# BEYOND CAPTIONS: CONTEXT-GROUNDED RECONSTRUCTION FOR BIOMEDICAL MULTIMODAL CONTINUED PRETRAINING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Biomedical figures are explained not by captions alone but by body-text passages that discuss them. Yet current multimodal corpora typically reduce figures to isolated image-caption pairs, discarding this crucial context. Existing pipelines either omit this context or append it without enforcing the figure references that support each attachment, which can create unsupported image-text attachments and incoherent discourse. We introduce context-grounded reconstruction, a source-grounded framework that converts PubMed Central Open Access (PMC-OA) records into referentially coherent interleaved sequences. It recovers captions and source text, attaches context only through article-native figure references, repairs non-contiguous context, and prunes unsupported images. Starting from these reconstructed sequences, PMC-InterCPT first filters records for text quality and medical relevance, then applies evidence-aware allocation to form a 9.63B-token corpus for continued pretraining (CPT) of generative medical MLLMs. With fixed supervised fine-tuning (SFT), PMC-InterCPT improves Qwen3.5-4B-Base by 1.46 medical-average points and 3.11 general/scientific-average points over a token-matched raw source control, and surpasses a 42% larger raw-data run. Gains transfer to Qwen3.5-2B-Base and LLaVA-OneVision-1.5-4B-Base. Controlled ablations show that context-grounded reconstruction, rather than simply appending article context, is central to useful biomedical multimodal CPT.

## 1 INTRODUCTION

Multimodal large language models (MLLMs) benefit from interleaved image-text pretraining that preserves rich visual and textual evidence in a common sequence (Lin et al., 2024). In biomedical literature, that evidence is rarely contained in a caption alone. The surrounding body text can identify cohorts, experimental settings, comparisons, and claims that make a figure interpretable. It also frequently refers to several figures jointly. A biomedical CPT corpus that retains this structure could provide substantially richer training signal than an image-caption archive.

Existing resources extracted from PubMed Central Open Access (PMC-OA), including BIOMEDICA (Lozano et al., 2025), expose biomedical figures at unprecedented scale. BIOMEDICA contains more than 24 million image-caption pairs and extensive metadata, but its primary training representation is a figure paired with a caption. This organization leaves figure-referencing article text unused. More importantly, simply concatenating nearby text is unsafe: a paragraph can mention multiple figures, occur far from other retained contexts, or remain after the figure it supports has been removed. We therefore formulate the task as *grounded reconstruction*: rebuilding article-native records into referentially coherent interleaved sequences while preserving the figure references that justify each image-context association.

We propose context-grounded reconstruction, a reference-constrained framework for constructing biomedical CPT corpora. Rather than treating article text as generic auxiliary text, it uses article-native figure references as structural evidence for context attachment. The framework begins with source-grounded recovery, which restores provenance-linked captions and normalizes source text while preserving reference anchors. Reference-constrained interleaving then forms multi-figure sequences without duplicating shared context. Coherence-constrained curation detects discontinuities



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

among retained source passages, preserves locally supported context. These operations yield a candidate pool in which each retained image–caption–context relation remains traceable to the source article. Semantic curation filters this pool for textual usability and medical relevance at full-corpus scale. Finally, evidence-aware allocation shapes the CPT learning signal by regulating the model's exposure to complementary forms of scientific evidence: biomedical visual observations, quantitative results, mechanistic structures, and auxiliary context.

We instantiate this process as PMC-InterCPT and evaluate it as CPT data for generative medical MLLMs under a fixed SFT stage. Rather than validating the corpus only through CLIP-style contrastive representation scores, we assess its effect on generative medical VQA, medical reasoning, and scientific figure understanding, using token-matched raw controls and controlled ablations.

Our main contributions are:

- **Reference-constrained interleaving**, a reconstruction formulation that converts PMC figure records and article text into referentially coherent interleaved sequences, including multi-figure relations and context-supported image retention.
- **Coherence-constrained curation**, a procedure that turns reconstructed sequences into usable biomedical CPT data. It repairs context, prunes unsupported images, and filters for text quality and medical relevance.
- **Evidence-aware allocation**, a four-bucket taxonomy and controlled allocation scheme that regulates the model's exposure to clinical visual, quantitative, and mechanistic scientific evidence. PMC-InterCPT is the resulting 9.63B-token corpus of this allocation.
- **Generative MLLM gains.** With fixed SFT, PMC-InterCPT outperforms the token-matched raw baseline by 1.46 points on medical tasks and 3.11 points on general/scientific tasks, with gains transferring across backbones.

## 2 Related Work

### 2.1 Medical Multimodal Large Language Models

Recent years have witnessed rapid development of medical multimodal large language models (MLLMs), including LLaVA-Med (Li et al., 2023), Med-Flamingo (Moor et al., 2023), HuatuoGPT-Vision (Chen et al., 2024), MAIRA-2 (Bannur et al., 2024), and LingShu (Xu et al., 2025), which adapt general-purpose VLMs to clinical and biomedical domains through instruction tuning or domain-adaptive pretraining. Despite strong performance on medical VQA tasks, these models typically rely on small curated instruction datasets or general-domain pretrained weights, without access to large-scale high-quality medical interleaved CPT data. Our work addresses this gap by constructing and validating a CPT dataset specifically designed for this purpose.

### 2.2 Biomedical Multimodal Datasets from Literature

PMC-OA (Lin et al., 2023), PMC-15M (Zhang et al., 2023a), Open-PMC-18M (Baghbanzadeh et al., 2025), and BIOMEDICA (Lozano et al., 2025) demonstrate the value of PMC as a scalable source of biomedical image-caption data. These resources differ in scale, subfigure handling, and metadata, but their predominant training unit is an image-caption pair. Their reported empirical validation focuses on CLIP-style contrastive medical vision-language representation learning, rather than evaluating whether the resulting data improves generative medical MLLMs. PubMedVision reformats PubMed image-text pairs into medical VQA data using an MLLM (Chen et al., 2024), which is tailored to instruction-style supervision rather than CPT. Open-PMC-18M additionally shows that curation and alignment can matter more than raw scale (Baghbanzadeh et al., 2025).

PMC-InterCPT is complementary to these efforts. Rather than treating adjacent article text as generic enrichment or converting records into question-answer pairs, we reconstruct figure-reference-supported interleaved sequences and explicitly repair incoherent context before CPT.

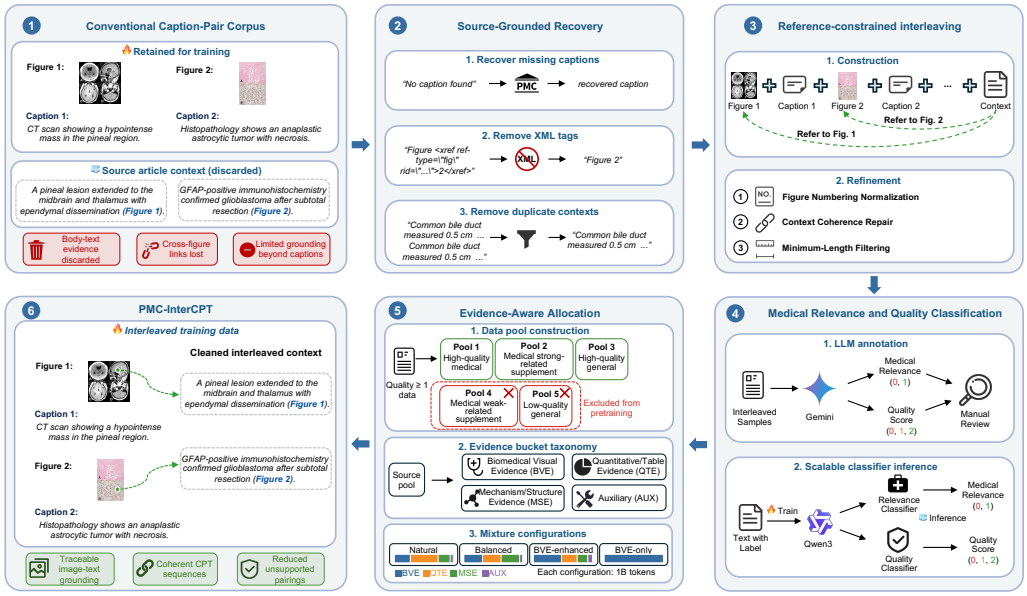


Figure 1: PMC-InterCPT construction. Source-grounded recovery restores captions and cleans article text Reference-constrained interleaving attaches contexts only when supported by article-native figure references. Quality/relevance curation and evidence-aware allocation form the final corpus.

### 2.3 Interleaved Image-Text Pretraining

Flamingo (Alayrac et al., 2022) demonstrated that interleaved image-text data enables strong few-shot multimodal generalization. Subsequent work including MMC4 (Zhu et al., 2023) and OBELICS (Laurençon et al., 2023) scaled interleaved web data to billions of documents, showing consistent benefits over image-caption pair training. These results support interleaved biomedical representations but leave open how article context should be grounded to figures. We address this through reference- and discourse-constrained reconstruction from PMC.

### 2.4 Data Quality and Mixture for Pretraining

FineWeb (Penedo et al., 2024), DCLM (Li et al., 2024b), and RefinedWeb (Penedo et al., 2023) have established that careful quality filtering of web text yields substantial gains for general LLM pretraining. DoReMi (Xie et al., 2023) showed that domain reweighting can significantly accelerate pretraining convergence. We extend these principles to medical multimodal CPT through a corpus construction process that begins with source-grounded reconstruction, filters records with LLM-supervised quality models, and then allocates the retained data with a modality-aware resampling strategy. We provide a systematic empirical study of these stages.

## 3 Method

BIOMEDICA is a large-scale archive of biomedical image-caption pairs extracted from the PMC Open Access subset. Its article-level provenance makes it possible to reconstruct a stronger training unit: a sequence in which each retained image and context segment is justified by an article-native figure reference. As Figure 1 illustrates, conventional caption-pair training retains figures and captions but discards the source article context that explicitly links them. We call the process that recovers this structure *context-grounded reconstruction*. It addresses missing source text, flattened markup and repetition, many-to-many figure-context references, and incoherent concatenation of distant paragraphs. The reconstructed sequences are subsequently curated for quality and medical relevance, then allocated across complementary evidence types to form the final CPT corpus.

### 3.1 SOURCE-GROUNDED RECOVERY

**Caption Recovery.** BIOMEDICA can contain records without a usable caption when the original PMC-OA extraction does not associate an image or table-derived media object with a directly recoverable caption field. For such records, we recover the caption directly from the corresponding PMC XML by mapping the image to its enclosing `<fig>` or `<table-wrap>` element. We retain the source figure identifier and use identifier variants as a fallback. This recovery is preferable to generating a caption because it preserves the publication's original figure-text association.

**Structure-preserving normalization.** We remove residual XML tags from both captions and context fields while retaining their inner text. We then normalize whitespace, remove local near-duplicate fragments, and deduplicate repeated contexts within an article. These operations correct artifacts introduced when XML paragraphs with inline elements are flattened, while retaining the references needed for subsequent reconstruction.

### 3.2 REFERENCE-CONSTRAINED INTERLEAVING

Each reconstructed record alternates image slots with caption and context slots. A context paragraph is attached only to figures that it explicitly references. When a paragraph refers to multiple figures, the corresponding figures are co-included rather than duplicating the paragraph across independent examples. We canonicalize caption prefixes with source figure numbers, use each context only once within an article, and propagate image modality metadata for mixture allocation.

Context segments for a figure can originate from distant article locations. Concatenating them can place unrelated statements next to one another, producing a synthetic context that may not faithfully represent the evidence associated with the figure. We therefore retain adjacent sequences and resolve a discontinuity by favoring a segment exclusively referring to the primary figure; otherwise, we retain the earlier segment. Any non-primary image no longer supported by the retained context is pruned, while the primary image is always preserved. We finally remove examples that do not contain sufficient caption and context for meaningful CPT signal using separate minimum-length criteria for captions and contexts. Exact thresholds are reported in the appendix.

### 3.3 MEDICAL RELEVANCE AND QUALITY CLASSIFICATION

After sample reconstruction, we filter by medical relevance and text quality using a two-stage approach.

**Stage 1: Annotation.** We use Gemini-3.1-pro-preview to assign two labels to sampled interleaved records: medical relevance $\in \{0, 1\}$ and text quality $\in \{0, 1, 2\}$. The annotator receives captions and figure-linked context but not images: structural grounding is enforced upstream through PMC-OA figure references, whereas these labels measure medical relevance and textual usability.

We do not use VLM-based image-text relevance as a hard filter. In preliminary experiments, visual recognition errors and hallucinated interpretations sometimes caused valid biomedical samples to be rejected, particularly for dense compound figures. The quality label penalizes malformed table or LaTeX dumps, repetition, and incoherent text. Sampled labels are reviewed before classifier training.

**Stage 2: Classifier Inference.** We use the reviewed annotations to train separate Qwen3-1.7B classifiers for medical relevance and text quality. The medical classifier predicts $m \in \{0, 1\}$, where $m = 1$ denotes medical relevance, while the quality classifier predicts $q \in \{0, 1, 2\}$. We apply these predicted labels to every reconstructed record and retain $q \geq 1$ records for source-pool construction. A quality score of $q = 0$ denotes severely corrupted text, such as heavy repetition, malformed table or LaTeX dumps, or incoherent passages, and is removed. Scores $q = 1$ and $q = 2$ denote usable-but-imperfect and clean, coherent text, respectively. Annotation prompts, classifier training details, and per-class results are provided in the appendix.

### 3.4 EVIDENCE-AWARE ALLOCATION

**Source Pool.** After decontaminating against the PMC-VQA benchmark, we partition the quality$\geq$1 data into five pools based on quality score, medical relevance, and image modality: (1) *high-quality medical* (quality=2, medical=1); (2) *medical strong-related supplement* (quality=1, medical=1, label $\in$ {Clinical Imaging, Immuno Assays, Laboratory Specimens and Cultures, Microscopy, PCR}); (3) *high-quality general* (quality=2, medical=0); (4) *medical weak-related supplement* (quality=1, medical=1, non-core labels); and (5) *low-quality general* (quality=1, medical=0). Pools 4 and 5 are excluded from pretraining: pool 4 contains lower-quality medical samples from weakly related visual categories, while pool 5 contributes off-domain content. The source pool for all subsequent experiments therefore consists of pools 1–3: high-quality medical, medical strong-related supplement, and high-quality general.

**Evidence Buckets.** Analysis of the natural token distribution within the source pool reveals severe imbalance: Plots and Charts account for 51.2% of tokens, while clinically relevant Clinical Imaging represents only 7.5%. To enable principled resampling, we define four *evidence buckets* that group image modalities by their functional role in biomedical communication:

- **Biomedical Visual Evidence (BVE):** Clinical Imaging, Microscopy, Immuno Assays, PCR, Laboratory Specimens and Cultures
- **Quantitative/Table Evidence (QTE):** Plots and Charts, Tables
- **Mechanism/Structure Evidence (MSE):** Illustrative Diagrams, Chemical Structures, Graphs and Networks, Scientific Formulae and Equations
- **Auxiliary (AUX):** Natural Images, Tools and Materials, Screen Based Visuals, Maps, Ambiguous

**Allocation Strategy.** Evidence-aware allocation controls the relative exposure to complementary forms of scientific evidence in the CPT corpus. All allocations draw from the same filtered source pool and differ only in their bucket-level token weights; within each bucket, samples are drawn in proportion to their natural token count. Natural preserves the source-pool distribution, Balanced approximately equalizes the evidence buckets, BVE-enhanced prioritizes biomedical visual evidence while retaining the other evidence types, and BVE-only restricts the corpus to biomedical visual evidence. The exact bucket proportions for the controlled 1B-token configurations are provided in the mixture-configuration analysis in the Experiments section. Unless stated otherwise, the final 9.63B-token PMC-InterCPT corpus uses the BVE-enhanced allocation (45% BVE, 30% QTE, 20% MSE, and 5% AUX), selected from the controlled study for its strongest medical average while retaining non-BVE evidence.

### 3.5 DATASET SUMMARY

Figure 2 summarizes both the scale changes and the final composition of PMC-InterCPT. Adding figure-linked context expands the raw 13.65B-token caption archive into a 19.89B-token interleaved pool. Length filtering, quality curation, decontamination, source-pool construction, and evidence-aware allocation then yield the final corpus of 10.11M samples and 9.63B tokens. This contraction reflects selective construction of context-grounded training sequences rather than raw-data scaling.

The final corpus remains structurally diverse: 13.2% of samples contain multiple images, preserving cross-figure evidence unavailable to caption-only training. Plots and charts are the largest individual modality (30.4% of samples), followed by microscopy (20.5%) and illustrative diagrams (16.4%). At the allocation, BVE receives the largest share of tokens, while QTE and MSE together account for half of the corpus. The representative image panels in Figure 2(b) illustrate this coverage across the four evidence buckets.

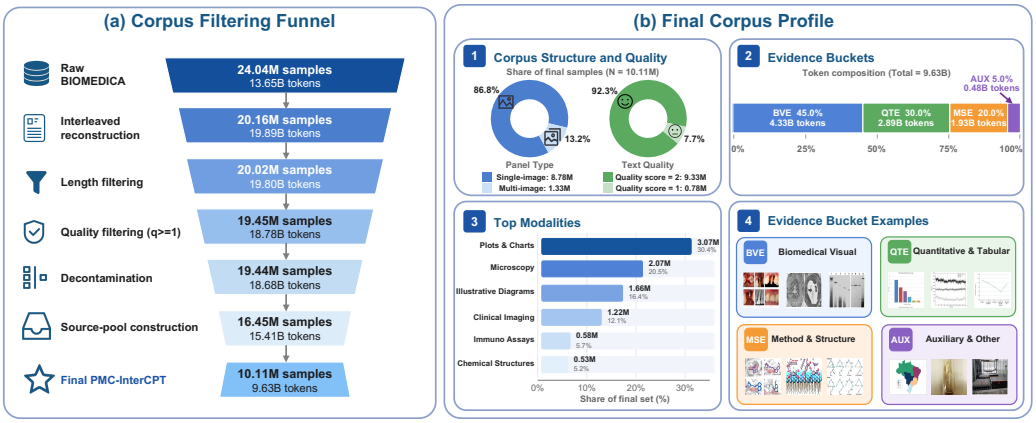


Figure 2: Corpus filtering funnel and final profile of PMC-InterCPT. (a) The funnel tracks sample and token counts from raw BIOMEDICA through interleaved reconstruction, length filtering, quality filtering, decontamination, source-pool construction, and final allocation. (b) The final profile reports panel structure and text quality over 10.11M samples, top image modalities by sample share, and the BVE-enhanced token allocation. Representative images illustrate each evidence bucket.

# 4 Experiments

## 4.1 Experimental Setup

**Model.** Our primary controlled data study uses Qwen3.5-4B-Base (Qwen Team, 2026) before a fixed SFT stage, so differences primarily reflect the CPT data source and mixture. We additionally report Qwen3.5-2B-Base results at the same full scale and a separate architecture-transfer ablation on LLaVA-OneVision-1.5-4B-Base.

**Evaluation.** We evaluate on five medical benchmarks: MMMU-Med-Test (Yue et al., 2024), MMMU-Pro-Med-10 (Yue et al., 2025), PMC-VQA-clean (Zhang et al., 2023b), OmniMed-VQA (Hu et al., 2024), and PretexEval (Zhou et al., 2025). We additionally evaluate on four general or scientific multimodal benchmarks: MMMU-All-Val (Yue et al., 2024), ChartQA (Masry et al., 2022), CharXiv-Val (Wang et al., 2024), and SciVQA (Borisova et al., 2025). The suite is intended to evaluate generative MLLM utility beyond contrastive representation learning, spanning medical visual question answering, medical factual knowledge, and scientific figure understanding. For PMC-VQA-clean, we remove all training samples whose source PMC articles overlap with questions. ChartQA, CharXiv-Val, and SciVQA are scored using Qwen2.5-72B as an LLM judge, and other benchmarks are scored by rule-based extraction.

**Training protocol.** All experiments use a two-stage CPT+SFT protocol. During CPT, we use a maximum sequence length of 8192 with sequence packing, freeze the ViT encoder, and train for one epoch with a warmup ratio of 0.05 to a maximum learning rate of $5.0 \times 10^{-6}$, followed by cosine decay to $1.0 \times 10^{-6}$, micro-batch size 2, and global batch size 256. CPT is run on 64 NVIDIA H800 GPUs. After CPT, all models are supervised fine-tuned on the same 50K examples randomly sampled from LLaVA-OneVision (Li et al., 2024a). During SFT, we train all model parameters for three epochs on 8 NVIDIA H800 GPUs with cosine learning-rate decay from $2.5 \times 10^{-6}$ to 0, a warmup ratio of 0.1, micro-batch size 2, and global batch size 64.

## 4.2 Main Results

Table 1 compares PMC-InterCPT with both full raw BIOMEDICA and a 9.63B-token random BIOMEDICA control under the same fixed SFT stage, across Qwen3.5-4B-Base and Qwen3.5-2B-Base. The token-matched control isolates the effect of reconstruction from CPT token budget.

Table 1: Full-scale CPT+SFT results across Qwen3.5 Base scales. Superscripts give CPT token budgets. Every row uses the same 50K-example LLaVA-OneVision SFT stage. Bold and underlining mark the best and second-best values. ChartQA, CharXiv, and SciVQA use the LLM judge. Other benchmarks use rule-based scoring.

| Backbone | CPT data | Medical benchmarks | | | | | | General/scientific benchmarks | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | MMMU Med | MMMU-Pro Med | PMC VQA | Omni Med | Pretex Eval | Medical avg. | MMMU All | Chart QA | Char Xiv | Sci VQA | General avg. |
| Qwen3.5-4B | None (SFT only) | **58.96** | **41.61** | 61.40 | 85.37 | 64.52 | 62.37 | **55.11** | 75.68 | 53.50 | **68.71** | 63.25 |
| | BIOMEDICA$^{13.65B}$ | 58.62 | 37.76 | **62.90** | 84.97 | 67.48 | 62.35 | 53.89 | 75.00 | 47.90 | 66.74 | 60.88 |
| | BIOMEDICA$^{9.63B}$ | 56.39 | 39.51 | 61.80 | **86.95** | 67.68 | 62.47 | 54.33 | 74.56 | 45.20 | 66.76 | 60.21 |
| | PMC-InterCPT$^{9.63B}$ | 58.73 | 40.56 | 62.75 | 86.51 | **71.10** | **63.93** | 55.00 | **77.00** | **54.00** | 67.29 | **63.32** |
| Qwen3.5-2B | None (SFT only) | 45.78 | **32.17** | 56.45 | 80.35 | 56.15 | 54.18 | 46.44 | **73.40** | 45.70 | 64.79 | 57.58 |
| | BIOMEDICA$^{13.65B}$ | 45.09 | 28.67 | 59.00 | 79.09 | **56.36** | 53.64 | 44.67 | 71.72 | 42.90 | 63.48 | 55.69 |
| | BIOMEDICA$^{9.63B}$ | 44.06 | 29.37 | 58.20 | 79.08 | 53.36 | 52.81 | 45.00 | 71.76 | 42.00 | 61.83 | 55.15 |
| | PMC-InterCPT$^{9.63B}$ | **47.20** | **32.17** | **59.30** | **81.93** | 55.23 | **55.17** | **46.78** | 73.12 | **46.90** | **65.33** | **58.03** |

Table 2: Controlled pipeline ablation. Superscripts give CPT token budgets. Rows introduce interleaving, context reconstruction, and quality filtering; BVE-enh. denotes the BVE-enhanced mixture.

| Format | Clean | QF | Mixture | Medical benchmarks | | | | | | General/scientific benchmarks | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | MMMU Med | MMMU-Pro Med | PMC VQA | Omni Med | Pretex Eval | Medical avg. | MMMU All | Chart QA | Char Xiv | Sci VQA | General avg. |
| Caption$^{1B}$ | – | – | – | 58.39 | **42.66** | 60.90 | 86.04 | 66.50 | 62.90 | 54.78 | 75.88 | 48.20 | 66.78 | 61.41 |
| Interleaved$^{1B}$ | No | No | Natural | 59.30 | 37.41 | 61.40 | 85.73 | 65.24 | 61.82 | 54.11 | 76.56 | 52.00 | **69.00** | 62.92 |
| Interleaved$^{1B}$ | Yes | No | Natural | 59.53 | 39.86 | 61.20 | 86.17 | 64.77 | 62.31 | 55.22 | 76.12 | 52.70 | 68.55 | 63.15 |
| Interleaved$^{1B}$ | Yes | Yes | Natural | **60.10** | 41.26 | 61.10 | **86.38** | 62.92 | 62.35 | 55.33 | **77.08** | **53.30** | 68.60 | **63.58** |
| Interleaved$^{1B}$ | Yes | Yes | BVE-enh. | 59.36 | 41.96 | **61.50** | 86.19 | **67.37** | **63.28** | **55.44** | 76.64 | 51.60 | 67.45 | 62.78 |

Against the token-matched raw control, PMC-InterCPT improves the medical average by 1.46 points and the general/scientific average by 3.11 points on Qwen3.5-4B-Base. The gains span distinct evaluation formats rather than a single benchmark protocol. The strongest matched-control gains are on MMMU-Med (+2.34), PretexEval (+3.42), ChartQA (+2.44), and CharXiv (+8.80). The full raw BIOMEDICA run does not close this gap: despite using 42% more CPT tokens than PMC-InterCPT, it remains 1.58 points lower on the medical average and 2.44 points lower on the general/scientific average. Relative to Base+SFT, PMC-InterCPT raises the medical average by 1.56 points while maintaining general/scientific performance.

The same data comparison transfers to Qwen3.5-2B-Base. PMC-InterCPT improves the medical average by 2.35 points and the general/scientific average by 2.88 points over its token-matched raw control, and by 0.99 and 0.45 points over SFT-only. Together, the two scale settings indicate that source-grounded curation contributes beyond simply increasing the volume of in-domain raw data.

### 4.3 Curation and Allocation Ablation

Table 2 reports benchmark-level results, while Figure 3 visualizes their aggregate trajectory and the resulting mixture positions. The controlled comparison holds the backbone, 1B-token CPT budget, SFT data, and training hyperparameters fixed while adding one pipeline stage at a time. This cumulative design distinguishes raw context addition from the subsequent repair, filtering, and evidence-allocation choices.

Naively interleaving article context improves the general/scientific average by 1.51 points but decreases the medical average by 1.08 points relative to raw caption pairs. Thus, article text alone is not a useful substitute for figure-grounded context. Context reconstruction recovers 0.49 points on the medical average and 0.23 points on the general/scientific average. The largest recovery is on MMMU-Pro Med (+2.45), with additional gains on OmniMedVQA (+0.44) and CharXiv (+0.70). Quality filtering then adds 0.05 points on the medical average and 0.43 points on the general/scientific average.

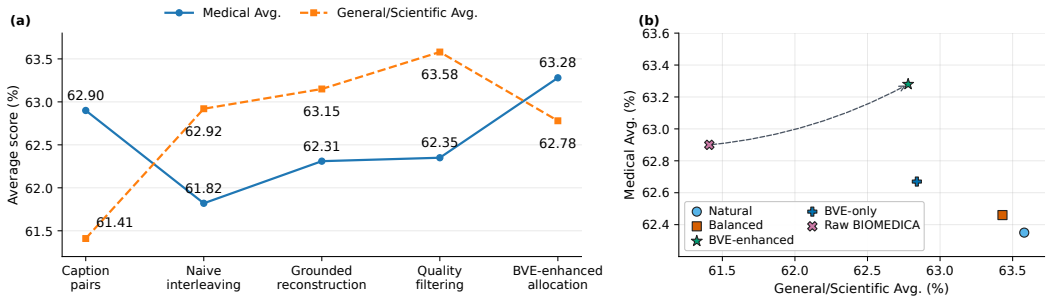


Figure 3: Component and mixture analysis under the fixed 1B-token CPT setting. (a) Cumulative effects of naive interleaving, context-grounded reconstruction, quality filtering, and BVE-enhanced allocation on medical and general/scientific averages. (b) Medical–general positions of the clean, quality-filtered mixture configurations; the dashed arrow connects the raw caption-pair control to BVE-enhanced PMC-InterCPT.

Table 3: Mixture configurations. Superscripts give CPT token budgets. All mixture rows use the same clean, quality-filtered PMC-InterCPT source pool and fixed LLaVA-OneVision SFT stage. Raw BIOMEDICA is a caption-pair control.

| Backbone | CPT data | Medical benchmarks | | | | | | General/scientific benchmarks | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | MMMU Med | MMMU-Pro Med | PMC VQA | Omni Med | Pretex Eval | Medical avg. | MMMU All | Chart QA | Char Xiv | Sci VQA | General avg. |
| Qwen3.5-4B-Base | Raw BIOMEDICA$^{1B}$ | 58.39 | **42.66** | 60.90 | 86.04 | 66.50 | 62.90 | 54.78 | 75.88 | 48.20 | 66.78 | 61.41 |
| | Natural$^{1B}$ | **60.10** | 41.26 | 61.10 | 86.38 | 62.92 | 62.35 | 55.33 | 77.08 | **53.30** | 68.60 | **63.58** |
| | Balanced$^{1B}$ | 59.42 | 39.86 | **62.80** | 85.56 | 64.65 | 62.46 | **55.44** | 76.92 | 52.60 | **68.74** | 63.43 |
| | BVE-enhanced$^{1B}$ | 59.36 | 41.96 | 61.50 | 86.19 | **67.37** | **63.28** | **55.44** | 76.64 | 51.60 | 67.45 | 62.78 |
| | BVE-only$^{1B}$ | 59.82 | 39.51 | 61.75 | **86.51** | 65.76 | 62.67 | 54.78 | **77.20** | 51.50 | 67.88 | 62.84 |

Compared with the Natural distribution, the BVE-enhanced distribution changes the balance by +0.92 points on the medical average and -0.80 points on the general/scientific average. We present evidence-aware allocation as a medical-prioritized tradeoff, not a universally dominant optimum.

**One-Billion-Token Mixture Configurations.** Figure 3(b) and Table 3 compare the set of 1B-token mixture configurations. Natural preserves the source-pool distribution (28% BVE, 52% QTE, 17% MSE, and 3% AUX). Balanced approximately equalizes the evidence buckets (33%, 33%, 29%, and 5%, respectively). BVE-enhanced increases BVE while retaining the other buckets (45% BVE, 30% QTE, 20% MSE, and 5% AUX), whereas BVE-only contains 100% BVE.

Balanced allocation provides only a 0.11-point medical-average increase over Natural, whereas BVE enhancement improves it by 0.92 points. BVE-only remains 0.61 points below BVE-enhanced, indicating that the benefit does not come from removing quantitative, mechanistic, and auxiliary evidence. The general results expose the corresponding tradeoff: Natural obtains the highest general/scientific average (63.58), while BVE-enhanced is 0.80 points lower (62.78). The per-benchmark optima also differ: Natural is strongest on MMMU-Med and CharXiv, Balanced on PMC-VQA-clean and SciVQA, BVE-only on OmniMedVQA and ChartQA, and BVE-enhanced on PretexEval and the overall medical average. These results suggest that BVE-enhanced sampling is a medical-oriented trade-off rather than a universally superior mixture.

### 4.4 Validation of Figure-Context Grounding

To test whether the benefit comes from figure-grounded context rather than additional in-domain text, Figure 4 reports a reference-breaking counterfactual from the BVE-enhanced 1B mixture. For each replaceable context segment, we retain the image, caption, article, and tokenizer-level length budget, but replace the segment with a randomly sampled text span from the same article with a matching token count. Thus, the control preserves article domain and context length while breaking the original figure-reference attachment.

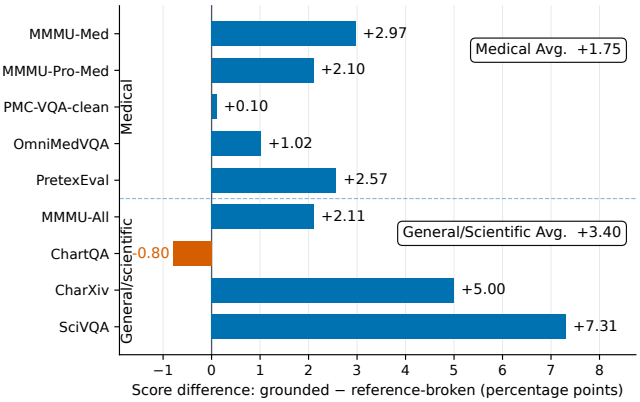


Figure 4: Effect of figure-context grounding in the BVE-enhanced 1B mixture. Bars show the score difference between PMC-InterCPT and a reference-broken control that uses token-matched text; positive bars indicate the loss caused by breaking the figure-context attachment.

Table 4: Cross-architecture ablation on LLaVA-OneVision-1.5-4B-Base. Superscripts give CPT token budgets; all CPT rows use the same SFT stage.

| Backbone | CPT data | Medical benchmarks | | | | | | General/scientific benchmarks | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | MMMU Med | MMMU-Pro Med | PMC VQA | Omni Med | Pretex Eval | Medical avg. | MMMU All | Chart QA | Char Xiv | Sci VQA | General avg. |
| LLaVA-OneVision 1.5-4B-Base | None (SFT only) | 45.95 | 25.87 | 48.45 | 62.44 | 56.62 | 47.87 | 47.11 | **59.28** | **35.70** | **53.95** | **49.01** |
| | BIOMEDICA$^{1B}$ | **48.69** | 24.13 | 50.25 | 64.21 | 55.75 | 48.61 | 48.11 | 58.24 | 34.90 | 46.07 | 46.83 |
| | PMC-InterCPT$^{1B}$ | 48.29 | **27.62** | **51.85** | **65.93** | **57.49** | **50.24** | **48.44** | 57.72 | 34.00 | 52.64 | 48.20 |

Breaking figure-context links decreases the medical average by 1.75 points and the general/scientific average by 3.40 points. The loss spans both benchmark groups: MMMU-Med decreases by 2.97 points, PretexEval by 2.57, CharXiv by 5.00, and SciVQA by 7.31. Since the two conditions retain the same images, captions, source articles, token budget, and CPT/SFT configuration, this comparison isolates the contribution of the original figure-reference attachment rather than text quantity or visual content alone. It shows that article context becomes useful CPT supervision only when it remains aligned to the figure it is meant to explain. The larger general/scientific drop highlights the value of figure-linked context beyond captions. Full results are in the appendix.

### 4.5 Cross-Architecture Ablation

Table 4 tests whether the corpus-construction benefit extends to LLaVA-OneVision-1.5-4B-Base under a fixed 1B-token CPT budget. Relative to the token-matched raw BIOMEDICA control, PMC-InterCPT improves the medical average by 1.63 points and the general/scientific average by 1.37 points. The improvement covers four of the five medical benchmarks and is especially pronounced on SciVQA (+6.57), suggesting that the benefit is not confined to the Qwen architecture or to biomedical-only evaluation.

## 5 Conclusion

We introduced PMC-InterCPT, a sequential framework for constructing biomedical interleaved CPT data. It uses source figure references to reconstruct and repair coherent sequences, filters the reconstructed records for textual usability and medical relevance, and then allocates the retained data through a controllable evidence mixture. Its 9.63B-token instantiation improves over the raw source pool on medical and general multimodal performance under a fixed CPT+SFT protocol. The component ablation further shows that article context becomes useful only after referent-aware reconstruction and curation. We hope this framework supports more reliable use of scientific literature.

### AI USE STATEMENT

Generative AI tools were used for brainstorming, information gathering, code development assistance, improving manuscript readability, identifying relevant literature, and formatting references. A generative model was also used as the annotation model described in the medical-relevance and text-quality classification stage. The authors reviewed and verified all AI-assisted code and annotations, and take responsibility for the final content and claims of this work.

### ETHICS STATEMENT

PMC-InterCPT is derived from scientific articles in PMC-OA and does not involve new data collection from human participants. Corpus construction preserves source-level license metadata, and any public redistribution should be restricted to records whose licenses permit it. Biomedical publications may contain sensitive clinical descriptions and demographic or publication biases; PMC-InterCPT is intended for research, not unvalidated clinical use.

### REPRODUCIBILITY STATEMENT

The main paper specifies the corpus-construction stages, controlled data comparisons, training protocol, and evaluation suite. The appendix provides detailed dataset statistics, reconstruction and filtering rules, annotation prompts, classifier settings, benchmark protocols, and complete auxiliary results. We also provide an anonymized code and data supplement containing the processing pipeline and representative reconstructed samples. If accepted, we will publicly release PMC-InterCPT in accordance with source-license and redistribution requirements.

## REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

Negin Baghbanzadeh, Mohammed Saidul Islam, Sajad Ashkezari, Elham Dolatabadi, and Arash Afkanpour. Open-pmc-18m: A high-fidelity large scale medical dataset for multimodal representation learning. *arXiv preprint arXiv:2506.02738*, 2025.

Shruthi Bannur, Kenza Bouzid, Daniel C Castro, Anton Schwaighofer, Anja Thieme, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Salvatelli, Harshita Sharma, et al. MAIRA-2: Grounded radiology report generation. *arXiv preprint arXiv:2406.04449*, 2024.

Ekaterina Borisova, Nikolas Rauscher, and Georg Rehm. Scivqa 2025: Overview of the first scientific visual question answering shared task. In *Proceedings of the Fifth Workshop on Scholarly Document Processing (SDP 2025)*, pp. 182–210, 2025.

Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Ruifei Zhang, Zhenyang Cai, Ke Ji, Guangjun Yu, Xiang Wan, and Benyou Wang. HuatuoGPT-Vision, towards injecting medical visual knowledge into multimodal llms at scale. *arXiv preprint arXiv:2406.19280*, 2024.

Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. OmniMed-VQA: A new large-scale comprehensive evaluation benchmark for medical LVLM. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22170–22183, 2024.

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, et al. OBELICS: An open web-scale filtered dataset of interleaved image-text documents. *Advances in Neural Information Processing Systems*, 36:71683–71702, 2023.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024a.

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.

Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Gadre, Hritik Bansal, Etash Guha, Sedrick Keh, Kushal Arora, et al. DataComp-LM: In search of the next generation of training sets for language models. *Advances in Neural Information Processing Systems*, 37: 14200–14282, 2024b.

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pretraining for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26689–26699, 2024.

Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-clip: Contrastive language-image pre-training using biomedical documents. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 525–536. Springer, 2023.

Alejandro Lozano, Min Woo Sun, James Burgess, Liangyu Chen, Jeffrey J Nirschl, Jeffrey Gu, Ivan Lopez, Josiah Aklilu, Anita Rau, Austin Wolfgang Katzer, et al. Biomedica: An open biomedical image-caption archive, dataset, and vision-language models derived from scientific literature. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 19724–19735, 2025.

Ahmed Masry, Do Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2263–2279, 2022.

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)*, pp. 353–367. PMLR, 2023.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The RefinedWeb dataset for falcon LLM: Outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*, 2023.

Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, Thomas Wolf, et al. The FineWeb datasets: Decanting the web for the finest text data at scale. *Advances in Neural Information Processing Systems*, 37:30811–30849, 2024.

Qwen Team. Qwen3.5: Towards native multimodal agents. Alibaba Cloud Community Blog, 2026. URL https://www.alibabacloud.com/blog/qwen3-5-towards-native-multimodal-agents_602894.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, et al. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. *Advances in Neural Information Processing Systems*, 37:113569–113697, 2024.

Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy S Liang, Quoc V Le, Tengyu Ma, and Adams Wei Yu. DoReMi: Optimizing data mixtures speeds up language model pretraining. *Advances in Neural Information Processing Systems*, 36:69798–69818, 2023.

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.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9556–9567, 2024.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15134–15186, 2025.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023a.

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*, 2023b.

Lianmin Zheng, Liangsheng Yin, Zhiqiang Xie, Chuyue Sun, Jeff Huang, Cody H Yu, Shiyi Cao, Christos Kozyrakis, Ion Stoica, Joseph E Gonzalez, et al. Sglang: Efficient execution of structured language model programs. *Advances in neural information processing systems*, 37:62557–62583, 2024a.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 3: system demonstrations)*, pp. 400–410, 2024b.

Yuxuan Zhou, Xien Liu, Chen Ning, Xiao Zhang, and Ji Wu. Reliable and diverse evaluation of llm medical knowledge mastery. In *International Conference on Learning Representations*, volume 2025, pp. 57899–57922, 2025.

Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. Multimodal C4: An open, billion-scale corpus of images interleaved with text. *Advances in Neural Information Processing Systems*, 36:8958–8974, 2023.

# A Additional Dataset Statistics

## A.1 Raw BIOMEDICA Statistics

BIOMEDICA is built from the PubMed Central Open Access subset and provides an article-level archive serialized into image-caption training records (Lozano et al., 2025). The released corpus contains over 6M full-text articles, over 24M image-caption pairs, and over 30M figure references. Each record includes article-level metadata and image-level annotations, including article identifiers, title, abstract, journal, license, keywords, MeSH terms, citation metadata, image identifiers, image hashes, panel labels, and global/local taxonomy labels. Table 5 summarizes the raw BIOMEDICA properties that are most relevant to our pipeline.

Table 5: Raw BIOMEDICA statistics reported by Lozano et al. (2025).

| Property | Value |
|---|---|
| Full-text articles | 6M+ |
| Image-caption pairs | 24M+ |
| Figure references | 30M+ |
| Metadata fields | 27 |
| Primary format in BIOMEDICA | Image-caption pairs |
| Additional signal in our pipeline | Figure-referencing text |

## A.2 Cleaned Dataset Statistics

Table 6 reports the missing-caption recovery statistics. We processed 2,537 tar archives containing 72,151,269 files, including 24,050,423 JSON records. Among these JSON records, 1,987,649 were modified during missing-caption recovery, corresponding to an 8.26% modification rate. For records whose caption field required recovery, 6,238,761 captions were successfully recovered and 3,639,453 were not recovered, giving a success rate of 63.2%. Because one JSON record can contain multiple image slots, the number of recovered captions can exceed the number of modified records. No XML files were missing for the affected articles. Most recovered captions correspond to table-like figure entries whose captions are available in the PMC XML. Unrecovered cases are largely entries such as graphical abstracts, formulae, or other media objects for which the source XML does not provide a caption that can be extracted.

Table 6: Missing-caption recovery statistics. The success rate is computed as recovered captions divided by recovered plus unrecovered captions.

| Statistic | Value |
|---|---|
| Processed tar archives | 2,537 |
| Processed files | 72,151,269 |
| JSON records | 24,050,423 |
| JSON records modified | 1,987,649 |
| Modification rate | 8.26% |
| Captions recovered | 6,238,761 |
| Captions unrecovered | 3,639,453 |
| Recovery success rate | 63.2% |
| Missing XML files | 0 |

Repeated-text cleanup modified 22,844,977 of 24,050,423 JSON records (94.99%). This high modification rate reflects the prevalence of repeated short fragments and duplicated local context spans introduced when PMC XML paragraphs with inline figure references are flattened into plain text.

Table 7 summarizes the major scale changes across the later pipeline stages. Interleaved reconstruction reduces the number of training records relative to raw BIOMEDICA by reorganizing article-level figures into context-grounded samples, but increases the token count by adding body-text context. Quality filtering, decontamination, source-pool selection, and final mixture sampling then progressively concentrate the corpus into the final PMC-InterCPT training set.

Table 7: Dataset scale across the main corpus construction stages.

| Stage | Samples | Tokens |
|---|---|---|
| Raw BIOMEDICA | 24.04M | 13.65B |
| After interleaved reconstruction | 20.16M | 19.89B |
| After minimum-length filtering | 20.02M | 19.80B |
| Quality score $\geq 1$ | 19.45M | 18.78B |
| After decontamination | 19.44M | 18.68B |
| Source-pool construction | 16.45M | 15.41B |
| PMC-InterCPT final mixture | 10.11M | 9.63B |

Table 8: Evidence-bucket distribution in the decontaminated source pool before final mixture sampling.

| Evidence bucket | Samples | Text tokens | Image tokens | Total tokens | Share |
|---|---|---|---|---|---|
| Biomedical Visual Evidence (BVE) | 4,036,696 | 2.07B | 2.27B | 4.33B | 28.13% |
| Quantitative/Table Evidence (QTE) | 8,681,434 | 3.72B | 4.30B | 8.02B | 52.03% |
| Mechanism/Structure Evidence (MSE) | 3,117,352 | 1.18B | 1.37B | 2.55B | 16.54% |
| Auxiliary (AUX) | 621,590 | 0.21B | 0.30B | 0.51B | 3.31% |
| Total | 16,457,072 | 7.18B | 8.23B | 15.41B | 100.00% |

### A.3 Evidence Bucket Distribution

Table 8 reports the evidence-bucket distribution after decontamination and source-pool construction, using the decontaminated global cross statistics. The source pool consists of high-quality medical samples, strong medical supplements from core biomedical labels, and high-quality general samples. Although QTE remains the largest bucket under the natural source-pool distribution, the BVE-enhanced final mixture deliberately increases BVE to 45% of tokens while retaining 30% QTE, 20% MSE, and 5% AUX.

Table 9 reports the final PMC-InterCPT bucket composition after modality-aware resampling. Compared with the natural source-pool distribution, the final mixture increases BVE from 28.13% to 45.0% while reducing QTE from 52.03% to 30.0%. This produces a medical-visual-evidence-enhanced corpus while retaining substantial QTE and MSE coverage.

Table 9: Final evidence-bucket composition of PMC-InterCPT after modality-aware resampling.

| Evidence bucket | Tokens | Final share |
|---|---|---|
| BVE | 4.33B | 45.0% |
| QTE | 2.89B | 30.0% |
| MSE | 1.93B | 20.0% |
| AUX | 0.48B | 5.0% |
| Total | 9.63B | 100.0% |

Table 10 gives the corresponding five-pool decomposition. Pools 1–3 are included in the source pool used for all mixture experiments, while pools 4–5 are excluded from pretraining.

## B Sample Reconstruction and Filtering Details

### B.1 Interleaved Sample Format

Each reconstructed interleaved sample is serialized as a Parquet row with three top-level fields: `images`, `texts`, and `metadata`. The `images` and `texts` fields are aligned arrays of equal length. An image slot stores a base64-encoded image in `images` and `null` in the corresponding `texts` position. A text slot stores `null` in `images` and the text string in `texts`. For a single-

Table 10: Five quality/medical pools after decontamination. HQ denotes high-quality, and supp. denotes supplement. Pools 1–3 form the source pool, and pools 4–5 are excluded from pretraining.

| Pool | Samples | Tokens |
|---|---|---|
| Pool 1: HQ medical | 12,389,828 | 11.67B |
| Pool 2: Med-strong supp. | 782,139 | 1.02B |
| Pool 3: HQ general | 3,285,105 | 2.73B |
| Pool 4: Med-weak supp. | 2,278,176 | 2.61B |
| Pool 5: Low-quality general | 704,879 | 0.67B |

image sample, the typical sequence is:

$$[\mathrm{IMG}_1] \rightarrow [\text{Figure } n. \text{ caption}] \rightarrow [\text{context paragraph}].$$

For multi-image samples, additional referenced figures are inserted with their captions before the shared context, so that the context can refer to all figures present in the sequence.

The `metadata` field is a JSON string that stores article-level provenance and bibliographic information together with aligned per-image attributes, including file names, identifiers, paths, captions, labels, panel annotations, cluster assignments, and image sizes. Captions receive explicit figure-number prefixes, e.g., "Figure 2.", to disambiguate subsequent context references.

### B.2 Context Coherence Repair and Length Filtering

Context coherence repair is applied after initial interleaved sample reconstruction. For samples with multiple context segments, we locate each normalized context string in the full article text and check whether consecutive segments are adjacent in the original article. Two segments are treated as adjacent when their spans are contiguous after whitespace normalization. The implementation also uses an edge-anchor fallback that searches for the concatenation of the previous segment's ending words and the next segment's beginning words in the article text.

When two consecutive contexts are not adjacent, we avoid placing unrelated statements next to one another in a synthetic context that may not faithfully represent the evidence associated with the figure. The repair rule compares the two segments using raw figure-reference information from the source `image_context`. If only one segment refers exclusively to the primary image, that segment is kept. If both segments are primary-image-only, or neither is primary-image-only, the earlier segment is kept. After context pruning, images and captions that are no longer referenced by the kept contexts are removed, while the primary image is always retained. The metadata arrays are subsetted consistently with the retained images.

Length filtering is applied after context cleanup. For each row, the first caption and all non-caption context segments are counted separately. The script removes tail figure/table-reference fragments from contexts, normalizes repeated whitespace, collapses duplicated periods, removes empty parentheses, and removes extra spaces before punctuation. Language is determined by detecting CJK characters in the combined caption and context. For Chinese samples, a row is dropped when both `caption_chars < 40` and `context_chars < 120`. For non-Chinese samples, a row is dropped when both `caption_tokens < 12` and `context_tokens < 30`.

## C Medical Relevance and Quality Classification

### C.1 Annotation Guidelines

The LLM annotation step uses text-only inputs: captions and figure-linked context are shown to the annotator, while images are withheld. This design makes the two labels conservative measures of textual usefulness for CPT rather than image-text completeness. One practical reason for this choice is that image-conditioned quality annotation was unstable in our preliminary trials: when images were included as part of the quality criterion, the annotator model sometimes over-interpreted subtle visual details or made incorrect judgments about image content, which introduced hallucination-like

errors into labels intended to measure corpus quality. In addition, the raw data are sourced from PMC papers, where the figure, caption, and in-text figure references are linked by the original publication structure. We therefore assume a baseline level of image-text association from the source corpus and restrict annotation to the two aspects that still require scalable filtering: medical relevance and text purity. The annotation contains two targets.

**Medical relevance.** `medical_relevance` is binary. A score of 1 indicates that the sample is medically, clinically, or biomedically relevant, including biomedical research, clinical imaging, pathology, laboratory assays, molecular biology, public health, pharmacology, and other life-science evidence. A score of 0 indicates that the text is not medical or biomedical, or is only incidentally related to science without useful medical-domain content.

**Text quality.** `quality` is a three-level text-purity score. The score does not measure whether the caption fully describes the image, because the annotator does not receive the image. Instead, it measures whether the text is clean enough to serve as CPT data:

- **0: severe quality problems.** The sample has heavy repetition, copy-paste artifacts, severe garbling, major incoherence, mixed unrelated passages, unreadable table dumps, malformed LaTeX source fragments, or text dominated by extraction noise.
- **1: usable but imperfect.** The sample remains mostly understandable, but contains noticeable redundancy, corruption, malformed joins, table-formatting damage, citation clutter, or limited extraction artifacts.
- **2: clean and coherent.** The sample is low-noise, low-redundancy, readable medical or biomedical text with coherent captions and context.

The guidelines deliberately penalize flattened tables and malformed formula extraction. Short, semantically meaningful equations or chemical notation are allowed, but raw source fragments such as `\documentclass`, `\usepackage`, long pasted TeX blocks, duplicated rendered/source equations, and broken table rows lower the quality score. If a table or formula artifact is present but limited, the maximum score is usually 1; if it dominates the sample or makes a large section difficult to read, the score is 0.

### C.2 LLM Annotation Prompt

We use Gemini-3.1-pro-preview with temperature 0 to label sampled interleaved records. The user message concatenates the scoring instruction with a text-only rendering of the sample. The sample rendering separates captions from additional context so that the annotator can distinguish figure captions from body-text evidence. The first block below shows the complete scoring instruction used in the annotation request.

**System-style instruction used in the annotation request**

You are grading one interleaved text-only sample for medical multimodal pretraining data filtering.
**Important:**

- You will NOT receive images.
- Judge ONLY from the provided text.
- Do NOT assess completeness because images are unavailable.
- The final quality score is just a text purity score.

**You must score TWO dimensions:**

1. `medical_relevance`: integer in {0,1}
2. `quality`: integer in {0,1,2}

**Definitions:**
**1) Medical Relevance:**

- 0: not medical / not biomedical / not clinically relevant.
- 1: medically or biomedically relevant.

**2) Quality (text purity only):**

- 0: severe text quality problems. Examples: heavy repetition, obvious copy-paste artifacts, severe garbling, major incoherence, mixed unrelated content, unreadable table dumps, or text dominated by noise.
- 1: usable but imperfect. Some redundancy, corruption, noticeable noise, or table-related formatting damage remains.
- 2: clean, coherent, low-noise, low-redundancy medical text.

**Very strict penalty rules for tables and readability:**

- If the text contains a flattened table, embedded raw table dump, broken row/column listing, mashed cell values, or missing-space table extraction that noticeably hurts readability, quality must be lowered.
- If table artifacts are present but limited and the sample is still mostly readable prose, use quality=1, not 2.
- If table artifacts dominate a substantial portion of the sample, or make a large section hard to read, use quality=0.
- If the text switches from normal prose into malformed table content, this should usually be treated as at least quality=1 and often quality=0 if the malformed part is long or noisy.
- A clean sentence that merely references a table is not a problem. Only penalize when table content itself appears in the text and damages readability.

**Very strict penalty rules for LaTeX / formula artifacts:**

- Clean, short, semantically normal mathematical expressions are acceptable and should NOT be penalized by themselves. Examples of acceptable content include inline or display math such as `\(a^2 - 4b = m^2\)`, chemical notation, or brief equations that naturally belong in scientific text.
- Penalize strongly when the text contains malformed, dumped, or extraction-corrupted LaTeX/TeX content that hurts readability. Examples include raw preamble or source fragments such as `\documentclass`, `\usepackage`, `\begin{document}`, `\end{document}`, duplicated rendered formula text, OCR-broken math markup, or long formula-source blocks pasted into prose.
- If a malformed LaTeX block appears briefly but the rest of the sample is still understandable, use quality=1.
- If malformed LaTeX or formula-source noise occupies a substantial part of the sample, interrupts sentence flow, or makes the sample feel corrupted or unnatural, use quality=0.
- When both a human-readable equation and a dumped LaTeX source block appear together, treat this as a formatting artifact and lower quality.

**Other strict penalty rules for quality:**

- Repeated short sentences, repeated clauses, repeated local phrases, or repeated caption fragments should lower quality.
- Context that mostly repeats the caption rather than adding useful text should lower quality.
- Fragmented text, malformed joins, truncation artifacts, broken sentence boundaries, or mixed unrelated passages should lower quality.
- Citation-heavy but still readable text can be acceptable; only penalize if it materially hurts readability or purity.
- Be conservative: if unsure whether table artifacts materially reduce readability, choose the lower quality score.

**Output requirements:**

- Return ONLY one JSON object.
- No markdown, no extra keys.

**Return this exact schema:**

```
{
"medical_relevance": 0|1,
"quality": 0|1|2,
"reasons": {
"medical_relevance": "one short sentence",
"quality": "one short sentence"
}
}
```

The second block shows the text-only sample template concatenated with the scoring instruction.

**Text-only sample format**

```
Sample texts for evaluation:
Captions:
- <caption text for each image>
Additional context texts:
- <figure-linked body-text context>
```

### C.3 Classifier Training Details

We train separate Qwen3-1.7B classifiers with LLaMA-Factory (Zheng et al., 2024b): a three-way text-quality classifier and a binary medical-relevance classifier. Both use full-parameter fine-tuning with a 2,048-token cutoff, per-device batch size 4, no gradient accumulation, learning rate $1 \times 10^{-5}$, six epochs, cosine learning-rate decay, warmup ratio 0.1, and bf16 precision. For each task, we retain the checkpoint with the strongest validation-set performance.

Table 11: Balanced split for the quality classifier. Labels 0, 1, and 2 correspond to severe-noise, usable-but-imperfect, and clean text quality.

| Split | Label 0 | Label 1 | Label 2 | Total |
|---|---|---|---|---|
| Train | 408 | 408 | 408 | 1,224 |
| Validation | 51 | 51 | 51 | 153 |
| Test | 52 | 52 | 52 | 156 |

Table 12: Balanced split for the medical-relevance classifier. In the classifier training files, label 0 denotes medical and label 1 denotes non-medical. Outputs are converted to the paper's `medical_relevance` convention during post-processing.

| Split | Label 0 | Label 1 | Total |
|---|---|---|---|
| Train | 1,134 | 1,134 | 2,268 |
| Validation | 141 | 141 | 282 |
| Test | 143 | 143 | 286 |

Large-scale classifier inference is run with SGLang (Zheng et al., 2024a). Both classifiers use the same serving configuration: data parallelism 8, tensor parallel size 1, batch size 512, maximum input length 4,096 characters, maximum new tokens 4, and temperature 0. The constrained output space is therefore a single label token for each record.

### C.4 Classifier Test-Set Performance Analysis

Table 13 and Table 14 report held-out test performance for the two classifiers. The quality classifier reaches 92.3% accuracy and 92.3% macro F1. Performance is strongest on label 0 and label 2, with F1 scores of 95.3% and 93.2%, respectively. The lower F1 for label 1 (88.2%) is expected because this middle class is a boundary category: limited table damage, mild repetition, or moderate formatting noise can be close to either severe-noise label 0 or clean label 2. This error pattern is acceptable for filtering because the main data pool keeps quality$\geq$1 records. The most important operational distinction is separating severely corrupted samples from usable ones.

The medical-relevance classifier reaches 93.4% accuracy and 93.7% macro F1. The class-level F1 scores are balanced, with 94.2% for medical samples and 93.1% for non-medical samples. This indicates that the classifier is not merely biased toward the dominant biomedical source domain, and can still identify off-domain or weakly related records introduced by broad PMC extraction.

Table 13: Quality classifier performance on the held-out test set.

| Metric | Value |
|---|---|
| Accuracy | 92.3% |
| Macro F1 | 92.3% |
| Label 0 F1 | 95.3% |
| Label 1 F1 | 88.2% |
| Label 2 F1 | 93.2% |

Table 14: Medical-relevance classifier performance on the held-out test set.

| Metric | Value |
|---|---|
| Accuracy | 93.4% |
| Macro F1 | 93.7% |
| Medical label F1 | 94.2% |
| Non-medical label F1 | 93.1% |

# D Qualitative Case Study

Figure 5 illustrates the transformation of one PMC-OA article-derived record. The raw image-context field contains residual `xref` markup and repeated text while referring jointly to an MRI figure and a pathology figure. Source-grounded recovery and reference-constrained interleaving retain the two figures with canonicalized captions and one cleaned shared context. Thus, the final sequence preserves the original article's cross-figure evidence while removing extraction artifacts, rather than generating new clinical content or attaching arbitrary nearby text.

# E Evaluation Details

## E.1 Benchmark Descriptions

MMMU (Yue et al., 2024) is a multi-discipline multimodal reasoning benchmark covering college-level expert subjects. We report MMMU-Med-Test using the Health and Medicine domain, and MMMU-All-Val using the full validation split across disciplines. MMMU-Pro (Yue et al., 2025) is a harder and more robust variant of MMMU with expanded option sets and reduced shortcut cues. We evaluate its medical subset with the 10-option configuration.

PMC-VQA (Zhang et al., 2023b) is a medical visual question answering benchmark built from PubMed Central figures. We evaluate a cleaned test split, PMC-VQA-clean, and remove CPT samples whose source PMC articles overlap with this benchmark before training. OmniMedVQA (Hu et al., 2024) is a broad medical VQA benchmark spanning multiple modalities and question categories, including clinical imaging, pathology, charts, and other biomedical visual evidence. Pretex-Eval (Zhou et al., 2025) evaluates medical knowledge mastery through predicate-equivalence transformations: it tests whether a model gives consistent answers to medically equivalent reformulations of a question.

ChartQA (Masry et al., 2022) evaluates chart understanding through human-written and machine-augmented questions over chart images. CharXiv (Wang et al., 2024) evaluates reasoning over figures from scientific papers, with our experiments using the validation split's reasoning questions. SciVQA (Borisova et al., 2025) evaluates scientific figure question answering with a mixture of multiple-choice and open-ended questions. Together, these general/scientific benchmarks test whether medical CPT improves biomedical reasoning without collapsing chart, scientific figure, and broad multimodal competence.

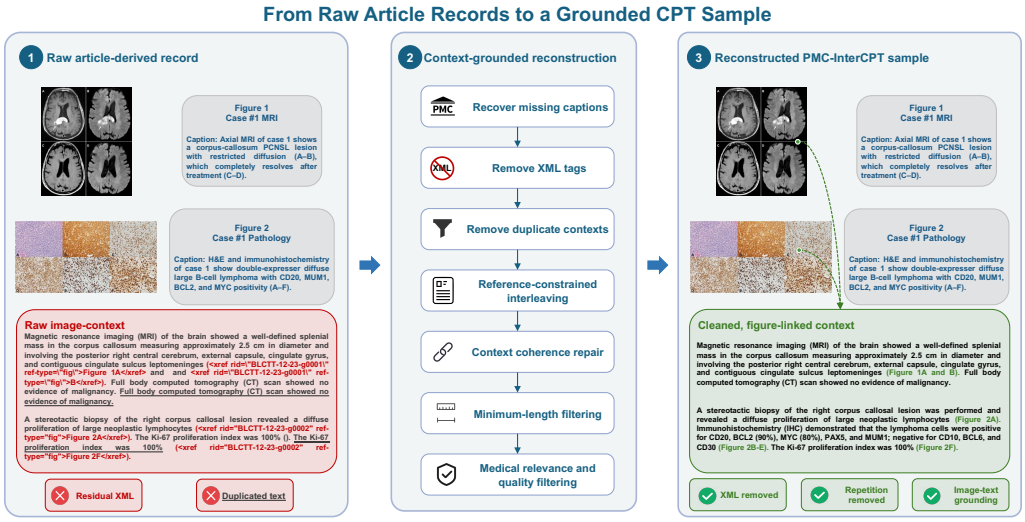


Figure 5: Illustrative reconstruction of a PMC-OA article record. The raw record contains residual `xref` markup and duplicated text in its image-context field. The reconstruction process removes these artifacts and preserves the jointly referenced MRI and pathology figures with normalized captions and a shared, figure-linked context.

### E.2 Evaluation Protocol

All benchmarks are evaluated with zero-shot prompting unless otherwise stated. For multiple-choice questions, the prompt lists the question followed by options and instructs the model to answer directly with the option letter:

**Multiple-choice evaluation prompt**

```
Question: <question>
Options:
A. <option A>
B. <option B>
...
Answer with the option's letter from the given choices
directly.
```

For open-ended questions, the prompt is:

**Open-ended evaluation prompt**

```
<question>
Answer the question using a single word or phrase.
```

Inference uses deterministic decoding with temperature 0. The Medevalkit default configuration uses maximum new tokens 1024, top-$p$ 0.001, repetition penalty 1.0, presence penalty 2.0.

**Rule-based scoring.** MMMU and MMMU-Pro use exact option-letter accuracy. PMC-VQA-clean and OmniMedVQA use rule-based multiple-choice matching against the gold option letter or answer text. PretexEval uses rule-based answer matching. No rule-based score is used for ChartQA, CharXiv-Val, or SciVQA in our reported results.

**LLM-judge scoring.** For ChartQA, CharXiv-Val, and SciVQA, we use Qwen2.5-72B-Instruct as an LLM judge for all reported scores. The judge receives the question, the standard answer, and the model response, and outputs a binary decision in the format `<judge>0</judge>` for correct or

Table 15: Post-warmup maximum CPT learning-rate comparison for Qwen3.5-4B-Base on Natural-1B CPT with fixed SFT. Bold and underlining mark the best and second-best values.

| Post-warmup max. CPT LR | Medical benchmarks | | | | | | General/scientific benchmarks | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | MMMU Med | MMMU-Pro Med | PMC VQA | Omni Med | Pretex Eval | Medical avg. | MMMU All | Chart QA | Char Xiv | Sci VQA | General avg. |
| $2.5 \times 10^{-6}$ | 59.30 | 39.86 | **61.85** | <u>85.04</u> | 62.48 | 61.71 | <u>54.44</u> | 76.44 | 50.40 | 66.74 | 62.01 |
| $5 \times 10^{-6}$ | **60.10** | <u>41.26</u> | <u>61.10</u> | **86.38** | <u>62.92</u> | **62.35** | **55.33** | **77.08** | **53.30** | **68.60** | **63.58** |
| $1 \times 10^{-5}$ | <u>59.42</u> | **41.96** | 60.80 | 84.45 | **63.84** | <u>62.09</u> | <u>54.44</u> | 76.20 | <u>52.00</u> | <u>67.29</u> | 62.48 |
| $3 \times 10^{-5}$ | 57.71 | 37.76 | 60.70 | 84.67 | 61.10 | 60.39 | 54.33 | **77.08** | 51.40 | 67.26 | <u>62.52</u> |
| $5 \times 10^{-5}$ | 54.34 | 34.27 | 60.45 | 81.46 | 62.49 | 58.60 | 53.33 | <u>76.48</u> | 47.20 | 65.26 | 60.57 |

Table 16: Reference-breaking counterfactual control. Superscripts give CPT token budgets. The control replaces figure-linked context with token-matched random text from the same source article. Bold marks the best values.

| Backbone | CPT condition | Medical benchmarks | | | | | | General/scientific benchmarks | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | MMMU Med | MMMU-Pro Med | PMC VQA | Omni Med | Pretex Eval | Medical avg. | MMMU All | Chart QA | Char Xiv | Sci VQA | General avg. |
| Qwen3.5-4B-Base | Reference-broken$^{1B}$ | 56.39 | 39.86 | 61.40 | 85.17 | 64.80 | 61.52 | 53.33 | **77.44** | 46.60 | 60.14 | 59.38 |
| | PMC-InterCPT$^{1B}$ | **59.36** | **41.96** | **61.50** | **86.19** | **67.37** | **63.28** | **55.44** | 76.64 | **51.60** | **67.45** | **62.78** |

`<judge>1</judge>` for incorrect. The judge prompt treats semantically equivalent answers as correct, which is important for open-ended scientific and chart questions where surface forms can differ from the reference.

# F Additional Results

Unless otherwise noted, bold and underlining mark the best and second-best values.

## F.1 Grounding Control

Table 16 provides the full per-benchmark values for the reference-breaking control. Breaking figure-context links lowers the medical average by 1.75 points and the general/scientific average by 3.40 points, indicating that the original figure-reference attachment contributes beyond in-domain article text alone.

## F.2 Learning-Rate Comparison Results

Figure 6 and Table 15 compare the maximum CPT learning rate reached after warmup on the Natural 1B mixture with fixed SFT. All entries use Qwen3.5-4B-Base and the fixed LLaVA-OneVision SFT stage. The $5\times10^{-6}$ configuration achieves the highest medical average (62.35) and general/scientific average (63.58), and is also best on MMMU-Med and OmniMedVQA. The neighboring $1 \times 10^{-5}$ configuration is strongest on MMMU-Pro Med and PretexEval, while $2.5 \times 10^{-6}$ is strongest on PMC-VQA-clean; neither improves both aggregate metrics. Raising the rate to $3\times10^{-5}$ and $5\times10^{-5}$ reduces both averages, especially at $5 \times 10^{-5}$.

## F.3 Per-Benchmark and Per-Subset Breakdown

Unless otherwise stated, Tables 17–19 compare Qwen3.5-4B-Base after the same fixed SFT stage: raw BIOMEDICA uses 13.65B CPT tokens and PMC-InterCPT uses 9.63B CPT tokens. Table 17 provides subject-level results for the two MMMU-style medical benchmarks. PMC-InterCPT recovers much of the performance drop caused by raw BIOMEDICA CPT on the harder MMMU-Pro-Medical-10 setting, improving over BIOMEDICA by 2.80 points overall. On MMMU-Pro-Med-10, the largest gains appear in Pharmacy (+8.77) and Public Health (+8.62), while Diagnostics and Laboratory Medicine improves over BIOMEDICA on both MMMU-Med-Test and MMMU-Pro-Med-10. The gains are not uniform across all subjects, which is expected because the final mixture is optimized for broad medical multimodal CPT rather than subject-specific overfitting.

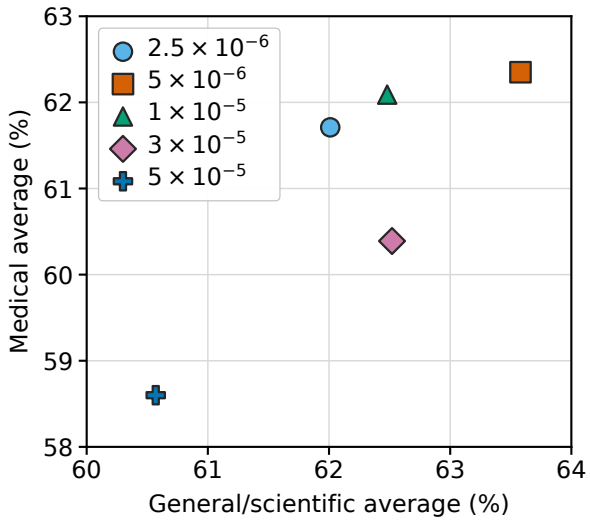


Figure 6: Aggregate medical and general results across post-warmup maximum CPT learning rates on Natural-1B with fixed SFT. Each color denotes one maximum learning rate.

Table 17: Per-subset medical benchmark results for Qwen3.5-4B-Base. Raw BIOMEDICA uses 13.65B CPT tokens and PMC-InterCPT uses 9.63B; both use the same fixed SFT stage. Bold marks the best values.

| Benchmark | Subset | BIOMEDICA | PMC-InterCPT |
|---|---|---|---|
| MMMU-Med | Total | 58.62 | **58.73** |
| MMMU-Med | Basic Medical Science | 66.87 | **67.18** |
| MMMU-Med | Clinical Medicine | **62.77** | 61.23 |
| MMMU-Med | Diagnostics and Laboratory Medicine | 46.91 | **48.77** |
| MMMU-Med | Pharmacy | 60.00 | **63.49** |
| MMMU-Med | Public Health | **53.24** | 50.88 |
| MMMU-Pro-Med-10 | Total | 37.76 | **40.56** |
| MMMU-Pro-Med-10 | Basic Medical Science | **50.00** | 46.15 |
| MMMU-Pro-Med-10 | Clinical Medicine | **35.59** | 32.20 |
| MMMU-Pro-Med-10 | Diagnostics and Laboratory Medicine | 23.33 | **26.67** |
| MMMU-Pro-Med-10 | Pharmacy | 49.12 | **57.89** |
| MMMU-Pro-Med-10 | Public Health | 32.76 | **41.38** |

Table 18 reports OmniMedVQA results by question type and image modality. PMC-InterCPT improves most clearly on Disease Diagnosis (+1.94 over BIOMEDICA) and Lesion Grading (+7.10), two categories that require clinically grounded visual interpretation rather than only recognizing the imaging modality. The modality-level results show a similar pattern: PMC-InterCPT improves over BIOMEDICA on CT (+2.19), MR (+2.81), and Fundus Photography (+1.37), while gains are smaller or do not appear for several other modalities. These category-level changes are consistent with the final curated, BVE-enhanced PMC-InterCPT improving clinically oriented visual questions, while not isolating the contribution of resampling from the other corpus-construction stages.

Table 19 gives additional breakdowns for the scientific and chart-oriented benchmarks. PMC-InterCPT improves ChartQA mainly on the human-written split (+3.52), while the augmented split changes only slightly. On SciVQA, PMC-InterCPT improves open-ended questions (+1.20) but slightly decreases multiple-choice accuracy (-1.08), producing a modest overall gain. The largest scientific-figure improvement appears on CharXiv-Val, where PMC-InterCPT improves every answer-type group and raises the total score by 6.10 points.

Table 18: OmniMedVQA breakdown by question type and image modality for Qwen3.5-4B-Base. Bold marks the best values.

| Group | Subset | N | BIOMEDICA | PMC-InterCPT |
|---|---|---|---|---|
| Question | Modality Recognition | 11,565 | 98.66 | **98.69** |
| Question | Disease Diagnosis | 55,387 | 84.27 | **86.21** |
| Question | Anatomy Identification | 16,448 | 80.74 | **81.36** |
| Question | Lesion Grading | 2,098 | 53.00 | **60.10** |
| Question | Other Biological Attributes | 3,498 | 89.88 | **91.08** |
| Modality | Fundus Photography | 5,398 | 79.49 | **80.86** |
| Modality | Microscopy Images | 5,680 | **84.91** | 84.37 |
| Modality | X-Ray | 7,916 | **87.44** | 87.41 |
| Modality | Dermoscopy | 6,679 | **83.74** | 82.45 |
| Modality | CT | 15,809 | 83.90 | **86.09** |
| Modality | Ultrasound | 10,991 | 80.47 | **82.08** |
| Modality | MR | 31,877 | 86.07 | **88.88** |
| Modality | OCT | 4,646 | **95.72** | 95.61 |

Table 19: Scientific and chart-oriented benchmark breakdowns for Qwen3.5-4B-Base. Bold marks the best values.

| Benchmark | Subset | N | BIOMEDICA | PMC-InterCPT |
|---|---|---|---|---|
| SciVQA | Total | 4,200 | 66.74 | **67.29** |
| SciVQA | Multiple choice | 1,200 | **70.25** | 69.17 |
| SciVQA | Open-ended | 3,000 | 65.33 | **66.53** |
| ChartQA | Total | 2,500 | 75.00 | **77.00** |
| ChartQA | Human | 1,250 | 63.76 | **67.28** |
| ChartQA | Augmented | 1,250 | 86.24 | **86.72** |
| CharXiv-Val | Total | 1,000 | 47.90 | **54.00** |
| CharXiv-Val | Answer type 1 | 440 | 49.55 | **55.68** |
| CharXiv-Val | Answer type 2 | 99 | 63.64 | **73.74** |
| CharXiv-Val | Answer type 3 | 232 | 51.29 | **57.33** |
| CharXiv-Val | Answer type 4 | 229 | 34.50 | **38.86** |