Where Not to Learn: Prior-Aligned Training with Subset-based Attribution Constraints for Reliable Decision-Making

Anonymous authors
Paper under double-blind review

Abstract

Reliable models should not only predict correctly, but also base their decisions on acceptable evidence. However, conventional supervised learning typically provides only class-level labels, allowing models to achieve high accuracy by exploiting shortcut correlations rather than intended decision evidence. Human priors, such as bounding boxes or target interface elements, can help constrain such behavior, but aligning model evidence with these priors remains challenging because learned decision evidence often diverges from human perception. In this work, we study attribution-guided human-prior alignment with subset-selection-based attribution. Motivated by prior deletion and insertion evaluations showing that subset-selection attribution can identify compact decision-supporting regions, we use it as a training-time signal to expose the model's decision evidence. When the top-attributed evidence deviates substantially from the prior region, we penalize off-prior reliance and encourage the model to shift its evidence toward the intended regions. This yields a selective prior-constrained objective that avoids uniformly suppressing all non-prior regions. We validate our method on both image classification and click decision tasks in MLLM-based GUI agents. Across discriminative classification and autoregressive decision-making settings, our method improves task accuracy while enhancing attribution reasonability.

1 Introduction

Machine learning has recently achieved remarkable progress, with large-scale vision and multimodal models delivering strong performance across a wide range of tasks (Li et al., 2025c;b). As these models are increasingly deployed in real-world applications, reliability becomes a central concern (Kuznietsov et al., 2024). Reliable models should not only predict correctly, but also rely on acceptable and task-relevant evidence. However, during training, models may learn shortcut correlations (Geirhos et al., 2020; Kauffmann et al., 2025), leading to seemingly correct outputs that are supported by inappropriate evidence and can fail unpredictably in safety-critical or interactive settings, as shown in Fig. 1.

Standard supervised learning typically provides only class-level supervision, specifying what the correct output should be while leaving the decision evidence largely unconstrained (D'Amour et al., 2022; Geirhos et al., 2018; Rosenfeld et al., 2021). As a result, even large-scale models can be driven to rely on the easiest or most statistically salient correlations rather than the intended causal or semantically meaningful features (D'Amour et al., 2022; Turpin et al., 2023). Human priors can mitigate this issue by constraining what a reasonable decision should rely on. Here, human priors refer to human-recognizable cues about which input components (e.g., objects/regions/attributes) should be relied on for the prediction, typically provided as weak supervision such as sparse clicks, bounding boxes, or saliency annotations. However, aligning models to such priors remains difficult, because model representations and internal decision processes often diverge from human perception (Feather et al., 2019; Poursabzi-Sangdeh et al., 2021; Ngo et al., 2024).

Attribution methods (Chen et al., 2024; 2026) aim to expose the input evidence that a trained model relies on for its predictions. In this work, we follow an intervention-based notion of attribution faithfulness, where important regions should cause larger prediction changes when removed and should better preserve the prediction when retained or inserted. Prior attribution studies (Chen et al., 2024; 2025a; 2026) have shown

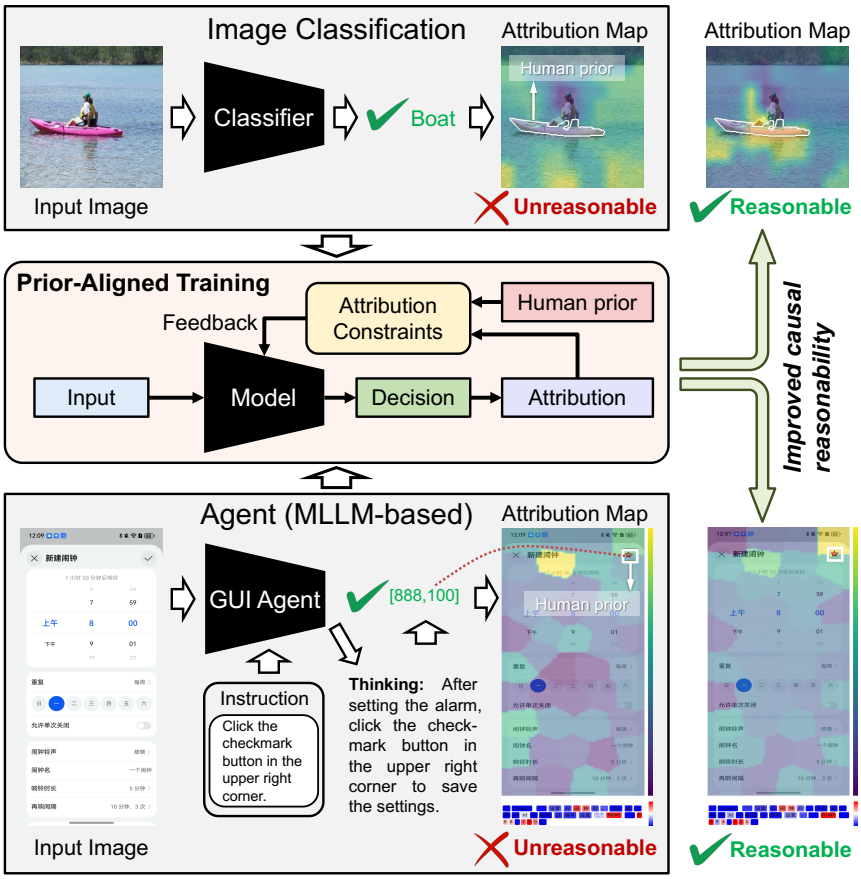

Figure 1: Correct outputs do not guarantee reasonable decision evidence: both a classifier and an MLLM-based GUI agent can succeed while their attribution maps violate human priors. We propose attribution-guided, prior-constrained alignment training to push evidence toward human-prior regions and improve causal reasonableness and decision reliability.

that subset-selection-based attribution performs strongly under such deletion and insertion evaluations, suggesting that it can identify compact decision-supporting evidence. However, these attribution methods are primarily post-hoc explanations; how to use such attribution signals to correct unreasonable model evidence during training remains less explored. Some works (Rao et al., 2023) leverage attribution signals for targeted model correction. RRR (Ross et al., 2017) and XIL (Schramowski et al., 2020) improve decision rationality and accuracy by suppressing gradients outside human-prior regions at the input/feature level, while MEGL (Zhang et al., 2024) encourages feature activation maps to align with human annotations to learn more plausible evidence. However, these methods (i) rely on low-faithfulness attributions that may fail to capture the true decision evidence, and (ii) enforce hard, uniform suppression/enhancement (e.g., pushing all non-prior regions toward zero), ignoring that different regions may contribute unequally.

Motivated by these challenges, we study how faithful subset-selection attribution can be used as an effective training signal for human-prior alignment. Rather than proposing explanation alignment as a new paradigm, our goal is to instantiate attribution-guided alignment with subset-selection-based attribution and examine whether more faithful decision evidence leads to more reliable prior-constrained training across different prediction settings, as shown in Fig. 1. We represent human priors as expected input regions, such as object bounding boxes or interface elements, and use subset-selection-based attribution (Chen et al., 2024; 2026) during training to identify compact decision-sufficient regions. We do not enforce a direct, global alignment to human priors. Instead, we penalize reliance on non-prior regions only when the model's most salient attributed evidence deviates from the prior. When the top-attributed evidence is consistent with human

expectations, we refrain from intervening and allow other regions to contribute freely. This yields a training objective that guides decision evidence toward the intended regions without sacrificing predictive performance.

We evaluate the proposed framework across both image classification tasks and click decision tasks in multimodal large language model–based GUI (Graphical User Interface) agent (Zhang et al., 2025c) settings. These experiments encompass both conventional discriminative prediction and autoregressive decision-making scenarios. Empirically, human prior alignment consistently improves task accuracy while simultaneously enhancing decision reasonability, indicating that constraining decision evidence can yield models that are not only more interpretable but also more robust and effective. In summary, the contributions of this paper are:

- We revisit attribution-guided human-prior alignment and instantiate it with subset-selection-based attribution, showing that compact decision-sufficient evidence can provide a more reliable training signal than conventional gradient- or activation-based explanations.
- We design a selective prior-constrained objective that intervenes only when the top-attributed decision evidence deviates from the human-prior region, avoiding hard uniform suppression of all non-prior regions.
- We evaluate the proposed instantiation on both image classification and MLLM-based GUI agent click decision tasks, demonstrating improved predictive performance and attribution reasonability across discriminative and autoregressive decision-making settings.

2 Related Work

Attribution technologies aim to explain the decision evidence of a model prediction by assigning relevance to input components such as pixels, regions, or tokens. Existing approaches differ in mechanism, including gradient-based methods (Selvaraju et al., 2020; Zhao et al., 2024; Zhang et al., 2025a; Xing et al., 2025), perturbation-based methods (Petsiuk et al., 2018; Novello et al., 2022), Shapley value–based methods (Lundberg & Lee, 2017; Sun et al., 2023), and attention-based methods (Li et al., 2025a). Despite their empirical success, these methods face a core challenge of faithfulness, namely, whether the attributed evidence reflects the causal factors driving the decision, which has motivated minimal sufficiency formulations that seek the smallest evidence subset preserving the original prediction. Recent subset selection-based methods Chen et al. (2024; 2025b;a; 2026) achieve higher faithfulness than other attribution methods, so we use them to guide model training toward more reasonable decision evidence.

Attribution-guided learning studies how attribution signals can be incorporated into training to shape model behavior beyond output supervision (Gao et al., 2024). Some works encourage sparsity or smoothness by regularizing gradient-based attributions during training, but often at the cost of accuracy (Erion et al., 2021; Han et al., 2021; Pillai et al., 2022). Other works use counterfactual attribution for data augmentation to improve generalization (Chen et al., 2025c;d), but do not directly improve attribution reasonableness. Attribution-based human prior alignment methods can improve the reasonableness of model decisions (Ross et al., 2017; Schramowski et al., 2020; Selvaraju et al., 2019; Zhang et al., 2023). However, they often rely on low-faithfulness attributions (e.g., Grad-CAM (Selvaraju et al., 2020) or LIME (Ribeiro et al., 2016)) to guide training, which may limit reasonability gains when the attributions fail to reflect true decision evidence. In this paper, we constrain training with highly faithful attribution methods (LIMA (Chen et al., 2024; 2025b) and EAGLE (Chen et al., 2026)) and human priors, improving both attribution reasonability and model performance.

3 Preliminaries and Problem Statement

3.1 Subset-selection based Attribution

Attribution methods seek to explain model decisions by quantifying the dependence of a prediction on individual input components. Subset-selection based attribution ranks sub-regions in the entire inputs by iteratively selecting compact decision-supporting subsets. Regions selected earlier are deemed more influential, defined as follows.

Definition 3.1 (Subset-Selection-Based Attribution)**.** Given an input $\mathbf{x}$, a trained model f, and an objective set function $\mathcal{F}(\cdot)$, subset-selection-based attribution sparsifies $\mathbf{x}$ into sub-regions $\mathcal{V} = \{v_1, \ldots, v_n\}$ and produces a ranking over $\mathcal{V}$ by solving

$$\pi = \max_{\pi \in \mathcal{P}(\mathcal{V})} \sum_{r=1}^{|\mathcal{V}|} \mathcal{F}(\pi_{:r}), \tag{1}$$

where π is an ordering of $\mathcal{V}$, $\pi_{:r}$ denotes the prefix set consisting of the first r elements in π, and $\mathcal{P}(\mathcal{V})$ is the set of all permutations of $\mathcal{V}$. This objective can be efficiently optimized via greedy search or its accelerated variants.

From this perspective, attribution is cast as a subset selection problem over $\mathcal{V}$, where decision evidence is characterized by compact, decision-supporting subsets and their induced ordering.

3.2 Problem Statement

Attribution-based constrained training is formulated by introducing an attribution regularization term $\mathcal{L}_{\text{human}}$ that encourages consistency between model attributions and human priors. The resulting optimization objective is

$$\min_{\theta} \mathbb{E}_{(x,y,H)\sim\mathcal{D}} \left[\underbrace{\mathcal{L}_{\text{task}}(f_\theta(x), y)}_{\text{task supervision}} + \underbrace{\lambda\, \mathcal{L}_{\text{human}}(\mathcal{A}(f_\theta(x), y)\,, H)}_{\text{human prior alignment}} \right], \tag{2}$$

where H denotes a human prior associated with sample x, and $\mathcal{A}$ is the attribution method. Importantly, H serves as weak guidance rather than exact causal ground truth, and is used to constrain the model's attributed decision evidence toward human-recognized regions. Such constraints aim to improve the causal rationality of model decisions, thereby enhancing model performance, robustness, and interpretability.

4 Method

This section introduces our attribution-based prior-constrained alignment algorithm. Section 4.1 presents the alignment principle. Section 4.2 then details the loss-function instantiation. Finally, Section 4.3 describes the overall training objective and optimization procedure.

4.1 Evidence-Level Alignment Principle

We align model behavior with human priors by constraining decision evidence rather than internal representations. This relies on subset-selection-based attribution, which identifies compact decision-supporting subsets and their induced ordering, and is more faithful than gradient- or attention-based methods in reflecting the evidence driving model decisions. Let $\mathcal{V}$ denote the set of input sub-regions and H denote a human prior specified over the input space, such as bounding boxes or masks. For a given prediction, attribution ranks regions in $\mathcal{V}$ by their influence on the decision. When the most influential evidence sufficiently overlaps with H, no constraint is imposed. When highly ranked evidence lies largely outside H, the model may rely on unintended cues, which should be discouraged during training.

Alignment is imposed asymmetrically: only off-prior decision evidence is penalized, while evidence consistent with the prior remains unconstrained. This avoids over-regularization and preserves flexibility within human-recognized regions. The same principle applies to both discriminative classification and autoregressive decision-making in MLLM-based GUI agents. In practice, we instantiate a black-box subset-selection–based attribution framework using LIMA Chen et al. (2024; 2025b) for image classification models and EAGLE Chen et al. (2026) for MLLM-based GUI agents. As attribution relies only on model inputs and outputs, the framework generalizes across diverse model architectures.

4.2 Alignment with Subset-based Attribution

We instantiate the prior constrained training using the subset-based attribution framework in Section 3.1, which produces an ordering over sub-regions $\mathcal{V}$ by decision influence. Let $\pi = (v_{\pi_1}, v_{\pi_2}, \ldots, v_{\pi_{|\mathcal{V}|}})$ denote the

ranking over sub-regions $\mathcal{V}$ induced by attribution, where regions appearing earlier are more influential. Let H denote the human prior specified over the input space (e.g., bounding boxes or masks). Since H may not lie in the same discrete space as $\mathcal{V}$, we define an overlap function $\phi(v, H) \in [0, 1]$, which measures the spatial consistency between a region v and the human prior H (e.g., IoU or mask coverage). A region is considered off-prior when $\phi(v, H)$ is small.

Deviation loss: To prevent the most influential attribution region from deviating from the human prior during training, we introduce a *Deviation Loss*. Since subset-selection–based attribution ranks regions according to a set function $\mathcal{F}(\cdot)$, deviations from the prior are addressed by suppressing the contribution of the top-ranked region. Specifically, when the most influential region v_{π_1} exhibits low consistency with the human prior, we reduce its utility score $\mathcal{F}(v_{\pi_1})$ to discourage reliance on this region. The resulting optimization objective is

$$\mathcal{L}_{\text{deviation}} = \sum_{i=1}^{b} \mathcal{F}\left(v_{(i,\pi_1)}\right) \cdot \mathbf{1}\left[\phi(v_{(i,\pi_1)}, H_i) < \tau\right], \tag{3}$$

where b denotes the batch size, $v_{(i,\pi_1)}$ is the most influential attribution region for the i-th sample, τ is a threshold determining consistency with the human prior, and $\mathbf{1}[\cdot]$ is the indicator function. Intuitively, no penalty is applied when the most influential attribution region lies within the human prior. When the primary attribution region falls outside the prior, its explanatory influence should be limited, and the corresponding utility score $\mathcal{F}$ is therefore suppressed.

Redundancy loss: Beyond constraining the primary attribution region, we further regulate *higher-order attribution regions*, referring to all attribution results beyond the top-ranked one. When such higher-order regions fall outside the human prior, their contribution to the model's decision should be limited, as accumulating evidence from unintended regions leads to redundant and potentially spurious decision support. Intuitively, off-prior regions should not provide substantial additional gains once the primary evidence has been identified. Since subset–based attribution constructs decision evidence sequentially via marginal gains of the set function $\mathcal{F}(\cdot)$, we suppress excessive marginal contributions from higher-order off-prior regions. This redundancy loss mitigates the accumulation effect in multi-region combinations, preventing off-prior regions from jointly contributing to the prediction and introducing shortcut cues. The objective is

$$\mathcal{L}_{\text{redundancy}} = \sum_{i=1}^{b} \sum_{r=2}^{k} \mathsf{ReLU}\left(\Delta_{i,r}\right) \cdot \mathbf{1}\left[\phi(v_{(i,\pi_r)}, H_i) < \tau\right], \tag{4}$$

where $\Delta_{i,r} = \mathcal{F}(\pi_{:r}) - \mathcal{F}(\pi_{:r-1})$ denotes the marginal gain contributed by the region at rank r, given the previously selected prefix regions, and k denotes the maximum number of sub-regions considered during attribution.

4.3 Overall Training Objective

We optimize the model using a mixed training objective that alternates between standard task supervision and evidence-level alignment. Specifically, alignment losses are applied only at regular intervals to reduce computational overhead and to avoid over-constraining the model during training.

Formally, let t denote the training step and T the alignment interval. When $t \bmod T = 0$, alignment is applied only to training samples that are correctly predicted by the model. For such samples, the optimization objective is

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda_1 \mathcal{L}_{\text{deviation}} + \lambda_2 \mathcal{L}_{\text{redundancy}}, \tag{5}$$

where $\mathcal{L}_{\text{task}}$ denotes the standard task loss. For samples that are incorrectly predicted, as well as for all steps where $t \bmod T \neq 0$, the model is optimized using only the task loss. This intermittent and conditional alignment strategy ensures that attribution-based constraints are imposed only when the model's predictions are reliable, allowing efficient learning of task-relevant representations while periodically correcting reliance on off-prior decision evidence, leading to stable training and improved generalization. The overall training procedure is summarized in Algorithm 1.

Algorithm 1: Prior constrained training with subset-based attribution

Input: Training data $(\mathbf{x}_i, y_i, H_i)$, alignment interval T, loss weights λ_1, λ_2, attribution length k.
Output: Trained model parameters θ.

```
Initialize model parameters θ;
for t = 1 to T_max do
    Sample a mini-batch {(x_i, y_i, H_i)}_{i=1}^b;
    Compute task loss L_task;
    if t mod T == 0 then
        L_deviation ← 0;
        L_redundancy ← 0;
        for each sample i in the batch do
            if the prediction of sample i is correct then
                Compute top-k attribution ranking π using LIMA or EAGLE;
                Compute deviation loss from the top-ranked region v_{π_1};
                Compute redundancy loss from higher-order regions;
        end
        L_total ← L_task + λ1 L_deviation + λ2 L_redundancy;
    else
        L_total ← L_task;
    end
    Update model parameters θ using ∇L_total;
end
return θ;
```

5 Experiments

5.1 Experimental Setup

Datasets. We evaluate the proposed method on two representative tasks: image classification and a MLLM-based GUI agent clicking task. For image classification, we use two datasets with high-quality object-level annotations. ImageNet-S (Gao et al., 2022) is a curated subset of ImageNet with 919 categories and pixel-level segmentation masks, while Saliency-Bench (Zhang et al., 2025b) is constructed from MS COCO with high-quality object annotations. These datasets enable a challenging evaluation of both predictive performance and attribution faithfulness. For the GUI agent task, UI elements (e.g., buttons and icons) encode human priors over actionable targets. We collect 936 single-step Android clicking tasks with annotations of click locations and UI element bounding boxes, enabling a controlled evaluation of decision rationality in domain-specific MLLMs. The dataset will be released.

Baselines. We compare with representative attribution-based prior alignment baselines, including RRR (Ross et al., 2017), which penalizes input-level gradients (Simonyan et al., 2014) on non-prior regions, XIL (Selvaraju et al., 2019), which suppresses Grad-CAM (Selvaraju et al., 2020) activations outside prior regions at the feature level, and MEGL (Zhang et al., 2024), which aligns Grad-CAM maps with mask annotations using an ℓ_1 loss. For ViT-based architectures, Grad-CAM is replaced with Grad-ECLIP (Zhao et al., 2024).

Implementation Details. For classification models, we compute attributions and perform prior alignment once every 10 training steps. For GUI-agent models, attributions are computed once every 5 steps. The loss balancing coefficients, λ_1 and λ_2, are both set to 0.5. During training, the subset-selection-based attribution sparsifies each image into 50 sub-regions. The attribution search selects at most 10 sub-regions and early-stops once the prediction confidence of the selected subset exceeds 0.8. We apply the attribution-alignment losses only to samples that are correctly predicted with confidence above 0.75, otherwise, the model is trained with the standard task loss only. More details please see the *Appendix*.

5.2 Evaluation on Image Classification

We first validate our method on image classification tasks, where the selected datasets provide both class labels and object masks as human priors. In addition to comparing against direct fine-tuning to assess the benefit of prior supervision, we include attribution-based baselines that adopt different attribution methods and alignment strategies. We report top-1 accuracy and decision rationality measured by Point Game (Zhang et al., 2018), which evaluates whether predictions attend to target objects rather than background regions. Since LIMA (Chen et al., 2024; 2025b) provides the highest attribution faithfulness among existing methods, we adopt LIMA-based attributions for evaluation to ensure a consistent and reliable assessment of decision rationality, regardless of the attribution strategies used during training.

As shown in Table 1, across backbones (CLIP (Radford et al., 2021), ViT (Dosovitskiy et al., 2021), and ResNet (He et al., 2016)) on Saliency-Bench, our method consistently improves Point Game (e.g., from 0.4363 to 0.5463 on ViT) while also increasing top-1 accuracy (e.g., from 0.5150 to 0.5694 on ViT and from 0.6076 to 0.6551 on CLIP). On ImageNet-S, the gains in top-1 accuracy are relatively modest but remain positive (e.g., improve 4.95 points on ViT and 1.74 points on ResNet-101), which we attribute in part to the limited number of training images per category in this subset, while rationality improves where available. We further report top-1 accuracy conditioned on successful Point Game outcomes. Notably, our method yields substantial gains on this metric (e.g., from 0.7093 to 0.8377 on CLIP for ImageNet-S), suggesting that when the model attends to the target object as expected, its predictions become markedly more reliable. Compared with prior-alignment baselines that rely on input gradients or Grad-CAM variants, the improvements are more consistent on rationality-related metrics, indicating a higher effective upper bound when enforcing priors with more faithful attributions.

Table 1: Evaluation of attribution-based prior alignment methods for image classification models on the Saliency-Bench and ImageNet-S datasets. Both model performance (accuracy) and decision rationality are reported, with rationality measured by the Point Game and accuracy conditioned on successful Point Game outcomes.

Datasets	Human Prior	Models	Methods	Attributions	Top-1 Acc.	Top-2 Acc.	Point Game	Top-1 Acc. (PG=1)
Saliency-Bench	Masks	CLIP	Fine-tuning	-	0.6076	0.7847	0.5231	0.9044
			RRR (Ross et al., 2017)	Input Gradient	0.6030	0.7821	0.5253	0.8943
			XIL (Schramowski et al., 2020)	Grad-ECLIP	0.6400	0.7891	0.5327	0.9045
			MEGL (Zhang et al., 2024)	Grad-ECLIP	0.6354	8180	0.5318	0.9004
			Ours	LIMA	**0.6551**	**0.8264**	**0.5648**	**0.9192**
		ViT (base)	Fine-tuning	-	0.5150	0.7350	0.4363	0.7786
			RRR (Ross et al., 2017)	Input Gradient	0.5370	0.7512	0.4509	0.6530
			XIL (Schramowski et al., 2020)	Grad-ECLIP	0.5139	0.6968	0.4397	0.8087
			MEGL (Zhang et al., 2024)	Grad-ECLIP	0.5359	0.7338	0.5145	0.8242
			Ours	LIMA	**0.5694**	**0.7639**	**0.5463**	**0.8519**
		ResNet-101	Fine-tuning	-	0.5498	0.7569	0.6235	0.7694
			RRR (Ross et al., 2017)	Input Gradient	0.5498	0.7604	0.6076	0.7857
			XIL (Schramowski et al., 2020)	Grad-CAM	0.5521	0.7616	0.6725	0.8679
			MEGL (Zhang et al., 2024)	Grad-CAM	0.5451	0.7662	0.6315	0.8344
			Ours	LIMA	**0.5590**	**0.7662**	**0.6984**	**0.8782**
ImageNet-S	Masks	CLIP	Fine-tuning	-	0.7969	0.8888	0.7001	0.7093
			RRR (Ross et al., 2017)	Input Gradient	0.7898	0.8861	0.7051	0.7642
			XIL (Schramowski et al., 2020)	Grad-ECLIP	0.7807	0.8786	0.7535	0.8042
			MEGL (Zhang et al., 2024)	Grad-ECLIP	0.7857	0.8795	0.7556	0.7942
			Ours	LIMA	**0.7974**	**0.8895**	**0.7712**	**0.8377**
		ViT (base)	Fine-tuning	-	0.6713	0.7728	0.8041	0.8762
			RRR (Ross et al., 2017)	Input Gradient	0.6868	0.7912	0.7923	0.8580
			XIL (Schramowski et al., 2020)	Grad-ECLIP	0.6952	0.7971	0.8035	0.8514
			MEGL (Zhang et al., 2024)	Grad-ECLIP	0.6969	0.8024	0.8143	0.8654
			Ours	LIMA	**0.7208**	**0.8087**	**0.8226**	**0.8878**
		ResNet-101	Fine-tuning	-	0.7071	0.8011	0.8453	0.8814
			RRR (Ross et al., 2017)	Input Gradient	0.7073	0.8076	0.8364	0.8532
			XIL (Schramowski et al., 2020)	Grad-CAM	0.7225	0.8182	0.8491	0.8904
			MEGL (Zhang et al., 2024)	Grad-CAM	0.7212	0.8158	0.8303	0.8522
			Ours	LIMA	**0.7245**	**0.8186**	**0.8672**	**0.9040**

Figure 2 shows qualitative results. We visualize LIMA attribution for all models, regardless of the attribution strategy used during training. Notably, image classification inputs may contain multiple co-occurring objects, where spurious or non-target objects can distract the decision evidence. Compared with prior-alignment

baselines, our method yields attributions that are more concentrated on the human-prior target regions, indicating that the resulting predictions rely less on irrelevant objects or background cues.

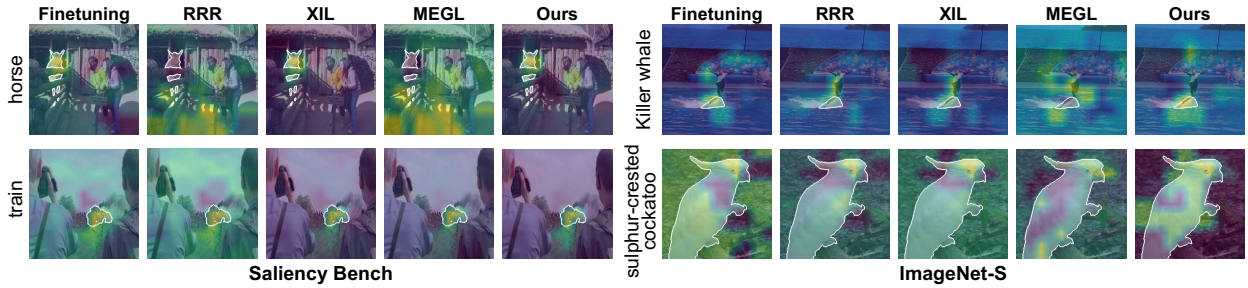

Figure 2: Qualitative comparison on Saliency-Bench and ImageNet-S. For each method, we visualize LIMA-based attributions on the same inputs; white masks indicate human priors (target object regions).

5.3 Ablation Study

Ablation of the components. Table 2 presents ablations of the deviation loss and redundancy loss across different backbones on the Saliency-Bench dataset. Overall, the deviation loss brings consistent gains in both accuracy and decision reasonability. This indicates that explicitly penalizing reliance on off-prior evidence can effectively steer the model to ground its most influential evidence on human-recognized regions, which improves not only prediction performance but also prior-consistent explanations. In contrast, the redundancy loss mainly affects the quality of the explanation: when combined with the deviation loss, it yields an additional (typically mild) improvement in Point Game, while its impact on accuracy is limited. This behavior aligns with its design goal, by accounting for the cumulative effect of selected regions, the redundancy term suppresses repeated/overlapping evidence and encourages more efficient evidence allocation, thereby slightly enhancing decision reasonability.

Table 2: Ablation studies on the deviation loss and the redundancy loss on the Saliency Bench dataset.

Models	Deviation loss	Redundancy loss	Accuracy	Point Game
CLIP	✗	✗	0.6076	0.5231
	✓	✗	0.6525	0.5575
	✓	✓	**0.6551**	**0.5648**
ViT	✗	✗	0.5150	0.4363
	✓	✗	0.5359	0.5238
	✓	✓	**0.5690**	**0.5463**
ResNet	✗	✗	0.5498	0.6235
	✓	✗	0.5535	0.6849
	✓	✓	**0.5590**	**0.6984**

Parameter sensitivity analysis. We conduct a parameter sensitivity study on ImageNet-S using the ViT backbone, focusing on (i) the step interval for applying attribution-based prior constraints during training and (ii) the weighting coefficient of the deviation loss. Figure 3A shows that applying the constraint more frequently (i.e., using a smaller interval) can improve classification accuracy, but enforcing it too often may disrupt optimization of the primary task and degrade performance. Figure 3B indicates that accuracy remains stable for small-to-moderate λ_1, while overly large λ_1 causes a clear drop, suggesting that the deviation loss should be weighted moderately to avoid overwhelming the main objective.

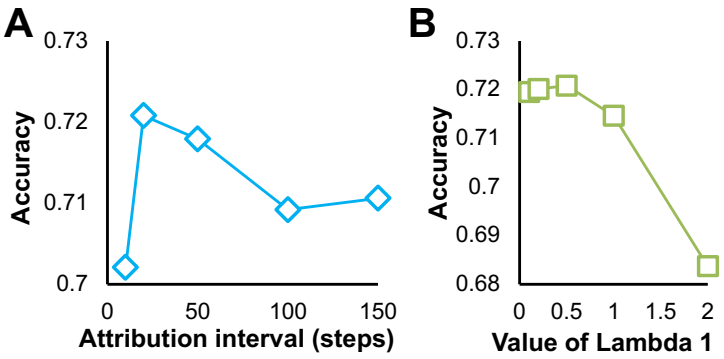

Figure 3: Impact of training hyperparameters on model performance. **A.** Effect of the attribution interval on validation accuracy. **B.** Effect of the loss balancing coefficient λ_1 on validation accuracy.

Robustness for noise. Table 3 compares robustness under Gaussian noise corruption at evaluation time on the Saliency Bench dataset. Specifically, we add random Gaussian noise to validation images and report the resulting accuracy. Our method consistently outperforms standard fine-tuning on the noisy validation set, indicating stronger noise robustness. This suggests that prior-aligned evidence constraints help the model rely on stable, semantically meaningful regions rather than brittle spurious cues, thereby improving robustness to input perturbations.

Table 3: Validation accuracy with Gaussian noise on the Salinecy Bench dataset.

Models	Methods	Clean Val. Accuracy	Noisy Val. Accuracy
CLIP	Fine-tuning	0.6076	0.5995
	Ours	**0.6551**	**0.6157**
ResNet	Fine-tuning	0.5498	0.4167
	Ours	**0.5590**	**0.4572**

5.4 Extension to MLLM-based GUI Agent

Next, we validate our method in a GUI agent setting. We use AgentCPM-GUI (Zhang et al., 2025c), a reasoning-oriented MLLM that produces both `thinking` and a final `decision` (click action). We adopt supervised fine-tuning (SFT) as the primary training paradigm, using data that contains the target decision together with GPT-distilled `thinking` traces. Our goal is to improve the consistency between the decision evidence expressed in `thinking` and the executed `decision` via attribution-based consistency regularization. We employ EAGLE (Chen et al., 2026) for attribution, which is tailored to MLLMs. Since there are no established attribution-alignment baselines for MLLMs in this GUI clicking setup, we mainly evaluate the gains of our method over standard SFT. Note that SFT already injects prior information to some extent, as the training supervision explicitly specifies the click target. The evaluation metrics are described in Appendix A.4.

Table 4 summarizes the results on the GUI agent clicking task. The findings are analyzed from three aspects: functional performance, content understanding with attribution consistency, and decision reliability. With supervised fine-tuning (SFT) and LoRA adaptation, AgentCPM-GUI achieves a click success rate of 84.61% and a distance error of 94.71. After introducing attribution constraints, the click success rate increases to 89.23% (an absolute gain of 4.62%), while the distance error decreases to 78.64 (a relative reduction of 16.96%). These results indicate that our attribution-prior alignment improves task performance and yields more stable clicks by encouraging attention to task-relevant regions. Figure 4 shows some examples.

Table 4: Evaluation on the GUI agent clicking task with AgentCPM-GUI. Standard SFT (LoRA) is compared with attribution-based alignment (LoRA). Task performance is reported by click success rate and distance error, and reliability is measured by Point Game and metrics conditioned on successful Point Game outcomes (click success rate and distance error when PG=1).

Methods	Task Performance		Point Game (↑)	Reliability Metrics	
	Click success rate (↑)	Distance error (↓)		Click success rate (PG=1) (↑)	Distance error (PG=1)
SFT (LoRA)	84.61%	94.71	0.8153	96.22%	7.11
Ours (LoRA)	**89.23%**	**78.64**	**0.8615**	**100%**	**0.0**

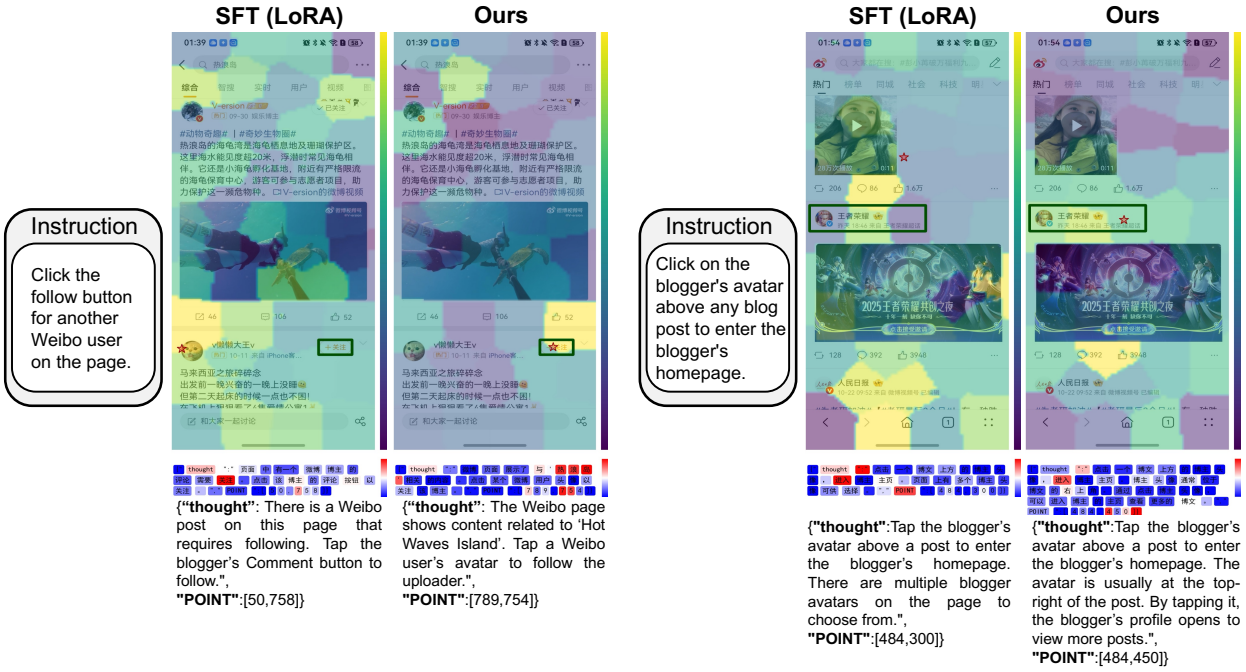

Figure 4: Qualitative GUI agent results comparing SFT with LoRA adaptation and the proposed our method, together with attribution heatmaps, predicted click locations (stars), and human-prior target bounding boxes.

Next, decision rationality is examined for the reasoning agent. EAGLE is used to attribute the generation process without additional annotations. As shown in Table 4, the Point Game (PG) score increases from 0.8153 with SFT to 0.8615 with our method (a relative gain of 5.67%), indicating stronger consistency between reasoning–decision evidence and human-prior target regions. Figure 5 further provides a qualitative comparison: although both methods produce correct clicks, SFT does not always attend to the target region during `thinking`, whereas our method yields more evidence-consistent `thinking` and `decision`.

Reliability is further analyzed on samples where attributions match human priors (PG= 1): as shown in Table 4, the click success rate increases from 96.22% to 100%, and the distance error drops from 7.11 to 0. While the 100% rate may be influenced by the limited evaluation set, the consistent trend indicates that prior-consistent evidence correlates with more robust behavior, and our method increases the coverage of such high-reliability decisions by improving `thinking` and `decision` evidence consistency.

Failure analysis. Next, failure cases are analyzed in Figure 6. In these examples, neither model outputs a `POINT`; instead, both return a `STATUS`. The attribution results indicate that the `thought` from both SFT and our method captures the instruction intent, but the supporting evidence differs: SFT focuses on the

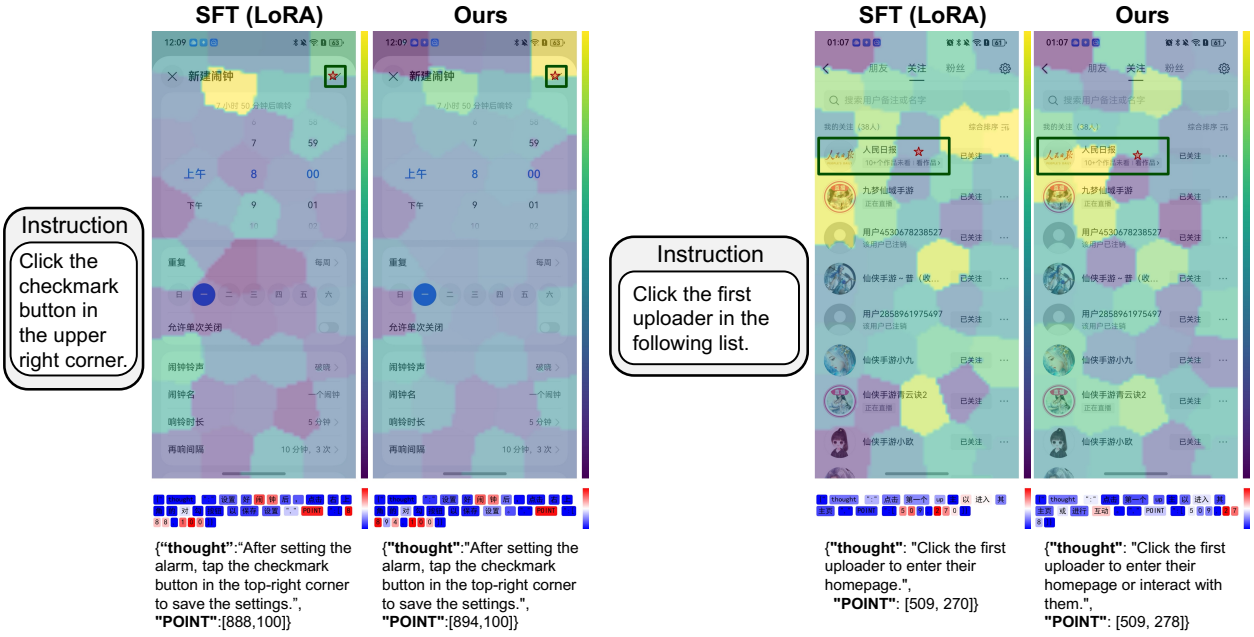

Figure 5: GUI clicking example comparing SFT with LoRA adaptation and our method, showing attribution maps and correct clicks on the target checkmark.

like button region, whereas our method concentrates on the *follow* button. This suggests that our method produces more semantically grounded `thought` evidence, even when the final action format is incorrect.

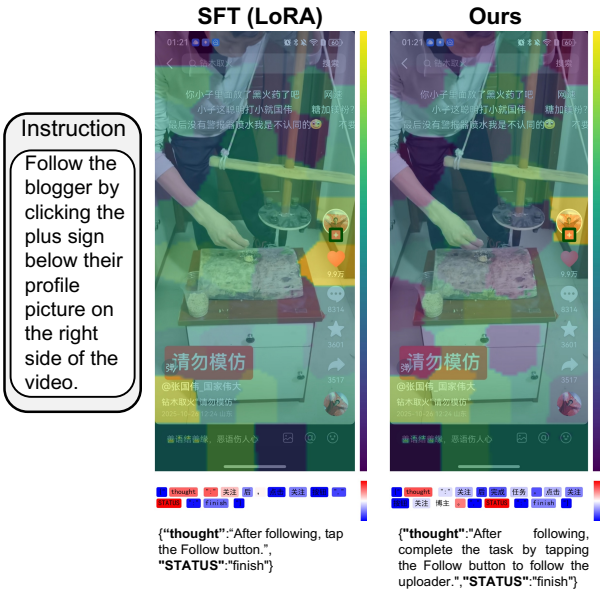

Figure 6: Failure-case comparison on the GUI agent task.

6 Conclusion

In this paper, we argued that reliable models should not only produce correct outputs, but also rely on acceptable, task-relevant evidence. We proposed a prior-aligned training framework that enforces evidence-level constraints using faithful subset-selection attribution. Human priors are encoded as expected input regions (e.g., object masks or UI-element bounding boxes), and the model's decision evidence is explicitly exposed during training via subset-based attributions. Experiments on both image classification and GUI clicking tasks show that prior-constrained attribution alignment consistently improves task performance while simultaneously enhancing decision reasonability. Our results suggest that aligning models with human-recognized evidence provides a practical path toward more causally reasonable decisions, leading to improved robustness and interpretability.

Broader Impact Statement

This work relates to *explanation-guided learning*, a training paradigm that leverages explanation signals as an additional form of supervision beyond labels. A potential positive impact lies in high-stakes and data-scarce domains such as healthcare, where interpretability requirements are stringent and purely label-driven training may amplify spurious correlations. In such settings, explanation-guided learning can provide a practical mechanism to incorporate expert knowledge during training, improve transparency for auditing and debugging, and potentially enhance robustness by discouraging reliance on unintended cues.

Limitations. However, prior signals can be imperfect and may reflect incomplete or biased human assumptions. If used indiscriminately, explanation-guided learning could constrain models in ways that reduce generalization, suppress valid evidence, or introduce systematic biases. Careful design of explanation supervision, validation across diverse populations, and domain-specific safeguards are therefore essential for responsible use, especially in clinical deployment.

References

Ruoyu Chen, Hua Zhang, Siyuan Liang, Jingzhi Li, and Xiaochun Cao. Less is more: Fewer interpretable region via submodular subset selection. In *ICLR*, 2024. 1, 2, 3, 4, 7

Ruoyu Chen, Siyuan Liang, Jingzhi Li, Shiming Liu, Maosen Li, Zhen Huang, Hua Zhang, and Xiaochun Cao. Interpreting object-level foundation models via visual precision search. In *CVPR*, 2025a. 1, 3

Ruoyu Chen, Siyuan Liang, Jingzhi Li, Shiming Liu, Li Liu, Hua Zhang, and Xiaochun Cao. Less is more: Efficient black-box attribution via minimal interpretable subset selection. *arXiv preprint arXiv:2504.00470*, 2025b. 3, 4, 7

Ruoyu Chen, Hua Zhang, Jingzhi Li, Li Liu, Zhen Huang, and Xiaochun Cao. Generalized semantic contrastive learning via embedding side information for few-shot object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(8):6496–6514, 2025c. 3

Ruoyu Chen, Xiaoqing Guo, Kangwei Liu, Siyuan Liang, Shiming Liu, Qunli Zhang, Laiyuan Wang, Hua Zhang, and Xiaochun Cao. Where mllms attend and what they rely on: Explaining autoregressive token generation. In *CVPR*, pp. 17057–17066, 2026. 1, 2, 3, 4, 9

Yannan Chen, Ruoyu Chen, Bin Zeng, Wei Wang, Shiming Liu, Qunli Zhang, Zheng Hu, Laiyuan Wang, Yaowei Wang, and Xiaochun Cao. Did models sufficient learn? attribution-guided training via subset-selected counterfactual augmentation. *arXiv preprint arXiv:2511.12100*, 2025d. 3

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, et al. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226):1–61, 2022. 1

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 7

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. 3

Jenelle Feather, Alex Durango, Ray Gonzalez, and Josh McDermott. Metamers of neural networks reveal divergence from human perceptual systems. In *NeurIPS*, pp. 10078–10089, 2019. 1

Shanghua Gao, Zhong-Yu Li, Ming-Hsuan Yang, Ming-Ming Cheng, Junwei Han, and Philip Torr. Large-scale unsupervised semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45 (6):7457–7476, 2022. 6

Yuyang Gao, Siyi Gu, Junji Jiang, Sungsoo Ray Hong, Dazhou Yu, and Liang Zhao. Going beyond xai: A systematic survey for explanation-guided learning. *ACM Computing Surveys*, 56(7):1–39, 2024. 3

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *ICLR*, 2018. 1

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11): 665–673, 2020. 1

Tao Han, Wei-Wei Tu, and Yu-Feng Li. Explanation consistency training: Facilitating consistency-based semi-supervised learning with interpretability. In *AAAI*, volume 35, pp. 7639–7646, 2021. 3

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pp. 770–778, 2016. 7

Jacob Kauffmann, Jonas Dippel, Lukas Ruff, Wojciech Samek, Klaus-Robert Müller, and Grégoire Montavon. Explainable ai reveals clever hans effects in unsupervised learning models. *Nature Machine Intelligence*, pp. 1–11, 2025. 1

Anton Kuznietsov, Balint Gyevnar, Cheng Wang, Steven Peters, and Stefano V Albrecht. Explainable ai for safe and trustworthy autonomous driving: A systematic review. *IEEE Transactions on Intelligent Transportation Systems*, 25(12):19342–19364, 2024. 1

Yi Li, Hualiang Wang, Xinpeng Ding, Haonan Wang, and Xiaomeng Li. Token activation map to visually explain multimodal llms. In *ICCV*, pp. 48–58, 2025a. 3

Yifan Li, Yuhang Chen, Anh Dao, Lichi Li, Zhongyi Cai, Zhen Tan, Tianlong Chen, and Yu Kong. Industryeqa: Pushing the frontiers of embodied question answering in industrial scenarios. In *NeurIPS*, 2025b. 1

Yifan Li, Zhixin Lai, Wentao Bao, Zhen Tan, Anh Dao, Kewei Sui, Jiayi Shen, Dong Liu, Huan Liu, and Yu Kong. Visual large language models for generalized and specialized applications. *arXiv preprint arXiv:2501.02765*, 2025c. 1

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *NeurIPS*, pp. 4765–4774, 2017. 3

Richard Ngo, Lawrence Chan, and Sören Mindermann. The alignment problem from a deep learning perspective. In *ICLR*, 2024. 1

Paul Novello, Thomas Fel, and David Vigouroux. Making sense of dependence: Efficient black-box explanations using dependence measure. In *NeurIPS*, pp. 4344–4357, 2022. 3

Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. In *BMVC*, pp. 151, 2018. 3

Vipin Pillai, Soroush Abbasi Koohpayegani, Ashley Ouligian, Dennis Fong, and Hamed Pirsiavash. Consistent explanations by contrastive learning. In *CVPR*, pp. 10213–10222, 2022. 3

Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pp. 1–52, 2021. 1

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pp. 8748–8763, 2021. 7

Sukrut Rao, Moritz Böhle, Amin Parchami-Araghi, and Bernt Schiele. Studying how to efficiently and effectively guide models with explanations. In *ICCV*, pp. 1922–1933, 2023. 2

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?" explaining the predictions of any classifier. In *SIGKDD*, pp. 1135–1144, 2016. 3

Elan Rosenfeld, Pradeep Kumar Ravikumar, and Andrej Risteski. The risks of invariant risk minimization. In *ICLR*, 2021. 1

Andrew Slavin Ross, Michael C Hughes, and Finale Doshi-Velez. Right for the right reasons: training differentiable models by constraining their explanations. In *IJCAI*, pp. 2662–2670, 2017. 2, 3, 6, 7

Patrick Schramowski, Wolfgang Stammer, Stefano Teso, Anna Brugger, Franziska Herbert, Xiaoting Shao, Hans-Georg Luigs, Anne-Katrin Mahlein, and Kristian Kersting. Making deep neural networks right for the right scientific reasons by interacting with their explanations. *Nature Machine Intelligence*, 2(8):476–486, 2020. 2, 3, 7

Ramprasaath R Selvaraju, Stefan Lee, Yilin Shen, Hongxia Jin, Shalini Ghosh, Larry Heck, Dhruv Batra, and Devi Parikh. Taking a hint: Leveraging explanations to make vision and language models more grounded. In *ICCV*, pp. 2591–2600, 2019. 3, 6

Ramprasaath R Selvaraju, Michael Cogswell, Das Abhishek, Vedantam Ramakrishna, Parikh Devi, and Batra Dhruv. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, 2020. 3, 6

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *ICLR 2014 Workshop*, 2014. 6

Ao Sun, Pingchuan Ma, Yuanyuan Yuan, and Shuai Wang. Explain any concept: Segment anything meets concept-based explanation. In *NeurIPS*, pp. 21826–21840, 2023. 3

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *NeurIPS*, pp. 74952–74965, 2023. 1

Xiaoying Xing, Chia-Wen Kuo, Li Fuxin, Yulei Niu, Fan Chen, Ming Li, Ying Wu, Longyin Wen, and Sijie Zhu. Where do large vision-language models look at when answering questions? *arXiv preprint arXiv:2503.13891*, 2025. 3

Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. *International Journal of Computer Vision*, 126(10):1084–1102, 2018. 7

Xiaofeng Zhang, Yihao Quan, Chen Shen, Xiaosong Yuan, Shaotian Yan, Liang Xie, Wenxiao Wang, Chaochen Gu, Hao Tang, and Jieping Ye. From redundancy to relevance: Information flow in lvlms across reasoning tasks. In *NAACL*, pp. 2289–2299, 2025a. 3

Yifei Zhang, Siyi Gu, Yuyang Gao, Bo Pan, Xiaofeng Yang, and Liang Zhao. Magi: Multi-annotated explanation-guided learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1977–1987, 2023. 3

Yifei Zhang, Tianxu Jiang, Bo Pan, Jingyu Wang, Guangji Bai, and Liang Zhao. Megl: Multimodal explanation-guided learning. *arXiv preprint arXiv:2411.13053*, 2024. 2, 6, 7

Yifei Zhang, James Song, Siyi Gu, Tianxu Jiang, Bo Pan, Guangji Bai, and Liang Zhao. Saliency-bench: A comprehensive benchmark for evaluating visual explanations. In *SIGKDD*, pp. 5924–5935, 2025b. 6

Zhong Zhang, Yaxi Lu, Yikun Fu, Yupeng Huo, Shenzhi Yang, Yesai Wu, Han Si, Xin Cong, Haotian Chen, Yankai Lin, et al. Agentcpm-gui: Building mobile-use agents with reinforcement fine-tuning. *arXiv preprint arXiv:2506.01391*, 2025c. 3, 9

Chenyang Zhao, Kun Wang, Xingyu Zeng, Rui Zhao, and Antoni B Chan. Gradient-based visual explanation for transformer-based clip. In *ICML*, pp. 61072–61091, 2024. 3, 6

A More Details for GUI-Agent Experiments

A.1 Dataset Format

Each sample is stored as a JSON object containing an `id`, a screenshot reference (`image`), and a multi-turn `conversations` list in a `system/user/assistant` format. The assistant output follows a strict action schema and includes a `thought` field (thinking) and an executable action such as a click `POINT` (decision). In addition, each sample provides a human-annotated `bounding_box` for the target UI element, which serves as a human prior for alignment and evaluation. All GUI tasks are collected and executed in Chinese, and the paper presents an English-translated version of the prompts for clarity.


```
{
  "id": "0",
  "image": {
    "<image_00>": "img/screenshot_0.jpg"
  },
  "conversations": [
    {
      "role": "system",
      "content": "# Role\nYou are an agent familiar with Android touch-based GUI operations. Given
      a user's request, analyze the GUI elements and layout on the current screen and produce the
      next action.\n\n# Task\nGiven the current screenshot, output the next operation to accomplish
      the user request.\n\n# Rule\n- Output in compact JSON format.\n- The action must follow the
      Schema constraints.\n\n# Schema\n{\"type\":\"object\",\"description\":\"Execute an action and
      decide the task
      status\",\"additionalProperties\":false,\"optional\":[\"thought\"],\"properties\":
      {\"thought\":{\"type\":\"string\",\"description\":\"The agent's reasoning\"},
      \"POINT\":{\"$ref\":\"#/$defs/Location\",\"description\":\"Click a specific position on the
      screen\"},\"to\":{\"description\":\"Movement / gesture parameters\",
      \"oneOf\":[{\"enum\":[\"up\",\"down\",\"left\",\"right\"],\"description\":\"Swipe from the
      current point (POINT) in one of four
      directions\"},{\"$ref\":\"#/$defs/Location\",\"description\":\"Move to a specific
      location\"}]},\"duration\":{\"type\":\"integer\",\"description\":\"Execution or wait time in
      milliseconds\",\"minimum\":0,\"default\":200},\"PRESS\":{\"type\":\"string\",\"description\":\"Trigger
      a special
      key\",\"enum\":[\"HOME\",\"BACK\",\"ENTER\"]},\"TYPE\":{\"type\":\"string\",\"description\":\"Input
      text\"},\"STATUS\":{\"type\":\"string\",\"description\":\"Task status: satisfied (no action
      needed), impossible, interrupt, need_feedback\",
      \"enum\":[\"continue\",\"finish\",\"satisfied\",\"impossible\",\"interrupt\",\"need_feedback\"],
      \"default\":\"continue\"}},\"$defs\":{\"Location\":{\"type\":\"array\",\"description\":\"Coordinates
      are relative to the top-left corner and scaled to [0,1000]; the first entry is x and the
      second is
      y\",\"items\":{\"type\":\"integer\",\"minimum\":0,\"maximum\":1000},\"minItems\":2,\"maxItems\":2}}}"
    },
    {
      "role": "user",
```

```
      "content": "<Question>Search bilibili.com, then search for 'Qianting Weiwei Mi' on the
      website and open the uploader list. Add the page to bookmarks and verify it appears in the
      bookmark manager.</Question>\nCurrent screenshot: <image_00>"
    },
    {
      "role": "assistant",
      "content": "{\"thought\":\"Locate and tap the browser icon on the home screen to open the
      browser.\",\"POINT\":[591,915]}"
    }
  ],
  "bounding_box": [706,2438,858,2590]
}
```


A.2 Attributing Thinking and Decision

Unlike approaches that attribute only the final action, this work treats a mobile agent's *reasoning process* (thinking) and *executed action* (decision) as a unified decision chain for attribution analysis. Concretely, given model parameters θ and input X, the proposed high-faithfulness attribution method EAGLE is used to jointly attribute the visual evidence relied upon when generating $\hat{Z}$ (the reasoning sequence) and $\hat{A}$ (the final action decision):

$$\mathcal{A} = \mathcal{A}(f_\theta, X, \{\hat{Z}, \hat{A}\}), \tag{6}$$

where $\mathcal{A}(\cdot)$ denotes the attribution operator and outputs a single attribution heatmap that quantifies the contribution of different screen regions to the *overall reasoning–decision process*. This attribution does not separate intermediate stages of reasoning and decision-making; instead, it directly captures the key set of visual evidence the model relies on to complete the current task. Owing to the explicit and stable spatial layout in GUI environments, the resulting attribution can be naturally mapped to screen coordinates, providing a unified and actionable supervision signal for subsequent attribution alignment and reliability-enhancing training.

A.3 Training Procedure

To control computational overhead, the attribution alignment loss is computed periodically rather than at every update step. Specifically, at pre-defined steps, the current model attributions are computed, the Top-k salient regions are extracted, and their overlap with the target action region is examined. If the salient attributed regions fail to sufficiently cover the target region, an attribution alignment penalty is applied. Algorithm 2 provides a formal description.

A.4 Evaluation Metrics

Two metrics are used to evaluate the GUI agent clicking task: click success rate and distance error. The **click success rate** measures whether the predicted click point falls inside the human-annotated target bounding box. The **distance error** quantifies how far the predicted point is from the ground-truth target region: it is set to 0 if the predicted point lies inside the bounding box; otherwise, it is computed as the minimum Euclidean distance from the point to the bounding box boundary (i.e., the closest point on the box). Figure A1 illustrates these metrics with representative examples.

B Limitations and Future Work

Limitations. Our method depends heavily on the quality and coverage of human-prior annotations. In large-scale settings, such fine-grained human labels are often unavailable or prohibitively expensive, and coarse or noisy priors may weaken the training signal and limit scalability.

Future work could mitigate this by developing scalable, automated prior acquisition schemes (e.g., weak/self-supervised cues, pseudo-labels from detection/segmentation models, or priors distilled from multi-model consensus). Another promising direction is to integrate attribution into reinforcement learning, using

Algorithm 2: Attribution-guided training for reliability enhancement of GUI agents

Input: Training sample (X, I, Z, A, B), where I is the screenshot, Z is the thinking process, A is the action decision, and B is the target UI bounding box; attribution operator $\mathcal{A}(\cdot)$; attribution interval K; number of top salient regions k, where $k = 2$.

Output: Trained model parameters θ.

Initialize model parameters θ;
for $t = 1$ **to** $T_{\max}$ **do**
  Predict thinking $\hat{Z}$ and action $\hat{A}$ under current parameters θ;
  Compute supervised loss $\mathcal{L}_{\text{CE}}$; `/* SFT cross-entropy loss */`
  if t mod $K == 0$ **then**
    $\mathbf{M} \leftarrow \mathcal{A}(f_\theta, I, \{\hat{Z}, \hat{A}\})$; `/* Compute joint attribution map */`
    Extract top-k salient regions $\mathcal{S} = \{S_1, \ldots, S_k\}$ from $\mathbf{M}$;
    $\mathcal{L}_{\text{attr}} \leftarrow 0$;
    for $i = 1$ **to** $|\mathcal{S}|$ **do**
      if $S_i \not\subset B$ **then**
        $\mathcal{L}_{\text{attr}} \leftarrow \mathcal{L}_{\text{attr}} + \text{Penalty}$;
        break;
      if $\text{Area}(S_i \cap B)/\text{Area}(B) \geq 0.75$ **then**
        break;
    end
    $\mathcal{L}_{\text{total}} \leftarrow \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{attr}}$;
  else
    $\mathcal{L}_{\text{total}} \leftarrow \mathcal{L}_{\text{CE}}$;
  end
  Update model parameters θ using $\nabla_\theta \mathcal{L}_{\text{total}}$;
end
return θ;

attribution-based signals to select or refine more evidence-consistent chains of thought (e.g., incorporating "evidence rationality" into the reward), thereby encouraging more reliable reasoning and decision-making.

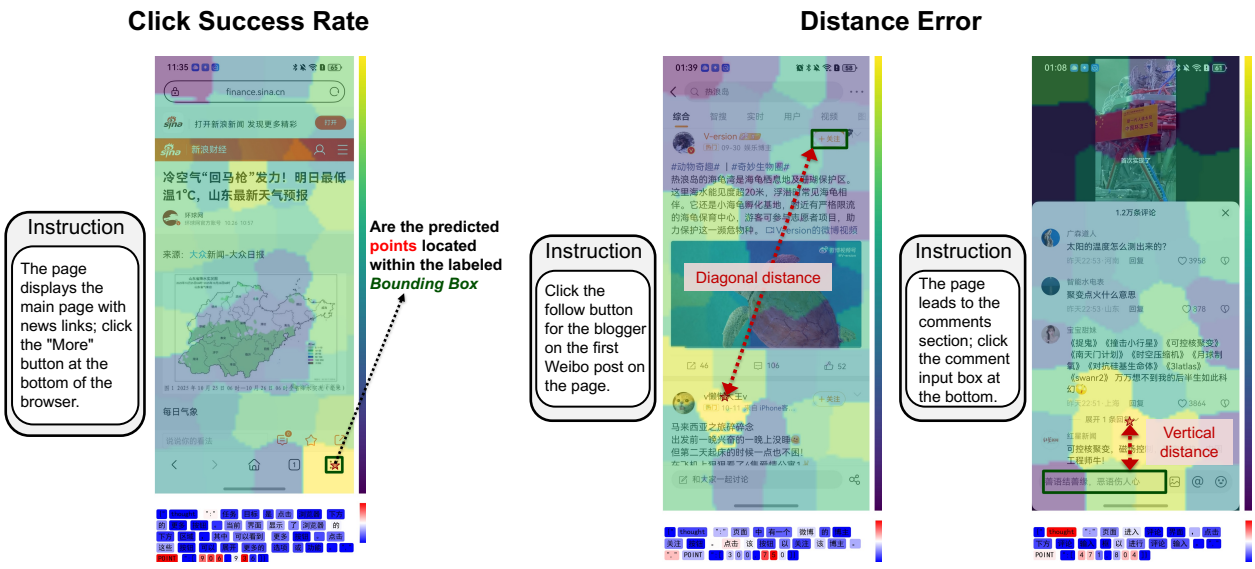

Figure A1: Illustration of evaluation metrics for the GUI agent clicking task. *Click Success Rate* counts a prediction as successful if the predicted click point falls inside the labeled target UI element bounding box. *Distance Error* measures the distance between the predicted click point and the target location (e.g., to the bounding box or target point), with examples showing diagonal and vertical distances.