From Attribution to Action: Jointly ALIGNing Predictions and Explanations

Dongsheng Hong[1] *, **Chao Chen**[2] *, **Yanhui Chen**[1], **Shanshan Lin**[1], **Zhihao Chen**[1], **Xiangwen Liao**[1]†

[1] College of Computer and Data Science, Fuzhou University, China
[2] School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen), China

Abstract

Explanation-guided learning (EGL) has shown promise in aligning model predictions with interpretable reasoning, particularly in computer vision tasks. However, most approaches rely on external annotations or heuristic-based segmentation to supervise model explanations, which can be noisy, imprecise and difficult to scale. In this work, we provide both empirical and theoretical evidence that low-quality supervision signals can degrade model performance rather than improve it. In response, we propose ALIGN, a novel framework that jointly trains a classifier and a masker in an iterative manner. The masker learns to produce soft, task-relevant masks that highlight informative regions, while the classifier is optimized for both prediction accuracy and alignment between its saliency maps and the learned masks. By leveraging high-quality masks as guidance, ALIGN improves both interpretability and generalizability, showing its superiority across various settings. Experiments on the two domain generalization benchmarks, VLCS and Terra Incognita, show that ALIGN consistently outperforms six strong baselines in both in-distribution and out-of-distribution settings. Besides, ALIGN also yields superior explanation quality concerning sufficiency and comprehensiveness, highlighting its effectiveness in producing accurate and interpretable models.

1 Introduction

To provide transparent and trustworthy explanations for the decisions made by deep neural networks, **Explanation-Guided Learning** (EGL) (Gao et al. 2024) integrates explanation signals (e.g., human-provided *masks*) into the training process to align model reasoning with interpretable semantics. These masks typically highlight regions of interest that correspond to task-relevant, informative components in the input (e.g., objects or salient structures). For example, CARE (Zhuang et al. 2019), GRADIA (Gao et al. 2022b), and MAGI (Zhang et al. 2023) penalize attributions to irrelevant regions based on human-annotated masks, thus encouraging the model to focus on relevant features and make more interpretable decisions.

Existing EGL methods rely heavily on *manual annotations*, which are labor-intensive and potentially inaccurate.

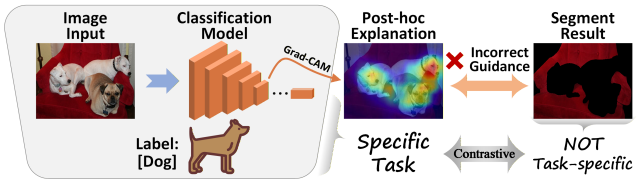


Figure 1: Segmentations as guidance may emphasize background, limiting reliability for task-specific explanations.

Recent approaches (Li et al. 2023b; Guesmi, Aswani, and Shafique 2024) attempt to eliminate manual supervision by enforcing explanation consistency during training. Leveraging segmentation results as guidance can improve explanation quality, but these upstream segmentation models are typically **not task-specific**, and may highlight irrelevant regions, yielding misleading explanations. As shown in Fig. 1, the segmentation mainly captures the surrounding environment rather than the target object `dogs`. This highlights the need for a **task-aware** masker that produces reliable, semantically aligned guidance. Moreover, prior work has primarily emphasized *empirical* performance, with limited *theoretical* insight into their generalization behavior.

In this paper, we first revisit the impact of mask quality on EGL by conducting a preliminary experiment, revealing that imprecise or low-quality masks can hinder prediction accuracy. We further reinforce the need for high-quality, task-relevant masks through a theoretical analysis under the *Probably Approximately Correct* (PAC) learning framework. Specifically, we show that better mask quality leads to tighter generalization bounds under domain shifts (e.g., under out-of-distribution settings), as well as lower in-distribution errors. Building on these insights, we propose **Attribution-Learning Iterative Guidance Network** (ALIGN), a novel framework that jointly trains a mask generator (termed *masker*) and a *classifier* in an iterative manner. Instead of relying on costly and potentially noisy annotations, ALIGN uses a learnable masker to produce soft masks that highlight semantically relevant regions of the input. Simultaneously, the classifier is optimized not only for predictive accuracy but also to align its own saliency maps with the generated masks. By explicitly guiding the model on which features to attend to, ALIGN enhances both inter-

*These authors contributed equally.
†Corresponding author: liaoxw@fzu.edu

pretability and generalizability.

Beyond theoretical justification, we conduct comprehensive experiments on two standard domain generalization benchmarks, VLCS and Terra Incognita. Compared with six state-of-the-art methods, including DRE (Li et al. 2023b), SGDrop (Bertoin et al. 2024), and SGT (Ismail, Corrada Bravo, and Feizi 2021), ALIGN achieves superior predictive performance on both in-distribution and out-of-distribution data. Furthermore, ALIGN produces more meaningful explanations regarding sufficiency and comprehensiveness. Qualitative visualizations and extensive ablation studies further demonstrate that ALIGN yields robust, interpretable predictions.

In summary, the contributions of this paper are threefold:

- We revisit the role of mask quality in EGL, providing both *empirical* evidence and *theoretical* justification that high-quality masks enhance generalization, while poor masks degrade predictive performance.
- We propose **ALIGN**, a novel annotation-free framework that jointly trains a masker and a classifier, aligning model attributions with learned masks to promote interpretability and generalizability.
- Extensive experiments on domain generalization benchmarks demonstrate that ALIGN achieves superior accuracy and explanation quality compared to sota methods, which is supported by visualizations and ablation studies.

2 Related Work

Depending on the source of supervision for explanation alignment, EGL methods can be categorized into human-annotated and annotation-free methods.

Human-Annotated Supervision. Methods such as CARE (Zhuang et al. 2019), GRADIA (Gao et al. 2022b), RES (Gao et al. 2022a), and MAGI (Zhang et al. 2023) use human-labeled masks or saliency cues to align model attributions with semantically meaningful regions. While effective, these methods rely on costly manual annotations, limiting scalability.

Annotation-Free Training. To avoid manual supervision, methods like SGT (Ismail, Corrada Bravo, and Feizi 2021), SMOOT (Karkehabadi, Homayoun, and Sasan 2024), and DRE (Li et al. 2023b) enforce consistency between explanations and predictions using gradient-based masking or stability constraints. However, these methods could rely on insentimantical regions and lack theoretical guarantees.

Besides images, EGL has been extended *graphs*, such as GazeGNN (Wang et al. 2024), GNES (Gao et al. 2021), and GG-NES (Etemadyrad et al. 2024), and *texts* (Li et al. 2023a, 2022). See more relevant works in (Hong et al. 2025).

3 Preliminaries

3.1 Explainable machine learning & Grad-CAM

Given an input-label pair (x, y), where $x \in \mathbb{R}^d$ and $y \in Y$ denote the input features and the corresponding class label, respectively, a classifier $f_\theta : \mathbb{R}^d \to \mathbb{R}^{|Y|}$ is parameterized by θ. For clarity, we omit θ and write $f(x)$. The model's predicted probability for class y is denoted by $f_y(x)$.

The goal of explainable machine learning (XML) is to compute an importance map $\Phi : \mathbb{R}^d \to \mathbb{R}^d$ that reflects the relevance of each input dimension (e.g., pixel) to the model's prediction. Specifically, $\Phi_y(x)$ estimates the contribution of each element in x to the prediction of class y.

In this paper, we adopt Grad-CAM (Selvaraju et al. 2020) as the explanation method. Let A^k denote the activation map of the k-th channel in the selected layer for the input x. Grad-CAM first computes the importance weight α_y^k for each channel using the gradient of the output score $f_y(x)$ with respect to A^k, averaged spatially over all locations (i, j), where Z is the total number of spatial positions:

$$\alpha_y^k = \frac{1}{Z} \sum_{i,j} \frac{\partial f_y(x)}{\partial A_{i,j}^k}. \tag{1}$$

The final explanation map $\Phi_y(x)$ is obtained by a weighted combination of the activation maps followed by a ReLU operation to retain only positive influences:

$$\Phi_y(x) = \text{ReLU}\left(\sum_k \alpha_y^k A^k \right). \tag{2}$$

3.2 Explanation-guided learning

EGL seeks to enhance the model's interpretability by integrating attribution-based signals during training (Ross, Hughes, and Doshi-Velez 2017). More specifically, EGL encourages the alignment of the explanation $\Phi_y(x)$ with a supervision signal, represented by a mask $M(x)$:

$$\mathcal{L}_{egl} = d\big(\Phi_y(x), M(x)\big), \tag{3}$$

where $d(\cdot, \cdot)$ denotes a divergence measure, such as L_1 norm or Binary Cross-entropy (BCE) (Ruby, Yendapalli et al. 2020), and the mask $M(x)$ is typically defined by fixed annotations (Ross, Hughes, and Doshi-Velez 2017; Rieger et al. 2020) or predefined segmented results. By aligning the explanation with these signals, EGL ensures that the model focuses on relevant regions, improving interpretability.

4 Methodology

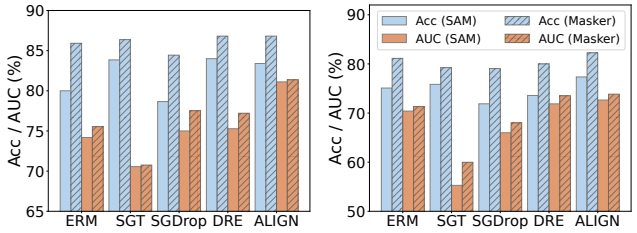
In this section, we begin with a *preliminary experiment* (Sec. 4.1) showing that annotation-free masks generated by off-the-shelf segmentation models can be imprecise and may *hinder* model performance. We then provide *theoretical analyses* (Sec. 4.2) to underscore the importance of learning high-quality, task-relevant masks to enhance both predictive accuracy and explanation reliability. Finally, we introduce **ALIGN** (Attribution-Learning Iterative Guidance Network) in Sec. 4.3, a framework that iteratively trains a classifier f and a masker M to focus on semantically meaningful regions of the input.

4.1 Precise segmentation could improve prediction performance

As previously discussed, obtaining high-quality human-annotated masks for training is both labor-intensive and costly. A common workaround is to leverage large pre-trained segmentation models such as the Segment Anything

Model (SAM) (Kirillov et al. 2023) to generate pseudo-masks in a zero-shot manner. While these models are capable of producing dense segmentations across diverse inputs, their outputs are not specifically optimized for the downstream prediction task and may include irrelevant or noisy regions. Such misaligned or imprecise masks can misguide the learning process, introducing spurious correlations and ultimately degrading predictive performance.

To empirically assess the impact of mask quality, we compare segmentation signals derived from SAM with those generated by our proposed task-driven masker. For both types of masks, we apply a controlled background perturbation: the background, those not highlighted by the mask, are blurred using Gaussian noise, while the foreground remains unaltered. In this way, the saliency signal identified by the mask is preserved, while background distractions are suppressed. Then, well trained classifiers are evaluated on these modified inputs[1]. In principle, a high-quality mask should preserve task-relevant features and thus lead to improved performance when the background is suppressed. Conversely, poor masks may obscure essential features or retain irrelevant ones, resulting in degraded performance.



(a) Result on VOC2007 (b) Result on LabelMe

Figure 2: Impact of mask quality for predictions.

Experiments are conducted on VOC2007 and LabelMe, two subsets of the VLCS benchmark (Fang, Xu, and Rockmore 2013). We evaluate five models: the standard backbone model (ERM), three state-of-the-art explanation-guided learning methods (SGT, SGDrop, and DRE), and our proposed ALIGN framework. As shown in Fig. 2, all models exhibit noticeable improvements in classification accuracy when using masks produced by our task-driven masker, compared to those generated by SAM[2].

The results suggest that segmentation quality plays a critical role in guiding model reasoning and prediction. The consistent gap between two cases indicates that task-driven masker is more effective than static, pre-trained segmentation outputs. It supports our hypothesis: *segmentation signals not optimized for the prediction task can hinder performance*, and precise and task-relevant masks are essential for effective learning. Next, we discuss the necessity of high-quality masks formally.

[1]See Sec. 5 for implementation details and Sec. 5.4 for ablation results when *training* with different masks.

[2]The numerical results can be found in (Hong et al. 2025).

4.2 Theoretical analysis

Basic notations. We consider the domain shift scenario, where the input distribution could change between source domain $\mathcal{S}$ (training) and a target domain $\mathcal{T}$ (testing). Assume that each input x can be decomposed into objective $x^{(obj)}$ and background $x^{(bg)}$ according to segmentation M:

$$x^{(obj)} = M \odot x, \quad x^{(bg)} = (1 - M) \odot x. \tag{4}$$

We compare three hypotheses (models) correspond to different assumptions of mask M.

- f_1: Vanilla models can potentially use *all* features in the image, including spurious features.
- f_2: Perfect guided models use *all relevant* regions to the prediction, e.g., $\frac{df_2(x^{(bg)})}{dx} \approx \mathbf{0}$.
- f_3: Strict guided models utilize a *strict subset* of $x^{(obj)}$, namely $x^{(sub)}$. SAM could lead to a mix of f_1 and f_3.

Bounding generalization error under domain shift. To analyze how model predictions vary across domains, Lemma 1 shows that less invariant features help reduce sensitivity. In other words, f_2 and f_3 are expected to be less sensitive than f_1 when background changes.

Lemma 1 *Let $x_{\mathcal{S}}, x_{\mathcal{T}} \in \mathbb{R}^d$ be two inputs with identical object features and differing background: $x_{\mathcal{S}}^{(obj)} = x_{\mathcal{T}}^{(obj)}$, $x_{\mathcal{S}}^{(bg)} \neq x_{\mathcal{T}}^{(bg)}$. Define the local Lipschitz constant between $x_{\mathcal{S}}$ and $x_{\mathcal{T}}$ for model f as:*

$$\kappa_f(x_{\mathcal{S}}, x_{\mathcal{T}}) := \frac{|f(x_{\mathcal{T}}) - f(x_{\mathcal{S}})|}{\|x_{\mathcal{T}} - x_{\mathcal{S}}\|_2}. \tag{5}$$

If f_1 is sensitive to all features and f_2 satisfies $dh_2(x^{(bg)})/dx \approx \mathbf{0}$, then:

$$\begin{aligned} |f_2(x_{\mathcal{T}}) - f_2(x_{\mathcal{S}})| < |f_1(x_{\mathcal{T}}) - f_1(x_{\mathcal{S}})|, \\ \kappa_{f_2}(x_{\mathcal{S}}, x_{\mathcal{T}}) < \kappa_{f_1}(x_{\mathcal{S}}, x_{\mathcal{T}}). \end{aligned} \tag{6}$$

All the proofs are provided in (Hong et al. 2025). Lemma 1 aligns with existing studies showing that adversarially robust models indeed exhibit smaller input-gradients on perceptually irrelevant features (Srinivas, Bordt, and Lakkaraju 2023; Tsipras et al. 2018).

Based on Lemma 1, we further derive discrepancy bounds on the change in mean squared error (MSE) Δ_{MSE} (Lemma 2) and cross entropy loss Δ_{CE} (Lemma 3) across domains. Both Lemmas imply that f_2, which does not rely on spurious features, shows a lower discrepancy bound, and thus better generalizability, than those use all features (f_1).

Lemma 2 (MSE Discrepancy Bound under Domain Shift) *If the conditional distribution $P(y \mid x^{(obj)})$ is the same across domains $\mathcal{S}$ and $\mathcal{T}$, and both the label values y and prediction $f(x)$ are bounded: $|y| \leq 1$ and $|f(x)| \leq 1$. Then, the MSE discrepancy under domain shifts:*

$$\begin{aligned} \Delta_{MSE} &:= \left|\mathbb{E}_{\mathcal{T}}[(f(x) - y)^2] - \mathbb{E}_{\mathcal{S}}[(f(x) - y)^2]\right| \\ &\leq 4\left|\mathbb{E}_{\mathcal{T}}[f(x)] - \mathbb{E}_{\mathcal{S}}[f(x)]\right| + \left|\mathbb{E}_{\mathcal{T}}[y^2] - \mathbb{E}_{\mathcal{S}}[y^2]\right|. \end{aligned} \tag{7}$$

When $\mathbb{E}_{\mathcal{T}}[y^2] = \mathbb{E}_{\mathcal{S}}[y^2]$ (i.e., no label distribution shift):

$$\Delta_{MSE} \leq 4\left|\mathbb{E}_{\mathcal{T}}[f(x)] - \mathbb{E}_{\mathcal{S}}[f(x)]\right|. \tag{8}$$

Lemma 3 (Cross-Entropy Stability to Small Shifts) *Suppose for every input x and its label y, the difference in predicted probabilities between the two domains is bounded: $\left|f(x_{\mathcal{T}}) - f(x_{\mathcal{S}})\right| \leq \epsilon$, then the absolute difference in cross-entropy risk is bounded as*

$$\Delta_{CE} := |CE_{\mathcal{T}}(f) - CE_{\mathcal{S}}(f)| \leq C \cdot \epsilon, \tag{9}$$

for some constant C that depends only on the range of $f(x)$, e.g., $C = \frac{1}{\min_i f(x)_i}$.

Note that f_3 outperforms f_1 by strictly restricting its input to a subset of the salient features $x^{(obj)}$, as supported by the preceding lemmas. However, as established in Lemma 4, f_2 shows superiority over f_3 concerning in-domain performance, achieving both lower mean squared error and cross-entropy loss by leveraging the complete set of relevant features. These analyses motivate the design of *a learnable masker to guide model training* toward more effective feature selection.

Lemma 4 (In-domain errors for feature inclusion) *Let $f_2^*(x^{(obj)}) := \mathbb{E}[y \mid x^{(obj)}]$ and $f_3^*(x^{(sub)}) := \mathbb{E}[y \mid x^{(sub)}]$ denote the Bayes optimal predictors solely using feature $x^{(obj)}$ and $x^{(sub)}$, respectively. Then, the associated Bayes risks under squared loss satisfy:*

$$\mathbb{E}_{\mathcal{S}}\left[(y - f_2^*(x^{(obj)}))^2\right] \leq \mathbb{E}_{\mathcal{S}}\left[(y - f_3^*(x^{(sub)}))^2\right], \tag{10}$$

and the following cross-entropy error inequality holds:

$$\mathbb{E}_{\mathcal{S}}[-\log f_2(x)_y] \leq \mathbb{E}_{\mathcal{S}}[-\log f_3(x)_y]. \tag{11}$$

Both of which with strict inequality if there exists a feature $x_j \in x^{(obj)} \setminus x^{(sub)}$ such that $x_j \not\perp\!\!\!\perp y \mid x^{(sub)}$.

4.3 The ALIGN framework

Motivated by our empirical observations and theoretical findings, we argue that learning a task-driven, high-quality masker during training is both necessary and beneficial for enhancing model interpretability and generalization. To this end, we propose **Attribution-Learning Iterative Guidance Network (ALIGN)**. As illustrated in Fig. 3, ALIGN jointly learns a masker M and a classifier f in an iterative manner. The classifier f is instantiated as a standard ResNet (He et al. 2016), while the masker M is a lightweight convolutional network that produces a soft mask $M(x) \in [0,1]^d$, highlighting task-relevant regions of the input.

ALIGN consists of two interleaved optimization steps: one for refining the *masker* to generate smooth and semantically meaningful regions that preserve prediction-relevant information, and the other for updating the *classifier* based on both prediction and attribution alignment objectives.

Masker objective. Predefined segmentation models often produce imprecise or task-irrelevant masks, introducing noise that can hinder prediction performance. Thus, we propose learning a task-driven masker M that automatically identifies input regions relevant to the model's prediction. The core goal of M is to highlight informative features while suppressing misleading background regions.

Formally, we encourage the model to predict high confidence on the masked input $x \odot M(x)$ (i.e., retained foreground), and low confidence on the complementary region $x \odot (1 - M(x))$ (i.e., suppressed background). This contrast is captured by the probability distance:

$$dist(x) = f_y(x \odot M(x)) - f_y(x \odot (1 - M(x))), \tag{12}$$

where $f_y(\cdot)$ denotes the predicted probability for the ground-truth class y. The main objective for the masker is:

$$\mathcal{L}_{dist} = MSE(dist(x), 1), \tag{13}$$

which encourages f to rely more on the foreground regions identified by M. Since $dist(x) \in (-1, 1)$, the mean squared error is used to enforce a high contrast.

To ensure that the learned masks are both interpretable and effective, we incorporate two regularization terms. The sparsity loss penalizes unnecessary activations, encouraging compact masks:

$$\mathcal{L}_{sparsity} = \|M(x)\|_1. \tag{14}$$

The smoothness loss enforces spatial continuity by penalizing abrupt changes between neighboring pixels:

$$\begin{aligned}\mathcal{L}_{smooth} = \frac{1}{Z}\sum_{i,j}(&|M_{i,j}(x) - M_{i+1,j}(x)| \\ &+ |M_{i,j}(x) - M_{i,j+1}(x)|).\end{aligned} \tag{15}$$

In summary, the overall loss for the masker M is:

$$\mathcal{L}_{mask} = \mathcal{L}_{dist} + \lambda_1 \mathcal{L}_{sparsity} + \lambda_2 \mathcal{L}_{smooth}, \tag{16}$$

where $\lambda_1 > 0$ and $\lambda_2 > 0$ are hyper-parameters.

Classifier objective. The classifier f is trained to minimize a composite objective:

$$\mathcal{L}_{clf} = \mathcal{L}_{cls} + \lambda_3 \mathcal{L}_{egl} + \lambda_4 \mathcal{L}_{reg}, \tag{17}$$

comprising the following components: (1) Classification loss $\mathcal{L}_{cls} = \text{CE}(f(x), y)$ is the standard cross-entropy loss used to ensure accurate predictions. (2) Explanation guided loss $\mathcal{L}_{egl} = \text{BCE}(\Phi_y(x), M(x))$ aligns the classifier's explanation $\Phi_y(x)$ with the mask $M(x)$ generated by the masker, using binary cross-entropy. (3) Mixup-based regularization: Inspired by (Li et al. 2023b), we use a mixup strategy for both input and explanation. Formally, for two inputs (x_i, y) and (x_j, y) with the same class, a synthetic sample is constructed as:

$$\tilde{x} = \beta x_i + (1 - \beta) x_j, \quad \beta \sim \text{Beta}(\alpha, \alpha). \tag{18}$$

The regularization loss is defined as:

$$\begin{aligned}\mathcal{L}_{reg} = &\|\beta\Phi(x_i) + (1-\beta)\Phi(x_j) - \Phi(\tilde{x})\|_1 \\ &+ \text{CE}(f(\tilde{x}), y) + \|\Phi(\tilde{x})\|_1,\end{aligned} \tag{19}$$

which encourages consistency in both prediction and attribution space, while promoting explanation sparsity.

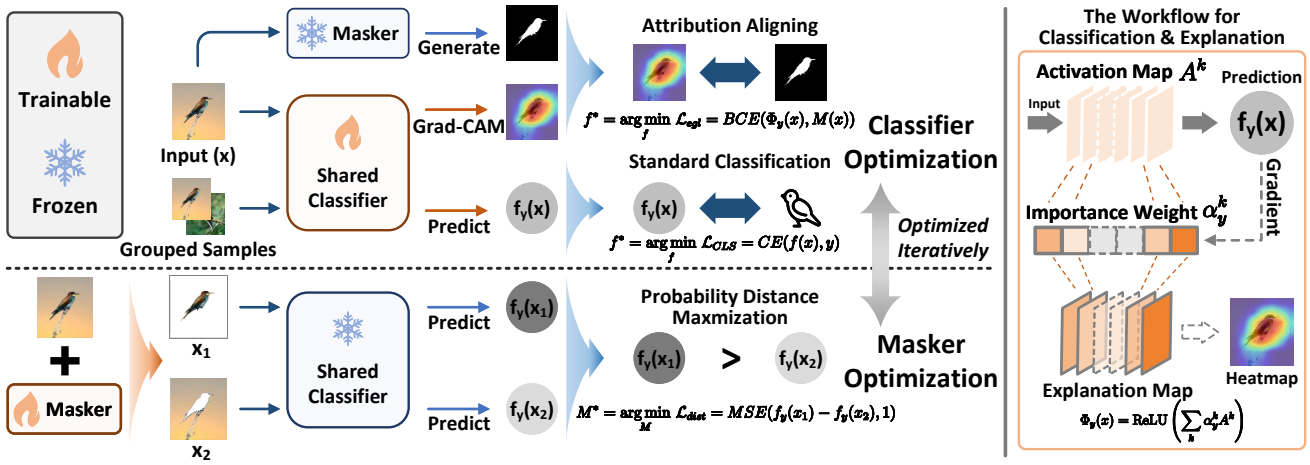


Figure 3: Overview of the proposed ALIGN framework.

Joint optimization. As illustrated in Fig. 3, the ALIGN framework is trained in an alternating optimization scheme. Specifically, during each training iteration, the masker M is updated while keeping the classifier f fixed. Subsequently, the classifier f is updated using the latest output from M, with M parameters held constant.

To mitigate cold-start issues and enhance training stability in the early stages, the process begins with classifier-only training. At this stage, only the standard classification loss $\mathcal{L}_{cls}$ and the mix-up based regularization $\mathcal{L}_{reg}$ are applied, and no explanation supervision is imposed (i.e., $\mathcal{L}_{egl} = 0$). This allows the classifier to form a reliable initial decision without being influenced by potentially unstable masks.

The masker is introduced after a predefined warm-up phase (set to 200 in our paper). The iterative training then continues with both $\mathcal{L}_{mask}$ and $\mathcal{L}_{clf}$ active, gradually aligning the model's reasoning with the learned masks.

5 Experiments

5.1 Datasets and settings

We evaluate the baselines on VLCS and Terra Incognita datasets. VLCS (Fang, Xu, and Rockmore 2013) includes four domains (VOC2007, LabelMe, Caltech101, SUN09) with about 25K images in 5 classes, while Terra Incognita (Xu et al. 2020) comprises four locations (38, 43, 46, 100) totaling around 11K images in 10 categories.

All sub-datasets were split into training, validation, and test sets with a ratio of 6:2:2. The classifier are based on ResNet-18 backbones (He et al. 2016). The masker is a CNN-based model that consists of 3 convolutional blocks: the first two use 3x3 kernels, with ReLU and Batch Normalization, while the final block applies a 1x1 convolution followed by a Sigmoid. *Masker* adopts $\lambda_1 = 10$ and $\lambda_2 = 1$ across all the experiments, and *classifier* uses $\lambda_3 = \lambda_4 = 0.1$ for VLCS, and $\lambda_3 = \lambda_4 = 0.2$ for Terra Incognita. More details can be found in (Hong et al. 2025).

5.2 Baselines and evaluation metrics

We compare ALIGN with several representative explanation guided learning baselines, including ERM (He et al. 2016), IRM (Arjovsky et al. 2019), Mixup (Xu et al. 2020), SGT (Ismail, Corrada Bravo, and Feizi 2021), SGDrop (Bertoin et al. 2024), and DRE (Li et al. 2023b).

In addition to prediction metrics such as **AUC** and **Acc**, the explanations are evaluated by two metrics: **Sufficiency (Suff)** and **Comprehensiveness (Comp)** (Gao et al. 2024).

5.3 Main results

To quantitatively assess **ALIGN**, we conduct experiments on VLCS and Terra Incognita, benchmarking against both standard and explanation-aware baselines. Table 1 summarizes the results for prediction and explanation metrics.

On VLCS, ALIGN achieves the best accuracy and AUC in most domains (VOC2007, CAltech101, SUN09), consistently outperforming all baselines. In terms of explanation quality, ALIGN yields competitive or superior results across both Suff and Comp metrics, indicating that its predictions rely on concise and interpretable evidence.

Similarly, in Terra Incognita dataset, ALIGN again leads in Acc and AUC for most locations (38, 43, 100) and consistently excels in interpretability metrics, demonstrating strong generalization in complex, real-world settings.

These improvements can be attributed to guidances provided by the masker in ALIGN, which effectively identifies task-relevant regions, enabling the classifier to focus on the most relevant input areas. To further assess its effect, we compare alternative masking strategies, including pretrained segmentation and heuristic methods (Sec. 5.4).

In a few cases, ALIGN may not achieve the absolute best performance but remains highly competitive. As noted in Lemma 4, this can occur when the generated mask inadvertently omits a few relevant features, causing slight degradation in in-distribution accuracy. Nonetheless, as shown in Sec. 5.5, ALIGN shows clear advantages under domain shifts by reducing dependence on spurious background cues.

Method	VOC2007				LabelMe				Caltech101				SUN09			
	Acc	AUC	Suff	Comp	Acc	AUC	Suff	Comp	Acc	AUC	Suff	Comp	Acc	AUC	Suff	Comp
ERM	85.35	76.95	17.18	15.64	**80.80**	71.08	1.22	15.21	99.73	96.86	33.71	2.22	80.87	**77.06**	29.42	11.20
IRM	81.45	75.46	26.38	11.83	77.47	68.36	8.49	14.10	98.86	94.71	11.41	**10.34**	81.45	75.46	26.38	11.83
Mixup	85.55	74.38	**10.87**	**24.47**	79.73	69.31	23.21	**17.14**	99.58	94.84	30.26	3.56	78.85	72.57	**17.75**	9.34
SGT	86.64	72.91	17.72	18.55	79.54	60.80	26.43	12.85	99.52	95.43	22.39	2.28	79.23	63.20	32.87	12.36
SGDrop	86.26	78.37	18.02	17.35	79.25	68.81	12.89	13.29	99.94	97.70	12.30	1.68	80.15	75.54	23.98	8.60
DRE	85.61	77.41	16.54	17.02	80.31	73.77	10.48	15.67	99.95	**99.34**	8.61	1.50	81.76	70.96	34.62	8.88
ALIGN	**86.91**	**82.18**	15.08	16.66	80.23	**74.29**	**1.00**	13.32	**99.98**	99.05	**4.71**	2.81	**82.54**	71.16	20.75	**12.54**

Method	Loc_38				Loc_43				Loc_46				Loc_100			
	Acc	AUC	Suff	Comp	Acc	AUC	Suff	Comp	Acc	AUC	Suff	Comp	Acc	AUC	Suff	Comp
ERM	77.89	61.03	**13.78**	35.59	**76.35**	56.25	13.93	43.62	72.69	58.95	14.03	35.74	88.47	77.88	19.12	36.92
IRM	80.50	59.47	11.58	36.47	74.05	63.31	9.95	49.15	72.69	59.46	21.32	38.18	87.53	71.63	13.38	37.58
Mixup	77.41	57.25	14.94	34.14	67.28	55.74	19.65	38.86	73.85	58.78	11.52	36.63	88.72	75.89	16.48	38.79
SGT	77.87	51.27	13.97	29.51	74.73	60.75	12.07	39.43	71.59	58.67	17.37	35.36	86.41	74.87	16.38	29.64
SGDrop	80.26	63.85	21.68	46.41	73.38	56.72	29.82	60.21	74.94	67.31	14.73	35.07	89.43	79.15	**9.43**	36.08
DRE	77.37	65.02	17.58	31.21	74.89	62.64	17.65	**60.60**	73.95	65.73	**10.87**	35.30	88.39	80.15	14.63	31.56
ALIGN	**83.62**	**66.83**	19.75	**51.66**	72.47	**65.05**	**8.20**	44.33	**77.27**	**69.83**	13.94	**40.54**	**90.54**	**84.13**	16.28	**42.88**

Table 1: Overvall performance comparison of ALIGN and baselines arcoss eight sub-datasets. **Bold** values indicate the winner for each metric and dataset. ALIGN consistently achieves the best or competitive performance across all metrics.

5.4 Ablation study: effectiveness of the masker

In Sec. 4.1, we found that "precise segmentation can better improve prediction performance", showing that masks generated by our proposed masker better support model *inference* than those produced by SAM. Here, we further investigate the impact of the masker during *training* by comparing ALIGN against three alternative variants. Each variant modifies the mask generation strategy to assess the contribution of the explanation-guided loss and mask quality:

- **w/o EG** disables the EGL component during training by setting $\lambda_3 = 0$ in Eq. (17).
- **m-SAM** replaces the learned mask $M(x)$ with segmentation maps generated by the pretrained SAM (Kirillov et al. 2023), as fixed targets for saliency alignment.
- **m-Gray** uses grayscale intensity values as the mask, i.e., $M(x) = \text{Gray}(x)$, encouraging the classifier's saliency map to align with low-level pixel brightness.

Results in Table 2 show that: (1) Incorporating external mask signals for EGL (m-SAM, m-Gray) improves performance over the variant without EGL (w/o EG), confirming the value of explanation as supervision. (2) ALIGN, which dynamically generates task-relevant masks, further outperforms all fixed alternatives, demonstrating the advantage of task-driven, end-to-end learning of explanation signals.

5.5 Out-of-distribution generalization

To complement the theoretical analysis presented in Sec. 4.2 and further assess the generalizability of ALIGN, we conduct out-of-distribution (OOD) experiments. In this setting, the model is trained on a source domain and evaluated on remaining target domains *without retraining*. This setup simulates real-world distribution shifts where models must generalize to unseen environments.

Table 3 reports accuracy across all source-target domain pairs in VLCS. ALIGN consistently outperforms all baselines across most OOD settings, often by a substantial margin. We attribute the improved generalization to the trainable masker, which aligns classifier saliency with learned task-relevant regions. This alignment helps identify domain-invariant features, reducing overfitting to spurious source correlations and enhancing performance on unseen domains.

5.6 Case Study

Post-hoc explanation. To better understand how ALIGN improves both prediction accuracy and interpretability, we conduct a case study comparing the attribution patterns of models. Fig. 4 visualizes the Grad-CAM saliency maps for two samples from the VLCS dataset. The ground truth is shown in brackets, with the corresponding predictions from each model are reported below each heatmap.

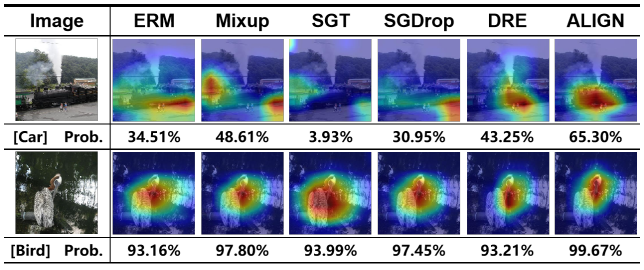


Figure 4: Two case studies from the VLCS dataset.

In the first example (label `car`), baselines exhibit scattered and misaligned attention, often focusing on irrelevant background regions (e.g., sky, smoke, or vegetation), leading to lower confidence. In contrast, ALIGN concentrates its

Method	VOC2007				LabelMe				Caltech101				SUN09			
	Acc	AUC	Suff	Comp	Acc	AUC	Suff	Comp	Acc	AUC	Suff	Comp	Acc	AUC	Suff	Comp
w/o EG	85.61	77.41	16.54	17.02	**80.31**	73.77	10.48	**15.67**	99.95	**99.34**	8.61	1.50	81.76	70.96	34.62	8.88
m-SAM	85.51	80.32	**13.11**	**18.43**	77.66	70.64	10.25	7.14	99.80	98.43	5.63	2.12	79.70	68.51	26.42	8.74
m-Gray	86.90	79.15	13.45	14.31	78.79	69.83	10.09	10.30	99.86	97.42	11.97	1.21	80.13	70.59	32.75	7.53
ALIGN	**86.91**	**82.18**	15.08	16.66	80.23	**74.29**	**1.00**	13.32	**99.98**	99.05	**4.71**	**2.81**	**82.54**	**71.16**	**20.75**	**12.54**

Method	Loc_38				Loc_43				Loc_46				Loc_100			
	Acc	AUC	Suff	Comp	Acc	AUC	Suff	Comp	Acc	AUC	Suff	Comp	Acc	AUC	Suff	Comp
w/o EG	77.37	65.02	17.58	31.21	**74.89**	62.64	17.65	**60.60**	73.95	65.73	**10.87**	35.30	88.39	80.15	14.63	31.56
m-SAM	80.02	**69.72**	**7.57**	30.10	72.02	63.07	13.86	44.13	73.06	67.12	23.09	40.48	89.13	79.51	**11.78**	38.54
m-Gray	78.04	58.34	11.03	30.20	70.09	55.72	11.05	39.75	75.92	**70.16**	17.20	35.42	89.45	82.19	21.09	38.39
ALIGN	**83.62**	66.83	19.75	**51.66**	72.47	**65.05**	**8.20**	44.33	**77.27**	69.83	13.94	**40.54**	**90.54**	**84.13**	16.28	**42.88**

Table 2: Ablation study on the role of the masker in ALIGN. The winners are in **bold**.

Train on	Test on	ERM	SGT	SGDrop	DRE	ALIGN
VOC2007	LabelMe	59.43	**62.12**	59.81	51.71	56.89
	Caltech101	98.31	99.48	98.73	99.52	**99.53**
	SUN09	73.02	69.23	71.91	73.74	**74.03**
LabelMe	VOC2007	67.27	70.12	61.39	58.53	**73.63**
	Caltech101	90.04	91.92	89.92	89.59	**96.63**
	SUN09	56.36	59.58	50.03	51.35	**61.36**
Caltech101	VOC2007	51.68	41.53	42.10	**53.11**	49.60
	LabelMe	38.89	33.68	33.32	41.64	**42.24**
	SUN09	42.03	34.21	37.61	43.87	**48.04**
SUN09	VOC2007	64.65	66.56	63.51	**66.76**	61.01
	LabelMe	59.66	57.86	52.82	57.78	**62.07**
	Caltech101	73.37	73.97	65.08	**83.24**	65.77

Table 3: Prediction accuracy on VLCS under OOD setting.

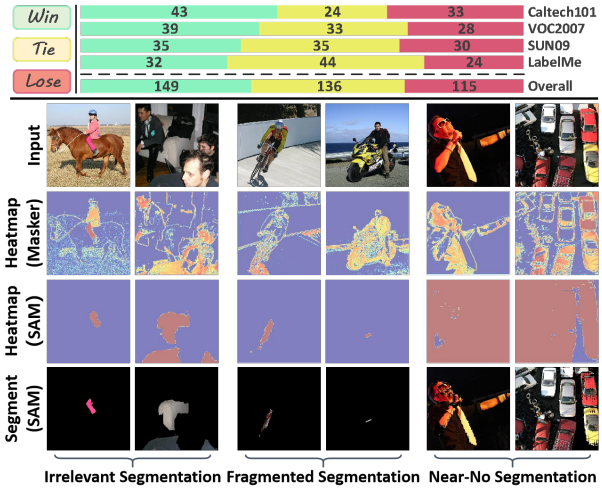


Figure 5: Mask quality evaluation results on VLCS dataset. Upper: human assessment; Lower: visualization examples.

attention around the vehicle, supporting a more accurate prediction (65.30%, the highest among all methods). In the second example (label `bird`), most baselines attend broadly to the surrounding area, incorporating unnecessary contextual features (e.g., background trees or shadows). ALIGN, however, focuses consistently on the bird's highly discriminative regions (head and beak), resulting in 99.67% confidence.

Overall, two real examples highlight ALIGN's ability to consistently ground its predictions in semantically meaningful regions, leading to superior performance in both accuracy and interpretability relative to baseline models.

Mask evaluation. To assess mask quality, we conducted a qualitative study using 100 randomly selected cases from each VLCS sub-dataset, including outputs from SAM and our masker. Four independent volunteers performed pairwise comparisons, labeling each case as Win, Tie, or Lose based on the masker's performance against SAM. Besides, Fig. 5 also visualizes representative heatmaps to compare the spatial attention and segmentation consistency.

From the human evaluation, our masker consistently outperformed SAM, especially on Caltech101 and VOC2007, achieving 149 wins versus 115 losses (**Win/Lose ratio** $\approx$ **1.30**), demonstrating superiority. Visualization analysis further revealed that SAM could produce (1) *irrelevant*, (2) *fragmented*, or (3) *near-empty* segmentations, while our masker precisely highlighted task-relevant target regions. These results suggest that SAM, being *task-agnostic*, generates suboptimal masks when used directly as guidance, whereas our *task-specific* masker yields more reliable and semantically aligned masks for downstream applications.

6 Conclusion

We propose ALIGN, a novel EGL framework that iteratively trains a masker alongside the classifier, effectively constraining the model's attention to object-relevant regions. This end-to-end approach not only enhances performance under domain shifts, but also yields more faithful and interpretable explanations, as demonstrated by both quantitative and qualitative evaluations. In future work, we plan to extend ALIGN to multi-object scenarios and explore alternative explanation mechanisms to further enrich interpretability.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62476060). Chao Chen was also supported by the National Key Research and Development Program of China (No. 2023YFB3106504) and Pengcheng-China Mobile Jointly Funded Project (No. 2024ZY2B0050). We appreciate all the co-workers' constructive comments, which significantly contributed to the development of this paper.

References

Arjovsky, M.; Bottou, L.; Gulrajani, I.; and Lopez-Paz, D. 2019. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*.

Bertoin, D.; Sanchez, E. H.; Zouitine, M.; and Rachelson, E. 2024. The Overfocusing Bias of Convolutional Neural Networks: A Saliency-Guided Regularization Approach. *arXiv preprint arXiv:2409.17370*.

Etemadyrad, N.; Gao, Y.; Manoj Pudukotai Dinakarrao, S.; and Zhao, L. 2024. Global explanation supervision for Graph Neural Networks. *Frontiers in big Data*, 7: 1410424.

Fang, C.; Xu, Y.; and Rockmore, D. N. 2013. Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In *ICCV*.

Gao, Y.; Gu, S.; Jiang, J.; Hong, S. R.; Yu, D.; and Zhao, L. 2024. Going beyond xai: A systematic survey for explanation-guided learning. *ACM Computing Surveys*.

Gao, Y.; Sun, T.; Bhatt, R.; Yu, D.; Hong, S.; and Zhao, L. 2021. Gnes: Learning to explain graph neural networks. In *ICDM*, 131–140. IEEE.

Gao, Y.; Sun, T. S.; Bai, G.; Gu, S.; Hong, S. R.; and Liang, Z. 2022a. Res: A robust framework for guiding visual explanation. In *SIGKDD*, 432–442.

Gao, Y.; Sun, T. S.; Zhao, L.; and Hong, S. R. 2022b. Aligning eyes between humans and deep neural network through interactive attention alignment. *HCI*, 1–28.

Guesmi, A.; Aswani, N. S.; and Shafique, M. 2024. Exploring the Interplay of Interpretability and Robustness in Deep Neural Networks: A Saliency-Guided Approach. In *ICIPCW*, 4066–4072. IEEE.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

Hong, D.; Chen, C.; Chen, Y.; Lin, S.; Chen, Z.; and Liao, X. 2025. From Attribution to Action: Jointly ALIGNing Predictions and Explanations. *arXiv preprint arXiv:2511.06944*.

Ismail, A. A.; Corrada Bravo, H.; and Feizi, S. 2021. Improving deep learning interpretability by saliency guided training. *NeurIPS*, 34: 26726–26739.

Karkehabadi, A.; Homayoun, H.; and Sasan, A. 2024. SMOOT: Saliency guided mask optimized online training. In *DCAS*, 1–6. IEEE.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; Dollár, P.; and Girshick, R. 2023. Segment Anything. *arXiv:2304.02643*.

Li, L. H.; Hessel, J.; Yu, Y.; Ren, X.; Chang, K.-W.; and Choi, Y. 2023a. Symbolic chain-of-thought distillation: Small models can also" think" step-by-step. *arXiv preprint arXiv:2306.14050*.

Li, S.; Chen, J.; Shen, Y.; Chen, Z.; Zhang, X.; Li, Z.; Wang, H.; Qian, J.; Peng, B.; Mao, Y.; et al. 2022. Explanations from large language models make small reasoners better. *arXiv preprint arXiv:2210.06726*.

Li, T.; Qiao, F.; Ma, M.; and Peng, X. 2023b. Are Data-driven Explanations Robust against Out-of-distribution Data? In *CVPR*, 3821–3831.

Rieger, L.; Singh, C.; Murdoch, W.; and Yu, B. 2020. Interpretations are useful: penalizing explanations to align neural networks with prior knowledge. In *ICML*, 8116–8126.

Ross, A. S.; Hughes, M. C.; and Doshi-Velez, F. 2017. Right for the right reasons: Training differentiable models by constraining their explanations. *IJCAI*.

Ruby, U.; Yendapalli, V.; et al. 2020. Binary cross entropy with deep learning technique for image classification. *Int. J. Adv. Trends Comput. Sci. Eng*, 9(10).

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2020. Grad-CAM: visual explanations from deep networks via gradient-based localization. *International journal of computer vision*, 128: 336–359.

Srinivas, S.; Bordt, S.; and Lakkaraju, H. 2023. Which models have perceptually-aligned gradients? an explanation via off-manifold robustness. *NeurIPS*, 36: 21172–21195.

Tsipras, D.; Santurkar, S.; Engstrom, L.; Turner, A.; and Madry, A. 2018. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*.

Wang, B.; Pan, H.; Aboah, A.; Zhang, Z.; Keles, E.; Torigian, D.; Turkbey, B.; Krupinski, E.; Udupa, J.; and Bagci, U. 2024. Gazegnn: A gaze-guided graph neural network for chest x-ray classification. In *WACV*, 2194–2203.

Xu, M.; Zhang, J.; Ni, B.; Li, T.; Wang, C.; Tian, Q.; and Zhang, W. 2020. Adversarial domain adaptation with domain mixup. In *AAAI*, volume 34, 6502–6509.

Zhang, Y.; Gu, S.; Gao, Y.; Pan, B.; Yang, X.; and Zhao, L. 2023. Magi: Multi-annotated explanation-guided learning. In *ICCV*, 1977–1987.

Zhuang, J.; Cai, J.; Wang, R.; Zhang, J.; and Zheng, W. 2019. Care: Class attention to regions of lesion for classification on imbalanced data. In *International Conference on Medical Imaging with Deep Learning*, 588–597. PMLR.