RefineNet: Elevating Medical Foundation Models through Quality-Centric Data Curation by MLLM-Annotated Proxy Distillation

Ningyi Zhang[1], Yuan Gao[2,3], Xin Wang[4], Ka-Hou Chan[1], Jian Wu[5,6,7], Chan-Tong Lam[1], Shanshan Wang[8], Yue Sun[1], Sio-Kei Im[1], and Tao Tan[1,⋆]

[1] Faculty of Applied Sciences, Macao Polytechnic University, Macao, China
taotan@mpu.edu.mo
[2] Imaging Division, University Medical Center Utrecht, Utrecht, The Netherlands
[3] Department of Radiology, The Netherlands Cancer Institute, The Netherlands
[4] GROW School, Maastricht University, Maastricht, The Netherlands
[5] The Second Affiliated Hospital and Liangzhu Laboratory, Zhejiang University School of Medicine, Hangzhou, China
[6] State Key Laboratory of Transvascular Implantation Devices and TIDRI, Hangzhou, China
[7] Zhejiang Key Laboratory of Medical Imaging Artificial Intelligence, Hangzhou, China
[8] Paul C. Lauterbur Research Center for Biomedical Imaging, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, Guangdong, China

Abstract. The rapid advancement of medical foundation models creates unprecedented demand for large-scale training data, yet existing medical repositories remain contaminated by heterogeneous mixtures of high- and low-quality image-text pairs—a severe data pollution problem that significantly bottlenecks model performance and optimization. While manual curation could theoretically ensure quality, it is impractical for managing large-scale datasets effectively. To address this critical challenge, we introduce RefineNet—a scalable framework that systematically refines data quality by distilling multimodal large language model (MLLM) insights into an offline reward model. RefineNet innovatively decouples human decision-making for quality assessment into two key dimensions: image-text fidelity and semantic consistency. By strategically filtering and curating datasets, RefineNet demonstrates remarkable performance improvements across diagnostic tasks. Specifically, our method selects 50% high-quality data subsets that outperform full-data baselines by 9.15% in Recall@10 (retrieval), 85.59 AUC (classification), and 72.59% accuracy (visual question answering). Moreover, RefineNet achieves notable agreement with human expert judgments (Pearson's r=0.67), providing clinicians an auditable bridge between automated curation and validation.

Keywords: Medical data curation · quality assessment · multimodal learning · foundation models.

[⋆] Corresponding author.

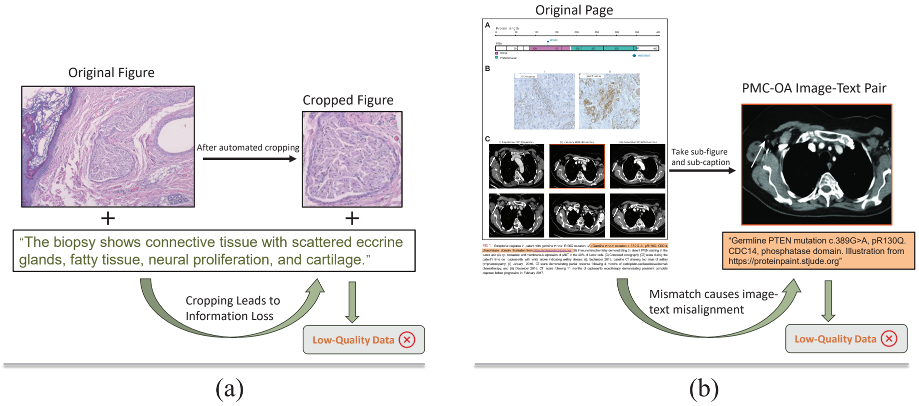

Original Figure
After automated cropping
Cropped Figure
+
+
"The biopsy shows connective tissue with scattered eccrine glands, fatty tissue, neural proliferation, and cartilage."
Cropping Leads to Information Loss
Low-Quality Data
(a)
Original Page
Take sub-figure and sub-caption
PMC-OA Image-Text Pair
"Germline PTEN mutation c.389G>A, pR130Q. CDC14, phosphatase domain. Illustration from https://proteinpaint.stjude.org"
Mismatch causes image-text misalignment
Low-Quality Data
(b)

Fig. 1. Image-text mismatch examples. (a) Cropped histopathology image paired with a general caption. (b) CT image incorrectly matched with a genetic mutation description. Both low-quality examples highlight the necessity of validation in image-text datasets.

1 Introduction

Reliable medical foundation models require precise image-text alignment, yet automated dataset curation often propagates systematic errors [1]. For instance, PMC-OA seeks diverse image-text data by auto-cropping multi-panel figures, but imprecise markers can cause misalignment and low-quality unverifiable data [2], as illustrated in Figure 1. Consequently, models trained on quality-filtered subsets outperform those using unfiltered datasets in diagnostic tasks, challenging the assumption that larger datasets necessarily lead to better performance [3].

These quality gaps in AI systems increase clinical risks, as image-text mismatches can amplify diagnostic errors through self-reinforcing feedback loops [4]. For instance, AI decisions using error-labeled data raised tumor staging errors 2.3-fold [5]. This undermines model reliability and patient safety, emphasizing the need for better assessment methods aligned with medical requirements [6].

However, current quality assessment methods fail to meet clinical needs. Rule-based heuristics lack semantic understanding [7], while metrics like ClipScore prioritize generic semantics, overlooking medical-specific details [8]. Human evaluation, though the gold standard, is cost-prohibitive for large-scale applications [9], creating a trilemma of balancing clinical validity, scalability, and cost efficiency.

To address these gaps, we propose RefineNet that bridges: translating the clinical reasoning capacity of multimodal large language model (MLLM) [10] into tractable quality dimensions and enabling auditability through interpretable scoring mirroring human workflows [11]. Evaluations show RefineNet-optimized data achieves superior performance (NIQE=6.18, Perplexity=26.95), with CLIP models trained on its curated small datasets outperforming full-data baselines. Unlike existing methods, RefineNet enhances evaluation accuracy, provides clinical interpretability, and aligns closely with human expert judgments, making it suitable for medical applications.

Our work makes three key contributions:

- Introduce a novel quality assessment framework utilizing MLLM-annotated datasets with RefineNet to enable efficient offline evaluation.
- Demonstrate that high-quality data curation consistently improves performance across quality metrics and downstream tasks, proving the value of quality-centric data scaling.
- RefineNet offers high human-aligned assessments, enabling reliable and scalable biomedical data curation with auditable human judgment.

2 Related Work

Data Quality Assessment Current quality assessment methods struggle to balance scalability with domain expertise. While rule-based approaches [12] provide transparency, they miss semantic understanding. Feature-alignment techniques like ClipScore [8] enhance cross-modal matching but retain pretraining biases unsuitable for medical imaging nuances [13]. Though human evaluation sets the medical gold standard [14], its expense prohibits large-scale application.

Automated Evaluation with MLLMs LLM-based frameworks like RLAIF enable scalable assessment [15], while multimodal models show diagnostic potential [16]. However, API dependencies and output variability hinder clinical deployment. Our solution distills MLLM capabilities into a deterministic reward model that preserves interpretability without cloud dependencies.

Data Scaling Strategies Traditional scaling laws emphasizing quantity [17] conflict with evidence of diminishing returns from low-quality data [18]. Though curriculum learning and coreset selection [19] address noise, they lack medical-specific quality dimensions. Our framework introduces biomedical-aware scaling through optimized multidimensional quality metrics.

3 Methodology

We construct our data quality assessment method through three steps: 1) Building a proxy dataset using decoupled criteria and MLLM annotations, 2) designing a parameter-efficient reward model, and 3) training the reward model to identify high-quality data. Figure 2 illustrates the pipeline.

3.1 Proxy Dataset: MLLM-Augmented Proxy Dataset Curation

Using MLLMs as human evaluator substitutes in automated assessments is effective, but advanced models like Gemini are often closed-source and API-dependent, limiting evaluation due to latency, costs, customization, and medical

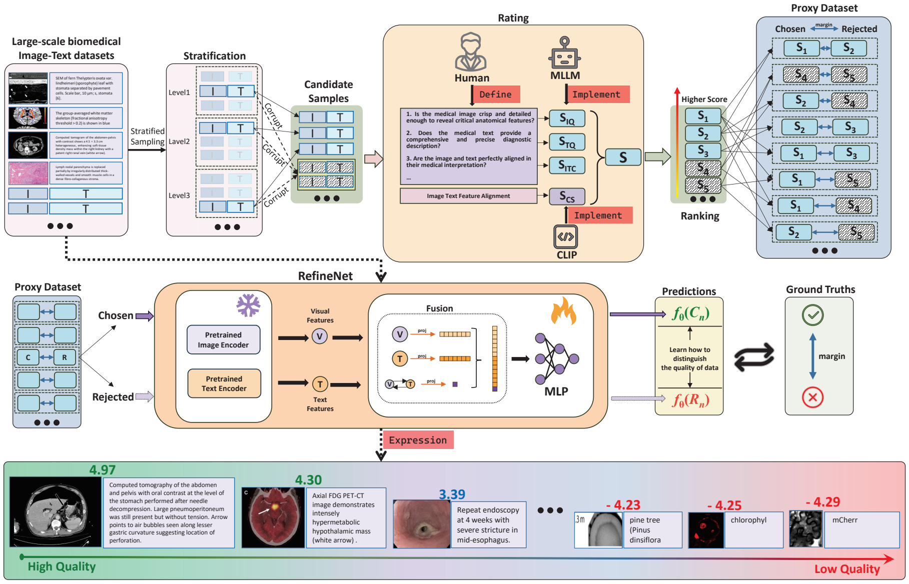

Large-scale biomedical Image-Text datasets
SEM of fern Thelypteris ovata var. lindheimeri (sporophyte) leaf with stomata separated by pavement cells. Scale bar, 10 μm; s, stomata [6].
The group-averaged white matter skeleton (fractional anisotropy threshold > 0.2) is shown in blue
Computed tomogram of the abdomen-pelvis with contrast shows a 4.5 × 3.3 cm heterogeneous, enhancing soft-tissue density mass within the right kidney with a patent right renal vein (white arrow).
Lymph nodal parenchyma is replaced partially by irregularly distributed thick-walled vessels and smooth muscle cells in a dense fibro-collagenous stroma.
I
T
Stratified Sampling
Stratification
Level1
Lavel2
Level3
Corrupt
Corrupt
Corrupt
Candidate Samples
Rating
Human
MLLM
Define
Implement
1. Is the medical image crisp and detailed enough to reveal critical anatomical features?
2. Does the medical text provide a comprehensive and precise diagnostic description?
3. Are the image and text perfectly aligned in their medical interpretation?
...
Image Text Feature Alignment
S_{IQ}
S_{TQ}
S_{ITC}
S_{CS}
S
Implement
CLIP
Higher Score
S_1
S_2
S_3
S_4
S_5
Ranking
Proxy Dataset
Chosen
margin
Rejected
Proxy Dataset
C
R
Chosen
Rejected
RefineNet
Pretrained Image Encoder
Pretrained Text Encoder
Visual Features
Text Features
V
T
Fusion
proj
MLP
Predictions
$f_\theta(C_n)$
Learn how to distinguish the quality of data
$f_\theta(R_n)$
Ground Truths
margin
Expression
4.97
Computed tomography of the abdomen and pelvis with oral contrast at the level of the stomach performed after needle decompression. Large pneumoperitoneum was still present but without tension. Arrow points to air bubbles seen along lesser gastric curvature suggesting location of perforation.
4.30
Axial FDG PET-CT image demonstrates intensely hypermetabolic hypothalamic mass (white arrow) .
3.39
Repeat endoscopy at 4 weeks with severe stricture in mid-esophagus.
- 4.23
pine tree (Pinus dinsiflora
- 4.25
chlorophyl
- 4.29
mCherr
High Quality
Low Quality

Fig. 2. The pipeline entails constructing a proxy dataset, training RefineNet, and scoring the image-text dataset.

data privacy concerns [20]. To address this, we built a proxy dataset to gather high- and low-quality human-centric samples with measurable quality margins, training an offline reward model to generalize across various medical data domains.

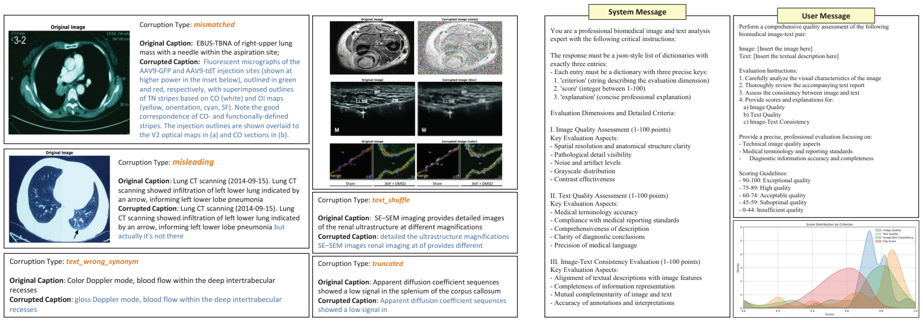
First, we use stratified sampling by ClipScore percentiles to create a diverse subset from 1.6 million image-text pairs. To improve discrimination, we add 10% negative samples by corrupting images, text, and creating mismatches, as in Figure 3 (a).

Subsequently, We analyzed 10,990 samples with Gemini-1.5-Flash [21], focusing on image quality, text accuracy, and visual-text relevance. We also included ClipScore for assessing concept similarity beyond human metrics. The instructions and distribution of scores are shown in Figure 3 (b). Finally, we normalized the scores and generated 4.85 million contrastive triplets as proxy dataset for reward model training.

3.2 Architecture: Multidimensional Quality-Aware Reward Model

We introduce RefineNet, a parameter-efficient reward model optimized with discriminative signals to select high-quality data from a proxy dataset, capturing the evaluation capabilities of MLLM and human judgment logic.

RefineNet employs a frozen CLIP encoder with ResNet50 [22] and PubMedBert [23], pre-trained on medical data, to extract features. A trainable scorer maps these to a shared space, combines their dot product, and fuses them. The



Corruption Type: *mismatched*

Original Caption: EBUS-TBNA of right-upper lung mass with a needle within the aspiration site;
Corrupted Caption: Fluorescent micrographs of the AAV9-GFP and AAV9-tdT injection sites (shown at higher power in the inset below), outlined in green and red, respectively, with superimposed outlines of TN stripes based on CO (white) and OI maps (yellow, orientation, cyan, SF). Note the good correspondence of CO- and functionally-defined stripes. The injection outlines are shown overlaid to the V2 optical maps in (a) and CO sections in (b).

Corruption Type: *misleading*

Original Caption: Lung CT scanning (2014-09-15). Lung CT scanning showed infiltration of left lower lung indicated by an arrow, informing left lower lobe pneumonia
Corrupted Caption: Lung CT scanning (2014-09-15). Lung CT scanning showed infiltration of left lower lung indicated by an arrow, informing left lower lobe pneumonia but actually it's not there

Corruption Type: *text_wrong_synonym*

Original Caption: Color Doppler mode, blood flow within the deep intertrabecular recesses
Corrupted Caption: gloss Doppler mode, blood flow within the deep intertrabecular recesses

Corruption Type: *text_shuffle*

Original Caption: SE–SEM imaging provides detailed images of the renal ultrastructure at different magnifications
Corrupted Caption: detailed the ultrastructure magnifications SE–SEM images renal imaging at of provides different

Corruption Type: *truncated*

Original Caption: Apparent diffusion coefficient sequences showed a low signal in the splenium of the corpus callosum
Corrupted Caption: Apparent diffusion coefficient sequences showed a low signal in

(a)

System Message

You are a professional biomedical image and text analysis expert with the following critical instructions:

The response must be a json-style list of dictionaries with exactly three entries:
- Each entry must be a dictionary with three precise keys:
  1. 'criterion' (string describing the evaluation dimension)
  2. 'score' (integer between 1-100)
  3. 'explanation' (concise professional explanation)

Evaluation Dimensions and Detailed Criteria:

I. Image Quality Assessment (1-100 points)
Key Evaluation Aspects:
- Spatial resolution and anatomical structure clarity
- Pathological detail visibility
- Noise and artifact levels
- Grayscale distribution
- Contrast effectiveness

II. Text Quality Assessment (1-100 points)
Key Evaluation Aspects:
- Medical terminology accuracy
- Compliance with medical reporting standards
- Comprehensiveness of description
- Clarity of diagnostic conclusions
- Precision of medical language

III. Image-Text Consistency Evaluation (1-100 points)
Key Evaluation Aspects:
- Alignment of textual descriptions with image features
- Completeness of information representation
- Mutual complementarity of image and text
- Accuracy of annotations and interpretations

User Message

Perform a comprehensive quality assessment of the following biomedical image-text pair:

Image: [Insert the image here]
Text: [Insert the textual description here]

Evaluation Instructions:
1. Carefully analyze the visual characteristics of the image
2. Thoroughly review the accompanying text report
3. Assess the consistency between image and text
4. Provide scores and explanations for:
  a) Image Quality
  b) Text Quality
  c) Image-Text Consistency

Provide a precise, professional evaluation focusing on:
- Technical image quality aspects
- Medical terminology and reporting standards
- Diagnostic information accuracy and completeness

Scoring Guidelines:
- 90-100: Exceptional quality
- 75-89: High quality
- 60-74: Acceptable quality
- 45-59: Suboptimal quality
- 0-44: Insufficient quality

(b)

Fig. 3. (a) Examples of corrupted images and texts. (b) Multidimensional assessment instructions and the distribution of scores across each dimension

fused features are processed through multilayer perceptron (MLP) layers to generate a scalar quality score for the given image-text pair:

$$f(x_i, x_t) = \text{MLP}([\phi_{\text{img}}(x_i) \oplus \phi_{\text{txt}}(x_t) \oplus (\phi_{\text{img}}(x_i) \odot \phi_{\text{txt}}(x_t))]) \tag{1}$$

where ϕ is the encoder, $\oplus$ denotes concatenation, and $\odot$ represents element-wise multiplication. $f(x_i, x_t)$ is RefineNet's quality score for the image-text pair (x_i, x_t).

3.3 Training Objective: Margin-Optimized Contrastive Learning

RefineNet aims to create a metric space that distinguishes sample quality by mapping features so that high-quality samples cluster together and low-quality samples are separated, using *Margin* to measure this difference. For the given training samples (*Chosen*, *Rejected*, *Margin*), where $c = (x_i^c, x_t^c)$ and $r = (x_i^r, x_t^r)$ represent chosen and rejected image-text pairs respectively, a binary rank loss with margin is used to optimize RefineNet:

$$\mathcal{L} = -\mathbb{E}_{(c,r,m)\sim\mathcal{D}} \log \sigma(f(c) - f(r) - m) \tag{2}$$

where σ is the sigmoid activation function, c/r are chosen/rejected samples and m their quality margin. This objective enforces proportional score differences aligned with human/MLLM assessments.

The final model enables efficient offline quality evaluation, distilling MLLM judgment capabilities into lightweight parametric form while maintaining interpretability through explicit quality dimensions.

4 Experiments

4.1 Experimental Setup

Dataset, Baselines and Evaluation Protocol We evaluate RefineNet on PMC-OA [24], one of the largest open-source biomedical dataset containing 1.6M

Table 1. Top 1k image-text pairs evaluated using image (NIQE, BRISQUE) and text (Perplexity, LLM-TQE) quality metrics, with **bold** for best and underline for second-best results. LLM-TQE scores from Llama-3.1-8b [28].

Model	Image Quality Assessment		Text Quality Assessment	
	NIQE ↓	BRISQUE ↓	Perplexity ↓	LLM-TQE ↑
ImageReward	7.18	30.15	2368.62	6.17
BlipScore	6.97	33.46	471.82	6.47
ClipScore	6.95	27.11	82.24	7.16
RefineNet	**6.18**	**26.43**	**26.95**	**7.86**

image-text pairs affected by contextual fragmentation. Baselines include: 1) Full dataset training; 2) Random/Bucket sampling; 3) Existing metrics (ClipScore [8], BlipScore [25], ImageReward [26]). Evaluation is conducted across three dimensions:

– **Image/Text Quality**: Evaluate the individual image and text quality of high-quality data from automated methods using no-reference metrics. NIQE and BRISQUE [27] for images, Perplexity and LLM-based Text Quality Evaluator (LLM-TQE) for text [28].
– **Downstream Tasks** A fixed CLIP model is trained from scratch with high-quality data collected via automated methods. Data quality is assessed indirectly using downstream task metrics: retrieval (Recall@K on Roco [29]), classification (AUC on MedMNIST [30]), and VQA (Accuracy on Slake [31]).
– **Human Alignment**: Evaluate the correlation of various quality assessment methods with human scoring results using Pearson and Spearman scores to determine their reliability and safety.

4.2 Evaluation of RefineNet for Image and Text Quality Refinement

RefineNet surpasses other data refinement methods, demonstrated by its superior performance in evaluating the top 1,000 high-quality images and texts from the same test set. For image quality, RefineNet achieves the lowest NIQE and BRISQUE scores, indicating high naturalness, clarity, and low distortion, with ClipScore ranking second. For text quality, RefineNet records the lowest Perplexity, reflecting natural and coherent text, and the highest LLM-TQE score, demonstrating superior text quality aligned with LLM standards.

4.3 Ablation Study of Multi-Dimensional Evaluation Criteria

Ablation studies demonstrated that Image-Text Consistency (ITC) surpassed isolated Image Quality (IQ) and Text Quality (TQ) in tasks such as Visual Question Answering (VQA), while ClipScore (CS) enhanced retrieval. Integrating ITC, IQ, TQ, and CS yielded the most effective results, balancing quality assessment and semantic alignment for robust CLIP training.

Table 2. Ablation study of multidimensional criteria (IQ: Image Quality, TQ: Text Quality, ITC: Image-Text Consistency, CS: ClipScore).

IQ	TQ	ITC	CS	Retrieval	VQA	Classification
✓	×	×	×	40.77	60.53	83.39
×	✓	×	×	39.72	58.83	83.74
×	×	✓	×	42.85	61.81	84.37
×	×	×	✓	49.30	59.95	84.90
✓	✓	×	×	42.55	58.67	84.48
✓	✓	✓	×	44.20	59.86	85.13
✓	✓	✓	✓	**52.13**	**66.14**	**85.59**

Table 3. Performance comparison of CLIP models trained on full-size and refined datasets at 75%, 50%, and 25% ratios across retrieval, VQA, and classification tasks.

Ratio	Model	CrossModel Retrieval						VQA		Classification
		Image to Text			Text to Image			Closed	Open	Average of 12 Datasets
		R@1	R@5	R@10	R@1	R@5	R@10			
100%	Fullsize	12.50	33.30	44.20	13.25	32.85	45.20	62.98	59.68	83.49
75%	Random	10.35	29.05	41.20	11.15	27.80	40.00	63.46	53.64	80.83
	Bucket	12.05	29.25	41.80	11.30	28.20	41.00	62.74	54.72	81.31
	Worst	4.30	14.30	22.00	3.50	12.30	20.40	62.98	53.79	80.89
	ImageReward	13.10	34.35	46.40	12.70	32.90	44.05	**66.34**	55.65	81.62
	ClipScore	14.50	36.10	48.85	14.25	35.15	46.70	63.70	**58.91**	83.33
	BlipScore	14.10	36.00	46.80	12.65	34.75	45.85	**66.34**	55.81	81.00
	RefineNet	**15.35**	**38.00**	**49.75**	**14.85**	**35.60**	**47.80**	65.14	55.81	**83.39**
50%	Random	8.80	24.80	35.15	8.50	23.60	33.90	66.82	57.51	81.82
	Bucket	8.60	25.30	36.45	8.20	24.35	34.95	65.86	52.55	82.04
	Worst	1.60	7.05	11.65	1.80	7.25	11.20	60.57	52.55	80.18
	ImageReward	13.10	32.10	42.50	11.35	31.00	42.35	64.42	54.41	83.24
	ClipScore	14.90	37.95	48.80	15.60	35.90	48.00	62.98	55.81	83.59
	BlipScore	13.00	34.15	45.65	13.40	33.25	43.80	63.94	55.34	84.16
	RefineNet	**18.60**	**41.90**	**53.35**	**16.80**	**40.00**	**50.90**	**72.59**	**59.68**	**85.59**
25%	Random	3.65	11.45	17.85	3.35	9.75	15.45	62.18	56.27	78.37
	Bucket	3.35	12.30	18.30	3.70	11.10	16.55	62.98	53.79	80.33
	Worst	0.15	0.60	1.40	0.05	0.95	1.80	52.88	42.10	72.44
	ImageReward	6.05	17.80	26.65	5.75	17.15	24.90	57.93	52.86	81.68
	ClipScore	9.85	25.95	36.85	8.10	24.00	34.45	62.25	57.20	80.88
	BlipScore	7.15	19.45	28.05	7.30	19.05	27.05	60.33	54.72	80.86
	RefineNet	**10.65**	**27.65**	**38.85**	**10.00**	**27.55**	**37.80**	**63.70**	**57.82**	**82.13**

4.4 Improving CLIP Performance with Less but High-Quality Data

RefineNet-curated data at 50% volume (777k pairs) outperformed full PMC-OA (1.6M pairs) across tasks (Table 3). Retrieval R@10 (Image→Text) improved by 9.15%, and VQA accuracy exceeded baselines despite 50% fewer samples. Classification achieved higher average AUC than full-data training, excelling in 9 of 12 datasets. Control groups (random/worst sampling) performed significantly worse, highlighting the necessity of quality-centric curation.

The results highlight the significant noise (at least 25%) in automatically collected, large-scale medical datasets like PMC-OA, but show that a curated high-quality subset can improve multimodal foundational models more effectively than merely increasing data volume.

Table 4. Comparison of correlation metrics between automated methods and human ratings. * sets the upper limit for RefineNet, as it is trained on distilled data from Gemini 1.5 Flash, resulting in superior correlation scores.

Methods	Correlation Metrics		
	Spearman	Pearson	P-Value
Gemini 1.5 Flash *	0.7158	0.7633	$<$ **0.01**
ImageReward	0.1824	0.2869	$>$ **0.05**
ClipScore	0.5171	0.4226	$<$ **0.01**
BlipScore	0.5987	0.6486	$<$ **0.01**
RefineNet	**0.6226**	**0.6688**	$<$ **0.01**

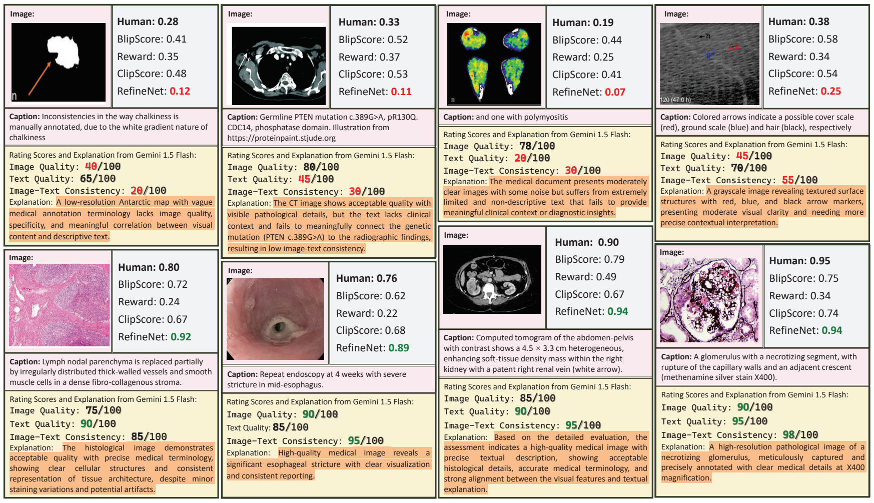


Image:
Human: 0.28
BlipScore: 0.41
Reward: 0.35
ClipScore: 0.48
RefineNet: **0.12**
Caption: Inconsistencies in the way chalkiness is manually annotated, due to the white gradient nature of chalkiness
Rating Scores and Explanation from Gemini 1.5 Flash:
Image Quality: **40**/100
Text Quality: **65**/100
Image-Text Consistency: **20**/100
Explanation: A low-resolution Antarctic map with vague medical annotation terminology lacks image quality, specificity, and meaningful correlation between visual content and descriptive text.

Image:
Human: 0.80
BlipScore: 0.72
Reward: 0.24
ClipScore: 0.67
RefineNet: **0.92**
Caption: Lymph nodal parenchyma is replaced partially by irregularly distributed thick-walled vessels and smooth muscle cells in a dense fibro-collagenous stroma.
Rating Scores and Explanation from Gemini 1.5 Flash:
Image Quality: **75**/100
Text Quality: **90**/100
Image-Text Consistency: **85**/100
Explanation: The histological image demonstrates acceptable quality with precise medical terminology, showing clear cellular structures and consistent representation of tissue architecture, despite minor staining variations and potential artifacts.

Image:
Human: 0.33
BlipScore: 0.52
Reward: 0.37
ClipScore: 0.53
RefineNet: **0.11**
Caption: Germline PTEN mutation c.389G>A, pR130Q. CDC14, phosphatase domain. Illustration from https://proteinpaint.stjude.org
Rating Scores and Explanation from Gemini 1.5 Flash:
Image Quality: **80**/100
Text Quality: **45**/100
Image-Text Consistency: **30**/100
Explanation: The CT image shows acceptable quality with visible pathological details, but the text lacks clinical context and fails to meaningfully connect the genetic mutation (PTEN c.389G>A) to the radiographic findings, resulting in low image-text consistency.

Image:
Human: 0.76
BlipScore: 0.62
Reward: 0.22
ClipScore: 0.68
RefineNet: **0.89**
Caption: Repeat endoscopy at 4 weeks with severe stricture in mid-esophagus.
Rating Scores and Explanation from Gemini 1.5 Flash:
Image Quality: **90**/100
Text Quality: **85**/100
Image-Text Consistency: **95**/100
Explanation: High-quality medical image reveals a significant esophageal stricture with clear visualization and consistent reporting.

Image:
Human: 0.19
BlipScore: 0.44
Reward: 0.25
ClipScore: 0.41
RefineNet: **0.07**
Caption: and one with polymyositis
Rating Scores and Explanation from Gemini 1.5 Flash:
Image Quality: **78**/100
Text Quality: **20**/100
Image-Text Consistency: **30**/100
Explanation: The medical document presents moderately clear images with some noise but suffers from extremely limited and non-descriptive text that fails to provide meaningful clinical context or diagnostic insights.

Image:
Human: 0.90
BlipScore: 0.79
Reward: 0.49
ClipScore: 0.67
RefineNet: **0.94**
Caption: Computed tomogram of the abdomen-pelvis with contrast shows a 4.5 × 3.3 cm heterogeneous, enhancing soft-tissue density mass within the right kidney with a patent right renal vein (white arrow).
Rating Scores and Explanation from Gemini 1.5 Flash:
Image Quality: **85**/100
Text Quality: **90**/100
Image-Text Consistency: **95**/100
Explanation: Based on the detailed evaluation, the assessment indicates a high-quality medical image with precise textual description, showing acceptable histological details, accurate medical terminology, and strong alignment between the visual features and textual explanation.

Image:
Human: 0.38
BlipScore: 0.58
Reward: 0.34
ClipScore: 0.54
RefineNet: **0.25**
Caption: Colored arrows indicate a possible cover scale (red), ground scale (blue) and hair (black), respectively
Rating Scores and Explanation from Gemini 1.5 Flash:
Image Quality: **45**/100
Text Quality: **70**/100
Image-Text Consistency: **55**/100
Explanation: A grayscale image revealing textured surface structures with red, blue, and black arrow markers, presenting moderate visual clarity and needing more precise contextual interpretation.

Image:
Human: 0.95
BlipScore: 0.75
Reward: 0.34
ClipScore: 0.74
RefineNet: **0.94**
Caption: A glomerulus with a necrotizing segment, with rupture of the capillary walls and an adjacent crescent (methenamine silver stain X400).
Rating Scores and Explanation from Gemini 1.5 Flash:
Image Quality: **90**/100
Text Quality: **95**/100
Image-Text Consistency: **98**/100
Explanation: A high-resolution pathological image of a necrotizing glomerulus, meticulously captured and precisely annotated with clear medical details at X400 magnification.

Fig. 4. Visualizations of data quality assessment examples show the input image (top-left), paired text (middle-top), automated/human scores (top-right, with green/red for highest/lowest), and Gemini 1.5 Flash's ratings and explanations (bottom).

4.5 Human Alignment and Clinical Quality Factors

Five biomedical PhDs evaluated 100 image-text samples to test RefineNet's alignment with human judgment (Table 4). Gemini 1.5 Flash, despite its high relevance, is a closed-source model limited to API calls, posing potential risks of medical data leakage. In contrast, RefineNet aligns closely with human judgment, supports offline operation, and requires minimal trainable parameters, ensuring data security and reliable quality assessment. Furthermore, Figure 4 identifies three key clinical quality indicators: (1) precise terminology (e.g., "methenamine silver stain X400"), (2) clear annotations (e.g., distinct markers like "white arrow"), and (3) information density alignment (e.g. exact matches such as "4×3.3 centimeters" and "right renal vein"). High-quality pairs (human score > 0.85)

use more specific terms and fewer ambiguities than low-quality samples. These attributes correlate strongly with cross-modal retrieval performance, highlighting their significance in developing reliable diagnostic models.

5 Conclusion & Disscusion

This work introduces RefineNet, a novel approach for medical data curation that prioritizes quality over quantity, creating more reliable multimodal foundation models. By distilling MLLM expertise into an offline reward model, it achieves privacy-compliant, efficient quality assessment. This approach challenges traditional scaling by showing that high-quality data subsets can boost model performance. RefineNet offers an auditable framework aligned with human expertise. Current limitations include domain biases and static assessments, while future research will focus on dynamic quality thresholds and clinician interventions to improve model safety.

Acknowledgments. This research was supported by the National Natural Science Foundation of China (Grant Nos. 62222118, U22A2040, and 12326612), Guangdong Provincial Key Laboratory of Artificial Intelligence in Medical Image Analysis and Application (No. 2022B1212010011), Shenzhen Science and Technology Program (Grant Nos. RCYX20210706092104034 and JCYJ20220531100213029), Youth Innovation Promotion Association of Chinese Academy of Sciences, Macao Polytechnic University Grant (RP/FCA-15/2022, RP/FCA-10/2023, and RP/FCA-08/2024), Science and Technology Development Fund of Macao (0041/2023/RIB2, 0021/2022/AGJ, and 0004/2024 /E1B1), submission control Macao Polytechnic University (fca.269a.bdb3.4), Zhejiang Key R&D Program of China (Grant No. 2023C03053), Zhejiang Key Laboratory of Medical Imaging Artificial Intelligence, State Key Laboratory of Transvascular Implantation Devices (Grant No. SKLTID2024003), and the Transvascular Implantation Devices Research Institute (TIDRI).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Huang, W., Li, C., Zhou, H.-Y., Yang, H., Liu, J., Liang, Y., Zheng, H., Zhang, S., Wang, S.: Enhancing representation in radiography-reports foundation model: A granular alignment algorithm using masked contrastive learning. Nature Communications **15**(1), 7620 (2024)
2. Kingston, B., Bailleux, C., Delaloge, S., Schiavon, G., Scott, V., Lacroix-Triki, M., Carr, T.H., Kozarewa, I., Gevensleben, H., Kemp, Z., et al.: Exceptional response to AKT inhibition in patients with breast cancer and germline PTEN mutations. JCO Precision Oncology **3**, 1–7 (2019)
3. Budach, L., Feuerpfeil, M., Ihde, N., Nathansen, A., Noack, N., Patzlaff, H., Naumann, F., Harmouch, H.: The effects of data quality on machine learning performance. arXiv preprint arXiv:2207.14529 (2022)

4. Jiang, D., Dou, W., Vosters, L., Xu, X., Sun, Y., Tan, T.: Denoising of 3D magnetic resonance images with multi-channel residual learning of convolutional neural network. Japanese Journal of Radiology **36**, 566–574 (2018)
5. Waite, S., Scott, J., Gale, B., Fuchs, T., Kolla, S., Reede, D.: Interpretive error in radiology. American Journal of Roentgenology **208**(4), 739–749 (2017)
6. Kozegar, E., Soryani, M., Behnam, H., Salamati, M., Tan, T.: Computer aided detection in automated 3-D breast ultrasound images: a survey. Artificial Intelligence Review **53**(3), 1919–1941 (2020)
7. Heinrich, B., Klier, M., Schiller, A., Wagner, G.: Assessing data quality–a probability-based metric for semantic consistency. Decision Support Systems **110**, 95–106 (2018). Elsevier
8. Hessel, J., Holtzman, A., Forbes, M., Bras, R. L., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning. arXiv preprint arXiv:2104.08718 (2021)
9. Carrell, D.S., Cronkite, D.J., Malin, B.A., Aberdeen, J.S., Hirschman, L.: Is the juice worth the squeeze? Costs and benefits of multiple human annotators for clinical text de-identification. Methods of Information in Medicine **55**(4), 356–364 (2016)
10. Saab, K., Tu, T., Weng, W.-H., Tanno, R., Stutz, D., Wulczyn, E., Zhang, F., Strother, T., Park, C., Vedadi, E., et al.: Capabilities of gemini models in medicine. arXiv preprint arXiv:2404.18416 (2024)
11. Waltersdorfer, L., Ekaputra, F.J., Miksa, T., Sabou, M.: AuditMAI: Towards an infrastructure for continuous AI auditing. arXiv preprint arXiv:2406.14243 (2024)
12. Müller-Budack, E., Theiner, J., Diering, S., Idahl, M., Hakimov, S., Ewerth, R.: Multimodal news analytics using measures of cross-modal entity and context consistency. International Journal of Multimedia Information Retrieval **10**(2), 111–125 (2021)
13. Wang, Y., Chen, Y., Yan, W., Fang, A., Zhou, W., Jamieson, K., Du, S. S.: CLIPLoss and Norm-Based Data Selection Methods for Multimodal Contrastive Learning. arXiv preprint arXiv:2405.19547 (2024)
14. Regenwetter, L., Srivastava, A., Gutfreund, D., Ahmed, F.: Beyond statistical similarity: Rethinking metrics for deep generative models in engineering design. Computer-Aided Design **165**, 103609 (2023)
15. Lee, H., Phatale, S., Mansoor, H., Lu, K. R., Mesnard, T., Ferret, J., Bishop, C., Hall, E., Carbune, V., Rastogi, A.: RLAIF: Scaling reinforcement learning from human feedback with AI feedback. (2023)
16. Khan, S., Biswas, M. R., Murad, A., Ali, H., Shah, Z.: An Early Investigation into the Utility of Multimodal Large Language Models in Medical Imaging. arXiv preprint arXiv:2406.00667 (2024)
17. Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., Amodei, D.: Scaling laws for neural language models. arXiv preprint arXiv:2001.08361 (2020)
18. Li, Z., Xie, C., Cubuk, E. D.: Scaling (Down) CLIP: A Comprehensive Analysis of Data, Architecture, and Training Strategies. arXiv preprint arXiv:2404.08197 (2024)
19. Zhou, X., Pi, R., Zhang, W., Lin, Y., Chen, Z., Zhang, T.: Probabilistic bilevel coreset selection. In: Proceedings of the International Conference on Machine Learning, pp. 27287–27302 (2022). PMLR
20. Riedemann, L., Labonne, M., Gilbert, S.: The path forward for large language models in medicine is open. npj Digital Medicine **7**(1), 339 (2024)

21. Team, G., Georgiev, P., Lei, V. I., Burnell, R., Bai, L., Gulati, A., Tanzer, G., Vincent, D., Pan, Z., Wang, S., et al.: Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv preprint **arXiv:2403.05530** (2024)
22. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
23. Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., Poon, H.: Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare (HEALTH) **3**(1), 1–23 (2021). ACM New York, NY
24. Lin, W., Zhao, Z., Zhang, X., Wu, C., Zhang, Y., Wang, Y., Xie, W.: Pmc-clip: Contrastive language-image pre-training using biomedical documents. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 525–536. Springer (2023)
25. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: Proceedings of the International Conference on Machine Learning, pp. 12888–12900. PMLR (2022)
26. Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J., Dong, Y.: Imagereward: Learning and evaluating human preferences for text-to-image generation. Advances in Neural Information Processing Systems **36**, 15903–15935 (2023)
27. Mittal, A., Moorthy, A.K., Bovik, A.C.: No-reference image quality assessment in the spatial domain. IEEE Transactions on Image Processing **21**(12), 4695–4708 (2012)
28. Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al.: The llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024)
29. Pelka, O., Koitka, S., Rückert, J., Nensa, F., Friedrich, C.M.: Radiology Objects in COntext (ROCO): A Multimodal Image Dataset. In: Stoyanov, D., et al. (eds.) Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis. LABELS CVII STENT 2018, Lecture Notes in Computer Science, vol. 11043, pp. 1–13. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01364-6_20
30. Yang, J., Shi, R., Ni, B.: Medmnist classification decathlon: A lightweight automl benchmark for medical image analysis. In: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), pp. 191–195. IEEE (2021)
31. Liu, B., Zhan, L.-M., Xu, L., Ma, L., Yang, Y., Wu, X.-M.: Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), pp. 1650–1654. IEEE (2021)