# PRISM: CLASS-CONSISTENT IMAGE GENERATION FROM TABULAR DATA

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Many real-world datasets are tabular. While recent open-domain foundation models condition image generation primarily on text, a common strategy for tabular data is to serialize rows into natural language and condition a text-to-image diffusion model. This clubs linguistically similar classes together even when they are visually distinct, giving incorrectly conditioned images. We propose PRISM, a tri-fold model, which maps tables directly to images without using intermediate mapping to text. We train a tabular conditioning encoder that converts each attribute into a dedicated cross-attention token for the diffusion model. We then fine-tune the model toward the target class, away from confusable alternatives using a GAN-style discriminative loss on generated images. At inference, we steer towards the correct class via classifier-free guidance maximizing target class alignment without any textual input. We evaluate PRISM on three diverse applications: medical imaging, fine-grained bird classification, and plant disease recognition. Evaluations reveal improvements of up to 116% in class consistency over the best language-based baseline. On the medical dataset, a board-certified dermatologist evaluated the generated images and judged most to be visually indistinguishable from real samples. `https://anonymous.4open.science/r/PRISM-8FC2`

## 1 INTRODUCTION

Many real-world datasets are tabular, containing a mix of categorical and discrete attributes. However, conditional image generation models accept visual or textual inputs and cannot directly use this structure. The naïve workaround would be to serialize the table into a text prompt via CLIP embeddings Radford et al. (2021), and then feed it into a text-to-image diffusion model Rombach et al. (2022). But this approach forces the table into a representation it was not designed for. Categorical variables are mapped to vocabulary tokens, and continuous values are converted into word sequences, with attribute relationships expressed through sentence structure. These transformations introduce vocabulary ambiguity, semantic mismatch, and information loss absent from the original representation Lee et al. (2025b).

Even if serialization was perfect, language-based conditioning faces a second limitation. Text encoders compress class identity into an embedding space optimized for linguistic similarity, not visual discrimination. Hence, classes that are linguistically similar but visually distinct are mapped to nearby regions, leading to incorrect generations. The generator then produces visually similar outputs regardless of how different the true classes actually are Qiao et al. (2019); Jiang et al. (2024); Gupta et al. (2024). For example, *"Bactrian camel"* and *"Dromedary camel"* share the dominant token *"camel,"* which can result in similar CLIP embeddings despite being visually distinguishable by the number of humps.

We propose PRISM (Per-attribute Representation for Image Synthesis from Metadata), a tri-fold framework for class-consistent image generation from tabular metadata, designed to replace CLIP-based conditioning with a structured representation that preserves both attribute identity and inter-class distinctions. PRISM is comprised of three sequential stages. First, we adapt a pre-trained text-to-image diffusion model using lightweight LoRA Hu et al. (2021) adapters and jointly train a *Tabular Conditioning Encoder* that converts each attribute into a dedicated cross-attention token, replacing text embeddings entirely. This provides a strong base while preserving attribute identity and avoiding interference between fields. Second, we freeze the Tabular Conditioning Encoder and apply



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.

*Similarity-Aware Class Refinement*, a GAN-style Goodfellow et al. (2014) discriminative loss to fine-tune the diffusion backbone towards separating visually confusable classes by pushing generated images toward the target class and away from nearby alternatives. Third, we apply *classifier-free guidance* (CFG) Ho & Salimans (2022) at *inference* to further improve class consistency.

We validate our approach on three benchmarks: a challenging subset of fifteen visually similar warbler (bird) species, fine-grained plant disease recognition across three crop species, and a medical imaging task comprising seven skin lesion categories. On the first two, PRISM substantially outperforms all language-based baselines in class consistency, achieving up to 96% alignment with the correct class while maintaining high generation quality. On the medical task, PRISM matches a task-specialized CLIP without any language representation or domain-specific pre-training. On this task, we also validate our results through a blind evaluation by a certified dermatologist, showing that the generated images achieve high realism and diagnostic accuracy, consistent with real images. As a consequence of our attribute-wise encoding, we also observe emergent structure in the learned conditioning space, enabling attribute transfer, foreground localization, and smooth interpolation between classes.

To summarize, our main contributions are: (1) we tackle the problem of class-consistent image generation directly from structured tabular data, highlighting the limitations of language-based conditioning (2) a novel tri-stage framework, PRISM, that replaces text conditioning with structured tabular representations, explicitly addressing class confusion through a combination of tabular conditioning, similarity-aware refinement, and classifier-free guidance (3) we provide strong empirical results across three diverse benchmarks, including a medical imaging task with human evaluation, showing that PRISM consistently outperforms language-based baselines in class consistency without sacrificing visual quality. We also observe that per-attribute conditioning induces a structured conditioning space, from which spatial and semantic properties emerge without explicit supervision.

## 2 Related Work

**Conditional image generation.** Latent diffusion models have established text-conditioned generation as the dominant paradigm for high-quality image synthesis Rombach et al. (2022); Ramesh et al. (2022), with subsequent work extending conditioning to spatial controls, personalization, and image editing Zhang et al. (2023); Ruiz et al. (2023); Brooks et al. (2023). More relevant to our setting, FineDiffusion Pan et al. (2024) scales class-conditional generation to thousands of visually similar categories, and TabLLM Hegselmann et al. (2023) explores processing structured tabular data by converting records into natural language for large language models. Despite these advances, all such approaches ultimately rely on text-based representations, either to encode class identity or to mediate structured inputs, inheriting the limitations of language embeddings and failing to preserve attribute-level structure for image generation.

**Domain-adapted vision-language encoders.** The limitations of general-purpose CLIP for domain-specific tasks have motivated a line of work that pretrains or fine-tunes vision-language encoders on domain-relevant image-text pairs. In the medical domain, MedCLIP Wang et al. (2022b) scales contrastive pretraining to unpaired medical images and reports, while PLIP Huang et al. (2023) leverages pathology image-text pairs sourced from medical social media, and BiomedCLIP Zhang et al. (2025) extends this further by pretraining on fifteen million biomedical image-text pairs spanning diverse imaging modalities. For the natural world, BioCLIP Stevens et al. (2024) pretrains on over ten million biology images spanning hundreds of thousands of taxa, achieving strong performance on fine-grained species classification. These encoders substantially improve downstream classification and retrieval within their respective domains by aligning visual representations with domain-specific language. However, when used as conditioning signals for image generation, they inherit the same fundamental limitation as general-purpose CLIP: tabular metadata must still be serialized into text before encoding, preserving the language bottleneck regardless of how well the encoder is adapted to the domain.

**Structured conditioning spaces in generative models.** A parallel line of work has investigated the geometric structure of latent and conditioning spaces in generative models. In GANs, it has been shown that the latent space of models such as StyleGAN Karras et al. (2019) organizes visual attributes into consistent directions, such that moving along a fixed direction in latent space smoothly changes a single attribute, for example age or lighting, while leaving others unchanged Shen et al.

(2020); Härkönen et al. (2020). These directions are typically discovered post-hoc through supervised or unsupervised analysis rather than being structurally guaranteed by the architecture. More recent work has sought to impose disentanglement explicitly through regularization objectives in VAEs Fil et al. (2021) and diffusion models Yang et al. (2023), encouraging independent factors of variation to be encoded in separate latent dimensions. However, these approaches focus on structuring the latent space itself, rather than designing conditioning mechanisms that preserve the structure of input attributes. As a result, they do not address how to maintain attribute-level control when conditioning on structured signals such as tabular metadata.

**Tabular data to image generation.** A small but growing body of work has begun conditioning image generation directly on tabular inputs. Tang et al. (2023) formally introduce this problem, benchmarking DALL-E and GPT-3 pipelines on realism and diversity but not class consistency; as their task generates a single summary image for an entire table rather than one image per row with an associated class, we do not include it as a class-consistency baseline. Table2Image Lee et al. (2025a) takes a different angle, converting tabular records into autoencoder-generated images as intermediate features for classification rather than as outputs. Most closely related to our work, CTIP Roh et al. (2025) pre-trains a tabular encoder via contrastive learning and uses the resulting embedding to condition a diffusion model for tire footprint generation. However, CTIP and Tang et al. (2023) both collapse tabular features into a single conditioning vector, losing attribute-level structure, and none of the three address the class confusion that arises when classes are linguistically or visually similar.

# 3 PRISM

Our tabular-to-image generator, PRISM, takes the tabular data as conditional inputs to synthesize the corresponding images. Specifically, we are given $\mathcal{D} = \{(x_i, m_i)\}_{i=1}^{N}$, where $x_i$ is an image and $m_i$ is a tabular record with categorical and continuous attributes. Our goal is to learn a model $p_\theta(x \mid m)$ that generates realistic images that match the class, specified by $m_j^1$ for each row $j$.

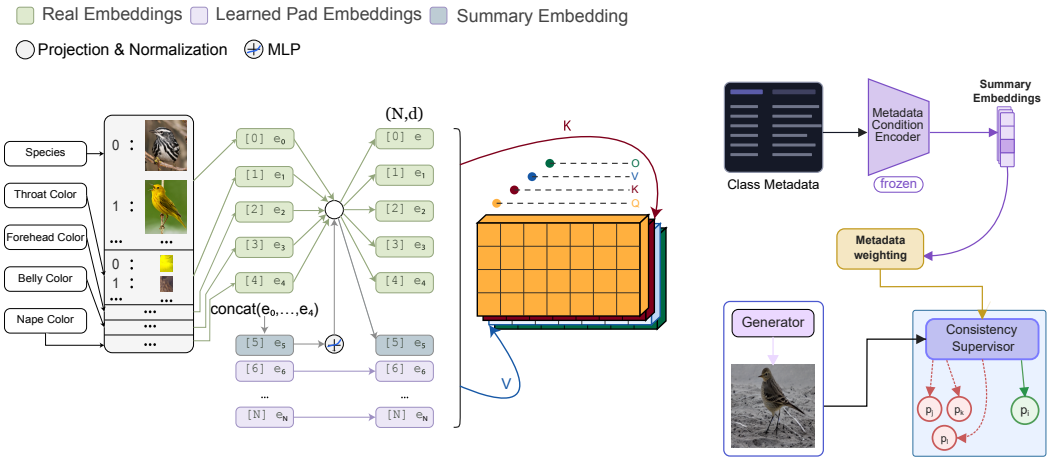
PRISM consists of three stages (see Figure 1). First, we replace the text encoder with a tabular encoder. Each metadata attribute is mapped to a dedicated cross-attention token and fed directly into the diffusion UNet (Sec. 3.1), replacing CLIP entirely. We keep the diffusion model mostly frozen and train only the encoder and LoRA adapters under the standard denoising objective, so the model learns to use tabular inputs without relearning image generation. Second, we fine-tune the model to separate visually-similar classes, through a GAN-style loss on fully generated images (Section 3.2). Third, at inference, we apply classifier-free guidance to amplify the conditioning signal, so generation is more strongly aligned with the target class (Section 3.3).

## 3.1 Stage 1: Training a Tabular Conditioning Encoder

The key design choice in PRISM is to represent each tabular attribute as a separate conditioning token. This diverges from existing approaches (e.g. CTIP Roh et al. (2025)), which maps all tabular attributes into a shared embedding that is then replicated across all cross-attention layers. A shared representation forces distinct attributes (e.g., species identity, coloration) together, resulting in trade-offs between competing attribute signals.

Interestingly, when conditioning signals are not entangled, variations in individual attributes produce consistent and localised changes in the generated image. Specifically, we observe that (i) interpolating between conditioning vectors yields smooth semantic transitions, and (ii) modifying a single attribute leads to targeted, coherent changes without disrupting the rest of the image. For example, conditioning a horse from the AWA2 dataset Xian et al. (2018) with `horns=1` yields anatomically plausible variations, despite such combinations being absent during training (Figure 2, right).

Motivated by this observation, we design a lightweight dataset-agnostic module that maps a tabular row of attributes $m$ directly into a conditioning sequence $c \in \mathbb{R}^{(K+1+N_{\text{pad}}) \times d}$, where $K$ is the number of attributes, $d$ is the cross-attention dimension, and $N_{\text{pad}}$ fills the sequence to the required context length of the cross-attention layers of Stable Diffusion v1.5 Rombach et al. (2022), as shown in Figure 1a. Each attribute is encoded independently into a dedicated token: categorical fields via learned embedding tables, and continuous fields via a two-layer multi-layer-perceptron (MLP) with SiLU activations Ramachandran et al. (2017), followed by a projection to dimension $d$. A summary




(a) Stage 1: Tabular Conditioning Encoder Training

(b) Stage 2: Similarity-Aware Class Refinement (fine-tuning)

Figure 1: PRISM architecture. (Left) Each metadata field is encoded independently into a token; a summary token $\tau_{\text{sum}}$ aggregates all fields through a two-layer MLP, and the resulting token sequence is injected into the UNet cross-attention layers, fine-tuned with LoRA adapters. (Right) Stage 2 pushes generated images toward the target class prototype $p_i$ and away from confusable prototypes, weighted by metadata similarity between summary embeddings.

token $\tau_{\text{sum}}$ is then computed from all attribute embeddings using a second MLP, producing a controlled global update of the conditioning signal. The final conditioning sequence is formed by concatenating all attribute tokens, the summary token, and a set of learned padding tokens. Together, this produces a conditioning representation where attributes are factorized at the token level rather than entangled in a single embedding.

The encoder is trained jointly with LoRA adapters inserted into the cross-attention layers of the UNet, while keeping the rest of the pretrained latent diffusion model, including the VAE, frozen. Training is performed with the standard denoising objective: $\mathcal{L}_{\text{diff}} = \mathbb{E}_{x,\, m,\, \epsilon,\, t}\left[\, \|\epsilon - \epsilon_\theta(z_t,\, t,\, c(m))\|^2 \,\right]$. Conditioning is dropped with probability ($p = 0.15$) to enable classifier-free guidance at inference.

**Example.** Consider a record with $K = 3$ attributes: species, throat color, and belly color. The encoder produces three attribute tokens $\tau_1, \tau_2, \tau_3$ and one summary token $\tau_{\text{sum}}$, padded to the required context length. If throat color changes from yellow to gray, only $\tau_2$ is updated while all other tokens remain identical, and the shift in $\tau_{\text{sum}}$ reflects only that attribute's contribution through the summary MLP. The cross-attention layers in the UNet can therefore learn to attend to $\tau_2$ specifically when denoising the throat region, independently of how the species or belly color tokens are set.

### 3.2 Stage 2: Similarity-Aware Class Refinement

Stage 1 alone is insufficient for classes that remain visually confusable after tabular conditioning. We observe this as residual class overlap in the learned feature space, where certain classes form tight clusters under t-SNE visualizations of pretrained visual embeddings and exhibit higher confusion rates under a standard evaluation model (see Appendix K).

We address this through a second refinement stage that increases separation between such classes in visual feature space. We fine-tune the model using a GAN-style contrastive objective on generated images. For each sample, the loss pulls the image toward its target class prototype while pushing it away from nearby class prototypes. Classes that are closer in the learned conditioning space are assigned higher weights, so harder distinctions receive a stronger training signal.

We freeze the tabular conditioning encoder and update only the LoRA adapters, as re-training the encoder at this stage can disrupt attribute-level structures learned during Stage 1 (Appendix L). The separation loss weighs hard negatives by their metadata similarity: the softmax-weighted similarity

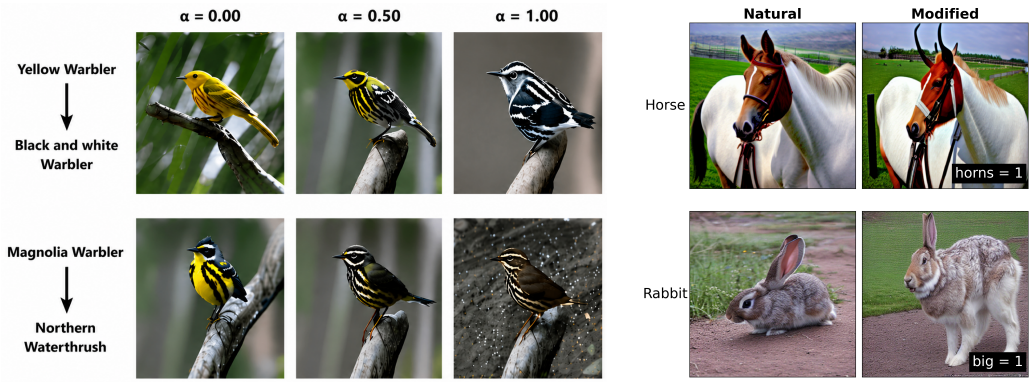



Figure 2: Per-attribute tokenization in PRISM. **(Left)** Each row shows images generated by interpolating between two species' conditioning vectors as $\mathbf{c}(\alpha) = \text{slerp}(\mathbf{c}_A, \mathbf{c}_B, \alpha)$, where $\alpha = 0$ is species A and $\alpha = 1$ is species B; intermediate values produce semantically coherent transitions. **(Right)** Each pair shows a naturally conditioned image alongside the same animal conditioned with a single binary attribute overridden (`horns=1` for horse, `big=1` for rabbit).

between the summary embeddings $\tau^i_{\text{sum}}$ and $\tau^j_{\text{sum}}$, computed from the same tabular records used for conditioning, determines the penalty strength of visual confusion between two classes.

As a reliable and independent metric, we use two separate visual backbones. One serves as the consistency supervisor (ConvNeXt-Large pre-trained on ImageNet-22k Liu et al. (2022)), which evaluates generated images during training, and one is reserved exclusively as the independent class consistency evaluator (DINOv2 Oquab et al. (2024)). Similar to a GAN discriminator, the consistency supervisor evaluates fully generated images and pushes them toward their target class prototype in visual feature space. The combined generator loss is $\mathcal{L}_G = \lambda_{\text{diff}}\mathcal{L}_{\text{diff}} + \lambda_{\text{sep}}(t)\,\mathcal{L}_{\text{sep}}$, where $\lambda_{\text{sep}}$ warms up linearly. The separation loss is:

$$\mathcal{L}_{\text{sep}} = \sum_{j \neq i} \text{softmax}\left(\frac{\text{sim}(\tau^i_{\text{sum}},\, \tau^j_{\text{sum}})}{T}\right) \max(0,\, \delta - \text{sim}(f_{\text{fake}}, p_i) + \text{sim}(f_{\text{fake}}, p_j))\,, \qquad (1)$$

where $f_{\text{fake}} = \text{ConvNeXt}(G(m_i))$ is the normalized feature of the generated image for target class $i$, $p_i$ and $p_j$ are mean prototype vectors over real training images, $\delta$ is the margin, and $T$ is a temperature controlling how sharply weight concentrates on the most metadata-similar classes.

### 3.3 Stage 3: Classifier-Free Guidance at Inference

PRISM's tabular conditioning encoder organizes the embedding space around class-discriminative attributes, giving classifier-free guidance (CFG) Ho & Salimans (2022) a well-structured signal to amplify. At inference, CFG runs the UNet twice per denoising step. Once with the metadata embedding $c(m)$ and once with a null embedding $\mathbf{0}$. The two predictions are combined as: $\epsilon = \epsilon_\emptyset + \gamma\,(\epsilon_{\text{cond}} - \epsilon_\emptyset)$, where $\gamma$ is a fixed guidance scale. This extrapolates away from the unconditional prediction and toward the conditional one, steering generation more strongly toward the target class. Higher values of $\gamma$ increase class consistency at the cost of sample diversity. Images are generated through DDIM Song et al. (2021) steps and decoded by the frozen VAE to the target spatial resolution (Figure 3).

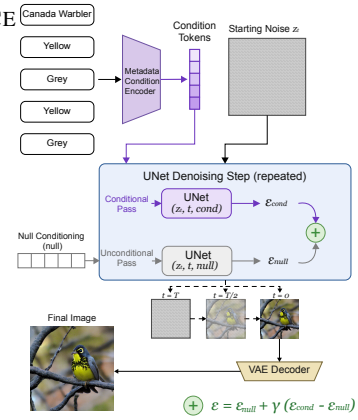



Figure 3: Stage 3: Inference-time CFG

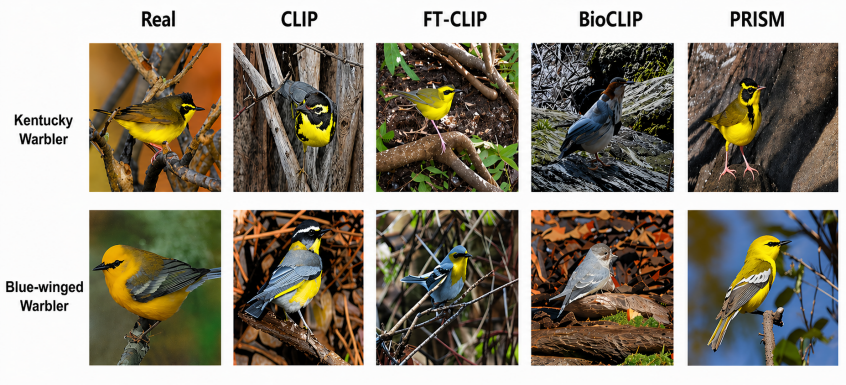



Figure 4: Qualitative comparison of generated images across conditioning methods. Each row shows one generated image per method for a target species, alongside a real reference image. CLIP, FT-CLIP and BioCLIP produce visually incorrect species, while PRISM generates images consistent with the target class. This qualitative pattern matches the class consistency results in Table 1.

# 4 Evaluation

## 4.1 Setup

We compare against three language-based conditioning baselines, all built on Stable Diffusion with LoRA Hu et al. (2021) adapters and trained with the same hyperparameters as PRISM. In all cases, tabular metadata is first serialized into natural language prompts (examples in Appendix G). **CLIP** encodes prompts into token embeddings fed into UNet cross-attention, with the encoder frozen and only UNet LoRA adapters trained. **Fine-tuned CLIP** additionally applies LoRA adapters to the CLIP encoder, enabling domain adaptation. **Specialized CLIP** replaces CLIP with a domain-adapted encoder (PLIP for HAM10000, BioCLIP for CUB-200-2011 and PlantVillage). A learned projection maps the encoder outputs to the cross-attention space, trained jointly with the UNet LoRA adapters while keeping the encoder frozen. We evaluate all methods on three benchmarks with tabular metadata: HAM10000 Tschandl et al. (2018), CUB-200-2011 Wah et al. (2011), and PlantVillage Hughes & Salathé (2015). The metrics used for this evaluation are class consistency and image realism (see Appendix O for details).

## 4.2 Generation quality results

Figure 4 illustrates the core failure mode of language-based conditioning. We use CUB-200-2011 for this qualitative comparison as bird plumage differences are visible to any reader. CLIP, FT-CLIP and BioCLIP all generate visually incorrect species, which is expected: the selected species are nearest neighbors in CLIP's conditioning space, and similar conditioning vectors produce similar generation behavior. Table 1 confirms this quantitatively. CLIP achieves 0.71 mean consistency and

Table 1: Class consistency and image realism metrics across all methods and datasets. For class consistency and CLIPIQA, higher is better. For FID, lower is better. Bold indicates the best result per dataset and metric. Per-class consistency results are shown in Appendix A

| Dataset | Metric | Real | CLIP | FT CLIP | Spec. CLIP | PRISM |
|---|---|---|---|---|---|---|
| HAM10000 | Consistency ↑ | – | $0.32 \pm 0.25$ | $0.35 \pm 0.23$ | $0.73 \pm 0.15$ | $\mathbf{0.73 \pm 0.13}$ |
| | CLIPIQA ↑ | 0.48 | $0.36 \pm 0.07$ | $0.38 \pm 0.04$ | $\mathbf{0.42 \pm 0.04}$ | $0.37 \pm 0.04$ |
| | FID (BiomedCLIP) ↓ | – | 0.154 | 0.136 | 0.153 | 0.094 |
| CUB-200-2011 | Consistency ↑ | – | $0.71 \pm 0.22$ | $0.83 \pm 0.17$ | $0.08 \pm 0.07$ | $\mathbf{0.96 \pm 0.07}$ |
| | CLIPIQA ↑ | 0.74 | $0.74 \pm 0.04$ | $0.72 \pm 0.03$ | $0.71 \pm 0.01$ | $\mathbf{0.75 \pm 0.04}$ |
| | FID ↓ | – | 28.79 | 29.13 | 32.38 | **24.16** |
| PlantVillage | Consistency ↑ | – | $0.35 \pm 0.28$ | $0.43 \pm 0.31$ | $0.12 \pm 0.09$ | $\mathbf{0.93 \pm 0.1}$ |
| | CLIPIQA ↑ | 0.76 | $\mathbf{0.77 \pm 0.04}$ | $0.74 \pm 0.05$ | $0.69 \pm 0.03$ | $0.76 \pm 0.07$ |
| | FID ↓ | – | 55.14 | 66.24 | 52.95 | **44.71** |

fine-tuning raises this to $0.83$, suggesting some species-level signal can be recovered through domain adaptation. Specialized CLIP collapses to near-random consistency ($0.08$), a failure we verified is structural rather than a prompt engineering artifact by evaluating both descriptive and taxonomic prompts aligned with BioCLIP's iNaturalist pretraining distribution (Appendix J). PRISM achieves $0.96$, outperforming all baselines by a substantial margin. Realism metrics are comparable across all methods, with PRISM matching the real image distribution on CLIPIQA ($0.75$) and achieving the best FID (24.16), confirming that consistency gains do not come at the cost of visual quality.

On HAM10000, PRISM achieves $0.73$, matching Specialized CLIP while outperforming all other language-based baselines. Fine-tuning the CLIP encoder provides no meaningful gain over frozen CLIP ($0.35$ vs $0.32$), demonstrating that the language bottleneck is the limiting factor rather than insufficient domain adaptation. Specialized CLIP improves substantially by leveraging PLIP's domain-specific pretraining, yet PRISM matches this performance without any language representation or domain-specific pretraining, relying solely on general-purpose tabular conditioning. Per-class results (Appendix A) show strong consistency across all seven classes, with residual confusion concentrated on `bcc` and `akiec`, which are known to be diagnostically challenging even for specialists. The dermatologist evaluation is consistent with this: PRISM achieves a fool rate of 67.9% and an expert diagnostic accuracy of 35.7%, compared to 57.1% on real images, with the most frequent confusions mirroring those of the classifier on real validation images (Appendix E). Since Inception-based FID relies on ImageNet-pretrained features that are not well suited to medical imagery, we evaluate image quality on HAM10000 using BiomedCLIP-FID, a medical-domain feature space, complemented by our dermatologist evaluation and the qualitative results in Figure 5.

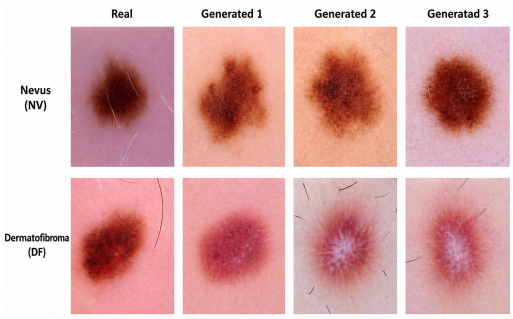



Figure 5: Representative generated images from PRISM for `df` and `nv`, selected as representative classes with distinctive visual patterns that allow direct assessment of realism. Images were judged as clinically real and were correctly classified by a board-certified dermatologist.

On PlantVillage, the same pattern repeats with even greater contrast. As observed on the other datasets, fine-tuning provides only marginal gains over frozen CLIP ($0.43$ vs $0.35$), and Specialized CLIP collapses to $0.12$ regardless of prompt format (Appendix J). PRISM achieves $0.93$, the highest consistency across all three datasets and all methods. This is particularly notable given that PlantVillage uses the simplest conditioning signal of the three, encoding only crop species and disease label, confirming that direct tabular conditioning can exploit even a minimal structured signal with high fidelity. The confusion matrix is nearly perfectly diagonal (Appendix E) and realism metrics follow the same pattern as the other datasets. Qualitative results are shown in Appendix C.

Finally, we examine the effect of the guidance scale $\gamma$ on the quality-consistency trade-off. As $\gamma$ increases, class consistency rises sharply and saturates around $\gamma = 6$, beyond which further increases yield diminishing returns on consistency while FID degrades steadily. This behavior arises because the conditioning encoder already organizes the embedding space around class-discriminative attributes: CFG amplifies a signal that is already well-structured, so modest guidance is sufficient to steer generation toward the target class without sacrificing visual diversity. We fix $\gamma = 6$ across all experiments as the optimal operating point (Appendix M). We further verify that Stage 2 refinement does not induce memorization or a meaningful loss of intra-class diversity relative to Stage 1 (Appendix N). Additionally, to isolate the contribution of per-attribute tokenization, we evaluate a naive baseline that encodes only the species label as a single embedding. Despite preserving coarse color statistics, it produces visually incoherent outputs, confirming that per-attribute structure is the key driver of class consistency (Appendix F). Lastly, to verify that class consistency results are not specific to the choice of evaluator, we additionally report it using a frozen EfficientNet classifier in Appendix B.

### 4.3 Comparison with Tabular-to-Image Generation

We additionally compare against non-language conditioning approaches on CUB-200-2011 (Table 2). ACGAN and FineDiffusion provide class-conditioned baselines, while Scene Graph and CTIP provide structured conditioning alternatives. ACGAN, FineDiffusion, and Scene Graph achieve consistencies of $0.093$, $0.053$, and $0.255$, respectively, compared with $0.960$ for PRISM. Among these methods, CTIP Roh et al. (2025) is the closest to our setting because it conditions a diffusion model directly on the same tabular metadata. Unlike PRISM, however, CTIP encodes the complete attribute row into a single conditioning vector rather than preserving individual attributes as separate cross-attention tokens. We therefore additionally evaluate CTIP on all three datasets. CTIP achieves class consistencies of $0.21$, $0.44$, and $0.91$ on CUB-200-2011, HAM10000, and PlantVillage, respectively, compared with $0.96$, $0.73$, and $0.93$ for PRISM. The difference is largest on fine-grained CUB, smaller on HAM10000, and negligible on PlantVillage, suggesting that per-attribute tokenization is particularly beneficial when the conditioning attributes must distinguish visually confusable classes.

Table 2: Class consistency of non-language conditioning baselines on CUB.

| **Method** | **Consistency ↑** |
|---|---|
| ACGAN | 0.093 |
| FineDiffusion | 0.053 |
| CTIP | 0.210 |
| Scene Graph | 0.255 |
| **PRISM** | **0.960** |

### 4.4 The Role of Language Separability

The failures observed in the previous section are not incidental, they have a structural cause. When two classes occupy nearby regions in CLIP's embedding space, the conditioning vectors that drive generation are nearly indistinguishable, regardless of how visually distinct the classes actually are. Figure 6 makes this explicit. For each dataset, we compute pairwise cosine distances between CLIP text embeddings of class names and measure the corresponding confusion reduction achieved by PRISM over CLIP. Confusion reduction is defined as the average off-diagonal confusion between a pair of classes under CLIP minus the same quantity under PRISM. Across all three benchmarks, the benefit is largest precisely in the low CLIP distance regime ($\leq 0.33$), where PRISM reduces pairwise confusion by up to 0.35 on PlantVillage, 0.15 on HAM10000 and 0.28 on CUB-200-2011. As class names grow more separable in embedding space, the gap narrows monotonically toward zero across all three datasets, confirming that both methods perform comparably once the language bottleneck disappears. This pattern holds across three domains with fundamentally different visual structures: bird plumage, skin lesions and plant disease, suggesting it reflects a general property of text-based conditioning rather than a dataset-specific artifact. The result offers a practical criterion for practitioners: direct tabular conditioning yields the greatest gains precisely when classes are fine-grained enough to be linguistically indistinguishable.

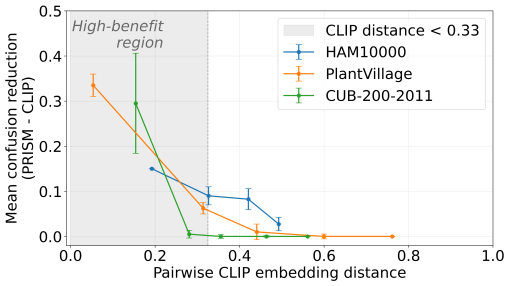



Figure 6: Confusion reduction between PRISM and CLIP as a function of pairwise class separation in CLIP embedding space. Confusion reduction is defined as the mean pairwise difference in off-diagonal confusion matrix entries between CLIP and PRISM between classes. Class pairs are binned into intervals of width 0.2; points show the median confusion reduction per bin with standard error bars. The shaded region (CLIP distance $\leq$ 0.33) identifies class pairs that are nearest neighbors in CLIP embedding space.

### 4.5 Emergent Foreground Localization

Without any explicit supervision, PRISM's conditioning space exhibits emergent geometric structure. Figure 22 shows how the summary embeddings $\tau^i_{\text{sum}}$ are fully separated in t-SNE space. The summary token $\tau_{\text{sum}}$ also reliably localizes the foreground subject. Extracting cross-attention maps at a fixed intermediate timestep and binarizing via Otsu thresholding achieves a mean IoU of $0.63$ across all 15

warbler species (Figure 7). This performance approaches weakly supervised segmentation methods trained only with image-level labels (Appendix D).

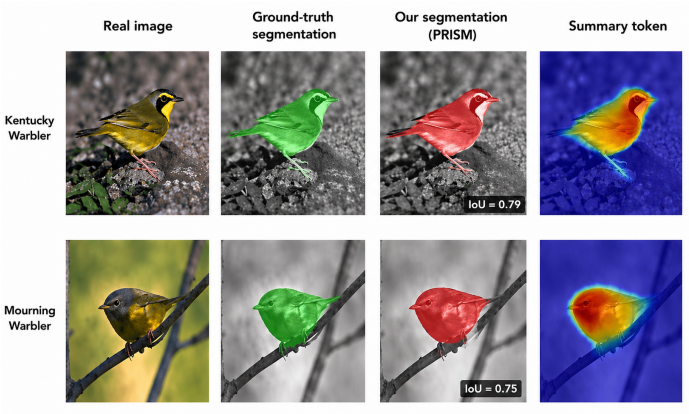



Figure 7: Emergent foreground localization from the summary token $\tau_{\text{sum}}$. Each row shows a held-out warbler image with its ground-truth CUB segmentation mask (green), the binarized cross-attention map via Otsu thresholding (red), and the raw attention heatmap. Mean IoU across all 15 warbler species is $0.63$ vs. a random baseline of $0.20$.

## 5 Discussion and Limitations

Our findings suggest that the widespread reliance on language as a universal conditioning interface may be suboptimal for structured data, as language embeddings inherently compress and entangle attribute information, limiting fine-grained control. In contrast, directly encoding structured variables as separate tokens enables precise alignment between conditioning inputs and generated outputs, leading to improved class consistency without sacrificing visual quality. The emergence of spatial localization and smooth semantic interpolation further indicates that structured conditioning induces a geometrically coherent and semantically grounded latent space. Despite these promising results, we acknowledge the following limitations. First, the method assumes access to high-quality structured metadata. In real-world settings, such data may be incomplete, noisy, or weakly correlated with visual features, which could degrade conditioning consistency. Empirically, however, PRISM degrades gracefully rather than collapsing when auxiliary attributes are corrupted or missing at inference (Appendix N). Second, the per-attribute token design, while empirically effective, may not scale to settings with many attributes or high-cardinality variables. Additionally, due to attribute-wise encoding, higher-order interactions between variables are not captured before cross-attention, which may limit the ability to model complex dependencies in real-world data. Third, evaluation of class consistency depends on external classifiers, which may not reflect perceptual correctness and can underestimate performance when the evaluator itself struggles with certain classes. Beyond these technical limitations, PRISM may also raise practical and ethical concerns. While the method can support data augmentation for underrepresented medical conditions, the $67.9\%$ dermatologist fool rate highlights the risk of synthetic images being mistaken for real clinical evidence.

## 6 Conclusion

We introduced PRISM, a three-stage framework for class-consistent image generation directly from tabular metadata, bypassing the language bottleneck that causes visually distinct but linguistically similar classes to collapse under text-based conditioning. By encoding each metadata attribute as an independent cross-attention token and refining class separation through similarity-aware contrastive training, PRISM achieves up to 96% class consistency across three diverse benchmarks, improving over the best language-based baseline by up to 116%. On medical imaging, generated images were judged indistinguishable from real samples by a board-certified dermatologist, and the learned conditioning space exhibits emergent structure, including foreground localization and smooth interpolation between classes, without any explicit supervision.

### AI USE STATEMENT

In this work, we used a generative AI tool (Claude) to help edit and polish the text of the paper and to write and debug code for several evaluation and baseline experiments, whose outputs we verified before including them. The core PRISM architecture and training procedure were designed and implemented by the authors without AI assistance. We reviewed all AI-assisted text, code, and outputs, and take full responsibility for the final content of this work.

## REFERENCES

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions, 2023. URL https://arxiv.org/abs/2211.09800.

Miroslav Fil, Munib Mesinovic, Matthew Morris, and Jonas Wildberger. Beta-vae reproducibility: Challenges and extensions, 2021. URL https://arxiv.org/abs/2112.14278.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Gunshi Gupta, Karmesh Yadav, Yarin Gal, Dhruv Batra, Zsolt Kira, Cong Lu, and Tim G. J. Rudner. Pre-trained text-to-image diffusion models are versatile representation learners for control, 2024. URL https://arxiv.org/abs/2405.05852.

Stefan Hegselmann, Alejandro Buendia, Hunter Lang, Monica Agrawal, Xiaoyi Jiang, and David Sontag. Tabllm: Few-shot classification of tabular data with large language models, 2023. URL https://arxiv.org/abs/2210.10723.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

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, 2021. URL https://arxiv.org/abs/2106.09685.

Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas J. Montine, and James Zou. A visual–language foundation model for pathology image analysis using medical Twitter. *Nature Medicine*, 29:2307–2316, 2023. doi: 10.1038/s41591-023-02504-3.

David P. Hughes and Marcel Salathé. An open access repository of images on plant health to enable the development of mobile disease diagnostics. *arXiv preprint arXiv:1511.08060*, 2015.

Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls, 2020. URL https://arxiv.org/abs/2004.02546.

Zutao Jiang, Guian Fang, Jianhua Han, Guansong Lu, Hang Xu, Shengcai Liao, Xiaojun Chang, and Xiaodan Liang. Realigndiff: Boosting text-to-image diffusion model with coarse-to-fine semantic re-alignment, 2024. URL https://arxiv.org/abs/2305.19599.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks, 2019. URL https://arxiv.org/abs/1812.04948.

Seungeun Lee, Il-Youp Kwak, Kihwan Lee, Subin Bae, Sangjun Lee, Seulbin Lee, and Seungsang Oh. Table2image: Interpretable tabular data classification with realistic image transformations, 2025a. URL https://arxiv.org/abs/2412.06265.

Simon A. Lee, Sujay Jain, Alex Chen, Kyoka Ono, Arabdha Biswas, Ákos Rudas, Jennifer Fang, and Jeffrey N. Chiang. Clinical decision support using pseudo-notes from multiple streams of ehr data. *npj Digital Medicine*, 8(1), July 2025b. ISSN 2398-6352. doi: 10.1038/s41746-025-01777-x. URL http://dx.doi.org/10.1038/s41746-025-01777-x.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s, 2022. URL https://arxiv.org/abs/2201.03545.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. URL https://arxiv.org/abs/2304.07193.

Ziying Pan, Kun Wang, Gang Li, Feihong He, and Yongxuan Lai. Finediffusion: Scaling up diffusion models for fine-grained image generation with 10,000 classes, 2024. URL https://arxiv.org/abs/2402.18331.

Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. Mirrorgan: Learning text-to-image generation by redescription, 2019. URL https://arxiv.org/abs/1903.05854.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.

Prajit Ramachandran, Barret Zoph, and Quoc V Le. Searching for activation functions. *arXiv preprint arXiv:1710.05941*, 2017.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. URL https://arxiv.org/abs/2204.06125.

Daeyoung Roh, Donghee Han, Jihyun Nam, Jungsoo Oh, Youngbin You, Jeongheon Park, and Mun Yong Yi. Ctip: Towards accurate tabular-to-image generation for tire footprint generation. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1166–1175, 2025. doi: 10.1109/WACV61041.2025.00121.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation, 2023. URL https://arxiv.org/abs/2208.12242.

Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. Interfacegan: Interpreting the disentangled face representation learned by gans, 2020. URL https://arxiv.org/abs/2005.09635.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=St1giarCHLP.

Samuel Stevens, Jiaman Wu, Matthew J Thompson, Elizabeth G Campolongo, Chan Hee Song, David Edward Carlyn, Li Dong, Wasila M Dahdul, Charles Stewart, Tanya Berger-Wolf, Wei-Lun Chao, and Yu Su. Bioclip: A vision foundation model for the tree of life, 2024. URL https://arxiv.org/abs/2311.18803.

Alex Tang, Gromit Yeuk-Yin Chan, Ryan A. Rossi, Chang Xiao, and Eunyee Koh. Tabular data to image generation: Benchmark data, approaches, and evaluation, 2023. URL https://openreview.net/forum?id=g7TXnKjn3Y.

Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5:180161, 2018. doi: 10.1038/sdata.2018.161.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. CUB-200-2011: Caltech-UCSD birds-200-2011 dataset, 2011.

Jianyi Wang, Kelvin C. K. Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images, 2022a. URL https://arxiv.org/abs/2207.12396.

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text, 2022b. URL https://arxiv.org/abs/2210.10163.

Yongqin Xian, Christoph H. Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—A comprehensive evaluation of the good, the bad and the ugly. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(8), 2018. doi: 10.1109/TPAMI.2018.2857768.

Tao Yang, Yuwang Wang, Yan Lv, and Nanning Zheng. Disdiff: Unsupervised disentanglement of diffusion probabilistic models, 2023. URL https://arxiv.org/abs/2301.13721.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. URL https://arxiv.org/abs/2302.05543.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Andrea Tupini, Yu Wang, Matt Mazzola, Swadheen Shukla, Lars Liden, Jianfeng Gao, Angela Crabtree, Brian Piening, Carlo Bifulco, Matthew P. Lungren, Tristan Naumann, Sheng Wang, and Hoifung Poon. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs, 2025. URL https://arxiv.org/abs/2303.00915.

## A PER-CLASS CONSISTENCY RESULTS

Table 3: Per-class consistency measured using a DINOv2-trained classifier across CUB-200-2011 and PlantVillage. Higher values indicate better performance.

**(a) HAM10000**

| Class | CLIP | FT CLIP | Spec. CLIP | PRISM |
|---|---|---|---|---|
| akiec | 0.16 | 0.05 | 0.75 | 0.66 |
| bcc | 0.32 | 0.19 | 0.60 | 0.53 |
| bkl | 0.15 | 0.40 | 0.55 | 0.62 |
| df | 0.06 | 0.13 | 0.91 | 0.91 |
| mel | 0.29 | 0.40 | 0.59 | 0.74 |
| nv | 0.84 | 0.62 | 0.95 | 0.80 |
| vasc | 0.44 | 0.67 | 0.74 | 0.86 |
| **Mean** | $0.32 \pm 0.25$ | $0.35 \pm 0.23$ | $0.73 \pm 0.15$ | $\mathbf{0.73 \pm 0.13}$ |

**(b) CUB-200-2011**

| Class | CLIP | FT CLIP | Spec. CLIP | PRISM |
|---|---|---|---|---|
| American Pipit | 1.00 | 1.00 | 0.01 | 1.00 |
| Black and white Warbler | 0.99 | 1.00 | 0.17 | 1.00 |
| Blue winged Warbler | 0.75 | 0.76 | 0.10 | 0.98 |
| Canada Warbler | 0.88 | 0.86 | 0.05 | 0.97 |
| Cape May Warbler | 0.66 | 0.77 | 0.10 | 0.96 |
| Golden winged Warbler | 0.35 | 0.58 | 0.05 | 0.98 |
| Kentucky Warbler | 0.81 | 0.99 | 0.02 | 1.00 |
| Louisiana Waterthrush | 0.63 | 0.86 | 0.25 | 0.99 |
| Magnolia Warbler | 0.89 | 0.92 | 0.18 | 1.00 |
| Mourning Warbler | 0.45 | 0.58 | 0.01 | 0.95 |
| Northern Waterthrush | 0.59 | 0.52 | 0.11 | 0.91 |
| Pine Warbler | 0.61 | 0.92 | 0.02 | 0.94 |
| Prairie Warbler | 0.75 | 0.91 | 0.00 | 0.98 |
| Tennessee Warbler | 0.27 | 0.73 | 0.05 | 0.73 |
| Yellow Warbler | 0.95 | 1.00 | 0.03 | 0.97 |
| **Mean**$\pm\,\sigma$ | $0.71 \pm 0.22$ | $0.83 \pm 0.17$ | $0.08 \pm 0.07$ | $\mathbf{0.96 \pm 0.07}$ |

**(c) PlantVillage**

| Class | CLIP | FT CLIP | Spec. CLIP | PRISM |
|---|---|---|---|---|
| Pepper bell Bacterial spot | 0.31 | 0.46 | 0.11 | 0.92 |
| Pepper bell healthy | 0.96 | 0.94 | 0.27 | 1.00 |
| Potato Early blight | 0.27 | 0.25 | 0.02 | 1.00 |
| Potato Late blight | 0.63 | 0.68 | 0.08 | 0.90 |
| Potato healthy | 0.81 | 0.94 | 0.19 | 1.00 |
| Tomato Bacterial spot | 0.00 | 0.05 | 0.02 | 0.99 |
| Tomato Early blight | 0.46 | 0.18 | 0.09 | 0.75 |
| Tomato Late blight | 0.37 | 0.72 | 0.32 | 0.98 |
| Tomato Leaf Mold | 0.34 | 0.69 | 0.06 | 0.98 |
| Tomato Septoria leaf spot | 0.01 | 0.10 | 0.05 | 0.67 |
| Tomato Spider mites | 0.02 | 0.01 | 0.09 | 0.84 |
| Tomato Target Spot | 0.09 | 0.05 | 0.03 | 0.96 |
| Tomato Yellow Leaf Curl | 0.53 | 0.75 | 0.08 | 1.00 |
| Tomato mosaic virus | 0.23 | 0.33 | 0.24 | 0.95 |
| Tomato healthy | 0.15 | 0.25 | 0.09 | 0.93 |
| **Mean**$\pm\sigma$ | $0.35 \pm 0.28$ | $0.43 \pm 0.31$ | $0.12 \pm 0.09$ | $\mathbf{0.93 \pm 0.1}$ |

## B CLASS CONSISTENCY USING EFFICIENTNET

Table 4: Per-class consistency measured by a frozen EfficientNet evaluator across all three datasets. Higher is better.

**(a) HAM10000**

| Class | CLIP | FT CLIP | Spec. CLIP | PRISM |
|---|---|---|---|---|
| akiec | 0.05 | 0.03 | 0.61 | 0.69 |
| bcc | 0.27 | 0.17 | 0.77 | 0.73 |
| bkl | 0.11 | 0.52 | 0.47 | 0.79 |
| df | 0.09 | 0.22 | 0.91 | 0.91 |
| mel | 0.33 | 0.44 | 0.67 | 0.78 |
| nv | 0.53 | 0.32 | 0.79 | 0.65 |
| vasc | 0.41 | 0.62 | 0.8 | 0.87 |
| **Mean** | $0.26 \pm 0.17$ | $0.33 \pm 0.19$ | $0.72 \pm 0.13$ | $\mathbf{0.77 \pm 0.09}$ |

**(b) CUB-200-2011**

| Class | CLIP | FT CLIP | Spec. CLIP | PRISM |
|---|---|---|---|---|
| American Pipit | 0.92 | 0.99 | 0.08 | 1.00 |
| Black and white Warbler | 1.00 | 1.00 | 0.12 | 0.99 |
| Blue winged Warbler | 0.53 | 0.77 | 0.07 | 0.94 |
| Canada Warbler | 0.30 | 0.38 | 0.05 | 0.59 |
| Cape May Warbler | 0.75 | 0.63 | 0.18 | 0.74 |
| Golden winged Warbler | 0.65 | 0.82 | 0.15 | 0.97 |
| Kentucky Warbler | 0.81 | 0.97 | 0.02 | 0.98 |
| Louisiana Waterthrush | 0.41 | 0.63 | 0.16 | 0.86 |
| Magnolia Warbler | 0.85 | 0.91 | 0.20 | 0.99 |
| Mourning Warbler | 0.12 | 0.28 | 0.02 | 0.71 |
| Northern Waterthrush | 0.64 | 0.68 | 0.18 | 0.80 |
| Pine Warbler | 0.41 | 0.65 | 0.05 | 0.71 |
| Prairie Warbler | 0.49 | 0.76 | 0.03 | 0.83 |
| Tennessee Warbler | 0.22 | 0.56 | 0.05 | 0.46 |
| Yellow Warbler | 0.98 | 0.98 | 0.01 | 0.94 |
| **Mean** | $0.61 \pm 0.26$ | $0.73 \pm 0.21$ | $0.09 \pm 0.07$ | $\mathbf{0.83 \pm 0.16}$ |

**(c) PlantVillage**

| Class | CLIP | FT CLIP | Spec. CLIP | PRISM |
|---|---|---|---|---|
| Pepper bell Bacterial spot | 0.25 | 0.42 | 0.05 | 0.96 |
| Pepper bell healthy | 0.97 | 0.95 | 0.37 | 1.00 |
| Potato Early blight | 0.35 | 0.24 | 0.03 | 0.99 |
| Potato Late blight | 0.55 | 0.76 | 0.07 | 0.90 |
| Potato healthy | 0.82 | 0.96 | 0.21 | 1.00 |
| Tomato Bacterial spot | 0.00 | 0.10 | 0.07 | 1.00 |
| Tomato Early blight | 0.21 | 0.05 | 0.04 | 0.81 |
| Tomato Late blight | 0.35 | 0.79 | 0.27 | 1.00 |
| Tomato Leaf Mold | 0.41 | 0.58 | 0.09 | 0.98 |
| Tomato Septoria leaf spot | 0.22 | 0.28 | 0.09 | 0.98 |
| Tomato Spider mites | 0.03 | 0.09 | 0.11 | 0.99 |
| Tomato Target Spot | 0.11 | 0.16 | 0.11 | 0.99 |
| Tomato Yellow Leaf Curl | 0.28 | 0.56 | 0.10 | 0.98 |
| Tomato mosaic virus | 0.39 | 0.59 | 0.34 | 1.00 |
| Tomato healthy | 0.12 | 0.34 | 0.15 | 0.94 |
| **Mean** | $0.34 \pm 0.28$ | $0.46 \pm 0.30$ | $0.14 \pm 0.10$ | $\mathbf{0.97 \pm 0.05}$ |

## C QUALITATIVE RESULTS FOR PLANTVILLAGE DATASET

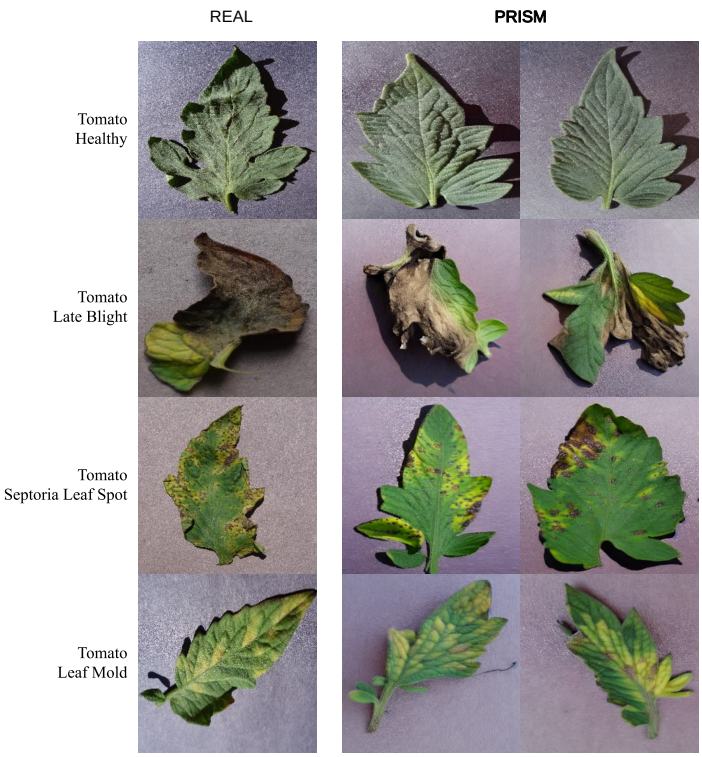



Figure 8: Qualitative results on PlantVillage. For each class, we compare a real reference image (left) with images generated by our method (right), illustrating class consistency.

## D SEGMENTATION

Table 5: Per-species IoU of the summary token $\tau_{\text{sum}}$ cross-attention map against ground-truth segmentation masks on real CUB-200-2011 validation images. The binary foreground mask is obtained by applying Otsu thresholding to the normalized attention map, followed by morphological closing to fill small holes. Results are averaged over five validation images per species. The random baseline is approximately 0.20, corresponding to the fraction of image area occupied by the bird.

| **Species** | **IoU** |
|---|---|
| American Pipit | 0.66 |
| Black and white Warbler | 0.69 |
| Blue winged Warbler | 0.53 |
| Canada Warbler | 0.63 |
| Cape May Warbler | 0.68 |
| Golden winged Warbler | 0.62 |
| Kentucky Warbler | 0.65 |
| Louisiana Waterthrush | 0.59 |
| Magnolia Warbler | 0.68 |
| Mourning Warbler | 0.72 |
| Northern Waterthrush | 0.53 |
| Pine Warbler | 0.65 |
| Prairie Warbler | 0.65 |
| Tennessee Warbler | 0.54 |
| Yellow Warbler | 0.67 |
| **Mean** | **0.63** |

## E EVALUATOR CONFUSION MATRICES

We report the evaluator's confusion matrix on real validation images to contextualize class consistency scores. Residual confusions (e.g., mel vs nv) reflect intrinsic dataset difficulty and limitations of the evaluator, rather than generation errors.

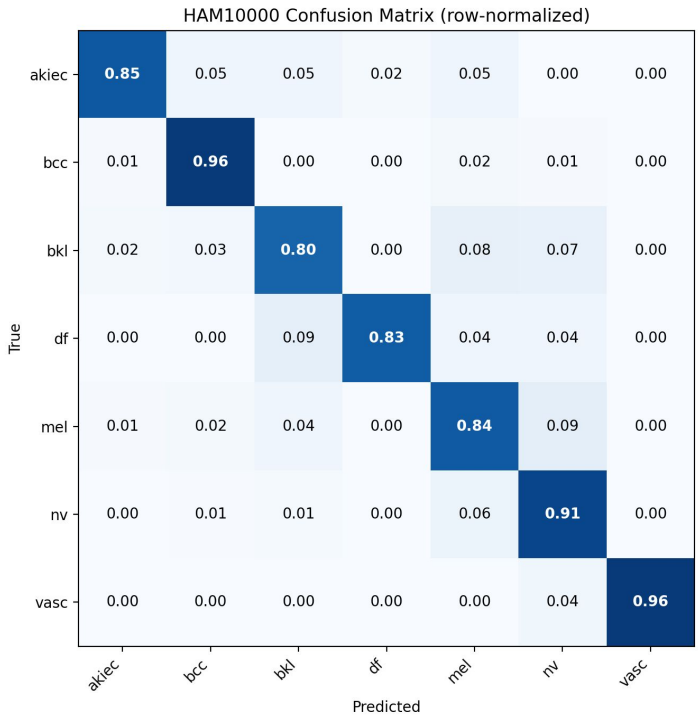



Figure 9: DinoV2 confusion matrix on HAM10000 validation images (row-normalized).

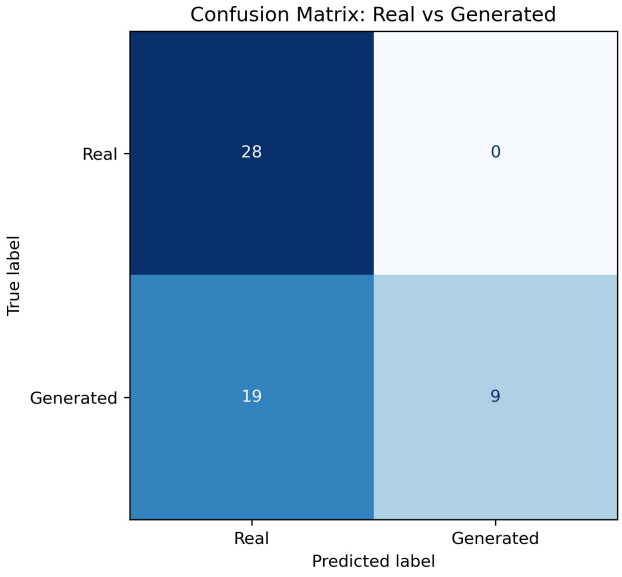



Figure 10: Confusion matrix of dermatologist annotations, comparing true labels with assigned labels for real and generated images

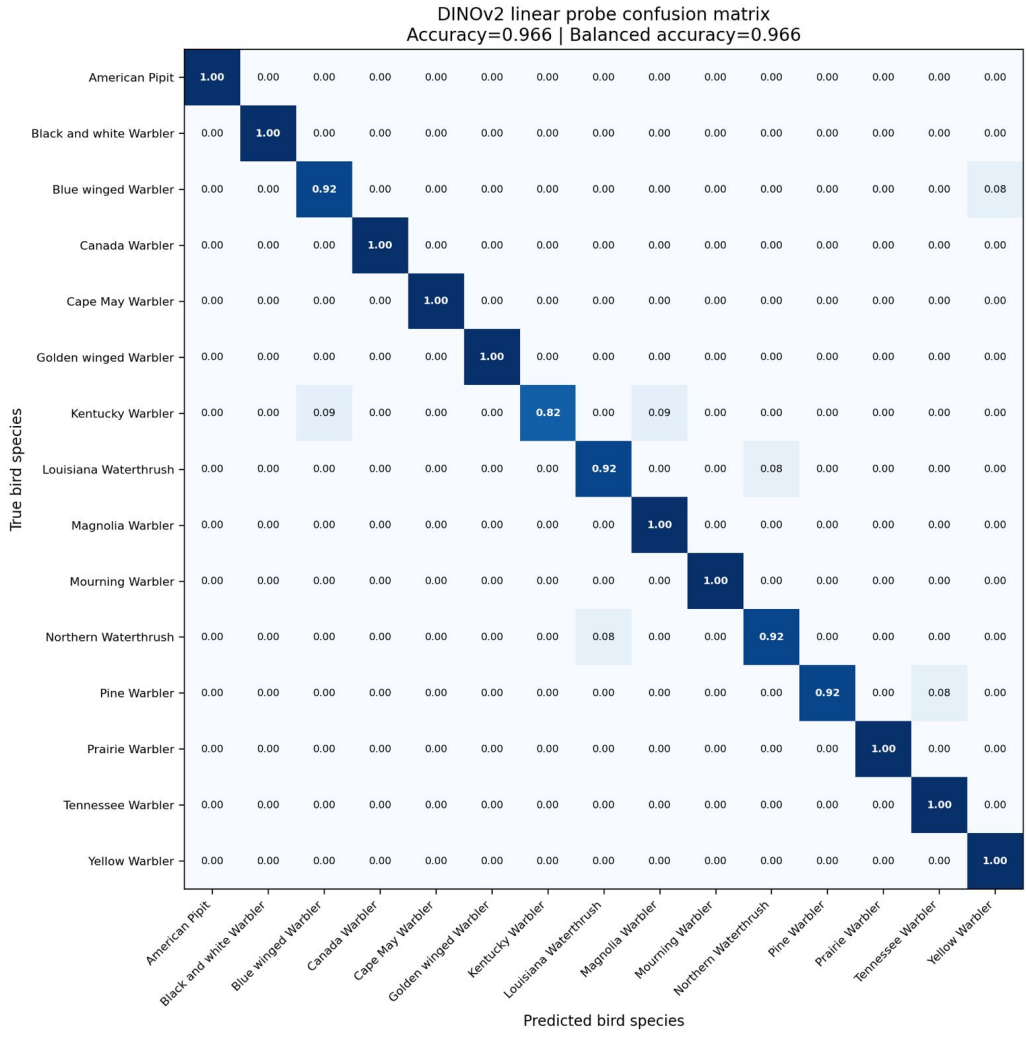



Figure 11: DinoV2 classifier confusion matrix on CUB-200-2011 validation images (row-normalized).

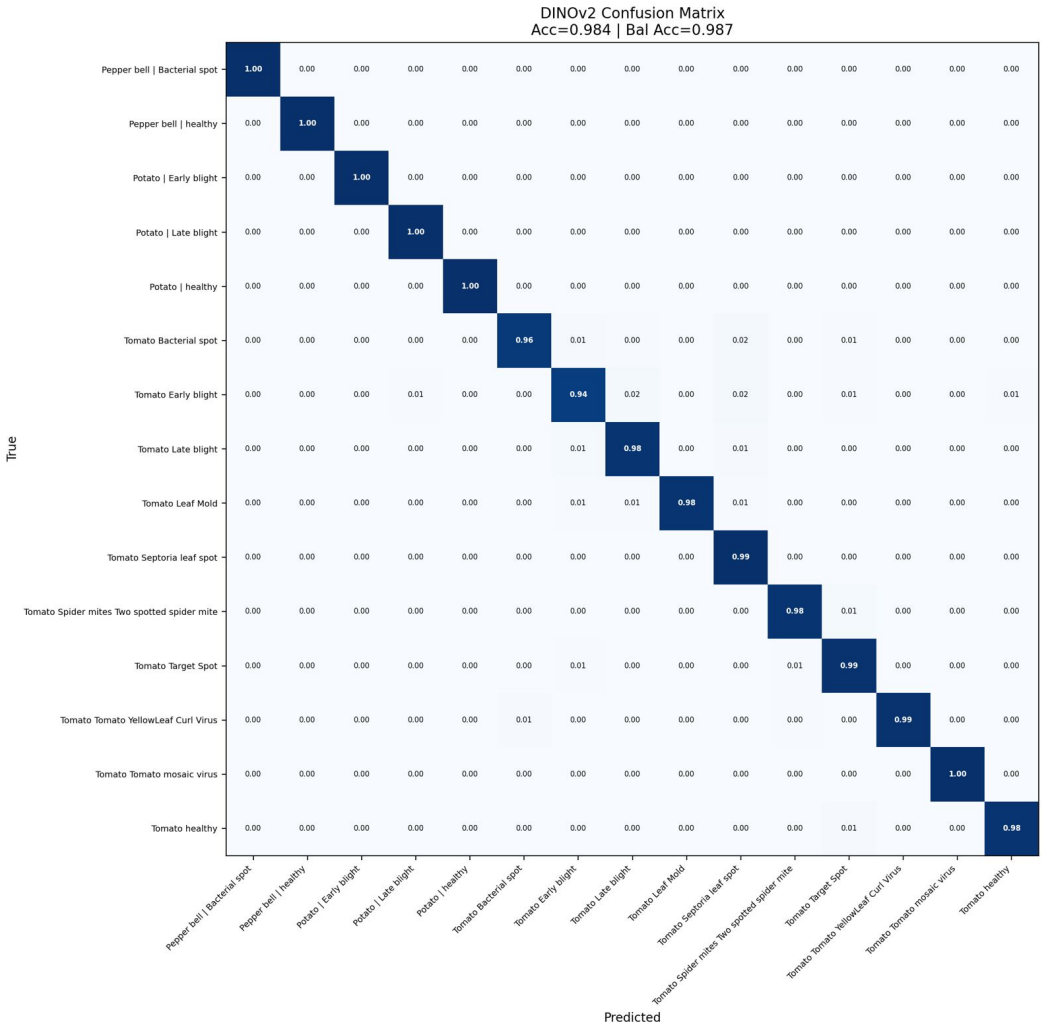



Figure 12: DinoV2 classifier confusion matrix on PlantVillage validation images (row-normalized).

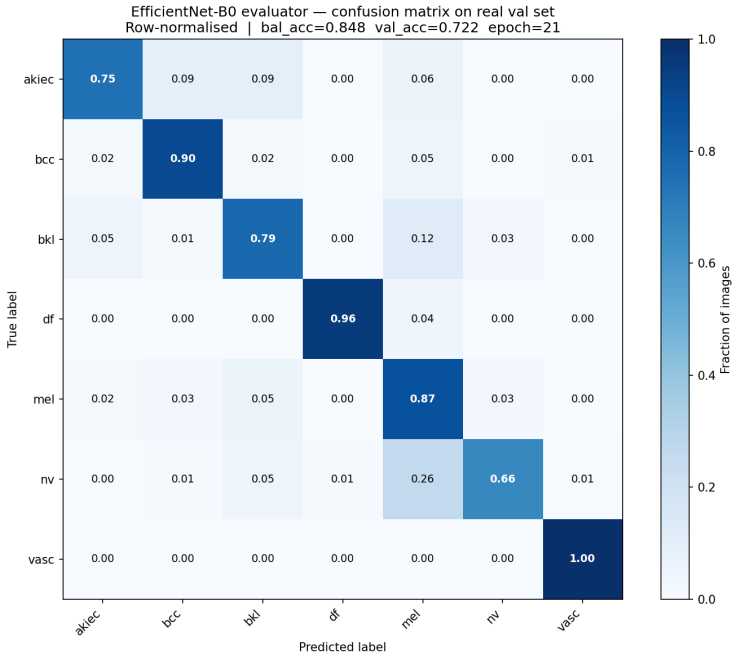



Figure 13: confusion matrix on HAM10000 validation images using EfficientNet (row-normalized).

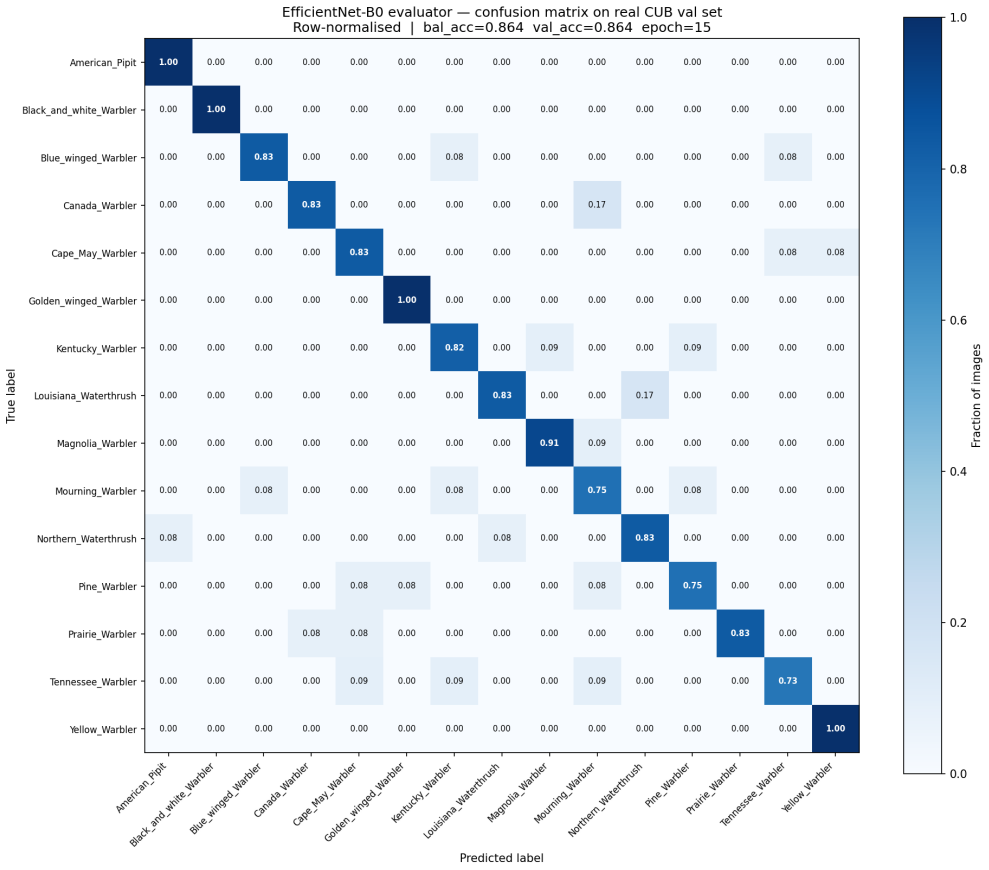



Figure 14: confusion matrix on CUB-200-2011 validation images using EfficientNet (row-normalized).

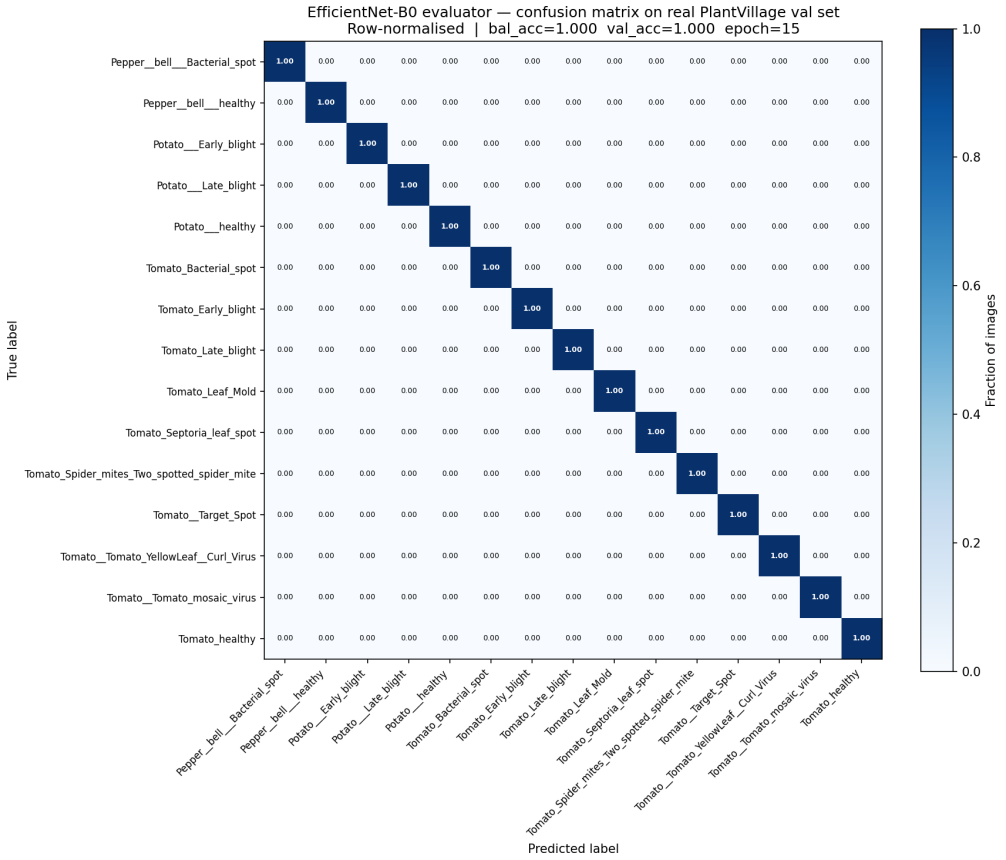



Figure 15: confusion matrix on PlantVillage validation images using EfficientNet (row-normalized).

## F NAIVE ARCHITECTURE

To isolate the contribution of per-attribute tokenization, we evaluate a baseline that encodes only the species label as a single embedding broadcast across all conditioning positions, discarding attribute information entirely. We evaluate this on CUB-200-2011, where plumage differences are visually unambiguous. Without per-attribute structure, the model produces visually incoherent outputs that fail to capture correct species morphology despite preserving coarse color statistics.

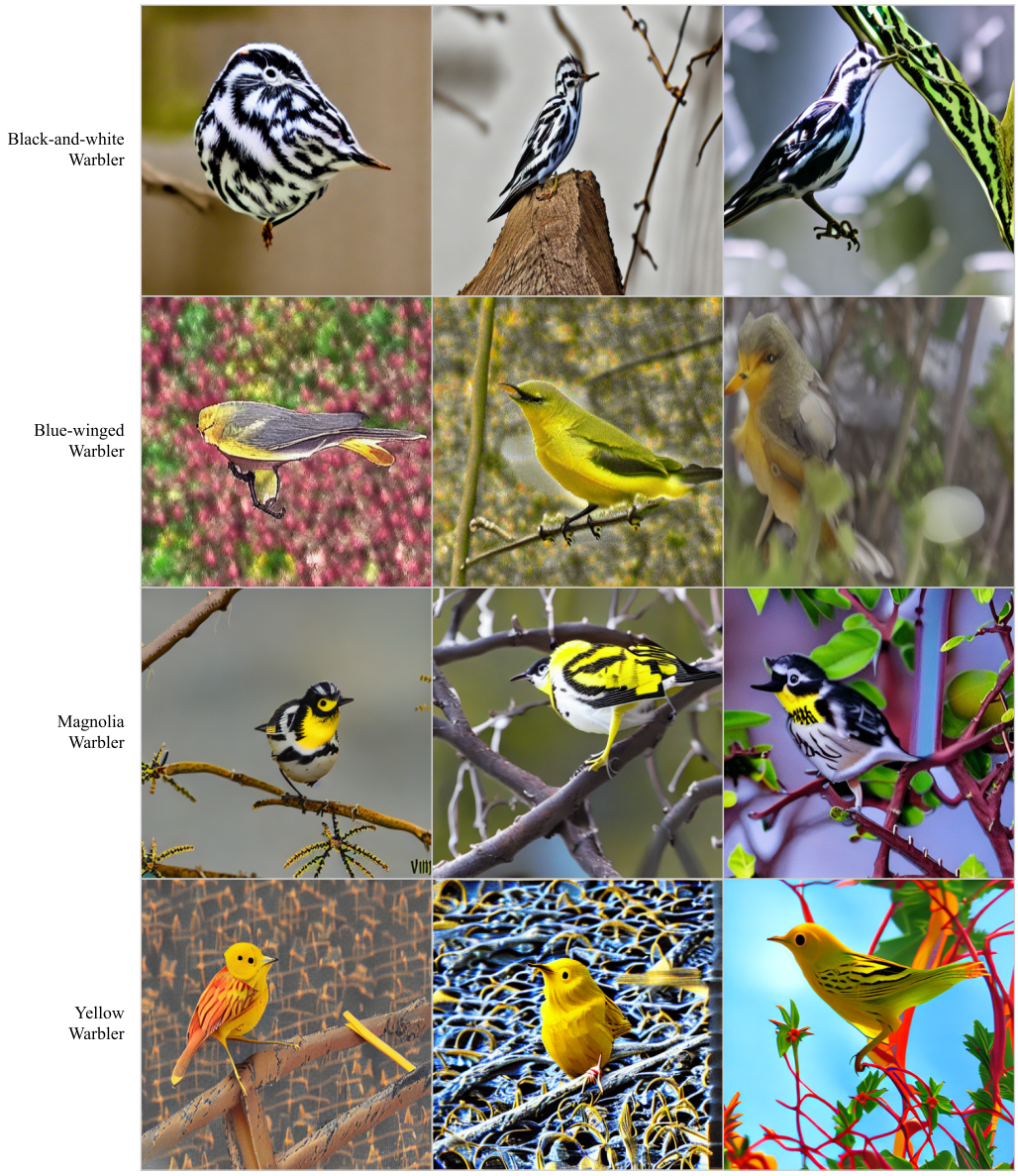



Figure 16: Qualitative results of the species-only naive baseline. Encoding class identity as a single broadcast embedding, without per-attribute structure, produces visually incoherent images that fail to capture correct species morphology.

## G PROMPT ENGINEERING

We use multiple prompt templates to convert structured attributes into natural language descriptions. Below we show a representative subset of prompts used in our experiments.

**CUB Prompt templates**

- "A wildlife photograph of a {species} with a {throat} throat, {forehead} forehead, {belly} belly, and {nape} nape."
- "A high-quality bird photograph of a {species}: {throat} throat, {forehead} forehead, {belly} belly, {nape} nape."
- "A {species} with {throat} throat and {forehead} forehead, {belly} belly, {nape} nape. Wildlife photograph."
- "Bird photo: {species} with {throat} throat coloring, {forehead} forehead, {belly} belly, and {nape} nape."
- "{species}, {throat} throat, {forehead} forehead, {belly} belly, {nape} nape."
- "{species} bird: {throat} throat, {forehead} forehead, {belly} belly, {nape} nape."

**Examples:**

- "A wildlife photograph of a Yellow Warbler with a yellow throat, yellow forehead, yellow belly, and yellow nape."
- "A Black-and-white Warbler: black throat, black forehead, white belly, black nape."

**CUB BioCLIP Scientific Templates.** BioCLIP was pretrained on over 10M iNaturalist biodiversity images using taxonomic conventions. Descriptive plumage sentences are outside its pretraining distribution. We therefore further tested using three taxonomic prompt styles:

- *Full*: "{common} {scientific}"
- *Common only*: "{common}"
- *Scientific only*: "{scientific}"

**Examples:**

- "Yellow Warbler *Setophaga petechia*"
- "Black-and-white Warbler *Mniotilta varia*"
- "*Geothlypis formosa*" (Kentucky Warbler, scientific only)

At inference, we always use the full format (common + scientific name).

**PlantVillage Prompt Templates.**

- "A close-up photograph of a {plant} leaf showing {condition}."
- "A high-quality plant disease photograph of a {plant} leaf with {condition}."
- "A detailed agricultural photograph of a {plant} plant leaf affected by {condition}."
- "A {plant} leaf with {condition}. Plant disease photograph."
- "Plant photo: {plant} leaf affected by {condition}."
- "{plant} leaf, {condition}."
- "{plant} plant: {condition} leaf."

**Examples:**

- "A close-up photograph of a tomato leaf showing bacterial spot."

- "A detailed agricultural photograph of a potato plant leaf affected by early blight."

For samples without disease, we use a separate set of templates to explicitly indicate the absence of pathology.

- "A close-up photograph of a healthy {plant} leaf."
- "A high-quality photograph of a {plant} plant with a healthy green leaf."
- "A detailed agricultural photograph of a healthy {plant} leaf."
- "A healthy {plant} leaf. Plant photograph."
- "Plant photo: {plant} leaf, healthy and green."
- "{plant} leaf, healthy."
- "{plant} plant: healthy leaf."

**Examples:**

- "A close-up photograph of a healthy tomato leaf."
- "Potato plant: healthy leaf."

**PlantVillage BioCLIP Prompt Templates.** BioCLIP encodes plant taxa from iNaturalist using scientific names. Disease common names such as "late blight" or "bacterial spot" are outside its pretraining distribution. We therefore also tested useing the pathogen's scientific names. We use three prompt styles:

- *Full*: "{common plant} {scientific plant} {scientific pathogen}"
- *Common*: "{common plant} {scientific pathogen}"
- *Scientific*: "{scientific plant} {scientific pathogen}"

**Examples (diseased):**

- "Tomato *Solanum lycopersicum Phytophthora infestans*" (late blight)
- "Potato *Alternaria solani*" (early blight, common style)
- "*Capsicum annuum Xanthomonas vesicatoria*" (pepper bacterial spot, scientific style)
- "Tomato *Solanum lycopersicum Septoria lycopersici*" (Septoria leaf spot)

**Examples (healthy):**

- "Tomato *Solanum lycopersicum*"
- "*Solanum tuberosum*, healthy" (potato, scientific style)

**HAM10000 Prompt Templates (Full Metadata).**

- "Dermoscopy image of a {dx} on the {site} of a {age}-year-old {sex}."
- "Dermoscopic photograph showing a {dx} on the {site} in a {age}-year-old {sex}."
- "Skin lesion ({dx}) located on the {site} of a {age}-year-old {sex}."

**Examples:**

- "Dermoscopy image of a melanoma on the back of a 45-year-old female."
- "Dermoscopic photograph showing a basal cell carcinoma on the face in a 67-year-old male."
- "Skin lesion (actinic keratosis) located on the scalp of a 72-year-old male."

## H ATTENTION MAPS OF PLANTVILLAGE

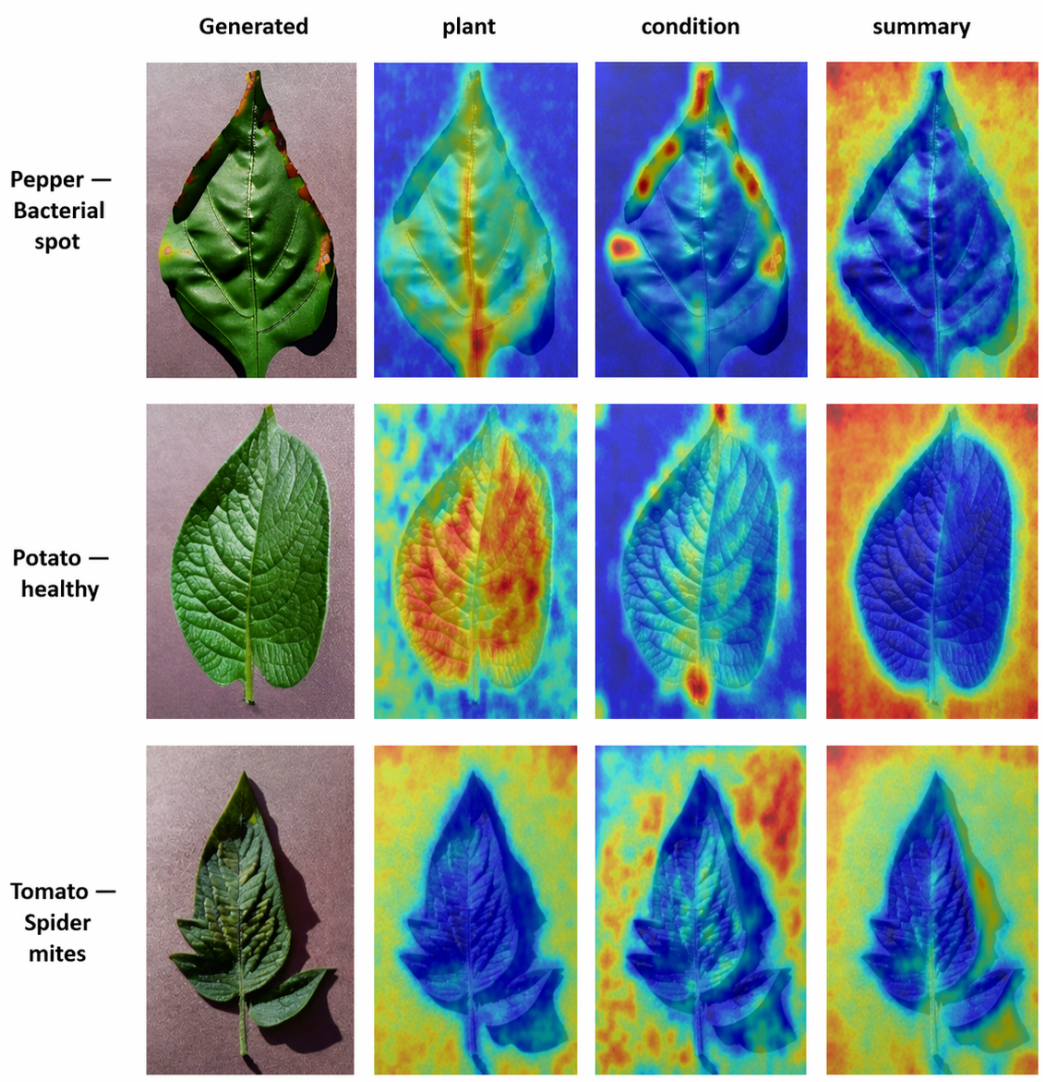



Figure 17: Cross-attention maps for PlantVillage-generated images. Each row shows a generated image alongside the spatial attention of three metadata tokens, `plant`, `condition`, and `summary`, extracted from the UNet cross-attention layers.

# I Analysis of Contrastive Refinement for Class Consistency

Table 6

**(a) HAM10000**

| Class | PRISM S1 | PRISM S1 More Steps | PRISM S2 |
|---|---|---|---|
| akiec | 0.69 | 0.65 | 0.66 |
| bcc | 0.53 | 0.56 | 0.53 |
| bkl | 0.72 | 0.63 | 0.62 |
| df | 0.91 | 0.89 | 0.91 |
| mel | 0.73 | 0.74 | 0.74 |
| nv | 0.68 | 0.73 | 0.80 |
| vasc | 0.86 | 0.90 | 0.86 |
| **Mean** | $0.731 \pm 0.116$ | $0.729 \pm 0.119$ | $0.732 \pm 0.126$ |

**(b) CUB-200-2011**

| Class | PRISM S1 | PRISM S1 More Steps | PRISM S2 |
|---|---|---|---|
| American Pipit | 0.99 | 1.00 | 1.00 |
| Black and white Warbler | 0.99 | 0.98 | 1.00 |
| Blue winged Warbler | 0.99 | 0.98 | 0.98 |
| Canada Warbler | 0.95 | 0.96 | 0.97 |
| Cape May Warbler | 1.00 | 0.96 | 0.96 |
| Golden winged Warbler | 0.97 | 0.99 | 0.98 |
| Kentucky Warbler | 1.00 | 1.00 | 1.00 |
| Louisiana Waterthrush | 0.99 | 0.99 | 0.99 |
| Magnolia Warbler | 1.00 | 1.00 | 1.00 |
| Mourning Warbler | 0.96 | 0.97 | 0.95 |
| Northern Waterthrush | **0.82** | **0.78** | **0.91** |
| Pine Warbler | 0.96 | 0.95 | 0.94 |
| Prairie Warbler | 0.97 | 1.00 | 0.98 |
| Tennessee Warbler | 0.75 | 0.78 | 0.73 |
| Yellow Warbler | 0.99 | 1.00 | 0.97 |
| **Mean $\pm$ Std** | $0.955 \pm 0.069$ | $0.956 \pm 0.069$ | $0.957 \pm 0.074$ |

**(c) PlantVillage**

| Class | PRISM S1 | PRISM S1 More Steps | PRISM S2 |
|---|---|---|---|
| Pepper bell Bacterial spot | 0.86 | 0.66 | 0.92 |
| Pepper bell healthy | 1.00 | 1.00 | 1.00 |
| Potato Early blight | 0.97 | 0.95 | 1.00 |
| Potato Late blight | 0.94 | 0.90 | 0.90 |
| Potato healthy | 1.00 | 1.00 | 1.00 |
| Tomato Bacterial spot | 0.96 | 0.95 | 0.99 |
| Tomato Early blight | 0.84 | 0.90 | 0.75 |
| Tomato Late blight | 0.99 | 0.99 | 0.98 |
| Tomato Leaf Mold | 1.00 | 0.99 | 0.98 |
| Tomato Septoria leaf spot | 0.85 | 0.69 | 0.67 |
| Tomato Spider mites | 0.87 | 0.92 | 0.84 |
| Tomato Target Spot | 0.98 | 0.98 | 0.96 |
| Tomato Yellow Leaf Curl | 1.00 | 1.00 | 1.00 |
| Tomato mosaic virus | 0.97 | 0.98 | 0.95 |
| Tomato healthy | 0.98 | 0.97 | 0.93 |
| **Mean** | $0.947 \pm 0.058$ | $0.925 \pm 0.104$ | $0.925 \pm 0.096$ |

## J SCIENTIFIC VS DESCRIPTIVE PROMPTS

Table 7: **Effect of prompt style on BioCLIP baseline performance.** DINOv2 class consistency for BioCLIP generated images using descriptive versus scientific prompts. Neither style yields strong class consistency, confirming that prompt engineering alone is insufficient.

| Dataset | BioCLIP Descriptive | BioCLIP Scientific |
|---|---|---|
| CUB-200-2011 | 0.08±0.07 | 0.07±0.07 |
| PlantVillage | 0.12±0.09 | $0.18 \pm 0.13$ |

## K STAGE 2 DIAGNOSIS

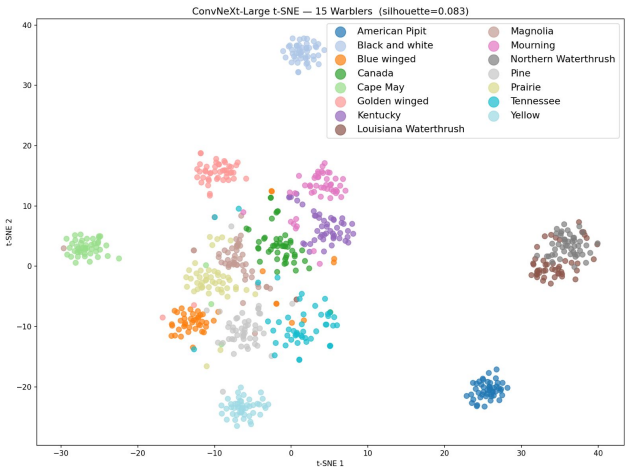



Figure 18: t-SNE visualization of ConvNeXt-Large image features for 15 warbler species. Each point represents an image, colored by species. Clusters indicate how well the model separates species in feature space.

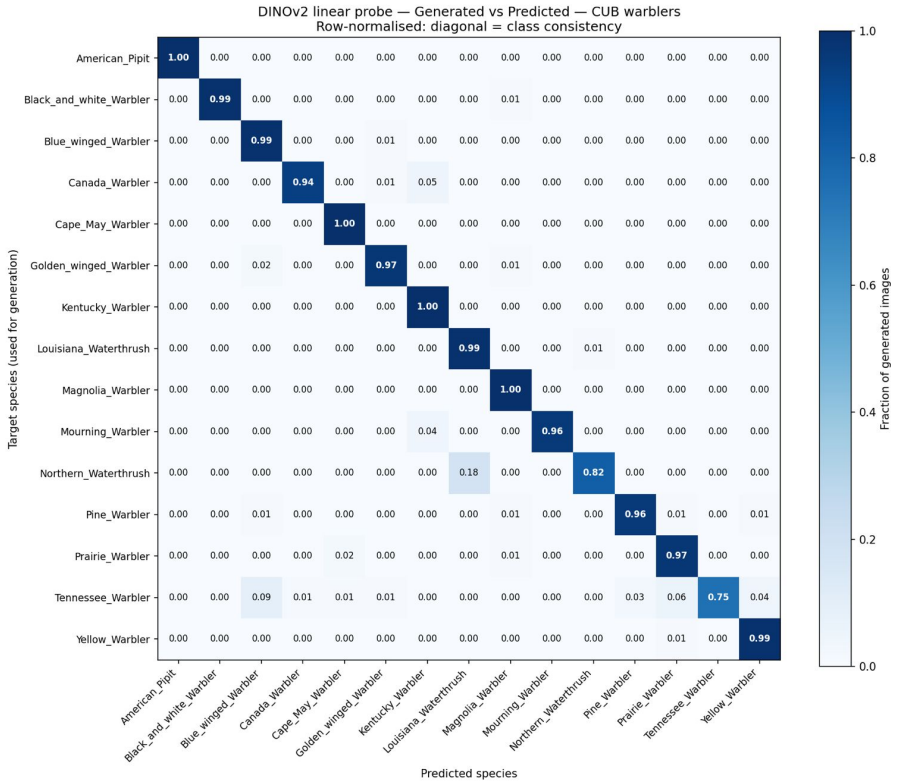



Figure 19: Row-normalized confusion matrix of a DINOv2 linear probe on Stage 1 outputs, used to derive metadata-weighted contrastive loss in Stage 2. Diagonal values indicate class consistency, while off-diagonal entries reveal visually confusable species. Notably, Louisiana Waterthrush and Northern Waterthrush exhibit significant mutual confusion, motivating higher weighting as hard negatives




Figure 20: Hard pair targeted by Stage 2. Louisiana and Northern Waterthrush are visually near-identical and share closely related metadata attributes, causing them to form overlapping clusters in both the tabular conditioning embedding space and the consistency supervisor's visual feature space (see Figures 18 and 22). Stage 2 applies contrastive refinement specifically to such confusable pairs, using the consistency supervisor as a training signal to push their class representations apart and improve generation fidelity.

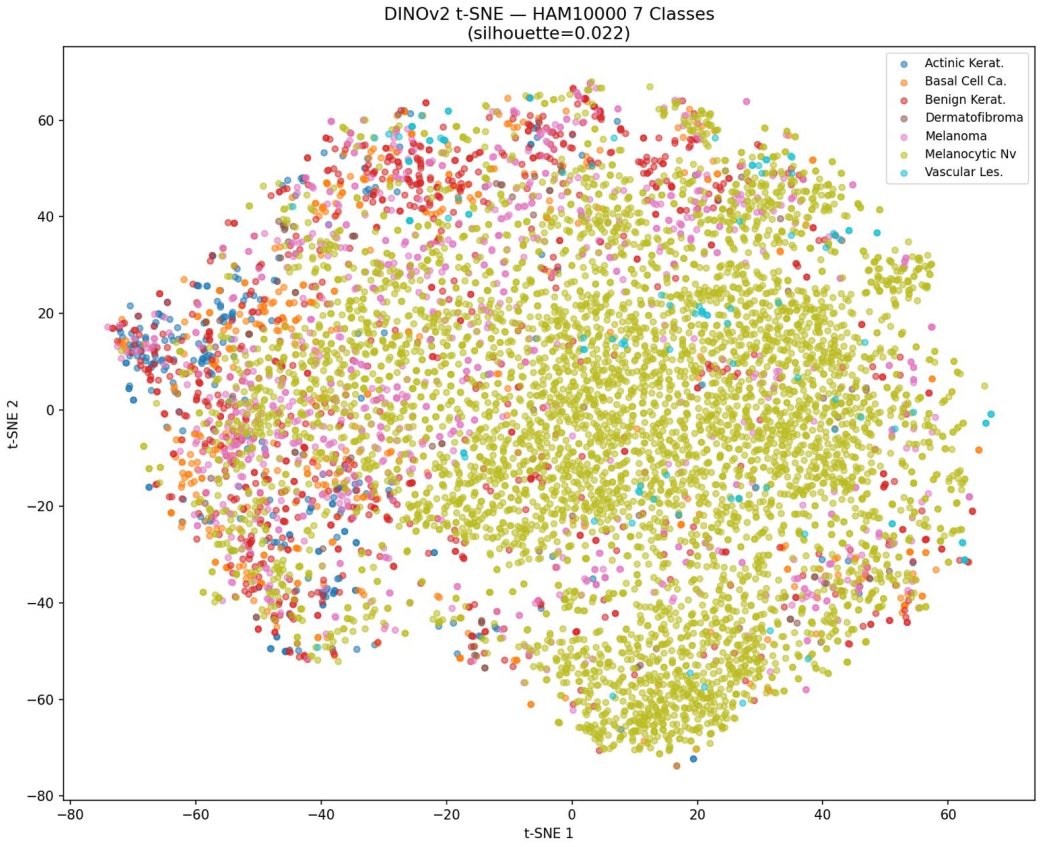



Figure 21: No visual clustering in HAM10000. Unlike CUB (Figure 18), the seven disease categories form no distinguishable clusters in the consistency supervisor's feature space (silhouette = 0.022), with all classes completely intermixed. This reflects the inherent visual ambiguity of dermoscopic images, where class distinctions are too subtle for a general-purpose visual encoder to separate.

# L PRISM EMBEDDING SPACE

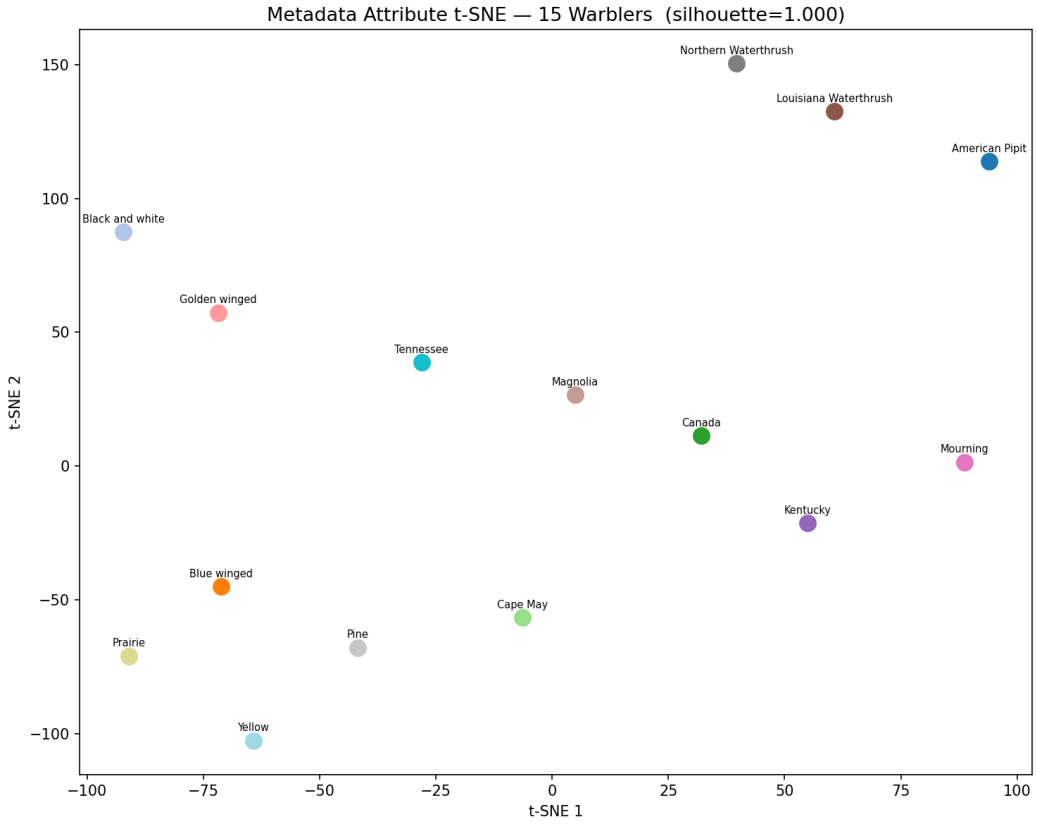



Figure 22: t-SNE of L2-normalized class-level summary embeddings $\tau^i_{\text{sum}}$ for 15 warbler species (one point per class, computed from modal metadata attributes). Silhouette score $= 1.00$ indicates perfect separation,emerging without any explicit metric-learning objective.

## M REALISM VS CLASS CONSISTENCY AT INFERENCE

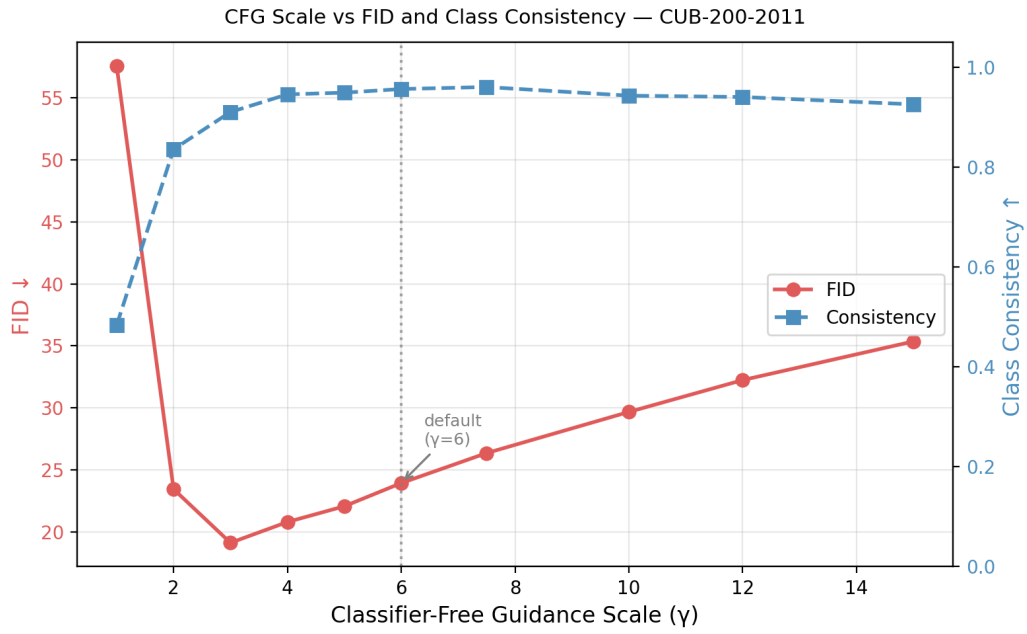



Figure 23: **CFG scale trade-off on CUB-200-2011.** Class consistency (blue) saturates by $\gamma \approx 4$ and remains stable, while FID (red) reaches its minimum near $\gamma = 3$ before rising as over-conditioning degrades realism. The default $\gamma = 6$ (dotted line) balances both objectives.

## N Diversity, Memorization, and Robustness to Imperfect Metadata

**Effect of Stage 2 on diversity and memorization.** On Stage 2's effect on diversity and memorization, we compare Stage 1 and Stage 2 using DINOv2 features. Across all 15 CUB species, gen-to-real nearest-neighbor ratios stay between 0.99 and 1.31, well above the 0.7 memorization threshold, for both stages. Intra-class diversity drops only modestly from Stage 1 to Stage 2, and consistently across every class rather than a few outliers.

Table 8: Effect of Stage 2 refinement on generation diversity and memorization on CUB-200-2011, measured with DINOv2 features. Gen-to-real nearest-neighbor ratios well above the 0.7 memorization threshold indicate that generated images are not near-duplicates of training images.

| | **Mean gen→real NN ratio** | **Mean intra-class diversity** |
|---|---|---|
| Stage 1 | 1.115 | 38.669 |
| Stage 2 | 1.088 | 36.904 |
| Memorization threshold | 0.700 | – |

**Robustness to imperfect metadata.** On robustness to imperfect metadata, we corrupt 0, 2, or all 3 auxiliary attributes (lesion site, sex, age) at inference on HAM10000, replacing them with missing embeddings while keeping the diagnosis label intact, and measure class consistency with EfficientNet-B0. Consistency degrades gradually and does not collapse even with all three attributes missing, indicating the model relies primarily on the diagnosis label for identity and treats auxiliary attributes as refinement rather than a hard requirement.

Table 9: Class consistency on HAM10000 as auxiliary attributes (lesion site, sex, age) are corrupted at inference, measured with EfficientNet-B0. The diagnosis label is kept intact in all cases.

| **Corrupted attributes** | **EfficientNet-B0** |
|---|---|
| 0 | 0.833 |
| 2 | 0.805 |
| 3 (all) | 0.729 |

# O DATASETS AND METRICS

**HAM10000.** Large collection of $10{,}015$ dermoscopy images spanning seven skin lesions: `akiec`, `bcc`, `bkl`, `df`, `mel`, `nv` and `vasc`. The dataset exhibits pronounced class imbalance: the most frequent class, `nv` , contains $6{,}705$ images while the rarest class, `df`, contains only $115$ images. Each image is accompanied by four structured attributes: diagnosis label, anatomical site of the lesion, patient sex, and patient age forming the conditioning signal $m$.

**CUB-200-2011** We select 15 similarly named species (e.g., warblers), where shared terminology is expected to induce proximity in CLIP text embeddings, to challenge its ability to distinguish fine-grained classes. To choose conditioning attributes, we group features and compute per-species distributions from the dominant attribute per image. Discriminability is measured with pairwise Jensen–Shannon divergence, with higher values indicating greater separability. We select the top 4 from distinct body regions to minimize redundancy: throat, forehead, belly and nape color. Together with the species label, these four attributes define the conditioning input $m$.

**PlantVillage.** We use a 15-class subset covering infections as well as healthy states across three crop species: pepper, potato and tomato. The resulting subset contains approximately $20{,}449$ images with moderate class imbalance: the most frequent class, Tomato Yellow Leaf Curl Virus, contains $3{,}209$ images while the rarest class, Potato healthy, contains only $152$ images. Each image is accompanied by two structured attributes: the crop species and the health status label, which together constitute the conditioning signal $m$.

**Class Consistency.** Measures whether generated images are faithful to the class specified by the conditioning signal. For each generated image intended to depict class $d$, we pass it through a classifier trained independently on real images and record whether the predicted class matches $d$. Class consistency is defined as the proportion of generated images conditioned on class $d$ that are correctly predicted as class $d$ by the evaluator. To ensure architecture-independent evaluation, we measure class consistency using external classifiers. We use DINOv2 for CUB and PlantVillage, and ConvNeXt for HAM10000, as both are trained to accurately classify real images.

**Image Realism.** We evaluate all methods using two image quality metrics: CLIPIQA Wang et al. (2022a) and FID. For HAM10000, we additionally conduct a blind evaluation with a certified dermatologist. The expert is presented with 56 images, 28 real and 28 generated, randomly interleaved covering all classes evenly. For each image, the expert first provides a diagnosis label, and only then judges whether the image is real or synthetic, ensuring that knowledge of the generation status does not contaminate the diagnostic assessment. We report two complementary metrics: the fool rate, defined as the fraction of generated images judged as real; and expert diagnostic accuracy, defined as the fraction of generated images for which the expert's diagnosis matches the conditioning label.

# P IMPLEMENTATION DETAILS

All experiments are implemented in PyTorch. Tables 10–15 summarize every hyperparameter and architectural choice used across all three stages.

Table 10: Base model and shared settings.

| **Parameter** | **Value** |
|---|---|
| Base model | Stable Diffusion v1.5 |
| Image resolution | $512 \times 512$ |
| Text encoder | CLIP ViT-L/14 |
| Conditioning sequence length | 77 tokens |
| Conditioning token dimension | 768 |
| Latent spatial resolution | $64 \times 64 \times 4$ |
| VAE scaling factor | 0.18215 |
| Noise scheduler | DDIM |
| Inference steps | 50 |
| Precision | fp16 |

Table 11: Tabular Conditioning Encoder architecture.

| **Parameter** | **Value** |
|---|---|
| Per-attribute embedding dim | 32 |
| Attribute types supported | Categorical (lookup), Continuous (MLP + SiLU) |
| Per-attribute projection | Linear(32, 768) + LayerNorm(768) |
| Summary input dim | $32 \times N_{\text{fields}}$ (concatenated) |
| Summary projection | Linear $\rightarrow$ SiLU $\rightarrow$ Linear(768) $\rightarrow$ LayerNorm |
| Summary hidden dim | 256 |
| Content tokens | $N_{\text{fields}}$ attribute + 1 summary |
| Padding tokens | $77 - (N_{\text{fields}} + 1)$, learnable $\mathcal{N}(0, 0.02)$ |
| Activation function | SiLU |

Table 12: Stage 1: LoRA fine-tuning.

| **Parameter** | **Value** |
|---|---|
| LoRA rank | 16 |
| LoRA target modules | `to_q, to_k, to_v, to_out` |
| CFG dropout probability | 0.15 |
| Null conditioning | $\mathbf{0} \in \mathbb{R}^{1\times77\times768}$ |
| Training steps | Until Convergence |
| Batch size | 4 |
| Learning rate | $3 \times 10^{-5}$ |
| Optimizer | AdamW |
| Class sampling | Inverse-frequency weighted sampler |

Table 13: Stage 2: Similarity-Aware Class Refinement.

| **Parameter** | **Value** |
|---|---|
| Consistency supervisor | ConvNeXt-Large (frozen) |
| Negative weighting temperature $\tau$ | 0.1 |
| Contrastive margin $\delta$ | 0.10 |
| $\lambda_{\text{diff}}$ | 1.0 |
| $\lambda_{\text{sep}}$ (start $\rightarrow$ end) | $0.02 \rightarrow 0.40$ |
| $\lambda_{\text{sep}}$ warmup steps | 1,000 |
| Contrastive interval | every 10 steps |
| Reward denoising steps | 6 (last steps of 50-step DDIM) |
| Gradient clipping | max norm 1.0 |
| CFG dropout probability | 0.15 |
| Training steps | 5,000 |
| Batch size | 4 |
| Learning rate | $5 \times 10^{-6}$ |
| Optimizer | AdamW |
| Random seed | 42 |

Table 14: Stage 3: Classifier-Free Guidance.

| **Parameter** | **Value** |
|---|---|
| Guidance scale $\gamma$ | 6.0 |
| CFG sweep range | 1.0 – 15.0 |
| Null conditioning | $\mathbf{0} \in \mathbb{R}^{1\times77\times768}$ |
| Inference steps | 50 |

Table 15: Evaluation settings.

| **Parameter** | **Value** |
|---|---|
| Class consistency evaluator | DINOv2 ViT-B/14 + linear probe |
| DINOv2 feature dimension | 768 |
| DINOv2 input resolution | $224 \times 224$ |
| Linear probe | Frozen backbone, trained on real train split |
| FID feature extractor | Inception-v3 (pool3 layer) |
| FID input resolution | $299 \times 299$ |
| FID batch size | 64 |
| Generated images per class | 300 (HAM10000), 100 (CUB-200-2011 & PlantVillage) |
| Hardware | NVIDIA RTX 4090 |
| Total runtime | $\approx$10 hours (all stages, generation, evaluation) |