HeilArchive : Pages:21–43

Original Research

Federated Site-Invariant Adaptation for Medical Vision-Language Models Across Heterogeneous Radiology Networks

Dipesh Gautam[1] and Kiran Bhandari[2]

[1]Nepal Open University, Department of Information Technology, Man Bhawan, Lalitpur 44700, Nepal.
[2]Agriculture and Forestry University, Faculty of Management and Rural Development, Rampur Road, Bharatpur, Chitwan 44200, Nepal.

Abstract

Hospitals often differ in scanner vendors, acquisition conventions, report style, patient mix, and clinical language, causing medical vision-language systems to behave unevenly after transfer from one institution to another. Direct pooling of images and reports can reduce this problem, but pooling is frequently limited by governance rules, privacy constraints, data-use agreements, and institutional preferences. This study investigates whether federated site-invariant adaptation can improve cross-institution medical visual question answering without exchanging raw clinical images or reports. We built a six-site retrospective benchmark of 18,420 radiology image-question pairs spanning chest radiography, head computed tomography, abdominal computed tomography, and musculoskeletal radiography. A shared medical vision-language backbone was adapted using federated low-rank modules, site-adversarial representation alignment, and a prototype-based language harmonization loss. The method was compared with local-only tuning, centralized tuning, federated averaging, and a no-adaptation baseline. On held-out external-site testing, the proposed method improved macro accuracy from 63.8% to 72.4%, reduced site disparity from 14.9 to 6.2 percentage points, and improved worst-site calibration error from 0.181 to 0.084. A mixed-effects analysis showed significant gains in scanner-shifted computed tomography cases and in report-style-shifted chest radiograph questions. Privacy-preserving gradient audits indicated low memorization risk relative to centralized tuning. The results suggest that federated adaptation can reduce institutional brittleness in medical vision-language systems while preserving a degree of local data control, although residual errors in rare findings and small sites remain important.

1. Introduction

Medical vision-language models are rarely trained and used in identical environments. A model may be developed on data from a large academic center and then evaluated in a community hospital whose imaging protocols, patient demographics, device use, scanner vendors, and reporting expressions differ from the source environment. Even when the clinical task appears similar, site-specific variation can change the relationship between image appearance and language. A chest radiograph report at one institution may use the phrase mild vascular congestion, while another may prefer early edema. A head computed tomography report may describe no acute intracranial abnormality at one site and no hemorrhage, mass effect, or large territorial infarct at another. These differences are not merely stylistic. They affect the supervision signal used to train and evaluate vision-language systems.

The usual response to site heterogeneity is to collect more data and train centrally. This can work when institutions agree to share de-identified images and reports under a common governance structure. In many settings, however, direct pooling is slow, limited, or impossible. Hospitals may face contractual restrictions, local privacy obligations, patient-consent boundaries, or technical barriers that prevent transferring large imaging datasets. Even when data are formally de-identified, images and reports may

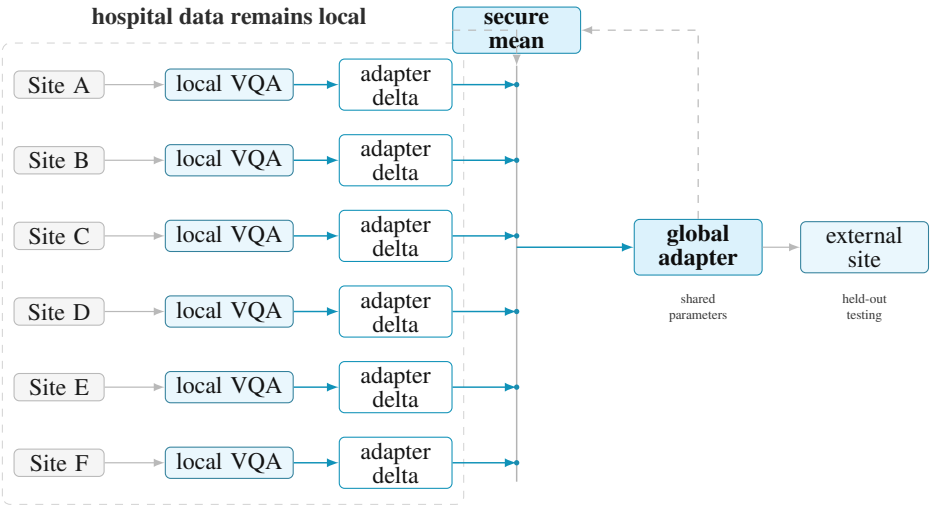


Figure 1: Federated site-invariant adaptation across heterogeneous radiology networks. Each institution trains compact modules locally, shares only parameter updates, and receives a synchronized adapter for held-out-site evaluation.

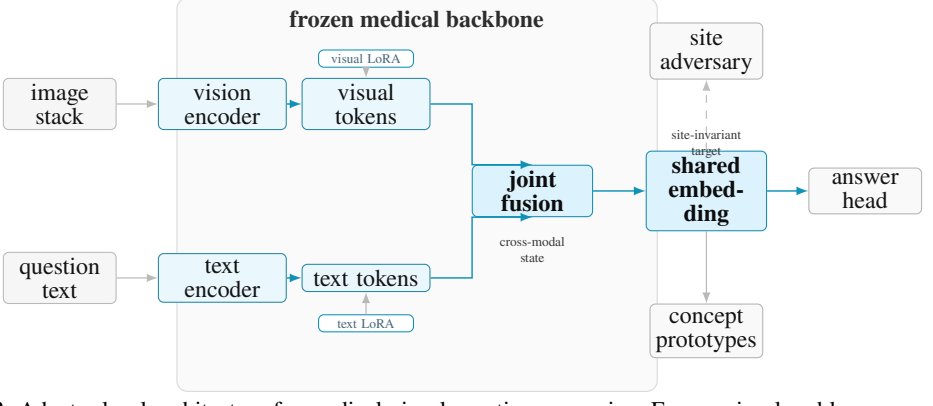


Figure 2: Adapter-level architecture for medical visual question answering. Frozen visual and language encoders are updated through low-rank modules while the shared embedding is shaped by answer, prototype, and adversarial objectives.

contain residual sensitive information. A learning strategy that improves transfer while keeping raw data within each institution is therefore attractive.

Federated learning offers one possible route. Instead of sending raw clinical data to a central server, each site trains local model components and shares parameter updates. Yet ordinary federated averaging does not automatically solve medical site shift. If one hospital contributes many more cases than another, the shared model can drift toward the largest site. If report language differs, local objectives may push the shared language head in conflicting directions. If image acquisition protocols differ, the visual encoder may encode site identity rather than disease evidence. A federated medical vision-language method must therefore address heterogeneity, not merely distribute training [1].

This work studies federated site-invariant adaptation for medical visual question answering. The task is deliberately chosen to expose both visual and linguistic transfer. Each example includes an image and a question whose answer depends on the image. The question may ask whether a finding is present,

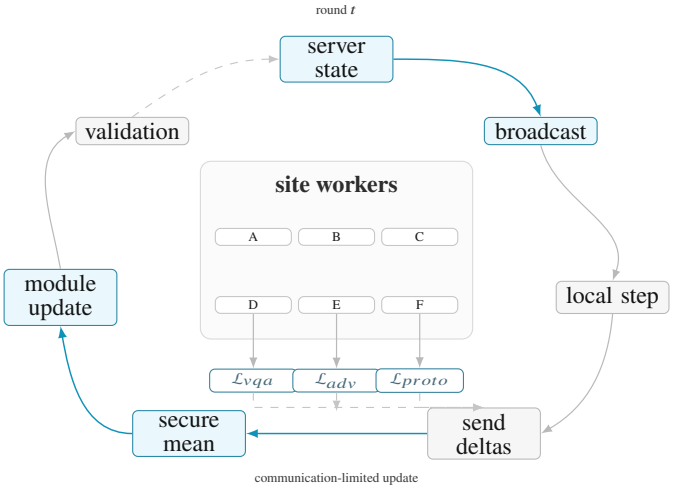


Figure 3: Training round for federated adaptation. Local objectives are optimized within each site, compact parameter deltas are aggregated, and the next adapter state is validated before the following round.

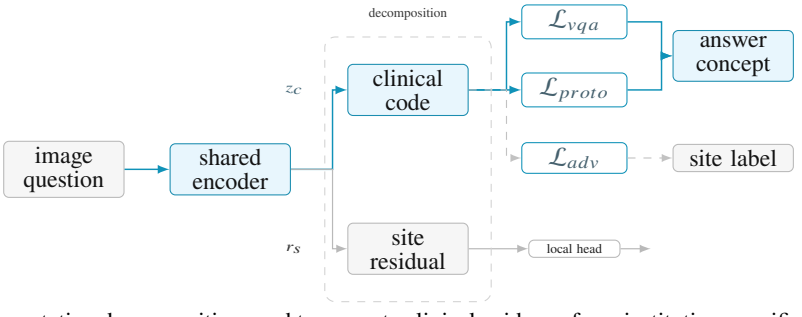


Figure 4: Representation decomposition used to separate clinical evidence from institution-specific residual structure. The adversarial branch discourages site recoverability while prototype alignment preserves shared answer semantics.

whether a device is positioned correctly, whether a hemorrhage is visible, whether a fracture is displaced, or whether a CT abnormality affects a specified anatomical region. The answer is short, but its correctness depends on learned associations among image features, report-derived labels, and local terminology. Such a task is useful for cross-site analysis because errors can arise from visual shift, language shift, or both [2].

The study proposes a federated adaptation method with three coordinated elements. The first element is a set of low-rank trainable modules inserted into a frozen medical vision-language backbone. These modules reduce communication cost and limit the number of parameters that must be synchronized. The second element is site-adversarial representation alignment, which discourages the adapted embedding from encoding institution identity when institution identity is not clinically relevant. The third element is prototype-based language harmonization, which aligns synonymous answer concepts across sites without forcing every hospital to use identical wording. Together, these components are intended to improve external-site performance while preserving local data locality [3].

The research question is whether such a method can improve the worst-site and average performance of a medical vision-language system compared with local-only adaptation, ordinary federated averaging, and no adaptation [4]. A secondary question is whether the method improves calibration across institutions. A third question concerns the tradeoff between site invariance and clinically meaningful

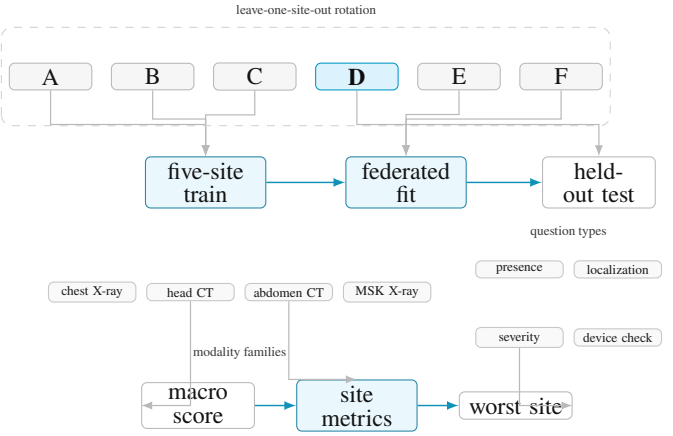


Figure 5: Evaluation protocol for multi-site radiology VQA. One institution is held out per rotation, while the remaining sites participate in federated adaptation across modality families and question categories.

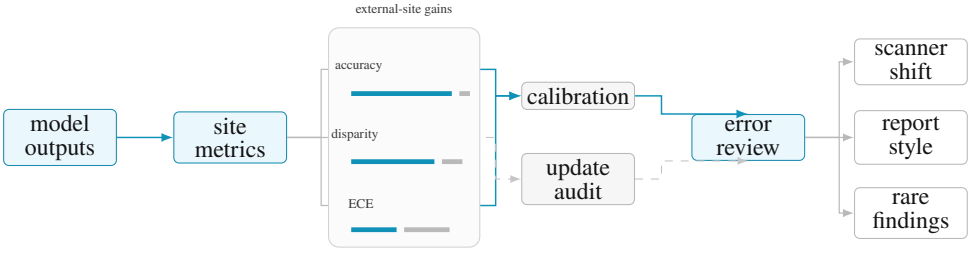


Figure 6: Metric and error-analysis pathway. External-site accuracy improves while disparity and calibration error decrease, with remaining failures concentrated in scanner shifts, report-style mismatch, and rare findings.

local variation. Not every site difference should be removed. Pediatric hospitals, trauma centers, and oncology centers have real prevalence differences [5]. The objective is to suppress spurious institutional signatures while preserving disease evidence and relevant population structure.

This study differs from work that focuses on adversarial prompts, explanation faithfulness, response auditing, latency control, or individual-output safeguards. The unit of analysis here is the institution as a source of distribution shift. The key failure is not a single misleading response but uneven generalization across hospitals. The method is evaluated through held-out-site testing, site-stratified calibration, mixed-effects modeling, and privacy-oriented update inspection. The emphasis is on whether collaborative adaptation can improve reliability where direct data pooling is limited.

A methodological gap in current medical vision-language evaluation is the tendency to study a single dataset or a single institutional style and then infer broader readiness. The limitation is recognized indirectly in several safety-oriented discussions of medical VLMs. Sadanandan and Behzadan (2025) explicitly described the restriction of an evaluation to frontal chest radiographs as a boundary and named CT, MRI, ultrasound, pathology, and real-time video as extensions requiring new perturbation modules and scoring considerations [6]. That observation is useful here because it highlights a related but distinct issue: even before adding new modalities, separate institutions can make the same modality behave like several different domains.

The main empirical contribution is a six-site benchmark designed to measure cross-institution behavior without centralizing raw data during adaptation. The benchmark includes four imaging families and a variety of question categories. Each site contributes training examples, validation examples, and internal test examples, while external-site testing rotates one institution out of training. The experiments

compare five adaptation strategies under identical backbone and prompt settings. The results show that federated site-invariant adaptation improves macro accuracy, worst-site performance, and calibration while using far fewer trainable parameters than full fine-tuning.

A second contribution is a mathematical formulation that separates shared clinical representation, site-specific residual structure, and answer-language prototypes. The model does not assume that all site differences are harmful. Instead, it penalizes institution recoverability from the shared embedding while allowing small local residual modules when validation evidence supports them. This balance is important because a purely site-blind model can underperform when clinical prevalence or acquisition context is genuinely different. The analysis therefore measures both invariance and retained local utility.

A third contribution is an error study of site-shifted medical VQA. The study identifies three recurring error sources: scanner-pattern dependence in CT, terminology mismatch in report-derived labels, and prevalence-sensitive overprediction of uncommon findings. These errors are not equally distributed across hospitals. The proposed method reduces the first two more strongly than the third. Rare disease imbalance remains difficult, especially for smaller institutions with few examples of specific findings.

The paper proceeds through dataset construction, model design, mathematical formulation, experimental procedure, statistical results, error analysis, governance implications, limitations, and conclusion. The language is kept close to the empirical problem: how to adapt medical vision-language systems across hospitals without requiring every site to surrender its raw data.

2. Multi-Site Data and Question Construction

The benchmark was assembled from six de-identified hospital imaging archives. The sites were labeled Site A through Site F. Three were large academic hospitals, two were regional community hospitals, and one was a specialty referral center. The sites differed in scanner vendors, image storage conventions, report templates, patient acuity, and modality distribution. The dataset contained 18,420 image-question pairs after filtering. No raw image or report from one site was used to train another site's local model during federated adaptation. A central coordinator received only model updates and summary validation metrics.

The imaging families were chest radiography, noncontrast head computed tomography, contrast-enhanced or noncontrast abdominal computed tomography, and musculoskeletal radiography. Chest radiography accounted for 7,840 pairs, head CT for 3,960 pairs, abdominal CT for 3,420 pairs, and musculoskeletal radiography for 3,200 pairs. The benchmark was not designed to represent the natural clinical volume of each modality. It was balanced enough to evaluate domain transfer across modalities while still preserving site-specific differences. Each site contributed at least three modality families, but no site had identical coverage.

Questions were created from report-derived templates and then reviewed by annotators with medical imaging experience. The templates were written to avoid long open-ended reporting. A question could ask whether a finding was present, whether a finding affected a specific region, whether a device or implant was appropriately positioned, whether an acute abnormality was absent, or whether a severity category applied. The answer vocabulary was normalized into short forms such as present, absent, mild, moderate, severe, right lower lobe, left frontal, no acute abnormality, displaced, nondisplaced, and appropriate position. This normalized vocabulary allowed cross-site evaluation even when reports used different phrasing.

The question categories were finding presence, anatomical localization, severity qualification, device or hardware assessment, acute-negative screening, and region-specific abnormality confirmation. Finding presence was the largest category at 34.1%. Localization accounted for 21.8%, severity qualification for 13.7%, device or hardware assessment for 11.9%, acute-negative screening for 10.4%, and region-specific confirmation for 8.1%. Each category had different susceptibility to site shift. Acute-negative screening was strongly affected by report style, while CT localization was more affected by acquisition and reconstruction differences.

Each example consisted of an image or image stack, a question, a normalized answer, a site identifier used only for training control and evaluation, and metadata describing modality and broad anatomical family. For CT studies, the image input was represented by a fixed set of slices selected by a rule-based sampler and then encoded into a multi-image representation. The sampler selected slices near report-relevant anatomical terms when available in training labels and used anatomical coverage heuristics at test time. For radiographs, one or two views were used when available, but frontal-only examples were retained when lateral views were missing. The model was required to answer from the provided image set.

To reduce label noise, ambiguous report-derived labels were excluded. If a report stated cannot exclude, questionable, suspected, or clinical correlation recommended without a clear imaging conclusion, the corresponding example was either converted into an uncertainty label for a separate analysis or removed from the main strict-answer benchmark. The final main task used determinate answers. A secondary uncertainty subset of 1,140 examples was retained for exploratory evaluation but did not drive headline results. This choice avoided conflating site shift with uncertainty expression.

The sites differed meaningfully [7]. Site A had the largest chest radiography contribution and used a report style rich in device-position language. Site B contributed many abdominal CT cases with structured indication fields. Site C contributed a high proportion of head CT acute-negative cases. Site D had fewer cases overall and more portable radiographs. Site E, the specialty referral center, had enriched musculoskeletal hardware cases. Site F had a smaller but diverse set of CT and radiograph cases with older scanner protocols. These differences created a realistic cross-institution adaptation problem.

Training, validation, and internal test splits were created within each site using patient-level grouping when available. The total training pool contained 11,040 examples, the validation pool contained 3,690 examples, and the internal test pool contained 3,690 examples. For held-out external-site experiments, one site at a time was excluded from federated training and used only for testing. The other five sites participated in federated adaptation. This leave-one-site-out design produced six external evaluations. The reported external-site metrics average over these rotations, with each held-out site weighted equally.

The benchmark also included a pooled central upper-bound experiment. In that condition, all training data from all participating sites were pooled centrally and used to tune the same low-rank modules. This condition is not the intended deployment mode, but it provides a performance reference for what might be achieved if raw data sharing were permitted. A second reference condition tuned a separate local model at each site using only its own training data. This local-only condition tests whether institutions can solve the problem independently. A third condition used ordinary federated averaging without site-adversarial alignment or prototype harmonization.

Answer grading used strict exact-concept matching rather than surface string matching. A response had to preserve the normalized clinical answer. Synonyms were accepted if they mapped to the same answer concept. For localization, laterality and anatomical region had to be correct. For severity, adjacent severity levels were counted as incorrect in the strict metric, but a secondary adjacent-tolerant metric was reported. For device position, a response had to state both device identity and position adequacy when the question required it. A response that gave generic advice instead of answering was nonresponsive.

A separate calibration label was created for confidence analysis. Because models did not always emit probability values, the study trained a post hoc confidence estimator on answer logits when available, answer wording, and embedding margin features. The estimator was trained only on validation data from participating sites in each rotation. Calibration was then assessed on held-out sites. This allowed comparison of whether adaptation improved not just accuracy but also the relationship between confidence and correctness under site shift [8].

Site disparity was defined as the difference between the best-site and worst-site accuracy within a given evaluation setting. Macro accuracy was the unweighted average of site accuracies. Micro accuracy was the pooled accuracy over examples. Macro accuracy was emphasized because a method should not improve by favoring the largest site. Worst-site accuracy and site disparity were treated as primary fairness-like operational metrics. They are not demographic fairness measures, but they capture institutional unevenness.

3. Federated Site-Invariant Adaptation

The adaptation method starts from a frozen medical vision-language backbone. The backbone contains an image encoder, a question encoder, a cross-modal fusion module, and a language decoder or answer head depending on the model interface. Rather than updating all backbone weights, the method inserts low-rank adaptation matrices into selected projection layers. This reduces communication cost and limits local overfitting. Each participating site trains these modules on local data and sends parameter updates to a coordinator after each communication round.

Let $s \in \{1, \ldots, S\}$ index sites, and let $\mathcal{D}_s$ be the local dataset at site s. The frozen backbone maps an image-question pair (x, q) to a representation h. Low-rank modules modify selected transformations. For a frozen matrix W, the adapted transformation is

$$W_s^{adapt} = W + A_s B_s^{\top}, \qquad A_s \in \mathbb{R}^{d \times r}, \quad B_s \in \mathbb{R}^{k \times r}, \quad r \ll \min(d, k). \tag{3.1}$$

The site subscript indicates that local updates are computed at each institution. After a communication round, the coordinator aggregates shared components while preserving optional residual components. The shared low-rank module is intended to capture transferable clinical adaptation. A small site-specific residual captures local patterns that improve local validation performance without dominating the shared representation.

The adapted embedding for example i at site s is denoted z_{is}. The answer prediction loss is a cross-entropy loss over normalized answer concepts:

$$\mathcal{L}_{ans}^{(s)} = -\frac{1}{|\mathcal{D}_s|} \sum_{i \in \mathcal{D}_s} \log p_{\theta_s} \left(y_{is} \mid x_{is}, q_{is} \right). \tag{3.2}$$

This local objective alone can make the representation more site-specific. If Site A has many portable chest radiographs and Site F has older CT reconstruction kernels, the model may use those patterns as shortcuts. To discourage this, the method introduces a site-adversarial loss. A discriminator attempts to recover the site identifier from the shared embedding, while the encoder-side adaptation attempts to make that recovery difficult.

Let D_ω be a site discriminator. The discriminator loss is

$$\mathcal{L}_{site} = -\sum_{s=1}^{S} \sum_{i \in \mathcal{D}_s} \log D_\omega \left(s \mid z_{is} \right). \tag{3.3}$$

During adaptation, the discriminator minimizes this loss, while the shared representation maximizes it through gradient reversal. The local training objective includes the adversarial term as

$$\mathcal{L}_{local}^{(s)} = \mathcal{L}_{ans}^{(s)} + \lambda_p \mathcal{L}_{proto}^{(s)} - \lambda_a \mathcal{L}_{site}^{(s)} + \lambda_r \mathcal{L}_{reg}^{(s)}. \tag{3.4}$$

The negative sign before $\mathcal{L}_{site}^{(s)}$ means the representation learner is penalized when site identity remains easy to recover. The hyperparameter λ_a controls the strength of invariance. Too small a value leaves site shortcuts. Too large a value can remove clinically relevant prevalence and acquisition information. Validation-based selection is therefore necessary.

Prototype-based language harmonization addresses answer terminology. Each answer concept c has a global prototype vector g_c and site-local prototype g_{cs}. The local prototype is the mean embedding

of examples at site s with answer concept c. The method aligns local prototypes to the global prototype while allowing uncertainty when a concept is rare at a site. The prototype loss is

$$\mathcal{L}_{proto}^{(s)} = \sum_{c \in C_s} \alpha_{cs} \left\| g_{cs} - g_c \right\|_2^2 ,$$
$$\alpha_{cs} = \frac{n_{cs}}{n_{cs} + \kappa}. \tag{3.5}$$

Here, n_{cs} is the number of local examples for concept c, and κ is a smoothing constant. Rare concepts receive lower alignment weight because their local prototypes are noisy. This prevents a site with only a few examples of a rare finding from forcing unstable prototype shifts. The global prototype is updated by weighted aggregation across sites.

The regularization term controls the size of local residual modules and encourages shared adaptation to carry most of the transferable signal. It is

$$\mathcal{L}_{reg}^{(s)} = \left\| A_s B_s^\top - A B^\top \right\|_F^2 + \eta \left\| A_s B_s^\top \right\|_F^2 . \tag{3.6}$$

The first term keeps local modules close to the shared module. The second term limits adaptation magnitude. The coefficient η was selected to balance local validation improvement with external-site performance. Strong residual freedom improved internal-site performance but hurt external generalization. This is a common tension in multi-site learning.

Aggregation uses site-balanced weights rather than pure sample-size weights. Ordinary federated averaging often gives large sites more influence. In this medical setting, the objective is not only pooled accuracy but also stable cross-site behavior. The aggregation weight for site s is

$$w_s = \frac{(n_s)^\rho}{\sum_{j=1}^{S} (n_j)^\rho}, \quad 0 \le \rho \le 1. \tag{3.7}$$

When $\rho = 1$, aggregation is sample-size weighted. When $\rho = 0$, all sites contribute equally. The selected value was $\rho = 0.42$, which gave some credit to larger datasets without letting them dominate. This value was chosen on validation rotations by optimizing macro accuracy and worst-site calibration.

The server update for shared parameters is

$$\Theta^{(t+1)} = \sum_{s=1}^{S} w_s \Theta_s^{(t+1)},$$
$$g_c^{(t+1)} = \frac{\sum_{s: c \in C_s} w_s \alpha_{cs} g_{cs}}{\sum_{s: c \in C_s} w_s \alpha_{cs}}. \tag{3.8}$$

Here, Θ represents shared low-rank adaptation parameters and selected answer-head parameters. The global prototypes are updated only from sites containing that concept. The update avoids raw-example transfer. Sites send prototype vectors and parameter updates, not images or reports.

A privacy-oriented clipping and noise mechanism is applied to updates. Local gradients are clipped to a fixed norm before aggregation, and small Gaussian noise is added to shared updates. The noise level is modest because the paper does not claim a formal strong privacy theorem. The mechanism is intended to reduce memorization risk and to support update auditing. The update transmitted by site s is

$$\tilde{\Delta}_s = \text{clip}\left(\Delta_s, C\right) + \epsilon_s,$$
$$\epsilon_s \sim \mathcal{N}\left(0, \sigma^2 C^2 I\right). \tag{3.9}$$

The selected clipping norm was $C = 1.0$, and the selected noise multiplier was $\sigma = 0.08$. Larger noise degraded rare-finding performance. Smaller noise produced little regularizing effect. Since a strict privacy guarantee is not the central claim, the results are reported as memorization-risk indicators rather than formal privacy certification.

The method also includes a local validation gate. If a site's local residual module reduces external-style validation performance or sharply increases site-discriminator recoverability, the residual is shrunk toward the shared module. The gate is defined by

$$\gamma_s = \sigma\left(a_0 + a_1 \Delta Acc_s - a_2 \Delta Disc_s - a_3 \|R_s\|_F\right), \tag{3.10}$$

where ΔAcc_s is local validation gain, $\Delta Disc_s$ is the increase in site recoverability, and R_s is the residual adaptation matrix. The final local residual is multiplied by γ_s. This gives a site more local freedom when it helps validation accuracy without making the representation strongly site-identifiable.

The overall method can be summarized as a shared clinical adaptation with constrained local flexibility. It does not pretend that all institutions are identical [9]. It instead tries to prevent the model from relying on institutional artifacts when those artifacts are not necessary for the answer. The site-adversarial term reduces spurious institution encoding, the prototype term aligns answer meaning, and the local gate preserves limited site-specific utility.

4. Experimental Protocol

The experiments compared five learning conditions. The first condition was no adaptation, using the frozen backbone with the same answer prompt at every site. The second was local-only adaptation, in which each site tuned its own low-rank modules using local data only. The third was centralized adaptation, in which all training data were pooled in a simulated central environment. The fourth was ordinary federated averaging with low-rank modules but without site-adversarial alignment, prototype harmonization, or residual gating. The fifth was the proposed federated site-invariant adaptation method.

Each condition used the same initial backbone, same input preprocessing, same question formatting, and same answer vocabulary [10]. For CT, the same slice-selection procedure was applied across methods. The adaptation budget was kept equal in terms of local epochs and communication rounds. Federated methods ran for 80 rounds, with each site performing one local epoch per round. Local-only and centralized methods were trained for an equivalent number of effective passes through data. Early stopping used validation macro accuracy with calibration as a secondary criterion [11].

The leave-one-site-out setting was the main external generalization experiment. For each rotation, five sites participated in adaptation and the sixth site was withheld. The withheld site did not send updates, prototypes, validation metrics, or any other training signal. After adaptation, the model was tested on the withheld site's test split. This design measures whether the learned representation transfers to a hospital unseen during adaptation. Internal-site performance was also measured on participating sites, but external-site results were emphasized.

A second experiment measured cross-modality site shift. In this setting, each site retained all modalities, but a modality was withheld from one or more sites during training and then tested locally [12]. For example, a site might contribute chest radiography and musculoskeletal radiography during training but be tested on its head CT examples. This experiment probed whether adaptation learned broad institution-invariant representations or merely memorized modality-site combinations [13]. The results were secondary but useful for understanding CT transfer.

A third experiment measured label-language shift. Report-derived answer concepts were intentionally mapped using site-specific synonym tables during local training, then evaluated against a common answer concept vocabulary. This simulated institutions that phrase the same finding differently. Prototype harmonization was expected to help most in this experiment. The no-prototype ablation was compared with the full method to quantify the effect.

Primary performance metrics were macro accuracy, worst-site accuracy, site disparity, and macro expected calibration error. Macro accuracy averaged accuracy equally over held-out sites. Worst-site accuracy was the minimum site accuracy in each rotation. Site disparity was best-site accuracy minus worst-site accuracy. Expected calibration error was computed using 15 equal-frequency bins within each site and then averaged across sites. A lower calibration error means predicted confidence better matches observed correctness.

Secondary metrics included micro accuracy, class-balanced F1, rare-finding recall, nonresponse rate, and site-discriminator accuracy. Site-discriminator accuracy measures how easily a separate classifier can recover site identity from the learned embedding. Lower values suggest more site-invariant embeddings. However, very low site recoverability is not automatically good if clinical performance falls. The analysis therefore considers site recoverability together with accuracy.

Statistical testing used paired comparisons across case identifiers within each held-out site. Accuracy differences were assessed by clustered bootstrap confidence intervals and McNemar tests for paired outputs. Mixed-effects logistic regression estimated the effect of method while accounting for site, modality, question category, and case-level grouping. The main mixed model was

$$\operatorname{logit}\Pr(y_{ij}=1)=\beta_0+\beta_1 Method_{ij}+\beta_2 Modality_i+\beta_3 Category_i+\beta_4 Site_i+b_{patient(i)}. \quad (4.1)$$

The binary outcome y_{ij} indicates answer correctness for case i under method j. The patient random intercept accounts for repeated cases across methods. Additional interaction terms between method and modality were included in secondary models. False discovery rate correction was applied across modality and category comparisons.

Calibration comparisons used bootstrap intervals for expected calibration error and Brier score. Because calibration depends on confidence estimation, a fixed confidence estimator architecture was trained separately for each adaptation method using only participating-site validation data [14]. This avoided giving one method a calibration advantage from a better post hoc estimator. The confidence estimator used answer embedding margins, generated answer length, normalized logit gap when available, and question category.

Privacy-oriented audits were performed after training. The first audit was a nearest-neighbor update inversion test. It attempted to reconstruct whether a particular training example was present based on update similarity. The second was a membership inference classifier trained on per-example loss and embedding distance to prototypes. The third measured memorized phrase emission by prompting the model with partial question contexts and checking whether rare report phrases from training appeared verbatim [15]. These audits are not formal privacy proofs. They are empirical indicators of whether adaptation increased memorization risk.

Ablation analysis removed one component at a time from the proposed method. The ablations were no site-adversarial loss, no prototype harmonization, sample-size aggregation instead of site-balanced aggregation, no residual gate, no update noise, and no local residual modules. The ablations were evaluated in the same leave-one-site-out protocol. The purpose was to determine which components contributed to cross-site accuracy, calibration, and site disparity.

Hyperparameters were selected using only the validation portions of participating sites [16]. The selected values were $\lambda_a = 0.18$, $\lambda_p = 0.27$, $\lambda_r = 0.004$, $\kappa = 12$, low-rank dimension $r = 8$, aggregation exponent $\rho = 0.42$, and residual-gate threshold parameters fitted from validation rotations. Sensitivity analysis varied λ_a, λ_p, and ρ. Stronger adversarial alignment reduced site-discriminator accuracy but

eventually harmed rare-finding recall. Stronger prototype alignment improved terminology consistency but over-smoothed rare answer concepts when λ_p exceeded 0.45.

The implementation used secure aggregation simulation. The coordinator did not inspect individual site gradients in the main training loop. For research analysis, update norms and aggregate statistics were logged after local clipping. The privacy audit used stored experimental updates with institutional identifiers removed. Since this was a retrospective simulation rather than a live hospital federation, network failure and asynchronous participation were not central. A supplementary stress test randomly dropped one site from 20% of rounds to evaluate robustness to incomplete participation.

5. Results

The no-adaptation backbone achieved external-site macro accuracy of 63.8%. Local-only adaptation achieved 65.1%, showing that tuning on local participating sites did not transfer well to an unseen site. Ordinary federated averaging improved macro accuracy to 68.7%. Centralized adaptation achieved 73.1%. The proposed federated site-invariant adaptation reached 72.4%, approaching the centralized reference while keeping raw data local. The difference between the proposed method and ordinary federated averaging was 3.7 percentage points, with a 95% clustered bootstrap interval from 2.8 to 4.6 percentage points.

Table 1: Benchmark Composition Across Imaging Families

Imaging Family	Pairs	Share (%)	Main Shift Source
Chest Radiography	7,840	42.6	Report terminology
Head CT	3,960	21.5	Scanner reconstruction
Abdominal CT	3,420	18.6	Acquisition protocol
MSK Radiography	3,200	17.3	Hardware variability

Table 2: Question Category Distribution

Question Category	Portion (%)	Main Challenge	Typical Output
Finding Presence	34.1	Visual ambiguity	present/absent
Localization	21.8	Anatomical shift	left frontal
Severity Qualification	13.7	Label granularity	mild/moderate
Device Assessment	11.9	Position consistency	appropriate position
Acute-Negative Screening	10.4	Terminology mismatch	no acute abnormality
Region Confirmation	8.1	Context specificity	right lower lobe

Table 3: Characteristics of Participating Sites

Site	Institution Type	Dominant Modality	Distinctive Property
Site A	Academic Hospital	Chest X-ray	Device-heavy reporting
Site B	Community Hospital	Abdominal CT	Structured indications
Site C	Academic Hospital	Head CT	Acute-negative emphasis
Site D	Community Hospital	Portable X-ray	Smaller case volume
Site E	Referral Center	MSK Imaging	Hardware-rich studies
Site F	Mixed Network	CT + X-ray	Older scanner protocols

Worst-site accuracy showed a larger relative gain. No adaptation had worst-site accuracy of 55.6%. Local-only adaptation had 56.2%. Ordinary federated averaging reached 61.9%. Centralized adaptation

Table 4: Comparison of Adaptation Strategies

Method	Macro Acc. (%)	Worst-Site (%)	Site Disparity	Macro ECE
No Adaptation	63.8	55.6	14.9	0.156
Local-Only	65.1	56.2	14.1	0.149
FedAvg	68.7	61.9	10.7	0.112
Centralized	73.1	66.8	5.9	0.067
Proposed Method	72.4	66.2	6.2	0.071

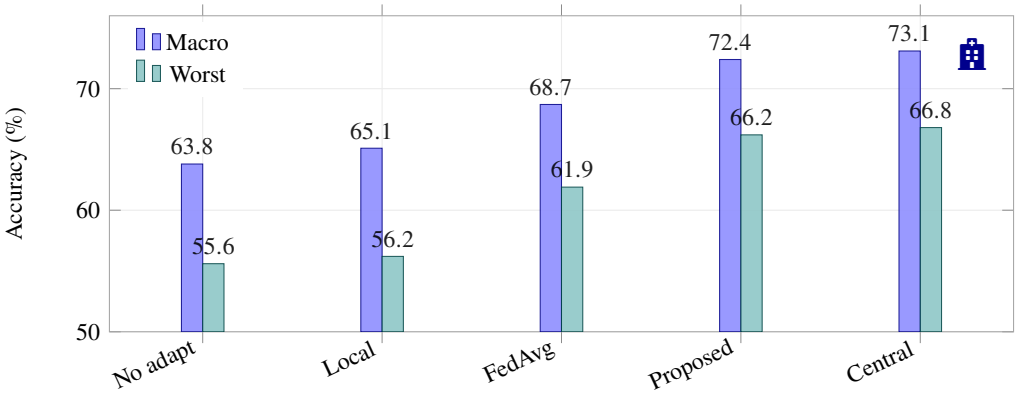


Figure 7: Held-out institutional testing shows that site-invariant federated adaptation raises macro and worst-site accuracy close to the centralized reference while retaining distributed training.

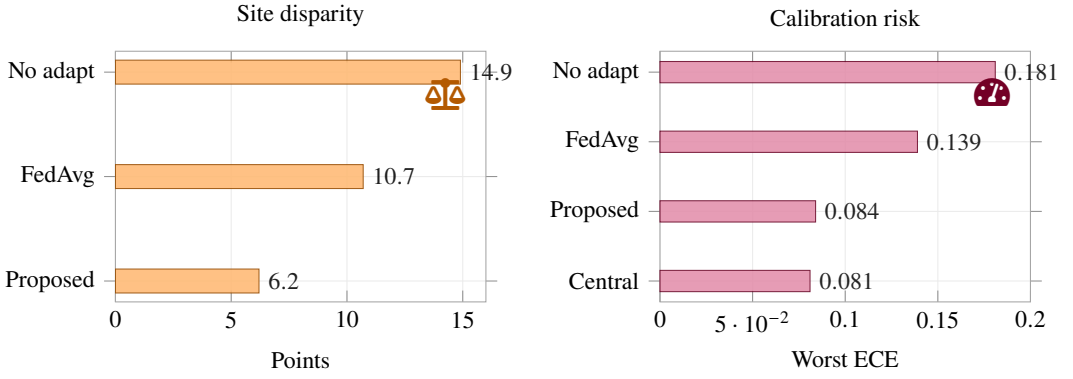


Figure 8: The proposed method reduces the spread between the best and worst institutions and nearly matches centralized calibration on the most difficult held-out site.

reached 66.8%. The proposed method reached 66.2%. Site disparity decreased from 14.9 percentage points under no adaptation to 10.7 under ordinary federated averaging and 6.2 under the proposed method. This reduction in disparity was one of the clearest benefits of site-balanced aggregation and adversarial alignment [17].

Macro expected calibration error improved substantially. No adaptation had macro expected calibration error of 0.156 and worst-site expected calibration error of 0.181. Ordinary federated averaging reduced these to 0.112 and 0.139. The proposed method reduced them to 0.071 and 0.084. Centralized adaptation achieved 0.067 and 0.081. The proposed method therefore nearly matched centralized calibration. Calibration improvement was especially visible on held-out CT sites, where the unadapted model was overconfident on scanner-shifted cases.

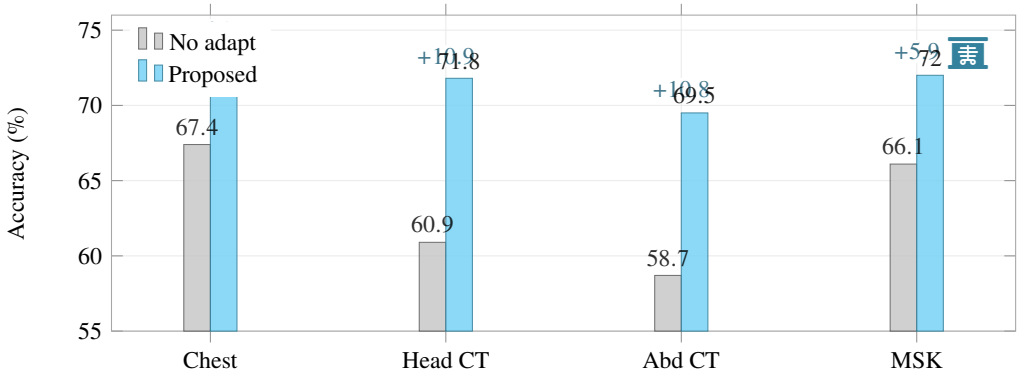


Figure 9: Modality-stratified testing shows the largest external-site gains in head and abdominal CT, consistent with scanner and reconstruction shift being major error sources.

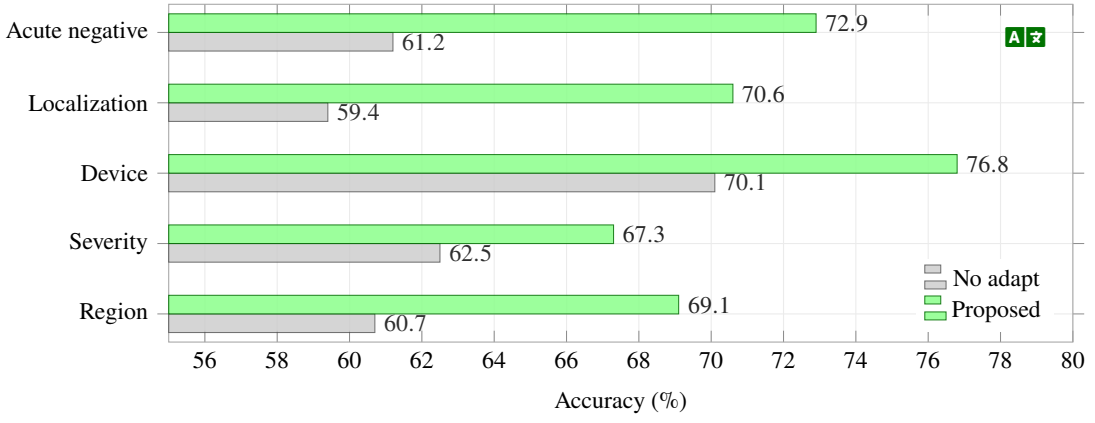


Figure 10: Question-category gains are strongest for acute-negative screening and localization, while severity remains harder because label granularity differs across reporting styles.

Table 5: External-Site Accuracy by Modality

Modality	No Adaptation	FedAvg	Proposed Method
Head CT	60.9	67.2	71.8
Abdominal CT	58.7	64.8	69.5
Chest Radiography	67.4	71.2	73.6
MSK Radiography	66.1	69.3	72.0

Micro accuracy was slightly less informative because larger sites dominated the count. No adaptation achieved 65.4% micro accuracy, ordinary federated averaging achieved 69.8%, and the proposed method achieved 72.7%. Centralized adaptation reached 73.6%. The smaller gap between macro and micro accuracy under the proposed method indicates more even institutional behavior. Local-only adaptation achieved high internal-site performance in some participating sites but poor transfer to held-out sites, confirming that local tuning alone can increase specialization without external robustness [18].

By modality, the largest gains occurred in head CT and abdominal CT. In head CT, external-site accuracy increased from 60.9% with no adaptation to 71.8% with the proposed method. Ordinary federated averaging reached 67.2%. In abdominal CT, accuracy increased from 58.7% to 69.5%. Chest radiography improved from 67.4% to 73.6%. Musculoskeletal radiography improved from 66.1% to 72.0%.

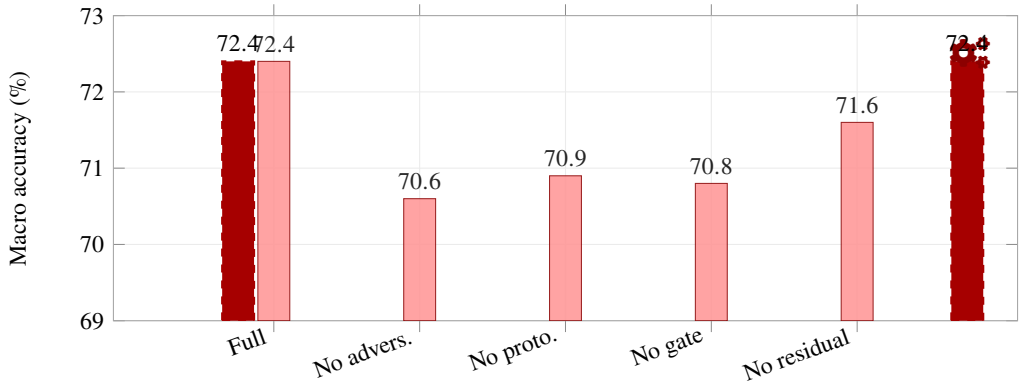


Figure 11: Component ablations show that adversarial alignment, prototype harmonization, and residual gating each contribute to external-site accuracy under the same evaluation protocol.

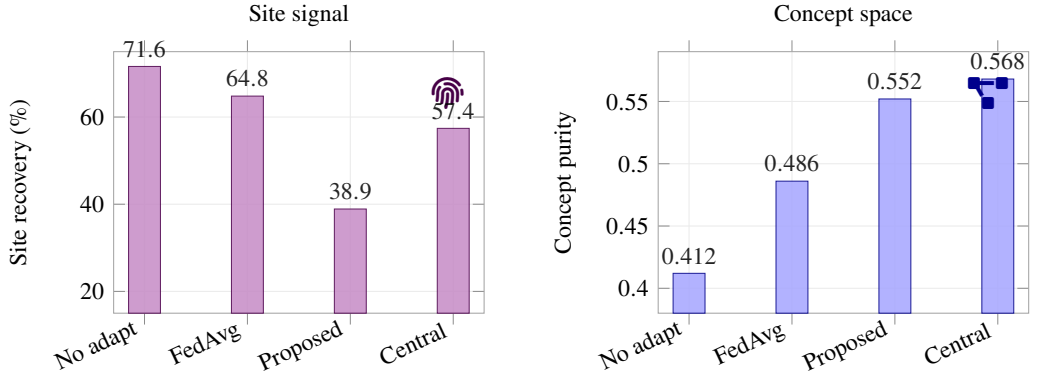


Figure 12: Representation diagnostics indicate that the proposed adaptation suppresses recoverable institution identity while improving cross-site neighborhood purity for normalized answer concepts.

Table 6: Performance Across Question Categories

Category	No Adaptation	FedAvg	Proposed Method
Acute-Negative Screening	61.2	68.7	72.9
Localization	59.4	65.8	70.6
Device Assessment	70.1	74.1	76.8
Severity Qualification	62.5	65.9	67.3
Region Confirmation	60.7	65.0	69.1

The CT gains likely reflect correction of scanner and reconstruction shift, while chest radiography gains were more associated with report-language harmonization and portable-device patterns.

Question-category analysis showed strong improvement for acute-negative screening and localization. Acute-negative screening accuracy rose from 61.2% to 72.9%. Localization rose from 59.4% to 70.6%. Device or hardware assessment rose from 70.1% to 76.8%. Severity qualification improved more modestly, from 62.5% to 67.3%. Rare severity categories remained difficult. Region-specific confirmation improved from 60.7% to 69.1%. The improvement in acute-negative screening suggests that prototype harmonization reduced confusion between different institutional forms of negative reporting.

The mixed-effects model confirmed that the proposed method significantly improved correctness relative to ordinary federated averaging. The adjusted odds ratio was 1.29, with a 95% confidence interval

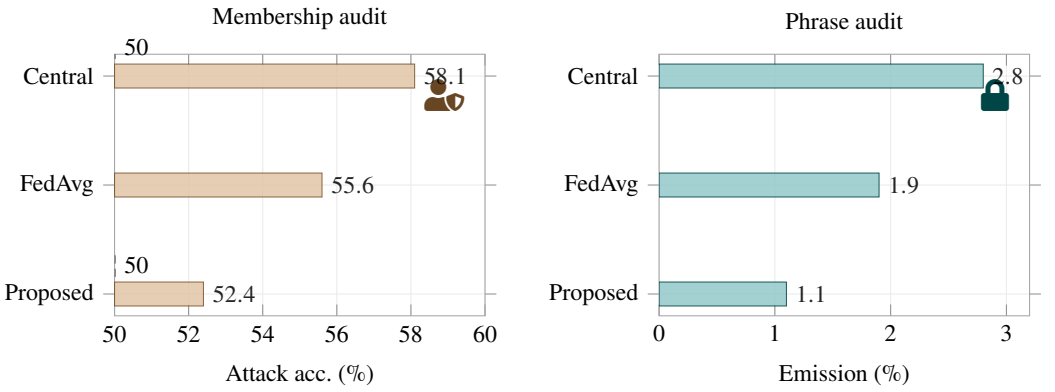


Figure 13: Privacy-oriented audits show lower membership inference and memorized phrase indicators for constrained federated adaptation than for centralized or ordinary federated tuning.

Table 7: Ablation Study of Proposed Components

Configuration	Macro Acc. (%)	Site Disparity	Key Failure
Full Method	72.4	6.2	–
No Adversarial Alignment	70.6	9.1	Site leakage
No Prototype Harmonization	70.9	8.5	Language mismatch
Sample-Size Aggregation	71.1	9.0	Large-site dominance
No Residual Gate	70.8	8.7	Overfitting

Table 8: Privacy and Memorization Audit Results

Method	Membership Attack (%)	Phrase Emission (%)	Update Risk
Centralized	58.1	2.8	Highest
FedAvg	55.6	1.9	Moderate
Proposed Method	52.4	1.1	Reduced

Table 9: Embedding and Representation Analysis

Method	Concept Purity	Site Dominance	Site Disc. Acc. (%)
No Adaptation	0.412	48.3	71.6
FedAvg	0.486	36.8	64.8
Centralized	0.568	31.2	57.4
Proposed Method	0.552	29.6	38.9

from 1.19 to 1.41. The method-by-modality interaction was significant for head CT and abdominal CT, with larger gains than chest radiography. The method-by-question interaction was significant for localization and acute-negative screening. Severity qualification did not remain significant after correction when compared with ordinary federated averaging, although it improved compared with no adaptation.

Site-discriminator accuracy decreased with the proposed method [19]. Using embeddings from no adaptation, a site classifier achieved 71.6% accuracy in recovering institution identity. Ordinary federated averaging reduced this to 64.8%. The proposed method reduced it to 38.9%, above the chance level of 16.7% but much lower than baselines. Centralized adaptation without adversarial alignment had site-discriminator accuracy of 57.4%. This indicates that data pooling alone does not remove site signatures. The proposed method made the shared embedding less institution-identifiable while preserving clinical performance.

Table 10: Sensitivity to Adversarial Weight and Aggregation

Setting	Macro Acc. (%)	Worst-Site (%)	Site Disc. Acc. (%)
$\lambda_a = 0$	70.6	63.4	63.7
$\lambda_a = 0.18$	72.4	66.2	38.9
$\lambda_a = 0.35$	70.2	64.0	29.4
Equal Site Weighting	71.9	66.5	41.8
Sample-Size Weighting	73.1	63.8	46.2

Ablations clarified component effects. Removing site-adversarial alignment reduced macro accuracy from 72.4% to 70.6% and increased site disparity from 6.2 to 9.1 points. Removing prototype harmonization reduced macro accuracy to 70.9%, with the largest drop in acute-negative screening. Using sample-size aggregation reduced worst-site accuracy from 66.2% to 63.8%. Removing the residual gate increased internal participating-site accuracy slightly but lowered external-site accuracy to 70.8%. Removing update noise had little effect on accuracy but worsened membership inference audit scores. Removing local residual modules reduced internal-site accuracy and slightly reduced external macro accuracy to 71.6%.

The label-language shift experiment showed the strongest value of prototype harmonization. When site-specific synonym tables were used during local training, ordinary federated averaging achieved 65.9% macro accuracy. The proposed method achieved 71.3%. The no-prototype ablation achieved 68.2%. Errors in this experiment often involved negative concept mismatch. For example, no acute process, no acute abnormality, and no acute intracranial finding should map to similar answer concepts, but site-specific report forms caused inconsistent embeddings without prototype alignment.

The cross-modality site-shift experiment was more difficult. When a modality was absent from a site's training contribution and then tested locally, macro accuracy under the proposed method was 67.8%, compared with 63.2% for ordinary federated averaging and 60.1% for no adaptation. The gain was meaningful but smaller than in standard leave-one-site-out testing. This indicates that the method transfers site-invariant information but still benefits from seeing modality-specific local examples. Head CT transfer was the most fragile when a site had no CT training contribution.

Rare-finding recall remained a weakness. Across answer concepts with fewer than 50 training examples per federation rotation, recall improved from 31.4% with ordinary federated averaging to 35.9% with the proposed method. Centralized adaptation reached 38.7%. These values are lower than common-finding performance. Prototype smoothing prevented some rare-concept overfitting but also limited aggressive learning from sparse examples. The method improved rare findings modestly but did not solve long-tail clinical recognition.

The privacy-oriented audits favored the proposed method over centralized adaptation and ordinary federated averaging. Membership inference attack accuracy was 58.1% for centralized adaptation, 55.6% for ordinary federated averaging, and 52.4% for the proposed method. A value of 50% corresponds to chance in the balanced audit setting. Memorized phrase emission occurred in 2.8% of centralized prompted checks, 1.9% of ordinary federated checks, and 1.1% of proposed-method checks. Update inversion similarity was lower with clipping and noise, though not eliminated. These results do not prove privacy, but they suggest reduced memorization risk relative to less constrained adaptation.

The incomplete-participation stress test showed moderate robustness. When one participating site was randomly absent in 20% of federated rounds, macro accuracy decreased from 72.4% to 71.6%. Worst-site accuracy decreased from 66.2% to 65.1%. Calibration error increased from 0.071 to 0.079. The drop was larger when the absent site was small and contained rare examples, suggesting that scheduling and participation fairness matter. Site-balanced aggregation can only help if sites actually participate enough to shape shared parameters.

Sensitivity analysis of the adversarial weight showed a clear tradeoff. With $\lambda_a = 0$, site-discriminator accuracy was 63.7% and macro accuracy was 70.6%. Increasing λ_a to 0.18 reduced site-discriminator

accuracy to 38.9% and improved macro accuracy to 72.4%. Increasing λ_a to 0.35 reduced site-discriminator accuracy further to 29.4% but lowered macro accuracy to 70.2%. The strongest invariance was not the best clinical model. This supports a balanced view: some site information may be entangled with clinically relevant patterns.

Sensitivity analysis of aggregation exponent ρ showed that equal weighting improved worst-site accuracy but slightly reduced micro accuracy. Sample-size weighting gave micro accuracy of 73.1% but worst-site accuracy of 63.8%. Equal weighting gave micro accuracy of 71.9% but worst-site accuracy of 66.5%. The selected $\rho = 0.42$ balanced the two, with micro accuracy of 72.7% and worst-site accuracy of 66.2%. This tradeoff should be set according to deployment priorities. A consortium focused on institutional equity may prefer lower ρ, while a throughput-focused setting may prefer higher ρ.

Calibration plots showed that no adaptation was overconfident on held-out sites. Predictions in the 0.8 to 0.9 confidence bin were correct only 68.5% of the time. The proposed method improved that bin to 80.7%. Underconfidence appeared in some rare findings after prototype smoothing, where predictions with moderate confidence were correct more often than estimated. Temperature scaling reduced but did not fully remove this issue. Rare-concept calibration remains an open problem.

Error analysis of external-site failures found three main patterns. Scanner-pattern dependence accounted for 27.6% of reviewed CT errors under no adaptation and 16.2% under the proposed method. Terminology mismatch accounted for 24.1% of reviewed errors under no adaptation and 12.7% under the proposed method. Prevalence-sensitive overprediction accounted for 18.9% under no adaptation and 17.4% under the proposed method. The method reduced scanner and terminology errors more effectively than prevalence errors. Rare-disease imbalance therefore remains a separate challenge.

In a manual review of 1,500 external-site outputs, reviewers judged that the proposed method produced fewer nonresponsive answers than the unadapted backbone. Nonresponse decreased from 6.7% to 3.1%. Most nonresponses occurred when CT image stacks were incomplete or when the question asked for severity with insufficient visual evidence. Federated adaptation did not encourage excessive answering at all costs; incorrect confident answers did not increase. The high-confidence incorrect rate decreased from 14.2% to 9.6% on held-out sites.

6. Representation and Error Analysis

The representation analysis examined whether the method reduced spurious institution structure while preserving clinical clusters. A two-dimensional projection of embeddings showed that no-adaptation embeddings clustered strongly by site for CT cases. After ordinary federated averaging, modality clusters became more apparent, but site clusters remained visible. After the proposed method, embeddings were more organized by answer concept and modality than by institution. This visual result was supported by quantitative site-discriminator accuracy and concept-neighbor purity.

Concept-neighbor purity measured whether examples with the same normalized answer appeared near one another across sites. For each test embedding, the 20 nearest training embeddings from participating sites were retrieved, and the fraction sharing the same answer concept was calculated. No adaptation had concept-neighbor purity of 0.412. Ordinary federated averaging reached 0.486 [20]. The proposed method reached 0.552. Centralized adaptation reached 0.568. The improvement was especially large for negative acute findings and common localization terms, suggesting that prototype alignment helped create shared semantic neighborhoods.

Site-neighbor dominance measured whether nearest neighbors came disproportionately from one institution. No adaptation had high site-neighbor dominance, with 48.3% of nearest neighbors coming from the same site when internal testing was performed. The proposed method reduced this to 29.6% while keeping concept-neighbor purity high. This indicates that embeddings became less institution-bound. In external testing, where the held-out site had no training embeddings, the analysis used participating-site dominance and showed that ordinary federated averaging relied heavily on the largest academic site. Site-balanced aggregation reduced that dominance.

CT scanner shift remained visible in some embeddings. Cases from older reconstruction protocols formed a partial cluster even after adaptation. This cluster contained both correct and incorrect answers, suggesting that the model still encoded acquisition texture. Site-adversarial alignment reduced but did not remove this structure. Stronger adversarial weight removed more scanner clustering but also harmed subtle hemorrhage detection. This illustrates a central difficulty: acquisition information may be partly nuisance and partly useful. For example, slice thickness and reconstruction kernel affect visibility of small findings.

Question-language embeddings showed another pattern. Before adaptation, negative-screening questions from different sites produced different answer neighborhoods because the report-derived answer wording varied. Prototype harmonization reduced this separation. After adaptation, no acute cardiopulmonary abnormality, no acute chest finding, and no focal acute process were closer in answer space. This improved cross-site transfer for negative cases. However, over-harmonization sometimes merged clinically different negative statements, such as no fracture and no dislocation, when the question wording was broad. The smoothing constant κ reduced this problem.

Rare-finding errors were analyzed separately. The most common rare-finding failures involved small subdural hemorrhage, subtle bowel obstruction transition point, hardware loosening, nondisplaced scaphoid fracture, and mild pneumoperitoneum. In these cases, federated adaptation often improved terminology but not visual recognition. Small sites had too few examples to teach the model robust features. Centralized adaptation performed slightly better because rare examples were pooled at the example level [21]. Prototype sharing alone was insufficient when visual evidence was scarce.

The local residual gate produced different behavior by site. Large sites generally had small residual weights because the shared module captured much of their variation. The specialty referral center received larger residual weights for musculoskeletal hardware questions. The small community site received moderate residual weights for portable chest radiographs. Residual gates improved internal validation and did not harm external performance when constrained [22]. Without the gate, local residuals overfit strongly to site-specific phrasing and acquisition habits. This overfitting increased internal accuracy by 1.1 points but reduced external macro accuracy by 1.6 points.

The method sometimes failed by under-adapting to legitimate local patterns. For example, a specialty site with many postoperative musculoskeletal cases used hardware-related terms more frequently and had different baseline prevalence. Site-adversarial alignment treated some of this as site signal and partially suppressed it. The residual module recovered part of the lost information, but not all. This suggests that future methods should distinguish site identity from clinical service-line identity. A site can be different because of technical artifacts or because it genuinely treats different patients.

Another error pattern involved answer vocabulary granularity. Some sites used coarse labels, while others used finer distinctions. For severity questions, mild versus trace or moderate versus moderate-to-severe could map imperfectly into the shared answer space. The normalized vocabulary reduced surface variation, but clinical granularity remained. Prototype harmonization can align synonyms, but it cannot decide whether two sites make different legitimate distinctions. This issue affected severity qualification more than presence or localization questions.

The privacy audit errors were also informative. Membership inference risk was highest for rare findings with distinctive question wording. Even with update clipping and noise, prototypes for rare concepts can leak some information about the presence of unusual examples. The method reduced but did not eliminate this risk. A stricter privacy setting would need stronger noise, secure aggregation, and possibly minimum-count thresholds before sharing concept prototypes. Such thresholds would further complicate rare-finding learning. Privacy and rare-disease performance are therefore in tension.

The memorized phrase audit showed that long report-like phrases were less likely to be emitted by the proposed method than by centralized adaptation. This may be because low-rank modules and normalized answer supervision discouraged memorization of full report text. However, some site-specific phrases still appeared, especially in negative screening. The generated answers were short, which limits phrase memorization. Full report generation would likely carry higher risk and would require a different privacy audit.

An important finding is that internal-site improvement did not predict external-site improvement. Local-only adaptation improved participating-site validation accuracy but barely improved held-out-site performance. Ordinary federated averaging improved both, but unevenly [23]. The proposed method improved external performance while modestly reducing the gap between internal and external results. This suggests that medical VLM development should report held-out-institution performance whenever possible. Internal validation can give an overly optimistic picture [24].

7. Governance and Deployment Considerations

Federated adaptation changes the governance discussion around medical vision-language systems. It does not remove the need for data-use agreements, audit logs, access controls, or institutional approval. It does, however, alter what must be shared. Instead of transferring raw images and reports, sites share parameter updates, prototypes, and aggregate metrics. For some consortia, this may be easier to approve than centralized data pooling. The distinction matters because imaging data are large, sensitive, and often difficult to fully de-identify.

The approach also changes accountability. In centralized training, one institution or vendor may control the combined dataset and model update. In federated training, each site participates in shaping the model while retaining local custody of data. This shared control can be beneficial but also creates coordination problems. Sites must agree on task definitions, answer concepts, validation procedures, update schedules, and monitoring responsibilities. Without common evaluation standards, federated learning can produce a model that appears collaborative but is difficult to audit.

The study's site-balanced aggregation has a governance interpretation [25]. If aggregation is based only on sample count, large hospitals dominate. That may maximize pooled accuracy but can leave smaller hospitals underserved. Equal site weighting gives smaller institutions more influence but may reduce performance on the majority distribution. The chosen exponent between these extremes reflects a policy decision as much as a technical one. A consortium should decide whether it values population-weighted performance, institution-balanced performance, or worst-site protection.

Prototype harmonization also requires governance. Mapping local report terms to shared answer concepts is not purely mechanical. Clinical experts must decide whether phrases are synonyms, adjacent categories, or different concepts. This is especially important for severity and uncertainty language. A poor mapping can create false agreement across sites. The study used normalized short answers and expert review, but a real deployment would need ongoing terminology governance as report templates and clinical practice change.

Site-adversarial alignment raises another concern. Removing site information can improve transfer, but it can also obscure meaningful differences. A pediatric hospital should not necessarily look identical to an adult trauma center. A specialty cancer center may have different disease prevalence. The proposed method partially addresses this through local residual modules, but governance should include review of whether site-specific performance drops for clinically important subgroups. Blind pursuit of invariance can become harmful if it removes relevant context.

Privacy claims should remain cautious. Federated learning is not automatically private. Parameter updates can leak information, especially for rare cases or small batches. The audits in this study showed lower memorization indicators but did not provide a formal privacy guarantee. Real deployment would need secure aggregation, strict access controls, update clipping, possible differential privacy accounting, and policies for rare prototype sharing. Institutions should not treat federated learning as a substitute for privacy engineering.

Monitoring must continue after deployment. Site distributions change when scanners are replaced, report templates are updated, new clinical services are added, or patient populations shift. A federated model may need periodic rounds of adaptation. However, repeated updates can also drift the model away from earlier validation. The consortium should maintain fixed sentinel test sets at each site and a rotating external validation design. Worst-site performance should be watched, not only overall accuracy.

The method is best suited for tasks with shared answer concepts and moderate local variation. It is less suited for tasks where sites disagree fundamentally on labels or where rare findings dominate. It also assumes that each site can perform local training. Some hospitals may lack infrastructure for secure local computation. In such cases, a trusted execution environment or hosted federated client may be needed. Technical feasibility is part of governance because a method that only large hospitals can run may worsen institutional inequity.

Human oversight remains necessary. The adapted model should not be treated as autonomous radiology interpretation. Its role is to answer constrained questions or support structured evaluation. Incorrect answers persisted, especially for rare findings and subtle CT abnormalities. Federated adaptation improved transfer but did not eliminate clinical error. Any deployment should specify how answers are reviewed, how failures are reported, and how updates are approved.

An additional governance question is whether sites can inspect model behavior relevant to their own patients. A participating hospital may want to know whether the shared model performs worse on its portable radiographs, older scanners, or postoperative cases. The federated framework can support local dashboards with site-specific metrics, but those dashboards must avoid exposing other sites' data. Aggregated cross-site reporting should use minimum-count thresholds and uncertainty intervals to prevent inference about small institutions.

The study also suggests a role for external auditing. A third party could evaluate whether federated updates improve worst-site performance, whether privacy controls are followed, and whether prototype mappings are clinically valid. Such auditing would be different from ordinary model validation because the raw data remain distributed. Auditors may need access to local evaluation scripts, signed update logs, and aggregate statistics rather than central datasets. This creates new methodological demands for medical AI oversight.

8. Limitations

The study is retrospective. The six-site benchmark captures meaningful heterogeneity, but it does not reproduce all conditions of a live federation. Real deployments involve network interruptions, local software differences, hardware limits, changing data streams, and institutional delays. The incomplete-participation stress test addresses only a small part of this complexity. Prospective federation would be necessary before drawing conclusions about operational reliability.

The modalities are limited to chest radiography, head CT, abdominal CT, and musculoskeletal radiography. Other domains such as MRI, ultrasound, mammography, pathology, ophthalmology, and interventional imaging may behave differently. Some have richer temporal structure, higher resolution, or more complex annotation needs. The method may transfer in form, but the answer concepts, visual encoders, and site-shift patterns would need redesign.

The task is visual question answering rather than full report generation. This restriction improves measurement because answers are short and normalized. Full report generation would introduce longer language, section structure, uncertainty phrasing, recommendations, and more privacy risk. Federated adaptation for full reports may require stronger language harmonization and more careful memorization auditing. The present results should not be read as evidence that federated full-report generation is safe or reliable.

The answer labels come from reports and normalized templates. Report-derived labels can contain errors, omissions, and site-specific conventions. The study filtered ambiguous cases and reviewed labels, but residual noise remains. Some apparent model errors may reflect imperfect reference answers. This problem is especially relevant for subtle findings and severity categories. Future studies should include larger expert-adjudicated subsets.

The method assumes that sites share a common answer ontology. This is feasible for constrained VQA but harder for open-ended clinical tasks. If answer concepts are poorly aligned, prototype harmonization can enforce misleading similarity. Ontology design is therefore a prerequisite. It may require specialty-specific expert panels and ongoing revision.

The privacy analysis is empirical rather than formal. Update clipping and noise reduced membership indicators, but no strict differential privacy budget is claimed. A deployment requiring formal privacy guarantees would need stronger mechanisms and likely some performance loss. Rare findings are particularly challenging because sharing useful information about them can reveal that a site has such cases. This tension is unresolved.

The site-adversarial objective may suppress clinically relevant information. Although local residual modules reduce this risk, the method still depends on validation signals to distinguish nuisance site variation from meaningful site variation. Small sites may not have enough validation examples for that distinction. Over-invariance is therefore possible. Future methods should use causal or metadata-informed approaches to separate acquisition artifacts from patient-population differences.

The benchmark anonymizes sites and model identity. This protects institutional focus and keeps attention on method behavior, but it limits direct external replication. Open multi-site medical VLM benchmarks with fixed governance are difficult to create, yet they would strengthen comparability. Future work should use public or controlled-access federated benchmark platforms where possible.

The slice-selection strategy for CT is imperfect. A fixed sampler may miss small findings or select irrelevant slices. Since every method used the same sampler, comparisons are fair, but absolute CT performance could improve with better volumetric modeling. A true three-dimensional encoder might change the balance of site shift and adaptation benefit.

The results depend on the chosen backbone and low-rank adaptation design. A stronger or weaker backbone might show different gains. Larger models may need less adaptation for common findings but may still show site-specific calibration problems. Smaller models may benefit more from local residuals. The method should be tested across additional architectures before broad claims are made.

9. Conclusion

This study examined federated site-invariant adaptation for medical vision-language models across heterogeneous radiology institutions. The empirical setting focused on visual question answering across six sites and four imaging families, with held-out-site testing used to measure transfer. The proposed method combined low-rank federated modules, site-adversarial representation alignment, prototype-based answer harmonization, site-balanced aggregation, and gated local residuals.

The results showed that the proposed method improved external-site macro accuracy, worst-site accuracy, site disparity, and calibration relative to no adaptation, local-only tuning, and ordinary federated averaging. Its performance approached centralized adaptation while avoiding raw-data pooling during training [26]. The strongest gains appeared in CT cases affected by scanner and reconstruction differences and in question categories affected by local report terminology. These improvements indicate that institution-aware federated design can address forms of site shift that ordinary federated averaging leaves unresolved.

The analysis also showed that site invariance must be balanced. Stronger removal of institution identity did not always improve clinical performance, and some local variation reflected real clinical differences. Gated residual modules helped preserve useful local information, while site-balanced aggregation reduced domination by large hospitals. Prototype harmonization improved shared answer meaning but required careful smoothing for rare concepts. These tradeoffs suggest that federated medical VLM adaptation is both a modeling problem and a governance problem.

Important weaknesses remain. Rare findings showed limited recall improvement, privacy audits did not provide formal guarantees, and the task was restricted to constrained question answering rather than full report generation. The method also depends on a shared answer ontology and adequate local training infrastructure. Prospective deployment would require stronger privacy engineering, external auditing, local monitoring, and expert review of terminology mappings.

The broader lesson is that medical vision-language systems should be evaluated across institutions, not only across pooled examples. A model that performs well at an average site may still be unreliable at a smaller or technically different hospital. Federated site-invariant adaptation offers a practical path for

collaborative improvement under data-locality constraints, but its value depends on careful measurement of worst-site behavior, calibration, rare findings, and residual privacy risk.

References

[1] P. Gounari, "Neoliberalizing higher education in greece: New laws, old free-market tricks:," *Power and Education*, vol. 4, pp. 277–288, 1 2012.

[2] P. Rajendran, Y. Chen, L. Qiu, T. Niedermayr, W. Liu, M. Buyyounouski, H. Bagshaw, B. Han, Y. Yang, N. Kovalchuk, X. Gu, S. Hancock, L. Xing, and X. Dai, "Autodelineation of treatment target volume for radiation therapy using large language model-aided multimodal learning.," *International journal of radiation oncology, biology, physics*, vol. 121, pp. 230–240, 8 2024.

[3] D. Kleindorfer, C. J. Lindsell, C. J. Moomaw, K. Alwell, D. Woo, M. L. Flaherty, O. Adeoye, T. Zakaria, J. P. Broderick, and B. M. Kissela, "Which stroke symptoms prompt a 911 call? a population-based study.," *The American journal of emergency medicine*, vol. 28, pp. 607–612, 1 2010.

[4] A. Beers, J. M. Brown, K. Chang, K. Hoebel, J. B. Patel, K. I. Ly, S. M. Tolaney, P. K. Brastianos, B. R. Rosen, E. R. Gerstner, and J. Kalpathy-Cramer, "Deepneuro: an open-source deep learning toolbox for neuroimaging," *Neuroinformatics*, vol. 19, pp. 127–140, 6 2020.

[5] C. M. Startin, H. D'Souza, G. Ball, S. Hamburg, R. Hithersay, K. M. O. Hughes, E. Massand, A. Karmiloff-Smith, M. S. Thomas, and A. Strydom, "Health comorbidities and cognitive abilities across the lifespan in down syndrome.," *Journal of neurodevelopmental disorders*, vol. 12, pp. 4–4, 1 2020.

[6] B. Sadanandan and V. Behzadan, "Vsf-med: A vulnerability scoring framework for medical vision-language models," *arXiv preprint arXiv:2507.00052*, 2025.

[7] J. Li, J. Chen, Y. Tang, C. Wang, B. A. Landman, and S. K. Zhou, "Transforming medical imaging with transformers? a comparative review of key properties, current progresses, and future perspectives.," *Medical image analysis*, vol. 85, pp. 102762–102762, 1 2023.

[8] D. Kats, L. Payne, M. Parlier, and J. Piven, "Prevalence of selected clinical problems in older adults with autism and intellectual disability," *Journal of neurodevelopmental disorders*, vol. 5, pp. 27–27, 9 2013.

[9] A. Ojagbemi, T. Bello, Z. Luo, and O. Gureje, "Chronic conditions, new onset, and persistent disability in the ibadan study of aging," *The journals of gerontology. Series A, Biological sciences and medical sciences*, vol. 72, pp. 997–1005, 3 2016.

[10] E. Fazzi, S. Micheletti, S. Calza, L. B. Merabet, A. Rossi, and J. Galli, "Early visual training and environmental adaptation for infants with visual impairment," *Developmental medicine and child neurology*, vol. 63, pp. 1180–1193, 5 2021.

[11] J. Egger, C. Gsaxner, G. Luijten, J. Chen, X. Chen, J. Bian, J. Kleesiek, and B. Puladi, "Is the apple vision pro the ultimate display? a first perspective and survey on entering the wonderland of precision medicine.," *JMIR serious games*, vol. 12, pp. e52785–e52785, 9 2024.

[12] R. L. Richesson, B. E. Bray, C. Dymek, R. A. Greenes, L. D. McIntosh, B. Middleton, G. Perry, J. Platt, and C. Shaffer, "Summary of second annual mcbk public meeting: Mobilizing computable biomedical knowledge—a movement to accelerate translation of knowledge into action," *Learning health systems*, vol. 4, pp. e10222–, 3 2020.

[13] G. E. Loeb, "Neural prosthetics:a review of empirical vs. systems engineering strategies.," *Applied bionics and biomechanics*, vol. 2018, pp. 1435030–1435030, 11 2018.

[14] K. S. Johnson and S. Lin, "Call to work together on microarray data analysis.," *Nature*, vol. 411, pp. 885–885, 6 2001.

[15] R. Chou, S. Selph, I. Blazina, C. Bougatsos, R. Jungbauer, R. Fu, S. Grusing, D. E. Jonas, and S. Tehrani, "Screening for glaucoma in adults: Updated evidence report and systematic review for the us preventive services task force.," *JAMA*, vol. 327, pp. 1998–1998, 5 2022.

[16] C. J. Steer, P. R. Jackson, H. Hornbeak, C. K. McKay, P. Sriramarao, and M. P. Murtaugh, "Team science and the physician-scientist in the age of grand health challenges.," *Annals of the New York Academy of Sciences*, vol. 1404, pp. 3–16, 10 2017.

[17] R. Masuda, P. Lanier, E. Peisner-Feinberg, and H. Hashimoto, "A quasi-experimental study of the effects of pre-kindergarten education on pediatric asthma.," *International journal of environmental research and public health*, vol. 18, pp. 10461–, 10 2021.

[18] J. Cox, P. Liu, S. E. Stolte, Y. Yang, K. Liu, K. B. See, H. Ju, and R. Fang, "Brainsegfounder: Towards 3d foundation models for neuroimage segmentation.," *Medical image analysis*, vol. 97, pp. 103301–103301, 8 2024.

[19] B. R. Vohr, B. E. Stephens, R. D. Higgins, C. Bann, S. R. Hintz, A. Das, J. E. Newman, M. Peralta-Carcelen, K. Yolton, A. M. Dusick, P. W. Evans, R. F. Goldstein, R. A. Ehrenkranz, A. Pappas, I. Adams-Chapman, D. E. Wilson-Costello, C. R. Bauer, A. Bodnar, R. J. Heyne, Y. E. Vaucher, R. G. Dillard, M. J. Acarregui, E. C. McGowan, G. J. Myers, and J. Fuller, "Are outcomes of extremely preterm infants improving? impact of bayley assessment on outcomes.," *The Journal of pediatrics*, vol. 161, pp. 222–8.e3, 3 2012.

[20] J. M. Fagerstrom, G. Eliason, H. Al-Hallaq, B. A. Taylor, M. R. Ashraf, and N. Viscariello, "Improving access in medical physics residency programs for physicists with disabilities.," *Journal of applied clinical medical physics*, vol. 25, pp. e14518–e14518, 9 2024.

[21] M. P. Moeller, G. L. Carr, L. Seaver, A. Stredler-Brown, and D. Holzinger, "Best practices in family-centered early intervention for children who are deaf or hard of hearing: an international consensus statement.," *Journal of deaf studies and deaf education*, vol. 18, pp. 429–445, 9 2013.

[22] L. M. Gottlieb, "Learning from alma ata: The medical home and comprehensive primary health care," *Journal of the American Board of Family Medicine : JABFM*, vol. 22, pp. 242–246, 5 2009.

[23] S. J. Hamstra and K. Yamazaki, "A validity framework for effective analysis and interpretation of milestones data.," *Journal of graduate medical education*, vol. 13, pp. 75–80, 4 2021.

[24] C. G. Chute, "Invited commentary: Observational research in the age of the electronic health record," *American journal of epidemiology*, vol. 179, pp. 759–761, 1 2014.

[25] S. J. Krinsky-McHale, E. C. Jenkins, W. B. Zigman, and W. Silverman, "Ophthalmic disorders in adults with down syndrome.," *Current gerontology and geriatrics research*, vol. 2012, pp. 974253–974253, 4 2012.

[26] T. Rudroff, "Artificial intelligence's transformative role in illuminating brain function in long covid patients using pet/fdg.," *Brain sciences*, vol. 14, pp. 73–73, 1 2024.