KEEP YOUR PEERS: TASK SELECTION FOR ROBUST EXPLANATIONS IN MULTI-TASK LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Reliable deployment of large vision-language models (LVLMs) requires explanations that remain stable under realistic input perturbations, not merely accurate predictions. Yet multi-task learning (MTL) and auxiliary-task selection have been studied predominantly through predictive objectives, with comparatively little attention to explainability robustness. This raises a fundamental question: which auxiliary tasks should be selected to enhance the robustness of explanations for a primary task? This work addresses the question by formulating an explanation instability functional and analyzing its first variation under changes in auxiliary-task participation. The resulting variational analysis is generic and attribution-agnostic: it operates at the level of the explanation functional and can be applied with any differentiable attribution map, rather than being tied to a particular explanation method, model architecture, or application. Motivated by this analysis, we introduce **PEER**(**Perturbation Explanation Error Relation**), a criterion based on the interaction between perturbation-induced explanation errors of the primary and auxiliary tasks. Under suitable regularity conditions, this relation identifies auxiliary tasks whose retention enhances the robustness of primary-task explanations. The framework is realized in LLaVA-PEER, a LLaVa model trained with PEER-MTL, for chest X-ray analysis in resource-constrained mHealth settings, where severe image compression provides a realistic source of perturbation. The study also releases a radiologist-annotated free-text reporting dataset for publicly available tuberculosis CXR benchmarks. Experiments show that PEER-MTL reduces explanation instability and yields substantially more reliable LVLM-generated reports.

1 INTRODUCTION

Multi-task learning (MTL) improves generalization by sharing representations across related tasks, raising data efficiency and enabling positive transfer. It does so only when the tasks cooperate. When they compete for capacity, or when no single shared representation serves all of them, joint training produces *negative transfer* and the auxiliary tasks degrade the primary one. For a given primary task, deciding which auxiliary tasks to train alongside it is therefore a central question, and existing answers—task-affinity estimates (Standley et al., 2020; Fifty et al., 2021; Li et al., 2024) and online gradient-conflict resolution (Yu et al., 2020; Wang et al., 2021; Liu et al., 2021a)—all answer it with respect to the same objective: predictive accuracy. However, in certain scenarios, due to input perturbations, although the prediction remains stable, the explanation maps generated become unstable.

To motivate the study of such a question we draw inspiration from the setting of mHealth (mobile Health) diagnosis under resource constraints. In rural tuberculosis (TB) screening, chest X-rays (CXRs) are relayed to a centralized AI service over a low-bandwidth messaging platform such as WhatsApp, a diagnosis is generated, and a report is returned over the same channel. The transport is lossy: WhatsApp compresses images by up to 90%. It has been noted that network intricacies induces image perturbation and it induces two failure modes in Vision Transformers—*prediction instability*, where the diagnosis disagrees between the original and compressed image, and *out-of-lung saliency* (OLS), where the model's attribution map shows regions outside the lungs as major contributors for prediction (DeGrave et al., 2021; Geirhos et al., 2020; Antony et al., 2025). Figure 1 shows the effect: images a radiologist would call identical yield entirely different attributions.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

The goal of this paper is to leverage multi-task learning (MTL) for explanation robustness of a primary task under image perturbations. Selecting auxiliary tasks for explanation robustness raises challenges such as: *(i) Search is exponential and evaluation is expensive.* There are 2^m candidate subsets, and scoring one on explanation robustness requires retraining the model and recomputing attributions which is computationally expensive. *(ii) Task-affinity measures are proxies for predictive performance.* Existing task-affinity measures do not characterize how a candidate auxiliary task will affect the stability of the primary task's explanations.

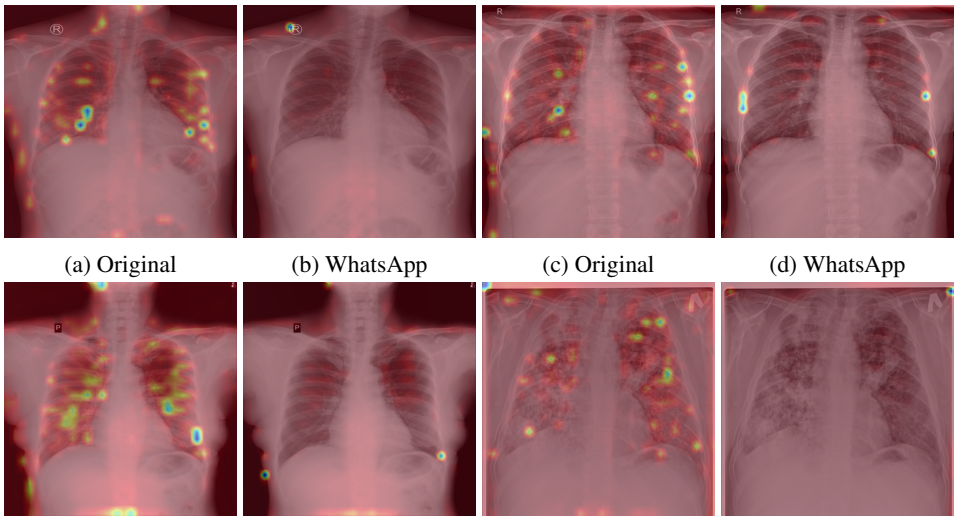
Contributions. We address these with a criterion computed from how explanations *move* under perturbations.

- **The PEER score.** We introduce the **Perturbation induced Explanation Error Relation**, the correlation between the perturbation-induced changes in the primary and an auxiliary task's attribution maps. To our knowledge, explanation behaviour has not previously been used as a signal of task relatedness.
- **The PEER-deletion principle.** We formalize **VAR**, the **Explanation Instability Functional**, as the measure of explanation instability defined by the second moment of the perturbation-induced change in an attribution map and study how it responds to deleting an auxiliary task. Assuming an explanation transfer relationship between **PEER** and **VAR**, validated by data, we arrive at a peer-deletion principle. For the two-task case it states that the change in **VAR** under deletion carries the *opposite* sign to the auxiliary task's **PEER** score (Proposition 1).
- **Keeping Explanation Enhancing Peers.** We extend the PEER-deletion principle to m tasks and prove that retaining exactly the *negatively* scored auxiliary tasks lowers `EIF` relative to training on all of them (Theorem 1). The resulting auxiliary task selection rule, **KEEP** (Keeping Explanation Enhancing Peers): the tasks worth keeping are those whose explanations move *against* the primary task's under perturbation, cancelling its perturbation response and acting as regularisers.
- **An explanation-robust medical LVLM. PEER**-MTL reduces out-of-lung saliency by up to 30% relative to training on all auxiliary tasks, and LLaVA-PEER, a LLaVa model trained with PEER-MTL, loses approximately 60% less report-generation quality under compression.
- **Curated radiology report.** Public TB CXR datasets (Montgomery, Shenzhen, and TBX11K) provide only binary labels. We curate radiologist-annotated free-text reports for their images and release them for research use.

2 Related Works

Medical foundation models. Recent advances in large vision-language models (LVLMs) have enabled the development of medical foundation models trained on paired medical image–text datasets. For instance, Med-Flamingo (Moor et al., 2023), CheXagent (Chen et al., 2024b), LLaVA-Rad (Zambrano Chaves et al., 2025) and XrayGPT (Thawkar et al., 2023) fine-tune LVLMs on chest X-ray (CXR) datasets such as MIMIC-CXR to support radiology report generation. Broader medical generalist models, including RadFM (Wu et al., 2025), BiomedGPT (Zhang et al., 2023), LLaVA-Med (Li et al., 2023a) and Uni-Med (Zhu et al., 2024) curates a medical multi-modal dataset spanning modalities including CXR, CT, and MRI. However, such generalist models require further fine-tuning on task-specific data to be effective for downstream clinical applications.

Explainability in Vision models. Feature attribution methods are a common class of explanation techniques that assign an importance score to each image pixel, indicating its contribution to the model's output. There are several feature attribution methods (Selvaraju et al., 2017; Carter et al., 2019; Lundberg & Lee, 2017) designed for convolutional and feed-forward architectures. For transformer architectures, one early approach, Attention Rollout (Abnar & Zuidema, 2020a), sequentially connects attention maps across all layers. Attention flow (Abnar & Zuidema, 2020b) contextualizes the attention mechanism as a max-flow problem; however, it is computationally demanding and has not been extensively evaluated for vision tasks. Chefer et al. (Chefer et al., 2021b) later introduced a rollout procedure that mean-weights attention activations using a combination of gradients and LRP-inspired relevances. Subsequent work (Chefer et al., 2021a) found that omitting LRP relevances and relying solely on positive mean-weighting of attention activations with gradients improved faithfulness within cross-attention layers. The same authors also adapted Grad-CAM (Selvaraju et al.,



(a) Original (b) WhatsApp (c) Original (d) WhatsApp

(e) Original (f) WhatsApp (g) Original (h) WhatsApp

Figure 1: Exemplified TB localization generated with Vit-B/16 architectures on original images and their corresponding WhatsApp compressed images. Although each pair looks visually identical and receives the same prediction, the localization maps differ substantially.

2017) to transformers by weighting the final attention map with gradients. More recently, LeGrad (Bousselham et al., 2025) uses gradients with respect to the attention maps of individual ViT layers as an explainability signal. A separate line of work (Ghorbani et al., 2019; Adebayo et al., 2018) construct perturbations that are imperceptible and leave the predictions unchanged, yet produce entirely different attribution map. To improve attribution robustness, (Chen et al., 2024a; Wang et al., 2020) propose to smoothen the model's decision surface by applying a regularizer.

Multi-Task Learning. Multi-task learning (MTL) has emerged as an effective strategy for improving generalization by sharing knowledge across related tasks (Evgeniou & Pontil, 2004). Earlier MTL methods based on optimization (Zhou et al., 2026b; Liu et al., 2021a) combine per-task gradients into a joint update direction using various heuristics (Sener & Koltun, 2018; Navon et al., 2022; Liu et al., 2021b; Kendall et al., 2018; Qin et al., 2025). However, these methods require computing and storing per-task gradients by performing a backward pass for each task, incurring memory and computation overhead at each step — infeasible for training LLM's and LVLM's on commodity hardware. Some recent approaches aim to find task relationships for improving predictive performance (Wang et al., 2025; Li et al., 2024; Fifty et al., 2021). Recent work has explored MTL in LVLM for diverse application domains, including fashion understanding (Han et al., 2023), image restoration (Luo et al., 2024), digital pathology (Ramanathan et al., 2025), Alzheimer's disease prediction (Zhou et al., 2026a), diagnosing upper gastro-intestinal diseases using endoscopy images (He et al., 2025), evaluating non-tumor gastric diseases indicators in whole slide images (Fu et al., 2025) and abusive language detection in the setting where the language under consideration has only limited resources for training (Gaim et al., 2026). However, MTL for improving explainability robustness has not been addressed in the literature.

3 PEER-GUIDED AUXILIARY-TASK SELECTION

Consider a multi-task learning problem consisting of **primary task**, indexed by t, and m **auxiliary tasks**, indexed by $1, \ldots, m$. All tasks share a common backbone parameterized by $\theta \in \mathbb{R}^p$. For each task i (where $i = t$ or $i \in [m]$), let $A_i(x; \theta) \in \mathbb{R}^d$ denote the Attribution map (e.g., saliency maps, Grad$\times$Attn, or GradCAM) produced for input x.

Assume that $(X, \widetilde{X}) \sim P$, where P is the joint distribution of an input X and its perturbed version $\widetilde{X}$. The perturbation may correspond to compression, noise, corruption, or any application-specific transformation.

For each task i, quantify the error in explanation under perturbation through

$$V_i(\theta) = A_i(\widetilde{X};\theta) - A_i(X;\theta) - \mathbb{E}_P\Big[A_i(\widetilde{X};\theta) - A_i(X;\theta)\Big].$$

Each explanation error, V_i is assumed to belong to $L_2(P;\mathbb{R}^d) = \Big\{Z : \mathbb{E}_P\|Z(X,\widetilde{X})\|_2^2 < \infty\Big\}$. For each task $i \in \{t, 1, \ldots, m\}$, let $\widehat{R}_i : \Theta \to \mathbb{R}$ denote its empirical risk. Define the all-task empirical objective $\widehat{F}(\theta) = \widehat{R}_t(\theta) + \sum_{j=1}^m \widehat{R}_j(\theta)$.

Throughout the paper, unless otherwise stated all explanation Errors are evaluated at the pretrained multitask model

$$\theta_0 = \arg\min_\theta \widehat{F}(\theta) \left(= \widehat{R}_t(\theta) + \sum_{j=1}^m \widehat{R}_j(\theta), \right), \qquad \texttt{(Vanilla MTL)}$$

Hence, $\nabla\widehat{F}(\theta_0) = 0$. For simplicity, we suppress the dependence on θ_0 and write V_i.

Finally, the **Perturbation induced Explanation Error Relation (PEER)** score between the primary task t and auxiliary task j is

$$\mathbf{PEER}_{\,tj}(\theta) = \frac{\mathbb{E}_P\left[V_t(\theta)^\top V_j(\theta)\right]}{\sigma_t\sigma_j}, \tag{1}$$

where $\sigma_i^2 = \mathbb{E}_P\|V_i(\theta)\|_2^2, i \in \{t, [m]\}$. **PEER** is the correlation between the perturbation-induced explanation responses of the primary task and an auxiliary task.

The objective is to investigate if the explanation robustness of an MTL network for the primary task can be improved when trained with a subset of auxiliary tasks chosen through **PEER** scores. To formulate the problem we introduce the Explanation Instability functional (EIF)

$$\mathbf{VAR}_t(\theta) = \frac{1}{2}\mathbb{E}_P\|V_t(\theta)\|^2. \qquad \texttt{(EIF)}$$

For a selected subset $S \subseteq [m]$, define

$$\theta_S = \underset{\theta\in\Theta}{\operatorname{argmin}} F_S(\theta) \left(= \widehat{R}_t(\theta) + \sum_{l\in S}\widehat{R}_l(\theta) = F(\theta) - \sum_{j\notin S}\widehat{R}_j(\theta)\right), \tag{2}$$

Let θ_S be the fully retrained model with tasks in S. We introduce the problem of keeping the tasks that enhance explainability. In particular, we introduce the following problem.

Problem Statement:KEEP The problem of Keeping Explanation Enhancing Peers (**KEEP**) can be understood as finding $S \subset [m]$, a subset of tasks, such that it produces improved Explanation robustness as follows.

$$\mathbf{VAR}_t(\theta_S) \leq \mathbf{VAR}_t(\theta_0)$$

where θ_0 is parameters of the vanilla MTL model and θ_S is the parameters of the MTL model trained on S. Existing literature has studied task relatedness to enhance performance but not explainability Sener & Koltun (2018); Navon et al. (2022); Liu et al. (2021b); Kendall et al. (2018); Qin et al. (2025), but to the best of our knowledge the problem of **KEEP** has not been studied.

4 Identifying Subset of Tasks for enhancing explainability

In this section we discuss strategies to solve **KEEP** using **PEER** scores evaluated at θ_0. We first study explanation robustness in a simplified setting where there is one auxiliary task and one primary task. This setup helps in characterizing the effect of shared features on Explanation robustness when an auxiliary task is deleted. We refer this as **PEER**-deletion principle. This principle is then used in the general setup to arrive at a strategy for solving **KEEP**.

4.1 EXPLANATION TRANSFER: MEASURING THE INFLUENCE OF ONE AUXILIARY TASK

Two-task setup To understand the influence of an auxiliary task on explanation robustness, we first consider a two-task problem consisting of a primary task t and one auxiliary task a. Introduce

$$\theta_\eta = \arg\min_\theta \widehat{F}_\eta(\theta), \quad \widehat{F}_\eta(\theta) = \widehat{R}_t(\theta) + \eta \widehat{R}_a(\theta),\ \eta \in [0,1], \tag{3}$$

The parameter η continuously interpolates between the primary-only model, θ_0, and the jointly trained model, θ_1. The curve $\{\theta_\eta : \eta \in [0,1]\}$ will be called the *auxiliary-weight homotopy*. Along this path, define the primary and auxiliary explanation responses by $V_t(\eta) = V_t(\theta_\eta),\ V_a(\eta) = V_a(\theta_\eta)$. Denote primary explanation instability(see equation `EIF`) by $\mathbf{VAR}_{t\eta} = \mathbf{VAR}_t(\theta(\eta))$ The effect of deleting the auxiliary task is

$$\Delta_a = \mathbf{VAR}_{t0} - \mathbf{VAR}_{t1}. \tag{4}$$

Thus, $\Delta_a < 0$ means that deleting the auxiliary task improves the explanation robustness of the primary task.

We introduce the following mild assumption on the relationship between change in V_t along the homotopy path.

Assumption 1 (Positive Explanation Transfer along the auxiliary-weight homotopy path)**.** *For every $\eta \in [0,1]$, the rate of change of* $\mathbf{VAR}$ *is related to the explanation alignment $C_{ta}(\eta) = \mathbb{E}_P\big[V_t(\eta)^\top V_a(\eta)\big]$ as follows,*

$$\dot{\mathbf{VAR}}_t(\eta) = \lambda(\eta) C_{ta}(\eta)$$

where $\lambda(\eta) > 0$.

The positivity of $\lambda(\eta)$, and hence the name positive explanation transfer, reflects the premise of multitask learning: auxiliary tasks are introduced to promote useful shared representations. Thus, increasing the weight of an auxiliary task reinforces, rather than reverses, the explanatory features it contributes to the shared backbone, causing the primary explanation to evolve locally in the same explanatory direction as the auxiliary explanation; a measure of this is $C_{ta}(\eta)$. These intuitions are further confirmed in actual experiments.

Under this assumption we can now characterize the relation of **PEER** and change in **VAR** if the auxiliary task is removed. More specifically,

Proposition 1 (**PEER** deletion principle)**.** *In the* **Two-task Setup***, suppose Assumption 1 holds. If* $\operatorname{sign}(\mathbf{PEER}(\theta_\eta)) = \operatorname{sign}(\mathbf{PEER}(\theta_1))$ *, for all* $\eta \in [0,1]$*, then*

$$\mathbf{PEER}_{ta}(\theta_1)\,\Delta_a \le 0. \tag{5}$$

If $\mathbf{PEER}_{ta}(\eta)$ *is nonzero for all $\eta \in [0,1]$ then the inequality is strict.*

The **KEEP** problem is naturally a combinatorial bilevel optimization problem in which the upper level selects auxiliary tasks while the lower level retrains the multitask model. Rather than repeatedly solving the lower-level problem, we study the sensitivity of its solution with respect to the auxiliary-task weight using a continuous homotopy. The resulting optimizer path enables a first-order characterization of the effect of task deletion on explanation instability, leading to the **PEER**–deletion principle. One could use the homotopy principle for the general case and assuming Assumption 1 to hold, one could develop a more general theory. But the exposition gets more complicated. Instead, going forward we will make the assumption that **PEER**-deletion principle itself holds in the general setting.

4.2 THE CASE OF MANY AUXILIARY TASKS

In this subsection we return to the original problem and apply the **PEER**-deletion principle to the setup of m auxiliary tasks.

We begin by stating some Assumptions, and will be followed by a choice of subset which will provably solve **KEEP** . There are four key components: (a) the initial model,(b)the re-trained model , (c) the `EIF` functional, and (d) **PEER** scores. We introduce the following assumptions, one per component.

Assumption 2 (Smoothness)**.** *Each $\widehat{R}_i$ is twice continuously differentiable in a neighborhood of θ_0. At θ_0 because of optimality the following holds*

$$\nabla \widehat{F}(\theta_0) = 0, \quad H = \nabla^2 \widehat{F}(\theta_0) \succeq \mu I \tag{6}$$

for some $\mu > 0$. Furthermore for every $S \subset [m]$,

1. *$H_S = \nabla^2 \widehat{F}(\theta_S) \succeq \mu_S I$ for some $\mu_S \geq 0$*
2. *There exists $L_H > 0$ such that, for all θ, θ' in a neighborhood of θ_0,*

$$\left\| \nabla^2 \widehat{F}_S(\theta) - \nabla^2 \widehat{F}_S(\theta') \right\|_{\text{op}} \leq L_H \|\theta - \theta'\|.$$

This assumption can be used to measure the deviation of θ_S and θ_0 as follows.

Lemma 1 (Exact local parameter perturbation under task deletion)**.** *Suppose Assumption 2 holds. Let θ_0 be defined in equation* `Vanilla MTL` *and θ_S be defined in equation 2.*

$$\theta_S - \theta_0 = H_S^{-1} \sum_{j \notin S} \nabla \widehat{R}_j(\theta_0) + r_S \quad r_S = O\big(\|\theta_S - \theta_0\|^2\big)\,.$$

where $H_S = \nabla^2 \widehat{F}_S(\theta_0) = H - \sum_{j \notin S} \nabla^2 \widehat{R}_j(\theta_0)$,

The lemma provides an expression involving H_S^{-1} which needs to be evaluated at θ_S. For ease of analysis it would be good to get a characterization involving H(the full hessian evaluated at θ_0) instead. The next assumption uses smoothness of the Hessian to control the error in replacing H_S by H.

Assumption 3 (Locality of retraining)**.** *For every $S \in [m]$, the retrained solution θ_S lies in a neighbourhood of θ_0 and satisfies*

$$\begin{aligned} &\theta_S = \theta_0 + h_S + r_S, \quad h_S = \sum_{j \notin S} h_j, \\ &h_j = H^{-1} \nabla \widehat{R}_j(\theta_0), j \in [m] \quad \|r_S\| \leq C \|h_S\|^2, \end{aligned} \tag{7}$$

for some constant $C > 0$.

Assumption 4 (Differentiability of `EIF`)**.** *The* `EIF` *functional $\mathbf{VAR}_t : \Theta \to \mathbb{R}$ is continuously Fréchet differentiable at θ_0 and there exists $L > 0$ such that for all $h \in \mathbb{R}^p$.*

$$|\mathbf{VAR}_t(\theta_0 + h) - \mathbf{VAR}_t(\theta_0) - D\mathbf{VAR}_t(\theta_0)[h]| \leq L\|h\|^2$$

$$D\mathbf{VAR}_t(\theta)[h] = \lim_{\varepsilon \to 0} \frac{\mathbf{VAR}_t(\theta + \varepsilon h) - \mathbf{VAR}_t(\theta)}{\varepsilon}$$

Furthermore $D\mathbf{VAR}_t(\theta)[h] = \nabla \mathbf{VAR}_t(\theta)^\top h \leq M\|h\|$ where $M = \|\nabla \mathbf{VAR}_t(\theta_0)\|_{op}$.

The assumption states that the first variation, $D\mathbf{VAR}_t$, is Lipschitz continuous. For any subset $S \subset [m]$

$$\Delta_S = D\mathbf{VAR}_t(\theta_0)[h_S] = \sum_{j \notin S} D\mathbf{VAR}_t(\theta_0)[h_j]$$

holds because of linearity of the first variation. Introduce

$$\Delta_j = D\mathbf{VAR}_t(\theta_0)[h_j] = \nabla \mathbf{VAR}_t(\theta_0)^\top h_j. \tag{8}$$

where h_j and h_S is defined in Assumption 3. Observe that j here denotes a task which is removed and Δ_j is the change in the $\mathbf{VAR}_t$ value after retraining. In the previous section we found that under positive explanation transfer one can prove the **PEER**-deletion principle for one primary and one auxiliary task. In the general case we will assume that the principle holds. More precisely,

Assumption 5 (**PEER**-deletion principle)**.** *For every auxiliary task j, $\mathbf{PEER}_{tj}$ $\Delta_j \leq 0$, with equality if and only if $\mathbf{PEER}_{tj} = 0$.*

Intuitively, since **PEER** is evaluated at θ_0(equation `Vanilla MTL`), one can expect that if V_t and V_j have a positive **PEER** score, then task j contributes to the current explanation in essentially the same direction as the existing explanation perturbation. Removing it should reduce the magnitude of the perturbation, leading to a decrease in $\mathbf{VAR}$. Opposite is expected for negative $\mathbf{PEER}$ score.

PEER-MTL:KEEPING AUXILIARY TASKS WITH NEGATIVE **PEER**

To solve the problem of **KEEP**, we consider the subset $S_{\mathbf{KEEP}} = \{j \in [m] | \mathbf{PEER}_{tj} < 0\}$ of tasks. In the following we establish $S_{\mathbf{KEEP}}$ leads to better explanation robustness than provided by the original model, as measured through EIFfunctional.

Theorem 1 ($S_{\mathbf{KEEP}}$ Improves Explanation Robustness)**.** *Let θ_0 denote the multitask model trained using all auxiliary tasks, and let θ_S denote the model obtained after retraining only with the tasks in $S = \{j | \mathbf{PEER}_{tj} < 0\} \subset [m]$. Suppose Assumptions 2–5 hold and C, L, M are defined therein. Define $h_S = \sum_{j \notin S} h_j,\ \Delta_j = D\mathbf{VAR}[h_j]$.*

1. *$\Delta_j \leq 0$ for all $j \notin S$*
2. *If $-\sum_{j \notin S} \Delta_j > \left(MC + \frac{L_I}{2}(1 + C\|h_S\|)^2\right) \|h_S\|^2$, then $\mathbf{VAR}_t(\theta_S) < \mathbf{VAR}_t(\theta_0)$.*

Proof is in the Appendix.

Remark 1. *The final claim 2 in Theorem 1 can be stated more cleanly if we assume that each task is well trained locally. For example if there exists $\varepsilon > 0$ such that $\|\nabla R_j(\theta_0)\| \leq \varepsilon, \quad j \in [m]$., then one can show that $\mathbf{VAR}_t(\theta_S) < \mathbf{VAR}_t(\theta_0)$ as soon as $-\sum_{j \notin S} \Delta_j \geq \left(MC + \frac{L}{2}(1 + C(m - |S|)\frac{\epsilon}{\mu})^2\right)(m - |S|)^2 \left(\frac{\epsilon}{\mu}\right)^2$ holds. This is straightforward to see by noticing that $\|h_j\| = \|H^{-1}\nabla R_j(\theta_0)\| \leq \frac{\epsilon}{\mu}$*

Remark 2. *The* **PEER***-deletion assumption can be weakened and the results can be re-derived by employing homotopy based methods as sketched in the section involving only one auxiliary task.*

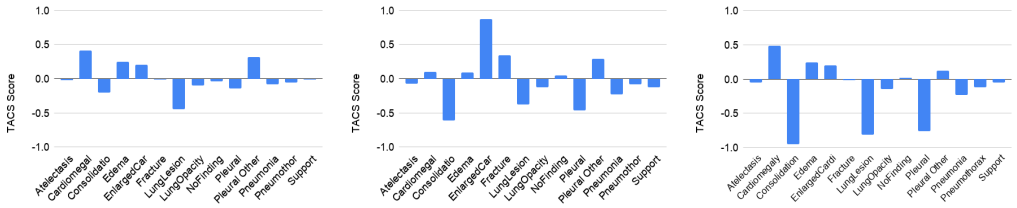
Remark 3. *The techniques would be of broader interest in the MTL setting. It could be potentially used to study other kinds of functionals.*

4.2.1 **PEER-MTL**: A MTL FRAMEWORK ROBUST TO EXPLANATION PERTURBATION

The MTL model obtained by solving equation 2 where $S = S_{\mathbf{KEEP}}$ will be referred in the sequel as **PEER-MTL**. It is expected that this model will produce explanations which are robust to image perturbations. We will use **PEER-MTL** to build a VLM model in the setting of CXR analysis where there is a primary task and there are a set of auxiliary tasks that can be augmented with primary task. **PEER-MTL** helps to choose the best set of auxiliary tasks that is robust to image perturbations. In the next section we will empirically evaluate the performance of this model.

Model	Method	Saliency		Grad-Attn		Gradcam		Legrad	
		Orig	WA	Orig	WA	Orig	WA	Orig	WA
Vit-B/14	TB	53.44	99.93	53.12	86.47	56.25	80.54	56.28	84.61
	Vanilla-MTL	48.53	56.47	44.22	55.86	49.71	61.25	46.51	59.13
	TAG-MTL	49.58	58.21	47.76	57.31	50.98	61.42	46.36	60.78
	PEER-MTL	45.63	47.47	35.86	37.62	44.04	46.21	43.62	45.83
Vit-B/16	TB	54.64	99.93	53.45	83.28	56.24	81.44	57.02	82.17
	Vanilla-MTL	49.53	57.44	46.62	57.34	49.40	59.22	46.34	59.87
	TAG-MTL	51.05	60.41	48.43	59.20	52.65	60.26	47.26	62.31
	PEER-MTL	45.60	48.22	44.32	45.18	44.64	46.07	43.64	45.91
Vit-B/32	TB	54.64	99.93	54.52	84.92	56.62	83.76	57.62	85.03
	Vanilla-MTL	48.60	58.44	47.62	58.34	50.23	59.45	49.82	58.25
	TAG-MTL	51.55	64.58	50.31	61.87	53.10	62.15	52.56	67.49
	PEER-MTL	46.60	48.19	44.64	46.06	44.69	45.52	43.71	45.58

Table 1: OLS Score for TB CXR images on original and WhatsApp-compressed (WA) test images, evaluated on ViT trained on TB dataset (TB), TB with all auxiliary tasks (Vanilla-MTL), TB with TAG-selected tasks (TAG-MTL) and TB with PEER-selected tasks (**PEER-MTL**). **PEER-MTL** (our approach) consistently achieves the lowest OLS scores.



(a) Vit-B/16 (b) Vit-B/32 (c) Vit-B/14

Figure 2: PEER score measured for all models using original and WhatsApp images for model trained used in CXR images in TB and MIMIC dataset and evaluated using original and WhatsApp test images in TB and MIMIC dataset

5 Experiments

The goal of our experiments is to build a Vision-Language Model (VLM) for Tuberculosis (TB) diagnosis from Chest X-ray (CXR) images, designed for integration with mHealth applications. TB remains a leading cause of death worldwide, with over 10 million new cases annually (Organization, 2023). AI-based diagnosis can address the severe shortage of radiologists in low- and middle-income countries (LMICs). Using our PEER-driven auxiliary task selection strategy, we construct a VLM termed LLaVA-**PEER**. Our experiments are designed to demonstrate: (i) the efficacy of explanation techniques—among different ViT explanation methods, which are most sensitive? (ii) Assumption 1 - is Assumption 1 valid? (iii) the selection of auxiliary tasks—which tasks are selected by **PEER**? (iv) the robustness of **PEER-MTL**—does **PEER-MTL** improve prediction and explanation robustness? and (v) the efficacy of report generation—does LLaVA-**PEER** generate high-quality reports?

5.1 Experiment Settings

Datasets and Evaluation. We utilize the Montgomery, Shenzhen, and TBX11K datasets (Jaeger et al., 2014; Liu et al., 2020), comprising a total of 5,400 TB and normal CXR images. These publicly available datasets provide binary labels. To enable VLM training, we curate radiology reports for these datasets, annotated by board-certified radiologists. To study robustness under real-world perturbations, we create a WhatsApp-CXR dataset by transmitting the test images over WhatsApp in an automated manner. The original test images occupy 8.0 GB, while the WhatsApp-compressed versions require only 1.1 GB. The original images, curated radiology reports, and WhatsApp-compressed test images collectively form the PEER-TB dataset.

For auxiliary tasks, we leverage MIMIC-CXR (Johnson et al., 2019), which contains 377,110 CXR training images annotated for 14 lung abnormalities alongside free-text radiology reports. Training is performed on the TB dataset (primary task) alongside PEER-selected auxiliary tasks from the MIMIC-CXR training set. Evaluation is conducted on both the original and WhatsApp-compressed test splits of the TB datasets. We assess prediction and explanation robustness using the Prediction Instability (PI) and Out-of-Lung Saliency (OLS) scores (DeGrave et al., 2021; Antony et al., 2025). Generated reports are evaluated for lexical similarity and factual accuracy using standard metrics: BLEU, ROUGE, and RadGraph-based F1 scores (Jain et al., 2021).

Models For prediction and explainability robustness evaluation, we consider three Vision Transformer variants: ViT-B/14, ViT-B/16, and ViT-B/32. To build our VLM, we adopt the LLaVA architecture (Liu et al., 2023a), comprising a ViT vision encoder (one of the three variants above), a fully connected projection layer, and a LLaMA2 Vicuna-7B-v1.5 (Touvron et al., 2023) language decoder. The resulting model, LLaVA-**PEER**, is fine-tuned with primary and PEER-selected auxiliary tasks.

5.2 Experiment Results

Efficacy of Explanation Techniques We evaluate the effectiveness of four Vision Transformer explanation techniques—Saliency, Grad$\times$Attn (Chefer et al., 2021a), GradCAM (Chefer et al., 2021b), and LeGrad (Bousselham et al., 2025)—using MoRF (Most Relevant First) and LeRF (Least Rel-

evant First) (Li et al., 2023b). MoRF measures the change in model activations when sequentially removing the most relevant features identified by each attribution method, while LeRF measures the change when removing the least relevant features. Evaluation is performed across increasing removal thresholds (0–9) for ViT-B/16, ViT-B/32, and ViT-L/14 (Appendix G.1). From Figure 4 and 5 in Appendix, it is evident that across all architectures, Saliency and Grad×Attn consistently achieve the highest or competitive scores, indicating superior identification of relevant features. In contrast, GradCAM consistently produces the lowest scores, suggesting inferior feature localization capability.

Validation of Assumption 1 We verify assumption 1 on two task setup. Using the objective in equation 3, we train a family of models along the auxiliary-weight homotopy at $\eta \in [0, 1]$. At each η compute $\mathbf{VAR}_t(\eta)$ and $C_{ta}(\eta)$ from the explanation responses on a fixed held-out set of original and compressed image pairs in PEER-TB dataset. Estimating $\dot{\mathbf{VAR}}_t(\eta)$ by equation `EIF` gives $\lambda(\eta) = \dot{\mathbf{VAR}}_t(\eta)/C_{ta}(\eta)$. We plot $\lambda(\eta)$ for different values of η for Vit-B/32, for both Consolidation, an auxiliary task with negative **PEER**, and Cardiomegaly, one with positive **PEER**. From the results in Figure 6 in Appendix, it can be seen that $\lambda(\eta) > 0, \forall \eta \in [0, 1]$, validating assumption 1.

Selection of Auxiliary Tasks In this paper, our aim is to select auxiliary tasks for the TB (primary task) diagnosis model that result in positive transfer. The auxiliary tasks are chosen from the MIMIC-CXR dataset. To accomplish this, we first train a vision transformer using the TB dataset together with the MIMIC-CXR training dataset. The **PEER** score (Equation 1) is then computed using original images and their corresponding WhatsApp-compressed (perturbed) versions from both the MIMIC-CXR and TB test datasets. The **PEER** scores are evaluated for ViT-B/14, ViT-B/16, and ViT-B/32. From the **PEER** values (Figure 2), several consistent patterns emerge across the three ViT variants. The following tasks exhibit relatively high negative **PEER** scores (favorable for inclusion) across architectures: Consolidation, Lung Lesion, Pleural Effusion, and Lung Opacity. In contrast, the following tasks show high positive **PEER** scores (excluded due to expected negative transfer): Cardiomegaly, Edema, Enlarged Cardiomediastinum, and Pleural Other.

PEER-MTL Improves Prediction and Explainability Robustness We trained three ViT variants under four configurations: TB-only, Vanilla-MTL (all auxiliary tasks), TAG-MTL (Fifty et al., 2021) and **PEER-MTL** (PEER-selected tasks). We evaluated prediction robustness via PI Score and explainability robustness via OLS Score. Prediction Robustness (Table 3 in Appendix). TB-only models exhibit high PI scores (12.46–14.82). Vanilla-MTL reduces these significantly (2.06–3.37). **PEER-MTL** achieves the lowest scores (1.25–2.28), consistently outperforming Vanilla-MTL. Regarding Explainability Robustness (Table 1), TB-only models show alarmingly high OLS on compressed images—Saliency approaches 100% across all variants. **PEER-MTL** substantially reduces OLS, by upto 30%, across all attribution methods. These results demonstrate that PEER-guided task selection effectively mitigates both prediction instability and out-of-lung saliency, producing models with more reliable predictions and explanations.

PEER-MTL Improves Report Generation. Table 4 (Appendix) compares report generation performance on original and WhatsApp-compressed CXR images in PEER-TB dataset. Llava-**PEER** achieves the highest scores across all metrics on both image types. Notably, on RadGraph—a clinically meaningful metric measuring entity extraction accuracy— Llava-**PEER** achieves 23.74 (original) and 22.34 (WA), outperforming LLaVA-CXR(Vanilla) by 7.15 and 10.09 points, respectively. This represents over 60% reduction in compression-induced degradation compared to Vanilla-MTL, confirming that **PEER**-guided task selection enhances both absolute performance and robustness to real-world image perturbations.

6 Conclusion

This paper initiates a study of explanation robustness in MTL framework. The goal of the study is to select auxiliary tasks to ensure the primary task provides explanations which are robust to perturbations. We characterize the effect of removing auxiliary task on explanation instability via the **PEER**-deletion principle. We believe the theoretical approach presented here will be of general interest to develop more understanding of MTL frameworks. We conduct an empirical study of **PEER-MTL** under mhealth settings and our results indicate that **PEER-MTL** improves both attribution robustness and downstream report generation.

AI USE STATEMENT

In this work, we have not used generative AI tools for any tasks with required disclosure. Additionally, we used generative AI tools for tasks with recommended disclosure such as polishing the writing to improve readability. We take responsibility for the final content of this work.

ETHICS STATEMENT

All chest X-rays used here come from publicly released, de-identified research datasets (Jaeger et al., 2014; Liu et al., 2020; Johnson et al., 2019), each used within the terms under which it was released. We release only radiologist-authored free-text reports for the public TB benchmarks, keyed by the original image filenames: no images are redistributed and no MIMIC-CXR material is included. Because the released reports add an annotation layer to data that is already public and de-identified, and no images are redistributed, the release introduces no new privacy risk and conforms to the terms of the source datasets.

REPRODUCIBILITY STATEMENT

All assumptions underlying our theoretical claims are stated explicitly in Section 4, and complete proofs of Proposition 1 and Theorem 1 are given in Appendix A, B. Training configurations, architectures, and optimisation hyperparameters for the vision encoder and for LLaVA-**PEER** are reported in Appendix E. The details of PEER-TB dataset is given in Table 2. The link to source code for computing **PEER**, the VLM pipeline and the curated dataset is included in the supplementary materials.

REFERENCES

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4190–4197, Online, July 2020a. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.385. URL https://aclanthology.org/2020.acl-main.385/.

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 4190–4197, 2020b.

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31, 2018.

David Alvarez-Melis and Tommi S Jaakkola. On the robustness of interpretability methods. *arXiv preprint arXiv:1806.08049*, 2018.

Christopher Anders, Plamen Pasliev, Ann-Kathrin Dombrowski, Klaus-Robert Müller, and Pan Kessel. Fairwashing explanations with off-manifold detergent. In *International Conference on Machine Learning*, pp. 314–323. PMLR, 2020.

Mariamma Antony, Rajiv Porana, Sahil M Lathiya, Siva Teja Kakileti, and Chiranjib Bhattacharyya. Chexwhatsapp: A dataset for exploring challenges in the diagnosis of chest x-rays through mobile devices. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 25887–25896, 2025.

Walid Bousselham, Angie Boggust, Sofian Chaybouti, Hendrik Strobelt, and Hilde Kuehne. Legrad: An explainability method for vision transformers via feature formation sensitivity. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20336–20345, 2025.

Brandon Carter, Jonas Mueller, Siddhartha Jain, and David Gifford. What made you do this? understanding black-box decisions with sufficient input subsets. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 567–576. PMLR, 2019.

Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 397–406, 2021a.

Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 782–791, 2021b.

Chao Chen, Chenghua Guo, Rufeng Chen, Guixiang Ma, Ming Zeng, Xiangwen Liao, Xi Zhang, and Sihong Xie. Training for stable explanation for free. *Advances in Neural Information Processing Systems*, 37:3421–3457, 2024a.

Zhihong Chen, Maya Varma, Justin Xu, Magdalini Paschali, Dave Van Veen, Andrew Johnston, Alaa Youssef, Louis Blankemeier, Christian Bluethgen, Stephan Altmayer, et al. A vision-language foundation model to enhance efficiency of chest x-ray interpretation. *arXiv e-prints*, pp. arXiv–2401, 2024b.

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.

Ann-Kathrin Dombrowski, Maximillian Alber, Christopher Anders, Marcel Ackermann, Klaus-Robert Müller, and Pan Kessel. Explanations can be manipulated and geometry is to blame. *Advances in neural information processing systems*, 32, 2019.

Theodoros Evgeniou and Massimiliano Pontil. Regularized multi–task learning. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 109–117, 2004.

Chris Fifty, Ehsan Amid, Zhe Zhao, Tianhe Yu, Rohan Anil, and Chelsea Finn. Efficiently identifying task groupings for multi-task learning. *Advances in Neural Information Processing Systems*, 34:27503–27516, 2021.

Mingxi Fu, Liming Liu, Fanglei Fu, Jingli Ouyang, Xueying Shi, Song Duan, Tian Guan, Yonghong He, Zhiqiang Cheng, and Lianghui Zhu. A multi-task learning model for evaluating non-tumor gastric diseases indicators in whole slide images. *Scientific Reports*, 2025.

Fitsum Gaim, Hoyun Song, Huije Lee, Changgeon Ko, Euijun Hwang, and Jong C Park. A multi-task benchmark for abusive language detection in low-resource settings. *Advances in Neural Information Processing Systems*, 38, 2026.

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

Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 3681–3688, 2019.

Xiao Han, Xiatian Zhu, Licheng Yu, Li Zhang, Yi-Zhe Song, and Tao Xiang. Fame-vil: Multi-tasking vision-language model for heterogeneous fashion tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2669–2680, 2023.

Yuxuan He, Qilei Chen, Benyuan Liu, and Yu Cao. Foundational multi-task multimodal model for upper gi endoscopy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6612–6621, 2025.

Stefan Jaeger, Sema Candemir, Sameer Antani, Yì-Xiáng J Wáng, Pu-Xuan Lu, and George Thoma. Two public chest x-ray datasets for computer-aided screening of pulmonary diseases. *Quantitative imaging in medicine and surgery*, 4(6):475, 2014.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven Truong, Du Nguyen Duong Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew Lungren, Andrew Ng, Curtis Langlotz, Pranav Rajpurkar, and Pranav Rajpurkar. Radgraph: Extracting clinical entities and relations from radiology reports. In J. Vanschoren and S. Yeung (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1. Curran, 2021. URL `https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/c8ffe9a587b126f152ed3d89a146b445-Paper-round1.pdf`.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019.

Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7482–7491, 2018.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36: 28541–28564, 2023a.

Dongyue Li, Aneesh Sharma, and Hongyang R Zhang. Scalable multitask learning using gradient-based estimation of task affinity. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1542–1553, 2024.

Xuhong Li, Mengnan Du, Jiamin Chen, Yekun Chai, Himabindu Lakkaraju, and Haoyi Xiong. M4: A unified xai benchmark for faithfulness evaluation of feature attribution methods across metrics, modalities, and models. 2023b.

Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. *Advances in Neural Information Processing Systems*, 34:18878–18890, 2021a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023a.

Liyang Liu, Yi Li, Zhanghui Kuang, J Xue, Yimin Chen, Wenming Yang, Qingmin Liao, and Wayne Zhang. Towards impartial multi-task learning. iclr, 2021b.

Yun Liu, Yu-Huan Wu, Yunfeng Ban, Huifang Wang, and Ming-Ming Cheng. Rethinking computer-aided tuberculosis diagnosis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2646–2655, 2020.

Yun Liu, Yu-Huan Wu, Shi-Chen Zhang, Li Liu, Min Wu, and Ming-Ming Cheng. Revisiting computer-aided tuberculosis diagnosis. *IEEE transactions on pattern analysis and machine intelligence*, 46(4):2316–2332, 2023b.

David G Luenberger. *Optimization by vector space methods*. John Wiley & Sons, 1997.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Ziwei Luo, Fredrik K Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas Schön. Controlling vision-language models for multi-task image restoration. In *International Conference on Learning Representations*, volume 2024, pp. 16226–16246, 2024.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, pp. 353–367. PMLR, 2023.

Aviv Navon, Aviv Shamsian, Idan Achituve, Haggai Maron, Kenji Kawaguchi, Gal Chechik, and Ethan Fetaya. Multi-task learning as a bargaining game. *arXiv preprint arXiv:2202.01017*, 2022.

World Health Organization. *Global tuberculosis report 2023*. World health organization, 2023.

Xiaohan Qin, Xiaoxing Wang, and Junchi Yan. Towards consistent multi-task learning: Unlocking the potential of task-specific parameters. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10067–10076, 2025.

Vishwesh Ramanathan, Tony Xu, Pushpak Pati, Faruk Ahmed, Maged Goubran, and Anne L Martel. Modaltune: Fine-tuning slide-level foundation models with multi-modal information for multi-task learning in digital pathology. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23912–23923, 2025.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. *Advances in neural information processing systems*, 31, 2018.

Trevor Standley, Amir Zamir, Dawn Chen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Which tasks should be learned together in multi-task learning? In *International conference on machine learning*, pp. 9120–9132. PMLR, 2020.

Omkar Thawkar, Abdelrahman Shaker, Sahal Shaji Mullappilly, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, Jorma Laaksonen, and Fahad Shahbaz Khan. Xraygpt: Chest radiographs summarization using medical vision-language models. *arXiv preprint arXiv:2306.07971*, 2023.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Fan Wang and Adams Wai Kin Kong. Exploiting the relationship between kendall's rank correlation and cosine similarity for attribution protection. *Advances in Neural Information Processing Systems*, 35:20580–20591, 2022.

Zedong Wang, Siyuan Li, and Dan Xu. Rep-mtl: Unleashing the power of representation-level task saliency for multi-task learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3413–3423, 2025.

Zifan Wang, Haofan Wang, Shakul Ramkumar, Piotr Mardziel, Matt Fredrikson, and Anupam Datta. Smoothed geometry for robust attribution. *Advances in neural information processing systems*, 33:13623–13634, 2020.

Zirui Wang, Yulia Tsvetkov, Orhan Firat, and Yuan Cao. Gradient vaccine: Investigating and improving multi-task optimization in massively multilingual models. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=F1vEjWK-lH_.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Hui Hui, Yanfeng Wang, and Weidi Xie. Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. *Nature Communications*, 16(1):7866, Aug 2025. ISSN 2041-1723. doi: 10.1038/s41467-025-62385-7. URL https://doi.org/10.1038/s41467-025-62385-7.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *Advances in Neural Information Processing Systems*, 33:5824–5836, 2020.

Juan Manuel Zambrano Chaves, Shih-Cheng Huang, Yanbo Xu, Hanwen Xu, Naoto Usuyama, Sheng Zhang, Fei Wang, Yujia Xie, Mahmoud Khademi, Ziyi Yang, et al. A clinically accessible small multimodal radiology model and evaluation metric for chest x-ray findings. *Nature Communications*, 16(1):3108, 2025.

Kai Zhang, Jun Yu, Eashan Adhikarla, Rong Zhou, Zhiling Yan, Yixin Liu, Zhengliang Liu, Lifang He, Brian Davison, Xiang Li, et al. Biomedgpt: A unified and generalist biomedical generative pre-trained transformer for vision, language, and multimodal tasks. *arXiv e-prints*, pp. arXiv–2305, 2023.

Wenchao Zhou, Zhao Xue, Jiaqi Liang, Jing Cui, Xiangjie Guo, and Yalu Wen. Modeling and application of alzheimer's disease complex trait prediction based on multi-task learning. *Scientific Reports*, 2026a.

Zhipeng Zhou, Ziqiao Meng, Pengcheng Wu, Peilin Zhao, and Chunyan Miao. Continual optimization with symmetry teleportation for multi-task learning. *Advances in Neural Information Processing Systems*, 38:171209–171230, 2026b.

Xun Zhu, Ying Hu, Fanbin Mo, Miao Li, and Ji Wu. Uni-med: a unified medical generalist foundation model for multi-task learning via connector-moe. *Advances in Neural Information Processing Systems*, 37:81225–81256, 2024.

A Appendix:Mathematical Details for the Auxiliary-Weight Homotopy

A.1 Assumptions

The assumption 1 is motivated by the shared representation learned by the multi-task model. Since the primary and auxiliary tasks are coupled through the same network parameters, an infinitesimal reduction in the auxiliary-task weight perturbs the shared representation before affecting the task-specific explanations.

A.2 Differentiability of $\mathbf{VAR}_t(\eta)$

For sake of completeness we sketch the proof of differentiability of $\mathbf{VAR}$ with η. The first-order optimality condition for θ_η is

$$\nabla \widehat{R}_t(\theta_\eta) + \eta \nabla \widehat{R}_a(\theta_\eta) = 0. \tag{9}$$

Differentiating with respect to η yields

$$\left(\nabla^2 \widehat{R}_t(\theta_\eta) + \eta \nabla^2 \widehat{R}_a(\theta_\eta)\right) \dot{\theta}_\eta + \nabla \widehat{R}_a(\theta_\eta) = 0. \tag{10}$$

Therefore, whenever H_η is nonsingular,

$$\dot{\theta}_\eta = -H_\eta^{-1} \nabla \widehat{R}_a(\theta_\eta). \tag{11}$$

Recall $V_t(\eta) = \mathcal{V}_t(\theta_\eta)$, where

$$\mathcal{V}_t(\theta) = A_t(\widetilde{X}; \theta) - A_t(X; \theta) - \mathbb{E}_P\Big[A_t(\widetilde{X}; \theta) - A_t(X; \theta)\Big]. \tag{12}$$

If $\mathcal{V}_t$ is Fréchet differentiable, the chain rule gives (see (Luenberger, 1997, Chapter 7, Sections 7.2–7.3))

$$\dot{V}_t(\eta) = D\mathcal{V}_t(\theta_\eta)[\dot{\theta}_\eta]. \tag{13}$$

Substituting equation 11

$$\dot{V}_t(\eta) = D\mathcal{V}_t(\theta_\eta)\left[H_\eta^{-1} \nabla \widehat{R}_a(\theta_\eta)\right]. \tag{14}$$

Recall

$$\mathbf{VAR}_t(\eta) = \frac{1}{2}\mathbb{E}_P\left[\|V_t(\eta)\|_2^2\right], \tag{15}$$

differentiation under the expectation yields

$$\dot{\mathbf{VAR}}_t(\eta) = \mathbb{E}_P\Big[V_t(\eta)^\top \dot{V}_t(\eta)\Big]. \tag{16}$$

Plugging in equation 14

$$\dot{\mathbf{VAR}}_t(\eta) = -\mathbb{E}_P\Big[V_t(\eta)^\top D\mathcal{V}_t(\theta_\eta)\left[H_\eta^{-1} \nabla \widehat{R}_a(\theta_\eta)\right]\Big]. \tag{17}$$

This is quite illuminating when we study the Assumption 1. It posits that the first variation of the Explaination response of the Target task is affected by the value of the Auxiliary task.

A.3 Proof of Proposition 1

Proposition 2 (**PEER** deletion principle)**.** *Suppose Assumption 1 holds and the sign of the explanation alignment is preserved along the auxiliary-weight homotopy. Then*

$$\mathbf{PEER}_{ta}(\theta_1)\,\Delta_a \le 0. \tag{18}$$

If the endpoint PEER score is nonzero and the pathwise alignment is nonzero on a set of positive measure, then the inequality is strict.

By the fundamental theorem of calculus,

$$\Delta_a = \mathbf{VAR}_t(0) - \mathbf{VAR}_t(1) = -\int_0^1 \mathbf{V\dot{A}R}_t(\eta)\,d\eta. \tag{19}$$

Under Assumption 1,

$$\Delta_a = -\int_0^1 \lambda(\eta)\mathbb{E}_P\big[V_t(\eta)^\top V_a(\eta)\big]\;d\eta. \tag{20}$$

Because $\lambda(\eta) > 0$ and the alignment preserves its sign along the path, the integral has the same sign as

$$\mathbb{E}_P[V_t(1)^\top V_a(1)]. \tag{21}$$

The normalization in the PEER score is positive whenever the two explanation responses are non-degenerate. Hence

$$\operatorname{sign}(\Delta_a) = -\operatorname{sign}(\mathrm{PEER}_{ta}(\theta_1))\,, \tag{22}$$

which proves

$$\mathrm{PEER}_{ta}(\theta_1)\Delta_a \le 0. \tag{23}$$

The inequality is strict whenever the endpoint PEER score is nonzero and the pathwise alignment is nonzero on a set of positive measure.

B Appendix:Identifying Subset of Tasks for enhancing explainability

In this section we discuss strategies to solve **KEEP** using **PEER** scores evaluated at θ_0. We begin by stating some Assumptions, which will be followed by a choice of subset which will be justified.

B.1 Proofs of Lemma 1 and Theorem 1 in main paper

We present the proof of the Lemma stated after Assumption 2. We begin by restating the Lemma.

Lemma 2 (Exact local parameter perturbation under task deletion)**.** *Suppose Assumption 2 holds. Let θ_0 be defined in equation* `Vanilla MTL` *and θ_S be defined in equation 2.*

$$\theta_S - \theta_0 = H_S^{-1}\sum_{j\notin S}\nabla\widehat{R}_j(\theta_0) + O\big(\|\theta_S-\theta_0\|^2\big)\,.$$

where $H_S = \nabla^2\widehat{F}_S(\theta_0) = H - \sum_{j\notin S}\nabla^2\widehat{R}_j(\theta_0)$,

Proof. Since θ_S is a stationary point, $\nabla\widehat{F}_S(\theta_S) = 0$.

Let $\delta_S = \theta_S - \theta_0$. By Assumption 2, Taylor expansion of $\nabla\widehat{F}_S$ around θ_0 gives

$$0 = \nabla\widehat{F}_S(\theta_0) + H_S\delta_S + q_S,$$

where $q_S := \int_0^1 \Big[\nabla^2\widehat{F}_S(\theta_0+\tau\delta_S) - H_S\Big]\,\delta_S\,d\tau$. This leads to the following

$$\begin{aligned}\|q_S\| &\le \int_0^1 \Big\|\nabla^2\widehat{F}_S(\theta_0+\tau\delta_S) - \nabla^2\widehat{F}_S(\theta_0)\Big\|\,\|\delta_S\|\,d\tau\\ &\le \int_0^1 L_H\tau\|\delta_S\|^2\,d\tau\\ &= \frac{L_H}{2}\|\delta_S\|^2.\end{aligned}$$

Since $\nabla \widehat{F}(\theta_0) = 0$, we have $\nabla \widehat{F}_S(\theta_0) = -\sum_{j \notin S} \nabla \widehat{R}_j(\theta_0)$.

Hence $H_S \delta_S = \sum_{j \notin S} \nabla \widehat{R}_j(\theta_0) - q_S$.

Multiplying both sides by H_S^{-1} yields

$$\delta_S = H_S^{-1} \sum_{j \notin D} \nabla \widehat{R}_j(\theta_0) - H_S^{-1} q_S.$$

Defining

$$r_S = -H_S^{-1} q_S,$$

we obtain

$$\theta_S - \theta_0 = H_S^{-1} \sum_{j \notin S} \nabla \widehat{R}_j(\theta_0) + r_S.$$

Noting that $\|H_S^{-1}\|_{\mathrm{op}} = \mu_S$

$$\begin{aligned} \|r_S\| &= \|H_S^{-1} q_S\| \\ &\leq \|H_S^{-1}\|_{\mathrm{op}} \, \|q_S\| \\ &\leq \frac{L_H}{2\mu_S} \|\delta_S\|^2. \end{aligned}$$

Hence we conclude that

$$\|r_S\| = O(\|\theta_S - \theta_0\|^2).$$

□

Next we restate Theorem 1 and then present the proof.

Theorem 2 ($S_{\mathbf{KEEP}}$ Selection Improves Explanation Robustness)**.** *Let θ_0 denote the multitask model trained using all auxiliary tasks, and let θ_S denote the model obtained after re-training with only the tasks in $S = \{j | \mathbf{PEER}_{tj} < 0\} \subset [m]$. Suppose Assumptions 2–5 hold and h_j, C, L, M are defined therein. Define*

$$h_S = \sum_{j \notin S} h_j, \qquad \Delta_j = D\mathbf{VAR}_t[h_j].$$

1. $\Delta_j \leq 0$ *for all* $j \notin S$
2. $|\mathbf{VAR}_t(\theta_S) - \mathbf{VAR}_t(\theta_0) - \sum_{j \notin S} \Delta_j| \leq \left(MC + \frac{L}{2}(1 + C\|h_S\|)^2\right) \|h_S\|^2$
3. *If* $-\sum_{j \in D} \Delta_j > \left(MC + \frac{L}{2}(1 + C\|h_S\|)^2\right) \|h_S\|^2$, *then* $\mathbf{VAR}_t(\theta_S) < \mathbf{VAR}_t(\theta_0)$.

Proof. For all $j \notin S$ one has $\mathbf{PEER}_{tj} \geq 0$ for all $j \notin S$. By Assumption 5 $\Delta_j < 0$ whenever $\mathbf{PEER}_{tj} < 0$ and $\Delta_j = 0$ if $\mathbf{PEER}_{tj} = 0$. This proves the first claim.

To prove the second claim, we recall that by Assumption 3

$$\theta_S - \theta_0 = h_S + r_S, \|r_S\| \leq C\|h_S\|^2. \tag{24}$$

Next by Assumption 4

$$\mathbf{VAR}_t(\theta_S) - \mathbf{VAR}_t(\theta_0) = D\mathbf{VAR}_t(\theta_0)[h_S + r_S] + e_S,$$

where $|e_S| \leq \frac{L}{2}\|h_S + r_S\|^2$. Using linearity of the Fréchet derivative,

$$D\mathbf{VAR}_t(\theta_0)[h_S + r_S] = D\mathbf{VAR}_t(\theta_0)[h_S] + D\mathbf{VAR}_t(\theta_0)[r_S] = \sum_{j \notin S} \Delta_j + D\mathbf{VAR}_t(\theta_0)[r_S]$$

The last equality is true because of equation 8 and $h_S = \sum_{j \notin S} h_j$.

Hence,

$$\mathbf{VAR}_t(\theta_D) - \mathbf{VAR}_t(\theta_0) = \sum_{j \in D} \Delta_j + \varepsilon_S, \tag{25}$$

where

$$\varepsilon_S = D\mathbf{VAR}_t(\theta_0)[r_S] + e_S.$$

By Assumption 4

$$|D\mathbf{VAR}_t(\theta_0)[r_S]| \leq M\|r_S\| \leq MC\|h_S\|^2. \tag{26}$$

The last inequality holds due to equation 24. Moreover,

$$\|h_S + r_S\| \leq \|h_S\| + \|r_S\| \leq \|h_S\| + C\|h_S\|^2.$$

Therefore,

$$|e_S| \leq \frac{L}{2}(1 + C\|h_S\|)^2\|h_S\|^2.$$

Combining it with the equation 26 gives

$$|\varepsilon_S| \leq \left(MC + \frac{L}{2}(1 + C\|h_S\|)^2\right)\|h_S\|^2.$$

The second claim claim follows by combining this with equation 25.

By hypothesis of the third claim. $-\sum_{j\notin S}\Delta_j > \left(MC + \frac{L}{2}(1 + C\|h_S\|)^2\right)\|h_S\|^2$ when plugged into the second claim implies

$$\mathbf{VAR}_t(\theta_S) - \mathbf{VAR}_t(\theta_0) < 0.$$

which proves the theorem. □

C Challenges of CXR Diagnosis

Definitions of PIP, OLS and corresponding evaluation metrics are as follows (DeGrave et al., 2021; Antony et al., 2025):

C.1 Prediction Instability Problem

Prediction Instability of a predictive model is defined as the probability of disagreement between the predictions on a randomly perturbed instance and the true instance.

Consider a dataset $\mathcal{D} = \{X_i, \mathtt{R}(X_i)\}_{i=1}^N$ consisting of N pairs of original and WhatsApp-compressed CXR images denoted by X_i, and $\mathtt{R}(X_i)$ respectively.

Definition 1. ***PI Score*** *The estimate of PI Score due to image perturbation by WhatsApp can now be defined as the fraction of pairs where the predictions differed expressed as percentage.*

$$PI\ Score(\mathtt{f}; \mathcal{D}) = \frac{1}{N}\left(\sum_{\{X_i, \mathtt{R}(X_i)\}\in\mathcal{D}} PI(X_i, \mathtt{R}(X_i), \mathtt{f})\right) \times 100 \tag{27}$$

and $PI(X_i, \mathtt{R}(X_i), \mathtt{f}) = \mathbf{I}(y(X_i; \mathtt{f})) \neq \mathbf{I}(y(\mathbf{R}(X_i); \mathtt{f}))$

C.2 Out-of-Lung Saliency (OLS)

OLS refers to models identifying regions outside the lungs as significant contributors to predictions of lung abnormalities.

Definition 2. ***OLS Score*** *of a model* $\mathtt{f}$ *on a dataset* D *is defined as the percentage of images in* D *with an Intersection Over Lung-Region (IOL) value less than a threshold* η*.*

$$OLS\ Score_\eta(D; \mathtt{f}) = \frac{1}{N}\left(\sum_{x\in D} IOL(x; \mathtt{f}) < \eta\right) \times 100$$

$$IOL(x; \mathtt{M}) = \frac{\mathtt{Pix}\left(heatmap\text{-}region(y(x; \mathtt{f})) \cap lung(x)\right)}{\mathtt{Pix}\left(heatmap\text{-}region(y(x; \mathtt{f}))\right)}$$

where $\mathtt{Pix}(R(x))$ *counts the number of Pixels in region* $R(x)$*, a subset of the pixels in the original image* x*.*

A high *OLS* suggests that the majority of the dataset is affected by OLS, while a low score implies lesser impact.

Dataset	TB	Normal
Montgomery Jaeger et al. (2014)	58	80
Shenzhen Jaeger et al. (2014)	336	326
TBX11K Liu et al. (2023b)	800	3800

Table 2: Distribution of TB and Normal CXR images in Montgomery, Shenzhen, and TBX11K datasets

D Additional Related work

Fragility of explanations A separate line of work establishes that attribution maps are unstable. (Ghorbani et al., 2019; Adebayo et al., 2018) construct perturbations that are imperceptible and leave the predictions unchanged, yet produce entirely different attribution map. (Dombrowski et al., 2019; Anders et al., 2020) shows that this susceptibility is governed by the curvature induced by the activation function and that explanations can be steered to an arbitrary target map, while keeping same prediction via targeted training. (Alvarez-Melis & Jaakkola, 2018) formalizes the phenomenon as a failure of local lipschitz continuity. To improve attribution robustness, (Chen et al., 2024a; Wang et al., 2020) propose to smoothen the model's decision surface based by applying a regularizer. (Wang & Kong, 2022) shows that the non-differentiable Kendall rank correlation is positively related to cosine similarity and regularizes the cosine similarity between natural and perturbed attributions.

E Additional Experimental Details

For TB-ViT, we follow a compact supervised training protocol for chest X-ray classification. The model is trained with AdamW using a learning rate of 1×10^{-4}, weight decay of 0.05, and a cosine annealing schedule over 20-30 epochs, optimized with binary cross-entropy loss. Training is performed in a distributed data-parallel setting, and model selection is based on the best validation accuracy.

All llava-mtl models are trained on NVIDIA A100 GPUs using DeepSpeed ZeRO Stage 2 with bfloat16 mixed precision. The vision encoder is a ViT-B/32/16/14 (input resolution 512×512 or 518×518) initialized from a chest X-ray domain-adapted checkpoint, connected to Vicuna-7B-v1.5 via a two-layer GELU MLP projector. Pre-training optimizes only the projector for 1 epoch (lr $= 1 \times 10^{-3}$, global batch size 256), followed by LoRA fine-tuning ($r = 64$, $\alpha = 128$) of the full model for 3 epochs (lr $= 1 \times 10^{-4}$), both using a cosine scheduler with 3% linear warmup and AdamW. Reports are generated via greedy decoding (max 256 tokens) and evaluated using BLEU-1/4, ROUGE-1/L, and RadGraph.

F Dataset Details

The distribution of TB/Normal images in PEER-TB dataset is given in Table 2. A sample image-report pair in PEER-TB dataset is given in Figure 3

G Additional Experimental Results

G.1 Analyzing efficacy of explanation techniques.

MoRF (Li et al., 2023b) measures the change in model activations when removing the most relevant features identified by each attribution method. Higher scores indicate better faithfulness, as removing truly important features should significantly alter model representations. Results are shown for Saliency, Grad$\times$Attn (Chefer et al., 2021a), GradCAM (Chefer et al., 2021b) and LeGrad (Bousselham et al., 2025) across increasing removal thresholds (0–9) for (a) ViT-B/16, (b) ViT-B/32, and (c) ViT-L/14. Saliency and Grad$\times$Attn achieves the highest or competitive scores across all architectures, particularly at mid-range thresholds, while GradCAM consistently produces the lowest scores, indicating inferior identification of relevant features. All methods converge at threshold 9 (maximum removal).

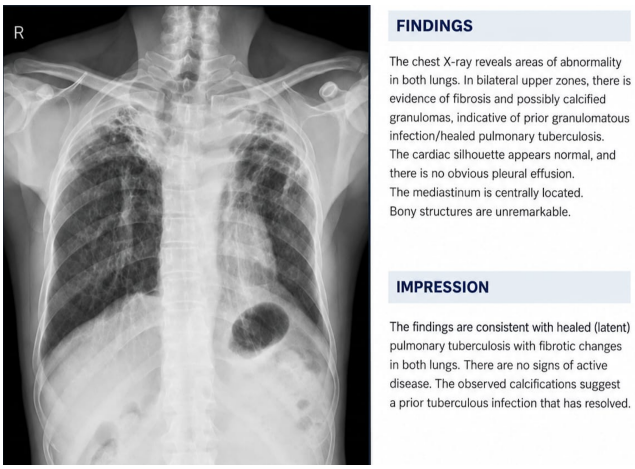
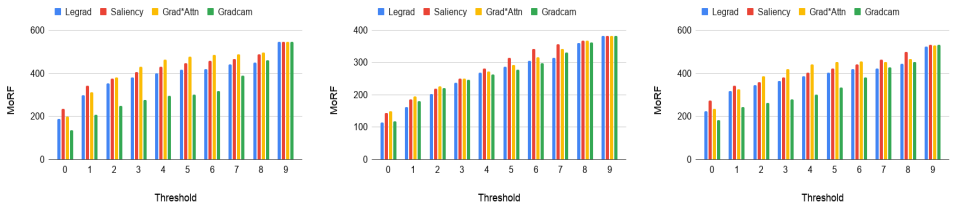


Figure 3: A sample image-radiology report pair in PEER-TB dataset



(a) Vit-B/16 (b) Vit-B/32 (c) Vit-B/14

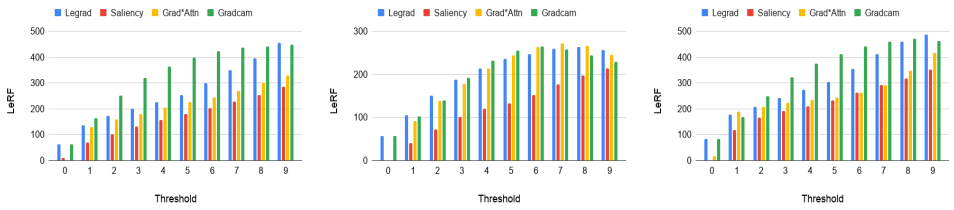
Figure 4: MORF scores (higher = better) measure activation change when removing most relevant features across increasing thresholds (0–9). Gradient×Attention and saliency achieves the best overall performance across ViT-B/16, ViT-B/32, and ViT-L/14, while GradCAM consistently underperforms. All methods converge at threshold 9.

LeRF (Li et al., 2023b) measures the change in model activations when removing the least relevant features identified by each attribution method. Lower scores indicate better faithfulness, as removing unimportant features should minimally alter model representations. Results are shown for Saliency, Grad×Attn (Chefer et al., 2021a), GradCAM (Chefer et al., 2021b) and LeGrad (Bousselham et al., 2025) across increasing removal thresholds (0–9) for (a) ViT-B/16, (b) ViT-B/32, and (c) ViT-L/14. Saliency and Grad×Attn achieve the lowest baseline scores (0.04–10.0), demonstrating superior precision in identifying truly irrelevant features. In contrast, GradCAM and Legrad start significantly higher (56.8–82.8), suggesting these methods inadvertently remove relevant features even when targeting least relevant ones.

G.2 Selection of Auxiliary Tasks

To accomplish this, we first train a multi-task vision transformer (encoder in Llava VLM) using the TB dataset together with the MIMIC-CXR training dataset. The PEER score is then computed using original images and their corresponding WhatsApp-compressed (perturbed) versions from both the MIMIC-CXR and TB test datasets. The TACS scores evaluated for ViT-B/14, ViT-B/16, and ViT-B/32 are presented in Figure 2.

From Figure 2, several consistent patterns emerge across the three ViT variants. The following tasks exhibit relatively high negative TACS scores (favorable for inclusion) across architectures: Consolidation, Lung Lesion, Pleural Effusion, and Lung Opacity. In contrast, the following tasks show high positive TACS scores (excluded due to expected negative transfer): Cardiomegaly, Edema, Enlarged Cardiomediastinum, and Pleural Other.



(a) Vit-B/16 (b) Vit-B/32 (c) Vit-B/14

Figure 5: LeRF scores (lower = better) measure activation change when removing least relevant features across thresholds (0–9). Saliency and Gradient×Attention achieve the lowest baseline scores (0.04–10.0), indicating superior precision in identifying irrelevant features. GradCAM and Legrad perform poorly, with baseline scores exceeding 56.8.

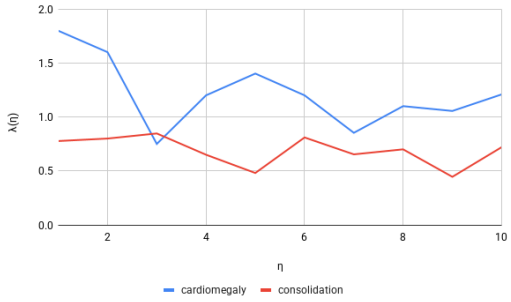


Figure 6: Plot of $\lambda(\eta)$ versus η along the auxiliary weight homotopy for TB as primary task and Consolidation and Cardiomegaly as auxiliary task for Vit-B/32 model

Model	TB	Vanilla-MTL	TAG-MTL	**PEER-MTL**
Vit-B/14	13.57	2.82	2.80	2.28
Vit-B/16	12.46	2.06	1.89	1.25
Vit-B/32	14.82	3.37	2.51	1.36

Table 3: PI Score of TB CXR images for the original and WhatsApp test images, evaluated on ViT trained on TB dataset, TB and all auxiliary tasks (Vanilla-MTL), TB and TAG-guided auxiliary tasks (TAG-MTL), TB and PEER-guided auxiliary tasks (**PEER-MTL**). The results show **PEER-MTL** is able to mitigate prediction instability problem compared to other approaches

Model	ROUGE-1		ROUGE-L		BLEU-1		BLEU-4		RadGraph	
	Orig	WA	Orig	WA	Orig	WA	Orig	WA	Orig	WA
Biomed-GPT	08.85	06.12	06.97	04.86	05.46	03.21	00.67	00.37	03.89	01.88
Rad-FM	13.64	10.08	10.01	07.20	09.43	06.89	01.22	00.76	05.14	02.73
LLaVA-Med	27.26	22.13	17.44	13.50	21.91	15.77	02.37	01.27	07.32	04.00
XrayGPT	16.63	13.20	13.88	10.83	13.10	10.27	04.09	03.59	08.02	05.68
LLaVA-Rad	28.15	24.02	18.88	15.64	18.45	14.33	02.38	01.28	13.84	09.61
CheXagent	25.08	20.96	16.61	12.81	17.17	12.28	02.92	01.80	11.80	07.88
LLaVA-VanillaMTL	29.83	26.75	21.45	17.33	22.67	18.40	09.35	06.22	16.59	12.25
LLaVA-TAGMTL	25.75	20.36	19.78	14.13	20.41	17.78	08.59	05.92	14.79	11.48
LLaVA-PEER	**33.01**	**31.79**	**22.33**	**21.18**	**25.43**	**23.39**	**10.37**	**09.21**	**23.74**	**22.34**

Table 4: Comparison of radiology report generation on original and WhatsApp-compressed (WA) CXR images. All baselines show substantial performance degradation under compression, particularly on RadGraph, indicating that compression harms clinical entity extraction. In contrast, LLaVA-**PEER** exhibits minimal degradation—over 60% less than Full-MTL and CheXagent—while achieving the highest absolute scores.

Model	ROUGE-1		ROUGE-L		BLEU-1		BLEU-4		RadGraph	
	Orig	WA	Orig	WA	Orig	WA	Orig	WA	Orig	WA
LLaVA-PEER(Vit-B/14)	33.01	31.79	22.33	21.18	25.43	23.39	10.37	09.21	23.74	22.34
LLaVA-PEER(Vit-B/16)	32.21	30.21	21.53	20.17	24.87	22.33	09.35	09.12	23.52	21.38
LLaVA-PEER(Vit-B/32)	32.07	30.14	20.25	20.18	25.21	23.03	09.54	09.11	19.39	17.38

Table 5: LLaVA-**PEER** results with different Vit model as vision encoder.