DGTUNE: DISCRIMINATIVE GEOMETRY TUNING FOR NOISE-AFFECTED MEDICAL VISION–LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Pre-training on large-scale datasets followed by fine-tuning on downstream tasks has become a popular paradigm for medical vision–language models. However, medical descriptions and answers provide imperfect upstream supervision, raising a fundamental question: does clean downstream learning overcome the noise inherited from pre-training? We investigate this problem through a controlled family of medical vision–language models trained with different levels of supervision noise. Our experiments reveal persistent downstream performance deficits, including under distribution shift, despite clean adaptation labels. Through Fisher analysis, we link these persistent deficits to weakened class-discriminative structure in the adapted representations. Crucially, strengthening discrimination on observed samples alone does not ensure that the same discriminative relations generalize to new samples. Motivated by this distinction, we propose Discriminative Geometry Tuning (DGTune), a lightweight adaptation method that jointly strengthens weak Fisher directions and promotes their cross-sample generalization. Its two complementary regularizers enhance weak class separation and preserve the corresponding directions and class-contrast patterns across balanced, disjoint sample groups. Experiments on controlled noisy and public medical models show improved in-distribution and out-of-distribution generalization with DGTune.

1 INTRODUCTION

Large-scale pre-training followed by task-specific adaptation has become a central paradigm in machine learning (Kornblith et al., 2019; Dosovitskiy et al., 2021; Radford et al., 2021). In medicine, image–text learning draws on radiology reports and biomedical publications to acquire transferable visual representations (Huang et al., 2021; Zhang et al., 2022; Wang et al., 2022; Zhang et al., 2023). Generative medical vision–language models (VLMs) extend this paradigm to image-grounded dialogue and question answering, learning from biomedical image–text corpora and, increasingly, medical instruction data (Li et al., 2023; Moor et al., 2023; Chen et al., 2024b).

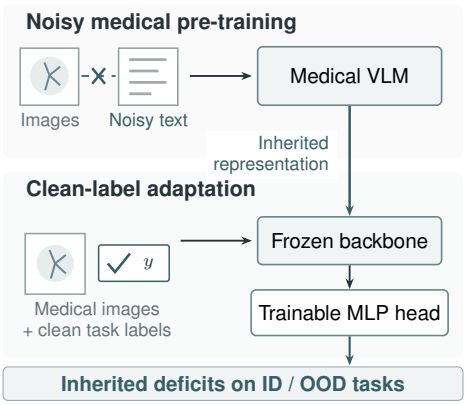

Figure 1: **Problem setting.** Upstream supervision noise can persist through a frozen medical VLM despite downstream adaptation with clean labels.

Yet scaling pre-training does not ensure reliable supervision. Vision–language models can learn from weakly curated image–text pairs (Jia et al., 2021), while data filtering substantially affects representation quality under a fixed training procedure (Gadre et al., 2023). This concern is particularly consequential in medicine, where images from different disease categories can share similar anatomy, with diagnostic distinctions depending on subtle pathological features (Zhang et al., 2022). Biomedical captions may also describe multiple panels or clinical context beyond the paired image (Chen et al., 2024b). A mismatched description can therefore remain

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

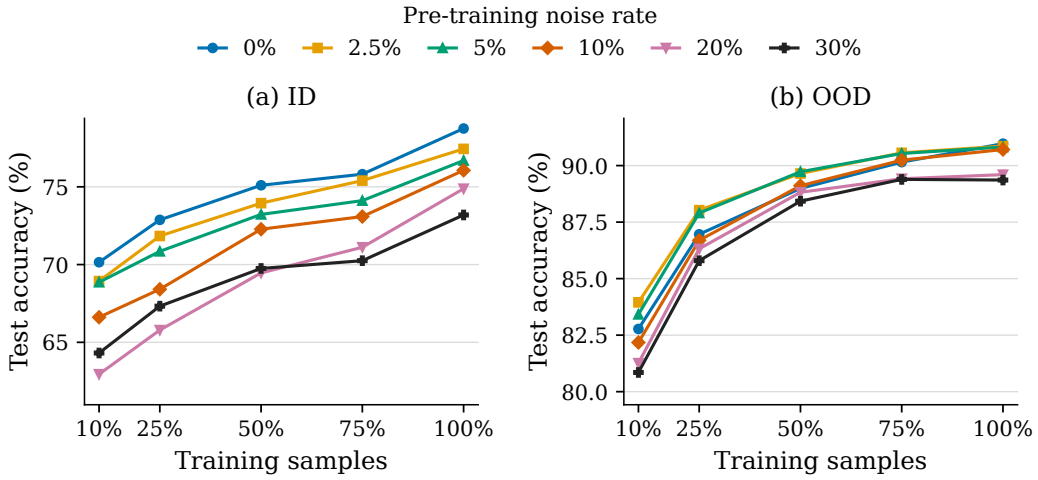

Figure 2: **Noise inheritance across downstream data volumes.** MLP adaptation on Kvasir-v2 (ID) and the cross-center PathMNIST-224 test set (OOD).

anatomically plausible while misrepresenting the finding that distinguishes disease states, potentially preserving coarse medical semantics while weakening fine-grained discrimination. Specialist annotation costs and privacy constraints further limit access to corrective supervision (Chen et al., 2024b). These imperfections enter the model before downstream adaptation, when users may have clean task labels but neither the upstream data nor the resources to repeat medical pre-training. A fundamental question is therefore whether *clean downstream supervision can overcome the effects of noisy medical pre-training*.

The difficulty lies in where the noise resides. Conventional noisy-label learning addresses corrupted task supervision through robust losses or sample selection (Ghosh et al., 2017; Han et al., 2018; Li et al., 2020). Related studies show that downstream adaptation itself can alter pre-trained features and their robustness under distribution shift (Kumar et al., 2022; Wortsman et al., 2022). Here, however, our concern is the imperfect upstream supervision already encoded in a model before clean downstream learning begins. Prior work on noisy model learning has established the importance of this distinction for transfer (Chen et al., 2024a). For medical VLMs, it concerns visual associations learned from medical descriptions and answers, not only categorical labels. The challenge is to mitigate the effects of inherited noise when adapting medical models to downstream tasks.

We investigate this problem directly in *billion-parameter medical VLMs*. Starting from Qwen2.5-VL-7B (Bai et al., 2025), we perform full-parameter, two-stage medical training on PubMedVision (Chen et al., 2024b), constructing six models with controlled upstream noise levels spanning 0% to 30%. Noise is introduced into medical captions and answers during model construction, while downstream supervision remains unchanged. This study reveals that adverse effects of upstream noise can persist after clean-label adaptation, including under distribution shift. We call this persistence *downstream noise inheritance*: clean labels provide a learning signal, but do not automatically erase the influence of the model's pre-training history. Figure 2 illustrates this persistence as more clean labels become available for within-dataset adaptation and cross-center transfer.

To investigate the mechanisms underlying this persistence, we adopt a Fisher-geometric view of adaptation (Fisher, 1936). Prior work connects spectral geometry to transferability (Chen et al., 2019) and strengthens weak discriminant directions (Dorfer et al., 2016). Our focus is the distinction between *Fisher discriminative strength on observed samples* and *its generalization to new samples*. Weak Fisher eigenvalues indicate limited class separation relative to within-class variation. Increasing them, however, need not yield generalizable discriminants: equally strong spectra can encode different class-separating directions. The crucial question is whether the same direction and class-contrast pattern remain discriminative on another sample group. We therefore couple weak-spectrum enhancement with cross-sample generalization of Fisher discriminants.

Building on this perspective, we propose *Discriminative Geometry Tuning* (DGTune), a lightweight adaptation method with two complementary regularizers. *Weak-spectrum strengthening* increases the weakest Fisher eigenvalues, enhancing class separation relative to within-class variation. *Cross-sample Fisher transfer* fixes the weak discriminant directions and their associated class-contrast patterns estimated from one sample group, then encourages their discriminative strength on another group. Applied bidirectionally to class-balanced, disjoint groups, this term promotes cross-sample Fisher generalization rather than merely matching independently estimated spectra. Together, the two terms strengthen weak discriminative structure and encourage it to transfer beyond the samples on which it was estimated. DGTune optimizes only an MLP adaptation head on frozen medical VLM features; neither geometric regularizer adds computation at inference.

Our main contributions are summarized as follows:

- **Noise inheritance in 7B medical VLMs.** We construct six noise-controlled 7B medical VLMs through full-parameter, two-stage training and demonstrate that upstream supervision noise can leave persistent downstream deficits despite clean-label adaptation.
- **Discriminative Geometry Tuning (DGTune).** We introduce two modules: *weak-spectrum strengthening* enhances weak class separation, while *cross-sample Fisher transfer* promotes generalization of the same discriminative directions and class-contrast patterns.
- **Extensive evaluation.** Experiments on models trained with synthetic supervision noise and publicly available medical backbones demonstrate DGTune's effectiveness across in-distribution adaptation and out-of-distribution transfer.

2 Understanding the Label Noise in Pre-trained Models

We examine how imperfect medical supervision affects a pre-trained model after it is adapted using clean downstream labels. Our study follows the noise from model construction to downstream prediction: we first build a controlled family of medical VLMs, compare their transfer performance under fixed adaptation protocols, and then examine the weak class-discriminative structure that remains after clean-label adaptation.

2.1 Experimental Design

Noisy medical supervision. Medical VLMs learn from textual targets rather than only categorical labels (Li et al., 2023; Chen et al., 2024b). For each medical training stage, let $\mathcal{D} = \{(v_i, q_i, a_i)\}_{i=1}^{N}$ denote image inputs, user prompts, and assistant targets. A prompt requests a detailed description during visual alignment and contains a medical question during instruction tuning. At noise rate ρ, recipients in $\mathcal{R}_\rho$ receive targets from other examples:

$$\widetilde{a}_i^{(\rho)} = \begin{cases} a_{d(i)}, & i \in \mathcal{R}_\rho, \quad d(i) \neq i, \\ a_i, & i \notin \mathcal{R}_\rho. \end{cases} \tag{1}$$

Images and user prompts remain unchanged. This intervention corrupts image–target or question–answer correspondence while retaining medical text as supervision. We use six noise rates, $\rho \in \{0, 0.025, 0.05, 0.10, 0.20, 0.30\}$, with $|\mathcal{R}_\rho| = \lfloor \rho N + 0.5 \rfloor$ and nested recipient sets within each stage. Recipient–donor pairs are retained as the noise rate increases. Here, 0% denotes the absence of *injected* noise, not an assumption that all original annotations are error-free.

Pre-trained medical models. Medical VLMs provide representations learned by connecting visual findings with language. We construct our model family from Qwen2.5-VL-7B-Instruct (Bai et al., 2025) through two stages of full-parameter training on PubMedVision (Chen et al., 2024b): image–text alignment using medical descriptions, followed by instruction tuning using medical questions and answers. At each noise level, instruction tuning builds on the corresponding alignment model, with the same nominal noise rate applied to both stages. The architecture and training protocol are otherwise shared. For downstream adaptation, we view each resulting model as a feature extractor $f_{\phi_\rho} : \mathcal{X} \to \mathbb{R}^D$, where $x \in \mathcal{X}$ pairs an image with a fixed task-specific prompt. Rather than using the language-modeling head to generate answers, we use the hidden representation $f_{\phi_\rho}(x)$ to learn a task-specific classifier from clean labels. The backbone remains frozen: linear

Noise (%)	Primary task tests (a) Kvasir-v2	(b) PathMNIST	(c) Camelyon17	(d) IDRiD	(e) Messidor-2	Cross-domain retinal transfer (f) APTOS → M2	(g) DeepDR → APTOS	(h) IDRiD → M2	(i) M2 → IDRiD
2.5	−11.4	+6.1	−8.5	+21.7	+21.4	+57.8	+70.3	+147.6	+16.6
5	−9.3	+15.8	−5.0	+37.5	+34.9	+23.0	+69.7	+102.0	+5.2
10	−12.4	−13.9	−9.8	−7.1	+21.0	+4.7	+14.5	+56.5	−29.4
20	−18.4	−7.5	−9.8	−7.9	+4.9	−27.4	+31.5	−12.8	−35.2
30	−10.5	−13.9	−20.3	−66.2	−14.1	−30.8	−22.6	+34.9	−27.6

Change in W (%)

−150 −75 0 75 150

Figure 3: **Weak Fisher strength after clean-label adaptation.** Relative change $100[W(\rho)/W(0)-1]$ results; negative values indicate weaker discrimination. The identically zero baseline is omitted. M2: Messidor-2.

probing learns a linear classifier, whereas MLP adaptation additionally learns a nonlinear feature transformation (Kornblith et al., 2019; Kumar et al., 2022). The motivating analysis uses cross-entropy alone, before introducing our regularization. This separation lets us examine how noisy medical supervision affects subsequent learning from clean labels.

ID and OOD evaluation. At each upstream noise rate ρ, we freeze the feature extractor f_{ϕ_ρ} and train a linear or MLP classifier on a clean downstream training set $\mathcal{D}'_{\mathrm{tr}} = \{(x_i, y_i)\}_{i=1}^{M}$ with K classes. ID evaluation uses separate training and test partitions from the same source distribution P_{s}, including Kvasir-v2 (Pogorelov et al., 2017) and four retinal domains—APTOS, DeepDR, IDRiD, and Messidor-2—organized using GDRBench (Che et al., 2023). OOD evaluation applies the source-trained classifier to a target distribution $P_{\mathrm{t}} \neq P_{\mathrm{s}}$ without further adaptation, using the cross-center PathMNIST-224 test set (Yang et al., 2023; Kather et al., 2019).

2.2 Results

Figure 2 compares mean downstream accuracy in ID and OOD settings across pre-training noise levels and downstream data volumes. These results reveal two main observations: **1)** High pre-training noise leaves a downstream accuracy deficit in both ID and OOD settings, despite adaptation with clean labels. Mild noise can be neutral or beneficial, but higher noise levels impair performance. **2)** More clean downstream data improve accuracy but do not eliminate the high-noise deficit: in both illustrated settings, a gap remains even when the full training set is used. These observations reveal that the influence of noisy medical pre-training persists beyond the upstream task. What properties of the learned representation account for this inherited deficit? We next examine class-discriminative structure through Fisher analysis to investigate its geometric basis.

2.3 Weak Discriminative Structure under Noisy Pre-training

To investigate why upstream noise remains consequential after clean-label adaptation, we turn from prediction accuracy to the class-discriminative structure of the learned representation. A feature direction may capture substantial variation without separating the target classes. Fisher analysis measures separation relative to within-class variation, directly connecting feature geometry to downstream labels (Fisher, 1936; Dorfer et al., 2016). We examine the frozen representation supplied to a linear probe and the hidden representation of an MLP, before introducing our method.

Measuring weak discrimination. For labeled representations $\{(z_i, y_i)\}_{i=1}^{n}$, with $z_i \in \mathbb{R}^d$ and K classes, let n_c be the number of examples in class c, $\mu_c = n_c^{-1} \sum_{i:y_i=c} z_i$ its mean, and $\bar{\mu} =$

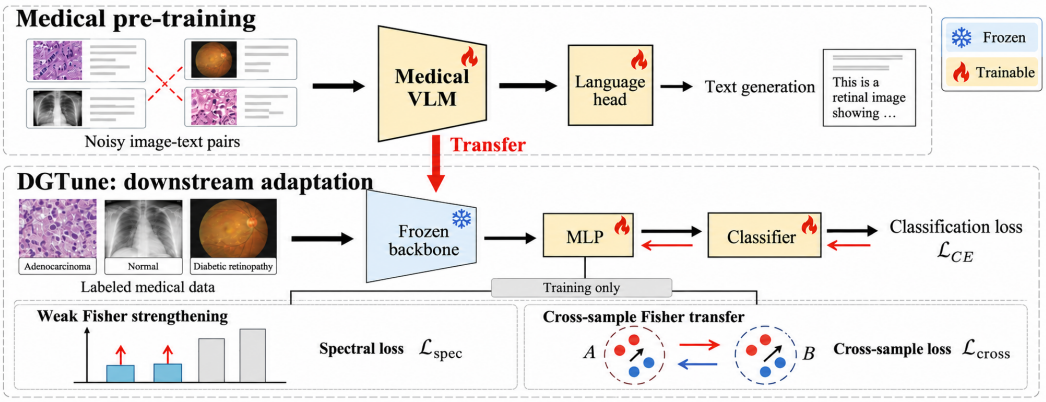

Figure 4: **DGTune framework.** Medical pre-training is followed by downstream adaptation with a frozen backbone and trainable MLP and classifier. Weak Fisher strengthening improves deficient class separation; cross-sample Fisher transfer encourages the same discriminative relations across balanced, disjoint groups A and B. Both regularizers operate only during training. Blue and yellow denote frozen and trainable modules, respectively; red backward arrows denote adaptation gradients. Medical thumbnails and geometric diagrams are illustrative.

$K^{-1}\sum_c \mu_c$ the equal-class mean. The class-mean matrix and within-class covariance are

$$M = \frac{1}{\sqrt{K}}[\mu_1 - \bar{\mu}, \ldots, \mu_K - \bar{\mu}],$$
$$\Sigma = \frac{1}{K}\sum_{c=1}^{K} \frac{1}{n_c - 1} \sum_{i:y_i=c} (z_i - \mu_c)(z_i - \mu_c)^\top. \tag{2}$$

Equal-class weighting prevents frequent classes from dominating the statistic. We stabilize the covariance with $S = (1-\alpha)\Sigma + [\alpha \operatorname{tr}(\Sigma)/d + \epsilon]I$, where α controls shrinkage and $\epsilon > 0$. Let $H \in \mathbb{R}^{K\times(K-1)}$ be an orthonormal basis for class contrasts, satisfying $H^\top H = I$ and $H^\top \mathbf{1} = 0$. The Fisher-whitened contrast matrix and weakest strength are

$$C = S^{-1/2}MH, \qquad T = C^\top C = H^\top M^\top S^{-1} M H,$$
$$W = \lambda_{\min}(T) = \min_{\|v\|_2=1} v^\top T v = \sigma^2_{\min}(C). \tag{3}$$

Whitening expresses class separation relative to within-class variation. The minimum over unit class-contrast vectors v isolates the weakest relation rather than averaging over well-separated directions; smaller W therefore indicates weaker discrimination. Zero eigenvalues are retained, so a collapsed contrast gives $W = 0$. We use every example in each original split, with sample identities fixed across noise levels and MLP checkpoints selected on validation data.

Analysis. Figure 3 shows how weakest Fisher strength changes with pre-training noise after clean-label MLP adaptation. Mild noise can increase W on some tasks, while high noise leaves a deficit in weak class discrimination. Across all results, most results at 30% noise are lower than at 0%, despite using clean labels for downstream learning. A smaller W means that the weakest class contrast is less distinguishable relative to within-class variability. Thus, fitting a downstream classifier does not eliminate the effect of upstream noise on weak discriminative structure. Intermediate responses and retinal-transfer trajectories vary with the target task, indicating that this geometry also depends on the distribution being evaluated. These observations motivate explicitly strengthening weak class-discriminative directions during adaptation, rather than relying on the classification objective alone.

3 Mitigating the Noise with Discriminative Geom- etry Tuning

The weak-discrimination deficit identified in Section 2.3 motivates an adaptation objective that improves the representation underlying the downstream decision boundary. We introduce *Discriminative Geometry Tuning* (DGTune; Figure 4), which strengthens weak class separation and encourages

the same discriminative relations to generalize across samples. Given a frozen medical backbone f_ϕ, we learn an MLP representation $z = h_\theta(f_\phi(x)) \in \mathbb{R}^d$ and a classification head $g_\omega(z)$. Two complementary regularizers act on this representation: one raises deficient Fisher singular values, while the other tests their associated directions and class contrasts on a separate sample group. Only θ and ω are updated, preserving the pre-trained medical model.

3.1 Strengthening Weak Discriminative Structure

Our analysis concerns class separation rather than feature variation alone. We therefore regularize the singular values of a *Fisher-whitened class-contrast matrix*, which measures between-class separation relative to within-class variation (Fisher, 1936; Dorfer et al., 2016). To estimate this geometry without allowing frequent classes to dominate, each update draws two balanced groups A and B, with m examples per class and no shared sampling units. Units follow available grouping identifiers, or individual images otherwise. Classification uses a separate shuffled minibatch $\mathcal{C}$ from the empirical training distribution; balanced sampling serves only the geometric objective. For $G \in \{A, B\}$, let $\mu_{G,c}$ denote its class means and $\bar{\mu}_G$ their average. We define

$$
\begin{aligned}
M_G &= \frac{1}{\sqrt{K}}[\mu_{G,1} - \bar{\mu}_G, \ldots, \mu_{G,K} - \bar{\mu}_G], \\
\Sigma_G &= \frac{1}{K(m-1)} \sum_{c=1}^{K} \sum_{i \in G_c} (z_i - \mu_{G,c})(z_i - \mu_{G,c})^\top, \\
S_G &= (1-\alpha)\Sigma_G + \left(\frac{\alpha \operatorname{tr}(\Sigma_G)}{d} + \epsilon \right) I,
\end{aligned}
\tag{4}
$$

where shrinkage, with $\alpha \in [0, 1]$ and $\epsilon > 0$, stabilizes covariance estimation from small groups. Using the orthonormal class-contrast basis H, the whitened matrix and its spectrum are

$$
C_G = S_G^{-1/2} M_G H, \qquad T_G = C_G^\top C_G, \qquad \lambda_{G,j} = \sigma_{G,j}^2, \quad j = 1, \ldots, K-1. \tag{5}
$$

The singular values are ordered increasingly, $\sigma_{G,1} \leq \cdots \leq \sigma_{G,K-1}$. Their squares recover the Fisher eigenvalues in Equation (3), with $W = \sigma_{G,1}^2$, directly connecting the regularizer to the weak structure examined in Section 2.3.

Cross-entropy can exploit well-separated directions without explicitly correcting weaker contrasts, while increasing total spectral energy can favor already dominant directions. Building on the emphasis on small discriminant eigenvalues in DeepLDA (Dorfer et al., 2016), we instead require deficient directions to attain a calibrated strength. For the q weakest directions and a squared-singular-value threshold $\tau > 0$, we use

$$
\mathcal{L}_{\text{spec}} = \frac{1}{2q} \sum_{G \in \{A,B\}} \sum_{j=1}^{q} \left[1 - \frac{\sigma_{G,j}^2}{\tau} \right]_+^2, \qquad [a]_+ = \max(a, 0). \tag{6}
$$

The normalized hinge raises weak separation without continuing to reward directions that already meet the threshold. Setting $q = 1$ directly targets W, while larger $q \leq K - 1$ applies the same principle to a broader lower spectrum. Even with all $K - 1$ directions included, only those below τ are penalized. This objective targets within-group strength; making that strength useful across samples requires a further constraint on the discriminative relations it represents.

3.2 Generalizing Fisher Discrimination across Samples

Singular values specify the strength of discrimination but not its orientation or class identity: two groups can have identical spectra and still separate classes along different directions. Consequently, strengthening each spectrum independently does not ensure that the same Fisher relation remains effective across groups. DGTune addresses this by holding a source group's feature direction and signed class contrast fixed while evaluating them on the other group. For a unit eigenvector $v_{G,j}$ associated with the jth smallest eigenvalue of T_G, construct

$$
\widetilde{u}_{G,j} = S_G^{-1} M_G H v_{G,j}, \qquad u_{G,j} = \operatorname{sg}\left(\frac{\widetilde{u}_{G,j}}{\|\widetilde{u}_{G,j}\|_2} \right), \qquad a_{G,j} = \operatorname{sg}\left(\frac{M_G^\top u_{G,j}}{\|M_G^\top u_{G,j}\|_2} \right), \tag{7}
$$

where sg denotes stop-gradient. The direction $u_{G,j}$ determines where separation is measured, and $a_{G,j}$ encodes the corresponding pattern of class means. Detaching both anchors prevents the source reference from moving to accommodate a weak response in the receiving group. For $G' \neq G$, their signed agreement and within-class scale are

$$n_{G\to G',j} = a_{G,j}^\top M_{G'}^\top u_{G,j}, \qquad s_{G\to G',j} = \sqrt{u_{G,j}^\top S_{G'} u_{G,j}}. \tag{8}$$

The sign of $n_{G\to G',j}$ distinguishes preservation of the source contrast from reversal, which a squared numerator would incorrectly reward. We retain this information through the transfer score

$$c_{G\to G',j} = \begin{cases} n_{G\to G',j}/\operatorname{sg}(s_{G\to G,j}), & n_{G\to G',j} \le 0, \\ n_{G\to G',j}/s_{G\to G',j}, & n_{G\to G',j} > 0. \end{cases} \tag{9}$$

Positive agreement is normalized by the receiving group's variation, measuring how reliably it expresses the source contrast. Nonpositive agreement instead uses the fixed source scale, preventing variance inflation in the receiving group from reducing the penalty without correcting the contrast. On its own group, a nondegenerate anchor satisfies $c_{G\to G,j} = \sqrt{\lambda_{G,j}} = \sigma_{G,j}$ (Appendix A); hence $\sqrt{\tau}$ provides a transfer threshold consistent with the spectral objective. The resulting bidirectional loss is

$$\mathcal{L}_{\text{cross}} = \frac{1}{2q} \sum_{(G,G')\in\{(A,B),(B,A)\}} \sum_{j=1}^{q} \left[1 - \frac{c_{G\to G',j}}{\sqrt{\tau}}\right]_+^2. \tag{10}$$

Each group supplies a detached reference and receives gradients when evaluated against the other. Repeatedly resampling these disjoint groups exposes the same type of discriminative relation to different examples, encouraging cross-sample Fisher generalization using only labeled training data. Multiple group pairs contribute averaged losses, with separate covariance estimates for each pair.

Together, the two regularizers complement the mean cross-entropy $\mathcal{L}_{\text{CE}}$ on $\mathcal{C}$: classification fits the downstream decision rule, spectral strengthening raises deficient separation, and cross-sample transfer requires that separation to remain effective beyond its reference group. We optimize

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + r(t)\big(\beta\mathcal{L}_{\text{spec}} + \gamma\mathcal{L}_{\text{cross}}\big). \tag{11}$$

After a cross-entropy warmup, we calibrate τ from training-group weak eigenvalues and keep it fixed, while a linear ramp $r(t)$ introduces regularization gradually. Geometry is measured on the MLP's hidden representation before dropout; efficient linear solves and a $(K-1)$-dimensional eigendecomposition avoid explicitly constructing $S_G^{-1/2}$. Derivations and optimization details appear in Appendix A. At inference, the group sampler and both regularizers are removed, leaving only the frozen medical backbone and the learned MLP.

3.3 Evaluation on Models with Synthetic Noise

We examine the downstream performance of DGTune under different pre-training noise ratios, comparing it with MLP tuning and NMTune on five in-distribution (ID) and two out-of-distribution (OOD) medical image classification tasks. Table 1 summarizes the reported accuracies at 0% and 30% noise, while Appendix B provides the complete results across all six noise levels, together with linear probing (LP) references.

For ID tasks, DGTune records the highest reported accuracy on all five datasets at both noise endpoints. At 30% noise, its average accuracy is 70.66%, compared with 57.83% for MLP tuning and 60.06% for NMTune, corresponding to differences of 12.83 and 10.61 percentage points, respectively. The differences are particularly pronounced on IDRiD, where DGTune reaches 57.28%, compared with 39.81% for MLP tuning and 40.78% for NMTune. Although NMTune yields higher accuracies than MLP tuning on several datasets, it remains below MLP tuning on APTOS and Messidor-2 across all six noise levels. For OOD tasks, DGTune records the highest reported accuracies on both PathMNIST and Camelyon17. At 30% noise, it achieves 93.16% and 91.62%, respectively, exceeding the corresponding NMTune results by 2.78 and 0.88 percentage points, and the MLP tuning results by 3.80 and 2.82 percentage points. Across the complete appendix table, DGTune has the highest reported accuracy in all 42 dataset–noise configurations.

Table 1: Comparison of downstream classification performance on five in-distribution (ID) and two out-of-distribution (OOD) medical tasks, measured by test accuracy (%).

		ID					OOD	
Noise	Method	Kvasir-v2	APTOS	DeepDR	IDRiD	Messidor-2	PathMNIST	Camelyon17
0%	MLP	78.75	72.89	58.33	45.63	62.36	90.97	91.55
	NMTune	86.25	70.95	58.59	48.54	60.06	91.95	92.59
	DGTune	**89.62**	**77.87**	**63.38**	**55.34**	**67.82**	**93.48**	**93.15**
30%	MLP	73.19	71.09	46.72	39.81	58.33	89.36	88.80
	NMTune	81.62	67.63	52.78	40.78	57.47	90.38	90.74
	DGTune	**88.12**	**78.01**	**62.37**	**57.28**	**67.53**	**93.16**	**91.62**

Table 2: Huatuo accuracy (%). $\Delta^\star$ subtracts the strongest baseline. Transfers abbreviate APTOS (A), DeepDR (D), IDRiD (I), and Messidor-2 (M).

Task / setting	MLP	NMTune	DGTune	$\Delta^\star$
Huatuo-7B:				
Kvasir-v2	79.08	**85.38**	84.79	-0.59
PathMNIST-224	90.25	91.62	**92.76**	+1.14
Camelyon17	86.47	89.38	**91.69**	+2.31
APTOS	74.00	74.04	**77.82**	+3.78
DeepDR	54.21	52.86	**60.35**	+6.14
IDRiD	45.63	47.25	**54.37**	+7.12
Messidor-2	57.66	58.05	**67.82**	+9.77
A $\rightarrow$ D	48.15	48.18	**52.68**	+4.50
A $\rightarrow$ I	48.26	49.68	**56.07**	+6.39
A $\rightarrow$ M	58.98	58.85	**61.09**	+2.11
D $\rightarrow$ A	45.03	47.43	**63.99**	+16.56
D $\rightarrow$ I	40.83	46.19	**50.13**	+3.94
D $\rightarrow$ M	48.76	49.22	**61.18**	+11.96
I $\rightarrow$ A	64.21	54.98	**68.72**	+4.51
I $\rightarrow$ D	42.61	37.17	**49.60**	+6.99
I $\rightarrow$ M	50.61	47.69	**58.96**	+8.35
M $\rightarrow$ A	52.59	52.17	**65.86**	+13.27
M $\rightarrow$ D	49.11	47.39	**52.61**	+3.50
M $\rightarrow$ I	49.42	43.80	**55.30**	+5.88

Table 3: General-purpose accuracy (%).

Backbone		MLP	NMTune	DGTune	$\Delta^\star$
Swin-L	ID	73.58	71.94	**74.53**	+0.95
	OOD	60.18	55.24	**63.04**	+2.86
ViT-L/14	ID	70.62	71.35	**72.81**	+1.46
	OOD	53.87	52.31	**61.25**	+7.38
ConvNeXt-L	ID	70.32	69.81	**74.31**	+3.99
	OOD	54.10	51.68	**61.12**	+7.02

Table 4: Biomedical NER (F1, %)

Dataset	MLP	NMTune	DGTune	$\Delta^\star$
PubMedBERT				
NCBI Disease	77.51	77.85	**79.39**	+1.54
BC5CDR	83.90	**85.03**	84.75	-0.28
BioBERT				
NCBI Disease	76.71	76.43	**79.19**	+2.48
BC5CDR	83.72	84.23	**84.35**	+0.12
ClinicalBERT				
NCBI Disease	70.53	73.29	**73.84**	+0.55
BC5CDR	81.41	81.84	**81.96**	+0.12

4 EXPERIMENTS

We extend the evaluation to public medical VLMs, general-purpose visual backbones, and biomedical language models. This tests the applicability of discriminative adaptation when the downstream user has access to the representation but not its upstream supervision.

4.1 MEDICAL VISION–LANGUAGE MODELS

Setup. We use the frozen HuatuoGPT-Vision-7B backbone (Chen et al., 2024b), abbreviated as Huatuo-7B. Huatuo-7B is a multimodal large language models trained using PubMedVision, which contains approximately 1.3 million medical visual question-answering samples derived from PubMed. The 7B variant adopts the LLaVA-v1.5 architecture with Qwen2-7B as its language backbone to integrate visual and textual information. We keep the backbone frozen and evaluate downstream adaptation on seven medical image classification tasks and twelve directed retinal transfers.

Results. Table 2 shows that DGTune outperforms the strongest baseline in eighteen of nineteen Huatuo-7B evaluation settings, including all twelve directed retinal transfers. On APTOS, DeepDR, and Messidor-2, it achieves accuracies of 77.82%, 60.35%, and 67.82%, respectively. The largest gain occurs on DeepDR-to-APTOS transfer, where DGTune reaches 63.99%, exceeding the

strongest baseline by 16.56 percentage points. On DeepDR-to-Messidor-2, it achieves 61.18%, a gain of 11.96 points. Improvements also extend to PathMNIST-224 and Camelyon17, with gains of 1.14 and 2.31 points over the strongest baseline, respectively. Kvasir-v2 remains the only exception, where DGTune trails NMTune by 0.59 points. Overall, these results support the effectiveness of DGTune in adapting frozen medical representations for ID and OOD tasks.

4.2 General-Purpose Visual Models

Setup. We next consider Swin-L (Liu et al., 2021) pre-trained on ImageNet-22K, and ViT-L/14 (Dosovitskiy et al., 2021) and ConvNeXt-L (Liu et al., 2022) trained with LAION-2B image–text supervision (Schuhmann et al., 2022). These frozen representations provide a complementary test beyond medically specialized pre-training. We averaged the five source-domain tests for ID and the two cross-center/hospital tests plus twelve retinal transfers for OOD.

Results. Table 3 shows that DGTune consistently outperforms the baselines on both ID and OOD tasks. For Swin-L, ViT-L/14, and ConvNeXt-L, the mean ID accuracies reach 74.53%, 72.81%, and 74.31%, respectively, while the corresponding OOD means are 63.04%, 61.25%, and 61.12%. DGTune improves OOD accuracy over the strongest baseline by 2.86, 7.38, and 7.02 percentage points, respectively, with larger gains than on ID tasks. These results support the effectiveness of DGTune across transformer and convolutional representations, particularly for cross-domain generalization.

4.3 Language Models and Datasets

Setup. We evaluate PubMedBERT (Gu et al., 2021), BioBERT (Lee et al., 2020), and Bio_ClinicalBERT (Alsentzer et al., 2019) on NCBI Disease (Doğan et al., 2014) and BC5CDR (Wei et al., 2016). Adaptation uses frozen contextual token features and the designated training, development, and test splits. We measure entity-level micro-F1 with exact boundaries and types.

Results. Table 4 extends the evaluation to biomedical named entity recognition, where DGTune improves over the strongest baseline in five of six settings. On NCBI Disease, it achieves F1 scores of 79.39% with PubMedBERT and 79.19% with BioBERT, yielding gains of 1.54 and 2.48 percentage points, respectively. On BC5CDR, however, improvements are limited to 0.12 and 0.12 points with BioBERT and ClinicalBERT, while PubMedBERT underperforms NMTune by 0.28 points. These results support the applicability of DGTune to biomedical language representations, although the gains are modest overall and not uniform across datasets and backbones.

4.4 Other experiments

Appendix B presents component ablations, parameter sensitivity analyses, and GPU update-time measurements. The ablation supports the complementary benefits of weak-spectrum strengthening and cross-sample Fisher transfer, while the sensitivity study examines the effects of β, γ, and q. DGTune averages 17.29 ms per GPU update, compared with 1.40 ms for MLP, but incurs little additional update cost relative to the single-component variant.

5 Conclusion

We studied how noisy medical pre-training affects downstream adaptation with clean labels. Controlled experiments with billion-parameter medical VLMs show that upstream noise can leave persistent effects on downstream performance. Motivated by the discriminative geometry of these representations, we introduced DGTune, which couples weak Fisher-spectrum strengthening with cross-sample transfer of discriminative directions and class contrasts. By optimizing only a lightweight MLP, DGTune improves adaptation without updating the pre-trained backbone. Experiments on controlled and publicly available models demonstrate its benefits across medical tasks, highlighting the importance of both upstream supervision quality and downstream discriminative geometry in building transferable medical models.

REFERENCES

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pp. 72–78. Association for Computational Linguistics, 2019. doi: 10.18653/v1/W19-1909. URL https://aclanthology.org/W19-1909/.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025. doi: 10.48550/arXiv.2502.13923. URL https://arxiv.org/abs/2502.13923.

Haoxuan Che, Yuhan Cheng, Haibo Jin, and Hao Chen. Towards generalizable diabetic retinopathy grading in unseen domains. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2023*, pp. 430–440. Springer, 2023. URL https://arxiv.org/abs/2307.04378.

Hao Chen, Jindong Wang, Ankit Shah, Ran Tao, Hongxin Wei, Xing Xie, Masashi Sugiyama, and Bhiksha Raj. Understanding and mitigating the label noise in pre-training on downstream tasks. In *International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=TjhUtloBZU.

Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Zhenyang Cai, Ke Ji, Xiang Wan, and Benyou Wang. Towards injecting medical visual knowledge into multimodal LLMs at scale. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 7346–7370. Association for Computational Linguistics, 2024b. doi: 10.18653/v1/2024.emnlp-main.418. URL https://aclanthology.org/2024.emnlp-main.418/.

Xinyang Chen, Sinan Wang, Mingsheng Long, and Jianmin Wang. Transferability vs. discriminability: Batch spectral penalization for adversarial domain adaptation. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1081–1090. PMLR, 2019. URL https://proceedings.mlr.press/v97/chen19i.html.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. NCBI disease corpus: A resource for disease name recognition and concept normalization. *Journal of Biomedical Informatics*, 47: 1–10, 2014. doi: 10.1016/j.jbi.2013.12.006.

Matthias Dorfer, Rainer Kelz, and Gerhard Widmer. Deep linear discriminant analysis. In *International Conference on Learning Representations*, 2016. URL https://arxiv.org/abs/1511.04707.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2010.11929.

R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2): 179–188, 1936. doi: 10.1111/j.1469-1809.1936.tb02137.x. URL https://doi.org/10.1111/j.1469-1809.1936.tb02137.x.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei Koh, Olga Saukh, Alexander Ratner, Shuran Song, Hannaneh Hajishirzi, Ali Farhadi, Romain Beaumont, Sewoong Oh, Alex Dimakis, Jenia Jitsev, Yair Carmon, Vaishaal Shankar, and Ludwig Schmidt. DataComp:

In search of the next generation of multimodal datasets. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/56332d41d55ad7ad8024aac625881be7-Abstract-Datasets_and_Benchmarks.html.

Aritra Ghosh, Himanshu Kumar, and P. S. Sastry. Robust loss functions under label noise for deep neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017. doi: 10.1609/aaai.v31i1.10894. URL https://ojs.aaai.org/index.php/AAAI/article/view/10894.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 2021. doi: 10.1145/3458754.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/a19744e268754fb0148b017647355b7b-Abstract.html.

Shih-Cheng Huang, Liyue Shen, Matthew P. Lungren, and Serena Yeung. GLoRIA: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3942–3951, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Huang_GLoRIA_A_Multimodal_Global-Local_Representation_Learning_Framework_for_Label-Efficient_Medical_ICCV_2021_paper.html.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4904–4916. PMLR, 2021. URL https://proceedings.mlr.press/v139/jia21b.html.

Jakob Nikolas Kather, Johannes Krisam, Pornpimol Charoentong, Tom Luedde, Esther Herpel, Cleo-Aron Weis, Timo Gaiser, Alexander Marx, Nektarios A. Valous, Dyke Ferber, Lina Jansen, Constantino Carlos Reyes-Aldasoro, Inka Zörnig, Dirk Jäger, Hermann Brenner, Jenny Chang-Claude, Michael Hoffmeister, and Niels Halama. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLOS Medicine*, 16(1): e1002730, 2019. doi: 10.1371/journal.pmed.1002730. URL https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1002730.

Simon Kornblith, Jonathon Shlens, and Quoc V. Le. Do better ImageNet models transfer better? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2661–2671, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Kornblith_Do_Better_ImageNet_Models_Transfer_Better_CVPR_2019_paper.html.

Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2202.10054.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020. doi: 10.1093/bioinformatics/btz682. URL https://academic.oup.com/bioinformatics/article/36/4/1234/5566506.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper

/2023/hash/5abcdf8ecdcacba028c6662789194572-Abstract-Datasets_and_Benchmarks.html.

Junnan Li, Richard Socher, and Steven C. H. Hoi. DivideMix: Learning with noisy labels as semi-supervised learning. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/2002.07394.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. URL https://arxiv.org/abs/2103.14030.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. URL https://arxiv.org/abs/2201.03545.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-Flamingo: A multimodal medical few-shot learner. In *Proceedings of the 3rd Machine Learning for Health Symposium*, volume 225 of *Proceedings of Machine Learning Research*, pp. 353–367. PMLR, 2023. URL https://proceedings.mlr.press/v225/moor23a.html.

Konstantin Pogorelov, Kristin Ranheim Randel, Carsten Griwodz, Sigrun Losada Eskeland, Thomas de Lange, Dag Johansen, Concetto Spampinato, Duc-Tien Dang-Nguyen, Mathias Lux, Peter Thelin Schmidt, Michael Riegler, and Pål Halvorsen. KVASIR: A multi-class image dataset for computer aided gastrointestinal disease detection. In *Proceedings of the 8th ACM on Multimedia Systems Conference*, pp. 164–169. ACM, 2017. doi: 10.1145/3083187.3083212. URL https://datasets.simula.no/kvasir/.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: An open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://papers.nips.cc/paper/2022/hash/a1859debfb3b59d094f3504d5ebb6c25-Abstract-Datasets_and_Benchmarks.html.

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. MedCLIP: Contrastive learning from unpaired medical images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 3876–3887. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.256. URL https://aclanthology.org/2022.emnlp-main.256/.

Chih-Hsuan Wei, Yifan Peng, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Jiao Li, Thomas C. Wiegers, and Zhiyong Lu. Assessing the state of the art in biomedical relation extraction: overview of the BioCreative V chemical–disease relation (CDR) task. *Database*, 2016: baw032, 2016. doi: 10.1093/database/baw032.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7959–7971, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Wortsman_Robust_Fine-Tuning_of_Zero-Shot_Models_CVPR_2022_paper.html.

Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. MedMNIST v2—a large-scale lightweight benchmark for 2D and 3D biomedical image classification. *Scientific Data*, 10(1):41, 2023. doi: 10.1038/s41597-022-01721-8. URL https://doi.org/10.1038/s41597-022-01721-8.

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. *arXiv preprint arXiv:2303.00915*, 2023. doi: 10.48550/arXiv.2303.00915. URL https://arxiv.org/abs/2303.00915.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D. Manning, and Curtis P. Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Proceedings of the 7th Machine Learning for Healthcare Conference*, volume 182 of *Proceedings of Machine Learning Research*, pp. 2–25. PMLR, 2022. URL https://proceedings.mlr.press/v182/zhang22a.html.

A DGTune: Derivation and Optimization Details

Connection between the spectrum and transfer scores. Write $X_G = M_G H$ and $Y_G = S_G^{-1} X_G$. The contrast basis satisfies $H^\top H = I$ and $HH^\top = I - \mathbf{1}\mathbf{1}^\top/K$. Because $M_G\mathbf{1} = 0$, we have $M_G = M_G HH^\top$. For a unit eigenvector v with $T_G v = \lambda v$ and $\lambda > 0$, let $\widetilde{u} = Y_G v$. Then

$$\widetilde{u}^\top S_G \widetilde{u} = \lambda, \qquad M_G^\top \widetilde{u} = HT_G v = \lambda H v, \qquad \|M_G^\top \widetilde{u}\|_2^2 = \lambda^2. \tag{12}$$

Normalizing $\widetilde{u}$ does not change the ratio of between-class to within-class strength. With $a = M_G^\top u / \|M_G^\top u\|_2$, the source-group score is therefore

$$\frac{a^\top M_G^\top u}{\sqrt{u^\top S_G u}} = \frac{\|M_G^\top u\|_2}{\sqrt{u^\top S_G u}} = \sqrt{\lambda} = \sigma. \tag{13}$$

This identity explains why the spectral loss uses threshold τ whereas the signed transfer loss uses $\sqrt{\tau}$. Flipping the sign of an eigenvector flips both u and a, leaving the transfer score unchanged. The score thus depends on the class-discriminative relation, not an arbitrary eigenvector sign.

Efficient Fisher computation. Stack the within-class residuals into $R_G \in \mathbb{R}^{Km \times d}$, scaled by $1/\sqrt{K(m-1)}$, so that $\Sigma_G = R_G^\top R_G$. Let $b_G = \alpha \|R_G\|_F^2/d + \epsilon$ and $V_G = \sqrt{1-\alpha}\, R_G$. We compute

$$Y_G = \frac{1}{b_G}\left[X_G - V_G^\top (V_G V_G^\top + b_G I)^{-1} V_G X_G\right], \qquad T_G = X_G^\top Y_G. \tag{14}$$

The inner system is solved by Cholesky factorization. This is exactly $S_G^{-1} X_G$ by the Woodbury identity and avoids forming the inverse or inverse square root of the $d \times d$ covariance. Scatter matrices, solves, and eigendecompositions use float64, while the MLP and classification objective use float32. We symmetrize T_G and clamp only negative eigenvalues within a numerical roundoff tolerance; materially negative or nonfinite values raise an error.

Degenerate reference directions. The spectral loss includes zero eigenvalues. If a source direction or its class-contrast vector has zero or nonfinite norm, its cross-group reference cannot be normalized. That reference contributes zero to the cross loss, with the denominator $2q$ retained, and its occurrence is logged. All valid references satisfy the source-score identity above, which is also checked in the implementation. Gradients of the spectral loss pass through the eigenvalues. In the cross loss, source eigenvectors and normalized anchors are detached, while gradients pass through the receiving group's means and, for positive agreement, covariance.

Training-only calibration and scheduling. Let T_{w} be the number of cross-entropy warmup updates and T_{r} the ramp length. From all selected weak eigenvalues collected over the latter half of warmup, compute the median $\widetilde{\lambda}$. For positive $\widetilde{\lambda}$, set $\tau = \max(10^{-6}, \kappa\widetilde{\lambda})$, where $\kappa > 0$ is a threshold multiplier. The implementation uses $\tau = 1$ if the median is zero, recording the median together with the resulting threshold. After calibration, τ remains fixed and $r(t) = \min\{1, \max\{0, (t - T_{\mathrm{w}})/T_{\mathrm{r}}\}\}$. Neither calibration nor gradient updates use validation or test labels.

Sampling and adaptation recipe. Each geometric panel contains its own balanced, unit-disjoint pair; estimates are never merged across panels. Units are partitioned before class-wise sampling, and $m \geq 2$ distinct units per class are sampled on each side. The classification batch is drawn by shuffling the original training examples, separately from these panels. This decoupled recipe keeps the MLP's supervised sampling distribution and adds geometric regularization without changing its architecture. The matched adaptation runs use a 512-dimensional hidden layer, dropout 0.1, and AdamW. Their geometric objective uses $q = K - 1$, with thresholding selecting deficient directions, and 20% warmup followed by a 20% ramp.

Table 5: **Complete six-level results.**

		ID					OOD	
Noise	Method	Kvasir-v2	APTOS	DeepDR	IDRiD	Messidor-2	PathMNIST	Camelyon17
0%	LP	58.00	49.93	45.45	44.66	58.33	71.87	58.09
	MLP	78.75	72.89	58.33	45.63	62.36	90.97	91.55
	NMTune	86.25	70.95	58.59	48.54	60.06	91.95	92.59
	DGTune	**89.62**	**77.87**	**63.38**	**55.34**	**67.82**	**93.48**	**93.15**
2.5%	LP	58.06	49.93	45.45	48.54	58.33	73.98	82.45
	MLP	77.44	72.75	56.06	44.66	62.07	90.86	90.98
	NMTune	83.75	70.54	58.33	47.57	59.77	91.87	91.52
	DGTune	**89.19**	**78.70**	**63.64**	**59.22**	**69.54**	**93.79**	**92.21**
5%	LP	59.13	49.93	45.45	44.66	58.33	71.98	81.83
	MLP	76.69	72.61	55.81	42.72	61.78	90.82	90.81
	NMTune	83.62	70.40	56.82	46.60	59.48	91.28	91.43
	DGTune	**89.12**	**79.94**	**65.15**	**63.11**	**66.38**	**93.52**	**92.69**
10%	LP	52.94	49.93	45.45	47.57	58.33	71.11	81.52
	MLP	76.06	72.48	52.78	41.75	60.34	90.71	89.65
	NMTune	82.12	69.71	54.55	45.63	59.20	90.92	91.28
	DGTune	**89.50**	**78.42**	**63.89**	**57.28**	**68.97**	**93.27**	**92.20**
20%	LP	51.31	49.93	45.45	46.60	58.33	69.64	81.56
	MLP	74.88	71.23	51.52	40.78	59.77	89.60	89.05
	NMTune	81.94	68.46	54.29	43.69	58.33	90.68	91.04
	DGTune	**87.44**	**78.56**	**62.88**	**62.14**	**66.67**	**93.20**	**92.40**
30%	LP	54.06	49.93	45.45	42.72	58.33	69.22	63.42
	MLP	73.19	71.09	46.72	39.81	58.33	89.36	88.80
	NMTune	81.62	67.63	52.78	40.78	57.47	90.38	90.74
	DGTune	**88.12**	**78.01**	**62.37**	**57.28**	**67.53**	**93.16**	**91.62**

B Evaluation Protocols and Results

B.1 Full results

Table 1 shows the 0% and 30% noise endpoints; Table 5 retains the complete six-level results.

MLP and DGTune. Both methods use the same 512-dimensional head, supervised sampling, and optimizer recipe. DGTune adds the spectral and cross-sample losses with $q = K - 1$. Retinal tasks use 100 epochs, learning rate 10^{-3}, CE batch size 256, and source-validation macro-F1 selection. The three non-retinal tasks use 50 epochs, learning rate 3×10^{-4}, CE batch size 1,024, and validation-accuracy selection.

NMTune reference. We adopt the publicly reported hyperparameter configuration from the original NMTune paper (Chen et al., 2024a). Using the official implementation, we train for 30 epochs with a batch size of 64, AdamW, a learning rate of 10^{-3}, weight decay of 10^{-4}, and a cosine learning-rate schedule. Its residual $D \to 4D \to D$ head uses batch normalization and ReLU. The loss weights are 1, 0.04, and 0.001, and the spectral objective uses the top-five singular-value fraction from the pinned official code.

B.2 Ablation study

Component ablation. Table 6 examines the contributions of the two regularizers on Kvasir-v2 across six upstream noise levels. Both single-component variants consistently improve over MLP, while combining them yields the highest accuracy at every noise level. At 30% noise, DGTune achieves 88.12% accuracy, exceeding the spectrum-only and cross-sample-only variants by 10.67 and 11.50 percentage points, respectively. Moreover, the accuracy drop from 0% to 30% noise is reduced from 5.56 points for MLP to 1.50 points for DGTune. These results support the com-

plementary benefits of weak-spectrum strengthening and cross-sample Fisher transfer, with their combination improving downstream accuracy and reducing sensitivity to upstream noise.

Table 6: Component ablation on Kvasir-v2 across six upstream noise levels.

Noise (%)	MLP	Spectrum only	Cross-sample only	DGTune
0	78.75	81.22	81.56	**89.62**
2.5	77.44	79.60	79.66	**89.19**
5	76.69	79.58	79.42	**89.12**
10	76.06	78.92	79.09	**89.50**
20	74.88	77.55	77.09	**87.44**
30	73.19	77.45	76.62	**88.12**

B.3 SENSITIVITY STUDY

Parameter sensitivity. Table 7 examines the sensitivity of DGTune to β, γ, and q on Kvasir-v2 by varying one parameter at a time while fixing the others. Across all six noise levels, the highest accuracy in each sweep is obtained at $\beta = 10$, $\gamma = 3$, and $q = 7$, respectively. Although $\beta = 10$ outperforms all smaller tested values, increasing γ beyond 3 reduces performance: at 30% noise, accuracy decreases from 88.12% to 82.38% when γ increases to 10. Increasing q from 1 to 7 improves accuracy at every noise level, including a gain of 7.12 percentage points at 30% noise. These results favor the same configuration across noise levels, while demonstrating that appropriate parameter settings remain important for achieving strong performance.

Table 7: One-factor parameter sensitivity on Kvasir-v2.

Parameter	Value	0%	2.5%	5%	10%	20%	30%
β	0.1	86.31	86.00	84.94	84.31	81.75	81.94
	0.3	86.31	86.19	83.81	84.38	82.12	82.44
	1	85.56	84.94	85.06	84.88	82.81	83.19
	3	87.50	86.12	85.50	85.81	83.62	84.75
	10	**89.62**	**89.19**	**89.12**	**89.50**	**87.44**	**88.12**
γ	0.1	88.62	87.81	86.19	85.50	84.50	84.00
	0.3	87.38	86.44	86.56	85.56	84.75	84.94
	1	88.12	85.94	84.62	85.56	83.69	85.50
	3	**89.62**	**89.19**	**89.12**	**89.50**	**87.44**	**88.12**
	10	85.69	85.00	84.69	85.31	82.88	82.38
q	1	85.62	84.19	83.44	83.06	79.75	81.00
	2	86.94	85.69	85.50	84.50	83.19	82.69
	7	**89.62**	**89.19**	**89.12**	**89.50**	**87.44**	**88.12**

B.4 GPU UPDATE TIME RESULTS

GPU update time. Table 8 reports GPU update times on Kvasir-v2 using an H800 with a maximum cross-entropy batch size of $1{,}024$. Across the six noise levels, DGTune requires an average of 17.29 ms per update, compared with 1.40 ms for MLP, corresponding to approximately $12.4\times$ the update time. However, its cost is comparable to the spectrum-only (16.92 ms) and cross-sample-only (17.54 ms) variants. Combining the two regularizers therefore incurs little additional update cost relative to using either component alone, while providing the stronger accuracy observed in the component ablation. These measurements characterize GPU update cost.

Table 8: GPU update time on Kvasir-v2 across six noise levels (ms). All variants use an H800 and a maximum cross-entropy batch size of 1,024.

Noise (%)	MLP	Spectrum only	Cross-sample only	DGTune
0	1.28	16.02	16.94	16.52
2.5	1.29	14.82	15.96	15.39
5	1.28	16.42	17.01	16.35
10	1.66	19.43	19.88	20.01
20	1.62	18.71	19.03	19.14
30	1.29	16.15	16.41	16.35
Across-noise mean	1.40	16.92	17.54	17.29