Med-CMR: A Fine-Grained Benchmark Integrating Visual Evidence and Clinical Logic for Medical Complex Multimodal Reasoning

Supplementary Material

1. Related Work

1.1. MLLMs in medical reasoning

MLLMs show strong capability and great potential in medical clinical reasoning across visual and textual modalities. Existing MLLMs can be categorized into general-domain and medical-domain models, which differ in their data sources and specialization for medical reasoning tasks.

General MLLMs. Multimodal large language models have evolved from perception-driven systems to adaptive reasoning-centered frameworks [18, 23, 24]. Early general-domain MLLMs were mainly developed for multimodal perception by aligning visual encoders with pretrained language models. [10, 13, 26]. Later models, including GPT-4 [1], adopt end-to-end multimodal training and support richer modalities, leading to stronger general multimodal reasoning. Building on these architectures, reasoning-oriented models introduced explicit internal reasoning phases [23], where the model generates and refines intermediate reasoning traces before producing the final answer. Examples include OpenAI's o1 [7] series, which allocates dedicated reasoning tokens. Unified frameworks such as Qwen3-VL [21], Gemini2.5 [4], and GPT-5 [16] further integrate a dynamic "thinking" mode into general multimodal models, allowing adaptive control over the amount of internal reasoning based on task complexity.

Medical MLLMs. Medical MLLMs are further trained on large-scale medical data and specifically developed for clinical and medical tasks. They have similarly evolved from a perception-aligned to a reasoning-enhanced paradigm. Models such as Med-Flamingo [14] and LLaVA-Med [9] are among the first to extend general multimodal frameworks to the medical domain through medical data adaptation and multimodal alignment. More recent domain models, including LingShu [20] and Medgemma [15], build upon medical semantic alignment and further incorporate multi-step reasoning data during training, enhancing implicit reasoning and factual consistency in medical applications.

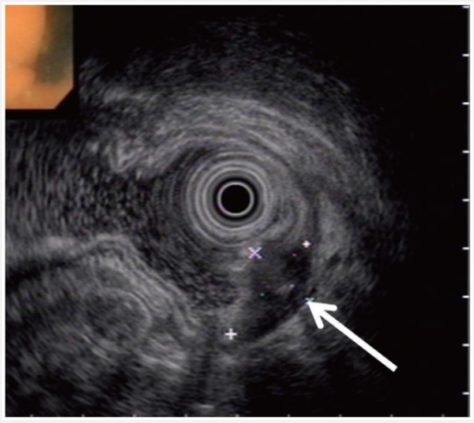
1.2. Multimodal Medical Benchmarks

Multimodal medical benchmarks have evolved from early datasets focusing on simple perception and conceptual understanding to recent ones that begin to address complex clinical reasoning. Early medical VQA benchmarks, such as VQA-RAD [8], VQA-Med [3] , Path-VQA [5], and SLAKE [12], focus on evaluating recognition and factual understanding, aiming at perception-level comprehension rather than reasoning. Specifically, VQA-RAD focuses on radiology, VQA-Med covers multiple medical specialties, Path-VQA centers on pathology slides, and SLAKE provides bilingual annotations for broader accessibility. Later, large-scale and domain-diverse benchmarks were introduced to broaden modality coverage and improve generalization. PMC-VQA [25] collects image–text pairs from biomedical literature, OmniMedVQA [6] covers multiple medical specialties, and GMAI-MMbench [22] integrates heterogeneous data for general multimodal evaluation. These benchmarks enable large-scale assessment but still focus on shallow comprehension and retrieval-based reasoning. More recent reasoning-oriented benchmarks start to explore complex medical reasoning. MedXpertQA [27] contains a large number of questions that involve complex reasoning but was not specifically developed for this purpose. HIE-Reasoning [2] focuses on clinical complex reasoning using 133 neonatal MRI cases, which confines its scope to a narrow clinical setting. Both benchmarks indicate that current models have difficulty with complex reasoning, yet neither offers a fine-grained evaluation or systematic analysis of medical reasoning complexity. To bridge this gap, we present Med-CMR, the first benchmark that provides a fine-grained evaluation of MLLMs in complex clinical reasoning. To enable fine-grained evaluation, we decompose the complexity of clinical medical reasoning into three visual dimensions and four reasoning dimensions, each evaluated by a corresponding task.

2. Case Studies

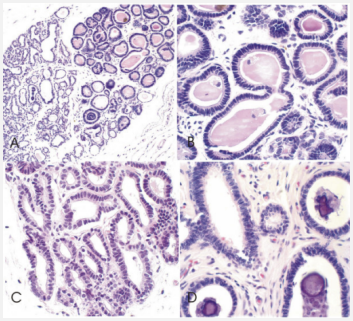
2.1. Case Studies Across Categories

Small-Object Detection



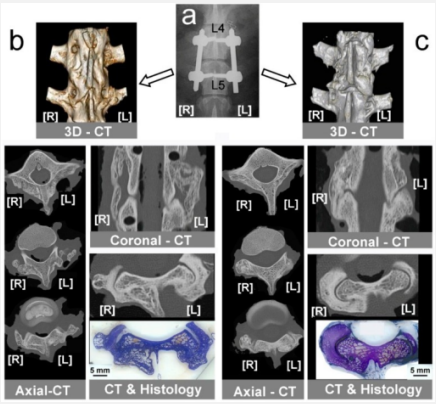
Question: Considering the mass's visual characteristics in the provided image and its relationship to surrounding structures, which feature best supports the conclusion that this is a pathologically significant tiny focus rather than an artifact or normal variation?
Options: A) The ovoid, heterogeneous hypoechoic focus with small echogenic foci centered in the submucosa and muscularis propria supports a true pathologic lesion. B) The ovoid, heterogeneous hypoechoic focus with small central echogenic foci confined to the mucosal layer is diagnostic of a true pathologic lesion. C) The ovoid, heterogeneous hypoechoic focus with central echogenic foci abutting only the muscularis mucosae suggests a focal pathology and therefore supports a true lesion rather than benign variation. D) The ovoid, heterogeneous hypoechoic focus with small central echogenic foci confined to the mucosal layer and superficial submucosa is indicative of a true pathologic lesion rather than an artifact. E) The small ovoid, heterogeneous hypoechoic focus with punctate echogenic foci localized predominantly to the mucosal and superficial submucosal layers is most consistent with a true pathologic lesion rather than an artifact.
Answer: A
Explanation: The mass shows an ovoid and heterogeneous hypoechoic focus with several small echogenic spots inside it. These features are centered in the submucosa and muscularis propria, which are deeper wall layers that do not normally contain such discrete structures. Artifacts on endoscopic ultrasonography usually do not form a stable ovoid shape, do not show internal echogenic foci arranged in a consistent pattern, and do not remain confined to specific wall layers. Normal variation of the gastric wall also does not create a sharply outlined hypoechoic mass in these deeper layers. Because the lesion has a defined shape, internal components, and a location that matches a true subepithelial abnormality, option A best supports the presence of a real small lesion rather than an artifact or normal tissue change.

Fine-Detail Discrimination



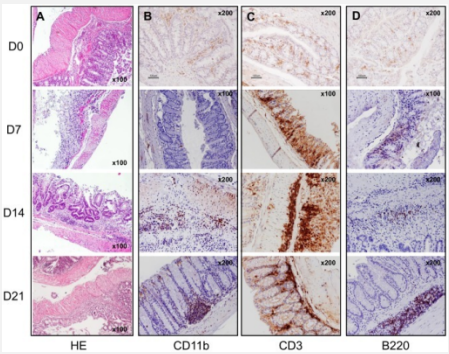
Question: Observe the overall pattern of the lesions in the tissue sections. Do the visual attributes suggest a central distribution or a pattern reflecting multicentric involvement?
Answer: The lesion pattern is consistent with multicentric involvement.
Explanation: The tissue sections show many separated glandular structures scattered throughout the sampled region. Each focus has nearly the same microscopic features, including a single layer of columnar epithelial cells, uniform nuclei, similar intraluminal secretions, and occasional calcifications. There is no visible gradient of change, no directional pattern, and no structure that can be identified as a central point of extension. Instead, the abnormalities appear as multiple small and discrete foci that share the same architecture. This pattern supports the interpretation that the lesions arise independently in multiple sites rather than spreading from one origin.

Spatial Understanding



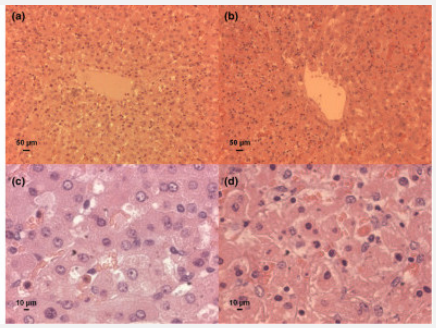
Question: Based on the comparison of structural and functional images in the affected region, how does the signal intensity correlate with the expected tissue viability in the reconstructed area?
Options: A) Signal intensity correlates with viable tissue on the left, reflecting optimal graft integration and successful fusion in the HA scaffold group. B) Signal intensity aligns with viable tissue on the right, consistent with successful fusion. C) Signal intensity corresponds to viable tissue on the left, indicating robust graft incorporation and successful fusion on the left side. D) Signal intensity aligns with viable tissue on the left, consistent with successful fusion. E) Signal intensity is greater on the left, consistent with preserved bone viability and fusion success localized to the left side.
Answer: B
Explanation: The CT images and histology sections show that only the right side has a continuous bridge of new bone. The right side appears denser on CT and matches the appearance of healthy, formed bone on the histology slice. The left side does not show the same continuous structure and has gaps instead of solid fusion. Because the structural views and the tissue views both point to good bone formation on the right side, the signal pattern matches viable tissue only on the right, which supports successful fusion on that side.

Temporal Prediction



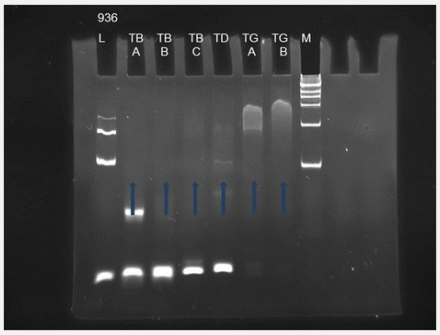
Question: What visual configuration most strongly indicates a therapy-related process rather than a baseline condition in the affected region across the different time-points?
Answer: Progressive accumulation of CD3+, followed by B220+ and Mac-1+ inflammatory cells at later time-points.
Explanation: The images show a time-dependent change in the inflammatory pattern of the colon. At day 0, the tissue architecture is mostly preserved and immune cell staining is sparse. By day 7, $CD3^+$ T cells begin to accumulate along the mucosa, indicating an early lymphocytic response. At day 14, $CD3^+$ staining becomes dense and widespread, showing that T-cell infiltration is the dominant feature at this stage. At the same time, $B220^+$ B cells start to appear, although at a lower level. By day 21, the inflammatory infiltrate becomes more mixed. $B220^+$ B cells and $Mac\text{-}1^+$ myeloid cells increase, filling deeper layers of the mucosa and submucosa. This progression in the order of appearance—first $CD3^+$ T cells, then $B220^+$ B cells, and finally $Mac\text{-}1^+$ cells—shows a dynamic sequence instead of a baseline condition. This temporal pattern is consistent with a therapy-related or injury-related immune process, where different immune cell populations are recruited at different stages.

Causal Reasoning



Question: What visual configuration most strongly indicates a therapy-related rather than baseline process in the affected region?
Answer: Sinusoidal infiltration by polymorphonuclear neutrophils and lymphocytes in the centrolobular liver
Explanation: The images compare two conditions that differ only by the use of the recruitment manoeuvre: pressure-controlled ventilation alone (PCV) and pressure-controlled ventilation plus recruitment (PCV+R). In the PCV group (a, c), centrolobular liver tissue shows largely preserved architecture with only a few scattered inflammatory cells in the sinusoids. In the PCV+R group (b, d), the same region of the liver now shows dense sinusoidal infiltration by polymorphonuclear neutrophils and lymphocytes. Because the animals, organ, stain, and magnification are otherwise matched, this new inflammatory pattern appears specifically in the group that received the additional manoeuvre. The most reasonable causal interpretation is that the recruitment manoeuvre is associated with therapy-related liver injury, and the visual configuration that captures this causal effect is the sinusoidal infiltration by neutrophils and lymphocytes in the centrolobular area.

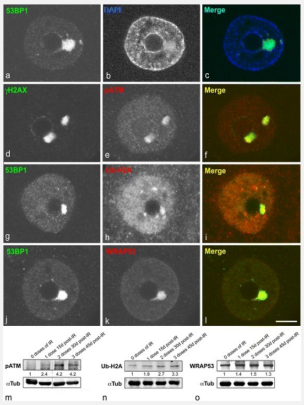
Long-Tail Generalization



Question: Which feature in the observed pattern most supports an unusual mechanism rather than a common diagnosis?
Options: A) Germline configuration of T-cell receptor loci interpreted as evidence for a de-differentiated T-cell neoplasm rather than an NK-cell origin. B) Germline configuration of T-cell receptor genes taken to exclude an NK-cell origin and instead indicate a cryptic T-cell clone with non-productive rearrangements. C) Germline T-cell receptor configuration definitively confirming lineage infidelity as a de-differentiated T-cell neoplasm rather than a potential NK-cell origin. D) Germline T-cell receptor loci interpreted as definitive evidence of lineage infidelity consistent with a de-differentiated T-cell neoplasm rather than an NK-cell lineage. E) Germline T-cell receptor configuration indicating a potential NK-cell origin.
Answer: E
Explanation: The gel shows that all tested samples have a germline pattern for the T-cell receptor (TCR) loci rather than a discrete clonal rearranged band. In other words, the TCR genes remain in their unrearranged configuration, similar to the control ladder, and there is no evidence of a clonal T-cell population. True T-cell lymphomas almost always show rearranged TCR genes, so a purely germline configuration argues against a conventional T-cell neoplasm. Natural killer (NK) cells do not rearrange TCR genes, so finding germline TCR in a lymphoid malignancy strongly supports an NK-cell origin. This unusual lineage assignment, based on the germline TCR configuration, is the key feature that points to a rare NK-cell lymphoma mechanism rather than a common T-cell lymphoma diagnosis.

Multi-source integration

Question: Which visual feature in the image series most strongly supports the conclusion that the observed lesions reflect a coordinated pathological process rather than random occurrences?
Answer: The recurring colocalization of multiple DNA damage and repair markers within the same PDDF in the affected nucleus.
Explanation: The images show that several DNA damage–related markers — 53BP1, γH2AX, pATM, Ub-H2A, and WRAP53 — are all concentrated within the same intranuclear structure, the persistent DNA damage focus (PDDF). In every panel, these markers appear in the same region of the nucleus, demonstrating repeated colocalization rather than scattered or independent lesions. This stable and shared localization pattern indicates that the abnormalities arise from a single coordinated DNA damage process instead of multiple random events. The Western blot data support this by showing IR-dependent increases in these same repair markers, matching the focal nuclear accumulation observed microscopically.

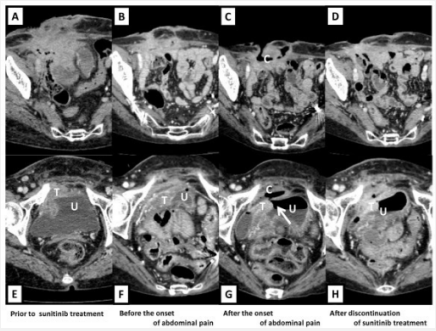


Med-CMR examples across seven categories.

2.2. GPT-5 Error Case Studies

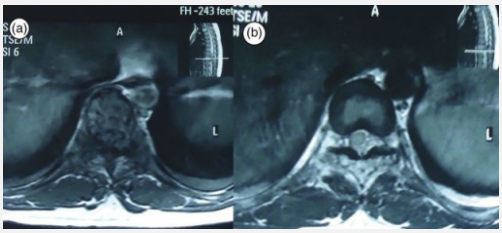
Question Understanding

Question: Considering the affected region's relationship with adjacent structures, which aspect most clearly supports the significance of the subtle finding?
Options: A) A focal mural discontinuity at the posterior bladder wall contiguous with the tumor, consistent with a tiny bladder dome perforation. B) A minute defect at the inferior (neck) aspect of the anterior bladder wall adjacent to the lesion suggesting intraperitoneal communication. C) A small defect at the anterior bladder wall adjacent to the lesion. D) A minute defect at the inferior (neck) aspect of the anterior bladder wall adjacent to the lesion suggesting intraperitoneal communication. E) A minute defect at the lateral anterior bladder dome adjacent to the lesion.



Correct Answer: C
Model Answer: E
Model Reasoning: The question asks which imaging detail, based on how the lesion relates to nearby structures, best confirms the clinical importance of a subtle CT finding. Serial axial CT images show a pelvic mass abutting the urinary bladder. At the lateral anterior bladder dome, right where the mass contacts the bladder wall, a minute defect in the bladder wall at that site. This focal defect at the dome adjacent to the lesion most strongly indicates a small but clinically important breach or early communication. Other observations either occur at locations that do not relate to the dome or do not show this discrete dome defect, so they lack the same direct anatomic relevance.
Error Interpretation: The model did not ground its interpretation in the specific structural detail that the question was designed to assess. It shifted the abnormality toward a dome-based location and responded as if the finding were meant to highlight a lateral dome defect, which is not what the question described. By doing so, it effectively answered a different question than the one asked. This indicates a question understanding error.

Image Recognition



Question: Analyze the relationship between the lesions and their surrounding anatomical structures. Are the lesions primarily acting as a primary process with satellites or are they indicative of secondary spread from another region?

Options: A) The pattern suggests a primary paraspinal soft-tissue neoplasm centered in the paraspinal musculature with contiguous spread into adjacent vertebrae and focal epidural impingement rather than a primary vertebral lesion with satellite deposits. B) The pattern suggests multifocal paravertebral soft tissue disease with superficial secondary vertebral scalloping and partial epidural compression, indicating primary soft-tissue origin with subsequent superficial bony remodeling rather than true bony invasion. C) The pattern suggests a primary vertebral lesion with satellite involvement, demonstrating extraosseous extension and epidural encasement of the cord. D) The pattern suggests a primary vertebral process with multifocal satellite lesions limited to the paravertebral soft tissues without true epidural encasement of the thecal sac. E) The pattern suggests multifocal paravertebral soft tissue disease with secondary involvement of the vertebral body margins and partial epidural displacement of the thecal sac without true vertebral origin.

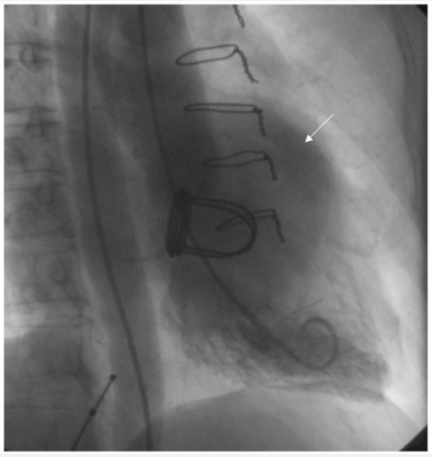
Correct Answer: C

Model Answer: B

Model Reasoning: The task is to decide whether the observed spinal lesions originate primarily from paravertebral soft tissues with subsequent spread to adjacent structures such as the bony vertebra or epidural space, or whether they represent secondary involvement of soft tissues due to a primary vertebral process. Axial MR images at the thoracic level show bulky, lobulated soft tissue masses in the paravertebral regions. The masses abut and indent the posterolateral vertebral body and pedicle with cortical indentation and shallow scalloping rather than a marrow centered expansile lesion. There is extension toward the epidural space causing partial effacement of the thecal sac and cord compression. Intervertebral disc contours appear relatively preserved, and involvement is asymmetric and multifocal along the paravertebral gutters. Primary vertebral pathology typically begins within the vertebral body marrow with centripetal expansion, endplate and disc involvement, and then secondarily extends into paravertebral or epidural spaces, but those features such as marrow centered destruction and disc or endplate collapse are not dominant here. The conspicuous finding is large paravertebral soft tissue masses that contact the bone surface and produce scalloping-type superficial remodeling, implying outside-in influence rather than true primary osseous destruction. Multifocal paravertebral lesions with only superficial secondary cortical scalloping and epidural encroachment favor a soft tissue origin with satellites rather than a primary osseous process. Preservation of disc spaces further argues against primary spondylodiscitis-like or marrow-origin lesions. The pattern suggests multifocal paravertebral soft tissue disease with superficial secondary vertebral scalloping and partial epidural compression, indicating primary soft-tissue origin with subsequent superficial bony remodeling rather than true bony invasion.

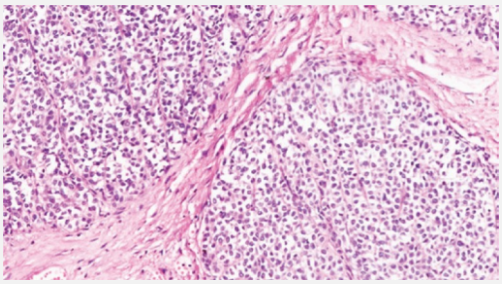
Error Interpretation: The model did not correctly identify the true site of origin of the lesions. The expected pattern reflects a primary vertebral process with secondary extraosseous extension, but the model interpreted the appearance as soft-tissue–dominant disease with only superficial effects on the vertebrae. This shows that the model misread the structural relationships on the images. The error comes from image recognition.

Reasoning Process



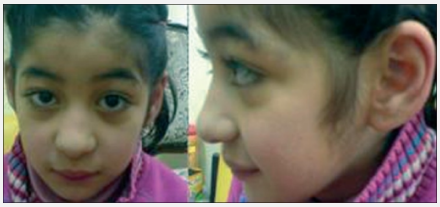
Question: What visual configuration in the image most strongly indicates a therapy-related adverse effect rather than a baseline process?
Options: A) Displacement and extrinsic narrowing of the left circumflex artery by an adjacent aneurysm. B) External indentation of the left circumflex artery by an adjacent aneurysmal sac accompanied by focal vessel irregularity and intraluminal filling defect consistent with intrinsic luminal disease rather than pure extrinsic compression. C) Displacement of the left circumflex artery by an adjacent aneurysm with preserved luminal caliber and absence of any vessel displacement or extrinsic narrowing, indicating post-procedural anatomical remodeling rather than acute compression. D) External indentation of the left circumflex artery by an adjacent aneurysmal sac without appreciable vessel indentation or displacement, suggesting no compression. E) Displacement of the left circumflex artery by an adjacent aneurysm with preserved luminal caliber and no adjacent aneurysmal sac, suggesting intrinsic vessel tortuosity rather than extrinsic narrowing.
Correct Answer: A
Model Answer: B
Model Reasoning: The task is to identify which visual pattern on the image points to a complication caused by therapy (iatrogenic or extrinsic effect) rather than a patient's baseline intrinsic coronary disease. The fluoroscopic cardiac image shows surgical hardware such as a ring or annuloplasty device and wires, indicating prior cardiac intervention. The arrow points to a coronary segment along the lateral wall. At the arrowed site the coronary contour shows a sharp outward concavity adjacent to a rounded external opacity consistent with an aneurysmal or pseudoaneurysmal sac. In addition to this external impression, the vessel segment at the point of contact appears irregular with a focal intraluminal filling defect, implying that the wall and lumen themselves have been altered rather than simply pushed aside. Baseline intrinsic coronary disease such as stable atherosclerosis usually appears as smooth luminal narrowing without an associated adjacent sac or combined pattern of external indentation plus intraluminal defect. A therapy-related adverse effect, such as a postoperative pseudoaneurysm with associated intraluminal thrombus, can both indent the artery from outside and create focal luminal irregularity. This combined appearance of an external aneurysmal sac, focal indentation, and intraluminal filling defect is best captured by option B: external indentation of the left circumflex artery by an adjacent aneurysmal sac accompanied by focal vessel irregularity and intraluminal filling defect consistent with intrinsic luminal disease rather than pure extrinsic compression.
Error Interpretation: The model correctly identifies the adjacent aneurysmal structure and its relationship to the left circumflex artery, but it misinterprets the vessel contour and luminal appearance during its reasoning process. While working through the options, it introduces features such as focal vessel irregularity and an intraluminal filling defect, even though these are not supported by the image. This overinterpretation leads it away from the actual pattern of displacement and extrinsic narrowing that indicates a therapy-related effect. The mistake therefore stems from a flawed reasoning process in which unwarranted details are added and the visual findings are overstated, causing the model to favor an option that does not match the observed configuration.

Domain Understanding Error



Question: Analyze the microscopic pattern within the affected region. What unusual feature could suggest a rare disease process rather than a common condition?
Options: A) Epithelioid cells with mildly irregular nuclei and low mitotic activity in an organoid pattern, with isolated tiny foci of degenerative change mischaracterized as necrosis. B) Irregular nuclei with infrequent mitoses in a disorganized architectural pattern, with diffuse inflammation without necrotic foci. C) Epithelioid cells with irregular nuclei and low mitotic activity in an organoid/epithelioid architectural pattern, with small confluent areas interpreted as ischemic-type necrosis absent. D) Irregular nuclei with infrequent mitoses in an epithelioid/organoid arrangement, with scant focal apoptosis without established coagulative necrosis. E) Irregular nuclei with numerous mitoses in an organoid/epithelioid architectural pattern, including focal necrosis.
Correct Answer: E
Model Answer: A
Model Reasoning: The task is to examine the histologic image and identify an uncommon microscopic feature that would point toward a rare disease process rather than a typical condition. The field shows solid nests or sheets of uniform epithelioid cells separated by fibrous septa. The nuclei are mildly irregular, the chromatin is fairly even, and there are only rare mitotic figures indicating low proliferative activity. The architecture is organoid or nested. There are tiny pale foci suggestive of degenerative change but not broad, established necrosis. Common reactive or inflammatory conditions would show mixed inflammatory cells and lack an organoid nesting pattern with even this low level of mitotic activity. Many common malignant processes would show marked pleomorphism, high mitotic rates, and overt coagulative necrosis, none of which is prominent here. The combination of epithelioid cytology, organoid growth, and low mitotic activity with only minute degenerative foci is more aligned with a rare epithelioid or neuroendocrine-type process rather than a common pathology. This pattern is best captured by option A: epithelioid cells with mildly irregular nuclei and low mitotic activity in an organoid pattern, with isolated tiny foci of degenerative change mischaracterized as necrosis.
Error Interpretation: The model's mistake stems from a gap in its pathological knowledge. Although the image shows epithelioid cells in an organoid pattern with mildly irregular nuclei, low mitotic activity, and only tiny foci of degenerative change, the model did not recognize that this specific combination is more suggestive of certain rare epithelioid or neuroendocrine-type processes rather than of a common condition. Instead, it interpreted these findings as nonspecific changes without appreciating their diagnostic implication. This indicates that the model lacks the domain knowledge required to distinguish the histologic pattern of a rare disease from patterns seen in more typical lesions.

Format Error



Question: By examining the visible face features, can you deduce the nature of the developmental anomaly and how it affects facial structure organization?
Options: A) The facial features indicate a developmental anomaly characterized by frontal hair upsweep, long face, high-set ears, sparse eyebrows, long eyelashes, epicanthic folds, and sparse eyebrows, reflecting altered craniofacial organization. B) The facial features suggest a developmental anomaly with frontal hair upsweep, short face, high-set ears, sparse eyebrows, short eyelashes, and absent epicanthic folds, indicating altered craniofacial organization. C) The facial features suggest a developmental anomaly with frontal hair upsweep, long face, high-set ears, sparse eyebrows, short eyelashes, epicanthic folds, and sparse eyebrows, indicating altered craniofacial organization. D) The facial features indicate a developmental anomaly characterized by frontal hair upsweep, short face, low-set ears, thick eyebrows, long eyelashes, and absent epicanthic folds, reflecting altered craniofacial organization. E) The facial features indicate a developmental anomaly characterized by frontal hair upsweep, long face, low-set ears, thick eyebrows, long eyelashes, epicanthic folds, and full eyebrows, reflecting altered craniofacial organization.
Correct Answer: E
Model Answer: /S
Model Reasoning: The question asks to infer, from the facial features shown, what developmental anomaly is present and how it alters the organization of facial structures. The child's face shows a frontal hair upsweep with an elongated facial contour, high-set external ears, sparse eyebrows, long eyelashes, and the presence of epicanthic folds. These findings occur bilaterally and affect multiple craniofacial units. The combination of facial elongation, altered ear position, reduced eyebrow density, eyelash prominence, and epicanthic folds points to a disturbance in craniofacial development that affects the proportional relationships and patterned organization of the face. Such abnormalities reflect altered growth or alignment of craniofacial elements derived from early embryologic processes governing facial morphology. Mechanistically, disruption in the coordinated development of these structures leads to an atypical arrangement of facial features, producing the observed craniofacial organization.
Error Interpretation: The model output "/S" instead of a valid option due to a breakdown in its answer formatting logic, likely caused by confusion between text generation and option selection steps during decoding.

GPT-5 error examples across five error types.

3. Data Leakage Analysis

To ensure that Med-CMR provides a fair and reliable evaluation of multimodal models, we analyze the dataset for potential leakage signals. This analysis shows that Med-CMR exhibits minimal evidence of leakage.

During benchmark curation, we follow a controlled process designed to reduce the chance of overlap with existing corpora. All questions are generated by instantiating human-designed templates that are created manually and adjusted only slightly to match each case, rather than adapted from any external materials. Correct answers are derived from the caption content and rewritten in a new form. For multiple-choice items, distractors are proposed by several models and then refined through human selection. This pipeline greatly reduces the chance of data leakage.

We further examine potential contamination using N-gram Accuracy, specifically ROUGE-L [11] and edit distance similarity, which have been adopted in earlier studies [19]. We sample 2,000 MCQs and 500 open-ended questions and convert the MCQs into open-ended form to collect free-text predictions so that lexical similarity can be measured directly. We evaluate three top-performing models, including both proprietary and open-source models, since these models offer a representative upper bound for potential overlap. We compare each model's responses with the reference answers to obtain the ROUGE-L and edit distance similarity, and the results are reported in Table 1 and Table 2. Using the criteria introduced in prior work [19], the measured similarity is far below the level associated with contamination, indicating that the leakage risk is minimal.

Model	MCQ-converted	OE	Overall
GPT-5 [16]	0.123	0.132	0.124
InternVL3.5-241B-A28B [17]	0.073	0.108	0.079
Qwen3-VL-235B-A22B [21]	0.047	0.072	0.051

Table 1. ROUGE-L (Longest Common Subsequence F-score) across MCQ-converted, open-ended, and overall subsets.

Model	MCQ-converted	OE	Overall
GPT-5 [16]	0.069	0.078	0.071
InternVL3.5-241B-A28B [17]	0.049	0.068	0.052
Qwen3-VL-235B-A22B [21]	0.032	0.050	0.035

Table 2. Edit distance similarity across MCQ-converted, open-ended, and overall subsets.

4. Medical Coverage and Distribution

4.1. Modality

Modality	Count	Prop.
Pathology / Microscopy	5662	0.264
Photography / Dermatology	4662	0.218
CT	3441	0.161
X-ray & Fluoroscopy	2593	0.121
MRI	2017	0.094
Ultrasound	896	0.042
Endoscopy	764	0.036
Ophthalmology / Optical	649	0.030
Nuclear Medicine / PET	436	0.020
Physiologic Signals	225	0.011
Radiotherapy	54	0.003
Derived / Meta	23	0.001

Table 3. Modality coverage list and distribution. Prop. denotes proportion.

4.2. Body System

Body System	Count	Prop.
Digestive	3843	0.186
Integumentary	3218	0.156
Nervous	2545	0.123
Skeletal	2266	0.110
Other / NA	1848	0.089
Respiratory	1824	0.088
Cardiovascular	1821	0.088
Reproductive	1274	0.062
Muscular	842	0.041
Urinary	813	0.039
Endocrine	360	0.017

Table 4. Body system coverage list and distribution. Prop. denotes proportion.

5. Task Mapping

Complexity Dimension	Corresponding Task	Abbr.
Visual Complexity		
Small-Object Detection	Small-Object Detection	SOD
Fine-Detail Discrimination	Fine-Detail Discrimination	FDD
Spatial Understanding	Spatial Understanding	SU
Reasoning Complexity		
Temporal Prediction	Temporal Prediction	TP
Causal Reasoning	Causal Reasoning	CR
Long-Tail Generalization	Long-Tail Generalization	LTG
Multi-Source Integration	Multi-Source Integration	MSI

Table 5. Mapping between multimodal medical reasoning complexity dimensions and tasks. Abbr. denotes task abbreviation.

6. Prompts

6.1. MCQ Response Generation Prompt

Please carefully observe this medical image and answer the following question:

Question: {question}
Options:
A) {option_Text_A}
B) {option_Text_B}
C) {option_Text_C}
D) {option_Text_D}
E) {option_Text_E}

Answer only with the option letter (A–E).

Table 6. Prompt of MCQ response generation.

6.2. Open-ended Response Generation Prompt

Please carefully observe this medical image and answer the following question:

Question: {question}

Think step by step, integrating both visual features and medical knowledge to reach your conclusion. Then provide the final answer to the question in one short sentence or a single medical term.

Output format (Must follow strictly):
Reasoning: <visual and diagnostic reasoning process>
Answer: <final answer to the question>

Table 7. Prompt of open-ended response generation.

6.3. Open-ended Response Scoring Prompt

You are a comprehensive and unbiased medical imaging evaluation assistant.

You are given the following inputs:
0. [QUESTION]: {question}
- This is the **original diagnostic or descriptive question** posed to the model. It defines what the model is expected to describe, explain, or conclude about the image. It provides the evaluation context and should be considered when judging relevance and reasoning alignment.

1. [IMAGE CAPTION]: {image_caption}
- This is a **factual visual description** of the image content. It objectively states what is seen in the image, including structures, densities, shapes, boundaries, or textures. It serves as the visual ground truth for all vision-related evaluation.

2. [GROUND TRUTH]: {ground_truth}
- This is the **expert-verified reference interpretation** describing the clinically correct diagnosis or observation derived from the image. It represents the gold standard for semantic and factual correctness, not the visual features themselves.

3. [MODEL REASONING]: {model_reasoning}
This is the **model's reasoning process** that explains how it interprets the visual information and arrives at its final answer. It may include descriptive, inferential, or diagnostic steps.

4. [MODEL ANSWER]: {model_answer}
- This is the **model's final conclusion or diagnostic output**, which summarizes its interpretation of the image based on the reasoning process.

You must evaluate the model output starting from Clarity and end with Ground Truth Consistency. Each aspect must be scored independently; do not let one aspect's judgment influence the others.

- The first two aspects (Language Clarity and Reasoning Coherence) are evaluated only on linguistic and logical quality, **not factual correctness**.
- The Vision Feature Accuracy aspect compares the model's visual understanding against the **IMAGE CAPTION only**, **without considering medical correctness or ground truth content**.
- Only the last aspect (Ground Truth Consistency) compares reasoning and answer with the **GROUND TRUTH** for factual and semantic accuracy.
There are four independent aspects to evaluate:
1. Language Clarity (10%)
2. Reasoning Coherence (10%)
3. Vision Feature Accuracy (40%)
4. Ground Truth Consistency (40%)

Step 1 — Language Clarity (10%)
Evaluate whether [MODEL REASONING] and [MODEL ANSWER] are clear and unambiguous in expression, without internal contradictions or vague wording.
Scores:
0: Ambiguous, contradictory, or difficult to understand; key meaning is unclear.
0.5: Generally understandable but includes ambiguous terms, unclear phrasing, or minor inconsistencies.
1: Fully clear, precise, and internally consistent with no ambiguity.

Step 2 — Reasoning Coherence (10%)
Evaluate whether [MODEL REASONING] is logically coherent and consistent, ensuring that the reasoning flow from description to conclusion is smooth and medically reasonable.
Scores:
0: Clear logical errors or incoherent reasoning flow.
0.5: Mostly coherent but includes small logical gaps or weak causal links.
1: Fully coherent, consistent, and logically well-structured reasoning.

Step 3 — Vision Feature Accuracy (40%)
Evaluate whether [MODEL REASONING] and [MODEL ANSWER] correctly reflect the key visual features described in [IMAGE CAPTION], including structure, density, morphology, boundary, and texture.
Do not consider medical correctness or the ground truth content — only check the consistency with the image caption.
Scores:
0: No relevant visual features match the caption.
1: Only a few minor features match; most are inaccurate or missing.
2: Some correct features, but major structures or textures are misunderstood.
3: Mostly correct; captures main visual characteristics but includes noticeable or clinically relevant inaccuracies.
4: Completely correct; all key visual features match the caption with no meaningful omissions.

Step 4 — Ground Truth Consistency (40%)
Evaluate how well the combination of [MODEL REASONING] and [MODEL ANSWER] matches the [GROUND TRUTH] in terms of semantic meaning, factual correctness, and overall medical interpretation.
Scores:
0: Completely incorrect; reasoning and answer have no alignment with the ground truth.
1: Mostly incorrect; minor semantic overlap but conceptually different.
2: Partially correct; some overlap but with major factual or interpretive errors.
3: Mostly consistent; captures main meaning but with clear omissions or inaccuracies.
4: Perfectly consistent; reasoning and answer fully align with the ground truth meaning.

Please output the results strictly in the following format (each score on a new line):
Language Clarity: [score]
Reasoning Coherence: [score]
Vision Feature Accuracy: [score]
Ground Truth Consistency: [score]

Table 8. Prompt of scoring the open-ended response.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1

[2] Rina Bao, Shilong Dong, Zhenfang Chen, Sheng He, Ellen Grant, and Yangming Ou. Visual and domain knowledge for professional-level graph-of-thought medical reasoning. In *Forty-second International Conference on Machine Learning*, 2025. 1

[3] Asma Ben Abacha, Sadid A Hasan, Vivek V Datla, Dina Demner-Fushman, and Henning Müller. Vqa-med: Overview of the medical visual question answering task at imageclef 2019. In *Proceedings of CLEF (Conference and Labs of the Evaluation Forum) 2019 Working Notes*. 9-12 September 2019, 2019. 1

[4] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 1

[5] Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020. 1

[6] Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22170–22183, 2024. 1

[7] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024. 1

[8] Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):180251, 2018. 1

[9] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564, 2023. 1

[10] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 1

[11] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 9

[12] Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, pages 1650–1654. IEEE, 2021. 1

[13] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 1

[14] Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, pages 353–367. PMLR, 2023. 1

[15] Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025. 1

[16] Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025. 1, 10

[17] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. 10

[18] Yun Xing, Xiaobin Hu, Qingdong He, Jiangning Zhang, Shuicheng Yan, Shijian Lu, and Yu-Gang Jiang. Boosting reasoning in large multimodal models via activation replay. *arXiv preprint arXiv:2511.19972*, 2025. 1

[19] Ruijie Xu, Zengzhi Wang, Run-Ze Fan, and Pengfei Liu. Benchmarking benchmark leakage in large language models. *arXiv preprint arXiv:2404.18824*, 2024. 9

[20] Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, et al. Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. *arXiv preprint arXiv:2506.07044*, 2025. 1

[21] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 1, 10

[22] Jin Ye, Guoan Wang, Yanjun Li, Zhongying Deng, Wei Li, Tianbin Li, Haodong Duan, Ziyan Huang, Yanzhou Su, Benyou Wang, et al. Gmai-mmbench: A comprehensive multimodal evaluation benchmark towards general medical ai. *Advances in Neural Information Processing Systems*, 37:94327–94427, 2024. 1

[23] Xinlei Yu, Chengming Xu, Guibin Zhang, Zhangquan Chen, Yudong Zhang, Yongbo He, Peng-Tao Jiang, Jiangning Zhang, Xiaobin Hu, and Shuicheng Yan. Vismem: Latent vision memory unlocks potential of vision-language models. *arXiv preprint arXiv:2511.11007*, 2025. 1

[24] Xinlei Yu, Chengming Xu, Guibin Zhang, Yongbo He, Zhangquan Chen, Zhucun Xue, Jiangning Zhang, Yue Liao, Xiaobin Hu, Yu-Gang Jiang, et al. Visual multi-agent system: Mitigating hallucination snowballing via visual flow. *arXiv preprint arXiv:2509.21789*, 2025. 1

[25] Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*, 2023. 1

[26] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 1

[27] Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. *arXiv preprint arXiv:2501.18362*, 2025. 1