KNOWLEDGE-GUIDED ON-POLICY SELF DISTILLATION FOR MEDICAL MLLMS CONTINUAL LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Medical multimodal large language models (MLLMs) need continuous adaptation to emerging clinical tasks while preserving previously acquired capabilities, yet collecting high-quality trajectories and designing explicit rewards remain difficult. On-policy self distillation (OPSD) enables continual learning by supervising the model on trajectories generated via its own evolving policy without complex reward design, but the effectiveness depends on the quality of teacher-side privileged signal. In this work, we observe that OPSD-based medical MLLMs continual learning is not simply to provide more medical context, but to construct privileged knowledge conditions that yield reliable teacher supervision on uncertain student-generated trajectories. With this insight, we propose knowledge-guided on-policy self distillation for medical MLLMs continual learning (MedKgSD), a reward-free and external-teacher-free framework that uses biomedical knowledge graphs to construct training-time privileged knowledge conditions for on-policy self-distillation. For each new task, MedKgSD builds a knowledge graph induced task-level concept space and derives sample-specific privileged conditions from biomedical knowledge relations to interpret student-visited states under structured medical context. Since knowledge-derived conditions could be noisy or mismatched, we introduce token-selective entropy calibration to identify reliable teacher signals on informative student states and distills the teacher behavior along student-generated trajectories. This converts structured medical relations into dense on-policy supervision for new-task adaptation, with prior-policy distillation reducing the forgetting of previous capabilities. Experiments across multiple medical backbones demonstrate that MedKgSD improves the adaptation-retention trade-off in continual medical adaptation.

1 INTRODUCTION

Medical multimodal large language models (MLLMs) have made substantial advances in clinical reasoning over heterogeneous medical inputs, including medical images, reports, and diagnostic questions. Recent medical post-training studies show that domain-specific supervision and policy optimization can improve medical visual question answering, report understanding, and diagnostic reasoning (Dai et al., 2026; Lai et al., 2026; Jiang et al., 2026). However, real clinical deployment is rarely limited to a fixed set of tasks, and new clinical tasks, annotation protocols as well as data distributions may continue to emerge after training, thereby requiring medical MLLMs to acquire new capabilities while preserving previously learned medical ability. Existing post-training paradigms only partially satisfy this continual learning requirement, since supervised fine-tuning (Chu et al., 2025; Liu et al., 2026a) relies on high-quality answers or expert trajectories that are costly to collect and tied to fixed offline data. Reinforcement learning (Szepesvári, 2022; Yu et al., 2026; Wei et al., 2026) provides on-policy adaptation but requires explicit medical rewards that are difficult to design and validate across evolving tasks.

A promising direction is on-policy distillation (OPD), which supervises trajectories sampled from the student policy with dense token-level guidance from a teacher, avoiding exclusive reliance on offline expert demonstrations or complex reward function (Agarwal et al., 2024; Gu et al., 2024; Chen et al., 2026). Recent work further extends this idea to on-policy self-distillation (OPSD), where a same-backbone teacher provides supervision by conditioning on privileged training-time information, such as verified solutions or self-generated targets (Zhao et al., 2026; Hübotter et al.,

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.

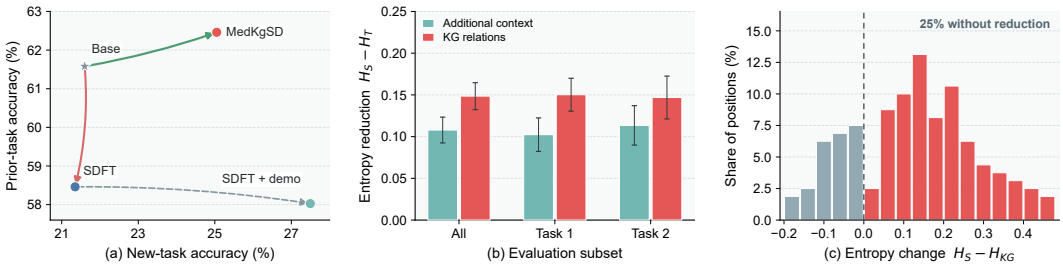

Figure 1: Empirical motivation for MedKgSD. (a) SDFT with training-only demonstrations from Qwen-3.5-122B-A10B improves new-task accuracy but lowers prior-task retention. (b) KG-conditioned teachers reduce entropy more than unstructured medical context at uncertain student positions. (c) KG conditioning yields no entropy reduction at 25% of sampled positions.

2026), demonstrations (Shenfeld et al., 2026), retrieved or contextual prompts (Ye et al., 2026), and privileged supporting contexts (Stein et al., 2026). This formulation reduces dependence on larger external teacher models, but shifts the central bottleneck from reward design to the construction of high-quality privileged signals. Existing privileged signals are often tied to samples, prompts, or documents, and recent analyses show that distillation can degrade when the teacher signal is poorly matched, weakly informative, or overconfident (Li et al., 2026; Zhu et al., 2026; Kim et al., 2026).

This work is motivated by the observation that effective teacher-side supervision for uncertain student trajectories requires reusable privileged knowledge rather than merely additional medical context. As shown in Fig. 1(a), training-only demonstrations improve new-task adaptation but compromise prior-task retention. To construct such privileged knowledge, we use biomedical knowledge graphs (KG), which organize diseases, phenotypes, anatomical entities, mechanisms, and treatments through structured clinical relations. As shown in Fig. 1(b), KG-conditioned teachers reduce uncertainty more effectively than unstructured medical context at uncertain student positions. Accordingly, KG-derived knowledge serves as training-time privileged conditions that induce same-backbone teacher policies for self distillation, rather than as deploy-time retrieved evidence for the student. Fig. 1(c) shows that KG conditioning fails to reduce uncertainty at a substantial fraction of sampled token positions, motivating token-selective calibration of the resulting teacher signal. Inspired from these observations, we propose knowledge-guided on-policy self distillation for Medical MLLMs continual Learning (MedKgSD), a reward-free and external-teacher-free framework that constructs KG-induced task-level concept spaces, derives sample-specific privileged conditions, and calibrates teacher signals using token-selective teacher-policy entropy. This calibration suppresses unreliable supervision when KG contexts are noisy or mismatched, while retaining informative teacher signals at certain positions where privileged knowledge reduces uncertainty. The student is trained on its own trajectories to the calibrated self-teacher, while prior-policy distillation preserves compatibility with previous model. In summary, our main contributions are as follows:

- We propose MedKgSD, a reward-free and external-teacher-free framework that exploits structured biomedical relations as privileged teacher-side conditions, using task-conditioned KG contexts to induce a same-backbone teacher and transform relational knowledge into dense on-policy supervision for uncertain student-generated trajectories.
- To address noisy or mismatched KG contexts, we introduce token-selective entropy calibration (TSE), which identifies student-visited positions where KG conditioning reduces uncertainty and uses them to weight candidate teacher contexts.
- Extensive experiments across multiple backbones show that MedKgSD improves the adaptation-retention trade-off over supervised fine-tuning, RL-based optimization, and self distillation fine-tuning for continual adaptation on medical tasks.

2 Related Work

2.1 Medical MLLM Post-Training

Medical MLLMs require post-training to adapt general vision-language backbones to clinical images, physiological signals, reports, and diagnostic question answering. Recent medical post-

training works, including QoQ-Med (Dai et al., 2026), Med-R1 (Lai et al., 2026), MedTVT-R1 (Zhang et al., 2026b), and MedVLThinker (Huang et al., 2025), further enhance medical reasoning, modality understanding, and diagnostic reliability through clinical instruction tuning and reinforcement-style reasoning optimization. Despite these advances, existing objectives remain largely tied to offline answer imitation or task-specific reward construction, which becomes less attractive in continual learning where each new clinical task requires both adaptation and retention. In contrast, on-policy self distillation provides a reward-free alternative by supervising the model on trajectories induced by its own current policy (Shenfeld et al., 2026; Liu et al., 2026b; Zhao et al., 2026).

2.2 On-Policy Self-Distillation with Privileged Context

On-policy self distillation improves model adaption by supervising trajectories sampled from its own current policy, thereby reducing the off-policy mismatch of supervised fine-tuning without explicit reward design. Recent works, including SDFT (Shenfeld et al., 2026), COPSD (Liu et al., 2026b), Self-Distilled Reasoner (Zhao et al., 2026), OPCD (Ye et al., 2026), and π-Play (Zhang et al., 2026a), show that a same-scale or self-derived teacher can provide dense token-level guidance when conditioned on demonstrations, verified traces, contextual prompts, or self-generated artifacts. These methods turn otherwise sparse, static, or unavailable supervision into feedback on student-visited states, but their effectiveness depends on whether the privileged context provides reliable information gain and remains compatible with the student policy (Li et al., 2026; Zhu et al., 2026; Jin et al., 2026). This limitation is more pronounced in medical continual learning, where the teacher signal should reflect reusable clinical relations rather than task-local hints or brittle prompt conditions. Therefore, MedKgSD grounds the privileged teacher context in structured biomedical knowledge and calibrates it with teacher-policy uncertainty, enabling the self-teacher to provide domain-relevant supervision on student-generated trajectories.

2.3 Biomedical Knowledge Graphs for Reasoning

Biomedical knowledge graphs (Chandak et al., 2023; Bodenreider, 2004) organize clinical and biological concepts through structured relations among diseases, phenotypes, anatomical entities, molecular mechanisms, drugs, and treatments. Recent knowledge-enhanced medical AI works use such graphs to support entity normalization, clinical retrieval, medical question answering, report understanding, and knowledge-grounded reasoning (Rezaei et al., 2025; Yan et al., 2024; Wang et al., 2025; Jiang et al., 2025). These methods show that structured knowledge can provide reusable domain priors beyond isolated task labels, but most of them treat the graph as an external retrieval source, static feature augmentation, or additional input context (Zhao et al., 2025; Jia et al., 2025; Chen et al., 2025). This leaves underexplored how biomedical knowledge can be transformed into a training-time supervision signal for continual medical MLLM adaptation. Therefore, MedKgSD uses structured biomedical relations to induce a knowledge-conditioned self-teacher and distills its calibrated behavior into the student on self-generated trajectories.

3 Method

3.1 Problem Formulation

We study continuous learning for medical MLLMs. The model receives a sequence of medical tasks $\mathcal{D}_1, \ldots, \mathcal{D}_T$, where each task $\mathcal{D}_t = \{(x_i, q_i, a_i)\}$ consists of a multimodal clinical input x_i, a question or instruction q_i, and an answer a_i. At stage t, the model needs adapt to $\mathcal{D}_t$ without erasing capabilities acquired from $\mathcal{D}_{<t}$. Let π_θ denote the current student policy. Supervised fine-tuning optimizes the likelihood of offline answers $\log \pi_\theta(a_i \mid x_i, q_i)$, where the supervision trajectory is fixed within the dataset rather than induced by the current model. Reinforcement learning provides an on-policy alternative, but requires a reward $r(x, q, y)$ that is difficult to define, validate, and scale in medical domains. We seek an on-policy self distillation signal induced by structured biomedical knowledge, without external teacher models or explicit reward engineering.

At stage t, the post-training problem can be viewed as updating $\pi_{\theta_{t-1}}$ to π_{θ_t}. Let $\bar{\theta}$ denote a frozen or temporally smoothed teacher copy, $\mathcal{C}_t$ denote a task-level KG concept space, $c_{i,k}$ denote a candidate context for sample i, y_i denote a student-generated trajectory, and $s_{i,j} = (x_i, q_i, y_{i,<j})$ denote

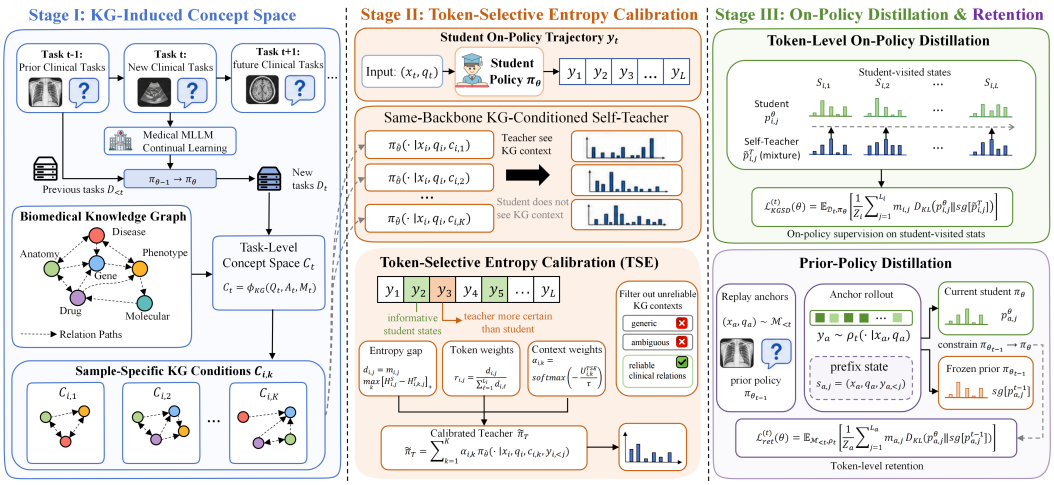

Figure 2: Overview of MedKgSD. The student first generates on-policy trajectories without access of KG, while the same backbone is reused as a teacher under multiple KG-derived privileged contexts. Token-selective entropy calibration weights these teacher policies by their confidence, and the resulting knowledge-conditioned supervision is distilled back to the student together with a prior-policy retention.

a student-visited prefix state. We use $U_{i,k}^{\mathrm{TSE}}$ and $\alpha_{i,k}$ to represent the token-selective teacher uncertainty and the resulting context weight, respectively. Under this notation, the method answers the following formulation-level questions:

- **How is a privileged self-teacher induced? (Sec. 3.2)** MedKgSD constructs KG-derived privileged conditions $c_{i,k} \subset \mathcal{C}_t$ from structured biomedical relations and exposes them merely to same-backbone teacher policies $\pi_{\bar{\theta}}$, while the student receives the original input (x_i, q_i).
- **Where is the teacher signal reliable? (Sec. 3.3)** Token-selective entropy evaluates KG-conditioned teachers on student-visited prefix states $s_{i,j}$, identifies states where structured knowledge reduces student uncertainty, and converts the resulting teacher uncertainty into context weights $\alpha_{i,k}$.
- **How is continual adaptation optimized? (Sec. 3.4)** The student distills the entropy-calibrated teacher distribution on its own generated trajectory, while prior-policy distillation on replay anchors constrains drift from $\pi_{\theta_{t-1}}$ to π_{θ_t}.

3.2 Knowledge-Conditioned Self-Teacher Induction

MedKgSD aims to expose KG-derived privileged knowledge conditions only to the teacher side. In medical continual learning, student trajectories often contain uncertain, incomplete, or partially formed clinical decisions. Conditioning the same-backbone teacher on structured biomedical relations allows these student-visited states to be interpreted under reusable medical semantics, while keeping the deployed student input unchanged. In practice, we instantiate the KG with PrimeKG (Chandak et al., 2023), a large-scale precision-medicine knowledge graph that integrates multiple biomedical resources and provides relation paths across diseases, phenotypes, anatomical entities, biological mechanisms, pathways, drugs, and treatment-related concepts. For an incoming task $\mathcal{D}_t$, we first induce a task-level concept space

$$\mathcal{C}_t = \Phi_{\mathrm{KG}}(Q_t, A_t, M_t), \tag{1}$$

where Q_t denotes the task-level question semantics, A_t denotes the answer or label vocabulary, M_t denotes the medical modality or task description, and Φ_{KG} maps these task cues to KG-derived biomedical concepts and relations. The concept space in Eq. 1 specifies task-relevant medical knowledge, but does not encode the gold answer for any individual sample. For a sample (x_i, q_i), we derive a set of candidate knowledge contexts

$$c_{i,1}, \ldots, c_{i,K} \subset \mathcal{C}_t. \tag{2}$$

Each context in Eq. 2 contains a compact set of biomedical concepts and relation paths. Conditioning the model copy on $c_{i,k}$ induces a context-specific self-teacher policy

$$\pi_{\bar{\theta}}(\cdot \mid x_i, q_i, c_{i,k}), \tag{3}$$

where $\bar{\theta}$ denotes a temporally smoothed or frozen copy of the current model. The policy in Eq. 3 receives the privileged KG-derived context, whereas the student policy receives only the standard task input (x_i, q_i). Details of KG concept construction, context selection, and teacher-student prompt are provided in Appendix C.1 and Appendix C.2.

3.3 Token-Selective Entropy Calibration

Not every KG-derived context is equally informative for a given sample, and not every token in a student trajectory is equally useful for judging that context. A concept context may be generic, weakly related to the input evidence, or ambiguous among multiple biomedical entities. Averaging teacher uncertainty over the whole trajectory can dilute the states where the KG-conditioned teacher actually changes the local policy. We calibrate the sample-level knowledge condition with token-selective entropy (TSE) and therefore context score is estimated from student-visited states where the KG-conditioned teacher is more certain than the student under the same prefix. The student first generates an on-policy trajectory

$$y_i = (y_{i,1}, \ldots, y_{i,L_i}) \sim \pi_\theta(\cdot \mid x_i, q_i). \tag{4}$$

For each prefix state, we evaluate both the student and the KG-conditioned teacher conditioned on the same fixed token prefix

$$p^S_{i,j} = \pi_\theta(\cdot \mid x_i, q_i, y_{i,<j}), \qquad p^T_{i,k,j} = \pi_{\bar{\theta}}(\cdot \mid x_i, q_i, c_{i,k}, y_{i,<j}). \tag{5}$$

The student distribution $p^S_{i,j}$ and the teacher distribution $p^T_{i,k,j}$ in Eq. 5 are defined on the same student-generated prefix, while the teacher-side input additionally contains the privileged KG context $c_{i,k}$. We denote their next-token entropies as $H^S_{i,j} = H(p^S_{i,j})$ and $H^T_{i,k,j} = H(p^T_{i,k,j})$, where $H(\cdot)$ denotes next-token entropy. TSE first identifies states where some KG-conditioned teacher becomes more certain than the student. With $m_{i,j} \in \{0, 1\}$ denoting the valid trajectory-token mask, we define

$$d_{i,j} = m_{i,j} \max_k \left[H^S_{i,j} - H^T_{i,k,j}\right]_+, \tag{6}$$

where $[z]_+ = \max(z, 0)$. A large $d_{i,j}$ indicates that at least one KG context induces a sharper teacher distribution at state $s_{i,j}$. We then normalize these scores into token-selection weights:

$$r_{i,j} = \begin{cases} \dfrac{d_{i,j}}{\sum_{\ell=1}^{L_i} d_{i,\ell}}, & \sum_{\ell=1}^{L_i} d_{i,\ell} > 0, \\ \dfrac{m_{i,j}}{\sum_{\ell=1}^{L_i} m_{i,\ell}}, & \text{otherwise.} \end{cases} \tag{7}$$

The fallback in Eq. 7 recovers whole-trajectory averaging when no teacher context reduces student uncertainty on the current trajectory.

For each candidate context, TSE uses the token-weighted teacher entropy $U^{\mathrm{TSE}}_{i,k} = \sum_{j=1}^{L_i} r_{i,j} H^T_{i,k,j}$, which summarizes the teacher uncertainty over the selected token positions. The context score and converts it into a context weight:

$$\alpha_{i,k} = \frac{\exp\Big(-(\sum_{j=1}^{L_i} r_{i,j} H^T_{i,k,j})/\tau\Big)}{\sum_{k'=1}^{K} \exp\Big(-(\sum_{j=1}^{L_i} r_{i,j} H^T_{i,k',j})/\tau\Big)}. \tag{8}$$

A low token-weighted entropy score means that context $c_{i,k}$ induces a more decisive teacher policy on the states where KG conditioning is informative relative to the student. This score estimates the usefulness of the teacher signal, not the factual correctness of the KG context. The temperature τ controls the sharpness of the context weighting. The entropy-calibrated self-teacher is the mixture

$$\tilde{\pi}_T(\cdot \mid x_i, q_i, y_{i,<j}) = \sum_{k=1}^{K} \alpha_{i,k} \pi_{\bar{\theta}}(\cdot \mid x_i, q_i, c_{i,k}, y_{i,<j}). \tag{9}$$

This formulation treats entropy as a policy-level uncertainty measure, not as an entity-linking confidence score. The weights in Eq. 8 emphasize the knowledge conditions rather than make the self-teacher in Eq. 9 more decisive on informative student-visited states. TSE is used to calibrate the context mixture, and it does not restrict the later distillation loss to selected tokens.

3.4 On-Policy Self-Distillation for Continual Learning

The entropy-calibrated mixture in Eq. 9 defines the teacher distribution, but the student should learn from this signal only on states induced by its own policy. Therefore we use the calibrated teacher as an on-policy distillation target and add a prior-preserving constraint for sequential learning. Let $p^{\theta}_{i,j} = \pi_\theta(\cdot \mid x_i, q_i, y_{i,<j})$, $\tilde{p}^{T}_{i,j} = \tilde{\pi}_{\bar{\theta}}(\cdot \mid x_i, q_i, y_{i,<j})$, and $Z_i = \sum_{j=1}^{L_i} m_{i,j}$. The knowledge-guided on-policy self distillation objective is

$$\mathcal{L}^{(t)}_{\mathrm{KGSD}}(\theta) = \mathbb{E}_{\mathcal{D}_t, \pi_\theta}\left[\frac{1}{Z_i}\sum_{j=1}^{L_i} m_{i,j}\, D_{\mathrm{KL}}\big(p^{\theta}_{i,j} \| \,\mathrm{sg}[\tilde{p}^{T}_{i,j}]\big)\right]. \tag{10}$$

Here the expectation is over $(x_i, q_i) \sim \mathcal{D}_t$ and a student trajectory $y_i \sim \mathrm{sg}[\pi_\theta(\cdot \mid x_i, q_i)]$. Unlike answer-level imitation, Eq. 10 transfers how structured biomedical knowledge changes the teacher distribution along student-visited states. The supervision is both on-policy and concept-structured, and it is evaluated on trajectories induced by the current student, while the target behavior is induced by biomedical relations in the KG.

Prior-preserving retention. To support continual learning, we preserve prior capabilities by distilling from the model before the current adaptation stage. Let $\pi_{\theta_{t-1}}$ denote the policy after learning task $t-1$, and $\mathcal{M}_{<t}$ denote a small set of anchor examples from previous tasks. For an anchor input (x_a, q_a), let $p^{\theta}_{a,j} = \pi_\theta(\cdot \mid x_a, q_a, y_{a,<j})$, $p^{t-1}_{a,j} = \pi_{\theta_{t-1}}(\cdot \mid x_a, q_a, y_{a,<j})$, and $Z_a = \sum_{j=1}^{L_a} m_{a,j}$. The prior-preserving objective is

$$\mathcal{L}^{(t)}_{\mathrm{ret}}(\theta) = \mathbb{E}_{\mathcal{M}_{<t}, \rho_t}\left[\frac{1}{Z_a}\sum_{j=1}^{L_a} m_{a,j} D_{\mathrm{KL}}\big(p^{\theta}_{a,j} \| \,\mathrm{sg}[p^{t-1}_{a,j}]\big)\right]. \tag{11}$$

Here the expectation is over replay anchors $(x_a, q_a) \sim \mathcal{M}_{<t}$ and replay trajectories $y_a \sim \rho_t(\cdot \mid x_a, q_a)$. The retention term is enabled only when $t > 1$. The objective in Eq. 11 does not introduce an external teacher. It constrains the current policy to remain compatible with its own previous behavior while the KG-conditioned self-teacher guides adaptation to the new task. The full training objective at stage t is

$$\mathcal{L}^{(t)}(\theta) = \mathcal{L}^{(t)}_{\mathrm{KGSD}}(\theta) + \lambda \mathcal{L}^{(t)}_{\mathrm{ret}}(\theta), \tag{12}$$

where λ is a hyperparameter controlling the contribution of prior-policy preservation during model updates. In Eq. 12, the adaptation signal comes from the entropy-calibrated self-distillation objective in Eq. 10, while Eq. 11 constrains the current policy against its own previous behavior.

4 Experiment

4.1 Experimental Setup

Benchmarks & Models. We evaluate MedKgSD on two medical MLLM backbones, QoQ-Med-7B (Dai et al., 2026) and MedVLThinker-3B (Huang et al., 2025). The continual-learning stream uses OmniMedVQA (Hu et al., 2024) to evaluate previously acquired medical capabilities across modality recognition, anatomy identification, disease diagnosis, lesion grading, and biological attribute recognition. MedXpertQA-MM (Zuo et al., 2025) provides the incoming tasks, organized into two sequential stages of reasoning and understanding. After each adaptation stage, we evaluate the resulting model on both prior and incoming tasks.

Baselines. We compare MedKgSD with the base model and three representative post-training paradigms. Supervised fine-tuning (SFT) directly optimizes the incoming-task demonstrations, RL-based optimization uses the corresponding medical reasoning baseline (Dai et al., 2026), and self distillation fine-tuning (SDFT) (Shenfeld et al., 2026) learns from self-teacher supervision.

Training Details. We train all methods with LoRA under the same two-stage continual-learning protocol. MedKgSD generates on-policy student trajectories and re-scores them with same-backbone teachers conditioned on task-specific PrimeKG contexts. We use one epoch per stage with at most 100 steps, a learning rate of 1×10^{-6}, LoRA rank 64, and bfloat16 precision. Responses are sampled with temperature 1.0 and top-$p = 0.95$, while evaluation uses greedy decoding. All experiments run on 8 NVIDIA A800 GPUs, with additional details in Appendix A.

Table 1: Main continual learning results. Bold values indicate the best result within each stage-specific comparison group, and underlined denote the second-best. Prior Avg calculates both Acc and F1 scores. Δ Prior is relative to the base model.

Method	Prior Tasks										New Tasks				Overall	
	Modality		Anatomy		Disease		Lesion		Biological		Reasoning		Understanding		Prior Avg	Δ Prior
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1		
QoQ-Med-7B																
Stage 1: after reasoning-task adaptation																
Base	**91.4**	**91.4**	63.5	51.4	66.8	**70.3**	36.3	43.2	50.0	56.0	16.6	17.1	19.2	18.1	62.0	–
+ SFT	90.2	90.1	64.1	51.9	65.2	68.8	35.0	41.5	51.2	56.4	16.9	16.6	–	–	61.4	-0.6
+ DRPO	90.8	90.7	65.0	52.6	66.6	69.9	35.2	41.9	51.4	56.6	15.9	16.2	–	–	62.1	+0.1
+ SDFT	91.0	91.1	63.7	51.6	62.9	67.0	32.4	39.0	50.0	55.9	15.9	15.1	–	–	60.5	-1.5
+ MedKgSD	90.2	90.2	**68.4**	**55.5**	**69.5**	70.2	**45.5**	**47.7**	**59.8**	**60.2**	**19.3**	**19.1**	–	–	**65.7**	**+3.7**
Stage 2: after understanding-task adaptation																
Base	**91.4**	**91.4**	63.5	51.4	66.8	**70.3**	36.3	43.2	50.0	56.0	16.6	17.1	19.2	18.1	62.0	–
+ SFT	90.8	91.1	64.1	52.0	63.9	66.9	34.0	40.6	52.5	57.7	14.8	14.5	19.8	18.0	61.4	-0.6
+ DRPO	91.2	91.1	64.8	52.6	65.0	68.5	34.8	41.8	50.0	55.6	17.6	17.1	**20.7**	**19.7**	61.6	-0.4
+ SDFT	90.8	90.7	63.1	51.2	61.1	65.6	33.2	41.2	49.8	55.7	16.2	15.6	19.8	18.5	60.2	-1.8
+ MedKgSD	90.4	90.5	**68.6**	**55.6**	**69.5**	70.1	**45.1**	**47.0**	**61.9**	**62.7**	**20.3**	**20.1**	**20.7**	18.9	**66.1**	**+4.1**
MedVLThinker-3B																
Stage 1: after reasoning-task adaptation																
Base	79.9	80.2	60.5	49.0	60.0	60.0	51.4	51.6	56.1	56.0	15.9	15.4	27.3	26.5	60.5	–
+ SFT	79.5	79.8	61.1	49.2	60.4	60.5	50.0	50.3	**57.2**	**57.2**	17.2	16.9	–	–	60.5	0.0
+ DRPO	80.3	80.6	60.0	48.6	**60.7**	**60.9**	49.0	49.2	54.5	54.5	17.9	17.6	–	–	59.8	-0.7
+ SDFT	80.3	80.6	60.7	49.2	58.0	58.2	51.2	51.3	54.3	54.2	17.2	16.6	–	–	59.8	-0.7
+ MedKgSD	**80.5**	**80.8**	**61.7**	**49.9**	60.2	60.4	**52.1**	**52.4**	56.6	56.6	**18.6**	**18.2**	–	–	**61.1**	**+0.6**
Stage 2: after understanding-task adaptation																
Base	79.9	80.2	60.5	49.0	**60.0**	60.0	51.4	51.6	56.1	56.0	15.9	15.4	27.3	26.5	60.5	–
+ SFT	80.9	81.1	60.3	**49.5**	58.6	58.8	47.9	48.0	56.4	56.4	14.5	13.7	29.1	26.5	59.8	-0.7
+ DRPO	79.7	80.0	59.8	48.6	58.2	58.4	47.7	47.5	52.7	52.7	21.4	20.4	25.5	23.8	58.5	-2.0
+ SDFT	79.7	80.0	50.8	41.0	56.8	56.9	39.5	38.7	53.3	53.4	18.3	18.1	25.5	25.0	55.0	-5.5
+ MedKgSD	**82.6**	**82.8**	**61.3**	48.2	**60.0**	**60.1**	**51.6**	**51.7**	**58.0**	**58.0**	**22.1**	**21.6**	**31.5**	**30.4**	**61.4**	**+0.9**

4.2 Main Results

We first analyze the main continual-learning results in Table 1 and derive three observations. (1) MedKgSD achieves the best adaptation-retention trade-off across both backbones. After the final adaptation stage, MedKgSD improves the prior-task average by 4.1% on QoQ-Med-7B and by 0.9% on MedVLThinker-3B, while SFT, reward-based optimization, and SDFT generally lead to prior-task degradation. (2) The gains are not obtained by self distillation alone. Standard SDFT reduces the prior-task average by 1.8% on QoQ-Med-7B and by 5.5% on MedVLThinker-3B after the second stage, suggesting that privileged trajectories without structured knowledge calibration can still overwrite previously acquired capabilities. (3) MedKgSD remains competitive on incoming tasks while improving retention. It obtains the best reasoning performance on both backbones after sequential adaptation, and achieves the strongest overall new-task performance on MedVLThinker-3B. These results indicate that KG-conditioned self-teacher induction provides useful adaptation signals, while prior-policy distillation prevents the new-task learning process from prior medical capabilities.

4.3 Ablation Study

As shown in Table 2, we evaluate each component by progressively adding OPSD, KG-conditioned teacher context, TSE calibration, and prior-policy retention. Standard OPSD alone brings limited benefit, indicating that self-distillation without structured medical context can still provide weak teacher guidance. Adding KG context improves the adaptation-retention trade-off by grounding the self-teacher in biomedical concept relations. TSE further strengthens this effect by selecting more reliable KG-conditioned teacher states according to teacher-policy uncertainty, while prior-policy retention stabilizes continual learning by preserving previous-task behavior. The full MedKgSD achieves the strongest overall performance, showing that these components are complementary rather than redundant.

4.4 Analysis

Training Dynamics As shown in Fig. 3, we further analyze the training dynamics of MedKgSD. Fig. 3 (a) compared with SDFT, MedKgSD maintains a consistently lower teacher entropy while producing a larger teacher-student KL. This indicates that the KG-conditioned teacher provides

Table 2: Ablation results at the final continual-learning stage. Std. OPSD denotes standard on-policy self distillation and Ret. denotes prior-policy distillation. Prior Avg averages Acc/F1 over prior tasks, while New Avg averages accuracy over new tasks. Underst. denotes Understanding; Δ Prior is relative to the base model.

Variant	Prior Tasks										New Tasks				Overall		
	Modality		Anatomy		Disease		Lesion		Biological		Reasoning		Understanding		Prior Avg	New Avg	Δ Prior
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1			
Base Model	**91.4**	**91.4**	63.5	51.4	66.8	**70.3**	36.3	43.2	50.0	56.0	16.6	17.1	19.2	18.1	62.0	17.9	–
Std. OPSD	90.0	90.8	62.7	50.3	63.1	65.7	35.4	42.2	54.2	59.5	16.9	17.1	20.5	18.4	61.4	18.7	-0.6
Std. OPSD w/ Ret.	90.6	90.8	63.2	51.1	64.4	67.1	36.2	43.0	54.8	60.3	16.8	17.0	20.2	18.2	62.2	18.5	+0.2
MedKgSD w/ KG	90.4	90.3	65.1	54.5	64.6	67.3	39.2	46.4	54.1	58.2	18.6	18.7	19.9	18.1	63.0	19.2	+1.0
MedKgSD w/ TSE	91.0	90.9	67.9	55.1	68.4	69.5	**45.8**	**47.6**	60.3	61.4	18.6	18.4	**21.6**	**19.8**	65.8	20.1	+3.8
MedKgSD w/ Ret.	90.4	90.5	**68.6**	**55.6**	**69.5**	70.1	45.1	47.0	**61.9**	**62.7**	**20.3**	**20.1**	20.7	18.9	**66.1**	**20.5**	**+4.1**

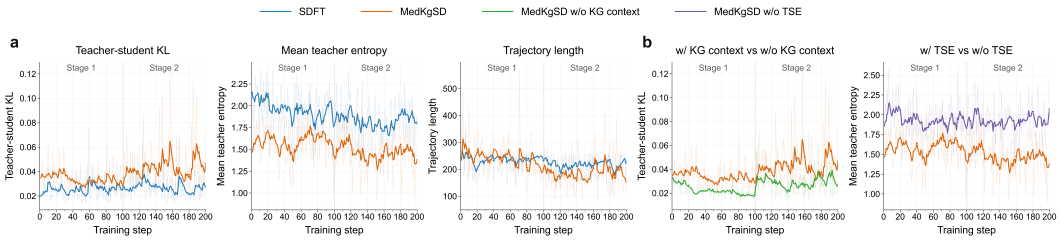

Figure 3: Training-dynamics analysis of MedKgSD. (a) Comparison between SDFT and MedKgSD on teacher-student KL, mean teacher entropy, and generated trajectory length across two continual-learning stages. (b) Component-level dynamics of MedKgSD, comparing models with and without KG context, and with and without TSE calibration. The vertical dashed line separates Stage 1 and Stage 2. Light curves show per-step values, and solid curves show smoothed trends.

a more decisive supervision signal that remains different from the current student policy, rather than collapsing to weak self-imitation. Meanwhile, the trajectory length of MedKgSD stays in a similar range to SDFT, suggesting that the gain is not simply caused by longer responses. Fig. 3 (b) further clarify the roles of KG context and TSE. Removing KG context substantially reduces the teacher-student KL, while removing TSE consistently increases teacher entropy, indicating that structured biomedical context strengthens task-specific guidance and entropy calibration improves the reliability of KG-conditioned teacher states. Together, these dynamics support that KG context strengthens the teacher signal, while TSE improves its reliability during continual adaptation.

Relational knowledge contributes beyond additional context. Table 3 compares teacher-side knowledge conditions after Stage 2. Retrieved text improves Prior Avg from 62.2% to 63.3%, whereas KG conditioning reaches 66.1. Relative to entities alone, KG relations add 1.5% in Prior Avg and 1.2% in reasoning accuracy. Verbalizing the triples retains much of the benefit, reaching 65.0% in Prior Avg, but remains 1.1% below the original KG condition. Rewiring the edges reduces Prior Avg to 61.7%, below the no-context condition. This pattern supports the importance of relational content and correct connections, with a smaller additional benefit from the original KG representation. The advantage is task-dependent: modality accuracy varies little, and the entity-only condition achieves slightly higher understanding accuracy than KG.

Calibration improves the adaptation-retention trade-off. Fig. 4 (a) compares policies for weighting KG-conditioned teachers. TSE reaches 66.14% in Prior Avg and 20.50% in new-task average accuracy, exceeding mean-entropy weightingand maximum-probability weighting. Thus, the observed benefit is not reproduced by either confidence summary alone. Fig. 4 (c) further shows that the largest accuracy gains over SDFT+KG and KG-OPSD occur on anatomy, disease, lesion, and biological-attribute tasks. Modality recognition shows little gain, indicating that KG-conditioned calibration is useful on tasks requiring biomedical distinctions rather than on every visual decision.

Calibration does not eliminate sensitivity to incorrect relations. In Fig. 4(b), increasing the relation mismatch ratio from 0% to 100% reduces TSE's Prior Avg from 66.14% to 62.40% and the new-task average accuracy from 20.50% to 18.90%, showing that corrupted biomedical relations weaken both retention and adaptation. Despite this degradation, TSE remains consistently stronger

Table 3: Effect of teacher-side knowledge context on QoQ-Med-7B after Stage 2. We compare no additional context, retrieved biomedical text, and KG context, together with KG variants that retain only entities, verbalize the triples, or rewire the edges. Acc and F1 are reported for each prior and incoming task. Prior Avg averages Acc and F1 over the five prior tasks.

Teacher context	Prior Tasks										New Tasks				Overall	
	Modality		Anatomy		Disease		Lesion		Biological		Reasoning		Understanding		Prior Avg	Δ Prior
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1		
None	90.6	90.8	63.2	51.1	64.4	67.1	36.2	43.0	54.8	60.3	16.8	17.0	20.2	18.2	62.2	+0.2
Retrieved text	90.6	90.6	65.0	52.2	65.8	68.1	39.1	43.9	56.8	61.0	18.4	18.1	20.5	18.8	63.3	+1.3
KG	90.4	90.5	68.6	55.6	69.5	70.1	45.1	47.0	61.9	62.7	20.3	20.1	20.7	18.9	66.1	**+4.1**
Entities only	90.5	90.7	66.9	53.7	67.3	69.1	42.2	45.4	58.7	61.6	19.1	18.8	21.0	19.0	64.6	+2.6
Verbalized triples	90.6	90.5	66.8	54.7	68.2	69.0	42.4	46.2	60.1	61.6	19.7	19.5	20.4	18.7	65.0	+3.0
Rewired edges	90.7	91.0	62.5	50.7	63.6	66.7	34.7	42.2	54.7	60.4	17.4	17.2	19.4	18.0	61.7	-0.3

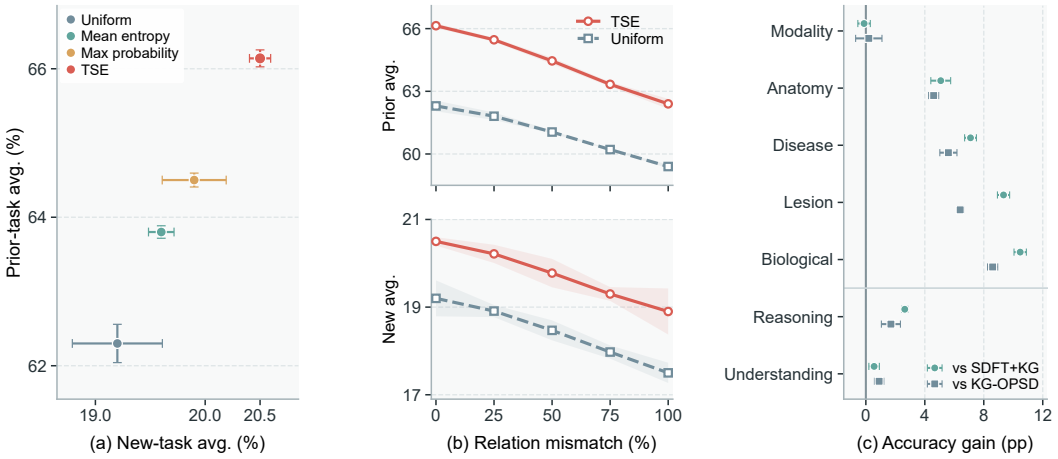

Figure 4: Analyses of KG-conditioned teacher supervision on QoQ-Med-7B. (a) Comparison of teacher-weighting policies in terms of prior-task and new-task performance. (b) Prior-task performance and new-task accuracy under increasing relation mismatch, comparing TSE with uniform weighting. (c) Task-wise accuracy gains of TSE over SDFT+KG and uniformly weighted KG-OPSD under the same KG context. Points and curves show the mean over three seeds, while error bars and shaded bands indicate the standard deviation.

than uniform weighting across all tested mismatch levels. At 100% mismatch, uniform weighting reaches only 59.40% Prior Avg and 17.50% new-task accuracy, while TSE retains advantages of 3.00% and 1.40%. These results indicate that accurate knowledge construction and entropy calibration play complementary roles, with TSE mitigating the effect of imperfect KG contexts.

5 Conclusion

We propose MedKgSD, a knowledge-guided on-policy self distillation framework for medical MLLMs continual learning. Our central idea is to use biomedical knowledge graphs as structured privileged context for inducing a same-backbone self-teacher, allowing external medical priors to be converted into dense token-level supervision on student-generated trajectories. This formulation provides a reward-free and external-teacher-free pathway for continual adaptation, while prior-policy distillation constrains harmful policy drift and helps preserve previously acquired medical capabilities. Across two medical MLLM backbones, MedKgSD consistently improves the adaptation-retention trade-off over supervised fine-tuning, RL-based post-training, and standard self distillation. Further analyses show that structured KG context provides stronger teacher guidance, while TSE and prior-policy retention improve supervision reliability and stabilize continual adaptation. Future work may extend this framework to broader clinical modalities, more heterogeneous knowledge sources, and dynamically updated biomedical knowledge graphs.

AI USE STATEMENT

We used generative AI tools to assist with language polishing and manuscript writing. The authors reviewed and revised the AI-assisted text and take full responsibility for the final content, accuracy, and claims of this paper.

REPRODUCIBILITY STATEMENT

The main paper specifies the knowledge-conditioned teacher construction, token-selective entropy calibration, and continual self-distillation objective. Appendix A describes the data sources, task order, training and evaluation procedures, hardware, and optimization settings, with hyperparameters summarized in Table 4. Appendix A.3 details the replay-anchor source, buffer size, sampling procedure, and prior-policy distillation objective, including the exclusion of held-out evaluation data. Appendix C.1 describes PrimeKG concept selection, the four teacher-side context views, and serialization limits; Appendix C.2 provides the student, teacher, and KG prompt templates. Additional results, training-time measurements, retention-coefficient sensitivity, and stage-wise continual-learning diagnostics are provided in Appendix B.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, volume 2024, pp. 21246–21263, 2024.

Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic Acids Research*, 32(Database issue):D267–D270, jan 2004. doi: 10.1093/nar/gkh061.

Payal Chandak, Kexin Huang, and Marinka Zitnik. Building a knowledge graph to enable precision medicine. *Scientific Data*, 10(67), 2023.

Hanzhu Chen, Xu Shen, Jie Wang, Zehao Wang, Qitan Lv, Junjie He, Rong Wu, Feng Wu, and Jieping Ye. Knowledge graph finetuning enhances knowledge manipulation in large language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *Advances in Neural Information Processing Systems*, 38:57654–57689, 2026.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. *arXiv preprint arXiv:2501.17161*, 2025.

David Dai, Peilin Chen, Chanakya Ekbote, and Paul Liang. Qoq-med: Building multimodal clinical foundation models with domain-aware grpo training. *Advances in Neural Information Processing Systems*, 38:37406–37453, 2026.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. In *International Conference on Learning Representations*, volume 2024, pp. 32694–32717, 2024.

Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22170–22183, June 2024.

Xiaoke Huang, Juncheng Wu, Hui Liu, Xianfeng Tang, and Yuyin Zhou. Medvlthinker: Simple baselines for multimodal medical reasoning. *arXiv preprint arXiv:2508.02669*, 2025.

Jonas Hübotter, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, et al. Reinforcement learning via self-distillation. *arXiv preprint arXiv:2601.20802*, 2026.

Mingyi Jia, Junwen Duan, Yan Song, and Jianxin Wang. medikal: Integrating knowledge graphs as assistants of llms for enhanced clinical diagnosis on emrs. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 9278–9298, 2025.

Pengcheng Jiang, Cao Danica Xiao, Minhao Jiang, Parminder Bhatia, Taha Kass-Hout, Jimeng Sun, and Jiawei Han. Reasoning-enhanced healthcare predictions with knowledge graph community retrieval. In *International Conference on Learning Representations*, volume 2025, pp. 81785–81830, 2025.

Zheng Jiang, Heng Guo, Chengyu Fang, Changchen Xiao, Xinyang Hu, Lifeng Sun, and Minfeng Xu. Medvr: Annotation-free medical visual reasoning via agentic reinforcement learning. *arXiv preprint arXiv:2604.08203*, 2026.

Yiqiao Jin, Yiyang Wang, Lucheng Fu, Yijia Xiao, Yinyi Luo, Haoxin Liu, B Aditya Prakash, Josiah Hester, Jindong Wang, and Srijan Kumar. Unisd: Towards a unified self-distillation framework for large language models. *arXiv preprint arXiv:2605.06597*, 2026.

Jeonghye Kim, Xufang Luo, Minbeom Kim, Sangmook Lee, Dohyung Kim, Jiwon Jeon, Dongsheng Li, and Yuqing Yang. Why does self-distillation (sometimes) degrade the reasoning capability of llms? *arXiv preprint arXiv:2603.24472*, 2026.

Yuxiang Lai, Jike Zhong, Ming Li, Shitian Zhao, Yuheng Li, Konstantinos Psounis, and Xiaofeng Yang. Med-r1: Reinforcement learning for generalizable medical reasoning in vision-language models. *IEEE transactions on medical imaging*, 2026.

Yaxuan Li, Yuxin Zuo, Bingxiang He, Jinqian Zhang, Chaojun Xiao, Cheng Qian, Tianyu Yu, Huanang Gao, Wenkai Yang, Zhiyuan Liu, et al. Rethinking on-policy distillation of large language models: Phenomenology, mechanism, and recipe. *arXiv preprint arXiv:2604.13016*, 2026.

Mingyang Liu, Gabriele Farina, and Asuman Ozdaglar. Uft: Unifying supervised and reinforcement fine-tuning. *Advances in Neural Information Processing Systems*, 38:101347–101383, 2026a.

Yihong Liu, Raoyuan Zhao, Michael A Hedderich, and Hinrich Schütze. Crosslingual on-policy self-distillation for multilingual reasoning. *arXiv preprint arXiv:2605.09548*, 2026b.

Mohammad Reza Rezaei, Reza Saadati Fard, Jayson L Parker, Rahul G Krishnan, and Milad Lankarany. Agentic medical knowledge graphs enhance medical question answering: Bridging the gap between llms and evolving medical knowledge. *arXiv preprint arXiv:2502.13010*, 2025.

Idan Shenfeld, Mehul Damani, Jonas Hübotter, and Pulkit Agrawal. Self-distillation enables continual learning. *arXiv preprint arXiv:2601.19897*, 2026.

Alex Stein, Furong Huang, and Tom Goldstein. Gates: Self-distillation under privileged context with consensus gating. *arXiv preprint arXiv:2602.20574*, 2026.

Csaba Szepesvári. *Algorithms for reinforcement learning*. Springer nature, 2022.

Ziqing Wang, Chengsheng Mao, Xiaole Wen, Yuan Luo, and Kaize Ding. Amanda: Agentic medical knowledge augmentation for data-efficient medical visual question answering. *arXiv preprint arXiv:2510.02328*, 2025.

Yuxiang Wei, Olivier Duchenne, Jade Copet, Quentin Carbonneaux, Lingming Zhang, Daniel Fried, Gabriel Synnaeve, Rishabh Singh, and Sida Wang. Swe-rl: Advancing llm reasoning via reinforcement learning on open software evolution. *Advances in Neural Information Processing Systems*, 38:78500–78525, 2026.

Quan Yan, Junwen Duan, and Jianxin Wang. Multi-modal concept alignment pre-training for generative medical visual question answering. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 5378–5389, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.319.

Tianzhu Ye, Li Dong, Xun Wu, Shaohan Huang, and Furu Wei. On-policy context distillation for language models. *arXiv preprint arXiv:2602.12275*, 2026.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *Advances in Neural Information Processing Systems*, 38:113222–113244, 2026.

Yaocheng Zhang, Yuanheng Zhu, Wenyue Chong, Songjun Tu, Qichao Zhang, Jiajun Chai, Xiaohan Wang, Wei Lin, Guojun Yin, and Dongbin Zhao. pi-play: Multi-agent self-play via privileged self-distillation without external data. *arXiv preprint arXiv:2604.14054*, 2026a.

Yuting Zhang, Kaishen Yuan, Hao Lu, Yutao Yue, Jintai Chen, and Kaishun Wu. Medtvt-r1: A multimodal llm empowering medical reasoning and diagnosis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 35248–35259, 2026b.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. *arXiv preprint arXiv:2601.18734*, 2026.

Xuejiao Zhao, Siyan Liu, Su-Yin Yang, and Chunyan Miao. Medrag: Enhancing retrieval-augmented generation with knowledge graph-elicited reasoning for healthcare copilot. In *Proceedings of the ACM on Web Conference 2025*, pp. 4442–4457, 2025.

Siqi Zhu, Xuyan Ye, Hongyu Lu, Weiye Shi, and Ge Liu. The many faces of on-policy distillation: Pitfalls, mechanisms, and fixes. *arXiv preprint arXiv:2605.11182*, 2026.

Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. *arXiv preprint arXiv:2501.18362*, 2025.

APPENDIX CONTENTS

A Implementation Details **14**

A.1 Data Details 14

A.2 Training and Evaluation Details 14

A.3 Retention Protocol and Replay Anchors 14

B Additional Analysis **15**

B.1 Expanded Main Results by Backbone 16

B.2 Training-Time Analysis 16

B.3 Retention-Coefficient Sensitivity 16

B.4 Additional Continual-Learning Diagnostics 17

B.5 Positive Backward Transfer and Trace-Length Dynamics 18

C Knowledge Context and Prompt Details **19**

C.1 KG Concept Construction and Context Serialization 19

C.2 Prompt Templates 19

D Qualitative Case Studies **21**

D.1 Comparison with SDFT 21

D.2 KG-Grounded Case Trajectory 22

D.3 Boundary Failure Case on Modality Recognition 23

E Limitations and Impact **24**

E.1 Limitations and Future Work 24

E.2 Impact Statement 25

A Implementation Details

A.1 Data Details

The continual-learning stream separates the prior-task evaluation suite from the incoming adaptation tasks. The two data sources have the following roles.

- **OmniMedVQA** (Hu et al., 2024) provides the prior-task evaluation suite. We evaluate modality recognition, anatomy identification, disease diagnosis, lesion grading, and biological attribute recognition to measure how adaptation changes existing medical visual question-answering capabilities. These tasks are evaluation-only: their examples are excluded from adaptation, replay, and KG context construction.
- **MedXpertQA-MM** (Zuo et al., 2025) supplies the incoming medical VQA tasks. We first adapt to the reasoning subset and then to the understanding subset. Training examples from the completed reasoning stage provide replay anchors for the understanding stage; held-out evaluation examples are excluded from replay and KG construction.

A.2 Training and Evaluation Details

Training setup. We implement MedKgSD in PyTorch on a single node with eight NVIDIA A800 GPUs (80GB each). Training uses LoRA adapters, DeepSpeed ZeRO-3, bfloat16 precision, and AdamW, without weight quantization. The student receives the original medical VQA input. KG contexts condition only the same-backbone teacher policies during training and are absent from the student input at inference.

On-policy optimization. At each training step, the current student policy first samples an on-policy response from the student prompt. The same generated trajectory is then re-scored by same-backbone teacher policies conditioned on KG-derived contexts. Teacher distributions are mixed with token-selective entropy calibration, so lower-uncertainty teacher states contribute more reliable token-level supervision. The student is optimized on its own generated tokens using the KG-conditioned self distillation loss. For continual retention, replay anchors from completed incoming adaptation tasks are additionally constrained by prior-policy distillation, which penalizes harmful drift from the stage-start policy. All student rollouts are used for training; we do not filter trajectories by answer correctness.

Hyperparameters. Table 4 summarizes the shared configuration across backbones and method variants. Unless otherwise stated, we use one optimization epoch per stage, a fixed maximum number of training steps per stage, sampling temperature 1.0, and nucleus sampling with top-$p = 0.95$ for on-policy trajectory generation. The same evaluation protocol is used for all compared methods.

Evaluation. After each adaptation stage, we evaluate the resulting checkpoint on all prior OmniMedVQA tasks and on the current MedXpertQA-MM task. We report accuracy and macro-F1 for individual tasks, together with prior-task average, new-task average, and the change in prior-task performance relative to the base checkpoint. For multiple-choice outputs, predictions are normalized to the option label before scoring. Evaluation uses greedy decoding to avoid introducing sampling variance into the continual-learning matrix.

A.3 Retention Protocol and Replay Anchors

We separate the retention protocol from the prior-task evaluation suite to avoid train–test contamination. OmniMedVQA is used only for prior-task evaluation in our two-stage stream. Each prior task is marked as evaluation-only and has no adaptation split in the continual-learning stream. These examples are therefore not treated as trainable stages and are never inserted into the replay memory. The incoming MedXpertQA-MM tasks provide the only training samples for adaptation and replay.

Replay memory is populated only after an incoming task has been completed. Before the first adaptation stage, the replay buffer is empty, so the retention term is inactive. After the MedXpertQA-MM reasoning stage, the buffer stores a deterministic bounded subset of reasoning training samples,

Hyperparameter	Value
Task stream and hardware	
Backbones	QoQ-Med-7B; MedVLThinker-3B
Prior-task evaluation stream	OmniMedVQA modality, anatomy, disease, lesion, and biological-attribute tasks
Incoming continual tasks	MedXpertQA-MM Reasoning → MedXpertQA-MM Understanding
Continual stages	2
GPU	8×NVIDIA A800 (80GB)
Optimization	
Distributed training	DeepSpeed-Zero3 with LoRA adapters
Precision	bfloat16
Quantization	None
Optimizer	AdamW
Learning rate	1×10^{-6}
Weight decay	0.01
Optimization epochs per stage	1
Max steps per stage	100
Per-rank train batch size	2
Training temperature	1.0
Training top-p	0.95
Max training response length	4096 tokens
Evaluation decoding	Greedy
Max evaluation response length	8192 tokens
LoRA adapter	
LoRA rank	64
LoRA α	128
LoRA dropout	0.05
LoRA target modules	`q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, `down_proj`

Table 4: Experiment details for MedKgSD. The table reports the shared training protocol unless otherwise stated. KG contexts are used only to induce teacher-side policies during training.

using the fixed seed and the same replay implementation for all retention-enabled variants. During the subsequent understanding stage, replay anchors are sampled uniformly from this buffer. In our default setting, the buffer stores at most 256 samples per completed task and draws a replay mini-batch of 8 anchors when the retention term is evaluated. Held-out evaluation examples from MedXpertQA-MM and all OmniMedVQA prior-task examples are excluded from replay.

The retention loss is policy distillation, not supervised replay. For each replay anchor, the current student first samples an on-policy trajectory. We then evaluate both the current policy and the frozen stage-start prior policy on the same generated tokens and minimize the masked forward KL, $\mathrm{KL}(\pi_{\theta_t} \| \pi_{\theta_{t-1}})$, over the generated response positions. This computation does not use gold answer labels, answer correctness, KG context, or an external teacher. The gold label may be present in the dataset schema for evaluation, but it is not read by the retention objective.

The same replay protocol is used by all retention-enabled variants. Variants without retention disable both replay sampling and the prior-policy KL term. Thus baselines or ablations that do not include retention are not given hidden replay supervision, while retention-enabled ablations use the same anchor source, sampling seed, buffer size, replay batch size, and prior-policy objective.

B Additional Analysis

We report final-stage results, training cost, retention sensitivity, and diagnostics derived from the stage-wise evaluation matrix. We also examine how changes in answer correctness relate to reasoning-trace length.

Table 5: Expanded final-stage continual-learning results by backbone. The table follows the main result layout but reports only the Stage 2 checkpoint after understanding-task adaptation. Underst. denotes Understanding; Δ Prior is relative to the base model.

Backbone	Method	Prior Tasks										New Tasks				Overall	
		Modality		Anatomy		Disease		Lesion		Biological		Reasoning		Underst.		Prior Avg	Δ Prior
		Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1		
QoQ-Med -7B	Base	**91.4**	**91.4**	63.5	51.4	66.8	**70.3**	36.3	43.2	50.0	56.0	16.6	17.1	19.2	18.1	62.0	–
	+ SFT	90.8	91.1	64.1	52.0	63.9	66.9	34.0	40.6	52.5	57.7	14.8	14.5	19.8	18.0	61.4	-0.6
	+ DRPO	91.2	91.1	64.8	52.6	65.0	68.5	34.8	41.8	50.0	55.6	17.6	17.1	20.7	19.7	61.6	-0.4
	+ SDFT	90.8	90.7	63.1	51.2	61.1	65.6	33.2	41.2	49.8	55.7	16.2	15.6	19.8	18.5	60.2	-1.8
	+ Std. OPSD	90.0	90.8	62.7	50.3	63.1	65.7	35.4	42.2	54.2	59.5	16.9	17.1	20.5	18.4	61.4	-0.6
	+ SDFT+demo	89.8	90.0	62.4	50.7	61.5	65.9	32.8	40.7	50.3	55.9	**23.2**	**22.8**	**24.6**	**23.1**	60.0	-2.0
	+ MedKgSD	90.4	90.5	**68.6**	**55.6**	**69.5**	70.1	**45.1**	**47.0**	**61.9**	**62.7**	20.3	20.1	20.7	18.9	**66.1**	**+4.1**
MedVL-Thinker-3B	Base	79.9	80.2	60.5	49.0	**60.0**	60.0	51.4	51.6	56.1	56.0	15.9	15.4	27.3	26.5	60.5	–
	+ SFT	80.9	81.1	60.3	**49.5**	58.6	58.8	47.9	48.0	56.4	56.4	14.5	13.7	29.1	26.5	59.8	-0.7
	+ DAPO	79.7	80.0	59.8	48.6	58.2	58.4	47.7	47.5	52.7	52.7	21.4	20.4	25.5	23.8	58.5	-2.0
	+ SDFT	79.7	80.0	50.8	41.0	56.8	56.9	39.5	38.7	53.3	53.4	18.3	18.1	25.5	25.0	55.0	-5.5
	+ Std. OPSD	80.1	80.4	60.4	49.1	59.2	59.4	50.4	50.5	56.2	56.2	17.5	16.8	28.7	26.2	60.2	-0.3
	+ SDFT+demo	79.4	79.7	52.0	42.5	57.2	57.4	41.0	40.2	56.7	56.2	21.6	20.9	**33.4**	**31.6**	56.2	-4.3
	+ MedKgSD	**82.6**	**82.8**	**61.3**	48.2	**60.0**	**60.1**	**51.6**	**51.7**	**58.0**	**58.0**	**22.1**	**21.6**	31.5	30.4	**61.4**	**+0.9**

Table 6: Logged training time on QoQ-Med. Per-step time is measured in seconds from adjacent training-metric records, and total time is the summed valid wall-clock interval in minutes. Stage 1 corresponds to reasoning-task adaptation and Stage 2 to understanding-task adaptation. The table reports coarse step-level timing rather than component-level profiling.

Method	Stage 1		Stage 2		Overall	
	s/step	Total min	s/step	Total min	Avg. s/step	Total min
SFT	2.0	6.5	3.3	9.3	2.6	15.8
Std. OPSD	21.3	37.7	20.6	33.9	21.0	71.6
DRPO	32.2	53.2	38.1	62.9	35.2	116.1
SDFT	56.1	92.5	73.1	120.6	64.6	213.1
MedKgSD	51.9	85.6	65.2	107.6	58.6	193.2

B.1 Expanded Main Results by Backbone

Table 5 expands Table 1 by separating the two backbones within a single appendix table. To keep the appendix view compact, we report only the final Stage 2 checkpoint after understanding-task adaptation. The layout follows the main table and adds standard OPSD and SDFT with demonstration-enriched trajectories to make the comparison set explicit.

B.2 Training-Time Analysis

Table 6 reports the QoQ-Med training time for each method. We estimate per-step wall-clock time from consecutive training telemetry records. Because several runs were resumed, total time is computed as the sum of valid adjacent step intervals within each stage rather than as the difference between the first and last timestamps. Intervals longer than 300 seconds are excluded as run-boundary gaps; no valid within-run intervals are removed by this threshold. These measurements are coarse step-level diagnostics: the logs do not decompose rollout generation, teacher forward passes, retention evaluation, and optimizer updates.

B.3 Retention-Coefficient Sensitivity

We further examine the coefficient λ in Eq. 12, which controls the strength of prior-policy distillation. As shown in Fig. 5, increasing λ from zero to a moderate value improves both prior-task and

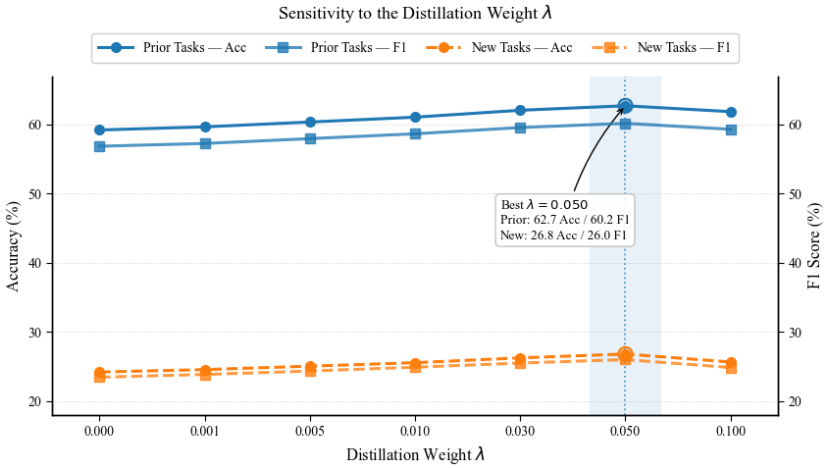

Figure 5: Sensitivity to the prior-policy retention coefficient λ. The sweep reports prior-task and new-task accuracy/F1 under different retention strengths. A moderate retention weight gives the best adaptation-retention balance, while too little retention weakens prior-task preservation and too much retention slightly suppresses both prior-task and new-task performance.

Table 7: Derived continual-learning diagnostics from the main stage-wise evaluation matrix. Prior BWT is the final prior-task change relative to the base model. Old-task BWT and average forgetting are computed over prior tasks and the reasoning task after the final understanding-stage update. AIA denotes average incremental accuracy over the two adaptation stages. Higher is better except for average forgetting.

Backbone	Method	Prior BWT	Old-task BWT	Avg. Forgetting	Reasoning Δ	AIA	Final Seen Avg.
QoQ-Med-7B	SFT	-0.6	-0.9	1.4	-2.1	51.3	48.6
	DRPO	-0.4	-0.2	0.6	1.3	51.9	49.3
	SDFT	-1.8	-1.4	1.5	0.4	50.5	48.0
	MedKgSD	4.1	3.6	0.2	1.0	55.5	53.0
MedVLThinker -3B	SFT	-0.7	-1.1	1.3	-2.9	51.0	48.7
	DAPO	-2.0	-1.1	1.6	3.2	50.6	48.3
	SDFT	-5.5	-4.3	4.6	1.3	49.1	45.5
	MedKgSD	0.9	1.4	0.1	3.5	52.7	51.4

new-task performance, suggesting that a mild retention constraint can stabilize continual adaptation without preventing the model from absorbing KG-conditioned teacher signals. The best trade-off appears at $\lambda = 0.05$, where the model obtains 62.7 prior-task accuracy, 60.2 prior-task F1, 26.8 new-task accuracy, and 26.0 new-task F1. Increasing the coefficient further to 0.1 slightly reduces all four metrics, indicating that overly strong retention can constrain the update and does not provide additional preservation benefit. We therefore use $\lambda = 0.05$ in the main experiments.

B.4 Additional Continual-Learning Diagnostics

We further derive continual-learning diagnostics from the stage-wise results in Table 1. These diagnostics re-express the same evaluation matrix using standard continual-learning views. We report only quantities that can be computed from the evaluated checkpoints. In particular, we do not report strict forward transfer because the protocol does not evaluate every incoming task immediately before its own adaptation stage. Reporting such a number from the base model alone would conflate zero-shot capability with transfer from the preceding stage.

For each method, prior BWT is the final prior-task average minus the base prior-task average. Old-task BWT and average forgetting are computed over tasks that are already available before the final

Table 8: Stage-wise continual-learning matrix derived from Table 1. Prior entries average the five prior tasks over Acc/F1. Reasoning and Understanding entries average Acc/F1 for the corresponding incoming task. Final Seen Avg. averages all tasks observed by the final checkpoint. Reason. and Underst. denote Reasoning and Understanding.

Backbone	Method	Base Prior	Stage 1 Prior	Stage 1 Reason.	Stage 2 Prior	Stage 2 Reason.	Stage 2 Underst.	Final Seen Avg.
QoQ-Med-7B	Base	62.0	–	–	–	–	–	–
	SFT	62.0	61.4	16.8	61.4	14.7	18.9	48.6
	DRPO	62.0	62.1	16.1	61.6	17.4	20.2	49.3
	SDFT	62.0	60.5	15.5	60.2	15.9	19.2	48.0
	MedKgSD	62.0	65.7	19.2	66.1	20.2	19.8	53.0
MedVLThinker -3B	Base	60.5	–	–	–	–	–	–
	SFT	60.5	60.5	17.1	59.8	14.1	27.8	48.7
	DAPO	60.5	59.8	17.8	58.5	20.9	24.7	48.3
	SDFT	60.5	59.8	16.9	55.0	18.2	25.3	45.5
	MedKgSD	60.5	61.1	18.4	61.4	21.9	31.0	51.4

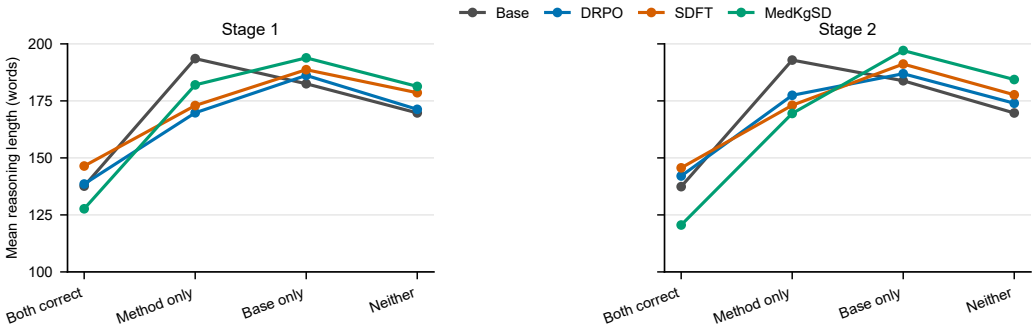

Figure 6: Trace-length dynamics for QoQ-Med prior-task transfer. Samples are grouped by correctness relative to the base model. Method-only buckets correspond to cases where the adapted method is correct but the base model is wrong; base-only buckets show the opposite. MedKgSD gains are not accompanied by uniformly longer traces, showing that the observed gains do not require uniformly longer traces.

stage: the five prior-task Acc/F1 pairs and the reasoning-task Acc/F1 pair after Stage 1. Average incremental accuracy (AIA) averages the mean performance over all tasks observed after Stage 1 and after Stage 2. All values are percentages and are derived from the displayed Acc/F1 values in Table 1.

Table 8 gives the corresponding lower-triangular view used to compute the diagnostics. The matrix separates the base prior-task anchor, the reasoning-stage checkpoint, and the final understanding-stage checkpoint. This view is useful because it distinguishes three effects that are collapsed in the main table: prior-task preservation, reasoning-task retention after the second stage, and adaptation to the final understanding task.

B.5 Positive Backward Transfer and Trace-Length Dynamics

We observed an initially counterintuitive pattern in the QoQ-Med continual-learning runs: after adaptation, several prior OmniMedVQA tasks improve rather than merely being retained. To audit whether this gain is only a byproduct of longer explanations, we recompute correctness from the extracted answer letter and measure the word count inside each generated `<think>` segment. For MedXpertQA, repeated rollout entries are normalized to the corresponding evaluation sample before computing the aggregate statistics.

Figure 6 separates samples by whether the base model and each adapted method are correct. The MedKgSD-only-correct bucket is consistently shorter than the corresponding base traces, whereas base-only and jointly incorrect samples tend to have longer adapted traces. This suggests that the prior-task improvement is not explained by verbose reasoning alone. The shorter traces are consistent with more concise answers on these samples, but length alone does not establish that the reasoning is more accurate or better grounded. The length statistic is only a proxy for trace behavior, so we use it as an observational diagnostic rather than as an optimization target.

C Knowledge Context and Prompt Details

C.1 KG Concept Construction and Context Serialization

We construct KG contexts in two stages: a task-level concept space is built once for each incoming adaptation task, and sample-specific teacher contexts are then selected from that concept space. The lexical matching procedure produces a bounded set of candidate concepts, which may include noisy or weakly related entries. TSE calibrates the resulting teacher distributions during distillation.

We instantiate the KG with PrimeKG and retain biomedical edges whose endpoints belong to the node types used by our method: disease, phenotype, anatomy, gene/protein, biological process, pathway, and drug. For each adaptation task, we extract candidate lookup terms from three task-level sources: the training questions, the task modality description, and the task-level answer vocabulary. The answer vocabulary is used only as a broad label-space prior for the task. It is aggregated across the adaptation split and is not the gold answer for any particular sample. OmniMedVQA prior-task examples and held-out MedXpertQA-MM examples are not used to construct KG contexts.

Candidate terms are normalized by lowercasing, splitting compact medical labels, removing generic prompt words, and adding modality priors such as retina-related terms for fundus images or cell/tissue terms for microscopy. We then retrieve PrimeKG nodes by token overlap and rank them with a deterministic score combining seed-term overlap, modality overlap, and a node-type prior. The default task-level limits are 64 seed terms, 128 retained concepts, and at most 16 neighboring relations per retained concept. This ranking produces a bounded concept space before any per-sample context is generated.

For each training sample, we generate four teacher-side KG views by cycling through label-space-centered, modality-centered, phenotype-centered, and mechanism-centered context types. Each view scores nodes from the task-level concept space using the sample question, the sample modality, and the task-level answer vocabulary. The context selector does not use the sample-specific answer choices or the gold label when selecting KG concepts. Each view keeps up to eight selected concepts and up to ten relations connected to those concepts, and the serialized text is capped at 2048 characters.

The serialized KG block begins with a safety sentence stating that the KG context is a biomedical prior and may not determine the image finding by itself. This sentence is part of the teacher prompt by construction. It is included because KG relations can define plausible clinical concepts without establishing whether the visual evidence is present in the image. The student prompt never receives this KG block during training or inference.

C.2 Prompt Templates

Figure 7 gives the student, teacher, and KG context wrappers. Both policies receive the image, question, and choices; the teacher additionally receives the serialized biomedical context. The answer-only wrappers require the same output format for answer extraction. The teacher wrapper instructs the model to ignore irrelevant KG concepts and ground its decision in the image.

The reasoning traces in the qualitative cases are presented separately from these answer-only templates. They illustrate output behavior, but do not specify the generation configuration used to obtain those traces.

Student system prompt.

- **Role.** You are a careful medical visual question-answering assistant. You answer medical image questions using only the provided image, question, and answer choices.
- **Task.** Inspect the image, understand the question, compare all candidate choices, and select the single option that is best supported by the visual evidence.
- **Grounding.** Do not rely on external knowledge, hidden context, or assumptions that are not supported by the image and question. If multiple choices seem plausible, choose the one most directly supported by the visible findings.
- **Output rule.** Return exactly one option from the provided choices. If choices are labeled, use the exact format `<letter>. <option text>`. Do not include explanations, reasoning steps, confidence scores, or alternative answers.

Student user template. The user message contains the image input followed by the original question and choices:

```
[image(s)]
[Question]
```

Answer with one of the following:

```
[Choice 1]
[Choice 2]
[...]
```

Instruction: Select exactly one answer from the options above. Return only the exact option as written. For labeled options, use `<letter>. <option text>`. Do not add any explanation.

Teacher system prompt.

- **Role.** You are a same-backbone teacher model for medical visual question answering. You receive the same image, question, and choices as the student, together with teacher-only biomedical KG context.
- **Task.** Re-evaluate the student-facing VQA input under the supplied KG-derived privileged condition. Use the KG context to clarify medically relevant diseases, anatomical entities, phenotypes, mechanisms, or relations when they help distinguish the choices.
- **Grounding.** Treat the KG context as biomedical prior knowledge, not as an answer key. The final decision must remain grounded in the image and question. Do not infer visual findings that contradict the image, and ignore KG concepts that are irrelevant to the current sample.
- **Output rule.** Do not expose the KG text in the final response. Return exactly one option in the same format required of the student, with no explanation, rationale, confidence score, or extra text.

Teacher user template. The teacher receives the same image, question, choices, and answer-format instruction as the student, followed by:

Teacher-only biomedical KG context:

```
[Serialized KG context]
```

Use the KG context only as biomedical prior knowledge. The final answer must still be exactly one option from the choices above.

KG context template.

- **Role.** This block represents teacher-only biomedical prior knowledge derived from KG concepts and relations. It is used only to condition the teacher during distillation and is not provided to the student at deployment.
- **Content.** The context contains a compact set of candidate biomedical concepts and relation triples, which may cover label-space concepts, modality-related entities, clinical phenotypes, anatomical structures, and disease mechanisms.
- **Constraint.** The KG context may be incomplete, noisy, or only partially relevant. It should support clinical disambiguation but should not override visual grounding. We keep a bounded number of concepts and relations and truncate the serialized text to a fixed length budget.

Serialized context template.

This KG context is a biomedical prior and may not determine the image finding by itself.

```
- Concepts: [concept 1; concept 2; ...; concept m]
- Relations:
1. [source concept] --[relation]--> [target concept]
2. [...]
```

Figure 7: Prompt templates used by MedKgSD. The student and teacher share the same image, question, choices, and answer-format instruction; only the teacher receives the additional KG context. This preserves the deployment interface while allowing the teacher policy to express how structured biomedical priors change its distribution over the student-generated trajectory.

D Qualitative Case Studies

D.1 Comparison with SDFT

We present a representative fundus VQA case in Fig. 8 to illustrate how KG-guided self distillation affects the model's reasoning. The question asks for the specific abnormality in a retinal fundus image, with the ground-truth answer being no diabetic retinopathy. For MedKgSD, teacher-side KG contexts organize candidate clinical concepts around diabetic retinopathy, retinal morphology, retinal hemorrhage, and abnormal retinal vascular morphology. Guided by these structured concepts and relations, MedKgSD checks for the absence of hemorrhages, exudates, macular edema, neovascularization, and vascular abnormalities before giving the correct answer. In contrast, SDFT produces a longer, less focused trace, first hypothesizing subtle diabetic retinopathy-related changes and then drifting toward congenital cataracts. This example illustrates a difference in the generated traces. A single qualitative comparison does not isolate the effect of KG conditioning; the controlled comparisons in the main paper provide the quantitative evidence.

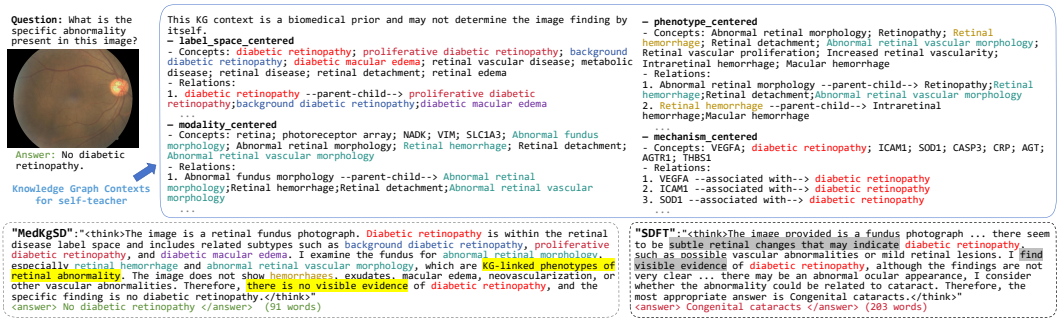

Figure 8: Case study of a representative retinal fundus VQA question. MedKgSD conditions the self-teacher with KG contexts centered on the label space, imaging modality, retinal phenotype, and biomedical mechanism. Guided by these contexts, MedKgSD focuses on clinically relevant retinal abnormalities and answers "no diabetic retinopathy", while SDFT produces an incorrect cataract diagnosis. The key KG-linked clinical entities are highlighted.

Complete KG-grounded trajectory for Diabetic Retinopathy (part 1)

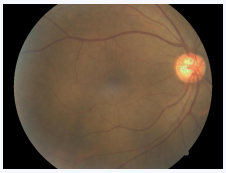

Image: retinal fundus photograph

Task. OmniMedVQA lesion-grading prior evaluation.
Question. What is the specific abnormality present in this image?
Answer choices. A. Congenital cataracts; B. Corneal ulcer; C. No diabetic retinopathy.; D. Retinal detachment.
Gold answer. C: No diabetic retinopathy.
MedKgSD output. <answer> No diabetic retinopathy </answer>

MedKgSD reasoning trace. The image is a retinal fundus photograph. Diabetic retinopathy is within the retinal disease label space and includes related subtypes such as background diabetic retinopathy, proliferative diabetic retinopathy, and diabetic macular edema. I examine the fundus for abnormal retinal morphology, especially retinal hemorrhage and abnormal retinal vascular morphology, which are KG-linked phenotypes of retinal abnormality. The image does not show hemorrhages, exudates, macular edema, neovascularization, or other vascular abnormalities. Therefore, there is no visible evidence of diabetic retinopathy, and the specific finding is no diabetic retinopathy.

Context 0: label-space-centered

- **Prior note:** This KG context is a biomedical prior and may not determine the image finding by itself.
- **Concepts:** diabetic retinopathy; proliferative diabetic retinopathy; background diabetic retinopathy; diabetic macular edema; retinal vascular disease; metabolic disease; retinal disease; retinal detachment; retinal edema.
- **Relations:**
  1. diabetic retinopathy –parent-child–> proliferative diabetic retinopathy
  2. diabetic retinopathy –parent-child–> background diabetic retinopathy
  3. diabetic retinopathy –parent-child–> diabetic macular edema
  4. retinal vascular disease –parent-child–> diabetic retinopathy
  5. metabolic disease –parent-child–> diabetic retinopathy
  6. retinal disease –parent-child–> retinal detachment
  7. retinal disease –parent-child–> retinal vascular disease
  8. retinal disease –parent-child–> retinal edema

Context 1: modality-centered

- **Prior note:** This KG context is a biomedical prior and may not determine the image finding by itself.
- **Concepts:** retina; photoreceptor array; NADK; VIM; SLC1A3; Abnormal fundus morphology; Abnormal retinal morphology; Retinal hemorrhage; Retinal detachment; Abnormal retinal vascular morphology.
- **Relations:**
  1. retina –parent-child–> photoreceptor array
  2. NADK –expression present–> retina
  3. VIM –expression present–> retina
  4. SLC1A3 –expression present–> retina
  5. Abnormal fundus morphology –parent-child–> Abnormal retinal morphology
  6. Abnormal retinal morphology –parent-child–> Retinal hemorrhage
  7. Abnormal retinal morphology –parent-child–> Retinal detachment
  8. Abnormal retinal morphology –parent-child–> Abnormal retinal vascular morphology

Figure 9: PrimeKG-grounded diabetic-retinopathy case trajectory, part 1. The panel shows the original fundus image, question, answer choices, extracted MedKgSD answer, reasoning trace, and the first two teacher-side KG views. The label-space context enumerates diabetic-retinopathy subtypes and related retinal diseases, while the modality-centered context anchors the sample to fundus and retinal morphology concepts.

D.2 KG-Grounded Case Trajectory

We include one complete retinal-fundus example to make the teacher-side KG condition auditable at the sample level. Figures 9 and 10 report the complete context rather than a shortened post-hoc summary. We keep label-space, modality, phenotype, and mechanism-centered views, including relations that are only indirectly useful, so that the reader can inspect what the teacher received. The displayed reasoning trace is used as case-level evidence; evaluation still depends on the extracted final answer span. This distinction matters because MedKgSD is not given the KG context at inference time, and the KG context is not treated as an oracle that bypasses image grounding.

Complete KG-grounded trajectory for Diabetic Retinopathy (part 2)

Context 2: phenotype-centered

- **Prior note:** This KG context is a biomedical prior and may not determine the image finding by itself.
- **Concepts:** Abnormal retinal morphology; Retinopathy; Retinal hemorrhage; Retinal detachment; Abnormal retinal vascular morphology; Retinal vascular proliferation; Increased retinal vascularity; Intraretinal hemorrhage; Macular hemorrhage.
- **Relations:**
  1. Abnormal retinal morphology –parent-child–> Retinopathy
  2. Abnormal retinal morphology –parent-child–> Retinal hemorrhage
  3. Abnormal retinal morphology –parent-child–> Retinal detachment
  4. Abnormal retinal morphology –parent-child–> Abnormal retinal vascular morphology
  5. Abnormal retinal vascular morphology –parent-child–> Retinal vascular proliferation
  6. Abnormal retinal vascular morphology –parent-child–> Increased retinal vascularity
  7. Retinal hemorrhage –parent-child–> Intraretinal hemorrhage
  8. Retinal hemorrhage –parent-child–> Macular hemorrhage

Context 3: mechanism-centered

- **Prior note:** This KG context is a biomedical prior and may not determine the image finding by itself.
- **Concepts:** VEGFA; diabetic retinopathy; ICAM1; SOD1; CASP3; CRP; AGT; AGTR1; THBS1.
- **Relations:**
  1. VEGFA –associated with–> diabetic retinopathy
  2. ICAM1 –associated with–> diabetic retinopathy
  3. SOD1 –associated with–> diabetic retinopathy
  4. CASP3 –associated with–> diabetic retinopathy
  5. CRP –associated with–> diabetic retinopathy
  6. AGT –associated with–> diabetic retinopathy
  7. AGTR1 –associated with–> diabetic retinopathy
  8. THBS1 –associated with–> diabetic retinopathy

Figure 10: PrimeKG-grounded diabetic-retinopathy case trajectory, part 2. The panel continues the same teacher-side context with phenotype-centered and mechanism-centered PrimeKG relations.

D.3 Boundary Failure Case on Modality Recognition

We include one failure case to clarify the boundary of KG-conditioned self distillation. The example comes from the modality-recognition prior task, where the correct decision depends mainly on image scale and acquisition format rather than on disease, anatomy, or phenotype relations. This setting separates a shallow visual-format cue from the type of structured biomedical prior that MedKgSD is designed to exploit. The case should therefore be read as a boundary condition for context reliability, not as evidence that KG conditioning is harmful in general.

Boundary case: shallow modality cue versus biomedical semantic prior

Case. `ALL Challenge_0058` from the QoQ-Med modality prior task.
Question. What modality of imaging is utilized to capture this picture?
Answer choices. A. Magnetic Resonance Imaging (MRI); B. PET scan; C. Ultrasound; D. Microscopy.
Ground truth. D. Microscopy.

Model	Answer	Status	Trace behavior
Base	D	Correct	Identifies the cell-level appearance and rules out whole-body modalities such as MRI, PET, and ultrasound.
DRPO	D	Correct	Also anchors the decision on cellular detail, color staining, and microscopy-scale structure.
SDFT	A	Wrong	Maps the isolated colored object to cross-sectional soft-tissue imaging and selects MRI.
MedKgSD	A	Wrong	Emphasizes MRI-like soft-tissue contrast and color-coded diagnostic imaging, despite the microscopy cue.

KG context audit. The curated context for this sample is partially correct rather than purely noisy: it contains *microscopy*, *optical microscope*, *hematology microscopy*, and *color-stained cellular image*, but it also contains MRI-related attractors such as *magnetic resonance imaging*, *MRI contrast*, *soft tissue image*, and *MRI signal generation*. Several relations explicitly support the correct modality, e.g., *optical microscope captures color-stained cellular image* and *cellular morphology supports microscopy classification*. At the same time, relations such as *MRI contrast can show irregular tissue-like regions* and *MRI signal generation supports magnetic resonance imaging* provide a competing diagnostic-imaging interpretation.
Interpretation. The error illustrates an over-semanticization mode: when the answer is determined by a low-level visual-format cue, a teacher conditioned on biomedical concepts can still be pulled toward a plausible clinical-imaging explanation. This does not contradict the positive transfer observed on anatomy, lesion, disease, and biological-attribute tasks, where structured medical concepts are more directly relevant. Instead, the case motivates the reliability checks used in MedKgSD: KG-conditioned teacher signals should be trusted most when the task requires biomedical relational knowledge, and down-weighted or audited when the decision is primarily a modality or acquisition-format cue.

Figure 11: Boundary failure case on modality recognition. The case isolates a shallow visual-format decision where direct microscopy cues are sufficient, but KG-conditioned reasoning can be attracted by clinically plausible MRI semantics.

E Limitations and Impact

E.1 Limitations and Future Work

MedKgSD is designed for continual medical VQA adaptation where the student follows the standard deployment input, while KG-derived privileged knowledge conditions are available only to the teacher side during training. This design keeps the inference interface unchanged, but it also limits what the method can claim. The present experiments evaluate two medical VQA backbones, a two-stage adaptation stream, and a fixed set of prior-task anchors. They do not establish behavior under longer clinical task streams, multi-turn clinical dialogue, longitudinal patient data, report generation, or institution-level distribution shift. Extending the evaluation to longer streams and more heterogeneous medical workflows is therefore a necessary next step.

The KG context used by MedKgSD should be interpreted as a structured biomedical prior rather than as a sample-level explanation or an answer oracle. Even when the selected concepts are clinically meaningful, the KG view may be incomplete, partially irrelevant, or too coarse for visually shallow decisions. The modality-recognition failure case illustrates this boundary: a low-level acquisition-format cue can be easier to solve from the image than from disease- or phenotype-centered semantics. Future work should improve multimodal entity linking, task-aware KG view selection, and context reliability estimation, especially for cases where visual format, staining, or acquisition scale determines the answer.

Token-selective entropy calibration provides a way to weight KG-conditioned teacher policies by their decisiveness on student-visited trajectories, but it is not a factuality verifier for the KG. A low-entropy teacher signal indicates that a context-conditioned policy is more confident on a particular trajectory; it does not prove that the retrieved biomedical relations are complete or causally sufficient for the visual finding. More direct factuality checks, adversarial context perturbations, and calibration analyses across correct and incorrect predictions would make the reliability mechanism more transparent.

The prior-policy retention term is evaluated here as a lightweight constraint against harmful drift. It can preserve prior capabilities and, in related medical tasks, may coincide with positive backward transfer. This should not be read as a guarantee against forgetting in all streams. In longer or more conflicting task sequences, the same constraint may trade off against plasticity. Future work should study adaptive retention schedules, memory-efficient prior-policy approximations, and standard continual-learning diagnostics over longer streams.

MedKgSD also introduces training-time overhead because teacher-side KG-conditioned policies require additional forward computation. This overhead is acceptable for the present experimental setting, but it may grow for larger backbones, larger context banks, or frequent model updates. Future work should study context caching, adaptive context budgets, and smaller teacher mixtures.

E.2 Impact Statement

MedKgSD aims to make medical multimodal self distillation more reliable under continual adaptation by using structured biomedical knowledge as teacher-side privileged conditioning. A positive impact of this direction is that medical MLLMs may adapt to new VQA-style tasks while retaining useful prior behavior, without requiring reward engineering or a separate external teacher model at inference time. Because the KG is used only during training, the deployed student remains compatible with the same input format as the base model.

The method also carries risks that are specific to medical AI. Biomedical KGs and medical VQA datasets can encode incomplete, biased, or outdated knowledge, and a model trained with such signals may produce overconfident explanations when the visual evidence is ambiguous. KG-conditioned reasoning can also over-semanticize cases whose answer depends on shallow visual cues, as shown in the modality boundary case. These risks are more consequential in clinical settings, where an incorrect answer or persuasive but unsupported rationale could affect downstream decisions.

We therefore view MedKgSD as a training methodology for controlled medical VQA research, not as a standalone diagnostic system. Any clinical use would require prospective validation, dataset-governance review, bias and subgroup evaluation, and human oversight. The analyses in this appendix, including prompt disclosure, KG trajectory inspection, training-time reporting, continual-learning diagnostics, and boundary failure cases, are intended to make the method's assumptions and failure modes visible rather than to imply clinical readiness.