RDFace: A Benchmark Dataset for Rare Disease Facial Image Analysis under Extreme Data Scarcity and Phenotype-Aware Synthetic Generation

Supplementary Material

Appendix Contents

A. Dataset documentation

A.1. Disease list and metadata

We provide a complete list of the 103 rare disease classes included in the RDFace dataset. As summarized in Table S1, each entry includes the disease name, abbreviation (Abbr) (used for labeling), associated gene (if available), clinical subcategory based on Orphanet (if available), Orphanet code, and the number of real facial images (# Img) curated for that class.

Table S1. Metadata of rare disease classes in the RDFace dataset.[1]

Disease Name	Abbr	Gene	Subcategory	Orphanet Code	# Img
Aarskog-Scott syndrome	AAR	FGD1	Delayed puberty	915	5
Aicardi-Goutieres	AIC	TREX1	?	51	5
Triple A syndrome	ALL	AAAS	multisystem disease	869	5
Allan-herndon-dudley Syndrome	ALLA	SLC16A2	Ataxia	59	5
Alport syndrome	ALP	?	Abnormal retinal morphology	63	4
Alpha-thalassemia	ALPH	?	Cholestasis	846	5
Alpha-mannosidase deficiency	ALPM	MAN2B1	lysosomal storage disease	61	5
Alstrom	ALS	ALSM1	multisystemic disorder	64	5
Angelman Syndrome	ANG	UBE3A	Intellectual disability	72	5
X-linked cleft palate and ankyloglossia	ANK	TBX22	developmental defect during embryogenesis syndrome	324601	2
Apert syndrome	APE	FGFR2	Cardiomyopathy	87	5
AR Polycystic Kidney Disease	ARP	PKHD1	hepatorenal fibrocystic syndrome	731	1
Arterial Tortuosity	ART	?	connective tissue disorder	3342	5
Ataxia Telangiectasia	ATA	ATM	combined dystonia	100	2
Atypical Rett Syndrome	ATYP	MECP2, GABBR2, STXBP1, CDKL5	Calcium nephrolithiasis	3095	5
Auriculocondylar Syndrome	AURI	EDN1, PLCB4, GNAI3	Abnormal soft palate morphology to abnormality of the uvulva	137888	5
Bainbridge-ropers Syndrome	BAI	ASXL3	Severe postnatal growth retardation	352577	5
Bardet-Biedl	BAR	BBS2	ciliopathy with multisystem involvement	110	5
Barber-say Syndrome	BARB	TWIST2	Hyperextensible skin	1231	5
Barth syndrome	BART	?	Cardiomyopathy	111	5
Beckwith-wiedemann Syndrome	BEC	?	Multiple renal cysts	116	5
Bohring-opitz Syndrome	BOH	ASXL1	Muscular hypotonia	97297	5
Boycott-Beaulieu-Innes	BOY	THOC6	syndromic intellectual disability disorder	363444	5
Congenital adrenal hyperplasia	CAH	?	Congenital adrenal hyperplasia	418	5
Canavan Disease	CAN	ASPA	Abnormality of serum amino acid level	141	5
COFS syndrome	CER	ERCC6	diseases of DNA repair	1466	5
Chudley-McCullough	CHU	GPSM2	syndromic deafness	314597	4
Cleidocranial Dysplasia	CLE	CBFA1	developmental abnormality of bone	1452	5
Clouston Syndrome	CLOU	GJB6	Ectodermal dysplasia	189	4
CODAS	COD	LONP1	multiple congenital anomalies syndrome	1458	5
Coffin-lowry Syndrome	COF	RPS6KA3	Spasticity	192	5
Combined pituitary hormone deficiencies, genetic forms	COM	PROP1	Congenital hypopituitarism	95494	4
SLC39A8-CDG	COND	SLC39A8	?	468699	5

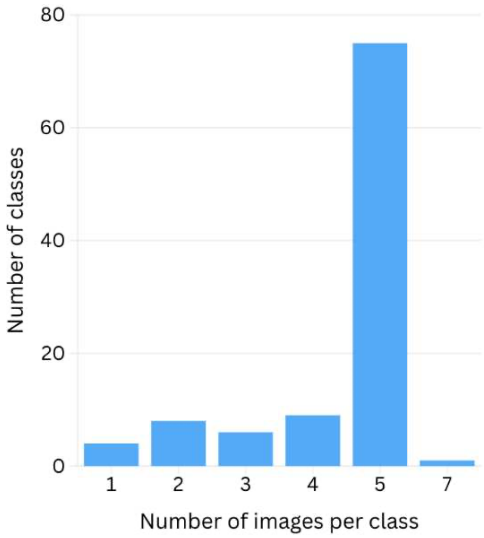
Disease Name	Abbr	Gene	Subcategory	Orphanet Code	# Img
Cranioectodermal dysplasia	CRA	DPH1	developmental disorder	1515	5
Crisponi Syndrome	CRIS	CLCF1, CRLF1	Malignant hyperthermia	1545	5
3C syndrome	CSY	CCDC22, WASHC5	?	7	5
Cushing disease	CUS	USP8, CDH23	Abnormal bleeding	96253	4
Cystic Fibrosis	CYS	CFTR	?	586	5
Diamond-blackfan Anemia	DIA	RPS19	Colon cancer	124	5
Donnai-barrow Syndrome	DON	LRP2	Partial agenesis of the corpus callosum	2143	5
DOORS syndrome	DOO	TBC1D24	Hypothyroidism	79500	5
Dopa-Responsive Dystonia	DOP	TH	group of diseases	255	3
Dysosteosclerosis	DYS	?	primary bone dysplasia disease	1782	2
Early infantile epileptic encephalopathy	EPIE	?	epileptic encephalopathy	1934	5
Fanconi Anemia	FANC	FANCC	DNA repair disorder	84	5
Geroderma Osteodysplastica	GEO	GORAB	?	2078	5
Glutaryl-CoA dehydrogenase deficiency	GLU	GCDH	neurometabolic disorder	25	5
HHH (hyperornithinemia-hyperammonemia-homocitrullinuria)	HHH	SLC25A15	disorder of urea cycle metabolism	415	1
Hypohidrotic Ectodermal Dysplasia	HYPE	EDA1	disorder of ectoderm development	238468	1
Hypophosphatasia	HYPO	ALPL	metabolic disorder	436	5
Severe combined immunodeficiency	IMMU	DCLRE1C	primary immunodeficiency	183660	5
Joubert Syndrome	JOUA	TMEM237	?	475	5
Juvenile Amyotrophic Lateral Sclerosis	JUV	ALS2, HNRNPA2B1, HNRNPA1	Motor neuron atrophy	300605	5
Kawasaki Disease	KAW	?	Arrhythmia	2331	5
Infantile Krabbe Disease	KRA	GALC, PSAP	Generalized myoclonic seizure	206436	5
Laron Syndrome	LAR	GHR	Hypoglycemia	633	3
Leigh	LEI	NDUFV1	progressive neurological disease	506	5
Leprechaunism	LEP	INSR	Hypoglycemia	508	7
Loeys-Dietz	LOE	TGFB2	connective tissue disorder	60030	5
Malignant hyperthermia of anesthesia	MAL	RYR1	pharmacogenetic disorder of skeletal muscle	423	3
Maple Syrup Urine Disease	MAP	BCKDHB	disorder of branched-chain amino acid metabolism	511	5
Marden-walker Syndrome	MAR	PIEZO2	Muscular dystrophy	2461	5
Marinesco-sjögren Syndrome	MARI	INPP5K, SIL1	Muscular dystrophy	559	5
Oculotrichoanal syndrome	MBO	FREM1	multiple congenital anomalies	2717	5
Mucopolysaccharidosis type 4	MOR	?	lysosomal storage disease	582	5
Mowat-wilson Syndrome	MOW	ZEB2	Abdominal distention	2152	5
Mucopolysaccharidosis type 7	MPS	GUSB	lysosomal storage disease	584	5
Multiple Sulfatase Deficiency	MUL	SUMF1	Progressive neurologic deterioration	585	5
Ochoa Syndrome	OCH	HPSE2, LRIG2	Renal insufficiency	2704	5
Oculocutaneous Albinism	OCU	TYR	disorders of melanin biosynthesis	55	5
Odonto-onycho-dermal dysplasia	ODO	WNT10A	ectodermal dysplasia	2721	5
Osteogenesis imperfecta	OST	SEPINF1	group of diseases	666	5
Parietal Foramina	PAR	?	?	60015	3

Disease Name	Abbr	Gene	Subcategory	Orphanet Code	# Img
Glycogen storage disease due to acid maltase deficiency, late-onset	POM	GAA	Glycogen storage disease	420429	5
Pontocerebellar Hypoplasia	PON	TOE1	?	98523	5
Primary Hyperoxaluria	PRI	GRHPR	disorder of glyoxylate metabolism	416	2
Microcephaly-lymphedema-chorioretinopathy syndrome	PRIM	?	?	2526	5
Propionic Acidemia	PRO	PCCB	organic aciduria	35	5
PRUNE1-related neurological syndrome	PRU	PRUNE1	?	544469	2
Pten Hamartoma Tumor Syndrome	PTEN	?	Endometrial carcinoma	306498	5
Pyruvate Carboxylase Deficiency	PYR	PC	neurometabolic disorder	3008	2
Renpenning	REN	PQBP1	intellectual disability syndrome	3242	4
Restrictive Dermopathy	RES	ZMPSTE24	congenital genodermatosis	1662	5
Rhizomelic Chondrodysplasia Punctata	RHO	PEX7	group of diseases	177	5
Roberts	ROB	ESCO2	?	3103	5
Spinal arteriovenous metameric syndrome	SAM	GSC	?	53721	1
Sandhoff Disease	SAND	?	Progressive psychomotor deterioration	796	5
Sialidosis type 2	SIA	NEU1	lysosomal storage disease	87876	5
Sickle Cell Anemia	SIC	HBB	Chronic hemolytic anemia	232	3
Proximal Spinal Muscular Atrophy	SPI	SMN1	neuromuscular disorder	70	4
Spondylodysplastic Ehlers-danlos Syndrome	SPO	?	Platyspondyly	536471	3
Congenital sucrase-isomaltase deficiency	SUC	SI	carbohydrate intolerance disorder	35122	2
Isolated sulfite oxidase deficiency	SUL	SO	?	99731	5
Temple syndrome	TEMP	?	Maturity-onset diabetes of the young	254516	5
Turner Syndrome	TUR	?	Biliary cirrhosis	881	4
Tyrosinemia Type 1	TYR	FAH	inborn error of tyrosine catabolism	882	5
Usher Syndrome type 1	USHB	MYO7A	?	231169	4
CACH syndrome	VAN	EIF2B5	?	135	5
Hyperostosis corticalis generalisata	VANB	?	craniotubular hyperostosis	3416	2
Walker Warburg	WAL	POMT1	congenital muscular dystrophy	899	5
Warsaw Breakage Syndrome	WAR	DDX11	psychomotor retardation	280558	5
Wolf-hirschhorn Syndrome	WOL	LETM1, NSD2	Rib segmentation abnormalities	280	5
Zellweger	ZEL	?	peroxisome biogenesis disorder	912	5

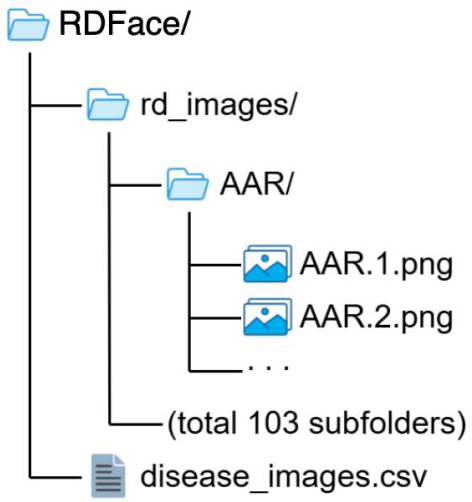
[1] ? indicates that the information is not available.

A.2. Dataset distribution and organization

Aside from the metadata table, we also provide two figures to illustrate the distribution and structure of the RDFace dataset. Figure S1a shows the number of images available for each disease class, highlighting the inherent imbalance in data availability across different rare diseases. Figure S1b depicts the organization of the dataset. These visualizations provide a clearer understanding of the dataset's composition and the challenges posed by data scarcity in rare disease research.



(a) Distribution of images across disease classes.



(b) Structure of the RDFace dataset.

Figure S1. **Overview of the RDFace dataset.**

B. Few-shot learning

Few-shot learning (FSL) addresses the problem of classifying instances when only a few labeled examples are available for each class. This setting is common in domains like rare disease diagnosis, where collecting large-scale labeled datasets is impractical due to the low prevalence and data privacy constraints.

In the standard n-way k-shot FSL setting, each learning episode involves n distinct classes, with only k labeled examples (support set) per class. The goal is to classify unlabeled examples (query set) drawn from the same n classes, using the limited support data for reference.

Formally, each episode consists of:

- A support set $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^{n \cdot k}$, containing k samples per class.
- A query set $\mathcal{Q} = \{(x_j, y_j)\}_{j=1}^{q}$, containing unseen samples from the same n classes.

We adopted Prototypical Networks to solve the few-shot classification task on RDFace. This approach learns an embedding function $f_\theta : \mathcal{X} \to \mathbb{R}^d$ that maps input images into a feature space. For each class c_i, a prototype (centroid) is computed as the mean embedding of its support samples:

$$\mu_i = \frac{1}{k} \sum_{x \in \mathcal{S}_i} f_\theta(x) \tag{1}$$

For each query image $x^{(q)}$, the model computes the distance between its embedding and each prototype μ_i, typically using the squared Euclidean distance:

$$d(x^{(q)}, \mu_i) = \left\| f_\theta(x^{(q)}) - \mu_i \right\|_2^2 \tag{2}$$

The query image is then assigned to the class with the closest prototype, and a cross-entropy loss is computed over the predictions for optimization.

B.1. Prototypical networks algorithm

We provide the algorithm used for episodic training on RDFace as below:

Algorithm 1 Prototypical Networks Episodic Training on RDFace

Require: Feature extractor f_θ, training classes $\mathcal{C}_{\text{train}}$, number of ways n, support size $k = 1$

Sample n classes $\{c_1, \ldots, c_n\} \sim \mathcal{C}_{\text{train}}$
for each class c_i **do**
    Sample support set $\mathcal{S}_i = \{x_i^{(s)}\}$, query set $\mathcal{Q}_i = \{x_i^{(q)}\}$
    Compute prototype: $\mu_i \leftarrow f_\theta(x_i^{(s)})$
end for
for each query $x_j^{(q)} \in \bigcup_i \mathcal{Q}_i$ **do**
    Compute distances: $d_{ij} \leftarrow \|f_\theta(x_j^{(q)}) - \mu_i\|_2$
    Predict label: $\hat{y}_j \leftarrow \arg\min_i d_{ij}$
end for
Compute loss: $\mathcal{L}_{\text{episode}} \leftarrow \text{CrossEntropy}(\hat{y}, y)$
Update f_θ via backpropagation

C. Synthetic image samples

Figure S2 presents the top 100 synthetic facial images generated by FastGAN, ranked by their similarity to real disease faces using a landmark-based cosine metric. These samples highlight the model's ability to capture coarse facial structure across diverse rare disease classes, although occasional artifacts and inconsistencies remain visible. In contrast, Figure S3 provides a more targeted comparison using DreamBooth of 50 classes choosen to display. For each disease class, we show the most and least phenotype-consistent synthetic images compared to real images based on 5×5 facial landmark cosine similarity. This comparison illustrates both the strength and the limitations of current generative models. While top-ranked DreamBooth samples often replicate key craniofacial traits, the bottom-ranked samples reveal potential risks of phenotype distortion or mode collapse. These visualizations support the use of generative models for phenotype augmentation, while also emphasizing the need for careful validation when applying synthetic images in clinical or diagnostic settings.

Figure S2. **Representative FastGAN images ranked most similar to diseases.**

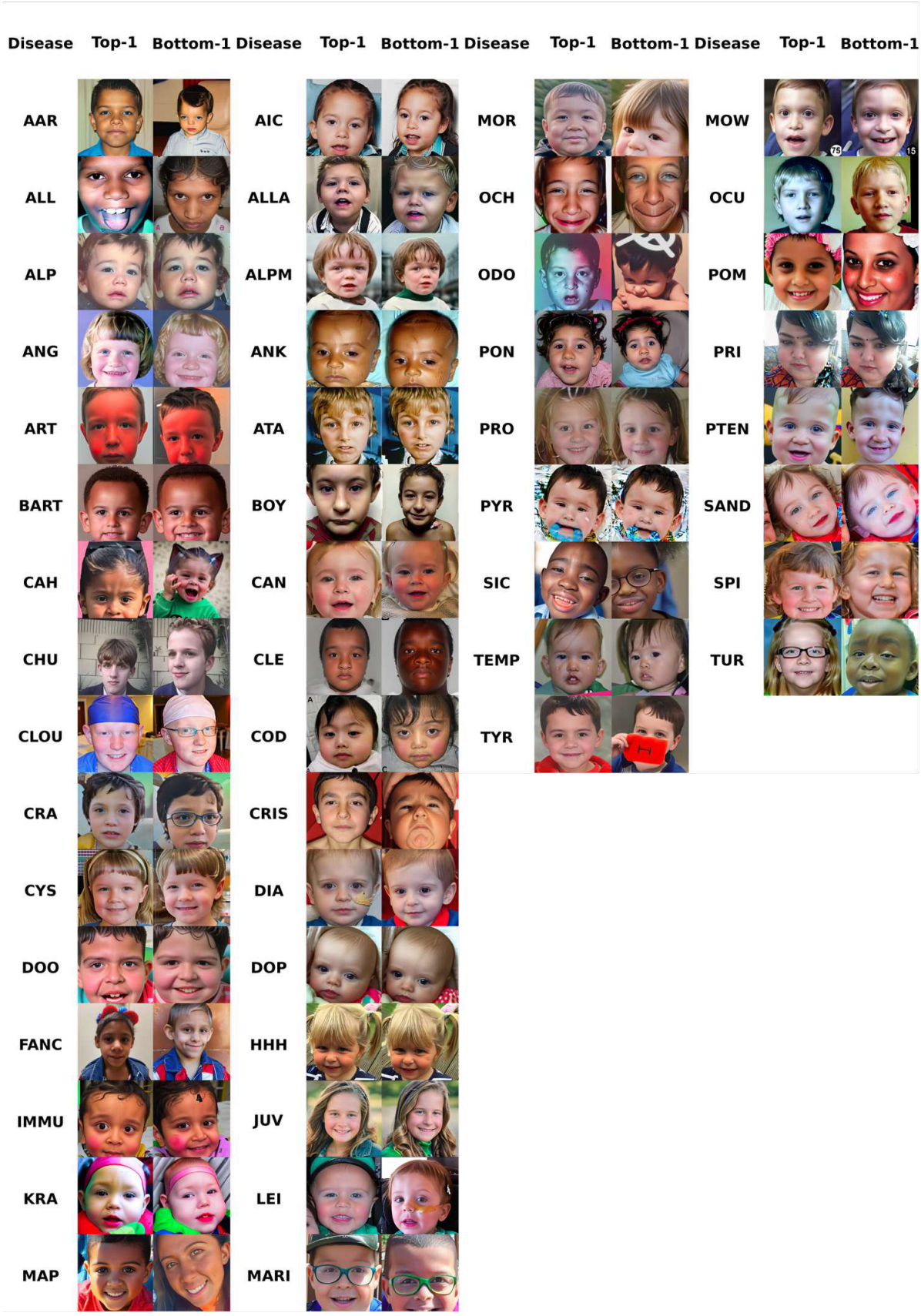


Figure S3. **Top-1 and Bottom-1 synthetic images based on similarity to real images.** Top-1 and Bottom-1 denote the most and least similar samples respectively.

D. Expert and automated evaluation of synthetic data

D.1. Landmark-based similarity analysis

We conducted our evaluation across all 103 rare disease classes in RDFace. For each class, DreamBooth-generated synthetic images were used, provided they passed quality checks based on facial structure and alignment. Images failing these checks, such as those with low RetinaFace detection confidence or invalid landmark configurations, were excluded at the image level, not the class level. All similarity analyses were performed using normalized 5×5 landmark distance matrices computed from RetinaFace keypoints.

D.1.1. Heatmap of landmark-based cosine similarities

Figure S4 presents a cosine similarity heatmap comparing the average DreamBooth-generated landmark structure for each class against all real disease class prototypes. The heatmap reveals a strong diagonal trend, reflecting high intra-class similarity, with several classes showing close alignment between synthetic and real facial geometry. However, for certain classes—particularly those with very limited or noisy training data—the diagonal intensity diminishes, indicating lower phenotype fidelity. Off-diagonal activity highlights cross-class resemblance, often reflecting overlapping craniofacial features among syndromes or mode collapse in synthesis.

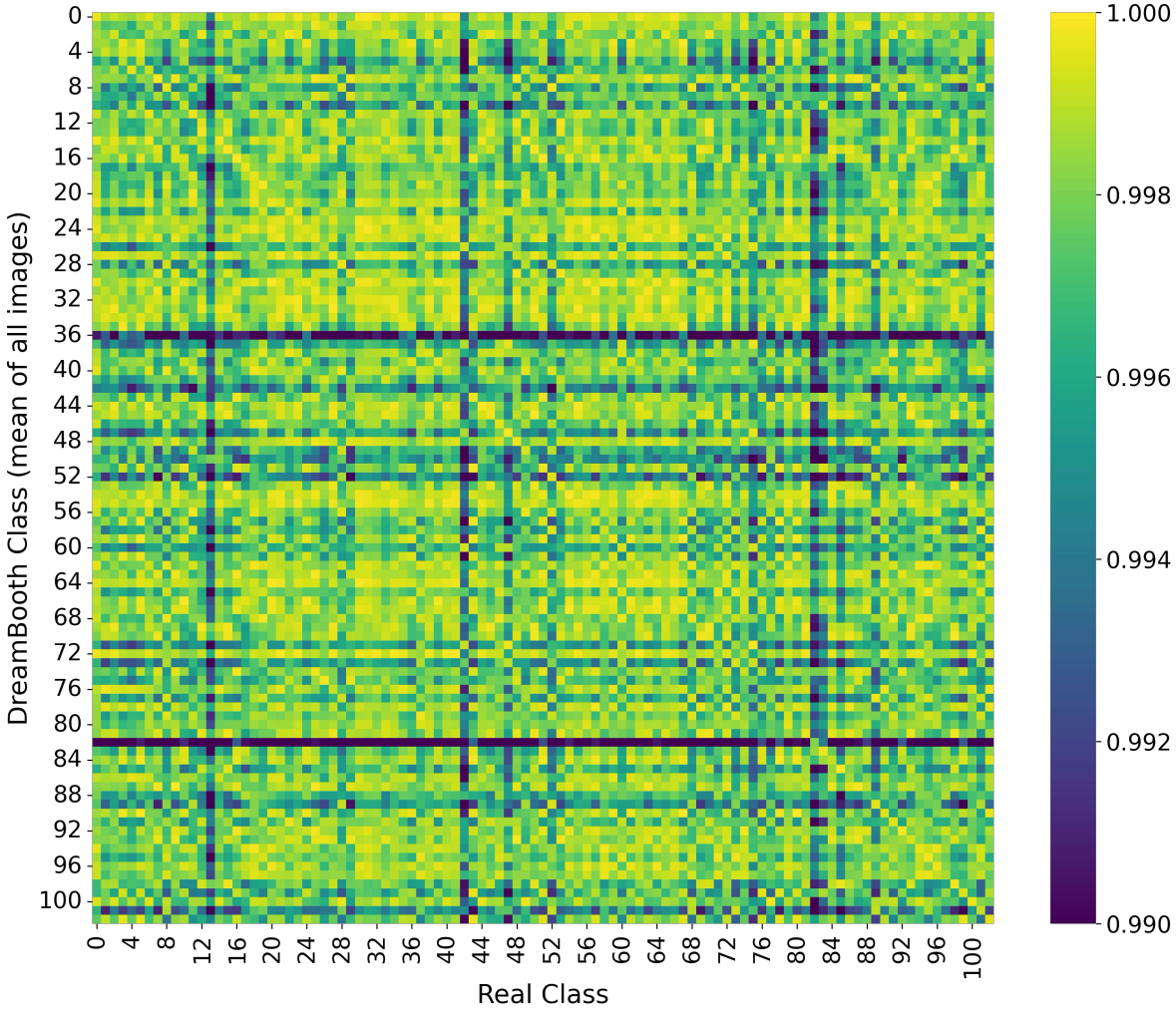


Figure S4. **Cosine similarity heatmap between DreamBooth-generated and real disease prototypes.** The heatmap compares the average landmark distance matrices of DreamBooth-generated images (rows) to those of real disease class prototypes (columns). Brighter values indicate greater similarity.

D.1.2. Ranking consistency of DreamBooth-generated Images

To quantify how consistently DreamBooth preserves class-specific features, we computed the rank position of the ground-truth class for each synthetic image among all real class prototypes based on cosine similarity. Figure S5 summarizes the average rank per class.

While many classes exhibit ranks close to 1, several fall well below the mean, suggesting inconsistency in capturing distinctive morphology. These cases often correspond to underrepresented or visually subtle conditions in the training set. The overall mean rank across all classes is 19.74, serving as a practical reference for DreamBooth's phenotype alignment under ultra-low-shot conditions.

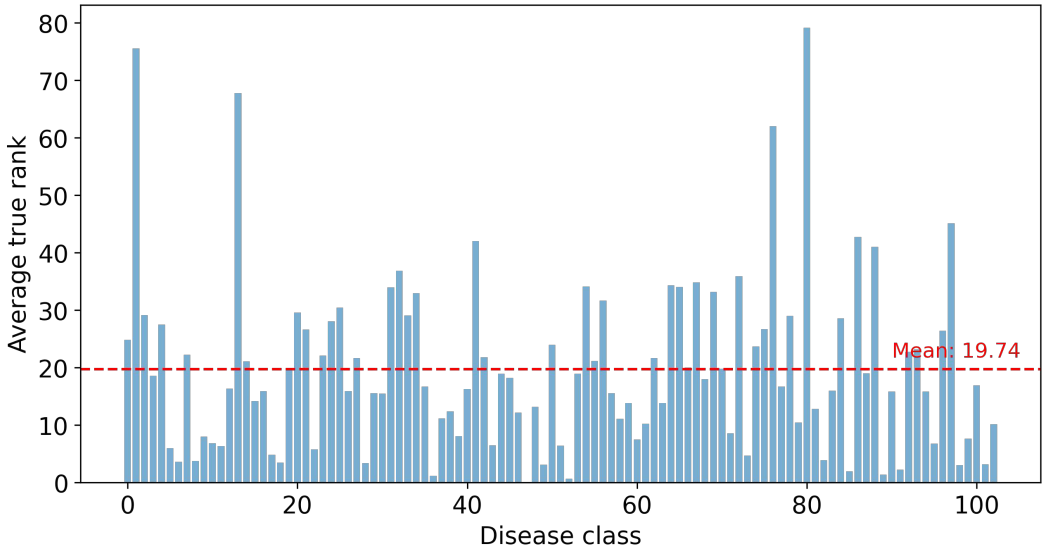


Figure S5. **Average rank of the true disease class for each DreamBooth-generated image.** Each rank is calculated using cosine similarity to the corresponding real class prototype. The red dashed line indicates the overall mean rank.

D.2. Expert review

In order to evaluate the clinical plausibility of synthetic image-label pairs, we conducted an expert review of 50 DreamBooth and 50 FastGAN samples. Each image was independently assessed by two medical doctor (MD) students (MD1 and MD2) and categorized as *Plausible*, *Implausible*, or *Uncertain*. The results of this expert evaluation are summarized in Table S2.

Table S2. Expert evaluation of the plausibility of synthetic image-label pairs generated by DreamBooth and FastGAN with two MDs.

Plausibility Rating	**DreamBooth**		**FastGAN**	
	MD1	MD2	MD1	MD2
Plausible	38	31	14	5
Implausible	2	3	15	30
Uncertain	10	16	21	15

To assess the consistency between the two raters, inter-rater reliability was computed for each model, and the pairwise confusion matrices are presented in Table S3. Each cell represents the number of images assigned to a given label combination by the two MD reviewers (*DreamBooth / FastGAN*). A strong concentration of counts along the diagonal indicates higher agreement on image plausibility.

Quantitatively, the inter-rater agreement results (Table S4) demonstrate that DreamBooth achieved a substantially higher degree of consensus between raters ($\kappa = 0.654$), corresponding to "substantial agreement" on the Landis–Koch scale, whereas FastGAN showed only minimal agreement ($\kappa = 0.069$). This difference highlights that DreamBooth-generated

Table S3. Expert validation confusion matrices for DreamBooth (DB) and FastGAN (FG). (DB / FG)

Label	Plausible	Implausible	Uncertain	Total
Plausible	31 / 2	0 / 0	0 / 3	31 / 5
Implausible	0 / 7	2 / 11	1 / 12	3 / 30
Uncertain	7 / 5	0 / 4	9 / 6	16 / 15
Total	38 / 14	2 / 15	10 / 21	50 / 50

samples were generally perceived as more clinically plausible and consistent across evaluators, while FastGAN outputs exhibited greater variability and uncertainty.

Table S4. Inter-rater agreement metrics for DreamBooth (DB) and FastGAN (FG).

Method	Observed Agreement (%)	Cohen's κ	SE	95% CI
DB	84.0 (42/50)	0.654	0.106	[0.446, 0.862]
FG	38.0 (19/50)	0.069	0.091	[–0.110, 0.248]

Overall, the expert review confirms that DreamBooth produces synthetic facial images that retain phenotypic plausibility and diagnostic relevance more consistently than FastGAN, aligning with the quantitative similarity metrics and qualitative visual inspection presented in the main manuscript.

D.3. Observations and implications

Our evaluation highlights the complementary strengths of automated and expert-based assessments for characterizing the quality of synthetic facial images. Landmark-based similarity analysis reveals that many disease classes exhibit distinct and consistent craniofacial geometry, validating the use of facial landmarks as a phenotypic signature. The observed variability in average rank and cosine similarity across classes reflects sensitivity to both morphological subtlety and the quality of training exemplars. These findings indicate that normalized landmark distance metrics offer an interpretable, spatially grounded method for quantifying phenotype preservation and filtering synthetic samples prior to downstream tasks.

However, structural alignment alone does not guarantee clinical plausibility. To address this, we conducted an expert review to evaluate whether synthetic image-label pairs appeared medically credible. Results show that DreamBooth-generated images were more frequently judged as plausible by at least one medical doctor, whereas FastGAN samples showed higher rates of uncertainty and implausibility. These findings underscore that perceptual realism, which is often emphasized in generative model benchmarks, does not always correlate with diagnostic fidelity. Expert review provides a critical semantic layer that captures domain-specific context often missed by automated metrics.

Together, these results suggest that effective synthetic data curation requires a multi-faceted evaluation strategy that integrates spatial similarity, visual quality, and human judgment. Such approaches are especially important in rare disease settings, where subtle phenotypic cues and label noise can dramatically impact model performance. In future applications, combining interpretable landmark filtering with lightweight expert-in-the-loop review may provide a scalable path for enhancing dataset quality and clinical utility.

E. Tradeoff between disease-specific structure and visual realism

To assess the quality and disease-relevance of generated synthetic images, we evaluate two image-level realism metrics: RetinaFace detection confidence and LPIPS perceptual similarity. These metrics are computed across Top-n subsets (ranging from 1000 to 6000 images), where the images are ranked by their landmark-based cosine similarity to real disease class prototypes.

DreamBooth Figure S6 shows the trend of RetinaFace and LPIPS scores across Top-n DreamBooth subsets. As n increases, RetinaFace confidence scores gradually decline, suggesting reduced alignment and detectability in lower-ranked images. Meanwhile, LPIPS scores increase, indicating a decrease in perceptual similarity to real images. These opposing trends reflect a trade-off between structural fidelity (as captured by landmarks) and low-level texture realism. The landmark-based ranking effectively promotes DreamBooth samples with coherent facial geometry and higher visual plausibility.

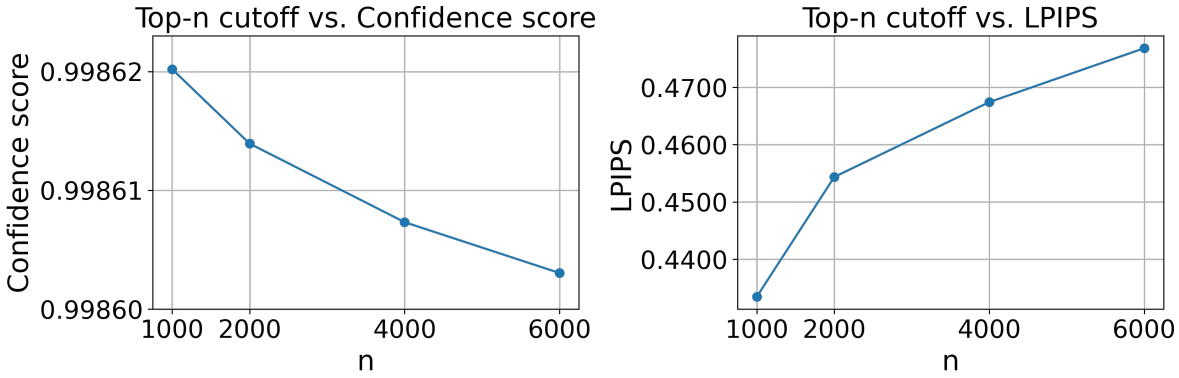


Figure S6. **DreamBooth – Correlation Between Top-n Ranking and Visual Realism.** RetinaFace detection confidence (left) and LPIPS similarity (right) across Top-n ranked DreamBooth images.

FastGAN FastGAN samples exhibit a broadly similar trend to DreamBooth in terms of ranking-based visual quality (see Figure S7). As the Top-n threshold increases, RetinaFace confidence scores slightly decline and LPIPS scores gradually increase, indicating reduced structural detectability and perceptual similarity at lower-ranked positions. However, some local fluctuations are observed—particularly around the Top-2000 cutoff—where both metrics deviate slightly from the overall trajectory. These irregularities suggest that landmark-based ranking is still somewhat effective for FastGAN, but may be less stable than for DreamBooth due to the lack of class-specific conditioning.

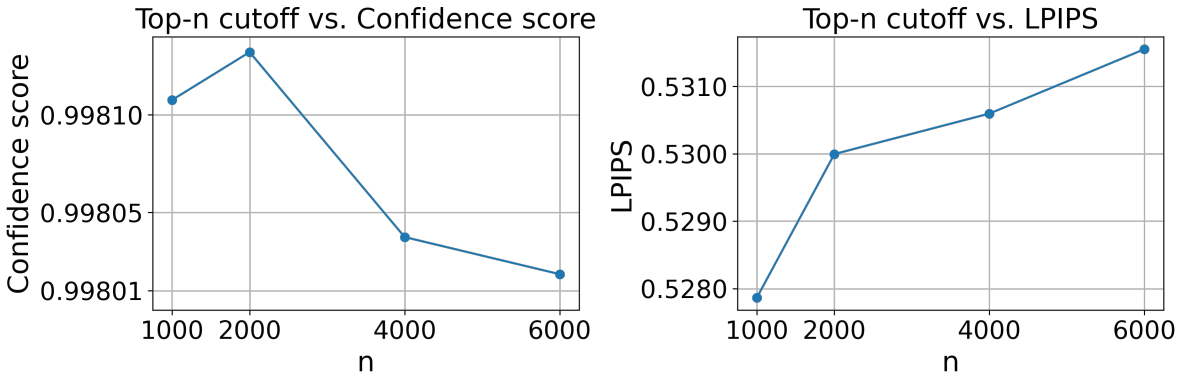


Figure S7. **FastGAN – Correlation Between Top-n Ranking and Visual Realism.** RetinaFace detection confidence (left) and LPIPS similarity (right) across Top-n ranked FastGAN images.

F. Synthetic data-involved downstream tasks results

F.1. Standard supervised classification and synthetic scaling effect

Table S5 and Table S6 below report Top-k classification accuracies across various backbones. Each row represents classification performance (Top-k accuracy) under a different training configuration. "Real only" refers to models trained exclusively on real RDFace data. "Top-n" rows correspond to training sets augmented with the Top-n synthetic images selected based on landmark similarity to real samples. The synthetic images are chosen to best align with phenotype-specific facial structure.

Table S5. Top-k accuracies (%) across backbones and synthetic cutoffs of DreamBooth samples.

ACC (%)	Top-n	ResNet	DenseNet	FaceNet	VGG	Swin-T	CLIP
Top-1	Real only	6.90 (1.45)	**15.93 (2.34)**	9.91 (1.81)	11.68 (1.58)	14.34 (2.61)	3.1 (1.48)
	Top-1000	12.21 (1.70)	**17.52 (2.29)**	15.04 (2.87)	16.64 (4.07)	16.81 (1.40)	9.03 (2.29)
	Top-2000	12.57 (2.61)	**20.35 (1.40)**	14.51 (1.34)	14.69 (1.61)	18.76 (2.20)	15.22 (4.35)
	Top-4000	13.27 (2.26)	**19.65 (0.74)**	16.46 (1.01)	17.35 (4.18)	18.76 (2.29)	16.81 (1.25)
	Top-6000	13.63 (1.61)	**21.06 (2.53)**	16.28 (2.91)	16.64 (2.61)	18.94 (2.70)	15.75 (2.58)
Top-5	Real only	18.58 (3.00)	**33.63 (3.70)**	24.60 (5.43)	29.91 (2.68)	26.19 (2.68)	12.74 (2.84)
	Top-1000	27.26 (3.39)	**35.58 (3.39)**	32.04 (1.92)	29.91 (1.45)	33.81 (2.29)	17.52 (4.26)
	Top-2000	30.09 (1.98)	**37.35 (2.37)**	33.63 (2.65)	33.10 (3.99)	36.81 (3.40)	25.13 (2.70)
	Top-4000	30.09 (4.29)	**40.35 (4.08)**	33.98 (2.97)	34.16 (2.84)	35.40 (4.15)	28.14 (0.97)
	Top-6000	33.45 (4.70)	**39.65 (2.29)**	32.74 (1.77)	33.81 (1.58)	36.46 (3.78)	29.38 (2.45)
Top-10	Real only	28.50 (3.67)	**43.01 (2.63)**	34.87 (4.75)	38.41 (1.34)	35.93 (3.17)	19.12 (5.18)
	Top-1000	38.41 (3.94)	**47.61 (3.30)**	44.25 (2.94)	39.82 (1.77)	46.19 (3.03)	26.02 (2.55)
	Top-2000	41.59 (5.16)	**47.26 (2.13)**	43.89 (2.55)	43.01 (1.34)	46.02 (1.08)	33.98 (2.25)
	Top-4000	40.53 (3.45)	**51.33 (3.00)**	43.72 (2.70)	44.25 (3.00)	46.19 (1.58)	34.87 (1.61)
	Top-6000	42.65 (2.68)	**49.20 (1.48)**	43.89 (2.47)	46.73 (1.15)	48.67 (2.26)	37.70 (2.03)
Top-30	Real only	54.34 (2.39)	**64.42 (1.92)**	58.23 (5.06)	60.88 (2.02)	58.41 (3.81)	42.30 (4.40)
	Top-1000	62.30 (2.22)	**70.62 (3.03)**	64.07 (3.99)	63.01 (2.83)	**70.62 (2.20)**	49.56 (2.65)
	Top-2000	62.12 (3.15)	**70.62 (3.39)**	64.60 (3.59)	66.73 (3.99)	69.56 (2.70)	57.17 (4.18)
	Top-4000	63.36 (2.91)	70.09 (1.48)	67.96 (1.58)	64.42 (2.37)	**70.80 (1.08)**	57.35 (2.02)
	Top-6000	63.19 (3.94)	**68.67 (3.23)**	63.01 (2.89)	66.37 (2.58)	68.50 (2.39)	61.24 (3.05)

Table S6. Top-k accuracies (%) across backbones and synthetic cutoffs of FastGAN samples.

ACC (%)	Top-n	ResNet	DenseNet	FaceNet	VGG	Swin-T	CLIP
Top-1	Real only	6.90 (1.45)	**15.93 (2.34)**	9.91 (1.81)	11.68 (1.58)	14.34 (2.61)	3.1 (1.48)
	Top-1000	8.50 (1.48)	**13.27 (3.13)**	6.55 (2.84)	7.26 (2.37)	10.44 (1.70)	1.42 (1.34)
	Top-2000	6.55 (2.13)	**10.44 (0.40)**	5.13 (0.74)	6.37 (3.67)	8.50 (2.63)	1.06 (1.15)
	Top-4000	4.07 (1.48)	8.14 (1.45)	4.96 (1.19)	7.96 (2.26)	**9.73 (1.40)**	0.71 (0.74)
	Top-6000	4.78 (0.79)	**9.73 (2.08)**	4.78 (1.34)	5.49 (2.61)	9.20 (1.84)	1.42 (1.01)
Top-5	Real only	18.58 (3.00)	**33.63 (3.70)**	24.60 (5.43)	29.91 (2.68)	26.19 (2.68)	12.74 (2.84)
	Top-1000	20.71 (1.34)	**27.79 (1.34)**	18.05 (2.97)	18.94 (3.40)	25.66 (1.88)	5.49 (2.37)
	Top-2000	19.12 (2.31)	23.54 (2.22)	18.05 (1.73)	18.76 (2.89)	**25.49 (4.03)**	6.90 (2.76)
	Top-4000	17.17 (2.31)	25.49 (2.89)	16.81 (1.25)	20.71 (2.04)	**26.19 (3.57)**	5.49 (1.70)
	Top-6000	18.58 (1.98)	**25.49 (2.02)**	15.58 (1.94)	21.95 (0.97)	23.89 (4.42)	6.19 (1.08)
Top-10	Real only	28.50 (3.67)	**43.01 (2.63)**	34.87 (4.75)	38.41 (1.34)	35.93 (3.17)	19.12 (5.18)
	Top-1000	29.20 (3.59)	**37.70 (3.29)**	30.27 (3.51)	29.03 (5.92)	36.64 (2.22)	11.15 (1.48)
	Top-2000	28.67 (4.18)	34.34 (0.74)	28.85 (3.79)	29.38 (2.20)	**35.58 (3.51)**	11.68 (2.20)
	Top-4000	29.03 (2.68)	35.93 (3.04)	28.32 (2.50)	31.33 (4.23)	**36.64 (4.58)**	11.15 (2.77)
	Top-6000	31.15 (3.15)	**35.93 (3.10)**	24.07 (2.29)	32.39 (2.97)	33.45 (3.62)	13.63 (0.79)
Top-30	Real only	54.34 (2.39)	**64.42 (1.92)**	58.23 (5.06)	60.88 (2.02)	58.41 (3.81)	42.30 (4.40)
	Top-1000	51.68 (2.04)	**61.59 (5.82)**	57.52 (0.63)	55.22 (5.07)	60.35 (1.92)	34.51 (2.73)
	Top-2000	53.10 (3.54)	58.41 (2.08)	55.04 (4.44)	53.27 (3.33)	**61.77 (4.90)**	32.39 (2.13)
	Top-4000	56.46 (3.51)	**60.00 (3.15)**	55.58 (1.70)	59.12 (4.12)	57.88 (3.40)	32.57 (2.46)
	Top-6000	55.75 (3.19)	54.34 (3.68)	55.40 (2.31)	**60.71 (1.94)**	56.46 (1.15)	38.23 (5.78)

Synthetic scaling effect The relationship between real-only and synthetic-augmented performance across Top-k settings and Top-n synthetic cutoffs is shown in Figure S8 and Figure S9. These plots visualize the downstream classification accuracies using DreamBooth and FastGAN generated samples, respectively, across six backbone models.

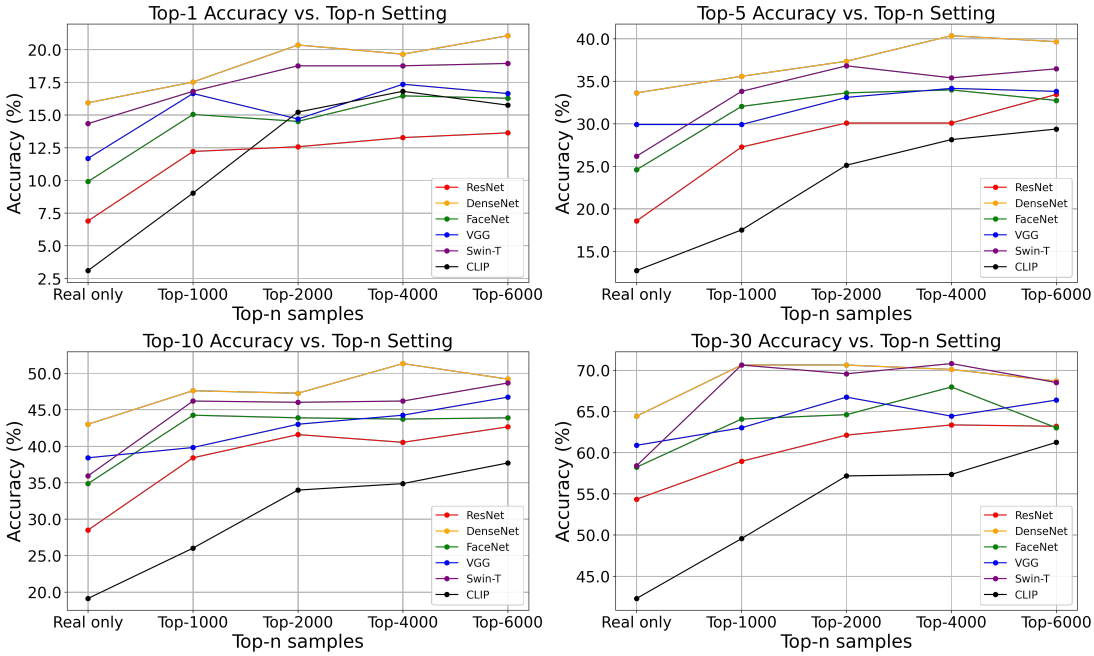


Figure S8. **Top-k accuracy comparison using DreamBooth-generated data.** Each subplot shows Top-1, Top-5, Top-10, and Top-30 accuracy across synthetic cutoffs for six backbone models. DreamBooth augmentation improves performance across most settings.

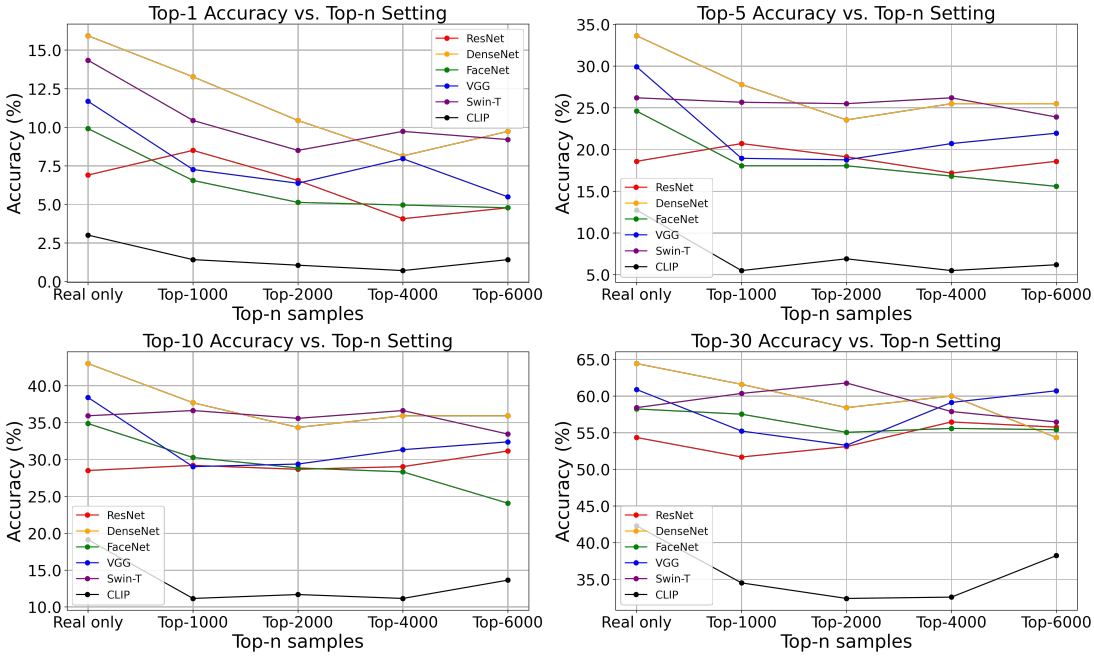


Figure S9. **Top-k accuracy comparison using FastGAN-generated data.** Each subplot shows Top-1, Top-5, Top-10, and Top-30 accuracy across synthetic cutoffs for six backbone models. Compared to DreamBooth, FastGAN augmentation results in less consistent or degraded performance across most settings.

F.2. Few-shot learning

Based on prior results, we focus our few-shot learning analysis on DreamBooth-augmented data. This choice is motivated by the consistent advantages of DreamBooth samples over FastGAN in terms of semantic fidelity, preservation of disease-relevant facial traits, and interpretability. These properties are especially important in ultra-low-shot settings, where maximizing the realism and diagnostic relevance of synthetic data is critical for generalization. Table S7 summarizes performance across multiple backbones under synthetic data augmentation. Each pair of rows shares the same pretrained backbone, where the first row (e.g., ResNet) uses only real training data, and the second row (e.g., ResNet (dream)) incorporates DreamBooth-generated synthetic images for data augmentation.

Table S7. Few-shot classification accuracies (%) under 5-way, 10-way, and 15-way settings with 1-shot and 5-shot support and query sets.

ACC (%)	**5-way**		**10-way**		**15-way**	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
ResNet	24.18 (2.56)	–	18.15 (2.58)	–	17.99 (1.26)	–
ResNet (dream)	**25.72 (1.62)**	33.62 (1.91)	**22.21 (2.60)**	22.63 (2.11)	**18.22 (1.91)**	20.04 (2.77)
DenseNet	26.20 (2.01)	–	17.36 (1.66)	–	**17.30 (3.45)**	–
DenseNet (dream)	**29.88 (1.51)**	33.40 (2.02)	**19.79 (3.19)**	24.43 (3.82)	16.06 (2.77)	18.35 (2.24)
FaceNet	**25.16 (4.89)**	–	12.93 (2.33)	–	8.03 (1.40)	–
FaceNet (dream)	23.60 (4.06)	28.30 (2.92)	**13.17 (2.33)**	17.31 (3.47)	**9.07 (1.44)**	12.17 (1.24)
VGG	**21.54 (3.72)**	–	10.82 (2.14)	–	9.05 (2.20)	–
VGG (dream)	21.02 (5.85)	26.76 (3.42)	**13.67 (2.08)**	13.68 (1.94)	**9.59 (1.48)**	11.20 (1.59)
Swin-T	22.24 (2.88)	–	10.94 (2.69)	–	8.03 (2.41)	–
Swin-T (dream)	**26.72 (4.34)**	25.78 (4.91)	**13.30 (1.56)**	13.43 (1.94)	**9.57 (2.57)**	8.82 (0.97)
CLIP	**23.48 (5.28)**	–	**12.33 (1.96)**	–	7.30 (1.90)	–
CLIP (dream)	22.30 (2.63)	24.80 (3.48)	12.14 (1.98)	13.04 (2.08)	**8.66 (2.82)**	8.11 (1.32)

F.3. Observations

Standard supervised classification The results in Table S5 and Table S6 highlight a clear distinction in the effectiveness of different generative models for data augmentation in classifications. DreamBooth consistently enhances model performance across Top-k accuracies and Top-n settings, demonstrating its ability to produce high-quality, semantically aligned images that support generalization. In contrast, FastGAN shows less consistent gains and, in some cases, degrades performance as more synthetic data is added—indicating a lower signal-to-noise ratio in the generated samples. These findings underscore the importance of not only using synthetic data, but selecting the right generation method.

Few-shot learning Table S7 further illustrates how the utility of synthetic augmentation depends on the interaction between model architecture and data quality. While DreamBooth-generated samples consistently improve performance, gains vary across backbones, with CNNs such as DenseNet and ResNet benefiting more than transformer-based models like Swin Transformer or CLIP. Additionally, the performance gap between 1-shot and 5-shot conditions emphasizes the value of even modest increases in labeled support. These results suggest that few-shot learning with synthetic data is most effective when both model and augmentation strategy are jointly optimized for the task.

Overall, high semantic fidelity, structural consistency, and disease-relevant visual features appear essential for synthetic augmentation to benefit rare disease classification, especially in ultra-low-shot regimes.

G. VLM-based report generation

G.1. Prompt design

To ensure consistency and clinical interpretability of the generated phenotype reports, we designed a structured prompt tailored to the rare disease diagnosis task. The prompt guides the multimodal vision language model (VLM) to assume the role of a professional clinical geneticist and produce structured, concise diagnostic reports based on input facial images. The prompt emphasizes anatomical coverage, medical reasoning, and alignment with the supervised landmark annotations used in other parts of our study. A prompt template is shown in Figure S10.

```
You are a professional clinical geneticist. Given the facial image, generate a concise
diagnostic report focusing on facial features.

**Instructions:**
  - Begin with an overall impression paragraph summarizing the patient's facial appearance.
  - Then, for each facial region, write a single paragraph to describe the facial features.
  - Finish with a prediction of potential disease and a diagnostic suggestion or next steps.

**Facial Diagnostic Report:**
    Overall Impression:
    Left Eye:
    Right Eye:
    Nose:
    Mouth/Lips:
    Diagnostic Suggestion:
```

Figure S10. **Prompt template for VLM-based report generation.**

G.2. Report evaluation

To evaluate the semantic consistency of synthetic facial images, we used a multimodal large language model (VLM) to generate structured phenotype reports for real and DreamBooth-generated images. This analysis assesses whether synthetic faces reflect clinically meaningful traits of their target disease class. We excluded FastGAN images, as they are generated unconditionally and lack one-to-one correspondence with real patients. Since the VLM-based evaluation depends on matched image pairs for assessing regional semantic consistency, it is not applicable to class-agnostic generative models.

To illustrate this evaluation process, we present selected examples spanning high and low semantic similarity scores. For each example, we compare the phenotype reports generated from a real image and a DreamBooth-generated synthetic image of the same disease class. Region-level descriptions are extracted from both reports. Matching phenotype terms are highlighted in green, while contradictory or inconsistent terms are shown in red. The examples cover a range of semantic similarity levels, allowing qualitative inspection of both faithful and divergent generations. Detailed comparisons are shown in Figure S11, S12, S13, and S14.

Rare Disease Facial Image Report

Overall Impression: The patient exhibits a distinctive facial appearance characterized by a broad forehead, prominent nasal bridge, and a relatively flat midface. The eyes appear to be set apart, and there is a noticeable lack of upper eyelid fold. The nose is wide with a broad base, and the mouth appears to be slightly open.

Left Eye: The left eye appears to have a mild ptosis (drooping) and a lack of upper eyelid fold.

Right Eye: The right eye also shows signs of ptosis and a lack of upper eyelid fold, similar to the left eye.

Nose: The nose is broad with a wide base and a relatively prominent nasal bridge.

Mouth/Lips: The mouth is slightly open, and the lips appear thin.

Diagnostic Suggestion: The patient's facial features—broad forehead, wide nose, and dental abnormalities—suggest Apert syndrome, a rare genetic disorder caused by FGFR2 mutations. Genetic testing and a multidisciplinary evaluation are recommended for confirmation and management.

Synthetic Image Report

Overall Impression: The patient exhibits a distinctive facial appearance characterized by a broad forehead, prominent nasal bridge, and a slightly upturned mouth. The cheeks appear full, and there is a noticeable asymmetry in the facial structure.

Left Eye: The left eye appears slightly smaller than the right eye, with a mild ptosis (drooping) noted. The eyelid margin is smooth, and there is no obvious epicanthal fold.

Right Eye: The right eye is larger and more prominent compared to the left. The eyelid margin is smooth.

Nose: The nose is broad with a prominent nasal bridge. The tip of the nose is slightly upturned, and there is a mild deviation to the right side.

Mouth/Lips: The mouth is slightly open, revealing a full set of teeth. The lips are full, and the corers of the mouth are turned upwards. There is a slight asymmetry in the mouth opening, with the left side appearing slightly wider.

Diagnostic Suggestion: The patient's facial features—broad face, upturned mouth, and small nose—are suggestive of Down syndrome; genetic testing (e.g., karyotype or aCGH) and a comprehensive evaluation by a pediatrician or geneticist are recommended for confirmation and further assessment.

Figure S11. **High-Similarity Case.** The real and synthetic reports both describe ptosis, a broad nose, and slighly open mouth appearance. Minor deviations in phrasing are present, but the overall diagnostic impression remains aligned.

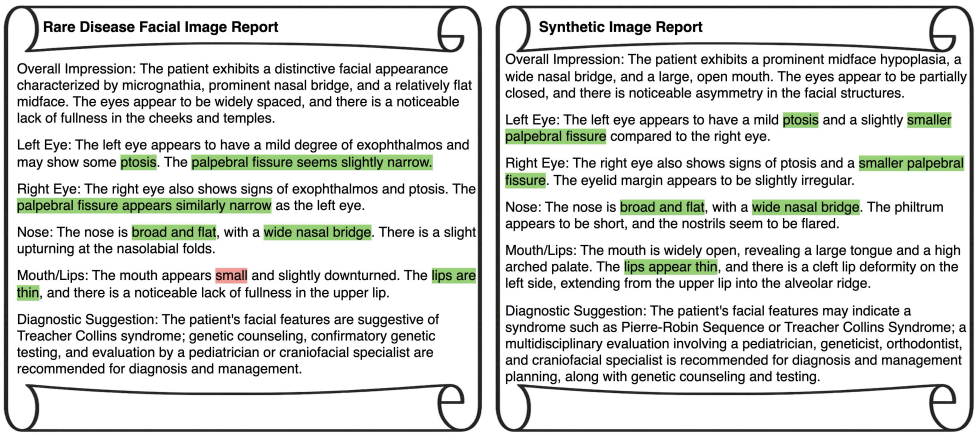


Figure S12. **High-Similarity Case.** Side-by-side comparison of phenotype reports from a real and synthetic image of the same disease class. The synthetic report closely mirrors the real one, especially in descriptions of palpebral fissures, ptosis, nasal shape, and lip structure.

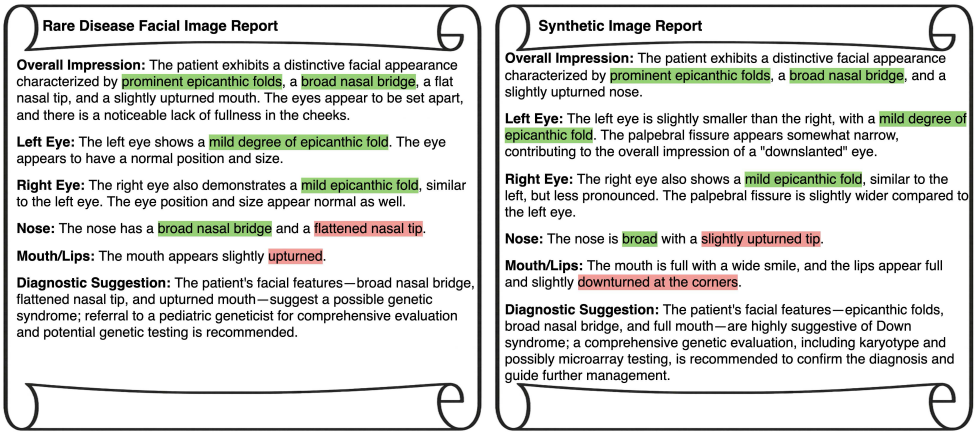


Figure S13. **Mixed-Agreement Case.** Although there is some alignment in features like nasal shape and bridge width, the synthetic report diverges in descriptors of lip configuration and diagnostic suggestion, showing partial inconsistency.

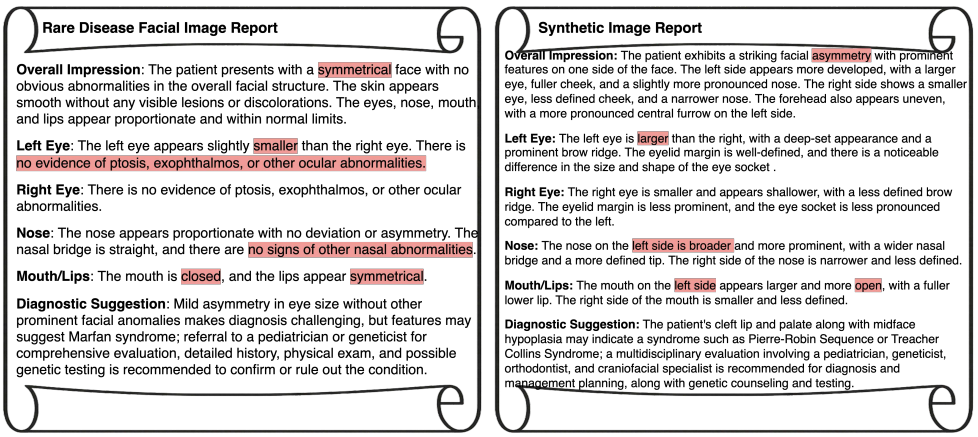


Figure S14. **Low-Similarity Case.** The real report describes normal symmetry, while the synthetic report notes asymmetry and deviation, showing the VLM interprets synthetic structure differently.

BioBERT semantic similarity analysis We report the detailed results of the semantic similarity analysis between real and synthetic phenotype descriptions. Similarity scores were computed using BioBERT embeddings across five facial regions and an overall report segment. Table S8 summarizes the mean and standard deviation of cosine similarities across the dataset. The highest alignment was observed in the nose and eye regions, while the mouth/lips showed slightly greater variability.

TF-IDF comparison To complement the analysis presented above, we conducted a parallel evaluation using a traditional text similarity method based on TF-IDF [2] (Table S8). For each facial region and the overall report, we computed cosine similarity between real and synthetic descriptions using scikit-learn [4]'s `TfidfVectorizer`.

Table S8. Semantic similarity and TF-IDF-based semantic similarity across five facial regions.

Region	BioBERT Similarity Score	TF-IDF Similarity Score
Overall	0.8404 (0.0748)	0.7630 (0.0707)
Left Eye	0.7485 (0.1228)	0.4364 (0.1358)
Right Eye	0.7535 (0.1423)	0.4578 (0.1352)
Nose	0.7712 (0.1344)	0.4315 (0.1432)
Mouth/Lips	0.7355 (0.1376)	0.4612 (0.1321)

These findings support the need for domain-specific semantic models when evaluating medical text. TF-IDF–based comparisons fail to fully capture conceptual similarity in phenotype language. While overall similarity trends are consistent, the lower region-wise scores highlight the limitations of surface-level lexical methods when applied to clinical narratives.

G.3. Uncertainty and robustness analysis

Figure S15 complements the uncertainty analysis in the main text by visualizing the distribution of uncertainty scores across all samples. 75% of the cases exhibit low uncertainty (< 0.14), indicating stable and consistent phenotype descriptions under stochastic sampling.

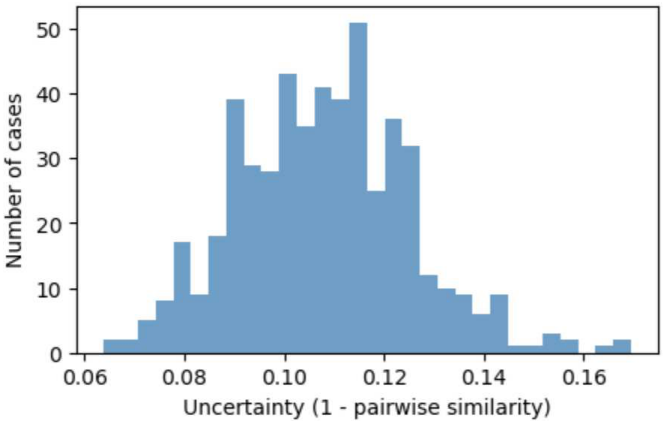


Figure S15. **Distribution of uncertainty scores across all cases.**

To further assess robustness, Table S9 reports cross-model phenotype similarity between Qwen2.5-VL and LLaVA-NeXT. The results show consistent similarity ranges across real and synthetic images, supporting that the evaluation is not sensitive to the choice of VLM.

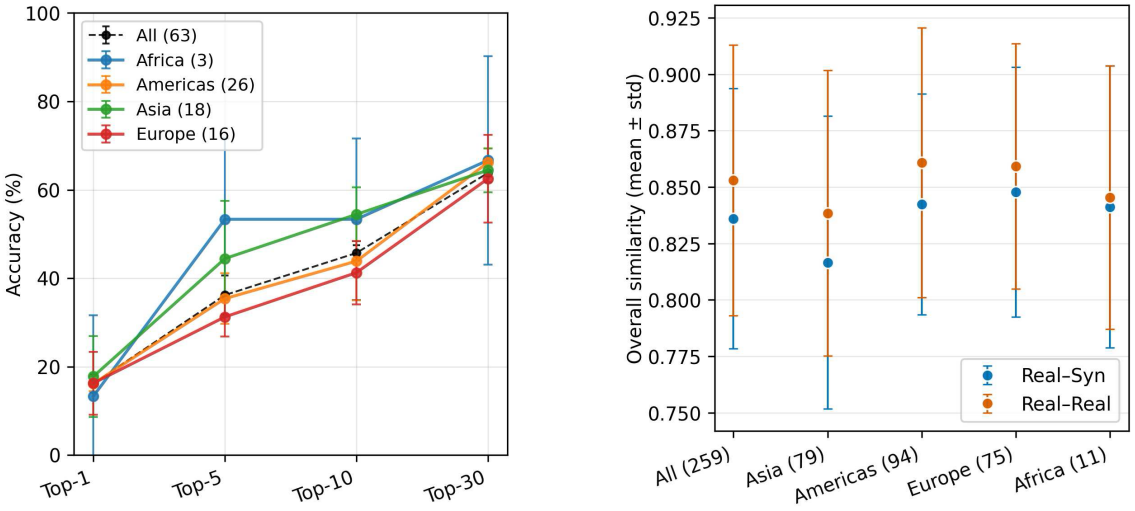
Table S9. Cross-model phenotype similarity.

	LLaVA (Real)	LLaVA (Synthetic)
Qwen (Real)	0.7053 (0.0711)	0.7176 (0.0691)
Qwen (Synthetic)	0.7204 (0.0726)	0.7355 (0.0730)

H. Regional analysis and potential bias

To assess potential bias in both classification performance and synthetic data evaluation, we analyze model behavior across geographic regions. However, population-level demographic attributes (e.g., ethnicity, ancestry, or skin tone) are not available in our dataset, as web-scraped rare disease case reports rarely provide standardized annotations. As a proxy, we stratify the data by geographic region (the only consistently recoverable attribute) and group samples into four regions: **Africa**, **Americas**, **Asia**, and **Europe**.

Regional diagnosis performance. We report Top-k accuracy for supervised learning across regions in Figure S16a. While minor variations are observed at lower k, performance trends are broadly consistent across regions and converge as k increases, indicating similar generalization behavior as the results in main text.



(a) Regional diagnosis results across Top-k accuracies.

(b) Regional phenotype similarity comparison.

Figure S16. **Regional analysis across geographic groups.**

Regional phenotype similarity. We further evaluate phenotype similarity across regions using VLM-generated reports (Figure S16b). Real–synthetic similarity is comparable to real–real similarity across all regions within statistical uncertainty, suggesting that synthetic data preserves phenotype characteristics without introducing substantial regional bias. Notably, the Americas region shows slightly higher similarity scores, which may reflect a larger representation of cases from this region in the dataset. However, the overall consistency across regions supports the generalizability of our findings.

Limitations. We note that geographic region is only a coarse proxy and does not directly correspond to demographic attributes such as skin tone or ancestry. In addition, potential collection bias in publicly available case imagery may influence both classification and generative models. Future work should prioritize the collection of more diverse and well-annotated datasets to enable more granular analysis of demographic bias and ensure equitable performance across all patient populations and the development of synthetic data generation methods that explicitly account for demographic diversity.

I. Statistical reporting details

All reported results in this paper include 1-sigma error bars, expressed as the standard deviation in parentheses (e.g., 6.90 (1.45)), computed over 5-fold cross-validation. The primary source of variability is the random train/test split across folds. For each configuration, we evaluate performance on the held-out fold and report the sample mean and standard deviation across the five runs. We assume approximately normally distributed metrics across folds, which is common in classification evaluation under low-data regimes. No additional sources of randomness (e.g., weight initialization or stochastic sampling) are varied unless explicitly stated.

Standard deviation is calculated using the unbiased estimator (i.e., Bessel's correction):

$$\mathrm{std}(x_1, \ldots, x_n) = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

This computation is implemented via NumPy [1]'s `np.std(..., ddof=1)` function.

J. Hyperparameter settings and training details

We summarize the key hyperparameters used in our benchmark experiments across three experimental settings in Table S10, S11, S12, and S13.

J.1. Standard supervised classification

Table S10. Hyperparameters for standard supervised classification experiments.

Parameter	**Value**
Optimizer	Adam
Learning rate	$1e^{-4}$
Batch size	32
Training folds	5-fold class split
Number of epochs	50
Loss function	CrossEntropyLoss
Weight decay	$1e^{-4}$
Image size	224×224
Pretrained backbone	ImageNet for ResNet, DenseNet, Swin-T, CLIP VggFace2 for FaceNet VggFace for VGG

J.2. Few-Shot learning

Table S11. Hyperparameters for Prototypical Networks.

Parameter	**Value**
Optimizer	Adam
Learning rate	$1e^{-3}$
Batch size	1
Training folds	5-fold class split
Episodes per fold	600 train / 100 val / 200 test
Loss function	CrossEntropyLoss
Distance metric	Euclidean Distance

J.3. Synthetic data generation

Table S12. Training settings for DreamBooth per disease class.

Parameter	Value / Description
Base model	SG161222/Realistic_Vision_V5.1_noVAE
Training setting	Class-conditioned (per disease)
Text prompt	`"a child with [DISEASE] disease"`
Training steps	800 per class
Batch size	1
Learning rate	$1e^{-6}$
Image resolution	512×512
Mixed precision	fp16
Safety checker	NSFW filter

Table S13. Training settings for FastGAN on pooled real images.

Parameter	Value / Description
Architecture	Original FastGAN (skip-layer excitation)
Training setting	Unconditional (no class labels)
Iterations	80,000
Batch size	8
Optimizer	Adam
Learning rate	$2e^{-4}$
Image resolution	512×512

J.4. Hardware and compute resources

Standard supervised classification, few-shot learning, DreamBooth, and FastGAN models were all implemented in PyTorch [3] and trained using a single NVIDIA A100 GPU with 80GB of memory (also used for VLM-based report generation). Specifically, standard supervised classification and few-shot learning required approximately 240 and 550 minutes of training time, respectively, for downstream analysis. DreamBooth required 30–50 minutes of training per class, depending on the number of input images and prompt complexity, while FastGAN training took approximately 15 hours to converge.

References

[1] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, 2020. 20

[2] Christopher D Manning. *An introduction to information retrieval*. Cambridge University Press, 2009. 18

[3] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. *PyTorch: an imperative style, high-performance deep learning library*. Curran Associates Inc., Red Hook, NY, USA, 2019. 21

[4] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. 18