MERIT: TOWARD RELIABLE GENERALIST–SPECIALIST COLLABORATION FOR MEDICAL VISUAL QUESTION ANSWERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Reliable integration of specialist evidence is important for extending medical vision–language models (VLMs) beyond the capabilities of any single receiver, yet the heterogeneous and sometimes conflicting outputs of classifiers, segmenters, retrievers, and domain-specific generators make this difficult. Existing methods often mix multiple sources in a shared context, entangling supportive and contradictory signals before their reliability can be assessed. We introduce MERIT, a training-free framework toward reliable incorporation of specialist evidence into a frozen medical VLM at inference time. Rather than directly fusing expert outputs, MERIT preserves each specialist's native observation and feeds it to a separate source-isolated branch of the same receiver, together with a no-expert anchor branch. Because all branches share the same committed answer prefix, heterogeneous evidence is compared through its effect on the receiver's next-token distribution. Adaptive BARD forms a candidate revision from centered expert-induced residuals and accepts it only when sufficient isolated source groups prefer the candidate to the anchor; the chosen token is then synchronized across branches. This design separates evidence acquisition, influence estimation, and answer-time commitment without learning cross-expert projections or updating model parameters. We evaluate MERIT across medical VQA, multidisciplinary multimodal reasoning, and report-generation benchmarks using two representative frozen medical VLMs (i.e., HuatuoGPT-Vision and LLaVA-Med). MERIT consistently improves greedy decoding across all evaluated receiver–benchmark pairs. Ablation studies further indicate that source-isolated, receiver-mediated integration provides a general strategy for reusing heterogeneous medical expertise without additional training.

1 INTRODUCTION

Medical vision–language models (VLMs) have shown growing capabilities in image-grounded question answering and multimodal reasoning (Li et al., 2023; Chen et al., 2024a), as well as report generation (Chen et al., 2024b). However, medical visual understanding spans diverse imaging modalities, anatomical regions, and diagnostic tasks, making it difficult for a single general-purpose model to capture all relevant expertise. In contrast, specialized models such as radiograph classifiers, biomedical segmenters, pathology encoders, retrievers, and domain-specific generators provide complementary task-specific knowledge (Cohen et al., 2022; Zhao et al., 2025; Lu et al., 2024; Jin et al., 2023; Chen et al., 2024b). Reusing such specialists at inference time is therefore an appealing way to augment a frozen medical VLM without retraining the receiver or consolidating all expertise into a single model.

The central challenge for reliable collaboration, however, is not simply *which specialist to invoke*, but *how heterogeneous and potentially conflicting specialist evidence should influence the prediction of a frozen generalist*. Different specialists expose fundamentally different outputs, such as classification labels and scores (Cohen et al., 2022), segmentation masks (Zhao et al., 2025), retrieved passages (Jin et al., 2023), or generated findings (Chen et al., 2024b), and their reliability can vary across inputs (Hong et al., 2024; Xia et al., 2025). Existing approaches address parts of this problem through tool orchestration, retrieval augmentation, or evidence prompting (Fallahpour et al., 2025; Xia et al., 2025; Li et al., 2025a). A common strategy is to place multiple observations into a shared

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

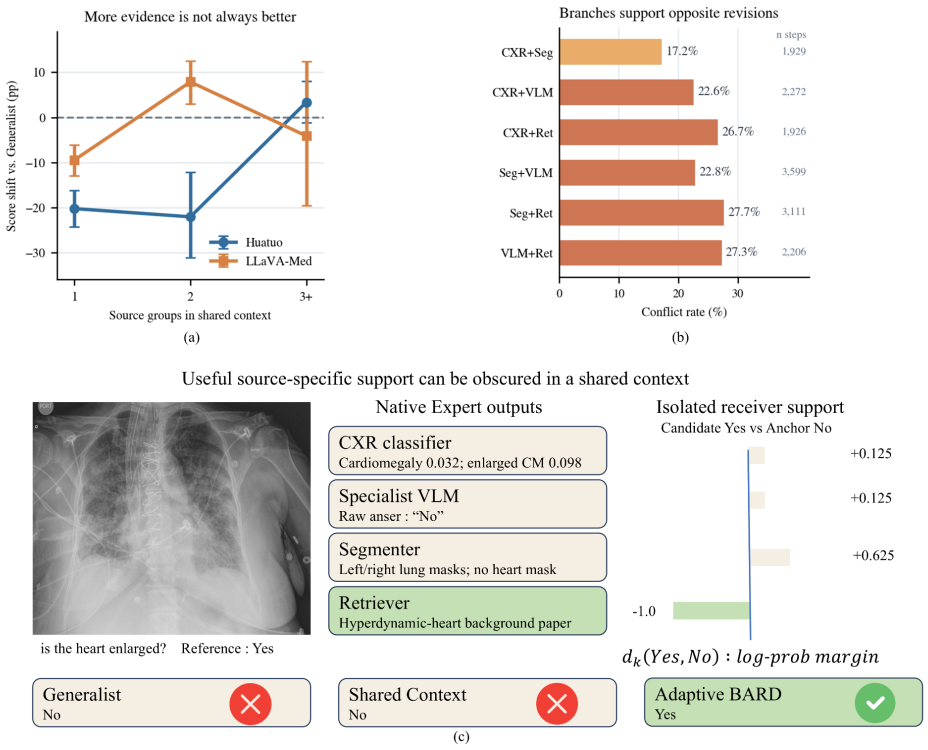

Figure 1: **Motivation: heterogeneous expert evidence can interact destructively in a shared context. (a)** Shared-context gains vary with the receiver and the number of source groups. **(b)** Isolated source groups can favor opposite revisions at the same decoding step; numbers on the right give evaluated steps. **(c)** In a representative VQA-RAD case, several isolated branches support the correct revision from anchor *No* to candidate *Yes*, while the generalist and shared-context inference remain incorrect. Bars show $d_k(b_t, a_t) = \log p_k(b_t) - \log p_k(a_t)$. These observations motivate measuring source effects before cross-source interaction.

context and let the receiver jointly reason over them (Fallahpour et al., 2025; Xia et al., 2025; Li et al., 2025a). While simple, this design couples the effects of all evidence sources before the receiver makes a prediction. Once heterogeneous observations are mixed, it becomes difficult to determine which source supports a revision, which source contradicts it, and whether useful source-specific evidence has been suppressed by interactions with other sources.

We empirically observe that such interactions can be consequential. As illustrated in Figure 1, adding more specialist evidence to a shared context does not yield monotonic improvements across receivers; isolated source groups can favor opposing revisions at the same decoding step; and source-specific evidence that supports the correct answer can be obscured after joint fusion. These observations suggest that the main difficulty is not necessarily the absence of useful specialist information, but the way in which that information is integrated. This motivates a different question: *rather than forcing heterogeneous specialists into a common representation or context, can their evidence be compared through the effect it induces on a common receiver?* Our key observation is that heterogeneous evidence need not share a native representation if its influence can be measured through the same frozen VLM.

Based on this observation, we introduce **MERIT**, a training-free framework toward *reliable, source-isolated, receiver-mediated* integration of heterogeneous specialist evidence. Instead of directly combining specialist outputs, MERIT preserves each observation in its task-native form and groups outputs that share a known source of dependence. Each resulting source group then conditions an isolated branch of the same frozen receiver, alongside a no-expert anchor branch. All branches receive the same image and question and, crucially, share the same committed answer prefix. Consequently, classifiers, segmenters, retrievers, and generators do not need to be projected into a common expert

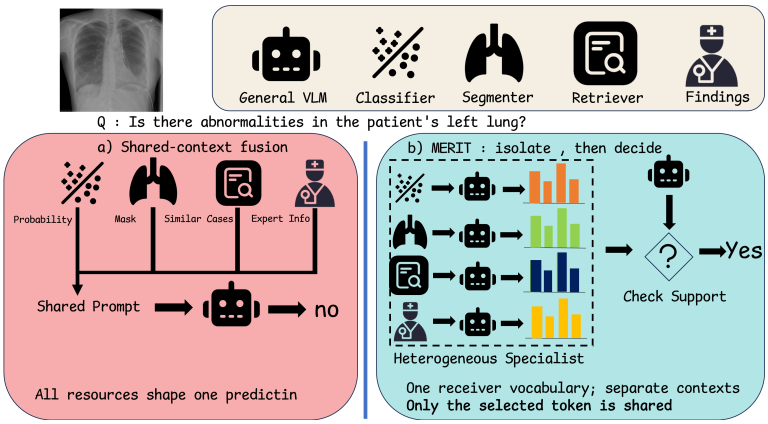

Figure 2: **Shared-context fusion versus MERIT.** Direct fusion allows all heterogeneous observations to shape one prediction before their individual effects can be identified. MERIT instead evaluates source groups in separate contexts of the same frozen receiver and commits only a supported revision. The high-resolution source image is embedded directly without re-encoding.

representation; instead, they become comparable through how they alter the receiver's next-token distribution. Figure 2 contrasts this design with direct shared-context fusion.

At each decoding step, MERIT represents the effect of a source group as a centered residual relative to the no-expert branch. **Adaptive BARD** (**B**yzantine-**A**nchored **R**esidual **D**ecoding) then separates *candidate construction* from *token commitment*. It first aggregates source-induced residuals through a geometric-median proposal to obtain a candidate token. Rather than immediately accepting the aggregated prediction, MERIT checks whether sufficiently many isolated source groups independently prefer the candidate to the no-expert anchor. Only a supported revision is committed; otherwise, the anchor token is retained. The selected token is then synchronized across all branches so that the next decoding step again compares evidence under the same realized prefix. In this way, MERIT separates *evidence acquisition*, *influence estimation*, and *answer-time commitment*, without learning cross-expert projections, training a gating module, or updating either the receiver or specialist parameters.

Our contributions are: (i) a receiver-mediated interface for heterogeneous task-native experts; (ii) a source-isolated, anchor-protected decoding rule; and (iii) cross-backbone, ablation, and report-generation evidence for training-free specialist collaboration.

2 RELATED WORK

Medical models and external expertise. LLaVA-Med and HuatuoGPT-Vision adapt visual instruction following to medical data (Li et al., 2023; Chen et al., 2024a). More specialized interfaces include chest-radiograph prediction with TorchXRayVision (Cohen et al., 2022), pathology image–text representations with CONCH (Lu et al., 2024), biomedical object parsing with BiomedParse (Zhao et al., 2025), and literature retrieval with MedCPT (Jin et al., 2023). MedRAX coordinates medical tools (Fallahpour et al., 2025), while visual evidence prompting converts specialist observations into contextual guidance (Li et al., 2025a). MERIT uses this form of external expertise but addresses how several fallible sources influence the receiver during decoding.

Tool orchestration and inference-time control. LLMCompiler separates planning, dependency-aware dispatch, and execution to coordinate function calls (Kim et al., 2024); OctoTools provides extensible tools and multi-step orchestration (Lu et al., 2026). Our separation concerns evidence-conditioned predictions rather than parallel task scheduling. VCD and ICD contrast input conditions (Leng et al., 2024; Wang et al., 2024), DoLa contrasts model layers (Chuang et al., 2024), and PAI, AvisC, and VISTA modify attention or visual information during inference (Liu et al., 2024;

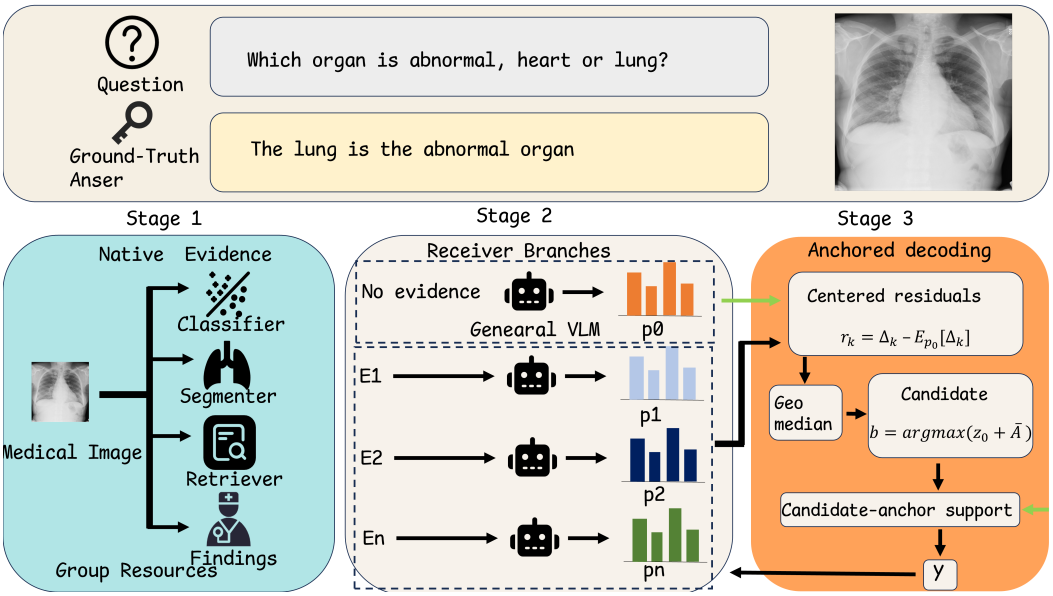

Figure 3: **MERIT at one decoding step.** Native specialist outputs are grouped by declared source dependence. Each group conditions an isolated instance of the same frozen receiver, alongside a no-expert branch. The decoder centers group-induced residuals, forms a geometric-median proposal, and tests candidate–anchor support. It then appends one selected token to every live branch. Model icons indicate shared weights, not separately trained models. Colored distributions and median geometry are illustrative; exact computations appear in Equations 3–7.

Woo et al., 2025; Li et al., 2025b). These approaches provide relevant inference-time baselines, but do not directly define the contribution of heterogeneous external experts.

Clinical Contrastive Decoding also uses specialist signals for training-free radiology generation (Zhang et al., 2026). MERIT instead compares isolated source-group effects through a shared receiver; the current tables do not establish superiority over CCD.

Isolation and robust aggregation. RobustRAG isolates retrieved contexts before aggregating model responses (Xiang et al., 2026). Geometric-median aggregation is an established robust estimation primitive, including in federated learning (Pillutla et al., 2022). We do not claim either primitive as new. Our design combines native medical evidence, a shared multimodal receiver, centered residuals relative to a no-expert anchor, and coverage-dependent support checks. The resulting rule is not a certificate of medical correctness: source groups can share errors, and reliable expertise can be sparse.

3 METHOD

Given an image I, question q, frozen receiver G_θ, and registry of specialists, MERIT generates a free-form answer $y = (y_1, \ldots, y_T)$. The framework has three components: a native evidence interface, isolated receiver branches, and an anchored residual decoder. Figure 3 shows their data flow. Specialist acquisition occurs before decoding; subsequent steps reuse the observations and synchronize only the committed answer prefix.

3.1 NATIVE EVIDENCE INTERFACE

A specialist is eligible when its declared modality, anatomical scope, input requirements, and native task match the request. An eligible expert produces $o_e = E_e(I, q)$ using its supported interface. The adapter preserves the observation's semantics: labels and native scores for classification, geometry for segmentation, text for generation, and provenance-bearing passages for retrieval. These outputs

need not constitute answers to the complete question. In the running example, identifying a region and describing an abnormality remain distinct observations.

Textual outputs enter the receiver through a deterministic evidence rendering. When supported, masks and activation maps additionally use the existing parameter-free spatial interface. An unsupported dense artifact is not silently treated as if its pixel values were visible in a text prompt. The original image remains an input to every receiver branch. No cross-model projection is learned.

Multiple tools may share a checkpoint, upstream prediction, or other known source of dependence. We combine such outputs into a declared source group rather than assign an independent vote to every tool call. Let $E_1, \ldots, E_n$ denote the resulting evidence contexts. Grouping removes obvious duplication but does not establish error independence. Both the group assignment and the admissible expert interfaces are fixed before evaluation.

3.2 Isolated Receiver Branches

The receiver provides the common output space. At step t, the no-expert branch defines

$$p_0^t(v) = p_\theta(v \mid I, q, y_{<t}), \qquad a_t = \arg\max_v p_0^t(v), \tag{1}$$

and source group k defines

$$p_k^t(v) = p_\theta(v \mid I, q, E_k, y_{<t}), \qquad k = 1, \ldots, n. \tag{2}$$

All branches use the same receiver vocabulary and valid token support. A different backbone instantiates a separate decoding process; its token IDs are never combined with another backbone's IDs. This is also not a comparison between independently generated sentences: the contexts differ in evidence, while their committed output tokens are identical.

Writing $z_k^t = \log p_k^t$, we represent a source's effect with a centered residual:

$$\Delta_k^t = z_k^t - z_0^t, \qquad r_k^t = \Delta_k^t - \mathbb{E}_{v \sim p_0^t}[\Delta_k^t(v)]\,\mathbf{1}. \tag{3}$$

Centering fixes a common offset convention before distances between vectors are measured. It leaves each branch's pairwise token preferences unchanged. Importantly, r_k^t measures how an observation changes this receiver, not the observation's probability of being correct. Native classifier and retrieval scores are not substituted for r_k^t.

The isolated contexts separate acquisition from influence. In the running example, the classifier and segmenter each influence a separate prediction, and both effects are compared with the same anchor. Neither expert needs language logits.

3.3 Anchored Residual Decoding

Adaptive BARD separates candidate construction from token commitment. With $n \geq 2$ available groups, it constructs a residual proposal using the Euclidean geometric median:

$$\bar{r}^t \in \arg\min_r \sum_{k=1}^{n} \|r - r_k^t\|_2, \qquad b_t = \arg\max_v \big[z_0^t(v) + \bar{r}^t(v)\big]. \tag{4}$$

The geometric median reduces the leverage of an isolated distant residual relative to an arithmetic mean (Pillutla et al., 2022). It cannot remove an error shared by several groups. With two residuals, the implementation selects their midpoint; the stricter acceptance rule below handles this sparse case.

Proposal versus protection. For arithmetic aggregation, $z_0^t + n^{-1}\sum_k r_k^t = n^{-1}\sum_k z_k^t + c_t\mathbf{1}$, where c_t is constant across tokens. The mean-residual proposal is therefore equivalent to averaging expert-conditioned log probabilities; the implemented two-source midpoint has the same property. Residual formation alone does not protect the baseline. That role belongs to the acceptance rule below.

Candidate support. A source supports b_t only if its isolated receiver branch prefers b_t to a_t. Its margin is

$$\begin{aligned} d_k^t(b_t, a_t) &= z_0^t(b_t) - z_0^t(a_t) + r_k^t(b_t) - r_k^t(a_t) \\ &= z_k^t(b_t) - z_k^t(a_t). \end{aligned} \tag{5}$$

A positive residual increase is insufficient: the complete evidence-conditioned branch must prefer the candidate to the anchor.

Coverage-dependent commitment. For a declared budget f, set $\hat{f}_n = \min\{f, \lfloor (n-1)/2 \rfloor\}$. Define the aggregate margin $g_t = [z_0^t + \bar{r}^t](b_t) - [z_0^t + \bar{r}^t](a_t)$. For a proposal distinct from the anchor, acceptance is

$$A_t = \begin{cases} 0, & n \le 1, \\ \mathbf{1}\{g_t > 0 \ \wedge \ \min_k d_k^t > 0\}, & n = 2, \\ \mathbf{1}\{g_t > 0 \ \wedge \ \sum_k \mathbf{1}[d_k^t > 0] \ge n - \hat{f}_n\}, & n \ge 3. \end{cases} \tag{6}$$

If $b_t = a_t$, the decoder emits the anchor directly. Otherwise, it commits

$$y_t = \begin{cases} b_t, & A_t = 1, \\ a_t, & A_t = 0. \end{cases} \qquad y_{<t+1} = [y_{<t}, y_t]. \tag{7}$$

For the default $f = 1$, two groups must be unanimous, while three groups require two positive margins. With one source, the decoder preserves the anchor instead of treating a single observation as sufficient redundancy. This policy is deliberately conservative: it can reject a uniquely correct specialist.

What the fault budget means. The budget declares how many source groups may be unreliable; it does not estimate the number of wrong experts. This is centralized aggregation, without the $3f + 1$ communication requirement of distributed agreement. The term "Byzantine" describes the robustness motivation, not a clinical guarantee.

3.4 Commit Once, Synchronize All Branches

Every live branch receives the token selected by Equation 7, including the no-expert branch. The following step therefore compares distributions at the same realized prefix. Separate contexts and caches preserve evidence isolation; separate unconstrained answer trajectories would not. Algorithm 1 summarizes the complete inference procedure.

The anchor is local to the current prefix. After an accepted change, a later anchor token is conditioned on that modified prefix; it does not restore the answer that an independent greedy run would have produced. Thus, next-token anchoring is not a guarantee of sequence-level baseline preservation.

4 Experiments

We evaluate MERIT around three questions. *First*, can the same training-free integration rule improve different frozen medical VLM receivers across tasks? *Second*, which aspects of evidence integration are associated with the observed gains: shared-context fusion, source isolation, robust aggregation, or anchor-protected commitment? *Third*, does the same inference-time mechanism extend beyond short-form question answering to open-ended autoregressive generation? We address these questions through cross-receiver benchmark comparisons, controlled evidence-integration ablations, and chest-radiograph report generation.

4.1 Experimental Setup

Receivers and benchmarks. We use two frozen medical vision–language models as receivers: HuatuoGPT-Vision-7B (Chen et al., 2024a) and LLaVA-Med-7B (Li et al., 2023). For medical visual question answering, we evaluate on PMC-VQA (Zhang et al., 2023), PathVQA (He et al., 2020), SLAKE (Liu et al., 2021), and VQA-RAD (Lau et al., 2018). We additionally evaluate broader multimodal reasoning on MMMU (Yue et al., 2024). To test whether the same integration mechanism applies beyond short-form answers, we evaluate free-form chest-radiograph report generation on MIMIC-CXR (Johnson et al., 2019).

Algorithm 1 MERIT with Adaptive BARD

Require: Image I, question q, frozen receiver, expert registry, budget T, fault budget f

Acquire native outputs; construct source-group contexts $\{E_k\}_{k=1}^{n}$
Initialize isolated receiver contexts and committed prefix $y \leftarrow [\,]$
for $t = 1, \ldots, T$ **do**
    Obtain p_0^t at prefix y; set $a \leftarrow \arg\max_v p_0^t(v)$ and $u \leftarrow a$
    if $n \geq 2$ **then**
        Obtain each p_k^t using I, q, E_k and the same prefix y
        Compute centered residuals and candidate b (Equations 3–4)
        Compute source margins and A_t (Equations 5–6)
        if $b \neq a$ and $A_t = 1$ **then**
            $u \leftarrow b$
        end if
    end if
    Append u to y and synchronize every live branch
    if u is EOS **then**
        break
    end if
end for
return y

Specialist pool. We construct the expert registry to cover complementary evidence types rather than maximize the number of nominal models. The configured pool includes chest-radiograph finding and anatomy models from TorchXRayVision (Cohen et al., 2022); pathology and biomedical image–text encoders from CONCH (Lu et al., 2024) and BiomedCLIP (Zhang et al., 2024); promptable biomedical segmentation with BiomedParse (Zhao et al., 2025); question-conditioned chest-radiograph findings from CheXagent (Chen et al., 2024b); and biomedical knowledge or source-case retrieval with MedCPT and BiomedCLIP-based retrieval (Jin et al., 2023; Zhang et al., 2024). Each registration declares its supported modality, anatomical or task scope, native output type, and source group. Only experts whose declared requirements match the current request and whose evidence is actually delivered become active branches. Capabilities that share a known underlying model family are grouped into the same source group to avoid treating obviously correlated outputs as independent support. All specialists remain frozen, and not every expert is applicable to every benchmark example.

Baselines and metrics. For the VQA and multimodal reasoning benchmarks, we compare MERIT with standard greedy decoding and a set of training-free or inference-time decoding approaches: VCD (Leng et al., 2024), ICD (Wang et al., 2024), DoLa (Chuang et al., 2024), PAI (Liu et al., 2024), AvisC (Woo et al., 2025), VISTA (Li et al., 2025b), and MedRAG (Xiong et al., 2024); For MIMIC-CXR report generation, we additionally include AGLA (An et al., 2024) and OPERA (Huang et al., 2024). We retain each benchmark's supplied scoring convention and report individual dataset scores as percentages. For a compact cross-dataset summary, Table 1 additionally reports an unweighted macro-average (Avg.) over the five benchmark scores within each receiver block. Because the underlying benchmark protocols are heterogeneous, Avg. is used only as a descriptive summary and not as a replacement for per-dataset evaluation. For MIMIC-CXR, we report BLEU-4, ROUGE-L, and METEOR. Higher values indicate better performance for all reported metrics.

Inference and implementation. All specialist models and both receiver models remain frozen throughout evaluation. Eligible specialists are executed before autoregressive decoding, and their native outputs are cached together with source identifiers. Only source groups that are actually delivered to the receiver contribute to the number of available groups n.

At each decoding step with $n \geq 2$, MERIT evaluates one no-expert anchor branch and n source-isolated branches under the same committed prefix. The geometric median used by Adaptive BARD is computed with at most 12 stabilized Weiszfeld iterations, a relative tolerance of 10^{-6}, and an inverse-distance floor of 10^{-12}. After candidate construction and the support test, the selected token is appended to every live branch before the next decoding step. This implementation preserves source isolation while ensuring that all branches are compared under the same realized answer prefix.

MERIT introduces additional inference-time computation relative to greedy decoding, and we do not claim a wall-clock speedup.

4.2 Main Results

We first examine whether the same collaboration rule transfers across different frozen receivers and benchmark settings. Table 1 reports results for both HuatuoGPT-Vision-7B and LLaVA-Med-7B across the five evaluated benchmarks. MERIT numerically improves over greedy decoding in all ten evaluated receiver–benchmark pairs. For HuatuoGPT-Vision, the largest gain over greedy occurs on MMMU, followed by SLAKE; for LLaVA-Med, the largest gains occur on VQA-RAD and SLAKE. Across the ten settings, MERIT achieves the highest score among the evaluated methods in seven.

The Avg. column provides a descriptive macro-average across the five benchmarks. MERIT obtains the highest average for both receivers, reaching 50.94 with HuatuoGPT-Vision and 32.65 with LLaVA-Med. This average advantage does not imply uniform per-dataset dominance: DoLa remains stronger on HuatuoGPT-Vision PMC-VQA, MedRAG on HuatuoGPT-Vision PathVQA, and PAI on LLaVA-Med PathVQA. We therefore interpret Table 1 as evidence of broad cross-dataset gains while retaining the individual benchmark results as the primary comparison.

Table 1: **Main VQA results across two medical VLM receivers.** Scores are percentages; higher is better. Avg. is the unweighted macro-average over the five benchmark scores. Best results within each receiver block are bold.

Method	PMC-VQA	PathVQA	SLAKE	VQA-RAD	MMMU	Avg.
HuatuoGPT-Vision-7B						
Greedy	52.35	38.66	54.69	62.82	32.70	48.24
VCD (Leng et al., 2024)	43.72	33.58	52.83	55.83	35.46	44.28
ICD (Wang et al., 2024)	42.34	31.86	51.52	52.37	35.07	42.63
DoLa (Chuang et al., 2024)	**53.81**	29.35	48.48	49.23	38.34	43.84
PAI (Liu et al., 2024)	52.15	37.23	52.73	60.40	30.56	46.61
AvisC (Woo et al., 2025)	44.83	34.48	49.98	57.63	38.19	45.02
VISTA (Li et al., 2025b)	53.70	35.40	54.37	57.84	36.60	47.58
MedRAG (Xiong et al., 2024)	46.23	**39.65**	49.88	48.88	29.14	42.76
MERIT	52.42	38.86	**57.16**	**63.06**	**43.19**	**50.94**
LLaVA-Med-7B						
Greedy	13.63	31.67	34.68	48.47	20.65	29.82
VCD (Leng et al., 2024)	11.75	31.33	32.54	47.19	17.72	28.11
ICD (Wang et al., 2024)	11.68	31.50	33.30	47.16	20.89	28.91
DoLa (Chuang et al., 2024)	13.60	31.64	34.78	48.81	20.64	29.89
PAI (Liu et al., 2024)	13.06	**32.24**	35.12	47.83	20.15	29.68
AvisC (Woo et al., 2025)	9.93	32.06	33.07	47.83	17.53	28.08
VISTA (Li et al., 2025b)	13.24	31.56	35.56	46.68	20.65	29.54
MedRAG (Xiong et al., 2024)	13.39	31.18	27.19	39.41	22.10	26.65
MERIT	**14.92**	32.14	**39.49**	**53.96**	**22.73**	**32.65**

4.3 Analysis

We complement the benchmark comparison with two analyses that target different aspects of the method. First, we isolate how evidence organization and commitment affect short-form VQA. Second, we test whether the same inference-time collaboration extends to open-ended chest-radiograph reporting. Table 2 collects both results.

Table 2: Additional analyses. (a) Evidence-integration ablation on SLAKE and VQA-RAD; Adaptive BARD is the complete MERIT decoder. **(b)** MIMIC-CXR report generation with LLaVA-Med. Higher is better.

(a) Evidence integration

Receiver	Data	Gen.	Joint-all	Iso-Mean	Iso-Geo	Adaptive BARD
Huatuo	SLAKE	54.67	49.31	48.08	48.00	**57.16**
LLaVA-Med	SLAKE	34.68	32.28	32.93	33.12	**39.49**
Huatuo	VQA-RAD	62.84	44.79	47.23	47.95	**63.06**
LLaVA-Med	VQA-RAD	48.47	45.44	43.52	43.74	**53.96**

(b) MIMIC-CXR report generation

Method	BLEU-4↑	ROUGE-L↑	METEOR↑
Greedy	1.02	12.87	9.90
VCD (Leng et al., 2024)	1.02	12.47	9.52
ICD (Wang et al., 2024)	1.21	13.03	10.37
DoLa (Chuang et al., 2024)	1.01	12.83	9.84
PAI (Liu et al., 2024)	1.06	12.85	9.64
AvisC (Woo et al., 2025)	1.48	13.09	13.26
VISTA (Li et al., 2025b)	1.00	13.16	9.47
AGLA (An et al., 2024)	1.18	13.69	10.89
OPERA (Huang et al., 2024)	1.41	13.80	12.73
MERIT	**2.53**	**17.11**	**14.82**

Ablation study The main results establish the cross-setting behavior of MERIT, but do not by themselves identify which part of the integration strategy is associated with the gains. We therefore compare four variants that differ in how heterogeneous evidence is exposed to and combined by the receiver.

Table 2(a) compares the unassisted generalist with four increasingly structured variants. *Joint-all* mixes all source groups in one context; *Iso-Mean* and *Iso-Geo* aggregate source-isolated branches without anchored commitment; *Adaptive BARD* is the complete decoder. Joint-all underperforms the matched generalist in all four settings, and isolated mean or geometric-median aggregation does not recover the loss. Adaptive BARD is the only variant that improves on every matched generalist, indicating that controlled commitment matters beyond adding or robustly aggregating evidence.

Generalization to Open-ended report generation Task. Table 2(b) evaluates LLaVA-Med on MIMIC-CXR benchmark for report generation. MERIT obtains the highest reported BLEU-4, ROUGE-L, and METEOR scores, exceeding the strongest non-MERIT result by 1.05, 3.31, and 1.56 points, respectively. This complements the VQA results by showing that the same inference-time mechanism can also alter longer autoregressive outputs.

Qualitative behavior. A representative MIMIC-CXR example is provided in Appendix A; we keep the main text focused on aggregate and controlled comparisons.

5 Conclusion

In this work, we introduced MERIT, a training-free framework for integrating heterogeneous specialist evidence into frozen medical vision–language models. Its core contribution is to replace direct cross-expert fusion with receiver-mediated influence comparison: task-native specialist observations are evaluated through source-isolated branches of the same frozen receiver, making their effects comparable in a shared next-token space without learned cross-expert projections or parameter updates. Building on this interface, Adaptive BARD further separates candidate construction from answer-time commitment by accepting a proposed revision only when it receives sufficient support relative to the no-expert anchor, while synchronizing the selected token across branches to preserve a common decoding prefix. Across two medical VLM receivers, MERIT numerically improves greedy decoding in all evaluated receiver–benchmark pairs, while ablations show that neither shared-context fusion nor unprotected aggregation of isolated source effects reproduces the behavior of the complete method; the same mechanism also improves the reported MIMIC-CXR generation metrics. Overall, MERIT provides a practical inference-time framework for reusing complementary medical expertise while keeping both specialists and the receiver frozen.

AI Assistance Disclosure

Generative AI tools were used to aid or polish writing, for retrieval and discovery (e.g., finding related work), and for research ideation or execution. All AI-assisted content was reviewed by the authors, who take responsibility for the final paper.

References

Wenbin An, Feng Tian, Sicong Leng, Jiahao Nie, Haonan Lin, QianYing Wang, Guang Dai, Ping Chen, and Shijian Lu. Agla: Mitigating object hallucinations in large vision-language models with assembly of global and local attention. *arXiv preprint arXiv:2406.12718*, 2024. URL https://arxiv.org/abs/2406.12718.

Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Ruifei Zhang, Zhenyang Cai, Ke Ji, Guangjun Yu, Xiang Wan, and Benyou Wang. Huatuogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale. *arXiv preprint arXiv:2406.19280*, 2024a. URL https://arxiv.org/abs/2406.19280.

Zhihong Chen, Maya Varma, Justin Xu, Magdalini Paschali, Dave Van Veen, Andrew Johnston, Alaa Youssef, Louis Blankemeier, Christian Bluethgen, Stephan Altmayer, Jeya Maria Jose Valanarasu, Mohamed Siddig Eltayeb Muneer, Eduardo Pontes Reis, Joseph Paul Cohen, Cameron Olsen, Tanishq Mathew Abraham, Emily B. Tsai, Christopher F. Beaulieu, Jenia Jitsev, Sergios Gatidis, Jean-Benoit Delbrouck, Akshay S. Chaudhari, and Curtis P. Langlotz. A vision-language foundation model to enhance efficiency of chest x-ray interpretation. *arXiv preprint arXiv:2401.12208*, 2024b. URL https://arxiv.org/abs/2401.12208.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R. Glass, and Pengcheng He. Dola: Decoding by contrasting layers improves factuality in large language models. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/edc36117f795ca52a0cbf6a7b3882859-Abstract-Conference.html.

Joseph Paul Cohen, Joseph D. Viviano, Paul Bertin, Paul Morrison, Parsa Torabian, Matteo Guarrera, Matthew P. Lungren, Akshay Chaudhari, Rupert Brooks, Mohammad Hashir, and Hadrien Bertrand. Torchxrayvision: A library of chest x-ray datasets and models. In *Proceedings of the 5th International Conference on Medical Imaging with Deep Learning*, volume 172, pp. 231–249, 2022. URL https://proceedings.mlr.press/v172/cohen22a.html.

Adibvafa Fallahpour, Jun Ma, Alif Munim, Hongwei Lyu, and Bo Wang. Medrax: Medical reasoning agent for chest x-ray. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 15661–15676, 2025. URL https://proceedings.mlr.press/v267/fallahpour25a.html.

Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020. URL https://arxiv.org/abs/2003.10286.

Giwon Hong, Jeonghwan Kim, Junmo Kang, Sung-Hyon Myaeng, and Joyce Jiyoung Whang. Why so gullible? enhancing the robustness of retrieval-augmented models against counterfactual noise. In *Findings of the Association for Computational Linguistics: NAACL*, pp. 2474–2495, 2024. doi: 10.18653/v1/2024.findings-naacl.159.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13418–13427, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Huang_OPERA_Alleviating_Hallucination_in_Multi-Modal_Large_Language_Models_via_Over-Trust_CVPR_2024_paper.html.

Qiao Jin, Won Kim, Qingyu Chen, Donald C. Comeau, Lana Yeganova, W. John Wilbur, and Zhiyong Lu. Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11):btad651, 2023. doi: 10.1093/bioinformatics/btad651.

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.

Sehoon Kim, Suhong Moon, Ryan Tabrizi, Nicholas Lee, Michael W. Mahoney, Kurt Keutzer, and Amir Gholami. An llm compiler for parallel function calling. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 24370–24391, 2024. URL https://proceedings.mlr.press/v235/kim24y.html.

Jason J. Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific Data*, 5:180251, 2018. doi: 10.1038/sdata.2018.251.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13872–13882, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Leng_Mitigating_Object_Hallucinations_in_Large_Vision-Language_Models_through_Visual_Contrastive_CVPR_2024_paper.html.

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 36, Datasets and Benchmarks Track*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/5abcdf8ecdcacba028c6662789194572-Abstract-Datasets_and_Benchmarks.html.

Wei Li, Zhen Huang, Houqiang Li, Le Lu, Yang Lu, Xinmei Tian, Xu Shen, and Jieping Ye. Visual evidence prompting mitigates hallucinations in large vision-language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4048–4080, 2025a. doi: 10.18653/v1/2025.acl-long.205. URL https://aclanthology.org/2025.acl-long.205/.

Zhuowei Li, Haizhou Shi, Yunhe Gao, Di Liu, Zhenting Wang, Yuxiao Chen, Ting Liu, Long Zhao, Hao Wang, and Dimitris N. Metaxas. The hidden life of tokens: Reducing hallucination of large vision-language models via visual information steering. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 35799–35819, 2025b. URL https://proceedings.mlr.press/v267/li25ca.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 *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pp. 1650–1654, 2021. doi: 10.1109/ISBI48211.2021.9434010.

Shi Liu, Kecheng Zheng, and Wei Chen. Paying more attention to images: A training-free method for alleviating hallucination in lvlms. In *European Conference on Computer Vision*, 2024.

Ming Y. Lu, Bowen Chen, Drew F. K. Williamson, Richard J. Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, Anil V. Parwani, Andrew Zhang, and Faisal Mahmood. A visual-language foundation model for computational pathology. *Nature Medicine*, 30:863–874, 2024. doi: 10.1038/s41591-024-02856-4. URL https://www.nature.com/articles/s41591-024-02856-4.

Pan Lu, Bowen Chen, Sheng Liu, Rahul Thapa, Joseph Boen, and James Zou. Octotools: A multi-agent framework with extensible tools for complex reasoning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1–86,

2026. doi: 10.18653/v1/2026.acl-long.1. URL https://aclanthology.org/2026.acl-long.1/.

Krishna Pillutla, Sham M. Kakade, and Zaid Harchaoui. Robust aggregation for federated learning. *IEEE Transactions on Signal Processing*, 70:1142–1154, 2022. doi: 10.1109/TSP.2022.3153135.

Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 15840–15853, 2024. doi: 10.18653/v1/2024.findings-acl.937. URL https://aclanthology.org/2024.findings-acl.937/.

Sangmin Woo, Donguk Kim, Jaehyuk Jang, Yubin Choi, and Changick Kim. Don't miss the forest for the trees: Attentional vision calibration for large vision language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 1927–1951, 2025. doi: 10.18653/v1/2025.findings-acl.99. URL https://aclanthology.org/2025.findings-acl.99/.

Peng Xia, Kangyu Zhu, Haoran Li, Tianze Wang, Weijia Shi, Sheng Wang, Linjun Zhang, James Zou, and Huaxiu Yao. Mmed-rag: Versatile multimodal rag system for medical vision language models. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/a559a5a8aa5ae6682ced009ad97cdb16-Abstract-Conference.html.

Chong Xiang, Tong Wu, Zexuan Zhong, David Wagner, Danqi Chen, and Prateek Mittal. Certifiably robust rag against retrieval corruption. In *IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, 2026. URL https://satml.org/2026/program/.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 6233–6251, 2024. doi: 10.18653/v1/2024.findings-acl.372. URL https://aclanthology.org/2024.findings-acl.372/.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9556–9567, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Yue_MMMU_A_Massive_Multi-discipline_Multimodal_Understanding_and_Reasoning_Benchmark_for_CVPR_2024_paper.html.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Andrea Tupini, Yu Wang, Matt Mazzola, Swadheen Shukla, Lars Liden, Jianfeng Gao, Angela Crabtree, Brian Piening, Carlo Bifulco, Matthew P. Lungren, Tristan Naumann, Sheng Wang, and Hoifung Poon. A multimodal biomedical foundation model trained from fifteen million image–text pairs. *NEJM AI*, 2(1), 2024. doi: 10.1056/AIoa2400640. URL https://ai.nejm.org/doi/full/10.1056/AIoa2400640.

Xi Zhang, Zaiqiao Meng, Jake Lever, and Edmond S. L. Ho. Ccd: Mitigating hallucinations in radiology mllms via clinical contrastive decoding. In *Findings of the Association for Computational Linguistics: ACL*, pp. 35174–35200, 2026. doi: 10.18653/v1/2026.findings-acl.1755.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*, 2023. URL https://arxiv.org/abs/2305.10415.

Theodore Zhao, Yu Gu, Jianwei Yang, et al. A foundation model for joint segmentation, detection and recognition of biomedical objects across nine modalities. *Nature Methods*, 22:166–176, 2025. doi: 10.1038/s41592-024-02499-w.

A QUALITATIVE REPORT-GENERATION EXAMPLE

Figure 4 illustrates MERIT on a representative MIMIC-CXR example. The greedy baseline generates a pleural effusion and additional unsupported etiological content, whereas the MERIT output more closely matches the reference finding. The example also illustrates the distinction between a specialist's native output and its effect on the receiver: the classifier contributes a coarse score, while the committed output reflects the evidence-conditioned receiver prediction rather than directly reproducing that score. This example is qualitative and is not used as an aggregate measure of clinical correctness.

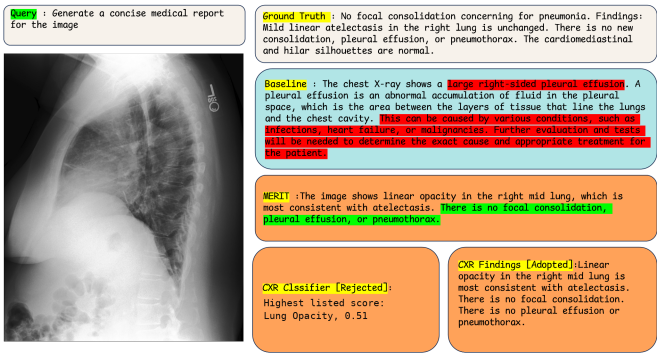

Figure 4: **Qualitative MIMIC-CXR example.** The greedy baseline generates findings unsupported by the reference (red), whereas the MERIT output more closely matches the reference finding (green). The classifier value is a native expert score rather than a calibrated probability or a directly generated answer.