Learning Where to Look and How to Judge: Resolution-agnostic Image Quality Assessment with Quality-aware Saliency

Supplementary Material

This supplementary material provides: (i) additional details on computational cost and compute-adaptive patch selection; (ii) architectural and implementation details for the IQA-specific saliency module $S(\cdot)$; (iii) additional qualitative saliency visualizations and comparisons with UNISAL; (iv) further experiments on AGIQA-3K; and (v) an ablation study on the training objective.

0.1. Additional Details on Computational Cost and Compute-Adaptive Patch Selection

Unlike most deep learning architectures, ReLIQS does not have a fixed computational cost, since it can operate on a variable number of patches. The only fixed-cost component is the PIE module that produces the perceptual importance map. PIE is intentionally lightweight and operates on a resized version of the input whose short side is 224 pixels, yielding a constant computational cost independent of the original image resolution.

The dominant cost in ReLIQS comes from the patch encoder, and thus scales primarily with the number of selected patches, denoted by k in Sec. 4.4. For low- and mid-resolution images, simple uniform spatial sampling with no overlap or with 50% overlap remains computationally manageable. For high-resolution images, however, this strategy quickly becomes prohibitive, motivating the need for compute-adaptive patch selection.

In our main experiments on the datasets in Tab. 1, we sample a single patch from the scale with short side 224, and uniformly sample patches with 50% overlap from the scales with short side 512 and at the original resolution. If the original image resolution is lower than 512, the original-resolution scale and the 512-scale coincide; in that case, we keep only the original-resolution scale, since ReLIQS can operate with an arbitrary number of scales.

For the UHD dataset in main text Tab. 3, retaining all patches with 50% overlap is computationally infeasible. We therefore always sample a single patch at the 224-scale and use Fig. 1 to decide how many patches to draw from the higher-resolution scales. This figure reports performance as a function of patch count per scale and shows, similar to Fig. 2 in the main text, that performance quickly saturates as the number of patches increases. Utilizing this saturation, we choose a total of $36 + 11 + 1 = 48$ patches per image, allocated from higher to lower scales, respectively, which attains near-maximum performance at substantially reduced computational cost. KonIQ-10K exhibits a similar trend, as shown in Fig. 2.

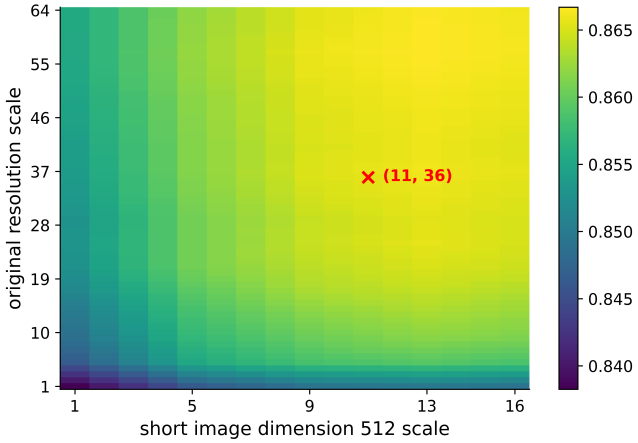

Figure 1. SRCC on UHD test set over patch counts at original resolution and short image dimension 512. We select the (11, 36) patch configuration as a good trade-off between performance and computational cost. With one additional patch sampled at short image dimension 224, this corresponds to a total of 48 patches.

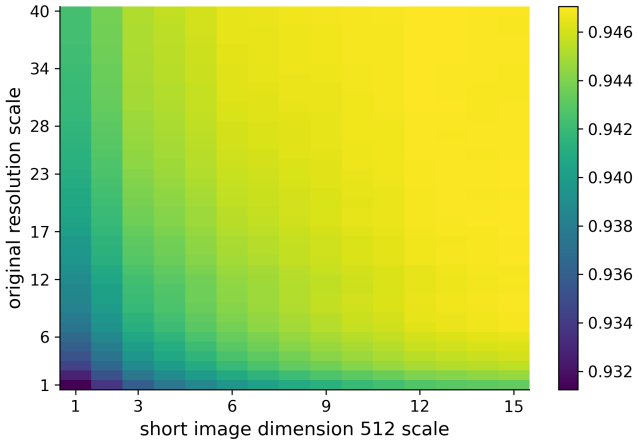

Figure 2. SRCC on KonIQ-10K test set over patch counts at original resolution and short image dimension 512. Results are shown for the first split (max SRCC $= 0.947$) rather than the median over 10 splits (SRCC $= 0.949$).

Although we report our main results in main text Tabs. 1 and 2 using the full 50% overlap patch set for simplicity, the corresponding curves indicate that the number of patches can be reduced substantially for KonIQ images with resolution 768×1024 while maintaining performance close to the reported scores. Overall, these analyses demonstrate that ReLIQS can flexibly trade computational cost for accuracy

Table 1. Single-dataset training ablation study on the training objective (PLCC / SRCC). Median performance over 10 splits is reported.

Setting	Authentic				Synthetic			AIGC
	KonIQ-10K	SPAQ	CLIVE	FLIVE	KADID	CSIQ	LIVE	AGIQA-3K
Trained on: KonIQ-10K								
Only PLCC	**0.958** / 0.941	**0.891** / 0.883	0.891 / 0.856	0.653 / 0.540	0.700 / 0.691	**0.842** / 0.810	**0.879** / 0.880	**0.770** / 0.695
Only MR	0.924 / 0.947	0.878 / 0.888	0.850 / **0.865**	0.601 / 0.548	0.695 / 0.705	0.794 / 0.813	0.860 / 0.871	0.752 / **0.705**
PLCC + MR (uncer.)	**0.958 / 0.949**	**0.891 / 0.894**	**0.892 / 0.865**	**0.654 / 0.549**	**0.701 / 0.707**	**0.842 / 0.818**	**0.879 / 0.894**	0.768 / **0.705**
Trained on: KonIQ-10K, SPAQ, KADID								
PLCC + MR (equal)	0.953 / 0.944	0.932 / 0.928	**0.890** / 0.867	- / -	0.952 / 0.950	0.893 / 0.855	- / -	0.758 / 0.714
PLCC + MR (uncer.)	**0.954 / 0.946**	**0.936 / 0.933**	**0.890 / 0.869**	- / -	**0.955 / 0.953**	**0.894 / 0.857**	- / -	**0.792 / 0.729**

Table 2. AGIQA-3K fine-tuning results on the test set (PLCC / SRCC).

Model	AGIQA-3K
OneAlign + LoRA [6]	0.920 / 0.880
ReLIQS	**0.933 / 0.892**

by adjusting the patch budget per scale.

Unless otherwise noted, we compute GMACs using the `thop` library for both ReLIQS and all baseline models.

0.2. Details on IQA-Specific Saliency

The lightweight network $S(\cdot)$ produces perceptual importance maps, which we refer to as IQA-specific saliency maps, from which relative patch weights are derived. $S(\cdot)$ uses a TinyCLIP ViT-8M visual encoder followed by a shallow convolutional head with two depthwise 5×5 convolutions, each followed by a GeLU nonlinearity, and a final pointwise 1×1 convolution. Perceptual importance maps are obtained by applying a softmax over all spatial locations to normalize the saliency values before patch sampling.

The input to $S(\cdot)$ is the image resized so that its short side is 224 pixels. Since the long side can vary, we interpolate the learned positional embeddings of the TinyCLIP ViT-8M visual encoder to the corresponding spatial grid before tokenization. When the long side is not divisible by the internal patch size (16 pixels), we pad the image, run the encoder, and then crop the padded regions from the dense feature map before feeding it to the convolutional head. Although both $S(\cdot)$ and $E(\cdot)$ use transformer-based CLIP backbones, $E(\cdot)$ outputs only the CLS token, whereas $S(\cdot)$ utilizes the full set of spatial (dense) features.

0.3. Additional Qualitative Results on IQA-Specific Saliency

$S(\cdot)$ outputs normalized saliency maps at a short side of 224 pixels. In the main text Fig. 3, we visualize these maps to highlight regions the model deems perceptually important. For visualization, we take the output of $S(\cdot)$, scale it by 255, round, cast to `uint8`, and replicate the single-channel map across the three RGB channels. The resulting visualizations appear as grayscale maps, where brighter regions correspond to higher predicted importance. In Figs. 4 and 5, we show additional examples of these IQA-specific saliency maps.

Alongside additional qualitative results, we visualize the learning progression of the IQA-specific saliency maps. Specifically, we plot $S(\cdot)$ predictions in the multi-dataset setting, where the model is trained on KonIQ-10K, CLIVE, BID, KADID, CSIQ, and LIVE. As shown in Fig. 3, the maps evolve from diffuse responses to more structured, semantically aligned patterns as training proceeds.

We also compared the learned IQA-specific saliency with conventional visual saliency using the UNISAL model [1]. In the examples shown in Fig. 6, the UNISAL saliency maps appear sharper and more spatially concentrated, whereas the IQA-specific saliency maps tend to highlight broader regions.

0.4. Further Evaluation on AGIQA-3K Dataset

Recall that in main text Tab. 1, with all models trained only on KonIQ-10K, the cross-dataset performance of ReLIQS on AGIQA-3K lags behind Q-Align and DeQA. In the main paper, we attributed this gap to pretraining differences: ReLIQS uses the OpenAI CLIP ViT-B/16 visual encoder, whereas Q-Align and DeQA build on more recent MLLMs that are more likely to have encountered AI-generated content during pretraining.

To probe this, we trained ReLIQS only on AGIQA-3K and compared it against the reported performance of OneAlign with LoRA fine-tuning on AGIQA-3K, where OneAlign is the Q-Align variant jointly trained on multiple IQA and video quality assessment datasets. Unless otherwise noted, we use the same training settings as in the main experiments (optimizer, learning rate schedule, etc.), but restrict supervision to AGIQA-3K. As shown in Tab. 2, ReLIQS outperforms the OneAlign + LoRA variant by +1.3 PLCC and +1.2 SRCC on the official test split. These results indicate that our architecture is not inherently ill-suited to AI-generated image quality assessment and can be effectively repurposed for this setting via fine-tuning. They are also consistent with our view that the cross-dataset gap in

Figure 3. Learning progression of IQA-specific saliency maps for models trained on KonIQ-10K, CLIVE, BID, KADID, CSIQ, and LIVE. Columns show the original image, $S(\cdot)$ output before training, and outputs after 10, 20, 30, 40, and 50 epochs (the best checkpoint for this run is at epoch 51). The maps become progressively more structured and semantically aligned over training.

Table 3. Median PLCC / SRCC of compared IQA models in the single-dataset training setting. Models are trained on KonIQ-10K. **Bold** indicates best performing model and underline indicates the next best.

Method	KonIQ-10K
GrepQ [5]	- / 0.855
QPT [8]	0.941 / 0.927
LODA [7]	0.944 / 0.932
QCN [4]	0.945 / 0.934
SHDIQA [3]	0.948 / 0.937
ATTIQA [2]	0.952 / 0.942
ReLIQS	**0.958 / 0.949**

main text Tab. 1 stems, at least in part, from differences in pretraining exposure and training data rather than fundamental architectural limitations.

0.5. Ablation Study on the Training Objective

Our training objective combines a margin-ranking term $\mathcal{L}_{\text{MR}}$ and a PLCC term $\mathcal{L}_{\text{PLCC}}$, with their relative weights learned via an uncertainty-based scheme. In this scheme, we initialize all loss-specific scales in Eq. (14) by setting the corresponding σ values to 1.

In Tab. 1, we ablate the training objective in both single-dataset and multi-dataset setups. In the single-dataset setting on KonIQ-10K, combining $\mathcal{L}_{\text{PLCC}}$ and $\mathcal{L}_{\text{MR}}$ consistently improves performance over using either term alone. Using only $\mathcal{L}_{\text{PLCC}}$ yields results closer to the combined case, but with noticeably lower SRCC, whereas using only $\mathcal{L}_{\text{MR}}$ attains SRCC comparable to the combined setting while substantially degrading PLCC.

In the multi-dataset setting, where the model is trained on KonIQ-10K, SPAQ, and KADID, uncertainty-based adaptive weighting of $\mathcal{L}_{\text{PLCC}}$ and $\mathcal{L}_{\text{MR}}$ provides a small but consistent performance gain over a simple variant that weights the two losses equally. In the single-dataset setting, however, this gain was negligible, likely because the model converges earlier when trained on a single dataset and we apply early stopping. Overall, the uncertainty-based combination of loss terms yields a slight but consistent performance boost, particularly in the multi-dataset setting.

0.6. Further Comparisons with SOTA IQA Methods

In addition to the baselines in the main text, we compared ReLIQS against several recently proposed SOTA IQA methods. When trained and evaluated on KonIQ-10K, as tabulated in Tab. 3, ReLIQS outperformed ATTIQA [2], SHDIQA [3], QCN [4], and LODA [7] by PLCC/SRCC margins of 0.006/0.007, 0.010/0.012, 0.013/0.015, and 0.014/0.017, respectively.

Figure 4. Learned IQA-specific saliency maps (middle) with input images (left) and overlays (right). Fine-tuning purely on MOS supervision yields strong bias toward semantically salient regions.

Figure 5. Learned IQA-specific saliency maps (middle) with input images (left) and overlays (right). Fine-tuning purely on MOS supervision yields strong bias toward semantically salient regions.

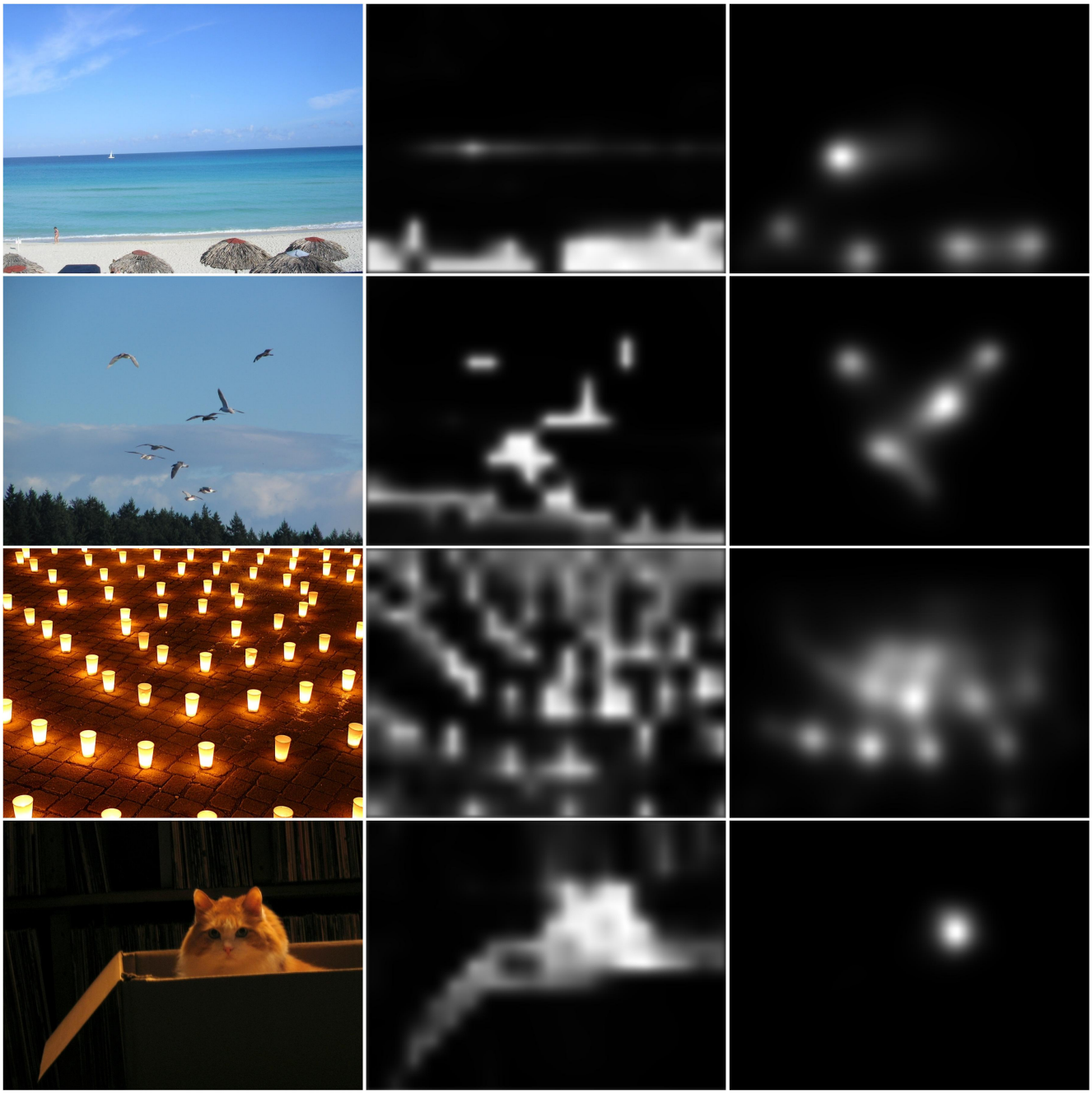

Figure 6. Input images (left), learned IQA-specific saliency maps (middle), and UNISAL visual saliency maps (right). In these examples, the IQA-specific saliency maps highlight broader regions, while the UNISAL maps are more spatially concentrated.

References

[1] Richard Droste, Jianbo Jiao, and J. Alison Noble. Unified image and video saliency modeling. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V*, page 419–435, Berlin, Heidelberg, 2020. Springer-Verlag. 2

[2] Daekyu Kwon, Dongyoung Kim, Sehwan Ki, Younghyun Jo, Hyong-Euk Lee, and Seon Joo Kim. Attiqa: Generalizable image quality feature extractor using attribute-aware pretraining. In *Computer Vision – ACCV 2024*, pages 284–300, Singapore, 2025. Springer Nature Singapore. 3

[3] Xudong Li, Wenjie Nie, Yan Zhang, Runze Hu, Ke Li, Xiawu Zheng, and Liujuan Cao. Distilling spatially-heterogeneous distortion perception for blind image quality assessment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2344–2354, 2025. 3

[4] Nyeong-Ho Shin, Seon-Ho Lee, and Chang-Su Kim. Blind image quality assessment based on geometric order learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12799–12808, 2024. 3

[5] Suhas Srinath, Shankhanil Mitra, Shika Rao, and Rajiv Soundararajan. Learning generalizable perceptual representations for data-efficient no-reference image quality assessment. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 22–31, 2024. 3

[6] 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*, pages 54015–54029. PMLR, 2024. 2

[7] Kangmin Xu, Liang Liao, Jing Xiao, Chaofeng Chen, Haoning Wu, Qiong Yan, and Weisi Lin. Boosting image quality assessment through efficient transformer adaptation with local feature enhancement. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2662–2672, 2024. 3

[8] Kai Zhao, Kun Yuan, Ming Sun, Mading Li, and Xing Wen. Quality-aware pretrained models for blind image quality assessment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22302–22313, 2023. 3