PhenoLIP: Phenotype Guided Medical Vision–Language Pretraining

Cheng Liang[1,2], Chaoyi Wu[1], Weike Zhao[1,2], Ya Zhang[1,2], Yanfeng Wang[1,2,†], and Weidi Xie[1,2,†]

[1] Shanghai Jiao Tong University, Shanghai, China
[2] Shanghai Artificial Intelligence Laboratory, Shanghai, China
[†] Corresponding authors

Code https://github.com/MAGIC-AI4Med/PhenoLIP
Model https://www.modelscope.cn/models/lcsjtu/PhenoLIP
Benchmark https://www.modelscope.cn/datasets/lcsjtu/PhenoBench

Abstract. Recent progress in CLIP-like vision-language models (VLMs) has greatly advanced medical image analysis. However, most existing medical VLMs still rely on coarse image-text contrastive objectives and fail to capture the systematic visual knowledge encoded in well-defined medical phenotype ontologies. To address this limitation, we construct **PhenoKG**, the first large-scale, phenotype-centric multimodal knowledge graph that encompasses around 524K high-quality image-text pairs linked to more than 3,000 phenotypes. Building upon PhenoKG, we propose **PhenoLIP**, a pretraining framework that explicitly incorporates structured phenotype knowledge into medical VLMs through a two-stage process. We first learn a knowledge-enhanced phenotype embedding space that captures the hierarchical structure of phenotype ontologies and then distill this textual knowledge into multimodal pretraining via a teacher-guided knowledge distillation objective. To support evaluation, we further introduce **PhenoBench**, an expert-verified benchmark designed for phenotype recognition, comprising over 7,800 image-caption pairs covering more than 1,000 phenotypes. Extensive experiments demonstrate that PhenoLIP outperforms previous state-of-the-art baselines, improving upon BIOMEDICA in phenotype classification accuracy by 8.06% and in cross-modal retrieval by 15.03%, underscoring the value of integrating phenotype-centric priors into medical VLMs for structured and interpretable medical image understanding.

1 Introduction

Large-scale vision-language model pretraining, such as CLIP [21], has made significant progress in computer vision, which has also inspired medical image analysis. Current medical VLMs [16,18,35], trained on massive collections of medical image–text pairs through contrastive learning, demonstrate strong cross-modal understanding and generalization capabilities. They have revolutionized a wide range of medical visual and multimodal tasks, *e.g.*, disease diagnosis [15, 16, 20],

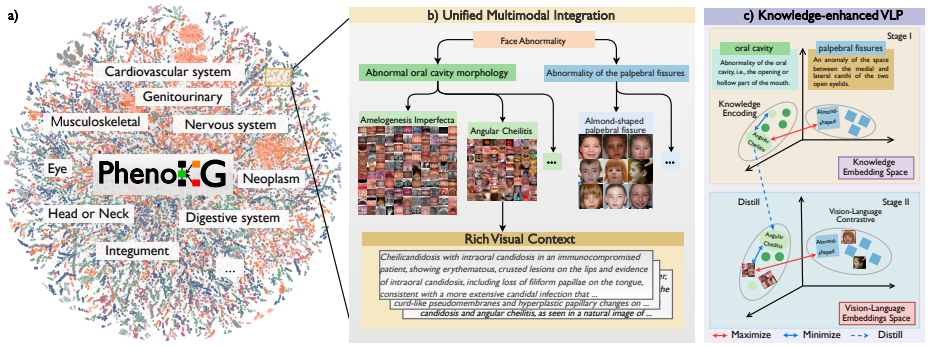

Fig. 1: The figure illustrates our core methods **PhenoKG** and **PhenoLIP**, from left to right: **(a)** macro level visualization of PhenoKG, the first large scale phenotype centric multimodal knowledge graph that hierarchically organizes diverse anatomical systems; **(b)** unified multimodal integration that aligns phenotype images, rich visual descriptions, and structured phenotype ontology knowledge into a single graph; and **(c) PhenoLIP**, our phenotype knowledge enhanced vision language pretraining framework, which first learns a structured phenotype embedding from the ontology and then guides the vision language pretraining via distillation.

radiology report generation [14, 29], and medical visual question answering [8], achieving expert-level performance on several benchmarks.

Conventional medical vision–language pretraining typically relies on heterogeneous supervision sources, such as clinical reports and large-scale image–caption pairs from the medical literature; however, the accompanying text is often free-form and inconsistently expressed rather than grounded in standardized clinical terminology or a shared ontology, that weakens image–text alignment and limits transfer to diagnosis-oriented tasks. Phenotype ontologies address this bottleneck by providing standardized terms and a hierarchy that captures semantic granularity: because these phenotype terms closely reflect the language clinicians use in diagnosis, they make the supervision space more clinically grounded and more consistent across specialties, while also normalizing diverse descriptions into unified phenotype concepts and separating findings that are linguistically similar but clinically distinct. As a result, phenotype-centric pretraining introduces structured semantic anchors beyond raw image–text pairing, enabling more precise and clinically meaningful medical image understanding.

In this paper, we construct the first, large-scale, phenotype-centric multimodal medical knowledge graph, termed as **PhenoKG**, providing the structured and multimodal knowledge that current VLMs lack, as shown in Figure 1. Our approach builds upon the Human Phenotype Ontology (HPO) [9], a well-established taxonomy of human phenotypes with rich clinical annotations and disease associations. To incorporate multimodal information, we develop a scalable pipeline that automatically mines and filters medical image-caption pairs from literature. The resulting high-quality image–caption pairs are linked to phenotype nodes, enriching the knowledge graph with multimodal evidence. Un-

like prior vision–language datasets built from unstructured web data [18, 23], **PhenoKG** unifies visual data with structured phenotypic knowledge, bridging medical image understanding with ontology-based priors. In total, **PhenoKG** comprises 5,839 phenotype groups and 13,812 specific phenotypes. Among these phenotypes, 3,096 are associated with image–caption pairs, yielding around 524K image–caption pairs that establish a systematic foundation for medical Vision-Language Pretraining.

In contrast to conventional medical VLMs such as BiomedCLIP [35] and BIOMEDICA [18], which are typically pretrained on large but unstructured collections of image-caption pairs, our approach is fundamentally knowledge-driven. We propose **P**henotype-centric **L**anguage–**I**mage **P**retraining, termed as **PhenoLIP**, a two-stage pretraining framework designed to gradually distill the structured multimodal medical knowledge into VLMs. We first train a phenotype knowledge encoder on the textual ontology within **PhenoKG** to map phenotype names, definitions, and relationships into a unified semantic space. Optimized with contrastive learning, the encoder pulls semantically related phrases, such as synonyms, definitions, and hierarchical relations, closer in the embedding space. In the second stage, during the main vision-language pretraining on the multimodal data within PhenoKG, we employ the frozen knowledge encoder as a teacher to guide the VLM's text encoder. Through an auxiliary knowledge distillation objective, the text encoder is regularized to align its outputs with the teacher's knowledge-enhanced embedding space. This process runs in parallel with the primary image-text contrastive alignment, effectively injecting phenotype ontology priors into the VLM and enabling it to ground visual evidence in fine-grained, structured phenotype concepts.

For evaluation, we introduce **PhenoBench**, an expert-verified benchmark specifically designed for phenotype recognition, that contains 7,819 image-caption pairs covering 1,187 distinct phenotypes. This task represents a fundamental capability in medical image analysis that underpins more advanced applications, such as diagnosis, medical visual question answering, and report generation. We evaluate our final pre-trained VLMs on both PhenoBench and multiple public datasets across diverse tasks, including phenotype classification, fine-grained image-to-phenotype retrieval, and rare facial phenotype identification. Our method outperforms state-of-the-art biomedical VLM by 8.06% average accuracy in phenotype classification and 15.03% in cross modal retrieval on 9 downstream datasets, including our proposed **PhenoBench**. The results underscore the effectiveness of our approach and highlight the transformative potential of integrating structured knowledge with multimodal pre-training to enable more precise and semantically-grounded reasoning in medical AI.

2 Related Work

VLP in medical image analysis. Vision–Language Pre-training (VLP) aims to learn joint representations from large-scale image–text data. General-domain models such as CLIP [21] and its successors [26,34] have achieved strong perfor-

mance by aligning images and text in a shared embedding space using contrastive learning. However, when transferred to specialized domains like medicine, these models suffer from a vocabulary gap and limited ability to capture fine-grained clinical semantics [23], underscoring the need for domain-specific pretraining. Recent medical VLMs have extended CLIP-style frameworks into biomedicine. BiomedCLIP [35] pre-trains on 15M biomedical image-text pairs from PubMed, while PMC-CLIP [16] leverages figure-caption pairs from PubMed Central to learn general biomedical representations.

Knowledge-enhanced medical VLMs. Beyond general medical VLMs, several approaches have explored incorporating structured medical knowledge into vision-language models. DermLIP [31] constructs a dermatology-specific dataset and employs ontology-based text refinement to enhance representations. KEP [39] and KEEP [38] focuses on pathology images and incorporates structured disease knowledge to enhance vision-language alignment. In radiology, MedKLIP [30] extracts report information into structured triplets and enriches entities with knowledge-based descriptions. KAD [36] leverages a medical knowledge graph (UMLS [4]) represented as concept–relation–concept triplets, and uses entities/relations extracted from reports to guide vision–language pretraining. Although existing knowledge-enhanced approaches excel in specialized domains, they neglect the critical role of *phenotype ontology knowledge*. This ontology defines atom-level, fine-grained pathology priors, providing a comprehensive knowledge foundation for medical imaging tasks across all modalities.

Medical knowledge graph. Medical Knowledge Graphs (KGs) are essential for organizing complex biomedical information. Large-scale KGs like UMLS [4], OrphaNet [28], and OMIM [1] have structured vast amounts of data on terminologies, rare diseases, and genes. iKraph [37], PrimgKG [6] links molecular and genetic factors to clinical outcomes by structuring 4 million relationships across entities like proteins, biological processes, phenotypes, and drugs. Among these, the Human Phenotype Ontology (HPO) [9] provides a standardized, hierarchical vocabulary for human phenotypic abnormalities. However, existing medical KGs, including HPO, are predominantly text-based and lack explicit links to visual data, which severely restricts their utility in medical image analysis. Our work aims to bridge this modality gap.

Medical vision-language datasets. In the recent literature, significant progress has been made for curating large-scale, high-quality image–caption data in medical domains. BIOMEDICA [18], PMC-CLIP [16], and PMC-15M [35] constructed datasets covering diverse medical images, charts, and related text descriptions through automated crawling strategies. Derm1M [31] focused on dermatology, providing rich image–caption pairs through fine-grained annotation and classification. Quilt-1M [13] collected pathology-related videos and their subtitles from YouTube, generating paired image–text data through multi-stage preprocessing. Although these datasets have greatly expanded medical multimodal resources, they remain fragmented and lack a systematic organization aligned with phenotype ontology systems. This absence of structured phenotype-centered knowledge limits their effectiveness for developing comprehensive medical VLMs.

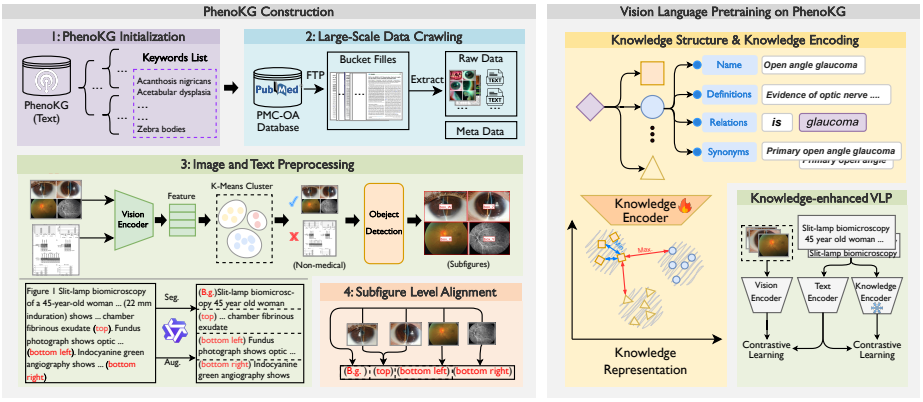

Fig. 2: Overview of the **PhenoKG** construction pipeline and the **PhenoLIP** training process. **Left**: PhenoKG Construction. This illustrates the four-stage pipeline for building our multimodal knowledge graph: (1) PhenoKG Initialization, (2) Crawling image-caption pairs from the PMC-OA database using phenotype keywords, (3) Preprocessing images (via clustering-based filtering and subfigure detection) and text (via LLM-based refinement), (4) Aligning subfigures with their corresponding descriptions. **Right**: PhenoLIP Pretraining. This depicts our two-stage pretraining framework. First, a knowledge encoder is trained on PhenoKG's textual ontology to learn a structured phenotype embedding. Second, during vision-language pretraining, this frozen knowledge encoder serves as a teacher to distill its structured knowledge into the VLM's text encoder, complementing the primary image-text contrastive alignment.

3 PhenoKG & PhenoBench Construction

In this section, we detail the construction of a large-scale multimodal phenotype knowledge graph. As illustrated in Figure 2 (left), the procedure contains two main stages. First, we initialize its textual part based on the Human Phenotype Ontology (HPO) (Sec. 3.1). Second, we develop an automated pipeline to augment this graph with around 524K aligned image-caption pairs obtained from biomedical literature (Sec. 3.2), the detailed statistics are listed in Appendix 9.2. Finally, by splitting the crawled articles, we also construct **PhenoBench**, an expert-annotated benchmark for phenotype recognition evaluation.

3.1 Knowledge Graph Initialization

To establish **PhenoKG**, we first construct its textual part based on the Human Phenotype Ontology (HPO) [9], a standardized ontology system for phenotypes. We collected 19,703 phenotype terms as graph nodes, together with their definitions, relational edges, and synonyms from the Human Phenotype Ontology (HPO). This collection serves as an initialized text-only phenotype knowledge graph, comprising 19,703 nodes and 23,528 edges that represent hierarchical relationships and correlations among phenotype terms. The construction details and statistics of this process are provided in Appendix 14.1.

3.2 Multimodal Data Collection and Processing

After constructing the textual phenotype KG, we enriched each node with multimodal evidence from the biomedical literature using the following pipeline.

Data crawling. Based on the constructed knowledge graph, we first filter a list of keywords. We retained only terminal nodes, *i.e.*, those without outgoing edges, and removed other non-terminal, categorical nodes such as "abnormality of the nervous system". Leveraging this keyword list, we processed the entire PubMed Central Open Access (PMC-OA) corpus [22], extracting figure–caption pairs whose captions contained exact matches to any phenotype keyword. In addition to captions, we retrieved the corresponding in-text figure-referencing paragraph (e.g., "...as seen in Figure 1..."), preserving contextual cues for later disambiguation and grounding.

Image processing. As illustrated in Figure 2, raw figure–caption pairs contain substantial noise, including compound figures and non-medical graphics (*e.g.*, charts and graphs). First, to remove non-medical images, we embedded all figures with DINOv3 and clustered the embeddings via K-means into 20 coarse clusters, each further subdivided into 20 sub-clusters (400 clusters in total) [19, 24]. We then manually audited the clusters, sampling representative images from each cluster to assign a semantic label, and retained only those dominated by medically relevant content. Second, to split compound figures, we applied a pretrained subfigure detector (DAB-DETR) [2, 17] to segment multi-panel figures into subfigures. Third, we re-applied the clustering-based filter to the extracted subfigures to eliminate residual non-medical content. This procedure yielded around 524K medical figure–caption pairs. On expert-annotated validation sets, our cluster-based filter attains 96.2%/93.5% precision/recall for identifying medical images, and the DAB-DETR subfigure detector reaches 94.1%/92.7% precision/recall (93.4% F1). Additional implementation details are provided in Appendix 14.2, and the full human-in-the-loop evaluation in Appendix 10.

Text processing. For each retained figure, we kept the original caption and its in-text paragraph. We then used Qwen3 [32] to synthesize a cleaned caption per image, removing non-informative tokens and numeric artifacts while preserving semantic content. For compound figures that were segmented into subfigures, we further prompted the LLM with a tailored instruction to produce subfigure-level textual descriptions consistent with the original semantics (Appendix 12.1).

Subfigure level alignment. We aligned subfigures with their corresponding subfigure texts. If the number of subfigures and candidate texts did not match, we conservatively retained the original compound figure–caption pair. Otherwise, we invoked Qwen2.5-VL [3] to perform fine-grained matching. Specifically, we rendered the compound image overlaid with subfigure bounding boxes and unique identifiers (*e.g.*, *box_1*, *box_2*) and prompted the model to map each identifier to the most relevant text segment (Fig. 2; Appendix 12.2). An ablation on overlaying boxes for prompting is reported in Appendix 10.2.

Multimodal integration. The final step was to integrate the collected image-caption pairs into the original textual phenotype KG nodes. Since each im-

age–caption pair is filtered by search keywords that are directly linked to the original phenotype nodes, it is straightforward to insert them into the textual KG structure based on keywords.

3.3 PhenoBench Construction.

In parallel with PhenoKG, we curated an independent evaluation suite, **PhenoBench**, using a matched but disjoint pipeline to prevent data leakage. We enforced a strict split at the document and figure levels by partitioning the source corpus using unique `PMCID` and `FigureID`. Articles and figures assigned to PhenoBench are excluded from any training or development sets, ensuring no overlap of images, captions, or near-duplicates across splits. All candidate image–caption pairs were manually reviewed by clinical experts to confirm medical relevance, caption fidelity to the visual content, and accurate phenotype linkage. The final benchmark contains 7,819 image–caption pairs linked to 1,187 distinct phenotype terms and can be used to evaluate phenotype recognition (based on the linked phenotype terms) and image–text retrieval (based on the original image–caption pair supervision).

4 Vision-Language Pretraining on PhenoKG

We present **PhenoLIP**, a vision–language pretraining framework that distills structured phenotype knowledge from PhenoKG into joint visual–textual representations (Figure 2). The training comprises two stages. First, we train a phenotype-aware text encoder to capture relational structure over the phenotype KG (Sec. 4.1). Second, we use this knowledge-enhanced encoder to guide vision–language contrastive pretraining, aligning images with text while injecting KG priors into the VLM (Sec. 4.2).

4.1 Phenotype Knowledge Encoding

To leverage the phenotype knowledge in PhenoKG, we learn text embeddings for phenotype concepts in a space explicitly regularized by ontology relations, hierarchical structure, and lexical semantics. As illustrated in Figure 2 (Right), we train a phenotype aware text encoder with metric learning to map each term, its definition, its synonyms, and its hierarchical is a relations into a latent space, enabling effective utilization of phenotype knowledge.

Problem formulation. Given a set of phenotype nodes, each can be denoted as (T_i, A_i), where T_i represents the phenotype term, and A_i represents its related attribute set, including various textual types: (i) the phenotype name, (ii) its detailed definition, (iii) its synonymous names, and (iv) relational descriptions derived from its related hierarchical (is-a) relations (e.g., "*[head phenotype] is a child phenotype of [tail phenotype]*"). Our goal is to learn a knowledge encoder $\Phi_{\mathrm{k}}(\cdot)$ that satisfies:

$$\mathrm{sim}(\Phi_{\mathrm{k}}(a_i), \Phi_{\mathrm{k}}(a_i^+)) \gg \mathrm{sim}(\Phi_{\mathrm{k}}(a_i), \Phi_{\mathrm{k}}(a_j)),\ i \neq j, \tag{1}$$

where $\mathrm{sim}(\cdot,\cdot)$ denotes the similarity function, a_i and a_i^+ represent two attributes sampled from the phenotype attribute set A_i, and a_j is an attribute sampled from another phenotype set A_j. The encoder can map related phenotype attributes to similar vector representations, while mapping unrelated concepts to distant vector representations.

Training pipeline. To optimize the objective in Eq. 1, we train the knowledge encoder with a contrastive objective that pulls together multiple textual descriptions of the same phenotype concept and pushes apart descriptions from different concepts. Specifically, given a sampled attribute a_i, we first sample another positive attribute pair a_i^+ from PhenoKG and encode them into textual embeddings with the knowledge encoder, termed as $z = \Phi_{\mathrm{k}}(a)$, resulting in (z_i, z_i^+). Given a mini training batch with $2B$ batch size, which consists of B phenotypes, each with two randomly sampled attribute descriptions, we adopt an InfoNCE-style loss with in-batch negatives:

$$\mathcal{L}_{\mathrm{k}} = -\frac{1}{2B}\sum_{i=1}^{2B}\log\frac{\exp(\mathrm{sim}(z_i, z_i^+)/\tau_1)}{\sum_{k=1}^{2B}\mathbf{1}_{k\neq i}\exp(\mathrm{sim}(z_i, z_k)/\tau_1)}, \tag{2}$$

where $\mathrm{sim}(\cdot,\cdot)$ denotes cosine similarity, τ_1 is a temperature hyperparameter. By minimizing $\mathcal{L}_{\mathrm{k}}$, the knowledge encoder Φ_{k} learns to project semantically related textual attribute descriptions of the same phenotype terms more closely in the embedding space.

Construction of positive pairs. Rather than concatenating all attributes of a phenotype into a single sequence, we form two types of positive pairs and encode each side *independently* with the shared encoder: **(i)** (*name*/*synonym*, *definition*) pairs; and **(ii)** (*head–relation*, *tail*) pairs taken from ontology edges. For type (ii), the *head* and *relation* texts are joined with a `[SEP]` token, and the `[CLS]` embedding of the joined sequence serves as the query, which is contrasted with the independently encoded *tail* text, so that the different attributes of each phenotype are aggregated around a shared semantic anchor.

Implementation details: We adopt the pretrained PubmedBERT [10] as the initialization of the knowledge encoder. More details are listed in Appendix 15.1.

4.2 Knowledge-Enhanced VLP

This section details of knowledge-enhanced visual representation learning, leveraging the former phenotype knowledge encoder Φ_{k} trained on PhenoKG.

Problem formulation. Given the image–caption pairs collected from PhenoKG, denoted as (x_i, c_i), our objective is to train a visual encoder Φ_{v} and a text encoder Φ_{t}, such that their joint embedding space satisfies:

$$\mathrm{sim}(\Phi_{\mathrm{v}}(x_i), \Phi_{\mathrm{t}}(c_i)) \gg \mathrm{sim}(\Phi_{\mathrm{v}}(x_i), \Phi_{\mathrm{t}}(c_j)),\ i \neq j. \tag{3}$$

Training pipeline. During training, as illustrated in Figure 2, we adopt InfoNCE-style contrastive learning to align visual and textual representations. In addition,

to ensure that the text embedding space effectively captures the hierarchical knowledge encoded in the knowledge graph (KG), we introduce an auxiliary knowledge distillation objective, which regularizes the learned text embeddings to align with the pretrained knowledge encoder.

Specifically, denoting the visual embedding as $v_i = \Phi_{\mathrm{v}}(x_i)$ and text embedding as $t_i = \Phi_{\mathrm{t}}(c_i)$, the multimodal contrastive learning loss can be defined as:

$$\mathcal{L}_{\mathrm{M}} = -\frac{1}{B}\sum_{i=1}^{B}\left[\log\frac{\langle v_i, t_i\rangle_{\tau_2}}{\sum_{j=1}^{B}\langle v_i, t_j\rangle_{\tau_2}} + \log\frac{\langle t_i, v_i\rangle_{\tau_2}}{\sum_{j=1}^{B}\langle t_i, v_j\rangle_{\tau_2}}\right], \tag{4}$$

where $\langle a, b\rangle_\tau = \exp(\mathrm{sim}(a, b)/\tau)$ denotes the exponentiated cosine similarity scaled by a temperature τ, and B is the training batch size. Additionally, we introduce a knowledge distillation branch with the pretrained knowledge encoder $\Phi_{\mathrm{k}}(\cdot)$ acting as a teacher model. For each caption c_i in the batch, the teacher produces a reference knowledge embedding $k_i = \Phi_{\mathrm{k}}(c_i)$. We then enforce consistency between the student's text embedding t_i and the teacher's knowledge embedding k_i using a text-knowledge contrastive loss, denoted as $\mathcal{L}_{\mathrm{KD}}$. This loss function encourages the student's text encoder to map captions into the pre-structured semantic space of the teacher by treating (t_i, k_i) as a positive pair and all other non-corresponding pairs (t_i, k_j) as negatives within the batch, formulated as:

$$\mathcal{L}_{\mathrm{KD}} = -\frac{1}{B}\sum_{i=1}^{B}\left[\log\frac{\langle t_i, k_i\rangle_{\tau_3}}{\sum_{j=1}^{B}\langle t_i, k_j\rangle_{\tau_3}} + \log\frac{\langle k_i, t_i\rangle_{\tau_3}}{\sum_{j=1}^{B}\langle k_i, t_j\rangle_{\tau_3}}\right], \tag{5}$$

where τ_3 is the distillation temperature, reusing the notation $\langle\cdot,\cdot\rangle_\tau$ defined above. The overall pretraining objective combines the multimodal contrastive loss and the knowledge distillation loss:

$$\mathcal{L}_{\mathrm{VLP}} = \mathcal{L}_{\mathrm{M}} + \alpha\mathcal{L}_{\mathrm{KD}}, \tag{6}$$

where α balances the two objectives. Further implementation details are provided in Appendix 15.2.

Implementation details. We adopt BiomedCLIP [35] as the initialization for the visual encoder, while the text encoder is initialized with the weights from the trained phenotype knowledge encoder. Details are in Appendix 15.2.

5 Experiments

In this section, we introduce our experiments in detail, including evaluation settings, related datasets, and metrics.

Evaluation settings. We assess three standard regimes: zero-shot classification, cross-modal retrieval, and linear probing. In **zero-shot classification**, the model assigns labels to unseen tasks by comparing image embeddings with class-specific textual prompts, without fine-tuning, following CLIP [21]; prompt

templates are in Appendix 13.1. In **cross-modal retrieval**, we evaluate image-to-text and text-to-image retrieval to measure alignment between image and phenotype descriptions. In **linear probing**, we freeze the encoders and train a single linear layer on top of the visual features to gauge representation quality.

Evaluation datasets. We benchmarked on diverse biomedical datasets spanning 7 imaging modalities and 10 medical image analysis tasks under the three regimes above. Dataset descriptions, splits, evaluation protocols, and their clinical features are detailed in Appendix 11.

- For **zero-shot classification**, we employ HAM10000 [25] and DermaMNIST [33] for dermatology, LC25000 [5] for pathology, BreastMNIST, ChestMNIST, and RSNA [27] for radiology, and BloodMNIST, TissueMNIST, RetinaMNIST, and OCTMNIST [33] for hematology, histology, and ophthalmology, respectively, along with our proposed PhenoBench for phenotype recognition.
- For **cross-modal retrieval**, we adopt PhenoBench across four distinct retrieval scenarios: I2T (image-to-text), T2I, I2P (image-to-phenotype) and P2I retrieval. Additionally, we involve Face2Gene [11], to evaluate the model's performance in rare disease related facial phenotype retrieval, denoting a fine-grained and long-tail benchmark that challenges the model to associate subtle visual traits with corresponding phenotype concepts.
- For **linear probing**, we adopt six benchmark datasets: RSNA [27], HAM10000 [25], and four datasets from MedMNIST v2 [33], including BreastMNIST, ChestMNIST, OCTMNIST, and RetinaMNIST. We report the results under 1%, 10%, and 100% of the training data, respectively.

Baselines. We consider two main categories of state-of-the-art vision–language models (VLMs) for comparison, *i.e.*, general-domain VLMs and biomedical VLMs. The former category includes Open-CLIP [7], SigLIP2 [26], and CoCa [34], which are pre-trained on large-scale general-domain web image–text data. The latter category comprises PMC-CLIP [16], BiomedCLIP [35], and BIOMEDICA [18], all pre-trained on domain-specific medical image–text data. We also compare with knowledge-enhanced medical VLMs, including DermLIP [31], KEEP [38], MedKLIP [30] and KAD [36], which inject medical knowledge through different mechanisms (details in Appendix 16.1). A comprehensive cross-domain performance comparison with these knowledge-enhanced baselines is provided in Appendix 16.2.

Metrics. Model performance is evaluated with metrics tailored to each setting. For zero-shot classification and linear probing, we report classification accuracy (ACC) to measure recognition and representation quality. For cross-modal retrieval tasks, including the rare facial phenotype matching on Face2Gene [11], we use Recall@K (R@K) as the main evaluation metric, along with Precision, Recall, and F1-Score to capture fine-grained matching performance in clinically relevant retrieval scenarios.

Table 1: Benchmarking results on zero-shot image classification (Acc). Each column reports accuracy on the datasets of that modality: Derm.: Dermatology (DermaMNIST, HAM10000); Path.: Pathology (LC25000); Radio.: Radiology (ChestMNIST, BreastMNIST, RSNA); Hema.: Hematology (BloodMNIST); Histo.: Histology (TissueMNIST); Ophthal.: Ophthalmology (OCTMNIST, RetinaMNIST); Avg.: Average. Best method per column is **bold**; second-best model is underlined.

Method	Encoder		Accuracy							
	Vision	Text	Derm.	Path.	Radio.	Hema.	Histo.	Ophthal.	PhenoBench	Avg.
General VLMs										
OpenCLIP [7]	ViT-B	GPT2	14.77	33.33	29.75	23.00	4.52	19.43	2.26	18.15
SigLIP2 [26]	So400m	SigLIP64	11.31	20.00	4.95	8.30	**9.00**	23.00	0.25	10.97
CoCa [34]	ViT-B	GPT2	12.63	33.00	25.87	6.55	3.39	20.55	1.38	14.77
Biomedical VLMs										
PMC-CLIP [16]	ResNet50	PubmedBert	41.35	45.12	29.10	21.50	5.68	43.75	7.50	27.71
BiomedCLIP [35]	ViT-B	PubmedBert	47.59	42.87	28.47	20.40	4.23	40.18	8.15	27.41
BIOMEDICA [18]	ViT-L	GPT2	**56.76**	57.40	37.72	9.35	5.40	26.15	6.69	28.50
PhenoLIP	ViT-B	PubmedBert	55.08	**58.33**	**49.03**	**23.24**	7.87	**51.63**	**10.76**	**36.56**

6 Results

In this section, we analyze the main results, comparing PhenoLIP with baselines across all evaluated settings, followed by ablation studies, as presented in Table 1, 2, 4, and 3.

6.1 Zero-Shot Classification

We first benchmark models on their zero-shot accuracy in classifying medical images into task-related classes.

PhenoLIP unlocks superior zero-shot capabilities. As shown in Table 1, PhenoLIP substantially outperforms both general-domain and biomedical VLMs across a wide range of medical imaging modalities, demonstrating its superior capabilities in medical image understanding and perception. Notably, it achieves the highest average accuracy of 36.56%, substantially surpassing the next best-performing model, BIOMEDICA [18], which attained an accuracy of 28.50%. The performance gain can be observed on 5 imaging modalities, indicating the model's strong generalization ability across diverse medical imaging domains.

PhenoKG delivers robust cross-specialty generalization. The strength of our approach is particularly evident in phenotype-rich domains. In Radiology, PhenoLIP attains 49.03% accuracy, creating a remarkable 11.31% gap with the second-best model. Strong performance is also seen in Pathology, where its 58.33% accuracy represents a 0.93% absolute improvement over BIOMEDICA. This trend of superior performance extends to other specialized fields like Hematology (23.24%) and Ophthalmology (51.63%). Conversely, the lack of improvement on the Histology (TissueMNIST) dataset is attributable to the nature of the task, which involves classifying broad tissue types rather than phenotype-wise abnormalities.

Table 2: Cross-modal retrieval results on PhenoBench (7,819 images, 1,187 phenotypes). I2T represents image-to-text retrieval and T2I represents text-to-image retrieval. I2P represents image-to-phenotype retrieval and P2I represents phenotype-to-image retrieval. The best-performing model for each setting is in **bold**, the second-best is underlined.

Method	Encoder		I2T		T2I		I2P		P2I	
	Vision	Text	R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50
General VLMs										
OpenCLIP [7]	ViT-B	GPT2	10.72	24.25	10.08	22.95	2.88	8.34	2.79	8.24
SigLIP2 [26]	So400m	SigLIP64	14.02	28.16	10.33	23.44	0.31	0.76	0.08	0.62
CoCa [34]	ViT-B	GPT2	9.27	22.16	7.57	18.93	2.02	6.42	1.82	5.59
Biomedical VLMs										
PMC-CLIP [16]	ViT-L	PubmedBert	40.00	64.82	36.68	61.83	7.22	23.44	6.64	18.92
BiomedCLIP [35]	ViT-B	PubmedBert	32.91	56.63	32.43	56.08	3.71	13.38	3.77	12.17
BIOMEDICA [18]	ViT-L	GPT2	40.51	66.28	40.03	67.38	8.12	25.27	6.82	19.60
PhenoLIP	ViT-B	PubmedBert	**63.30**	**81.92**	**66.61**	**87.68**	**13.84**	**36.88**	**12.77**	**31.30**

6.2 Cross-modal Retrieval.

In this section, we evaluate the cross-modal retrieval performance of PhenoLIP on two fronts: first, on the diverse PhenoBench dataset for both standard image-caption and phenotype-centric retrieval, and second, on the challenging Face2Gene [11] benchmark for identifying rare phenotypes.

PhenoLIP enhances cross-modal retrieval performance on PhenoBench. In Table 2, we present the cross-modal retrieval results on PhenoBench, covering diverse medical image-caption pairs. Its diversity is summarized in Table 2. As shown by the results, in the standard I2T retrieval, PhenoLIP achieves an R@10 of 63.30%, surpassing the second-best model, BIOMEDICA [18], by a remarkable margin of 22.79%. Similar improvement can be observed in the T2I task, where our model's R@10 of 66.61% exceeds BIOMEDICA's 40.03%. This demonstrates that PhenoLIP has learned a significantly more accurate and robust alignment between various medical images and their corresponding textual descriptions.

PhenoLIP enables better alignment between visual features and phenotypes. Crucially, PhenoLIP's advantage becomes even more pronounced in the more challenging and clinically meaningful phenotype-centric retrieval tasks. In the I2P retrieval setting, our model attains an R@10 of 13.84%, nearly doubling the performance of the runner-up, BIOMEDICA (8.12%). Similarly, for P2I retrieval, PhenoLIP reaches an R@10 of 12.77%, markedly surpassing BIOMEDICA's 6.82%. Its substantial lead in both I2P and P2I tasks, across a challenging vocabulary of 1187 distinct phenotype classes, confirms that our PhenoLIP method has successfully established a robust, bidirectional alignment between visual features and their specific clinical phenotype representations.

PhenoLIP demonstrates better perception of rare phenotypes. To assess the model's ability to identify rare phenotypes from the long tail of the distribution, we introduce a rare phenotype retrieval task, linking medical facial images with related phenotype terms. As shown in Table 4, PhenoLIP consis-

Table 3: Benchmarking results on Linear Evaluation (Acc) across different data ratios. The best-performing model for each setting is in **bold**, and the second-best is underlined.

	RSNA			BreastMNIST			ChestMNIST			OCTMNIST			RetinaMNIST			HAM10000			Avg.		
	1%	10%	100%	1%	10%	100%	1%	10%	100%	1%	10%	100%	1%	10%	100%	1%	10%	100%	1%	10%	100%
General VLMs																					
OpenCLIP [7]	77.07	79.52	81.61	32.05	73.72	83.33	51.31	52.88	53.69	67.40	73.20	73.10	43.50	55.00	61.75	52.81	68.65	77.23	54.02	67.16	71.79
SigLIP2 [26]	54.55	68.90	77.50	32.50	73.50	83.00	51.05	52.50	53.80	67.00	72.90	72.50	43.80	54.50	61.00	53.24	66.67	72.57	50.36	64.83	70.06
CoCa [34]	77.29	79.59	81.36	36.54	78.21	82.05	51.62	52.85	53.48	43.75	55.75	59.50	**52.24**	56.92	58.37	55.45	66.67	72.94	52.82	65.00	67.95
Biomedical VLMs																					
PMC-CLIP [16]	82.49	**83.74**	83.76	26.92	75.00	82.05	49.50	53.70	**55.19**	**76.20**	69.90	71.60	45.00	56.50	60.00	54.13	64.36	75.25	55.71	67.20	71.31
BiomedCLIP [35]	81.99	83.66	83.99	**51.92**	**79.49**	84.62	51.31	54.46	55.03	68.20	73.80	74.10	43.50	55.75	59.00	59.74	65.02	72.61	59.44	68.70	71.56
BIOMEDICA [18]	80.09	82.94	83.44	37.82	77.56	85.90	51.99	53.57	54.45	61.40	76.40	76.70	43.50	55.50	**63.50**	57.76	**69.31**	73.60	55.43	69.21	72.93
PhenoLIP	**83.06**	83.36	**84.11**	49.36	74.36	**86.54**	**52.65**	**54.71**	55.15	73.90	**78.30**	**77.10**	43.50	**57.75**	61.75	**62.71**	**69.31**	**77.89**	**60.86**	**69.63**	**73.76**

Table 4: Rare facial phenotype identification results on the Face2Gene [11] dataset (321 images). Each image is associated with 21 phenotypes on average, and the retrieval pool contains 993 candidate phenotypes. The best result per column is in **bold**, and the second best is underlined.

Method	Retrieval Metrics (%)			Matching Metrics (%)		
	R@5	R@10	R@50	Prec.	Recall	F1
General VLMs						
OpenCLIP [7]	6.97	11.94	39.80	1.63	2.56	1.81
SigLIP2 [26]	6.85	11.80	39.50	1.50	2.10	1.75
CoCa [34]	4.48	7.46	38.31	1.04	2.21	1.25
Biomedical VLMs						
PMC-CLIP [16]	6.75	11.60	49.50	1.65	3.20	2.17
BiomedCLIP [35]	6.84	11.73	49.84	1.80	3.77	2.22
BIOMEDICA [18]	3.48	7.46	43.28	1.10	2.37	1.36
PhenoLIP	**7.49**	**12.05**	**55.05**	**2.08**	**4.56**	**2.62**

tently outperforms the other evaluated methods across all metrics in this highly challenging setting. The model's advantage is most evident in the retrieval, where it achieves the highest scores with an R@5 of 7.49% and R@50 of 55.05%. It also leads in direct phenotype matching, attaining the best F1-score of 2.62%. The low absolute matching scores across all models reflect the task's extreme difficulty, as each image has, on average, 21 related phenotypes within a vast pool of 993 facial phenotype candidates. Despite this challenge, PhenoLIP's consistent top-ranking performance demonstrates its enhanced ability to capture subtle visual signatures of rare diseases.

6.3 Linear Probing

Table 3 presents the linear probing results, which evaluate the quality and transferability of the learned visual representations. Our PhenoLIP model demonstrates remarkable performance across 18 evaluation settings (6 datasets at three data ratios), PhenoLIP achieves the best average performance at all three data ratios, including 1%, 10%, and 100%.

Table 5: Ablation study results. We evaluate the impact of different encoder initializations, knowledge distillation, and data curation strategies on image-to-text (I2T) retrieval, text-to-image (T2I) retrieval, and classification performance. BCLIP denotes vision encoder of BiomedCLIP, PMB and KB denotes PubmedBert and our pretrained knowledge encoder. Our full model, combining all components, achieves the best results across all tasks. The Best results and related configurations are in **bold**.

Encoder		Know.	Data	I2T		T2I		Cls.
Vision	Text	Distill.	Cur.	R@10	R@50	R@10	R@50	Acc
Scratch	KB			28.15	45.33	31.04	49.81	2.13
CLIP	KB			47.20	68.91	51.72	73.05	6.44
BCLIP	KB			50.11	71.22	53.68	75.90	7.02
CLIP	PMB			49.53	70.88	54.88	78.14	6.95
BCLIP	PMB			53.42	74.10	58.03	81.25	7.81
CLIP	PMB	✓		58.91	78.54	62.19	84.33	9.57
CLIP	PMB	✓	✓	60.13	79.62	63.55	85.18	9.98
BCLIP	PMB	✓	✓	**63.30**	**81.92**	**66.61**	**87.68**	**10.76**

The representations from PhenoLIP demonstrate exceptional data efficiency in low-data scenarios. For instance, using just 1% of the labels for the ChestMNIST and HAM10000 datasets, it achieved the top position with accuracies of 52.65% and 62.71%, respectively. As the data ratio increases to 10%, its superiority in the ophthalmology modality becomes prominent, achieving first place on both OCTMNIST and RetinaMNIST with 78.30% and 57.75% accuracy. When using the full 100% dataset, the model sustains its leading performance, achieving the top rank on several datasets, including RSNA, BreastMNIST, and HAM10000, with respective accuracies of 84.11%, 86.54%, and 77.89%. These leading results across diverse modalities, spanning X-ray, dermatoscopy, ophthalmology, and ultrasound, provide compelling evidence that the representations learned by PhenoLIP are highly effective and robust in low-data regimes and high-data scenarios, indicating excellent generalization capabilities.

6.4 Ablation Studies

To demonstrate the impact of each component of our method, we conduct a series of ablation studies with the cross-modal retrieval task on PhenoBench for: (*i*) different encoder initializations, (*ii*) our proposed knowledge distillation loss, and (*iii*) data curation strategies. As shown in Table 5, we can draw the following key observations (detailed discussion is listed in Appendix 18):

- **Encoder initialization:** Domain-specific pre-trained encoders, such as BiomedCLIP and PubmedBERT, serve as stronger baselines, confirming the value of biomedical-informed representations.
- **Knowledge distillation loss:** The proposed $\mathcal{L}_{\text{KD}}$ contributes significantly to performance gains. By aligning the VLM's text encoder with the frozen phenotype knowledge encoder, it effectively injects structured ontological knowledge into the model.

- **Data curation strategies:** Our fine-grained data curation pipeline, including subfigure detection and LLM-based caption refinement, consistently improves model performance. These steps reduce noise from compound figures and enhance alignment between images and text, leading to more robust vision-language representations.
- **Knowledge graph components:** We further evaluate the contribution of different components (definitions, synonyms, and relational structure) in the phenotype knowledge encoder, detailed in Appendix 18.1. The results show that both semantic definitions and hierarchical graph topology significantly contribute to the overall vision-language pretraining.

7 Conclusion

In this work, we presented **PhenoLIP**, a phenotype-ontology-aware vision–language pretraining framework that integrates structured phenotype knowledge into medical VLMs. Built upon our newly curated large-scale multimodal knowledge graph **PhenoKG** and evaluated with the expert-verified benchmark **PhenoBench**, PhenoLIP achieves substantial gains in phenotype recognition, diagnostic classification, and retrieval tasks. These results highlight the importance of combining medical VLMs with phenotype-centric knowledge to achieve more accurate and interpretable medical image understanding.

Acknowledgements.

This work was supported in part by Shanghai Special Fund for Promoting High-Quality Industrial Development: Pilot Industry Innovation and Development Program (Artificial Intelligence Track, ID: 2025-GZL-RGZN-02020).

References

1. Amberger, J.S., Bocchini, C.A., Schiettecatte, F., Scott, A.F., Hamosh, A.: Omim.org: Online mendelian inheritance in man (omim®), an online catalog of human genes and genetic disorders. Nucleic Acids Research **43**(D1), D789–D798 (2015)
2. Baghbanzadeh, N., Ashkezari, S., Dolatabadi, E., Afkanpour, A.: Open-pmc-18m: A high-fidelity large scale medical dataset for multimodal representation learning. arXiv preprint arXiv:2506.02738 (2025)
3. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
4. Bodenreider, O.: The unified medical language system (umls): integrating biomedical terminology. Nucleic Acids Research **32**(suppl_1), D267–D270 (2004)
5. Borkowski, A.A., Bui, M.M., Thomas, L.B., Wilson, C.P., DeLand, L.A., Mastorides, S.M.: Lung and colon cancer histopathological image dataset (lc25000). arXiv preprint arXiv:1912.12142 (2019)
6. Chandak, P., Huang, K., Zitnik, M.: Building a knowledge graph to enable precision medicine. Scientific Data **10**(1), 67 (2023)
7. Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L., Jitsev, J.: Reproducible scaling laws for contrastive language-image learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2818–2829 (2023)
8. Fan, Z., Liang, C., Wu, C., Zhang, Y., Wang, Y., Xie, W.: Advancing radiology foundation models with reasoning through step-by-step verification from daily reports. Communications Medicine (2026)
9. Gargano, M.A., Matentzoglu, N., Coleman, B., Addo-Lartey, E.B., Anagnostopoulos, A.V., Anderton, J., Avillach, P., Bagley, A.M., Bakštein, E., Balhoff, J.P., et al.: The human phenotype ontology in 2024: phenotypes around the world. Nucleic Acids Research **52**(D1), D1333–D1346 (2024)
10. 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)
11. Gurovich, Y., Hanani, Y., Bar, O., Nadav, G., Fleischer, N., Gelbman, D., Basel-Salmon, L., Krawitz, P.M., Kamphausen, S.B., Zenker, M., et al.: Identifying facial phenotypes of genetic disorders using deep learning. Nature Medicine **25**(1), 60–64 (2019)
12. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
13. Ikezogwo, W., Seyfioglu, S., Ghezloo, F., Geva, D., Sheikh Mohammed, F., Anand, P.K., Krishna, R., Shapiro, L.: Quilt-1m: One million image-text pairs for histopathology. Advances in Neural Information Processing Systems **36**, 37995–38017 (2023)
14. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al.: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 590–597 (2019)
15. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day. Advances in Neural Information Processing Systems **36**, 28541–28564 (2023)

16. 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: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 525–536. Springer (2023)
17. Liu, S., Li, F., Zhang, H., Yang, X., Qi, X., Su, H., Zhu, J., Zhang, L.: Dab-detr: Dynamic anchor boxes are better queries for detr. In: International Conference on Learning Representations (2022)
18. Lozano, A., Sun, M.W., Burgess, J., Chen, L., Nirschl, J.J., Gu, J., Lopez, I., Aklilu, J., Rau, A., Katzer, A.W., 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)
19. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al.: Scikit-learn: Machine learning in python. The Journal of Machine Learning Research **12**, 2825–2830 (2011)
20. Qiu, P., Wu, C., Liu, J., Zheng, Q., Liao, Y., Wang, H., Yue, Y., Fan, Q., Zhen, S., Wang, J., et al.: Evolving diagnostic agents in a virtual clinical environment. arXiv preprint arXiv:2510.24654 (2025)
21. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763. PmLR (2021)
22. Roberts, R.J.: Pubmed central: The genbank of the published literature (2001)
23. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. Advances in Neural Information Processing Systems **35**, 25278–25294 (2022)
24. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025)
25. Tschandl, P., Rosendahl, C., Kittler, H.: The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific Data **5**(1), 1–9 (2018)
26. Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., et al.: Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786 (2025)
27. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2097–2106 (2017)
28. Weinreich, S.S., Mangon, R., Sikkens, J., Teeuw, M.E., Cornel, M.: Orphanet: a european database for rare diseases. Nederlands Tijdschrift voor Geneeskunde **152**(9), 518–519 (2008)
29. Wu, C., Zhang, X., Zhang, Y., Hui, H., Wang, Y., Xie, W.: Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. Nature Communications **16**(1), 7866 (2025)
30. Wu, C., Zhang, X., Zhang, Y., Wang, Y., Xie, W.: Medklip: Medical knowledge enhanced language-image pre-training for x-ray diagnosis. In: Proceedings of the

IEEE/CVF International Conference on Computer Vision (ICCV). pp. 21372–21383 (October 2023)
31. Yan, S., Hu, M., Jiang, Y., Li, X., Fei, H., Tschandl, P., Kittler, H., Ge, Z.: Derm1m: A million-scale vision-language dataset aligned with clinical ontology knowledge for dermatology. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 12681–12690 (October 2025)
32. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
33. Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H., Ni, B.: Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. Scientific Data **10**(1), 41 (2023)
34. Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., Wu, Y.: Coca: Contrastive captioners are image-text foundation models. Transactions on Machine Learning Research (2022)
35. Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., et al.: A multimodal biomedical foundation model trained from fifteen million image–text pairs. NEJM AI **2**(1), AIoa2400640 (2025)
36. Zhang, X., Wu, C., Zhang, Y., Xie, W., Wang, Y.: Knowledge-enhanced visual-language pre-training on chest radiology images. Nature Communications **14**(1), 4542 (2023)
37. Zhang, Y., Sui, X., Pan, F., Yu, K., Li, K., Tian, S., Erdengasileng, A., Han, Q., Wang, W., Wang, J., et al.: A comprehensive large-scale biomedical knowledge graph for ai-powered data-driven biomedical research. Nature Machine Intelligence pp. 1–13 (2025)
38. Zhou, X., Sun, L., He, D., Guan, W., Wang, G., Wang, R., Wang, L., Yuan, X., Sun, X., Zhang, Y., et al.: Knowledge-enhanced pretraining for vision-language pathology foundation model on cancer diagnosis. Cancer Cell **44**(4), 777–791 (2026)
39. Zhou, X., Zhang, X., Wu, C., Zhang, Y., Xie, W., Wang, Y.: Knowledge-enhanced visual-language pretraining for computational pathology. In: European Conference on Computer Vision. pp. 345–362. Springer (2024)