
VISFIS: Visual Feature Importance Supervision with Right-for-the-Right-Reason Objectives

Zhuofan Ying,* Peter Hase,* and Mohit Bansal
Department of Computer Science
University of North Carolina at Chapel Hill
{zfyang, peter, mbansal}@cs.unc.edu

Abstract

Many past works aim to improve visual reasoning in models by supervising feature importance (estimated by model explanation techniques) with human annotations such as highlights of important image regions. However, recent work has shown that performance gains from feature importance (FI) supervision for Visual Question Answering (VQA) tasks persist even with *random* supervision, suggesting that these methods do not meaningfully align model FI with human FI. In this paper, we show that model FI supervision can meaningfully improve VQA model accuracy as well as performance on several Right-for-the-Right-Reason (RRR) metrics by optimizing for four key model objectives: (1) accurate predictions given limited but sufficient information (Sufficiency); (2) max-entropy predictions given no important information (Uncertainty); (3) invariance of predictions to changes in unimportant features (Invariance); and (4) alignment between model FI explanations and human FI explanations (Plausibility). Our best performing method, Visual Feature Importance Supervision (VISFIS), outperforms strong baselines on benchmark VQA datasets in terms of both in-distribution and out-of-distribution accuracy. While past work suggests that the mechanism for improved accuracy is through improved explanation plausibility, we show that this relationship depends crucially on explanation *faithfulness* (whether explanations truly represent the model’s internal reasoning). Predictions are more accurate when explanations are plausible *and* faithful, and not when they are plausible but not faithful. Lastly, we show that, surprisingly, RRR metrics are not predictive of out-of-distribution model accuracy when controlling for a model’s in-distribution accuracy, which calls into question the value of these metrics for evaluating model reasoning.²

1 Introduction

Many past works aim to teach models to ignore spurious features by making use of additional information about which features in an input are important [31, 46, 61]. For example, individual words can be annotated as (un)important in NLP tasks [46, 66], or regions of image pixels can be highlighted by humans as extra supervision for vision tasks [11, 51]. In this broad class of feature importance (FI) supervision methods, human annotations of important features are typically provided for individual datapoints, and methods often use data augmentation or gradient supervision to encourage models to rely only on important features when making predictions. Such approaches have seen performance improvements in image classification [50, 7], text classification [35, 66, 58], and multimodal visual question answering (VQA) tasks [47, 64, 36].

*Equal contribution.

²All supporting code for experiments in this paper is available at <https://github.com/zfyang/visfis>.

Visual Feature Importance Supervision

$$\text{VISFIS} = \mathcal{L}_{\text{Task}} + \mathcal{L}_{\text{Suff}} + \mathcal{L}_{\text{Unc}} + \mathcal{L}_{\text{Inv-FI}} + \mathcal{L}_{\text{Align}}$$

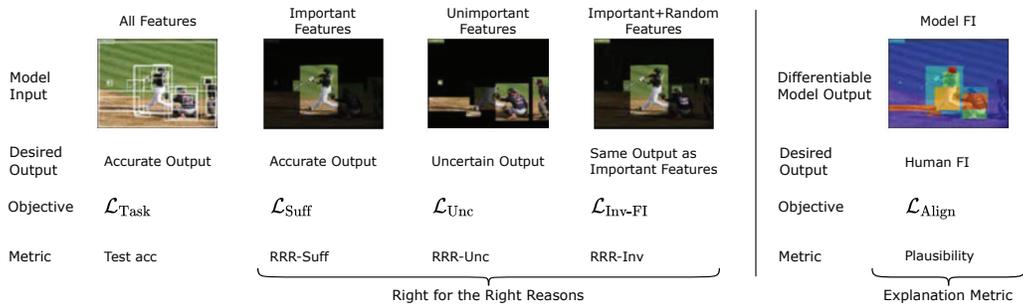


Figure 1: We depict five core desiderata for VQA models with associated metrics and objectives. We seek models that (1) are accurate given the full visual input, (2) are accurate given only important features, (3) are appropriately uncertain given only unimportant features, (4) are invariant to unimportant features, and (5) yield FI estimates (explanations) that align with human FI. VISFIS combines the five objectives. Note objectives #2–#5 require additional FI annotations. Darkened image regions correspond to bounding box representations that have been Replaced (see Sec. 3).

One of the primary motivations behind these approaches is to improve task accuracy by making models “Right for the Right Reasons” [46]. While existing FI supervision methods for VQA models can improve accuracy [47, 64], recent work has cast doubt on whether human supervision is the source of these improvements. Specifically, Shrestha et al. [48] show that VQA improvements persist even when *random image FI annotations* are used as supervision, suggesting that existing approaches may not extract meaningful signal from the human annotations. Motivated by this shortcoming, we explore several aspects of the FI supervision problem in the context of VQA tasks:

Improving VQA accuracy with FI supervision via four key model objectives (Sec. 7.1). Past VQA methods focus on data augmentation techniques [36] or ways to directly supervise model feature importance [47, 64]. We make use of four key objectives (represented in Fig. 1): (1) a Sufficiency objective encouraging the model to predict the correct label given only important input features [7]; (2) an Uncertainty objective encouraging max-entropy outputs when given only unimportant features; (3) an Invariance objective encouraging model outputs to be invariant to changes in unimportant features [46]; and (4) an Align objective that penalizes the model when its FI estimates differ from human FI annotations [47]. Our best performing method, termed Visual Feature Importance Supervision (VISFIS), combines these strategies in a novel manner, improving both in-distribution and out-of-distribution accuracy. Following guidelines from Shrestha et al. [48], we show that this improvement does not occur with random supervision, meaning VISFIS learns from human supervision itself. Lastly, after analyzing how explanation plausibility and faithfulness [23] relate to accuracy at the datapoint level, we suggest that FI supervision improves prediction accuracy by improving the plausibility of *faithful* FI explanations, rather than plausibility alone as past work has suggested [47, 7].

Evaluating models on new Right for the Right Reason (RRR) metrics (Sec. 7.4). Beyond measuring model accuracy, past works evaluate models on a few Right for the Right Reason metrics in order to understand whether model reasoning is correct [47, 48, 65, 7, 19, 16]. Correct reasoning is valuable because it suggests that models will generalize to test data that we might not be able to verify their performance on, which can occur e.g. when there are exponentially many cases we wish to test or when such data is prohibitively expensive to collect. We propose a broad set of RRR metrics for model evaluation, with similar motivation to our key model objectives above (see Fig. 1). Specifically, in addition to measuring existing metrics for (1) in-distribution accuracy, (2) out-of-distribution accuracy [1, 10, 6, 8, 7], and (3) model accuracy on sufficient feature subsets (RRR-Suff) [7], we also evaluate (4) model uncertainty given uninformative inputs (RRR-Unc), and (5) model invariance to changes in unimportant features (RRR-Inv). These metrics help verify that models can: arrive at correct answers relying only on features that are actually important (metric #3), are invariant to the addition or removal of unimportant features that should not affect the label (metric #4), and are appropriately uncertain about the model class when the input contains no meaningful evidence for any class (metric #5).

Predicting model generalization to OOD data with RRR metrics (Sec. 7.5). The practical value of the above RRR metrics can be considered in terms of their ability to inform us about model performance on data that we are not able to test on. We simulate this situation by evaluating models on in-distribution (ID) data and predicting whether the models will generalize to OOD data based on their accuracy and RRR metrics for ID data. Surprisingly, we find that both existing RRR metrics and our new ones do not better predict OOD accuracy than ID accuracy does on its own. This finding suggests that these metrics may not be a good evaluation of model reasoning, and that there is no good replacement yet for evaluating model accuracy on OOD data in addition to ID data.

In summary of our contributions, we show that:

1. FI supervision can improve both ID and OOD model accuracy on several benchmark VQA datasets. In particular, VISFIS improves over unsupervised baselines and the previous state-of-the-art on CLEVR-XAI by up to 4.7 points on OOD data.
2. Explanation plausibility correlates with model accuracy only when explanations are also faithful, which sheds light on the mechanism by which FI supervision improves model accuracy.
3. FI supervision improves model performance on several RRR metrics, including new invariance and uncertainty metrics.
4. RRR metrics do *not* correlate better with OOD accuracy than ID model accuracy does on its own. Consequently, RRR metrics may not be as valuable as previously thought.

2 Related Work

Supervising FI explanations. Past works primarily supervise gradient-based [46] or attention-based model explanations [66, 52, 17, 9]. For example, Ross et al. [46] enforce an ℓ_2 norm on the gradient of the loss w.r.t. the model input for features marked as unimportant by a human FI explanation. This method appears in several later works [50, 18, 51]. In a VQA setting, Selvaraju et al. [47] and Wu and Mooney [64] align the entire input gradient with human FI. In addition to using input gradients (termed Vanilla Gradient), we consider the Expected Gradients method [15], a computationally efficient implementation of Integrated Gradients [54]. Omission or perturbation-based approaches have seen more limited use. Kennedy et al. [28] regularize omission-based FI toward 0 for group identifiers in hate speech detection. In addition to simple leave-one-out [33] and keep-one-in methods, we propose a differentiable version of the popular linear method, SHAP [37]. For a survey of methods we refer readers to Friedrich et al. [16]. Following the analysis of Shrestha et al. [48], we use a random supervision baseline to show that VISFIS succeeds by virtue of additional supervision and not simply via model regularization.

Supervised data augmentation. This line of work uses human explanations to guide data augmentation, sometimes in a human-in-the-loop manner. For instance, Teney et al. [58] present LIME explanations to people and solicit feedback that is converted into counterexamples for model training. Liang et al. [34] use expert natural language counterfactual explanations to manufacture new labeled inputs.

We build on prior work for our data augmentation objectives. Similar to Chang et al. [7] and concurrent work from Singla et al. [51], our Sufficiency objective encourages models to be accurate given inputs with sufficient features selected according to human FI. Our Uncertainty objective reduces model confidence when no important features are provided, while the most related objectives from Chang et al. [7] and Liu et al. [36] encourage a *different* answer rather than an uncertain output. We are not aware of objectives encouraging invariance to changes in unimportant features as our invariance objective does. Other concurrent work encourages models to always predict the true label even when unimportant features are swapped with other unimportant features from the data distribution [19].

Right for the Right Reason metrics. Only a few past works explicitly evaluate RRR metrics in addition to test set accuracy. Our metrics include the existing RRR-Suff metric [7] and our RRR-Inv and RRR-Unc metrics. While explanation plausibility is regularly proposed as an RRR metric [47, 48, 65, 43, 16], we show that the relationship between accuracy and plausibility is controlled by explanation faithfulness, meaning that plausibility on its own should not be an RRR metric. Recently, Joshi et al. [27] propose a number of distinct distribution shifts in text classification for evaluating FI supervision techniques according to model OOD accuracy. We use a “changing prior” distribution shift standardly used in past work for VQA [11, 59], and moreover, the focus of our work is on novel RRR objectives, metrics, and analysis of how supervision improves models.

New machine learning metrics are often justified from first principles or on the basis of a strong correlation with some other good metric, like a human rating [41]. Here, we assess RRR metrics on the basis of their correlation with OOD accuracy, while controlling for model ID accuracy. This means that we measure the correlation between ID and OOD accuracy like in past studies [57, 39, 38, 60], but we also consider RRR metrics as additional explanatory variables for OOD generalization.

3 Terminology and Notation

FI Explanations. We distinguish between a human explanation e and a model explanation \tilde{e} that is obtained algorithmically to explain how a model arrived at its prediction for some datapoint. In this paper, human explanations are real-valued annotations for input features (which are bounding box representations for each of our datasets). The score for each bounding box is an indication of its importance to determining the datapoint label, which could roughly be thought of as an answer to the question, “why did data point x receive label y ” [40]. For several objectives and metrics, we binarize the explanations, selecting a threshold based on the data distribution (see Appendix E).

Replace Functions. Both generating and evaluating model explanations often require “hiding” input features from a model. In practice, we must replace features with some baseline value [53]. One simple and common way to replace features is to use an all-zeros feature [32, 54, 4]. We compare among several Replace functions to find the best function for learning from FI supervision (see Appendix B). We ultimately select the All-Negative-Ones function, which replaces a bounding box feature vector with the all negative ones vector, $\{-1\}^d$. We use x_e to denote a version of the input x where features where e is 0 are replaced via our Replace function.

Model Notation. We parametrize a distribution $p(y|x) = f_\theta(x)$ for classification purposes. Here, $\hat{y} = \arg \max_y f_\theta(x)_y$ is the model prediction, $f_\theta(x)_{\hat{y}}$ is the predicted probability, and \mathcal{Y} is the space of eligible answers, which is a large set that is shared across all questions in our VQA tasks.

4 Methods for Learning from Human Feature Importance Supervision

We now describe how to optimize for several key model desiderata using human FI supervision (represented in Fig. 1). In Sec. 4.5, we give the overall objective for Visual Feature Importance Supervision (VISFIS), which combines the objective terms below in order to improve model generalization.

4.1 Accuracy Given Sufficient Information

Goal. Like Chang et al. [7], we hope for image processing models to make accurate predictions given subsets of image features that are sufficient for arriving at the correct label, since this suggests that a model recognizes that the important features are in fact important.

Method. Access to human explanations should enable us to automatically construct sufficient inputs with some amount of unimportant information removed [7]. In particular, for every input we can create another datapoint by using the human explanation e to Replace unimportant features, while keeping the same label. The corresponding objective is given as:

$$\mathcal{L}_{\text{Suff}}(\theta, x, y, e) = \text{CrossEnt}(f_\theta(x_e), y). \quad (1)$$

This objective differs from previous instantiations [7] by virtue of the Replace function used (see Sec. 4.5). We compare $\mathcal{L}_{\text{Suff}}$ against an unsupervised baseline using random feature subsets. That is, a random distribution \mathcal{D}_s specifies how likely it is that we Replace a feature:

$$\mathcal{L}_{\text{Suff-Random}}(\theta, x, y, \mathcal{D}_s) = \mathbb{E}_{s \sim \mathcal{D}_s} \text{CrossEnt}(f_\theta(x_s), y) \quad (2)$$

which, when training via SGD, is estimated using one sample per datapoint per batch.

4.2 Uncertainty Given Only Unimportant Information

Goal. We would prefer for a model to give uncertain outputs for inputs with no important features, meaning the model should give a near-uniform distribution over classes. Since there is no evidence for any given class, the model should not be confident the input belongs in a particular class.

Method. With this goal in mind, Chang et al. [7] train models to give less confident outputs for images with important foreground features removed. More specifically, they encourage the model to predict any class *except* the image’s true class. In contrast, we penalize a KL divergence between the model output distribution and a uniform distribution,

$$\mathcal{L}_{\text{Unc}}(\theta, x, e) = \text{KL}(\text{Unif}(|\mathcal{Y}|), f_{\theta}(x_u)) \quad (3)$$

where $\text{Unif}(|\mathcal{Y}|)$ is the uniform distribution and $u = 1 - e$ indicates unimportant features.

4.3 Invariance to Unimportant Information

Goal. We would like models to be invariant to changes in an image’s unimportant features. This property is desirable because it means that a model correctly treats unimportant features as unimportant.

Method. We first describe a simple data augmentation approach, then describe an FI supervision approach similar to past work [50, 7].

In a data augmentation approach, we train a model to produce the same outputs for two inputs that share the same important information while differing in the unimportant information they contain. Specifically, we use e to obtain an input with both important and unimportant features, denoted by $x_{e \cup u} = \text{Replace}(x, e \cup u)$. Then, we penalize the KL divergence between the output distributions on the two inputs x_e and $x_{e \cup u}$. The resulting objective is then:

$$\mathcal{L}_{\text{Inv-DA}}(\theta, x, e, \mathcal{D}_u) = \mathbb{E}_{u \sim \mathcal{D}_u} \text{KL}(f_{\theta}(x_e), f_{\theta}(x_{e \cup u})) \quad (4)$$

where \mathcal{D}_u is a distribution over binary vectors. $\mathcal{L}_{\text{Inv-DA}}$ is estimated with one sample like $\mathcal{L}_{\text{Suff-Random}}$.

In an FI supervision approach, we first obtain model explanations at the datapoint level as $\tilde{e} = \text{Explain}(f_{\theta}, x, \hat{y})$, where Explain is a differentiable explanation method (possible methods described below). Then we seek to directly penalize models for treating unimportant features as important. To do so, we encourage FI scores for unimportant features to be 0:

$$\mathcal{L}_{\text{Inv-FI}}(\tilde{e}_u) = \|\tilde{e}_u\|_1 \quad (5)$$

where \tilde{e}_u is the subset of the explanation over features marked as unimportant by e . Past work uses an ℓ_2 distance for this objective [50, 7], while we use an ℓ_1 penalty after normalizing explanations to unit length, since explanations from different FI methods have different scales.

We consider a few options for differentiable explanation methods. Past work has primarily used gradient-based [46, 47, 50, 18, 64, 7] and attention-based explanations [66, 52, 17, 9]. We adopt existing gradient/attention methods and provide new differentiable perturbation-based methods.

1. *Gradient-based explanations.* One can optimize objectives involving gradient-based explanations w.r.t. θ by computing second derivatives like $\nabla_{\theta} \nabla_x f_{\theta}(x)$ in a framework like PyTorch [42]. We use a simple Vanilla Gradient method and the Expected Gradients method (see Appendix A).
2. *Attention-based explanations.* We supervise bounding box attention weights in the UpDn model [2], but early experiments suggest this is not an effective method and we do not explore it further.
3. *Perturbation-based explanations.* Perturbation-based methods like SHAP [37] are very popular explanation methods, but have seen only limited use for FI supervision [28]. We consider a leave-one-out method (LOO), a keep-one-in method (KOI), Average Effect, and SHAP (see Appendix A).

In Appendix D, we discuss limits on the compute budget used for each method during model training.

4.4 Aligned Model and Human Feature Importance

Goal. Alignment between human and model explanations has frequently been proposed as a goal for models [47, 48, 65, 16]. In general, past works assume that model explanations are *faithful*, meaning they accurately communicate a model’s internal reasoning [23]. This assumption is necessary for the alignment between model and human explanations, termed *plausibility* by Jacovi and Goldberg [23], to be evidence that model reasoning is similar to human reasoning. Of course, model explanations are not guaranteed to be faithful. To the extent that they are faithful, however, encouraging explanation plausibility during training may help align model reasoning with human reasoning.

Method. We first obtain model explanations at the datapoint level as $\tilde{e} = \text{Explain}(f_\theta, x, \hat{y})$ (see Sec. 4.3 above). Then, we can measure the difference between \tilde{e} and the human explanation e using an l_p distance, cosine similarity, or a differentiable ranking function [47, 64]. We use a cosine similarity since model explanations and human explanations do not share the same scale. Our objective is thus:

$$\mathcal{L}_{\text{align}}(\theta, x, e, \tilde{e}) = \text{cos-sim}(e, \tilde{e}) \quad (6)$$

4.5 Overall Objective for VISFIS: Visual Feature Importance Supervision

We combine the supervised objective terms from above to achieve the corresponding model desiderata simultaneously. Following objective tuning experiments showing that Inv-FI outperforms Inv-DA (see Appendix Table 11), we use Inv-FI rather than Inv-DA, and therefore our final VISFIS objective is:

$$\lambda_1 \mathcal{L}_{\text{Task}} + \lambda_2 \mathcal{L}_{\text{Suff}} + \lambda_3 \mathcal{L}_{\text{Unc}} + \lambda_4 \mathcal{L}_{\text{Align}} + \lambda_5 \mathcal{L}_{\text{Inv-FI}} \quad (7)$$

where $\mathcal{L}_{\text{Task}}$ is a standard supervised cross-entropy loss. Besides tuning the values of λ_i one at a time, we also tune the Replace function and FI method used in this objective, making sure to use comparable compute budgets across FI methods. Replace functions we consider are listed in Appendix B (results in Table 6), and FI methods in Appendix A (results in Tables 9 and 10). Following tuning, we find that it is preferable to Replace bounding box representations with the negative ones vector, $\{-1\}^d$, and surprisingly, we find that Vanilla Gradient is the best performing FI method, surpassing all perturbation-based methods as well as the Expected Gradients method.

5 Metrics

Next, we describe the RRR and explanation metrics for each of our model desiderata outlined above. We also measure model ID and OOD accuracy (distribution shifts described in Sec. 6). As with the model objectives, we use the All-Negative-Ones Replace function as needed.

RRR-Sufficiency. We measure model accuracy on inputs containing only features selected as important by their respective human explanation (similar to [7]). The remaining features are Replaced.

RRR-Uncertainty. We propose to measure how uncertain the model prediction is given only unimportant information. Specifically, we report the average model predicted probability when we provide only unimportant features to the model (according to the human explanation), so lower is better.

RRR-Invariance. We propose to calculate the agreement between model predictions with the input x_e and three $x_{e \cup u}$ that each include a random number of unimportant features. The final metric is averaged over three random u for each test point, then over all test points.

Explanation Plausibility. Our explanation plausibility metric is the Spearman’s rank correlation between the human and model feature importance vectors, similar to past work [11, 47]. We use continuous FI estimates in order to calculate the rank correlation. A rank correlation is preferable here because human and model FI explanations do not lie in the same space.

Explanation Faithfulness. We use two standard faithfulness metrics [13]. Sufficiency measures whether *keeping* important features (according to model explanation \tilde{e}) leads the model to *retain* its confidence in its original prediction: $\text{Suff}(f_\theta, x, \tilde{e}) = f_\theta(x)_{\hat{y}} - f_\theta(x_{\tilde{e}})_{\hat{y}}$. Comprehensiveness measures whether *removing* important features from an input leads to a *decline* in model confidence, $\text{Comp}(f_\theta, x, \tilde{e}) = f_\theta(x)_{\hat{y}} - f_\theta(x_{\tilde{e}^C})_{\hat{y}}$, where $x_{\tilde{e}^C} = \text{Replace}(x, 1 - \tilde{e})$ is the *complement* of features in $x_{\tilde{e}}$. We average these score over several sparsity levels of \tilde{e} , keeping or removing the top 10%, 25%, or 50% of features [13]. Note we compute these metrics using the best available explanation method per dataset, as measured by explanation faithfulness (comparison in Appendix G).

6 Experiment Setup

Datasets. We perform experiments on three benchmark datasets: CLEVR-XAI [5], GQA [21], and VQA-HAT [11]. CLEVR-XAI is an algorithmically generated dataset based on CLEVR [26] and provides ground truth visual segmentation masks for each question. CLEVR-XAI is limited in visual varieties and vocabularies, but it offers FI supervision in a controlled, low-noise setting. GQA

contains compositional reasoning questions over naturalistic images. GQA also includes the program for generating the questions and the ground-truth scene graph from the Visual Genome dataset [30]. This allows us to obtain bounding boxes of relevant objects identified through the question program, which we use as FI supervision. VQA-HAT is based on VQAv1 [3], including naturalistic images and questions with mouse tracking de-blurring used to collect image FI annotations from humans. For VQA, we report model performance on the more challenging *other* type questions as recommended by Teney et al. [59].

Distribution Shifts. We create both ID and OOD test sets for each dataset, so we always have four data splits: Train, Dev, Test-ID, and Test-OOD (split sizes shown in Table 1). To obtain OOD data, we use distribution shifts similar to those in VQA-CP, which are intended to vary the linguistic bias between ID and OOD

Table 1: Dataset split sizes.

Dataset	Train	Dev	Test-ID	Test-OOD
CLEVR-XAI	83k	14k	21k	22k
GQA-101k	101k	20k	20k	20k
VQA-HAT	36k	6k	9k	9k

splits [1]. We apply the same procedure for distribution shift on all three datasets for comparability. In detail, we create groups of questions according to the first few words in each question (indicating the type of question), and allocate groups unevenly into ID and OOD sets, randomly assigning 80% of each group to one set and 20% to the other. The ID set is split into Train, Dev, and Test-ID. Model selection is done according to Dev set performance. We further downsample the very large GQA dataset from to about 100k for training and 20k for other splits. See Appendix Fig. 5 for training size ablation analysis. We note that we avoid several pitfalls in evaluating VQA models against distribution shifts, as outlined by Teney et al. [59]. See Appendix Table 7 for sensitivity analysis with randomly resplit data.

Human Feature Importance. For all datasets, we obtain human FI scores at the bounding box (BB) level for detected BBs from the Faster-RCNN detector [45]. Following Selvaraju et al. [47], for both VQA-HAT and CLEVR-XAI we obtain importance scores from pixel-level annotations as $s^k = E_i^k / (E_i^k + E_o^k)$, where s^k is the score for the k^{th} detected BB and E_i^k and E_o^k are the average pixel-level importance score inside and outside the BB, respectively. VQA-HAT has real-valued pixel-level scores, while for CLEVR-XAI, we set the pixel-level score to 1 for pixels within the segmentation mask and 0 elsewhere. For GQA, since we have BB level annotations, we calculate the importance score based on the intersection over union (IoU) between ground-truth important BBs and detected BBs: $s^k = \max_{l \in \mathcal{G}} \text{IoU}(B_d^k, B_{\text{gt}}^l)$ where B_d^k is the BB of the k^{th} detected object and B_{gt}^l is the BB of the l -th ground-truth object. With importance scores for each BB, we manually set a threshold for determining important and unimportant objects (0.85, 0.55, and 0.3 for CLEVR-XAI, VQA-HAT, and GQA respectively). See Appendix E for sensitivity analysis for this threshold.

Models. We run experiments with UpDn [2] and LXMERT [56]. Both models rely on bounding box representations generated by a pretrained Faster R-CNN model [45] (further details in Appendix F).

Hypothesis Testing. We conduct hypothesis tests via a bootstrap resampling model seeds and datapoints 10k times [14]. We obtain 95% confidence intervals in the same way.

7 Experiment Results

7.1 Can FI Supervision Improve Model Accuracy for VQA?

Design. Using UpDn on our three datasets, we compare VISFIS with previous state-of-the-art FI supervision methods for VQA tasks [47, 64] as well as for image classification [50, 7]. We give results for LXMERT only on CLEVR-XAI, since GQA and VQA are a part of the pretraining data for LXMERT [56]. Note we test on the more challenging *other* type questions only for VQA, following Teney et al. [59]. Selvaraju et al. [47] use $\mathcal{L}_{\text{align}}$ with a ranking loss to align Vanilla Gradient explanations and human FI supervision. Wu and Mooney [64] propose a relaxed version of the ranking loss that binarizes important and unimportant features according to human FI supervision and encourages higher model FI for important objects than unimportant ones. The other methods we consider all use an $\mathcal{L}_{\text{FI-Inv}}$ objective with an l_2 penalty on Vanilla Gradient explanations. On top of this, Chang et al. [7] add an $\mathcal{L}_{\text{Suff}}$ objective with a Shuffle Replace function that randomly permutes features rather than replacing them, to preserve the marginal data distribution, and Singla et al. [51] add an $\mathcal{L}_{\text{Suff}}$ objective with a Gaussian noise Replace function. Our unsupervised baselines are models trained with only label supervision or using $\mathcal{L}_{\text{Suff-Random}}$.

Table 2: Test accuracy across FI supervision methods and datasets with an UpDn model. We bold/underline numbers higher than the best unsupervised baseline at a significance threshold of $p < .05$ (and bold is better than underline at $p < .05$).

Method	CLEVR-XAI		GQA-101k		VQA-HAT	
	ID	OOD	ID	OOD	ID	OOD
Baseline	71.37±0.57	36.80±1.00	51.82±0.62	31.80±0.64	37.53±1.32	28.76±1.10
Suff-Random	71.72±0.57	39.08±0.80	51.59±0.65	31.65±0.82	37.99±1.35	29.34±1.03
Selvaraju et al. [47]	71.32±0.58	37.96±1.00	51.38±0.62	31.99±0.77	36.93±1.37	27.38±1.27
Wu and Mooney [64]	71.48±0.64	37.31±0.86	51.54±0.67	31.61±0.78	37.24±1.32	28.26±1.15
Simpson et al. [50]	71.22±0.60	37.54±0.71	52.10±0.68	31.99±0.77	37.66±1.30	28.73±1.44
Chang et al. [7]	70.77±0.56	35.38±0.92	50.29±0.65	30.40±0.86	32.55±1.41	17.98±1.75
Singla et al. [51]	71.54±0.58	38.25±1.39	<u>52.42±0.66</u>	<u>32.58±0.59</u>	38.28±1.37	29.25±2.12
ViSFIS	72.82±0.56	43.78±1.11	54.81±0.61	34.88±0.80	38.75±1.35	31.21±1.28
w/ Rand. Supervis.	69.70±0.67	33.28±1.03	49.82±0.62	29.93±0.89	37.16±1.30	27.51±1.17

Results. We show results for UpDn in Table 2 and for LXMERT in Table 3. First, we find that FI supervision can meaningfully improve model accuracy. With UpDn on CLEVR-XAI, ViSFIS improves ID accuracy by 1.1 points (± 0.5 ; $p=1e-4$) and OOD accuracy by 4.7 points (± 1.4 ; $p<1e-4$) over the strongest baseline without supervision, Suff-Random (see Appendix Fig. 9 for breakdown in improvements by CLEVR question type). Trends are similarly positive on the other datasets and with LXMERT, where ViSFIS outperforms the baseline by 0.48 points (± 0.35 ; $p<.01$) on ID data and 1.07 points ($p<1e-4$) on OOD data (for results on all VQA question types, see Appendix Table 15). These improvements do not persist when using random explanations (last row), meaning they are caused by the human supervision. Finally, we observe that ViSFIS is the best overall method across datasets and architectures, as other methods typically do not improve accuracy over an unsupervised baseline. The next best method is that of Singla et al. [51], which improves over an unsupervised baseline only for the GQA dataset with UpDn, but ViSFIS still outperforms Singla et al. [51] there by 2.39 (± 0.55 ; $p<1e-4$) points on ID data and 2.31 points (± 0.66 ; $p<1e-4$) on OOD data.

Table 3: LXMERT + CLEVR-XAI results.

Method	ID Acc	OOD Acc
Baseline	86.91±0.43	73.76±0.72
Suff-Random	86.53±0.47	73.52±1.07
Selvaraju et al. [47]	87.03±0.43	74.56±0.58
Wu and Mooney [64]	86.73±0.46	73.97±0.75
Simpson et al. [50]	86.22±0.57	73.12±1.28
Chang et al. [7]	85.05±0.57	67.27±2.27
Singla et al. [51]	87.08±0.46	74.10±0.75
ViSFIS	87.39±0.45	74.83±0.70
w/ Rand. Supervis.	85.84±0.83	71.81±1.34

7.2 How Does FI Supervision Improve Accuracy?

Design. Past work hypothesizes that FI supervision improves accuracy by aligning model and human FI [47, 7]. Surprisingly, we find that the relationship between model test accuracy and average explanation plausibility is fairly weak (linear correlation on UpDn+CLEVR-XAI models is $\rho=0.14\pm 0.19$). Here, we argue that plausible explanations alone are not evidence of correct model reasoning, but plausible *and* faithful explanations are. Using 4 million ID/OOD test predictions from UpDn+CLEVR-XAI models, we visualize trendlines from logistic regressions predicting model accuracy based on plausibility and faithfulness at the datapoint level, grouped into Worst, Middle, and Best faithfulness categories based on Sufficiency/Comprehensiveness metrics (see Appendix E).

Results. Fig. 2 shows that as an explanation for a datapoint becomes more plausible, the model is more likely to correctly predict that point’s label, but *only when the explanation is also faithful*. Indeed, a maximally plausible and faithful explanation has about a 90% chance of being correct,

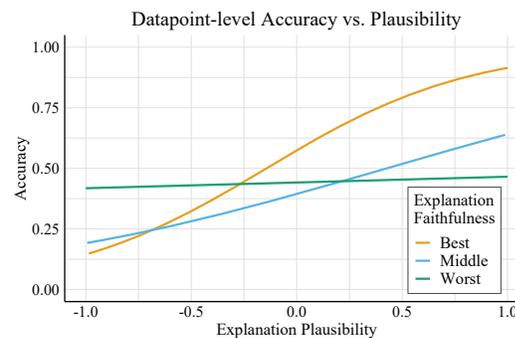


Figure 2: Datapoint-level accuracy by explanation plausibility, averaged across CLEVR-XAI models. Trendlines are logistic regressions. When explanations are more faithful, their alignment with human explanations better correlates with model accuracy.

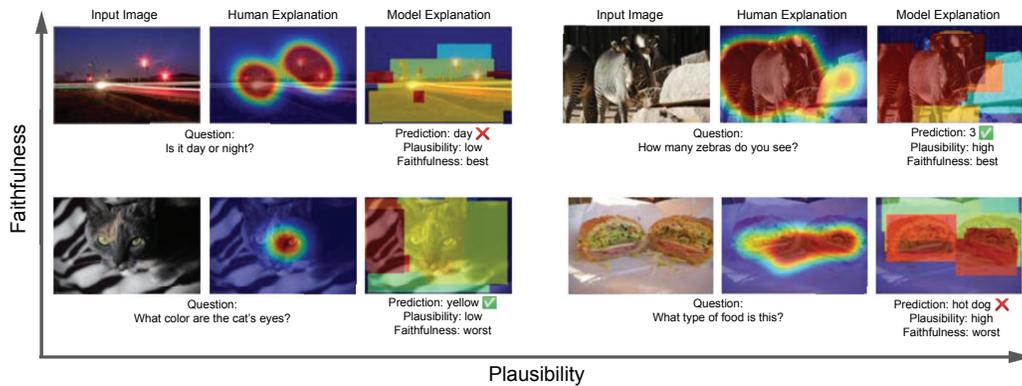


Figure 3: Qualitative visualization of the relationship between accuracy, plausibility, and faithfulness represented in Fig. 2. In a low faithfulness setting (in terms of explanation sufficiency), a data point with an implausible explanation can still have a correct prediction (bottom left), while a data point with highly plausible explanation can have an incorrect prediction (bottom right). Among higher faithfulness points (top row), data with more plausible explanations tend to be correctly predicted.

while a minimally plausible but highly faithful explanation has closer to a 12.5% of being correct. For unfaithful explanations, plausibility has essentially no relationship with accuracy. Though these trends are not necessarily causal, they are consistent with the view that when model predictions are correct, it is because their true reasoning (as revealed by *faithful* explanations) aligns with human reasoning. Fig. 3 qualitatively illustrates this relationship among faithfulness, plausibility, and accuracy with example data points and model predictions. We emphasize that while past work has treated plausibility as an RRR metric [47, 48, 65, 43, 16], the results here demonstrate that plausibility alone cannot be a measure of model correctness.

7.3 Which FI Supervision Objectives Improve Accuracy?

Design. We ablate across objective terms from Sec. 4 for UpDn on CLEVR-XAI. The weight for each objective term is tuned while using only that objective, then kept fixed when objectives are combined (further details in Appendix F). We consider another kind of ablation experiment where we use random supervision for one objective at a time in VISFIS, with results in Appendix Table 12.

Results. In Table 4, we find that each individual objective is valuable on its own, and they do well when combined. Relative to the Baseline OOD accuracy, Suff-Human adds 4.1 points, Unc adds 1.54 points, Inv-FI adds 2.08 points, and Align adds 4.81 points. When the four objectives are combined in VISFIS, the improvement rises to 6.98 points.

7.4 Can FI Supervision Make Models Right for the Right Reasons?

Design. We report RRR metrics as well as explanation plausibility for the UpDn+CLEVR-XAI models from our objective ablation above.

Results. In Table 4, we find that FI supervision generally improves RRR metric scores. Compared to the Baseline, VISFIS achieves 27.8 points better Sufficiency, 13.8 points better Invariance, and 11.5 points better Uncertainty. The best unsupervised method closes the gap slightly on RRR-Suff and RRR-Inv. Specifically, Suff-Random is only 3.97 points worse than Suff-Human on RRR-Suff, and only 0.53 points worse than Inv-DA on RRR-Inv. It suggests that FI supervision noticeably improves RRR metrics, meaning model behavior better fulfills the theoretical desiderata from Sec. 4.

Table 4: Objective ablation for UpDn + CLEVR-XAI

Objective	Acc		RRR Metrics			Expl.
	ID	OOD	Suff	Inv	Unc	Plaus.
Baseline	71.37	36.80	48.82	77.89	55.17	28.82
Inv-DA	71.17	35.91	72.53	93.12	76.29	14.33
Inv-FI	71.41	38.88	45.31	76.34	71.41	28.60
Unc	71.30	38.34	10.75	86.58	4.16	8.56
Align	72.04	41.61	61.19	79.51	64.22	37.20
Suff-Random	71.73	39.08	73.59	92.59	60.93	17.32
Suff-Human	71.87	40.91	76.94	90.82	81.42	16.27
+ Align	72.42	41.63	78.55	89.69	80.02	35.73
+ Unc	72.33	41.54	77.83	89.70	41.68	23.41
VISFIS	72.82	43.78	76.65	91.72	43.64	22.67

7.5 Do RRR Metrics Predict OOD Generalization?

Design. We measure the correlation between RRR metrics (calculated with ID data) and OOD accuracy across a large set of models. We report results here for all UpDn models on CLEVR-XAI, with similar results for GQA/VQA and LXMERT given in Appendix Table 16. We consider a few possible model metrics, including several composite metrics that combine model-level metrics. To optimally weight the individual metrics, we fit statistical models to the data that predict OOD accuracy given the available metrics. Since this risks overfitting the composite metrics to the data we have, we perform a cross-validation resampling model-level statistics 10k times, using 90 models’ metrics as training data and 10 for testing each time. The final metrics we consider are: (1) ID accuracy on its own as a baseline, (2) RRR metrics on their own, (3) ID accuracy plus average model confidence, (4) ID accuracy plus explanation metrics (for plausibility and faithfulness), (5) ID accuracy plus RRR metrics, and (6) All Metrics, which uses all available metrics.

Results. In Table 5, we show the average correlations between each metric and model OOD accuracy achieved in our cross-validation. Interestingly, we find that **RRR metrics do not achieve a better correlation with OOD accuracy than ID accuracy does on its own.** ID accuracy alone has a correlation of 0.863 with OOD accuracy, while using ID Acc + RRR metrics achieves a correlation of 0.852. In fact, the only additional metric that improves one’s ability to predict OOD accuracy is the average model confidence on ID data (more confident models have slightly better OOD accuracy), though this does not hold for LXMERT models (see Appendix Table 16). These results cast doubt on the value of RRR metrics. If ID accuracy on its own is a better predictor of OOD accuracy than RRR metrics, then RRR metrics may not be a better measure of the quality of model reasoning than ID accuracy is. We believe the RRR metrics considered in this paper are still theoretically justified as model desiderata, but we cannot recommend them as measures of model generalization to OOD data.

Table 5: Correlations between metrics and OOD accuracy, with 95% confidence intervals.

Metric	$\rho(\text{Metric, OOD Acc})$	
	Train	Test
RRR-Suff	0.278	0.333 \pm .0021
RRR-Inv	0.149	0.157 \pm .0058
RRR-Unc	0.029	0.021 \pm .0063
ID Acc	0.870	0.863 \pm .0018
+ Model Conf.	0.909	0.907 \pm .0010
+ Expl. Metrics	0.875	0.861 \pm .0046
+ RRR-all	0.874	0.852 \pm .0033
All Metrics	0.925	0.891 \pm .0014

8 Discussion & Conclusion

Limitations. Though we evaluate with two standard model architectures and three datasets, our conclusions may be limited to settings using Faster-RCNN [45] bounding box representations as the feature space rather than pixel space. Additionally, though we follow existing guidelines with our distribution shifts [11, 59], we do not measure model generalization across all typical kinds of shifts [44]. Lastly, we note that FI supervision methods are limited by the need for additional annotations.

Ethics. We hope that our findings regarding model accuracy and explanation plausibility/faithfulness will help dispel the notion that models reason like humans (or are more grounded) simply because model explanations look similar to human explanations, which can cause unwarranted trust in ML models [24]. We do not foresee specific ethical risks arising from this work that do not already apply to the general use of machine learning for visual question answering tasks, such as the potential deployment of ML models in settings where they may harm people [62, 55].

Conclusions. In this paper, we show that (1) FI supervision can improve VQA model accuracy via our VISFIS method, (2) accuracy improvements appear to stem from improving explanation plausibility specifically for faithfully explained data, (3) FI supervision can improve RRR metric performance, and (4) RRR metrics do not actually correlate well with OOD accuracy.

Acknowledgements

We thank Jaemin Cho for helpful discussion of this work, as well as Derek Tam, Xiang Zhou, and Archiki Prasad for useful feedback. This work was supported by ARO Award W911NF2110220, DARPA Machine-Commonsense (MCS) Grant N66001-19-2-4031, NSF-AI Engage Institute DRL-211263, and a Google PhD Fellowship. The views contained in this article are those of the authors and not of the funding agency.

References

- [1] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4971–4980, 2018. 2, 7
- [2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/papers/Anderson_Bottom-Up_and_Top-Down_CVPR_2018_paper.pdf. 5, 7, 17
- [3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 7
- [4] Leila Arras, Ahmed Osman, Klaus-Robert Müller, and Wojciech Samek. Evaluating recurrent neural network explanations. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 113–126, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4813. URL <https://www.aclweb.org/anthology/W19-4813>. 4
- [5] Leila Arras, Ahmed Osman, and Wojciech Samek. Clevr-xai: A benchmark dataset for the ground truth evaluation of neural network explanations. *Information Fusion*, 81:14–40, 2022. ISSN 1566-2535. doi: <https://doi.org/10.1016/j.inffus.2021.11.008>. URL <https://github.com/ahmedmagdiosman/clevr-xai>. 6, 19
- [6] Remi Cadene, Corentin Dancette, Matthieu Cord, Devi Parikh, et al. Rubi: Reducing unimodal biases for visual question answering. *Advances in neural information processing systems*, 32, 2019. 2
- [7] Chun-Hao Chang, George Alexandru Adam, and Anna Goldenberg. Towards robust classification model by counterfactual and invariant data generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15212–15221, 2021. URL <https://arxiv.org/pdf/2106.01127.pdf>. 1, 2, 3, 4, 5, 6, 7, 8, 26
- [8] Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. Counterfactual samples synthesizing for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10800–10809, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/papers/Chen_Counterfactual_Samples_Synthesizing_for_Robust_Visual_Question_Answering_CVPR_2020_paper.pdf. 2
- [9] George Chrysostomou and Nikolaos Aletras. Enjoy the salience: Towards better transformer-based faithful explanations with word salience. *arXiv preprint arXiv:2108.13759*, 2021. URL <https://arxiv.org/pdf/2108.13759.pdf>. 3, 5
- [10] Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. Don't take the easy way out: ensemble based methods for avoiding known dataset biases. *arXiv preprint arXiv:1909.03683*, 2019. 2
- [11] Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. Human attention in visual question answering: Do humans and deep networks look at the same regions? *Computer Vision and Image Understanding*, 163:90–100, 2017. URL <https://computing.ece.vt.edu/~abhshkdz/vqa-hat/>. 1, 3, 6, 10, 19
- [12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019. 21
- [13] Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. Eraser: A benchmark to evaluate rationalized nlp models. In *ACL 2020*, volume abs/1911.03429, 2020. URL <https://arxiv.org/pdf/1911.03429.pdf>. 6

- [14] Bradley Efron and Robert J Tibshirani. *An Introduction to the Bootstrap*. CRC press, 1994. 7
- [15] Gabriel Erion, Joseph D Janizek, Pascal Sturmfels, Scott M Lundberg, and Su-In Lee. Improving performance of deep learning models with axiomatic attribution priors and expected gradients. *Nature machine intelligence*, 3(7):620–631, 2021. URL <https://arxiv.org/pdf/1906.10670.pdf>. 3, 17
- [16] Felix Friedrich, Wolfgang Stammer, Patrick Schramowski, and Kristian Kersting. A typology to explore and guide explanatory interactive machine learning. *arXiv preprint arXiv:2203.03668*, 2022. URL <https://arxiv.org/pdf/2203.03668.pdf>. 2, 3, 5, 9
- [17] Yuyang Gao, Tong Sun, Liang Zhao, and Sungsoo Hong. Aligning eyes between humans and deep neural network through interactive attention alignment. *arXiv preprint arXiv:2202.02838*, 2022. URL <https://arxiv.org/pdf/2202.02838.pdf>. 3, 5
- [18] Reza Ghaeini, Xiaoli Z Fern, Hamed Shahbazi, and Prasad Tadepalli. Saliency learning: Teaching the model where to pay attention. In *NAACL*, 2019. URL <https://arxiv.org/pdf/1902.08649.pdf>. 3, 5
- [19] Vipul Gupta, Zhuowan Li, Adam Kortylewski, Chenyu Zhang, Yingwei Li, and Alan Yuille. Swapmix: Diagnosing and regularizing the over-reliance on visual context in visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5078–5088, June 2022. 2, 3
- [20] Peter Hase, Harry Xie, and Mohit Bansal. The out-of-distribution problem in explainability and search methods for feature importance explanations. *NeurIPS*, 34, 2021. URL <https://arxiv.org/pdf/2106.00786.pdf>. 23
- [21] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. URL <https://cs.stanford.edu/people/dorarad/gqa/index.html>. 6, 19, 20
- [22] Aya Abdelsalam Ismail, Hector Corrada Bravo, and Soheil Feizi. Improving deep learning interpretability by saliency guided training. *NeurIPS*, 34, 2021. URL <https://arxiv.org/pdf/2111.14338.pdf>. 25
- [23] Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? In *ACL 2020*, 2020. URL <https://www.aclweb.org/anthology/2020.acl-main.386.pdf>. 2, 5
- [24] Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in ai. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 624–635, 2021. URL <https://arxiv.org/pdf/2010.07487.pdf>. 10
- [25] Sarthak Jain and Byron C Wallace. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, 2019. 17
- [26] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017. 6
- [27] Brihi Joshi, Aaron Chan, Ziyi Liu, Shaoliang Nie, Maziar Sanjabi, Hamed Firooz, and Xiang Ren. Er-test: Evaluating explanation regularization methods for nlp models. *arXiv preprint arXiv:2205.12542*, 2022. 3
- [28] Brendan Kennedy, Xisen Jin, Aida Mostafazadeh Davani, Morteza Dehghani, and Xiang Ren. Contextualizing hate speech classifiers with post-hoc explanation. In *ACL*, 2020. URL <https://aclanthology.org/2020.acl-main.483.pdf>. 3, 5

- [29] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 21
- [30] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017. URL <https://arxiv.org/abs/1602.07332>. 7
- [31] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th international conference on intelligent user interfaces*, pages 126–137, 2015. URL https://dl.acm.org/doi/pdf/10.1145/2678025.2701399?casa_token=X9Yst59NvP4AAAAA:sC4BWZt6dIEKDLLODG4YZa9TkiCnRCn6C4N6G8TphVw2C9UQj7s0L2j_WyvPpzpd746zZSvNYQc. 1
- [32] Jiwei Li, Will Monroe, and Dan Jurafsky. Understanding neural networks through representation erasure. *arXiv preprint arXiv:1612.08220*, 2016. URL <https://arxiv.org/pdf/1612.08220.pdf>. 4, 17
- [33] Jiwei Li, Will Monroe, and Dan Jurafsky. Understanding Neural Networks through Representation Erasure. *arXiv:1612.08220 [cs]*, 2016. URL <http://arxiv.org/abs/1612.08220>. arXiv: 1612.08220. 3
- [34] Weixin Liang, James Zou, and Zhou Yu. ALICE: active learning with contrastive natural language explanations. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *EMNLP*, pages 4380–4391. Association for Computational Linguistics, 2020. URL <https://www.aclweb.org/anthology/2020.emnlp-main.355/>. 3
- [35] Frederick Liu and Besim Avci. Incorporating priors with feature attribution on text classification. In *ACL*, 2019. URL <https://aclanthology.org/P19-1631.pdf>. 1
- [36] Yibing Liu, Yangyang Guo, Jianhua Yin, Xueming Song, Weifeng Liu, Liqiang Nie, and Min Zhang. Answer questions with right image regions: A visual attention regularization approach. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 18(4):1–18, 2022. URL <https://arxiv.org/pdf/2102.01916.pdf>. 1, 2, 3
- [37] Scott M Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774, 2017. URL <http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf>. 3, 5, 17, 18
- [38] Horia Mania and Suvrit Sra. Why do classifier accuracies show linear trends under distribution shift? *arXiv preprint arXiv:2012.15483*, 2020. URL <https://arxiv.org/pdf/2012.15483.pdf>. 4
- [39] John Miller, Karl Krauth, Benjamin Recht, and Ludwig Schmidt. The effect of natural distribution shift on question answering models. In *International Conference on Machine Learning*, pages 6905–6916. PMLR, 2020. URL <https://arxiv.org/pdf/2004.14444.pdf>. 4
- [40] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.*, 267:1–38, 2019. doi: 10.1016/j.artint.2018.07.007. URL <https://doi.org/10.1016/j.artint.2018.07.007>. 4
- [41] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL 2002*, 2002. URL <https://www.aclweb.org/anthology/P02-1040.pdf>. 4
- [42] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NeurIPS*, 2017. URL <https://openreview.net/pdf?id=BJJsrmfCZ>. 5, 21

- [43] Grusha Prasad, Yixin Nie, Mohit Bansal, Robin Jia, Douwe Kiela, and Adina Williams. To what extent do human explanations of model behavior align with actual model behavior? In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, 2021. URL <https://aclanthology.org/2021.blackboxnlp-1.1.pdf>. 3, 9
- [44] Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset shift in machine learning*. Mit Press, 2008. URL <http://www.acad.bg/ebook/ml/The.MIT.Press.Dataset.Shift.in.Machine.Learning.Feb.2009.eBook-DDU.pdf>. 10
- [45] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 7, 10
- [46] Andrew Slavin Ross, Michael C. Hughes, and Finale Doshi-Velez. Right for the right reasons: Training differentiable models by constraining their explanations. In *IJCAI*, pages 2662–2670, 2017. doi: 10.24963/ijcai.2017/371. URL <https://www.ijcai.org/proceedings/2017/0371.pdf>. 1, 2, 3, 5
- [47] Ramprasaath Ramasamy Selvaraju, Stefan Lee, Yilin Shen, Hongxia Jin, Shalini Ghosh, Larry P. Heck, Dhruv Batra, and Devi Parikh. Taking a HINT: leveraging explanations to make vision and language models more grounded. In *ICCV*, pages 2591–2600. IEEE, 2019. doi: 10.1109/ICCV.2019.00268. URL https://openaccess.thecvf.com/content_ICCV_2019/papers/Selvaraju_Taking_a_HINT_Leveraging_Explanations_to_Make_Vision_and_Language_ICCV_2019_paper.pdf. 1, 2, 3, 5, 6, 7, 8, 9, 21, 26
- [48] Robik Shrestha, Kushal Kafle, and Christopher Kanan. A negative case analysis of visual grounding methods for VQA. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8172–8181, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.727. URL <https://aclanthology.org/2020.acl-main.727>. 2, 3, 5, 9
- [49] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *Workshop at International Conference on Learning Representations.*, 2013. URL <https://arxiv.org/pdf/1312.6034.pdf>. 17
- [50] Becks Simpson, Francis Dutil, Yoshua Bengio, and Joseph Paul Cohen. Gradmask: Reduce overfitting by regularizing saliency. In *MIDL 2019 Extended Abstracts*, 2019. URL <https://arxiv.org/pdf/1904.07478.pdf>. 1, 3, 5, 7, 8, 26
- [51] Sahil Singla, Mazda Moayeri, and Soheil Feizi. Core risk minimization using salient imagenet. *arXiv preprint arXiv:2203.15566*, 2022. URL <https://arxiv.org/pdf/2203.15566.pdf>. 1, 3, 7, 8, 26
- [52] Joe Stacey, Yonatan Belinkov, and Marek Rei. Supervising model attention with human explanations for robust natural language inference. In *AAAI*, 2022. URL <https://arxiv.org/pdf/2104.08142.pdf>. 3, 5, 17
- [53] Pascal Sturmfels, Scott Lundberg, and Su-In Lee. Visualizing the impact of feature attribution baselines. *Distill*, 5(1):e22, 2020. URL <https://distill.pub/2020/attribution-baselines/>. 4
- [54] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pages 3319–3328, 2017. URL <https://arxiv.org/pdf/1703.01365.pdf>. 3, 4, 17
- [55] Harini Suresh and John Guttag. A framework for understanding sources of harm throughout the machine learning life cycle. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–9, 2021. URL <https://dl.acm.org/doi/pdf/10.1145/3465416.3483305>. 10

- [56] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *EMNLP*, 2019. 7
- [57] Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, 33: 18583–18599, 2020. URL <https://proceedings.neurips.cc/paper/2020/file/d8330f857a17c53d217014ee776bfd50-Paper.pdf>. 4
- [58] Damien Teney, Ehsan Abbasnejad, and Anton van den Hengel. Learning what makes a difference from counterfactual examples and gradient supervision. In *European Conference on Computer Vision*, pages 580–599. Springer, 2020. URL <https://arxiv.org/pdf/2004.09034.pdf>. 1, 3
- [59] Damien Teney, Ehsan Abbasnejad, Kushal Kaffe, Robik Shrestha, Christopher Kanan, and Anton Van Den Hengel. On the value of out-of-distribution testing: An example of goodhart’s law. *Advances in Neural Information Processing Systems*, 33:407–417, 2020. URL <https://proceedings.neurips.cc/paper/2020/file/045117b0e0a11a242b9765e79cbf113f-Paper.pdf>. 3, 7, 10
- [60] Damien Teney, Yong Lin, Seong Joon Oh, and Ehsan Abbasnejad. Id and ood performance are sometimes inversely correlated on real-world datasets. *arXiv preprint arXiv:2209.00613*, 2022. URL <https://arxiv.org/pdf/2209.00613.pdf>. 4
- [61] Stefano Teso and Kristian Kersting. Explanatory interactive machine learning. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 239–245, 2019. URL https://ml-research.github.io/papers/teso2019aies_XIML.pdf. 1
- [62] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*, 2021. URL <https://arxiv.org/pdf/2112.04359.pdf>. 10
- [63] Sarah Wiegrefe and Yuval Pinter. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, 2019. URL <https://arxiv.org/pdf/1908.04626.pdf>. 17
- [64] Jialin Wu and Raymond Mooney. Self-critical reasoning for robust visual question answering. *Advances in Neural Information Processing Systems*, 32, 2019. URL <https://proceedings.neurips.cc/paper/2019/file/33b879e7ab79f56af1e88359f9314a10-Paper.pdf>. 1, 2, 3, 5, 6, 7, 8, 21, 26
- [65] Jianwei Yang, Jiayuan Mao, Jiajun Wu, Devi Parikh, David D Cox, Joshua B Tenenbaum, and Chuang Gan. Object-centric diagnosis of visual reasoning. *arXiv preprint arXiv:2012.11587*, 2020. URL <https://arxiv.org/pdf/2012.11587.pdf>. 2, 3, 5, 9
- [66] Ruiqi Zhong, Steven Shao, and Kathleen McKeown. Fine-grained sentiment analysis with faithful attention. *arXiv preprint arXiv:1908.06870*, 2019. URL <https://arxiv.org/pdf/1908.06870.pdf>. 1, 3, 5

Checklist

1. For all authors...
 - (a) Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope? [Yes] All claims made in the abstract and introduction are properly supported by our paper’s results and contributions.
 - (b) Did you describe the limitations of your work? [Yes] See Section 8.
 - (c) Did you discuss any potential negative societal impacts of your work? [Yes] Also see Section 8.

- (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
- 2. If you are including theoretical results...
 - (a) Did you state the full set of assumptions of all theoretical results? [N/A]
 - (b) Did you include complete proofs of all theoretical results? [N/A]
- 3. If you ran experiments...
 - (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] The code and instructions for reproducing the main results are in supplement. The datasets are freely available online (urls in Appendix E).
 - (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 6 and Appendix E and F.
 - (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] All main experiments are conducted with 5 random seeds. Hypothesis testing are conducted with 10k bootstrapping. See Section 7.
 - (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Appendix F.
- 4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
 - (a) If your work uses existing assets, did you cite the creators? [Yes] All creators of existing assets are properly cited.
 - (b) Did you mention the license of the assets? [Yes] All assets are used in accordance with their licenses (listed in Appendix E).
 - (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
 - (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
 - (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] See Section 8.
- 5. If you used crowdsourcing or conducted research with human subjects...
 - (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
 - (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
 - (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]