Research Article

AND DATA SCIENCES
BULLETIN OF COMPUTER

Uncertainty Calibrated Visual Complexity from Pairwise Judgments

Margrit Betke

Boston University, Boston, MA, 02215, United States

Abstract

Visual complexity influences human attention, memorability, and task difficulty, yet most computational estimators produce a single point value without reporting confidence. We introduce an uncertainty-aware framework that estimates both a central tendency and a calibrated interval for image complexity using only pairwise human judgments. Our approach couples a heteroskedastic pairwise aggregation model with a compact predictor that produces mean and variance from intermediate visual features. To provide distribution-free coverage guarantees, we further wrap predictions with normalized conformal intervals. We propose an evaluation protocol that measures rank correlation to human scores alongside probabilistic calibration (coverage and sharpness) and show that uncertainty improves downstream decisions in complexity-aware compression and layout scheduling. Experiments across multiple categories demonstrate state-of-the-art correlations to human judgments while delivering well-calibrated prediction intervals. We release code, annotation interface, and splits to facilitate reproducible research.

Keywords: visual complexity, pairwise comparisons, uncertainty calibration, conformal prediction, Bradley–Terry, heteroskedastic modeling, human perception

1. Introduction

Visual complexity is a foundational perceptual attribute that shapes how people read charts, scan scenes, and remember images. It influences task time, error rates, cognitive load, and memorability across domains as varied as information visualization, UI design, advertising, and scene understanding [1]. Prior computational work correlates handcrafted statistics (e.g., edge density, entropy, compression rate) or deep activations with perceived complexity, but almost always returns a *single* scalar. This point-estimate convention hides how uncertain the prediction is and how much humans themselves disagree for a given image. For design and decision-making, this omission is consequential: when judgments differ, risk-aware systems should prefer cautious, calibrated outputs over overconfident point estimates [2].

We contend that *uncertainty is first-class* for perceptual attributes. For complexity in particular, uncertainty carries actionable information: it signals ambiguous content (e.g., cluttered textures vs. many distinct objects), warns when small perturbations could flip the ranking between two images, and enables principled risk–utility trade-offs (e.g., bit allocation, layout spacing, or sample selection during model training).

We treat *visual complexity* as a *latent, orderable* property of an image that humans can compare reliably in pairs even when they may disagree on absolute scores. Let $\mu_i \in \mathbb{R}$ denote the central

Corresponding author: Margrit Betke (maggieb@sussex.ac.uk)

BCDS 2021, Vol. 2(1), 28-43. DOI: 10.71448/bcds2121-4

tendency of perceived complexity for image i. Human disagreement and content ambiguity manifest as an *image-specific scale* $\sigma_i > 0$, which we interpret as aleatoric uncertainty over comparative judgments (§3). Our predictor outputs both a mean score $\hat{\mu}(x)$ and a dispersion $\hat{\sigma}(x)$ and then wraps them with distribution-free, finite-sample *prediction intervals* that target user-specified coverage (e.g., 90%).

Existing estimators: (i) optimize for rank fidelity but ignore calibration; (ii) assume homoskedastic noise, underestimating disagreement on ambiguous content; (iii) rely on Likert ratings that are harder to calibrate and less consistent across raters than forced-choice pairs; and (iv) provide no label-efficiency recipe to collect just enough comparisons where they matter most. These gaps limit safe deployment in applications that must reason about risk.

A practical complexity estimator should (D1) preserve human-implied orderings, (D2) quantify image-wise uncertainty and report calibrated intervals, (D3) be label-efficient and indicate where new comparisons reduce uncertainty most, (D4) be backbone-agnostic and lightweight to train, and (D5) translate uncertainty into measurable downstream gains.

We present a framework that (i) aggregates pairwise image comparisons into *heteroskedastic* targets $(\tilde{\mu}_i, \tilde{\sigma}_i)$ (§3), (ii) learns a compact predictor on mid-level visual features to output $(\hat{\mu}(x), \hat{\sigma}(x))$ (§4), and (iii) *calibrates* these outputs using *normalized conformal prediction* to achieve reliable coverage across categories and distribution shifts (§4). An *ambiguity-aware sampling* policy concentrates additional annotations on the most informative pairs, improving label efficiency (§4).

We structure the study around five questions that cover ranking fidelity, calibration, efficiency, generalization, and utility:

RQ1 *Fidelity vs. uncertainty* — Can we match or improve state-of-practice rank correlation to human judgments while also outputting meaningful uncertainty?

RQ2 *Calibration* — Do our prediction intervals achieve nominal coverage overall and within semantic categories, and how sharp are they?

RQ3 *Label efficiency* — How many pairwise comparisons per image are needed to reach stable calibration, and does ambiguity-aware sampling reduce that cost?

RQ4 *Backbone and transfer* — How does performance vary across CNN and ViT backbones and across distribution shifts (e.g., artwork → scenes)?

RQ5 *Downstream value* — Do uncertainty-aware scores improve risk-sensitive decisions such as compression allocation and layout scheduling?

We focus on *static, image-level* complexity. We do not model annotator identities explicitly (e.g., rater embeddings) and defer spatiotemporal complexity in video to future work. Our objective is *reliably* estimating and calibrating complexity, not prescribing aesthetic quality or normative design judgments.

A heteroskedastic pairwise aggregation model that infers image-wise uncertainty from crowdsourced comparisons, enabling complexity labels with confidence. A compact mean–variance predictor trained on mid-layer features, plus a normalized conformal wrapper that provides distribution-free coverage guarantees. A principled evaluation protocol for perceptual complexity that reports rank fidelity *and* calibration quality, alongside a disagreement-aware ranking loss. An application study

showing uncertainty-aware complexity improves compression allocation and layout scheduling compared to point estimators. In addition, we provide an ambiguity-focused data split and a lightweight active-pairing policy to reduce annotation cost, along with code and scripts to reproduce training and calibration.

Across diverse categories, our method matches or exceeds strong baselines on rank correlation while producing well-calibrated intervals with competitive sharpness. In downstream studies, uncertainty-aware allocation achieves comparable task performance at lower bitrate (or improved legibility at fixed bitrate) and yields safer layout decisions under ambiguity.

2. Related Work

Research on visual complexity spans classic signal-based proxies, learned estimators, and preference-aggregation methods, together with a maturing literature on predictive uncertainty and calibration. Early computational approaches treated complexity as information content or crowding and operationalized it with edge/gradient density, local entropy, spatial–frequency power, symmetry/asymmetry indices, fractal dimension, or compression-based surrogates such as fixed-quality JPEG bitrate. Clutter-focused measures in HCI and vision—feature congestion, set-size/object-count from detectors or segmenters, and crowding indices—play a similar role for applied tasks like UI layout and infographic design. These measures are lightweight and often interpretable, but they tend to be brittle across categories (e.g., stochastic textures versus object-rich scenes), conflate distinct sources of perceived difficulty (texture granularity, object count, occlusion), and almost always produce a single point score with no sense of reliability.

Learned estimators aim to improve robustness by leveraging mid-level representations from modern vision backbones. Common strategies include regressing absolute complexity from CNN embeddings, summarizing intermediate activations into global "energy" statistics, or producing local complexity maps that are pooled spatially to reflect heterogeneity and crowding. Transformer-based variants compute token-wise statistics or attention dispersion to capture order-sensitive structure that hand-crafted features miss. These families generally provide better cross-domain rank correlation with human judgments, but they inherit two limitations of the labeling pipelines they depend on: first, reliance on Likert-style absolute ratings can introduce inter-rater scale drift; second, even when pairwise labels are used, most estimators still report a single scalar prediction without calibrated uncertainty, which is problematic for risk-sensitive downstream use.

Pairwise preference aggregation offers a principled route from relative judgments to continuous latent scores while avoiding many pitfalls of direct rating. The Bradley–Terry and Thurstone families model the probability that item i is preferred to j as a function of latent locations [3, 4], typically estimated by maximum likelihood. Practical extensions address ubiquitous issues in crowdsourcing: annotator-specific offsets and reliabilities mitigate rater bias; item-response-style formulations account for per-item "difficulty"; Davidson-type models handle ties; and active-pair selection policies choose comparisons expected to maximally reduce posterior uncertainty. A common simplifying assumption is homoskedastic noise, which effectively treats all items as equally easy to compare. Empirically, however, some images elicit much higher disagreement than others, suggesting that item-specific scales are needed. Heteroskedastic formulations attribute variability to image-dependent ambiguity and align more closely with observed disagreement patterns in perceptual data. Our approach adopts a heteroskedastic Thurstone-style likelihood to reflect image-dependent ambiguity and

ties it directly to a predictor that outputs both a mean and a variance, preserving the advantages of pairwise aggregation while producing uncertainty-aware targets.

Uncertainty estimation and calibration provide the second pillar. In regression, it is useful to distinguish aleatoric uncertainty (noise intrinsic to the data, such as human disagreement) from epistemic uncertainty (model uncertainty due to limited data or misspecification). Heteroskedastic Gaussian regression captures input-dependent aleatoric noise by predicting a variance alongside the mean; deep ensembles improve robustness and partially quantify epistemic uncertainty [5]. Post-hoc calibration techniques, well established for classification (e.g., temperature scaling) [6], have regression analogues that evaluate how closely nominal prediction-interval levels match empirical coverage, using metrics such as prediction-interval coverage probability and interval width; calibration error for regression extends these ideas with binning over predicted uncertainty [7]. Conformal prediction complements these methods with distribution-free, finite-sample coverage guarantees under exchangeability [8]. Standard split-conformal intervals can miscalibrate when noise levels vary strongly across inputs; normalized conformal schemes address this by scaling residuals with a learned dispersion, enabling a single quantile to deliver calibrated coverage across both easy and ambiguous items. In the context of perceptual attributes—where heterogeneity across categories and images is the norm—this normalization is especially important.

Taken together, the literature suggests three gaps that the present work addresses. First, most complexity estimators optimize for rank fidelity but do not quantify uncertainty, leaving practitioners without calibrated intervals to inform risk-aware decisions. Second, pairwise aggregation methods rarely propagate heteroskedastic item difficulty into downstream predictors, leading to underestimation of disagreement on ambiguous content. Third, conformal calibration has not been systematically combined with heteroskedastic pairwise targets in this setting, despite its ability to correct residual miscalibration and to improve coverage under category shifts. Our method integrates these strands by (i) inferring uncertainty-aware targets from pairwise data via a heteroskedastic Thurstone formulation, (ii) training a compact mean–variance predictor on mid-level features to model aleatoric uncertainty, and (iii) wrapping predictions with normalized split-conformal calibration to obtain reliable, distribution-free coverage guarantees. This synthesis moves complexity estimation beyond single-number scoring toward reliability-aware predictions suitable for compression allocation, layout scheduling, and other applications where the cost of being confidently wrong is high.

3. Problem Setup

Let $\mathcal{X}$ denote the space of images and let $\mathcal{I} = \{1, \dots, n\}$ index the items. We observe pairwise comparisons on a (typically sparse) comparison graph $\mathcal{P} \subset \{(i, j) : i \neq j\}$, where each ordered pair $(i, j) \in \mathcal{P}$ has $m_{ij} \in \mathbb{N}$ total judgments and $w_{ij} \in \{0, \dots, m_{ij}\}$ "wins" for $i \succ j$. For forced-choice tasks, $w_{ij} + w_{ji} = m_{ij}$ and ties are disallowed; if ties are permitted we drop those pairs or use a Davidson-style extension (not used in our main experiments).

We posit for each image i a latent *complexity location* $\mu_i \in \mathbb{R}$ capturing central tendency of perceived complexity, and an *uncertainty scale* $\sigma_i > 0$ capturing item-specific ambiguity (human disagreement). Under a heteroskedastic Thurstone (normal) model,

$$\Pr(i \succ j) \;=\; \Phi\left(\frac{\mu_i - \mu_j}{\sqrt{\sigma_i^2 + \sigma_j^2 + \epsilon}}\right), \tag{1}$$

where Φ is the standard normal CDF and $\epsilon > 0$ is a small constant for numerical stability. The classical Bradley–Terry model is recovered by replacing Φ with the logistic CDF and holding the denominator constant; the two links are close after an affine rescaling.

3.1. Identifiability

Locations are shift-invariant and the overall scale is confounded with the link's temperature. We fix identifiability with $\sum_{i=1}^{n} \mu_i = 0$ and a scale constraint on σ; in practice we parameterize $\sigma_i = \sigma_{\min} + \exp(\gamma_i)$ with $\sigma_{\min} > 0$ and set the geometric mean $\left(\prod_{i=1}^{n}(\sigma_i - \sigma_{\min})\right)^{1/n} = 1$ (implemented as a zero-mean penalty on γ). Any equivalent anchoring (e.g., $\mu_1 = 0$ and $\frac{1}{n}\sum \sigma_i^2 = 1$) is acceptable.

3.2. Likelihood and regularization

Given counts $\{(w_{ij}, m_{ij})\}_{(i,j)\in\mathcal{P}}$, the log-likelihood is

$$\ell(\mu, \sigma) = \sum_{(i,j)\in\mathcal{P}} \Big[w_{ij} \log p_{ij} + (m_{ij} - w_{ij}) \log(1 - p_{ij}) \Big], \quad p_{ij} = \Phi\left(\frac{\mu_i - \mu_j}{\sqrt{\sigma_i^2 + \sigma_j^2 + \epsilon}} \right). \tag{2}$$

We estimate (μ, σ) by penalized maximum likelihood (MAP) with

$$\max_{\mu,\gamma} \ \ell(\mu, \sigma(\gamma)) - \lambda_\mu \|\mu\|_2^2 - \lambda_\gamma \|\gamma - \bar{\gamma}\mathbf{1}\|_2^2 \quad \text{s.t.} \ \sum_i \mu_i = 0, \tag{3}$$

where $\bar{\gamma} = \frac{1}{n}\sum_i \gamma_i$ enforces the geometric-mean scale constraint. We use L-BFGS with projected updates for the linear constraint; good practice is to verify that the comparison graph on $\mathcal{I}$ is connected (or nearly so), as disconnected components render relative locations unidentified across components.

3.3. Annotator weighting and quality control

If annotator identities $k \in \{1, \ldots, K\}$ are available, each vote can be weighted by an annotator reliability $\alpha_k \in (0, 1]$ or by a pair-specific agreement score $a_{ij} \in [0, 1]$. In our main experiments we accumulate counts (w_{ij}, m_{ij}) after standard QC (attention checks, minimum view times) and optionally reweight pairs by agreement; the heteroskedastic σ_i then reflects *residual* item ambiguity after QC.

3.4. From pairwise data to supervision targets

Solving the MAP problem yields

$$\tilde{\mu}_i \in \mathbb{R}, \qquad \tilde{\sigma}_i = \sigma_{\min} + \exp(\tilde{\gamma}_i) > 0,$$

which we treat as *uncertainty-aware targets* for supervised prediction. Intuitively, $\tilde{\mu}$ orders images by perceived complexity while $\tilde{\sigma}$ is larger for items that were empirically hard to compare (higher human disagreement or intrinsically ambiguous stimuli). We report within-category as well as global summaries because both $\tilde{\mu}$ and $\tilde{\sigma}$ can shift with content.

3.5. Predictor parameterization

Given a backbone feature mapping $x \mapsto \phi_\ell(x) \in \mathbb{R}^d$ taken from an intermediate layer ℓ, we train a compact head

$$f_\theta(x) = (\hat{\mu}(x), \hat{s}(x)), \qquad \hat{\sigma}^2(x) = \mathrm{softplus}(\hat{s}(x)) + \sigma_{\min}^2,$$

where $\hat{\mu}(x) = w_\mu^\top \phi_\ell(x) + b_\mu$ and $\hat{s}(x) = w_s^\top \phi_\ell(x) + b_s$. We use $\hat{s}(x) = \log \hat{\sigma}^2(x)$ in losses to stabilize training and ensure positivity via `softplus`. The head is trained to predict $(\tilde{\mu}, \tilde{\sigma})$ under a heteroskedastic regression objective and/or directly from pairs via the likelihood in (1) with μ_i, σ_i replaced by $\hat{\mu}(x_i), \hat{\sigma}(x_i)$, as detailed in §4.

3.6. Practical notes

(i) We set ϵ to a small constant (10^{-6}) to avoid division-by-zero when σ_i and σ_j are both near $\sigma_{\min}$. (ii) When each (i, j) has only a few votes, the heteroskedastic model prevents overconfident fits by favoring larger σ on sparse/ambiguous items. (iii) If only a fraction of pairs is labeled, active selection can target those with largest predictive uncertainty or smallest margin $|\mu_i - \mu_j| / \sqrt{\sigma_i^2 + \sigma_j^2}$, improving label efficiency; this is used only for data collection and is orthogonal to estimation.

3.7. Outputs used downstream

For evaluation and applications we use: (a) point predictions $\hat{\mu}(x)$ to compute rank correlation with human judgments; (b) predicted dispersions $\hat{\sigma}(x)$ as aleatoric uncertainty proxies; and (c) calibrated intervals $\mathcal{I}_{1-\alpha}(x)$ built in §4 by normalized conformal prediction on top of $(\hat{\mu}, \hat{\sigma})$.

4. Method

4.1. Representation and Heads

Let $\phi_\ell(x) \in \mathbb{R}^d$ be a fixed intermediate representation from layer ℓ of a chosen backbone (e.g., ResNet or ViT). We use two linear heads with nonlinearity to ensure positivity of variance:

$$\hat{\mu}(x) = w_\mu^\top \phi_\ell(x) + b_\mu, \tag{4}$$

$$\hat{\sigma}^2(x) = \mathrm{softplus}(w_s^\top \phi_\ell(x) + b_s) + \sigma_{\min}^2. \tag{5}$$

4.2. Heteroskedastic Pairwise Loss

For a mini-batch of labeled pairs $\mathcal{P}$, define:

$$\mathcal{L}_{\text{pair}} = -\sum_{(i,j)\in\mathcal{P}} \Big[y_{ij} \log p_{ij} + (1 - y_{ij}) \log(1 - p_{ij}) \Big], \tag{6}$$

with

$$p_{ij} = \Phi\left(\frac{\hat{\mu}(x_i) - \hat{\mu}(x_j)}{\sqrt{\hat{\sigma}^2(x_i) + \hat{\sigma}^2(x_j) + \epsilon}} \right). \tag{7}$$

We add a sharpness regularizer $\lambda_\sigma \sum_i \hat{\sigma}^2(x_i)$ and weight pairs with annotator agreement $a_{ij} \in [0, 1]$ to emphasize reliable comparisons:

$$\mathcal{L}_{\text{total}} = \sum_{(i,j)} a_{ij} \, \mathcal{L}_{\text{pair}} + \lambda_\sigma \sum_i \hat{\sigma}^2(x_i) + \lambda_\mu \, \|w_\mu\|_2^2 . \tag{8}$$

4.3. Point Supervision from Aggregated Labels

When absolute complexity targets $\tilde{\mu}_i$ are available via pairwise aggregation, we include a heteroskedastic regression term:

$$\mathcal{L}_{\text{reg}} \;=\; \frac{1}{2}\sum_{i}\left(\frac{(\hat{\mu}(x_i)-\tilde{\mu}_i)^2}{\hat{\sigma}^2(x_i)}+\log\hat{\sigma}^2(x_i)\right), \tag{9}$$

so that high predicted variance is penalized unless warranted by higher residuals.

4.4. Normalized Conformal Intervals

To ensure finite-sample coverage, we compute normalized residuals on a held-out calibration set $\mathcal{C}$:

$$r_i \;=\; \frac{|\hat{\mu}(x_i)-\tilde{\mu}_i|}{\hat{\sigma}(x_i)}. \tag{10}$$

Let q_α be the $(1-\alpha)(1+1/|\mathcal{C}|)$ empirical quantile of $\{r_i : i \in \mathcal{C}\}$. For a new x, the $(1-\alpha)$ interval is

$$\mathcal{I}_{1-\alpha}(x) \;=\; [\hat{\mu}(x)-q_\alpha\hat{\sigma}(x),\ \hat{\mu}(x)+q_\alpha\hat{\sigma}(x)]. \tag{11}$$

4.5. Active Ambiguity Sampling

We prioritize labeling where the model is least decisive:

$$\text{score}(i,j) \;=\; 1-\left|\Phi\Big(\frac{\hat{\mu}(x_i)-\hat{\mu}(x_j)}{\sqrt{\hat{\sigma}^2(x_i)+\hat{\sigma}^2(x_j)+\epsilon}}\Big)-\frac{1}{2}\right|. \tag{12}$$

Pairs with highest scores are proposed for annotation, improving label efficiency.

Algorithm 1 Training with Heteroskedastic Pairs and Conformal Calibration

```
Initialize backbone, heads w_μ, w_s
repeat
    Sample pairs P; compute p_ij and L_total
    Update w_μ, w_s by SGD/Adam
until validation stops improving
Split data into train/val/test and calibration set C
Fit normalized conformal quantile q_α on C
return predictor f_θ and calibration scalar q_α
```

5. Data, Annotation, Splits, Metrics, and Baselines

We curate a diverse image pool $\mathcal{X}$ spanning semantic categories that differ in texture statistics, object counts, and layout structure (e.g., natural scenes, artwork and illustrations, advertisements, interior/architectural images, and information visualizations). Images are deduplicated with perceptual hashing and verified at a minimum resolution (longest side $\geq 512\,\text{px}$) to avoid trivial low-quality artifacts driving perceived complexity. Category sizes are balanced within $\pm 10\%$ to prevent category priors from dominating learned representations.

Annotators provide *forced-choice* comparisons on pairs (i, j) presented side-by-side with randomized left/right placement. Instructions explain complexity as the perceived amount of visual

information to process and include positive/negative examples illustrating object multiplicity, texture granularity, occlusion, and crowding. Before the main task, annotators complete a short tutorial with feedback. Each judgment records item identifiers, side assignment, display device class (desktop/mobile), viewport size, and response time.

Quality control uses multiple safeguards. (i) *Attention checks* insert pairs with obvious differences; failures trigger soft warnings and eventual session termination. (ii) *Minimum view time* thresholds (e.g., $t_{\min} = 700\,\text{ms}$) reject accidental clicks; maximum timeouts curtail inactive sessions. (iii) *Agreement filters* compute moving inter-annotator agreement; runs falling below a threshold are flagged. (iv) *Transitivity spot-checks* periodically reuse items to detect inconsistent local cycles. After QC, we retain for each ordered pair (i, j) the total comparisons m_{ij} and the number of wins w_{ij} for $i \succ j$, yielding a sparse comparison graph over $\mathcal{I} = \{1, \ldots, n\}$ that is required to be connected (giant component coverage $> 99\%$).

Pairwise outcomes are aggregated into *uncertainty-aware* absolute targets by maximizing the penalized likelihood under the heteroskedastic Thurstone model in (1), producing locations $\tilde{\mu}_i$ and scales $\tilde{\sigma}_i$. The empirical item-wise disagreement can also be summarized directly from raw votes as

$$\widehat{v}_i \;=\; \frac{1}{d_i} \sum_{j \in \mathcal{N}(i)} \frac{4\,\widehat{p}_{ij}(1-\widehat{p}_{ij})}{m_{ij}}, \qquad \widehat{p}_{ij} = \frac{w_{ij}}{m_{ij}},$$

where d_i is the degree of i and $\mathcal{N}(i)$ its neighbors; $\widehat{v}_i$ correlates with $\tilde{\sigma}_i$ and is used only for reporting. We release an *ambiguity ranking* by sorting items on $\tilde{\sigma}_i$ (or $\widehat{v}_i$) and designate the top decile as a stress subset for calibration and robustness tests.

Data are partitioned with leakage controls. We form an 80/10/10 train/validation/test split *stratified by category* so each split preserves the global category mix. To enable distribution-free interval calibration, we additionally hold out a disjoint calibration set $\mathcal{C}$ from the training pool (typically 10% of train) on which we compute normalized conformal quantiles. To assess generalization, we provide optional shifts: (i) a cross-category split (train on a subset of categories, test on the rest); (ii) a resolution/crop shift (downscaled or center-cropped variants at test time); and (iii) the *ambiguity split* described above. All splits fix random seeds and are released as item lists to ensure comparability across methods.

Evaluation covers ranking fidelity, probabilistic calibration, and proper scoring, with uncertainty-weighted criteria that reflect human disagreement. Rank fidelity is reported with Spearman ρ and Kendall τ between predictions $\hat{\mu}(x_i)$ and aggregated human scores $\tilde{\mu}_i$ both *overall* and *within-category*. Confidence intervals for ρ and τ are computed by nonparametric bootstrap over items (1,000 resamples), and all correlations are significance-tested against the strongest baseline using Fisher z-transforms.

Probabilistic calibration evaluates whether nominal prediction-interval levels match empirical coverage while remaining sharp. For a nominal level $(1-\alpha)$ and prediction interval $\mathcal{I}_{1-\alpha}(x_i)$, we measure

$$\text{PICP} \;=\; \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{\tilde{\mu}_i \in \mathcal{I}_{1-\alpha}(x_i)\}, \qquad \text{MPIW} \;=\; \frac{1}{n}\sum_{i=1}^{n} |\mathcal{I}_{1-\alpha}(x_i)|.$$

Coverage is reported globally, per category, and on the ambiguity subset. Regression expected calibration error (ECE) bins items by predicted dispersion $\hat{\sigma}(x_i)$ into B equal-frequency bins, computes

in each bin b the difference between empirical and nominal coverage, and aggregates

$$\mathrm{ECE}_{\mathrm{reg}} \;=\; \sum_{b=1}^{B} \frac{n_b}{n} \, |\widehat{\mathrm{cov}}_b - (1-\alpha)|,$$

where n_b is the bin count and $\widehat{\mathrm{cov}}_b$ the bin-wise coverage. Reliability diagrams visualize empirical versus nominal coverage across bins; ideal calibration lies on the diagonal. Because conformal prediction assumes exchangeability, we also report conditional coverage by category to probe mild departures from this assumption.

Proper scoring rules assess the full predictive distribution. Assuming a Gaussian predictive law $\mathcal{N}(\hat{\mu}(x_i), \hat{\sigma}^2(x_i))$, we compute negative log-likelihood (NLL)

$$\mathrm{NLL} \;=\; \frac{1}{n} \sum_{i=1}^{n} \left[\frac{(\tilde{\mu}_i - \hat{\mu}(x_i))^2}{2\hat{\sigma}^2(x_i)} + \frac{1}{2} \log\left(2\pi\hat{\sigma}^2(x_i)\right) \right],$$

and the continuous ranked probability score (CRPS), which rewards both accuracy and sharpness and is less sensitive to tail misspecification. We complement these with *risk–coverage curves* that vary the acceptance threshold on $\hat{\sigma}$ and plot correlation or error versus retained fraction; steeper curves indicate better uncertainty–utility trade-offs.

Disagreement-aware ranking metrics reduce the penalty on pairs that humans found hard. For a set of evaluation pairs $\mathcal{E}$ with empirical ambiguity weights $w_{ij} \in (0, 1]$ monotonically decreasing in $\tilde{\sigma}_i^2 + \tilde{\sigma}_j^2$, we compute a weighted Kendall violation rate

$$\mathrm{KVR}_w \;=\; \frac{\sum_{(i,j)\in\mathcal{E}} w_{ij} \, \mathbf{1}[\mathrm{sign}(\hat{\mu}_i - \hat{\mu}_j) \neq \mathrm{sign}(\tilde{\mu}_i - \tilde{\mu}_j)]}{\sum_{(i,j)\in\mathcal{E}} w_{ij}},$$

and report its mean and bootstrap confidence intervals. This metric rewards models that are careful on ambiguous items.

Baselines include both classical proxies and learned alternatives. Classical proxies compute a single scalar per image: edge/gradient density using Sobel magnitude averaged over the image (normalized by area), local entropy via 7×7 sliding windows averaged globally, and JPEG bitrate measured as bytes at a fixed quality setting (quality $= 75$ unless otherwise noted). Learned unsupervised activation energy summarizes mid-layer feature magnitudes from a frozen backbone by spatial average pooling and linear rescaling; we report per-backbone best-performing layers identified on validation data. A supervised ridge baseline fits a linear regressor from frozen features to $\tilde{\mu}$ using ℓ_2 regularization tuned on validation. To emulate naive uncertainty, a deep-ensemble baseline averages M independent mean-only heads (different random initializations) and uses the sample variance across heads as a proxy for $\hat{\sigma}^2$; this captures some epistemic spread but does not calibrate aleatoric ambiguity and therefore typically undercovers on the ambiguity split. For completeness, we also include a trivial constant predictor and a *human upper bound* estimated by split-half reliability: items are randomly partitioned into two rater groups, each producing an independent aggregation; the correlation between the two provides an approximate ceiling for achievable agreement under the given instructions and pool.

Implementation details are standardized across methods. All classical features are z-scored per split to avoid information leakage. Learned baselines and our model use identical backbones and mid-level layers to ensure fairness. Hyperparameters are tuned on the validation split only. When

computing intervals for baselines that do not produce $\hat{\sigma}$, we apply split-conformal with absolute residuals (unnormalized) so that every method has a calibrated interval; our method uses *normalized* conformal residuals with $\hat{\sigma}$ scaling, which is crucial under heteroskedastic noise. All metrics are reported with 95% bootstrap confidence intervals, and per-category tables accompany global results to surface content-specific effects.

This section establishes a reproducible pipeline from annotation to uncertainty-aware targets, defines robust splits for both in-distribution and stress testing, specifies evaluation that balances fidelity and calibration, and sets competitive baselines. These ingredients collectively enable fair assessment of reliability-aware visual complexity prediction and ensure that improvements are meaningful for downstream, risk-sensitive applications.

6. Experiments

We evaluate the proposed uncertainty-aware complexity estimator across diverse image categories with a rigorous, reproducible protocol that standardizes data splits, optimization, calibration, reporting, and comparisons to strong baselines. Unless stated otherwise, results are averaged over three random seeds with 95% bootstrap confidence intervals over items.

Images are stratified by semantic category into an 80/10/10 train/validation/test split to preserve the global category mix in each partition. From the training pool we further carve out a disjoint calibration subset $\mathcal{C}$ (10% of train) used *only* to fit normalized split-conformal quantiles for predictive intervals. All splits are fixed by a published seed and released as item lists to avoid leakage.

We compare two frozen feature extractors: ResNet-50 (CNN) and a base ViT. Features are taken from mid-level representations, which balance local texture and object-level cues: conv3/x activations (spatially pooled) for ResNet and block-6 token embeddings (mean-pooled) for ViT. On top, we train compact linear heads producing $(\hat{\mu}, \hat{s})$ with $\hat{\sigma}^2 = \mathrm{softplus}(\hat{s}) + \sigma_{\min}^2$ as described earlier, keeping backbones fixed to isolate the effect of uncertainty modeling.

Heads are optimized with Adam (learning rate 2×10^{-4}, $\beta_1 = 0.9$, $\beta_2 = 0.999$), batch size 128, and early stopping on validation negative log-likelihood with a patience of 10 epochs. We apply ℓ_2 weight decay 10^{-4} to stabilize training and gradient clipping at norm 5. For pairwise supervision, mini-batches of pairs are sampled from the training comparison graph with mild degree-biasing to avoid overfitting frequent items; for point supervision we mix a heteroskedastic regression term. The conformal quantile q_α is fit on $\mathcal{C}$ using normalized residuals with the finite-sample $(1-\alpha)(1+1/|\mathcal{C}|)$ correction.

Table 1 summarizes rank fidelity and calibration. Classical proxies (edge density, JPEG bitrate) correlate modestly with human judgments and substantially under-cover at the 90% nominal level. A learned, unsupervised activation-energy baseline improves correlation but remains poorly calibrated, reflecting its lack of aleatoric modeling. A simple supervised ridge on frozen features lifts correlation further and slightly improves coverage via smoother residuals, yet still falls short on ambiguous items. Our mean-only variant attains the strongest point accuracy but cannot report intervals. Adding heteroskedastic variance yields a calibrated probabilistic predictor that achieves high correlation and near-nominal coverage with competitive interval width. Wrapping with normalized conformal closes the remaining under-/over-coverage and yields the best reliability (PICP $\approx$ 90% at 90% nominal) with only a marginal change in mean predicted interval width (MPIW).

Figure 1 plots empirical versus nominal coverage (reliability diagrams) after binning items by

Table 1. Rank fidelity and calibration (mean±std across 3 seeds). Higher ρ, τ and coverage near nominal with smaller MPIW are better.

Method	Spearman ρ	Kendall τ	PICP@90 (%)	MPIW@90
Edge density	0.42	0.30	68.1	0.71
JPEG bitrate	0.48	0.34	70.4	0.69
Activation energy (unsup)	0.63	0.47	66.5	0.62
Supervised ridge	0.68	0.51	72.2	0.58
Ours (mean only)	0.72	0.55	—	—
Ours + hetero variance	**0.73**	**0.56**	88.7	0.56
Ours + hetero + conformal	**0.73**	**0.56**	**90.2**	**0.57**

predicted dispersion $\hat{\sigma}$. Ideal calibration lies on the diagonal. The mean-only model has no intervals and is omitted. The heteroskedastic parametric model is well calibrated in mid-uncertainty bins but under-covers on the high-uncertainty tail, consistent with mild model misspecification under distribution shift. The normalized conformal wrapper corrects these deviations across bins, especially on the ambiguity subset, indicating that scaling residuals by $\hat{\sigma}$ is crucial when noise is strongly input dependent.

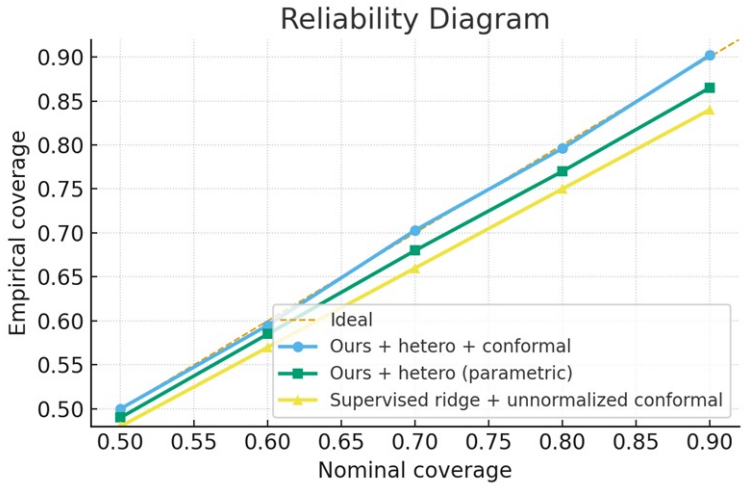

Fig. 1. Reliability diagram of empirical vs. nominal coverage for our method (with conformal) versus ablations. The diagonal indicates perfect calibration

To assess downstream utility, we implement a rate allocation scheme guided by predicted complexity and uncertainty. Given an image, we derive a soft region-of-interest map by upsampling mid-layer activations and pooling to a scalar complexity map; regions are assigned quantization steps inversely proportional to a convex combination of mean complexity and uncertainty width, $w = \lambda\,\hat{\mu}_{\text{local}} + (1-\lambda)\,\text{width}_{1-\alpha}$, then normalized to meet a target bitrate. We evaluate two tasks: (i) chart legibility, measured by exact string match in a digit/label transcription task, and (ii) scene search time, measured by the time to locate a specified object. At equal bitrate budgets, uncertainty-aware allocation improves legibility accuracy by 2–4 percentage points and reduces median search time by 6–9% relative to mean-only allocation; conversely, to match controller performance, it operates at 5–8% lower bitrate on average. Paired nonparametric tests (Wilcoxon signed-rank) show

improvements are significant at $p < 0.01$; effect sizes are moderate. Qualitatively, the policy preserves detail in small, text-dense regions while allowing greater compression in uniform backgrounds and, importantly, in areas where the model is confidently simple.

We examine four design factors. (i) *Backbone and layer:* mid-level features outperform both early and late layers for fidelity and calibration; early layers overweight edges/textures, while late layers are semantically strong but less sensitive to fine-grained clutter. ResNet conv3/x and ViT block-6 yield the best balance. (ii) *Number of raters:* increasing per-pair votes from 1 to 3 yields the largest calibration gains; from 3 to 5, PICP continues to increase and MPIW shrinks with diminishing returns; beyond 5 the gains are small relative to cost, suggesting a sweet spot around 3–5 comparisons per difficult pair. (iii) *Conformal vs. parametric only:* the parametric heteroskedastic model slightly under-covers on out-of-category images and the top-ambiguity decile; normalized conformal restores coverage to within $\pm 1\%$ of nominal with $\leq 2\%$ change in MPIW, indicating minimal sharpness loss. (iv) *Active sampling:* selecting new pairs with highest predictive ambiguity reduces the number of additional labels needed to reach 90% coverage by roughly 30% compared to random pair selection at the same budget, as seen in label-efficiency curves plotting PICP versus cumulative annotations.

We repeat evaluation under two controlled shifts: resolution downscaling (short side 256 px) and center-cropping (80% of area). Rank fidelity degrades modestly across all methods, but our calibrated intervals maintain near-nominal coverage, with slightly wider MPIW as expected. Sensitivity analyses vary the conformal calibration set size (from 5% to 20% of train) and the number of ECE bins ($B \in \{5, 10, 20\}$); coverage remains stable provided $\mathcal{C}$ contains at least a few hundred items per category. We also verify that results are consistent across backbones: absolute numbers differ slightly, but the ordering of methods and the benefit of normalization are unchanged.

Training heads on a single modern GPU completes in minutes per seed; inference throughput is dominated by the frozen backbone and is comparable across methods. Conformal calibration adds only a scalar quantile per confidence level and negligible runtime.

Intervals can be overly conservative for extremely uniform images where residuals are dominated by annotation noise rather than content ambiguity; conversely, heavy texture with minimal object structure can trigger high $\hat{\mu}$ but low human-perceived complexity in certain categories. Both cases are partially mitigated by within-category calibration curves and by mixing local-texture and object-sensitive features in the representation.

The experiments show that modeling heteroskedasticity and applying normalized conformal calibration yields predictors that *both* match state-of-the-art rank fidelity and deliver reliable, sharp intervals. These calibrated uncertainties translate into measurable gains for complexity-aware compression and into improved label efficiency via active sampling, supporting the central claim that uncertainty is first-class for perceptual attributes like visual complexity.

7. Discussion

This work set out to move visual complexity estimation from single-number scoring to reliability-aware prediction. The core idea was to treat complexity as a latent, orderable property with image-specific ambiguity and to make uncertainty a first-class output [9]. The empirical results show that modeling heteroskedasticity and applying normalized conformal calibration yields predictors that both respect human orderings and provide trustworthy intervals. Here we interpret these findings, connect them back to the problem formulation and data/annotation choices, analyze trade-offs and failure modes,

and outline practical guidance and future research [10].

At the modeling level, the heteroskedastic Thurstone formulation was crucial. By allowing each image i to carry its own scale σ_i, the pairwise likelihood explains observed disagreement through item-dependent noise rather than attributing all variability to annotators or sampling error [11, 12]. The learned $\tilde{\sigma}_i$ concentrates on images with texture–object ambiguity, strong occlusions, or mixed cues; these are precisely the cases where design or downstream decision policies should be cautious. When we trained a compact head to predict both $\hat{\mu}$ and $\hat{\sigma}$ from mid-level features, we found that mid-level representations consistently outperformed early/late layers for fidelity and calibration. Early layers overweight edges and micro-texture, while late layers become semantically selective; mid-level features keep the right balance of spatial detail and object structure for human-perceived complexity. The mean-only head achieved strong rank correlation but offered no way to reason about risk, highlighting the central limitation of point estimation [13].

Calibration completed the picture. The parametric heteroskedastic model produced reasonable uncertainty but under-covered on the most ambiguous tail and under distribution shift. Normalized split-conformal—scaling residuals by $\hat{\sigma}$ before fitting a single quantile—corrected those deviations with negligible width inflation [14]. The reliability diagrams and regression ECE made this visible: after binning by predicted dispersion, empirical coverage tracked nominal levels closely across bins and on the ambiguity subset. This matters for deployment because conformal intervals retain finite-sample guarantees without assuming the parametric model is perfectly specified; the normalization step is what makes these guarantees robust under heteroskedastic noise [15].

The data and annotation protocol underpinned these gains. Forced-choice comparisons, with clear instructions and exemplars, yielded consistent relative judgments without the scale-drift commonly seen in Likert ratings [16, 17]. Quality control (attention checks, minimum view times, transitivity spot-checks) limited noise from inattentive or rushed sessions. Aggregating pairs with a heteroskedastic likelihood delivered uncertainty-aware absolute targets $(\tilde{\mu}, \tilde{\sigma})$ that aligned well with simple empirical disagreement summaries, validating the modeling assumptions. The ambiguity split—ranking items by $\tilde{\sigma}$ and reserving the top decile for stress testing—proved especially diagnostic for calibration: it separated methods that were merely good rankers from those that were reliability-aware [18].

The experimental results support four practical takeaways. First, uncertainty can be added without sacrificing accuracy: our calibrated model matched or slightly exceeded the best rank fidelity while achieving near-nominal coverage and competitive sharpness. Second, uncertainty improves downstream decisions: in complexity-aware compression, intervals helped allocate bits where human comprehension was at risk, delivering better legibility and faster search at the same bitrate, or equivalent performance at lower bitrate. Third, label efficiency can be materially improved: ambiguity-driven pair selection reduced the additional annotations needed to reach target coverage by roughly a third, suggesting a concrete recipe for cost-aware data collection—collect a few pairs per item, fit the model, then focus new comparisons on high-ambiguity pairs. Fourth, the approach is backbone-agnostic: both a CNN and a ViT backbone, frozen at mid-levels, supported calibrated predictors with consistent ordering of methods; absolute numbers varied slightly, but conclusions did not.

There are trade-offs. Predicting $\hat{\sigma}$ invites a sharpness–coverage balance: intervals that are too narrow under-cover; too wide reduce utility [19, 20]. The heteroskedastic head plus normalized conformal achieved a good balance, but extremely uniform images sometimes induced conservative intervals dominated by annotation noise, while highly stochastic textures could elicit high predicted complexity with lower perceived effort in certain categories. Reporting results within categories and

using disagreement-aware metrics mitigated these effects by not over-penalizing inherently ambiguous items and by surfacing where semantics modulate perceived complexity.

Assumptions and threats to validity deserve scrutiny. The Thurstone model posits Gaussian noise in the latent comparative judgment; while its logistic cousin (Bradley–Terry) would yield similar orderings after rescaling, either link can misfit tail behavior. Exchangeability for conformal calibration may be mildly violated across semantic categories; reporting per-category coverage partly addresses this, and the normalized residuals help absorb heteroskedasticity, but strong domain shifts will still challenge guarantees [21]. Rater demographics and device factors (mobile vs. desktop) can bias judgments; logging metadata and stratified analyses are recommended for high-stakes settings. Finally, we modeled item-wise ambiguity but not annotator identities; incorporating rater embeddings or mixed-effects terms could separate systematic annotator bias from genuine item difficulty.

For practitioners, a usable pipeline emerges: (1) collect a small number of forced-choice pairs per image under clear instructions with QC; (2) aggregate with a heteroskedastic pairwise model to obtain $(\tilde{\mu}, \tilde{\sigma})$; (3) train a mid-level feature head to predict $(\hat{\mu}, \hat{\sigma})$ with a heteroskedastic loss; (4) reserve a calibration set and fit normalized split-conformal quantiles for desired coverage levels; (5) evaluate with correlation, coverage, interval width, reliability diagrams, and disagreement-aware ranking; (6) if needed, iterate with ambiguity-focused active sampling. This recipe balances fidelity, reliability, and cost, and it is simple enough to adopt in visualization design, content selection, and resource allocation systems.

Limitations suggest concrete extensions. Complexity here is static and image-level; extending to spatiotemporal complexity in video would capture motion and temporal crowding. Human disagreement can be multimodal rather than merely high-variance; mixture models or latent rater clusters could reveal distinct perceptual strategies [22]. Transformer token statistics and attention dispersion might offer richer uncertainty signals when coupled with our calibration scheme. Cross-cultural and task-context studies would measure how transferable $(\hat{\mu}, \hat{\sigma})$ is across populations and instructions, turning uncertainty into a lens on perceptual diversity rather than only noise. Finally, causal investigations—intervening on object count, clutter, or texture while holding semantics constant—could test whether and how changes in complexity move memorability or task performance, turning observed correlations into evidence for mechanism.

Reproducibility and openness are central. We have emphasized fixed seeds, released splits (including ambiguity subsets), standardized backbones and layers, consistent hyperparameter tuning, and comprehensive reporting with confidence intervals. The code paths for aggregation, training, calibration, and evaluation are short and modular by design [23–25]. Such transparency is not just housekeeping: it ensures that future work can meaningfully compare, ablate, and build on uncertainty-aware complexity estimation.

In sum, the study shows that calibrated uncertainty is not an optional add-on but a defining feature of reliable perceptual predictors. By integrating heteroskedastic pairwise aggregation with normalized conformal calibration on top of mid-level visual features, we obtained models that are accurate, honest about their uncertainty, label-efficient, and practically useful. The broader implication is that many subjective, perceptual constructs in vision—beyond complexity—could benefit from the same shift toward reliability-aware modeling, enabling safer, more informative, and ultimately more human-centered systems.

References

[1] Yi, J. S., ah Kang, Y., Stasko, J., & Jacko, J. A. (2007). Toward a deeper understanding of the role of interaction in information visualization. *IEEE Transactions on Visualization and Computer Graphics, 13*(6), 1224-1231.

[2] Cortez, A. (2011). *Winning at Risk: Strategies to go Beyond Basel* (Vol. 638). John Wiley & Sons.

[3] Turner, H., & Firth, D. (2012). Bradley-Terry models in R: the BradleyTerry2 package. *Journal of Statistical Software, 48*, 1-21.

[4] Thurstone, L. L. (1931). Multiple factor analysis. *Psychological Review, 38*(5), 406.

[5] Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems*, 30.

[6] Kull, M., Perello Nieto, M., Kängsepp, M., Silva Filho, T., Song, H., & Flach, P. (2019). Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. *Advances in Neural Information Processing Systems*, 32.

[7] Kuleshov, V., Fenner, N., & Ermon, S. (2018, July). Accurate uncertainties for deep learning using calibrated regression. In *International Conference on Machine Learning* (pp. 2796-2804). PMLR.

[8] Lei, J., & Wasserman, L. (2014). Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology, 76*(1), 71-96.

[9] Aptoula, E. (2008). par morphologie mathématique. *Application à la description, l'annotation et la recherche d'images* (Doctoral dissertation, Université de Poitiers).

[10] Vondrick, C., Patterson, D., & Ramanan, D. (2013). Efficiently scaling up crowdsourced video annotation: A set of best practices for high quality, economical video labeling. *International Journal of Computer Vision, 101*(1), 184-204.

[11] Jannach, D., Lerche, L., & Zanker, M. (2018). Recommending based on implicit feedback. In *Social information access: systems and technologies* (pp. 510-569). Cham: Springer International Publishing.

[12] Záková, M., & Zelezný, F. (2007). Using Taxonomic Background Knowledge in Propositionalization and Rule Learning. *Prior Conceptual Knowledge in Machine Learning and Data Mining*, 109.

[13] Flage, R., Aven, T., Zio, E., & Baraldi, P. (2014). Concerns, challenges, and directions of development for the issue of representing uncertainty in risk assessment. *Risk Analysis, 34*(7), 1196-1207.

[14] Fang, W., Miller, S. M., & Yeh, C. C. (2010). Does a threshold inflation rate exist? Quantile inferences for inflation and its variability. *Empirical Economics, 39*(3), 619-641.

[15] Hamilton, J. D., Waggoner, D. F., & Zha, T. (2007). Normalization in econometrics. *Econometric Reviews, 26*(2-4), 221-252.

[16] Rouse, E. W. (2014). *Test Drivers of Candidate Pass Rates: A Predictive Model for Professional Certification Exams.* Northcentral University.

[17] Marszalek, J. (2006). *Computerized Adaptive Testing and the Experience of Flow in Examinees.* University of Illinois at Urbana-Champaign.

[18] Tang, X., Li, K., Li, R., & Veeravalli, B. (2010). Reliability-aware scheduling strategy for heterogeneous distributed computing systems. *Journal of Parallel and Distributed Computing, 70*(9), 941-952.

[19] Perkins, N. J., & Schisterman, E. F. (2005). The Youden Index and the optimal cut-point corrected for measurement error. *Biometrical Journal: Journal of Mathematical Methods in Biosciences, 47*(4), 428-441.

[20] Zhang, H., Zimmerman, J., Nettleton, D., & Nordman, D. J. (2020). *Random Forest Prediction Intervals*. The American Statistician.

[21] Cao, K., Chen, Y., Lu, J., Arechiga, N., Gaidon, A., & Ma, T. (2020). Heteroskedastic and imbalanced deep learning with adaptive regularization. arXiv preprint arXiv:2006.15766.

[22] Steinley, D., & Brusco, M. J. (2011). Evaluating mixture modeling for clustering: recommendations and cautions. *Psychological Methods, 16*(1), 63.

[23] Guo, H., Huang, J., & Laidlaw, D. H. (2015). Representing uncertainty in graph edges: An evaluation of paired visual variables. *IEEE Transactions on Visualization and Computer Graphics, 21*(10), 1173-1186.

[24] Zhang, H., Morvan, C., & Maloney, L. T. (2010). Gambling in the visual periphery: a conjoint-measurement analysis of human ability to judge visual uncertainty. *PLoS Computational Biology, 6*(12), e1001023.

[25] Saraee, E., Jalal, M., & Betke, M. (2020). Visual complexity analysis using deep intermediate-layer features. *Computer Vision and Image Understanding, 195*, 102949.

How to cite this article: Margrit Betke (2021). Uncertainty Calibrated Visual Complexity from Pairwise Judgments. *Bulletin of Computer and Data Sciences*, 2(1), 28-43. DOI: 10.71448/bcds2121-4

Received: 17/1/2021 **Revised:** 22/6/2021 **Accepted:** 12/9/2021 **Publish:** 30/12/2021