GESTALTCLIP: CONCEPT-ENHANCED LEARNING OF CLINICALLY STRUCTURED FACIAL REPRESENTATION FROM HETEROGENEOUS PHENOTYPE SUPERVISION

Anonymous authors
Paper under double-blind review

ABSTRACT

Facial morphology provides important diagnostic cues for many rare genetic disorders, yet learning facial representations structured by clinically meaningful phenotypic concepts remains difficult because available datasets are small and supervision is highly fragmented. Rare disease facial representation learning is further limited by scarce image-level phenotype annotations, while richer but heterogeneous supervision exists across multiple levels, including Human Phenotype Ontology (HPO) concepts, patient phenotype sets, and disease level phenotype knowledge. We introduce **GestaltCLIP**, a concept-enhanced vision–language framework that integrates these heterogeneous phenotype signals to learn clinically structured facial representations without requiring complete image–phenotype pairing. GestaltCLIP unifies concept-, patient-, and disease-level phenotype supervision through ontology-aware HPO learning, phenotype-set modeling, direct image–phenotype grounding, and disease-mediated weak supervision. The resulting vision encoder operates with images alone at inference time. On GestaltMatcher Database, GestaltCLIP consistently improves disease recognition across four disease recognition metrics, yielding an average relative improvement of 15.58%, and achieves a 7.73-fold increase in mean average precision (mAP) for atomic image-to-HPO retrieval. This demonstrates substantially improved phenotype-concept grounding. Under strict external evaluation on RD-Face, Top-1 accuracy improves from 8.93% to 15.18%. These results demonstrate that integrating heterogeneous phenotype supervision through explicit concept learning produces clinically structured, phenotype-grounded, and transferable facial representations for rare disease analysis.

1 INTRODUCTION

Characteristic facial morphology provides important phenotypic cues for many rare genetic diseases. This has motivated computational approaches for learning disease related facial representations (Hsieh et al., 2022). However, supervision in this domain is scarce and heterogeneous. Facial images are commonly associated with disease labels, while patient-specific phenotype annotations are much less available. In contrast, structured phenotype knowledge is abundant but is typically not paired with individual images. The Human Phenotype Ontology (HPO) provides such knowledge through clinical concepts, patient phenotype sets, and disease-level phenotype associations (Köhler et al., 2021). The challenge is therefore not only limited supervision, but also a mismatch in the level and image correspondence of phenotype information.

Modern vision–language learning offers a natural way to inject semantic knowledge into visual representations. CLIP (Radford et al., 2021) established large-scale image–text contrastive pretraining, which has since been adapted to biomedical imaging through methods such as ConVIRT (Zhang et al., 2022), BiomedCLIP (Zhang et al., 2023), and PhenoLIP (Liang et al., 2026). Rare disease phenotype supervision, however, does not naturally satisfy this assumption that image and text are completely paired. As illustrated in Figure 1, available phenotype information may be clinically related to a disease without describing the same patient. A disease can be associated with many HPO terms, while an individual patient expresses only a subset of them, and only some expressed pheno-

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

types may be visible in a facial photograph. Directly aligning the full disease phenotype profile with every image therefore introduces systematic supervision noise.

This creates a broader representation-learning problem: *how can we learn useful facial representations when phenotype information is available at different levels?* Existing approaches address only parts of this setting. GestaltMatcher learns transferable facial representations from disease-level supervision (Hsieh et al., 2022), while biomedical vision–language models (VLM) primarily rely on paired or semantically matched image–text supervision (Zhang et al., 2022; Wang et al., 2022; Zhang et al., 2023). Most closely related, PhenoLIP incorporates phenotype ontology knowledge into medical vision–language pretraining (Liang et al., 2026). Our setting differs in that most facial images lack patient-level HPO annotations, while phenotype supervision is distributed across HPO concepts, real patient phenotype sets, and limited verified image–phenotype pairs.

We introduce **GestaltCLIP**, a concept-enhanced vision–language framework for learning from these heterogeneous sources of phenotype supervision. GestaltCLIP first learns an ontology-aware phenotype concept space that preserves graded semantic and hierarchical relationships. It then models dependencies among phenotypes from real patient HPO sets and grounds verified image–phenotype pairs. For images without patient-level phenotype annotations, disease-level knowledge provides a coarse phenotype prior that is refined using image-specific evidence and learned phenotype set structure, yielding patient-conditioned weak supervision instead of assigning the full disease phenotype profile to every image.

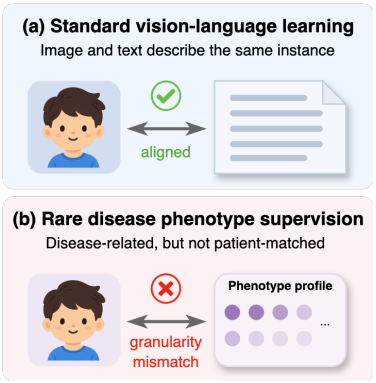


Figure 1: Supervisions in standard vision–language learning and facial phenotype-related learning.

Importantly, phenotype knowledge is used to shape the visual representation during training rather than as a required input at inference. After training, GestaltCLIP exports a standalone vision encoder that operates on facial images alone. This allows us to evaluate whether structured phenotype supervision is genuinely internalized into the learned representation through disease recognition, phenotype grounding, cross-dataset transfer, and unseen disease recognition. Empirically, GestaltCLIP improves disease recognition across four metrics, substantially strengthens image-to-HPO grounding, and improves strict cross-dataset transfer and unseen disease generalization.

Overall, our main contributions are:

- We formulate rare-disease facial representation learning as a **cross-granularity supervision problem**, where phenotype knowledge at the concept, patient, and disease levels is not equally specific to an individual image.
- We introduce **GestaltCLIP**, which learns a shared facial–phenotype representation by combining ontology-aware HPO concepts, patient concept phenotype set modeling, and patient-conditioned disease-mediated alignment.
- We show that this **phenotype-structured representation** improves disease recognition, image–HPO grounding, cross-dataset transfer, and unseen disease generalization beyond matched vision–language baselines.

2 Related Work

Rare Disease Facial Representation Learning. Characteristic craniofacial morphology provides important phenotypic cues for recognizing genetic diseases. Early computational work organized facial images into phenotype spaces reflecting disease similarity (Ferry et al., 2014), while DeepGestalt later demonstrated large-scale supervised recognition across hundreds of genetic diseases (Gurovich et al., 2019). GestaltMatcher further shifted the focus toward transferable representations by constructing a Clinical Face Phenotype Space. In this space, distances between facial embeddings capture disease similarity, enabling matching even for diseases not represented during supervised encoder training (Hsieh et al., 2022). Subsequent work has improved metric learning and incorporated clinical phenotype information into multimodal disease prediction (Hustinx et al., 2023; Wu et al., 2024). Structured phenotypes have also been incorporated into rare disease diagnos-

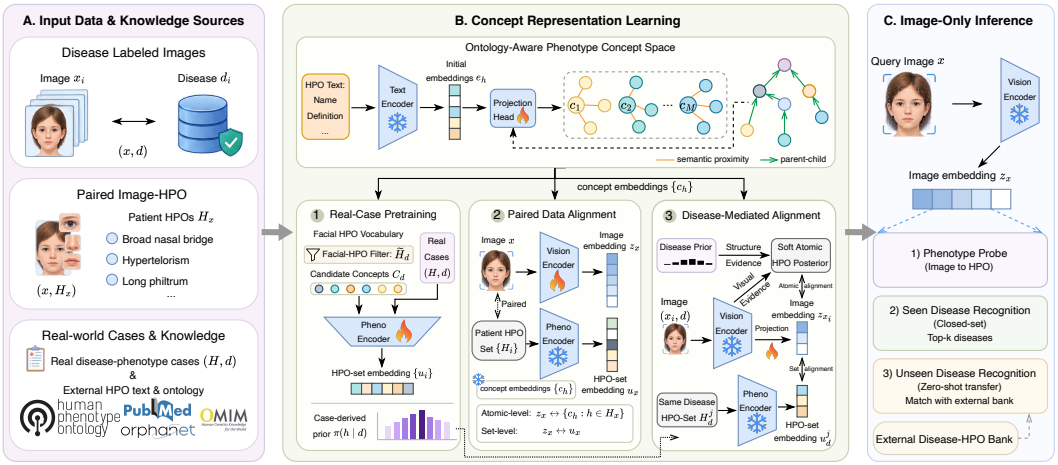


Figure 2: **Overview of GestaltCLIP.** An ontology-shaped HPO concept space supports three-stage training: (1) concept set pretraining on clinical cases, (2) multi-granularity alignment of paired images and patient HPO sets, and (3) disease-mediated alignment of weakly labelled images.

tic systems, including recent agentic frameworks (Zhao et al., 2026; Chen et al., 2026b). However, existing facial representations are still primarily organized around patient or disease similarity rather than explicit phenotype-concept structure. They do not specify which phenotypic traits explain patient similarity or how those traits relate through a structured phenotype ontology.

Biomedical Vision–Language Pretraining. Contrastive vision–language learning enables visual representations to be supervised by richer semantic information than categorical labels. CLIP established large-scale image–text contrastive pretraining as a transferable representation learning paradigm (Radford et al., 2021). FaRL extended visual–linguistic pretraining specifically to facial representation learning, combining face–text contrastive learning with masked image modeling (Zheng et al., 2022). BiomedCLIP further adapted vision–language pretraining to large-scale biomedical image–text data (Zhang et al., 2023), while MedCLIP explored medical contrastive learning with decoupled image–text pairing and knowledge-driven semantic similarity (Wang et al., 2022). Together, these approaches show that visual representation learning can benefit from semantic supervision beyond fixed class labels. However, rare disease phenotype knowledge poses a different correspondence problem. Disease-level HPO profiles summarize population-level manifestations rather than describing a specific image, and may include phenotypes that are non-facial. Their correspondence to an individual face is therefore only partial, making direct image–profile alignment less well specified than conventional image–text supervision.

Phenotype Ontology and Phenotype Learning. HPO provides a standardized representation of phenotypic abnormalities and their semantic relationships (Robinson et al., 2008). It is widely used in computational analysis of Mendelian disease (Köhler et al., 2021). Unlike flat diagnostic labels, HPO concepts form a directed acyclic graph whose structure supports semantic-similarity-based disease prioritization (Köhler et al., 2009; Robinson et al., 2014). In addition, structured phenotype knowledge has begun to inform multimodal representation learning. For example, PhenoLIP learns phenotype-centric representations for vision–language pretraining (Liang et al., 2026). Rare disease facial representation learning, however, must integrate supervision across concept, patient, and disease levels. These levels provide complementary but distinct forms of knowledge through ontology relations, phenotype co-occurrence, and population-level associations.

3 Methodology

3.1 Overview

Let $\mathcal{H}$ denote the fixed facial-HPO vocabulary of photographically assessable phenotypes. We consider facial images x_i with disease labels d_i, with patient-level HPO annotations $H_i \subseteq \mathcal{H}$ available for only a subset of images. These sources differ in semantic granularity and image correspondence. GestaltCLIP performs *concept representation learning* by integrating phenotype knowledge at the

Algorithm 1 Training procedure of GestaltCLIP

Require: HPO ontology $\mathcal{H}$, clinical sets $\mathcal{D}_1$, paired data $\mathcal{D}_2$, weak data $\mathcal{D}_3$, disease profiles $\{C_d\}$
Ensure: Trained visual pathway (f_{vision}, W)

$\psi \leftarrow \arg\min_{\psi}\big(\mathcal{L}_{\text{geo}} + \mathcal{L}_{\text{hier}}\big)$; construct and freeze $\{c_h\}_{h\in\mathcal{H}}$

Stage 1: Real-case pretraining

$M \subset H, \quad V = H \setminus M, \quad u_V = f_{\text{pheno}}(V), \quad H \sim \mathcal{D}_1$

$\mathcal{L}_{S1} = \mathcal{L}_{\text{mask}} + \lambda_{\text{mean}}(1 - u_V^{\top} m_V)$

$f_{\text{pheno}} \leftarrow \arg\min_{f_{\text{pheno}}} \mathcal{L}_{S1}$; freeze f_{pheno}

Stage 2: Paired multi-granularity alignment

$z_i = \text{norm}(W f_{\text{vision}}(x_i)), \quad u_i = f_{\text{pheno}}(H_i), \quad (x_i, H_i, d_i) \sim \mathcal{D}_2$

$\mathcal{L}_{S2} = \mathcal{L}_{\text{NCE}} + \lambda_{\text{PU}}\mathcal{L}_{\text{PU}} + \lambda_{\text{set}}\mathcal{L}_{\text{set}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}}$

$(f_{\text{vision}}, W) \leftarrow \arg\min_{f_{\text{vision}}, W} \mathcal{L}_{S2}$; freeze teacher $(\bar{f}_{\text{vision}}, \bar{W})$

Stage 3: Disease-mediated weak alignment

$\bar{z}_i = \text{norm}(\bar{W} \bar{f}_{\text{vision}}(x_i)), \quad s_{ih}^{\text{vis}} = \bar{z}_i^{\top} c_h$

$q_i(h) \propto \exp\big(s_{ih}^{\text{vis}} + \lambda_{\text{prior}} \text{logit}(\pi_{d_i h}) + \lambda_{\text{str}} s_{ih}^{\text{str}}\big)$

$\mathcal{L}_{S3} = \lambda_{\text{atomic}}\mathcal{L}_{\text{atomic}} + \lambda_{\text{weakset}}\mathcal{L}_{\text{weakset}} + \lambda_{\text{replay}}\mathcal{L}_{S2}^{\text{replay}}$

$W \leftarrow \arg\min_{W} \mathcal{L}_{S3}$ with f_{vision} frozen

return (f_{vision}, W)

concept, patient, and disease levels. This learning produces a phenotype-structured facial representation that requires only the facial image at inference. Figure 2 illustrates the framework, and Algorithm 1 summarizes training.

3.2 Ontology-Aware Phenotype Concept Space

For each facial HPO concept $h \in \mathcal{H}$, we serialize its textual description T_h and obtain an initial embedding from a frozen text encoder $e_h = \text{norm}(f_{\text{text}}(T_h))$. We then inject ontology structure through a trainable residual projection $c_h = \text{norm}((1-\rho)e_h + \rho\ \text{norm}(r_\psi(e_h)))$.

The projection is trained to preserve both graded semantic proximity, measured by Lin similarity (Lin, 1998), and hierarchical ordering among HPO concepts through $\mathcal{L}_{\text{geo}}$ and $\mathcal{L}_{\text{hier}}$, respectively. The resulting concept embeddings $\{c_h\}$ are then frozen.

3.3 Stage 1: Real-Case Pretraining

Given a patient phenotype set H, we mask $M \subset H$ and denote the visible subset by $V = H \setminus M$. The visible concepts are represented by an additive anchor and a learned concept set embedding:

$$m_V = \text{norm}\left(\frac{1}{|V|}\sum_{h\in V} c_h\right), \qquad u_V = f_{\text{pheno}}(\{c_h : h \in V\}). \tag{1}$$

Stage 1 trains f_{pheno} to recover the masked concepts from V. Because unreported HPOs are not necessarily true negatives, masked-HPO prediction uses a positive–unlabeled objective $\mathcal{L}_{\text{mask}}$. We additionally preserve the additive structure of the concept space by regularizing u_V toward m_V:

$$\mathcal{L}_{S1} = \mathcal{L}_{\text{mask}} + \lambda_{\text{mean}}\left(1 - u_V^{\top} m_V\right). \tag{2}$$

After pretraining, f_{pheno} is frozen for Stages 2 and 3.

3.4 Stage 2: Paired Multi-Granularity Alignment

Stage 2 grounds visual representations in the phenotype space using paired image–HPO annotations. A vision backbone f_{vision} followed by a trainable projection W produces

$$z_i = \text{norm}(W f_{\text{vision}}(x_i)), \qquad u_i = f_{\text{pheno}}(\{c_h : h \in H_i\}). \tag{3}$$

Set-level alignment combines a symmetric multi-positive contrastive loss $\mathcal{L}_{\text{NCE}}$ (van den Oord et al., 2018), with same-disease cases treated as additional positives, and a matched image–set cosine loss

$\mathcal{L}_{\text{set}}$. Atomic alignment uses image–concept similarity $z_i^\top c_h$ with a PU objective $\mathcal{L}_{\text{PU}}$. We further preserve the pretrained visual geometry using $\mathcal{L}_{\text{reg}}$. The overall objective is

$$\mathcal{L}_{S2} = \mathcal{L}_{\text{NCE}} + \lambda_{\text{PU}}\mathcal{L}_{\text{PU}} + \lambda_{\text{set}}\mathcal{L}_{\text{set}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}}. \tag{4}$$

3.5 Stage 3: Disease-Mediated Weak Alignment

Stage 3 extends phenotype grounding to disease-labeled images without patient-level HPO annotations. Disease identity defines a facial-HPO candidate set C_{d_i}, while a frozen Stage-2 teacher $(\bar{f}_{\text{vision}}, \bar{W})$ provides image-specific evidence:

$$\bar{z}_i = \text{norm}\big(\bar{W}\bar{f}_{\text{vision}}(x_i)\big), \qquad s_{ih}^{\text{vis}} = \bar{z}_i^\top c_h, \quad h \in C_{d_i}. \tag{5}$$

The frozen phenotype concept set encoder further evaluates structural compatibility, yielding s_{ih}^{str}. Visual evidence, disease-level phenotype priors, and phenotype concept set structure are combined into an image-conditioned atomic posterior:

$$q_i(h) = \text{softmax}_{h\in C_{d_i}}\left(s_{ih}^{\text{vis}} + \lambda_{\text{prior}}\,\text{logit}(\pi_{d_i h}) + \lambda_{\text{str}}s_{ih}^{\text{str}}\right). \tag{6}$$

During Stage 3, f_{vision} remains frozen and only W is optimized. The projection is trained using atomic alignment to q_i, set-level alignment with real same-disease phenotype sets, and replay of paired Stage-2 examples:

$$\mathcal{L}_{S3} = \lambda_{\text{atomic}}\mathcal{L}_{\text{atomic}} + \lambda_{\text{weakset}}\mathcal{L}_{\text{weakset}} + \lambda_{\text{replay}}\mathcal{L}_{S2}^{\text{replay}}. \tag{7}$$

Image-only inference. After Stage 3, the learned visual pathway, consisting of f_{vision} and W, operates directly on the query image without patient-level HPO input. The resulting representation is used for seen-disease recognition and image retrieval. For phenotype and unseen disease evaluation, the image representation is instead compared against fixed HPO or disease–HPO knowledge. Therefore, phenotype knowledge structures the representation during training and candidate matching, but is not required as patient-side input at inference. See Appendix A for more method details.

4 Experiments and Results

We evaluate whether GestaltCLIP improves disease recognition, supports atomic- and set-level HPO concept retrieval, transfers to external datasets, and enables unseen disease recognition from image-derived phenotype evidence. Unless otherwise stated, downstream evaluations freeze the final vision encoder and projection, separating representation quality from task-specific adaptation.

4.1 Experimental Setup

Image-based datasets. We use the GestaltMatcher Database (GMDB) (Lesmann et al., 2023) as the rare disease facial benchmark. Following GestaltMatcher (Hsieh et al., 2022), we stratify diseases by *disease frequency*. Diseases with at least 7 distinct patients form the *Common* partition (275 diseases), while those with one to 6 patients form the *Rare* partition (306 diseases). Both are split at the patient level into 65% training, 15% validation, and 20% test sets. RDFace (Feng et al., 2026) is used exclusively for external evaluation. 25 diseases overlap the GMDB training space, while 77 are unseen. 56 of the unseen diseases have disease-level HPO profiles and constitute the 56-way unseen-disease benchmark. No RDFace image is used for representation learning or model selection. Exact image duplicates are excluded across GMDB partitions and between GMDB and RDFace.

Phenotype knowledge. Stage 1 uses real patient-level phenotype sets from RareArena (Chen et al., 2026a), RareBench (Chen et al., 2024), and FGDD (Song et al., 2025). After normalization and filtering, 7,234 cases remain for phenotype concept set pretraining and for estimating the training-set disease–HPO distribution used in Stage 3. We use the HPO (Gargano et al., 2024) as the phenotype vocabulary and ontology, with a fixed set of 1,127 photographically observable facial HPO concepts. See Appendix A and B.1 for more details on the datasets, HPO profile and facial-HPO filtering.

Baselines. We compare GestaltCLIP with several baselines, including OpenCLIP (Cherti et al., 2023), BiomedCLIP (Zhang et al., 2023), MedCLIP (Wang et al., 2022), PhenoLIP (Liang et al.,

Table 1: Disease recognition on GMDB *common* and *rare* partitions. Results are mean ± standard deviation (SD) across seeds with **Bold** and underline denoting the best and second-best results.

	Classification			Image Retrieval		
Method	Bal. Acc. (%)	Top-5 Acc. (%)	Macro-F1 (%)	mAP (%)	R@1 (%)	R@5 (%)
Common diseases						
OpenCLIP	23.30 ± 0.17	55.89 ± 0.49	23.71 ± 0.44	4.35 ± 0.06	20.07 ± 0.26	38.81 ± 0.60
Adapted OpenCLIP	22.33 ± 0.16	58.55 ± 0.27	23.18 ± 0.20	5.23 ± 0.01	21.72 ± 0.11	41.42 ± 0.16
BiomedCLIP	7.03 ± 0.65	32.54 ± 0.63	6.38 ± 0.68	3.00 ± 0.00	11.54 ± 0.00	27.07 ± 0.00
PhenoLIP	6.81 ± 0.09	31.99 ± 0.08	5.98 ± 0.06	2.86 ± 0.00	11.17 ± 0.00	27.77 ± 0.00
MedCLIP	4.02 ± 0.40	25.96 ± 0.05	3.67 ± 0.47	1.79 ± 0.00	6.28 ± 0.00	16.01 ± 0.00
FaRL	20.59 ± 0.25	53.00 ± 0.13	20.70 ± 0.38	4.13 ± 0.00	20.53 ± 0.00	36.65 ± 0.00
ViT	8.14 ± 0.26	33.99 ± 0.28	7.90 ± 0.09	2.41 ± 0.00	10.64 ± 0.00	21.81 ± 0.00
GestaltMatcher-Arc	19.89 ± 0.06	53.40 ± 0.32	20.55 ± 0.23	5.41 ± 0.00	19.47 ± 0.00	38.03 ± 0.00
GestaltCLIP	**25.93 ± 0.68**	**60.20 ± 0.63**	**26.78 ± 0.82**	**6.77 ± 0.17**	**25.53 ± 0.35**	**45.28 ± 0.27**
Rare diseases						
OpenCLIP	21.83 ± 1.14	38.80 ± 1.13	17.13 ± 0.87	15.32 ± 0.74	19.53 ± 1.70	30.86 ± 1.95
Adapted OpenCLIP	20.34 ± 0.81	39.58 ± 0.60	15.75 ± 0.79	15.89 ± 1.27	19.92 ± 1.35	35.03 ± 0.45
BiomedCLIP	10.03 ± 0.08	22.27 ± 1.79	7.67 ± 1.28	9.73 ± 0.00	8.20 ± 0.00	21.48 ± 0.00
PhenoLIP	6.70 ± 1.25	21.48 ± 0.68	4.72 ± 1.02	8.92 ± 0.00	7.81 ± 0.00	21.88 ± 0.00
MedCLIP	0.49 ± 0.00	4.30 ± 0.78	0.01 ± 0.01	4.10 ± 0.00	3.52 ± 0.00	10.16 ± 0.00
FaRL	20.10 ± 1.29	38.80 ± 1.85	15.87 ± 1.41	14.90 ± 0.00	17.97 ± 0.00	34.77 ± 0.00
ViT	9.14 ± 0.28	22.92 ± 0.45	6.20 ± 0.52	8.52 ± 0.00	8.20 ± 0.00	19.14 ± 0.00
GestaltMatcher-Arc	12.72 ± 0.50	24.48 ± 2.60	9.59 ± 0.30	13.21 ± 0.00	16.41 ± 0.00	28.91 ± 0.00
GestaltCLIP	**26.97 ± 1.19**	**44.40 ± 1.97**	**21.68 ± 0.36**	**19.57 ± 0.25**	**23.31 ± 0.81**	**41.02 ± 0.00**

2026), FaRL (Zheng et al., 2022), ViT (ImageNet) (Dosovitskiy et al., 2021), and GestaltMatcher-Arc (Hustinx et al., 2023) following ArcFace (Deng et al., 2019). To isolate the effect of the training from simple in-domain adaptation, we additionally include an *adapted OpenCLIP* control, which uses the same OpenCLIP vision backbone and is fine-tuned on the paired GMDB image–HPO training data. Appendix C.4 also compares disease-supervised GestaltCLIP with native GestaltMatcher-Arc, while Appendix B.2 provides baseline implementation details.

Evaluation protocols. For **disease recognition**, we evaluate *disease classification* with a trained linear probe on the GMDB training split, with model selection on validation data. We report Top-1, Top-5, balanced accuracy (Bal. Acc.), and macro-F1. *Disease-level image retrieval* uses each test image to query the patient-disjoint training gallery by cosine similarity, with mean average precision (mAP) and Recall@K ($K \in \{1, 5, 10\}$). For **phenotype grounding**, we evaluate *atomic HPO concept retrieval*, *HPO concept linear probing*, and *concept set retrieval*. Atomic concept retrieval ranks all facial HPO concepts directly from the image representation, while set concept embeddings are produced by the frozen f_{pheno}. We report mAP and Recall@K. **Cross-dataset transfer and unseen disease generalization** are evaluated on RDFace without using RDFace for training or model selection. For shared diseases, the GMDB trained linear probe is applied directly to RDFace. For the unseen diseases with available HPO profiles, image-derived phenotype scores are matched against fixed disease-level HPO profiles. We report Top-1, Top-5, Bal. Acc., and macro-F1. Full metric definitions are provided in Appendix B.3.

Implementation details. The GestaltCLIP vision backbone is initialized from OpenCLIP ViT-B/16 pretrained on LAION-2B (Cherti et al., 2023), and the phenotype concept encoder uses SapBERT (Liu et al., 2021). Facial-HPO filtering is performed with GPT-5.5 (OpenAI, 2026) and validated by two human reviewers (see Appendix A). All stages are optimized with AdamW and validation early stopping. Stage 2 adapts only the final two vision blocks and image projection, whereas Stage 3 freezes the vision backbone and optimizes only the projection head. All experiments are run with three seeds 42, 43, 44. Other implementation details such as learning rates and loss weights are provided in the reproducibility tables in Appendix H.

4.2 Disease Recognition

Table 1 first evaluates whether phenotype-structured training improves disease-level utility of the learned representation. GestaltCLIP consistently outperforms all baselines on both classification and image retrieval. On *common* diseases, it achieves the highest balanced accuracy, Top-5 accuracy, and Macro-F1 (25.93%, 60.20%, and 26.78%), together with the strongest retrieval performance.

Table 2: Phenotype grounding on the GMDB *common* and *rare* partitions. All tasks are evaluated using **mAP** ± SD across seeds. *Note that *ViT* and *GestaltMatcher-Arc* are reported only for the linear-probe setting, as they do not define a native cross-modal embedding space for direct HPO or concept set retrieval. **Bold** and underline denote the best and second-best values.

	HPO Concept Retrieval	**HPO Concept Linear Probe**	**Concept Set Retrieval**
Method	mAP (%)	mAP (%)	mAP (%)
Common diseases			
OpenCLIP	1.69 ± 0.00	23.54 ± 0.39	1.53 ± 0.00
Adapted OpenCLIP	9.87 ± 0.52	24.92 ± 0.14	2.97 ± 0.08
BiomedCLIP	1.58 ± 0.00	19.72 ± 0.01	1.43 ± 0.00
PhenoLIP	1.77 ± 0.00	19.48 ± 0.01	1.42 ± 0.00
MedCLIP	0.74 ± 0.00	18.26 ± 0.03	1.02 ± 0.00
FaRL	1.34 ± 0.00	23.15 ± 0.01	1.05 ± 0.00
ViT*	–	19.36 ± 0.01	–
GestaltMatcher-Arc*	–	24.40 ± 0.02	–
GestaltCLIP	**13.07 ± 0.16**	**25.89 ± 0.15**	**6.05 ± 0.41**
Rare diseases			
OpenCLIP	1.98 ± 0.00	16.53 ± 0.37	8.82 ± 0.00
Adapted OpenCLIP	7.62 ± 0.37	17.05 ± 0.24	12.12 ± 0.22
BiomedCLIP	1.38 ± 0.00	14.50 ± 0.04	7.59 ± 0.00
PhenoLIP	2.03 ± 0.00	15.46 ± 0.13	10.51 ± 0.00
MedCLIP	0.68 ± 0.00	16.54 ± 0.09	7.53 ± 0.00
FaRL	1.37 ± 0.00	17.69 ± 0.08	11.13 ± 0.00
ViT*	–	15.10 ± 0.15	–
GestaltMatcher-Arc*	–	18.59 ± 0.10	–
GestaltCLIP	**11.91 ± 0.16**	**21.04 ± 0.47**	**20.69 ± 1.44**

Table 3: External and unseen disease evaluation on RDFace. Results are mean ± SD across seeds. **Bold** and underline denote the best and second-best values.

	External Test			**Unseen Disease Recognition**		
Method	Bal. Acc. (%)	Top-5 (%)	Macro-F1 (%)	Bal. Acc. (%)	Top-5 (%)	Macro-F1 (%)
OpenCLIP	8.89 ± 2.38	17.86 ± 0.89	12.11 ± 2.86	1.94 ± 0.00	8.61 ± 0.00	0.79 ± 0.00
Adapted OpenCLIP	13.96 ± 0.83	22.92 ± 1.86	19.75 ± 1.08	2.50 ± 0.00	17.62 ± 0.00	0.48 ± 0.01
BiomedCLIP	3.67 ± 0.46	7.14 ± 0.89	5.03 ± 0.31	2.74 ± 0.00	9.02 ± 0.00	0.96 ± 0.00
PhenoLIP	5.33 ± 0.99	8.93 ± 0.00	7.38 ± 1.26	2.26 ± 0.00	11.07 ± 0.00	1.22 ± 0.00
MedCLIP	2.33 ± 0.46	8.33 ± 0.52	3.40 ± 0.60	2.11 ± 0.00	6.97 ± 0.00	0.82 ± 0.00
FaRL	14.71 ± 0.54	26.19 ± 1.36	19.19 ± 0.82	2.11 ± 0.00	9.43 ± 0.00	1.43 ± 0.00
ViT	8.33 ± 0.46	15.48 ± 1.03	11.38 ± 0.77	1.79 ± 0.00	12.43 ± 0.24	0.08 ± 0.00
GestaltMatcher-Arc	15.51 ± 2.61	25.89 ± 0.89	20.31 ± 2.71	3.96 ± 0.00	13.93 ± 0.00	1.33 ± 0.00
GestaltCLIP	**15.78 ± 0.89**	**29.17 ± 1.36**	**20.73 ± 1.14**	**5.89 ± 0.93**	**21.04 ± 0.63**	**2.72 ± 0.57**

The gains are larger on *rare* diseases. Relative to OpenCLIP, balanced accuracy increases from 21.83% to 26.97% and Macro-F1 from 17.13% to 21.68%, while retrieval mAP improves from 15.32% to 19.57%. GestaltCLIP also exceeds the adapted OpenCLIP control trained on the same paired image–HPO data, suggesting that in-domain adaptation alone is insufficient to account for these gains. The gains in both linear classification and embedding-based retrieval suggest that phenotype supervision improves facial representations in complementary ways. It enhances disease separability and the local structure of the embedding space. Full metrics are reported in Appendix C.1.

4.3 Phenotype Grounding

We next ask whether disease-level gains are accompanied by phenotype structure in the embedding space. As shown in Table 2, GestaltCLIP achieves the strongest phenotype grounding across GMDB partitions. The largest improvements appear in cross-modal retrieval. Relative to the adapted OpenCLIP control, HPO concept retrieval's mAP increases from 9.87% to 13.07% on *common* diseases and from 7.62% to 11.91% on *rare* diseases. Patient-level set concept retrieval improves from 2.97% to 6.05% and from 12.12% to 20.69%, respectively.

Linear-probe gains are smaller but remain consistent across strong baselines, reaching 25.89% and 21.04% mAP on *common* and *rare* diseases. Linear probing measures whether phenotype information can be decoded from the representation, while direct retrieval requires the image embedding to share the geometry of the phenotype space. The larger gains in direct retrieval suggest that heterogeneous phenotype supervision does more than make HPO information recoverable. It organizes the

Table 4: Ablation study of GestaltCLIP on disease recognition and phenotype grounding. Results are mean ± SD across seeds (%). **Bold** and underline denote the best and second-best values.

Method	**GMDB *Common***	**GMDB *Rare***	**RDFace External**	**RDFace Unseen**
Disease Recognition	Bal. Acc. (%)	Bal. Acc. (%)	Bal. Acc. (%)	Bal. Acc. (%)
Raw concept space	24.13 ± 0.63	23.62 ± 0.73	12.98 ± 1.57	3.86 ± 0.06
Stage 1: Mean-pooled sets	23.77 ± 0.28	24.76 ± 0.60	12.98 ± 1.63	4.40 ± 0.68
Stage 2: w/o set	23.93 ± 0.30	24.79 ± 0.09	14.73 ± 0.81	4.82 ± 0.36
Stage 2: w/o atomic	25.20 ± 0.26	25.23 ± 1.57	14.80 ± 0.80	3.56 ± 0.53
Stage 2: full	25.43 ± 0.37	25.69 ± 1.80	14.53 ± 1.01	3.93 ± 0.94
Stage 3: w/o set	23.91 ± 0.24	26.28 ± 0.24	15.60 ± 0.00	4.84 ± 0.41
Stage 3: w/o atomic	24.18 ± 0.44	26.14 ± 0.81	13.60 ± 1.74	4.31 ± 1.36
Stage 3: w/o replay	24.23 ± 0.48	22.50 ± 2.62	13.89 ± 1.26	3.75 ± 0.36
GestaltCLIP	**25.93 ± 0.68**	**26.97 ± 1.19**	**15.78 ± 0.89**	**5.89 ± 0.93**
Method	**GMDB *Common* Atomic**	**GMDB *Rare* Atomic**	**GMDB *Common* Set**	**GMDB *Rare* Set**
Phenotype Grounding	mAP (%)	mAP (%)	mAP (%)	mAP (%)
Raw concept space	9.62 ± 0.46	8.88 ± 0.09	4.34 ± 0.35	13.29 ± 0.47
Stage 1: Mean-pooled sets	9.91 ± 0.25	9.48 ± 0.48	5.23 ± 0.27	14.56 ± 0.77
Stage 2: w/o set	7.48 ± 0.23	8.18 ± 0.10	2.24 ± 0.18	9.65 ± 0.27
Stage 2: w/o atomic	6.78 ± 0.11	3.80 ± 0.21	4.68 ± 0.22	15.93 ± 1.90
Stage 2: full	9.52 ± 0.14	6.98 ± 0.38	4.71 ± 0.40	18.18 ± 1.37
Stage 3: w/o set	12.49 ± 0.45	9.52 ± 0.59	4.59 ± 0.06	17.61 ± 1.18
Stage 3: w/o atomic	9.76 ± 0.25	7.06 ± 0.15	5.32 ± 0.15	19.42 ± 1.31
Stage 3: w/o replay	12.02 ± 0.44	10.49 ± 0.24	2.53 ± 0.12	11.75 ± 0.54
GestaltCLIP	**13.07 ± 0.16**	**11.91 ± 0.16**	**6.05 ± 0.41**	**20.69 ± 1.44**

visual embedding around phenotype concepts and patient-level phenotype structure. This effect is particularly evident for structured phenotype retrieval. Full metrics are reported in Appendix C.2.

4.4 Cross-Dataset Transfer and Unseen Disease Generalization

We next evaluate whether the learned phenotype structure transfers beyond the GMDB development domain. Table 3 reports results on the independent RDFace dataset, which is used only for external evaluation. On the shared-disease subset, GestaltCLIP gives modest but consistent gains over the strongest face-specialized baselines, improving Top-5 accuracy to 29.17% and Macro-F1 to 20.73%.

The unseen disease setting is stricter. No facial images from the target diseases are used during representation learning, and prediction is performed by matching image-derived phenotype scores to fixed disease-HPO profiles. GestaltCLIP improves balanced accuracy to 5.89% and Top-5 accuracy to 21.04%, compared with 4.29% and 13.93% for GestaltMatcher-Arc, while the random Top-5 performance is approximately 8.9% for this 56-way task. Appendix D further separates visual-only, prior-only, structure-only, and full variants. The full model gives the highest balanced accuracy, supporting a contribution from image-conditioned phenotype evidence beyond disease-HPO priors alone. Full metrics are reported in Appendix C.3.

4.5 Ablation studies

Table 4 examines how different supervision levels contribute to disease utility and phenotype grounding. In Stage 2, atomic- and set-level supervision show complementary behavior. Set-level alignment generally provides stronger utility for disease recognition. Combining the two objectives, however, yields the strongest GMDB disease recognition and improves grounding to *rare* patient profiles compared with either objective alone. This suggests that atomic concept correspondence and patient-level phenotype structure provide distinct but complementary signals.

Stage 3 makes this separation more pronounced. Removing set-level weak alignment retains strong atomic HPO retrieval (12.49% and 9.52% mAP), whereas removing atomic alignment yields the strongest ablated patient-profile retrieval (5.32% and 19.42% mAP). Removing paired replay preserves competitive atomic grounding but substantially reduces GMDB *rare* disease recognition and patient-profile retrieval. This suggests that replay is particularly important for retaining previously learned phenotype structure during weak refinement. With all components retained, With all components retained, GestaltCLIP provides the most consistent performance across disease recognition and phenotype grounding metrics. Full results and additional ablations are in Appendix D.

4.6 Representation Analysis and Visualization

We further examine how phenotype structure is reflected in the learned representation. Figure 3 compares image–phenotype similarity distributions for observed patient phenotypes, same disease unobserved phenotypes, and unrelated HPO concepts. GestaltCLIP shows clearer separation between observed and unrelated phenotypes than the adapted OpenCLIP control, providing direct evidence that phenotype supervision reorganizes the visual embedding space.

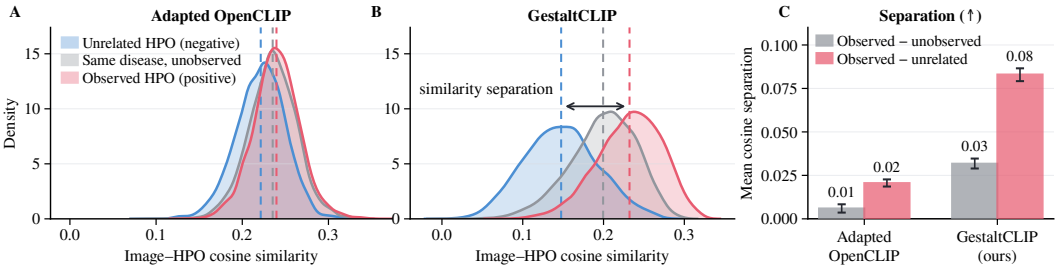


Figure 3: **Image–phenotype similarity separation. (A–B)** Image–HPO cosine-similarity distributions for Adapted OpenCLIP and GestaltCLIP. **(C)** Mean separation between observed HPOs and unrelated or same-disease unobserved HPOs. Error bars show 95% confidence interval.

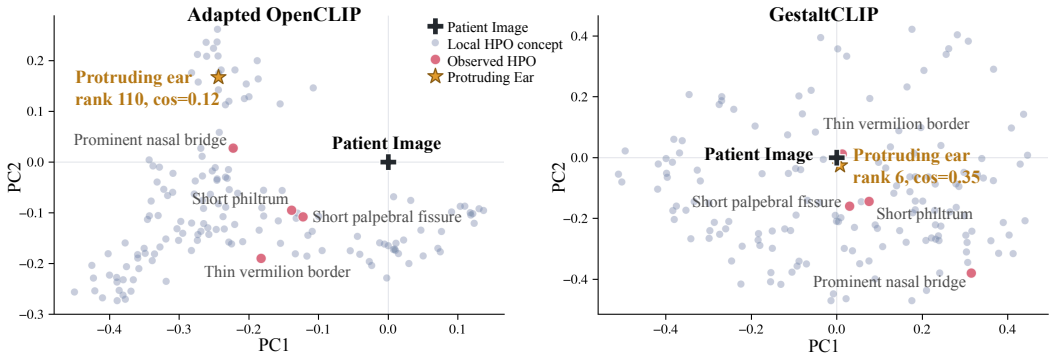


Figure 4: **Patient-level qualitative analysis of phenotype-grounded representation for *Protruding Ear* (HP:0000411).** Patient-centered image–HPO embedding neighborhoods for Adapted OpenCLIP and GestaltCLIP, visualized using the first two principal components (PC1 and PC2).

A qualitative comparison against the strongest baseline further illustrates phenotype grounding. GestaltCLIP moves *Protruding Ear* (HP:0000411) substantially closer to the patient embedding in Figure 4 (rank 110 vs. 6; similarity 0.12 vs. 0.35). In Figure 5, the concept-conditioned occlusion sensitivity maps (Zeiler & Fergus, 2014) show the same pattern of attribution shifting toward auricular regions. These results suggest that phenotype supervision reshapes the visual representation toward clinically relevant facial cues.

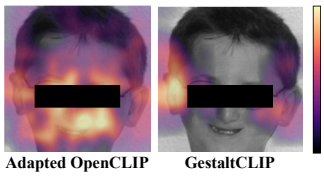


Figure 5: Occlusion sensitivity maps for *Protruding Ear*.

5 Conclusion

We introduced GestaltCLIP, a framework for learning phenotype-structured facial representations from heterogeneous supervision spanning HPO concepts, patient phenotype sets, and disease-level phenotype knowledge. Rather than treating phenotype information only as auxiliary labels, GestaltCLIP uses it to organize the geometry of the visual representation across multiple levels of clinical structure. Across GMDB and the independent RDFace benchmark, this structured representation improves disease recognition, retrieval, phenotype grounding, and cross-dataset transfer, with particularly strong gains for rare and visually unseen diseases. Representation analyses further show stronger image–phenotype separation and anatomically localized phenotype evidence. More broadly, our results suggest that structured semantic knowledge can bridge the mismatch between scarce image-level supervision and richer domain knowledge, providing a general direction for representation learning when supervision is heterogeneous or available at different levels of granularity.

AI USE STATEMENT

We used generative AI tools for brainstorming, language editing, and structured assessment of facial-HPO photographic observability under a predefined annotation protocol. The AI-assisted HPO filtering was validated against blinded human review as described in Appendix A.6. We did not use generative AI tools to generate experimental results or patient data. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This study involves human participant data, including patient facial images and associated phenotype information, and was conducted under institutional Research Ethics Board (REB) approval. To preserve double-blind review, identifying approval information is withheld during submission and will be disclosed in the final version upon acceptance. All patient data and associated phenotype information were obtained, accessed, and used in accordance with applicable ethical approvals, data-use agreements, and dataset-specific governance requirements.

REPRODUCIBILITY STATEMENT

We provide the complete training objectives, implementation details, dataset processing procedures, evaluation protocols, ablation settings, and hyperparameters in the Appendix H. Subject to dataset-access restrictions, we will release training code, preprocessing scripts, split files, HPO vocabulary construction metadata, and configuration files upon acceptance. Facial images cannot be redistributed, but reported splits and disease–HPO mappings will be documented for authorized users.

REFERENCES

Haichao Chen, Zhengyun Zhao, Songchi Zhou, Shikai Hu, Jinyuan Wang, Ye Jin, Xianghong Jin, Yih Chung Tham, Xiaofei Wang, Weizhi Ma, Honghan Wu, Bin Sheng, Shuyang Zhang, Sheng Yu, and Tien Yin Wong. Rarearena: a comprehensive benchmark dataset unveiling the potential of large language models in rare disease diagnosis. *The Lancet Digital Health*, 8(2):100953, 2026a. doi: 10.1016/j.landig.2025.100953.

Xuanzhong Chen, Xiaohao Mao, Qihan Guo, Lun Wang, Shuyang Zhang, and Ting Chen. RareBench: Can LLMs serve as rare diseases specialists? In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 4850–4861. ACM, 2024. doi: 10.1145/3637528.3671576.

Xuanzhong Chen, Ye Jin, Xiaohao Mao, Lun Wang, Shuyang Zhang, and Ting Chen. Rareagents: Autonomous multi-disciplinary team for rare disease diagnosis and treatment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 101–109, 2026b. doi: 10.1609/aaai.v40i1.36969.

Mehdi Cherti, Romain Beaumont, Ross Wightman, et al. 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.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4690–4699, 2019. doi: 10.1109/CVPR.2019.00482.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Ganlin Feng, Yuxi Long, Hafsa Ali, Erin Lou, Fahad Butt, Qian Liu, Yang Wang, and Pingzhao Hu. RDFace: A benchmark dataset for rare disease facial image analysis under extreme data

scarcity and phenotype-aware synthetic generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 42976–42986, 2026.

Yuhao Feng, Lei Qi, and Weidong Tian. Phenobert: A combined deep learning method for automated recognition of human phenotype ontology. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 20(2):1269–1277, 2023. doi: 10.1109/TCBB.2022.3170301.

Quentin Ferry, Julia Steinberg, Caleb Webber, David R. FitzPatrick, Chris P. Ponting, Andrew Zisserman, and Christoffer Nellaker. Diagnostically relevant facial gestalt information from ordinary photos. *eLife*, 3:e02020, 2014. doi: 10.7554/eLife.02020.

Michael A. Gargano, Nicolas Matentzoglu, Ben Coleman, et al. The human phenotype ontology in 2024: Phenotypes around the world. *Nucleic Acids Research*, 52(D1):D1333–D1346, 2024. doi: 10.1093/nar/gkad1005.

Yaron Gurovich, Yair Hanani, Omri Bar, Guy Nadav, Nicole Fleischer, Dekel Gelbman, Lina Basel-Salmon, Peter M. Krawitz, Susanne B. Kamphausen, Martin Zenker, Lynne M. Bird, and Karen W. Gripp. Identifying facial phenotypes of genetic disorders using deep learning. *Nature Medicine*, 25(1):60–64, 2019. doi: 10.1038/s41591-018-0279-0.

Tzung-Chien Hsieh, Aviram Bar-Haim, Shahida Moosa, Nadja Ehmke, Karen W. Gripp, Jean Tori Pantel, Magdalena Danyel, Martin Atta Mensah, Denise Horn, Stanislav Rosnev, et al. Gestaltmatcher facilitates rare disease matching using facial phenotype descriptors. *Nature Genetics*, 54 (3):349–357, 2022. doi: 10.1038/s41588-021-01010-x.

Alexander Hustinx, Fabio Hellmann, Ömer Sümer, Behnam Javanmardi, Elisabeth André, Peter Krawitz, and Tzung-Chien Hsieh. Improving deep facial phenotyping for ultra-rare disorder verification using model ensembles. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5018–5028, 2023.

Sebastian Köhler, Marcel H. Schulz, Peter Krawitz, Sebastian Bauer, Sandra Dölken, Claus E. Ott, Christine Mundlos, Denise Horn, Stefan Mundlos, and Peter N. Robinson. Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *The American Journal of Human Genetics*, 85(4):457–464, 2009. doi: 10.1016/j.ajhg.2009.09.003.

Sebastian Köhler, Michael Gargano, Nicolas Matentzoglu, Leigh C. Carmody, David Lewis-Smith, Nicole A. Vasilevsky, Daniel Danis, Ganna Balagura, Gareth Baynam, Amy M. Brower, et al. The human phenotype ontology in 2021. *Nucleic Acids Research*, 49(D1):D1207–D1217, 2021. doi: 10.1093/nar/gkaa1043.

Hellen Lesmann, Alexander Hustinx, Shahida Moosa, et al. Gestaltmatcher database: A global reference for facial phenotypic variability in rare human diseases. *medRxiv*, 2023. doi: 10.1101/2023.06.06.23290887.

Cheng Liang, Chaoyi Wu, Weike Zhao, Ya Zhang, Yanfeng Wang, and Weidi Xie. Phenolip: Integrating phenotype ontology knowledge into medical vision-language pretraining, 2026. URL https://arxiv.org/abs/2602.06184.

Dekang Lin. An information-theoretic definition of similarity. In *Proceedings of the Fifteenth International Conference on Machine Learning*, pp. 296–304. Morgan Kaufmann, 1998.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. Self-alignment pretraining for biomedical entity representations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4228–4238, 2021. doi: 10.18653/v1/2021.naacl-main.334.

X. Mao, Y. Huang, Y. Jin, et al. A phenotype-based ai pipeline outperforms human experts in differentially diagnosing rare diseases using ehrs. *npj Digital Medicine*, 8:68, 2025. doi: 10.1038/s41746-025-01452-1.

OpenAI. Gpt-5.5 system card. OpenAI, April 2026. URL https://openai.com/index/gpt-5-5-system-card/. Published April 23, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Peter N. Robinson, Sebastian Köhler, Sebastian Bauer, Dominik Seelow, Denise Horn, and Stefan Mundlos. The human phenotype ontology: A tool for annotating and analyzing human hereditary disease. *The American Journal of Human Genetics*, 83(5):610–615, 2008. doi: 10.1016/j.ajhg.2008.09.017.

Peter N. Robinson, Sebastian Köhler, Anika Oellrich, Kai Wang, Christopher J. Mungall, Suzanna E. Lewis, Nicole Washington, Sebastian Bauer, Dominik Seelow, Peter Krawitz, Christian Gilissen, Melissa Haendel, and Damian Smedley. Improved exome prioritization of disease genes through cross-species phenotype comparison. *Genome Research*, 24(2):340–348, 2014. doi: 10.1101/gr.160325.113.

Elena Rojano, José Córdoba-Caballero, Francisco M. Jabato, David Gallego, María Serrano, Belén Pérez, Àngels Parés-Aguilar, James R. Perkins, Juan A. G. Ranea, and Paula Seoane-Zonjic. Evaluating, filtering and clustering genetic disease cohorts based on human phenotype ontology data with cohort analyzer. *Journal of Personalized Medicine*, 11(8):730, 2021. doi: 10.3390/jpm11080730.

Jie Song, Mengqiao He, Shumin Ren, and Bairong Shen. An explainable dataset linking facial phenotypes and genes to rare genetic diseases. *Scientific Data*, 12(1):634, 2025. doi: 10.1038/s41597-025-04922-z.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2018. URL https://arxiv.org/abs/1807.03748.

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 3876–3887. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.256.

Da Wu, Jingye Yang, Cong Liu, Tzung-Chien Hsieh, Elaine Marchi, Justin Blair, Peter Krawitz, Chunhua Weng, Wendy Chung, Gholson J. Lyon, Ian D. Krantz, Jennifer M. Kalish, and Kai Wang. GestaltMML: Enhancing rare genetic disease diagnosis through multimodal machine learning combining facial images and clinical texts, 2024. URL https://arxiv.org/abs/2312.15320v2.

Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Computer Vision – ECCV 2014*, pp. 818–833. Springer, 2014. doi: 10.1007/978-3-319-10590-1_53.

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

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D. Manning, and Curtis P. Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Proceedings of the 7th Machine Learning for Healthcare Conference*, volume 182 of *Proceedings of Machine Learning Research*, pp. 2–25. PMLR, 2022.

Weike Zhao, Chaoyi Wu, Yanjie Fan, Pengcheng Qiu, Xiaoman Zhang, Yuze Sun, Xiao Zhou, Shuju Zhang, Yu Peng, Yanfeng Wang, Xin Sun, Ya Zhang, Yongguo Yu, Kun Sun, and Weidi Xie. An agentic system for rare disease diagnosis with traceable reasoning. *Nature*, 651:775–784, 2026. doi: 10.1038/s41586-025-10097-9.

Yinglin Zheng, Hao Yang, Ting Zhang, Jianmin Bao, Dongdong Chen, Yangyu Huang, Lu Yuan, Dong Chen, Ming Zeng, and Fang Wen. General facial representation learning in a visual-linguistic manner. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18697–18709, 2022.

GestaltCLIP Supplementary Appendix

Contents

A Additional Method Details 2

B Additional Experimental Details 5

C Additional Experimental Results 10

D Additional Ablation Results 15

E Analysis of the Learned Representation 23

F Robustness and Sensitivity Analyses 26

G Limitations and Broader Considerations 28

H Reproducibility Checklist 29

A Additional Method Details

A.1 HPO profile construction and facial HPO filtering

We use the HPO (Gargano et al., 2024) as the phenotype vocabulary and ontology. We construct the facial HPO vocabulary through three sequential steps: (a) canonical case cleaning, (2) ontology-based craniofacial preselection, and (3) photographic-observability filtering. These steps are deliberately separated where ontology membership determines whether a phenotype is anatomically related to the craniofacial region, whereas photographic observability determines whether it can be assessed from facial imagery.

Canonical HPO normalization.

Patient-level phenotype annotations are first standardized against the HPO release of June 23, 2026. Explicitly negated mentions are excluded and alternative HPO identifiers are mapped to their current canonical IDs. Identifiers absent from the current ontology are discarded, and duplicate HPO terms within each patient are collapsed. Cases with fewer than three HPO terms were excluded because extremely sparse phenotype profiles provide limited information for phenotype-based representation learning, following prior phenotype-based cohort filtering and rare disease modeling practice (Rojano et al., 2021; Mao et al., 2025). Importantly, this threshold is applied to the complete phenotype profile rather than to facial phenotypes only.

Ontology-based craniofacial preselection.

To construct a high-recall facial candidate set, we first define a master phenotype universe from the union of positive disease-associated HPO terms and canonical phenotype terms observed in the study datasets. Disease-level HPO associations with a `NOT` qualifier or a non-phenotypic-abnormality aspect are excluded.

A phenotype is retained as an ontology-based craniofacial candidate if it is equal to, or descends through HPO `is_a` relations from, at least one of eight prespecified craniofacial roots: *Abnormality of the face*, *Abnormality of the mouth*, *Abnormal ocular adnexa morphology*, *Abnormal pinna morphology*, *Abnormal hair morphology*, *Abnormal scalp morphology*, *Abnormal neck morphology*, and *Abnormal skull morphology*. Formally, for phenotype h and root set $\mathcal{R}$,

$$h \in \mathcal{H}^{\text{onto}}_{\text{facial}} \iff \exists r \in \mathcal{R} : r = h \ \vee \ r \in \text{Ancestors}(h).$$

We intentionally exclude broader roots such as *Abnormality of the head*, *Abnormality of the eye*, and *Abnormality of the skin*, which contain substantial non-facial subtrees. This ontology step is used only for candidate generation and does not imply visibility in a standard facial photograph.

Photographic observability.

We subsequently assess facial relevance and photographic observability as separate properties using GPT-5.5 (OpenAI, 2026) under a predefined structured annotation protocol. For each HPO candidate, the model is provided with the HPO name and definition, ontology context, structured observability criteria, and literature-supported examples. It assigns a facial-relevance label (FACIAL, PERIFACIAL, NONFACIAL, or BROAD) and an observability label (DIRECT, CONDITIONAL, or NOT-ASSESSABLE). Absence from the literature is not treated as evidence of non-observability.

Table A.1: Summary of the facial-HPO vocabulary construction.

Filtering stage	# HPO terms
Master canonical phenotype universe	12,528
Ontology-based craniofacial candidates	1,390
Photographically assessable within ontology candidates	1,127
Direct	285
Conditional	842

DIRECT phenotypes can in principle be assessed from ordinary facial photographs, whereas CONDITIONAL phenotypes require additional conditions such as profile view, visible ears or hair, an open mouth, or other image-specific requirements. Phenotypes requiring non-visual clinical information or modalities unavailable from facial photographs are labeled NOT-ASSESSABLE.

The primary phenotype vocabulary is restricted to terms that both (i) satisfy the ontology-based craniofacial criterion and (ii) are classified as either DIRECT or CONDITIONAL. This yields 1,127 HPO concepts, comprising 285 DIRECT and 842 CONDITIONAL terms as shown in Table A.1.

HUMAN VALIDATION OF FACIAL PHENOTYPE FILTERING

To validate the automated facial-phenotype filtering used to construct the facial HPO vocabulary, we randomly sampled 200 HPO concepts from the filtered candidate set and conducted a blinded human review. Two human readers independently evaluated each concept based only on its HPO term and official definition. Reviewers determined whether the phenotype could be visually assessed from a frontal facial image and therefore should be retained as a facial phenotype.

A.2 ONTOLOGY-AWARE PHENOTYPE CONCEPT SPACE

Each concept text T_h contains the HPO preferred name, definition, and synonyms. The frozen text encoder produces a 768-dimensional representation, while only the residual projection r_ψ is optimized.

Let $s^{\mathrm{Lin}}(h_a, h_b) \in [0, 1]$ denote Lin similarity computed from HPO information content. The geometry loss is

$$\mathcal{L}_{\mathrm{geo}} = \mathbb{E}_{(a,b)} \left[\left(\frac{1 + c_{h_a}^\top c_{h_b}}{2} - s^{\mathrm{Lin}}(h_a, h_b) \right)^2 \right]. \tag{8}$$

For triplets (a, p, n) in which p is ontologically closer to a than n, hierarchical ordering is enforced by

$$\mathcal{L}_{\mathrm{hier}} = \mathbb{E}_{(a,p,n)} \left[m - c_{h_a}^\top c_{h_p} + c_{h_a}^\top c_{h_n} \right]_+, \qquad m = 0.1. \tag{9}$$

The concept-space objective is $\mathcal{L}_{\mathrm{geo}} + \mathcal{L}_{\mathrm{hier}}$.

A.3 STAGE-1 MASKED-HPO OBJECTIVE

Let M denote the masked observed concepts and U a sampled set of unannotated concepts. Given prediction logits ℓ_h, masked-HPO pretraining uses

$$\mathcal{L}_{\mathrm{mask}} = -\frac{1}{|M|} \sum_{h \in M} \log \sigma(\ell_h) - \lambda_{\mathrm{PU}} \frac{1}{|U|} \sum_{h \in U} \log \sigma(-\ell_h). \tag{10}$$

The reduced weight $\lambda_{\mathrm{PU}} < 1$ reflects that unannotated HPO concepts are not necessarily true negatives. Only f_{pheno} and its masked-HPO prediction head are updated during Stage 1.

A.4 STAGE-2 ALIGNMENT DETAILS

For atomic alignment, image–concept similarity is

$$a_{ih} = z_i^\top c_h. \tag{11}$$

The PU objective $\mathcal{L}_{\mathrm{PU}}$ treats annotated HPO concepts as positives and unannotated concepts as unlabeled, following the same principle as Stage 1.

At the set level, $\mathcal{L}_{\mathrm{NCE}}$ is a symmetric multi-positive contrastive objective between image embeddings and phenotype concept set embeddings, with paired cases and same disease cases treated as positives. $\mathcal{L}_{\mathrm{set}}$ additionally preserves direct matched image–set correspondence.

The geometry-preserving term $\mathcal{L}_{\mathrm{reg}}$ constrains adaptation of the pretrained visual representation during phenotype grounding. Stage 2 updates the trainable visual parameters and projection W, while the phenotype concept bank and f_{pheno} remain fixed.

A.5 STAGE-3 WEAK-ALIGNMENT OBJECTIVES

The atomic target q_i is defined only over disease-compatible candidates C_{d_i}, whereas the student distribution is normalized over the complete facial-HPO vocabulary:

$$p_i(h) = \frac{\exp(z_i^\top c_h/\tau)}{\sum_{u\in\mathcal{H}} \exp(z_i^\top c_u/\tau)}. \tag{12}$$

The atomic alignment loss is

$$\mathcal{L}_{\text{atomic}} = -\frac{1}{B}\sum_i w_i \sum_{h\in C_{d_i}} q_i(h)\log p_i(h), \tag{13}$$

where w_i denotes the sample weight used for weak supervision.

For set-level weak supervision, a real phenotype set $\widetilde{H}_i$ is sampled from the same disease and encoded as

$$\widetilde{u}_i = f_{\text{pheno}}(\widetilde{H}_i). \tag{14}$$

Disease-conditioned set alignment is

$$\mathcal{L}_{\text{weakset}} = \mathcal{L}_{\text{NCE}}(\{z_i\},\{\widetilde{u}_i\}) + \lambda_{\text{set}}\frac{1}{B}\sum_i \left(1 - z_i^\top \widetilde{u}_i\right). \tag{15}$$

Finally, paired Stage-2 samples are replayed during Stage 3 through $\mathcal{L}_{S2}^{\text{replay}}$ to preserve direct image–phenotype grounding while learning from weak disease-mediated supervision.

A.6 HUMAN VALIDATION OF FACIAL PHENOTYPE FILTERING

As shown in Table A.2, the GPT-based facial phenotype filtering showed good agreement with independent human judgments. Reviewer 1 and reviewer 2 agreed with GPT on 73.5% and 79.0% of concepts, respectively, while the two reviewers agreed with each other moderately on 74.5% of concepts ($\kappa = 0.34$).

Table A.2: Blinded human validation of GPT-based facial phenotype observability filtering in a stratified sample of 200 HPO concepts.

Metric	Result
HPO concepts reviewed	200
Expert 1–GPT agreement (%)	73.5
Expert 2–GPT agreement (%)	79.0
Expert 1–Expert 2 agreement (%)	74.5
Expert inter-rater Cohen's κ	0.34
Human-consensus cases evaluated	149
GPT accuracy vs. human consensus (%)	85.2
GPT precision (%)	92.7
GPT recall (%)	89.8
GPT F1 score (%)	91.3

Among the 149 concepts for which the two reviewers reached consensus, GPT achieved 85.2% accuracy for facial observability classification, with 92.7% precision, 89.8% recall, and 91.3% F1. Most disagreements involved phenotypes that were only partially observable from facial appearance or whose definitions included both facial and non-facial manifestations, suggesting that ambiguity was concentrated near the boundary of facial observability. Overall, these results support GPT-based filtering as a scalable approximation to consensus human assessment for constructing the facial HPO vocabulary.

B Additional Experiment Details

B.1 Datasets and splits

In this section, we provide additional details about the datasets used in our experiments.

Image-Based Datasets

GestaltMatcher Database (GMDB). GMDB (Lesmann et al., 2023) serves as the primary facial-image dataset with 10,965 images and 581 diseases for representation learning and in-domain evaluation. Table B.1 and B.2 summarize the dataset statistics and split across stages. All splits are constructed at the patient level, so images from the same individual remain within a single split.

Table B.1: GMDB training, validation, and test splits for *Common* and *Rare* partitions. HPO-annotated images denote samples with usable patient-level HPO annotations after filtering.

Partition	Split	# Images	# Diseases	# HPO-annotated
Common	Train	6,292	275	2,627
	Validation	1,485	275	632
	Test	1,880	275	798
	Total	9,657	275	4,057
Rare	Train	781	306	257
	Validation	271	204	69
	Test	256	204	73
	Total	1,308	306	399

Table B.2: Allocation of GMDB across GestaltCLIP training and evaluation. Stage 2 uses HPO-annotated images, whereas Stage 3 uses disease-labeled images without usable HPO annotations.

Data subset / use	*Common*	*Rare*	Total
Training			
Stage 2	2,627	257	2,884
Stage 3	3,511	448	3,959
Not used	154	76	230
Full training split	6,292	781	7,073
Held-out HPO-annotated data			
Validation: model selection	632	69	701
Test: phenotype-related evaluation	798	73	871

RDFace. RDFace (Feng et al., 2026) is reserved exclusively for external evaluation. Table B.3 summarizes its composition with respect to the GMDB training disease space. No RDFace image is used for representation learning, probe fitting, model selection, or hyperparameter tuning.

Table B.3: Composition of the RDFace external evaluation set.

Subset	# Diseases	# Images
Overall	102	437
Shared with GMDB	25	112
Unseen from GMDB	77	325
Unseen with disease-level HPO profiles	56	244

Figure B.1 characterizes the disease distribution across the image-based datasets used in this study, combining GMDB and RDFace after canonical disease-name normalization. The resulting distribution is strongly long-tailed, where a small number of disease groups contain relatively many images, whereas most diseases are represented by only a few examples. This imbalance motivates separate

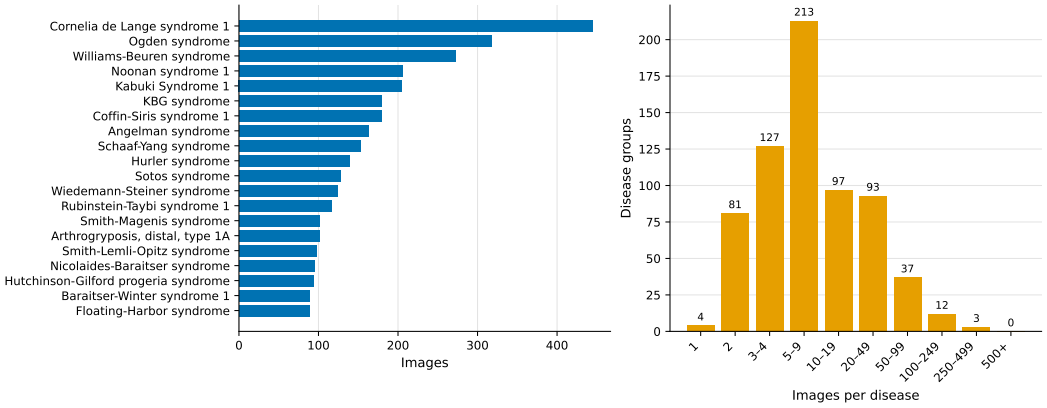


Figure B.1: Distribution of disease-labeled facial images across GMDB and RDFace. Left: the 20 most frequent canonical disease groups. Right: the number of disease groups across image-count bins, illustrating the strongly long-tailed distribution of the combined image collection.

evaluation of *common*, *rare*, and externally held-out diseases and highlights the need for representation learning under severe class-level data scarcity.

PHENOTYPE-BASED DATASETS

Patient-level phenotype data are collected from RareArena (Chen et al., 2026a), RareBench (Chen et al., 2024), and FGDD (Song et al., 2025). RareBench and FGDD provide structured HPO annotations, whereas RareArena consists of free-text clinical cases. We use PhenoBERT (Feng et al., 2023) to extract HPO concepts from RareArena before mapping all annotations to canonical HPO identifiers. Table B.4 summarizes the statistics of the resulting patient-level phenotype sources. The resulting profiles are then filtered to the predefined facial-HPO vocabulary before Stage 1.

Table B.4: Statistics of the patient-level phenotype sources used to construct the Stage 1 pretraining corpus. Numbers in parentheses ( ) indicate cases, or corresponding diseases, for which no curated disease–phenotype associations are available in the official HPO annotation file.

Dataset	Annotation Source	Cases	Diseases
RareArena	Free text → PhenoBERT HPO	72,661 (31,609)	4,596 (2,074)
RareBench	Structured HPO IDs	1,061 (0)	429 (4)
FGDD	Structured HPO IDs	951 (364)	211 (0)
Overall	Canonical HPO profiles	74,673 (31,974)	4,912 (2,076)

Figure B.2 illustrates the heterogeneity of the phenotype-text sources. The available metadata differ largely across datasets. Some sources provide structured HPO codes, while others primarily provide disease labels or additional gene information. Disease terminology is also highly variable before ontology normalization, particularly in RareArena, further motivating canonical mapping of disease and phenotype annotations before constructing the training corpus.

Figure B.3 summarizes the scale of the structured phenotype knowledge used by GestaltCLIP. The HPO ontology, typically the `phenotype.hpoa` file, contains tens of thousands of concepts and relations, while the disease annotation bank provides a larger collection of disease–phenotype associations. These resources support both ontology-aware concept space construction and the disease-level phenotype priors used during weak alignment.

B.2 BASELINE IMPLEMENTATION DETAILS

We use publicly released pretrained checkpoints for external baselines when doing so does not introduce overlap with our evaluation data. When an official checkpoint may have been trained using data that overlap with our evaluation setting, we reproduce the corresponding architecture using an independent pretraining dataset instead. No rare disease facial images from the evaluation cohorts are

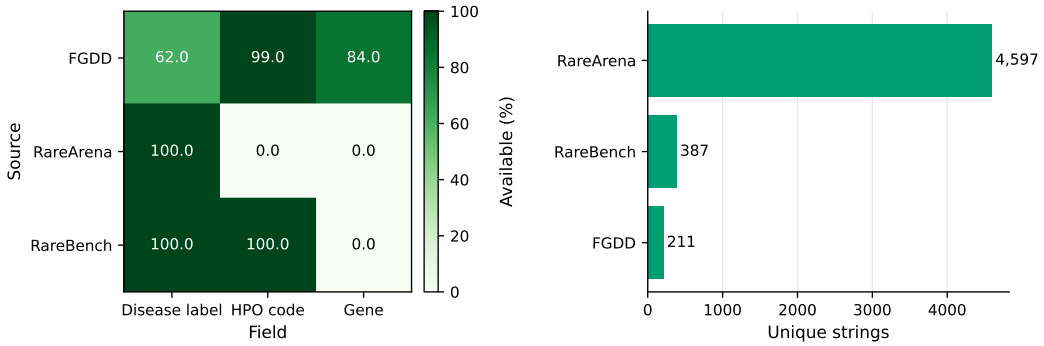


Figure B.2: Characteristics of the phenotype-text sources. Left: availability of disease labels, HPO codes, and gene information across datasets. Right: the number of unique disease-label strings before ontology-based normalization, illustrating substantial heterogeneity in source terminology.

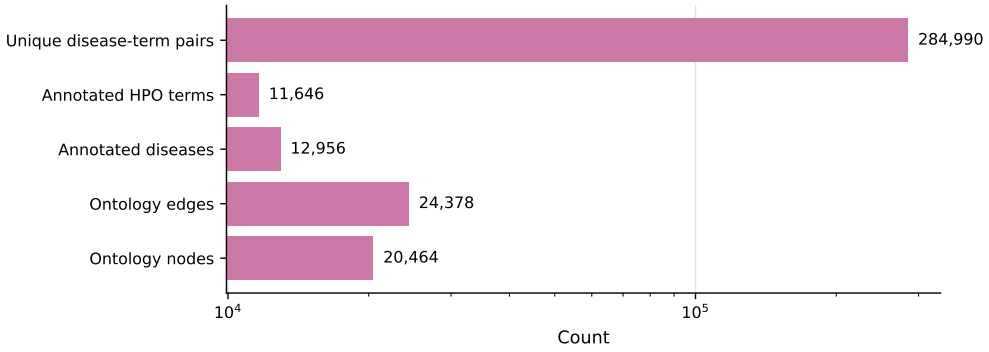


Figure B.3: Scale of the HPO ontology and disease–phenotype annotation bank used by Gestalt-CLIP, including ontology nodes and edges, annotated diseases and HPO terms, and unique disease–term associations.

used during baseline pretraining or adaptation. The pretrained encoder is kept frozen, and only the downstream linear prediction head is trained for representation evaluation unless otherwise stated. A summary of the implementation details for each baseline is provided in Table B.5.

OpenCLIP. We use the pretrained OpenCLIP ViT-B/16 model with the LAION-2B checkpoint `laion2b_s34b_b88k`. The visual encoder takes 224×224 images, which are processed using the checkpoint-specific resizing, cropping, and normalization pipeline.

BiomedCLIP. We use the publicly released BiomedCLIP checkpoint, which employs a ViT-B/16 visual encoder (`BiomedCLIP-PubMedBERT_256-vit_base_patch16_224`) with an input resolution of 224×224. Images are processed using the preprocessing configuration provided with the released checkpoint.

MedCLIP. We use the officially released MedCLIP-ViT checkpoint. Images are processed using the preprocessing pipeline provided with the official implementation.

PhenoLIP. We use the publicly released PhenoLIP model, which employs a ViT-B visual encoder and PubMedBERT text encoder and is pretrained using the phenotype-oriented PhenoKG corpus. We follow the authors' released preprocessing and inference procedure and use the output of the pretrained visual encoder as the image representation.

FaRL. We use the publicly released `FaRL-Base-Patch16-LAIONFace20M-ep16` checkpoint, corresponding to the ViT-B/16 model pretrained for 16 epochs on LAION-Face-20M and used

Table B.5: Implementation details of the evaluated representation models. "Adaptation" denotes additional training on GMDB data used in this study.

Method	Backbone / checkpoint	Input	In-domain adaptation
OpenCLIP	ViT-B/16, LAION-2B	224^2	None
BiomedCLIP	ViT-B/16, PMC-15M	224^2	None
MedCLIP	MedCLIP-ViT, official checkpoint	224^2	None
PhenoLIP	ViT-B/16, PhenoKG	224^2	None
FaRL	ViT-B/16, LAION-Face-20M	224^2	None
ViT (ImageNet)	ViT-B/16, ImageNet-21K/1K	224^2	None
ArcFace	ArcFace iResNet-100, Glint360K	112^2	Glint360K
Adapted OpenCLIP	ViT-B/16, LAION-2B	224^2	Paired GMDB image–HPO
GestaltCLIP	**ViT-B/16, OpenCLIP initialization**	$\mathbf{224^2}$	**Stages 2–3 phenotype training**

in the original paper. We use the CLIP ViT-B/16 preprocessing associated with the released model and extract the resulting visual representation for downstream evaluation.

ViT (ImageNet). We use a ViT-B/16 model pretrained on ImageNet-1K as a conventional visual representation baseline. Specifically, we use the `ViT_B_16_Weights.IMAGENET1K_V1` weights from Torchvision. Images are resized and center-cropped to 224×224 and normalized using the corresponding ImageNet preprocessing.

GestaltMatcher-Arc. We do not use the publicly released GestaltMatcher-Arc checkpoint because the official model is further trained on GMDB, which is also part of our evaluation setting. Using this checkpoint could therefore introduce overlap between representation training and our patient-level evaluation splits. Instead, we use an ArcFace iResNet-100 (Deng et al., 2019) initialized with the publicly available Glint360K-pretrained `glint360k_r100` weights, corresponding to the facial backbone initialization used by GestaltMatcher-Arc. To preserve potentially disease-relevant craniofacial context, including peripheral cues such as the ears and overall head contour, we use the same context-preserving facial field of view as the other methods rather than applying a strict landmark-aligned tight crop. Images are resized to the ArcFace-native 112×112 input resolution before feature extraction. We additionally evaluate GestaltMatcher-Arc under its native landmark-aligned preprocessing and disease-supervised fine-tuning protocol as a strong disease-recognition baseline. Full results are provided in Appendix C.4.

Adapted OpenCLIP. To control for direct adaptation to paired phenotype supervision, we fine-tune the same OpenCLIP ViT-B/16 visual backbone on the paired GMDB training set. Each HPO concept is independently encoded from its preferred name using the frozen OpenCLIP text encoder. For each patient, we construct a phenotype concept set target by mean-pooling the embeddings of the annotated HPO concepts. Training uses a weighted combination of symmetric image–set InfoNCE and a multi-positive image–HPO contrastive objective, in which all annotated HPO concepts for an image are treated as positives. No phenotype concept set encoder, ontology-structured concept representation, or disease-mediated weak supervision is used. The resulting visual encoder is evaluated using the same downstream protocol as the other representations.

B.3 Evaluation Metrics

Let x_i denote a query image, d_i its disease label, and H_i its observed facial-HPO set. Let $\pi_i(k)$ denote the item ranked at position k for query i, and let TopK_i denote the set of the Top-K retrieved items. For multi-positive retrieval, $\mathcal{P}_i$ denotes the set of relevant gallery items for query i. Metrics are macro-averaged over queries unless otherwise specified. The definition of each metric is provided in Table B.6.

Recall@K is defined according to the relevance structure of each task. For multi-positive image/set retrieval, we use a retrieval-success definition in which a query is counted as successful if at least one relevant same disease item appears within the Top K. For atomic HPO concept retrieval, Recall@K instead measures the fraction of the patient's observed HPO concepts recovered within the Top

K. For patient-profile retrieval, each query has exactly one relevant profile, so the two definitions coincide.

Table B.6: Definitions of the evaluation tasks and metrics. All retrieval metrics are computed per query and then averaged across queries.

Task	Metric	Definition	Interpretation
Disease classification	Top-K Acc.	$\mathrm{Acc@K} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\left[d_i \in \mathrm{TopK}_i\right]$	Fraction of test images for which the true disease is among the K highest-scoring disease classes.
Disease classification	Bal. Acc.	$\mathrm{BAcc} = \frac{1}{C}\sum_{c=1}^{C} \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FN}_c}$	Mean per-disease recall, giving each disease equal weight irrespective of class frequency.
Disease classification	Macro-F1	$\mathrm{MacroF1} = \frac{1}{C}\sum_{c=1}^{C} \frac{2P_cR_c}{P_c + R_c}$	Mean class-wise F1 score, where P_c and R_c are precision and recall for disease c.
Image retrieval	Recall@K	$\mathrm{R@K} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\left[\mathrm{TopK}_i \cap \mathcal{P}_i \neq \emptyset\right],$ $\mathcal{P}_i = \{j : d_j = d_i\}.$	Whether at least one training image from the same disease as the query appears among the Top-K retrieved images.
Image retrieval	mAP	$\mathrm{mAP} = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{\lvert\mathcal{P}_i\rvert} \sum_{r:\pi_i(r)\in\mathcal{P}_i} \mathrm{Prec}_i(r)$	Measures the full ranking quality of all same disease training images.
atomic HPO concept retrieval	Recall@K	$\mathrm{R@K} = \frac{1}{N}\sum_{i=1}^{N} \frac{\lvert\mathrm{TopK}_i \cap H_i\rvert}{\lvert H_i\rvert}$	Fraction of the query patient's observed HPO concepts recovered among the Top-K individually ranked HPO concepts.
atomic HPO concept retrieval	mAP	$\mathrm{mAP} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{AP}_i$	Mean average precision over patient images. Unlike patient-profile retrieval, each query can have multiple relevant HPO concepts.
HPO concept linear probe	Recall@K	$\mathrm{R@K} = \frac{1}{N}\sum_{i=1}^{N} \frac{\lvert\mathrm{TopK}_i \cap H_i\rvert}{\lvert H_i\rvert}$	Fraction of the query patient's observed HPO concepts recovered among the top-K HPO predictions produced by a multi-label linear probe trained on the frozen image representation.
HPO concept linear probe	mAP	$\mathrm{mAP} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{AP}_i$	Mean average precision over the full HPO ranking produced by the linear probe, measuring how well phenotype information is linearly recoverable from the frozen image representation.
Concept set retrieval	Recall@K	$\mathrm{R@K} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\left[r_i \leq K\right]$	r_i is the rank of the *same patient's* unique observed HPO profile. Each query has exactly one positive profile.
Concept set retrieval	mAP	$\mathrm{mAP} = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{r_i}$	Because each query has exactly one relevant profile, AP reduces to reciprocal rank and mAP is numerically identical to MRR.
Unseen disease recognition	Top-K Acc.	$\mathrm{Acc@K} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\left[d_i \in \mathrm{TopK}_i\right]$	The query image is ranked against disease candidates not represented by training images; candidate scores are obtained by matching the predicted HPO profile to disease HPO profiles.

C Additional Experimental Results

In this section, we report the full experimental results that could not be included in the main text due to space constraints. We provide complete evaluation metrics for the experiments summarized in the main paper, together with additional supporting analyses where applicable.

This section reports the complete metrics for the downstream evaluation tasks summarized in the main text. Each table corresponds to a single evaluation task so that results obtained under different prediction or retrieval protocols are not conflated. For GMDB evaluations, the *common* and *rare* disease partitions are reported together within the same table. Results are mean ± SD across seeds and are reported as percentages.

Table C.1: Disease classification on GMDB. **Bold** and underline denote the best and second-best.

Method	Top-1 (%)	Top-5 (%)	Bal. Acc. (%)	Macro-F1 (%)
Common diseases				
OpenCLIP	32.61 ± 0.05	55.89 ± 0.49	23.30 ± 0.17	23.71 ± 0.44
Adapted OpenCLIP	33.33 ± 0.13	58.55 ± 0.27	22.33 ± 0.16	23.18 ± 0.20
BiomedCLIP	13.46 ± 0.38	32.54 ± 0.63	7.03 ± 0.65	6.38 ± 0.68
PhenoLIP	13.23 ± 0.06	31.99 ± 0.08	6.81 ± 0.09	5.98 ± 0.06
MedCLIP	10.69 ± 0.44	25.96 ± 0.05	4.02 ± 0.40	3.67 ± 0.47
FaRL	29.50 ± 0.12	53.00 ± 0.13	20.59 ± 0.25	20.70 ± 0.38
ViT	15.44 ± 0.22	33.99 ± 0.28	8.14 ± 0.26	7.90 ± 0.09
GestaltMatcher-Arc	32.91 ± 0.22	53.40 ± 0.32	19.89 ± 0.06	20.55 ± 0.23
GestaltCLIP	**35.85 ± 0.09**	**60.20 ± 0.63**	**25.93 ± 0.68**	**26.78 ± 0.82**
Rare diseases				
OpenCLIP	22.79 ± 1.63	38.80 ± 1.13	21.83 ± 1.14	17.13 ± 0.87
Adapted OpenCLIP	21.48 ± 0.68	39.58 ± 0.60	20.34 ± 0.81	15.75 ± 0.79
BiomedCLIP	9.77 ± 0.39	22.27 ± 1.79	10.03 ± 0.08	7.67 ± 1.28
PhenoLIP	7.42 ± 1.56	21.48 ± 0.68	6.70 ± 1.25	4.72 ± 1.02
MedCLIP	1.17 ± 1.03	4.30 ± 0.78	0.49 ± 0.00	0.01 ± 0.01
FaRL	20.96 ± 0.90	38.80 ± 1.85	20.10 ± 1.29	15.87 ± 1.41
ViT	9.24 ± 0.81	22.92 ± 0.45	9.14 ± 0.28	6.20 ± 0.52
GestaltMatcher-Arc	12.37 ± 0.23	24.48 ± 2.60	12.72 ± 0.50	9.59 ± 0.30
GestaltCLIP	**26.95 ± 1.03**	**44.40 ± 1.97**	**26.97 ± 1.19**	**21.68 ± 0.36**

Table C.2: Disease image retrieval on GMDB. **Bold** and underline denote the best and second-best.

Method	mAP (%)	R@1 (%)	R@5 (%)	R@10 (%)
Common diseases				
OpenCLIP	4.35 ± 0.06	20.07 ± 0.26	38.81 ± 0.60	48.49 ± 0.79
Adapted OpenCLIP	5.23 ± 0.01	21.72 ± 0.11	41.42 ± 0.16	51.38 ± 0.19
BiomedCLIP	3.00 ± 0.00	11.54 ± 0.00	27.07 ± 0.00	36.49 ± 0.00
PhenoLIP	2.86 ± 0.00	11.17 ± 0.00	27.77 ± 0.00	36.65 ± 0.00
MedCLIP	1.79 ± 0.00	6.28 ± 0.00	16.01 ± 0.00	22.66 ± 0.00
FaRL	4.13 ± 0.00	20.53 ± 0.00	36.65 ± 0.00	45.96 ± 0.00
ViT	2.41 ± 0.00	10.64 ± 0.00	21.81 ± 0.00	29.15 ± 0.00
GestaltMatcher-Arc	5.41 ± 0.00	19.47 ± 0.00	38.03 ± 0.00	47.82 ± 0.00
GestaltCLIP	**6.77 ± 0.17**	**25.53 ± 0.35**	**45.28 ± 0.27**	**54.65 ± 0.08**
Rare diseases				
OpenCLIP	15.32 ± 0.74	19.53 ± 1.70	30.86 ± 1.95	38.28 ± 1.03
Adapted OpenCLIP	15.89 ± 1.27	19.92 ± 1.35	35.03 ± 0.45	42.58 ± 0.39
BiomedCLIP	9.73 ± 0.00	8.20 ± 0.00	21.48 ± 0.00	29.69 ± 0.00
PhenoLIP	8.92 ± 0.00	7.81 ± 0.00	21.88 ± 0.00	32.03 ± 0.00
MedCLIP	4.10 ± 0.00	3.52 ± 0.00	10.16 ± 0.00	14.06 ± 0.00
FaRL	14.90 ± 0.00	17.97 ± 0.00	34.77 ± 0.00	40.62 ± 0.00
ViT	8.52 ± 0.00	8.20 ± 0.00	19.14 ± 0.00	28.52 ± 0.00
GestaltMatcher-Arc	13.21 ± 0.00	16.41 ± 0.00	28.91 ± 0.00	31.25 ± 0.00
GestaltCLIP	**19.57 ± 0.25**	**23.31 ± 0.81**	**41.02 ± 0.00**	**46.48 ± 1.03**

Table C.3: HPO concept retrieval on GMDB. **Bold** and underline denote the best and second-best.

Method	mAP (%)	R@1 (%)	R@5 (%)	R@10 (%)
Common diseases				
OpenCLIP	1.69 ± 0.00	0.04 ± 0.00	0.79 ± 0.00	2.22 ± 0.00
Adapted OpenCLIP	9.87 ± 0.52	3.85 ± 0.22	7.69 ± 0.74	11.68 ± 0.72
BiomedCLIP	1.58 ± 0.00	0.19 ± 0.00	0.30 ± 0.00	0.81 ± 0.00
PhenoLIP	1.77 ± 0.00	0.10 ± 0.00	0.33 ± 0.00	1.11 ± 0.00
MedCLIP	0.74 ± 0.00	0.03 ± 0.00	0.12 ± 0.00	0.27 ± 0.00
FaRL	1.34 ± 0.00	0.13 ± 0.00	0.45 ± 0.00	1.02 ± 0.00
ViT	–	–	–	–
GestaltMatcher-Arc	–	–	–	–
GestaltCLIP	**13.07 ± 0.16**	**5.66 ± 0.26**	**12.23 ± 0.14**	**17.04 ± 0.09**
Rare diseases				
OpenCLIP	1.98 ± 0.00	0.44 ± 0.00	1.13 ± 0.00	1.97 ± 0.00
Adapted OpenCLIP	7.62 ± 0.37	1.48 ± 0.20	5.99 ± 0.32	10.50 ± 1.11
BiomedCLIP	1.38 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	1.28 ± 0.00
PhenoLIP	2.03 ± 0.00	0.34 ± 0.00	0.96 ± 0.00	1.73 ± 0.00
MedCLIP	0.68 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
FaRL	1.37 ± 0.00	0.15 ± 0.00	0.61 ± 0.00	0.72 ± 0.00
ViT	–	–	–	–
GestaltMatcher-Arc	–	–	–	–
GestaltCLIP	**11.91 ± 0.16**	**5.04 ± 0.21**	**11.56 ± 0.59**	**15.35 ± 0.62**

Table C.4: HPO concept linear-probe results on GMDB. **Bold** and underline denote the best and second-best.

Method	mAP (%)	R@1 (%)	R@5 (%)	R@10 (%)
Common diseases				
OpenCLIP	23.54 ± 0.39	5.90 ± 0.48	20.05 ± 0.30	30.37 ± 0.35
Adapted OpenCLIP	24.92 ± 0.14	6.76 ± 0.07	20.97 ± 0.12	31.87 ± 0.33
BiomedCLIP	19.72 ± 0.01	5.18 ± 0.01	16.20 ± 0.04	25.88 ± 0.10
PhenoLIP	19.48 ± 0.01	5.57 ± 0.01	15.33 ± 0.11	25.08 ± 0.05
MedCLIP	18.26 ± 0.03	5.05 ± 0.08	13.79 ± 0.02	23.02 ± 0.03
FaRL	23.15 ± 0.01	6.25 ± 0.01	19.26 ± 0.04	29.61 ± 0.09
ViT	19.36 ± 0.01	4.53 ± 0.02	15.98 ± 0.06	25.71 ± 0.11
GestaltMatcher-Arc	24.40 ± 0.02	6.70 ± 0.01	20.43 ± 0.09	32.63 ± 0.10
GestaltCLIP	**25.89 ± 0.15**	**7.38 ± 0.27**	**21.46 ± 0.21**	**32.88 ± 0.29**
Rare diseases				
OpenCLIP	16.53 ± 0.37	2.26 ± 0.23	15.41 ± 1.02	21.40 ± 0.07
Adapted OpenCLIP	17.05 ± 0.24	2.39 ± 0.33	15.08 ± 0.42	22.28 ± 0.52
BiomedCLIP	14.50 ± 0.04	2.25 ± 0.11	11.61 ± 0.76	20.77 ± 0.24
PhenoLIP	15.46 ± 0.13	1.87 ± 0.08	16.12 ± 0.14	21.64 ± 0.00
MedCLIP	16.54 ± 0.09	4.78 ± 0.00	12.94 ± 0.07	22.73 ± 0.11
FaRL	17.69 ± 0.08	4.60 ± 0.20	13.97 ± 0.10	23.44 ± 0.24
ViT	15.10 ± 0.15	2.75 ± 0.13	12.43 ± 0.13	20.50 ± 0.39
GestaltMatcher-Arc	18.59 ± 0.10	4.30 ± 0.11	14.53 ± 0.06	25.64 ± 0.32
GestaltCLIP	**21.04 ± 0.47**	**5.12 ± 0.00**	**17.08 ± 0.24**	**28.38 ± 0.22**

C.1 Disease Recognition

Disease Classification

Table C.1 reports all four classification metrics. In addition to Bal. Acc. and Macro-F1, which account for class imbalance, we include Top-1 and Top-5 accuracy to show performance under both exact and candidate-list recognition. Reporting the two GMDB partitions together also allows the behavior of each representation to be inspected under different levels of disease sample availability without changing the downstream evaluation protocol.

Table C.5: Concept set retrieval on GMDB. **Bold** and underline denote the best and second-best.

Method	mAP (%)	R@1 (%)	R@5 (%)	R@10 (%)
Common diseases				
OpenCLIP	1.53 ± 0.00	0.25 ± 0.00	1.00 ± 0.00	2.13 ± 0.00
Adapted OpenCLIP	2.97 ± 0.08	0.58 ± 0.14	3.26 ± 0.38	6.18 ± 0.19
BiomedCLIP	1.43 ± 0.00	0.13 ± 0.00	1.38 ± 0.00	2.51 ± 0.00
PhenoLIP	1.42 ± 0.00	0.25 ± 0.00	1.63 ± 0.00	2.26 ± 0.00
MedCLIP	1.02 ± 0.00	0.00 ± 0.00	0.88 ± 0.00	1.63 ± 0.00
FaRL	1.05 ± 0.00	0.13 ± 0.00	0.63 ± 0.00	1.38 ± 0.00
ViT	–	–	–	–
GestaltMatcher-Arc	–	–	–	–
GestaltCLIP	**6.05 ± 0.41**	**1.84 ± 0.29**	**7.77 ± 0.70**	**13.87 ± 0.82**
Rare diseases				
OpenCLIP	8.82 ± 0.00	2.74 ± 0.00	8.22 ± 0.00	16.44 ± 0.00
Adapted OpenCLIP	12.12 ± 0.22	2.74 ± 0.00	14.61 ± 0.79	36.53 ± 0.79
BiomedCLIP	7.59 ± 0.00	0.00 ± 0.00	9.59 ± 0.00	20.55 ± 0.00
PhenoLIP	10.51 ± 0.00	2.74 ± 0.00	12.33 ± 0.00	21.92 ± 0.00
MedCLIP	7.53 ± 0.00	0.00 ± 0.00	9.59 ± 0.00	19.18 ± 0.00
FaRL	11.13 ± 0.00	0.00 ± 0.00	17.81 ± 0.00	27.40 ± 0.00
ViT	–	–	–	–
GestaltMatcher-Arc	–	–	–	–
GestaltCLIP	**20.69 ± 1.44**	**9.13 ± 2.09**	**31.51 ± 0.00**	**48.40 ± 1.58**

DISEASE IMAGE RETRIEVAL

Unlike the linear classification evaluation, image retrieval introduces no task-specific prediction head and therefore directly probes the neighborhood structure of the frozen representation. Table C.2 reports mAP together with Recall at progressively larger retrieval depths, providing a more complete view than any single nearest-neighbor cutoff.

C.2 PHENOTYPE GROUNDING

ATOMIC HPO CONCEPT RETRIEVAL

Table C.3 evaluates direct image-to-concept alignment against the fixed HPO concept space. This task differs from HPO linear probing because no classifier is learned from image features to phenotype labels. Successful retrieval therefore requires compatible cross-modal geometry. Methods without a compatible phenotype embedding space are marked as unavailable rather than being assigned an artificial score.

HPO CONCEPT LINEAR PROBING

Table C.4 reports the complete atomic HPO concept linear-probe results. Because the visual encoder is frozen, this experiment measures the phenotype information accessible from the learned representation rather than the quality of direct image–concept alignment. The four metrics additionally distinguish overall multi-label ranking quality from recovery of relevant HPO concepts at different cutoff depths.

PATIENT-LEVEL CONCEPT SET RETRIEVAL

Table C.5 evaluates concept retrieval at the level of complete patient phenotype profiles rather than individual HPO concepts. This provides a stricter test of whether the visual embedding reflects combinations of phenotypes that co-occur within patients. As in atomic retrieval, methods for which a compatible concept set representation is unavailable are indicated with a dash.

C.3 Cross-dataset transfer and unseen disease generalization

RDFace shared disease transfer

Table C.6 reports the complete shared disease transfer evaluation on RDFace. No RDFace-specific learning is performed. Therefore, this evaluation tests how well the learned facial representation transfers across datasets while the target disease identities remain represented during development.

Table C.6: Cross-dataset disease recognition on the RDFace shared disease subset. **Bold** and underline denote the best and second-best.

Method	Top-1 (%)	Top-5 (%)	Bal. Acc. (%)	Macro-F1 (%)
OpenCLIP	8.93 ± 1.79	17.86 ± 0.89	8.89 ± 2.38	12.11 ± 2.86
Adapted OpenCLIP	12.80 ± 0.52	22.92 ± 1.86	13.96 ± 0.83	19.75 ± 1.08
BiomedCLIP	3.87 ± 0.52	7.14 ± 0.89	3.67 ± 0.46	5.03 ± 0.31
PhenoLIP	5.65 ± 1.03	8.93 ± 0.00	5.33 ± 0.99	7.38 ± 1.26
MedCLIP	2.38 ± 0.52	8.33 ± 0.52	2.33 ± 0.46	3.40 ± 0.60
FaRL	13.99 ± 1.03	$\underline{26.19 \pm 1.36}$	14.71 ± 0.54	19.19 ± 0.82
ViT	8.63 ± 0.52	15.48 ± 1.03	8.33 ± 0.46	11.38 ± 0.77
GestaltMatcher-Arc	$\underline{14.88 \pm 2.25}$	25.89 ± 0.89	$\underline{15.51 \pm 2.61}$	$\underline{20.31 \pm 2.71}$
GestaltCLIP	$\mathbf{15.18 \pm 1.55}$	$\mathbf{29.17 \pm 1.36}$	$\mathbf{15.78 \pm 0.89}$	$\mathbf{20.73 \pm 1.14}$

RDFace unseen disease recognition

Table C.7 reports all four metrics for the unseen disease evaluation. This protocol differs from cross-dataset transfer because the target diseases are excluded from facial representation learning. The evaluation therefore tests whether the learned image representation can be related to disease hypotheses through phenotype in the absence of visual training examples from the target classes.

Table C.7: Unseen disease recognition on RDFace. **Bold** and underline denote the best and second-best.

Method	Top-1 (%)	Top-5 (%)	Bal. Acc. (%)	Macro-F1 (%)
OpenCLIP	2.46 ± 0.00	8.61 ± 0.00	1.94 ± 0.00	0.79 ± 0.00
Adapted OpenCLIP	2.87 ± 0.00	$\underline{17.62 \pm 0.00}$	2.50 ± 0.00	0.48 ± 0.01
BiomedCLIP	2.87 ± 0.00	9.02 ± 0.00	2.74 ± 0.00	0.96 ± 0.00
PhenoLIP	2.05 ± 0.00	11.07 ± 0.00	2.26 ± 0.00	1.22 ± 0.00
MedCLIP	2.05 ± 0.00	6.97 ± 0.00	2.11 ± 0.00	0.82 ± 0.00
FaRL	2.05 ± 0.00	9.43 ± 0.00	2.11 ± 0.00	$\underline{1.43 \pm 0.00}$
ViT	2.05 ± 0.00	12.43 ± 0.24	1.79 ± 0.00	0.08 ± 0.00
GestaltMatcher-Arc	$\underline{4.10 \pm 0.00}$	13.93 ± 0.00	$\underline{3.96 \pm 0.00}$	1.33 ± 0.00
GestaltCLIP	$\mathbf{5.19 \pm 0.63}$	$\mathbf{21.04 \pm 0.63}$	$\mathbf{5.89 \pm 0.93}$	$\mathbf{2.72 \pm 0.57}$

C.4 Disease-Supervised Comparison with Native GestaltMatcher-Arc

Our main experiments evaluate facial representations under a common downstream protocol, including frozen-encoder linear probing and retrieval. For that comparison, the ArcFace backbone underlying GestaltMatcher-Arc is evaluated using the same context-preserving facial field of view as the other methods, rather than the native landmark-aligned GestaltMatcher preprocessing.

We therefore conduct an additional **method-faithful comparison** in which GestaltMatcher-Arc follows its native ArcFace-based preprocessing and disease-supervised training procedure. Specifically, faces are landmark-aligned and resized to 112×112, and the ArcFace iResNet-100 backbone initialized from Glint360K is fine-tuned using disease labels according to the GestaltMatcher-Arc training setting. GestaltCLIP is likewise evaluated after disease-supervised fine-tuning.

Table C.8 reports the resulting performance across the full downstream task suite. Please not that the linear probe is the only phenotype grouding task reported for *GestaltMatcher-Arc*, as it does not define a native cross-modal embedding space for direct HPO concept or concept set retrieval.

Native GestaltMatcher-Arc provides a particularly strong closed-set disease recognition baseline, achieving higher GMDB *common* disease linear classification accuracy than GestaltCLIP. The

two methods, however, exhibit markedly different behavior on representation-oriented evaluations. GestaltCLIP performs substantially better on GMDB disease image retrieval and several HPO concept and concept set retrieval tasks. It also shows better transfer to unseen RDFace diseases. These results are consistent with the two training objectives inducing different representation properties. The native GestaltMatcher-Arc pipeline is highly effective for disease discrimination, whereas GestaltCLIP yields an embedding space that better supports representation-oriented tasks.

Table C.8: Disease-supervised comparison with native GestaltMatcher-Arc and GestaltCLIP. **Bold** denotes the best performance.

Evaluation	Metric	GestaltMatcher-Arc	GestaltCLIP
GMDB common diseases			
Disease classification	Top-1 (%)	**50.43 ± 0.16**	40.41 ± 0.56
	Top-5 (%)	**70.94 ± 0.06**	61.83 ± 0.17
	Bal. Acc. (%)	32.15 ± 0.11	**33.15 ± 1.20**
	Macro-F1 (%)	**32.91 ± 0.27**	32.82 ± 0.95
Image retrieval	mAP (%)	9.74 ± 0.00	**31.31 ± 0.93**
	R@1 (%)	33.46 ± 0.34	**38.10 ± 0.66**
	R@5 (%)	**53.56 ± 0.79**	48.24 ± 0.42
	R@10 (%)	**62.82 ± 0.37**	52.87 ± 0.28
Image-to-HPO probe	mAP (%)	26.32 ± 0.23	**26.84 ± 0.04**
	R@1 (%)	7.19 ± 0.06	**7.49 ± 0.33**
	R@5 (%)	21.41 ± 0.33	**22.69 ± 0.08**
	R@10 (%)	33.06 ± 0.56	**35.04 ± 0.06**
Native atomic HPO retrieval	mAP (%)	–	6.85 ± 0.15
	R@1 (%)	–	2.87 ± 0.17
	R@5 (%)	–	6.78 ± 0.16
	R@10 (%)	–	9.43 ± 0.29
Patient-set retrieval	mAP (%)	–	6.57 ± 0.02
	R@1 (%)	–	2.38 ± 0.00
	R@5 (%)	–	9.06 ± 0.07
	R@10 (%)	–	13.95 ± 0.07
GMDB rare diseases			
Disease classification	Top-1 (%)	25.13 ± 0.45	**25.65 ± 1.63**
	Top-5 (%)	38.80 ± 1.19	**42.71 ± 3.63**
	Bal. Acc. (%)	24.63 ± 0.66	**25.37 ± 1.61**
	Macro-F1 (%)	19.64 ± 1.42	**19.69 ± 0.14**
Image retrieval	mAP (%)	22.78 ± 0.00	**25.94 ± 0.70**
	R@1 (%)	24.87 ± 0.45	**28.12 ± 0.96**
	R@5 (%)	36.20 ± 0.60	**39.06 ± 0.64**
	R@10 (%)	43.88 ± 0.60	**47.27 ± 0.55**
Image-to-HPO probe	mAP (%)	20.24 ± 0.11	**20.92 ± 0.32**
	R@1 (%)	3.30 ± 0.22	**5.81 ± 0.17**
	R@5 (%)	15.91 ± 0.23	**17.08 ± 0.51**
	R@10 (%)	25.60 ± 2.02	**26.05 ± 0.19**
Native atomic HPO retrieval	mAP (%)	–	5.35 ± 0.38
	R@1 (%)	–	1.76 ± 0.57
	R@5 (%)	–	5.38 ± 0.87
	R@10 (%)	–	8.20 ± 1.38
Patient-set retrieval	mAP (%)	–	20.74 ± 2.04
	R@1 (%)	–	8.22 ± 1.37
	R@5 (%)	–	31.05 ± 2.85
	R@10 (%)	–	42.92 ± 2.85
RDFace external evaluation			
Shared diseases	Top-1 (%)	12.20 ± 1.86	**14.88 ± 0.52**
	Top-5 (%)	**29.76 ± 1.36**	29.17 ± 2.25
	Bal. Acc. (%)	12.40 ± 1.78	**15.87 ± 0.61**
	Macro-F1 (%)	16.11 ± 2.23	**20.18 ± 0.82**
Unseen diseases	Top-1 (%)	4.37 ± 0.19	**6.15 ± 0.71**
	Top-5 (%)	14.62 ± 0.39	**21.72 ± 0.71**
	Bal. Acc. (%)	4.51 ± 0.00	**5.51 ± 0.59**
	Macro-F1 (%)	1.69 ± 0.00	**3.77 ± 0.47**

D Full ablation Studies

This section provides the complete results for the **ablation experiments summarized in the main text**, and the **additional ablation studies**. Each table isolates a single downstream evaluation task, while the GMDB *common* and *rare* disease partitions are reported together when applicable. This organization makes it possible to examine whether the effect of each component is consistent across evaluation criteria without conflating classification, embedding retrieval, and phenotype grounding.

All results are reported as mean ± SD across seeds and are expressed as percentages. Bold and underlined values denote the best and second-best mean results, respectively, within each partition and metric. For concept set retrieval, we use exact patient-profile retrieval where each query image is matched against its own patient's observed facial-HPO profile rather than treating profiles from the same disease as interchangeable.

D.1 Full Ablation Results in main text

Disease classification

Table D.1 reports the complete frozen-linear disease-classification ablation. The comparison separates changes introduced before paired image–phenotype alignment, alternative Stage 2 objectives, and the contributions of the Stage 3 weak-alignment components. Reporting all four classification metrics also reveals cases in which a component improves candidate-list recognition without producing the same ordering under Top-1 or class-balanced metrics.

Table D.1: Ablation results for frozen-linear disease classification on GMDB. **Bold** and underline denote the best and second-best values.

Variant	Top-1 (%)	Top-5 (%)	Bal. Acc. (%)	Macro-F1 (%)
Common diseases				
Raw concept space	32.48 ± 0.25	57.48 ± 0.50	24.13 ± 0.63	24.38 ± 0.58
Stage 1: Mean-pooled sets	33.37 ± 0.33	58.01 ± 0.19	23.77 ± 0.28	24.13 ± 0.49
Stage 2: w/o set	33.51 ± 0.40	58.21 ± 0.37	23.93 ± 0.30	24.47 ± 0.49
Stage 2: w/o atomic	35.41 ± 0.21	**60.23 ± 0.22**	25.20 ± 0.26	26.22 ± 0.38
Stage 2: full	35.43 ± 0.37	59.84 ± 0.59	25.43 ± 0.37	26.25 ± 0.48
Stage 3: w/o set	34.08 ± 0.08	58.37 ± 0.13	23.91 ± 0.24	24.64 ± 0.33
Stage 3: w/o atomic	34.47 ± 0.33	60.14 ± 0.31	24.18 ± 0.44	24.95 ± 0.49
Stage 3: w/o replay	34.72 ± 0.57	60.11 ± 0.23	24.23 ± 0.48	25.01 ± 0.53
Full GestaltCLIP	**35.85 ± 0.09**	60.20 ± 0.63	**25.93 ± 0.68**	**26.78 ± 0.82**
Rare diseases				
Raw concept space	24.61 ± 1.03	39.97 ± 1.19	23.62 ± 0.73	18.57 ± 0.54
Stage 1: Mean-pooled sets	25.78 ± 0.68	40.76 ± 0.23	24.76 ± 0.60	19.17 ± 0.65
Stage 2: w/o set	25.13 ± 0.60	40.49 ± 0.60	24.79 ± 0.09	19.49 ± 0.41
Stage 2: w/o atomic	25.91 ± 1.63	42.58 ± 1.17	25.23 ± 1.57	19.61 ± 0.74
Stage 2: full	26.43 ± 0.98	43.88 ± 0.23	25.69 ± 1.80	20.27 ± 0.91
Stage 3: w/o set	25.65 ± 2.00	43.49 ± 0.90	26.28 ± 0.24	21.02 ± 1.41
Stage 3: w/o atomic	26.30 ± 1.26	43.62 ± 1.58	26.14 ± 0.81	20.60 ± 0.70
Stage 3: w/o replay	23.44 ± 2.56	40.89 ± 0.98	22.50 ± 2.62	17.39 ± 1.86
Full GestaltCLIP	**26.95 ± 1.03**	**44.40 ± 1.97**	**26.97 ± 1.19**	**21.68 ± 0.36**

Disease image retrieval

Table D.2 evaluates how the same ablations affect disease-level neighborhood structure without fitting a downstream classifier. The results show that the component producing the strongest linear classification performance is not necessarily optimal at every retrieval cutoff.

Atomic HPO concept retrieval

Table D.3 isolates direct atomic phenotype grounding. This task is particularly informative for distinguishing the effects of concept-level and set-level supervision because evaluation is performed directly against the fixed HPO concept bank rather than through a fitted phenotype classifier.

Table D.2: Ablation results for disease image-to-image retrieval on GMDB. **Bold** and underline denote the best and second-best values.

Variant	mAP (%)	R@1 (%)	R@5 (%)	R@10 (%)
Common diseases				
Raw concept space	4.84 ± 0.05	18.62 ± 0.30	36.24 ± 0.16	46.33 ± 0.38
Stage 1: Mean-pooled sets	4.72 ± 0.01	17.46 ± 0.39	35.92 ± 0.22	45.16 ± 0.37
Stage 2: w/o set	4.58 ± 0.06	20.73 ± 0.16	40.41 ± 0.30	49.77 ± 0.75
Stage 2: w/o atomic	5.34 ± 0.15	23.55 ± 0.64	42.68 ± 0.54	52.87 ± 0.82
Stage 2: full	5.39 ± 0.12	24.40 ± 0.79	43.09 ± 0.95	52.87 ± 1.19
Stage 3: w/o set	6.11 ± 0.26	24.82 ± 0.51	44.41 ± 0.88	54.61 ± 0.70
Stage 3: w/o atomic	6.39 ± 0.13	25.14 ± 0.25	**45.34 ± 0.21**	**55.28 ± 0.13**
Stage 3: w/o replay	5.67 ± 0.12	23.71 ± 0.38	44.50 ± 0.41	53.99 ± 0.47
Full GestaltCLIP	**6.77 ± 0.17**	**25.53 ± 0.35**	45.28 ± 0.27	54.65 ± 0.08
Rare diseases				
Raw concept space	12.66 ± 0.31	11.20 ± 0.98	28.12 ± 1.03	36.33 ± 0.39
Stage 1: Mean-pooled sets	18.60 ± 0.25	21.61 ± 0.45	35.94 ± 0.68	44.27 ± 1.37
Stage 2: w/o set	15.81 ± 0.09	20.44 ± 0.60	32.29 ± 0.81	39.32 ± 0.60
Stage 2: w/o atomic	18.22 ± 0.39	20.96 ± 1.26	35.68 ± 0.81	44.66 ± 1.37
Stage 2: full	18.61 ± 0.59	22.40 ± 0.98	39.32 ± 1.85	46.48 ± 1.41
Stage 3: w/o set	18.18 ± 0.57	20.96 ± 1.26	34.77 ± 1.17	44.01 ± 1.63
Stage 3: w/o atomic	19.55 ± 0.20	22.92 ± 0.45	38.67 ± 1.79	**47.79 ± 0.23**
Stage 3: w/o replay	17.76 ± 0.84	23.18 ± 1.48	35.29 ± 0.98	44.01 ± 1.37
Full GestaltCLIP	**19.57 ± 0.25**	**23.31 ± 0.81**	**41.02 ± 0.00**	46.48 ± 1.03

Table D.3: Ablation results for atomic HPO concept retrieval on GMDB. **Bold** and underline denote the best and second-best values.

Variant	mAP (%)	R@1 (%)	R@5 (%)	R@10 (%)
Common diseases				
Raw concept space	9.62 ± 0.46	3.91 ± 0.27	7.17 ± 0.44	11.23 ± 0.77
Stage 1: Mean-pooled sets	9.91 ± 0.25	4.23 ± 0.42	7.67 ± 0.74	11.22 ± 1.17
Stage 2: w/o set	7.48 ± 0.23	1.45 ± 0.15	6.03 ± 0.33	10.41 ± 0.47
Stage 2: w/o atomic	6.78 ± 0.11	2.80 ± 0.14	6.71 ± 0.46	9.56 ± 0.25
Stage 2: full	9.52 ± 0.14	3.86 ± 0.18	9.30 ± 0.26	12.87 ± 0.10
Paired replay only	9.90 ± 0.14	4.37 ± 0.19	9.33 ± 0.15	12.77 ± 0.09
Stage 3: w/o set	12.49 ± 0.45	5.26 ± 0.34	11.93 ± 0.17	16.70 ± 0.59
Stage 3: w/o atomic	9.76 ± 0.25	4.17 ± 0.28	9.10 ± 0.23	12.96 ± 0.27
Stage 3: w/o replay	12.02 ± 0.44	4.96 ± 0.41	11.63 ± 0.29	16.34 ± 0.27
Full GestaltCLIP	**13.07 ± 0.16**	**5.66 ± 0.26**	**12.23 ± 0.14**	**17.04 ± 0.09**
Rare diseases				
Raw concept space	8.88 ± 0.09	2.15 ± 0.37	6.99 ± 0.69	12.60 ± 1.24
Stage 1: Mean-pooled sets	9.48 ± 0.48	2.69 ± 0.83	9.75 ± 0.38	14.57 ± 1.30
Stage 2: w/o set	8.18 ± 0.10	2.20 ± 0.19	6.20 ± 0.62	10.05 ± 1.22
Stage 2: w/o atomic	3.80 ± 0.21	1.05 ± 0.15	2.64 ± 0.38	4.80 ± 0.66
Stage 2: full	6.98 ± 0.38	2.01 ± 0.82	6.50 ± 1.09	10.50 ± 1.75
Stage 3: w/o set	9.52 ± 0.59	2.36 ± 0.60	9.14 ± 0.55	13.87 ± 0.61
Stage 3: w/o atomic	7.06 ± 0.15	2.09 ± 0.36	6.26 ± 1.18	11.16 ± 0.91
Stage 3: w/o replay	10.49 ± 0.24	3.09 ± 0.36	**11.69 ± 0.12**	15.19 ± 0.05
Full GestaltCLIP	**11.91 ± 0.16**	**5.04 ± 0.21**	11.56 ± 0.59	**15.35 ± 0.62**

HPO CONCEPT LINEAR PROBING

Table D.4 evaluates phenotype information that can be decoded from each frozen representation. In contrast to atomic retrieval, this evaluation allows a fitted multilabel probe and therefore measures phenotype accessibility rather than direct image–concept geometric alignment.

Table D.4: Ablation results for frozen HPO concept linear probing on GMDB. **Bold** and underline denote the best and second-best values.

Variant	mAP (%)	R@1 (%)	R@5 (%)	R@10 (%)
Common diseases				
Raw concept space	24.45 ± 0.15	6.23 ± 0.22	20.33 ± 0.14	31.16 ± 0.23
Stage 1: Mean-pooled sets	24.70 ± 0.20	6.39 ± 0.31	20.53 ± 0.48	31.46 ± 0.21
Stage 2: w/o set	24.83 ± 0.11	6.52 ± 0.12	20.90 ± 0.34	31.98 ± 0.24
Stage 2: w/o atomic	25.57 ± 0.04	6.92 ± 0.17	20.83 ± 0.14	32.25 ± 0.07
Stage 2: full	25.71 ± 0.13	6.87 ± 0.19	21.31 ± 0.12	32.66 ± 0.35
Stage 3: w/o set	25.51 ± 0.03	6.91 ± 0.17	20.76 ± 0.11	32.11 ± 0.12
Stage 3: w/o atomic	25.71 ± 0.13	6.89 ± 0.17	21.24 ± 0.33	32.56 ± 0.44
Stage 3: w/o replay	25.34 ± 0.08	7.00 ± 0.12	20.97 ± 0.08	32.36 ± 0.16
Full GestaltCLIP	$\mathbf{25.89 \pm 0.15}$	$\mathbf{7.38 \pm 0.27}$	$\mathbf{21.46 \pm 0.21}$	$\mathbf{32.88 \pm 0.29}$
Rare diseases				
Raw concept space	17.08 ± 0.90	2.68 ± 1.02	14.69 ± 0.67	22.25 ± 0.26
Stage 1: Mean-pooled sets	17.48 ± 0.53	2.82 ± 1.11	16.16 ± 0.57	22.63 ± 1.33
Stage 2: w/o set	16.94 ± 0.53	2.59 ± 0.95	16.64 ± 0.41	21.57 ± 0.15
Stage 2: w/o atomic	19.77 ± 0.61	3.93 ± 0.28	16.33 ± 1.13	26.39 ± 1.04
Stage 2: full	19.41 ± 0.05	3.93 ± 0.00	15.83 ± 0.22	26.59 ± 0.74
Stage 3: w/o set	19.79 ± 0.76	4.80 ± 0.56	16.08 ± 1.56	25.33 ± 0.34
Stage 3: w/o atomic	20.30 ± 0.33	4.87 ± 0.26	16.85 ± 0.52	24.65 ± 0.44
Stage 3: w/o replay	17.18 ± 0.25	2.86 ± 0.36	16.11 ± 0.23	21.66 ± 0.41
Full GestaltCLIP	$\mathbf{21.04 \pm 0.47}$	$\mathbf{5.12 \pm 0.00}$	$\mathbf{17.08 \pm 0.24}$	$\mathbf{28.38 \pm 0.22}$

Table D.5: Ablation results for patient-level concept set retrieval on GMDB. **Bold** and underline denote the best and second-best values.

Variant	mAP (%)	R@1 (%)	R@5 (%)	R@10 (%)
Common diseases				
Raw concept space	4.34 ± 0.35	1.09 ± 0.26	4.89 ± 0.55	9.31 ± 0.92
Stage 1: Mean-pooled sets	5.23 ± 0.27	1.29 ± 0.19	6.60 ± 0.38	11.86 ± 0.19
Stage 2: w/o set	2.24 ± 0.18	0.33 ± 0.19	2.51 ± 0.22	4.05 ± 0.51
Stage 2: w/o atomic	4.68 ± 0.22	1.29 ± 0.32	5.64 ± 0.66	10.48 ± 1.34
Stage 2: full	4.71 ± 0.40	1.38 ± 0.38	5.14 ± 0.33	9.90 ± 0.82
Stage 3: w/o set	4.59 ± 0.06	1.34 ± 0.14	5.51 ± 0.25	9.69 ± 0.32
Stage 3: w/o atomic	5.32 ± 0.15	1.67 ± 0.07	6.56 ± 0.19	11.74 ± 0.59
Stage 3: w/o replay	2.53 ± 0.12	0.63 ± 0.22	2.38 ± 0.13	4.76 ± 0.33
Full GestaltCLIP	$\mathbf{6.05 \pm 0.41}$	$\mathbf{1.84 \pm 0.29}$	$\mathbf{7.77 \pm 0.70}$	$\mathbf{13.87 \pm 0.82}$
Rare diseases				
Raw concept space	13.29 ± 0.47	3.65 ± 2.09	26.94 ± 2.09	44.75 ± 1.58
Stage 1: Mean-pooled sets	14.56 ± 0.77	3.20 ± 0.79	25.57 ± 2.85	43.49 ± 1.48
Stage 2: w/o set	9.65 ± 0.27	0.91 ± 0.79	14.61 ± 2.85	29.22 ± 0.79
Stage 2: w/o atomic	15.93 ± 1.90	3.20 ± 1.58	24.66 ± 3.62	39.27 ± 2.85
Stage 2: full	18.18 ± 1.37	5.02 ± 0.79	30.14 ± 2.37	47.95 ± 1.37
Stage 3: w/o set	17.61 ± 1.18	5.02 ± 0.79	29.22 ± 2.85	46.58 ± 4.94
Stage 3: w/o atomic	19.42 ± 1.31	6.85 ± 2.37	30.14 ± 1.37	42.47 ± 0.00
Stage 3: w/o replay	11.75 ± 0.54	2.28 ± 1.58	14.61 ± 2.09	36.99 ± 1.37
Full GestaltCLIP	$\mathbf{20.69 \pm 1.44}$	$\mathbf{9.13 \pm 2.09}$	$\mathbf{31.51 \pm 0.00}$	$\mathbf{48.40 \pm 1.58}$

PATIENT-LEVEL CONCEPT SET RETRIEVAL

Table D.5 evaluates retrieval of each patient's exact observed facial-HPO concept set. Each query therefore has one unique patient profile as the positive target. This criterion is stricter than disease-level phenotype matching because patients with the same diagnosis may have different observed phenotype sets.

Table D.6: Ablation results for cross-dataset external evaluation on RDFace. **Bold** and underline denote the best and second-best values.

Variant	Top-1 (%)	Top-5 (%)	Bal. Acc. (%)	Macro-F1 (%)
Raw concept space	12.39 ± 0.21	25.89 ± 0.89	12.98 ± 1.57	17.81 ± 1.85
Stage 1: Mean-pooled sets	12.50 ± 1.79	24.70 ± 0.52	12.98 ± 1.63	17.44 ± 1.79
Stage 2: w/o set	13.39 ± 0.89	25.30 ± 0.52	14.73 ± 0.81	$\underline{20.60 \pm 0.74}$
Stage 2: w/o atomic	13.69 ± 0.52	$\underline{28.57 \pm 1.55}$	14.80 ± 0.80	18.86 ± 0.33
Stage 2: full	13.99 ± 0.52	28.57 ± 0.89	14.53 ± 1.01	18.89 ± 1.44
Stage 3: w/o set	$\underline{14.29 \pm 0.00}$	26.79 ± 0.00	$\underline{15.60 \pm 0.00}$	20.51 ± 0.00
Stage 3: w/o atomic	13.39 ± 0.89	26.19 ± 2.87	13.60 ± 1.74	17.79 ± 2.48
Stage 3: w/o replay	12.80 ± 0.52	25.89 ± 2.36	13.89 ± 1.26	19.28 ± 1.72
Full GestaltCLIP	$\mathbf{15.18 \pm 1.55}$	$\mathbf{29.17 \pm 1.36}$	$\mathbf{15.78 \pm 0.89}$	$\mathbf{20.73 \pm 1.14}$

RDFace cross-dataset transfer

Table D.6 examines whether the effects observed on GMDB persist under external transfer to RDFace. This evaluation is particularly useful for distinguishing components that improve the development domain from those whose effects survive a dataset shift.

RDFace unseen disease recognition

Table D.7 evaluates compositional generalization when target diseases are not visually observed during representation learning and prediction is performed through their concept profiles. Unlike the closed-set evaluations, individual ablations can trade off exact recognition, class balance, and candidate-list performance, so all four metrics are retained.

Table D.7: Ablation results for RDFace unseen disease recognition. **Bold** and underline denote the best and second-best values.

Variant	Top-1 (%)	Top-5 (%)	Bal. Acc. (%)	Macro-F1 (%)
Raw concept space	3.96 ± 0.47	18.03 ± 1.08	3.86 ± 0.06	1.15 ± 0.20
Stage 1: Mean-pooled sets	4.10 ± 0.41	$\underline{20.49 \pm 1.64}$	4.40 ± 0.68	1.26 ± 0.25
Stage 2: w/o set	4.92 ± 0.41	13.80 ± 0.85	4.82 ± 0.36	1.35 ± 0.26
Stage 2: w/o atomic	3.96 ± 0.47	17.90 ± 0.85	3.56 ± 0.53	2.16 ± 0.82
Stage 2: full	4.51 ± 1.08	18.99 ± 1.18	3.93 ± 0.94	2.58 ± 0.23
Stage 3: w/o set	$\underline{5.05 \pm 0.47}$	18.31 ± 2.02	$\underline{4.84 \pm 0.41}$	2.56 ± 0.15
Stage 3: w/o atomic	4.78 ± 1.18	20.22 ± 1.03	4.31 ± 1.36	$\mathbf{3.23 \pm 0.84}$
Stage 3: w/o replay	3.69 ± 0.41	13.93 ± 0.41	3.75 ± 0.36	1.29 ± 0.48
Full GestaltCLIP	$\mathbf{5.19 \pm 0.63}$	$\mathbf{21.04 \pm 0.63}$	$\mathbf{5.89 \pm 0.93}$	$\underline{2.72 \pm 0.57}$

D.2 Additional Ablation Analyses

In this section, we provide additional fine-grained ablation analyses that complement the main ablation studies in the paper. The main experiments focus on the major stage-level design choices of GestaltCLIP, while the analyses here further decompose the internal objectives and evidence sources within each stage.

For ontology-aware phenotype concept space, we analyze the effects of the two loss terms ($\mathcal{L}_{\text{geo}}$ and $\mathcal{L}_{\text{hier}}$). For Stage 2, we analyze the effects of the four loss terms ($\mathcal{L}_{\text{NCE}}$, $\mathcal{L}_{\text{PU}}$, $\mathcal{L}_{\text{set}}$, and $\mathcal{L}_{\text{reg}}$). For Stage 3, we analyze the effects of evidence (visual, prior, or structural) to produce disease-mediated labels for unpaired images. The task settings are the same as in the main paper. All results are reported as mean $\pm$ SD across seeds and are expressed as percentages. Bold and underlined values denote the best and second-best mean results, respectively, within each partition and metric.

Table D.8: Ablation results for frozen-linear disease classification on GMDB. **Bold** and underline denote the best and second-best values.

Variant	Top-1 (%)	Top-5 (%)	Bal. Acc. (%)	Macro-F1 (%)
Common diseases				
Upstream: $\mathcal{L}_{\text{geo}}$ only	34.61 ± 0.46	60.11 ± 0.11	24.14 ± 0.37	24.89 ± 0.39
Upstream: $\mathcal{L}_{\text{hier}}$ only	34.72 ± 0.56	60.02 ± 0.44	24.21 ± 0.34	25.02 ± 0.40
Stage 2: w/o $\mathcal{L}_{\text{NCE}}$	33.99 ± 0.00	58.37 ± 0.06	23.85 ± 0.32	24.50 ± 0.33
Stage 2: w/o $\mathcal{L}_{\text{PU}}$	$\mathbf{35.85 \pm 0.09}$	60.14 ± 0.29	$\underline{25.73 \pm 0.26}$	$\underline{26.44 \pm 0.10}$
Stage 2: w/o $\mathcal{L}_{\text{set}}$	35.73 ± 0.27	60.09 ± 0.16	25.37 ± 0.40	26.12 ± 0.36
Stage 2: w/o $\mathcal{L}_{\text{reg}}$	$\underline{35.74 \pm 0.24}$	60.05 ± 0.19	25.31 ± 0.28	26.12 ± 0.36
Stage 3: visual only	34.77 ± 0.43	$\underline{60.50 \pm 0.33}$	24.67 ± 0.59	25.59 ± 0.47
Stage 3: prior only	34.66 ± 0.08	$\mathbf{60.51 \pm 0.43}$	24.54 ± 0.50	25.46 ± 0.41
Stage 3: structure only	34.66 ± 0.38	60.23 ± 0.36	24.60 ± 0.47	25.51 ± 0.26
Full GestaltCLIP	$\mathbf{35.85 \pm 0.09}$	60.20 ± 0.63	$\mathbf{25.93 \pm 0.68}$	$\mathbf{26.78 \pm 0.82}$
Rare diseases				
Upstream: $\mathcal{L}_{\text{geo}}$ only	26.04 ± 1.63	43.49 ± 0.81	26.63 ± 1.76	21.44 ± 0.89
Upstream: $\mathcal{L}_{\text{hier}}$ only	25.39 ± 1.03	43.23 ± 1.26	26.08 ± 1.10	20.80 ± 0.79
Stage 2: w/o $\mathcal{L}_{\text{NCE}}$	25.13 ± 0.60	40.89 ± 0.60	24.63 ± 0.94	19.17 ± 0.40
Stage 2: w/o $\mathcal{L}_{\text{PU}}$	25.91 ± 1.63	42.58 ± 1.17	25.23 ± 1.57	19.61 ± 0.74
Stage 2: w/o $\mathcal{L}_{\text{set}}$	26.43 ± 1.58	43.88 ± 1.19	25.73 ± 1.43	19.99 ± 1.12
Stage 2: w/o $\mathcal{L}_{\text{reg}}$	$\underline{26.56 \pm 1.41}$	$\underline{44.40 \pm 0.81}$	25.65 ± 1.25	19.86 ± 1.08
Stage 3: visual only	25.26 ± 0.90	$\mathbf{44.66 \pm 1.80}$	26.13 ± 0.97	21.04 ± 0.57
Stage 3: prior only	26.43 ± 0.23	43.23 ± 1.48	25.81 ± 0.55	19.87 ± 0.72
Stage 3: structure only	25.91 ± 0.45	44.27 ± 1.80	$\underline{26.79 \pm 0.49}$	$\underline{21.62 \pm 0.14}$
Full GestaltCLIP	$\mathbf{26.95 \pm 1.03}$	$\underline{44.40 \pm 1.97}$	$\mathbf{26.97 \pm 1.19}$	$\mathbf{21.68 \pm 0.36}$

D.3 Full Ablation Results

Disease classification

Table D.8 reports the frozen-linear disease-classification ablation on the GMDB. The full model achieves the best Top-1, Bal. Acc., and Macro-F1 on both *common* and *rare* diseases, although the highest Top-5 accuracy is obtained by the Stage 3 prior-only variant on *common* diseases and the visual-only variant on *rare* diseases. Among the Stage 2 objectives, removing $\mathcal{L}_{\text{NCE}}$ produces the largest overall degradation, while removing $\mathcal{L}_{\text{PU}}$ has a comparatively smaller effect on *common* disease classification.

Disease image retrieval

Table D.9 reports disease-level image retrieval using the frozen representations. On *common* diseases, the full model achieves the highest mAP and R@1 and remains competitive at larger retrieval depths, while the structure-only Stage 3 variant gives the highest R@10. On *rare* diseases, the full model achieves the best R@1 and R@5 and ties for the highest mAP, whereas removing $\mathcal{L}_{\text{reg}}$ gives the highest R@10. The pronounced drop after removing $\mathcal{L}_{\text{NCE}}$ indicates that multi-positive alignment is particularly important for preserving disease-level neighborhood structure.

Atomic HPO concept retrieval

Table D.10 reports direct atomic HPO concept retrieval against the fixed concept bank. The full model achieves the highest mAP and R@1 on both *common* and *rare* diseases, and also reaches the best R@5 on *common* diseases. However, the strongest deeper-rank recall is not always obtained by the full model: the hierarchy-only upstream variant slightly exceeds it at *common* disease R@10, while geometry-only and structure-only variants remain competitive on *rare* disease R@5 and R@10. Removing $\mathcal{L}_{\text{PU}}$ causes the largest degradation, especially for *rare* diseases, showing that the PU objective is important for maintaining direct atomic phenotype grounding.

Table D.9: Ablation results for disease image-to-image retrieval on GMDB. **Bold** and underline denote the best and second-best values.

Variant	mAP (%)	R@1 (%)	R@5 (%)	R@10 (%)
Common diseases				
Upstream: $\mathcal{L}_{\text{geo}}$ only	6.47 ± 0.03	24.43 ± 0.45	43.67 ± 0.19	54.27 ± 0.11
Upstream: $\mathcal{L}_{\text{hier}}$ only	6.45 ± 0.06	24.52 ± 0.24	43.63 ± 0.38	53.88 ± 0.24
Stage 2: w/o $\mathcal{L}_{\text{NCE}}$	4.58 ± 0.08	20.78 ± 0.29	40.32 ± 0.42	50.02 ± 0.58
Stage 2: w/o $\mathcal{L}_{\text{PU}}$	6.11 ± 0.26	24.82 ± 0.51	44.41 ± 0.88	54.61 ± 0.70
Stage 2: w/o $\mathcal{L}_{\text{set}}$	6.27 ± 0.24	25.05 ± 0.24	45.05 ± 0.74	55.07 ± 0.37
Stage 2: w/o $\mathcal{L}_{\text{reg}}$	6.11 ± 0.02	23.85 ± 0.62	42.71 ± 0.56	52.89 ± 0.16
Stage 3: visual only	6.36 ± 0.09	25.07 ± 0.08	$\mathbf{45.35 \pm 0.25}$	$\underline{55.23 \pm 0.11}$
Stage 3: prior only	$\underline{6.66 \pm 0.02}$	24.96 ± 0.29	44.40 ± 0.31	54.93 ± 0.33
Stage 3: structure only	6.61 ± 0.04	$\underline{25.20 \pm 0.24}$	44.66 ± 0.03	$\mathbf{55.25 \pm 0.43}$
Full GestaltCLIP	$\mathbf{6.77 \pm 0.17}$	$\mathbf{25.53 \pm 0.35}$	$\underline{45.28 \pm 0.27}$	54.65 ± 0.08
Rare diseases				
Upstream: $\mathcal{L}_{\text{geo}}$ only	18.18 ± 0.27	20.96 ± 0.90	35.94 ± 1.03	44.53 ± 1.41
Upstream: $\mathcal{L}_{\text{hier}}$ only	17.99 ± 0.24	20.44 ± 0.98	34.90 ± 0.60	44.40 ± 1.80
Stage 2: w/o $\mathcal{L}_{\text{NCE}}$	15.78 ± 0.10	20.44 ± 0.23	32.16 ± 0.90	39.84 ± 1.17
Stage 2: w/o $\mathcal{L}_{\text{PU}}$	18.61 ± 0.59	22.40 ± 0.98	39.32 ± 1.85	46.48 ± 1.41
Stage 2: w/o $\mathcal{L}_{\text{set}}$	$\underline{19.33 \pm 0.58}$	22.92 ± 0.81	39.97 ± 1.13	$\underline{47.27 \pm 0.39}$
Stage 2: w/o $\mathcal{L}_{\text{reg}}$	$\mathbf{19.57 \pm 0.24}$	$\underline{23.05 \pm 0.68}$	$\underline{40.62 \pm 0.39}$	$\mathbf{47.66 \pm 0.39}$
Stage 3: visual only	18.59 ± 0.44	21.48 ± 1.03	35.55 ± 0.39	45.70 ± 0.39
Stage 3: prior only	18.93 ± 0.50	22.14 ± 1.26	36.98 ± 0.23	47.01 ± 1.19
Stage 3: structure only	18.78 ± 0.17	22.27 ± 0.00	35.81 ± 1.37	45.70 ± 1.79
Full GestaltCLIP	$\mathbf{19.57 \pm 0.25}$	$\mathbf{23.31 \pm 0.81}$	$\mathbf{41.02 \pm 0.00}$	46.48 ± 1.03

Table D.10: Ablation results for atomic HPO concept retrieval on GMDB. **Bold** and underline denote the best and second-best values.

Variant	mAP (%)	R@1 (%)	R@5 (%)	R@10 (%)
Common diseases				
Upstream: $\mathcal{L}_{\text{geo}}$ only	$\underline{12.56 \pm 0.30}$	5.26 ± 0.31	$\underline{11.95 \pm 0.30}$	16.81 ± 0.33
Upstream: $\mathcal{L}_{\text{hier}}$ only	12.55 ± 0.43	5.26 ± 0.36	$\mathbf{12.23 \pm 0.39}$	$\mathbf{17.11 \pm 0.51}$
Stage 2: w/o $\mathcal{L}_{\text{NCE}}$	7.42 ± 0.23	1.24 ± 0.06	6.04 ± 0.23	10.71 ± 0.46
Stage 2: w/o $\mathcal{L}_{\text{PU}}$	6.78 ± 0.11	2.80 ± 0.14	6.71 ± 0.46	9.56 ± 0.25
Stage 2: w/o $\mathcal{L}_{\text{set}}$	9.25 ± 0.16	3.54 ± 0.09	8.95 ± 0.07	12.29 ± 0.18
Stage 2: w/o $\mathcal{L}_{\text{reg}}$	9.38 ± 0.08	3.69 ± 0.11	9.22 ± 0.18	12.83 ± 0.10
Stage 3: visual only	11.98 ± 0.28	4.95 ± 0.32	11.63 ± 0.26	16.34 ± 0.39
Stage 3: prior only	12.20 ± 0.16	$\underline{5.30 \pm 0.23}$	11.34 ± 0.37	16.15 ± 0.27
Stage 3: structure only	12.47 ± 0.05	5.08 ± 0.18	11.92 ± 0.06	17.00 ± 0.29
Full GestaltCLIP	$\mathbf{13.07 \pm 0.16}$	$\mathbf{5.66 \pm 0.26}$	$\mathbf{12.23 \pm 0.14}$	$\underline{17.04 \pm 0.09}$
Rare diseases				
Upstream: $\mathcal{L}_{\text{geo}}$ only	10.52 ± 0.20	2.86 ± 0.40	$\mathbf{11.90 \pm 0.20}$	$\underline{15.37 \pm 0.29}$
Upstream: $\mathcal{L}_{\text{hier}}$ only	10.24 ± 0.28	2.96 ± 0.26	11.02 ± 0.67	15.19 ± 0.26
Stage 2: w/o $\mathcal{L}_{\text{NCE}}$	8.22 ± 0.26	2.08 ± 0.12	6.53 ± 0.49	10.47 ± 1.75
Stage 2: w/o $\mathcal{L}_{\text{PU}}$	3.80 ± 0.21	1.05 ± 0.15	2.64 ± 0.38	4.80 ± 0.66
Stage 2: w/o $\mathcal{L}_{\text{set}}$	6.76 ± 0.26	1.85 ± 0.14	6.11 ± 0.55	9.65 ± 0.70
Stage 2: w/o $\mathcal{L}_{\text{reg}}$	6.93 ± 0.27	1.64 ± 0.18	7.33 ± 1.45	10.75 ± 1.67
Stage 3: visual only	9.57 ± 0.17	2.91 ± 0.55	9.74 ± 0.19	$\mathbf{15.42 \pm 0.36}$
Stage 3: prior only	9.66 ± 0.13	$\underline{3.05 \pm 0.45}$	10.01 ± 0.55	14.53 ± 0.63
Stage 3: structure only	$\underline{10.65 \pm 0.11}$	3.03 ± 0.23	$\underline{11.68 \pm 0.63}$	15.04 ± 0.25
Full GestaltCLIP	$\mathbf{11.91 \pm 0.16}$	$\mathbf{5.04 \pm 0.21}$	11.56 ± 0.59	15.35 ± 0.62

HPO CONCEPT LINEAR PROBING

Table D.11 evaluates phenotype information recoverable from the frozen visual representation with a fitted multilabel linear probe. On *common* diseases, the full model obtains the best R@1 and ties for the best mAP, R@5, and R@10, indicating that the learned representation retains phenotype information that is readily accessible to a simple downstream classifier. On *rare* diseases, the full model achieves the best mAP and Macro-F1, whereas the upstream concept-space variants obtain slightly higher R@1 or R@5. This difference from direct atomic retrieval suggests that preserving

Table D.11: Ablation results for frozen HPO concept linear probing on GMDB. **Bold** and underline denote the best and second-best values.

Variant	mAP (%)	R@1 (%)	R@5 (%)	R@10 (%)
Common diseases				
Upstream: $\mathcal{L}_{\text{geo}}$ only	25.53 ± 0.01	6.82 ± 0.16	20.92 ± 0.32	32.08 ± 0.21
Upstream: $\mathcal{L}_{\text{hier}}$ only	25.61 ± 0.04	7.09 ± 0.09	20.77 ± 0.13	32.02 ± 0.21
Stage 2: w/o $\mathcal{L}_{\text{NCE}}$	24.86 ± 0.13	6.57 ± 0.02	20.83 ± 0.07	31.99 ± 0.24
Stage 2: w/o $\mathcal{L}_{\text{PU}}$	25.20 ± 0.07	$\underline{7.17 \pm 0.25}$	20.69 ± 0.24	31.77 ± 0.60
Stage 2: w/o $\mathcal{L}_{\text{set}}$	$\mathbf{25.89 \pm 0.02}$	7.10 ± 0.08	$\underline{21.41 \pm 0.11}$	$\underline{32.72 \pm 0.24}$
Stage 2: w/o $\mathcal{L}_{\text{reg}}$	$\underline{25.78 \pm 0.07}$	6.95 ± 0.18	21.33 ± 0.14	32.63 ± 0.36
Stage 3: visual only	25.62 ± 0.04	6.95 ± 0.11	20.94 ± 0.16	32.30 ± 0.03
Stage 3: prior only	25.47 ± 0.11	6.79 ± 0.11	20.88 ± 0.08	31.97 ± 0.03
Stage 3: structure only	25.51 ± 0.08	6.91 ± 0.25	20.91 ± 0.06	31.99 ± 0.28
Full GestaltCLIP	$\mathbf{25.89 \pm 0.15}$	$\mathbf{7.38 \pm 0.27}$	$\mathbf{21.46 \pm 0.21}$	$\mathbf{32.88 \pm 0.29}$
Rare diseases				
Upstream: $\mathcal{L}_{\text{geo}}$ only	$\underline{20.42 \pm 0.27}$	$\underline{5.12 \pm 0.00}$	$\mathbf{17.30 \pm 0.06}$	24.78 ± 0.38
Upstream: $\mathcal{L}_{\text{hier}}$ only	20.35 ± 0.29	$\mathbf{5.18 \pm 0.11}$	17.02 ± 0.42	24.85 ± 0.25
Stage 2: w/o $\mathcal{L}_{\text{NCE}}$	16.51 ± 0.06	1.79 ± 0.12	16.40 ± 0.09	21.53 ± 0.11
Stage 2: w/o $\mathcal{L}_{\text{PU}}$	19.77 ± 0.61	3.93 ± 0.28	16.33 ± 1.13	26.39 ± 1.04
Stage 2: w/o $\mathcal{L}_{\text{set}}$	19.30 ± 0.65	3.92 ± 0.39	15.66 ± 1.23	$\underline{26.87 \pm 0.43}$
Stage 2: w/o $\mathcal{L}_{\text{reg}}$	19.38 ± 0.09	3.93 ± 0.00	15.73 ± 0.25	26.83 ± 0.82
Stage 3: visual only	20.35 ± 0.37	4.89 ± 0.36	16.19 ± 0.14	25.35 ± 0.82
Stage 3: prior only	20.13 ± 0.09	4.77 ± 0.32	16.67 ± 0.77	25.59 ± 1.08
Stage 3: structure only	20.20 ± 0.21	4.75 ± 0.23	16.53 ± 0.11	25.66 ± 1.21
Full GestaltCLIP	$\mathbf{21.04 \pm 0.47}$	$\underline{5.12 \pm 0.00}$	$\underline{17.08 \pm 0.24}$	$\mathbf{28.38 \pm 0.22}$

decodable phenotype information and preserving exact cross-modal geometry are related but distinct properties of the representation.

Patient-level concept set retrieval

Table D.12 evaluates retrieval of each patient's exact observed facial-HPO profile. On *common* diseases, the full model achieves the highest mAP, R@5, and R@10, with only R@1 slightly higher when $\mathcal{L}_{\text{set}}$ is removed. On *rare* diseases, the full model achieves the highest mAP, R@1, and R@10, while the variant without $\mathcal{L}_{\text{reg}}$ gives the highest R@5. The especially large gains over the variant without $\mathcal{L}_{\text{NCE}}$ show that the paired multi-positive objective is important not only for atomic concepts but also for recovering complete patient-specific phenotype profiles.

RDFace cross-dataset transfer

Table D.13 reports external transfer to RDFace for diseases shared with GMDB. The full model achieves the best Top-1 accuracy, Bal. Acc., and Macro-F1, while the hierarchy-only upstream variant gives the highest Top-5 accuracy. The gains of the full model are therefore most consistent for exact recognition rather than candidate-list recall alone, suggesting that the combined representation remains useful under cross-dataset shift.

RDFace unseen disease recognition

Table D.14 evaluates transfer to diseases whose facial images are not observed during representation learning. The full model achieves the highest Bal. Acc. and the second-highest Top-1 and Top-5 accuracy, but does not dominate every metric: the structure-only variant gives the highest Top-1, removing $\mathcal{L}_{\text{reg}}$ gives the highest Top-5, and the prior-only variant gives the highest Macro-F1. These results indicate that unseen disease transfer depends on several complementary properties of the learned space, with the full model providing the strongest class-balanced performance rather than uniformly optimizing every ranking metric.

Table D.12: Ablation results for exact patient-level concept set retrieval on GMDB. **Bold** and underline denote the best and second-best values.

Variant	mAP (%)	R@1 (%)	R@5 (%)	R@10 (%)
Common diseases				
Upstream: $\mathcal{L}_{\text{geo}}$ only	4.73 ± 0.24	1.42 ± 0.19	5.51 ± 0.55	9.98 ± 0.71
Upstream: $\mathcal{L}_{\text{hier}}$ only	4.57 ± 0.24	1.25 ± 0.22	5.30 ± 0.19	9.82 ± 0.36
Stage 2: w/o $\mathcal{L}_{\text{NCE}}$	2.06 ± 0.28	0.33 ± 0.26	2.09 ± 0.19	3.84 ± 0.52
Stage 2: w/o $\mathcal{L}_{\text{PU}}$	4.68 ± 0.22	1.29 ± 0.32	5.64 ± 0.66	10.48 ± 1.34
Stage 2: w/o $\mathcal{L}_{\text{set}}$	5.38 ± 0.26	$\mathbf{1.88 \pm 0.13}$	6.35 ± 0.69	11.53 ± 0.99
Stage 2: w/o $\mathcal{L}_{\text{reg}}$	5.44 ± 0.17	1.63 ± 0.00	6.89 ± 0.22	12.24 ± 0.85
Stage 3: visual only	5.13 ± 0.20	1.25 ± 0.13	6.56 ± 0.32	11.95 ± 0.59
Stage 3: prior only	5.46 ± 0.22	1.71 ± 0.29	6.73 ± 0.32	12.24 ± 0.07
Stage 3: structure only	5.14 ± 0.16	1.80 ± 0.07	6.02 ± 0.33	10.78 ± 0.45
Full GestaltCLIP	$\mathbf{6.05 \pm 0.41}$	1.84 ± 0.29	$\mathbf{7.77 \pm 0.70}$	$\mathbf{13.87 \pm 0.82}$
Rare diseases				
Upstream: $\mathcal{L}_{\text{geo}}$ only	18.20 ± 1.47	5.02 ± 0.79	31.05 ± 3.45	47.49 ± 0.79
Upstream: $\mathcal{L}_{\text{hier}}$ only	18.48 ± 1.41	5.48 ± 1.37	31.51 ± 2.37	46.12 ± 0.79
Stage 2: w/o $\mathcal{L}_{\text{NCE}}$	9.97 ± 0.41	0.46 ± 0.79	14.16 ± 4.40	30.14 ± 1.37
Stage 2: w/o $\mathcal{L}_{\text{PU}}$	15.93 ± 1.90	3.20 ± 1.58	24.66 ± 3.62	39.27 ± 2.85
Stage 2: w/o $\mathcal{L}_{\text{set}}$	18.75 ± 1.78	6.39 ± 1.58	29.22 ± 3.45	43.38 ± 3.45
Stage 2: w/o $\mathcal{L}_{\text{reg}}$	19.38 ± 1.41	6.39 ± 1.58	$\mathbf{32.88 \pm 3.62}$	42.92 ± 0.79
Stage 3: visual only	19.57 ± 0.34	7.31 ± 0.79	30.14 ± 2.74	44.29 ± 2.85
Stage 3: prior only	18.70 ± 0.83	5.02 ± 0.79	29.22 ± 1.58	46.12 ± 2.85
Stage 3: structure only	17.56 ± 1.71	4.57 ± 1.58	30.59 ± 2.85	47.03 ± 2.85
Full GestaltCLIP	$\mathbf{20.69 \pm 1.44}$	$\mathbf{9.13 \pm 2.09}$	31.51 ± 0.00	$\mathbf{48.40 \pm 1.58}$

Table D.13: Ablation results for cross-dataset external evaluation on RDFace. **Bold** and underline denote the best and second-best values.

Variant	Top-1 (%)	Top-5 (%)	Bal. Acc. (%)	Macro-F1 (%)
Upstream: $\mathcal{L}_{\text{geo}}$ only	14.29 ± 0.89	28.57 ± 1.55	14.58 ± 1.08	19.75 ± 1.82
Upstream: $\mathcal{L}_{\text{hier}}$ only	13.69 ± 0.52	$\mathbf{30.06 \pm 0.52}$	14.04 ± 0.41	18.72 ± 0.51
Stage 2: w/o $\mathcal{L}_{\text{NCE}}$	13.39 ± 0.89	26.49 ± 1.03	14.73 ± 0.81	20.17 ± 0.87
Stage 2: w/o $\mathcal{L}_{\text{PU}}$	14.29 ± 0.00	27.38 ± 1.03	15.60 ± 0.00	20.60 ± 0.08
Stage 2: w/o $\mathcal{L}_{\text{set}}$	14.29 ± 0.00	27.38 ± 0.52	15.60 ± 0.00	20.60 ± 0.08
Stage 2: w/o $\mathcal{L}_{\text{reg}}$	14.29 ± 0.00	27.38 ± 0.52	15.60 ± 0.00	20.56 ± 0.08
Stage 3: visual only	13.39 ± 0.89	28.27 ± 1.03	13.96 ± 1.37	18.69 ± 1.61
Stage 3: prior only	13.69 ± 1.36	28.57 ± 0.89	14.22 ± 1.82	18.92 ± 1.99
Stage 3: structure only	13.69 ± 1.36	28.57 ± 0.89	14.22 ± 1.82	18.92 ± 1.99
Full GestaltCLIP	$\mathbf{15.18 \pm 1.55}$	29.17 ± 1.36	$\mathbf{15.78 \pm 0.89}$	$\mathbf{20.73 \pm 1.14}$

Table D.14: Ablation results for RDFace unseen disease recognition. **Bold** and underline denote the best and second-best values.

Variant	Top-1 (%)	Top-5 (%)	Bal. Acc. (%)	Macro-F1 (%)
Upstream: $\mathcal{L}_{\text{geo}}$ only	5.05 ± 0.63	18.72 ± 1.32	4.84 ± 0.55	2.55 ± 0.23
Upstream: $\mathcal{L}_{\text{hier}}$ only	5.05 ± 1.03	18.85 ± 1.48	4.51 ± 0.95	2.71 ± 0.53
Stage 2: w/o $\mathcal{L}_{\text{NCE}}$	4.10 ± 0.00	14.07 ± 1.03	3.93 ± 0.31	0.79 ± 0.05
Stage 2: w/o $\mathcal{L}_{\text{PU}}$	3.96 ± 0.47	17.90 ± 0.85	3.56 ± 0.53	2.16 ± 0.82
Stage 2: w/o $\mathcal{L}_{\text{set}}$	4.23 ± 0.63	20.77 ± 0.85	3.90 ± 0.61	2.64 ± 0.50
Stage 2: w/o $\mathcal{L}_{\text{reg}}$	4.51 ± 1.08	$\mathbf{21.58 \pm 0.63}$	4.11 ± 1.09	2.73 ± 0.59
Stage 3: visual only	5.05 ± 1.03	20.90 ± 1.48	4.90 ± 1.00	2.79 ± 0.61
Stage 3: prior only	4.78 ± 0.85	20.49 ± 2.28	4.66 ± 0.85	$\mathbf{2.82 \pm 0.61}$
Stage 3: structure only	$\mathbf{5.33 \pm 1.08}$	18.72 ± 1.03	5.50 ± 1.27	2.80 ± 0.70
Full GestaltCLIP	5.19 ± 0.63	21.04 ± 0.63	$\mathbf{5.89 \pm 0.93}$	2.72 ± 0.57

E Analysis of the Learned Representation

This section provides additional analyses of the learned representation, including visualization of the embedding space, examination of phenotype structure in the visual space, and comparison of performance on *common* versus *rare* diseases.

E.1 HPO–HPO Concept Geometry

Table E.1: Intrinsic evaluation of HPO concept geometry before and after ontology-aware refinement on held-out ontology nodes.

Concept representation	LIN Spearman ↑	Hierarchy accuracy ↑	Semantic preservation
Raw concept bank	0.8399	0.9488	–
+ Ontology-aware refinement	**0.8677**	**0.9886**	0.9942

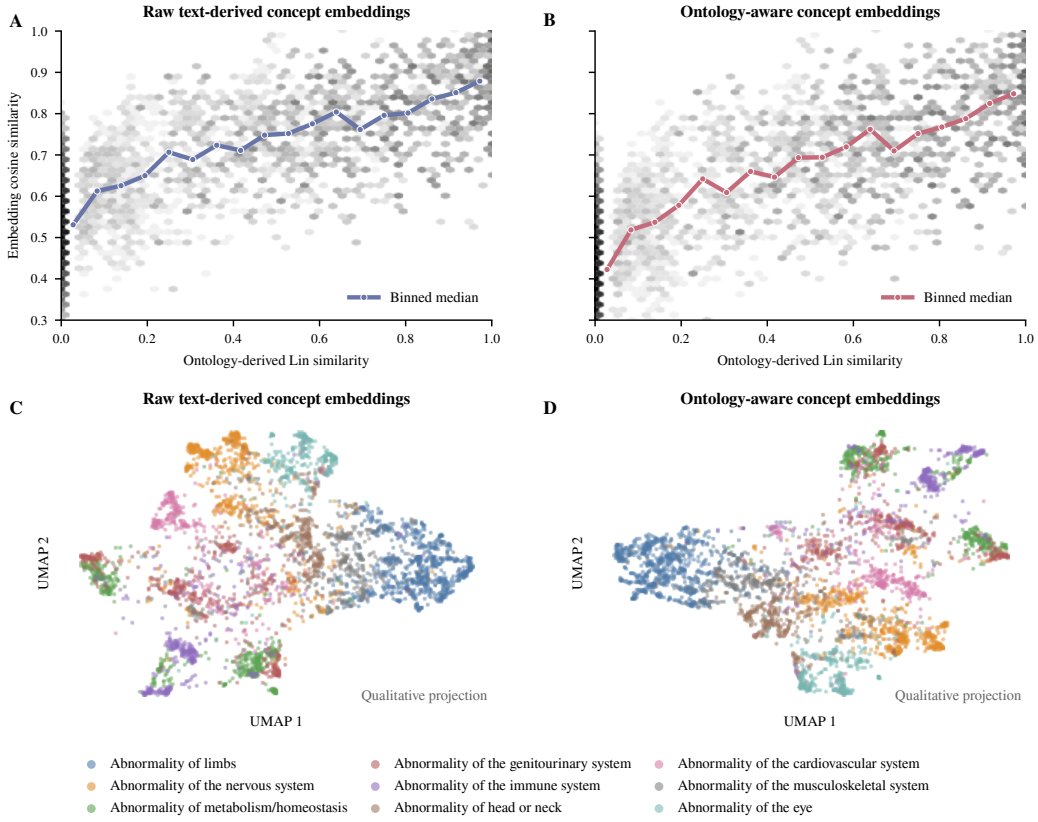


Figure E.1: **HPO concept geometry before and after ontology-aware refinement. (A–B)** Pairwise embedding similarity as a function of ontology-derived Lin similarity, showing increased separation of ontology-aware concepts after refinement. **(C–D)** Qualitative UMAP projections colored by top-level HPO phenotype category.

We first evaluate whether ontology-aware refinement improves the intrinsic geometry of the HPO concept space. Table E.1 compares the raw text-derived concept bank with the ontology-refined representation on held-out ontology nodes. LIN Spearman measures the Spearman rank correlation between pairwise cosine similarity in the learned HPO embedding space and ontology-derived Lin semantic similarity across HPO concept pairs. Hierarchy triplet accuracy evaluates whether ontology-near concepts are ranked consistently with the HPO hierarchy. Semantic preservation measures how strongly the refined concepts retain the original text-derived semantics.

Ontology-aware refinement consistently improves the intrinsic HPO concept geometry on held-out nodes. LIN Spearman increases from 0.8399 to 0.8677, while hierarchy triplet accuracy improves

from 0.9488 to 0.9886. At the same time, semantic preservation remains high at 0.9942, indicating that the refinement strengthens ontology consistency without substantially altering the original text-derived semantics.

Visualization of HPO concept geometry. Figure E.1 provides complementary views of the HPO concept space before and after ontology-aware refinement. In panels A–B, refinement increases the separation of ontology-distant concept pairs while preserving high similarity among ontology-near concepts, consistent with the improved Lin–cosine Spearman correlation in Table E.1. Panels C–D show the corresponding UMAP projections, where the refined embeddings exhibit a more structured global organization across major phenotype branches.

E.2 Image–HPO Alignment

We further examine the global organization of patient image embeddings and phenotype concepts in the shared representation space. This complements the main-text image–HPO similarity and qualitative grounding analyses by visualizing how patient embeddings are positioned relative to phenotype concept embeddings.

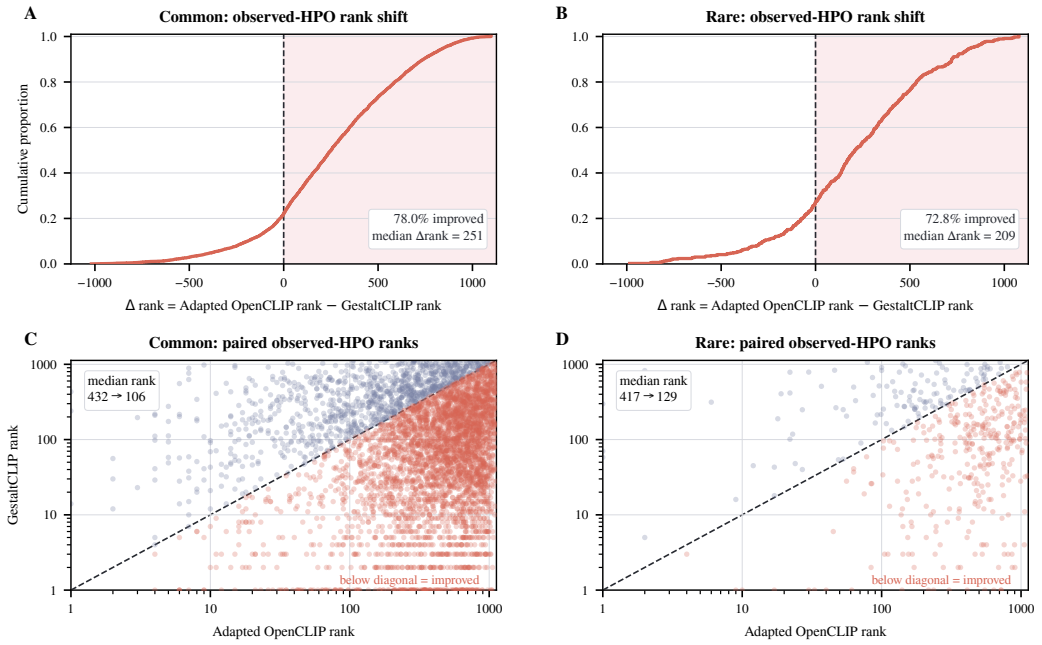


Figure E.2: **Observed-HPO rank shifts between adapted OpenCLIP and GestaltCLIP. (A–B)** Empirical cumulative distributions of the paired rank shift for *common* and *rare* diseases. Positive values indicate improved alignment. **(C–D)** Paired observed-HPO ranks between adapted OpenCLIP and GestaltCLIP on logarithmic axes, where points below the diagonal correspond to improved phenotype ranking.

We quantify how GestaltCLIP changes the retrieval rank of clinically observed HPO concepts relative to the adapted OpenCLIP baseline. For each observed image–HPO pair, we compute the rank shift as the adapted OpenCLIP rank minus the GestaltCLIP rank, such that positive values indicate improved alignment. As shown in Figure E.2, the majority of observed phenotype concepts move upward in the ranked HPO list in GestaltCLIP. 78.0% of pairs improve in the *common* partition and 72.8% improve in the *rare* partition, with median rank shifts of +251 and +209, respectively. Consistent with this paired analysis, the median observed-HPO rank decreases from 432 to 106 for *common cases* and from 417 to 129 for *rare cases*. These results indicate that GestaltCLIP's objective does not merely alter the global embedding geometry, but systematically brings patient images closer to their observed phenotype concepts in the shared representation space.

E.3 Image–Image Phenotype Geometry

Finally, we examine whether phenotype structure is reflected within the learned patient image representation itself. Specifically, we test whether patients with more similar phenotype profiles also have more similar image embeddings.

We first measure phenotype coherence among cross-disease image neighbors (Figure E.3). For each test query, we retrieve the top-K nearest images in the learned embedding space while excluding patients with the same disease label, and compare their observed HPO sets using Jaccard similarity. GestaltCLIP yields consistently higher neighborhood phenotype similarity across all evaluated K. This shows that the visual representation captures phenotype structure beyond disease identity.

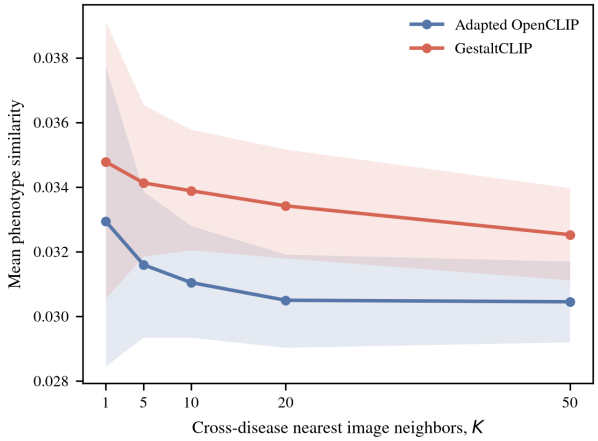


Figure E.3: **Cross-disease phenotype coherence of image nearest neighbors.** Mean Jaccard similarity between each query patient's observed HPO set and those of its top-K cross-disease image neighbors. Shaded regions denote 95% bootstrap confidence intervals over query patients.

Figure E.4 examines the per-patient consistency of this improvement. At $K = 5$, 10, 20, and 50, GestaltCLIP improves neighborhood phenotype coherence for 47.5%, 55.1%, 58.6%, and 59.4% of held-out patients. This shows that the aggregate gain is distributed across a substantial fraction of the cohort rather than being driven by a small number of queries.

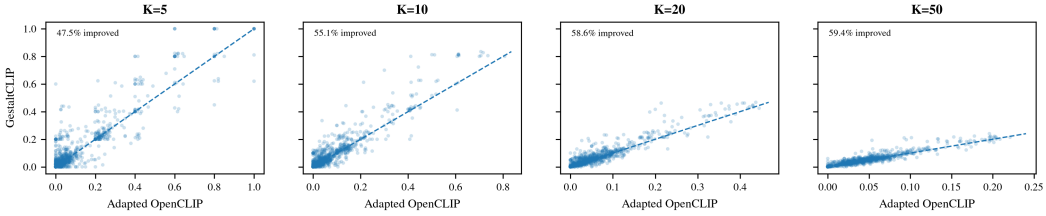


Figure E.4: **Paired patient-level comparison of neighborhood phenotype coherence.** Each point compares Adapted OpenCLIP and GestaltCLIP for the same query patient at a fixed K; points above the dashed identity line indicate improvement under GestaltCLIP.

F Robustness and Sensitivity Analyses

We examine whether the learned representations remain stable under changes to key modeling choices. Specifically, we study sensitivity to the phenotype text encoder, the Stage 1 HPO masking ratio, and the number of phenotype candidates retained during Stage 3 weak-label construction.

F.1 Text Encoder Sensitivity

To assess whether the learned HPO concept geometry depends strongly on the choice of text encoder, we repeat the intrinsic evaluation with multiple biomedical and vision–language text encoders. Table F.1 reports results on held-out ontology nodes.

LIN Spearman measures the agreement between learned embedding similarity and Lin semantic similarity, while hierarchy accuracy evaluates whether ontology-near concepts are ranked consistently with the HPO hierarchy. Semantic preservation measures how strongly the ontology-shaped representations retain the original text-encoder semantics. Base LIN Spearman denotes the corresponding similarity structure before ontology-aware refinement.

Across all encoders, ontology-aware refinement yields consistently high LIN Spearman correlation and hierarchy accuracy while preserving the original text semantics. The final concept geometry remains relatively stable across encoder choices, and LIN Spearman improves over the corresponding base representation for every encoder, suggesting that the ontology-aware refinement is not tied to a particular text backbone.

Table F.1: Intrinsic evaluation of HPO concept geometry on held-out ontology nodes. For BioMedCLIP and OpenCLIP, the base representations are taken after their native CLIP text projection layers and before ontology-aware refinement.

Text encoder	LIN Spearman ↑	Base LIN Spearman ↑	Hierarchy accuracy ↑	Semantic preservation ↑
BiomedBERT	0.8169	0.7245	0.9588	0.9938
BioLORD-2023	0.8701	0.8463	0.9904	0.9943
SapBERT	0.8677	0.8399	0.9886	0.9942
BioMedCLIP	0.8212	0.7905	0.9606	0.9939
OpenCLIP	0.8254	0.8068	0.9664	0.9942

F.2 Hyperparameter Sensitivity

The default hyperparameters were selected using the development data and fixed before evaluation on the held-out test sets; the alternative settings below are evaluated only to characterize sensitivity around these choices.

Figure F.1 summarizes the overall sensitivity pattern across downstream tasks, while Tables F.2 and F.3 provide the corresponding numerical results.

Stage 1 masking ratio sensitivity

Stage 1 learns concept set structure through masked-HPO prediction. To assess whether downstream performance depends strongly on the masking difficulty, we compare masking ratios of 15% and 50% while keeping the remaining training and evaluation protocol unchanged. Table F.2 summarizes the primary metric for each downstream evaluation.

The default 30% masking ratio provides the strongest overall downstream performance, particularly for phenotype-profile retrieval and the *common* and *rare* disease representation tasks. Nevertheless, the 15% and 50% variants remain broadly competitive on several classification and HPO-probing metrics, indicating that Stage 1 does not require a narrowly tuned masking ratio. Performance degrades more clearly at the two extremes for exact phenotype-profile retrieval, suggesting that an intermediate masking level provides a better balance between preserving sufficient phenotype context and requiring meaningful reconstruction.

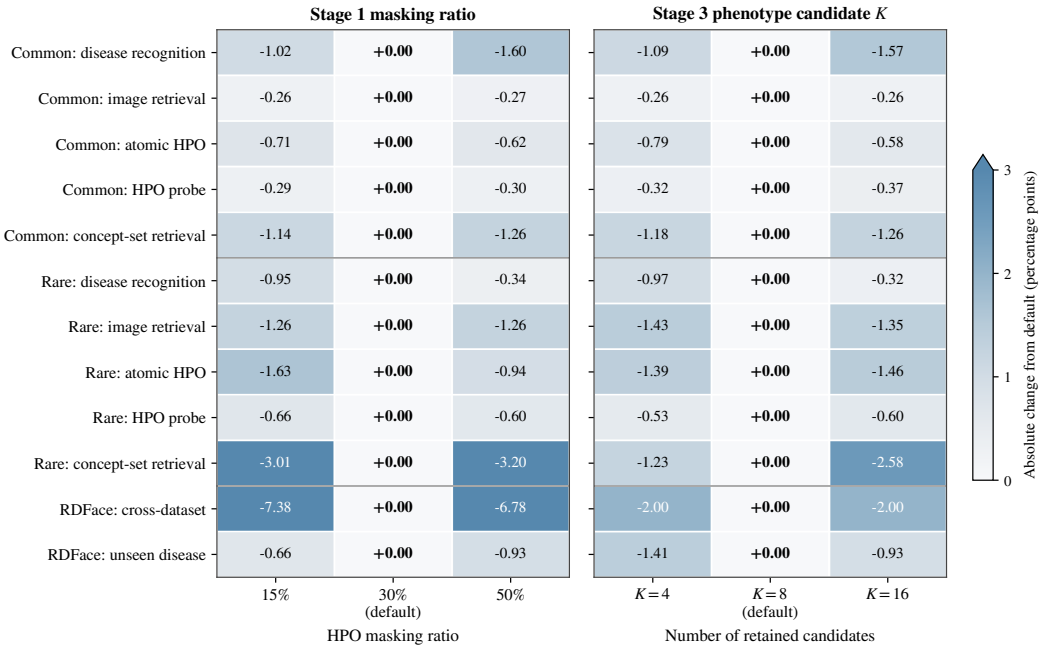


Figure F.1: Sensitivity of GestaltCLIP to the Stage 1 HPO masking ratio and the number of phenotype candidates retained in Stage 3. Each cell reports the change in the representative downstream metric, in percentage points, relative to the default configuration (30% masking for Stage 1 and $K = 8$ for Stage 3). Darker cells indicate larger absolute deviations from the default setting, while the numerical annotations preserve the signed performance differences.

Table F.2: Sensitivity to the Stage 1 HPO masking ratio. The default GestaltCLIP configuration uses a 30% masking ratio.

Dataset / task	Metric	15%	**30% (default)**	50%
GMDB *Common* disease recognition	Bal. Acc.	24.91 ± 0.38	**25.93 ± 0.68**	24.33 ± 0.34
GMDB *Common* image retrieval	mAP	6.51 ± 0.10	**6.77 ± 0.17**	6.50 ± 0.02
GMDB *Common* atomic HPO concept retrieval	mAP	12.36 ± 0.48	**13.07 ± 0.16**	12.45 ± 0.17
GMDB *Common* HPO concept linear probe	mAP	25.60 ± 0.04	**25.89 ± 0.15**	25.59 ± 0.11
GMDB *Common* concept set retrieval	mAP	4.91 ± 0.15	**6.05 ± 0.41**	4.79 ± 0.23
GMDB *Rare* disease recognition	Bal. Acc.	26.02 ± 0.57	**26.97 ± 1.19**	26.63 ± 2.13
GMDB *Rare* image retrieval	mAP	18.31 ± 0.55	**19.57 ± 0.25**	18.31 ± 0.36
GMDB *Rare* atomic HPO concept retrieval	mAP	10.28 ± 0.14	**11.91 ± 0.16**	10.97 ± 0.50
GMDB *Rare* HPO concept linear probe	mAP	20.38 ± 0.34	**21.04 ± 0.47**	20.44 ± 0.15
GMDB *Rare* concept set retrieval	mAP	17.68 ± 0.65	**20.69 ± 1.44**	17.49 ± 0.79
RDFace cross-dataset transfer	Bal. Acc.	8.40 ± 1.56	**15.78 ± 0.89**	9.00 ± 0.53
RDFace unseen disease recognition	Bal. Acc.	5.23 ± 0.18	**5.89 ± 0.93**	4.96 ± 0.98

Stage 3 Phenotype Candidate Sensitivity

Table F.3 evaluates sensitivity to the number of phenotype candidates retained during Stage 3 patient-specific weak-label construction.

Across the representative metrics, the default intermediate candidate size ($K = 8$) consistently outperforms both smaller ($K = 4$) and larger ($K = 16$) candidate sets. This pattern is observed not only for disease recognition but also for image retrieval, atomic phenotype grounding, phenotype linear probing, patient-profile retrieval, and external transfer. A smaller candidate set may discard relevant phenotype evidence, whereas a larger set can introduce lower-confidence or less patient-specific disease-associated concepts into the weak target. The intermediate setting therefore appears

Table F.3: Sensitivity to the number of phenotype candidates retained in Stage 3. The default GestaltCLIP configuration uses $K = 8$.

Dataset / task	Metric	$K = 4$	$K = 8$ **(default)**	$K = 16$
GMDB Common disease recognition	Bal. Acc.	24.84 ± 0.42	$\mathbf{25.93 \pm 0.68}$	24.36 ± 0.42
GMDB Common image retrieval	mAP	6.51 ± 0.08	$\mathbf{6.77 \pm 0.17}$	6.51 ± 0.06
GMDB Common atomic HPO concept retrieval	mAP	12.28 ± 0.40	$\mathbf{13.07 \pm 0.16}$	12.49 ± 0.49
GMDB Common HPO concept linear probe	mAP	25.57 ± 0.05	$\mathbf{25.89 \pm 0.15}$	25.52 ± 0.02
GMDB Common concept set retrieval	mAP	4.87 ± 0.31	$\mathbf{6.05 \pm 0.41}$	4.79 ± 0.17
GMDB Rare disease recognition	Bal. Acc.	26.00 ± 0.83	$\mathbf{26.97 \pm 1.19}$	26.65 ± 2.02
GMDB Rare image retrieval	mAP	18.14 ± 0.55	$\mathbf{19.57 \pm 0.25}$	18.22 ± 0.36
GMDB Rare atomic HPO concept retrieval	mAP	10.52 ± 0.26	$\mathbf{11.91 \pm 0.16}$	10.45 ± 0.23
GMDB Rare HPO concept linear probe	mAP	20.51 ± 0.26	$\mathbf{21.04 \pm 0.47}$	20.44 ± 0.33
GMDB Rare concept set retrieval	mAP	19.46 ± 1.88	$\mathbf{20.69 \pm 1.44}$	18.11 ± 1.40
RDFace cross-dataset transfer	Bal. Acc.	13.78 ± 0.31	$\mathbf{15.78 \pm 0.89}$	13.78 ± 0.31
RDFace unseen disease recognition	Bal. Acc.	4.48 ± 0.52	$\mathbf{5.89 \pm 0.93}$	4.96 ± 0.55

to provide a favorable balance between phenotype coverage and patient-specificity during Stage 3 weak-label construction.

G Limitations and Broader Considerations

G.1 Limitations

GestaltCLIP is evaluated on rare disease facial datasets that remain limited in scale and demographic diversity. Some diseases contain only a small number of patients, and phenotype annotations vary in completeness across cases. Although the model improves transfer to unseen diseases, this setting still relies on the availability and quality of disease–HPO knowledge and may be less reliable for disorders with sparse or incomplete phenotype profiles. In addition, the current framework uses global facial representations and does not explicitly localize which facial regions support individual phenotype concepts. These factors may limit robustness under stronger domain shifts or in cases with subtle, age-dependent, or weakly visible facial features.

G.2 Future Work

Future work will address these limitations through larger and more diverse external evaluations, with particular attention to age, sex, and geographic subgroups. Phenotype supervision can also be strengthened by incorporating additional patient-level annotations and updated disease–HPO associations, while preserving the current weakly supervised setting. For unseen diseases with incomplete profiles, confidence-aware retrieval such as conformal prediction could be added to identify cases with insufficient phenotype support. Finally, the global representation can be extended with lightweight region-aware features to test whether localized facial evidence improves phenotype grounding without substantially changing the current training framework.

G.3 Broader Considerations

Facial images contain sensitive biometric and health-related information, and rare-disease applications require particular care because small patient populations can increase re-identification risk. The intended use of GestaltCLIP is research on representation learning and phenotype-aware retrieval, not autonomous clinical diagnosis. Dataset access, model development, and reporting should therefore follow appropriate consent, privacy, and data-governance procedures. Performance should also be examined across demographic and geographic subgroups before clinical use is considered. More broadly, phenotype-aware facial representations may support earlier screening or case retrieval, but their outputs should be interpreted together with clinical findings and genetic evidence rather than as standalone diagnostic conclusions.

H Reproducibility Check

This section summarizes the implementation and optimization settings used to reproduce the GestaltCLIP experiments. All stages use the same fixed 1,127-term facial HPO vocabulary, patient-grouped data splits, and seed-specific concept banks.

Experiments were implemented in Python 3.11 and PyTorch and were executed on four NVIDIA A100 GPUs. Mixed-precision training was enabled for the image-based stages using FP16 automatic mixed precision.

All experiments were repeated with random seeds 42, 43, and 44. The seeds controlled Python, NumPy, PyTorch, CUDA, data shuffling, and stochastic negative sampling. Reported aggregate results are the mean and sample standard deviation across the three seeds. Model selection and early stopping used only the predefined GMDB validation partitions. Neither the GMDB test partitions nor RDFace were used for representation training, checkpoint selection, threshold selection, or hyperparameter tuning.

The same OpenCLIP ViT-B/16 image backbone pretrained on LAION-2B (`laion2b_s34b_b88k`) was used throughout the image-based stages. Training images were processed using a 224×224 random resized crop with scale $(0.8, 1.0)$ and random horizontal flipping. Validation and test images were resized and center-cropped to 224×224. All images were normalized using the OpenCLIP mean and standard deviation. Optimization used AdamW, gradient clipping, and validation-based early stopping. The hyperparameters for each stage are summarized in the following tables.

Table H.1: Shared architecture and reproducibility settings for GestaltCLIP.

Parameter	**Value**
Primary random seeds	42, 43, and 44
Facial phenotype vocabulary	1,127 HPO concepts
Concept text encoder	SapBERT
Concept text encoder checkpoint	`cambridgeltl/SapBERT-from-PubMedBERT-fulltext`
Vision backbone	OpenCLIP ViT-B/16
Vision pretraining	LAION-2B `laion2b_s34b_b88k`
Image resolution	224×224
Shared embedding temperature	$\tau = 0.07$
Concept set hidden dimension	256
Concept set Transformer layers	2
Attention heads	4
Dropout	0.1
Concept set representation	Normalized mean + bounded residual
Maximum concept set residual ratio	0.10
Training precision	FP16 automatic mixed precision

Table H.2: Hyperparameters for Stage 1 real cases pretraining.

Parameter	**Value**
Training input	Patient-level clinical HPO concept sets
Image input	None
Trainable components	Bounded-residual PhenoEncoder and masked-HPO head
Optimizer	AdamW
Learning rate	3×10^{-4}
Weight decay	1×10^{-4}
Batch size	64
Maximum epochs	50
Validation fraction	0.10
Early-stopping patience	7 epochs
Minimum validation improvement	1×10^{-4}
Gradient-norm clipping	1.0
Masked-HPO fraction	0.30
Sampled unlabeled HPOs per case	64
Unlabeled-negative weight	0.05
Mean-preservation loss weight	1.0
Checkpoint criterion	Minimum validation loss

Table H.3: Hyperparameters for Stage 2 paired image–phenotype alignment.

Parameter	**Value**
Training input	Paired GMDB image–HPO cases
Training partitions	GMDB *common* and *rare* training sets
Optimizer	AdamW
Vision-backbone learning rate	1×10^{-6}
Projection-head learning rate	3×10^{-4}
Weight decay	0.01
Training batch size	32
Validation batch size	128
Maximum epochs	60
Early-stopping patience	8 epochs
Minimum validation improvement	1×10^{-4}
Gradient-norm clipping	1.0
Backbone adaptation	Last two visual Transformer blocks
Vision projection initialization	Identity-padded projection
Maximum projection residual ratio	0.15
Contrastive temperature	0.07
Multi-positive contrastive weight	1.0
PU atomic-concept loss weight	0.5
Image-to-set cosine loss weight	0.25
Projection-isometry weight	0.1
Sampled unlabeled HPOs per image	128
Unlabeled-negative weight	0.05
PU ranking weight	0.5
PU ranking margin	0.2

Table H.4: Hyperparameters for Stage 3 disease-mediated weak alignment.

Parameter	Value
Training input	Disease-linked GMDB images and paired replay cases
Training partitions	GMDB common and rare training sets
Optimizer	AdamW
Projection-head learning rate	5×10^{-5}
Weight decay	0.01
Batch size	16
Diseases per batch	8
Images per disease	2
Maximum epochs	40
Early-stopping patience	6 epochs
Minimum validation improvement	1×10^{-4}
Gradient-norm clipping	1.0
Vision-backbone adaptation	Frozen
Contrastive temperature	0.07
Pseudo-label candidate budget	Top-8 HPO concepts
Pseudo-label temperature	1.0
Visual evidence weight	1.0
Disease-prior weight	0.5
HPO-structure weight	0.5
Prior smoothing	1.0
Confidence weighting	Enabled
Minimum confidence weight	0.25
Maximum normalized entropy	1.0
Weak contrastive weight	1.0
Weak set-alignment weight	0.25
Overall weak-supervision weight	0.25
Paired Stage 2 replay weight	1.0
Atomic posterior weight	1.0
Atomic-objective warm-up	10 epochs
Weak-objective warm-up	10 epochs

Table H.5: Hyperparameters for downstream evaluation of frozen representations.

Parameter	Value
Representation	Frozen Stage 3 projected image embedding
Probe type	Linear softmax classifier
Optimizer	AdamW
Learning rate	0.01
Weight decay	0.01
Label smoothing	0.1
Training batch size	256
Embedding extraction batch size	128
Maximum probe epochs	100
Early-stopping patience	10 epochs
Minimum validation improvement	1×10^{-4}
Diagnostic retrieval cutoffs	$k \in \{1, 5, 10\}$
Model-selection data	Matching GMDB validation partition only
External evaluation	RDFace, evaluation only
Aggregate reporting	Mean $\pm$ sample SD over three seeds