APPLENMR-TEXT: A COST-SENSITIVE LF-NMR BENCHMARK FOR APPLE MOLDY-CORE SCREENING

Anonymous authors
Paper under double-blind review

ABSTRACT

Apple moldy-core screening is a small-data, cost-sensitive task: the defect is internal, destructive verification is needed for labels, and expert diagnostic notes are unavailable as inputs for new fruit. We introduce AppleNMR-Text, a benchmark of 237 Red Fuji apples with LF-NMR slices, destructively verified labels, and expert-reviewed text, together with fixed cross-validation and quality-gated utility evaluation. As a documented image-input benchmark model, we study a two-stage pipeline: dual-memory retrieval guides structured rationale generation, then a classifier combines the image, generated rationale, and sanitized evidence. On the reported five-fold evaluation, this model obtains TAAPM 393.733 ± 13.973, versus 158.40 ± 224.01 for vision-only VGG19 and 155.33 ± 61.98 for directly fine-tuned Qwen3-VL-8B. Dual-memory retrieval outperforms either single memory on TAAPM. Balanced accuracy is 0.700 ± 0.019, so the utility advantage should not be read as uniform superiority on conventional metrics. AppleNMR-Text provides a controlled testbed for image-only screening and for separating an expert-text oracle from a deployable input protocol; its single-season cohort does not establish cross-domain generalization.

1 INTRODUCTION

Moldy core can develop inside an apple with little visible surface change. LF-NMR images expose internal structure, and destructive inspection supplies the final disease label. A screening benchmark must separate imaging, expert interpretation, ground truth, and the operational decision rule.

AppleNMR-Text retains 237 of 273 Red Fuji apples (79 diseased, 158 healthy), with expert-reviewed diagnostic text. They share a season, cold-storage protocol, operator, and acquisition day; transfer to other conditions remains untested. The data and corresponding labels are openly available in the supplementary materials.

At test time, the external input is an LF-NMR image; the expert note for that apple is privileged. Alongside conventional scores, TAAPM applies fixed contract parameters and a fixed decision threshold of 0.5 to distinguish defective shipments from unnecessary rejection of healthy fruit.

Our benchmark model retrieves fold-isolated evidence, generates a structured rationale, and classifies from the image, generated text, and sanitized evidence. The evaluation distinguishes this image-input pipeline from an expert-text oracle; the component algorithms are established methods.

Our contributions are:

1. **A controlled dataset and protocol.** AppleNMR-Text pairs LF-NMR imaging with destructive labels and expert-reviewed descriptions. We specify sample-level folds, image-only external inference, an image-and-text expert comparison, and a fixed cost-sensitive evaluation setting.
2. **An inspectable benchmark model.** The reported two-stage pipeline uses fold-isolated retrieval from an expert knowledge memory and a sanitized case memory, structured rationale generation, and a separate image-text classifier. Its intermediate objects expose the source of retrieved claims; citation validity alone is not a measure of factual support.
3. **Protocol-specific evidence.** Comparisons and ablations report both conventional metrics and TAAPM. They show a utility gain for the complete two-stage configuration on this

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.

cohort, reveal that dual memory is stronger than either individual memory in the reported ablation, and show where the method loses ground on balanced accuracy or when a component is used alone.

2 Related Work and Positioning

Prior apple moldy-core studies have investigated vibration–Vis/NIR fusion (Liu et al., 2024) and spectral-model transfer across origins (Zhang et al., 2023). Our benchmark studies LF-NMR images with expert text and a fixed utility protocol.

Class imbalance is addressed through reweighting, classifier adjustment, and class-aware objectives Cui et al. (2019); Menon et al. (2021); Ren et al. (2020); Kang et al. (2020). Supervised contrastive learning shapes embeddings Khosla et al. (2020), while EMA teachers stabilize learning Tarvainen & Valpola (2017). Vision-language pretraining links images and text Radford et al. (2021); Li et al. (2022; 2023); RAG provides retrieved evidence for generation Lewis et al. (2020). Our contribution combines a documented benchmark, expert-text protocol, cost-sensitive evaluation, and an image-input model with an intermediate evidence record. The experiments do not establish an algorithmic advantage outside this domain.

3 AppleNMR-Text: Data and Task

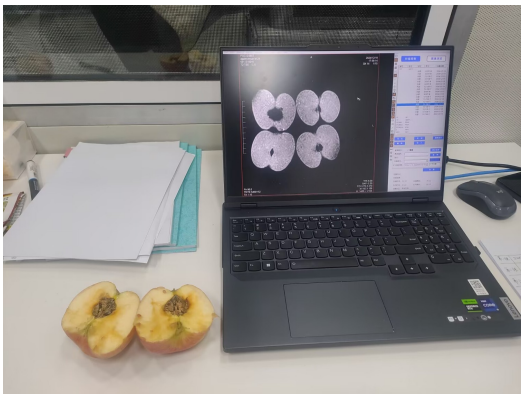
Cohort and ground truth. The study collected 273 individual apples and retained 237 with complete imaging and ground-truth records. Disease status was established after scanning by destructive inspection. The observed study prevalence is $79/237$, whereas the operational prevalence parameter used for TAAPM is $r_{\mathrm{bad}} = 0.1$ (Section 5). They represent different quantities. The 36 other samples were excluded for data completeness or quality control before training and evaluation.

Expert records and expert comparison. Three domain experts assigned a pre-dissection diagnosis using LF-NMR information, while the final disease label was determined after dissection. The authors report majority-vote and adjudication for disagreements. Expert-reviewed diagnostic text is the canonical paired text used by this benchmark; its controlled vocabulary records observed patterns, potential confounders, and uncertainty. For the image-and-text expert comparison in Table 3, experts review the LF-NMR image together with text, but do not see the ground-truth label. This reference therefore has a different input condition from the automated image-input pipeline. No inter-expert agreement statistic is claimed in this paper.

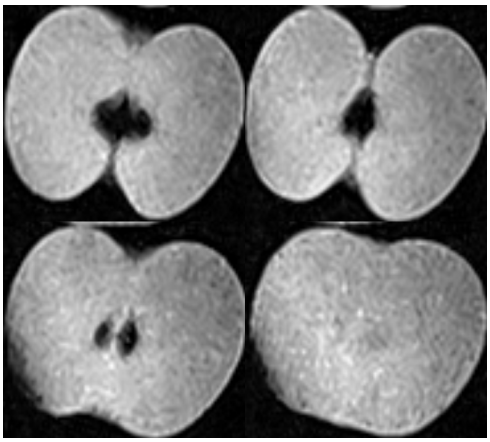
Acquisition and preprocessing. Scanning used 68 mT, a T2-weighted turbo-spin-echo protocol, 15 slices per apple, and 10 mm inter-slice spacing. Four apples were scanned in a fixed acquisition layout. For each apple, a detector finds the maximal outer bounding box across its stack. The selected box is propagated to all slices; four consecutive slices around the maximal response are assembled into a 2×2 mosaic and resized. Figure 1 illustrates acquisition and the processed image. A unique sample identifier links all slices, the mosaic, the expert record, and the destructive label.

Splits and augmentation. Five folds each hold out one test partition. Only the training partition is oversampled to approximately balance the two classes; the held-out partition keeps its observed composition. Dataset partitioning precedes augmentation, which is limited to rotations by multiples of 90° and horizontal or vertical reflection within training. The source fold counts are shown in Table 1; each fold's entries still refer to 237 distinct apples before training oversampling. Each seed (2024, 2025, and 2026) runs five-fold cross-validation. The five-fold results are first aggregated within each seed; the reported mean and standard deviation are then calculated across the three seed-level results.

Evaluation task. For apple i, the recorded triple is (x_i, t_i, y_i): x_i is the image mosaic, t_i is its expert diagnostic record, and $y_i \in \{0, 1\}$ is the post-dissection label. The deployed prediction receives x_i as its external input and may retrieve training-fold knowledge or cases through the fixed model pipeline. The expert text t_i is available for training and oracle analysis, but not for the image-input test prediction. This distinction is necessary to interpret the rows of Table 3.



(a) LF-NMR acquisition and dissected example.



(b) Processed 2×2 image mosaic.

Figure 1: Acquisition and LF-NMR representation from the supplied examples.

Table 1: Five-fold split counts supplied with AppleNMR-Text. "Balanced train" is the reported training size after diseased-class oversampling; test sizes are unchanged.

Fold	Training			Test			Balanced train
	Total	Healthy	Diseased	Total	Healthy	Diseased	size
1	190	127	63	47	31	16	254
2	190	127	63	47	31	16	254
3	190	126	64	47	32	15	252
4	189	126	63	48	32	16	252
5	189	126	63	48	32	16	252

4 Image-Input Benchmark Model

Figure 2 presents the two-stage information flow, and Figure 3 reproduces the detailed source overview. Knowledge and case memories are constructed for the fold; the test sample contributes an image but not its expert note or label. Stage I produces a structured rationale and an evidence record. Stage II makes the final binary prediction. We report the full descriptive interface and additional training detail in the supplementary material.

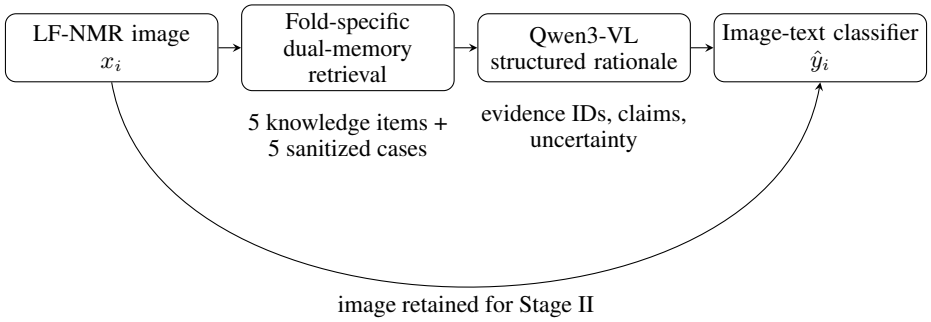


Figure 2: The external test input is an LF-NMR image. Stage I retrieves fold-specific evidence and generates a rationale; Stage II uses the image, rationale, and sanitized evidence for classification.

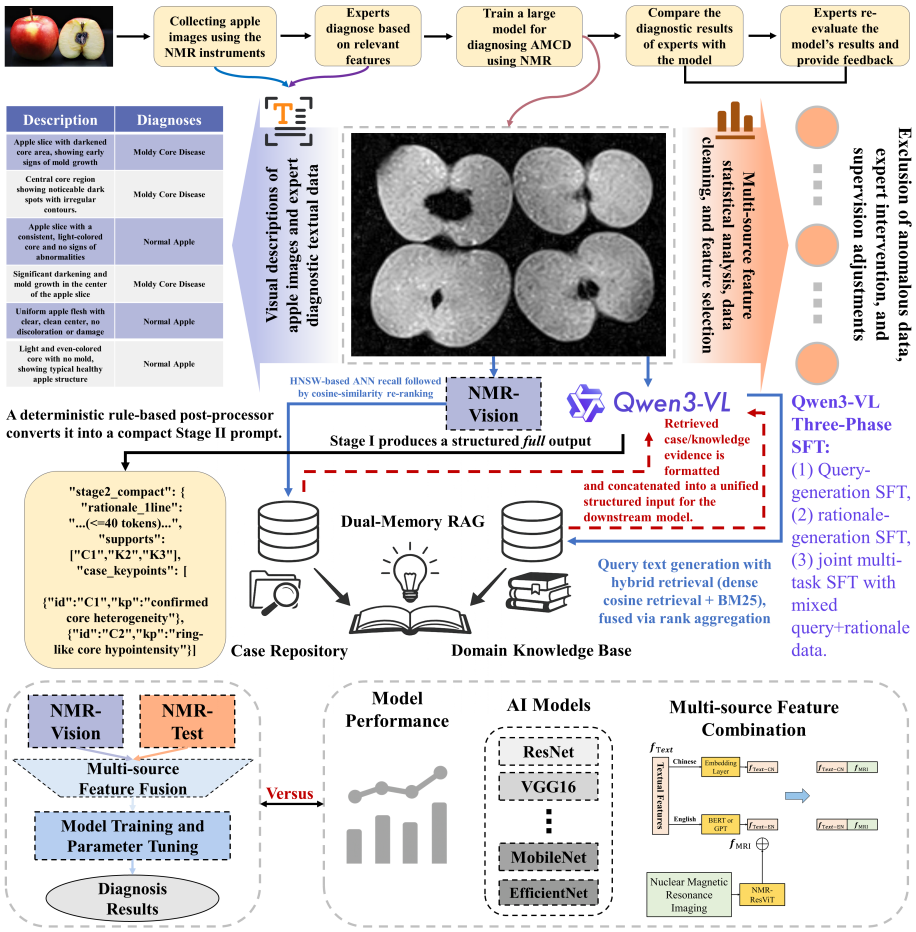


Figure 3: LF-NMR acquisition and screening pipeline. Embedded structured-text snippets illustrate the interface format; they are not held-out case outputs.

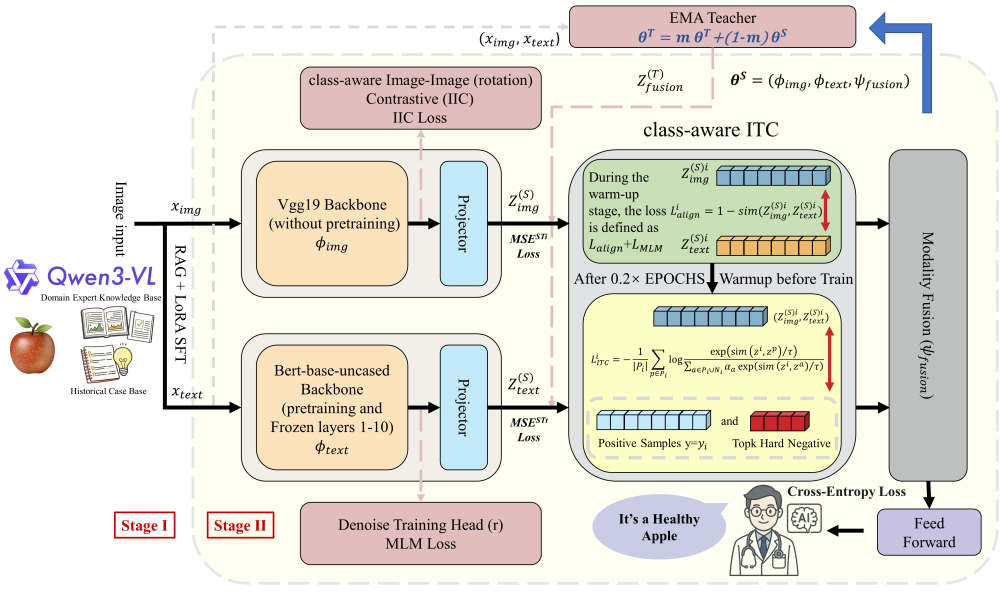


Figure 4: Stage II architecture with image and text encoders, contrastive objectives, EMA teacher, and fusion classifier.

4.1 STAGE I: FOLD-ISOLATED EVIDENCE AND RATIONALE

Let $\mathcal{M}_K$ contain expert diagnostic cues and confounders and $\mathcal{M}_C^{(k)}$ contain cases from training fold k. The retrieval and generation interface is

$$E_i^{(k)} = \mathcal{R}^{(k)}(x_i; \mathcal{M}_K, \mathcal{M}_C^{(k)}), \qquad r_i^{(k)} = \mathcal{G}_\theta^{(k)}(x_i, E_i^{(k)}). \tag{1}$$

The case branch searches image embeddings, then reranks and filters near-duplicate candidates. The knowledge branch combines dense retrieval and sparse term matching. The retrieved units form an Evidence Pack with persistent source identifiers. Retrieval selects five knowledge items and five case items ($k_K = k_C = 5$), for a total of ten items in the Evidence Pack.

The held-out fold is excluded from its retrieval bank; its expert notes and labels are never query inputs. Training queries also exclude every image of the same apple by apple ID. Case summaries remove verdict-equivalent fields before retrieval evidence reaches Stage I or Stage II. Retrieval is present in generator training and inference. The supplement specifies the audit rules.

Evidence Pack. Each retrieved knowledge snippet or sanitized case has an evidence ID, visible text, and provenance metadata. Case verdicts are excluded from model-visible text; a valid citation ID alone does not establish factual support for a generated claim.

Table 2: Stage I output and its use. Field names are taken from the reported interface; this table describes the protocol and does not report a measured fidelity rate.

Field	Role in the reported pipeline
`image_findings`	Descriptive findings conditioned on the test image.
`supported_claims`	Claims with identifiers of retrieved evidence items.
`uncertainties`	Ambiguity and confounders retained in the rationale.
`used_evidence_ids`	Citation set checked against the retrieved pack and kept for audit.
`confidence`, `final_rationale`	Stated confidence and compact natural-language interpretation.

Validation boundary. The generated record is checked against the retrieved identifier set before Stage II. This ensures traceability of a citation, but does not measure whether the cited source supports the generated claim.

The final two-stage system uses the 8B generator, trained with LoRA (Hu et al., 2022): rank 64, learning rate 10^{-4}, batch size 32, and 50 epochs. The 2B/4B/8B settings in Table 3 are direct-classification scale comparisons.

4.2 STAGE II: DECISION FROM IMAGE AND GENERATED TEXT

The adopted image backbone is VGG19 (Simonyan & Zisserman, 2015) trained from scratch on LF-NMR mosaics at 384×384 resolution. A BERT-base-uncased text encoder (Devlin et al., 2019) freezes its lower ten transformer layers while updating the upper two and a projection head. Image and text features are projected to a shared embedding space of dimension 2048 and combined by a gated fusion module. The Stage II input uses the image, compact generated rationale, and retrieved evidence text. Retrieved case labels are excluded from the classifier text stream.

The objective combines binary classification with class-aware image-image contrast (IIC), image-text contrast (ITC), masked-language modeling (MLM), and EMA-teacher consistency:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda_{\text{IIC}}\mathcal{L}_{\text{IIC}} + \lambda_{\text{ITC}}\mathcal{L}_{\text{ITC}} + \lambda_{\text{MLM}}\mathcal{L}_{\text{MLM}} + \lambda_{\text{dist}}\mathcal{L}_{\text{dist}}. \tag{2}$$

IIC uses two augmented views and same-class positives; MLM masks 10% of text tokens. Stage II also trains for 50 epochs using AdamW (Loshchilov & Hutter, 2019). The first 20% warm up image-text alignment; later training uses class-aware ITC and top-16 hard negatives. EMA momentum is 0.95, and runs use seeds 2024–2026. Figure 4 shows Stage II.

At inference, the classifier produces $p_i = \mathcal{F}_\phi(x_i, r_i, E_i)$ and $\hat{y}_i = \mathbb{I}[p_i \geq \tau]$. The decision threshold is fixed at $\tau = 0.5$ for all experiments. The expert-text oracle replaces generated text with the expert record at test time and therefore answers a different, privileged-input question.

5 COST-SENSITIVE EVALUATION PROTOCOL

We report balanced accuracy, F1, MCC, and accuracy in addition to TAAPM. The latter evaluates a shipping decision with an asymmetric quality requirement. Given a defect score and the fixed threshold $\tau = 0.5$, an apple with score at least τ is rejected; otherwise it is shipped. Define $R_{\mathrm{bad}}(\tau)$ as the rejection rate of diseased fruit and $F_{\mathrm{good}}(\tau)$ as the rejection rate of healthy fruit. Under the fixed operational prevalence r_{bad}, the fraction shipped and the defective fraction of that shipment are

$$k_{\mathrm{ship}}(\tau) = (1 - r_{\mathrm{bad}})(1 - F_{\mathrm{good}}(\tau)) + r_{\mathrm{bad}}(1 - R_{\mathrm{bad}}(\tau)), \tag{3}$$

$$d_{\mathrm{ship}}(\tau) = \frac{r_{\mathrm{bad}}(1 - R_{\mathrm{bad}}(\tau))}{k_{\mathrm{ship}}(\tau)}. \tag{4}$$

TAAPM assigns zero utility if the buyer's maximum defective fraction α is exceeded; when the gate is satisfied, it subtracts the procurement required to ship a fixed quantity Q:

$$\mathrm{TAAPM}(\tau) = \begin{cases} Qp_{\mathrm{sell}} - \dfrac{Q}{k_{\mathrm{ship}}(\tau)} c_{\mathrm{buy}}, & d_{\mathrm{ship}}(\tau) \le \alpha, \\ 0, & d_{\mathrm{ship}}(\tau) > \alpha. \end{cases} \tag{5}$$

The supplied configuration fixes $Q = 1000$, $p_{\mathrm{sell}} = 1$, $c_{\mathrm{buy}} = 0.4$, $c_{\mathrm{extra}} = 0$, $\alpha = 0.05$, and $r_{\mathrm{bad}} = 0.1$. The extra-cost term is zero and does not appear in Eq. 5. The r_{bad} parameter is an operational prior, not the 79/237 diseased proportion in the curated cohort. The quantities and units should be interpreted within this contract model, not as measured factory profit.

The cost parameters were fixed before evaluation and applied identically across methods. The decision threshold is fixed at $\tau = 0.5$ for all experiments; validation and test labels are not used to select it. TAAPM is evaluated at this fixed threshold. Test-selected weights in the separate "different weight vote" diagnostic are excluded from comparisons. The supplement derives TAAPM in full.

Interpreting the quality gate. Missing diseased fruit can violate the defective-shipment cap; false rejection of healthy fruit increases the procurement required for Q accepted apples. TAAPM can therefore rank models differently from accuracy, and zero TAAPM does not imply zero accuracy. It is a contract-based score, not measured production-line profit.

6 EXPERIMENTS AND FINDINGS

6.1 COMPARISONS UNDER DIFFERENT INPUT CONDITIONS

The automated image-input comparison includes vision-only VGG19, direct image prompting or fine-tuning of Qwen3-VL, GPT-5 API prompting, and our two-stage model. The text-only BERT row consumes expert text and the Stage II oracle consumes expert text at inference; these rows are privileged-input references. Human experts review images and text without the ground-truth label. Table 3 reports the complete comparison under these distinct input conditions. The GPT-based models are GPT-5-Instant and GPT-5-Thinking; their calls were made on November 23, 2025.

Table 3: Overall comparison. Values are mean ± standard deviation across three seeds after aggregating the five folds within each seed. The BERT and Stage II oracle rows use expert text. Human experts review the LF-NMR image and text, without the ground-truth label. The two-stage model uses only the image as its external test input. ZS: zero-shot; FT: fine-tuned.

Model	Bal-acc	F1-score	MCC	ACC	TAAPM
Unimodal baselines					
Text-only (BERT)	0.771±0.029	0.712±0.043	0.600±0.065	0.771±0.029	147.60±119.88
Vision-only (VGG19)	0.703±0.034	0.604±0.047	0.389±0.053	0.681±0.035	158.40±224.01
Comparison with human experts					
Human experts	0.851±0.011	0.811±0.015	0.733±0.021	0.885±0.009	413.067±15.032
Stage II oracle	0.704±0.001	0.611±0.002	0.386±0.003	0.699±0.002	409.067±0.377
Two-stage (ours)	0.700±0.019	0.608±0.022	0.382±0.038	0.700±0.019	393.733±13.973
GPT-5 API baselines					
GPT-5-Instant	0.452±0.017	0.396±0.016	−0.095±0.033	0.404±0.014	0.000±0.000
GPT-5-Thinking	0.681±0.039	0.565±0.054	0.383±0.092	0.737±0.041	166.533±43.404
Qwen3-VL baselines					
Qwen3-VL-2B (ZS)	0.412±0.000	0.442±0.000	−0.228±0.000	0.329±0.000	0.000±0.000
Qwen3-VL-2B (FT)	0.594±0.011	0.431±0.021	0.202±0.023	0.665±0.011	−45.875±79.443
Qwen3-VL-4B (ZS)	0.505±0.000	0.488±0.000	0.0188±0.000	0.362±0.000	0.000±0.000
Qwen3-VL-4B (FT)	0.568±0.020	0.404±0.046	0.142±0.036	0.632±0.004	0.000±0.000
Qwen3-VL-8B (ZS)	0.500±0.000	0.495±0.000	—	0.329±0.000	0.000±0.000
Qwen3-VL-8B (FT)	0.629±0.019	0.478±0.032	0.284±0.040	0.718±0.012	155.33±61.98

Among the listed automated image-input models that follow the fixed protocol, the two-stage system obtains the highest reported TAAPM of 393.733 ± 13.973. Its mean exceeds vision-only VGG19 (158.40 ± 224.01) by 235.333 metric units, and directly fine-tuned Qwen3-VL-8B (155.33 ± 61.98) by 238.403. These are differences of reported means, not paired significance tests. The two-stage model's balanced accuracy of 0.700 ± 0.019 is below vision-only VGG19's 0.703 ± 0.034 and far below the text-only BERT reference's 0.771 ± 0.029. Its ACC of 0.700 ± 0.019 is also lower than the image-and-text expert comparison's 0.885 ± 0.009. Utility and conventional discrimination emphasize different error patterns under the quality gate.

Table 3 lists the remaining API and Qwen3-VL variants. Several zero-shot rows have zero TAAPM despite nonzero accuracy; Qwen3-VL-8B zero-shot, for example, has ACC 0.329 ± 0.000. This distinction follows from the quality gate.

The privileged expert-text oracle reaches TAAPM 409.067 ± 0.377, a mean gap of 15.334 above the image-input system; human experts with image-and-text inputs reach 413.067 ± 15.032. These results neither establish human-level image-only diagnosis nor identify which pipeline component explains the gap.

6.2 Classifier objectives and operating utility

Table 4 shows that IIC alone lowers TAAPM from 358.533 ± 17.925 to 345.200 ± 32.811, whereas the full configuration reaches 409.067 ± 0.377. A possible explanation for the IIC result is weaker alignment or calibration, but that mechanism was not independently tested. MLM and combined objective results are in Table 4.

Table 4: Stage II objective ablation. Values are mean ± standard deviation across three seeds after aggregating the five folds within each seed. Colors compare each metric with mid-fusion: green is baseline, blue is lower, and red is higher. Class-aware ITC omits HardNeg and EMA; the next row adds HardNeg, and full Stage II adds the EMA teacher.

Variant	Bal-acc	F1-score	MCC	ACC	TAAPM
Baseline mid-fusion	0.712±0.007	0.617±0.009	0.410±0.012	0.727±0.005	358.533±17.925
+ IIC only	0.705±0.022	0.608±0.028	0.398±0.041	0.723±0.017	345.200±32.811
+ MLM only	0.715±0.010	0.622±0.011	0.413±0.021	0.724±0.014	379.600±7.222
+ IIC & MLM only	0.722±0.004	0.630±0.005	0.428±0.007	0.733±0.002	382.133±9.309
+ Class-aware ITC	0.707±0.014	0.613±0.018	0.393±0.026	0.706±0.011	399.067±34.832
+ ITC + HardNeg	0.693±0.004	0.599±0.004	0.365±0.008	0.685±0.005	403.467±5.507
Full Stage II (+ EMA)	0.704±0.001	0.611±0.002	0.386±0.003	0.699±0.002	409.067±0.377

Balanced accuracy is 0.704 ± 0.001 for full Stage II, below mid-fusion (0.712 ± 0.007) and IIC with MLM (0.722 ± 0.004). The TAAPM gain does not imply each loss improves every metric. Full Stage II has ACC 0.699 ± 0.002.

The class-aware ITC configuration without HardNeg or EMA has TAAPM 399.067 ± 34.832. Adding HardNeg gives 403.467 ± 5.507, and the full configuration with EMA gives 409.067 ± 0.377. These rows show the utility of the reported combined training configuration. Conventional metrics capture a different trade-off: IIC with MLM has F1 0.630 ± 0.005 and ACC 0.733 ± 0.002, compared with 0.611 ± 0.002 and 0.699 ± 0.002 for full Stage II. Thus the utility ranking should not be presented as a uniform improvement in discrimination.

6.3 What the evidence-source ablations show

Table 5 gives TAAPM 288.667 without RAG, 338.933 with knowledge retrieval, 345.733 with case retrieval, and 393.733 with both. The dual-memory benefit is specific to TAAPM: no RAG has balanced accuracy 0.704 ± 0.014 versus 0.700 ± 0.019 for dual memory.

Removing or shuffling the rationale lowers mean TAAPM to 130.266 ± 225.628 or 260.266 ± 225.585, with high variability. Adding RAG only at inference yields 263.333 ± 228.108, below 393.733 ± 13.973 with retrieval in generator training. These tests do not measure sentence-level factual fidelity.

The output-format rows provide a further comparison: free-form rationale reports TAAPM 263.467 ± 228.228, while structured rationale reports 288.667 ± 83.236. The large variation limits claims about the reliability of this difference. In the Stage I optimization rows, SFT alone reaches 350.133 ± 21.000, compared with 288.667 ± 83.236 for zero-shot generation. The full training-time retrieval configuration reaches 393.733 ± 13.973. These are results for the stated configurations, not independent estimates of each component's causal effect.

Table 5: Full pipeline ablation. Values are mean ± standard deviation across three seeds after aggregating the five folds within each seed. Stage II backbone and oracle rationale are expert-text references. ZS: zero-shot; FT: fine-tuned.

Variant	Bal-acc	F1-score	MCC	ACC	TAAPM
Two-stage (Full: Stage I + Dual-memory RAG + Stage II)					
Two-stage (ours)	0.700±0.019	0.608±0.022	0.382±0.038	0.700±0.019	393.733±13.973
(1) Necessity of two-stage: single-stage and Qwen3-VL Classification baselines					
Stage II backbone	0.704±0.001	0.611±0.002	0.386±0.003	0.699±0.002	409.067±0.377
Qwen3-VL-8B (ZS)	0.500±0.000	0.495±0.000	—	0.329±0.000	0.000±0.000
Qwen3-VL-8B (FT)	0.629±0.019	0.478±0.032	0.284±0.040	0.718±0.012	155.33±61.98
(2) Rationale contribution: expert text vs no text vs shuffled text (sanity checks)					
No rationale input	0.679±0.015	0.584±0.017	0.340±0.030	0.679±0.015	130.266±225.628
Shuffled rationale	0.689±0.020	0.596±0.023	0.360±0.040	0.690±0.021	260.266±225.585
Oracle expert rationale	0.704±0.001	0.611±0.002	0.386±0.003	0.699±0.002	409.067±0.377
(3) Text provision design: output format and dual-memory evidence source					
(3.1) Output format: free-form vs structured rationale					
free-form rationale	0.692±0.018	0.559±0.020	0.366±0.035	0.693±0.018	263.467±228.228
structured rationale	0.704±0.014	0.600±0.023	0.406±0.012	0.736±0.004	288.667±83.236
(3.2) Evidence source ablation: No-RAG vs knowledge-only vs case-only					
No-RAG	0.704±0.014	0.600±0.023	0.406±0.012	0.736±0.004	288.667±83.236
Knowledge-only RAG	0.700±0.008	0.602±0.010	0.388±0.016	0.719±0.009	338.933±5.021
Case-only RAG	0.701±0.018	0.603±0.024	0.389±0.033	0.717±0.012	345.733±37.324
Dual-memory RAG	0.700±0.019	0.608±0.022	0.382±0.038	0.700±0.019	393.733±13.973
(4) Stage I optimization: improving rationale generation quality					
zero-shot	0.704±0.014	0.600±0.023	0.406±0.012	0.736±0.004	288.667±83.236
SFT only	0.710±0.011	0.613±0.014	0.406±0.020	0.727±0.007	350.133±21.000
+ RAG at inference	0.695±0.017	0.601±0.020	0.370±0.034	0.696±0.017	263.333±228.108
+ RAG during training	0.700±0.019	0.608±0.022	0.382±0.038	0.700±0.019	393.733±13.973

7 Discussion, Limitations, and Release

AppleNMR-Text separates image-only inference, destructive ground truth, expert text, sample-level cross-validation, and cost-sensitive evaluation. The two-stage implementation attains the highest reported TAAPM among comparable automated image-input systems on this cohort. Its expert-text oracle and expert comparison have additional text inputs; no universal accuracy or cross-domain advantage is established.

The cohort spans one cultivar, season, storage protocol, operator, and acquisition day. Cross-site, cross-scanner, cross-season, and production-line performance are untested. Citation IDs do not prove rationale fidelity. Data completeness and quality control excluded 36 collected apples before training. The operational prior 0.1 differs from the cohort prevalence; TAAPM depends on the stated contract and threshold rule, not measured profit at other prevalences.

The data and corresponding labels are openly released with the supplementary materials. A GitHub repository is being prepared.

Reproducibility Statement

The data and corresponding labels accompany the supplementary materials. Additional configurations and predictions are planned for release.

AI Use Statement

Generative AI assisted with editing, organization, formatting, and critical assessment of the reported results and methods. It did not generate new experimental results for this revision. The authors remain responsible for the final content.

REFERENCES

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9268–9277, 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pp. 4171–4186. Association for Computational Linguistics, 2019. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *International Conference on Learning Representations (ICLR)*, 2020. URL https://openreview.net/forum?id=r1gRTCVFvB.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning (ICML)*, pp. 12888–12900. PMLR, 2022.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning (ICML)*, pp. 19730–19742. PMLR, 2023.

Zhi Liu, Dexiang Le, Tianyu Zhang, Qingrong Lai, Jiansheng Zhang, Bin Li, Yunfeng Song, and Nan Chen. Detection of apple moldy core disease by fusing vibration and Vis/NIR spectroscopy data with dual-input MLP-Transformer. *Journal of Food Engineering*, 382:112219, 2024. doi: 10.1016/j.jfoodeng.2024.112219.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://openreview.net/forum?id=37nvvqkCo5.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pp. 8748–8763. PMLR, 2021.

Jiawei Ren, Cunjun Yu, Shunan Sheng, Xiao Ma, Haiyu Zhao, Shuai Yi, and Hongsheng Li. Balanced meta-softmax for long-tailed visual recognition. *Advances in Neural Information Processing Systems*, 33:4175–4186, 2020.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2015. URL https://www.robots.ox.ac.uk/~vgg/publications/2015/Simonyan15/.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in Neural Information Processing Systems*, 30, 2017.

Zhongxiong Zhang, Haoling Liu, Zichao Wei, Miao Lu, Yuge Pu, Liulei Pan, Zuojing Zhang, Juan Zhao, and Jin Hu. A transfer learning method for spectral model of moldy apples from different origins. *Food Control*, 150:109731, 2023. doi: 10.1016/j.foodcont.2023.109731.

SUPPLEMENTARY MATERIAL

A AUDITABILITY AND LEAKAGE CONTROL

The proposed two stage framework is designed under two coupled requirements: deployment faithful image only inference and end to end auditability. Stage I constructs an Evidence Pack from a knowledge memory and a case memory, while Stage II predicts the final class from the image, the generated rationale, and the sanitized evidence. The purpose of the present section is to make the corresponding guarantees explicit, formally inspectable, and reproducible.

For a test sample x_i in fold k, the auditable inference chain is

$$E_i^{(k)} = R^{(k)}(x_i; M_K^{(k)}, M_C^{(k)}), \qquad r_i^{(k)} = G_\theta^{(k)}(x_i, E_i^{(k)}), \qquad \hat{y}_i^{(k)} = F_\phi^{(k)}(x_i, r_i^{(k)}, E_i^{(k)}). \tag{S1}$$

An output is considered *auditable* when every diagnostic claim in $r_i^{(k)}$ cites at least one evidence identifier from $E_i^{(k)}$, every cited identifier can be traced to a concrete retrieved source unit, and the final decision can be reconstructed from the logged image, sanitized Evidence Pack, validated rationale, and model checkpoint. An intermediate artifact is considered *label clean* when it contains no field or expression whose primary semantic function is to disclose the target label of any retrieved case.

The scope of the guarantee is intentionally operational. Predictive evidence remains available when it is expressed as image grounded observations, modality specific cues, or confounder descriptions. What is excluded are verdict style signals that reveal the class of a retrieved precedent directly or through a semantic equivalent.

A.1 THREAT MODEL FOR LABEL LEAKAGE

We consider two classes of leakage risk.

Direct leakage. Direct leakage occurs when a retrieved field explicitly stores the class of a case, or a one to one encoding of that class. Let $y_j \in \{0, 1\}$ be the ground truth label of retrieved case j, and let z_j denote a field appearing in its retrieved summary. Direct leakage is present if there exists a trivial decoder h such that

$$h(z_j) = y_j \tag{S2}$$

and z_j belongs to a field whose intended semantics is diagnosis, conclusion, verdict, status, result, or any equivalent administrative outcome.

Implicit leakage. Implicit leakage occurs when the label is recoverable from a textual span that does not use the canonical class token yet remains semantically equivalent to a final verdict. Examples include synonymous diagnostic conclusions, binary screening outcomes, accept or reject style decisions, or summary statements such as "consistent with disease" and "no evidence of disease". Formally, let $\mathcal{N}(\cdot)$ denote the canonical normalization used by the audit pipeline, including case folding, punctuation removal, whitespace normalization, and lexical alias mapping. Implicit leakage is present if

$$\mathcal{N}(u_j) \in \mathcal{V}_{\text{forbid}} \tag{S3}$$

for some text span u_j contained in a retrieved case summary, where $\mathcal{V}_{\text{forbid}}$ is the forbidden lexical ontology defined in Sec. A.4.

Leakage channels covered by the audit. The audit addresses four concrete channels:

1. retrieved case fields that explicitly store the class label;
2. verdict style or conclusion style sentences embedded inside otherwise descriptive case notes;
3. synonymous or administratively encoded label equivalents;
4. downstream reuse of any excluded case field inside the generator input, the classifier input, or the packaged Evidence Pack.

Leakage channels outside the intended exclusion set. The audit does not remove legitimate diagnostic content expressed as observations. For example, signal pattern descriptions, region level findings, confounder notes, and uncertainty statements remain valid because they support grounded reasoning without collapsing the retrieved precedent into a hidden gold answer. The target of the audit is therefore verdict disclosure, not observational evidence.

A.2 Fold Isolated Retrieval Protocol

Let the five fold protocol define mutually disjoint training and testing partitions

$$\mathcal{D}^{(k)}_{\text{train}} \cap \mathcal{D}^{(k)}_{\text{test}} = \varnothing, \qquad \mathcal{D}^{(k)}_{\text{train}} \cup \mathcal{D}^{(k)}_{\text{test}} = \mathcal{D}. \tag{S4}$$

For each fold k, all retrieval related components that depend on data are instantiated within the training partition of that fold. This rule applies to the case memory index, similarity normalization, reranking components, query side normalization modules, and any auxiliary models used by retrieval. No component is fit on, adapted to, or calibrated with held out fold labels.

In addition to fold isolation, training-time retrieval excludes every candidate with the same apple ID as the query. The original image, augmented versions, and any other images of that apple therefore cannot be retrieved as evidence for its own query.

The fold isolated protocol is defined as follows.

Case memory construction. The case memory for fold k is built only from training fold cases:

$$M^{(k)}_C = \left\{ \tilde{s}_j \ : \ (x_j, r_j, y_j) \in \mathcal{D}^{(k)}_{\text{train}} \right\}, \tag{S5}$$

where $r_j = (t_j, m_j, \pi_j)$ denotes the raw case record containing expert text t_j, optional normalized non label metadata m_j, and provenance only fields π_j, and

$$\tilde{s}_j = S(r_j) \tag{S6}$$

is the sanitized case summary obtained from the deterministic audit transform $S(\cdot)$ defined in Sec. A.5. The full raw case record may be retained for internal provenance, whereas only the sanitized projection $\tilde{s}_j$ is eligible to enter retrieval presentation and evidence packaging.

Knowledge memory construction. The knowledge memory M_K is version locked and independent of test fold labels. Its content is curated from external expert knowledge, modality specific criteria, and confounder descriptions. Because it is not derived from held out case labels, it does not constitute a case label leakage channel. Any version change of the knowledge memory is recorded by a memory version identifier.

Query construction constraint. For any test image $x_i \in \mathcal{D}^{(k)}_{\text{test}}$, the retrieval query is built from image derived content only:

$$q^{(k)}_i = Q^{(k)}(x_i, m^*_i), \tag{S7}$$

where m^*_i denotes normalized non label metadata when available. The query constructor is prohibited from accessing y_i, the human expert record t_i, or any test time annotation.

For clarity, the fold specific retrieval operator is decomposed into a knowledge branch and a case branch:

$$R^{(k)}_K(x_i) = \text{TopK}_K\left(q^{(k)}_i; M^{(k)}_K\right), \qquad R^{(k)}_C(x_i) = \text{TopK}_C\left(q^{(k)}_i; M^{(k)}_C\right), \tag{S8}$$

and the final Evidence Pack is formed by

$$E^{(k)}_i = \text{Pack}\Big(R^{(k)}_K(x_i),\, R^{(k)}_C(x_i)\Big). \tag{S9}$$

The two branches use $k_K = k_C = 5$: five knowledge items and five sanitized case items form the ten-item Evidence Pack. Here $R^{(k)}_C(x_i)$ denotes the set of sanitized retrieved case summaries selected from the case memory for query $q^{(k)}_i$, while $R^{(k)}_K(x_i)$ denotes the retrieved knowledge snippets from the knowledge memory.

Field	Type	Constraint and semantics	Primary role
`evidence_id`	string	Persistent identifier for citation and provenance tracing.	provenance, generator, validator
`case_uid`	string	Anonymized case identifier, unique within the benchmark release.	provenance only
`retrieval_score`	float	Similarity or aggregated rank score, optional, never interpreted as a label.	provenance, ranking trace
`observation_summary`	short text	Descriptive image grounded summary of visible or inferred LF NMR patterns. Verdict style wording is forbidden.	generator, classifier
`signal_pattern_tags`	list[string]	Controlled vocabulary for signal distribution, heterogeneity, attenuation, asymmetry, and related descriptive cues.	generator, classifier
`region_tags`	list[string]	Controlled vocabulary for spatial or slice level localization cues.	generator, classifier
`confounder_tags`	list[string]	Storage or acquisition related confounders, ambiguity factors, or physiological variation descriptors.	generator, classifier
`uncertainty_note`	short text	Optional ambiguity statement that explicitly avoids final decision wording.	generator, classifier
`metadata_tags`	list[string]	Optional normalized non label metadata used for retrieval disambiguation.	retrieval, generator
`source_pointer`	string	Pointer to the internal full record or memory entry for audit reconstruction.	provenance only

Table S1: Whitelisted case summary template exposed to retrieval grounded rationale generation. Any field outside this schema is excluded from the presented case summary. Fields whose semantics correspond to diagnosis, verdict, result, class, or outcome are forbidden.

Retrieval boundary constraint. Retrieved precedents for a held out sample must satisfy

$$\mathrm{src}(e) \in \mathcal{D}_{\mathrm{train}}^{(k)} \quad \forall e \in R_C^{(k)}(x_i), \tag{S10}$$

and every retrieved case summary must pass the leakage audit before being packed into $E_i^{(k)}$.

Stage II input constraint. The classifier receives only the image, the validated rationale, and the sanitized evidence text. Fields excluded by the audit, especially label revealing case fields, are never serialized into the Stage II input tensor or text stream.

Together, these rules guarantee that retrieval for fold k remains train fold bounded, label clean at presentation time, and reproducible under the same split.

A.3 Case Summary Template for Retrieved Cases

To preserve precedent level grounding while avoiding verdict disclosure, each retrieved case is exposed through a sanitized summary template whose fields are observation centered and provenance aware. The template is intentionally narrower than the full internal case record. Only the whitelisted projection may appear in the Evidence Pack.

Table S1 defines the formal retrieved case summary schema.

The corresponding sanitized case projection can be written as

$$\begin{aligned} \tilde{s}_j &= \Pi_{\mathrm{allow}}(r_j), \\ \mathrm{Fields}(\tilde{s}_j) &= \{\texttt{evidence_id}, \texttt{case_uid}, \texttt{retrieval_score}, \texttt{observation_summary}, \\ &\qquad \texttt{signal_pattern_tags}, \texttt{region_tags}, \texttt{confounder_tags}, \\ &\qquad \texttt{uncertainty_note}, \texttt{metadata_tags}, \texttt{source_pointer}\}. \end{aligned} \tag{S11}$$

where r_j is the raw stored case record and Π_{allow} denotes field level whitelist projection.

For clarity, the whitelisted schema serves two distinct purposes. First, it preserves sufficient descriptive information for precedent grounded reasoning. Second, it sharply separates *descriptive evidence* from *decision fields*. This separation is essential because Stage I is allowed to reason from comparable precedents, whereas it is not allowed to consume a retrieved gold verdict.

A.4 Forbidden Label Equivalent Fields and Synonymous Expressions

Leakage prevention is implemented as a strict exclusion specification. Two complementary sets are maintained: a forbidden field set and a forbidden lexical ontology.

Forbidden field set. The following fields are excluded from any retrieved case presentation, Evidence Pack serialization, or classifier input:

$$\begin{aligned}\mathcal{F}_{\text{forbid}} = \{&\texttt{label}, \texttt{class}, \texttt{diagnosis}, \texttt{diagnostic_conclusion},\\ &\texttt{final_assessment}, \texttt{verdict}, \texttt{result}, \texttt{outcome},\\ &\texttt{status}, \texttt{disease_status}, \texttt{abnormality_flag},\\ &\texttt{screening_decision}, \texttt{accept_reject}, \texttt{pass_fail}\}.\end{aligned} \tag{S12}$$

Any case field whose semantics is equivalent to a final diagnostic or administrative decision is mapped into $\mathcal{F}_{\text{forbid}}$ and removed.

Forbidden lexical ontology. At the text level, all retrieved summaries are screened against a version controlled lexical ontology $\mathcal{V}_{\text{forbid}}$, which includes the canonical class terms, their inflectional and synonymous forms, and verdict style semantic equivalents. For the present binary task, $\mathcal{V}_{\text{forbid}}$ contains at least the following classes of expressions:

1. explicit class names, for example terms denoting moldy core, healthy fruit, diseased fruit, infected fruit, normal fruit, or abnormal fruit;
2. binary outcome expressions, for example positive, negative, pass, fail, accept, reject, qualified, unqualified;
3. verdict phrases, for example consistent with disease, diagnostic of disease, confirmed disease, ruled out disease, no evidence of disease;
4. summary conclusion prefixes, for example diagnosis:, conclusion:, verdict:, result:, final assessment:, screening decision:;
5. administrative action phrases whose semantics implies a class decision, for example remove from shipment, retain for sale, reject batch item.

Template level exclusions. In addition to isolated tokens, the audit removes whole sentence templates whose discourse function is a final decision. The template library $\mathcal{T}_{\text{forbid}}$ includes regular expressions of the form

$$\texttt{\^{}(diagnosis|conclusion|verdict|result|status)\textbackslash s*[::]} \tag{S13}$$

and

$$\begin{aligned}&\texttt{(consistent with|diagnostic of|confirmed|}\\ &\qquad\texttt{ruled out|no evidence of).}\end{aligned} \tag{S14}$$

A sentence matching $\mathcal{T}_{\text{forbid}}$ is removed in its entirety rather than partially edited, because partial editing can leave residual label equivalence.

Priority rule. When a field or sentence is ambiguous between descriptive content and verdict content, the audit resolves the ambiguity conservatively and excludes it from the retrieved case summary. This rule preserves label cleanliness under borderline phrasing.

A.5 Automatic Audit Rules for Leakage Prevention

The case sanitization and Evidence Pack audit are fully automatic and deterministic once the whitelist, forbidden field set, forbidden lexical ontology, and verdict template library are fixed. The audit procedure is applied before a retrieved case is allowed to enter the generator context or the classifier input.

For a raw case record r_j, the sanitization map is

$$S(r_j) = V_{\text{post}}\Big(T_{\text{strip}}\big(L_{\text{screen}}(\Pi_{\text{allow}}(r_j))\big)\Big). \tag{S15}$$

where Π_{allow} is whitelist projection, L_{screen} is lexical screening, T_{strip} removes verdict style sentences, and V_{post} is the final post processing validator.

The complete audit rules are listed below.

Rule 1: field whitelist projection. Only fields defined in Table S1 are retained. Every field name in $\mathcal{F}_{\text{forbid}}$ is removed before any text processing.

Rule 2: lexical normalization. All retained string fields are normalized using lowercasing, punctuation folding, whitespace collapsing, and alias canonicalization. The normalized text is the input to all subsequent lexical checks.

Rule 3: forbidden ontology screening. For every textual span u in the retained case summary, the audit computes

$$\mathbf{1}_{\text{lex}}(u) = \begin{cases} 1, & \text{Lex}(\mathcal{N}(u)) \cap \mathcal{V}_{\text{forbid}} \neq \varnothing, \\ 0, & \text{otherwise.} \end{cases} \tag{S16}$$

where $\text{Lex}(\mathcal{N}(u))$ denotes the set of normalized tokens and matched n-gram phrases extracted from span u. If $\mathbf{1}_{\text{lex}}(u) = 1$, the offending sentence is removed. If the offending content is attached to a field whose semantics becomes unclear after removal, the entire field is removed.

Rule 4: verdict template stripping. Each sentence is matched against the forbidden template library $\mathcal{T}_{\text{forbid}}$. A matched sentence is deleted in full. This step is applied even when the sentence no longer contains an exact lexical hit after normalization, because template level leakage often survives paraphrase.

Rule 5: empty field cleanup. After Rules 1 to 4, empty strings, empty lists, and structurally incomplete records are removed. If removal causes a case summary to lose all descriptive content, the retrieved case is discarded from presentation and the retriever falls back to the next eligible candidate.

Rule 6: Evidence Pack admission test. A sanitized case summary $\tilde{s}_j$ is admitted into the Evidence Pack only if

$$\begin{aligned} \text{AuditPass}(\tilde{s}_j) = 1 \iff & \Big(\text{Fields}(\tilde{s}_j) \subseteq \mathcal{F}_{\text{allow}}\Big) \\ & \wedge \Big(\text{LexHit}(\tilde{s}_j) = 0\Big) \wedge \Big(\text{TemplateHit}(\tilde{s}_j) = 0\Big). \end{aligned} \tag{S17}$$

Otherwise, the case is excluded and replaced by the next ranked candidate that passes the audit.

Rule 7: rationale citation validation. After Stage I generation, every cited evidence identifier must belong to the retrieved evidence set:

$$\text{Cite}(r_i^{(k)}) \subseteq \text{IDs}(E_i^{(k)}). \tag{S18}$$

Outputs containing out of set citations, malformed identifiers, or missing required evidence references are rejected.

Rule 8: Stage II packaging constraint. The classifier input is built only from the validated rationale, the image, and the sanitized evidence text. Provenance only fields such as internal source pointers may be retained in the audit log, yet they are not serialized into the model input stream.

Rule 9: log based reproducibility. For each test sample, the system stores the retrieval results before and after sanitization, the final Evidence Pack, the rationale validation outcome, and the final prediction record. These logs permit retrospective inspection of where evidence entered, how it was filtered, and which identifiers supported the final decision.

B Formal Definitions of Stage I Interfaces

The main paper formulates the proposed framework as a two stage decomposition in which evidence construction is explicitly separated from downstream decision making. In Stage I, the model retrieves a dual memory evidence set and generates a structured rationale grounded in that retrieved evidence. In Stage II, the final prediction is conditioned on the image, the generated rationale, and the retrieved evidence while preserving the image only deployment assumption for external inputs. To make this interface fully reproducible and auditable, this section formalizes the Stage I intermediate objects, including the Evidence Pack, the structured generator output, the claim to evidence linking constraints, and the compact rationale projection used by Stage II.

B.1 Formal Definition of the Evidence Pack

Let $x_i \in \mathcal{X}$ denote the LF-NMR mosaic of sample i. Let M_K denote the knowledge memory and M_C denote the case memory. As in the main paper, Stage I first retrieves an evidence set

$$E_i = \mathcal{R}(x_i; M_K, M_C) = \{e_{i,m}\}_{m=1}^{K_i}, \tag{S19}$$

where K_i is the number of retained evidence items for sample i.

We define the *Evidence Pack* for sample i as the ordered collection

$$\mathcal{P}_i = \big(e_{i,1}, e_{i,2}, \ldots, e_{i,K_i}\big), \tag{S20}$$

where each evidence item $e_{i,m}$ is a typed record

$$e_{i,m} = \Big(\mathsf{id}_{i,m}, \mathsf{src}_{i,m}, \mathsf{text}_{i,m}, \mathsf{prov}_{i,m}, \mathsf{meta}_{i,m}\Big). \tag{S21}$$

Each component is defined as follows.

Persistent evidence identifier. $\mathsf{id}_{i,m} \in \Sigma^*$ is a persistent evidence identifier represented as a string. It is unique within the active memory snapshot used in the current fold and remains stable across retrieval, prompting, output parsing, and post generation validation. This identifier is the only valid citation handle that the generator may use when linking claims to evidence.

Source type. $\mathsf{src}_{i,m}$ specifies the memory origin of the evidence item:

$$\mathsf{src}_{i,m} \in \{\texttt{knowledge}, \texttt{case}\}. \tag{S22}$$

Here, `knowledge` denotes a retrieved knowledge snippet from M_K, and `case` denotes a sanitized case summary from M_C.

Content field. $\mathsf{text}_{i,m} \in \Sigma^*$ is the textual content exposed to the generator. Its semantics depend on the source type.

For $\mathsf{src}_{i,m} = \texttt{knowledge}$, $\mathsf{text}_{i,m}$ is a chunked knowledge snippet describing modality specific imaging cues, confounders, interpretation criteria, or related domain knowledge.

For $\mathsf{src}_{i,m} = \texttt{case}$, $\mathsf{text}_{i,m}$ is a sanitized summary derived from a historical case record rather than the raw record itself. Let the original case tuple be (x_j, y_j, t_j). The exposed case text is

$$\bar{t}_j = \mathcal{S}(t_j), \tag{S23}$$

where $\mathcal{S}(\cdot)$ is a deterministic sanitization operator that removes any field or expression that can reveal the final class directly or near deterministically. This includes explicit labels, direct diagnostic verdicts, synonymous label statements, and summary conclusions that collapse the case into the target label. The purpose of $\mathcal{S}$ is to preserve precedent level descriptive evidence while preventing label leakage.

Field name	Data type	Req.	Semantic meaning
`image_findings`	list[`Finding`]	Yes	Image conditioned descriptive findings extracted from the LF-NMR input. Each finding should describe observable patterns rather than the final class label.
`supported_claims`	list[`Claim`]	Yes	Diagnostic or interpretive claims supported by retrieved evidence. Every claim must cite at least one valid evidence identifier.
`uncertainties`	list[`Uncertainty`]	No	Explicit statements about ambiguity, insufficient evidence, confounding factors, or quality limitations relevant to the current sample.
`used_evidence_ids`	list[string]	Yes	Deduplicated union of all evidence identifiers cited anywhere in the structured output. This field is useful for audit and validation.
`confidence`	real in $[0, 1]$ or ordinal string	Yes	Confidence associated with the generated rationale under the retrieved evidence context. It is a self assessment field rather than a final classification probability.
`final_rationale`	string	Yes	Coherent natural language rationale summarizing findings, evidence grounded interpretation, and uncertainty in a compact human readable form.

Table S2: Formal top level schema of the Stage I structured output.

Provenance metadata. $\mathsf{prov}_{i,m}$ stores the provenance record needed for audit and traceability. It is represented as a structured tuple

$$\mathsf{prov}_{i,m} = \big(\mathsf{memory_name}, \mathsf{record_id}, \mathsf{chunk_id}, \mathsf{fold_scope}, \mathsf{memory_version}, \mathsf{retrieval_rank}\big). \tag{S24}$$

For knowledge memory items, record_id refers to the originating knowledge document and chunk_id identifies the retrieval unit. For case memory items, record_id refers to the historical sample identifier and chunk_id may be null when the case summary is treated as one atomic unit. fold_scope records the fold specific isolation boundary under which the item is valid.

Optional metadata. $\mathsf{meta}_{i,m}$ contains optional retrieval side information such as similarity score, dense rank, sparse rank, reranking score, diversity flag, confidence surrogate, or source specific tags. These fields are allowed for reproducibility and audit, yet they are auxiliary and are not required for the logical validity of claim citations.

Accordingly, the Evidence Pack is an interface object with two simultaneous roles. First, it is the full set of textual evidence made available to the Stage I generator. Second, it is an auditable record that binds every retrieved text fragment to a persistent identifier and explicit provenance metadata.

For later reference, define the valid identifier set induced by the retrieved pack as

$$\mathrm{ID}(E_i) = \{\mathsf{id}_{i,m} \mid e_{i,m} \in E_i\}. \tag{S25}$$

B.2 Formal Definition of the Structured Output Schema

Given the image x_i and the Evidence Pack $\mathcal{P}_i$, the Stage I generator produces a structured output

$$o_i = \mathcal{G}_\theta(x_i, \mathcal{P}_i), \tag{S26}$$

where o_i is a schema constrained object. We formalize it as

$$o_i = \Big(\mathcal{F}_i, \mathcal{C}_i, \mathcal{U}_i, \mathcal{I}_i, q_i, r_i\Big), \tag{S27}$$

with the following top level fields.

The nested record types are defined as follows.

Finding record. Each element $f \in \mathcal{F}_i$ is a tuple

$$f = \big(\texttt{finding_text}, \texttt{evidence_ids}\big), \tag{S28}$$

where `finding_text` is a short textual description of an observed image pattern and `evidence_ids` is a list of evidence identifiers that ground this observation or its interpretation.

Claim record. Each element $c \in \mathcal{C}_i$ is a tuple

$$c = \big(\texttt{claim_text}, \texttt{claim_type}, \texttt{evidence_ids}\big), \tag{S29}$$

where `claim_text` is a proposition derived from the image and retrieved evidence, `claim_type` is an application level tag such as `supportive`, `cautionary`, or `differential`, and `evidence_ids` is a nonempty list of supporting identifiers.

Uncertainty record. Each element $u \in \mathcal{U}_i$ is a tuple

$$u = \big(\texttt{uncertainty_text}, \texttt{uncertainty_type}, \texttt{evidence_ids}\big), \tag{S30}$$

where `uncertainty_text` states the source of ambiguity, `uncertainty_type` may encode categories such as image quality, confounder overlap, or retrieval coverage, and `evidence_ids` is optional when the uncertainty statement is linked to explicit retrieved evidence.

Evidence union field. The field $\mathcal{I}_i = \texttt{used_evidence_ids}$ is defined deterministically as

$$\mathcal{I}_i = \mathrm{Unique}\left(\bigcup_{f \in \mathcal{F}_i} \Gamma(f) \cup \bigcup_{c \in \mathcal{C}_i} \Gamma(c) \cup \bigcup_{u \in \mathcal{U}_i} \Gamma(u) \right), \tag{S31}$$

where $\Gamma(\cdot)$ extracts the cited identifier list from a record and $\mathrm{Unique}(\cdot)$ preserves the first occurrence order while removing duplicates.

Confidence field. q_i is the rationale confidence field. It may be implemented either as a scalar in $[0, 1]$ or as an ordinal category such as `low`, `medium`, and `high`. Since the classifier in Stage II predicts the final label independently, this field is interpreted as confidence in the explanatory rationale rather than as the deployed decision score.

Final rationale. $r_i \in \Sigma^*$ is a human readable rationale that compresses the most important information from $\mathcal{F}_i, \mathcal{C}_i, \mathcal{U}_i$, and q_i into one coherent explanatory text span.

B.3 Claim to Evidence Linking Constraints

The main paper requires every diagnostic claim to cite retrieved evidence identifiers. We formalize this requirement as a hard interface constraint.

For every claim $c \in \mathcal{C}_i$, let $\Gamma(c)$ denote the list of cited identifiers. The claim is *citation valid* with respect to the retrieved Evidence Pack E_i if and only if

$$\Gamma(c) \subseteq \mathrm{ID}(E_i) \quad \text{and} \quad |\Gamma(c)| \geq 1. \tag{S32}$$

This yields the following three validity conditions.

Identifier membership. Every cited identifier must belong to the retrieved identifier set $\mathrm{ID}(E_i)$. Any citation to an identifier outside the retrieved pack is invalid.

Nonempty support. Every claim must retain at least one valid supporting identifier after validation. A claim with zero valid citations is considered unsupported and cannot be preserved as a supported claim.

Schema consistency. The cited identifiers listed inside each claim, finding, or uncertainty item must be consistent with the top level field `used_evidence_ids`. In particular,

$$\bigcup_{c \in \mathcal{C}_i} \Gamma(c) \subseteq \mathcal{I}_i \subseteq \mathrm{ID}(E_i). \tag{S33}$$

We define the deterministic post generation validator as

$$(\hat{o}_i, \delta_i) = \mathcal{V}(o_i, E_i), \tag{S34}$$

where $\hat{o}_i$ is the validated structured output and δ_i is an audit log recording any removed citation or rejected field.

The validator $\mathcal{V}$ performs the following checks.

Syntactic checks. It verifies that all required top level fields are present, that field types match the declared schema, and that every nested record is well formed.

Citation checks. For every cited identifier, it verifies membership in $\mathrm{ID}(E_i)$. Any out of set citation is deleted from the record. This formalizes the main paper statement that invalid citations are removed during validation.

Support checks. After citation cleanup, any claim with an empty support set is rejected from the set of supported claims. The same conservative rule may be applied to findings when a grounded finding loses all supporting identifiers.

Output rejection conditions. The validated output is rejected and regenerated if any of the following conditions hold:

$$\begin{aligned} &\text{a required top level field is missing,} \\ &\mathcal{C}_i = \varnothing \text{ after validation,} \\ &\texttt{final_rationale} \text{ is empty or malformed,} \\ &\mathcal{I}_i \nsubseteq \mathrm{ID}(E_i). \end{aligned} \tag{S35}$$

The scope of $\mathcal{V}$ is intentionally deterministic and interface oriented. It guarantees schema well formedness, identifier validity, and support nonemptiness. It does not constitute a complete semantic verifier of scientific truth. Semantic faithfulness is encouraged upstream by retrieval grounding and prompt constraints, while formal post generation validation focuses on inspectable structural guarantees.

B.4 Compact Rationale Construction for Stage II

The full structured output is richer than what is necessary for Stage II classification. We therefore define a projection from the audit oriented Stage I output to a compact textual rationale consumed by Stage II.

Let the validated structured output be $\hat{o}_i$. The compact rationale is defined as

$$\tilde{r}_i = \Pi(\hat{o}_i), \tag{S36}$$

where $\Pi(\cdot)$ is a deterministic linearization and compression operator.

Specifically, $\Pi(\cdot)$ retains the semantic contents of the following fields:

$$\begin{aligned} &\texttt{image_findings}, \ \texttt{supported_claims}, \ \texttt{uncertainties}, \\ &\texttt{confidence}, \ \texttt{final_rationale}. \end{aligned} \tag{S37}$$

In contrast, the following fields are treated as audit only metadata and are excluded from the classifier text input:

$$\texttt{used_evidence_ids}, \quad \mathsf{prov}_{i,m}, \quad \mathsf{meta}_{i,m}, \quad \mathsf{src}_{i,m}, \tag{S38}$$

together with raw record identifiers, chunk identifiers, retrieval ranks, memory versions, and any sanitization flags. These fields remain available for auditing and reproducibility, yet they are not injected into the Stage II textual stream.

The projection $\Pi(\cdot)$ is implemented in four steps.

Step 1: field selection. Keep only validated findings, validated supported claims, uncertainty notes, confidence, and the final rationale text.

Step 2: canonical ordering. Arrange the retained contents in a fixed order:

$$\texttt{Findings} \rightarrow \texttt{Claims} \rightarrow \texttt{Uncertainty} \rightarrow \texttt{Confidence} \rightarrow \texttt{Summary}. \tag{S39}$$

This deterministic ordering removes formatting variance across samples.

Step 3: textual compression. Apply sentence level deduplication, remove repeated evidence phrases, and normalize wording into short declarative spans. Let $L_{\max}$ denote the maximum rationale length used by Stage II. The final text is truncated or compressed so that

$$|\tilde{r}_i| \leq L_{\max}. \tag{S40}$$

Step 4: audit separation. Store the full Evidence Pack $\mathcal{P}_i$ and the validated structured output $\hat{o}_i$ as the audit trace, while only the compressed text $\tilde{r}_i$ is passed as the rationale channel to Stage II.

Accordingly, the Stage II inputs can be written as

$$\big(x_i,\, \tilde{r}_i,\, \{\mathsf{text}_{i,m}\}_{m=1}^{K_i}\big), \tag{S41}$$

where the evidence pathway uses evidence text only, and any label revealing fields from retrieved cases remain excluded by construction.

This compact projection serves two purposes. It preserves the explanatory signal needed for multimodal classification, and it keeps audit metadata separate from the classifier facing text interface. As a result, the system maintains claim level traceability at the Stage I audit layer while exposing to Stage II only the semantically relevant rationale and evidence text.

C Prompt Templates and Generation Protocol

This section provides the complete prompt specification used in Stage I for retrieval grounded rationale generation. The prompt design follows four principles required by the main method: evidence grounded generation from a dual memory Evidence Pack, persistent evidence identifier preservation, structured output compliance, and explicit uncertainty reporting. To reduce implementation ambiguity, we report both the system level instruction and the user side formatting template used to inject the retrieved evidence into the generation context.

C.1 Stage I System Prompt

The Stage I generator is instructed by a fixed system prompt shared across all folds. The purpose of this prompt is to constrain the model toward auditable rationale generation rather than free narrative description. In particular, the model is required to ground each diagnostic claim in one or more evidence identifiers drawn from the retrieved Evidence Pack, avoid unsupported assertions, report uncertainty explicitly when the evidence is insufficient or partially conflicting, and return a schema compliant structured output that can be parsed deterministically.

A reproducible version of the system prompt is given below.

> **System Prompt**
>
> You are an expert assistant for apple moldy core diagnosis from LF-NMR mosaic images. Your task is to generate a structured diagnostic rationale from the input image and the retrieved Evidence Pack.
>
> You must obey the following rules.
>
> (1) Use only the provided image and the provided Evidence Pack. Do not invent external facts, hidden labels, or unsupported conclusions.
>
> (2) Every diagnostic claim must be supported by at least one valid evidence identifier copied exactly from the Evidence Pack.
>
> (3) Evidence identifiers are persistent provenance tags. Never alter, paraphrase, merge, or fabricate an identifier.

(4) Distinguish clearly between observed image findings, evidence supported claims, and uncertainty statements.

(5) When the available evidence is weak, incomplete, or conflicting, state the uncertainty explicitly. Do not convert uncertainty into false certainty.

(6) Do not reveal any training label, gold diagnosis, or hidden metadata. Retrieved case items are precedents and can support reasoning, yet they must not be treated as direct ground truth for the current sample.

(7) Return the answer in the required structured format only. The output must be machine parsable and must contain all required fields.

(8) The final rationale must be concise, evidence grounded, and suitable for downstream multimodal classification.

Required output fields:

1. `image_findings`: a list of direct image observations stated in clinical or diagnostic language.
2. `supported_claims`: a list of claim objects. Each object must contain the claim text and the supporting evidence identifiers.
3. `uncertainties`: a list of uncertainty statements, each linked to the relevant evidence identifiers when applicable.
4. `confidence`: one of `low`, `medium`, or `high`.
5. `final_rationale`: a compact synthesis grounded in the preceding fields.

Output constraints:

1. Every element in `supported_claims` must cite at least one evidence identifier.
2. Every cited identifier must belong to the current Evidence Pack.
3. The `final_rationale` must summarize only supported findings and acknowledged uncertainty.
4. Do not output explanations outside the schema.

In our implementation, the system prompt remains fixed across training and inference so that rationale generation is conditioned by the same structural constraints in both phases. This choice reduces train inference mismatch at the prompt level and makes the generated rationale a stable interface for Stage II.

C.2 User Prompt Template with Evidence Pack Formatting

The user prompt injects the current LF-NMR sample together with the retrieved Evidence Pack. The formatting is designed to preserve source separation between knowledge memory and case memory, while keeping evidence identifiers visible in plain text so that they can be copied exactly into the generated output.

The braced terms below are runtime variables populated from the current sample and retrieval output; they are not missing manuscript text. The knowledge block is repeated for five retrieved knowledge items and the case block for five retrieved cases. Ellipses indicate this repetition. The JSON block specifies the output format: its strings are schematic fields, not observed model outputs or a held-out example.

A reproducible template is given below.

User Prompt Template

Sample identifier: {`sample_id`}
Task: Generate a structured diagnostic rationale for the current LF-NMR mosaic.
Use only the evidence listed below.

Image Context

LF-NMR mosaic image of one apple specimen is provided as the visual input.

Metadata Tags

{`normalized_metadata_tags`}

Evidence Pack Begins

Section A. Knowledge Memory Evidence
[K_01] Source: knowledge
Title: {knowledge_title_1}
Text: {knowledge_text_1}

[K_02] Source: knowledge
Title: {knowledge_title_2}
Text: {knowledge_text_2}

...
Section B. Case Memory Evidence
[C_01] Source: case
Case summary: {case_summary_1}
Observation notes: {case_text_1}

[C_02] Source: case
Case summary: {case_summary_2}
Observation notes: {case_text_2}

...
Evidence Pack Ends
Please return a JSON object with the following schema:

```
{
  "image_findings": [
    "..."
  ],
  "supported_claims": [
    {
      "claim_text": "...",
      "evidence_ids": ["K_01", "C_02"]
    }
  ],
  "uncertainties": [
    {
      "text": "...",
      "evidence_ids": ["K_02"]
    }
  ],
  "confidence": "low | medium | high",
  "final_rationale": "..."
}
```

Important requirements:

1. Copy evidence identifiers exactly.
2. Use only identifiers appearing in this Evidence Pack.
3. Do not output any extra keys.
4. Do not output markdown.
5. Do not output explanatory text before or after the JSON object.

Several implementation details are important for reproducibility. First, evidence items from the two memories are grouped by source to preserve provenance semantics. Second, case memory entries expose textual summaries and observations only, while label fields are hidden before insertion into the prompt. Third, the model receives evidence identifiers in plain text rather than latent tags so that post generation citation checking can be implemented by exact string matching. Fourth, the requested output schema is presented explicitly within the prompt to reduce format drift and improve deterministic parsing.

C.3 Validation and Rejection Rules

Post generation validation is applied to every Stage I output before it is accepted for downstream use. The validator operates as a deterministic filter with three possible outcomes: accepted, corrected, or discarded.

Accepted outputs. An output is accepted directly when all of the following conditions hold:

1. the response is valid JSON and can be parsed without ambiguity
2. all required keys are present, namely `image_findings`, `supported_claims`, `uncertainties`, `confidence`, and `final_rationale`
3. `confidence` belongs to the closed set {`low`, `medium`, `high`}
4. every supported claim contains at least one evidence identifier
5. every cited identifier belongs to the retrieved Evidence Pack of the current sample
6. the final rationale is non empty

Corrected outputs. An output is corrected by a deterministic repair step when the deviation is purely syntactic and does not change semantic content. The following corrections are permitted:

1. whitespace normalization and Unicode normalization
2. removal of duplicated evidence identifiers within the same field
3. coercion of singleton strings to one element lists where the schema expects a list
4. normalization of confidence tokens such as `Low` or `HIGH` to the canonical lowercase form
5. insertion of an empty list for optional list fields when the field is present conceptually yet omitted syntactically

No semantic repair is applied to unsupported claims. In particular, the validator never fabricates new evidence identifiers, never rewrites claim content to force support, and never infers hidden fields.

Discarded outputs. An output is discarded when any of the following conditions holds:

1. the response cannot be parsed into valid JSON after one deterministic repair pass
2. one or more required fields remain missing
3. any supported claim has no evidence identifier
4. any cited evidence identifier is outside the retrieved Evidence Pack
5. the output contains free text outside the structured object that prevents deterministic parsing
6. the final rationale is empty, vacuous, or reduced to a placeholder string

Unsupported claim handling. A claim is treated as unsupported when it asserts a diagnostic property or modality specific interpretation yet provides no valid evidence identifier. Outputs containing such claims are rejected rather than partially trusted. This strict rule is important because Stage I is designed to produce auditable rationales with explicit provenance, and partial acceptance of unsupported claims would weaken the traceability objective.

Out of set citation handling. All cited identifiers are matched exactly against the identifier set of the current Evidence Pack. Any identifier that is missing, truncated, reformatted, or copied from another sample is considered invalid. Outputs with invalid identifiers are rejected. In the implementation, this check is performed before rationale compression so that only validated rationale content is passed to Stage II.

Schema stabilization for downstream use. After acceptance, the structured rationale is serialized into a canonical field order and compressed into a compact rationale string for Stage II. This serialization step is deterministic and does not add new semantic content. Its purpose is to ensure that downstream text encoding is driven by validated fields only.

D Human Expert and Generated Text Review Protocol

This section describes expert text curation, the image-and-text expert comparison, and the generated-rationale review protocol. The expert comparison uses images and text without ground-truth labels. This differs from the automated pipeline, whose external test input is only the LF-NMR image.

D.1 Human Expert Evaluation Setup

A panel of three domain experts with experience in apple internal defect assessment and LF-NMR interpretation participated in the image-and-text expert comparison. For each test case, experts were provided with the standardized 2×2 LF-NMR mosaic and accompanying text. Ground-truth labels were not shown. This is an image-and-text expert comparison. The accompanying text was derived from existing sample-associated descriptions; it was not Stage-I-generated reasoning and contained no ground-truth diagnostic labels. The experts in this comparison were different from the experts responsible for text annotation.

The expert comparison was evaluated under the same five fold protocol used for the automated systems. In each fold, experts reviewed only the held out test apples, ensuring that human judgments and model predictions were compared on the identical case set. For each apple, each expert completed a structured review form containing three fields: (i) a binary decision, *healthy* or *moldy core*; (ii) an optional confidence rating on a three level scale, *low*, *medium*, or *high*; and (iii) a concise textual justification describing the most salient LF-NMR findings supporting the decision.

For quantitative comparison with model predictions, the final human panel decision was aggregated at the case level by majority vote across the three experts. Performance was then summarized with the same metrics used for automated systems, namely balanced accuracy, F1 score, Matthews correlation coefficient, accuracy, and TAAPM. The human expert results therefore represent an image-and-text expert comparison with ground-truth labels withheld. Sharing the test cases and metrics does not make its input condition identical to the image-input automated pipeline.

D.2 Expert Adjudication for Dataset Text Records

During dataset construction, each apple was paired with an expert verified diagnostic record anchored to the corresponding LF-NMR mosaic and the destructive verification outcome. Text curation followed a controlled adjudication procedure intended to standardize semantics, improve reproducibility, and preserve deployment relevance.

First, a primary expert drafted or revised the diagnostic record using a controlled vocabulary for lesion presence, image findings, confounding patterns, and uncertainty expressions. Second, synonymous or locally idiosyncratic wording was normalized so that semantically equivalent descriptions were expressed consistently across records. Third, ambiguous statements were flagged whenever they contained unsupported causal language, unresolved referents, internally conflicting clauses, or wording that could not be mapped reliably to observable LF-NMR evidence. Fourth, flagged cases were sent to adjudication.

Adjudication was performed by a second expert and, when required, by a senior third expert. The purpose of adjudication was to standardize terminology and interpretation rather than to redefine the ground truth label. The final accepted text record had to satisfy four requirements: (i) its terminology matched the controlled vocabulary; (ii) each diagnostic statement could be traced to observable image findings or established domain knowledge; (iii) uncertainty language was explicit whenever the evidence was weak, mixed, or potentially confounded; and (iv) the final wording contained no extraneous information outside the intended diagnostic record schema.

When disagreement concerned wording only, the more specific and image grounded formulation was retained. When disagreement concerned interpretation, the adjudicator selected the formulation most consistent with the LF-NMR image and the verified outcome used for dataset construction. This procedure ensured that the final AppleNMR Text records were semantically standardized, auditable, and suitable for both supervision and retrieval grounded generation.

D.3 Generated Text Review Criteria

Generated rationales were reviewed with a case indexed rubric designed to match the structure of Stage I outputs. Each review was performed jointly over the input LF-NMR image, the retrieved Evidence Pack, and the generated structured rationale. Five review dimensions were used.

First, **factual support**. Every diagnostic claim had to be supported either by directly observable LF-NMR evidence or by retrieved evidence items relevant to the case. Unsupported lesion descriptions, unsupported causal mechanisms, and unsupported generalizations were marked as errors.

Second, **citation correctness**. Each cited evidence identifier had to exist in the retrieved Evidence Pack and had to support the claim for which it was invoked. Out of set citations, citation reuse with mismatched semantics, or unsupported claim to citation links were marked as critical errors.

Third, **terminology appropriateness**. The rationale had to use normalized domain terminology consistent with the curated dataset style. Vague colloquial wording, inconsistent naming of the same finding, or expressions that conflated disease cues with confounders were marked down.

Fourth, **uncertainty calibration**. Confidence statements had to match the strength of available evidence. Overconfident language in equivocal cases and unjustified hesitation in visually clear cases were both treated as calibration errors.

Fifth, **image consistency**. The narrative had to remain consistent with the observed LF-NMR appearance of the input mosaic, including lesion related signal changes, spatial patterning, and plausible confounding factors. Statements contradicted by the visible image evidence were treated as major errors.

Each dimension was rated as *acceptable*, *borderline*, or *unacceptable*. A generated rationale was accepted only if factual support and citation correctness were both rated acceptable and no image consistency violation was present. Borderline cases on terminology or uncertainty calibration were not automatically rejected and were instead forwarded to adjudication. In addition to the categorical judgment, reviewers could record a short failure tag such as *unsupported claim*, *citation mismatch*, *terminology drift*, or *miscalibrated certainty* to facilitate error analysis.

D.4 Conflict Resolution Rules

Disagreement handling followed a pre specified hierarchy. For the expert comparison diagnosis task, the final case decision was obtained by majority vote across the three experts. Since the panel size was odd, no tie occurred at the final label level. When the majority label differed from a reviewer who expressed low confidence or raised a more detailed qualitative concern, the majority decision was retained for metric computation, while the dissenting note was preserved in the audit log.

For dataset text curation and generated rationale review, a two stage conflict resolution procedure was used. If the first two reviewers agreed, their decision was accepted directly. If they disagreed, the case was escalated to the third expert for adjudication. The adjudicator had access to the image, the retrieved evidence when applicable, the structured rationale output, and the review comments of the initial reviewers. The final accepted record or rationale version was the adjudicated version entered into the audit log.

Certain failure types were treated as mandatory escalation events even when initially raised by only one reviewer. These included out of set citations, claims unsupported by either the image or the retrieved evidence, contradictions between the text and the observed LF-NMR image, and wording that would materially change the implied diagnosis. These rules were introduced to prioritize safety, traceability, and deployment fidelity in AppleNMR Text.

E Full Definition of TAAPM

The main paper compares models using TAAPM at the fixed decision threshold $\tau = 0.5$. The operational parameters are fixed at $Q = 1000$, $p_{\mathrm{sell}} = 1$, $c_{\mathrm{buy}} = 0.4$, $c_{\mathrm{extra}} = 0$, $\alpha = 0.05$, and $r_{\mathrm{bad}} = 0.1$. The last quantity is an operational prior, distinct from the empirical cohort prevalence of $79/237$. The threshold is not selected using validation or test labels.

TAAPM is designed for a deployment setting in which each incoming apple is assigned to one of two operational actions, namely *ship* or *reject*. The metric combines two requirements that arise in practice. First, the delivered batch must satisfy a buyer specified quality constraint on the tolerated fraction of defective fruit. Second, when that constraint is satisfied, the metric accounts for expected economic return under procurement and shipment requirements. All reported comparisons use the same fixed decision threshold.

E.1 Full Utility Formulation

Let $y \in \{0, 1\}$ denote the ground truth state of an incoming apple, where $y = 1$ indicates a defective apple and $y = 0$ indicates a healthy apple. Let $s(x) \in [0, 1]$ denote the model score for image x, interpreted as the confidence that the apple is defective. At the fixed decision threshold $\tau = 0.5$, the deployment action is defined as

$$a_\tau(x) = \begin{cases} \text{reject}, & s(x) \geq \tau, \\ \text{ship}, & s(x) < \tau. \end{cases} \tag{S42}$$

Let

$$r_{\text{bad}} = \Pr(y = 1) \tag{S43}$$

denote the defective prevalence in the incoming apple stream, so that $1 - r_{\text{bad}} = \Pr(y = 0)$ is the healthy prevalence.

Using $\tau = 0.5$, define

$$R_{\text{bad}}(\tau) = \Pr\big(a_\tau(x) = \text{reject} \mid y = 1\big), \tag{S44}$$

which is the recall on defective apples, and

$$F_{\text{good}}(\tau) = \Pr\big(a_\tau(x) = \text{reject} \mid y = 0\big), \tag{S45}$$

which is the false reject rate on healthy apples.

These quantities induce the following population level outcome probabilities:

$$\text{TP}(\tau) = r_{\text{bad}}\, R_{\text{bad}}(\tau), \tag{S46}$$
$$\text{FN}(\tau) = r_{\text{bad}}\, [1 - R_{\text{bad}}(\tau)], \tag{S47}$$
$$\text{FP}(\tau) = (1 - r_{\text{bad}})\, F_{\text{good}}(\tau), \tag{S48}$$
$$\text{TN}(\tau) = (1 - r_{\text{bad}})\, [1 - F_{\text{good}}(\tau)]. \tag{S49}$$

Here, $\text{FN}(\tau)$ corresponds to defective apples that are mistakenly shipped, while $\text{TN}(\tau)$ corresponds to healthy apples that are correctly shipped. Since only shipped apples contribute to delivered volume, the expected shipping rate is

$$k_{\text{ship}}(\tau) = \Pr\big(a_\tau(x) = \text{ship}\big) = \text{TN}(\tau) + \text{FN}(\tau). \tag{S50}$$

The defective proportion among shipped apples is therefore

$$d_{\text{ship}}(\tau) = \Pr\big(y = 1 \mid a_\tau(x) = \text{ship}\big) = \frac{\text{FN}(\tau)}{\text{TN}(\tau) + \text{FN}(\tau)} = \frac{\text{FN}(\tau)}{k_{\text{ship}}(\tau)}. \tag{S51}$$

A buyer specified quality gate is imposed on the delivered batch. Let $\alpha \in (0, 1)$ denote the maximum tolerated defective fraction in the shipped apples. The shipment at $\tau = 0.5$ is acceptable only if

$$d_{\text{ship}}(\tau) \leq \alpha. \tag{S52}$$

If (S52) is violated, the shipment is regarded as operationally unacceptable and the utility is set to zero.

Let Q denote the required number of shipped apples during a contract period, p_{sell} the selling price per shipped apple, and c_{buy} the procurement cost per purchased apple. Since only a fraction $k_{\text{ship}}(\tau)$ of purchased apples is expected to be shipped, the expected number of apples that must be procured to deliver Q shipped apples is

$$Q_{\text{buy}}(\tau) = \frac{Q}{k_{\text{ship}}(\tau)}. \tag{S53}$$

When the quality gate is satisfied, the expected contract profit under threshold τ is

$$U(\tau) = Q\, p_{\text{sell}} - Q_{\text{buy}}(\tau)\, c_{\text{buy}} = Q\, p_{\text{sell}} - \frac{Q}{k_{\text{ship}}(\tau)}\, c_{\text{buy}}. \tag{S54}$$

Symbol	Definition and practical interpretation
r_{bad}	Defective prevalence in the incoming apple stream. This quantity describes the base rate of risky fruit entering the grading line. A larger r_{bad} increases the chance that missed defects contaminate the shipped batch.
$R_{\mathrm{bad}}(\tau)$	Recall on defective apples, that is, the probability of correctly rejecting a defective apple at threshold τ. Larger values improve shipment safety by reducing the mass of missed defective fruit.
$F_{\mathrm{good}}(\tau)$	False reject rate on healthy apples, that is, the probability of wrongly rejecting a healthy apple at threshold τ. Larger values reduce usable throughput and increase procurement demand.
$\mathrm{FN}(\tau)$	Population fraction of defective apples that are shipped. This is the most critical failure term because it directly increases the defective proportion in the delivered batch.
$\mathrm{TN}(\tau)$	Population fraction of healthy apples that are shipped. This is the productive throughput term that contributes to fulfilling the contract efficiently.
$k_{\mathrm{ship}}(\tau)$	Expected shipping rate, equal to the fraction of incoming apples that are shipped. It determines how much upstream procurement is required to deliver a fixed shipment target.
$d_{\mathrm{ship}}(\tau)$	Defective fraction among shipped apples. This is the delivered batch quality statistic inspected by the buyer or auditor.
α	Maximum tolerated defective fraction in the shipment. This parameter encodes the buyer acceptance criterion. Smaller α values impose a stricter quality regime.
Q	Required number of shipped apples in the contract period. This converts per sample classification behavior into batch level operational utility.
p_{sell}	Selling price per shipped apple. This determines the revenue side of the contract.
c_{buy}	Procurement cost per purchased apple. This quantifies the upstream acquisition cost, including fruit that will later be rejected by the grading system.
$Q_{\mathrm{buy}}(\tau)$	Expected number of apples that must be purchased in order to ship Q apples. It expands as $k_{\mathrm{ship}}(\tau)$ decreases, thereby capturing the cost of excessive rejection.
τ	Decision threshold on the model score, fixed at 0.5 for all experiments. Validation and test labels are not used to select it.

Table S3: Parameters used in TAAPM and their deployment interpretation.

The Task Aware Apple Pathology Metric is therefore defined as

$$\mathrm{TAAPM}(\tau) = \begin{cases} Q\,p_{\mathrm{sell}} - \dfrac{Q}{k_{\mathrm{ship}}(\tau)}\,c_{\mathrm{buy}}, & d_{\mathrm{ship}}(\tau) \le \alpha, \\ 0, & d_{\mathrm{ship}}(\tau) > \alpha. \end{cases} \tag{S55}$$

Equation (S55) clarifies the compact form shown in the main paper. The expression

$$\mathrm{TAAPM} = Q\,p_{\mathrm{sell}} - \frac{Q}{k_{\mathrm{ship}}}\,c_{\mathrm{buy}} \tag{S56}$$

is precisely the feasible case utility, namely the branch of (S55) that applies after the shipment has passed the quality gate.

All reported TAAPM values are evaluated at the fixed threshold:

$$\mathrm{TAAPM}_{\mathrm{reported}} = \mathrm{TAAPM}(0.5). \tag{S57}$$

For each experimental setting, the five-fold evaluation results are first aggregated within each seed. The reported mean and standard deviation are then calculated across the three seed-level results. Conventional accuracy, balanced accuracy, F1 score, and MCC do not encode either the shipment acceptability constraint in (S52) or the procurement efficiency term in (S54).

E.2 Parameter Definitions and Practical Interpretation

For clarity, Table S3 summarizes all quantities appearing in TAAPM and their operational meaning.

The practical meaning of TAAPM can be understood from two coupled operational pressures. The first pressure is shipment quality. If $R_{\mathrm{bad}}(\tau)$ is too low, then $\mathrm{FN}(\tau)$ becomes large, $d_{\mathrm{ship}}(\tau)$ rises,

and the quality gate in (S52) may be violated. In that case, the deployment value collapses to zero regardless of conventional classification scores. This mechanism captures the severe consequence of shipping defective apples in an industrial contract setting.

The second pressure is procurement efficiency. If $F_{\mathrm{good}}(\tau)$ is too high, then many healthy apples are unnecessarily rejected. This lowers $k_{\mathrm{ship}}(\tau)$, increases $Q_{\mathrm{buy}}(\tau)$, and reduces expected profit through the procurement term in (S54). Even when the quality gate remains satisfied, an overly conservative model can therefore achieve poor operational utility because it consumes too much healthy inventory to meet the required shipment target.

TAAPM thus combines a hard acceptance constraint with a soft economic objective. The acceptance constraint is mainly governed by the quadruple $(r_{\mathrm{bad}}, R_{\mathrm{bad}}(\tau), \mathrm{FN}(\tau), \alpha)$, while the economic objective is mainly governed by $(Q, p_{\mathrm{sell}}, c_{\mathrm{buy}}, k_{\mathrm{ship}}(\tau))$. At the fixed threshold of 0.5, higher feasible utility corresponds to greater economically useful throughput under the stated contract.

A further advantage of this formulation is that the asymmetry of decision consequences is built into the structure of the metric itself. There is no need to introduce manually chosen per error penalties. Missed defects are already amplified by the buyer acceptance gate, whereas unnecessary rejection of healthy apples is already reflected by the procurement expansion factor $1/k_{\mathrm{ship}}(\tau)$. As a result, TAAPM remains economically interpretable and directly aligned with deployment.

In the reported experiments, all methods are evaluated at $\tau = 0.5$ under the same fixed parameterization $(r_{\mathrm{bad}}, \alpha, Q, p_{\mathrm{sell}}, c_{\mathrm{buy}})$. Under this protocol, differences in TAAPM arise only from the model dependent operating characteristics $R_{\mathrm{bad}}(\tau)$ and $F_{\mathrm{good}}(\tau)$. Consequently, TAAPM can be read as the expected deployment value of a model under a specified grading contract, rather than as a generic classification score.

F Full Training Details

This section consolidates the full implementation details used in all reported experiments. The complete pipeline follows the deployment constrained setting of AppleNMR Text, where each sample consists of a standardized LF-NMR mosaic, an expert verified text record, and a binary label, while only the LF-NMR image is assumed available at inference time. Training is organized into two stages. Stage I performs retrieval grounded rationale generation from the image, and Stage II performs the final multimodal classification conditioned on the image, the generated rationale, and the retrieved evidence text. All preprocessing, retrieval, model selection, and final evaluation are conducted under the same fold boundaries.

Implementation summary. The final two-stage system uses the 8B Stage I generator. The 2B, 4B, and 8B settings in the direct-classification comparison are model-scale baselines, not alternative generator configurations of the final system. Stage I uses LoRA rank 64, a learning rate of 1×10^{-4}, batch size 32, and 50 training epochs. Stage II is also trained for 50 epochs. Retrieval uses 2048-dimensional embeddings and a total top-k of 10: five knowledge items and five case items.

Repeated evaluation. All reported experimental settings use three random seeds (2024, 2025, and 2026), each with a complete five-fold cross-validation procedure. The five-fold results are first aggregated into one result per seed. The reported mean and standard deviation are then calculated across these three seed-level results, rather than directly across the 15 fold-level results.

F.1 Data Processing and Augmentation

A total of 273 apples were collected. Data-completeness and quality-control screening retained 237 samples, including 79 positive samples, for the final cohort. The other 36 samples were excluded before model training and evaluation on data-quality or completeness grounds, rather than model predictions or downstream outcomes.

For each apple, the raw LF-NMR acquisition contains 15 slice images. A standardized 2×2 mosaic is constructed so that all downstream models receive a unified visual input. The preprocessing pipeline first detects the largest outer apple bounding box over the 15 slices. The slice with the maximal detected area defines the reference crop, and this maximal box is then propagated to all

slices of the same sample, ensuring a consistent field of view across the full slice stack. Four consecutive slices centered around the maximal response position are then selected, with boundary correction applied when the maximal slice lies near the beginning or end of the stack. These four slices are tiled into a 2×2 composite image and resized to the target network resolution.

Multimodal pairing is resolved at the sample level through a unique identifier. Each apple identifier links its LF-NMR mosaic, expert record, and binary label. Any augmented version of an image inherits the same identifier, label, and paired text record as the original image. Dataset partitioning is always performed before augmentation, and all image text pairings are resolved only within the corresponding fold after the split has been fixed.

Augmentation is applied only to the training portion of each fold. In order to preserve physically meaningful NMR structure, only geometry preserving transforms are used, including rotations by integer multiples of 90°, horizontal flips, and vertical flips. No random cropping, color jitter, elastic deformation, or intensity perturbation is introduced. Test images undergo only deterministic preprocessing, namely crop propagation, 2×2 mosaic construction, resizing, and normalization.

Class imbalance is handled only inside the training folds. The minority diseased class is oversampled until the effective class distribution becomes approximately balanced, while the validation and test partitions preserve the original prevalence. For Stage II, the final image input is resized to 384×384 and normalized channel wise to zero mean and unit variance.

F.2 Stage I Training Configuration

Stage I uses the 8B Qwen3-VL generator for image-conditioned rationale construction. It is fine-tuned with LoRA rank 64, a learning rate of 1×10^{-4}, batch size 32, and 50 epochs. The 2B/4B/8B direct-classification baselines do not represent alternative generators of the final two-stage system.

Stage I uses two external memories. The knowledge memory stores curated domain knowledge units, where each chunk is assigned a persistent evidence identifier. The case memory stores historical labeled cases and expert verified records, indexed by image embeddings so that visually similar precedents can be retrieved during both training and inference. All retrieval components are strictly fold isolated. For a given outer fold, the vector index, reranker, preprocessing pipeline, and all retrieval assets are built only from the training partition of that fold. The held out test fold is excluded from index construction, retrieval, and any auxiliary retrieval stage. During training, an additional apple-ID filter excludes the query apple's original image, augmented images, and all other images from that apple. Retrieved case summaries also hide label fields before being exposed to the generator or the classifier, preventing target leakage.

The Stage I query is deterministically constructed from the input LF-NMR image through a fixed rule based tokenizer. Labels, expert records, and any test time annotations are never used to form the query. When metadata are available, normalized metadata tags may be appended to reduce retrieval drift under storage or acquisition variation. Retrieval uses embeddings of dimension 2048 and proceeds through dual memory search. In the case memory, approximate nearest neighbor search is first conducted in image embedding space, followed by similarity reranking and diversity filtering to suppress near duplicate precedents. In the knowledge memory, dense semantic retrieval is fused with sparse lexical matching through rank aggregation. The final retrieved items are serialized into an Evidence Pack with persistent evidence identifiers. The knowledge memory contributes five items and the case memory contributes five items, giving ten retrieved items in total. The reranking depth, diversity threshold, and token budget are not specified in the available records.

A key protocol detail is that retrieval is executed during Stage I training rather than introduced only at inference time. For every training and validation sample, evidence is freshly retrieved within the current fold, packed into the same Evidence Pack schema used at deployment, and then provided to the LoRA adapted generator together with the input image. This design ensures that the generator is optimized under the same retrieval conditioned input distribution that it encounters at inference time, thereby avoiding train inference mismatch.

The Stage I output is both schema constrained and citation constrained. Each generated record contains image findings, evidence identifiers, supported claims, uncertainty statements, a confidence field, and a final rationale. After generation, post hoc validation checks whether every cited evidence identifier belongs to the retrieved evidence set E_i. Any citation that falls outside the retrieved set

is either deterministically repaired or the sample is marked invalid according to the implemented validation rule. The final rationale is then normalized into a compact rationale $\tilde{r}_i$ for downstream Stage II use.

Other Stage I optimization settings and the checkpoint-selection implementation require verification from the original experiment records.

F.3 Stage II Training Configuration

Stage II is the final multimodal classifier used for binary diagnosis. The adopted visual encoder is VGG19 trained from scratch to match LF-NMR image statistics. All original fully connected classification layers are removed, and a global average pooling layer is appended after the final convolutional block to obtain a compact visual representation. The input to the visual branch is the preprocessed LF-NMR mosaic resized to 384×384 and normalized channel wise to zero mean and unit variance.

The adopted text encoder is `bert base uncased`. The corresponding tokenizer is used, and the input sequence is truncated or padded to a fixed maximum length (not specified in the supplied files). Transformer layers 1 through 10 are frozen. Only the top two transformer layers, layer normalizations, and a lightweight projection head are updated during training. These backbone choices correspond to the final adopted configuration, while the cited preliminary comparison was not present in the supplied archive.

The visual and textual features are mapped by modality specific projector MLPs into a shared embedding space of dimension 2048, yielding z_{img} and z_{text}. These embeddings are fused by a gated fusion module ψ_{fusion} to produce the final multimodal representation z_{fusion}, which is then passed to a lightweight binary classification head. The Stage II input contains the LF-NMR image, the compact rationale generated by Stage I, and the retrieved evidence text. Label fields from retrieved cases are excluded from the classifier input.

In addition to the classification loss, Stage II uses four auxiliary objectives. First, an intra modal image image supervised contrastive loss is applied to two augmented views of each training image. Positives are defined by shared class labels, so the anchor uses both another view of the same image and views from other images in the same class. Second, a masked language modeling objective is attached to the text branch in order to improve robustness to noisy phrasing. The masking ratio is fixed to 10%, and masked positions follow the standard 80 10 10 corruption rule. Third, a curriculum image text contrastive objective is employed to stabilize early alignment and later encourage class aware cross modal separation. Fourth, an EMA teacher regularizes the fused representation and improves optimization stability.

The overall Stage II objective is

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda_{\text{IIC}}\mathcal{L}_{\text{IIC}} + \lambda_{\text{ITC}}\mathcal{L}_{\text{ITC}} + \lambda_{\text{MLM}}\mathcal{L}_{\text{MLM}} + \lambda_{\text{MSE}}\mathcal{L}_{\text{MSE}}.$$

Stage II training lasts 50 epochs and proceeds in two phases. During the warmup phase, which occupies the first 20% of these 50 epochs, the classifier head and the EMA teacher are disabled. The optimization objective in this phase is

$$\mathcal{L}_{\text{warm}} = \lambda_{\text{IIC}}\mathcal{L}_{\text{IIC}} + \lambda_{\text{MLM}}\mathcal{L}_{\text{MLM}} + \lambda_{\text{align}}\mathcal{L}_{\text{align}},$$

with fixed coefficients

$$\lambda_{\text{IIC}} = \lambda_{\text{MLM}} = \lambda_{\text{align}} = 1.$$

This stage is designed to stabilize unimodal representation learning and positive image text alignment before stronger discriminative pressure is introduced.

After warmup, the classifier head, the full class aware ITC objective, and the EMA teacher are activated. Let $T = 50$ denote the number of Stage II epochs and T_w the warmup length. For each epoch index $t > T_w$, a normalized progress variable is defined as

$$\rho(t) = \frac{t - T_w}{T - T_w}, \qquad \rho(t) \in [0, 1],$$

and the shared auxiliary weight follows

$$\lambda(t) = \lambda_0\big(1 - \rho(t)\big), \qquad \lambda_0 = 0.1.$$

The total loss in the main phase is therefore

$$\mathcal{L}_{\text{total}}(t) = \mathcal{L}_{\text{CE}} + \lambda(t)\big(\mathcal{L}_{\text{IIC}} + \mathcal{L}_{\text{MLM}} + \mathcal{L}_{\text{ITC}} + \mathcal{L}_{\text{MSE}}\big).$$

This schedule encourages strong alignment and denoising signals during early main phase training, while progressively shifting emphasis toward sharpening the decision boundary of the classifier.

Hard negative mining is activated only during the main phase of class aware image text contrastive learning. For an anchor embedding z_i, the positive pool contains the paired text embedding together with all text embeddings from the same class, while the negative pool contains text embeddings from the opposite class. The hard negative set is defined by the top K most similar negatives under cosine similarity, with $K = 16$ in the final adopted configuration. Only the positives and these hard negatives participate in the contrastive denominator. Their weights are normalized so that negative contribution remains balanced relative to the total positive mass.

The EMA teacher is inactive during warmup and enabled only in the main stage. After each student update, the teacher parameters are updated by exponential moving average with momentum $m = 0.95$,

$$\theta^T \leftarrow m\theta^T + (1-m)\theta^S.$$

The teacher supplies a smoothed target for the fused representation, and the student is regularized toward this target through a mean squared error term.

Stage II optimization uses AdamW in both phases and runs for 50 epochs with seeds 2024, 2025, and 2026. Model selection is conducted strictly within each outer fold, using validation TAAPM evaluated at the fixed threshold of 0.5 as the primary checkpoint criterion. This checkpoint selection does not change the classification threshold. Balanced accuracy, F1 score, MCC, and validation loss are recorded as secondary diagnostics. The held out outer test fold is never used for checkpoint selection, early stopping, retrieval construction, or hyperparameter tuning.

F.4 GPT-Based Evaluation

The GPT-based experiments use the actual model configurations GPT-5-Instant and GPT-5-Thinking. The model calls were conducted on November 23, 2025. The experimental input protocol is applied consistently within each corresponding model condition.