CONTEXT-SPACE ADAPTATION OF FROZEN MULTIMODAL LARGE LANGUAGE MODELS FOR MEDICAL IMAGE DIAGNOSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models (MLLMs) offer a promising paradigm for medical image diagnosis, yet their performance can degrade substantially when applied to medical datasets with heterogeneous visual distributions and diagnostic tasks. Fine-tuning can improve domain adaptation, but requires repeated model updates for each target dataset. In this work, we investigate gradient-free adaptation of frozen MLLMs through the construction of dataset-specific context states. We propose Clinician Mimetic Workflow, a context-space adaptation framework that adapts frozen MLLMs by jointly constructing complementary visual and experiential context states. The framework comprises two coordinated context construction processes. Discriminative Exemplar Coreset Selection (DECS) organizes representative diagnostic cases in a frozen DINOv2 feature space and refines their adaptive retrieval keys to construct the visual context state. Self-Refined Experience Summarization (SRES) distills reusable diagnostic experience from diverse reasoning trajectories and maintains it through adaptive experience-state updates to construct the experiential context state. During inference, the visual exemplars provided by DECS and the diagnostic experience accumulated by SRES are jointly incorporated into the multimodal context, enabling the frozen MLLM to perform dataset-specific diagnostic reasoning without updating pretrained model parameters. We systematically evaluate the proposed framework on a subset of the public MedMNIST 2D datasets and a private nasopharyngeal carcinoma MRI dataset. Experimental results show that our method consistently improves over representative gradient-free in-context learning methods and achieves competitive diagnostic performance on real-world clinical data.

1 INTRODUCTION

Multimodal large language models (MLLMs) have shown rapidly growing capabilities in medical image understanding and diagnostic reasoning. Models such as LLaVA-Med, Med-Flamingo, Med-PaLM M, and BiomedGPT have extended multimodal language models to biomedical visual question answering, image interpretation, diagnostic reasoning, and general-purpose biomedical analysis (Li et al., 2023; Moor et al., 2023; Tu et al., 2024; Zhang et al., 2024). At the same time, medical imaging data are inherently heterogeneous. Variations across institutions, scanners, acquisition protocols, and patient populations can induce substantial distribution shifts and consequently affect cross-dataset generalization (Yoon et al., 2024). Enabling an MLLM with general medical capabilities to adapt effectively to different target datasets therefore remains an important problem for practical medical applications.

A common solution is to adapt model parameters to each target task or dataset. Supervised fine-tuning can encode target-domain information into model parameters (Li et al., 2023; Zhang et al., 2024), while parameter-efficient approaches reduce the optimization cost through low-rank updates or lightweight trainable components (Hu et al., 2021). Nevertheless, when a model is expected to serve multiple medical datasets, such approaches still require separate gradient-based optimization for different target distributions, increasing computational cost as well as model maintenance and deployment complexity. This motivates a different question: can a frozen MLLM acquire dataset-

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.

specific diagnostic capability by organizing target-dataset knowledge entirely in its context, without modifying pretrained parameters?

In-context learning (ICL) provides a natural mechanism for such adaptation, allowing a model to adjust its predictions according to task information and demonstrations provided in the input context (Brown et al., 2020). For visual tasks, demonstration selection has a substantial influence on ICL performance, and retrieval based on pretrained visual representations generally provides more informative examples than random selection (Zhang et al., 2023). Ferber et al. (2024) further demonstrated this mechanism for medical pathology classification, showing that query-specific visual retrieval can substantially improve the diagnostic performance of MLLMs. In parallel, Self-Refine (Madaan et al., 2023) and Reflexion (Shinn et al., 2023) have shown that reasoning trajectories and feedback generated during previous attempts can be distilled into information that improves subsequent reasoning. Together, these directions reveal two complementary forms of target-dataset knowledge: representative diagnostic cases that characterize its visual distribution, and reusable diagnostic experience extracted from prior reasoning. Existing studies have largely developed these two forms of context independently, leaving open how they can be organized into a unified and reusable dataset-specific context state.

Our contributions are summarized as follows:

- **Context-space adaptation.** We formulate dataset adaptation of frozen MLLMs for medical image classification as the construction of dataset-specific context states, shifting adaptation from parameter space to a reusable context space composed of complementary visual and experiential knowledge.
- **Coordinated visual and experiential context construction.** We propose Clinician Mimetic Workflow, in which DECS constructs a discriminative visual context state and SRES distills reusable diagnostic experience into an experiential context state. The two processes are coordinated through representative diagnostic cases and jointly support gradient-free adaptation of frozen MLLMs.
- **Evaluation across heterogeneous medical imaging datasets.** We systematically evaluate the framework on a subset of the public MedMNIST 2D datasets (Yang et al., 2023) and a private nasopharyngeal carcinoma MRI dataset, demonstrating consistent improvements over representative gradient-free in-context learning methods across heterogeneous imaging modalities and a real-world clinical setting.

2 Related Work

2.1 Medical Multimodal Models for Image Diagnosis

Medical multimodal foundation models have substantially expanded the ability of large language models to interpret biomedical images. LLaVA-Med and Med-Flamingo explored biomedical visual instruction tuning and medical multimodal few-shot learning, respectively (Li et al., 2023; Moor et al., 2023), while BiomedCLIP established general-purpose biomedical vision-language representations through large-scale contrastive pretraining (Zhang et al., 2025). Generalist models such as Med-PaLM M and BiomedGPT subsequently extended multimodal modeling to heterogeneous medical modalities and task families (Tu et al., 2024; Zhang et al., 2024), whereas specialized systems including PathChat, CheXagent, and RadFM further strengthened domain-specific capabilities in pathology and radiology (Lu et al., 2024; Chen et al., 2024; Wu et al., 2025). Despite these advances, model performance remains sensitive to imaging modality, clinical dataset, and task definition (Yoon et al., 2024). Such heterogeneity makes effective use of target-dataset information an important component of adapting general medical MLLMs to specific diagnostic settings.

2.2 In-Context Adaptation through Exemplar Selection and Retrieval

The effectiveness of visual in-context learning depends strongly on the demonstrations presented to the model. Zhang et al. (2023) systematically studied visual exemplar selection and showed that retrieval using pretrained visual representations can substantially outperform random demonstration selection, with supervised retrievers providing further gains in demonstration utility. In medical

imaging, Ferber et al. (2024) demonstrated that MLLMs can perform cancer pathology classification through ICL and found query-specific nearest-neighbor retrieval in visual embedding space to be markedly more effective than random demonstrations. Subsequent studies have explored increasingly informative strategies for context construction. APT and AICL investigate task-aware or active exemplar acquisition (Kandiyana et al., 2025; Wang et al., 2025), while RAICL constructs query-conditioned contexts by retrieving relevant demonstrations from an external case pool (Zhan et al., 2026). Collectively, these studies show a progression from random demonstrations toward retrieval-based and actively constructed visual contexts. However, the retrieved context is generally determined primarily by the current query and does not explicitly maintain a reusable dataset-specific context state across cases.

2.3 Context Refinement and Experiential Memory

A complementary line of research improves model reasoning by refining generated outputs or retaining information from previous reasoning processes. Self-Refine iteratively revises model outputs using self-generated feedback (Madaan et al., 2023), while Reflexion converts trajectory feedback into natural-language reflections that are stored in episodic memory for subsequent attempts (Shinn et al., 2023). In multimodal settings, reflection and iterative correction have likewise been used to improve visual reasoning and grounding (Cheng et al., 2025; Liao et al., 2025). CIRCLE further improves multimodal in-context classification by iteratively refining pseudo-labeled context examples (Garosi et al., 2026).

More recent memory-augmented methods extend this idea from local refinement toward reusable external states that persist across examples. Memory-of-Thought stores prior reasoning thoughts for retrieval on future problems (Li & Qiu, 2023), Agent Workflow Memory abstracts historical trajectories into reusable workflows (Wang et al., 2024), and A-MEM maintains structured memories that evolve through continued interaction (Xu et al., 2026). Notes to Self further extracts experiential abstractions, such as strategies and cautions, from solution traces for reuse in subsequent reasoning (Liu et al., 2026). These studies collectively indicate a broader transition from instance-specific correction toward explicit and reusable external experiential states.

Existing work has therefore developed two complementary mechanisms for context-based adaptation: selecting informative visual demonstrations and accumulating reusable reasoning experience. Yet these forms of knowledge are typically modeled separately. In contrast, **Clinician Mimetic Workflow** treats them as complementary components of a single dataset-specific context state. DECS constructs a discriminative visual context state from representative diagnostic cases, while SRES builds an experiential context state from reusable diagnostic experience distilled from reasoning trajectories. Their coordination enables target-dataset knowledge to be organized in context space and reused by a frozen MLLM for medical image diagnosis without gradient-based parameter adaptation.

3 Method

3.1 Context-Space Adaptation Framework

Given a target medical image classification dataset $\mathcal{D}$, let $\mathcal{D}_{\mathrm{tr}} = \{(x_i, y_i)\}_{i=1}^{N}$ denote its training split. We adapt a frozen MLLM without updating its pretrained parameters by constructing a reusable dataset-specific context state:

$$\mathcal{A}_{\mathcal{D}} = (\mathcal{V}_{\mathcal{D}}, \mathcal{E}_{\mathcal{D}}), \tag{1}$$

where $\mathcal{V}_{\mathcal{D}}$ is a visual context state composed of representative cases and their adaptive retrieval keys, and $\mathcal{E}_{\mathcal{D}}$ is an experiential context state distilled from diagnostic reasoning over $\mathcal{D}_{\mathrm{tr}}$. Both context states are constructed exclusively from $\mathcal{D}_{\mathrm{tr}}$; validation and test samples are not used during context construction. Both the visual encoder and the MLLM remain frozen throughout adaptation.

The two states encode target-dataset knowledge at different levels: the visual state preserves instance-level appearance and local distribution structure, whereas the experiential state captures reusable diagnostic regularities abstracted across cases. This separation allows retrieved exemplars to provide query-specific visual evidence while the Experience Bank supplies cross-case reasoning cues that are not tied to any individual demonstration.

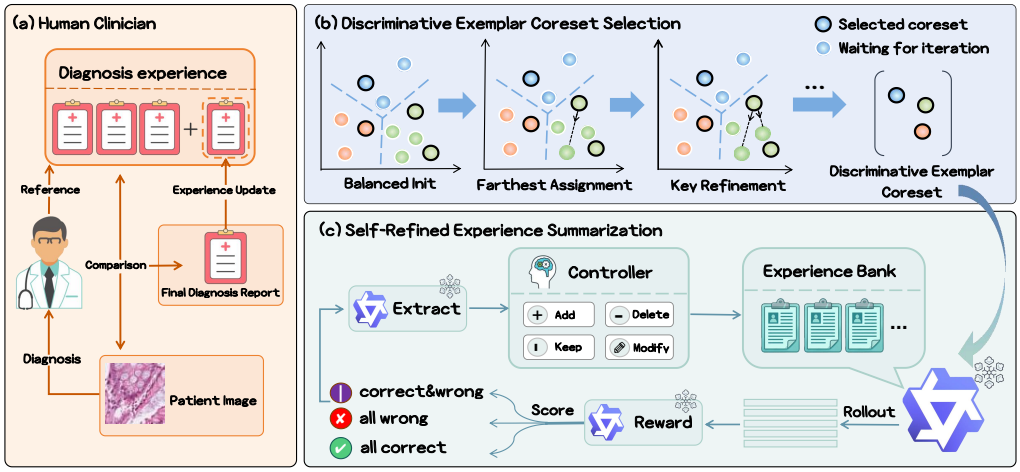

Figure 1: Overview of the proposed **Clinician Mimetic Workflow**. DECS constructs a dataset-specific visual context from representative labeled cases, while SRES distills reusable diagnostic experience through iterative Experience Bank updates. The two context states jointly condition the frozen MLLM at inference. Snowflakes denote frozen pretrained models.

Clinician Mimetic Workflow constructs these two states in a coordinated manner. Discriminative Exemplar Coreset Selection (DECS) builds a compact visual context and supplies representative cases for Self-Refined Experience Summarization (SRES), which converts diagnostic reasoning over these cases into a reusable Experience Bank. At inference, query-specific visual exemplars and the fixed Experience Bank jointly condition the frozen MLLM, as illustrated in Figure 1.

3.2 Discriminative Exemplar Coreset Selection

DECS constructs a compact class-balanced exemplar bank in a frozen feature space. Each exemplar is paired with an adaptive retrieval key initialized from its DINOv2 feature and subsequently refined using complementary same-class training samples. The real exemplar image-label pairs remain fixed, while their retrieval keys are adapted to support exemplar retrieval and experience construction.

Frozen Feature Extraction and Exemplar Initialization. For each $(x_i, y_i) \in \mathcal{D}_{\text{tr}}$, a frozen DINOv2 encoder (Oquab et al., 2023) extracts an ℓ_2-normalized feature $\mathbf{f}_i$. We initialize a class-balanced exemplar bank

$$\mathcal{C} = \{(x_{e_j}, y_{e_j}, \mathbf{p}_j)\}_{j=1}^{S},$$

where (x_{e_j}, y_{e_j}) is a real labeled training case and $\mathbf{p}_j$ is its associated adaptive retrieval key, initialized from the corresponding DINOv2 feature.

To accommodate datasets of different sizes, the exemplar-bank size S and refinement budget T_{opt} are adaptively scaled as

$$S = \min\left(N, \max\left(C, \left\lfloor B_{\text{size}}\sqrt{\frac{N}{N_{\text{ref}}}}\right\rfloor\right)\right), \qquad T_{\text{opt}} = \max\left(1, \left\lfloor B_{\text{epoch}}\sqrt{\frac{N_{\text{ref}}}{N}}\right\rfloor\right), \tag{2}$$

where N and C are the numbers of training samples and classes in $\mathcal{D}_{\text{tr}}$, respectively, and B_{size}, B_{epoch}, and N_{ref} are base scaling parameters.

Hard-Target Refinement. To avoid repeatedly reinforcing already similar visual patterns and encourage each retrieval key to absorb complementary intra-class information, each training sample is assigned to the least similar same-class retrieval key:

$$a(i) = \arg\min_{j:y_{e_j}=y_i} \langle \mathbf{f}_i, \mathbf{p}_j \rangle . \tag{3}$$

Let $\mathcal{Q}_j = \{i \mid a(i) = j\}$ denote the samples assigned to the j-th exemplar slot. For non-empty $\mathcal{Q}_j$, the retrieval key is updated by

$$\mathbf{p}_j \leftarrow \mathrm{norm}\left[(1-\alpha)\mathbf{p}_j + \alpha \frac{1}{|\mathcal{Q}_j|} \sum_{i \in \mathcal{Q}_j} \mathbf{f}_i\right], \tag{4}$$

where α is the update rate and $\mathrm{norm}(\cdot)$ denotes ℓ_2 normalization. Empty slots remain unchanged, and refinement is repeated for T_{opt} epochs. The least-similar assignment encourages each key to integrate class-consistent information across heterogeneous appearances rather than repeatedly reinforce visual patterns already well represented by that key.

Only the retrieval key $\mathbf{p}_j$ is refined; the associated real exemplar image x_{e_j} and label y_{e_j} remain fixed. After refinement, $\mathbf{p}_j$ is not intended to reproduce the visual embedding of x_{e_j} itself. Instead, it serves as an adaptive retrieval key associated with that exemplar, integrating complementary same-class information from $\mathcal{D}_{\mathrm{tr}}$. Query-key similarity is therefore used to determine which fixed real labeled exemplars are presented to the MLLM. The resulting exemplar bank constitutes $\mathcal{V}_{\mathcal{D}}$. For each query, DECS retrieves the Top-K cases according to cosine similarity with the refined keys, and the same bank is reused to organize SRES experience construction.

3.3 Self-Refined Experience Summarization

SRES converts reasoning outcomes over DECS-selected cases into reusable diagnostic experience and maintains them as an explicit experiential context state.

Experience Case Organization. Rather than independently resampling the full training set, SRES organizes DECS exemplars into class-balanced experience rounds, each containing at most one previously unprocessed case from each class. For the t-th experience case x_t, DECS-retrieved exemplars $\mathcal{C}_K(x_t)$ and the current Experience Bank $\mathcal{B}^{(t)}$ form the multimodal context

$$P_t = f_{\mathrm{vis}}\left(x_t, \mathcal{C}_K(x_t)\right) \oplus f_{\mathrm{text}}\left(\mathcal{B}^{(t)}\right), \tag{5}$$

where the current experience case itself is excluded from the retrieval pool when constructing $\mathcal{C}_K(x_t)$ to prevent label leakage. The ground-truth label y_t is not exposed during rollout generation.

Reward-Guided Reasoning Rollouts. Given P_t, the frozen MLLM generates G diverse reasoning trajectories $\mathcal{R}_t = \{r_{t,g}\}_{g=1}^{G}$. Each trajectory is evaluated according to five criteria: diagnostic correctness, validity, evidence quality, differential reasoning, and consistency. These criteria correspond to the `is_correct`, `valid`, `evidence`, `differential`, and `consistency` fields shown in Figure 2, respectively. Let $\Omega = \{\text{correctness}, \text{validity}, \text{evidence}, \text{differential}, \text{consistency}\}$.

The trajectory score is computed as

$$s_{t,g} = \frac{1}{|\Omega|} \sum_{\omega \in \Omega} M_{\mathrm{eval}}\left(\omega \mid r_{t,g}, P_t, y_t\right), \tag{6}$$

where Ω denotes the evaluation criteria. The ground-truth label y_t is used only by the evaluator, separating diagnostic generation from supervised assessment.

Contrastive Experience Summarization. SRES contrasts higher- and lower-scoring trajectories to identify diagnostic observations and reasoning patterns associated with successful or erroneous decisions. Their critical semantic differences are summarized into a candidate heuristic h_t intended for reuse across subsequent cases rather than case-specific correction, as illustrated in Figure 2.

Experience-State Transition and Saturation. Instead of append-only accumulation, each candidate heuristic is integrated through an explicit state transition:

$$\mathcal{T} : \left(\mathcal{B}^{(t)}, h_t\right) \mapsto \left(a_t, \mathcal{B}^{(t+1)}\right), \qquad a_t \in \{\mathrm{Add}, \mathrm{Delete}, \mathrm{Keep}, \mathrm{Modify}\}. \tag{7}$$

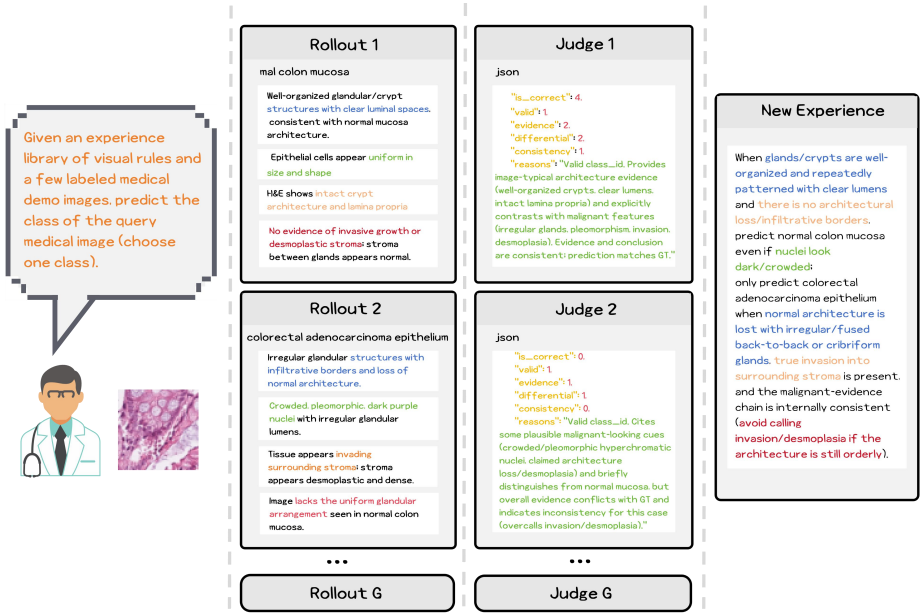

Figure 2: Illustration of **Self-Refined Experience Summarization**. Higher- and lower-quality reasoning trajectories are contrasted to synthesize reusable diagnostic heuristics. Matching colors indicate the semantic evidence linking trajectory differences to the resulting experience.

Add introduces new experience, Delete removes redundant or unreliable information, Keep preserves the current state, and Modify revises existing experience. This mechanism allows the bank to consolidate new evidence while controlling redundancy and inconsistency.

Experience construction terminates when a complete class-balanced round induces no state changes, i.e., $\Delta_r = \sum_{t \in \mathcal{G}_r} \mathbb{I}[a_t \neq \text{Keep}] = 0$, or when all DECS exemplars have been processed once. The final Experience Bank $\mathcal{B}^\star$ defines the experiential context state $\mathcal{E}_\mathcal{D}$.

3.4 Inference

After context construction, the MLLM, DECS exemplar bank, and Experience Bank remain fixed. For a test query x_q, its normalized DINOv2 representation $\mathbf{f}_q$ is used to retrieve

$$\mathcal{C}_K(x_q) = \text{TopK}_{(x_{e_j}, y_{e_j}, \mathbf{p}_j) \in \mathcal{C}} \langle \mathbf{f}_q, \mathbf{p}_j \rangle . \tag{8}$$

The retrieved real images and labels are combined with the fixed Experience Bank $\mathcal{B}^\star$ to form the query context, and the frozen MLLM produces the final prediction.

No reasoning rollouts, trajectory evaluation, retrieval-key refinement, or Experience Bank updates are performed during testing. Dataset adaptation is therefore carried entirely by the reusable visual and experiential context states constructed before inference.

4 Experiments

4.1 Datasets

We first evaluate the proposed context-space adaptation framework on a subset of the public MedMNIST 2D benchmark (Yang et al., 2023). The selected datasets consist of single-label medical image classification tasks spanning multiple imaging modalities, including pathology, dermoscopy, optical coherence tomography (OCT), chest X-ray, fundus imaging, ultrasound, and CT. They also cover diverse class structures and visual distributions. We use the 224×224 image version provided by MedMNIST and follow the official training, validation, and test splits.

Table 1: Classification accuracy on 11 MedMNIST 2D datasets. The best result in each dataset is shown in bold.

Method	Path MNIST	Derma MNIST	OCT MNIST	Pneumonia MNIST	Retina MNIST	Breast MNIST
Zero-Shot	0.347	0.426	0.422	0.766	0.426	0.500
Ferber	0.807	0.754	0.590	0.833	0.426	0.783
RAICL	0.884	0.747	0.559	0.812	0.365	0.737
CIRCLE	0.731	0.687	0.496	0.794	0.338	0.684
Ours	**0.905**	**0.819**	**0.703**	**0.851**	**0.513**	**0.814**

Method	Blood MNIST	Tissue MNIST	OrganA MNIST	OrganC MNIST	OrganS MNIST	Mean
Zero-Shot	0.267	0.306	0.206	0.221	0.203	0.372
Ferber	0.766	0.395	0.565	0.515	0.483	0.629
RAICL	0.847	0.461	0.740	0.701	0.623	0.680
CIRCLE	0.724	0.371	0.556	0.510	0.465	0.578
Ours	**0.912**	**0.535**	**0.830**	**0.770**	**0.682**	**0.758**

We further evaluate the framework on a private nasopharyngeal carcinoma (NPC) MRI dataset using patient-level Stage III/IV classification under the seventh-edition AJCC/UICC staging system. The dataset is split at the patient level into training, validation, and independent test sets at a ratio of 60%/20%/20%.

4.2 Implementation Details

All experiments are implemented with PyTorch and vLLM and conducted on a single NVIDIA RTX 6000 GPU. We use Qwen3-VL-8B as the MLLM backbone and keep all pretrained model parameters frozen throughout adaptation and testing. DECS uses a frozen DINOv2 encoder for visual feature extraction, with no gradient updates applied to the visual encoder.

For DECS, we use a batch size of 128 and set the retrieval-key update rate to $\alpha = 0.2$. We set $N_{\text{ref}} = 1000$, $B_{\text{size}} = 50$, and $B_{\text{epoch}} = 10$ for adaptive exemplar-bank scaling. At inference time, $K = 5$ visual exemplars are retrieved for each query. For SRES, $G = 5$ reasoning trajectories are generated for each experience case with a sampling temperature of 0.8. Both trajectory evaluation and experience-state transition are implemented using the same frozen Qwen3-VL-8B backbone with separately designed task-specific prompts. No additional trainable models are introduced.

Accuracy (ACC) is used as the primary evaluation metric for both the public datasets and the private NPC dataset. For the NPC dataset, we additionally report per-class recall for Stage III and Stage IV to characterize class-specific performance. Overall ACC is computed over all NPC patients, whereas Stage III/IV recall denotes the proportion of correctly classified patients within the corresponding stage. We report the mean results over five independent runs.

For a fair comparison among gradient-free methods, all approaches use the same frozen MLLM, dataset splits, label verbalization, output parsing, and test configuration. Methods involving visual demonstrations use the same frozen DINOv2 encoder and an identical exemplar budget of $K = 5$. Under this controlled reproduction setting, RAICL performs query-specific Top-K visual retrieval in the DINOv2 feature space. The DINOv2-kNN ablation therefore uses the same query-specific retrieval configuration.

4.3 Comparison Experiments

We first compare different gradient-free in-context learning approaches on public medical image classification tasks. Table 1 reports the results on the evaluated subset of MedMNIST 2D. Our method achieves the highest ACC on every evaluated dataset and reaches a mean accuracy of 0.758, exceeding the strongest competing method, RAICL, which obtains 0.680, by 7.8 percentage points.

Table 2: Overall classification accuracy and per-class recall on the private nasopharyngeal carcinoma MRI dataset. Best results are shown in bold.

Method	Stage III Recall	Stage IV Recall	Overall ACC
Zero-Shot	0.594	0.194	0.551
Ferber	0.771	0.264	0.716
RAICL	0.736	0.318	0.691
CIRCLE	0.719	0.234	0.667
Ours	**0.831**	**0.458**	**0.791**

Table 3: Ablation of visual and experiential context states. Best results are shown in bold.

Visual Context	Experiential Context	MedMNIST Mean ACC	Stage III Recall	Stage IV Recall
None	None	0.372	0.594	0.194
Random ICL	None	0.556	0.665	0.229
DINOv2-kNN	None	0.680	0.736	0.318
DECS	None	0.710	0.763	0.374
DECS	SRES	**0.758**	**0.831**	**0.458**

Importantly, the improvement is not driven by only one or two datasets. Relative to RAICL, our method improves OCTMNIST from 0.559 to 0.703 and RetinaMNIST from 0.365 to 0.513, corresponding to gains of 14.4 and 14.8 percentage points, respectively. It also improves OrganAMNIST from 0.740 to 0.830, while retaining a gain on PathMNIST, where the competing methods already achieve relatively high accuracy. These results indicate that the benefit of the constructed context states persists across substantially different visual distributions and imaging modalities rather than being concentrated in a particular task family.

We further evaluate all methods on the private NPC MRI dataset. As shown in Table 2, our method achieves an overall ACC of 0.791, compared with 0.716 for the strongest baseline, corresponding to a 7.5 percentage-point improvement. The gain is observed for both staging categories: Stage III recall reaches 0.831, while Stage IV recall increases to 0.458. In particular, compared with the strongest baseline Stage IV recall of 0.318 from RAICL, our method improves Stage IV recall by 14.0 percentage points. The improvement on the lower-performing Stage IV category suggests that jointly incorporating retrieved visual evidence and dataset-level diagnostic experience can provide useful complementary information for this challenging subgroup.

4.4 Ablation Studies

We conduct two sets of ablation studies to examine the respective contributions of the visual and experiential context states, as well as the design choices used to construct the SRES experience state.

Table 3 first decomposes the contribution of visual context construction. Random ICL improves the MedMNIST mean ACC from 0.372 to 0.556, confirming that labeled visual demonstrations already provide substantial task information to the frozen MLLM. Replacing random demonstrations with query-specific DINOv2 retrieval further raises performance to 0.680, a gain of 12.4 percentage points, demonstrating the importance of visually relevant exemplar selection.

DECS further improves the MedMNIST mean from 0.680 to 0.710 under the same exemplar budget and frozen DINOv2 feature space. The improvement is also observed on the NPC dataset, where Stage III and Stage IV recall increase from 0.736/0.318 to 0.763/0.374. Since DINOv2-kNN and DECS share the same visual encoder and query-specific retrieval setting, these gains isolate the benefit of refining the retrieval keys into a dataset-specific visual context rather than relying directly on the pretrained feature geometry.

Table 4: Ablation of experience-state construction in SRES. Best results are shown in bold.

Experience Cases	Bank Update	Stopping	MedMNIST Mean ACC	Stage III Recall	Stage IV Recall
DECS	Append-only	Full sweep	0.518	0.613	0.204
DECS	State transition	Full sweep	0.742	0.820	0.408
Random	State transition	Saturation	0.714	0.767	0.363
DECS	State transition	Saturation	**0.758**	**0.831**	**0.458**

Adding SRES on top of DECS yields a further improvement from 0.710 to 0.758 on MedMNIST. On NPC, Stage III recall increases from 0.763 to 0.831, while Stage IV recall increases from 0.374 to 0.458. The consistent gains after fixing the DECS visual context indicate that the Experience Bank contributes information complementary to retrieved exemplars rather than merely duplicating the effect of visual retrieval.

We next examine how the experiential context itself should be constructed. As shown in Table 4, replacing append-only accumulation with state-transition updating increases MedMNIST mean ACC from 0.518 to 0.742. Similar improvements are observed on NPC, where Stage III and Stage IV recall increase from 0.613/0.204 to 0.820/0.408. This large gap indicates that simply retaining all newly generated heuristics is insufficient: effective experiential context requires consolidation and revision of previously accumulated information. The Add/Delete/Keep/Modify transition provides an explicit mechanism for removing redundancy and integrating potentially overlapping or conflicting diagnostic experience.

Representative experience cases also matter. With state-transition updating and saturation-based stopping fixed, replacing randomly selected experience cases with DECS exemplars improves MedMNIST mean ACC from 0.714 to 0.758 and raises Stage IV recall from 0.363 to 0.458. This supports using the same representative visual structure for both exemplar retrieval and experience acquisition.

Finally, saturation-based stopping improves over processing the full exemplar sweep under the same DECS cases and state-transition mechanism, increasing the MedMNIST mean from 0.742 to 0.758 and Stage IV recall from 0.408 to 0.458. Thus, the absence of Experience Bank updates over a complete class-balanced round provides a practical signal for early stopping, avoiding unnecessary processing once no further state change is observed over the current round.

5 Limitations

Our approach relies on labeled target data for exemplar construction and trajectory evaluation. Trajectory evaluation and experience-state transition use the same frozen Qwen3-VL-8B backbone, which may introduce correlated model biases; stronger or heterogeneous evaluators may further improve experience construction. Finally, our evaluation focuses primarily on medical image classification; extending context-space adaptation to 3D imaging and more complex generative clinical tasks remains an important direction for future work.

6 Conclusion

We introduced Clinician Mimetic Workflow, a gradient-free context-space adaptation framework in which DECS constructs a discriminative visual context and SRES distills reusable diagnostic experience for frozen MLLMs. Experiments on MedMNIST and a private NPC MRI dataset show consistent gains over representative gradient-free ICL baselines, with ablations confirming the complementary contributions of the two context states. These results position context-space adaptation as a practical alternative to repeated parameter updates for dataset-specific medical image diagnosis.

Reproducibility Statement

To facilitate reproducibility, we provide an anonymous code repository containing the implementation of DECS and SRES as well as the experimental pipelines used in this work. Further implementation details, algorithmic procedures, dataset descriptions, and baseline reproduction settings are provided in the appendix.

AI use statement

After completing the initial draft of this paper, we used generative AI tools to assist with language polishing, clarity, and manuscript formatting. Generative AI tools were not used for research ideation, hypothesis formulation, experimental design, data analysis, interpretation of results, or drawing scientific conclusions. Multimodal large language models (MLLMs) used within the proposed framework constitute part of the research methodology and are therefore distinct from the use of generative AI tools for manuscript preparation. All AI-assisted edits were reviewed by the authors, who take full responsibility for the final content of the paper.

References

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

Zhihong Chen, Maya Varma, Justin Xu, Magdalini Paschali, Dave Van Veen, Andrew Johnston, Alaa Youssef, Louis Blankemeier, Christian Bluethgen, Stephan Altmayer, et al. A vision-language foundation model to enhance efficiency of chest x-ray interpretation. *arXiv preprint arXiv:2401.12208*, 2024.

Kanzhi Cheng, Li YanTao, Fangzhi Xu, Jianbing Zhang, Hao Zhou, and Yang Liu. Vision-language models can self-improve reasoning via reflection. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 8876–8892, 2025.

Dyke Ferber, Georg Wölflein, Isabella C Wiest, Marta Ligero, Srividhya Sainath, Narmin Ghaffari Laleh, Omar SM El Nahhas, Gustav Müller-Franzes, Dirk Jäger, Daniel Truhn, et al. In-context learning enables multimodal large language models to classify cancer pathology images. *Nature Communications*, 15(1):10104, 2024.

Marco Garosi, Matteo Farina, Alessandro Conti, Massimiliano Mancini, and Elisa Ricci. Large multimodal models as general in-context classifiers. *arXiv preprint arXiv:2602.23229*, 2026.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Abhiram Kandiyana, Peter R Mouton, Yaroslav Kolinko, Lawrence O Hall, and Dmitry Goldgof. Active prompt tuning enables gpt-4o to do efficient classification of microscopy images. In *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*, pp. 01–05. IEEE, 2025.

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

Xiaonan Li and Xipeng Qiu. Mot: Memory-of-thought enables chatgpt to self-improve. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6354–6374, 2023.

Yuan-Hong Liao, Rafid Mahmood, Sanja Fidler, and David Acuna. Can large vision-language models correct semantic grounding errors by themselves? In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14667–14678. IEEE, 2025.

Chang Liu, Xinyu Li, and Artur Dubrawski. Notes to self: Can llms benefit from experiential abstractions? *arXiv preprint arXiv:2607.20372*, 2026.

Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Melissa Zhao, Aaron K Chow, Kenji Ikemura, Ahrong Kim, Dimitra Pouli, Ankush Patel, et al. A multimodal generative ai copilot for human pathology. *Nature*, 634(8033):466–473, 2024.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in neural information processing systems*, 36:46534–46594, 2023.

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

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems*, 36:8634–8652, 2023.

Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, et al. Towards generalist biomedical ai. *Nejm Ai*, 1(3):AIoa2300138, 2024.

Fan Wang, Zhongyi Han, and Yilong Yin. Empowering multimodal models via active in-context learning for test-time medical imaging. In *2025 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 4135–4140. IEEE, 2025.

Zora Zhiruo Wang, Jiayuan Mao, Daniel Fried, and Graham Neubig. Agent workflow memory. *arXiv preprint arXiv:2409.07429*, 2024.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Hui Hui, Yanfeng Wang, and Weidi Xie. Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. *Nature Communications*, 16(1):7866, 2025.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-mem: Agentic memory for llm agents. *Advances in Neural Information Processing Systems*, 38:17577–17604, 2026.

Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific data*, 10(1):41, 2023.

Jee Seok Yoon, Kwanseok Oh, Yooseung Shin, Maciej A Mazurowski, and Heung-Il Suk. Domain generalization for medical image analysis: A review. *Proceedings of the IEEE*, 112(10):1583–1609, 2024.

Zaifu Zhan, Shuang Zhou, Xiaoshan Zhou, Yongkang Xiao, Jun Wang, Jiawen Deng, He Zhu, Yu Hou, Yiran Song, Mingquan Lin, et al. Retrieval-augmented in-context learning for multimodal large language models in disease classification. *Journal of Biomedical Informatics*, pp. 105017, 2026.

Kai Zhang, Rong Zhou, Eashan Adhikarla, Zhiling Yan, Yixin Liu, Jun Yu, Zhengliang Liu, Xun Chen, Brian D Davison, Hui Ren, et al. A generalist vision–language foundation model for diverse biomedical tasks. *Nature medicine*, 30(11):3129–3141, 2024.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. A multimodal biomedical foundation model trained from fifteen million image–text pairs. *Nejm Ai*, 2(1):AIoa2400640, 2025.

Yuanhan Zhang, Kaiyang Zhou, and Ziwei Liu. What makes good examples for visual in-context learning? *Advances in Neural Information Processing Systems*, 36:17773–17794, 2023.

A ALGORITHMIC PROCEDURES

For completeness of implementation, Algorithms 1 and 2 summarize the execution procedures of DECS and SRES, respectively. They correspond directly to the components described in Section 3.2 and Section 3.3.

Algorithm 1 Discriminative Exemplar Coreset Selection (DECS)

Require: Training split $\mathcal{D}_{\mathrm{tr}} = \{(x_i, y_i)\}_{i=1}^{N}$; frozen visual encoder ϕ; number of classes C; $B_{\mathrm{size}}, B_{\mathrm{epoch}}, N_{\mathrm{ref}}$; update rate α
Ensure: Exemplar bank $\mathcal{C}$ with refined retrieval keys

Extract and ℓ_2-normalize visual features $\mathbf{f}_i \leftarrow \mathrm{norm}(\phi(x_i))$ for all $x_i \in \mathcal{D}_{\mathrm{tr}}$
$S \leftarrow \min\Big(N, \max\Big(C, \Big\lfloor B_{\mathrm{size}}\sqrt{N/N_{\mathrm{ref}}}\Big\rfloor\Big)\Big)$
$T_{\mathrm{opt}} \leftarrow \max\Big(1, \Big\lfloor B_{\mathrm{epoch}}\sqrt{N_{\mathrm{ref}}/N}\Big\rfloor\Big)$
Initialize a class-balanced exemplar bank $\mathcal{C} = \{(x_{e_j}, y_{e_j}, \mathbf{p}_j)\}_{j=1}^{S}$, with $\mathbf{p}_j \leftarrow \mathbf{f}_{e_j}$
for $\tau = 1$ to T_{opt} **do**
    Initialize $\mathcal{Q}_j \leftarrow \emptyset$ for all $j \in \{1, \ldots, S\}$
    for each $(x_i, y_i) \in \mathcal{D}_{\mathrm{tr}}$ **do**
        $a(i) \leftarrow \arg\min_{j: y_{e_j} = y_i} \langle \mathbf{f}_i, \mathbf{p}_j \rangle$
        $\mathcal{Q}_{a(i)} \leftarrow \mathcal{Q}_{a(i)} \cup \{i\}$
    end for
    for $j = 1$ to S **do**
        if $\mathcal{Q}_j \neq \emptyset$ **then**
            $\mathbf{p}_j \leftarrow \mathrm{norm}\Big[(1-\alpha)\mathbf{p}_j + \frac{\alpha}{|\mathcal{Q}_j|}\sum_{i \in \mathcal{Q}_j} \mathbf{f}_i\Big]$
        end if
    end for
end for
return $\mathcal{C} = \{(x_{e_j}, y_{e_j}, \mathbf{p}_j)\}_{j=1}^{S}$

Algorithm 2 Self-Refined Experience Summarization (SRES)

Require: DECS exemplar bank $\mathcal{C}$; frozen MLLM M; number of retrieved exemplars K; number of reasoning trajectories G
Ensure: Final Experience Bank $\mathcal{B}^{\star}$

Organize DECS exemplars into class-balanced experience rounds $\{\mathcal{G}_r\}_{r=1}^{R}$
Initialize Experience Bank $\mathcal{B}^{(0)} \leftarrow \emptyset$ and set $t \leftarrow 0$
for each experience round $\mathcal{G}_r$ **do**
    $\Delta_r \leftarrow 0$
    for each experience case $(x_t, y_t) \in \mathcal{G}_r$ **do**
        Retrieve Top-K exemplars $\mathcal{C}_K(x_t)$ from $\mathcal{C}$ using the DECS retrieval keys, excluding the current experience case
        Construct $P_t \leftarrow f_{\mathrm{vis}}(x_t, \mathcal{C}_K(x_t)) \oplus f_{\mathrm{text}}\left(\mathcal{B}^{(t)}\right)$
        Generate $\mathcal{R}_t = \{r_{t,g}\}_{g=1}^{G}$ using M without exposing y_t
        for $g = 1$ to G **do**
            $s_{t,g} \leftarrow \dfrac{1}{|\Omega|}\displaystyle\sum_{\omega \in \Omega} M_{\mathrm{eval}}\left(\omega \mid r_{t,g}, P_t, y_t\right)$
        end for
        Contrast higher- and lower-scoring trajectories and summarize their critical differences into a candidate diagnostic heuristic h_t
        $(a_t, \mathcal{B}^{(t+1)}) \leftarrow \mathcal{T}(\mathcal{B}^{(t)}, h_t), a_t \in \{\mathrm{Add}, \mathrm{Delete}, \mathrm{Keep}, \mathrm{Modify}\}$
        if $a_t \neq \mathrm{Keep}$ **then**
            $\Delta_r \leftarrow \Delta_r + 1$
        end if
        $t \leftarrow t + 1$
    end for
    if $\Delta_r = 0$ **then**
        break
    end if
end for
$\mathcal{B}^{\star} \leftarrow \mathcal{B}^{(t)}$
return $\mathcal{B}^{\star}$

B Additional Dataset Details

The private nasopharyngeal carcinoma (NPC) MRI dataset is used for patient-level Stage III/IV classification under the seventh edition of the AJCC/UICC staging system. All data partitions are constructed at the patient level, ensuring that images from the same patient do not appear across different splits. The dataset is divided into training, validation, and independent test sets with a ratio of 60%/20%/20%.

Distinguishing Stage III from Stage IV NPC is clinically important because the two stages reflect different extents of advanced disease and are associated with different levels of disease burden and clinical risk. This distinction contributes to patient risk stratification, treatment planning, and prognosis assessment. From a modeling perspective, the task therefore requires the model to capture subtle but clinically meaningful imaging differences related to disease extent, rather than merely separating broad disease categories, providing a practical setting for evaluating dataset-specific context adaptation in fine-grained medical image classification.

All data used in this study were de-identified prior to analysis. No personally identifiable information was available to the authors or provided to the models.

C Baseline Reproduction

All baseline methods are reimplemented according to their original methodological descriptions. To ensure a controlled comparison, components shared across methods are standardized wherever applicable. Specifically, all methods use the same Qwen3-VL-8B backbone, identical dataset splits, label verbalization, output parsing, and test protocol. Methods involving visual retrieval use the same frozen DINOv2 encoder, and the number of retrieved visual exemplars is fixed to $K = 5$. Method-specific adaptation procedures are otherwise preserved.

For RAICL, replacing its backbone and visual encoder with the common Qwen3-VL-8B and DINOv2 configuration results in query-specific Top-K nearest-neighbor retrieval in the DINOv2 feature space. This retrieval procedure is identical to the DINOv2-kNN configuration used in our ablation study. We therefore report it as RAICL in the main comparison, following the baseline method, and as DINOv2-kNN in the ablation study to explicitly denote the retrieval-only configuration.

D Analysis of Experience-State Dynamics

To further characterize the behavior of SRES, we analyze the Experience Bank produced in one representative run. These statistics are intended to reveal the internal dynamics of experience-state construction and are complementary to the five-run averaged performance results reported in the main text.

D.1 Experience Bank Size

Figure 3 compares the final number of experience entries retained by append-only accumulation and state-transition updating. Append-only accumulation continuously preserves newly synthesized experience, whereas the proposed state-transition mechanism allows the existing bank to be updated through Add, Delete, Keep, and Modify operations.

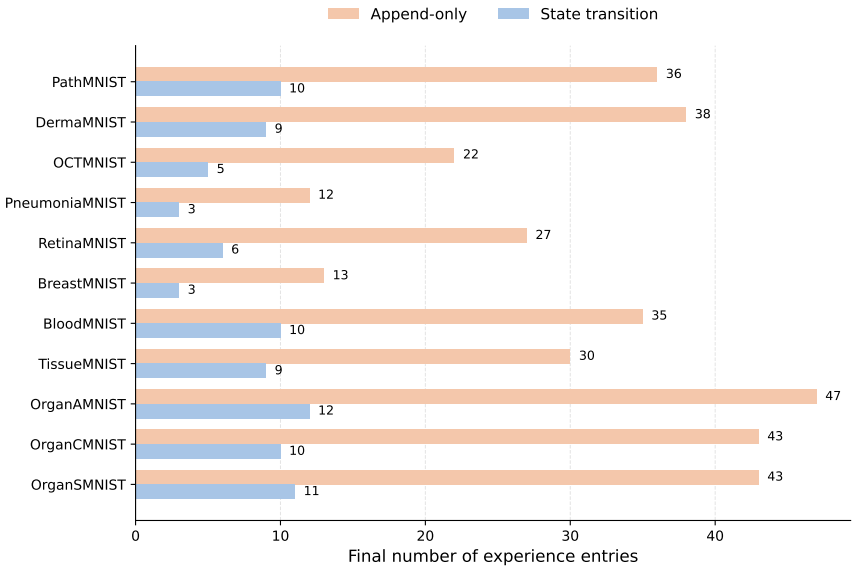

Figure 3: Final number of Experience Bank entries under append-only accumulation and state-transition updating in one representative run.

Across the evaluated MedMNIST datasets, append-only accumulation consistently produces larger Experience Banks than state-transition updating. Together with the performance results in Table 4, this observation indicates that the improvement obtained by state-transition updating is not attributable to retaining a larger number of experience entries. Instead, SRES maintains a substantially more compact experience state while achieving higher diagnostic accuracy, consistent with its objective of integrating and revising accumulated diagnostic experience rather than continuously appending new summaries.

D.2 State-Transition Statistics

Table 5 reports the frequency of the four state-transition actions in the same representative run. All four operations are observed across every evaluated dataset, indicating that the Experience Bank undergoes both structural and content-level updates during construction.

Table 5: Experience-state transition statistics from one representative run. Final entries denote the number of experience entries retained after state construction.

Dataset	Final Entries	Add	Delete	Keep	Modify	Total Actions
PathMNIST	10	12	2	9	8	31
DermaMNIST	9	12	3	8	8	31
OCTMNIST	5	7	2	5	4	18
PneumoniaMNIST	3	4	1	3	2	10
RetinaMNIST	6	8	2	6	6	22
BreastMNIST	3	4	1	3	3	11
BloodMNIST	10	12	2	8	7	29
TissueMNIST	9	11	2	8	6	27
OrganAMNIST	12	15	3	12	10	40
OrganCMNIST	10	13	3	11	9	36
OrganSMNIST	11	14	3	11	10	38
Total	**88**	**112**	**24**	**84**	**73**	**293**

Across all evaluated datasets, SRES performs 112 Add, 24 Delete, 84 Keep, and 73 Modify operations. Add is the most frequent action, reflecting the continued acquisition of new diagnostic experience, while Keep and Modify are also frequently invoked as the bank evaluates whether newly synthesized heuristics should preserve or revise the current state. Delete occurs less frequently but is consistently observed across datasets, allowing previously retained information to be removed when it becomes redundant or unsuitable.

The transition statistics are consistent with the resulting Experience Bank sizes. Since the bank is initialized as empty, Add increases the number of entries, Delete decreases it, and Keep and Modify leave the number of entries unchanged. Accordingly, the 112 Add and 24 Delete operations result in 88 retained entries. These results characterize SRES as a dynamically maintained experience state rather than an append-only collection of independently generated heuristics.