Derm1M: A Million-scale Vision-Language Dataset Aligned with Clinical Ontology Knowledge for Dermatology
–Supplemental Material–

Siyuan Yan[1*] Ming Hu[1*] Yiwen Jiang[1*] Xieji Li[1]
Hao Fei[2] Philipp Tschandl[3] Harald Kittler[3] Zongyuan Ge[1](✉)
[1] Monash University [2] National University of Singapore [3] Medical University of Vienna
{siyuan.yan,zongyuan.ge}@monash.edu

Abstract

In this supplementary material, we present the following contents: (1) more detailed data curation pipelines. (2) additional dataset statistics. (3) downstream dataset details. (4) additional ablation studies, and (5) additional implementation details.

1. More Detailed Data Curation Pipelines

1.1. YouTube Video

The main curation pipeline for YouTube is illustrated in Fig. 1 and Fig. 2 and detailed below.

Collecting Representative Channels and Videos. We constructed a comprehensive list of over 355 dermatology-related terms by consulting relevant literature and publicly available datasets. These terms encompass skin disease-related concepts, common names for various skin conditions, and associated synonyms. For each keyword, we retrieved the top 200 recommended videos from search queries. Additionally, based on our empirical observation that channel-based searches yield more focused and higher-quality explanatory videos compared to keyword searches, we manually curated 50 YouTube channels dedicated to dermatology explanations and downloaded their videos. During the downloading process, we prioritized the highest-resolution version of each video while filtering out videos shorter than 30 seconds or with a resolution below 224p. In total, we collected approximately 51k videos.

Filtering for Narrative-Style Videos. We assessed each video to determine: (1) whether it contains a sufficient number of usable dermatological images, and (2) whether it qualifies as a narrated video with rich explanatory voiceovers.

*Equal contribution

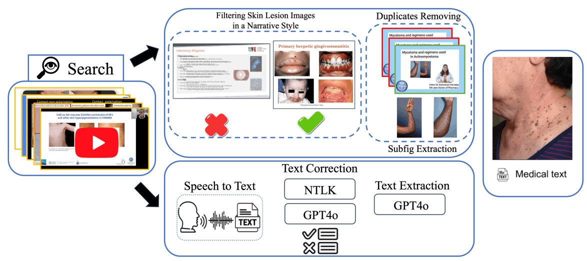

Figure 1. **Curation pipeline for YouTube content.** Our process begins with searching and collecting 51k videos from educational channels, followed by filtering to identify narrative-style content with high-quality explanations. We then extract and denoise text using a combination of speech-to-text models, handcrafted algorithms, and LLMs. Finally, we align the processed text with corresponding image pairs to create a curated dataset.

For criterion (1), we employed keyframe extraction with a predefined threshold to ensure a minimum level of visual change required for keyframe selection. For newly acquired videos, we extracted keyframes using FFmpeg by computing inter-frame color histogram differences. The threshold was determined via linear interpolation between 0.008 for 5-minute videos and 0.25 for 200-minute videos. We then trained and applied a DenseNet121 image classifier to identify keyframes containing dermatological images. Videos where more than 50% of keyframes were classified as containing dermatological content were labeled as valid.

For criterion (2), we utilized inaSpeechSegmenter to estimate the proportion of human speech within each video, setting a threshold of 0.2. Videos falling below this threshold were marked as silent or lacking sufficient explanatory narration.

Text Extraction and Denoising. To address the challenges of automatic speech recognition (ASR) for medical terminology in YouTube subtitles [2], we employed the large-

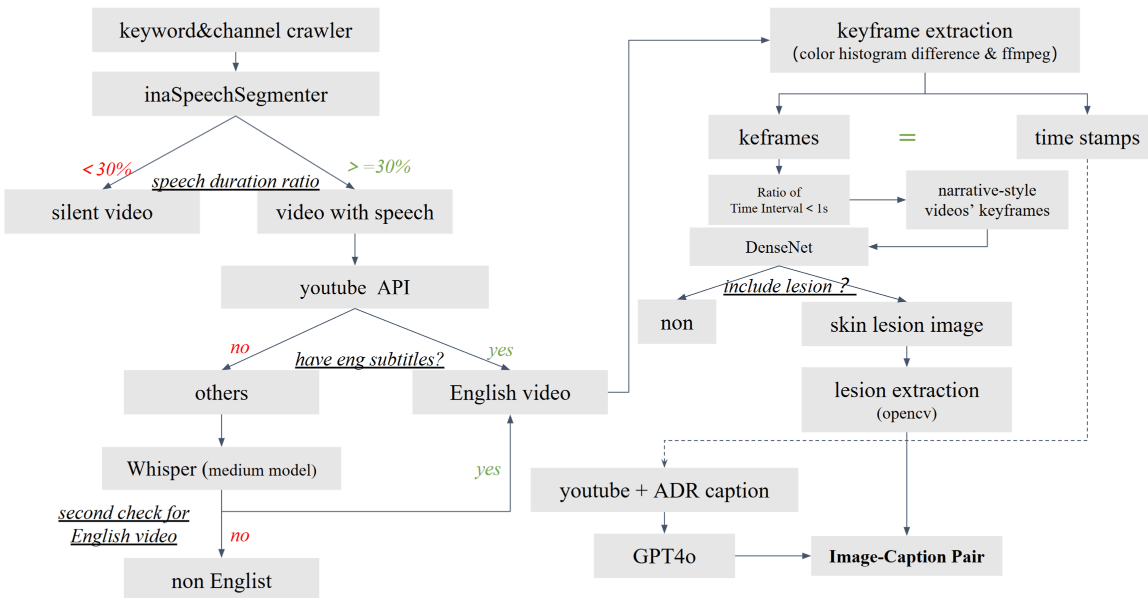

Figure 2. Flow chart of the curation pipeline for YouTube content.

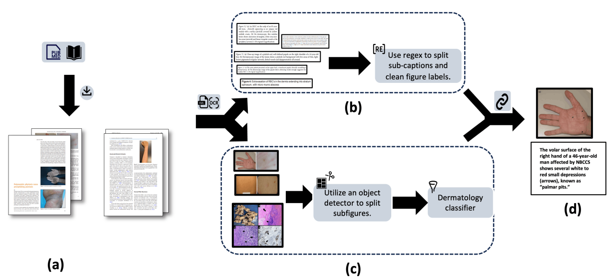

Figure 3. **Curation Pipeline for PubMed Source.** The process involves: (a) Searching and downloading articles from PubMed Open Access using pre-defined dermatology terms; (b) Extracting, splitting, and cleaning image captions; (c) Splitting subfigures using object detection and classifying skin images; and (d) Aligning image-text pairs to create the final dataset.

scale open-source Whisper Large-V3 model [6] to perform speech-to-text conversion by directly transcribing entire audio segments. We then developed a transcription denoising and quality control pipeline consisting of three key steps:

(i) Applying the RAKE algorithm to identify key phrases (up to four words) and optimizing them by removing stop words using NLTK;

(ii) Utilizing GPT-4o to verify and correct each entry, correcting transcription errors, and refining the alignment of complete descriptive statements;

(iii) Prompting a language model to generate a structured summary of the subtitles for improved readability and organization.

Aligning Image and Text Pairs. To achieve precise alignment between images and their corresponding text, we defined the timestamp interval between consecutive keyframes as an image chunk and treated each transcribed sentence as a text chunk. When the temporal overlap between an image chunk and a text chunk exceeded 50%, we considered them a matched pair. Note that one image chunk may map to multiple text chunks, in which case we concatenated these text chunks into a single longer description.

For image chunks without any matching text chunks (i.e., those with zero temporal overlaps), we observed they typically depicted content similar to previous frames that already had mapped text (e.g., continuous discussion of the same type of lesion). Consequently, we aligned such unmatched chunks to the most recent preceding keyframe with an associated text description. To mitigate potential mismatches in fine-grained details, we additionally filtered out image-specific details from the textual content, preserving only high-level descriptive information.

1.2. PubMed

Following [3, 4], the main curation pipeline for PubMed OA is illustrated in Fig. 3 and detailed below.

Collecting Image-text Pairs. We retrieved dermatology-related articles published between 1990 and 2024 from the PMC Open Access Subset using 356 domain-specific terms. This query yielded 566,571 articles with approximately 3.6 million images. To filter out non-dermatology-related figures (e.g., diagrams, flow charts, cartoon illustrations, and X-rays), we implemented a combination of clustering and manual inspection. After filtering, we matched the selected images with their captions from the provided XML format files to construct approximately 50K dermatology-focused image-text pairs.

Filtering Process. We used EfficientNetV2-S for feature extraction and applied PCA to reduce feature dimensions to 50. Using these features, we performed hierarchical K-means clustering, first grouping images into 20 major clusters, each further divided into 20 subclusters. We manually inspected 50 representative images per subcluster, iteratively removing non-dermatology clusters over five rounds until only dermatology-related clusters remained.

1.3. Educational Material

Collecting Image-text Pairs from Educational Material. We curated image-text pairs from 68 materials using the Fitz optical scanning module from the PyMuPDF Python package. For each detected image, the nearest caption box containing figure-related patterns was automatically retrieved to form an image-text pair. When direct extraction from scanned PDFs was not feasible, Optical Character Recognition (OCR) converted them into a vectorized format before extraction.

Removing Non-Dermatology Images. To remove non-dermatology images from curated image-text pairs, we partitioned the curated image-text pairs into 20,000 image chunks for computational efficiency. EfficientNetV2-S served as a feature extractor, encoding image features that were subsequently reduced to a 50-dimensional space using Principal Component Analysis (PCA). Manual inspection and iterative non-dermatology cluster removal were performed three times to ensure the elimination of the most irrelevant images.

Subfigure Detection and Segmentation. For subfigure detection, we trained a DINO [12] object detector using the MMDetection framework on 1,072 training images and 213 validation images sampled for each data source. The trained model then detected all subfigures, systematically cropping and arranging them in a structured left-to-right, top-to-bottom order.

Subcaption Detection and Image-text Pairing. Subcaptions were extracted using regular expressions that identified common subfigure markers (e.g., A) and (a)), facilitating automated detection and segmentation. Each subfigure was matched sequentially with its corresponding subcaption. If discrepancies arose between the number of subfigures and subcaptions, the original images and captions remained intact to preserve data integrity.

Automated Filtering of Non-Dermatology Subfigures. To further remove non-dermatology subfigures, we trained a DenseNet-121 classifier on a manually annotated dataset of 2,200 images sourced from educational materials (2,000 dermatology, 200 non-dermatology). Using a weighted random sampler and the Adam optimizer, we trained with a batch size of 128 and a learning rate of 9e-3. Early stopping was applied, halting training if validation AUROC showed no improvement after 22 epochs. By applying this classifier, we effectively excluded non-dermatology images from our dataset, ultimately ensuring a high-quality collection of image-text pairs sourced from educational materials.

1.4. Medical Forums

Extracting Image-Text Pairs from Twitter Posts. We began by manually reviewing content associated with 58 dermatology-related keywords to identify highly relevant channels. Through this process, we curated 27 dermatology channels comprising 14,099 posts, including both the tweet content and the three most-liked replies under each tweet. To ensure the dataset contained high-quality dermatology images, we applied the classifier mentioned in the educational materials-curated pipeline, refining the dataset to 6,726 images.

Text Cleaning and Processing. The accompanying text underwent extensive cleaning, removing @usernames, hashtags (‘#’), newline (‘$\backslash n$’) and carriage return (‘$\backslash r$’) symbols, HTML links, URLs, bold and italicized text, and other invalid characters. Additionally, all sentences containing question marks or beginning with “What is” were eliminated to enhance textual clarity.

Manual Removal of Advertisements. Further refinement was carried out through manual inspection, leading to the removal of advertisement-related tweets, reducing the dataset to 6,532 image-text pairs.

Text Standardization. To standardize the text, we reconstructed it by concatenating the original tweet content with its longest reply. Finally, image-text pairs containing fewer than three words were discarded, resulting in a final dataset of 6,431 high-quality image-text pairs. For other medical forums such as IIYI and Reddit, we followed similar workflows.

1.5. Public Dataset

Additionally, we created handcrafted image-text pairs using the publicly available SCIN [9] and MSKCC [1] datasets. For the MSKCC dataset, we generated text descriptions by integrating anatomic site, lesion type, and diagnosis results into a structured template, yielding 10,619 image-text pairs. Similarly, for the SCIN dataset, we constructed text descriptions by incorporating image modality, skin tone, age, gender, skin texture, symptoms, and diagnosis into a handcrafted template, resulting in 6,518 image-text pairs.

1.6. Initial Text Processing and Quality Control

General Processing. As shown in Fig. 4, we processed non-forum text through language detection, information block filtering, and abbreviation expansion. Non-English text was identified using SpaCy and translated into English using GPT. For data sourced from PubMed and educational

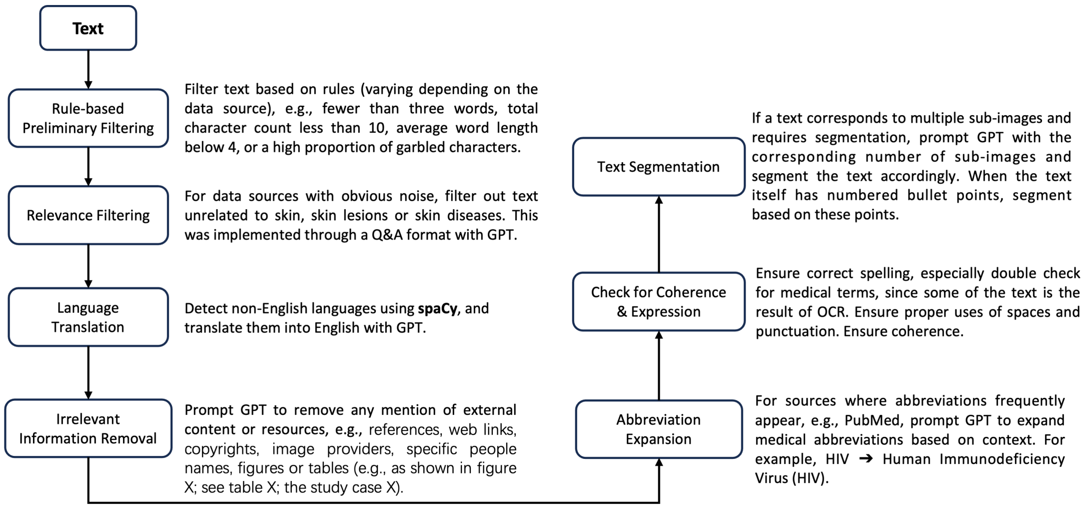

Figure 4. Workflow for general text processing.

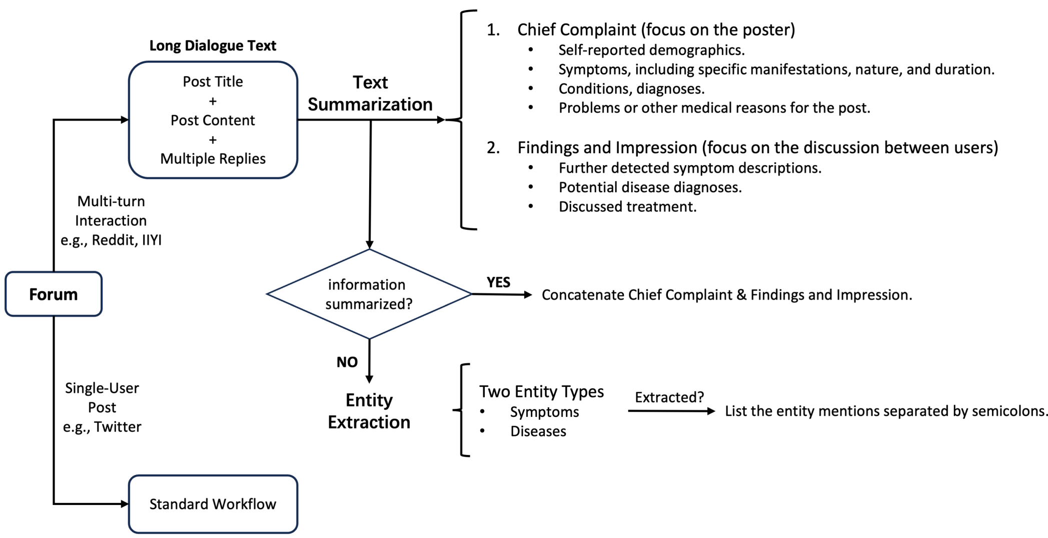

Figure 5. Workflow for forum-specific text processing.

materials, we instructed GPT to filter out non-medical information blocks, such as citations, figure references, hyperlinks, copyright statements, and personal names. For knowledge-dense texts, GPT recognized and expanded abbreviations contextually (e.g., HIV $\rightarrow$ Human Immunodeficiency Virus (HIV)).

Forum-Specific Processing. As illustrated in Fig. 5, for multi-turn discussions in medical forums, we extracted structured summaries, including: (1) the poster's chief complaint (demographics, symptoms, diagnosis, and medical intent) and (2) clinical findings and impressions from replies (symptom elaborations, differential diagnostic discussions, and treatment suggestions). These components were concatenated to form the final captions. If summarization failed, we instead extracted symptom and disease entities, connecting them with semicolons.

Quality Control. We applied rule-based filtering specific to data sources, including but not limited to discarding captions with fewer than three words or ten characters, as well as those consisting solely of garbled text. For noisy sources, we instructed GPT to evaluate dermatological relevance using a QA-based approach. OCR-derived text underwent spell-checking, punctuation correction, and coherence refinement. This ensured captions were precise, standardized, and dermatology-focused.

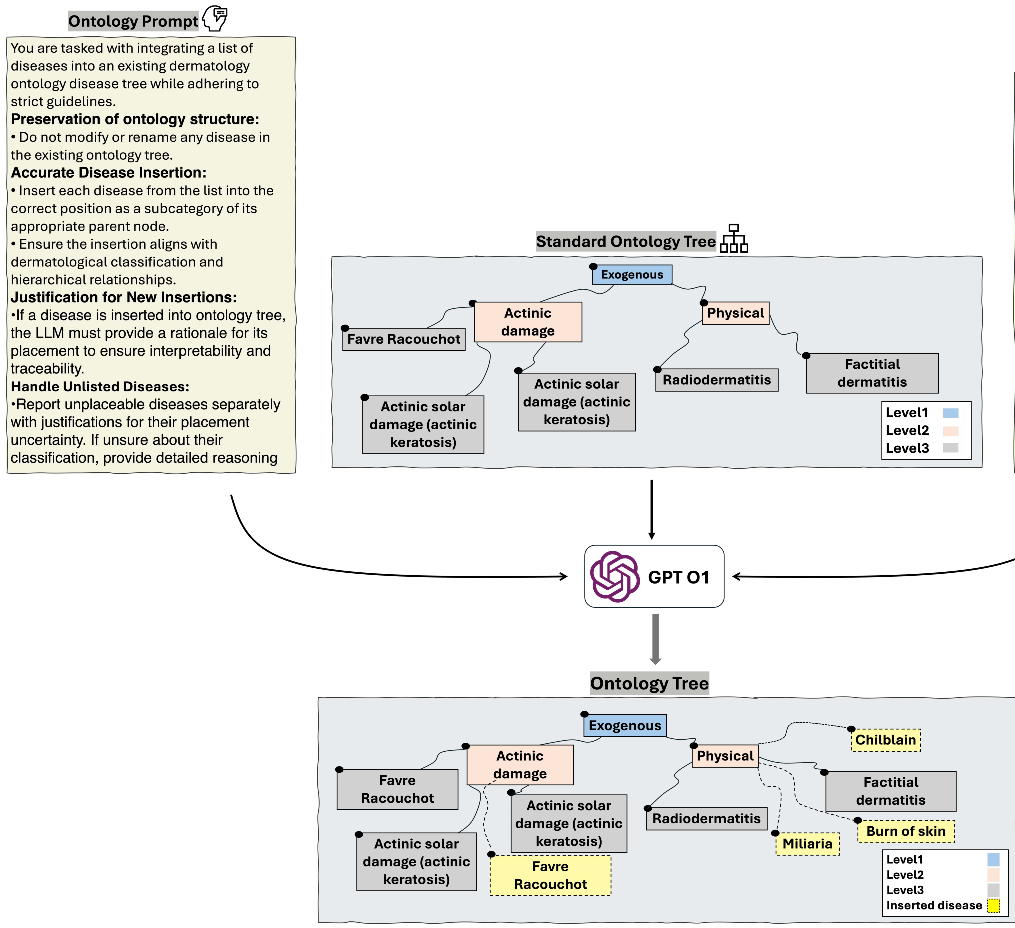

Figure 6. **Example of ontology tree construction.** A pipeline for developing a comprehensive ontology tree from a standard expert-constructed ontology tree. The LLM is provided with the standard ontology tree pre-defined by medical experts, the ontology prompt, and the standardized disease list, ensuring the original ontology structure is maintained while accurately inserting diseases into the correct hierarchical positions.

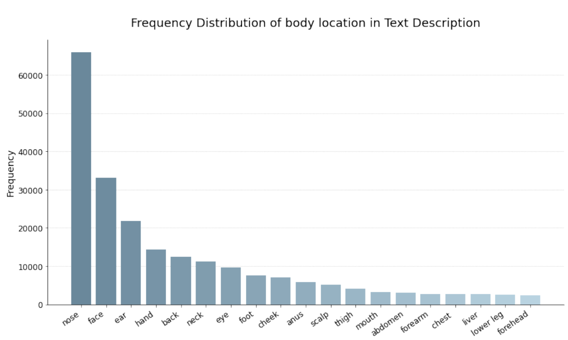

Figure 7. Frequency distribution of body sites.

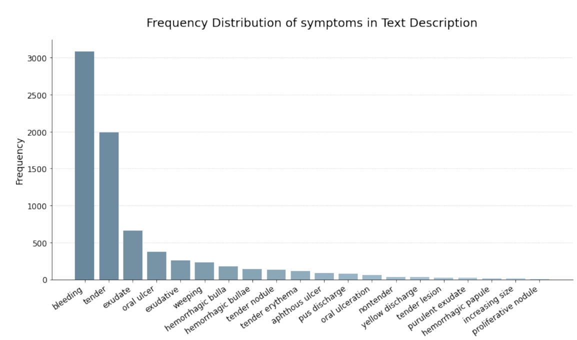

Figure 8. Frequency distribution of symptoms.

1.7. Ontology Knowledge Augmentation

1.7.1. Standardized Disease and Clinical Concept

For all data sources, the LLM-extracted content was often noisy and unformatted. We standardized diseases and concepts to avoid medical term ambiguity issues.

Construction of the Standardized Disease List. We constructed a standardized disease list by compiling disease labels from the F17k, SD128, SNU134, SCIN, and HAM

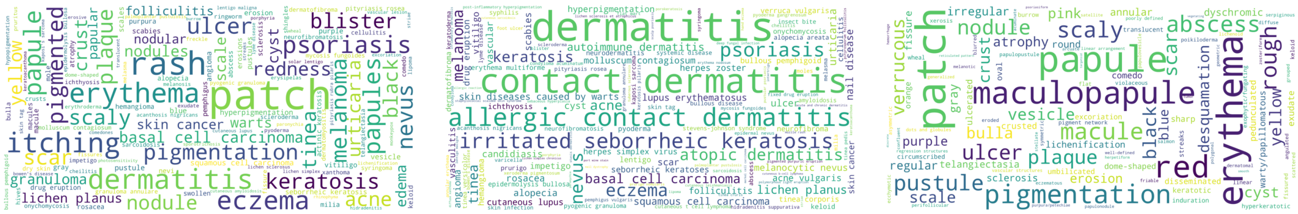

Figure 9. Word cloud of medical terms: medical term (left), diseases (middle), and clinical concept (right).

datasets. Additionally, we leveraged LLM to automatically identify and merge diseases with identical content but different names, ensuring consistency and reducing redundancy. This process resulted in a standardized disease list containing 407 unique disease labels.

Construction of Clinical Concept List. We established a standardized concept list by compiling labels from public skin condition datasets, including Derm7pt and SkinCon. To further expand this list, we prompted the LLM to generate additional skin-related visual concepts based on key dermatological categories: Basic Morphology, Secondary Changes, Basic Colors, Color Characteristics, Shape Characteristics, Surface Features, Distribution Patterns, Border Characteristics, and Special Morphology. We conducted this process multiple times and manually removed any unrelated concepts. This resulted in a standardized clinical concept list containing 130 unique clinical concept labels as shown in Table 5.

Alignment between LLM-extracted contents and standardized lists. To align LLM-extracted content with standardized lists, we implemented two distinct pipelines for disease and clinical concepts:

1) Standardized Disease List Alignment: We constructed a mapping framework using a Word2Vec-based approach, employing BioMedBERT as a word encoder to transform LLM-generated disease names and the standardized disease list into vector representations. To ensure accurate mappings, we iterated through the LLM-extracted content list, computing similarity scores against the standardized disease list. If the highest similarity score exceeded 0.7, the LLM-generated disease name was mapped to its corresponding standardized term. This process successfully aligned LLM-generated content with 390 unique standardized diseases as shown in Table 5, bridging the connection between downstream classification datasets and pretrained image-text pairs.

2) Standardized Clinical Concept List Alignment: We used two methods to match LLM-extracted concepts with the standardized concept list. First, the Substring Matching Algorithm identified overlapping terms, successfully aligning most LLM-extracted concepts with standardized clinical concepts (e.g., "erythematous-violaceous macule" mapped to "erythematous," "violaceous," and "macule"). Second, for the remaining unmatched concepts, we employed LLM-assisted alignment, providing the LLM with both lists to iteratively refine matches through multi-turn dialogues. This process enabled the alignment of LLM-extracted concepts with 130 standardized clinical concepts.

1.7.2. Ontology Construction and Augmentation

To construct a dermatology ontology tree, we built upon an initial standard ontology tree (Fig.1e) curated by four dermatology experts, encompassing 128 dermatological diseases from the SD128[7] dataset. We then utilized a specialized ontology prompt strategy that enabled the LLM to systematically integrate diseases from the standardized disease list into the ontology structure while maintaining hierarchical integrity.

The ontology construction follows four key principles: 1) *Preservation of the standard ontology structure* – The LLM must retain the original hierarchy and avoid modifying the positions of existing nodes. 2) *Accurate disease insertion* – Each disease from the standardized disease list must be correctly placed, considering its hierarchical relationship with existing nodes in the ontology tree. 3) *Justification for new insertions* – If a disease is inserted into the ontology tree, the LLM must provide a rationale for its placement to ensure interpretability and traceability. 4) *Handling uncertain classifications* – If the LLM is unsure of a disease's placement, it defers the decision by adding it to a separate list with an accompanying explanation.

LLM-driven Ontology Integration. As shown in Fig. 6, we provided the LLM with three key inputs: the standard ontology tree, the ontology prompt, and the standardized disease list. The LLM then automatically integrated diseases from the standardized disease list into the ontology tree, generating a refined structure that captured rich and diverse hierarchical relationships. For instance, Miliaria was correctly inserted as a child node under Physical and Exogenous conditions. To ensure stability and consistency, we repeated this automatic LLM-driven integration for five iterations, refining the ontology tree through iterative manual adjustments and validation. As a result, we successfully constructed an ontology tree comprising 371 skin disease conditions, while 19 general diseases remained unplaced due to the LLM's uncertainty regarding their classification. This structured methodology ensured that ontology tree development remained systematic, transparent, and aligned with expert-defined standards, while effectively leveraging the LLM's capabilities for hierarchical reasoning and disease classification.

Ontology Caption Construction. Once the standardized disease list was integrated, we used the augmented ontology tree to retrieve all parent nodes of each disease, generating hierarchical disease paths (e.g., folliculitis: inflammatory → infectious → bacterial → folliculitis). We then transformed the hierarchy into ontology-augmented captions using a series of predefined templates, such as 'This is a skin photo diagnosed as {inflammatory, infectious, bacterial, folliculitis}.' This approach ensured that ontology captions accurately represented hierarchical relationships within the ontology tree, providing a structured and standardized description of dermatological conditions.

Knowledge Augmentation Caption Construction. Finally, the knowledge-augmented caption was constructed by appending the ontology caption and clinical concept caption to the end of the original caption. Similar to ontology caption construction, the clinical concept captions were generated using a handcrafted template: "This is a skin photo showing {concept_a, concept_b, concept_c}."

2. Additional Dataset Statistics

Fig. 7 and 8 illustrate the frequency distribution of anatomical locations and symptoms in Derm1M. The analysis reveals that skin conditions predominantly manifest on the face, nose, and ears, while common symptoms include bleeding and tenderness. These distributions offer valuable insights into prevalence patterns within the dataset. Additionally, Fig. 9 displays word clouds highlighting frequent terms across three categories: medical terminology, dermatological conditions, and clinical concepts. Fig. **??**–**??** showcase representative image-text pairs from the Derm1M dataset. Table 5 and 6 show the complete list of the 390 skin conditions and 130 clinical concepts covered in Derm1M.

3. Downstream Dataset Details

Daffodil: This dataset is distinguished by its comprehensive collection of 9,548 dermatoscopic images across five skin conditions (acne, vitiligo, hyperpigmentation, nail psoriasis, and SJS-TEN), offering a valuable resource for non-melanoma skin disease classification that complements existing skin cancer-focused datasets like ISIC2019 [1] and HAM10000 [8].

4. Additional Ablation Studies

We explore the performance differences between training methods on the Derm1M dataset, comparing SigLIP [11], CoCa [10], and CLIP [5]. Tables 1–3 present downstream performance across various tasks. CLIP consistently outperforms on zero-shot disease classification and few-shot/full-shot linear evaluation, achieving the highest accuracy in most settings. However, SigLIP and CoCa demonstrate superior performance on cross-modal retrieval tasks.

5. Additional Implementation Details

Training Details. We pretrain a series of models called DermLIP on the Derm1M dataset following CLIP [5]'s contrastive learning objective. Each model is trained for 30 epochs on a single NVIDIA H200 GPU. We swap hyper-parameters including batch size and learning rate, selecting the best-performing models based on validation loss.

Prompt Details for Zero-shot Classification We adhere to the zero-shot classification method of the OpenCLIP framework, utilizing a prompt ensemble strategy for evaluation. The specific prompt templates employed in this process are detailed in Table 7.

Hyper-parameter tables for main models We present the pre-training hyper-parameters for the DermLIP models in Table 8. The table includes all critical training hyper-parameters, while the remaining parameters adhere to the default settings of the OpenCLIP framework.

Training methods	Pretrained Data	Vision Enc.	Text Enc.	HAM	F17K	PAD	Daffodil	Average
#class				7	113	6	5	
SigLIP	Derm1M	ViT-B16	SigLIP	0.6068	0.2249	0.5857	0.7058	0.5308
CoCa	Derm1M	ViT-B32	GPT77	0.4098	0.1700	0.5466	**0.7262**	0.4632
CLIP	Derm1M	ViT-B16	GPT77	**0.6820**	**0.2278**	**0.6074**	0.7257	**0.5607**

Table 1. Ablation on different training methods for zero-shot disease classification (Acc).

Labeling Ratio	Methods	Vision Enc.	Text Enc.	HAM	F17K	PAD	Daffodil	Average
#class				7	113	6	5	
1%	SigLIP	ViT-B16	SigLIP	0.6986	0.1394	0.5098	0.7476	0.5239
	CoCa	ViT-B32	GPT77	0.7212	0.1349	0.5076	0.7974	0.5403
	CLIP	ViT-B16	GPT77	**0.7458**	**0.1602**	**0.5184**	**0.8545**	**0.5697**
10%	SigLIP	ViT-B16	SigLIP	0.8037	0.2980	0.6312	0.8759	0.6522
	CoCa	ViT-B32	GPT77	0.7532	0.2967	0.6551	0.8681	0.6433
	CLIP	ViT-B16	GPT77	**0.8110**	**0.3555**	**0.6594**	**0.9372**	**0.6908**
100%	SigLIP	ViT-B16	SigLIP	**0.8550**	0.4433	0.6703	0.9330	0.7254
	CoCa	ViT-B32	GPT77	0.7591	0.4933	0.7115	0.8743	0.7096
	CLIP	ViT-B16	GPT77	0.8523	**0.5102**	**0.7614**	**0.9644**	**0.7720**

Table 2. Ablation on different training methods for linear evaluation (Acc).

Training methods	Vision Enc.	Text Enc.	Holdout (n=9806)				SkinCAP (n=3989)			
			I2T (%)		T2I (%)		I2T (%)		T2I (%)	
			R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50
SigLIP	ViT-b16	SigLIP	0.3763	0.5614	0.3818	0.5716	**0.1860**	**0.3896**	**0.1908**	**0.4141**
CoCa	ViT-B32	GPT77	**0.4150**	**0.6102**	**0.4182**	**0.6116**	0.1564	0.3643	0.1737	0.3818
CLIP	ViT-b16	GPT77	0.4069	0.6021	0.3966	0.5992	0.1567	0.3632	0.1594	0.3567

Table 3. Ablation on different training methods for cross-modal retrieval results. I2T represents image-to-text retrieval and T2I represents text-to-image retrieval.

Method	**Encoder**		**AUROC**		
	Vision	**Text**	**SkinCon (32)**	**Derm7pt (7)**	**Average**
CLIP-B16 [5]	ViT-B16	GPT77	0.6643	0.5594	0.6119
SigLIP [11]	ViT-B16	SigLIP	0.6769	0.5631	0.6200
CoCa [10]	ViT-B32	GPT77	0.6041	0.5677	0.5859
PMC-CLIP [4]	ResNet50	GPT77	0.6251	0.5820	0.6036
BiomedCLIP [13]	ViT-B16	PMB256	0.6817	0.6092	0.6455
MONET [3]	ViT-L14	GPT77	0.7502	**0.6889**	0.7196
DermLIP	ViT-B16	GPT77	**0.7728**	0.6877	**0.7303**
DermLIP	PanDerm-B	PMB256	0.7299	0.6148	0.6724

Table 4. Zero-shot concept annotation (AUROC).

A-E	F-M	N-R	S-Z
abscess	fissure	necrosis	salmon
acuminate	fissured	nodule	satellite
angulated	flat	orange	scale
annular	flat topped	oval	scaly
arciform arrangement	follicular-centered	papule	scar
arcuate	friable	papulonodule	scattered
asymmetric	generalized	papulopustule	sclerosis
atrophy	geometric	papulovesicle	serpiginous
black	gray	patch	sharp
blue	grouped	pedunculated	smooth
blue whitish veil	hemorrhage	perifollicular	stellate
blurred	herpetiform	pigment network	streaks
brown(hyperpigmentation)	hyperkeratotic	pigmentation	symmetric
bulla	induration	pink	targetoid
burrow	irregular	plaque	telangiectasia
circumscribed	keloid	poikiloderma	translucent
clustered	keratotic	polygonal	tumor
comedo	leaf-shaped	poorly defined	ulcer
confluent	lichenification	psoriasiform	ulcerated
crust	lichenoid	purple	umbilicated
crusted	linear	purpura/petechiae	vascular structures
cyst	linear arrangement	pustule	vegetating
dermatomal	localized	raised	verrucous
desquamation	macule	red	vesicle
diffuse	maculopapule	regression structures	violaceous
discrete	maculopatch	regular	warty/papillomatous
disseminated	melanotic	reticular	wedge-shaped
dome-shaped	molluscoid	reticulated pattern	well-defined
dots and globules		rough	wheal
dyschromic		round	white(hypopigmentation)
ecchymotic			xanthomatous
eczematous			xerosis
eroded			yellow
erosion			zosteriform
erythema			
excoriation			
exophytic/fungating			
exudate			

Table 5. Full List of 130 standardized clinical concepts.

A-E	F-M
abscess	factitial dermatitis
acanthosis nigricans	favre racouchot
acne	fibroma molle
acne keloidalis nuchae	fixed drug eruption
acne urticata	fixed eruptions
acne vulgaris	flat wart
acquired autoimmune bullous diseaseherpes gestationis	flushing
acrokeratosis verruciformis	follicular mucinosis
actinic granuloma	folliculitis
actinic solar damage(actinic keratosis)	foot ulcer
actinic solar damage(cutis rhomboidalis nuchae)	foreign body reaction of the skin
actinic solar damage(pigmentation)	fox-fordyce disease
actinic solar damage(solar elastosis)	freckle
actinic solar damage(solar purpura)	fungal dermatitis
actinic solar damage(telangiectasia)	fungal dermatosis
acute and chronic dermatitis	furuncle
acute constitutional eczema	geographic tongue
acute dermatitis	granulation tissue
acute dermatitis, nos	granuloma annulare
acute generalized exanthematous pustulosis	granuloma faciale
acute vesicular dermatitis	grover’s disease
adnexal neoplasm	guttate psoriasis
ageing skin	hailey hailey disease
allergic contact dermatitis	halo nevus
allergic reaction	hand eczema
alopecia	hand foot and mouth disease
alopecia areata	hemangioma
alopecia mucinosa	hematoma of skin
amyloidosis	hemosiderin pigmentation of lower limb due to varicose veins of lower limb
angiofibroma	hemosiderin pigmentation of skin due to venous insufficiency
angiokeratoma	herpes simplex virus
angioma	herpes zoster
angular cheilitis	hidradenitis suppurativa
animal bite - wound	histiocytosis of skin
annular erythema	hormonal acne
apocrine hydrocystoma	hyperkeratosis palmaris et plantaris
arsenical keratosis	hyperpigmentation
atopic dermatitis	hypersensitivity
atopic winter feet	hypertrichosis
autoimmune dermatitis	hypertrophic scar
basal cell carcinoma	ichthyosis
beau’s lines	idiopathic guttate hypomelanosis
becker nevus	impetigo
behcets disease	infantile atopic dermatitis
benign keratosis	infected eczema
blister	inflammatory dermatosis
blue nevus	insect bite
bowen’s disease	intertrigo
bullous disease	inverse psoriasis
bullous pemphigoid	irritant contact dermatitis
burn of forearm	irritated seborrheic keratosis (from ”sk/isk”)
burn of skin	junction nevus
café au lait macule	juvenile plantar dermatosis
calcinosis cutis	juvenile xanthogranuloma
callus	kaposi sarcoma
campbell de morgan spots	kaposi’s sarcoma of skin
candida intertrigo	keloid
candidiasis	keratoacanthoma
cellulitis	keratoderma
central centrifugal cicatricial alopecia	keratolysis exfoliativa of wende
cheilitis	keratosis
chilblain	keratosis pilaris
childhood bullous pemphigoid	keratosis pilaris rubra faciei
cholestasis of pregnancy	kerion
chondrodermatitis nodularis helicis	knuckle pads

chronic actinic dermatitis	koilonychia
chronic dermatitis, nos	langerhans cell histiocytosis
clubbing of fingers	leg veins
compound nevus	lentigo
condyloma	lentigo maligna
condyloma acuminatum	lentigo maligna melanoma
confluent and reticulated papillomatosis	leukocytoclastic vasculitis
congenital nevus	leukonychia
contact dermatitis	lichen amyloidosis
contact dermatitis caused by rhus diversiloba	lichen nitidus
contact dermatitis, nos	lichen planus
contact purpura	lichen sclerosis et atrophicus
crowe’s sign	lichen simplex chronicus
cutaneous b-cell lymphoma	lichen spinulosus
cutaneous horn	lichen striatus
cutaneous larva migrans	lipoma
cutaneous leishmaniasis	livedo reticularis
cutaneous lupus	local infection of wound
cutaneous sarcoidosis	localized cutaneous vasculitis
cutaneous t cell lymphoma	localized skin infection
cyst	lupus erythematosus
darier-white disease	lyme disease
dariers disease	lymphangioma
deep fungal infection	lymphocytic infiltrate of jessner
degos disease	majocchi granuloma
dermatitis	median nail dystrophy
dermatitis herpetiformis	medication-induced cutaneous pigmentation
dermatofibroma	melanin pigmentation due to exogenous substance
dermatosis papulosa nigra	melanocytic nevus
desquamation	melanoma
diffuse xanthoma	melasma
digital fibroma	merkel cell carcinoma
dilated pore of winer	metastatic carcinoma
discoid eczema	milia
disseminated actinic porokeratosis	miliaria
drug eruption	moles
drug eruptions & reactions	molluscum contagiosum
drug-induced pigmentary changes	morphea
dry skin	mucinosis
dyshidrosiform eczema	mucocele
dysplastic nevus	mucosal melanotic macule
ecthyma	muzzle rash
ecthyma gangrenosum	mycosis fungoides
eczema	myxoid cyst
eczema herpeticum	
ehlers danlos syndrome	
elephantiasis nostras	
epidermal nevus	
epidermoid cyst	
epidermolysis bullosa	
erosion of skin	
erosive pustular dermatosis of the scalp	
eruptive odontogenic cyst	
eruptive xanthoma	
erythema ab igne	
erythema annulare centrifugum	
erythema craquele	
erythema dyschromicum perstans	
erythema elevatum diutinum	
erythema gyratum repens	
erythema migrans	
erythema multiforme	
erythema nodosum	
exfoliative dermatitis	
exfoliative erythroderma	
N-R	**S-Z**
naevus comedonicus	sand-worm eruption

nail disease
nail dystrophy
nail psoriasis
necrobiosis lipoidica
nematode infection
neurodermatitis
neurofibroma
neurofibromatosis
neutrophilic dermatoses
nevus
nevus depigmentosus
nevus sebaceous of jadassohn
nevus spilus
no definitive diagnosis
nummular eczema
onycholysis
onychomycosis
onychoschizia
organoid nevus
ota nevus
others
palmoplantar pustulosis
palpable migrating erythema
papular dermatoses of pregnancy
parapsoriasis
paronychia
parvovirus b19 infection
pemphigus vulgaris
phototherapy
phytophotodermatitis
pigmentation of pregnancy
pigmented progressive purpuric dermatosis
pigmented purpuric eruption
pilar cyst
pincer nail deformity
pityriasis alba
pityriasis lichenoides
pityriasis lichenoides chronica
pityriasis lichenoides et varioliformis acuta
pityriasis rosea
pityriasis rubra pilaris
pityrosporum folliculitis
poikiloderma
poikiloderma of civatte
poisoning by nematocyst
polymorphic eruption of pregnancy
polymorphous light eruption
porokeratosis
porokeratosis of mibelli
poroma
porphyria
port wine stain
post-inflammatory hyperpigmentation
post-inflammatory hypopigmentation
post-inflammatory pigmentation
pressure ulcer
prurigo
prurigo gravidarum
prurigo nodularis
prurigo of pregnancy
prurigo pigmentosa
pruritic urticarial papules and plaques of pregnancy
pruritus ani
pseudo-glucagonoma syndrome
pseudofolliculitis barbae
pseudorhinophyma
psoriasis
sarcoidosis
scabies
scalp psoriasis
scar
scleroderma
scleromyxedema
sebaceous hyperplasia
seborrheic keratoses
sixth disease
skin and soft tissue atypical mycobacterial infection
skin cancer
skin diseases caused by warts
skin infection
skin lesion in drug addict
skin tag
spider veins
squamous cell carcinoma
staphylococcal scalded skin syndrome
stasis dermatitis
stasis edema
stasis ulcer
steatocystoma multiplex
steroid acne
steroid use abusemisuse dermatitis
stevens-johnson syndrome
strawberry birthmarks
striae
subungual hematoma
sun spots
sunburn
superficial gyrate erythema
superficial spreading melanoma ssm
superficial wound of body region
sweet syndrome
sweet’s syndrome
syphilis
syringoma
systemic disease
telangiectasia macularis eruptiva perstans
tick bite
tinea
tinea corporis
tinea cruris
tinea manus
tinea pedis
tinea versicolor
transient acantholytic dermatosis
traumatic blister
traumatic ulcer
tuberous sclerosis
tungiasis
ulcer
unilateral laterothoracic exanthem
urticaria
urticaria pigmentosa
urticarial vasculitis
varicella
varicose veins of lower extremity
vascular
vasculitis
venous lake
verruca vulgaris
viral exanthem
viral exanthems: roseola
vitiligo
wound/abrasion
xanthelasma

pustular psoriasis pyoderma pyoderma gangrenosum pyogenic granuloma radiodermatitis raynaud phenomenon red stretch marks relapsing polychondritis rheumatoid nodule rhinophyma riehl melanosis rosacea	xeroderma pigmentosum xerosis xerotic eczema

Table 6. Full list of 390 standardized skin conditions.

ID	Template
1	`This is a skin image of {CLASS_LABEL}.`
2	`This is a skin image of {CLASS_LABEL}.`
3	`A skin image of {CLASS_LABEL}.`
4	`An image of {CLASS_LABEL}, a skin condition.`
5	`{CLASS_LABEL}, a skin disorder, is shown in this image.`
6	`The skin lesion depicted is {CLASS_LABEL}.`
7	`The skin cancer in this image is {CLASS_LABEL}.`
8	`This image depicts {CLASS_LABEL}, a type of skin cancer.`

Table 7. Prompt templates for zero-shot classification.

Hyper-parameters	ViT-B16 + GPT77	PanDerm-B + PMB256	ViT-B16 + SigLIP	ViT-B32 + GPT77
warmup	1000	1000	1000	1000
weight decay	0.1	0.1	0.1	0.1
LR Scheduler	cosine	cosine	cosine	cosine
batch size	4096	2048	2048	512
learning rate	1e-4	1e-4	1e-4	1e-4
epochs	30	30	30	30
Pretrain	openai	PanDerm	webli	laion2b_s13b_b90k
Vision Encoder	ViT-B16	PanDerm-B	ViT-B16	ViT-B32
Text Encoder	GPT77	PMB256	SigLIP	GPT77

Table 8. Hyperparameters for DermLIP models pretraining.

References

[1] Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pages 168–172. IEEE, 2018. 3, 7

[2] Wisdom Ikezogwo, Saygin Seyfioglu, Fatemeh Ghezloo, Dylan Geva, Fatwir Sheikh Mohammed, Pavan Kumar Anand, Ranjay Krishna, and Linda Shapiro. Quilt-1m: One million image-text pairs for histopathology. *Advances in neural information processing systems*, 36, 2024. 1

[3] Chanwoo Kim, Soham U Gadgil, Alex J DeGrave, Jesutofunmi A Omiye, Zhuo Ran Cai, Roxana Daneshjou, and Su-In Lee. Transparent medical image ai via an image–text foundation model grounded in medical literature. *Nature Medicine*, pages 1–12, 2024. 2, 8

[4] 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*, pages 525–536. Springer, 2023. 2, 8

[5] Alec Radford, Jong Wook Kim, Chris Hallacy, and et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 7, 8

[6] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023. 2

[7] Xiaoxiao Sun, Jufeng Yang, Ming Sun, and Kai Wang. A benchmark for automatic visual classification of clinical skin disease images. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14*, pages 206–222. Springer, 2016. 6

[8] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9, 2018. 7

[9] Abbi Ward, Jimmy Li, Julie Wang, and et al. Creating an empirical dermatology dataset through crowdsourcing with web search advertisements. *JAMA Network Open*, 7(11): e2446615–e2446615, 2024. 3

[10] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 7, 8

[11] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023. 7, 8

[12] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection, 2022. 3

[13] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. A multimodal biomedical foundation model trained from fifteen million image–text pairs. *NEJM AI*, 2(1):AIoa2400640, 2025. 8