DISTILLING REASONING CHAIN AND CHAIN-LOCAL EXPERIENCES FOR FEW-SHOT MULTIMODAL IN-CONTEXT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal in-context learning (ICL) offers a promising avenue for adapting Large Vision-Language Models (LVLMs) to specialized downstream vision-language tasks with only a few in-context demonstrations (ICDs). However, existing few-shot multimodal ICL methods either rely on superficial image-question-answer imitation or induce ICD-individual reasoning guidance, limiting their ability to capture shared task-level reasoning structures. Moreover, these methods are largely inherited from pure language scenarios, overlooking perceptual bottlenecks of models in multimodal tasks and failing to provide sufficient perceptual support for reliable step-wise execution. In this paper, we propose *Distill-ICL*, a training-free framework that distills few-shot ICDs into two complementary forms of guidance: a *Normative Reasoning Chain* and *Chain-Local Experiences*. Specifically, we introduce *ICDs-MCTS*, an ICD-set-oriented variant of Monte Carlo Tree Search that collectively expands and evaluates reasoning paths over the entire demonstration set, inducing a shared reasoning chain that specifies the sequence of reasoning steps for the target task. To support accurate visual perception when executing these steps, Distill-ICL further performs controlled rollouts along the induced chain and distills local guidance at critical nodes: Inter-ICD Pivotal Experiences capture answer-discriminative visual attributes to guide critical decisions, while Intra-ICD Fragile Experiences reinforce visual attributes prone to misinterpretation to help mitigate step-local perceptual errors. Extensive experiments on five specialized multimodal benchmarks show that Distill-ICL consistently outperforms existing training-free ICL methods across various LVLMs, while exhibiting favorable computational efficiency and improved robustness to ICD selection and ordering perturbations. Code will be released upon publication.

1 INTRODUCTION

Recent advances in Large Vision-Language Models (LVLMs) have demonstrated transformative potential in general vision-language tasks. Through multimodal In-Context Learning (ICL), these models can rapidly adapt to novel task specifications from only a few In-Context Demonstrations (ICDs), without requiring any parameter updates. This property makes ICL a compelling paradigm for efficiently adapting general-purpose LVLMs to specialized downstream scenarios, such as medical image and financial chart analysis, where their out-of-the-box performance is often constrained by insufficient domain-specific reasoning and perception capabilities.

Standard few-shot ICL (Brown et al., 2020) treats ICDs as "black-box" templates, expecting the model to mimic the correct answers by imitating superficial image-question-answer correspondences. However, this superficial imitation is extremely fragile when adapting to challenging downstream tasks, as LVLMs struggle to autonomously derive the implicit logical reasoning and visual parsing mechanisms embedded in the ICDs. To this end, recent studies (Wei et al., 2022; Wu et al., 2026) have attempted to utilize reasoning strategies (*e.g.*, Chain-of-Thought (Wei et al., 2022), Tree-of-Thought (Yao et al., 2023), and Tree-search-based reasoning (Coulom, 2006)) to augment each ICD with specific explicit reasoning processes, aiming to help models derive the underlying logic of problem-solving.

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.

Despite their advances, existing reasoning-augmented ICL approaches still face two key limitations. *First*, they typically induce reasoning guidance for each ICD independently, without treating the ICD set as a collective source of task supervision. Consequently, the resulting guidance often captures instance-specific reasoning patterns rather than a shared task-level reasoning structure applicable across ICD instances. *Second*, these approaches are largely inherited from pure language scenarios, overlooking the perceptual support required to execute reasoning steps reliably in multimodal tasks. Executing these reasoning steps reliably requires accurate interpretation of the relevant visual cues. However, limited domain-specific perceptual capabilities can cause general-purpose LVLMs to make errors at individual steps, which may accumulate along the reasoning process and ultimately lead to incorrect predictions.

In this work, we propose **Distill-ICL**, a training-free ICL framework that **distills** two complementary contexts from few-shot ICDs: a *Normative Reasoning Chain* and *Chain-Local Experiences*. *First*, to move beyond instance-wise reasoning induction, we introduce *ICDs-MCTS*, an ICD-set-oriented variant of Monte Carlo Tree Search (MCTS) (Coulom, 2006). ICDs-MCTS collectively expands reasoning paths considering the entire ICD set and evaluates them by overall effectiveness across demonstrations, thereby inducing a shared reasoning chain that captures task-level reasoning patterns generalizable across demonstrations. *Second*, to further improve the visual interpretation reliability of step-wise execution, Distill-ICL performs multiple controlled rollouts along the final searched chain and distills two types of chain-local experiences. Inter-ICD Pivotal Experiences capture answer-discriminative visual attributes revealed by contrasting ICDs with different answers, thereby guiding the LVLM to make critical distinctions at pivotal reasoning steps. Intra-ICD Fragile Experiences identify error-prone visual attributes within a single ICD by contrasting successful and failed rollouts, providing safeguards against local perceptual failures. During test time, the LVLM follows the normative chain and uses chain-local experiences as step-specific guidance for reliable multimodal reasoning.

Our contributions are summarized as follows:

- We propose Distill-ICL, a training-free framework that can reliably perform few-shot multimodal ICL, which enables general-purpose LVLMs to adapt effectively to specialized downstream multimodal scenarios.
- We introduce ICDs-MCTS, an ICD-set-oriented MCTS variant that searches for a Normative Reasoning Chain at the ICD-set level, yielding a shared reasoning guidance generalizable across ICDs. Based on this chain, we further distill Chain-Local Experiences, which associate critical nodes with Inter-ICD Pivotal Experiences for answer-critical decisions and Intra-ICD Fragile Experiences for preventing local perceptual failures.
- Extensive experiments on five specialized multimodal benchmarks demonstrate that Distill-ICL consistently achieves state-of-the-art (SOTA) performance on five open-source and proprietary LVLMs. Moreover, Distill-ICD exhibits leading test-time computational efficiency and robustness to ICD perturbation.

2 Related Work

2.1 Multimodal In-Context Learning

In-context learning (ICL) enables large models to adapt to specific tasks by conditioning on contextual information Tai et al. (2024); Dou et al. (2026); Dong et al. (2026). In this work, we focus on the setting where the context is defined as few-shot multimodal in-context demonstrations (ICDs), each specified as an image-question-answer triplet.

Recent studies improve multimodal ICL either by training LVLMs with dedicated architectures and large-scale interleaved image-text data (Alayrac et al., 2022; Awadalla et al., 2023; Zhao et al., 2024), or by enhancing demonstrations at test time in a training-free manner Monea et al. (2025); Ma et al. (2025). Since training-dependent approaches lack flexibility in downstream adaptation, our work focuses on the latter setting. Existing training-free methods Wei et al. (2022); Wu et al. (2026) typically augment ICDs with explicit reasoning processes. However, these methods typically derive ICD-individual reasoning guidance rather than the shared task-level reasoning structure across the ICD set. Moreover, they fail to support for reliable step-wise execution, especially at answer-pivotal or execution-fragile nodes.

2.2 Reasoning-Augmented Prompting

Reasoning-augmented prompting improves model inference by eliciting intermediate reasoning processes rather than relying on direct answer generation. Classically, Chain-of-Thought Wei et al. (2022) elicits step-by-step rationales, while Tree-of-Thoughts Yao et al. (2023) and Graph-of-Thoughts Besta et al. (2024) organize candidate thoughts into structured search spaces. RAP Hao et al. (2023) and LATS Zhou et al. (2024) further use Monte Carlo Tree Search (MCTS) Coulom (2006) to search over reasoning trajectories.

These advances have also inspired recent ICL methods to augment ICDs with explicit reasoning contents. The closest work to ours is ThoughtICR (Wu et al., 2026), which applies MCTS over a fixed set of abstract reasoning-actions to construct thought cards for ICDs. However, it still induces reasoning guidance independently for each ICD. In contrast, we introduce ICDs-MCTS, an ICD-set-oriented MCTS variant that defines path expansion and reward estimation over the entire ICD set, thereby inducing a Normative Reasoning Chain that captures shared task-level reasoning structure. Furthermore, we build on this chain to distill Chain-Local Experiences, providing localized visual-attribute guidance for more reliable multimodal reasoning.

3 Methodology

3.1 Problem Formulation

In this work, we study few-shot multimodal ICL for specialized visual question answering. Given a small set of task-specific in-context demonstrations (ICDs), our objective is to adapt a general-purpose LVLM to a target downstream task, so that it can solve query image-question pairs similar to those ICDs. Formally, let $\mathcal{D} = \{D_m\}_{m=1}^{M}$ denote the ICD set, where each demonstration $D_m = (I_m, Q_m, A_m)$ denotes an image-question-answer triplet. Given a target query (I_q, Q_q), the standard ICL paradigm constructs the model context by directly concatenating all ICDs with the target query:

$$\mathcal{C}_{\text{base}} = \text{Concat}\Big[\{(I_m, Q_m, A_m)\}_{m=1}^{M},\ (I_q, Q_q)\Big]. \tag{1}$$

While straightforward, this paradigm offers neither explicit guidance on the reasoning steps to be performed nor safeguards for their correct execution.

3.2 Overview

To address these limitations, as shown in Fig. 1, we propose **Distill-ICL**, a training-free framework that **distills** two complementary contextual augmentations from few-shot ICDs: a *Normative Reasoning Chain* and *Chain-Local Experiences*. The normative reasoning chain is searched by ICDs-MCTS over a broad reasoning-action space, yielding a task-level reasoning-action trajectory generalizable across the ICD set. Conditioned on this chain, chain-local experiences are further distilled from multiple rollouts. Each experience supports the reliable execution of a specific critical step, such as a pivotal step that strongly affects the final prediction or a fragile step that is prone to errors.

3.3 ICDs-MCTS for Normative Reasoning Chain Search

In this stage, we aim to search for a single *Normative Reasoning Chain* that generalizes across the ICD set. Unlike standard MCTS, which typically optimizes a trajectory for a single instance, our ICDs-MCTS performs collective search over all ICDs. Each tree node v represents a reasoning-action prefix $h(v) = (a^{(1)}, \ldots, a^{(t)})$, and the final chain is obtained by identifying a high-reward path shared by the ICD set.

ICD-Collective Expansion. At each expandable node v, we query the LVLM with an expansion prompt that conditions on the current prefix $h(v)$ and each ICD D_m. The generated next-step actions are then canonicalized and aggregated across all ICDs to form the candidate action set:

$$\mathcal{A}(v) = \Phi\left(\bigcup_{m=1}^{M} \text{LVLM}\big(\mathcal{P}_{\text{exp}}(D_m, h(v))\big)\right), \tag{2}$$

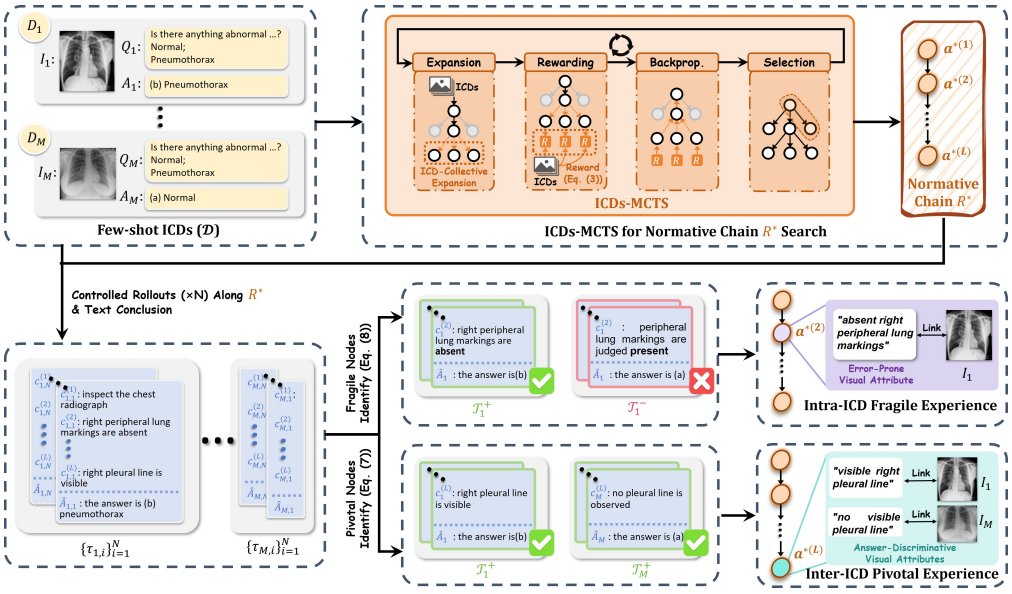


Figure 1: Overview of Distill-ICL. From a few-shot ICD set, ICDs-MCTS collectively searches over the entire ICD set to induce a task-level Normative Reasoning Chain R^*. Distill-ICL then performs controlled rollouts along R^* to identify Inter-ICD Pivotal nodes and Intra-ICD Fragile nodes, from which it distills Chain-Local Experiences with image-linked visual attributes. These experiences provide step-specific guidance for answer-critical decisions and safeguards against local perceptual failures during inference.

where $\mathcal{P}_{\text{exp}}$ denotes the expansion prompt and $\Phi(\cdot)$ maps semantically equivalent action descriptions into the same canonical action. This collective expansion allows the search space to cover diverse reasoning cues from different ICDs, rather than being biased toward a single demonstration.

Universality-Aware Rewarding. For a newly expanded child node $u = v \oplus a$, we evaluate whether the extended prefix $h(u)$ is useful for solving the ICD set. Let $r_m(u)$ denote the execution score of $h(u)$ on ICD D_m, computed by verifying the model prediction against the ground-truth (GT) answer A_m. To favor actions that are broadly effective across ICDs, we define the reward as:

$$R(u) = \frac{1}{M}\sum_{m=1}^{M} r_m(u) - \lambda \cdot \sqrt{\frac{1}{M}\sum_{m=1}^{M}\left(r_m(u) - \bar{r}(u)\right)^2}, \tag{3}$$

where $\bar{r}(u) = \frac{1}{M}\sum_{m=1}^{M} r_m(u)$ and λ controls the penalty on cross-ICD variance. The first term measures overall effectiveness, while the second term discourages actions that work only for a small subset of ICDs.

Backpropagation and Selection. Following this, $R(u)$ is backpropagated to all nodes along the selected path. For each node w on this path, we increment its visit count $N(w)$ by one and update its value estimate $Q(w)$ as:

$$Q(w) = \frac{1}{N(w)}\sum_{j=1}^{N(w)} R_j(w), \tag{4}$$

where $R_j(w)$ denotes the reward backpropagated through node w at its j-th visit. The updated statistics are then used in the next selection phase. Specifically, starting from the root, ICDs-MCTS recursively selects the child node with the highest upper-confidence score:

$$u^* = \arg\max_{u \in \text{Ch}(v)} \left(Q(u) + c\sqrt{\frac{\log(N(v)+1)}{N(u)+1}} \right), \tag{5}$$

where $\text{Ch}(v)$ denotes the children of node v and c controls the exploration strength. By alternating selection, collective expansion, universality-aware rewarding, and backpropagation, ICDs-MCTS

progressively concentrates the search on reasoning-action prefixes that are both effective and broadly supported across the ICD set.

Chain Decoding. After the search budget is exhausted, we decode the Normative Reasoning Chain by selecting the leaf node with the highest length-normalized path score:

$$v^* = \arg\max_{v \in \mathcal{S}_{\text{leaf}}} \left(\frac{1}{|h(v)|} \sum_{w \in \text{Path}(v)} Q(w) \right), \tag{6}$$

where $\mathcal{S}_{\text{leaf}}$ denotes the leaf-node set, $\text{Path}(v)$ denotes the path from the root v_0 to node v, and $|h(v)|$ is the length of the corresponding action prefix. The resulting chain is formulated as $R^* = h(v^*) = (a^{*(1)}, \ldots, a^{*(L)})$, which serves as a task-level reasoning trajectory for subsequent chain-local experience distillation.

3.4 Chain-Local Experience Distillation

The Normative Reasoning Chain R^* specifies the reasoning actions to follow, but does not ensure their reliable execution. In specialized tasks, errors often arise from local failures along an otherwise reasonable chain, such as drawing an incorrect intermediate conclusion at an answer-pivotal step or an execution-fragile step. To address this issue, we distill *Chain-Local Experiences*, which attach step-specific guidance to critical nodes of R^*.

Controlled Rollouts Along R^*. Following R^*, we perform N controlled rollouts for each ICD D_m: $\{\tau_{m,i}\}_{i=1}^{N} = \left\{ \left(c_{m,i}^{(1)}, \ldots, c_{m,i}^{(L)}, \hat{A}_{m,i}\right) \right\}_{i=1}^{N}$. Here, $c_{m,i}^{(k)}$ denotes the intermediate conclusion at step $a^{*(k)}$, and $\hat{A}_{m,i}$ is the final prediction. We omit the free-form reasoning text, as rollout variations mainly reflect local conclusion differences while the action sequence is fixed. We then split the rollouts into successful and failed sets $\mathcal{T}_m^+$ and $\mathcal{T}_m^-$ by verifying $\hat{A}_{m,i}$ against A_m.

Inter-ICD Pivotal Experience. Some steps are pivotal because different conclusions at these steps can substantially affect the final answers. To identify such steps, we compare successful rollouts from two ICDs D_m and $D_{m'}$ with $A_m \neq A_{m'}$, and measure the Jensen-Shannon divergence between their step-wise conclusion distributions:

$$\Delta_{m,m'}^{(k)} = D_{\text{JS}} \left(P(c^{(k)} | \mathcal{T}_m^+) \parallel P(c^{(k)} | \mathcal{T}_{m'}^+) \right). \tag{7}$$

If $\Delta_{m,m'}^{(k)} > \gamma_p$, the step $a^{*(k)}$ is identified as an inter-ICD pivotal node. For this node, we prompt the LVLM to condense the two divergent conclusions into answer-discriminative visual attributes specifically for the two ICDs, and link these to the corresponding images as attribute-level textual descriptions, forming the *Inter-ICD Pivotal Experience*.

Intra-ICD Fragile Experience. Unlike inter-ICD pivotal nodes, intra-ICD fragile nodes characterize local execution instability within the same ICD. At such steps, the LVLM is prone to hallucinations, causing different rollouts on the same ICD to reach inconsistent conclusions and produce different final predictions. We identify these nodes by contrasting successful and failed rollouts of each ICD D_m:

$$\delta_m^{(k)} = D_{\text{JS}} \left(P(c^{(k)} | \mathcal{T}_m^+) \parallel P(c^{(k)} | \mathcal{T}_m^-) \right). \tag{8}$$

A step $a^{*(k)}$ is identified as an intra-ICD fragile node if $\delta_m^{(k)} > \gamma_f$. For this node, we prompt the LVLM to abstract the shared visual attributes (*i.e.*, error-prone visual attributes) of successful rollouts and link them to I_m, forming the *Intra-ICD Fragile Experience*. These experiences emphasize visually challenging attributes that are easily missed or misinterpreted, guiding the model to execute the corresponding step more reliably.

Model Test-Time Inference. At test time, Distill-ICL augments the base context $\mathcal{C}_{\text{base}}$ with R^* and the distilled Chain-Local Experiences. The LVLM is instructed to follow R^* step by step, while experiences associated with pivotal or fragile nodes provide localized, image-linked guidance for executing the corresponding reasoning steps.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Benchmarks We evaluate Distill-ICL on five specialized multimodal benchmarks spanning four downstream domains: medical imaging (SMMILE (Rieff et al., 2025) and EndoBench (Liu et al., 2025)), finance (FinMME (Luo et al., 2025)), industrial anomaly detection (MMAD (Jiang et al., 2025)), and plant pathology (PlantVillageVQA (Sakib et al., 2025)). Except for SMMILE, which is originally designed for multimodal ICL, we construct the ICL setting for the remaining benchmarks by randomly grouping samples of the same task type into ICDs and query instances. Note that these benchmarks require domain-specific perception and structured reasoning, making them well suited to assess whether different methods can adapt general-purpose LVLMs to downstream scenarios beyond their native competence. More dataset details are provided in the Appendix A.

Evaluation Protocol. Considering the inherent instability of few-shot ICL, we repeat evaluation over five randomly sampled ICL episodes, where each episode contains a newly sampled set of ICDs and query instances from the same task type. To ensure a fair and reliable comparison, all baselines and our method are evaluated on the same ICL episodes, with all tasks formulated as multiple-choice question answering. We report the mean accuracy across episodes, and further analyze ICD-selection and ordering robustness in Sec. 4.6.

Baselines. To ensure a comprehensive evaluation, we use five representative LVLM backbones that support multimodal inputs, including three open-source LVLMs[1]: Qwen3-VL-8B and Qwen3-VL-32B (Bai et al., 2025), Llama-3.2-11B (Meta AI, 2024), as well as two proprietary LVLMs: Gemini-3.1-Flash (Google DeepMind, 2025) and GPT-5.6-terra (OpenAI, 2026). We compare our Distill-ICL with five competitive training-free ICL inference baselines: *Zero-shot* directly answers the query without demonstrations. *Few-shot ICL* (Brown et al., 2020) uses image-question-answer triples as standard multimodal demonstrations. *Few-shot CoT* (Wei et al., 2022) extends Few-shot ICL by augmenting step-by-step rationales. *ThoughtICR* (Wu et al., 2026) induces a thought card for each ICD, which specifies a sequence containing five types of human-like reasoning actions and serves as high-level guidance for test-time reasoning. *Stochastic ICRL* (Monea et al., 2025) searches for robust in-context compositions by retaining successful sampled ICD rollouts. We implement ThoughtICR and Stochastic ICRL following their official codes.

Implementation Details. Distill-ICL is implemented as a training-free framework. Open-source LVLMs are deployed with vLLM framework (Kwon et al., 2023), while proprietary LVLMs are accessed through their official APIs. For ICDs-MCTS, each expandable node samples three candidate actions per ICD using temperature 0.7 and top-p 0.9, followed by action canonicalization and merging. We run ICDs-MCTS with a search budget S of 50 simulations. For Chain-Local Experience Distillation, we conduct 20 controlled rollouts per ICD along R^* using temperature 0.7 and top-p 0.9 to expose diverse local execution behaviors. The pivotal-node and fragile-node thresholds are set to $\gamma_p = 0.3$ and $\gamma_f = 0.25$, respectively, and their search experiments are shown in Appendix C.1. All canonicalization, summarization, and visual-attribute abstraction are performed by the corresponding LVLM backbone itself, without introducing additional external models. During final inference, we use decoding with temperature 0 and top-p 1.0 for stable generation.

4.2 MAIN RESULTS

Table 1 reports the performance of different LVLM backbones and ICL inference methods across five specialized multimodal benchmarks. We summarize the two main observations as follows:

(I) *Efficacy of Rationale-Augmented Demonstrations.* The standard Few-shot ICL fails to produce consistent improvements over the Zero-shot baseline across different backbones and benchmarks. In some cases, such as Llama-3.2-11B, directly utilizing image-question-answer triples even degrades the average performance, suggesting that simple input-output demonstrations alone are insufficient for eliciting the domain-specific reasoning required by specialized visual tasks. In contrast, methods that introduce explicit reasoning guidance, including Few-shot CoT, ThoughtICR, and Distill-

[1] All open-source LVLMs are evaluated in their instruction-tuned versions.

Table 1: Accuracy (%) comparison across five domain-specific multimodal benchmarks. Avg. is computed over all benchmark scores, and the best result is highlighted in **bold**.

LVLMs	Methods	Medical		Finance	Anomaly	Plant	Avg.
		SMMILE	EndoBench	FinMME	MMAD	PlantVillageVQA	
Qwen3-VL-8B	Zero-shot	53.3	40.1	54.8	46.4	19.8	42.9
	Few-shot ICL	51.8	45.4	53.0	52.3	25.8	45.7
	Few-shot CoT	55.7	47.2	56.8	55.3	29.6	48.9
	ThoughtICR	58.4	50.3	61.1	56.5	31.2	51.5
	Stochastic ICRL	56.0	48.4	58.9	54.7	29.0	49.4
	Ours	**61.2**	**53.7**	**64.6**	**61.2**	**35.1**	**55.2**
Qwen3-VL-32B	Zero-shot	59.1	41.6	63.0	52.4	27.9	48.8
	Few-shot ICL	60.7	46.8	65.0	60.7	34.3	53.5
	Few-shot CoT	63.9	50.0	66.3	65.4	38.5	56.8
	ThoughtICR	67.3	53.4	68.2	69.0	40.7	59.7
	Stochastic ICRL	64.3	48.6	66.7	66.3	39.4	57.1
	Ours	**69.5**	**56.3**	**70.5**	**72.8**	**44.7**	**62.8**
Llama-3.2-11B	Zero-shot	42.3	31.3	49.4	37.8	11.7	34.5
	Few-shot ICL	40.5	29.3	49.5	37.3	12.6	33.8
	Few-shot CoT	44.2	32.0	51.5	39.7	15.6	36.6
	ThoughtICR	45.7	32.8	53.1	41.2	17.8	38.1
	Stochastic ICRL	42.9	32.2	51.9	40.6	16.5	36.8
	Ours	**47.0**	**35.5**	**55.7**	**44.1**	**20.1**	**40.5**
Gemini-3.1	Zero-shot	68.9	48.0	67.0	54.4	43.0	56.3
	Few-shot ICL	68.9	52.4	67.3	62.6	44.1	59.1
	Few-shot CoT	70.7	54.9	68.5	65.6	47.0	61.3
	ThoughtICR	71.8	56.5	69.8	67.9	48.4	62.9
	Stochastic ICRL	71.4	56.7	68.6	68.5	48.9	62.8
	Ours	**72.5**	**58.5**	**71.4**	**71.0**	**50.2**	**64.7**
GPT-5.6-terra	Zero-shot	73.7	50.3	72.4	55.7	45.3	59.5
	Few-shot ICL	74.1	53.7	74.2	58.8	47.9	61.7
	Few-shot CoT	75.5	55.0	75.6	60.4	49.4	63.2
	ThoughtICR	77.0	56.9	77.1	62.3	51.7	65.0
	Stochastic ICRL	76.2	55.6	76.1	61.0	50.5	63.9
	Ours	**78.7**	**59.4**	**79.0**	**65.0**	**54.5**	**67.3**

ICL, generally exhibit more stable gains over the Zero-shot baseline. This indicates that explicit structured reasoning signals are important for adapting LVLMs to domains that require fine-grained perception and multi-step decision-making.

(II) *Superiority of Distill-ICL.* Compared to existing training-free ICL methods, Distill-ICL delivers consistent improvements across LVLM backbones and specialized benchmarks. We attribute these gains to two aspects. First, ICDs-MCTS searches for a normative reasoning chain jointly supported by the entire ICD set over a broad, task-specific reasoning-action space, rather than relying on a single demonstration or a fixed abstract action space as ThoughtICR. Second, Chain-Local Experiences provide step-local guidance at answer-pivotal and execution-fragile nodes, helping the model alleviate local visual perception errors. The significance analysis is presented in Appendix C.2.

Overall, these results show that Distill-ICL more fully exploits the limited information in few-shot ICDs by transforming sparse demonstrations into structured reasoning guidance and localized image-linked experiences.

4.3 Ablation Studies

To validate the efficacy of each component proposed in our Distill-ICL: the normative chain R^* searched by ICDs-MCTS algorithm and Chain-local Experiences, we conduct extensive ablation studies. In this section, all ablation experiments are performed with Qwen3-VL-32B on three representative benchmarks: SMMILE, EndoBench, and FinMME.

Effectiveness of ICDs-MCTS. Block I of Table 2 evaluates different strategies for obtaining reasoning-chain guidance. We compare ICDs-MCTS with three representative alternatives: CoT+Best-of-N (BoN), Tree-of-Thought (ToT), and standard MCTS. For a fair comparison, we ensure that the rollout numbers or search budgets are matched across all methods. Unlike our ICDs-MCTS, which induces a unified Normative Reasoning Chain shared by the entire ICD set, these alternatives are implemented to derive instance-specific reasoning trajectories for individual ICDs. The results show that our ICDs-MCTS achieves the best performance, suggesting that, under few-shot settings, inducing a shared task-level reasoning trajectory from limited ICDs is more effective than selecting or aggregating instance-specific trajectories.

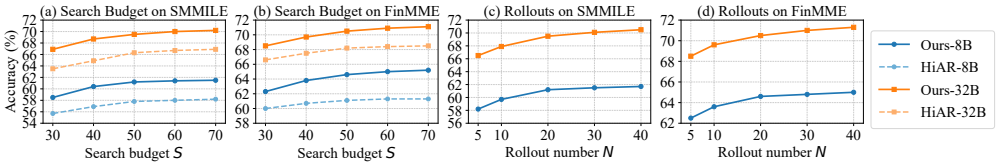


Figure 2: Test-time scaling of Distill-ICL with respect to search and rollout budgets: S and N.

Effectiveness of Chain-Local Experiences. In Block II of Table 2, based on the searched Normative Chain R^*, we separately introduce two types of distilled Chain-Local Experiences: Inter-ICD Pivotal Experiences and Intra-ICD Fragile Experiences. Starting from R^*, the introduction of both types of experiences improves model performance, validating their effectiveness. We find that Intra-ICD Fragile Experiences bring more significant improvements, which indicates that it is crucial to provide explicit guidance for visual attributes that are easily overlooked or misinterpreted during local execution. The complete Distill-ICL achieves the best overall performance, demonstrating that the two types of experience are complementary: pivotal experiences capture inter-ICD answer-discriminative cues, whereas fragile experiences provide intra-ICD safeguards for reliable step-wise reasoning.

Table 2: Ablation on the effectiveness of the two proposed components: ICDs-MCTS and Chain-Local Experiences. Qwen3-VL-32B is selected as the LVLM backbone.

Methods	Medical		Finance	Avg.
	SMMILE	EndoBench	FinMME	
Baselines:				
Few-shot ICL	60.7	46.8	65.0	57.5
Block I: Ablation on ICDs-MCTS:				
CoT+BoN	64.4	50.7	66.6	60.6
ToT	65.0	51.3	66.9	61.1
MCTS	65.4	52.0	67.4	61.6
ICDs-MCTS (R^*)	66.3	53.2	68.0	62.5
Block II: Ablation on Chain-Local Experiences:				
R^*+Pivotal Exp.	67.5	54.2	68.7	63.5
R^*+Fragile Exp.	68.4	55.4	69.8	64.5
Ours	**69.5**	**56.3**	**70.5**	**65.4**

4.4 Test-Time Computational Cost

We evaluate test-time computational cost using Qwen3-VL-32B across two separate stages: the context construction for each ICD set and the online inference for each query. Table 3 reports token usage and latency averaged over the five benchmarks.

Table 3: Test-time computational cost with Qwen3-VL-32B. Context construction is measured per ICD set, and online inference per query.

Methods	Context Construction		Online Inference	
	Tokens ↓	Latency (s) ↓	Tokens ↓	Latency (s) ↓
Few-shot ICL	0	0	6,159	2.8
Few-shot CoT	21,794	21.4	23,468	10.1
ThoughtICR	63,973	224.5	22,483	11.6
Stochastic ICRL	25,779	31.6	18,940	9.0
Ours	47,732	48.7	8,031	3.7

In the context construction stage, compared to the second-best (accuracy) method, ThoughtICR, our Distill-ICL improves accuracy by 3.1% (see Table 1) while reducing token consumption by 25.4% and latency by 78.3%. These savings reflect its unified search over the ICD set, rather than separate searches and reasoning trajectories for individual ICDs.

It is important to note that context construction is performed only once per ICD set, whereas online inference is performed for each query; consequently, the online inference cost per query is particularly critical for evaluating real-world deployment efficiency. Compared to the most naive method, Few-shot ICL, our Distill-ICL improves accuracy by 9.3% (see Table 1) while adding only 0.9s online latency, resulting in computational costs much lower than all other reasoning-augmented methods. This efficiency demonstrates that the context distilled by Distill is compact and effective, rather than merely increasing computation.

4.5 Test-Time Scaling

We study the test-time scaling behavior of Distill-ICL on two benchmarks, SMMILE and FinMME, using two LVLM backbones, Qwen3-VL-8B and Qwen3-VL-32B. We consider two key computational budgets: the ICDs-MCTS search budget S for Normative Reasoning Chain search, and the number of controlled rollouts N for Chain-Local Experience Distillation. In Fig. 2, increasing either S or N can consistently improve performance. A larger S enables ICDs-MCTS to explore a broader

task-specific reasoning-action space, while a larger N provides more observations of chain-local execution behaviors along R^*, leading to more sufficient Chain-Local Experiences. Furthermore, Fig. 2 shows that Distill-ICL exhibits better search-budget scalability than ThoughtICR, another MCTS-based ICL method. Comparisons between Qwen3-VL-8B and Qwen3-VL-32B indicate that stronger LVLMs benefit more from test-time scaling, suggesting that Distill-ICL can better unlock the reasoning potential of more capable backbones. Balancing performance and computational cost, we set $S = 50$ and $N = 20$ as the default configuration.

4.6 Robustness to ICD Perturbations

Numerous studies (Rieff et al., 2025; Tian et al., 2025) have shown that standard Few-shot ICL is sensitive to the selection and ordering of ICDs, severely compromising its reliability. To evaluate whether different methods mitigate such instability, we conduct two perturbation experiments: (I) we fix queries and randomly resample ICDs from the same task type, evaluating robustness to demonstration selection; and (II) we fix selected ICDs for each query but randomly permute their order in the context, evaluating robustness to position bias. Since SMMILE contains limited same-task instances, we perform this analysis on the remaining four benchmarks. We set Qwen3-VL-32B as the LVLM backbone and repeat each perturbation five times. For each trial, we average accuracy over the four benchmarks, and Fig. 3 reports the minimum, maximum, and average accuracy across the five trials. The standard Few-shot ICL exhibits large performance fluctuations, confirming its sensitivity to both demonstration selection and ordering. In contrast, Distill-ICL significantly reduces this variability while improving both minimum and average accuracy. This robustness is attributed to abstracting raw ICDs into a shared Normative Reasoning Chain and Chain-Local Experiences, which provide order-insensitive reasoning guidance rather than relying on position-sensitive instance imitation.

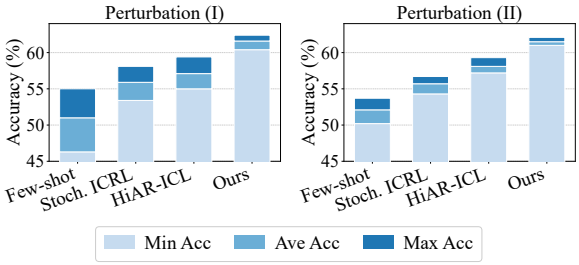


Figure 3: Multiple evaluations of the two types of ICD perturbations. We report the average performance over four benchmarks excluding SMMILE.

4.7 Cross-Model Transferability

We further evaluate the cross-model transferability of the explicit guidance distilled by Distill-ICL. Specifically, we use a source LVLM to distill the Normative Reasoning Chain and Chain-Local Experiences, and apply them to a different target LVLM during inference. As shown in Table 4, distilled guidance is generally transferable between models: distilled guidance from a source model can also improve the performance of another target LVLM on the corresponding few-shot ICL baseline. Moreover, we observe a "teacher-student" effect: the guidance distilled from the stronger model (Qwen3-VL-32B) has a better improvement effect (55.9 *vs.* 55.2) on Qwen3-VL-8B than the guiding information generated by Qwen3-VL-8B itself. This effect does not emerge when transferring from GPT-5.6-terra to Qwen3-VL-8B, possibly due to larger discrepancies in training data or model capabilities across model families. However, compared to the standard few-shot ICL setting, this cross-family transfer can also lead to a significant performance improvement (54.3 *vs.* 45.7).

Table 4: Cross-model transferability of distilled guidance. We report average results across five benchmarks.

Target Model	Source Model	Avg.	△
Qwen3-VL-8B	Few-shot ICL	45.7	0
	Qwen3-VL-8B	55.2	+9.5
	Qwen3-VL-32B	55.9	+10.2
	GPT-5.6-terra	54.3	+8.6

5 Conclusion

In this work, we propose *Distill-ICL*, a training-free framework for reliable few-shot multimodal ICL in specialized downstream tasks. Move beyond generating reasoning guidance independently for each ICD, Distill-ICL distills the ICD set into a shared Normative Reasoning Chain and further derives Chain-Local Experiences along this chain. This enables LVLMs to follow a task-level reasoning guidance while receiving localized visual-attribute support at critical steps. Experiments on five benchmarks show that Distill-ICL consistently achieves SOTA performance across various LVLMs and exhibits improved robustness to ICD perturbations. Distill-ICL also shows high computational efficiency, which is of great practical significance for deployment.

AI USE STATEMENT

In this work, we used generative AI tools for [assist in the writing of proofs, assist with translation]. We have not used generative AI tools for [help develop theoretical models or conceptual frameworks, design or provide feedback on research methodology or experiments, implement methods, clean and reformat dataset, support qualitative and thematic data analysis, interpret results], and [Generate synthetic data sets, formulate mathematical claims, provide critical ingredients for proving mathematical claims, propose or refine hypotheses] are not applicable to this work. Additionally, we used generative AI tools for [create or modify scientific figures or images, brainstorming, sourcing/searching for information, edit a research paper to improve readability]. We have reviewed all AI-assisted work. [We have reviewed all the text content generated by AI, and aligned the tables produced by it with the actual experimental results.]. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Sec. 3 describes the overall framework and its individual components, while the appendix provides the prompt templates used in our method. Hyperparameter settings are reported in the "Implementation Details" paragraph of Sec. 4.1. Our experiments use five publicly available datasets, and we only reorganized them to accommodate the few-shot ICL setting. "Benchmarks" paragraph of Sec. 4.1 details the data reorganization strategy and evaluation protocol, and the corresponding experimental subsections describe the implementation of method variants and the evaluation metrics. To account for variability in ICL performance, we report results averaged over five randomly sampled ICL episodes. We will release the code and data splits upon acceptance and include a link to the public repository in the camera-ready version.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 23716–23736, 2022.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, et al. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 17682–17690, 2024.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901, 2020.

Rémi Coulom. Efficient selectivity and backup operators in monte-carlo tree search. In *International conference on computers and games*, pp. 72–83. Springer, 2006.

Yuhao Dong, Shulin Tian, Shuai Liu, Shuangrui Ding, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Jiaqi Wang, and Ziwei Liu. Demo-icl: In-context learning for procedural video knowledge acquisition. *arXiv preprint arXiv:2602.08439*, 2026.

Shihan Dou, Ming Zhang, Zhangyue Yin, Chenhao Huang, Yujiong Shen, Junzhe Wang, Jiayi Chen, Yuchen Ni, Junjie Ye, Cheng Zhang, et al. Cl-bench: A benchmark for context learning. *arXiv preprint arXiv:2602.03587*, 2026.

Google DeepMind. Gemini 3.1 flash-lite. technical report, 2025.

Shibo Hao, Yi Gu, Haodi Ma, Joshua Jiahua Hong, Zhen Wang, Daisy Zhe Wang, and Zhiting Hu. Reasoning with language model is planning with world model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 8154–8173, 2023.

Xi Jiang, Jian Li, Hanqiu Deng, Yong Liu, Bin-Bin Gao, Yifeng Zhou, Jialin Li, Chengjie Wang, and Feng Zheng. Mmad: A comprehensive benchmark for multimodal large language models in industrial anomaly detection. In *International conference on learning representations*, volume 2025, pp. 87273–87295, 2025.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pp. 611–626, 2023.

Shengyuan Liu, Boyun Zheng, Wenting Chen, Zhihao Peng, Zhenfei Yin, Jing Shao, Jiancong Hu, and Yixuan Yuan. Endobench: A comprehensive evaluation of multi-modal large language models for endoscopy analysis. *arXiv preprint arXiv:2505.23601*, 2025.

Junyu Luo, Zhizhuo Kou, Liming Yang, Xiao Luo, Jinsheng Huang, Zhiping Xiao, Jingshu Peng, Chengzhong Liu, Jiaming Ji, Xuanzhe Liu, et al. Finmme: Benchmark dataset for financial multimodal reasoning evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 29465–29489, 2025.

Zehong Ma, Shiliang Zhang, Longhui Wei, and Qi Tian. Efficient multi-modal long context learning for training-free adaptation. In *International Conference on Machine Learning*, pp. 42236–42251. PMLR, 2025.

Meta AI. Llama-3.2-11B-Vision-Instruct. Hugging Face model card, September 2024. URL https://huggingface.co/meta-llama/Llama-3.2-11B-Vision-Instruct.

Giovanni Monea, Antoine Bosselut, Kianté Brantley, and Yoav Artzi. LLMs are in-context bandit reinforcement learners. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=c0RsezY2D1.

OpenAI. Using GPT-5.6. OpenAI API Documentation, 2026. URL https://developers.openai.com/api/docs/guides/latest-model?model=gpt-5.6. Accessed: 2026-09-25.

Melanie Rieff, Maya Varma, Ossian Rabow, Subathra Adithan, Julie Kim, Ken Chang, Hannah Lee, Nidhi Rohatgi, Christian Bluethgen, Mohamed S Muneer, et al. Smmile: An expert-driven benchmark for multimodal medical in-context learning. In *39th Annual Conference on Neural Information Processing Systems (NeurIPS 2025)*, 2025.

Syed Nazmus Sakib, Nafiul Haque, Mohammad Zabed Hossain, and Shifat E Arman. Plantvillagevqa: A visual question answering dataset for benchmarking vision-language models in plant science. *arXiv preprint arXiv:2508.17117*, 2025.

Yan Tai, Weichen Fan, Zhao Zhang, and Ziwei Liu. Link-context learning for multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27176–27185, 2024.

Xinyu Tian, Shu Zou, Zhaoyuan Yang, and Jing Zhang. Identifying and mitigating position bias of multi-image vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10599–10609, 2025.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Jinyang Wu, Mingkuan Feng, Shuai Zhang, Feihu Che, Zhengqi Wen, Chonghua Liao, Ling Yang, Haoran Luo, Zheng Lian, and Jianhua Tao. Beyond examples: Towards automated thought-level in-context reasoning for large language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2955–2995, 2026.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822, 2023.

Haozhe Zhao, Zefan Cai, Shuzheng Si, Xiaojian Ma, Kaikai An, Liang Chen, Zixuan Liu, Sheng Wang, Wenjuan Han, and Baobao Chang. MMICL: Empowering vision-language model with multi-modal in-context learning. In *International Conference on Learning Representations*, 2024.

Andy Zhou, Kai Yan, Michal Shlapentokh-Rothman, Haohan Wang, and Yu-Xiong Wang. Language agent tree search unifies reasoning acting and planning in language models. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

APPENDIX

A DATASET DETAILS

Table 5 summarizes the details of the five benchmarks and the corresponding ICD-query grouping strategies used in this work. Fig. 4 visually presents three representative ICD-query groups from three datasets. In each ICD sample within the ICD set, there is an image, a question, and an answer. The target query consists of an image, a question, and the answer options.

Table 5: Dataset details and ICD–query grouping strategy.

Dataset	Coverage	Grouping Strategy
SMMILE	111 medical problems across 6 specialties and 13 imaging modalities	Same medical problem
EndoBench	4 endoscopy scenarios and 12 major clinical tasks types	Same clinical task type
FinMME	18 financial domains and 6 asset classes	Same financial domain & asset classes
MMAD	38 product categories	Same product category
PlantVillageVQA	20 normalized disease label and a health label	Same disease/normal label

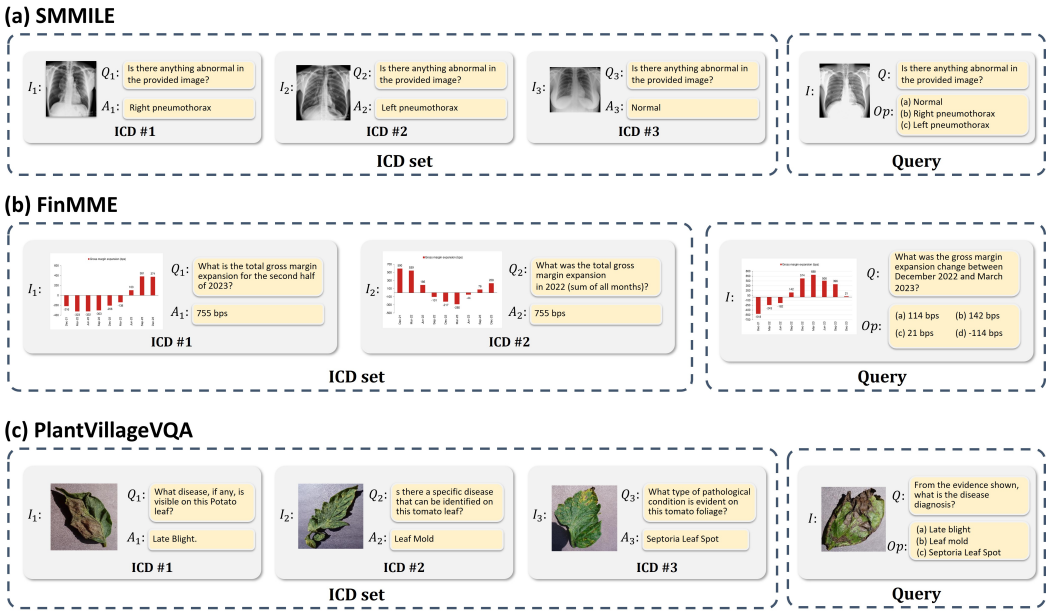


Figure 4: ICD–query visualization examples (a) SMMILE, (b) FinMME, and (c) PlantVillageVQA.

Here, we detail the five benchmarks used in this work:

SMMILE. SMMILE (Rieff et al., 2025) is an expert-curated benchmark for multimodal medical in-context learning. The public benchmark contains 111 problems and 517 image–question–answer triplets across six medical specialties and 13 imaging modalities. Each problem provides a target query and a small set of related demonstrations. We keep the official ICD-query organization. Fig 4(a) presents an ICD-query group related to the interpretation of chest X-ray images, consisting of three demonstrations and one query.

EndoBench. EndoBench (Liu et al., 2025) evaluates multimodal understanding of gastroscopy, colonoscopy, capsule endoscopy, and surgical endoscopy. It contains 12 major clinical tasks: (1) anatomical structure recognition (organ identification, landmark identification); (2) lesion analysis

and grading (lesion quantification, lesion type identification, lesion severity grading); (3) spatial localization and region understanding (visual grounding, region selection, region recognition); and (4) surgical workflow and operation analysis (preoperative assessment, macro phases identification, micro operation analysis, instrument management). We reorganize and obtain the ICD-query group according to the officially defined question types of each samples.

FinMME. FinMME (Luo et al., 2025) is a financial multimodal reasoning benchmark that contains 11,099 examples across 18 financial domains and six asset classes. The ICD set and the query sample are paired within the financial domain and asset classes. Fig. 4(b) shows arithmetic ICD-query gropu about gross margin expansion in basis points.

MMAD. MMAD (Jiang et al., 2025) evaluates industrial anomaly understanding across 38 product categories from four source datasets. Each product category has its specific defect vocabulary as well as normal. We organize the ICD-query groups according to the product categories.

PlantVillageVQA. PlantVillageVQA (Sakib et al., 2025) contains 55,448 plant images, covering 14 types of crops and 21 normalized labels, including 20 disease labels and one healthy label. Table 6 lists all 21 normalized labels. Fig. 4(c) illustrates disease identification using potato and tomato leaves. We converted the original free-text questions into multiple-choice questions for evaluation.

Table 6: List of 21 normalized labels of PlantVillageVQA.

	Label		Label
1	Apple scab	12	Leaf blight (Isariopsis leaf spot)
2	Bacterial spot	13	Leaf scorch
3	Black rot	14	Northern leaf blight
4	Cedar apple rust	15	Powdery mildew
5	Cercospora leaf spot / gray leaf spot	16	Septoria leaf spot
6	Common rust	17	Two-spotted spider-mite damage
7	Early blight	18	Target spot
8	Esca (black measles)	19	Tomato yellow leaf curl virus
9	Huanglongbing (citrus greening)	20	Tomato mosaic virus
10	Late blight	21	Healthy
11	Leaf mold		

B Prompt Templates

This section reports the prompt templates used by DISTILL-ICL, including prompt templates for ICDs-MCTS expansion, chain-local visual-attribute abstraction, and final model inference.

B.1 Prompt for ICDs-MCTS Expansion

Prompt 1: ICDs-MCTS Expansion

System Prompt

You are a multimodal reasoning-action proposer inside an ICDs-MCTS controller. Given one in-context demonstration and the current reasoning-action prefix, propose candidate next actions that could extend the prefix. Each action must describe a reusable visual or semantic check, not a final answer. The MCTS controller will canonicalize, aggregate, score, backpropagate, and select actions outside this prompt.

User Prompt

Generate candidate next reasoning actions for expanding the current ICDs-MCTS node.

Continued on next page...

(Continued from previous page)

<Task Description>
[Insert task type, answer format, and option schema]
</Task Description>

<ICD Demonstration D_m>
Image: [Insert ICD image]
Question: [Insert ICD question]
Options: [Insert option texts, if available]
Gold answer: [Insert ICD answer text]
</ICD Demonstration D_m>

<Current MCTS Prefix>
[Insert current prefix $h(v) = (a^{(1)}, ..., a^{(t)})$, or NONE at the root]
</Current MCTS Prefix>

<Existing Canonical Actions>
[Insert existing canonical actions under this node, if any]
</Existing Canonical Actions>

Propose diverse next-step actions that are useful for solving this ICD and can transfer to other ICDs from the same task. Avoid actions that directly mention the gold answer unless the answer text itself is a visible class label to be verified. Do not estimate rewards or choose the final tree branch.

Output Constraint

Return a JSON object with these keys:

- `candidate_actions`: a list of candidate next actions.
- For each action, include `action_text`, `visual_check`, `why_next`, and `merge_hint`.
- `forbidden_fields`: confirm that no final answer, reward, value estimate, or branch-selection decision is returned.

B.2 Prompt for Chain-Local Experience Abstraction

Prompt 2: Chain-Local Visual Attribute Abstraction

System Prompt

You are a chain-local experience abstractor. Given a critical node on the normative chain, summarize the visual attributes that explain reliable execution at that node. The critical node has already been identified algorithmically by Jensen-Shannon divergence over rollout conclusions; do not select a new node. This single template supports both Inter-ICD Pivotal Experience and Intra-ICD Fragile Experience.

User Prompt

Abstract chain-local visual attributes from the supplied rollout evidence.

<Task Description>
[Insert task type, category, answer format, and option schema]
</Task Description>

<Experience Type>
[Insert inter_icd_pivotal or intra_icd_fragile]
</Experience Type>

<Normative Chain>
[Insert $R^* = (a^{*(1)}, ..., a^{*(L)})$]
</Normative Chain>

<Selected Critical Node>
Node index: [Insert k]
Chain action: [Insert $a^{*(k)}$]

Continued on next page...

(Continued from previous page)

Node statistics: [Insert divergence score and supporting rollout counts]
</Selected Critical Node>

<Rollout Evidence>
For `inter_icd_pivotal`: [Insert two or more ICDs with image, question, options, answer text, and successful conclusions at node k]
For `intra_icd_fragile`: [Insert one ICD with image, question, options, answer text, successful conclusions at node k, and failed conclusions at node k]
</Rollout Evidence>

If the experience type is `inter_icd_pivotal`, condense the divergent successful conclusions into answer-discriminative visual attributes for the compared ICDs and link each attribute to its image and answer text. If the experience type is `intra_icd_fragile`, abstract the visual attributes shared by successful rollouts, contrast them with failed conclusions, and link the missed or misinterpreted attribute to the same ICD image. Use option text or semantic answer names rather than option letters.

Output Constraint

Return a JSON object with these keys:

- `experience_type`: `inter_icd_pivotal` or `intra_icd_fragile`.
- `selected_node`: node index, chain action, and supplied divergence statistic.
- `visual_attributes`: concise attributes grounded in the provided image evidence.
- `image_linked_experience`: image or sample ID, node action, attribute text, and associated answer text when applicable.
- `usage_guidance`: a short step-local cue, contrast, or warning to be used during test-time inference.
- `scope`: task type, category, and option schema where the experience applies.

B.3 Prompt for Test-Time Guided Inference

Prompt 3: Test-Time Guided Inference

System Prompt

You are a multimodal reasoning assistant at test time. Answer the target question by combining three sources of context: few-shot ICDs, the normative reasoning chain, and chain-local experiences. Follow the normative chain step by step. Use experiences only as localized cues, contrasts, or warnings for executing the corresponding chain steps. The final answer must be grounded in the target image and target options, not copied from support labels or experience examples.

User Prompt

Answer the target query using the provided context.

<Task Description>
Task type: [Insert task type]
Answer format: [Insert answer format]
Option schema: [Insert target option schema]
</Task Description>

<In-Context Demonstrations>
[Insert each ICD image, question, options, and answer]
</In-Context Demonstrations>

<Normative Chain R^* >
[Insert $R^* = (a^{*(1)}, ..., a^{*(L)})$]
</Normative Chain R^* >

<Chain-Local Experiences>
[Insert Inter-ICD Pivotal Experiences and Intra-ICD Fragile Experiences, each linked to chain node, support image or sample ID,

Continued on next page...

(Continued from previous page)

`visual attribute, and usage guidance]`
</Chain-Local Experiences>

<Target Sample>
Image: `[Insert target image here]`
Question: `[Insert target question here]`
Options: `[Insert target option texts here]`
</Target Sample>

Follow R^* step by step. At each node, inspect the target image, write the target-image observation for that node, and then check whether any chain-local experience gives a relevant cue or warning. Apply an experience only when its visual condition is actually observed in the target image. Do not use support-image labels, answer frequency, or rollout success labels as evidence for the target.

Output Constraint

Return a JSON object with these keys:

- `chain_observations`: one target-image observation for each node in R^*.
- `experience_usage`: experiences used or ignored, with a short reason.
- `target_evidence`: decisive evidence from the target image only.
- `final_answer`: one option letter and answer text.
- `answer_tag`: `<answer>LETTER</answer>`.

C Additional Experimental Results

C.1 Hyperparameters Choices

Fig. 5 reports the average accuracy over five benchmarks under different hyperparameter choices: the pivotal-node threshold γ_p and the fragile-node threshold γ_f. When tuning one threshold, the other is fixed to its default value. The results show that Distill-ICL achieves the best average performance at $\gamma_p = 0.30$ and $\gamma_f = 0.25$, and thus this configuration is adopted as the default.

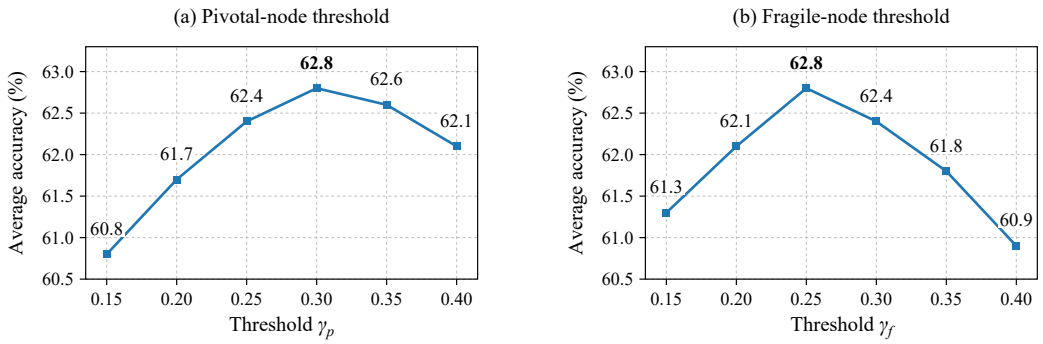


Figure 5: Sensitivity to the (a) pivotal-node threshold γ_p and (b) the fragile-node threshold γ_f using Qwen3-VL-32B. We report average accuracy over five benchmarks. When varying one threshold, the other is fixed at its default value ($\gamma_p = 0.30$, $\gamma_f = 0.25$).

C.2 Consistency of Performance Gains and Statistical Significance

Due to the relatively unstable performance of the few-shot ICL setting, we assess the reliability of the performance gains through paired comparisons at two complementary levels: episode-level consistency of performance gains and query-level statistical significance.

Experimental Setup. We evaluate Distill-ICL against the strongest baseline on five benchmarks using Qwen3-VL-32B and Gemini-3.1. All methods are evaluated on the same five ICL episodes,

Table 7: Paired comparisons between Distill-ICL and the strongest baselines. "Mean Gain ± Std" reports the mean and standard deviation of episode-level accuracy gains in percentage points. "Win Episodes" counts positive gains over five shared episodes. The final column reports query-level McNemar p-values. "*Avg.*" rows compare overall performance against ThoughtICR.

			Episode Level		Query Level
LVLMs	**Dataset**	**Strongest Baseline**	**Mean Gain ± Std**	**Win Episodes**	**McNemar p**
Qwen3-VL-32B	SMMILE	ThoughtICR	$+2.2 \pm 0.7$	$5/5$	1.2×10^{-4}
	EndoBench	ThoughtICR	$+2.9 \pm 0.9$	$5/5$	3.5×10^{-6}
	FinMME	ThoughtICR	$+2.3 \pm 1.0$	$5/5$	2.7×10^{-5}
	MMAD	ThoughtICR	$+3.8 \pm 1.2$	$5/5$	4.3×10^{-5}
	PlantVillageVQA	ThoughtICR	$+4.0 \pm 1.2$	$5/5$	6.1×10^{-6}
	Avg.	ThoughtICR	$+3.1 \pm 0.3$	$5/5$	4.4×10^{-8}
Gemini-3.1	SMMILE	ThoughtICR	$+0.7 \pm 0.3$	$5/5$	8.7×10^{-3}
	EndoBench	Stochastic ICRL	$+1.8 \pm 0.6$	$5/5$	1.9×10^{-5}
	FinMME	ThoughtICR	$+1.6 \pm 0.5$	$5/5$	4.0×10^{-4}
	MMAD	Stochastic ICRL	$+2.5 \pm 1.0$	$5/5$	7.8×10^{-5}
	PlantVillageVQA	Stochastic ICRL	$+1.3 \pm 0.6$	$5/5$	1.3×10^{-4}
	Avg.	ThoughtICR	$+1.8 \pm 0.2$	$5/5$	1.7×10^{-7}

with identical demonstrations and query sets within each episode. For each dataset–backbone pair, we select the baseline with the highest mean accuracy across the five episodes and use this same baseline throughout the paired comparison. For the overall comparisons reported in the *Avg.* rows, ThoughtICR (Wu et al., 2026) serves as the reference for both backbones, as it achieves the highest average baseline accuracy across the five benchmarks.

Episode-Level Consistency. For each episode e, we compute the paired accuracy gain:

$$\Delta_e = \mathrm{Acc}^{(e)}_{\text{Distill-ICL}} - \mathrm{Acc}^{(e)}_{\text{baseline}}, \qquad e \in \{1, \ldots, 5\}, \tag{9}$$

where accuracy is expressed as a percentage, so that Δ_e is measured in percentage points. We report two statistics: *Mean Gain ± Std*, the mean and standard deviation of the five paired gains, and *Win Episodes*, the number of episodes with $\Delta_e > 0$. The standard deviation characterizes variability across episodes and should not be interpreted as a confidence interval. Given the limited number of episodes, we use these statistics to describe the consistency of the gains rather than conduct episode-level significance tests.

Query-Level Significance. For each dataset–backbone pair, we pool predictions from the five shared ICL episodes and apply McNemar's test to the paired binary correctness outcomes of Distill-ICL and the selected baseline. The test compares two types of discordant outcomes: queries answered correctly only by Distill-ICL and queries answered correctly only by the baseline. The null hypothesis is that these two outcomes have equal probabilities. For the *Avg.* rows, we further pool paired predictions across all five benchmarks and compare Distill-ICL with ThoughtICR, the strongest baseline in terms of average benchmark accuracy. These aggregate p-values assess differences in correctness over the pooled queries, complementing the episode-level analysis of gain consistency.

Results. As shown in Table 7, Distill-ICL outperforms the selected baseline in all five episodes for every dataset–backbone pair. The mean gains range from 2.2% to 4.0% for Qwen3-VL-32B and from 0.7% to 2.5% for Gemini-3.1. All reported McNemar p-values fall below the significance threshold of 0.05, including the smallest observed gain on Gemini-3.1/SMMILE ($p = 8.7 \times 10^{-3}$). Together, these results provide evidence of consistent gains across the evaluated episodes and differences in paired query-level correctness.

D LIMITATIONS

Despite its strong performance in few-shot multimodal ICL, Distill-ICL has two main limitations: First, Distill-ICL primarily operates within a standardized ICD setting, requiring ICDs and queries

to share the same underlying task format. It is not intended for general context learning with heterogeneous instructions, broad background knowledge, or loosely related multimodal contexts. Second, Distill-ICL does not implement fine-grained visual grounding to local image regions. Because not all LVLMs possess reliable localization capabilities, we purposefully avoid this mechanism to maintain the broad compatibility with various LVLMs.