COCAGENT: CHAIN-OF-CAUSALITY MULTI-AGENT REASONING FOR EVIDENCE-BASED MEDICAL DECISION MAKING

Anonymous authors

ABSTRACT

Strabismus subtype diagnosis requires integrating ocular alignment across gaze positions, deviation patterns, and clinical criteria, yet existing deep learning and vision-language models learn direct mappings from images to predictions, with reasoning that remains implicit and explanations weakly grounded in clinical evidence. We propose **COCAgent** (**C**hain-**O**f-**C**ausality Agent), an multi-agent framework for evidence-based strabismus subtype diagnosis. COCAgent links multimodal ophthalmic evidence through a Chain of Causality: it constructs subtype-specific causal chains, verifies candidate diagnoses against structured clinical criteria, and generates the final diagnostic report and subtype prediction, while a verification mechanism filters out unsupported or inconsistent reasoning. Such causality-grounded diagnosis requires data that expose causal structure, which no existing strabismus dataset provides. We therefore further construct **StrabBench**, a clinical benchmark comprising the photograph of the nine cardinal positions of gaze, subtype annotations, doctor-in-the-loop diagnostic reports, and task-specific causal graphs with node-level analysis. On StrabBench, COCAgent achieves 92% weighted F1, 96.1% BERTScore, and 83.4% CosSim, improving both subtype prediction and report quality over existing approaches.

1 INTRODUCTION

Strabismus, defined as any form of binocular misalignment (Repka et al., 2018; Qiu et al., 2026), affects 0.8%–6.0% of the population (Robaei et al., 2006; Zhu et al., 2015; McKean-Cowdin et al., 2013; Friedman et al., 2009) and arises from imbalances in extraocular muscle function or disruptions in neural control. Although traditionally considered a pediatric condition (Chia et al., 2007), it can severely impair visual development in children, making early screening and timely intervention critical for preventing amblyopia and binocular vision impairment.

For automated diagnostic systems, however, accuracy alone is not enough: a strabismus system is expected to identify the subtype and produce a structured report with supporting evidence, which clinicians rely on for verification and treatment planning. Recent vision-language models (VLMs) perform strongly on medical image understanding (Google, 2026; Dubey et al., 2024; Bai et al., 2025; OpenAI, 2025; Zhang et al., 2024), yet they operate as black boxes whose outputs are difficult to audit and cannot yield such evidence-grounded reports. Clinicians, by contrast, examine the photograph of the nine cardinal positions of gaze in a characteristic order—Position 5 first, then Positions 2 and 8 (upgaze and downgaze) to expose A/V patterns, and finally the remaining positions—whereas a traditional VLM produces an incorrect subtype with hallucinated findings (Figure 1). Recent efforts incorporate external knowledge, domain-specific supervision, and region-level visual grounding (Xia et al., 2025; Wang et al., 2026b; Ghosh et al., 2025; Wang et al., 2025a; Yang et al., 2025; Holland et al., 2025; Shi et al., 2025), but retrieved knowledge or improved alignment alone does not guarantee that the evidence actually supports the diagnostic hypothesis.

Chain-of-thought (CoT) reasoning has improved VLM problem solving (Yao et al., 2023; Jiang et al., 2026), and causal reasoning has extended it from association toward intervention and counterfactual analysis (Bhasuran et al., 2026; Vashishtha et al., 2026; Yu et al., 2025; Li et al., 2025a), inspiring structured reasoning for medical diagnosis (Jiang et al., 2025a; Lai et al., 2026; Liu et al., 2024; Qiu et al., 2025; Bao et al., 2025; Li et al., 2026). These methods equip medical VLMs with free-text

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.

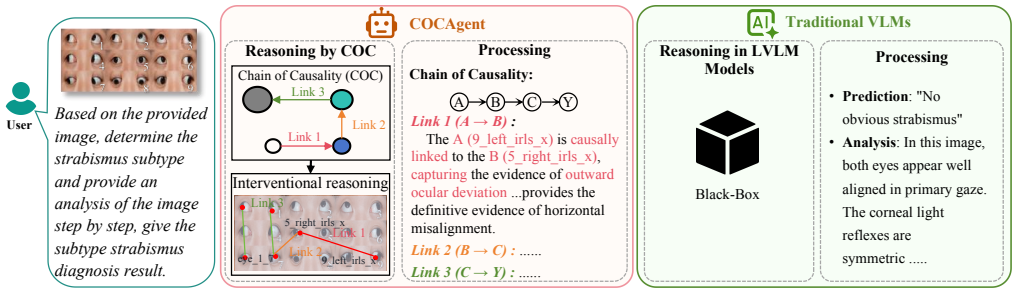

Figure 1: Comparison of reasoning over the photograph of the nine cardinal positions of gaze between COCAgent and traditional VLMs. COCAgent reasons along Chain-of-Causality, examining the abnormal evidence at gaze positions. Traditional VLMs answers in a black-box and produces an incorrect subtype with hallucination.

rationales via distillation, reinforcement learning, or structured search, but three limitations remain: the reasoning is confined to text and rarely grounded in explicit visual evidence (Parcalabescu & Frank, 2025; Liu et al., 2026a), leaving rationales prone to hallucination; it proceeds in a single forward pass without external tools or verification, so intermediate errors propagate unchecked; and the rationale is often post hoc rather than faithful, describing the decision instead of driving it. Knowledge-graph-augmented reasoning retrieves query-relevant triples, paths, or subgraphs as context (Baek et al., 2023; Luo et al., 2024; Li et al., 2025b; Wang et al., 2025b; Liu et al., 2026b; Zhang et al., 2026; Zhou et al., 2026), explores the graph as an agent (Sun et al., 2024; Chen et al., 2024; Shi et al., 2026), or integrates graph structure into learning (Chen et al., 2025; Luo et al., 2026; Yan et al., 2026), yet evidence is still selected by semantic relevance or graph connectivity, with no guarantee that it forms a causally coherent chain supporting the final diagnosis.

To address these limitations, we propose **COCAgent** (**C**hain-**O**f-**C**ausality Agent), an evidence-based multi-agent framework that transforms VLM-assisted strabismus diagnosis from black-box generation into a causality-chained, knowledge-constrained, and evidence-based decision-making process. COCAgent mirrors this clinical examination paradigm: its Chain of Causality examines the gaze positions sequentially, each link conditioning the next, rather than answering in a single forward pass (Figure 1). Building such a system, however, requires data that expose causal structure rather than labels alone, and no existing strabismus dataset provides causal graphs or evidence-level analysis. We therefore further construct **StrabBench**, a benchmark that pairs each case with subtype annotations, diagnostic reports, and task-specific causal graphs with node-level analysis.

Our technical contributions can be summarized as follows:

(1) We propose COCAgent, an evidence-based multi-agent framework that reformulates VLM-assisted strabismus diagnosis from black-box generation into a chain-of-causality, knowledge-constrained, and decision making process.

(2) We develop a causal-graph-based Chain-of-Causality construction method that localizes subtype-specific evidential paths by integrating causal relationships with class-specific feature contributions and combines these paths with clinical criteria for evidence-grounded diagnostic reasoning.

(3) We construct StrabBench, a fine-grained strabismus benchmark comprising photographs of the nine cardinal positions of gaze, subtype annotations, doctor-in-the-loop diagnostic reports, and task-specific causal graphs with node-level analysis.

(4) Extensive experiments on StrabBench show that COCAgent achieves 92% weighted F1, 96.1% BERTScore, and 83.4% CosSim, improving both fine-grained diagnostic performance and report quality over existing approaches.

2 Related Work

Causal and Structured Reasoning for Medical Decision Making. Medical decision support is moving from direct prediction toward explicit reasoning: KARE and Med-R1 add knowledge-graph retrieval and reinforcement learning to medical VLM reasoning (Jiang et al., 2025a; Lai et al., 2026); causal evaluation along Pearl's ladder (Bhasuran et al., 2026), executable counter-

factuals (Vashishtha et al., 2026), causal sufficiency and necessity (Yu et al., 2025), and multimodal counterfactual refinement (Li et al., 2025a); and structured methods shape the reasoning trajectory, from Tree of Thoughts to MedCoT, Med-CoE, Clinical Graph of Thoughts, TumorChain, and M3CoTBench (Yao et al., 2023; Liu et al., 2024; Qiu et al., 2025; Bao et al., 2025; Li et al., 2026; Jiang et al., 2026). None, however, verifies that each intermediate hypothesis is both reasoning-consistent and supported by case-specific multimodal evidence. **Evidence-Based and Explainable Vision-Language Models.** Medical VLMs increasingly rely on external knowledge and explicit grounding: retrieval and guideline-coupled tool use (Xia et al., 2025; Wang et al., 2026b) provide context without establishing hypothesis-specific support from the patient's data; grounding-oriented work exposes weak prediction–evidence links (Parcalabescu & Frank, 2025; Liu et al., 2026a) and strengthens description-, region-, and supervision-level alignment (Ghosh et al., 2025; Wang et al., 2025a; Yang et al., 2025; Holland et al., 2025; Shi et al., 2025), but none verifies that the identified evidence actually supports the diagnostic hypothesis. **KG-Augmented Reasoning.** Knowledge graphs (KGs) complement the parametric knowledge of large language models (LLMs) with explicit, traceable relational evidence, retrieved as query-relevant triples, paths, or subgraphs (Baek et al., 2023; Luo et al., 2024; Li et al., 2025b; Wang et al., 2025b; Liu et al., 2026b; Zhang et al., 2026; Zhou et al., 2026), explored by agent-style LLMs (Sun et al., 2024; Chen et al., 2024; Shi et al., 2026), or injected into training (Chen et al., 2025; Luo et al., 2026; Yan et al., 2026). All select evidence by semantic relevance or connectivity, not by whether it forms a causally coherent chain for the target decision, so irrelevant or spurious evidence can enter clinical reasoning unchecked. COCAgent instead organizes retrieved evidence into an explicit causal chain and verifies each intermediate finding against the final subtype decision.

3 StrabBench: A Strabismus Assessment Benchmark

Existing strabismus datasets are often limited in scale and lack sample-level evidence structure and clinically grounded diagnostic reports. To fill this gap, we construct StrabBench, a strabismus assessment benchmark built on clinical data from 1,065 patients collected at a university-affiliated eye center. The construction pipeline comprises standardized photography and a doctor-in-the-loop annotation and report refinement process (Figure 2, upper panel).

Photography Protocol and Collection. We establish a standardized clinical photography protocol, with trained personnel collecting images under consistent patient positioning, observation conditions, and gaze targets. For each patient, photographs are captured at the nine cardinal positions of gaze and composited into a single image I, preserving the gaze information required for fine-grained strabismus assessment.

Doctor-in-the-Loop Annotation. Senior ophthalmologists review the photographs of the nine cardinal positions of gaze and provide fine-grained subtype annotations. For report drafting, we compared three LVLMs as annotation assistants: Gemini-3-Flash-Preview (Google, 2026), GPT-5.2 (OpenAI, 2025), and Qwen3-VL (Bai et al., 2025). Consequently, we adopted Gemini-3-Flash-Preview, whose drafts the ophthalmologists judged to be the most clinically acceptable. The model generates a preliminary diagnostic report for each case, which the ophthalmologists review and revise; this doctor-in-the-loop process iterates until the report is clinically acceptable.

Evidence and Node-Level Analysis. Beyond the image, subtype label, and diagnostic report, each case in StrabBench is paired with a sample-level causal graph. The key abnormal gaze positions identified in the diagnostic report are extracted as graph nodes. For each case, every selected position is associated with a specific analysis describing the abnormal finding at that gaze position and explaining how it supports the subtype diagnosis. To our knowledge, StrabBench is the first strabismus dataset that exposes sample-level evidence structure, enabling research on evidence-based, causality-grounded diagnosis rather than label prediction alone.

Benchmark Composition and Distribution. The resulting benchmark contains 1,065 patients across six clinically defined subtypes: exotropia, A-pattern exotropia, V-pattern exotropia, esotropia, V-pattern esotropia, and vertical strabismus. The study was approved by the Ethics Committee of the data-collection institution and registered with the Chinese Clinical Trial Registry; the approval and registration numbers are withheld to preserve anonymity and will be disclosed in the camera-ready version. A waiver of informed consent was granted because the study used a retrospective cohort of de-identified clinical images. More details are given in the supplementary material.

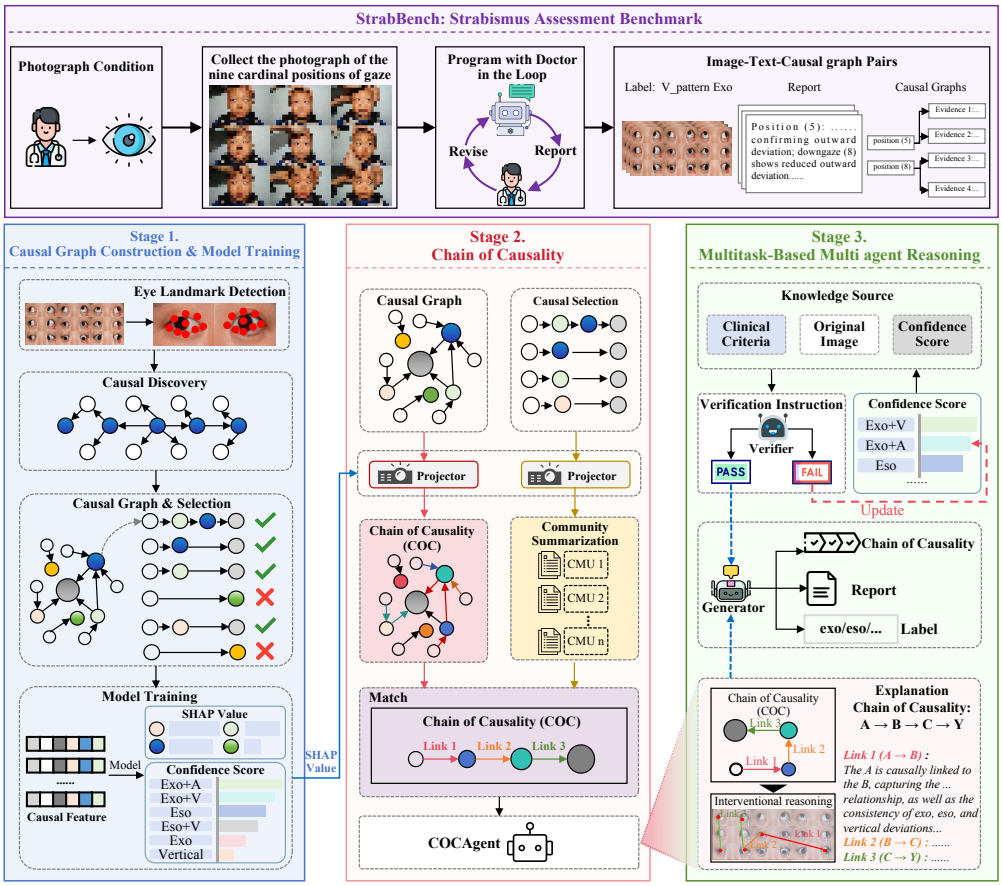

Figure 2: Overview of StrabBench (upper panel): photograph collection, doctor-in-the-loop annotation, and the image-text-causal graph pairs. COCAgent (lower panel): Stage 1 constructs a causal graph over features extracted from the photograph and trains a classifier; Stage 2 builds candidate Chains of Causality with two selection strategies and matches them against Causal Memory Units (CMUs); Stage 3 verifies the candidate diagnosis with clinical criteria R and generates the report.

4 Chain-Of-Causality Agent: An Evidence-Based Multi-Agent Framework

COCAgent (Figure 2) targets the three gaps identified in Section 1: rationales ungrounded in explicit visual evidence, single-pass generation without verification, and explanations that describe rather than drive the decision. Three sequential steps close one gap each: (1) causal graph construction and model training ground reasoning in explicit, clinically meaningful features and their causal relations; (2) chain-of-causality construction lets the evidence chain drive the diagnosis; (3) verification against clinical criteria gates the final report. Formally, for each patient, the photograph of the nine cardinal positions of gaze is acquired under the Hirschberg test and composited into a single image I. Eye landmarks are detected from I and encoded as a feature matrix:

$$\mathbf{X} \in \mathbb{R}^{9\times N}, \tag{1}$$

where 9 is the number of positions and N is the number of causally selected features (N=20 for the A/V-pattern task and N=24 for the deviation-type task). The candidate label space is $\mathcal{Y}$={*esotropia*, *V-pattern esotropia*, *exotropia*, *A-pattern exotropia*, *V-pattern exotropia*, *vertical strabismus*}.

4.1 Step 1: Causal Graph Construction and Classifier Training

Clinicians reason over well-defined measurements, not raw pixels, and a rationale is auditable only if built from such quantities. Step 1 converts the photograph of the nine cardinal positions of gaze into landmark features, from the training population, which of them causally support each subtype;

the resulting graph is the structural prior for later reasoning, defining which evidence may support prediction. The classifier trained on these features yields the three outputs downstream reasoning needs: a candidate diagnosis, a confidence-ranked candidate list, and per-feature Shapley values.

Causal Graph Construction. To identify the features that causally contribute to diagnosis, we perform causal discovery over the extracted feature set and the candidate predictions Y. We select the HCM algorithms (Li et al., 2022; Colombo et al., 2014) and employ a greedy algorithm for causal graph construction. The resulting a causal graph is defined as:

$$G = (V, E), \quad V = \{X_1, \ldots, X_m, Y\}, \quad E \subseteq V \times V, \tag{2}$$

where X_m denotes the m-th causal feature and each $(X_i, X_j) \in E$ is a directed edge $X_i \to X_j$, indicating that X_i is discovered as a direct cause of X_j. G characterizes the discovered causal relationships among the extracted features and the candidate predictions Y (Figure 2, Stage 1).

To obtain clinically meaningful causal features, we retain only the features that lie on a directed path to Y: a feature X_i is kept only if a directed path from X_i to Y exists in G. This removes discovered variables that make no contribution to the predictions. Accordingly, the updated causal graph is:

$$G_{\text{update}} = (V', E'), \quad V' = \{X_i \in V : \exists \text{ a directed path } X_i \rightsquigarrow Y \text{ in } G\} \cup \{Y\}, \quad E' = E \cap (V' \times V'). \tag{3}$$

where V' retains the causal features that admit a directed path to Y, and E' contains the edges of G with both endpoints in V'.

Classifier Training. Based on G_{update}, we train a classifier on the selected causal features. Let $\mathcal{F}_{cls}$ denote the classifier parameterized by θ_{cls}:

$$\mathcal{F}_{cls}\left(\mathbf{x}; \theta_{cls}\right) = \left(Y,\, p(Y \mid \mathbf{x}),\, \{\varphi_i\}_{X_i \in V' \setminus \{Y\}}\right), \tag{4}$$

where $p(Y \mid \mathbf{x})$ is the confidence of the candidate prediction given the full causal feature vector, and φ_i is the Shapley value of feature X_i. This design establishes a direct connection between the candidate predictions Y and the causal features X_i.

4.2 Step 2: Chain-of-Causality Construction and Evidence-Based Analysis

A flat set of important features is not yet a reasoning process: clinicians examine the gaze positions in order, letting each observation condition the next, and Step 2 turns the Shapley-weighted graph into such a trajectory (Figure 1). The two strategies mirror how clinicians weigh evidence—anchor at the strongest finding, or average over the chain with emphasis near the conclusion. Since one patient's chain may contain idiosyncratic links, the CMUs accumulate population-level chain statistics, and evidence matching keeps only the chain the population supports.

After causal feature selection and classifier training in Step 1, we construct the Chain of Causality (COC) by associating each Shapley value φ_i produced by the classifier with its corresponding causal feature in G_{update}, forming the Shapley-weighted graph G_{shapley}. Candidate directed paths from causal features X_i to the candidate predictions Y are then identified on this graph. Formally:

$$G_{\text{shapley}} = \widehat{F}(G_{\text{update}}, \varphi) = (V', E', \varphi), \quad \varphi : V' \setminus \{Y\} \to \mathbb{R},\ \varphi(X_i) = \varphi_i, \tag{5}$$

where $\varphi : V' \setminus \{Y\} \to \mathbb{R}$ is a node-weight function that assigns each causal feature X_i its Shapley value φ_i produced by the classifier $\mathcal{F}_{cls}$ (Eq. 4), whose magnitude reflects the contribution strength of X_i to the prediction; the edge structure E' is preserved.

We then select a unique chain of causality using two complementary strategies: (1) the strongest-evidence anchoring strategy emphasizes the most influential causal feature, and (2) the dynamically weighted average strategy evaluates the collective evidence along the entire chain, with greater emphasis on features closer to Y. A match function compares the two strategies against Causal Memory Units (CMUs) and yields the final COC, while the CMUs preserve recurrent causal patterns. Finally, the selected COC is passed to the COCAgent, which follows the chain sequentially and integrates the corresponding ocular evidence to produce an evidence-based diagnostic explanation.

Chain of Causality. Based on G_{shapley}, we identify directed paths that connect causal features X_i to the candidate predictions Y. A chain of causality is represented as:

$$L_{\text{coc}} = (v_0, v_1, \ldots, v_k), \quad (v_{j-1}, v_j) \in E',\ v_j \in V',\ j = 1, \ldots, k, \tag{6}$$

where $v_k = Y, v_0, \ldots, v_{k-1}$ are causal features, each associated with its Shapley weight $\varphi(v_j)$, and each consecutive pair (v_{j-1}, v_j) is a directed edge of G_{shapley}. Each chain thus preserves two types of information: the directed causal connections and the contribution of each feature to the candidate prediction. Since multiple directed paths may reach Y, a unique COC is selected from the candidate paths according to the strength and distribution of their evidence.

(1) Strongest-Evidence Anchoring. The first strategy anchors the COC to the strongest causal feature of the current sample. Among all feature nodes that can reach Y through a directed causal path, the feature with the maximum absolute Shapley value is selected as the starting node:

$$X_r = \arg \max_{X_i \in V' \setminus \{Y\}} |\varphi_i| , \tag{7}$$

where r indexes causal features ranked by Shapley value. Starting from X_r, we greedily follow the directed causal structure of G_{shapley} toward Y: the successor with the maximum absolute Shapley weight among the outgoing neighbors is selected, until Y is reached, yielding the chain:

$$L_{\text{COC max}} = (v_0, v_1, \ldots, v_k), \quad v_{j+1} = \arg \max_{u:\,(v_j,\,u) \in E'} |\varphi(u)|, \tag{8}$$

where $v_0 = X_r$, $v_k = Y$, $v_1, \ldots, v_{k-1}$ are the subsequent causal features along the directed path from X_r to Y, and the j-th reasoning link corresponds to the directed edge $(v_{j-1}, v_j) \in E'$. This strategy identifies the most influential causal feature while preserving its causal connection to the candidate prediction, thereby producing the strongest evidence.

(2) Dynamically Weighted Average. The second strategy considers the evidence contribution of the entire chain rather than relying on a single strongest feature. For each candidate chain p, we compute its weighted-average Shapley score $\text{score}(p)$ and select the chain with the highest score:

$$L_{\text{COC ave}} = \arg \max_{p} \text{score}(p), \tag{9}$$

$$\text{score}(p) = \frac{\sum_{X_i \in p} w_i \, |\varphi_i|}{\sum_{X_i \in p} |\varphi_i|}, \tag{10}$$

$$w_i = |\varphi_i|^{\rho} \cdot \gamma^{1-d_i}. \tag{11}$$

Here, d_i denotes the number of directed hops from X_i to Y, with γ=2 and ρ=1. Evidence with a stronger contribution and a shorter causal distance to Y therefore receives a larger weight w_i. The resulting ratio optimization is solved with Dinkelbach's algorithm (Dinkelbach, 1967). Compared with strongest-evidence anchoring, this strategy evaluates evidence distributions over the whole chain.

Causal Memory Units (CMUs). The chains of causality selected on the training set are further accumulated into Causal Memory Units (CMUs). For each diagnostic subtype, we record the occurrences of each directed edge and causal feature in its selected chains:

$$\text{CMU}_Y = \{\text{cnt}\,(l_t, Y)\,, \alpha\}\,, \tag{12}$$

where $\text{cnt}(l_t, Y)$ counts the directed edges and causal features appearing in the chains selected for subtype Y. Each subtype maintains its own CMU, and Laplace smoothing with α=0.01 avoids zero-frequency estimates. These statistics preserve the recurrent chains and causal features associated with each subtype and serve as memory for subsequent reasoning.

Evidence Matching and COC Reasoning. The two candidate chains, from strongest-evidence anchoring and dynamically weighted averaging, are matched against the CMUs to select the final COC. The COCAgent then follows the chain link by link, $\text{Link}_1, \ldots, \text{Link}_n$ (Figure 2), where each Link_j is a directed edge $(v_{j-1}, v_j) \in E'$ of the selected chain (Eq. 6), interpreting how the corresponding causal feature supports the candidate prediction. Every reasoning link is thus grounded in its visual evidence and gaze position within the chain, explicitly connecting the observed evidence, causal features, and the prediction.

4.3 Step 3: Knowledge-Grounded Verification and Report Generation

Even a well-grounded chain can start from a wrong candidate: the classifier's top prediction is only a hypothesis. Step 3 therefore gates generation on verification. The verifier applies the clinical criteria R that specialists use (listed in the supplementary material) and examines the original image rather than trusting the classifier's confidence score, rejecting unsupported candidates until one passes.

Verification Instruction. For each candidate diagnosis Y, the corresponding clinical criteria R are retrieved and provided to the verifier together with the original image I and the candidate confidence score. The verifier examines whether the visual evidence in I satisfies the clinical criteria and returns a binary result, *Pass* or *Fail*. On a *Fail*, the current candidate is rejected and the next-ranked candidate is verified with the same procedure, continuing until a candidate passes verification.

COC-Guided Generation. For the verified candidate diagnosis Y, the verification result is combined with the selected Chain of Causality L_{coc} and its evidence-based analysis E_{coc} produced by the COCAgent. Together with the original image I, these inputs are provided to the generator:

$$\text{Report} = \text{Generate}\left(I,\, R,\, L_{\text{coc}},\, E_{\text{coc}}\right). \tag{13}$$

The generated report integrates the verified clinical evidence with the Chain-of-Causality reasoning from the COCAgent, and the predicted label is the clinically verified candidate.

5 Experiments

5.1 Experimental Setup

Dataset. The benchmark contains 1,065 patients, divided at the patient level into training, validation, and test sets (857/104/104). The benchmark supports both six-class subtype classification and evidence-based diagnostic reasoning and report generation; dataset statistics, preprocessing details, and representative samples are provided in the supplementary material. **Comparison Methods.** For diagnostic reasoning, we compare COCAgent with general-purpose LVLMs (Qwen3.5 (Team, 2026), InternVL3.5 (Wang et al., 2025c)) and medical VLMs (HealthGPT (Lin et al., 2025), MedGemma 1.5 (Sellergren et al., 2026), Hulu-Med (Jiang et al., 2025b)). For subtype classification, we compare with image-based backbones (ResNet-50 (He et al., 2016), DenseNet (Huang et al., 2017), ViT (Dosovitskiy et al., 2021), Swin Transformer (Liu et al., 2021), MambaVision (Hatamizadeh et al., 2025), DINOv3 (Siméoni et al., 2025), EUPE (Zhu et al., 2026), ViT-5 (Wang et al., 2026a), Jumbo (Fuller et al., 2026), ThinkingViT (Hojjat et al., 2026)) and the task-specific, causality-inspired graph model CI-GNN (Zheng et al., 2025). All methods are trained and evaluated under the same patient-level split and evaluation protocol. **Implementation Details.** The visual classifier is trained in PyTorch on a NVIDIA RTX 4090 GPU for 500 epochs with a cosine learning-rate schedule. All VLMs, including the LVLM components of COCAgent, are evaluated zero-shot without task-specific fine-tuning: Qwen3.5 (4B, 9B) (Team, 2026), InternVL3.5 (4B, 8B) (Wang et al., 2025c), HealthGPT-Pro (4B, 8B) (Lin et al., 2025), MedGemma-1.5 (4B) (Sellergren et al., 2026), and Hulu-Med (4B, 7B) (Jiang et al., 2025b); only the classifier and the CMU statistics are derived from the training split. Hardware, augmentation, and the full prompt templates are given in the supplementary material.

5.2 Evaluation Metrics

Report Generation Metrics. We assess report quality with BLEU-1 (Papineni et al., 2002), ROUGE-1 (Lin, 2004), BERTScore (Zhang et al., 2020) (F1 over contextual token embeddings, roberta-large), and CosSim (cosine similarity of sentence embeddings from the all-MiniLM-L6-v2 encoder (Reimers & Gurevych, 2019)). We further introduce *Position Alignment* to measure recovery of abnormal gaze positions and clinically relevant Position Groups via exact position-set matching. It is the average recall of ground-truth abnormal position nodes, with a score of one when both reports contain none. **Classification Metrics.** We derive the deviation direction and the angle-pattern category (A-pattern, V-pattern, or non-A/V), and report Acc_{Dir} and Acc_{Ang}, together with weighted precision, recall, and F1 over the six subtypes (W-P, W-R, W-F1) (Pedregosa et al., 2011). **Ablation Metrics.** We score each chain with Ave_link (links per chain), Cov@10 (fraction of chain edges among the ten most frequent CMU edges), and Concept F1 over strabismus-specific concepts (gaze position) extracted from the generated reports, following Smit et al. (2020) and Jain et al. (2021). Finally, following ERASER (DeYoung et al., 2020), let $m(x)_j$ denote the probability that LVLMs assigns to the correct diagnosis j given input x. Comprehensiveness, $m(x)_j - m(x \backslash r)_j$, is the confidence drop when the chain evidence r is removed; a higher value means the chain is more necessary for the prediction. We report a sufficiency variant, $m(x_r)_j$, the confidence obtained from the chain evidence alone; a higher value means the chain alone suffices for the diagnosis. Full mathematical definitions are given in the supplementary material.

Table 1: Report generation and subtype classification on the test set. All metrics are reported as percentages (%); best results are in **bold**. COCAgent uses Gemini-3-Flash-Preview as its LVLMs backbone. Position Alignment measures the proportion of ground-truth Position/Position Group nodes correctly recovered by the prediction. A predicted node is considered matched only when its complete position index set exactly matches the ground truth.

Model	Position Alignment	Report Quality				Subtype Classification				
		BLEU-1	ROUGE-1	BERTScore	CosSim	Acc_{Dir}	Acc_{Ang}	W-P	W-R	W-F1
General VLM										
Qwen3.5-4B (Team, 2026)	20.4	8.3	29.8	82.0	26.5	29.8	79.8	47.3	17.3	19.4
Qwen3.5-9B (Team, 2026)	0.0	0.5	14.0	75.7	9.1	8.7	27.9	3.2	2.9	3.0
InternVL3.5-4B (Wang et al., 2025c)	28.1	6.0	28.1	80.9	26.7	48.1	79.8	48.9	44.2	46.5
InternVL3.5-8B (Wang et al., 2025c)	28.5	6.5	21.6	78.3	18.3	16.3	79.8	0.5	6.7	0.8
Medical VLM										
HealthGPT-Pro-4B (Lin et al., 2025)	6.7	7.9	28.8	80.6	28.8	13.5	78.8	59.5	12.5	18.1
HealthGPT-Pro-8B (Lin et al., 2025)	3.4	11.0	30.6	81.5	26.3	29.8	79.8	66.0	20.2	23.3
MedGemma-1.5-4B (Sellergren et al., 2026)	4.8	10.4	25.1	79.0	23.7	44.2	66.3	51.3	35.6	40.7
Hulu-Med-4B (Jiang et al., 2025b)	0.0	3.2	22.7	79.2	21.3	41.3	79.8	47.4	35.6	39.5
Hulu-Med-7B (Jiang et al., 2025b)	0.8	22.6	29.5	78.5	26.7	3.8	79.8	0.1	2.9	0.2
COCAgent (Ours)	**62.0**	**37.2**	**40.9**	**96.1**	**83.4**	**94.2**	**97.1**	**91.8**	**93.3**	**92.0**

Table 2: Performance comparison of representative classification methods. Acc_{Dir} and Acc_{Ang} denote direction and angle-pattern accuracy; W-P, W-R, and W-F1 are weighted precision, recall, and F1. All metrics are reported as percentages (%). Best results are in **bold**.

Model	Acc_{Dir}	Acc_{Ang}	W-P	W-R	W-F1
CI-GNN (Zheng et al., 2025)	90.4	77.9	63.9	68.3	65.8
DenseNet-121 (Huang et al., 2017)	89.4	80.8	70.2	73.1	71.4
DINOv3 ViT-B (Siméoni et al., 2025)	92.3	81.7	62.1	76.0	68.0
EUPE-ViT-B (Zhu et al., 2026)	92.3	76.0	71.3	69.2	69.3
Jumbo-Base (Fuller et al., 2026)	89.4	83.7	72.3	76.9	72.6
Swin Transformer (Liu et al., 2021)	87.5	79.8	64.8	70.2	66.2
ThinkingViT-DeiT (Hojjat et al., 2026)	92.3	82.7	77.7	76.0	70.4
ViT-5-Base (Wang et al., 2026a)	89.4	81.7	68.7	73.1	69.5
COCAgent (Ours)	**94.2**	**97.1**	**91.8**	**93.3**	**92.0**

5.3 Main Results

Diagnostic Report Generation. Table 1 shows that COCAgent outperforms every baseline on all four report-quality metrics by a wide margin, reaching 37.2 BLEU-1, 40.9 ROUGE-1, 96.1 BERTScore, and 83.4 CosSim, against the best baseline values of 22.6 (Hulu-Med-7B), 30.6 (HealthGPT-Pro-8B), 82.0 (Qwen3.5-4B), and 28.8 (HealthGPT-Pro-4B), respectively. More importantly, COCAgent achieves 62.0% Position Alignment, substantially exceeding the best baseline of 28.5% (InternVL3.5-8B), indicating that it more reliably identifies the abnormal gaze positions and clinically relevant Position Groups underlying the generated report. Fluent reports, however, do not imply correct diagnoses: Hulu-Med-7B ranks first among baselines in BLEU-1 yet reaches only 3.8% Acc_{Dir}, and several VLMs collapse to the majority angle class (Acc_{Ang} 79.8%). COCAgent instead couples accurate abnormal-position grounding and the best report quality with 94.2% Acc_{Dir} and 97.1% Acc_{Ang}, matching its supervised results in Table 2.

Strabismus Subtype Classification Performance. In table 2 COCAgent achieves the best results on all metrics, 94.2% Acc_{Dir}, 97.1% Acc_{Ang}, 91.8% W-P, 93.3% W-R, and 92.0% W-F1, improving over the strongest baselines (92.3% Acc_{Dir}, 83.7% Acc_{Ang}, 72.6% W-F1) by 1.9, 13.4, 14.1, 16.4, and 19.4 points, respectively. The largest gain is in Acc_{Ang}, indicating that COCAgent integrates gaze variation more effectively than direct visual backbones.

5.4 Ablation Study

Ablation on COCAgent Components. Table 3 ablates the components of COCAgent. Verification brings the largest gain: from (a) to (c), the exact subtype accuracy (W-R) rises from 76.0% to 90.4% and Acc_{Ang} from 84.6% to 98.1%, since filtering candidate findings before they enter the report matters most for the deviation angle, the noisiest quantity to estimate from the photograph of the nine cardinal positions of gaze. The no-rule controls localize this gain: dropping the rules costs 9.6

Table 3: Ablation of COCAgent components on StrabBench. Each row toggles one or more components: (a) causal features only; (b) rule-free verification; (c) verification with the knowledge-guided rules; (d) (c) plus the COC chain; (e) the full model without the rules; (f) the full COCAgent. Feat.: causal features; Verif.: verification; Evid. match: evidence matching. W-R equals the exact six-class accuracy (joint correctness of direction and pattern). All metrics are reported as percentages (%); best results in **bold**, ties at one decimal place bolded together. All variants use Gemini-3-Flash-Preview as the default LVLM. Row (a)produces no report, hence "–".

	Components					Report Quality				Subtype Classification				
Setting	Feat.	Rule	Verif.	COC	Evid. match	BLEU-1	ROUGE-1	BERTScore	CosSim	$\mathrm{Acc}_{\mathrm{Dir}}$	$\mathrm{Acc}_{\mathrm{Ang}}$	W-P	W-R	W-F1
(a)	✓	×	×	×	×	–	–	–	–	90.4	84.6	75.8	76.0	71.1
(b)	✓	×	✓	×	×	32.4	35.9	95.8	81.5	87.5	86.5	84.9	80.8	82.1
(c)	✓	✓	✓	×	×	36.9	40.4	96.0	82.1	92.3	**98.1**	89.4	90.4	89.6
(d)	✓	✓	✓	✓	×	**37.7**	**41.1**	**96.1**	82.0	**94.2**	93.3	89.6	89.4	89.4
(e)	✓	×	✓	✓	✓	33.1	36.4	95.8	81.6	91.3	92.3	**93.0**	87.5	88.1
(f)	✓	✓	✓	✓	✓	37.2	40.9	**96.1**	**83.4**	**94.2**	97.1	91.8	**93.3**	**92.0**

Table 4: Comparison of the chain selection strategies. $L_{\mathrm{COC\,max}}$ anchors the chain at the feature with the maximum Shapley value; $L_{\mathrm{COC\,ave}}$ selects the chain that maximizes the dynamically weighted score $w_i = |\varphi_i|^{\rho}\gamma^{1-d_i}$ over all candidate paths; Evidence Match selects per sample between the two candidates according to CMUs. All metrics except Ave_link are reported as percentages (%), averaged over the 104 evaluation samples; higher values are better. All variants use Gemini-3-Flash-Preview as the default LVLM. Best results are in **bold**.

Strategy	Ave_link	Cov@10	Concept F1	Comp	Suff
$L_{\mathrm{COC\,max}}$	1.31	92.5	20.2	4.4	73.7
$L_{\mathrm{COC\,ave}}$	1.85	80.8	**20.6**	**8.8**	72.6
Evidence Match	1.19	**94.1**	20.1	6.2	**74.2**

points of exact subtype accuracy at the verification stage ((b) vs. (c)) and 5.8 points in the full model ((e) vs. (f)), and weakens the clinical wording of reports (BLEU-1 36.9%/32.4% and 37.2%/33.1%). Conditioning generation on the COC chain (d) yields the peak surface metrics (BLEU-1 37.7%, ROUGE-1 41.1%) but loses 1.0 point of subtype accuracy, since the 7.5% of chain edges uncovered by the population statistics can carry unverified findings into the report. Evidence matching (f) removes exactly those links, recovering the best W-R (93.3%), W-F1 (92.0%), and CosSim (83.4%) for only 0.5 and 0.2 points of BLEU-1 and ROUGE-1.

Comparison of Chain Strategies. Table 4 compares the two chain selection strategies with evidence matching. The $L_{\mathrm{COC\,ave}}$ chains are the most informative (Concept F1 20.6%, Comprehensiveness 8.8%) but the least population-covered (Cov@10 80.8%), as their additional links are often absent from the CMU statistics; the anchored $L_{\mathrm{COC\,max}}$ chains are better covered (92.5%) but less informative. $L_{\mathrm{COC\,ave}}$ alone thus favors informativeness over validity: in Table 3, uncovered chains reduce the exact subtype accuracy by 1.0 point despite the best surface metrics. Evidence matching resolves this tension, selecting the population-supported candidate per sample and yielding the most compact chains (Ave_link 1.19) with the highest Cov@10 (94.1%) and Sufficiency (74.2%) for 0.5 points of Concept F1. COCAgent therefore combines $L_{\mathrm{COC\,ave}}$ with evidence matching and verification; four case studies appear in the supplementary material.

6 Conclusion

We presented COCAgent, an evidence-based multi-agent framework for strabismus subtype diagnosis, together with StrabBench, to our knowledge the first benchmark pairing strabismus cases with sample-level causal graphs and node-level analysis. Instead of generating reports directly, COCAgent selects causal features, builds a Chain-of-Causality over Shapley-weighted evidence, verifies each finding against clinical criteria, and grounds the chain in population-level evidence, making every diagnostic step traceable to explicit evidence. COCAgent substantially improves both subtype prediction and report quality on StrabBench, provides an effective solution for evidence-supported and clinically reliable decision making.

AI USE STATEMENT

In this work, we used generative AI tools to draft and edit parts of the manuscript for readability, suggest the paper structure and title, identify relevant literature for standard evaluation metrics, and format tables and references; we did not use such tools for any other task requiring disclosure, in particular not for generating data, developing the methodology, formulating mathematical claims, implementing methods, or designing experiments. Large vision-language models also appear in the research itself rather than as research aids: they drafted the preliminary diagnostic reports during StrabBench construction, each reviewed and revised by senior ophthalmologists before inclusion (Section 3), and they serve as components of the proposed framework (Section 4). We have reviewed all AI-assisted work: all AI-assisted text was checked by the authors against the experimental outputs and the clinical terminology standards adopted in this study, and all model-generated benchmark content was verified by ophthalmologists. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

StrabBench is built from retrospective, de-identified clinical images collected at a university-affiliated eye center. The study was approved by the center's Ethics Committee and registered with the Chinese Clinical Trial Registry; the approval and registration numbers are withheld to preserve anonymity and will be disclosed in the camera-ready version. All benchmark reports were reviewed by senior ophthalmologists, and no identifiable patient information is released. Because the data come from a single center, the benchmark may not represent other populations, and models trained on it should be re-validated before use in different clinical settings. COCAgent is a decision-support research prototype, not a medical device: its predictions and generated reports can be wrong and must not replace professional diagnosis. Any clinical use requires prospective validation, regulatory approval, and oversight by qualified clinicians.

REPRODUCIBILITY STATEMENT

The StrabBench construction protocol, including the photography standard and the doctor-in-the-loop annotation procedure, is described in Section 3. The full COCAgent pipeline—causal graph construction, the two chain selection strategies, evidence matching, and the verification and generation stages—is specified in Section 4, with all hyperparameters (γ=2, ρ=1, α=0.01) stated inline. Data splitting, baselines, and implementation details are given in Section 5.1, and all evaluation metrics are defined in Section 5.2. All experiments use the same patient-level 8:1:1 split, and the chain-level metrics are computed on the test set.

REFERENCES

Jinheon Baek et al. Knowledge-augmented language model prompting for zero-shot knowledge graph question answering. *arXiv preprint arXiv:2306.04136*, 2023.

Shuai Bai et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Rina Bao et al. Visual and domain knowledge for professional-level graph-of-thought medical reasoning. In *ICML*, 2025.

Balu Bhasuran et al. Evaluation of causal reasoning for large language models in contextualized clinical scenarios of laboratory test interpretation. *npj Digital Medicine*, 2026.

Hanzhu Chen et al. Knowledge graph finetuning enhances knowledge manipulation in large language models. In *ICLR*, 2025.

Liyi Chen et al. Plan-on-graph: Self-correcting adaptive planning of large language model on knowledge graphs. In *NeurIPS*, 2024.

Audrey Chia, Lipika Roy, and Linley Seenyen. Comitant horizontal strabismus: an asian perspective. *Br. J. Ophthalmol.*, 2007.

Diego Colombo et al. Order-independent constraint-based causal structure learning. *JMLR*, 2014.

Jay DeYoung et al. ERASER: A benchmark to evaluate rationalized NLP models. In *ACL*, 2020.

Werner Dinkelbach. On nonlinear fractional programming. *Management Science*, 1967.

Alexey Dosovitskiy et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

Abhimanyu Dubey et al. The llama 3 herd of models. *arXiv e-prints*, 2024.

David S Friedman et al. Prevalence of amblyopia and strabismus in white and african american children aged 6 through 71 months: the baltimore pediatric eye disease study. *Ophthalmology*, 2009.

Anthony Fuller et al. Thicker and quicker: A jumbo token for fast plain vision transformers. In *ICLR*, 2026.

Sreyan Ghosh et al. Visual description grounding reduces hallucinations and boosts reasoning in lvlms. In *ICLR*, 2025.

Google. Gemini 3 flash preview, 2026.

Ali Hatamizadeh et al. Mambavision: A hybrid mamba-transformer vision backbone. In *CVPR*, 2025.

Kaiming He et al. Deep residual learning for image recognition. In *CVPR*, 2016.

Ali Hojjat et al. ThinkingViT: Matryoshka thinking vision transformer for elastic inference. In *CVPR*, 2026.

Robbie Holland, Thomas RP Taylor, Christopher Holmes, et al. Specialized curricula for training vision language models in retinal image analysis. *npj Digital Medicine*, 2025.

Gao Huang et al. Densely connected convolutional networks. In *CVPR*, 2017.

Saahil Jain et al. RadGraph: Extracting clinical entities and relations from radiology reports. *arXiv preprint arXiv:2106.14463*, 2021.

Juntao Jiang et al. M3cotbench: Benchmark chain-of-thought of mllms in medical image understanding. In *ICLR*, 2026.

Pengcheng Jiang et al. Reasoning-enhanced healthcare predictions with knowledge graph community retrieval. In *ICLR*, 2025a.

Songtao Jiang et al. Hulu-med: A transparent generalist model towards holistic medical vision-language understanding. *arXiv preprint arXiv:2510.08668*, 2025b.

Yuxiang Lai et al. Med-r1: Reinforcement learning for generalizable medical reasoning in vision-language models. *IEEE TMI*, 2026.

Jin Li et al. Revealing multimodal causality with large language models. In *NeurIPS*, 2025a.

Mufei Li et al. Simple is effective: The roles of graphs and large language models in knowledge-graph-based retrieval-augmented generation. In *ICLR*, 2025b.

Sijing Li et al. Tumorchain: Interleaved multimodal chain-of-thought reasoning for traceable clinical tumor analysis. In *ICLR*, 2026.

Yan Li et al. A hybrid causal structure learning algorithm for mixed-type data. In *AAAI*, 2022.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, 2004.

Tianwei Lin et al. Healthgpt: A medical large vision-language model for unifying comprehension and generation via heterogeneous knowledge adaptation, 2025.

Guimeng Liu et al. How do medical mllms fail? a study on visual grounding in medical images. In *ICLR*, 2026a.

Jiaxiang Liu, Yuan Wang, Jiawei Du, Joey Tianyi Zhou, and Zuozhu Liu. Medcot: Medical chain of thought via hierarchical expert. In *EMNLP*, 2024.

Yu Liu et al. Pathmind: A retrieve-prioritize-reason framework for knowledge graph reasoning with large language models. In *AAAI*, 2026b.

Ze Liu et al. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021.

Linhao Luo et al. Reasoning on graphs: Faithful and interpretable large language model reasoning. In *ICLR*, 2024.

Linhao Luo et al. G-reasoner: Foundation models for unified reasoning over graph-structured knowledge. In *ICLR*, 2026.

Roberta McKean-Cowdin et al. Prevalence of amblyopia or strabismus in asian and non-hispanic white preschool children: multi-ethnic pediatric eye disease study. *Ophthalmology*, 2013.

OpenAI. Update to gpt-5 system card: Gpt-5.2, 2025.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translation. In *ACL*, 2002.

Letitia Parcalabescu and Anette Frank. Do vision and language decoders use images and text equally? how self-consistent are their explanations? In *ICLR*, 2025.

Fabian Pedregosa et al. Scikit-learn: Machine learning in Python. *JMLR*, 2011.

Chen Qiu et al. Explainable medical visual question answering via chain of evidence. *Knowledge-Based Systems*, 2025.

Dehui Qiu et al. A vr-based automated strabismus diagnosis system with progressive semi-supervised learning. *IEEE JBHI*, 2026.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *EMNLP-IJCNLP*, pp. 3982–3992, 2019.

Michael X Repka et al. Strabismus, strabismus surgery, and reoperation rate in the united states: analysis from the iris registry. *Ophthalmology*, 2018.

Dana Robaei et al. Factors associated with childhood strabismus: findings from a population-based study. *Ophthalmology*, 2006.

Andrew Sellergren et al. Medgemma 1.5 technical report. *arXiv preprint arXiv:2604.05081*, 2026.

Yucheng Shi, Quanzheng Li, Jin Sun, Xiang Li, and Ninghao Liu. Enhancing cognition and explainability of multimodal foundation models with self-synthesized data. In *ICLR*, 2025.

Yuxin Shi et al. Plan-answer-refine-on-graph: Structured planning and self-refinement for large language model reasoning on knowledge graphs. In *ICLR*, 2026.

Oriane Siméoni et al. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

Akshay Smit et al. Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. In *EMNLP*, 2020.

Jiashuo Sun et al. Think-on-graph: Deep and responsible reasoning of large language model on knowledge graph. In *ICLR*, 2024.

Qwen Team. Qwen3.5: Towards native multimodal agents, 2026.

Aniket Vashishtha et al. Executable counterfactuals: Improving llms' causal reasoning through code. In *ICLR*, 2026.

Feng Wang et al. ViT-5: Vision transformers for the mid-2020s. *arXiv preprint arXiv:2602.08071*, 2026a.

Lehan Wang et al. Interpretable bilingual multimodal large language model for diverse biomedical tasks. In *ICLR*, 2025a.

Song Wang et al. Reasoning of large language models over knowledge graphs with super-relations. In *ICLR*, 2025b.

Weiyun Wang et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025c.

Ziyue Wang et al. Medagent-pro: Towards evidence-based multi-modal medical diagnosis via reasoning agentic workflow. In *ICLR*, 2026b.

Peng Xia et al. Mmed-rag: Versatile multimodal rag system for medical vision language models. In *ICLR*, 2025.

Shiqi Yan et al. Explore-on-graph: Incentivizing autonomous exploration of large language models on knowledge graphs with path-refined reward modeling. In *ICLR*, 2026.

Yuguang Yang et al. Prompt as knowledge bank: Boost vision-language model via structural representation for zero-shot medical detection. In *ICLR*, 2025.

Shunyu Yao et al. Tree of thoughts: Deliberate problem solving with large language models. In *NeurIPS*, 2023.

Xiangning Yu et al. Causal sufficiency and necessity improves chain-of-thought reasoning. In *NeurIPS*, 2025.

Jiajin Zhang, Ge Wang, Mannudeep K Kalra, and Pingkun Yan. Disease-informed adaptation of vision-language models. *IEEE TMI*, 2024.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In *ICLR*, 2020.

Yaoze Zhang et al. Leanrag: Knowledge-graph-based generation with semantic aggregation and hierarchical retrieval. In *AAAI*, 2026.

Jiawen Zheng et al. Causality-inspired graph neural network for interpretable strabismus subtype classification. In *MICCAI*, 2025.

Zhuoping Zhou et al. Query-aware flow diffusion for graph-based rag with retrieval guarantees. In *ICLR*, 2026.

Chenchen Zhu et al. Efficient universal perception encoder. *arXiv preprint arXiv:2603.22387*, 2026.

Hui Zhu et al. Association between childhood strabismus and refractive error in chinese preschool children. *PLOS ONE*, 2015.