Case2Flow: Bridging Patient Cases and Guideline Flowcharts through Multimodal Retrieval

Jiale Wei[1,2], Yufan Chen[1], Alexander Jaus[1,2], Zdravko Marinov[1,2], Julian Friedrich[3], Simon Reiß[1], Jens Kleesiek[3], Rainer Stiefelhagen[1],

[1]Karlsruhe Institute of Technology, Karlsruhe, Germany
[2]Helmholtz Information and Data Science School for Health (HIDSS4Health), Karlsruhe/Heidelberg, Germany
[3]University Hospital Essen, Essen, Germany

Correspondence: jiale.wei@kit.edu

Abstract

Medical guidelines encode rich, evidence-based decision logic, yet the specific decision artifact a clinician needs is hard to locate within a guideline, let alone across guidelines covering plausible diseases and treatments. While guideline passages have supported end-to-end question answering, flowcharts remain largely underused in decision support despite their ability to encode actionable clinical pathways. We therefore introduce *Case2Flow*, a task designed to retrieve the most relevant guideline flowchart for a given patient case from a collection of guideline documents. To support it, we construct *FlowAtlas*, a curated corpus of 202 flowcharts extracted from 2,080 medical guidelines, together with a pipeline that synthesises 1,911 aligned case-flowchart pairs. Our evaluation of multimodal retrieval methods reveals systematic failure modes, including over-reliance on keywords and spurious token-patch matches induced by uninformative background regions in flowcharts. Motivated by this, we propose *CRISP*, a training-free scoring method that sharpens late-interaction retrieval by suppressing uninformative patches, discounting ambiguous token matches, and incorporating bidirectional query-image alignment. CRISP improves Recall@1 by up to $+18.71$ pp, while a blinded physician assessment on published case narratives provides preliminary feasibility evidence beyond synthetic queries. Code and datasets are publicly available at Case2Flow.

1 Introduction

The volume of new medical guidelines is growing rapidly, with annual releases surging by up to 29%[1], enriching evidence-based practice yet also increasing the burden on clinicians who must navigate and stay current with an ever-larger body of documents. This makes it harder to put the decision support to practical use at the point of care.

One approach to making guideline knowledge accessible is training end-to-end language models such as BioBERT (Lee et al., 2020), adapted for question answering, where guidelines serve as contextual inputs (Li et al., 2025b) or training data (Li et al., 2025a,b) to ground diagnoses and treatment decisions. Retrieval-augmented methods (Wang et al., 2026; Sohn et al., 2025; Staniek et al., 2025) further link patient data with guideline passages. However, these approaches operate on unstructured text and do not address visual decision artifacts within guidelines, such as flowcharts encoding clinical pathways. Multimodal document retrieval has advanced through late-interaction models (Faysse et al., 2025; Huang and Tan, 2025), and prior work on clinical flowcharts has focused on extracting decision logic (Li et al., 2023) or benchmarking LLM agents on predefined workflows (Xiao et al., 2024), but retrieving guideline flowcharts from patient cases remains unexplored.

Rather than predicting a single *'correct'* decision, an alternative is to retrieve the most relevant decision pathways for a given patient, letting practitioners make informed choices grounded in guideline logic. We pursue this direction: specifically, we target the retrieval of flowcharts from medical guidelines, whose decision tree structures capture best practices while exposing multiple treatment pathways. Our task *Case2Flow*, shown in Figure 1, aims to find the most relevant flowchart from a large set of medical guidelines, given a patient's clinical profile, including history, symptoms, and key findings.

To enable *Case2Flow*, we first close the data gap with a synthetic generation pipeline that produces (*textual patient case*, *flowchart*) pairs. We generate 1,911 such pairs from 202 flowcharts filtered from

[1]https://www.guidelinecentral.com/insights/2024-in-review-guidelines/

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 50976–50992
October 24-29, 2026 ©2026 Association for Computational Linguistics

over two thousand candidate guidelines across four sources, spanning clinical contexts from treatment to prevention. The resulting dataset *FlowAtlas* allows us to benchmark multimodal retrieval systems on their ability to retrieve the source flowchart from a patient description. We evaluate six methods on *FlowAtlas*, investigate their failure modes, and also report results on the MedGUIDE dataset (Li et al., 2025b), originally proposed for question answering. Finally, we show that shortcomings in the visual embeddings of late-interaction models can be addressed by our training-free *CRISP* scoring, which suppresses uninformative image patches and verifies bidirectional query-image alignment. Because the benchmark itself is synthetic, we additionally assess the retrieved charts on published clinical case narratives under blinded physician rating, which tests whether the task transfers beyond our generation pipeline. Our contributions can be summarised as follows:

- We define the novel retrieval task *Case2Flow* (Case-based medical flowchart retrieval), which locates the decision artifact a clinician may wish to inspect.
- We close the data gap with a pipeline that extracts flowcharts from medical guidelines and produces flowchart-grounded synthetic patient cases, yielding *FlowAtlas* with 1,911 case-flowchart pairs.
- We benchmark existing retrieval methods on *Case2Flow*, highlight current challenges, and show how our training-free *CRISP* scoring improves Recall@1 by up to +18.71 pp.

2 Related Work

Clinical decision support and guideline retrieval. Clinical decision support has gradually moved from closed-form prediction toward systems that ground recommendations in patient records, biomedical evidence, and clinical guidelines. Retrieval-augmented generation has been used to reduce hallucination and improve the use of evidence in medical question answering and guideline-based reasoning (Wang et al., 2026; Sohn et al., 2025; Staniek et al., 2025; Li et al., 2025a). Recent oncology agents further show that language models can coordinate tools, knowledge bases, and guideline citations for complex treatment decisions (Ferber et al., 2025). These systems demonstrate the value of connecting patient information to external clinical knowledge, but most of them retrieve textual passages, produce final answers, or assume that the relevant guideline context is already available. They therefore leave open a more basic access problem: given a patient case, how can a system locate the specific decision artifact that a clinician may wish to inspect?

Multimodal document image retrieval. This access problem is closely related to document image retrieval, where recent work has moved beyond OCR pipelines toward direct visual retrieval. Benchmarks such as NL-DIR (Guo et al., 2025) and MMDocIR (Dong et al., 2025) evaluate whether models can retrieve visually rich pages or layout elements from natural language queries. Visual RAG systems such as VisRAG (Yu et al., 2025) and VDocRAG (Tanaka et al., 2025) show that representing pages as images preserves layout, figures, tables, and charts that text extraction may weaken. Late-interaction models such as ColPali (Faysse et al., 2025) further improve retrieval by matching query tokens against visual patch representations rather than collapsing the page into a single embedding. These studies, however, are primarily designed for general documents or open-domain question answering. A clinical flowchart is more than a visually rich page: its relevance depends on whether a case satisfies its branching conditions and follows its pathway semantics.

Flowchart understanding and retrieval. Prior work on medical flowcharts has mainly studied how to extract, formalise, or execute decision logic once the relevant source is known. MedDM (Li et al., 2023) and FT-MDT (Li et al., 2025c) convert medical text or guidelines into decision trees and executable guidance rules, while recent decision-tree extraction methods improve structured clinical rule construction (Hou et al., 2025). FlowBench (Xiao et al., 2024) evaluates whether agents can follow workflow knowledge during planning, and MedGUIDE benchmarks whether large language models can make guideline-consistent decisions over curated NCCN decision-tree diagrams (Li et al., 2025b). These studies are complementary to ours, but they largely assume a selected workflow, a converted tree, or a predefined decision diagram. In practice, before a clinician or system can inspect a guideline pathway, the relevant visual artifact must first be located in a heterogeneous guideline collection. Case2Flow addresses this earlier access problem by retrieving the raw guideline flowchart

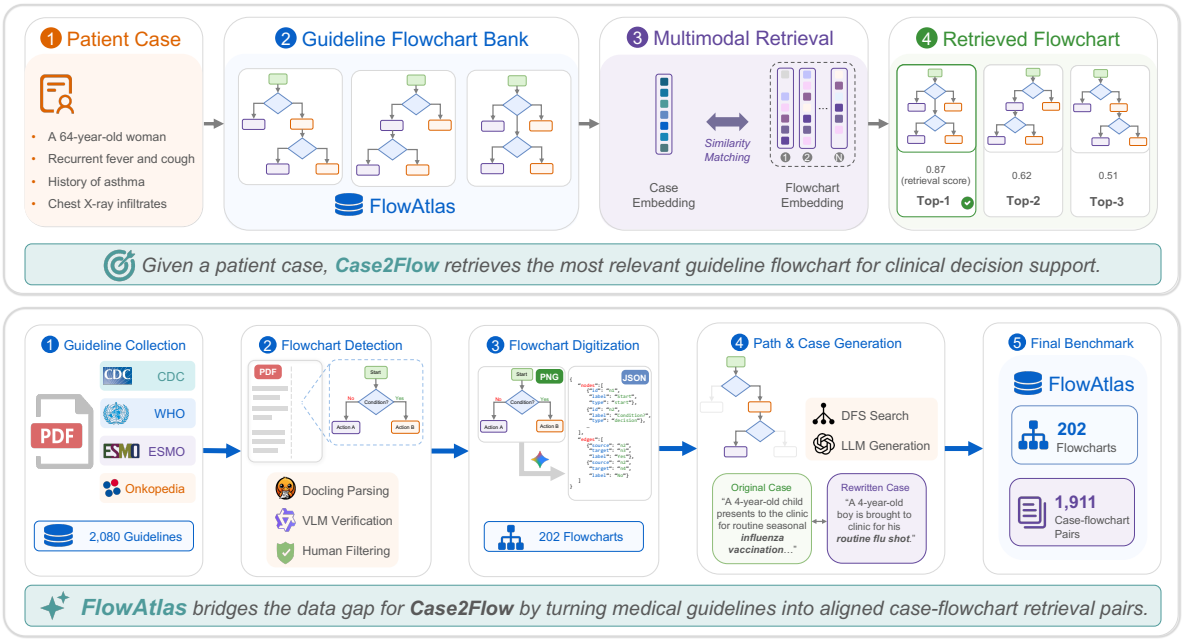

Figure 1: The task *Case2Flow*: given a textual patient case (history, symptoms, key findings), retrieve the most relevant guideline flowchart from a large cross-provider candidate pool. We build *FlowAtlas* from 2,080 guidelines by extracting and digitising 202 flowcharts and synthesising 1,911 aligned case-flowchart pairs through DFS path enumeration, case generation, and clinical rewriting.

that best matches an unstructured patient case. Our task therefore differs from executing known workflows or generating final clinical decisions, while preserving the visual form in which guideline logic is often published.

3 Methodology

3.1 *Case2Flow*: Case-based Flowchart Retrieval

The *Case2Flow* task is to identify the best-matching flowchart from a set of candidate flowcharts $\mathcal{F} = \{f_1, \ldots, f_n\}$ given a textual patient case. A query c concatenates the case description with a focused clinical question; the reference answer is withheld from the retriever. Retrieval is carried out by a scoring function θ that maps a query-flowchart pair to a real-valued score and should rank the target flowchart $f^\star \in \mathcal{F}$ first, i.e., $f^\star = \arg\max_{f_j \in \mathcal{F}} \theta(c, f_j)$. We define the best match as the flowchart whose decision logic most directly applies to the patient's clinical situation. For *FlowAtlas*, $f^\star$ is the chart containing the decision path from which the case was synthesised, fixed before retrieval. It is therefore a reproducible provenance target rather than a claim that no other chart in $\mathcal{F}$ could be clinically appropriate; Section 4.5 examines the latter question directly. MedGUIDE uses its own supplied case-chart mapping. Obtaining such pairs in practice is costly, because experts must be familiar with every flowchart in $\mathcal{F}$, potentially beyond their specialty, to determine whether (c, f_j) is the best match. We address this data gap below.

3.2 Building *FlowAtlas*

Guideline collection. We crawl 2,080 guideline PDFs from four public providers chosen for both topical breadth and stylistic diversity: CDC[2] (infectious disease, vaccination, public health), WHO[3] (global health, prevention), ESMO[4] (oncology), and Onkopedia[5] (haematology/oncology). Per-source statistics and a stylistic comparison are provided in Table 6 and Appendix B.

Flowchart detection and digitisation. Candidate flowchart pages are first localised with Docling (Auer et al., 2024), which combines caption keyword matching with its visual classifier. A Qwen3-VL (Bai et al., 2025) verifier then checks each candidate for clinical content, a valid structure of nodes and arrows, and branching logic. A final manual sweep removes residual false positives such as figures that are not decision charts or

[2] https://stacks.cdc.gov/
[3] https://www.who.int/publications
[4] https://www.esmo.org/guidelines
[5] https://www.onkopedia-guidelines.info

charts without a clear clinical purpose. Each retained chart is digitised by Gemini-2.5-Flash into a structured graph: nodes (start, decision, action, terminal) carry text labels, and labelled edges encode branch conditions.

Path enumeration and case synthesis. From each flowchart graph, we enumerate start-to-terminal decision paths via DFS with cycle detection. For each path, GPT-5-mini generates a synthetic patient case describing demographics, presenting complaints, and findings, using only information attached to nodes along the path; the prompt prohibits clinical details that do not appear in the flowchart. The same pass also emits a focused clinical question about the next decision and the reference answer read off the terminal node of the path; only the case and the question enter the retrieval query. A second LLM pass then rewrites the case and the question into natural clinical language and substitutes flowchart-specific phrasing with lay equivalents (e.g. *"influenza vaccination"* → *"flu shot"*; *"CD4 count below 200 cells/μL"* → *"severely immunodeficient"*), leaving the target flowchart and the answer unchanged.

Two evaluation modes. We retain both versions of each case and use them as parallel evaluation modes. `Original` cases share lexical material with the source flowchart and measure retrieval accuracy when surface tokens are available as a cue. `Rewritten` cases require recognition of the same clinical situation through paraphrases closer to bedside phrasing. Reporting both modes distinguishes accuracy driven by lexical overlap from robustness to clinically equivalent phrasing.

3.3 CRISP Scoring

Late-interaction VLMs such as ColPali (Faysse et al., 2025) and ColQwen3 (Huang and Tan, 2025) encode a query c into T token embeddings $\mathbf{q}_1, \ldots, \mathbf{q}_T$ and a flowchart f into S per-patch embeddings $\mathbf{p}_1, \ldots, \mathbf{p}_S$, and score relevance by MaxSim: $\mathrm{MaxSim}(c, f) = \frac{1}{T}\sum_{t=1}^{T} \max_{s\in[S]} \mathbf{q}_t^\top \mathbf{p}_s$. Inspection of per-patch similarities on clinical flowcharts (cf. Figure 3 and Figure 2) reveals two recurring sources of error. First, flowcharts contain substantial whitespace and background. Their patch embeddings cluster near a common background direction, yet their inner products with frequent query tokens can remain large enough for MaxSim to select them as best matches. Second, frequent query tokens such as *"patient"*, *"history of"*, and *"mg"* match many patches similarly. Because MaxSim takes the maximum for each token, these matches can contribute high scores that reflect token frequency rather than relevance.

CRISP (**C**ross-directional **R**efined **I**nformative **S**coring of **P**atches) is a scoring modification that addresses both effects and adds a reverse-direction coverage term. Let $\mathcal{A} = \{s \in [S] : \|\mathbf{p}_s\| > \varepsilon\}$ denote the active (non-padding) patches. The full CRISP score combines three components, abbreviated *PW*, *Adj*, and *Bi* below.

Patch weighting (PW). We compute the mean active embedding $\bar{\mathbf{p}} = \frac{1}{|\mathcal{A}|}\sum_{s\in\mathcal{A}} \mathbf{p}_s$ and weight each patch by its deviation from this mean:

$$w_s = \frac{d_s}{\sum_{s'\in\mathcal{A}} d_{s'}} \cdot |\mathcal{A}|, \quad d_s = \|\mathbf{p}_s - \bar{\mathbf{p}}\|_2, \quad (1)$$

with $w_s = 0$ for $s \notin \mathcal{A}$. Patches that carry text or decision-node content deviate from the average and receive higher weights; background patches that cluster near $\bar{\mathbf{p}}$ are suppressed. The weights average to one over active patches, preserving the overall score magnitude.

Adjusted forward scoring (Adj). Instead of taking the single best weighted patch per token, we measure how much that best match stands out from the token's weighted average over all patches:

$$a_t = \max_{s\in\mathcal{A}} w_s \cdot \mathbf{q}_t^\top \mathbf{p}_s \; - \; \frac{1}{|\mathcal{A}|}\sum_{s\in\mathcal{A}} w_s \cdot \mathbf{q}_t^\top \mathbf{p}_s \,. \quad (2)$$

A token with a sharp, distinctive patch match yields a large a_t; a token that matches everything near-uniformly produces $a_t \approx 0$ and contributes little to the score.

Bidirectional coverage (Bi). The forward term checks whether every query token finds a confident match in the flowchart, but it does not check whether the query addresses the flowchart content. We therefore add a reverse MaxSim, denoted *Bi*, in which each active patch selects its best query token, and we combine the two terms:

$$\theta_{\text{CRISP}}(c, f) = \frac{1}{T}\sum_{t=1}^{T} a_t \; + \; \frac{1}{|\mathcal{A}|}\sum_{s\in\mathcal{A}} \max_{t\in[T]} \mathbf{q}_t^\top \mathbf{p}_s \,. \quad (3)$$

The first term captures query-to-image precision; the second captures image-to-query coverage. Normalising by T and $|\mathcal{A}|$ keeps both on a comparable scale. CRISP requires no training and can replace MaxSim in late-interaction retrieval models.

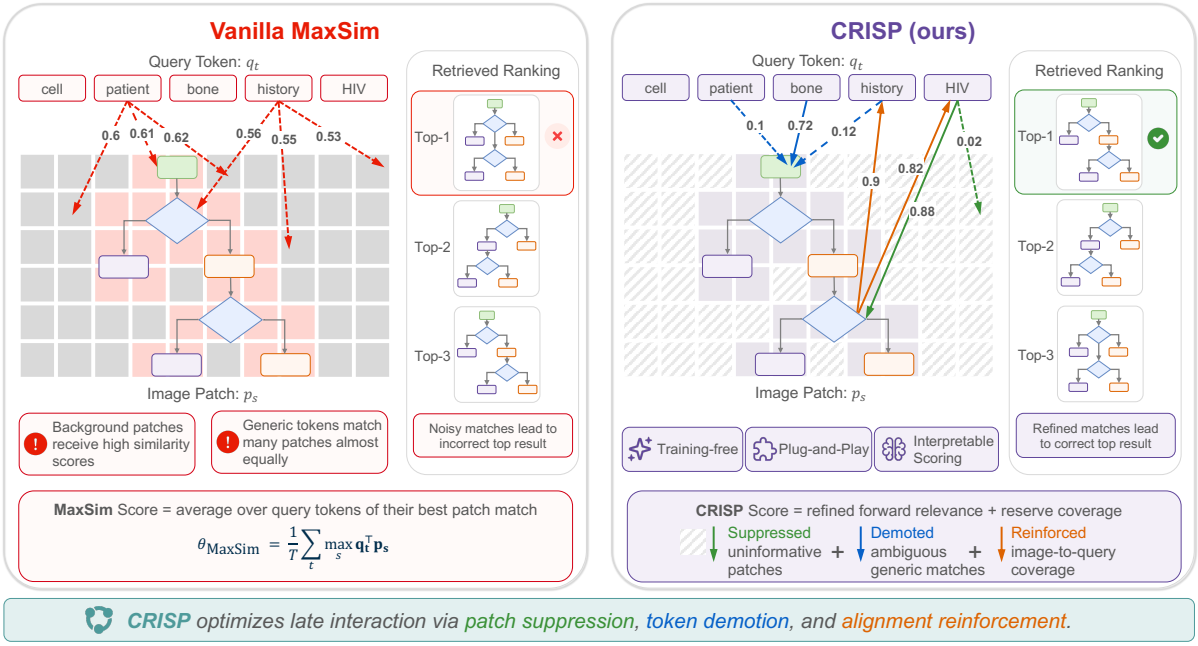

Figure 2: *CRISP* sharpens late-interaction retrieval by reducing the influence of uninformative background patches, subtracting a soft uniform-match baseline from each query token to demote generic matches, and adding a reverse image-to-query coverage term. All three components are training-free and operate only at scoring time.

4 Experiments

4.1 Setup

Settings. All models are evaluated zero-shot. We report Recall@1 under three regimes: (i) *per-source*, where the candidate pool contains only flowcharts from a single provider and isolates in-domain retrieval; (ii) *mixed*, pooling all 202 *FlowAtlas* flowcharts with the 55 flowcharts from MedGUIDE (Li et al., 2025b) into 257 candidates spanning four sources and an external benchmark style; and (iii) *external*, evaluating directly on the MedGUIDE pool with its 7,747 cases, none generated by our pipeline. In the `rewritten` Mixed setting the rewriting pass is applied only to the 1,911 FlowAtlas queries; the 7,747 MedGUIDE queries are external and carried over unchanged. Recall@$\{1, 3, 5, 10\}$ and MRR for all methods are reported in Appendix D.

Baselines. We compare two families of methods. *Textual pipelines* extract text from each flowchart and retrieve with BM25: **OCR + BM25** (Tesseract) and **VLM Captioning + BM25** (Qwen3-VL-2B captions); stronger dense and hybrid variants are added in Table 2. *Direct image retrieval* computes cross-modal similarity directly: **CLIP** (Radford et al., 2021), **Qwen3-VL-Emb** (Bai et al., 2025), **ColPali** (Faysse et al., 2025), and **ColQwen3** (Huang and Tan, 2025). For ColPali and ColQwen3 we additionally report *CRISP* (Section 3.3); since it only changes the scoring function, its embeddings are identical to the corresponding baseline.

4.2 Quantitative Results

Table 1 reports Recall@1 across all settings.

Per-source retrieval. Within a single provider, most methods clear 77% on `original` cases, and ColQwen3 reaches 96.14% on Onkopedia, indicating strong performance of modern late-interaction VLMs on familiar layouts. CLIP is a clear outlier (e.g. 13.16% on Onkopedia), reflecting how poorly its web caption-style pre-training transfers to clinical descriptions and flowcharts.

Original vs. rewritten cases. Across the 1,911 pairs, the second LLM pass yields a mean token-set Jaccard similarity of 0.44 between the two modes, with the substitutions concentrated on flowchart-specific vocabulary, and the two paradigms react differently. *Text pipelines decline substantially*: OCR + BM25 drops by −23.89 pp on Onkopedia and −8.71 pp on ESMO once lay vocabulary replaces flowchart phrasing. *Strong VLMs are more robust*: ColQwen3 and Qwen3-VL-Emb lose at most about 5 pp per source. ColPali sits between the two, in line with its smaller backbone, which is also where CRISP brings the largest gains.

Method	Size	Type	CDC	WHO	ESMO	Onko.	NCCN	Mixed
RANDOM	–	–	7.69	1.92	1.12	2.08	1.82	0.39
Textual Pipeline-Based Flowchart Retrieval								
Tesseract OCR + BM25	–	`Original`	83.82	79.55	77.65	88.70	50.59	62.55
		`Rewritten`	75.00	72.16	68.94	64.81	N/A	58.55
VLM Captioning + BM25	2B	`Original`	**89.71**	84.66	80.56	86.12	71.28	71.83
		`Rewritten`	83.82	76.42	75.00	67.38	N/A	68.62
Direct Image-Based Flowchart Retrieval								
CLIP	0.4B	`Original`	45.59	50.57	17.68	13.16	18.86	15.17
		`Rewritten`	45.59	46.59	13.76	10.01	N/A	14.73
Qwen3-VL-Emb	8B	`Original`	80.88	85.80	82.83	92.70	79.24	74.39
		`Rewritten`	76.47	82.10	79.17	87.70	N/A	73.05
ColPali	3B	`Original`	82.35	86.08	80.93	80.40	54.67	55.48
		`Rewritten`	82.35	76.70	68.81	61.66	N/A	52.09
ColQwen3	4B	`Original`	85.29	86.65	86.36	96.14	79.91	78.44
		`Rewritten`	83.82	82.67	82.95	91.70	N/A	76.93
ColPali + *CRISP* (Ours)	3B	`Original`	85.29	85.80	82.07	83.69	73.38	70.98
		`Rewritten`	**85.29**	74.72	69.44	69.96	N/A	68.23
ColQwen3 + *CRISP* (Ours)	4B	`Original`	82.35	**87.22**	**87.88**	**96.85**	**84.85**	**83.69**
		`Rewritten`	82.35	**83.24**	**83.96**	**92.13**	N/A	**81.88**

Table 1: Recall@1 (%) across guideline sources and query types. *Per-source*: candidate pool from one provider. *NCCN*: external MedGUIDE (Li et al., 2025b) flowcharts. *Mixed*: all 257 flowcharts pooled. MedGUIDE supplies its own case-chart pairs and has no rewriting pass, so the `Rewritten` row is not applicable (N/A) in the NCCN column. Best results per column in **bold**.

External benchmark. On the external NCCN-based MedGUIDE dataset, ColQwen3 (79.91%) and Qwen3-VL-Emb (79.24%) lead, while the best textual pipeline, VLM captioning, reaches 71.28%. These trends mirror those on FlowAtlas, suggesting that the observed difficulty is not restricted to cases generated by our pipeline.

Cross-source mixed pool. With all 257 flowcharts pooled, retrieval becomes substantially harder: ColQwen3 drops from 85–96% per source to 78.44%, and ColPali from 80–86% to 55.48%. This setting is closest to deployment, where the relevant flowchart is rarely known to live in a single provider's pool.

CRISP. Applying CRISP as a drop-in scoring change improves both ColPali and ColQwen3 across nearly every setting, with the largest gains in the harder, larger-pool regimes. On ColPali, mixed-pool Recall@1 improves by +15.50 pp (55.48 → 70.98) and NCCN by +18.71 pp (54.67 → 73.38). On ColQwen3, mixed improves by +5.25 pp (78.44 → 83.69) and NCCN by +4.94 pp (79.91 → 84.85). The larger improvement on ColPali is consistent with its weaker baseline being more affected by uniform weighting and background patches; ColQwen3 appears to mitigate part of this effect through stronger patch embeddings (cf. Table 3). On the small CDC pool (13 candidates), CRISP slightly lowers ColQwen3 (85.29 → 82.35): a net loss of two correct top-1 predictions over 68 cases, better read as noise at this sample size than as a reliable regression, but marking the regime in which CRISP has least to offer. Across all other settings, ColQwen3 + CRISP achieves the best Recall@1, surpassing the 2× larger Qwen3-VL-Emb. A cluster bootstrap confirms that the ColQwen3 Recall@1 gains are significant in all three settings (Appendix F). On Mixed, CRISP costs 1.26× the scoring time on ColQwen3 and 1.49× on ColPali, with under 0.2% additional peak GPU memory.

Alternative pipelines. Comparing visual retrieval against BM25 alone leaves open whether the gap reflects the modality or merely an under-tuned text pipeline, so Table 2 varies both how the text is matched and what text is matched over. Dense retrieval with BGE-M3 (Chen et al., 2024) over the same OCR output is not by itself an improvement (56.74 against 62.55 for BM25 on `original` Mixed), but their equal-weight hybrid is, gaining

Method	Mixed		NCCN
	orig	rew	orig
OCR + BGE-M3	56.74	54.23	57.45
OCR hybrid	70.25	66.90	68.41
Graph-text + BGE-M3	55.50	53.46	59.06
Graph-text hybrid	67.60	65.82	70.12
OCR hybrid $\rightarrow$ BGE-reranker	51.34	49.53	54.30
OCR hybrid + MaxSim	77.15	74.35	77.73
OCR hybrid + *CRISP*	80.48	77.68	81.18
ColQwen3 (MaxSim)	78.44	76.93	79.91
ColQwen3 + *CRISP*	**83.69**	**81.88**	**84.85**

Table 2: Recall@1 (%) of stronger text, reranking, and fusion baselines on the full query sets. *Hybrid*: equal-weight fusion of BM25 and BGE-M3 scores after per-query z-normalisation; *graph-text*: the digitised flowchart graph serialised instead of OCR text.

7.7 and 13.5 pp over the two respectively and leading every other text-only configuration on Mixed. Serialising the digitised flowchart graph instead of the OCR text trades one weakness for another, winning on NCCN (70.12 versus 68.41) and losing on Mixed (67.60 versus 70.25): the structure our pipeline recovers is not what the text channel is missing. Taking the best text-only configuration in each setting, the gap to ColQwen3 + CRISP still stands at 13.4 to 15.0 pp, so the multimodal advantage is not an artifact of a weak text baseline.

Two further ways of adding capacity on the text side fail as well. A text cross-encoder (`bge-reranker-v2-m3`) (Chen et al., 2024) applied zero-shot to the hybrid top-10 lowers Recall@1 by 14 to 19 pp while leaving Recall@10 unchanged by construction: the shortlist it receives already contains the correct chart in over 93% of queries, and it reorders that shortlist worse than the first-stage scores do. A domain mismatch is the likely cause, since the checkpoint is trained on natural-language passage pairs whereas flowchart OCR yields fragmentary node labels; we report it for this checkpoint and setup only, not for rerankers trained on the task. Equal-weight score fusion of the OCR hybrid with ColQwen3 likewise degrades the visual scorer in all six settings, by 1.3 to 2.6 pp under MaxSim and by 3.2 to 4.2 pp under CRISP. The larger degradation under CRISP is consistent with the rest of this section: sharpening the visual score leaves the text channel less to add.

A multimodal second stage, by contrast, does help, and is complementary to CRISP rather than an alternative to it. Reranking the ColQwen3 top-10 with Qwen3-VL-Reranker-8B on a fixed sample of 500 queries per setting, stratified by provider and initial-rank bucket, raises Recall@1 from 78.60 to 91.20 (`original` Mixed), 76.40 to 90.00 (`rewritten` Mixed), and 80.00 to 94.60 (NCCN). These are paired with MaxSim on the identical sample and are not comparable to the full-set numbers in Tables 1 and 2; they show that a much larger cross-encoder recovers a large part of the residual error, at a cost far above that of a scoring-time change.

4.3 Ablation

Table 3 reports all combinations of the three CRISP components on MedGUIDE. Each component is enabled on top of the MaxSim baseline, and we report Recall@1 and MRR on both backbones. *Adj* is ablated independently by setting all active-patch weights to one in Eq. (2), so no combination is tied to the presence of *PW*.

Bi is the dominant single component. Adding *Bi* alone yields the largest single-step gain on both backbones, with +10.18 pp R@1 on ColPali and +2.68 pp on ColQwen3. This accounts for roughly half of the full CRISP improvement: 54% of the R@1 lift on each backbone, and 56% (ColPali) and 62% (ColQwen3) of the MRR lift. The image-to-query direction therefore provides the largest missing signal in vanilla MaxSim, and reverse MaxSim recovers part of it.

PW depends on the backbone's patch representations. *PW* alone lifts ColPali by +7.64 pp R@1 but only +0.86 pp on ColQwen3. ColQwen3 already L_2-normalises patch embeddings and explicitly masks non-image tokens during scoring, so the active-patch variation available to *PW* is much smaller. On the weaker, unmasked ColPali backbone, *PW* down-weights background patches that MaxSim would otherwise score highly.

Max-minus-mean needs the reverse term. *Adj* is not a standalone improvement: alone it gains +3.26 pp R@1 on ColPali but loses 3.61 pp on ColQwen3, and *PW* + *Adj* loses 4.67 pp on ColQwen3 relative to the baseline. Replacing per-token MaxSim with $\max - \text{mean}$ removes absolute similarity magnitude in favour of contrast, which on its own is too noisy to rank flowcharts. Reintroducing the reverse term restores the absolute-magnitude signal in the image-to-query direction, and *Adj* then becomes the largest contributor: *Adj* + *Bi* exceeds *Bi* alone by +7.22 pp R@1 on

ColPali and $+1.66$ pp on ColQwen3, and exceeds *PW + Bi* on both.

The components interact rather than simply add. The full factorial shows the components to be complementary but not independently monotone. On ColQwen3 the single-component gains sum to -0.07 pp R@1 while the full method reaches $+4.94$ pp; on ColPali the same combination is sub-additive ($+21.08$ summed versus $+18.71$ achieved), because *PW* and *Adj* partly address the same background-dominance effect, and once *Adj + Bi* is in place *PW* adds only $+1.31$ and $+0.60$ pp. That the pattern holds across a weak and a strong backbone indicates that CRISP addresses a limitation of MaxSim aggregation rather than of a specific encoder.

	ColPali		ColQwen3	
Scoring variant	R@1	MRR	R@1	MRR
MaxSim (baseline)	54.67	66.91	79.91	87.55
+ PW	62.31	72.86	80.77	88.25
+ Adj	57.93	68.70	76.30	84.49
+ Bi	64.85	75.65	82.59	89.74
+ PW + Adj	59.27	69.69	75.24	84.12
+ PW + Bi	67.76	78.47	83.93	90.63
+ Adj + Bi	72.07	81.16	84.25	90.58
+ **CRISP** (PW+Adj+Bi)	**73.38**	**82.55**	**84.85**	**91.08**
Δ vs. MaxSim	*+18.71*	*+15.64*	*+4.94*	*+3.53*

Table 3: All combinations of the CRISP components on MedGUIDE (7,747 cases, 55 flowcharts; Recall@1 and MRR in %). *PW*: patch weighting; *Adj*: adjusted forward scoring; *Bi*: bidirectional coverage.

4.4 Behavioural Analysis

We examine two questions about CRISP's effect on the Mixed pool: how it changes the per-query rank of the ground truth, and which errors remain after it is applied (Table 4).

Per-query rank changes. CRISP improves the rank of the ground-truth flowchart for 14.5% to 39.9% of queries and worsens it for only 4.0% to 6.3%, depending on the backbone and mode. For most queries (53.8% to 81.4%), the rank is unchanged. Among queries that MaxSim ranks below the first position, CRISP recovers 33.5% to 40.9% to rank 1 across all four combinations of model and mode. The mean gain when CRISP helps also exceeds the mean loss when it hurts ($+16$ versus -2 positions on ColPali, $+3$ versus -2 on ColQwen3), so CRISP concentrates its effect on difficult, incorrectly ranked queries rather than broadly reshuffling rankings.

	ColPali		ColQwen3	
	`orig`	`rew`	`orig`	`rew`
Per-query rank change vs. MaxSim (%)				
Improved	37.4	39.9	14.5	15.2
Unchanged	57.1	53.8	81.4	80.2
Hurt	5.4	6.3	4.0	4.6
Recovered to rank 1†	40.9	39.6	35.8	33.5
Top-1 errors remaining after CRISP				
Number of errors	2,803	3,068	1,575	1,750
Same document (%)	3.7	4.1	5.1	6.1
Same provider (%)	51.4	49.4	67.4	62.2
Cross-provider (%)	44.9	46.4	27.5	31.7

Table 4: Behavioural analysis of CRISP on the Mixed pool. *Top*: per-query rank change of the ground-truth flowchart relative to MaxSim; † Recovery is a fraction of the queries MaxSim ranks below top-1, not of all queries. *Bottom*: the top-1 errors remaining under CRISP, classified by the deepest level at which the prediction matches the ground truth.

Where the residual errors lie. When CRISP retrieves the wrong flowchart, we ask whether the prediction is clinically unrelated or a topical neighbour of the ground truth. Every flowchart has three nested levels of identity: a *provider*, a *guideline document* within that provider's corpus, and a *chart* within that document. We partition each residual top-1 error by the deepest level at which the prediction matches the ground truth: *same document* (a different chart in the same guideline, typically another stage or line of treatment), *same provider* (a different document from the same body on a related subject), and *cross-provider*.

Table 4 (bottom panel) reports this breakdown on the Mixed pool. Among the 1,575 ColQwen3+CRISP residual errors with `original` queries, only 27.5% are cross-provider; 67.4% retrieve a different guideline document on an overlapping topic and 5.1% a different chart within the same document. The same-provider class dominates every combination of model and mode and is concentrated in the NCCN subset, which contributes 95% of these errors and contains many oncology flowcharts differing only in tumour subtype or staging. Separately, a recurring CDC confusion occurs among three seasonal-influenza vaccination guidelines whose charts share most of their text. ColPali shows the same pattern with more cross-provider errors (44.9%), consistent with its weaker baseline ranking. Most remaining errors are therefore topical neighbours rather than unrelated charts,

matching the steep Recall@k curve in Appendix D: when the top-1 result is wrong, the correct chart is usually within a short shortlist.

4.5 Clinical Validation on Published Cases

FlowAtlas measures whether a system recovers the chart a case was generated from, which is reproducible but not the same as whether that chart is clinically usable; we therefore also evaluate on real clinical narratives under a clinician's judgement.

Top-1 system	**Accept.** (≥ 2)	**Applic.** ($= 3$)
ColPali (MaxSim)	9/50 (18%)	6/50 (12%)
+ *CRISP*	13/50 (26%)	6/50 (12%)
ColQwen3 (MaxSim)	22/50 (44%)	14/50 (28%)
+ *CRISP*	**24/50** (48%)	14/50 (28%)

Table 5: Blinded physician rating of the top-1 chart on 50 published PMC-CaseReport narratives (Wu et al., 2025), scored 0 (irrelevant) to 3 (directly applicable).

Protocol. We draw 50 article-deduplicated case contexts from PMC-CaseReport (Wu et al., 2025), a corpus of published case reports. Candidates are screened for library coverage by a text and title matching step independent of the four evaluated systems, and the sample is frozen before any system runs; the systems receive only the published case context. The 133 deduplicated case-chart pairs so returned are rated by one physician, blinded to system identity and to the screening information, from 0 (irrelevant) to 3 (directly applicable); multiple or no candidates may be acceptable.

Results. Table 5 reports the outcome. ColQwen3+CRISP returns an acceptable chart for 48% of cases and a directly applicable one for 28%, ahead of both ColPali variants, and the ordering of the four systems matches their Recall@1 ordering on FlowAtlas. Across the union of all returned candidates, 31 of 50 cases received at least one acceptable chart and 19 received none. Two limits apply: it rates only the returned candidates, so it measures what the systems surface rather than whether a suitable chart exists elsewhere in the library; and the MaxSim-CRISP differences are within rating noise here, so it supports the feasibility of the task rather than the specific gain from CRISP.

4.6 Qualitative Results

Figure 3 visualises per-patch similarity on a representative CDC flowchart with ColPali. Under MaxSim, empty page background receives activations comparable to the decision nodes and a few generic tokens drive the score; CRISP suppresses those patches and concentrates similarity on decision nodes and clinical-term tokens, reflecting the two artifacts that motivated its design. The pattern holds across providers, and Appendix E adds one heatmap per CRISP component.

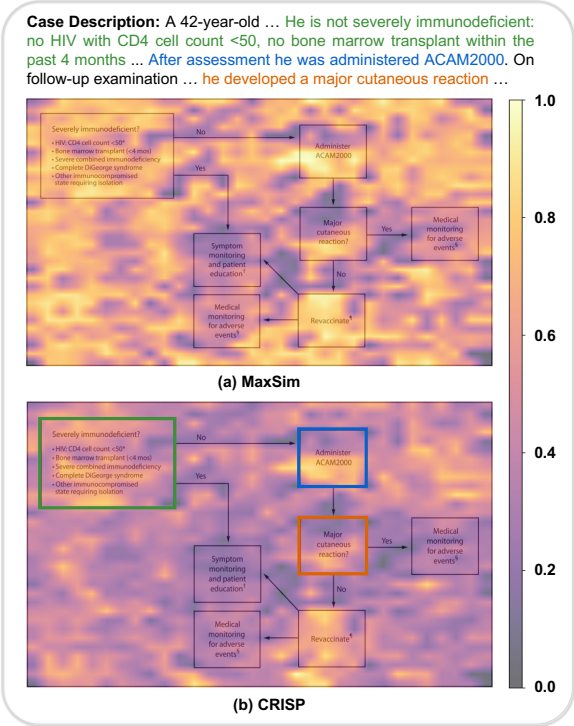

Figure 3: Per-patch similarity heatmaps on a CDC flowchart (ColPali). Top: standard MaxSim assigns high scores almost uniformly. Bottom: CRISP suppresses background and highlights decision nodes.

5 Conclusion

We introduced *Case2Flow*, the task of retrieving the guideline flowchart that applies to a patient case, and *FlowAtlas*, 1,911 case-flowchart pairs built over four guideline sources. Benchmarking six retrieval methods exposed a failure mode of late-interaction scoring, in which uninformative background patches attract spurious matches; our training-free *CRISP* scoring corrects it, improving Recall@1 by up to +18.71 pp at a small scoring-time cost. A blinded physician assessment on published case narratives provides preliminary feasibility evidence beyond our synthetic queries, while leaving clear headroom for future work on real clinical notes.

Limitations

Synthetic cases. FlowAtlas cases are LLM-generated from the labels along enumerated decision paths and then rewritten by a second LLM pass into more natural clinical phrasing. The pipeline is grounded in the flowcharts and remains consistent with the encoded decision logic, but it does not reproduce the noise, ambiguity, and incomplete information of real clinical notes. We mitigate this limitation in three ways: we report both `Original` and `Rewritten` modes, whose difference provides a proxy for paraphrase robustness; we evaluate on the external MedGUIDE benchmark, where all methods show trends consistent with those observed on FlowAtlas; and we assess retrieved charts on published clinical case narratives (Section 4.5). That assessment has its own limits: a single rater, 50 cases screened for library coverage and concentrated in oncology (no CDC or WHO targets), and labels on the returned candidates only. It is feasibility evidence, not an unconditional estimate of real-world accuracy. Validation on retrospective clinical notes requires governed clinical data and is beyond the scope of this study.

Closed-set evaluation. We evaluate retrieval on a closed pool of 202 FlowAtlas flowcharts (and 55 MedGUIDE flowcharts in the Mixed and external regimes). This matches a deployment in which an institution curates and updates a guideline library, but it does not measure how methods scale to thousands of candidates or behave when no relevant chart exists in the pool. CRISP's reverse-coverage term provides a per-query coverage-based score that could be evaluated for abstention, although we do not calibrate or evaluate that use here.

Scope of CRISP. CRISP is a scoring modification on top of late-interaction VLMs and does not alter the underlying embeddings, so the headroom it can recover is bounded by what the backbone already represents. When the backbone fails to encode a piece of clinical content (for example, a dosage rendered as a small footnote), CRISP cannot recover it, because the relevant patch embedding lacks the information. The small CDC regression on ColQwen3 (Section 4) is better treated as noise at this sample size; it occurs in the regime where CRISP has the least headroom—a small pool and a strong backbone. As a practical rule, the settings in which CRISP helps most are those with many competing candidates and a backbone that leaves headroom; we do not have evidence for a pool-size threshold at which it should be disabled, and we do not propose an adaptive switch between MaxSim and CRISP. CRISP is therefore complementary to backbone improvements rather than a replacement.

Source and language coverage. FlowAtlas covers four guideline providers and one external benchmark, all in English. The four sources span layout style (text-heavy vs. diagrammatic), topical scope (public health vs. oncology), and provider type (national agency, international body, professional society, specialist consortium), but are not exhaustive. Eligibility filtering also affects the providers very unevenly ($776 \rightarrow 13$ for CDC, $1{,}159 \rightarrow 52$ for WHO, $91 \rightarrow 89$ for ESMO, $54 \rightarrow 48$ for Onkopedia), so the corpus is concentrated in oncology: ESMO and Onkopedia together contribute 137 of the 202 charts. National-specialty guidelines outside oncology and non-English guidelines are not represented. The construction pipeline is provider-agnostic and released in full to facilitate extensions to further providers and languages.

Ethical Considerations

Data and consent. FlowAtlas is built from publicly accessible guideline documents (CDC, WHO, ESMO, Onkopedia) and the publicly released MedGUIDE flowcharts (Li et al., 2025b). The patient cases are synthesised from flowchart labels and contain no real patient records or personally identifying information. The clinical validation in Section 4.5 uses published case narratives from PMC-CaseReport (Wu et al., 2025); we redistribute case identifiers, not article content.

Human assessment. The blinded rating in Section 4.5 was carried out by one clinician among the authors. No external annotators were recruited or compensated and no human-subject data were collected, so ethics board approval was not required.

Intended use and release. FlowAtlas and CRISP are released for research on case-based guideline retrieval. Retrieving a flowchart is an access aid, not a clinical recommendation, and Section 4.5 shows the top-1 chart to be unusable in roughly half of the assessed cases, so the resources are not suitable for automated clinical decision-making without expert oversight. We release the pipeline, prompts, and code under an explicit licence with

a per-component rights table, and provide reconstruction scripts where source documents may not be redistributed.

Acknowledgments

The present contribution is supported by the Helmholtz Association under the joint research school "HIDSS4Health – Helmholtz Information and Data Science School for Health". This work was performed on the HoreKa supercomputer, funded by the Ministry of Science, Research and the Arts Baden-Württemberg and by the Federal Ministry of Education and Research.

References

Christoph Auer, Maksym Lysak, Ahmed Nassar, Michele Dolfi, Nikolaos Livathinos, Panos Vagenas, Cesar Berrospi Ramis, Matteo Omenetti, Fabian Lindlbauer, Kasper Dinkla, Lokesh Mishra, Yusik Kim, Shubham Gupta, Rafael Teixeira de Lima, Valery Weber, Lucas Morin, Ingmar Meijer, Viktor Kuropiatnyk, and Peter W. J. Staar. 2024. Docling technical report. Technical Report arXiv:2408.09869, IBM Research.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.

Kuicai Dong, Yujing Chang, Derrick Goh Xin Deik, Dexun Li, Ruiming Tang, and Yong Liu. 2025. MM-DocIR: Benchmarking multimodal retrieval for long documents. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 30971–31005, Suzhou, China. Association for Computational Linguistics.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. 2025. ColPali: Efficient document retrieval with vision language models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025*. OpenReview.net.

Dyke Ferber, Omar S. M. El Nahhas, Georg Wölflein, Isabella C. Wiest, Jan Clusmann, Marie-Elisabeth Leßmann, Sebastian Foersch, Jacqueline Lammert, Maximilian Tschochohei, Dirk Jäger, Manuel Salto-Tellez, Nikolaus Schultz, Daniel Truhn, and Jakob Nikolas Kather. 2025. Development and validation of an autonomous artificial intelligence agent for clinical decision-making in oncology. *Nature Cancer*, 6(8):1337–1349.

Hao Guo, Xugong Qin, Jun Jie Ou Yang, Peng Zhang, Gangyan Zeng, Yubo Li, and Hailun Lin. 2025. Towards natural language-based document image retrieval: New dataset and benchmark. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 29722–29732.

Ruihui Hou, Xiaojun Wang, Weiyan Zhang, Zhexin Song, Kai Wang, Yifei Chen, Jingping Liu, and Tong Ruan. 2025. Decision tree extraction for clinical decision support system with if-else pseudocode and PlanSelect strategy. *IEEE Journal of Biomedical and Health Informatics*, 29(5):3642–3653.

Xin Huang and Kye Min Tan. 2025. Beyond text: Unlocking true multimodal, end-to-end RAG with tomoro ColQwen3. Model card, `TomoroAI/tomoro-colqwen3-embed-4b`.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Binbin Li, Tianxin Meng, Xiaoming Shi, Jie Zhai, and Tong Ruan. 2023. MedDM: LLM-executable clinical guidance tree for clinical decision-making. *arXiv preprint arXiv:2312.02441*.

Haolin Li, Tianjie Dai, Zhe Chen, Siyuan Du, Jiangchao Yao, Ya Zhang, and Yanfeng Wang. 2025a. RAD: Towards trustworthy retrieval-augmented multi-modal clinical diagnosis. In *Advances in Neural Information Processing Systems 38 (NeurIPS 2025)*.

Xiaomin Li, Mingye Gao, Yuexing Hao, Taoran Li, Guangya Wan, Zihan Wang, and Yijun Wang. 2025b. MedGUIDE: Benchmarking clinical decision-making in large language models. *arXiv preprint arXiv:2505.11613*.

Yuheng Li, Jiechao Gao, Wei Han, Wenwen Ouyang, Wei Zhu, and Hui Yi Leong. 2025c. FT-MDT: Extracting decision trees from medical texts via a novel low-rank adaptation method. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 65–76, Suzhou, China. Association for Computational Linguistics.

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. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Jiwoong Sohn, Yein Park, Chanwoong Yoon, Sihyeon Park, Hyeon Hwang, Mujeen Sung, Hyunjae Kim, and Jaewoo Kang. 2025. Rationale-guided retrieval augmented generation for medical question answering. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 12739–12753, Albuquerque, New Mexico. Association for Computational Linguistics.

Michael Staniek, Artem Sokolov, and Stefan Riezler. 2025. Training and evaluation of guideline-based medical reasoning in LLMs. *arXiv preprint arXiv:2512.03838*.

Ryota Tanaka, Taichi Iki, Taku Hasegawa, Kyosuke Nishida, Kuniko Saito, and Jun Suzuki. 2025. VDocRAG: Retrieval-augmented generation over visually-rich documents. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24827–24837.

Ziyue Wang, Junde Wu, Linghan Cai, Chang Han Low, Xihong Yang, Qiaxuan Li, and Yueming Jin. 2026. MedAgent-Pro: Towards evidence-based multi-modal medical diagnosis via reasoning agentic workflow. In *The Fourteenth International Conference on Learning Representations, ICLR 2026*. OpenReview.net.

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

Ruixuan Xiao, Wentao Ma, Ke Wang, Yuchuan Wu, Junbo Zhao, Haobo Wang, Fei Huang, and Yongbin Li. 2024. FlowBench: Revisiting and benchmarking workflow-guided planning for LLM-based agents. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10883–10900, Miami, Florida, USA. Association for Computational Linguistics.

Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, and Maosong Sun. 2025. VisRAG: Vision-based retrieval-augmented generation on multi-modality documents. In *The Thirteenth International Conference on Learning Representations, ICLR 2025*. OpenReview.net.

A FlowAtlas Statistics

Table 6 reports per-source counts for *FlowAtlas* and for the external MedGUIDE pool used in the *Mixed* setting. Path length denotes the number of decision nodes along the start-to-terminal path used to seed each synthetic case. Orig. / Rew. Jaccard is the token-level Jaccard similarity between the two query modes, computed from sets of lower-cased alphabetic tokens; lower values indicate more surface rewriting.

Source	Guide.	Flow.	Cases	Path (avg)	Words (o/r)	Jacc.
CDC	776	13	68	5.3	113/102	0.45
WHO	1,159	52	352	6.3	114/104	0.44
ESMO	91	89	792	5.8	103/103	0.44
Onkopedia	54	48	699	7.7	106/105	0.45
FlowAtlas	**2,080**	**202**	**1,911**	6.4	109/104	0.44
MedGUIDE	N/A	55	7,747	N/A	N/A	N/A
Mixed	N/A	**257**	**9,658**	N/A	N/A	N/A

Table 6: FlowAtlas dataset statistics per guideline source. *Guide.*: number of crawled guideline PDFs; *Flow.*: retained flowcharts; *Cases*: synthesised case-flowchart pairs; *Path*: mean number of decision nodes per case; *Words*: mean word count of (case + question) in `original/rewritten` mode; *Jacc.*: mean token-level Jaccard similarity between the two modes. MedGUIDE is included for the Mixed and external evaluations but is not part of FlowAtlas.

B Per-Source Characteristics

The four FlowAtlas sources differ in layout style, caption convention, and topic mix, and these differences affect retrieval difficulty. Figure 4 shows one representative chart from each of the five candidate pools. CDC is text-light and prescriptive. Its candidate pool is small and most methods approach saturation, leaving little room for CRISP to help. WHO and ESMO cover heterogeneous topics with a mix of flow-diagram and table-like layouts; ColPali experiences its largest within-pool drops on these two sources. Onkopedia is oncology-specialised and pharmacology-dense, and has the longest decision paths in FlowAtlas (up to 20 nodes). ColQwen3+CRISP still reaches 96.85% within this pool, suggesting that long, specialty-specific paths are not the primary limitation. MedGUIDE is an external benchmark of cancer-staging flowcharts whose tabular layouts differ markedly from those of the four FlowAtlas sources; it is also where CRISP improves ColPali the most ($+18.71$ pp R@1). The *Mixed* pool combines all five sources and requires methods to discriminate among layout styles simultaneously. The spread between methods is widest in this setting, which most closely approximates cross-provider deployment.

C FlowAtlas Construction Pipeline

Guideline collection. We collect $2{,}080$ guideline PDFs from four public providers (CDC, ESMO, Onkopedia, and WHO) and exclude non-English documents and non-guideline artifacts. The four corpora vary in size and topical scope by design: CDC and WHO cover broad, population-oriented topics, whereas ESMO and Onkopedia are oncology-specialised.

Flowchart detection and digitisation. Candidate flowchart pages are first localised with Docling (Auer et al., 2024), which combines caption keyword matching (`{algorithm, flowchart, flow chart, pathway, decision tree, schema}`) with its visual layout classifier. Pages without a caption match but with a high visual-classifier score are also kept as candidates. Each candidate is then passed to a Qwen3-VL-32B (Bai et al., 2025) verifier with a structured three-criteria prompt: (i) clinical content, (ii) a valid structure of nodes and arrows, and (iii) at least one branch. A final manual sweep by the authors removes residual false positives such as tabular figures, organisational charts, and charts without a clear clinical purpose. The three-stage cascade yields 202 retained flowcharts; most filtering occurs at the verification stage rather than during detection. Each retained chart is digitised by Gemini-2.5-Flash into a JSON graph with typed nodes (*start*, *decision*, *action*, *terminal*) and labelled edges that carry the branch condition. Charts whose digitised graph fails a connectivity check are discarded.

Path enumeration and case synthesis. We enumerate start-to-terminal decision paths in each graph via depth-first search with cycle detection. Long charts can produce hundreds of paths; we cap the number per chart at 40 and prefer paths that visit a previously unused node or edge in order to maintain case diversity. For each path, GPT-5-mini writes a synthetic patient case strictly grounded in the labels along the path. The system prompt restricts the model to demographics, history, pre-

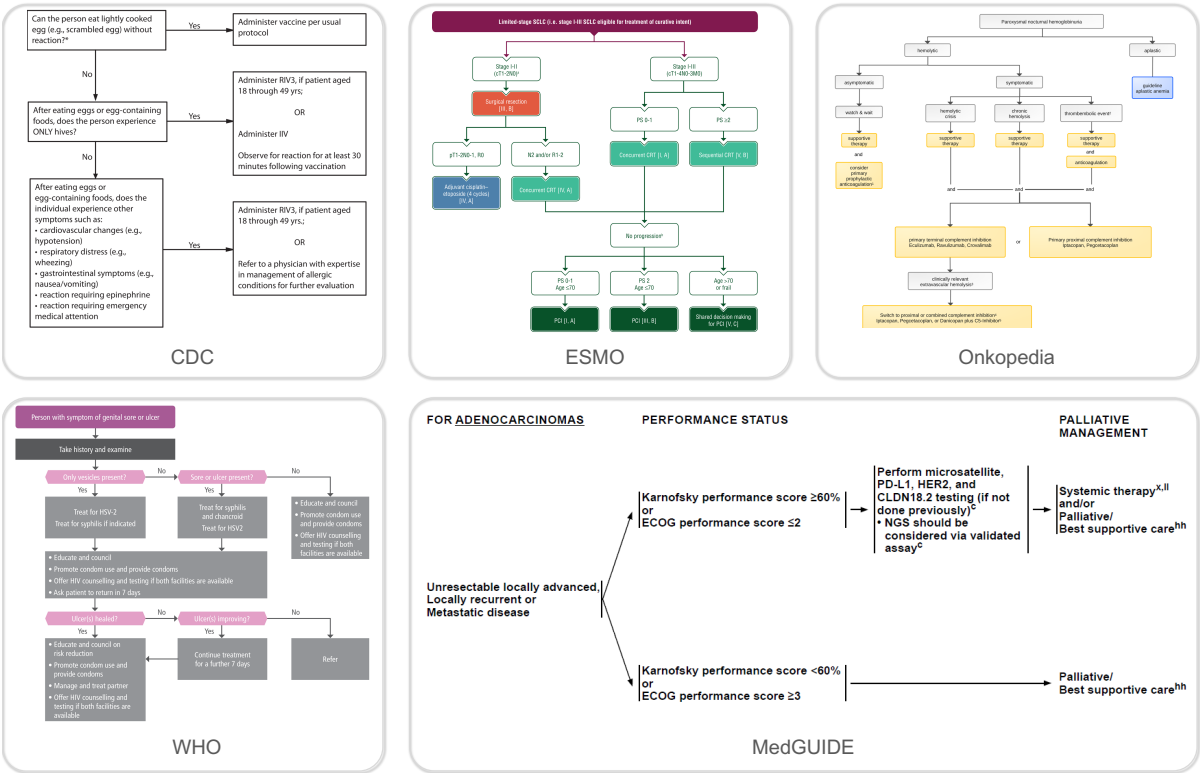

Figure 4: One representative flowchart from each of the five candidate pools (CDC, ESMO, Onkopedia, WHO and MedGUIDE).

senting complaints, and findings that appear on the path; it prohibits clinical details not present in the chart (lab values, comorbidities, treatments). A second pass then rewrites each case into natural clinical phrasing and substitutes flowchart vocabulary with lay equivalents (e.g., *"influenza vaccination"* → *"flu shot"*; *"CD4 count below 200 cells/μL"* → *"severely immunodeficient"*). The two passes are kept separate so that the rewritten case preserves the same clinical situation while reducing measurable surface overlap (mean Jaccard 0.44; Table 6).

Post-generation filtering. Among the FlowAtlas cases we discard (i) empty or truncated outputs from either LLM pass, (ii) cases with fewer than 30 tokens of clinical content, and (iii) cases whose rewritten and original versions are byte-identical, since these would collapse the two evaluation modes. MedGUIDE cases are not passed through the rewriting pass and are therefore exempt from (iii). The remaining 1,911 pairs form the FlowAtlas evaluation set; per-source counts and word statistics are reported in Table 6.

Post-generation validation. Manual review is applied at a single point in the pipeline: the sweep over detected flowchart pages described above, which decides chart eligibility. Graph connectivity checking, path enumeration, case synthesis, and the post-generation filtering above are fully automated, and the generated cases were not clinically adjudicated. Clinical judgement enters the study only in the separate assessment of Section 4.5.

Reproducibility. All four prompts (flowchart verification, graph extraction, case synthesis, and clinical rewriting) are released. LLM calls use temperature 1.0 for case generation (to encourage stylistic diversity) and temperature 0.3 for verification and rewriting (where determinism matters more). The rewriting pass also passes a fixed seed to the API, which makes regenerated cases close to the released ones but does not guarantee bit-exact reproduction.

D Per-Source and Mixed-Pool Results

For both query modes on the five candidate pools (four FlowAtlas sources and the external MedGUIDE pool), Tables 7 and 8 report per-source Recall@k ($k \in \{1, 3, 5, 10\}$) and MRR. Table 9 provides the corresponding numbers on the *Mixed* pool that combines all 257 flowcharts. Two patterns hold across the tables. First, CRISP improves

R@1 and MRR more than R@5 or R@10, sharpening the top of the ranking rather than reshuffling lower-ranked results. Second, R@10 is already high for the strongest models (above 97% for ColQwen3+CRISP on the Mixed pool in both query modes). The central challenge is therefore placing the correct flowchart at rank 1 rather than retrieving a relevant shortlist, which is precisely the objective that CRISP targets.

E Visualising the CRISP Score

The main-text heatmap (Figure 3) contrasts MaxSim with CRISP, but it combines the three CRISP terms (Section 3.3) into a single "before" and "after" view. Figure 5 expands the comparison on the same chart and case by isolating one component at a time, allowing the contributions reported in the ablation table (Table 3) to be inspected on the patch grid.

(a) MaxSim. This panel renders $\max_t \mathbf{q}_t^\top \mathbf{p}_s$ on the image, the per-patch quantity that the token-side MaxSim baseline aggregates and that also appears as the (unweighted) second term of Eq. (3). On the CDC chart the map is bright across whitespace strips and connector arrows, with similarity scores comparable to those of the decision boxes themselves. This is precisely the saturation failure mode that motivates the patch-weighting and adjusted-forward terms.

(b) Patch weights w_s. This panel renders the patch-weight map w_s from Eq. (1); it is query-independent and depends only on the L_2 distance between each patch embedding and the active-patch mean $\bar{\mathbf{p}}$. High-weight regions trace the text-bearing decision and outcome nodes (*Severely immunodeficient?*, *Administer ACAM2000*, *Major cutaneous reaction?*, *Revaccinate*), whereas uniform whitespace lies near $\bar{\mathbf{p}}$ and receives little weight. The contrast between (a) and (b) illustrates the geometric prior that PW contributes before any query token is matched, and it provides a spatial account of the +7.64 pp R@1 improvement attributed to PW on ColPali in Table 3.

(c) Patch weights + adjusted forward. The adjusted-forward term in Eq. (2) is per-token rather than per-patch: each query token t contributes the contrast $a_t = \max_s w_s \, \mathbf{q}_t^\top \mathbf{p}_s - \frac{1}{|\mathcal{A}|} \sum_s w_s \, \mathbf{q}_t^\top \mathbf{p}_s$. We render it spatially by attributing a_t to the patch at which the maximum is achieved, so at most T patches carry a non-zero value and the resulting map is sparse by construction. The credited patches sit on the path-specific tokens that distinguish the case from generic clinical text: *ACAM2000* in the administration node, *cutaneous reaction* in the adverse-event branch, and the immunodeficiency-related tokens in the entry box. On its own this term is too noisy to rank flowcharts (cf. Table 3); when combined with the reverse term, it contributes per-token contrast that PW alone cannot provide.

(d) Complete CRISP. The full score in Eq. (3) combines the per-token forward term in (c) with the unweighted reverse term in (a), each normalised by T or $|\mathcal{A}|$. For a single image-side proxy, we render $w_s \cdot \max_t \mathbf{q}_t^\top \mathbf{p}_s$, which keeps the same convention as (a) and (b) and shows how the weights from (b) reshape the coverage map from (a): whitespace activations are suppressed while decision-node activations are preserved. The closed-form score additionally accumulates the per-token contrast in (c), but that contribution is a scalar per candidate and has no per-patch signature of its own.

F Statistical Reliability and Cost

Significance. Queries are not independent, since many are generated from different decision paths through the same flowchart, so we resample *ground-truth flowcharts* as clusters (255 on Mixed, 53 on NCCN) in a paired bootstrap over 10,000 replicates. The 95% intervals for the ColQwen3 Recall@1 improvement are $[2.35, 8.29]$, $[2.03, 8.07]$, and $[1.24, 8.85]$ pp on `original` Mixed, `rewritten` Mixed, and NCCN. None includes zero; an exact paired McNemar test gives $p < 10^{-40}$ in all three, with CRISP correcting 602–747 queries that MaxSim ranks wrongly against 220–269 in the opposite direction.

Cost. CRISP reuses the backbone embeddings, so its cost is confined to scoring. On one RTX 2080 Ti, scoring 9,658 Mixed queries against 257 candidates from cached embeddings takes 101.2 versus 151.0 s for ColPali (1.49×, +5.2 ms per query) and 222.3 versus 280.0 s for ColQwen3 (1.26×, +6.0 ms), with peak GPU memory rising by under 0.2% (266.2 → 266.7 and 355.2 → 355.7 MiB). Embedding generation, OCR, and model loading are identical under both scoring functions and excluded.

Method	CDC					WHO					ESMO				
	R@1	R@3	R@5	R@10	MRR	R@1	R@3	R@5	R@10	MRR	R@1	R@3	R@5	R@10	MRR
`Original` *queries*															
OCR + BM25	83.82	95.59	100.00	100.00	90.00	79.55	91.48	93.18	93.18	85.52	77.65	87.75	89.14	91.16	83.12
VLM Cap. + BM25	89.71	94.12	100.00	100.00	92.89	84.66	95.45	96.59	96.59	90.10	80.56	90.03	90.28	91.54	85.30
CLIP	45.59	83.82	92.65	98.53	65.68	50.57	76.42	88.92	95.74	65.94	17.68	26.89	32.20	43.06	26.08
Qwen3-VL-Emb	80.88	**98.53**	**100.00**	**100.00**	89.58	85.80	95.17	96.59	96.59	90.47	82.83	91.54	92.42	93.43	87.38
ColPali	82.35	97.06	**100.00**	**100.00**	89.46	86.08	95.17	96.02	96.02	90.41	80.93	90.28	91.67	91.79	85.71
+ *CRISP*	85.29	95.59	**100.00**	**100.00**	90.56	85.80	**96.31**	**98.01**	**99.72**	91.33	82.07	90.91	91.92	92.42	86.57
ColQwen3	**85.29**	97.06	**100.00**	**100.00**	**91.18**	86.65	95.17	96.59	96.59	90.94	86.36	91.92	92.93	94.07	89.59
+ *CRISP*	82.35	94.12	**100.00**	**100.00**	88.97	**87.22**	95.17	96.59	96.59	**91.37**	**87.88**	**94.07**	**94.19**	**94.32**	**91.09**
`Rewritten` *queries*															
OCR + BM25	75.00	92.65	100.00	100.00	84.04	72.16	87.22	90.06	92.05	80.16	68.94	78.79	82.83	87.37	75.28
VLM Cap. + BM25	83.82	94.12	100.00	100.00	90.12	76.42	91.19	94.60	96.31	84.23	75.00	84.85	86.87	89.02	80.54
CLIP	45.59	80.88	92.65	100.00	65.96	46.59	69.32	85.23	92.61	60.97	13.76	20.20	23.99	33.84	21.00
Qwen3-VL-Emb	76.47	**97.06**	**100.00**	**100.00**	86.52	82.10	94.89	96.59	97.44	88.52	79.17	91.04	92.42	93.56	85.17
ColPali	82.35	**97.06**	**100.00**	**100.00**	89.22	76.70	90.91	94.32	95.17	84.16	68.81	82.20	86.74	89.77	76.61
+ *CRISP*	**85.29**	95.59	**100.00**	**100.00**	**91.05**	74.72	94.60	**96.88**	**99.15**	84.49	69.44	85.23	89.14	90.78	78.00
ColQwen3	83.82	**97.06**	**100.00**	**100.00**	90.44	82.67	94.03	96.02	96.31	88.21	82.95	91.54	91.92	93.43	87.58
+ *CRISP*	82.35	**97.06**	**100.00**	**100.00**	89.22	**83.24**	**94.32**	96.31	96.59	**88.76**	**83.96**	**93.56**	**93.81**	**94.19**	**88.72**

Table 7: Per-source Recall@$\{1, 3, 5, 10\}$ and MRR (%) on CDC, WHO and ESMO. Best score per column among the direct-image rows in **bold**. Continued in Table 8.

Method	Onkopedia					MedGUIDE				
	R@1	R@3	R@5	R@10	MRR	R@1	R@3	R@5	R@10	MRR
`Original` *queries*										
OCR + BM25	88.70	93.85	96.42	98.14	91.92	50.59	68.40	76.20	88.00	62.49
VLM Cap. + BM25	86.12	92.27	93.56	95.28	89.69	71.28	89.14	92.89	96.48	81.02
CLIP	13.16	21.89	28.18	37.63	22.29	18.86	35.52	46.44	62.68	32.46
Qwen3-VL-Emb	92.70	96.57	98.86	**100.00**	95.16	79.24	94.45	98.36	99.56	87.10
ColPali	80.40	87.84	90.56	92.85	85.19	54.67	74.62	82.66	89.14	66.91
+ *CRISP*	83.69	90.56	93.28	97.14	88.13	73.38	90.59	93.93	97.17	82.55
ColQwen3	96.14	99.28	99.71	**100.00**	97.65	79.91	94.44	97.81	99.68	87.55
+ *CRISP*	**96.85**	**99.43**	**100.00**	**100.00**	**98.16**	**84.85**	**97.39**	**99.46**	**99.95**	**91.08**
`Rewritten` *queries*										
OCR + BM25	64.81	79.40	85.41	91.42	73.83	N/A	N/A	N/A	N/A	N/A
VLM Cap. + BM25	67.38	85.26	88.84	92.70	77.27	N/A	N/A	N/A	N/A	N/A
CLIP	10.01	19.31	23.89	33.91	19.06	N/A	N/A	N/A	N/A	N/A
Qwen3-VL-Emb	87.70	93.56	96.57	99.14	91.49	N/A	N/A	N/A	N/A	N/A
ColPali	61.66	72.53	76.11	80.26	68.76	N/A	N/A	N/A	N/A	N/A
+ *CRISP*	69.96	76.11	78.68	84.98	75.05	N/A	N/A	N/A	N/A	N/A
ColQwen3	91.70	97.42	98.71	**100.00**	94.67	N/A	N/A	N/A	N/A	N/A
+ *CRISP*	**92.13**	**98.14**	**99.00**	**100.00**	**95.18**	N/A	N/A	N/A	N/A	N/A

Table 8: Per-source Recall@$\{1, 3, 5, 10\}$ and MRR (%) on Onkopedia and the MedGUIDE pool. MedGUIDE does not provide separate rewritten cases, so its rewritten rows are marked N/A. Continued from Table 7.

Method	Original queries					Rewritten queries				
	R@1	R@3	R@5	R@10	MRR	R@1	R@3	R@5	R@10	MRR
OCR + BM25	62.55	77.74	83.23	88.75	71.71	58.55	74.64	80.90	86.85	68.40
VLM Cap. + BM25	71.83	87.71	91.20	93.86	80.50	68.62	85.69	89.86	92.97	78.07
CLIP	15.17	28.32	35.78	46.02	25.48	14.73	27.57	34.96	45.05	24.85
Qwen3-VL-Emb	74.39	90.33	93.95	96.75	82.90	73.05	89.98	93.69	96.57	82.03
ColPali	55.48	73.75	80.61	86.43	66.22	52.09	71.00	78.44	84.71	63.36
+ *CRISP*	70.98	86.46	90.00	93.30	79.52	68.23	84.23	87.91	91.69	77.16
ColQwen3	78.44	91.83	95.15	97.91	85.72	76.93	91.26	94.68	97.64	84.68
+ *CRISP*	**83.69**	**95.30**	**97.62**	**99.01**	**89.75**	**81.88**	**94.64**	**97.25**	**98.85**	**88.56**

Table 9: Recall@$\{1, 3, 5, 10\}$ and MRR (%) on the *Mixed* pool (257 candidates, 9,658 queries per mode).

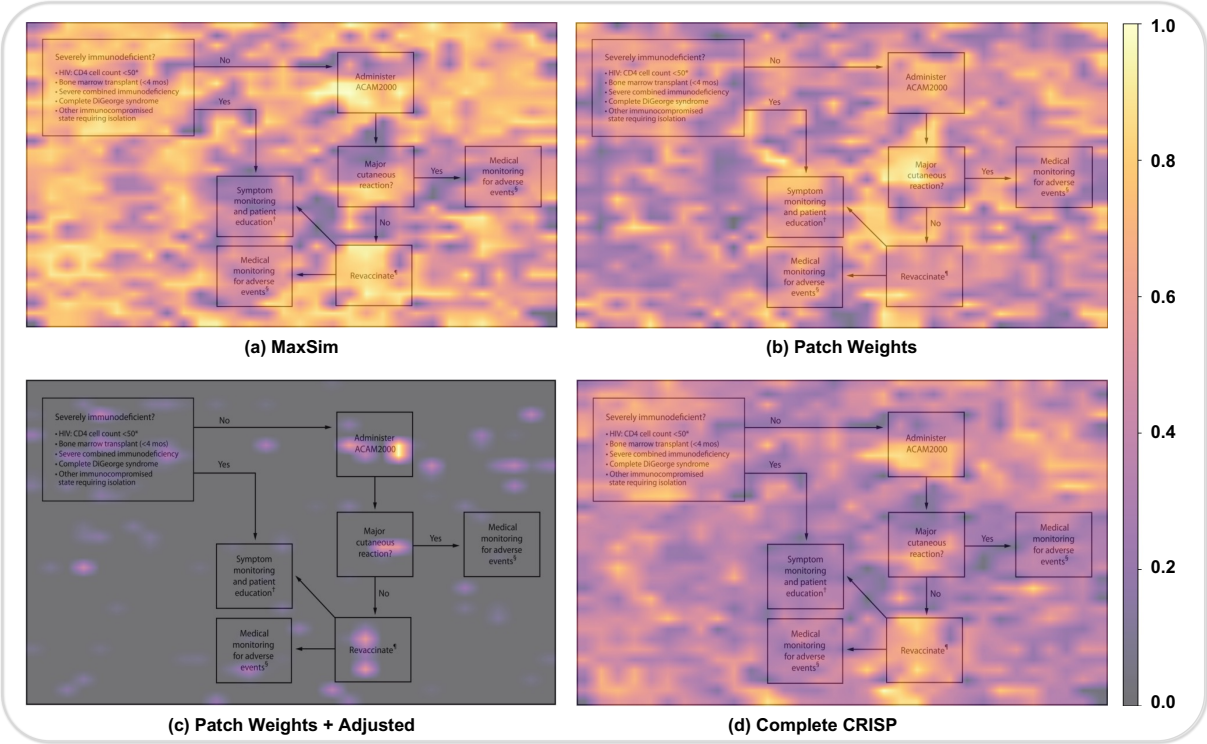

Figure 5: Per-patch decomposition of CRISP; each panel isolates one ingredient from the ablation table (Table 3). **(a) MaxSim**: $\max_t \mathbf{q}_t^\top \mathbf{p}_s$ saturates over whitespace. **(b) Patch weights**: w_s from Eq. (1); query-independent, concentrated on decision and outcome boxes. **(c) Patch weights + Adjusted**: per-token contrast a_t from Eq. (2) credited to its $\arg\max_s$ patch; sparse by construction. **(d) Complete CRISP**: image-side proxy $w_s \cdot \max_t \mathbf{q}_t^\top \mathbf{p}_s$ for Eq. (3); the additive reverse term contributes a scalar per candidate with no separate per-patch signature. Each panel is normalised to its minimum and maximum values independently.