ASSET: Acquisition-Sensitive Subspace Estimation for Test-Time Adaptation of Medical VLMs

Mohan Zhang[1] **Zhen Tan**[2] **Songyuan Sui**[3] **Tianlong Chen**[1]
[1]University of North Carolina at Chapel Hill
[2]Stevens Institute of Technology [3]Rice University

Abstract

Medical vision-language models are often adapted under high-quality acquisition conditions. At target deployment, the same model may face lower-quality scanners, protocols, reconstructions, or artifacts. These shifts change the visual evidence, while the intended clinical question remains unchanged. As a result, a source-adapted model can fail under this acquisition shift. Methods are proposed to solve this problem. Medical acquisition-shift correction mainly targets image-only predictors, leaving the multimodal model underexplored. Robust training depends on anticipated degradations, and generic test-time tuning can easily overfit the small target training set, leaving unknown acquisition shifts insufficiently corrected. In this paper, we hypothesize that acquisition-like augmentations of target examples can expose parameter directions responsible for acquisition-shift failures. Based on it, we propose ASSET, an *Acquisition-Sensitive Subspace Estimation* method for test-time adaptation. ASSET estimates an acquisition-sensitive parameter subspace from prediction and gradient changes under these augmentations. It then adapts the model within this subspace through iterative estimate-and-update rounds. Comprehensive experiments show the advantages of ASSET. For example, ASSET consistently improves target accuracy under acquisition-quality shift across datasets and backbones by 2.62% on average, while preserving performance when no shift is present. Under the strongest acquisition-quality shift, ASSET even improves target accuracy by 4.12%, showing that the correction generalizes to severe shifts.

1 Introduction

Multimodal medical models are commonly adapted or trained on curated medical image-text data (Qin et al., 2023; Li et al., 2023; Zhang et al., 2025). The same model may later be deployed at a site with different scanner hardware, acquisition protocols, dose levels, reconstruction pipelines, or artifact profiles, producing lower-quality images. We refer to the development condition as the high-quality source setting and the deployment condition as the lower-quality target setting. Such acquisition differences can alter image appearance without altering the clinical question the model is expected to answer (Guo et al., 2024; Hognon et al., 2024). The source-target acquisition mismatch matters for multimodal medical models because their predictions remain grounded in image evidence, even when the text input specifies the intended clinical query. Figure 1a illustrates this deployment problem.

Prior work in medical imaging shows that scanner and protocol mismatch can reduce model performance across radiological modalities, and that acquisition-shift reduction is a practical problem in downstream medical image analysis (Guo et al., 2024; Hognon et al., 2024). Recent medical vision-language studies also show that artifacts, noise, and corruptions can degrade multimodal performance, while evaluations on curated or artifact-free images may overstate reliability under deployment conditions (Xia et al., 2024; Cheng et al., 2025; Imam et al., 2025). Together, these findings make acquisition quality a concrete deployment variable rather than a purely synthetic perturbation.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

We therefore study directional acquisition-quality shift, where a source-adapted multimodal model needs test-time adaptation for a lower-quality target setting after deployment.

Several research lines address parts of acquisition-shift adaptation, but their interventions occur at different levels of the deployment pipeline. Medical image harmonization and acquisition-shift correction reduce scanner or protocol mismatch through image-side transformations or image-only predictors, leaving correction of deployed multimodal models underexplored (Mali et al., 2021; Guan and Liu, 2022; Hognon et al., 2024). Robust training and fine-tuning can improve robustness under anticipated shifts, but they intervene before deployment and require access to the training pipeline, anticipated shifts, or specialized objectives (Fang et al., 2022; Shu et al., 2023; Han et al., 2024; Schlarmann et al., 2024). Test-time adaptation methods update deployed models from target data, often by optimizing confidence, augmentation consistency, prompts, caches, or pseudo-labels (Wang et al., 2021; Zhang et al., 2022a; Shu et al., 2022; Karmanov et al., 2024; Kaya et al., 2026). However, these objectives do not estimate parameter directions tied to acquisition-quality change, and can easily overfit the few target examples without isolating acquisition-related correction. This leaves a central research question (RQ):

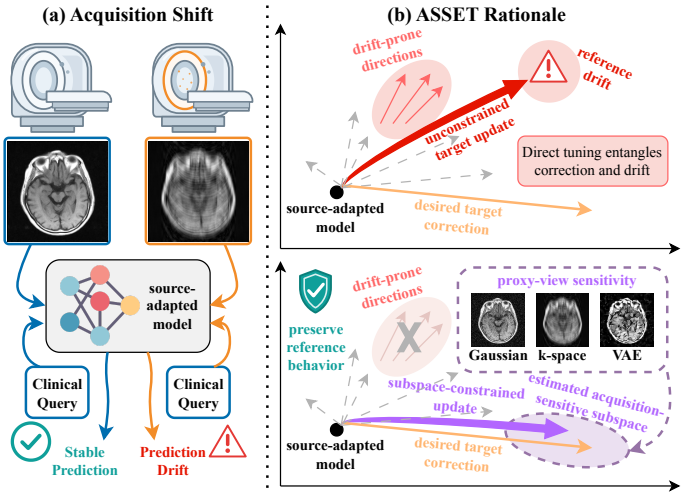

Figure 1: Acquisition-quality shift and ASSET intuition. (a) The clinical query remains comparable, but image acquisition changes from a high-quality development setting to a lower-quality deployment setting. (b) Direct target tuning can follow a mixed update that causes drift under the original acquisition condition, whereas ASSET estimates acquisition-sensitive parameter directions and constrains test-time adaptation to the estimated subspace.

> ***(RQ)*** *How can a source-adapted multimodal medical model adapt at test time to lower-quality acquisition using only a few target examples without overfitting or drifting?*

We formalize this RQ as test-time adaptation under directional acquisition-quality shift. Given a source-adapted multimodal medical model and a small labeled target support set from the lower-quality acquisition distribution, the goal is to improve performance under the lower-quality target condition while preserving behavior under the original high-quality acquisition condition. Direct fine-tuning on the support set is a natural strategy, but the labels do not identify which parameter directions correspond to acquisition-quality correction. An unconstrained update can therefore entangle the desired correction with support-set overfitting and drift under the original acquisition condition. This motivates the hypothesis that acquisition-like perturbations of target examples can reveal update directions along which the model's predictions are sensitive to acquisition quality, even when the exact target acquisition process is unknown.

To address directional acquisition-quality shift, we propose ASSET, an *Acquisition-Sensitive Subspace Estimation* method for test-time adaptation. At a high level, ASSET treats adaptation as selective correction in the model's weight space: first estimate which parameter directions are tied to the acquisition shift, then update only along those directions. A source-adapted model contains task-relevant behavior that should be retained, but some directions in its weight space may correspond to acquisition-sensitive behavior that becomes unreliable when image quality changes. As illustrated in Figure 1b, ASSET estimates a compact acquisition-sensitive parameter subspace from the target support set. Adaptation is then restricted to the estimated subspace, while source-side reference constraints limit drift. The method, therefore, separates subspace estimation from target fitting.

ASSET obtains its sensitivity signal from acquisition-like augmentations of target support examples. We call the resulting augmented inputs proxy views of the original examples. The proxy views are not assumed to reproduce the true scanner, protocol, or reconstruction process. Instead, proxy views

reveal prediction and gradient sensitivity to visual changes that resemble acquisition-quality variation. ASSET uses these sensitivity signals to estimate an acquisition-sensitive parameter subspace, and then optimizes target supervision and consistency under the subspace constraint. Because the relevant directions are local to the current model, ASSET repeats the estimate-and-update process multiple times as the model changes. The support examples are also weighted by reliability and informativeness, rather than treated as equally useful signals.

We evaluate ASSET on SLAKE, RadImageNet-VQA, and GMAI-MMBench (Liu et al., 2021; Butsanets et al., 2025; Chen et al., 2024) with InternVL3.5-8B and MedGemma-4B-IT (Wang et al., 2025; Sellergren et al., 2025). Across 30 model–dataset–shift combinations, ASSET improves target accuracy under acquisition-quality shift by 2.62% on average while preserving no-shift performance. Under the strongest shifts, ASSET improves accuracy by 4.12%, showing that the correction generalizes to severe acquisition-quality changes. On a real chest X-ray protocol shift from posteroanterior to anteroposterior views, ASSET improves target accuracy by 2.98% while preserving accuracy on the source view. Existing generic test-time and image-side correction baselines remain limited because they either fit target examples too directly or leave the deployed multimodal model unchanged. Mechanism analyses show that ASSET-estimated subspaces align better with target-correction directions than random subspaces. Because ASSET relies on proxy-view sensitivity rather than a fixed medical noise model, the same framework can extend beyond acquisition shift. We also evaluate ASSET on the Visual Robustness Benchmark (Ishmam et al., 2025) and observe strongest-shift improvements in 5 of 6 corruption families. Our contributions are:

•We formulate directional acquisition-quality shift as a test-time adaptation problem for multimodal medical models, where a source-adapted model must adapt to lower-quality target acquisition from a small labeled support set without drifting under no-shift conditions.

•We introduce ASSET, a test-time adaptation method that estimates acquisition-sensitive update directions from proxy-view prediction and gradient sensitivity, restricts updates to the resulting subspace, and iteratively reweights support signals.

•We demonstrate on SLAKE, RadImageNet-VQA, and GMAI-MMBench, using InternVL3.5-8B and MedGemma-4B-IT, that ASSET improves target accuracy under acquisition-quality shift by 2.62% on average and by 4.12% under the strongest acquisition-quality shift while preserving no-shift performance. ASSET also corrects a real chest X-ray protocol shift without source drift and improves strongest-shift accuracy in 5 of 6 general VQA shift families.

2 Related Work

Medical acquisition-quality shift. Medical acquisition-quality shift is usually treated as image-distribution heterogeneity caused by scanner hardware, acquisition protocols, reconstruction pipelines, dose, or site differences (Guan and Liu, 2022; Mali et al., 2021). Existing approaches mainly correct the image side, including feature harmonization after feature extraction (Fortin et al., 2018), image harmonization before prediction (Hognon et al., 2024), and modality-specific degradation models for MRI noise, k-space undersampling, or low-dose CT artifacts (Gudbjartsson and Patz, 1995; Zbontar et al., 2018; Chen et al., 2017). Medical domain adaptation moves closer to model-level correction, but typically targets image-only segmentation or diagnosis models (Kamnitsas et al., 2017; Dou et al., 2018; Yang et al., 2019). ASSET instead adapts the deployed medical VLM itself at test time.

Robustness evaluation and robust training. Robustness benchmarks expose failures under controlled corruptions and perturbations (Hendrycks and Dietterich, 2019; Imam et al., 2025; Chen et al., 2023). Robust training and robust fine-tuning improve behavior before deployment through pretraining diversity (Fang et al., 2022), OOD fine-tuning objectives (Shu et al., 2023), semantic anchors (Han et al., 2024), or adversarial vision-encoder fine-tuning (Schlarmann et al., 2024). These works either evaluate robustness under controlled distribution shifts or build robustness into a model before deployment. ASSET instead adapts an already deployed medical VLM by inferring a correction signal from a small target support set.

Test-time adaptation for vision-language models. Test-time adaptation commonly uses entropy minimization (Wang et al., 2021), augmentation consistency (Zhang et al., 2022a), continual self-training (Wang et al., 2022), prompts (Shu et al., 2022), caches (Karmanov et al., 2024), or prediction

aggregation (Kaya et al., 2026). These signals can be ambiguous under acquisition shift: fitting a few target examples may capture support-specific artifacts or move the model away from no-shift behavior. ASSET therefore estimates parameter directions sensitive to acquisition-quality changes and restricts supervised adaptation to the estimated subspace. Appendix A provides further discussion.

3 Problem Statement

This section first formulates acquisition-quality shift as a test-time correction problem for source-adapted medical VLMs. We then use controlled degradations to demonstrate that this shift leads to systematic failures. Finally, we analyze direct target fitting to show why target labels alone are insufficient for selecting correction directions.

Acquisition-quality shift. Let f_{θ_0} denote a multimodal medical model adapted under a high-quality source acquisition condition P_s. At deployment, the same model may receive inputs from a lower-quality target condition P_t, where scanner, protocol, dose, reconstruction, or artifact differences alter the image evidence. For an input $u = (x, q)$, the image x carries acquisition-dependent visual information, while the query q specifies the clinical task. We study the directional correction problem where the task is unchanged but the visual evidence shifts, so a model competent under P_s must restore accurate predictions under P_t.

In medical deployment, target images from a new scanner, protocol, or reconstruction pipeline may be available, but annotation requires expert interpretation and site-specific validation. Large labeled target sets are therefore impractical, while a small calibration set is realistic. We assume access to a small labeled target support set $S_t = \{(u_i^t, y_i^t)\}_{i=1}^n$ sampled from P_t. Let $F_\theta(u)$ denote the model prediction for u, and let $\ell(F_\theta(u), y)$ denote the supervised loss. These labels provide target supervision, but limited examples make direct fine-tuning unstable.

The desired update should improve performance under the target acquisition condition while preserving behavior under the original acquisition condition. We write this correction objective as

$$\min_{\Delta \in \mathcal{A}} R_t(\theta_0 + \Delta) \quad \text{s.t.} \quad D_{\text{ref}}(\theta_0 + \Delta, \theta_0) \leq \varepsilon_{\text{ref}},$$

where R_t is target risk under P_t, $\Delta \in \mathcal{A}$ is the allowed update, and D_{ref} measures drift from the source-adapted model under the original acquisition condition. The support set is the empirical signal for this correction, but its labels alone do not specify which model changes correspond to acquisition-quality correction rather than support-specific fitting or reference drift.

Controlled degradations reveal systematic failures. The formulation above treats P_t as a lower-quality acquisition condition, but real scanner, protocol, reconstruction, and post-processing shifts are difficult to parameterize as reproducible target transformations. We therefore instantiate P_t with controlled target degradations that keep the clinical query and label fixed while changing the image evidence. Following prior work on medical acquisition shift and modality-specific image degradation, we use four physics-inspired target families: CT-Gaussian, CT-Poisson, MRI-Rician, and MRI-k-space (Chen et al., 2017; Gudbjartsson and Patz, 1995; Zbontar et al., 2018; Guo et al., 2024; Hognon et al., 2024). We also use a VAE-based learned degradation to probe quality changes not captured by explicit CT/MRI noise models (Kingma and Welling, 2014; Rombach et al., 2022; Moyer et al., 2020; Rais et al., 2024). Each family is evaluated at multiple severity levels to represent increasing target acquisition-quality shift; Appendix D gives the construction details.

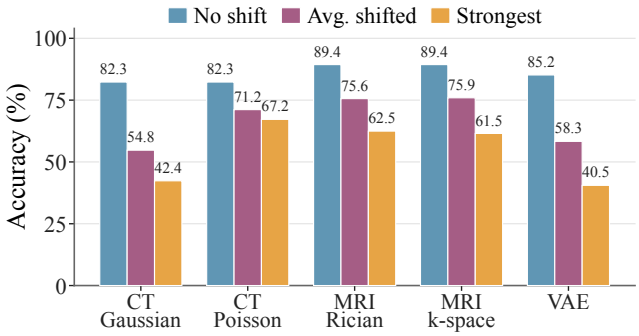

Figure 2: **Accuracy under controlled acquisition-quality degradations.** Bars average over two backbones and three datasets, comparing no-shift, average shifted severities, and the strongest severity. CT and MRI pairs share modality-specific no-shift sets; VAE uses mixed images. Controlled acquisition changes degrade source-adapted medical VLMs.

Figure 2 empirically demonstrates that these controlled acquisition-quality shifts affect source-adapted medical VLMs under the experimental setup in Sec. 5.1. Accuracy decreases from the no-shift condition to the average over shifted severities and drops further under the strongest severity across 30 model–dataset–family combinations. This establishes the empirical premise of our problem: acquisition-quality shift alone can systematically degrade source-adapted medical VLMs, motivating post-deployment correction for the target acquisition condition.

Direct target fitting entangles correction and drift. Given the degradation observed above, a natural correction strategy is supervised optimization on the labeled support set S_t:

$$\Delta_{\text{fit}} = \arg\min_{\Delta \in \mathcal{A}} \frac{1}{|S_t|} \sum_{(u_i^t, y_i^t) \in S_t} \ell(F_{\theta_0+\Delta}(u_i^t), y_i^t).$$

Here Δ is the update applied to the source-adapted model. The objective fits S_t, but it does not distinguish acquisition-quality correction from support-specific fitting or reference drift.

Figure 3 shows this ambiguity empirically. Direct fitting is worse than the source-adapted model in the no-shift condition and under mild shifts, but becomes better as the shift becomes stronger. This behavior shows that target labels specify what the model should predict on the support set, but not which parameter directions provide acquisition-sensitive correction.

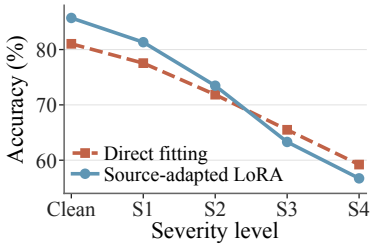

Figure 3: **Preservation–correction trade-off under direct target fitting.** Clean is the no-shift condition, and S1–S4 denote increasing target-shift severity. Direct target fitting trades no-shift preservation for shifted-target correction.

4 Method

We develop ASSET as a subspace-constrained adaptation method for acquisition-quality correction. ASSET applies proxy transformations to target support examples, differentiates the resulting prediction changes to estimate a compact update subspace, and optimizes target and source-side objectives only within this subspace. It further weights support examples by their current adaptation signal and iterates between direction estimation and constrained updates.

Proxy views for acquisition-sensitive probing. In deployment, the exact acquisition shift is rarely available as a reproducible image transformation. Rather than modeling the target process directly, ASSET probes target support images with proxy transformations and measures how the current prediction changes. For a target support input $u_i^t = (x_i^t, q_i^t)$, let $\mathcal{T}$ denote a bank of proxy transformations. ASSET samples a small view batch $\mathcal{V}_i = \{\tau_{i,b}\}_{b=1}^{n_v}$ from $\mathcal{T}$, where $u_i^{t,\tau} = (\tau(x_i^t), q_i^t)$ keeps the query fixed and transforms only the image. This approximates the perturbation average without enumerating $\mathcal{T}$, so prediction changes capture image-evidence sensitivity under a fixed query. We quantify this response with a prediction discrepancy $D(\cdot,\cdot)$, differentiable in its first argument. Let $\text{sg}[\cdot]$ denote stop-gradient. We define proxy-view sensitivity as

$$d_\theta(u_i^t, \tau) = D\big(F_\theta(u_i^{t,\tau}), \text{sg}[F_\theta(u_i^t)]\big), \quad \bar{d}_\theta(u_i^t; \mathcal{V}_i) = \frac{1}{|\mathcal{V}_i|} \sum_{\tau \in \mathcal{V}_i} d_\theta(u_i^t, \tau). \tag{1}$$

For closed-ended tasks, D can compare distributions over candidate labels; for sequence outputs, it can compare teacher-forced token distributions. Large $\bar{d}_\theta(u_i^t; \mathcal{V}_i)$ indicates that the model prediction is sensitive to acquisition-like changes in the image evidence. The stop-gradient treats the untransformed prediction as a local reference, so the sensitivity gradient captures directions that change the proxy-view response rather than directions that move both predictions together. ASSET uses this sensitivity to decide where adaptation should act in parameter space.

The bank $\mathcal{T}$ can combine general image-quality transformations, such as blur, noise, or compression, with domain-informed transformations when coarse domain knowledge is available. For medical deployment, CT- or MRI-informed transformations such as low-dose noise or k-space perturbations provide tractable approximations of plausible acquisition-quality variation, giving ASSET evidence for subspace estimation even when the exact target process is unknown.

Estimating acquisition-sensitive directions. Proxy-view sensitivity identifies target support examples for which the current model is sensitive to acquisition-like perturbations, but adaptation requires parameter directions that control this sensitivity. ASSET obtains these directions by differentiating the view-batch discrepancy with respect to the adaptable parameters. Let ϕ_m denote the adaptable parameters in model module m. For the current model θ, ASSET computes

$$g_{i,m} = \nabla_{\phi_m} \bar{d}_\theta(u_i^t; \mathcal{V}_i), \tag{2}$$

where each gradient describes how module m affects average proxy-view sensitivity. For each module m, ASSET stacks the sensitivity gradients into

$$G_m = [g_{1,m}^\top; \ldots; g_{i,m}^\top; \ldots], \tag{3}$$

where rows are target support examples and columns are adaptable parameters. ASSET then extracts parameter-space eigendirections using the dual Gram form

$$K_m = G_m G_m^\top = U_m \Lambda_m U_m^\top, \qquad V_m = G_m^\top U_m \Lambda_m^{-1/2}. \tag{4}$$

Here K_m is a Gram matrix over support examples, and the columns of V_m are parameter-space eigendirections of $G_m^\top G_m$ for nonzero eigenvalues. Let $\lambda_{m,1} \geq \lambda_{m,2} \geq \cdots$ be the eigenvalues. ASSET selects the smallest $k_{m,\rho}$ whose cumulative energy reaches threshold ρ, then applies a per-module rank cap $k_{\max}$:

$$k_{m,\rho} = \min\left\{ k : \frac{\sum_{a=1}^{k} \lambda_{m,a}}{\sum_a \lambda_{m,a}} \geq \rho \right\}, \qquad k_m = \min(k_{m,\rho}, k_{\max}). \tag{5}$$

The rank cap keeps optimization efficient when many small directions satisfy the energy threshold. We set $\rho = 0.95$ and $k_{\max} = 16$ throughout. The first k_m columns of V_m form B_m, the acquisition-sensitive subspace for module m. This basis is not a generic low-rank approximation; it is estimated from prediction and gradient changes induced by acquisition-like proxy views.

Subspace-constrained adaptation. The estimated subspace identifies where acquisition-sensitive correction should act, but adaptation still needs an objective that fits the target support set without freely updating all adaptable parameters. ASSET therefore optimizes only the coefficients of the module-wise subspace directions. For module m, let α_m denote the trainable coefficients in basis B_m, and define

$$\Delta_m(\alpha_m) = -B_m \alpha_m. \tag{6}$$

The full update $\Delta(\alpha)$ applies $\Delta_m(\alpha_m)$ to each adaptable module. Within this constrained space, ASSET uses target labels through the supervised loss

$$\mathcal{L}_{\text{tgt}}(\theta; S_t) = \frac{1}{|S_t|} \sum_{(u_i^t, y_i^t) \in S_t} \ell(F_\theta(u_i^t), y_i^t). \tag{7}$$

This term specifies target behavior, while proxy-view consistency stabilizes target predictions:

$$\mathcal{L}_{\text{cons}}(\theta; S_t, \mathcal{T}) = \frac{1}{|S_t|} \sum_{u_i^t \in S_t} \mathbb{E}_{\tau \in \mathcal{T}} D\big(F_\theta(u_i^{t,\tau}), \text{sg}[F_\theta(u_i^t)]\big). \tag{8}$$

This term reduces prediction changes caused by acquisition-like perturbations of the same target example. When source or no-shift reference examples are available, ASSET can further anchor the update through reference supervision. For $S_s = \{(u_j^s, y_j^s)\}_{j=1}^{n_s}$, this gives

$$\mathcal{L}_{\text{src}}(\theta; S_s) = \frac{1}{|S_s|} \sum_{(u_j^s, y_j^s) \in S_s} \ell(F_\theta(u_j^s), y_j^s). \tag{9}$$

This keeps the adapted model aligned with labeled no-shift behavior. A preservation loss can also penalize deviation from the source-adapted model:

$$\mathcal{L}_{\text{pres}}(\theta; S_s, \theta_0) = \frac{1}{|S_s|} \sum_{u_j^s \in S_s} D\big(F_\theta(u_j^s), \text{sg}[F_{\theta_0}(u_j^s)]\big). \tag{10}$$

The final subspace objective combines these terms:

$$\mathcal{L}(\alpha) = \mathcal{L}_{\text{tgt}} + \lambda_{\text{cons}} \mathcal{L}_{\text{cons}} + \lambda_{\text{src}} \mathcal{L}_{\text{src}} + \lambda_{\text{pres}} \mathcal{L}_{\text{pres}} + \lambda_\alpha \|\alpha\|_2^2. \tag{11}$$

ASSET optimizes $\alpha^\star = \arg\min_\alpha \mathcal{L}(\alpha)$, with each loss evaluated at $\theta + \Delta(\alpha)$, yielding $\theta + \Delta(\alpha^\star)$. Terms involving S_s are omitted when no reference set is available. We set $\lambda_{\text{cons}} = \lambda_{\text{src}} = \lambda_{\text{pres}} = 1$ and $\lambda_\alpha = 10^{-3}$ throughout.

Iterative estimate-and-update with influence weighting. The acquisition-sensitive subspace is local to the current model. After a constrained update, previous proxy-view sensitivities may no longer reflect current behavior, making both parameter directions and example weights outdated. ASSET therefore applies the estimate-and-update procedure iteratively. At round r, proxy discrepancies, sensitivity gradients, and the constrained objective are instantiated under $\theta^{(r)}$.

Different target examples provide different evidence for the next subspace estimate: a useful example should be reliably interpreted while still exposing proxy-view sensitivity. Let $C(F_\theta(u), y)$ be a task reliability functional that is larger when the model prediction is more reliable for label y. At round r, ASSET samples $\mathcal{V}_i^{(r)}$ and computes

$$c_i^{(r)} = C(F_{\theta^{(r)}}(u_i^t), y_i^t), \qquad s_i^{(r)} = \bar{d}_{\theta^{(r)}}(u_i^t; \mathcal{V}_i^{(r)}) \tag{12}$$

as the current reliability and proxy-view sensitivity scores. To reduce oscillations from the small support set, ASSET smooths them across rounds. With $\bar{c}_i^{(0)} = c_i^{(0)}$ and $\bar{s}_i^{(0)} = s_i^{(0)}$, for $r > 0$

$$\bar{c}_i^{(r)} = \beta \bar{c}_i^{(r-1)} + (1-\beta) c_i^{(r)}, \qquad \bar{s}_i^{(r)} = \beta \bar{s}_i^{(r-1)} + (1-\beta) s_i^{(r)}. \tag{13}$$

ASSET assigns larger influence to examples whose current reliability and proxy-view sensitivity exceed the corresponding smoothed scores:

$$\omega_i^{(r)} = \left[\sigma\left(\frac{c_i^{(r)} - \bar{c}_i^{(r)}}{\mathrm{std}(c^{(r)}) + \epsilon} \right) \cdot \sigma\left(\frac{s_i^{(r)} - \bar{s}_i^{(r)}}{\mathrm{std}(s^{(r)}) + \epsilon} \right) \right]^\gamma \quad \text{and} \quad w_i^{(r)} = \frac{\omega_i^{(r)}}{\sum_{i'=1}^{|S_t|} \omega_{i'}^{(r)}}. \tag{14}$$

where $\sigma(\cdot)$ is the sigmoid function, the standard deviations are computed across target support examples in the current round, ϵ is a small constant for numerical stability, and we set $\beta = 0.7$ and $\gamma = 1$ throughout. The normalized weights enter the next subspace estimate by scaling each example gradient:

$$G_m^{(r)} = [\sqrt{w_1^{(r)}}\, g_{1,m}^{(r)\top}; \ldots; \sqrt{w_i^{(r)}}\, g_{i,m}^{(r)\top}; \ldots]. \tag{15}$$

ASSET repeats the dual-Gram eigendirection selection on these weighted stacks, optimizes Eq. (11) under the current model $\theta^{(r)}$, and updates

$$\theta^{(r+1)} = \theta^{(r)} + \Delta^{(r)}(\alpha^{(r)}). \tag{16}$$

Repeating estimation and constrained adaptation keeps the subspace aligned with the current adapted model rather than a single initial estimate. Algorithm 1 in Appendix B summarizes the full procedure.

Although the main setting is medical acquisition-quality shift, ASSET is defined through proxy transformations, prediction sensitivity, and subspace-constrained adaptation rather than a medical-specific degradation model. The same estimate-and-update procedure can be applied to broader distribution shifts by choosing an appropriate proxy bank, as validated in Sec. 5.2.

5 Experiments

5.1 Experimental Setup

Model and datasets. We evaluate ASSET on SLAKE, RadImageNet-VQA, and GMAI-MMBench (Liu et al., 2021; Butsanets et al., 2025; Chen et al., 2024), using InternVL3.5-8B and MedGemma-4B-IT as medical VLM backbones (Wang et al., 2025; Sellergren et al., 2025). For each dataset, we train a LoRA adapter on the no-shift source split to obtain the source-adapted model. We evaluate this model under 5 acquisition-quality shift families: CT-Gaussian, CT-Poisson, MRI-Rician, MRI-k-space, and VAE. We further evaluate a real acquisition-protocol shift from posteroanterior (PA) to anteroposterior (AP) chest radiographs in MIMIC-CXR-VQA (Bae et al., 2023; Johnson et al., 2019). Appendix D details the dataset construction and shift grids.

Baselines. We compare against baselines spanning 5 adaptation mechanisms under visual distribution shift: image correction before VLM inference, prediction aggregation without model updates, distribution alignment of target statistics or calibration parameters, cache-based adaptation from support features, and test-time parameter updates. The baselines are histogram matching, non-local means denoising (Buades et al., 2005), TTAug, ZERO (Kaya et al., 2026; Farina et al., 2024), UPA,

Table 1: Main results on medical VQA acquisition-quality shifts. LoRA is the source-adapted reference. Baselines are grouped by adaptation mechanism. Each dataset reports no-shift accuracy (Clean), average shifted accuracy (Avg.), and strongest-shift accuracy (Str.), averaged over 2 backbones and 5 shift families. All accuracies are percentages. Arrows indicate higher is better. Best results in each column are in bold, and second-best results are underlined.

Mechanism	Method	SLAKE			RadImageNet			GMAI			Overall		
		Clean↑	Avg.↑	Str.↑	Clean↑	Avg.↑	Str.↑	Clean↑	Avg.↑	Str.↑	Clean↑	Avg.↑	Str.↑
Reference	LoRA	88.1	79.0	73.0	96.3	71.8	51.7	72.7	50.8	39.8	85.7	67.2	54.8
Image correction	HistMatch	73.6	67.2	63.6	84.2	60.7	43.1	56.3	41.9	34.7	71.4	56.6	47.2
	NLM	73.4	69.1	66.2	83.7	62.9	47.5	57.3	43.8	35.4	71.4	58.6	49.7
Prediction aggregation	TTAug	87.7	78.7	72.6	96.5	72.1	52.2	73.3	50.9	39.7	85.8	67.2	54.8
	ZERO	87.8	78.9	72.6	96.5	72.2	51.9	**73.6**	**51.4**	39.8	**86.0**	67.5	54.8
Distribution alignment	UPA	73.7	68.3	64.4	74.0	56.8	41.6	53.2	40.1	31.9	67.0	55.1	46.0
	BCA	74.3	68.8	64.9	84.8	61.0	42.8	58.9	42.8	34.0	72.7	57.5	47.2
	CORAL	83.4	74.8	68.5	92.1	70.4	50.8	54.6	41.6	33.5	76.7	62.3	50.9
Cache adaptation	Tip-Adapter	67.5	61.0	57.1	92.3	62.4	43.1	70.3	49.6	39.9	76.7	57.7	46.7
	TDA	63.5	59.7	57.2	84.9	46.8	32.8	52.8	37.5	32.5	67.0	48.0	40.8
Test-time adaptation	Tent	87.8	78.5	71.8	96.3	71.6	51.4	73.0	50.6	**40.3**	85.7	66.9	54.5
	MEMO	**88.5**	79.4	73.5	96.6	71.9	52.3	72.2	51.1	39.5	85.8	67.5	55.1
	SAR	87.7	78.5	71.7	96.3	71.6	51.4	72.7	50.7	39.6	85.6	66.9	54.2
	TTAdapt	86.0	75.4	68.5	94.6	71.8	54.3	69.0	48.5	38.7	83.2	65.2	53.8
Ours	**ASSET**	88.1	**79.6**	**73.9**	**96.7**	**78.9**	**62.8**	72.8	50.9	40.0	85.9	**69.8**	**58.9**

BCA, CORAL (Roschewitz et al., 2023; Zhou et al., 2025; Sun and Saenko, 2016), Tip-Adapter, TDA (Zhang et al., 2022b; Karmanov et al., 2024), Tent, MEMO, SAR, and TTAdapt (Wang et al., 2021; Zhang et al., 2022a; Niu et al., 2023; Kaya et al., 2026). All methods start from the same source-adapted LoRA model.

Training and evaluation. At test time, ASSET uses a small labeled target support set and a no-shift reference set, with default size 32 for each. ASSET remains effective with smaller or resampled support sets (Appendix E). We instantiate the prediction discrepancy with JS divergence and run 5 estimate-and-update rounds with 8 optimization steps per round. Appendix C summarizes the fixed LoRA and ASSET training configurations.

5.2 Experimental Results

Acquisition-shift performance. Table 1 presents the main results under medical acquisition-quality shift. ASSET achieves the best shifted-target accuracy while preserving the no-shift accuracy of the source-adapted model. Compared with LoRA, ASSET improves average shifted accuracy from 67.2% to 69.8% and strongest-shift accuracy from 54.8% to 58.9%. It also outperforms MEMO, the strongest adaptation baseline, by 2.3% and 3.8% on these 2 shifted metrics.

The baselines show 2 common failure modes: prediction aggregation and most test-time adaptation methods stay close to LoRA, while image correction, cache adaptation, and distribution alignment often reduce accuracy. These results suggest that generic correction signals are either too weak or act at the wrong level for multimodal acquisition shift. ASSET instead turns the target support signal into a directional constraint, correcting shifted targets without unconstrained target fitting.

Real acquisition-protocol shift. Acquisition shift in clinical practice also arises from changes in imaging protocol. In MIMIC-CXR-VQA, PA radiographs form the source condition and AP radiographs form the target condition. The 2 views reverse the X-ray projection direction and change image appearance, while the proxy bank contains no view-related transformation. We adapt InternVL3.5-8B with 32 labeled AP support examples and 32 PA reference examples, and average results over 3 draws. Table 2 shows that ASSET improves AP accuracy from 63.5% to 66.5% while keeping PA accuracy at 80.4%. ASSET w/o subspace uses the same data

Table 2: Real acquisition-protocol shift results on MIMIC-CXR-VQA with InternVL3.5-8B. Accuracies are percentages averaged over 3 draws.

Method	AP (target)	PA (source)
LoRA	63.5	79.9
ASSET w/o subspace	63.0	66.2
ASSET	**66.5**	**80.4**

and objectives but lowers PA accuracy to 66.2%. The subspace constraint therefore separates target correction from source drift under a real acquisition change.

Broader distribution shifts. We further evaluate whether ASSET can correct broader visual distribution shifts beyond medical acquisition-quality shift. We use InternVL3.5-8B on 14 corruption types from the Visual Robustness Benchmark (Ishmam et al., 2025), grouped into the benchmark's 6 corruption families. For each target corruption, the proxy bank samples transformations from the other corruption families, excluding the target family. Table 3 shows that ASSET improves strongest-shift accuracy in 5 of 6 families and is slightly below LoRA on noise corruptions. These results further validate the broader claim in Sec. 4: proxy-sensitive subspace estimation can provide useful correction signals beyond modality-specific medical shifts.

Table 3: Distribution-shift results on the Visual Robustness Benchmark. n denotes corruption types; accuracies are percentages.

Family	n	LoRA		ASSET		ΔStr.
		Clean	Str.	Clean	Str.	
Noise	4	87.3	75.6	86.9	75.2	-0.4
Blur	2	87.3	74.8	87.3	75.7	+0.8
Weather	1	87.3	75.7	87.0	77.0	+1.3
Image Attr.	3	87.3	80.3	87.2	81.2	+0.9
Physical	2	87.3	81.7	86.8	82.2	+0.5
Digital	2	87.3	82.7	87.2	82.8	+0.2

Estimated subspaces capture correction directions. ASSET's subspace constraint is intended to select correction directions, not simply to reduce the number of trainable directions. We evaluate the eigendirection basis B_m defined by Eqs. (4)–(5) with a projection diagnostic. For each shift family, we compute supervised target-correction gradients on strongest-shift target test examples under the source-adapted model. We then project these gradients onto the ASSET basis estimated from support examples and report the fraction of gradient energy captured by the projection.

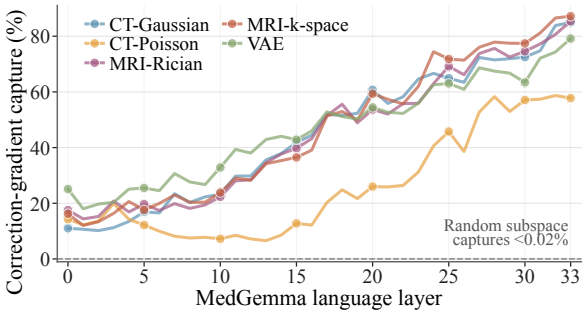

Figure 4: Projection diagnostic for ASSET-estimated subspaces on MedGemma/RadImageNet. The curves show target-correction gradient capture on target test examples; the dashed line is a random-direction control with the same module-wise ranks.

We use a random-direction control to validate the eigendirection selection in Eq. (5). This control keeps the same module-wise ranks as ASSET but uses random directions. Figure 4 shows that ASSET captures 25.9–46.4% of target-correction gradient energy across the 5 shift families, while the random-direction control captures less than 0.02%. This gap shows that proxy-view sensitivity selects structured directions aligned with target correction, rather than merely imposing a low-dimensional update. The captured fraction also increases in later language layers, indicating that the estimated directions concentrate where target-correction gradients are strongest.

Iterative estimate-and-update. ASSET treats the acquisition-sensitive subspace as a local estimate tied to the current model, rather than a static update space fixed before adaptation. To quantify the value of updating this estimate during adaptation, we track no-shift and strongest-shift accuracy across rounds. Figure 5a shows that strongest-shift accuracy rises from 46.7% before ASSET adaptation to 60.5% by round 4, while no-shift accuracy stays near 97% throughout. This trend indicates that iterative re-estimation progressively strengthens acquisition-quality correction while maintaining no-shift behavior.

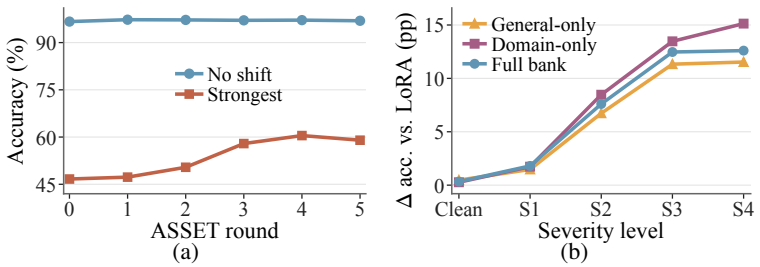

Figure 5: Ablation studies on MedGemma/RadImageNet. (a) Round dynamics: no-shift and strongest-shift accuracies are averaged over 5 acquisition shift families. (b) Proxy bank: accuracy changes are measured relative to the source-adapted LoRA model across severity levels.

Proxy transformation bank. To isolate the contribution of modality-specific proxy knowledge, we compare general image-quality transformations, modality-informed transformations, and their combination. Figure 5b shows that general transformations already improve shifted accuracy, indicating that generic image-quality probes can provide useful correction evidence. Modality-informed transformations contribute more at larger severities, suggesting that coarse modality information enables proxy views that better approximate plausible CT/MRI acquisition-quality changes. The full bank combines these complementary signals and is used in the main experiments. Appendix E further ablates the subspace constraint, training objectives, and influence weighting.

6 Conclusion

In this work, we study test-time adaptation for source-adapted medical VLMs under directional acquisition-quality shift. We first formulate the setting as target correction from a small labeled support set while preserving no-shift behavior. We next propose ASSET, which estimates acquisition-sensitive subspaces from proxy-view sensitivity and constrains adaptation to the estimated subspace. Experiments across medical VQA benchmarks and a real chest X-ray protocol shift show improved shifted-target accuracy with stable no-shift performance. Mechanism analyses further show that the estimated subspaces align with target-correction directions. Additional results on a broader visual robustness benchmark suggest that ASSET can extend beyond modality-specific medical shifts. A current limitation is that most of our medical target shifts are controlled degradations. Validating ASSET on real acquisitions from more clinical sites before deployment remains future work.

Acknowledgments and Disclosure of Funding

This work was supported in part by National Institutes of Health (NIH) grants R01EB037101, R01MH143267, and 1OT2OD038051-01. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

Seongsu Bae, Daeun Kyung, Jaehee Ryu, Eunbyeol Cho, Gyubok Lee, Sunjun Kweon, Jungwoo Oh, Lei Ji, Eric I. Chang, Tackeun Kim, and Edward Choi. EHRXQA: A multi-modal question answering dataset for electronic health records with chest X-ray images. In *Advances in Neural Information Processing Systems*, 2023. URL https://papers.nips.cc/paper_files/paper/2023/hash/0c007ebef1d11fd48da6ce4f54687db6-Abstract-Datasets_and_Benchmarks.html.

Antoni Buades, Bartomeu Coll, and Jean-Michel Morel. A non-local algorithm for image denoising. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2005. doi: 10.1109/CVPR.2005.38.

Léo Butsanets, Charles Corbière, Julien Khlaut, Pierre Manceron, and Corentin Dancette. RadImageNet-VQA: A large-scale CT and MRI dataset for radiologic visual question answering. *arXiv preprint arXiv:2512.17396*, 2025. URL https://arxiv.org/abs/2512.17396.

Chentao Cao, Zhun Zhong, Zhanke Zhou, Tongliang Liu, Yang Liu, Kun Zhang, and Bo Han. Noisy test-time adaptation in vision-language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=iylpeTI0Ql.

Hu Chen, Yi Zhang, Mannudeep K. Kalra, Feng Lin, Yang Chen, Peixi Liao, Jiliu Zhou, and Ge Wang. Low-dose CT with a residual encoder-decoder convolutional neural network. *IEEE Transactions on Medical Imaging*, 2017. URL https://arxiv.org/abs/1702.00288.

Pengcheng Chen, Jin Ye, Guoan Wang, Yanjun Li, Zhongying Deng, Wei Li, Tianbin Li, Haodong Duan, Ziyan Huang, Yanzhou Su, Benyou Wang, Shaoting Zhang, Bin Fu, Jianfei Cai, Bohan Zhuang, Eric J. Seibel, Yu Qiao, and Junjun He. GMAI-MMBench: A comprehensive multimodal evaluation benchmark towards general medical AI. In *Advances in Neural Information Processing Systems*, 2024. URL https://papers.nips.cc/paper_files/paper/2024/hash/ab7e02fd60e47e2a379d567f6b54f04e-Abstract-Datasets_and_Benchmarks_Track.html.

Shuo Chen, Jindong Gu, Zhen Han, Yunpu Ma, Philip Torr, and Volker Tresp. Benchmarking robustness of adaptation methods on pre-trained vision-language models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/a2a544e43acb8b954dc5846ff0d77ad5-Paper-Datasets_and_Benchmarks.pdf.

Zijie Cheng, Ariel Yuhan Ong, Siegfried K. Wagner, David A. Merle, Lie Ju, Hanyuan Zhang, Ruinian Chen, Linze Pang, Boxuan Li, Tiantian He, Anran Ran, Hongyang Jiang, Dawei Gabriel Yang, Ke Zou, Jocelyn Hui Lin Goh, Sahana Srinivasan, Andre Altmann, Daniel C. Alexander, Carol Y. Cheung, Yih Chung Tham, Pearse A. Keane, and Yukun Zhou. Understanding the robustness of vision-language models to medical image artefacts. *npj Digital Medicine*, 8:727, 2025. doi: 10.1038/s41746-025-02108-w. URL https://www.nature.com/articles/s41746-025-02108-w.

Qi Dou, Cheng Ouyang, Cheng Chen, Hao Chen, and Pheng-Ann Heng. Unsupervised cross-modality domain adaptation of convnets for biomedical image segmentations with adversarial loss. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pages 691–697, 2018. doi: 10.24963/ijcai.2018/96. URL https://www.ijcai.org/Proceedings/2018/96.

Alex Fang, Gabriel Ilharco, Mitchell Wortsman, Yuhao Wan, Vaishaal Shankar, Achal Dave, and Ludwig Schmidt. Data determines distributional robustness in contrastive language-image pre-training (CLIP). In *Proceedings of the 39th International Conference on Machine Learning*, pages 6216–6234, 2022. URL https://proceedings.mlr.press/v162/fang22a.html.

Matteo Farina, Gianni Franchi, Giovanni Iacca, Massimiliano Mancini, and Elisa Ricci. Frustratingly easy test-time adaptation of vision-language models. In *Advances in Neural Information Processing Systems*, 2024. doi: 10.52202/079017-4099. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/e92cb6f981a2cacb2a710ecaa0d7b141-Abstract-Conference.html.

Jean-Philippe Fortin, Nicholas Cullen, Yvette I. Sheline, Warren D. Taylor, Irem Aselcioglu, Philip A. Cook, Phil Adams, Crystal Cooper, Maurizio Fava, Patrick J. McGrath, Melvin McInnis, Mary L. Phillips, Madhukar H. Trivedi, Myrna M. Weissman, and Russell T. Shinohara. Harmonization of cortical thickness measurements across scanners and sites. *NeuroImage*, 167:104–120, 2018. doi: 10.1016/j.neuroimage.2017.11.024. URL https://doi.org/10.1016/j.neuroimage.2017.11.024.

Hao Guan and Mingxia Liu. Domain adaptation for medical image analysis: A survey. *IEEE Transactions on Biomedical Engineering*, 69(3):1173–1185, 2022. doi: 10.1109/TBME.2021.3117407. URL https://arxiv.org/abs/2102.09508.

Hao Guan, Yunbi Liu, Erkun Yang, Pew-Thian Yap, Dinggang Shen, and Mingxia Liu. Multi-site MRI harmonization via attention-guided deep domain adaptation for brain disorder identification. *Medical Image Analysis*, 71:102076, 2021. doi: 10.1016/j.media.2021.102076. URL https://doi.org/10.1016/j.media.2021.102076.

Hákon Gudbjartsson and Samuel Patz. The rician distribution of noisy MRI data. *Magnetic Resonance in Medicine*, 1995. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC2254141/.

Brian Guo, Darui Lu, Gregory Szumel, Rongze Gui, Tingyu Wang, Nicholas Konz, and Maciej A. Mazurowski. The impact of scanner domain shift on deep learning performance in medical imaging: An experimental study. *arXiv preprint arXiv:2409.04368*, 2024. URL https://arxiv.org/abs/2409.04368.

Jinwei Han, Zhiwen Lin, Zhongyisun Sun, Yingguo Gao, Ke Yan, Shouhong Ding, Yuan Gao, and Gui-Song Xia. Anchor-based robust finetuning of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26919–26928, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Han_Anchor-based_Robust_Finetuning_of_Vision-Language_Models_CVPR_2024_paper.html.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *The Seventh International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HJz6tiCqYm.

Clément Hognon, Pierre-Henri Conze, Vincent Bourbonne, Olivier Gallinato, Thierry Colin, Vincent Jaouen, and Dimitris Visvikis. Contrastive image adaptation for acquisition shift reduction in medical imaging. *Artificial Intelligence in Medicine*, 148:102747, 2024. doi: 10.1016/j.artmed.2023.102747. URL https://doi.org/10.1016/j.artmed.2023.102747.

Raza Imam, Rufael Marew, and Mohammad Yaqub. On the robustness of medical vision-language models: Are they truly generalizable? *arXiv preprint arXiv:2505.15425*, 2025. URL https://arxiv.org/abs/2505.15425.

Farhan Ishmam, Ishmam Tashdeed, Talukder Asir Saadat, Hamjajul Ashmafee, Abu Raihan Mostofa Kamal, and Azam Hossain. Visual robustness benchmark for visual question answering (VQA). In *Proceedings of the Winter Conference on Applications of Computer Vision*, pages 6623–6633, 2025. URL https://openaccess.thecvf.com/content/WACV2025/html/Ishmam_Visual_Robustness_Benchmark_for_Visual_Question_Answering_VQA_WACV_2025_paper.html.

Alistair E. W. 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*, 6:317, 2019. doi: 10.1038/s41597-019-0322-0. URL https://www.nature.com/articles/s41597-019-0322-0.

Konstantinos Kamnitsas, Christian Baumgartner, Christian Ledig, Virginia F. J. Newcombe, Joanna P. Simpson, Andrew D. Kane, David K. Menon, Aditya Nori, Antonio Criminisi, Daniel Rueckert, and Ben Glocker. Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. In *Information Processing in Medical Imaging*, pages 597–609, 2017. doi: 10.1007/978-3-319-59050-9_47. URL https://arxiv.org/abs/1612.08894.

Adilbek Karmanov, Dayan Guan, Shijian Lu, Abdulmotaleb El Saddik, and Eric Xing. Efficient test-time adaptation of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Karmanov_Efficient_Test-Time_Adaptation_of_Vision-Language_Models_CVPR_2024_paper.html.

Mehmet Onurcan Kaya, Desmond Elliott, and Dim P. Papadopoulos. Efficient test-time scaling for small vision-language models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=rClkte0ZTp.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014. URL https://arxiv.org/abs/1312.6114.

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. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2306.00890.

Fanfei Li, Thomas Klein, Wieland Brendel, Robert Geirhos, and Roland S. Zimmermann. LAION-C: An out-of-distribution benchmark for web-scale vision models. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025. URL https://proceedings.mlr.press/v267/li25aw.html.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. SLAKE: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *IEEE International Symposium on Biomedical Imaging*, 2021. URL https://arxiv.org/abs/2102.09542.

Shruti Atul Mali, Abdalla Ibrahim, Henry C. Woodruff, Vincent Andrearczyk, Henning Müller, Sergey Primakov, Zohaib Salahuddin, Avishek Chatterjee, and Philippe Lambin. Making radiomics more reproducible across scanner and imaging protocol variations: A review of harmonization methods. *Journal of Personalized Medicine*, 11(9):842, 2021. doi: 10.3390/jpm11090842. URL https://www.mdpi.com/2075-4426/11/9/842.

Daniel Moyer, Greg Ver Steeg, Chantal M. W. Tax, and Paul M. Thompson. Scanner invariant representations for diffusion MRI harmonization. *Magnetic Resonance in Medicine*, 84(4):2174–2189, 2020. doi: 10.1002/mrm.28243. URL https://doi.org/10.1002/mrm.28243.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=g2YraF75Tj.

Fanny Orlhac, Jakoba J. Eertink, Anne-Ségolène Cottereau, Josée M. Zijlstra, Catherine Thieblemont, Michel Meignan, Ronald Boellaard, and Irène Buvat. A guide to ComBat harmonization of imaging biomarkers in multicenter studies. *Journal of Nuclear Medicine*, 63(2):172–179, 2022. doi: 10.2967/jnumed.121.262464. URL https://jnm.snmjournals.org/content/63/2/172.

Ziyuan Qin, Huahui Yi, Qicheng Lao, and Kang Li. Medical image understanding with pretrained vision language models: A comprehensive study. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=txlWziuCE5W.

Khadija Rais, Mohamed Amroune, Abdelmadjid Benmachiche, and Mohamed Yassine Haouam. Exploring variational autoencoders for medical image generation: A comprehensive study. *arXiv preprint arXiv:2411.07348*, 2024. URL https://arxiv.org/abs/2411.07348.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Rombach_High-Resolution_Image_Synthesis_With_Latent_Diffusion_Models_CVPR_2022_paper.html.

Mélanie Roschewitz, Galvin Khara, Joe Yearsley, Nisha Sharma, Jonathan J. James, Éva Ambrózay, Adam Heroux, Peter Kecskemethy, Tobias Rijken, and Ben Glocker. Automatic correction of performance drift under acquisition shift in medical image classification. *Nature Communications*, 14:6608, 2023. doi: 10.1038/s41467-023-42396-y. URL https://www.nature.com/articles/s41467-023-42396-y.

Christian Schlarmann, Naman Deep Singh, Francesco Croce, and Matthias Hein. Robust CLIP: Unsupervised adversarial fine-tuning of vision embeddings for robust large vision-language models. In *Proceedings of the 41st International Conference on Machine Learning*, pages 43685–43704. PMLR, 2024. URL https://proceedings.mlr.press/v235/schlarmann24a.html.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, et al. MedGemma technical report. *arXiv preprint arXiv:2507.05201*, 2025. URL https://arxiv.org/abs/2507.05201.

Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. In *Advances in Neural Information Processing Systems*, 2022. URL https://papers.nips.cc/paper_files/paper/2022/hash/5bf2b802e24106064dc547ae9283bb0c-Abstract-Conference.html.

Yang Shu, Xingzhuo Guo, Jialong Wu, Ximei Wang, Jianmin Wang, and Mingsheng Long. CLIPood: Generalizing CLIP to out-of-distributions. In *Proceedings of the 40th International Conference on Machine Learning*, pages 31716–31731. PMLR, 2023. URL https://proceedings.mlr.press/v202/shu23a.html.

Baochen Sun and Kate Saenko. Deep CORAL: Correlation alignment for deep domain adaptation. In *European Conference on Computer Vision Workshops*, 2016. URL https://arxiv.org/abs/1607.01719.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *The Ninth International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=uXl3bZLkr3c.

Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Wang_Continual_Test-Time_Domain_Adaptation_CVPR_2022_paper.html.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. URL https://arxiv.org/abs/2508.18265.

Peng Xia, Ze Chen, Juanxi Tian, Yangrui Gong, Ruibo Hou, Yue Xu, Zhenbang Wu, Zhiyuan Fan, Yiyang Zhou, Kangyu Zhu, Wenhao Zheng, Zhaoyang Wang, Xiao Wang, Xuchao Zhang, Chetan Bansal, Marc Niethammer, Junzhou Huang, Hongtu Zhu, Yun Li, Jimeng Sun, Zongyuan Ge, Gang Li, James Zou, and Huaxiu Yao. Cares: A comprehensive benchmark of trustworthiness in medical vision language models. In *Thirty-eighth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/fde7f40f8ced5735006810534dc66b33-Paper-Datasets_and_Benchmarks_Track.pdf.

Zehao Xiao, Shilin Yan, Jack Hong, Jiayin Cai, Xiaolong Jiang, Yao Hu, Jiayi Shen, Qi Wang, and Cees G. M. Snoek. DynaPrompt: Dynamic test-time prompt tuning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=EFZEdHB3Mp.

Junlin Yang, Nicha C. Dvornek, Fan Zhang, Julius Chapiro, MingDe Lin, and James S. Duncan. Unsupervised domain adaptation via disentangled representations: Application to cross-modality liver segmentation. In *Medical Image Computing and Computer Assisted Intervention*, pages 255–263, 2019. doi: 10.1007/978-3-030-32245-8_29. URL https://arxiv.org/abs/1907.13590.

Xu Yang, Xuan Chen, Moqi Li, Kun Wei, and Cheng Deng. A versatile framework for continual test-time domain adaptation: Balancing discriminability and generalizability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Yang_A_Versatile_Framework_for_Continual_Test-Time_Domain_Adaptation_Balancing_Discriminability_CVPR_2024_paper.html.

Hee Suk Yoon, Eunseop Yoon, Joshua Tian Jin Tee, Mark Hasegawa-Johnson, Yingzhen Li, and Chang D. Yoo. C-TPT: Calibrated test-time prompt tuning for vision-language models via text feature dispersion. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=jzzEHTBFOT.

Jure Zbontar, Florian Knoll, Anuroop Sriram, Tullie Murrell, Zhengnan Huang, Matthew J. Muckley, Aaron Defazio, Ruben Stern, Patricia Johnson, Mary Bruno, Marc Parente, Krzysztof J. Geras, Joe Katsnelson, Hersh Chandarana, Zizhao Zhang, Michal Drozdzal, Adriana Romero, Michael Rabbat, Pascal Vincent, Nafissa Yakubova, James Pinkerton, Duo Wang, Erich Owens, C. Lawrence Zitnick, Michael P. Recht, Daniel K. Sodickson, and Yvonne W. Lui. fastMRI: An open dataset and benchmarks for accelerated MRI. *arXiv preprint arXiv:1811.08839*, 2018. URL https://arxiv.org/abs/1811.08839.

Marvin Zhang, Sergey Levine, and Chelsea Finn. MEMO: Test time robustness via adaptation and augmentation. In *Advances in Neural Information Processing Systems*, 2022a. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/fc28053a08f59fccb48b11f2e31e81c7-Abstract-Conference.html.

Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-Adapter: Training-free adaption of CLIP for few-shot classification. In *European Conference on Computer Vision*, 2022b. URL https://arxiv.org/abs/2207.09519.

Ziyang Zhang, Yang Yu, Yucheng Chen, Xulei Yang, and Si Yong Yeo. Medunifier: Unifying vision-and-language pre-training on medical data with vision generation task using discrete visual representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Zhang_MedUnifier_Unifying_Vision-and-Language_Pre-training_on_Medical_Data_with_Vision_Generation_CVPR_2025_paper.html.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. LlamaFactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the*

Association for Computational Linguistics (Volume 3: System Demonstrations), pages 400–410, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-demos.38. URL https://aclanthology.org/2024.acl-demos.38/.

Lihua Zhou, Mao Ye, Shuaifeng Li, Nianxin Li, Xiatian Zhu, Lei Deng, Hongbin Liu, and Zhen Lei. Bayesian test-time adaptation for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Zhou_Bayesian_Test-Time_Adaptation_for_Vision-Language_Models_CVPR_2025_paper.html.

A Extended Related Work

Medical acquisition-quality shift. Medical acquisition-quality shift is usually treated as image-distribution heterogeneity introduced by scanner hardware, acquisition protocols, reconstruction pipelines, dose levels, or site differences (Guan and Liu, 2022; Mali et al., 2021). Existing methods mainly differ in which part of the imaging pipeline they modify. Feature harmonization corrects site effects after image features or quantitative measurements are extracted, while image harmonization reduces visual mismatch before prediction (Fortin et al., 2018; Orlhac et al., 2022; Guan et al., 2021). Among these works, Contrastive Image Adaptation is particularly relevant. It explicitly targets acquisition-shift reduction, but its correction remains image-side rather than model-level (Hognon et al., 2024). Medical domain adaptation moves closer to model-level correction, but it mainly targets image-only segmentation or diagnosis models through adversarial alignment, domain transfer, or disentangled representation learning (Kamnitsas et al., 2017; Dou et al., 2018; Yang et al., 2019). Modality-specific correction can encode acquisition physics, such as MRI noise, k-space undersampling, or low-dose CT artifacts, but it usually requires a predefined degradation model (Gudbjartsson and Patz, 1995; Zbontar et al., 2018; Chen et al., 2017). ASSET instead performs acquisition-quality correction by adapting the deployed medical VLM itself at test time, rather than by preprocessing images, harmonizing extracted measurements, or adapting image-only predictors.

Robustness evaluation and robust training. Robustness benchmarks make deployment failures visible under controlled distribution shifts. ImageNet-C and ImageNet-P standardize the evaluation of common corruptions and perturbations (Hendrycks and Dietterich, 2019). Newer VLM benchmarks examine multimodal corruptions, web-scale OOD corruptions, and medical artifacts (Li et al., 2025; Xia et al., 2024; Cheng et al., 2025; Imam et al., 2025). A related benchmark studies this issue for pretrained VLM adaptation methods, showing that stronger adaptation does not necessarily yield stronger performance under visual or textual corruptions (Chen et al., 2023). A separate line tries to build robustness before deployment. Fang et al. (2022) show that CLIP's distributional robustness is largely shaped by the diversity of its pretraining data. CLIPood improves OOD behavior during downstream fine-tuning through a new objective and temporal ensemble (Shu et al., 2023). Anchor-based robust fine-tuning and Robust CLIP further preserve or improve robustness with semantic anchors or adversarial vision-encoder fine-tuning (Han et al., 2024; Schlarmann et al., 2024). These works either evaluate robustness under controlled distribution shifts or build robustness into a model before deployment. ASSET instead adapts an already deployed medical VLM from a small target support set. Since the acquisition shift cannot be fully characterized by a predefined degradation model, the correction signal must be inferred from the support examples.

Test-time adaptation for vision-language models. Test-time adaptation updates a deployed model using target data available at inference. Conventional methods use self-supervised signals such as entropy minimization, augmentation consistency, or continual self-training (Wang et al., 2021; Zhang et al., 2022a; Wang et al., 2022; Yang et al., 2024). In VLMs, many methods adapt the text prompt or the prediction rule at inference time. Prompt-based methods tune text embeddings or prompt buffers, while cache and aggregation methods refine predictions with stored features, augmented views, or pseudo-label consensus (Shu et al., 2022; Yoon et al., 2024; Xiao et al., 2025; Karmanov et al., 2024; Farina et al., 2024; Cao et al., 2025). Recent test-time scaling methods begin to address generative VLMs by aggregating token predictions or adapting from consensus pseudo-labels (Kaya et al., 2026). However, these methods still rely on target predictions, consistency, or pseudo-labels as the adaptation signal. Under acquisition-quality shift, this signal can be ambiguous: optimizing on a small target support set can match the support labels while also capturing support-specific artifacts, or

it may move the model away from its no-shift behavior. ASSET therefore estimates which parameter directions are sensitive to acquisition-quality changes and restricts supervised test-time adaptation to that subspace.

B ASSET Algorithm

Algorithm 1 ASSET: Acquisition-Sensitive Subspace Estimation

Require: Source-adapted model f_{θ_0}, target support S_t, optional reference set S_s, proxy bank $\mathcal{T}$, rounds R, gradient steps T, view-batch size n_v, energy threshold ρ, rank cap $k_{\max}$

Initialize $\theta^{(0)} \leftarrow \theta_0$
for $r = 0, \ldots, R-1$ **do**
  Sample view batches $\mathcal{V}_i^{(r)} \subset \mathcal{T}$ of size n_v and construct $u_i^{t,\tau}$
  Compute $c_i^{(r)}$ and $s_i^{(r)} = \bar{d}_{\theta^{(r)}}(u_i^t; \mathcal{V}_i^{(r)})$ using Eqs. (1) and (12)
  Initialize or smooth scores using Eq. (13) and compute $w_i^{(r)}$ using Eq. (14)
  for each adaptable module m **do**
    Compute sensitivity gradients $g_{i,m}^{(r)}$ using Eq. (2)
    Form the weighted stack $G_m^{(r)}$ using Eq. (15)
    Extract $B_m^{(r)}$ from the leading eigendirections in Eqs. (4)–(5)
  end for
  Parameterize the module-wise update $\Delta^{(r)}(\alpha)$ using Eq. (6)
  Initialize $\alpha^{(r)} \leftarrow 0$; update $\alpha^{(r)}$ for T steps on Eq. (11)
  Update $\theta^{(r+1)} \leftarrow \theta^{(r)} + \Delta^{(r)}(\alpha^{(r)})$
end for

Ensure: Adapted model $f_{\theta^{(R)}}$

C Implementation Details

Source adaptation. For each backbone and dataset, we train a source-adapted LoRA model on the no-shift source split using LLaMA-Factory (Zheng et al., 2024). LoRA uses rank 16, scaling alpha 32, and `lora_target=all`, while the vision tower and multimodal projector are frozen. We train for 2 epochs with learning rate 10^{-4}, cosine scheduling, warmup ratio 0.1, cutoff length 2048, bfloat16 precision, and gradient checkpointing. InternVL3.5-8B uses per-device batch size 2 with gradient accumulation 8; MedGemma-4B-IT uses per-device batch size 4 with gradient accumulation 4. Both settings give effective batch size 16. The resulting epoch-2 adapter is used for all medical acquisition-shift experiments. Data sampling and training seeds are fixed to 2026.

ASSET adaptation. ASSET uses up to 32 labeled target support examples and 32 no-shift reference examples. Subspace estimation samples 2 proxy views per target support example, and constrained adaptation samples 2 proxy views for the consistency objective. We run 5 estimate-and-update rounds, with 8 coefficient-optimization steps per round. The target and reference batch sizes are 2 and 4. The coefficient optimizer uses learning rate 10^{-3}, while LoRA parameters remain constrained to the estimated subspace. We use the JS divergence for proxy-view sensitivity, consistency, and preservation. The eigendirection energy threshold is $\rho = 0.95$, the per-module rank cap is $k_{\max} = 16$, and the module energy floor is 10^{-8}. For influence weighting, we set $\beta = 0.7$ and $\gamma = 1$. The objective weights are $\lambda_{\text{cons}} = \lambda_{\text{src}} = \lambda_{\text{pres}} = 1$, with coefficient regularization $\lambda_\alpha = 10^{-3}$.

For closed-ended VQA evaluation and training losses, the model is scored over the candidate answer set. The reliability functional C used for influence weighting is instantiated as the candidate-score margin

$$C(F_\theta(u), y) = s_\theta(u, y) - \max_{y' \neq y} s_\theta(u, y'), \tag{17}$$

where $s_\theta(u, y)$ is the model score for candidate answer y. For single-token candidates this is the corresponding next-token logit; for multi-token candidates it is the summed candidate sequence score.

Table 4: Severity grids for the controlled medical acquisition-quality shifts.

Family	Modality	Severity parameter	Values
CT-Gaussian	CT	image-noise std.	60, 120, 180, 240
CT-Poisson	CT	dose fraction	0.50, 0.25, 0.10, 0.05
MRI-Rician	MRI	Rician noise σ	0.05, 0.10, 0.15, 0.20, 0.25
MRI-k-space	MRI	acceleration factor R	2, 4, 6, 8
VAE	mixed CT/MRI	latent-noise σ	1, 2, 3, 4, 5

Proxy banks. The proxy bank is designed to span plausible acquisition-quality mechanisms rather than to match a single benchmark degradation. General transformations cover common image-quality axes that can arise across imaging pipelines, including resolution, blur, contrast, stochastic noise, and compression. Modality-informed transformations cover coarse CT/MRI acquisition and reconstruction effects that are commonly modeled in medical imaging, such as dose/noise variation, reconstruction smoothness or sharpness, partial-volume effects, Rician noise, ringing, motion, k-space artifacts, and anisotropic resolution.

Concretely, the medical proxy bank contains 5 general image-quality transformations: resolution resampling, low-pass blur, gamma contrast, Gaussian noise, and JPEG compression. It also contains 4 CT-informed transformations: window/level shift, reconstruction blur, reconstruction sharpening, and partial-volume simulation; and 5 MRI-informed transformations: Rician noise, Gibbs ringing, motion ghosting, k-space spike artifact, and anisotropic resolution. CT shift families sample from the general and CT-informed transformations, MRI shift families sample from the general and MRI-informed transformations, and VAE shift families sample from the full bank.

When a target family coincides with a bank member, ASSET removes that identical transformation from the proxy pool and estimates directions from related but non-identical probes. Thus Gaussian-noise proxy views are excluded for CT-Gaussian targets, and Rician-noise proxy views are excluded for MRI-Rician targets. K-space spike proxy views remain available for MRI-k-space targets because they are distinct from the benchmark's k-space undersampling degradation.

D Data and Shift Construction

Medical VQA protocol. The main medical experiments use SLAKE, RadImageNet-VQA, and GMAI-MMBench. We keep closed-ended VQA examples and canonicalize each dataset into a no-shift source split for LoRA training, a no-shift reference pool, a labeled target support pool, and a clean evaluation split. Evaluation always keeps the question and answer fixed and replaces only the image with its degraded counterpart. Modality assignments are taken from dataset metadata. CT-Gaussian and CT-Poisson use CT examples, MRI-Rician and MRI-k-space use MRI examples, and the VAE degradation is modality-agnostic and uses mixed CT/MRI images.

Support and reference pools contain up to 32 examples per experiment. For mixed VAE settings, support and reference examples are balanced between CT and MRI when enough examples are available; otherwise we use all available examples for the smaller subset. RadImageNet-VQA and GMAI-MMBench use 150 CT examples, 150 MRI examples, and 300 mixed examples for VAE evaluation. SLAKE is smaller, so we use all available modality-matched evaluation examples: 186 CT rows from 40 images, 73 MRI rows from 26 images, and 259 mixed rows from 66 images.

Medical acquisition-quality shifts. The controlled medical shifts are applied to both target support and evaluation images. Table 4 summarizes the severity parameter and grid for each family. "Clean" denotes the original no-shift image, and the values in each row are ordered from mild to strong severity.

CT-Gaussian adds zero-mean Gaussian image noise on the 8-bit intensity scale. CT-Poisson simulates low-dose CT from reconstructed images by constructing a pseudo-sinogram with the Radon transform, applying Poisson noise to Beer–Lambert transmission counts, and reconstructing with inverse Radon transform. We use nominal photon count $I_0 = 10^5$ and maximum attenuation 10; smaller dose fraction denotes stronger degradation. MRI-Rician applies the standard magnitude model $\sqrt{(x+n_1)^2+n_2^2}$.

Table 5: Subspace ablation on RadImageNet-VQA with InternVL3.5-8B. Entries are strongest-shift accuracies (%), and Avg. averages the 5 shift families.

Method	CT-Gaussian	CT-Poisson	MRI-Rician	MRI-k-space	VAE	Avg.
LoRA	50.0	84.7	62.0	51.3	36.0	56.8
Random subspace	49.3	84.0	62.0	52.0	36.0	56.7
ASSET w/o subspace	76.0	94.0	67.3	**59.3**	41.3	67.6
Fixed one-shot subspace	76.7	92.7	**70.0**	56.0	37.3	66.5
ASSET	**84.7**	**94.7**	68.7	58.0	**42.3**	**69.7**

Table 6: Objective, weighting, and rank ablation on RadImageNet-VQA with InternVL3.5-8B. Entries are strongest-shift accuracies (%), and Avg. averages the 5 shift families.

Method	CT-Gaussian	CT-Poisson	MRI-Rician	MRI-k-space	VAE	Avg.
w/o $\mathcal{L}_{\text{cons}}$	76.0	90.0	65.3	56.7	40.0	65.6
w/o $\mathcal{L}_{\text{src}}$	81.3	87.3	58.7	60.0	37.7	65.0
w/o $\mathcal{L}_{\text{pres}}$	79.3	93.3	68.0	56.7	41.0	67.7
Uniform weighting	78.0	88.7	**69.3**	56.7	40.3	66.6
$k_{\max} = 2$	76.0	90.0	64.7	58.7	36.0	65.1
$k_{\max} = 4$	82.0	90.7	68.0	**62.7**	35.0	67.7
$k_{\max} = 8$	77.3	91.3	68.0	58.0	37.7	66.5
ASSET	**84.7**	**94.7**	68.7	58.0	**42.3**	**69.7**

MRI-k-space performs 2-D Fourier undersampling, using center-line fractions 0.16, 0.08, 0.06, and 0.04 for acceleration factors $R = 2, 4, 6, 8$, respectively.

The VAE degradation probes learned quality changes beyond the explicit CT/MRI models. We use the pretrained Stable-Diffusion VAE `stabilityai/sd-vae-ft-mse`, implemented as an AutoencoderKL model from the latent diffusion family (Kingma and Welling, 2014; Rombach et al., 2022). Each image is converted to grayscale, resized to 256×256, encoded to the latent mean z, perturbed as $z + \sigma\epsilon$ with $\epsilon \sim \mathcal{N}(0, I)$, decoded, and resized back to the original image size.

Real acquisition-protocol shift. The real-shift experiment uses MIMIC-CXR-VQA, which is built on MIMIC-CXR. We keep clinical yes/no questions and remove questions about the view position. Each image is assigned to the PA source or AP target condition by its annotated view. The source-adapted LoRA model is trained on PA examples. Each of 3 draws samples 32 labeled AP support examples and 32 PA reference examples. Evaluation uses 782 AP and 146 PA questions whose images do not overlap with the support or reference sets. Proxy views are sampled from the full medical proxy bank, which contains no view-position or projection transformation.

General VQA shift benchmark. The broader distribution-shift experiment uses InternVL3.5-8B on the Visual Robustness Benchmark (Ishmam et al., 2025). We use 1000 VQAv2 yes/no source training examples, 300 non-overlapping target evaluation examples, and 32 target support examples from the validation split. The benchmark contains 14 corruptions grouped into 6 families: noise, blur, weather, image property, physical, and digital. Each corruption is evaluated at severities 1–5. For each target corruption, ASSET builds the proxy bank from corruptions outside the target family and samples proxy views at severity 3, so the proxy bank does not include the exact held-out corruption family being evaluated.

E Extended Ablations

Model components. We ablate the components of ASSET with InternVL3.5-8B on RadImageNet-VQA and report the strongest-shift accuracy of each shift family. Table 5 examines the subspace constraint. Random subspace replaces the estimated directions with random ones. ASSET w/o subspace updates all adaptable parameters with the same data and objectives. Fixed one-shot subspace keeps the first estimated subspace across rounds. Table 6 removes each training objective,

replaces influence weighting with uniform weights, and lowers the rank cap $k_{\max}$. Full ASSET achieves the highest average accuracy in both tables, while random directions bring no gain over LoRA.

Support set. We vary the number of labeled support examples with InternVL3.5-8B on GMAI-MMBench under the strongest CT-Gaussian shift. Table 7 shows that ASSET improves over LoRA at every support size and reaches its highest accuracy with 32 examples. We further evaluate 5 random draws of the support and reference sets on SLAKE and GMAI-MMBench. In every draw, ASSET improves both average shifted accuracy and strongest-shift accuracy over LoRA.

Table 7: Support-size ablation on GMAI-MMBench with InternVL3.5-8B. Entries are strongest-shift accuracies (%) under CT-Gaussian shift, and columns give the number of labeled support examples.

Method	4	8	16	32
LoRA	25.3	25.3	25.3	25.3
ASSET	**26.0**	**28.0**	**27.3**	**29.3**

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes].

   Justification: See Sec. 1.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes].

   Justification: See Sec. 6.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.

- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A].

   Justification: This paper focuses on empirical analysis to validate the effectiveness of the proposed method, and does not include formal theoretical assumptions or proofs. See Sec. 5.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes].

   Justification: See Secs. 4 and 5, and Appendices C and D.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes].

   Justification: The code and data preparation scripts are available at `https://github.com/UNITES-Lab/ASSET`.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes].

   Justification: See Sec. 5 and Appendices C and D.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.

- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes].

   Justification: The paper reports aggregate results across multiple models, datasets, shift families, and severity levels. See Sec. 5.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes].

   Justification: Experiments were run on an internal Slurm GPU cluster using NVIDIA A100 and NVIDIA RTX PRO 6000 GPUs, with jobs using up to 8 GPUs in parallel.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes].

   Justification: The work uses existing research datasets and models, reports controlled benchmark evaluations, and does not collect new human-subject data.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes].

    Justification: The paper studies robustness of medical VLMs under acquisition-quality shift, which can improve reliability, while the conclusion notes that validation on real acquisitions from more clinical sites remains future work.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A].

    Justification: The paper does not release a new dataset, pretrained model, or generated image model.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes].

    Justification: Existing datasets, models, and tools are credited through citation and used under their public or research-use terms.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A].

    Justification: The paper does not release new datasets or models.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A].

    Justification: The paper does not involve crowdsourcing or new human-subject experiments.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A].

    Justification: The paper uses existing datasets and does not collect new human-subject data.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes].

    Justification: The paper studies adaptation of vision-language models and specifies the VLM backbones in Sec. 5.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.