MetaRanker: Human-in-the-loop Active Ranking for Metalens Image Quality

Yujin Park
Hanyang University
Seoul, Republic of Korea
yujin1019a@hanyang.ac.kr

Haejun Chung*
Hanyang University
Seoul, Republic of Korea
haejun@hanyang.ac.kr

Ikbeom Jang*
Hankuk University of Foreign Studies
Yongin, Republic of Korea
ijang@hufs.ac.kr

Abstract

Image quality in modern imaging systems emerges from the coupled effects of the sensor, optics, and computational reconstruction. Ultra-thin metalenses offer a path toward substantial miniaturization, but practical designs often exhibit pronounced chromatic and field-dependent aberrations that decouple standard fidelity metrics (e.g., PSNR, SSIM) from human preference and downstream utility—a phenomenon we term *recognizability collapse*. We introduce MetaRanker, a human-in-the-loop active ranking framework that formalizes metalens image quality in terms of semantic interpretability, defined as the degree to which humans can reliably recognize objects and structures in the presence of optical artifacts. MetaRanker combines a probabilistic preference model with uncertainty-aware query selection and leverages vision–language models (VLMs) to provide lightweight semantic priors. These priors are used only to guide the sampling of informative comparisons; human judgments remain the primary supervision signal throughout. Crucially, the system is self-correcting. Across real-world and synthetic metalens datasets, MetaRanker reduces the annotation cost by approximately 80% relative to exhaustive evaluation while achieving the best rank-aggregated inter-rater reliability. Our results confirm that standard metrics fail to capture interpretability in this domain, positioning MetaRanker as a practical step toward perceptually grounded metalens evaluation and future co-design.

CCS Concepts

• **Applied computing** → **Physical sciences and engineering**; • **Information systems** → *Rank aggregation*; *Wrappers (data mining)*; • **Human-centered computing** → **Human computer interaction (HCI)**.

Keywords

Metalens, Image Quality Assessment, Human-in-the-loop Annotation, Pairwise Ranking, Human-AI Collaboration, Active Learning

ACM Reference Format:
Yujin Park, Haejun Chung, and Ikbeom Jang. 2026. MetaRanker: Human-in-the-loop Active Ranking for Metalens Image Quality. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (KDD '26), August 09–13, 2026, Jeju Island, Republic of Korea.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3770855.3818912

*Corresponding author.



This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.
KDD '26, Jeju Island, Republic of Korea
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2259-2/2026/08
https://doi.org/10.1145/3770855.3818912

1 Introduction

Modern imaging systems have evolved through advances in sensors and optics, yet a critical gap remains between physical fidelity and perceptual quality. While neural image signal processing (ISP) has bridged this gap for conventional cameras, the misalignment between distortion metrics (e.g., PSNR) and human utility is particularly acute for metalenses. Despite offering a scalable, planar form factor, metalenses suffer from severe chromatic and off-axis aberrations [17, 40]. Current reconstruction pipelines typically rely on proxy objectives that fail to capture *semantic interpretability*—whether a human can reliably recognize objects amidst these unique artifacts. We observe a "recognizability collapse" where standard metrics correlate poorly with actual human visual task performance. [19, 33]

Bridging this gap requires treating human judgment as the ground-truth signal. However, collecting pairwise human preferences at scale is prohibitively expensive ($O(N^2)$), creating a significant data bottleneck for AI-driven scientific discovery. For example, even a modest set of $N = 30$ images requires $N(N-1)/2 = 435$ exhaustive pairwise comparisons, whereas our protocol uses a fixed budget of only $B = 90$ comparisons per session. To iterate effectively, metalens research needs a method to acquire high-quality preference signals under strict budget constraints.

To address this, we introduce **MetaRanker**, a budget-efficient active ranking framework. MetaRanker defines image quality via semantic interpretability and combines a probabilistic rating model with an uncertainty-aware acquisition strategy to prioritize informative decision boundaries. To mitigate cold-start inefficiency, we utilize a vision-language prior as a sampling guide—biasing early candidates without treating the prior as ground truth. Validated on the DRMI and Metaformer datasets, our approach achieves superior inter-rater reliability with a fraction of the annotation effort. Our contributions are:

- We introduce **MetaRanker**, a human-in-the-loop active ranking framework for metalens image quality.
- We propose an **uncertainty-aware acquisition function** to obtain human-preference labels only for decision-relevant pairs.
- We incorporate a **VLM-based semantic prior** to guide early sampling, effectively reducing cold-start waste without compromising ground-truth integrity.
- We demonstrate **budget-efficient performance**, achieving reliable preference supervision under limited budgets across diverse metalens datasets.

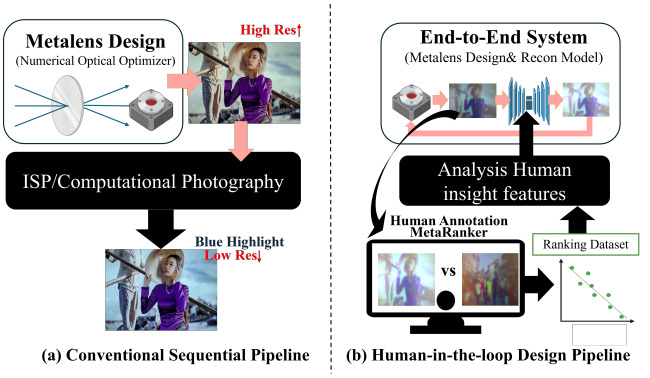

Figure 1: Objective misalignment across the imaging pipeline. Proxy metrics can improve while human interpretability degrades; our ranking annotations enable a human-aligned signal that can be integrated into future closed-loop co-design.

2 Related Work

Metalens Imaging and the Perception–Distortion Trade-off. Metalenses implement a desired wavefront by patterning arrays of subwavelength resonators (meta-atoms), enabling an ultra-thin, planar form factor. However, because a metalens must jointly provide optical power and dispersion control, it is subject to more stringent bandwidth, aperture, and efficiency constraints than conventional multi-element refractive lens assemblies, which distribute aberration correction across multiple surfaces. In particular, enforcing the required phase profile over a broad, continuous spectral band (e.g., the visible) at practical apertures (e.g., millimeter to centimeter scale) generally demands large, spatially varying group delays that are difficult to realize within a thin metasurface and are bounded by fundamental delay–bandwidth and efficiency limits.

As a result, many designs either accept a wavelength-dependent focal length, as in diffractive optics, or attempt to maintain an approximately wavelength-invariant focal length over a limited band via dispersion engineering, often at the cost of reduced bandwidth, numerical aperture, or focusing efficiency. The former inherently introduces strong *chromatic aberration* through wavelength-dependent defocus. Moreover, due to field-angle sensitivity and limited degrees of freedom in a single surface, practical metalenses can exhibit substantial *off-axis aberrations* (e.g., coma and astigmatism), producing spatially varying blur that worsens toward the image periphery. Consequently, practical metalens cameras often rely on computational reconstruction to recover semantically relevant details, making the end-to-end pipeline (optics → sensor → reconstruction) the appropriate unit of evaluation [11, 16].

Recent approaches have advanced this restoration front, ranging from deep end-to-end learning [30] to transformer-based aberration correction (e.g., MetaFormer) [18] and spatially varying deconvolution (e.g., EigenCWD) [40]. However, a key lesson from computational photography is that optimizing for distortion-based fidelity (e.g., PSNR) does not necessarily enhance perceptual quality, a phenomenon formalized as the perception–distortion trade-off [5]. Learned metrics like LPIPS [41] and neural ISP studies [13, 29] further demonstrate that deep feature consistency aligns better with human preference than pixel-wise error. Our work bridges this gap in the metalens domain by addressing the *data bottleneck* in human-grounded supervision: we introduce a budget-efficient ranking pipeline that targets semantic interpretability and can provide human-aligned supervision for future reconstruction or design optimization.



Figure 2: Metric–perception misalignment. Left images yield higher SSIM despite severe perceptual artifacts (e.g., chromatic aberration), illustrating that standard fidelity metrics can misalign with human-recognized image quality. Examples are adapted from DRMI [30], whose clean scenes are derived from DIV2K datasets [1].

Subjective IQA and Pairwise Judgments. Subjective image quality assessment (IQA) is typically conducted via controlled protocols that collect human opinions under standardized viewing conditions. While mean opinion score (MOS) [23] is common, pairwise comparisons are often preferred when absolute scoring is unreliable due to scale ambiguity or cognitive load. [37] Pairwise judgments can be converted into global rankings through probabilistic models such as Bradley–Terry [6] or Thurstone-style formulations, enabling robust aggregation under rater noise.

To further reduce the burden of data collection, recent works have explored progressive boosting [20, 38] and sorting-based protocols that lower comparison complexity to $O(n \log n)$ [14, 22]. While hybrid approaches like EZ-Sort [27] optimize this via zero-shot priors, such models are often brittle under domain shift, and logarithmic scaling remains prohibitive for large datasets. This necessitates budget-aware, uncertainty-driven active preference learning, particularly in noisy domains like metalens imaging; unlike fixed sorting schedules, this approach dynamically prioritizes only the most informative comparisons to maximize efficiency under a strict budget.

Ranking Models with Uncertainty. Pairwise preference learning is closely related to classical paired-comparison models and modern skill-rating systems. In online settings, rating systems such as Elo and the Glicko [8, 9] family maintain per-item ratings with uncertainty, updating them sequentially with each comparison. Uncertainty-aware models are attractive for human-in-the-loop pipelines because they can distinguish between "confirmed low quality" and "insufficiently explored" configurations.

Task- and Semantics-Oriented Quality. Beyond perceptual pleasantness, an important line of work evaluates image quality through *utility* for recognition, detection, or decision-making. This perspective aligns with practical requirements in scientific imaging, where an image must be interpretable to support analysis. Accordingly, semantic interpretability provides a complementary axis to

distortion- or perceptual-similarity metrics, particularly when the dominant failure mode is the loss of recognizable structures under severe degradations.

Active Sampling for Preference Learning. Because pairwise labeling is expensive, active preference learning seeks to minimize comparisons by selecting informative pairs, often prioritizing high uncertainty or high expected information gain. [4] GURO [3] provides a principled acquisition criterion that accounts for both epistemic and aleatoric uncertainty, enabling efficient preference learning with contextual attributes. These ideas are directly relevant to budget-limited human ranking pipelines. MetaRanker builds on this direction while targeting metalens-specific artifacts and interpretability-driven objectives.

3 Methods

3.1 Problem Setup: Metalens Quality Ranking

Let $\mathcal{I} = \{1, \ldots, N\}$ denote a set of metalens images generated under varying optical design parameters and reconstruction hyperparameters. Unlike natural images, metalens outputs suffer from unique degradations such as chromatic aberrations, radial blur, and reduced contrast, which often render traditional metrics (PSNR, SSIM) ineffective. We define the ranking objective not merely as perceptual pleasantness, but as **semantic interpretability**: the ability of an observer to correctly identify objects and structures despite optical artifacts. We collect pairwise comparisons (i, j, y) where $y \in \{0, 0.5, 1\}$ denotes the outcome: $y = 1$ if i is preferred, $y = 0$ if j is preferred, and $y = 0.5$ indicates a tie. This explicitly handles ambiguity in recognizability under standardized annotation guidelines (Table 1). The goal is to infer a global utility score vector $\mathbf{r} \in \mathbb{R}^N$ while minimizing the annotation budget via active sampling.

3.2 Probabilistic Ranking with Adaptive Volatility

We adopt a probabilistic model based on Glicko-2 [9], characterizing each image i by a latent quality rating r_i and a rating deviation (uncertainty) RD_i. This formulation is particularly suitable for our setup because it explicitly quantifies epistemic uncertainty, allowing the system to distinguish between "confirmed poor quality" and "unexplored parameters."

Pairwise Probability and Update. The probability that image i is preferred over j is modeled as:

$$\Pr(i \succ j) \;=\; \frac{1}{1 + 10^{-g(\mathrm{RD}_j)\,(r_i - r_j)/400}}, \tag{1}$$

where $q = \ln(10)/400$ and $g(\mathrm{RD}) = \left(1 + 3q^2\mathrm{RD}^2/\pi^2\right)^{-1/2}$. Upon observing outcome y, we update r_i and RD_i following standard Glicko-2 procedures; ties are encoded as y=0.5, yielding a near-zero update when the model already predicts near-equal quality, while a tie against an overconfident prediction still triggers the adaptive volatility mechanism below. Crucially, we incorporate an *adaptive volatility* mechanism to handle the "ambiguous artifacts" often found in metalens imaging.

Self-Correcting Property. Even when the VLM prior is imperfect and biases early sampling toward suboptimal regions, human feedback induces large prediction errors for inconsistent pairs. Our adaptive volatility mechanism inflates the item's volatility σ_i (and thus RD_i) when the prediction error $e = |y - \Pr(i \succ j)|$ is high. This increases the chance that uncertain items are revisited by the acquisition function (Eq. 3), enabling rapid recovery without manual intervention:

$$\sigma_i^{\text{new}} \;=\; \sigma_i + \alpha \cdot \max\big(0,\; |y - \Pr(i \succ j)| - \theta\big), \tag{2}$$

where α controls the correction magnitude and θ is a dead-zone threshold that prevents unnecessary inflation from minor prediction errors.

Algorithm 1 MetaRanker Active Ranking (default configuration)

Require: Image set $\mathcal{I}$, VLM prior scores s_i, budget B
Init: $r_i \leftarrow 1500,\ \mathrm{RD}_i \leftarrow \mathrm{RD}_0,\ \sigma_i \leftarrow \sigma_0$ ▷ uninformative init
for $t = 1$ to B **do**
  $C \leftarrow \textsc{SampleCandidatePairs}(\mathcal{I}; s)$ ▷ prior-guided candidate pool (hybrid)
  $(i^*, j^*) \leftarrow \arg\max_{(i,j)\in C}\ \mathcal{A}(i, j)$ ▷ Eq. 3
  if $\textsc{AutoOn} \wedge \textsc{Conf}(i^*, j^*) \geq \tau_{\text{auto}}$ **then**
    $y \leftarrow \textsc{AutoLabel}(i^*, j^*;\ s)$ ▷ weak supervision (optional)
  else
    $y \leftarrow \textsc{QueryHuman}(i^*, j^*)$
  end if
  $p \leftarrow \Pr(i^* \succ j^*)$ ▷ before rating update
  Update $(r_{i^*}, r_{j^*}, \mathrm{RD}_{i^*}, \mathrm{RD}_{j^*})$ by Glicko-2 using outcome y
  $\sigma_{i^*}, \sigma_{j^*} \leftarrow \sigma_{i^*}, \sigma_{j^*} + \alpha \cdot \max(0,\ |y - p| - \theta)$ ▷ Eq. 2
end for
return ranking by sorting r_i (descending)

3.3 Uncertainty-Aware Active Sampling

To efficiently navigate the metalens design space under a fixed budget, we employ an acquisition function $\mathcal{A}(i, j)$ that combines uncertainty with a preference for near-boundary pairs (Fig. 3(B)). Intuitively, MetaRanker targets comparisons with overlapping posteriors (a high-uncertainty region), which tend to be most informative for refining the ordering. For a candidate pair (i, j), the score is defined as:

$$\mathcal{A}(i, j) = \underbrace{(\mathrm{RD}_i + \mathrm{RD}_j)}_{\text{Uncertainty}} \cdot \underbrace{\left(1 + \frac{|r_i - r_j|}{\kappa}\right)^{-1.5}}_{\text{Fine-grained Closeness}} \cdot \underbrace{\Big(1 + \lambda(c_i + c_j)\Big)^{-1}}_{\text{Novelty}}, \tag{3}$$

where c_i is the comparison count, $\kappa = 100$ is the closeness scaling factor, and λ is the novelty decay weight.

Optimization Rationale. Standard active learning often prioritizes pure uncertainty. However, in metalens optimization, many configurations yield clearly unusable (collapsed) images. Comparing two "noise" images is uninformative. Our strategy introduces a sharpened closeness term with an exponent of 1.5 and a scaling factor $\kappa = 100$. This focuses the budget on the *Pareto frontier* of the design space distinguishing between "slightly blurry" and "sharp but aberrant" images rather than wasting queries on obvious decisions or varying grades of noise.

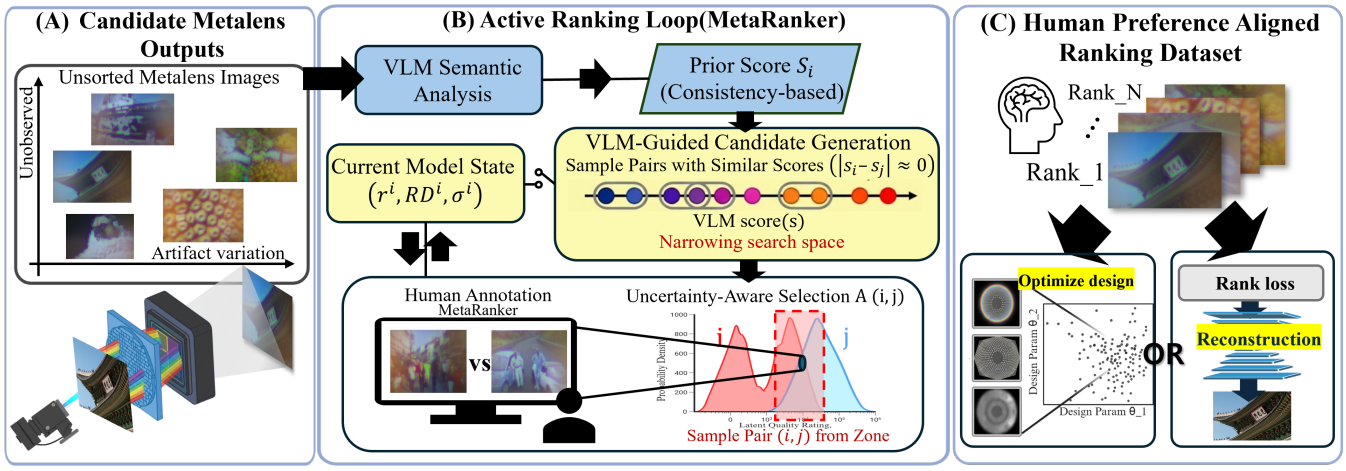

Figure 3: Overview of MetaRanker. (A) Input: Metalens outputs exhibit diverse, metric-misaligned artifacts (e.g., chromatic aberration, haze). (B) Active ranking: A VLM-guided prior and uncertainty-aware sampling select informative pairs for human comparison. (C) Use: The resulting human-aligned rankings provide evaluation ground truth and can support future reconstruction training (solid) or design optimization (dashed).

Table 1: Metalens image quality assessment guidelines for the pairwise annotation. Designed by a domain expert.

Step	Checklist
0	**Category fixed:** Each session is conducted within a single category (Animal or Landscape); all pairs shown to annotators are pre-filtered accordingly.
1	**Recognizable?** Which image lets you identify the main object/scene *more confidently* (less ambiguity)?
2	**(Animal) Texture cues:** Which shows clearer discriminative textures (fur/skin pattern, facial/limb contours) rather than just "looking sharp"? **(Landscape) Structure cues:** Which preserves coherent scene layout (boundaries/horizon/building edges) without breaking understanding?
3	**Artifact penalty:** If both are recognizable, pick the one with fewer metalens artifacts that harm understanding (color fringing/field blur/haze).
4	**Tie rule:** If still tied, prefer lower artifact severity; if both are unrecognizable, choose the "less collapsed" one (any stable structure).

VLM Prompt for Quality Assessment

System Instruction: You analyze image recognizability and description stability. Return ONLY a JSON object.

Response Format: {
`"caption": "1-2 sentences describing content",`
`"objects": ["list of clearly identifiable objects"],`
`"visibility": "clear | somewhat | unclear",`
`"confidence": 0.0 - 1.0`
} **Constraint:** If the image is blurry or unclear, set visibility="unclear" and keep the object list minimal.

Figure 4: The structured prompt used to extract semantic metadata from LLaVA. By enforcing a JSON output with an explicit object list, we enable quantitative consistency checks.

Theoretical Convergence. Active ranking is most informative when $\Pr(i \succ j) \approx 0.5$, where Fisher information is maximized. With uninformative initialization, purely random sampling wastes budget on trivially separable pairs. Our prior-guided *candidate generation* increases the likelihood of sampling near-boundary comparisons early, while the Glicko-2 uncertainty terms drive subsequent refinement. This converts the budget from global discovery to local boundary refinement, accelerating convergence under fixed annotation budgets.

3.4 VLM-based Semantic Prior

Cold-start ranking is inefficient because random initial pairs often have obvious quality differences. To accelerate convergence, we inject domain knowledge using an **off-the-shelf Vision-Language Model (VLM)**. While our framework is compatible with various foundational models, we employ LLaVA-7b [21] in our implementation to estimate a proxy quality score s_i before human annotation begins.

$$s_i = w_1 \cdot \bar{v}_i + w_2 \cdot \Omega_i + w_3 \cdot n_i, \qquad n_i = \frac{\log(1 + \bar{N}_{\text{obj}})}{\log(1 + N_{\max})}, \tag{4}$$

We use $N_{\max} = 8$, map visibility labels as `clear`$\mapsto 1.0$, `somewhat`$\mapsto 0.5$, and `unclear`$\mapsto 0.0$, and set $(w_1, w_2, w_3) = (1/3, 1/3, 1/3)$ unless otherwise stated, so $s_i \in [0, 1]$.

Semantic Consistency as Quality Proxy. We hypothesize that high-quality metalens images should yield consistent semantic descriptions, whereas degraded images lead to hallucinations or failure to detect objects. For each image i, we prompt the VLM K times with a fixed temperature of 0.7 to induce stochasticity, generating: (1) a list of detected objects O_k, (2) a visibility label $v_k \in \{\text{clear, somewhat, unclear}\}$, and (3) a confidence level c_k.

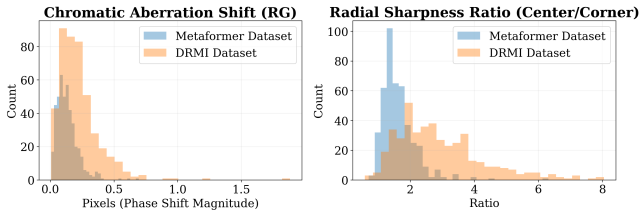

Figure 5: Dataset-level optical severity. CA Shift (px) measures chromatic misregistration, and radial sharpness ratio measures center-to-corner sharpness degradation.

We define the *semantic consistency* Ω_i using the Jaccard index of detected objects across sampled runs:

$$\Omega_i = \frac{1}{K(K-1)} \sum_{k \neq m} \frac{|O_k \cap O_m|}{|O_k \cup O_m|}. \tag{5}$$

A high Ω_i indicates that the optical features are distinct enough to be robustly recognized.

Prior Score Formulation and Usage. We convert the VLM outputs into a scalar prior score $s_i \in [0, 1]$ (Eq. 4). However, we empirically found that initializing rating values with s_i introduces anchoring bias (see Section 5.3). Therefore, we initialize all items with an uninformative rating $r_i^{(0)} = 1500$ and a common uncertainty $\mathrm{RD}_i^{(0)} = \mathrm{RD}_0$. The prior is used *solely* to guide early **candidate pair generation**: we preferentially sample pairs with similar prior scores (or adjacent ranks under s_i), thereby increasing the probability of querying informative, near-boundary comparisons without committing to the prior as ground truth. This strategy efficiently mitigates the risk of hallucinated priors in specialized scientific domains, where off-the-shelf VLMs usually lack the domain-specific calibration required for reliable absolute scoring.

Robustness to Imperfect VLM and Domain Gaps. The VLM prior may be imperfect because off-the-shelf VLMs are not calibrated for metalens artifacts. MetaRanker limits this risk by using the prior only for candidate generation, not as a label or rating initializer. When human feedback contradicts the current model prediction, adaptive volatility increases uncertainty and redirects later queries toward unresolved items (Eq. 2).

Optional Auto-comparison. To reduce human queries, we optionally auto-label a comparison when the model is highly confident (e.g., based on a large predicted margin or high prior confidence). Auto-labeled outcomes are treated as weak supervision and can be disabled without changing the core ranking model.

4 Experiments

4.1 Datasets

We evaluate MetaRanker on two metalens datasets with distinct degradation processes, and a controlled simulation used only for the human–metric agreement analysis.

DRMI (real images). DRMI [30] contains images directly captured through a custom fabricated metalens, exhibiting chromatic and field-dependent aberrations as well as sensor noise and real-world imperfections. Following the DRMI data acquisition protocol, the underlying clean scenes are cropped from DIV2K datasets [1], displayed on a monitor, and captured through the metalens imaging system. We use DRMI as a real-capture benchmark and cite both the benchmark and the underlying source-image dataset.

Metaformer (PSF-based synthesized images). Following the released MetaFormer pipeline [18], degraded inputs are generated by convolving clean Open Images V7 source images [2] with calibrated spatially varying PSFs and associated noise models, yielding more regular and controllable degradations than real capture. We use this synthetic degradation setting because the private real-capture subset reported in the original work is not publicly accessible.

Sim (PSF-Angle, controlled). To isolate optical degradation from content variability, we fix a single scene and convolve it with PSFs sampled at N_{sim}=30 uniformly spaced field angles (from on-axis to maximum field height), generating images with strictly monotonically increasing aberration severity. This controlled set is used exclusively for Table 3. One annotator repeats multiple ranking runs on Sim while varying only the algorithmic seed; we report mean±std across runs.

4.2 Setup and Evaluation

For each dataset–category–method condition, we use the same N=30 item set and allocate a fixed budget of B=90 pairwise comparisons ($3N$) per annotator session. The evaluation used a custom web interface with randomized left/right positions, neutral background, and fixed display dimensions. Human annotation was performed by seven annotators (one co-author and six independent experts in visual analysis and/or metalens imaging). Each annotator completed one independent $B = 90$ session per dataset–category–method condition; all $B = 90$ comparisons in the main benchmark were human-labeled, and auto-comparison was disabled. We compare MetaRanker against GURO [3], Bayesian Comparative Judgment (Bayesian CJ) [10], and LBPS-EIC [25] under the same budget constraint. We report inter-rater reliability (IRR) using Kendall's τ and Spearman's ρ; descriptive uncertainty summaries are provided in Appendix C. **Baselines.** GURO [3] is a contextual preference learner that models pairwise outcomes with a feature-dependent logistic model and selects pairs by an information criterion in the learned embedding space. Its performance therefore depends on whether the chosen visual features form a meaningful geometry for metalens-specific artifacts. Bayesian CJ [10] is a Bayesian Bradley–Terry-style model that maintains a per-item posterior (mean and uncertainty) and performs *active* pair selection by prioritizing comparisons that are both uncertain (high posterior variance) and near the decision boundary (high $p(1-p)$), with additional novelty/balance terms. We treat Bayesian CJ as a strong uncertainty-aware active-ranking baseline rather than a passive rank aggregation method. LBPS-EIC [25] is a learning-based pairwise comparison method that leverages perceptual embeddings and information-theoretic criteria for active pair selection; we include it as an additional competitive baseline following reviewer recommendation. To manage annotation cost in this specialized domain, we applied a two-stage screening procedure: all candidate methods were first evaluated via simulation, and the top-performing methods were advanced to the human annotation phase. For descriptive comparison between methods, we report paired non-parametric bootstrap confidence

intervals and Cliff's δ effect sizes for Kendall's τ and Spearman's ρ differences (Appendix C). To standardize judgments toward semantic interpretability (recognizability), annotators followed a concise guideline formulated by a domain expert (15 years of experience) grounded in established perceptual research [19, 28] (Table 1).

5 Results

5.1 Ranking Performance and Metric Alignment

Inter-Rater Reliability under Fixed Budget Table 2 reports IRR and condition-wise rank aggregation under the fixed budget constraint ($B = 3N$). MetaRanker achieves the highest Kendall's τ in all four conditions, ranging from 0.695 (DRMI Animal) to 0.840 (MetaFormer Landscape). When Kendall and Spearman ranks are aggregated, MetaRanker also obtains the best Final Rank in every dataset–category condition. For Kendall's τ, descriptive bootstrap summaries (Appendix Table 8) show directionally positive improvements in all 12 baseline-condition comparisons (3 baselines × 4 conditions). Against Bayesian CJ, Kendall's τ gains are +0.077 on DRMI Animal, +0.336 on DRMI Landscape, +0.182 on Metaformer Animal, and +0.196 on Metaformer Landscape. Against GURO, all conditions show positive Kendall gains (e.g., DRMI Animal $\Delta\tau = +0.269$; DRMI Landscape $\Delta\tau = +0.363$). In terms of annotation efficiency, the human benchmark uses $B = 90$ comparisons instead of 435 exhaustive pairs for $N = 30$, a 79.3% reduction, while preserving the observed Kendall and Final-Rank gains.

IQA Metrics versus Human Rankings Table 3 compares standard IQA metrics with reference-free human rankings. In DRMI (diverse scenes), full-reference metrics are near-zero or negative (e.g., PSNR/SSIM/LPIPS Kendall $\tau \approx -0.08$), indicating structural misalignment. In the controlled Sim (PSF-Angle) setting where content is fixed, alignment improves only modestly: NIQE reaches Kendall $\tau = 0.28 \pm 0.07$ (Spearman $\rho = 0.40 \pm 0.11$), while full-reference metrics remain low ($\tau < 0.15$). This shows that the mismatch is not explained solely by content diversity. Figure 2 illustrates representative cases where pixel-level similarity fails to reflect semantic interpretability under chromatic/field-dependent aberrations.

5.2 Annotation Difficulty Structure across Conditions

Figure 5 shows that DRMI exhibits markedly higher optical severity (CA shift and radial deviation) than Metaformer. Table 4 characterizes the annotation difficulty at the condition level. DRMI Landscape exhibits a *polarized* distribution (Easy 63.5%, Hard 15.7%), DRMI Animal is *diffuse* (Easy 49.4%, Hard 33.6%), and both Metaformer conditions show *plateau* distributions (Easy > 72%, Hard < 7%). Notably, MetaRanker's gain over Bayesian CJ does not scale monotonically with overall optical severity or mean difficulty; the largest gain(+0.336) occurs in the polarized regime. Figure 6 reports convergence against an exhaustive ground-truth ranking determined by PSNR on a larger set (N=500, B=1,500; 3 runs).

We employ PSNR here solely as a deterministic oracle to benchmark search efficiency, isolating algorithmic convergence from the human label noise and perceptual misalignment discussed earlier.

Given this complete ground truth, MetaRanker and Bayesian CJ converge to similar levels ($\tau \approx 0.77$–0.74, $\Delta\tau < 0.03$), whereas feature-based methods (GURO, RankNet) converge substantially lower ($\tau \leq 0.6$ on Metaformer) and exhibit high run-to-run variance. This performance gap is consistent with a separate feature space analysis (available in the released code repository): NSS features separate metalens outputs from the natural image manifold, but deep ResNet50 embeddings entangle metalens artifacts with standard distortions, providing no discriminative boundaries for quality-based active selection.

5.3 Ablation Study

Table 6 in Appendix A reports ablation results under a controlled synthetic setup (N=600, B=1,800). H1 shows that the VLM prior is beneficial when used as a sampling guide rather than as a rating warm-start: the flat-start sampling-guide variant reaches $\tau = 0.804$, improving over the flat-start baseline ($\tau = 0.797$) by $\Delta\tau = +0.007$, whereas direct VLM-biased warm-starting decreases performance to $\tau = 0.785$. H2 confirms the value of active acquisition, with the hybrid strategy improving over random selection by $\Delta\tau = +0.069$. H3 further shows that explicit rating uncertainty is essential: Glicko-2 with adaptive volatility outperforms Elo with $K = 32$ by $\Delta\tau = +0.427$. Because these ablations use a synthetic Bradley–Terry oracle, we interpret them as mechanism checks supporting flat initialization, prior-guided candidate generation, active acquisition, and Glicko-2-style uncertainty tracking.

6 Discussion

This study found that human-aligned ranking is both necessary and achievable under strict budget constraints. In this section, we analyze the structural causes of metric–human misalignment, explain why feature-based active learning struggles in this domain, and characterize the difficulty regimes where MetaRanker offers the greatest advantage. Finally, we discuss implications for perceptually grounded metalens co-design.

Metric–Human Misalignment in Metalens Imaging. Across our experiments, standard IQA metrics exhibit weak or even negative correlations with human-derived rankings in the metalens domain (Section 5.1). While content diversity can amplify this discrepancy, it does not explain it: in the controlled Sim setting, where scene content is fixed and only optical degradation varies, full-reference metrics (PSNR, SSIM, LPIPS) remain poorly aligned with human rankings (Kendall's $\tau < 0.15$), and only NIQE achieves moderate correlation.

We argue that the root cause is structural. A t-SNE analysis of ResNet50 embeddings reveals that metalens-specific aberrations (e.g., chromatic misregistration, spatially varying blur, and contrast loss) are entangled with conventional distortions such as noise or compression. Under such entanglement, proximity in feature space does not reliably correspond to proximity in perceived semantic interpretability. NSS features separate metalens outputs from the natural-image manifold more clearly, but they primarily capture statistical naturalness rather than recognizability. This misalignment has implications beyond evaluation. Any model-selection or optimization pipeline that uses these metrics as proxies for human interpretability inherits the same objective mismatch, potentially

Table 2: Inter-rater reliability (IRR) of metalens image quality assessment. The mean±std of pairwise session agreement is shown. Rank Sum is computed as Rank(Kendall)+Rank(Spearman) within each dataset–category condition; lower is better. Final Rank is determined by Rank Sum, with ties resolved by Kendall rank and then Spearman rank. Bold indicates the best metric value or best rank; underline indicates the second-best metric value.

Dataset	Category	Method	Kendall	Spearman	Rank Sum↓	Final Rank↓
DRMI [30]	Animal	GURO [3]	0.426±0.394	0.358±0.596	8	4
		Bayesian CJ [10]	0.618±0.240	**0.846±0.065**	4	2
		LBPS-EIC [25]	<u>0.637±0.228</u>	0.814±0.072	5	3
		Ours	**0.695±0.190**	<u>0.830±0.056</u>	**3**	**1**
	Landscape	GURO [3]	0.423±0.385	0.443±0.752	8	4
		Bayesian CJ [10]	0.450±0.337	0.633±0.109	6	3
		LBPS-EIC [25]	<u>0.773±0.133</u>	**0.912±0.032**	**3**	2
		Ours	**0.786±0.128**	<u>0.859±0.076</u>	**3**	**1**
MetaFormer [18]	Animal	GURO [3]	0.272±0.491	0.098±0.893	8	4
		Bayesian CJ [10]	<u>0.612±0.236</u>	<u>0.862±0.043</u>	4	2
		LBPS-EIC [25]	0.508±0.298	0.713±0.103	6	3
		Ours	**0.794±0.152**	**0.897±0.058**	**2**	**1**
	Landscape	GURO [3]	<u>0.783±0.158</u>	**0.935±0.029**	**3**	2
		Bayesian CJ [10]	0.644±0.334	0.854±0.039	7	4
		LBPS-EIC [25]	0.657±0.194	0.813±0.072	7	3
		Ours	**0.840±0.095**	<u>0.877±0.066</u>	**3**	**1**

Table 3: Human–metric agreement showing the failure of standard metrics. DRMI (diverse): Real-world images with varied scenes. Sim (controlled): Simulation with fixed content to isolate optical degradation. Agreement is consistently low across all metric types, including recent learned metrics (CLIP-IQA+, FGResQ), confirming that the metric–human gap cannot be closed by stronger automated IQA alone.

		Diverse (DRMI)		Controlled (Sim)	
Type	Metric	Kendall	Spearman	Kendall	Spearman
No-Reference	NIQE [24]	0.13±0.08	0.18±0.12	0.28±0.07	0.40±0.11
	MUSIQ [15]	0.05±0.03	0.07±0.05	0.06±0.06	0.09±0.08
	MANIQA [39]	−0.06±0.07	−0.09±0.10	0.06±0.07	0.08±0.10
	CLIP-IQA+ [35]	−0.05±0.06	−0.07±0.09	−0.06±0.05	−0.08±0.07
	FGResQ [31]	−0.32±0.11	−0.45±0.14	−0.11±0.08	−0.15±0.10
Full-Reference	LPIPS [41]	−0.08±0.10	−0.11±0.12	0.14±0.08	0.19±0.10
	SSIM [36]	−0.08±0.09	−0.11±0.11	0.09±0.09	0.12±0.13
	PSNR [12]	−0.08±0.08	−0.12±0.11	0.06±0.09	0.08±0.12

converging to numerically favorable solutions that do not preserve the semantic structures that matter for deployment utility. This concern is directly reflected in our simulation analysis.

Implications for Feature-Based Active Preference Learning. The representational mismatch above directly affects feature-based active learning strategies. GURO identifies informative pairs using contextual embeddings, implicitly assuming that feature proximity reflects proximity in the target quality ordering. When this assumption fails—as the ResNet50 embedding analysis confirms—the acquisition policy can repeatedly select pairs that are close in embedding space yet uninformative for ordering by semantic interpretability. This mechanism explains the pronounced oscillations and early plateaus observed in simulation (Fig. 6) and the instability of feature-based methods across random seeds.

Table 4: Annotation difficulty structure. Easy/Middle/Hard: fraction of pairs by annotator agreement (≥0.8 / between / ≤0.6). Regime: qualitative summary of the difficulty distribution pattern. The difficulty structure helps explain the varying effectiveness of active ranking across conditions.

Condition	Easy†	Middle	Hard‡	Regime
DRMI-Landscape	63.5%	20.8%	15.7%	Polarized
DRMI-Animal	49.4%	17.0%	33.6%	Diffuse
Metaformer-Landsc.	72.4%	20.9%	6.7%	Plateau
Metaformer-Animal	74.9%	20.2%	4.9%	Plateau

†Agreement ≥ 0.8. ‡Agreement ≤ 0.6.

MetaRanker is designed to reduce dependence on external feature geometry. Its acquisition function (Eq. 3) operates on uncertainty internal to the probabilistic rating model, rather than on distances in a pretrained embedding space. The VLM-based semantic prior plays a complementary role: it constrains candidate generation by grouping images with similar semantic-consistency scores, increasing the likelihood of near-boundary comparisons early. Notably, the benefit arises less from substituting a stronger backbone per se, and more from using a task-relevant semantic-consistency signal to narrow the search space in a way that does

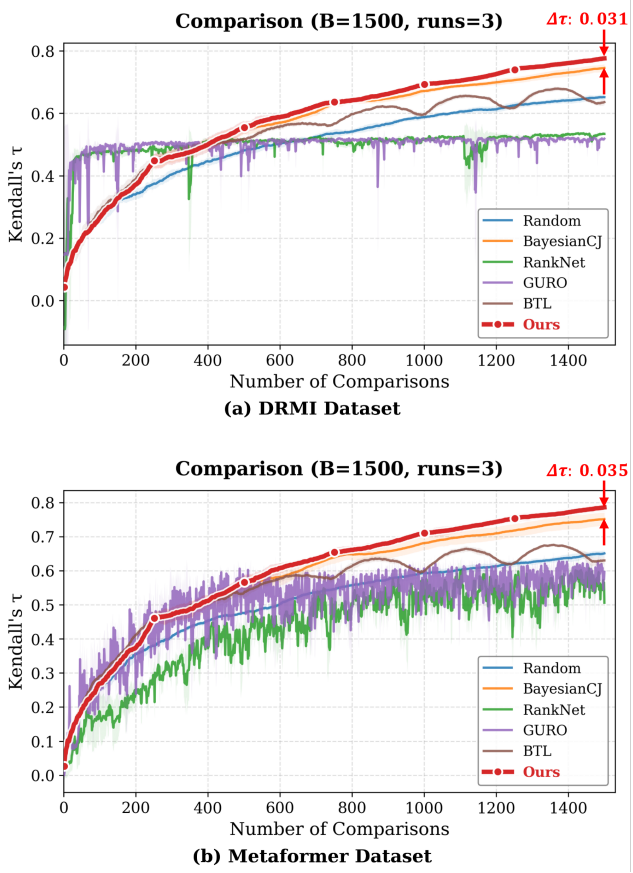

Figure 6: Controlled simulation against exhaustive PSNR ground truth (N=500, B=1500, 3 runs). Under this deterministic proxy target, Ours and Bayesian CJ converge similarly, indicating comparable search efficiency when the target ranking is internally consistent. This experiment evaluates algorithmic convergence only and does not validate PSNR as a perceptual quality target. In contrast, feature-based methods (GURO, RankNet) plateau at $\tau \leq 0.6$ on Metaformer, revealing the limitations of deep feature representations for metalens artifacts.

not require a well-calibrated embedding geometry for metalens artifacts.

Regime-Dependent Benefits of Active Ranking. In this small expert-panel setting, Kendall's τ is the most appropriate primary IRR criterion for our main analysis. Each condition is evaluated by only seven annotators, with one session per annotator, so we avoid relying on strong assumptions about calibrated score magnitudes or stable rank-distance scales across raters. Kendall's τ measures pairwise concordance between induced rankings, matching both the pairwise annotation primitive and the active-ranking objective. Spearman's ρ is retained as a complementary rank-consistency measure because it summarizes global rank displacement, but we treat it as secondary to the pairwise-order agreement captured by Kendall's τ. We therefore focus our main analysis on these two rank-based IRR metrics: Kendall's τ for ordinal pairwise agreement and Spearman's ρ for global rank consistency.

Although MetaRanker achieves the highest Kendall's τ across all conditions, the magnitude of improvement depends strongly on the baseline and condition. Relative to Bayesian CJ, the gain ranges from a modest margin on DRMI Animal to a large gap on DRMI Landscape (Section 5.1). We attribute this variation to how the adaptive volatility mechanism (Eq. 2) interacts with the condition-specific difficulty structure (Table 4).

In the *polarized* regime of DRMI Landscape, hard pairs concentrate near a narrow decision boundary. When human feedback contradicts the model's current belief, the volatility mechanism inflates uncertainty for the relevant items, and the acquisition function revisits them. Because the informative region is compact, the feedback loop repeatedly allocates queries where information gain is highest. This yields a large Kendall gap against Bayesian CJ (Ours 0.786 vs. 0.450), while the stronger LBPS-EIC baseline narrows the margin in this condition.

In the *plateau* regimes of Metaformer, most pairs are unambiguous (Easy > 72%, Hard < 7%), so large prediction errors and thus volatility reactivation occur infrequently. With limited near-boundary mass, the marginal benefit over simpler uncertainty-aware baselines is naturally smaller (Metaformer Animal: Ours 0.794 vs. Bayesian CJ 0.612). The *diffuse* regime of DRMI Animal lies between these extremes: informative pairs exist but are broadly distributed, requiring wider exploration and yielding a moderate advantage. Overall, MetaRanker provides the largest gains when the quality landscape contains a meaningful yet localized boundary region, which commonly arises in metalens design spaces near recognizability thresholds.

Objective Alignment and Implications for Co-Design. Our simulation experiments reveal a complementary insight. Under a consistent PSNR-based simulation target, MetaRanker and Bayesian CJ converge to similar Kendall's τ. When the target signal is internally consistent, asymptotic performance is determined more by the validity of the target than by small differences in acquisition design; crucially, this distortion-based proxy is structurally misaligned with human perception in the metalens domain (Table 3). High sample efficiency under such an objective mismatch leads to fast convergence toward the *wrong target*. This motivates caution for end-to-end metalens pipelines that optimize pixel-fidelity proxies (e.g., PSNR/SSIM) as primary objectives [32, 34], since they may reduce pixel-wise distortion without improving semantic interpretability. Importantly, this paper does not present closed-loop co-design. Rather, we position MetaRanker as addressing the evaluation bottleneck required for perceptually aligned objectives. This framing follows established precedents: LPIPS [41] was first validated as an evaluation metric before being adopted as a loss, and preference-based reward modeling [7] preceded its integration into RLHF pipelines [26]. Integrating MetaRanker into a co-design loop is technically feasible, e.g., by training a differentiable surrogate loss from rankings or by using the ranking as a fitness signal in derivative-free optimization [42], but empirical validation of such closed-loop integration remains future work.

Limitations. This study has constraints that limit the generality of its conclusions. We evaluated two metalens settings (DRMI and Metaformer); because optical severity is confounded with dataset

identity, relationships among severity, difficulty regime, and ranking advantage should be interpreted as suggestive rather than causal. The limited annotation data (n=7 sessions per dataset–category–method condition) restricts statistical resolution and increases sensitivity to outlier judgments, motivating our emphasis on descriptive bootstrap intervals and effect sizes rather than fine-grained p-values. Methodologically, gaps remain between simulation proxies and real-world deployment: ablations rely on synthetic rankings, and the simulation is driven by the PSNR metric, which is misaligned with human interpretability (Table 3). Finally, MetaRanker is validated as an evaluation protocol; demonstrating end-to-end closed-loop integration remains out of scope.

Future Work. Future work will address these limitations along three axes. First, we will extend evaluation to additional metalens systems spanning intermediate severity levels and characterize budget sensitivity via sweeps over $B \in [N, 5N]$, including studies with non-expert annotators. Second, we will close the loop between evaluation and optimization: as discussed in Section 6, this could involve training a differentiable surrogate model using pairwise preferences from MetaRanker (analogous to reward modeling in RLHF [7]) or using the ranking directly as a fitness signal in gradient-free optimization [42]. In parallel, we will scale the VLM prior to larger candidate pools (N>100) and benchmark stronger foundation models to better characterize the trade-off between prior quality and annotation reduction. Third, we will explore transferring the same uncertainty-aware ranking protocol to related diffractive/holographic imaging settings where field-dependent aberrations and metric misalignment are also common.

7 Limitations and Ethical Considerations

The DRMI dataset is an image dataset captured using a metalens that we independently designed and fabricated (Figshare, open access [30]); the MetaFormer dataset is a simulated dataset [18]. Both are publicly available. No personally identifiable or sensitive data were collected from annotators; the study involved only image-quality preference judgments under the protocol in Section 4.2.

8 Conclusion

MetaRanker addresses the prohibitive $O(N^2)$ annotation bottleneck in metalens image quality evaluation by introducing a "supervision-format shift" that formulates semantic interpretability as a human-grounded ranking problem rather than a distortion-metric optimization target. MetaRanker uses a fixed budget of only $B = 3N$ pairwise comparisons per session, corresponding to an approximately 80% reduction relative to exhaustive pairwise evaluation for $N = 30$. Under this constrained labeling budget, MetaRanker achieves the top Final Rank across all animal and landscape conditions in both the DRMI and MetaFormer datasets (Table 2). These results confirm that reliable human-aligned preference supervision can be efficiently obtained under strict annotation constraints.

A central design principle is that the VLM prior is used only as an imperfect sampling guide, not as a supervisory label or rating initializer. This flat-start strategy prevents anchoring bias, keeping human comparisons as the primary supervision signal throughout the ranking process. When observed human feedback contradicts the model's current preference prediction, the adaptive volatility mechanism raises uncertainty for the relevant items and redirects subsequent acquisition toward unresolved comparisons. This self-correcting behavior allows MetaRanker to benefit from "weak" AI semantic guidance without anchoring the final ranking to potentially hallucinated VLM prior predictions.

Overall, our findings suggest that standard IQA metrics do not accurately reflect human perception in the metalens domain, especially when semantic recognizability is degraded by chromatic and field-dependent aberrations that cause "recognizability collapse." MetaRanker therefore provides a practical evaluation protocol for perceptually grounded metalens assessment. Integrating the resulting human-aligned rankings into reconstruction training or closed-loop optical co-design is a natural next step. While our preliminary reconstruction pilots demonstrate that supervising pipelines with MetaRanker rankings yields significant gains in downstream perceptual (LPIPS) and naturalness (NIQE) metrics, full end-to-end empirical validation of such optimization pipelines remains future work.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grants funded by the Ministry of Science and ICT (MSIT) (RS-2024-00338048, RS-2024-00455720, RS-2026-25487796); the National Institute of Health (NIH) research project (2026-ER0904-00); the Advanced GPU Utilization Support Program and the High Performance Computing Support project (RQT-25-070083), both funded by MSIT; Hankuk University of Foreign Studies Research Fund of 2026; the Culture, Sports and Tourism R&D Program through KOCCA funded by MCST (RS-2024-00332210); the Artificial Intelligence Graduate School Program, Hanyang University (RS-2020-II201373); and the Artificial Intelligence Semiconductor Support Program (IITP-2025-RS-2023-00253914), supervised by IITP and funded by the Korean government.

GenAI Disclosure

Generative AI tools were not used for algorithm development, data collection, human annotation, statistical analysis, or result generation. Claude was used for code debugging assistance, and GPT-5 was used for grammatical and stylistic refinement. The authors reviewed and verified all AI-assisted content and take full responsibility for the work.

References

[1] Eirikur Agustsson and Radu Timofte. 2017. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 126–135.

[2] Rodrigo Benenson and Vittorio Ferrari. 2022. From colouring-in to pointillism: revisiting semantic segmentation supervision. *arXiv preprint arXiv:2210.14142* (2022).

[3] Herman Bergström, Emil Carlsson, Devdatt Dubhashi, and Fredrik D. Johansson. 2024. Active Preference Learning for Ordering Items In- and Out-of-Sample. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[4] Erdem Biyik, Nima Anari, and Dorsa Sadigh. 2024. Batch active learning of reward functions from human preferences. *ACM Transactions on Human-Robot Interaction* 13, 2 (2024), 1–27.

[5] Yochai Blau and Tomer Michaeli. 2018. The perception-distortion tradeoff. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6228–6237.

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

[7] Paul F Christiano, Jan Leike, Tom Brown, Miljan Marber, Buck Shlegeris, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, Vol. 30.

[8] Mark E Glickman. 1995. The glicko system. *Boston University* 16, 8 (1995), 9.

[9] Mark E. Glickman. 2012. *Example of the Glicko-2 System.* Technical Report. Boston University. 1–6 pages. https://www.glicko.net/glicko/glicko2.pdf Online technical report (commonly cited as 2012)..

[10] Andy Gray, Alma Rahat, Tom Crick, and Stephen Lindsay. 2024. A Bayesian active learning approach to comparative judgement within education assessment. *Computers and Education: Artificial Intelligence* 6 (2024), 100245.

[11] Xuemei Hu, Weizhu Xu, Qingbin Fan, Tao Yue, Feng Yan, Yanqing Lu, and Ting Xu. 2024. Metasurface-based computational imaging: a review. *Advanced Photonics* 6, 1 (2024), 014002–014002.

[12] Quan Huynh-Thu and Mohammed Ghanbari. 2008. Scope of validity of PSNR in image/video quality assessment. *Electronics Letters* 44, 13 (2008), 800–801.

[13] Andrey Ignatov, Nikolay Kobyshev, Radu Timofte, Kenneth Vanhoey, and Luc Van Gool. 2017. DSLR-quality photos on mobile devices with deep convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*. 3277–3285.

[14] Ikbeom Jang, Garrison Danley, Ken Chang, and Jayashree Kalpathy-Cramer. 2022. Decreasing annotation burden of pairwise comparisons with human-in-the-loop sorting: Application in medical image artifact rating. *arXiv preprint arXiv:2202.04823* (2022).

[15] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. 2021. MUSIQ: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5148–5157.

[16] Mohammadreza Khorasaninejad and Federico Capasso. 2017. Metalenses: Versatile multifunctional photonic components. *Science* 358, 6367 (2017), eaam8100.

[17] Joohoon Kim, Junhwa Seong, Wonjoong Kim, Gun-Yeal Lee, Seokwoo Kim, Hongyoon Kim, Seong-Won Moon, Dong Kyo Oh, Younghwan Yang, Jeonghoon Park, et al. 2023. Scalable manufacturing of high-index atomic layer–polymer hybrid metasurfaces for metaphotonics in the visible. *Nature Materials* 22, 4 (2023), 474–481.

[18] Byeonghyeon Lee, Youbin Kim, Yongjae Jo, Hyunsu Kim, Hyemi Park, Yangkyu Kim, Debabrata Mandal, Praneeth Chakravarthula, Inki Kim, and Eunbyung Park. 2024. Aberration Correcting Vision Transformers for High-Fidelity Metalens Imaging. *arXiv preprint arXiv:2412.04591* (2024).

[19] Chunyi Li, Yuan Tian, Xiaoyue Ling, Zicheng Zhang, Haodong Duan, Haoning Wu, Ziheng Jia, Xiaohong Liu, Xiongkuo Min, Guo Lu, et al. 2025. Image Quality Assessment: From Human to Machine Preference. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 7570–7581.

[20] Suiyi Ling, Jing Li, Anne Flore Perrin, Zhi Li, Lukáš Krasula, and Patrick Le Callet. 2020. Strategy for boosting pair comparison and improving quality assessment accuracy. *arXiv preprint arXiv:2010.00370* (2020).

[21] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in Neural Information Processing Systems* 36 (2023), 34892–34916.

[22] Lucas Maystre and Matthias Grossglauser. 2017. Just sort it! A simple and effective approach to active preference learning. In *International Conference on Machine Learning (ICML)*. PMLR, 2344–2353.

[23] Aliaksei Mikhailiuk, María Pérez-Ortiz, and Rafal Mantiuk. 2018. Psychometric scaling of TID2013 dataset. In *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 1–6.

[24] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. 2012. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters* 20, 3 (2012), 209–212.

[25] Shima Mohammadi and João Ascenso. 2025. Uncertainty-driven Sampling for Efficient Pairwise Comparison Subjective Assessment. *IEEE Transactions on Multimedia* (2025).

[26] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, et al. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, Vol. 35.

[27] Yujin Park, Haejun Chung, and Ikbeom Jang. 2025. EZ-Sort: Efficient Pairwise Comparison via Zero-Shot CLIP-Based Pre-Ordering and Human-in-the-Loop Sorting. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*. 5120–5124.

[28] Judith Alice Redi, Tobias Hoßfeld, Pavel Korshunov, Filippo Mazza, Isabel Povoa, and Christian Keimel. 2013. Crowdsourcing-based multimedia subjective evaluations: a case study on image recognizability and aesthetic appeal. In *Proceedings of the 2nd ACM International Workshop on Crowdsourcing for Multimedia*. 29–34.

[29] Eli Schwartz, Raja Giryes, and Alex M Bronstein. 2018. DeepISP: Toward learning an end-to-end image processing pipeline. *IEEE Transactions on Image Processing* 28, 2 (2018), 912–923.

[30] Joonhyuk Seo, Jaegang Jo, Joohoon Kim, Joonho Kang, Chanik Kang, Seong-Won Moon, Eunji Lee, Jehyeong Hong, Junsuk Rho, and Haejun Chung. 2024. Deep-learning-driven end-to-end metalens imaging. *Advanced Photonics* 6, 6 (2024), 066002–066002.

[31] Xiangfei Sheng, Xiaofeng Pan, Zhichao Yang, Pengfei Chen, and Leida Li. 2026. Fine-grained image quality assessment for perceptual image restoration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 40. 8914–8922.

[32] Vincent Sitzmann, Steven Diamond, Yifan Peng, Xiong Dun, Stephen Boyd, Wolfgang Heidrich, Felix Heide, and Gordon Wetzstein. 2018. End-to-end optimization of optics and image processing for achromatic extended depth of field and super-resolution imaging. *ACM Transactions on Graphics (TOG)* 37, 4 (2018), 1–13.

[33] Tianshu Song, Leida Li, Hancheng Zhu, and Jiansheng Qian. 2021. IE-IQA: Intelligibility enriched generalizable no-reference image quality assessment. *Frontiers in Neuroscience* 15 (2021), 739138.

[34] Ethan Tseng, Shane Colburn, James Whitehead, Luocheng Huang, Seung-Hwan Baek, Arka Majumdar, and Felix Heide. 2021. Neural nano-optics for high-quality thin lens imaging. *Nature Communications* 12, 1 (2021), 6493.

[35] Jianyi Wang, Kelvin C.K. Chan, and Chen Change Loy. 2023. Exploring CLIP for Assessing the Look and Feel of Images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 2555–2563.

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

[37] Long Xu, Jia Li, Weisi Lin, Yun Zhang, Yongbing Zhang, and Yihua Yan. 2016. Pairwise comparison and rank learning for image quality assessment. *Displays* 44 (2016), 21–26.

[38] Miao Yang, Ge Yin, Yixiang Du, and Zhiqiang Wei. 2021. Pair comparison based progressive subjective quality ranking for underwater images. *Signal Processing: Image Communication* 99 (2021), 116444.

[39] Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. 2022. MANIQA: Multi-Dimension Attention Network for No-Reference Image Quality Assessment. In *CVPR Workshops (NTIRE)*.

[40] Joel Yeo, N Duane Loh, Ramon Paniagua-Dominguez, and Arseniy I Kuznetsov. 2025. EigenCWD: a spatially varying deconvolution algorithm for single metalens imaging. *Optics Express* 33, 13 (2025), 28481–28492.

[41] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 586–595.

[42] Simone Zini and Marco Buzzelli. 2025. Bayesian nights: Optimizing night photography rendering with Bayesian derivative-free methods. *Pattern Recognition* 161 (2025), 111314.

Appendix

A Ablation Study Details

A.1 Hyperparameter Configuration

Table 5 lists the exact values used in all main experiments. Sensitivity to α and RD_0 is higher than for other parameters because these govern the adaptive volatility correction, which is designed to buffer against human label noise; κ and θ are stable across a wide range.

Table 5: Hyperparameter values used in all experiments. $\Delta\tau$ reports the maximum deviation from the default across the one-factor sweep range shown.

Symbol	Default	Sweep range	$\Delta\tau$
α	0.5	{0.1, 0.3, 0.5, 1.0}	0.130
θ	0.05	{0.01, 0.05, 0.10, 0.20}	0.070
RD_0	350	{150, 250, 350, 500}	0.118
σ_0	0.06	{0.02, 0.06, 0.10, 0.20}	0.019
λ	0.8	{0.2, 0.5, 0.8, 1.5}	0.038
κ	100	{50, 100, 200, 400}	0.025
τ_{auto}	0.85	{0.80, 0.85, 0.90, 0.95}	0.014

The ablation study results are summarized in Table 6. The simulation used a synthetic Bradley–Terry ground truth with $N = 600$ items.

H1: Role of VLM Prior (Initialization vs. Sampling Guide). The results demonstrate that the VLM prior is most effective when used

as a *sampling guide* rather than for *initialization*. The configuration **Sampling guide (Flat start,** $s = 0$**)** achieves the highest performance ($\tau = 0.804$). In contrast, directly initializing ratings with VLM scores (**Warm-start only,** $s = 100$) yields a lower τ (0.785) than even the baseline (0.797). This indicates that biasing the initial state with imperfect VLM predictions creates a "bad prior" that the ranking algorithm struggles to unlearn under a limited budget. Instead, the sampling guide strategy uses the VLM solely to identify candidate pairs with similar scalar prior scores, e.g., small $|s_i - s_j|$ or adjacent ranks under s_i. By prioritizing such pairs for comparison, the algorithm filters out uninformative "obvious" pairs (e.g., severe blur vs. sharp) and focuses the budget on likely decision boundaries from the start, all while maintaining an unbiased flat initialization ($r = 1500$) for the rating engine itself.

H2 & H3: Necessity of Uncertainty-Awareness. Hypothesis H2 confirms that active acquisition is critical; **Random** selection performs significantly worse ($\tau = 0.661$) compared to **Uncertainty** or **Hybrid** strategies ($\tau \approx 0.730$), validating the efficiency of targeting decision boundaries. Furthermore, H3 highlights the importance of modeling rating uncertainty explicitly. **Elo**, which lacks a dynamic uncertainty parameter (*RD*), fails catastrophically ($\tau = 0.298$), whereas **Glicko-2** maintains high performance. This proves that the acquisition function relies heavily on the variance of the posterior distribution (*RD*) to identify informative pairs.

H4 & H5: Optional Automation and Robustness. H4 shows that auto-comparison is not the primary source of the main performance gains. Conservative settings, such as Threshold 0.90 and Adaptive, remain comparable to the Off setting, whereas the more aggressive Threshold 0.85 setting reduces performance. Because this table reports final ranking accuracy rather than the number of saved human labels, we treat auto-comparison as an optional efficiency component rather than as a core supervision mechanism. H5 evaluates robustness to synthetic comparison-label noise. The results remain within a narrow range across tested noise levels, but they do not show a monotonic degradation pattern. Therefore, the slightly higher value at Noise $p = 0.6$ should not be interpreted as evidence that noise improves ranking. Rather, the result supports a bounded robustness claim: adaptive volatility can prevent severe collapse when observed comparison outcomes are inconsistent with the model's current predictions. Adversarial VLM-prior corruption is evaluated separately in Section E.

B VLM Prompting and Post-processing Details

We summarize the VLM decoding configuration and deterministic post-processing rules used to derive the VLM prior in Section 3.4. The structured prompt is shown in Fig. 4; the score is computed from fixed decoded fields rather than free-form caption quality.

B.1 Scope and design choice

The `caption` field is logged for auditability but is not a primary signal in our score. The prior in Eq. 4 uses mapped visibility labels, object-set consistency, and normalized object count with $N_{\max} = 8$ and equal weights unless otherwise stated.

C Descriptive Uncertainty Analysis

We summarize performance differences between MetaRanker and competing baselines using the seven human annotation sessions from Table 2. For each (dataset, category, method) cell, the seven annotator-specific rankings yield $\binom{7}{2} = 21$ pairwise inter-session correlations for Kendall's τ and Spearman's ρ. Mean Diff. is the mean correlation of MetaRanker minus that of the baseline; positive values indicate higher IRR for MetaRanker. We report paired nonparametric bootstrap CIs with 20,000 resamples and Cliff's δ effect sizes. Because the 21 correlations are not fully independent—the same annotator appears in multiple pairs—we interpret the intervals

Table 6: Ablation summary (mean±std). $N = 600$, $B = 1{,}800$. **Note that** s **denotes the warm-start spread:** $s = 0$ **indicates a flat initialization (**$r = 1500$**), while** $s = 100$ **indicates a VLM-biased initialization.**

Hypothesis	Method	Kendall τ
H1: Prior (Init vs Guide)	**Sampling guide (Flat start,** $s = 0$**)**	**0.804±0.004**
	Baseline (Flat start, $s = 0$)	0.797±0.002
	RD init (Flat start, $s = 0$)	0.784±0.009
	Full prior (Flat start, $s = 0$)	0.781±0.005
	Warm-start only ($s = 100$)	0.785±0.009
	Warm-start + Sampling ($s = 100$)	0.767±0.008
	Warm-start + RD ($s = 100$)	0.741±0.004
	Full prior + warm-start ($s = 100$)	0.724±0.002
H2: Acquis.	Hybrid	0.730±0.012
	Uncertainty	0.729±0.009
	Boundary	0.724±0.002
	Random	0.661±0.009
H3: Rating	Glicko-2 (Adaptive)	0.725±0.001
	Glicko-2 ($\tau = 0.5$)	0.724±0.002
	Elo ($K = 32$)	0.298±0.002
	Elo ($K = 16$)	0.153±0.001
H4: Auto	Threshold 0.90	0.724±0.002
	Adaptive	0.724±0.002
	Off	0.721±0.003
	Threshold 0.85	0.710±0.004
H5: Label noise	Noise $p = 0.6$	0.730±0.017
	Clean ($p = 0$)	0.724±0.002
	Noise $p = 0.4$	0.722±0.018
	Noise $p = 0.8$	0.720±0.005
	Noise $p = 0.2$	0.714±0.011

Table 7: VLM prompting configuration (fixed across datasets). We use an off-the-shelf VLM (LLaVA-7b via an Ollama API) and decode K **stochastic samples per image for semantic stability estimation.**

Item	Setting
VLM model	LLaVA-7b (Ollama backend)
Image encoding	JPEG, max side 768 px, quality 85
Quality samples (K)	K=2 (default), fixed per experiment
Quality temperature	T_{VLM}=0.7
Category temperature	0.1 (when category prompt is used)
Seeds	$17 + 101k$ for $k \in \{0, \ldots, K-1\}$
Max objects	up to 8 nouns (lowercased)
Truncation	caption: 400 chars; readable_text: 200 chars
Fallback on parse failure	visibility=unclear, objects=[], confidence=0

Table 9: Robustness to adversarial prior corruption. Final Kendall's τ degrades gracefully under moderate corruption, supporting a bounded robustness claim for the self-correcting mechanism.

Corruption p	Initial τ (prior)	Final τ	Δ from clean
0.0 (clean)	0.170	0.747	—
0.5	0.003	0.710	−0.037
1.0 (inverted)	−0.170	0.669	−0.078

Table 8: Descriptive comparison between MetaRanker and baselines. Mean Diff. is Ours minus baseline; bold highlights a 95% bootstrap CI excluding 0.

Condition	Method	M	Mean Diff. [95% CI]	Cliff's δ [95% CI]
DRMI Land.	GURO	τ	**+.363 [.180, .546]**	+0.72 [+0.44, +0.94]
		ρ	**+.416 [.072, .760]**	+0.86 [+0.65, +1.00]
	Bayes CJ	τ	**+.336 [.174, .498]**	+0.72 [+0.45, +0.92]
		ρ	**+.226 [.167, .285]**	+0.95 [+0.85, +1.00]
	LBPS-EIC	τ	+.013 [−.068, .094]	+0.17 [−0.19, +0.51]
		ρ	**−.053 [−.090, −.016]**	−0.58 [−0.84, −0.27]
DRMI Animal	GURO	τ	**+.269 [.074, .464]**	+0.57 [+0.27, +0.83]
		ρ	**+.472 [.200, .744]**	+0.82 [+0.54, +1.00]
	Bayes CJ	τ	+.077 [−.058, .212]	+0.27 [−0.09, +0.60]
		ρ	−.016 [−.054, .022]	−0.36 [−0.68, −0.00]
	LBPS-EIC	τ	+.058 [−.073, .189]	+0.06 [−0.30, +0.41]
		ρ	+.016 [−.024, .056]	−0.00 [−0.37, +0.35]
MetaF Animal	GURO	τ	**+.522 [.290, .754]**	+0.68 [+0.37, +0.93]
		ρ	**+.799 [.392, 1.206]**	+0.72 [+0.43, +0.97]
	Bayes CJ	τ	**+.182 [.058, .306]**	+0.71 [+0.46, +0.90]
		ρ	**+.035 [.003, .067]**	+0.37 [+0.01, +0.70]
	LBPS-EIC	τ	**+.286 [.137, .435]**	+0.53 [+0.21, +0.81]
		ρ	**+.184 [.131, .237]**	+0.80 [+0.58, +0.96]
MetaF Land.	GURO	τ	+.057 [−.025, .139]	−0.01 [−0.37, +0.35]
		ρ	**−.058 [−.090, −.026]**	−0.87 [−0.99, −0.69]
	Bayes CJ	τ	**+.196 [.039, .353]**	+0.59 [+0.27, +0.86]
		ρ	+.023 [−.011, .057]	+0.04 [−0.32, +0.40]
	LBPS-EIC	τ	**+.183 [.087, .279]**	+0.87 [+0.70, +0.98]
		ρ	**+.064 [.021, .107]**	+0.45 [+0.11, +0.76]

as descriptive uncertainty summaries rather than formal hypothesis tests.

D Preliminary Reconstruction-Adaptation Pilot

As a preliminary qualitative check, we fine-tuned three DRMI reconstruction variants using MetaRanker-, PSNR-, or SSIM-ranked supervision and evaluated them on held-out DRMI images. The pilot suggests that MetaRanker-ranked supervision can improve perceptual metrics such as LPIPS and NIQE, whereas SSIM-ranked supervision tends to favor distortion-oriented scores. Because this experiment is limited in scale, we treat it as suggestive evidence only; large-scale reconstruction training and closed-loop co-design validation remain future work (Section 6).

E Prior Corruption Stress Test

A potential concern is whether MetaRanker's self-correcting property holds when the VLM prior is severely miscalibrated. To stress-test this behavior, we construct adversarially corrupted priors by replacing a fraction p of the prior scores with inverted scores, where p=0 denotes the original VLM prior and p=1.0 denotes a fully inverted prior. We then measure the final Kendall's τ under a controlled synthetic-oracle setting (N=600, B=1,800, 3 runs per level). We use PSNR only as a deterministic oracle for this stress test, not as a proxy for human perceptual quality.

The ranking degrades gracefully under moderate corruption ($\Delta\tau$=−0.037 at p=0.5), and remains usable even under a fully adversarial prior ($\Delta\tau$=−0.078). This supports a *bounded* robustness claim: the prior is used only to guide early candidate pair generation, not as a supervisory label or rating initialization. When human feedback contradicts the prior-guided sampling trajectory, the prediction error inflates volatility and rating uncertainty (Eq. 2), redirecting subsequent queries toward uncertain items. A fully inverted prior is an extreme stress case unlikely in practice, since even poorly calibrated VLMs typically retain some semantic signal on non-degenerate images.