TRANSDIFF: PLUG-AND-PLAY ADVERSARIAL DEFENSE FOR NO-REFERENCE IMAGE QUALITY ASSESSMENT VIA CALIBRATED SCORE CONSENSUS

Anonymous authors
Paper under double-blind review

ABSTRACT

Adversarial perturbations can substantially alter the predictions of no-reference image quality assessment (NR-IQA) models while remaining nearly imperceptible. Because these models guide image processing and quality-based selection, manipulated scores can compromise downstream decisions without corresponding changes in perceived quality. We introduce TransDiff, a method for estimating a model's pre-attack score directly from an attacked image. The method evaluates multiple transformed versions of the image to obtain complementary observations related to its original score. Crucially, transformations themselves affect quality predictions. We account for these effects using score-shift distributions calibrated on images without adversarial perturbations, then combine the calibrated observations through distributional agreement. Our theoretical analysis establishes conditions under which this procedure accurately recovers the pre-attack score despite image mismatch, finite calibration and transformation sampling, residual attack effects, and optimization error. TransDiff retains the shape of the calibrated score-shift distributions rather than reducing them to low-order statistics. Experiments on KADID-10k and KonIQ-10k across nine NR-IQA models and white-box and black-box attacks show that TransDiff improves score recovery over test-time image defenses and matched score-recovery baselines. Relative to a matched Gaussian-shift baseline, it reduces aggregate recovery error by 11.5% under white-box attacks and 7.5% under black-box attacks on the nested stress subset. These results demonstrate the value of retaining calibrated score-shift distribution shape for recovering a model's original prediction.

1 INTRODUCTION

No-reference image quality assessment (NR-IQA) metrics are increasingly used as scalar decision signals in image-processing, monitoring, and quality-control pipelines. This creates an unusual adversarial vulnerability: a perturbation can substantially change the predicted quality score while leaving perceived image quality nearly unchanged (Antsiferova et al., 2024; Korhonen & You, 2022; Yang et al., 2024). In such systems, manipulating the score alone can be enough to alter downstream decisions.

NR-IQA is also awkward for conventional test-time defenses. Image purification suppresses adversarial perturbations by modifying the input, but blur, denoising, compression, texture changes, and reconstruction artifacts are themselves properties that an image-quality metric is meant to measure. The original image may already contain genuine distortions, whose effect on the metric should be preserved. We therefore consider a different objective: given an attacked image x_{adv} and a frozen metric f, recover the unavailable pre-attack prediction $y^* = f(x)$ rather than reconstructing the original image or re-estimating perceptual quality.

Transformations provide a natural source of information about this latent score. Evaluating transformed versions of x_{adv} yields multiple observations of the same underlying image, but these observations cannot simply be averaged. NR-IQA metrics are not invariant to cropping, resizing, filtering, geometric changes, or mild photometric perturbations. A transformed prediction $a_\phi = f(\phi(x_{\mathrm{adv}}))$

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

therefore reflects the unknown original score, the transformation's normal effect on the metric, and the residual attack effect.

Our key idea is to treat each transformation as a *calibrated measurement channel*. Before deployment, for every transformation ϕ we measure its clean score shifts $\Delta_\phi = f(\phi(X)) - f(X)$ over calibration images and retain their distribution. At inference, a candidate original score v implies a shift $a_\phi - v$. If v is compatible with the original prediction, these implied shifts should be typical under the calibrated transformation responses. TransDiff aggregates this agreement across transformed views and returns the candidate with the strongest support. The metric remains frozen, no recovery network is trained, and no image reconstruction is required.

Calibration must capture more than a single bias. Transformation-induced score shifts can be asymmetric, multimodal, or contain low-density regions that distinguish otherwise plausible candidate scores. TransDiff therefore retains the empirical shape of each shift distribution through kernel smoothing rather than reducing it to location or variance. We isolate the value of this information using matched baselines that share the same transformations, calibration data, observations, and inference budget but replace the empirical distributions with simpler summaries.

Our analysis characterizes when this recovery is possible. At population level, the correct score uniquely maximizes the expected agreement objective. For a particular attacked image, this recovery signal competes with image-specific mismatch, finite calibration error, transformation sampling, residual attack effects, and optimization error. The analysis separates these terms and exposes an important limitation: more calibration data and transformed views reduce sampling errors, but image-specific effects shared across views and coherent attack-induced shifts do not vanish simply by adding more transformations.

We evaluate TransDiff on KADID-10k and KonIQ-10k across nine frozen NR-IQA metrics under white-box and black-box attacks. The default method uses only four transformations sampled from a fixed bank of 48. In the white-box setting, the attacker knows the metric, transformation bank, calibration data, sampling distribution, and recovery rule, and optimizes the expected defended score with expectation over transformations; inference uses fresh samples. TransDiff substantially improves score recovery over existing test-time image defenses while introducing less drift on unattacked inputs. In a matched comparison, retaining the empirical shift distributions reduces recovery error by 11.5% relative to a moment-matched Gaussian model under white-box attacks and by 7.5% under black-box attacks.

Our contributions are:

- We formulate adversarial robustness of frozen NR-IQA metrics as *pre-attack score recovery*, separating recovery of the metric's original output from image purification or re-estimation of perceptual quality.
- We introduce TransDiff, which treats image transformations as calibrated measurement channels and recovers the latent score through distributional agreement across randomly sampled transformed views, without retraining the metric or reconstructing the image.
- We characterize when recovery succeeds by separating the population signal from image mismatch, calibration and sampling errors, residual attack effects, and optimization error, and validate the method across nine NR-IQA metrics, two datasets, and white-box and black-box attacks.

2 Related work

Adversarial robustness of NR-IQA Attacks on NR-IQA target individual scores as well as their ordering across images (Antsiferova et al., 2024; Yang et al., 2024). Existing responses include training with gradient regularization (Liu et al., 2024), median randomized smoothing (Shumitskaya et al., 2025a), and feature-space smoothing (Shumitskaya et al., 2025b); broader comparisons cover training, purification, and certification (Gushchin et al., 2025). TransDiff instead recovers the pre-attack output of a frozen metric using calibrated transformation-induced score shifts.

Transformation aggregation and smoothing E-LPIPS studies random transformation ensembles for perceptual similarity (Kettunen et al., 2019), while IQA smoothing aggregates randomized

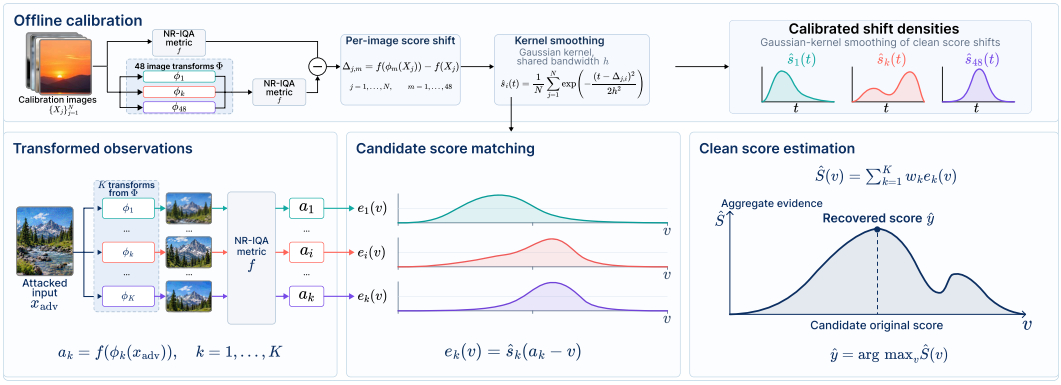

Figure 1: **Overview of TransDiff**. Offline, score-shift distributions are calibrated for each transformation using clean images. At test time, transformed scores of the adversarial image are converted into evidence over candidate clean scores using the corresponding calibrated distributions. Evidence from multiple transformations is then aggregated, and the maximizer is returned as the recovered score.

predictions or representations (Shumitskaya et al., 2025a;b). Interval fusion provides another established approach to combining continuous measurements (Marzullo, 1990; Brooks & Iyengar, 1996). These ideas motivate our aggregation and interval baselines. Our comparison asks what is gained by retaining the empirical shift distributions when the observations, calibration images, and metric-evaluation budget are fixed.

Distributional estimation Kernel smoothing represents a distribution at a chosen resolution, and maximum-correntropy estimation uses bounded kernel similarity for robust fitting (Bako, 2018). Maximum mean discrepancy (MMD) compares distributions through kernel mean embeddings (Gretton et al., 2012). TransDiff applies these tools to pre-attack score recovery. MMD characterizes the population margin; image mismatch captures how the shift pattern of a particular image can differ from the calibration population, including through shared effects across transformations.

Diffusion purification Diffusion purification reconstructs partially noised inputs to reduce adversarial effects (Nie et al., 2022; Zhang et al., 2023), but reconstruction can also alter genuine image-quality attributes. TransDiff instead returns a scalar estimate of the pre-attack metric score.

3 Method

3.1 Recovering the pre-attack score

The target is the pre-attack score $y^* = f(x)$, and the observations are the transformed predictions $a_\phi = f(\phi(x_{\text{adv}}))$. A candidate score v implies the shift $a_\phi - v$ for transformation ϕ. TransDiff compares these implied shifts with clean calibration and combines the agreement across transformations.

3.2 Target and calibration

Let $f : \mathcal{X} \to \mathbb{R}$ be a frozen NR-IQA metric, x the image before attack, and $x_{\text{adv}} = x + \delta$ the observed input. We recover the pre-attack score $y^* = f(x)$. Fix a bank Φ of K transformations and weights $q_\phi \geq 0$ summing to one. Each transformation includes all parameters and any internal random seed. On n iid clean calibration images, compute the score shifts

$$\Delta_{\phi,j} = f(\phi(X_j)) - f(X_j). \tag{1}$$

The same images are used for every transformation, preserving the shift matrix for subsequent analysis. No independence across transformations is assumed. The bank, weights, and a common bandwidth $h > 0$ are fixed on separate development data. Calibration is repeated for each metric; its weights remain unchanged.

3.3 AGREEMENT AND RECOVERY

For $k_h(u, v) = \exp(-(u-v)^2/(2h^2))$, define

$$\widehat{s}_\phi(t) = \frac{1}{n}\sum_{j=1}^{n} k_h(t, \Delta_{\phi,j}). \tag{2}$$

For an observed prediction $a_\phi = f(\phi(x_{\text{adv}}))$, the candidate v implies the shift $a_\phi - v$. The quantity $\widehat{s}_\phi(a_\phi - v)$ measures its agreement with the calibrated shifts. It retains modes, asymmetry, and low-density gaps at resolution h. The Gaussian kernel smooths the empirical distribution; it does not assume Gaussian score shifts. Up to the common factor $h\sqrt{2\pi}$, Eq. (2) is a Gaussian kernel density estimate. Because h is shared across transformations, this factor does not affect the maximizing score.

By default, TransDiff draws $m = 4$ transformations independently from the calibrated 48-entry bank according to q, with replacement. Full-bank inference evaluates all 48 transformations and is reported separately as a view-budget ablation. More generally, sampled inference draws $\phi_1, \ldots, \phi_m$ iid from q and uses their empirical frequencies. Write $w_\phi = q_\phi$ in the full-bank case and $w_\phi = m^{-1}\sum_i \mathbf{1}\{\phi_i = \phi\}$ in the sampled case. Both use

$$\widehat{S}(v) = \sum_{\phi\in\Phi} w_\phi \widehat{s}_\phi(a_\phi - v), \qquad \widehat{y} \in \arg\max_{v\in\mathbb{R}} \widehat{S}(v). \tag{3}$$

Thus TransDiff returns the candidate with the largest aggregate agreement. It requires no learned recovery model.

Global score selection. The recovery objective is one-dimensional but can be multimodal, so we search for its global rather than a local maximum. We return a score whose agreement is within ρ of the optimum,

$$\widehat{S}(\widehat{y}) \geq \sup_v \widehat{S}(v) - \rho, \tag{4}$$

where ρ is the optimization error used in the analysis below. In practice, a grid over the data-dependent range of candidate scores provides a controlled value of ρ; Appendix B.2 gives the search interval and optimization tolerance.

4 WHY RECOVERY WORKS

For a clean image, replacing the pre-attack score y^* by $y^* + d$ moves every implied shift by $-d$. We first measure the population margin against this score error, then bound how much the margin can change for a fixed attacked image.

4.1 POPULATION RECOVERY SIGNAL

Let P_ϕ be the score-shift distribution for transformation ϕ, and let $s_\phi(t) = \mathbb{E}_{D\sim P_\phi} k_h(t, D)$ denote its population agreement. For independent shifts $R, D \sim P_\phi$, define

$$G(d) = \mathbb{E}_{\phi\sim q}\mathbb{E}_{R,D\,\text{iid}\sim P_\phi} k_h(R - D, d), \quad H(d) = G(0) - G(d). \tag{5}$$

Here $G(d)$ is the expected aggregate agreement at score error d, and $H(d)$ is the population margin over that candidate. The margin has an exact expression in terms of maximum mean discrepancy (MMD) (Gretton et al., 2012).

Proposition 1 (Population identification)**.** *Let $T_d P$ denote translation of a distribution by d. Then*

$$H(d) = \tfrac{1}{2}\mathbb{E}_{\phi\sim q}\,\mathrm{MMD}^2_{k_h}(P_\phi, T_d P_\phi). \tag{6}$$

Consequently, G has its unique maximum at zero. For every $r > 0$, $\inf_{|d|\geq r} H(d) > 0$.

Thus the population objective uniquely identifies y^*. The size of $H(d)$ determines how strongly a score error is rejected. While distributional shape can help distinguish candidates from Gaussian or location-only baselines, broad distributions or competing modes can make the margin small. Appendix A.1 gives the proof.

4.2 IMAGE-SPECIFIC MISMATCH

Calibration describes how transformations change scores *on average* over clean images, whereas inference must recover the score of one particular image. To see the gap, suppose every transformation has shift $\Delta_\phi(X) = b_\phi + U_X$, where b_ϕ is its characteristic effect and the same $U_X \sim \mathcal{N}(0, \sigma^2)$ affects every transformation of image X. Even without an attack, the observations on image x are $a_\phi = y^* + b_\phi + U_x$. Each calibrated agreement curve then peaks at $v = y^* + U_x$, so their sum does too. More calibration data can identify the marginal shift distributions exactly, and more views can average away sampling noise, but neither separates U_x from y^*.

We use χ_x to measure this gap between the behavior of image x and the calibration population. Let X, X' be independent clean images and let $\phi \sim q$ be independent of them, with the same transformation used within each difference:

$$\begin{aligned} Q_x &= \mathcal{L}(\Delta_\phi(x) - \Delta_\phi(X')), \\ Q_0 &= \mathcal{L}(\Delta_\phi(X) - \Delta_\phi(X')), \\ \chi_x &= \mathrm{MMD}_{k_h}(Q_x, Q_0). \end{aligned} \tag{7}$$

Q_x contains shift differences involving the target image, while Q_0 contains the corresponding differences between two random calibration images. Their kernel distance χ_x measures how far the target image's pattern deviates from ordinary clean-image variation. It remains present even if every P_ϕ is known exactly; Appendix A.2 derives its contribution to the recovery bound.

4.3 FROM SIGNAL TO FINITE-SAMPLE RECOVERY

Population recovery provides the signal $H(d)$ against a candidate with score error d. Conceptually, recovery requires this signal to exceed image mismatch, calibration error, sampling error, the attack effect, and optimization error. The theorem below makes this decomposition precise.

The attack shift for transformation ϕ is

$$t_\phi = f(\phi(x_{\mathrm{adv}})) - f(\phi(x)), \qquad \tau = \mathbb{E}_{\phi \sim q}|t_\phi|. \tag{8}$$

τ is its average magnitude across transformations. If $L \geq \sup_{\phi,t} |s'_\phi(t)|$ and $M \geq \sup_{\phi,t} |s''_\phi(t)|$, the attack changes the agreement difference between scores d apart by at most $\tau \min\{2L, M|d|\}$. The Gaussian kernel gives $L = e^{-1/2}/h$ and $M = 1/h^2$. A shared attack shift is an exact failure mode: if $t_\phi = t$ for every active transformation, then $\widehat{S}_{\mathrm{adv}}(v) = \widehat{S}_{\mathrm{clean}}(v - t)$ for the same active weights, so the maximizing scores shift by t.

Finite calibration and sampled inference add two errors. For failure probabilities $\alpha \in (0, 1)$ and, for sampled inference, $\beta \in (0, 1)$, set

$$\begin{aligned} \eta_n &= \frac{1 + \sqrt{2\log(K/\alpha)}}{\sqrt{n}}, \\ \varepsilon_m &= \frac{1 + \sqrt{2\log(1/\beta)}}{\sqrt{m}}, \qquad B_{n,m} = \chi_x + \eta_n + \varepsilon_m. \end{aligned} \tag{9}$$

Here η_n is the calibration error and ε_m is the sampling error. Full-bank inference sets $\varepsilon_m = 0$ and omits β. Together with the optimization error ρ from Eq. (4), these terms yield the finite-sample guarantee.

Theorem 1 (Finite-sample recovery)**.** *Fix x, the bank, q, and h, independently of the iid calibration sample. An attack may depend on that sample. For sampled inference, draw the transformations after the attacked input is fixed. With probability at least $1-\alpha-\beta$, the output of Eq. (4), with $\widehat{d} = \widehat{y} - y^*$, satisfies*

$$\boxed{\begin{aligned} &H(\widehat{d}) \le w_h(\widehat{d}) B_{n,m} + \tau \min\{2L, M|\widehat{d}|\} + \rho, \\ &w_h(d) = \sqrt{2(1 - e^{-d^2/(2h^2)})}. \end{aligned}} \tag{10}$$

For full-bank inference the probability is at least $1-\alpha$. On this event, $|\widehat{y} - y^*| < r$ whenever

$$\inf_{|d| \ge r} [H(d) - w_h(d) B_{n,m} - \tau \min\{2L, M|d|\}] > \rho. \tag{11}$$

Equation (11) rules out every score error beyond r whose population margin exceeds the combined image, calibration, sampling, attack, and optimization errors. The infimum over $|d| \ge r$ is essential because a distant mode may compete with the peak near zero. The proof follows the same decomposition: Proposition 1 supplies the signal; the first three errors perturb the agreement comparison in the Gaussian RKHS; smoothness controls the attack effect; and ρ accounts for approximate maximization. Appendix A gives the complete argument and local scaling.

The theorem is a finite-sample explanation of when the method succeeds, not an extra inference procedure: χ_x, τ, and the other diagnostic terms are not required to compute $\widehat{y}$.

5 Experiments

We evaluate whether TransDiff recovers the pre-attack metric score better than existing defenses, whether the shape of the calibrated shift distributions helps beyond matched summaries, and how recovery changes with attack strength and the number of transformed views. Unless stated otherwise, TransDiff uses four transformations sampled with replacement from the fixed 48-entry bank. The same four-view budget applies to its transformation-based score-recovery baselines.

5.1 Experimental setup

Datasets. We use KADID-10k for synthetic distortions and KonIQ-10k for authentic distortions (Lin et al., 2019; Hosu et al., 2020), matching the evaluation domain of the prior robustness benchmark (Gushchin et al., 2025). The main comparison uses 1,000 images per dataset; a nested subset of 200 images per dataset is used for the matched score-recovery comparison and black-box attacks. The PaQ-2-PiQ image dataset (Ying et al., 2020) supplies disjoint development, calibration, and reference-image splits, separate from the evaluation datasets. KADID images sharing an undistorted reference are grouped for sampling and analysis. "Clean" denotes an image before attack, which may itself be distorted.

NR-IQA metrics. We evaluate nine frozen NR-IQA metrics: Meta-IQA (Zhu et al., 2020), MANIQA (Yang et al., 2022), CLIP-IQA+ (Wang et al., 2023), TOPIQ (Chen et al., 2024), KonCept (Hosu et al., 2020), SPAQ (Fang et al., 2020), PAQ2PIQ (Ying et al., 2020), Linearity (Li et al., 2020), and FPR (Chen et al., 2022). PAQ2PIQ here denotes the metric, distinct from the PaQ-2-PiQ image dataset. We use the fixed score bounds of the prior benchmark (Gushchin et al., 2025) (Appendix D). Every score is oriented so that larger values indicate higher quality.

Calibration protocol. For each metric, calibration evaluates all 48 entries of a differentiable transformation bank on a disjoint set of PaQ-2-PiQ images. The bank and bandwidth are fixed using the development split. Sampling is uniform over the 48 bank entries. At inference, we draw four transformations with replacement; Figure 3 also studies larger samples and the deterministic full bank. Attack optimization and final evaluation use independent draws from the same bank. Appendix B specifies the bank.

Attacks. The white-box suite contains I-FGSM, Optimised-UAP, the attacks of Korhonen & You (2022) and Yang et al. (2024), and FACPA. Each attack seeks to increase the final defended score and

Table 1: Comparison with test-time defenses across both datasets and nine NR-IQA metrics. Clean and one-run white-box results use the main subset (1,000 images/dataset); black-box results use the nested stress subset (200/dataset). $D = D_{\text{score}}$, $R = R_{\text{score}}$; SROCC is measured against MOS. Cost is inference time per image. Bold and underlines mark the best and second-best values in each column, respectively; the clean no-defense error is zero by definition.

	Clean (main)		White-box (main)			Black-box (stress)			Cost
Method	$D\downarrow$	SROCC↑	$D\downarrow$	$R\uparrow$	SROCC↑	$D\downarrow$	$R\uparrow$	SROCC↑	ms/img↓
No defense	**0.00**	**0.624**	64.29	0.57	0.471	35.26	0.66	**0.541**	—
Flip	8.17	0.582	57.14	0.48	0.565	26.90	0.55	0.458	**57.05**
DiffJPEG	14.45	0.601	27.18	0.70	0.572	16.13	0.74	0.463	88.11
MPRNet	27.86	0.571	45.69	0.55	0.514	22.61	0.59	0.451	<u>64.67</u>
DiffPure	16.23	0.497	21.52	<u>0.76</u>	0.543	14.43	0.83	0.516	690.13
Purify++	13.01	0.516	<u>20.04</u>	0.75	0.569	<u>12.03</u>	0.81	0.502	253.08
InstantPure	14.43	0.547	20.78	<u>0.76</u>	<u>0.577</u>	12.52	<u>0.84</u>	0.514	223.12
TransDiff	<u>7.93</u>	<u>0.611</u>	**11.04**	**0.79**	**0.595**	**6.16**	**0.88**	<u>0.518</u>	92.16

is optimized separately against every method. The white-box attacker knows the metric, transformation bank, sampling distribution, calibration, and score-recovery rule, but does not know the four transformations sampled for a future inference call. The primary attack estimates the expected-score gradient with expectation over transformations (EOT), drawing new four-view sets during optimization; final evaluation draws a fresh set. The nested subset also hosts NES, Parsimonious, and Square black-box attacks.

Evaluation measures. The target is the pre-attack output $y_j^* = f(x_j)$. We report D_{score} (normalized MAE in percent of the fixed metric range), R_{score} (the benchmark's logarithmic recovery measure), and SROCC with MOS; Appendix D defines them. SROCC measures ranking against human ratings. We average image-level errors within each dataset, metric, and attack, then give equal weight to attacks, metrics, and datasets; SROCC is computed within each dataset–metric–attack cell before averaging.

5.2 Comparison with existing test-time defenses

Table 1 compares the default four-view TransDiff with six test-time image defenses. The white-box results average over the attack suite on the main subset; black-box results use the nested stress subset. Each attack targets the final score returned by the corresponding method.

Recovery under attack. TransDiff has the lowest aggregate D_{score}: 11.04 under white-box attacks and 6.16 under black-box attacks. Relative to Purify++, the strongest image defense on this measure, error falls by 44.9% and 48.8%, respectively. TransDiff also achieves the highest R_{score} in both settings (0.79 and 0.88).

Clean inputs and human-quality ranking. On clean inputs, TransDiff has the least score drift among the evaluated defenses ($D_{\text{score}} = 7.93$) while nearly preserving the undefended SROCC (0.611 versus 0.624). White-box SROCC rises from 0.471 without defense to 0.595, above the strongest image defense (0.577 for InstantPure). Under black-box attacks, the undefended metric ranks images better (0.541 versus 0.518) even as its score error is much larger. This distinction matters: the target is the original metric score, whose absolute scale is not captured by SROCC.

Inference cost. TransDiff takes 92.16 ms per image, approximately $2.4\times$ faster than InstantPure and $2.7\times$ faster than Purify++. Its inference time is close to DiffJPEG (88.11 ms), while achieving lower score-recovery error under both white-box and black-box attacks.

Table 2: Comparison of score-recovery methods using the same transformation bank and shared calibration data where applicable. All results are calculated on the stress subset (200 images per dataset) with 5 reruns per attack. $D = D_{\text{score}}$, $R = R_{\text{score}}$; SROCC is measured against MOS. Bold and underline mark the best and second-best values in each column, respectively.

	Clean		White-box			Black-box		
Method / baseline	$D \downarrow$	SROCC↑	$D \downarrow$	$R \uparrow$	SROCC↑	$D \downarrow$	$R \uparrow$	SROCC↑
No defense	**0.00**	**0.621**	63.03	0.56	0.473	36.17	0.65	**0.544**
Uncorrected mean	10.02	0.530	15.83	0.61	0.537	8.16	0.72	0.505
Calibrated mean	9.87	0.547	15.87	0.61	0.538	8.10	0.73	0.507
Calibrated median	9.85	0.545	15.02	0.74	0.538	7.32	0.79	0.509
Trimmed median	9.21	0.562	14.79	0.73	0.561	7.34	0.79	0.511
Interval overlap	9.15	0.573	14.65	<u>0.76</u>	0.569	7.04	0.83	0.515
Gaussian shifts	8.11	0.589	<u>12.56</u>	0.75	<u>0.582</u>	<u>6.71</u>	<u>0.85</u>	0.516
TransDiff	<u>7.91</u>	<u>0.607</u>	**11.12**	**0.79**	**0.589**	**6.21**	**0.88**	<u>0.517</u>

5.3 What does distribution shape contribute?

Table 2 compares recovery rules on the nested stress subset at the same four-view budget, averaging five runs of each attack. Transformation-based methods use the same bank and share calibration data where applicable, while attacks are optimized separately against each method. The Gaussian-shift baseline matches TransDiff's calibration samples, weights, bandwidth, and aggregation, replacing each empirical shift distribution with its fitted mean and variance.

Aggregation and calibration. Uncorrected averaging already reduces white-box D_{score} from the undefended 63.03 to 15.83, and black-box error from 36.17 to 8.16. Combining transformed observations is therefore a strong baseline. Subtracting calibrated mean shifts changes these values only slightly, to 15.87 and 8.10. Median and interval-based rules improve further, with interval overlap reaching 14.65 and 7.04. These comparisons suggest that a fixed location correction alone captures little of the additional recovery benefit available from calibration; the rule used to combine the observations matters.

Beyond location and variance. Gaussian shifts improve on the summary-based rules, reaching $D_{\text{score}} = 12.56$ under white-box attacks and 6.71 under black-box attacks. Retaining the empirical distributions reduces these errors further to 11.12 and 6.21, gains of 11.5% and 7.5%. Corresponding R_{score} values increase from 0.75 to 0.79 and from 0.85 to 0.88. Because the Gaussian baseline matches the transformations, calibration data, bandwidth, and aggregation, this gain is consistent with useful structure beyond the fitted mean and variance. Black-box SROCC is almost unchanged (0.516 versus 0.517), so the clearest difference is in score fidelity. The empirical model also reduces clean-image score drift relative to Gaussian shifts (7.91 versus 8.11).

5.4 Robustness under increasing attack strength

Figure 2 isolates I-FGSM and varies its ℓ_∞ radius, complementing the attack-averaged tables. TransDiff uses its default four-view configuration throughout this sweep. TransDiff has the lowest plotted error at every radius on both datasets; Gaussian shifts are its closest comparator. The gap persists as ϵ grows, although TransDiff's error also increases, most visibly between $8/255$ and $16/255$. The two datasets show the same ordering despite their different distortion sources.

5.5 Effect of the number of transformed views

Figure 3 varies the number of sampled transformations from the fixed 48-entry bank. It compares TransDiff, Gaussian shifts, and the calibrated median, and includes a separate endpoint that evaluates the complete bank. The steepest improvement occurs between one and four views. Gains then taper off, while evaluating all 48 transformations still gives the lowest plotted error. Four views capture

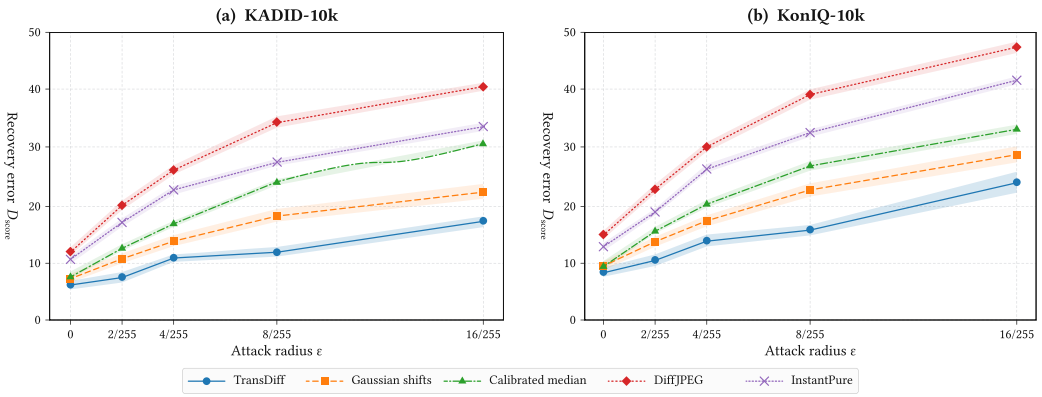

Figure 2: Score-recovery error under increasing I-FGSM attack strength on KADID-10k and KonIQ-10k. Attacks maximize each method's final defended score. TransDiff uses four sampled transformations. The $\epsilon = 0$ point measures score drift without an attack; lower D_{score} is better. Shaded bands show 95% confidence intervals across images.

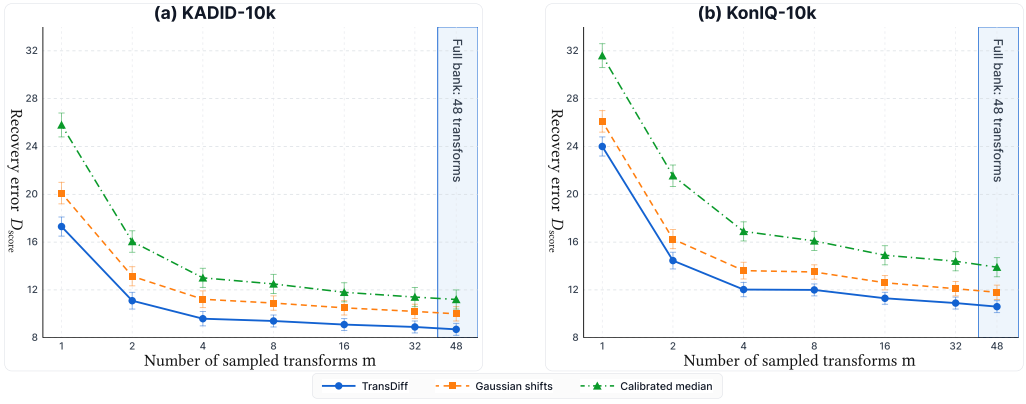

Figure 3: Score-recovery error versus the number of transformed views. Results are averaged over white-box attacks on the stress subset, with five runs per attack and a perturbation budget of $\epsilon = 8/255$. Error bars show 95% confidence intervals across images.

much of the benefit at one twelfth of the full-bank evaluation count. TransDiff remains ahead of Gaussian shifts and the calibrated median across the sweep: more observations reduce sampling variability, but their number alone cannot correct an image-specific score shift shared across views.

6 Conclusion

We introduced TransDiff, a method for recovering the pre-attack score of a frozen NR-IQA metric through calibrated agreement across transformed views. By accounting for transformation-induced score changes, it combines these observations without assuming score invariance or reconstructing the input image. Across nine metrics and two datasets, four sampled views improve score recovery over the evaluated test-time image defenses under white-box and black-box attacks. Matched comparisons show that retaining empirical distribution shape provides additional gains beyond location and variance. Our analysis characterizes when the recovery signal dominates calibration, sampling, image-specific, and attack-induced errors, while exposing a fundamental limitation: adding views cannot resolve score shifts shared across transformations. Together, these findings support calibrated distributional agreement as a practical approach to recovering original metric outputs when transformations themselves alter the quantity being measured.

AI USE STATEMENT

AI models were used for polishing the text and finding related works. They were also used for checking the correctness of the theory presented in the paper.

REPRODUCIBILITY STATEMENT

Appendix A gives the full mathematical analysis. Appendices B–D specify the recovery rule, white-box objective, and evaluation measures.

REFERENCES

Anastasia Antsiferova, Khaled Abud, Aleksandr Gushchin, Ekaterina Shumitskaya, Sergey Lavrushkin, and Dmitriy Vatolin. Comparing the robustness of modern no-reference image- and video-quality metrics to adversarial attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 700–708, 2024. doi: 10.1609/aaai.v38i2.27827.

Laurent Bako. Robustness analysis of a maximum correntropy framework for linear regression. *Automatica*, 87:218–225, 2018. doi: 10.1016/j.automatica.2017.09.006.

R. Brooks and Sundararaj Iyengar. Robust distributed computing and sensing algorithm. *Computer*, 29:53–60, 07 1996. doi: 10.1109/2.507632.

Baoliang Chen, Lingyu Zhu, Chenqi Kong, Hanwei Zhu, Shiqi Wang, and Zhu Li. No-reference image quality assessment by hallucinating pristine features. *IEEE Transactions on Image Processing*, 31:6139–6151, 2022. doi: 10.1109/TIP.2022.3205770.

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

Yuming Fang, Hanwei Zhu, Yan Zeng, Kede Ma, and Zhou Wang. Perceptual quality assessment of smartphone photography. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3674–3683, 2020. doi: 10.1109/CVPR42600.2020.00373.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012.

Aleksandr Gushchin, Khaled Abud, Georgii Bychkov, Ekaterina Shumitskaya, Anna Chistyakova, Sergey Lavrushkin, Bader Rasheed, Kirill Malyshev, Dmitriy S. Vatolin, and Anastasia Antsiferova. Guardians of image quality: Benchmarking defenses against adversarial attacks on image quality metrics. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 21444–21470. PMLR, 2025.

Vlad Hosu, Hanhe Lin, Tamás Szirányi, and Dietmar Saupe. Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on Image Processing*, 29:4041–4056, 2020. doi: 10.1109/TIP.2020.2967829.

Markus Kettunen, Erik Härkönen, and Jaakko Lehtinen. E-lpips: Robust perceptual image similarity via random transformation ensembles. *arXiv preprint arXiv:1906.03973*, 2019. URL `https://arxiv.org/abs/1906.03973`.

Jari Korhonen and Junyong You. Adversarial attacks against blind image quality assessment models. In *QoEVMA Workshop at ACM Multimedia*, pp. 3–11, 2022.

Dingquan Li, Tingting Jiang, and Ming Jiang. Norm-in-norm loss with faster convergence and better performance for image quality assessment. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 789–797, 2020. doi: 10.1145/3394171.3413804.

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

Yujia Liu, Chenxi Yang, Dingquan Li, Jianhao Ding, and Tingting Jiang. Defense against adversarial attacks on no-reference image quality models with gradient norm regularization, 2024. URL https://arxiv.org/abs/2403.11397.

Keith Marzullo. Tolerating failures of continuous-valued sensors. *ACM Transactions on Computer Systems*, 8(4):284–304, 1990. doi: 10.1145/128733.128735.

Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Anima Anandkumar. Diffusion models for adversarial purification. In *International Conference on Machine Learning (ICML)*, 2022.

Ekaterina Shumitskaya, Mikhail Pautov, Dmitriy Vatolin, and Anastasia Antsiferova. Stochastic biqa: Median randomized smoothing for certified blind image quality assessment. *Computer Vision and Image Understanding*, 259:104447, 2025a. doi: 10.1016/j.cviu.2025.104447.

Ekaterina Shumitskaya, Dmitriy Vatolin, and Anastasia Antsiferova. Fs-iqa: Certified feature smoothing for robust image quality assessment, 2025b.

Jianyi Wang, Kelvin C. K. Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 2555–2563, 2023. doi: 10.1609/aaai.v37i2.25353.

Chenxi Yang, Yujia Liu, Dingquan Li, Yan Zhong, and Tingting Jiang. Beyond score changes: Adversarial attack on no-reference image quality assessment from two perspectives. *arXiv preprint arXiv:2404.13277*, 2024.

Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1191–1200, 2022.

Zhenqiang Ying, Haoran Niu, Praful Gupta, Dhruv Mahajan, Deepti Ghadiyaram, and Alan Bovik. From patches to pictures (paq-2-piq): Mapping the perceptual space of picture quality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3575–3585, 2020. doi: 10.1109/CVPR42600.2020.00363.

Boya Zhang, Weijian Luo, and Zhihua Zhang. Purify++: Improving diffusion-purification with advanced diffusion models and control of randomness, 2023.

Hancheng Zhu, Leida Li, Jinjian Wu, Weisheng Dong, and Guangming Shi. Metaiqa: Deep meta-learning for no-reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14143–14152, 2020. doi: 10.1109/CVPR42600.2020.01415.

A PROOFS OF THE RECOVERY RESULTS

We prove Proposition 1 and Theorem 1, using the notation of Sections 3–4. Fix the clean image x, the finite bank, q, and h independently of the iid calibration images. Write $\Delta_\phi(z) = f(\phi(z)) - f(z)$. Calibration outputs may be dependent across transformations. The attack may depend on calibration but is fixed before inference draws. The output satisfies Eq. (4).

A.1 KERNEL MEANS AND THE POPULATION IDENTITY

Let $\mathcal{H}_h$ be the Gaussian reproducing-kernel Hilbert space (RKHS), $\psi_h(t) = k_h(t, \cdot)$, and $\mu_P = \mathbb{E}_{D\sim P}\psi_h(D)$. Then $\|\psi_h(t)\| = 1$, $\|\mu_P\| \le 1$, and

$$\mathrm{MMD}_{k_h}(P, Q) = \|\mu_P - \mu_Q\|_{\mathcal{H}_h}.$$

Define

$$G(d) = \sum_\phi q_\phi \mathbb{E}_{R,D\,\mathrm{iid}\sim P_\phi} k_h(R - D, d),$$
$$H(d) = G(0) - G(d).$$

Proof. Fix ϕ. The two within-law terms in the squared MMD both equal $\mathbb{E}k_h(R, D)$ by translation invariance. The cross term is $\mathbb{E}k_h(R, D+d) = \mathbb{E}k_h(R-D, d)$. Subtracting gives the stated identity after averaging over ϕ.

The Gaussian kernel is characteristic. A probability law on $\mathbb{R}$ cannot equal a nonzero translate: if $P = T_d P$, its characteristic function $\varphi(t)$ satisfies $\varphi(t) = e^{itd}\varphi(t)$; continuity and $\varphi(0) = 1$ force $e^{itd} = 1$ throughout a neighborhood of zero, hence $d = 0$. Thus each positive-weight term is strictly positive at nonzero d. Dominated convergence gives continuity of G and $G(d) \to 0$ as $|d| \to \infty$. Since $G(0) > 0$, positivity on compact sets bounded away from zero implies a positive infimum outside every radius $r > 0$. □

A.2 IMAGE MISMATCH

For Q_x, Q_0, and χ_x from Eq. (7), the same transformation is used within each shift difference. The population agreement for image x obeys

$$S_x(y^* + d) = \sum_\phi q_\phi s_\phi(\Delta_\phi(x) - d) = \langle \mu_{Q_x}, \psi_h(d)\rangle,$$

whereas $G(d) = \langle \mu_{Q_0}, \psi_h(d)\rangle$. Consequently,

$$|S_x(y^*) - S_x(y^* + d) - H(d)| \le w_h(d)\chi_x, \tag{12}$$

with

$$w_h(d) = \|\psi_h(0) - \psi_h(d)\| = \sqrt{2(1 - e^{-d^2/(2h^2)})}.$$

The shared factor $w_h(d)$ makes the bound vanish at $d = 0$; bounding the two candidate scores separately would lose this local behavior. In particular, $w_h(d) \le \min\{\sqrt{2}, |d|/h\}$.

A.3 UNIFORM CALIBRATION AND SAMPLING CONCENTRATION

For iid Hilbert-space-valued random elements $Z_1, \ldots, Z_N$ with $\|Z_j\| \le 1$ and mean μ,

$$\Pr\left\{\left\|N^{-1}\sum_j Z_j - \mu\right\| > \frac{1 + \sqrt{2\log(1/u)}}{\sqrt{N}}\right\} \le u. \tag{13}$$

The centered average has expected norm at most $1/\sqrt{N}$, and replacing one observation changes the norm by at most $2/N$. McDiarmid's inequality then gives Eq. (13).

Apply this to $\psi_h(\Delta_\phi(X_j))$ and take a union bound over K entries. With probability at least $1 - \alpha$,

$$\max_\phi \|\widehat{\mu}_\phi - \mu_{P_\phi}\| \le \eta_n, \qquad \eta_n = \frac{1 + \sqrt{2\log(K/\alpha)}}{\sqrt{n}}. \tag{14}$$

The union bound does not assume independent calibration columns.

The calibration event also covers attacks chosen after observing the calibration sample. Define the RKHS isometry $U_a\psi_h(t) = \psi_h(a-t)$. For any response a_ϕ, the candidate-centered empirical and population kernel means differ by $U_{a_\phi - y^*}(\widehat{\mu}_\phi - \mu_{P_\phi})$, whose norm is bounded by η_n independently of a_ϕ. Averaging over q preserves the bound.

Conditioned on calibration and the attacked input, define for one sampled transformation

$$Z_\phi = \frac{1}{n}\sum_j \psi_h(a_\phi - y^* - \Delta_\phi(X_j))$$

This element has norm at most one. Equation (13) gives sampling error at most

$$\varepsilon_m = \frac{1+\sqrt{2\log(1/\beta)}}{\sqrt{m}}$$

with conditional probability at least $1-\beta$. The calibration and sampling events therefore hold jointly with probability at least $1-\alpha-\beta$, uniformly over candidate scores for the fixed attacked input. The transformations must be drawn after the attack is fixed. Full-bank inference has no sampling error and uses only the calibration event.

A.4 Effect of the attack shift

Let $t_\phi = f(\phi(x_{\mathrm{adv}})) - f(\phi(x))$ and $\tau = \sum_\phi q_\phi |t_\phi|$. Suppose $|s'_\phi| \le L$ and $|s''_\phi| \le M$ uniformly. Direct differentiation of the kernel gives valid universal constants $L = e^{-1/2}/h$ and $M = 1/h^2$. Mixtures inherit these bounds.

For $C_{\phi,d}(u) = s_\phi(u) - s_\phi(u-d)$, the fundamental theorem of calculus gives

$$|C'_{\phi,d}(u)| = |s'_\phi(u) - s'_\phi(u-d)| \le \min\{2L, M|d|\}.$$

It follows that

$$|C_{\phi,d}(u+t_\phi) - C_{\phi,d}(u)| \le |t_\phi| \min\{2L, M|d|\}. \tag{15}$$

Averaging proves the attack-shift term in the main theorem.

A.5 Proof of the recovery theorem

On the concentration events, let $B = \chi_x + \eta_n + \varepsilon_m$, with $\varepsilon_m = 0$ for full-bank inference. The image mismatch, calibration error, and sampling error form a signed RKHS element with norm at most B. Their effect on the agreement difference between y^* and $y^* + d$ is at most $w_h(d)B$. Equation (15) bounds the attack-shift term. Therefore, simultaneously for every $d \in \mathbb{R}$,

$$\begin{aligned} H(d) \le \widehat{S}(y^*) - \widehat{S}(y^*+d) \\ + w_h(d)B + \tau \min\{2L, M|d|\}. \end{aligned}$$

Substituting $d = \widehat{y} - y^*$ and using $\widehat{S}(y^*) - \widehat{S}(\widehat{y}) \le \rho$ proves Eq. (10). If the left-hand margin exceeds ρ for every $|d| \ge r$, then the output error lies inside $(-r, r)$.

Local scaling. If the global condition has localized $\widehat{d}$ to a region where $H(d) \ge ad^2$ for some $a > 0$, then $w_h(d) \le |d|/h$ yields, for exact optimization ($\rho = 0$),

$$|\widehat{d}| \le \frac{(\chi_x + \eta_n + \varepsilon_m)/h + M\tau}{a}.$$

Thus calibration and sampling errors decrease with n and m, while image mismatch and the attack shift remain.

B SCORE-RECOVERY IMPLEMENTATION

B.1 CALIBRATION AND TRANSFORMATION CONFIGURATION

For each frozen metric, retain the score shifts from Eq. (1) on the PaQ-2-PiQ calibration split. Choose the bank and common bandwidth h on the disjoint development split; h is expressed in metric-score units. Each bank entry fixes all transformation parameters and internal seeds. Table 3 lists the bank; evaluation subsets and repeated-run aggregation are specified in Section 5.

Table 3: The fixed differentiable transformation bank used for full-bank inference (48 entries). Spatial and frequency parameters are given at the model input resolution.

Family (entries)	Fixed variants
Flip (1)	Horizontal.
Rotation (6)	$\pm1^\circ, \pm3^\circ, \pm5^\circ$.
Translation (8)	$(\pm2, 0)$, $(0, \pm2)$ and $(\pm2, \pm2)$ pixels.
Crop + resize (6)	Central crops retaining 98% and 94%; four corner crops retaining 96%.
Downscale + upscale (4)	Scale factors $0.95, 0.90, 0.85, 0.80$.
Gaussian blur (4)	$\sigma = 0.4, 0.7, 1.0, 1.4$ pixels.
Bilateral filter (3)	$(\sigma_s, \sigma_r) = (0.8, 0.06), (1.2, 0.10), (1.6, 0.14)$.
Frequency filtering (3)	Smooth masks with relative cutoffs $0.92, 0.82, 0.72$ of Nyquist.
Smooth DCT compression (4)	AC shrinkage with $\lambda = 0.003, 0.006, 0.012, 0.024$.
Cross-channel mixing (2)	$R \leftrightarrow G$ and $G \leftrightarrow B$ with mixing coefficient 0.03.
Tone curve (4)	$T_a(x) = x + a\,x(1-x)(2x-1)$, $a = \pm0.08, \pm0.16$.
Saturation (3)	Multipliers $0.97, 0.93, 0.87$.
Total	48 transformations.

The identity is not a bank entry. Each entry has probability $q_\phi = 1/48$, so a family's total weight is proportional to its number of variants.

B.2 AGREEMENT EVALUATION AND GLOBAL SEARCH

For the active weights w_ϕ, compute $a_\phi = f(\phi(x_{\mathrm{adv}}))$ and

$$\widehat{S}(v) = \sum_\phi \frac{w_\phi}{n} \sum_j \exp\left(-\frac{(a_\phi - \Delta_{\phi,j} - v)^2}{2h^2}\right).$$

The active centers are $z_{\phi j} = a_\phi - \Delta_{\phi,j}$. The objective is a continuous Gaussian mixture whose derivative is positive below the smallest active center and negative above the largest. Hence its global maxima lie in $[z_{\min}, z_{\max}]$; for an exact tie, select the smallest maximizer.

A grid including both endpoints with maximum spacing g contains a point within $g/2$ of a global maximizer. Since $|\widehat{S}'(v)| \le L_h = e^{-1/2}/h$, choosing the best grid point guarantees $\rho \le L_h g/2$. A refinement that does not lower its agreement preserves this bound. If all centers coincide, their common value is the exact solution.

Inference modes. Default inference uses the empirical frequencies of four independent draws with replacement, including duplicate multiplicity. The view-budget ablation increases the number of draws; its full-bank endpoint uses $w = q$ over all 48 entries. Appendix C describes EOT and fresh evaluation draws.

B.3 BASELINE DEFINITIONS

Use the same active weights and transformed predictions for all transformation-based comparisons. The uncorrected mean is $\sum_\phi w_\phi a_\phi$. The corrected mean subtracts the empirical mean shift from each prediction. The corrected median is the weighted median of $a_\phi - \widehat{b}_\phi$, where $\widehat{b}_\phi$ is the empirical median shift; define a weighted median as the smallest value whose cumulative weight reaches one half. At the default $m = 4$, the trimmed-median baseline drops the observation farthest from the weighted median of calibrated predictions (ties follow bank order), then takes the weighted median of the other three. The uncalibrated variants apply the corresponding summary directly to a_ϕ.

For the moment-matched Gaussian baseline, let $\widehat{\mu}_\phi$ and $\widehat{\sigma}^2_\phi$ be the empirical shift mean and variance (variance divisor n). Replace the empirical agreement by

$$\widehat{s}^{\mathrm{G}}_\phi(t) = \frac{h}{\sqrt{h^2 + \widehat{\sigma}^2_\phi}} \exp\left(-\frac{(t - \widehat{\mu}_\phi)^2}{2(h^2 + \widehat{\sigma}^2_\phi)} \right).$$

The zero-variance case is well-defined. Use the same h, aggregation weights, and controlled global maximization as for TransDiff; the resulting Gaussian components can have different widths. This baseline retains location and variance while removing higher-order distribution shape.

For the interval baseline, let ℓ_ϕ and u_ϕ be the empirical 5th and 95th percentiles of the calibrated shift samples. Form $[a_\phi - u_\phi, a_\phi - \ell_\phi]$ and maximize weighted overlap, returning the midpoint of a maximal-overlap interval. Quantile summaries are used only in this comparison, not in TransDiff.

Existing test-time defenses. The image defenses in Table 1 follow the prior benchmark (Gushchin et al., 2025); each returns the frozen metric score of its processed image.

C Attack protocol

For every defense, the white-box attacker maximizes the final deployed score: $f(P(z))$ for an image defense P, or the returned score for a score-recovery method. The attacker knows the frozen metric and, for TransDiff, the calibrated shifts, transformation bank, sampling probabilities, bandwidth, and recovery rule. The suite contains I-FGSM, Optimised-UAP, the attacks of Korhonen & You (2022) and Yang et al. (2024), and FACPA. The primary pixel constraint is $\|\delta\|_\infty \leq 8/255$ for images in $[0, 1]$, with an I-FGSM sweep over $\{2, 4, 8, 16\}/255$; attacked images are projected to the valid image range. Each method is attacked independently.

C.1 Objective and EOT

Let $T_{\pi,\omega}(z)$ be the final defended score under method π and sampling randomness ω. The attack selects a single perturbation by maximizing

$$\max_{\|\delta\|_\infty \leq \epsilon_{\mathrm{atk}}} J_\pi(\delta), \qquad J_\pi(\delta) = \mathbb{E}_{\omega \sim q^4}[T_{\pi,\omega}(x + \delta)].$$

Here ω is a four-view sample with replacement from the public bank. For a deterministic image defense or full-bank recovery, the expectation is omitted. For sampled inference, the white-box attacker estimates this expectation from independent four-view samples and differentiates through the transformations and recovery rule. Evaluation draws fresh views after the perturbation is fixed; these draws are unknown to the attacker during optimization.

C.2 Attacking the global maximum

The aggregate agreement is smooth in the candidate score, but selection of its global maximum can switch branches. At an isolated stationary maximum with nonzero curvature and differentiable transformations, a local implicit derivative is

$$\nabla_z T(z) = -\frac{\nabla_z \partial_v \widehat{S}_z(v)|_{v=T(z)}}{\partial_v^2 \widehat{S}_z(T(z))}.$$

This is a local derivative of a continuous stationary branch. At a branch switch, the attack uses the final returned score to select its iterate.

Attack budget. We use 10 projected I-FGSM iterations with step size $\epsilon_{\mathrm{atk}}/10$ and, at each step, average gradients over eight independently sampled four-view sets; the iteration count follows an existing NR-IQA defense evaluation (Shumitskaya et al., 2025b). The same sampled-inference EOT principle applies to the other iterative white-box attacks; their published attack-specific optimization rules are retained. Full-bank attacks differentiate through the deterministic weighted bank. Select perturbations by the final defended score.

D EVALUATION MEASURES

For N evaluation images, let $y_j^* = f(x_j)$ and let $\widehat{y}_j$ denote the defended score. We follow the score-recovery measures of the prior NR-IQA benchmark:

$$D_{\text{score}} = \frac{100}{N} \sum_{j=1}^{N} \frac{|\widehat{y}_j - y_j^*|}{\text{diam}(f)},$$

$$R_{\text{score}} = \frac{1}{N} \sum_{j=1}^{N} \log \frac{\max\{\beta_1 - y_j^*,\, y_j^* - \beta_2\}}{|\widehat{y}_j - y_j^*|},$$

The range $\text{diam}(f)$ is the fixed upper-minus-lower bound reported by Gushchin et al. (2025, Appendix A.3.2, Table 6), not the observed spread in KADID, KonIQ, or the PaQ-2-PiQ image dataset. That benchmark obtained its bounds from DIV2K validation images, JPEG quality-10 versions, and random noise. Its R_{score} convention uses β_1 and β_2 as the maximum and minimum pre-attack metric scores in the declared dataset; use the same convention within each declared evaluation subset. Report SROCC between recovered scores and dataset MOS separately for clean and attacked inputs.

Table 4: Fixed lower and upper NR-IQA metric score bounds from Gushchin et al. (2025, Table 6). For each metric, $\text{diam}(f)$ is the difference between these bounds. These values are used for D_{score}.

Metric	Lower	Upper
Meta-IQA	0.00	1.00
MANIQA	0.00	1.00
CLIP-IQA+	0.00	1.00
TOPIQ	0.22	0.82
KonCept	26.40	66.87
SPAQ	21.75	77.75
PAQ2PIQ metric	58.38	84.17
Linearity	25.78	83.23
FPR	47.22	77.05

The cost column in Table 1 reports online inference per image, including the transformed metric calls and score recovery.