Exploiting Unpaired Dermatology Data for Multimodal Foundation Models via Transitive Cross-Modal Linking

Supplementary Material

6. Zero-shot classification: 15-classes setup

To assess semantic generalization beyond the training taxonomy, the 15-classes zero-shot classification is introduced, including fifteen diagnostic concepts that reflect a finer-grained stratification of skin lesion categories, including morphological subtypes and pre-malignant entities not explicitly represented during training: *actinic keratosis*, *basal cell carcinoma*, *benign melanocytic nevus*, *blue nevus*, *Bowen disease / squamous cell carcinoma in situ*, *congenital / special-pattern nevi*, *dermatofibroma / fibrous lesions*, *dysplastic / atypical nevus (Clark-type)*, *lentigo maligna*, *lichenoid keratosis*, *melanoma*, *seborrheic keratosis & pigmented keratoses / solar lentigo*, *squamous cell carcinoma (invasive)*, and *vascular lesions*. Unlike the 8-class setup, which match the diagnostic categories adopted during training, this configuration evaluates whether the shared latent space encodes clinically relevant semantic distinctions at a level of granularity that the model was not explicitly optimized to capture.

7. Synthetic note generation

The generation of clinical notes from images aims to pair a small dermatology image dataset with textual descriptions. This step is motivated by two considerations: pairing images with text allows the development of downstream MM or foundation models and the use of textual descriptions as prompts for image generation. Since this process involves LLMs, which may hallucinate, synthetic clinical notes are generated exploiting the metadata associated with the original images. Public dermatology datasets provide heterogeneous annotations that mainly encode lesion class information (e.g., melanocytic nevus, melanoma) and, in some cases, finer-grained subclass specifications that can be mapped to a shared medical terminology spanning the eight diagnostic categories considered in this study. Metadata are harmonized through mapping to a common vocabulary to mitigate inconsistencies across datasets. Although these annotations rarely describe lesion morphology, lesion type and subclass information can be used to populate a predefined clinical note template. In parallel, clinical descriptions are generated by prompting a LLM to improve the finding description with lesion attributes such as asymmetry, chromatic variation, and border morphology. Four LLMs are employed: GPT-4o-mini, MedGemma [53], SkinGPT4 [63], and DermLIP [59, 60]. GPT-4o-mini is a general-purpose LLM, MedGemma is specialized on medical data, and SkinGPT4 and DermLIP are specialized on dermatology data. Prompts are explicitly designed to produce dermatology-specific terminology including in the metadata, reducing hallucinations. During SD fine-tuning, clinical notes are adopted only if they contain dermatological terminology expressed within a predefined vocabulary, preventing the propagation of hallucinated content.

8. Single repository adopted in real dataset

Table 4 shows the composition of the dataset including real images, highlighting the single repositories composing it.

Table 4. Overview of the real dataset, showing the single repositories adopted.

Dataset	BEK	DF	NEV	VASC	ACK	BCC	MEL	SCC	Total
Training partition									
BCN20000 (D)	796	86	2,944	77	515	1,966	1,999	391	8,774
Derm12345 (D)	456	128	2,116	214	56	300	271	241	3,782
Derm7pt (C)	46	14	18	17	0	26	160	0	281
DermNet (C)	154	40	39	0	38	133	62	21	487
FLUO_SC (C)	176	0	110	0	392	347	13	54	1,092
MIDAS (C)	215	39	555	26	135	419	164	259	1,812
TOTAL training	**1,843**	**307**	**5,782**	**334**	**1,136**	**3,191**	**2,669**	**966**	**16,228**
Validation partition									
BCN20000 (D)	170	18	630	16	110	421	428	57	1,850
Derm12345 (D)	75	12	891	17	4	57	37	15	1,108
Derm7pt (C)	21	6	8	8	0	12	69	0	124
DermNet (C)	22	7	7	0	6	19	10	7	78
FLUO_SC (C)	26	0	17	0	56	50	3	9	161
MIDAS (C)	32	7	80	5	20	61	24	37	266
TOTAL validation	**346**	**50**	**1633**	**46**	**196**	**620**	**571**	**125**	**3,587**
Testing partition									
BCN20000 (D)	172	20	632	18	112	422	430	111	1,917
Derm12345 (D)	175	40	1,733	51	14	66	92	47	2,218
Derm7pt (C)	13	4	5	5	0	7	45	0	79
DermNet (C)	44	11	11	0	11	38	18	7	140
FLUO_SC (C)	50	0	31	0	112	99	3	15	310
MIDAS (C)	61	11	158	7	38	119	47	74	515
HAM10000 (D)	1338	160	7737	180	149	622	1305	229	11,720
SKINL2 (D)	33	14	97	0	0	40	28	0	212
Fitzpatrick17k (C)	30	9	39	0	5	36	115	141	375
HIBA (D)	88	61	602	51	63	340	253	158	1,616
PAD UFES 20 (C)	235	0	244	0	730	845	52	192	2,298
SD198 (C)	60	18	32	3	62	13	38	0	226
MSK (D)	566	18	2595	67	251	741	1311	296	5,845
Milk10k(C)	544	52	746	47	303	2522	450	473	5,137
Milk10k (D)	544	52	746	47	303	2522	450	473	5,137
TOTAL Testing	**3,953**	**470**	**15,408**	**476**	**2,153**	**8,432**	**4,637**	**2,216**	**37,745**

9. Performance on single datasets

Performance on single datasets are reported considering cross-modal retrieval (Table 5) and zero-shot learning (Table 6). For both tasks, the three linking strategies are reported, all combined with P data: P + Single, P + Double, P + Cross. Furthermore, also the performance of state-of-the-art foundation models: specific for medical data, BioMedCLIP[62], MedImgInsights[9], specific for dermatology, MONET[34] or Derm1M[59]. Results are reported as a comparison, in Table 7 (cross-modal retrieval) and Table 8 (zero-shot learning).

Table 5. Results on the cross-modal retrieval task, considering the Image-to-Text setup (left part of the Table) and the Text-to-Image setup (right part) setups. The three linking strategies are reported: P + Single, P + Double, P + Cross. Table reports the mAP (mean pm std of ten runs).

	Image-to-Text			**Text-to-Image**		
	Single	**Double**	**Cross**	**Single**	**Double**	**Cross**
BCN20000	0.652 ± 0.022	0.732 ± 0.024	$\mathbf{0.737 \pm 0.034}$	0.625 ± 0.023	0.681 ± 0.018	$\mathbf{0.684 \pm 0.018}$
derm12345	0.850 ± 0.011	0.861 ± 0.010	$\mathbf{0.862 \pm 0.011}$	0.806 ± 0.010	$\mathbf{0.834 \pm 0.012}$	0.831 ± 0.014
Derm7pt	0.686 ± 0.010	0.733 ± 0.031	$\mathbf{0.752 \pm 0.040}$	0.483 ± 0.019	$\mathbf{0.593 \pm 0.011}$	0.575 ± 0.024
DermNet	0.523 ± 0.027	0.593 ± 0.029	$\mathbf{0.630 \pm 0.056}$	0.443 ± 0.005	$\mathbf{0.538 \pm 0.009}$	0.515 ± 0.021
FLUO_SC	0.633 ± 0.058	0.679 ± 0.017	$\mathbf{0.694 \pm 0.038}$	0.464 ± 0.024	0.534 ± 0.017	$\mathbf{0.544 \pm 0.020}$
MRA_MIDAS	0.478 ± 0.006	0.518 ± 0.016	$\mathbf{0.548 \pm 0.018}$	0.574 ± 0.027	0.631 ± 0.012	$\mathbf{0.633 \pm 0.023}$
HAM10000	$\mathbf{0.770 \pm 0.006}$	0.760 ± 0.008	0.737 ± 0.028	0.387 ± 0.020	$\mathbf{0.620 \pm 0.080}$	0.572 ± 0.099
SKINL2	0.731 ± 0.027	$\mathbf{0.767 \pm 0.014}$	0.752 ± 0.018	0.637 ± 0.008	$\mathbf{0.706 \pm 0.008}$	0.669 ± 0.015
Fitzpatrick17k	$\mathbf{0.579 \pm 0.027}$	0.576 ± 0.015	0.559 ± 0.016	0.495 ± 0.025	$\mathbf{0.583 \pm 0.021}$	0.579 ± 0.023
HIBA	0.537 ± 0.014	$\mathbf{0.557 \pm 0.013}$	0.554 ± 0.021	0.648 ± 0.021	$\mathbf{0.704 \pm 0.017}$	0.699 ± 0.015
PAD-UFES-20	0.615 ± 0.024	$\mathbf{0.627 \pm 0.007}$	0.615 ± 0.013	0.488 ± 0.031	0.560 ± 0.028	$\mathbf{0.587 \pm 0.029}$
SD198	0.399 ± 0.023	0.400 ± 0.019	$\mathbf{0.435 \pm 0.023}$	0.467 ± 0.023	0.532 ± 0.016	$\mathbf{0.541 \pm 0.020}$
MSK	0.620 ± 0.010	$\mathbf{0.627 \pm 0.013}$	0.608 ± 0.011	0.662 ± 0.021	0.709 ± 0.015	$\mathbf{0.718 \pm 0.018}$
Milk10k_clinic	0.519 ± 0.021	$\mathbf{0.567 \pm 0.023}$	0.556 ± 0.011	0.579 ± 0.031	0.645 ± 0.018	$\mathbf{0.654 \pm 0.022}$
Milk10k_dermo	0.597 ± 0.031	$\mathbf{0.647 \pm 0.020}$	0.622 ± 0.015	0.579 ± 0.031	0.645 ± 0.018	$\mathbf{0.654 \pm 0.022}$

Table 6. Results on the zero-shot learning task, considering the 8- (left part of the Table) and the 15-classes (right part) setups. The three linking strategies are reported: P + Single, P + Double, P + Cross. Table reports the weighted F1-score (mean pm std of ten runs).

	Image-to-Text			**Text-to-Image**		
	Single	**Double**	**Cross**	**Single**	**Double**	**Cross**
BCN20000	0.539 ± 0.035	0.639 ± 0.030	$\mathbf{0.666 \pm 0.038}$	0.487 ± 0.061	$\mathbf{0.619 \pm 0.033}$	0.610 ± 0.062
derm12345	0.769 ± 0.020	0.820 ± 0.017	$\mathbf{0.823 \pm 0.021}$	0.279 ± 0.040	$\mathbf{0.226 \pm 0.017}$	0.214 ± 0.031
Derm7pt	0.438 ± 0.023	0.546 ± 0.133	$\mathbf{0.639 \pm 0.06}$	0.465 ± 0.071	0.522 ± 0.161	$\mathbf{0.568 \pm 0.092}$
DermNet	0.519 ± 0.064	0.579 ± 0.074	$\mathbf{0.535 \pm 0.054}$	0.413 ± 0.062	$\mathbf{0.455 \pm 0.071}$	0.426 ± 0.027
FLUO_SC	0.624 ± 0.045	0.792 ± 0.073	$\mathbf{0.815 \pm 0.089}$	$\mathbf{0.487 \pm 0.095}$	0.404 ± 0.020	0.415 ± 0.072
MRA_MIDAS	0.397 ± 0.041	0.396 ± 0.021	$\mathbf{0.451 \pm 0.031}$	0.263 ± 0.060	0.312 ± 0.032	$\mathbf{0.325 \pm 0.053}$
HAM10000	0.693 ± 0.010	$\mathbf{0.718 \pm 0.011}$	0.711 ± 0.011	0.680 ± 0.023	$\mathbf{0.699 \pm 0.016}$	0.653 ± 0.046
SKINL2	0.604 ± 0.036	$\mathbf{0.727 \pm 0.029}$	0.722 ± 0.028	0.520 ± 0.035	$\mathbf{0.606 \pm 0.045}$	0.560 ± 0.059
Fitzpatrick17k	$\mathbf{0.527 \pm 0.040}$	0.411 ± 0.043	0.436 ± 0.081	$\mathbf{0.404 \pm 0.024}$	0.324 ± 0.047	0.369 ± 0.075
HIBA	0.449 ± 0.034	0.497 ± 0.019	$\mathbf{0.508 \pm 0.025}$	0.316 ± 0.044	$\mathbf{0.367 \pm 0.039}$	0.340 ± 0.093
PAD-UFES-20	0.478 ± 0.008	$\mathbf{0.548 \pm 0.017}$	0.505 ± 0.048	$\mathbf{0.416 \pm 0.019}$	0.376 ± 0.018	0.361 ± 0.020
SD198	0.228 ± 0.024	0.302 ± 0.025	$\mathbf{0.307 \pm 0.045}$	0.216 ± 0.022	0.205 ± 0.022	$\mathbf{0.224 \pm 0.033}$
MSK	0.529 ± 0.023	0.551 ± 0.016	$\mathbf{0.556 \pm 0.016}$	0.223 ± 0.009	$\mathbf{0.226 \pm 0.009}$	0.216 ± 0.013
Milk10k_clinic	$\mathbf{0.542 \pm 0.010}$	0.524 ± 0.013	0.527 ± 0.013	0.456 ± 0.017	0.437 ± 0.007	$\mathbf{0.432 \pm 0.020}$
Milk10k_dermo	0.595 ± 0.018	$\mathbf{0.605 \pm 0.012}$	0.580 ± 0.049	0.502 ± 0.031	$\mathbf{0.538 \pm 0.007}$	0.500 ± 0.042

Table 7. Results on the cross-modal retrieval, considering four foundation models used as benchmark: BioMedCLIP, MedImgInsights (MedImg), MONET or Derm1M. Results are reported for both setups (Image-to-Text) and (Text-to-Image), reporting the mAP of the model, used for inference.

	Image-to-Text				**Text-to-Image**			
	BioMedClip	**MedImg**	**MONET**	**Derm1M**	**BioMedClip**	**MedImg**	**MONET**	**Derm1M**
BCN20000	0.320	0.682	0.333	0.403	0.320	0.588	0.397	0.463
derm12345	0.534	0.751	0.443	0.444	0.507	0.769	0.645	0.664
Derm7pt	0.168	0.501	0.230	0.411	0.205	0.525	0.237	0.306
DermNet	0.209	0.483	0.308	0.438	0.167	0.438	0.193	0.263
FLUO_SC	0.236	0.59	0.311	0.422	0.161	0.426	0.196	0.297
MRA_MIDAS	0.289	0.400	0.288	0.307	0.285	0.584	0.350	0.410
HAM10000	0.470	0.752	0.374	0.431	0.517	0.711	0.506	0.571
SKINL2	0.407	0.686	0.407	0.430	0.367	0.654	0.459	0.502
Fitzpatrick17k	0.212	0.406	0.304	0.331	0.186	0.443	0.223	0.291
HIBA	0.351	0.528	0.325	0.425	0.323	0.597	0.409	0.478
PAD-UFES-20	0.271	0.708	0.416	0.490	0.176	0.445	0.236	0.331
SD198	0.217	0.474	0.267	0.333	0.190	0.417	0.231	0.287
MSK	0.383	0.597	0.336	0.398	0.375	0.632	0.459	0.501
Milk10k clinic	0.28	0.457	0.359	0.440	0.224	0.559	0.297	0.401
Milk10k dermo	0.278	0.564	0.329	0.417	0.224	0.559	0.297	0.401

Table 8. Results on the zero-shot learning, considering four foundation models used as benchmark: BioMedCLIP, MedImgInsights (MedImg), MONET or Derm1M. Results are reported for both setups, 8- and 15-classes, reporting the weighted F1-score of the model, used for inference.

	8-classes				**15-classes**			
	BioMedClip	**MedImg**	**MONET**	**Derm1M**	**BioMedClip**	**MedImg**	**MONET**	**Derm1M**
BCN20000	0.251	0.006	0.149	0.439	0.154	0.006	0.155	0.037
derm12345	0.506	0	0.426	0.763	0.309	0	0.1	0.335
Derm7pt	0.322	0	0.343	0.59	0.135	0	0.226	0.383
DermNet	0.307	0.011	0.386	0.491	0.323	0.011	0.313	0.488
FLUO_SC	0.238	0.192	0.294	0.573	0.346	0.192	0.224	0.244
MRA_MIDAS	0.189	0.010	0.177	0.252	0.108	0.010	0.124	0.096
HAM10000	0.406	0	0.341	0.698	0.261	0	0.049	0.018
SKINL2	0.428	0	0.491	0.543	0.21	0	0.347	0.318
Fitzpatrick17k	0.395	0	0.287	0.41	0.22	0	0.247	0.34
HIBA	0.196	0.003	0.24	0.512	0.109	0.003	0.098	0.142
PAD-UFES-20	0.35	0.153	0.295	0.624	0.384	0.153	0.267	0.276
SD198	0.327	0.118	0.404	0.472	0.166	0.048	0.385	0.458
MSK	0.353	0.004	0.186	0.523	0.125	0.004	0.128	0.193
Milk10k clinic	0.182	0.001	0.13	0.273	0.211	0.001	0.101	0.105
Milk10k dermo	0.174	0.001	0.089	0.269	0.156	0.001	0.106	0.05