

Right for the Wrong Reasons? Evaluating Plausibility and Behavioral Change in Explanation-Guided Learning

David Dembinsky^{1,2}, Adriano Lucieri¹, and Andreas Dengel^{1,2}

¹ German Research Center for Artificial Intelligence (DFKI) GmbH, Germany,
`{firstname}.{lastname}@dfki.de`

² RPTU University Kaiserslautern-Landau, Germany

Abstract. Explanation-Guided Learning (EGL) promises models that are not only accurate but also rely on task-relevant evidence. However, whether improved agreement between model explanations and human-provided rationales also implies a changed decision strategy remains unclear. We study this tension under controlled synthetic shortcut introduction using a constructed shortcut dataset. Inspired by real-world applications, our benchmark covers shortcuts of different kinds, including localized, background-based, and global cues. This allows us to evaluate whether different EGL objectives produce behavioral changes and explanatory changes. Across these axes, we find that the effects of EGL are strongly objective-specific. Rationale-based data augmentation provides the clearest robustness gains when shortcut evidence can be separated from task-relevant foreground information, reaching average accuracy improvements of up to +4.8 percentage points (pp) together with increased foreground reliance. Attribution-constraining objectives primarily improve explanation plausibility, increasing Importance Mass by up to +14 pp on average, but provide weaker and less consistent behavioral correction, with average accuracy gains of at most +1.6 pp. Moreover, neither improved robustness nor improved plausibility reliably implies higher perturbation-based fidelity. Overall, our results show that EGL can be useful, but its success cannot be inferred from rationale agreement alone and should instead be evaluated along complementary behavioral and explanation-level axes. We publicly release all experimental results to enable further analysis.³

Keywords: Explanation-Guided Learning · shortcut bias · plausibility · fidelity · robustness.

1 Introduction

Ever more complex machine learning models are deployed in high-stakes domains such as medicine and finance, where opaque decision processes complicate human oversight and may obscure biases encoded by the model. In response, the

³ <https://right-for-wrong-reasons.dfki.de/>

community has developed a broad range of methods for explaining model predictions. This growing emphasis on transparency is also reflected in regulation such as the EU AI Act, which establishes transparency requirements for high-risk AI systems [7].

While explanations are commonly used as post-hoc diagnostic tools, transparency can also be pursued during model design or training, for example through inherently interpretable architectures. A complementary strategy is to retain flexible models but guide them during training toward user-specified evidence. This idea underlies *Explanation-Guided Learning* (EGL), which augments the classification loss with auxiliary objectives derived from user-provided rationales [8,19]. In its classical form, such guidance directly constrains model explanations to align with these rationales, aiming to encourage models to base their predictions on task-relevant evidence and follow the aspiration of being “right for the right reasons.” However, whether optimizing explanations actually changes the model’s decision strategy in this intended way remains an empirical question. The promise of explanation-guided training has led to a growing body of work in computer vision and natural language processing [8,21,25,31,34].

Assessing whether EGL fulfills its intended goals is further complicated by a gap between the apparent quality of an explanation and the model behavior it is meant to reflect. *Plausibility*, i.e. agreement between an explanation and human annotations, does not necessarily imply *fidelity*, i.e. causal relevance of the highlighted features to the model’s decision [4,9]⁴. A model may therefore be trained to produce plausible explanations while still relying on shortcuts or spurious correlations [10]. For EGL, this raises a central question: do objectives that improve rationale agreement also induce the intended behavioral changes, such as greater robustness, reduced shortcut reliance, and more faithful explanations, or do they primarily optimize the appearance of explanations? This question is especially relevant when comparing classical attribution-constraining objectives with rationale-based data-augmentation objectives, since both use rationales during training but may affect the learned model differently. Yet prior work has mostly considered isolated shortcut settings and often emphasizes plausibility metrics, leaving the relation between objective type, model behavior, and explanation quality insufficiently understood.

Contributions. We conduct a multi-axis evaluation of EGL under controlled synthetic shortcut bias. To this end, we construct a benchmark that injects three shortcut types: localized patch cues, class-correlated backgrounds, and global color-cast cues. Using this benchmark, we compare several EGL objectives for a convolutional and a transformer-based model, distinguishing attribution-constraining objectives from rationale-based data-augmentation objectives. We evaluate the resulting models along four complementary axes, grouped into behavioral and explanation-level effects: robustness to shortcut shift (RQ1) and reliance on foreground versus shortcut evidence (RQ2) as behavioral axes, and explanation plausibility with respect to object masks (RQ3) and explanation

⁴ The terms *faithfulness* and *fidelity* are used interchangeably in this work.

fidelity to the model’s decision process (RQ4) as explanation-level axes. Our results show that these desiderata can diverge substantially. The Attribution-constraining objectives yield the clearest plausibility gains, whereas rationale-based data-augmentation objectives are more effective for improving robustness and reducing shortcut reliance in separable shortcut settings. These findings show that evaluating EGL primarily through explanation plausibility is insufficient: improved rationale agreement does not show that the model relies on the intended features, and objectives that directly intervene on the training data may be more effective for changing model behavior than attribution alignment.

2 Related Work

EGL provides a broad family of approaches that aim to shape model behavior by incorporating explanation-related feedback into training; for a comprehensive overview of objectives, feedback types, and application domains, we refer to *Gao et al.* [8]. In the classical attribution-based formulation, EGL augments standard task supervision with an additional explanation-supervision term:

$$\min \underbrace{\mathcal{L}_{\text{Pred}}(f(X), Y)}_{\text{task supervision}} + \underbrace{\lambda \mathcal{L}_{\text{Exp}}(\xi(f, X, Y), \tilde{E})}_{\text{explanation supervision}}, \quad (1)$$

where an explanation function ξ produces an explanation for the model f , which is compared against a supposed ground-truth rationale $\tilde{E}$ [8].⁵ In computer vision and vision-language reasoning, such ground-truth rationales $\tilde{E}$ are commonly represented as bounding boxes or segmentation masks [21,31,34]. While intuitive, these targets may be noisy, incomplete, or overly coarse, and even accurate spatial annotations need not coincide with the evidence actually used by a neural network. This raises the concern that optimizing explanations toward human-interpretable localization targets may improve the appearance of explanations without necessarily improving the model’s decision process [31]. A more domain-specific alternative is the use of causally or physically grounded continuous explanation targets, as in *Talies et al.*, who derive target explanations from the stress field around a crack tip [30].

The explanation function ξ is usually chosen to be differentiable, since the explanation loss must be optimized during training [8]. Prominent choices include input gradients or Saliency [24], as used by *Ross et al.* and *Simpson et al.* [19,25], and Grad-CAM [20], as used by *Selvaraju et al.* and *Talies et al.* [21,30]. These classical explanation-guidance objectives directly constrain the explanation signal, for example by aligning it with the annotated rationale or by penalizing attribution outside the rationale. VisFIS is particularly relevant for the present work because it combines such attribution-constraining terms with rationale-based data-augmentation terms [34]. Its attribution-based terms

⁵ In their survey, *Gao et al.* additionally include an *explanation regularization* term [8], which we omit for simplicity, as it is not central to the present work.

align explanations with ground-truth feature-importance annotations and penalize saliency outside the annotated rationale, whereas its sufficiency and uncertainty objectives use the rationale to construct augmented training inputs by preserving or removing rationale regions. Thus, VisFIS provides a natural objective family for separating explanation-alignment effects from rationale-based data-augmentation effects: both use “ground-truth” rationales during training, but only the former directly optimizes explanations, while the latter changes the available input evidence.

Empirically, EGL is often reported to improve predictive performance, but the effect is neither uniform nor independent of the chosen explanation target or objective type. In standard supervised settings, explanation guidance can lead to moderate performance improvements, for instance in medical imaging or crack-tip segmentation, but can also reduce performance in individual settings [25,30]. A central motivation is its potential to mitigate shortcut learning and improve generalization under distribution shift, which has been studied in controlled settings with known irrelevant features [19], vision-language benchmarks with shifted answer distributions [21,34], and image-classification settings with spatial confounders [31]. Overall, these works indicate that explanation-related supervision can reduce reliance on certain shortcuts, but the evidence remains limited to a relatively small set of shortcut types. Moreover, *Viviano et al.* show that attribution priors may improve generalization under covariate shift, but can also hurt performance when no shift is present, and that improved attribution localization does not necessarily imply improved out-of-distribution performance [31]. Thus, it remains unclear whether the observed benefits transfer to broader shortcut scenarios, and whether they are driven by explanation alignment itself or by objectives that intervene on the training data.

The effect of EGL on the resulting post-hoc explanations is evaluated less consistently than its effect on predictive performance. Most works primarily assess plausibility, i.e. whether explanations overlap with human- or domain-provided target regions [19,21,25,34]. Across these studies, EGL generally improves agreement with the annotated rationale, but this alone does not establish that the explanation faithfully represents the model’s decision process. Only a smaller subset of works evaluates explanation fidelity explicitly, for example through comparisons with occlusion-based importance [21], sufficiency and comprehensiveness [34], or deletion- and insertion-based metrics [30]. These studies suggest that plausibility, fidelity, and predictive performance are related but distinct: an increase in explanation plausibility can only be expected to improve predictive performance if the explanation is also faithful to the model’s behavior [34].

Taken together, prior work motivates EGL as a promising strategy for improving generalization and steering model behavior, but two limitations remain. First, shortcut mitigation has mostly been investigated in isolated benchmark settings rather than through a systematic comparison of different shortcut paradigms. Second, it remains insufficiently clear whether objectives that improve explanation alignment also change the evidence used by the model, or whether they can merely produce explanations that better resemble the pro-

vided rationale. These limitations motivate a controlled evaluation of classical attribution-constraining objectives and rationale-based data-augmentation objectives across shortcut-shift robustness, shortcut reliance, explanation plausibility, and explanation fidelity.

3 Experiments

We evaluate EGL under controlled shortcut bias, reporting representative results in the main text and the full experimental data on the supplementary website at <https://right-for-wrong-reasons.dfki.de/>.

Setup

Dataset. We construct customized-FMNIST (cFMNIST), a synthetic benchmark inspired by previous work [3], by placing Fashion-MNIST foreground objects [32] on CIFAR-10 backgrounds [14]. The task is to classify the foreground object while independently introducing shortcut cues.

To cover qualitatively different shortcut structures, we consider three real-world-inspired variants: *cFMNIST-patch* uses a colored square in the upper-left corner, reflecting marker-like or border artifacts reported in skin-lesion analysis [2,22]; *cFMNIST-background* correlates the CIFAR-10 background class with the foreground label, reflecting context biases in vision datasets [18,33]; and *cFMNIST-color* applies a class-correlated global color cast, abstracting acquisition- or site-specific shifts [13,28]. *cFMNIST-clean* uses the same approach for foreground-background construction, but without a class-correlated shortcut.

For each shortcut variant, we use correlation strengths $p \in \{0.5, 0.8, 0.95\}$; $p = 0.1$ corresponds to random assignment in the ten-class task and is used for shortcut-independent evaluation. The foreground object mask serves as reference rationale $\tilde{E}$. Construction details, split sizes, and diagnostic variants are given in Section A.

Models. We evaluate a ResNet18 [12] trained from scratch and an ImageNet-pretrained ViT-B/16 (ViT) [6], representing complementary convolutional and transformer-based architectures. We report ResNet18 results in the main text and provide the corresponding ViT analyses in the supplementary material; while effect sizes and individual settings differ, the main qualitative conclusions remain consistent across architectures.

EGL objectives. We follow the VisFIS objective family [34], which combines rationale-based data augmentation with classical explanation guidance and relates to earlier attribution-guided approaches [25,30,31]. We evaluate four objectives individually: *Sufficiency* and *Uncertainty* construct modified training inputs from the rationale, whereas *Invariance* and *Alignment* directly constrain the explanation signal. Let x_e denote the foreground-only input, x_u the input

with the foreground removed, $\xi_y(f, x)$ the nonnegative, spatially normalized attribution map for ground-truth class y , and $U_{\mathcal{Y}}$ the uniform distribution over the label set $\mathcal{Y}$. The objectives are

$$\begin{aligned}\mathcal{L}_{\text{Suff}} &= \text{CE}(f(x_e), y), & \mathcal{L}_{\text{Unc}} &= \text{KL}(f(x_u) \| U_{\mathcal{Y}}), \\ \mathcal{L}_{\text{Inv}} &= \left\| \xi_y(f, x) \odot (1 - \tilde{E}) \right\|_1, & \mathcal{L}_{\text{Align}} &= 1 - \text{cos-sim}(\xi_y(f, x), \tilde{E}).\end{aligned}\quad (2)$$

Thus, Sufficiency trains directly on the foreground, Uncertainty discourages confidence without foreground evidence, Invariance penalizes attribution outside $\tilde{E}$, and Alignment matches the attribution map to $\tilde{E}$. Each objective is added as $\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda \mathcal{L}_{\text{EGL}}$. We further evaluate VisFIS*, $\mathcal{L}_{\text{VisFIS}^*} = \mathcal{L}_{\text{CE}} + \sum_j \lambda_j^* \mathcal{L}_j$, with $j \in \{\text{Suff}, \text{Unc}, \text{Inv}, \text{Align}\}$, using the individually selected coefficients without an additional joint hyperparameter search. Attribution-constraining objectives use Vanilla Gradient as the training-time explanation signal, following Ying et al. [34], who report it to perform best among their considered signals with modest differences overall. We train a grid of λ values and report medians across random seeds; implementation, selection, and training details are provided in Section B.1 and Section B.2.

3.1 RQ1: Which EGL objectives improve generalization under shortcut shift?

RQ1 evaluates generalization when shortcut correlations learned during training are removed at test time. Test_{ID} preserves the training correlation, whereas $\text{Test}_{\text{rand}}$ uses random shortcut assignment with $p = 0.1$; for *Clean*, the two splits are interchangeable. Figure 1 reports ResNet18 accuracy changes relative to the unguided baseline. For each objective and setting, λ^* is selected by highest $\text{Test}_{\text{rand}}$ accuracy, and both $\text{Test}_{\text{rand}}$ and corresponding Test_{ID} changes are reported.

The benchmark induces the intended shortcut reliance. For the unguided baseline, increasing shortcut correlation raises Test_{ID} accuracy but lowers $\text{Test}_{\text{rand}}$ accuracy, most strongly for Patch and Color. Thus, the shortcuts are not merely present but actively used by the model, creating the generalization gap that EGL is intended to mitigate.

EGL is not a general-purpose robustness booster. Improvements concentrate in high-correlation settings, where shortcut reliance and the Test_{ID} – $\text{Test}_{\text{rand}}$ gap are largest. Clean and weakly correlated settings show little benefit, indicating that guidance mainly helps when there is substantial shortcut reliance to correct.

Sufficiency is strongest for separable shortcuts. Among individual objectives, Sufficiency yields the clearest $\text{Test}_{\text{rand}}$ gains for Patch and high-correlation background shortcuts. Foreground masking removes these spatially separable cues while preserving the task-relevant object. For Color, where the shortcut is distributed across the image, gains are weaker and less systematic; Invariance and

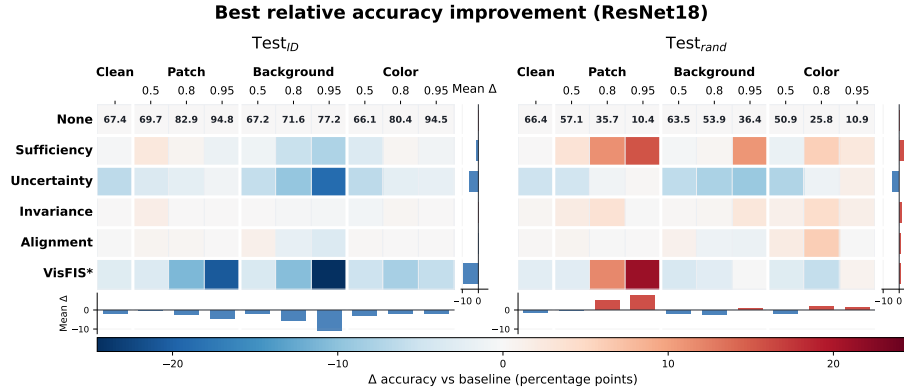


Fig. 1. RQ1. Relative ResNet18 accuracy change under EGL. For each objective and dataset setting, λ^* is selected by highest Test_{rand} accuracy; Test_{rand} and corresponding Test_{ID} changes are reported for the same λ^* . Columns group shortcut types and correlations p , rows show objectives, and baseline accuracies are shown above each column.

Alignment become comparatively more competitive but remain less consistent than the strongest Sufficiency effects. Uncertainty rarely improves robustness.

Robustness gains can trade off against Test_{ID} accuracy. Several Test_{rand} improvements coincide with neutral or negative Test_{ID} changes, especially at high correlations, since Test_{ID} continues to reward shortcut use. This mirrors prior findings that attribution priors can improve generalization under shift while reducing in-distribution performance [31]. The combined VisFIS* objective is also not automatically additive: unlike the strong OOD gains reported by Ying et al. [34], it improves some patch settings here but does not consistently preserve the strongest individual effects and can incur substantial Test_{ID} reductions, suggesting interference rather than additive benefits.

Answer to RQ1. EGL can improve shortcut-shift generalization, but its effectiveness depends on the objective and shortcut structure. Sufficiency provides the clearest gains when shortcut evidence can be suppressed without removing task-relevant foreground information, whereas attribution-constraining objectives are weaker and less consistent. The supplementary ViT results support the same conclusion, with less room for improvement from a stronger baseline.

3.2 RQ2: Which EGL objectives reduce shortcut reliance?

RQ2 asks whether robustness gains correspond to a shift in the evidence used by the model. The clearest effect comes from Sufficiency, which most consistently reduces shortcut reliance when shortcut and foreground evidence are separable, particularly for patch and high-correlation background shortcuts. To mea-

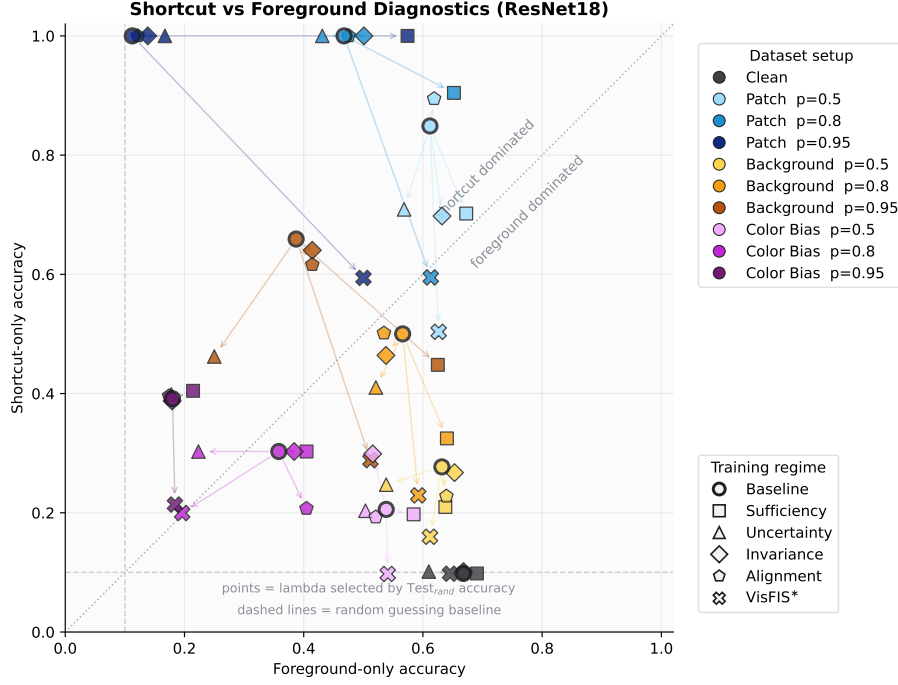


Fig. 2. RQ2. Foreground-only versus shortcut-only diagnostic accuracy for ResNet18. Arrows connect the unguided baseline to the configuration selected by highest $\text{Test}_{\text{rand}}$ accuracy in RQ1. Movement toward the lower-right indicates stronger foreground preference; dashed lines indicate random guessing.

sure shortcut reliance directly, we evaluate diagnostic inputs that isolate task-relevant foreground or shortcut evidence; details are given in Section A.1. The *Foreground-Preference* $\Phi_\lambda = A_{\text{FG},\lambda} - A_{\text{SC},\lambda}$ is the difference between foreground-only and shortcut-only accuracy, with positive values indicating foreground-dominated performance, whereas negative values indicate shortcut-dominated performance. For consistency with RQ1, we evaluate the λ^* selected by highest $\text{Test}_{\text{rand}}$ accuracy; reliance-oriented selections are reported on the supplementary website.

Figure 2 shows the ResNet18 diagnostics. Each arrow connects the unguided baseline to its robustness-selected EGL configuration; movement toward the lower-right indicates stronger foreground preference through higher foreground-only accuracy, lower shortcut-only accuracy, or both.

Baseline reliance depends on shortcut type and strength. The unguided ResNet18 is strongly foreground-dominated for Clean but shortcut-dominated for Patch, increasingly so at higher correlations. Background and Color, however, depend more strongly on p : low correlations remain mostly foreground-oriented, whereas

high correlations increasingly favor shortcut evidence. These patterns mirror the shortcut-shift failures in RQ1.

Sufficiency most consistently shifts reliance toward the foreground. The clearest changes occur for patch and high-correlation background shortcuts, where shortcut and foreground evidence are spatially separable. Sufficiency increases foreground-only performance, reduces the relative dominance of shortcut-only evidence, or both. These settings also yield some of its strongest $\text{Test}_{\text{rand}}$ gains, linking improved robustness to a behavioral shift away from shortcut evidence. For globally distributed color shortcuts, the shifts are substantially weaker.

Other objectives produce less systematic shifts. Uncertainty, Invariance, and Alignment alter reliance in individual settings but less consistently than Sufficiency. Uncertainty can reduce shortcut-only accuracy at the cost of foreground-only accuracy, while Invariance and Alignment generally produce smaller shifts. VisFIS* can strongly suppress shortcut reliance for some patch and background settings, but is less reliable for Color and does not consistently yield the strongest robustness gains.

Answer to RQ2. Sufficiency most consistently reduces shortcut reliance, particularly when shortcut and task-relevant evidence are separable. The effect is clearest for patch and high-correlation background shortcuts and coincides with the strongest robustness gains in RQ1. Other objectives show weaker or less systematic effects, and globally distributed color shortcuts remain difficult to correct. The supplementary ViT results support the same qualitative conclusion.

3.3 RQ3: Which EGL objectives improve explanation plausibility?

RQ1 and RQ2 showed that robustness and reliance shifts are strongest for objectives that can suppress shortcut evidence during training. RQ3 asks whether the robustness-selected models also produce more plausible post-hoc explanations. The clearest gains instead come from Invariance and Alignment, showing that the objectives that most improve model behavior are not necessarily those that most improve rationale agreement. As before, we evaluate the λ^* selected by highest $\text{Test}_{\text{rand}}$ accuracy; plausibility-oriented selections are provided on the supplementary website.

Plausibility measures the agreement between an explanation and a human-expected reference rationale [4], here the foreground mask. We evaluate Saliency [24], SmoothGrad [27], Integrated Gradients [29], Occlusion [35], LIME [18], and SHAP [15]. Plausibility is measured using Importance Mass (IM), AUROC, and IoU AUC, which respectively capture attribution mass inside the foreground, foreground-background ranking, and spatial overlap across thresholds. We focus on IM, which directly measures foreground concentration, and ResNet18 in Figure 3; ViT, plausibility-oriented selections, AUROC, and IoU AUC are provided on the supplementary website, with implementation details in Section C.1 and Section C.2.

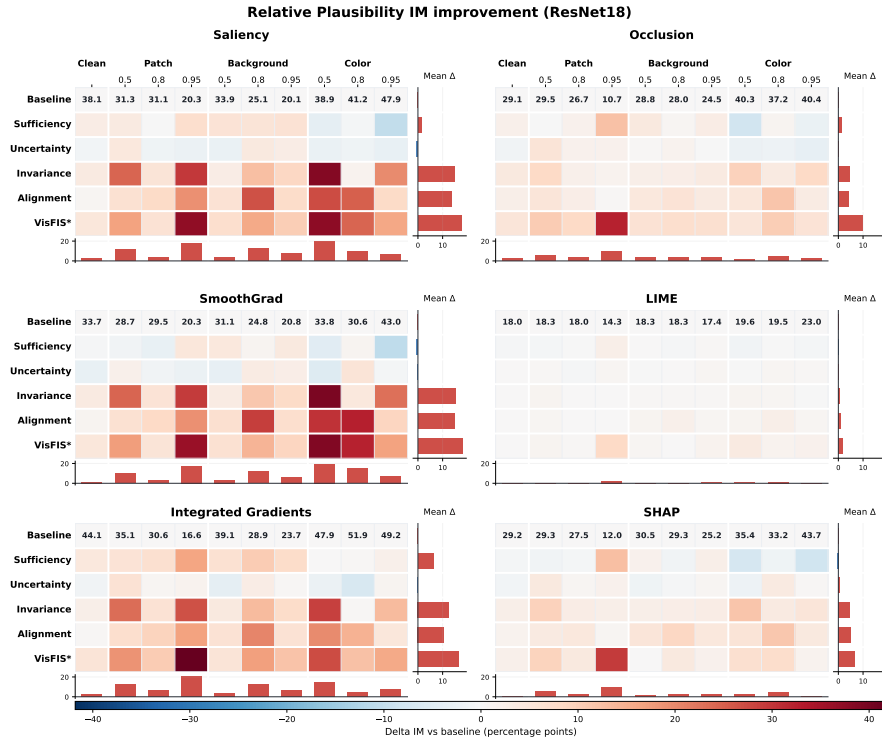


Fig. 3. RQ3. Relative ResNet18 plausibility change measured by Importance Mass. λ^* is selected by highest $\text{Test}_{\text{rand}}$ accuracy as in RQ1. Columns group shortcut types and correlations p , rows show objectives, and panels correspond to post-hoc explanation methods. Baseline Importance Mass values are shown above each column.

Plausibility gains transfer unevenly across explanation families. Saliency, SmoothGrad, and Integrated Gradients show similar qualitative patterns, whereas Occlusion, LIME, and SHAP improve more weakly and less consistently. This is consistent with Saliency being the training-time guidance signal: supervision transfers most reliably within the gradient-based explanation family and less to perturbation-based and surrogate-based methods.

Attribution-constraining objectives yield the clearest gains. Invariance and Alignment produce the strongest and most consistent IM improvements, matching their direct optimization of foreground-aligned attributions. Sufficiency, despite its stronger robustness and reliance effects, yields more modest plausibility gains, while Uncertainty shows no consistent improvement. Hence, behavioral correction and explanation plausibility favor different objectives.

Objective type matters more than shortcut type. Plausibility does not show the systematic shortcut-type or correlation pattern seen for robustness and reliance.

Because λ^* is selected by $\text{Test}_{\text{rand}}$ accuracy, these results describe robustness-selected rather than plausibility-optimal models; plausibility-oriented selection strengthens the improvements while preserving the advantage of Invariance and Alignment. The additional metrics support the same overall pattern: AUROC shows similar improvements in foreground-background ranking, whereas IoU AUC is more conservative, suggesting that EGL more reliably shifts attribution toward the foreground than produces uniformly better object-shaped maps.

Answer to RQ3. Invariance and Alignment most consistently improve explanation plausibility, whereas Sufficiency drives the strongest behavioral changes in separable shortcut settings. Thus, improved robustness does not imply more plausible explanations, and improved plausibility does not imply reduced shortcut reliance, consistent with evidence that localization and OOD performance can diverge [31] and that plausibility should be interpreted through faithfulness [34]. The supplementary ViT results, additional metrics, and plausibility-oriented selections support the same conclusion.

3.4 RQ4: Do EGL objectives improve explanation fidelity?

RQ4 asks whether the behavioral or plausibility improvements observed above are accompanied by more faithful explanations. No objective yields a consistent fidelity improvement across architectures, methods, and shortcut settings. This is particularly relevant because VisFIS argues that plausibility and predictive performance are meaningfully connected only when explanations are faithful to model behavior [34].

We evaluate the same six explanation methods using Decaying Degradation Score (DDS) [23], a guided perturbation fidelity metric [4]. DDS tests whether perturbing highly attributed features affects the model output earlier and more strongly than perturbing low-attribution features, with respect to the predicted class. As in RQ2 and RQ3, we evaluate the robustness-selected λ^* from RQ1. Figure 4 reports ResNet18 DDS with mean replacement; formal details, ViT results, Gaussian perturbations, and fidelity-oriented selections are provided in Section C.3 and on the supplementary website.

Fidelity effects vary across architectures, methods, and shortcuts. Mean replacement and Gaussian perturbation both yield similar qualitative conclusions. For ResNet18, baseline DDS is often already high, particularly for perturbation-based and surrogate explanations, leaving little room for improvement; lower ViT baselines produce larger apparent gains in several settings. Neither shortcut type nor correlation strength determines the direction of the effect, and Patch $p = 0.95$ illustrates the mismatch particularly clearly: qualitative inspection suggests that most gradient explanations move toward the foreground, while the corresponding DDS changes remain mixed. Unlike robustness, reliance, or plausibility, fidelity therefore shows no clear global response to EGL.



Fig. 4. RQ4. Relative ResNet18 fidelity change measured by DDS with mean replacement. λ^* is selected by highest $\text{Test}_{\text{rand}}$ accuracy as in RQ1. Columns group shortcut types and correlations p , rows show objectives, and panels correspond to post-hoc explanation methods. Baseline DDS values are shown above each column.

Attribution-constraining objectives improve fidelity selectively. For ResNet18, improvements occur mainly for gradient-based explanations under Invariance and Alignment, partially mirroring their plausibility gains in RQ3. The effect is nevertheless weaker and less systematic, with no stable objective ordering across datasets, architectures, or explanation methods. ViT shows more frequent improvements, including for initially weaker perturbation-based and surrogate explanations, but likewise no consistent ranking.

Behavioral improvements do not imply higher fidelity. Sufficiency provides the clearest counterexample. Despite some of the strongest robustness and foreground-reliance gains in RQ1 and RQ2, it often reduces ResNet18 DDS in settings where model behavior changes substantially. On Background $p = 0.95$, for example, Sufficiency improves $\text{Test}_{\text{rand}}$ accuracy and foreground preference while DDS decreases for all six explanation methods. Robustness and foreground reliance therefore do not imply higher perturbation-based fidelity.

Answer to RQ4. Fidelity does not improve consistently under any objective and cannot be inferred from robustness, reliance, or plausibility. Invariance and Alignment yield selective gains, especially for gradient-based explanations, whereas Sufficiency can improve robustness and foreground reliance while reducing DDS. Fidelity therefore constitutes a separate evaluation axis: higher DDS indicates that an explanation better captures the model’s current decision process, but not that the model relies on semantically intended evidence. The supplementary ViT results, Gaussian perturbations, and fidelity-oriented selections support the same qualitative conclusion.

4 Discussion

Together, our results show that explanation-guided training can satisfy the visible rationale objective without reliably inducing the intended behavioral change. Attribution-constraining objectives such as Invariance and Alignment most clearly improve explanation plausibility, especially for gradient-family explanations, but they provide weaker and less consistent gains in shortcut-shift robustness and foreground reliance. This challenges a central motivation of EGL, namely that optimizing explanations should steer the model toward the intended evidence. Viewed through Goodhart’s law, rationale agreement becomes a proxy for the desired behavioral change: once optimized directly, it can improve as a measured explanation property without ensuring that the model’s evidence use has changed accordingly. This is not surprising in light of broader evidence that explanation maps can be manipulated without proportionally changing model behavior, for example through adversarial attacks or models trained to expose misleading explanation interfaces [5,4,26,36].

The strongest behavioral correction comes from Sufficiency, which uses the rationale as data augmentation rather than as an explanation-alignment target. By training on foreground-only inputs, Sufficiency changes the evidence available during optimization and can suppress shortcut reliance when shortcut and task evidence are separable. This explains its strong effects for localized patches and many background shortcuts, where foreground masking removes shortcut evidence while preserving the object. For global color shortcuts, the cue is distributed across the image and less cleanly separable from useful foreground information, leading to weaker and less stable improvements. Thus, our results qualify VisFIS-style supervision [34]: its objectives should not be interpreted as a uniformly additive package, since attribution constraints primarily improve plausibility whereas rationale-based data augmentation more directly changes model behavior. At the same time, this should not be read as a general argument against EGL. When rationales are causally grounded and well aligned with the data-generating process, even attribution-constraining objectives can be useful [30]. Our results instead caution against treating any plausible or human-annotated rationale as a sufficient ground truth for the model’s intended decision process.

Fidelity further complicates this picture. An illustrative case is ResNet18 on Patch $p = 0.95$ under the robustness-selected VisFIS* objective. Here, several evaluation axes suggest a successful shift from shortcut-based to foreground-based evidence: VisFIS* improves $\text{Test}_{\text{rand}}$ accuracy (RQ1), increases foreground-only accuracy and reduces shortcut-only accuracy (RQ2), and produces substantially more foreground-aligned explanations (RQ3). Yet DDS decreases across explanation methods (RQ4). This does not necessarily mean that explanation and behavior are completely misaligned. Rather, EGL appears to move the explanation more strongly toward the foreground than it moves the model behavior itself. The resulting explanation can become more plausible while becoming less aligned with the remaining sensitivities of the model, which is precisely what perturbation-based fidelity detects. Conversely, baseline explanations can be highly faithful under perturbation because they identify the shortcut patch used by the shortcut-reliant model. Perturbation fidelity should therefore be interpreted as a test of model sensitivity under a specified perturbation process, not as a generic certificate of semantic correctness. This underlines a broader challenge in feature attribution evaluation, where different metrics can encode different assumptions about what constitutes a good explanation [1,4].

A practical implication is that EGL claims depend on the selected evaluation axis. Our main analyses select λ^* by shortcut-shift robustness and then test whether other axes improve as well. The supplementary model-selection analyses show the complementary view: selecting for reliance, plausibility, or fidelity can change which objective appears most successful, and improvements on one axis do not reliably recover gains on the others.⁶ Accordingly, EGL evaluations should state whether they target behavioral robustness, evidence reliance, explanation plausibility, fidelity, or a specific combination of these goals. The underlying concern is not unique to feature attribution in computer vision. Related work on NLP rationales and concept-based models has reported analogous cases in which constraining the visible explanation does not necessarily constrain the predictor to use only the intended evidence [11,16]. This makes large language models a particularly relevant setting to investigate: natural-language rationales can provide convincing explanation interfaces, while the extent to which they reflect or constrain the model’s actual decision process remains an open question.

Limitations and future work. Our benchmark intentionally trades ecological realism for experimental control, allowing us to isolate how shortcut type and strength affect the behavior of different EGL objectives. Whether these findings transfer to natural shortcuts, real-world confounding, or rationale supervision in language and multimodal models remains to be established. We focus on Saliency as the training-time guidance signal and on a fixed set of post-hoc explanation methods; other guidance signals, explanation methods, or attribution objectives may change the balance between robustness, reliance, plausibility, and fidelity. Similarly, our combined VisFIS* objective uses the individually selected coeffi-

⁶ We encourage interested readers to inspect the supplementary material for a more holistic inspection: <https://right-for-wrong-reasons.dfki.de/>.

cients for the four objective terms rather than a dedicated joint hyperparameter search, which may underestimate the attainable performance of the combined objective. Future work could also make the proxy failure more explicit by training models toward intentionally uninformative or implausible rationales and testing how strongly explanation appearance can be decoupled from model behavior. Finally, our results call for further work on whether perturbation-based fidelity metrics capture the right notion of faithfulness for explanation-guided models, and whether the community can develop more robust ways of assessing feature attribution fidelity. This question is especially important as explanation-guided objectives and rationale-based supervision are extended from controlled vision benchmarks to broader settings, including language models and multimodal systems.

5 Conclusion

We presented a multi-axis evaluation of EGL under controlled shortcut bias, jointly analysing shortcut-shift robustness, diagnostic reliance, explanation plausibility, and explanation fidelity. Our results show that direct explanation supervision can improve visible rationale agreement without reliably inducing the intended behavioral change. Attribution-constraining objectives such as Invariance and Alignment primarily improve explanation plausibility, especially for gradient-family explanations, but provide weaker and less consistent robustness and reliance gains. In contrast, rationale-based data augmentation, especially Sufficiency, more directly changes model behavior when shortcut evidence can be separated from task-relevant foreground information.

Fidelity further complicates this interpretation rather than resolving it. Even when robustness, reliance, and plausibility jointly suggest a successful shift from shortcut-based to foreground-based evidence, perturbation-based fidelity can decrease. Thus, EGL should not be judged from rationale agreement alone, but neither should fidelity be treated as a final certificate of semantic correctness. Our results support a more cautious view: explanation-guided training can be useful, but its success depends on how the rationale enters training, whether it changes the evidence available to the model, and how behavioral and explanation-level evaluations are interpreted together.

Acknowledgments. The authors thank the anonymous reviewers for their constructive feedback. This work was funded by the Ministry of Science and Health of the State of Rhineland-Palatinate (Ministerium für Wissenschaft und Gesundheit Rheinland-Pfalz) within the project BioMaX (grant no. 724-0019#2025/0004-1501 15404).

Use of generative AI. The authors used ChatGPT and Codex by OpenAI to support manuscript preparation, software implementation, and development of the supplementary website. This included assistance with structuring and revising text, and implementing code for experiments, analysis, visualization, and the website. All research design decisions, result interpretations, scientific claims, and manuscript revisions were made and checked by the authors, who remain fully responsible for the work.

References

1. Barr, B., Fatsi, N., Hancox-Li, L., Richter, P., Proano, D., Mok, C.: The disagreement problem in faithfulness metrics. *arXiv:2311.07763* (2023)
2. Bissoto, A., Valle, E., Avila, S.: Debiasing skin lesion datasets and models? not so fast. In: *CVPR Workshops* (2020)
3. Bykov, K., Hedström, A., Nakajima, S., Höhne, M.: NoiseGrad — enhancing explanations by introducing stochasticity to model weights. *AAAI* (2022)
4. Dembinsky, D., Lucieri, A., Frolov, S., Najjar, H., Watanabe, K., Dengel, A.: Unifying VXAI: A systematic review and framework for the evaluation of explainable AI. *TMLR* (2026)
5. Dombrowski, A.K., Alber, M., Anders, C., Ackermann, M., Müller, K.R., Kessel, P.: Explanations can be manipulated and geometry is to blame. *Advances in Neural Information Processing Systems* (2019)
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv:2010.11929* (2020)
7. European Parliament and Council of the European Union: Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union* (2024), <https://eur-lex.europa.eu/eli/reg/2024/1689/oj>
8. Gao, Y., Gu, S., Jiang, J., Hong, S.R., Yu, D., Zhao, L.: Going beyond XAI: A systematic survey for explanation-guided learning. *ACM CSUR* (2024)
9. Hase, P., Bansal, M.: Evaluating explainable AI: Which algorithmic explanations help users predict model behavior? In: *ACL* (2020)
10. Hase, P., Zhang, S., Xie, H., Bansal, M.: Leakage-adjusted simulatability: Can models generate non-trivial explanations of their behavior in natural language? In: *Findings of ACL: EMNLP* (2020)
11. Havasi, M., Parbhoo, S., Doshi-Velez, F.: Addressing leakage in concept bottleneck models. In: *NeurIPS* (2022)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016)
13. Howard, F.M., Dolezal, J., Kochanny, S., Schulte, J., Chen, H., Heij, L., Huo, D., Nanda, R., Olopade, O.I., Kather, J.N., et al.: The impact of site-specific digital histology signatures on deep learning model accuracy and bias. *Nat. Commun.* (2021)
14. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
15. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems* (2017)
16. Lyu, Q., Apidianaki, M., Callison-Burch, C.: Towards faithful model explanation in NLP: A survey. *Comput. Linguist.* (2024)
17. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: PyTorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems* (2019)
18. Ribeiro, M.T., Singh, S., Guestrin, C.: "why should i trust you?" explaining the predictions of any classifier. In: *KDD* (2016)

19. Ross, A.S., Hughes, M.C., Doshi-Velez, F.: Right for the right reasons: training differentiable models by constraining their explanations. In: IJCAI (2017)
20. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: ICCV (2017)
21. Selvaraju, R.R., Lee, S., Shen, Y., Jin, H., Ghosh, S., Heck, L., Batra, D., Parikh, D.: Taking a hint: Leveraging explanations to make vision and language models more grounded. In: ICCV (2019)
22. Sies, K., Winkler, J.K., Fink, C., Bardehle, F., Toberer, F., Kommoss, F.K., Buhl, T., Enk, A., Rosenberger, A., Haenssle, H.A.: Dark corner artefact and diagnostic performance of a market-approved neural network for skin cancer classification. JDDG (2021)
23. Šimić, I., Sabol, V., Veas, E.: Perturbation effect: a metric to counter misleading validation of feature attribution. In: CIKM (2022)
24. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. In: ICLR Workshop (2014)
25. Simpson, B., Dutil, F., Bengio, Y., Cohen, J.P.: GradMask: Reduce overfitting by regularizing saliency (2019)
26. Slack, D., Hilgard, S., Singh, S., Lakkaraju, H.: Feature attributions and counterfactual explanations can be manipulated. arXiv:2106.12563 (2021)
27. Smilkov, D., Thorat, N., Kim, B., Viégas, F., Wattenberg, M.: SmoothGrad: removing noise by adding noise. arXiv preprint arXiv:1706.03825 (2017)
28. Soylu, U., Oelze, M.L.: Calibrating data mismatches in deep learning-based quantitative ultrasound using setting transfer functions. IEEE TUFFC (2023)
29. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: ICML (2017)
30. Talies, J., Breitbarth, E., Melching, D.: Trustworthy AI-based crack-tip segmentation using domain-guided explanations. Mach. Learn.: Sci. Technol. (2026)
31. Viviano, J.D., Simpson, B., Dutil, F., Bengio, Y., Cohen, J.P.: Saliency is a possible red herring when diagnosing poor generalization. In: International Conference on Learning Representations (2021)
32. Xiao, H., Rasul, K., Vollgraf, R.: Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. arXiv:1708.07747 (2017)
33. Xiao, K.Y., Engstrom, L., Ilyas, A., Madry, A.: Noise or signal: The role of image backgrounds in object recognition. In: International Conference on Learning Representations (2021)
34. Ying, Z., Hase, P., Bansal, M.: VisFIS: Visual feature importance supervision with right-for-the-right-reason objectives. Advances in Neural Information Processing Systems (2022)
35. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: ECCV (2014)
36. Zhang, X., Wang, N., Shen, H., Ji, S., Luo, X., Wang, T.: Interpretable deep learning under fire. In: USENIX Security (2020)

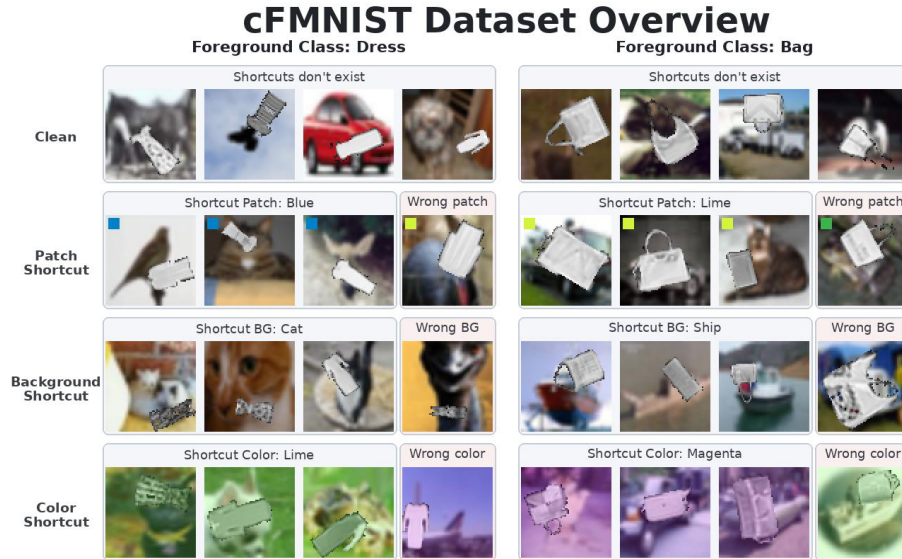


Fig. 5. Examples of clean cFMNIST and the patch, background, and color shortcut variants.

A Dataset and Diagnostic Inputs

Synthetic dataset construction. Each cFMNIST image places a Fashion-MNIST object on a CIFAR-10 background. The foreground mask $\tilde{E}$ is obtained by thresholding the Fashion-MNIST image at pixel intensity $t = 32$, while retaining the original pixel intensities within the mask. Object and mask are jointly resized, rotated, and placed at a random image location.

Splits and shortcut correlations. For each shortcut type and $p \in \{0.5, 0.8, 0.95\}$, the shortcut cue matches the foreground class with probability p ; $p = 0.1$ corresponds to random assignment in the ten-class task. Each dataset contains 20,000 training, 5,000 validation, and 5,000 images each for Test_{ID} and $\text{Test}_{\text{rand}}$. Test_{ID} uses the training correlation, whereas $\text{Test}_{\text{rand}}$ uses $p = 0.1$. Unless stated otherwise, RQ3 and RQ4 are evaluated on Test_{ID} . RQ1 evaluates Test_{ID} and $\text{Test}_{\text{rand}}$, while RQ2 uses the diagnostic inputs described in Section A.1.

A.1 Foreground vs. Shortcut Diagnostics

Foreground-only inputs preserve the transformed Fashion-MNIST object and replace all other pixels with the dataset mean background. Shortcut-only inputs preserve only the regime-specific cue: an uninformative mean image for Clean, the colored patch on the mean background for Patch, the CIFAR-10 background for Background, and the global tint applied to the mean background for Color. Shortcut-only inputs use perfectly class-correlated cues ($p = 1$).



Fig. 6. Foreground-only and shortcut-only diagnostic inputs used in Section 3.2.

B Training and EGL Objectives

B.1 Model training

Architectures and protocol. We train ResNet18 [12] from scratch and use the ImageNet-pretrained PyTorch `vit_b_16` [17]. For each objective, we evaluate $\lambda \in \{0.05, 0.1, 0.25, 0.5, 1\}$ with five random seeds per value; the shared $\lambda = 0$ baseline uses ten seeds.

Training hyperparameters. Architecture-specific hyperparameters are fixed across guidance regimes. Models are trained for up to 80 epochs with early stopping. ResNet18 uses Adam with learning rate 10^{-3} , weight decay 10^{-4} , batch size 64, cosine scheduling without warm-up, and patience 12. ViT uses AdamW with learning rate 10^{-5} , weight decay 0.05, batch size 4, label smoothing 0.05, a 5-epoch linear warm-up followed by cosine scheduling, and patience 20. EGL checkpoints are selected by the full weighted validation objective; the $\lambda = 0$ baseline is selected by validation accuracy.

B.2 Explanation guidance implementation

For Sufficiency, x_e preserves the transformed Fashion-MNIST object and replaces non-foreground pixels by the dataset mean background. For Uncertainty, x_u covers the foreground with a circle enclosing its bounding box to avoid object-shape leakage through the removal boundary. For Invariance and Alignment, Vanilla Gradient attributions are computed for the ground-truth class y , converted to nonnegative spatial mass by absolute channel-summing, and normalized before applying the loss.

The evaluated λ range covers the relevant scale of the guidance terms: $\mathcal{L}_{\text{Inv}}$ is typically comparable to the classification loss or up to two orders smaller, particularly for ViT. Empirically, small nonzero λ values already alter the optimized explanations, with the effect generally plateauing within the evaluated range.

C Post-hoc Explanations and Evaluation Metrics

C.1 Explanation methods

Explanations are computed on Test_{ID} for the predicted class $\hat{y}$; checks on $\text{Test}_{\text{rand}}$ showed no meaningful qualitative differences. Hyperparameters were selected by manual inspection of representative examples, not by optimizing an evaluation metric.

Saliency [24] uses the absolute input gradient of the predicted-class logit. **SmoothGrad** [27] averages gradients over 64 noisy inputs with Gaussian noise $\sigma = 0.15$ in normalized input space, followed by absolute channel-summing. **Integrated Gradients** [29] uses a zero baseline and 64 trapezoidally weighted integration steps; attributions are multiplied by the input-baseline difference and reduced by absolute channel-sum. **Occlusion** [35] replaces 12×12 patches with zero at stride 8 and assigns the resulting predicted-logit change, averaged across overlapping patches. **LIME** [18] uses SLIC with 32 segments, compactness 10, and $\sigma = 0$; 256 binary masks are sampled, masked regions are set to zero, and a locally weighted ridge surrogate is fitted to predicted-class logits. **SHAP** [15] uses the same SLIC segmentation and 512 coalition masks, including all-on and all-off coalitions, and fits a KernelSHAP-weighted linear surrogate to predicted-class logits. For LIME and SHAP, superpixel coefficients are projected back to pixels.

C.2 Plausibility metrics

Importance Mass measures the fraction of absolute attribution mass inside the foreground: $\text{IM} = (\sum_{i \in \tilde{E}} |\xi_i|) / (\sum_i |\xi_i|)$. **AUROC** evaluates whether foreground pixels receive higher attribution scores than background pixels. For **IoU AUC**, each explanation is min-max normalized to $[0, 1]$ and thresholded at 101 levels; we report the normalized trapezoidal area under the resulting IoU-threshold curve. Constant attribution maps are normalized to zero.

C.3 Fidelity metrics

We evaluate fidelity using Decaying Degradation Score (DDS) [23] for the predicted class $\hat{y}$. DDS compares Most Relevant First (MoRF) and Least Relevant First (LeRF) perturbations, which remove features in descending and ascending attribution order, respectively. Let M_i and L_i indicate whether the original predicted class is preserved after step i of the respective perturbation sequence. Then

$$\text{DDS} = \sum_{i=1}^n (L_i - M_i) \left(\frac{n - i + 1}{n} \right)^3,$$

where the cubic decay emphasizes early perturbations. Pixel-based explanations use steps of 0.05 of all pixels; LIME and SHAP use one superpixel per step. Mean replacement sets selected pixels to 0 in normalized input space, while the Gaussian variant adds zero-mean noise with $\sigma = 0.25$.