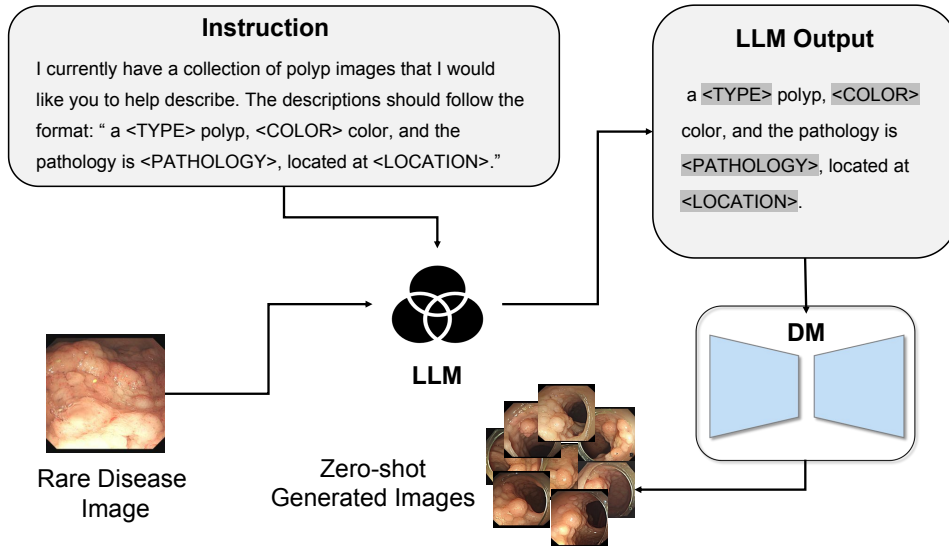


SUPPLEMENTARY INFORMATION

1 Supplementary Figures



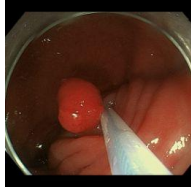
Supplementary Fig. 1: Zero-Shot method Pipeline. Rare disease images are provided as input and processed through a large language model (LLM) using zero-shot prompting. The instruction specifies the target description format ("a TYPE polyp, COLOR color, and the pathology is PATHOLOGY, located at LOCATION"). The LLM generates structured textual outputs, which are further integrated with a diffusion model (DM) to produce zero-shot generated polyp images.

INPUT IMAGE**LLM OUTPUT**

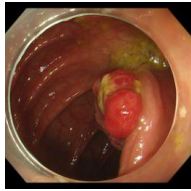
a sessile polyp, pink color, and the pathology is undetermined, located at the colon.



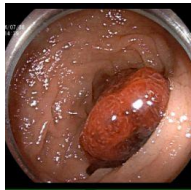
a pedunculated polyp, red color, and the pathology is undetermined, located at the colon.



a pedunculated polyp, bright red color, and the pathology is undetermined, located at the sigmoid colon.

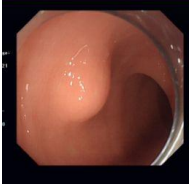
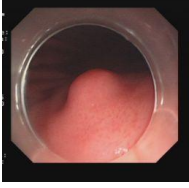
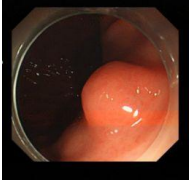


a pedunculated polyp, red color with yellowish exudate, and the pathology is undetermined, located at the descending colon.



a sessile polyp, dark reddish-brown color, and the pathology is undetermined, located at the transverse colon.

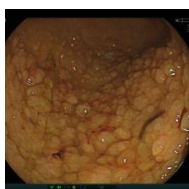
Supplementary Fig. 2: Zero-shot description examples for Juvenile Polyposis Syndrome (JPS). Input endoscopic images (left) are provided to the LLM, which generates structured textual descriptions (right) following the instruction template. These outputs serve as text conditions to guide the diffusion model for rare-lesion image synthesis (see Fig. 1).

INPUT IMAGE	LLM OUTPUT
	a sessile polyp, light brownish color, and the pathology is undetermined, located at the transverse colon.
	a sessile polyp, pale pink color, and the pathology is undetermined, located at the sigmoid colon.
	a sessile polyp, smooth pink color, and the pathology is undetermined, located at the descending colon.
	a sessile polyp, light pink color, and the pathology is undetermined, located at the ascending colon.
	a pedunculated polyp, smooth reddish color, and the pathology is undetermined, located at the sigmoid colon.

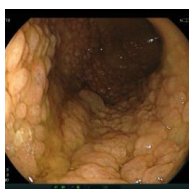
Supplementary Fig. 3: Zero-shot description examples for Calcifying Fibrous Tumor (CFT). Input endoscopic images (left) are provided to the LLM, which generates structured textual descriptions (right) following the instruction template. These outputs serve as text to guide the diffusion model for rare-lesion image synthesis (see Fig. 1).

INPUT IMAGE**LLM OUTPUT**

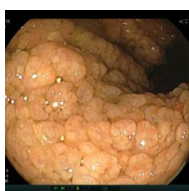
a nodular polyp, orange color, and the pathology is serrated adenoma, located at the gastric antrum.



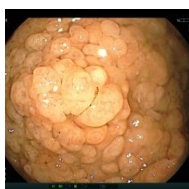
a villous polyp, yellowish color, and the pathology is tubulovillous adenoma, located at the rectum.



a villous polyp, yellowish color, and the pathology is tubulovillous adenoma, located at the descending colon.

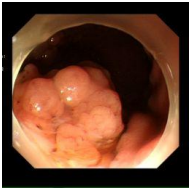


a villous polyp, yellowish color, and the pathology is tubulovillous adenoma, located at the sigmoid colon.



a villous polyp, yellowish color, and the pathology is tubulovillous adenoma, located at the rectosigmoid junction.

Supplementary Fig. 4: Zero-shot description examples for Familial Adenomatous Polyposis (FAP). Input endoscopic images (left) are provided to the LLM, which generates structured textual descriptions (right) following the instruction template. These outputs serve as text conditions to guide the diffusion model for rare-lesion image synthesis (see Fig. 1).

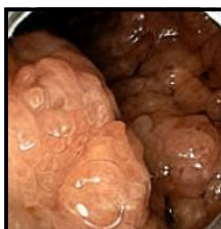
INPUT IMAGE	LLM OUTPUT
	a lobulated polyp,pinkish-yellow color,and the pathology is villous adenoma,located at the rectum.
	a villous polyp,dark reddish color,and the pathology is villous adenoma,located at the sigmoid colon.
	a villous polyp,reddish color,and the pathology is villous adenoma,located at the descending colon.
	a villous polyp,reddish-brown color,and the pathology is villous adenoma,located at the sigmoid colon.
	a villous polyp,light reddish color,and the pathology is villous adenoma,located at the descending colon.

Supplementary Fig. 5: Zero-shot description examples for Peutz–Jeghers Syndrome (PJS). Input endoscopic images (left) are provided to the LLM, which generates structured textual descriptions (right) following the instruction template. These outputs serve as text conditions to guide the diffusion model for rare-lesion image synthesis (see Fig. 1).

Ground Truth
Disease Image



DreamBooth Generated Samples

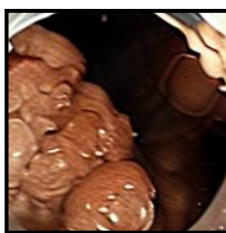


Supplementary Fig. 6: Results from the DreamBooth. Images are representative synthetic samples from the 100 images generated per prototype, with similar results throughout; the same applies to Supplementary Figs. 7–12.

Ground Truth
Disease Image



Textual Inversion Generated Samples



Supplementary Fig. 7: Results from the Textual Inversion model.

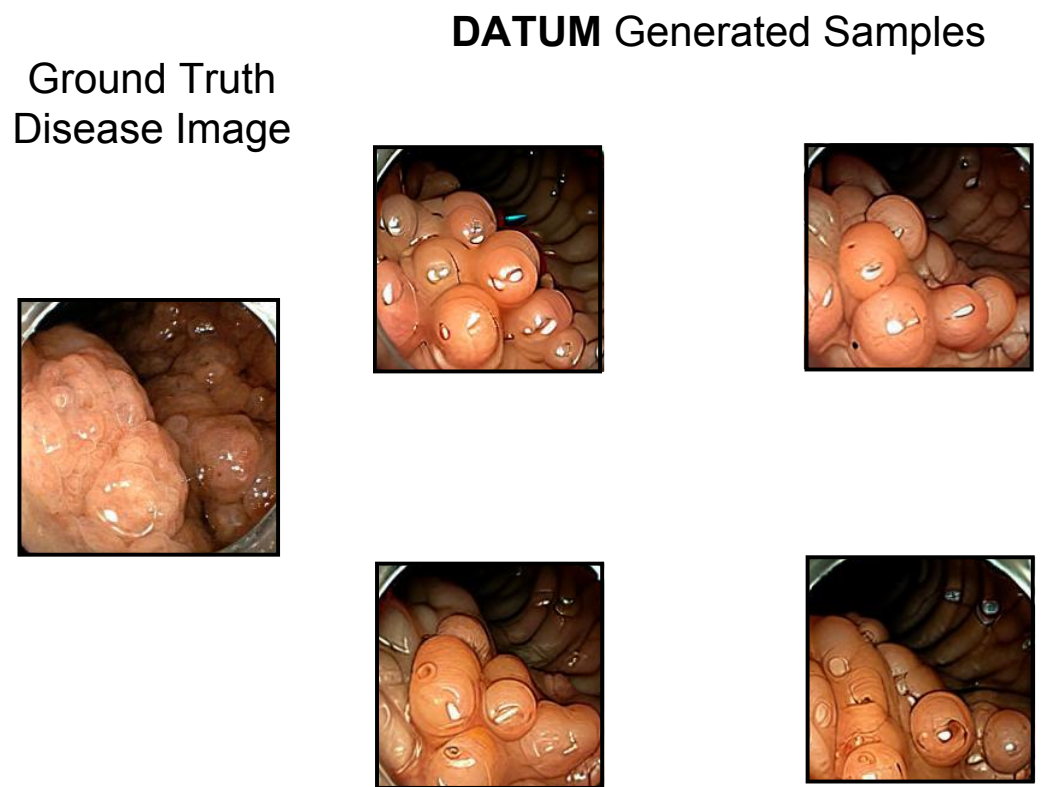
Ground Truth
Disease Image



Zero Shot Generated Samples



Supplementary Fig. 8: Results from the Zero shot Method.

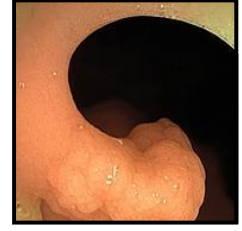
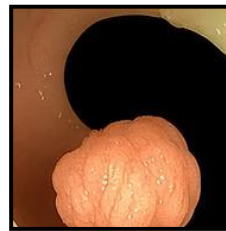
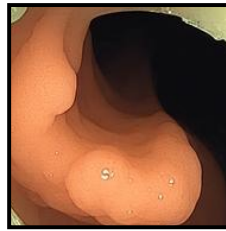


Supplementary Fig. 9: Results from the DATUM.

Ground Truth
Disease Image



SeedSelect Generated Samples

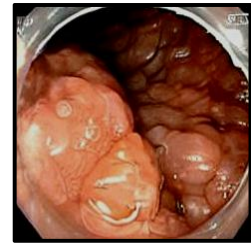
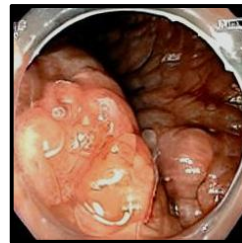
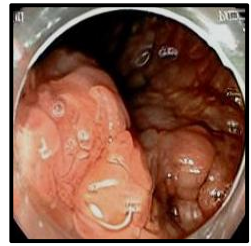
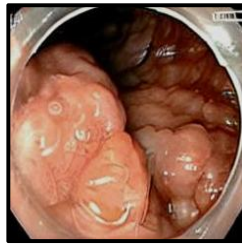


Supplementary Fig. 10: Results from the Seed Select.

Ground Truth
Disease Image



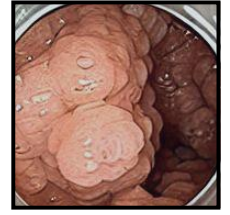
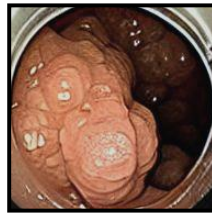
Finetune SD Generated Samples



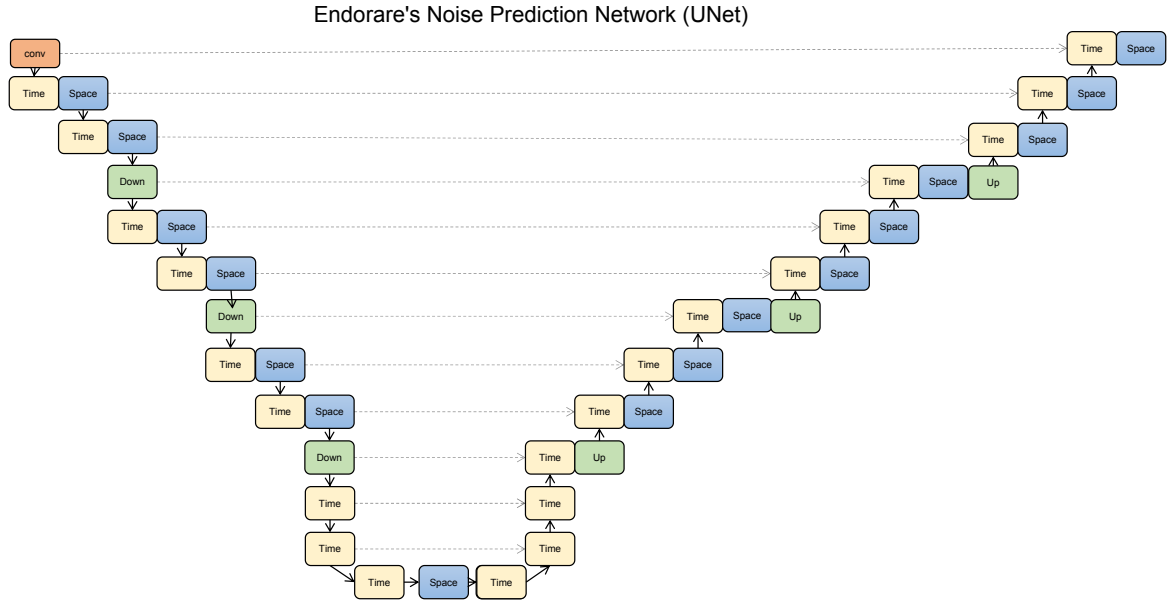
Supplementary Fig. 11: Results from the Finetune SD.

Ground Truth
Disease Image

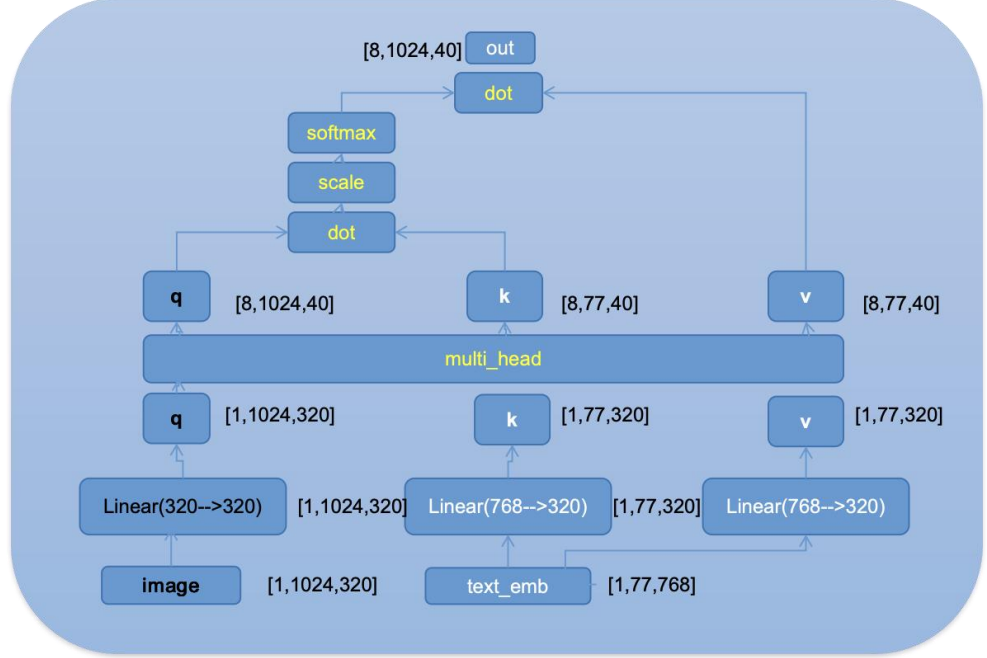
EndoRare Generated Samples



Supplementary Fig. 12: Results from the EndoRare.

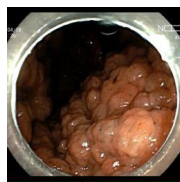


Supplementary Fig. 13: The schematic illustrates the U-Net backbone employed to predict noise ε from a noisy input $\mathbf{x}_t$ at timestep t .

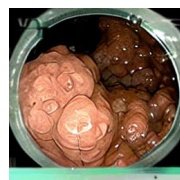
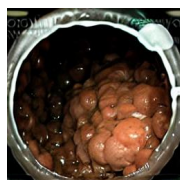
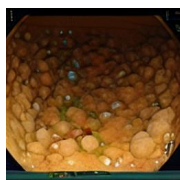


Supplementary Fig. 14: Schematic of the cross-attention mechanism within the U-Net denoiser. The intermediate feature maps $\phi(\mathbf{x}_t)$ from the U-Net decoder are projected to form the **Query** (Q). The conditional context embedding (e.g., text prompts encoded by a CLIP text encoder) $\mathbf{c}$ is projected to form the **Key** (K) and **Value** (V). The attention output $\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V$ is then fused back into the U-Net features, guiding the denoising process towards the conditional input.

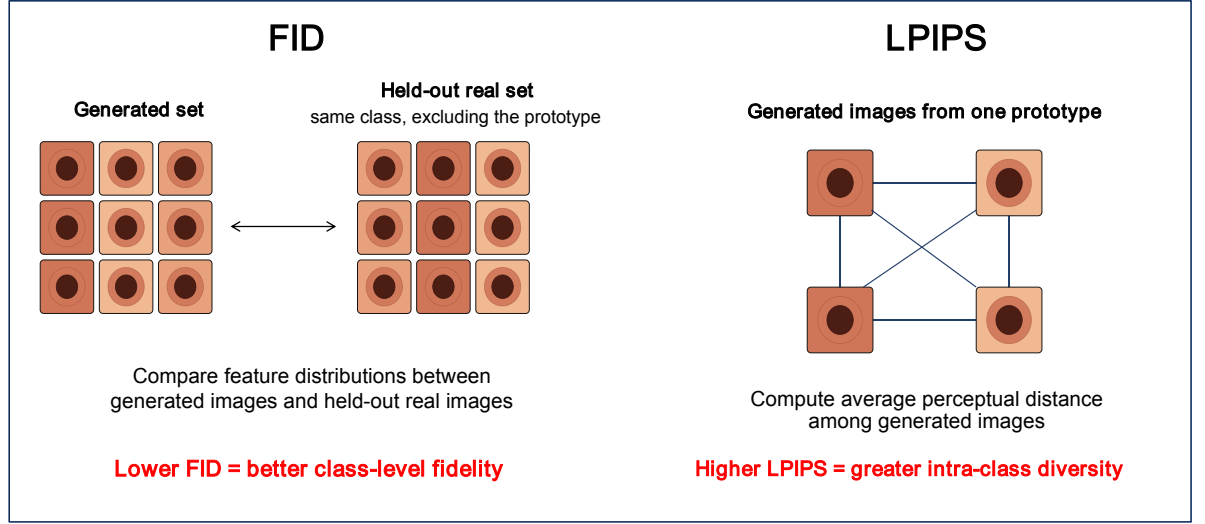
Ground Truth



**Generated
Sample by PSE**



Supplementary Fig. 15: Image Reconstruction via Learned Prototype-Specific Embedding (PSE). **Top:** Ground truth images. **Bottom:** Reconstructed images from the learned PSE.



Supplementary Fig. 16: Computation scheme for FID and intra-class LPIPS. FID was computed by comparing the distribution of generated images with held-out real images from the same lesion class, excluding the reference prototype used for one-shot generation. Intra-class LPIPS was computed as the average pairwise perceptual distance among generated images within the same prototype-conditioned subset.

2 Supplementary Tables

Supplementary Table 1: Summary of Clinical Features for Juvenile Polyposis Syndrome (JPS), Calcifying Fibrous Tumor (CFT), Familial Adenomatous Polyposis (FAP), and Peutz–Jeghers Syndrome (PJS).

Feature	JPS	CFT	FAP	PJS
Incidence rate	1 in 100,000 individuals	Rare; exact incidence unknown	1 in 8000 to 1 in 10,000 individuals	1 in 25,000 to 1 in 280,000 individuals
Main cause	Mutation in <i>SMAD4</i> or <i>BMPR1A</i>	Mostly sporadic; rare JAK–STAT–related (e.g., <i>JAK1</i>)	<i>APC</i> gene mutation	Mutation in <i>STK11</i>
Primary risk	Colorectal and gastric cancer	Systemic multiple calcifying fibrous tumors may occur	Colorectal, gastric, and duodenal cancer	Gastrointestinal and breast cancer
Age of onset	Childhood; often under 20	Median 22.5 y (14–34)	Teens to early adulthood	Childhood; often around 10 years

Supplementary Table 2: Quantitative comparison on **Juvenile Polyposis Syndrome (JPS)** across different one-shot rare-lesion synthesis methods. We report FID (lower is better), CLIP-I similarity (higher is better), and LPIPS diversity (higher is better) to jointly evaluate image fidelity, semantic alignment, and intra-class variability.

Method	FID ($\downarrow$)	CLIP-I ($\uparrow$)	LPIPS ($\uparrow$)
Finetune SD	250.3971	0.8832	0.2254
Textual Inversion	363.6228	0.7270	0.5008
DreamBooth	244.3869	0.8975	0.2305
DATUM	311.6627	0.8485	0.3865
SeedSelect	319.2532	0.8445	0.3632
Zero Shot	247.5746	0.8544	0.5017
Ours	276.8537	0.8335	0.4745

Supplementary Table 3: Quantitative comparison on **Calcifying Fibrous Tumor (CFT)** across different one-shot rare-lesion synthesis methods. The table shows FID, CLIP-I similarity, and LPIPS diversity, reflecting visual realism, class faithfulness, and diversity of generated samples relative to the clinical reference.

Method	FID ($\downarrow$)	CLIP-I ($\uparrow$)	LPIPS ($\uparrow$)
Finetune SD	275.3940	0.9115	0.1958
Textual Inversion	368.2366	0.7523	0.4270
DreamBooth	312.5440	0.9134	0.2231
DATUM	318.5397	0.8416	0.3991
SeedSelect	328.7758	0.8633	0.3466
Zero Shot	279.3189	0.8653	0.4982
Ours	295.0912	0.8615	0.4353

Supplementary Table 4: Quantitative comparison on **Familial Adenomatous Polyposis (FAP)** using competing generative strategies. Results are summarized in terms of FID (image fidelity), CLIP-I (semantic similarity), and LPIPS (sample diversity), highlighting differences in quality and variability among approaches.

Method	FID ($\downarrow$)	CLIP-I ($\uparrow$)	LPIPS ($\uparrow$)
Finetune SD	284.2080	0.8919	0.2920
Textual Inversion	386.4716	0.7634	0.5148
DreamBooth	276.3108	0.9198	0.2041
DATUM	420.3747	0.7833	0.5062
SeedSelect	401.9167	0.8518	0.3839
Zero Shot	381.9915	0.8313	0.5059
Ours	300.3042	0.8427	0.5199

Supplementary Table 5: Quantitative comparison on **Peutz–Jeghers Syndrome (PJS)** with multiple rare-lesion synthesis methods. Evaluation metrics include FID, CLIP-I, and LPIPS, which together assess generation fidelity, clinical concept alignment, and intra-class diversity.

Method	FID ($\downarrow$)	CLIP-I ($\uparrow$)	LPIPS ($\uparrow$)
Finetune SD	243.7507	0.8965	0.2628
Textual Inversion	372.5007	0.7491	0.5218
DreamBooth	180.4905	0.9348	0.1999
DATUM	374.9242	0.7844	0.4471
SeedSelect	334.1716	0.8418	0.3827
Zero Shot	330.3781	0.8265	0.4944
Ours	248.7765	0.8491	0.5249

Supplementary Table 6: Blinded clinical expert scores (3-point scale, higher is better) across four rare entities synthesized by different one-shot methods.

Method	JPS	CFT	FAP	PJS
Finetune SD	2.25	2.04	2.05	2.00
Textual Inversion	1.05	1.23	1.00	1.00
DreamBooth	2.05	2.00	2.00	2.00
SeedSelect	1.31	1.28	1.21	1.08
DATUM	1.61	1.49	1.09	1.40
Zero Shot	2.23	1.97	1.72	1.68
Ours	2.36	2.23	2.40	2.29

Supplementary Table 7: Supplementary multi-site reader-validation analysis. Metrics include Precision, Recall, F1, MCC, and Balanced Accuracy, all reported as percentages. Values are pooled across all four rare entities for each of the $n = 10$ novice endoscopists, who are independent participants assessed once before and once after synthetic exposure on the same fixed image set. The bottom block reports the mean improvement (Δ , Post–Pre), its 95% confidence interval and the p value of a two-sided paired t -test computed on the $n = 10$ reader-level values, with 9 degrees of freedom and no adjustment for multiple comparisons. The corresponding t statistics are $t(9) = 5.16$ for precision, 18.98 for recall, 9.95 for F1, 9.86 for MCC and 18.75 for balanced accuracy.

Exp	Cond.	Precision	Recall	F1	MCC	Bal.Acc.
Reader 1	Pre	64.29	56.25	60.00	59.76	77.97
	Post	68.18	93.75	78.95	79.73	96.66
Reader 2	Pre	35.00	43.75	38.89	38.45	71.47
	Post	82.35	87.50	84.85	84.73	93.66
Reader 3	Pre	66.67	50.00	57.14	57.37	74.88
	Post	88.24	93.75	90.91	90.86	96.81
Reader 4	Pre	52.94	56.25	54.55	54.10	77.88
	Post	83.33	93.75	88.24	88.27	96.78
Reader 5	Pre	100.00	75.00	85.71	86.49	87.50
	Post	100.00	100.00	100.00	100.00	100.00
Reader 6	Pre	71.43	62.50	66.67	66.51	81.13
	Post	100.00	100.00	100.00	100.00	100.00
Reader 7	Pre	71.43	62.50	66.67	66.51	81.13
	Post	94.12	100.00	96.97	96.98	99.97
Reader 8	Pre	50.00	43.75	46.67	46.27	71.66
	Post	93.75	93.75	93.75	93.69	96.84
Reader 9	Pre	52.94	56.25	54.55	54.10	77.88
	Post	83.33	93.75	88.24	88.27	96.78
Reader 10	Pre	58.33	43.75	50.00	50.09	71.72
	Post	76.47	81.25	78.79	78.61	90.50
Mean	Pre	62.30	55.00	58.08	57.97	77.32
Mean	Post	86.98	93.75	90.07	90.11	96.80
Δ (Post–Pre)	—	24.68	38.75	31.98	32.15	19.48
Δ 95% CI	—	[13.86, 35.49]	[34.13, 43.37]	[24.71, 39.26]	[24.77, 39.52]	[17.13, 21.83]
Paired t p	—	5.9×10^{-4}	1.4×10^{-8}	3.7×10^{-6}	4.0×10^{-6}	1.6×10^{-8}
Significance	—	Yes	Yes	Yes	Yes	Yes

Readers are novice endoscopists from Zhongshan Hospital, Zhongshan Hospital Xiamen Branch, Zhongshan Geriatric Medical Center, and the First People's Hospital of Nanchang. Metrics are reported as percentages.

Supplementary Table 8: Juvenile polyposis syndrome (JPS): ROC AUC (mean and variance).

Method	Mean	Variance
No Augment	0.0449	0.0003
Oversample	0.8964	0.0044
Zero Shot	0.9386	0.0007
DreamBooth	0.8651	0.0024
Seed Select	0.7036	0.0128
Textual Inversion	0.8973	0.0006
DATUM	0.9299	0.0032
Ours	0.9701	0.0002

Supplementary Table 9: Juvenile polyposis syndrome (JPS): PR AUC (mean and variance).

Method	Mean	Variance
No Augment	0.0064	0.0000
Oversample	0.4130	0.0254
Zero Shot	0.5561	0.0084
DreamBooth	0.3223	0.0174
Seed Select	0.0327	0.0002
Textual Inversion	0.3067	0.0150
DATUM	0.3409	0.0098
Ours	0.5628	0.0215

Supplementary Table 10: Calcifying fibrous tumor (CFT): ROC AUC (mean and variance).

Method	Mean	Variance
No Augment	0.4821	0.0039
Oversample	0.9150	0.0059
Zero Shot	0.9274	0.0011
DreamBooth	0.9128	0.0050
Seed Select	0.8680	0.0114
Textual Inversion	0.9033	0.0091
DATUM	0.9534	0.0015
Ours	0.9658	0.0018

Supplementary Table 11: Calcifying fibrous tumor (CFT): PR AUC (mean and variance).

Method	Mean	Variance
No Augment	0.0126	0.0000
Oversample	0.3978	0.0511
Zero Shot	0.2222	0.0061
DreamBooth	0.3546	0.0256
Seed Select	0.2276	0.0579
Textual Inversion	0.3980	0.0459
DATUM	0.3750	0.0225
Ours	0.4984	0.0449

Supplementary Table 12: Familial adenomatous polyposis (FAP): ROC AUC (mean and variance).

Method	Mean	Variance
No Augment	0.3408	0.0027
Oversample	0.9586	0.0016
Zero Shot	0.8988	0.0020
DreamBooth	0.9379	0.0046
Seed Select	0.9128	0.0031
Textual Inversion	0.9795	0.0003
DATUM	0.9658	0.0003
Ours	0.9749	0.0003

Supplementary Table 13: Familial adenomatous polyposis (FAP): PR AUC (mean and variance).

Method	Mean	Variance
No Augment	0.0091	0.0000
Oversample	0.7881	0.0010
Zero Shot	0.5419	0.0080
DreamBooth	0.7822	0.0051
Seed Select	0.6369	0.0360
Textual Inversion	0.7955	0.0008
DATUM	0.7540	0.0051
Ours	0.8003	0.0034

Supplementary Table 14: Peutz–Jeghers syndrome (PJS): ROC AUC (mean and variance).

Method	Mean	Variance
No Augment	0.3109	0.0028
Oversample	0.9908	0.0002
Zero Shot	0.9928	0.0000
DreamBooth	0.9874	0.0006
Seed Select	0.9953	0.0000
Textual Inversion	0.9993	0.0000
DATUM	0.9958	0.0000
Ours	0.9996	0.0000

Supplementary Table 15: Peutz–Jeghers syndrome (PJS): PR AUC (mean and variance).

Method	Mean	Variance
No Augment	0.0086	0.0000
Oversample	0.7464	0.0850
Zero Shot	0.5184	0.0069
DreamBooth	0.8250	0.1225
Seed Select	0.7583	0.0075
Textual Inversion	0.9458	0.0117
DATUM	0.7620	0.0168
Ours	0.9700	0.0006

Supplementary Table 16: Downstream classification performance of EndoRare ablation variants. Rows follow the incremental module order used in the ablation. PSE denotes prototype-specific embedding, TLE denotes tailored lesion embedding, and FM denotes fusion module. Delta values are computed relative to the full EndoRare model. Higher PR-AUC, ROC-AUC, and TPR values indicate better performance.

Method	PSE	TLE	FM	PR-AUC		ROC-AUC		TPR at fixed FPR			
				Value	Δ	Value	Δ	1%	5%	10%	20%
Text only	–	–	–	0.3317	-0.3302	0.9015	-0.0752	0.2500	0.5000	0.5000	0.9167
PSE only	✓	–	–	0.4270	-0.2348	0.9042	-0.0725	0.5000	0.5000	0.5833	0.7500
PSE + TLE	✓	✓	–	0.3952	-0.2666	0.9069	-0.0698	0.5000	0.5833	0.6667	0.8333
Full EndoRare	✓	✓	✓	0.6618	0.0000	0.9767	0.0000	0.5833	0.8333	0.9167	1.0000

3 Supplementary Methods

3.1 Algorithm of Our Method

Algorithm 1: EndoRare: Knowledge-Enhanced One-Shot Rare Disease Image Synthesis

Input: Source dataset $\mathcal{D}_S = \{(I_{S_i}, R_{S_i})\}$ with images and reports; target prototype image I_T ; CLIP text encoder τ_θ ; CLIP image encoder; diffusion model ϵ_θ .

Output: Synthetic images $\{\tilde{I}\}$ for the target rare disease.

// (A) Knowledge-Enhanced Diffusion Pretraining

- 1 **foreach** $(I_{S_i}, R_{S_i}) \in \mathcal{D}_S$ **do**
- 2 $Attr_i \leftarrow \text{LLM}(R_{S_i})$ **// a) extract**
 morphology/color/pathology/location
- 3 $y_i \leftarrow \text{format}(Attr_i)$ **// b) template: <MORPHOLOGY>, <COLOR>, <PATHOLOGY>, <LOCATION>**
- 4 $\mathbf{e}_i^k \leftarrow \tau_\theta(attr_i^k)$, $\mathbf{v}_i^k \leftarrow \text{CLIP-Img}(I_{S_i})^k$ **// c) attribute embeddings**
- 5 sample $t \sim \mathcal{U}(\{1, \dots, T\})$, $\epsilon \sim \mathcal{N}(0, 1)$, and obtain latent $\mathbf{z}_t$.
- 6 $\mathcal{L}_{\text{DM}} \leftarrow \|\epsilon - \epsilon_\theta(\mathbf{z}_t, t, \tau_\theta(y_i))\|^2$ **// d) LDM loss**
- 7 $\mathcal{L}_{\text{Align}} \leftarrow \sum_k \|\mathbf{v}_i^k - \mathbf{e}_i^k\|^2$ **// e) cross-modal alignment**
- 8 update ϵ_θ by minimizing $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{Align}} + \lambda \mathcal{L}_{\text{DM}}$.

// (B) One-Shot Prototype-Specific Embedding (PSE)

- 9 $\mathbf{p}_T \leftarrow \tau_\theta(\text{name}(I_T))$ **// initialize from class-name text embedding**
- 10 **for** $s = 1$ **to** S **do**
- 11 sample $t, \epsilon, \mathbf{z}_t$; $\mathcal{L}_{\text{Recon}} \leftarrow \|\epsilon - \epsilon_\theta(\mathbf{z}_t, t, \mathbf{p}_T)\|^2$.
- 12 update *only* $\mathbf{p}_T$ with $\nabla_{\mathbf{p}_T} \mathcal{L}_{\text{Recon}}$; freeze ϵ_θ .

// (C) Tailored Lesion Embedding (TLE) & Fusion

- 13 obtain visual attributes $\{\mathbf{v}_T^k\}$ from I_T via CLIP image encoders.
- 14 format attributes $\rightarrow$ text $\rightarrow \{\mathbf{r}_T^k\}$.
- 15 $s_k \leftarrow \text{sim}(\mathbf{r}_T^k, \mathbf{p}_T)$; $w_k \leftarrow \text{softmax}(\{s_k\})$.
- 16 $\mathbf{e}_T \leftarrow (\sum_k w_k \mathbf{r}_T^k) \oplus \mathbf{p}_T$ **// $\oplus$: concat or weighted average**

// (D) Guided Denoising Sampling

- 17 **for** $n = 1$ **to** N **do**
- 18 sample $\mathbf{z}_T \sim \mathcal{N}(0, \mathbf{I})$.
- 19 **for** $t = T, T-1, \dots, 1$ **do**
- 20 $\hat{\epsilon} \leftarrow \epsilon_\theta(\mathbf{z}_t, t, \mathbf{e}_T)$.
- 21 $\mathbf{z}_{t-1} \leftarrow \sqrt{\bar{\alpha}_{t-1}} \left(\frac{\mathbf{z}_t - \sqrt{1 - \bar{\alpha}_t} \hat{\epsilon}}{\sqrt{\bar{\alpha}_t}} \right) + \sqrt{1 - \bar{\alpha}_{t-1} - \sigma^2} \hat{\epsilon} + \sigma \epsilon$.
- 22 $\tilde{I} \leftarrow \mathcal{D}(\mathbf{z}_0)$; add $\tilde{I}$ to output set.
- 23 **return** $\{\tilde{I}\}$

3.2 EndoRare Implementation Details

The proposed framework is based on Stable Diffusion v1.5, using DDIM sampling for 200 steps and classifier-free guidance with a scale of 7.5. The model is trained with a base learning rate of 1.0×10^{-4} , employing a linear learning rate scheduler that starts at 0.00085 and linearly increases to 0.012 over the course of training. The training is conducted over 1000 timesteps with conditioning applied at each step. The model is trained for a total of 100 epochs, with intermediate states saved regularly using checkpointing to facilitate analysis and model recovery if necessary.

The UNet model with four input channels, corresponding to the image and its associated conditioning information. The UNet architecture includes 320 model channels and employs attention mechanisms at three different resolutions to improve image fidelity. The model also incorporates multiple residual blocks and a spatial transformer for efficient handling of spatial information. We also adopt ChatGPT-4o as the LLM component in our pipeline.

The vision concept learning stage of the model employs the CLIP-ViT-Large-Patch14 encoder, trained for 50 epochs with a learning rate of 1×10^{-3} . This encoder aligns visual features with textual descriptions, facilitating the generation of images based on textual input.

3.3 LLM Evaluation and Sensitivity Analysis

In our pipeline, we employ OpenAI’s GPT-4o—a state-of-the-art large language model proven to achieve near-perfect accuracy when extracting attributes from clinical reports. During evaluation, the LLM produced no erroneous labels on our test set. To assess the impact of potential extraction errors, we conducted a sensitivity analysis in which we introduced small perturbations to key attributes (e.g., polyp size, morphology, or location) and fed them into our diffusion model. We observed that minor inconsistencies produced only subtle, localized changes in the synthesized images and did not result in any statistically significant drop in clinical validation or classification metrics.

Moreover, we compared two LLM backbones (GPT-4o and DeepSeek) and found virtually identical attribute outputs and end-to-end performance, underscoring the robustness of our approach to the choice of language model.

3.4 Baselines and implementation details

We evaluate our approach against six comparison methods that represent common strategies for one-shot/few-shot generation in diffusion models. For clarity, *each method is presented with a brief description followed by its concrete implementation details.*

Shared setup (applies to all baselines).

Base checkpoint. Unless otherwise stated, *all baselines use the same Stable Diffusion backbone checkpoint as EndoRare*, i.e., the diffusion model weights *domain-adapted via fine-tuning on endoscopic image-text pairs*. Any original off-the-shelf Stable Diffusion checkpoint referenced by a method is replaced by this *in-domain* checkpoint to ensure fairness and isolate methodological differences.

Reproducibility. Across all baselines, we *strictly keep the hyperparameter configurations as released by the original authors or official repositories*. We do not tune learning rates, schedulers, diffusion step counts, model architectures (VAE, text encoder, or U-Net), classifier-free guidance scales, or regularization settings. We fix a global random seed and enable deterministic execution in PyTorch (seed = 0; deterministic cuDNN; benchmarking disabled), applying the same seed across sampling, data loading, and on-the-fly augmentations.

Unified inference. Unless a method explicitly dictates otherwise, we use a DDIM sampler with 200 steps, classifier-free guidance (CFG) = 7.5, $\eta = 1.0$, unconditional branch from the empty string, and generate 100 images per prototype, decoded through the first-stage autoencoder without post-hoc filtering.

Zero-shot (text-to-image).

Description. We use a text-to-image diffusion model (the same in-domain endoscopy-adapted checkpoint as *EndoRare*) prompted only with a GPT-generated description of the target rare disease—no personalization or additional training.

Implementation. Rare-disease endoscopic images are described via a structured instruction to a large language model, producing standardized textual descriptors (e.g., “a TYPE polyp, COLOR color, pathology is PATHOLOGY, located at LOCATION”). These prompts condition the diffusion model; no parameter updates, prompt engineering, rejection sampling, or sample curation are applied beyond the unified inference setup.

Fine-tuned Stable Diffusion (“Finetuned SD”).

Description. The diffusion backbone (U-Net) is fine-tuned for each target prototype image to better memorize instance-specific appearance, though this risks overfitting and typically yields limited diversity from a single example.

Implementation. We fine-tune only the diffusion U-Net while freezing both the VAE and the text encoder, using the in-domain endoscopy-adapted checkpoint for initialization. The objective is standard MSE on noise prediction with timesteps sampled uniformly. We train with Adam at a learning rate of 1.0×10^{-6} for 300 iterations on the single prototype image using the base prompt embedding (no additional token

learning). All other training details—batching, augmentation, mixed-precision policy, gradient clipping, and parameter initialization—follow the official Stable Diffusion training recipe. Sampling follows the unified inference.

Textual Inversion.

Description. A new text token is optimized while the diffusion model is kept frozen; this parameter-efficient approach compresses the target concept into an embedding.

Implementation. We follow the official textual-inversion recipe and train *only* the learned placeholder token (no U-Net or VAE updates) on top of the in-domain endoscopy-adapted checkpoint. The embedding manager is initialized with a single placeholder string ("*") and an initializer word, and is optimized using Adam with a base learning rate of 5×10^{-3} . The diffusion backbone remains fixed with a linear noise schedule ($\beta_{\min} = 8.5 \times 10^{-4}$, $\beta_{\max} = 1.2 \times 10^{-2}$), 1,000 timesteps, and the standard latent autoencoder scaling (0.18215). The U-Net follows the default Stable Diffusion configuration (model channels = 320; attention resolutions {4, 2, 1}; channel multipliers [1, 2, 4]; two residual blocks per level). Training is conducted for up to 6,100 steps using the official **PersonalizedBase** loader (image size 256×256 ; train/val repeats as provided). Inference follows the unified protocol.

DreamBooth.

Description. The diffusion model is fine-tuned with a unique identifier tied to the target concept and a class prior for regularization; it improves identity preservation but requires non-trivial training.

Implementation. Initialized from the in-domain endoscopy-adapted checkpoint, we preserve the official DreamBooth protocol with prior-preserving regularization and joint model adaptation. The diffusion U-Net is updated with Adam at a base learning rate of 1×10^{-6} ; the first-stage autoencoder remains frozen and the CLIP text encoder stays frozen by default. The noise schedule, timestep count (1,000), attention layout, and autoencoder settings follow the official recipe. Training proceeds for the default maximum step budget in the single-image setting (with the prescribed regularization set), adhering to the released batch size, augmentation, and logging cadence. Sampling follows the unified inference.

DATUM.

Description. A one-shot diffusion scheme that encodes the appearance of the target and relies on prompt engineering to encourage diversity; it avoids heavy backbone updates and promotes variation, yet attribute-level controllability is limited by prompt quality.

Implementation. We mirror the official two-stage adaptation recipe on top of the in-domain checkpoint: first optimize a text embedding on a single prototype image, then perform brief U-Net fine-tuning while keeping the VAE and text encoder frozen. The adaptation stage updates only the diffusion U-Net using Adam with a learning rate of 1.0×10^{-6} for 300 iterations, under an MSE noise-prediction objective with

uniformly sampled timesteps. The learned instance token ($\mathbf{V}^*$) is used in the conditional prompt and can be composed with a class cue (e.g., “polyp”) to promote generalization. Inference follows the unified protocol.

SeedSelect.

Description. A training-free strategy that selects or adjusts initial noise seeds to enforce semantic and appearance consistency with the prototype; it is efficient and simple to deploy, but controllability is restricted and diversity hinges on seed search.

Implementation. Using the same in-domain Stable Diffusion checkpoint, we reproduce the official seed-selection baseline. Semantic scoring follows the released configuration with a CLIP backbone (ViT-H/14). Optimization settings: outer-loop iterations = 30 with learning rate 0.1 and step-wise decay; semantic loss weight = 10.0 with a three-term weighting [0.05, 0.25, 0.7]; appearance loss as MSE; and the auxiliary NAO-style selection (candidate points = 50, inner optimization iterations = 5, learning rate = 0.05). Selection-time generations use the default low-step DDIM setting for efficient scoring; final samples for evaluation follow the unified inference above.

3.5 Evaluation Criteria

In this section, we present the evaluation criteria used to assess the performance of the models. These criteria are designed to comprehensively measure various aspects of model accuracy, similarity, and robustness.

3.5.1 Fréchet Inception Distance (FID)

The Fréchet Inception Distance (FID) is a metric used to evaluate the quality of generated images by comparing the distribution of real and generated images in the feature space of a pre-trained Inception model. FID is computed as follows:

$$\text{FID} = \|\mu_r - \mu_g\|_2^2 + \text{Tr} \left(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{1/2} \right) \quad (1)$$

where μ_r and μ_g represent the mean vectors, and Σ_r and Σ_g represent the covariance matrices of the real and generated image distributions, respectively.

3.5.2 CLIP Image Similarity

CLIP (Contrastive Language-Image Pretraining) is a model that learns a joint representation of images and text. To measure image similarity, we use the cosine similarity between the image embeddings and the reference image embeddings from the CLIP model. The similarity score is computed as:

$$\text{similarity} = \frac{\mathbf{I}_r^\top \mathbf{I}_g}{\|\mathbf{I}_r\|_2 \|\mathbf{I}_g\|_2} \quad (2)$$

where $\mathbf{I}_r$ and $\mathbf{I}_g$ are the feature vectors of the real and generated images, respectively.

3.5.3 Learned Perceptual Image Patch Similarity (LPIPS)

LPIPS measures the perceptual similarity between two images by evaluating differences in their deep feature representations. This metric is based on a pre-trained neural network that captures perceptual image quality differences:

$$\text{LPIPS}(x, y) = \frac{1}{N} \sum_{i=1}^N \|f_i(x) - f_i(y)\|_2 \quad (3)$$

where f_i represents the features extracted from the i -th layer of the pre-trained network, and N is the number of layers used in the calculation.

3.5.4 Precision-Recall Area Under the Curve (PR-AUC)

Precision-Recall Area Under the Curve (PR-AUC) evaluates the trade-off between precision and recall at various thresholds. It is calculated as the area under the precision-recall curve:

$$\text{PR-AUC} = \int_0^1 \text{Precision}(r) \, dr \quad (4)$$

where $\text{Precision}(r)$ is the precision at recall r .

3.5.5 Receiver Operating Characteristic Area Under the Curve (ROC-AUC)

Receiver Operating Characteristic Area Under the Curve (ROC-AUC) is a metric used to evaluate the model's ability to distinguish between positive and negative samples. It is computed as the area under the ROC curve, which plots the true positive rate (TPR) against the false positive rate (FPR):

$$\text{ROC-AUC} = \int_0^1 \text{TPR}(f) \, df \quad (5)$$

where $\text{TPR}(f)$ is the true positive rate at a given false positive rate f .

3.5.6 True Positive Rate at 20% False Positive Rate (TPR@20%FPR)

The True Positive Rate at 20% False Positive Rate (TPR@20%FPR) evaluates the model's sensitivity to detect true positives while maintaining a fixed false positive rate of 20%. It is computed as:

$$\text{TPR@20\%FPR} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \quad (6)$$

where True Positives and False Negatives are counted at the point where the false positive rate is 20%.

3.5.7 Precision

Precision measures the proportion of true positive results in all positive predictions. It is calculated as:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \quad (7)$$

3.5.8 Recall

Recall, also known as Sensitivity or True Positive Rate, measures the proportion of true positives correctly identified by the model. It is calculated as:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \quad (8)$$

3.5.9 F1-Score

The F1-Score is the harmonic mean of precision and recall. It is particularly useful when there is an imbalance between classes. It is computed as:

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (9)$$

3.6 Prevalence-aware operating-point analysis

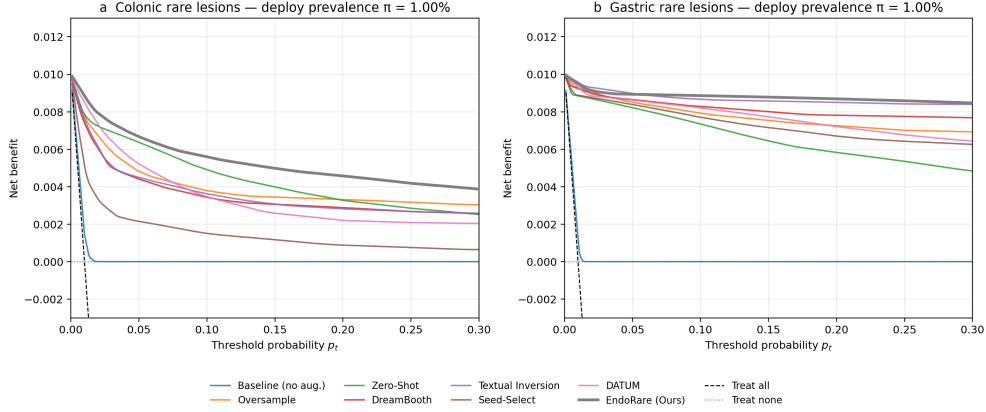
Aggregate discrimination metrics such as PR-AUC and ROC-AUC do not, by themselves, quantify the workflow consequences of rare-lesion detection under low-prevalence conditions. This issue is particularly important for rare gastrointestinal lesions, where even a small false-positive rate can produce a non-negligible number of additional flagged cases. To contextualize the sensitivity–false-positive trade-off, we performed supplementary exploratory analyses, including fixed-FPR operating-point summaries, prevalence-aware estimates of true-positive and false-positive yields, positive predictive value, and exploratory net-benefit curves. These analyses were intended for workflow interpretation and method comparison.

3.6.1 Exploratory net-benefit analysis

We used net-benefit curves as an exploratory way to visualize the trade-off between sensitivity and false-positive burden under assumed deployment prevalences. For a deployment prevalence π , threshold probability p_t , and operating threshold τ , net benefit was computed as:

$$\text{NB}(p_t) = \text{Sens}(\tau) \pi - \text{FPR}(\tau)(1 - \pi) \frac{p_t}{1 - p_t}. \quad (10)$$

Here, p_t represents the clinical threshold probability at which the harm of one missed rare lesion is considered comparable to the harm of unnecessary additional review. Because probability calibration was not the focus of this study and the classifiers were trained as comparative augmentation models, these curves should be interpreted as



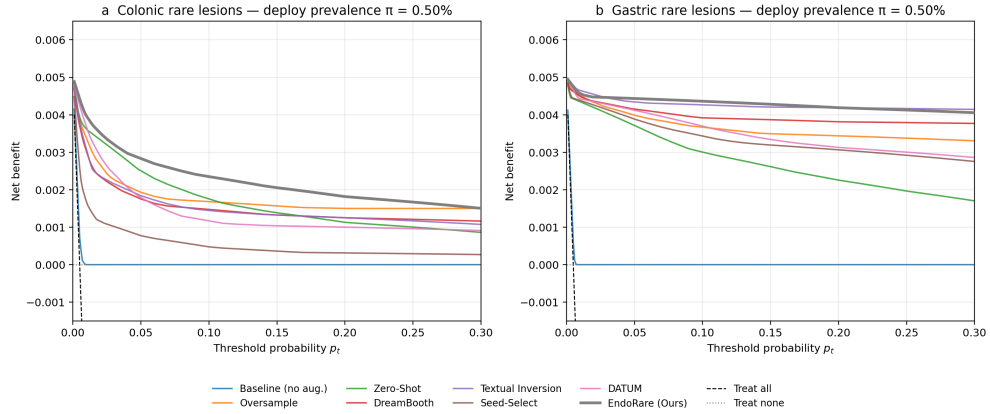
Supplementary Fig. 17: Exploratory net-benefit curves at 1% assumed deployment prevalence. Net benefit is shown for colonic and gastric rare-lesion classifiers across threshold probabilities. The x-axis is the threshold probability p_t and the y-axis is the net benefit $\text{NB}(p_t)$ defined in the equation above.

exploratory summaries of sensitivity–FPR trade-offs rather than calibrated clinical decision curves. We therefore use the net-benefit analysis only as a supplementary visualization and rely primarily on fixed-FPR operating-point summaries to make the false-positive burden transparent. At assumed prevalences of $\pi = 1\%$ and $\pi = 0.5\%$, EndoRare showed favourable exploratory net-benefit trends relative to competing augmentation strategies in the tested low-prevalence settings (Supplementary Figs. 17 and 18).

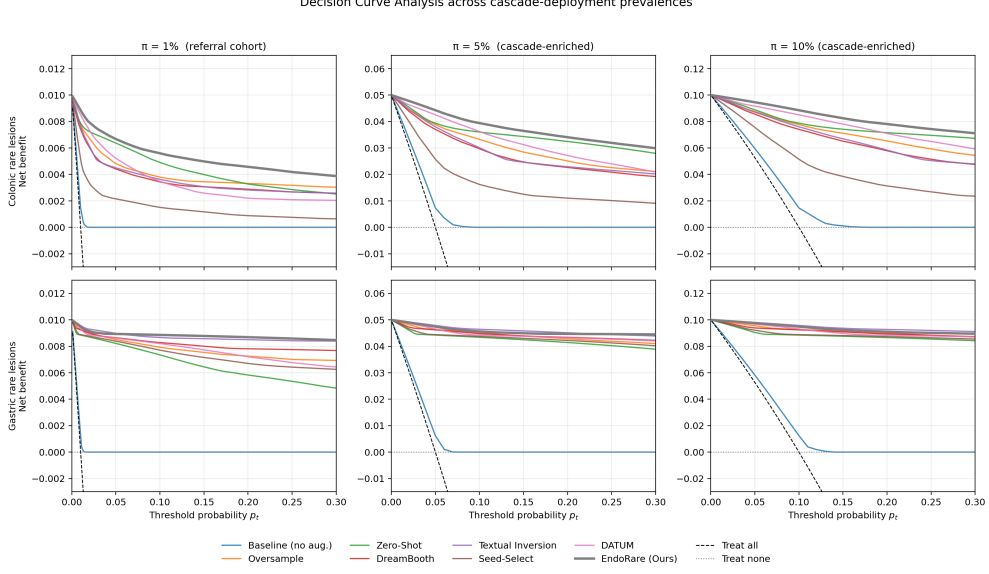
3.6.2 Fixed-FPR operating points and false-positive burden

To make threshold-dependent performance more interpretable for workflow planning, we summarized operating points at fixed false-positive-rate budgets of 1%, 5%, 10%, and 20%. These fixed-FPR budgets were used for standardized comparison across augmentation strategies. At the same FPR budget, methods incur comparable false-positive workloads; therefore, differences in TPR indicate how much sensitivity is retained under the same false-positive constraint.

At the strict 1% FPR budget and an assumed prevalence of $\pi = 1\%$, EndoRare recovered 50% of colonic rare lesions and 90% of gastric rare lesions, corresponding to approximately 9–10 false-positive alerts per 1,000 examinations and approximately 5–9 true-positive detections per 1,000 examinations (Supplementary Table 17). These summaries provide an interpretable estimate of workflow burden, but prospective validation and site-specific threshold calibration would be required before any clinical use.



Supplementary Fig. 18: Exploratory net-benefit curves at 0.5% assumed deployment prevalence. This lower-prevalence setting was used as a sensitivity-analysis scenario. The x-axis is the threshold probability p_t and the y-axis is the net benefit $NB(p_t)$ defined in the equation above.

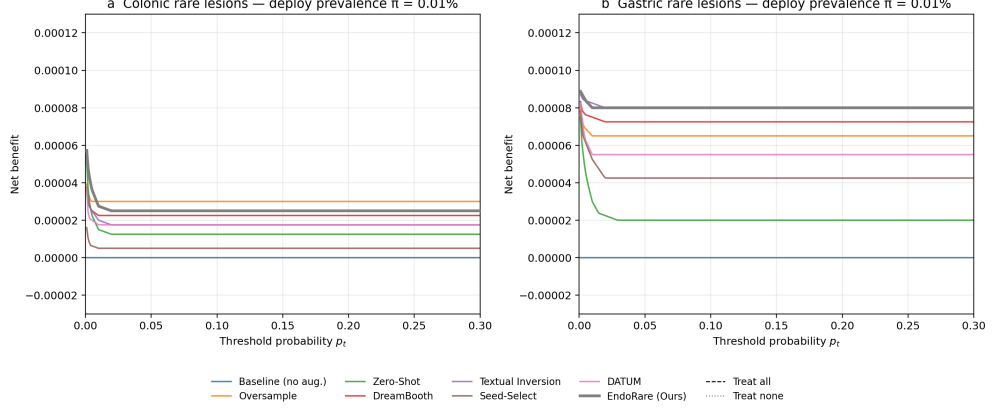


Supplementary Fig. 19: Exploratory net-benefit curves under hypothetical enriched-cohort prevalences. Net benefit is shown at $\pi = 1\%$, 5% , and 10% for colonic and gastric rare lesions. The x-axis is the threshold probability p_t and the y-axis is the net benefit $NB(p_t)$ defined in the equation above.

3.6.3 Illustrative enriched-cohort analysis

The entities studied here have very low population prevalence. We therefore do not position EndoRare-augmented classifiers as autonomous first readers for unselected populations. A more plausible future use case would be a second-reader or case-prioritization aid applied after routine clinical assessment or after an upstream detection system has already identified potentially abnormal examinations. This scenario is discussed as a plausible deployment hypothesis and was not empirically validated in the present study.

To illustrate how conditional prevalence may affect workflow burden, we repeated the operating-point summaries under hypothetical enriched-cohort prevalences of $\pi = 1\%$, 5% , and 10% (Supplementary Fig. 19). These values were used as sensitivity-analysis settings rather than as estimates of real-world deployment prevalence. Under higher assumed conditional prevalence, the number of flagged cases requiring review per confirmed rare lesion decreases, as expected from the increase in positive predictive value (Supplementary Table 18). These analyses should be interpreted only as exploratory workflow illustrations. Prospective studies with real upstream triage systems and site-specific prevalence estimates are required to evaluate any cascade workflow.



Supplementary Fig. 20: Very-low-prevalence stress test at 1/10,000 assumed prevalence. Net benefit is shown for colonic and gastric rare-lesion classifiers at an assumed prevalence of 1/10,000. The x-axis is the threshold probability p_t and the y-axis is the net benefit $NB(p_t)$ defined in the equation above.

3.6.4 Very-low-prevalence stress test

For completeness, we also performed an extreme low-prevalence stress test at $\pi = 1/10,000$. As expected, the absolute net benefit of all single-stage classifiers decreased substantially under this prevalence assumption (Supplementary Fig. 20). This analysis reinforces the limited scope of our deployment interpretation: EndoRare-augmented classifiers should not be interpreted as autonomous first-reader tools for unselected population-level screening at extreme rare-disease prevalence. Instead, any future clinical use would require site-specific prevalence estimation, probability calibration, threshold selection, and prospective workflow validation.

Supplementary Table 17: Fixed-FPR operating-point summary at 1% assumed prevalence. Values are macro-averages across rare entities within each organ. False-positive and true-positive yields are reported per 1,000 examinations for EndoRare at the indicated FPR budget.

Organ	FPR budget	EndoRare TPR	Best comparator TPR	EndoRare FP/1,000	EndoRare TP/1,000
Colon	1%	0.50	0.45 (Zero-Shot)	9.4	5.0
Colon	5%	0.75	0.75 (Zero-Shot)	48.8	7.5
Colon	10%	0.88	0.80 (DATUM)	99.0	8.8
Colon	20%	0.95	0.95 (DATUM)	198.0	9.5
Stomach	1%	0.90	0.88 (Textual Inversion)	9.9	9.0
Stomach	5%	0.90	0.95 (Textual Inversion)	48.0	9.0
Stomach	10%	0.95	0.95 (Textual Inversion)	99.0	9.5
Stomach	20%	1.00	1.00 (Textual Inversion)	198.0	10.0

Supplementary Table 18: Illustrative number-needed-to-evaluate under hypothetical enriched-cohort prevalences. NNE is defined as $1/PPV$ and gives the average number of flagged cases requiring review to identify one true rare lesion under the specified assumed prevalence. NE indicates that NNE was not estimable because no true positives were recovered at the selected operating point.

Deployment prevalence	EndoRare NNE, colon	EndoRare NNE, stomach	Baseline NNE, colon	Baseline NNE, stomach
1%	7.5	6.3	NE	NE
5%	2.2	2.0	NE	NE
10%	1.6	1.5	NE	NE

3.7 Downstream classifier training protocol

For downstream rare-lesion classification, we used an ImageNet-pretrained ResNet-50 in a controlled linear-probe setting. All backbone parameters were frozen, and only the final fully connected layer was replaced and trained for binary classification. Images were resized to 224×224 and normalized using standard ImageNet statistics. Training used cross-entropy loss and stochastic gradient descent with momentum. The fixed training settings were: learning rate = 0.01, momentum = 0.9, batch size = 64, weight decay = 1×10^{-4} , scheduler step size = 30 epochs, and learning-rate decay factor $\gamma = 0.1$. These settings were specified a priori, and no method-specific hyperparameter optimization was performed. The same classifier architecture, preprocessing pipeline, optimizer, scheduler, and hyperparameters were used across all augmentation strategies and baselines to isolate the effect of the generated training images. Every configuration was evaluated by five-fold cross-validation. The folds were generated once with a single fixed random seed per organ (seed 555 for the colonic entities, juvenile polyposis syndrome and calcifying fibrous tumour, and seed 42 for the gastric entities, familial adenomatous polyposis and Peutz–Jeghers syndrome) and were then reused unchanged for every augmentation strategy, so that all methods are compared on identical splits. The five fold-level values obtained in this way are the replicates that underlie the means and error bars shown in Fig. 3 and Fig. 4b of the main text and in Supplementary Table 17; they are technical replicates of the training and evaluation procedure on a fixed image set and are not independent patient cohorts.