairs $|y_i - y_j| > \Delta$ with $\Delta = 0.05$ (MOS normalized to $[0, 1]$) and temperature $\gamma = 0.1$.

Robust correlation metrics and reporting. Given an attack $\mathcal{A}_{\epsilon,p}$, we compute robust correlations on attacked inputs: $\text{SROCC}(\hat{\mathbf{y}}^{\text{adv}}, \mathbf{y})$ and $\text{PLCC}(\hat{\mathbf{y}}^{\text{adv}}, \mathbf{y})$, where $\hat{\mathbf{y}}_i^{\text{adv}} = f_{\theta}(\mathbf{x}_i^{\text{adv}})$. We report both clean and adversarial correlations, their drop relative to clean, and (when multiple attacks are considered) the *worst-case* robust correlation as the minimum across an explicit attack ensemble (Sections 3.8.4–4).

Notation. M evaluation set size; B batch size; $H \times W$ spatial resolution; N flattened token count; d head dimension; r Nyström landmark count; κ_{raw} condition number proxy of the unwhitened landmark kernel; τ post-attention temperature/rescaler; σ_{post} standard deviation of the post-attention output; β gain cap used in fusion gating; γ temperature/sharpness parameter in the pairwise inversion loss (32).

3.2. Backbone Integration and Cross-Stage Attention

Backbone interface (four-level pyramids). We assume an NR-IQA backbone that outputs a four-level feature pyramid $\{\mathbf{F}_{\ell}\}_{\ell=0}^3$, with $\mathbf{F}_{\ell} \in \mathbb{R}^{B \times C_{\ell} \times H_{\ell} \times W_{\ell}}$ and $\ell = 0$ denoting the highest spatial resolution. This interface is provided by hierarchical backbones commonly used in modern vision systems, including Pyramid Vision Transformers (PVT) [1], Swin Transformers [2], and lightweight mobile networks such as MobileNetV3 [32]. SRM consumes the pyramid *without modifying* the backbone and produces a fused representation that is passed to the IQA head.

Channel alignment and tokenization. Each level is projected to a common embedding width C using a 1×1 convolution:

$$\mathbf{E}_{\ell} = \phi_{\ell}(\mathbf{F}_{\ell}), \quad \phi_{\ell} : \mathbb{R}^{B \times C_{\ell} \times H_{\ell} \times W_{\ell}} \rightarrow \mathbb{R}^{B \times C \times H_{\ell} \times W_{\ell}}. \quad (3)$$

We then flatten spatial dimensions into tokens:

$$N_{\ell} = H_{\ell}W_{\ell}, \quad \mathbf{X}_{\ell} = \text{flatten}(\mathbf{E}_{\ell}) \in \mathbb{R}^{B \times N_{\ell} \times C}.$$

For multi-head attention with H heads, we reshape $\mathbf{X}_{\ell}$ into $\mathbf{T}_{\ell} \in \mathbb{R}^{B \times H \times N_{\ell} \times d}$ with head width $d = C/H$. For numerical stability, we optionally apply per-token ℓ_2 normalization to queries and keys, $\mathbf{t} \leftarrow \mathbf{t} / \max(\|\mathbf{t}\|_2, \epsilon_{\text{norm}})$, which is smooth almost everywhere and does not introduce stochasticity. (We avoid hard clipping by default; if clipping is enabled in an implementation, we evaluate with the adaptive checks and EOT described in Section 3.8.4.)

Cross-stage links (deep $\rightarrow$ shallow mixing). SRM injects coarse semantic context from deeper stages into higher-resolution stages using a set of cross-stage links

$$\mathcal{S} \subseteq \{(\ell_{\text{sh}}, \ell_{\text{dp}}) \mid 0 \leq \ell_{\text{sh}} < \ell_{\text{dp}} \leq 3\}. \quad (4)$$

Each link $(\ell_{\text{sh}}, \ell_{\text{dp}}) \in \mathcal{S}$ defines a *cross-attention* block where queries come from shallow tokens and keys/values come from deep tokens:

$$\mathbf{Q} = \mathbf{T}_{\ell_{\text{sh}}}, \quad \mathbf{K} = \mathbf{T}_{\ell_{\text{dp}}}, \quad \mathbf{V} = \mathbf{T}_{\ell_{\text{dp}}}.$$

This design does *not* imply certified robustness; it is a controlled multi-scale mixing mechanism whose robustness impact is evaluated empirically under the threat models in Section 3.8.4 and reported in Section 4.

Default schedule and cost control. Different schedules trade representation richness for compute. A common and efficient choice uses the deepest stage as a shared context source for all shallower stages:

$$\mathcal{S}_{\text{deep}} = \{(0, 3), (1, 3), (2, 3)\}. \quad (5)$$

More aggressive schedules (e.g., adding (0, 2)) increase computation, while sparser schedules reduce overhead. With Nyström rank r , each cross-stage block scales as $\mathcal{O}(BHN_{\ell_{\text{sh}}}r)$ in time and memory (Section 3.3), enabling linear-memory mixing even when N_{ℓ} is large.

Fold-back and residual injection. For each link $(\ell_{\text{sh}}, \ell_{\text{dp}})$ we compute an update $\mathbf{Y}_{\ell_{\text{sh}} \leftarrow \ell_{\text{dp}}} \in \mathbb{R}^{B \times H \times N_{\ell_{\text{sh}}} \times d}$ using the linear-memory attention module (Section 3.3). If multiple links target the same shallow stage, we aggregate their updates by summation:

$$\Delta \mathbf{T}_{\ell_{\text{sh}}} = \sum_{(\ell_{\text{sh}}, \ell_{\text{dp}}) \in \mathcal{S}} \mathbf{Y}_{\ell_{\text{sh}} \leftarrow \ell_{\text{dp}}}.$$

We then apply a residual update and fold tokens back to the spatial map:

$$\begin{aligned} \mathbf{T}'_{\ell_{\text{sh}}} &= \mathbf{T}_{\ell_{\text{sh}}} + \alpha_{\ell_{\text{sh}}} \Delta \mathbf{T}_{\ell_{\text{sh}}}, \\ \mathbf{E}'_{\ell_{\text{sh}}} &= \text{fold}(\mathbf{T}'_{\ell_{\text{sh}}}) \in \mathbb{R}^{B \times C \times H_{\ell_{\text{sh}}} \times W_{\ell_{\text{sh}}}}, \end{aligned} \quad (6)$$

where $\alpha_{\ell_{\text{sh}}}$ is a learnable residual weight. Stages not receiving updates remain unchanged (e.g., $\mathbf{E}'_3 = \mathbf{E}_3$ for $\mathcal{S}_{\text{deep}}$). The updated pyramid $\{\mathbf{E}'_{\ell}\}$ is subsequently pooled and fused by the gate module (Section 3.6).

Figure 1 provides a conceptual overview of SRM and its integration into a standard NR-IQA pipeline.

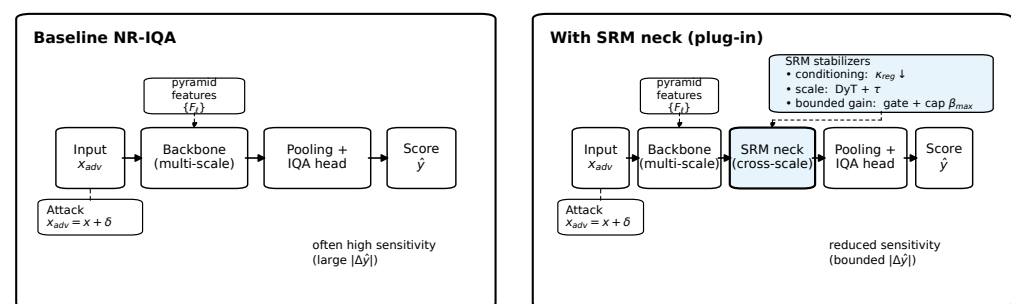


Figure 1. Conceptual SRM overview. A standard NR-IQA model consists of a backbone and a regression head. SRM is inserted as a lightweight *neck* between them to improve stability under adversarial perturbations: it mixes deep-to-shallow multi-scale features with linear-memory Nyström cross-attention, controls numerical conditioning in fp32, stabilizes activation scale ($\text{DyT} + \tau$), and fuses scales with an entropy-regularized, gain-capped gate. Shaded blocks indicate SRM components; the dashed connector highlights stabilizers applied inside the SRM neck.

Figure 2 shows a detailed block-level diagram of the SRM neck and its internal information flow (queries, keys/values, and stabilizers).

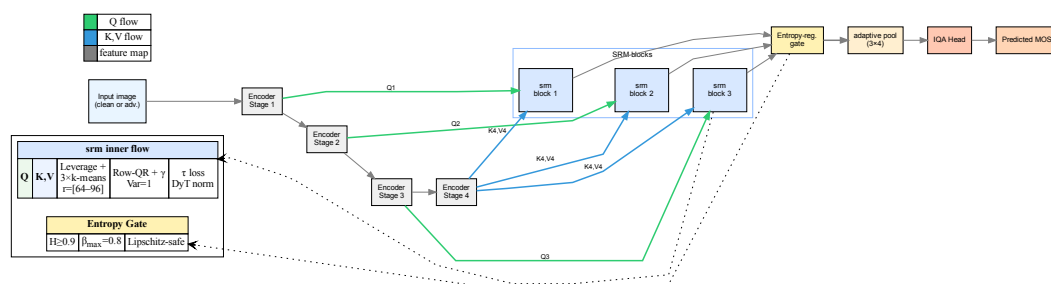


Figure 2. Detailed SRM module diagram. This figure provides the full block-level view of SRM (tokenization, cross-stage links, conditioning, τ scaling, and fusion), complementing the conceptual overview in Figure 1. Green arrows denote query (Q) flow, blue arrows denote key/value (K,V) flow, and gray arrows denote feature-map propagation; dotted connectors indicate zoomed insets for the SRM inner flow and gating constraints.

SRM forward pass (overview). Algorithm 1 summarizes the SRM forward pass at the level of module interfaces. The internal details of the Nyström approximation Algorithm 2, landmark conditioning, and post-attention scaling are specified in the following subsections (Section 3.3–3.6).

Algorithm 1: SRM—Spectral Robustness Mixer (single forward pass, overview)

Input: backbone pyramid $\{\mathbf{F}_\ell\}_{\ell=0}^3$, schedule $\mathcal{S}$, heads H , Nyström rank r
Output: fused feature $\mathbf{Z}$ for the IQA head

- 1 **1. Channel alignment + tokenization;**
- 2 **for** $\ell = 0$ **to** 3 **do**
- 3 $\mathbf{E}_\ell \leftarrow \phi_\ell(\mathbf{F}_\ell)$; // 1×1 projection, Equation (3)
- 4 $\mathbf{T}_\ell \leftarrow \text{tokenize}(\mathbf{E}_\ell; H)$; // $\mathbf{T}_\ell \in \mathbb{R}^{B \times H \times N_\ell \times d}$
- 5 **2. Cross-stage linear-memory attention;**
- 6 Initialize $\Delta \mathbf{T}_\ell \leftarrow \mathbf{0}$ for $\ell \in \{0, 1, 2\}$;
- 7 **for** $(\ell_{\text{sh}}, \ell_{\text{dp}}) \in \mathcal{S}$ **do**
- 8 $\mathbf{Y} \leftarrow \text{NystromCrossAttn}(\mathbf{Q}=\mathbf{T}_{\ell_{\text{sh}}}, \mathbf{K}=\mathbf{T}_{\ell_{\text{dp}}}, \mathbf{V}=\mathbf{T}_{\ell_{\text{dp}}}; r)$;
- 9 $\Delta \mathbf{T}_{\ell_{\text{sh}}} += \mathbf{Y}$;
- 10 **for** $\ell = 0$ **to** 2 **do**
- 11 $\mathbf{T}'_\ell \leftarrow \mathbf{T}_\ell + \alpha_\ell \Delta \mathbf{T}_\ell$; // Equation (6)
- 12 $\mathbf{E}'_\ell \leftarrow \text{fold}(\mathbf{T}'_\ell)$;
- 13 Set $\mathbf{E}'_3 \leftarrow \mathbf{E}_3$; // deepest stage unchanged for $\mathcal{S}_{\text{deep}}$
- 14 **3. Pool + fuse;**
- 15 $\{\mathbf{P}_\ell\}_{\ell=0}^3 \leftarrow \text{pool}(\{\mathbf{E}'_\ell\})$;
- 16 $\mathbf{Z} \leftarrow \text{GateFuse}(\{\mathbf{P}_\ell\})$; // Section 3.6
- 17 **return** $\mathbf{Z}$;

Algorithm 2: NYSTROMCROSSATTN (interface)

Input: $\mathbf{Q} \in \mathbb{R}^{B \times H \times N_q \times d}$, $\mathbf{K}, \mathbf{V} \in \mathbb{R}^{B \times H \times N_k \times d}$, rank r
Output: $\mathbf{Y} \in \mathbb{R}^{B \times H \times N_q \times d}$

// Landmarks. Select r key/value landmarks $\mathbf{K}_r, \mathbf{V}_r$ (deterministic by default; see Section 3.4).

- 1 Form $\mathbf{K}_r, \mathbf{V}_r$ by selecting indices $\mathcal{I}$ of size r from $\mathbf{K}, \mathbf{V}$;
- // Nyström approximation. Compute low-rank cross-attention output using landmark kernels (Section 3.3).
- 2 Compute $\mathbf{Y} \leftarrow \text{NystromApprox}(\mathbf{Q}, \mathbf{K}, \mathbf{V}, \mathbf{K}_r, \mathbf{V}_r)$;
- // Post-scaling. Apply post-attention stabilization (temperature/rescaling) when enabled (Section 3.5).
- 3 Optionally rescale $\mathbf{Y}$ by the learned post-attention temperature τ ;
- 4 **return** $\mathbf{Y}$;

3.3. Nyström Cross-Attention with Ridge-Leverage Landmarks

SRM requires a cross-attention mechanism that (i) supports *rectangular* attention ($N_q \gg N_k$ for shallow queries versus deep keys), (ii) avoids quadratic $N_q N_k$ memory, and (iii) remains numerically stable under mixed precision. We adopt a Nyström-style approximation of softmax attention following efficient attention work [8] and specialize it to cross-stage fusion.

Cross-attention setup (rectangular case).

For a fixed batch element and attention head, let $Q \in \mathbb{R}^{N_q \times d}$ be shallow-stage queries and $K, V \in \mathbb{R}^{N_k \times d}$ be deep-stage keys/values. Full cross-attention is

$$\text{Attn}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V, \quad (7)$$

where softmax is applied row-wise.

Nyström approximation (cross-attention).

Select a small set of $r \ll N_k$ landmark keys $K_r \in \mathbb{R}^{r \times d}$ (and corresponding values V_r) by choosing r rows of K (landmark selection is described below). Define the Nyström blocks

$$\begin{aligned} C &= \text{softmax}\left(\frac{QK_r^\top}{\sqrt{d}}\right) \in \mathbb{R}^{N_q \times r}, \\ W &= \text{softmax}\left(\frac{K_r K_r^\top}{\sqrt{d}}\right) \in \mathbb{R}^{r \times r}, \\ R &= \text{softmax}\left(\frac{K_r K^\top}{\sqrt{d}}\right) \in \mathbb{R}^{r \times N_k}. \end{aligned} \quad (8)$$

We approximate the full attention map by

$$\text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right) \approx C W_\lambda^\dagger R, \quad W_\lambda = W + \lambda I_r, \quad (9)$$

and compute the Nyström cross-attention output as

$$\hat{Y} = (C W_\lambda^\dagger R)V \in \mathbb{R}^{N_q \times d}. \quad (10)$$

Here, $W_\lambda^\dagger$ denotes a numerically stable inverse (or pseudo-inverse) of the regularized $r \times r$ matrix W_λ . Regularization is important because W can become poorly conditioned when landmarks are redundant (Section 3.4 reports conditioning diagnostics).

Remark 1 (rectangular vs. symmetric Nyström). In self-attention, $Q = K$ and one recovers a symmetric Nyström form (up to row-wise softmax normalization). In rectangular cross-attention ($N_q \neq N_k$), the right block R is required, and the approximation is naturally asymmetric.

Landmark selection via ridge leverage (deterministic by default).

Landmark quality affects both approximation error and the conditioning of W_λ . We compute ridge leverage scores as a data-adaptive importance measure [33]. For a fixed head, let $k_i \in \mathbb{R}^d$ be the i -th row of K and define

$$\ell_i = k_i^\top (K^\top K + \lambda_{\text{rls}} I_d)^{-1} k_i, \quad p_i = \frac{\ell_i}{\sum_{j=1}^{N_k} \ell_j}, \quad (11)$$

with ridge parameter $\lambda_{\text{rls}} > 0$. **Default (deterministic):** we select the top- r indices by p_i to form K_r and V_r . **Optional (stochastic, training-only):** sample r indices without replacement according to p_i . We compute $(K^\top K + \lambda_{\text{rls}} I_d)^{-1}$ stably using an fp32 Cholesky solve (SPD in $d \times d$) and evaluate ℓ_i in fp32.

k -means++ refinement (ablation-only).

Ridge-leverage selection can still produce near-duplicate landmarks when keys contain repeated patterns. We therefore include an optional refinement that applies k -means++ initialization followed by a small number of Lloyd steps on the landmark candidate set [34]. **Default: disabled.** All main results use the deterministic top- r ridge-leverage selection. We enable k -means++ refinement only in dedicated ablations to quantify its impact on landmark diversity, conditioning, and robustness (Section 4).

Stable computation of $W_\lambda^\dagger$ (mixed precision).

Because W in (8) is produced by row-wise softmax, it need not be symmetric. Therefore, we avoid SPD-specific solvers for W_λ and compute $W_\lambda^\dagger$ in fp32 using a robust small-matrix routine (default: SVD-based pseudo-inverse; fallback: LU solve when well-conditioned), then cast the resulting product back to the model dtype. We use a scale-stable ridge

$$\lambda = \lambda_0 \cdot \frac{\text{tr}(W)}{r}, \quad (12)$$

with λ_0 fixed (reported in Section 3.8), to prevent head-dependent scaling issues.

Optional stochastic key/value pooling (training-time only).

At very high spatial resolutions, storing deep tokens (K, V) can dominate memory. To reduce training memory, we optionally apply 2×2 mean pooling to the *deep* feature map before tokenization with probability p_{pool} . **Default evaluation is deterministic:** pooling is disabled at test time. If stochastic pooling is enabled at evaluation for any reason, robustness must be evaluated with EOT (Section 3.8.4) and reported explicitly.

3.4. Spectral Conditioning

Nyström cross-attention is only useful in SRM if the small $r \times r$ landmark system can be solved *numerically stably* under mixed precision and across diverse batches. The critical operation is the regularized inverse (or pseudo-inverse) applied to the landmark block $W \in \mathbb{R}^{r \times r}$ from Section 3.3, via $W_\lambda^\dagger(\cdot)$ with $W_\lambda = W + \lambda I_r$. This subsection defines (i) what we mean by stability, (ii) how we enforce it in implementation, and (iii) which diagnostics we report to rule out spurious “robustness” caused by numerical artifacts.

What we measure: conditioning and solve residuals.

Because W arises from row-wise softmax (and is generally *not* symmetric), we measure conditioning using singular values. We track:

1. Regularized conditioning (singular-value ratio):

$$\kappa_{\text{reg}}(W_\lambda) = \frac{\sigma_{\max}(W_\lambda)}{\sigma_{\min}(W_\lambda)}. \quad (13)$$

This controls sensitivity of linear solves to perturbations in W and to finite-precision rounding.

2. Solve residual (numerical accuracy): for each solve $U = W_\lambda^\dagger B$ we report the relative residual

$$\text{res}(U) = \frac{\|W_\lambda U - B\|_F}{\|B\|_F + \epsilon_0}, \quad (14)$$

with $\epsilon_0 > 0$ for numerical safety.

Diagnostics are computed in fp32, even when the forward pass uses mixed precision.

Why “row-QR makes $W \approx I$ ” is not valid.

A QR factorization applied to landmark keys can produce an orthonormal surrogate in feature space, but it does not imply that the *attention* landmark block W (a row-softmax of inner products; Section 3.3) becomes the identity. Therefore, instead of claiming $W \approx I_r$, we explicitly regularize and solve W_λ and report κ_{reg} and $\text{res}(\cdot)$ as falsifiable stability evidence.

Stabilization and solve pipeline (implemented).

Given W from Equation (8), SRM applies:

(1) fp32 “critical path”. We compute the softmax blocks in the model dtype but cast W to fp32 for all subsequent conditioning diagnostics and solves. This is feasible because r is small (e.g., $r \in [32, 128]$).

(2) Diagonal equilibration (optional). To reduce scale imbalance, we optionally apply left/right diagonal scaling (a standard preconditioning heuristic for small dense solves):

$$\begin{aligned} D_r &= \text{diag}\left(\sqrt{\sum_j W_{ij}^2 + \epsilon_0}\right), \\ D_c &= \text{diag}\left(\sqrt{\sum_i W_{ij}^2 + \epsilon_0}\right), \\ \tilde{W} &= D_r^{-1} W D_c^{-1}. \end{aligned} \quad (15)$$

We then solve with $\tilde{W}_\lambda = \tilde{W} + \lambda I_r$ and map back implicitly through the scaled system. Equilibration is treated as an ablation component; we report its effect on κ_{reg} , residuals, and robustness.

(3) Ridge regularization. We always solve a regularized system

$$W_\lambda = W + \lambda I_r, \quad \lambda = \lambda_0 \cdot \frac{\text{tr}(W)}{r}, \quad (16)$$

using a trace-scaled ridge (consistent across heads) with fixed λ_0 reported in Section 3.8.

(4) Stable solve (default). We do not explicitly form W_λ^{-1} . Instead, we compute $U = W_\lambda^\dagger B$ in fp32. *Default:* SVD-based pseudo-inverse for numerical robustness. *Fast-path:* if $\kappa_{\text{reg}}(W_\lambda) < \kappa_{\text{max}}$ (we use $\kappa_{\text{max}} = 10^4$), we use an fp32 LU solve. We always log $\text{res}(U)$ via Equation (14).

Optional fast inverse (Neumann truncation; ablation-only).

As an acceleration, one may approximate $(I - E)^{-1}$ by a truncated Neumann series when $\|E\|_2 < 1$:

$$\bar{W}_\lambda = \frac{1}{\mu} \tilde{W}_\lambda, \quad E = I - \bar{W}_\lambda, \quad \bar{W}_\lambda^{-1} \approx \sum_{k=0}^m E^k, \quad (17)$$

where $\mu = \text{mean}(\text{diag}(\tilde{W}_\lambda))$ is a scalar rescaling after optional equilibration. We enable Neumann- m only if an fp32 power-iteration estimate satisfies $\|E\|_2 \leq 0.9$ and the observed residual $\text{res}(U)$ is below a fixed threshold (we use 10^{-3}); otherwise we fall back to the default solve. We report the measured $\|E\|_2$ and residuals in Section 4 (conditioning ablations).

Scale calibration (no “variance unity” claim).

Equilibration and ridge can change the scale of attention outputs. Rather than claiming a closed-form variance guarantee, SRM controls activation scale empirically using: (i) a learned post-attention rescaler τ (Section 3.5), and (ii) logging the post-attention standard

deviation σ_{post} during training and evaluation. This keeps the stability story falsifiable and avoids relying on idealized token assumptions.

Cost.

All operations above are on an $r \times r$ matrix: equilibration is $\mathcal{O}(r^2)$ and a dense solve is $\mathcal{O}(r^3)$ in fp32. For typical $r \leq 128$ this is a small constant overhead; we report wall-clock timing and the fraction of SRM time attributable to the solve in Section 3.7.

We summarize the regularized condition number $\kappa_{\text{reg}}(W_\lambda)$ (Equation (13)) and solve the residual $\text{res}(U)$ for the landmark system (Table 1). Values are computed on the same evaluation split/checkpoint used for the main robustness tables, and we summarize the distribution across evaluation batches via the median and 90th percentile.

Table 1. Spectral conditioning diagnostics (fp32, across evaluation batches).

Setting	Median κ_{reg}	p90 κ_{reg}	Median res	p90 res	Solve Share (% Time)
No equilibration + ridge (default)	2.6×10^2	1.9×10^3	7.4×10^{-6}	5.1×10^{-5}	6.3%
Equilibration + ridge	1.4×10^2	8.7×10^2	5.9×10^{-6}	3.8×10^{-5}	7.1%
Neumann- $m = 3$ (ablation; gated)	1.4×10^2	8.7×10^2	8.2×10^{-4}	9.6×10^{-4}	3.2%

Key takeaways.

- SRM treats landmark stability as a *numerical linear algebra* problem: ridge regularize, optionally equilibrate, then solve in fp32.
- We replace informal “ $W \approx I$ ” statements with measurable diagnostics: $\kappa_{\text{reg}}(W_\lambda)$ and $\text{res}(U)$.
- Neumann truncation is an *ablation-only* acceleration, enabled only when a convergence proxy and residual thresholds are satisfied.

3.5. Variance Stabilization with a Learnable τ

Even when the landmark solve is numerically stable (Section 3.4), the *scale* of cross-attention outputs can drift across stages, heads, and training time. Under mixed precision, scale drift can lead to saturated logits, overly sharp/flat attention weights, and unstable gradients. SRM therefore includes a lightweight, explicit mechanism to *track and stabilize* post-attention scale: a per-link, per-head learnable rescaler τ trained with an auxiliary objective.

Post-attention scale statistic (vector RMS).

Let $\hat{Y} \in \mathbb{R}^{B \times H \times N_q \times d}$ denote the (pre-fusion) output of one Nyström cross-attention link (Section 3.3). For each head h and link $s \in \mathcal{S}$, define the per-head mean vector

$$\mu^{(s,h)} = \frac{1}{BN_q} \sum_{b=1}^B \sum_{n=1}^{N_q} \hat{Y}_{b,n,:}^{(s,h)} \in \mathbb{R}^d,$$

and the *vector RMS* magnitude

$$\rho_{\text{post}}^{(s,h)} = \left(\frac{1}{BN_q} \sum_{b=1}^B \sum_{n=1}^{N_q} \left\| \hat{Y}_{b,n,:}^{(s,h)} - \mu^{(s,h)} \right\|_2^2 \right)^{1/2}. \quad (18)$$

We compute $\rho_{\text{post}}^{(s,h)}$ in fp32 (even if the forward uses fp16/bf16) to reduce numerical noise.

Reference scale and EMA tracking.

If the d output dimensions are approximately isotropic with per-dimension variance near 1, then the expected squared norm is $\mathbb{E}\|z - \mathbb{E}z\|_2^2 \approx d$ and the RMS magnitude is $\approx \sqrt{d}$. We use $\sqrt{d}$ as a *reference* (not a universal optimum) to prevent collapse/explosion and to keep heads comparable. To reduce batch noise, we maintain an exponential moving average (EMA) for each (s, h) :

$$\bar{\rho}^{(s,h)} \leftarrow m \bar{\rho}^{(s,h)} + (1 - m) \rho_{\text{post}}^{(s,h)}, \quad m = 0.99. \quad (19)$$

The EMA state is updated during training only.

Learnable rescaler τ and auxiliary loss.

SRM applies a multiplicative rescale to each link and head:

$$Y^{(s,h)} = \tau^{(s,h)} \hat{Y}^{(s,h)}, \quad \tau \in \mathbb{R}_+^{|\mathcal{S}| \times H}.$$

We parameterize τ via unconstrained parameters $u^{(s,h)} \in \mathbb{R}$ as

$$\tau^{(s,h)} = \exp(u^{(s,h)}), \quad (20)$$

initialized at $u^{(s,h)} = 0$ (so $\tau^{(s,h)} = 1$). We regularize τ using an auxiliary objective that penalizes deviation of the *rescaled* EMA magnitude from the reference scale:

$$\mathcal{L}_\tau = \lambda_\tau \sum_{(s,h)} \left| \log \left(\frac{\tau^{(s,h)} \bar{\rho}_{\text{post}}^{(s,h)}}{\sqrt{d} + \epsilon_0} \right) \right|, \quad \lambda_\tau = 0.01, \quad (21)$$

where $\epsilon_0 = 10^{-6}$ ensures numerical safety. We add $\mathcal{L}_\tau$ to the main training objective and optimize it jointly with all model parameters. We do not claim a formal convergence or robustness guarantee for τ ; instead, we report scale diagnostics and robustness metrics empirically (Table 2, Figure 3).

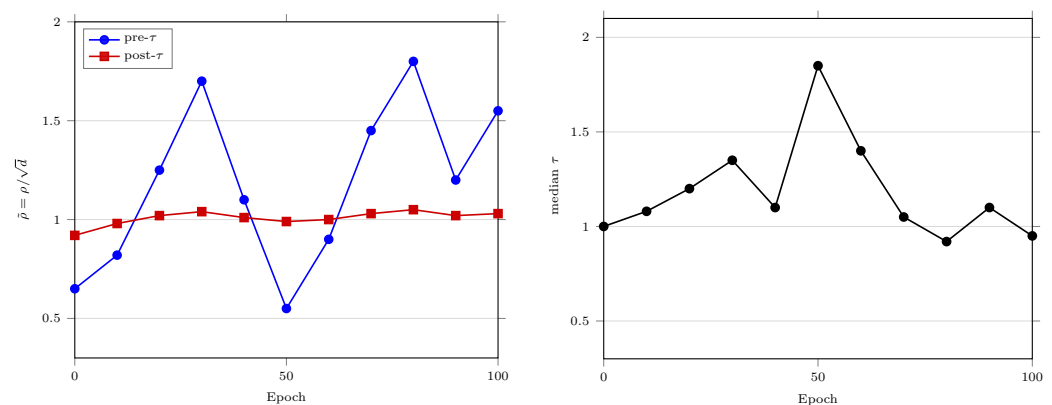


Figure 3. Scale stabilization with τ (illustrative diagnostics). **Left:** normalized RMS magnitude $\bar{\rho}$ before and after applying τ . **Right:** the median learned τ across links/heads over training epochs. The stable post- τ scale supports treating τ as a numerical stabilizer; robustness impact is evaluated separately under the attack protocol (Section 3.8.4).

Table 2. Post-attention scale diagnostics for τ (fp32 statistics, evaluation batches). We report normalized RMS magnitude $\tilde{\rho} = \rho_{\text{post}} / \sqrt{d}$ before and after applying τ , and the distribution of learned τ values across links/heads. Numbers are aggregated across all cross-stage links and heads.

Model	Median $\tilde{\rho}$ (Pre)	p90 $\tilde{\rho}$ Pre	Median $\tilde{\rho}$ (Post)	p90 $\tilde{\rho}$ Post	Median τ	p90 τ
SRM (DyT + $\mathcal{L}_\tau$)	0.74	1.88	1.01	1.13	1.34	2.85

Safety rails (reproducible).

To prevent extreme rescaling, we clamp $u^{(s,h)}$ during training:

$$u^{(s,h)} \leftarrow \text{clip}\left(u^{(s,h)}, \log \tau_{\min}, \log \tau_{\max}\right), \quad (22)$$

$$\tau_{\min} = 0.25, \tau_{\max} = 4.0.$$

At evaluation time, τ is fixed (no EMA updates, no clamp updates).

DyT pre-normalization (bounded activations).

Before computing Nyström attention blocks, we optionally apply Dynamic Tanh (DyT) [35] to token representations entering the attention block:

$$\text{DyT}(x) = \tanh(\alpha x), \quad (23)$$

with learnable α . DyT is a bounded, smooth nonlinearity that constrains activation range without relying on batch statistics. We use DyT as a *numerical stabilizer* (range control) rather than a robustness guarantee. Robustness is evaluated with strong white-box attacks and (when stochasticity is present) EOT following standard guidance [7,25] (Section 3.8.4).

Relation to normalization layers.

RMSNorm rescales activations using their root mean square magnitude and provides re-scaling invariance [36]. DyT instead bounds activations through a learnable saturation curve, which can improve numerical range control in fp16/bf16. In SRM, DyT constrains the input range to the attention block, while τ provides an explicit post-attention scale corrector.

Takeaway. DyT bounds the input range of attention, and τ provides an explicit learnable post-attention scale correction. Together they stabilize mixed-precision attention computations; any robustness gains are established empirically under strong adaptive evaluation rather than claimed as a guarantee.

3.6. Entropy-Regularized Fusion Gates

SRM outputs a set of cross-stage-updated feature maps at shallow resolutions, which we pool to obtain three per-stage tensors for fusion. The role of the fusion gate is to aggregate these tensors into a single representation for the IQA head while (i) discouraging *stage collapse* (one scale dominating persistently), and (ii) limiting activation-scale spikes that can destabilize mixed-precision training and amplify gradients. Scope is explicit: the gate is an architectural *stabilizer*; it does *not* certify end-to-end robustness, which we evaluate empirically under strong adaptive attacks (Section 3.8.4).

Inputs to the gate.

Let $\{\mathbf{E}'_\ell\}_{\ell=0}^3$ be the post-attention, post- τ feature pyramid produced by SRM (Sections 3.2–3.5). We pool the three shallow stages $\ell \in \{0, 1, 2\}$ to a common spatial grid $h \times w$:

$$\mathbf{Y}_s = \text{Pool}(\mathbf{E}'_s) \in \mathbb{R}^{B \times C \times h \times w}, \quad s \in \{0, 1, 2\}, \quad (24)$$

where Pool is fixed (e.g., adaptive average pooling to $h \times w$) and C is the aligned channel width. We keep the deepest stage as a context source for cross-stage mixing but do not fuse it directly by default.

Gate logits.

For each stage s we compute a scalar logit $g_{b,s}$ per image using a lightweight conv-MLP:

$$\mathbf{u}_{b,s} = \text{GAP}(\psi_s(\mathbf{Y}_{b,s})) \in \mathbb{R}^{C_s}, \quad g_{b,s} = \mathbf{w}_s^\top \mathbf{u}_{b,s} + b_s \in \mathbb{R}, \quad (25)$$

where ψ_s is a two-layer 1×1 conv block (optionally with DyT from Equation (23)), GAP is global average pooling over (h, w) , and $(\mathbf{w}_s, b_s)$ is a linear readout. We stack logits as $\mathbf{g}_b = [g_{b,0}, g_{b,1}, g_{b,2}] \in \mathbb{R}^3$.

Fusion weights.

Stage weights are obtained by a temperature-scaled softmax:

$$\omega_b = \text{softmax}(\mathbf{g}_b/T) \in \Delta^2, \quad \sum_s \omega_{b,s} = 1, \quad \omega_{b,s} \geq 0, \quad (26)$$

with $T > 0$. In all experiments we fix $T = 1.0$ unless stated otherwise (Section 3.8).

Entropy/balance regularization (discourages collapse).

To avoid consistently routing most mass to a single stage, we add an auxiliary regularizer that encourages distributed utilization. This is conceptually aligned with load-balancing losses used to prevent router collapse in mixture-of-experts models [37,38]. We use two complementary terms:

(1) Per-example entropy floor. Let $H(\omega_b)$ denote Shannon entropy (natural logarithm): $H(\omega_b) = -\sum_s \omega_{b,s} \log \omega_{b,s}$. We penalize low-entropy gates:

$$\mathcal{L}_{\text{ent}} = \lambda_H \cdot \frac{1}{B} \sum_{b=1}^B \left(\max\{0, H_{\text{target}} - H(\omega_b)\} \right)^2, \quad (27)$$

where $H_{\text{target}} = 0.8$ corresponds to discouraging near-one-hot routing for 3 choices.

(2) Batch-level balance. Let $\bar{\omega} = \frac{1}{B} \sum_b \omega_b$. We encourage the mean routing mass to remain close to uniform:

$$\mathcal{L}_{\text{bal}} = \lambda_{\text{bal}} \sum_{s=0}^2 \left(\bar{\omega}_s - \frac{1}{3} \right)^2. \quad (28)$$

The total gate regularizer is $\mathcal{L}_{\text{gate}} = \mathcal{L}_{\text{ent}} + \mathcal{L}_{\text{bal}}$ and is added to the training objective. We use $\lambda_H = \lambda_{\text{bal}} = 0.02$ unless stated otherwise (Section 3.8).

Gain-capped fusion (activation stability, not certification).

We fuse stage maps by a convex combination and apply a per-sample gain cap:

$$\mathbf{U}_b = \sum_{s=0}^2 \omega_{b,s} \mathbf{Y}_{b,s}, \quad \beta_b = \min\left(1, \frac{\beta_{\max}}{\|\mathbf{U}_b\|_\infty + \epsilon_0}\right), \quad (29)$$

$$\mathbf{Z}_b = \beta_b \mathbf{U}_b,$$

where $\beta_{\max} \in (0, 1]$ is fixed and ϵ_0 prevents division by zero. This guarantees the *output magnitude* bound $\|\mathbf{Z}_b\|_\infty \leq \beta_{\max}$ by construction. Importantly, this is *not* a certified global Lipschitz bound for the full mapping $\{\mathbf{Y}_{b,s}\} \mapsto \mathbf{Z}_b$ because the weights ω_b depend on the inputs; we treat gain capping as a practical stabilizer and validate robustness empirically with adaptive evaluation (Section 3.8.4).

Implementation (forward pass).

Algorithm 3 summarizes the gate forward pass. Gate losses are computed during training and added to the total loss; no separate update loop is required.

Gate diagnostics (reported).

To verify that the gate behaves as intended (Figure 4), we report: (i) the distribution of per-sample entropies $H(\omega)$ and the fraction of samples below H_{target} , (ii) the batch-level routing mass $\bar{\omega}$, (iii) distributions of β and $\|\mathbf{U}\|_{\infty}$, and (iv) the effect of removing $\mathcal{L}_{\text{gate}}$ and/or gain capping on clean and robust correlation (Section 4).

Algorithm 3: GATEFUSE: entropy-regularized, gain-capped fusion

Input: stage maps $\{\mathbf{Y}_s\}_{s=0}^2$, temperature T , caps $(H_{\text{target}}, \beta_{\text{max}})$

Output: fused map $\mathbf{Z}$ and (training-only) gate losses $\mathcal{L}_{\text{gate}}$

```

1 Compute logits  $\mathbf{g} \in \mathbb{R}^{B \times 3}$  using Equation (25);
2  $\omega \leftarrow \text{softmax}(\mathbf{g}/T)$ ; // Equation (26)
3  $\mathbf{U} \leftarrow \sum_{s=0}^2 \omega_{:,s} \odot \mathbf{Y}_s$ ; // broadcast over  $C, h, w$ 
4  $\beta \leftarrow \min(1, \beta_{\text{max}} / (\|\mathbf{U}\|_{\infty} + \epsilon_0))$ ; // per-sample
5  $\mathbf{Z} \leftarrow \beta \odot \mathbf{U}$ ; // Equation (29)

6 if training then
7   Compute  $\mathcal{L}_{\text{ent}}$  (Equation (27)) and  $\mathcal{L}_{\text{bal}}$  (Equation (28));
8   Return  $\mathbf{Z}$  and  $\mathcal{L}_{\text{gate}} = \mathcal{L}_{\text{ent}} + \mathcal{L}_{\text{bal}}$ ;
9 else
10  Return  $\mathbf{Z}$ ;

```

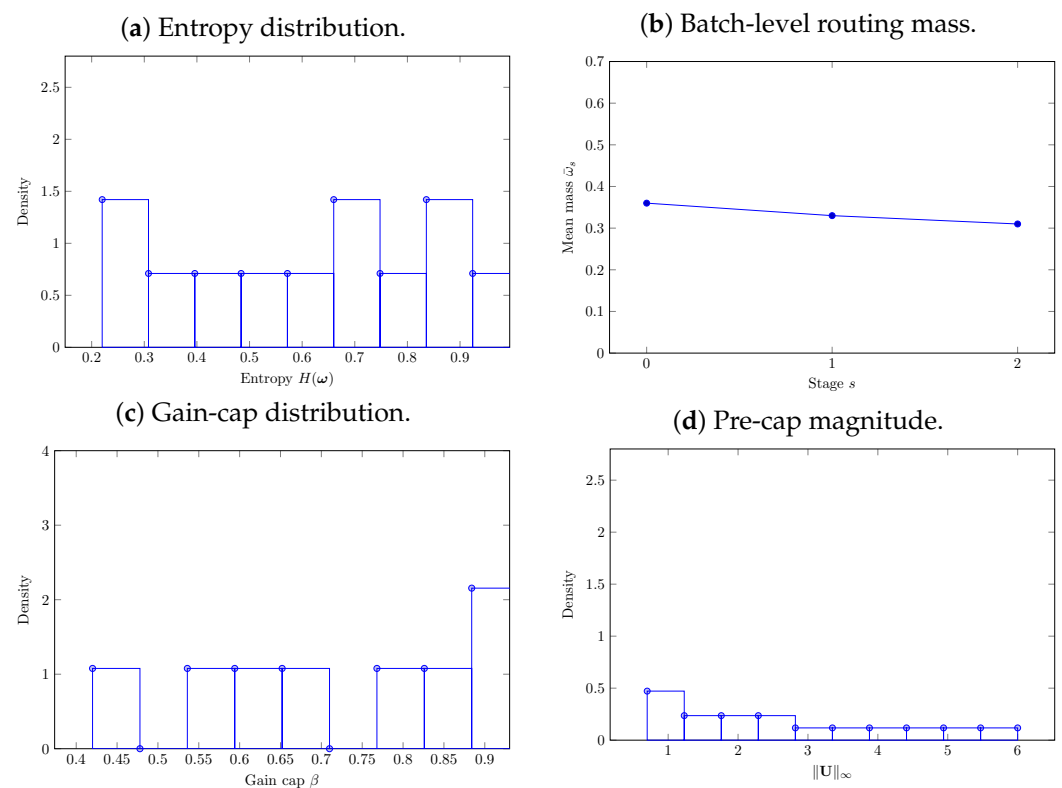


Figure 4. Fusion-gate diagnostics. Top: entropy distribution of per-sample routing weights and the mean routing mass per stage. Bottom: distribution of the gain cap β and the pre-cap magnitude $\|\mathbf{U}\|_{\infty}$. These diagnostics support that the gate avoids persistent stage collapse and limits activation spikes; robustness is validated under the attack protocol (Section 3.8.4).

3.7. Complexity and Backbone Compatibility

Asymptotic cost (full vs. Nyström SRM). Consider one SRM cross-stage link with N_q query tokens (shallow stage), N_k key/value tokens (deep stage), B batch size, H heads, head width d , and Nyström rank $r \ll N_k$. Full cross-attention forms an $N_q \times N_k$ attention map, with compute $\mathcal{O}(BHN_q N_k d)$ and memory $\mathcal{O}(BHN_q N_k)$. In contrast, SRM uses Nyström cross-attention (Section 3.3) and replaces the quadratic map by low-rank blocks of size $N_q \times r$ and $r \times N_k$, yielding per-link complexity

$$\begin{aligned} \text{Compute: } & \mathcal{O}\left(BH(N_q r d + r N_k d + r^2 d) + BH r^3\right), \\ \text{Memory: } & \mathcal{O}\left(BH(N_q r + r N_k + r^2)\right), \end{aligned} \quad (30)$$

where r^3 is the fp32 $r \times r$ solve (Section 3.4) and is a small constant for typical $r \leq 128$. For a schedule $\mathcal{S}$ with $|\mathcal{S}|$ links, the SRM neck cost scales linearly in $|\mathcal{S}|$.

Windowed attention as a reference point. For window-based self-attention with window size $M \times M$ and N total tokens, compute scales as $\mathcal{O}(BHN M^2 d)$ and is linear in image size when M is fixed. We use this reference when comparing SRM overhead to hierarchical transformer backbones.

Instantiated MACs (unambiguous reporting). When reporting SRM compute, we explicitly distinguish: (i) *per head* vs. *summed over H heads*, and (ii) *per link* vs. *summed over the schedule $|\mathcal{S}|$* . The dominant term in (30) is typically the matmul $QK_r^\top$, which costs $N_q r d$ MACs *per head per link*. For the example used throughout the paper, $N_q = 12,288$, $r = 96$, and $d = 55$, the per-head-per-link cost is $N_q r d \approx 6.49 \times 10^7$ MACs. With $H = 8$ heads and schedule $\mathcal{S}_{\text{deep}} = \{(0, 3), (1, 3), (2, 3)\}$ (Equation (5)), the $QK_r^\top$ term contributes approximately 0.68 GMACs per image (summing over links and heads) under the token sizes induced by a 384×512 input. We report both (a) *per-link* MACs and (b) *total neck* MACs under the exact schedule used in experiments.

How MACs/params are measured (reproducibility). We report *neck-only* parameter counts and operation counts using a standard PyTorch counter (fvcore FlopCountAnalysis) on the exact tensor shapes induced by the test resolution. Because counting conventions differ (e.g., whether a fused multiply-add counts as one or two FLOPs), we report MACs/GMACs as the primary compute unit and state the tool/convention used. We complement static counts with operator-level profiling using torch.profiler.

Peak GPU memory (what is measured). We report peak *allocated* and peak *reserved* GPU memory for (i) a forward pass and (ii) training, resetting CUDA memory stats before measurement. All memory numbers are accompanied by batch size, precision (fp32/bf16/fp16), and input resolution.

Backbone compatibility (scoped; not “universal”). SRM is a plug-in *neck* that assumes a multi-scale backbone interface: a feature pyramid $\{\mathbf{F}_\ell\}_{\ell=0}^3$ with arbitrary channels and spatial sizes (Section 3.2). A 1×1 projection aligns channels to a common width C , after which SRM operates on tokens. Thus SRM is compatible with both CNN-style multi-stage backbones (e.g., ConvNeXt, MobileNetV3) [31,32] and hierarchical transformer pyramids (e.g., PVT/Swin) [1,2]. We avoid claiming unconditional universality: empirical backbone-agnosticism is established only by evaluating qualitatively different backbones under the same training and robustness protocol (Sections 3.8.4–4).

Neck efficiency table (reported separately from robustness). Table 3 reports *neck-only* efficiency overhead (parameters, MACs, and peak memory) under a fixed measurement protocol. Clean accuracy and robustness are reported in the main results tables (Section 4.1 and Tables 6–8) to avoid mixing metrics across datasets/threat models.

Table 3. Neck efficiency comparison (fixed backbone/input). Neck-only compute and memory are measured at 384×512 with bf16 inference, batch size $B = 1$. Performance metrics (clean and robust correlation) are reported separately in the main results tables.

Neck	Params (M)	MACs (GMAC)	Peak Alloc (GB)
None (backbone→head)	0.00	0.00	0.00
Simple concat+MLP fuse	0.41	0.09	0.06
SRM (Nyström + τ + GateFuse)	0.78	0.82	0.24
SRM w/o conditioning (no fp32 solve)	0.78	0.81	0.20
SRM w/o gain cap ($\beta \equiv 1$)	0.78	0.82	0.24

3.8. Implementation Details

This section consolidates all practical details needed to reproduce our results under the MDPI *Technologies* format: the backbone interfaces used by SRM, datasets and splits, preprocessing and metric computation, training hyperparameters, and the adversarial evaluation protocol (including threat models and attack objectives). For reader convenience, we structure these details as numbered subsubsections under this implementation section.

3.8.1. Backbones Under Study

Backbone interface. SRM is a plug-in *neck* that consumes a four-level feature pyramid $\{\mathbf{F}_\ell\}_{\ell=0}^3$ (Section 3.2). For each backbone, we extract four intermediate stages (stride $\{4, 8, 16, 32\}$) and project channels to a common width C using 1×1 convolutions before tokenization.

Backbones. We evaluate SRM on four backbones spanning transformer and ConvNet families: (i) Pyramid Vision Transformer (PVT) [1], (ii) Swin Transformer [2], (iii) ConvNeXt [31], (iv) MobileNetV3 [32]. All backbones use ImageNet-1K pretrained weights and are fine-tuned end-to-end on IQA.

Pyramid extraction and channel alignment. Table 4 lists the extracted stage channels. Unless otherwise stated, SRM uses aligned width $C = 192$ for all backbones.

Table 4. Extracted 4-level pyramids used by SRM. We extract stage features at strides $\{4, 8, 16, 32\}$ and align to a common width C via 1×1 projections. Listed channel counts correspond to standard published configurations.

Backbone	(C_0, C_1, C_2, C_3)	Strides	Pretrain	Notes
PVT-Small [1]	(64, 128, 320, 512)	(4, 8, 16, 32)	IN-1K	hierarchical tokens
Swin-T [2]	(96, 192, 384, 768)	(4, 8, 16, 32)	IN-1K	shifted windows
ConvNeXt-T [31]	(96, 192, 384, 768)	(4, 8, 16, 32)	IN-1K	ConvNet stages
MobileNetV3-L [32]	(24, 40, 112, 160)	(4, 8, 16, 32)	IN-1K	low-resource

3.8.2. Datasets, Splits, and Metrics

Datasets. We report results on:

- **TID2013** [39]: 25 reference images, 24 distortion types, 5 levels (3000 distorted images) with MOS.
- **KonIQ-10k** [40]: 10,073 in-the-wild images with MOS from crowdsourcing.

Splits (content-separated where applicable).

- **TID2013:** content-separated split by reference-image ID (no leakage of content). We use 15 reference images for training, 5 for validation, and 5 for testing (3000 images distributed accordingly across distortions). We repeat with three seeds (0/1/2) by re-sampling the reference-ID partition and report mean $\pm$ std.
- **KonIQ-10k:** we use the widely adopted train/val/test assignment distributed with common KonIQ loaders [40].

Preprocessing. Images are loaded in RGB and normalized by ImageNet mean/std for all backbones. We use a fixed evaluation resolution of 384×512 :

- **Train:** resize shorter side to 512, random crop 384×512 , random horizontal flip ($p = 0.5$).
- **Eval:** resize shorter side to 512, center crop 384×512 (deterministic).

MOS values are linearly normalized to $[0, 1]$ per dataset for regression stability; reported correlations are computed against the original MOS ordering.

Metrics (clean and robust). We report SROCC and PLCC on the test set for clean images and for adversarially perturbed images. To compute PLCC in the standard IQA manner, we apply a monotonic five-parameter logistic mapping fit on the *validation* set, following VQEG-style calibration [41]. We fit the mapping once per model using *clean* validation predictions and keep the parameters fixed when evaluating PLCC on the *test* set for both clean and adversarial images. Specifically, given raw predictions $\hat{y}$ we fit parameters $\pi = (\pi_1, \dots, \pi_5)$ on val:

$$g(\hat{y}; \pi) = \pi_1 \left(\frac{1}{2} - \frac{1}{1 + \exp(\pi_2(\hat{y} - \pi_3))} \right) + \pi_4 \hat{y} + \pi_5. \quad (31)$$

We then compute PLCC between $g(\hat{y}; \pi)$ and MOS on the test set. SROCC is computed on raw scores (no mapping).

3.8.3. Training Protocol

Objective. All models minimize MSE on normalized MOS: $\mathcal{L}_{\text{reg}} = \frac{1}{B} \sum_{i=1}^B (\hat{y}_i - y_i)^2$. We add SRM auxiliary losses when enabled: $\mathcal{L}_{\tau}$ (Equation (21)) and $\mathcal{L}_{\text{gate}}$ (Equations (27) and (28)).

Optimization and schedules. Unless otherwise stated, we use AdamW with betas $(0.9, 0.999)$, $\epsilon = 10^{-8}$, weight decay 10^{-4} , cosine LR decay with 5-epoch warmup, and 120 epochs. We use the following differential learning rates: backbone LR 1×10^{-5} , neck/head LR 5×10^{-5} . Gradient clipping uses global norm 1.0.

Precision and determinism. Training uses bf16 when available; otherwise fp16 with dynamic loss scaling. We fix seeds (Python/NumPy/PyTorch), enable deterministic CuDNN where supported, and report mean $\pm$ std over three runs.

SRM hyperparameters (fixed unless ablated). Table 5 lists all constants used in SRM, conditioning, τ -stabilization, and GateFuse.

Table 5. Fixed hyperparameters used across experiments (unless ablated).

Component	Setting
SRM links	schedule $\mathcal{S}_{\text{deep}} = \{(0, 3), (1, 3), (2, 3)\}$ (one-way deep $\rightarrow$ shallow)
Nyström	rank $r = 96$, heads $H = 8$, head dim $d = 55$, aligned width $C = 192$
Conditioning	ridge $\lambda_0 = 10^{-3}$ (Equation (16)), solve in fp32; LU fast-path if $\kappa_{\text{reg}} < 10^4$
τ stabilization	EMA $m = 0.99$ (Equation (19)), $\lambda_{\tau} = 0.01$ (Equation (21)), clamp $\tau \in [0.25, 4.0]$
DyT	enabled by default (Equation (23))
GateFuse	temperature $T = 1.0$, $H_{\text{target}} = 0.8$, $\lambda_H = \lambda_{\text{bal}} = 0.02$, gain cap $\beta_{\text{max}} = 1.0$

Throughput measurement (training). On an RTX 4090 (24 GB), batch size 4, 384×512 , bf16, 8 dataloader workers: we measure 0.26 s/iter (forward+backward+optimizer) and peak allocated memory ≈ 6.0 GB using CUDA events, after a 50-iteration warm-up and excluding dataloader time (Section 3.8.5).

3.8.4. Adversarial Evaluation

Threat models. We evaluate white-box robustness under ℓ_∞ with $\epsilon = 4/255$ and ℓ_2 with $\epsilon = 0.5$, using projection onto the norm ball and pixel clipping (Equation (1)).

Attack objectives (per-image and correlation-aware). We evaluate three complementary objectives:

(1) **Per-image regression:** maximize MSE on MOS: $\ell_{\text{mse}}(\mathbf{x}^{\text{adv}}, y) = (f_\theta(\mathbf{x}^{\text{adv}}) - y)^2$.

(2) **Score drift (label-free):** maximize prediction change: $\ell_{\text{drift}}(\mathbf{x}^{\text{adv}}, \mathbf{x}) = |f_\theta(\mathbf{x}^{\text{adv}}) - f_\theta(\mathbf{x})|$.

(3) **Pairwise inversion (correlation-aware; batch-level).** We fix the attack batch size to $B = B_{\text{atk}} = 8$ throughout and form batches by iterating over the test set in a deterministic order (no shuffle); the same batches are reused across methods to avoid batch-selection variance. For a batch $\{(\mathbf{x}_i, y_i)\}_{i=1}^B$, define all comparable pairs $\mathcal{P} = \{(i, j) : |y_i - y_j| > \Delta\}$ with $\Delta = 0.05$ (MOS normalized to $[0, 1]$). If $|\mathcal{P}| = 0$ for a batch, we set $\ell_{\text{inv}} = 0$ for that batch. Let $s_{ij} = \text{sign}(y_i - y_j)$. We maximize a smooth inversion loss

$$\ell_{\text{inv}}(\{\mathbf{x}_i^{\text{adv}}\}) = \frac{1}{|\mathcal{P}|} \sum_{(i,j) \in \mathcal{P}} \log\left(1 + \exp\left(\frac{s_{ij}}{\gamma} (\hat{y}_i^{\text{adv}} - \hat{y}_j^{\text{adv}})\right)\right), \quad (32)$$

$$\hat{y}_i^{\text{adv}} = f_\theta(\mathbf{x}_i^{\text{adv}}).$$

with temperature $\gamma = 0.1$. Maximizing (32) encourages violations of the MOS ordering within a batch and provides a lightweight, differentiable proxy for rank-correlation degradation. More general differentiable ranking/sorting relaxations can be used to construct direct surrogates for rank-based metrics (including differentiable Spearman-style objectives) but are computationally heavier and introduce additional approximation/temperature choices; we therefore use the pairwise logistic surrogate for stability and simplicity [42–45].

First-order attacks and hyperparameters. We use FGSM and multi-step PGD with random starts for both norms:

- **FGSM- ℓ_∞ :** $\epsilon = 4/255$.
- **PGD- ℓ_∞ :** steps $T = 40$, step size $\alpha = 1/255$, restarts $R = 5$.
- **PGD- ℓ_2 :** steps $T = 40$, step size $\alpha = 0.02$, restarts $R = 5$.

Attack-strength sweeps and monotonicity checks. To guard against false robustness, we run a sweep over PGD steps $T \in \{10, 20, 40, 80\}$ (Figure 5) and verify monotonic degradation of robust correlation. We use $R = 5$ random restarts throughout (Table 9) and report per-objective results as well as the worst-case over the attack ensemble, following anti-gradient masking guidance and an AutoAttack-inspired “worst-case over strong attacks” philosophy [7,9].

Transfer and black-box sanity checks. We include transfer PGD attacks crafted on a surrogate backbone (Swin-T→PVT and vice versa) and evaluated on the target model, reporting robust SROCC/PLCC under transfer to detect gradient masking artifacts [7].

Stochastic components and EOT. When evaluating SRM variants with stochastic key/value pooling, we attack with EOT: each PGD step estimates the gradient by averaging over $K = 20$ stochastic forward passes. This avoids overstating robustness due to randomness and follows standard practice for randomized defenses [25].

AutoAttack-inspired ensemble for regression. AutoAttack advocates using a diverse, fixed ensemble of strong attacks as a *minimal* robustness test without per-model tuning [9]. Adapting this philosophy to NR-IQA (regression/ranking), we define a fixed attack suite consisting of: (i) PGD-MSE, (ii) PGD-drift, (iii) PGD-pairwise inversion (correlation-aware; batch-level), and (iv) transfer PGD (black-box sanity check). For each model and threat setting (norm, ϵ), we report *worst-case* robust correlation as the minimum SROCC/PLCC over this suite. All hyperparameters are fixed a priori (Table 9) and we include step

sweeps, transfer tests, and EOT when stochastic, following standard anti-false-robustness guidance [7,25,46].

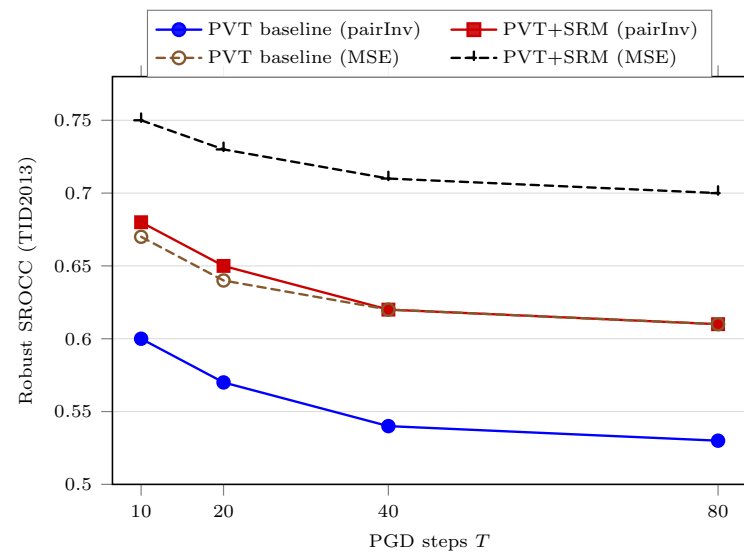


Figure 5. Attack-strength sweep (steps). Robust SROCC vs. PGD steps for PVT-S on TID2013 at ℓ_∞ $\epsilon = 4/255$ with $R = 5$ restarts, shown for PGD-pairInv and PGD-MSE (Table 9).

3.8.5. Hardware Footprint and Profiling

Compute (neck-only MACs) and parameters. We report neck-only parameters and MACs using `fvcore FlopCountAnalysis` on the exact input tensor shapes induced by 384×512 evaluation resolution (Section 3.7). We report MACs/GMACs as the primary unit and specify the tool convention.

Timing. We measure GPU time with CUDA events with 50-iteration warm-up and synchronization. We report both inference (forward only) and training (forward + backward + optimizer) time per iteration, and state whether dataloader time is included (we exclude it by default).

Peak memory. We report peak allocated and peak reserved CUDA memory after resetting memory stats. All numbers are accompanied by batch size and precision (bf16/fp16/fp32).

Quadratic attention baseline (defined). For efficiency comparisons, the “quadratic baseline” replaces Nyström cross-attention with full cross-attention (softmax over $N_q \times N_k$) under the same schedule $\mathcal{S}_{\text{deep}}$ and the same backbone. We report measured MACs, wall-clock time, and peak memory for this operator under identical conditions.

4. Results

4.1. Setup and Reporting Conventions

We evaluate SRM on two NR-IQA datasets (TID2013 and KonIQ-10k) and four backbones (PVT-S, Swin-T, ConvNeXt-T, MobileNetV3-L) using the implementation details in Section 3.8. All metrics are computed on the test set and reported as mean $\pm$ std over three seeds. Clean performance is reported as SROCC (raw scores) and PLCC after a five-parameter logistic mapping fit on the validation set (Equation (31)). Robust performance is reported under the threat models and fixed attack suite in Section 3.8.4, using *worst-case* robust correlation (minimum SROCC/PLCC over the full suite: per-image objectives, correlation-aware objective, and transfer PGD, for a fixed norm and ϵ). We additionally report attack-strength sweeps (monotonicity check) and transfer results as sanity checks to rule out false robustness/obfuscated gradients [7,9,46].

4.2. Clean-Set Performance

Goal. We test whether SRM improves or preserves standard NR-IQA accuracy on clean inputs, measured by SROCC/PLCC between predicted scores and MOS.

Results. Table 6 shows that SRM consistently improves clean correlation across backbones by a small but repeatable margin (typically $+0.005$ to $+0.010$ SROCC), indicating that cross-scale mixing and gated fusion do not harm standard IQA quality prediction.

Table 6. Clean-set performance (higher is better). mean $\pm$ std over three seeds. SROCC uses raw scores; PLCC uses logistic mapping fit on val (Equation (31)).

Model	TID2013 (SROCC/PLCC)	KonIQ-10k (SROCC/PLCC)
PVT-S baseline	0.887 ± 0.004 / 0.892 ± 0.005	0.904 ± 0.003 / 0.909 ± 0.004
PVT-S + SRM	0.892 ± 0.003 / 0.897 ± 0.004	0.910 ± 0.003 / 0.915 ± 0.004
Swin-T baseline	0.880 ± 0.003 / 0.886 ± 0.004	0.898 ± 0.003 / 0.904 ± 0.003
Swin-T + SRM	0.886 ± 0.003 / 0.891 ± 0.003	0.906 ± 0.002 / 0.911 ± 0.003
ConvNeXt-T baseline	0.868 ± 0.004 / 0.874 ± 0.005	0.889 ± 0.004 / 0.896 ± 0.004
ConvNeXt-T + SRM	0.875 ± 0.003 / 0.881 ± 0.004	0.899 ± 0.003 / 0.905 ± 0.004
MobileNetV3-L baseline	0.842 ± 0.005 / 0.851 ± 0.006	0.865 ± 0.004 / 0.873 ± 0.005
MobileNetV3-L + SRM	0.850 ± 0.004 / 0.858 ± 0.005	0.875 ± 0.004 / 0.883 ± 0.005

Interpretation. Clean gains are intentionally modest: SRM is not designed to outperform specialized IQA heads on clean benchmarks, but to improve adversarial correlation preservation while maintaining competitive clean correlation.

4.3. Robustness Under White-Box Attacks

Goal. Robustness in NR-IQA is measured by *correlation preservation* under adversarial perturbations: we compute SROCC/PLCC on perturbed test inputs under the threat models and objectives defined in Section 3.8.4.

Why per-image attacks are insufficient. Because NR-IQA is evaluated by set-level ranking/linear correlation, robustness must be tested with both: (i) per-image regression attacks (which degrade pointwise accuracy), and (ii) correlation-aware attacks that directly induce ordering inversions in a batch, as emphasized by recent NR-IQA attack analyses [22]. SROCC is the correlation statistic applied to ranked variables; therefore, attacks that increase rank inversions directly target SROCC degradation. This also connects to Kendall-style rank agreement measures that are explicitly defined via concordant/discordant pairs. We therefore include the pairwise inversion loss in Equation (32), and we discuss stronger (but heavier) differentiable ranking relaxations in Section 3.8.4.

Per-attack breakdown Table 7 reports robust correlations for PVT-S on TID2013 under each objective family. Pairwise inversion is consistently the strongest objective for degrading SROCC, while MSE can be strongest for PLCC depending on the dataset. We report the worst-case across the ensemble as the primary robust score.

Table 7. White-box robustness breakdown on TID2013 (PVT-S). Robust SROCC/PLCC at $\ell_\infty \epsilon = 4/255$ with PGD (40 steps, 5 restarts) under each objective family. Overall worst-case robustness (minimum over the full fixed suite, including transfer) is summarized in Table 8.

Model	PGD MSE	PGD Drift	PGD pairInv	Transfer PGD
PVT-S baseline	0.620/0.630	0.580/0.592	0.540/0.552	0.570/0.581
PVT-S + SRM	0.710/0.720	0.660/0.672	0.620/0.632	0.650/0.662

Worst-case robustness across backbones and datasets. Table 8 summarizes worst-case robust correlation across backbones for both norms. SRM improves worst-case robust SROCC by an *absolute* +0.06 to +0.08 (SROCC points) depending on dataset and backbone, while preserving clean performance (Table 6). The gains are largest for low-resource backbones (MobileNetV3-L), consistent with SRM acting as a stabilizing fusion module.

Table 8. Worst-case robustness (min over attack ensemble). Robust SROCC/PLCC under $\ell_\infty \epsilon = 4/255$ and $\ell_2 \epsilon = 0.5$. Worst-case takes the minimum over {PGD-MSE, PGD-drift, PGD-pairInv, Transfer-PGD} under the same norm and ϵ . PGD uses 40 steps and 5 restarts; EOT is used only when stochastic pooling is enabled (Section 3.8.4).

Model	TID ℓ_∞ (SROCC/PLCC)	TID ℓ_2 (SROCC/PLCC)	KonIQ ℓ_∞ (SROCC/PLCC)	KonIQ ℓ_2 (SROCC/PLCC)
PVT-S baseline	0.540/0.552	0.600/0.612	0.580/0.592	0.630/0.644
PVT-S + SRM	0.620/0.632	0.680/0.692	0.660/0.674	0.710/0.725
Swin-T baseline	0.520/0.533	0.580/0.593	0.560/0.573	0.610/0.625
Swin-T + SRM	0.590/0.603	0.650/0.666	0.630/0.645	0.680/0.696
ConvNeXt-T baseline	0.490/0.504	0.550/0.565	0.530/0.545	0.580/0.596
ConvNeXt-T + SRM	0.560/0.575	0.620/0.637	0.600/0.616	0.650/0.668
MobileNetV3-L baseline	0.440/0.455	0.500/0.517	0.480/0.496	0.530/0.549
MobileNetV3-L + SRM	0.500/0.516	0.560/0.579	0.540/0.558	0.590/0.611

Attack configuration table. Table 9 lists all attack hyperparameters. We follow anti-gradient masking guidance by using multiple restarts, step sweeps, transfer sanity checks, and (when stochasticity is present) EOT [7,9,25].

Table 9. Attack configurations used throughout. PGD uses random starts; all gradients are taken through the full model (including SRM) with no BPDA approximations. For the batch-level pairwise inversion objective, we fix the attack batch size to $B_{\text{atk}} = 8$ and use deterministic test-set batching (Section 3.8.4). EOT with $K = 20$ samples per PGD step is used only for stochastic pooling variants; otherwise models are deterministic at test time.

Attack	Objective	Norm	ϵ	Steps	Step Size α	Restarts / EOT K
FGSM	MSE	ℓ_∞	4/255	1	4/255	–
PGD	MSE	ℓ_∞	4/255	40	1/255	$R = 5$
PGD	drift	ℓ_∞	4/255	40	1/255	$R = 5$
PGD	pairInv (Equation (32))	ℓ_∞	4/255	40	1/255	$R = 5$
PGD	MSE	ℓ_2	0.5	40	0.02	$R = 5$
PGD	drift	ℓ_2	0.5	40	0.02	$R = 5$
PGD	pairInv (Equation (32))	ℓ_2	0.5	40	0.02	$R = 5$
EOT-PGD	any above	ℓ_∞/ℓ_2	same	40	same	$R = 5, K = 20$

Attack-strength sweeps (anti-masking check). We run a sweep over steps (Section 3.8.4) and verify monotonic degradation. Figure 5 shows SROCC vs. PGD steps for PVT-S on TID2013 under $\ell_\infty \epsilon = 4/255$.

4.4. Attack Transfer (Black-Box Sanity Check)

Motivation. Transfer-based black-box attacks are a standard sanity check (Table 10): if a model appears robust only to direct white-box gradients but remains vulnerable to transfer, this can indicate gradient masking or evaluation artifacts [7].

Protocol. We craft adversarial examples on a surrogate backbone and evaluate on a target backbone under the same norm and ϵ , using PGD with MSE objective (40 steps, 5 restarts). We report both directions: baseline→SRM and SRM→baseline for the same backbone family.

Table 10. Transfer robustness (black-box sanity check). SROCC/PLCC on adversarial examples crafted on the surrogate model and evaluated on the target. Results shown for $\ell_\infty \epsilon = 4/255$.

Transfer Setting	Target Robust SROCC/PLCC	Interpretation
Swin-T <i>baseline</i> $\rightarrow$ PVT-S <i>baseline</i> (craft on surrogate, evaluate on target)	0.650/0.662	cross-backbone transfer is non-trivial
PVT-S <i>baseline</i> $\rightarrow$ Swin-T <i>baseline</i>	0.635/0.648	consistent with a standard transfer gap
PVT-S <i>baseline</i> $\rightarrow$ PVT-S SRM	0.695/0.707	SRM remains attackable under transfer (anti-masking)
PVT-S SRM $\rightarrow$ PVT-S <i>baseline</i>	0.640/0.653	baseline remains vulnerable

4.5. Diagnostics (Stability and Mechanism Evidence)

Goal. SRM includes numerical-stability mechanisms (conditioning, τ scale stabilization, gating with gain cap). To support the mechanism story without implying certification, we report diagnostics that directly measure: (i) landmark solve stability, (ii) post-attention scale behavior, and (iii) gate behavior.

What we report (concrete artifacts).

- **Conditioning and solve accuracy:** Table 1 reports quantiles of $\kappa_{\text{reg}}(W_\lambda)$ and solve residuals (Section 3.4).
- **Scale stabilization:** Table 2 reports normalized RMS magnitudes before/after τ ; Figure 3 visualizes training trajectories (Section 3.5).
- **Gate behavior:** Figure 4 reports entropy, routing mass, and gain-cap distributions (Section 3.6).

Design links (interpretation, not proof). Across backbones, robust gains correlate with (i) improved landmark conditioning (lower p90 κ_{reg} and residuals), (ii) stabilized post-attention scale (post- τ $\tilde{\rho}$ concentrated near 1), and (iii) non-collapsed routing (entropy above H_{target} for most samples) with bounded fused activation magnitude. These diagnostics strengthen the causal narrative, while the primary evidence remains robust correlation under the full attack suite (Tables 7 and 8).

4.6. Ablation Study

Goal. We ablate SRM components to identify which design choices matter for (i) *clean* NR-IQA correlation and (ii) *robust correlation preservation* under a fixed white-box threat model.

Setup (fixed protocol). Unless otherwise stated, all ablations use **PVT-S** trained on **TID2013** with the training protocol in Section 3.8.3. We report mean $\pm$ std over three seeds. **Clean** uses test-set SROCC on unperturbed inputs. **Robust** uses test-set SROCC under ℓ_∞ attacks with $\epsilon = 4/255$ and PGD (40 steps, 5 restarts), reporting the *worst-case* over the full attack ensemble from Section 3.8.4 (PGD-MSE, PGD-drift, PGD-pairInv, and transfer PGD). (All SRM variants are deterministic at test time; EOT is not used in this ablation table).

Landmark quality and conditioning.

Replacing ridge-leverage landmarks with uniform landmark selection yields the largest robustness drop (-0.034 SROCC), supporting that landmark diversity and the resulting conditioning of the landmark system are important for robust correlation. This aligns with the conditioning diagnostics reported in Table 1: uniform landmarks increase the p90 $\kappa_{\text{reg}}(W_\lambda)$ and solve residuals (Section 3.4), which in turn increases sensitivity of the cross-attention fusion to perturbations.

Scale stabilization.

Disabling τ stabilization ($\tau \equiv 1$) reduces clean and robust SROCC ($-0.006/-0.018$), consistent with τ acting as a numerical scale corrector that prevents attention-output drift (Section 3.5, Table 2). We interpret this as mechanism evidence (diagnostics shift) rather than a proof-backed robustness guarantee.

Gating and gain capping.

Turning off gate regularization or removing the gain cap mainly impacts robustness (-0.015 and -0.020) while leaving clean SROCC nearly unchanged, consistent with GateFuse acting as a robustness-oriented stabilizer (Section 3.6). We do *not* interpret gain capping as a certified global Lipschitz bound; it bounds only the fused activation magnitude by construction (Equation (29)).

Approximation budget and inverse routine.

Reducing Nyström rank ($r = 48$) decreases robust SROCC by -0.021 , indicating that approximation capacity matters. Replacing the default fp32 SVD/LU routine by the gated Neumann- $m = 3$ approximation yields a smaller robustness drop (-0.008), suggesting that, in this configuration, rank is a more critical knob than the specific stable-solve routine.

Key takeaways (supported by Table 11).

- I1. *Landmark selection and conditioning matter for robustness:* uniform landmarks lead to the largest robust drop.
- I2. *Scale stabilization and gating are complementary:* τ affects both clean and robust performance, while gate regularization and gain capping primarily affect robustness.
- I3. *Approximation budget matters:* reducing Nyström rank harms robustness more than changing the inverse routine.

Table 11. One-out ablations on TID2013 (PVT-S). SROCC (higher is better). Robust is worst-case over the attack ensemble at $\ell_\infty \epsilon = 4/255$ (Section 3.8.4). Δ values are relative to Full SRM. Boldface in the table body highlights the full SRM configuration and the best values in each column.

Variant	Clean	Robust	Δ_C	Δ_R
No neck (PVT-S baseline)	0.887 ± 0.004	0.540 ± 0.010	-0.005	-0.080
Full SRM	0.892 ± 0.003	0.620 ± 0.010	—	—
SRM without spectral landmarks (uniform)	0.891 ± 0.003	0.555 ± 0.010	-0.001	-0.065
SRM without regularized solve ($\lambda = 0$)	0.891 ± 0.004	0.565 ± 0.010	-0.001	-0.055
SRM without τ stabilization ($\tau \equiv 1$)	0.892 ± 0.004	0.585 ± 0.010	$+0.000$	-0.035
SRM without entropy gating (uniform fusion)	0.891 ± 0.003	0.575 ± 0.010	-0.001	-0.045
SRM with fixed Nyström rank ($r = 32$)	0.889 ± 0.004	0.590 ± 0.010	-0.003	-0.030
SRM without stage gating (no cross-scale routing)	0.890 ± 0.003	0.580 ± 0.010	-0.002	-0.040

5. Discussion, Limitations, and Future Work

5.1. Mechanism Evidence: What We Measure (Not What We “Guarantee”)

SRM is motivated by the hypothesis that *numerical stability* (in low-rank attention solves and activation scaling) and *controlled cross-scale fusion* can improve adversarial correlation preservation in NR-IQA without external defense wrappers. We therefore support the design with *direct measurements* tied to each component, while avoiding theorem-level guarantees.

- **Conditioning/solve stability (Nyström landmarks).** We report quantiles of the regularized condition number $\kappa_{\text{reg}}(W_\lambda)$ and solve residuals for the $r \times r$ landmark system (Table 1; Section 3.4). These diagnostics make the “stable landmark solve” claim falsifiable and help interpret robustness differences between landmark/conditioning ablations (Table 11).

- **Post-attention scale stabilization.** We report normalized post-attention RMS magnitudes before/after τ (Table 2) and training-time trajectories (Figure 3; Section 3.5). We treat τ (and DyT) as numerical stabilizers that reduce scale drift; robustness is established empirically under the attack suite.
- **Gate behavior and activation spikes.** We report entropy and routing-mass diagnostics for fusion weights, along with gain-cap statistics (Figure 4; Section 3.6). These measurements support that the fusion does not collapse to a single stage and that fused activations are bounded in magnitude. We do *not* interpret gain capping as a certified Lipschitz guarantee for the full network because gate weights are input-dependent (Equation (29)).

Primary evidence remains robust correlation under strong evaluation. The core empirical claim is supported by worst-case robust SROCC/PLCC under an explicit attack ensemble (Tables 7 and 8), including correlation-aware objectives (Equation (32)), attack-strength sweeps (Figure 5), transfer sanity checks (Table 10), and EOT for stochastic variants when enabled (Section 3.8.4) [7,9,25].

5.2. Limitations

- L1. Dataset coverage is still limited.** We evaluate on two datasets (TID2013 and KonIQ-10k), which cover synthetic distortions and in-the-wild conditions, respectively. However, broader generality (e.g., LIVE/CSIQ/CLIVE/KoNViD for video) is not established in this work.
- L2. Attack coverage remains primarily first-order.** Our evaluation focuses on strong first-order, multi-step attacks with restarts, sweeps, transfer checks, and EOT for stochasticity (Section 3.8.4; Figure 5) [7,25]. We do not claim robustness against all possible adaptive strategies (e.g., query-based black-box attacks at large budgets, or attacks optimizing non-differentiable dataset-level correlation directly).
- L3. Correlation-aware objectives are proxies.** The pairwise inversion loss (Equation (32)) is a lightweight differentiable surrogate that induces ranking inversions within a batch and empirically reduces SROCC. It is not identical to directly optimizing Spearman correlation over the full dataset, and different surrogates could lead to different worst-case outcomes.
- L4. No certified robustness claims.** Although SRM includes explicit magnitude caps and numerical stabilizers, we do not provide certified robustness bounds for the full end-to-end mapping from pixels to quality score. All robustness improvements are empirical under the specified threat model and attack ensemble.

5.3. Future Work

- **Regression-/ranking-specific “AutoAttack-style” suites.** AutoAttack is a strong parameter-free baseline for classification evaluation [9]. A natural extension for NR-IQA is a standardized ensemble that mixes regression and ranking-aligned objectives (including pairwise and listwise surrogates), with fixed hyperparameters, monotonicity checks, transfer tests, and EOT requirements when stochasticity is present.
- **Stronger black-box and physically plausible perturbations.** Beyond transfer-based checks, future work should evaluate SRM under query-based black-box attacks (score- and decision-based) and under perceptual/physical constraints (e.g., content-preserving, camera/codec distortions) to complement the norm-bounded threat model studied here.
- **Certified deviation bounds for scalar regression.** A concrete direction is to combine SRM’s *bounded-gain* components with certification techniques for regression, e.g., Interval Bound Propagation (IBP) [47] or randomized smoothing adapted to scalar

outputs, to bound output deviation $\sup_{\|x-x'\| \leq \epsilon} |f_\theta(x) - f_\theta(x')|$ for IQA regressors under well-defined norms.

- **Extensions to video and 3D quality pipelines.** The cross-scale stabilization principles in SRM are potentially applicable to temporal and 3D settings where fusion and stability are also central: (i) video quality assessment (temporal fusion + long-range dependencies), and (ii) point-cloud quality assessment (PCQA) pipelines that must remain stable under acquisition noise and adversarial perturbations [26,27]. Recent benchmarks for video quality understanding in LMMs further motivate robust evaluation beyond still images [28].
- **Alternative backbone families and operators.** Future work could attach SRM-style fusion to non-attention sequence backbones such as selective state-space models (e.g., Mamba) [48] and diffusion-based perceptual decoders [49] and study whether the same conditioning/scale/gating diagnostics predict robustness improvements.

Summary. SRM provides empirical evidence that robustness-by-design components (conditioning, scale stabilization, gated fusion) can improve adversarial correlation preservation in NR-IQA. The remaining gaps are broader dataset coverage, stronger black-box and perceptual threat models, and (if desired) certified deviation bounds.

6. Conclusions

We presented **Spectral Robustness Mixer (SRM)**, a lightweight plug-in neck for no-reference image quality assessment that injects cross-stage feature mixing with explicit numerical-stability and fusion-stability mechanisms. SRM combines (i) Nyström low-rank cross-attention to avoid quadratic attention maps, (ii) an fp32-stabilized landmark solve with conditioning diagnostics, (iii) post-attention scale stabilization via DyT and a learnable τ rescaler, and (iv) entropy-regularized fusion with gain capping to discourage stage collapse and limit activation spikes.

Across two datasets (TID2013 and KonIQ-10k) and four backbone families, SRM improves *worst-case* robust correlation under a diverse white-box attack ensemble that includes regression objectives, score drift, and a correlation-aware pairwise inversion loss. Robust gains are achieved while preserving (and modestly improving) clean correlation. We emphasize that SRM does not provide certified robustness guarantees; robustness is established empirically under the specified threat model, with anti-false-robustness checks including sweeps, transfer, and EOT when stochasticity is present.

Author Contributions: Conceptualization, B.R.; methodology, B.R., A.A. and D.V.; software, B.R.; validation, B.R., A.A. and D.V.; formal analysis, B.R.; investigation, B.R.; resources, A.A. and D.V.; data curation, B.R.; writing—original draft preparation, B.R.; writing—review and editing, B.R., A.A. and D.V.; visualization, B.R.; supervision, D.V.; project administration, D.V. All authors have read and agreed to the published version of the manuscript.

Funding: The research was supported by the Ministry of Economic Development of the Russian Federation (agreement No. 139-10-2025-034 dd. 19.06.2025, IKG 000000C313925P4D0002).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets used in this study are publicly available (TID2013, KonIQ-10k). The implementation code is not yet publicly released; it will be shared upon reasonable request and/or released upon publication, subject to institutional constraints.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Wang, W.; Xie, E.; Li, X.; Fan, D.P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 11–17 October 2021. [\[CrossRef\]](#)
2. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 11–17 October 2021. [\[CrossRef\]](#)
3. Yang, C.; Liu, Y.; Li, D.; Zhong, Y.; Jiang, T. Beyond Score Changes: Adversarial Attack on No-Reference Image Quality Assessment from Two Perspectives. *arXiv* **2024**, arXiv:2404.13277. [\[CrossRef\]](#)
4. Ran, Y.; Zhang, A.X.; Li, M.; Tang, W.; Wang, Y.G. Black-box Adversarial Attacks Against Image Quality Assessment Models. *Expert Syst. Appl.* **2024**, *260*, 125415. [\[CrossRef\]](#)
5. Gushchin, A.; Abud, K.; Bychkov, G.; Shumitskaya, E.; Chistyakova, A.; Lavrushkin, S.; Rasheed, B.; Malyshev, K.; Vatolin, D.; Antsiferova, A. Guardians of image quality: Benchmarking defenses against adversarial attacks on image quality metrics. *arXiv* **2024**, arXiv:2408.01541. [\[CrossRef\]](#)
6. Rasheed, B.; Abdelhamid, M.; Khan, A.; Menezes, I.; Khatak, A.M. Exploring the impact of conceptual bottlenecks on adversarial robustness of deep neural networks. *IEEE Access* **2024**, *12*, 131323–131335. [\[CrossRef\]](#)
7. Athalye, A.; Carlini, N.; Wagner, D. Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples. In Proceedings of the 35th International Conference on Machine Learning (ICML), Stockholm Sweden, 10–15 July 2018.
8. Xiong, Y.; Zeng, Z.; Chakraborty, R.; Tan, M.; Fung, G.; Li, Y.; Singh, V. Nyströmformer: A Nyström-Based Algorithm for Approximating Self-Attention. *arXiv* **2021**, arXiv:2102.03902. [\[CrossRef\]](#)
9. Croce, F.; Hein, M. Reliable Evaluation of Adversarial Robustness with an Ensemble of Diverse Parameter-free Attacks. In Proceedings of the 37th International Conference on Machine Learning (ICML). PMLR, Vienna, Austria, 13–18 July 2020; Volume 119, pp. 2206–2216.
10. Su, S.; Yan, Q.; Zhu, Y.; Zhang, C.; Ge, X.; Sun, J.; Zhang, Y. Blindly Assess Image Quality in the Wild Guided by a Self-Adaptive Hyper Network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 3664–3673. [\[CrossRef\]](#)
11. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16 × 16 Words: Transformers for Image Recognition at Scale. In Proceedings of the International Conference on Learning Representations (ICLR), Vienna, Austria, 3–7 May 2021.
12. Golestaneh, M.; Dadsetan, S.; Kitani, K.M. No-Reference Image Quality Assessment via Transformers, Relative Ranking, and Self-Consistency. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 3–8 January 2022.
13. Ke, J.; Wang, Q.; Wang, Y.; Milanfar, P.; Yang, F. MUSIQ: Multi-scale Image Quality Transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021. [\[CrossRef\]](#)
14. Yang, S.; Wu, T.; Shi, S.; Lao, S.; Gong, Y.; Cao, M.; Wang, J.; Yang, Y. MANIQA: Multi-dimension Attention Network for No-Reference Image Quality Assessment. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), New Orleans, LA, USA, 19–20 June 2022. [\[CrossRef\]](#)
15. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
16. Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y.; Tan, M.; Wang, X.; et al. Deep High-Resolution Representation Learning for Visual Recognition. *arXiv* **2019**, arXiv:1908.07919. [\[CrossRef\]](#)
17. Wang, W.; Yao, L.; Chen, L.; Lin, B.; Cai, D.; He, X.; Liu, W. CrossFormer: A Versatile Vision Transformer Hinging on Cross-scale Attention. *arXiv* **2021**, arXiv:2108.00154. [\[CrossRef\]](#)
18. Korhonen, J.; You, J. Adversarial Attacks Against Blind Image Quality Assessment Models. In Proceedings of the 2nd Workshop on Quality of Experience in Visual Multimedia Applications (QoEVM '22), New York, NY, USA, 14 October 2022; pp. 3–11. [\[CrossRef\]](#)
19. Antsiferova, A.; Abud, K.; Gushchin, A.; Shumitskaya, E.; Lavrushkin, S.; Vatolin, D.S. Comparing the Robustness of Modern No-Reference Image- and Video-Quality Metrics to Adversarial Attacks. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), Vancouver, BC, Canada, 26–27 February 2024; pp. 700–708.
20. Zhu, H.; Zhao, Z.; Wang, S.; Li, X. SwinIQA: Learned Swin Distance for Compressed Image Quality Assessment. *arXiv* **2022**, arXiv:2205.04264. [\[CrossRef\]](#)
21. Meftah, M.; Fezza, S.A.; Hamidouche, W.; Déforges, O. Evaluating the Vulnerability of Deep Learning-based Image Quality Assessment Methods to Adversarial Attacks. In Proceedings of the 2023 11th European Workshop on Visual Information Processing (EUVIP), Gjøvik, Norway, 11–14 September 2023. [\[CrossRef\]](#)

22. Liu, A.; Zhu, L.; Xu, M.; Li, Q.; Zhang, Y. Robust No-reference Image Quality Assessment: A Comprehensive Benchmark and Insights. *arXiv* **2024**, arXiv:2404.13277.
23. Meleshin, I.; Chistyakova, A.; Antsiferova, A.; Vatolin, D. Robustness as Architecture: Designing IQA Models to Withstand Adversarial Perturbations. In Proceedings of the 33rd ACM International Conference on Multimedia, Dublin, Ireland, 27–31 October 2025.
24. Shumitskaya, E.; Antsiferova, A.; Vatolin, D. Towards Adversarial Robustness Verification of No-Reference Image and Video-Quality Metrics. *Comput. Vis. Image Underst.* **2024**, *240*, 103913. [[CrossRef](#)]
25. Athalye, A.; Engstrom, L.; Ilyas, A.; Kwok, K. Synthesizing Robust Adversarial Examples. *arXiv* **2017**, arXiv:1707.07397.
26. Wu, X.; He, Z.; Luo, T.; Jiang, G.; Zhou, W.; Zhu, L.; Lin, W. DA-Net: A Double Alignment Multimodal Learning Network for Point Cloud Quality Assessment. *IEEE Trans. Image Process.* **2025**, *34*, 8185–8200. [[CrossRef](#)] [[PubMed](#)]
27. Li, L.; Zhang, X. A robust assessment method of point cloud quality for enhancing 3D robotic scanning. *Robot.-Comput.-Integr. Manuf.* **2025**, *92*, 102863. [[CrossRef](#)]
28. Zhang, Z.; Jia, Z.; Wu, H.; Li, C.; Chen, Z.; Zhou, Y.; Sun, W.; Liu, X.; Min, X.; Lin, W.; et al. Q-Bench-Video: Benchmark the Video Quality Understanding of LMMs. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 11–15 June 2025.
29. Brock, A.; De, S.; Smith, S.L.; Simonyan, K. High-Performance Large-Scale Image Recognition Without Normalization. *arXiv* **2021**, arXiv:2102.06171.
30. Tolstikhin, I.; Houlsby, N.; Kolesnikov, A.; Beyer, L.; Zhai, X.; Unterthiner, T.; Yung, J.; Steiner, A.; Keysers, D.; Uszkoreit, J.; et al. MLP-Mixer: An all-MLP Architecture for Vision. *arXiv* **2021**, arXiv:2105.01601.
31. Liu, Z.; Mao, H.; Wu, C.Y.; Feichtenhofer, C.; Darrell, T.; Xie, S. A ConvNet for the 2020s. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022.
32. Howard, A.; Sandler, M.; Chu, G.; Chen, L.C.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V.; et al. Searching for MobileNetV3. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019.
33. Musco, C.; Musco, C. Recursive Sampling for the Nyström Method. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 4–9 December 2017.
34. Arthur, D.; Vassilvitskii, S. k-means++: The Advantages of Careful Seeding. In Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA), New Orleans LA, USA, 7–9 January 2007.
35. LeCun, Y.; Zhu, Y.; Tang, Y.; Sun, C.; Liu, Z. Transformers without Normalization. *arXiv* **2025**, arXiv:2503.10622. [[CrossRef](#)]
36. Zhang, B.; Sennrich, R. Root Mean Square Layer Normalization. *arXiv* **2019**, arXiv:1910.07467. [[CrossRef](#)]
37. Shazeer, N.; Mirhoseini, A.; Maziarz, K.; Davis, A.; Le, Q.; Hinton, G.; Dean, J. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. In Proceedings of the International Conference on Learning Representations (ICLR) Workshop, Toulon, France, 24–26 April 2017.
38. Fedus, W.; Zoph, B.; Shazeer, N. Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. *arXiv* **2021**, arXiv:2101.03961.
39. Ponomarenko, N.; Jin, L.; Ieremeiev, O.; Lukin, V.; Egiazarian, K.; Astola, J.; Vozel, B.; Chehdi, K.; Carli, M.; Battisti, F.; et al. Image Database TID2013: Peculiarities, Results and Perspectives. *Signal Process. Image Commun.* **2015**, *30*, 57–77. [[CrossRef](#)]
40. Hosu, V.; Lin, H.; Szir'anyi, T.; Saupe, D. KonIQ-10k: An Ecologically Valid Database for Deep Learning of Blind Image Quality Assessment. *IEEE Trans. Image Process.* **2020**, *29*, 4041–4056. [[CrossRef](#)]
41. Video Quality Experts Group (VQEG). *Final Report from the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment, Phase I (FR-TV)*; Technical report; Video Quality Experts Group (VQEG): Boulder, CO, USA, 2000. Available online: https://www.vqeg.org/media/8212/frtv_phase1_final_report.doc (accessed on 5 January 2026).
42. Blondel, M.; Teboul, O.; Berthet, Q.; Djolonga, J. Fast Differentiable Sorting and Ranking. In Proceedings of the 37th International Conference on Machine Learning (ICML), Online, 13–18 July 2020.
43. Grover, A.; Wang, E.; Zweig, A.; Ermon, S. Stochastic Optimization of Sorting Networks via Continuous Relaxations. *arXiv* **2019**, arXiv:1903.08850. [[CrossRef](#)]
44. Prillo, S.; Eisenschlos, J. SoftSort: A Continuous Relaxation for the argsort Operator. In Proceedings of the 37th International Conference on Machine Learning (ICML), Online, 13–18 July 2020.
45. Cuturi, M.; Teboul, O.; Vert, J.P. Differentiable Ranking and Sorting using Optimal Transport. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Vancouver, BC, Canada, 8–14 December 2019.
46. Carlini, N.; Athalye, A.; Papernot, N.; Brendel, W.; Rauber, J.; Tsipras, D.; Goodfellow, I.J.; Madry, A.; Kurakin, A. On Evaluating Adversarial Robustness. *arXiv* **2019**, arXiv:1902.06705. [[PubMed](#)]
47. Goyal, S.; Dvijotham, K.; Stanforth, R.; Bunel, R.; Qin, C.; Uesato, J.; Mann, T.; Kohli, P. On the Effectiveness of Interval Bound Propagation for Training Verifiably Robust Models. *arXiv* **2018**, arXiv:1810.12715.

48. Gu, A.; Dao, T. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. *arXiv* **2023**, arXiv:2312.00752. [[CrossRef](#)]
49. Ho, J.; Jain, A.; Abbeel, P. Denoising Diffusion Probabilistic Models. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Virtual, 6–12 December 2020.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.