RAREVA: ADAPTIVE AGENTIC VERIFICATION FOR RARE DISEASE CANDIDATE PRIORITIZATION FROM FACIAL IMAGES UNDER DATA SCARCITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Rare disease candidate prioritization from facial images is particularly challenging under extreme data scarcity, where genetic testing and structured clinical annotations may be unavailable and visually similar disorders can exhibit overlapping craniofacial phenotypes. Existing approaches typically formulate this setting as either direct visual classification or end-to-end vision–language reasoning, providing limited mechanisms for explicitly verifying why one candidate disease should be favored over competing alternatives. We propose **RareVA**, an image-first system that formulates rare disease candidate prioritization as candidate-conditioned disease verification. Rather than directly selecting a disease from the visual prediction space, RareVA first identifies a small set of plausible candidates, extracts image-derived facial phenotypes through a vision–language model, and verifies each candidate through agentic reasoning over evidence retrieved from a disease–phenotype knowledge graph. To improve computational efficiency, RareVA further introduces Adaptive Candidate Selection (ACS), which dynamically allocates the number of candidates requiring verification according to case difficulty. We evaluate RareVA on RDFace and further assess generalization on the independent GMDB benchmark. RareVA improves Top-1 accuracy from 25.66% to 34.51% on RDFace and from 7.56% to 12.25% on GMDB. ACS further reduces average output-token cost by 17.15% relative to the best-performing fixed-budget configuration while achieving comparable observed accuracy. These results support candidate-conditioned verification as a formulation for evidence-grounded rare disease candidate prioritization from facial images under severe data scarcity.

1 INTRODUCTION

Early rare disease assessment remains one of the most challenging problems in clinical decision support because individual disorders are rare, phenotypically heterogeneous, and supported by limited labeled data, often resulting in prolonged diagnostic odysseys before a definitive diagnosis is established Castro et al. (2017); Baynam et al. (2016); Anderson et al. (2013); Zurynski et al. (2017). In many real-world scenarios, particularly during early pediatric screening and in low-resource settings, genetic testing and structured clinical annotations are not immediately available. Facial appearance therefore becomes one of the earliest, and sometimes the only, sources of evidence for differential diagnosis and referral decisions Pavisich et al. (2024); Oniūnaitė et al. (2025); Qiang et al. (2022).

This image-first setting is inherently difficult because it combines extreme data scarcity, subtle inter-disease visual similarity, and incomplete visual evidence Kováč et al. (2024); Song et al. (2025a). It also raises a fundamental question for AI: *how can rare disease candidates be reliably prioritized when only one to several facial images are available for each disease?* Simply increasing supervision is often impractical because of the rarity of individual disorders. Likewise, grouping diseases into broader clinical categories does not fully resolve the problem, since visually similar disorders frequently span different clinical taxonomies while diseases within the same subgroup may exhibit markedly different facial characteristics Hsieh & Krawitz (2023). These observations suggest that existing image-first approaches remain limited because they primarily focus on improving direct visual recognition rather than determining how evidence should be used to compare and reprioritize competing rare disease hypotheses.

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

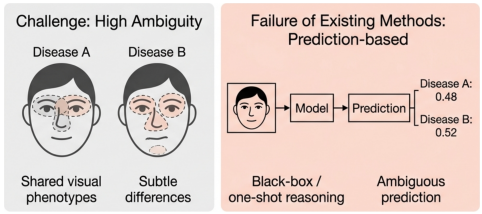


Figure 1: Illustration of ambiguity and failure of direct prediction.

Existing AI approaches largely follow two paradigms. Deep visual classifiers achieve promising recognition performance but operate as direct predictors with limited support for evidence-based comparison among competing disease candidates Gurovich et al. (2019); Hsieh et al. (2022); van der Velden et al. (2022). More recently, large language model (LLM)- and vision-language model (VLM)-based agent systems have demonstrated strong reasoning capabilities in medical decision support Cheng et al. (2025); Zhao et al. (2026); Chen et al. (2026), yet most assume structured clinical evidence is already available and are not designed for directly reasoning over facial images under scarce rare disease knowledge Jin et al. (2025); Hamza et al. (2025). Consequently, neither paradigm explicitly supports structured, evidence-based verification and prioritization of a bounded set of competing disease hypotheses (Figure 1).

We argue that the limitation lies in the formulation itself. In clinical practice, facial appearance is generally not sufficient to establish a definitive diagnosis in isolation. Instead, it can be used to generate a set of plausible disease hypotheses, after which each candidate is evaluated against disease-specific phenotypic evidence before subsequent clinical assessment. Inspired by this workflow, we formulate image-first rare disease analysis as a *candidate-conditioned verification and prioritization* problem rather than direct recognition. Given a facial image and an initial set of rare disease candidates, the objective is to verify phenotype–disease consistency and reprioritize the candidate set using explicit evidence. This formulation supports structured comparison among competing hypotheses while enabling adaptive allocation of reasoning resources according to case difficulty.

Based on this formulation, we propose **RareVA** (**Rare** Disease **V**ision **A**gent), an image-first multi-agent system for rare disease candidate verification and prioritization. RareVA explicitly separates visual perception from candidate-level reasoning through an intermediate phenotype representation. A visual classifier first generates a bounded rare disease candidate set, a VLM extracts image-derived facial phenotypes, and an agentic reasoning system verifies each candidate using evidence retrieved from a curated disease–phenotype knowledge graph (KG). The final system reprioritizes the candidate set according to the accumulated verification evidence rather than performing open-set disease detection. To improve computational efficiency, we further introduce **Adaptive Candidate Selection (ACS)**, an optimization-based policy that dynamically allocates reasoning budgets according to case difficulty rather than using a fixed verification budget for every case Teerapittayanon et al. (2016). Finally, we perform uncertainty-aware evaluation using bootstrap confidence intervals to quantify the statistical reliability of observed performance differences.

In summary, this work makes the following contributions:

- We formulate image-first rare disease analysis as a **candidate-conditioned verification and prioritization** problem, shifting the focus from direct visual recognition to explicit evidence comparison and re-prioritization of competing disease hypotheses under uncertain visual evidence.
- We propose **RareVA**, an image-first agentic system that grounds candidate verification using image-derived facial phenotypes and a curated disease–phenotype KG through multi-agent collaboration and reasoning.
- We introduce **ACS**, an adaptive reasoning policy that reduces unnecessary agentic inference by allocating verification effort according to case difficulty, and evaluate its cost–accuracy trade-off together with uncertainty estimates obtained through bootstrap confidence intervals.

2 Related Work

Facial analysis for rare disease diagnosis. Automated diagnosis from facial images has evolved from geometric feature analysis to deep learning–based approaches Hsieh & Krawitz (2023); Sherif et al. (2024); Germanese et al. (2023). Early works relied on handcrafted anthropometric landmarks Narayanan et al. (2019), while deep convolutional neural networks (DCNNs) enabled supervised classification paradigms such as DeepGestalt Gurovich et al. (2019). To address data scarcity, GestaltMatcher Hsieh et al. (2022) reframed the task as metric learning, and recent benchmarks further support evaluation under low-resource settings Hsieh et al. (2022); Feng et al. (2026). De-

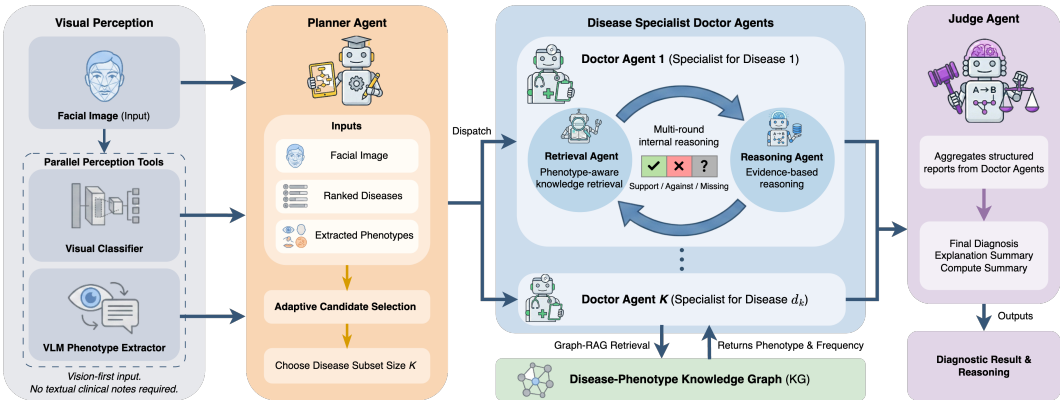


Figure 2: Overall architecture pipeline of RareVA.

spite strong discriminative performance, these methods primarily provide visual similarity scores or direct disease rankings, but do not explicitly verify disease-specific evidence for each candidate.

Multimodal and VLM-based approaches. Recent methods incorporate additional modalities such as clinical text or demographic attributes Wu et al. (2024), which are often unavailable in image-first settings. Vision-language models (VLMs) bridge perception and semantics by generating structured representations from images Li et al. (2022; 2023); Yang et al. (2025). However, end-to-end VLM reasoning remains sensitive to subtle visual variations and typically relies on global or single-pass inference Zhang et al. (2025); Hartsock & Rasool (2024); Cheng et al. (2025). Although prompt engineering can improve VLM outputs, it does not by itself change the underlying inference structure from global diagnosis to candidate-wise evidence verification Gu et al. (2023); Gao et al. (2023).

LLM-based multi-agent reasoning systems. Knowledge graphs and retrieval-augmented generation have been used to provide external medical evidence for LLM-based diagnosis Zuo et al. (2025); Song et al. (2025b). In rare disease diagnosis, RareAgents Chen et al. (2026) and DeepRare Zhao et al. (2026) simulate expert collaboration and support traceable reasoning. However, these systems primarily operate on structured textual inputs. They assume textual cases or curated phenotypes are already available, which is not the case in image-first settings where phenotypes must be extracted from visual inputs and reasoning must remain robust to noisy intermediate representations.

3 Methodology

3.1 Problem Formulation

Let $x \in \mathcal{X}$ denote an input facial image and $\mathcal{Y}$ the predefined rare disease label space. Rather than performing open-set disease detection or directly predicting over the entire label space, RareVA operates on an initial set of candidate diseases proposed from the image. A visual classifier first produces a ranked candidate set $C(x) \subset \mathcal{Y}$. To control computational cost, agentic verification is performed over a selected subset $D(x) \subseteq C(x)$. For each candidate $y \in D(x)$, RareVA estimates an evidence-consistency score $f(x, y)$ reflecting the support for that candidate from image-derived phenotypes and disease-specific knowledge. The highest priority candidate after verification is

$$\hat{y} = \arg \max_{y \in D(x)} f(x, y). \tag{1}$$

where $\hat{y}$ denotes the top-ranked candidate after verification.

RareVA decouples visual perception from agentic reasoning through an intermediate phenotype representation and follows a two-stage process. (1) *Visual perception:* a vision classifier (VisCls) proposes candidate diseases, while a VLM extracts phenotypic evidence from the image. (2) *Agentic reasoning:* the system performs candidate-conditioned multi-agent reasoning to determine which proposed disease is best supported by the visual evidence and disease-specific knowledge.

This formulation separates candidate generation (whether the true disease is proposed), from candidate-conditioned verification (how effectively the system re-prioritizes among the candidates).

This separation enables controlled evaluation of agentic reasoning independent of candidate recall while exposing the limitations imposed by the visual classifier prior. The overall architecture is illustrated in Figure 2, and the procedure is in Algorithm 1. The prompt sets are provided in Appendix A.

Algorithm 1 Overall RareVA Candidate Verification and Prioritization

Require: Input facial image x; candidate limit $K_{\max}$; VisCls f_θ; VLM phenotype extractor $\mathcal{V}$; ACS policy π; disease–phenotype KG $\mathcal{G}$; agent $\mathcal{L}$; max round $T_{\max}$

$C(x) \leftarrow f_\theta(x, K_{\max})$
$P(x) \leftarrow \textsc{ConsensusAndNormalize}(\mathcal{V}(x))$
$\psi(x) \leftarrow \textsc{RoutingFeatures}(x, C(x), P(x))$
$K(x) \leftarrow \pi(\psi(x)) \in 2, \ldots, K_{\max}$ {adaptive allocation of verification budget}
$D(x) \leftarrow \mathcal{L}_{\textsc{Planner}}(x, P(x), C(x), K(x))$
for each disease $d \in D(x)$ **do**
  while round $< T_{\max}$ **and** $\neg$, $\textsc{Stop}(\mathcal{L}\textsc{Doctor})$ **do**
    $E_d \leftarrow \mathcal{L}\textsc{Retrieval}(\mathcal{G}, d)$
    $A_d \leftarrow \mathcal{L}\textsc{Reasoning}(x, P(x), d, E_d)$
  end while
end for; Top-priority candidate $y^\star$, rationale $r \leftarrow \mathcal{L}\textsc{Judge}(A_{d\,d\in D(x)})$

3.2 Visual Perception

The perception stage transforms raw facial appearance into structured intermediate representations suitable for downstream reasoning. A VisCls f_θ produces a ranked candidate set by estimating class probabilities $p_\theta(y \mid x)$ and selecting $C(x) = \text{Top-}K_{\max}(p_\theta(\cdot \mid x))$. In parallel, a VLM extracts phenotype terms from the image through N runs, followed by consensus aggregation, $P_{\text{term}}(x) = \textsc{Consensus}(\{\mathcal{V}^{(n)}(x)\}_{n=1}^{N})$, where $\mathcal{V}^{(n)}(x)$ denotes the phenotype term set from the n-th run.

The consensus term set is then mapped to canonical Human Phenotype Ontology (HPO) terms Gargano et al. (2023) (a standardized vocabulary) using Doc2HPO Liu et al. (2019), yielding $P(x) = \textsc{Doc2HPO}(P_{\text{term}}(x))$. The resulting HPO phenotype set serves as a structured intermediate representation linking visual evidence to disease-specific knowledge. Additional details on consensus aggregation and ontology mapping are provided in Appendix D.

3.3 Multi-Agent Agentic Reasoning

After the perception stage, we turn to agentic reasoning that verifies and aggregates evidence via a multi-agent system of three agent roles: *Planner*, *Doctor*, and *Judge*.

3.3.1 Planner Agent

The *Planner Agent* formulates adaptive reasoning allocation as an *optimization* problem, determining how much computation cost should be allocated to each case ($D(x)$). Cases with confident visual evidence generally require fewer verifications, whereas ambiguous cases benefit from additional verifications. This optimization is implemented via ACS, where a policy π maps a difficulty representation $\psi(x)$ to a candidate budget: $K(x) = \pi(\psi(x)), K(x) \in \mathcal{K}$. We formulate ACS as a utility maximization objective balancing performance and computational cost:

$$\max_{\pi} \; \mathbb{E}_x \left[\mathbb{I}(\hat{y}_{K(x)} = y_{\text{gt}}) - \lambda \cdot \text{Cost}(x, K(x)) \right], \tag{2}$$

where $\text{Cost}(\cdot)$ measures inference overhead (e.g., number of output tokens), and λ controls the accuracy–efficiency trade-off. Implementation details are provided in Appendix E.

3.3.2 Hierarchical Doctor Agents

As shown in Figure 3, for each candidate disease $d \in D(x)$, we instantiate a *Doctor Agent* that performs candidate-conditioned verification via iterative retrieval and reasoning. The pseudocode is provided in Algorithm G.1 with disease–phenotype KG details provided in Appendix F.

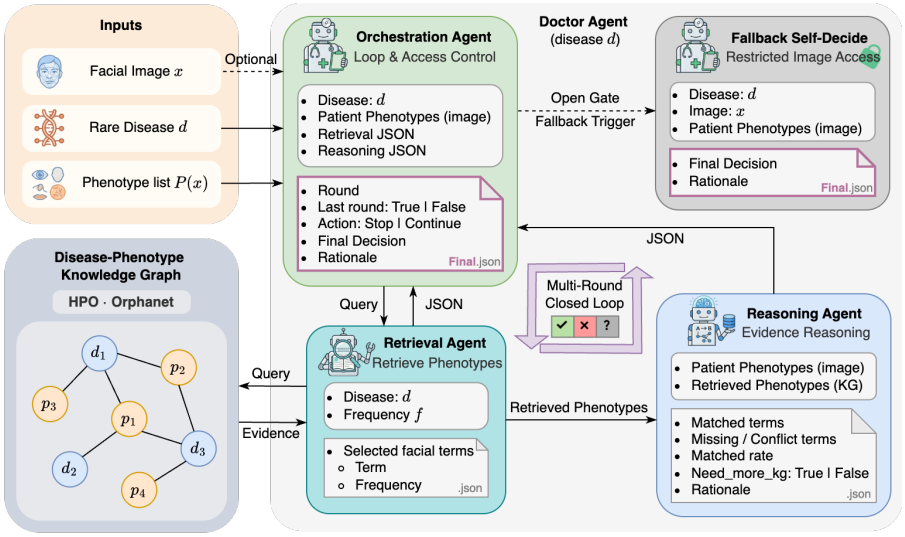


Figure 3: Architecture of a Hierarchical Doctor Agent.

At round t, the orchestration agent in the Doctor Agent maintains a *memory state* $M_d^{(t)}$ that accumulates retrieved evidence across rounds, updated as:

$$M_d^{(t)} = M_d^{(t-1)} \cup KG(d, \rho^{(t)}), \tag{3}$$

where $KG(d, \rho^{(t)})$ retrieves phenotype evidence under frequency constraint $\rho^{(t)}$. Given the updated evidence, the agent evaluates phenotype–disease compatibility:

$$A_d^{(t)} = \mathcal{R}\big(P(x), M_d^{(t)}\big), \tag{4}$$

where $\mathcal{R}$ computes an evidence-grounded assessment based on the observed phenotype set $P(x)$. This retrieval–reasoning loop iterates until a stopping condition is met (e.g., round limit), following a *ReAct-style* interaction pattern Yao et al. (2023). The final decision is taken from the first stopping round:

$$A_d^{\text{final}} = \begin{cases} A_d^{(t^\star)}, & \exists\, t^\star : \ \textsc{Stop}(A_d^{(t^\star)}), \\ \textsc{Fallback}(x, P(x), d), & \text{otherwise.} \end{cases} \tag{5}$$

The fallback mechanism enables direct image-based assessment when phenotype-grounded reasoning is insufficient. More details on the Doctor Agent and its fallback mechanism are in Appendix G.

3.3.3 Judge Agent

The *Judge Agent* receives the structured verification reports produced by all Doctor Agents and performs comparative evidence aggregation. Rather than independently predicting a disease from the original image, the Judge compares the consistency, completeness, and disease specificity of the evidence associated with each candidate and produces the final prioritized ordering. When multiple Doctor Agents report strong support, the Judge resolves their relative ordering through comparative evidence analysis. When support is weak or conflicting across candidates, the Judge records this uncertainty while still returning the highest-priority candidate required by the closed-set evaluation protocol. Detailed decision templates are provided in Appendix H.

4 Experiment

4.1 Experimental Setup

Datasets and splits. We adopt a two-stage evaluation protocol consisting of a controlled development benchmark and an external validation cohort. We use *RDFace* Feng et al. (2026) for method development. The dataset contains 456 facial images spanning 103 genetically confirmed rare diseases. Each image is collected from published clinical case reports and linked to a literature source,

Table 1: Overall performance under the candidate-bounded protocol. Top-1 accuracy with 95% bootstrap CI and the improvement in percentage points (pp) over the prior with 95% bootstrap CI.

Method	Top-1 Acc. (%)	Top-1 Acc. CI	Δ (pp)	Δ CI
Perception Prior - VisCls				
DenseNet Huang et al. (2017)	<u>25.66</u>	[17.67, 33.63]	—	—
CLIP Radford et al. (2021)	17.70	[10.62, 24.78]	—	—
Gestalt Hsieh et al. (2022)	13.27	[7.08, 19.47]	—	—
Baselines (same candidate set)				
Non-agentic reranker	22.12	[15.04, 30.09]	− 3.54	[-7.08, -0.88]
Structured phenotype scoring	25.66	[17.70, 33.63]	+ 0.00	[-2.65, 2.65]
Single VLM (Gemini)	26.55	[18.58, 34.51]	+ 0.89	[-4.42, 6.19]
Single VLM + Pheno	21.24	[14.16, 29.20]	− 4.42	[-8.85, 0.00]
Single VLM + KG-RAG	28.32	[20.35, 36.28]	+ 2.66	[-2.65, 7.96]
Single VLM + KG-RAG + multi-round	29.20	[21.24, 37.17]	+ 3.54	[-1.77, 8.85]
Our Method – RareVA (full)	**34.51**	**[25.66, 43.36]**	**+ 8.85**	**[2.65, 15.93]**

providing disease labels together with curated metadata. We use a fixed 65%/10%/25% train/validation/test split, with all hyperparameter tuned on the validation split. The test set is held out for final test. To evaluate generalization beyond the development benchmark, we further use *GMDB* Hsieh et al. (2022) as an external validation set. More details on datasets are provided in Appendix B.

Phenotype annotation subset. For 261 RDFace images, we construct reference ground truth (GT) phenotype sets ($P_{\mathrm{GT}}(x)$) from the corresponding sources. These annotations are used exclusively for evaluating phenotype extraction quality and are never exposed to RareVA during inference.

Budgeting protocol. We set a maximum candidate budget $K_{\max} = 5$ and compare fixed-budget inference with ACS. The ACS policy is trained on the validation set and remains fixed during testing.

Baselines. We compare against multiple baselines. All operate on the same candidate set $C(x)$:

- **Perception prior.** Use the VisCls Top-1 prediction as the final prediction.
- **Non-agentic reranker.** Rank candidates by cosine similarity between extracted phenotypes and disease embeddings derived from the KG using BioBERT Lee et al. (2019).
- **Structured phenotype scoring.** Apply a rule-based scoring scheme over matched, spurious, and missing phenotypes using KG statistics and BioBERT. $score(d) = \sum_{\text{matched}} w(p,d) - \alpha \sum_{\text{missing}} w(p,d) - \beta \sum_{\text{spurious}} w(p)$
- **Single VLM.** A single VLM selects the final disease from the candidate set in a zero-shot manner.
  - **+ Phenotype.** The VLM is additionally provided with extracted phenotype information.
  - **+ KG-RAG.** The VLM is further provided with retrieved KG evidence.
  - **+ multi-round.** The VLM is further provided with multi-round retrieved KG evidence.

Evaluation metrics. We use Top-1 accuracy (Acc.) as the primary metric for performance. Since prediction is performed over a candidate set, we additionally report the Top-5 Acc. of the VisCls to indicate the upper bound of candidate coverage. To quantify the *statistical uncertainty* of performances, we report bootstrap 95% confidence intervals (CI) and standard errors (SE) (Appendix I.2). For *efficiency*, we use the selected candidate budget $K(x)$ and the total number of output tokens.

4.2 Overall Performance

Table 1 reports the overall performance on the RDFace test set under a shared candidate protocol. DenseNet is fixed as the VisCls backbone for all methods, achieving 25.66% Top-1 accuracy and 43.36% Top-5 candidate coverage.

RareVA achieves 34.51% Top-1 accuracy, improving over the VisCls prior by 8.85 pp. In contrast, direct phenotype-based reranking reaches 22.12%, while structured phenotype scoring matches the visual prior at 25.66%. These results show that extracted phenotypes are not sufficiently reliable as a direct decision signal. A single VLM adjudicator with Gemini GeminiTeam (2025) achieves only marginal improvement (26.55%), and incorporating phenotype prompts can even degrade performance, suggesting that directly using phenotypes does not reliably improve decision making. Adding KG-based retrieval provides limited gains, indicating that knowledge alone is insufficient.

Table 2: External validation on GMDB.

Method	Top-1 (%)	95% CI	Δ (pp)
VisCls	7.56	[5.60, 9.68]	—
Filtered TENT	9.98	[7.72, 12.25]	+2.42
Strongest VLM	10.74	[8.62, 13.46]	+3.18
RareVA	**12.25**	[9.98, 15.13]	**+4.69**

Table 3: Phenotype extraction performance.

Method	F1	Prec.	Rec.
Best-of-3 Oracle	0.6376	0.5264	0.9183
Semantic Majority	**0.6299**	**0.5486**	**0.8369**
Single Run	0.5837	0.4733	0.8775
Semantic Union	0.4811	0.3507	0.9258

Because all reasoning methods operate on the same VisCls-generated candidate set, these results isolate the effect of the candidate generation. Comparisons across additional visual backbones and VLM variants are reported in Appendix C and exhibit the same overall trend.

4.3 Candidate-Conditioned Verification

Overall accuracy combines two distinct sources of error: (a) whether the correct disease is proposed by the visual classifier and (b) whether the reasoning system selects it once proposed. To isolate the second factor, we evaluate all methods on the subset of test cases for which the GT disease appears in the VisCls Top-5 candidate set.

Within this subset, the VisCls Top-1 accuracy is 59.18%, whereas RareVA achieves 79.59% accuracy. The paired improvement is 20.41 pp, with a 95% bootstrap CI of $[6.12, 34.69]$, excluding zero. The strongest global VLM baseline with multi-round KG retrieval reaches 67.35%, remaining 12.24 pp below RareVA (Figure 4). To view the full results, including full baselines and component ablation comparisons, see Table I.2.

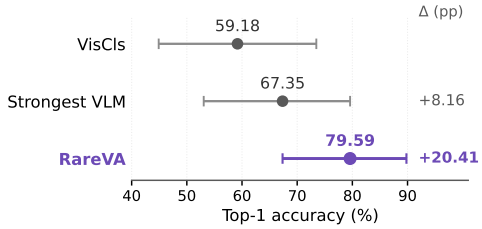


Figure 4: Performance on the GT disease in VisCls Top-5 candidate set.

These results provide a more direct evaluation of the proposed reasoning system. Once the correct disease is available for consideration, RareVA resolves more cases than the strongest VLM adjudication in this subset. The result supports candidate-conditioned verification, rather than unrestricted global reasoning or phenotype matching, as the primary source of downstream improvement.

4.4 Generalization to External Cohort

We next evaluate whether the improvement transfers beyond the RDFace development benchmark. Following the protocol described in Appendix B, we first retain only GMDB samples whose disease labels exactly match diseases present in RDFace. We then remove the overlapping patients identified through manual inspection of the original literature sources to avoid cross-dataset contamination.

As shown in Table 2, direct visual inference achieves 7.56% Top-1 accuracy. Applying filtered TENT test-time adaptation Niu et al. (2022) improves accuracy to 9.98%. The strongest VLM-based method achieves 10.74% while RareVA reaches 12.25%, corresponding to an absolute improvement of 4.69 pp over direct inference. More detailed baseline comparisons are provied in Appendix C.3.

Although absolute performance is lower than on RDFace, the direction of improvement is consistent across the two independently datasets. This suggests that RareVA relies less on dataset-specific visual statistics and instead exploits disease-specific phenotype evidence that transfers across cohorts.

Table 4: Direct paired comparison with the strongest VLM baseline. Δ denotes the paired Top-1 Acc. difference (RareVA − strongest VLM).

Setting	Strongest VLM	RareVA	Δ (pp)	RareVA-only / VLM-only	McNemar p
RDFace (all, $n = 113$)	29.20	34.51	+5.31 [0.89, 10.62]	7 / 1	0.070
GT ∈ Top-5 ($n = 49$)	67.35	79.59	+12.24 [2.04, 24.49]	7 / 1	0.070
GMDB external ($n = 661$)	10.74	12.25	+1.51 [0.61, 2.57]	11 / 1	0.006

4.5 Paired Comparison with the Strongest VLM

To assess whether RareVA improves beyond global VLM reasoning on the same cases, we perform paired comparisons against the strongest VLM baseline (Table 4). RareVA improves Top-1 accuracy

by 5.31 pp on RDFace and by 12.24 pp within the GT-in-Top-5 subset, with positive paired-bootstrap CIs in both settings. On the external GMDB cohort, RareVA also improves accuracy by 1.51 pp and corrects 11 cases missed by the strongest VLM while losing only one case, yielding a significant McNemar exact test ($p = 0.006$). Together, these results indicate that candidate-conditioned verification provides consistent case-level gains over direct global VLM reasoning.

4.6 Reliability of the Phenotype Representation

We evaluate phenotype extraction on 261 RDFace images for which patient-specific reference phenotypes can be recovered from the corresponding clinical reports. As shown in Table 3, Qwen3-VL QwenTeam (2025) achieves high recall while maintaining moderate precision using semantic-majority aggregation, indicating that the extracted phenotype representation captures most reported phenotypes at the cost of introducing false-positive evidence. Semantic-majority provides the most favorable precision–recall trade-off among the evaluated consensus strategies. Additional comparisons such as different consensus mechanisms and VLM extractors are in Appendix D.2 and D.3.

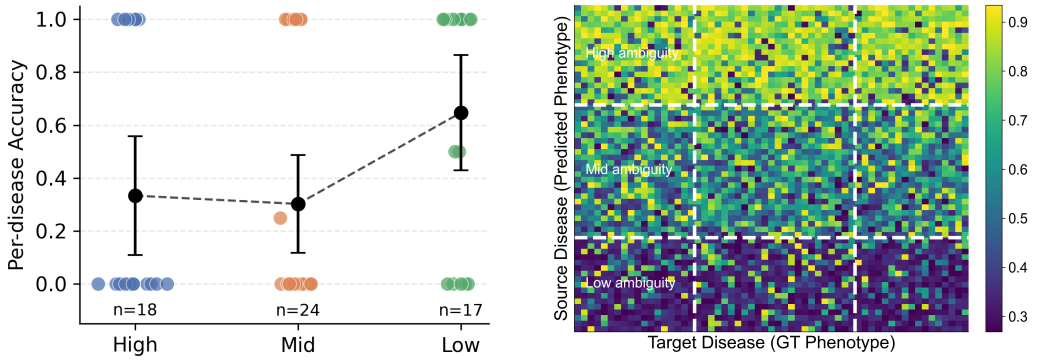


Figure 5: **Structured ambiguity in rare disease phenotype.** Left: Per-disease accuracy by ambiguity group. Right: Cross-disease phenotype alignment matrix sorted by ambiguity.

Exploratory disease ambiguity analysis. We further examine whether Top-1 accuracy varies with cross-disease phenotype similarity. Diseases are grouped according to their maximum alignment with another disease profile. Accuracy decreases from 69.23% in the lowest-similarity group to 26.67% in the highest-similarity group (Figure 5). This association is consistent with greater prediction difficulty among phenotypically overlapping diseases. Although it does not by itself separate the effects of extraction error, candidate recall, and class frequency.

4.7 Accuracy–Cost Optimization Analysis

We evaluate ACS against fixed candidate budgets $K \in \{2, 3, 4, 5\}$ under the same downstream reasoning pipeline. Fixed-budget performance is non-monotonic, increasing from 30.09% at $K = 2$ to 35.40% at $K = 4$ before dropping to 28.32% at $K = 5$, suggesting that evaluating additional candidates may introduce confusable hypotheses rather than improve performance.

As shown in Figure 6, ACS reduces the average output tokens from 10,064 to 8,338 (17.15%) relative to the best-performing fixed-budget policy ($K = 4$), while achieving a comparable observed Top-1 accuracy (34.51% vs. 35.40%). Table E.2 further shows that ACS selects intermediate budgets for most cases and reserves larger budgets for more challenging instances. Consistent with this observation, the paired bootstrap 95% CI for the accuracy difference between ACS and $K = 4$ spans zero ($[-6.19, 4.42]$ pp), and McNemar's exact test detects no statistically significant difference ($p = 1.0$). These results indicate that ACS provides an adaptive accuracy–cost trade-off by allocating verification budgets according to case difficulty, reducing computational cost while maintaining a competitive operating point relative to fixed-budget inference.

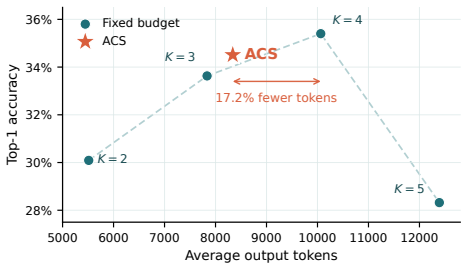


Figure 6: Accuracy–cost trade-off of ACS.

Table 5: Ablation of RareVA components under the candidate-bounded protocol.

Method	Top-1 Acc. (%)	Top-1 Acc. CI	Δ (pp)	Δ CI
Perception Prior				
VisCls (DenseNet)	25.66	[17.67, 33.63]	—	—
Multi-Agent Variants				
w/o KG-RAG	30.97	[22.12, 39.82]	+ 5.31	[-0.88, 11.50]
w/o ACS (fixed budget K)				
$K = 2$	30.09	[22.10, 38.05]	+ 4.43	[-0.88, 10.62]
$K = 3$	33.63	[24.78, 42.48]	+ 7.97	[2.65, 13.27]
$K = 4$	35.40	[26.55, 44.25]	+ 9.74	[4.42, 15.93]
$K = 5$ (same as w/o Planner Agent)	28.32	[20.35, 37.17]	+ 2.66	[-2.65, 7.96]
w/o multi-round Doctor Agents	30.97	[23.01, 39.82]	+ 5.31	[0.00, 11.50]
w/o hierarchical Doctor Agents	29.20	[21.24, 37.17]	+ 3.54	[-1.77, 8.85]
w/o Judge Agents	29.20	[21.24, 38.05]	+ 3.54	[-1.77, 8.85]
Full Model –RareVA (Ours)	**34.51**	**[25.66, 43.36]**	**+ 8.85**	**[2.65, 15.93]**

4.8 Ablation Study Analysis

Table 5 evaluates the contribution of each major component under the shared candidate-bounded protocol. Removing KG retrieval reduces Top-1 accuracy from 34.51% to 30.97%, indicating that disease-specific evidence provides complementary information beyond the image and extracted phenotype set. Replacing multi-round Doctor reasoning with a single-round procedure yields the same performance (30.97%), suggesting that iterative evidence accumulation is important for resolving diagnostically ambiguous cases. Removing either the hierarchical Doctor decomposition or the Judge Agent further decreases accuracy to 29.20%, highlighting the importance of both structured disease-specific verification and cross-candidate evidence aggregation. Overall, these results suggest that RareVA benefits from progressively combining complementary sources of evidence throughout the candidate-conditioned verification pipeline. Appendix I shows additional uncertainty analyses.

We additionally compare ACS with a VisCls-entropy heuristic under the same pipeline. The heuristic uses fewer candidates and tokens on average (7,154 tokens) but achieves only 29.20% accuracy, versus 34.51% for ACS. This suggests that a fixed uncertainty threshold may under-allocate reasoning when additional candidate verification is beneficial. Overall, the components provide complementary rather than redundant functionality within the candidate-bounded verification pipeline.

4.9 Qualitative Traceable Reasoning Analysis

Appendix J provides representative success and failure traces, including candidate generation, phenotype extraction, ACS selection, Doctor verification, KG evidence, and Judge comparison. Successful cases show how RareVA rejects visually plausible but phenotype-inconsistent candidates, while failures mainly arise from (a) exclusion of the GT disease from the initial candidate set, (b) misleading extracted phenotypes, or (c) insufficient disease-specific KG coverage.

5 Conclusion

We presented RareVA, an image-first agentic system for rare disease candidate prioritization that formulates reasoning as candidate-conditioned verification. By combining image-derived phenotype extraction, disease-specific knowledge retrieval, and multi-agent verification, RareVA enables explicit evidence-grounded comparison among competing disease candidates and improves Top-1 accuracy on both RDFace and the external GMDB cohort. ACS further provides an adaptive accuracy–cost trade-off by allocating verification effort on a case-specific basis. Our results support candidate-conditioned verification as a practical alternative to direct visual prediction or global VLM adjudication under severe data scarcity. RareVA remains dependent on the quality of its upstream visual and phenotype representations, motivating future work on more tightly integrated perception and verification as well as broader evaluation across heterogeneous rare disease settings. Overall, RareVA demonstrates that explicitly verifying competing disease candidates can provide a more structured and efficient alternative to direct prediction in data-scarce image-first settings.

AI USE STATEMENT

Generative AI models were used as components of the proposed system and experimental baselines, including the VLMs and LLM backends described in the Methods and Appendix. Separately, generative AI tools were used for brainstorming and language editing, which both fall under the recommended disclosure categories. We have verified and tested all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This study involves human participant data, including patient facial images and associated phenotype information. Ethical approval was obtained from the relevant institutional Health Sciences Research Ethics Board prior to study initiation (**Reference Number: REDACTED for double-blind review**), and data usage permissions were granted for RDFace and GMDB dataset used in this study. Our use of the dataset complies with all access and usage policies, including restrictions on redistribution and re-identification. All experiments were conducted in accordance with these guidelines.

REPRODUCIBILITY STATEMENT

We provide the complete training objectives, implementation details, dataset processing procedures, evaluation protocols, ablation settings, and hyperparameters in the Appendix M Table M.1, M.2, M.3, and M.4. We will release training code, preprocessing scripts, and configuration files upon acceptance.

REFERENCES

Matilda Anderson, Elizabeth J. Elliott, and Yvonne A. Zurynski. Australian families living with rare disease: experiences of diagnosis, health services use and needs for psychosocial support. *Orphanet Journal of Rare Diseases*, 8:22, 2013. doi: 10.1186/1750-1172-8-22.

Gareth Baynam, Nicholas Pachter, Fiona McKenzie, Sharon Townshend, Jennie Slee, Cathy Kiraly-Borri, Anand Vasudevan, Anne Hawkins, Stephanie Broley, Lyn Schofield, Hedwig Verhoef, Caroline E. Walker, Caron Molster, Jenefer M. Blackwell, Sarra Jamieson, Dave Tang, Timo Lassmann, Kym Mina, John Beilby, Mark Davis, Nigel Laing, Lesley Murphy, Tarun Weeramanthri, Hugh Dawkins, and Jack Goldblatt. The rare and undiagnosed diseases diagnostic service – application of massively parallel sequencing in a state-wide clinical service. *Orphanet Journal of Rare Diseases*, 11:77, 2016. doi: 10.1186/s13023-016-0462-7.

Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pp. 67–74. IEEE Press, 2018. doi: 10.1109/FG.2018.00020.

Raquel Castro, Juliette Senecat, Myriam de Chalendar, Ildikó Vajda, Dorica Dan, and Béata Boncz. *Bridging the Gap between Health and Social Care for Rare Diseases: Key Issues and Innovative Solutions*, pp. 605–627. Springer International Publishing, Cham, 2017. ISBN 978-3-319-67144-4. doi: 10.1007/978-3-319-67144-4_32.

Xuanzhong Chen, Ye Jin, Xiaohao Mao, Lun Wang, Shuyang Zhang, and Ting Chen. Rareagents: autonomous multi-disciplinary team for rare disease diagnosis and treatment. In *Proceedings of the Fortieth AAAI Conference on Artificial Intelligence and Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence and Sixteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'26/IAAI'26/EAAI'26. AAAI Press, 2026. ISBN 978-1-57735-906-7. doi: 10.1609/aaai.v40i1.36969. URL https://doi.org/10.1609/aaai.v40i1.36969.

Zijie Cheng, Ariel Yuhan Ong, Siegfried K. Wagner, David A. Merle, Lie Ju, Hanyuan Zhang, Ruinian Chen, Linze Pang, Boxuan Li, Tiantian He, Anran Ran, Hongyang Jiang, Dawei Gabriel

YANG, Ke Zou, Jocelyn Hui Lin Goh, Sahana Srinivasan, Andre Altmann, Daniel C. Alexander, Carol Y. Cheung, Yih Chung Tham, Pearse A. Keane, and Yukun Zhou. Understanding the robustness of vision-language models to medical image artefacts. *npj Digital Medicine*, 8(1):727, 2025. doi: 10.1038/s41746-025-02108-w.

Ganlin Feng, Yuxi Long, Hafsa Ali, Erin Lou, Fahad Butt, Qian Liu, Yang Wang, and Pingzhao Hu. Rdface: A benchmark dataset for rare disease facial image analysis under extreme data scarcity and phenotype-aware synthetic generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 42976–42986, June 2026.

Yanjun Gao, Ruizhe Li, Emma Croxford, Samuel Tesch, Daniel To, John Caskey, Brian W. Patterson, Matthew M. Churpek, Timothy Miller, Dmitriy Dligach, and Majid Afshar. Large language models and medical knowledge grounding for diagnosis prediction. *medRxiv*, 2023. doi: 10.1101/2023.11.24.23298641.

Michael A Gargano, Nicolas Matentzoglu, Ben Coleman, Eunice B Addo-Lartey, Anna V Anagnostopoulos, Joel Anderton, Paul Avillach, Anita M Bagley, Eduard Bakštein, James P Balhoff, Gareth Baynam, Susan M Bello, Michael Berk, Holli Bertram, Somer Bishop, Hannah Blau, David F Bodenstein, Pablo Botas, Kaan Boztug, Jolana Čady, Tiffany J Callahan, Rhiannon Cameron, Seth J Carbon, Francisco Castellanos, J Harry Caufield, Lauren E Chan, Christopher G Chute, Jaime Cruz-Rojo, Noémi Dahan-Oliel, Jon R Davids, Maud de Dieuleveult, Vinicius de Souza, Bert B A de Vries, Esther de Vries, J Raymond DePaulo, Beata Derfalvi, Ferdinand Dhombres, Claudia Diaz-Byrd, Alexander J M Dingemans, Bruno Donadille, Michael Duyzend, Reem Elfeky, Shahim Essaid, Carolina Fabrizzi, Giovanna Fico, Helen V Firth, Yun Freudenberg-Hua, Janice M Fullerton, Davera L Gabriel, Kimberly Gilmour, Jessica Giordano, Fernando S Goes, Rachel Gore Moses, Ian Green, Matthias Griese, Tudor Groza, Weihong Gu, Julia Guthrie, Benjamin Gyori, Ada Hamosh, Marc Hanauer, Kateřina Hanušová, Yongqun (Oliver) He, Harshad Hegde, Ingo Helbig, Kateřina Holasová, Charles Tapley Hoyt, Shangzhi Huang, Eric Hurwitz, Julius O B Jacobsen, Xiaofeng Jiang, Lisa Joseph, Kamyar Keramatian, Bryan King, Katrin Knoflach, David A Koolen, Megan L Kraus, Carlo Kroll, Maaike Kusters, Markus S Ladewig, David Lagorce, Meng-Chuan Lai, Pablo Lapunzina, Bryan Laraway, David Lewis-Smith, Xiarong Li, Caterina Lucano, Marzieh Majd, Mary L Marazita, Victor Martinez-Glez, Toby H McHenry, Melvin G McInnis, Julie A McMurry, Michaela Mihulová, Caitlin E Millett, Philip B Mitchell, Veronika Moslerová, Kenji Narutomi, Shahrzad Nematollahi, Julian Nevado, Andrew A Nierenberg, Nikola Novák Čajbiková, Jr. Nurnberger, John I, Soichi Ogishima, Daniel Olson, Abigail Ortiz, Harry Pachajoa, Guiomar Perez de Nanclares, Amy Peters, Tim Putman, Christina K Rapp, Ana Rath, Justin Reese, Lauren Rekerle, Angharad M Roberts, Suzy Roy, Stephan J Sanders, Catharina Schuetz, Eva C Schulte, Thomas G Schulze, Martin Schwarz, Katie Scott, Dominik Seelow, Berthold Seitz, et al. The human phenotype ontology in 2024: phenotypes around the world. *Nucleic Acids Research*, 52(D1):D1333–D1346, 11 2023. ISSN 0305-1048. doi: 10.1093/nar/gkad1005.

GeminiTeam. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025.

Danila Germanese, Sara Colantonio, Marco Del Coco, Pierluigi Carcagnì, and Marco Leo. Computer vision tasks for ambient intelligence in children's health. *Information*, 14(10), 2023. ISSN 2078-2489. doi: 10.3390/info14100548.

Jindong Gu, Zhen Han, Shuo Chen, Ahmad Beirami, Bailan He, Gengyuan Zhang, Ruotong Liao, Yao Qin, Volker Tresp, and Philip Torr. A systematic survey of prompt engineering on vision-language foundation models, 2023.

Yaron Gurovich, Yair Hanani, Omri Bar, Guy Nadav, Nicole Fleischer, Dekel Gelbman, Lina Basel-Salmon, Peter M Krawitz, Susanne B Kamphausen, Martin Zenker, Lynne M Bird, and Karen W Gripp. Identifying facial phenotypes of genetic disorders using deep learning. *Nature medicine*, 25(1):60—64, January 2019. ISSN 1078-8956. doi: 10.1038/s41591-018-0279-0.

Ameer Hamza, Abdullah, Yong Hyun Ahn, Sungyoung Lee, and Seong Tae Kim. Llava needs more knowledge: retrieval augmented natural language generation with knowledge graph for explaining thoracic pathologies. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial*

Intelligence, AAAI'25. AAAI Press, 2025. ISBN 978-1-57735-897-8. doi: 10.1609/aaai.v39i3.32342.

Iryna Hartsock and Ghulam Rasool. Vision-language models for medical report generation and visual question answering: a review. *Frontiers in Artificial Intelligence*, Volume 7 - 2024, 2024. ISSN 2624-8212. doi: 10.3389/frai.2024.1430984.

Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, June 2016.

Tzung-Chien Hsieh and Peter M. Krawitz. Computational facial analysis for rare mendelian disorders. *American Journal of Medical Genetics Part C: Seminars in Medical Genetics*, 193(3): e32061, 2023. doi: https://doi.org/10.1002/ajmg.c.32061.

Tzung-Chien Hsieh, Aviram Bar-Haim, Shahida Moosa, Nadja Ehmke, Karen W Gripp, Jean Tori Pantel, Magdalena Danyel, Martin Atta Mensah, Denise Horn, Stanislav Rosnev, Nicole Fleischer, Guilherme Bonini, Alexander Hustinx, Alexander Schmid, Alexej Knaus, Behnam Javanmardi, Hannah Klinkhammer, Hellen Lesmann, Sugirthan Sivalingam, Tom Kamphans, Wolfgang Meiswinkel, Frédéric Ebstein, Elke Krüger, Sébastien Küry, Stéphane Bézieau, Axel Schmidt, Sophia Peters, Hartmut Engels, Elisabeth Mangold, Martina Kreiß, Kirsten Cremer, Claudia Perne, Regina C Betz, Tim Bender, Kathrin Grundmann-Hauser, Tobias B Haack, Matias Wagner, Theresa Brunet, Heidi Beate Bentzen, Luisa Averdunk, Kimberly Christine Coetzer, Gholson J Lyon, Malte Spielmann, Christian P Schaaf, Stefan Mundlos, Markus M Nöthen, and Peter M Krawitz. Gestaltmatcher facilitates rare disease matching using facial phenotype descriptors. *Nature genetics*, 54(3):349—357, March 2022. ISSN 1061-4036. doi: 10.1038/s41588-021-01010-x.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2261–2269, 2017. doi: 10.1109/CVPR.2017.243.

Yizhang Jin, Jian Li, Tianjun Gu, Yexin Liu, Bo Zhao, Jinxiang Lai, Zhenye Gan, Yabiao Wang, Chengjie Wang, Xin Tan, and Lizhuang Ma. Efficient multimodal large language models: a survey. *Visual Intelligence*, 3(1):27, 2025. doi: 10.1007/s44267-025-00099-6.

Philipp Koehn. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pp. 388–395, Barcelona, Spain, July 2004. Association for Computational Linguistics.

Peter Kováč, Peter Jackuliak, Alexandra Bražinová, Ivan Varga, Michal Aláč, Martin Smatana, Dušan Lovich, and Andrej Thurzo. Artificial intelligence-driven facial image analysis for the early detection of rare diseases: Legal, ethical, forensic, and cybersecurity considerations. *AI*, 5 (3):990–1010, 2024. ISSN 2673-2688. doi: 10.3390/ai5030049.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 09 2019. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz682. URL https://doi.org/10.1093/bioinformatics/btz682.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day, 2023.

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*, 2022.

Cong Liu, Fabricio Sampaio Peres Kury, Ziran Li, Casey Ta, Kai Wang, and Chunhua Weng. Doc2hpo: a web application for efficient and accurate hpo concept curation. *Nucleic Acids Research*, 47(W1):W566–W570, 05 2019. ISSN 0305-1048. doi: 10.1093/nar/gkz386.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10012–10022, October 2021.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Cyril Zakka, Yash Dalmia, Eduardo Pontes Reis, Pranav Rajpurkar, and Jure Leskovec. Med-flamingo: A multimodal medical few-shot learner, July 2023.

Dhanya Lakshmi Narayanan, Prajnya Ranganath, Shagun Aggarwal, Ashwin B. Dalal, Shubha R. Phadke, and Kaushik Mandal. Computer-aided facial analysis in diagnosing dysmorphic syndromes in indian children. *Indian Pediatrics*, 56:1017 – 1019, 2019. doi: 10.1007/s13312-019-1682-4.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *International conference on machine learning*, pp. 16888–16905. PMLR, 2022.

Urtė Oniūnaitė, Ozge Yilmaz, Rasa Ugenskiene, and Lina Jankauskaite. Editorial: Rare diseases: from basic science to clinical practice and public health. *Frontiers in Pediatrics*, Volume 13 - 2025, 2025. ISSN 2296-2360. doi: 10.3389/fped.2025.1579709.

Kascia Pavisich, Hannah Jones, and Gareth Baynam. The diagnostic odyssey for children living with a rare disease – caregiver and patient perspectives: A narrative review with recommendations. *Rare*, 2:100022, 2024. ISSN 2950-0087. doi: https://doi.org/10.1016/j.rare.2024.100022.

Jiaqi Qiang, Danning Wu, Hanze Du, Huijuan Zhu, Shi Chen, and Hui Pan. Review on facial-recognition-based applications in disease diagnosis. *Bioengineering*, 9(7), 2022. ISSN 2306-5354. doi: 10.3390/bioengineering9070273.

QwenTeam. Qwen3-vl technical report, 2025.

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 *Proceedings of the 38th International Conference on Machine Learning, ICML 2021*, volume 139, pp. 8748–8763. PMLR, 2021.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22500–22510, 2023. doi: 10.1109/CVPR52729.2023.02155.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 815–823, June 2015.

Andrew Sellergren, Chufan Gao, Fereshteh Mahvar, Timo Kohlberger, Fayaz Jamil, Madeleine Traverse, Alberto Tono, Bashir Sadjad, Lin Yang, et al. Medgemma 1.5 technical report, 2026.

Fayroz F. Sherif, Nahed Tawfik, Doaa Mousa, Mohamed S. Abdallah, and Young-Im Cho. Automated multi-class facial syndrome classification using transfer learning techniques. *Bioengineering*, 11(8), 2024. ISSN 2306-5354. doi: 10.3390/bioengineering11080827.

Jie Song, Mengqiao He, Xin Zheng, Yuxin Zhang, Cheng Bi, Jinhua Feng, Jiale Du, Hang Li, and Bairong Shen. Face-based machine learning diagnostics: applications, challenges and opportunities. *Artificial Intelligence Review*, 58(243), 2025a.

Jie Song, Zhichuan Xu, Mengqiao He, Jinhua Feng, and Bairong Shen. Graph retrieval augmented large language models for facial phenotype associated rare genetic disease. *npj Digital Medicine*, 8(543), 2025b.

Surat Teerapittayanon, Bradley McDanel, and H.T. Kung. Branchynet: Fast inference via early exiting from deep neural networks. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pp. 2464–2469, 2016. doi: 10.1109/ICPR.2016.7900006.

Bas H.M. van der Velden, Hugo J. Kuijf, Kenneth G.A. Gilhuijs, and Max A. Viergever. Explainable artificial intelligence (xai) in deep learning-based medical image analysis. *Medical Image Analysis*, 79:102470, 2022. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2022.102470.

Da Wu, Jingye Yang, Cong Liu, Tzung-Chien Hsieh, Elaine Marchi, Justin Blair, Peter Krawitz, Chunhua Weng, Wendy Chung, Gholson J Lyon, et al. Gestaltmml: Enhancing rare genetic disease diagnosis through multimodal machine learning combining facial images and clinical texts, 2024.

Tao Yang, Dandan Huang, Yunting Lin, Pengfei Wu, Zhikun Wu, Gangyuan Ma, Yulan Lu, Xinran Dong, Dingpeng Li, Junshuang Ge, et al. A specialized large language model for clinical reasoning and diagnosis in rare diseases, 2025.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

Jiarui Zhang, Mahyar Khayatkhoei, Prateek Chhikara, and Filip Ilievski. Mllms know where to look: Training-free perception of small visual details with multimodal llms, 02 2025.

Weike Zhao, Chaoyi Wu, Yanjie Fan, Xiaoman Zhang, Pengcheng Qiu, Yuze Sun, Xiao Zhou, Yanfeng Wang, Ya Zhang, Yongguo Yu, Kun Sun, and Weidi Xie. An agentic system for rare disease diagnosis with traceable reasoning. *Nature*, 651(775-784), 2026.

Kaiwen Zuo, Yirui Jiang, Fan Mo, and Pietro Lio. Kg4diagnosis: A hierarchical multi-agent llm framework with knowledge graph enhancement for medical diagnosis. In *AAAI Bridge Program on AI for Medicine and Healthcare*, pp. 195–204. PMLR, 2025.

Yvonne Zurynski, Aranzazu Gonzalez, Marie Deverell, Amy Phu, Helen Leonard, John Christodoulou, and Elizabeth Elliott. Rare disease: a national survey of paediatricians' experiences and needs. *BMJ Paediatrics Open*, 1(1):e000172, 2017. doi: 10.1136/bmjpo-2017-000172.

RareVA Supplementary Appendix

Contents

A Prompt Sets 2

B Dataset Details 9

C Additional Baseline Model Comparison and Details 10

D VLM Phenotype Extraction 11

E Planner Agent and ACS Optimization 14

F Knowledge Graph Construction and Querying 16

G Hierarchical Doctor Agent Multi-Round Protocol 17

H Judge Agent Details 20

I Uncertainty and Stability Quantification 20

J Qualitative Case Studies and Error Analysis 24

K Additional Analysis 29

L Limitations and Future Work 31

M Reproducibility Checklist 32

A Prompt Sets

Below, we show all the prompts we used in the phenotype extraction and the multi-agent system with structured JSON outputs and prompts used for baseline comparisons (VisCls + VLM and VisCls + Phenotypes + VLM).

A.1 Prompts for RareVA

Prompt 1. Phenotype Extraction

You are a clinical dysmorphology expert.

Your task:
– Extract facial phenotypes from one pediatric face photo.
– Aim for 6 – 10 phenotypes (do NOT exceed 10).

Evidence rules:
– For EACH phenotype, "evidence" must be a short image–grounded phrase (8 – 20 words),
mentioning a visible cue AND the facial region.
– Do NOT use placeholders like "Observed", "Present", "Noted".

Hard rules:
– Output MUST be a SINGLE valid JSON object.
– Extract ONLY phenotypes that are visually observable on the face.
– Do NOT output diagnoses, syndromes, genes, or non–visible attributes.
– Avoid near–duplicates. Keep the most standard HPO term.

Return ONLY JSON with this schema:

```
{
  "image_id": "{image_id}",
  "phenotypes": [
    {
      "term": "",
      "evidence": "",
    }
  ],
  "negative_findings": [],
  "notes": []
}
```

Prompt 2. Planner Agent

You are the MAIN ORCHESTRATOR in a multi–agent rare disease diagnosis system.

You receive FOUR inputs:
1) A patient facial image.
2) A patient–specific phenotype list extracted from this image (only patient–specific, no general syndrome text).
3) A preselection set (PS) of candidate diseases (with scores/metadata).
4) A fixed subset size K (provided by policy).

Your task:
– Select exactly K diseases from the preselection set to call disease–specific Doctor Agents.
– Output ONLY a single JSON object following the required schema. No extra text.

Hard rules:
– You MUST output valid JSON only.
– Select ONLY from the preselection candidates.
– You must NOT invent phenotypes or visual findings.
– If image evidence is unclear, say "uncertain" rather than hallucinating.
– Prefer candidates that are jointly supported by PS prior/metadata and are broadly compatible with the provided patient phenotype list (no external lookup).
– Use the image ONLY to sanity−check whether the PROVIDED phenotypes are visually plausible; do NOT perform free−form diagnosis from the image.

Tie−breaking:
– When similar, prefer candidates compatible with more salient patient phenotypes.

FIXED_K: {K}
PS_PRESELECT: {PS_PRESELECT_JSON}
PHENOTYPES: {PHENOTYPE_BULLETS}

Return ONLY JSON with this schema:
```
{
  "fixed_k": {K},
  "selected_candidates": [
    {"disease_id": "...", "reason_ps": "...", "reason_pheno": "...", "reason_image": "..."},
    ...
  ],
  "diagnosis_summary": ["...", "..."],
  "call_agents": ["DoctorAgent::<disease_id>", "..."],
}
```

Prompt 3. Doctor – Orchestration Agent

You are the DOCTOR AGENT (orchestrator). You supervise two sub−agents:
1) Retrieval Agent output (facial KG phenotypes for this disease)
2) Reasoning Agent output (match/no_match/uncertain + need_more_kg)

Your task:
– Decide whether to STOP now or CONTINUE to request broader KG tiers.
– Provide a final binary decision ONLY when you stop.
– Output ONLY JSON. No extra text.

Hard rules:
– You MUST output valid JSON only.
– "final_decision" MUST be either "match" or "no_match" WHEN action="stop".
– If need_more_kg="true", you MUST set action="continue" and final_decision to null.

Decision policy:
– Use Reasoning Agent's decision as primary signal.

DISEASE: {DISEASE_NAME}
ROUND: {ROUND_INDEX}
LAST_ROUND: {IS_LAST_ROUND}

PATIENT_PHENOTYPES: {PATIENT_PHENO_BULLETS}
RETRIEVAL_JSON: {RETRIEVAL_JSON}
REASONING_JSON: {REASONING_JSON}

Return ONLY JSON with this schema:
```
{
```

```
  "disease_name": "{DISEASE_NAME}",
  "action": "stop" | "continue",
  "final_decision": "match" | "no_match" | null,
  "stop_reason": "...",
  "rationale": ["...","..."]
}
```

Prompt 4. Doctor – Retrieval Agent

You are the RETRIEVAL AGENT for early rare disease diagnosis.

Your task:
– From KG_CANDIDATES_JSON, select ONLY craniofacial / facial−visible phenotypes: (head, face, forehead, eyes, eyebrows, nose, mouth, lips, philtrum, ears, jaw, chin, hairline, skin/hair of head/face).
– Remove non−facial/systemic phenotypes (height, neck, limbs, heart, kidney, neurodevelopment, seizures, etc.).
– Prefer phenotypes that are specific.

Hard rules:
– Output ONLY valid JSON.

DISEASE: {DISEASE_NAME}
KG_CANDIDATES_JSON: {KG_CANDIDATES_JSON}

Return ONLY JSON with this schema:

```
{
  "disease_name": "{DISEASE_NAME}",
  "selected_facial_terms": [
    {
      "term": "...",
      "frequency": "...",
    }
  ],
  "notes": "..."
}
```

Prompt 5. Doctor – Reasoning Agent

You are the REASONING AGENT for a disease−specific Doctor Agent. You are an expert in {DISEASE_NAME}, mastering at analyzing phenotypes and diagnosis of this disease.

Inputs:
1) Disease Name
2) Patient phenotype bullets (patient−specific, already extracted; do NOT invent new phenotypes).
3) Disease facial phenotype candidates (from retrieval agent).

Your task:
– Based on the two list of phenotypes decide if the disease is a MATCH or NO_MATCH for this patient.
– If phenotypes are empty or evidence is insufficient, MUST request more KG phenotypes by setting need_more_kg=true.

Hard rules:
– Output ONLY valid JSON.

– "rationale" MUST contain at least 2 non–empty items.
– If evidence is insufficient, rationale must explicitly say so.
– Do NOT introduce new phenotype terms not present in PATIENT_PHENOTYPES into matched_terms/missing_key_terms/conflicts.

Decision policy:
– Output "no_match" ONLY if there is strong negative evidence (clear conflicts) or the patient's phenotypes strongly contradict the disease facial profile.
– If overlap is low OR evidence is insufficient/uncertain (e.g., missing key facial terms, patient phenotypes are sparse, or terms may be synonyms), output "uncertain" and set need_more_kg=true.
– Do NOT require exact string match; allow synonym / near–equivalent matches (e.g., "prominent forehead" ~ "broad forehead") BUT only if it is plausible.

DISEASE: {DISEASE_NAME}
PATIENT_PHENOTYPES: {PATIENT_PHENO_BULLETS}
DISEASE_FACIAL_CANDIDATES_JSON: {DISEASE_FACIAL_JSON}

Return ONLY JSON with this schema:
{
  "disease_name": "{DISEASE_NAME}",
  "decision": "match" | "no_match" | "uncertain",
  "matched_rate": float in range of [0.0–1.0],
  "matched_terms": ["..."],
  "missing_key_terms": ["..."],
  "conflicts": ["..."],
  "need_more_kg": false,
  "rationale": ["...","..."]
}

Prompt 6. Doctor – Fallback Self-decide Prompt

You are the DOCTOR AGENT expert in {DISEASE_NAME}.

Your task:
– Make a binary decision for {DISEASE_NAME} disease.

Context:
– The retrieval agent failed to extract ANY facial phenotypes from the KG after multiple frequency tiers. Therefore: DO NOT claim "KG supports X". KG facial evidence is missing.
– You MAY use: patient phenotype list + patient image + general medical knowledge.
– If evidence is weak, still choose the more plausible binary decision, but reflect uncertainty in rationale.

Hard rules:
– Output ONLY JSON. No extra text.
– Decide MATCH or NO_MATCH.

DISEASE: {DISEASE_NAME}
PATIENT_IMAGE: {PATIENT_IMAGE}
PATIENT_PHENOTYPES: {PATIENT_PHENO_BULLETS}

Return ONLY JSON with this schema:
{
  "disease_name": "{DISEASE_NAME}",
  "final_decision": "match" | "no_match",

```
  "decision_basis": ["patient_terms","image","knowledge_based"],
  "rationale": ["...","..."]
}
```

Prompt 7. Judge Agent [MATCH MODE]

```
You are a clinical Judge Agent expert in rare disease field.

Your task:
- Choose exactly ONE disease from the candidate list for the patient.

MODE: {MATCH}

Rules:
1) Consider ONLY candidates with 'match'.
2) Score each eligible candidate (0-1 each):
   A. Alignment: explains salient facial findings in IMAGE.
   B. Specificity: case-specific (not generic).
   C. Consistency: no contradiction to IMAGE.
3) Pick highest total. Tie-break: higher C, then A, then B.
4) Override a Doctor "match" ONLY if rationale clearly contradicts visible evidence in
     IMAGE; if override, state in rationale.

Hard rules:
- Output ONLY valid JSON. No extra text.

IMAGE: {IMAGE_NAME}
CANDIDATES_LIST: {CANDIDATES_BLOCK}
DOCTOR_OUTPUT: {DOCTOR_OUTPUT_JSON}
PHENOTYPES: {PHENOTYPE_BULLETS}

Return ONLY JSON with this schema:
{
  "image_name": "{image_name}",
  "chosen_disease": "... (must be EXACTLY one from the list)",
  "score": average number in range of [0.0-1.0],
  "rationale": "...",
}
```

Prompt 8. Judge Agent [NO_MATCH MODE]

```
You are a clinical Judge Agent expert in rare disease field.

Your task:
- Choose exactly ONE disease from the candidate list for the patient.

MODE: {NO_MATCH}

Rules:
1) Consider ALL candidates in CANDIDATE_LIST, but do NOT do free-form diagnosis
     from IMAGE.
2) First do a conservative IMAGE-phenotype check on PHENOTYPES:
   - Split PHENOTYPES into SUPPORTED (clearly supported by IMAGE) vs
       UNCLEAR (unclear/not supported).
   - Do NOT add new phenotypes.
3) Score each candidate (0-1 each):
   A. Alignment: disease is consistent with SUPPORTED phenotypes.
```

B. Prior support: disease is reasonably supported by its prior position in CANDIDATE_LIST.
C. Consistency: no strong contradiction to salient IMAGE evidence / SUPPORTED phenotypes.
4) Pick highest total. Tie−break: higher C, then A, then B.
5) Output an uncertainty note grounded in: (i) unanimous no_match from Doctors, and (ii) limited/unclear phenotype support.

Hard rules:
− Output ONLY valid JSON.

IMAGE: {IMAGE_NAME}
CANDIDATE_LIST: {CANDIDATES_BLOCK}
DOCTOR_OUTPUT: {DOCTOR_OUTPUT_JSON}
PHENOTYPES: {PHENOTYPE_BULLETS}

Return ONLY JSON with this schema:
```
{
  "image_name": "{image_name}",
  "chosen_disease": "...",
  "score": average number in range of [0.0−1.0],
  "phenotype_check": { "supported": ["..."], "unclear": ["..."] },
  "rationale": "...",
  "uncertainty_note": "..."
}
```

A.2 Prompts for Baseline/Ablation

Prompt 9. VisCls + VLM

You are an expert in rare disease field.

Your task:
– Given a pediatric facial photograph and a candidate disease list, answer the most likely diagnosis.

Hard rules:
– Return EXACTLY one disease from the candidate list and rationale.
– No extra text, no bullet points, no numbering.
– Use ONLY information that can be inferred from the facial image.
– Do NOT assume access to genetic tests, clinical notes, or prior diagnoses.

PATIENT_IMAGE: {DISEASE_IMAGE}
CANDIDATE_LIST: {CANDIDATE_LIST}

Return ONLY JSON with this schema:
```
{
  "diagnosis": "",
  "rationale": ""
}
```

Prompt 10. VisCls + Phenotype + VLM

You are an expert in rare disease field.

Your task:
– Given a pediatric facial photograph, a phenotype list related to this patient, and a candidate disease list, answer the most likely diagnosis.

Hard rules:
– Return EXACTLY one disease from the candidate list and rationale.
– No extra text, no bullet points, no numbering.
– Use ONLY information that can be inferred from the facial image and phenotype list.
– Do NOT assume access to genetic tests, clinical notes, or prior diagnoses.

PATIENT_IMAGE: {DISEASE_IMAGE}
PATIENT_PHENO: {PATIENT_PHENO}
CANDIDATE_LIST: {CANDIDATE_LIST}

Return ONLY JSON with this schema:
```
{
  "diagnosis": "",
  "rationale": ""
}
```

B Dataset Details

B.1 RDFace Split Statistics

RDFace contains 456 facial images spanning 103 rare diseases. We follow the original RDFace benchmark protocol, which defines the data partition at the image level. The dataset is divided into training, validation, and test splits (65%, 10%, and 25%), with disease classes containing only a single image assigned to the training split to maximize the amount of supervision available for visual representation learning. The complete disease taxonomy and dataset construction procedure are described in the original RDFace paper. Here we summarize the statistics of each split in Table B.1.

Table B.1: RDFace dataset split statistics.

Split	# Images	# Diseases	GT Phenotypes Available
Train	297	103	Yes
Validation	46	35	No
Test	113	59	No

The original RDFace benchmark defines image-level partitions rather than explicit patient-identity partitions. We therefore performed an additional source-linked overlap audit using the subset of cases for which patient-specific phenotype descriptions could be recovered. All 261 such images belong to the training split, and no phenotype-linked patient overlap was identified between the training and test splits. We therefore treat the RDFace test set as held out with respect to the available source-linked patient annotations, while noting that the official benchmark split itself is defined at the image level.

B.2 GT Phenotype Annotations

RDFace is intentionally designed as an image-only benchmark and therefore does not include paired phenotype annotations. To evaluate the reliability of the phenotype extraction module, we additionally collected patient-specific phenotype descriptions for the source-linked training cases by revisiting the original source links associated with each image, including published case reports and clinical literature referenced in the RDFace metadata. Phenotype terms were manually extracted from the corresponding patient descriptions and serve exclusively as reference annotations for evaluating phenotype extraction accuracy.

To prevent any use of GT phenotypes during final prediction, these annotations are never provided to RareVA at inference time. They are not used to supervise the visual classifier, the VLM phenotype extractor, the ACS policy, or the downstream multi-agent reasoning modules. During end-to-end system operation, RareVA receives only the facial image as input and must infer phenotype evidence directly from visual appearance. The usage of each annotation type is summarized in Table B.2.

Table B.2: Usage of RDFace data and annotations in RareVA.

Data / Annotation	Training	Validation	Test-time Input
Facial images	Yes	Yes	Yes
Disease labels	Yes	Yes	Evaluation only
GT phenotypes	Evaluation only	No	No
VLM-extracted phenotypes	No	No	Yes
Disease–phenotype KG	Retrieval only	Retrieval only	Retrieval only

B.3 External Validation on GMDB

To evaluate the generalization of RareVA beyond RDFace, we additionally conduct external validation on GMDB, which contains 10,980 facial images spanning 581 genetic disorders. Since RDFace and GMDB adopt different disease taxonomies, we first perform disease-name matching between the two datasets. This results in 26 shared disease classes comprising 676 GMDB images. We then checked for overlap between RDFace and GMDB. Fifteen overlapping images were identified,

each corresponding to a distinct patient identity. Therefore, the 15 overlapping images represent 15 unique patients. All 15 were removed from the GMDB evaluation cohort before any external evaluation was performed, yielding the final external test set.

GMDB is used exclusively for evaluation and is never involved in training, model selection, or hyperparameter tuning. Following the protocol described in Section 4, all competing methods are evaluated under the same image-first verification setting using only the facial image as input, without access to patient-specific phenotype annotations.

C Additional Baseline Model Comparison and Details

C.1 Sensitivity to the VisCls Backbone

We evaluate the sensitivity of the candidate-generation stage to the choice of visual backbone. DenseNet was specified as the default VisCls backbone in the main RareVA pipeline prior to conducting this comparison. The experiments in this section were added subsequently as a sensitivity analysis and were not used for backbone selection or hyperparameter tuning.

Under a common supervised training protocol, we compare DenseNet with several alternative visual-classification backbones on RDFace and report Top-K accuracy for $K \in \{1, 5, 10, 30\}$. We additionally include a Gestalt-style architecture as a reference. The results characterize how candidate-generation performance varies across backbone choices rather than serving as a model-selection procedure.

Table C.1: Supervised classification results across different backbones.

Acc (%)	**Top-1**	**Top-5**	**Top-10**	**Top-30**
ResNet152 He et al. (2016)	10.62	31.86	44.25	63.72
DenseNet169 Huang et al. (2017)	**25.66**	**43.36**	**50.44**	**71.68**
FaceNet Schroff et al. (2015)	17.70	31.86	43.36	64.60
VGG16 Cao et al. (2018)	15.04	30.09	41.59	69.91
Swin Transformer-Tiny Liu et al. (2021)	18.58	29.20	43.36	68.18
CLIP (ViT-B/32) Radford et al. (2021)	17.70	33.63	40.71	60.18
Gestalt Gurovich et al. (2019); Hsieh et al. (2022)	13.27	26.55	37.17	63.71

C.2 Sensitivity to the VLM Backend

To assess whether RareVA's performance depends critically on a particular VLM backend, we perform a VLM sensitivity analysis with the VisCls candidate generator fixed to DenseNet. We substitute several VLMs into the same candidate-bounded reasoning protocol while holding the candidate set, prompts, and evaluation procedure fixed, and report Top-1 accuracy on the RDFace test split.

Gemini-2.5-Flash serves as the default backend throughout the main experiments and was fixed before this comparison was conducted. This analysis was added subsequently to assess sensitivity to the choice of VLM backend and was not used for model selection or hyperparameter tuning.

Table C.2: Comparison of different VLMs as baseline adjudicators under the candidate-bounded protocol. Δ denotes absolute change in pp for Top-1 accuracy relative to VisCls prior.

Setting	Top-1 Acc. (%)	Δ (pp)
llava-med-v1.5-mistral-7b Li et al. (2023)	22.12	- 3.54
Med-Flamingo-7B Moor et al. (2023)	22.12	- 3.54
Med-Gemma-1.5-4B Sellergren et al. (2026)	25.66	+ 0.00
Qwen3-VL-8B-Instruct QwenTeam (2025)	24.78	- 0.88
Gemini-2.5-Flash-Image GeminiTeam (2025)	**26.55**	+ 0.89

C.3 External Validation on GMDB

We report the full performance of RareVA on the external validation GMDB in Table C.3.

Table C.3: **Overall performance under the candidate-bounded protocol.** Top-1 accuracy with 95% bootstrap CI and the improvement in pp over the perception prior on GMDB.

Method	Top-1 Acc. (%)	Top-1 Acc. CI	Δ (pp)
Perception Prior – DenseNet	7.56	[5.60, 9.68]	—
TTA methods – filtered TENT	9.98	[7.72, 12.25]	+ 2.42
VLM Baselines			
Single VLM (Gemini)	7.87	[5.90, 9.98]	+ 0.31
Single VLM + Pheno	7.26	[5.30, 9.38]	− 0.30
Single VLM + KG-RAG	9.38	[7.26, 11.65]	+ 1.82
Single VLM + KG-RAG + multi-round	10.74	[8.62, 13.46]	+ 3.18
Our Method			
RareVA (full)	**12.25**	**[9.98, 15.13]**	**+ 4.69**

D VLM PHENOTYPE EXTRACTION

Although one could attempt image-only prediction using VLMs, we instead extract facial phenotypes as an explicit interface between perception and reasoning to improve modularity, interpretability, and efficiency. This structured intermediate representation enables the multi-agent system to ground verification and rationales in case-specific, image-observable traits rather than implicit image impressions, consistent with prior evidence that phenotype representations are useful for decision support and reasoning Wu et al. (2024); Song et al. (2025b).

D.1 SEMANTIC CONSENSUS DETAILS

To mitigate stochastic hallucinations inherent in VLMs, we perform N independent extraction runs, yielding phenotype term sets $\{P^{(n)}(x)\}_{n=1}^{N}$. To account for lexical variation across runs, we first apply text normalization and then aggregate terms via a *semantic majority* voting scheme. Let $\mathrm{sim}(\cdot,\cdot)$ denote cosine similarity between text embeddings, and define a matching predicate $\mathcal{M}(p,q) \in \{0,1\}$ that outputs 1 if $\mathrm{sim}(p,q) \geq \tau$. Under $\mathcal{M}$, terms across runs are grouped into semantic clusters $\{g\}$.

For each cluster g, the cumulative vote count is defined as:

$$v(g) = \sum_{n=1}^{N} \mathbb{I}\Big[\exists\, p \in P^{(n)}(x) \text{ s.t. } p \in g\Big]. \tag{6}$$

We retain clusters that meet a consensus threshold and form the consensus term set:

$$P_{\text{term}}(x) = \{\, \mathrm{repr}(g) \;:\; v(g) \geq \lceil N/2 \rceil \,\}, \tag{7}$$

where $\mathrm{repr}(g)$ is a representative term for cluster g (e.g., the cluster medoid).

Evaluation Protocol. We evaluate extracted phenotypes on a GT-annotated subset using exact HPO term matching. Let TP, FP, and FN denote true positives, false positives, and false negatives. We report Precision, Recall, and F1:

$$\mathrm{Precision} = \mathrm{TP}/(\mathrm{TP}+\mathrm{FP}), \quad \mathrm{Recall} = \mathrm{TP}/(\mathrm{TP}+\mathrm{FN}), \tag{8}$$

$$\mathrm{F1} = 2\,\mathrm{Precision} \times \mathrm{Recall}/(\mathrm{Precision}+\mathrm{Recall}), \tag{9}$$

as well as hallucination rate $(1-\mathrm{Precision})$ and missing rate $(1-\mathrm{Recall})$. Standard deviations are computed across samples.

D.2 Consensus Mechanism Variants

To aggregate multiple stochastic phenotype extraction outputs produced for the same image, we evaluate a set of consensus mechanisms operating at either the semantic or exact-string level.

Single-run baselines. `single_run1, 2, 3` each use the phenotype list from one individual stochastic run without any cross-run aggregation. These serve as minimal baselines to quantify variance induced by sampling.

Semantic consensus variants. `sem_majority` performs semantic deduplication across runs and retains only phenotype clusters supported by a majority of runs (i.e., > 2). `sem_union` constructs a semantic-level union of all extracted phenotypes across runs, maximizing recall at the cost of increased hallucinations.

Exact matching variants. `exact_all` aggregates phenotypes using exact string matching without semantic merging. `exact_union` further relaxes this constraint by taking the union over all exact strings across runs, yielding the highest recall but the largest number of false positives.

Selection-based variants. `lpa_selector` and `aws_selector` select a single run from multiple candidates using learned or heuristic scoring functions, respectively, and return only the chosen run's output. These methods trade off consensus robustness for simplicity by avoiding explicit aggregation.

Oracle upper bound. `best_of_3_oracle` retrospectively selects, for each sample, the single run that achieves the highest F1 score with respect to GT. This variant is not deployable and is reported solely as an upper bound on achievable performance under multiple sampling.

Metrics. After ontology normalization with Doc2HPO, we evaluate extracted phenotypes against GT phenotypes using exact matching. Let TP denote the number of predicted phenotypes whose HPO term names appear in the GT, FP the number of predicted phenotypes whose HPO term names do not appear in the GT, and FN the number of GT phenotypes whose HPO term names are missing from the prediction. Precision and recall are defined as

$$\mathrm{Precision} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FP}) \quad \text{and} \quad \mathrm{Recall} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN}). \tag{10}$$

The F1 score is the harmonic mean

$$\mathrm{F1} = 2\,\mathrm{Precision} \times \mathrm{Recall}/(\mathrm{Precision} + \mathrm{Recall}). \tag{11}$$

We additionally report $\mathrm{FP}/(\mathrm{TP} + \mathrm{FP})$ $[1 - \mathrm{Precision}]$ as hallucination rate and $\mathrm{FN}/(\mathrm{TP} + \mathrm{FN})$ $[1 - \mathrm{Recall}]$ as missing rate. Standard deviations (sd) are computed across test samples.

Table D.1: Consensus mechanisms for phenotype extraction. Values are reported as mean ± sd.

mode	F1	Precision	Recall	Halluc.	Missing
best_of_3_oracle	0.6376 ± 0.2203	0.5264 ± 0.2409	0.9183 ± 0.1859	0.4736 ± 0.2409	0.0817 ± 0.1859
sem_majority	$\mathbf{0.6299 \pm 0.2342}$	$\mathbf{0.5486 \pm 0.2560}$	$\mathbf{0.8369 \pm 0.2506}$	$\mathbf{0.4514 \pm 0.2560}$	$\mathbf{0.1631 \pm 0.2506}$
lpa_selector	0.5929 ± 0.2256	0.4857 ± 0.2352	0.8615 ± 0.2304	0.5143 ± 0.2352	0.1385 ± 0.2304
aws_selector	0.5881 ± 0.2293	0.4825 ± 0.2404	0.8568 ± 0.2401	0.5175 ± 0.2404	0.1432 ± 0.2401
single_run3	0.5837 ± 0.2189	0.4733 ± 0.2316	0.8775 ± 0.2196	0.5267 ± 0.2316	0.1225 ± 0.2196
single_run1	0.5795 ± 0.2287	0.4743 ± 0.2347	0.8502 ± 0.2454	0.5257 ± 0.2347	0.1498 ± 0.2454
single_run2	0.5723 ± 0.2310	0.4654 ± 0.2372	0.8537 ± 0.2478	0.5307 ± 0.2377	0.1463 ± 0.2478
exact_all	0.4881 ± 0.3520	0.5959 ± 0.4153	0.4637 ± 0.3703	0.2240 ± 0.3244	0.5363 ± 0.3703
sem_union	0.4811 ± 0.2048	0.3507 ± 0.1911	0.9226 ± 0.1838	0.6493 ± 0.1911	0.0774 ± 0.1838
exact_union	0.4099 ± 0.1896	0.2826 ± 0.1655	0.9258 ± 0.1803	0.7174 ± 0.1655	0.0742 ± 0.1803

Discussion. Complete quantitative results are summarized in Table D.1. We adopt `sem_majority` as our default consensus mechanism because it yields a robust precision–recall trade-off without relying on oracle selection and avoids the error amplification of overly aggressive aggregation. The standard deviations across metrics indicate substantial per-case variability in phenotype extraction, which we attribute to heterogeneous image difficulty and annotation sparsity/heterogeneity in long-tail diseases. Under this regime, consensus is beneficial primarily as a *stabilizer*: `sem_majority` reduces run-to-run stochasticity by suppressing idiosyncratic terms that appear in only a minority of runs, while still retaining consistently detected phenotypes. Finally, we believe the relatively high hallucination rates (measured after consensus) are consistent with a recall-oriented extraction behavior of VLMs. When uncertain, the model tends to propose

additional plausible facial phenotypes rather than omit them, prioritizing coverage at the cost of extra false positives.

D.3 VLM Robustness and Sensitivity Analysis

To assess the robustness of VLM-based phenotype extraction, we additionally run a second VLM (Gemini-2.5-Flash-Image) on the same image set under the same prompt and decoding configuration (3 runs per image; identical temperature/top-p/max tokens). We apply the same `sem_majority` consensus mechanism to aggregate the three runs for each model. We then evaluate robustness from two complementary perspectives: (i) *pre-normalization* and (ii) *post-normalization*.

Pre-normalization. We report the cross-model semantic agreement between the two consensus term sets using BioBERT embeddings and cosine similarity, which isolates extraction consistency from Doc2HPO. For semantic agreement, we embed each extracted term with BioBERT and compute a cosine similarity matrix between the two sets. We perform one-to-one maximum matching and treat a matched pair as consistent if its cosine similarity exceeds a threshold $\tau = 0.7$.

Post-normalization. We report the accuracy against GT annotation after Doc2HPO mapping (Qwen3-VL vs GT; Gemini-2.5-Flash-Image vs GT), using the same sample-level Precision/Recall/F1 and hallucination/missing metrics as in Table D.1.

Table D.2: Robustness analysis of VLM-based phenotype extraction before and after ontology normalization.

Pre-normalization: Cross-VLM Semantic Agreement (No Doc2HPO)				
Model Pair	F1	Precision	Recall	Mean Cosine
Qwen vs Gemini	0.7224 ± 0.1332	0.7992 ± 0.1536	0.6724 ± 0.1489	0.8046 ± 0.0930
Post-normalization: Accuracy vs GT (With Doc2HPO)				
Model	F1	Precision	Recall	Halluc.	Missing
Qwen	0.6299 ± 0.2342	0.5486 ± 0.2560	0.8369 ± 0.2506	0.4514 ± 0.2560	0.1631 ± 0.2506
Gemini	0.6179 ± 0.2191	0.5049 ± 0.2299	0.8978 ± 0.2049	0.4951 ± 0.2299	0.1022 ± 0.2049

Discussion. Table D.2 assesses how sensitive our phenotype-extraction stage is to the choice of VLM backbone, both before and after ontology normalization. In the *pre-normalization* term space (no Doc2HPO), Qwen3-VL and Gemini-2.5-Flash-Image exhibit substantial cross-model agreement (F1 $= 0.67 \pm 0.15$; cosine similarity $= 0.80 \pm 0.09$), indicating that the two backbones tend to surface largely overlapping facial findings under the same prompt and decoding configuration. After Doc2HPO mapping, both models achieve comparable *post-normalization* accuracy against GT (F1 $\approx$ 0.62–0.63), suggesting that the end-to-end phenotype set quality is not strongly backbone-dependent under our evaluation protocol. Notably, the two backbones reflect a consistent precision–recall trade-off: Gemini is more recall-oriented (higher recall and lower missing rate) while incurring slightly higher hallucination (false-positive) rates, whereas Qwen is relatively more conservative (higher precision). The relatively large per-image standard deviations are expected in this setting due to heterogeneous case difficulty and long-tail variability in the number and salience of visible facial cues. Overall, these results support the robustness of our perception-to-phenotype interface and motivate evaluating downstream agentic reasoning under a fixed consensus and normalization protocol rather than attributing gains to a particular VLM choice.

E Planner Agent and ACS Optimization

E.1 Planner Agent Details

Objective. The *Planner Agent* allocates an inference-time verification budget by selecting a case-specific subset of candidates for downstream verification, maximizing end-to-end utility under a cost budget.

Inputs and Outputs. Given an input image x, a pre-selected ranked candidate set $C(x)$, an extracted phenotype set $P(x)$, and a fixed budget K, the Planner outputs $D(x) \subseteq C(x)$ with $|D(x)| = K$, which is dispatched to disease-specific Doctor Agents. The Planner selects *only* from $C(x)$ and never introduces new diseases to keep the disease search space grounded in the perception stage. Selection is not a strict Top-K truncation; instead, the Planner performs candidate-conditioned routing within $C(x)$ by combining the PS prior/metadata with compatibility to $P(x)$ (Appendix A.1), without any external disease–phenotype retrieval/lookup at routing time.

E.2 Routing Features $\psi(x)$ Definition

Feature scope and leakage control. The ACS policy is trained and evaluated using features constructed exclusively from the visual perception stage prior to any multi-agent reasoning. Specifically, all routing features are derived from the initial VisCls-based image encoder and the VLM-based phenotype extraction outputs, including image-level embeddings, candidate set statistics, and phenotype-level summary features. No information from downstream agents (e.g., doctor, retrieval, or judge stages), intermediate reasoning traces, or final decisions is used in feature construction, ensuring that the routing policy operates without access to future signals and avoiding any form of information leakage. Specifically, the routing feature vector $\psi(x)$ for an input image x is constructed as follows:

- **Candidate Set Features:**
  - `candidates_count` denotes the number of disease candidates proposed by the router. This value is bounded by a predefined maximum candidate budget.
  - `candidates_confidence` summarizes the router's confidence over the proposed candidates, computed directly from the VisCls softmax outputs.
- **Phenotype-based Features:**
  - `patient_phenotype_embeddings` is a pooled embedding computed from the phenotype terms available at the routing stage, providing a semantic summary of the extracted patient phenotypes.
  - `patient_phenotype_count` is the number of phenotype terms available at routing time.
  - `patient_chars` measures the total character length of the phenotype descriptions and serves as a lightweight proxy for phenotype evidence volume.

Together, these five features provide a minimal yet informative representation of visual evidence strength, candidate uncertainty, and phenotype availability at the routing stage, while remaining fully independent of downstream reasoning.

E.3 ACS Policy

Offline rollout supervision. To learn the budget policy π despite the non-differentiable downstream multi-agent execution, we construct supervised targets from offline rollouts over the discrete action set $\mathcal{K}$. For each case x in the policy-training split, we execute the full downstream pipeline once for each $K \in \mathcal{K}$, producing a final result $\hat{y}_K$ and a realized cost. We define the realized utility as

$$u(x, K) = \mathbf{1}(\hat{y}_K = y_{\text{gt}}) - \lambda \cdot \text{cost}(x, K), \tag{12}$$

where y_{gt} denotes the GT disease label and $\text{cost}(x, K)$ is measured by output token usage, which is strongly correlated with K in practice. We train a lightweight MLP to regress the utility estimate $\hat{u}(\psi(x), K)$ from the routing state $\psi(x)$ (Appendix E.2) and the candidate budget K (one-hot encoded), using mean squared error. At inference time, the policy π is defined as

$$K(x) = \pi(\psi(x)) = \arg\max_{K \in \mathcal{K}} \hat{u}(\psi(x), K), \tag{13}$$

Table E.1: Fixed-budget Top-1 accuracy and average output tokens on the test set ($n = 113$).

K	# Correct	Top-1 Acc. (%)	Avg. Tokens
2	34	30.09	5,513
3	38	33.63	7,834
4	40	35.40	10,064
5	32	28.32	12,390
ACS	39	34.51	8,338

Table E.2: Distribution of candidate budgets selected by ACS.

Selected $K(x)$	# Cases
2	27
3	48
4	26
5	12
Total	113

which selects the candidate budget forwarded to the *Planner Agent*.

Leakage-averse training protocol. To avoid performance leakage into end-to-end system evaluation, we keep the test set fixed ($n = 113$) and exclude it from all policy training and hyperparameter selection. We also exclude the training split used for VisCls from ACS policy learning to reduce overfitting to downstream agent behavior. Specifically, we train the ACS policy using only the held-out validation split (10% of data; $n = 46$), as shown in Appendix B, and apply the finalized policy once to the fixed test set for final reporting.

Setting the trade-off λ. Due to the limited size of the validation split, we do not perform performance-driven hyperparameter tuning on λ. Instead, we fix a priori a target average cost penalty η (set to $\eta = 0.10$ in all experiments) and calibrate the scale by normalizing with the mean validation rollout cost:

$$\lambda = \eta \Big/ \mathbb{E}_{x\sim \mathrm{Val},\, K\sim\mathcal{K}}\big[\mathrm{cost}(x, K)\big]. \tag{14}$$

This yields an expected penalty of approximately $-\eta$ at the mean validation cost and keeps the test set unused for tuning.

E.4 Fixed-Budget Accuracy and Token Statistics

We report the exact Top-1 accuracy and token usage statistics for fixed candidate budgets $K \in \{2, 3, 4, 5\}$ on the test set ($n = 113$). Table E.1 provides a reference for interpreting the accuracy–cost trade-off of ACS. Figure K.2 compares ACS with the best-performing fixed-budget policy ($K = 4$). While increasing K initially improves accuracy, performance saturates and degrades at larger budgets despite substantially higher token costs, illustrating the inefficiency of uniform fixed-budget evaluation. A plausible explanation is that larger K introduces more visually confusable candidates into downstream verification, increasing the chance of false matches and making Judge adjudication over longer rationales noisier. This non-monotonicity motivates adaptive budgeting, which allocates larger differentials only on harder cases.

F KNOWLEDGE GRAPH CONSTRUCTION AND QUERYING

Data Sources and Preprocessing. We construct a disease–phenotype knowledge graph (KG) for our 103-disease cohort using the official HPO annotation file (`phenotype.hpoa`) and the HPO ontology (`hp.obo`) for term definitions (release: 2025-10-22). For each disease in our curated list, we retrieve all available phenotype annotations from `phenotype.hpoa` and represent each phenotype by its HPO identifier. Term names are resolved from `hp.obo` and retained for display only. The extracted associations are exported as a canonical JSON snapshot for portability and then loaded into Neo4j for indexed querying. We optionally remove isolated entities with no disease–phenotype associations after import.

Graph Schema. We load the JSON registry into Neo4j as the query backend. The Neo4j graph contains two node types: `Disease` and `Phenotype`. Disease nodes store `disease name` and metadata fields (`orphanet_code`, `disease_abbr`, `subcategory`, `gene`) and Phenotype nodes store `hpo_id` and `hpo_name`. Each disease–phenotype association is represented as a directed relationship `(Disease)-[:HAS_PHENOTYPE]->(Phenotype)` with properties `frequency`. We enforce uniqueness on `Disease(orphanet_code)` and `Phenotype(hpo_id)` and index the corresponding key fields to support efficient lookup.

Frequency Annotation Encoding. We retain the raw frequency label from `phenotype.hpoa` on each `HAS_PHENOTYPE` relationship. For downstream ranking and filtering in retrieval, we additionally map the raw label to an ordinal code `freq_code` under a fixed ordering (higher indicates more frequent): `Obligate` > `Very frequent` > `Frequent` > `Occasional` > `Very rare` > `Unknown`. When a label is missing or not provided by the annotation, we set it to `Unknown`.

KG Coverage and Statistics. After importing HPO annotations for our curated 103-disease cohort into Neo4j, HPOA provides at least one phenotype association for 91 diseases, which form the connected graph used for the Retrieval Agent. This KG thus contains 91 `Disease` nodes, 1,817 `Phenotype` nodes, and 4,259 `HAS_PHENOTYPE` relationships. We report node/edge counts over diseases with at least one associated phenotype. To characterize heterogeneity in evidence availability, which directly affects retrieval difficulty and downstream reasoning reliability, we report the per-disease number of associated phenotypes as in Figure F.1. The mean number of phenotypes per disease is around 47, with a long tail to 140 phenotypes for well-studied conditions. The pronounced variability in annotation density across diseases reflects uneven evidence availability, which directly impacts retrieval difficulty and motivates our design of adaptive, phenotype-driven reasoning.

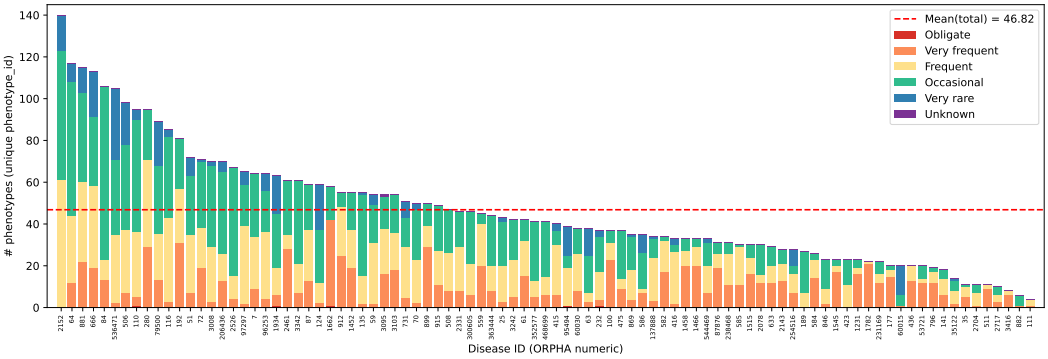


Figure F.1: Stacked bars show the number of unique phenotypes per disease, stratified by HPO frequency. The red dashed line marks the cohort mean.

Query Templates. All runtime retrieval is executed via indexed Neo4j queries:

```
MATCH (d:Disease {disease_name: $disease})-[r:HAS_PHENOTYPE]->(p:Phenotype)
WHERE r.frequency IN $tiers
RETURN p.name AS term, p.phenotype_id AS phenotype_id, r.frequency AS frequency
LIMIT $limit
```

G HIERARCHICAL DOCTOR AGENT MULTI-ROUND PROTOCOL

G.1 HIERARCHICAL DOCTOR AGENTS DETAILS

For each selected candidate disease $d \in D(x)$, we instantiate a per-disease *Doctor Agent* that operates under an explicit *Orchestration–Retrieval–Reasoning* protocol with a multi-round closed loop. We refer to this design as *Hierarchical* because each Doctor agent is itself a compact multi-agent system. While it acts as one component of the global multi-agent pipeline, it internally orchestrates specialized sub-agents to produce a per-disease assessment. Doctor Agents are grounded in a disease–phenotype KG, which provides structured disease-specific evidence for verification. Building on this KG, the Doctor Agent implements a *graph-based RAG* procedure. It retrieves relevant KG facts and conditions reasoning agent to produce auditable per-disease rationales.

Disease-Phenotype KG. Doctor Agents rely on a disease-phenotype KG $\mathcal{G}$ constructed from publicly verified resources (e.g., HPO), which provides disease-phenotype associations. For a candidate d, the graph induces a structured phenotype set $P_{\mathcal{G}}(d)$ and supports phenotype-conditioned retrieval for evidence collection. This external evidence is used to ground the agent's assessment and to make the reasoning trace auditable.

Orchestration Agent. The *Orchestration Agent* initializes a per-disease state for candidate disease using the extracted patient phenotype list $P(x)$ and the its identifier d, while keeping the raw image x restricted by default (Figure 3). Direct access to x is only granted through an explicit fallback gate when phenotype-grounded evidence is insufficient. It maintains a per-disease memory state $M_d^{(t)}$ that accumulates retrieved evidence and intermediate reasoning across rounds. It also records the round index t and termination status (`Round`, `LastRound`, `Action`) and produces a final structured output `Final.json` containing the per-disease binary decision (`match` or `no_match`) and rationale. Importantly, the orchestrator enforces a strong separation between phenotype-grounded reasoning and image perception By default, downstream modules operate on $P(x)$ and KG evidence only, and direct image access is disallowed unless a strict fallback trigger is met.

Retrieval Agent. The *Retrieval Agent* retrieves candidate-conditioned evidence from the KG to support graph-based retrieval-augmented reasoning. At round t, given a candidate disease d and a predefined frequency constraint $\rho^{(t)}$, it extracts a filtered facial phenotype profile from the KG $\tilde{P}_d^{(t)} = P_{\mathcal{G}}(d; \rho^{(t)})$, where $\rho^{(t)}$ is progressively relaxed across rounds to expand the evidence scope from highly typical to rarer phenotypes (details in Appendix G.3). All retrieved phenotypes $\tilde{P}_d^{(t)}$ and the active frequency set $\rho^{(t)}$ are serialized into `retrieval.json` for transparent inspection.

Reasoning Agent. After the Retrieval Agent returns the KG evidence $\tilde{P}_d^{(t)}$, the *Reasoning Agent* is invoked to compare it against the extracted patient phenotype list $P(x)$, perform evidence-based compatibility analysis and provide reasoning. It outputs: (i) matched terms, (ii) missing or conflicting terms, (iii) a matched-rate between 0 to 1, and (iv) a boolean signal `Need_more_kg` indicating whether additional KG retrieval is required. It also produces a short rationale grounded in the matched vs. missing/conflict phenotypes. All fields are serialized in `reasoning.json`, forming an auditable per-round reasoning trace.

Multi-round Closed Loop. The Doctor agent executes a multi-round closed loop in which the orchestrator alternates between retrieval and reasoning until (a) the evidence becomes sufficient, or (b) a strict round cap is reached. This alternating retrieval-reasoning loop follows the same high-level **ReAct**-style agentic prompting, where actions (here, KG retrieval) are interleaved with reasoning to iteratively refine evidence before termination Yao et al. (2023). When `Need_more_kg` is true or the missing/conflict signal remains high, the orchestrator initiates another retrieval–reasoning round with more flexible $\rho^{(t)}$. Otherwise, it stops and emits `Final.json` for disease d, containing the final decision and rationale derived from the last round.

Restricted Fallback Self-Decide. While the default pathway avoids direct image usage, some cases may remain under-supported due to insufficient phenotype extraction or incomplete KG evidence. When the maximum round budget is reached but the evidence is still deemed insufficient (e.g., Orchestrator's `Action` = `Continue`) the orchestrator opens an explicit *fallback gate* (Figure 3). In this mode, a *Fallback Self-Decide* module is granted restricted access to $(d, x, P(x))$ and directly

judges whether the patient matches disease d by jointly considering the raw facial image and the extracted phenotypes. The resulting per-disease decision and rationale are serialized to `Final.json`.

G.2 Doctor Agent Protocol

Following the architecture in Appendix G.1, a detailed Doctor Agent pseudocode is provided in Algorithm G.1.

Algorithm G.1 Hierarchical Doctor Agent with LLM-Governed Multi-round KG Retrieval

Require: Candidate disease d; normalized phenotypes $P(x)$; KG fetcher $\mathcal{K}$; round cap $T_{\max}$; frequency schedule $\{\rho^{(t)}\}$; image x (optional).
Ensure: `Final.json` decision for d and per-round trace files.

$t \leftarrow 1$
Initialize trace list $\mathcal{T}$
`final_decision` $\leftarrow$ **null**
while $t \le T_{\max}$ **do**
  Set round flag: `last_round` $\leftarrow (t = T_{\max})$
  Retrieval Agent (round t):
    $K^{(t)} \leftarrow \mathcal{K}.\textsc{FetchCandidates}(d, \rho^{(t)})$
  Write `retrieval.json` containing $\{d, t, \rho^{(t)}, K^{(t)}\}$
  Reasoning Agent:
    $R^{(t)} \leftarrow \mathcal{L}.\textsc{Reason}(d, P(x), K^{(t)})$
  Parse from $R^{(t)}$:
      `decision`, `matched_rate`, `missing`, `conflicting`, `need_more_kg`
  Write `reasoning.json` for round t
  Orchestration Agent:
    $O^{(t)} \leftarrow \mathcal{L}.\textsc{Orchestrate}(d, P(x), K^{(t)}, R^{(t)}, t, \texttt{last_round})$
  Parse from $O^{(t)}$: `action` $\in \{$`continue`, `stop`$\}$, `final_decision`
  Write `orchestration.json` for round t
  Append $(t, \rho^{(t)}, R^{(t)}, O^{(t)})$ to trace list $\mathcal{T}$
  if $O^{(t)}$.`action` $=$ `continue` **and** `last_round` $=$ **false then**
    $t \leftarrow t + 1$
    continue
  end if
  if $O^{(t)}$.`action` $=$ `continue` **and** `last_round` $=$ **true then**
    Fallback due to insufficient evidence:
      $F \leftarrow \mathcal{L}.\textsc{FallbackDecide}(d, x, P(x), \mathcal{T})$
    Write `Final.json` $\leftarrow F$ and **return**
  end if
  if $O^{(t)}$.`action` $=$ `stop` **then**
    Write `Final.json` $\leftarrow O^{(t)}$ and **return**
  end if
end while

Per-round artifacts. At each round t, the doctor agent writes (i) `retrieval.json` containing {`disease`, `round`, `freq_tiers`, `retrieved_terms`}, and (ii) `reasoning.json` containing structured fields {`matched`, `missing`, `conflicting`, `matched_rate`, `need_more_kg`, `prelim_decision`}. The final output `Final.json` stores {`decision` $\in$ {`match`,`no_match`}, `rationale`} for the disease candidate.

G.3 Retrieval Schedule and Frequency Relaxation Across Rounds

We control KG retrieval using a fixed frequency schedule $\rho^{(t)}$ that expands monotonically across rounds. At round t, the KG fetcher returns only disease–phenotype edges whose frequency label is contained in $\rho^{(t)}$:

$$K^{(t)} = \{(p, \texttt{freq}) : (d) - [\texttt{HAS_PHENOTYPE}] -> (p) \ \wedge \ \texttt{freq} \in \rho^{(t)}\}. \tag{15}$$

The round-wise schedule is:

$$\rho^{(1)} = \{\texttt{Obligate}, \texttt{Very frequent}\},$$
$$\rho^{(2)} = \{\texttt{Obligate}, \texttt{Very frequent}, \texttt{Frequent}\},$$
$$\rho^{(3)} = \{\texttt{Obligate}, \texttt{Very frequent}, \texttt{Frequent}, \texttt{Occasional}\},$$

Unless stated otherwise, we set $T_{\max} = 3$ as in Algorithm G.1. Missing frequency annotations can optionally be included via an additional predicate `OR r.frequency IS NULL` in the Neo4j query. When enabled, it is applied only in the final round to avoid bypassing the frequency-controlled relaxation.

G.4 RAG Illustration

In Figure G.1, we illustrate how our Doctor Agent instantiates ***KG-based RAG*** using the disease–phenotype KG as the retrieval or evidence source. Concretely, for each candidate disease d, the agent retrieves KG evidence and augments the reasoning context to perform evidence-grounded verification in a multi-round loop.

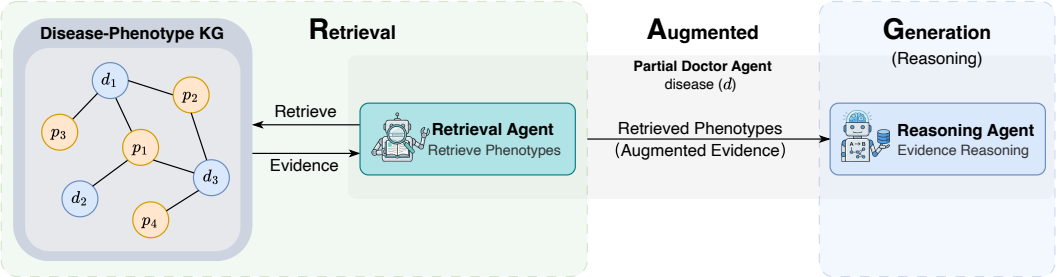


Figure G.1: KG-powered RAG illustration for a per-disease Doctor Agent. The retrieval agent queries the disease–phenotype KG to obtain evidence, which is then appended (augmented) to the reasoning context for evidence-grounded verification and decision generation.

G.5 Robustness and Behavior Analysis

Robustness to incomplete KG coverage. We analyze the impact of incomplete KG (KG) coverage from multiple perspectives.

First, fallback reasoning is triggered in 34.5% (39/113) of test cases, defined as cases where at least one of the Top-5 candidates lacks KG support. A position-wise analysis shows that missing KG coverage is relatively limited among high-ranked candidates (Top-1: 8.0%, Top-2: 4.4%, Top-3: 5.3%), indicating that most confident candidates remain KG-supported.

Second, we compare performance between cases with and without missing KG coverage. RareVA achieves 33.3% accuracy on cases with incomplete KG and 35.1% on cases without, showing comparable performance across both groups.

These results suggest that while KG provides useful structured evidence when available, the overall performance gains are not solely driven by KG-covered diseases. Instead, the fallback mechanism enables robust reasoning when KG support is incomplete, improving applicability in realistic settings where knowledge bases are inherently sparse. All statistics are computed on the test set ($n = 113$) under the same candidate-bounded protocol as in Table 1.

H JUDGE AGENT DETAILS

The *Judge Agent* aggregates per-disease Doctor assessments and selects the highest-priority candidate $\hat{y}$ among the candidate set $D(x)$. The Judge operates in two modes based on the Doctor outputs $\{A_d(z_d, r_d)\}_{d \in D(x)}$, where A_d contains a $z_d \in \{0, 1\}$ denotes a binary `match` / `no_match` verdict and r_d is the corresponding evidence-grounded rationale. In addition to the final prioritized candidate and explanation, the Judge Agent also outputs a lightweight compute summary (e.g., token usage) for downstream compute-aware analysis.

MATCH mode. If at least one Doctor Agent predicts `match` ($\exists d : z_d = 1$), the Judge treats the matched diseases as strong priors and selects a single highest-priority candidate $\hat{y}$ from this subset. It ranks matched candidates using a fixed rubric that scores (A) alignment of the Doctor rationale to facial findings, (B) case-specificity of the rationale, and (C) consistency with the raw image evidence. The Judge may override a Doctor `match` only when the stated rationale clearly contradicts observable facial evidence in x, in which case it must explicitly justify the override.

NO_MATCH mode. If all Doctor Agents predict `no_match` ($\forall d : z_d = 0$), the Judge must still output a single candidate $\hat{y}$ from $D(x)$ under weak evidence. Specifically, it uses the raw image x to conduct a conservative image–phenotype check, splitting $P(x)$ into phenotypes that are supported by observable facial cues versus those that are unclear/not supported. It then ranks candidates in $D(x)$ using a constrained rubric that scores (A) consistency with the supported phenotypes, (B) prior support from the candidate ordering in $D(x)$, and (C) absence of strong contradictions to salient image evidence. Finally, the Judge appends an explicit uncertainty note grounded in (i) the unanimous `no_match` outputs from Doctors and (ii) limited or inconsistent phenotype support from $P(x)$.

I UNCERTAINTY AND STABILITY QUANTIFICATION

I.1 STABILITY ANALYSIS

RDFace dataset is limited in size. however, it is a realistic testbed for rare disease verification and prioritization under extreme data scarcity. To assess the stability of our results, we perform a **stochasticity analysis** by running RareVA on the same test set for multiple runs with different random seeds. This approach captures variability due to stochastic elements in the model (e.g., LLM sampling) while keeping the dataset fixed. Across four stochastic runs, we report the reports of correct predictions among the test set. Results are 39, 39, 39, and 38 / 113 ($\pm$ 0.38 pp), demonstrating high stability given a fixed dataset. We do not perform repeated dataset splits due to extremely limited per-class samples, as stratified or leave-one-out splits would further reduce coverage and increase variance. In this regime, bootstrap resampling provides a more reliable alternative.

I.2 BOOTSTRAP CONFIDENCE INTERVAL PROTOCOL

We quantify uncertainty due to the finite size of the test set by applying **bootstrap resampling** to Top-1 accuracy on the fixed 113-case test set ($N = 113$), following standard practice Koehn (2004). For each method, we represent per-case correctness as a binary vector $y \in \{0, 1\}^N$ (e.g., Top-1 correctness) on this fixed evaluation set. We draw $B = 10{,}000$ bootstrap resamples by sampling N cases with replacement from $\{1, \ldots, N\}$. For each resample b, we recompute the accuracy as the mean $\hat{m}^{(b)} = \frac{1}{N} \sum_{i \in S_b} y_i$. We report the point estimate $\hat{m}$ on the original test set together with the bootstrap standard error (SE) (the standard deviation of $\{\hat{m}^{(b)}\}_{b=1}^{B}$) and a 95% percentile confidence interval (CI) given by the 2.5th and 97.5th percentiles of $\{\hat{m}^{(b)}\}$. For method comparisons against a reference (VisCls prior), as in our ablations, we use **paired bootstrap resampling**. For each resample S_b, we evaluate methods on the same resampled cases and form the difference $\Delta^{(b)} = \hat{m}^{(b)}_{\text{method}} - \hat{m}^{(b)}_{\text{VisCls}}$ and report the 95% percentile CI of $\{\Delta^{(b)}\}$ and its SE in pp.

I.3 Conditional Uncertainty on Test Set

To extend the baseline and ablation results with bootstrap uncertainty analysis in Table 1 and Table 5 in the main text, we additionally report the standard error (SE) and 95% CI of the improvement compared to the VisCls prior and their SEs for all methods in Table I.1.

Table I.1: **Uncertainty quantification on test set ($n = 113$).** We report Top-1 accuracy with 95% bootstrap CI (SE) and the improvement over the VisCls prior in pp with 95% bootstrap CI (SE).

Method	Top-1 Acc. (%)	Top-1 Acc. CI (SE) (%)	Δ (pp)	Δ CI (SE) (pp)
Perception Prior				
VisCls (DenseNet)	25.66	[17.70, 33.63] (4.09)	—	—
Baselines				
Non-agentic reranker	22.12	[15.04, 30.09] (3.88)	-3.54	[-7.08, -0.88] (1.73)
Structured phenotype scoring	25.66	[17.70, 33.63] (4.10)	+0.00	[-2.65, 2.65] (1.25)
Single VLM (Gemini)	26.55	[18.58, 34.51] (4.13)	+0.88	[-4.42, 6.19] (2.65)
+ Pheno	21.24	[14.16, 29.20] (3.85)	-4.42	[-8.85, 0.00] (2.28)
+ KG-RAG	28.32	[20.35, 36.28] (4.21)	+2.65	[-2.65, 7.96] (2.63)
+ multi-round	29.20	[21.24, 37.17] (4.26)	+3.54	[-1.77, 8.85] (2.77)
Ablations				
w/o KG-RAG	30.97	[22.12, 39.82] (4.36)	+5.31	[-0.88, 11.50] (3.02)
w/o ACS (fixed budget K)				
$K = 2$	30.09	[22.10, 38.05] (4.31)	+4.42	[-0.88, 10.62] (2.90)
$K = 3$	33.63	[24.78, 42.48] (4.45)	+7.96	[2.65, 13.27] (2.81)
$K = 4$	35.40	[26.55, 44.25] (4.51)	+9.73	[4.42, 15.93] (3.05)
$K = 5$	28.32	[20.35, 37.17] (4.23)	+2.65	[-2.65, 7.96] (2.93)
w/o multi-round Doctor Agents	30.97	[23.01, 39.82] (4.36)	+5.31	[0.00, 11.50] (3.02)
w/o hierarchical Doctor Agents	29.20	[21.24, 37.17] (4.27)	+3.54	[-1.77, 8.85] (2.77)
w/o Judge Agents	29.20	[21.24, 38.05] (4.28)	+3.54	[-1.77, 8.85] (2.76)
Full Model				
RareVA (Ours)	**34.51**	**[25.66, 43.36] (4.47)**	**+8.85**	**[2.65, 15.93] (3.42)**

Table I.2: **Conditional uncertainty with GT ∈ Top-5 candidates ($n = 49$).** Results are reported using the same bootstrap metrics as Table I.1.

Method	Top-1 Acc. (%)	Top-1 Acc. CI (SE) (%)	Δ (pp)	Δ CI (SE) (pp)
Perception Prior				
VisCls (DenseNet)	59.18	[44.90, 73.47] (7.09)	—	—
Baselines				
Non-agentic reranker	51.02	[36.73, 65.31] (7.21)	-8.16	[-16.33, -2.04] (3.85)
Structured phenotype scoring	59.18	[44.90, 73.47] (7.11)	+0.00	[-6.12, 6.12] (2.89)
Single VLM (Gemini)	61.22	[46.94, 75.51] (7.05)	+2.04	[-10.20, 14.29] (6.08)
+ Pheno	48.98	[34.69, 63.27] (7.14)	-10.20	[-20.41, 0.00] (5.20)
+ KG-RAG	65.31	[51.02, 77.55] (6.89)	+6.12	[-6.12, 18.37] (6.03)
+ multi-round	67.35	[53.06, 79.59] (6.78)	+8.16	[-4.08, 20.41] (6.31)
Ablations				
w/o KG-RAG	71.43	[59.18, 83.67] (6.51)	+12.24	[0.00, 26.53] (6.79)
w/o ACS (fixed budget K)				
$K = 2$	69.39	[57.14, 81.63] (6.66)	+10.20	[-2.04, 22.45] (6.56)
$K = 3$	77.55	[65.31, 87.76] (5.96)	+18.37	[6.12, 30.61] (6.24)
$K = 4$	81.63	[69.39, 91.84] (5.51)	+22.45	[10.20, 34.69] (6.64)
$K = 5$	65.31	[51.02, 77.55] (6.88)	+6.12	[-6.12, 18.37] (6.63)
w/o multi-round Doctor Agents	71.43	[59.18, 83.67] (6.52)	+12.24	[0.00, 26.53] (6.84)
w/o hierarchical Doctor Agents	67.35	[53.06, 79.59] (6.79)	+8.16	[-4.08, 20.41] (6.34)
w/o Judge Agents	67.35	[53.06, 79.59] (6.79)	+8.16	[-4.08, 20.41] (6.34)
Full Model				
RareVA (Ours)	**79.59**	**[67.35, 89.80] (5.85)**	**+20.41**	**[6.12, 34.69] (7.62)**

I.4 Conditional Uncertainty When the GT Is in the Candidate List

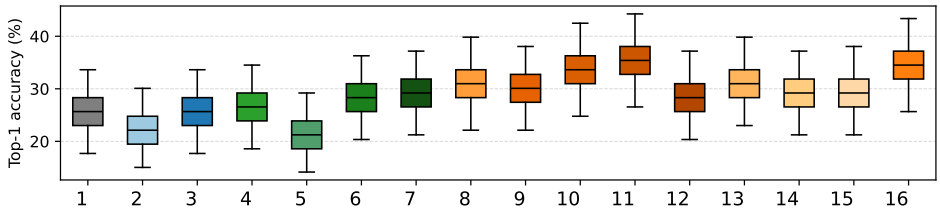
A key question is whether the *verification* stage provides benefit once the correct disease appears among the VisCls-proposed candidates. We therefore condition on cases where the GT lies in the Top-5 list (coverage region), yielding a subset of n=49 cases (cov@5 = 43.36% overall). Table I.2 reports bootstrap uncertainty for this subset.

Two observations follow. First, the agent exhibits a substantially higher Top-1 accuracy than the VisCls prior on this subset, with a paired improvement Δ whose CI excludes zero. Second, naïve phenotype-driven shortcuts remain inferior in this controlled regime, reinforcing the interpretation that phenotypes should act as an *interface* for evidence retrieval and hypothesis testing rather than a direct decision signal. While CIs are wider than on the full set due to smaller n, the magnitude and sign of the paired improvements remain informative.

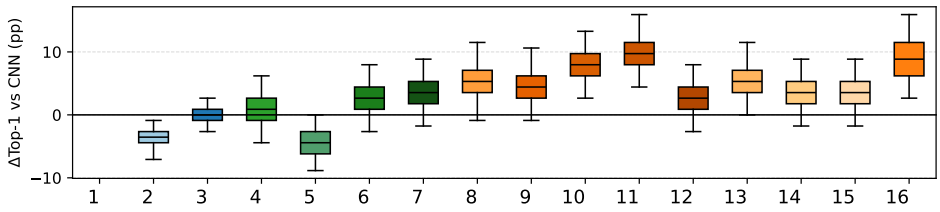
I.5 Bootstrap Uncertainty Visualization

Below we visualize the full bootstrap distributions of Top-1 accuracy and paired Δ for all methods on both the full test set and the GT $\in$ Top-5 subset. Importantly, the methods are indexed from 1–16 for readability. Full names are provided below:

1. VisCls
2. Non-Agentic Reranker
3. Structured Phenotype Match Scoring
4. Single VLM
5. Single VLM + Pheno
6. Single VLM + KG-RAG
7. Single VLM + KG-RAG + multi-round
8. RareVA w/o KG-RAG
9. RareVA w/o ACS (K=2)
10. RareVA w/o ACS (K=3)
11. RareVA w/o ACS (K=4)
12. RareVA w/o ACS (K=5)
13. RareVA w/o multi-round Doctor Agent
14. RareVA w/o hierarchical Doctor Agent
15. RareVA w/o judge Agent
16. RareVA (Ours)



(a) **ALL** Top-1 accuracy (%) with 95% bootstrap CI.



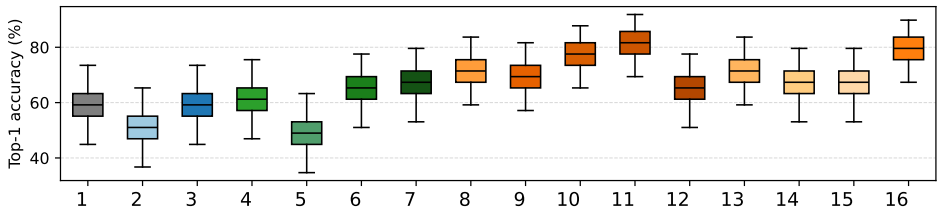
(b) **ALL** Paired Δ vs.VisCls (pp) with 95% bootstrap CI.

Figure I.1: **All test cases bootstrap uncertainty visualization.** Whiskers indicate percentile 95% CIs from bootstrap resampling. The horizontal line at 0 indicates no improvement.

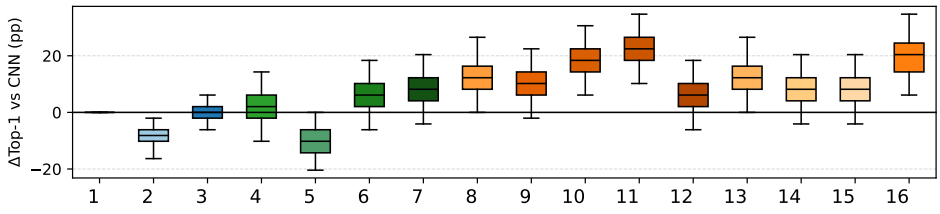
I.6 Discussion and Implications

Figure I.1 and Figure I.2 complement Table I.1 and Table I.2 by visualizing the full bootstrap uncertainty, revealing not only confidence intervals but also the *shape*, *spread*, and *relative ordering* of methods. This provides a more nuanced view of robustness under small-n uncertainty. We highlight four key observations.

Paired Δ reveals consistent improvement patterns. With n=113, per-method accuracy distributions exhibit substantial overlap, making absolute comparisons visually ambiguous. However, the paired Δ distributions (bottom panels) show a clear rightward shift for RareVA and its stronger variants, with most of their mass lying above 0. In contrast, weaker baselines cluster tightly around or below 0. This confirms that the improvement over the VisCls prior is not driven by a few favorable



(a) **GT** $\in$ **Top-5** Top-1 accuracy (%) with 95% bootstrap CI.



(b) **GT** $\in$ **Top-5** Paired Δ vs.VisCls (pp) with 95% bootstrap CI.

Figure I.2: **GT $\in$ Top-5 bootstrap uncertainty visualization.** Whiskers indicate percentile 95% CIs from bootstrap resampling. The horizontal line at 0 indicates no improvement.

resamples, but reflects a *systematic shift* in performance across bootstrap draws. Thus, paired Δ provides a more reliable signal than absolute accuracy under small-sample uncertainty.

Verification gains are stable once coverage is satisfied. Conditioning on GT $\in$ Top-5 (Figure I.2) leads to two consistent changes: (i) accuracy distributions shift upward across all methods, and (ii) Δ distributions become more separated, with stronger methods (e.g., RareVA and larger K variants) exhibiting clearly positive and less overlapping distributions. This indicates that once the correct hypothesis is included, the multi-round verification process yields *stable and repeatable gains*, not just improvements in expectation. The tighter clustering of high-performing methods further suggests reduced variance in this regime. Therefore, the main bottleneck lies in the proposal stage rather than verification.

Performance improvements follow a monotonic structure. Across both figures, methods exhibit a consistent ordering: from non-agentic baselines, to single-pass VLM variants, to retrieval-augmented and multi-round variants, and finally to full agentic systems. In particular, the ACS variants ($K = 2 \rightarrow 4$) show a clear monotonic improvement in both accuracy and Δ, while overly large K ($K = 5$) degrades performance, indicating diminishing returns and potential noise introduction. This pattern is visible not only in mean values but also in the entire distribution, supporting the interpretation that ACS balances exploration and noise control.

Shortcut strategies are consistently unstable. The bootstrap distributions further reveal that shortcut methods (phenotype-only or shallow VLM combinations) have wider spreads and often center near or below the VisCls baseline. Their Δ distributions frequently cross or fall below 0, even when point estimates appear competitive. This indicates that such methods are highly sensitive to resampling and lack consistent gains across cases. In contrast, structured multi-agent reasoning yields both higher medians and more stable distributions, reinforcing that phenotypes are more effective as *intermediate representations* rather than standalone decision signals.

Evaluation implications. The bootstrap distributions highlight the importance of analyzing *distributional behavior* rather than relying solely on point estimates and interval bounds. Future evaluations would benefit from larger test sets and/or multiple splits to further reduce uncertainty. From a modeling perspective, improving the proposal stage (e.g., stronger candidate generators or calibrated expansion policies) is likely to yield compounding gains, as it directly increases the proportion of cases that can benefit from the demonstrated robustness of the verification stage.

J Qualitative Case Studies and Error Analysis

In this section, we present representative success and failure cases to illustrate how the proposed pipeline behaves end-to-end (VisCls pre-selection → VLM phenotype extraction → planner candidate budgeting → multi-round doctor verification with retrieval → judge aggregation).

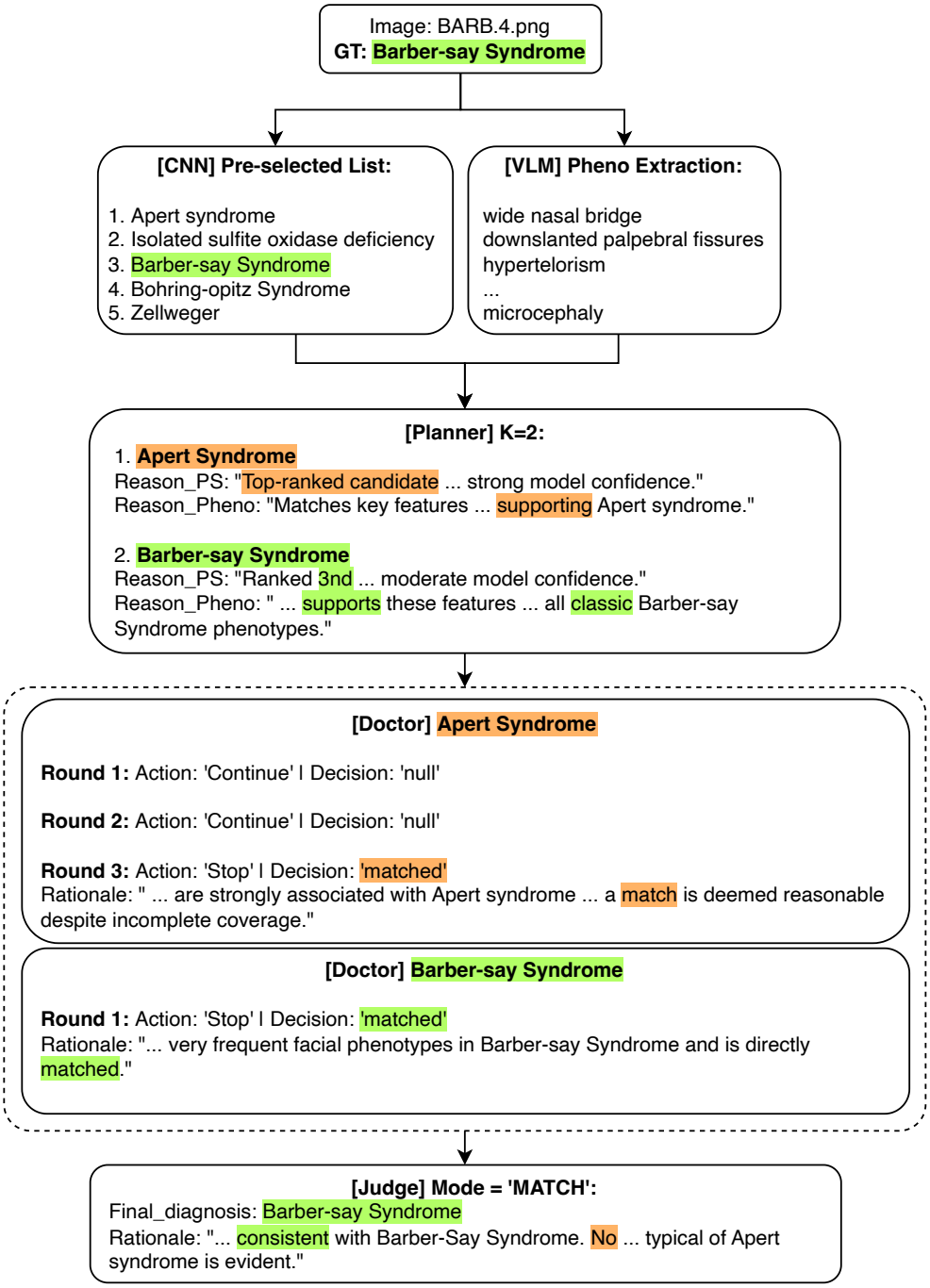


Figure J.1: **BARB.4 (GT: Barber–Say Syndrome)**. Planner selects $K{=}2$. GT Doctor stops at Round 1 with `match`. Non-GT Doctor stops at Round 3 also with `match`. Judge finalizes Barber–Say after rejecting alternatives with insufficient phenotype overlap.

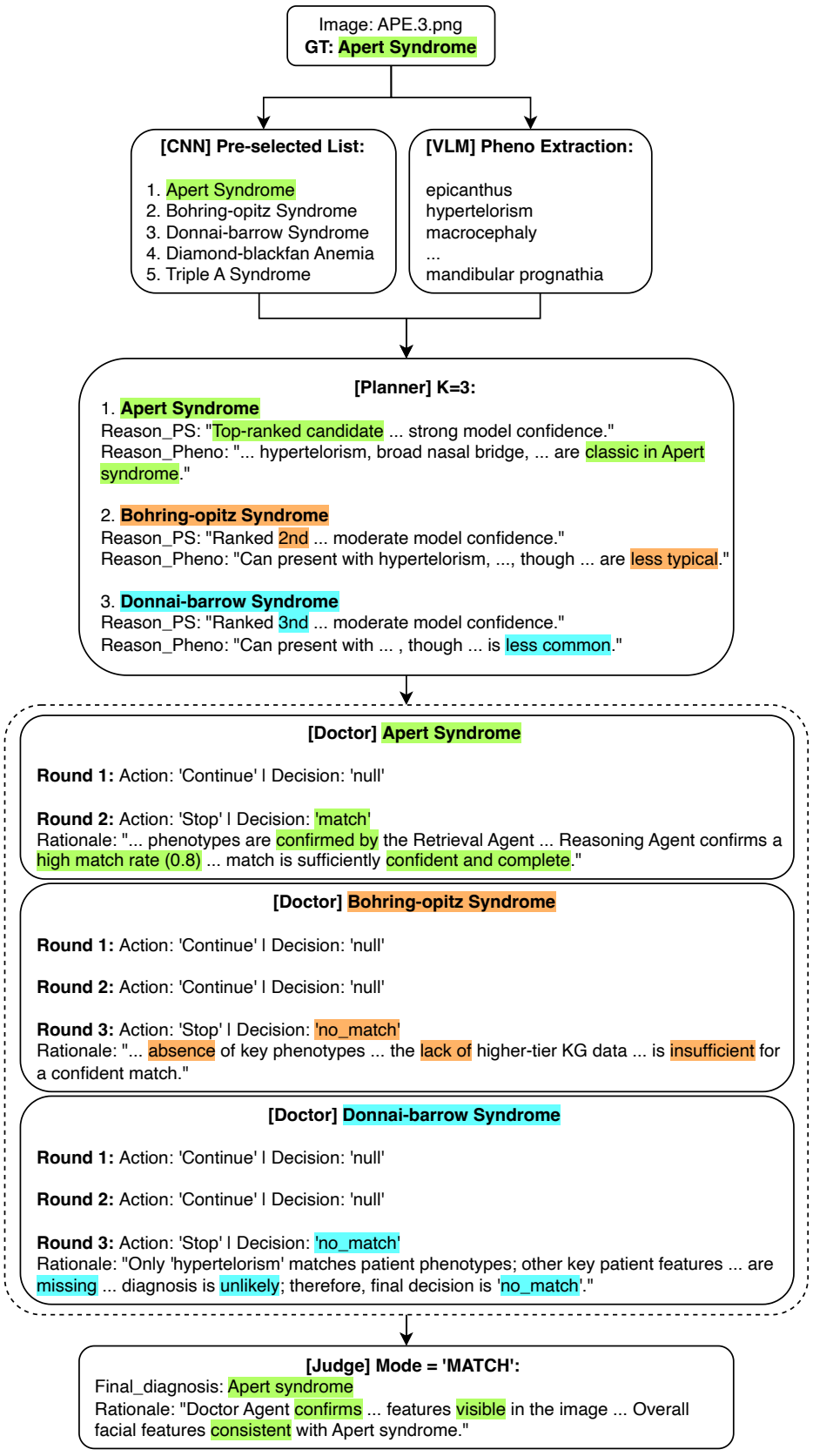


Figure J.2: **APE.3 (GT: Apert Syndrome)**. Planner selects $K{=}3$. GT Doctor stops at Round 2 with `match`. Non-GT Doctors rejects all others at Round 3. Judge finalizes Apert syndrome.

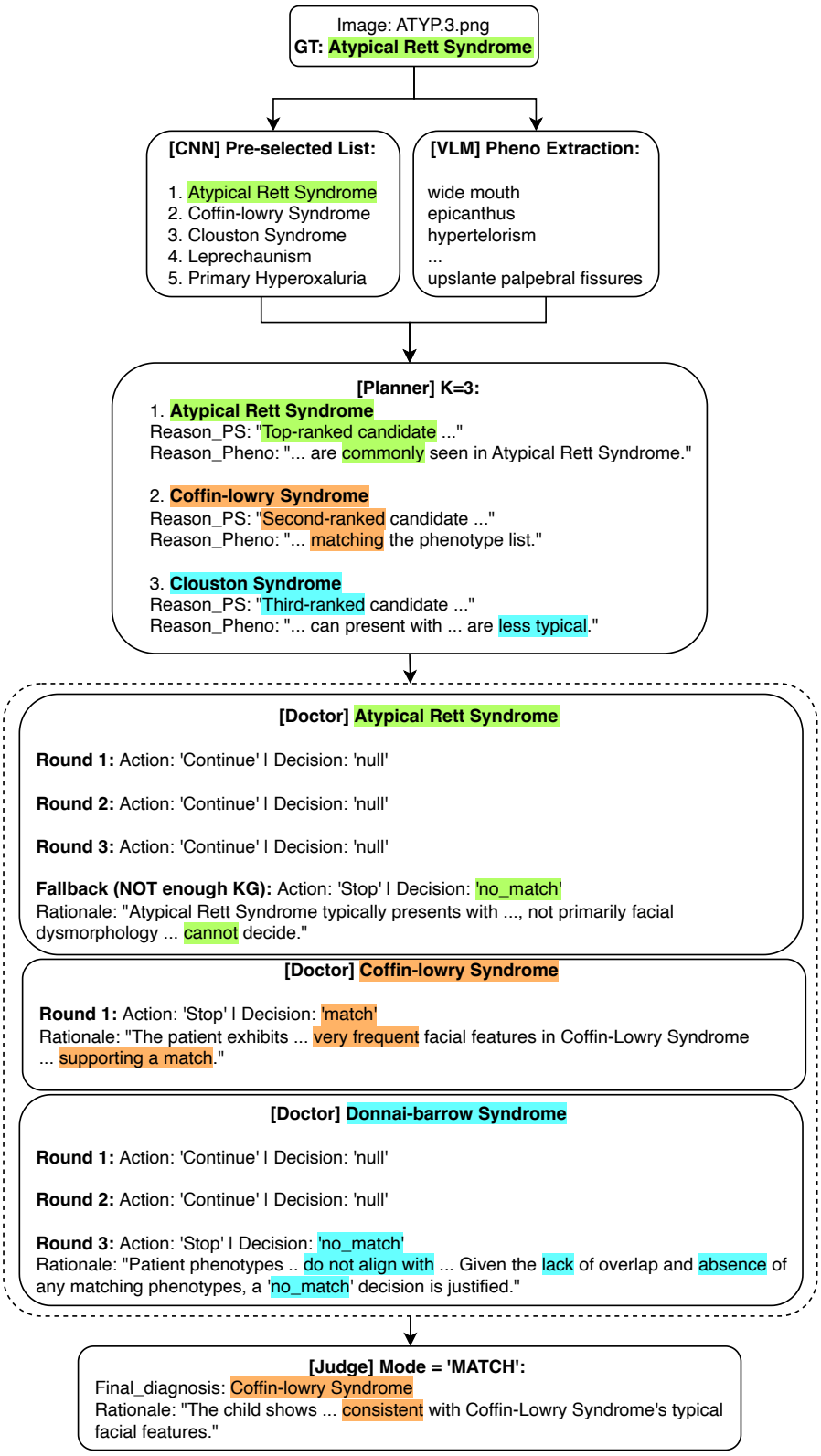


Figure J.3: **ATYP.3 (GT: Atypical Rett Syndrome).** Planner selects $K{=}3$. GT remains in candidates but verification cannot ground a confident facial match by Doctor (KG facial evidence insufficient $\rightarrow$ fallback `no_match`). A competing syndrome matches the extracted phenotype list and is selected by the Judge.

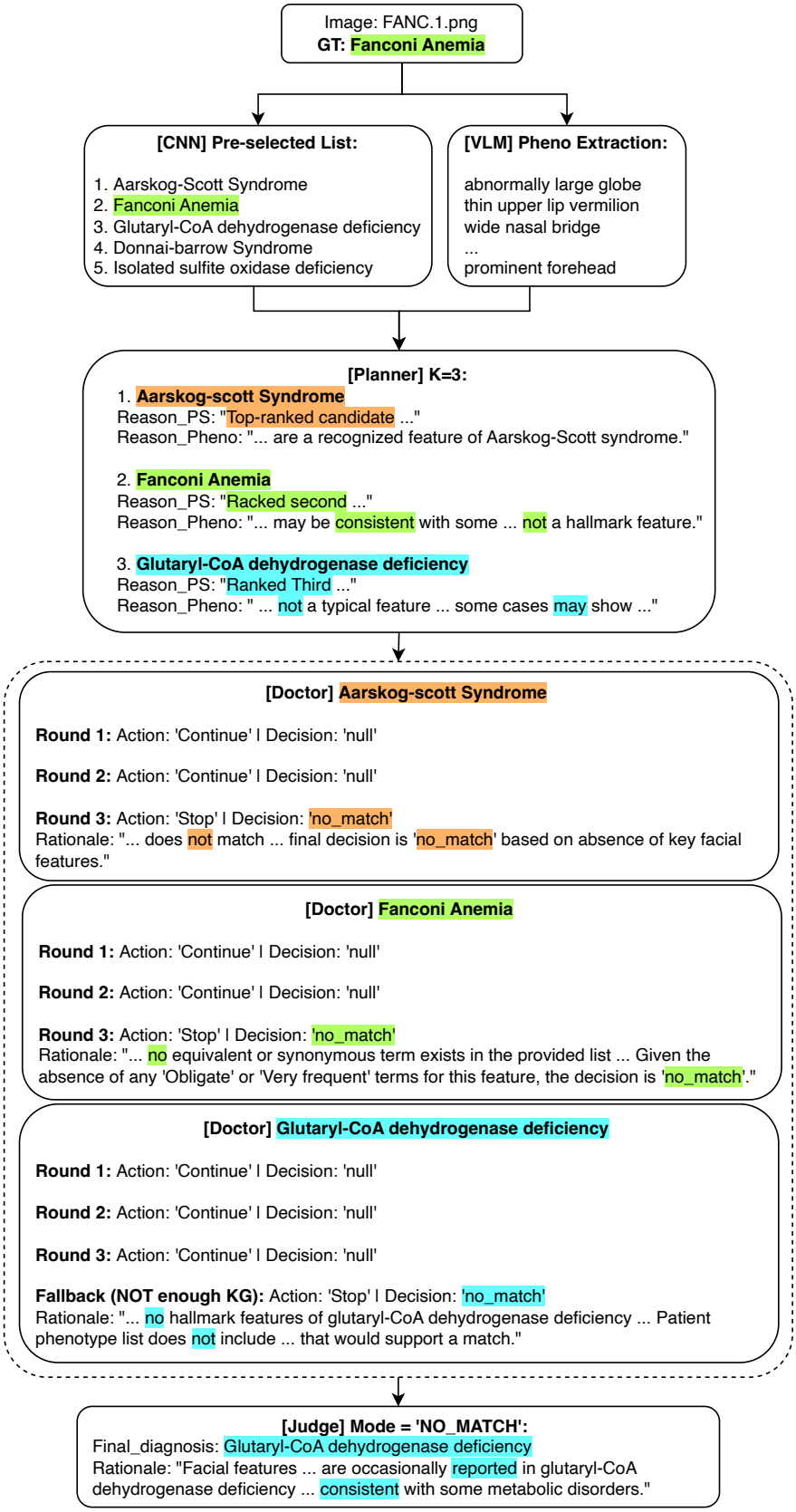


Figure J.4: **FANC.1(GT: Fanconi anemia)**. Planner selects $K=3$. All top candidates rejected by Doctors (`no_match`) due to missing hallmark/very-frequent facial terms or unmatched synonyms. Judge runs in `NO_MATCH` mode and outputs an alternative disease with weak facial support.

How to read the traces and color coding. In all qualitative case figures, *green* denotes the *GT* disease. All other colors (orange & cyan) are used *only* to visually distinguish different **doctor-agent traces** (one per candidate disease) within the same case. Importantly, a color does *not* indicate correctness. It is a consistent identifier for a specific candidate's verification trajectory (e.g., extracted phenotypes → retrieved evidence → per-round decisions and stop reasons). This visualization is designed for ***traceable causal reasoning***, where readers can follow, per candidate, which evidence was consulted, what was accepted or rejected, and why the agent made decision. These results also showed that our RareVA exposes structured intermediate states and decision fields, thereby improving **interpretability** and **auditability** of the final result compared to black-box end-to-end models and single-round VLM-based systems.

J.1 Representative Success Cases

Figures J.1 and J.2 shows two representative success cases.

In **BARB.4**, although Apert is ranked highly by the VisCls, the phenotype signal and verification quickly support Barber–Say, and the judge finalizes the correct disease. This case highlights the benefit of explicit verification that the system does not simply follow the top VisCls candidate when phenotype-grounded evidence contradicts it.

In **APE.3**, the doctor reaches a confident `match` for the correct candidate by Round 2, while rejecting other plausible alternatives due to missing key facial phenotypes.

J.2 Representative Failure Examples

Figures J.3 and J.4 illustrates two failure modes that recur in our error analysis.

First, in **ATYP.3**, the correct disease remains in the candidate set, but the doctor cannot establish a confident `match` under the evidence-grounded protocol. The trace indicates that Atypical Rett is not primarily characterized by distinctive facial dysmorphology, and the available KG facial evidence is insufficient, triggering the *fallback* stop with `no_match`. Meanwhile, a competing candidate (Coffin–Lowry Syndrome) achieves a `match` due to stronger alignment with the phenotype list, causing the judge to finalize the wrong disease.

Second, for the **FANC.1** case, none of the top candidates receives a confident `match`: multiple candidates are rejected because hallmark facial phenotypes are absent or because no synonymous/high-frequency KG terms can be matched to the extracted phenotype list. The judge accordingly operates in `NO_MATCH` mode and outputs an alternative disease with only weak/occasional facial support, reflecting a conservative-but-imperfect behavior under limited evidence.

Overall Discussion. Overall, the failure cases suggest that errors are less about the orchestration mechanics and more about *information insufficiency and misalignment* under a face-centric evidence protocol. When disease-specific facial evidence in the KG is sparse or non-discriminative, the Doctor Agent may be forced into a conservative `no_match`/fallback decision even if the GT is present in the candidate set. Conversely, in visually confusable settings, an alternative syndrome with stronger or more easily matchable facial phenotypes can receive higher evidence support and dominate the Judge aggregation, yielding a wrong-but-plausible disease. A second recurring driver is phenotype signal quality. Extraction noise and term-mapping constraints (e.g., missing synonyms or overly strict "very-frequent/obligate" matching) can prevent otherwise relevant evidence from being credited, amplifying `no_match` outcomes. These observations motivate an actionable failure taxonomy and corresponding mitigation knobs, including (i) strengthening candidate coverage/ranking when uncertainty is high, (ii) relaxing or adapting the retrieval schedule and evidence-matching strictness under evidence sparsity, (iii) improving phenotype verification/normalization and synonym coverage, and (iv) refining Judge aggregation rules to reduce over-reliance on facial-only support for disorders known to have weak facial signatures. Importantly, RareVA makes these failure modes *inspectable*. The per-candidate, per-round traces localize where the pipeline breaks, enabling targeted debugging and principled future improvements rather than opaque end-to-end errors.

K Additional Analysis

K.1 Robustness under Synthetic Distribution Shift

To probe generalization beyond the original test distribution, we conduct a pilot study on **unseen perturbed samples** derived from held-out test cases.

Specifically, we select 5 classes with sufficient data (≥ 4 samples) and generate 10 synthetic variations per class using the few-shot diffusion model DreamBooth Ruiz et al. (2023), resulting in 50 perturbed samples.

This setting evaluates robustness to image-level distribution shift rather than unseen disease classes. The result remains within the candidate-bounded label space.

Table K.1: Robustness under synthetic distribution shift.

Method	# Correct Cases
VisCls (DenseNet)	6
Single VLM (Gemini)	7
RareVA (Ours)	11

Under this setup, RareVA retains more correct predictions under perturbations than both VisCls and Single VLM (Table K.1), indicating improved robustness to image-level distribution shift.

While preliminary, this pilot study provides complementary evidence that candidate-wise, evidence-grounded reasoning improves stability on shifted unseen samples.

K.2 Per-disease Performance Analysis

In this section, we show the additional per-disease Top-1 accuracy for RareVA on RDFace.

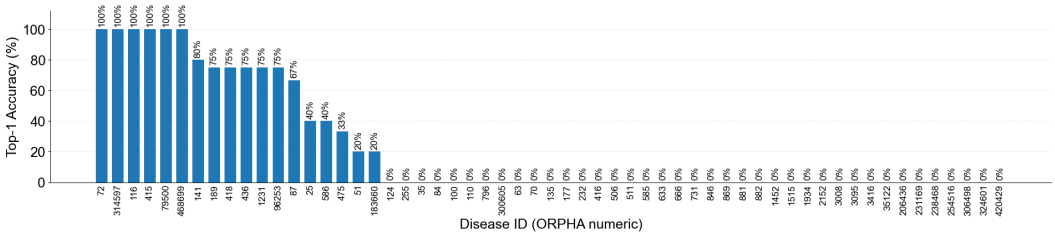


Figure K.1: Per-disease Top-1 accuracy (%) for RareVA on RDFace.

Figure K.1 shows the per-disease Top-1 accuracy. We observe that performance varies across diseases, highlighting the influence of disease-specific characteristics.

K.3 Comparison with the Best Fixed Budget

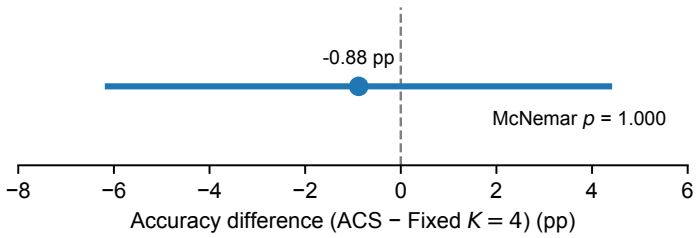


Figure K.2: Paired comparison between ACS and the best fixed-budget policy ($K = 4$).

Figure K.2 compares ACS with the best-performing fixed-budget policy ($K = 4$). The paired bootstrap 95% confidence interval for the accuracy difference spans zero ($[-6.19, 4.42]$ pp), and

McNemar's exact test detects no significant difference ($p = 1.0$). These results further support that ACS achieves comparable performance while substantially reducing computational cost through.

K.4 Subgroup Analysis

We provide a preliminary subgroup analysis across region, sex, and age in Table K.2. Because demographic metadata are missing for some samples, this analysis includes only cases with available metadata for the corresponding subgroup attribute. Due to small and imbalanced sample sizes, these results are descriptive rather than conclusive. We do not observe consistent or systematic performance disparities across subgroups.

Table K.2: Subgroup-wise accuracy.

Group	Category	Top-1 Acc. (%)	# Cases
Region	Africa	0.0	2
	Americas	33.3	18
	Asia	0.0	4
	Europe	44.4	9
Sex	Male	26.5	49
	Female	42.9	42
Age	0–3	35.3	17
	4–7	12.5	8
	8–11	25.0	8
	12+	28.6	7

While these subgroup analyses are limited by small sample sizes and potential confounders, they do not reveal any clear performance disparities. However, the low accuracy in certain regions (e.g., Africa, Asia) may reflect data scarcity or distributional differences, warranting further investigation with larger and more balanced datasets. The preliminary nature of these results underscores the need for more comprehensive fairness evaluations in future work.

K.5 ACS Broader Impact

While ACS is introduced in the context of candidate-conditioned verification, its underlying formulation is more general and can be interpreted as an instance of adaptive computation allocation for structured reasoning systems.

From candidate selection to computation allocation. In the main text, ACS selects the number of candidate diseases $K(x)$ based on an input-dependent difficulty representation $\psi(x)$:

$$K(x) = \pi(\psi(x)).$$

This can be viewed more broadly as allocating a computation budget conditioned on input uncertainty. The action space need not be restricted to candidate set size; instead, $K(x)$ can be interpreted as a proxy for the amount of reasoning effort applied to a given input. Under this view, ACS implements an instance-wise policy that trades off predictive performance and computational cost:

$$\max_{\pi} \ \mathbb{E}_x\big[\mathbb{I}(y^* = y_{\text{gt}}) - \lambda \cdot \text{Cost}(\pi(\psi(x)))\big].$$

This formulation is agnostic to the specific reasoning structure and can be applied to a broader class of adaptive inference problems.

Generalization to agentic reasoning systems. Although instantiated here as candidate subset selection, the same principle extends naturally to other dimensions of agent-based reasoning. For example, the policy π can be used to control the number of reasoning rounds in iterative agent loops, the depth of verification or self-reflection steps, the number of retrieved knowledge items in retrieval-augmented pipelines, or the number of tools or sub-agents invoked during inference.

In each case, the policy determines where additional computation is most beneficial, enabling efficient allocation of reasoning resources under constrained budgets.

Relation to existing adaptive computation strategies. ACS differs from prior approaches to adaptive computation such as early-exit mechanisms or entropy-based heuristics. Early-exit methods

(e.g., BranchyNet) adjust computation depth within a single feed-forward model, while heuristic strategies rely on predefined uncertainty thresholds. In contrast, ACS operates at the level of structured reasoning, learning a data-driven policy over discrete reasoning actions (e.g., candidate expansion) and directly optimizing an explicit accuracy–cost objective. This allows ACS to adapt to task-specific uncertainty patterns that may not be captured by simple confidence measures.

Compatibility with evolving reasoning systems. ACS is orthogonal to mechanisms that improve reasoning quality over time, such as self-refinement, memory augmentation, or self-evolving agent frameworks. In such settings, ACS governs how much computation is allocated per instance at inference time, while self-evolving mechanisms improve the quality of the underlying reasoning process across instances. This separation suggests that ACS can serve as a lightweight control layer for managing computational efficiency in more complex adaptive systems.

L Limitations and Future Work

L.1 Limitations

Candidate-bounded closed-set protocol. The current study follows a candidate-bounded, closed-set protocol to enable controlled comparison under the same visual candidate set. This design isolates the downstream verification problem and makes dependence on upstream candidate recall explicit. However, it assumes that the correct disease is included among the proposed candidates, making final performance dependent on candidate proposal quality. Future deployment should combine RareVA with stronger candidate generation, open-set recognition, abstention, and calibrated uncertainty estimates.

Noisy phenotype interface. RareVA uses VLM-extracted phenotypes as the interface between visual perception and agentic reasoning. This improves interpretability, but phenotype extraction remains noisy: subtle facial traits may be missed, and some extracted terms may be imprecise after ontology normalization. RareVA mitigates this through repeated phenotype extraction, consensus aggregation, HPO normalization, and candidate-wise verification rather than direct VLM prediction, but future work should incorporate confidence-aware phenotype extraction and expert-refined phenotype annotations.

KG coverage. The disease–phenotype KG is constructed from available HPO-style disease profiles and may be incomplete for under-studied diseases. This can limit evidence retrieval for diseases with sparse phenotype annotations. In the current system, KG evidence is used for candidate-specific verification rather than as the sole source, reducing but not eliminating the impact of missing knowledge. Future work should expand the KG with curated case reports, literature-derived phenotype descriptions, and frequency-aware disease profiles.

Generalization under distribution shift. Although RareVA generalizes to an external cohort (GMDB as shown in Section 4.4), broader evaluation across more diverse clinical settings remains necessary. Real-world deployment may involve differences in imaging devices, illumination, pose, demographic composition, and disease prevalence that are not fully represented in the current benchmarks.

L.2 Future Work

Future work can extend RareVA in several directions. First, deployment-oriented evaluation should incorporate abstention, open-set recognition, calibrated uncertainty estimates, and prospective clinician-in-the-loop workflows, where the system supports candidate prioritization rather than making standalone decisions. Second, candidate proposal can be improved through stronger visual backbones, metric-learning retrieval, ensemble candidate generation, or adaptive candidate expansion when the current candidates are inconsistent with the extracted evidence. Third, the phenotype interface can be strengthened by incorporating domain-adapted phenotype extractors, confidence-weighted phenotype scoring, and expert-refined phenotype annotations. Fourth, the disease–phenotype KG can be expanded using curated case reports, literature-derived phenotype descriptions, and frequency-aware disease profiles. Finally, RareVA can be extended beyond the image-first setting by incorporating additional clinical information when available, such as age, sex, symptoms, family history, or genetic findings.

M REPRODUCIBILITY CHECK

This section summarizes the practical details required to reproduce the main experiments, including compute resources, software environment, and model configuration. All baselines and RareVA use the same fixed data splits, prompt templates, and evaluation scripts to ensure controlled comparison.

All experiments are run on a single NVIDIA A100 GPU using PyTorch and standard Python libraries for preprocessing, training, evaluation, and bootstrap analysis. VLM-based phenotype extraction and agentic reasoning use fixed model versions and decoding settings. The full RareVA pipeline requires approximately 90–150 seconds per case (mean $\sim$120 s), depending on the ACS-selected candidate budget and the number of Doctor Agent reasoning rounds, with an estimated cost of approximately \$0.10 USD per case based on Gemini-2.5-Flash API usage.

Hyperparameter development was restricted to the RDFace training and validation splits. Neither the RDFace test set nor the external GMDB benchmark was used for model selection or tuning. Trainable components use random seed 42, including data shuffling, model initialization, and ACS training. Inference uses fixed prompts and decoding configurations throughout. Detailed model, training, retrieval, and inference hyperparameters are reported in the following tables.

Table M.1: Hyperparameters for supervised classification of VisCls.

Parameter	**Value**
Optimizer	AdamW
Learning rate	$1e^{-4}$
Batch size	32
Number of epochs	50
Loss function	CrossEntropyLoss
Weight decay	$1e^{-4}$
Image size	224×224

Table M.2: Hyperparameters for VLM-based phenotype extraction.

Parameter	**Value**
VLM backbone	Qwen3-VL (main)
Decoding	temperature $= 0.5$
Top-p	1.0
Max new tokens	1024
# consensus runs	3 (semantic consensus)
Consensus rule	Majority / semantic clustering (Appendix D.2)
Phenotype normalization	Doc2HPO (raw $\rightarrow$ HPO)

Table M.3: Hyperparameters for agents in RareVA.

Parameter	**Value**
VLM backbone	Gemini-2.5-Flash (API)
Model Version	001, stable version released June 2025
Decoding	temperature $= 0.1$
Top-p	1.0
Max new tokens	1024
Doctor Agent rounds	$T_{\max} = 3$

Table M.4: Hyperparameters for ACS policy.

Parameter	**Value**
ACS target penalty	$\eta = 0.10$
ACS trade-off	$\lambda = \eta / \mathbb{E}_{\text{Val}}[\text{cost}] \approx 1.16 \times 10^{-5}$