

Appendix Contents

A. Limitations and Future Work	1
B. Ethical Statement	1
C. Dataset Detailed Analysis	2
C.1. Overview of Datasets	2
C.2. Data Preprocessing and Quality Control	2
C.3. Performance of Data Preprocessing	3
D. Ablation Study	3
E. Reward Collapse	3
E.1. Normalized Contribution Index (NCI)	3
E.2. Quantitative Analysis based on NCI	5
F. Implement Details	5
F.1. Hyperparameters	5
F.2. Models	5
F.3. Medical Thinking Knowledge Injection	6
G. Supplementary Quantitative Results	6

A. Limitations and Future Work

Scope of Evaluation. Our current evaluation is limited to publicly available medical VQA benchmarks. While these datasets encompass diverse modalities and question types, they may not fully reflect the complexity and heterogeneity of real-world clinical scenarios. In future work, we plan to extend the evaluation to larger-scale and more diverse clinical datasets, including proprietary hospital archives and cross-institutional benchmarks, to better assess generalizability.

Model Scale and Efficiency. We adopt a 3B-parameter backbone to ensure reproducibility and computational efficiency. Although this scale sufficiently demonstrates the effectiveness of our adaptive reinforcement learning framework, it may limit the upper bound of reasoning capability. Future research will explore scaling to larger model backbones and investigating lightweight distillation strategies to balance performance and efficiency in deployment.

Reward Design Generality. The adaptive semantic reward in ARMed is primarily designed for medical reasoning tasks. While the formulation is conceptually general, its applicability to other multimodal or non-medical reasoning domains remains untested. Future directions include extending ARMed to broader visual reasoning tasks such as scientific VQA, document understanding, or instructional vision-language reasoning.

Evaluation Metrics. We rely mainly on conventional text-based metrics (BLEU, ROUGE, BERTScore, and CosSim)

to evaluate open-ended answers. These automatic measures may not adequately capture the depth, correctness, or interpretability of clinical reasoning. Future work will involve designing more semantically grounded, human-aligned, and task-specific evaluation protocols—potentially integrating expert review or reasoning trace analysis.

B. Ethical Statement

This research strictly adheres to the ethical standards and best practices for medical AI research.

Data Usage and Privacy. All datasets used in this work (VQA-RAD, SLAKE, PathVQA, VQA-Med, PMC-VQA, and MedXpertQA) are publicly released for academic research purposes. All data were fully anonymized and de-identified by the original providers prior to public release, containing no personally identifiable information (PII) or Protected Health Information (PHI). Our data usage fully complies with the respective dataset licenses and data use agreements. No private or clinical patient data were accessed or used in this study.

Intended Use and Potential Misuse. Our proposed framework, **ARMed**, is designed solely for academic research on reinforcement learning and medical reasoning under open-ended visual question answering (VQA) settings. It is not intended for any direct clinical diagnosis or patient care. Although the model demonstrates improved reasoning interpretability and factual accuracy, it is a research prototype that should only be used for benchmarking, algorithmic analysis, and educational purposes. Any deployment in clinical environments requires additional regulatory validation and expert oversight. We explicitly discourage any misuse such as generating synthetic clinical advice, manipulating diagnostic workflows, or producing deceptive “medical deepfakes.”

Algorithmic Fairness and Bias. ARMed’s performance is influenced by the quality and balance of its training datasets. Despite including multiple benchmark datasets to increase diversity, potential biases may still exist—for instance, in demographic, imaging modality, or disease-type distributions. These biases could be inadvertently learned and propagated by the model. We recommend that future work systematically investigate fairness, demographic generalization, and the ethical implications of reinforcement learning signals in medical reasoning systems.

Societal Impact. By enhancing the interpretability and reliability of medical vision-language reasoning, this research aims to advance transparent and accountable medical AI. However, we recognize that such technologies could also amplify systemic biases or be misused if applied irresponsibly. We therefore emphasize the importance of human-in-the-loop oversight, transparent evaluation, and continued ethical scrutiny throughout model development and deploy-

answers, often stemming from vague or underspecified questions. This granularity inconsistency leads to cases where multiple semantically distinct answers are all valid, but only under different implicit interpretations of the same question. Such ambiguity not only complicates answer evaluation but also destabilizes policy learning in reinforcement-based training paradigms, where reward signals are sensitive to subtle semantic discrepancies.

To enhance the consistency of open-ended medical QA pairs, we take conceptual inspiration from prior work on VQA consistency auditing [7], which conducted refinement for both closed- and open-ended questions. In contrast, our approach focuses exclusively on open-ended questions, and we refine and extend the auditing details to better support the redefinition process tailored for reinforcement learning. This targeted design allows for a more nuanced and context-aware QA refinement strategy, as illustrated in Figure S3:

1. **Semantic Coverage Assessment:** ensuring that each question captures the full semantic scope of its corresponding answer;
2. **Preservation of Expressiveness:** retaining free-form, descriptive phrasing that supports nuanced and clinically relevant reasoning;
3. **Specificity Calibration:** aligning the granularity of the question with the level of detail in the answer to avoid under- or over-generalization.

These refinements, performed as a preprocessing step, enhance question-answer alignment and improve semantic fidelity, thereby providing a more stable foundation for policy optimization under the GRPO framework.

To illustrate the impact of our refinement strategy, several representative examples of question rewriting are presented in Figure S4. Each example showcases the transformation of an ambiguous or underspecified original question into a more precise and semantically aligned version, while preserving its open-ended nature. These refinements enhance the alignment between the question and its corresponding answer, thereby reducing potential mismatches that could mislead reinforcement-based policy updates. Specifically, the revisions clarify vague phrasing, increase semantic coverage, and calibrate specificity to match the expert-provided answers. This process serves as a critical preprocessing step, ensuring higher consistency and reasoning quality during training.

C.3. Performance of Data Preprocessing

To evaluate the effectiveness of our proposed VQA Refinement framework, we conduct ablation studies on three representative methods—SFT, GRPO, and ARMed-I—across multiple medical VQA datasets. As shown in Table S1 and Table S2, VQA Refinement consistently and significantly improves model performance under all settings.

D. Ablation Study

Due to space limitations, we only present part of the ablation study results in the main text. The complete results of the ablation experiments are provided in Table S3 and Table S4.

E. Reward Collapse

E.1. Normalized Contribution Index (NCI)

Motivation. In Group Relative Policy Optimization (GRPO), the policy update is governed by the advantage A_i , obtained via group-wise normalization of the total reward:

$$A_i = \frac{r_i - \bar{r}}{\sigma(r)}. \quad (\text{S1})$$

Let the total reward be a weighted sum of distinct components, $r = \sum_k \gamma_k R_k$. Substituting this expression into the advantage formulation and rearranging terms yields:

$$A_i = \sum_k \underbrace{\left[\gamma_k \frac{\sigma(R_k)}{\sigma(r)} \right]}_{\text{Effective Coefficient}} \underbrace{\left(\frac{R_{k,i} - \bar{R}_k}{\sigma(R_k)} \right)}_{\text{Normalized Component}}. \quad (\text{S2})$$

Equation S2 reveals a key structural insight: the *effective optimization weight* of component k is not simply the scalar γ_k , but is modulated by the ratio $\frac{\sigma(R_k)}{\sigma(r)}$. This leads to what we term *Reward Collapse*—if a particular reward component (e.g., a semantic constraint) exhibits much lower variance than the overall reward ($\sigma(R_k) \ll \sigma(r)$), its effective coefficient approaches zero. Consequently, the advantage signal A_i —and thus the policy gradient—becomes insensitive to that component, rendering it effectively invisible during optimization.

From Advantage to Component Contribution. Starting from Eq. S2, we can rewrite the normalized advantage as

$$\begin{aligned} A_i &= \sum_k \alpha_k Z_{k,i}, \\ \alpha_k &= \gamma_k \frac{\sigma(R_k)}{\sigma(r)}, \\ Z_{k,i} &= \frac{R_{k,i} - \bar{R}_k}{\sigma(R_k)}. \end{aligned} \quad (\text{S3})$$

Each $Z_{k,i}$ represents a standardized (zero-mean, unit-variance) version of reward component R_k , while α_k determines its effective scaling in the normalized advantage.

Since A_i is normalized within each group, we have $\text{Var}(A) = 1$. Taking the covariance between A and itself yields:

$$\begin{aligned} 1 &= \text{Cov}(A, A) = \text{Cov} \left(\sum_k \alpha_k Z_k, A \right) \\ &= \sum_k \alpha_k \text{Cov}(Z_k, A). \end{aligned} \quad (\text{S4})$$

Table S1. Effect of the refine process on model performance across Path-VQA, SLAKE, and PMC VQA datasets. Results are shown for SFT, GRPO, and ARMed-I with and without refinement. Gray-shaded columns indicate dataset-wise averages.

Method	Refine	Path-VQA					SLAKE					PMC VQA
		BLEU-1	ROUGE-1	BERTScore	CosSim	Avg.	BLEU-1	ROUGE-1	BERTScore	CosSim	Avg.	(Acc.)
SFT	✗	48.28	48.94	96.74	90.98	71.24	74.00	75.00	99.13	97.49	86.41	37.55
	✓	55.43	56.62	97.74	94.36	76.04	70.58	71.54	99.10	97.67	84.72	46.80
GRPO	✗	42.37	43.51	97.63	92.32	68.96	64.49	66.50	99.08	97.30	81.84	39.55
	✓	52.36	54.28	98.04	95.60	75.07	68.30	69.49	99.13	97.89	83.70	48.10
ARMed-I	✗	44.76	46.64	95.25	89.09	68.94	68.86	70.62	98.72	97.17	83.84	46.70
	✓	54.88	58.77	97.49	94.74	76.47	68.44	70.49	98.91	97.54	83.85	47.30

Table S2. Effect of the refine process on model performance across VQA-RAD, VQA-Med, and MedXpert datasets. Results are shown for SFT, GRPO, and ARMed-I with and without refinement. Gray-shaded columns indicate dataset-wise averages.

Method	Refine	VQA-RAD					VQA-Med					MedXpert
		BLEU-1	ROUGE-1	BERTScore	CosSim	Avg.	BLEU-1	ROUGE-1	BERTScore	CosSim	Avg.	(Acc.)
SFT	✗	56.60	58.87	98.13	95.52	77.28	21.83	23.39	95.47	90.09	57.70	16.65
	✓	56.58	58.68	98.36	96.27	77.47	22.44	24.55	95.54	90.11	58.16	18.35
GRPO	✗	47.26	49.81	98.44	96.20	72.93	15.81	18.37	94.42	90.74	54.84	17.95
	✓	51.43	53.68	97.54	97.28	74.98	19.95	20.85	96.32	91.71	57.21	20.65
ARMed-I (Ours)	✗	49.34	52.50	97.39	95.27	73.63	17.01	19.62	96.41	92.39	56.36	19.30
	✓	54.04	58.89	98.36	96.84	77.03	18.72	21.48	95.15	91.28	56.66	20.90

Table S3. Ablation study results on Path-VQA, SLAKE, and PMC VQA datasets. **Aug** denotes *augmentation*, **Text** denotes *text reward*, **Sem** denotes *semantic reward*, and **Ada** denotes *adaptive semantic reward*. Complete results on all test sets can be found in the Appendix. For each metric column, **bold** denotes the best performance and underline denotes the second-best. Gray-shaded columns correspond to dataset-specific averages.

Model	Method				Path-VQA					SLAKE					PMC VQA
	w. Aug	w. Text	w. Sem	w. Ada	BLEU-1	ROUGE-1	BERTScore	CosSim	Avg.	BLEU-1	ROUGE-1	BERTScore	CosSim	Avg.	(Acc.)
Qwen2.5-VL-3B	✗	✗	✗	✗	5.30	8.20	86.91	85.10	46.38	13.34	19.09	90.35	89.14	52.98	45.70
	✗	✓	✗	✗	49.62	56.07	96.21	93.80	73.93	64.91	68.76	98.48	97.19	82.34	46.40
	✗	✓	✓	✗	52.36	54.28	98.04	95.60	75.07	68.30	69.49	99.13	97.89	83.70	48.10
	✗	✓	✗	✓	54.88	58.77	97.49	94.74	76.47	68.44	70.49	98.91	97.54	83.85	47.30
	✓	✗	✗	✗	40.72	43.06	96.49	94.11	68.60	59.39	60.54	97.76	97.32	78.75	43.40
	✓	✓	✗	✗	<u>61.72</u>	61.92	97.65	95.74	79.26	73.16	<u>74.87</u>	98.85	97.15	86.01	46.60
	✓	✓	✓	✗	61.23	<u>62.53</u>	97.64	96.03	<u>79.36</u>	<u>73.22</u>	73.96	<u>99.15</u>	<u>97.97</u>	<u>86.08</u>	48.05
	✓	✓	✗	✓	63.61	64.96	98.46	<u>95.81</u>	80.71	76.14	76.65	99.46	98.13	87.60	48.75

Table S4. Ablation study results on VQA-RAD, VQA-Med, and MedXpertQA datasets. **Aug** denotes *augmentation*, **Text** denotes *text reward*, **Sem** denotes *semantic reward*, and **Ada** denotes *adaptive semantic reward*. For each metric column, **bold** denotes the best performance and underline denotes the second-best. Gray-shaded columns correspond to dataset-specific averages.

Model	Method				VQA-RAD					VQA-Med					MedXpert
	w. Aug	w. Text	w. Sem	w. Ada	BLEU-1	ROUGE-1	BERTScore	CosSim	Avg.	BLEU-1	ROUGE-1	BERTScore	CosSim	Avg.	(Acc.)
Qwen2.5-VL-3B	✗	✗	✗	✗	9.58	13.38	88.46	87.64	49.77	4.88	6.39	69.47	67.01	36.94	11.60
	✗	✓	✗	✗	52.55	59.18	97.84	96.48	76.51	13.90	17.39	94.53	90.96	54.20	19.90
	✗	✓	✓	✗	51.43	53.68	97.54	<u>97.28</u>	74.98	19.95	20.85	<u>96.32</u>	91.71	57.21	20.65
	✗	✓	✗	✓	54.04	58.89	98.36	96.84	77.03	18.72	21.48	95.15	91.28	56.66	20.90
	✓	✗	✗	✗	49.61	51.53	97.62	96.93	73.92	20.52	<u>22.50</u>	95.29	92.04	57.59	19.25
	✓	✓	✗	✗	57.63	<u>61.09</u>	98.66	97.23	78.65	20.75	22.42	95.30	91.96	57.61	20.50
	✓	✓	✓	✗	<u>57.78</u>	60.93	<u>98.68</u>	97.26	<u>78.66</u>	<u>20.84</u>	22.33	96.31	<u>92.34</u>	57.96	<u>20.95</u>
	✓	✓	✗	✓	58.72	61.10	98.69	97.29	78.95	21.51	23.17	96.56	92.36	58.40	22.30

Relating to Reward Components. Because $A = (r - \bar{r})/\sigma(r)$ is a linear normalization of r , the covariance scales

$$\text{Cov}(R_k, A) = \frac{\text{Cov}(R_k, r)}{\sigma(r)}. \quad (\text{S5})$$

Table S5. Hyperparameters for Adaptive Semantic Alignment Reward

Hyperparameter	Description	Example Value
ρ	Minimum ratio for filtering valid rewards	0.8
L_{\max}	Max length of historical reward buffer	2000
p	Percentile to compute dynamic threshold	0.5
δ_{\max}	Max threshold change per step	0.01
T_{\min}	Minimum allowed threshold value	0.0
T_{\max}	Maximum allowed threshold value	0.995
ϵ	Small constant to avoid division by zero	1×10^{-8}
α_{pos}	Steepness of S-curve (positive side)	5.0
α_{neg}	Steepness of S-curve (negative side)	2.0
τ	Word frequency threshold	5

Since linear normalization preserves correlation, it follows that

$$\text{Cov}(Z_k, A) = \rho(Z_k, A) = \rho(R_k, A) = \rho(R_k, r). \quad (\text{S6})$$

In other words, the standardized component Z_k has the same correlation with the normalized advantage A as R_k has with the total reward r .

Definition of NCI. Substituting this relation into Eq. S4 gives:

$$1 = \sum_k \alpha_k \rho(R_k, r) = \sum_k \gamma_k \frac{\sigma(R_k)}{\sigma(r)} \rho(R_k, r). \quad (\text{S7})$$

This motivates defining the **Normalized Contribution Index (NCI)** for each component as

$$\text{NCI}_k = \gamma_k \frac{\sigma(R_k)}{\sigma(r)} \rho(R_k, r), \quad (\text{S8})$$

which satisfies $\sum_k \text{NCI}_k = 1$.

Interpretation. Each NCI_k quantifies the proportional contribution of component k to the total variance of the normalized advantage A :

- γ_k — design weight (intended importance);
- $\frac{\sigma(R_k)}{\sigma(r)}$ — relative signal strength;
- $\rho(R_k, r)$ — alignment with the overall reward.

Intuitively, NCI_k measures how much the component R_k contributes to the unit-variance signal of A . When $\text{NCI}_k \approx 0$, that component has effectively collapsed and no longer influences the policy update.

E.2. Quantitative Analysis based on NCI

From Figure S2, we can clearly observe the evolution of NCI values for each reward component during training under both GRPO and ARMed-I. In GRPO, the NCI of the format reward (R_f) rapidly drops to nearly zero, indicating that it barely contributes to the gradient update. The

NCI of the static semantic reward (R_s) also gradually decreases to around 0.2, suggesting that its guidance becomes much weaker compared to the textual reward (R_c). This phenomenon is a clear manifestation of the *reward collapse* problem discussed earlier. In contrast, under our proposed ARMed-I, the NCI values of the textual reward (R_c) and the adaptive semantic alignment reward (R_{as}) remain at comparable levels after stabilization, indicating that both semantic and textual signals are effectively guiding the model toward the desired optimization direction.

F. Implement Details

F.1. Hyperparameters

GRPO. The detailed training configurations for GRPO are provided in the implementation details section in main text. Specifically, the hyperparameters used in our design of the adaptive semantic alignment reward are listed in Table S5. These include thresholding criteria, buffer limits, and the shaping parameters of the reward transformation function, all of which are carefully selected to ensure both training stability and semantic discrimination.

Supervise Fine-Tuning. For Supervised Fine-Tuning (SFT), we employ the LLaMA Factory with full-parameter tuning, using the same base model and hardware setup. The batch size is set to 16 per GPU without gradient accumulation. A cosine learning rate schedule is adopted with an initial rate of 1×10^{-6} as same as GRPO.

F.2. Models

BERTScore. To compute the BERTScore, we adopt PubMedBERT [3], a transformer-based language model that is pretrained from scratch on the PubMed abstracts and full-text biomedical articles. Unlike general-domain models such as BERT-base or RoBERTa, PubMedBERT is specifically optimized for understanding biomedical terminology, abbreviations, and domain-specific sentence structures. Its

pretraining corpus ensures a closer alignment with the linguistic patterns found in medical visual question answering tasks, thereby enabling more accurate semantic similarity estimation between generated answers and ground truth responses.

Cosine Similarity. For cosine similarity evaluation, we utilize BioBERT-mnli-snli-scinli-scitail-mednli-stsb [2], a domain-adapted sentence embedding model that integrates multiple biomedical natural language inference and semantic textual similarity datasets during training. This model is designed to capture nuanced relational semantics in the biomedical context, making it well-suited for reward evaluation tasks in medical reinforcement learning. By leveraging embeddings from this multi-task fine-tuned BioBERT variant, we aim to provide a robust and context-aware similarity signal that reflects medical factual consistency and reasoning coherence.

F.3. Medical Thinking Knowledge Injection

Our training set consists of a total of 26,366 questions. Following our selection strategy, we curated a subset of 4,581 representative questions. Among them, 940 questions were identified as high-frequency types, while the remaining 3,641 were categorized as low-frequency types. A word cloud visualization of the answers corresponding to high-frequency questions is shown in Figure S1.

G. Supplementary Quantitative Results

To further validate the effectiveness and interpretability of the proposed **ARMed** framework, this section presents several representative demonstration results (see Figure S5-S8). These examples cover diverse medical visual question answering (VQA) scenarios, illustrating the model’s reasoning capability across different modalities and question types. It can be observed that ARMed maintains both semantic consistency and clinical plausibility throughout the reasoning process. The reasoning trace (`<think>` block) and the corresponding final answer (`<answer>` block) jointly constitute a transparent and analyzable decision pathway, further confirming the stability and generalizability of the proposed method in open-ended medical reasoning tasks.

VQA Refinement Prompt

```
ori_q: {question}  
ori_a: {answer}
```

Role: “QA-Consistency Auditor” – an expert data curator. Your task is to audit and refine **open-ended** and **binary (yes/no)** visual-question-answering (VQA) pairs by evaluating the alignment between the question (`ori_q`) and the answer (`ori_a`). Revise them when necessary to ensure clarity, precision, and consistency. Visual content is unavailable—base all decisions solely on textual information.

Process:

- 1 Parse the original question (`ori_q`) and original answer (`ori_a`).
- 2 Based solely on linguistic content, infer the most likely intended meaning (`Expert_Guess`).
- 3 Compare `Expert_Guess` with `ori_a` for consistency in scope, granularity, semantic focus, and answerability.
- 4 Assign a status from the following:
 - **consistent**: The question already elicits exactly the content and structure found in the answer.
 - **needs_fix**: The question is underspecified, ambiguous, or mismatched with the answer.
 - **drop**: The QA pair is incoherent, contradictory, or cannot be repaired.

Guidelines for Fixing (if `status = "needs_fix"`):

General Constraints:

- Ensure the revised question fully supports and specifies all semantic information in the answer.
- Maintain alignment in scope, granularity, structure, and ordering between question and answer.
- Preserve the form of the answer (e.g., short phrase, description, yes/no).

For Open-Ended Questions:

- Begin with a directive verb (e.g., “Describe”, “Explain”, “Identify”, “What is”, “Why is”).
- Explicitly request each semantic element contained in the answer.
- Match the level of detail; avoid over-generalization or unnecessary specificity.
- Keep the ordering of components consistent between question and answer.
- Avoid vague references such as “this”, “that”, “there”, or “it” unless the referent is clear from text alone.

For Yes/No Questions:

- Ensure the question expresses a clear, testable proposition that can be answered by “yes” or “no”.
- Explicitly include the subject and the attribute or action being confirmed or denied.
- Remove ambiguity and implicit assumptions (e.g., clarify “Is it raining?” to “Is the sky raining in this scene?” if needed).
- Do not convert yes/no answers into descriptive ones or vice versa.
- Maintain grammatical completeness and clarity.

Output Format:

Return exactly **one JSON object** with the schema:

```
{  
  "status": "consistent | needs_fix | drop",  
  "ori_q": "<string>",  
  "ori_a": "<string>",  
  "new_q": "<string>",  
  "new_a": "<string>",  
  "notes": "rationale for decision"  
}
```

Do not return anything other than the JSON object.

Figure S3. Prompt template for open-ended and binary (yes/no) VQA refinement.

Examples of VQA Refinement

Example 1

Original Question: Where does the trabecular bone forming the marrow space show?

Refined Question: In what location does the trabecular bone forming the marrow space typically appear?

Answer: At the margins.

Example 2

Original Question: What does the cortical bone forming the outer shell show?

Refined Question: Describe the structural features visible in the cortical bone forming the outer shell, including the arrangement of lamellae, presence of osteocytic lacunae, and relationship to blood vessels.

Answer: Concentric lamellae along with osteocytic lacunae surrounding central blood vessels.

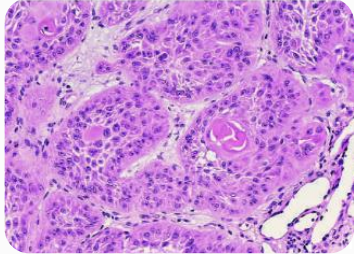
Example 3


Original Question: What does process begin as?


Refined Question: Describe how the process begins, including its initial form and location within the body.

Answer: The process begins as a focus of microabscess in a vascular loop in the marrow, which then expands to stimulate further activity.

Figure S4. Illustration of open-ended VQA refinement examples. Each pair consists of an original question, a refined version with improved specificity and semantic clarity, and the corresponding expert-provided answer.



: What specific findings indicate that these tumor cells are differentiating towards a normal squamous cell phenotype based on the histological structures visible in the image?

: **<think>** The image provided is a histological section showing tumor cells that closely resemble normal squamous epithelial cells. Key features include:

1. Cellular Similarity: The tumor cells have a morphology that is very similar to normal squamous epitheliacytoplasm.
2. Intercellular Connections: A specific structural feature to note is the presence of intercellular bridges. These are visible connections between the cells (representing desmosomes), which are a hallmark of squamous epithelium and contribute to the "prickle cell" appearance seen in normal tissue.
3. Keratinization: The prominent, bright pink structures (often concentric) are not fluid-filled vesicles, but rather nests of keratin (also known as keratin pearls). This signifies that the tumor cells are retaining the functional ability of normal squamous cells to produce keratin. These features collectively—specifically the intercellular bridges and keratin nests—demonstrate that the tumor cells are strikingly mimicking the characteristics of normal squamous epithelial cells. **</think>**

<answer> The tumor cells exhibit a striking resemblance to normal squamous epithelial cells, characterized distinctively by the presence of intercellular bridges and the formation of keratin nests. **</answer>**

Figure S5. A quantitative result of our ARMed



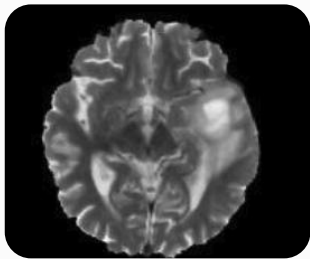
: Interpret this radiologic image, clarifying its projection method and the primary observable skeletal pathology.



<think>The image is a standard radiograph of the pelvis. The view is consistent with an Anteroposterior (AP) projection, commonly used for evaluating both hip joints. The pathology is localized to the right femoral head (which appears on the left side of the image). The femoral head has lost its smooth, spherical contour and shows signs of mechanical failure, appearing flattened and collapsed. Internally, there is a mixed pattern of increased density (sclerosis) and decreased density (lucent areas/cystic changes). This appearance—subchondral collapse, structural failure, and mixed density—is the classic radiographic presentation of Avascular Necrosis (Osteonecrosis). It is distinct from an acute acetabular fracture, which would present with a defined fracture line in the socket. Based strictly on the visual evidence, the diagnosis is confirmed as AVN of the femoral head.</think>

<answer>This image displays an AP radiograph of the pelvis, showing significant structural changes in the right femoral head. The observation of collapse, flattening, and mixed areas of sclerosis is consistent with a diagnosis of Avascular Necrosis (Osteonecrosis) of the femoral head.</answer>

Figure S6. A quantitative result of our ARMed



: Describe the main abnormalities that can be observed in the left hemisphere.



<think>This image shows an axial MRI scan of the brain. On the left side of the brain, there are regions with abnormal signal intensity. The affected area appears swollen, suggesting the presence of brain edema, which usually results from tissue injury, tumor, or inflammation. Additionally, part of the lesion demonstrates contrast enhancement, which indicates a brain enhancing tumor — typically associated with an active tumor region or disrupted blood-brain barrier. In the surrounding area, there are also regions without enhancement, which may represent non-enhancing tumor components, such as necrotic tissue or infiltrative tumor margins. Overall, these imaging features suggest multiple pathological changes on the left side of the brain, including edema and both enhancing and non-enhancing tumor regions.</think>

<answer>Brain Edema, Brain Enhancing Tumor, Brain Non-enhancing Tumor.</answer>

Figure S7. A quantitative result of our ARMed



: Describe the key finding and the likely medical intervention being demonstrated.



<think>The image shows an angiographic view of the abdominal vasculature. Contrast material highlights the arterial branches, and there is evidence of intervention within a branch of the inferior mesenteric or internal iliac artery. The presence of a metallic density resembling a microcoil suggests a therapeutic embolization procedure. Such procedures are commonly performed to control active bleeding by occluding the bleeding vessel. Given the location and vascular anatomy, this finding is most consistent with a microcoil embolization performed to treat a lower gastrointestinal bleed. The most concerning feature in this angiogram is the presence of an actively bleeding vessel requiring endovascular intervention.</think>

<answer>Microcoil embolization of a lower gastrointestinal bleed.</answer>

Figure S8. A quantitative result of our ARMed

References

- [1] 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. 2
- [2] Pritam Deka, Anna Jurek-Loughrey, and Deepak P. Evidence extraction to validate medical claims in fake news detection. In *International conference on health information science*, pages 3–15. Springer, 2022. 6
- [3] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing, 2020. 5
- [4] 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. 2
- [5] 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):1–10, 2018. 2
- [6] 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. 2
- [7] Shaohao Rui, Kaitao Chen, Weijie Ma, and Xiaosong Wang. Improving medical reasoning with curriculum-aware reinforcement learning. *arXiv preprint arXiv:2505.19213*, 2025. 3
- [8] 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. 2
- [9] 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. 2