

---

# A Causal Framework for Aligning Image Quality Metrics and Deep Neural Network Robustness

---

**Nathan Drenkow<sup>1,2,\*</sup>**

<sup>1</sup>The Johns Hopkins University Applied Physics Laboratory  
Laurel, MD USA

\* e-mail: ndrenko1@jhu.edu

**Mathias Unberath<sup>2</sup>**

<sup>2</sup>The Johns Hopkins University  
Baltimore, MD USA

## Abstract

Image quality plays an important role in the performance of deep neural networks (DNNs) that have been widely shown to exhibit sensitivity to changes in imaging conditions. Conventional image quality assessment (IQA) seeks to measure and align quality relative to human perceptual judgments, but we often need a metric that is not only sensitive to imaging conditions but also well-aligned with DNN sensitivities. We first ask whether conventional IQA metrics are also informative of DNN performance. We show theoretically and empirically that conventional IQA metrics are weak predictors of DNN performance for image classification. Using our causal framework, we then develop metrics that exhibit strong correlation with DNN performance, thus enabling us to effectively estimate the quality distribution of large image datasets relative to targeted vision tasks.

## 1 Introduction

Ensuring the robustness of deep neural networks (DNNs) to real-world imaging conditions is crucial for safety- and cost-critical applications. Extensive research has shown that DNNs remain sensitive to natural distortions [1–4] despite efforts to close the gap between performance on clean and naturally-distorted images. While much effort has focused primarily on the design and optimization of robust DNNs, there is now growing interest in developing a deeper understanding of how the properties of the image data itself influence robustness during training and evaluation [5–8].

Since image quality is known to influence DNN behavior, a first step in analyzing image data is to examine the relationship between image quality (IQ) and DNN robustness. Here *quality* describes the absence of distortion but more generally relates to the ability to extract task-relevant information from the image. Image *quality* and *difficulty* are closely related where quality measures properties of the imaging conditions while difficulty involves content and composition in addition to the conditions. While image content and quality remain intimately related, prior work in evaluating DNN robustness [9–13] has shown that changes in image quality can result in diverse DNN behavior even when content is held fixed. For this reason, examining the relationship between IQ metrics and DNN behavior remains an important question that is theoretically and empirically unanswered prior to this work.

In an ideal case, image quality metrics will strongly correlate with task DNN performance while remaining relatively independent of knowledge or assumptions about specific downstream task models that will consume the data. Despite decades of research into IQ metrics [14–17], we examine for the first time in-depth the explicit connection between IQ metrics and task DNN performance. Due to the lack of prior theoretical work in this area, we first analyze this link empirically. Then, *the primary goal of this work is to provide a mathematical framework for identifying and analyzing the conditions under which IQ metrics and DNN performance are correlated.*

In providing a mathematical means to establish a link between IQ and DNN performance, our method enables new pathways for quantitatively analyzing dataset composition. Using IQ metrics as a proxy for DNN performance, we can estimate the quality distribution of datasets to understand how the range of “easy” to “hard” images influences observed DNN robustness. This becomes increasingly useful as image datasets continue to grow in size to the extent that the cost and feasibility of using human annotators to assess and annotate properties of each data point is becoming intractable. With pre-training datasets for foundation models and other large-scale vision models approaching hundreds of millions to billions of images [18–20], new automated methods for analysis are needed for quantitatively assessing dataset composition.

In this work, we focus specifically on *natural robustness* which considers how images are distorted due to real-world factors such as lighting, weather, sensor settings, and/or motion. Image quality assessment (IQA) metrics have been developed over several decades of research [14–17, 21–25] and provide quantitative measures of quality calibrated with respect to human perceptual judgments. To the best of our knowledge, little work has been done to understand how these IQA metrics can help relate image difficulty and DNN performance. To make this connection explicit, we state our primary research question: **What is the extent of the relationship between IQ and DNN performance metrics?**

Our primary motivation is to identify image quality metrics that allow us to assess the distribution of image quality in large-scale datasets and establish quality-driven priors for DNN performance independent of any specific trained task models. We propose the following desiderata for IQ metrics towards achieving these objectives.

- **D1 - Sensitive:** IQ metrics should be sufficiently sensitive to changes in image conditions
- **D2 - Blind:** IQ metrics should work in No Reference IQA (NR-IQA) settings where images are assessed *without* knowledge of a reference image captured under “clean” conditions
- **D3 - Predictive:** IQ metrics should be correlated with DNN task performance
- **D4 - Task Model Agnostic:** IQ metrics should be designed/trained/calibrated without *a priori* knowledge of the downstream DNN models/architectures to be trained or evaluated on the data under consideration

The first criterion (**D1**) is a baseline condition requiring that the metric is actually sensitive to the natural conditions likely in the imaging domain. **D2** operates under the assumption that real-world datasets will not consist of pairs of clean/distorted images and will instead contain images collected in diverse conditions. **D3** stems from the idea that quality metrics should measure general properties of the data that influence performance metrics (e.g., if the quality metric decreases, then the performance metric should also decrease, although not necessarily at the same rate). Lastly, **D4** comes from the desire to use image quality to assess the composition of the dataset *independent* of any task-specific model training and without making assumptions about the type of DNN to be trained downstream. In other words, we want to avoid IQ metrics that are biased towards specific task models and/or require pre-training on each dataset to be analyzed.

To determine the extent to which IQA metrics satisfy the above desiderata, our work makes the following contributions:

- Our primary contribution is a causal framework for analyzing the relationship between image quality and DNN performance in a range of IQA settings
- We use the framework to establish theoretically and empirically the independence of image quality and DNN performance under general conditions
- We identify specific conditions under which IQA metrics can be predictive of DNN performance
- We conduct a first-of-its-kind evaluation of the relation between conventional IQ metrics and DNN performance and find that conventional metrics are weakly predictive of performance
- We use the framework in the context of image classification tasks to develop a new task-guided IQA metric that enables quantitative assessments of image quality that are strongly predictive of downstream DNN task performance
- We show that our framework and novel metric provide the means to expose subtle differences in dataset quality distributions that directly impact on DNN performance

## 2 Results

### 2.1 Experiment setup

Given the causal interpretation of IQA and the IQ metric desiderata (**D1-D4**), we examine how conventional NR-IQA metrics relate to DNN performance. Our primary hypothesis is that if quality ( $Q$ ) and task model performance ( $M$ ) metrics are sensitive to a common set of visual features  $Z$  derived from images  $X$  (Fig. 1), then we should observe that  $Q$  is correlated with  $M$  and even predictive of  $M$  given  $X$ . Plainly stated, if image quality is high in general, then DNN performance should be similarly high (and vice versa).

We focus the following experiments on image classification tasks since they have available benchmark datasets and have been well-studied within the deep learning field. We show how our framework can be used to identify the relationship between IQA metrics and DNN performance as well as how it can guide the development of new metrics that satisfy all desiderata. For image classification, our experiments show that common NR-IQA methods are very weakly predictive of DNN performance, and while they satisfy **D1**, **D2**, **D4** of our IQ desiderata, they fail to satisfy **D3** and may not be suitable for estimating priors on DNN performance.

In the experiments in this section we use the following basic setup. In order to have precise control and knowledge of the type and severity of image distortion, we use the ImageNet validation (IN-val) and ImageNet-C (IN-C) [9] datasets for evaluating IQ/performance on clean and corrupted images respectively. For reference, we provide a common corruptions causal Directed Acyclic Graph (DAG) in Appendix C for comparison with the ones in Figures 1 and 2.

For each experiment, we compute the IQ metric ( $Q$ ) and DNN correctness ( $M$ ) for each image of the IN-C evaluation dataset. We use the following common and high-performing NR-IQA metrics ( $Q$ ): CLIP-IQA [21], ARNIQA [17], BRISQUE [26], and Total Variation (TV). Here, CLIP-IQA and ARNIQA represent the state-of-the-art in learning-based IQA metrics while BRISQUE and TV represent conventional non-deep learning baselines. For DNNs, we evaluate the correctness ( $M$ ) using pretrained ResNet34 [27], ConvNext-B [28], EfficientNet-V2-M [29], MobileNet-V3-L [30], Vision Transformers [31], and Swin-B [32] models provided via the `torchvision` package [33]. This set of models covers a wide range of architectural design, scale, efficiency, and performance characteristics in order to capture potential variability in the relationship between DNN performance and IQ metrics. Across all experiments, 95% confidence intervals (CI) are obtained via bootstrapping with 1000 resamples.

### 2.2 Correlation and Predictability of $Q$ , $M$ (**D3**)

We start by examining the correlation between  $Q$ ,  $M$  for NR-IQA metrics. Figure 3 shows the general relationship between  $Q$ ,  $M$  where each point in the figure is the average accuracy (over 50k images) for each corruption and severity in IN-C. Similarly, Table 1 computes the Kendall Rank Correlation Coefficient (KRCC), Spearman Rank Correlation Coefficient (SRCC), and Pearson Linear Correlation Coefficient (PLCC) between IQ and average accuracy across all corruption/severity pairs (75 total).

These results provide a look at the group-wise association between  $Q$ ,  $M$  where the groups capture general trends in performance/IQ based on corruption type and severity. The low correlation between  $Q$ ,  $M$  suggests that these NR-IQA metrics likely fall under the model described by Figure 2 where  $Q$ ,  $M$  are conditionally independent given  $X$ .

We also examine the point-wise relationship between  $Q$ ,  $M$ . We aggregate DNN predictions and IQ values for all images in IN-C across all corruptions/severities and then randomly split the dataset (by image ID) into 80% training and 20% testing. We train a logistic regression classifier to predict  $P(M|Q)$  and test on the hold-out set. We measure the predictability of  $M$  using Area Under the Curve (AUC) and average cross-entropy (CE).

Table 1 shows that at the per-image level,  $Q$  is still weakly predictive of  $M$  (i.e.,  $AUC \approx 0.5$ ). This result is consistent with the theoretical analysis in Sec. 4.2 and the weak correlation observed empirically between  $Q$ ,  $M$  measured at the group level.

While the causal DAG in Figure 2 would suggest that conditioning on the class label  $Y$  should not change the result, we test this empirically as follows. We re-run the logistic regression for each label value in  $\mathcal{Y}$  separately (1000 total) and compute the mean AUC ( $mAUC$ ) and CE ( $mCE$ )

across all labels. While we observe some variability in results when fixing  $Y$ , we find  $mAUC = 0.5652$  ( $\sigma = 0.08$ ) and  $mCE = 0.6176$  ( $\sigma = 0.1094$ ) suggesting that even when we control for  $Y$ , the predictability of DNN performance from the NR-IQA metrics remains weak. More detailed results from experiments controlling for image content can be found in Appendix G.

These results suggest that NR-IQA metrics are likely sensitive to a different set of image features than task DNNs (i.e., no shared  $Z$ ) and thus are barely, if at all, predictive of performance (i.e., they do not satisfy criterion **D3**). In particular, Table 1 shows that the AUC for predicting DNN performance given conventional NR-IQA metrics is close to chance (i.e., 0.5). We also observe that this result holds across a range of NR-IQA metrics including those that are minimally tunable (TV, BRISQUE) to optimized through large-scale training (CLIP-IQA, ARNIQA), but all of which are calibrated against human judgments. This provides strong evidence that this form of calibration is a likely cause of the lack of correlation with DNN performance. *The primary implication of this result is that if we intend to use IQ metrics to measure image quality/difficulty from the DNN perspective, common NR-IQA metrics may not be well-suited to this task and alternative approaches are needed.*

Lastly, we acknowledge that estimating image quality independent of the image content remains a challenging task. In our experiment, we found that by holding image- and class-level content fixed (Appendix G), we find that NR-IQA metrics are still weakly predictive of task DNN performance. While this does not completely disentangle the quality/content relationship, it provides clear evidence that the relationship between conventional NR-IQA metrics is weak.

### 2.3 Restoring the association between $Q, M$ via strong task-guidance (D3)

The previous results indicated that existing NR-IQA meet desiderata **D1**, **D2**, **D4**, but the lack of predictability (**D3**) between NR-IQA metrics and DNN accuracy/correctness is a major limitation in using these metrics for assessing dataset quality relative to potential downstream task models. Focusing specifically on **D3**, we next consider an alternative formulation of the causal model that will allow us to recover a dependence between  $M, Q$  when conditioning on  $X$ .

In the case where a pre-trained DNN  $f_\theta$  is given, Figure 4 describes a scenario where the predictions from this DNN may also be used as indicators of quality. This parallels other work [34] which shows that uncertainty in the output predictions is often a good predictor of the OOD nature of the input. Note here that while  $Q, M$  both depend on  $\hat{Y}$ ,  $Q$  requires no knowledge of the labels. In this case, it is possible that  $\hat{Y}$  can be incorrect from the perspective of the ground truth label  $Y$  but still provide information about  $Q$  (e.g., via a low confidence prediction).

Because this approach uses a model for  $Q$  that is already trained for the classification task, we consider this **strong** task-guided IQA (TG-IQA). Clearly, this provides an alternative to the conventional NR-IQA metrics but now violates **D4** since  $Q$  is informed directly by the same model trained for the task and measured by  $M$ . Nonetheless, our (temporary) goal here is to use the causal framework to show there exists a case where  $Q, M$  are associated through a common set of features  $Z$ . Our hypothesis is that with **strong** TG-IQA we should observe a clear correlation between  $M, Q$ .

We examine the case where  $Q$  is determined directly from predictions generated by a pre-trained task DNN. In this case, let  $f_\theta$  be pre-trained to predict  $P(Y|X)$ . Then, let  $z \in \mathbb{R}^k$  be the pre-softmax logits obtained from  $f_\theta$  and  $\hat{y} = \text{softmax}(z)$  where each  $\hat{y}_i = P(Y = i|X)$  for  $i \in 1, \dots, K$ . We consider three possible variants of  $Q$  in this setting: (1) Max probability:  $Q_p := \max_i \hat{y}_i$ , (2) Entropy:  $Q_h := H(\hat{y}) = -\sum_i \hat{y}_i \log \hat{y}_i$ , and (3) Max logit:  $Q_l := \max_i z_i$ . While all three cases are inherently tied to the underlying label set  $\mathcal{Y}$ , the values of  $Q$  do **not** have access to the ground truth label  $Y$ . Each of these  $Q$  implicitly capture a DNN’s confidence about its prediction and the natural underlying hypothesis is that confidence and image quality are positively correlated (i.e., as quality decreases, confidence also tends to decrease). These choices for  $Q$  are driven by their use in out-of-distribution [34–36] and distribution shift detection [37].

Using the same setup as in Section 2.1, we now replace the NR-IQA metrics with  $Q_p, Q_h, Q_l$ . As in Section 2.2, we examine the group-wise correlation and point-wise predictability of  $M$  from  $Q$ . To ensure our test of predictability is fair, we use separate models for obtaining  $M$  and  $Q$  (e.g., ConvNext-B and Swin-B respectively). We provide additional results for other model pairs in Appendix E. Figure 5 shows the group-wise relationship between  $Q, M$  where groups are averages over all images for the corresponding corruption, severity.

The results in Figure 5 and Table 2 show that strong task-guidance for  $Q$  results in high correlation between  $Q$ ,  $M$  and predictability of  $M$  from  $Q$  (**D3**). *This result is important to show that by using the causal framework it is possible to find a metric  $Q$  that relies on a similar set of features as separate task models and is predictive of  $M$ .*

The high correlation between  $M$ ,  $Q$  exists across all task DNN architectures and variants of the metric which is consistent with the scenario shown in Fig. 4. In particular, given the DAG in Fig. 4, we see that without placing strong requirements on the specific task model architecture/design (e.g., convnet vs. transformer), if the task DNN producing predictions  $\hat{Y}$  is sensitive to the imaging conditions,  $Q$  will be strongly correlated with  $M$ .

However, like the previous section, this approach is only a partial solution since it satisfies **D1**, **D2**, **D3** but clearly violates **D4** by requiring a model already trained for the classification task. In this case, modeling/optimization choices for the model used for  $Q$  may result in unintended biases (e.g., performance disparities across (image class, distortion type) pairs) that might skew estimates of  $Q$ .

#### 2.4 Restoring the association between $Q$ , $M$ via weak task-guidance (**D3**, **D4**)

So far, Sec. 2.2 showed that common NR-IQA metrics are weakly predictive of DNN performance and are therefore not viable candidates for supporting image/dataset-level analysis given our desiderata. Then, we were able to address the predictability issue (**D3**) in Sec. 2.3 using strong task-guidance, but at the cost of requiring a task model already trained for the classification task (a violation of **D4**). To satisfy all desiderata, we use the causal framework to design a task-guided metric (ZSCLIP-IQA) that uses the weaker zero-shot formulation of the image classification task to align  $Q$  and  $M$  (Sec. 4.4).

Using the setup from Sec. 2.2 we now replace the NR-IQA metrics with  $Q_p$ ,  $Q_h$ ,  $Q_l$  based on the weak task-guided ZSCLIP-IQA method described in Section 4.4. We again examine the group-wise correlation and point-wise predictability of  $M$  from  $Q$ . Figure 6 and Table 3 show that weak task-guidance is enough to restore the association between  $Q$  and  $M$  without requiring a new task model to be trained on the dataset of interest.

We also note that while the CLIP backbone for ZSCLIP-IQA is pre-trained on a self-supervised task that resembles classification, it was not exposed to ImageNet (or IN-C) data during its training (see Sec. 5 in [18]) and can be effectively used here in a zero-shot setting to satisfy **D4**. In fact, while methods like CLIP-IQA and ARNIQA also rely on pre-trained backbones, the results of Tables 1 and 3 show that only ZSCLIP-IQA is “guided” (via our causal framework) to be a stronger predictor of DNN performance compared to other methods calibrated to human perceptual judgment.

The results of this experiment confirm that using the causal framework we can design an IQA metric that satisfies all of the desiderata. We note that while the group-wise correlation between  $M$ ,  $Q$  is high, the predictability (measured via AUC) is not as significant given that the DNN used as the backbone for the IQ metric is no longer the same as that used for the task. Nonetheless, the weak task-guided metric still achieves above chance AUC which is a significant improvement over conventional NR-IQA metrics (see Table 1). Future research may be able to further use the weak task-guided formulation and causal framework to improve this result. *The primary implication of this result is that the causal framework enabled us to develop an IQ metric that satisfies all of our desiderata and thus allows us to effectively predict task DNN performance on an image dataset by analyzing its quality distribution. We perform one such analysis in Section 2.5.*

#### 2.5 Predictability of DNN performance for mildly corrupted datasets

In the previous experiments, the use of IN-C allowed us to investigate the large-scale effect of image corruptions on the predictability of performance by using multiple corrupted versions of the validation set with multiple levels of severity. In real-world datasets, we expect that only a small fraction of images will be corrupted. We next examine the extent to which IQA metrics can be used to show differences between the quality distributions of datasets containing varying levels of corruption while still satisfying **D1-D4** in these more realistic settings.

To answer this question, we generate new variants of IN-val consisting of mixtures of clean and corrupted images. For each variant, we specify a set of valid corruptions  $\mathcal{C}$ , severities  $\mathcal{S}$ , and a corruption probability  $p_c$ . We choose a fraction  $1 - p_c$  of the original IN-val image IDs to remain as clean images and a fraction  $p_c$  to be corrupted. The corrupted images are sampled uniformly amongst

the corruptions  $c \in \mathcal{C}$  and severities  $s \in \mathcal{S}$ . The resulting variant consists of the original 50k image IDs with a mixture of clean and corrupted images. We choose  $\mathcal{C}$  to consist of all 15 corruptions in the IN-C dataset and limit severity to  $\mathcal{S} = \{1, 2, 3\}$  in order to further test the sensitivity of the IQA metrics (**D1**). We create variants of the IN-val dataset for  $p_c = N/100$  for  $N \in [1, \dots, 20]$ . We evaluate the DNNs on these dataset variants and estimate predictability using logistic regression as in previous experiments.

We compute  $mAUC$  over all  $p_c$  variants and find ZSCLIP-IQA ( $Q_l$ ) outperforms all other NR-IQA metrics with  $mAUC = 0.64$  with the next best (CLIP-IQA) achieving only  $mAUC = 0.57$ . The full results are found in Appendix H and show that predictability with ZSCLIP-IQA is stable with respect to changes in the proportion of clean/corrupted images in the dataset whereas more traditional NR-IQA metrics remain near random chance  $AUC$  and exhibit higher variance as  $p_c$  changes.

This experiment examines how IQ metrics can be used to measure the variability of image quality within and across image *datasets* and to provide insight into the overall difficulty of the dataset for downstream task DNNs. The results show that while all metrics in this experiment can distinguish between differences in the quality distributions of the dataset variants, only ZSCLIP-IQA achieves high predictability over all variants. Conventional IQA metrics improve only as the number of distorted images in the dataset increases (where it becomes easier to separate clean and corrupted images).

The broader impact of these results is that we can use IQ metrics that satisfy the desiderata as a means to audit datasets to provide hypotheses about how their composition relates to task DNN performance. For instance, if an image quality distribution is estimated using a weak task-guided metric and found to skew towards higher-quality images, it may be difficult to draw strong conclusions about task DNN robustness evaluated on that data. In contrast, if the quality distribution skews towards lower-quality images, then there is stronger evidence that a task DNN that performs well on the data is a robust model. *The primary implication of this experiment is that we show how the weak task-guided metric, developed using our causal framework, provides a means to effectively estimate and compare dataset-level image quality distributions in this way.*

### 3 Discussion

In this work, we were motivated to identify measures of image quality that allow us to produce IQ-driven priors on DNN performance. We presented a causal inference framework for this problem and proposed a set of IQ metric desiderata to guide our analysis (Sec. 4). Using our causal framework, we show conditions where image quality measures can be predictive of DNN performance. We then provide a first-of-its-kind detailed examination of the relationship between conventional NR-IQA metrics and DNN performance. We use our causal framework and extensive empirical evaluations in the context of image classification to demonstrate that common NR-IQA do not satisfy our desired IQ criteria (Sec. 2.2). We then use the causal approach to develop introduce the notion of task-guided IQA metrics (Sec. 2.3). In particular, the ZSCLIP-IQA (Sec. 4.4) metric provides a causality-driven proof-of-concept that satisfies all IQ desiderata and paves the way for future research to improve the alignment between IQA metrics and DNN performance.

One broad implication from this work is that our framework exposed the existence of cases where IQ metrics and task DNNs are not sensitive to the same features of image quality. In the context of using IQ metrics as priors for DNN performance (e.g., to estimate image understanding difficulty), we demonstrated for the first time that conventional NR-IQA metrics are generally not appropriate for this task. However, while our task-guided IQA metrics were shown to be suitable in contexts where task DNNs are the downstream consumers of the image data, conventional NR-IQA metrics (e.g., BRISQUE, TV, CLIP-IQA, ARNIQIA) may still be more appropriate for applications where humans are the target consumers.

Similarly, while the task-guided metrics were found to be predictive of task DNN performance, there was not much disparity in the degree of predictability observed across different classes of task DNN architectures (e.g., convolutional nets, transformers). As such, the task-guided metrics could be used to aid in making choices between task model architectures in downstream applications by first predicting baseline performance on targeted image datasets using the TG-IQA metric and then comparing actual task DNN performance against those predictions to determine which architecture exhibits the greatest improvement over the baseline.

Lastly, the results of Section 2.5 demonstrate how our work lays the foundation for future research to examine how to analyze dataset-level properties of images. Such analyses can be used to determine how properties of datasets (e.g., image quality distribution) impact downstream consumers of the data. For instance, by first understanding the quality distribution of robustness benchmark datasets, downstream evaluation results obtained using these benchmarks can be better contextualized. Such findings will allow researchers to better de-couple properties of datasets from properties of their models, thus enabling them to make more targeted improvements to either. This capability is of growing interest as imaging datasets continue to grow in size beyond the limits of qualitative assessment by humans.

As a tool for analysis, the proposed causal framework poses minimal broader risks. While causal models require assumptions about the data generating process, these assumptions are made explicitly in the causal graph and improve the overall transparency of the analysis. Of greater concern is the possibility that using quality metrics to prune/resample datasets may lead to unintended consequences such as removing poor-quality images in a way that disproportionately affects protected groups. While our work does not address the question of dataset pruning/resampling, we mention this to help ensure that future researchers consider these possibilities in their own work.

We also recognize that image quality alone is insufficient to predict task model performance as both image content and composition play a role in task difficulty and the relationship between IQA and performance may be confounded by other factors. While the experiment in Appendix G examined two approaches to controlling for content, future work should more deeply address this question. Nonetheless, we show that our causal framework still allows us to analyze the conditions where quality properties of our dataset may be correlated with DNN performance. Our ZSCLIP-IQA method provides one solution that satisfies the proposed IQ desiderata but we believe there are many opportunities for improving on this approach in future research.

We also acknowledge that our experiments only addressed image classification tasks. We focused initially on classification since it is well-studied and clearly defined, with many public benchmarks available for evaluation. Even in this context, we are the first to show that the notion of quality is task-dependent (i.e., perceptual judgment vs. classification). Our primary contribution in this work is the causal framework for IQA and we believe this provides a strong foundation for supporting future research that examines similar questions for a wider range of vision tasks.

## 4 Methods

Image quality assessment has been long-studied in the computer vision and image processing literature. Full Reference IQA (FR-IQA) [14, 15, 38] assume the availability of a reference or “clean” image against which the test image is compared and the quality is measured. In contrast, the NR-IQA [14–17, 21–24, 26] setting (aka Blind IQA) uses only features of the test image to estimate a quality score. In both settings, conventional IQA methods are calibrated and compared against human perceptual judgments of quality such as Mean Opinion Scores (MOS). These measures are task-agnostic and humans are not required to make judgments about the content of the image but only to measure subjective “quality” (typically on a scale of 1-5, Poor-High). This motivates our investigation into whether these metrics can also provide task-relevant assessments of image quality.

In addition, separate from IQA research, considerable attention has been paid to understanding and characterizing the impact of image quality on DNN robustness [39]. In conventional settings, robustness is evaluated by exposing DNNs to images degraded by precisely controlled synthetic corruptions that target specific isolated imaging conditions [1, 2, 9–11, 40–43]. Later evaluations of robustness have targeted imaging conditions that capture more challenging, naturally-occurring degradations [3, 12, 13, 44]. In all of these works, quantitative measures of IQ were not explicitly correlated with DNN robustness. Our work is the first to our knowledge to systematically examine this relationship and paves the way for correlating image quality and DNN robustness when control over image degradation is not possible.

A key question of this work centers on whether IQA metrics calibrated against human MOS are sensitive to any of the same image features that DNNs use for downstream tasks. Outside of the IQA literature, prior works have shown differences in humans and DNNs in the context of shape/texture bias [45, 46], shortcut learning [47–50], and error consistency [4, 51], but the question remains open whether IQA metrics aligned with human MOS correlate with DNN performance.

Our work is strongly motivated by the growing interest in automated methods for dataset analysis. In particular, new methods focus on dataset pruning [52–54], identifying difficult/important examples [5, 55] or data slices [56–59], and dataset auditing for shortcuts [7, 8]. Other results have shown that understanding dataset composition matters for analyzing model robustness [3, 12]. Our work takes a positive step towards automated methods for analyzing the quality distribution of image datasets and establishing priors on DNN performance.

#### 4.1 Causal Framework for IQA

To understand the relationship between IQA metrics and DNN performance, we provide here a causal inference perspective on the IQA problem. We use causal directed acyclic graphs (DAGs) to illustrate our assumptions about the imaging generating process, quality metric, and performance metric as well as the interactions between all associated variables. This causal framework provides a means for identifying the specific conditions under which quality metrics are predictive of DNN performance.

We specify a causal DAG  $\mathcal{G}$  via a set of nodes  $\mathcal{V}$  and directed edges  $\mathcal{E}$ . To obtain the causal interpretation, directed edges imply a causal relationship such that for a variable/node  $V \in \mathcal{V}$ ,  $V$  is a function of its parents ( $V = f_V(pa(V), U)$  where  $U$  is an exogenous noise term).

For defining causal models in the IQA context, we start from a set of factors  $A \in \mathcal{A}$  that capture the key variables in the data generating process affecting the image conditions (e.g., lighting, focal length, aperture, exposure, weather). Let  $X \in \mathcal{X}$  be the resulting images, and for a task  $T$ , let  $Y \in \mathcal{Y}$  be the label associated with  $X$  for the task. For this work, we focus on classification tasks where  $\mathcal{Y}$  consists of a discrete set of  $K$  classes ( $\mathcal{Y} = \{1, \dots, K\}$ ). Our quality metric  $Q : \mathcal{X} \rightarrow \mathbb{R}$  maps images to real number scores (typically in  $[0, 1]$  where 1 is the highest quality). We also assume a downstream task DNN  $f_\theta : \mathcal{X} \rightarrow \mathcal{Y}$  that maps images to class probabilities and is parameterized by  $\theta$ . We write the predicted probabilities  $\hat{Y} = f_\theta(X)$  where  $\hat{Y} \in \mathbb{R}^K$ . Given one-hot encoded labels  $Y$  and predictions  $\hat{Y}$ , we can compute a performance metric (e.g., accuracy)  $M : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$ . In the general case and without loss of generality, we assume that  $\hat{Y}$  is the prediction from a deterministic DNN. Similarly, we also assume that  $Q, M$  are both deterministic functions of their parents in the causal DAG.

#### 4.2 Baseline IQA formulation

We start with the baseline formulation of the IQA problem as shown in Figure 2. We assume the labels  $Y$  are determined from interpreting  $X$  and that an oracle labeling function exists such that  $Y$  can always be determined from  $X$ .

This model is a general formulation and makes no assumptions about the nature of the functions that compute  $Q, \hat{Y}$ . This causal model is also consistent with conventional NR-IQA settings [26, 60] where the determination of  $Q$  given  $X$  is based on a function calibrated to human perceptual judgment and without knowledge of the task or labels.

**Conditional independence of  $Q, M$ :** The causal graph of Figure 2 illustrates that under the baseline formulation,  $Q \perp M|X$  and  $X$  is said to *d-separate*  $Q, M$ . The interpretation here is that given any image  $X$ , there is no expected relationship between  $Q$  and  $M$  by construction. This is not to say that a relationship cannot exist, but simply that there is nothing in this model that ensures it directly.

In addition to observing the d-separation of  $Q, M$  by  $X$ , we can also compute the average causal effect (ACE) of  $Q \rightarrow M$ . We use the potential outcomes notation  $M(Q = q)$  (or  $M(q)$ ) to indicate the value of  $M$  if  $Q$  had been set to the value of  $q$ .

$$\begin{aligned}
 ACE(Q \rightarrow M) &= \mathbb{E}[M(Q = q) - M(Q = q')] = \mathbb{E}[M(q)] - \mathbb{E}[M(q')] && \text{linearity of expectation} \\
 &= \mathbb{E}_X[\mathbb{E}[M(q)|X] - \mathbb{E}[M(q')|X]] && \text{law of total expectation} \\
 &= \mathbb{E}_X[\mathbb{E}[M(q)|Q = q, X] - \mathbb{E}[M(q')|Q = q', X]] && \text{unconfoundedness} \\
 &= \mathbb{E}_X[\mathbb{E}[M|Q = q, X] - \mathbb{E}[M|Q = q', X]] && \text{consistency} \\
 &= \mathbb{E}_X[\mathbb{E}[M|X] - \mathbb{E}[M|X]] = 0. && \text{conditional independence}
 \end{aligned}$$

The absence of causal effect and association between  $Q, M$  in this formulation suggests that, without further assumptions, traditional IQA metrics should not be predictive of DNN performance. Furthermore, while No Reference and Full Reference (FR) IQA differ in their assumptions and setup,

we provide their causal models as special cases of Figure 2 in Appendix A to show that  $Q \perp M | X$  holds in both cases.

### 4.3 Shared features IQA formulation

Ideally, we would like  $Q$  and  $M$  to become dependent when conditioning on  $X$  (i.e., we can learn about  $M$  by observing only  $Q$ ). This occurs in the case where there exists a common set of features  $Z$  that are utilized both for the prediction function of  $\hat{Y}$  and the quality score  $Q$  as shown in Figure 1. This scenario does not presume a singular set of  $Z$  that serves the task model and quality metric, but rather, the  $Z$  shown here represents the intersection of features used by both. The existence of  $Z$  ensures  $Q, M$  are no longer independent given  $X$ . The primary question is whether such a  $Z$  exists or whether  $Q, M$  are related only as shown in Figure 2. An expanded discussion relating the baseline and shared features models can be found in Appendix B.

**Remark: Correlation vs. Causation** While we typically use causal models to estimate cause-effect relationships, a key clarification here is that we seek a weaker criterion, namely to establish the conditions under which quality and performance are *at least* correlated. Since we know that  $X$  is a common cause for both  $Q, M$ , we first want to ensure that when we estimate  $Q, M$  from  $X$  we know the conditions under which  $Q, M$  will be related via the same features of  $X$  (e.g., via  $Z$ ).

To address the aforementioned issues, we consider instead a weaker form of task-guidance where quality metrics can be aligned with task-specific information without requiring the expense of training a new task model directly on the dataset of interest (**D4**). In this setting shown in Figure 7, the computation of  $Q$  is dependent not only on the image  $X$  but on the label set  $\mathcal{Y}$ . The task  $T$  is used as a selection variable [61, 62] on which the dataset is conditioned, and as a collider in the DAG,  $T$  creates an association between  $M, Q$ .

### 4.4 Zero-Shot CLIP IQA

Using the causal framework, we propose a new quality metric that uses Zero-Shot (ZS) capabilities of the multi-modal CLIP foundation model [18] in order to address all desiderata. In particular, we derive a new image quality metric (ZSCLIP-IQA) based on a zero-shot classification problem for our data and task of interest (Figure 7).

Let  $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$  be our dataset with images  $x$  and labels  $y$ ,  $f : \mathcal{X} \rightarrow \mathcal{Z}$  be our CLIP image embedding network ( $\mathcal{Z} \in \mathbb{R}^d$ ), and  $g : \mathcal{T} \rightarrow \mathcal{W}$  be our CLIP text/token-embedding network ( $\mathcal{W} \in \mathbb{R}^d$ ).

We define a set of task-relevant classes/tokens  $\mathcal{T} = \{T_i\}_{i=1}^K$  that capture the text labels for concepts or entities likely to occur in the images (e.g.,  $K = 1000$  classes in the ImageNet dataset). We embed each of the text tokens  $g(T_i) = w_i$  and normalize to get a unit vector representation for each token. Note when using CLIP as our text embedding network, we may also augment  $T_i$  to include additional words (e.g., "A picture of a <token>"). The full set  $\mathbf{W} = [w_0; w_1; \dots; w_K] \in \mathbb{R}^{d \times K}$  constitutes the ZS weights.

To evaluate image quality, we compute image embeddings  $f_\theta(x) = z$  for  $x \in \mathcal{D}$  which we normalize to be unit length. For each image, we compute the cosine similarity between the image embedding and each of the tokens  $s = z\mathbf{W}$  (with  $z \in \mathbb{R}^d$ ,  $s \in \mathbb{R}^K$ ) and compute estimated class probabilities via a softmax over similarity scores ( $\hat{y} = \text{softmax}(s)$ ). As in the strong TG-IQA scenario, we implement three variants of  $Q$ : (1) Max probability:  $Q_p := \max_i \hat{y}_i$ , (2) Entropy:  $Q_h := H(\hat{y}) = -\sum_i \hat{y}_i \log \hat{y}_i$ , and (3) Max-logit:  $Q_l := \max_i s_i$ .

## Data Availability

The datasets generated and/or analyzed during the current study are available in the ImageNet-C repository (<https://zenodo.org/records/2235448>).

## Code Availability

The code used for this study is available from the corresponding author upon reasonable request.

## Acknowledgments

This work is funded by Independent Research & Development (IR&D) program funds in the Research and Exploratory Development Department (REDD) at the Johns Hopkins University Applied Physics Laboratory.

## Author Contributions

ND developed the method, designed and ran the experiments, analyzed the results and wrote the main manuscript. MU supported the development of the method, design of experiments, and contributions to the main manuscript. All authors reviewed the manuscript.

## Competing Interests

The authors declare no competing interests.

## References

1. Taori, R., Dave, A., Shankar, V., Carlini, N., Recht, B. & Schmidt, L. *Measuring Robustness to Natural Distribution Shifts in Image Classification* in *Advances in Neural Information Processing Systems* (eds Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F. & Lin, H.) **33** (Curran Associates, Inc., 2020), 18583–18599.
2. Djolonga, J., Yung, J., Tschannen, M., Romijnders, R., Beyer, L., Kolesnikov, A., Puigcerver, J., Minderer, M., D’Amour, A. N., Moldovan, D., Gelly, S., Houlsby, N., Zhai, X. & Lučić, M. *On Robustness and Transferability of Convolutional Neural Networks* in *Conference on Computer Vision and Pattern Recognition* (2021).
3. Ibrahim, M., Garrido, Q., Morcos, A. & Bouchacourt, D. The robustness limits of SoTA vision models to natural variation. *arXiv [cs.CV]* (2022).
4. Geirhos, R., Narayananappa, K., Mitzkus, B., Thieringer, T., Bethge, M., Wichmann, F. A. & Brendel, W. *Partial success in closing the gap between human and machine vision* in *Advances in Neural Information Processing Systems* (2021).
5. Ilyas, A., Park, S. M., Engstrom, L., Leclerc, G. & Madry, A. Datamodels: Predicting predictions from training data. *arXiv [stat.ML]* (2022).
6. Lin, J., Zhang, A., Lécuyer, M., Li, J., Panda, A. & Sen, S. *Measuring the Effect of Training Data on Deep Learning Predictions via Randomized Experiments* in *Proceedings of the 39th International Conference on Machine Learning* (eds Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G. & Sabato, S.) **162** (PMLR, 2022), 13468–13504.
7. Pavlak, M. F., Drenkow, N., Petrick, N., Farhangi, M. M. & Unberath, M. Data AUDIT: Identifying attribute utility- and detectability-induced bias in task models. *Medical Image Computing and Computer-Assisted Intervention*, 442–452 (2023).
8. Drenkow, N., Pavlak, M., Harrigan, K., Zirikly, A., Subbaswamy, A., Farhangi, M. M., Petrick, N. & Unberath, M. Detecting dataset bias in medical AI: A generalized and modality-agnostic auditing framework. *arXiv [cs.LG]* (2025).
9. Hendrycks, D. & Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv [cs.LG]* (2019).
10. Dodge, S. & Karam, L. Understanding how image quality affects deep neural networks. *arXiv [cs.CV]* (2016).
11. Laugros, A., Caplier, A. & Ospici, M. Are adversarial robustness and common perturbation robustness independent attributes? *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 1045–1054 (2019).
12. Drenkow, N. & Unberath, M. *RobustCLEVR: A Benchmark and Framework for Evaluating Robustness in Object-centric Learning* in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (2024), 4518–4527.
13. Drenkow, N., Ribaudo, C. & Unberath, M. Causality-driven audits of model robustness. *arXiv [cs.CV]* (2024).

14. Wang, Z., Bovik, A. C., Sheikh, H. R. & Simoncelli, E. P. Image quality assessment: from error visibility to structural similarity. en. *IEEE Trans. Image Process.* **13**, 600–612 (4 2004).
15. Zhang, L., Zhang, L., Mou, X. & Zhang, D. FSIM: a feature similarity index for image quality assessment. en. *IEEE Trans. Image Process.* **20**, 2378–2386 (8 2011).
16. Xu, S., Jiang, S. & Min, W. No-reference/Blind Image Quality Assessment: A Survey. *IETE Tech. Rev.* **34**, 223–245 (2017).
17. Agnolucci, L., Galteri, L., Bertini, M. & Del Bimbo, A. *Arniqa: Learning distortion manifold for image quality assessment* in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (2024), 189–198.
18. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G. & Sutskever, I. Learning transferable visual models from natural language supervision. *arXiv [cs.CV]* (2021).
19. Sun, C., Shrivastava, A., Singh, S. & Gupta, A. Revisiting unreasonable effectiveness of data in deep learning era. *ICCV*, 843–852 (2017).
20. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S., Crowson, K., Schmidt, L., Kaczmarczyk, R. & Jitsev, J. LAION-5B: An open large-scale dataset for training next generation image-text models. *Adv. Neural Inf. Process. Syst.* **abs/2210.08402** (2022).
21. Wang, J., Chan, K. C. K. & Loy, C. C. Exploring CLIP for assessing the look and feel of images. *National Conference on Artificial Intelligence*, 2555–2563 (2022).
22. Mittal, A., Soundararajan, R. & Bovik, A. C. Making a “completely blind” image quality analyzer. en. *IEEE Signal Process. Lett.* **20**, 209–212 (3 2013).
23. Ye, P., Kumar, J., Kang, L. & Doermann, D. *Unsupervised feature learning framework for no-reference image quality assessment* en. in *2012 IEEE Conference on Computer Vision and Pattern Recognition* 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Providence, RI (IEEE, 2012), 1098–1105.
24. Xu, J., Ye, P., Li, Q., Du, H., Liu, Y. & Doermann, D. Blind image quality assessment based on High Order Statistics Aggregation. en. *IEEE Trans. Image Process.* **25**, 4444–4457 (9 2016).
25. Zhang, W., Ma, K., Yan, J., Deng, D. & Wang, Z. Blind image quality assessment using A deep bilinear convolutional neural network. *arXiv [eess.IV]* (2019).
26. Mittal, A., Moorthy, A. K. & Bovik, A. C. No-reference image quality assessment in the spatial domain. en. *IEEE Trans. Image Process.* **21**, 4695–4708 (12 2012).
27. He, K., Zhang, X., Ren, S. & Sun, J. *Deep residual learning for image recognition* in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), 770–778.
28. Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T. & Xie, S. A ConvNet for the 2020s. *arXiv [cs.CV]* (2022).
29. Tan, M. & Le, Q. V. EfficientNetV2: Smaller models and faster training. *arXiv [cs.CV]* (2021).
30. Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., Le, Q. V. & Adam, H. Searching for MobileNetV3. *arXiv [cs.CV]* (2019).
31. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. & Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv [cs.CV]* (2020).
32. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S. & Guo, B. Swin Transformer: Hierarchical vision Transformer using shifted windows. *arXiv [cs.CV]*, 10012–10022 (2021).
33. Marcel, S. & Rodriguez, Y. *Torchvision the machine-vision package of torch* in *Proceedings of the 18th ACM international conference on Multimedia* (Association for Computing Machinery, Firenze, Italy, 2010), 1485–1488.
34. Hendrycks, D., Basart, S., Mazeika, M., Zou, A., Kwon, J., Mostajabi, M., Steinhardt, J. & Song, D. Scaling out-of-distribution detection for real-world settings. *arXiv [cs.CV]* (2019).
35. Hendrycks, D. & Gimpel, K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv [cs.NE]* (2016).
36. Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T. L., Parajuli, S., Guo, M., Song, D., Steinhardt, J. & Gilmer, J. The many faces of robustness: A critical analysis of out-of-distribution generalization. *ICCV*, 8320–8329 (2020).

37. Wang, D., Shelhamer, E., Liu, S., Olshausen, B. & Darrell, T. Tent: Fully test-time adaptation by entropy minimization. *arXiv [cs.LG]* (2020).
38. Zhang, R., Isola, P., Efros, A. A., Shechtman, E. & Wang, O. *The unreasonable effectiveness of deep features as a perceptual metric* in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2018), 586–595.
39. Drenkow, N., Sani, N., Shpitser, I. & Unberath, M. A systematic review of robustness in deep learning for computer vision: Mind the gap? *arXiv [cs.CV]* (2021).
40. Michaelis, C., Mitzkus, B., Geirhos, R., Rusak, E., Bringmann, O., Ecker, A. S., Bethge, M. & Brendel, W. Benchmarking Robustness in Object Detection: Autonomous Driving when Winter is Coming. *arXiv [cs.CV]* (2019).
41. Laugros, A., Caplier, A. & Ospici, M. Using synthetic corruptions to measure robustness to natural distribution shifts. *arXiv [cs.CV]* (2021).
42. Varga, D. Understanding how image quality affects transformer neural networks. en. *Signals* **5**, 562–579 (2024).
43. Oguine, K. J., Mukul, R. D. S., Drenkow, N. & Unberath, M. *From generalization to precision: exploring SAM for tool segmentation in surgical environments* in *Medical Imaging 2024: Image Processing* **12926** (2024), 7–12.
44. Zhao, B., Yu, S., Ma, W., Yu, M., Mei, S., Wang, A., He, J., Yuille, A. & Kortylewski, A. OOD-CV: A benchmark for robustness to out-of-distribution shifts of individual nuisances in natural images. *arXiv [cs.CV]* (2021).
45. Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A. & Brendel, W. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv [cs.CV]* (2018).
46. Hermann, K. L., Chen, T. & Kornblith, S. The origins and prevalence of texture bias in convolutional neural networks. *arXiv [cs.CV]* (2019).
47. Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M. & Wichmann, F. A. Shortcut learning in deep neural networks. en. *Nat. Mach. Intell.* **2**, 665–673 (2020).
48. Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J. & Oermann, E. K. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. en. *PLoS Med.* **15**, e1002683 (2018).
49. Brown, A., Tomasev, N., Freyberg, J., Liu, Y., Karthikesalingam, A. & Schrouff, J. Detecting shortcut learning for fair medical AI using shortcut testing. en. *arXiv [cs.LG]*, 4314 (2022).
50. Ong Ly, C., Unnikrishnan, B., Tadic, T., Patel, T., Duhamel, J., Kandel, S., Moayedi, Y., Brudno, M., Hope, A., Ross, H. & McIntosh, C. Shortcut learning in medical AI hinders generalization: method for estimating AI model generalization without external data. en. *npj Digital Medicine* **7**, 1–10 (2024).
51. Wichmann, F. A. & Geirhos, R. Are Deep Neural Networks Adequate Behavioral Models of Human Visual Perception? en. *Annu Rev Vis Sci* **9**, 501–524 (2023).
52. Tan, H., Wu, S., Du, F., Chen, Y., Wang, Z., Wang, F. & Qi, X. Data pruning via moving-one-sample-out. *arXiv [cs.LG]* (2023).
53. He, M., Yang, S., Huang, T. & Zhao, B. Large-scale dataset pruning with dynamic uncertainty. *arXiv [cs.LG]* (2023).
54. Abbas, A., Rusak, E., Tirumala, K., Brendel, W., Chaudhuri, K. & Morcos, A. S. Effective pruning of web-scale datasets based on complexity of concept clusters. *arXiv [cs.CV]* (2024).
55. Kwok, D., Anand, N., Frankle, J., Dziugaite, G. K. & Rolnick, D. Dataset difficulty and the role of inductive bias. *arXiv [cs.LG]* (2024).
56. Eyuboglu, S., Varma, M., Saab, K., Delbrouck, J.-B., Lee-Messer, C., Dunnmon, J., Zou, J. & Ré, C. Domino: Discovering systematic errors with cross-modal embeddings. *arXiv [cs.LG]* (2022).
57. Chung, Y., Kraska, T., Polyzotis, N., Tae, K. H. & Whang, S. E. *Slice Finder: Automated Data Slicing for Model Validation* in *2019 IEEE 35th International Conference on Data Engineering (ICDE)* (IEEE, 2019), 1550–1553.
58. Sohoni, N. S., Dunnmon, J. A., Angus, G., Gu, A. & Ré, C. No subclass left behind: Fine-grained robustness in coarse-grained classification problems. *arXiv [cs.LG]* (2020).

59. Chen, V. S., Wu, S., Weng, Z., Ratner, A. & Ré, C. Slice-based Learning: A Programming Model for Residual Learning in Critical Data Slices. en. *Adv. Neural Inf. Process. Syst.* **32**, 9392–9402 (2019).
60. Ma, K., Liu, W., Zhang, K., Duanmu, Z., Wang, Z. & Zuo, W. End-to-end blind image quality assessment using deep neural networks. en. *IEEE Trans. Image Process.* **27**, 1202–1213 (3 2018).
61. Zadrozny, B. *Learning and evaluating classifiers under sample selection bias* in *Twenty-first international conference on Machine learning - ICML '04* (ACM Press, Banff, Alberta, Canada, 2004).
62. Bareinboim, E., Tian, J. & Pearl, J. in *Probabilistic and Causal Inference* 433–450 (ACM, New York, NY, USA, 2022).

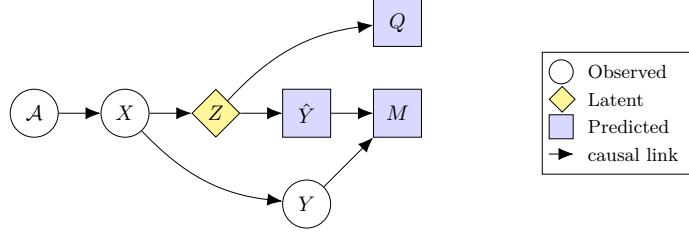


Figure 1: **Baseline IQA formulation with shared latent features.** Causal diagram relating model accuracy ( $M$ ) with IQ metrics ( $Q$ ). In this case,  $Q$  and  $M$  are related via a common cause  $Z$  that represents features in the image  $X$  that influence both the prediction and quality.

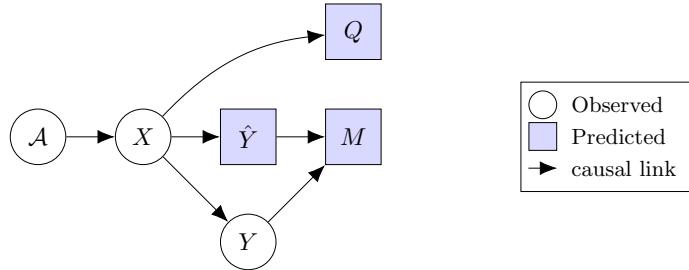


Figure 2: **Baseline IQA formulation.** Causal diagram relating model accuracy ( $M$ ) with IQ metrics ( $Q$ ).

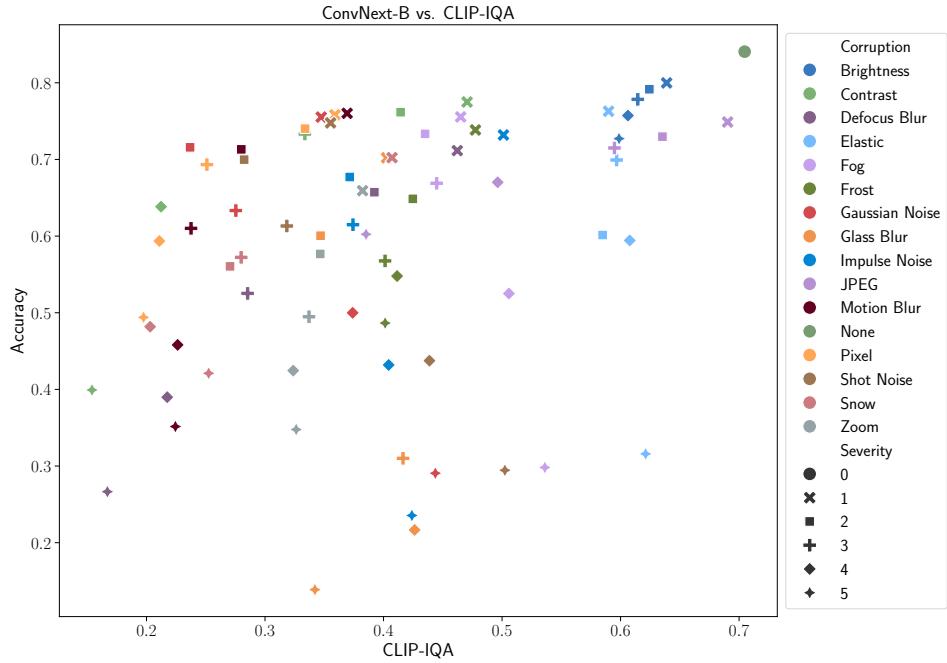


Figure 3: **Relationship between conventional NR-IQA metric and classification accuracy.** Accuracy ( $M$ ) vs. IQ ( $Q$ ) for ConvNext-B and CLIP-IQA respectively. Each point represents the average accuracy over all images in the ImageNet val set corrupted with the corresponding corruption/severity. Little correlation is observed between  $M$ ,  $Q$  across all corruptions/severities and  $Q$  is weakly predictive of  $M$ .

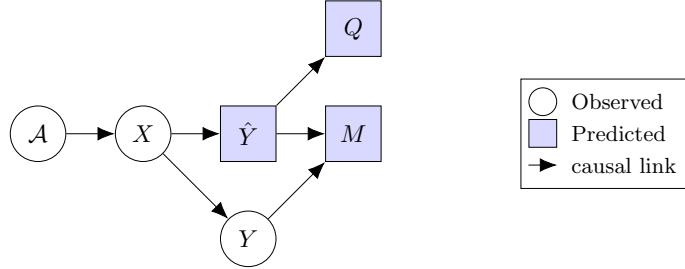


Figure 4: **Strong task-guided IQA.** Causal diagram relating model accuracy ( $M$ ) with IQ metrics ( $Q$ ) with the additional dependence  $\hat{Y} \rightarrow Q$ .

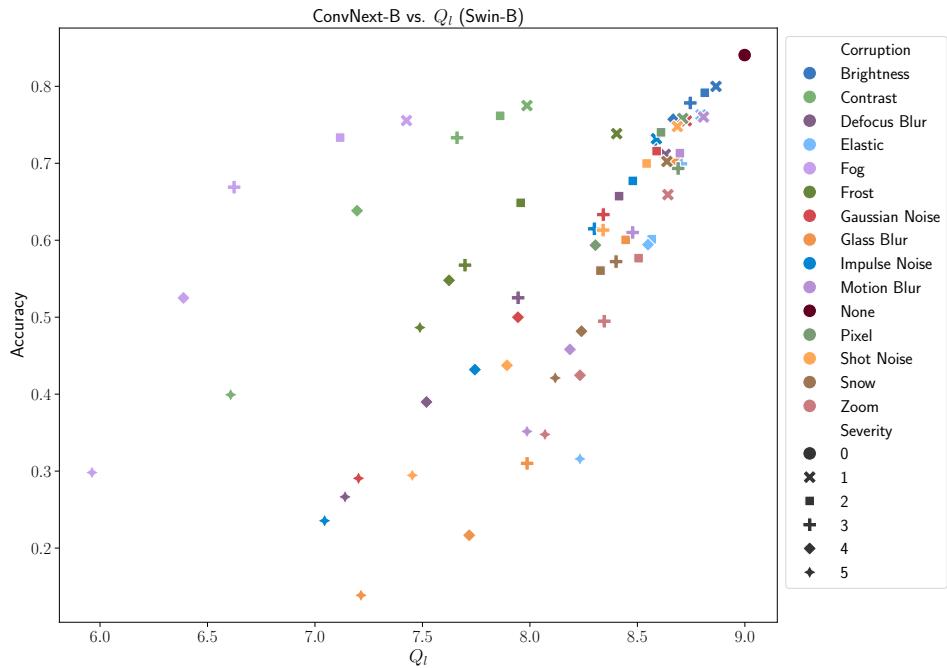


Figure 5: **Relationship of strong task-guided IQA and classification accuracy.** Accuracy vs. max logit using ConvNext-B for the task model and Swin-B for  $Q_l$  which generally outperforms the other variants

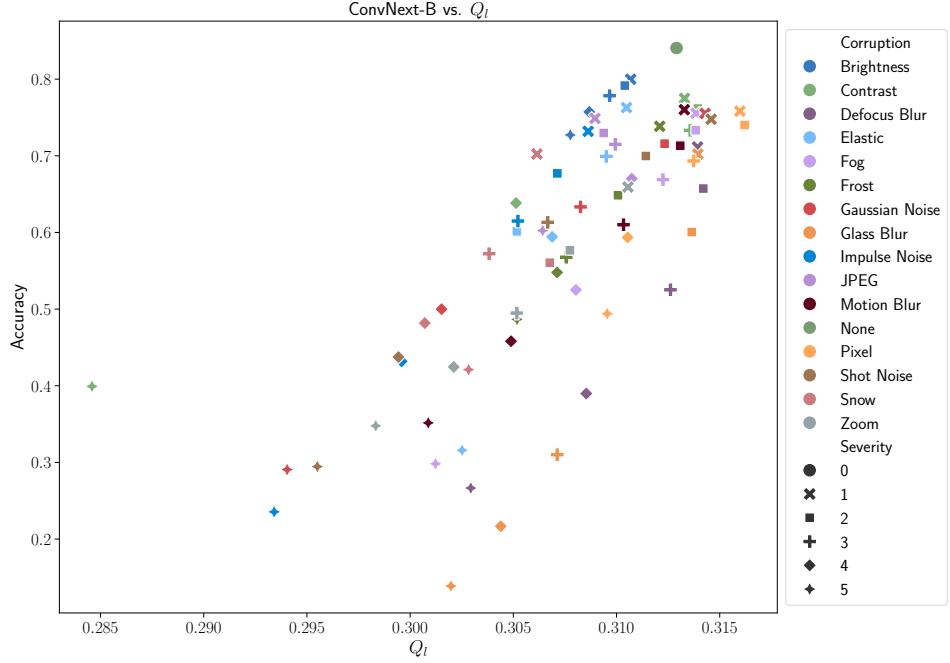


Figure 6: **Relationship of weak task-guided IQA and classification accuracy.** Accuracy vs. IQ with ConvNext-B as the task model  $M$  and ZSCLIP-IQA max-logit as the quality metric  $Q_l$  which generally outperforms the other variants.

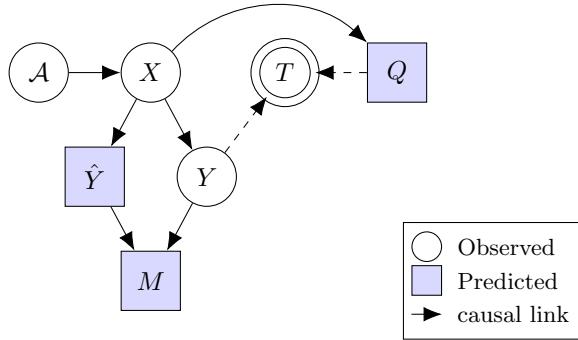


Figure 7: **Weak task-guided IQA.** Causal DAG relating model performance ( $M$ ) with IQ metrics ( $Q$ ).  $Q$  uses information about the label set  $\mathcal{Y}$ . The task  $T$  is viewed as a selection variable which influences both the labels  $Y$  and  $Q$ .

Table 1: Correlation between IQ and accuracy, correctness. SRCC, PLCC computed using average accuracy for each (corruption, severity). AUC and CE based on point-wise predictions (95% CI within  $\pm 0.001$ ). SRCC, PLCC values have  $p < 0.05$ .

Model	IQA Metric	AUC $\uparrow$	CE $\downarrow$	$ KRCC  \uparrow$	$ PLCC  \uparrow$	$ SRCC  \uparrow$
ConvNext-B	ARNIQA	0.517	0.677	$0.088 \pm 0.149$	$0.168 \pm 0.215$	$0.127 \pm 0.214$
	BRISQUE	0.568	0.670	$0.255 \pm 0.129$	$0.398 \pm 0.190$	$0.374 \pm 0.182$
	CLIP-IQA	0.567	0.670	$0.273 \pm 0.154$	$0.328 \pm 0.202$	$0.378 \pm 0.212$
	TV	0.477	0.676	$0.108 \pm 0.183$	$0.138 \pm 0.294$	$0.151 \pm 0.255$
ResNet34	ARNIQA	0.499	0.663	$0.003 \pm 0.155$	$0.006 \pm 0.205$	$0.007 \pm 0.225$
	BRISQUE	0.552	0.658	$0.175 \pm 0.140$	$0.278 \pm 0.186$	$0.254 \pm 0.202$
	CLIP-IQA	0.599	0.647	$0.307 \pm 0.159$	$0.467 \pm 0.188$	$0.429 \pm 0.207$
	TV	0.500	0.657	$0.051 \pm 0.194$	$0.291 \pm 0.256$	$0.047 \pm 0.264$
EfficientNet-V2-M	ARNIQA	0.522	0.674	$0.098 \pm 0.149$	$0.204 \pm 0.213$	$0.143 \pm 0.214$
	BRISQUE	0.579	0.665	$0.271 \pm 0.137$	$0.431 \pm 0.186$	$0.393 \pm 0.188$
	CLIP-IQA	0.574	0.666	$0.276 \pm 0.158$	$0.332 \pm 0.209$	$0.389 \pm 0.213$
	TV	0.468	0.675	$0.109 \pm 0.181$	$0.075 \pm 0.275$	$0.158 \pm 0.254$
MobileNet-V3-L	ARNIQA	0.503	0.682	$0.008 \pm 0.148$	$0.024 \pm 0.215$	$0.019 \pm 0.216$
	BRISQUE	0.562	0.676	$0.198 \pm 0.137$	$0.314 \pm 0.191$	$0.285 \pm 0.196$
	CLIP-IQA	0.586	0.669	$0.300 \pm 0.159$	$0.412 \pm 0.192$	$0.415 \pm 0.215$
	TV	0.506	0.676	$0.038 \pm 0.191$	$0.298 \pm 0.271$	$0.035 \pm 0.261$
Swin-B	ARNIQA	0.510	0.675	$0.069 \pm 0.155$	$0.118 \pm 0.230$	$0.098 \pm 0.222$
	BRISQUE	0.574	0.667	$0.291 \pm 0.131$	$0.443 \pm 0.183$	$0.426 \pm 0.178$
	CLIP-IQA	0.571	0.667	$0.290 \pm 0.161$	$0.361 \pm 0.199$	$0.410 \pm 0.211$
	TV	0.485	0.674	$0.090 \pm 0.180$	$0.153 \pm 0.299$	$0.123 \pm 0.255$
ViT-L16	ARNIQA	0.508	0.674	$0.067 \pm 0.157$	$0.098 \pm 0.226$	$0.097 \pm 0.224$
	BRISQUE	0.562	0.668	$0.274 \pm 0.125$	$0.430 \pm 0.172$	$0.407 \pm 0.175$
	CLIP-IQA	0.575	0.665	$0.293 \pm 0.152$	$0.414 \pm 0.180$	$0.422 \pm 0.202$
	TV	0.485	0.673	$0.103 \pm 0.182$	$0.181 \pm 0.289$	$0.136 \pm 0.247$

Table 2: Correlation between IQ and accuracy, correctness. KRCC, SRCC, PLCC computed using average accuracy for each (corruption, severity). AUC and CE based on point-wise predictions (95% CI within  $\pm 0.001$ ). KRCC, SRCC, PLCC values have  $p < 0.05$ . Full table in Appendix E.

Model	IQA Metric	AUC $\uparrow$	CE $\downarrow$	$ KRCC  \uparrow$	$ PLCC  \uparrow$	$ SRCC  \uparrow$
ConvNext-B	$Q_h$	0.772	0.562	$0.660 \pm 0.070$	$0.822 \pm 0.070$	$0.854 \pm 0.063$
	$Q_l$	0.778	0.555	$0.660 \pm 0.067$	$0.826 \pm 0.067$	$0.854 \pm 0.063$
	$Q_p$	0.826	0.504	$0.738 \pm 0.045$	$0.888 \pm 0.045$	$0.910 \pm 0.044$
ResNet34	$Q_h$	0.848	0.470	$0.862 \pm 0.028$	$0.930 \pm 0.028$	$0.969 \pm 0.023$
	$Q_l$	0.827	0.492	$0.870 \pm 0.015$	$0.951 \pm 0.015$	$0.973 \pm 0.020$
	$Q_p$	0.850	0.461	$0.886 \pm 0.015$	$0.960 \pm 0.015$	$0.977 \pm 0.021$
EfficientNet-V2-M	$Q_h$	0.831	0.497	$0.888 \pm 0.020$	$0.956 \pm 0.020$	$0.981 \pm 0.012$
	$Q_l$	0.831	0.496	$0.876 \pm 0.029$	$0.937 \pm 0.029$	$0.977 \pm 0.014$
	$Q_p$	0.862	0.456	$0.900 \pm 0.014$	$0.968 \pm 0.014$	$0.984 \pm 0.010$
MobileNet-V3-L	$Q_h$	0.833	0.496	$0.906 \pm 0.019$	$0.973 \pm 0.019$	$0.984 \pm 0.013$
	$Q_l$	0.835	0.496	$0.895 \pm 0.031$	$0.951 \pm 0.031$	$0.982 \pm 0.015$
	$Q_p$	0.853	0.470	$0.922 \pm 0.010$	$0.983 \pm 0.010$	$0.988 \pm 0.010$
Swin-B	$Q_h$	0.766	0.578	$0.532 \pm 0.207$	$0.483 \pm 0.207$	$0.654 \pm 0.174$
	$Q_l$	0.732	0.597	$0.485 \pm 0.203$	$0.458 \pm 0.203$	$0.611 \pm 0.181$
	$Q_p$	0.807	0.529	$0.603 \pm 0.184$	$0.620 \pm 0.184$	$0.732 \pm 0.142$
ViT-L16	$Q_h$	0.799	0.545	$0.633 \pm 0.124$	$0.690 \pm 0.124$	$0.819 \pm 0.092$
	$Q_l$	0.780	0.556	$0.592 \pm 0.124$	$0.672 \pm 0.124$	$0.786 \pm 0.092$
	$Q_p$	0.828	0.506	$0.694 \pm 0.104$	$0.765 \pm 0.104$	$0.864 \pm 0.075$

Table 3: Correlation between IQ and accuracy. KRCC, SRCC, PLCC based on average accuracy for each (corruption, severity) combination. AUC and CE are computed based on point-wise predictions and all have 95% CI within  $\pm 0.001$ . All KRCC, SRCC, PLCC values have  $p < 0.005$ .

Model	ZSCLIP-IQA	AUC $\uparrow$	CE $\downarrow$	KRCC   $\uparrow$	PLCC   $\uparrow$	SRCC   $\uparrow$
ConvNext-B	$Q_h$	0.349	0.677	$0.738 \pm 0.067$	$0.869 \pm 0.059$	$0.906 \pm 0.048$
	$Q_l$	0.675	0.632	$0.573 \pm 0.084$	$0.764 \pm 0.080$	$0.783 \pm 0.085$
	$Q_p$	0.602	0.677	$0.788 \pm 0.056$	$0.884 \pm 0.048$	$0.937 \pm 0.034$
ResNet34	$Q_h$	0.666	0.663	$0.800 \pm 0.059$	$0.936 \pm 0.028$	$0.945 \pm 0.032$
	$Q_l$	0.692	0.609	$0.630 \pm 0.087$	$0.814 \pm 0.055$	$0.820 \pm 0.084$
	$Q_p$	0.368	0.663	$0.834 \pm 0.049$	$0.949 \pm 0.024$	$0.959 \pm 0.026$
EfficientNet-V2-M	$Q_h$	0.344	0.675	$0.721 \pm 0.074$	$0.844 \pm 0.069$	$0.890 \pm 0.061$
	$Q_l$	0.674	0.631	$0.545 \pm 0.084$	$0.717 \pm 0.079$	$0.757 \pm 0.089$
	$Q_p$	0.604	0.675	$0.759 \pm 0.062$	$0.860 \pm 0.051$	$0.920 \pm 0.042$
MobileNet-V3-L	$Q_h$	0.655	0.682	$0.740 \pm 0.065$	$0.896 \pm 0.047$	$0.907 \pm 0.046$
	$Q_l$	0.686	0.629	$0.559 \pm 0.087$	$0.765 \pm 0.084$	$0.768 \pm 0.091$
	$Q_p$	0.377	0.682	$0.785 \pm 0.055$	$0.918 \pm 0.039$	$0.936 \pm 0.033$
Swin-B	$Q_h$	0.354	0.676	$0.735 \pm 0.072$	$0.864 \pm 0.063$	$0.904 \pm 0.054$
	$Q_l$	0.672	0.633	$0.538 \pm 0.094$	$0.718 \pm 0.096$	$0.743 \pm 0.098$
	$Q_p$	0.601	0.676	$0.778 \pm 0.059$	$0.879 \pm 0.050$	$0.935 \pm 0.034$
ViT-L16	$Q_h$	0.363	0.674	$0.789 \pm 0.053$	$0.915 \pm 0.042$	$0.938 \pm 0.032$
	$Q_l$	0.673	0.630	$0.592 \pm 0.086$	$0.775 \pm 0.088$	$0.791 \pm 0.083$
	$Q_p$	0.599	0.674	$0.845 \pm 0.045$	$0.920 \pm 0.042$	$0.962 \pm 0.023$

## A Causal model for NR-/FR-IQA

We provide here the causal models for the NR-IQA and FR-IQA settings in Figure 8.

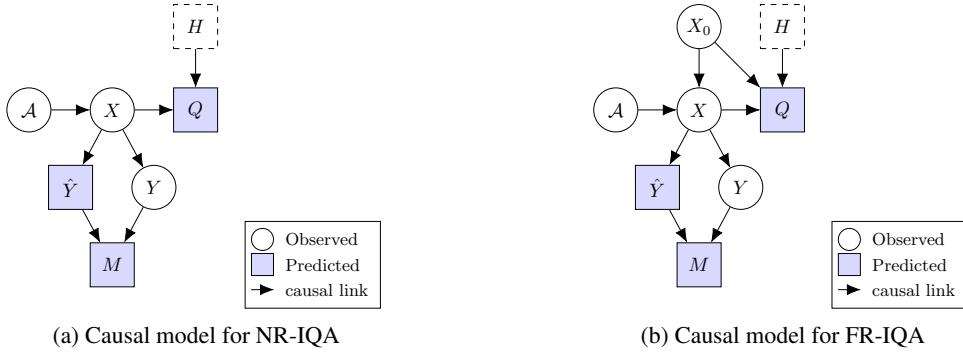


Figure 8: Causal model for the NR-/FR-IQA settings. The FR-IQA setting includes  $X_0$  which is the reference image. In both models,  $H$  indicates human annotator guidance which reflects that IQ metrics are typically calibrated against human perceptual judgements (dashed-line box indicates  $H$  is not used directly in the calculation of  $Q$ ).

While the main manuscript focused on the NR-IQA setting, we can see from the causal models here that the results generalize to the FR-IQA case as well. In particular, the independence of  $Q, M$  given  $X$  is not affected by whether a “clean” reference image ( $X_0$ ) is available for computing  $Q$ . Also, these models also account for the influence of human annotators  $H$  in calibrating the function for computing  $Q$ , but not that this does not change the relationship between  $M, Q$ .

## B Causal model for IQA with latent features

In understanding the difference between the baseline IQA formulation in Figure 2 and the shared features formulation of Figure 1 in Section 4, we provide an expanded version of the baseline DAG in Figure 9. Here we show that the task DNN for computing  $\hat{Y}$  and the function for computing  $Q$  rely on latent features  $Z_{\hat{Y}}$  and  $Z_Q$  respectively.

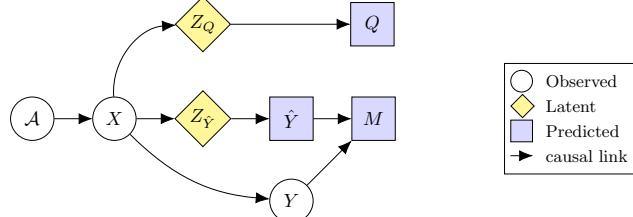


Figure 9: Causal model for IQA that accounts for the use of latent features by the task DNN and IQ metric towards computing  $M$  and  $Q$  respectively.

In this expanded model,  $Z_{\hat{Y}}$  and  $Z_Q$  are independent given  $X$  and not “shared”, and therefore  $Q \perp M \mid X$  as discussed in §4. In contrast, Figure 1 considers the case where  $Z$  represents the features derived from  $X$  that are common between  $Z_{\hat{Y}}$  and  $Z_Q$  shown in the baseline case above. Thus, Figure 1 shows the case where  $X$  does not block all paths between  $Q, M$  since a path exists from  $Q$  to  $M$  through  $Z$ . This ensures that  $Q$  and  $M$  will be correlated given  $X$  unlike in the baseline case above.

## C Causal model for common corruptions robustness evaluation

The common corruptions framework [9] is used in our experiments to ensure full control of the image distortion types and severity. Figure 10 shows a version of the baseline IQA causal model customized

to account for the corruption process used by this evaluation framework. Here, the corrupted image  $X$  is determined by the corruption function (e.g., Gaussian noise, defocus blur, fog, contrast, brightness, JPEG compression), the severity ( $S \in \{1, 2, 3, 4, 5\}$ ), and the “clean” image  $X_0$ .

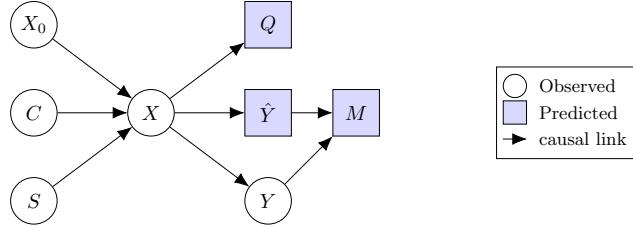


Figure 10: Causal model for the common corruptions framework where  $C$  refers to the corruption type,  $S$  refers to the corruption severity, and  $X_0$  is the unperturbed, “clean” image.

In this setting, we see that  $C, S$  replace the original set of imaging factors  $\mathcal{A}$  in the graph in Figure 2. As such, the analysis from §4 holds in the common corruptions framework and allows us to study the relationship between  $Q, M$  in a setting where we can precisely control the imaging conditions.

## D Relationship of NR-IQA and DNN performance metrics

In Figures 11, 12, and 13 we show the relationship of additional NR-IQA metrics with DNN performance for additional architectures and metrics. In general, we see weak trends in accuracy vs. average IQ suggesting that these metrics are most consistent with the causal model in Figure 2.

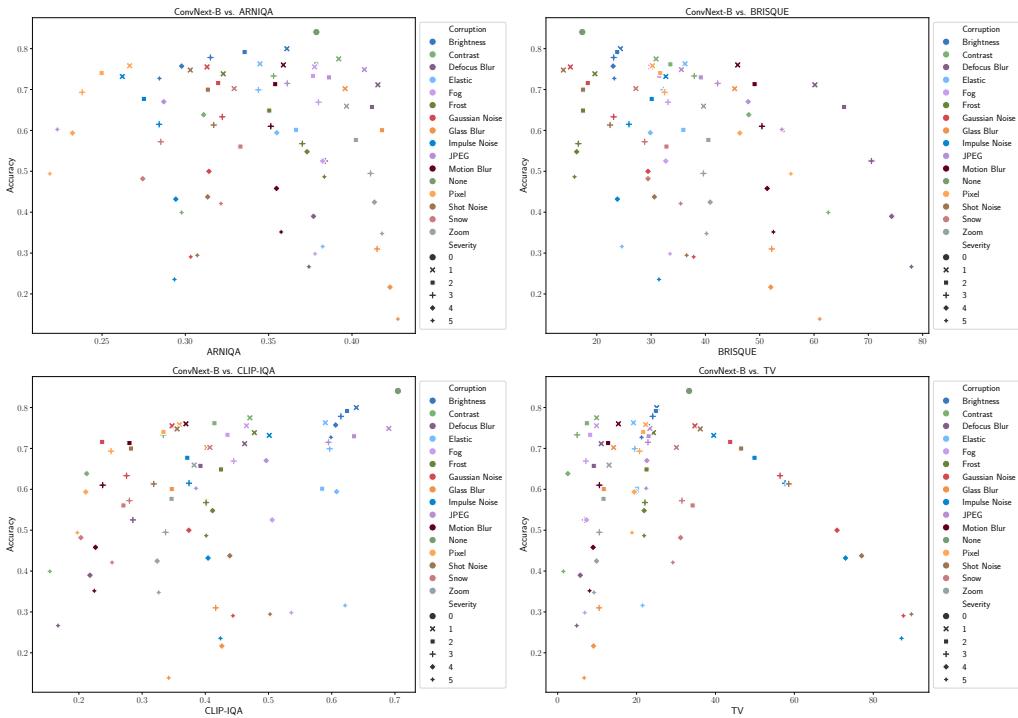


Figure 11: Comparison of ConvNext-B accuracy with (clockwise) ARNIQA, BRISQUE, CLIP-IQA, and TV. Little correlation is observed between group-wise accuracy and each NR-IQA metric.

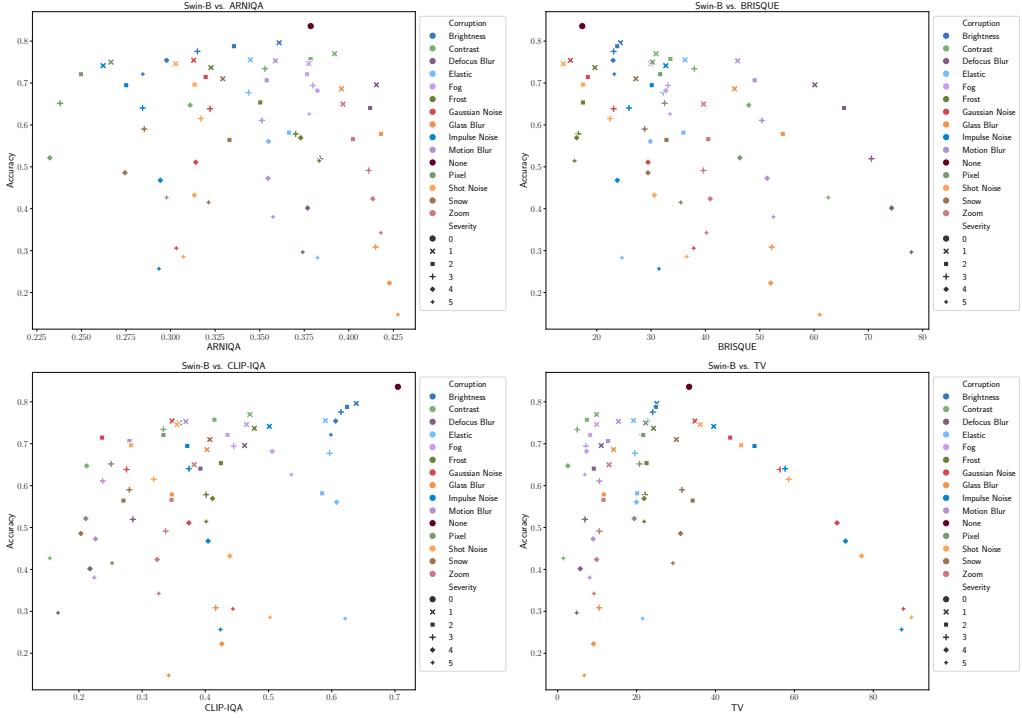


Figure 12: Comparison of Swin-B accuracy with (clockwise) ARNIQA, BRISQUE, CLIP-IQA, and TV. Little correlation is observed between group-wise accuracy and each NR-IQA metric.

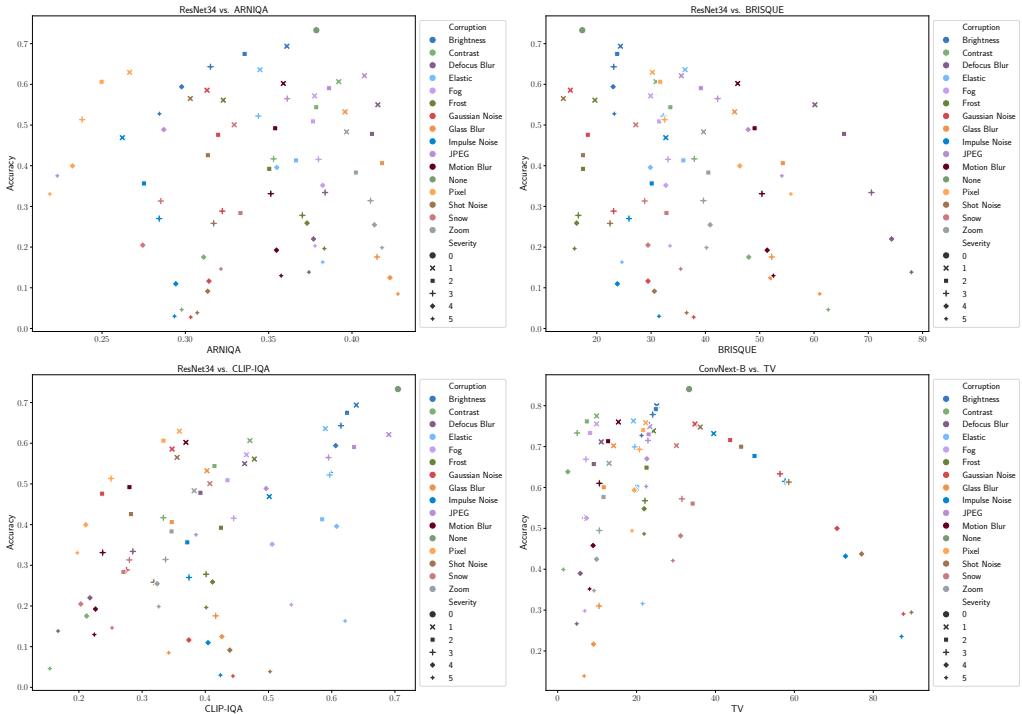


Figure 13: Comparison of ResNet34 accuracy with (clockwise) ARNIQA, BRISQUE, CLIP-IQA, and TV. Little correlation is observed between group-wise accuracy and each NR-IQA metric.

## E Relationship of strong task-guided IQA and DNN performance metrics

In Figures 14, 15, 16, we examine the relationship between DNN performance and the strong task-guided metrics ( $Q_p$ ,  $Q_h$ ,  $Q_l$ ) described in §2.3. Each figure pairs the task DNN under consideration with a pre-trained task model used to compute the quality metric.

Tables 4, 5, and 6 also show the point-wise predictability results for the strong task-guided IQA case. These tables extend Table 2 for additional task DNNs. Results here show that strong task-guided IQA metrics are highly correlated with DNN performance and that predictability remains high regardless of whether the pre-trained DNN used to compute  $Q$  is the same DNN used to obtain  $M$ .

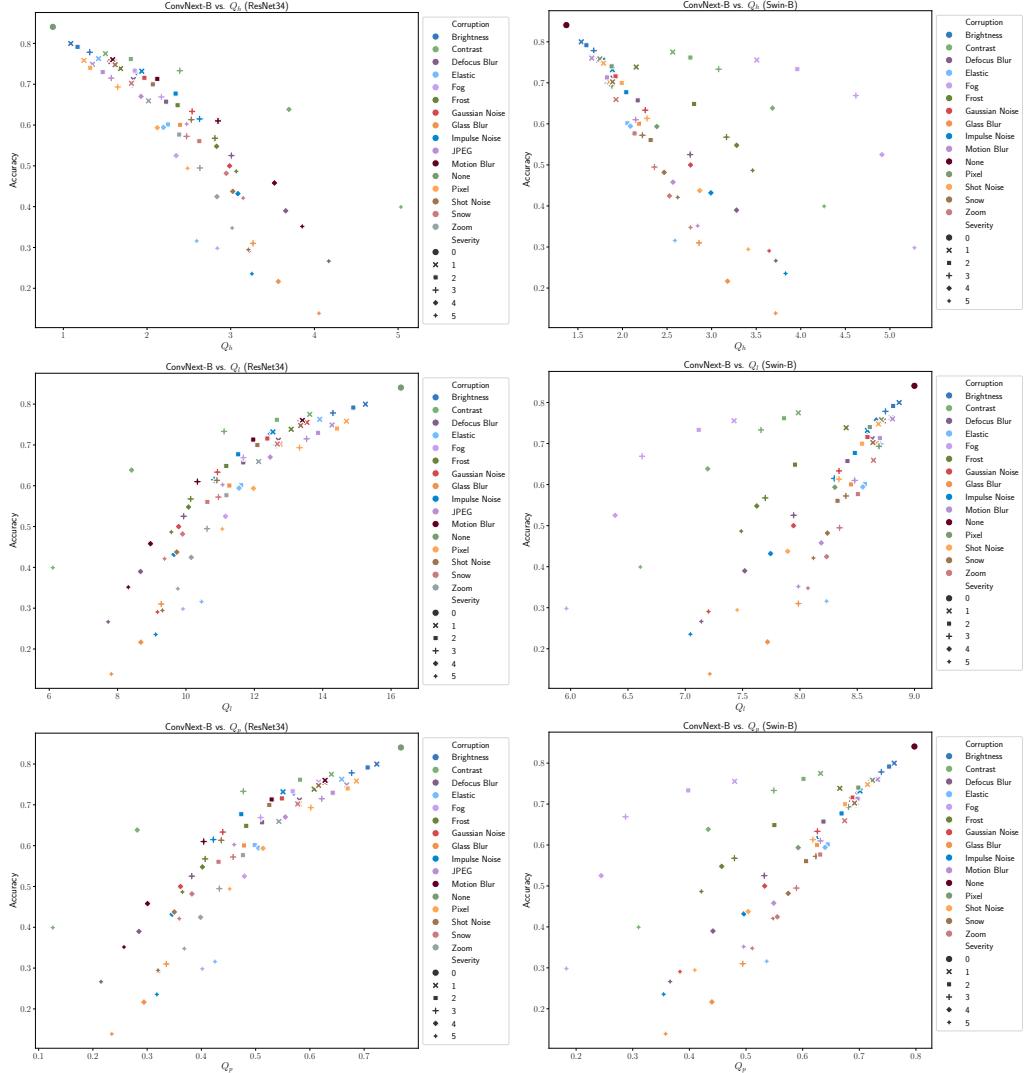


Figure 14: Comparison of ConvNext-B accuracy with (row)  $Q_h$ ,  $Q_l$ ,  $Q_p$  computed using (col) ResNet34, Swin-B. High correlation is observed between each IQA metric and accuracy.

## F Relationship of weak task-guided IQA and DNN performance metrics

We provide Figures 17, 18, 19 showing the relationship between DNN performance the weak task-guided ZSCLIP-IQA metric from §2.4.

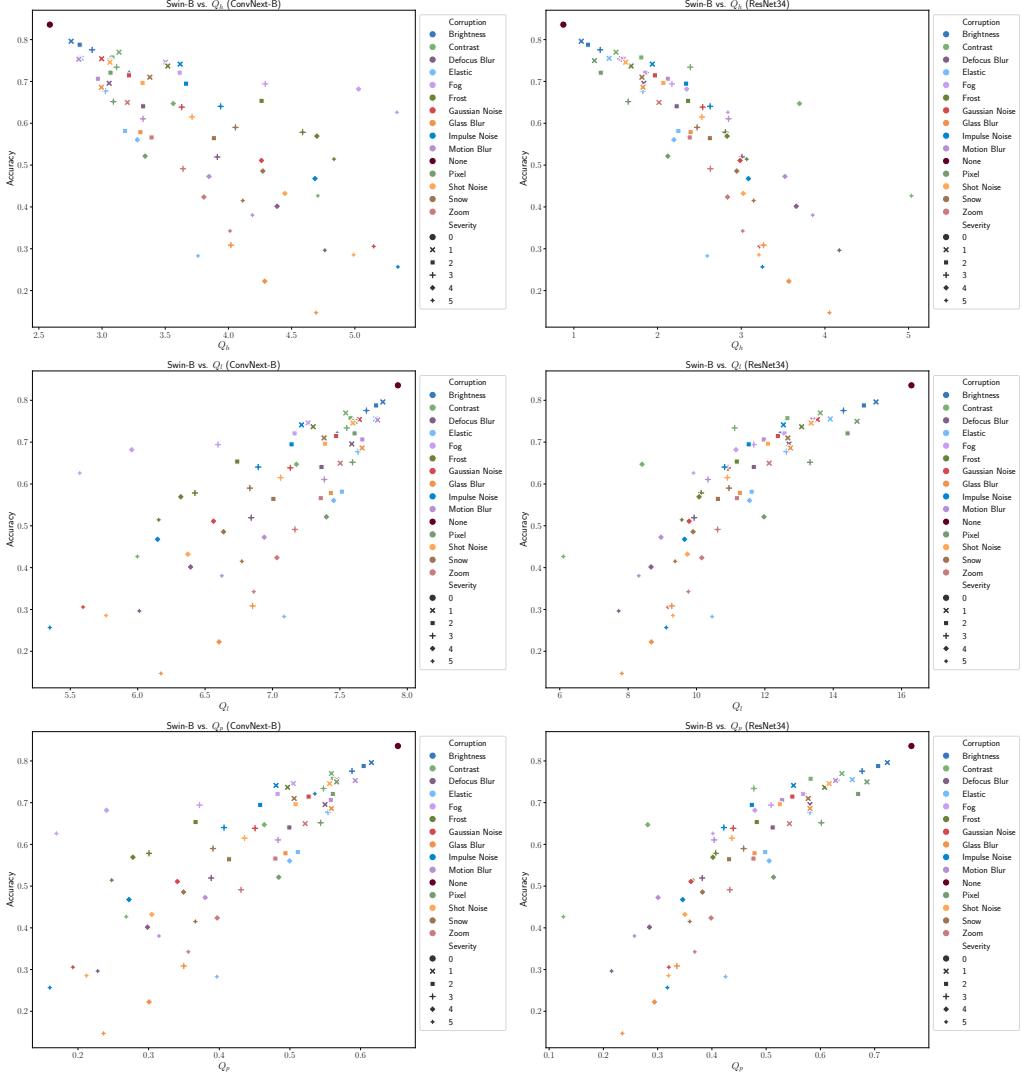


Figure 15: Comparison of Swin-B accuracy with (row)  $Q_h$ ,  $Q_l$ ,  $Q_p$  computed using (col) ConvNext-B, ResNet34. High correlation is observed between each IQA metric and accuracy.

## G Controlling for image content when evaluating predictability

The experiments of §2.2-2.4 examined predictability by modeling  $P(M|Q, X)$ . Since the content of  $X$  may be a confounder for both  $M, Q$ , we attempt to control for it in two ways.

In the first case, we take advantage of the synthetic nature of the IN-C dataset by looking at the predictability of  $M$  from  $Q$  for each image in the dataset separately which allows us to control the content precisely and only change the quality characteristics. For this experiment, we train a logistic regression classifier to predict  $P(M|Q)$  for individual image IDs trained using only  $M, Q$  computed from the set of distorted variants of each specific image ID. Given the original clean image and 15 corruptions with 5 severity levels each, we run 5-fold cross-validation with an 80/20 train/test split of the 76 total images. We repeat this for all 50k image IDs in the ImageNet validation set. The results shown in Figure 20a are an average of the AUC over all image IDs and folds.

We see that even when controlling for the image content the weak task-guided IQA generally achieves the highest  $mAUC$  with the lowest variance. Overall, this supports our hypothesis and causal analysis that weak task-guidance provides a means to associate  $M, Q$  even when conditioning on the image directly.

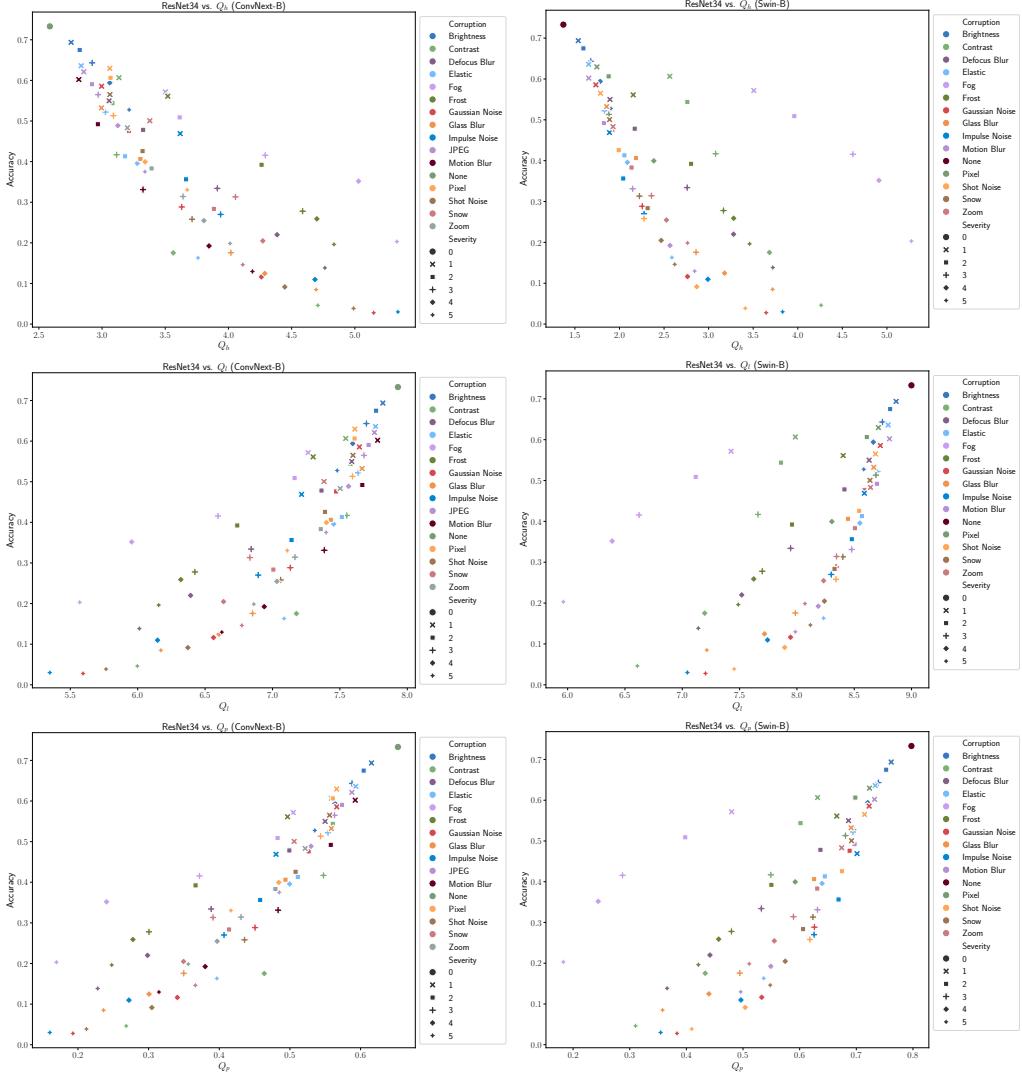


Figure 16: Comparison of ResNet34 accuracy with (row)  $Q_h$ ,  $Q_l$ ,  $Q_p$  computed using (col) ConvNext-B, Swin-B. High correlation is observed between each IQA metric and accuracy.

In the second case, we adjust for image content by controlling for the image label  $Y$ . Here, we train a separate classifier to model  $P(M|Q)$  for each of the 1000 labels in the ImageNet dataset. Each classifier is trained on the aggregate of 50 images per label along with all 15 corruptions at 5 severity levels (3751 total per label). We again use an 80/20 train/test split and perform 5-fold cross-validation. The results shown in Figure 20b are an average of the  $AUC$  over all labels and folds.

We find here that across all IQ metrics evaluated,  $mAUC$  is barely above chance. For the traditional NR-IQA metrics, this supports our analysis and main experiments which show little correlation between  $M, Q$ . For the ZSCLIP-IQA metric, we refer again to the causal model (Fig. 7) and see that while the task selection variable ensures the association between  $M, Q$ , it is the conditioning on  $Y$  (and  $X$ ), as done here, that blocks all paths between  $M, Q$  and once again removes the association.

## H Predictability of DNN performance for mildly corrupted datasets

To show that **D1** is satisfied even in the case of mildly corrupted data, we plot the distributions of  $Q$  in Figure 21. Across all variants, even in cases where the likelihood is low, each IQA metric exhibits sensitivity to corruption (**D1**).

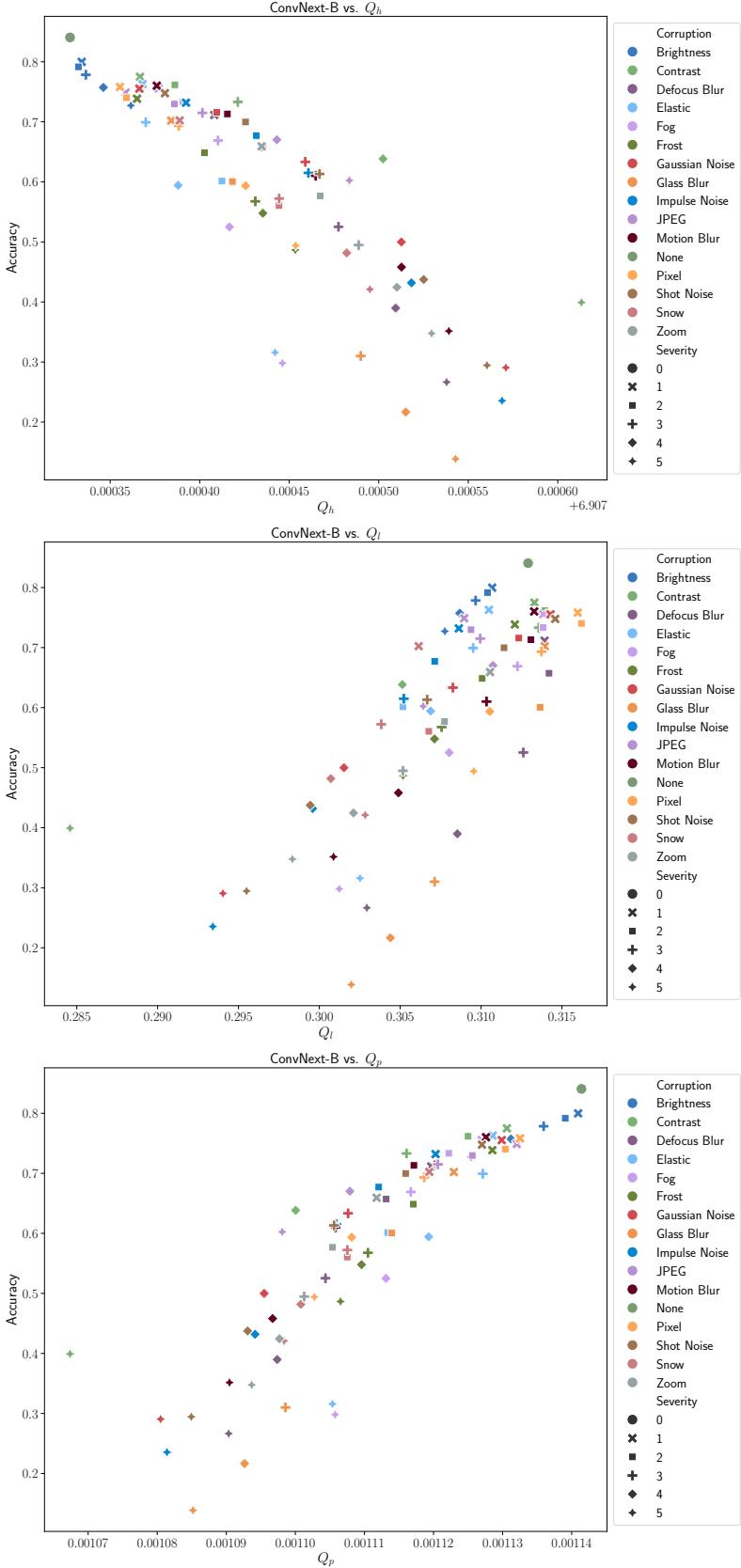


Figure 17: Comparison of ConvNext-B accuracy with (top to bottom)  $Q_h$ ,  $Q_l$ ,  $Q_p$  based on ZSCLIP-IQA. High correlation is observed between each ZSCLIP-IQA variant and accuracy.

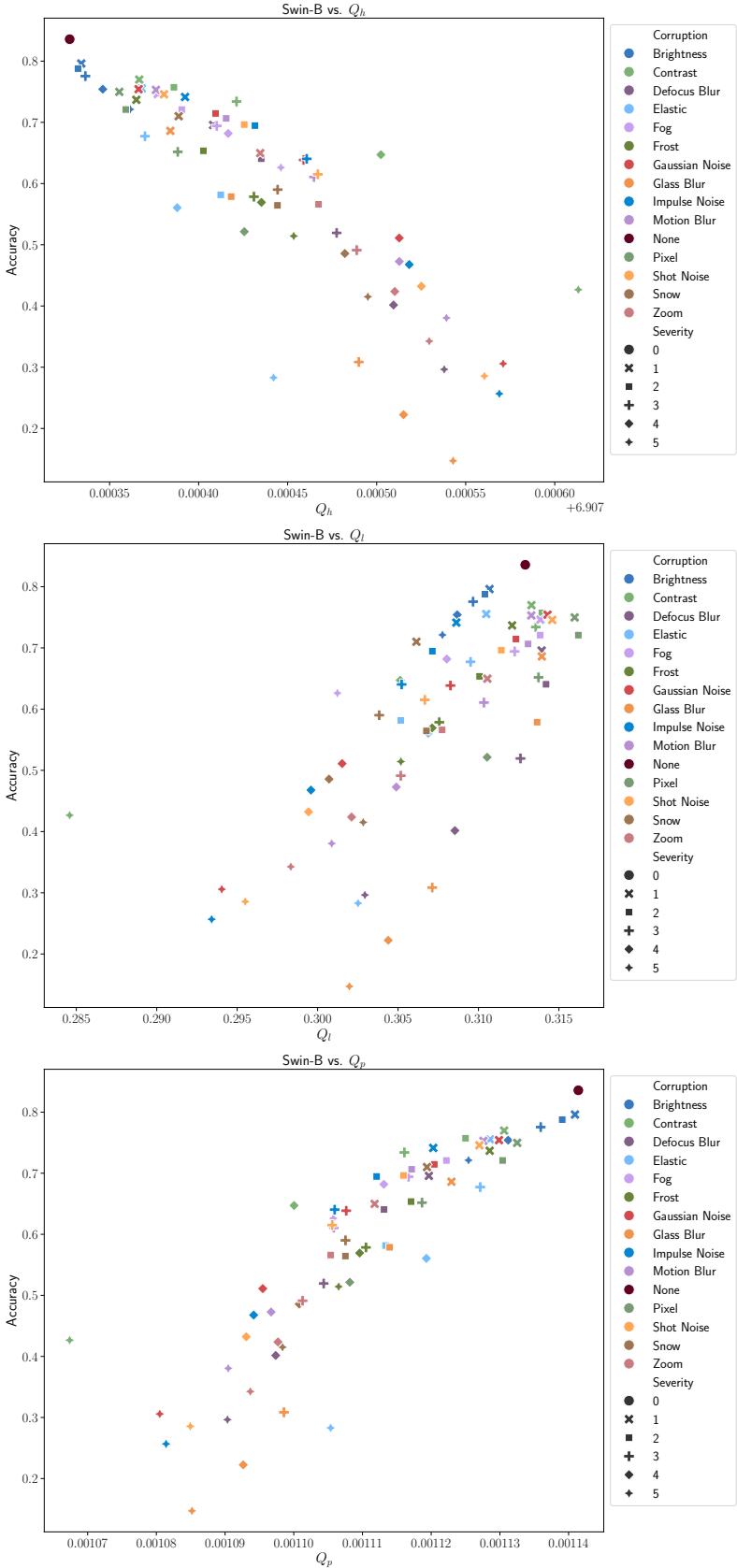


Figure 18: Comparison of Swin-B accuracy with (top to bottom)  $Q_h$ ,  $Q_l$ ,  $Q_p$  based on ZSCLIP-IQA. Some correlation is observed between each ZSCLIP-IQA variant and accuracy.

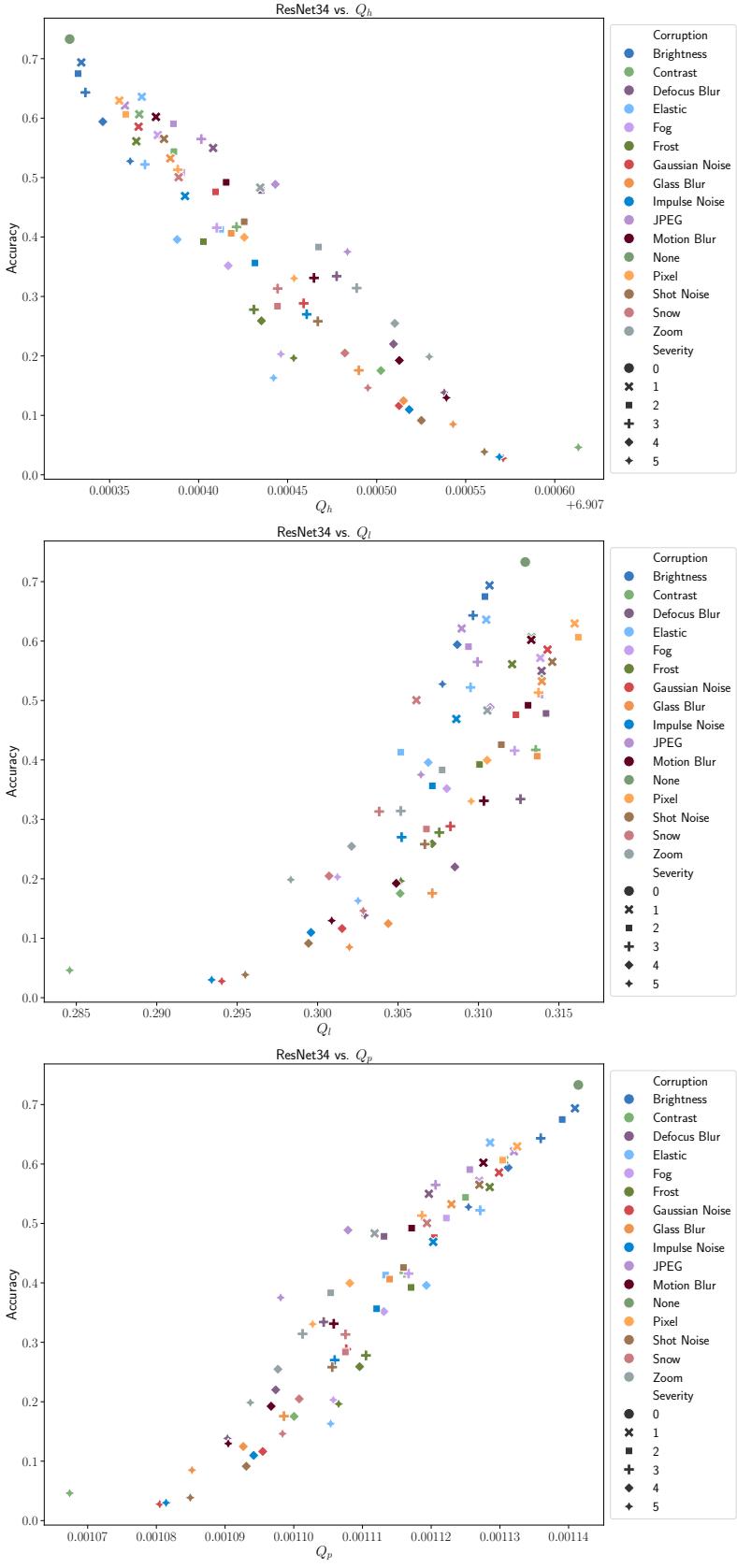


Figure 19: Comparison of ResNet34 accuracy with (top to bottom)  $Q_h$ ,  $Q_l$ ,  $Q_p$  based on ZSCLIP-IQA. High correlation is observed between each ZSCLIP-IQA variant and accuracy.

Table 4: **Strong Task-Guidance:** Correlation between IQ and accuracy for convolutional task DNN architectures. KRCC, SRCC, PLCC computed using average accuracy for each (corruption, severity). AUC and CE based on point-wise predictions (95% CI within  $\pm 0.001$ ). KRCC, SRCC, PLCC values have  $p < 0.05$ .

Model	IQA Metric	AUC $\uparrow$	CE $\downarrow$	$ PLCC  \uparrow$	$ SRCC  \uparrow$
ConvNext-B	ConvNext-B $Q_h$	0.772	0.562	0.660 $\pm$ 0.070	0.822 $\pm$ 0.070
	ConvNext-B $Q_l$	0.778	0.555	0.660 $\pm$ 0.067	0.826 $\pm$ 0.067
	ConvNext-B $Q_p$	0.826	0.504	0.738 $\pm$ 0.045	0.888 $\pm$ 0.045
	EfficientNet-V2-M $Q_h$	0.805	0.529	0.855 $\pm$ 0.030	0.945 $\pm$ 0.030
	EfficientNet-V2-M $Q_l$	0.801	0.538	0.846 $\pm$ 0.038	0.927 $\pm$ 0.038
	EfficientNet-V2-M $Q_p$	0.831	0.500	0.867 $\pm$ 0.025	0.956 $\pm$ 0.025
	MobileNet-V3-L $Q_h$	0.746	0.581	0.847 $\pm$ 0.028	0.927 $\pm$ 0.028
	MobileNet-V3-L $Q_l$	0.743	0.586	0.840 $\pm$ 0.039	0.909 $\pm$ 0.039
	MobileNet-V3-L $Q_p$	0.751	0.575	0.858 $\pm$ 0.025	0.929 $\pm$ 0.025
	ResNet34 $Q_h$	0.725	0.601	0.770 $\pm$ 0.069	0.859 $\pm$ 0.069
	ResNet34 $Q_l$	0.717	0.603	0.776 $\pm$ 0.051	0.866 $\pm$ 0.051
	ResNet34 $Q_p$	0.719	0.601	0.775 $\pm$ 0.051	0.875 $\pm$ 0.051
	Swin-B $Q_h$	0.760	0.579	0.601 $\pm$ 0.165	0.624 $\pm$ 0.165
	Swin-B $Q_l$	0.724	0.601	0.554 $\pm$ 0.166	0.604 $\pm$ 0.166
	Swin-B $Q_p$	0.791	0.547	0.672 $\pm$ 0.132	0.742 $\pm$ 0.132
ResNet34	ViT-L16 $Q_h$	0.781	0.562	0.615 $\pm$ 0.117	0.716 $\pm$ 0.117
	ViT-L16 $Q_l$	0.757	0.577	0.571 $\pm$ 0.125	0.696 $\pm$ 0.125
	ViT-L16 $Q_p$	0.800	0.540	0.671 $\pm$ 0.093	0.786 $\pm$ 0.093
	ConvNext-B $Q_h$	0.767	0.557	0.720 $\pm$ 0.060	0.858 $\pm$ 0.060
	ConvNext-B $Q_l$	0.760	0.563	0.733 $\pm$ 0.059	0.853 $\pm$ 0.059
	ConvNext-B $Q_p$	0.801	0.522	0.785 $\pm$ 0.044	0.904 $\pm$ 0.044
	EfficientNet-V2-M $Q_h$	0.796	0.527	0.786 $\pm$ 0.035	0.903 $\pm$ 0.035
	EfficientNet-V2-M $Q_l$	0.784	0.539	0.785 $\pm$ 0.041	0.877 $\pm$ 0.041
	EfficientNet-V2-M $Q_p$	0.814	0.508	0.789 $\pm$ 0.031	0.913 $\pm$ 0.031
	MobileNet-V3-L $Q_h$	0.814	0.506	0.864 $\pm$ 0.015	0.958 $\pm$ 0.015
	MobileNet-V3-L $Q_l$	0.807	0.515	0.858 $\pm$ 0.019	0.937 $\pm$ 0.019
	MobileNet-V3-L $Q_p$	0.823	0.495	0.872 $\pm$ 0.016	0.970 $\pm$ 0.016
	ResNet34 $Q_h$	0.848	0.470	0.862 $\pm$ 0.028	0.930 $\pm$ 0.028
	ResNet34 $Q_l$	0.827	0.492	0.870 $\pm$ 0.015	0.951 $\pm$ 0.015
	ResNet34 $Q_p$	0.850	0.461	0.886 $\pm$ 0.015	0.960 $\pm$ 0.015
	Swin-B $Q_h$	0.751	0.574	0.636 $\pm$ 0.153	0.643 $\pm$ 0.153
	Swin-B $Q_l$	0.709	0.600	0.624 $\pm$ 0.146	0.628 $\pm$ 0.146
	Swin-B $Q_p$	0.774	0.551	0.673 $\pm$ 0.129	0.747 $\pm$ 0.129
	ViT-L16 $Q_h$	0.782	0.546	0.687 $\pm$ 0.102	0.771 $\pm$ 0.102
	ViT-L16 $Q_l$	0.751	0.567	0.655 $\pm$ 0.100	0.757 $\pm$ 0.100
	ViT-L16 $Q_p$	0.794	0.534	0.721 $\pm$ 0.081	0.825 $\pm$ 0.081

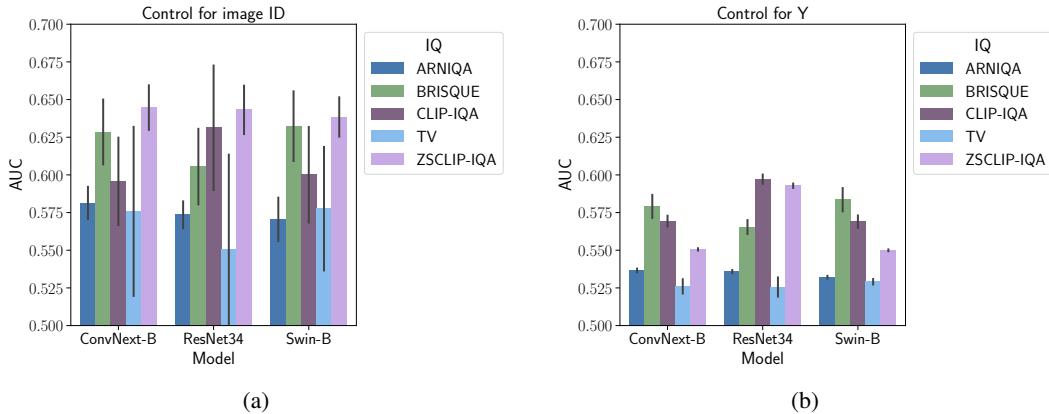


Figure 20: Mean AUC (mAUC) for classifiers trained (a) per image ID ( $X_i$ ) to model  $P(M|Q, X_i)$  and (b) per label  $Y$  to model  $P(M|Q, Y)$ . Averages are taken over all images/labels respectively and cross-validation folds with error bars indicating one standard deviation. Higher variance in (a) is attributed to lower sample sizes for training the logistic regression classifier.

While some IQA metrics are more sensitive to the overall image corruption, this does not necessarily translate to higher predictability. In fact, ZSCLIP-IQA appears to have smaller differences in IQA distribution across variants compared to other IQA metrics, yet the highest predictability of  $M$ . Figure 22 shows the predictability of  $M$  from  $Q$  for variants of the IN-C benchmark created as described in §2.5.

**Table 5: Strong Task-Guidance:** Correlation between IQ and accuracy for efficient task DNN architectures. KRCC, SRCC, PLCC computed using average accuracy for each (corruption, severity). AUC and CE based on point-wise predictions (95% CI within  $\pm 0.001$ ). KRCC, SRCC, PLCC values have  $p < 0.05$ .

Model	IQA Metric	AUC $\uparrow$	CE $\downarrow$	PLCC   $\uparrow$	SRCC   $\uparrow$
EfficientNet-V2-M	ConvNext-B $Q_h$	0.753	0.578	0.604 $\pm$ 0.102	0.742 $\pm$ 0.102
	ConvNext-B $Q_l$	0.753	0.578	0.601 $\pm$ 0.101	0.740 $\pm$ 0.101
	ConvNext-B $Q_p$	0.798	0.534	0.679 $\pm$ 0.075	0.820 $\pm$ 0.075
	EfficientNet-V2-M $Q_h$	0.831	0.497	0.888 $\pm$ 0.020	0.956 $\pm$ 0.020
	EfficientNet-V2-M $Q_l$	0.831	0.496	0.876 $\pm$ 0.029	0.937 $\pm$ 0.029
	EfficientNet-V2-M $Q_p$	0.862	0.456	0.900 $\pm$ 0.014	0.968 $\pm$ 0.014
	MobileNet-V3-L $Q_h$	0.749	0.578	0.827 $\pm$ 0.035	0.905 $\pm$ 0.035
	MobileNet-V3-L $Q_l$	0.745	0.583	0.819 $\pm$ 0.045	0.885 $\pm$ 0.045
	MobileNet-V3-L $Q_p$	0.752	0.573	0.827 $\pm$ 0.034	0.908 $\pm$ 0.034
	ResNet34 $Q_h$	0.731	0.595	0.766 $\pm$ 0.066	0.865 $\pm$ 0.066
	ResNet34 $Q_l$	0.724	0.597	0.762 $\pm$ 0.051	0.864 $\pm$ 0.051
	ResNet34 $Q_p$	0.723	0.597	0.761 $\pm$ 0.052	0.873 $\pm$ 0.052
	Swin-B $Q_h$	0.748	0.589	0.549 $\pm$ 0.189	0.533 $\pm$ 0.189
	Swin-B $Q_l$	0.711	0.609	0.503 $\pm$ 0.184	0.508 $\pm$ 0.184
	Swin-B $Q_p$	0.778	0.558	0.621 $\pm$ 0.162	0.663 $\pm$ 0.162
MobileNet-V3-L	ViT-L16 $Q_h$	0.760	0.579	0.565 $\pm$ 0.151	0.607 $\pm$ 0.151
	ViT-L16 $Q_l$	0.735	0.593	0.521 $\pm$ 0.155	0.582 $\pm$ 0.155
	ViT-L16 $Q_p$	0.779	0.560	0.620 $\pm$ 0.134	0.684 $\pm$ 0.134
	ConvNext-B $Q_h$	0.778	0.561	0.678 $\pm$ 0.074	0.842 $\pm$ 0.074
	ConvNext-B $Q_l$	0.773	0.567	0.679 $\pm$ 0.072	0.840 $\pm$ 0.072
	ConvNext-B $Q_p$	0.815	0.520	0.749 $\pm$ 0.051	0.901 $\pm$ 0.051
	EfficientNet-V2-M $Q_h$	0.815	0.521	0.836 $\pm$ 0.024	0.938 $\pm$ 0.024
	EfficientNet-V2-M $Q_l$	0.804	0.534	0.833 $\pm$ 0.032	0.913 $\pm$ 0.032
	EfficientNet-V2-M $Q_p$	0.834	0.497	0.847 $\pm$ 0.018	0.951 $\pm$ 0.018
	MobileNet-V3-L $Q_h$	0.833	0.496	0.906 $\pm$ 0.019	0.973 $\pm$ 0.019
	MobileNet-V3-L $Q_l$	0.835	0.496	0.895 $\pm$ 0.031	0.951 $\pm$ 0.031
	MobileNet-V3-L $Q_p$	0.853	0.470	0.922 $\pm$ 0.010	0.983 $\pm$ 0.010
	ResNet34 $Q_h$	0.783	0.558	0.778 $\pm$ 0.061	0.872 $\pm$ 0.061
	ResNet34 $Q_l$	0.769	0.563	0.783 $\pm$ 0.046	0.891 $\pm$ 0.046
	ResNet34 $Q_p$	0.778	0.555	0.787 $\pm$ 0.044	0.904 $\pm$ 0.044
	Swin-B $Q_h$	0.756	0.587	0.613 $\pm$ 0.180	0.587 $\pm$ 0.180
	Swin-B $Q_l$	0.714	0.614	0.583 $\pm$ 0.175	0.567 $\pm$ 0.175
	Swin-B $Q_p$	0.785	0.556	0.670 $\pm$ 0.149	0.710 $\pm$ 0.149
	ViT-L16 $Q_h$	0.788	0.557	0.655 $\pm$ 0.122	0.745 $\pm$ 0.122
	ViT-L16 $Q_l$	0.757	0.578	0.619 $\pm$ 0.121	0.731 $\pm$ 0.121
	ViT-L16 $Q_p$	0.804	0.539	0.706 $\pm$ 0.103	0.807 $\pm$ 0.103

Results show that weak task-guided IQA metrics are able to achieve higher  $AUC$  even when the number of corrupted images in the dataset is low. In comparison, conventional NR-IQA metrics achieve lower  $AUC$  and are more sensitive to the total level of corruption.

## I Compute resources

All experiments were run using a single NVIDIA A40 GPU with 48GB of memory. Predictability analysis can be conducted on CPU-only machine with at least 8 cores.

**Table 6: Strong Task-Guidance:** Correlation between IQ and accuracy for transformer task DNN architectures. KRCC, SRCC, PLCC computed using average accuracy for each (corruption, severity). AUC and CE based on point-wise predictions (95% CI within  $\pm 0.001$ ). KRCC, SRCC, PLCC values have  $p < 0.05$ .

Model	IQA Metric	AUC $\uparrow$	CE $\downarrow$		PLCC  $\uparrow$	SRCC  $\uparrow$
Swin-B	ConvNext-B $Q_h$	0.744	0.586		0.574 $\pm$ 0.129	0.706 $\pm$ 0.129
	ConvNext-B $Q_l$	0.746	0.586		0.573 $\pm$ 0.127	0.709 $\pm$ 0.127
	ConvNext-B $Q_p$	0.791	0.542		0.649 $\pm$ 0.102	0.788 $\pm$ 0.102
	EfficientNet-V2-M $Q_h$	0.797	0.537		0.835 $\pm$ 0.026	0.943 $\pm$ 0.026
	EfficientNet-V2-M $Q_l$	0.793	0.547		0.826 $\pm$ 0.037	0.923 $\pm$ 0.037
	EfficientNet-V2-M $Q_p$	0.822	0.510		0.853 $\pm$ 0.018	0.956 $\pm$ 0.018
	MobileNet-V3-L $Q_h$	0.742	0.583		0.812 $\pm$ 0.039	0.911 $\pm$ 0.039
	MobileNet-V3-L $Q_l$	0.739	0.589		0.804 $\pm$ 0.049	0.893 $\pm$ 0.049
	MobileNet-V3-L $Q_p$	0.748	0.576		0.814 $\pm$ 0.033	0.913 $\pm$ 0.033
	ResNet34 $Q_h$	0.722	0.603		0.724 $\pm$ 0.078	0.828 $\pm$ 0.078
	ResNet34 $Q_l$	0.713	0.604		0.721 $\pm$ 0.061	0.831 $\pm$ 0.061
	ResNet34 $Q_p$	0.716	0.602		0.727 $\pm$ 0.062	0.845 $\pm$ 0.062
	Swin-B $Q_h$	0.766	0.578		0.532 $\pm$ 0.207	0.483 $\pm$ 0.207
	Swin-B $Q_l$	0.732	0.597		0.485 $\pm$ 0.203	0.458 $\pm$ 0.203
	Swin-B $Q_p$	0.807	0.529		0.603 $\pm$ 0.184	0.620 $\pm$ 0.184
	ViT-L16 $Q_h$	0.768	0.576		0.542 $\pm$ 0.177	0.584 $\pm$ 0.177
	ViT-L16 $Q_l$	0.744	0.589		0.502 $\pm$ 0.176	0.563 $\pm$ 0.176
	ViT-L16 $Q_p$	0.789	0.551		0.596 $\pm$ 0.158	0.663 $\pm$ 0.158
ViT-L16	ConvNext-B $Q_h$	0.736	0.590		0.635 $\pm$ 0.086	0.780 $\pm$ 0.086
	ConvNext-B $Q_l$	0.739	0.591		0.642 $\pm$ 0.085	0.784 $\pm$ 0.085
	ConvNext-B $Q_p$	0.781	0.551		0.709 $\pm$ 0.065	0.850 $\pm$ 0.065
	EfficientNet-V2-M $Q_h$	0.781	0.553		0.804 $\pm$ 0.028	0.939 $\pm$ 0.028
	EfficientNet-V2-M $Q_l$	0.776	0.565		0.800 $\pm$ 0.037	0.920 $\pm$ 0.037
	EfficientNet-V2-M $Q_p$	0.803	0.530		0.815 $\pm$ 0.022	0.949 $\pm$ 0.022
	MobileNet-V3-L $Q_h$	0.732	0.589		0.807 $\pm$ 0.033	0.926 $\pm$ 0.033
	MobileNet-V3-L $Q_l$	0.729	0.595		0.800 $\pm$ 0.046	0.908 $\pm$ 0.046
	MobileNet-V3-L $Q_p$	0.740	0.580		0.813 $\pm$ 0.025	0.931 $\pm$ 0.025
	ResNet34 $Q_h$	0.718	0.604		0.772 $\pm$ 0.074	0.861 $\pm$ 0.074
	ResNet34 $Q_l$	0.707	0.607		0.771 $\pm$ 0.057	0.868 $\pm$ 0.057
	ResNet34 $Q_p$	0.713	0.602		0.777 $\pm$ 0.055	0.882 $\pm$ 0.055
	Swin-B $Q_h$	0.742	0.594		0.572 $\pm$ 0.164	0.558 $\pm$ 0.164
	Swin-B $Q_l$	0.706	0.613		0.533 $\pm$ 0.163	0.535 $\pm$ 0.163
	Swin-B $Q_p$	0.774	0.562		0.630 $\pm$ 0.138	0.685 $\pm$ 0.138
	ViT-L16 $Q_h$	0.799	0.545		0.633 $\pm$ 0.124	0.690 $\pm$ 0.124
	ViT-L16 $Q_l$	0.780	0.556		0.592 $\pm$ 0.124	0.672 $\pm$ 0.124
	ViT-L16 $Q_p$	0.828	0.506		0.694 $\pm$ 0.104	0.765 $\pm$ 0.104

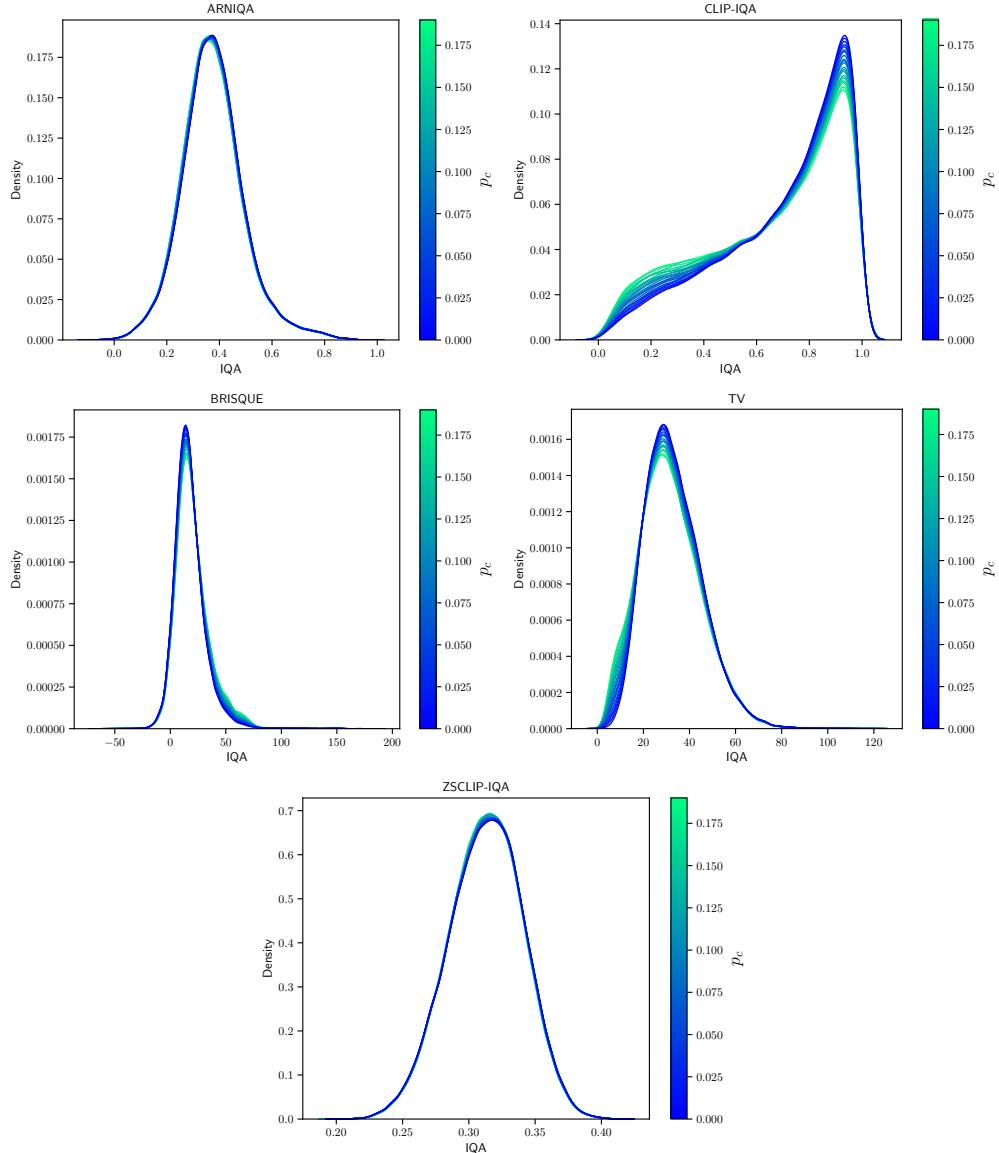


Figure 21: Distribution of IQA for each mildly corrupted variant of IN-val. Line color indicates the likelihood of image corruption  $p_c$  for each variant. Note that the amount of similarity/difference in the IQ distribution across variants does not explain the predictability which is determined by the causal DAG such as in §4, §2.3, and §2.4. See Figure 22 for predictability results.

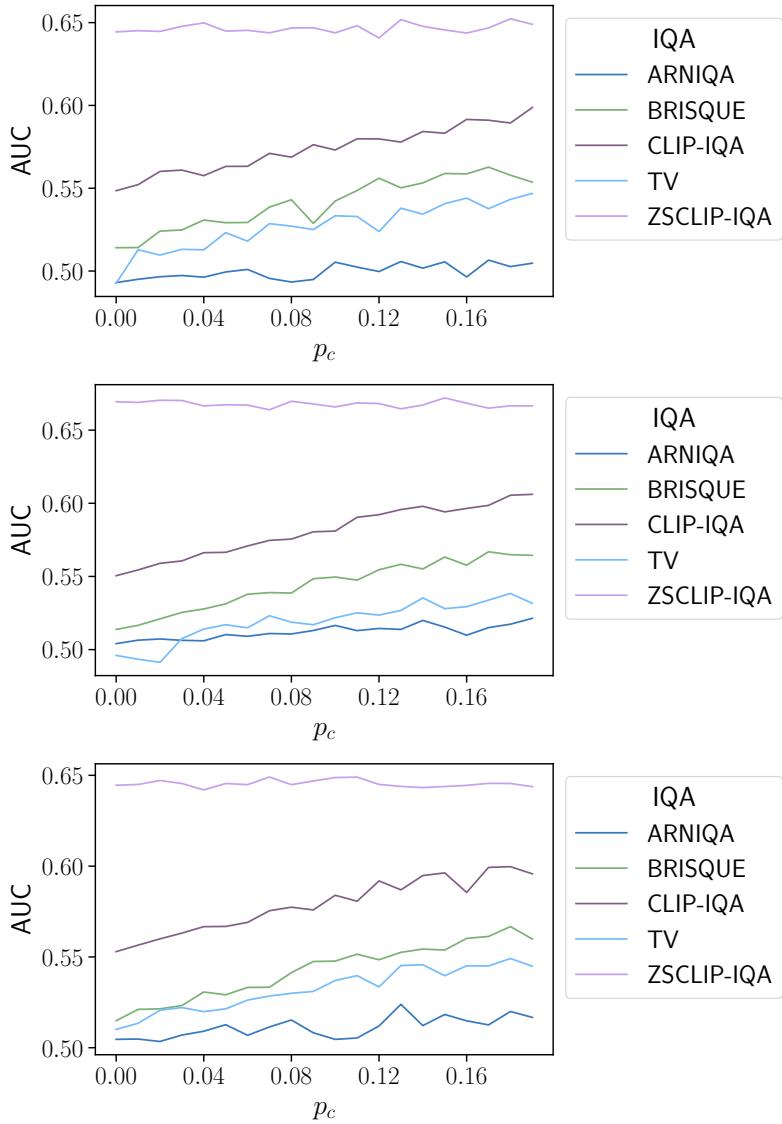


Figure 22: AUC vs.  $p_c$  where  $p_c$  represents the fraction of images in the test set that are mildly corrupted. Results are listed top to bottom: ConvNet-B, ResNet34, Swin-B. Predictability for ZSCLIP-IQA is relatively insensitive to the proportion of corrupted images whereas other metrics only improve as the proportion and diversity of corruptions increases.