

Step-CoT: Stepwise Visual Chain-of-Thought for Medical Visual Question Answering

Supplementary Material

A. Step-CoT Data Access and Format

The data can be accessed on HuggingFace at... The benchmark and code can be accessed on GitHub at... The dataset is organised in one main folder corresponding to three datasets. The dataset structure is shown as follows:

```
Step-CoT/  
|-- train.xlsx      # train paths  
|-- val.xlsx        # val paths  
|-- test.xlsx       # test paths  
|  
|-- dataset/  
|   |-- data.json  
|   |-- images/  
|       |-- 1_IM-0001-4001.dcm.png  
|       |-- 2_IM-0652-1001.dcm.png  
|       |-- 100..._370g35.png  
|       |-- 100..._ydw2jy.png  
|       |-- 1_1.png  
|       |-- 2_1.png
```

1. StepCoT/dataset/images: Contains the frontal-view original CXR images from all datasets. The image file-names follow the naming conventions of their respective source datasets.
2. StepCoT/dataset/data.json: Includes the stepwise VQA questions, answers, and associated reasoning for each image. The format of entries in the JSON file is shown as follows:

data.json

```
{  
  "patient_id": "***",  
  "image_path": "***",  
  "origin": "***",  
  "report": "The cardiac  
silhouette and mediastinum  
size are within normal  
limits. There is no  
pulmonary edema. There is  
no focal consolidation.  
There are no signs of a  
pleural effusion. There is  
no evidence of  
pneumothorax.",
```

```
"vqa_chain": [  
  {  
    "step": "Step 1",  
    "question": "Is there any  
abnormal radiodensity  
in the lungs?",  
    "options": [  
      "No abnormality",  
      "A1. Increased opacity",  
      "A2. Decreased opacity",  
      "A3. Mixed"  
    ],  
    "answer": "No abnormality",  
    "reasoning": "The report  
explicitly states no  
focal consolidation or  
other lung  
abnormalities, and the  
impression confirms a  
normal chest x-ray."  
  },  
  {  
    "step": "Step 2",  
    "question": "What is the  
distribution pattern of  
the abnormal findings?",  
    "options": [  
      "B1. Focal",  
      "B2. Scattered",  
      "B3. Diffuse"  
    ],  
    "answer": "N/A",  
    "reasoning": "No  
abnormalities exist,  
making distribution  
patterns irrelevant."  
  },  
  {  
    "step": "Step 3",  
    "question": "What is the  
predominant imaging  
pattern?",  
    "options": [  
      "C1. Consolidation",  
      "C2. Nodule",  
      "C3. Reticular",  
      "C4. Cavity",  
      "C5. Ground-glass  
opacity",  
    ],
```

```

    "answer": "N/A",
    "reasoning": "No
      abnormalities exist,
      making imaging patterns
      irrelevant."
  },
  {
    "step": "Step 4",
    "question": "Where is the
      main abnormality
      located?",
    "options": [
      "D1. Pleura/Chest wall",
      "D2. Lower lobes",
      "D3. Bilateral diffuse",
      "D4. Left upper lobe",
      "D5. Right upper lobe",
      "D6. Mediastinum"
    ],
    "answer": "N/A",
    "reasoning": "No
      abnormalities exist,
      making location
      irrelevant."
  }
}

"step": "Step 5",
"question": "Are the
  lesions well-defined or
  have any internal
  characteristics?",
"options": [
  "E1. Scarring/Fibrosis",
  "E2. Well-circumscribed",
  "E3. Spiculated",
  "E4. Cavitory",
],
"answer": "N/A",
"reasoning": "No
  abnormalities exist,
  making lesion
  characteristics
  irrelevant."
},
{
  "step": "Step 6",
  "question": "Do the
    lesions affect adjacent
    structures or cause
    structural changes?",
  "options": [
    "F1. Hyperinflation",
    "F2. Pleural effusion",
    "F3. Volume
      loss/atelectasis",
    "F4. No effect",

```

```

    "F5. Pneumothorax",
    "F6. Mediastinal shift",
  ],
  "answer": "N/A",
  "reasoning": "No
    abnormalities exist,
    making structural
    effects irrelevant."
},
{
  "step": "Step 7",
  "question": "What is the
    most likely
    radiographic
    diagnosis?",
  "options": [
    "1. Normal",
    "2. Infiltration",
    "3. Effusion",
    "4. Pneumonia",
    "5. Atelectasis",
    "6. Nodule",
    "7. Cardiomegaly",
    "8. Mass",
    "9. Pneumothorax",
  ],
  "answer": "Normal",
  "reasoning": "The
    impression concludes a
    normal chest x-ray with
    no acute findings."
}
}
}

```

The key columns are described as follows:

- **patient_id**: Each anonymous patient identifier corresponds to a single sample, and each sample contains only one anteroposterior (frontal) chest X-ray image. These identifiers ensure subject anonymity while allowing each CXR instance to be uniquely and consistently tracked throughout the dataset.
- **image_path**: A file reference pointing to the corresponding radiographic image for each sample. This field provides the exact storage location of the CXR image within the dataset directory structure, enabling reliable retrieval and consistent linkage between metadata entries and their associated medical images.
- **origin**: Dataset source information indicating the original dataset from which each sample was collected, ensuring traceability across heterogeneous data sources and enabling proper dataset-level stratification or analysis when required.
- **report**: Original radiology report text containing the

clinician’s narrative description of the image, including lesion characteristics, anatomical location, and other relevant diagnostic observations. All patient-identifiable or sensitive personal information has been fully removed to ensure compliance with privacy.

- **vqa_chain**: Seven-step diagnostic reasoning sequence:
 - (a) Detection step
 - (b) Lesion distribution step
 - (c) Radiographic pattern step
 - (d) Anatomical location step
 - (e) Morphologic feature step
 - (f) Secondary effects/associated signs step
 - (g) Diagnosis step

Each VQA step contains the question, multiple-choice options, selected answer, and clinical reasoning, creating a comprehensive framework for structured radiological interpretation and AI model training.

B. Detailed Information of Step-CoT

B.1. Data Acquisition

This work comprises original frontal CXR images and associated diagnostic text drawn from three public sources (totaling 10,068 CXR studies): (i) IU X-Ray [9], from which we use a subset of 3,749 CXR studies; (ii) PadChest-GR [4], from which we use a subset of 3,230 CXR studies; and (iii) Med-Image-Reports², from which we use a subset of 3,089 CXR studies. The combined corpus enables experiments on image–report alignment, grounded report generation, and stepwise VQA supervision across a broad mix of normal and abnormal cases.

B.1.1. IU X-Ray

The IU X-Ray dataset was collected by Indiana University and contains a large corpus of chest radiographs with associated radiology reports. Reports are organized under several headings (“Findings”, “Impression”, “Comparison”, and “Indication”); for this study, we use captions from the “Findings” section to provide descriptive image-level text. From the available corpus, we selected a curated subset of 3,749 frontal CXR studies that meet our inclusion criteria.

B.1.2. PadChest-GR

PadChest-GR is a bilingual chest X-ray benchmark derived from PadChest and tailored for Grounded Radiology Report Generation. The dataset includes clinician-validated annotations, bounding-box grounding for findings, and structured metadata; reports were processed (including sentence extraction, English translation, and label linking) to produce high-quality, sentence-level finding annotations. A team of radiologists further refined the corpus by removing low-

²<https://huggingface.co/datasets/zirui3/med-image-reports>

Table 7. Components of individual case analysis

Component	Description and Implementation
Clinical Questions	Sequential questions derived from radiological diagnostic pathways, progressing from abnormality detection to final diagnosis.
Answer Logic	<ul style="list-style-type: none"> • Direct answers from explicit report content. • “N/A” when the question is clinically irrelevant. • “No Answer” when information is missing, with inference attempted from available evidence.
Reasoning Chains	Each reasoning step required explicit justifications derived directly from report content or clinically valid inferences. This requirement ensures that model outputs remain grounded in available clinical evidence while maintaining the structured analytical rigor characteristic of expert radiological interpretation.

quality studies and annotating bounding boxes and categorical labels. For our experiments, we use a subset of 3,230 frontal-view CXR studies drawn from the PadChest-GR release.

B.1.3. Med-Image-Reports

The Med-Image-Reports benchmark aggregates chest X-ray studies and radiology-style captions from multiple public sources (OpenI, MIMIC-CXR, and PadChest). Original reports were preprocessed into concise diagnostic-style captions that describe both normal structures and clinically relevant abnormalities (e.g., cardiomegaly, pulmonary opacity, pleural effusion, pneumothorax, presence of devices). We adopt these standardized captions to ensure consistent supervision across heterogeneous origins and use a subset of 3,089 CXR studies from the Med-Image-Reports collection.

Preprocessing and harmonization Across all three sources, we restrict to frontal-view studies, extract or select the diagnostic caption, normalize common clinical terms, and remove records with missing or unusable captions. After harmonization, the resulting corpus contains 10,068 CXR studies used throughout the experiments reported in this paper.

B.2. Data Annotation

To construct a unified stepwise VQA supervision protocol across heterogeneous chest X-ray datasets, we perform automated annotation based on the corresponding radiology reports. For each CXR study, we utilize a large-scale language model (DeepSeek) to parse the paired report and extract clinically grounded information aligned with our seven-step diagnostic reasoning framework. Specifically, the model is prompted to identify key radiological observations, synthesize diagnostic cues, and populate each step

of the VQA schema with structured outputs, including the step-specific question, the corresponding answer, and a concise reasoning explanation. The structured components of the analysis framework are summarized in Table. 7. This automated annotation pipeline ensures consistent interpretation across datasets while preserving the clinical semantics embedded in expert-written reports. The complete prompt used for generating Step-CoT annotations is provided below.

LLM Prompt

Please analyze the chest X-ray findings based on the provided patient report (including patient_id, image_path, report, etc. (Replace according to the information contained in different datasets.) The output must strictly adhere to JSON format, containing patient_id, image_path, report, and vqa_chain. vqa_chain is an array where each element represents a step, including step, question, options, answer, and reasoning.

Answer rules: 1. Use "No Answer" if the report does not mention the relevant answer. 2. Use "N/A" if there is no tumor or the question is not applicable (e.g., when Step 1 answer indicates no abnormality, subsequent questions about tumors should be "N/A"). 3. For "No Answer" cases, attempt to infer the answer based on information from label, locations, sentence_en, label_group, classification, etc., if possible; otherwise, maintain "No Answer". 4. When providing answers, must consider previous steps' answers to establish a chain of thought.

Reasoning rules: 1. Must be written in English as complete sentences. 2. Based on report content or logical inference, directly analyze the reasons without additional modifiers (e.g., "the report describes"). 3. In reasoning, must consider previous steps' answers to establish a chain of thought (e.g., referencing Step 1 conclusions to explain Step 2 answers).

Format rules: 1. step, question, and options in vqa_chain must be directly copied from the following template to ensure textual consistency. 2. The entire output should be a JSON object.

Template vqa_chain structure: ["step": "Step 1", "question": "Is there any abnormal radiodensity in the lungs?", "options": ["No abnormality", "Increased opacity", "Decreased opacity", "Mixed"], "step": "Step 2", "question": "What is the distribution pattern of the abnormal findings?", "options": ["Focal", "Scattered", "Diffuse"], "step": "Step 3", "question": "What is the predominant imaging pattern?", "options": ["Consolidation", "Ground-glass opacity", "Reticular", "Cavity", "Nodule"], "step": "Step 4", "question": "Where

is the main abnormality located?", "options": ["Right upper lobe", "Left upper lobe", "Lower lobes", "Bilateral diffuse", "Pleura/Chest wall", "Mediastinum"], "step": "Step 5", "question": "Are the lesions well-defined or have any internal characteristics?", "options": ["Well-circumscribed", "Spiculated", "Cavitary", "Scarring/Fibrosis"], "step": "Step 6", "question": "Do the lesions affect adjacent structures or cause structural changes?", "options": ["No effect", "Mediastinal shift", "Volume loss/atelectasis", "Pleural effusion", "Pneumothorax", "Hyperinflation"], "step": "Step 7", "question": "What is the most likely radiographic diagnosis?", "options": ["Atelectasis", "Cardiomegaly", "Effusion", "Infiltration", "Mass", "Nodule", "G. Pneumonia", "Pneumothorax", "Normal"]]

B.2.1. LLM Prompt Consistency Experiment

We conduct an evaluation to quantify how different LLMs respond to an identical prompt for generating stepwise VQA annotations from chest X-ray reports. The same prompt and input set are submitted, verbatim, to three representative models (DeepSeek, ChatGPT, and G-Mini) under fixed decoding settings. For each report, we collect the structured JSON outputs and present side-by-side comparisons of the vqa_chain entries; the analysis focuses on per-step categorical agreement as well as differences in the free-text reasoning. This experimental design isolates prompt-driven variance by holding inputs and decoding parameters constant and provides direct, empirical evidence of how different LLMs interpret identical clinical text.

DeepSeek output

```
{
  "patient_id": "1",
  "image_path": "",
  "report": "The cardiac silhouette and mediastinum size are within normal limits. There is no pulmonary edema. There is no focal consolidation. There are no signs of a pleural effusion. There is no evidence of pneumothorax.",
  "vqa_chain": [
    {
      "step": "Step 1",
      "question": "Is there any abnormal radiodensity in the lungs?",
      "options": ["No abnormality", "Increased opacity", "Decreased opacity", "Mixed"],
      "answer": "No abnormality",
      "reasoning": "The report explicitly states no focal consolidation or other lung abnormalities, and the impression confirms a normal chest x-ray."
    }
  ]
}
```

```

},
{
  "step": "Step 2",
  "question": "What is the distribution pattern of the abnormal findings?",
  "options": ["Focal", "Scattered", "Diffuse"],
  "answer": "N/A",
  "reasoning": "No abnormalities exist, making distribution patterns irrelevant."
},
{
  "step": "Step 3",
  "question": "What is the predominant imaging pattern?",
  "options": ["Consolidation", "Ground-glass opacity", "Reticular", "Cavity", "Nodule"],
  "answer": "N/A",
  "reasoning": "No abnormalities exist, making imaging patterns irrelevant."
},
{
  "step": "Step 4",
  "question": "Where is the main abnormality located?",
  "options": ["Right upper lobe", "Left upper lobe", "Lower lobes", "Bilateral diffuse", "Pleura/Chest wall", "Mediastinum"],
  "answer": "N/A",
  "reasoning": "No abnormalities exist, making location irrelevant."
},
{
  "step": "Step 5",
  "question": "Are the lesions well-defined or have any internal characteristics?",
  "options": ["Well-circumscribed", "Spiculated", "Cavitary", "Scarring/Fibrosis"],
  "answer": "N/A",
  "reasoning": "No abnormalities exist, making lesion characteristics irrelevant."
},
{
  "step": "Step 6",
  "question": "Do the lesions affect adjacent structures or cause structural changes?",
  "options": ["No effect", "Mediastinal shift", "Volume loss/atelectasis", "Pleural effusion", "Pneumothorax", "Hyperinflation"],
  "answer": "N/A",
  "reasoning": "No abnormalities exist, making structural effects irrelevant."
},
{
  "step": "Step 7",
  "question": "What is the most likely radiographic diagnosis?",

```

```

"options": ["Atelectasis", "Cardiomegaly", "Effusion", "Infiltration", "Mass", "Nodule", "Pneumonia", "Pneumothorax", "Normal"],
"answer": "Normal",
"reasoning": "The impression concludes a normal chest x-ray with no acute findings."
}
]
}

```

ChatGPT output (example)

```

{
  "patient_id": "1",
  "image_path": "",
  "report": "The cardiac silhouette and mediastinum size are within normal limits. There is no pulmonary edema. There is no focal consolidation. There are no signs of a pleural effusion. There is no evidence of pneumothorax.",
  "vqa_chain": [
    {
      "step": "Step 1",
      "question": "Is there any abnormal radiodensity in the lungs?",
      "options": ["No abnormality", "Increased opacity", "Decreased opacity", "Mixed"],
      "answer": "A",
      "reasoning": "The report repeatedly negates pulmonary edema and focal consolidation; these explicit negative findings indicate no abnormal lung radiodensity on the frontal CXR ."
    },
    {
      "step": "Step 2",
      "question": "What is the distribution pattern of the abnormal findings?",
      "options": ["Focal", "Scattered", "Diffuse"],
      "answer": "N/A",
      "reasoning": "Given the absence of pathological findings in Step 1, distribution is not applicable for this case."
    },
    {
      "step": "Step 3",
      "question": "What is the predominant imaging pattern?",
      "options": ["Consolidation", "Ground-glass opacity", "Reticular", "Cavity", "Nodule"],
      "answer": "N/A",
      "reasoning": "No opacities or abnormal imaging patterns are described; therefore, pattern

```

```

classification cannot be assigned."
},
{
"step": "Step 4",
"question": "Where is the main abnormality located?",
"options": ["Right upper lobe", "Left upper lobe", "Lower lobes", "Bilateral diffuse", "Pleura/Chest wall", "Mediastinum"],
"answer": "N/A",
"reasoning": "No focal abnormality was detected in earlier steps, so localization is not applicable."
},
{
"step": "Step 5",
"question": "Are the lesions well-defined or have any internal characteristics?",
"options": ["Well-circumscribed", "Spiculated", "Cavitary", "Scarring/Fibrosis"],
"answer": "N/A",
"reasoning": "There are no lesions reported; lesion characterization is therefore not relevant."
},
{
"step": "Step 6",
"question": "Do the lesions affect adjacent structures or cause structural changes?",
"options": ["No effect", "Mediastinal shift", "Volume loss/atelectasis", "Pleural effusion", "Pneumothorax", "Hyperinflation"],
"answer": "N/A",
"reasoning": "Absence of abnormalities implies no structural effects such as effusion or pneumothorax."
},
{
"step": "Step 7",
"question": "What is the most likely radiographic diagnosis?",
"options": ["Atelectasis", "Cardiomegaly", "Effusion", "Infiltration", "Mass", "Nodule", "Pneumonia", "Pneumothorax", "Normal"],
"answer": "Normal",
"reasoning": "The combined negative findings and normal cardiac/mediastinal contours support a final impression of a normal chest radiograph."
}
]
}

```

Gemini output (example)

```

{
"patient_id": "1",

```

```

"image_path": "",
"report": "The cardiac silhouette and mediastinum size are within normal limits. There is no pulmonary edema. There is no focal consolidation. There are no signs of a pleural effusion. There is no evidence of pneumothorax.",
"vqa_chain": [
{
"step": "Step 1",
"question": "Is there any abnormal radiodensity in the lungs?",
"options": ["No abnormality", "Increased opacity", "Decreased opacity", "Mixed"],
"answer": "No abnormality",
"reasoning": "The report explicitly excludes edema, consolidation, effusion, and pneumothorax; therefore, no abnormal pulmonary radiodensity is present."
},
{
"step": "Step 2",
"question": "What is the distribution pattern of the abnormal findings?",
"options": ["Focal", "Scattered", "Diffuse"],
"answer": "N/A",
"reasoning": "No abnormalities are identified at Step 1, so distribution is not applicable."
},
{
"step": "Step 3",
"question": "What is the predominant imaging pattern?",
"options": ["Consolidation", "Ground-glass opacity", "Reticular", "Cavity", "Nodule"],
"answer": "N/A",
"reasoning": "Because no pathological patterns are described in the report, imaging pattern selection is not relevant."
},
{
"step": "Step 4",
"question": "Where is the main abnormality located?",
"options": ["Right upper lobe", "Left upper lobe", "Lower lobes", "Bilateral diffuse", "Pleura/Chest wall", "Mediastinum"],
"answer": "N/A",
"reasoning": "There is no focal abnormality to localize given the normal findings above."
},
{
"step": "Step 5",
"question": "Are the lesions well-defined or have any internal characteristics?",
"options": ["Well-circumscribed", "Spiculated", "Cavitary", "Scarring/Fibrosis"],

```

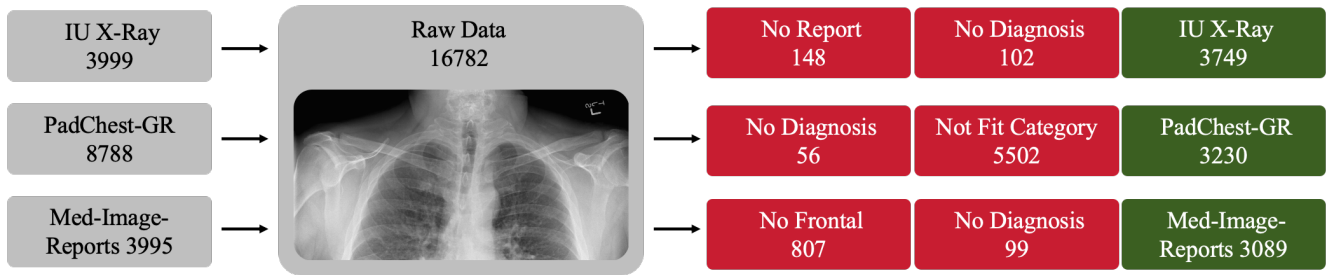


Figure 5. This study collected a total of 16,782 CXR samples in PNG format from three datasets, containing 3,999, 8,788, and 3,995 samples, respectively. After filtering, 10,068 samples were retained, yielding 10,068*7 QA pairs for training the stepwise Med-VQA task.

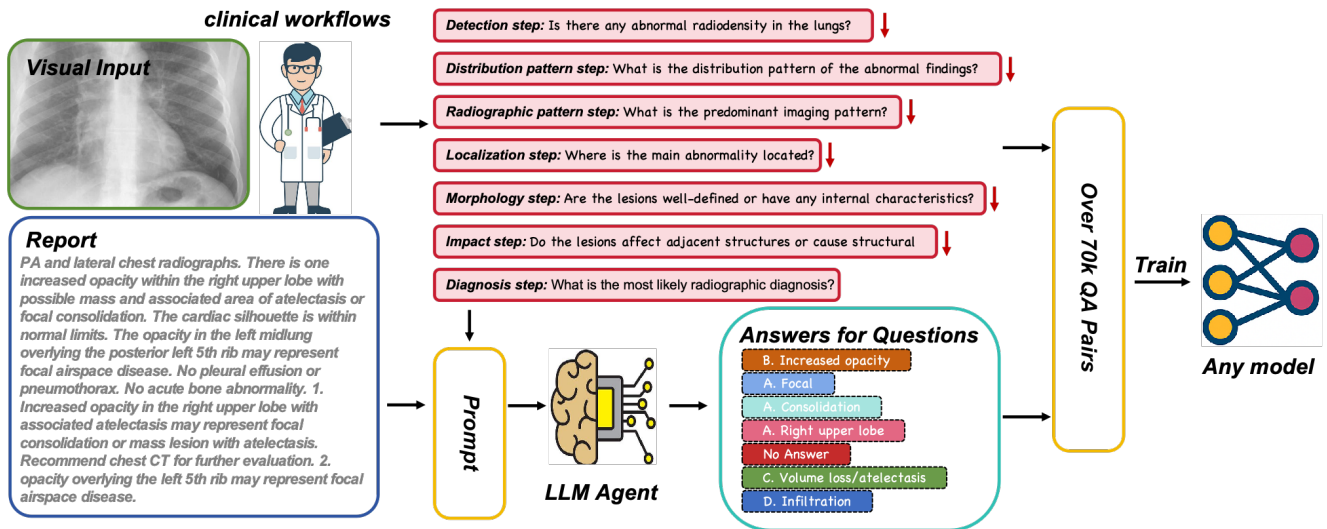


Figure 6. Preprocessing pipeline of the Step-CoT dataset.

```

"answer": "N/A",
"reasoning": "No lesions are reported; internal lesion
  characterization cannot be performed."
},
{
"step": "Step 6",
"question": "Do the lesions affect adjacent structures
  or cause structural changes?",
"options": ["No effect", "Mediastinal shift", "Volume
  loss/atelectasis", "Pleural effusion", "
  Pneumothorax", "Hyperinflation"],
"answer": "N/A",
"reasoning": "Given the absence of radiographic
  abnormalities, there is no evidence of structural
  impact."
},
{
"step": "Step 7",
"question": "What is the most likely radiographic
  diagnosis?",

```

```

"options": ["Atelectasis", "Cardiomegaly", "Effusion
", "Infiltration", "Mass", "Nodule", "Pneumonia
", "Pneumothorax", "Normal"],
"answer": "Normal",
"reasoning": "All reported findings are negative for
  acute pathology and cardiac/mediastinal
  contours are normal, yielding a radiographic
  impression of normal."
}
]
}

```

Comparative analysis of model outputs. All three models produce identical categorical answers for every step in this sample, indicating perfect per-step agreement on this case. Differences are confined to the free-text reasoning, which are stylistic and mildly variable in focus:

- DeepSeek (used in this study): concise, directly references the explicit negative findings; reasoning is compact and clinically focused.

- ChatGPT: slightly more formal and explicit about cross-step justification.
- Gemini: more narrative and slightly more verbose, highlighting report completeness.

Overall, the inter-model discrepancy for this case is minimal and largely stylistic. For downstream uses that depend only on categorical answers (the vqa_chain labels), the three models are functionally equivalent on this example. When evaluating the quality of the free-text reasoning, DeepSeek appears marginally preferable here because its explanations are concise and tightly coupled to explicit report statements. Moreover, DeepSeek has been shown in medical benchmark studies to excel in clinical reasoning and medical task performance [32, 39], surpassing or matching other leading LLMs in diagnostic and multi-modal reasoning tasks. This prior validation supports our choice of DeepSeek for medical report analysis, especially in a radiology VQA setting where faithful, medically grounded reasoning is critical.

B.3. Data Pre-Processing and Construct-processing

For the pre-processing of Step-CoT, we removed samples without diagnostic answers, without frontal CXR images, without reports, or those that did not fit the final diagnostic taxonomy. The remaining samples constitute the Step-CoT dataset, as illustrated in the Fig. 5.

For the Construct-processing of Step-CoT (shown as Fig. 6), we performed the following steps: (i) Clinical experts designed a stepwise question schema based on standard diagnostic workflows. (ii) We collected the clinical diagnostic reports corresponding to each CXR study. (iii) We constructed prompts by combining the clinical reports with the stepwise question schema and fed them into an LLM agent to derive step-specific answers from the report. (iv) The extracted answers were then paired with their corresponding questions to form QA supervision pairs, which were used for model training together with the associated CXR images.

B.4. Data Split

The dataset comprises 10,068 chest X-ray reports with comprehensive diagnostic annotations, systematically partitioned into training (70%), validation (15%), and test (15%) sets to ensure robust model evaluation. As detailed in Table 8, the dataset exhibits a natural class imbalance reflective of real-world clinical prevalence. The stratified partitioning strategy successfully maintained proportional representation of each diagnostic category across all splits, with minimal deviation from the target 70:15:15 distribution. This careful partitioning mitigates potential biases in model training and evaluation, particularly important given the substantial class imbalance.

Data provenance analysis (Table 8) reveals balanced contributions from three distinct sources, with each file propor-

tionally represented across the dataset splits. The source three datasets contributed 3,749 (37.2%), 3,230 (32.1%), and 3,089 (30.7%) samples, respectively, with consistent distribution patterns across training, validation, and test partitions. This multi-source composition enhances dataset diversity and reduces source-specific biases.

C. Detailed Analysis of Experimental Results

C.1. Stepwise benchmark results

In this section, we present the stepwise benchmark results on the Step-CoT dataset (shown in Table 9) and provide a detailed description of the experimental setup. The experimental setup is:

- Experimental details for vision foundation models. We compare with VisualBERT [24], CLIP [31], ALBEF [21], BLIP [22], FLAVA [35], and biomedclip [46]. Each instance provides a radiograph and the corresponding seven-step vqa_chain, with step questions formatted as “Question: [text]”. Labels follow the uniform mapping A–I → 0–8, with N/A as the final label; missing answers are set to -100. The data split is 70%/15%/15%. Images are resized to 224×224 and normalized, while questions are tokenized to a maximum of 128 tokens. Six representative visual foundation models are evaluated, where pooled visual and textual embeddings are fused via concatenation and fed into a lightweight MLP (two FC layers, hidden dim 768, ReLU). Training is performed with AdamW ($\text{lr} = 1 \times 10^{-5}$, weight decay 1×10^{-4}), batch size 8 for 50 epochs, using cosine learning rate scheduling and gradient clipping ($\text{max-norm} = 1.0$).

Across all seven diagnostic steps, the expanded benchmark reveals a clear and interpretable performance stratification across model families. Vision foundation models (VLMs) such as CLIP and BLIP improve moderately over pure visual models (shown in Table. 9), especially in specific steps requiring coarse visual pattern recognition (e.g., Morphologic feature step), but still exhibit systematically low sensitivity and overly conservative prediction behavior, leading to high specificity but failure to detect positive cases. Even domain-adapted BiomedCLIP—the strongest VLM—shows only partial gains: while accuracy and AUC improve across most steps, sensitivity remains $<25\%$ for nearly all tasks, indicating that contrastive alignment pre-training alone is insufficient for reconstructing intermediate reasoning. In contrast, enabling Step-CoT supervision consistently enhances performance across visual foundation models: models become less conservative, gain sensitivity, and improve F1 by leveraging structured intermediate reasoning. These results collectively confirm that Step-CoT introduces clinically meaningful reasoning supervision that bridges the gap between low-level visual recognition and high-level diagnostic inference, yielding improvements in

Table 8. Dataset distribution analysis for Step-CoT.

Category	Absolute Counts			Relative Distribution (%)		
	Train	Validation	Test	Train	Validation	Test
Diagnosis Categories						
Normal	4,750	1,018	1,019	70.0	15.0	15.0
Nodule	595	127	129	69.9	14.9	15.2
Atelectasis	511	109	111	69.9	14.9	15.2
Infiltration	506	108	109	69.9	14.9	15.2
Effusion	308	66	66	70.0	15.0	15.0
Pneumonia	193	41	42	69.9	14.9	15.2
Cardiomegaly	91	19	21	69.5	14.5	16.0
Mass	61	13	14	69.3	14.8	15.9
Pneumothorax	28	6	7	68.3	14.6	17.1
Diagnosis Subtotal	7,043	1,507	1,518	70.0	15.0	15.0
Data Sources						
IU X-Ray	2,584	563	602	70.0	15.3	16.3
PadChest-GR	2,297	461	472	71.1	14.3	14.6
Med-Image-Report	2,162	483	444	70.0	15.6	14.4
Source Subtotal	7,043	1,507	1,518	70.0	15.0	15.0

both factual precision and interpretability.

C.2. Comprehensive Evaluation of the Proposed Benchmark Method

In the experiments shown in Table 10, the Teacher model consistently achieved the highest and most stable multi-step performance across Accuracy, AUC, Sensitivity, Specificity, F1, and Precision. The distilled Student generally tracked the Teacher closely but exhibited step-dependent variability: in some steps the Student matched or slightly exceeded the Teacher, whereas in others it fell behind by a modest margin. Ablation analyses reveal a clear hierarchy of component importance. Removing the memory mechanism resulted in a consistent decrease in Sensitivity and F1, indicating that cross-step state accumulation supports the integration of evidence along the diagnostic chain. Removing the textual prompt led to the most pronounced reductions, particularly in AUC, Precision, and Sensitivity, confirming the necessity of question-guided multi-modal grounding. Notably, steps dominated by strong visual cues retained relatively high Accuracy and Specificity even under ablations, whereas steps that require subtle inter-step or contextual reasoning (e.g., Radiographic pattern step and Diagnosis step) showed marked declines. Across all models and steps, Sensitivity remains lower than Accuracy or Specificity, reflecting intrinsic challenges in the dataset: the prevalence of negative or normal cases, the subtlety of certain pathological manifestations, and the compounding effect of multi-step reasoning where early-stage uncertainty can reduce downstream detection of true positives. These characteristics highlight that the dataset encapsulates clinically relevant difficulties, making it a valuable benchmark

for evaluating multi-step diagnostic reasoning and for guiding the development of methods that can better handle low-prevalence, subtle abnormalities.

C.3. Clinical Expert Evaluation

We evaluated 200 randomly sampled cases and compared three outputs per case (Clinician, Teacher, Student) across the seven-step VQA chain (Table 11). The Teacher model yields the strongest overall performance: it improves accuracy and F1 for nearly every step relative to both the Student and clinician baselines (e.g., Detection accuracy 88.51% vs. Clinician 72.12% and Student 80.37%; Detection F1 57.24% vs. Clinician 37.86% and Student 50.84%). These gains are accompanied by substantially higher sensitivity (Detection sensitivity: Teacher 58.36% vs. Clinician 36.84%). The Student—distilled from the Teacher—retains most of this structured competence: on mid-level reasoning steps such as Distribution and Location, the Student surpasses clinicians (Distribution accuracy 72.63% vs. Clinician 66.02%; Location accuracy 69.46% vs. Clinician 66.03%), and its per-step performance typically lies within roughly 5–10 percentage points of the Teacher. Notable exceptions remain: for the Secondary Effects step, the clinician’s accuracy (90.97%) exceeds both Teacher (75.24%) and Student (73.97%), suggesting that certain effect-related judgments still rely on expert clinical context or multi-view/longitudinal information not available to the models. In sum, these results demonstrate that (i) explicit stepwise supervision (Teacher) materially improves recall, F1 and overall coherency compared to standard baselines, (ii) knowledge distillation produces a compact Student that preserves most gains with modest performance loss, and

Table 9. Stepwise benchmark results: Vision foundation models (%). The best value in each column is **bold** and the second best is underlined.

Model	Step	Accuracy	AUC	Sensitivity	Specificity	F1-Score	Precision
VisualBERT	Detection step	75.0	71.0	13.5	93.5	33.1	34.0
	Lesion distribution step	70.6	58.1	13.0	93.0	30.0	28.1
	Radiographic pattern step	74.1	62.0	14.6	93.2	20.1	17.5
	Anatomical location step	68.0	60.0	13.0	92.8	17.5	15.0
	Morphologic feature step	84.0	58.5	13.5	93.1	33.0	32.1
	Secondary effects/associated signs step	66.1	61.0	12.8	92.5	16.0	14.5
	Diagnosis step	65.5	63.1	10.1	91.8	9.6	8.6
CLIP	Detection step	74.9	65.2	30.4	80.2	40.8	46.1
	Lesion distribution step	78.7	62.5	29.0	79.0	28.3	29.8
	Radiographic pattern step	80.8	66.1	18.1	84.7	17.6	21.2
	Anatomical location step	76.6	63.5	17.1	88.4	16.2	15.9
	Morphologic feature step	87.7	67.0	21.1	81.8	20.7	22.5
	Secondary effects/associated signs step	75.9	59.5	16.4	87.6	15.6	15.6
	Diagnosis step	69.2	52.6	12.0	89.6	11.0	11.7
ALBEF	Detection step	82.5	80.5	17.5	95.0	40.5	42.1
	Lesion distribution step	79.0	76.1	18.0	<u>94.4</u>	42.1	44.5
	Radiographic pattern step	81.6	82.6	<u>19.0</u>	94.7	31.6	34.0
	Anatomical location step	75.5	70.0	16.5	94.1	20.5	18.6
	Morphologic feature step	89.5	78.1	18.6	<u>95.0</u>	48.1	<u>54.1</u>
	Secondary effects/associated signs step	73.5	78.5	17.8	93.8	25.6	27.0
	Diagnosis step	72.0	75.1	18.5	93.2	19.1	20.0
BLIP	Detection step	81.8	78.0	15.3	94.9	38.4	40.1
	Lesion distribution step	76.8	73.5	16.2	94.3	40.9	43.2
	Radiographic pattern step	80.9	81.7	17.1	94.6	29.8	33.0
	Anatomical location step	74.1	69.3	15.3	94.0	19.6	17.7
	Morphologic feature step	88.4	77.3	16.8	94.6	45.4	52.9
	Secondary effects/associated signs step	73.0	77.2	16.3	93.7	23.5	25.1
	Diagnosis step	71.0	74.9	17.2	92.9	18.2	18.3
FLAVA	Detection step	76.2	72.3	14.2	93.9	35.1	35.2
	Lesion distribution step	72.0	59.9	13.9	93.4	32.2	30.0
	Radiographic pattern step	77.0	63.3	15.0	93.8	21.1	18.9
	Anatomical location step	69.6	61.0	14.4	93.3	18.0	16.2
	Morphologic feature step	85.7	59.6	14.7	93.8	35.6	34.7
	Secondary effects/associated signs step	67.7	62.4	14.1	93.0	17.7	15.8
	Diagnosis step	67.1	64.2	10.9	92.3	10.4	9.3
biomedclip	Detection step	83.1	82.5	22.1	95.2	45.1	50.0
	Lesion distribution step	80.3	81.8	21.5	94.8	43.1	<u>46.1</u>
	Radiographic pattern step	82.5	84.0	23.0	<u>95.0</u>	33.1	40.1
	Anatomical location step	76.2	70.5	21.1	94.5	25.1	22.1
	Morphologic feature step	88.9	82.1	22.6	95.1	<u>47.1</u>	60.1
	Secondary effects/associated signs step	74.6	82.0	22.0	94.0	30.1	35.0
	Diagnosis step	73.1	76.0	21.7	93.5	22.1	25.0
Teacher	Detection step	91.8	88.3	63.1	94.6	61.6	<u>60.3</u>
	Lesion distribution step	84.6	<u>89.7</u>	51.4	93.8	49.8	61.7
	Radiographic pattern step	84.8	90.2	45.0	96.5	43.6	53.2
	Anatomical location step	77.1	89.5	45.7	95.3	44.9	57.7
	Morphologic feature step	<u>89.4</u>	95.1	46.1	94.1	40.5	39.2
	Secondary effects/associated signs step	79.7	86.4	41.4	95.3	40.0	41.4
	Diagnosis step	78.3	<u>89.5</u>	46.0	96.6	45.3	<u>48.7</u>
Student	Detection step	91.8	<u>85.0</u>	<u>62.4</u>	<u>93.8</u>	61.6	60.9
	Lesion distribution step	<u>83.4</u>	90.2	<u>46.6</u>	92.3	<u>42.6</u>	<u>39.5</u>
	Radiographic pattern step	<u>84.2</u>	<u>89.2</u>	<u>40.1</u>	94.2	<u>39.9</u>	53.2
	Anatomical location step	<u>76.9</u>	<u>88.7</u>	<u>44.1</u>	94.7	<u>42.6</u>	<u>45.7</u>
	Morphologic feature step	89.3	<u>94.2</u>	<u>41.4</u>	<u>88.3</u>	40.2	39.2
	Secondary effects/associated signs step	<u>79.6</u>	<u>85.6</u>	<u>39.1</u>	<u>94.2</u>	<u>37.9</u>	41.4
	Diagnosis step	<u>77.5</u>	90.0	<u>41.8</u>	<u>96.0</u>	<u>39.5</u>	49.9

Table 10. Comprehensive performance comparison (per-step metrics) between Teacher, Student, and ablation variants (%). The best value in each column is **bold** and the second best is underlined.

Model	Step	Accuracy	AUC	Sensitivity	Specificity	F1-Score	Precision
Teacher	Detection step	91.8	88.3	63.1	94.6	61.6	<u>60.3</u>
	Lesion distribution step	84.6	<u>89.7</u>	51.4	93.8	49.8	61.7
	Radiographic pattern step	84.8	90.2	45.0	96.5	43.6	53.2
	Anatomical location step	77.1	89.5	45.7	95.3	44.9	57.7
	Morphologic feature step	89.4	95.1	46.1	94.1	40.5	<u>39.2</u>
	Secondary effects/associated signs step	79.7	86.4	41.4	95.3	40.0	41.4
	Diagnosis step	78.3	<u>89.5</u>	46.0	96.6	45.3	<u>48.7</u>
Student	Detection step	91.8	<u>85.0</u>	<u>62.4</u>	<u>93.8</u>	61.6	60.9
	Lesion distribution step	<u>83.4</u>	90.2	<u>46.6</u>	<u>92.3</u>	<u>42.6</u>	<u>39.5</u>
	Radiographic pattern step	<u>84.2</u>	<u>89.2</u>	<u>40.1</u>	<u>94.2</u>	<u>39.9</u>	53.2
	Anatomical location step	<u>76.9</u>	<u>88.7</u>	<u>44.1</u>	<u>94.7</u>	<u>42.6</u>	<u>45.7</u>
	Morphologic feature step	<u>89.3</u>	<u>94.2</u>	<u>41.4</u>	<u>88.3</u>	<u>40.2</u>	39.2
	Secondary effects/associated signs step	<u>79.6</u>	<u>85.6</u>	<u>39.1</u>	<u>94.2</u>	<u>37.9</u>	41.4
	Diagnosis step	<u>77.5</u>	90.0	<u>41.8</u>	<u>96.0</u>	<u>39.5</u>	49.9
w/o Memory	Detection step	73.7	78.5	49.2	88.6	48.2	47.6
	Lesion distribution step	69.6	83.0	38.0	86.5	35.8	34.1
	Radiographic pattern step	72.3	84.5	33.5	89.0	33.1	46.8
	Anatomical location step	63.2	81.2	36.4	88.0	35.5	38.2
	Morphologic feature step	70.2	87.5	35.1	83.9	33.9	33.0
	Secondary effects/associated signs step	67.4	80.0	33.0	88.5	32.6	35.9
	Diagnosis step	65.5	84.8	35.2	89.8	34.9	44.1
w/o Text	Detection step	81.5	82.5	54.8	91.8	53.4	52.8
	Lesion distribution step	76.1	86.5	42.5	90.0	39.8	37.1
	Radiographic pattern step	77.3	87.3	37.2	92.1	37.0	49.6
	Anatomical location step	69.3	86.0	40.8	91.5	39.5	42.3
	Morphologic feature step	79.6	91.5	39.0	87.6	36.8	35.9
	Secondary effects/associated signs step	73.2	83.5	35.7	91.4	35.0	38.0
	Diagnosis step	72.1	87.2	38.6	93.0	38.1	47.5

Table 11. Clinical expert evaluation (%), per-step comparison between Clinicians, Teacher, and Student (N=200). The best value in each column is **bold**.

Model	Step	Accuracy	Sensitivity	Specificity	F1-Score
Clinician	Detection step	72.1	36.8	82.7	37.9
	Lesion distribution step	66.0	28.0	85.3	27.1
	Radiographic pattern step	69.1	25.1	88.7	27.1
	Anatomical location step	66.0	34.0	91.1	24.7
	Morphologic feature step	66.1	21.3	85.3	19.0
	Secondary effects/associated signs step	91.0	33.4	93.7	25.0
	Diagnosis step	73.1	37.2	93.8	34.4
Teacher	Detection step	88.5	58.4	92.1	57.2
	Lesion distribution step	78.4	45.6	91.3	43.8
	Radiographic pattern step	79.6	40.8	94.2	39.2
	Anatomical location step	72.8	41.3	93.6	40.1
	Morphologic feature step	84.8	42.3	91.8	36.9
	Secondary effects/associated signs step	<u>75.2</u>	36.9	<u>93.5</u>	35.4
	Diagnosis step	79.8	41.6	95.1	41.3
Student	Detection step	<u>80.4</u>	<u>52.1</u>	<u>89.8</u>	<u>50.8</u>
	Lesion distribution step	<u>72.6</u>	<u>38.3</u>	<u>88.0</u>	<u>35.4</u>
	Radiographic pattern step	<u>70.2</u>	<u>34.6</u>	<u>91.8</u>	<u>32.8</u>
	Anatomical location step	<u>69.5</u>	<u>36.8</u>	<u>91.3</u>	<u>35.4</u>
	Morphologic feature step	<u>83.3</u>	<u>35.1</u>	<u>88.5</u>	<u>30.3</u>
	Secondary effects/associated signs step	74.0	<u>32.5</u>	92.2	30.8
	Diagnosis step	<u>68.5</u>	<u>35.9</u>	<u>93.9</u>	<u>34.6</u>

Table 12. Computational Requirements Comparison (batch-size=4). **Note:** Bold = best (lowest); underline = 2nd best (2nd lowest).

Model	Params	Inf_sample	Inf_batch	Memory_peak
CLIP	151.81 M	1.55 ms	6.20 ms	<u>599.43 M</u>
FLAVA	242.54 M	4.39 ms	17.57 ms	971.42 M
BLIP	225.25 M	<u>1.46 ms</u>	<u>5.86 ms</u>	904.27 M
VisualBERT	111.98 M	0.03 ms	0.12 ms	438.22 M
Teacher	283.66 M	22.81 ms	91.23 ms	1219.06 M
Student	<u>151.56 M</u>	1.51 ms	6.02 ms	1219.72 M

(iii) remaining gaps (especially on clinically nuanced steps) point to limits of single-view supervision and motivate combining Step-CoT with richer context or human-in-the-loop verification for deployment.

C.4. Computational Efficiency Analysis

The computational efficiency analysis reveals significant insights across model architectures (as shown in Table 12). We compare with ResNet18 [17], ResNet50 [17], DenseNet121 [18], EfficientNet-B3 [37], CLIP [31], FLAVA [35], BLIP [22], and VisualBERT [24]. Traditional CNN models (ResNet18, ResNet50, DenseNet121, EfficientNet-B3) demonstrate lightweight parameter footprints (7.03-23.67M) with excellent inference speeds (0.20-1.30ms per sample), though DenseNet121 shows relatively higher latency. Multimodal models exhibit substantial parameter increases, with CLIP (151.81M) and FLAVA (242.54M) requiring significantly more computational resources, while BiomedCLIP stands out as exceptionally efficient (0.53M parameters, 0.07ms inference). Notably, our proposed Teacher-Student framework achieves remarkable efficiency gains: the Student model reduces parameters by 46.6% compared to the Teacher (283.66M to 151.56M) while achieving a 15 \times speedup in both single-sample (22.81ms to 1.51ms) and batch inference (91.23ms to 6.02ms). The Student model demonstrates competitive efficiency with CLIP despite similar parameter counts, though memory consumption remains a challenge (1219.72MB) across both our models, suggesting future optimization opportunities for deployment in resource-constrained environments.

D. Detail Method

This appendix provides a complete and reproducible description of the teacher model, the student model, the GAT-based memory, the distillation losses (KD and CH), and the training procedure used in our experiments.

D.1. Notation

Let B denote batch size, S be the number of reasoning steps, and d denote the hidden dimension used in the teacher (d_T) and d_S in the student. For step s teacher logits are $\ell_t^{(s)} \in \mathbb{R}^{B \times C_s}$ and student logits $\ell_s^{(s)} \in \mathbb{R}^{B \times C_s}$ where C_s is the number of classes for step s .

D.2. Teacher Model

Text encoding. A shared transformer-based encoder (we use BERT) maps each step prompt to a CLS embedding:

$$\mathbf{t}_s^{(b)} = \text{BERT}(\text{prompt} * s^{(b)})[CLS] \in \mathbb{R}^{d^*T}, \quad (7)$$

$$s = 1, \dots, S, b = 1, \dots, B.$$

Collect the S step vectors into $\mathbf{T}^{(b)} = [\mathbf{t}_1^{(b)}, \dots, \mathbf{t}_S^{(b)}]$.

Memory node initialization. A learnable memory vector $\mathbf{m} * 0 \in \mathbb{R}^{d^*T}$ is registered and expanded to the batch: $\mathbf{m}_0^{(b)} = \mathbf{m}_0, \forall b$.

Node set. For each example, we form nodes

$$\mathcal{N}^{(b)} = \mathbf{t}_1^{(b)}, \dots, \mathbf{t}_S^{(b)}, \mathbf{m}_0^{(b)}, \quad (8)$$

ordered so that the memory node is the last one.

Stacked GAT memory update. We apply L stacked multi-head GAT layers. For a single head, we first linearly project node features:

$$\tilde{\mathbf{h}} * i = W\mathbf{h} * i \in \mathbb{R}^{d'}, \quad (9)$$

and compute pairwise attention logits

$$e * ij^{(h)} = \text{LeakyReLU}(\mathbf{a} * \text{src}^{(h)\top} \tilde{\mathbf{h}} * i + \mathbf{a} * \text{dst}^{(h)\top} \tilde{\mathbf{h}} * j), \quad (10)$$

where h indexes attention heads and $\mathbf{a} * \text{src}^{(h)}, \mathbf{a} * \text{dst}^{(h)} \in \mathbb{R}^{d'}$ are learned. Normalize across destination nodes:

$$\alpha * ij^{(h)} = \frac{\exp(e_{ij}^{(h)})}{\sum_{j'} \exp(e_{ij'}^{(h)})}. \quad (11)$$

Head outputs are aggregated and (optionally) concatenated across heads to produce updated node features. Residual projection and LayerNorm follow each GAT layer. After L layers we obtain updated nodes $\mathbf{t}'_s, \mathbf{m}'$ where \mathbf{m}' is the updated memory node.

Step context fusion and prediction. For step s we extract the updated step node \mathbf{t}'_s and the updated memory node \mathbf{m}' , then fuse:

$$\mathbf{c}_s = \text{Fusion}([\mathbf{t}'_s; \mathbf{m}']) \in \mathbb{R}^{d^*T}. \quad (12)$$

In implementation, the *fusion_proj* is a linear layer with LayerNorm, ReLU, and dropout.

The step-specific prediction head (implemented in teacher model) uses a CLIP image encoder to compute an image embedding $\mathbf{v} \in \mathbb{R}^{d_T}$ (projected to d_T), then predicts logits

$$\ell_t^{(s)} = f^{(s)}(\mathbf{v}, \mathbf{c}_s) \in \mathbb{R}^{C_s}, \quad (13)$$

where $f^{(s)}$ concatenates \mathbf{v} and \mathbf{c}_s and passes them through a small MLP classifier.

Memory write-back. After producing logits $\ell_t^{(s)}$, we convert logits $\ell_t^{(s)}$ to a prediction embedding that is written back to memory. Concretely:

$$\mathbf{p}^{(s)} = \text{softmax}(\ell_t^{(s)}) \in \mathbb{R}^{B \times C_s}, \quad (14)$$

and using the classifier weights $W_{\text{cls}}^{(s)} \in \mathbb{R}^{C_s \times d_T}$ we form

$$\mathbf{e}^{(s)} = \mathbf{p}^{(s)} W_{\text{cls}}^{(s)} \in \mathbb{R}^{B \times d_T}. \quad (15)$$

A learned linear map *pred2mem* projects $\mathbf{e}^{(s)}$ to memory space, and a GRUCell updates:

$$\mathbf{m} * \text{new} = \text{GRUCell}(\text{pred2mem}(\mathbf{e}^{(s)}), \mathbf{m}'). \quad (16)$$

The updated memory $\mathbf{m} * \text{new}$ replaces the last node before processing the next step, enabling sequential flow.

D.3. Student Model

The student uses a frozen CLIP visual encoder to extract image features $\mathbf{v} * S \in \mathbb{R}^{d^*S}$ followed by a projection to the student’s hidden dim. A sequence of S light linear heads $g^{(s)}$ produce logits $\ell_s^{(s)} = g^{(s)}(\mathbf{v}_S)$. The student updates its internal feature between steps via a light residual update to simulate information flow (chain-style).

D.4. Distillation losses

For each valid example (where we mask invalid steps using a dataset-provided mask) and for each step s , we employ three losses.

1. Supervised cross-entropy.

$$\mathcal{L} * \text{CE}^{(s)} = -\frac{1}{N_s} \sum_{*i \in \mathcal{I} * s} \log p * s^{(i)}(y^{(i,s)}), \quad (17)$$

where \mathcal{I}_s indexes valid examples in the batch for step s and $N_s = |\mathcal{I}_s|$.

2. Soft label KD with temperature T . Let $\tilde{p}_t^{(i,s)} = \text{softmax}(\ell_t^{(i,s)}/T)$ and $\tilde{p} * S^{(i,s)} = \text{softmax}(\ell_S^{(i,s)}/T)$. The KD loss is

$$\mathcal{L} * \text{KD}^{(s)} = T^2 \cdot \text{KL}(\log \tilde{p}_S^{(s)}, |\tilde{p}_t^{(s)}) \quad (18)$$

computed over the valid subset.

3. Channel/relation alignment (CH). This term encourages the student to match the teacher’s inter-example similarity structure of image features. Denote teacher image features for step s as $U \in \mathbb{R}^{n \times d_p}$ and student image features as $V \in \mathbb{R}^{n \times d_p}$ after projecting to a common $\text{proj_dim} = p$ (via learnable linear maps for teacher and student). We apply a softmax across feature-dim to make per-example distributions:

$$K_U = \text{softmax}(U/T), \quad K_V = \text{softmax}(V/T) \quad (19)$$

(Each row sums to 1.) Define the empirical Gram matrices

$$M_U = K_U K_U^\top \quad \text{and} \quad M_V = K_V K_V^\top. \quad (20)$$

With centering matrices $C = I - \frac{1}{n} \mathbf{1}\mathbf{1}^\top$ we compute the HSIC-style scalars:

$$\begin{aligned} h_{UU} &= \text{tr}(CM_U C), \\ h_{VV} &= \text{tr}(CM_V C), \\ h_{UV} &= \sum ((CM_U C) \odot (CM_V C)) \end{aligned} \quad (21)$$

(Implementation uses elementwise product, then sum to form an inner product of centered Gram matrices). We form a similarity weight

$$w_{\text{fw}} = \frac{h_{UV}}{\sqrt{(h_{UU} + \varepsilon)(h_{VV} + \varepsilon)}}, \quad (22)$$

Then define the CH loss as a weighted divergence between the projected soft features:

$$\mathcal{L} * \text{CH}^{(s)} = w * \text{fw} \cdot \text{KL}(\log K_V, |, K_U). \quad (23)$$

Intuitively, if teacher/student feature similarity structures align strongly (w_{fw} large), we penalize their per-example soft feature mismatch more strongly.

Total student step loss. For step s we combine:

$$\mathcal{L} * \text{student}^{(s)} = \mathcal{L} * \text{CE}^{(s)} + \alpha_{\text{KD}} \mathcal{L} * \text{KD}^{(s)} + \alpha * \text{CH} \mathcal{L}_{\text{CH}}^{(s)}. \quad (24)$$

Teacher loss. Optionally, the teacher is trained with supervised cross-entropy:

$$\mathcal{L} * \text{teacher} = \sum *s = 1^S \mathcal{L}_{\text{CE}}^{(s)}. \quad (25)$$

In our implementation, we support an initial *teacher pre-train* stage (a small number of epochs) where only the teacher is updated to stabilize downstream distillation.

D.5. Training algorithm (implementation details)

1. Initialize teacher, student, and two separate optimizers (teacher and student). Initialize teacher memory and GAT parameters.
2. Optionally, run E_{pre} teacher-only epochs where only \mathcal{L}^* teacher is minimized (teacher supervised pretrain).
3. For each training batch:
 - (a) Teacher update (if enabled): compute teacher logits for each step and supervised CE loss, backpropagate, and step the teacher optimizer.
 - (b) Teacher forward (fresh): run the teacher (no grads) to obtain detached logits and projected image features used as KD/CH targets.
 - (c) Student update: compute student logits and student image features; for each step compute \mathcal{L}^* student^(s) using detached teacher outputs; sum over steps and step the student optimizer.
4. Validate teacher and student separately and save the best models using mean validation accuracy across steps.

D.6. Key hyper-parameters (defaults used in our experiments)

- Teacher hidden dim: $d_T = 768$.
- Student hidden dim: $d_S = 512$.
- GAT heads: $H = 4$, GAT layers $L = 2$, hence final GAT output dimension $H \times (\text{per-head-dim}) = d_T$.
- Temperature: $T = 2.0$.
- CH projection dimension: $p = 256$.
- KD weight: $\alpha_{\text{KD}} = 0.5$, CH weight $\alpha_{\text{CH}} = 1.0$.
- Teacher pretrain epochs: $E_{\text{pre}} = 2$ (optional).
- Optimizers: AdamW with teacher LR $\approx 5e-5$, student LR $\approx 1e-4$.