CRAFT: CLINICAL REWARD-ALIGNED FINETUNING FOR MEDICAL IMAGE SYNTHESIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Generating an image for a requested medical condition requires more than plausible appearance: the condition must be visible in the image. Medical-domain adaptation, richer descriptions, and reward or preference alignment improve different aspects of generation, but agreement with text or visual plausibility alone does not necessarily make the requested condition recognizable in a new sample. We ask whether direct feedback on generated images improves expression of the requested condition beyond equal additional text-guided reconstruction. CRAFT (Clinical Reward-Aligned Finetuning for Medical Image Synthesis) tests this question by retaining reconstruction and scoring sampled images against an instance description, class-shared visual criteria, and the target label using a classifier trained on real images. In a matched CheXpert continuation from the same TI+LoRA initialization, CRAFT exceeds reconstruction-only training in DINOv2 balanced accuracy by 10.70 percentage points after five additional epochs [paired source-bootstrap 95% CI 7.28, 14.12]. Across three CheXpert training runs, removing diagnostic feedback lowers balanced accuracy for label recognition, while MetaCLIP2 description and checklist similarities barely change. Across four domains, CRAFT has the highest synthetic-method point estimates for DINOv2 balanced accuracy and both MetaCLIP2 similarities. Two physicians selected CRAFT in 55.3% versus TI+LoRA in 26.3% of 800 method-anonymous four-way judgments. Feedback on sampled images thus makes requested conditions more recognizable to a held-out classifier trained on real images while improving alignment with instance descriptions and shared clinical criteria.

1 INTRODUCTION

Controllable medical-image generation asks a model to depict a specified condition, not merely the right modality or a plausible anatomy. A chest radiograph can look realistic while omitting the requested abnormality; image translation can likewise add or remove clinically relevant features without losing its plausible appearance (Cohen et al., 2018). In either case, the prompt and assigned label state what the image was meant to show, not what it visibly contains. Condition-specific synthetic images therefore require visual evidence for their assigned diagnosis (Sun et al., 2023; Wang et al., 2025).

Textual inversion supplies learned condition tokens (Gal et al., 2022; De Wilde et al., 2023). LoRA provides low-rank adaptation (Hu et al., 2022) and has been applied to medical-image diffusion (Wang et al., 2025). Image-grounded descriptions supply more detailed conditioning than a disease name alone (Huang et al., 2025). Reward and preference methods can steer generated outputs using assessments of images or favored alternatives (Black et al., 2023; Wallace et al., 2024; Wang et al., 2025; Clark et al., 2024). These approaches address domain fit, conditioning, and output alignment from different directions.

Reconstruction fits observed images, while text alignment measures agreement with supplied wording and visual plausibility concerns coherent appearance. A new sample can meet these aims yet lack visual cues that distinguish the requested condition from alternatives. Can feedback on generated samples improve those diagnostic cues beyond continued text-guided reconstruction alone?

CRAFT (Clinical Reward-Aligned Finetuning for Medical Image Synthesis) is our proposed answer. Starting from a validation-selected textual-inversion and LoRA (TI+LoRA) model, it retains recon-

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

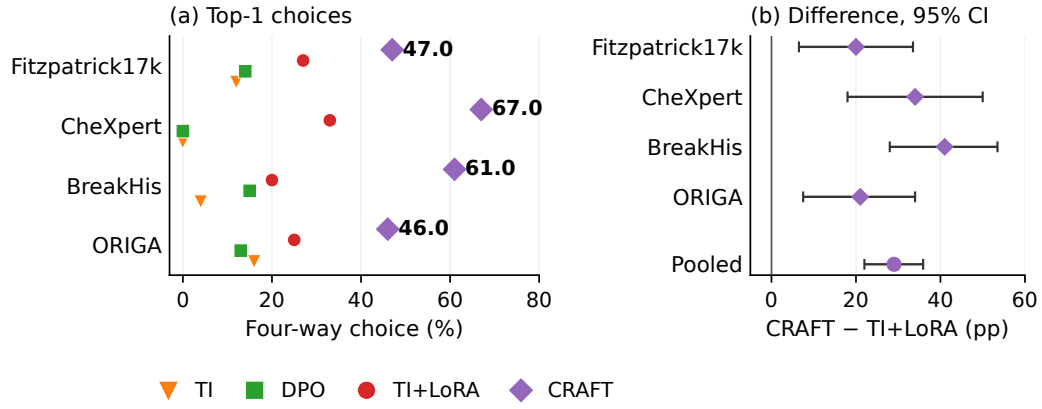

Figure 1: **Physician preference across four domains.** (a) Method-anonymous four-way top-1 shares from two physicians (100 cases per domain; 800 judgments). (b) CRAFT minus TI+LoRA shares by domain and pooled, in points, with case-bootstrap 95% intervals. Both judgments per case are resampled together; pooled resampling is stratified by modality. The pooled difference is $+29.0$ points [22.0, 35.9]. TI denotes textual inversion; DPO denotes direct preference optimization.

struction while adding feedback on images sampled from noise. An instance-specific description specifies the depicted appearance; a class-shared checklist describes visual criteria expected for the condition; and a classifier fitted on labeled real-image features scores support for the target label. These become Visual Description Consistency (VDC), Clinical Criteria Satisfaction (CCS), and Diagnostic Discriminability (DD), optimized with differentiable image feedback (Clark et al., 2024).

On CheXpert, CRAFT and reconstruction-only training share a TI+LoRA checkpoint, prompts, generation seeds, reconstruction optimizer and learning rate, batch size, and update count. After five additional epochs, CRAFT's balanced accuracy for requested-label recognition, measured by a held-out DINOv2 classifier trained on real images, is 10.70 percentage points higher [paired source-bootstrap 95% CI 7.28, 14.12]. We call percentage-point differences points below. The DINOv2 classifier and MetaCLIP2 description and checklist evaluator are unused in training or model selection, unlike the MedSigLIP training critic. Across four domains, CRAFT leads synthetic comparators in these automatic evaluations, and physicians favor it in blinded four-way choices (Figure 1).

Our contributions are threefold. **First**, we combine instance-level text, class-shared criteria, and real-trained diagnostic feedback with reconstruction in a sampled-image adaptation objective. **Second**, matched CheXpert continuation shows improved target-label recognition beyond equal additional reconstruction. **Third**, removal studies identify distinct effects of diagnostic and text feedback, while four-domain evaluations show stronger recognition and text alignment alongside physician preference for CRAFT.

2 Related Work

Medical-image adaptation. Medical diffusion models include domain-specific generators and parameter-efficient adaptations of general-purpose models (Chambon et al., 2022; De Wilde et al., 2023). VAP-Diffusion uses multimodal language models to enrich medical descriptions (Huang et al., 2025). CRAFT uses fixed image-grounded descriptions for conditioning and adds feedback on the resulting images.

Reward-based generation. Diffusion alignment has used reinforcement learning, preference optimization, and direct differentiation through denoising (Black et al., 2023; Wallace et al., 2024; Clark et al., 2024). CXRL uses policy-based RL with posture, diagnostic-classifier, and report-consistency rewards (Han et al., 2024); RL4Med-DDPO uses RL with vision-language guidance (Saremi et al., 2025); and Doctor Approved uses checklist-ranked preference pairs with DPO (Wang et al., 2025). CRAFT instead combines an instance-level description, a class-shared checklist, and a real-image

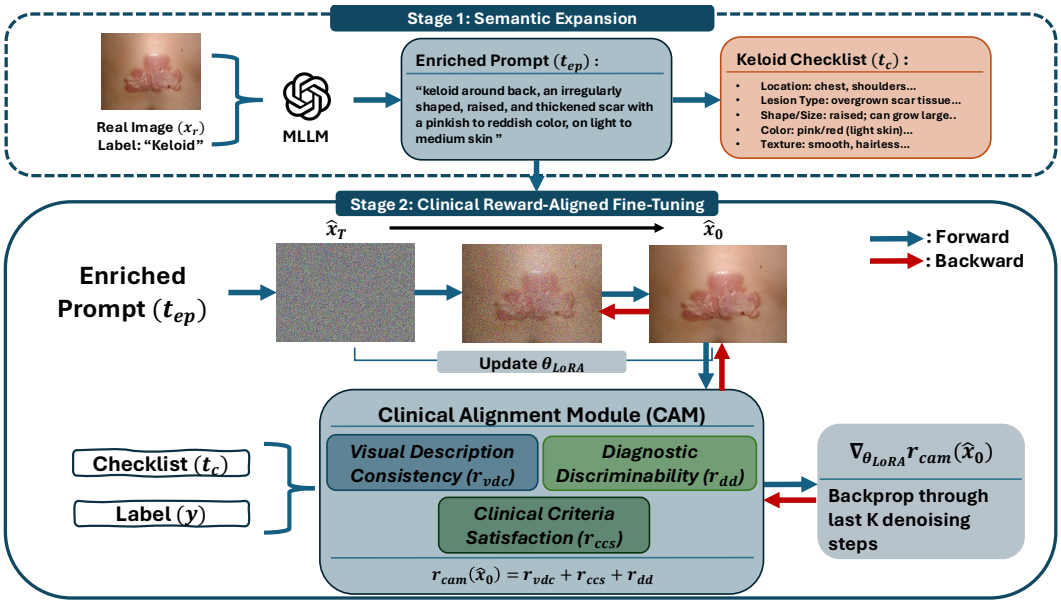

Figure 2: **CRAFT adaptation pipeline.** Stage 1 uses each image–label pair to produce an instance description and the label to produce a class-shared checklist. Stage 2 samples from noise and scores the image with a frozen Clinical Alignment Module (CAM) for description, checklist, and diagnostic feedback. Gradients pass through the final K denoising steps and update only LoRA parameters. A separate reconstruction path uses real images conditioned on the token-substituted enriched description.

diagnostic probe as differentiable feedback, optimized by direct backpropagation through the final sampling steps while retaining reconstruction. We evaluate it across four imaging domains. The CheXpert reconstruction+DD control tests the added text targets at CRAFT's retained coefficients.

Evaluation. We report target-label recognition, text alignment, feature-distribution discrepancy, perceptual variation, and human preference as separate endpoints. DINOv2, MetaCLIP2, KID, and LPIPS are complementary measures (Oquab et al., 2023; Chuang et al., 2025; Bińkowski et al., 2018; Zhang et al., 2018).

3 CRAFT: FEEDBACK ON SAMPLED MEDICAL IMAGES

Let $\mathcal{D} = \{(x_i, y_i)\}$ be the labeled real training set. We initialize CRAFT from a validation-selected TI+LoRA model. Alongside reconstruction training, CRAFT adds feedback on images sampled from independent noise. Its three targets address instance-specific appearance, visual criteria shared across a class, and features that support target-class prediction. Figure 2 summarizes the pipeline.

3.1 INSTANCE DESCRIPTIONS AND SHARED CLASS CUES

For each training image x_i with label y_i, GPT-4o produces a fixed enriched description $t_{\text{ep}}^{(i)}$ of its visible attributes. A label-conditioned checklist $t_c^{(y)}$ provides a second text target shared by all examples of class y. In dermatology, descriptions vary with a lesion's color, border, and texture, while the checklist specifies visual cues expected across examples of the same condition. Both targets are generated before reward adaptation and remain fixed during training.

The two text targets serve complementary purposes: the instance description asks whether a generated image shows the specified appearance, while the checklist asks whether it is consistent with class-level visual cues. CCS scores the checklist as a whole rather than adjudicating individual items. Diagnostic feedback, defined below, adds a signal learned from labeled real images rather than text alone.

Textual inversion introduces learned disease tokens. We call $t_{\mathrm{ep}}^{(i)}$ the human-readable enriched description and $\widetilde{t_{\mathrm{ep}}}^{(i)}$ its token-substituted version, with the disease name replaced by a learned token. Reconstruction and adapted-generator evaluation use the token-substituted version. The reward path uses the human-readable version, which is also the target for description feedback. Appendices B–C provide the prompt templates and class checklists; Appendix A gives the optimization settings.

3.2 THREE FEEDBACK TERMS

Let E_I and E_T denote frozen MedSigLIP image and text encoders (Sellergren et al., 2025), and let $\widehat{x}$ denote a sampled image. Visual Description Consistency (VDC) and Clinical Criteria Satisfaction (CCS) are

$$r_{\mathrm{vdc}}(\widehat{x}, t_{\mathrm{ep}}) = \cos\big(E_I(\widehat{x}), E_T(t_{\mathrm{ep}})\big), \tag{1}$$

$$r_{\mathrm{ccs}}(\widehat{x}, t_c) = \cos\big(E_I(\widehat{x}), E_T(t_c)\big). \tag{2}$$

Both rewards use the same encoder, but VDC has an instance-specific text target and CCS has a class-shared target.

For Diagnostic Discriminability (DD), a linear classifier ϕ is fitted on frozen MedSigLIP features extracted from real training images. Its target-class log probability is

$$r_{\mathrm{dd}}(\widehat{x}, y) = \log \frac{\exp\big(\phi(E_I(\widehat{x}))_y\big)}{\sum_{k=1}^{C} \exp\big(\phi(E_I(\widehat{x}))_k\big)}. \tag{3}$$

The classifier remains frozen during generator adaptation. Unlike the text-similarity rewards, DD penalizes a sample when the decision rule learned from real images favors another class. We define $\ell_{\mathrm{vdc}} = 1 - r_{\mathrm{vdc}}$, $\ell_{\mathrm{ccs}} = 1 - r_{\mathrm{ccs}}$, and $\ell_{\mathrm{dd}} = -r_{\mathrm{dd}}$.

3.3 JOINT ADAPTATION WITH RECONSTRUCTION

CRAFT uses separate reconstruction and reward paths. The reconstruction path encodes a real image as $z_i = \mathcal{E}_{\mathrm{VAE}}(x_i)$, adds diffusion noise, and predicts the corresponding target under $\widetilde{t_{\mathrm{ep}}}^{(i)}$. For noise prediction,

$$\mathcal{L}_{\mathrm{diff}} = \mathbb{E}_{i,t,\epsilon}\left[\left\|\epsilon - \epsilon_\theta\big(\alpha_t z_i + \sigma_t \epsilon, t, \widetilde{t_{\mathrm{ep}}}^{(i)}\big)\right\|_2^2\right]. \tag{4}$$

We use the corresponding velocity target when required by the generator parameterization.

Independently, the reward path draws $z_T \sim \mathcal{N}(0, I)$ and generates

$$\widehat{x}_i = \mathcal{D}_{\mathrm{VAE}}\Big(\mathcal{S}_\theta^{(K)}(z_T, t_{\mathrm{ep}}^{(i)})\Big), \tag{5}$$

where $\mathcal{S}_\theta^{(K)}$ follows the complete sampling trajectory while tracking gradients only through the final K denoising steps. Real-image pixels therefore affect the reconstruction loss but never enter the generated-image reward trajectory.

Averaging the three feedback losses over training examples and sampled trajectories gives $\mathcal{L}_{\mathrm{vdc}}$, $\mathcal{L}_{\mathrm{ccs}}$, and $\mathcal{L}_{\mathrm{dd}}$. The training objective is

$$\mathcal{L}(\theta) = 0.2\mathcal{L}_{\mathrm{diff}} + 0.2\mathcal{L}_{\mathrm{vdc}} + 0.2\mathcal{L}_{\mathrm{ccs}} + 0.2\mathcal{L}_{\mathrm{dd}}. \tag{6}$$

The four coefficients are fixed at 0.2 each. Only attention-layer LoRA parameters are updated. The base generator, VAE, text encoders, and diagnostic classifier remain frozen. Following DRaFT-K, image-level feedback is backpropagated through the final K denoising steps (Clark et al., 2024). CRAFT requires neither a physician-preference model nor synthetic winner–loser pairs.

Model selection. Generator checkpoints for all methods, including DPO, are selected on validation data. For TI+LoRA and CRAFT, selection uses the equal-weight mean of validation SigLIP description similarity, checklist similarity, and hard target-class accuracy. Selection uses a distinct SigLIP model in the MedSigLIP training critic's family, which may favor CRAFT-style optimization in checkpoint ranking. DINOv2, MetaCLIP2, and reader judgments are not used to select generator checkpoints. Appendix L reports the TI+LoRA selection trajectory and selected CRAFT epochs. Matched continuation evaluates fixed additional horizons without selecting on external endpoints.

4 EXPERIMENTS

We first compare CRAFT with equal additional reconstruction on CheXpert, then examine the feedback terms. Four-domain recognition and alignment results, physician preferences, and FLUX transfer and augmentation complete the evaluation.

4.1 EVALUATION SETUP

Data and methods. We evaluate a 20-condition Fitzpatrick17k subset, four frontal-view CheXpert classes, eight BreakHis classes at $100\times$ magnification, and binary glaucoma classification on ORIGA (Groh et al., 2021; Irvin et al., 2019; Spanhol et al., 2016; Zhang et al., 2010). The held-out sets contain 3,100, 495, 1,041, and 196 source images, respectively. Appendix A reports the training and validation partitions. Primary comparisons use the Stable Diffusion v2.1-base generator (`Manojb/stable-diffusion-2-1-base`) at 512×512. We compare TI, TI+LoRA, CRAFT, and a DPO baseline trained on 1,024 MedGemma-ranked preference pairs following Wang et al. (2025). We evaluate all methods, including adapted DPO, by text-to-image generation from noise on common held-out source panels and sampling seeds, unlike the cited method's source-conditioned image-to-image setting.

Held-out diagnostic evaluation. Using frozen DINOv2 CLS features (Oquab et al., 2023), we fit a StandardScaler and multinomial logistic regression on real training images, selecting regularization C on a real-only validation split, then freeze the classifier. We compare its predictions on synthetic images with requested labels, reporting balanced accuracy as primary and macro-F1 alongside it. Synthetic methods share five generation seeds; held-out real images provide a reference.

Held-out semantic evaluation. A frozen MetaCLIP2 encoder measures cosine similarity between each generated image and (i) its human-readable instance description and (ii) its class checklist (Chuang et al., 2025). We report the two similarities separately. All adapted-method semantic estimates use five generation seeds.

Generation prompts. Primary SD2.1 evaluation uses source-matched, token-substituted enriched descriptions and common sampling seeds; CRAFT's training reward uses human-readable descriptions. MetaCLIP2 scores against the human-readable descriptions and class checklists (Appendix A).

Distributional similarity and sample variation. We compute KID in DINOv2 feature space against held-out real images, matching sample counts and class composition (Bińkowski et al., 2018). The KID table uses generation seed 42. Within-prompt LPIPS uses 100 held-out prompts and four common sampling seeds per dataset; we average all six generated-image pairs within each prompt before aggregating (Zhang et al., 2018).

Uncertainty. Diagnostic contrast intervals use paired source-level 95% bootstrap resampling, keeping all generations from each source together. MetaCLIP2 standard deviations summarize generation-seed means; KID and LPIPS are point estimates. The intervals condition on evaluated checkpoints and exclude variation across training runs.

4.2 FEEDBACK IMPROVES RECOGNITION BEYOND MORE RECONSTRUCTION

From the same CheXpert TI+LoRA checkpoint, CRAFT and reconstruction-only training share prompts, seeds, reconstruction optimizer, learning rate, batch size, and update count; only CRAFT adds VDC, CCS, and DD. Its DINOv2 balanced-accuracy advantage grows across five additional epochs, from $+2.23$ $[0.62, 3.84]$ to $+10.70$ $[7.28, 14.12]$ points (Figure 3(a,b)). Reconstruction-only accuracy remains between 31.65% and 32.99%, while CRAFT reaches 42.64% at epoch $+5$, its validation-selected reward-stage checkpoint. At that horizon, macro-F1 is 38.52% versus 28.46% ($+10.06$ $[6.52, 13.60]$ points). Both diagnostic intervals exceed zero at every horizon (Appendix D). MetaCLIP2 description and checklist differences also favor CRAFT throughout, ending at $+0.1312$ $[0.0949, 0.1675]$ and $+0.0200$ $[+0.0087, +0.0314]$; DINOv2-KID is lower at every horizon, ending at 0.0641 versus 0.0771 on the $\times 10^3$ scale. At this shared initialization, sampled-image feedback improves recognition and text alignment beyond equal additional reconstruction.

Table 1: CheXpert component removal at the original checkpoint. BA (balanced accuracy) and macro-F1 are percentages from the real-trained DINOv2 classifier; description and checklist are MetaCLIP2 cosine similarities. Shading marks the lowest value in each column, the largest drop from full CRAFT. Each removal is a separate trained arm.

Variant	BA	Macro-F1	Description	Checklist
Full CRAFT	42.64	38.52	.2773	.2479
Without DD	33.63	30.25	.2767	.2478
Without reconstruction	38.96	35.12	.2723	.2422
Without CCS	41.63	37.51	.2717	.2274
Without VDC	42.44	38.34	.1586	.2363

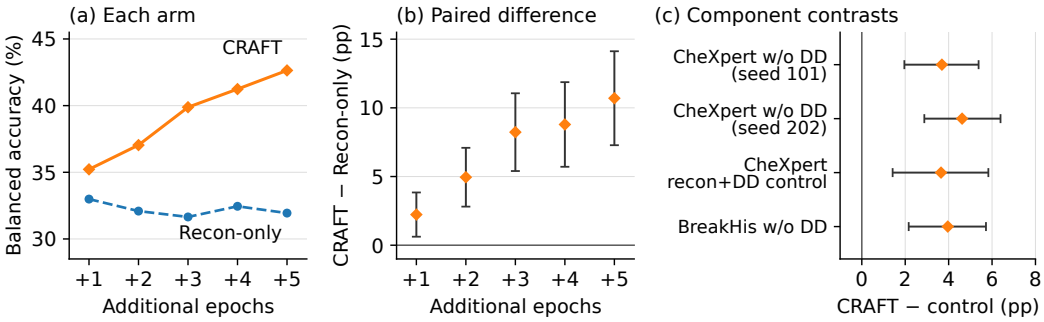

Figure 3: **Sampled-image feedback improves DINOv2 balanced accuracy.** (a) CRAFT and reconstruction-only continuation from a shared CheXpert TI+LoRA initialization across epochs $+1$ to $+5$; the vertical axis begins above zero. (b) CRAFT-minus-reconstruction-only differences in points. (c) Full CRAFT contrasts with w/o DD at two additional CheXpert training seeds and on BreakHis, and with CheXpert reconstruction+DD. Whiskers show paired source-level 95% bootstrap intervals on the differences.

4.3 The feedback terms have distinct measured effects

The shaded cells of Table 1 identify the largest drops when each CheXpert term is removed; each retained coefficient stays at 0.20. At the original checkpoint, removing DD lowers DINOv2 balanced accuracy by 9.01 points and macro-F1 by 8.27 points, while description and checklist similarities change by only 0.0006 and 0.0001. Removing VDC instead lowers description similarity by 0.1187 with a 0.20-point BA change; removing CCS lowers checklist similarity by 0.0205. Reconstruction removal also lowers recognition and both text similarities. The diagnostic term has the largest measured effect on recognition, while the text terms dominate their corresponding alignment scores.

DD removal lowers balanced accuracy at two more CheXpert training seeds (Figure 3(c)); their description differences are $+0.0012$ and $+0.0024$, and checklist differences are $+0.0006$ and $+0.0027$ (Appendix G). On BreakHis, full CRAFT exceeds the variant without DD by $+3.96$ $[2.16, 5.72]$ BA points, while that variant is near TI+LoRA ($+0.20$ $[-1.53, +1.93]$ points); the full-minus-without-DD macro-F1 difference is $+1.45$ $[-0.41, 3.36]$ points. Against the CheXpert reconstruction+DD control, adding VDC and CCS yields $+3.65$ $[1.42, 5.83]$ BA points, $+0.1210$ description similarity, and $+0.0203$ checklist similarity, with paired intervals above zero. DD supports measured recognition across these runs, and the VDC+CCS bundle adds recognition and text alignment in the CheXpert control.

4.4 Four-domain recognition and alignment

The requested labels are more recognizable to a held-out classifier. CRAFT leads all synthetic methods in DINOv2 balanced accuracy on Fitzpatrick17k, CheXpert, BreakHis, and ORIGA (Table 2). Relative to TI+LoRA, its gains range from $+3.29$ to $+9.99$ points, with paired intervals above zero

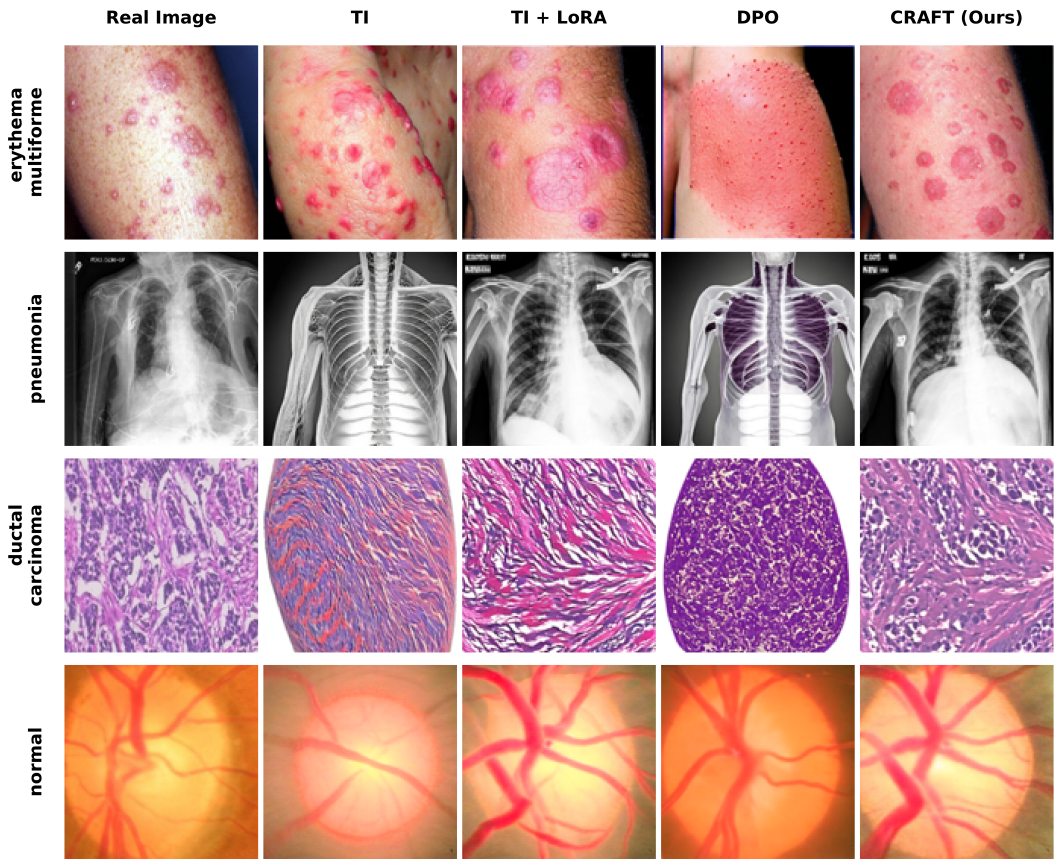

Figure 4: **Illustrative outputs across four domains.** Rows show erythema multiforme (Fitzpatrick17k), pneumonia (CheXpert), ductal carcinoma (BreakHis), and normal (ORIGA). Columns show Real Image, TI, TI+LoRA, DPO, and CRAFT. Larger galleries appear in Appendix K.

in every domain (Appendix E). On CheXpert, TI and DPO score 16.76% and 17.05% BA, below four-class chance (25%). Macro-F1 gains are $+4.40$, $+9.25$, $+1.83$, and $+3.95$ points in the same domain order; the BreakHis interval is $[-0.06, 3.65]$, while the other three intervals exclude zero. Across the four evaluated domains, the held-out classifier more often recognizes CRAFT's requested labels than those of the other synthetic methods.

Both text targets improve. CRAFT also has the highest MetaCLIP2 description and checklist similarity among synthetic methods in every domain (Table 2). Relative to TI+LoRA, description similarity increases by 0.0086–0.1189 and checklist similarity by 0.0117–0.0207. The two text scores can diverge: TI+LoRA improves Fitzpatrick17k checklist similarity over TI while reducing description similarity. Recognition can diverge from text alignment as well: ORIGA DPO exceeds TI+LoRA on checklist similarity while trailing on BA (Appendix F). CRAFT's lead across all four domains holds for each measured text target.

Distribution and variation. CRAFT's DINOv2-KID point estimate is lower than TI+LoRA's in all four domains and lowest among synthetic methods in three (Table 2). On Fitzpatrick17k, TI and DPO have lower KID than CRAFT despite lower recognition and text alignment, illustrating why we report condition-sensitive measures alongside distributional similarity. Within-prompt LPIPS changes little relative to TI+LoRA: the four differences range from -0.0008 to $+0.0114$. In the near-copy screen, no generation exceeds SSIM 0.95, while pHash flags two TI+LoRA and two CRAFT CheXpert generations (Appendix J). Figure 4 places illustrative generations beside real references; Appendix K

Table 2: Four-dataset results. DINOv2 balanced accuracy (BA) and macro-F1 are percentages; MetaCLIP2 description (Desc.) and checklist (Check.) are cosine similarities. DINOv2-KID is multiplied by 10^3 (lower is better); LPIPS measures within-prompt variation. Synthetic diagnostics and semantics use five seeds; KID uses seed 42; LPIPS uses 100 prompts and four seeds. Real images are a separate reference with no KID or LPIPS entry. Bold marks the best synthetic value for BA, F1, Desc., Check., and KID.

Dataset	Method	BA↑	F1↑	Desc.↑	Check.↑	KID↓	LPIPS
Fitzpatrick17k	Real	55.12	55.94	.2705	.2315	–	–
	TI	35.68	36.21	.1843	.2310	.1333	.6217
	DPO	36.91	37.00	.1846	.2331	**.1332**	.6241
	TI+LoRA	42.29	40.89	.1798	.2402	.2031	.5714
	CRAFT	**45.98**	**45.29**	**.2759**	**.2519**	.1937	.5828
CheXpert	Real	47.06	47.06	.2795	.2520	–	–
	TI	16.76	16.64	.1586	.2137	1.2348	.5686
	DPO	17.05	16.93	.1598	.2167	1.2299	.5687
	TI+LoRA	32.65	29.27	.1584	.2272	.0727	.4449
	CRAFT	**42.64**	**38.52**	**.2773**	**.2479**	**.0641**	.4457
BreakHis	Real	67.94	69.42	.2780	.2478	–	–
	TI	31.08	30.78	.2334	.2234	.4929	.5489
	DPO	31.47	31.81	.2492	.2342	.3451	.5584
	TI+LoRA	32.28	31.84	.2473	.2413	.2107	.5767
	CRAFT	**36.44**	**33.67**	**.2815**	**.2542**	**.1140**	.5759
ORIGA	Real	68.70	70.20	.2659	.2450	–	–
	TI	52.73	48.29	.2649	.2322	.6021	.5346
	DPO	50.39	45.23	.2658	.2477	.4050	.5607
	TI+LoRA	56.29	56.06	.2692	.2366	.4753	.5927
	CRAFT	**59.58**	**60.01**	**.2778**	**.2541**	**.3680**	.5993

adds galleries and CheXpert error cases. The gains in recognition and alignment accompany lower KID than TI+LoRA and little change in within-prompt variation.

4.5 Physicians prefer CRAFT

Two physicians each chose among four method-anonymous candidates for 100 cases per domain (400 cases, 800 judgments). CheXpert images were displayed at 512 pixels and the other domains at 128 pixels (Appendix H). CRAFT receives 55.3% of four-way top-1 choices versus 26.3% for TI+LoRA, a $+29.0$ $[22.0, 35.9]$-point difference under a modality-stratified case bootstrap (Figure 1). Its domain differences range from $+20.0$ to $+41.0$ points, with all four intervals above zero; inter-rater κ ranges from 0.29 to 0.56. CRAFT is the preferred candidate in this blinded four-way comparison.

4.6 FLUX alignment transfers; augmentation favors CRAFT

On FLUX.1-schnell, CRAFT improves the MetaCLIP2 text mean over enriched-prompt TI+LoRA in all four domains (Table 3). On BreakHis, the text mean improves by $+0.00316$ $[0.00160, 0.00471]$, driven by a $+0.00835$ description-similarity gain. Full-data augmentation with 20% synthetic examples per classifier batch favors CRAFT by $+0.6$ to $+4.6$ accuracy points across the four domains. Within CRAFT, accuracy increases from 10-shot to 20-shot to full-data training (Appendix M); FLUX within-prompt LPIPS results appear in Appendix I. These results extend the measured alignment pattern to FLUX and favor CRAFT in full-data augmentation.

5 Discussion

Together, the results suggest that better target-label recognition comes from feedback on sampled images, rather than more fitting of real images at the matched CheXpert initialization. Component removals show a division of labor: the real-trained diagnostic probe drives measured recognition, while the text targets drive their corresponding alignment scores and add recognition beyond recon-

Table 3: FLUX transfer and downstream augmentation. FLUX.1-schnell columns show CRAFT minus enriched-prompt TI+LoRA MetaCLIP2 description, checklist, and text-mean similarities; text-mean intervals are paired-source 95% bootstrap intervals. Within-prompt FLUX LPIPS results appear in Appendix I. Full-data augmentation uses 20% synthetic examples per classifier batch; accuracy values and point differences are point estimates (Appendix M).

	FLUX.1-schnell: CRAFT − TI+LoRA			Full-data augmentation accuracy (%)		
Dataset	Desc.	Check.	Text mean [95% CI]	TI+LoRA	CRAFT	Δ (points)
Fitzpatrick17k	+0.02164	+0.04640	+0.03402 [0.03274, 0.03531]	65.6	66.8	+1.2
CheXpert	+0.00825	+0.00850	+0.00838 [0.00765, 0.00910]	56.6	57.2	+0.6
BreakHis	+0.00835	−0.00203	+0.00316 [0.00160, 0.00471]	79.5	84.1	+4.6
ORIGA	+0.03609	+0.04404	+0.04007 [0.03670, 0.04344]	79.1	80.2	+1.1

struction plus diagnostic feedback. The four-domain recognition and alignment pattern, lower KID than TI+LoRA across those domains, and text-alignment gains on FLUX suggest that the effect is not confined to one dataset or generator. Physician preferences provide a separate, relative assessment of the generated images.

Evidence is limited to two readers, evaluator-defined recognition and alignment that may benefit from transfer among pretrained encoders even though DINOv2 and MetaCLIP2 differ from the MedSigLIP training critic, image-key splits for Fitzpatrick17k and ORIGA without established patient-level separation (Appendix A), and one CheXpert initialization; the single-checkpoint removal arms cannot resolve reward interactions. Broader reader evaluation, patient-level split verification, and repeated training would test how well these findings generalize.

6 Conclusion

CRAFT shows that feedback on sampled medical images makes requested conditions more recognizable to a held-out classifier trained on real images while improving alignment with instance descriptions and shared clinical criteria. At a matched CheXpert initialization, this feedback raised held-out DINOv2 balanced accuracy by 10.70 points over equal additional reconstruction. The four-domain classifier and text-alignment gains, together with physician preference for CRAFT in the blinded four-way comparison, support combining text and diagnostic feedback to generate condition-specific medical images.

ETHICS STATEMENT

The experiments use existing medical-image research datasets. The physician preference study evaluates synthetic images and does not guide patient care. The two physicians evaluated method-anonymous, randomly ordered candidates without additional compensation. Data and pretrained-model access remain subject to their original terms. Synthetic images can contain plausible but incorrect findings and should not be represented as real patient observations or as validated clinical evidence.

REPRODUCIBILITY STATEMENT

The supplementary code package provides the TI-conditioned LoRA stage and CRAFT reward-stage trainer (reconstruction, VDC, CCS, and DD), the MedSigLIP-448 DD-head trainer, the sampling script and per-dataset launch schedules (Tables A1–A2), DINOv2 and MetaCLIP2 evaluation with paired source-level bootstrap, matched-continuation and control recipes, and the downstream augmentation script. It also includes shared checklists and prompt construction, learned textual-inversion embeddings, fixed per-image descriptions, trained linear heads, static protocol checks, and unit tests. The textual-inversion trainer, GPT-4o enrichment runner (whose prompts appear in Appendix B), DPO comparator trainer, and KID, LPIPS, FLUX, and physician-study analyses are outside the package. Licensed medical images and base-model weights are not redistributed.

AI USE STATEMENT

GPT-4o was used to prepare fixed descriptions and class-level checklists. MedGemma provided offline preference labels for the DPO comparator. Generative AI also assisted with analysis planning, code, manuscript drafting, editing, and figure preparation. Human reader judgments are reported separately from model-generated feedback. The authors take responsibility for the experimental records and scientific claims.

REFERENCES

Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2023.

Pierre Chambon, Christian Bluethgen, Jean-Benoit Delbrouck, Rogier van der Sluijs, Małgorzata Połacin, Juan Manuel Zambrano Chaves, Tanishq Mathew Abraham, Shivanshu Purohit, Curtis P. Langlotz, and Akshay Chaudhari. RoentGen: Vision-language foundation model for chest x-ray generation. *arXiv preprint arXiv:2211.12737*, 2022.

Yung-Sung Chuang, Yang Li, Dong Wang, Ching-Feng Yeh, Kehan Lyu, Ramya Raghavendra, James Glass, Lifei Huang, Jason Weston, Luke Zettlemoyer, et al. Meta clip 2: A worldwide scaling recipe. *arXiv preprint arXiv:2507.22062*, 2025.

Kevin Clark, Paul Vicol, Kevin Swersky, and David Fleet. Directly fine-tuning diffusion models on differentiable rewards. In *International Conference on Learning Representations*, pp. 4793–4822, 2024.

Joseph Paul Cohen, Margaux Luck, and Sina Honari. Distribution matching losses can hallucinate features in medical image translation. In *International conference on medical image computing and computer-assisted intervention*, pp. 529–536. Springer, 2018.

Bram De Wilde, Anindo Saha, Maarten de Rooij, Henkjan Huisman, and Geert Litjens. Medical diffusion on a budget: textual inversion for medical image generation. *arXiv preprint arXiv:2303.13430*, 2023.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

Matthew Groh, Caleb Harris, Luis Soenksen, Felix Lau, Rachel Han, Aerin Kim, Arash Koochek, and Omar Badri. Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 1820–1828, 2021.

Woojung Han, Chanyoung Kim, Dayun Ju, Yumin Shim, and Seong Jae Hwang. Advancing text-driven chest x-ray generation with policy-based reinforcement learning. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume 15003 of *Lecture Notes in Computer Science*, pp. 56–66. Springer Nature Switzerland, oct 2024. doi: 10.1007/978-3-031-72384-1_6.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Peng Huang, Junhu Fu, Bowen Guo, Zeju Li, Yuanyuan Wang, and Yi Guo. Vap-diffusion: Enriching descriptions with mllms for enhanced medical image generation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume 15970 of *Lecture Notes in Computer Science*, pp. 624–633. Springer Nature Switzerland, 2025. doi: 10.1007/978-3-032-05141-7_60.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 590–597, 2019.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Parham Saremi, Amar Kumar, Mohamed Mohamed, Zahra TehraniNasab, and Tal Arbel. RL4Med-DDPO: Reinforcement learning for controlled guidance towards diverse medical image generation using vision-language foundation models. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume 15963 of *Lecture Notes in Computer Science*, pp. 478–488. Springer Nature Switzerland, sep 2025. doi: 10.1007/978-3-032-04965-0_45.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

Fabio A. Spanhol, Luiz S. Oliveira, Caroline Petitjean, and Laurent Heutte. A dataset for breast cancer histopathological image classification. *IEEE Transactions on Biomedical Engineering*, 63 (7):1455–1462, 2016. doi: 10.1109/TBME.2015.2496264.

Shenghuan Sun, Greg Goldgof, Atul Butte, and Ahmed M Alaa. Aligning synthetic medical images with clinical knowledge using human feedback. In *Advances in Neural Information Processing Systems*, volume 36, pp. 13408–13428, 2023.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8228–8238, 2024.

Janet Wang, Yunbei Zhang, Zhengming Ding, and Jihun Hamm. Doctor approved: Generating medically accurate skin disease images through ai-expert feedback. In Danielle Belgrave, Cheng Zhang, Howard Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, and Nanbo Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, pp. 19971–20003. Curran Associates, Inc., 2025. doi: 10.52202/085713-0674.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Zhuo Zhang, Feng Shou Yin, Jiang Liu, Wing Kee Wong, Ngan Meng Tan, Beng Hai Lee, Jun Cheng, and Tien Yin Wong. Origa-light: An online retinal fundus image database for glaucoma analysis and research. In *2010 Annual international conference of the IEEE engineering in medicine and biology*, pp. 3065–3068. IEEE, 2010.

Appendix

Contents

A. Implementation and evaluation protocols 14
Training splits, text inputs, sampling and statistical units

B. Prompt enrichment pipeline 16
Domain-specific system prompts and schema constraints

C. Class-shared clinical checklists 18
Fitzpatrick17k (20), CheXpert (4), BreakHis (8), ORIGA (2)

D. Matched reconstruction-only continuation 24
Five-horizon DINOv2, MetaCLIP2, and KID results

E. Complete held-out diagnostic evaluation 26
Full methods, per-domain contrasts, real reference

F. Complete held-out semantic evaluation 28
MetaCLIP2 description/checklist means and seed variation

G. Reward-removal studies 30
CheXpert and BreakHis external contrasts

H. Blinded reader preference 31
Method shares, reader agreement, and ranking instructions

I. Backbone transfer and prompt sensitivity 32
FLUX contrasts and label-substitution analysis

J. Copy and proximity screens 33
SSIM, perceptual hash, and DINOv2 nearest neighbors

K. Extended qualitative results 35
Four modality galleries and CheXpert error cases

L. Checkpoint selection 41
Validation-selection rule and TI+LoRA trajectory

M. Downstream augmentation 42
Three-scale accuracy and full-data comparisons

N. Held-out label composition 43
Per-class test counts and synthetic evaluation counts

Table A1: Data partitions and validation-selected epochs. TI+LoRA epochs belong to the standard adaptation trajectory; CRAFT epochs count its additional reward-adaptation stage.

Dataset	Classes	Train	Val.	Test	TI+LoRA	CRAFT
Fitzpatrick17k	20	2781	310	3100	4	10
CheXpert	4	445	50	495	4	5
BreakHis	8	936	104	1041	8	5
ORIGA	2	408	46	196	1	5

Table A2: Primary optimization schedules. T is the reward-sampling trajectory length, K the number of differentiable denoising steps, and M the trajectories per update.

Setting	Fitzpatrick17k	CheXpert	BreakHis	ORIGA
Stage-one learning rate	10^{-4}	10^{-4}	10^{-5}	10^{-5}
Stage-one schedule (epochs)	10	10	10	10
Reward-stage schedule (epochs)	10	10	5	5
Reward-stage learning rate	10^{-5}	10^{-5}	10^{-5}	10^{-5}
$T/K/M$	20/1/2	20/1/1	40/1/1	20/1/1

A Implementation and evaluation protocols

CRAFT combines three generated-image feedback terms with real-image reconstruction. This appendix documents the primary benchmark, matched continuation, reader evaluation, and supporting experiments.

A.1 Data partitions and conditions

Validation comprises 10% of the training pool, rounded upward per dataset. The Fitzpatrick17k subset follows the 20-condition setup of Wang et al. (2025). A local split audit describes class-stratified image-level sampling; retained train/test image keys are disjoint, but the records contain no patient identifiers. CheXpert uses frontal-view radiographs with exactly one unambiguous target label among No Finding, Pleural Effusion, Cardiomegaly, and Pneumonia; uncertain labels and records with multiple positive target labels were excluded. CheXpert is partitioned by patient across train, validation, and test, although the retained split files omit patient IDs and do not permit an independent patient-ID check. BreakHis uses eight categories at $100\times$ magnification and is partitioned by patient across train, validation, and test. ORIGA uses glaucoma and non-glaucoma labels; its retained train/test image keys are disjoint, but patient or eye grouping is not established by the available split records.

The real-image reference in the diagnostic table is held out from classifier fitting. Logistic-regression regularization is selected on a real-only validation split; the final classifier is then fitted on all designated real training images. This procedure is separate from generator checkpoint selection. Classifier predictions on generated images are compared with the conditioning class, not a specialist diagnosis assigned to each synthetic image.

A.2 Generator recipe

The primary recipe loads Stable Diffusion v2.1-base from the community-hosted Hugging Face repository `Manojb/stable-diffusion-2-1-base`[1]. It learns textual-inversion tokens for 3,000 steps, then adapts attention-layer LoRA parameters. The training configuration uses rank 4, batch size 2 per device, gradient accumulation 2, and training seed 42. The reward-stage learning rate is 10^{-5}. The reconstruction, description, checklist, and diagnostic coefficients are each 0.2. Evaluation uses five generation seeds.

The training critic is MedSigLIP-448. Its linear diagnostic head is fitted on real training features with Adam for 50 epochs at learning rate 10^{-3} and batch size 256. Encoders and probes remain

[1] `https://huggingface.co/Manojb/stable-diffusion-2-1-base`

Table A3: Description fields and content constraints in the prompt schemas.

Domain	Fields	Content
Dermatology	Body part; lesion features	Shape, texture, and color of visible features; bounded-length descriptions.
Radiology	Devices; key findings; visual summary	Visible radiographic descriptors; no pathology terms for No Finding.
Histopathology	Tissue context; key findings; visual summary	Patch-visible glandular, stromal, fibroepithelial, infiltrative, or mixed morphology.
Ophthalmology	Disc appearance; key findings; visual summary	Cup, rim, vessels, and peripapillary appearance, not pressure, visual-field, or prognostic claims.

frozen during generator adaptation. Gradients pass through the frozen image encoder to the generated image and LoRA parameters, not to the critic. The DPO comparator adapts the checklist-ranked preference/DPO protocol of Wang et al. (2025), using 1,024 MedGemma-ranked offline preference pairs per dataset. LoRA is applied at generation. All methods here, including adapted DPO, are evaluated by text-to-image generation from noise. In contrast, the cited work constructs preference pairs and primarily generates augmentation images by partially noising real source images (denoising strength $\gamma = 0.3$), then changing the target skin condition while preserving most source anatomical context. Our diagnostic scores therefore assess the adapted comparator under the common from-noise task, not the cited method's performance in its native source-conditioned image-to-image setting.

A.3 Text inputs

GPT-4o constructs fixed descriptions from image–label pairs. The ground-truth label constrains the diagnosis; the image supplies instance-specific attributes. Shared checklists are label-conditioned. Reward sampling uses human-readable descriptions, while disease-name-substituted versions condition reconstruction and adapted-generator evaluation. The reader study instead uses class-level label-plus-criteria prompts. Appendices B and C provide the full prompt logic and class checklists.

The checklist reward embeds the complete text. Its value is not the fraction of individually satisfied criteria. The same class-level checklist text is used for training and evaluation (Appendix C). The generator samples from noise; an image-derived prompt is textual conditioning, not an image-to-image generation input.

A.4 Sampling and evaluation units

The adapted-method target comparison uses the same source rows and five common generation seeds within each dataset, including Fitzpatrick17k DINOv2. The Fitzpatrick17k synthetic diagnostic panel has 3,100 underlying sources and 15,500 predictions per method; its Real row is a separate 3,100-image reference. Fitzpatrick17k BA intervals use 20,000 class-stratified source-bootstrap resamples, while macro-F1 intervals use 2,000. The five generations from each source remain nested in each resample. All adapted-method semantic tables also use five seeds. Semantic standard deviations describe variation across generation-seed means; these seeds are not independent training runs, and their SD does not measure source or training-run uncertainty.

KID uses generation seed 42 and equal real/synthetic sample counts with matched class composition. DINOv2 feature-space KID is not numerically interchangeable with Inception-FID or Inception-KID. DINOv2-KID and DINOv2 diagnostic classification share an encoder family and are not independent evaluator families. Within-prompt LPIPS uses 100 held-out prompts and four shared seeds per dataset, yielding six unordered generated/generated pairs per prompt. Pair distances are averaged within each prompt and then across prompts; they are not generated/real reconstruction distances. Higher LPIPS can reflect useful variation or unwanted inconsistency, so LPIPS direction is not a quality ranking. These are different evaluation panels from the full source-level diagnostic comparison. The real reference has no generator-training seed distribution and is not assigned a training-run standard deviation.

B PROMPT ENRICHMENT PIPELINE

The four system prompts below specify the enrichment logic used in CRAFT. GPT-4o receives an image and its ground-truth label and returns structured text before reward adaptation.

B.1 LABEL CONDITIONING AND DOMAIN SCHEMAS

Each box specifies the domain's fields, vocabulary constraints, and style rules. Labels and allowed-vocabulary placeholders are preserved verbatim. The instructions constrain the requested description; they do not ask the captioner to predict a diagnosis.

B.2 DERMATOLOGY (FITZPATRICK17K)

Dermatology (Fitzpatrick17k)

You're an expert dermatologist. You extract visual morphology from ONE clinical image. The ground truth diagnosis is provided below as **{label}**. Do not invent findings outside of this diagnosis.

Return STRICTLY valid JSON with keys:

- body_part (MUST be one of: {allowed_body_parts})
- lesion_features (concise phrase $\leq$ 20 words: shape, texture, color)

Style constraints:

- Use plain language. Avoid medical jargon (say "red" not "erythematous").
- Focus on visual attributes of **{label}** (shape, border, color).
- Output JSON only. No markdown.

B.3 HISTOPATHOLOGY (BREAKHIS)

Histopathology (BreakHis)

You're an expert breast pathologist. You describe ONE H&E breast histopathology patch at 100x magnification.

The ground-truth diagnosis is provided below as **{label}**. Do not predict a diagnosis. Do not invent findings outside of this diagnosis. Only mention morphology that is reasonably visible in a single 100x patch.

Return STRICTLY valid JSON with keys:

- tissue_context
- key_findings
- visual_summary

Schema rules:

- tissue_context: MUST be one of ["glandular", "fibroepithelial", "stromal", "papillary", "mucinous", "infiltrative", "mixed"]
- key_findings: comma-separated list chosen ONLY from **{allowed_key_findings}**
- visual_summary: concise phrase $\leq$ 18 words

Style constraints:

- Use visual morphology only.
- Prefer short pathology terms such as "enlarged acini", "slit-like ducts", "leaf-like stromal fronds", "tightly packed tubules", "irregular ducts", "single-file cells", "mucin pools", "papillary fronds", and "fibrovascular cores".

Histopathology (BreakHis) (continued)

- Do not mention biomarkers, grade, mitotic rate, receptor status, prognosis, or treatment.
- Do not mention features that cannot be supported by the patch.
- Avoid long sentences and differential diagnosis language.
- Output JSON only. No markdown.

B.4 RADIOLOGY (CHEXPERT)

Radiology (CheXpert)

You're an expert radiologist. Describe a frontal chest x-ray using simple visual language. The ground truth diagnosis is provided below as **{label}**. Do not invent findings outside of this diagnosis.

Return STRICTLY valid JSON with keys:

- devices (e.g., "pacemaker", "picc line", or "none")
- key_findings (comma-separated list from allowed pool)
- visual_summary (concise phrase $\leq$ 16 words)

Style constraints:

- Avoid Negations: Do NOT say "no pleural effusion". Instead, say "sharp costophrenic angles".
- Prefer positive descriptors (e.g., "clear lungs") for normal regions.
- If label="No Finding", forbid pathology terms entirely.
- Output JSON only. No markdown.

B.5 OPHTHALMOLOGY (ORIGA)

Ophthalmology (ORIGA)

You're an expert ophthalmologist. You describe ONE retinal fundus image.

The ground-truth diagnosis is provided below as **{label}**. Do not predict a diagnosis. Do not invent findings outside of this diagnosis. Only mention morphology that is reasonably visible in a single fundus photograph.

Return STRICTLY valid JSON with keys:

- disc_appearance
- key_findings
- visual_summary

Schema rules:

- disc_appearance: MUST be one of ["normal disc", "enlarged cup", "rim thinning", "notched rim", "glaucomatous disc"]
- key_findings: comma-separated list chosen ONLY from **{allowed_key_findings}**
- visual_summary: concise phrase $\leq$ 18 words

Style constraints:

- Use only visible fundus morphology.
- Focus on optic cup size, neuroretinal rim, vessel course, and peripapillary changes.
- Do not mention visual field loss, intraocular pressure, OCT, prognosis, or treatment.
- Do not mention features not clearly supported by the image.
- Output JSON only. No markdown.

C CLASS-SHARED CLINICAL CHECKLISTS

These tables list the class-level criteria used by CRAFT across 20 dermatology conditions, four radiology classes, eight histopathology classes, and two retinal classes. CRAFT embeds the complete class checklist for feedback; a checklist-similarity score is not an item-by-item clinical adjudication.

C.1 FITZPATRICK17K CHECKLISTS

Table A4: Fitzpatrick17k class-shared checklist text.

Condition	Attribute	Visual description
Acne	Location	Face, forehead, chest, shoulders, upper back
	Lesion Type	Bumps including comedones (whiteheads, blackheads) and inflamed pimples
	Shape/Size	Small clogged-pore bumps; larger tender nodules/cysts
	Color	Red or skin-colored bumps; blackheads have dark plug
	Texture	Oily or shiny skin; pus or crust if ruptured
Actinic Keratosis	Location	Sun-exposed areas (face, scalp, ears, hands)
	Lesion Type	Rough, scaly patch or small crusty bump
	Shape/Size	Flat or slightly raised, under 2.5 cm
	Color	Pink, red, or brownish; yellowish crust
	Texture	Dry, coarse, sandpaper-like surface
Allergic Contact Dermatitis	Location	Contact sites (hands, face, neck)
	Lesion Type	Red patches with blisters or swelling
	Shape/Size	Irregular shape following exposure
	Color	Pink to red; purple/brown on dark skin
	Texture	Weepy, crusty, or scaly
Basal Cell Carcinoma	Location	Sun-exposed areas (face, nose, neck)
	Lesion Type	Pearly bump or flat scaly patch
	Shape/Size	Small round/oval; central depression
	Color	Translucent/pearly; brown/black on dark skin
	Texture	Smooth, shiny; can crust or ulcerate
Eczema	Location	Flexural areas, hands, neck, eyelids
	Lesion Type	Patches/plaques, sometimes blisters
	Shape/Size	Ill-defined patches; often symmetric
	Color	Red/pink; purple/gray/brown on dark skin
	Texture	Dry, flaky, leathery (lichenification)
Erythema Multiforme	Location	Hands, feet, limbs, mucous membranes
	Lesion Type	Target (bull's-eye) lesions
	Shape/Size	Round (1-3 cm) with concentric rings
	Color	Dark center, pale ring, red outer ring
	Texture	Flat with raised center
Folliculitis	Location	Beard, scalp, legs, buttocks
	Lesion Type	Pustules centered on hair follicles
	Shape/Size	Clusters of 2-5 mm bumps
	Color	Red/pink; pus is white/yellow
	Texture	Dome-shaped, fluid-filled top
Granuloma Annulare	Location	Hands, feet, wrists, ankles
	Lesion Type	Smooth firm bumps forming rings
	Shape/Size	Annular (ring-shaped); papules
	Color	Skin-colored, pink, reddish
	Texture	Smooth, non-scaly

Continued on next page

Table A4 continued

Condition	Attribute	Visual description
Keloid	Location	Scars on chest, ears, jawline
	Lesion Type	Overgrown scar tissue
	Shape/Size	Raised, irregular, extends beyond wound
	Color	Pink/red; purple/brown on dark skin
	Texture	Smooth, firm, rubbery, shiny
Lichen Planus	Location	Wrists, ankles, mouth, genitals
	Lesion Type	Flat-topped papules
	Shape/Size	Polygonal 2-10 mm papules
	Color	Violaceous (purple); gray-brown
	Texture	Shiny with fine white lines (Wickham's striae)
Lupus Erythematosus	Location	Face (butterfly rash), scalp, ears
	Lesion Type	Malar rash or discoid plaques
	Shape/Size	Butterfly shape; coin-shaped discoid
	Color	Pink-red; hyperpigmented borders
	Texture	Smooth or scaly/scarring
Melanoma	Location	Anywhere; palms/soles in dark skin
	Lesion Type	Atypical mole/patch
	Shape/Size	Asymmetric, >6mm, irregular borders
	Color	Variegated (brown, black, red, blue)
	Texture	Smooth or ulcerated/crusted
Mycosis Fungoides	Location	Sun-protected areas (buttocks, thighs)
	Lesion Type	Patches, plaques, or tumors
	Shape/Size	Irregular patches; thick plaques
	Color	Reddish-brown; hyperpigmented
	Texture	Scaly, wrinkled (cigarette paper)
Pityriasis Rosea	Location	Trunk (Christmas tree pattern)
	Lesion Type	Herald patch then smaller spots
	Shape/Size	Oval patches
	Color	Salmon/pink; purple/brown
	Texture	Fine collarette scale
Prurigo Nodularis	Location	Limbs, back (reachable areas)
	Lesion Type	Firm itchy nodules
	Shape/Size	Round 1-3 cm nodules
	Color	Pink, brown, black
	Texture	Hard, rough, crusted/scabbed
Psoriasis	Location	Elbows, knees, scalp, lower back
	Lesion Type	Well-demarcated plaques
	Shape/Size	Round/oval plaques
	Color	Pink/red with silvery scale
	Texture	Thick dry scale; bleeding if peeled
Sarcoidosis	Location	Face, shins, scars
	Lesion Type	Firm plaques or nodules
	Shape/Size	Broad plaques; 1-5 cm nodules
	Color	Red-brown; purple; discoloration
	Texture	Smooth, firm, rubbery
Scabies	Location	Finger webs, wrists, waist
	Lesion Type	Burrows and small bumps
	Shape/Size	Linear burrows (5-15 mm)
	Color	Skin-toned/pink
	Texture	Scratch marks, slight ridges

Continued on next page

Table A4 continued

Condition	Attribute	Visual description
Squamous Cell Carcinoma	Location	Sun-exposed areas
	Lesion Type	Crusty bump or ulcer
	Shape/Size	Firm nodule; may grow rapidly
	Color	Pink/red base; white scale
	Texture	Rough, thick, crusted, tender
Vitiligo	Location	Face, hands, genitals
	Lesion Type	Depigmented patches
	Shape/Size	Irregular, well-defined
	Color	White/pale
	Texture	Normal smoothness

C.2 CHEXPERT CHECKLISTS

Table A5: CheXpert class-shared checklist text.

Condition	Attribute	Visual description
No Finding	Location	Entire chest field (Lungs, Heart, Pleura)
	Lesion Type	Normal anatomy; no pathological opacities
	Shape/Size	Heart size normal (less than half chest width); sharp triangular costophrenic angles
	Color	Lungs are dark (radiolucent); bones and heart are bright white
	Texture	Clear lung fields with fine, branching vascular markings; no haziness
Cardiomegaly	Location	Central chest (Mediastinum/Heart)
	Lesion Type	Enlargement of the cardiac silhouette
	Shape/Size	Globular or widened heart shadow; width $>$ 50% of the rib cage
	Color	Enlarged central white opacity
	Texture	Smooth, distinct heart borders; lung fields may be clear or show vascular congestion
Pneumonia	Location	Focal area in lungs (often asymmetric, lobar, or segmental)
	Lesion Type	Consolidation or Infiltrate (filled airspaces)
	Shape/Size	Patchy, irregular cloud-like shape; borders may be fluffy or ill-defined
	Color	Increased whiteness (opacity) against the dark lung background
	Texture	Fluffy or hazy 'ground glass' appearance; may show air bronchograms (dark tubes visible inside white area)
Pleural Effusion	Location	Lung bases (bottom corners) or lining the chest wall
	Lesion Type	Fluid accumulation masking the diaphragm
	Shape/Size	Meniscus sign (U-shaped curve) at the edges; blunting of the sharp costophrenic angle
	Color	Dense, homogeneous white opacity (blocks view of diaphragm/ribs behind it)
	Texture	Smooth, uniform density (water-like); obscures the sharp angle of the lung base

C.3 BREAKHIS CHECKLISTS

Table A6: BreakHis class-shared checklist text.

Condition	Attribute	Visual description
Adenosis	Tissue Context	Lobulocentric glandular units with crowded acini in fibrous stroma
	Lesion Type	Benign proliferative glandular lesion with enlarged acini
	Shape/Size	Numerous small round acini with relatively regular lumina; clustered lobular pattern
	Color	Pink eosinophilic stroma with blue-purple uniform epithelial nuclei
	Texture	Orderly gland architecture; smooth outlines; bland cytology
Fibroadenoma	Tissue Context	Fibroepithelial lesion with ducts embedded in fibrous stroma
	Lesion Type	Well-circumscribed biphasic benign lesion
	Shape/Size	Round to slit-like ducts compressed by stroma; pushing contours
	Color	Pale pink collagenous stroma with evenly basophilic epithelial nuclei
	Texture	Smooth fibrous background; bland epithelium; low atypia
Phyllodes Tumor	Tissue Context	Fibroepithelial lesion with epithelial-lined stromal clefts
	Lesion Type	Leaf-like stromal fronds projecting into clefted spaces
	Shape/Size	Elongated branching fronds and cleft-like spaces; broader architecture than fibroadenoma
	Color	Pink stromal fronds with blue-purple epithelial and stromal nuclei
	Texture	Undulating leaf-like contours; more cellular stroma; mixed epithelial-stromal pattern
Tubular Adenoma	Tissue Context	Gland-forming lesion dominated by tightly packed tubules
	Lesion Type	Benign tubular epithelial proliferation
	Shape/Size	Numerous small round or oval tubules with open lumina; minimal intervening stroma
	Color	Pink scant stroma with blue-purple uniform nuclei around luminal spaces
	Texture	Dense orderly tubular pattern; smooth gland borders; bland cytology
Ductal Carcinoma	Tissue Context	Infiltrative epithelial lesion within fibrotic or desmoplastic stroma
	Lesion Type	Malignant duct-forming carcinoma
	Shape/Size	Irregular angulated ducts, nests, or cords with variable lumina
	Color	Hyperchromatic blue-purple nuclei in pink desmoplastic stroma
	Texture	Crowded pleomorphic cells; jagged infiltrative architecture; reduced uniformity
Lobular Carcinoma	Tissue Context	Discohesive infiltrating cells in fibrous stroma, often around ducts or lobules
	Lesion Type	Malignant lobular-type infiltrate
	Shape/Size	Single-file cords and targetoid periductal arrangement; little tubule formation
	Color	Small dark nuclei scattered through pale to pink stroma
	Texture	Loose discohesive pattern; subtle infiltrative spread; monotonous small cells

Continued on next page

Table A6 continued

Condition	Attribute	Visual description
Mucinous Carcinoma	Tissue Context	Tumor cell clusters suspended within extracellular mucin pools
	Lesion Type	Mucin-producing carcinoma
	Shape/Size	Rounded clusters or strips of cells floating in large mucin lakes
	Color	Pale blue to lightly basophilic mucin with darker tumor cell nuclei
	Texture	Smooth gelatinous background; low-density floating cellular islands; soft lobulated spaces
Papillary Carcinoma	Tissue Context	Papillary epithelial proliferation organized around fibrovascular cores
	Lesion Type	Malignant papillary lesion
	Shape/Size	Branching papillary fronds with central fibrovascular stalks
	Color	Pink fibrovascular cores lined by blue-purple crowded epithelial cells
	Texture	Frond-like branching surface; layered epithelium; delicate core structures

C.4 ORIGA CHECKLISTS

Table A7: ORIGA class-shared checklist text.

Condition	Attribute	Visual description
Normal	Optic Disc	Small to moderate optic cup with balanced cup-to-disc ratio and preserved neuroretinal rim
	Rim	Neuroretinal rim appears circumferentially intact without focal thinning or notching
	Vessels	Retinal vessels near the disc follow a normal course without marked nasalization or bayoneting
	Peripapillary Region	No obvious nerve fiber layer defect, disc hemorrhage, or glaucomatous peripapillary change
	Overall Fundus	Overall retinal fundus appearance is consistent with a non-glaucomatous eye
Glaucoma	Optic Disc	Enlarged optic cup or increased vertical cup-to-disc ratio is visible
	Rim	Neuroretinal rim shows focal or diffuse thinning, notching, or clear rim loss, especially in the superior or inferior rim
	Vessels	Retinal vessels near the disc show nasalization, bayoneting, or displacement around the cup edge
	Peripapillary Region	Peripapillary nerve fiber layer loss, disc hemorrhage, or localized glaucomatous change is visible or suspiciously present
	Overall Fundus	Overall optic nerve head appearance is consistent with glaucomatous optic neuropathy

Table A8: CRAFT exceeds matched reconstruction-only continuation in DINOv2 balanced accuracy and macro-F1. Values are percentages and differences are percentage points across additional training horizons. Intervals are paired source-level 95% bootstrap intervals.

Additional epoch	Recon-only	CRAFT	Difference [95% CI]
Balanced accuracy			
+1	32.99	35.22	$+2.23$ $[+0.62, +3.84]$
+2	32.09	37.04	$+4.95$ $[+2.81, +7.09]$
+3	31.65	39.88	$+8.23$ $[+5.40, +11.06]$
+4	32.45	41.24	$+8.79$ $[+5.71, +11.87]$
+5	31.94	42.64	$+10.70$ $[+7.28, +14.12]$
Macro-F1			
+1	29.40	31.63	$+2.23$ $[+0.48, +3.98]$
+2	29.37	34.93	$+5.56$ $[+3.22, +7.90]$
+3	27.49	35.45	$+7.96$ $[+4.92, +11.00]$
+4	28.81	37.68	$+8.87$ $[+5.60, +12.14]$
+5	28.46	38.52	$+10.06$ $[+6.52, +13.60]$

D MATCHED RECONSTRUCTION-ONLY CONTINUATION

D.1 COMPARISON AND EVALUATION UNITS

Both arms begin from the selected CheXpert TI+LoRA epoch-4 initializer and share source prompts, generation seeds, reconstruction learning rate and optimizer, batch size, and update count. The reconstruction coefficient is .20 in both arms, as in the method objective; the three generated-image rewards are enabled only in CRAFT. The reported horizons are additional epochs $+1$ through $+5$; e1–e5 are local checkpoint labels for each arm. DINOv2 balanced accuracy is the primary endpoint. The probe is fitted on real training images only, and the reported 95% intervals use source-level bootstrap resampling. The validation selection score is not used in this comparison.

The matched-continuation evaluation is a separate CheXpert experiment from the four-domain evaluation and component-removal study.

D.2 DIAGNOSTIC OUTCOMES AT EACH HORIZON

The balanced-accuracy difference is positive at every horizon and increases across the five checkpoints (Table A8). Macro-F1 contrasts are likewise positive at each horizon.

Table A9: MetaCLIP2 semantic outcomes at matched additional horizons. The two text targets are reported separately. The $+5$ CRAFT checklist mean is .2479, as in Table 2; its paired interval uses the corresponding source-level evaluation record.

Additional epoch	Recon-only	CRAFT	Difference [95% CI]
Description similarity			
+1	0.1545	0.2219	$+0.0674$ $[+0.0528, +0.0809]$
+2	0.1490	0.2432	$+0.0942$ $[+0.0651, +0.1256]$
+3	0.1483	0.2707	$+0.1224$ $[+0.0459, +0.1989]$
+4	0.1541	0.2713	$+0.1172$ $[+0.0586, +0.1716]$
+5	0.1461	0.2773	$+0.1312$ $[+0.0949, +0.1675]$
Checklist similarity			
+1	0.2256	0.2440	$+0.0184$ $[+0.0151, +0.0219]$
+2	0.2188	0.2451	$+0.0263$ $[+0.0200, +0.0326]$
+3	0.2291	0.2421	$+0.0130$ $[+0.0065, +0.0195]$
+4	0.2196	0.2427	$+0.0231$ $[+0.0130, +0.0332]$
+5	0.2279	0.2479	$+0.0200$ $[+0.0087, +0.0314]$

Table A10: DINOv2-KID point estimates for the continuation panel, multiplied by 10^3. Lower KID is better. Differences are CRAFT minus reconstruction-only; no interval is supplied.

Additional epoch	Recon-only	CRAFT	Difference
+1	0.0807	0.0749	-0.0058
+2	0.0910	0.0759	-0.0151
+3	0.0757	0.0736	-0.0021
+4	0.0807	0.0704	-0.0103
+5	0.0771	0.0641	-0.0130

D.3 Semantic outcomes at each horizon

Description and checklist differences are positive at all horizons (Table A9), although their magnitudes need not increase monotonically. Source-bootstrap intervals are distinct from the generation-seed standard deviations in the primary semantic table.

D.4 Feature-distribution discrepancy

KID differences (Table A10) are arithmetic contrasts of point estimates. No KID interval or matched-panel LPIPS measurement is included. The primary SD2.1 diversity results are not reassigned to these continuation checkpoints.

Table A11: CRAFT leads synthetic methods in DINOv2 balanced accuracy across four domains. Accuracy, balanced accuracy, and macro-F1 are percentages; AUROC remains a fraction. Real is a separate held-out reference. Synthetic methods use five common generation seeds per dataset.

Dataset	Method	Accuracy	Bal. accuracy	Macro-F1	AUROC
Fitzpatrick17k	Real	55.84	55.12	55.94	0.9199
	TI	35.81	35.68	36.21	0.8560
	DPO	36.70	36.91	37.00	0.8613
	TI+LoRA	44.64	42.29	40.89	0.9155
	CRAFT	50.71	45.98	45.29	0.9220
CheXpert	Real	47.07	47.06	47.06	0.7166
	TI	16.77	16.76	16.64	0.4140
	DPO	17.05	17.05	16.93	0.4161
	TI+LoRA	32.69	32.65	29.27	0.5897
	CRAFT	42.71	42.64	38.52	0.6951
BreakHis	Real	75.31	67.94	69.42	0.9516
	TI	42.19	31.08	30.78	0.7708
	DPO	36.16	31.47	31.81	0.5288
	TI+LoRA	47.78	32.28	31.84	0.7514
	CRAFT	52.41	36.44	33.67	0.8054
ORIGA	Real	79.08	68.70	70.20	0.8074
	TI	75.20	52.73	48.29	0.7271
	DPO	72.86	50.39	45.23	0.5289
	TI+LoRA	73.88	56.29	56.06	0.7524
	CRAFT	78.37	59.58	60.01	0.8065

E COMPLETE HELD-OUT DIAGNOSTIC EVALUATION

Frozen DINOv2 CLS features are passed through a StandardScaler and multinomial logistic-regression classifier. The regularization parameter C is selected on real-only validation data, and the final classifier is fitted on all designated real training images. The same fitted evaluator scores every method within a dataset. Evaluation runs directly on generated images, not on a real-plus-synthetic augmented classifier-training set.

Balanced accuracy and macro-F1 are the primary diagnostic endpoints; accuracy and AUROC are included for completeness. The full comparison (Table A11) shows that CheXpert TI and DPO have below-chance balanced accuracy. The real-image reference is not a guaranteed upper bound on every metric.

Table A12: CRAFT minus each listed comparator under DINOv2, in percentage points. Reported intervals are paired 95% intervals; BreakHis DPO entries are point differences without intervals. BA denotes balanced accuracy. Synthetic methods use five seeds in every domain. ORIGA TI+LoRA point differences use rounded table entries; their paired intervals come from a separate source-level evaluation record.

Dataset	Comparator	Metric	Difference [95% CI]
Fitzpatrick17k	TI+LoRA	BA	$+3.69$ $[+2.53, +4.85]$
	TI+LoRA	Macro-F1	$+4.40$ $[+3.17, +5.62]$
	DPO	BA	$+9.07$ $[+7.82, +10.33]$
	DPO	Macro-F1	$+8.29$ $[+6.99, +9.49]$
	TI	BA	$+10.30$ $[+9.06, +11.54]$
	TI	Macro-F1	$+9.08$ $[+7.84, +10.33]$
CheXpert	TI+LoRA	BA	$+9.99$ $[+7.90, +12.12]$
	TI+LoRA	Macro-F1	$+9.25$ $[+7.13, +11.44]$
	DPO	BA	$+25.59$ $[+23.33, +27.84]$
	DPO	Macro-F1	$+21.59$ $[+19.32, +23.87]$
	TI	BA	$+25.87$ $[+23.58, +28.09]$
	TI	Macro-F1	$+21.88$ $[+19.56, +24.28]$
BreakHis	TI+LoRA	BA	$+4.15$ $[+2.42, +5.88]$
	TI+LoRA	Macro-F1	$+1.83$ $[-0.06, +3.65]$
	DPO	BA	$+4.97$
	DPO	Macro-F1	$+1.86$
	TI	BA	$+5.36$ $[+3.41, +7.29]$
	TI	Macro-F1	$+2.89$ $[+0.80, +4.94]$
ORIGA	TI+LoRA	BA	$+3.29$ $[+0.87, +5.68]$
	TI+LoRA	Macro-F1	$\approx +3.95$ $[+1.23, +6.71]$
	DPO	BA	$+9.19$ $[+5.21, +13.17]$
	DPO	Macro-F1	$+14.78$ $[+9.92, +19.64]$
	TI	BA	$+6.84$ $[+3.79, +9.98]$
	TI	Macro-F1	$+11.72$ $[+7.10, +16.28]$

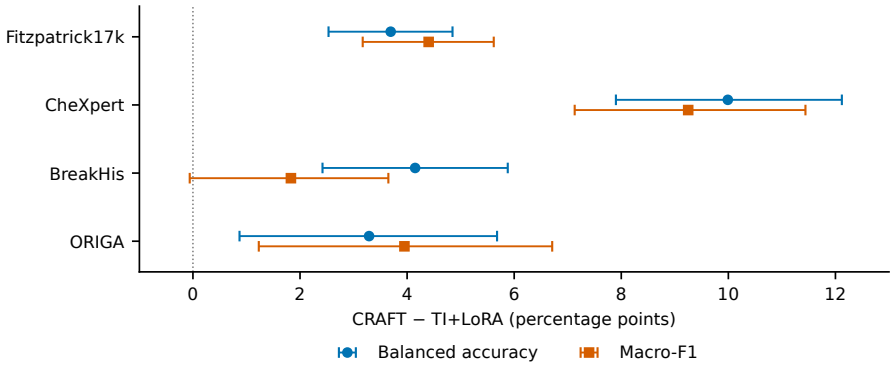

Figure A1: Diagnostic differences between CRAFT and TI+LoRA across the four domains. Whiskers show paired 95% intervals in percentage points. This view supplements the complete comparisons above.

E.1 Paired method contrasts

A contrast computed from the underlying values can differ slightly from subtracting two four-decimal display values. BreakHis CRAFT-minus-DPO BA and macro-F1 are point differences because paired intervals are unavailable for those contrasts. The BreakHis macro-F1 contrast with TI+LoRA includes zero, whereas its balanced-accuracy interval is entirely positive. ORIGA BA and macro-F1 paired intervals are $[+0.87, +5.68]$ and $[+1.23, +6.71]$ percentage points, respectively.

Table A13: Fitzpatrick17k MetaCLIP2 semantic scores. Synthetic entries are mean ± SD across generation-seed means. Real has one reference evaluation.

Method	Seeds	Description	Checklist
Real	1	0.2705	0.2315
TI	5	0.1843 ± 0.0004	0.2310 ± 0.0005
DPO	5	0.1846 ± 0.0003	0.2331 ± 0.0003
TI+LoRA	5	0.1798 ± 0.0001	0.2402 ± 0.0004
CRAFT	5	0.2759 ± 0.0002	0.2519 ± 0.0003

Table A14: CheXpert MetaCLIP2 semantic scores. Synthetic entries are mean ± SD across generation-seed means. Real has one reference evaluation.

Method	Seeds	Description	Checklist
Real	1	0.2795	0.2520
TI	5	0.1586 ± 0.0004	0.2137 ± 0.0013
DPO	5	0.1598 ± 0.0004	0.2167 ± 0.0013
TI+LoRA	5	0.1584 ± 0.0001	0.2272 ± 0.0004
CRAFT	5	0.2773 ± 0.0004	0.2479 ± 0.0006

F COMPLETE HELD-OUT SEMANTIC EVALUATION

MetaCLIP2 scores each generated image against its fixed human-readable description and its class checklist; the two semantic metrics are separate cosine similarities. The evaluator is not used for reward optimization or checkpoint selection. The tables below report five generation seeds for each adapted method.

Table A15: BreakHis MetaCLIP2 semantic scores. Synthetic entries are mean ± SD across generation-seed means. Real has one reference evaluation.

Method	Seeds	Description	Checklist
Real	1	0.2780	0.2478
TI	5	0.2334 ± 0.0003	0.2234 ± 0.0005
DPO	5	0.2492 ± 0.0005	0.2342 ± 0.0007
TI+LoRA	5	0.2473 ± 0.0004	0.2413 ± 0.0006
CRAFT	5	0.2815 ± 0.0003	0.2542 ± 0.0002

Table A16: ORIGA MetaCLIP2 semantic scores. Synthetic entries are mean ± SD across generation-seed means. Real has one reference evaluation.

Method	Seeds	Description	Checklist
Real	1	0.2659	0.2450
TI	5	0.2649 ± 0.0019	0.2322 ± 0.0024
DPO	5	0.2658 ± 0.0004	0.2477 ± 0.0007
TI+LoRA	5	0.2692 ± 0.0012	0.2366 ± 0.0017
CRAFT	5	0.2778 ± 0.0006	0.2541 ± 0.0003

Cross-domain numerical differences should not be read as differences on a common calibrated scale of medical quality.

Table A17: Full CRAFT improves CheXpert balanced accuracy and macro-F1 over w/o DD at two additional training seeds. Diagnostic differences are percentage points; text-similarity differences remain cosine units. Brackets give paired source-bootstrap 95% intervals over 495 held-out sources per pair.

Training seed	BA	Macro-F1	Description	Checklist
101	$+3.69$ $[+1.96, +5.38]$	$+4.89$ $[+2.99, +6.89]$	$+.0012$ $[+.0008, +.0015]$	$+.0006$ $[+.0002, +.0010]$
202	$+4.62$ $[+2.88, +6.39]$	$+6.07$ $[+4.17, +7.97]$	$+.0024$ $[+.0021, +.0028]$	$+.0027$ $[+.0023, +.0031]$

Table A18: Full CRAFT improves CheXpert balanced accuracy and macro-F1 over reconstruction+DD after five additional epochs. Diagnostic differences are percentage points; text-similarity differences remain cosine units. Brackets give paired source-bootstrap 95% intervals over 495 held-out sources. The .20/.20 arm matches Full's retained coefficients; .5/.5 changes them.

Weights	BA	Macro-F1	Description	Checklist
.20/.20 (primary)	$+3.65$ $[+1.42, +5.83]$	$+4.45$ $[+2.01, +6.82]$	$+.1210$ $[+.1191, +.1230]$	$+.0203$ $[+.0185, +.0220]$
.5/.5 (secondary)	$+3.73$ $[+1.47, +5.95]$	$+4.42$ $[+2.17, +6.75]$	$+.1221$ $[+.1205, +.1235]$	$+.0196$ $[+.0181, +.0212]$

G Reward-removal studies

The original-checkpoint absolute outcomes are reported in Table 1.

Each row removes one objective term: diagnostic feedback (DD), reconstruction, checklist feedback (CCS), or description feedback (VDC). The removed term is zeroed and each retained coefficient remains .20, without redistribution. The w/o DD arm retains reconstruction, VDC, and CCS. All five variants use the same CheXpert evaluation protocol (DINOv2 classifier and MetaCLIP2).

At the original checkpoint, Full minus w/o diagnostic is +9.01 BA and +8.27 macro-F1 percentage points, while its description and checklist differences are +.0006 and +.0001. Full minus w/o description is +.1187 in description similarity. These point differences follow from Table 1; paired intervals for these original removal contrasts are not reported here.

Two additional Full-versus-w/o-DD training-seed pairs use seeds 101 and 202. Both arms in each pair start from the selected CheXpert TI+LoRA epoch-4 initializer, continue for five reward-stage epochs under the same training/evaluation setup, and keep retained coefficients at .20 without redistribution. Full and w/o-DD DINOv2 BA are 40.47% versus 36.78% at seed 101 and 42.01% versus 37.40% at seed 202. Table A17 reports each pair separately. Both pairs show positive MetaCLIP2 description and checklist differences with paired source-level intervals above zero.

All arms in Table A18 start from CheXpert TI+LoRA epoch 4 and use five additional reward-stage epochs. Full BA/macro-F1 is 42.64%/38.52% versus 38.99%/34.07% for reconstruction+DD .20/.20; .5/.5 changes coefficients. This tests the VDC+CCS bundle in CheXpert, not individual-term effects.

G.1 BreakHis diagnostic-feedback removal

In a separate BreakHis comparison, DD is set to zero while reconstruction, VDC, and CCS retain coefficients of .20. The two fixed checkpoints are evaluated on 1,041 held-out sources with five generation seeds nested within each source. CRAFT and w/o DD have DINOv2 balanced accuracy 36.44% and 32.48%, respectively, a paired source-level difference of $+3.96$ $[2.16, 5.72]$ points. Their macro-F1 values are 33.67% and 32.22%, a difference of $+1.45$ $[-0.41, 3.36]$ points; this interval includes zero. AUROC differs by +.0521 without an interval. Paired MetaCLIP2 description and checklist differences are $+.0032$ $[.0029, .0035]$ and $+.0036$ $[.0032, .0040]$, respectively. W/o DD is close to TI+LoRA on BA ($+0.20$ $[-1.53, +1.93]$ points) and macro-F1 ($+0.38$ $[-1.40, +2.17]$ points).

Table A19: Blinded four-way top-1 preference (percent). The final column is CRAFT minus TI+LoRA in percentage points, with case-bootstrap 95% intervals. Both judgments remain nested within each case.

Dataset	TI	DPO	TI+LoRA	CRAFT	Difference [95% CI]
Fitzpatrick17k	12.0	14.0	27.0	47.0	$+20.0$ $[6.5, 33.5]$
CheXpert	0.0	0.0	33.0	67.0	$+34.0$ $[18.0, 50.0]$
BreakHis	4.0	15.0	20.0	61.0	$+41.0$ $[28.0, 53.5]$
ORIGA	16.0	13.0	25.0	46.0	$+21.0$ $[7.5, 34.0]$
Pooled	8.0	10.5	26.3	55.3	$+29.0$ $[22.0, 35.9]$

Table A20: Agreement between the two reader judgments within cases.

Dataset	Raw agreement	Cohen's κ
Fitzpatrick17k	.58	.386
CheXpert	.76	.466
BreakHis	.75	.558
ORIGA	.51	.291

H BLINDED READER PREFERENCE

The two physicians each evaluated four method-anonymous candidates per case: TI, DPO, TI+LoRA, and the three-feedback CRAFT variant (VDC + CCS + DD + reconstruction). Candidates within each case shared class-level prompt content and the generation seed; positions were randomized and method mappings were not provided. The endpoint was each physician's preferred candidate. Using disease-specific checklist guidance, they considered target-label consistency, clinical plausibility, anatomical or tissue realism, and overall image quality. They received no additional compensation. CheXpert images were displayed at 512 pixels; Fitzpatrick17k, BreakHis, and ORIGA at 128 pixels.

The study has 100 cases per domain, 400 cases overall, and two judgments per case. Case-bootstrap intervals preserve both reader judgments within each case. Pooled resampling is stratified by modality. The bootstrap uses 20,000 resamples. The reported preference shares are four-way top-1 choices, not head-to-head win rates or absolute correctness rates.

Reader labels apply only to the images displayed in the study; the automatic benchmark uses a separate final generated-image panel, so the two evaluations do not provide per-image joint validation.

H.1 CHEXPERT RANKING INSTRUCTION

The CheXpert reader study used the following instruction:

> "For each case, you will see the target diagnosis and anonymized synthetic images generated by different methods in randomized order. Please rank the images from best to worst according to clinical plausibility, diagnostic consistency with the target label, anatomical realism, and overall image quality. Do not attempt to identify the generation method."

The reader interface used a randomized grid with rank-entry fields.

Table A21: FLUX.1-schnell transfer: CRAFT minus enriched-prompt TI+LoRA under MetaCLIP2. Component differences are shown separately; the final column averages the two text similarities with a paired-source interval.

Dataset	Description	Checklist	Text mean [95% CI]
Fitzpatrick17k	+0.02164	+0.04640	$+0.03402\ [0.03274, 0.03531]$
CheXpert	+0.00825	+0.00850	$+0.00838\ [0.00765, 0.00910]$
BreakHis	+0.00835	-0.00203	$+0.00316\ [0.00160, 0.00471]$
ORIGA	+0.03609	+0.04404	$+0.04007\ [0.03670, 0.04344]$

Table A22: Within-prompt LPIPS differences in the separate FLUX experiment. Positive values indicate greater within-prompt perceptual distance, not necessarily higher overall quality.

Dataset	LPIPS difference	95% CI
Fitzpatrick17k	-0.03833	$[-0.04145, -0.03527]$
CheXpert	-0.02287	$[-0.02627, -0.01952]$
BreakHis	+0.02316	$[+0.01970, +0.02660]$
ORIGA	-0.07327	$[-0.07945, -0.06702]$

I Backbone transfer and prompt sensitivity

I.1 FLUX.1-schnell

The FLUX transfer experiment trains one FLUX.1-schnell CRAFT/TI+LoRA model pair per dataset using common enriched prompts and five generated samples per source. Generation uses four sampling steps at 512×512 resolution with guidance scale 0.0. Source counts are 3,100, 495, 1,041, and 196 for Fitzpatrick17k, CheXpert, BreakHis, and ORIGA. Paired-source intervals average the five generations within each source before resampling.

The final column averages the two MetaCLIP2 semantic components for compactness; on BreakHis the positive average combines a positive description change with a small negative checklist change. FLUX and SD2.1 are reported separately because their evaluations use different generator experiments.

I.2 Prompt-sensitivity study

A separate CheXpert prompt experiment evaluates 200 cases with three generation seeds under a real-only DINOv2 classifier. Relative to a modality-only prompt, adding the disease label increases recognition of the requested disease by $+10.2\ [4.7, 15.8]$ points. Substituting an incorrect class decreases original-target accuracy by $-12.3\ [-17.8, -6.8]$ points and increases agreement with the substituted class by $+8.8\ [2.5, 15.2]$ points. These results demonstrate sensitivity to the requested class within that panel, not an additional full-method benchmark or a finding-presence evaluation.

Table A23: Pixel-space copy screen. Counts flag SSIM $> .95$, pHash Hamming distance ≤ 5, or either criterion (Any). Threshold outcomes do not establish the absence of memorization outside the screened panel.

Dataset	Method	SSIM > 0.95	pHash ≤ 5	Any	Max SSIM
Fitzpatrick17k	TI	0	0	0	0.901
	DPO	0	0	0	0.902
	TI+LoRA	0	0	0	0.909
	CRAFT	0	0	0	0.910
CheXpert	TI	0	0	0	0.843
	DPO	0	0	0	0.834
	TI+LoRA	0	2	2	0.872
	CRAFT	0	2	2	0.883
BreakHis	TI	0	0	0	0.469
	DPO	0	0	0	0.678
	TI+LoRA	0	0	0	0.490
	CRAFT	0	0	0	**0.418**
ORIGA	TI	0	0	0	0.934
	DPO	0	0	0	0.886
	TI+LoRA	0	0	0	0.929
	CRAFT	0	0	0	**0.874**

J COPY AND PROXIMITY SCREENS

We screen generated images for proximity to real training images in both pixel space and DINOv2 feature space. These analyses address near-copying in the evaluated panels; they are not formal privacy or membership-inference tests.

J.1 PIXEL-SPACE NEAR-DUPLICATE DETECTION

SSIM and perceptual-hash distances are computed between generated samples and training images at thresholds of SSIM $> .95$ and pHash Hamming distance ≤ 5. The CheXpert screen covers 200 generations; screened sample counts for the remaining datasets are not reported. No images exceed the SSIM threshold (Table A23). The pHash criterion flags two TI+LoRA and two CRAFT generations in the CheXpert subset. These are screening results, not proof that every generated image is novel.

Table A24: DINOv2 nearest-neighbor distances. The signed gap is training-neighbor distance minus test-neighbor distance. These distances are computed for the generation panel used in the copy analysis, not for the matched-continuation checkpoints.

Dataset	Method	d_{NN}^{train}	d_{NN}^{test}	$\Delta = d_{NN}^{train} - d_{NN}^{test}$
Fitzpatrick17k	TI+LoRA	0.1267	0.1350	−0.0083
	CRAFT	0.1303	0.1370	−0.0067
CheXpert	TI+LoRA	0.0255	0.0255	0.0000
	CRAFT	0.0254	0.0259	−0.0005
BreakHis	TI+LoRA	0.1768	0.1757	0.0011
	CRAFT	0.0437	0.0439	−0.0002
ORIGA	TI+LoRA	0.1761	0.1872	−0.0111
	CRAFT	0.1382	0.1419	−0.0037

J.2 DINOv2 NEAREST-NEIGHBOR PROXIMITY

The feature-space analysis uses frozen DINOv2-base embeddings, measuring the cosine distance from each generated image to its nearest training image and nearest held-out test image. The table reports both distances and their signed difference. Because unequal real-set sizes and the choice of embedding affect nearest-neighbor distances, similar train/test proximity is treated as a descriptive check.

K EXTENDED QUALITATIVE RESULTS

Each modality gallery below includes real references and synthetic outputs labeled by method. A displayed real reference is not a pixel input to text-to-image generation; these illustrations are drawn from the main generation panels, not the matched-continuation evaluation. The selected examples do not estimate failure prevalence.



Figure A2: Fitzpatrick17k qualitative comparisons. Columns, from left to right: Real Image, TI, TI+LoRA, DPO, and CRAFT. Six example rows show conditions including pityriasis rosea, psoriasis, melanoma, folliculitis, and squamous cell carcinoma. These examples illustrate visual behavior.



Figure A3: CheXpert qualitative comparisons. Columns, from left to right: Real Image, TI, TI+LoRA, DPO, and CRAFT. These examples illustrate visual behavior.

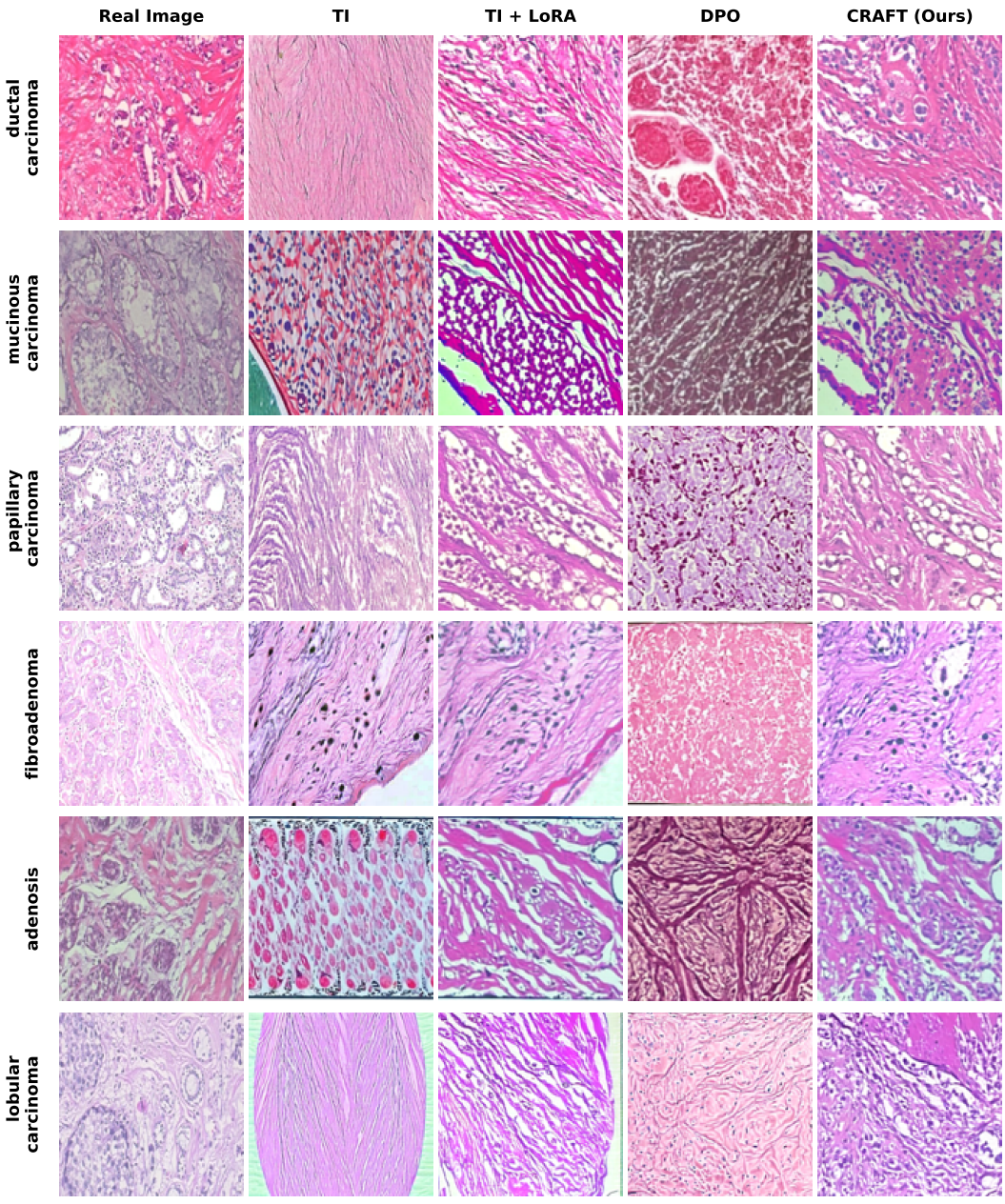

Figure A4: BreakHis qualitative comparisons. Columns, from left to right: Real Image, TI, TI+LoRA, DPO, and CRAFT. The six rows show ductal carcinoma, mucinous carcinoma, papillary carcinoma, fibroadenoma, adenosis, and lobular carcinoma labels. These examples illustrate visual behavior.

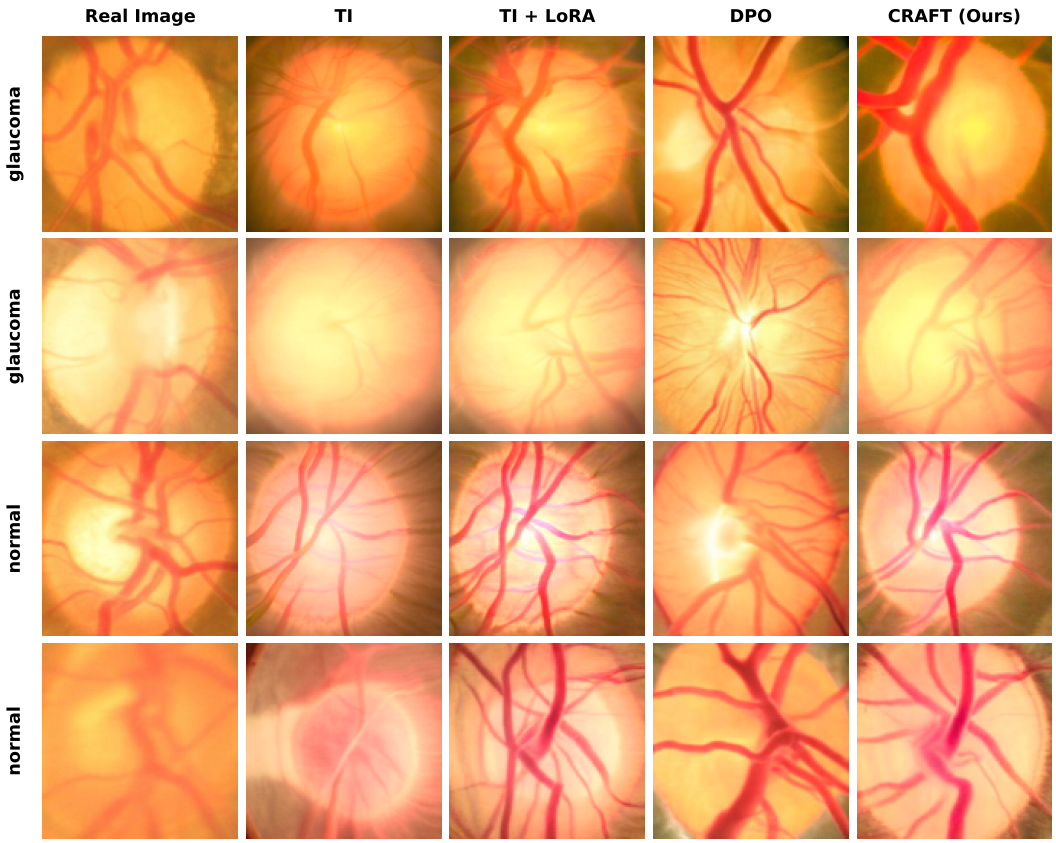

Figure A5: ORIGA qualitative comparisons. Columns, from left to right: Real Image, TI, TI+LoRA, DPO, and CRAFT. Two glaucoma and two normal examples are shown. These examples illustrate visual behavior.

K.1 CheXpert residual error cases



Figure A6: CheXpert error-case panel. Case A shows incomplete contour evidence for a requested cardiomegaly example, Case B shows absent characteristic focal consolidation for a requested pneumonia example, and Case C shows ambiguous basilar opacity for a requested pleural-effusion example.

The descriptions below summarize the intended qualitative interpretation of each example; they are not additional diagnostic annotations.

Case A: missing fine-grained evidence. The image shows a prominent cardiomediastinal silhouette but insufficient detail to delineate the cardiac contour confidently, with adjacent basilar/lateral opacity and indistinct borders. This example illustrates a broadly plausible image whose fine-grained evidence is incomplete.

Case B: mismatch with the requested diagnosis. The requested pneumonia example lacks a clear focal air-space consolidation, convincing lobar or segmental opacity, or visible air bronchograms. The image reproduces a chest-radiograph appearance without expressing the requested pattern.

Case C: ambiguous diagnostic content. The image shows basilar hazy opacity that can resemble pleural fluid, but without a definite meniscus sign or clearly assessable costophrenic angles. This example illustrates ambiguity between visual plausibility and complete finding expression.

These case illustrations do not form a systematic error audit.

Table A25: TI+LoRA validation-selection trajectory. Bold marks the checkpoint used for both baseline generation and CRAFT initialization. Each value averages five generation seeds; missing entries are unreported.

Dataset	e1	e2	e3	e4	e5	e6	e7	e8	e9	e10
Fitzpatrick17k	–	–	0.2221	**0.2398**	0.2325	0.2315	0.2254	0.2237	0.2241	0.2328
CheXpert	0.1410	0.1505	0.1531	**0.1544**	0.1421	0.1395	0.1379	0.1394	0.1388	0.1426
BreakHis	0.2113	0.2100	0.2183	0.2312	0.2369	0.2455	0.2493	**0.2511**	0.2410	0.2386
ORIGA	**0.3556**	0.3510	0.3414	0.3455	0.3476	0.3507	0.3414	0.3396	–	–

L CHECKPOINT SELECTION

For TI+LoRA and CRAFT checkpoint selection, we use

$$S_{\text{val}} = \frac{1}{3}\left(v_{\text{desc}} + v_{\text{check}} + a_{\text{target}}\right), \tag{7}$$

where the two similarities and the hard target-class prediction are computed under the validation SigLIP evaluation configuration. This equal-weight score is used for TI+LoRA and CRAFT selection; it is not a calibrated measure of medical correctness.

The selected TI+LoRA epochs are 4, 4, 8, and 1; these also serve as CRAFT's initialization. Selected CRAFT reward-stage epochs are 10, 5, 5, and 5. All later TI+LoRA checkpoints in the evaluated trajectory score below the selected maximum, confirming the stated selection rule for that trajectory but not implying that every possible reconstruction-only schedule would behave identically. The primary results compare the full CRAFT adaptation with the selected TI+LoRA baseline; the separate matched reconstruction-only result is reported in Section 4.2 and Appendix D, and its local e1–e5 labels do not replace the selected epochs in Table A1.

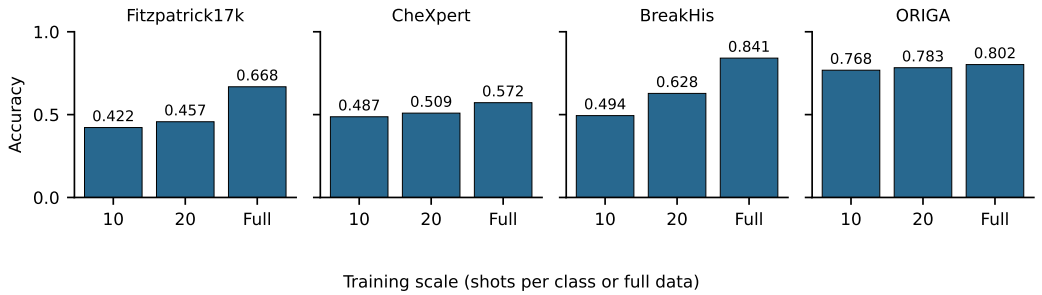

Figure A7: **Downstream accuracy of three-feedback CRAFT across training scales.** Each panel shows 10-shot, 20-shot, and full-data accuracy with 20% synthetic examples per classifier batch. Values are point estimates.

Table A26: Three-feedback CRAFT downstream accuracy (percent) by training scale. The 10-shot and 20-shot columns are per-class training budgets.

Dataset	10-shot	20-shot	Full
Fitzpatrick17k	42.2	45.7	66.8
CheXpert	48.7	50.9	57.2
BreakHis	49.4	62.8	84.1
ORIGA	76.8	78.3	80.2

Table A27: Retained full-data real-plus-synthetic augmentation accuracy (percent). CRAFT uses the three-feedback variant; differences are percentage-point contrasts of the reported means.

Dataset	TI+LoRA	CRAFT	Difference (pp)
Fitzpatrick17k	65.6	66.8	+1.2
CheXpert	56.6	57.2	+0.6
BreakHis	79.5	84.1	+4.6
ORIGA	79.1	80.2	+1.1

M DOWNSTREAM AUGMENTATION

The augmentation setting adds 20% synthetic examples per classifier batch and evaluates on real held-out data. This tests downstream classifier accuracy, unlike the synthetic-image classification experiment, which fits its evaluator on real images only. Table A26 gives the three-feedback CRAFT scores at each data scale.

The retained full-data comparisons are shown separately in Table A27. Full-data contrasts have no confidence intervals; the 10-shot and 20-shot results have no matched comparator values.

Table A28: Fitzpatrick17k held-out class counts.

Condition	Test	Synthetic per panel
Psoriasis	327	327
Squamous Cell Carcinoma	291	291
Lichen Planus	245	245
Basal Cell Carcinoma	234	234
Allergic Contact Dermatitis	215	215
Lupus Erythematosus	205	205
Sarcoidosis	175	175
Folliculitis	171	171
Scabies	169	169
Melanoma	131	131
Erythema Multiforme	118	118
Granuloma Annulare	105	105
Eczema	102	102
Pityriasis Rosea	97	97
Acne	91	91
Mycosis Fungoides	91	91
Actinic Keratosis	87	87
Prurigo Nodularis	85	85
Vitiligo	83	83
Keloid	78	78
Total	3100	3100

Table A29: CheXpert held-out class counts.

Condition	Test	Synthetic per panel
No Finding	124	124
Pleural Effusion	123	123
Cardiomegaly	123	123
Pneumonia	125	125
Total	495	495

Table A30: BreakHis held-out class counts.

Condition	Test	Synthetic per panel
Adenosis	57	57
Ductal Carcinoma	451	451
Fibroadenoma	130	130
Lobular Carcinoma	85	85
Mucinous Carcinoma	111	111
Papillary Carcinoma	71	71
Phyllodes Tumor	61	61
Tubular Adenoma	75	75
Total	1041	1041

N HELD-OUT LABEL COMPOSITION

The tables below report per-class test counts and the matching number of synthetic images generated per evaluation panel. Synthetic counts are per panel, before multiplication by generation seeds. Training and validation counts appear separately in Table A1.

Table A31: ORIGA held-out class counts.

Condition	Test	Synthetic per panel
Non-Glaucoma	145	145
Glaucoma	51	51
Total	196	196