DISRQAD: DIFFUSION IMAGE SUPER-RESOLUTION QUALITY ASSESSMENT DATASET AND BENCHMARK

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion-based image super-resolution (SR) can create visually plausible detail that is not supported by the low-resolution input. We introduce DISRQAD, a subjective-quality dataset and diagnostic benchmark for this setting. It contains mean opinion scores (MOS) for 14,000 SR outputs from ten diffusion and four non-diffusion methods, spanning four low-resolution degradation conditions and $\times 2/\times 4$ upscaling. We evaluate 51 standard full-reference and no-reference metric configurations and 11 adapted variants. Agreement with MOS is substantially weaker on diffusion outputs: the strongest standard no-reference baseline reaches 0.431 SRCC on diffusion SR versus 0.813 on non-diffusion SR. As a case study in benchmark use, a pruned and distilled Q-ReAlign-mini student reaches 0.496 SRCC on diffusion SR. DISRQAD measures perceived output quality, not faithfulness to the input; it enables analysis of metric behavior across generator families and input conditions. Our findings reveal a substantial gap in the assessment of diffusion-based SR and provide a basis for developing quality models sensitive to diffusion-specific artifacts.

1 INTRODUCTION

Super-resolution (SR) reconstructs high-resolution (HR) images from low-resolution (LR) inputs, with deep neural networks driving recent progress Yue et al. (2025); Chen et al. (2026). However, SR methods can alter perceptually salient structures and introduce artifacts such as oversharpening, ringing, and unnatural textures Xie et al. (2023); Molodetskikh et al. (2026). Reliable image quality assessment (IQA) is therefore needed to compare reconstructed images at scale and guide the development of SR methods Sheng et al. (2026); Chen et al. (2025b). This task becomes particularly challenging when generated details look plausible but do not match the underlying scene. A mismatch between metric scores and human judgments can therefore distort comparisons between SR methods, even when they are evaluated on the same inputs.

In this paper, we use "diffusion-based SR" as an umbrella term for the evaluated generative restoration methods that employ diffusion- or flow-based priors. Diffusion-based Super-Resolution methods Yue et al. (2023); Wu et al. (2024b); Yu et al. (2024) have demonstrated promising performance by generating images of higher quality. However, they still struggle to reconstruct plausible structures and textures, often hallucinating objects or modifying highly salient regions, such as human faces. Several works study perceptual artifacts generated by various SR models, including diffusion ones Molodetskikh et al. (2026); Xie et al. (2023); Borisov et al. (2026b). Because a visually attractive output may differ from both the LR evidence and the HR reference, neither fidelity to the reference nor appearance alone provides an unambiguous quality target. Figure 1 illustrates the range of perceived quality among diffusion-based SR outputs. In this work, we measure the visual quality of each output without showing raters its LR input or HR reference; fidelity to the underlying scene is a separate assessment target. This ambiguity may affect full-reference metrics, which penalize departures from the HR image, and no-reference metrics, which may reward plausible but unsupported detail. Human ratings are therefore needed to test how well existing IQA metrics rank generative SR outputs.

IQA metrics can be broadly divided into three categories: Full-Reference (FR) Zhang et al. (2018); Wang et al. (2004); Kirillova et al. (2022), Reduced-Reference (RR) Zheng et al. (2021), and No-Reference (NR) Wu et al. (2024a); Agnolucci et al. (2024b); Ying et al. (2020). FR metrics require

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

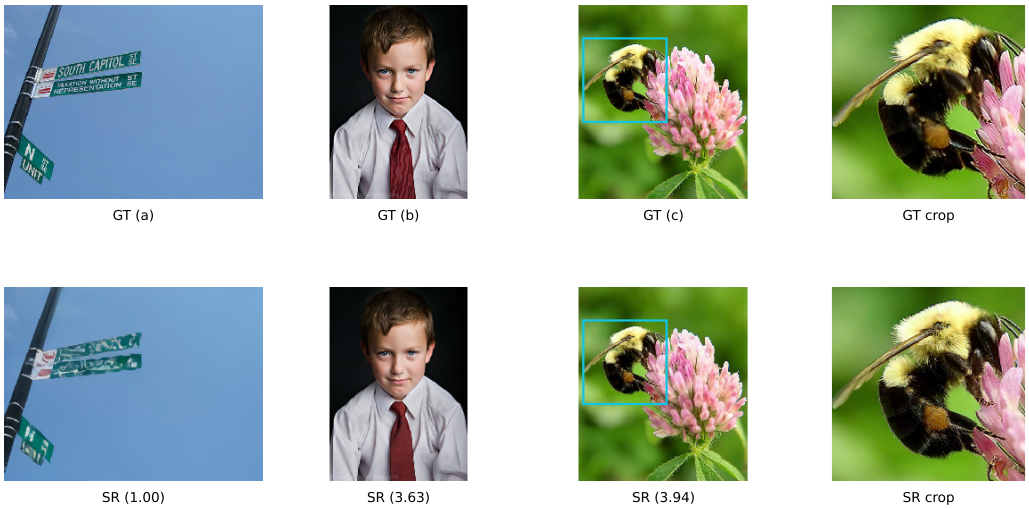


Figure 1: Comparison of diffusion-based super-resolution methods. The number in parentheses below each super-resolved image denotes its mean opinion score (MOS). (a) The super-resolved image exhibits significant artifacts. (b) The super-resolved image is perceptually comparable to the original image. (c) The super-resolved image is perceived as having higher visual quality than the original image.

access to an undistorted reference image, NR metrics estimate image quality using only the distorted image, whereas RR metrics rely on limited information extracted from the reference image. Recent studies Antsiferova et al. (2022) have shown that existing computational metrics exhibit only moderate agreement with human judgments. Furthermore, their limitations become even more pronounced when assessing super-resolved images Shi et al. (2019). Existing subjective studies that include diffusion SR cover relatively few generative methods Chen et al. (2025b); Li et al. (2025). It remains unclear whether the agreement between metric scores and human judgments is consistent across a broader range of generators and input conditions.

Therefore, one of the main objectives of this work is to determine which IQA metrics are most suitable for evaluating images produced by diffusion-based SR methods. We also investigate how well SR-specific IQA metrics perform on such data and analyze how fine-tuning on diffusion-generated images affects metric performance.

To evaluate IQA methods on generative SR, we first construct a diverse pool of LR inputs and outputs from ten diffusion and four non-diffusion SR methods. We then obtain MOS annotations for 10,000 diffusion and 4,000 non-diffusion outputs using more than 100,000 quality judgments. Finally, we compare 51 FR and NR IQA configurations against these human judgments and analyze the effects of SR-specific adaptation.

Our main contributions are as follows:

1. **DISRQAD: a large-scale comprehensive dataset** of 10,000 diffusion-based SR images and 4,000 classic SR images with 205,578 judgments. HR images constitute a comprehensive dataset of 10036 semantically diverse and structurally complex high-quality images.

2. **A comparative evaluation** of 51 IQA metrics of different types (FR and NR) and purposes, including SR-specific methods. We evaluated metrics from different families: traditional, CNN-based, and VLM-based.

3. **Analysis and enhancement of representative IQA metrics:** we fine-tune representative metrics on diffusion-SR specific data to evaluate how SR-specific training helps to align metrics with human visual perception.

2 RELATED WORK

2.1 SUBJECTIVE DATASETS FOR SUPER-RESOLUTION

Early SR-IQA datasets established human judgments as a basis for comparing reconstruction methods. CVIU-2017, QADS, and SISRSet vary algorithms and upscaling factors (Ma et al., 2017; Zhou et al., 2019; Shi et al., 2019); RealSRQ adds captured low-resolution inputs (Jiang et al., 2022). SISAR extends coverage by inferring quality scores from a smaller set of direct ratings (Zhao et al., 2022). These datasets chiefly represent non-diffusion SR, where blur, ringing, and loss of detail are common quality concerns.

Recent studies also evaluate generative SR. SRIQA-Bench includes five diffusion methods and examines cases in which an enhanced image is preferred to its reference (Chen et al., 2025b). ISRGen-QA provides direct mean opinion scores for five diffusion methods (Li et al., 2025). Our benchmark places more diffusion generators alongside non-diffusion SR within one rating procedure. Table 1 summarizes the datasets and their annotation methods.

Table 1: SR datasets with per-image scores. Ratings are absolute judgments; pairs are preferences. Sources are distinct LR inputs for SISAR and HR inputs otherwise.

Dataset	Sources	SR outputs	Non-diff. SR	Diffusion SR	Human input
CVIU-2017 (Ma et al., 2017)	30	1,620	9	0	Ratings
QADS (Zhou et al., 2019)	20	980	21	0	Pairs
SISRSet (Shi et al., 2019)	15	360	8	0	Pairs
RealSRQ (Jiang et al., 2022)	60	1,620	10	0	Pairs
SISAR (Zhao et al., 2022)	100	12,600*	10*	0	Ratings
SRIQA-Bench (Chen et al., 2025b)	100	1,000	5	5	Pairs
ISRGen-QA (Li et al., 2025)	19	720	10	5	Ratings
RQI study (Su et al., 2026)	312	2,184	4	3	Pairs
DISRQAD (ours)	**10,036**†	**14,000**	4	**10**	Ratings

* SISAR directly rates only 1,800 of its 12,600 outputs; the remaining 10,800 scores are inferred. Its method count comprises six SR algorithms and four combined operations.
† DISRQAD reports the full source pool before SR generation, not the number of unique sources represented in the 14,000 subjectively evaluated outputs.

2.2 IQA BENCHMARKS

Subjective protocols ask different questions. SBS180K records which of two SR outputs from the same input is preferred (Khrulkov & Babenko, 2021). FGRestore and LL-Bench combine such preferences with image-level scores across restoration tasks (Sheng et al., 2026; Liu et al., 2026). They broaden task coverage; our analysis holds the task to SR and varies the generator family. In DISRQAD, raters score each SR output directly on a five-category scale. We average their ratings into MOS.

Su et al. (2026) document cases where FR and NR metrics disagree with pairwise human preferences on SR outputs that include diffusion. Agreement with image-level scores is commonly measured by Pearson (PLCC), Spearman (SRCC), or Kendall (KRCC) correlation (Pearson, 1895; Spearman, 1904; Kendall, 1938). We report signed SRCC and PLCC against direct MOS for non-diffusion and diffusion SR separately. This asks whether a metric that follows human ratings on non-diffusion SR also follows them on diffusion SR. Video-SR benchmarks such as VSRQAD (Borisov et al., 2026a) and Herb et al. (2026) also examine generative restoration, but temporal consistency and video-specific artifacts distinguish their annotation setting from our still-image benchmark.

Related work on video SR has compared quality metrics with human preferences for codec-compressed outputs Bogatyrev et al. (2024) and benchmarked real-time SR on streaming content Bogatyrev et al. (2026).

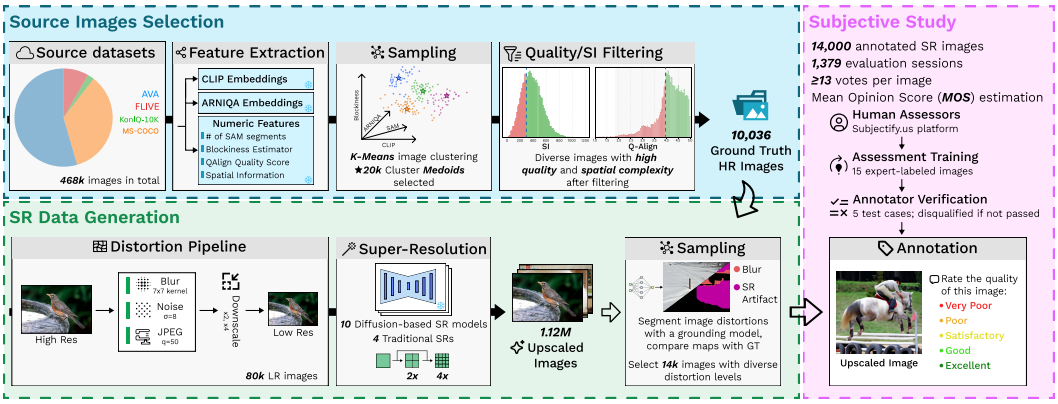


Figure 2: DISRQAD construction: source selection, LR condition generation, SR outputs, and subjective subset selection.

2.3 Image Quality Metrics

Full-reference (FR) metrics compare an SR output with a high-resolution reference; no-reference (NR) metrics judge the output alone. In generative SR, synthesized texture may look natural but differ from the reference. FR metrics measure correspondence with that reference; NR metrics assess appearance without it.

Full-reference metrics. FR methods do not share a single notion of correspondence. PSNR and SSIM reward pixel or structural similarity (Wang et al., 2004), whereas LPIPS and DISTS compare feature representations (Zhang et al., 2018; Ding et al., 2022). ERQA focuses on restored edges (Kirillova et al., 2022); A-FINE considers the case of an imperfect reference (Chen et al., 2025b). Their assumptions matter when SR synthesizes plausible detail that does not reproduce the reference texture.

No-reference metrics. NR methods differ in their supervision. NIQE models natural-scene statistics without human scores (Mittal et al., 2013); MUSIQ learns from subjective ratings (Ke et al., 2021), while Q-Align and DEQA use vision–language representations (Wu et al., 2024a; You et al., 2025). NeuralSBS learns SR preferences (Khrulkov & Babenko, 2021); FGResQ combines scores and rankings for restoration (Sheng et al., 2026). These methods can assess appearance without an HR image, but cannot compare generated details with that image. Table 3 includes the full range of FR and NR metrics.

3 DISRQAD Benchmark

Figure 2 summarizes the pipeline. We clustered more than 468,000 candidate images from KonIQ-10K, COCO, AVA, and FLIVE using structural and semantic features, then selected 10,036 sources using spatial-information (SI) and Q-Align thresholds. Q-Align filtered source-image quality before SR generation. The SI/Q-Align selection favors detailed, visually clean sources and excludes low-SI or low-Q-Align images from the rated benchmark. Appendix A gives the features and thresholds.

We constructed four LR conditions per source and scale: resized inputs and inputs with blur, additive noise, or JPEG compression, at $\times 2$ and $\times 4$. These yield 80,288 LR images. We applied ten diffusion and four non-diffusion SR methods, then selected 10,000 and 4,000 outputs, respectively, by clustering output features. Each output received a direct MOS via subjective comparison. The open part of DISRQAD is available at `https://anonymous-hf.com/a/rg0jfpcdpmvv/`. Further construction details appear in Appendix A.

3.1 Subjective Target and Protocol

Our final target is perceived output quality, not reconstruction fidelity. Assessors viewed each SR output without its LR input or HR reference and rated it on a five-category scale from “very

Table 2: SROCC between MOS computed from different numbers of ratings and the 20-rating baseline, with 95% confidence intervals.

Ratings per image	SROCC	95% confidence interval	
		Lower bound	Upper bound
10	0.982547	0.969290	0.990110
11	0.978173	0.961658	0.987619
12	0.973870	0.954175	0.985165
13	**0.983084**	**0.970229**	**0.990415**
14	0.983596	0.971124	0.990706
15	0.988575	0.979851	0.993534
16	0.991869	0.985641	0.995402
17	0.993483	0.988485	0.996316
18	0.997356	0.997356	0.998507
19	0.998654	0.997617	0.999240

poor" to "excellent." We averaged the ratings to obtain MOS. Thus, a visually attractive but scene-inconsistent image may receive a high score.

A 50-output pilot compared pairwise preference, FR ratings, and NR ratings; its limited sample motivated the selected NR protocol but does not establish general equivalence between protocols. Training and control items screened annotators. The final annotation collected at least 13 ratings per output, yielding more than 205,000 judgments from more than 80 observers. In the pilot's vote-count analysis (Table 2), the 13-rating MOS had SRCC 0.983 with the 20-rating baseline (95% CI $[0.970, 0.990]$); this measures agreement within that pilot, not uncertainty for every final MOS. Appendix B gives the protocol, categories, and pilot results.

4 Experiments

4.1 Metric selection and evaluation

We evaluate 51 standard FR and NR configurations, spanning conventional, learned, SR-oriented, and vision–language IQA, plus 11 adapted variants. Each standard configuration and its citation appear in Table 3. We used each metric's default preprocessing; lower-is-better scores were sign-reversed. SRCC measures ranking agreement and PLCC measures linear agreement with MOS, without nonlinear fitting. Correlations are computed separately on 10,000 diffusion and 4,000 non-diffusion outputs. The MOS-annotated DISRQAD subsets are used for evaluation: the adapted baselines are not trained on their MOS. The 110-image SRIQA-Bench set used to select the student checkpoint is a validation set, not an independent test set.

4.2 Benchmark results

Agreement with MOS decreases substantially on diffusion SR. Across the 51 configurations in Table 3, the median SRCC decreases from 0.65 on non-diffusion outputs to 0.27 on diffusion outputs, and the median PLCC decreases from 0.60 to 0.31. The decrease spans both FR and NR methods with widely different designs and supervision, rather than being confined to one metric family.

The strongest baselines narrow but do not close this gap. TOPIQ(FR) decreases from 0.81/0.79 SRCC/PLCC on non-diffusion SR to 0.40/0.42 on diffusion SR. Q-Align, a strong NR baseline, decreases from 0.81/0.76 to 0.42/0.49. The conventional FR metric FSIM also remains competitive on diffusion SR, reaching 0.39/0.40. These results show that neither access to an HR reference nor general-purpose learned quality representations alone ensures agreement with human judgments of diffusion SR. Q-Align's comparatively strong performance without SR-specific adaptation motivates the metric developed in Section 6.

The fine-tuned examples in the leaderboard further show that SR adaptation is not uniformly beneficial. Fine-tuning LPIPS-AlexNet changes diffusion SRCC only from 0.28 to 0.29, while LPIPS-VGG decreases from 0.30 to 0.27. MUSIQ improves on diffusion SR from 0.27 to 0.31 but decreases on non-diffusion SR from 0.60 to 0.50. A complete comparison of the adapted variants is therefore

Table 3: Overall signed correlations with MOS for 51 standard full-reference (FR) and no-reference (NR) IQA configurations on non-diffusion and diffusion SR. Rows are ordered by non-diffusion SRCC within each type. Best and second-best results within each type and column are shown in bold and underlined, respectively. Values are rounded to three decimal places.

		Non-diffusion SR		Diffusion SR	
Metric	Type	SRCC	PLCC	SRCC	PLCC
TOPIQ(FR) (Chen et al., 2024)	FR	**0.811**	**0.789**	**0.397**	**0.417**
AFINE(FR) (Chen et al., 2025b)	FR	<u>0.787</u>	<u>0.738</u>	0.310	<u>0.410</u>
VSI (Zhang et al., 2014)	FR	0.765	0.663	0.355	0.366
GMSD (Xue et al., 2014)	FR	0.765	0.719	0.317	0.325
LPIPS-VGG (Zhang et al., 2018)	FR	0.763	0.719	0.300	0.329
FSIM (Zhang et al., 2011)	FR	0.758	0.681	<u>0.386</u>	0.405
DISTS (Ding et al., 2022)	FR	0.751	0.695	0.232	0.283
DMM (Chen et al., 2025a)	FR	0.739	0.466	0.309	0.308
MS-SSIM (Wang et al., 2003)	FR	0.725	0.611	0.308	0.288
LPIPS-AlexNet (Zhang et al., 2018)	FR	0.723	0.655	0.282	0.311
WaDIQaM(FR) (Bosse et al., 2018)	FR	0.715	0.701	0.366	0.373
VIF (Sheikh & Bovik, 2006)	FR	0.687	0.669	0.344	0.349
MAD (Larson & Chandler, 2010)	FR	0.685	0.641	0.260	0.257
NLPD (Laparra et al., 2016)	FR	0.678	0.603	0.290	0.282
AHIQ (Lao et al., 2022)	FR	0.640	0.594	0.291	0.322
PSNR	FR	0.630	0.621	0.257	0.259
SSIM (Wang et al., 2004)	FR	0.597	0.566	0.271	0.272
PieAPP (Prashnani et al., 2018)	FR	0.567	0.451	0.214	0.269
CW-SSIM (Sampat et al., 2009)	FR	0.504	0.437	0.183	0.156
ERQA (Kirillova et al., 2022)	FR	0.493	0.388	0.244	0.289
CKDN	FR	0.093	0.173	0.334	0.373
Q-Align (Wu et al., 2024a)	NR	**0.813**	**0.758**	<u>0.415</u>	**0.487**
Q-ReAlign-lite	NR	<u>0.772</u>	0.719	0.405	0.458
DEQA (You et al., 2025)	NR	0.770	<u>0.728</u>	0.324	0.392
Q-ReAlign-pro	NR	0.760	0.716	**0.431**	<u>0.469</u>
LIQE (Zhang et al., 2023)	NR	0.719	0.684	0.279	0.342
NIMA (Talebi & Milanfar, 2018)	NR	0.702	0.671	0.329	0.357
HyperIQA (Su et al., 2020)	NR	0.688	0.662	0.208	0.306
Q-ReAlign-mini	NR	0.683	0.597	0.341	0.393
TRIQA (Sureddi et al., 2025)	NR	0.678	0.636	0.385	0.438
CLIPIQA (Wang et al., 2023)	NR	0.670	0.643	0.332	0.370
TOPIQ(NR) (Chen et al., 2024)	NR	0.668	0.623	0.256	0.344
MACLIP (Liao et al., 2026)	NR	0.651	0.625	0.274	0.317
DBCNN (Zhang et al., 2020)	NR	0.642	0.601	0.225	0.315
MANIQA (Yang et al., 2022)	NR	0.629	0.602	0.313	0.348
AFINE(NR) (Chen et al., 2025b)	NR	0.623	0.561	−0.001	0.008
MUSIQ (Ke et al., 2021)	NR	0.605	0.501	0.273	0.332
FGResQ (Sheng et al., 2026)	NR	0.593	0.560	0.006	0.183
QualiCLIP (Agnolucci et al., 2024a)	NR	0.590	0.562	0.032	0.121
TReS (Golestaneh et al., 2022)	NR	0.567	0.482	0.158	0.241
CNNIQA (Kang et al., 2014)	NR	0.566	0.416	0.159	0.232
WaDIQaM(NR) (Bosse et al., 2018)	NR	0.558	0.506	0.122	0.179
ARNIQA (Agnolucci et al., 2024b)	NR	0.533	0.489	0.103	0.190
PI (Blau et al., 2019)	NR	0.419	0.374	−0.101	−0.039
NRQM (Ma et al., 2017)	NR	0.403	0.355	−0.074	−0.004
PaQ-2-PiQ (Ying et al., 2020)	NR	0.401	0.283	0.148	0.228
ILNIQE (Zhang et al., 2015)	NR	0.396	0.305	−0.229	−0.112
PIQE (Venkatanath et al., 2015)	NR	0.380	0.363	−0.061	−0.046
BRISQUE (Mittal et al., 2012)	NR	0.370	0.354	0.015	0.009
NIQE (Mittal et al., 2013)	NR	0.360	0.327	−0.077	−0.044
NeuralSBS (Khrulkov & Babenko, 2021)	NR	−0.184	−0.087	0.238	0.258

reported separately: improvements in one SR family should not be treated as evidence of transfer to the other.

5 Analysis by Condition and SR Adaptation

5.1 Performance by content and input condition

All results described below are presented in Table 4. Metric performance differs substantially between classic and diffusion-based SR. On classic outputs, Q-Align reaches an overall SRCC of

Table 4: Within-subset Spearman correlation (SRCC) with MOS across SR generators, input degradations, and upscaling factors. PSNR and TOPIQ-FR use a reference image (FR); Q-Align and distilled Q-ReAlign-mini do not (NR). Each correlation is computed separately on its indicated subset. In the *SR method* rows, bold and underlined values mark the highest and lowest SRCC within each SR family and metric, respectively; they do not indicate statistical significance.

		SRCC with MOS			
Factor	Subset	PSNR (FR)	TOPIQ-FR	Q-Align (NR)	Q-ReAlign-mini (distilled, NR)
Diffusion SR					
Overall	All outputs	0.257	0.397	0.415	0.496
SR method	ResShift	**0.412**	**0.587**	0.571	0.613
	SinSR	0.342	0.506	0.428	0.515
	SeeSR	0.301	0.367	0.301	0.444
	PASD	0.239	0.232	0.223	0.326
	PASD-SDXL	0.255	0.393	0.447	0.468
	SUPIR	0.248	0.407	0.249	0.356
	DiT4SR	<u>0.183</u>	<u>0.152</u>	<u>0.178</u>	<u>0.313</u>
	InvSR	0.332	0.405	0.402	0.527
	CODSR	0.196	0.247	0.287	0.421
	PassionSR	0.317	0.571	**0.714**	**0.666**
LR degradation	Blur	0.251	0.448	0.473	0.542
	Noise	0.268	0.396	0.403	0.491
	JPEG	0.234	0.363	0.414	0.495
	Resized	0.274	0.393	0.382	0.466
Upscaling	$\times 2$	0.204	0.245	0.314	0.442
	$\times 4$	0.147	0.399	0.439	0.505
Non-diffusion SR					
Overall	All outputs	0.630	0.811	0.813	0.729
SR method	Real-ESRGAN	<u>0.500</u>	<u>0.665</u>	<u>0.633</u>	**0.621**
	RealSR	0.606	0.711	**0.769**	<u>0.306</u>
	SwinIR	0.504	0.701	0.633	0.589
	HAT	**0.631**	**0.715**	0.738	0.485
LR degradation	Blur	0.403	0.768	0.737	0.711
	Noise	0.711	0.864	0.842	0.802
	JPEG	0.603	0.831	0.851	0.757
	Resized	0.699	0.858	0.863	0.804
Upscaling	$\times 2$	0.467	0.689	0.710	0.724
	$\times 4$	0.467	0.669	0.719	0.578

0.813, while TOPIQ-FR reaches 0.811. Their correlations on diffusion outputs fall to 0.415 and 0.397, respectively. This gap extends beyond the leading methods: overall SRCC is higher on classic SR for 60 of the 62 evaluated metric variants. Thus, strong performance on classic SR provides limited evidence that a metric will rank diffusion outputs according to human judgments.

The degradation used to produce the low-resolution input also affects the results. For diffusion SR, the blur subset yields the highest SRCC for almost all of metrics. Upscaling factor and SR generator introduce further variation. On diffusion outputs, 52 of 62 metric variants have higher SRCC at ×4 than at ×2. The direction is less consistent on classic SR, where only 25 variants score higher at ×4.

The difficulty of quality prediction also varies substantially among diffusion SR generators. PassionSR and ResShift yield the highest within-method SRCC for 36 and 20 of the 62 metric variants, respectively. For example, Q-Align reaches 0.714 on PassionSR, while distilled Q-ReAlign-mini reaches 0.666 on PassionSR and 0.613 on ResShift. DiT4SR presents the opposite pattern: it produces the lowest within-method SRCC for 38 metric variants. CODSR and PASD are also challenging, with median SRCCs across metric variants of 0.135 and 0.158. The gap across generators shows why results pooled over all diffusion outputs should be accompanied by within-method correlations.

The differences among classic SR methods are smaller and depend more on the metric. RealSR has the highest median within-method SRCC across metric variants, while the medians for SwinIR, HAT, and Real-ESRGAN lie lower. Q-Align performs particularly well on RealSR and HAT, whereas TOPIQ-FR leads on Real-ESRGAN and SwinIR. Real-ESRGAN gives the lowest SRCC

Table 5: Overall signed correlations with MOS for 10 fine-tuned baselines and one distilled student on non-diffusion and diffusion SR. Rows are ordered by non-diffusion SRCC within each type. Best and second-best results within each type and column are shown in bold and underlined, respectively. Values are rounded to three decimal places.

		Non-diffusion SR		Diffusion SR	
Metric	Type	SRCC	PLCC	SRCC	PLCC
LPIPS-VGG(FT) (Zhang et al., 2018)	FR	**0.778**	**0.727**	0.267	0.306
LPIPS-AlexNet(FT) (Zhang et al., 2018)	FR	<u>0.752</u>	<u>0.691</u>	0.294	0.331
AHIQ(FT) (Lao et al., 2022)	FR	0.720	0.687	<u>0.334</u>	<u>0.388</u>
TOPIQ(FR)(FT) (Chen et al., 2024)	FR	0.694	0.680	**0.436**	**0.474**
PieAPP(FT) (Prashnani et al., 2018)	FR	0.539	0.443	0.306	0.315
Q-ReAlign-mini(distilled)	NR	**0.729**	**0.747**	**0.496**	**0.542**
Q-ReAlign-mini(FT)	NR	<u>0.727</u>	<u>0.674</u>	<u>0.450</u>	<u>0.470</u>
TOPIQ(NR)(FT) (Chen et al., 2024)	NR	0.637	0.594	0.369	0.412
MUSIQ(FT) (Ke et al., 2021)	NR	0.497	0.487	0.308	0.348
ARNIQA(FT) (Agnolucci et al., 2024b)	NR	0.379	0.314	−0.164	−0.155
LIQE(FT) (Zhang et al., 2023)	NR	0.309	0.144	−0.028	−0.016

for 25 of the 62 variants, yet it is not uniformly the hardest case. Thus, no single ranking of classic SR methods captures the behavior of every IQA metric.

Adaptation to diffusion SR improves some results but can reduce performance on classic outputs. Fine-tuning TOPIQ-FR raises its diffusion SRCC from 0.397 to 0.436, while its classic SRCC falls from 0.811 to 0.694. Distilled Q-ReAlign-mini achieves the highest overall diffusion SRCC (0.496) and retains a classic SRCC of 0.729. These findings motivate reporting results by SR family, generator, degradation, and scale alongside aggregate scores.

5.2 Performance gain after fine-tuning

We fine-tuned five full-reference (FR) methods (AHIQ, LPIPS-AlexNet, LPIPS-VGG, PieAPP, and TOPIQ-FR) and five no-reference (NR) methods (ARNIQA, MUSIQ, LIQE, TOPIQ-NR, and Q-ReAlign-mini) on SRIQA-Bench. Fine-tuning was performed under the following settings: a batch size of 8, 30 epochs, an initial learning rate of 1×10^{-4}, MSE loss against the MOS labels, and the Adam optimizer, with the best checkpoint selected according to SRCC. LIQE required a lower learning rate (5×10^{-6}) to ensure convergence. Due to the large model size and rapid convergence of Q-ReAlign-mini, its maximum number of epochs was limited to 2. All training was carried out using the IQA-PyTorch interface.

We measured the performance gain on the classic and diffusion tracks of our benchmark, as well as the decline in quality on non-SR distortions. The effect of fine-tuning depends on the metric and the evaluation domain. Table 5 shows the change in SRCC/PLCC metrics after fine-tuning. For most metrics the effect is positive (the correlation improved on at least one track, and for 3 metrics the correlation improved on both tracks), but there are also negative examples. Presumably, individual fine-tuning configurations need to be selected for them.

On non-SR distortions (the KonIQ-10K, KADID, and SPAQ datasets), the quality of the metrics dropped significantly. Compared with the specialized pipeline of a hierarchically distilled metric, simple fine-tuning on the SR domain reduces the universality of the metric for image quality assessment across different domains. Thus, improving quality on the diffusion SR domain requires the development of specialized approaches that do not neglect quality on classic SR and other types of distortions.

6 Efficient SR-adapted metric

Q-Align is a strong NR baseline but has 8.2B parameters. We use the 0.8B Q-ReAlign-mini (Q-Future, 2026) as teacher and starting point, prune its language stack to obtain a 340M-parameter student, and combine a five-token quality head with a regression head. Hierarchical distillation on KonIQ-10K, KADID-10K, and SPAQ precedes SR adaptation on 990 SRIQA-Bench images with

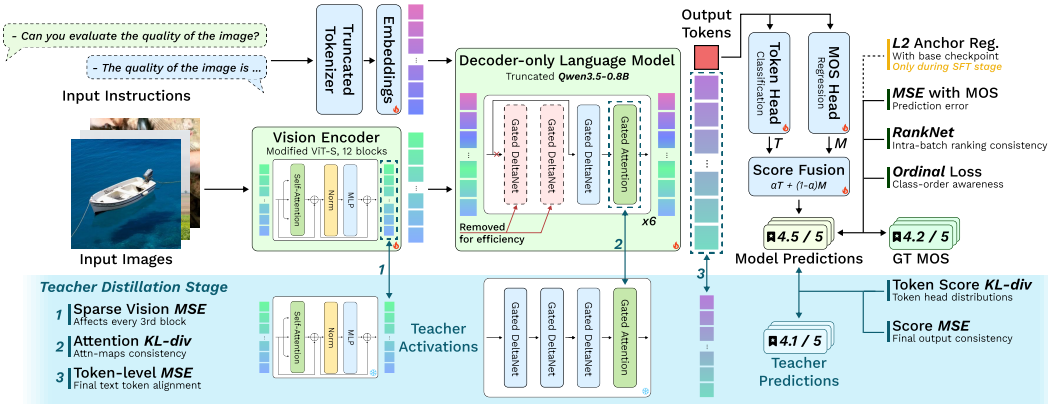


Figure 3: The process of Q-Align pruning and fine-tuning.

Table 6: Accuracy and efficiency of the proposed student and its vision–language baselines. Accuracy is reported as SRCC/PLCC. SRIQA-Bench results use the 110-image validation split. Throughput and peak VRAM are measured on 512×512 inputs using an NVIDIA RTX A6000 in BF16.

Model	Params	VRAM	FPS	SRIQA-Bench	Diffusion	Non-diff.
Q-ReAlign-mini	0.8B	2.3 GB	3.76	0.479/0.504	0.341/0.393	0.683/0.597
Q-ReAlign-lite	4.0B	10.7 GB	2.69	0.475/0.519	0.405/0.458	0.772/0.719
Q-ReAlign-pro	9.0B	19.2 GB	2.65	0.428/0.488	0.431/0.469	0.760/0.716
Q-Align	8.2B	16.0 GB	5.39	0.624/0.677	0.415/0.487	**0.813/0.758**
Ours	**0.34B**	**1.3 GB**	**6.37**	**0.755/0.794**	**0.496/0.542**	0.729/0.747

248 general-IQA replay images. The remaining 110 SRIQA-Bench images are used for checkpoint selection. The architecture, losses, and training settings are in Appendix C. The student retains Q-ReAlign-mini's visual encoder and reduces its 24 language modules to 12 by pruning two Gated DeltaNet modules from each four-module group; its vocabulary is restricted to 30 tokens. One head scores five ordered quality tokens and a second regresses a continuous score from the last hidden state. A learned convex combination gives the final prediction. This compact description makes the reported parameter and throughput changes interpretable; the layer matching, losses, and optimizer settings remain in Appendix C. Figure 3 summarizes the pruning and two-stage training.

7 Conclusion

We introduced DISRQAD, a human-annotated benchmark of 14,000 SR outputs from ten diffusion and four non-diffusion methods, with 205,578 quality judgments. We evaluated 51 standard IQA configurations and 11 adapted variants against mean opinion scores for perceived output quality. For 60 of the 62 evaluated variants, SRCC was lower on diffusion outputs than on non-diffusion outputs. This pattern occurs across full-reference and no-reference metrics, while its magnitude varies with the SR generator, input degradation, and upscaling factor. Performance on conventional SR alone, therefore, provides an incomplete picture of a metric's suitability for generative SR.

As a case study, our pruned and distilled Q-ReAlign-mini student reaches an SRCC of 0.496 on diffusion SR and 0.729 on non-diffusion SR with 340M parameters. The broader fine-tuning results show that adaptation can improve agreement with human ratings, but its effects differ across metrics and SR families. Together, DISRQAD and these analyses provide a basis for developing and testing quality predictors for generative SR. The benchmark measures perceived output quality; assessing whether synthesized details remain faithful to the LR input is a complementary direction for future work.

AI DISCLOSURE

We used generative AI tools for translation and language editing, assistance with drafting and revising code, and identification of relevant literature. The authors reviewed and edited AI-assisted material and take full responsibility for the code, analyses, results, and conclusions.

REPRODUCIBILITY STATEMENT

Section 3 describes the construction of DISRQAD, the subjective quality target, and access to the released portion of the dataset. Section 4 specifies the evaluated IQA configurations and correlation measures. Appendix A details source selection, input degradations, SR methods, and output sampling; Appendix B documents the annotation protocol and pilot study; and Appendix C provides the student architecture and training settings.

REFERENCES

Lorenzo Agnolucci, Leonardo Galteri, and Marco Bertini. Quality-aware image-text alignment for opinion-unaware image quality assessment. *arXiv preprint arXiv:2403.11176*, 2024a. doi: 10.48550/arXiv.2403.11176.

Lorenzo Agnolucci, Leonardo Galteri, Marco Bertini, and Alberto Del Bimbo. Arniqa: Learning distortion manifold for image quality assessment. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 188–197, 2024b. doi: 10.1109/WACV57701.2024.00026.

Anastasia Antsiferova, Sergey Lavrushkin, Maksim Smirnov, Alexander Gushchin, Dmitriy Vatolin, and Dmitriy Kulikov. Video compression dataset and benchmark of learning-based video-quality metrics. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088. doi: 10.52202/068431-1004.

Yochai Blau, Roey Mechrez, Radu Timofte, Tomer Michaeli, and Lihi Zelnik-Manor. The 2018 pirm challenge on perceptual image super-resolution. In *Computer Vision – ECCV 2018 Workshops*, pp. 334–355, 2019. doi: 10.1007/978-3-030-11021-5_21.

Evgeney Bogatyrev, Ivan Molodetskikh, and Dmitriy S. Vatolin. SR+Codec: a benchmark of super-resolution for video compression bitrate reduction. In *35th British Machine Vision Conference (BMVC)*, 2024. URL https://papers.bmvc2024.org/0959.pdf.

Evgeney Bogatyrev, Khaled Abud, Ivan Molodetskikh, Nikita Alutis, and Dmitriy Vatolin. Exploring real-time super-resolution: Benchmarking and fine-tuning for streaming content. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/34fd1a124705103a145451d88e6df231-Abstract-Conference.html.

Artem Borisov, Evgeney Bogatyrev, Khaled Abud, Ivan Molodetskikh, and Dmitriy Vatolin. Vsrqad: Video super-resolution quality assessment dataset and benchmark. *IEEE Access*, 14:60229–60251, 2026a. doi: 10.1109/ACCESS.2026.3679554.

Artem Borisov, Evgeney Bogatyrev, Khaled Abud, and Dmitriy Vatolin. Sr-ground: Image quality grounding for super-resolved content. *arXiv preprint arXiv:2605.21244*, 2026b. doi: 10.48550/arXiv.2605.21244.

Sebastian Bosse, Dominique Maniry, Klaus-Robert Müller, Thomas Wiegand, and Wojciech Samek. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Transactions on Image Processing*, 27:206–219, 01 2018. doi: 10.1109/TIP.2017.2760518.

Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Baoliang Chen, Hanwei Zhu, Lingyu Zhu, Shanshe Wang, Jingshan Pan, and Shiqi Wang. Debiased mapping for full-reference image quality assessment. *IEEE Transactions on Multimedia*, 27: 2638–2649, 2025a. doi: 10.1109/TMM.2025.3535280.

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. *Trans. Img. Proc.*, 33:2404–2418, January 2024. ISSN 1057-7149. doi: 10.1109/TIP.2024.3378466. URL https://doi.org/10.1109/TIP.2024.3378466.

Du Chen, Tianhe Wu, Kede Ma, and Lei Zhang. Toward Generalized Image Quality Assessment: Relaxing the Perfect Reference Quality Assumption . In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12742–12752, Los Alamitos, CA, USA, Jun 2025b. IEEE Computer Society. doi: 10.1109/CVPR52734.2025.01189.

Hao Chen, Junyang Chen, Jinshan Pan, and Jiangxin Dong. Bridging fidelity-reality with controllable one-step diffusion for image super-resolution. In *CVPR*, pp. 30584–30594, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Chen_Bridging_Fidelity-Reality_with_Controllable_One-Step_Diffusion_for_Image_Super-Resolution_CVPR_2026_paper.html.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2567–2581, 2022. doi: 10.1109/TPAMI.2020.3045810.

Zheng-Peng Duan, Jiawei Zhang, Xin Jin, Ziheng Zhang, Zheng Xiong, Dongqing Zou, Jimmy Ren, Chun-Le Guo, and Chongyi Li. Dit4sr: Taming diffusion transformer for real-world image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. doi: 10.1109/ICCV51701.2025.01761.

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

S Alireza Golestaneh, Saba Dadsetan, and Kris M Kitani. No-reference image quality assessment via transformers, relative ranking, and self-consistency. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3209–3218, 2022. doi: 10.1109/WACV51458.2022.00404.

Benjamin Herb, Steve Göring, Alexander Raake, and Rakesh Rao Ramachandra Rao. How accurate are video quality models for diffusion-based video super-resolution? *arXiv preprint arXiv:2605.25940*, 2026.

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

Qiuping Jiang, Zhentao Liu, Ke Gu, Feng Shao, Xinfeng Zhang, Hantao Liu, and Weisi Lin. Single image super-resolution quality assessment: A real-world dataset, subjective studies, and an objective metric. *IEEE Transactions on Image Processing*, 31:2279–2294, 2022. doi: 10.1109/TIP.2022.3154588.

Le Kang, Peng Ye, Yi Li, and David Doermann. Convolutional neural networks for no-reference image quality assessment. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1733–1740, 2014. doi: 10.1109/CVPR.2014.224.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5128–5137, 2021. doi: 10.1109/ICCV48922.2021.00510.

M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1–2):81–93, 1938. doi: 10.1093/biomet/30.1-2.81.

Valentin Khrulkov and Artem Babenko. Neural side-by-side: Predicting human preferences for no-reference super-resolution evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4988–4997, June 2021. doi: 10.1109/CVPR46437.2021.00495.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. doi: 10.1109/ICCV51070.2023.00371.

Anastasia Kirillova, Eugene Lyapustin, Anastasia Antsiferova, and Dmitry Vatolin. Erqa: Edge-restoration quality assessment for video super-resolution. In *Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 4: VISAPP,*, pp. 315–322. INSTICC, SciTePress, 2022. ISBN 978-989-758-555-5. doi: 10.5220/0010780900003124.

Shanshan Lao, Yuan Gong, Shuwei Shi, Sidi Yang, Tianhe Wu, Jiahao Wang, Weihao Xia, and Yujiu Yang. Attentions Help CNNs See Better: Attention-based Hybrid Image Quality Assessment Network . In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1139–1148, Los Alamitos, CA, USA, June 2022. IEEE Computer Society. doi: 10.1109/CVPRW56347.2022.00123. URL https://doi.ieeecomputersociety.org/10.1109/CVPRW56347.2022.00123.

Valero Laparra, Johannes Ball, Alexander Berardino, and Eero Simoncelli. Perceptual image quality assessment using a normalized laplacian pyramid. *Electronic Imaging*, 2016:1–6, 02 2016. doi: 10.2352/ISSN.2470-1173.2016.16.HVEI-103.

Eric Larson and Damon Chandler. Most apparent distortion: Full-reference image quality assessment and the role of strategy. *J. Electronic Imaging*, 19:011006, 01 2010. doi: 10.1117/1.3267105.

Yixiao Li, Xin Li, Chris Wei Zhou, Shuo Xing, Hadi Amirpour, Xiaoshuai Hao, Guanghui Yue, Baoquan Zhao, Weide Liu, Xiaoyuan Yang, et al. Vquala 2025 challenge on image super-resolution generated content quality assessment: Methods and results. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 3341–3351, 2025. doi: 10.1109/ICCVW69036.2025.00354.

Zhicheng Liao, Dongxu Wu, Zhenshan Shi, Sijie Mai, Hanwei Zhu, Lingyu Zhu, Yuncheng Jiang, and Baoliang Chen. Beyond cosine similarity: Magnitude-aware clip for no-reference image quality assessment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 6934–6942, 2026. doi: 10.1609/aaai.v40i9.37627.

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

Lu Liu, Huiyu Duan, Chenxin Zhu, Jintong Lu, Haoyun Jiang, Liu Yang, Qiang Hu, Guangtao Zhai, and Xiaoyun Zhang. Ll-bench: Rethinking low-level vision evaluation in the era of large-scale generative models. *arXiv preprint arXiv:2606.02535*, 2026. doi: 10.48550/arXiv.2606.02535. URL https://arxiv.org/abs/2606.02535.

Chao Ma, Chih-Yuan Yang, Xiaokang Yang, and Ming-Hsuan Yang. Learning a no-reference quality metric for single-image super-resolution. *Computer Vision and Image Understanding*, 158:1–16, 2017. doi: 10.1016/j.cviu.2016.12.009.

Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, 2012. doi: 10.1109/TIP.2012.2214050.

Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2013. doi: 10.1109/LSP.2012.2227726.

Ivan Molodetskikh, Kirill Malyshev, Mark Mirgaleev, Nikita Zagainov, Evgeney Bogatyrev, and Dmitriy Vatolin. Sr-prominence: A crowdsourced protocol and dataset suite for perceptually-weighted super-resolution artifact evaluation. *arXiv preprint arXiv:2605.14847*, 2026. doi: 10.48550/arXiv.2605.14847.

Jorge J. Moré. The levenberg-marquardt algorithm: Implementation and theory. In G. A. Watson (ed.), *Numerical Analysis*, pp. 105–116, Berlin, Heidelberg, 1978. Springer Berlin Heidelberg. ISBN 978-3-540-35972-2.

Karl Pearson. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58:240–242, 1895. doi: 10.1098/rspl.1895.0041.

Ekta Prashnani, Hong Cai, Yasamin Mostofi, and Pradeep Sen. Pieapp: Perceptual image-error assessment through pairwise preference. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. doi: 10.1109/CVPR.2018.00194.

Q-Future. Q-ReAlign: Lightweight human-aligned multimodal judges built on modern vision-language models. Software and model release, 2026. URL https://github.com/Q-Future/Q-ReAlign.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html. Preprint DOI: https://doi.org/10.48550/arXiv.2103.00020.

Mehul P. Sampat, Zhou Wang, Shalini Gupta, Alan Conrad Bovik, and Mia K. Markey. Complex wavelet structural similarity: A new image similarity index. *IEEE Transactions on Image Processing*, 18(11):2385–2401, 2009. doi: 10.1109/TIP.2009.2025923.

Hamid Sheikh and Alan Bovik. Image information and visual quality. *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society*, 15:430–44, 03 2006. doi: 10.1109/TIP.2005.859378.

Xiangfei Sheng, Xiaofeng Pan, Zhichao Yang, Pengfei Chen, and Leida Li. Fine-grained image quality assessment for perceptual image restoration. In *Proceedings of the Fortieth AAAI Conference on Artificial Intelligence and Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence and Sixteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'26/IAAI'26/EAAI'26. AAAI Press, 2026. ISBN 978-1-57735-906-7. doi: 10.1609/aaai.v40i11.37846. URL https://doi.org/10.1609/aaai.v40i11.37846.

Guangming Shi, Wenfei Wan, Jinjian Wu, Xuemei Xie, Weisheng Dong, and Hong Ren Wu. Sisrset: Single image super-resolution subjective evaluation test and objective quality assessment. *Neurocomputing*, 360:37–51, 2019. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2019.06.027. URL https://www.sciencedirect.com/science/article/pii/S0925231219308549.

C. Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904. doi: 10.2307/1412159.

Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. doi: 10.1109/CVPR42600.2020.00372.

Shaolin Su, Josep M. Rocafort, Danna Xue, David Serrano-Lozano, Lei Sun, and Javier Vazquez-Corral. Bridging the perception gap in image super-resolution evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30532–30542, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Su_Bridging_the_Perception_Gap_in_Image_Super-Resolution_Evaluation_CVPR_2026_paper.html. Preprint DOI: https://doi.org/10.48550/arXiv.2503.13074.

Rajesh Sureddi, Saman Zadtootaghaj, Nabajeet Barman, and Alan C. Bovik. Triqa: Image quality assessment by contrastive pretraining on ordered distortion triplets. In *2025 IEEE International Conference on Image Processing (ICIP)*, pp. 1744–1749, 2025. doi: 10.1109/ICIP55913.2025.11084443.

Hossein Talebi and Peyman Milanfar. Nima: Neural image assessment. *IEEE Transactions on Image Processing*, 27(8):3998–4011, August 2018. ISSN 1941-0042. doi: 10.1109/tip.2018.2831899. URL http://dx.doi.org/10.1109/TIP.2018.2831899.

N. Venkatanath, D. Praneeth, Bh. Maruthi Chandrasekhar, Sumohana S. Channappayya, and Swarup S. Medasani. Blind image quality evaluation using perception based features. In *2015 Twenty First National Conference on Communications (NCC)*, pp. 1–6, 2015. doi: 10.1109/NCC.2015.7084843.

Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *AAAI*, 2023. doi: 10.1609/aaai.v37i2.25353.

Yufei Wang, Wenhan Yang, Xinyuan Chen, Yaohui Wang, Lanqing Guo, Lap-Pui Chau, Ziwei Liu, Yu Qiao, Alex C Kot, and Bihan Wen. Sinsr: diffusion-based image super-resolution in a single step. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25796–25805, 2024. doi: 10.1109/CVPR52733.2024.02437.

Z. Wang, E.P. Simoncelli, and A.C. Bovik. Multiscale structural similarity for image quality assessment. In *The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pp. 1398–1402 Vol.2, 2003. doi: 10.1109/ACSSC.2003.1292216.

Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. doi: 10.1109/TIP.2003.819861.

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtao Zhai, and Weisi Lin. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 54015–54029. PMLR, 2024a. URL https://proceedings.mlr.press/v235/wu24ah.html. Preprint DOI: https://doi.org/10.48550/arXiv.2312.17090.

H.R. Wu and M. Yuen. A generalized block-edge impairment metric for video coding. *IEEE Signal Processing Letters*, 4(11):317–320, 1997. doi: 10.1109/97.641398.

Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. Seesr: Towards semantics-aware real-world image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 25456–25467, 2024b. doi: 10.1109/CVPR52733.2024.02405.

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 *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 38204–38226. PMLR, 2023. URL https://proceedings.mlr.press/v202/xie23c.html.

Wufeng Xue, Lei Zhang, Xuanqin Mou, and Alan C. Bovik. Gradient magnitude similarity deviation: A highly efficient perceptual image quality index. *IEEE Transactions on Image Processing*, 23(2):684–695, 2014. doi: 10.1109/TIP.2013.2293423.

Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1191–1200, 2022. doi: 10.1109/CVPRW56347.2022.00126.

Tao Yang, Rongyuan Wu, Peiran Ren, Xuansong Xie, and Lei Zhang. Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization. In *The European Conference on Computer Vision (ECCV) 2024*, 2024. doi: 10.1007/978-3-031-73247-8_5.

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

Zhiyuan You, Xin Cai, Jinjin Gu, Tianfan Xue, and Chao Dong. Teaching large language models to regress accurate image quality scores using score distribution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14483–14494, 2025. doi: 10.1109/CVPR52734.2025.01350.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25669–25680, June 2024. doi: 10.1109/CVPR52733.2024.02425.

Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Resshift: Efficient diffusion model for image super-resolution by residual shifting. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. doi: 10.52202/075280-0583.

Zongsheng Yue, Kang Liao, and Chen Change Loy. Arbitrary-steps image super-resolution via diffusion inversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23153–23163, June 2025. doi: 10.1109/CVPR52734.2025.02156.

Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. Fsim: A feature similarity index for image quality assessment. *IEEE Transactions on Image Processing*, 20(8):2378–2386, 2011. doi: 10.1109/TIP.2011.2109730.

Lin Zhang, Ying Shen, and Hongyu Li. Vsi: A visual saliency-induced index for perceptual image quality assessment. *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society*, 23, 08 2014. doi: 10.1109/TIP.2014.2346028.

Lin Zhang, Lei Zhang, and Alan C. Bovik. A feature-enriched completely blind image quality evaluator. *IEEE Transactions on Image Processing*, 24(8):2579–2591, 2015. doi: 10.1109/TIP.2015.2426416.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. doi: 10.1109/CVPR.2018.00068.

Weixia Zhang, Kede Ma, Jia Yan, Dexiang Deng, and Zhou Wang. Blind image quality assessment using a deep bilinear convolutional neural network. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(1):36–47, 2020. doi: 10.1109/TCSVT.2018.2886771.

Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 14071–14081, 2023. doi: 10.1109/CVPR52729.2023.01352.

Tiesong Zhao, Yuting Lin, Yiwen Xu, Weiling Chen, and Zhou Wang. Learning-based quality assessment for image super-resolution. *Trans. Multi.*, 24:3570–3581, January 2022. ISSN 1520-9210. doi: 10.1109/TMM.2021.3102401. URL https://doi.org/10.1109/TMM.2021.3102401.

Heliang Zheng, Huan Yang, Jianlong Fu, Zheng-Jun Zha, and Jiebo Luo. Learning conditional knowledge distillation for degraded-reference image quality assessment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10242–10251, October 2021. doi: 10.1109/ICCV48922.2021.01008.

Fei Zhou, Rongguo Yao, Bozhi Liu, and Guoping Qiu. Visual quality assessment for super-resolved images: Database and method. *Trans. Img. Proc.*, 28(7):3528–3541, July 2019. ISSN 1057-7149. doi: 10.1109/TIP.2019.2898638. URL https://doi.org/10.1109/TIP.2019.2898638.

Fei Zhou, Wei Sheng, Zitao Lu, and Guoping Qiu. video super-resolution quality assessment database (vsr-qad)-1, 2024. URL https://dx.doi.org/10.21227/j87d-w690.

Libo Zhu, Jianze Li, Haotong Qin, Yulun Zhang, Yong Guo, and Xiaokang Yang. PassionSR: Post-training quantization with adaptive scale in one-step diffusion based image super-resolution. In *CVPR*, 2025. doi: 10.1109/CVPR52734.2025.01192.

A Dataset construction

A.1 Source selection

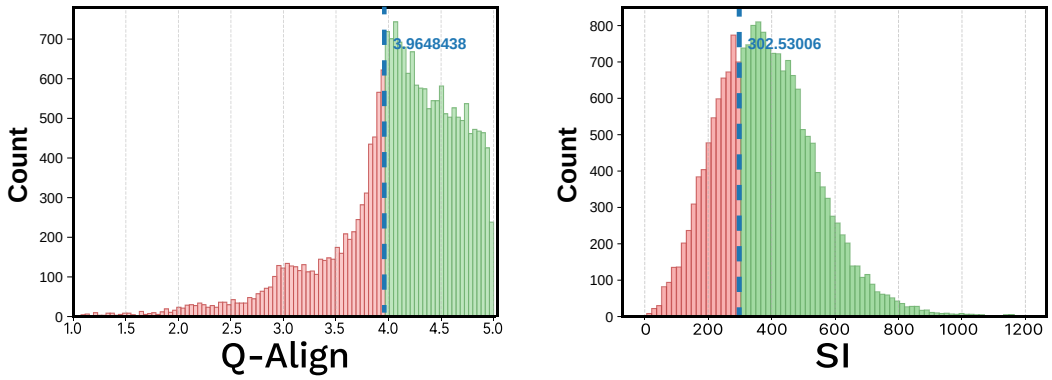
Table 7 lists the features used in source selection.

Table 7: Calculated image features and their descriptions.

Feature	Type	Description
SI (Spatial Information)	Numeric	Characterizes the amount of detail and texture in the scene.
Blockiness Wu & Yuen (1997)	Numeric	Evaluates how pronounced the artificial boundaries between adjacent image blocks are.
SAM Kirillov et al. (2023)	Numeric	Number of objects detected by the model in the image.
Q-Align	Numeric	Expresses not only the visual quality of the image but also its aesthetic component, as it was trained on both tasks.
ARNIQA Agnolucci et al. (2024b)	Embedding	Reflects information about image quality and visual distortions.
CLIP Radford et al. (2021)	Embedding	Expresses semantic information about the image.

The source-selection portion of Figure 2 yields 10,036 selected images from the candidate pool. We collected images from KonIQ-10K, CoCo (train, test, val splits), AVA, FLIVE; resulting in more than 468,000 images. Then, SI (Spatial Information), blockiness Wu & Yuen (1997), SAM Kirillov et al. (2023), Q-Align numeric features, ARNIQA Agnolucci et al. (2024b), and CLIP Radford et al. (2021) embeddings were calculated. SI characterizes the amount of detail and texture in the scene, blockiness evaluates how pronounced the artificial boundaries between adjacent image blocks are. As a feature of the SAM, we took the number of objects that the model detects in the image. Q-Align expresses not only the visual quality of the image, but also its aesthetic component, since it was trained on both of these tasks. If ARNIQA embeddings reflect information about image quality and visual distortions, then CLIP embeddings express semantic information about the image.

We set a goal to take the most diverse images in terms of semantics and the number of objects presented, their dimensional features, natural and artificial distortions, while achieving maximum structural complexity and visual and aesthetic quality in order to evaluate SR methods in a highly competitive environment. We used K-means clustering to cluster images by blockiness, SAM, ARNIQA and CLIP. We identified 20,000 clusters from the entire image set. A medoid image was first determined for each cluster to serve as a representative exemplar. Subsequently, we applied a selection algorithm designed to identify images with the highest Q-Align and Spatial Information (SI) scores. Q-Align is a large multimodal model-based quality metric that has demonstrated state-of-the-art performance on image quality assessment benchmarks Wu et al. (2024a), while SI quantifies the amount of spatial detail in an image based on the standard deviation of Sobel-filtered intensities. We employed a binning procedure to determine a percentage threshold T at which the intersection of the top $T\%$ of images ranked independently by SI and Q-Align most closely approximated the target sample size. Using the resulting quantile, we fixed the Q-Align threshold and, from the remaining candidate pool, selected the maximum SI threshold such that the resulting number of images was greater than or equal to 10,000. This procedure yielded a final set of 10,036 images, characterized by thresholds of SI $= 302.53006$ and Q-Align $= 3.9648438$. The distributions of these features are presented in Figure 4; its Q-Align and SI panels are shown in Figures 4a and 4b.



(a) Distribution of Q-Align feature in medoid images.

(b) Distribution of SI feature in medoid images.

Figure 4: Feature distributions in medoid images.

The 10,036 selected sources yield 80,288 LR images across four conditions and two scales. We apply next types of distortions: gaussian blur with quadratic kernel of 7 pixels and standard deviation 1.2, random noise with $X \sim \mathcal{N}(0, 8^2)$ and JPEG with quality 50%. We also applied 2x and 4x downscaling using cubic interpolation. This gives eight LR images per source.

A.2 SR DATASET CONSTRUCTION

We selected ten representative diffusion-based super-resolution methods (ResShift Yue et al. (2023), SinSR Wang et al. (2024), SeeSR Wu et al. (2024b), SUPIR Yu et al. (2024), PASD, PASD-SDXL Yang et al. (2024), InvSR Yue et al. (2025), CODSR Chen et al. (2026), DiT4SR Duan et al. (2025) and PassionSR Zhu et al. (2025)) to ensure broad coverage of the current design space of diffusion SR models. The benchmark includes methods based on different diffusion formulations, network architectures, conditioning strategies, and training paradigms. Specifically, the selected models range from early diffusion-based restoration approaches (e.g., ResShift) to recent latent diffusion and foundation-model-based methods (e.g., PASD-SDXL, InvSR, CODSR).

Thus, each method was applied to all images. If a model did not support configurable upscaling factors or lacked ×2 or ×4 scaling, we first upscaled the image to the smallest available scale exceeding the target resolution of reference image and then resized the output back to the target resolution using bicubic interpolation. Using this method, we obtained 802,880 SR images.

We also add non-diffusion SR models in our benchmark to compare performance of IQA metrics on images produced by both diffusion and classic SR approaches. Therefore, we included in the dataset images obtained using SwinIR, RealSR, HAT and Real-ESRGAN methods. Using these methods, we obtained 321,152 images.

Since resources for conducting subjective comparison are limited, we need to choose a small subset for comprehensive evaluation of IQA metrics. To include various types of distortions generated by SR methods in the dataset, we use an SR-grounding model Borisov et al. (2026b) that generates segmentation maps for the different distortion types. We calculate class areas of every distortion and cw-ssim metric Sampat et al. (2009) between such maps as a reference and the corresponding SR images. We also calculate Q-Align Wu et al. (2024a) and TOPIQ-FR Chen et al. (2024) because they have good correlation with human assessments, but the lowest correlation with each other according to Borisov et al. (2026a). We applied K-means clustering to select 10,000 medoids from diffusion SR outputs and 4,000 from non-diffusion SR outputs separately. This second selection is based on output features; semantic-cluster coverage after selection has not been established.

B SUBJECTIVE PROTOCOL AND PILOT

We designed the subjective protocol to rate perceived output quality at a feasible annotation cost. NR MOS does not measure content fidelity to the LR input or HR reference. To select the protocol,



Figure 5: Examples of HR–SR image pairs filtered by the highest-standard-deviation criterion during in-lab annotation. This filtering was performed when selecting examples for the assessors' training and examination sets.

we compared pairwise preferences with two forms of direct quality rating: full-reference (FR) and no-reference (NR).

Candidate protocols. In pairwise assessment, assessors select the better of two SR images, and the preferences are aggregated using the Bradley–Terry (BT) model Bradley & Terry (1952). Choosing between two images is generally easier for non-expert assessors than assigning an absolute quality score. We compared SR outputs derived from the same GT image. However, exhaustive pairwise comparison requires a number of judgments that grows quadratically with the number of outputs per reference. Moreover, BT scores estimated independently for different reference images are not directly comparable across references. We therefore used pairwise assessment as a baseline for selecting the rating protocol, rather than for annotating the entire dataset.

Direct quality ratings provide scores on a common scale, with an annotation cost that grows linearly with the number of images. Assessors assign each SR image a score from 1 to 5, with higher values indicating better quality, and the ratings are averaged to obtain its mean opinion score (MOS). Because assigning absolute scores can be difficult for non-experts, we evaluated both of the following protocols:

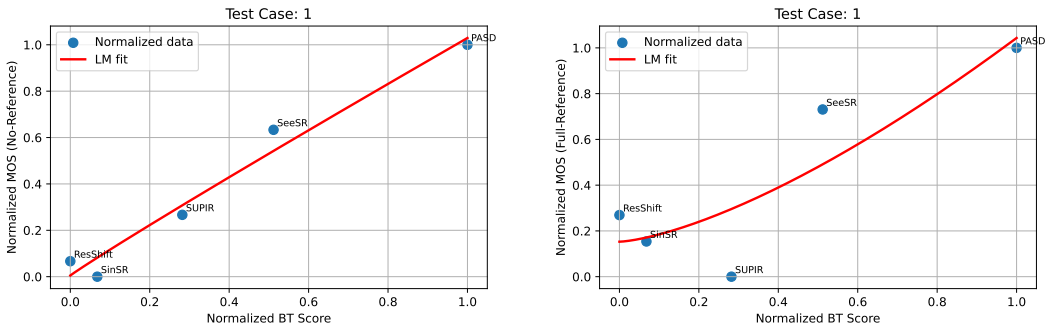
- **Full-reference (FR).** Assessors view both the GT reference and the SR image. The five categories, corresponding to scores 1 through 5, are "very poor quality compared to the reference," "significantly worse than the reference," "slightly worse than the reference," "comparable to the reference," and "better than the reference."
- **No-reference (NR).** Assessors view only the SR image. The five categories, corresponding to scores 1 through 5, are "very poor," "poor," "satisfactory," "good," and "excellent."

Pilot study and quality control. We evaluated all three protocols on a pilot subset of 50 SR images spanning different distortions and quality levels: 5 outputs for each of 10 reference images. We collected at least 20 judgments per image for each MOS protocol and at least 20 judgments per pair for pairwise assessment. Pairwise comparisons were conducted on Subjectify.us, while MOS ratings were collected on Yandex Tasks.

We used control items with predetermined answers to screen out unreliable assessors. Only assessors who successfully completed the training were admitted to the MOS tasks. To construct the control items and training examples, we first conducted an in-lab study on 20 images spanning different distortions and quality levels. 9 experts independently rated each image using the NR protocol. We excluded the 5 images with the highest variance in expert ratings and used the remaining 15 images to prepare the training examples and control items. Examples of images filtered out during the in-lab annotation stage are shown in Figure 5.

Selection of the MOS protocol. Following the strategy in Zhou et al. (2024), we evaluated how well each MOS protocol agreed with the pairwise scores after fitting a nonlinear mapping. We fitted the mapping separately for each reference image and each MOS protocol:

$$(\hat{a}_i, \hat{b}_i, \hat{c}_i) = \arg\min_{a_i, b_i, c_i} \sum_n \left(a_i x_{in}^{b_i} + c_i - z_{in} \right)^2, \tag{1}$$



(1) (2)

Figure 6: LM-fitted relationships between MOS and pairwise Bradley–Terry scores for SR outputs associated with one reference image: (1) no-reference MOS and (2) full-reference MOS.

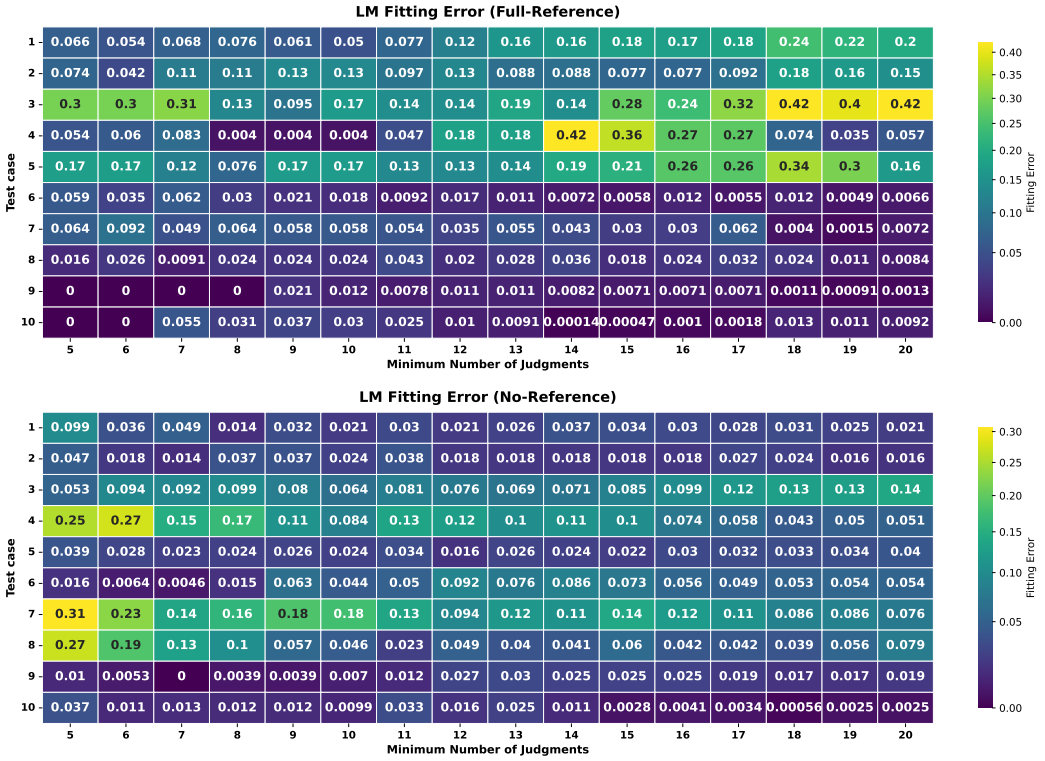


Figure 7: LM-fitting error between MOS and pairwise Bradley–Terry scores as a function of the minimum number of judgments per test case. The upper panel uses full-reference MOS, and the lower panel uses no-reference MOS. The panels use separate color scales.

where x_{in} is the MOS of the nth SR image associated with reference i, and z_{in} is its BT score obtained from pairwise assessment. We estimated the parameters using the Levenberg–Marquardt (LM) algorithm Moré (1978) and assessed agreement using the average fitting error. In this 50-output pilot, NR MOS showed closer fitted agreement with pairwise scores than FR MOS, so we selected the NR protocol for subsequent experiments. This small pilot does not establish general superiority of NR ratings. The fitting objective in equation 1 and Figure 6 illustrate the fitted relationships for the SR outputs associated with one reference image. Figure 7 shows how the LM-fitting error for each set of SR images associated with a single HR reference varies with the number of judgments and the annotation protocol (no-reference or full-reference).

Number of ratings per image. To determine the annotation budget, we compared MOS computed from different numbers of ratings with baseline scores obtained from at least 20 ratings per image. Table 2 reports Spearman's rank-order correlation coefficient (SROCC) and the corresponding 95% confidence intervals. With 13 ratings per image, the pilot table reports SROCC 0.983084 with a 95% confidence interval of $[0.970229, 0.990415]$. We selected a minimum of 13 ratings per image for the final annotation, recognizing that the pilot confidence interval does not establish the precision of each final MOS.

Final protocol. We annotated the entire dataset using NR MOS on the five-category scale described above, collecting at least 13 ratings per image. We retained the training examples and control items used in the pilot study.

C STUDENT ARCHITECTURE AND TRAINING

Q-Align provides a strong NR baseline in Table 3, but its 8.2B parameters make direct adaptation and deployment expensive. Directly fine-tuning the smaller 0.8B Q-ReAlign-mini checkpoint (Q-Future, 2026) on SRIQA-Bench raises its diffusion SRCC from 0.341 to 0.450, indicating that a compact vision–language metric can benefit from SR-specific supervision. As shown in Figure 3, we construct a smaller student from Q-ReAlign-mini and train it in two stages: hierarchical distillation on general IQA data followed by adaptation to SR.

C.1 STUDENT ARCHITECTURE

We initialize the student from the Q-ReAlign-mini 0.8B checkpoint. Like Q-Align, it receives an image and a fixed quality-assessment prompt. The student retains the visual encoder but reduces the language component of the Qwen3.5-VL backbone. Its 24 language modules form six repeated groups of three Gated DeltaNet modules followed by one full-attention module. We remove the first two Gated DeltaNet modules from each group, leaving 12 language modules in total, and restrict the text vocabulary to 30 tokens. These changes reduce the model to 340M parameters.

The model predicts quality through two heads. The token head converts the probabilities of five ordered quality tokens into a score T. The regression head applies a $1024 \to 256$ projection, GELU, dropout, a $256 \to 1$ projection, and a sigmoid to the last hidden state, producing a continuous estimate M. The final prediction is the convex combination $\hat{y} = \alpha T + (1 - \alpha)M$, where α is learned jointly with the model. This preserves Q-Align's discrete quality representation while allowing direct regression to MOS.

C.2 HIERARCHICAL DISTILLATION AND SR ADAPTATION

In the first stage, the unpruned Q-ReAlign-mini teacher provides supervision at several levels of the network. We match every third visual-layer output using mean squared error (MSE), match every corresponding language-layer output using MSE, and minimize the KL divergence between the teacher and student full-attention maps. At the prediction level, we match the teacher score using MSE and its distribution over the five quality tokens using KL divergence.

The student is simultaneously trained against subjective labels from KonIQ-10K (Hosu et al., 2020), KADID-10K (Lin et al., 2019), and SPAQ (Fang et al., 2020), using the dataset splits adopted by Q-Align. The label-based terms comprise MSE between $\hat{y}$ and MOS, an in-batch RankNet loss, and an ordinal loss over the levels $\{0, 0.25, 0.5, 0.75, 1\}$. When MOS falls between two adjacent levels, their target probabilities are obtained by linear interpolation; the ordinal loss is the mean squared distance between their cumulative distribution and that of the five predicted tokens. An additional MSE term encourages the two heads to produce consistent scores. In total, the first-stage objective contains nine terms, weighted as follows: MOS 1.0, ranking 0.3, ordinal 0.2, head consistency 0.05, teacher score 0.4, teacher tokens 0.5, text features 0.02, visual features 0.01, and attention 0.01. We train for eight epochs with an effective batch size of 8 (physical batch size 2) and select the checkpoint with the highest validation SRCC.

The second stage removes the teacher and adapts the student to SR. Its training set combines 990 SRIQA-Bench images with 248 images from the three general-IQA datasets. SRIQA-Bench assigns

MOS to both SR outputs and the corresponding HR images. Its 1,100 rated images were split at the individual-image level. The objective retains the first four label-based terms and adds L2 regularization toward the first-stage checkpoint with weight 10^{-4}. Both stages use AdamW with weight decay 0.01, 5% linear warm-up followed by cosine decay, and learning rates 2×10^{-5}, 2×10^{-6}, and 10^{-4} for the language, vision, and head parameters, respectively. This stage runs for two epochs with an effective batch size of 8 (physical batch size 4). The remaining 110 SRIQA-Bench images are used as validation for checkpoint selection.

C.3 DIFFUSION SR ADAPTATION

Starting from the SRIQA-Bench-adapted checkpoint, we further fine-tune the student on the open part of DISRQAD, following the same fine-tuning strategy used for SRIQA-Bench. We then evaluate it on the held-out private portion. On the 4,000 diffusion SR images in the private split, the model achieves an SRCC of 0.761 and a PLCC of 0.773. On the 1,600 non-diffusion SR images, it achieves an SRCC of 0.890 and a PLCC of 0.879. These results demonstrate that learning to predict perceived quality in the diffusion SR domain is feasible. The remaining performance gap relative to non-diffusion SR indicates that diffusion outputs pose distinct assessment challenges, motivating further domain-specific adaptation, specialized architectures, and systematic study.

D LIMITATIONS

DISRQAD focuses on the perceived visual quality of SR outputs under a no-reference protocol. Faithfulness to the LR input and semantic preservation are complementary assessment targets. The benchmark emphasizes detailed, high-quality source images and controlled single-degradation inputs, so applying its findings to compound real-world degradations calls for separate evaluation.