FAITHFUL OR HALLUCINATED? A FIDELITY METRIC AND BENCHMARK FOR REAL-WORLD SUPER-RESOLUTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Real-world image super-resolution aims to restore high-resolution images from low-resolution inputs under unknown degradations. Reliable SR should be both perceptually appealing and faithful to the LR observation, meaning that restored content should not contradict the visual evidence in the input. Automatically assessing such fidelity is challenging without paired HR references. Existing metrics either ignore LR consistency or rely on degraded LR proxies that may penalize valid restoration changes. Human judgments can identify input-inconsistent content but are costly to collect at scale. These challenges motivate an HR-reference-free metric for LR–SR fidelity assessment. To this end, we propose **FAIM**, a **F**idelity-**A**ware **I**nput-consistency **M**etric for real-world super-resolution. FAIM leverages frozen multi-layer DINOv3 features to assess LR–SR fidelity by jointly modeling spatial discrepancies and region-level contextual inconsistencies. We further introduce **FAIM-Bench**, a human-annotated benchmark containing 3,720 LR–SR pairs constructed from 465 real-world LR images and eight diffusion-based $\times 4$ SR models. It includes separate ratings of LR quality, SR perceptual quality, and LR–SR fidelity, together with 13,020 unique pairwise fidelity comparisons. Experiments show that FAIM agrees more strongly with human fidelity judgments than existing metrics, including on unseen SR methods. In addition, FAIM supports zero-shot artifact localization without additional localization supervision.



Input SR-A (DiT4SR) SR-B (HYPIR)

Metrics	SR-A	SR-B	Higher Fidelity	
Human ↑	1.2180	**2.3000**	B	
FAIM ↑ **(Ours)**	-8.9691	**-6.2901**	B	✓
PSNR ↑	21.2638	**22.3572**	B	✓
SSIM ↑	**0.6510**	0.5920	A	✗
LPIPS ↓	**0.6144**	0.6466	A	✗
DISTS ↓	**0.4353**	0.4725	A	✗
NIQE ↓	**3.3482**	3.4631	A	✗
MANIQA ↑	**0.5436**	0.4810	A	✗
CLIP-IQA ↑	**0.6820**	0.6502	A	✗
MUSIQ ↑	**70.0341**	67.0414	A	✗
Q-Align ↑	**4.1914**	4.1289	A	✗
TOPIQ-NR ↑	**0.6803**	0.6623	A	✗
LIQE ↑	**4.6872**	4.0164	A	✗

Figure 1: Mismatch between existing metrics and human fidelity judgments in real-world SR. NR-IQA ignores LR consistency, while proxy-reference metrics may penalize valid changes such as degradation removal or plausible detail synthesis.

1 INTRODUCTION

Image super-resolution (SR) aims to reconstruct a high-resolution (HR) image from its low-resolution (LR) counterpart. As a fundamental problem in low-level vision, SR has been widely studied due to its broad applications in surveillance, remote sensing, and medical imaging. Recent advances in generative modeling, particularly diffusion-based models pretrained on large-scale data Duan et al. (2025); Fei et al. (2026), have significantly reshaped real-world super-resolution (RWSR).

Unlike earlier regression-based methods Lim et al. (2017), which often produce over-smoothed results, modern generative SR models can synthesize realistic textures and fine details. However,

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.

such generation may conflict with evidence in the LR input, producing input-inconsistent content. This makes fidelity an important complement to perceptual quality. We define LR–SR fidelity as the degree to which restored content remains consistent with the available LR evidence, whereas perceptual quality describes how visually satisfactory the SR output appears. An over-smoothed result may therefore be faithful but visually poor, while a realistic-looking result may contain LR-inconsistent structures or details and thus have low fidelity. Importantly, fidelity does not require a unique reconstruction or strict similarity to the LR appearance. Since SR is inherently ambiguous, multiple HR reconstructions may be equally faithful. Details that cannot be resolved from the LR input are not considered unfaithful simply because they are unverifiable. Fidelity decreases only when the restored content contradicts observable LR evidence. Likewise, degradations such as blur, noise, and compression artifacts should be removed rather than reproduced as image content.

Existing image quality assessment (IQA) metrics do not directly assess LR–SR fidelity. No-reference (NR) IQA metrics Yang et al. (2022); Wang et al. (2023); Wu et al. (2024a) evaluate the SR output alone and therefore cannot determine whether the restored content is supported by the LR observation. Full-reference (FR) IQA metrics can compare the SR output with a bicubic-upsampled LR proxy (ILR), but they measure similarity to the degraded proxy rather than fidelity to the underlying LR evidence. As shown in Fig. 1, neither NR-IQA scores nor ILR–SR proxy-reference similarities consistently agree with human fidelity judgments. Since paired HR references are generally unavailable in RWSR, fidelity assessment should rely only on the LR observation and SR output while determining whether the restored content remains compatible with the visual evidence available in the LR input.

Motivated by this need, we propose **FAIM**, a **F**idelity-**A**ware **I**nput-consistency **M**etric for RWSR. FAIM takes the bicubic-upsampled LR image and the SR output as inputs and assesses their fidelity by modeling whether the restored content remains compatible with the visual evidence available in the LR observation. It leverages frozen multi-layer DINOv3 features to model spatial discrepancies and region-level contextual inconsistencies. We further introduce **FAIM-Bench**, a densely annotated benchmark for generative RWSR fidelity assessment, containing 3,720 matched-input LR–SR pairs constructed from 465 real-world LR images across five datasets and eight representative diffusion-based $\times 4$ SR models. Through a two-stage annotation protocol, we collect human ratings of LR quality, SR perceptual quality, and LR–SR fidelity, together with 13,020 unique pairwise fidelity comparisons among SR outputs generated from the same LR inputs. Experiments show that FAIM agrees more strongly with human fidelity judgments than existing NR-IQA and ILR–SR proxy-reference metrics. The resulting inconsistency maps further support zero-shot artifact localization.

In summary, our contributions are three-fold:

- We propose **FAIM**, a fidelity-aware input-consistency metric for real-world super-resolution. It evaluates LR–SR fidelity by jointly modeling spatial discrepancies and region-level contextual inconsistencies, without requiring paired HR references.
- We introduce **FAIM-Bench**, a densely annotated benchmark for generative RWSR fidelity assessment, comprising 3,720 matched-input LR–SR pairs from 465 real-world LR images and eight diffusion-based $\times 4$ SR models, with human annotations of LR quality, SR perceptual quality, and LR–SR fidelity, together with 13,020 unique pairwise fidelity comparisons.
- Extensive experiments show that FAIM agrees more strongly with human fidelity judgments than existing NR-IQA and ILR–SR proxy-reference metrics, including on unseen SR methods.

2 Related Work

2.1 Single Image Super-Resolution

Deep learning has advanced single-image super-resolution in both reconstruction accuracy and perceptual quality. Earlier methods Kim et al. (2016) mainly optimized pixel-level agreement with ground-truth HR images, which often produced over-smoothed results with limited high-frequency detail. Subsequent methods Ledig et al. (2017); Wang et al. (2021); Zhang et al. (2021) employed perceptual and adversarial objectives to encourage sharper and more realistic restoration. More recently, diffusion models have become a prominent direction for real-world SR because of their strong ability to synthesize realistic details. StableSR Wang et al. (2024) introduces pretrained

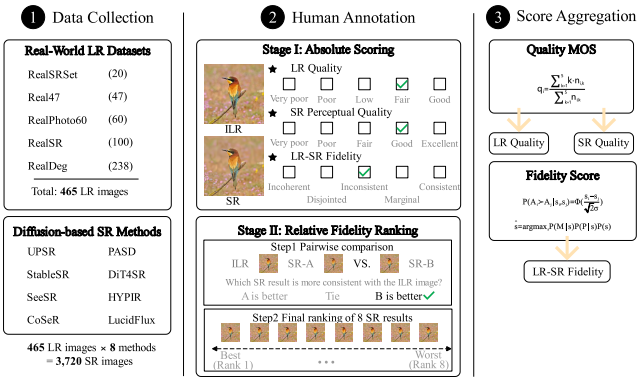


Figure 2: Construction pipeline of FAIM-Bench.

diffusion priors into RWSR through time-aware LR conditioning, while SeeSR Wu et al. (2024b) incorporates degradation-aware semantic prompts to reduce semantic errors. CoSeR Sun et al. (2024) leverages image-language priors and reference guidance, and PASD Yang et al. (2024) improves pixel-level control through pixel-aware cross-attention and degradation-insensitive guidance. DiT4SR Duan et al. (2025) adapts diffusion transformers to RWSR through LR-latent interaction and local cross-stream guidance, whereas UPSR Zhang et al. (2025) preserves reliable LR evidence using uncertainty-guided regional perturbation. HYPIR Lin et al. (2025) distills diffusion priors into a fast single-pass restoration model, and LucidFlux Fei et al. (2026) extends large diffusion transformers to caption-free universal restoration. Although these methods substantially improve perceptual quality, their generative priors can also introduce perceptually plausible but input-inconsistent content.

2.2 Image Quality Assessment

Image quality assessment is widely used to evaluate the visual quality of SR results. Full-reference metrics such as PSNR, SSIM Wang et al. (2004), LPIPS Zhang et al. (2018), and DISTS Ding et al. (2022) measure similarity to a reference image, while no-reference metrics such as NIQE Mittal et al. (2013), MUSIQ Ke et al. (2021), MANIQA Yang et al. (2022), CLIP-IQA Wang et al. (2023), LIQE Zhang et al. (2023), Q-Align Wu et al. (2024a), and TOPIQ Chen et al. (2024) assess image quality from the output alone. In RWSR, where paired HR references are typically unavailable, NR-IQA metrics are commonly used to assess perceptual quality. For input-consistency evaluation, the upsampled LR image is sometimes used as a proxy reference with full-reference metrics. However, NR-IQA does not consider the LR observation, while proxy-reference metrics measure similarity to the degraded LR proxy rather than whether the restored content remains compatible with the visual evidence available in the LR observation. Neither directly assesses LR–SR fidelity.

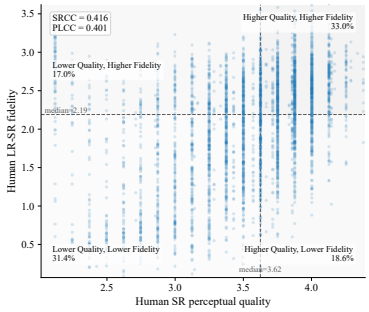
2.3 Segmentation with Self-Supervised Visual Features

Self-supervised visual features have been shown to capture spatial and semantic structures that are useful for segmentation without dense annotations. DINO Caron et al. (2021) demonstrates that self-supervised ViT features encode semantic organization and object boundaries. DINOv3 Siméoni et al. (2025) further improves dense feature quality at scale and introduces Gram anchoring to preserve localized representations during training. Building on these features, INSID3 Cuttano et al. (2026) demonstrates that frozen DINOv3 representations can support training-free segmentation. These findings motivate our use of frozen DINOv3 features to independently derive coarse regions for the ILR and SR images, providing structured context for fidelity assessment.

3 FAIM Benchmark

We construct **FAIM-Bench**, a densely annotated benchmark for real-world SR fidelity assessment based on human judgments. FAIM-Bench contains 3,720 LR–SR pairs constructed from 465 real-world LR source images and eight diffusion-based $\times 4$ SR models, as illustrated in Fig. 2. We collect

(a) Perceptual quality versus fidelity



(b) FAIM-Bench statistics

Dataset-level statistics

Dataset	#LR	#SR	LR quality	SR quality	Fidelity
RealDeg	238	1,904	3.501 ± 0.467	3.556 ± 0.464	1.955 ± 0.726
RealSRSet	20	160	3.269 ± 0.420	3.688 ± 0.368	2.558 ± 0.628
RealSR	100	800	3.353 ± 0.368	3.441 ± 0.323	2.019 ± 0.655
RealPhoto60	60	480	2.473 ± 0.294	3.099 ± 0.706	2.186 ± 0.680
Real47	47	376	2.544 ± 0.376	3.692 ± 0.327	2.131 ± 0.634

Overall distributions

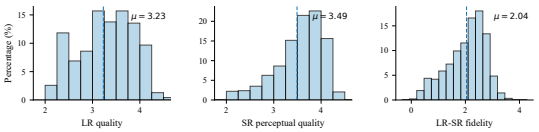


Figure 3: FAIM-Bench annotation analysis. (a) Human SR perceptual quality versus LR–SR fidelity. Dashed lines indicate the median perceptual-quality and fidelity scores, dividing the samples into four regimes with lower/higher perceptual quality and lower/higher fidelity. (b) Dataset-level and overall annotation statistics.

human judgments of LR image quality, SR perceptual quality, and LR–SR fidelity, together with 13,020 unique pairwise fidelity comparisons among SR outputs generated from the same LR inputs. The subjective study involves 12 annotators, with each LR–SR pair receiving at least eight valid judgments, resulting in approximately 186 annotator-hours of human evaluation.

3.1 Data Collection

We collect 465 real-world LR images from five RWSR datasets: RealDeg Chen et al. (2025), RealSRSet Zhang et al. (2021), RealSR Cai et al. (2019), RealPhoto60 Yu et al. (2024), and Real47 Lin et al. (2024). These images cover diverse real-world content and degradation conditions and are retained at their original resolutions. For each LR image, we generate $\times 4$ SR outputs using eight recent diffusion-based models: UPSR Zhang et al. (2025), StableSR Wang et al. (2024), SeeSR Wu et al. (2024b), CoSeR Sun et al. (2024), PASD Yang et al. (2024), DiT4SR Duan et al. (2025), HYPIR Lin et al. (2025), and LucidFlux Fei et al. (2026). Applying all eight models to every LR image yields $465 \times 8 = 3{,}720$ LR–SR pairs under matched input observations. We focus on diffusion-based SR models because current methods often produce perceptually plausible yet input-inconsistent content, making fidelity assessment particularly important for their evaluation. Regression-based methods Lim et al. (2017) are generally more conservative and produce less perceptually enhanced results, yielding fewer challenging cases for fine-grained human fidelity assessment. GAN-based methods Wang et al. (2021) can also exhibit fidelity issues, but their synthesized details are typically less realistic than those produced by recent diffusion-based methods. We therefore use diffusion-based models as the primary setting for constructing FAIM-Bench. For annotation and metric computation, each LR image is bicubic-upsampled by $\times 4$ to obtain the interpolated LR image at the same spatial resolution as its corresponding SR outputs. For a small number of ultra-high-resolution cases, we follow the default patch-wise inference strategy of the corresponding SR model.

3.2 Human Annotation Protocol

We collect human annotations for FAIM-Bench using a two-stage protocol with 12 annotators. Each LR–SR pair receives at least eight valid judgments, corresponding to approximately 186 annotator-hours in total. To reflect the real-world setting where paired HR references are typically unavailable, no ground-truth HR images are shown to the annotators. Fidelity is judged solely from the LR observation and its corresponding SR result, based on whether the restored content remains compatible with the visual evidence available in the LR input. In Stage I, annotators independently rate *LR quality*, *SR perceptual quality*, and *LR–SR fidelity* on five-level scales. Detailed criteria are provided in Appendix B.2. In Stage II, annotators compare the eight SR outputs generated from each LR image and provide relative fidelity judgments over all $\binom{8}{2} = 28$ output pairs. Across FAIM-Bench, this corresponds to 13,020 unique SR-output pairs, each evaluated by multiple annotators and aggregated into a pairwise fidelity preference. LR quality and SR perceptual quality are aggregated

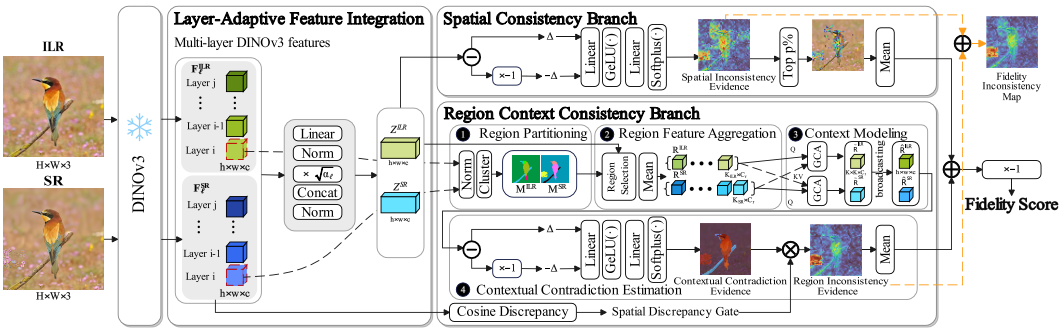


Figure 4: Overview of FAIM. LAFI adaptively integrates multi-layer DINOv3 features, which are evaluated by the Spatial Consistency Branch (SCB) and Region Context Consistency Branch (RCCB). Their aggregated inconsistency evidence yields the Fidelity Score after sign reversal, while the spatial responses are fused into the Fidelity Inconsistency Map.

using mean opinion scores (MOS) Wu et al. (2024a). For LR–SR fidelity, we combine absolute ratings and pairwise preferences following the hybrid subjective IQA formulation Ye & Doermann (2014) to obtain the final human fidelity scores. Details are provided in Appendix B.3.

3.3 Data Analysis

Perceptual quality versus fidelity. Fig. 3(a) shows that human SR perceptual quality and LR–SR fidelity are related but distinct, exhibiting only moderate correlation (SRCC=0.416, PLCC=0.401). Using the median scores as descriptive reference lines, 35.6% of the samples fall into the two off-diagonal regimes: 17.0% exhibit above-median fidelity but below-median perceptual quality, whereas 18.6% exhibit above-median perceptual quality but below-median fidelity. This substantial disagreement indicates that perceptual quality does not reliably reflect whether the restored content remains consistent with the visual evidence available in the LR input, motivating fidelity as a complementary dimension for SR evaluation.

Benchmark statistics. Fig. 3(b) summarizes the annotation statistics of FAIM-Bench. The human annotations reveal clear differences across the five source datasets in LR quality, SR perceptual quality, and LR–SR fidelity, while the overall distributions span a broad range for all three dimensions. These results highlight the diversity of FAIM-Bench in both input quality and SR restoration behavior.

4 A Fidelity-Aware Input-Consistency Metric for Real-World SR

4.1 Overview

The overall framework of FAIM is illustrated in Fig. 4. Given a real-world LR image I_{LR} and its super-resolved image I_{SR}, FAIM first bicubic-upsamples I_{LR} to the resolution of I_{SR}, yielding I_{ILR}. The two images are then passed through a frozen DINOv3 backbone Siméoni et al. (2025) to extract multi-layer features, which are adaptively integrated by the Layer-Adaptive Feature Integration (LAFI) module. The Spatial Consistency Branch (SCB) focuses on localized discrepancies at spatially aligned positions and aggregates the strongest responses, whereas the Region Context Consistency Branch (RCCB) captures broader contextual inconsistencies across regions through spatially gated region aggregation. The two aggregated signals are combined to obtain the final fidelity score. Their spatial response maps can additionally be combined to form a Fidelity Inconsistency Map for zero-shot artifact localization.

4.2 Layer-Adaptive Feature Integration Module

Passing I_{ILR} and I_{SR} separately through the frozen DINOv3 backbone yields multi-layer features $\{F_\ell^{\mathrm{ILR}}, F_\ell^{\mathrm{SR}}\}_{\ell\in\mathcal{L}}$. These features are projected into a shared embedding space and ℓ_2-normalized to obtain $\widehat{F}_\ell^{\mathrm{ILR}}$ and $\widehat{F}_\ell^{\mathrm{SR}}$. Since different DINO layers encode complementary visual cues, as shown

in Fig. 10, we assign each layer a learnable normalized weight α_ℓ. The weighted features are concatenated and normalized as $Z^X = \frac{[\sqrt{\alpha_\ell}\,\widehat{F}_\ell^X]_{\ell\in\mathcal{L}}}{\|[\sqrt{\alpha_\ell}\,\widehat{F}_\ell^X]_{\ell\in\mathcal{L}}\|_2}, X \in \{\mathrm{ILR}, \mathrm{SR}\}$. Since each projected feature is ℓ_2-normalized, we scale it by $\sqrt{\alpha_\ell}$ so that its squared magnitude in the concatenated representation is weighted by α_ℓ rather than α_ℓ^2. The resulting representations preserve spatial correspondence and are subsequently used by both SCB and RCCB.

4.3 Spatial Consistency Branch

SCB evaluates local discrepancies between spatially aligned SR and ILR representations. These discrepancies provide learned evidence of potential input inconsistency at each spatial location. For location i, we define $\Delta_i^S = Z_i^{\mathrm{SR}} - Z_i^{\mathrm{ILR}}$. To avoid introducing an arbitrary subtraction direction, we use a symmetric discrepancy scorer $\mathcal{S}_\eta(\Delta) = \frac{1}{2}[s_\eta(\Delta) + s_\eta(-\Delta)]$, where s_η is a lightweight two-layer MLP with a Softplus output. The spatial inconsistency evidence at location i is $E_i^S = \mathcal{S}_\eta(\Delta_i^S)$. We retain the largest p fraction of the evidence values and average them to obtain the spatial inconsistency score: $E^S = \frac{1}{|\mathcal{T}_p|}\sum_{i\in\mathcal{T}_p} E_i^S$.

4.4 Region Context Consistency Branch

SCB focuses on local discrepancies at spatially aligned locations. RCCB further captures region-level inconsistencies by modeling contextual relations between ILR and SR regions. We first partition I_{ILR} and I_{SR} independently using training-free agglomerative clustering over high-level DINOv3 features, following INSID3 Cuttano et al. (2026). Features within each region are pooled and projected to obtain Region Tokens $R^X \in \mathbb{R}^{K_X\times C_r}$, where K_X denotes the number of regions. Each token is further associated with a geometry descriptor $G^X \in \mathbb{R}^{K_X\times 3}$ containing its normalized centroid and log relative area. We then employ bidirectional geometry-aware cross-attention (GCA) to model contextual relations between ILR and SR regions. For each attention direction, let Y_1 denote the query side and Y_2 the source side. The m-th attention head projects their Region Tokens into $Q_m^{Y_1} \in \mathbb{R}^{K_{Y_1}\times d}$ and $K_m^{Y_2}, V_m^{Y_2} \in \mathbb{R}^{K_{Y_2}\times d}$, where $d = C_r/N_a$ and N_a is the number of attention heads. For each query–source region pair, the signed and absolute spatial offsets together with the relative log-area are encoded and mapped to a head-specific geometry bias $B^{Y_1,Y_2} \in \mathbb{R}^{N_a\times K_{Y_1}\times K_{Y_2}}$. The attention weights of head m are

$$A_m^{Y_1\to Y_2} = \mathrm{softmax}\Big(\frac{Q_m^{Y_1}(K_m^{Y_2})^\top}{\sqrt{d}} + B_m^{Y_1,Y_2}\Big), \tag{1}$$

where $A_m^{Y_1\to Y_2} \in \mathbb{R}^{K_{Y_1}\times K_{Y_2}}$. The corresponding response $A_m^{Y_1\to Y_2}V_m^{Y_2}$ is concatenated across all heads and projected back to the Region Token dimension. Applying GCA in both directions, $\mathrm{ILR}\to\mathrm{SR}$ and $\mathrm{SR}\to\mathrm{ILR}$, yields contextualized representations $\widetilde{R}^{\mathrm{ILR}}$ and $\widetilde{R}^{\mathrm{SR}}$. The contextualized Region Tokens are then broadcast back to their corresponding locations on the $h\times w$ feature grid $\widehat{R}_i^X = \widetilde{R}_{M^X(i)}^X, \widehat{R}^X \in \mathbb{R}^{h\times w\times C_r}$. Since the ILR and SR broadcast representations are spatially aligned, we define the contextual residual at location i as $\Delta_i^R = \widehat{R}_i^{\mathrm{SR}} - \widehat{R}_i^{\mathrm{ILR}}, \Delta_i^R \in \mathbb{R}^{C_r}$. An independent symmetric scorer produces the contextual contradiction response $C_i = \mathcal{S}_{\eta_R}(\Delta_i^R)$. Contextual contradiction alone may arise from region-level interactions without sufficient discrepancy in the original ILR–SR features. We therefore introduce a parameter-free Spatial Discrepancy Gate: $D_i = \frac{1}{|\mathcal{L}|}\sum_{\ell\in\mathcal{L}}[1-\cos(F_{\ell,i}^{\mathrm{ILR}}, F_{\ell,i}^{\mathrm{SR}})]$. The region inconsistency evidence at location i is then $E_i^R = D_iC_i$. We spatially average this map to obtain the region inconsistency score: $E^R = \frac{1}{hw}\sum_{i=1}^{hw} E_i^R$.

4.5 Fidelity Score Aggregation

SCB and RCCB yield the spatial inconsistency score E^S and region inconsistency score E^R, respectively. The final fidelity score is obtained by summing the two inconsistency scores and reversing their sign: $S_{\mathrm{FAIM}} = -(E^S + E^R)$. A higher score therefore indicates lower overall inconsistency and stronger fidelity to the LR observation. In addition to the scalar fidelity score, the Spatial Inconsistency Evidence and Region Inconsistency Evidence are combined element-wise to form the Fidelity Inconsistency Map: $E_i^F = E_i^S + E_i^R$. This map is used only for zero-shot artifact localization and does not affect the computation of the fidelity score.

Table 1: Comparison with existing metrics under the leave-one-SR-method-out evaluation protocol.

Metric	RealDeg-test		RealSR		RealPhoto60		Average	
	SRCC↑	PLCC-log↑	SRCC↑	PLCC-log↑	SRCC↑	PLCC-log↑	SRCC↑	PLCC-log↑
PSNR↑	0.133	0.244	0.145	0.266	0.357	0.478	0.212	0.329
SSIM↑	0.151	0.281	0.155	0.264	0.357	0.480	0.221	0.342
LPIPS↓	0.235	0.302	0.164	0.273	0.379	0.491	0.259	0.355
DISTS↓	0.359	**0.481**	0.083	0.240	0.442	0.518	0.295	0.413
NIQE↓	-0.121	0.220	-0.166	0.311	-0.290	0.456	-0.192	0.329
MANIQA↑	0.042	0.256	0.064	0.283	0.100	0.336	0.069	0.292
CLIP-IQA↑	0.107	0.263	0.072	0.246	-0.106	0.311	0.024	0.273
MUSIQ↑	0.030	0.219	-0.162	0.297	-0.153	0.433	-0.095	0.316
Q-Align↑	0.206	0.331	0.195	0.271	0.066	0.486	0.156	0.363
TOPIQ-NR↑	-0.007	0.248	-0.053	0.275	-0.145	0.421	-0.068	0.315
LIQE↑	0.041	0.224	0.070	0.244	-0.050	0.402	0.021	0.290
FAIM↑	**0.401**	0.474	**0.412**	**0.475**	**0.562**	**0.638**	**0.458**	**0.529**

4.6 Ranking-based Objectives

We train FAIM with pairwise RankNet Burges et al. (2005) supervision so that samples judged more faithful by humans receive higher predicted fidelity scores Sheng et al. (2026). Training uses only pairwise fidelity ordering, without regressing absolute MOS or fidelity scores. We construct two types of training pairs. **Same-image ranking** compares different SR results generated from the same LR image, using their human fidelity rankings to learn which restoration is more faithful under the same input. **Cross-image ranking** compares SR results from different LR images within the same dataset, using their human fidelity scores to make FAIM scores comparable across image contents. Stages I and II use both types of ranking supervision, while Stage III uses cross-image ranking to jointly fine-tune the final FAIM score. The detailed RankNet formulation and pair construction are provided in Appendix D.6.

5 Experiments

We compare FAIM with existing metrics under a leave-one-SR-method-out protocol on three real-world test sets, and conduct core ablations of its main components. We further demonstrate its zero-shot artifact localization capability. Appendix G provides complementary evaluation under the known-method LR-held-out setting, together with ground-truth HR analysis, artifact localization, statistical analyses, computational cost, and comprehensive ablations.

5.1 Experimental Details

Datasets. We split FAIM-Bench at the LR-image level. RealSRSet, Real47, and half of RealDeg form the development pool, containing 186 LR images, while the remaining half of RealDeg together with RealSR and RealPhoto60 forms the test split with 279 LR images. The development pool is further divided into 149 training and 37 validation LR images.

Evaluation protocol. Our main evaluation adopts an eight-fold leave-one-SR-method-out protocol, where each SR method is excluded entirely from FAIM training and validation and evaluated on the fixed test LR split. All baseline metrics are evaluated on the same held-out-method subsets. We compare FAIM with PSNR, SSIM Wang et al. (2004), LPIPS Zhang et al. (2018), DISTS Ding et al. (2022), NIQE Mittal et al. (2013), MANIQA Yang et al. (2022), CLIP-IQA Wang et al. (2023), MUSIQ Ke et al. (2021), Q-Align Wu et al. (2024a), TOPIQ-NR Chen et al. (2024), and LIQE Zhang et al. (2023), using ILR as the proxy reference for PSNR, SSIM, LPIPS, and DISTS. Agreement with human fidelity judgments is measured by SRCC and PLCC-log Sheikh et al. (2006), computed separately for each held-out method on each test dataset and macro-averaged over methods. Detailed definitions are provided in Appendix F.

Implementation. We use DINOv3-Large Siméoni et al. (2025) as the frozen feature backbone and extract features from layers $\mathcal{L} = \{6, 12, 18, 24\}$. SCB uses top-$10\%$ spatial aggregation. Region

Table 2: Ablation study of FAIM under the leave-one-SR-method-out protocol.

Variant	RealDeg-test		RealSR		RealPhoto60		Average	
	SRCC↑	PLCC-log↑	SRCC↑	PLCC-log↑	SRCC↑	PLCC-log↑	SRCC↑	PLCC-log↑
SCB only	**0.404**	0.491	0.305	0.385	0.454	0.540	0.388	0.472
RCCB only	0.376	0.468	**0.418**	0.461	0.562	0.657	0.452	**0.529**
DINOv3 last layer only	0.345	0.419	0.356	0.397	0.497	0.584	0.399	0.466
w/o Contextual Contradiction	0.397	0.470	0.375	0.416	0.563	**0.658**	0.445	0.514
w/o Spatial Discrepancy Gate	0.370	0.475	0.380	0.434	0.490	0.569	0.414	0.493
w/o Cross-Attention	0.338	0.417	0.314	0.400	0.516	0.626	0.390	0.481
w/o Geometry	0.387	**0.494**	0.405	0.438	**0.568**	0.622	0.453	0.518
Full	0.401	0.474	0.412	**0.475**	0.562	0.638	**0.458**	**0.529**

Tokens have dimension $C_r = 32$, and GCA uses $N_a = 2$ attention heads. FAIM is trained with the three-stage strategy described above. Further implementation details are provided in Appendix E.

5.2 Comparison with Existing Metrics

Table 1 compares FAIM with existing metrics under the leave-one-SR-method-out protocol. Output-only NR-IQA metrics show weak agreement with human LR–SR fidelity judgments, with average SRCC values ranging from -0.192 to 0.156. ILR–SR proxy-reference metrics perform better overall, with DISTS achieving the strongest baseline performance at an average SRCC of 0.295 and PLCC-log of 0.413. FAIM achieves the best average performance, reaching an SRCC of 0.458 and PLCC-log of 0.529, improving over DISTS by 0.163 and 0.116, respectively. FAIM also achieves the highest SRCC on all three test datasets. On RealDeg-test, DISTS obtains a slightly higher PLCC-log (0.481 versus 0.474), while FAIM retains the stronger rank correlation (0.401 versus 0.359). Because the held-out SR method is absent from both training and validation, these results indicate that FAIM captures transferable LR–SR fidelity cues rather than relying primarily on method-specific score offsets or restoration artifacts.

5.3 Ablation Study

Tab. 2 examines the main components of FAIM under the same leave-one-SR-method-out protocol as the main comparison. Each variant is evaluated using the same held-out-method folds and macro-averaging procedure, allowing us to assess how individual components contribute to generalization across unseen SR methods. More extensive design analyses under the known-method LR-held-out setting are provided in Appendix G.4.

Branch contribution. SCB and RCCB achieve average SRCC/PLCC-log of $0.388/0.472$ and $0.452/0.529$, respectively, showing that both spatial and region-level inconsistency evidence are informative. RCCB performs better on average, particularly on RealSR and RealPhoto60, while SCB remains stronger on RealDeg-test. This difference is consistent with their aggregation mechanisms. SCB focuses on the top-10% spatial responses and is therefore more sensitive to shifts in artifact extent and response distribution, whereas RCCB aggregates broader region-level contextual evidence. Combining both branches yields the highest average SRCC of 0.458 and matches RCCB in PLCC-log at 0.529, confirming their complementarity under unseen-method evaluation.

Multi-layer representation. Using only the final DINOv3 layer reduces the average SRCC/PLCC-log from $0.458/0.529$ to $0.399/0.466$, with consistent degradation across all three datasets. This demonstrates that fidelity cannot be fully characterized by high-level semantic representations alone. LR–SR inconsistencies may arise at different levels, ranging from local appearance and structure to semantic content. Integrating features across multiple depths therefore provides complementary evidence across these levels, leading to more robust fidelity assessment on unseen SR methods.

RCCB design. We ablate the two components that determine how region-level inconsistency is formed within RCCB while keeping SCB unchanged. Removing Contextual Contradiction reduces the average SRCC/PLCC-log from $0.458/0.529$ to $0.445/0.514$, while removing the Spatial Discrepancy Gate leads to a larger decline to $0.414/0.493$. This suggests that region-level fidelity benefits from combining contextual reasoning with direct evidence of LR–SR discrepancy. In particular, the

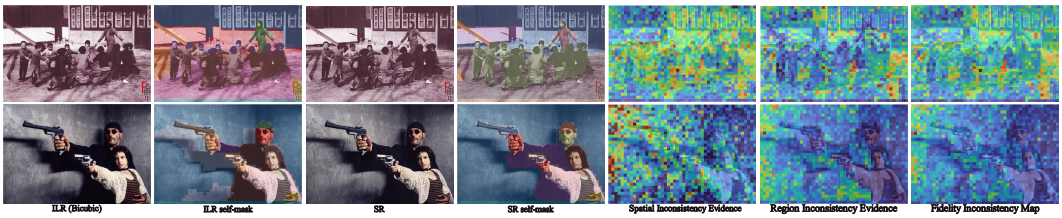


Figure 5: Zero-shot artifact localization of FAIM.

Spatial Discrepancy Gate provides an important constraint by emphasizing regions whose contextual differences are also supported by observable spatial disagreement, while Contextual Contradiction captures complementary inconsistencies in regional content and relations.

Region-context modeling. We further examine how RCCB models relations between ILR and SR regions. Removing cross-attention decreases the average SRCC/PLCC-log from $0.458/0.529$ to $0.390/0.481$, with degradation across all three datasets, highlighting the importance of explicitly comparing corresponding regional representations rather than treating regions independently. Removing the geometry bias results in a smaller decrease to $0.453/0.518$, indicating that relative position and region size provide useful complementary cues but are not the primary source of region-level discrimination. Overall, these results show that RCCB mainly benefits from relational comparison across regions, while geometry further refines this comparison by introducing spatial context.

5.4 Zero-Shot Artifact Localization

Fig. 5 presents qualitative results of FAIM for artifact localization. The independently obtained ILR and SR self-segmentation masks capture coherent regions around major objects and scene structures without segmentation supervision. The Spatial Inconsistency Evidence primarily highlights local LR–SR discrepancies, while the Region Inconsistency Evidence captures inconsistencies over broader regions. Combining the two produces the Fidelity Inconsistency Map, which localizes regions where the SR output is inconsistent with the LR observation. These results show that FAIM provides useful localization cues without spatial artifact supervision. Additional qualitative and quantitative evaluations on a dedicated artifact localization benchmark are provided in Appendix G.3.

6 Conclusion and Limitations

We proposed **FAIM**, a fidelity-aware input-consistency metric for real-world image super-resolution that assesses whether restored content remains compatible with the visual evidence available in the LR observation through complementary spatial and region-level consistency modeling. We further introduced **FAIM-Bench**, a human-annotated benchmark containing 3,720 LR–SR pairs constructed from 465 real-world LR images and eight diffusion-based $\times 4$ SR models. It includes separate ratings of LR quality, SR perceptual quality, and LR–SR fidelity, together with 13,020 unique pairwise fidelity comparisons. Experiments show that FAIM agrees more strongly with human fidelity judgments than existing NR-IQA and ILR–SR proxy-reference metrics, while its resulting inconsistency maps further enable zero-shot artifact localization. FAIM has three main limitations. First, its patch-level DINOv3 representation may miss fine-grained pixel artifacts. Second, FAIM does not explicitly model the perceptual importance of different image regions. Although RCCB partitions the image into coherent regions and models their contextual inconsistencies across regions, it does not distinguish foreground from background or determine which regions are more important to human perception. Consequently, spatially extensive inconsistencies in less important regions may have a stronger influence on the final fidelity assessment than localized errors in regions of greater perceptual importance, such as faces or foreground objects. Third, the current annotation scale remains limited, motivating broader human validation in future work.

AI USE STATEMENT

Generative AI tools were used solely for English language polishing, including improving grammar, wording, and readability.

ETHICS STATEMENT

FAIM-Bench involves low-risk subjective image assessment. Annotators were informed about the study, compensated for their participation, and their responses were anonymized and reported only in aggregate. Formal IRB review was not required under the applicable institutional policy. Additional details are provided in Appendix J and Appendix I.

REPRODUCIBILITY STATEMENT

We provide anonymized code and FAIM-Bench data in Appendix A. Details of the annotation protocol, implementation and training, evaluation protocol, ablation studies, and computational resources are provided in Appendices B.2, E, F, G.4, and G.6, respectively.

REFERENCES

Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *Computer Vision and Pattern Recognition*, pp. 6228–6237, 2018.

Christopher J. C. Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Gregory N. Hullender. Learning to rank using gradient descent. In *ICML*, volume 119, pp. 89–96, 2005.

Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, and Lei Zhang. Toward real-world single image super-resolution: A new benchmark and a new model. In *International Conference on Computer Vision*, pp. 3086–3095, 2019.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, pp. 9630–9640, 2021.

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.

Junyang Chen, Jinshan Pan, and Jiangxin Dong. Faithdiff: Unleashing diffusion priors for faithful image super-resolution. In *Computer Vision and Pattern Recognition*, pp. 28188–28197, 2025.

Regev Cohen, Idan Kligvasser, Ehud Rivlin, and Daniel Freedman. Looks too good to be true: An information-theoretic analysis of hallucinations in generative restoration models. In *Advances in Neural Information Processing Systems*, 2024.

Claudia Cuttano, Gabriele Trivigno, Christoph Reich, Daniel Cremers, Carlo Masone, and Stefan Roth. INSID3: training-free in-context segmentation with dinov3. In *Computer Vision and Pattern Recognition*, 2026.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(5):2567–2581, 2022.

Zheng-Peng Duan, Jiawei Zhang, Xin Jin, Ziheng Zhang, Zheng Xiong, Dongqing Zou, Jimmy S. Ren, Chun-Le Guo, and Chongyi Li. Dit4sr: Taming diffusion transformer for real-world image super-resolution. In *International Conference on Computer Vision*, 2025.

Song Fei, Tian Ye, Lujia Wang, and Lei Zhu. Lucidflux: Caption-free universal image restoration via a large-scale diffusion transformer. In *International Conference on Learning Representations*, 2026.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: multi-scale image quality transformer. In *International Conference on Computer Vision*, pp. 5128–5137, 2021.

Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Computer Vision and Pattern Recognition*, pp. 1646–1654, 2016.

Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Computer Vision and Pattern Recognition*, pp. 4681–4690, 2017.

Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Computer Vision and Pattern Recognition Workshops*, pp. 136–144, 2017.

Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Ben Fei, Bo Dai, Wanli Ouyang, Yu Qiao, and Chao Dong. Diffbir: Towards blind image restoration with generative diffusion prior. In *European Conference on Computer Vision*, 2024.

Xinqi Lin, Fanghua Yu, Jinfan Hu, Zhiyuan You, Wu Shi, Jimmy S. Ren, Jinjin Gu, and Chao Dong. Harnessing diffusion-yielded score priors for image restoration. *ACM Trans. Graph.*, 44(6): 208:1–208:21, 2025.

Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Process. Lett.*, 20(3):209–212, 2013.

Ivan Molodetskikh, Kirill Malyshev, Mark Mirgaleev, Nikita Zagainov, Evgeney Bogatyrev, and Dmitriy S. Vatolin. Sr-prominence: A crowdsourced protocol and dataset suite for perceptually-weighted super-resolution artifact evaluation. *CoRR*, abs/2605.14847, 2026.

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

Xiangfei Sheng, Xiaofeng Pan, Zhichao Yang, Pengfei Chen, and Leida Li. Fine-grained image quality assessment for perceptual image restoration. In *AAAI*, pp. 8914–8922, 2026.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, 2025. URL `https://arxiv.org/abs/2508.10104`.

Haoze Sun, Wenbo Li, Jianzhuang Liu, Haoyu Chen, Renjing Pei, Xueyi Zou, Youliang Yan, and Yujiu Yang. Coser: Bridging image and language for cognitive super-resolution. In *Computer Vision and Pattern Recognition*, pp. 25868–25878, 2024.

Jianyi Wang, Kelvin C. K. Chan, and Chen Change Loy. Exploring CLIP for assessing the look and feel of images. In *Association for the Advancement of Artificial Intelligence*, pp. 2555–2563, 2023.

Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin C. K. Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *International Journal of Computer Vision*, 2024.

Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data. In *International Conference on Computer Vision Workshops*, pp. 1905–1914, 2021.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Chunyi Li, Liang Liao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtai Zhai, and Weisi Lin. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. In *ICML*, 2024a.

Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. Seesr: Towards semantics-aware real-world image super-resolution. In *Computer Vision and Pattern Recognition*, pp. 25456–25467, 2024b.

Liangbin Xie, Xintao Wang, Xiangyu Chen, Gen Li, Ying Shan, Jiantao Zhou, and Chao Dong. Desra: Detect and delete the artifacts of gan-based real-world super-resolution models. In *ICML*, volume 202, pp. 38204–38226, 2023.

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 *Computer Vision and Pattern Recognition*, pp. 1191–1200, 2022.

Tao Yang, Rongyuan Wu, Peiran Ren, Xuansong Xie, and Lei Zhang. Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization. In *European Conference on Computer Vision*, pp. 74–91, 2024.

Peng Ye and David S. Doermann. Active sampling for subjective image quality assessment. In *Computer Vision and Pattern Recognition*, pp. 4249–4256, 2014.

Fanghua Yu, Jinjin Gu, Zheyuan Li, Jinfan Hu, Xiangtao Kong, Xintao Wang, Jingwen He, Yu Qiao, and Chao Dong. Scaling up to excellence: Practicing model scaling for photo-realistic image restoration in the wild. In *Computer Vision and Pattern Recognition*, pp. 25669–25680. IEEE, 2024.

Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *International Conference on Computer Vision*, pp. 4791–4800, 2021.

Leheng Zhang, Weiyi You, Kexuan Shi, and Shuhang Gu. Uncertainty-guided perturbation for image super-resolution diffusion model. In *Computer Vision and Pattern Recognition*, pp. 17980–17989, 2025.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

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

A Code, Dataset, and Sample Availability

To support reproducibility while preserving double-blind review, we provide anonymized links to the FAIM code and FAIM-Bench dataset:

Code: `https://anonymous.4open.science/r/FAIM-463B/`
Dataset: `https://kaggle.com/datasets/bfe00078ea5e1f59ad336d5f1eb66916c97e63373b92cc2b26dba55d8ec3d070`

The dataset link provides the released FAIM-Bench annotation files and images required for training and evaluating FAIM.

B Annotation Details

B.1 Data Collection

Table 3: Composition of the LR images in FAIM-Bench.

Source dataset	# LR images	Content and characteristics
RealDeg Chen et al. (2025)	238	Old photos, social-media images, movie stills; people, buildings, animals, natural scenes
RealSRSet Zhang et al. (2021)	20	Widely used real-world SR benchmark images with mixed real degradations
RealSR Cai et al. (2019)	100	DSLR-captured indoor and outdoor scenes with real optical degradations
RealPhoto60 Yu et al. (2024)	60	Real-world photos from common benchmarks and web sources
Real47 Lin et al. (2024)	47	Internet images covering landscapes, old photos, architecture, humans, plants, and animals
Total	465	Diverse real-world contents and degradation characteristics

We collect 465 real-world LR images from five widely used RWSR datasets: RealDeg Chen et al. (2025), RealSRSet Zhang et al. (2021), RealSR Cai et al. (2019), RealPhoto60 Yu et al. (2024), and Real47 Lin et al. (2024). These datasets cover diverse content, including natural scenes, portraits, and movie scenes. The number of LR images from each dataset is summarized in Tab. 3. We retain all LR images at their original resolutions without additional preprocessing. For each LR image, we generate $\times 4$ SR results using eight representative diffusion-based super-resolution models: UPSR Zhang et al. (2025), StableSR Wang et al. (2024), SeeSR Wu et al. (2024b), CoSeR Sun et al. (2024), PASD Yang et al. (2024), DiT4SR Duan et al. (2025), HYPIR Lin et al. (2025), and LucidFlux Fei et al. (2026). Applying each model to each LR image forms a complete source–generator Cartesian product, resulting in $465 \times 8 = 3{,}720$ SR outputs. This matched-input construction ensures that all restoration methods are evaluated under exactly the same set of LR observations. Due to GPU memory constraints, for a small number of ultra-high-resolution images whose $\times 4$ outputs exceed 4K resolution, certain models perform SR using overlapping patch-wise generation. We follow the default inference strategy of each model. For annotation and metric computation, we bicubic-upsample each LR image by $\times 4$ to obtain the interpolated LR image, which has the same spatial resolution as the corresponding SR outputs.

B.2 Five-Level Rating Criteria

We collect absolute ratings for three aspects: *LR quality*, *SR perceptual quality*, and *LR–SR fidelity*. Each aspect is rated on a five-level discrete scale. All subjective evaluation is conducted using only the LR observation and the corresponding SR result, without access to an HR reference. Specifically, the LR image is bicubic-upsampled by $\times 4$ to obtain the interpolated LR image, which is spatially aligned with the SR output and is used as the reference observation for LR–SR comparison. We adopt this no-reference-to-HR setting because paired HR images are often unavailable in real-world super-resolution scenarios. This protocol therefore reflects the practical setting in which fidelity must be assessed from the compatibility between the restored content and the visual evidence available in the observed LR input, rather than by comparison with an unavailable ground-truth HR image. Because the LR observation may be ambiguous, the protocol does not assume that a unique HR reconstruction can be inferred from the input.

LR quality. LR quality measures how much visual evidence is available in the LR input itself. Annotators judge the recognizability and informativeness of the LR image without considering any SR result. The five levels are defined as follows:

- **Very poor**: the image is almost unrecognizable, with little usable visual evidence.

- **Poor**: only coarse color or shape cues are visible, and most content is difficult to identify.
- **Low**: the main content can be partially inferred, but important structures or details remain unclear.
- **Fair**: the main content is generally recognizable despite visible blur, noise, compression, or other degradations.
- **Good**: the image is relatively clear and provides rich visual evidence for restoration.

SR perceptual quality. SR perceptual quality measures the visual quality of the restored image. Annotators evaluate whether the SR image looks clear, natural, and artifact-free, regardless of whether it is faithful to the LR input. The five levels are defined as follows:

- **Very poor**: the image contains severe blur, distortion, or artifacts and is difficult to use.
- **Poor**: the main content is visible, but the image has obvious artifacts, unnatural textures, or low visual quality.
- **Fair**: the image is visually acceptable, but noticeable blur, noise, artifacts, or unnatural details remain.
- **Good**: the image is mostly clear and natural, with only minor local imperfections.
- **Excellent**: the image is sharp, natural, and visually convincing, with few visible artifacts.

LR–SR fidelity. LR–SR fidelity measures the degree to which the SR output remains compatible with the visual evidence available in the LR input. Annotators compare restored structures, textures, identities, text, layouts, and local details against what can actually be observed from the LR image, rather than judging visual quality alone. Importantly, the LR observation may be ambiguous and does not uniquely determine all HR details. Therefore, a synthesized detail should not be penalized merely because it is not explicitly visible or verifiable in the LR input. Fidelity is reduced when the SR output introduces content that conflicts with observable LR evidence. Removing degradations such as blur, noise, or compression is likewise not considered a fidelity violation. The five levels are defined as follows:

- **Incoherent**: The SR image contains severe semantic, structural, identity, text, or layout changes that clearly conflict with observable evidence in the LR input.
- **Disjointed**: Many regions contain restored structures or details that are difficult to reconcile with the observable LR evidence, producing substantial LR–SR inconsistency.
- **Inconsistent**: The SR image largely preserves the LR-observable content but contains several noticeable local structures or details that conflict with available LR evidence.
- **Marginal**: The SR image is mostly compatible with the LR observation, with only minor local inconsistencies. Details that are ambiguous or not explicitly resolvable from the LR input are not penalized unless they conflict with observable evidence.
- **Consistent**: The SR image remains compatible with the visual evidence available in the LR observation, with no clear LR–SR contradiction. Synthesized fine details may be acceptable even when they are not explicitly visible in the LR input, provided that they do not conflict with that evidence.

B.3 Score Aggregation

For LR quality and SR perceptual quality, we aggregate absolute ratings using the mean opinion score (MOS) Wu et al. (2024a). Let $n_{i,k}$ denote the number of annotators assigning score k to image i, where $k \in \{1, 2, 3, 4, 5\}$. The aggregated quality score is computed as

$$q_i = \frac{\sum_{k=1}^{5} k \cdot n_{i,k}}{\sum_{k=1}^{5} n_{i,k}}. \tag{2}$$

For LR–SR fidelity, we aggregate absolute fidelity ratings and pairwise preferences into a single latent human fidelity score. Let M_{ik} denote the number of absolute ratings assigning image A_i to

ordinal level k, where $k \in \{1, \ldots, 5\}$. We encode the five ordinal levels as $c_k = k - 3, \quad c_k \in \{-2, -1, 0, 1, 2\}$. The ordinal thresholds are fixed rather than learned:

$$\gamma_0 = -\infty, \quad \gamma_5 = +\infty, \quad \gamma_k = \frac{c_k + c_{k+1}}{2}, \quad k = 1, \ldots, 4, \tag{3}$$

i.e., $(\gamma_1, \gamma_2, \gamma_3, \gamma_4) = (-1.5, -0.5, 0.5, 1.5)$. Given a latent score s_i, the absolute rating likelihood is modeled by an ordinal probit model:

$$P(o_i = k \mid s_i) = \Phi(\frac{\gamma_k - s_i}{\sigma_{\text{abs}}}) - \Phi(\frac{\gamma_{k-1} - s_i}{\sigma_{\text{abs}}}), \tag{4}$$

where o_i denotes the ordinal absolute fidelity rating assigned to item A_i and $\Phi(\cdot)$ is the standard Gaussian CDF. We fix $\sigma_{\text{abs}} = 1.0$. For pairwise preferences, if A_i is judged more faithful than A_j, we use the Thurstone–Mosteller formulation:

$$P(A_i \succ A_j \mid s_i, s_j) = \Phi(\frac{s_i - s_j}{\sqrt{2}\sigma_{\text{pair}}}), \tag{5}$$

where we fix $\sigma_{\text{pair}} = 1.0$. Let P_{ij} denote the number of pairwise comparisons in which A_i is preferred over A_j. We use a weak zero-mean Gaussian prior on the latent scores:

$$P(\mathbf{s}) = \prod_i \mathcal{N}(s_i; 0, \tau_s^2), \quad \tau_s = 3.0. \tag{6}$$

This prior only regularizes the global scale and prevents degenerate score drift; it does not impose any ordering constraint. The final scores are estimated by maximizing the following log posterior:

$$\begin{aligned} \hat{\mathbf{s}} = \arg\max_{\mathbf{s}} \quad & \sum_i \sum_{k=1}^{5} M_{ik} \log[P(o_i = k \mid s_i) + \epsilon] \\ & + \sum_{i \neq j} P_{ij} \log[\Phi(\frac{s_i - s_j}{\sqrt{2}\sigma_{\text{pair}}}) + \epsilon] - \sum_i \frac{s_i^2}{2\tau_s^2}, \end{aligned} \tag{7}$$

where $\epsilon = 10^{-8}$ is used for numerical stability. We optimize only the latent scores $\mathbf{s}$; the ordinal thresholds $\{\gamma_k\}$, probit noise scales σ_{abs} and σ_{pair}, and prior scale τ_s are fixed. The scores are initialized by the mean absolute rating after mapping the five ordinal labels to $\{-2, -1, 0, 1, 2\}$. The optimization is unconstrained with respect to $\mathbf{s}$, while the ordering of the ordinal thresholds is guaranteed by the fixed threshold construction above. The resulting latent scores $\hat{\mathbf{s}}$ lie on an unconstrained interval scale and are used as the final human fidelity scores in FAIM-Bench. Compared with using absolute ratings alone, this hybrid aggregation preserves both coarse subjective fidelity levels and fine-grained relative preferences among the SR results corresponding to the same LR input.

B.4 Inter-Annotator Agreement and Annotation Variability

We analyze inter-annotator agreement (IAA) using raw annotations from 12 anonymous annotators to characterize the uncertainty underlying the subjective judgments in FAIM-Bench. We do not assume that any individual annotator provides an objective ground-truth fidelity score. Since annotators labeled partially overlapping subsets of FAIM-Bench, agreement is computed only on shared items. For absolute ratings, LR quality is treated as an LR-image-level item, whereas SR perceptual quality and LR–SR fidelity are treated as LR/SR-pair-level items. For LR quality, duplicate entries associated with different SR outputs of the same LR image are collapsed into a single LR-level score; no within-item conflicts are observed after this grouping.

For absolute ratings, we report pairwise SRCC, PLCC, and quadratic weighted Cohen's kappa (QWK), macro-averaged over annotator pairs satisfying the overlap threshold. We require at least 20 overlapping LR images for LR quality and at least 100 overlapping LR/SR pairs for SR perceptual quality and LR–SR fidelity. We additionally report Krippendorff's alpha, computed jointly across all annotators while allowing missing labels.

As shown in Tab. 4, LR–SR fidelity obtains pairwise SRCC and PLCC values of 0.424 and 0.454, respectively, with QWK of 0.296 and Krippendorff's alpha of 0.307. Although these values are consistently higher than those for SR perceptual quality under the reported measures, the absolute

Table 4: Inter-annotator agreement for absolute ratings on overlapping items. Pairwise SRCC, PLCC, and QWK are macro-averaged over annotator pairs satisfying the overlap threshold. Krippendorff's alpha is computed across all annotators with missing labels allowed.

Target	Pairs	Avg. overlap	Min. overlap	SRCC	PLCC	QWK	Alpha
LR quality	57	233.7 LR images	26	0.373	0.377	0.280	0.363
SR perceptual quality	60	1777.2 LR/SR pairs	144	0.276	0.298	0.201	0.226
LR–SR fidelity	60	1777.2 LR/SR pairs	144	0.424	0.454	0.296	0.307

Table 5: Inter-annotator agreement for LR–SR fidelity rankings on overlapping LR images. Agreement is computed after aligning common SR methods for each overlapping LR image, and the reported correlations are macro-averaged over annotator pairs.

Target	Pairs	Avg. overlapping LR images	Avg. valid LR images	Avg. common methods	Agreement
LR–SR fidelity ranking	57	233.5	233.4	7.94	Kendall's $\tau_b = 0.312$, SRCC=0.399

agreement remains limited. This indicates substantial annotator variability in judging LR–SR fidelity and motivates interpreting the aggregated scores as consensus human judgments rather than noise-free ground truth.

For relative fidelity rankings, each item corresponds to one LR image with multiple SR outputs. For each overlapping LR image, we align the SR methods ranked by both annotators and compute Kendall's τ_b and SRCC over the common methods. These correlations are first averaged over overlapping LR images within each annotator pair and are then macro-averaged across annotator pairs.

Tab. 5 shows a mean Kendall's τ_b of 0.312 and a mean SRCC of 0.399. When correlations are averaged directly over all valid annotator-pair–LR-image cases rather than first averaging within each annotator pair, the corresponding values are 0.344 and 0.439. These results indicate substantial variability in the fine-grained relative ordering of SR outputs corresponding to the same LR observation.

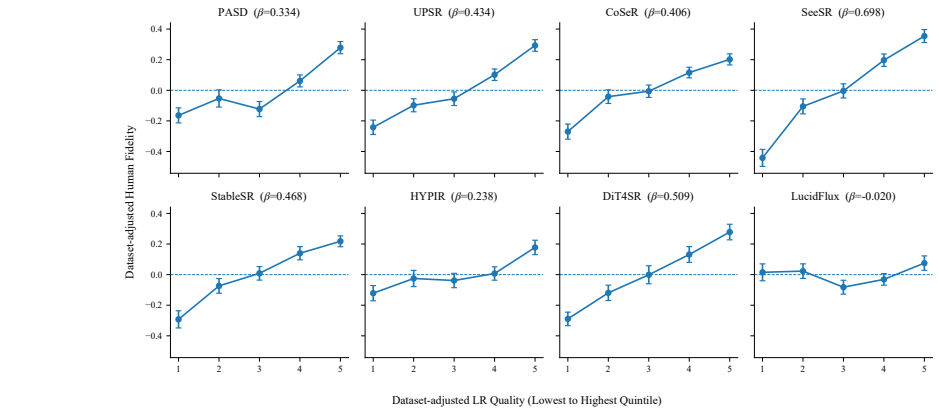
We also observe systematic differences in how individual annotators use the absolute rating scale, particularly for LR–SR fidelity. Multiple annotations and score aggregation reduce dependence on any single annotator, but they do not eliminate the subjective uncertainty inherent in fidelity assessment. Accordingly, the consensus fidelity scores in FAIM-Bench should be interpreted as aggregated human judgments rather than objective fidelity values, and small differences in metric correlations should be interpreted with appropriate caution.

C FAIM-Bench Data Analysis

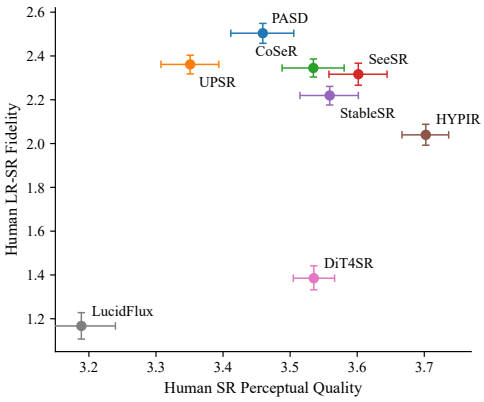
This section uses FAIM-Bench to characterize the behavior of current diffusion-based RWSR models beyond aggregate quality scores. We analyze the relationship between perceptual quality and LR–SR fidelity, model sensitivity to the available LR evidence, cross-domain consistency, low-fidelity tail risk, and scene-level pairwise dominance. We then examine how conventional perceptual-quality and proxy-reference metrics relate to human fidelity judgments and provide detailed method-level results across the five source datasets.

C.1 What FAIM-Bench Reveals about Generative RWSR

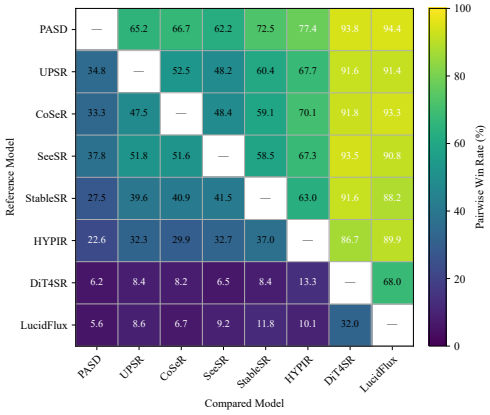
Perceptual Quality and Fidelity Define Different Model Operating Points. Fig. 6(a) places the eight SR models in the two-dimensional space of human SR perceptual quality and human LR–SR fidelity. The model ordering changes substantially across the two axes. HYPIR attains the highest mean perceptual quality (3.702) but only a moderate mean fidelity (2.040), whereas PASD attains the highest mean fidelity (2.504) with a lower perceptual-quality score (3.459). The contrast between CoSeR and DiT4SR is particularly revealing: their mean perceptual-quality scores are virtually identical (3.535 after rounding), yet their mean fidelity scores differ by nearly one point (2.345 versus 1.385). Thus, stronger perceptual quality does not imply a corresponding ordering in LR–SR fidelity.



(b) Sensitivity to LR evidence



(a) Quality–fidelity operating points



(c) Pairwise fidelity dominance

Figure 6: Model-level behavior revealed by FAIM-Bench. (a) Mean human SR perceptual quality versus mean human LR–SR fidelity for the eight SR models, with 95% bootstrap confidence intervals. (b) Dataset-adjusted fidelity across LR-quality quintiles; β denotes the model-specific dataset-adjusted sensitivity of fidelity to LR quality. (c) Pairwise win rates derived from the final human fidelity rankings for each LR image; each cell gives the percentage of LR images on which the row model is ranked more faithful than the column model.

Current generative RWSR methods instead occupy distinct operating points that reflect different balances between perceptual enhancement and preservation of content consistent with observable LR evidence. The coupling between the two quantities also varies substantially within individual models. The within-model SRCC between human perceptual quality and human fidelity ranges from 0.190 for PASD to 0.804 for LucidFlux. Perceptual-quality evaluation alone can therefore mask important differences in how strongly individual restoration models alter or synthesize content relative to the LR evidence.

Models Differ in Their Ability to Exploit Reliable LR Evidence. A global LR-quality–fidelity analysis can be confounded by systematic differences among the source datasets. To isolate within-domain behavior, we subtract the dataset-level mean LR-quality offset and, for each model–dataset combination, the corresponding mean fidelity offset before estimating their relationship. Fig. 6(b) groups the resulting dataset-adjusted LR quality into quintiles. Seven of the eight models show increasing fidelity as more reliable LR evidence becomes available, but the strength of this dependence varies markedly. SeeSR has the largest sensitivity ($\beta = 0.698$), followed by DiT4SR (0.509), StableSR (0.468), UPSR (0.434), CoSeR (0.406), PASD (0.334), and HYPIR (0.238). LucidFlux, by contrast, shows essentially no dataset-adjusted dependence on LR quality ($\beta = -0.020$). These trends indicate that RWSR models differ not only in average fidelity but also in their ability to translate clearer LR evidence into more faithful restoration. For most methods, better LR evidence improves fidelity, but the magnitude of this benefit is highly model dependent.

Table 6: Cross-domain human LR–SR fidelity for the eight diffusion-based SR models. Values in parentheses denote within-dataset fidelity ranks, where rank 1 corresponds to the highest mean fidelity. "Overall" is the mean over all 465 LR images, and "Avg. rank" is the average rank across the five source datasets.

Method	RealDeg	RealSRSet	RealSR	RealPhoto60	Real47	Overall	Avg. rank
PASD	**2.523 (1)**	2.778 (3)	2.413 (2)	**2.600 (1)**	2.359 (3)	**2.504**	**2.0**
UPSR	2.431 (2)	2.509 (5)	2.203 (3)	2.489 (3)	2.119 (6)	2.361	3.8
CoSeR	2.333 (3)	2.817 (2)	2.147 (4)	2.491 (2)	**2.441 (1)**	2.345	2.4
SeeSR	2.183 (5)	2.763 (4)	**2.447 (1)**	2.405 (4)	2.414 (2)	2.317	3.2
StableSR	2.233 (4)	2.428 (6)	2.145 (5)	2.144 (6)	2.317 (4)	2.220	5.0
HYPIR	1.943 (6)	**2.878 (1)**	1.944 (6)	2.145 (5)	2.240 (5)	2.040	4.6
DiT4SR	1.194 (7)	1.984 (8)	1.759 (7)	1.130 (8)	1.630 (7)	1.385	7.4
LucidFlux	0.799 (8)	2.305 (7)	1.093 (8)	2.088 (7)	1.526 (8)	1.167	7.6

Table 7: Low-fidelity tail statistics over the 465 LR images for each model. P_{10} is the model-specific 10th percentile. "Below T_{10}" uses the global 10th-percentile threshold $T_{10} = 0.954$, and "Below T_5" uses the global 5th-percentile threshold $T_5 = 0.652$. "Tail share" denotes the fraction of the 372 global-bottom-decile outputs contributed by each model.

Method	Mean fidelity	Model P_{10}	Below T_{10}	Tail rate	Tail share	Below T_5
PASD	**2.504**	**1.849**	4	**0.9%**	1.1%	0.0%
UPSR	2.361	1.736	4	**0.9%**	1.1%	0.4%
CoSeR	2.345	1.797	5	1.1%	1.3%	0.2%
SeeSR	2.317	1.576	8	1.7%	2.2%	0.6%
StableSR	2.220	1.583	7	1.5%	1.9%	0.6%
HYPIR	2.040	1.397	8	1.7%	2.2%	0.4%
DiT4SR	1.385	0.612	107	23.0%	28.8%	11.0%
LucidFlux	1.167	0.465	229	49.2%	61.6%	26.7%

Cross-Domain Fidelity Robustness. Table 6 reports human fidelity across the five source datasets together with within-dataset ranks. The rankings are substantially, though not perfectly, consistent across domains. Kendall's coefficient of concordance over the five dataset-specific rankings is $W = 0.741$ ($p = 5.2 \times 10^{-4}$). Across all dataset pairs, the mean Spearman rank correlation is 0.676 and the mean Kendall's τ is 0.529. PASD remains among the top three on all five datasets and achieves the best average rank (2.0); CoSeR is similarly consistent, with an average rank of 2.4. Other methods show stronger domain dependence. SeeSR ranks first on RealSR but fifth on RealDeg, while HYPIR ranks first on RealSRSet but sixth on both RealDeg and RealSR. In contrast, DiT4SR and LucidFlux remain near the bottom on almost every dataset. The results therefore point to both persistent model-level fidelity differences and meaningful interactions with content and degradation domain. These domain differences cannot be explained by LR quality alone. For example, RealDeg has a relatively high mean LR-quality MOS of 3.501 but a lower mean fidelity of 1.955, whereas Real47 has a lower mean LR-quality MOS of 2.544 but a higher mean fidelity of 2.131. This contrast suggests that fidelity is shaped by the type of real-world degradation and content distribution, not merely by the apparent cleanliness of the LR observation.

Low-Fidelity Tails Reveal Severe Failure Risk. Mean fidelity alone does not capture model reliability because severe fidelity failures can be concentrated in a relatively small subset of inputs. We therefore define a common low-fidelity threshold from the 10th percentile of all $3{,}720$ human fidelity scores, yielding $T_{10} = 0.954$, and additionally report the global 5th-percentile threshold $T_5 = 0.652$. Table 7 reveals large differences in the frequency of extreme low-fidelity outputs. PASD and UPSR each fall below the global 10th-percentile threshold on only 4 of 465 LR images (0.9%), and every PASD output remains above the global 5th-percentile threshold. By comparison, DiT4SR falls below T_{10} on 107 images (23.0%), and LucidFlux on 229 images (49.2%). Together, these two methods contribute 336 of the 372 outputs in the global bottom decile, accounting for 90.3% of all severe low-fidelity cases. The same qualitative pattern persists at the stricter 5th-percentile threshold. Model differences therefore extend well beyond mean fidelity and include sharply different risks of severe fidelity failure.

Table 8: Dataset-level analysis on FAIM-Bench. Fidelity denotes the human LR–SR fidelity score. PSNR, SSIM, LPIPS, and DISTS are computed between SR and the bicubic-upsampled LR image (ILR) and are therefore proxy-reference metrics rather than HR-reference measures.

Dataset	#LR	#SR	Fidelity↑	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	NIQE↓	MANIQA↑	CLIP-IQA↑	MUSIQ↑	Q-Align↑	TOPIQ-NR↑	LIQE↑
RealDeg	238	1904	1.955	28.361	**0.790**	0.517	0.263	4.246	0.334	0.542	40.574	4.164	0.454	2.320
RealSRSet	20	160	**2.558**	25.014	0.746	0.544	0.252	5.108	0.458	0.685	67.807	3.921	0.597	3.771
RealSR	100	800	2.019	24.740	0.741	**0.433**	**0.230**	4.447	0.430	0.620	64.093	4.041	0.569	3.479
RealPhoto60	60	480	2.186	**28.920**	0.766	0.553	0.367	4.988	0.359	0.551	49.135	3.362	0.468	2.697
Real47	47	376	2.131	25.600	0.744	0.489	0.246	**3.914**	**0.489**	**0.712**	**71.252**	**4.404**	**0.642**	**4.280**

Table 9: Correlation between NR-IQA metrics and human SR perceptual-quality judgments. The sign of NIQE is reversed before computing correlations so that higher values consistently indicate better predicted quality.

Metric	NIQE	MANIQA	CLIP-IQA	MUSIQ	Q-Align	TOPIQ-NR	LIQE
SRCC	0.196	0.177	0.271	0.191	**0.541**	0.281	0.346
PLCC	0.309	0.227	0.329	0.279	**0.648**	0.366	0.384

Pairwise Fidelity Dominance Across LR Images. Fig. 6(c) complements the mean-score analysis with scene-level pairwise dominance. For each LR image, we compare the final human fidelity ranks of every pair of SR models, yielding all $465 \times \binom{8}{2} = 13{,}020$ unique LR-conditioned model-pair comparisons. PASD achieves a majority win rate against every competing model, including 65.2% against UPSR, 66.7% against CoSeR, and 62.2% against SeeSR. By contrast, comparisons among UPSR, CoSeR, and SeeSR remain close to balanced, with win rates near 50%. Their relative performance is therefore strongly scene dependent despite differences in their average fidelity scores. The matrix also reveals a clear separation in the lower-fidelity regime. DiT4SR and LucidFlux rarely outrank the other six models, although DiT4SR still exceeds LucidFlux on 68.0% of LR images. Hence, the overall ordering is not driven solely by a small number of extreme cases; several major fidelity differences persist across a substantial fraction of the 465 LR observations.

C.2 Mismatch Between Existing Metrics and Human Fidelity Judgments

We next examine how commonly used IQA and proxy-reference metrics align with human judgments of LR–SR fidelity. Table 8 reports dataset-level human fidelity together with conventional metric scores. Because the primary FAIM-Bench protocol does not assume paired HR references, PSNR, SSIM, LPIPS, and DISTS are computed between the SR output and the bicubic-upsampled LR image (ILR); accordingly, they are interpreted as ILR–SR proxy-reference metrics rather than conventional HR-reference measures. Perceptual quality and fidelity are related but distinct evaluation dimensions. The perception–distortion trade-off suggests that improvements in perceptual realism can increase distortion from a reference signal Blau & Michaeli (2018), while recent analyses of generative restoration further associate aggressive perceptual enhancement with greater uncertainty and hallucination risk Cohen et al. (2024). The FAIM-Bench results reflect this distinction. Real47 obtains the strongest scores on several NR-IQA metrics without achieving the highest human fidelity, whereas RealSRSet has the highest dataset-level fidelity despite less competitive NR-IQA scores. The proxy-reference metrics show a similar mismatch: RealPhoto60 has the highest PSNR, RealDeg the highest SSIM, and RealSR the best LPIPS and DISTS, yet none has the highest human fidelity. Thus, neither output-only perceptual quality nor similarity to an ILR proxy can be treated as a substitute for human LR–SR fidelity.

Human Perceptual Quality and NR-IQA Metrics. To separate fidelity-specific limitations from a general inability to model perceptual quality, Table 9 compares NR-IQA scores directly with human SR perceptual-quality judgments. Because lower NIQE values indicate better quality, we reverse its sign before computing correlations. Several NR-IQA metrics correlate positively with human perceptual quality, with Q-Align showing the strongest agreement (SRCC=0.541, PLCC=0.648). Thus, output-only IQA metrics capture perceptual quality to a meaningful degree, even though they do not directly assess whether restored content remains compatible with the LR observation.

Table 10: Effect of LR quality on human SR perceptual-quality judgments. "SR-quality std." and "SR-quality range" are computed across the eight SR outputs of each LR image and then averaged within each LR-quality group.

LR quality group	#LR	#SR	Mean LR quality	Mean SR perceptual quality	SR-quality std.	SR-quality range
Low ≤ 2.5	80	640	2.311	3.209	0.468	1.387
Medium $(2.5, 3.5]$	231	1848	3.126	3.440	0.307	0.983
High > 3.5	154	1232	3.862	3.716	0.331	1.088

Table 11: Effect of LR quality on FAIM–human fidelity correlation on held-out test samples.

LR group	#LR	#SR	Mean fidelity	SRCC	PLCC
Low ≤ 2.5	48	384	2.094	0.481	0.539
Medium $(2.5, 3.5]$	136	1088	1.908	0.540	0.570
High > 3.5	95	760	2.187	0.636	0.722

C.3 Effect of LR Quality on Human SR Assessment

At the LR-image level, LR quality is positively correlated with the average human SR perceptual-quality score (SRCC=0.582, PLCC=0.582). As shown in Table 10, mean SR perceptual quality increases from 3.209 for low-quality LR inputs to 3.716 for high-quality inputs. Cleaner observations therefore tend to support perceptually stronger SR outputs. We also quantify the within-image variation across the eight SR models. The low-LR-quality group has the largest average standard deviation and range, indicating that the choice of restoration model has a stronger perceptual impact when the input is severely degraded. LR quality, however, is only weakly correlated with human LR–SR fidelity (SRCC=0.115, PLCC=0.083). A clean LR image therefore does not by itself guarantee faithful restoration, while a severely degraded input can still yield high fidelity when the recovered content remains consistent with the available evidence. This weak aggregate relationship motivates the model-specific, dataset-adjusted analysis in Fig. 6(b), which shows that individual SR models differ substantially in how much they benefit from improved LR evidence.

C.4 FAIM Performance under Different LR-Quality Levels

On the held-out test set, FAIM–human agreement improves with LR quality: SRCC increases from 0.481 in the low-quality group to 0.636 in the high-quality group, while PLCC increases from 0.539 to 0.722. Severely degraded LR observations provide weaker semantic and textural evidence, making input-conditioned fidelity assessment more ambiguous. Table 11 therefore reports the correlation statistics without relying on the absolute scale of the FAIM score.

C.5 Detailed Method-Level Results

Finally, Tables 12–16 provide detailed results for all eight diffusion-based SR models on each of the five source datasets. Fidelity denotes the human LR–SR fidelity score. Because the primary FAIM-Bench protocol does not assume paired HR references, PSNR, SSIM, LPIPS, and DISTS are computed between ILR and SR and should be interpreted as proxy-reference metrics. The detailed results reinforce the aggregate trends. PASD attains the highest human fidelity on RealDeg and RealPhoto60, HYPIR on RealSRSet, SeeSR on RealSR, and CoSeR on Real47. The methods favored by human fidelity frequently differ from those favored by ILR–SR proxy-reference or NR-IQA metrics. This recurring disagreement indicates that perceptual quality, proxy similarity, and LR-conditioned fidelity describe complementary aspects of generative RWSR behavior and should be evaluated separately rather than treated as interchangeable criteria.

Table 12: Method-level performance on RealDeg. Fidelity denotes the human LR–SR fidelity score. PSNR, SSIM, LPIPS, and DISTS are computed between SR and ILR and should be interpreted as proxy-reference metrics.

Method	Fidelity↑	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	NIQE↓	MANIQA↑	CLIP-IQA↑	MUSIQ↑	Q-Align↑	TOPIQ-NR↑	LIQE↑
UPSR	2.431	32.439	0.896	0.406	0.212	4.369	0.302	0.456	40.704	4.050	0.427	1.824
StableSR	2.233	29.564	0.801	0.583	0.263	3.322	0.334	0.589	40.877	4.096	0.487	2.416
SeeSR	2.183	30.017	0.859	0.450	0.272	4.311	0.370	0.609	**44.268**	4.129	**0.532**	2.699
CoSeR	2.333	30.970	0.884	0.428	0.230	4.175	0.312	0.476	38.211	4.162	0.412	2.237
PASD	**2.523**	**34.273**	**0.917**	**0.283**	**0.189**	6.072	0.283	0.392	34.279	4.031	0.336	1.721
DiT4SR	1.194	23.980	0.695	0.643	0.305	**3.147**	**0.391**	**0.649**	43.533	4.257	**0.532**	**3.439**
HYPIR	1.943	25.954	0.747	0.608	0.291	3.632	0.340	0.628	43.902	**4.313**	0.512	2.800
LucidFlux	0.799	19.695	0.525	0.736	0.339	4.942	0.339	0.540	38.815	4.277	0.396	1.427

Table 13: Method-level performance on RealSRSet. Fidelity denotes the human LR–SR fidelity score. PSNR, SSIM, LPIPS, and DISTS are computed between SR and ILR and should be interpreted as proxy-reference metrics.

Method	Fidelity↑	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	NIQE↓	MANIQA↑	CLIP-IQA↑	MUSIQ↑	Q-Align↑	TOPIQ-NR↑	LIQE↑
UPSR	2.509	26.783	0.788	0.522	**0.228**	**4.729**	0.412	0.661	66.979	3.799	0.588	3.601
StableSR	2.428	25.207	0.723	0.585	0.261	5.244	0.466	0.707	67.529	3.939	0.605	3.921
SeeSR	2.763	25.939	0.776	0.521	0.269	5.541	**0.536**	**0.739**	71.110	4.025	**0.680**	**4.219**
CoSeR	2.817	26.035	0.800	0.502	0.232	5.108	0.419	0.653	65.578	3.906	0.540	3.516
PASD	2.778	**27.887**	**0.811**	**0.441**	**0.228**	5.338	0.404	0.629	64.411	3.791	0.541	3.387
DiT4SR	1.984	23.845	0.693	0.604	0.260	4.745	0.418	0.687	67.027	3.844	0.563	3.731
HYPIR	**2.878**	23.152	0.715	0.576	0.265	5.164	0.487	0.712	67.901	4.018	0.620	3.803
LucidFlux	2.305	21.264	0.658	0.597	0.271	4.992	0.520	0.694	**71.920**	**4.047**	0.640	3.991

D Detailed Formulation of FAIM

D.1 Overview

Given a real-world LR image I_{LR} and its super-resolved image I_{SR}, FAIM evaluates the degree to which the restored content remains compatible with the visual evidence available in the LR observation. We first bicubic-upsample I_{LR} to the spatial resolution of I_{SR}, yielding the interpolated LR image I_{ILR}. Both images have spatial resolution $H \times W$, while their DINOv3 feature maps are defined on a spatial grid of size $h \times w$. A frozen DINOv3 backbone Siméoni et al. (2025) extracts multi-layer features from I_{ILR} and I_{SR}. Layer-Adaptive Feature Integration (LAFI) combines these features into spatially aligned representations Z^{ILR} and Z^{SR}. FAIM then evaluates their consistency through two complementary branches. The Spatial Consistency Branch (SCB) scores discrepancies at aligned feature locations and aggregates the strongest responses, while the Region Context Consistency Branch (RCCB) models region-level contextual inconsistency and weights it by the observed spatial discrepancy. Their inconsistency scores are combined and sign-reversed to obtain the Fidelity Score, while the inconsistency evidence maps are fused into the Fidelity Inconsistency Map.

D.2 Layer-Adaptive Feature Integration Module

Let $\{F_\ell^{\mathrm{ILR}}, F_\ell^{\mathrm{SR}}\}_{\ell\in\mathcal{L}}$ denote DINOv3 feature maps extracted from the selected layers, all defined on the $h \times w$ spatial grid. A single bias-free linear projection $P : \mathbb{R}^C \to \mathbb{R}^d$ is shared across all selected layers and between the ILR and SR streams. We define $\widehat{F}_\ell^X = \frac{P(F_\ell^X)}{\|P(F_\ell^X)\|_2}, \ell \in \mathcal{L}, X \in \{\mathrm{ILR}, \mathrm{SR}\}$. The projection parameters are shared between the ILR and SR streams so that the two sides remain directly comparable. Different DINOv3 layers encode complementary visual cues. We therefore assign each selected layer a learnable normalized weight α_ℓ satisfying $\alpha_\ell \geq 0, \sum_{\ell\in\mathcal{L}} \alpha_\ell = 1$. The layer features are scaled, concatenated, and normalized as

$$Z^X = \frac{[\sqrt{\alpha_\ell}\,\widehat{F}_\ell^X]_{\ell\in\mathcal{L}}}{\|[\sqrt{\alpha_\ell}\,\widehat{F}_\ell^X]_{\ell\in\mathcal{L}}\|_2}, \qquad X \in \{\mathrm{ILR}, \mathrm{SR}\}, \tag{8}$$

where $[\cdot]_{\ell\in\mathcal{L}}$ denotes concatenation along the feature dimension. Because each projected layer feature is unit-normalized, scaling by $\sqrt{\alpha_\ell}$ makes its contribution to the squared norm of the concatenated representation proportional to α_ℓ rather than α_ℓ^2. The resulting Z^{ILR} and Z^{SR} remain aligned on the same $h \times w$ feature grid and are shared by SCB and RCCB.

Table 14: Method-level performance on RealSR. Fidelity denotes the human LR–SR fidelity score. PSNR, SSIM, LPIPS, and DISTS are computed between SR and ILR and should be interpreted as proxy-reference metrics.

Method	Fidelity↑	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	NIQE↓	MANIQA↑	CLIP-IQA↑	MUSIQ↑	Q-Align↑	TOPIQ-NR↑	LIQE↑
UPSR	2.203	27.583	0.821	0.391	0.207	4.200	0.377	0.534	64.078	4.012	0.557	2.982
StableSR	2.145	24.663	0.739	0.472	0.240	4.318	0.436	0.645	62.853	4.127	0.590	3.695
SeeSR	**2.447**	27.181	0.812	0.393	0.238	4.872	0.491	0.659	65.576	3.958	**0.637**	3.593
CoSeR	2.147	25.631	0.789	0.403	0.207	4.439	0.420	0.628	64.337	4.146	0.553	3.701
PASD	2.413	**28.527**	**0.838**	**0.308**	**0.180**	4.986	0.378	0.523	58.699	4.000	0.497	2.829
DiT4SR	1.759	23.517	0.677	0.499	0.238	4.268	0.427	0.636	65.204	3.919	0.559	3.839
HYPIR	1.944	22.339	0.680	0.494	0.253	**3.981**	**0.492**	**0.699**	**66.965**	**4.186**	0.636	**4.128**
LucidFlux	1.093	18.475	0.572	0.502	0.274	4.512	0.417	0.633	65.036	3.979	0.525	3.067

Table 15: Method-level performance on RealPhoto60. Fidelity denotes the human LR–SR fidelity score. PSNR, SSIM, LPIPS, and DISTS are computed between SR and ILR and should be interpreted as proxy-reference metrics.

Method	Fidelity↑	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	NIQE↓	MANIQA↑	CLIP-IQA↑	MUSIQ↑	Q-Align↑	TOPIQ-NR↑	LIQE↑
UPSR	2.489	33.164	0.899	0.412	0.303	5.808	0.302	0.426	39.780	2.804	0.366	1.576
StableSR	2.144	29.851	0.752	0.655	0.372	4.015	0.338	0.566	46.760	3.003	0.493	2.699
SeeSR	2.405	29.196	0.770	0.588	0.405	4.518	0.434	0.669	56.149	3.488	0.586	3.244
CoSeR	2.491	34.582	0.909	0.374	0.303	5.639	0.282	0.362	32.580	2.555	0.325	1.663
PASD	**2.600**	**36.625**	**0.917**	**0.260**	**0.262**	8.675	0.306	0.373	31.769	2.694	0.300	1.741
DiT4SR	1.130	22.120	0.616	0.747	0.429	**3.348**	**0.459**	**0.734**	**65.594**	4.030	**0.617**	**4.322**
HYPIR	2.145	25.504	0.674	0.691	0.417	3.632	0.407	0.698	59.095	3.890	0.604	3.660
LucidFlux	2.088	20.321	0.596	0.701	0.445	4.267	0.343	0.582	61.354	**4.434**	0.455	2.672

D.3 Spatial Consistency Branch

SCB evaluates local LR–SR discrepancy at spatially corresponding feature locations. For location i, we define $\Delta_i^S = Z_i^{\mathrm{SR}} - Z_i^{\mathrm{ILR}}$. A directional residual would depend on the arbitrary ordering of the two inputs. We therefore use a symmetric discrepancy scorer $\mathcal{S}_\eta(\Delta) = \frac{1}{2}[s_\eta(\Delta) + s_\eta(-\Delta)]$, where s_η is a lightweight two-layer MLP with a non-negative Softplus output,

$$s_\eta(\Delta) = \mathrm{Softplus}(W_2\ \mathrm{GELU}(W_1\Delta + b_1) + b_2). \tag{9}$$

The spatial inconsistency evidence at location i is $E_i^S = \mathcal{S}_\eta(\Delta_i^S)$. Thus, $\{E_i^S\}_{i=1}^{hw}$ forms a spatial response map over the aligned feature grid. For image-level scoring, SCB focuses on the strongest discrepancies rather than averaging over all locations. Let $p \in (0, 1]$ denote the retained spatial proportion, and let $\mathcal{T}_p$ contain the indices of the largest $\lceil p\,hw \rceil$ responses in $\{E_i^S\}_{i=1}^{hw}$. The spatial inconsistency score is

$$E^S = \frac{1}{|\mathcal{T}_p|} \sum_{i \in \mathcal{T}_p} E_i^S. \tag{10}$$

This top-p aggregation prevents large regions with negligible discrepancy from dominating the branch score, while the full map E_i^S remains available for localization.

D.4 Region Context Consistency Branch

SCB compares aligned locations directly, whereas RCCB evaluates whether LR–SR differences remain compatible with broader region-level context. RCCB consists of region partitioning, region feature aggregation, bidirectional geometry-aware cross-attention, and spatially gated contextual contradiction estimation.

Region partitioning. Following INSID3 Cuttano et al. (2026), we construct regions independently for I_{ILR} and I_{SR} using training-free agglomerative clustering over high-level frozen DINOv3 features. For side $X \in \{\mathrm{ILR}, \mathrm{SR}\}$, let $\bar{F}^X_{\ell_p,i}$ denote the ℓ_2-normalized feature at location i from the partitioning layer ℓ_p. Pairwise feature dissimilarity is measured by cosine distance,

$$d_{ij}^X = 1 - (\bar{F}^X_{\ell_p,i})^\top \bar{F}^X_{\ell_p,j}. \tag{11}$$

For two candidate regions $\mathcal{A}$ and $\mathcal{B}$, average linkage defines

$$d^X(\mathcal{A}, \mathcal{B}) = \frac{1}{|\mathcal{A}|\,|\mathcal{B}|} \sum_{i \in \mathcal{A}} \sum_{j \in \mathcal{B}} d_{ij}^X. \tag{12}$$

Table 16: Method-level performance on Real47. Fidelity denotes the human LR–SR fidelity score. PSNR, SSIM, LPIPS, and DISTS are computed between SR and ILR and should be interpreted as proxy-reference metrics.

Method	Fidelity↑	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	NIQE↓	MANIQA↑	CLIP-IQA↑	MUSIQ↑	Q-Align↑	TOPIQ-NR↑	LIQE↑
UPSR	2.119	27.431	0.789	0.493	**0.218**	**3.452**	0.448	0.704	70.487	4.286	0.615	3.997
StableSR	2.317	26.512	0.756	0.502	0.253	4.339	0.534	0.750	71.392	**4.510**	0.677	4.456
SeeSR	2.414	27.434	0.787	0.460	0.255	3.992	0.529	0.709	72.492	4.432	**0.692**	4.434
CoSeR	**2.441**	26.400	0.793	0.452	0.221	3.988	0.447	0.687	69.069	4.398	0.585	4.049
PASD	2.359	**28.916**	**0.825**	**0.388**	**0.218**	4.011	0.433	0.642	68.548	4.384	0.602	4.049
DiT4SR	1.630	23.780	0.689	0.535	0.260	3.990	0.459	0.718	71.791	4.293	0.616	4.424
HYPIR	2.240	23.334	0.680	0.533	0.273	3.717	**0.540**	**0.752**	72.738	4.450	0.687	**4.508**
LucidFlux	1.526	20.994	0.636	0.550	0.269	3.821	0.519	0.738	**73.493**	4.480	0.660	4.325

Agglomerative merging under a clustering criterion produces a partition map $M^X \in \{1, \ldots, K_X\}^{h \times w}$, where K_X is the number of regions on side X. The two images are partitioned independently, allowing their region organizations to differ when the SR result changes local structure or grouping.

Region feature aggregation. For region k on side X, let $\Omega_k^X = \{i \mid M^X(i) = k\}$ denote the set of feature-grid locations belonging to the region. We pool the LAFI representation over the region and project it into the Region Token space $\bar{R}_k^X = W_r(\frac{1}{|\Omega_k^X|} \sum_{i \in \Omega_k^X} Z_i^X)$, followed by normalization, $R_k^X = \frac{\bar{R}_k^X}{\|\bar{R}_k^X\|_2}$. Collecting all region tokens gives $R^X \in \mathbb{R}^{K_X \times C_r}$. Each region is also associated with a geometry descriptor $g_k^X = [c_k^{x,X}, c_k^{y,X}, \log a_k^X]$, where $(c_k^{x,X}, c_k^{y,X})$ is the normalized region centroid and $a_k^X = \frac{|\Omega_k^X|}{hw}$ is its relative area. Stacking the descriptors gives $G^X \in \mathbb{R}^{K_X \times 3}$.

Geometry-aware cross-attention. RCCB models contextual interactions in both directions between the ILR and SR regions. For one attention direction, let Y_1 denote the query side and Y_2 the source side, with $Y_1, Y_2 \in \{\mathrm{ILR}, \mathrm{SR}\}$ and $Y_1 \neq Y_2$. For attention head m, the Region Tokens are projected into $Q_m^{Y_1} \in \mathbb{R}^{K_{Y_1} \times d}$, $K_m^{Y_2}, V_m^{Y_2} \in \mathbb{R}^{K_{Y_2} \times d}$, where $d = C_r/N_a$ and N_a is the number of attention heads. For query region u and source region v, we encode their relative geometry as $r_{uv}^{Y_1,Y_2} = [\Delta x_{uv}, \Delta y_{uv}, |\Delta x_{uv}|, |\Delta y_{uv}|, \Delta \log a_{uv}]$, with $\Delta x_{uv} = c_u^{x,Y_1} - c_v^{x,Y_2}$, $\Delta y_{uv} = c_u^{y,Y_1} - c_v^{y,Y_2}$, and $\Delta \log a_{uv} = \log a_u^{Y_1} - \log a_v^{Y_2}$. A learnable mapping $\psi : \mathbb{R}^5 \to \mathbb{R}^{N_a}$ converts this relation into a head-specific geometry bias, $B_{:,u,v}^{Y_1,Y_2} = \psi(r_{uv}^{Y_1,Y_2})$. The attention matrix for head m is then

$$A_m^{Y_1 \to Y_2} = \mathrm{softmax}\left(\frac{Q_m^{Y_1}(K_m^{Y_2})^\top}{\sqrt{d}} + B_m^{Y_1,Y_2}\right), \tag{13}$$

where the softmax is taken over source regions. The corresponding contextual response is $O_m^{Y_1 \to Y_2} = A_m^{Y_1 \to Y_2} V_m^{Y_2}$. Responses from all heads are concatenated and projected back to the Region Token dimension, $O^{Y_1 \to Y_2} = W_o[O_m^{Y_1 \to Y_2}]_{m=1}^{N_a}$. The query representation is updated through a residual connection, $\widetilde{R}^{Y_1} = R^{Y_1} + \gamma\, O^{Y_1 \to Y_2}$, where γ is a learnable residual scaling factor. Applying this operation in both directions, ILR $\to$ SR and SR $\to$ ILR, yields contextualized Region Tokens $\widetilde{R}^{\mathrm{ILR}}$ and $\widetilde{R}^{\mathrm{SR}}$.

Contextual contradiction and spatial discrepancy gating. The contextualized Region Tokens are broadcast back to their corresponding positions on the $h \times w$ feature grid, $\widehat{R}_i^X = \widetilde{R}_{M^X(i)}^X, X \in \{\mathrm{ILR}, \mathrm{SR}\}$. Because the broadcast representations share the same feature grid, their contextual residual is defined location-wise as $\Delta_i^R = \widehat{R}_i^{\mathrm{SR}} - \widehat{R}_i^{\mathrm{ILR}}$. An independent symmetric scorer produces the Contextual Contradiction response $C_i = \mathcal{S}_{\eta_R}(\Delta_i^R)$, where $\mathcal{S}_{\eta_R}$ has the same symmetric form as the SCB scorer but uses independent parameters. Contextual contradiction alone may reflect region-level interactions without sufficient discrepancy in the original ILR–SR observation. We therefore compute a parameter-free Spatial Discrepancy Gate directly from the frozen DINOv3 features,

$$D_i = \frac{1}{|\mathcal{L}|} \sum_{\ell \in \mathcal{L}} [1 - \cos(F_{\ell,i}^{\mathrm{ILR}}, F_{\ell,i}^{\mathrm{SR}})]. \tag{14}$$

The spatial Region Inconsistency Evidence is $E_i^R = D_i C_i$. The multiplicative gate attenuates contextual responses at locations with weak evidence of an ILR–SR discrepancy while preserving contextual

responses that coincide with stronger observed feature differences. The region inconsistency score is obtained by spatial averaging, $E^R = \frac{1}{hw}\sum_{i=1}^{hw} E_i^R$.

D.5 Fidelity Score Aggregation

SCB and RCCB produce the spatial and region inconsistency scores E^S and E^R, respectively. Their sum gives the image-level inconsistency, $E = E^S + E^R$. FAIM reverses its sign so that larger values consistently indicate stronger fidelity, $S_{\mathrm{FAIM}} = -(E^S + E^R)$. Thus, the image-level score depends on the top-p aggregation of SCB and the spatial mean of the gated RCCB evidence; it is not obtained by averaging the localization map described below.

For zero-shot localization, we retain the Spatial Inconsistency Evidence and Region Inconsistency Evidence from the two branches. The Fidelity Inconsistency Map is defined by their element-wise sum: $E_i^F = E_i^S + E_i^R, i = 1, \ldots, hw$. This map is an auxiliary spatial output used to visualize where the two branches detect LR–SR inconsistency. It does not alter the computation of S_{FAIM}.

D.6 Ranking-based Objective

FAIM is trained with pairwise RankNet supervision to preserve the relative ordering of human fidelity judgments. Let S_i and S_j denote the predicted fidelity scores of two samples and let $q_{ij} \in \{0, 1\}$ indicate their target ordering. The pairwise loss is

$$\ell_{\mathrm{rank}}(i, j) = -q_{ij} \log \sigma(\frac{S_i - S_j}{\tau_r}) - (1 - q_{ij}) \log[1 - \sigma(\frac{S_i - S_j}{\tau_r})], \tag{15}$$

where $\sigma(\cdot)$ is the sigmoid function and τ_r is a ranking temperature.

Same-image ranking. For each LR image, we compare SR candidates generated from that same observation. Let r_i denote the human within-image fidelity rank of candidate i, where a smaller rank indicates higher fidelity. For a pair with unequal ranks,

$$q_{ij}^{\mathrm{within}} = \mathbf{1}(r_i < r_j). \tag{16}$$

Pairs with identical ranks are ignored. If $\mathcal{P}_{\mathrm{within}}$ denotes the set of valid within-image pairs, the corresponding objective is

$$\mathcal{L}_{\mathrm{within}} = \frac{1}{|\mathcal{P}_{\mathrm{within}}|} \sum_{(i,j)\in\mathcal{P}_{\mathrm{within}}} \ell_{\mathrm{rank}}(i, j). \tag{17}$$

This supervision compares fidelity under the same LR evidence and therefore emphasizes relative differences among alternative restorations of a common observation.

Cross-image ranking. Cross-image ranking compares SR samples associated with different LR images from the same source dataset. Let y_i and y_j denote their aggregated human fidelity scores. To avoid constructing supervision from nearly tied scores, pairs satisfying $|y_i - y_j| \le \delta$ are excluded. For the remaining pairs,

$$q_{ij}^{\mathrm{cross}} = \mathbf{1}(y_i > y_j). \tag{18}$$

With $\mathcal{P}_{\mathrm{cross}}$ denoting the valid cross-image pairs,

$$\mathcal{L}_{\mathrm{cross}} = \frac{1}{|\mathcal{P}_{\mathrm{cross}}|} \sum_{(i,j)\in\mathcal{P}_{\mathrm{cross}}} \ell_{\mathrm{rank}}(i, j). \tag{19}$$

Restricting these comparisons to the same source dataset reduces the influence of dataset-dependent human-score scales while encouraging fidelity scores to remain comparable across image contents.

The first two training stages use both forms of ranking supervision,

$$\mathcal{L}_{\mathrm{I/II}} = \mathcal{L}_{\mathrm{within}} + \lambda_{\mathrm{cross}}\mathcal{L}_{\mathrm{cross}}, \tag{20}$$

whereas the final joint fine-tuning stage optimizes the full FAIM score using cross-image ranking,

$$\mathcal{L}_{\mathrm{III}} = \mathcal{L}_{\mathrm{cross}}. \tag{21}$$

The numerical optimization settings and concrete hyperparameter values are reported separately in Appendix E.

E IMPLEMENTATION DETAILS

We provide the implementation and optimization settings of FAIM. The DINOv3 backbone remains frozen throughout training, while LAFI, SCB, and the learnable components of RCCB are optimized according to the three-stage training strategy described below. The same architecture and optimization settings are used across evaluation protocols; only the composition of SR methods in the training and validation sets differs.

Feature extraction and integration. We use DINOv3-Large with a patch size of 16 as the frozen feature backbone and extract features from layers $\mathcal{L} = \{6, 12, 18, 24\}$. For LAFI, a single bias-free linear projection $P : \mathbb{R}^{1024} \to \mathbb{R}^{32}$ is shared across all selected DINOv3 layers and between the ILR and SR streams. Each projected feature is ℓ_2-normalized before multi-layer integration, and the concatenated representation is normalized again after applying the learnable layer weights. The layer weights are also shared between the ILR and SR streams.

Spatial Consistency Branch. The symmetric discrepancy scorer in SCB is implemented as a two-layer MLP with a hidden dimension of 32 and a Softplus output. SCB retains the top-10% spatial inconsistency responses, corresponding to $p = 0.1$, and averages them to obtain the branch-level spatial inconsistency evidence E^S.

Region Context Consistency Branch. Region partitioning is performed independently for ILR and SR using the frozen layer-24 DINOv3 features. We use average-linkage agglomerative clustering based on cosine distance and parameterize the hierarchy cut by a cosine-similarity threshold τ. Starting from $\tau = 0.88$, we increase the threshold in steps of 0.01 up to 0.98 and select the first threshold that produces at least six regions. This procedure is applied independently to the ILR and SR images and introduces no learnable parameters.

Features within each region are mean-pooled and projected to Region Tokens of dimension $C_r = 32$. Geometry-aware cross-attention uses $N_a = 2$ attention heads. For each query–source region pair, the pairwise relative geometry vector is five-dimensional, consisting of the signed centroid offsets, their absolute values, and the relative log-area. The corresponding geometry bias is predicted by a two-layer MLP with a hidden dimension of 16. The Contextual Contradiction scorer is implemented as a symmetric two-layer MLP with a hidden dimension of 16 and a Softplus output. The Spatial Discrepancy Gate is parameter-free and is computed from cosine discrepancies of the frozen DINOv3 features across the four selected layers. The final RCCB uses bidirectional geometry-aware cross-attention and combines the contextual contradiction response C_i with the Spatial Discrepancy Gate D_i as $E_i^R = D_i C_i$. The resulting region inconsistency evidence is spatially averaged as $E^R = \frac{1}{hw} \sum_{i=1}^{hw} E_i^R$.

Ranking supervision. FAIM is optimized using RankNet objectives constructed from human fidelity annotations. Same-image ranking compares different SR outputs generated from the same LR image, whereas cross-image ranking compares samples from different LR images within the same source dataset. For same-image ranking, pairs assigned the same human rank are excluded. For cross-image ranking, pairs whose human fidelity scores differ by at most $\delta = 0.05$ are excluded. We set the RankNet temperature to $\tau_r = 1$. When same-image and cross-image ranking losses are jointly used, they are equally weighted, corresponding to $\lambda_{\text{cross}} = 1$.

Training protocol. Training proceeds in three stages. In Stage I, LAFI and RCCB are optimized while SCB is not used. The branch score $S^R = -E^R$ is trained using both same-image and cross-image ranking losses. We use a learning rate of 4×10^{-4} and train for up to 40 epochs.

In Stage II, LAFI and RCCB are frozen and only SCB is optimized. The branch score $S^S = -E^S$ is trained using the same two ranking objectives. The learning rate remains 4×10^{-4}, and training proceeds for up to 60 epochs.

In Stage III, all learnable components are jointly fine-tuned using the complete FAIM score

$$S_{\text{FAIM}} = -\left(E^S + E^R\right). \tag{22}$$

Only cross-image ranking is used in this stage. The learning rate is reduced to 2×10^{-4}, and joint fine-tuning proceeds for up to 40 epochs.

All stages use AdamW with a weight decay of 1×10^{-4} and gradient clipping with a maximum norm of 5.0. Each training batch contains four distinct LR images together with their corresponding SR

outputs, enabling both same-image and cross-image ranking pairs to be constructed within a batch. For the main leave-one-SR-method-out evaluation, we train eight fold-specific FAIM models, each excluding one SR method entirely from training and validation. For the comprehensive ablations in Appendix G.4, we instead use the known-method LR-held-out setting, where all eight SR methods are represented during training and validation while the test LR images remain unseen.

F Evaluation Protocol Details

Correlation measures. We evaluate all metrics using the Spearman rank-order correlation coefficient (SRCC) and the Pearson linear correlation coefficient after logistic mapping (PLCC-log) Sheikh et al. (2006). SRCC measures the rank-order agreement between metric scores and human fidelity scores:

$$\mathrm{SRCC} = \mathrm{Pearson}(R(\mathbf{m}), R(\mathbf{y})), \tag{23}$$

where $\mathbf{m}$ denotes metric scores, $\mathbf{y}$ denotes human fidelity scores, and $R(\cdot)$ denotes the ranking operation. PLCC-log first applies a five-parameter logistic mapping to calibrate metric scores to the human-score scale:

$$f(m) = \beta_2 + \frac{\beta_1 - \beta_2}{1 + \exp(-\frac{m-\beta_3}{|\beta_4|+\epsilon})} + \beta_5 m, \tag{24}$$

where $\{\beta_i\}_{i=1}^{5}$ are fitted by least squares within each evaluation subset and $\epsilon = 10^{-12}$ is a small constant for numerical stability. Under the leave-one-SR-method-out protocol, the logistic mapping and correlation are computed independently for each held-out method within each dataset before macro-averaging. Under the known-method LR-held-out setting, they are computed on the pooled outputs of all eight SR methods within each dataset. We then compute the Pearson correlation between $f(\mathbf{m})$ and $\mathbf{y}$. We use PLCC-log rather than raw PLCC because different metrics operate on heterogeneous and nonlinear score scales, making direct linear correlation less comparable across metrics.

Evaluation settings. The main experiments use a leave-one-SR-method-out protocol. For each of the eight folds, one SR method is excluded entirely from FAIM training and validation and evaluated on the fixed test LR split. Metric correlations are computed separately for each held-out SR method on each test dataset and then macro-averaged over the eight methods. The overall result is obtained by averaging across RealDeg-test, RealSR, and RealPhoto60. Baseline metrics are evaluated on the same held-out-method subsets and aggregated in the same manner.

The comprehensive appendix ablations use a known-method LR-held-out setting, where all eight SR methods are represented during training and validation while the test LR images remain unseen. In this setting, correlations are computed by pooling the outputs of all eight SR methods within each test dataset.

G Additional Experimental Analyses

G.1 Known-Method LR-Held-Out Evaluation

As a complementary evaluation, we consider a known-method LR-held-out setting, where all eight SR methods are represented during training and validation, while the test LR images remain unseen. Unlike the main leave-one-SR-method-out evaluation, correlations are computed by pooling the outputs of all eight SR methods within each test dataset.

Tab. 17 compares FAIM with existing metrics on RealDeg-test, RealSR, and RealPhoto60. The seven NR-IQA metrics show poor agreement with human fidelity judgments. Six yield negative average SRCC, while Q-Align is only weakly positive at 0.033. On RealPhoto60, all seven NR-IQA metrics have negative SRCC. These results show that perceptual quality assessed from the SR output alone does not reliably reflect LR–SR fidelity.

ILR–SR proxy-reference metrics perform substantially better, showing that incorporating the LR observation is important for fidelity assessment. However, direct ILR–SR similarity should not be treated as fidelity itself. The human annotation statistics in Fig. 3(b) further illustrate this distinction: RealDeg has relatively high mean LR quality (3.501) and SR perceptual quality (3.556), but the lowest mean fidelity (1.955), whereas RealPhoto60 has the lowest mean LR quality (2.473) and

Table 17: Comparison with existing metrics under the known-method LR-held-out setting. All eight SR methods are represented during training and validation, while test LR images remain unseen. Correlations are computed by pooling outputs from all eight SR methods within each dataset. PSNR, SSIM, LPIPS, and DISTS use ILR as the proxy reference.

Metric	RealDeg-test		RealSR		RealPhoto60		Average	
	SRCC↑	PLCC-log↑	SRCC↑	PLCC-log↑	SRCC↑	PLCC-log↑	SRCC↑	PLCC-log↑
PSNR↑	0.651	0.719	0.474	0.541	0.518	0.531	0.547	0.597
SSIM↑	0.638	0.697	0.440	0.464	0.512	0.528	0.530	0.563
LPIPS↓	0.610	0.649	0.356	0.385	0.505	0.520	0.491	0.518
DISTS↓	0.672	0.686	0.242	0.283	0.511	0.542	0.475	0.504
NIQE↓	-0.157	0.197	-0.181	0.200	-0.433	0.462	-0.257	0.286
MANIQA↑	-0.190	0.246	0.046	0.173	-0.194	0.270	-0.113	0.230
CLIP-IQA↑	-0.209	0.304	-0.021	0.086	-0.366	0.448	-0.199	0.279
MUSIQ↑	-0.046	0.139	-0.147	0.207	-0.375	0.428	-0.189	0.258
Q-Align↑	0.105	0.118	0.160	0.188	-0.167	0.344	0.033	0.216
TOPIQ-NR↑	-0.120	0.257	0.051	0.166	-0.357	0.514	-0.142	0.313
LIQE↑	-0.052	0.138	0.037	0.114	-0.377	0.429	-0.131	0.227
FAIM↑	**0.798**	**0.852**	**0.661**	**0.672**	**0.698**	**0.774**	**0.719**	**0.766**

Table 18: Method-level metric scores on RealSR, together with the corresponding HR-reference scores. For PSNR, SSIM, LPIPS, DISTS, and FAIM, the ILR is used as the reference/input counterpart. The remaining no-reference IQA metrics are directly evaluated on the SR or HR image.

Metric	UPSR	StableSR	SeeSR	CoSeR	PASD	DiT4SR	HYPIR	LucidFlux	HR
PSNR↑	27.583	24.663	27.181	25.631	28.527	23.517	22.339	18.475	**29.362**
SSIM↑	0.821	0.739	0.812	0.789	0.838	0.677	0.680	0.572	**0.843**
LPIPS↓	0.391	0.472	0.393	0.403	0.308	0.499	0.494	0.502	**0.251**
DISTS↓	0.207	0.240	0.238	0.207	0.180	0.238	0.253	0.274	**0.147**
NIQE↓	4.200	4.318	4.872	4.439	4.986	4.268	**3.981**	4.512	5.091
MANIQA↑	0.377	0.436	0.491	0.420	0.378	0.427	**0.492**	0.417	0.327
CLIP-IQA↑	0.534	0.645	0.659	0.628	0.523	0.636	**0.699**	0.633	0.489
MUSIQ↑	64.078	62.853	65.576	64.337	58.699	65.204	**66.965**	65.036	58.227
Q-Align↑	4.012	4.127	3.958	4.146	4.000	3.919	**4.186**	3.979	3.997
TOPIQ-NR↑	0.557	0.590	**0.637**	0.553	0.497	0.559	0.636	0.525	0.408
LIQE↑	2.982	3.695	3.593	3.701	2.829	3.839	**4.128**	3.067	2.596
FAIM↑ **(Ours)**	-4.932	-4.991	-3.871	-4.912	-3.745	-7.285	-5.084	-7.603	**-2.572**

SR perceptual quality (3.099), but a higher mean fidelity (2.186). Thus, LR quality, SR perceptual quality, and LR–SR fidelity do not follow the same ordering. Proxy-reference metrics also vary substantially across the three test sets; for example, the SRCC of DISTS decreases from 0.672 on RealDeg-test to 0.242 on RealSR, while that of LPIPS decreases from 0.610 to 0.356.

FAIM achieves stronger agreement with human fidelity judgments across all three test sets, reaching an average SRCC of 0.719 and PLCC-log of 0.766. Compared with the strongest baseline averages, FAIM improves SRCC from 0.547 to 0.719 and PLCC-log from 0.597 to 0.766, corresponding to absolute gains of 0.172 and 0.169, respectively. These results further support the benefit of modeling the compatibility between restored content and observable LR evidence rather than relying on output perceptual quality or direct ILR–SR similarity.

The correlations in this setting are higher than those under the main unseen-method evaluation. This is expected because all SR methods are observed during training and because pooling outputs across methods can capture systematic between-method fidelity differences. We therefore report this setting as a complementary analysis, while using the leave-one-SR-method-out setting as the primary evaluation of cross-method generalization.

G.2 Fidelity Assessment on Ground-Truth HR Images

A useful sanity check for fidelity assessment is how a metric evaluates the paired ground-truth HR image when it is available. Although FAIM is designed for real-world SR settings without paired HR

references, RealSR provides paired LR–HR images, enabling this additional analysis. This setting is particularly informative because the ground-truth HR image represents a valid high-resolution realization of the observed LR input while containing genuine structures and textures that may not be explicitly recoverable from the LR image itself. We therefore treat each ground-truth HR image as a valid realization corresponding to its LR observation and evaluate it using the same set of metrics as the SR outputs. For PSNR, SSIM, LPIPS, and DISTS, the HR image is compared with the bicubic-upsampled LR image, following the same ILR-based proxy-reference protocol. No-reference IQA metrics directly evaluate the HR image itself, while FAIM measures its consistency with the corresponding LR observation. Table 18 also reports the average scores of the eight SR methods on RealSR to provide a common reference scale for comparison.

FAIM assigns the ground-truth HR images the highest average score of -2.572, exceeding all eight SR methods. This provides a useful sanity check for the proposed notion of fidelity. The paired HR images contain genuine high-frequency structures and textures that may be weakly represented or not explicitly resolvable in the LR input, yet they remain valid realizations of the same underlying scene. Their high FAIM scores therefore suggest that FAIM does not simply penalize details because they are absent or unverifiable in the LR observation. Instead, it can assign high fidelity to detailed high-resolution content as long as that content remains compatible with the visual evidence preserved in the LR input. This behavior is consistent with our definition of fidelity, under which ambiguity in the LR observation allows multiple plausible HR realizations rather than requiring strict similarity to the degraded input.

The ground-truth HR images also achieve the best scores on all four ILR-based proxy-reference metrics, indicating that similarity to the ILR does contain useful information about LR–HR consistency. However, such similarity should not be interpreted as fidelity itself, because a faithful high-resolution reconstruction may legitimately differ from the degraded ILR appearance through the removal of blur, noise, compression artifacts, and other degradations. In this sense, proxy-reference similarity captures only part of the relationship between the LR observation and a valid HR realization. The no-reference IQA results provide a complementary view of perceptual quality: several SR methods obtain higher scores than the ground-truth HR images on some NR-IQA metrics, reflecting differences in perceived sharpness, naturalness, or enhancement rather than input consistency. Together, these observations further illustrate the distinction among proxy-reference similarity, output perceptual quality, and input-conditioned fidelity, which capture related but non-equivalent aspects of SR evaluation.

G.3 Zero-Shot Artifact Localization on SR-Prominence

We further investigate whether the spatial inconsistency responses produced by FAIM can provide zero-shot localization cues for perceptually salient super-resolution artifacts on the SR-Prominence benchmark Molodetskikh et al. (2026). In this context, "artifact localization" refers to identifying fidelity-related SR artifacts through FAIM's LR–SR inconsistency responses, rather than treating FAIM as a dedicated or supervised artifact detector. We conduct this evaluation in a strictly zero-shot manner, using the pretrained FAIM model without any adaptation or parameter tuning on SR-Prominence. Experiments are performed on the official OpenImages and Urban100 evaluation subsets. Given an LR–SR pair, FAIM produces spatial responses from the Spatial Consistency Branch (SCB) and the Region Context Consistency Branch (RCCB). Specifically, we denote the SCB response by s_i, the discrepancy-gated contextual response by $d_i c_i$, and their combined fidelity-inconsistency response by

$$H_i^{\mathrm{SCB}} = s_i, \qquad H_i^{\mathrm{RCCB}} = d_i c_i, \qquad H_i^{\mathrm{Full}} = s_i + d_i c_i. \tag{25}$$

Importantly, FAIM requires only the interpolated LR observation (ILR) and the corresponding SR output, where ILR is obtained by bicubic upsampling of the observed LR image and introduces no additional reference information. FAIM does not access either an HR reference or any surrogate reference image. We compare its spatial responses with the artifact-localization baselines provided by SR-Prominence, including `bd_jup` Molodetskikh et al. (2026), a block-wise distortion map that combines LPIPS- and ERQA-based discrepancy cues, DeSRA Xie et al. (2023), DISTS Ding et al. (2022), and the learned SR-Prominence baseline Molodetskikh et al. (2026). The compared methods are evaluated under different reference settings. On OpenImages, the released baseline heatmaps are computed using an RLFN reconstruction as a pseudo-reference, i.e., an SR image generated by RLFN is used as a surrogate for the unavailable HR ground truth. On Urban100, the baselines instead use the true HR image as reference. In contrast, FAIM uses only ILR and SR in both datasets, without

Table 19: Artifact-prominence agreement on SR-Prominence. We report the official contrast-$p50$ Spearman correlation with human artifact prominence. FAIM is evaluated strictly zero-shot, without training or adaptation on SR-Prominence, and uses only ILR and SR. In contrast, the released comparison baselines use pseudo-reference information on OpenImages and true HR references on Urban100; SR-Prominence NN is additionally a learned artifact-localization model.

Method	Reference	OpenImages	Urban100	Avg.
`bd_jup`	Pseudo-GT / HR	0.202	0.236	0.219
DeSRA	Pseudo-GT / HR	0.379	0.019	0.199
DISTS	Pseudo-GT / HR	0.384	0.236	0.310
SR-Prominence NN	Pseudo-GT / HR	0.497	0.256	0.376
FAIM-SCB	ILR + SR	0.409	0.295	0.352
FAIM-RCCB	ILR + SR	**0.618**	0.358	**0.488**
FAIM-Full	ILR + SR	0.560	**0.359**	0.459

relying on either true or pseudo HR references. It is also important to distinguish the supervision settings: the SR-Prominence NN is a learned artifact-localization model, whereas FAIM is transferred to SR-Prominence strictly zero-shot, without any training or adaptation on this benchmark. Thus, the evaluation examines whether spatial responses learned solely for LR–SR fidelity assessment can transfer to artifact localization under a weaker reference-information setting.

Following the official SR-Prominence evaluation protocol, we measure the response contrast between annotated artifact regions and the background and compute its Spearman correlation with human perceptual prominence, denoted as ***Prominence SRCC***. We additionally report pixel-level AUROC and AUPRC together with the proportion of annotations for which the artifact region receives a higher median response than the background,

$$\Delta_{50} = \mathrm{median}_{i \in M} H_i - \mathrm{median}_{i \notin M} H_i, \tag{26}$$

where M denotes the human-annotated artifact region. AUROC and AUPRC quantify the pixel-level agreement between each spatial response and the annotated artifact regions, while Prominence SRCC evaluates whether the magnitude of the predicted fidelity-inconsistency response agrees with the perceptual severity of the artifact.

As shown in Table 19, FAIM exhibits strong agreement with human perceptual judgments of artifact prominence despite being transferred to SR-Prominence without artifact-localization supervision. On OpenImages, FAIM-RCCB achieves a Prominence SRCC of 0.618, exceeding the learned SR-Prominence baseline at 0.497, while the full FAIM response reaches 0.560. Notably, these results are obtained without any training or adaptation on SR-Prominence and without access to the pseudo-reference used by the comparison baselines. The same trend is observed on Urban100, where RCCB and Full obtain 0.358 and 0.359, respectively, compared with 0.256 for the learned baseline. In this case, FAIM still uses only ILR and SR, whereas the comparison baselines are evaluated with access to the true HR image. Averaged across the two datasets, RCCB achieves 0.488 and Full achieves 0.459, compared with 0.376 for the strongest comparison method. These results indicate that the magnitude of FAIM's spatial inconsistency response is closely aligned with how strongly humans perceive a localized artifact, even under a strictly weaker reference-information setting. The particularly strong performance of RCCB further supports the role of discrepancy-gated contextual contradiction: a local change receives a strong response when it is simultaneously inconsistent with the observed LR content and incompatible with its regional context.

The pixel-level results in Table 20 reveal a complementary behavior. Dedicated artifact-localization baselines, particularly the learned SR-Prominence model and DISTS, achieve higher AUROC and AUPRC, indicating more accurate overlap with the precise spatial extent of the human annotations. This comparison should also be interpreted in light of the different supervision and reference settings: the released comparison baselines are evaluated with pseudo-HR references on OpenImages or true HR references on Urban100, and SR-Prominence NN is specifically learned for artifact localization, whereas FAIM is transferred zero-shot and operates only on ILR–SR consistency. FAIM instead shows a stronger correspondence between response magnitude and perceptual prominence. This distinction is consistent with its objective: the spatial response of FAIM is not trained as an artifact segmentation mask, but arises as an explanatory decomposition of fidelity inconsistency. Nevertheless,

Table 20: Pixel-level artifact localization on SR-Prominence. AUPRC is particularly informative because artifact regions generally occupy only a small fraction of the image. These metrics evaluate spatial overlap and are complementary to the perceptual-prominence agreement reported in Table 19. FAIM is evaluated zero-shot using only ILR and SR, while the comparison baselines use pseudo-HR references on OpenImages and true HR references on Urban100.

	OpenImages			Urban100		
Method	AUPRC ↑	AUROC ↑	$P(\Delta_{50} > 0)$ ↑	AUPRC ↑	AUROC ↑	$P(\Delta_{50} > 0)$ ↑
`bd_jup`	0.569	0.781	**0.961**	0.537	0.743	**0.978**
DeSRA	0.567	0.711	0.667	0.469	0.632	0.541
DISTS	0.631	0.765	0.933	0.573	0.753	0.976
SR-Prominence NN	**0.677**	**0.807**	0.933	**0.594**	**0.769**	0.971
FAIM-SCB	0.483	0.666	0.846	0.394	0.616	0.806
FAIM-RCCB	0.569	0.739	0.906	0.411	0.647	0.836
FAIM-Full	0.536	0.720	0.885	0.413	0.647	0.838

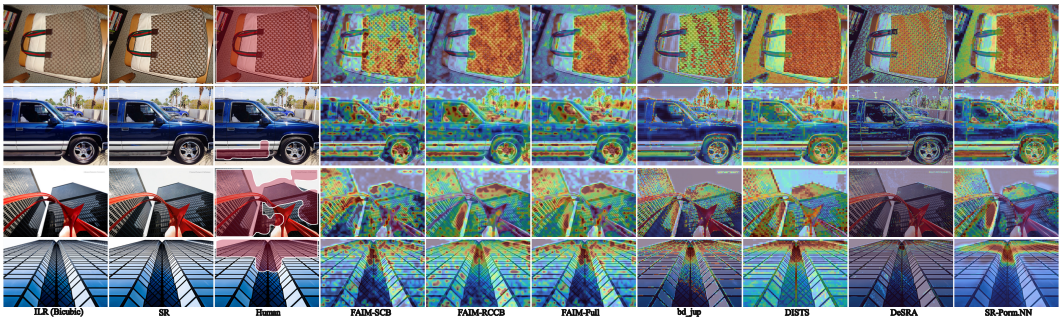


Figure 7: Qualitative artifact localization on SR-Prominence. FAIM is applied strictly zero-shot using only ILR and SR, without access to pseudo-HR or true HR reference images.

the artifact regions receive a higher response than the background in a large majority of cases, with $P(\Delta_{50} > 0)$ reaching 0.906 on OpenImages for RCCB and 0.838 on Urban100 for Full. The branch-level comparison further shows that RCCB is substantially more informative than SCB on OpenImages, whereas RCCB and Full perform similarly on Urban100. This suggests that contextual contradiction constitutes the principal source of artifact-sensitive localization, while local spatial inconsistency provides complementary evidence.

Figure 7 provides qualitative comparisons between FAIM, human artifact annotations, and the SR-Prominence baselines. Across both datasets, FAIM responses tend to concentrate on perceptually inconsistent regions identified by human annotators, while RCCB often produces more spatially selective responses than SCB. The examples also illustrate the distinction observed quantitatively: comparison methods can yield sharper agreement with artifact boundaries, whereas FAIM emphasizes regions whose inconsistency is perceptually salient relative to the surrounding image content. Together, these results demonstrate a favorable zero-shot transfer property of FAIM. Although FAIM is trained only for LR–SR fidelity assessment rather than artifact localization, its spatial responses provide meaningful artifact-localization cues on SR-Prominence without any dataset-specific adaptation and without access to the pseudo-HR or true HR references available to the comparison baselines.

G.4 Comprehensive Ablation Study

We complement the main leave-one-SR-method-out ablation with more extensive design analyses under the known-method LR-held-out setting. In this setting, all eight SR methods are represented during training and validation, while the test LR images remain unseen. Correlations are computed by pooling outputs from all eight SR methods within each test dataset. We use this setting for controlled analysis of a larger number of architectural and training choices; therefore, the absolute values reported in this section are not directly comparable to those in the main leave-one-SR-method-out

Table 21: Core architectural ablations of FAIM, including branch contribution, region-evidence decomposition, LAFI sharing, SCB–RCCB evidence fusion, and geometry-aware region reasoning.

Variant	RealDeg-test		RealSR		RealPhoto60		Average	
	SRCC↑	PLCC-log↑	SRCC↑	PLCC-log↑	SRCC↑	PLCC-log↑	SRCC↑	PLCC-log↑
SCB only	0.808	0.863	0.626	0.647	0.664	0.729	0.699	0.746
SCB only + mean spatial aggregation	0.779	0.835	0.583	0.581	0.668	0.734	0.677	0.716
RCCB only	0.765	0.811	0.610	0.622	0.675	0.764	0.684	0.732
w/o Contextual Contradiction	0.799	0.855	0.634	0.632	0.674	0.744	0.702	0.744
w/o Spatial Discrepancy Gate	0.750	0.794	0.545	0.582	0.680	0.764	0.658	0.713
Shared LAFI, Stage II frozen (FAIM)	0.798	0.852	**0.661**	**0.672**	**0.698**	**0.774**	**0.719**	**0.766**
Shared LAFI, Stage II trainable	0.808	**0.869**	0.608	0.629	0.655	0.736	0.690	0.745
Separate SCB/RCCB LAFI	**0.809**	0.864	0.624	0.653	0.669	0.729	0.701	0.749
Learned SCB–RCCB evidence weighting	0.790	0.843	0.618	0.632	0.681	0.748	0.697	0.741
Full FAIM w/o geometry bias	0.788	0.833	0.607	0.621	0.669	0.734	0.688	0.730
Full FAIM: Top-10% SCB aggregation + geometry	0.798	0.852	**0.661**	**0.672**	**0.698**	**0.774**	**0.719**	**0.766**

ablation. Unless stated otherwise, each ablation changes one design factor relative to the default FAIM configuration. Model selection at each trainable stage is based on SRCC on the validation split. Computational cost is summarized in Tab. 29.

Overall architecture and evidence decomposition. SCB and RCCB achieve average SRCC/PLCC-log of $0.699/0.746$ and $0.684/0.732$, respectively, while combining them improves the average to $0.719/0.766$. The gains are most evident on RealSR and RealPhoto60, indicating that the two branches provide complementary fidelity evidence. Replacing Top-10% aggregation in the SCB-only control with spatial mean aggregation reduces the average to $0.677/0.716$, showing that selective aggregation better captures informative spatial discrepancies than uniform averaging.

We further isolate the region-level components while retaining SCB. Removing Contextual Contradiction reduces the average SRCC from 0.719 to 0.702, removing geometry reduces it to 0.688, and removing the Spatial Discrepancy Gate reduces it to 0.658. These results show that contextual comparison, explicit geometry, and discrepancy-aware gating all contribute to region-level fidelity modeling, with the largest degradation caused by removing the gate.

The LAFI controls favor a shared feature space. Keeping the shared LAFI frozen in Stage II gives $0.719/0.766$ on average, compared with $0.690/0.745$ when it remains trainable and $0.701/0.749$ with separate SCB/RCCB LAFI. The separate design also increases the FAIM-specific parameter count from 0.046M to 0.079M and model size from 0.175MB to 0.300MB (Tab. 29). A learned scalar for SCB–RCCB fusion similarly reduces the average to $0.697/0.741$. These results support sharing LAFI across the two branches and using the fixed evidence sum for their final fusion.

LAFI: projection dimensionality. The projection sweep separates the effect of learned adaptation from dimensionality reduction. At 1024 dimensions, replacing the identity mapping with a learned projection improves the average SRCC/PLCC-log from $0.688/0.737$ to $0.712/0.765$, showing that feature adaptation is beneficial even without compression. Performance is otherwise non-monotonic: 256-D reaches $0.715/0.765$, while both 32-D and 8-D achieve an average SRCC of 0.719. The 32-D setting gives the highest average PLCC-log in this sweep (0.766) and performs better on RealSR than 8-D ($0.661/0.672$ versus $0.645/0.661$). Very low-dimensional projections ($d \leq 4$) consistently underperform the 8-D and 32-D settings, indicating that excessive compression removes useful fidelity information.

The computational comparison further supports this choice. The learned 1024-D projection increases the FAIM-specific parameter count from 0.046M to 1.316M and model size from 0.175MB to 5.019MB, while also increasing GMACs, without improving the average correlations (Tab. 29). The 8-D variant is more compact, but is weaker on RealSR and has slightly lower average PLCC-log. We therefore use 32-D as the default because it provides a better balance between accuracy and model complexity.

LAFI: DINOv3 depth, normalization, and weighting. Using only the final DINOv3 layer reduces the average SRCC/PLCC-log to $0.564/0.616$, showing that the deepest representation alone is insufficient for LR–SR fidelity assessment. Among the four-layer settings, $\{4, 10, 16, 22\}$ reaches $0.718/0.770$, while the default $\{6, 12, 18, 24\}$ reaches $0.719/0.766$ and performs better on RealSR. Increasing the number of selected layers does not lead to monotonic improvements: the 16-layer variant reaches the highest average of $0.728/0.780$, whereas using all 24 layers gives $0.718/0.763$.

Table 22: Effect of projection dimensionality in LAFI.

Variant	RealDeg-test		RealSR		RealPhoto60		Average	
	SRCC↑	PLCC-log↑	SRCC↑	PLCC-log↑	SRCC↑	PLCC-log↑	SRCC↑	PLCC-log↑
Identity 1024-D	0.805	0.864	0.562	0.585	0.696	0.763	0.688	0.737
Learned 1024-D	**0.820**	**0.878**	0.622	0.658	0.694	0.757	0.712	0.765
512-D	0.793	0.839	0.636	0.650	0.685	0.768	0.705	0.752
256-D	0.809	0.867	0.637	0.661	0.700	0.767	0.715	0.765
128-D	0.804	0.856	0.644	0.667	0.678	0.764	0.709	0.762
64-D	0.759	0.825	0.567	0.598	0.671	0.748	0.666	0.724
32-D (FAIM)	0.798	0.852	**0.661**	**0.672**	0.698	0.774	**0.719**	**0.766**
16-D	0.797	0.847	0.633	0.640	0.687	0.760	0.706	0.749
8-D	0.802	0.854	0.645	0.661	**0.710**	**0.779**	**0.719**	0.765
4-D	0.735	0.775	0.563	0.593	0.693	0.775	0.664	0.714
2-D	0.726	0.773	0.547	0.583	0.679	0.745	0.651	0.700
1-D	0.758	0.807	0.544	0.578	0.675	0.759	0.659	0.715

Table 23: Ablation of DINOv3 layer selection, normalization, and layer weighting in LAFI.

Variant	RealDeg-test		RealSR		RealPhoto60		Average	
	SRCC↑	PLCC-log↑	SRCC↑	PLCC-log↑	SRCC↑	PLCC-log↑	SRCC↑	PLCC-log↑
Layer 24 only	0.649	0.687	0.397	0.422	0.646	0.739	0.564	0.616
4 layers: $\{1, 7, 13, 19\}$	0.788	0.837	0.627	0.641	0.697	0.758	0.704	0.745
4 layers: $\{2, 8, 14, 20\}$	0.803	0.856	0.630	0.648	0.683	0.762	0.705	0.755
4 layers: $\{3, 9, 15, 21\}$	**0.813**	**0.868**	0.632	0.666	0.668	0.750	0.705	0.761
4 layers: $\{4, 10, 16, 22\}$	0.799	0.855	0.644	0.667	0.711	0.786	0.718	0.770
4 layers: $\{5, 11, 17, 23\}$	0.798	0.859	0.577	0.609	0.712	0.756	0.696	0.741
4 layers: $\{6, 12, 18, 24\}$ (FAIM)	0.798	0.852	0.661	0.672	0.698	0.774	0.719	0.766
8 layers: $\{1, 4, 7, 10, 13, 16, 19, 22\}$	0.778	0.808	0.529	0.560	0.655	0.743	0.654	0.704
8 layers: $\{2, 5, 8, 11, 14, 17, 20, 23\}$	0.799	0.854	0.598	0.624	0.716	0.768	0.704	0.749
8 layers: $\{3, 6, 9, 12, 15, 18, 21, 24\}$	0.782	0.825	0.611	0.624	0.674	0.757	0.689	0.735
16 layers: fixed 2-of-3 sampling	0.806	0.862	**0.665**	**0.688**	0.713	**0.791**	**0.728**	**0.780**
All 24 layers	0.801	0.860	0.637	0.653	**0.717**	0.775	0.718	0.763
Per-layer + post-concatenation ℓ_2 normalization (FAIM)	0.798	0.852	0.661	0.672	0.698	0.774	0.719	0.766
w/o per-layer ℓ_2 normalization	0.799	0.856	0.615	0.639	0.670	0.729	0.695	0.741
w/o post-concatenation ℓ_2 normalization	0.809	0.862	0.609	0.645	0.677	0.752	0.698	0.753
w/o both ℓ_2 normalizations	0.753	0.790	0.549	0.585	0.685	0.765	0.662	0.713
Uniform DINOv3 layer weighting	0.793	0.843	0.634	0.649	0.680	0.747	0.702	0.746
Learned DINOv3 layer weighting (FAIM)	0.798	0.852	0.661	0.672	0.698	0.774	0.719	0.766

The 16-layer variant improves the average SRCC/PLCC-log by $0.009/0.014$ over four-layer FAIM, but increases the FAIM-specific parameter count from 0.046M to 0.071M and peak GPU memory from 1.520GB to 2.312GB (Tab. 29). Using all 24 layers further increases peak memory to 2.848GB without improving the average over the four-layer model. We therefore use four layers as the default, since they retain strong cross-dataset accuracy with a substantially lower memory and parameter cost.

Both normalization stages are important. Removing per-layer normalization, post-concatenation normalization, or both reduces average SRCC from 0.719 to 0.695, 0.698, and 0.662, respectively. Replacing learned layer weights with uniform weights similarly lowers the average SRCC/PLCC-log to $0.702/0.746$. These results support normalized, adaptively weighted integration of features from multiple depths.

Spatial Consistency Branch. The spatial-aggregation sweep shows that SCB benefits from retaining a moderate fraction of the strongest discrepancy responses. Top-10% aggregation gives the best average result, $0.719/0.766$. Top-50% and mean aggregation reach $0.706/0.753$ and $0.699/0.745$, respectively, while Top-0.1% gives $0.688/0.735$. These results indicate that informative discrepancies are spatially concentrated, but using either too few locations or averaging the full response is less effective.

The remaining SCB ablations further support the default design. Symmetric ILR$rightarrow$SR comparison outperforms either one-way direction, and the residual representation performs better on average than either ordered concatenation. The learned discrepancy scorer also outperforms fixed ℓ_2 and cosine measures. Removing the Softplus constraint gives $0.708/0.751$, showing a modest benefit from non-negative scoring. Mean fusion of the two directions is also stronger than maximum or learned directional fusion. These results support symmetric residual comparison, learned non-negative scoring, and Top-10% spatial aggregation.

Table 24: Ablations of spatial evidence aggregation and discrepancy modeling in the Spatial Consistency Branch. Each row changes one factor relative to the Top-10% reference configuration.

Variant	RealDeg-test		RealSR		RealPhoto60		Average	
	SRCC↑	PLCC-log↑	SRCC↑	PLCC-log↑	SRCC↑	PLCC-log↑	SRCC↑	PLCC-log↑
Top-0.1% spatial aggregation	0.800	0.847	0.593	0.613	0.672	0.745	0.688	0.735
Top-0.5% spatial aggregation	**0.802**	0.854	0.635	0.668	0.685	0.762	0.708	0.761
Top-1% spatial aggregation	0.794	0.840	0.646	0.654	0.687	0.761	0.709	0.752
Top-5% spatial aggregation	0.792	0.846	0.598	0.621	0.652	0.724	0.681	0.730
Top-10% spatial aggregation (FAIM)	0.798	0.852	**0.661**	**0.672**	0.698	0.774	**0.719**	**0.766**
Top-15% spatial aggregation	0.800	0.862	0.623	0.650	0.657	0.740	0.694	0.751
Top-20% spatial aggregation	0.791	0.843	0.642	0.656	0.681	0.754	0.705	0.751
Top-25% spatial aggregation	0.793	0.848	0.630	0.644	0.695	0.773	0.706	0.755
Top-50% spatial aggregation	0.792	0.846	0.642	0.655	0.684	0.757	0.706	0.753
Top-75% spatial aggregation	0.793	0.846	0.611	0.630	0.688	0.761	0.698	0.746
Mean spatial aggregation (Top-100%)	0.795	0.850	0.610	0.629	0.692	0.755	0.699	0.745
Maximum spatial aggregation	0.785	0.831	0.625	0.642	0.656	0.743	0.689	0.739
ILR→SR discrepancy only	0.797	**0.869**	0.548	0.607	0.622	0.698	0.656	0.724
SR→ILR discrepancy only	0.800	0.851	0.620	0.637	0.700	0.760	0.707	0.750
Symmetric discrepancy (FAIM)	0.798	0.852	**0.661**	**0.672**	0.698	0.774	**0.719**	**0.766**
Concatenation $[\mathrm{ILR}, \mathrm{SR}]$	0.791	0.855	0.648	0.668	0.671	0.732	0.704	0.752
Concatenation $[\mathrm{SR}, \mathrm{ILR}]$	0.782	0.834	0.649	0.660	0.692	0.769	0.708	0.754
Residual discrepancy (FAIM)	0.798	0.852	**0.661**	**0.672**	0.698	0.774	**0.719**	**0.766**
ℓ_2 discrepancy	0.785	0.834	0.631	0.644	0.698	0.779	0.705	0.752
Cosine discrepancy	0.748	0.791	0.576	0.620	**0.704**	**0.784**	0.676	0.732
Learned discrepancy scorer (FAIM)	0.798	0.852	**0.661**	**0.672**	0.698	0.774	**0.719**	**0.766**
w/o Softplus	0.792	0.844	0.643	0.657	0.688	0.752	0.708	0.751
Symmetric fusion by mean (FAIM)	0.798	0.852	**0.661**	**0.672**	0.698	0.774	**0.719**	**0.766**
Symmetric fusion by maximum	0.782	0.832	0.605	0.628	0.681	0.735	0.689	0.732
Learned directional fusion	0.787	0.825	0.622	0.635	0.699	0.780	0.702	0.746

Table 25: Ablations of the Region Context Consistency Branch.

Variant	RealDeg-test		RealSR		RealPhoto60		Average	
	SRCC↑	PLCC-log↑	SRCC↑	PLCC-log↑	SRCC↑	PLCC-log↑	SRCC↑	PLCC-log↑
w/o Cross-Attention	0.794	0.846	0.616	0.648	0.680	0.768	0.697	0.754
Cross-Attention w/o geometry bias	0.788	0.833	0.607	0.621	0.669	0.734	0.688	0.730
Geometry-Aware Cross Attention (FAIM)	0.798	0.852	**0.661**	**0.672**	0.698	**0.774**	**0.719**	**0.766**
ILR→SR only	0.798	0.848	0.612	0.642	0.687	0.757	0.699	0.749
SR→ILR only	0.800	**0.860**	0.632	0.662	0.696	0.758	0.709	0.760
Bidirectional GCA (FAIM)	0.798	0.852	**0.661**	**0.672**	0.698	**0.774**	**0.719**	**0.766**
1 attention head	**0.804**	0.858	0.637	0.655	0.685	0.760	0.708	0.757
2 attention heads (FAIM)	0.798	0.852	**0.661**	**0.672**	0.698	**0.774**	**0.719**	**0.766**
4 attention heads	0.751	0.794	0.546	0.583	0.681	0.761	0.659	0.712
Mean Region Token pooling (FAIM)	0.798	0.852	**0.661**	**0.672**	0.698	**0.774**	**0.719**	**0.766**
Medoid Region Token pooling	0.751	0.794	0.546	0.583	0.681	0.761	0.659	0.712
Centrality-weighted Region Token pooling	0.794	0.841	0.634	0.648	0.679	0.755	0.702	0.748
Attention Region Token pooling	0.796	0.847	0.643	0.657	**0.712**	0.772	0.717	0.759
Concatenated contextual representation	0.751	0.794	0.547	0.584	0.682	0.761	0.660	0.713
Contextual residual (FAIM)	0.798	0.852	**0.661**	**0.672**	0.698	**0.774**	**0.719**	**0.766**
ℓ_2 Contextual Contradiction	0.743	0.815	0.619	0.633	0.672	0.739	0.678	0.729
Cosine Contextual Contradiction	0.732	0.784	0.532	0.555	0.637	0.727	0.634	0.688
Learned Contextual Contradiction scorer (FAIM)	0.798	0.852	**0.661**	**0.672**	0.698	**0.774**	**0.719**	**0.766**
w/o Contextual Contradiction Softplus	0.793	0.848	0.637	0.646	0.691	0.738	0.707	0.744
Full RCCB configuration	0.798	0.852	**0.661**	**0.672**	0.698	**0.774**	**0.719**	**0.766**

Region Context Consistency Branch. RCCB ablations are evaluated with SCB retained. Removing Cross-Attention reduces the average SRCC/PLCC-log from $0.719/0.766$ to $0.697/0.754$, while removing geometry gives $0.688/0.730$. These results show that both cross-region interaction and explicit spatial relations contribute to region-level fidelity modeling. Although the full model has slightly higher computational cost than the variant without Cross-Attention, the consistent accuracy gain supports retaining geometry-aware relational reasoning.

The remaining ablations refine the RCCB design. Bidirectional context exchange gives $0.719/0.766$, compared with $0.699/0.749$ and $0.709/0.760$ for the two one-way alternatives, and two attention heads give the best average result among the tested settings. Mean Region Token pooling also performs best overall, while attention pooling remains competitive at $0.717/0.759$. Residual contextual representation with a learned contradiction scorer outperforms concatenation and fixed ℓ_2/cosine alternatives. Removing the Softplus constraint lowers the average to $0.707/0.744$, indicating a modest benefit from non-negative scoring. These results support the two-head, bidirectional, geometry-aware RCCB used in FAIM.

Table 26: Ablation of Region Partitioning granularity and DINOv3 feature source.

Variant	RealDeg-test		RealSR		RealPhoto60		Average	
	SRCC↑	PLCC-log↑	SRCC↑	PLCC-log↑	SRCC↑	PLCC-log↑	SRCC↑	PLCC-log↑
No Region Partitioning	0.792	0.845	0.628	0.653	0.685	0.750	0.702	0.749
Target ≥ 6 regions (FAIM)	0.798	0.852	0.661	0.672	0.698	0.774	**0.719**	0.766
Target ≥ 8 regions	0.803	0.859	0.626	0.644	0.676	0.749	0.701	0.751
Target ≥ 16 regions	0.774	0.818	0.584	0.600	0.707	0.770	0.688	0.729
Target ≥ 32 regions	0.804	0.862	0.629	0.652	0.689	0.769	0.707	0.761
DINOv3 layer 6	0.787	0.842	0.632	0.650	0.706	0.778	0.709	0.757
DINOv3 layer 12	0.806	0.865	0.639	0.669	0.697	0.770	0.714	**0.768**
DINOv3 layer 18	0.804	0.859	0.630	0.653	0.700	0.772	0.712	0.762
DINOv3 layer 24 (FAIM)	0.798	0.852	0.661	0.672	0.698	0.774	**0.719**	0.766
Layers $\{12, 24\}$, mean fusion	0.777	0.830	0.592	0.616	0.678	0.747	0.682	0.731
Layers $\{12, 24\}$, learned fusion	**0.808**	**0.866**	0.640	0.662	0.700	0.777	0.716	**0.768**
Layers $\{6, 12, 18, 24\}$, mean fusion	0.775	0.840	0.622	0.642	0.681	0.750	0.693	0.744
Layers $\{6, 12, 18, 24\}$, learned fusion	0.790	0.848	0.650	0.663	**0.709**	**0.782**	0.717	0.764
8 uniformly sampled layers, mean fusion	0.785	0.835	0.622	0.639	0.689	0.761	0.699	0.745
8 uniformly sampled layers, learned fusion	0.800	0.856	**0.663**	**0.676**	0.687	0.764	0.717	0.765
All 24 layers, mean fusion	0.792	0.842	0.633	0.643	0.706	0.780	0.711	0.755
All 24 layers, learned fusion	0.791	0.843	0.633	0.645	0.692	0.768	0.706	0.752

Table 27: Ablation of the three-stage optimization strategy.

Variant	RealDeg-test		RealSR		RealPhoto60		Average	
	SRCC↑	PLCC-log↑	SRCC↑	PLCC-log↑	SRCC↑	PLCC-log↑	SRCC↑	PLCC-log↑
Joint optimization only	0.750	0.794	0.545	0.582	0.680	0.764	0.659	0.713
Stage I → Stage III	0.733	0.778	0.576	0.595	0.653	0.740	0.654	0.704
Stage II → Stage III	0.806	0.864	0.581	0.609	0.636	0.679	0.674	0.717
Stage I → Stage II	0.788	0.832	0.612	0.620	0.697	0.769	0.699	0.740
Stage II → Stage I → Stage III	**0.808**	0.868	0.589	0.628	0.637	0.687	0.678	0.728
Stage I → Stage II (LAFI trainable) → Stage III	**0.808**	**0.869**	0.608	0.629	0.655	0.736	0.690	0.745
Stage I → Stage II → Stage III (LAFI frozen in Stage III)	0.776	0.826	0.646	0.645	0.663	0.755	0.695	0.742
Stage I → Stage II → Stage III (FAIM)	0.798	0.852	**0.661**	**0.672**	**0.698**	**0.774**	**0.719**	**0.766**

Region Partitioning. Region Partitioning contributes independently of the subsequent contextual comparison: removing it lowers the average SRCC/PLCC-log from $0.719/0.766$ to $0.702/0.749$. Increasing the target number of regions beyond six does not provide a consistent benefit, with average SRCC values of 0.701, 0.688, and 0.707 for targets of 8, 16, and 32 regions. The 32-region setting also has slightly higher computational cost without improving accuracy (Tab. 29). These results support using six regions as the default.

For self-classification, single-layer features remain competitive with multi-layer fusion. Layer 24 gives the highest average SRCC (0.719), while layer 12 gives a slightly higher PLCC-log (0.768) but lower SRCC. Learned $\{12, 24\}$ fusion reaches $0.716/0.768$ and requires more peak GPU memory than the single-layer setting (Tab. 29). Larger source sets do not improve the overall rank correlation. We therefore use layer 24 as the default because it gives the strongest average SRCC with a simpler and more memory-efficient feature source.

Three-stage optimization. We first examine the optimization schedule before analyzing the ranking objective used at each stage. FAIM follows a progressive three-stage strategy: Stage I optimizes LAFI and RCCB, Stage II freezes the shared LAFI and optimizes SCB, and Stage III jointly fine-tunes all learnable modules. The complete Stage I→Stage II→Stage III schedule reaches an average SRCC/PLCC-log of $0.719/0.766$. In comparison, joint optimization alone gives $0.659/0.713$, while removing Stage II, Stage I, or Stage III reduces the average to $0.654/0.704$, $0.674/0.717$, and $0.699/0.740$, respectively. Reversing the first two stages also decreases performance to $0.678/0.728$. These results show that separately learning region-level and spatial evidence before joint refinement provides a more effective optimization path than either direct joint training or incomplete stage combinations.

The parameter-update strategy within the three stages is also important. Keeping LAFI trainable in Stage II reduces the average to $0.690/0.745$, while freezing feature integration in Stage III gives $0.695/0.742$. Thus, stabilizing the shared representation while learning SCB and allowing all learnable modules to adapt during the final joint stage are both beneficial.

Stage-specific ranking supervision. With the three-stage optimization structure fixed, we next examine the ranking supervision used at each stage. Following Appendix D.6, we denote same-

Table 28: Ablation of stage-specific ranking supervision. S and C denote same-image and cross-image ranking, respectively, while S+C denotes $\mathcal{L}_{\text{within}} + \lambda_{\text{cross}}\mathcal{L}_{\text{cross}}$.

Stage I	Stage II	Stage III	RealDeg-test		RealSR		RealPhoto60		Average	
			SRCC↑	PLCC-log↑	SRCC↑	PLCC-log↑	SRCC↑	PLCC-log↑	SRCC↑	PLCC-log↑
S+C	S+C	C	0.798	0.852	**0.661**	**0.672**	**0.698**	0.774	**0.719**	**0.766**
S	S+C	C	**0.809**	**0.867**	0.616	0.649	0.681	0.743	0.702	0.753
C	S+C	C	0.764	0.808	0.604	0.628	**0.698**	**0.777**	0.689	0.738
S+C	S	C	0.807	0.861	0.618	0.652	0.682	0.751	0.702	0.755
S+C	C	C	0.751	0.794	0.545	0.583	0.681	0.764	0.659	0.714
S+C	S+C	S	0.794	0.845	0.630	0.637	0.685	0.754	0.703	0.746
S+C	S+C	S+C	**0.809**	0.866	0.600	0.640	0.678	0.748	0.696	0.751

image ranking by S and cross-image ranking by C. The notation S+C refers to the joint objective $\mathcal{L}_{\text{within}} + \lambda_{\text{cross}}\mathcal{L}_{\text{cross}}$. FAIM adopts S+C/S+C/C for Stages I–III. We report the main FAIM result as the reference and vary the ranking objective at one stage at a time to examine the role of stage-specific supervision.

The results reveal a clear asymmetry between the first two stages and the final joint stage. FAIM with S+C/S+C/C achieves the strongest overall performance, reaching an average SRCC/PLCC-log of $0.719/0.766$. In Stage I, replacing the joint S+C objective with same-image ranking alone gives $0.702/0.753$, whereas using cross-image ranking alone decreases the average performance further to $0.689/0.738$. This indicates that same-image comparisons provide the more important supervision when the shared representation and region-context branch are initially established. Because these comparisons rank alternative SR reconstructions conditioned on the same LR observation, they emphasize restoration-dependent fidelity differences while controlling for image content. Cross-image ranking provides an additional constraint by encouraging the learned fidelity representation to remain comparable across different images.

This dependence on same-image supervision becomes even stronger in Stage II. Using S alone remains close to the joint objective at $0.702/0.755$, whereas using C alone substantially reduces the average correlations to $0.659/0.714$. Stage II specializes the spatial-consistency branch, for which comparisons among SR candidates generated from the same LR observation provide particularly direct supervision for restoration-induced local inconsistencies. Cross-image comparisons alone additionally contain content variation between different LR images and are therefore less effective for this stage.

The preferred supervision changes in Stage III, where the spatial-consistency and region-context branches are jointly adapted through the final FAIM score. FAIM uses cross-image ranking and obtains $0.719/0.766$. Replacing C with S gives $0.703/0.746$, while using the joint S+C objective gives $0.696/0.751$. This behavior is consistent with the role of the final stage: after the two branches have acquired fidelity-sensitive representations, cross-image ranking directly encourages the fused score to remain comparable across different image contents within the same source dataset. Overall, the ablation supports a progression from same-image-driven representation and spatial learning in Stages I and II to cross-image calibration during final joint adaptation.

Accuracy–efficiency trade-off. Tab. 29 compares the computational cost of representative FAIM variants under the same inference setting. The parameter count and model size refer to FAIM-specific modules, while the shared DINOv3-Large backbone is excluded from these two columns. We report GMACs, inference latency, and peak GPU memory together with average correlation performance. This comparison is separate from the cross-metric benchmark in Sec. G.6, which uses a different image resolution.

The results show a clear trade-off between accuracy and cost. The 16-layer variant gives the highest average accuracy ($0.728/0.780$), but increases the FAIM-specific parameter count from 0.046M to 0.071M and peak GPU memory from 1.520GB to 2.312GB. The learned 1024-D projection is also substantially larger than the 32-D default without improving the average correlations. In contrast, the 8-D projection is more compact and matches the baseline average SRCC, but performs worse on RealSR and is slightly lower in average PLCC-log. These results support the four-layer, 32-D configuration as a balanced choice between accuracy and computational cost.

Table 29: Accuracy–efficiency trade-off for representative FAIM variants. Params and Size denote FAIM-specific modules and exclude the shared DINOv3-Large backbone. Average denotes the macro-average over the three test datasets.

Variant	Avg. SRCC↑	Avg. PLCC-log↑	FAIM Params (M)↓	FAIM Size (MB)↓	GMACs↓	Latency (s)↓	Peak GPU (GB)↓
SCB only	0.699	0.746	0.046	0.175	**1399.403**	**0.615**	1.520
RCCB only	0.684	0.732	0.046	0.175	1410.263	0.918	1.520
Full FAIM	0.719	**0.766**	0.046	0.175	1410.282	0.943	1.520
Projection 8-D	0.719	0.765	**0.015**	**0.058**	1409.815	0.841	1.520
Projection 128-D	0.709	0.762	0.169	0.644	1412.152	0.836	1.521
Projection 1024-D	0.712	0.765	1.316	5.019	1429.602	0.842	1.623
1 selected DINOv3 layer	0.564	0.616	0.040	0.152	1409.811	0.856	**1.464**
8 selected DINOv3 layers $\{2, 5, 8, 11, 14, 17, 20, 23\}$	0.704	0.749	0.054	0.206	1410.910	0.831	1.795
16 selected DINOv3 layers	**0.728**	**0.780**	0.071	0.269	1412.166	0.857	2.312
All 24 DINOv3 layers	0.718	0.763	0.087	0.332	1413.422	0.894	2.848
w/o Cross-Attention	0.697	0.754	0.046	0.175	1410.282	0.852	1.520
Target ≥ 32 regions	0.707	0.761	0.046	0.175	1410.283	0.960	1.520
Self-classification $\{12, 24\}$, learned fusion	0.716	0.768	0.046	0.175	1410.734	0.926	1.794
Separate SCB/RCCB LAFI	0.701	0.749	0.079	0.300	1410.886	0.863	1.520

Table 30: Paired cluster-bootstrap confidence intervals for FAIM and DISTS under the leave-one-SR-method-out setting. Correlations are computed separately for each held-out SR method and macro-averaged over the eight methods. Values are reported as point estimate [95% confidence interval]. "Diff." denotes FAIM minus DISTS on the same bootstrap samples.

Dataset	Method	SRCC	PLCC-log
RealDeg-test	FAIM	0.401 [0.332, 0.465]	0.474 [0.434, 0.544]
RealDeg-test	DISTS	0.359 [0.271, 0.437]	0.481 [0.399, 0.582]
RealDeg-test	Diff.	0.042 [−0.039, 0.116]	−0.007 [−0.107, 0.102]
RealSR	FAIM	0.412 [0.303, 0.502]	0.475 [0.391, 0.567]
RealSR	DISTS	0.083 [−0.023, 0.197]	0.240 [0.199, 0.321]
RealSR	Diff.	0.328 [0.190, 0.447]	0.234 [0.126, 0.320]
RealPhoto60	FAIM	0.562 [0.467, 0.633]	0.637 [0.558, 0.716]
RealPhoto60	DISTS	0.442 [0.344, 0.532]	0.518 [0.445, 0.603]
RealPhoto60	Diff.	0.120 [0.033, 0.209]	0.119 [0.036, 0.196]
Average	FAIM	0.458 [0.401, 0.500]	0.529 [0.495, 0.583]
Average	DISTS	0.295 [0.236, 0.347]	0.413 [0.384, 0.474]
Average	Diff.	0.163 [0.101, 0.218]	0.116 [0.058, 0.163]

G.5 Statistical Significance

We conduct paired cluster-bootstrap analyses under both the main leave-one-SR-method-out setting and the complementary known-method LR-held-out setting. In both cases, LR images are treated as bootstrap clusters so that all associated predictions are resampled jointly. We use 1,000 bootstrap resamples and report point estimates together with percentile-based 95% confidence intervals.

Leave-one-SR-method-out setting. We first evaluate statistical significance under the same unseen-method protocol as the main comparison in Tab. 1. Within each dataset, LR images are resampled with replacement, and the same resampled LR images are used for FAIM and DISTS across all eight held-out-method folds. SRCC and PLCC-log are computed separately for each held-out SR method and then macro-averaged over the eight methods, following the main evaluation protocol. We compare FAIM with DISTS, the strongest baseline in Tab. 1.

As shown in Tab. 30, FAIM improves the macro-average SRCC over DISTS by 0.163, with a 95% confidence interval of $[0.101, 0.218]$, and improves PLCC-log by 0.116, with a confidence interval of $[0.058, 0.163]$. Both intervals remain strictly above zero, providing statistical evidence for the overall advantage of FAIM when generalizing to unseen SR methods. The improvement is particularly clear on RealSR and RealPhoto60, where the paired confidence intervals are above zero for both correlation measures. On RealDeg-test, the differences are smaller and the confidence intervals include zero, indicating comparable performance between FAIM and DISTS on this dataset.

Known-method LR-held-out setting. We additionally retain the cluster-bootstrap analysis under the known-method LR-held-out setting as a complementary evaluation. Here, all eight SR methods are represented during FAIM training and validation, while the test LR images remain unseen. Each sampled LR image contributes all eight associated SR outputs, and correlations are computed by

Table 31: Paired cluster-bootstrap confidence intervals for FAIM and PSNR under the known-method LR-held-out setting. Correlations are computed by pooling outputs from all eight SR methods within each dataset. Values are reported as point estimate [95% confidence interval]. "Diff." denotes FAIM minus PSNR on the same bootstrap samples.

Dataset	Method	SRCC	PLCC-log
RealDeg-test	FAIM	0.798 [0.771, 0.824]	0.852 [0.832, 0.872]
RealDeg-test	PSNR	0.651 [0.606, 0.692]	0.719 [0.681, 0.757]
RealDeg-test	Diff.	0.148 [0.110, 0.187]	0.133 [0.103, 0.163]
RealSR	FAIM	0.661 [0.608, 0.712]	0.672 [0.621, 0.722]
RealSR	PSNR	0.474 [0.388, 0.560]	0.541 [0.468, 0.616]
RealSR	Diff.	0.187 [0.108, 0.266]	0.131 [0.062, 0.199]
RealPhoto60	FAIM	0.698 [0.642, 0.748]	0.774 [0.732, 0.811]
RealPhoto60	PSNR	0.518 [0.439, 0.594]	0.531 [0.469, 0.590]
RealPhoto60	Diff.	0.179 [0.132, 0.228]	0.243 [0.196, 0.291]
Average	FAIM	0.719 [0.693, 0.744]	0.766 [0.744, 0.788]
Average	PSNR	0.547 [0.507, 0.587]	0.597 [0.564, 0.633]
Average	Diff.	0.171 [0.136, 0.205]	0.169 [0.139, 0.199]

Table 32: Computational cost comparison on 2048×2048 images using a single NVIDIA RTX A6000 GPU with batch size 1.

Metric	PSNR	SSIM	LPIPS	DISTS	NIQE	MANIQA	MUSIQ	CLIP-IQA	TOPIQ-NR	Q-Align	LIQE	FAIM
Time (s/sample) ↓	0.3420	0.3407	0.3797	0.5687	0.2185	0.5569	0.3293	0.2425	0.2820	0.3446	0.2010	0.8771
Memory (GB) ↓	0.188	0.539	1.132	5.969	0.491	2.693	1.131	1.360	2.886	15.435	2.554	1.520

pooling the outputs of all methods within each test dataset. Following the original analysis, we compare FAIM with PSNR.

Tab. 31 shows that FAIM improves the macro-average SRCC over PSNR by 0.171 (95% CI: [0.136, 0.205]) and PLCC-log by 0.169 (95% CI: [0.139, 0.199]). The paired confidence intervals remain strictly above zero for all three test datasets and both correlation measures. These results complement the unseen-method analysis by showing that FAIM also exhibits a consistent advantage when the evaluated SR methods are represented during training, while the leave-one-SR-method-out results above provide the primary statistical evidence for cross-method generalization.

G.6 Computational Resources

We compare the computational cost of FAIM with existing IQA metrics on 2048×2048 images from RealPhoto60. All methods are evaluated on a single NVIDIA RTX A6000 GPU with a batch size of 1 under the same setting. As shown in Table 32, FAIM requires 0.8771 seconds per ILR–SR pair and 1.520 GB of peak GPU memory. Although its inference time is higher than that of the compared metrics due to DINOv3 feature extraction and region-level relational modeling, its memory consumption remains moderate, enabling evaluation of high-resolution SR images on a single GPU.

H Additional Visualizations

H.1 Annotation Interface

Fig. 8 and Fig. 9 show the two-stage annotation interface used in FAIM-Bench. Stage 1 collects absolute ratings for LR quality, SR perceptual quality, and LR–SR fidelity using five-level discrete scales. Stage 2 collects relative fidelity preferences by asking annotators to compare SR outputs generated from the same LR input. The absolute ratings provide coarse quality and fidelity levels, while the pairwise comparisons provide fine-grained ranking supervision for LR–SR fidelity.

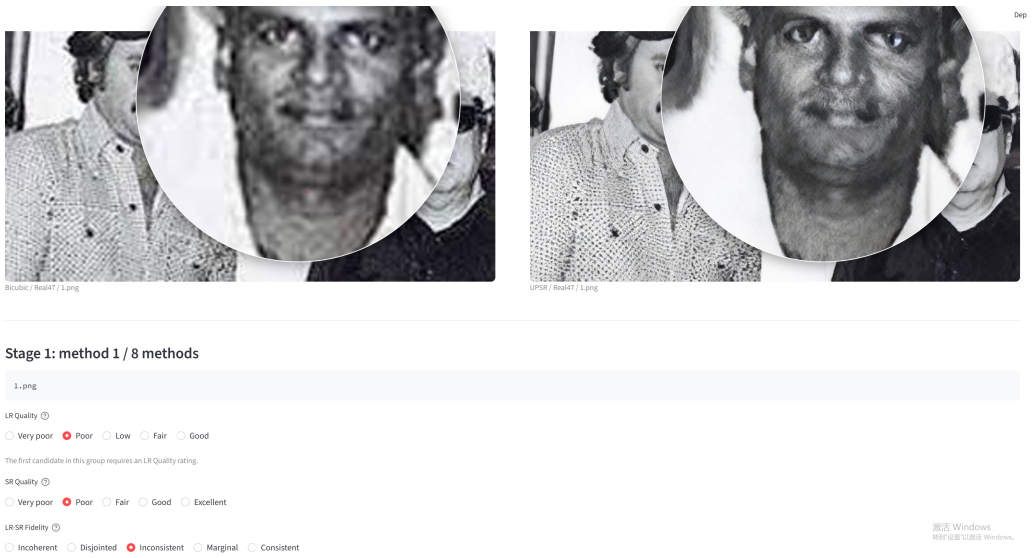


Figure 8: Stage-1 annotation interface. Annotators provide five-level absolute ratings for LR quality, SR perceptual quality, and LR–SR fidelity.

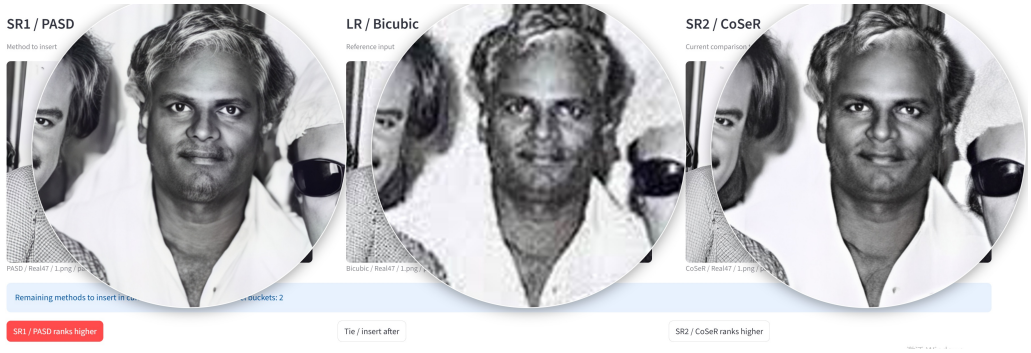


Figure 9: Stage-2 annotation interface. Annotators compare SR outputs generated from the same LR input and provide relative fidelity preferences.

H.2 Layer-wise DINOv3 Feature Visualization

Fig. 10 visualizes the representations produced by all 24 layers of DINOv3. For each ILR or SR image, the first panel shows the input image, followed by PCA visualizations of features from layers L1–L24. Shallow layers exhibit more locally noisy and texture-sensitive responses, preserving fine-grained appearance variations and spatial details. As the network becomes deeper, the representations gradually suppress local texture variations and form more coherent object- and region-level structures, reflecting increasingly semantic information. These observations demonstrate that different DINOv3 depths encode complementary visual cues, motivating the multi-layer representation used in FAIM.

H.3 Visualization of LR–SR Fidelity

We provide additional visual examples to illustrate how FAIM evaluates LR–SR fidelity. As shown in Fig. 11, each example contains the interpolated LR image (ILR), the SR output, the human LR–SR fidelity score, and the FAIM score. These examples show that perceptually plausible SR results can still receive low fidelity scores when their generated structures or textures conflict with visual evidence observable in the LR image. Conversely, SR outputs receive higher fidelity when their restored content remains compatible with the observable structures and layouts of the input, regardless

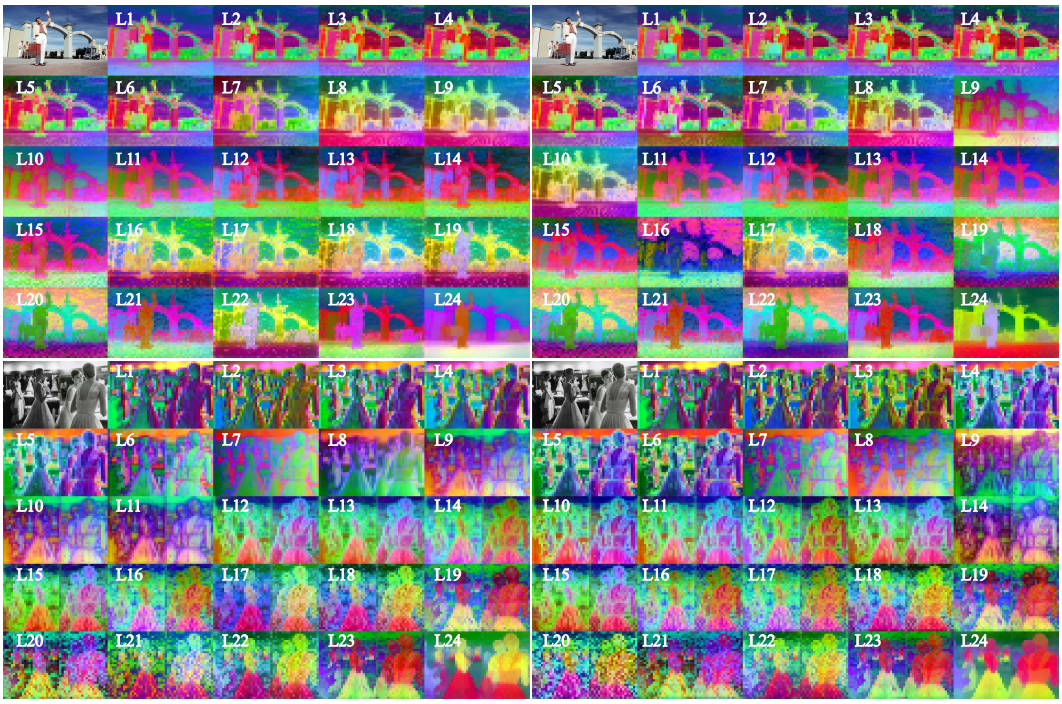


Figure 10: Layer-wise visualization of DINOv3 features for representative ILR–SR pairs. For each image, the first panel shows the input, followed by PCA visualizations of features from layers L1–L24. Shallow layers primarily capture fine-grained texture and local variations, while deeper layers exhibit progressively more coherent semantic structures.

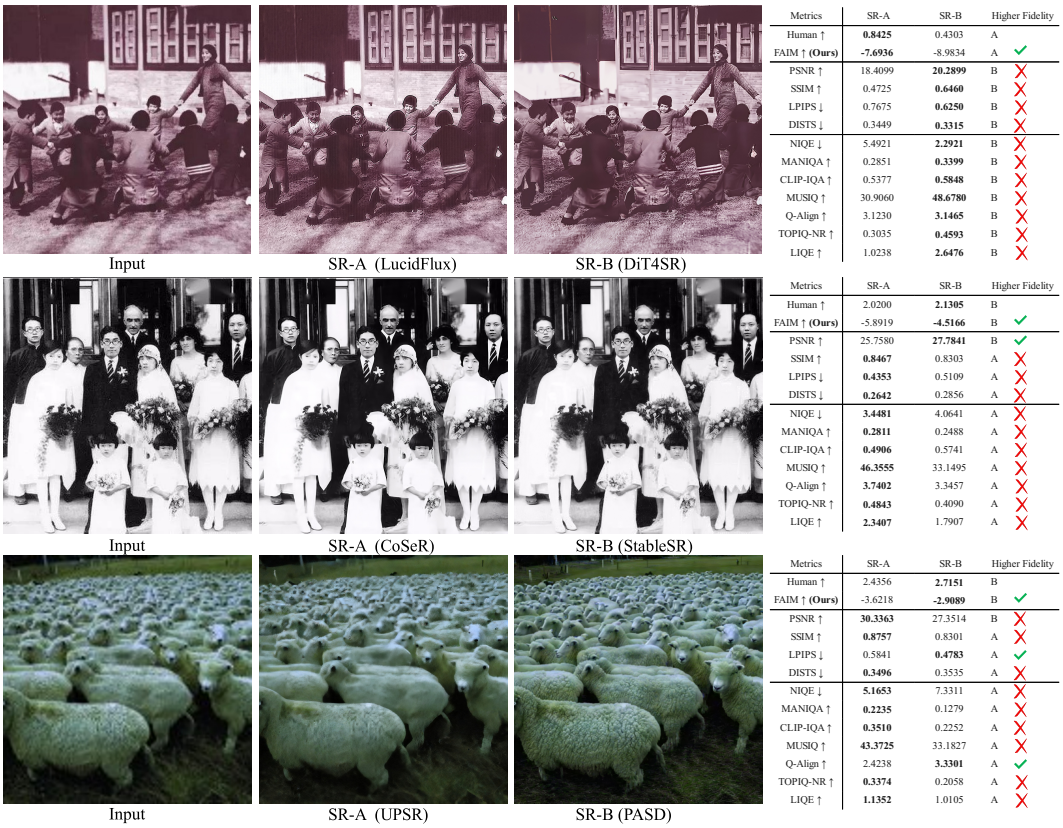
of how aggressively fine details are synthesized. The visualization highlights the difference between perceptual quality and input fidelity. FAIM assigns higher scores to SR results that maintain semantic structures, local textures, and spatial layouts consistent with the LR evidence, rather than simply rewarding sharper or more natural-looking outputs.

H.4 Zero-Shot Artifact Localization

We further visualize the spatial fidelity evidence produced by FAIM in Fig. 12. Although FAIM is trained for fidelity assessment, its spatial discrepancy responses can be directly reused for artifact localization without additional supervision. Given an ILR–SR pair, SCB highlights locations with strong aligned feature discrepancies, while RCCB identifies regions whose SR representations become inconsistent with the LR observation under region-level contextual reasoning. The resulting localization maps tend to emphasize spatially concentrated LR–SR inconsistencies, including structurally or semantically inconsistent regions introduced during generative restoration. These results provide a spatial interpretation of the predicted fidelity score by indicating where FAIM predicts the strongest LR–SR inconsistency evidence.

H.5 Failure Case Analysis

Fig. 13 presents two representative failure cases of FAIM. In both examples, the Fidelity Inconsistency Maps successfully localize visible deviations between the SR results and the input, yet the final rankings disagree with human judgments. A key reason is that the current formulation does not explicitly distinguish the perceptual severity of different inconsistencies. Human observers generally assign greater importance to distortions on salient foreground subjects, such as faces, bodies, or major objects, than to comparable inconsistencies in background regions. FAIM, in contrast, aggregates inconsistency evidence without such semantic importance weighting, allowing widespread but less perceptually critical background deviations to compete with localized yet severe foreground errors.



Metrics	SR-A	SR-B	Higher Fidelity
Human ↑	**0.8425**	0.4303	A
FAIM ↑ **(Ours)**	**-7.6936**	-8.9834	A ✓
PSNR ↑	18.4099	**20.2899**	B ✗
SSIM ↑	0.4725	**0.6460**	B ✗
LPIPS ↓	0.7675	**0.6250**	B ✗
DISTS ↓	0.3449	**0.3315**	B ✗
NIQE ↓	5.4921	**2.2921**	B ✗
MANIQA ↑	0.2851	**0.3399**	B ✗
CLIP-IQA ↑	0.5377	**0.5848**	B ✗
MUSIQ ↑	30.9060	**48.6780**	B ✗
Q-Align ↑	3.1230	**3.1465**	B ✗
TOPIQ-NR ↑	0.3035	**0.4593**	B ✗
LIQE ↑	1.0238	**2.6476**	B ✗

Metrics	SR-A	SR-B	Higher Fidelity
Human ↑	2.0200	**2.1305**	B
FAIM ↑ **(Ours)**	-5.8919	**-4.5166**	B ✓
PSNR ↑	25.7580	**27.7841**	B ✓
SSIM ↑	**0.8467**	0.8303	A ✗
LPIPS ↓	**0.4353**	0.5109	A ✗
DISTS ↓	**0.2642**	0.2856	A ✗
NIQE ↓	**3.4481**	4.0641	A ✗
MANIQA ↑	**0.2811**	0.2488	A ✗
CLIP-IQA ↑	**0.4906**	0.5741	A ✗
MUSIQ ↑	**46.3555**	33.1495	A ✗
Q-Align ↑	**3.7402**	3.3457	A ✗
TOPIQ-NR ↑	**0.4843**	0.4090	A ✗
LIQE ↑	**2.3407**	1.7907	A ✗

Metrics	SR-A	SR-B	Higher Fidelity
Human ↑	2.4356	**2.7151**	B
FAIM ↑ **(Ours)**	-3.6218	**-2.9089**	B ✓
PSNR ↑	**30.3363**	27.3514	B ✗
SSIM ↑	**0.8757**	0.8301	A ✗
LPIPS ↓	0.5841	**0.4783**	A ✓
DISTS ↓	**0.3496**	0.3535	A ✗
NIQE ↓	**5.1653**	7.3311	A ✗
MANIQA ↑	**0.2235**	0.1279	A ✗
CLIP-IQA ↑	**0.3510**	0.2252	A ✗
MUSIQ ↑	**43.3725**	33.1827	A ✗
Q-Align ↑	2.4238	**3.3301**	A ✓
TOPIQ-NR ↑	**0.3374**	0.2058	A ✗
LIQE ↑	**1.1352**	1.0105	A ✗

Figure 11: Additional visualization of LR–SR fidelity.

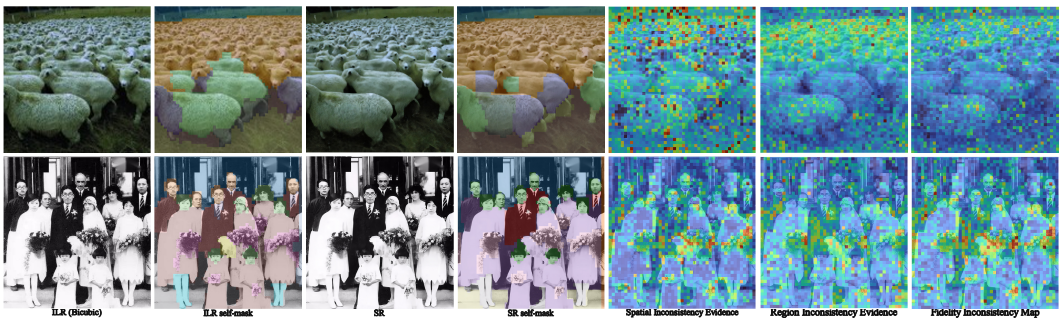


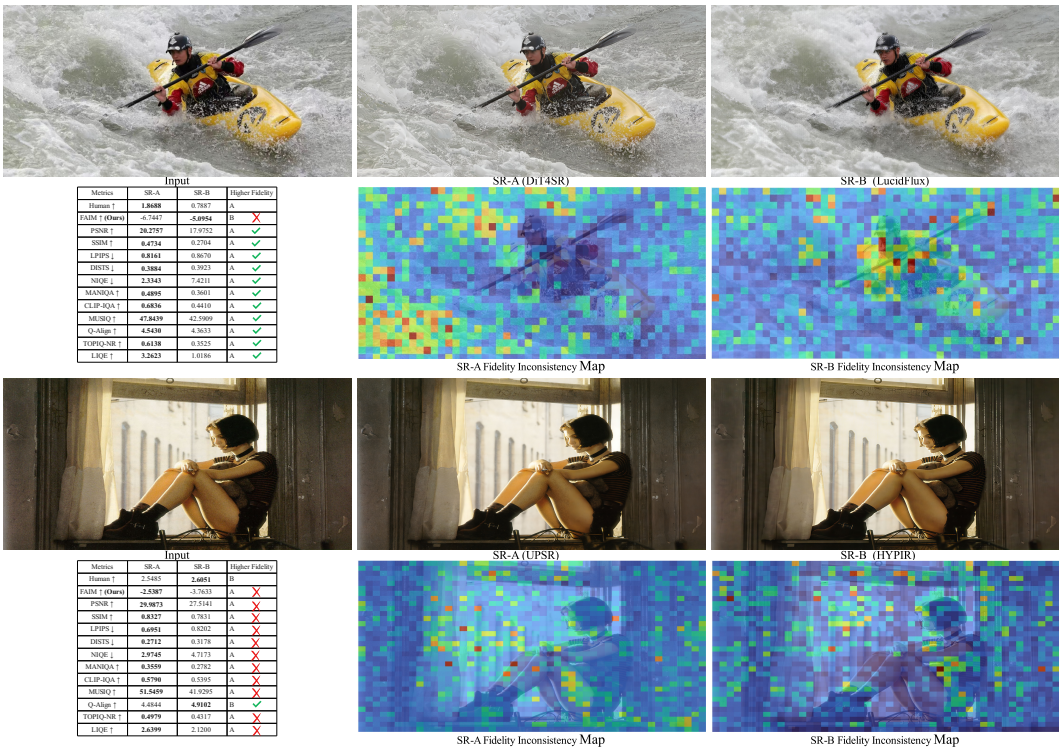
Figure 12: Additional zero-shot artifact localization results.

These cases suggest that incorporating perceptual or semantic saliency into inconsistency aggregation could further improve alignment with human fidelity judgments.

I Societal Impact

FAIM and FAIM-Bench aim to improve the evaluation of real-world image super-resolution, especially for generative SR models that may hallucinate visually plausible but input-inconsistent details. By measuring LR–SR fidelity and localizing potential artifacts, FAIM can support more transparent assessment of restoration models and encourage cautious review in sensitive domains such as surveillance, medical imaging, remote sensing, historical restoration, and scientific imaging.

This work also has potential risks. More accurate fidelity assessment may indirectly support stronger restoration systems, which could be misused for privacy-invasive surveillance, misleading visual



Metrics	SR-A	SR-B	Higher Fidelity
Human ↑	**1.8688**	0.7887	A
FAIM ↑ **(Ours)**	-6.7447	**-5.0954**	B ✗
PSNR ↑	**20.2757**	17.9752	A ✓
SSIM ↑	**0.4734**	0.2704	A ✓
LPIPS ↓	**0.8161**	0.8670	A ✓
DISTS ↓	**0.3884**	0.3923	A ✓
NIQE ↓	**2.3343**	7.4211	A ✓
MANIQA ↑	**0.4895**	0.3601	A ✓
CLIP-IQA ↑	**0.6836**	0.4410	A ✓
MUSIQ ↑	**67.8439**	42.5909	A ✓
Q-Align ↑	**4.5430**	4.3633	A ✓
TOPIQ-NR ↑	**0.6138**	0.3525	A ✓
LIQE ↑	**3.2623**	1.0186	A ✓

Metrics	SR-A	SR-B	Higher Fidelity
Human ↑	2.5485	**2.6051**	B
FAIM ↑ **(Ours)**	**-2.5387**	-3.7633	A ✗
PSNR ↑	**29.9873**	27.5141	A ✗
SSIM ↑	**0.8327**	0.7831	A ✗
LPIPS ↓	**0.6951**	0.8202	A ✗
DISTS ↓	**0.2712**	0.3178	A ✗
NIQE ↓	**2.9745**	4.7173	A ✗
MANIQA ↑	**0.3559**	0.2782	A ✗
CLIP-IQA ↑	**0.5790**	0.5395	A ✗
MUSIQ ↑	**51.5459**	41.9295	A ✗
Q-Align ↑	4.4844	**4.9102**	B ✓
TOPIQ-NR ↑	**0.4979**	0.4317	A ✗
LIQE ↑	**2.6399**	2.1200	A ✗

Figure 13: Failure cases of FAIM. The Fidelity Inconsistency Maps correctly localize visible LR–SR deviations, but the final rankings disagree with human judgments. This is mainly because FAIM does not explicitly model the perceptual importance of different regions, allowing widespread background inconsistencies to outweigh localized but more severe foreground errors.

reconstruction, or manipulation of low-quality images in forensic and evidentiary contexts. Moreover, users may over-trust enhanced images or interpret high fidelity scores as proof of factual correctness. Because FAIM relies on pretrained visual representations and human annotations, its scores may also reflect biases related to image content, acquisition conditions, or annotator preferences.

FAIM should therefore be used as an auxiliary evaluation tool rather than a certification of image authenticity. In high-stakes settings, SR outputs should be accompanied by uncertainty-aware analysis, expert review, and clear disclosure that the images are algorithmically reconstructed.

J Human Annotation Ethics and IRB Status

FAIM-Bench involves human annotation for LR quality, SR perceptual quality, and LR–SR fidelity assessment. The annotation task is limited to subjective visual evaluation of image pairs and does not involve intervention, behavioral experiments, collection of sensitive personal information, or assessment of annotators themselves. Annotators were asked to judge image quality and fidelity according to predefined instructions and five-level rating scales, and their responses were used only in aggregated form for benchmark construction. The potential risks to annotators are minimal. The main foreseeable risk is visual fatigue caused by repeatedly inspecting LR/SR image pairs. Before participation, annotators were informed of the task requirements, annotation procedure, expected workload, and their right to stop participation. No personally identifiable information is reported in the paper, and all annotations are anonymized and aggregated before analysis. According to the applicable institutional policy, this annotation-only benchmark construction does not require formal IRB approval or equivalent human-subjects review. The study was therefore conducted without formal IRB review, while following standard ethical practices for low-risk annotation tasks, including informed participation, privacy protection, and appropriate compensation for annotation labor.