Proceedings of Machine Learning Research **318** (2026)
The 39th Canadian Conference on Artificial Intelligence

Interactive Learning from Explanations with Adaptive Guidance

Hadi Moazen [†,*], Flavie Lavoie-Cardinal [†, ‡], Audrey Durand[†,◇]
[†] Université Laval
[‡] CERVO Brain Research Center
[◇] Canada CIFAR AI Chair (Mila)

Abstract

Explanatory Interactive Learning (XIL) has emerged as a promising paradigm to bridge the gap between machine learning models and human understanding by integrating Explainable Artificial Intelligence (XAI) methods directly into the training process. Traditionally, XIL methods in computer vision rely on expert annotations specifying the evidence present in the input, collected before training starts and regardless of the model behavior during training. This can be detrimental to the interactive nature of XIL and miss out on the opportunity of taking advantage of the intermediate information about the model during training. In this paper, we formalize XIL as an interactive learning paradigm to provide guidance on model explanations through a series of interactions with an expert user during training. Furthermore, we introduce an approach to approximate the evidence from sparse adaptive interactions collected as guiding points indicating where explanations were deemed irrelevant by the expert during training. We evaluate the proposed framework using a simulated interactive loop to explore interactions in an adaptive setting. Our results show that by taking advantage of the information provided by the model explanations during training, the proposed adaptive framework is able to match, or even exceed, the performance and explainability of XIL methods trained with access to the ground-truth evidence with fewer interactions.

Keywords: Explainable Artificial Intelligence (XAI), Explainable Interactive Learning (XIL), Computer Vision

1. Introduction

Deep neural networks have achieved strong predictive results in a variety of computer vision benchmarks and applications [1–4]. However, this high performance comes at the cost of explainability [5]. Deep models often behave like black boxes, undermining user trust and raising practical and ethical concerns when their decision-making processes lack transparency [6]. Explainable Artificial Intelligence (XAI) [7] was developed to address the need for greater understandability of these models and to shed light on their decision-making processes. Unlike interpretable (glass-box) models [8], which are designed with inherent transparency in their decision-making processes, explainability methods provide post-hoc attributions and visualizations for otherwise opaque black-box models [7]. However, XAI is only a diagnostics tool and does not provide approaches for fixing the underlying problems once they are discovered [9].

Explanatory Interactive Learning (XIL) [9] was introduced as a framework to close this gap by using explanations produced by XAI methods as training feedback to guide models throughout training. Since neural networks are over-parameterized, for a given task and dataset, the model parameters can converge to different modes in the Rashomon set [8] in the model space, achieving similar performance while exhibiting different characteristics [10]. XIL offers the opportunity of guiding the model parameters towards modes where the *explanations make sense to a human* by integrating user feedback on explanations directly into the training loop. In this work, we focus on computer vision tasks, where explanations are typically provided in the form of saliency maps using post-hoc explanations (see Fig. 1).

[*]hadi.moazen.1@ulaval.ca

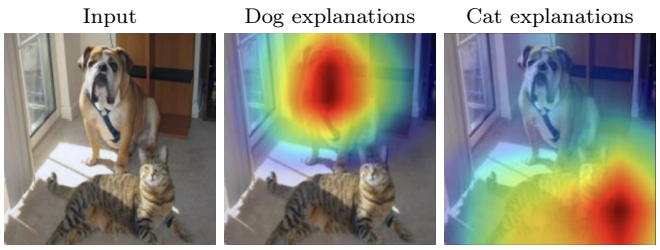

Figure 1. Example of Grad-CAM explanations [15].

Current XIL methods consider cases where underlying evidence is approximated through detailed expert feedback provided for every input-label pair in the dataset [11–14], which can make the evidence approximation step time-consuming and cumbersome.

In this work, we consider an XIL framework where the evidence is approximated through a series of low-effort interactions between the expert and the model explanations throughout training. Focusing on computer vision tasks, we propose an approach to approximate the underlying evidence from sparse guiding points. Guiding points are arguably the simplest form of feedback an expert can provide quickly and interactively. In the proposed framework, the guiding points provided by the experts are influenced by the current explanations of the model and the expert feedback adapts to model explanations for each input. We perform experiments in which we simulate experts with different levels of tolerance to irrelevant explanations. We show that the proposed interactive framework unlocks new trade-offs between model performance and explainability, while reducing the annotation burden compared to the traditional setting.

2. Interactive Learning from Explanations with Adaptive Guidance

In this section, we formally define the problem of learning a model guided by feedback on model explanations interactively provided by an expert in the loop. We focus on computer vision tasks, where $\mathcal{X}$ and $\mathcal{Y}$ respectively denote the input (image) and label spaces. The unknown function $f^\star : \mathcal{X} \mapsto \mathcal{Y}$ determines the underlying mapping form inputs $x \in \mathcal{X}$ to labels $y \in \mathcal{Y}$. Given a dataset $\mathcal{D} = \{x^i, y^i\}_{i=0}^N$ of N samples, the goal is to find a model $f_\theta : \mathcal{X} \mapsto \mathcal{Y}$ that approximates the real mapping $f^\star$ by relying, according to its explanations, on information deemed relevant by humans.

Given an input image x of dimension $W \times H$, an explanation typically corresponds to a matrix $E_\theta(x, y) \in [0, 1]^{W \times H}$ that indicates the contribution of each pixel to the decision of the model (closer to 1 means more contribution). We assume that human knowledge defines *evidence* $A(x, y) \in \{0, 1\}^{W \times H}$ indicating, for each pixel, whether it should contribute (1) or not (0) to the decision. Unfortunately, in practice, evidence is not directly accessible, and must be approximated, $\tilde{A}(x, y) \approx A(x, y)$, from expert interactions. *When learning from adaptive guidance, we assume that the expert is available to provide interactive feedback on model explanations associated with input-label pairs during training.* The objective is to minimize the number of expert interactions required to learn a model that maximizes performance and relevant explanations without relying on the unknown evidence.

Note 1 (Adaptive guidance)**.** *The proposed setting relies on expert feedback that is adapted in real-time to the current model explanation. This contrasts with the typical XIL settings [12, 13] in which the approximate evidence is solely extracted from the data (e.g., binary masks) and thus can be treated as any regular annotation.*

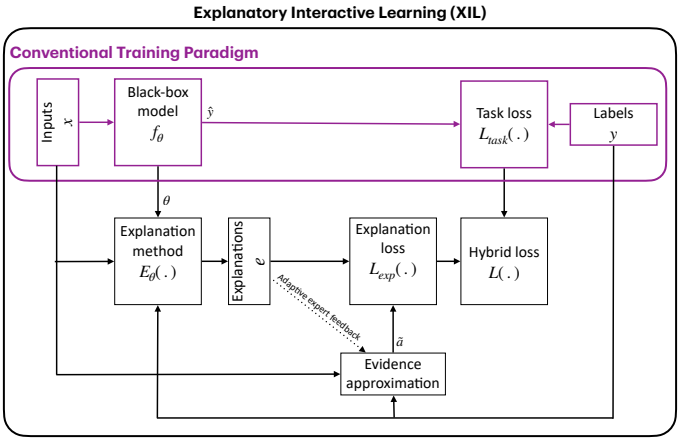

Figure 2. Typical guided training paradigm.

2.1. **Evaluation**

When evaluating an explainable model, we are interested in both the model performance on the considered task and the quality of the model explanations.

Performance metrics. The notion of performance is task-dependent and therefore any standard metric can be used. Usual metrics in computer vision tasks include accuracy or F1-score for classification [16, 17], and Average Precision (AP) or Intersection over Union (IoU) for object detection [18, 19].

Explainability metrics. The quality of model explanations provided as saliency maps can be evaluated by measuring the concentration of explanations (saliency) within regions deemed relevant by a human expert. Given an input-label pair $(x, y) \in \mathcal{D}$, the Energy-Based Pointing Game (EPG) [20] measures the proportion of model explanation $e = E_\theta(x, y)$ deemed relevant according to the approximated evidence $\tilde{a} = \tilde{A}(x, y)$[1]:

$$\text{EPG}(e, \tilde{a}) = \frac{\|e \odot \tilde{a}\|_1}{\|e\|_1}. \tag{2.1}$$

The EPG score lies in $[0, 1]$, with higher scores indicating more relevant explanations.

2.2. **Training**

Following prior work in XIL [13], the model is trained jointly for performance maximization and quality of explanations using a hybrid loss on input-label pairs $(x, y) \in \mathcal{D}$:

$$\mathcal{L}(f_\theta, x, y, e, \tilde{a}) = \mathcal{L}_{\text{task}}(f_\theta(x), y) + \lambda \mathcal{L}_{\text{exp}}(e, \tilde{a}), \text{where } \lambda \geq 0, \tag{2.2}$$

where $\mathcal{L}_{\text{task}}$ denotes a loss that captures the model performance on the task and $\mathcal{L}_{\text{exp}}$ denotes a loss that captures the relevance of explanation $e = E_\theta(x, y)$ according to the approximated evidence $\tilde{a} = \tilde{A}(x, y)$. Figure 2 illustrates the computation of such hybrid loss on a given learning iteration. While conventional training relies solely on a task loss to train the model, XIL also includes an explanation loss based on approximate evidence. Under adaptive guidance, the evidence is approximated from expert feedback interactively provided on the current model explanation.

Task loss. The task loss is used to train the model to perform well on the main task. Consequently, it is task-dependent and its formulation depends on the task at hand. In our experiments, we will use cross-entropy loss across the board for the multi-label classification.

[1]The Hadamard product $A \odot B$ denotes the element-wise product between two matrices A and B.

Explanation loss. The explanation loss can been formulated in different ways, which can encourage different traits in the trained model. For example, the *L1 loss* [11, 12] measures the absolute difference between the explanation $e = E_\theta(x, y)$ and the approximated evidence $\tilde{a} = \tilde{A}(x, y)$ for each pixel:

$$\mathcal{L}_{\text{exp}}^{(\text{L1})}(e, \tilde{a}) = \|e - \tilde{a}\|_1. \tag{2.3}$$

Alternatively, the *point-wise cross entropy (PCE) loss* [21] measures the point-wise binary cross-entropy between the binarized explanation $e = E_\theta(x, y) \in \{0, 1\}^{W \times H}$ and the approximated evidence $\tilde{a} = \tilde{A}(x, y)$ for each pixel:

$$\mathcal{L}_{\text{exp}}^{(\text{PCE})}(e, \tilde{a}) = \sum_{i,j} \left(\tilde{a}_{ij} \log e_{ij} + (1 - \tilde{a}_{ij} \log(1 - e_{ij})\right). \tag{2.4}$$

Both the L1 and PCE formulations encourage the model to *mimic* the approximated evidence. More recently, the *Energy Pointing Game (EPG) loss* [14] introduced the idea of penalizing the model for irrelevant explanations (according to the approximated evidence $\tilde{a}$) by using the negative EPG score (Eq. 2.1) as the explanation loss:

$$\mathcal{L}_{\text{exp}}^{(\text{EPG})}(e, \tilde{a}) = -\,\text{EPG}(e, \tilde{a}). \tag{2.5}$$

In our experiments we will consider all three losses.

3. Adaptive Guidance from Guiding Points

Under the proposed XIL setting with adaptive guidance, the expert is asked to provide feedback on the current model explanation for each input-label pair, at each learning iteration. This could rapidly become an unbearable burden for the expert. In this section, we introduce a procedure to approximate the evidence from weak, low-effort, interactions.

3.1. Approximating evidence from expert feedback

Every time the model processes an input and produces the associate explanations, we ask the expert to click on the locations where the model explanation is irrelevant. We then use the provided single-pixel clicked points as *guiding points* to construct a *mask of irrelevance*, Fig 3 provides samples of such guiding points and resulting masks.

The mask of irrelevance is created by expanding guiding points to regions whose latent representation is similar to the latent representation of a guiding point. Concretely, let $h_{ij} \in \mathcal{H}^d$ denote the d-dimensional latent representations of an input at the coordinates (i, j). We compute the cosine similarity between the latent representation at each spatial location (i, j) and the latent representation at guiding point location (p, q):

$$S_c(h_{ij}, h_{pq}) = \frac{h_{ij} \cdot h_{pq}}{\|h_{ij}\|_2 \, \|h_{pq}\|_2}.$$

Given a similarity threshold $\delta \geq 0$, we consider that location (i, j) is irrelevant if the latent representation at this position is sufficiently similar to the latent representation at the guiding point location (p, q), that is $S_c(h_{ij}, h_{pq}) > \delta$. Given a set of guiding points, a location is deemed irrelevant if it is irrelevant according to any guiding point in the set. Algorithm 1 summarizes the procedure for creating a mask of irrelevant regions guided by a set of guiding points. Figure 3 provides examples of masks of irrelevance extracted from given guiding points on two different inputs with a ResNet backbone and similarity threshold $\delta = 0.995$.

In practice, the set of guiding points associated to a given sample (x, y) grows over training epochs, as the expert iteratively provides adaptive feedback on the associated explanation $E_\theta(x, y)$ (which evolves with the model over training epochs). As the expert is only asked to provide feedback where explanations are deemed irrelevant, the expert stops providing

Algorithm 1 Identifying irrelevant regions for an input x of dimension $W \times H$

Input: Latent representation $h(x) \in \mathcal{H}^{W \times H \times d}$, threshold δ, set of guiding points G
Output: Mask m
Initialize an empty matrix $m \leftarrow \mathbf{0}^{W \times H}$
for every location (i, j) in input x **do**
    for every guiding point location (p, q) in G **do**:
        Update mask $m_{pq} \leftarrow \max(m_{pq}, S_c(h_{ij}, h_{pq}) > \delta)$
    end for
end for
Return m

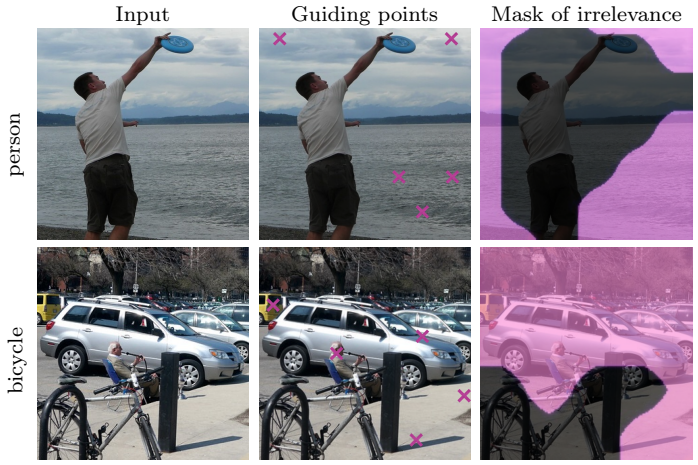

Figure 3. Masks of irrelevance obtained from sampled guiding points.

feedback on an input when its explanation converge to relevant regions (according to the expert). Given a mask of irrelevance m for an input x of dimension $W \times H$, we approximate the evidence as

$$\tilde{A}(x, y) = \mathbf{1}^{W \times H} - m, \tag{3.1}$$

which distinguishes *possibly relevant* regions (1) from *confidently irrelevant* regions (0).

3.2. **Training from adaptive guidance with guiding points**

With the proposed approach, the expert is only asked to provide feedback on irrelevant explanations. Therefore, the resulting approximated evidence (Eq. 3.1) contains information with variable confidence. Indeed, while regions marked as irrelevant by the expert can be treated as such, regions not marked as irrelevant may still be irrelevant. Such imperfection in the approximate evidence should be taken into account in the choice of training loss. As such, we focus on the EPG loss (Eqs. 2.5 and 2.1), as it relies only on regions confidently marked as irrelevant in the approximate evidence.

By construction, the approximate evidence for a given sample will contain much more signal when the explanation is less relevant. As such, gradients from samples with smaller irrelevant regions can be low and get ignored at training. To avoid this situation, we propose a weighted variant of the EPG loss, where the EPG score is adjusted based on the input

dimension $W \times H$ and the number of guiding points n collected on that input:

$$\mathcal{L}_{\text{exp}}^{(\text{EPG}_w)}(e, \tilde{a}) = -\frac{H \times W}{n} \text{EPG}(e, \tilde{a}). \tag{3.2}$$

4. Experiments

We conduct a series of experiments to evaluate the potential of the proposed approach for learning from explanations using adaptive expert guidance. We consider the task of multi-label classification using a backbone pre-trained on ImageNet [22] (without explanations) and fine-tuned using explanations on a target dataset. We carry out experiments on two large-scale computer vision datasets, MS COCO [23] and PASCAL VOC [24]. These datasets come with per-class segmentation masks for each sample (x, y) which we use as underlying evidence $A(x, y)$.

Backbones and explanations. We consider three model backbones with a corresponding explanation method for each: 1) ResNet50 [25] with Grad-CAM [15] explanations; 2) B-Cos [26] with its inherent bcos explanations; and 3) X-DNN [27] backbone with Integrated Gradient [28] explanations, which can be calculated with a single backward pass of the model. Following the setting in [14], backbones were fine-tuned for 10 epochs on the COCO dataset, and for 50 epochs on the PASCAL VOC dataset. For each model configuration, hyperparameter fine-tuning is carried out for $\lambda \in \{10^{-4}, 5 \times 10^{-4}, 10^{-3}, 5 \times 10^{-3}\}$ to select the models with the best performance and explainability on the validation set. We report the performance of the selected models on the test set.

Baselines. As the unguided baseline, we train the backbones without any guidance for explanations (Eq. 2.2 with $\lambda = 0$) on the target dataset. As evidence-guided baselines, we fine-tune the unguided baseline using guidance from the *perfect* underlying evidence: $\tilde{A}(x, y) = A(x, y)$. This is consistent with traditional XIL settings for classification [12, 13], where approximate evidence is often defined as precise segmentation masks or bounding boxes around objects of interest. For each backbone, we train three such evidence-guided baselines using the hybrid loss (Eq. 2.2) combined with the L1 loss (Eq. 2.3), the PCE loss (Eq. 2.4), and the EPG loss (Eq. 2.5), respectively. Note that this requires one fully detailed annotation for each sample in the dataset, unlike the proposed adaptive setting where the expert provides interactive feedback on each sample and its explanation during training.

4.1. Simulating user interactions

To make the experiments reproducible, we introduce a procedure for simulating expert interactions. We assume that when an expert is shown the current explanations of the model and is asked to provide guiding points, they will avoid the relevant region and choose the points in irrelevant regions proportional to their attributed saliency. We simulate this interaction by sampling guiding points outside the class-relevant region given evidence $A(x, y)$, with probability given by the saliency of the explanation. The selected points are added to the set of guiding points for the sample (x, y).

As the saliency is continuous in $[0, 1]$, it is very likely to remain positive, although low, in irrelevant regions. In practice, the decision to correct the model depends on the expert's interpretation and tolerance regarding the extent to which explanations highlight irrelevant regions. We simulate this expert tolerance using a threshold $\tau \in [0, 1)$ such that the expert keeps providing feedback until the saliency in irrelevant regions falls below this threshold. As a result, when the explanation meets the expert tolerance everywhere on a sample, no further feedback is provided for that sample. Algorithm 2 describes the general simulation procedure of expert interaction with a sample. All such samples start with an empty set of guiding points and the guiding points are accumulated for each sample as they are provided by the expert during training.

Algorithm 2 Simulating an interaction of the expert with sample (x, y).

Input: Evidence $a = A(x, y)$, explanations $e = E_\theta(x, y)$, guiding points G, threshold τ
Output: Updated guiding points G
Compute normalized irrelevant explanations $e^- \leftarrow e \odot (\mathbf{1}^{W \times H} - a)/\|e\|_1$
Compute mask of irrelevance m using guiding points G (Alg. 1)
Initialize fraction of irrelevant saliency $s \leftarrow \|e^- \odot m\|_1$
while s $> \tau$ **do**:
    Sample a guiding point g using e^- as the distribution
    $G \leftarrow G \cup g$
    Update m using guiding points G (Alg. 1)
    Update $s \leftarrow \|e^- \odot m\|_1$
end while
Return G

We conduct our experiments using threshold $\tau \in \{0, 0.25, 0.5, 0.75\}$ to simulate different levels of user tolerance when providing feedback. In this case, a threshold $\tau = 0$ would correspond to a scrutinizing user who is extremely determined in removing all saliency from irrelevant regions, whereas a threshold $\tau = 0.75$ would define a lenient user.

4.2. **Results**

We report the performance on the classification task (F1-score) and a measure of the relevance of explanations (EPG score, Eq. 2.1) for all training strategies. Figures 4 and 5 respectively display the results on COCO and PASCAL VOC, in which dominated and dominating regions are indicated with respect to the unguided baseline. When models with the highest F1-score are selected (performance first, Figs. 4 and 5, top-row), we observe that adaptive guidance from guiding points can result in models that perform similarly to

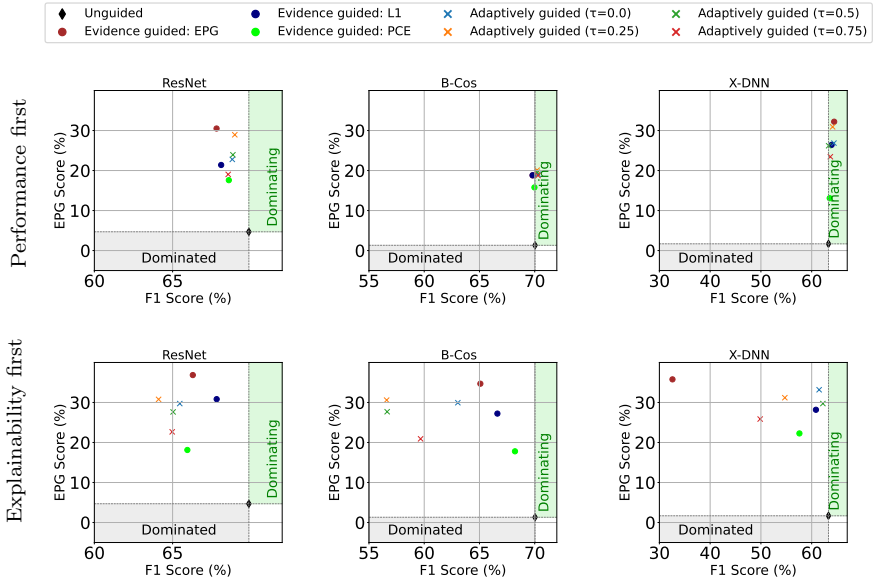

Figure 4. Performance and explainability on COCO for all baselines and the proposed adaptive guidance under different levels of simulated user tolerance (τ).

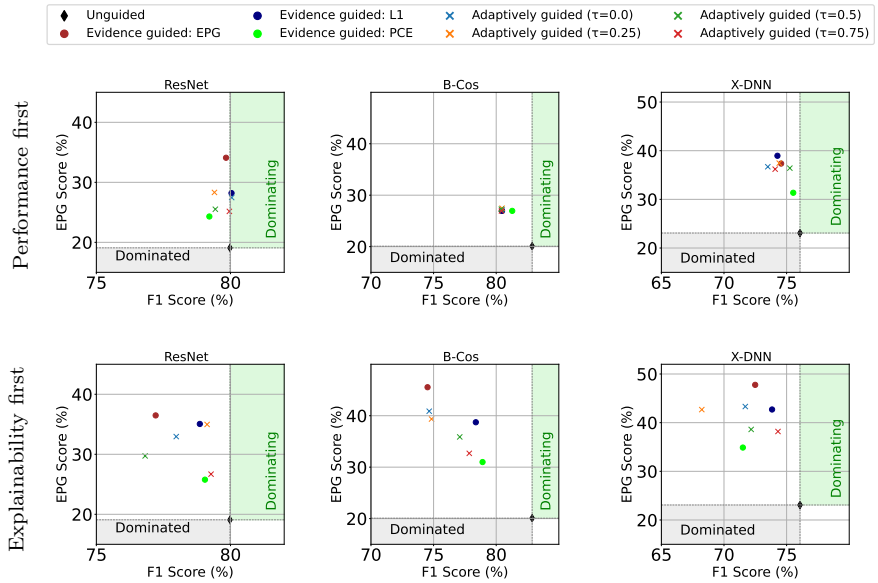

Figure 5. Performance and explainability on PASCAL VOC for all baselines and the proposed adaptive guidance under different levels of simulated user tolerance (τ).

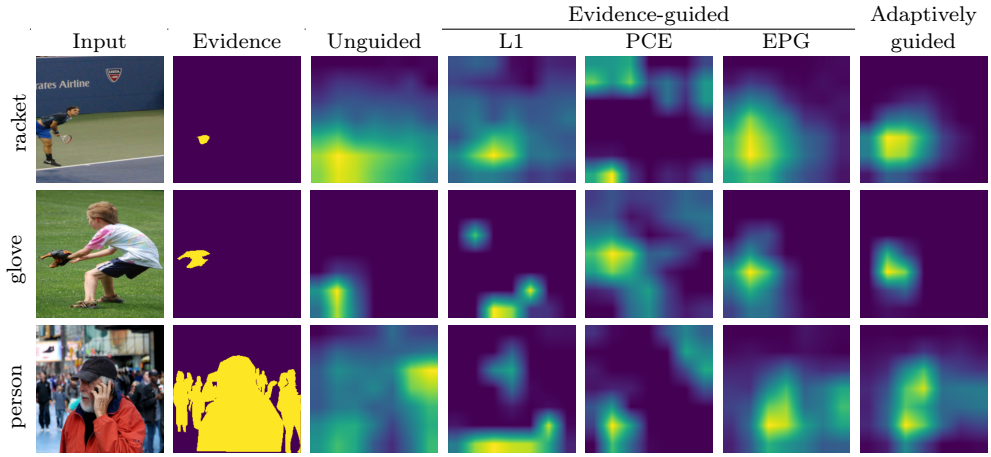

Figure 6. Samples of explanations obtained with baselines and with adaptive guidance.

evidence-guided models, while sometimes improving the explainability (for example, using ResNet with GradCam explanations on both datasets, Figs. 4 and 5, first-column). When models are selected based on highest EPG score (explainability first, Figs. 4 and 5, bottom-row), we observe that adaptive guidance from guiding points often results in an explicit trade-off between performance and explainability. Figure 6 provides some samples of Grad-CAM explanations with the ResNet backbone on the COCO dataset, given a semi-lenient simulated expert ($\tau = 0.5$) for adaptive guidance. We observe that the adaptively guided model can guide model explanations away from irrelevant evidence and towards the relevant regions of the image similar to evidence-guided models, while requiring fewer interactions (see Sec. 4.3).

Adaptive guidance from guiding points relies on interactions between the expert and the explanations. As such, the annotation burden and the quality of the approximate evidence are highly dependent on the feedback strategy of the expert. Given our simulated user interactions, we would expect that a scrutinizing expert ($\tau \to 0$) should incur a higher annotation burden compared with a lenient expert ($\tau \to 1$) willing to tolerate more saliency in irrelevant regions, which in turn should lead the former to better approximate the evidence. We investigate these aspects in the next sections.

4.3. **Annotation burden**

Figure 7 displays the number of interactions required per sample using guidance from evidence and adaptive guidance from guiding points under different levels of user tolerance. In the case of adaptive guidance, interactions correspond to the number of clicks on explanations associated with a given sample. In the case of guidance from evidence, interactions correspond to the number of clicks on an input required to produce a detailed segmentation mask. As expected, we observe that increasing the tolerance to errors ($\tau \to 1$) decreases the number of interactions. As a result, lenient experts would suffer less from the annotation burden. More importantly, lenient experts may even completely skip feedback on samples where they deem the explanation to be sufficiently accurate. Conversely, scrutinizing experts would end up providing more feedback, increasing their annotation burden. Overall,

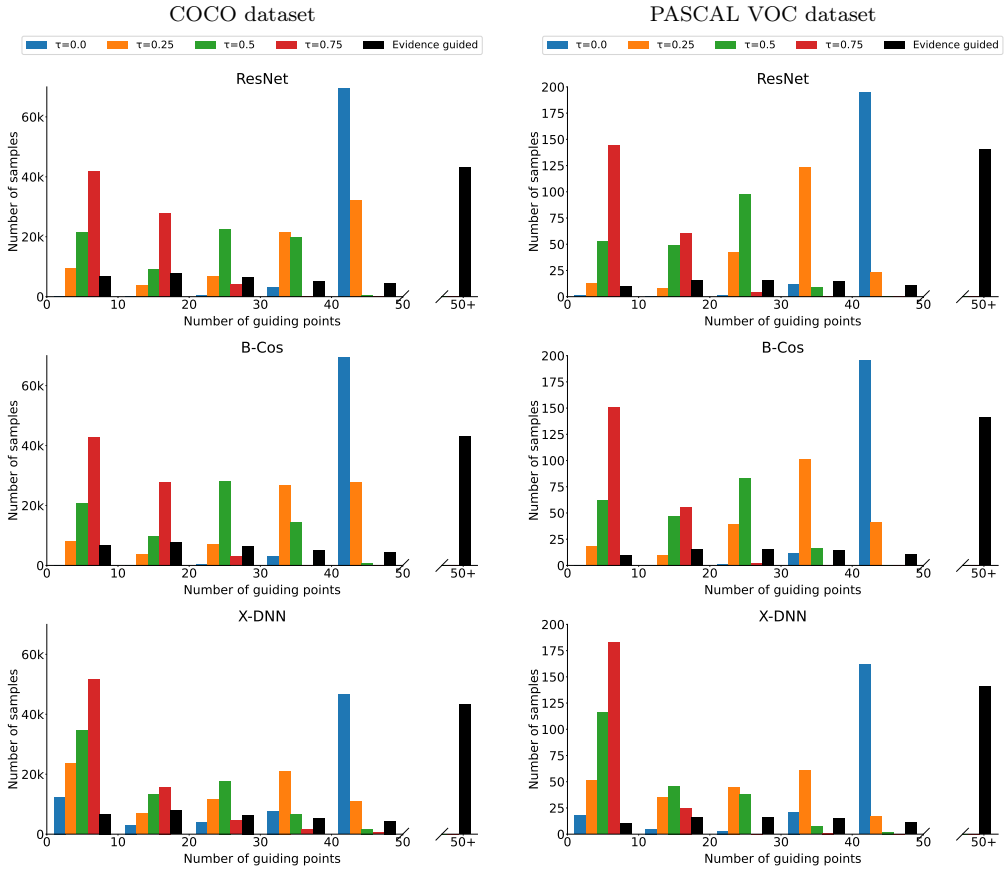

Figure 7. Number of interactions using guidance from evidence and adaptive guidance from guiding points under different levels of simulated user tolerance (τ).

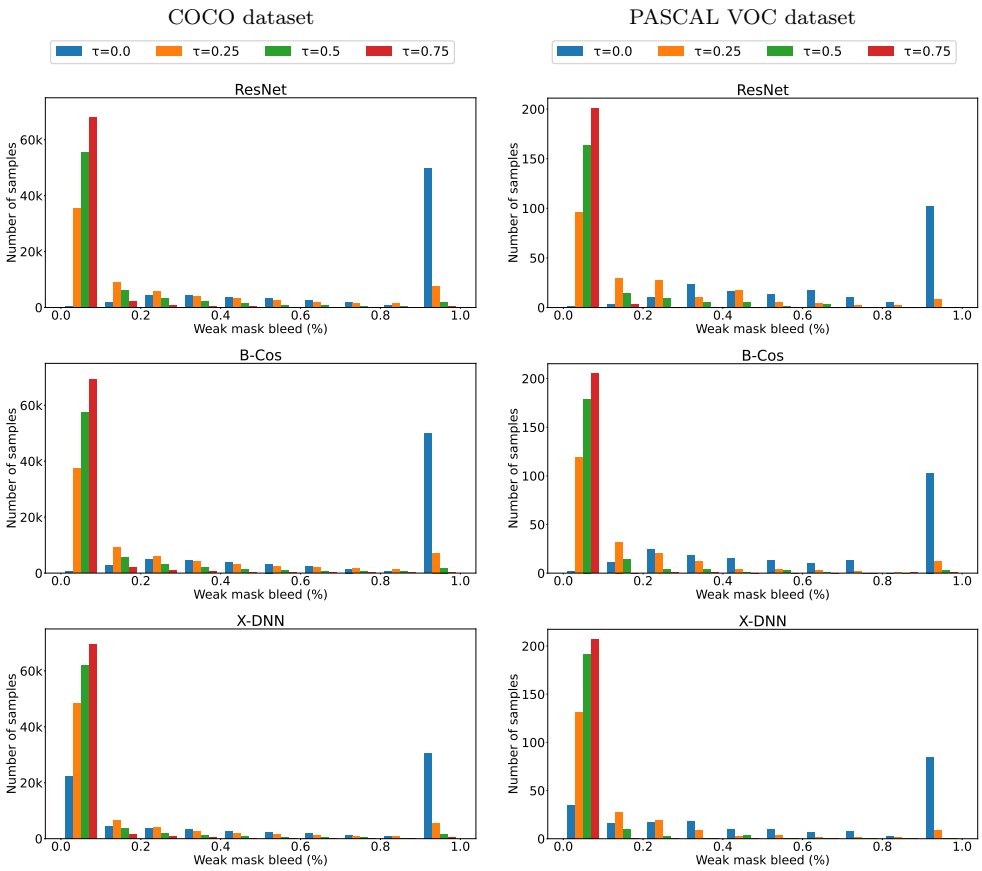

Figure 8. Fraction of relevant regions falsely marked as irrelevant due to mask bleed based for different levels of simulated user tolerance (τ), for both datasets.

we observe that adaptive guidance significantly reduces the amount of interactions compared with evidence guided methods.

4.4. **Evidence approximation**

It is expected that increasing the number of interactions between an expert and samples should result in a better approximation of the evidence (Eq. 3.1). However, our results surprisingly show that the EPG score obtained under adaptive guidance from scrutinizing experts does not necessarily outperform the EPG score obtained with lenient experts. To explain this, we investigate the quality of the resulting approximate evidence. Recall that the proposed approach for adaptive guidance propagates the feedback provided by the expert as guiding points to the entire input. This step relies on the cosine similarity between the latent representation of the input at guiding points and the latent representation at all other locations in the input (Alg. 1). While this reduces the need to collect dense feedback everywhere in the input, the trade-off is that nothing prevents the mask of irrelevance created from guiding points from propagating to relevant regions. We refer to this phenomenon as *bleeding* and investigate to what extent it is present.

Figure 8 displays the proportion of the relevant region (according to the evidence) marked as irrelevant according to the approximate evidence on the COCO dataset. We observe that bleeding is insignificant for the majority of samples when the expert has some leniency

($\tau \in \{0.25, 0.5, 0.75\}$). However, we notice major bleeding when the expert is scrutinizing ($\tau = 0.0$). As the explanations are dependent on the latent representation, a low but positive saliency in irrelevant regions may be unavoidable. This indicates that experts providing adaptive feedback within the proposed approach should demonstrate some tolerance and not expect to reach null saliency in irrelevant regions.

Conclusion

In this paper, we formalized XIL as a paradigm to provide guidance on model explanations through a series of interactions with an expert user during training. Focusing on computer vision tasks, we proposed an approach to approximate the evidence from sparse adaptive interactions collected as guiding points indicating where explanations were deemed irrelevant by the expert during training. To evaluate this framework, we provided a simulated user interaction loop, which was then used to fine-tune different backbones on a multi-label classification task on two large-scale computer vision datasets. The experiments demonstrated the efficacy of the proposed framework, which was able to match or exceed the performance and explainability of the models trained guided with the ground-truth evidence.

Since masks of irrelevant regions in the proposed framework are calculated based on the latent representation of guiding points, exploring whether guiding points can be transferred across samples is an interesting step forward. If possible, this would further reduce annotation burden by reusing guidance across samples. We also aim to extend our experiments by applying the adaptive guidance framework on a real-world dataset with human annotators in the loop to further validate the results obtained using simulated feedback.

Acknowledgements

We acknowledge funding from DEEL and the Canada CIFAR AI Chair program.

References

[1] N. Bendre, N. Ebadi, J. J. Prevost, and P. Najafirad. "Human action performance using deep neuro-fuzzy recurrent attention model". In: *IEEE Access* 8 (2020), pp. 57749–57761.

[2] Y. Chen, D. Zhao, L. Lv, and Q. Zhang. "Multi-task learning for dangerous object detection in autonomous driving". In: *Information Sciences* 432 (2018), pp. 559–571.

[3] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. "Learning deep features for discriminative localization". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 2921–2929.

[4] Y. Lu, Y. Chen, D. Zhao, B. Liu, Z. Lai, and J. Chen. "CNN-G: Convolutional neural network combined with graph for image segmentation with theoretical analysis". In: *IEEE Transactions on Cognitive and Developmental Systems* 13.3 (2020), pp. 631–644.

[5] Z. Zhao, P. Xu, C. Scheidegger, and L. Ren. "Human-in-the-loop extraction of interpretable concepts in deep learning models". In: *IEEE Transactions on Visualization and Computer Graphics* 28.1 (2021), pp. 780–790.

[6] A. Das and P. Rad. "Opportunities and challenges in explainable artificial intelligence (xai): A survey". In: *arXiv preprint arXiv:2006.11371* (2020).

[7] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI". In: *Information fusion* 58 (2020), pp. 82–115.

[8] C. Rudin, C. Chen, Z. Chen, H. Huang, L. Semenova, and C. Zhong. "Interpretable machine learning: Fundamental principles and 10 grand challenges". In: *Statistic Surveys* 16 (2022), pp. 1–85.

[9] F. Friedrich, W. Stammer, P. Schramowski, and K. Kersting. "A typology for exploring the mitigation of shortcut behaviour". In: *Nature Machine Intelligence* 5.3 (2023), pp. 319–330.

[10] H. Bang, A. Boggust, and A. Satyanarayan. "Explanation Alignment: Quantifying the Correctness of Model Reasoning At Scale". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2024, pp. 288–315.

[11] Y. Gao, T. S. Sun, L. Zhao, and S. R. Hong. "Aligning eyes between humans and deep neural network through interactive attention alignment". In: *Proceedings of the ACM on Human-Computer Interaction* 6.CSCW2 (2022), pp. 1–28.

[12] Y. Gao, T. S. Sun, G. Bai, S. Gu, S. R. Hong, and Z. Liang. "Res: A robust framework for guiding visual explanation". In: *Proceedings of the Conference on Knowledge Discovery and Data Mining*. 2022, pp. 432–442.

[13] Y. Gao, S. Gu, J. Jiang, S. R. Hong, D. Yu, and L. Zhao. "Going beyond xai: A systematic survey for explanation-guided learning". In: *ACM Computing Surveys* 56.7 (2024), pp. 1–39.

[14] S. Rao, M. Böhle, A. Parchami-Araghi, and B. Schiele. "Studying how to efficiently and effectively guide models with explanations". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2023, pp. 1922–1933.

[15] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. "Grad-cam: Visual explanations from deep networks via gradient-based localization". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2017, pp. 618–626.

[16] P. K. Mall, P. K. Singh, S. Srivastav, V. Narayan, M. Paprzycki, T. Jaworska, and M. Ganzha. "A comprehensive review of deep neural networks for medical image processing: Recent developments and future opportunities". In: *Healthcare Analytics* 4 (2023), p. 100216.

[17] Y. Zhang, T. Jiang, B. Pan, J. Wang, G. Bai, and L. Zhao. "MEGL: Multimodal Explanation-Guided Learning". In: *arXiv preprint arXiv:2411.13053* (2024).

[18] T. Diwan, G Anirudh, and J. V. Tembhurne. "Object detection using YOLO: challenges, architectural successors, datasets and applications". In: *multimedia Tools and Applications* 82.6 (2023), pp. 9243–9275.

[19] R. Padilla, S. L. Netto, and E. A. Da Silva. "A survey on performance metrics for object-detection algorithms". In: *Proceedings of the International Conference on Systems, Signals and Image Processing (IWSSIP)*. 2020, pp. 237–242.

[20] H. Wang, Z. Wang, M. Du, F. Yang, Z. Zhang, S. Ding, P. Mardziel, and X. Hu. "Score-CAM: Score-weighted visual explanations for convolutional neural networks". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 2020, pp. 24–25.

[21] H. Shen, K. Liao, Z. Liao, J. Doornberg, M. Qiao, A. Van Den Hengel, and J. W. Verjans. "Human-AI interactive and continuous sensemaking: A case study of image classification using scribble attention maps". In: *Extended abstracts of the Conference on Human Factors in Computing Systems*. 2021, pp. 1–8.

[22] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. "Imagenet large scale visual recognition challenge". In: *International journal of computer vision* 115.3 (2015), pp. 211–252.

[23] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. "Microsoft coco: Common objects in context". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2014, pp. 740–755.

[24] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. "The pascal visual object classes (voc) challenge". In: *International journal of computer vision* 88.2 (2010), pp. 303–338.

[25] K. He, X. Zhang, S. Ren, and J. Sun. "Deep residual learning for image recognition". In: *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 770–778.

[26] M. Böhle, M. Fritz, and B. Schiele. "B-cos networks: Alignment is all we need for interpretability". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 10329–10338.

[27] R. Hesse, S. Schaub-Meyer, and S. Roth. "Fast axiomatic attribution for neural networks". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 19513–19524.

[28] M. Sundararajan, A. Taly, and Q. Yan. "Axiomatic attribution for deep networks". In: *Proceedings of Machine Learning Research (MLR)*. 2017, pp. 3319–3328.

Appendix A. Hyperparameter Tuning

In this appendix, we present extensive results with different values for λ.

Backbone	Guidance	Setting	Lambda							
			$\lambda = 10^{-4}$		$\lambda = 5 \times 10^{-4}$		$\lambda = 10^{-3}$		$\lambda = 5 \times 10^{-3}$	
			F1	EPG	F1	EPG	F1	EPG	F1	EPG
ResNet	Evidence	EPG	68.32	21.96	**68.80**	30.06	68.64	30.31	67.13	35.56
		L1	**68.92**	21.03	68.80	25.39	68.75	27.35	68.81	28.11
		PCE	**68.90**	17.78	68.78	15.82	68.89	15.66	68.61	11.97
	Adaptive	$\tau = 0.0$	**69.67**	22.42	68.64	26.43	67.56	26.44	59.63	25.44
		$\tau = 0.25$	**69.84**	28.51	68.11	25.48	67.70	25.78	58.14	25.52
		$\tau = 0.5$	**69.77**	23.55	68.81	22.22	68.19	22.19	61.52	22.11
		$\tau = 0.75$	**69.41**	18.95	68.36	18.47	68.26	18.26	61.96	18.47
B-Cos	Evidence	EPG	**70.28**	18.39	70.26	21.22	70.19	28.36	68.25	30.43
		L1	**70.32**	18.41	70.31	20.02	70.27	21.34	69.59	25.61
		PCE	70.45	17.31	**70.78**	15.41	70.39	14.56	66.15	12.12
	Adaptive	$\tau = 0.0$	**70.99**	18.66	70.75	20.93	70.20	22.81	65.27	25.91
		$\tau = 0.25$	**70.99**	19.69	70.10	22.55	69.27	23.76	64.35	25.22
		$\tau = 0.5$	**70.95**	19.15	70.46	20.70	70.01	21.30	66.94	22.04
		$\tau = 0.75$	**70.96**	18.36	70.32	18.77	69.76	18.80	67.91	18.19
X-DNN	Evidence	EPG	64.16	27.37	**64.96**	31.77	63.86	34.10	36.42	34.67
		L1	64.11	24.06	**64.54**	26.04	63.77	27.43	61.78	31.28
		PCE	63.74	18.49	58.29	12.07	**64.17**	12.75	38.80	10.10
	Adaptive	$\tau = 0.0$	**65.03**	26.43	63.82	30.64	62.07	32.78	29.30	28.03
		$\tau = 0.25$	63.96	26.80	**64.92**	30.43	57.64	29.54	31.95	25.65
		$\tau = 0.5$	**64.37**	25.85	63.50	28.42	53.40	28.47	35.25	26.10
		$\tau = 0.75$	**63.99**	23.04	63.30	23.25	54.37	23.48	27.87	6.76

Table 1. performance-first experiments on COCO validation set.

Backbone	Guidance	Setting	Lambda							
			$\lambda = 10^{-4}$		$\lambda = 5 \times 10^{-4}$		$\lambda = 10^{-3}$		$\lambda = 5 \times 10^{-3}$	
			F1	EPG	F1	EPG	F1	EPG	F1	EPG
ResNet	Evidence	EPG	**67.79**	21.43	67.81	30.50	67.50	30.73	66.30	36.49
		L1	68.10	21.39	**68.11**	25.71	67.87	27.30	67.70	28.53
		PCE	**68.59**	17.59	68.25	15.86	68.08	16.04	68.18	12.30
	Adaptive	$\tau = 0.0$	**68.81**	22.78	67.95	26.75	66.58	26.75	58.46	25.78
		$\tau = 0.25$	**68.97**	28.94	67.31	25.83	66.97	26.13	57.60	25.85
		$\tau = 0.5$	**68.85**	23.94	67.88	22.60	67.42	22.56	60.85	22.47
		$\tau = 0.75$	**68.55**	19.03	67.82	18.89	67.29	18.67	61.52	18.89
B-Cos	Evidence	EPG	69.81	18.79	**69.83**	20.61	69.80	28.56	67.47	30.91
		L1	**69.81**	18.81	69.80	20.41	69.72	21.73	69.09	25.63
		PCE	69.91	17.70	**69.97**	15.78	69.84	14.96	65.54	12.46
	Adaptive	$\tau = 0.0$	**70.30**	19.07	70.07	21.35	69.56	23.21	64.42	26.34
		$\tau = 0.25$	**70.22**	20.09	69.60	22.94	68.82	24.15	63.57	25.62
		$\tau = 0.5$	**70.24**	19.55	69.91	21.11	69.49	21.71	66.42	22.46
		$\tau = 0.75$	**70.30**	18.76	69.95	19.03	69.42	19.05	67.49	18.61
X-DNN	Evidence	EPG	63.57	27.74	**64.43**	32.20	63.31	34.56	35.95	35.10
		L1	63.54	24.41	**63.92**	26.41	63.35	27.79	61.33	31.12
		PCE	62.83	18.86	57.65	12.39	**63.55**	13.09	37.95	10.39
	Adaptive	$\tau = 0.0$	**64.39**	26.86	63.08	31.13	61.79	33.12	28.56	28.57
		$\tau = 0.25$	63.69	25.25	**64.12**	30.92	56.99	30.03	31.61	26.13
		$\tau = 0.5$	**63.34**	26.19	62.68	28.97	52.83	29.02	34.33	26.52
		$\tau = 0.75$	**63.68**	23.45	62.43	23.70	53.42	24.04	27.54	6.92

Table 2. performance-first experiments on COCO test set.

Highest F1 (EPG) value in the performance-(explainability-) first experiments on the validation set is used to determine the λ value. Results corresponding to that λ on the test sets are used in Figs. 4 and 5.

Backbone	Guidance	Setting	Lambda							
			$\lambda = 10^{-4}$		$\lambda = 5 \times 10^{-4}$		$\lambda = 10^{-3}$		$\lambda = 5 \times 10^{-3}$	
			F1	EPG	F1	EPG	F1	EPG	F1	EPG
ResNet	Evidence	EPG	68.32	21.96	68.80	30.06	68.36	32.98	67.02	**36.47**
		L1	68.92	21.03	68.80	25.39	68.75	27.35	68.64	**30.46**
		PCE	68.90	**17.78**	60.46	17.36	66.35	16.64	56.59	16.26
	Adaptive	$\tau = 0.0$	69.53	25.43	68.13	28.85	66.49	**29.46**	56.98	27.88
		$\tau = 0.25$	69.84	28.51	64.81	**30.58**	64.14	30.11	43.75	28.30
		$\tau = 0.5$	69.16	25.74	65.40	**27.28**	64.70	26.92	53.26	24.34
		$\tau = 0.75$	68.78	20.83	66.00	**22.27**	64.17	22.05	54.81	19.97
B-Cos	Evidence	EPG	68.44	19.83	67.77	24.91	67.99	25.27	65.88	**34.28**
		L1	69.01	19.29	68.50	21.87	68.46	23.33	67.23	**26.79**
		PCE	70.45	**17.31**	70.50	15.84	69.04	15.76	63.74	12.89
	Adaptive	$\tau = 0.0$	70.04	20.27	68.94	25.30	68.77	26.81	63.68	**29.53**
		$\tau = 0.25$	69.81	21.79	68.98	25.91	67.49	27.78	57.81	**29.78**
		$\tau = 0.5$	69.72	20.76	68.79	23.81	68.32	24.91	56.84	**27.27**
		$\tau = 0.75$	70.19	18.84	69.36	19.36	67.64	19.94	60.36	**20.52**
X-DNN	Evidence	EPG	62.08	27.90	61.42	32.50	59.29	34.97	33.21	**35.35**
		L1	61.67	24.33	61.80	26.32	61.18	**27.83**	61.80	26.32
		PCE	58.15	**21.85**	58.29	12.07	50.40	13.50	25.66	10.66
	Adaptive	$\tau = 0.0$	60.98	27.75	62.75	30.76	62.07	**32.78**	26.43	28.34
		$\tau = 0.25$	60.19	27.85	64.92	30.43	55.32	**30.52**	28.16	27.81
		$\tau = 0.5$	61.01	26.29	62.85	**29.20**	52.77	29.04	28.23	27.44
		$\tau = 0.75$	62.06	24.26	62.58	24.47	50.78	**25.34**	8.58	11.45

Table 3. Explainability-first experiments on COCO validation set.

Backbone	Guidance	Setting	Lambda							
			$\lambda = 10^{-4}$		$\lambda = 5 \times 10^{-4}$		$\lambda = 10^{-3}$		$\lambda = 5 \times 10^{-3}$	
			F1	EPG	F1	EPG	F1	EPG	F1	EPG
ResNet	Evidence	EPG	67.70	22.25	67.81	30.50	67.62	33.45	66.29	**36.85**
		L1	68.10	21.39	67.71	25.99	67.95	27.75	67.81	**30.83**
		PCE	65.94	**18.13**	60.28	17.80	65.46	17.04	55.91	16.61
	Adaptive	$\tau = 0.0$	68.99	25.73	66.53	29.28	65.46	**29.73**	56.02	28.16
		$\tau = 0.25$	68.97	28.94	64.10	**30.78**	62.93	30.65	42.74	28.81
		$\tau = 0.5$	68.53	26.15	65.03	**27.64**	61.55	27.54	49.09	24.95
		$\tau = 0.75$	68.10	21.11	64.97	**22.66**	63.33	22.45	54.40	20.32
B-Cos	Evidence	EPG	66.81	20.53	66.90	25.34	66.80	29.77	65.08	**34.69**
		L1	66.72	20.08	66.72	22.61	66.74	24.11	66.62	**27.22**
		PCE	68.21	**17.80**	69.83	16.21	67.72	15.28	63.49	13.23
	Adaptive	$\tau = 0.0$	68.73	21.13	68.22	25.70	67.79	27.43	63.04	**29.94**
		$\tau = 0.25$	68.50	22.55	67.86	26.66	66.95	28.38	56.60	**30.59**
		$\tau = 0.5$	68.52	21.31	68.04	24.22	67.26	25.39	56.65	**27.69**
		$\tau = 0.75$	68.60	19.46	68.23	20.10	67.15	20.36	59.67	**20.92**
X-DNN	Evidence	EPG	61.47	28.32	59.18	32.92	59.39	35.54	32.58	**35.78**
		L1	61.24	24.64	59.94	26.83	60.84	**28.18**	62.29	25.62
		PCE	57.58	**22.27**	51.31	12.50	49.79	13.83	25.35	10.97
	Adaptive	$\tau = 0.0$	60.00	28.37	60.64	31.42	61.47	**33.18**	25.74	28.81
		$\tau = 0.25$	59.37	28.35	61.86	**31.20**	54.71	31.19	27.40	28.29
		$\tau = 0.5$	60.46	26.63	62.20	**29.73**	52.20	29.69	27.69	27.98
		$\tau = 0.75$	57.80	24.67	61.82	24.97	49.85	**25.86**	8.38	11.58

Table 4. Explainability-first experiments on COCO test set.

Backbone	Guidance	Setting	Lambda							
			$\lambda = 10^{-4}$		$\lambda = 5 \times 10^{-4}$		$\lambda = 10^{-3}$		$\lambda = 5 \times 10^{-3}$	
			F1	EPG	F1	EPG	F1	EPG	F1	EPG
ResNet	Evidence	EPG	**79.29**	24.99	78.77	30.42	78.96	29.79	79.43	34.80
		L1	78.89	25.47	79.72	26.57	**79.89**	28.06	78.77	33.82
		PCE	78.79	24.35	78.79	25.92	**78.91**	25.01	78.30	25.54
	Adaptive	$\tau = 0.0$	79.07	25.21	79.29	28.25	**79.34**	27.87	79.16	31.28
		$\tau = 0.25$	78.78	26.34	78.93	26.68	**79.34**	28.79	79.32	32.41
		$\tau = 0.5$	**79.34**	25.30	79.28	26.43	78.82	28.76	78.66	29.72
		$\tau = 0.75$	79.18	24.37	**79.39**	24.64	78.80	25.14	78.86	27.15
B-Cos	Evidence	EPG	79.89	26.78	79.89	26.80	**79.93**	26.84	79.81	28.83
		L1	79.91	26.78	79.91	26.81	**79.94**	26.84	79.87	28.17
		PCE	79.95	26.76	79.76	28.83	79.98	29.11	**80.49**	26.98
	Adaptive	$\tau = 0.0$	79.89	26.78	79.93	26.80	79.95	26.82	**80.02**	27.00
		$\tau = 0.25$	79.91	26.78	79.93	26.83	79.94	26.89	**80.07**	27.35
		$\tau = 0.5$	79.89	26.78	79.91	26.81	79.96	26.85	**80.12**	27.16
		$\tau = 0.75$	79.90	26.78	79.92	26.79	**79.93**	26.80	79.89	26.90
X-DNN	Evidence	EPG	**73.37**	37.32	72.56	39.97	72.44	39.82	72.34	42.37
		L1	72.45	34.29	72.77	37.25	73.53	35.94	**73.89**	39.51
		PCE	73.88	32.86	73.80	32.16	**74.91**	30.81	73.87	29.62
	Adaptive	$\tau = 0.0$	**72.88**	35.81	72.86	37.14	72.86	37.65	71.94	38.35
		$\tau = 0.25$	**73.31**	37.02	72.74	38.69	72.06	37.13	71.84	40.94
		$\tau = 0.5$	**73.53**	35.94	72.89	35.74	72.76	37.92	72.51	36.52
		$\tau = 0.75$	72.90	36.00	**73.39**	34.65	72.47	36.07	72.83	35.65

Table 5. Performance-first experiments on VOC validation set.

Backbone	Guidance	Setting	Lambda							
			$\lambda = 10^{-4}$		$\lambda = 5 \times 10^{-4}$		$\lambda = 10^{-3}$		$\lambda = 5 \times 10^{-3}$	
			F1	EPG	F1	EPG	F1	EPG	F1	EPG
ResNet	Evidence	EPG	79.51	25.84	79.55	29.86	79.15	29.22	**79.83**	34.09
		L1	79.71	25.53	79.68	26.56	80.04	28.20	**80.23**	31.36
		PCE	79.47	24.77	**79.50**	24.80	79.21	24.31	78.44	24.51
	Adaptive	$\tau = 0.0$	79.67	25.17	79.85	27.56	**80.04**	27.46	79.29	31.56
		$\tau = 0.25$	**79.81**	24.83	79.57	27.35	79.40	28.32	79.61	32.83
		$\tau = 0.5$	79.43	25.52	**79.44**	27.19	79.39	28.01	79.37	29.09
		$\tau = 0.75$	79.55	25.03	**79.96**	25.16	79.12	25.24	79.12	26.60
B-Cos	Evidence	EPG	80.46	26.90	80.44	26.93	80.44	26.96	**80.48**	27.20
		L1	80.46	26.90	80.46	26.93	**80.47**	26.97	80.46	27.24
		PCE	80.45	26.89	80.43	26.85	80.21	28.66	**81.28**	26.95
	Adaptive	$\tau = 0.0$	80.46	26.90	80.45	26.92	80.45	26.94	**80.47**	27.12
		$\tau = 0.25$	80.46	26.91	80.45	26.96	80.20	27.78	**80.47**	27.47
		$\tau = 0.5$	80.46	26.91	80.46	26.94	**80.46**	26.98	80.41	27.29
		$\tau = 0.75$	**80.48**	26.90	80.46	26.91	80.42	26.93	80.33	27.02
X-DNN	Evidence	EPG	**74.56**	37.36	73.68	39.64	73.63	40.68	73.58	43.26
		L1	73.54	35.20	**74.29**	36.31	74.19	37.44	74.27	38.96
		PCE	75.11	33.51	74.71	31.98	**75.53**	31.35	75.19	30.72
	Adaptive	$\tau = 0.0$	73.50	36.70	**73.89**	37.72	73.66	39.02	73.05	41.91
		$\tau = 0.25$	**74.41**	37.44	74.02	39.49	73.19	38.01	72.88	42.02
		$\tau = 0.5$	**75.26**	36.42	74.89	36.93	73.88	39.08	73.15	37.85
		$\tau = 0.75$	74.21	36.79	74.09	36.20	**74.60**	37.24	73.89	36.41

Table 6. Performance-first experiments on VOC test set.

Backbone	Guidance	Setting	Lambda							
			$\lambda = 10^{-4}$		$\lambda = 5 \times 10^{-4}$		$\lambda = 10^{-3}$		$\lambda = 5 \times 10^{-3}$	
			F1	EPG	F1	EPG	F1	EPG	F1	EPG
ResNet	Evidence	EPG	78.93	27.75	78.77	30.42	77.97	35.75	76.67	**37.88**
		L1	77.77	26.50	78.55	29.77	77.58	31.82	78.05	**35.37**
		PCE	77.77	25.07	78.79	**25.92**	75.50	25.85	76.52	25.80
	Adaptive	$\tau = 0.0$	77.86	26.76	77.87	28.79	78.22	29.82	77.41	**32.84**
		$\tau = 0.25$	76.97	28.66	77.80	29.61	77.60	31.27	78.85	**34.62**
		$\tau = 0.5$	69.45	27.85	75.34	29.06	78.20	29.29	78.45	**30.46**
		$\tau = 0.75$	78.43	24.88	77.17	26.54	78.36	26.87	78.60	**27.42**
B-Cos	Evidence	EPG	78.24	33.25	77.91	36.35	77.39	39.35	73.57	**46.33**
		L1	78.10	32.79	77.92	34.29	78.86	34.89	77.58	**39.05**
		PCE	78.75	**31.26**	79.61	29.78	79.98	29.11	80.49	26.98
	Adaptive	$\tau = 0.0$	78.04	33.33	77.45	35.82	76.78	37.81	74.14	**41.90**
		$\tau = 0.25$	77.70	33.77	78.48	35.43	77.34	37.28	73.09	**39.75**
		$\tau = 0.5$	78.20	32.78	78.79	33.78	77.41	35.32	75.49	**36.26**
		$\tau = 0.75$	77.92	32.30	77.95	32.29	78.57	32.33	78.03	**32.76**
X-DNN	Evidence	EPG	71.89	38.79	70.46	44.51	69.34	47.25	70.38	**48.36**
		L1	71.83	34.91	72.61	37.85	72.96	36.72	72.78	**41.87**
		PCE	69.74	**34.33**	70.73	33.77	73.17	34.01	70.90	33.06
	Adaptive	$\tau = 0.0$	71.27	36.36	70.57	42.29	72.07	40.76	70.87	**43.31**
		$\tau = 0.25$	71.63	38.30	72.74	38.69	70.05	39.79	66.50	**42.49**
		$\tau = 0.5$	72.15	36.76	70.32	37.32	70.95	38.35	71.53	**38.69**
		$\tau = 0.75$	72.90	36.00	70.56	34.91	72.06	36.31	72.45	**36.51**

Table 7. Explainability-first experiments on VOC validation set.

Backbone	Guidance	Setting	Lambda							
			$\lambda = 10^{-4}$		$\lambda = 5 \times 10^{-4}$		$\lambda = 10^{-3}$		$\lambda = 5 \times 10^{-3}$	
			F1	EPG	F1	EPG	F1	EPG	F1	EPG
ResNet	Evidence	EPG	79.37	28.47	78.26	30.39	78.98	34.81	77.21	**36.48**
		L1	78.81	27.17	78.77	29.75	78.38	30.91	78.86	**35.06**
		PCE	78.39	25.14	79.05	**25.77**	76.41	25.59	77.05	25.42
	Adaptive	$\tau = 0.0$	78.96	26.39	78.08	28.49	78.56	29.93	77.98	**32.95**
		$\tau = 0.25$	77.80	27.96	78.11	30.40	78.38	31.48	79.13	**34.97**
		$\tau = 0.5$	70.51	27.59	75.70	29.02	78.42	28.80	76.82	**29.73**
		$\tau = 0.75$	68.82	26.04	78.21	26.41	78.60	26.42	79.27	**26.69**
B-Cos	Evidence	EPG	78.46	33.27	78.34	35.97	78.18	39.02	74.52	**45.54**
		L1	78.48	32.74	78.20	34.36	79.02	35.04	78.37	**38.73**
		PCE	78.91	**30.99**	79.98	29.83	79.96	28.98	80.83	27.33
	Adaptive	$\tau = 0.0$	78.41	33.03	77.93	35.52	77.50	37.07	74.65	**40.87**
		$\tau = 0.25$	78.29	33.60	79.32	35.48	78.39	37.31	74.84	**39.35**
		$\tau = 0.5$	78.53	33.05	79.28	34.14	77.45	35.54	77.09	**35.88**
		$\tau = 0.75$	78.40	32.44	78.17	32.62	78.43	32.42	77.85	**32.69**
X-DNN	Evidence	EPG	73.83	39.20	71.53	44.90	70.70	47.25	72.49	**47.77**
		L1	71.97	36.30	73.83	38.65	74.21	37.38	73.84	**42.71**
		PCE	71.51	**34.88**	72.40	33.91	74.58	33.99	71.89	32.73
	Adaptive	$\tau = 0.0$	70.52	36.70	71.61	42.71	73.55	41.50	71.70	**43.31**
		$\tau = 0.25$	73.07	38.70	74.02	39.49	71.25	40.61	68.23	**42.68**
		$\tau = 0.5$	73.52	37.72	72.49	38.80	72.43	**40.01**	72.17	38.61
		$\tau = 0.75$	73.53	37.33	72.92	35.47	73.62	37.87	74.31	**38.19**

Table 8. Explainability-first experiments on VOC validation set.