LVLM-Aided Alignment of Task-Specific Vision Models

Alexander Koebler[1,*] Lukas Kuhn[1,3,4] Ingo Thon[2] Florian Buettner[1,3,4]

[1]Goethe University Frankfurt [2]Siemens AG
[3]German Cancer Research Center (DKFZ) [4]German Cancer Consortium (DKTK)

[*]Corresponding author: alexander.koebler@gmx.de

Abstract

In high-stakes domains, small task-specific vision models are crucial due to their low computational requirements and the availability of numerous methods to explain their results. However, these explanations often reveal that the models do not align well with human domain knowledge, relying instead on spurious correlations. This might result in brittle behavior once deployed in the real-world. To address this issue, we introduce a novel and efficient method for aligning small task-specific vision models with human domain knowledge by leveraging the generalization capabilities of a Large Vision Language Model (LVLM). Our LVLM-Aided Visual Alignment (LVLM-VA) method provides a bidirectional interface that translates model behavior into natural language and maps human class-level specifications to image-level critiques, enabling effective interaction between domain experts and the model. Our method demonstrates substantial improvement in aligning model behavior with human specifications, as validated on both synthetic and real-world datasets. We show that it effectively reduces the model's dependence on spurious features and on group-specific biases, without requiring fine-grained feedback.

1. Introduction

In an era of increasingly large general-purpose models being able to interpret and translate visual inputs in natural language, reliable small task-specific vision models for narrow classification tasks are still of vital importance. This is especially true in many high-stakes domains where interpretability and trustworthiness demands are rigorous. Examples include the medical and manufacturing domains, where misclassifications can have severe downstream impact, requiring high robustness and explainability. For these non-functional requirements, current Large Vision Language Models (LVLMs) fall short [8, 33]. However, ensuring the continued reliability of small task-specific models and making their predictions interpretable to subject-matter

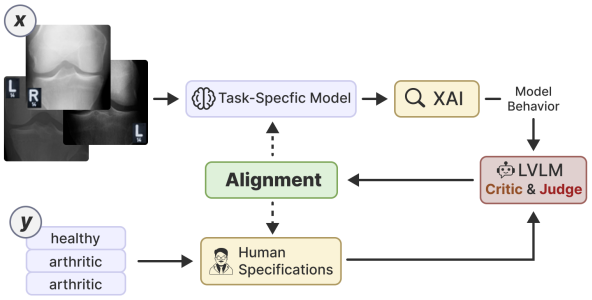

Figure 1. LVLM-Aided Visual Alignment (LVLM-VA) of a small task-specific vision model steered by human domain knowledge, using Explainable AI (XAI) in conjunction with a Large Vision Language Model (LVLM) Critic & Judge pair. The domain knowledge is induced into the system via human specifications on a class level supporting the LVLM to identify relevant core features within the input images and detect spurious shortcuts based on the model explanations. The Critic & Judge assessment is used to correct the original model in an alignment step but can also be used to provide feedback to the human expert.

experts remain challenging [5]. Spurious correlations in relatively small training datasets for narrow tasks can cause a model to learn shortcuts that yield good performance on the training distribution but result in brittle behavior when the model is deployed in the real world [17, 24]. One way to tackle this issue and increase the reliability of models is to explicitly incorporate human domain knowledge into the model training pipeline [30] and by this align the model with how a human would solve the task. While Explainable AI (XAI) techniques can be utilized to identify the learned shortcuts and make targeted corrections, interpreting the explanations generated by widely used XAI methods is often very difficult for domain experts. Furthermore, current methods rely on instance-wise feedback on the model's explanations [23, 25], which is too time-consuming for experts who are often highly specialized professionals, such as medical doctors. However, the bidirectional process of aligning an ML model, consisting of providing feedback

and interpreting the model's reasoning, is important not only for incorporating human knowledge and values but also for increasing user trust [26].

In this work, we introduce a synergistic approach that leverages recent advances in the capabilities of LVLMs to align small, task-specific vision models with human domain knowledge. The LVLM acts as a bidirectional translator. First, it translates explanations of the current model's behavior from image space into natural language, highlighting spurious correlations. Second, it translates human domain knowledge about the vision task, expressed in natural language, into instance-wise critiques in image space. Thus, the LVLM provides domain experts with a more intuitive interface through which they can actively steer the model and critically evaluate its reasoning. To achieve this, we make the following contributions:

- We propose LVLM-Aided Visual Alignment (LVLM-VA) as a novel approach allowing for automated instance-wise correction from class-level human specifications to efficiently align a neural network with human domain knowledge, reducing its reliance on spurious correlations.
- We introduce Positive Predictive Effect Probabilistic Segmentation via Weighted Gaussian Mixtures (PPEPS-WGM) to facilitate an LVLM to translate model behavior into natural language to detect spurious features.
- We demonstrate on different synthetic and real-world scenarios that our approach can effectively reduce the reliance of vision models on shortcuts by aligning them with human domain knowledge, without requiring any fine-grained feedback.

2. Related Work

Previous works have addressed the challenge of debugging models relying on spurious correlations by fine-tuning the model with human critique based on explanations of the current model behavior [23, 29]. Thereby, these methods improve the alignment of the model with human reasoning. However, they often require extensive fine-grained feedback for each image [25]. Furthermore, explanations of the current model behavior and feedback on potential errors must be provided directly in the image space [23]. This results in an inefficient interaction with the model. Stammer et al. [28] introduce a method to allow a Vision Language Model (VLM) to internally critique its own explanations independent of external human input increasing the model's performance but not explicitly aiming for aligning the model with human domain knowledge. Furthermore, the method does not translate to general small task-specific vision models for broad classification tasks. The authors in Zheng et al. [34] utilize a general-purpose captioning model to extract textual concepts without human steering from images and define a spuriousness score for each concept based on the accuracy of the classifier with and without that concept. However, the proposed captioning models might be incapable of identifying concepts for settings not explicitly included in the training data, e.g., medical or industrial images, and are limited to relatively discrete, co-occurring concepts, making it unsuitable for spurious features that manifest as subtle, continuous variations such as slight color differences. Gu et al. [7] introduce an approach to use an LVLM to provide explanations of the model's decision in natural language. However, they do not consider the natural language interface to inject human feedback back into the model. In contemporary work, Kuhn et al. [16] proposed a highly specialized VLM-based method for mitigating shortcuts in vision transformers. Aside from the previously mentioned approaches, non-human-centred methods focus on mitigating shortcuts without describing them directly in image space. These methods instead address the issue by balancing the model's performance between groups categorized by the class label and the presence or absence of spurious features. These approaches do not require instance-wise feedback about the location of spurious features; however, they do need additional annotations about the presence of spurious features per image. Kirichenko et al. [12] propose Deep Feature Reweighting (DFR), in which they only retrain the final layer of the vision model on a small, balanced dataset. This is based on the assumption that core features are often already learned during the initial training phase and simply need to be reweighted to improve performance on the test set. Idrissi et al. [11] demonstrate that straightforward data balancing techniques, such as subsampling or reweighting based on group frequencies, can deliver competitive worst group accuracy without the need for sophisticated training procedures. Lastly, Liu et al. [20] introduce Just Train Twice (JTT), whereby an initially on a few epochs trained model identifies challenging examples, and a second model is then trained on a reweighted dataset that up-samples these examples, aiming to reduce reliance on spurious correlations. These non-human-centric methods solely target equal performance across groups, regardless of the features used to achieve it. More precisely, those methods do not aim to directly align the model with human domain knowledge and thus lack explainability.

3. Problem Setting

Assume we have a vision model $f : \mathcal{X} \to \mathcal{Y}$ trained on a labeled training dataset $D_s = (x_s, y_s)_{i=1}^{n_s}$ with input images $x_s \in \mathbb{R}^{C\times H\times W}$ and labels $y_s \in \{1, ..., K\}$. For every class there exists a human specification $\mathcal{V}_k$ elaborating on important features to identify the specific class k. Further, explanation function $\Phi : \mathcal{X}\times\mathcal{Y}\times f \to \mathbb{R}^{H\times W}$ generates explanation maps in the original image space indicating which regions the model f considers important for predicting the output y given the input x. However, the trained model f might focus on areas identified by the explanations $\Phi(x, y, f)$ which do

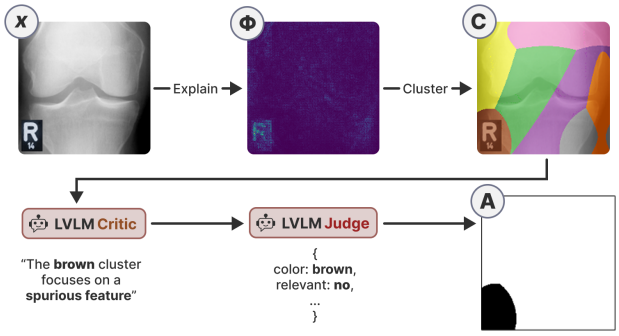

Figure 2. Correction mask generation process by the Critic & Judge pair for a vision model trained on a knee radiograph dataset. The image shows a hospital tag in the bottom left, which the model learned as a shortcut to classify the condition of the knee. All images are of size 224×224.

not match the description $\mathcal{V}_k$ indicating insufficient alignment of the model with domain knowledge and the reliance on spurious features. This results in reduced performance when the model is applied to test samples $D_t = (x_t, y_t)_{i=1}^{n_t}$ not subject to spurious correlations.

4. LVLM-Aided Visual Alignment (LVLM-VA)

We introduce our LVLM-Aided Visual Alignment (LVLM-VA) method as a two step approach to reduce the reliance of any vision classification model on spurious features.

4.1. Detecting Spurious Correlations

In the initial steps of our LVLM-VA approach (Fig. 2), a combination of XAI and an LVLM is used to generate an instance-wise correction signal aligning class-level human specifications and instance-level model explanations.

Sampling Strategy: While our procedure for detecting spurious features and aligning the original model can, in principle, be applied to the full training set, doing so would cause an unnecessary computational overhead on large datasets due to the reliance on a LVLM. To address this, we preferably generate the steering signal only for those samples on which the model relies on shortcuts. We propose an unsupervised sampling strategy based on the model's output entropy, motivated by the assumption that shortcuts are easier to learn than robust core features [10]. As a result, the model tends to exhibit lower output entropy on shortcut dependent training examples. Consequently, in subsequent steps, our alignment dataset D_{align} is defined as the N training samples with the lowest output entropy under the original model f. In contrast to other shortcut-mitigation approaches [11, 12], our method does not require additional group labels to identify samples affected by spurious features.

Positive Predictive Effect Probabilistic Segmentation via Weighted Gaussian Mixtures (PPEPS-WGM):

Based on the alignment set D_{align}, we use a large vision language model (LVLM) to identify potential spurious features. Following Yang et al. [32], we introduce a pre-segmentation step on the alignment images. The authors in [32] employ a Segment Anything Model (SAM) [13] to partition images based on visual content, showing that such pre-segmentation improves LVLM's capability of spatial location. However, we seek segmentation to support the detection of spurious features. Thus, we generate explanation maps $\Phi(x, y, f)$ in image space on the alignment dataset D_{align} as a proxy for the current decision process of the vision model f. Following, we introduce *Positive Predictive Effect Probabilistic Segmentation via Weighted Gaussian Mixtures (PPEPS-WGM)*, which performs model-centric segmentation by fitting a weighted Gaussian mixture whose components cluster regions according to their positive predictive effect shown in $\Phi(x, y, f)$. This steers the pre-segmentation and the LVLM's subsequent spatial allocation toward regions that most influence f's predictions rather than object boundaries, enabling more targeted identification of spurious features, i.e., regions of high positive attribution.

Setting: Let f be a trained model and let $b(x)$ denote the *additive* quantity we explain, i.e., the output logit of the target class. As an explanation method, we use DeepLIFT-SHAP [21] because of the convenient theoretical properties of Shapley values. More precisely, given an input image $x \in \mathbb{R}^{H \times W}$ with $M = H \cdot W$ pixels indexed by $i \in \{1, \dots, M\}$, Shapley values produce per-pixel attributions $\{\Phi_i(x)\}_{i=1}^{M}$ that satisfy *local accuracy*:

$$\sum_{i=1}^{M} \Phi_i(x) \;=\; b(x) - \mathbb{E}[b(X)].$$

Thus, $\{\Phi_i\}$ form a finite signed measure over pixels, where the attribution of a region equals the sum of its constituents and the total effect mass is conserved. In the following steps we focus on positive contributions as we consider spurious correlations as signals that falsely increase the score of the given prediction. We define the positive part of the model attribution and its total mass as:

$$\Phi_i^+(x) \;=\; \max\{\Phi_i(x),\, 0\}, \qquad Z^+(x) \;=\; \sum_{i=1}^{M} \Phi_i^+(x).$$

We subsequently normalize $\Phi_i^+(x)$ to obtain a discrete probability mass function (PMF) over pixels

$$p_i(x) \;=\; \frac{\Phi_i^+(x)}{Z^+(x)}, \qquad \sum_{i=1}^{M} p_i(x) = 1.$$

With that $p_i(x)$ is the probability that a randomly sampled unit of *positive predictive effect* (positive additive change in b) lies at pixel i.

Weighted Gaussian Mixture on the Positive-Effect Distribution: Let $z_i \in \mathbb{R}^d$ denote per-pixel features used for clustering. As we would like to segment where in the image the positive effect mass lives, not what the underlying content is, we instantiate z_i as *normalized spatial coordinates* on the equidistant image lattice,

$$z_i = \big((u_i + \tfrac{1}{2})/H,\ (v_i + \tfrac{1}{2})/W\big) \in [0,1]^2,$$

with $(u_i, v_i) \in \{0, \dots, H-1\} \times \{0, \dots, W-1\}$ the pixel indices. We fit a J-component Gaussian mixture to the discrete distribution p by weighted maximum likelihood:

$$\max_{\Theta} \mathcal{L}(\Theta) \;=\; \sum_{i=1}^{M} w_i \,\log\!\left(\sum_{j=1}^{J} \pi_j \,\mathcal{N}\big(z_i \mid \mu_j, \Sigma_j\big)\right),$$
$$w_i \;=\; M \cdot p_i(x),$$

where $\Theta = \{(\pi_j, \mu_j, \Sigma_j)\}_{j=1}^{J}$ are the mixture parameters with $\pi_j \geq 0$ and $\sum_j \pi_j = 1$. We define the responsibilities

$$r_{ij} \;=\; \frac{\pi_j \,\mathcal{N}(z_i \mid \mu_j, \Sigma_j)}{\sum_{\ell=1}^{J} \pi_\ell \,\mathcal{N}(z_i \mid \mu_\ell, \Sigma_\ell)}.$$

The M-step sufficient statistics are formed with effective weights $w_i r_{ij}$ for pixel i and component j. The mixture-weight update equals the positive-effect share captured by component j:

$$\pi_j^{\text{new}} \;=\; \frac{\sum_{i=1}^{M} w_i r_{ij}}{\sum_{i=1}^{M} w_i} \;=\; \sum_{i=1}^{M} p_i(x)\, r_{ij} \;=\; S_j,$$

where $S_j \in [0,1]$ and $\sum_j S_j = 1$. Hence π_j at optimum is interpretable as the *fraction of total positive effect* explained by component j.

Lastly, we derive a segmentation by assigning each pixel to its most probable component:

$$c_i \;=\; \arg\max_{j \in \{1,\dots,J\}} r_{ij}, \qquad C \in \{1, \dots, J\}^{H \times W}$$

given after reshaping $\{c_i\}_{i=1}^{M}$.

Generating LVLM-based Alignment Verdicts: Following the clustering step, the segmented explanation map C, together with the original image x and the ground truth label y, are provided to the LVLM-based Critic g. Further, a third image showing the original image overlapped with the segmentation is provided to the Critic in order to support the correct localization of the segments. To facilitate g in detecting if the vision model f relies on spurious features, it is instructed to utilize a chain-of-thought process [31]. The introduced prompt guides the model to (1) examine the original image, (2) identify which regions belong to the ground truth class y, (3) determine for each segmented cluster which parts of the original image are included, (4) combine both insights, and (5) describe if a cluster covers a relevant region, and (6) lastly provide a verdict whether a cluster is relevant based on the previous insights. To further allow to steer the LVLM in this process and emphasize what important concepts define a particular class, class-specific prompts include human-defined descriptions $\mathcal{V}_k$ about how to accurately recognize the class k. As these descriptions allow scaling class-level human feedback to instance-wise critique, they drastically decrease human effort for aligning the model. To seamlessly integrate the LVLM assessment into an automatic training pipeline and reduce the overall complexity of the task, we employ an LLM Judge h (which may be instantiated as the same model as g) that maps the free-form output of g for x to a final binary verdict R. It determines whether each cluster corresponds to a spurious feature, yielding a single binary verdict R_j for every cluster j in C. A class-agnostic prompt for h further steers this verdict by providing example pairs of Critic assessments and their associated binary human judgments. Evaluating prototypical outputs of g and specifying aligned verdicts establishes another option for bidirectional feedback, giving the human expert a mechanism to influence the final decision. In addition to aligning the outcome with human knowledge, prior work shows that such few-shot exemplars can substantially increase LLM performance on specified tasks [2].

4.2. Visual Alignment using LVLM Verdicts

Different previous works have focused on correcting model explanations [23, 27] by aligning them with fine-grained human feedback [25] in the form of instance-wise corrections in image space. In our LVLM-VA approach, we utilize the Right for the Right Reasons (RRR) loss function introduced by Ross et al. [23] for the alignment:

$$L(\theta, X, y, A) = \sum_{n=1}^{N}\sum_{k=1}^{K} -y_{nk}\log(\hat{y}_{nk}) + \lambda \sum_{n=1}^{N}\sum_{i=1}^{M}\left(A_{ni}\frac{\partial}{\partial x_{ni}}\sum_{k=1}^{K}\log(\hat{y}_{nk})\right)^2 + \gamma\sum_i \theta_i^2$$

Here, N is the number of used alignment samples, K refers to the number of classes, and M is the dimensionality of the input x. The first term "right answers" corresponds to the cross-entropy loss, optimizing the model to make correct classification predictions. The second term "right reasons" ensures that the model's decisions are based on relevant features by reducing the gradient in regions deemed irrelevant by experts via a binary mask A, steering the model to focus on important features and avoid spurious correlations. Additionally, an optional term "regularization" on the model

parameters θ can be added to prevent overfitting.
We automatically transfer the binary verdicts generated via the Critic & Judge pair into the correction maps A:

$$A = \sum_{j=1}^{J} R_j \cdot \mathbf{1}\left[C = j\right] \ ,$$

where the cluster verdict R_j is applied to the corresponding cluster j in the segmented explanation map C such that A only features clusters considered to be spurious. By this, we render the previously required tedious per-instance interaction to generate the expert maps obsolete. For the fine-tuning using the RRR loss, the alignment samples x_a are mixed with the training samples x_s in each batch of size I with a ratio of $\frac{I_{x_a}}{I_{x_s}}$. An epoch for N_{Train} training samples consists of $\frac{N_{Train}}{I_{x_s}}$ train iterations, in the case that this is greater than $\frac{N}{I_{x_a}}$, the alignment samples are over-sampled. With this procedure, we aim to avoid catastrophic forgetting of previously learned core features. In summary, fine-tuning the original model f using the RRR loss with instance-wise masks generated automatically based on human specifications allows our method to significantly reduce manual effort whilst still allowing human steering of the vision model.

5. Experiments

We evaluate our approach using three different datasets and two shortcut learning settings. In the first multi-class classification setting, artificial spurious decoys occur throughout the entire training set and their appearance correlates with the classes, but they are absent in the test set, which leads to low test performance. This setting may reflect a systematic bias in the data generation process, such as different camera settings when the model is trained and deployed. In the second setting, we evaluate two real-world binary classification tasks where the spurious features remain similar in both the training and test sets, but their frequency of occurrence in the training set depends on the class, whereas they are equally distributed when the model is deployed. This could be attributed to a shift between the training and test distributions. Within these settings, we benchmark our approach against constrained optimization with instance-level human feedback [23] and shortcut mitigation strategies that aim to reduce differences between group accuracies [11, 12, 20]. We use DeepLiftSHAP [21] to generate the explanation maps $\Phi(x, y, f)$ across all experiments. For the Critic and Judge, we use a GPT-4o model. The remainder of this section first presents the multi-class setting with artificial decoys before moving on to the real-world medical setting. More details are documented in the Appendix and code is provided at: `https://github.com/alexanderkoebler/LVLM-VA`.

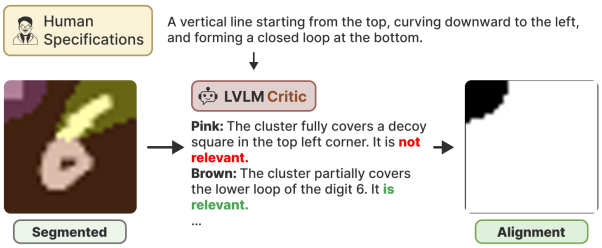

Figure 3. Intermediate results for aligning an MLP model for classifying DecoyMNIST. Based on the input of the original image, the segmentation map and the class level description, the LVLM-Critic correctly identifies that the top left corner includes the spurious decoy. The LLM-Judge assigns the right binary label which is subsequently transferred into the correction mask where black refers to 'not relevant'.

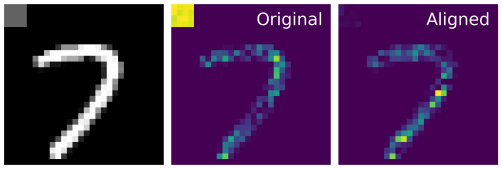

Figure 4. The explanations generated for a test example for an MLP model trained on DecoyMNIST before and after the alignment step. The original model clearly focuses on the spurious decoy in the upper left corner whereas the attribution of the aligned model is almost fully distributed across the actual digit.

5.1. Mitigating Systematic Decoys Across the Entire Training Set

In the first scenario, we evaluate our approach using an artificial decoy dataset based on digit classification [19] with added artificial decoys. This dataset is frequently used in the literature to study model debugging [1, 23].
Experimental Setting: Each image in the dataset contains a grey patch in a random corner. While the shade of grey for the samples in the training set depends on the digit k $(255 - 25 \cdot k)$, it is chosen randomly in the test set. As a result, these patches represent simple shortcut candidates for the model across all classes in the training set but act as harmful confounders in the test set. For this experiment we train a two-layer Multi-Layer-Perceptron (MLP) of width 256. For the alignment set we use $N = 256$ samples x_a from the training set using the previously described sampling strategy. The number of clusters J for PPEPS-WGM is set to three which, as shown in Fig. 3, is sufficient for allowing to clearly identify the spurious features.

Results: Subsequently, we first qualitatively evaluate the effect of the alignment step on the model's reliance on spurious decoys. The explanations of the MLP model in Fig. 4 show a complete reassignment of the model's attention to the actual digit. This indicates that the model is significantly less affected by the spurious features introduced during training.

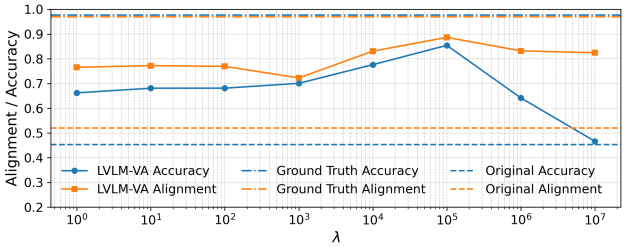

Figure 5. Alignment and accuracy on the test set across different λ values weighing the influence of the RRR loss term during alignment. Alignment as well as accuracy rise with increase in λ until 10^5 where the accuracy drops substantially. At $\lambda = 10^5$ both metrics are close to the upper bound reached when using the ground truth correction masks.

To also quantitatively evaluate the benefit of our method in this setting, we use the intuitive lower and upper bounds for human involvement as a baseline. The least human involvement is given by not aligning the model at all and only using the simple target labels during the original training process. In contrast, using instance-wise human-generated expert masks $A^{(GT)}$ in combination with the RRR loss requires significant manual effort if not automatically generated as in this synthetic setting. The availability of information about the location of the spurious features in this scenario allows us to also quantitatively evaluate the alignment of the model. For this, we introduce an alignment metric adapted from [14, 15] between the ground truth masks $A^{(GT)}$, i.e., the artificial decoys in the corner, and the absolute explanation maps after alignment $|\Phi(x, y, f)|$. This metric measures the fraction of attribution mass that lies outside the ground-truth spurious region and therefore on the relevant digit features. For N_t test samples it is defined as

$$\mu_{Align} = 1 - \frac{1}{N_t} \sum_{n=1}^{N_t} \frac{\sum_{i=1}^{M} A_{n,i}^{(GT)} |\Phi_i(x_n, y_n, f)|}{\sum_{i=1}^{M} |\Phi_i(x_n, y_n, f)|} \,.$$

As shown in Fig. 5, our LVLM-VA approach improves both performance and model alignment. This effect is emphasized by increasing the influence of the RRR loss up to $\lambda = 10^5$. However, at this point, performance drops drastically as the cross-entropy loss becomes negligible. LVLM-VA achieves values approaching those obtained using ground truth instance-wise feedback within the RRR loss, which requires substantial manual labelling effort.

5.2. Mitigating Spurious Correlations in Real-World Medical Datasets

In the following two experiments, we investigate the effectiveness of LVLM-VA to mitigate learned shortcuts in two real-world medical datasets shown in Fig. 6. In the medical domain, it is of vital importance to ensure consistent model performance despite often limited data quality and

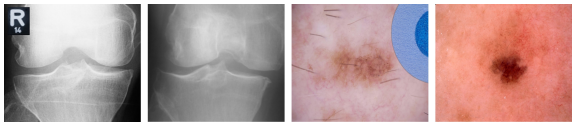

Figure 6. Prototypical images for the medical datasets. Some of the knee radiograph images (left) include spurious hospital tags whereas the skin lesion images (right) include colored bandages. Both of those spurious features might be learned as simple shortcuts compared to the complex original classification task.

quantity. Therefore, the model's performance should be independent of any spurious elements introduced in a subset of the available images that could lead to biases for any given (potentially protected) group. For this reason, we measure the effectiveness of our approach by its ability to improve the accuracy of the worst performing group (a standard metric for shortcut mitigation strategies [12, 34]) whilst at least maintaining overall accuracy. As the manual curation of medical datasets is very laborious and the time of medical experts is limited, LVLM-VA can help to automatically align the model with high-level expert input.

Experimental Setting: First, we evaluate our method on the *International Skin Imaging Collaboration (ISIC) skin lesion dataset* [4], which is also used in multiple other shortcut mitigation works [1, 18, 22]. This dataset contains images of real skin lesions, which are either benign or malignant tumors. Some of the images contain bandages of different colours, which are located randomly next to the skin lesions. Our training set consists of 1,800 samples per class. For the benign class, there are an equal number of images containing and not containing coloured bandages. In contrast, only ten images in the malignant class contain one or multiple bandages. This difference in the occurrence of bandages across the classes renders them a spurious feature that can easily be learned as a shortcut during the initial training phase. In the test set, bandages occur equally frequently, leading to low accuracy for images of malignant lesions containing bandages. For our LVLM-VA approach, we use an alignment dataset of size $N = 1024$ and provide short human descriptions for both classes.

Secondly, we use a *knee osteoarthritis radiograph dataset* [3, 16] that includes images depicting various stages of osteoarthritis. We evaluate the binary classification task of distinguishing between 'no' and 'moderate' osteoarthritis. Similar as done by DeGrave et al. [6], shortcuts are added in the form of hospital tags specified as 'L' or 'R' to the edges of the images, indicating the right or left knee. The occurrence of these tags in the training set is class-dependent, with 50% of healthy knees and only 2.5% of arthritic knees co-occurring with hospital tags. In total, there are 1,000 training samples. Furthermore, there are 400 test samples, distributed equally across the groups. As expected, the hospital tags lead to spurious shortcuts, resulting in low group accuracy

for arthritic knee images with hospital tags. The alignment dataset includes $N = 256$ samples, and we provide short human descriptions for both classes.

For both experiments, we use a ResNet50 model [9] and set $\lambda = 1$. We set the number of clusters J to seven. The performance is stable for reasonably large J (see Appendix). We benchmark LVLM-VA against the three commonly applied shortcut mitigation methods: sub-sampling groups (SUBG) [11], Deep Feature Reweighting [12] (DFR), and Just Train Twice [20] (JTT).

Results: Fig. 7 illustrates the initial stage of identifying potential spurious features and generating the corresponding correction maps A for the skin dataset. The proposed clustering approach, based on model explanations combined with PPEPS-WGM, generates one or more clusters covering irrelevant features without substantially overlapping the core features relevant for classification (i.e. the knee or skin lesion). The provided human descriptions help the LVLM-Critic identify which parts of the image should be considered core features, given a particular class. Based on this information, the LVLM-Critic provides intermediate assessments, enforced by the chain-of-thought prompt, as well as a clear, combined statement about the relevance of the considered cluster. This enables the Judge to efficiently generate a structured binary verdict for each cluster. These verdicts are then translated into the displayed correction masks. For knee radiographs, where the LVLM is unable to judge and is not intended to identify which part of the knee is relevant, it only declares the cluster that actually includes the spurious hospital tag as irrelevant (see Fig. 2). In contrast, for the ISIC dataset, where it is clear that clusters not covering the skin lesion should be irrelevant for the classification, the LVLM also labels clusters as irrelevant even though they do not include a bandage (see Fig. 7). This approach enables spurious features that are neither described by humans nor detected by the critic to be removed (e.g. slight shadows or reflections) without eliminating potentially important core features and sacrificing overall performance. In Fig. 8 and Fig. 9, the change in the model's average group / overall accuracy (ΔAGA) and the worst group accuracy (ΔWGA) before and after the shortcut mitigation step for LVLM-VA is compared to that of the three baselines. Across both experiments, LVLM-VA enables a significant improvement in the worst group accuracy without diminishing the overall accuracy. The two step JTT approach enables a slight improvement in WGA for the skin dataset, but does not consistently improve performance for both datasets. Subsampling groups (SUBG) improves WGA, even outperforming LVLM-VA on the knee dataset, but significantly sacrifices overall accuracy, rendering it invalid for most applications. Deep Feature Reweighting (DFR) was not beneficial in either experiment. The experiments on medical datasets demonstrate that LVLM-VA is the most effective way of mitigating group biases caused by spurious features, without requiring to specify which samples are affected by shortcuts. These instance-wise group labels are significantly more labor intensive to generate than human descriptions at the class level.

Figure 7. Prototypical shortcut detection results for the *skin lesion dataset*. Only the green cluster in the segmented input image covers the actual skin lesion. This enables the LVLM-Critic, which is informed by the human class description, to deem all other clusters as irrelevant. This produces a correction mask A, where gradients will be penalized except in the green segment during the fine-tuning step.

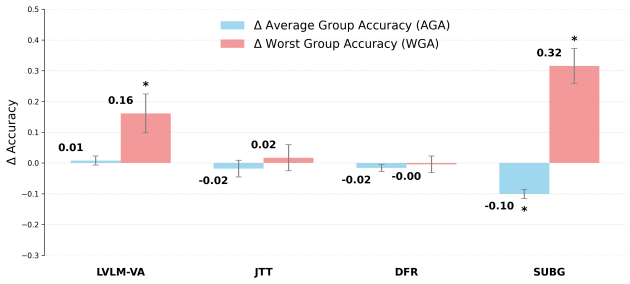

Figure 8. Change in Average Group Accuracy (AGA) and Worst Group Accuracy (WGA) relative to the original model after shortcut mitigation on the *knee radiographs dataset*. Results are averaged over seven random seeds (mean ± std). LVLM-VA is the only method which increases the WGA whilst maintaining overall accuracy. (*: Wilcoxon Signed-Rank Test $p < 0.05$)

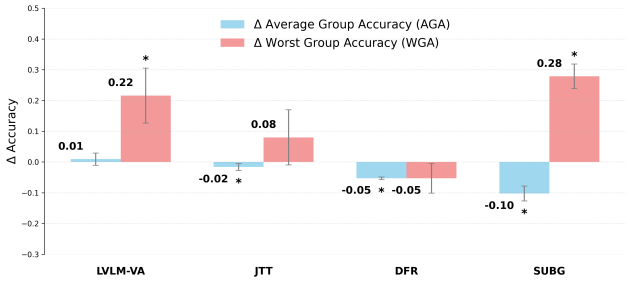

Figure 9. Change in Average Group Accuracy (AGA) and Worst Group Accuracy (WGA) relative to the original model after shortcut mitigation on the *skin lesion dataset*. Results are averaged over seven random seeds (mean ± std). LVLM-VA is the only method which increases the WGA whilst maintaining overall accuracy. (*: Wilcoxon Signed-Rank Test $p < 0.05$).

6. Ablations and Discussion

In this section, we present a user study to evaluate the effect of LVLM-VA on the perceived alignment of vision models. In addition, we conducted several experiments that examine the choice of sub-components of LVLM-VA and further discuss the limitations and scope.

User Study on Alignment: The conducted user study includes 18 participants with a data science background and familiarity with medical imaging. They assess three aspects across 26 questions covering both medical datasets: (1) *Cluster relevance*: Participants are asked to select irrelevant/spurious clusters given the original and the segmented image. They agreed with the LVLM-selected clusters in 88% of cases, which is in line with the measured verdict accuracy of 87% on the entire alignment set. (2) *Critic reasoning*: Participants are presented with the Critic's natural-language argumentation in addition to the previous images and tasked to assess its correctness. They agreed with the LVLM's argumentation in 87% of cases. (3) *Explanation alignment*: The participants are shown the target model explanations before and after the alignment step. Participants conclude that the post-alignment explanations are better aligned with their expectations in 86% of cases. This is consistent with the alignment analysis on DecoyMNIST (Fig. 4 & Fig. 5). Together, these results underline that LVLM-VA can significantly contribute to a better alignment of vision models with human expectations.

LVLM Choice: Our method leverages recent advances in rapidly evolving LVLMs. During the main development phase of this work, GPT-4o was used. However, Table 1 shows that newer models outperform older ones in verdict accuracy, which measures whether a cluster flagged as spurious actually contains part of the spurious feature. This higher verdict accuracy leads to an increased WGA, while at the same time, newer models continue to decrease in cost, thereby steadily improving the accessibility of our method.

Sampling Strategy: We further reduced the LVLM costs by introducing a low-entropy sampling strategy, which preferentially applies the Critic & Judge only to samples containing spurious features. On the knee dataset, this strategy produces an alignment set in which 56% of images contain spurious features, compared to only 25% under random sampling and 2% when sampling based on high entropy.

Table 1. Benchmark verdict accuracy (knee) and ΔWGA for different Critic/Judge LVLMs. Costs in USD per one million input tokens (Nov. 2025). (*: Wilcoxon Signed-Rank Test for the improvement in WGA w.r.t. the original model across 7 seeds $p < 0.05$)

Model	Cost	Verdict Acc.	ΔWGA
GPT-4o (used)	2.50	0.87	0.16 ± 0.06*
GPT-5	1.25	1.00	0.20 ± 0.09*
GPT-4o-mini	0.15	0.42	0.09 ± 0.02*

Segmentation Method: To make the most effective use of the limited alignment set, we introduced PPEPS-WGM as a segmentation method that targets clusters with high positive attribution density, rather than segmenting the underlying image content directly. To assess the benefit of PPEPS-WGM, we compare it to segmenting the input images with a Segment Anything Model (SAM) [13] as done by Yang et al. [32]. Table 2 shows that, although the verdict accuracy of the two approaches is similar, the overall improvement in ΔWGA achieved with SAM is smaller. This is because SAM frequently groups the spurious feature together with the relevant knee structures into a single segment. In contrast, PPEPS-WGM more effectively isolates the spurious feature by clustering spatially separated positive attribution.

Table 2. Verdict accuracy and corresponding increase in ΔWGA (knee) for SAM_VIT_B model (SAM) and PPEPS-WGM. (*: Wilcoxon Signed-Rank Test for the improvement in WGA w.r.t. the original model across 7 seeds $p < 0.05$)

Method	Verdict Acc.	Δ WGA
PPEPS-WGM	0.87	0.16 ± 0.06*
SAM	0.87	0.11 ± 0.04*

Limitations: Although LVLM-VA removes the need for fine-grained annotations, it relies on class-level descriptions provided by domain experts. In some cases, these descriptions may be difficult to formalize, as experts have rather learned those patterns intuitively. Furthermore, the distinction between core and spurious features is not always clear or spatially separable. We address this issue by having differing degrees of intervention between the skin lesion and knee radiograph datasets. While the LVLM intervenes on clear spurious features in the case of the synthetic decoy and knee dataset, it rather focuses on preserving clearly described core features in the form of skin lesions for the ISIC dataset. Being able to explicitly describe either core or spurious features applies to most real-world use-cases making our method applicable to a wide variety of settings.

7. Conclusion

We have proposed a novel approach, LVLM-Aided Visual Alignment (LVLM-VA), to correct spurious correlations and increase the overall and worst group accuracy of small, task-specific vision models. LVLM-VA translates model behavior into natural language and incorporates human descriptions at the class level via instance-wise critique into the model. This provides an efficient human-centered interface for aligning vision models with domain knowledge, eliminating the need for expensive fine-grained feedback or group labels. LVLM-VA fosters synergies between generative AI models and more explainable, established discriminative approaches.

References

[1] Yanzhe Bekkemoen and Helge Langseth. Correcting classification: A bayesian framework using explanation feedback to improve classification abilities. *arXiv preprint arXiv:2105.02653*, 2021.

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[3] Pingjun Chen, Linlin Gao, Xiaoshuang Shi, Kyle Allen, and Lin Yang. Fully automatic knee osteoarthritis severity grading using deep neural networks with a novel ordinal loss. *Computerized Medical Imaging and Graphics*, 75:84–92, 2019.

[4] Noel C. F. Codella, David Gutman, M. Emre Celebi, Brian Helba, Michael A. Marchetti, Stephen W. Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, and Allan Halpern. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 168–172, 2018.

[5] Thomas Decker, Ralf Gross, Alexander Koebler, Michael Lebacher, Ronald Schnitzer, and Stefan H Weber. The thousand faces of explainable ai along the machine learning life cycle: industrial reality and current state of research. In *International Conference on Human-Computer Interaction*, pages 184–208. Springer, 2023.

[6] Alex J. DeGrave, Joseph D. Janizek, and Su-In Lee. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3(7):610–619, 2021.

[7] Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Ming Tang, and Jinqiao Wang. Anomalygpt: Detecting industrial anomalies using large vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1932–1940, 2024.

[8] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14375–14385, 2024.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[10] Katherine L. Hermann, Hossein Mobahi, Thomas Fel, and Michael Curtis Mozer. On the foundations of shortcut learning. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

[11] Badr Youbi Idrissi, Martin Arjovsky, Mohammad Pezeshki, and David Lopez-Paz. Simple data balancing achieves competitive worst-group-accuracy. In *Conference on Causal Learning and Reasoning*, pages 336–351. PMLR, 2022.

[12] Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. *ICLR 2023*, 2023.

[13] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.

[14] Alexander Koebler, Christian Greisinger, Jan Paulus, Ingo Thon, and Florian Buettner. Through the eyes of the expert: Aligning human and machine attention for industrial ai. In *Artificial Intelligence in HCI: 5th International Conference, AI-HCI 2024, Held as Part of the 26th HCI International Conference, HCII 2024, Washington, DC, USA, June 29–July 4, 2024, Proceedings, Part II*, page 407–423, Berlin, Heidelberg, 2024. Springer-Verlag.

[15] Maximilian Kohlbrenner, Alexander Bauer, Shinichi Nakajima, Alexander Binder, Wojciech Samek, and Sebastian Lapuschkin. Towards best practice in explaining neural network decisions with lrp. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2020.

[16] Lukas Kuhn, Sari Sadiya, Jörg Schlötterer, Florian Buettner, Christin Seifert, and Gemma Roig. Efficient unsupervised shortcut learning detection and mitigation in transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2217–2226, 2025.

[17] Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking clever hans predictors and assessing what machines really learn. *Nature Communications*, 10, 2019.

[18] Phuong Quynh Le, Jörg Schlötterer, and Christin Seifert. Is last layer re-training truly sufficient for robustness to spurious correlations? *arXiv preprint arXiv:2308.00473*, 2023.

[19] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[20] Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*, pages 6781–6792. PMLR, 2021.

[21] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, page 4768–4777, Red Hook, NY, USA, 2017. Curran Associates Inc.

[22] Meike Nauta, Ricky Walsh, Adam Dubowski, and Christin Seifert. Uncovering and correcting shortcut learning in machine learning models for skin cancer diagnosis. *Diagnostics*, 12(1), 2022.

[23] Andrew Slavin Ross, Michael C. Hughes, and Finale Doshi-Velez. Right for the right reasons: training differentiable models by constraining their explanations. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, page 2662–2670. AAAI Press, 2017.

[24] Johannes Rueckel, Lena Trappmann, Balthasar Schachtner, Philipp Wesp, Boj Hoppe, Nicola Fink, Jens Ricke, julien Dinkel, Michael Ingrisch, and Bastian Sabel. Impact of confounding thoracic tubes and pleural dehiscence extent on artificial intelligence pneumothorax detection in chest radiographs. *Investigative Radiology*, Publish Ahead of Print, 2020.

[25] 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.

[26] Hua Shen, Tiffany Knearem, Reshmi Ghosh, Kenan Alkiek, Kundan Krishna, Yachuan Liu, Ziqiao Ma, Savvas Petridis, Yi-Hao Peng, Li Qiwei, et al. Towards bidirectional human-ai alignment: A systematic review for clarifications, framework, and future directions. *arXiv preprint arXiv:2406.09264*, 2024.

[27] Emanuel Slany, Yannik Ott, Stephan Scheele, Jan Paulus, and Ute Schmid. Caipi in practice: towards explainable interactive medical image classification. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pages 389–400. Springer, 2022.

[28] Wolfgang Stammer, Felix Friedrich, David Steinmann, Manuel Brack, Hikaru Shindo, and Kristian Kersting. Learning by self-explaining. *arXiv preprint arXiv:2309.08395*, 2023.

[29] Stefano Teso and Kristian Kersting. Explanatory interactive machine learning. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, page 239–245, New York, NY, USA, 2019. Association for Computing Machinery.

[30] Laura Von Rueden, Sebastian Mayer, Katharina Beckh, Bogdan Georgiev, Sven Giesselbach, Raoul Heese, Birgit Kirsch, Julius Pfrommer, Annika Pick, Rajkumar Ramamurthy, et al. Informed machine learning–a taxonomy and survey of integrating prior knowledge into learning systems. *IEEE Transactions on Knowledge and Data Engineering*, 35(1):614–633, 2021.

[31] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[32] Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. *arXiv preprint arXiv:2310.11441*, 2023.

[33] Xu Yang, Hanwang Zhang, Guojun Qi, and Jianfei Cai. Causal attention for vision-language tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9847–9857, 2021.

[34] Guangtao Zheng, Wenqian Ye, and Aidong Zhang. Learning robust classifiers with self-guided spurious correlation mitigation. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 5599–5607, 2024.