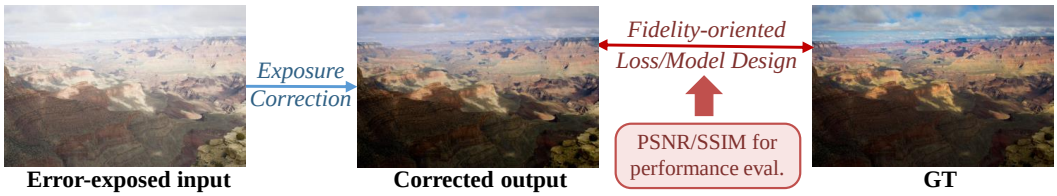
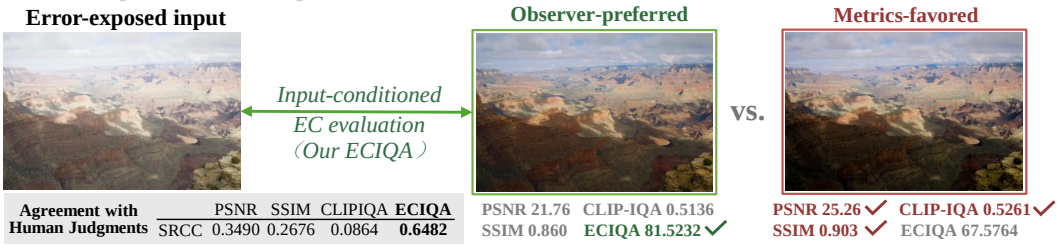
BEYOND GROUND-TRUTH FIDELITY: RETHINKING EXPOSURE CORRECTION THROUGH INPUT-CONDITIONED PERCEPTUAL EVALUATION

Anonymous authors
Paper under double-blind review



(a) PSNR/SSIM-driven ground truth (GT) fidelity may favor exposure corrections that discard local scene information preserved in the input.



Agreement with Human Judgments	PSNR	SSIM	CLIPIQA	ECIQA
SRCC	0.3490	0.2676	0.0864	**0.6482**

(b) Higher PSNR/SSIM may conflict with human preference, while our input-conditioned evaluation (named ECIQA) better matches observer judgments.

Figure 1: Motivation for our input-conditioned evaluation of exposure correction.

ABSTRACT

Prevailing evaluation protocols for exposure correction (EC) prioritize fidelity to a designated ground truth (GT), implicitly treating the GT as a perfect perceptual reference. Yet a mis-exposed input can retain reliable scene content absent from the GT, raising the question of whether higher target fidelity necessarily implies a better correction. To examine this premise, we conduct an extensive human preference study with more than 30 observers, collecting over 60,000 pairwise judgments over 5,200 images. We generate visually diverse candidate corrections using representative EC models trained with different objectives that induce distinct trade-offs in exposure, color, and content preservation. In each trial, observers view the input alongside two candidate corrections and provide pairwise preferences among all three images. The resulting judgments reveal limited agreement between conventional metric rankings and human preferences, motivating us to reformulate EC quality in terms of input-conditioned perceptual improvement. We construct ECIQA-Pref from these judgments and develop ECIQA, a degraded-reference metric that models directed input–candidate relations using global appearance and native-resolution local structural evidence. ECIQA achieves the highest agreement with subjective scores among the evaluated full-, no-, and degraded-reference metrics, reaching 0.6482 SRCC and 0.8169 PLCC. A qualitative pilot study further suggests that ECIQA-guided training can better preserve input-visible structure, providing preliminary support for optimizing EC beyond target fidelity.

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.

1 INTRODUCTION

Improper exposure is a common photographic defect that obscures scene content and reduces perceptual quality. Exposure correction (EC) aims to transform an under- or over-exposed photograph into a visually satisfactory rendition. Existing approaches range from traditional methods based on handcrafted priors to learning-based correction (Pizer et al., 1987; Wang et al., 2013; Afifi et al., 2021; Huang et al., 2022a;b; Li et al., 2024; Huang et al., 2025; Li et al., 2025a; Liu et al., 2024; Wang et al., 2025; Li et al., 2026). Despite this methodological diversity, progress is predominantly assessed by comparing the corrected output with a designated high-quality target using full-reference (FR) criteria such as PSNR (Huynh-Thu & Ghanbari, 2008), SSIM (Wang et al., 2004), LPIPS (Zhang et al., 2018), VIF (Sheikh & Bovik, 2006), and DISTS (Ding et al., 2020). This reliance on target agreement encourages GT-fidelity-oriented training objectives and model designs, raising the question of **whether improved GT fidelity necessarily translates into perceptual benefits.**

This issue is particularly important for EC. As shown in Figure 1a, a brighter exposure may reveal shadow details that are absent from a darker GT. Color biases in expert-retouched targets may also be learned by the model, leading to visible color shifts despite close numerical agreement with the target (Li et al., 2026). Thus, GT fidelity-oriented evaluation may favor outputs that lose reliable details preserved in the input simply because they better match the GT. No-reference (NR) methods, such as CLIP-IQA (Wang et al., 2023), HyperIQA (Su et al., 2020), MANIQA (Yang et al., 2022), and MUSIQ (Ke et al., 2021), avoid the need for a target at inference but assess the corrected image in isolation. They therefore lack direct evidence for judging whether a correction improves perceptual quality relative to the input while preserving useful scene information. These limitations motivate evaluating EC through human judgments that explicitly consider the original capture.

To investigate this issue empirically, we conduct an extensive human preference study with more than 30 observers. In each trial, observers view the original input together with two candidate corrections, allowing the input to serve as scene context when judging the perceptual benefits and losses introduced by correction. Psychophysical evidence reported by Kraft & Brainard (1999) shows that human observers maintain relatively stable surface-color perception under changes in illumination, providing a perceptual basis for comparing differently exposed renditions of the same scene. To obtain diverse correction outcomes, we train representative EC models with different objectives, including multi-exposure guidance that encourages the recovery of local scene structure. The resulting candidates exhibit different trade-offs in exposure, color, and content preservation. Observers provide preferences for all three image pairs, allowing the original input to outrank an unsuccessful correction.

The collected human judgments show limited agreement with conventional metric rankings, as reported in Figure 1b and Table 1. These findings motivate us to reformulate EC quality in terms of input-conditioned perceptual improvement. Based on these images and judgments, we construct ECIQA-Pref, which is, to our knowledge, the first perceptual preference dataset for EC. The dataset comprises 5,200 images and more than 60,000 individual pairwise judgments. Section 3 provides further details on the study protocol and benchmark construction.

Based on the collected data, we further develop ECIQA, a degraded-reference (DR) metric for EC evaluation. Building on vision–language quality assessment (Wang et al., 2023; Li et al., 2025b), ECIQA uses prompt-adapted visual and textual encoders (Jia et al., 2022; Zhou et al., 2022) and jointly models the mis-exposed input and corrected output through cross-attention (Vaswani et al., 2017). Spatially aligned native-resolution tiles are further used to capture local changes in shadow and highlight details. ECIQA achieves the highest agreement with subjective scores among the evaluated FR, NR, and DR metrics, including learned baselines fine-tuned on ECIQA-Pref. A qualitative pilot study further suggests that ECIQA-guided training can better preserve structures visible in the input, providing preliminary support for optimizing EC beyond target fidelity.

The contributions of this work are summarized below.

- We identify the limitation of GT-fidelity-oriented exposure correction, showing that improved agreement with the designated GT does not necessarily imply greater perceptual benefit. We further conduct an extensive human preference study with more than 30 observers to examine this mismatch.

- We construct ECIQA-Pref, to our knowledge the first human-perception-based preference dataset for exposure correction and the first preference benchmark collected under a degraded-reference protocol, containing 5,200 images and over 60,000 individual pairwise judgments.
- We introduce, to our knowledge, the first perceptual preference formulation for EC evaluation and develop ECIQA, which models input–candidate relations through cross-attention, combining with native-resolution local comparison. ECIQA achieves the highest agreement with human preferences among the evaluated FR, NR, and DR metrics.

2 Related Work

2.1 Exposure Correction

EC methods have evolved from handcrafted photometric adjustment (Pizer et al., 1987; Yuan & Sun, 2012; Wang et al., 2013; Zhang et al., 2019) to learning-based restoration. MSEC (Afifi et al., 2021) pioneered a unified framework for correcting both under- and overexposure through Laplacian-pyramid decomposition. Subsequent studies have explored exposure-normalized representations (Nsampi et al., 2021; Huang et al., 2022a;c; 2023; Liu et al., 2024), Fourier-domain processing (Huang et al., 2022b; Li et al., 2025a), decoupled luminance modeling and color transformation (Baek et al., 2023; Li et al., 2024), and CLIP-derived priors for perceptual restoration (Huang et al., 2025; Wang et al., 2025). Some approaches also reduce reliance on paired supervision through self-supervision or fusion-derived pseudo-supervision (Cui et al., 2024; Li et al., 2026).

Despite this diversity in architectures and supervision, EC evaluation still relies predominantly on agreement with the GT. This emphasis on GT fidelity has misdirected method development by rewarding closer target agreement even at the expense of perceptual quality. Therefore, a more perceptually aligned evaluation criterion is needed to guide the EC toward improvements consistent with human preference.

2.2 Image Quality Assessment

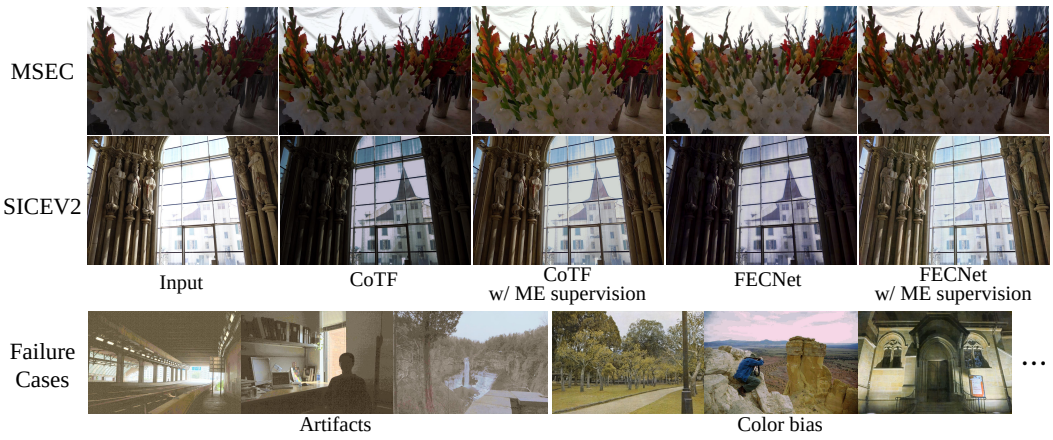
Low-level vision methods routinely report diverse image quality metrics as quantitative evidence of performance. Full-reference image quality assessment (FR-IQA) remains predominant (Huynh-Thu & Ghanbari, 2008; Wang et al., 2004; Zhang et al., 2011; 2018; Sheikh & Bovik, 2006; Ding et al., 2020), with metrics that capture different aspects of pixel fidelity, structural similarity, information fidelity, or perceptual resemblance but consistently quantify agreement with the designated GT and inherit the Perfect Reference Quality Assumption (Chen et al., 2025). No-reference image quality assessment (NR-IQA) requires no GT at inference and assesses an image in isolation (Su et al., 2020; Ke et al., 2021; Yang et al., 2022; Wang et al., 2023). Degraded-reference image quality assessment (DR-IQA) conditions prediction on a degraded observation, commonly treating it as a proxy for pristine content (Lin et al., 2024; Zheng et al., 2021). In contrast, the mis-exposed input in EC should not be treated merely as a proxy for the GT, since it may preserve local detail better than the target. We therefore formulate EC quality in terms of input-conditioned perceptual gain.

The quality priors learned by IQA methods also depend on their training data. TID2013 (Ponomarenko et al., 2015) and KADID-10k (Lin et al., 2019) contain reference-derived synthetic distortions, while PIPAL (Jinjin et al., 2020) extends coverage to outputs of perceptual image-restoration algorithms. EC, however, presents a distinctive challenge, requiring perceptual improvement to be assessed alongside the preservation of input-supported content. We therefore adopt a subjective data collection protocol in which the degraded input serves as the reference for perceptual judgment.

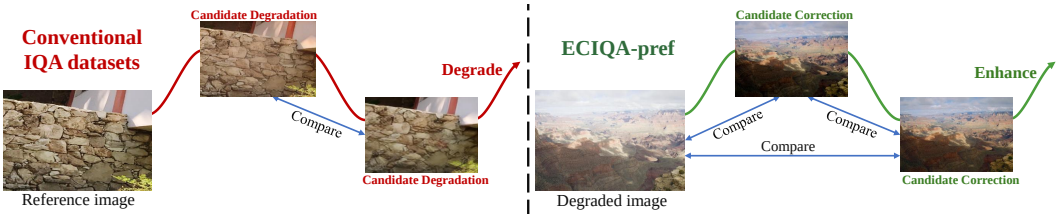
3 The ECIQA-Pref Dataset: A Human Study of Exposure Correction

3.1 Candidate Construction

We draw samples from two widely used EC datasets, MSEC (Afifi et al., 2021) and SICEV2 (Huang et al., 2022a), and train representative EC models (Li et al., 2024; Huang et al., 2022b) to generate



(a) Candidate corrections generated from MSEC (Afifi et al., 2021) and SICEV2 (Huang et al., 2022a) under distinct multi-exposure (ME) supervision strategies exhibit diverse trade-offs in exposure, color, and content preservation. Naturally occurring failure cases, including artifacts and color casts, are retained to broaden the quality spectrum and support robust evaluation across diverse correction outcomes.



(b) For each trial, two candidate corrections are sampled from the set and presented with the corresponding input for three pairwise comparisons. Unlike conventional reference-based IQA protocols, this input-conditioned design treats the input as both scene evidence and a comparison candidate.

Figure 2: Candidate construction and input-conditioned preference protocol in ECIQA-Pref.

candidate corrections. To obtain candidates with distinct trade-offs in exposure, color, and content preservation, we exploit multiple exposures of the same scene to incorporate information not fully represented by the designated GT. Specifically, we retain the original supervision in one setting and incorporate multi-exposure information in the other, using different strategies for the two datasets.

On MSEC, we augment the original training objective with MEF-SSIM (Ma et al., 2015), a perceptual metric for multi-exposure fusion that assesses structural preservation relative to the source exposure sequence. For a candidate Y and exposure sequence $\mathcal{E} = \{E_n\}$, we define the additional loss as:

$$\mathcal{L}_{\mathrm{MEF}}(\mathcal{E}, Y) = 1 - Q_{\mathrm{MEF}}(\mathcal{E}, Y), \tag{1}$$

where Q_{MEF} denotes the multiscale MEF-SSIM score. We set the weight of $\mathcal{L}_{\mathrm{MEF}}$ to half that of the original ℓ_1 loss. Appendix C details how MEF-SSIM constructs the structural representation from the source exposures.

On SICEV2, we instead incorporate multi-exposure information through the training target, using the fused GT from the corresponding SICE scene (Cai et al., 2018) as an alternative to the original middle-exposure GT.

Incorporating multi-exposure information into supervision does not guarantee better perceptual quality, as the resulting candidates are less preferred than their original-supervision counterparts in 23.51% image pairs. Figure 2a presents visual examples of candidate corrections, including naturally occurring failure cases. Retaining these cases broadens ECIQA-Pref to include color shifts, contrast variations, and visible artifacts alongside differences in local detail, allowing the study to examine both perceptual improvements and degradations introduced by correction.

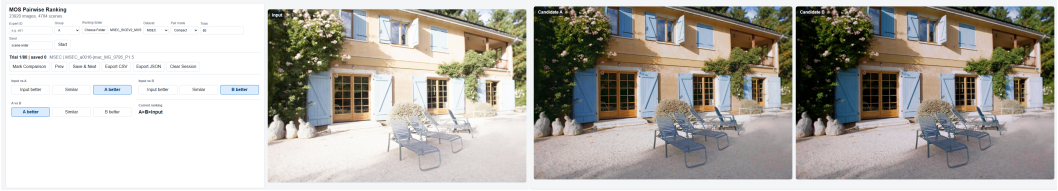


Figure 3: Annotation interface for the human preference study.

3.2 Input-Conditioned Subjective Testing

Building on the psychophysical evidence (Kraft & Brainard, 1999) and input-conditioned formulation discussed above, we design the subjective protocol to assess correction quality in the context of the original capture. As shown in Figure 3, each trial presents the input alongside two candidate corrections and elicits preferences for all three image pairs. Observers may prefer either image in a pair or judge the two to be of similar quality. For each sequence, we query a subset of possible trials, each initially assessed independently by five observers. Aggregated votes determine the queried pairwise preferences, and the remaining relations are inferred through transitive closure. If contradictory preferences prevent a consistent closure, five additional observers assess the affected sequence, and we re-aggregate the votes. Each sequence therefore receives judgments from at least five observers, with additional annotation allocated to ambiguous cases. Throughout the study, observers receive no suggestive cues and base their preferences solely on their own visual perception. In contrast to previous IQA datasets (Figure 2b), ECIQA-Pref is, to our knowledge, the first preference dataset to incorporate a degraded reference during subjective data collection.

3.3 Dataset Properties and Empirical Findings

The study yields ECIQA-Pref, comprising 5,200 images at their native spatial resolution and more than 60,000 individual pairwise judgments from over 30 observers of diverse ages and backgrounds. The annotation campaign spanned three months, with sessions limited to 30 minutes to mitigate visual fatigue. Following the Elo rating framework of PIPAL (Jinjin et al., 2020), we aggregate the pairwise judgments into subjective preference scores. Appendix A provides additional benchmark properties.

The collected judgments reveal limited agreement with conventional metrics, as quantified in Table 1. The input is also preferred to at least one candidate correction in 14.0% of exposure instances, indicating that correction does not invariably improve perceived quality. These findings motivate evaluating perceptual improvement in the context of the input, since neither correction itself nor closer agreement with the GT guarantees a perceptual benefit.

4 The ECIQA Model

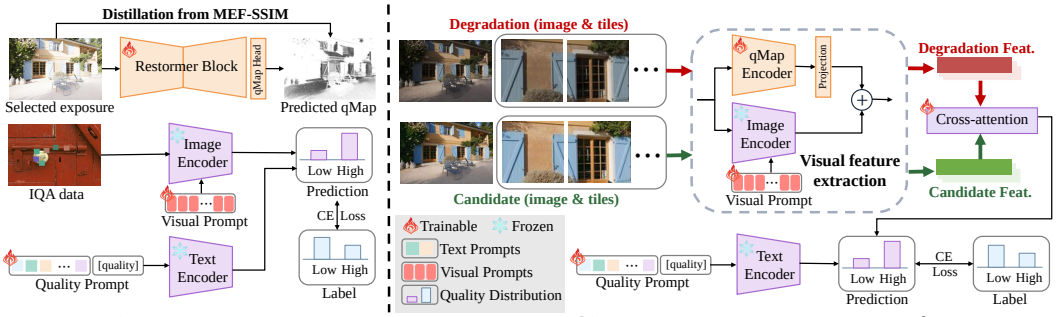
Using the preferences collected in ECIQA-Pref, we develop ECIQA, a degraded-reference metric for assessing exposure correction quality. As shown in Figure 4, the framework builds on vision–language quality assessment (Wang et al., 2023; Li et al., 2025b) and is trained in two stages. Stage I establishes basic quality sensitivity and distills structural cues into the qMap encoder, while Stage II adapts the model to human preferences in ECIQA-Pref. Throughout training, the pretrained CLIP (Radford et al., 2021) visual and textual backbones remain frozen, while the learnable prompts and task-specific modules are optimized according to the objectives of each stage.

4.1 Problem Formulation

Let X denote the mis-exposed input, Y a candidate rendition, and t its scene-relative preference target derived from the judgments in ECIQA-Pref. We learn an input-conditioned predictor:

$$\widehat{t} = f_{\Theta}(X, Y). \tag{2}$$

The input X serves two complementary roles. It first provides scene evidence for assessing appearance changes and content preservation in Y. Meanwhile, X itself is treated as a valid ren-



(a) Stage I: Pretraining (b) Stage II: Training on ECIQA-Pref

Figure 4: ECIQA learns perceptual preferences through joint input–candidate analysis. Stage I initializes quality-aware visual and textual representations and distills structural cues into the qMap encoder. Stage II adapts these representations to ECIQA-Pref, combining global appearance with local detail from native-resolution tiles.

dition and evaluated by the same predictor as $f_\Theta(X, X)$. This score is included together with the scores $\{f_\Theta(X, Y_i)\}$ of all candidate corrections from the same exposure instance, providing a scene-specific anchor for ranking the original capture and its corrected versions in a common preference space. Consequently, the formulation does not assume that every correction is better than its input: an unsuccessful correction can naturally receive a lower score than X. The target t thus represents relative perceptual preference within each exposure instance, rather than an absolute quality level shared across scenes.

4.2 Stage I: Pretraining

In our human study, candidates generated with multi-exposure supervision are preferred to their original-supervision counterparts in 72.50% of comparisons. Motivated by this observation, we pretrain a single-image structural predictor (Zamir et al., 2022) by distilling MEF-SSIM-derived quality maps (Ma et al., 2015). To establish basic sensitivity to perceptual quality before learning EC-specific preferences, we adapt the CLIP visual and textual encoders (Radford et al., 2021) through prompt tuning (Jia et al., 2022; Zhou et al., 2022) to distinguish quality variations across diverse distortions.

4.3 Stage II: Training on ECIQA-Pref

For each input–candidate pair (X, Y), we encode the full image and spatially aligned tiles for both images. CLIP visual features and qMap features are fused for each view. Let $\mathbf{F}_I^{\text{img}}$ denote the full-image features and $\mathbf{F}_I^{(k)}$ the features of the k-th tile for $I \in \{X, Y\}$. We concatenate these features along the token dimension:

$$\mathbf{S}_I = \text{concat}\left(\mathbf{F}_I^{\text{img}}, \mathbf{F}_I^{(1)}, \ldots, \mathbf{F}_I^{(K)}\right), \qquad I \in \{X, Y\}, \tag{3}$$

where K is the number of tiles. The candidate sequence then queries the input sequence through multi-head cross-attention (Vaswani et al., 2017). The corresponding query, key, and value projections are:

$$\mathbf{Q}_Y = W_Q \mathbf{S}_Y, \quad \mathbf{K}_X = W_K \mathbf{S}_X, \quad \mathbf{V}_X = W_V \mathbf{S}_X. \tag{4}$$

Applying $\text{MHA}(\mathbf{Q}_Y, \mathbf{K}_X, \mathbf{V}_X)$ updates the candidate sequence to $\widetilde{\mathbf{S}}_Y$. We concatenate $\widetilde{\mathbf{S}}_Y$ with the input features $\mathbf{S}_X$ and map the resulting features to the relation quality representation $\mathbf{R}$. We then compare this representation with the learnable high- and low-quality text prototypes $\mathbf{e}^{\text{high}}$ and $\mathbf{e}^{\text{low}}$ to obtain the quality score of candidate Y conditioned on input X:

$$\widehat{t}(X, Y) = \frac{\exp\left(\alpha\, \mathbf{R}^{\mathrm{T}} \mathbf{e}^{\text{high}}\right)}{\exp\left(\alpha\, \mathbf{R}^{\mathrm{T}} \mathbf{e}^{\text{high}}\right) + \exp\left(\alpha\, \mathbf{R}^{\mathrm{T}} \mathbf{e}^{\text{low}}\right)}, \tag{5}$$

where α is a learned logit scale. The model is trained with soft cross-entropy against the scene-relative preference targets collected in ECIQA-Pref.

Table 1: Agreement with ECIQA-Pref scores on under-exposed (Under), over-exposed (Over), and all (Overall) test images. † indicates fine-tuning on the ECIQA-Pref dataset. The best and second-best results are shown in **bold** and underlined.

Methods	Under		Over		Overall	
	SRCC ↑	PLCC ↑	SRCC ↑	PLCC ↑	SRCC ↑	PLCC ↑
Full-reference IQA						
PSNR (Huynh-Thu & Ghanbari, 2008)	0.4435	0.5967	0.2651	0.4074	0.3490	0.4953
SSIM (Wang et al., 2004)	0.3216	0.4177	0.2202	0.3488	0.2676	0.3806
FSIM (Zhang et al., 2011)	0.4446	0.6180	0.1770	0.2731	0.3088	0.4505
LPIPS (Zhang et al., 2018)	0.3261	0.3924	0.2319	0.3405	0.2716	0.3577
VIF (Sheikh & Bovik, 2006)	0.2904	0.3374	0.1041	0.1500	0.1633	0.2094
DISTS (Ding et al., 2020)	0.3658	0.4800	0.2419	0.3453	0.2955	0.4032
A-FINE (Chen et al., 2025)	0.1876	0.1520	0.1734	0.1967	0.1746	0.1706
No-reference IQA						
HyperIQA (Su et al., 2020)	0.1969	0.2240	0.1707	0.1704	0.1814	0.1934
HyperIQA†	0.6441	0.7652	0.5422	0.6213	0.5944	0.6915
MUSIQ (Ke et al., 2021)	0.3366	0.4165	0.2062	0.2232	0.2670	0.3190
MUSIQ†	0.5031	0.6887	0.4488	0.5713	0.4666	0.6229
MANIQA (Yang et al., 2022)	0.0800	0.1065	0.1278	0.1464	0.1035	0.1249
MANIQA†	0.6507	0.7562	0.6035	0.6662	0.6275	0.7087
CLIP-IQA (Wang et al., 2023)	0.0921	0.1041	0.0993	0.1067	0.0864	0.0942
CLIP-IQA+†	0.5726	0.8657	0.5117	0.6884	0.5436	0.7747
LIQE (Zhang et al., 2023)	0.3206	0.4068	0.2420	0.2778	0.2722	0.3335
LIQE†	0.6502	0.7978	<u>0.6071</u>	0.6689	<u>0.6293</u>	0.7293
ARNIQA (Agnolucci et al., 2024)	0.2861	0.3842	0.1961	0.2532	0.2351	0.3110
ARNIQA†	0.5738	0.7907	0.5152	0.6639	0.5447	0.7268
TOPIQ (Chen et al., 2024)	0.2645	0.3044	0.2353	0.2505	0.2494	0.2729
TOPIQ†	<u>0.6655</u>	0.8587	0.5502	0.6724	0.6096	0.7642
GRMP-IQA (Li et al., 2025b)	0.2414	0.2674	0.1963	0.2009	0.2124	0.2229
GRMP-IQA†	0.5487	0.8120	0.4436	0.6317	0.5019	0.7236
Degraded-reference IQA						
MEF-SSIM (Ma et al., 2015)	0.4050	0.4828	0.0872	0.2177	0.2083	0.3394
CKDN (Zheng et al., 2021)	-0.0622	-0.1110	0.0113	0.0140	-0.0255	-0.0495
CKDN†	0.6124	<u>0.8845</u>	0.6040	<u>0.7023</u>	0.6162	<u>0.7916</u>
ECIQA (Ours)	**0.6712**	**0.8967**	**0.6195**	**0.7387**	**0.6482**	**0.8169**

5 EXPERIMENTS

5.1 DATASETS

ECIQA-Pref comprises 507 under-exposed and 533 over-exposed instances. We allocate 80% of the instances to training and 20% to testing. The test set is balanced across exposure conditions, containing 104 under-exposed and 104 over-exposed instances.

For qMap pretraining, we use multi-exposure sequences from MSEC (Afifi et al., 2021), SICE (Cai et al., 2018), REED (Wang et al., 2025), HDR-Eye (Nemoto et al., 2015), and MEFLUT (Jiang et al., 2023). The visual and textual prompts are pretrained on TID2013 (Ponomarenko et al., 2015) and KADID-10k (Lin et al., 2019). Only the training splits of these datasets are used for pretraining, whereas ECIQA-Pref draws its inputs from the test splits of MSEC and SICEV2. All images included in ECIQA-Pref are excluded from pretraining.

5.2 IMPLEMENTATION DETAILS

Both full-image views and local tiles are provided to the encoder at a resolution of 224×224. Local tiles are cropped directly from the original-resolution images, with padding applied only at image boundaries. We sample four tiles during training and use all valid tiles during evaluation. ECIQA is fine-tuned for 20 epochs using AdamW (Loshchilov & Hutter, 2019), with learning rates of 1×10^{-5}

Table 2: Effectiveness of input features, qMap features, and native-resolution tiling. All variants use the same training settings as the full ECIQA model.

Variant	SRCC ↑	PLCC ↑
w/o input features	0.5875	0.6892
w/o qMap features	0.6236	0.7753
w/o native-resolution tiling	0.6395	0.8024
ECIQA (full)	**0.6482**	**0.8169**

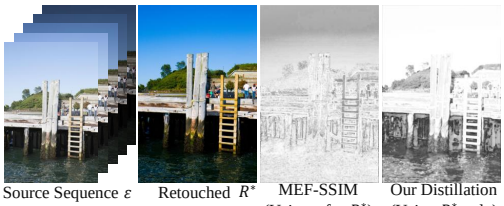


Figure 5: Quality maps from MEF-SSIM and our single-image distillation.

for the learnable prompts and qMap encoder and 1×10^{-4} for newly initialized layers. We use a batch size of one and accumulate gradients over eight steps.

ECIQA is implemented in PyTorch, and the annotation interface is developed using HTML, CSS, and JavaScript. All experiments are conducted on a single NVIDIA RTX 4090 GPU. The ECIQA-Pref dataset, the ECIQA model weights, and associated tools will be made publicly available following the double-blind review process.

5.3 Agreement with Human Preferences

We compare ECIQA with representative FR, NR, and DR metrics in terms of agreement with observer preferences. The FR group includes PSNR (Huynh-Thu & Ghanbari, 2008), SSIM (Wang et al., 2004), FSIM (Zhang et al., 2011), LPIPS (Zhang et al., 2018), VIF (Sheikh & Bovik, 2006), DISTS (Ding et al., 2020), and A-FINE (Chen et al., 2025). The NR group comprises HyperIQA (Su et al., 2020), MUSIQ (Ke et al., 2021), MANIQA (Yang et al., 2022), CLIP-IQA (Wang et al., 2023), LIQE (Zhang et al., 2023), ARNIQA (Agnolucci et al., 2024), TOPIQ (Chen et al., 2024), and GRMP-IQA (Li et al., 2025b). The DR group comprises MEF-SSIM (Ma et al., 2015), which receives the exposure sequence, and CKDN (Zheng et al., 2021), which receives the same input–candidate pair as ECIQA. We additionally fine-tune the learned baselines on ECIQA-Pref using their official implementations and include the resulting models in the comparison.

We report the Spearman rank-order correlation coefficient (SRCC) and Pearson linear correlation coefficient (PLCC). As shown in Table 1, ECIQA achieves the highest SRCC and PLCC on both under- and over-exposed subsets and overall among all evaluated methods. PSNR and SSIM, the most widely used metrics for EC evaluation, exhibit limited agreement with observer preferences, underscoring the limitations of target fidelity as a guide to perceptual improvement. Released NR models likewise show poor agreement, whereas fine-tuning on ECIQA-Pref substantially improves both correlations, suggesting that existing training data do not adequately capture exposure-related quality variations. Among the fine-tuned baselines, CKDN achieves the relatively better agreement with observer preferences, consistent with the value of using the input as a reference for assessing correction quality.

MEF-SSIM exhibits limited agreement with observer preferences despite using the entire exposure sequence as a reference. As illustrated in Figure 5, its reliance on consistency with the source exposures can penalize color deviations. It does not distinguish between perceptually beneficial color adjustments and undesirable color shifts. Agreement with a multi-exposure sequence therefore does not necessarily imply improved perceptual quality. This limitation explains why we retain the original GT fidelity loss alongside MEF-SSIM supervision during candidate generation.

5.4 Ablation Study

To examine the contributions of input context and local evidence to preference prediction, we remove input features, qMap features, or native-resolution tiling one at a time, keeping the remaining architecture, training protocol, and subjective supervision fixed.

Removing input features yields the largest decrease in both correlations, highlighting the importance of the original capture as scene evidence for preference prediction. Removing qMap features also reduces agreement with human preferences, indicating that distilled structural cues complement the

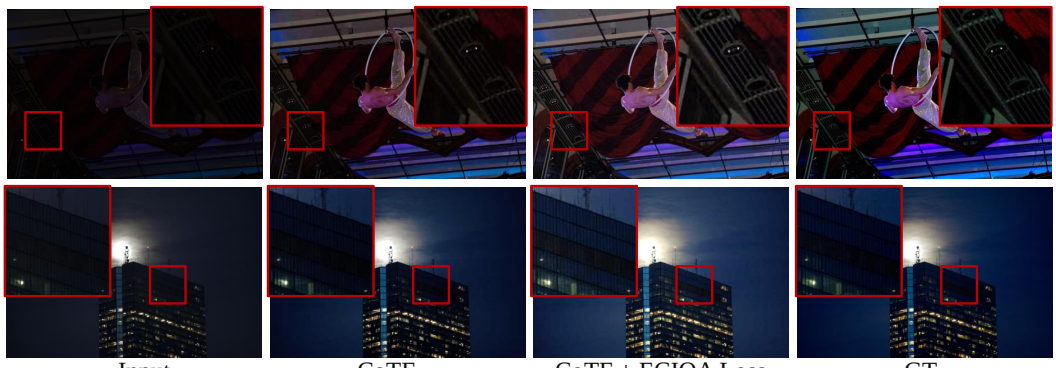


Figure 6: Visual example of ECIQA-guided EC training. Compared with the original CoTF model, the ECIQA-guided variant preserves more scene detail, including structures absent from the GT.

visual representations. The performance decline without native-resolution tiling further supports the value of preserving fine local detail when assessing correction quality. Together, these ablations support combining input context with local structural content when assessing corrections. Additional ablation results can be found in Appendix D.

5.5 A Pilot Study of ECIQA-Guided EC Training

We conduct a pilot study to examine whether the learned preference criterion can guide EC training beyond target fidelity. We augment the original training objective of CoTF (Li et al., 2024) with an ECIQA loss. Let $\overline{Q}(X, Y)$ denote the ECIQA prediction standardized using the mean and standard deviation of the preference targets used to fit the scorer. The additional objective is

$$\mathcal{L}_{ECIQA} = \lambda T \,\mathrm{softplus}\left(-\frac{\overline{Q}(X, Y)}{T}\right), \tag{6}$$

where $\lambda > 0$ weights the ECIQA term and $T > 0$ controls the smoothness of the loss. We empirically set $\lambda = 0.08$ and $T = 10$.

Figure 6 presents two examples in which ECIQA-guided training preserves more scene detail while correcting the overall exposure. These results provide preliminary qualitative support for guiding EC beyond target fidelity through learned perceptual preferences. More visual comparisons of ECIQA-guided EC training can be found in Appendix B.

6 Conclusion and Discussion

Our study identifies a key limitation of GT-fidelity-oriented exposure correction: closer agreement with a designated GT does not necessarily imply greater perceptual benefit. Human preference judgments reveal clear discrepancies with conventional metric rankings, motivating input-conditioned perceptual evaluation. Based on these judgments, we construct ECIQA-Pref and develop ECIQA, which models directed input–candidate relations together with global and native-resolution local evidence. The strong agreement of ECIQA with human preferences supports the value of retaining the original input as scene context, while a qualitative pilot study suggests its potential for guiding EC beyond target fidelity.

ECIQA-Pref currently covers two source datasets and two EC architectures, and the split is not strictly source-content disjoint. Its Elo scores are scene-relative, while the MEF-SSIM-based qMap may be less sensitive to chromatic or semantic failures. Future work will expand correction types and capture conditions, establish stricter evaluation splits, and further study input-conditioned preference supervision for EC learning.

AI USE STATEMENT

In this work, we used generative AI tools to assist with coding, manuscript drafting and language revision. We have not used generative AI tools for figures generation. All figures in this paper were prepared manually by the authors. We have reviewed all AI-assisted work, including the code and manuscript text. We take full responsibility for the final manuscript, code, and reported results, including content produced with the assistance of generative AI.

ETHICS STATEMENT

This work involves a human preference study. All participants provided informed consent and agreed to the public release of their annotation records.

REFERENCES

Mahmoud Afifi, Konstantinos G Derpanis, Bjorn Ommer, and Michael S Brown. Learning multi-scale photo exposure correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9157–9167, 2021.

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. IEEE, 2024.

Jong-Hyeon Baek, DaeHyun Kim, Su-Min Choi, Hyo-jun Lee, Hanul Kim, and Yeong Jun Koh. Luminance-aware color transform for multiple exposure correction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6156–6165, 2023.

Jianrui Cai, Shuhang Gu, and Lei Zhang. Learning a deep single image contrast enhancer from multi-exposure images. *IEEE transactions on image processing*, 27(4):2049–2062, 2018.

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.

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. IEEE, 2025.

Ruodai Cui, Li Niu, and Guosheng Hu. Unsupervised exposure correction. In *European Conference on Computer Vision*, pp. 252–268. Springer, 2024.

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, 2020.

Jie Huang, Yajing Liu, Xueyang Fu, Man Zhou, Yang Wang, Feng Zhao, and Zhiwei Xiong. Exposure normalization and compensation for multiple-exposure correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6043–6052, 2022a.

Jie Huang, Yajing Liu, Feng Zhao, Keyu Yan, Jinghao Zhang, Yukun Huang, Man Zhou, and Zhiwei Xiong. Deep fourier-based exposure correction network with spatial-frequency interaction. In *European Conference on Computer Vision*, pp. 163–180. Springer, 2022b.

Jie Huang, Man Zhou, Yajing Liu, Mingde Yao, Feng Zhao, and Zhiwei Xiong. Exposure-consistency representation learning for exposure correction. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 6309–6317, 2022c.

Jie Huang, Feng Zhao, Man Zhou, Jie Xiao, Naishan Zheng, Kaiwen Zheng, and Zhiwei Xiong. Learning sample relationship for exposure correction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9904–9913, 2023.

Xiang Huang, Qing Zhang, Jian-Fang Hu, and Wei-Shi Zheng. Clip-restorex: restore image structure and perception in exposure correction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 3760–3768, 2025.

Quan Huynh-Thu and Mohammed Ghanbari. Scope of validity of PSNR in image/video quality assessment. *Electronics Letters*, 44(13):800–801, 2008. doi: 10.1049/el:20080522.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European conference on computer vision*, pp. 709–727. Springer, 2022.

Ting Jiang, Chuan Wang, Xinpeng Li, Ru Li, Haoqiang Fan, and Shuaicheng Liu. Meflut: Unsupervised 1d lookup tables for multi-exposure image fusion. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10508–10517. IEEE, 2023.

Gu Jinjin, Cai Haoming, Chen Haoyu, Ye Xiaoxing, Jimmy S Ren, and Dong Chao. Pipal: a large-scale image quality assessment dataset for perceptual image restoration. In *European conference on computer vision*, pp. 633–651. Springer, 2020.

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. IEEE, 2021.

James M Kraft and David H Brainard. Mechanisms of color constancy under nearly natural viewing. *Proceedings of the National Academy of Sciences*, 96(1):307–312, 1999.

Ao Li, Zhenyu Wang, Chen Chen, Tao Huang, Fangfang Wu, and Weisheng Dong. Loopexpose: An unsupervised framework for arbitrary-length exposure correction. *IEEE Transactions on Image Processing*, 2026.

Gehui Li, Bin Chen, Chen Zhao, Lei Zhang, and Jian Zhang. Osmamba: Omnidirectional spectral mamba with dual-domain prior generator for exposure correction. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7480–7490. IEEE, 2025a.

Xudong Li, Zihao Huang, Yan Zhang, Yunhang Shen, Ke Li, Xiawu Zheng, Liujuan Cao, and Rongrong Ji. Few-shot image quality assessment via adaptation of vision-language models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10442–10452. IEEE, 2025b.

Ziwen Li, Feng Zhang, Meng Cao, Jinpu Zhang, Yuanjie Shao, Yuehuan Wang, and Nong Sang. Real-time exposure correction via collaborative transformations and adaptive sampling. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2984–2994. IEEE, 2024.

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

Xinying Lin, Xuyang Liu, Hong Yang, Xiaohai He, and Honggang Chen. Perception-and fidelity-aware reduced-reference super-resolution image quality assessment. *IEEE Transactions on Broadcasting*, 71(1):323–333, 2024.

Jin Liu, Huiyuan Fu, Chuanming Wang, and Huadong Ma. Region-aware exposure consistency network for mixed exposure correction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 3648–3656, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Kede Ma, Kai Zeng, and Zhou Wang. Perceptual quality assessment for multi-exposure image fusion. *IEEE transactions on image processing*, 24(11):3345–3356, 2015.

Hiromi Nemoto, Pavel Korshunov, Philippe Hanhart, and Touradj Ebrahimi. Visual attention in ldr and hdr images. In *9th International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM)*, 2015.

Ntumba Elie Nsampi, Zhongyun Hu, and Qing Wang. Learning exposure correction via consistency modeling. In *BMVC*, pp. 12, 2021.

Stephen M Pizer, E Philip Amburn, John D Austin, Robert Cromartie, Ari Geselowitz, Trey Greer, Bart ter Haar Romeny, John B Zimmerman, and Karel Zuiderveld. Adaptive histogram equalization and its variations. *Computer vision, graphics, and image processing*, 39(3):355–368, 1987.

Nikolay Ponomarenko, Lina Jin, Oleg Ieremeiev, Vladimir Lukin, Karen Egiazarian, Jaakko Astola, Benoit Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, and C.-C. Jay Kuo. Image database TID2013: Peculiarities, results and perspectives. *Signal Processing: Image Communication*, 30: 57–77, 2015. doi: 10.1016/j.image.2014.10.009.

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 Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021.

Hamid R Sheikh and Alan C Bovik. Image information and visual quality. *IEEE Transactions on image processing*, 15(2):430–444, 2006.

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 *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3664–3673. IEEE, 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Bo Wang, Huiyuan Fu, Zhiye Huang, Siru Zhang, Xin Wang, and Huadong Ma. From abyssal darkness to blinding glare: A benchmark on extreme exposure correction in real world. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7666–7675. IEEE, 2025.

Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 2555–2563, 2023.

Shuhang Wang, Jin Zheng, Hai-Miao Hu, and Bo Li. Naturalness preserved enhancement algorithm for non-uniform illumination images. *IEEE transactions on image processing*, 22(9):3538–3548, 2013.

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.

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 *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1190–1199. IEEE, 2022.

Lu Yuan and Jian Sun. Automatic exposure correction of consumer photographs. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part IV 12*, pp. 771–785. Springer, 2012.

Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5728–5739, 2022.

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.

Qing Zhang, Yongwei Nie, and Wei-Shi Zheng. Dual illumination estimation for robust exposure correction. In *Computer graphics forum*, volume 38, pp. 243–252. Wiley Online Library, 2019.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 586–595. IEEE, 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 *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14071–14081. IEEE, 2023.

Heliang Zheng, Huan Yang, Jianlong Fu, Zheng-Jun Zha, and Jiebo Luo. Learning conditional knowledge distillation for degraded-reference image quality assessment. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10222–10231. IEEE, 2021.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International journal of computer vision*, 130(9):2337–2348, 2022.

A Additional Details of ECIQA-Pref

A.1 Candidate Refinement

CoTF trained with additional MEF-SSIM supervision (CoTF-ME) exhibits a systematic tendency toward green color casts. This tendency may be related to the luminance-based formulation of MEF-SSIM (Ma et al., 2015), which evaluates structure using the Y channel in YUV color space. The green channel carries the largest weight in the RGB-to-Y conversion, making the luminance signal more sensitive to variations in this channel. We therefore apply color correction to all CoTF-ME candidates and replace their uncorrected versions in ECIQA-Pref. By combining well-preserved local structure with improved color rendition, these refined candidates are intended to approach the best perceptual quality attainable from each input.

A.2 Benchmark Composition and Preference Aggregation

ECIQA-Pref comprises 1,040 exposure instances, including 980 from MSEC and 60 from SICEV2. All images are retained at their native spatial resolution. The annotation campaign spanned three months, with sessions limited to 30 minutes to mitigate visual fatigue.

The reported total of more than 60,000 individual pairwise judgments counts each observer's directly recorded decision. We aggregate the resulting pairwise preferences into Elo ratings independently for each exposure instance, treating the input and its four corrections as equally eligible candidates. For a pair of renditions (i, j), the outcome $o_{ij} \in \{1, \frac{1}{2}, 0\}$ indicates a preference for i, similar perceived quality, or a preference for j, respectively. All five ratings are initialized to $1{,}500$. Given the current ratings ξ_i and ξ_j, the expected outcome and rating update are

$$\pi_{ij} = \frac{1}{1 + 10^{(\xi_j - \xi_i)/400}}, \qquad \xi_i \leftarrow \xi_i + K(o_{ij} - \pi_{ij}), \qquad K = 32. \tag{7}$$

Rendition j receives the opposite rating increment, with both updates computed from the same pre-update ratings. Vote-aggregated preferences and relations inferred through transitive closure follow the same update rule.

To reduce sensitivity to comparison order, we average the final ratings over 1,000 single-pass runs with deterministically shuffled comparison orders, resetting all ratings before each run. The averaged ratings are then min–max normalized to $[0, 100]$ within each exposure instance. If all averaged ratings are equal, each image receives a score of 50. These scores represent relative perceptual preferences among renditions of the same input, rather than a globally calibrated quality scale. The input is assigned no predefined quality level, and the designated GT does not participate in score aggregation.

B Additional Qualitative Results of ECIQA-Guided EC Training

We further examine the effect of ECIQA-guided training on color rendition and the preservation of input-visible structure, using the original CoTF model as the comparison baseline.

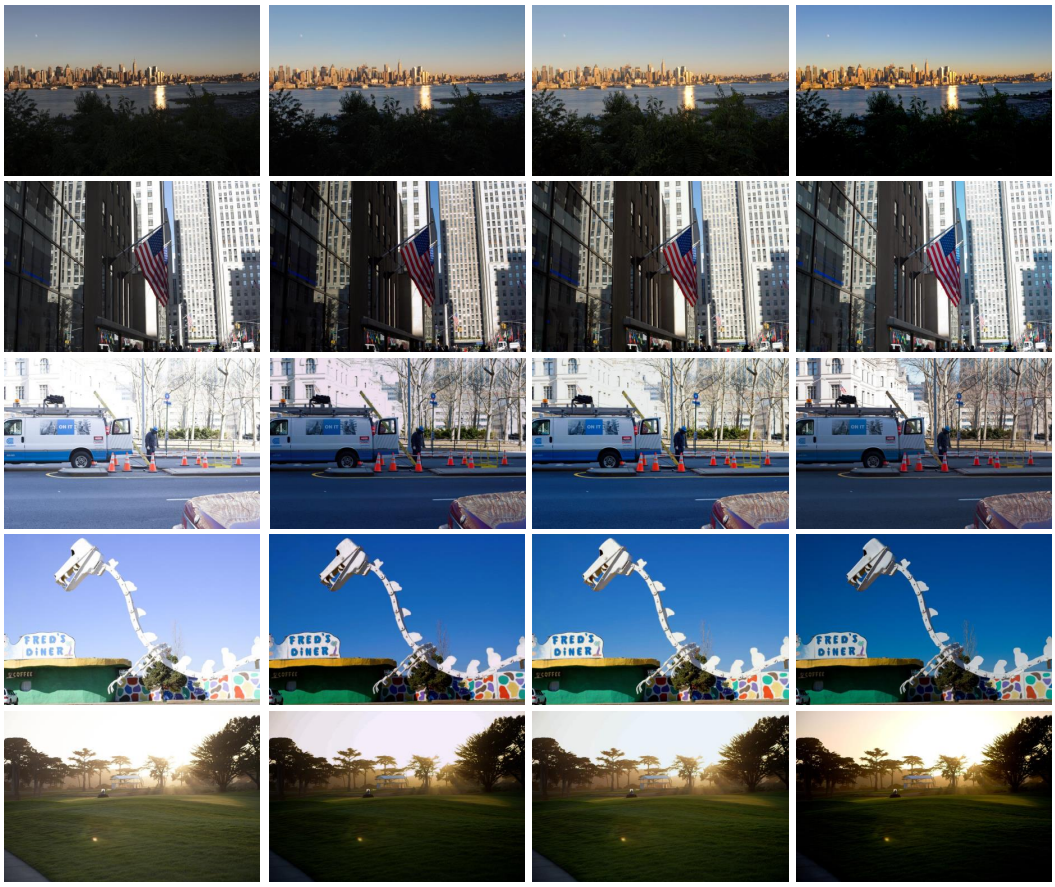


Figure 7: Additional visual comparisons between the original CoTF model and its ECIQA-guided variant.

C MEF-SSIM Supervision and Structural Pretraining

C.1 MEF-SSIM Formulation

MEF-SSIM (Ma et al., 2015) evaluates a fused image through local contrast and structure comparisons with the source exposures. We follow its patch decomposition and quality formulation below.

Let $\mathbf{x}_n$ denote a luminance patch from exposure $E_n \in \mathcal{E}$ and $\mathbf{y}$ the corresponding patch from candidate Y. With mean intensity μ_n and mean-removed patch $\widetilde{\mathbf{x}}_n = \mathbf{x}_n - \mu_n$, each source patch is decomposed as

$$\mathbf{x}_n = c_n \mathbf{s}_n + \mu_n, \qquad c_n = \|\widetilde{\mathbf{x}}_n\|_2, \qquad \mathbf{s}_n = \frac{\widetilde{\mathbf{x}}_n}{c_n}, \tag{8}$$

where c_n measures local contrast and $\mathbf{s}_n$ describes the unit-length structural direction. The local comparison omits the luminance term because the source patches have different exposure levels.

The desired contrast $\widehat{c}$ is the largest contrast among the source patches, while the desired structure $\widehat{\mathbf{s}}$ combines their structural directions. The resulting target patch is

$$\widehat{c} = \max_n c_n, \qquad \widehat{\mathbf{s}} = \frac{\sum_n w_n \mathbf{s}_n}{\|\sum_n w_n \mathbf{s}_n\|_2}, \qquad \widehat{\mathbf{x}} = \widehat{c}\widehat{\mathbf{s}}, \tag{9}$$

where $w_n = c_n^p$ and the exponent p adapts to structural consistency across exposures, following the original formulation. This construction favors visible local detail while constraining the target contrast to that available in the source sequence.

The candidate patch is then evaluated using an SSIM-style contrast–structure comparison,

$$S(\{\mathbf{x}_n\}, \mathbf{y}) = \frac{2\sigma_{\widehat{\mathbf{x}}\mathbf{y}} + C}{\sigma_{\widehat{\mathbf{x}}}^2 + \sigma_{\mathbf{y}}^2 + C}, \tag{10}$$

where $\sigma_{\widehat{\mathbf{x}}\mathbf{y}}$ is the local covariance, $\sigma_{\widehat{\mathbf{x}}}^2$ and $\sigma_{\mathbf{y}}^2$ are local variances, and C is a positive stabilizing constant. Spatial averaging produces a quality score at each scale, and a weighted geometric mean across scales yields Q_{MEF}. Its complement defines the supervision loss in Equation 1.

C.2 Distillation Objective and Encoder Transfer

We distill MEF-SSIM-derived structural quality maps into a single-image predictor (Zamir et al., 2022). Let M denote the teacher map computed using the corresponding exposure sequence and $\widehat{M}$ the prediction from the individual RGB image. We optimize

$$\mathcal{L}_{\mathrm{distill}} = \|\widehat{M} - M\|_1 + 0.1\,\|\nabla\widehat{M} - \nabla M\|_1. \tag{11}$$

The map term is the mean absolute prediction error. The gradient term sums the mean absolute errors of horizontal and vertical first-order finite differences, encouraging agreement in local spatial variation as well as map values.

After pretraining, we retain the four encoder stages as the qMap feature extractor for ECIQA. The decoder and map-prediction head are used only for distillation. The encoder is shared between the input and candidate and is subsequently adapted on ECIQA-Pref, as described in the main text. Exposure sequences are required to construct the teacher maps during pretraining, whereas ECIQA extracts structural features from each individual image at inference.

D Additional Ablation Study

Table 3 reports additional results for freezing the qMap encoder, fine-tuning the CLIP backbone, and increasing the number of training tiles from four to sixteen. Freezing qMap or using sixteen tiles yields a higher SRCC but a lower PLCC than the full ECIQA model. The full model retains the highest PLCC among these configurations.

Table 3: Additional comparisons of qMap adaptation, CLIP backbone fine-tuning, and the number of training tiles. The best result in each column is shown in **bold**.

Configuration	SRCC ↑	PLCC ↑
Frozen qMap encoder	0.6534	0.8014
Fine-tuned CLIP backbone	0.6432	0.8110
16 training tiles	**0.6571**	0.8011
ECIQA (full)	0.6482	**0.8169**