



Flip-Robust Neural Image Assessment (FR-NIMA) for Spatially Consistent IQA

Rawesak Tanawongsuwan¹ and Sukanya Phongsuphap²

ABSTRACT

Neural Image Assessment (NIMA) has become a widely adopted approach for blind image quality assessment (BIQA), yet it remains sensitive to simple spatial transformations such as horizontal flips. Such variation can lead to inconsistent predictions, even when the perceived visual content remains largely unchanged. To address this issue, we introduce Flip-Robust Neural Image Assessment (FR-NIMA), an enhanced training strategy that enhances the spatial robustness of BIQA models. Instead of modifying network architectures, FR-NIMA incorporates a flip-consistency regularization term that penalizes discrepancies between the predicted quality distributions of an image and its horizontally flipped counterpart. Two variants—one-branch and two-branch formulations—are explored, both introducing no additional model parameters. FR-NIMA is evaluated across four CNN backbones (MobileNetV2, VGG19, Xception, InceptionV3) and one Vision Transformer (ViT-Small) using the LIVE dataset and two additional test sets representing distinct scene types. Performance is assessed using complementary metrics, including the Test Loss (EMD²), Absolute Flip Gap (|FlipGap|), Flip-Consistency Win Rate (FCWR), Average Flip-Gap Delta (AFGD), and Average Flip-Gap Ratio (AFGR). Experimental results demonstrate that FR-NIMA effectively reduces flip-gap magnitude and variability while maintaining comparable test accuracy across all backbones. These findings establish FR-NIMA as a simple yet effective framework for enhancing the stability, spatial consistency, and trustworthiness of deep IQA models.

Article information:

Keywords: Image Quality Assessment, Neural Image Assessment, Flip-Consistency Regularization, Vision Transformer, Convolutional Neural Networks

Article history:

Received: November 14, 2025

Revised: February 17, 2026

Accepted: April 23, 2026

Published: April 25, 2026

(Online)

DOI: 10.37936/ecti-cit.2026202.264823

1. INTRODUCTION

Image Quality Assessment (IQA) plays a central role in many vision-based applications, including photo curation, where it helps automatically select aesthetically pleasing or technically superior images; camera benchmarking, where it quantifies perceptual fidelity to evaluate imaging systems and smartphone cameras; and image-enhancement pipelines, where it provides a perceptual feedback signal to guide the optimization of restoration and enhancement algorithms.

The introduction of Neural Image Assessment (NIMA) [1] marked a turning point in this domain, leveraging deep convolutional neural networks (CNNs) to predict the distribution of human opinion

scores and to achieve high correlation with subjective perception.

Following NIMA, subsequent deep-learning IQA models—such as RichIQA [2], TopIQ [3], TRIQA [4], and UniQA [5]—have further advanced distribution-based and transformer-driven approaches for both technical and aesthetic quality prediction.

Despite these advances, modern IQA models share a subtle yet important weakness—sensitivity to minor spatial transformations. In particular, a horizontal flip, while generally preserving low-level perceptual attributes such as sharpness, noise, and texture, can occasionally alter the semantic interpretation of an image—for example, when text, signage, or faces with directional lighting are present. Nevertheless, in standard BIQA benchmarks, horizontal flipping does

^{1,2}The authors are with the Faculty of Information and Communication Technology, Mahidol University, Thailand, Email: rawesak.tan@mahidol.ac.th and sukanya.pho@mahidol.ac.th

¹Corresponding author: rawesak.tan@mahidol.ac.th

not alter the ground-truth mean opinion score (MOS) annotations. Therefore, model predictions are expected to remain stable under such transformations. However, existing models such as NIMA often exhibit flip sensitivity, producing noticeably different quality estimates for mirrored images. This instability undermines the reliability and trustworthiness of IQA systems, particularly in downstream or real-time applications where stable and repeatable scores are essential. Similar concerns have been reported in image-classification research, motivating regularization strategies such as Visual Attention Consistency [6] and Attention-Guided CAM [7], which enforce prediction invariance under geometric perturbations. Yet, such flip-robustness has not been systematically investigated in the context of image-quality prediction.



Fig.1: A photograph of Koh Nang Yuan Island in Thailand and its horizontally flipped counterpart.

To illustrate this issue, Fig. 1 shows a photograph of Koh Nang Yuan Island in Thailand and its horizontally flipped version. Although both appear perceptually identical to human observers, baseline NIMA models produce noticeably different quality predictions. Table 1 reports the mean scores from three representative backbones—VGG19 [8], MobileNetV2 [9], and ViT-Small [10]—trained with the original NIMA framework. Despite unchanged perceptual content, their predicted means differ substantially, yielding absolute flip gaps of 0.66, 0.12, and 0.27, respectively. These discrepancies reveal that existing IQA models, while accurate overall, lack spatial robustness and may produce inconsistent evaluations under perceptually invariant transformations.

Table 1: Mean quality scores and flip gaps for an image and its flipped version. FR-NIMA-2 ($\lambda = 0.5$) yields more stable predictions than baseline NIMA across VGG19, MobileNetV2, and ViT-Small.

Model	Type	Mean (Original)	Mean (Flip)	FlipGap
VGG19	Original NIMA	7.43	8.09	0.66
	FR-NIMA-2	7.35	7.53	0.18
MobileNetV2	Original NIMA	7.67	7.79	0.12
	FR-NIMA-2	8.47	8.53	0.07
ViT-Small	Original NIMA	7.54	7.81	0.27
	FR-NIMA-2	7.03	6.95	0.09

To address this limitation, we propose **Flip-Robust Neural Image Assessment (FR-NIMA)**, a training strategy that augments NIMA

with a flip-consistency regularization term. This term penalizes differences between the predicted quality distributions of an image x and its horizontally flipped version $T(x)$, encouraging the model to learn spatially invariant representations. It is important to note that FR-NIMA enforces consistency at the model output level and does not constitute a direct experimental validation of human perceptual invariance.

Two complementary formulations are introduced:

- **FR-NIMA-1 (One-Branch Supervision):** the supervised EMD² loss is applied only to the original image x ; the flipped image $T(x)$ is used solely to compute a consistency loss between their predicted distributions $f(x)$ and $f(T(x))$.
- **FR-NIMA-2 (Two-Branch Supervision):** both x and $T(x)$ receive supervised EMD² losses, and their predictions are jointly constrained to agree.

Both variants share all parameters and introduce no additional learnable weights. As shown in Table 1, FR-NIMA with $\lambda = 0.5$ significantly lowers the absolute flip gaps (0.18 for VGG19, 0.07 for MobileNetV2, and 0.09 for ViT-Small) and produces more consistent scores without shifting their overall level.

We train and evaluate FR-NIMA on the LIVE In the Wild Image Quality Challenge Database [11, 12], sweeping $\lambda \in \{0.00, 0.25, 0.50, 0.75, 1.00\}$ to analyze the effect of regularization strength. To examine generalization, all models are further tested on two unseen public datasets—a scenery dataset [13] and a flowers dataset [14] of object-centric natural images.

To comprehensively evaluate flip robustness, we employ five complementary metrics—**Test Loss (EMD²)**, **Absolute Flip Gap (|FlipGap|)**, **Flip-Consistency Win Rate (FCWR)**, **Average Flip-Gap Delta (AFGD)**, and **Average Flip-Gap Ratio (AFGR)**. Together, these metrics capture both the magnitude and relative scale of robustness improvements, enabling a holistic assessment of model reliability beyond traditional accuracy measures. Full definitions of these metrics are provided in section 4.

This work makes the following contributions:

- 1) Flip-Robust Neural Image Assessment (FR-NIMA): a unified framework introducing two forms of flip-consistency regularization (one-branch and two-branch supervision) for IQA models.
- 2) Comprehensive evaluation metrics: we define and employ |FlipGap|, FCWR, AFGD, and AFGR as interpretable measures of prediction stability under flipping.
- 3) Extensive experiments across CNN and ViT backbones: results on LIVE, scenery, and flowers datasets show consistent reductions in flip sensitivity.
- 4) Empirical insight into robustness: the two-branch variant (FR-NIMA-2) yields the best

trade-off between accuracy and flip-robustness, while ViT-Small demonstrates the strongest overall consistency.

Collectively, these findings highlight FR-NIMA as an effective approach for improving the stability and spatial reliability of deep IQA systems at the prediction level.

2. RELATED WORKS

2.1 Image Quality Assessment

Image Quality Assessment (IQA) quantifies perceptual fidelity by modeling human visual judgment. According to a recent comprehensive survey [15], IQA methods are categorized by reference dependency—full-reference (FR), reduced-reference (RR), and no-reference (NR) or blind IQA (BIQA)—as well as by their intended application domain.

Early approaches relied on handcrafted statistical features such as gradient magnitude, phase congruency, or natural-scene statistics (NSS). With the advent of large-scale subjective rating collections and improved computational models, objective IQA has become a standard benchmark for evaluating image restoration, compression, and enhancement techniques.

The survey also distinguishes general-purpose IQA from task-specific IQA, reflecting a shift from universal perceptual modeling toward domain-adapted and reasoning-aware assessment. This broader perspective establishes the conceptual foundation for exploring flip-robustness as an additional dimension of IQA reliability.

2.2 Deep-Learning-Based IQA

Deep architectures replaced engineered features with learned perceptual representations. The pioneering Neural Image Assessment (NIMA) [1] predicted a ten-bin distribution of human ratings using EMD² loss, unifying aesthetic and technical quality prediction. Extensions such as RichIQA [2] and Descriptive IQA [16] incorporated richer supervision, exploiting full rating histograms or textual quality attributes.

Transformer-based architectures improved global reasoning: AHIQ [17] combined CNN and ViT branches for hybrid local-global features, and MANIQA [18] used multi-dimensional attention to handle subtle distortions in generated and restored images.

Hierarchical designs like TopIQ [3] and TRIQA [4] leveraged distortion-aware priors and contrastive triplet learning to boost cross-dataset generalization, while UniQA [5] unified aesthetic and technical IQA through vision-language pre-training.

Recent studies have advanced IQA toward reasoning and explainability. For example, VisualQuality-

R1 [19] integrates reinforcement learning with textual justifications to enhance interpretability, while DepictQA [20] leverages multi-modal large language models to generate natural-language explanations alongside predicted quality scores.

Together, these systems transform IQA from numeric regression into semantic and descriptive evaluation, improving alignment with human perception. Despite this progress, most deep IQA models emphasize accuracy and cross-dataset generalization rather than prediction stability under perceptually invariant transformations.

2.3 Task-Specific IQA

Parallel to generic BIQA, a surge of task-specific IQA tailors evaluation to domain or sensor constraints [15]. In medical imaging, MedQ-Bench [21] and FetalCLIP-CLS [22] assess diagnostic reliability using vision-language adaptation and parameter-efficient tuning. For face recognition, CR-FIQA [23] predicts relative classifiability rather than perceptual fidelity. Perceptual contexts also include Assessor360 [24] for omnidirectional/VR imagery, modeling the viewer’s browsing trajectory, and SEAGULL [25] for region-of-interest quality guided by segmentation. Restoration-oriented methods such as RADN [26] introduced deformable filters to tolerate spatial misalignment, and Re-IQA [27] used mixture-of-experts disentanglement for content and quality.

Image Intrinsic Scale Assessment [28] explores scale-consistency, identifying each image’s intrinsic perceptual scale where predicted quality stabilizes across magnifications.

Collectively, these works demonstrate IQA’s diversification—from global scalar prediction to context-aware, domain-adapted, and reasoning-driven assessment—and motivate reliability analysis across different forms of perceptual invariance.

2.4 Robustness and Consistency in Vision Models

Robustness to geometric and photometric perturbations remains fundamental for trustworthy vision models. In classification, Visual Attention Consistency [6] and Attention-Guided CAM [7] enforced attention-map invariance under flips and rotations. Consistency regularization is also common in self-supervised learning, penalizing divergence between augmented pairs.

In IQA, related efforts are sparse but growing: scale-consistency in IISA [28], spatial-alignment tolerance in RADN [26], and multi-view browsing coherence in Assessor360 [24] implicitly promote prediction stability. Meanwhile, reasoning-oriented approaches such as DepictQA [20] enhance semantic consistency by ensuring linguistic explanations align with perceptual quality judgments.

While these developments advance different forms of robustness, flip-consistency—the invariance of predicted quality between an image and its mirrored counterpart—remains largely unexplored in the IQA domain. Unlike classification, where labels remain identical under flipping, distribution-based BIQA models can exhibit asymmetric predictions due to feature-level spatial biases. Our proposed Flip-Robust Neural Image Assessment (FR-NIMA) extends this line of inquiry by introducing distribution-level geometric consistency, encouraging aligned quality predictions between x and its flipped version $T(x)$. This formulation broadens the notion of perceptual consistency—from scale and semantic reasoning to spatial symmetry—offering a complementary perspective on reliability in modern BIQA systems.

3. METHODOLOGY

3.1 Baseline: NIMA with EMD² Loss

The baseline of this work is the Neural Image Assessment (NIMA) model, which predicts a probability distribution over possible image quality scores rather than a single scalar rating.

Each NIMA model is built upon a pretrained backbone network—specifically, MobileNetV2 [9], VGG19 [8], Xception [29], InceptionV3 [30], or ViT-Small [10, 31]—followed by a regression head that outputs a 10-bin probability distribution corresponding to human quality ratings.

For a given input image x , the model produces a predicted distribution:

$$p = f(x) = [p_1, p_2, \dots, p_K] \quad (1)$$

where p_i represents the predicted probability that the image’s perceptual quality corresponds to rating i (from 1 to $K = 10$).

The ground-truth human ratings are represented as a normalized histogram:

$$q = [q_1, q_2, \dots, q_K] \quad (2)$$

where each q_i denotes the fraction of human raters assigning score i .

The baseline model is trained by minimizing the Earth Mover’s Distance squared (EMD²) between the predicted distribution p and the ground-truth distribution q :

$$\mathcal{L}_{base} = \frac{1}{K} \sum_{i=1}^K (F_p(i) - F_q(i))^2 \quad (3)$$

where

$$F_p(i) = \sum_{j=1}^i p_j, \quad F_q(i) = \sum_{j=1}^i q_j \quad (4)$$

are the cumulative distributions of the predicted and ground-truth ratings.

Intuitively, EMD² measures how much “mass” of the predicted distribution must be shifted to match the ground-truth distribution. Unlike cross-entropy or Kullback–Leibler divergence, which penalize only pointwise differences, EMD² respects the ordered nature of quality ratings—assigning higher penalties to predictions that deviate farther from the true score.

This formulation provides a smooth, perceptually meaningful supervision signal that aligns predicted and subjective quality distributions. The resulting NIMA models, trained independently for each backbone, serve as the **baselines** for evaluating the proposed flip-consistent extensions.

3.2 Flip-Robust NIMA (FR-NIMA)

Although the baseline NIMA model achieves strong correlation with human ratings, it remains sensitive to simple geometric transformations such as horizontal flipping. Ideally, an image and its mirrored version should produce nearly identical quality predictions, since flipping generally preserves perceptual content. However, baseline NIMA often yields inconsistent results, revealing a lack of spatial invariance.

To address this limitation, we propose Flip-Robust Neural Image Assessment (FR-NIMA), which augments NIMA with a flip-consistency regularization term that encourages prediction agreement between an image x and its horizontally flipped counterpart $T(x)$.

The overall training loss combines the baseline perceptual accuracy term and a flip-consistency penalty:

$$\mathcal{L}_{FR} = \mathcal{L}_{base} + \lambda \mathcal{D}(f(x), f(T(x))) \quad (5)$$

where

- $f(\cdot)$ denotes the predicted 10-bin quality distribution of the model,
- $\mathcal{D}(\cdot, \cdot)$ measures the distributional distance between the two predictions, and
- λ controls the strength of the regularization.

In this work, $\mathcal{D}(f(x), f(T(x)))$ is implemented as the mean-squared error (MSE) between the two predicted distributions:

$$\mathcal{D}(f(x), f(T(x))) = \frac{1}{10} \sum_{i=1}^{10} (f_i(x) - f_i(T(x)))^2 \quad (6)$$

where $f_i(x)$ and $f_i(T(x))$ denote the predicted probabilities of quality rating i for the original and flipped images, respectively.

Unlike the baseline EMD² loss, which models perceptual distance to ground-truth distributions, the flip-consistency term enforces prediction agreement and does not require ordinal weighting. Therefore, MSE provides a simple and effective regularization.

This formulation directly penalizes deviations between the predicted quality histograms, promoting spatially consistent outputs without introducing additional learnable parameters.

3.3 FR-NIMA Variants

We consider two complementary formulations that differ in how the supervised and consistency terms are applied.

(a) FR-NIMA-1: One-Branch Supervision

In this variant, the supervised loss is computed only on the original image x , while the flipped image $T(x)$ is used only for the consistency term:

$$\mathcal{L}_{FR-1} = \mathcal{L}_{base}(x) + \lambda \mathcal{D}(f(x), f(T(x))) \quad (7)$$

This one-branch design avoids redundant supervision of flipped samples, reducing computational cost while still improving flip consistency.

(b) FR-NIMA-2: Two-Branch Supervision

Here, both the original and flipped images are directly supervised by the baseline loss, and their predictions are constrained to align:

$$\mathcal{L}_{FR-2} = 0.5\mathcal{L}_{base}(x) + 0.5\mathcal{L}_{base}(T(x)) + \lambda \mathcal{D}(f(x), f(T(x))) \quad (8)$$

Since horizontal flipping does not alter perceptual quality, the flipped image shares the same ground-truth rating distribution as the original image. This dual-supervision approach promotes consistency in both perceptual accuracy and spatial prediction, resulting in more stable training and improved robustness to geometric transformations.

Both FR-NIMA variants share identical architectures and parameters; they differ only in how supervision is applied during training. A sweep over $\lambda \in [0, 1.0]$ is performed to examine the trade-off between perceptual accuracy and flip-robust consistency.

The proposed framework does not introduce architectural modifications; both original and flipped images are processed by the same backbone and regression head with shared parameters. The distinction between FR-NIMA-1 and FR-NIMA-2 lies solely in how the supervised loss is applied during training. Since the model structure remains identical to baseline NIMA and the optimization procedure is fully specified in Eqs. (5)–(8), the method can be directly implemented without structural changes.

FR-NIMA-1 introduces flip-consistency as a lightweight regularization applied only to the original image, making it computationally efficient. FR-NIMA-2 additionally supervises both original and flipped inputs, enforcing stronger invariance at the cost of increased training computation. Thus, FR-NIMA-1 is suitable for efficiency-sensitive settings,

while FR-NIMA-2 is preferable when maximal stability is desired.

3.4 Backbone Selection Rationale

The five pretrained backbones used in the baseline NIMA models—MobileNetV2, VGG19, Xception, InceptionV3, and ViT-Small—are retained for all FR-NIMA experiments. This choice ensures that every flip-consistent variant is directly comparable to its corresponding baseline and that performance differences can be attributed solely to the proposed training strategy rather than architectural changes.

These backbones were selected to provide both architectural diversity and continuity with prior work, spanning conventional convolutional networks and a transformer-based design.

3.4.1 CNN Backbones

The four convolutional networks (VGG19, MobileNetV2, Xception, and InceptionV3) were selected to represent different families of CNN design principles:

- VGG19 embodies a classic deep convolutional hierarchy with large receptive fields.
- MobileNetV2 offers a compact and efficient architecture built on depthwise separable convolutions.
- Xception and InceptionV3 incorporate multi-branch and multi-scale feature extraction, capturing spatial information at different granularities.

Several of these networks (VGG19, MobileNetV2, and InceptionV3) were also used in the original NIMA (2018) framework, ensuring direct comparability with earlier IQA baselines.

In addition, their trust characteristics were previously examined in a study on the evaluation of the trustworthiness of CNNs on transfer learning [32], where human evaluators perceived varying levels of reliability among CNN architectures based on heatmap explanations. By including both high-trust and low-trust CNNs, this work investigates whether similar trust differences appear in the context of image-quality assessment.

3.4.2 Vision Transformer Backbone

For transformer-based evaluation, we adopt ViT-Small as a representative attention-based architecture. MATLAB provides several pretrained Vision Transformer (ViT) variants on ImageNet [31], including ViT-Tiny (5.7M parameters), ViT-Small (22.1M), and ViT-Base (86.8M). Among these, ViT-Small offers the best trade-off between capacity, computational efficiency, and training stability. While ViT-Tiny tends to underfit IQA data and ViT-Base requires significantly more computational resources, ViT-Small provides stable convergence and sufficient representational power for perceptual quality prediction. Its inclusion complements the CNN backbones,

enabling analysis of how attention-based representations compare with convolutional inductive biases under flip-consistency regularization.

4. EXPERIMENTAL SETUP

All experiments were performed using the LIVE dataset [11], which served as the primary benchmark. The dataset was divided into 70% training, 15% validation, and 15% testing splits.

To assess generalization, the trained models were further evaluated on two additional datasets obtained from a public dataset repository: a scenery dataset [13] containing diverse outdoor images and a flowers dataset [14] comprising object-centric natural scenes. For both datasets, only the test splits provided by the original sources were used in our experiments.

4.1 Training Configuration

For each backbone architecture, we trained one baseline NIMA model using the standard EMD² loss, five FR-NIMA-1 (one-branch) models with $\lambda \in \{0.00, 0.25, 0.50, 0.75, 1.00\}$, and five FR-NIMA-2 (two-branch) models using the same λ sweep. In total, 55 models were trained (5 baselines, 25 one-branch, and 25 two-branch). All training was conducted on NVIDIA DGX-A100 and H100 servers.

All models were optimized using Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 10^{-8}$) with a base learning rate of 1×10^{-4} . In a small number of cases where convergence was unstable, the learning rate was reduced to 1×10^{-5} to ensure stable training. The mini-batch size was set to 32 for CNN backbones (MobileNetV2, VGG19, Xception, InceptionV3) and 16 for Vision Transformer backbones. Input resolution followed the requirement of each pretrained architecture. During preprocessing, images were resized to $1.1 \times$ the target input size and randomly cropped to the final resolution. No random horizontal flipping was used in baseline augmentation to avoid interfering with the proposed flip-consistency formulation.

During FR-NIMA training, paired samples $(x, T(x))$ were generated dynamically via horizontal flipping within the training loop. A linear warm-up schedule was applied to λ during the first five epochs.

Early stopping was applied using a shared validation EMD² threshold of 0.13, requiring three consecutive epochs below the threshold after a minimum of 10 epochs. Preliminary experiments indicated that lower thresholds resulted in inconsistent convergence behavior across different backbones and λ settings. The selected threshold provided a stable and comparable stopping criterion across all models, ensuring fair evaluation under identical training conditions.

4.2 Evaluation Metrics

We evaluate with five complementary metrics. Let the test set contain N images $\{x_i\}_{i=1}^N$. For any image x_i , let $T(x_i)$ be its horizontal flip. The model's predicted mean opinion score (MOS) for an image z (the expected value of its predicted rating distribution) is denoted $\mu(z)$.

- 1) **Test Loss (EMD²):** Measures alignment between predicted and ground-truth rating distributions (Eq. (3)).
- 2) **Absolute Flip Gap (|FlipGap|):** Measures the sensitivity of each image to a horizontal flip:

$$g(i) = |\text{FlipGap}(i)| = |\mu(x_i) - \mu(T(x_i))| \quad (9)$$

where $\mu(x_i)$ and $\mu(T(x_i))$ are the predicted mean opinion scores (MOS) of the original and flipped versions of image x_i . We report the dataset mean and standard deviation of $g(i)$ over $i = 1, \dots, N$.

- 3) **Flip-Consistency Win Rate (FCWR):** Indicates the percentage of test images for which the flip-consistent (FR) model achieves a smaller flip gap than the baseline.

Let $g_B(i)$ and $g_F(i)$ denote the absolute flip gaps of the baseline and FR models, respectively. Define an improvement indicator $I(i) = 1$ if $g_F(i) < g_B(i)$ (otherwise $I(i) = 0$). Then

$$\text{FCWR} = \frac{1}{N} \sum_{i=1}^N I(i) \times 100\% \quad (10)$$

where N is the number of test images. A higher FCWR implies more frequent robustness improvements.

- 4) **Average Flip-Gap Delta (AFGD):** Measures the average absolute reduction in flip gaps achieved by the FR model:

$$\text{AFGD} = \frac{1}{N} \sum_{i=1}^N (g_B(i) - g_F(i)) \quad (11)$$

Positive AFGD indicates that the FR model reduces flip sensitivity on average, while negative values imply degradation.

- 5) **Average Flip-Gap Ratio (AFGR):** Captures the relative improvement factor of flip robustness. With a small stabilizer $\varepsilon = 10^{-6}$ to avoid division by zero, define

$$r(i) = \frac{g_F(i) + \varepsilon}{g_B(i) + \varepsilon} \quad (12)$$

The geometric mean of these ratios summarizes the global effect:

$$\text{AFGR} = \exp \left(\frac{1}{N} \sum_{i=1}^N \ln r(i) \right) \quad (13)$$

Interpretation:

- $\text{AFGR} = 1$ implies no change
- $\text{AFGR} < 1$ indicates improvement (the FR model produces smaller gaps on average)
- $\text{AFGR} > 1$ indicates degradation

The geometric mean is used because it is scale-invariant and less influenced by extreme values.

Together, these metrics provide a holistic assessment of both perceptual accuracy and spatial robustness: **EMD²** quantifies predictive alignment with human ratings; **|FlipGap|** measures flip sensitivity; **FCWR** and **AFGD** describe the frequency and magnitude of improvements; and **AFGR** reflects the overall multiplicative gain in flip consistency.

5. RESULTS

5.1 Performance on the LIVE Test Dataset

Table 3 summarizes the quantitative results of all 30 trained models—five backbones evaluated under the baseline NIMA and the two proposed flip-consistent variants (FR-NIMA-1 and FR-NIMA-2)—on the LIVE test split. Five complementary metrics were used: Test Loss (EMD²), mean and standard deviation of the absolute flip gap (**|FlipGap|**), Flip-Consistency Win Rate (FCWR), Average Flip-Gap Delta (AFGD), and Average Flip-Gap Ratio (AFGR).

5.1.1 Test Loss (EMD²)

Among the baseline models, MobileNetV2 (0.11), Xception (0.11), InceptionV3 (0.11), and ViT-Small (0.12) exhibited lower test losses (EMD²) than VGG19 (0.14). These results are consistent with a study on CNN trustworthiness [32], which found that CNN models with more stable explanatory heatmaps—including MobileNetV2, Xception, and InceptionV3—were also perceived as more reliable by human evaluators.

Introducing flip-consistency regularization influenced performance differently across architectures:

- For the one-branch FR-NIMA-1 models, improvements in test loss were limited; only the VGG19 backbone showed noticeable gains in predictive accuracy (from 0.14 to 0.12-0.13), whereas other CNNs experienced minor degradations.
- In contrast, the two-branch FR-NIMA-2 models generally maintained or improved accuracy, especially for MobileNetV2 (from 0.11 to 0.10-0.11), Xception (from 0.11 to 0.09-0.10), and InceptionV3 (from 0.11 to 0.09-0.10), indicating that dual supervision can stabilize learning without harming performance.
- For the ViT-Small model, the test loss remained stable across λ values (0.12-0.13), suggesting that transformer-based IQA models are less sensitive to the choice of flip-consistency strength.

5.1.2 Absolute Flip Gap (**|FlipGap|**)

For the baseline models, the mean flip gaps were smallest for Xception (0.17), InceptionV3 (0.25), and MobileNetV2 (0.20), and largest for VGG19 (0.45). The ViT-Small baseline showed comparable robustness, with a flip gap of 0.23—similar to the most stable CNNs.

ViT-Small consistently exhibits smaller flip gaps than most CNN backbones. A possible explanation lies in architectural inductive biases. CNNs rely on local convolutional filters that are not reflection-equivariant, so horizontal reversal may alter orientation-sensitive activations. In contrast, Vision Transformers use global self-attention over patch tokens, emphasizing relational content rather than strictly local spatial patterns. Since flipping preserves global distortion characteristics, attention-based representations may remain more stable. While this interpretation is hypothesis-driven, it aligns with the comparatively weaker spatial inductive bias of transformer architectures.

Applying flip-consistency regularization markedly improved robustness across all architectures.

- Both FR-NIMA-1 and FR-NIMA-2 substantially reduced mean flip gaps, confirming that enforcing spatial consistency yields more stable quality predictions.
- The magnitude of improvement varied with the regularization strength (λ). For instance, under FR-NIMA-2, MobileNetV2 achieved its lowest flip gap (0.17) at $\lambda = 1.00$, while ViT-Small reached the lowest value (0.06) at $\lambda = 0.75$.
- The performance difference between the one-branch and two-branch variants was small for CNNs but more noticeable for ViT-Small, where two-branch supervision consistently achieved the strongest robustness (mean **|FlipGap|** = 0.06 at $\lambda = 0.75$).

5.1.3 FCWR, AFGD, and AFGR

Considering all three metrics—FCWR, AFGD, and AFGR—the results indicate that VGG19 shows the largest relative improvement, while ViT-Small achieves the strongest overall performance. Consistent with the reduction in flip gaps, both FCWR and AFGD increased substantially, indicating frequent and sizable per-image improvements over the baseline, whereas AFGR decreased, reflecting proportional gains in flip robustness.

Under FR-NIMA-1, VGG19 reached its peak improvement (FCWR = 76.14%, AFGD = 0.28, AFGR = 0.37) at $\lambda = 1.00$, while ViT-Small achieved comparable robustness (FCWR = 75.57%, AFGD = 0.14, AFGR = 0.39) at $\lambda = 0.75$.

Under FR-NIMA-2, VGG19 again performed strongly (FCWR = 73.30%, AFGD = 0.29, AFGR = 0.35) at $\lambda = 0.50$, and ViT-Small achieved the highest stability overall (FCWR = 81.82%, AFGD = 0.17,

AFGR = 0.27) at $\lambda = 0.75$.

Across all backbones, the average AFGR values were typically below 1, the reductions in flip gap (AFGD) were consistently positive, and the win rates (FCWR) remained more than 50%—confirming that flip-consistency training enhances both the frequency and magnitude of robustness improvements without compromising accuracy.

5.1.4 Additional Analysis: Comparison with Standard Flip Augmentation and Test-Time Flip Averaging

To better isolate the contribution of the proposed flip-consistency regularization, we conducted an additional controlled experiment comparing FR-NIMA with two commonly used strategies: (1) training baseline NIMA models with standard random horizontal flip augmentation, and (2) applying test-time flip averaging (TTFA) without consistency loss. For each backbone, two baseline models were trained—one without flip augmentation (the original baseline used throughout this study) and one with standard random horizontal flip augmentation. In addition, test-time flip averaging was applied by averaging predictions of the original and flipped images during inference.

Table 2: Evaluation of standard flip augmentation and test-time flip averaging (TTFA) compared to the original baseline NIMA models on the LIVE test dataset.

Model	Augmentation	EMD ²	FlipGap		EMD ² TTFA
			Mean	Std	
MobileNetV2	NoFlipAug	0.11	0.20	0.19	0.11
	FlipAug	0.11	0.23	0.18	0.11
VGG19	NoFlipAug	0.14	0.45	0.40	0.13
	FlipAug	0.13	0.29	0.30	0.13
Xception	NoFlipAug	0.11	0.17	0.14	0.11
	FlipAug	0.10	0.15	0.12	0.10
InceptionV3	NoFlipAug	0.11	0.25	0.19	0.11
	FlipAug	0.10	0.20	0.19	0.10
ViT-Small	NoFlipAug	0.12	0.23	0.25	0.12
	FlipAug	0.11	0.12	0.14	0.11

The additional experiment shown in Table 2 reveals that standard random horizontal flip augmentation already improves flip stability for several backbones. For example, VGG19 shows a substantial reduction in mean |FlipGap| from 0.45 to 0.29, and ViT-Small decreases from 0.23 to 0.12. Xception improves from 0.17 to 0.15, and InceptionV3 from 0.25 to 0.20. However, MobileNetV2 does not benefit from flip augmentation, with mean |FlipGap| slightly increasing from 0.20 to 0.23.

While flip augmentation generally reduces the mean flip gap, the magnitude of improvement varies across architectures and does not guarantee consistent stability gains. Moreover, the reduction achieved by augmentation alone remains smaller than that obtained by FR-NIMA at moderate λ values. For instance, under FR-NIMA-2, ViT-Small achieves mean

|FlipGap| values as low as 0.06, and VGG19 reaches 0.16 at $\lambda = 0.50$ —both significantly lower than their flip-augmented baselines.

Test-time flip averaging (TTFA) maintains similar EMD² values to the baseline models (e.g., MobileNetV2: 0.11; VGG19: 0.13), but it reduces flip discrepancy only by averaging outputs at inference time. Since TTFA does not modify the learned representation or training objective, it enforces symmetry externally rather than promoting intrinsic consistency.

These results indicate that while flip augmentation can partially reduce flip sensitivity, explicit flip-consistency regularization in FR-NIMA yields stronger and more systematic robustness improvements across architectures.

Based on these observations, all baseline comparisons in Table 3 – Table 5 use models trained without random horizontal flip augmentation. This ensures that improvements attributed to FR-NIMA arise from explicit consistency regularization rather than implicit augmentation effects, providing a clearer evaluation of the proposed method.

5.2 Generalization to External Datasets

To assess generalization, the trained models were evaluated on two external test datasets—scenery and flowers—using the same λ sweep as in the LIVE experiments. Each dataset contains its own predefined test split obtained from a public repository. Table 4 and Table 5 summarize the results, showing that both FR-NIMA variants improve flip robustness across diverse visual domains.

It is important to note that the Scenery and Flowers datasets do not contain MOS annotations. Therefore, they are used exclusively to evaluate prediction stability under horizontal flipping rather than supervised perceptual accuracy.

5.2.1 Scenery dataset

For the scenery dataset (Table 4), both FR-NIMA-1 and FR-NIMA-2 outperform their respective baselines across all architectures.

For MobileNetV2, the baseline mean |FlipGap| is 0.21. With FR-NIMA-2 at $\lambda = 0.25$, it improves to 0.09, accompanied by FCWR = 76.67%, AFGD = 0.12, and AFGR = 0.39.

For ViT-Small, the baseline mean |FlipGap| is 0.16, while FR-NIMA-2 at $\lambda = 1.00$ reduces it to 0.04, with FCWR = 86.67%, AFGD = 0.12, and AFGR = 0.19.

Overall, both FR-NIMA variants demonstrate strong generalization to unseen natural scenes, with FR-NIMA-2 providing slightly higher robustness across most backbones.

5.2.2 Flowers dataset

For the flowers dataset (Table 5), both FR-NIMA-1 and FR-NIMA-2 also outperform the baselines.

For VGG19, the baseline mean $|\text{FlipGap}|$ is 0.57, while FR-NIMA-2 at $\lambda = 1.00$ reduces it substantially to 0.12, with FCWR = 78.57%, AFGD = 0.45, and AFGR = 0.33.

For ViT-Small, the baseline mean $|\text{FlipGap}|$ is 0.19, and FR-NIMA-2 at $\lambda = 0.25$ lowers it to 0.04, with FCWR = 78.57%, AFGD = 0.15, and AFGR = 0.31.

The performance difference between FR-NIMA-1 and FR-NIMA-2 remains modest, suggesting that both formulations effectively regularize predictions even in domains with strong object structure and spatial symmetry.

5.3 Qualitative Illustration of Flip Robustness

Fig. 2 and Fig. 3 show qualitative examples from the scenery and flowers datasets using the ViT-Small backbone. Each figure presents one “win” case (largest reduction in flip gap) and one “loss” case (largest increase).

Since these datasets are outside the LIVE training distribution and do not provide MOS supervision, minor shifts in absolute predicted score levels between baseline and FR-NIMA models are expected. Such variations reflect cross-dataset generalization effects rather than systematic bias introduced by the proposed method.

In win cases, FR-NIMA predictions for the original and flipped images are nearly identical, indicating strong invariance to geometric reversal.

In loss cases, moderate differences remain, reflecting situations where spatial asymmetry or fine texture details still challenge prediction stability.

These examples do not aim to show perceptual correctness but rather illustrate how flip-consistency training affects model behavior. Together with the quantitative results (Table 3 - Table 5), they demonstrate how FR-NIMA stabilizes predictions across different content types.

It should be emphasized that this qualitative analysis focuses on prediction stability under spatial transformation rather than perceptual ranking performance. Correlation-based IQA metrics such as SROCC and PLCC are not explicitly analyzed in this study, as the primary objective is robustness enhancement while maintaining comparable predictive accuracy on the LIVE benchmark.

6. SUMMARY AND CONCLUSION

This study presented Flip-Robust Neural Image Assessment (FR-NIMA), an enhanced variant of NIMA that introduces flip-consistency regularization

to improve prediction stability under horizontal reversal. Two formulations—FR-NIMA-1 (one-branch) and FR-NIMA-2 (two-branch)—were explored, both sharing identical architectures without additional model parameters.

Comprehensive experiments across five backbones (four CNNs and one Vision Transformer) and three datasets demonstrated that FR-NIMA effectively improves spatial robustness while maintaining predictive accuracy. On the LIVE benchmark, both variants consistently reduced mean $|\text{FlipGap}|$, increased FCWR and AFGD values, and lowered the AFGR ratio, confirming that enforcing spatial consistency leads to more stable quality predictions. On the scenery and flowers datasets, similar patterns were observed, indicating stable prediction behavior across different image domains. Additional comparisons with standard flip augmentation and test-time flip averaging further clarify the role of explicit flip-consistency regularization in improving prediction stability.

Across all architectures, VGG19 exhibited the largest relative improvements, while ViT-Small achieved the highest overall stability.

While horizontal flipping may alter semantic interpretation in certain cases (e.g., images containing readable text, directional symbols, or asymmetric lighting), standard NR-IQA benchmarks are designed to evaluate perceptual distortions such as blur, noise, and compression artifacts rather than orientation-dependent semantics. MOS annotations are assigned per image without explicitly modeling geometric reversal. Accordingly, the proposed flip-consistency regularization reflects a modeling assumption that distortion-based quality should remain stable under horizontal flipping. In scenarios where directional meaning contributes to perceived quality, strict invariance may not fully align with human interpretation. In such cases, the consistency weight λ can be reduced or disabled depending on application requirements.

While LIVE In the Wild serves as the primary labeled dataset for supervised evaluation, extending validation to additional large-scale NR-IQA benchmarks would further strengthen cross-dataset assessment. Future work will include evaluation on other standard datasets to examine robustness beyond the current training domain and to further investigate the relationship between flip-consistency robustness and perceptual ranking performance using standard correlation-based IQA metrics such as SROCC and PLCC.

In conclusion, FR-NIMA enhances the spatial reliability and prediction robustness of blind IQA models by explicitly enforcing geometric consistency at the model output level. It extends the reliability dimension of IQA beyond predictive accuracy, providing a simple and generalizable framework for developing

more stable, interpretable, and robust image-quality prediction systems.

ACKNOWLEDGEMENT

In this work, training CNN and ViT models on DGX-A100 and H100 servers was supported by Mahidol University and the Office of the Ministry of Higher Education, Science, Research, and Innovation under the Reinventing University project: the Center of Excellence in AI-Based Medical Diagnosis (AI-MD) sub-project.

AUTHOR CONTRIBUTIONS

Conceptualization, R.T. and S.P.; methodology, R.T.; software, R.T.; validation, R.T. and S.P.; formal analysis, R.T. and S.P.; investigation, R.T. and S.P.; data curation, R.T.; writing—original draft preparation, R.T.; writing—review and editing, R.T. and S.P.; visualization, R.T.; supervision, R.T.; All authors have read and agreed to the published version of the manuscript.

References

- [1] H. Talebi and P. Milanfar, "NIMA: Neural Image Assessment," in *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 3998-4011, Aug. 2018.
- [2] X. Min *et al.*, "Exploring Rich Subjective Quality Information for Image Quality Assessment in the Wild," in *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 35, no. 8, pp. 7778-7791, Aug. 2025.
- [3] C. Chen *et al.*, "TOPIQ: A Top-Down Approach From Semantics to Distortions for Image Quality Assessment," in *IEEE Transactions on Image Processing*, vol. 33, pp. 2404-2418, 2024.
- [4] R. Suredi, S. Zadtootaghaj, N. Barman and A. C. Bovik, "TRIQA: Image Quality Assessment by Contrastive Pretraining on Ordered Distortion Triplets," *arXiv preprint arXiv:2507.12687*, 2025.
- [5] H. Zhou *et al.*, "UniQA: Unified Vision-Language Pre-training for Image Quality and Aesthetic Assessment," *arXiv preprint arXiv:2406.01069*, 2024.
- [6] H. Guo, K. Zheng, X. Fan, H. Yu and S. Wang, "Visual Attention Consistency Under Image Transforms for Multi-Label Image Classification," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, pp. 729-739, 2019.
- [7] S. Leem and H. Seo, "Attention Guided CAM: Visual Explanations of Vision Transformer Guided by Self-Attention," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, pp. 2956-2964, 2024.
- [8] S. Liu and W. Deng, "Very deep convolutional neural network based image classification using small training sample size," *3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, pp. 730-734, 2015.
- [9] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov and L. C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4510-4520, 2018.
- [10] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *arXiv preprint arXiv:2010.11929*, 2020.
- [11] LIVE In the Wild Image Quality Challenge Database, [Online] Available: <https://live.ece.utexas.edu/research/ChallengeDB/index.html>
- [12] D. Ghadiyaram and A. C. Bovik, "Massive Online Crowdsourced Study of Subjective and Objective Picture Quality," in *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 372-387, Jan. 2016.
- [13] Scenery dataset, [Online] Available: <https://universe.roboflow.com/shivv-thtaj/out-painting-using-gans-2/dataset/1>
- [14] Flowers dataset, [Online] Available: <https://universe.roboflow.com/lab-wtcwr/flowers-7rv11/dataset/1>
- [15] C. Ma, Z. Shi, Z. Lu, S. Xie, F. Chao and Y. Sui, "A Survey on Image Quality Assessment: Insights, Analysis, and Future Outlook," *arXiv preprint arXiv:2502.08540*, 2025.
- [16] Z. You *et al.*, "Descriptive Image Quality Assessment in the Wild," *arXiv preprint arXiv:2405.18842*, 2024.
- [17] S. Lao *et al.*, "Attentions Help CNNs See Better: Attention-based Hybrid Image Quality Assessment Network," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, New Orleans, LA, USA, pp. 1139-1148, 2022.
- [18] S. Yang *et al.*, "MANIQA: Multi-dimension Attention Network for No-Reference Image Quality Assessment," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, New Orleans, LA, USA, pp. 1190-1199, 2022.
- [19] T. Wu, J. Zou, J. Liang, L. Zhang and K. Ma, "VisualQuality-R1: Reasoning-Induced Image Quality Assessment via Reinforcement Learning to Rank," *arXiv preprint arXiv:2505.14460*, 2025.
- [20] Z. You, Z. Li, J. Gu, Z. Yin, T. Xue and C. Dong, "Depicting Beyond Scores: Advancing Image Quality Assessment Through Multi-modal Language Models," *European Conference on Computer Vision (ECCV)*, pp. 259-276, 2024.
- [21] J. Liu *et al.*, "MedQ-Bench: Evaluating

- and Exploring Medical Image Quality Assessment Abilities in MLLMs,” *arXiv preprint arXiv:2510.01691*, 2025.
- [22] D. He, H. Wang and M. Yaquub, “Advancing Fetal Ultrasound Image Quality Assessment in Low-Resource Settings,” *arXiv preprint arXiv:2507.22802*, 2025.
- [23] F. Boutros, M. Fang, M. Klemm, B. Fu and N. Damer, “CR-FIQA: Face Image Quality Assessment by Learning Sample Relative Classifiability,” *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, BC, Canada, pp. 5836-5845, 2023.
- [24] T. Wu *et al.*, “Assessor360: Multi-Sequence Network for Blind Omnidirectional Image Quality Assessment,” *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pp. 64957-64970, 2023.
- [25] Z. Chen *et al.*, “SEAGULL: No-reference Image Quality Assessment for Regions of Interest via Vision-Language Instruction Tuning,” *arXiv preprint arXiv:2411.10161*, 2024.
- [26] S. Shi *et al.*, “Region-Adaptive Deformable Network for Image Quality Assessment,” *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Nashville, TN, USA, pp. 324-333, 2021.
- [27] A. Saha, S. Mishra and A. C. Bovik, “Re-IQA: Unsupervised Learning for Image Quality Assessment in the Wild,” *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, BC, Canada, pp. 5846-5855, 2023.
- [28] V. Hosu, L. Agnolucci, D. Iso and D. Saupe, “Image Intrinsic Scale Assessment: Bridging the Gap Between Quality and Resolution,” *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.
- [29] F. Chollet, “Xception: Deep Learning with Depthwise Separable Convolutions,” *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, pp. 1800-1807, 2017.
- [30] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna, “Rethinking the Inception Architecture for Computer Vision,” *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, pp. 2818-2826, 2016.
- [31] Matlab documentation for Vision Transformer, [Online] Available: <https://www.mathworks.com/help/vision/ref/visiontransformer.html>
- [32] R. Tanawongsuwan, S. Phongsuphap and P. Mongkolwat, “Evaluating Trust in CNN Transfer Learning with Flower Image Classification via Heatmap-Based XAI,” *ECTI Transactions on Computer and Information Technology (ECTI-CIT)*, vol. 19, no. 3, pp. 392-405, 2025.

Table 3: Quantitative comparison of all 30 trained models on the LIVE In the Wild test dataset. The baseline models correspond to NIMA trained without random horizontal flip augmentation and serve as the reference for evaluating flip-consistent variants. Each backbone (MobileNetV2, VGG19, Xception, InceptionV3, and ViT-Small) is evaluated under the baseline NIMA framework and its flip-consistent extensions, FR-NIMA-1 (one-branch) and FR-NIMA-2 (two-branch). For each λ setting, the table reports test loss (EMD^2), the mean and standard deviation of the absolute flip gap, and three flip-robustness metrics—FCWR, AFGD, and AFGR. Lower EMD^2 and AFGR (< 1), together with higher FCWR and AFGD, indicate stronger robustness to horizontal flips.

Model	Type	FR-NIMA-1						FR-NIMA-2					
		EMD ²	FlipGap		FCWR	AFGD	AFGR	EMD ²	FlipGap		FCWR	AFGD	AFGR
			Mean	Std	(%)				Mean	Std	(%)		
MobileNetV2	Baseline	0.11	0.20	0.19				0.11	0.20	0.19			
	λ 0.00	0.14	0.21	0.18	44.32	-0.01	1.15	0.11	0.21	0.19	48.30	-0.01	1.08
	λ 0.25	0.14	0.18	0.15	52.84	0.02	0.92	0.1	0.19	0.16	50.57	0.01	0.91
	λ 0.50	0.14	0.15	0.12	54.55	0.05	0.86	0.11	0.18	0.17	53.41	0.02	0.92
	λ 0.75	0.14	0.14	0.10	58.52	0.06	0.82	0.11	0.18	0.15	58.52	0.03	0.94
	λ 1.00	0.14	0.15	0.11	53.98	0.05	0.83	0.11	0.17	0.15	50.57	0.03	0.94
VGG19	Baseline	0.14	0.45	0.40				0.14	0.45	0.40			
	λ 0.00	0.12	0.23	0.21	68.18	0.21	0.53	0.12	0.21	0.21	69.89	0.24	0.46
	λ 0.25	0.12	0.26	0.23	67.05	0.19	0.58	0.12	0.25	0.24	65.91	0.20	0.51
	λ 0.50	0.13	0.23	0.20	67.61	0.22	0.53	0.12	0.16	0.15	73.30	0.29	0.35
	λ 0.75	0.12	0.23	0.23	67.61	0.22	0.51	0.13	0.18	0.16	69.89	0.27	0.39
	λ 1.00	0.13	0.17	0.16	76.14	0.28	0.37	0.12	0.24	0.24	67.61	0.21	0.50
Xception	Baseline	0.11	0.17	0.14				0.11	0.17	0.14			
	λ 0.00	0.13	0.15	0.13	56.82	0.02	0.82	0.1	0.15	0.10	56.25	0.02	0.96
	λ 0.25	0.13	0.14	0.12	57.95	0.03	0.77	0.1	0.14	0.11	59.66	0.03	0.80
	λ 0.50	0.13	0.13	0.10	62.50	0.04	0.72	0.1	0.14	0.13	59.66	0.03	0.70
	λ 0.75	0.13	0.13	0.10	59.66	0.04	0.74	0.09	0.15	0.12	52.84	0.02	0.85
	λ 1.00	0.13	0.14	0.11	60.23	0.03	0.78	0.09	0.13	0.11	61.36	0.04	0.75
InceptionV3	Baseline	0.11	0.25	0.19				0.11	0.25	0.19			
	λ 0.00	0.13	0.22	0.19	56.25	0.03	0.91	0.09	0.22	0.24	59.09	0.04	0.79
	λ 0.25	0.13	0.21	0.18	63.64	0.04	0.85	0.09	0.22	0.20	58.52	0.03	0.80
	λ 0.50	0.13	0.17	0.16	68.75	0.08	0.64	0.1	0.17	0.16	67.05	0.08	0.61
	λ 0.75	0.13	0.16	0.16	67.61	0.09	0.56	0.09	0.20	0.17	56.82	0.05	0.80
	λ 1.00	0.13	0.16	0.14	68.18	0.09	0.62	0.09	0.15	0.13	68.18	0.10	0.60
ViT-Small	Baseline	0.12	0.23	0.25				0.12	0.23	0.25			
	λ 0.00	0.13	0.10	0.10	69.32	0.12	0.48	0.13	0.06	0.05	80.11	0.16	0.30
	λ 0.25	0.12	0.09	0.10	71.59	0.14	0.42	0.12	0.07	0.07	77.84	0.16	0.33
	λ 0.50	0.12	0.09	0.10	75.57	0.13	0.45	0.12	0.07	0.07	75.00	0.15	0.35
	λ 0.75	0.12	0.09	0.09	75.57	0.14	0.39	0.12	0.06	0.05	81.82	0.17	0.27
	λ 1.00	0.12	0.14	0.15	70.45	0.09	0.57	0.12	0.09	0.08	68.75	0.14	0.44

Table 4: Evaluation of all models on Dataset 1 (Scenery Images). Each backbone is tested under baseline NIMA and both FR-NIMA variants using the same λ sweep. Reported metrics include the mean and standard deviation of the absolute flip gap, FCWR (%), AFGD, and AFGR. Positive AFGD values, higher FCWR percentages, and $AFGR < 1$ indicate that the flip-consistent models consistently outperform the baseline in maintaining prediction stability across horizontally flipped scenes. The baseline models are those trained without flip augmentation.

Model	Type	FR-NIMA-1					FR-NIMA-2				
		FlipGap		FCWR (%)	AFGD	AFGR	FlipGap		FCWR (%)	AFGD	AFGR
		Mean	Std				Mean	Std			
MobileNetV2	Baseline	0.21	0.17				0.21	0.17			
	λ 0.00	0.21	0.16	46.67	0.00	0.97	0.14	0.13	60.00	0.07	0.54
	λ 0.25	0.19	0.14	60.00	0.03	0.70	0.09	0.08	76.67	0.12	0.39
	λ 0.50	0.14	0.12	70.00	0.07	0.63	0.12	0.08	63.33	0.09	0.57
	λ 0.75	0.14	0.13	70.00	0.07	0.55	0.11	0.09	80.00	0.11	0.40
	λ 1.00	0.14	0.11	76.67	0.07	0.62	0.13	0.11	60.00	0.08	0.53
VGG19	Baseline	0.37	0.32				0.37	0.32			
	λ 0.00	0.22	0.14	60.00	0.15	0.66	0.17	0.14	73.33	0.20	0.42
	λ 0.25	0.25	0.24	63.33	0.12	0.67	0.16	0.12	73.33	0.22	0.41
	λ 0.50	0.16	0.13	83.33	0.21	0.43	0.13	0.12	73.33	0.24	0.27
	λ 0.75	0.19	0.22	73.33	0.18	0.45	0.14	0.13	83.33	0.23	0.31
	λ 1.00	0.17	0.18	76.67	0.20	0.33	0.19	0.18	73.33	0.18	0.50
Xception	Baseline	0.16	0.16				0.16	0.16			
	λ 0.00	0.14	0.10	53.33	0.02	1.05	0.10	0.10	63.33	0.05	0.78
	λ 0.25	0.13	0.10	50.00	0.03	1.01	0.10	0.09	60.00	0.05	0.84
	λ 0.50	0.13	0.12	43.33	0.02	0.85	0.09	0.09	70.00	0.07	0.65
	λ 0.75	0.13	0.11	56.67	0.03	0.81	0.10	0.07	63.33	0.05	0.87
	λ 1.00	0.14	0.12	46.67	0.02	0.95	0.08	0.07	60.00	0.08	0.56
InceptionV3	Baseline	0.14	0.13				0.14	0.13			
	λ 0.00	0.13	0.09	43.33	0.01	1.21	0.12	0.09	63.33	0.02	1.15
	λ 0.25	0.14	0.13	53.33	0.01	1.05	0.09	0.11	63.33	0.06	0.65
	λ 0.50	0.09	0.09	56.67	0.05	0.57	0.09	0.08	70.00	0.06	0.73
	λ 0.75	0.09	0.10	66.67	0.05	0.55	0.12	0.10	56.67	0.02	0.95
	λ 1.00	0.07	0.08	73.33	0.08	0.49	0.10	0.10	63.33	0.04	0.69
ViT-Small	Baseline	0.16	0.14				0.16	0.14			
	λ 0.00	0.06	0.06	80.00	0.10	0.36	0.06	0.05	80.00	0.11	0.34
	λ 0.25	0.07	0.05	63.33	0.09	0.48	0.06	0.04	83.33	0.11	0.35
	λ 0.50	0.06	0.06	76.67	0.10	0.34	0.04	0.03	86.67	0.12	0.25
	λ 0.75	0.06	0.05	80.00	0.10	0.40	0.04	0.03	83.33	0.12	0.27
	λ 1.00	0.10	0.11	73.33	0.06	0.61	0.04	0.05	86.67	0.12	0.19

Table 5: Evaluation of all models on Dataset 2 (Flowers Images). Results mirror those in Table 4, showing the flip-robust performance of the baseline NIMA and both FR-NIMA variants under different λ values. The Mean Gap and Std Gap reflect the degree and variability of flip-induced prediction changes, while FCWR (%) and AFGD quantify how often and how much each FR-NIMA model improves upon its baseline counterpart. Larger FCWR and AFGD values and consistently low AFGR values (< 1) signify greater flip consistency and reliability in assessing localized object-centric image content. Baseline results correspond to models trained without flip augmentation to ensure consistent comparison with FR-NIMA variants.

Model	Type	FR-NIMA-1					FR-NIMA-2				
		FlipGap		FCWR (%)	AFGD	AFGR	FlipGap		FCWR (%)	AFGD	AFGR
		Mean	Std				Mean	Std			
MobileNetV2	Baseline	0.17	0.23				0.17	0.23			
	λ 0.00	0.20	0.13	35.71	-0.03	1.76	0.20	0.16	35.71	-0.03	1.50
	λ 0.25	0.14	0.09	42.86	0.03	1.20	0.17	0.10	35.71	0.00	1.55
	λ 0.50	0.14	0.08	42.86	0.03	1.34	0.12	0.11	53.57	0.05	0.87
	λ 0.75	0.11	0.07	50.00	0.06	0.94	0.13	0.12	64.29	0.04	0.92
	λ 1.00	0.10	0.07	53.57	0.07	0.78	0.10	0.08	57.14	0.07	0.71
VGG19	Baseline	0.57	0.66				0.57	0.66			
	λ 0.00	0.13	0.11	75.00	0.43	0.32	0.18	0.19	71.43	0.39	0.39
	λ 0.25	0.18	0.16	67.86	0.38	0.37	0.19	0.18	75.00	0.38	0.37
	λ 0.50	0.21	0.21	67.86	0.36	0.45	0.17	0.18	71.43	0.39	0.36
	λ 0.75	0.22	0.20	67.86	0.35	0.51	0.15	0.14	75.00	0.42	0.35
	λ 1.00	0.17	0.14	71.43	0.40	0.42	0.12	0.09	78.57	0.45	0.33
Xception	Baseline	0.14	0.11				0.14	0.11			
	λ 0.00	0.10	0.08	75.00	0.04	0.86	0.13	0.11	60.71	0.00	1.10
	λ 0.25	0.10	0.08	67.86	0.04	0.83	0.11	0.12	53.57	0.03	0.73
	λ 0.50	0.08	0.07	75.00	0.05	0.56	0.14	0.12	50.00	-0.01	1.21
	λ 0.75	0.09	0.07	71.43	0.05	0.63	0.16	0.14	50.00	-0.02	1.05
	λ 1.00	0.10	0.08	67.86	0.04	0.83	0.12	0.12	53.57	0.02	0.89
InceptionV3	Baseline	0.18	0.14				0.18	0.14			
	λ 0.00	0.11	0.08	57.14	0.07	0.59	0.13	0.19	57.14	0.05	0.64
	λ 0.25	0.14	0.10	53.57	0.04	0.89	0.10	0.08	64.29	0.08	0.53
	λ 0.50	0.09	0.08	75.00	0.08	0.50	0.10	0.08	71.43	0.08	0.57
	λ 0.75	0.09	0.07	71.43	0.09	0.54	0.15	0.12	57.14	0.03	0.75
	λ 1.00	0.07	0.07	78.57	0.10	0.35	0.08	0.05	67.86	0.10	0.52
ViT-Small	Baseline	0.19	0.22				0.19	0.22			
	λ 0.00	0.06	0.04	71.43	0.13	0.40	0.06	0.05	82.14	0.13	0.39
	λ 0.25	0.06	0.05	75.00	0.13	0.39	0.04	0.03	78.57	0.15	0.31
	λ 0.50	0.05	0.05	78.57	0.14	0.28	0.06	0.05	67.86	0.14	0.32
	λ 0.75	0.07	0.04	67.86	0.12	0.54	0.05	0.04	78.57	0.15	0.25
	λ 1.00	0.09	0.11	78.57	0.10	0.40	0.06	0.05	75.00	0.13	0.38

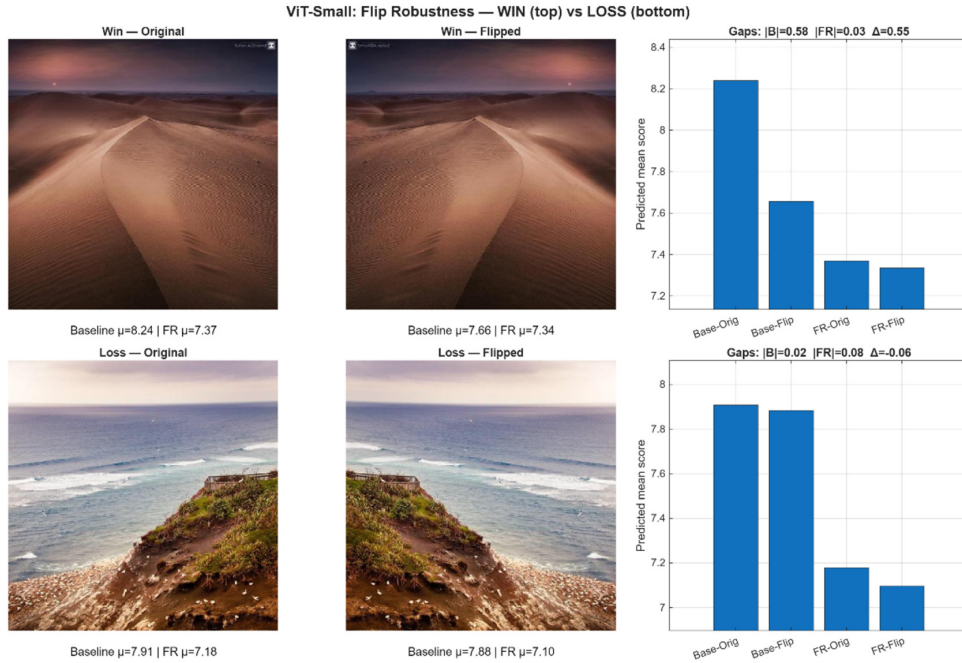


Fig.2: Qualitative examples of flip robustness on the Scenery dataset using the ViT-Small backbone. Each row shows an example image and its horizontally flipped counterpart, together with the predicted mean opinion scores (μ) from the baseline NIMA and FR-NIMA-2 ($\lambda = 0.75$) models. The top row shows the image with the largest reduction in flip gap (strongest win), while the bottom row shows the image with the largest increase in flip gap (strongest loss). These examples illustrate how the proposed flip-consistency regularization alters model behavior under horizontal flips, independent of human perceptual expectation.

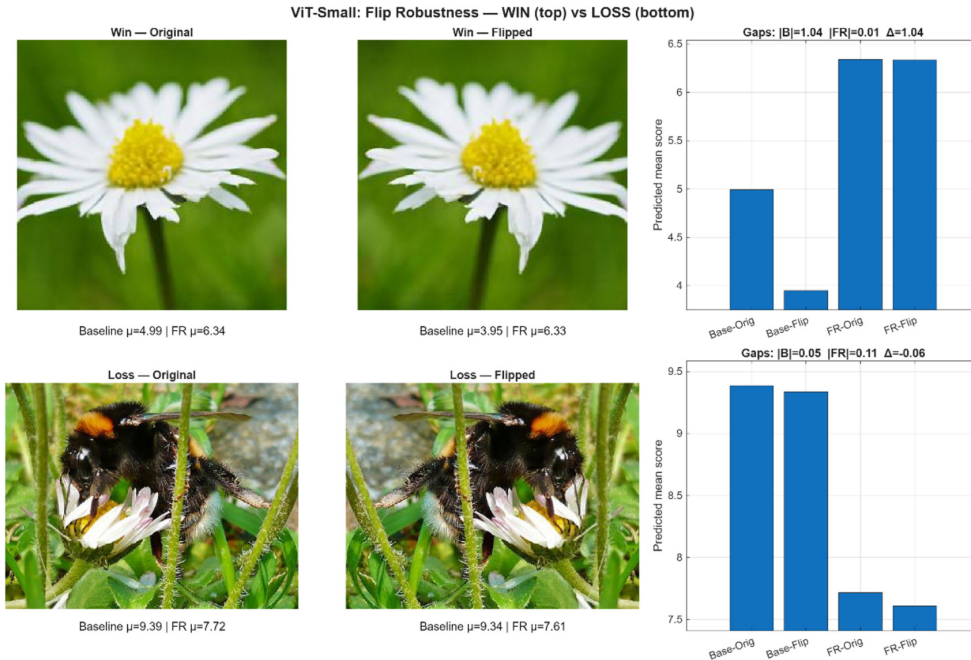


Fig.3: Qualitative examples of flip robustness on the Flower dataset using the ViT-Small backbone. The top row shows the strongest win case (largest flip-gap reduction), while the bottom row shows the strongest loss case (largest flip-gap increase). Although human observers may not perceive quality differences between the two views, the examples provide insight into how FR-NIMA’s regularization influences model sensitivity to mirrored spatial configurations.



Rawesak Tanawongsuwan is a faculty member in the faculty of Information and Communication Technology, Mahidol University, Thailand. His main research interests include computer vision, image processing, and computer graphics.



Sukanya Phongsuphap is a faculty member in the faculty of Information and Communication Technology, Mahidol University, Thailand. Her main research interests include image processing, pattern Recognition, digital signal processing, and heart rate variability.