Doctor Approved: Generating Medically Accurate Skin Disease Images through AI–Expert Feedback

Janet Wang [*1] **Yunbei Zhang** [*1] **Zhengming Ding** [1] **Jihun Hamm** [1]

Abstract

Paucity of medical data severely limits the generalizability of diagnostic ML models, as the full spectrum of disease variability can not be represented by a small clinical dataset. To address this, diffusion models (DMs) have been considered as a promising avenue for synthetic image generation and augmentation. However, they frequently produce *medically inaccurate* images, deteriorating the model performance. Expert domain knowledge is critical for synthesizing images that correctly encode clinical information, especially when data is scarce and quality outweighs quantity. Existing approaches for incorporating human feedback, such as reinforcement learning (RL) and Direct Preference Optimization (DPO), rely on robust reward functions or demand labor-intensive expert evaluations. Recent progress in Multimodal Large Language Models (MLLMs) reveals their strong visual reasoning capabilities, making them adept candidates as evaluators. In this work, we propose a novel framework, coined **MAGIC** (**M**edically **A**ccurate **G**eneration of **I**mages through AI-Expert **C**ollaboration), that synthesizes clinically accurate skin disease images for data augmentation. Our method creatively translates expert-defined criteria into actionable feedback for image synthesis of DMs, significantly improving clinical accuracy while reducing the direct human workload. Experiments demonstrate that our method greatly improves the clinical quality of synthesized skin disease images, with outputs aligning with dermatologist assessments. Additionally, augmenting training data with these synthesized images improves diagnostic accuracy by $+9.02\%$ on a challenging 20-condition skin disease classification task.

[*]Equal contribution [1]Tulane University. Correspondence to: Janet Wang <swang47@tulane.edu>, Jihun Hamm <jhamm3@tulane.edu>.

ICML 2025 Workshop on Models of Human Feedback for AI Alignment

1. Introduction

Recent advances in deep learning have made dermatological diagnosis increasingly accessible, offering significant potential for teledermatology in rural regions (Brinker et al., 2019; Esteva et al., 2017; Liu et al., 2020; Soenksen et al., 2021). However, privacy constraints and proprietary rights over skin images often lead to data scarcity, especially for rare conditions, making it difficult to capture the full complexity and variability of skin diseases for training robust diagnostic models. In response, various data augmentation strategies have been proposed—most straightforwardly, by aggregating open-source dermatological images (Aggarwal, 2019; Wang et al., 2024b). Yet, this approach does not guarantee access to high-quality samples of the precise clinical presentations needed, such as specific combinations of skin tones, body sites, and other lesion characteristics.

Image synthesis by Text-to-Image (T2I) Diffusion Models (DMs) (Dhariwal & Nichol, 2021) has emerged as a promising solution to enrich datasets under the guidance of prompts. Such controlled generation helps mitigate long-tail distributions, reduce biases against underrepresented groups, and improve model generalization—essential aspects of building reliable diagnostic systems (Ktena et al., 2024; Shin et al., 2023; Wang et al., 2024a). While the effectiveness of diffusion-based synthetic augmentation for common objects is debatable compared to retrieval-based methods, their value in the medical domain remains significant due to the proprietary nature of medical data and the general infeasibility of retrieval (Geng et al., 2024). T2I DMs have been employed to augment medical datasets across various imaging modalities (Ali et al., 2022; Huang et al., 2024; Khader et al., 2023; Pinaya et al., 2022). Previous works have also attempted to fine-tune DMs on skin disease images to enhance subsequent diagnostic model performance. However, these approaches typically involved end-to-end generation without expert participation during the training process, relegating expert assessment or filtering to a post-generation stage, rather than actively guiding the model to create clinically accurate images. (Akrout et al., 2023; Sagers et al., 2023; 2022; Wang et al., 2024a).

Aligning DMs via Reinforcement Learning from Human Feedback (RLHF) has been explored to adapt these mod-

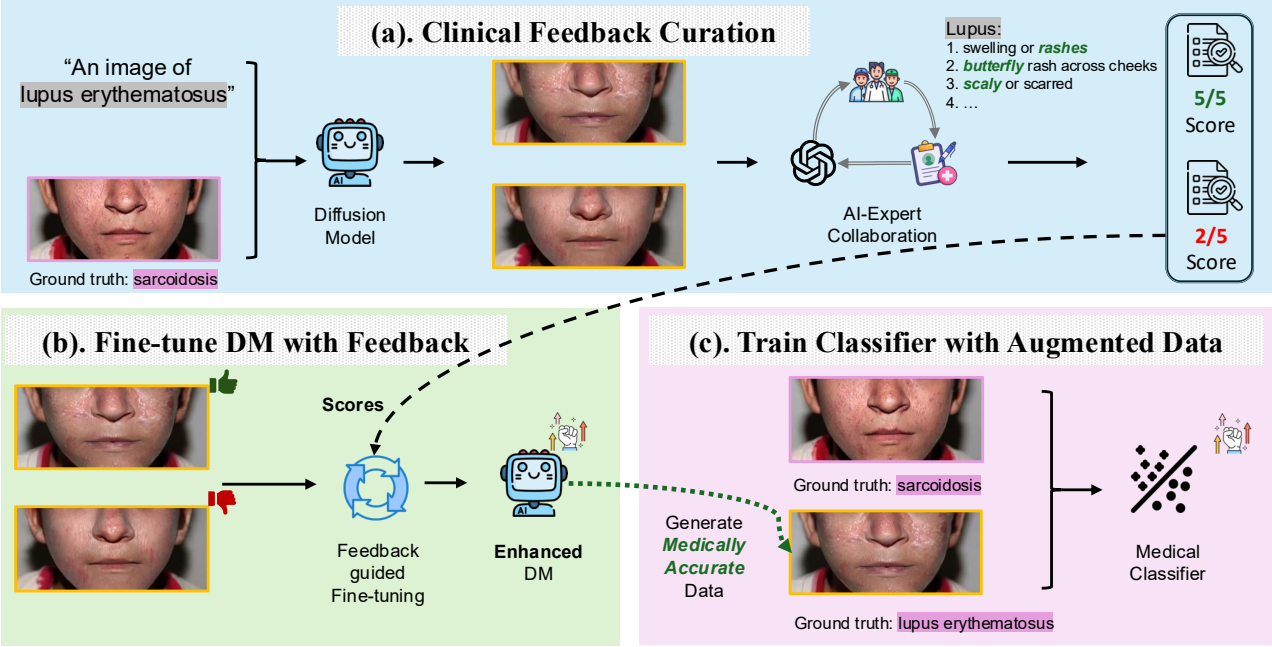


Figure 1. Illustration of our proposed **MAGIC**: (a) A preliminary fine-tuned diffusion model (DM) transforms a source image (e.g., sarcoidosis) to a target condition (e.g., lupus erythematosus); an MLLM then provides expert checklist-based feedback scores on the generated image pair. (b) This feedback guides the subsequent fine-tuning (e.g., RFT or DPO) of the DM. (c) The feedback-enhanced DM synthesizes medically accurate dermatological images for robust classifier training.

els and generate images that meet human preferences. In particular, (Lee et al., 2023c) proposes reward-weighted likelihood maximization to achieve alignment. Building on this, (Sun et al., 2023a) engages expert pathologists to assess sampled bone marrow images against a clinical plausibility checklist and train a reward function on binary feedback to emulate clinician assessments when fine-tuning a class-conditional DM. More recently, (Black et al., 2023a; Fan et al., 2023a) considers the denoising process as a multi-step Markov Decision Process (MDP) and adopts policy gradient optimization to fine-tune DMs based on human feedback. However, such methods still require reliable reward functions, whose training demands substantial computational resources and vast amounts of human-labeled feedback. To address these limitations, (Yang et al., 2023) proposes using Direct Preference Optimization (DPO) (Rafailov et al., 2023), which enables DM fine-tuning directly on preference data, bypassing the need for an explicit reward model and allowing iterative parameter updates based on human feedback at each timestep of the denoising process.

Inspired by recent advances in Reinforcement Learning from AI Feedback (RLAIF) (Lee et al., 2023a) and the strong visual reasoning capabilities of MLLMs, we propose **MAGIC** (**M**edically **A**ccurate **G**eneration of **I**mages through AI-Expert **C**ollaboration), a semi-automated framework that utilizes MLLMs for visual evaluation. In this framework, human experts are primarily required to: (1) craft, from credible sources, checklists that are easily verifiable by a MLLM, and (2) oversee the MLLM's feedback on synthetic images during the training of T2I DMs. By iteratively learning from the feedback enhanced with expert knowledge, MAGIC steers the T2I DMs toward more medically consistent generations. This approach highlights the potential of AI-expert collaboration, as MAGIC effectively leverages existing domain knowledge without labor-intensive annotation. Moreover, MAGIC incorporates an Image-to-Image (I2I) module within its training pipeline to initiate denoising from intermediate timesteps rather than pure Gaussian noise. This accelerates the sampling stage while ensuring factorized lesion transformations that do not deviate excessively from the real data distribution.

Through rigorous experiments, we demonstrate that our MAGIC framework performs effectively with both reward-based fine-tuning (RFT) and DPO, exhibiting particular strength with DPO. The MAGIC-DPO pipeline optimizes DMs to generate synthetic data that accurately represent each condition's unique visual features, with improvements observed as training progresses and more image-feedback pairs are used (Fig. 2). This is also validated by increasing dermatologist evaluation scores (Fig. 4(d)) and decreasing Fréchet Inception Distance (FID) scores (Fig. 4(c)), indicating improved clinical accuracy and fidelity. As a result, we also observe significant improvements in classification performance over baseline, highlighting MAGIC's potential

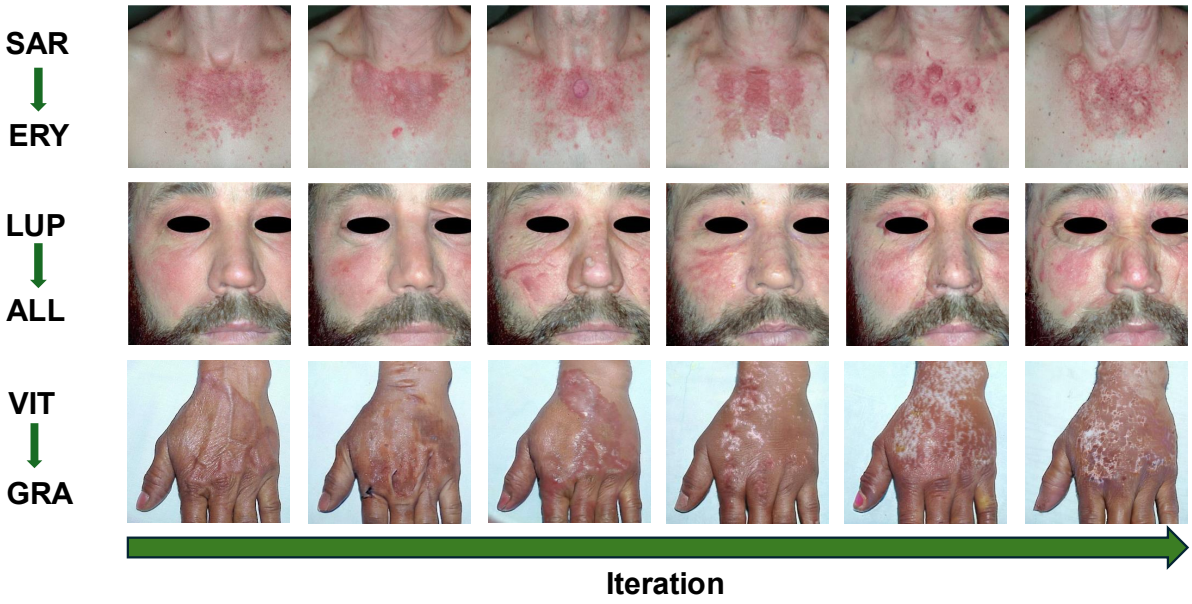


Figure 2. Evolution of synthetic skin conditions generated by MAGIC-DPO, illustrating its ability to learn unique visual features from feedback across training iterations. The Top Row demonstrates the model transforming Sarcoidosis (SAR) into Erythema Multiforme (ERY), learning features like "target" lesions with rings. The Middle Row demonstrates the model transforming Allergic Contact Dermatitis (ALL) into Lupus Erythematosus (LUP), developing a butterfly rash covering the cheeks. The Bottom Row demonstrates the model transforming Granuloma Annulare (GRA) into Vitiligo (VIT), evolving to show characteristic depigmented patches.

.

to advance AI dermatology. Overall, our main contributions are: **(i)** We propose **MAGIC**, a novel fine-tuning framework that integrates expert knowledge into DMs, enabling their subsequent fine-tuning with both DPO and RFT. The framework incorporates an I2I module to efficiently align the model for producing medically accurate images. **(ii)** Our framework employs an AI-Expert collaboration paradigm that offloads the work of visual evaluation to a powerful MLLM under minimal expert supervision, significantly reducing time and labor required from medical experts. **(iii)** MAGIC, particularly when combined with DPO (MAGIC-DPO), generates high-quality, clinically accurate images, achieving notable improvements in FID scores and classification performance. It yields a $+9.02\%$ boost in accuracy on a challenging 20-condition classification task and a $+13.89\%$ improvement in few-shot scenarios.

2. Related Works

DM-based Augmentation for Skin Disease Classification. Existing studies have explored diffusion models (DMs) to generate synthetic dermatological images for augmenting the training data of diagnostic models. Along this line, (Sagers et al., 2022) implemented a seed-based approach, sampling a small set of real images from the Fitzpatrick17k dataset (Groh et al., 2021) and generating synthetic data using the inpainting feature of OpenAI's DALL·E 2. Subsequently, (Sagers et al., 2023) leveraged Stable Diffusion's T2I pipeline, fine-tuned with Dreambooth, to produce images of specific disease conditions. Other related works (Akrout et al., 2023; Ktena et al., 2024) have similarly employed DM-based augmentation to enhance diagnostic accuracy and generalization on their internal skin disease datasets. Building on these advances, (Wang et al., 2024a) proposed a diffusion augmentation framework specifically targeting minority skin types. Their approach involved Textual Inversion (Gal et al., 2022b) and Low-Rank Adaptation (LoRA) (Hu et al., 2022) for fine-tuning, coupled with image-to-image generation for inference. This method enabled the creation of images depicting novel lesion concepts previously unseen by the DM. Their study revealed that images synthesized using this dual-guidance strategy improved the diagnostic performance of subsequent classifiers for minority skin types, even when reference data from these groups was absent from the training set. However, expert involvement in these previously proposed methods, if any, is typically confined to post-generation assessment or filtering, rather than actively guiding the image creation process.

Fine-tune Diffusion Models (DMs) with Feedback. Approaches to fine-tuning DMs with human feedback broadly fall into two categories: reward-based and preference-based. Reward-based methods (Black et al., 2023b; Fan & Lee, 2023; Fan et al., 2023b; Lee et al., 2023b; Xu et al., 2023) depend on robust reward models, the training of which typ-

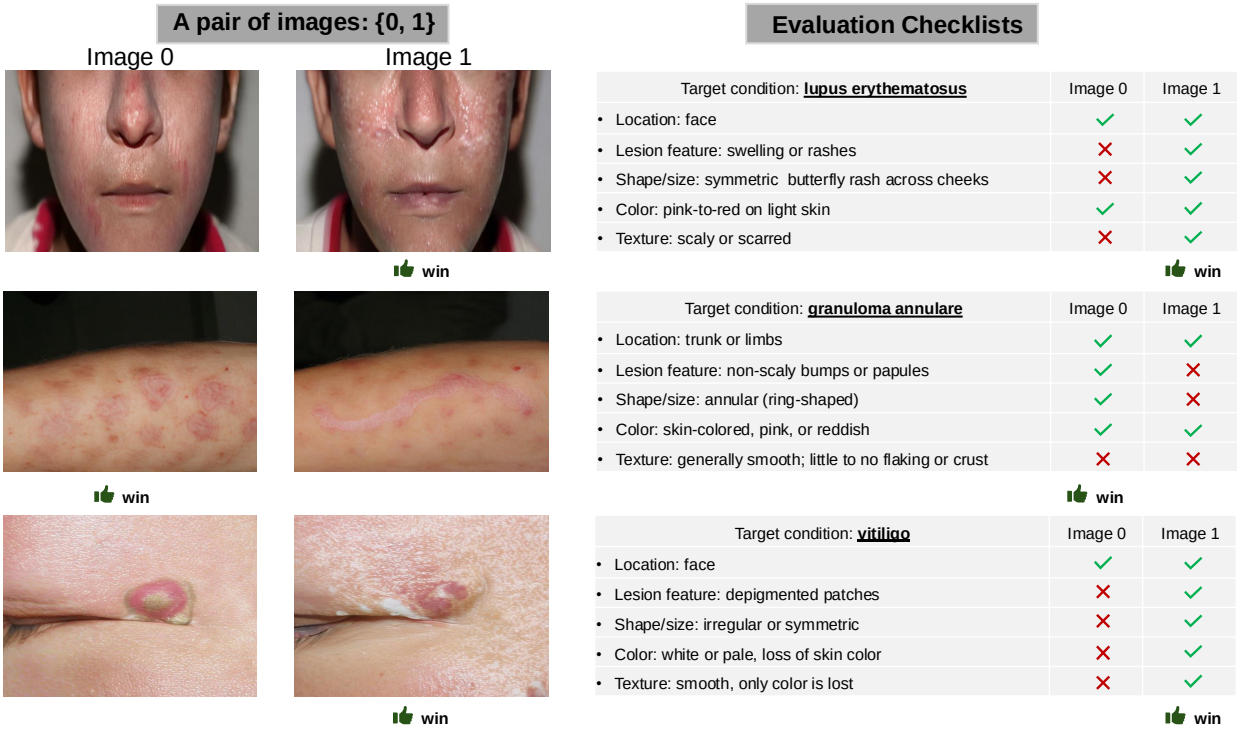


Figure 3. Illustration of the image assessment process by OpenAI's GPT-4o using condition-specific checklists for target skin conditions such as lupus erythematosus, granuloma annulare, and vitiligo. Each generated image in a pair is evaluated against five clinical criteria. The image with more satisfied criteria is considered the preferred sample in a comparison. Additional examples are in Appendix 6.

ically requires substantial datasets and extensive human evaluations. In the medical domain, for instance, (Sun et al., 2023a) leveraged reward-weighted maximization to synthesize plausible bone marrow images, by fine-tuning a class-conditional DM with a pathologist's feedback on synthetic images. In contrast, preference-based approaches aim to derive policies directly from preference data, thereby bypassing the need for explicit reward functions (Christiano et al., 2017; Dudík et al., 2015; Lee et al., 2023a). A key development in this area is Direct Preference Optimization (DPO) (Rafailov et al., 2023), originally proposed for fine-tuning language models directly using preferences. While DPO adaptations for diffusion models have primarily been tested for image-feedback alignment (Wallace et al., 2024; Yang et al., 2023), their application to medical image generation remains largely unexplored, especially for clinical images of skin diseases, which exhibit complex variations.

MLLMs-as-a-Judge. Collecting high-quality feedback has traditionally relied on human labelers, an approach that is both costly and difficult to scale. Recent research demonstrates that powerful proprietary MLLMs, such as GPT-4V and GPT-4o (OpenAI, 2024), can serve as effective generalist evaluators for vision-language tasks (Chen et al., 2024; Ge et al., 2023; Zhang et al., 2023). These models have proven particularly valuable in complex tasks requiring human-like judgment, including visual conversations and detailed image captioning, where MLLMs are often incorporated into evaluation benchmarks to assess model responses (Sun et al., 2023b; Zhang et al., 2024; Zheng et al., 2023). More recently, these models have shown capabilities in encoding clinical knowledge and acting as evaluators in medical reasoning (Singhal et al., 2023). Although employing MLLMs as collaborators in AI dermatology holds great potential to enhance the reliability of diagnostic models, the optimal paradigm for their collaboration with medical experts still remains underexplored.

3. Method

3.1. Preliminaries

Diffusion Models (DMs). DMs are designed to learn the probability distribution $p(x)$ by reversing a Markovian forward process, denoted as $q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1})$, which incrementally introduces noise into the images. The reversal, a denoising process, is implemented through a neural network tasked with predicting either the mean of $\boldsymbol{x}_{t-1}$ or the noise $\boldsymbol{\epsilon}_{t-1}$ from the forward process. In our approach, we utilize a network $\boldsymbol{\mu}_\theta(\boldsymbol{x}_t; t)$ to predict the mean of $\boldsymbol{x}_{t-1}$, rather than the added noise. We employ the Mean Squared Error (MSE) as a performance metric, defining the objective function of

Table 1. Performance of ResNet18-based classifiers trained on real and synthetic data.

Method	Acc	F1	Prec	Rec
Real	29.31	28.73	28.61	29.13
+ T2I	25.57	24.63	24.44	25.16
	-3.74	-4.11	-4.17	-3.97
+ I2I	31.45	31.09	31.03	31.49
	+2.14	+2.35	+2.42	+2.36
+ MAGIC	33.49	30.40	29.12	29.67
(RFT)	+4.18	+1.67	+0.51	+0.54
+ MAGIC	**38.33**	**37.01**	**38.41**	**36.06**
(DPO)	**+9.02**	**+8.28**	**+9.80**	**+6.94**

Table 2. Performance of DINOv2-based classifiers trained on real and synthetic data.

Method	Acc	F1	Prec	Rec
Real	49.89	49.43	50.03	49.31
+ T2I	47.73	47.26	47.51	47.43
	-2.16	-2.17	-2.52	-1.88
+ I2I	50.71	50.17	51.04	49.89
	+0.82	+0.74	+1.01	+0.58
+ MAGIC	51.16	52.66	52.17	52.69
(RFT)	+1.27	+3.23	+2.14	+3.38
+ MAGIC	**55.01**	**54.05**	**54.96**	**53.70**
(DPO)	**+5.12**	**+4.62**	**+4.93**	**+4.39**

our network as follows:

$$\mathcal{L}_{\mathrm{DM}} = \mathbb{E}_{t\sim[1,T],\boldsymbol{x}_0\sim p(\boldsymbol{x}_0),\boldsymbol{x}_t\sim q(\boldsymbol{x}_t|\boldsymbol{x}_0)}\left[\|\tilde{\boldsymbol{\mu}}(\boldsymbol{x}_0,\boldsymbol{x}_t) - \boldsymbol{\mu}_\theta(\boldsymbol{x}_t,t)\|^2\right], \tag{1}$$

where $\tilde{\boldsymbol{\mu}}_\theta(\boldsymbol{x}_t,\boldsymbol{x}_0)$ represents the posterior mean of the forward process.

In conditional generative modeling, diffusion models are adapted to learn the conditional distribution $p(x|\boldsymbol{c})$, where $\boldsymbol{c}$ represents conditioning information, such as image categories or captions. This adaptation involves augmenting the denoising network with additional input, $\boldsymbol{c}$, resulting in $\boldsymbol{\mu}_\theta(\boldsymbol{x}_t,t;\boldsymbol{c})$. To generate a sample from the learned distribution $p_\theta(x|\boldsymbol{c})$, we initiate the process by drawing a sample $\boldsymbol{x}_T \sim \mathcal{N}(\mathbf{0},\mathbf{I})$, which is then progressively denoised through iterative application of $\boldsymbol{\epsilon}_\theta$, based on specific samplers adopted (Ho et al., 2020). The reverse process is modeled as:

$$p_\theta(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t,\boldsymbol{c}) = \mathcal{N}\left(\boldsymbol{x}_{t-1};\boldsymbol{\mu}_\theta(\boldsymbol{x}_t,\boldsymbol{c},t),\sigma_t^2\mathbf{I}\right). \tag{2}$$

In our skin disease image generation framework, we leverage the I2I pipeline of Stable Diffusion (Rombach et al., 2022) to transform lesion features while preserving body part information in the image. This strategy effectively reduces semantic distortion during generation and ensures factorized translation of lesions, thereby enhancing medical plausibility. Specifically, we start with a real input dermatological image $\boldsymbol{x}_0$ (e.g., sarcoidosis), add partial noise to it, and transform it into a different target skin condition (e.g., lupus erythematosus), by denoising this partly noised images. And the denoising process is governed by $\boldsymbol{\mu}_\theta$ and denoise strength parameter γ.

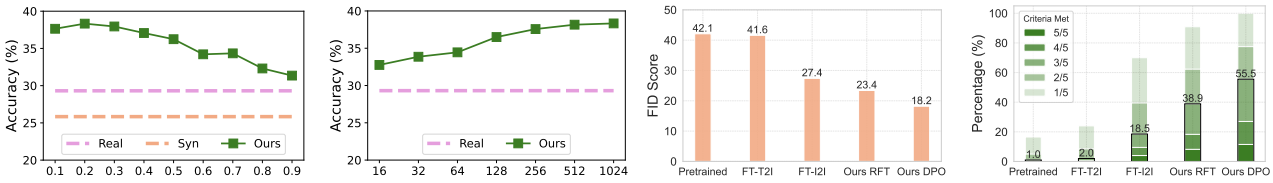
Multi-Step MDP Formulation. We formulate the diffusion model's denoising process as a multi-step Markov Decision Process (MDP), following (Black et al., 2023b; Sutton et al., 1998). In our model, the state $s \in \mathcal{S}$ includes the current denoising time step, denoised image data and prompt. The action space $\mathcal{A}$ includes possible image transformations at each time step. The state transition function $P(s'|s,a)$ describes the image evolution, and the reward function $r(s,a)$ assigns values based on the image quality at each time step, aiming to maximize cumulative returns $\mathcal{J}(\pi) = \mathbb{E}_\tau[\sum_{t=0}^{T-1} r(s_t,a_t)]$. The MDP is formulated as

$$\begin{aligned}
&\mathbf{s}_t \triangleq (\boldsymbol{c},t,\boldsymbol{x}_{T-t}), \quad P(\mathbf{s}_{t+1} \mid \mathbf{s}_t,\mathbf{a}_t) \triangleq (\delta_{\boldsymbol{c}},\delta_{t+1},\delta_{\boldsymbol{x}_{T-1-t}});\\
&\mathbf{a}_t \triangleq \boldsymbol{x}_{T-1-t}, \quad \pi(\mathbf{a}_t \mid \mathbf{s}_t) \triangleq p_\theta(\boldsymbol{x}_{T-1-t} \mid \boldsymbol{c},t,\boldsymbol{x}_{T-t});\\
&\rho_0(\mathbf{s}_0) \triangleq (p(\boldsymbol{c}),\delta_0,\mathcal{N}(\mathbf{0},\mathbf{I}));\\
&r(\mathbf{s}_t,\mathbf{a}_t) \triangleq r((\boldsymbol{c},t,\boldsymbol{x}_{T-t}),\boldsymbol{x}_{T-t-1}),
\end{aligned} \tag{3}$$

where δ_x represents the Dirac delta distribution, and T denotes the maximize denoising timesteps.

3.2. Preliminary Diffusion Models Fine-tuning

Previous studies have shown that off-the-shelf diffusion models struggle to represent skin lesion concepts, making preliminary fine-tuning necessary before aligning with expert feedback. Following (Wang et al., 2024a), we employ Latent Diffusion Models (LDMs) (Rombach et al., 2022), which operate in autoencoder latent space to reduce computational demands while maintaining generation quality. For simplicity, we abuse notation and use $\boldsymbol{x}$ to represent the latent input to the diffusion process rather than the original image. Our framework utilizes Textual Inversion (Gal et al., 2022a) to derive unique embeddings that capture the semantics of each condition extracted from training data. Each image is paired with a descriptive string containing placeholders (e.g., 'an image of $\{S_*\}$') as input. The optimal embedding v_*, encapsulating the lesion concept S_*, is then obtained by minimizing reconstruction loss while keeping the LDM fixed. To enhance the efficiency of the LDM fine-tuning process, we employ LoRA (Hu et al., 2022), adapting the model with the discovered tokens from Textual Inversion. This approach maintains the pre-trained model weights while introducing only two compact matrices A and B (where $A \in \mathbb{R}^{n\times r}, B \in \mathbb{R}^{r\times n}$). These matrices are embedded within the attention layers, enabling the detailed capture of skin lesion characteristics previously unrepresented in the initial model, aligned with the learned target embedding v_*.



((a)) Ratio ρ ((b)) #Feedback ((c)) FID Score ((d)) Synthetic Data Eval

Figure 4. Experimental results showing (a) the impact of ratio ρ, (b) feedback volume on accuracy, (c) FID score comparison across different methods, and (d) evaluation results on synthetic data showing the percentage of criteria met. Our method consistently outperforms baseline methods in most metrics, achieving lower FID scores and higher criteria satisfaction rates.

3.3. Expert Feedback Curation

While diffusion models can synthesize visually realistic medical images, their clinical validity often remains questionable (Sun et al., 2023a). Incorporating medical expertise is therefore crucial for guiding these models to generate medically accurate images. To provide this clinical guidance, our framework leverages structured feedback derived from checklists that are designed by an experienced dermatologist. These checklists evaluates five distinct aspects of each condition: `[Location, Lesion Type, Shape/Size, Color, Texture]` (see Appendix B for complete details). Assessment against these aspects yields a binary outcome (e.g., satisfied/not satisfied) for each criterion. To automate this evaluation, we instructed an MLLM to analyze each synthesized image based on the target condition's checklist and return a 5-dimensional binary score list, where each dimension corresponds to a criterion's satisfaction (see Appendix C for instruction details). To accommodate both reward-based and preference-based alignment strategies, we generate a pair of images from each text prompt and submit each single image to the MLLM for this assessment. Thus, the MLLM's score list for each image in a pair individually stands as a sample for RFT, while the pair of score lists can be used for DPO. Examples of this MLLM assessment using OpenAI's GPT-4o are illustrated in Fig. 3, showing yielded score lists such as `[1,0,0,1,0]` and `[1,1,1,1,1]` for a given pair. Ultimately, each 5-dimensional MLLM-generated score list is aggregated into an overall binary score (e.g., 0 for negative example, 1 for positive example) using a predefined algorithm (detailed in Appendix A.2). This semi-automated pipeline allows us to significantly accelerate the curation of expert feedback. Notably, only synthetic images are sent to API services and no real patient images are processed by the MLLM, to preserve privacy.

3.4. Finetuning with Expert Feedback

After collecting pairwise preferences, we explore two complementary ways to integrate them into optimizing the diffusion model parameters θ.

Reward-model guided fine-tuning (RFT) Let $\mathcal{R}_\phi : \mathbb{R}^{H\times W\times 3} \times \mathcal{C} \to \mathbb{R}_{\geq 0}$ be a learned scalar that predicts the likelihood an image x conditioned on class c satisfies every checklist item. We follow (Lee et al., 2023c; Sun et al., 2023a) and mix real and synthetic images when training $\mathcal{R}_\phi$ with an MSE loss. Formally, with feedback labels $y \in \{0, 1\}$ we minimize $\mathcal{L}_{\text{RM}}(\phi) = \sum_{(x,c,y)}(y - \mathcal{R}_\phi(x, c))^2$. After fitting ϕ, we refine θ by maximising the expected reward-weighted log-probability of the action sequence generated along each denoising trajectory $\sigma = \{(s_t, a_t)\}_{t=0}^{T-1}$:

$$\mathcal{L}_{\text{RFT}}(\theta) = \mathbb{E}_{(x,c)\sim\mathcal{D}_\text{s}}\Big[-\mathcal{R}_\phi(x, c)\sum_{t=0}^{T-1}\log \pi_\theta(a_t \mid s_t)\Big] + \beta_r\, \mathbb{E}_{(x,c)\sim\mathcal{D}_\text{r}}\Big[-\sum_{t=0}^{T-1}\log \pi_\theta(a_t \mid s_t)\Big], \tag{4}$$

where $\mathcal{D}_\text{s}$ and $\mathcal{D}_\text{r}$ denote synthetic and real image pools, respectively, and β_r balances fidelity to expert feedback against faithfulness to the original data distribution.

Direct Preference Optimization (DPO) Given a pair of trajectories (σ^w, σ^l) that yield a *winner* image x^w and a *loser* image x^l under expert comparison, DPO increases the likelihood of every action a_i^w on the winning branch while decreasing the likelihood of the corresponding a_i^l on the losing branch. Similar to reinforcement learning methods (Brown & Sandholm, 2019; Silver et al., 2016; 2017), rewards are assigned by $\forall s_t, a_t \in \sigma, r(s_t, a_t) = 1$ for winning the game and $\forall t \in \sigma, r(s_t, a_t) = -1$ for losing the game. Following (Yang et al., 2023), we also assume that if the final image is preferred, then any state-action pair in its generation path is superior to the corresponding pair in the non-preferred path. To maximize learning from each generation process under this assumption, we construct $t' = \gamma T$ sub-segments that allow the model to learn from intermediate states

$$\mathcal{L}^i_{\text{DPO}}(\theta) = -E_{(s_i,\sigma_w,\sigma_l)}[\log \rho(\beta \log \tfrac{\pi_\theta(a_i^w|s_i^w)}{\pi_{\text{ref}}(a_i^w|s_i^w)} - \beta \log \tfrac{\pi_\theta(a_i^l|s_i^l)}{\pi_{\text{ref}}(a_i^l|s_i^l)})], \tag{5}$$

where $i \in [0, t' - 1]$, effectively increasing data utilization by a factor of t'.

3.5. Synthetic Augmentation for Classifier Training

After fine-tuning a DM with expert-enhanced feedback, we leverage the model to synthesize images for dataset augmentation, primarily through an image-to-image translation approach. For any given real sample x with label y, we first randomly select a different target label y' from the label set. We then use the text prompt "an image of $\{y'\}$"—incorporating the specific text embedding for y' learned via Textual Inversion—to guide the DM in generating a new image x'. This process is designed so that x' preserves most of the anatomical context of the original sample x while primarily displaying the lesion semantics of the target label y', thereby achieving a factorized transformation. This I2I generation strategy offers a key benefit: it helps mitigate the risk of the classifier learning spurious correlations by preventing it from associating lesions with specific body locations, encouraging a focus on the intrinsic characteristics of the skin lesions. During the subsequent

classifier training phase, we intentionally control the influence of synthetic data using a ratio parameter $\rho \in (0, 1)$, which determines the percentage of synthetic images added to each training batch. While our method aims to generate medically accurate images, potential domain shifts between real and synthetic data remain an important consideration. Indeed, our experiments indicate that varying the proportion of synthetic data can significantly affect classifier performance on real test data (see Fig. 4(a)).

4. Experiments

Dataset. Following prior work (Wang et al., 2024a), we use the Fitzpatrick17k dataset to evaluate our synthetic augmentation pipeline (Groh et al., 2021). Fitzpatrick17k contains clinical photos of 114 skin conditions, each annotated with a condition label and a Fitzpatrick Skin Type (FST). Although there are other datasets of clinical photos (e.g., SCIN (Ward et al., 2024) and DDI(Daneshjou et al., 2022)), they are primarily collected within the United States and feature lighter skin tones. Fitzpatrick17k encompasses a wider range of skin types, making it particularly suitable for evaluating generalizable diagnostic approaches. For our experiments, we focus on a subset of the Fitzpatrick17k dataset consisting of 20 skin conditions. We chose these based on two criteria: (1) they present the largest class sizes in the dataset, and (2) they have well-established descriptions available from reputable clinical sources (e.g., Mayo Clinic, Cleveland Clinic), which allowed dermatologists to craft reliable diagnostic checklists of key visual features for these diseases. These checklists, verified by clinicians, distill essential visual cues for each condition, detailed in the Appendix B. The distribution of the selected classes is provided in the Appendix A.

Models and Baselines. We utilize Stable Diffusion v2-1 (Rombach et al., 2022) for image generation. For classification tasks, we employ ResNet18 (He et al., 2016) and DINOv2 (Oquab et al., 2023) as backbone architectures. For medical image generation, we evaluate four different methods: (1) diffusion model fine-tuned with Textual Inversion and LoRA, generating images via text-to-image (+ T2I); (2) the same fine-tuned model but generating via image-to-image (+ I2I); and (3/4) our proposed MAGIC (RFT/DPO) with expert feedback. We assess synthetic image quality using both FID score and human evaluation. For classification experiments, we first establish a baseline by training a classifier solely on real data. We then generate an equivalent number of synthetic images using each generation method (excluding the off-the-shelf DM due to its lack of domain-specific knowledge (Wang et al., 2024a)), and train classifiers on combined real and synthetic datasets. Implementation details are provided in Appendix A.

Implementation Details. To adapt the model to skin lesion concepts, our preliminary fine-tuning process proceeds in two stages: (i) We learn unique disease-related tokens by updating the text encoder via Textual Inversion (Gal et al., 2022b), thereby introducing new vocabulary specific to each condition; and (ii) we tie the newly learned tokens to fine-grained visual cues within the images by updating the UNet parameters via LoRA (Hu et al., 2022). Further details on prompts and hyperparameters can be found in the Appendix A.

For training with expert feedback, all experiments share a unified *sampling–feedback* pipeline. For each mini-batch of image–prompt pairs drawn from the real set, the current diffusion model generates two synthetic variants via the Stable-Diffusion image-to-image path, intentionally targeting skin-disease classes that differ from the originals to maximise diversity. Each synthetic image is then scored with the condition-specific checklists (Appendix B), which we submit to GPT-4o (OpenAI, 2024). The API returns binary vectors indicating whether each criterion is met; if the lesion is deemed invalid, an all-zero vector is assigned. From every pair of vectors we derive a *winner–loser* label and store the associated latents, timesteps, and prompt embeddings. We subsequently branch into two finetuning regimes: (i) in the *reward-model route* we fit a scalar network $\mathcal{R}_\phi$ to these binary outcomes and update θ by the reward-weighted likelihood of Eq. (4); (ii) in the *DPO route* we treat each preference tuple as in (Yang et al., 2023) and optimize the multi-segment loss of Eq. (5). Both routes draw from the same pool of feedback pairs, subsequent comparisons isolate the effect of the finetuning algorithm itself. Examples are visualised in Fig. 3.

For classifier training, we randomly split the dataset into training and hold-out sets at a 50/50 ratio, resulting in 3,100 training and 3,100 test images. The baseline classifier is trained exclusively on this 3,100-image training set. During inference, we apply the same hyperparameters used in the DPO sampling stage when generating synthetic images with the DPO fine-tuned model. We generate one synthetic image for each real image, intentionally assigning a target label that differs from the real image's original label while corresponding to the same body region. Following established practices, we combine synthetic and real images to optimize performance, maintaining a fixed ratio of synthetic to real examples in each training batch. All experiments are conducted *five* rounds on RTX 6000 Ada GPUs. Our experimental evaluation encompasses both CNN-based and Transformer-based classifier architectures, fine-tuned according to protocols outlined in previous work (Wang et al., 2024a).

5. Analysis

5.1. Experimental Results

Classification results. We comprehensively evaluate synthetic image quality by its impact on downstream classification using ResNet18 and DINOv2 architectures (Tables 1 and 2). Our MAGIC framework markedly enhances performance across both models compared to baselines. Standard fine-tuned Text-to-Image (T2I) generation degrades ResNet18 accuracy by -3.74% and DINOv2 by -2.16%, while the fine-tuned Image-to-Image (I2I) approach offers modest gains, increasing ResNet18 accuracy by $+2.14\%$ and DINOv2 by $+0.82\%$. The feedback integrated via our MAGIC framework proves beneficial for both Reward-model guided Fine-Tuning (RFT) and Direct Preference Optimization (DPO) strategies. Specifically, MAGIC-RFT improved accuracy over the real data baseline by $+4.18\%$ for ResNet18 and $+2.21\%$ for DINOv2. MAGIC-DPO demonstrated even more substantial gains, boosting accuracy by $+9.02\%$ for ResNet18 (from 29.31% to 38.33%) and by $+5.12\%$ for DINOv2 (from 49.89% to 55.01%), with similar improvements in F1, precision, and recall. We further validate the MAGIC framework on an additional medical dataset, SCIN, with results detailed in Appendix D.3.

The DPO approach within the MAGIC framework (MAGIC-DPO) shows particular strength. Its advantage may stem from directly optimizing for preference alignment without an intermediate reward model. This can be more robust and generalize better, proving especially advantageous when the number of feedback pairs is limited, as is common in specialized medical domains, thus sidestepping potential instabilities in reward modeling. The quality of expert guidance remains crucial for generating synthetic images that are not only visually plausible but also encode clinically relevant di-

Table 3. Performance of DINOv2-based classifiers in *few-shot* setting.

Method	Acc	F1	Prec	Rec
Real (all)	49.89	49.43	50.03	49.31
Real (310)	26.45	19.50	21.86	20.19
+ T2I	25.58	19.58	20.87	19.27
	-2.17	+0.08	-0.99	-0.92
+ I2I	30.10	27.26	28.07	27.00
	+3.65	+7.76	+6.21	+6.81
+ MAGIC	**37.39**	**36.90**	**37.95**	**36.94**
(DPO)	**+10.94**	**+17.40**	**+16.09**	**+16.75**
+ MAGIC-A	**40.34**	**39.43**	**42.20**	**38.77**
(DPO)	**+13.89**	**+19.93**	**+20.34**	**+18.58**

Table 4. Performance of classifiers across different backbones and Coarse/Structured checklists.

Model	Method	Acc	F1	Prec	Rec
RN18	Real	29.31	28.73	28.61	29.13
	+ MAGIC	32.83	30.58	29.75	31.18
	Coarse	+3.52	+1.85	+1.14	+2.05
	+ MAGIC	**38.33**	**37.01**	**38.41**	**36.06**
	Structured	**+9.02**	**+8.28**	**+9.80**	**+6.94**
DINO	Real	49.89	49.43	50.03	49.31
	+ MAGIC	51.16	52.66	52.17	52.69
	Coarse	+1.27	+3.23	+2.14	+3.38
	+ MAGIC	**55.01**	**54.05**	**54.96**	**53.70**
	Structured	**+5.12**	**+4.62**	**+4.93**	**+4.39**

agnostic features. This enhanced alignment is reflected across our evaluations, including improved qualitative outputs (Fig. 2), FID scores (Fig. 4(c)), and expert preference measures (Fig. 4(d)).

Expert evaluation on generated images. To further assess the quality and medical plausibility of images generated by our methods, we engaged medical experts to evaluate the synthetic data based on our specific checklist criteria. For each method, we sampled 10 images per skin condition, resulting in 200 images per method. Each image was evaluated against 5 criteria, with binary outcomes (satisfied/not satisfied). Fig. 4(d) summarizes these evaluation results, displaying the percentage of images meeting different numbers of criteria (with details in Appendix B). The results show that images from the pretrained diffusion model rarely satisfied more than one criterion, and none met more than three. Standard Text-to-Image (T2I) generation showed minimal improvement, with only 2.0% of images meeting 3 or more criteria and only a single image meeting 4 criteria overall. Fine-tuned Image-to-Image (I2I) generation yielded better outputs, with 18.5% of its images meeting 3 or more criteria, underscoring I2I's greater suitability for medical tasks. Our MAGIC framework significantly builds on this; MAGIC-RFT (Ours RFT) further increased the proportion of high-quality images, with 38.9% meeting 3 or more criteria. Notably, MAGIC-DPO (Ours DPO) demonstrated the best performance, with 55.5% of its images satisfying 3 or more criteria. This substantial improvement over both fine-tuned I2I and MAGIC-RFT correlates directly with the observed enhancements in classifier performance.

Few-shot Setting. We further evaluate our framework in a few-shot setting where only a small number of labeled data are available. This scenario better reflects real-world conditions, as collecting and labeling medical data is costly. We simulate this setting by randomly selecting 10% of the DINOv2 training set (310 images) while keeping the test set fixed. We fine-tuned the diffusion model on these 310 real images using our DPO-based approach (MAGIC-DPO) and other baselines. As shown in Table 3, MAGIC-DPO improves classifier accuracy by $+10.94\%$ (from 26.45% to 37.39%) compared to training with only the limited real data, significantly outperforming standard T2I and I2I augmentation baselines in this data-scarce context. Moreover, in practical scenarios, unlabeled medical data from the same distribution may be available even when expert labeling is cost-prohibitive. Our MAGIC framework can effectively utilize such *unlabeled* data; specifically, during the DPO fine-tuning stage, unlabeled data is processed by the diffusion model with randomly selected skin conditions, and feedback is evaluated solely based on the target condition. This makes our framework well-suited for leveraging unlabeled data. This augmented approach, termed **MAGIC-A** (also DPO-based), demonstrates that by incorporating an equal number of unlabeled samples (310), we can further improve accuracy by an additional 2.95% over MAGIC-DPO, reaching 40.34% accuracy.

5.2. Abaltion Study

Effect of Checklist Quality. We investigate the impact of checklist detail level on the MAGIC framework's efficacy in DPO training, by comparing two types of expert-designed checklists: a "Coarse" version using single-sentence descriptions for each condition, and a more detailed, "Structured" version (as used throughout the main paper and detailed in Appendix B). Table 4 shows that the quality of the checklist is crucial to feedback quality. For the ResNet18 (RN18), augmenting with MAGIC-DPO using Coarse checklists improved accuracy by +3.52% over the real data baseline (from 29.31% to 32.83%), whereas Structured checklists led to a much larger gain of +9.02% (to 38.33%). A similar trend was observed with the DINOv2 (DINO): Coarse checklists yielded a +1.27% accuracy improvement (from 49.89% to 51.16%), while Structured checklists achieved a +5.12% boost (to 55.01%). These results underscore that more detailed and well-structured expert guidance in the checklists significantly enhances the quality of synthetic images and subsequent classifier performance.

Effect of feedback volume. In addition to feedback quality, we also investigate how the quantity of feedback influences image quality and classifier performance. As DPO training progresses, more image pairs are used, providing additional feedback to guide the diffusion model. We visually demonstrate the evolution of generated images across epochs in Fig. 2. Additionally, we also train classifiers using synthetic data that is generated from different training stages. Results in Fig. 4(b) show that accuracy consistently improves as DPO training accumulates more feedback, with performance stabilizing after receiving feedback from approximately 512 image pairs. Based on these findings, we fix the feedback volume at 1024 image pairs for all our experiments.

Effect of the ratio ρ of synthetic data. We investigate how the ratio ρ of synthetic data affects classifier performance. Initial experiments with purely synthetic data failed to achieve performance comparable to real data-trained classifiers. It's expected that, without the guidance of real data, classifiers tend to overfit to the synthetic data distribution. We therefore systematically controlled the percentage of synthetic data used in each training batch across different values of ρ, while keeping the total volume of synthetic data constant. As shown in Fig. 4(a), performance improves when ρ is less than 0.5 (when synthetic data constitutes less than half of the training data). The performance remains stable when $\rho \in [0.1, 0.3]$. We adopt $\rho = 0.2$ for all our experiments.

6. Conclusion

In this work, we addressed the critical challenge of generating medically accurate synthetic images for augmenting scarce dermatological datasets, a key limitation in developing robust diagnostic models. We introduced MAGIC, a novel semi-automated framework designed to refine Diffusion Models by effectively integrating expert-enhanced clinical knowledge. Our approach uniquely leverages the visual reasoning capabilities of MLLMs to interpret and apply expert-defined checklists, thereby guiding DMs to produce images with high clinical fidelity while significantly reducing the burden on human experts. Our experiments demonstrate that MAGIC, substantially improves the clinical quality of synthesized skin disease images, as validated by both quantitative metrics like FID scores and qualitative assessments by dermatologists. Furthermore, augmenting training data with images generated by MAGIC led to significant enhancements in downstream classification accuracy for skin diseases, even in few-shot scenarios. These results underscore the efficacy of our AI-Expert collaboration paradigm in translating nuanced clinical criteria into actionable feedback for generative models. The findings highlight the considerable potential of combining expert-verified clinical knowledge with automated MLLM-based evaluations to create more reliable, scalable, and clinically valid synthetic data augmentation pipelines. This work paves the way for more robust AI-driven diagnostic tools in dermatology and other medical imaging domains where data scarcity and the need for high-fidelity synthetic data are paramount.

References

Aggarwal, s. L. P. Data augmentation in dermatology image recognition using machine learning. *Skin Research and Technology*, 25(6):815–820, 2019.

Akrout, M., Gyepesi, B., Holló, P., Poór, A., Kincső, B., Solis, S., Cirone, K., Kawahara, J., Slade, D., Abid, L., et al. Diffusion-based data augmentation for skin disease classification: Impact across original medical datasets to fully synthetic images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 99–109. Springer, 2023.

Ali, H., Murad, S., and Shah, Z. Spot the fake lungs: Generating synthetic medical images using neural diffusion models. In *Irish Conference on Artificial Intelligence and Cognitive Science*, pp. 32–39. Springer, 2022.

Black, K., Janner, M., Du, Y., Kostrikov, I., and Levine, S. Training diffusion models with reinforcement learning. 2023a.

Black, K., Janner, M., Du, Y., Kostrikov, I., and Levine, S. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2023b.

Brinker, T., Hekler, A., Enk, A., Klode, J., Hauschild, A., Berking, C., Schilling, B., Haferkamp, S., Utikal, J., Kalle, C., Fröhling, S., and Weichenthal, M. A convolutional neural network trained with dermoscopic images performed on par with 145 dermatologists in a clinical melanoma image classification task. *European Journal of Cancer*, 111:148–154, 03 2019. doi: 10.1016/j.ejca.2019.02.005.

Brown, N. and Sandholm, T. Superhuman ai for multiplayer poker. *Science*, 365(6456):885–890, 2019.

Chen, D., Chen, R., Zhang, S., Wang, Y., Liu, Y., Zhou, H., Zhang, Q., Wan, Y., Zhou, P., and Sun, L. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. In *Forty-first International Conference on Machine Learning*, 2024.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Daneshjou, R., Vodrahalli, K., Novoa, R. A., Jenkins, M., Liang, W., Rotemberg, V., Ko, J., Swetter, S. M., Bailey, E. E., Gevaert, O., et al. Disparities in dermatology ai performance on a diverse, curated clinical image set. *Science advances*, 8(31):eabq6147, 2022.

Dhariwal, P. and Nichol, A. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Dudík, M., Hofmann, K., Schapire, R. E., Slivkins, A., and Zoghi, M. Contextual dueling bandits. In *Conference on Learning Theory*, pp. 563–587. PMLR, 2015.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J. M., Swetter, S. M., Blau, H. M., and Thrun, S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542:115–118, 2017.

Fan, Y. and Lee, K. Optimizing ddpm sampling with shortcut fine-tuning. *arXiv preprint arXiv:2301.13362*, 2023.

Fan, Y., Watkins, O., Du, Y., Liu, H., Ryu, M., Boutilier, C., Abbeel, P., Ghavamzadeh, M., Lee, K., and Lee, K. Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 79858–79885. Curran Associates, Inc., 2023a. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/fc65fab891d83433bd3c8d966edde311-Paper-Conference.pdf.

Fan, Y., Watkins, O., Du, Y., Liu, H., Ryu, M., Boutilier, C., Abbeel, P., Ghavamzadeh, M., Lee, K., and Lee, K. Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models. *arXiv preprint arXiv:2305.16381*, 2023b.

Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A. H., Chechik, G., and Cohen-Or, D. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022a.

Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A. H., Chechik, G., and Cohen-Or, D. An image is worth one word: Personalizing text-to-image generation using textual inversion, 2022b. URL https://arxiv.org/abs/2208.01618.

Ge, W., Chen, S., Chen, G., Chen, J., Chen, Z., Yan, S., Zhu, C., Lin, Z., Xie, W., Wang, X., et al. Mllm-bench, evaluating multimodal llms using gpt-4v. *arXiv preprint arXiv:2311.13951*, 2023.

Geng, S., Hsieh, C.-Y., Ramanujan, V., Wallingford, M., Li, C.-L., Koh, P. W. W., and Krishna, R. The unmet promise of synthetic training images: Using retrieved real images performs better. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 7902–7929. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/0f25eb6e9dc26c933a5d7516abf1eb8c-Paper-Conference.pdf.

Groh, M., Harris, C., Soenksen, L., Lau, F., Han, R., Kim, A., Koochek, A., and Badri, O. 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*, pp. 1820–1828, 2021.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Hu, X., Wang, J., Hamm, J., Yotsu, R. R., and Ding, Z. Enhancing skin disease diagnosis: Interpretable visual concept discovery with sam. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 172–181. IEEE, 2025.

Huang, P., Gao, X., Huang, L., Jiao, J., Li, X., Wang, Y., and Guo, Y. Chest-diffusion: a light-weight text-to-image model for report-to-cxr generation. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 1–5. IEEE, 2024.

Khader, F., Müller-Franzes, G., Tayebi Arasteh, S., Han, T., Haarburger, C., Schulze-Hagen, M., Schad, P., Engelhardt, S., Baeßler, B., Foersch, S., et al. Denoising diffusion probabilistic models for 3d medical image generation. *Scientific Reports*, 13 (1):7303, 2023.

Ktena, I., Wiles, O., Albuquerque, I., Rebuffi, S.-A., Tanno, R., Roy, A. G., Azizi, S., Belgrave, D., Kohli, P., Cemgil, T., et al. Generative models improve fairness of medical classifiers under distribution shifts. *Nature Medicine*, pp. 1–8, 2024.

Lee, H., Phatale, S., Mansoor, H., Mesnard, T., Ferret, J., Lu, K., Bishop, C., Hall, E., Carbune, V., Rastogi, A., et al. Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*, 2023a.

Lee, K., Liu, H., Ryu, M., Watkins, O., Du, Y., Boutilier, C., Abbeel, P., Ghavamzadeh, M., and Gu, S. S. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*, 2023b.

Lee, K., Liu, H., Ryu, M., Watkins, O., Du, Y., Boutilier, C., Abbeel, P., Ghavamzadeh, M., and Gu, S. S. Aligning text-to-image models using human feedback, 2023c. URL https://arxiv.org/abs/2302.12192.

Liu, Y., Jain, A., Eng, C., Way, D. H., Lee, K., Bui, P., Kanada, K., de Oliveira Marinho, G., Gallegos, J., Gabriele, S., et al. A deep learning system for differential diagnosis of skin diseases. *Nature medicine*, 26(6):900–908, 2020.

Liu, Z., Miao, Z., Zhan, X., Wang, J., Gong, B., and Yu, S. X. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2537–2546, 2019.

OpenAI. Gpt-4o system card, 2024. URL https://arxiv.org/abs/2410.21276.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H. V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Howes, R., Huang, P.-Y., Xu, H., Sharma, V., Li, S.-W., Galuba, W., Rabbat, M., Assran, M., Ballas, N., Synnaeve, G., Misra, I., Jegou, H., Mairal, J., Labatut, P., Joulin, A., and Bojanowski, P. Dinov2: Learning robust visual features without supervision, 2023.

Pinaya, W. H., Tudosiu, P.-D., Dafflon, J., Da Costa, P. F., Fernandez, V., Nachev, P., Ourselin, S., and Cardoso, M. J. Brain imaging generation with latent diffusion models. In *MICCAI Workshop on Deep Generative Models*, pp. 117–126. Springer, 2022.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022.

Sagers, L. W., Diao, J. A., Groh, M., Rajpurkar, P., Adamson, A., and Manrai, A. K. Improving dermatology classifiers across populations using images generated by large diffusion models. In *NeurIPS 2022 Workshop on Synthetic Data for Empowering ML Research*, 2022. URL https://openreview.net/forum?id=Vzdbjtz6Tys.

Sagers, L. W., Diao, J. A., Melas-Kyriazi, L., Groh, M., Rajpurkar, P., Adamson, A. S., Rotemberg, V., Daneshjou, R., and Manrai, A. K. Augmenting medical image classifiers with synthetic data from latent diffusion models. *arXiv preprint arXiv:2308.12453*, 2023.

Shin, J., Kang, M., and Park, J. Fill-up: Balancing long-tailed data with generative models. *arXiv preprint arXiv:2306.07200*, 2023.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587): 484–489, 2016.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., et al. Large language models encode clinical knowledge. *Nature*, 620 (7972):172–180, 2023.

Soenksen, L. R., Kassis, T., Conover, S. T., Marti-Fuster, B., Birkenfeld, J. S., Tucker-Schwartz, J., Naseem, A., Stavert, R. R., Kim, C. C., Senna, M. M., et al. Using deep learning for dermatologist-level detection of suspicious pigmented skin lesions from wide-field images. *Science Translational Medicine*, 13(581):eabb3652, 2021.

Sun, S., Goldgof, G., Butte, A., and Alaa, A. M. Aligning synthetic medical images with clinical knowledge using human feedback. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 13408–13428. Curran Associates, Inc., 2023a. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/2b1d1e5affe5fdb70372cd90dd8afd49-Paper-Conference.pdf.

Sun, Z., Shen, S., Cao, S., Liu, H., Li, C., Shen, Y., Gan, C., Gui, L.-Y., Wang, Y.-X., Yang, Y., et al. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*, 2023b.

Sutton, R. S., Barto, A. G., et al. Introduction to reinforcement learning. 1998.

Wallace, B., Dang, M., Rafailov, R., Zhou, L., Lou, A., Purushwalkam, S., Ermon, S., Xiong, C., Joty, S., and Naik, N. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8228–8238, 2024.

Wang, J., Chung, Y., Ding, Z., and Hamm, J. From majority to minority: A diffusion-based augmentation for underrepresented groups in skin lesion analysis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 14–23. Springer, 2024a.

Wang, J., Zhang, Y., Ding, Z., and Hamm, J. Achieving reliable and fair skin lesion diagnosis via unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5157–5166, 2024b.

Ward, A., Li, J., Wang, J., Lakshminarasimhan, S., Carrick, A., Campana, B., Hartford, J., Sreenivasaiah, P. K., Tiyasirisokchai, T., Virmani, S., Wong, R., Matias, Y., Corrado, G. S., Webster, D. R., Smith, M. A., Siegel, D., Lin, S., Ko, J., Karthikesalingam, A., Semturs, C., and Rao, P. Creating an empirical dermatology dataset through crowdsourcing with web search advertisements. *JAMA Network Open*, 7(11):e2446615–e2446615, 11 2024. ISSN 2574-3805. doi: 10.1001/jamanetworkopen.2024.46615. URL https://doi.org/10.1001/jamanetworkopen.2024.46615.

Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J., and Dong, Y. Imagereward: Learning and evaluating human preferences for text-to-image generation. *arXiv preprint arXiv:2304.05977*, 2023.

Yang, K., Tao, J., Lyu, J., Ge, C., Chen, J., Li, Q., Shen, W., Zhu, X., and Li, X. Using human feedback to fine-tune diffusion models without any reward model. *arXiv preprint arXiv:2311.13231*, 2023.

Zhang, K., Li, B., Zhang, P., Pu, F., Cahyono, J. A., Hu, K., Liu, S., Zhang, Y., Yang, J., Li, C., et al. Lmms-eval: Reality check on the evaluation of large multimodal models. *arXiv preprint arXiv:2407.12772*, 2024.

Zhang, X., Lu, Y., Wang, W., Yan, A., Yan, J., Qin, L., Wang, H., Yan, X., Wang, W. Y., and Petzold, L. R. Gpt-4v (ision) as a generalist evaluator for vision-language tasks. *arXiv preprint arXiv:2311.01361*, 2023.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023.

Appendix

A. Additional Implementation Details

In this section, we present additional implementation details of our proposed method.

A.1. Pre-Feedback Fine-tuning

For textual inversion, we learn the text embedding for each skin condition through various prompts. These prompts are used to ensure robust learning of the text embedding across different phrasings and contexts:

```
skin_disease_prompt = [
    "a photo of a {}",
    "a rendering of a {}",
    "a cropped photo of the {}",
    "the photo of a {}",
    "a close-up photo of a {}",
    "a cropped photo of a {}",
    "a photo of the {}",
    "a photo of one {}",
    "a close-up photo of the {}",
    "a rendition of the {}",
    "a rendition of a {}" ]
```

The text embeddings are learned don't the entire training set. The AdamW optimizer is used with a learning rate of 5×10^{-4}.

For LoRA, the rank r is set to 32, and the learning rate is 5×10^{-6} for AdamW optimizer.

A.2. MLLM Score Processing for Preference Pairs

To translate 5-dimensional binary MLLM scores into preference signals for DPO, pairs of generated images are processed. For each image in a pair, its 5 binary scores are summed to get S_1 and S_2. If $\max(S_1, S_2) \leq 2$, both images are deemed low quality (outcome e.g., $[0, 0]$). If $\min(S_1, S_2) = 5$, or if $S_1 = S_2 > 2$, the pair is marked "both win" (e.g., $[1, 1]$). Otherwise, if $S_1 > S_2$, the first image is the "winner" (e.g., $[1, 0]$); if $S_2 > S_1$, the second wins (e.g., $[0, 1]$). This determines preferred/non-preferred samples for DPO loss computation. The distribution of these outcomes is in Table 7.

A.3. DPO fine-tuning

We conduct DPO fine-tuning for 128 iterations and for each iteration, 8 pairs (16 images) will be sampled. The denoise strength γ is set to 0.3. The DPO loss will be computed with the feedback. We utilize AdamW optimizer with a learning rate of 0.0001.

A.4. Classifier Training

We utilize the Adam optimizer with a learning rate of 0.01 and a step learning rate scheduler that reduces the learning rate to 0.1 of its previous value every 50 epochs. The classifier is trained for 200 epochs to ensure stable results. Each result reported in the table represents the average of five runs with different random seeds.

B. Expert Designed Checklist

We enclose the checklist we used in the experiment in this section. For each skin condition, we design 5 checklist evaluations from the perspective of `[Location, Lesion Type, Shape/Size, Color, Texture]` to capture the visual concept from the synthetic data. The details are shown in Table 10.

C. Automate Evaluation via MLLMs

For each pair of data, we use the following prompt to collect feedback from ChatGPT-4o:

```
prompt = f'''Evaluate images against the
following checklist:
{condition_checklist}
Return a list indicating whether
it satisfies each checklist
```

```
item (1 for satisfied, 0 otherwise).
Only the list of results should
be returned. Expected format:
[1, 0, 1, 0, 0]'''
```

D. Addtional Results

D.1. Distribution of Feedback

For each pair of data, our approach categorizes feedback into three types: both win ($[w = 0, w = 1]$), both lose ($[l = 0, l = 1]$), and one better than the other ($[w = 0, l = 1]$ or $[l = 0, w = 1]$). We present the distribution of feedback received during DPO training in Table 7.

D.2. More examples of image pairs

We provide two more image pairs in Fig. 6

D.3. Results on SCIN

The SCIN dataset (Ward et al., 2024), collected via a voluntary image donation platform from Google Search users in the United States, typically includes up to three images per case, each evaluated by up to three dermatologists. This diagnostic process yields a weighted skin condition label for each case. To ensure label accuracy for our study, we selected the condition with the highest weight as the definitive label, discarding ambiguous cases where multiple conditions had equal probabilities. Our analysis concentrated on the 10 most prevalent classes in the real world. Given that the SCIN dataset exhibits an imbalanced class distribution, we first sampled a uniformly distributed test set, following methodologies similar to ImageNet-LT (Liu et al., 2019). Furthermore, guided by approaches like that of (Shin et al., 2023), we employed our MAGIC-DPO framework to generate additional synthetic images for each condition, aiming to augment the test set towards a more uniform distribution. Further details on the dataset distribution are provided in Table 9. However, experiments conducted with this augmented SCIN dataset yielded suboptimal results, potentially attributable to inherent noise within the dataset, a challenge noted in works such as (Hu et al., 2025).

Our MAGIC framework's effectiveness is further validated on the SCIN dataset, with detailed performance for both ResNet18 and DINOv2 classifiers presented in Table 6. For the ResNet18 classifier on SCIN, models trained on real data achieved an accuracy of 23.13%. Standard T2I augmentation slightly decreased this to 22.60% (-0.5%), while I2I augmentation offered a modest improvement to 24.13% ($+1.0\%$). In contrast, our MAGIC framework demonstrated more substantial gains: MAGIC-RFT increased accuracy to 26.58% ($+3.5\%$), and MAGIC-DPO further improved it to 29.43% ($+6.3\%$). A similar trend was observed with the DINOv2 classifier, which had a baseline accuracy of 30.61% on real SCIN data. T2I augmentation reduced accuracy to 28.18% (-2.4%), and I2I provided a small increase to 32.15% ($+1.5\%$). Both MAGIC strategies again outperformed these: MAGIC-RFT achieved 33.82% accuracy ($+3.2\%$), while MAGIC-DPO led with 35.65% ($+5.0\%$). These results on the SCIN dataset consistently show the advantages of leveraging MAGIC, with both RFT and DPO components enhancing performance over standard augmentation techniques, and DPO often yielding the highest accuracy.

D.4. Score change during training

Figure 5 illustrates how the clinical quality of generated images, assessed by the number of satisfied expert-defined criteria, evolves throughout the feedback-guided training phase of our MAGIC framework. Initially, images from the Pre-trained model and the fine-tuned Text-to-Image (T2I) model satisfy very few criteria, with average scores of 0.3 and 0.5, respectively. Even the fine-tuned Image-to-Image (I2I) model, at the beginning of feedback training (Iteration 0), achieves an average of only 1.4 criteria met. As the model receives more feedback and training progresses (Iterations 32 through 128), a significant improvement is observed. The distribution of scores progressively shifts towards satisfying a higher number of clinical criteria, with the average number of criteria met increasing steadily from 1.4 to 3.0 by Iteration 128. This trend clearly demonstrates the diffusion model's ability to learn from and adapt to the expert-derived feedback over time, resulting in generated images that are increasingly more aligned with clinical requirements for medical accuracy.

E. Limitations

The efficacy of our MAGIC framework, like similar feedback-driven approaches, is naturally guided by the detail within the expert-crafted checklists and the continually advancing interpretive capabilities of Multimodal Large Language Models (MLLMs). The scope of conditions and populations within the dermatology datasets utilized (Fitzpatrick17k and SCIN) provides the foundation for the current findings, and extending this work to even broader and more varied datasets presents an exciting avenue for future research. While MAGIC demonstrates considerable potential in dermatology, its promising AI-Expert collaboration paradigm also invites future exploration and adaptation to enhance synthetic data generation in other medical imaging fields, each with its unique visual characteristics and clinical requirements.

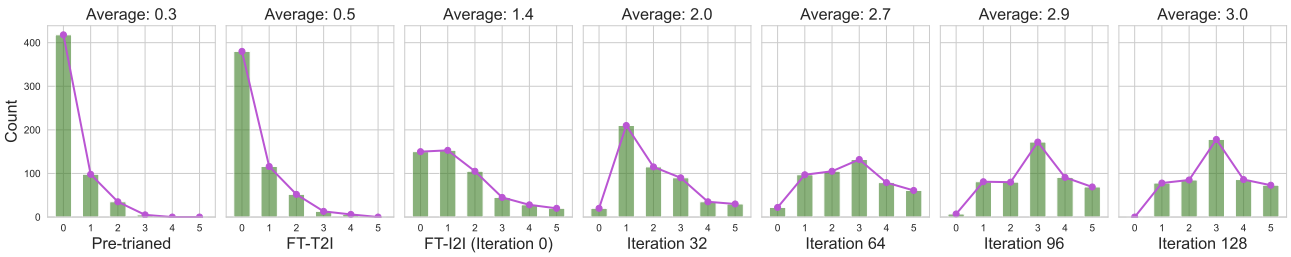


Figure 5. Feedback distribution as training progresses.

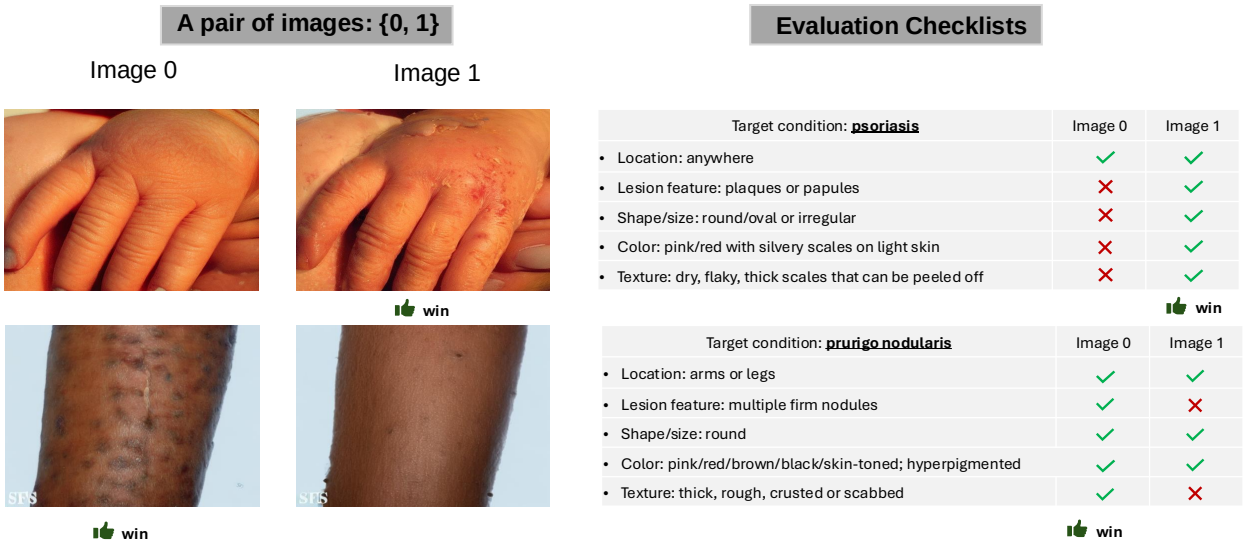


Figure 6. Two image pairs with the corresponding checklist.

Table 5. Performance of ResNet18-based classifiers trained on real and synthetic data for SCIN.

Training data	Acc	F1	Prec	Rec
Real	23.13	10.94	12.20	10.70
+ T2I	22.60	10.44	12.43	10.96
	-0.5	-0.5	+0.2	+0.3
+ I2I	24.13	10.90	12.03	11.06
	+1.0	0.0	-0.2	+0.4
+ MAGIC	26.58	11.69	15.79	11.89
RFT	+3.5	+0.7	+3.6	+1.2
+ MAGIC	**29.43**	**12.16**	**18.18**	**11.47**
DPO	**+6.3**	**+1.2**	**+6.0**	**+0.8**

Table 6. Performance of DINOv2-based classifiers trained on real and synthetic data for SCIN.

Training data	Acc	F1	Prec	Rec
Real	30.61	18.37	21.23	17.45
+ T2I	28.18	17.48	20.23	16.15
	-2.4	-0.9	-1.0	-1.3
+ I2I	32.15	20.10	23.80	19.06
	+1.5	+1.7	+2.6	+1.6
+ MAGIC	33.82	20.08	24.16	18.70
RFT	+3.2	+1.7	+2.9	+1.2
+ MAGIC	**35.65**	**21.39**	**24.00**	**19.40**
DPO	**+5.0**	**+3.0**	**+2.8**	**+1.9**

Table 7. Distribution of feedback

	both win	only one win	both lose
count	295	397	332

Table 8. Skin Condition Distribution for Fitzpatrick17k

Skin Condition	Real Training	Real Test	Synthetic
Acne	92	91	93
Actinic Keratosis	88	87	164
Allergic Contact Dermatitis	215	215	181
Basal Cell Carcinoma	234	234	154
Eczema	102	102	166
Erythema Multiforme	118	118	155
Folliculitis	171	171	114
Granuloma Annulare	106	105	148
Keloid	78	78	135
Lichen Planus	246	245	151
Lupus Erythematosus	205	205	172
Melanoma	130	131	155
Mycosis Fungoides	91	91	165
Pityriasis Rosea	96	97	156
Prurigo Nodularis	85	85	152
Psoriasis	326	327	165
Sarcoidosis	174	175	162
Scabies	170	169	176
Squamous Cell Carcinoma	290	291	175
Vitiligo	83	83	161
Total	3100	3100	3100

Table 9. Skin Condition Distribution for SCIN

Skin Condition	Real Training	Real Test	Synthetic
Eczema	409	36	0
Urticaria	178	34	0
Folliculitis	104	35	33
Tinea	72	34	58
Psoriasis	57	39	70
Herpes Simplex	49	36	76
Acne	44	31	80
Herpes Zoster	41	29	82
Pityriasis rosea	41	32	82
Tinea Versicolor	27	34	93
Total	1022	340	574

Table 10: Skin Conditions and Their Checklist Properties

Skin Condition	Checklist Details
Acne	1. Location: Face, forehead, chest, shoulders, upper back (areas with many oil glands) 2. Lesion Type: Bumps including comedones (whiteheads, blackheads) and inflamed pimples (papules, pustules, nodules) 3. Shape/Size: Small clogged-pore bumps; larger tender nodules/cysts in severe cases 4. Color: Red or skin-colored bumps (may appear purple/brown on dark skin); blackheads have dark plug, whiteheads have white tip 5. Texture: Oily or shiny skin with multiple bumps; some lesions with pus or crust if ruptured
Actinic keratosis	1. Location: Sun-exposed areas (face, scalp, ears, neck, forearms, backs of hands) 2. Lesion Type: Rough, scaly patch or small crusty bump 3. Shape/Size: Flat or slightly raised lesion, usually under 2.5 cm 4. Color: Pink, red, or brownish, possibly with a yellowish crust; on darker skin can appear gray or dark 5. Texture: Dry, coarse, sandpaper-like surface; may have a hard or wart-like feel
Allergic contact dermatitis	1. Location: Where allergen contacts skin (hands, face, eyelids, neck, etc.) 2. Lesion Type: Red patches often with small blisters (vesicles) or swelling 3. Shape/Size: Irregular shape following exposure pattern; size depends on contact area 4. Color: Pink to red on light skin; can be darker, purple, or brownish on dark skin 5. Texture: May be weepy, crusty, or scaly; inflamed and swollen in acute cases
Basal cell carcinoma	1. Location: Sun-exposed areas (face, nose, ears, neck, scalp, shoulders) 2. Lesion Type: Pearly or waxy bump/nodule, or flat scaly patch with a raised edge 3. Shape/Size: Small, round/oval; can ulcerate or develop a central depression 4. Color: Translucent or pearly on fair skin; brown/black or glossy dark on darker skin 5. Texture: Smooth, shiny surface; can crust or scab with central ulceration
Eczema	1. Location: Flexural areas (inner elbows, behind knees), hands, ankles, neck, eyelids, cheeks 2. Lesion Type: Patches or plaques, sometimes with small blisters or bumps 3. Shape/Size: Ill-defined patches varying in size; often bilateral or symmetric 4. Color: Red or pink on lighter skin; purple, gray, or dark brown on darker skin 5. Texture: Dry, flaky, or scaly; can become thick and leathery (lichenification)

Continued on next page

Table 10 continued from previous page

Skin Condition	Checklist Details
Erythema multiforme	1. Location: Hands, feet, arms, legs, can involve mucous membranes (lips, mouth, eyes) 2. Lesion Type: Target (bull's-eye) lesions with concentric rings 3. Shape/Size: Round lesions (1–3 cm) with a dark center, pale ring, and outer red ring 4. Color: Center is dark red/purple, ring is lighter or pink, outer zone is red; on dark skin, may be grayish or hyperpigmented center 5. Texture: Mostly flat but can have a blistered or raised center
Folliculitis	1. Location: Hair-bearing areas prone to friction or shaving (beard, scalp, underarms, legs, buttocks) 2. Lesion Type: Small pustules or red papules centered around hair follicles 3. Shape/Size: Clusters of 2–5 mm bumps; each with a central hair 4. Color: Red or pink on light skin; darker or hyperpigmented on dark skin; pus may appear white/yellow 5. Texture: Dome-shaped, often with a fluid-filled top; can crust if ruptured
Granuloma annulare	1. Location: Hands, feet, wrists, ankles (localized); can appear on trunk/limbs if generalized 2. Lesion Type: Smooth, firm bumps (papules) forming rings; typically non-scaly 3. Shape/Size: Annular (ring-shaped) up to a few cm wide; papules are a few mm each 4. Color: Skin-colored, pink, or reddish; can appear purple on darker skin 5. Texture: Generally smooth; little to no flaking or crust
Keloid	1. Location: Scars on chest, shoulders, earlobes, jawline, or any site of skin injury 2. Lesion Type: Overgrown scar tissue extending beyond the original wound 3. Shape/Size: Raised, irregularly shaped scar; can be small or grow large over time 4. Color: Pink or red on lighter skin; darker, purple or brown on darker skin 5. Texture: Smooth, hairless, firm/rubbery; shiny surface
Lichen planus	1. Location: Wrists, forearms, ankles, scalp, nails, mouth, genitals 2. Lesion Type: Flat-topped papules; can form plaques or lines from scratching 3. Shape/Size: Polygonal, 2–10 mm papules 4. Color: Violaceous (purple) on light skin; gray-brown or hyperpigmented on dark skin 5. Texture: Shiny surface with fine white lines (Wickham's striae); can be scaly if scratched
Lupus erythematosus	1. Location: Face (butterfly rash across cheeks/nose); can appear on scalp/ears; photosensitive areas 2. Lesion Type: Flat or slightly raised rash (malar/butterfly); discoid lesions can be scaly and scarred 3. Shape/Size: Butterfly rash covers the bridge of nose and both cheeks; discoid lesions are coin-shaped (1–3 cm) 4. Color: Pink-red on light skin; can be darker red or hyperpigmented on darker skin 5. Texture: Malar rash smooth or slightly raised; discoid can be rough/scaly with scarring

Continued on next page

Table 10 continued from previous page

Skin Condition	Checklist Details
Melanoma	1. Location: Can appear anywhere (trunk, limbs, face, nails); in darker skin, often on palms/soles or under nails 2. Lesion Type: Atypical mole or patch; irregular shape and color 3. Shape/Size: Asymmetric, often >6 mm, with notched/bumpy edges 4. Color: Multiple shades (brown, black, red, white, blue); on dark skin, often very dark with variation 5. Texture: Smooth early; may become raised, crusted, or ulcerated if advanced
Mycosis fungoides	1. Location: Usually non-sun-exposed areas (buttocks, lower abdomen, thighs); can spread more widely later 2. Lesion Type: Patches (like eczema), plaques (thickened), or tumor nodules (advanced) 3. Shape/Size: Irregular shapes, patches often a few cm wide; plaques larger/thicker; nodules can be several cm 4. Color: Pink-red to reddish-brown; darker or hyperpigmented on darker skin 5. Texture: Dry, scaly for patches; plaques thicker/scaly; nodules can be smooth or ulcerated
Pityriasis rosea	1. Location: Trunk (back, chest, abdomen) primarily; occasionally upper arms, thighs 2. Lesion Type: Herald patch (large oval) followed by multiple smaller oval patches/papules 3. Shape/Size: Herald patch ˜2-6 cm; daughter lesions ˜1-2 cm; often align in 'Christmas tree' pattern 4. Color: Pink/salmon on light skin; gray, brown, or purplish on dark skin 5. Texture: Fine collarette scale at inner edge; not typically thick or crusty
Prurigo nodularis	1. Location: Arms, legs, upper back, shoulders, scalp, areas easily reached for scratching 2. Lesion Type: Firm, itchy nodules, often with a crusted or scabbed top 3. Shape/Size: Round nodules 1-3 cm; multiple lesions often present 4. Color: May be pink, red, brown, black, or skin-toned; older lesions can be hyperpigmented 5. Texture: Thick, rough; scabs from scratching; firm to touch
Psoriasis	1. Location: Elbows, knees, scalp, lower back; can affect nails, palms, soles, or be widespread 2. Lesion Type: Well-demarcated plaques with thick, scaly surface; can also be smaller papules 3. Shape/Size: Round/oval or irregular plaques; can range from small patches to large areas 4. Color: On light skin, pink/red with silvery scales; on dark skin, purple/dark brown with grayish scales 5. Texture: Dry, flaky scales that can be peeled off; underlying skin may bleed (Auspitz sign)
Sarcoidosis	1. Location: Face (nose, cheeks - lupus pernio), shins (erythema nodosum), scars/tattoos, can be widespread 2. Lesion Type: Firm plaques, nodules, or discolored patches; red bumps on shins in erythema nodosum 3. Shape/Size: Plaques are broad and raised; nodules can be 1-5 cm; patchy discolorations vary 4. Color: Purplish or red-brown lumps; can be lighter/darker patches on dark skin; scars can turn red 5. Texture: Smooth, firm or rubbery; some lesions (erythema nodosum) are tender lumps under the skin

Continued on next page

Table 10 continued from previous page

Skin Condition	Checklist Details
Scabies	1. Location: Finger webs, wrists, waist, buttocks, genitals, armpits; in infants: palms, soles, scalp 2. Lesion Type: Tiny burrows (thin, wavy lines) plus small itchy bumps or vesicles 3. Shape/Size: Burrows ˜5–15 mm long; bumps ˜1–2 mm in clusters 4. Color: Skin-toned to pink/red; on darker skin, may appear darker or hyperpigmented 5. Texture: Scratch marks, crusted spots from itching; burrows feel like slight ridges
Squamous cell carcinoma	1. Location: Sun-exposed areas (face, ears, lips, hands), chronic scars, or wounds; can appear on mucosal surfaces 2. Lesion Type: Crusty or scaly bump, ulcer, or plaque; can have raised borders or a central depression 3. Shape/Size: Firm nodule or patch, >1 cm if untreated; may grow rapidly 4. Color: Pink/red on lighter skin; brown or darker on brown/Black skin; can show white/yellow keratin 5. Texture: Rough, thick, crusted surface; may bleed or ulcerate; firm on palpation
Vitiligo	1. Location: Face (around eyes, mouth), hands, feet, arms, legs, genitals; can occur anywhere on body 2. Lesion Type: Depigmented patches with well-defined borders; hair may turn white in affected area 3. Shape/Size: Irregular shapes; can start small and enlarge over time, often symmetrical 4. Color: Completely white or pale compared to surrounding skin; high contrast on darker skin 5. Texture: Normal skin texture (no scaling or thickening), only color is lost