Novel Finetuning Strategies for Adapting Biomedical Vision Language Models to Organ-Centered Pathology Microscopy Tasks

Siddharth Venkatesh[1]* **Benjamin Liu**[1, 2] **Ayman Sheikh**[1] **Anne Essien**[1] **Pratibh**[1]
Rayhan Roswendi[1] **Jeremiah Zhang**[1] **Kevin Zhu**[1] **Sunishchal Dev**[1]
[1]Algoverse AI Research
[2]Stanford University
venkatesh.sidd@gmail.com, bencliu@stanford.edu

Abstract

Biomedical vision-language models (VLMs) struggle with performance deterioration on earlier domains after fine-tuning and limited generalization under domain diversity and dataset imbalance. We propose an adapter-level framework combining Low-Rank Adaptation (LoRA) for efficient domain-specific tuning with model souping for cross-domain adaptability in microscopy images. Using BioMedCLIP and organ-specific domains from μ-Bench, adapter soups mitigate low generalization and improve robustness, achieving gains of up to 15% on fine-grained and 38% on coarse-grained tasks over baseline BioMedCLIP. The process is data- and resource-efficient, and hyperparameter analysis reveals sensitivities to domain similarity and dataset imbalance. Adapter merging offers a lightweight scalable approach for organ-specific accuracy and cross-domain stability in biomedical VLMs.

1 Introduction

Medical imaging is central to clinical practice, but automated analysis faces challenges such as annotation cost, [1], organ variability, [2], and poor model generalization, [3]. Vision-language models (VLMs) show promise but degrade in biomedical microscopy due to domain shifts and specialized reasoning needs, [4]. The μ-Bench benchmark highlights these issues, showing even biomedical foundation models like BioMedCLIP struggle with organ-specific tasks and suffer catastrophic forgetting, [5]. To address this, parameter-efficient methods like LoRA, [6] and model souping have emerged, [7; 8], enabling specialization with minimal parameters and improved generalization via weight merging. This work systematically evaluates LoRA-based model soups in biomedical microscopy across four organ datasets, testing five merging strategies. Results show hybrid generalist–specialist approaches, particularly SLERP, effectively balance specialization and generalization, improving μ-Bench performance by up to 38%, while divergence-driven methods often fail. Overall, we provide the first comprehensive evidence of how souping and LoRA interact in biomedical VLMs, offering scalable, lightweight, and robust solutions across organ systems.

*Correspondence to: venkatesh.sidd@gmail.com

39th Conference on Neural Information Processing Systems (NeurIPS 2025): The 3rd Workshop on Imageomics: Discovering Biological Knowledge from Images Using AI.

2 Related Work

2.1 Vision–Language Models in Biomedical Imaging

General-purpose VLMs like CLIP [9], BLIP [10], and Flamingo [11] inspire biomedical variants such as BioMedCLIP [12]. Yet, evaluation on μ-Bench [5] shows organ-specific degradation and reduction in a model's generalization ability under fine-tuning [4]. We address the specialization and generalization tradeoff through adapter-level merging.

2.2 Parameter-Efficient Fine-Tuning with LoRA

LoRA was introduced as a PEFT enabling specialization with 1% parameters in general NLP tasks [6]. Recent studies have extended to multi-task merging in NLP and vision [7; 8]. LoRA Soups [7] average adapters to combine skills without retraining, while Multi-LoRA Meets Vision [8] shows adapter merging for multi-task vision backbones. Unlike previous studies, our study extends LoRA's application to biomedical microscopy by systematically evaluating adapter-level merging strategies to address organ-specific variability, data imbalance, and cross-domain generalization challenges.

2.3 Model Merging and Weight Interpolation

Model soups improve generalization by merging fine-tuned models [13], with extensions like SLERP [14] and TIES merging [15]. Unlike prior work, we merge LoRA adapters instead of full weights. This provides the first systematic evidence of how merging strategies behave in imbalanced, structurally similar biomedical domains.

3 Methodology

3.1 Dataset

μ-Bench [5] collected licensed biomedical microscopy data, reviewed by pathologists for quality. It evaluates VLMs via VQA across cardiovascular, gastrointestinal, hematopathology, and neuropathology. Each image yields 5 coarse-grained (domain, modality, stain, subdomain, submodality) and 1 fine-grained (pathology classification) questions. Each image–question pair was treated as a training instance. Data was split 70/15/15 using stratified sampling. Dataset sizes: cardiovascular 2,400; gastrointestinal 24,844; hematopathology 15,600; neuropathology 7,746. For a detailed breakdown of data refer to Appendix A.

3.2 Fine-Tuning with Low-Rank Adaptation

We fine-tuned BioMedCLIP [12] with LoRA [6], a parameter-efficient fine-tuning method. In LoRA, a pre-trained weight matrix $W \in \mathbb{R}^{d\times k}$ is updated by adding a low-rank perturbation $\Delta W = BA$, where $A \in \mathbb{R}^{r\times k}$ and $B \in \mathbb{R}^{d\times r}$ are trainable matrices, and the rank r is chosen to be much smaller than $\min(d, k)$. LoRA adapters were placed on all layers, reducing the parameters used in training by 99.47%, as the base weights remain frozen, and only small adapter matrices are updated, enabling single-GPU training. For a detailed description of LoRA refer to Appendix D.

3.3 Model Souping

We merged LoRA adapters instead of full weights. Best organ checkpoints were selected by validation accuracy. Two families were explored:

Organ–Base Hybrids (WiSE-FT) [13]: Interpolating BioMedCLIP with one organ adapter using scale τ.

Multi-Organ Soups: We merged four organ adapters–Cardiovascular, Gastrointestinal, Hematopathology, and Neuropathology, using five strategies - (1) Linear Average [13], (2) SLERP [14], (3) Task Arithmetic [15], (4) TIES [15], (5) WiSE-FT after averaging

These methods span uniform averaging, geometry-aware interpolation, sparsity-based merging, and generalist–specialist balancing. Full derivations appear in Appendix C.

4 Results

4.1 Cross Domain Evaluation Results

Models were fine-tuned on organs fine-, coarse- and combined- tasks and were evaluated against the complete dataset of each organ alongside the complete pathology to evaluate cross-domain generalization. Combined-task fine-tuning yields a more balanced trade-off, demonstrating substantially higher robustness across domains and mitigating over-specialization, offering a promising strategy for more transferable medical foundation models. Based on these results from Figure 5, the LoRA-tuned model shows clear gains across all pathology subdomains, achieving substantially higher fine-grained, coarse-grained, and combined accuracies compared to the base BioMedCLIP. Detailed results exploring cross-domain evaluation can be found in Tables 1, 2, 3 in Appendix F.

After evaluating the organ-specific models - models fine-tuned on an organs complete dataset - across other domains (Appendix F) we provide an objective metric to compare our models. Using the evaluation metrics of arithmetic average and harmonic mean (as discussed in Appendix B), these were the results of each fine tuned model across each task type. Fine-tuning BioMedCLIP with LoRA yields clear improvements within target pathology domains while maintaining robustness across other organ-specific tasks, by restricting updates to only the low-rank adapters.

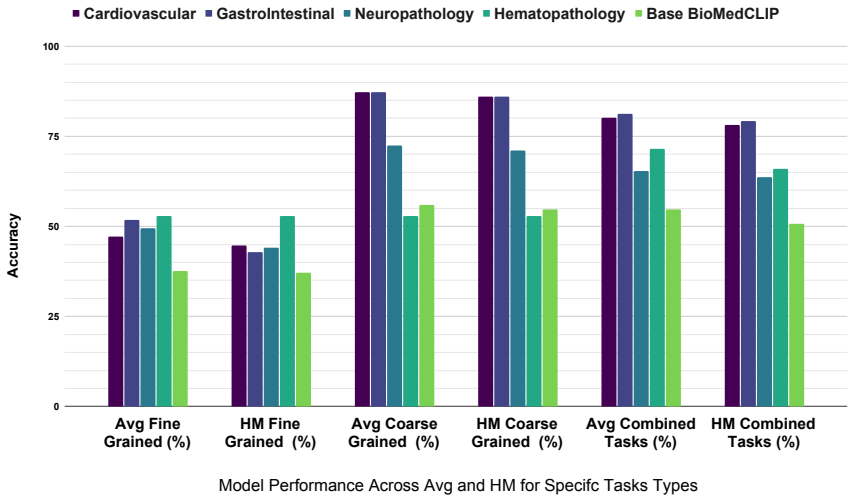

Figure 1: Cross-domain evaluation of organ-specific BioMedCLIP fine-tuned models compared to the baseline model using average (Avg) and harmonic mean (HM) accuracies. A detailed overview can be found in Table 4 in Appendix F. Accuracy is reported across fine-grained tasks, coarse-grained tasks, and combined tasks (fine- and coarse- grained tasks). Results are shown for cardiovascular, gastrointestinal, neuropathology, and hematopathology fine-tuned models, alongside the untuned base BioMedCLIP baseline.

As shown in Figure 1, almost every organ-specific models outperform the untuned BioMedCLIP baseline on fine- and coarse-grained tasks across all domains. The improvements are the most pronounced in coarse-grained accuracies, while the fine-grained accuracies of our fine tuned models outperform the base BioMedClip. This demonstrates not only improved specialization within pathology but also consistent robustness across diverse organ systems, indicating that LoRA provides a lightweight mechanism enhancing generalization. Overall, the table demonstrates a consistent trend: LoRA systematically boosts performance across all evaluation regimes, underscoring its value for domain adaptation in pathology-focused vision-language modeling.

4.2 Model Souping Results

Before evaluating soups, we fine-tuned BioMedCLIP with LoRA adapters on each organ dataset (Cardiovascular, Gastrointestinal, Hematopathology, Neuropathology). This produced strong organ-

specialized models, which we then merged into soups for comparison. As shown in Table 5 and Figure 2, almost every organ-specific BioMedCLIP hybrid substantially outperforms the untuned BioMedCLIP baseline across fine-, coarse-, and combined-pathology tasks. For example, the Gastrointestinal + BioMedCLIP hybrid achieves an average coarse accuracy of 87.9% and combined accuracy of 82.1%, far exceeding the baseline BioMedCLIP's 55.9% coarse and 51.8% combined. Hematopathology + BioMedCLIP likewise improves combined accuracy to 69.2%, while Cardiovascular + BioMedCLIP delivers strong coarse accuracy (85.5%) despite limited training data. These results demonstrate that two-way hybridization between the generalist BioMedCLIP model and a single specialized adapter effectively balances broad coverage with targeted discriminative power. By contrast, the All-Organ Linear Average soup achieves only 67.3% average combined accuracy, trailing the best organ-specific hybrids, reflecting the negative effect of dataset imbalance. This degradation highlights the effect of dataset imbalance: high-signal domains such as Gastrointestinal (≈4k samples) are diluted when merged with weaker ones such as Cardiovascular (≈400 samples). In other words, N-way averaging across structurally similar but uneven domains introduces interference without yielding the diversity-driven benefits reported in more heterogeneous benchmarks.

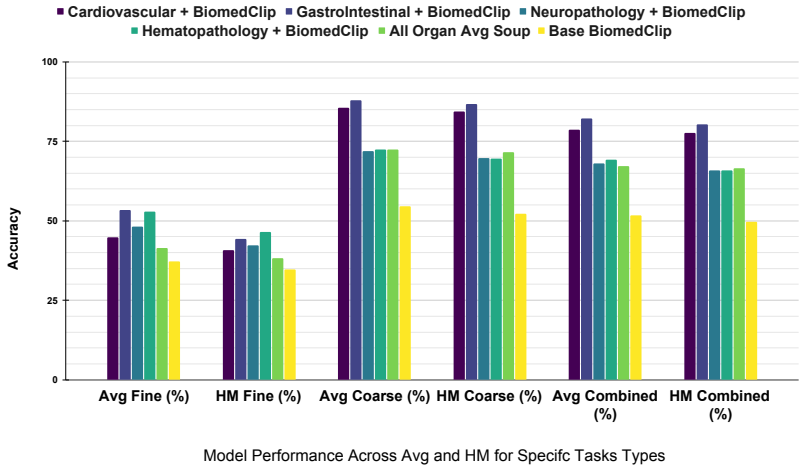

Figure 2: Performance of organ-specific BioMedCLIP hybrids (Base + Organ) compared to the base BioMedCLIP baseline and an all-organ average soup. Accuracy is reported across fine-grained tasks, coarse-grained tasks, and combined tasks (coarse- and fine-grained tasks) for all organ datasets using average (Avg) and harmonic mean (HM). Detailed results can be found in Table 5 in Appendix G.

4.3 Multi-Organ Souping Results

Turning to multi-organ soups in Figure 3, we observe that more sophisticated interpolation methods yield modest but consistent gains over uniform averaging. SLERP achieves the strongest results, with 91.2% coarse accuracy and 84.8% combined accuracy, outperforming Linear Average by nearly 17 percentage points on coarse and 18 points on combined tasks. WiSE-FT, when applied post-SLERP, performs comparably, confirming that angular-aware interpolation avoids some of the dilution effects of uniform averaging. By contrast, Task Arithmetic remains weak even at high β, achieving only 55.9% combined accuracy, while TIES collapses almost entirely, falling to 32.6% combined accuracy due to excessive trimming of shared biomedical features. These outcomes reinforce that organ-specific adapters are not highly divergent: their embeddings cluster in overlapping representational subspaces, limiting the utility of divergence-driven strategies. Methods such as Task Arithmetic and TIES fail in this setting - the high structural similarity across organ pathology images leaves little complementary signal to exploit, undermining approaches designed to leverage divergence. In aggregate, results establish two families of behavior: (1) hybrid soups (Base + Organ) consistently outperform both baselines and multi-way merges, showing that combining a generalist with a single specialist yields the best tradeoff between generalization and fine-grained sensitivity; and (2) multi-organ soups provide limited improvements beyond Linear Average, with SLERP and WiSE-FT Slerp offering the only robust gains.

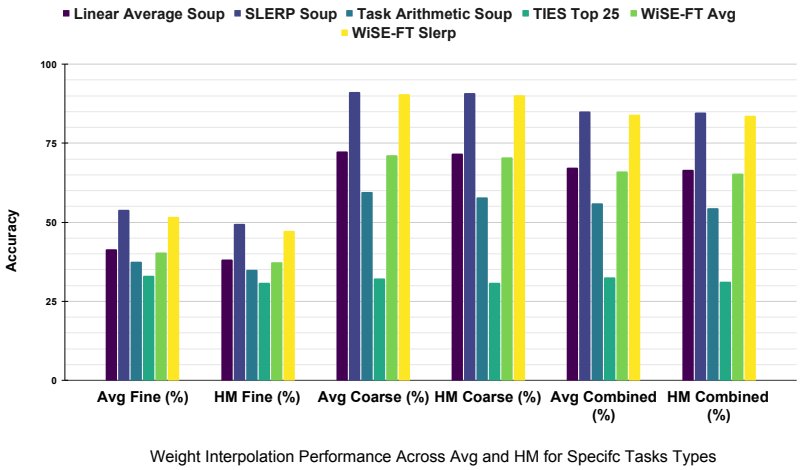

Figure 3: Performance of multi-organ models weight interpolated through methods such as the Linear Average Soup, SLERP Soup, Task Arithmetic Soup, TIES Top 25, WiSE-FT Avg, and WiSE-FT slerp. Accuracy is reported using average and harmonic mean across fine-grained tasks, coarse-grained tasks, and combined tasks (coarse- and fine-grained tasks).

4.3.1 Comparative Analysis of Various Hyperparameters and Souping Strategies

Task Arithmetic degraded accuracy at low β (0.1–0.3) and only partially recovered at higher values (0.7–0.9), never matching SLERP or Linear Average. Lower thresholds favored TIES (`keep_top_p=0.25`) for stability; high thresholds (`keep_top_p=0.9`) briefly improved fine-grained accuracy but collapsed coarse tasks. WiSE-FT was stable across $\tau \in \{0.3, 0.5, 0.7\}$, with stronger effects in two-way merges than multi-way soups. These sensitivities underscore that biomedical souping is less about extracting complementary expertise and more about carefully preserving shared structure. Small hyperparameter changes can suppress critical biological signals or exaggerate dataset imbalance, particularly when one domain dominates (e.g., Gastrointestinal). SLERP consistently outperformed Linear Average by preserving angular geometry, while WiSE-FT contributed little beyond averaging in multi-way settings. Collectively, these results confirm that hybridization (generalist + specialist) offers the most stable tradeoff between fine-grained accuracy and broad generalization highlighting its efficiency. Multi-organ soups remain fragile under biomedical constraints, and future work should probe adapter geometry and gradient conflicts directly to better explain the collapse of divergence-driven strategies. For a further detailed analysis of the representational geometry of the different adapter weights, please refer to Appendix J.

5 Limitations and Future Work

Our results reveal broader limitations and opportunities for extension. Potential negative impacts include biased performance across different demographic groups. While organ-specific hybrids consistently outperformed the baseline, N-way soups struggled to preserve these gains. In particular, dataset imbalance allowed large domains to dominate smaller ones, leading hybrids like Gastrointestinal + BioMedCLIP to outperform multi-organ merges. Future work should address these constraints through adaptive weighting (e.g., Fisher-weighted or dynamically scaled soups) and progressive merging pipelines that preserve both specialist and generalist strengths. Moreover, real-world pathology datasets such as TCGA, CAMELYON, and PANDA are noisier than μ-Bench, making them critical testbeds to evaluate robustness under realistic conditions. Although divergence-driven strategies such as Task Arithmetic and TIES struggled in our setting due to the high structural similarity across organ pathology images, they may prove more successful in these noisier, less structurally aligned datasets. Our current experiments focus on a one foundation model (BioMedCLIP), but extending evaluation to other biomedical VLMs like PLIP or QuiltCLIP [5], and even general-purpose VLMs, would better establish the framework's generalizability. Prior work has shown that biomedical VLMs sometimes

underperform compared to generalist models [5]; testing whether our strategies narrow or reverse this gap would provide stronger evidence for scalability and broader applicability.

6 Conclusion

We introduced a two-stage framework for biomedical VLMs using LoRA adapters and adapter-level soups, achieving consistent gains over BioMedCLIP on μ-Bench, with SLERP soups showing the strongest generalization. This is the first systematic study of LoRA-based merging in biomedical microscopy, demonstrating adapter soups as a lightweight alternative to full retraining and a general recipe for balancing specialization and transferability in domain-adapted foundation models.

7 Appendix

A Data Processing

To facilitate supervised VQA training, each question–image pair was treated as an independent training instance. Each instance was annotated with its corresponding task type (coarse or fine) and subsequently partitioned to support evaluation of both task-specific and overall performance. The datasets were divided into training, validation, and test splits in a 70/15/15 ratio using stratified sampling by question type to maintain representation across all categories.

All data used in this project was taken from the benchmark created in the μ-Bench dataset [5].The study used only pre-existing, publicly available biomedical microscopy datasets from repositories such as Zenodo, Dataverse, Dryad, and BBBC. All data were shared under permissive licenses (CC-BY-SA-4.0) that allow derivatives and redistribution. The μ-Bench dataset comprises 2,400 cardiovascular samples (2,000 coarse-grained, 400 fine-grained), 24,844 gastrointestinal samples (20,730 coarse-grained, 4,114 fine-grained), 15,600 hematopathology samples (13,000 coarse-grained, 2,600 fine-grained), and 7,746 neuropathology samples (6,455 coarse-grained, 1,291 fine-grained). The coarse-grained perception tasks tests basic image properties which are visually distinct and relatively straightforward even for non-biologists, but are important to assess whether VLMs have intuitive knowledge of biology and microscopy. Fine-grained perception include the identification of cell type, and disease classifications that are visually distinct and important for reasoning about biological images. Solving fine-grained perception relies on finer-grained visual features and is more challenging for humans [5].

B Evaluation Metrics

To summarize the performance of the model in multiple organ-specific tasks, we report both the **arithmetic average** and the **harmonic mean (HM)** of precision. The average provides a straightforward measure of overall performance, while the harmonic mean emphasizes consistent performance across all tasks, penalizing models that perform poorly in any single domain. Together, these metrics give a balanced view of both overall accuracy and robustness across domains.

1. Arithmetic Average:

$$\text{Avg Accuracy} = \frac{1}{n}\sum_{i=1}^{n} A_i$$

2. Harmonic Mean (HM):

$$\text{HM Accuracy} = \frac{n}{\sum_{i=1}^{n} \frac{1}{A_i}}$$

C Additional Details on Model Souping Strategies

For completeness, we provide the full mathematical formulations of the five model souping strategies described in Section 3.3. These details are omitted from the main text for brevity.

C.1 Linear Average Soup

The simplest approach averages parameters across organ adapters:

$$\theta_{\text{soup}} = \sum_{i=1}^{n} w_i \theta_i, \quad \sum_i w_i = 1,$$

where θ_i are the LoRA parameters for each organ. In our experiments, equal weights ($w_i = 0.25$) were used for the four organs, [13].

C.2 SLERP Soup

Spherical linear interpolation (SLERP) interpolates in angular space rather than Euclidean space, [14]. For two adapters θ_a, θ_b:

$$\text{SLERP}(\theta_a, \theta_b; \alpha) = \frac{\sin((1-\alpha)\theta)}{\sin\theta}\,\theta_a + \frac{\sin(\alpha\theta)}{\sin\theta}\,\theta_b.$$

Pairwise reduction was applied to extend SLERP across all four organ adapters.

C.3 Task Arithmetic Soup

Following vector arithmetic in representation space, [15], we first center each adapter's deltas:

$$\theta_i' = \theta_i - \overline{\theta}, \quad \overline{\theta} = \tfrac{1}{n}\sum_{i=1}^{n}\theta_i,$$

then recombine:

$$\theta_{\text{soup}} = \sum_i w_i\theta_i' + \beta\overline{\theta},$$

where β is a shared-knowledge coefficient (default $\beta = 0.5$).

C.4 TIES Merging

TIES (Trim–Intersect–Expand–Sign), [15] resolves conflicts between adapters by enforcing sparsity and sign consistency:

- **Trim:** keep only the top-p fraction of parameters by magnitude per adapter.
- **Intersect:** retain entries where all contributing adapters agree in sign.
- **Resolve:** in cases of disagreement, select the dominant contributor.

C.5 WiSE-FT Soup

WiSE-FT, [13] merges a general base model with a fine-tuned model by interpolation:

$$\theta_{\text{WiSE}} = (1-\tau)\theta_{\text{base}} + \tau\theta_{\text{fine}},$$

where τ balances generalization (θ_{base}) and specialization (θ_{fine}). For LoRA adapters, we scaled the low-rank matrices A, B by $\sqrt{\tau}$ to ensure that the effective update

$$\Delta W = \tfrac{\alpha}{r}(\sqrt{\tau}B)(\sqrt{\tau}A)$$

is correctly scaled.

D LoRA Architecture

In LoRA, a pre-trained weight matrix $W \in \mathbb{R}^{d\times k}$ is updated by adding a low-rank perturbation $\Delta W = BA$, where $A \in \mathbb{R}^{r\times k}$ and $B \in \mathbb{R}^{d\times r}$ are trainable matrices, and the rank r is chosen to be much smaller than $\min(d, k)$. Here, d and k denote the output and input dimensions of the layer, respectively. Notably, in the original method, matrix A is initialized using random values drawn from a normal distribution $A \sim \mathcal{N}(0, \sigma^2)$, as indicated in Figure 4. During adaptation, only A and B are learned, while the original weight matrix W remains frozen. This drastically reduces the number of trainable parameters from $d \times k$ (full matrix) to $r \times (d + k)$, and all symbols and initialization choices are captured in the schematic illustration. LoRA thus enables efficient adaptation to new tasks while retaining the pre-trained model's representations and knowledge.

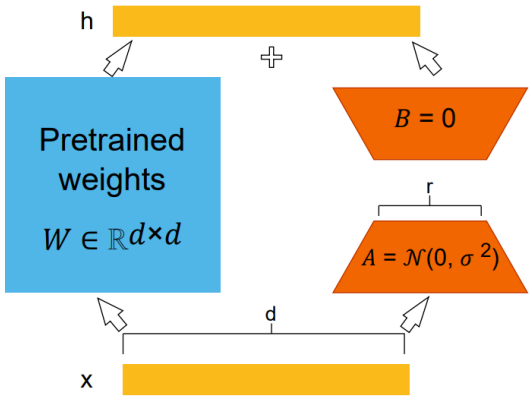

Figure 4: Redrawn from **(author?)** [6]. Illustration of the LoRA architecture.

E LoRA Fine-tuning Configuration

We fine-tuned BioMedCLIP (`microsoft/BiomedCLIP-PubMedBERT_256-vit_base_patch16_224`) using Low-Rank Adaptation (LoRA). The base model parameters were frozen, and LoRA adapters were applied to the text encoder's attention and feed-forward layers, as well as embedding projections. Below we summarize the key hyperparameters:

LoRA configuration: LoRA ranks were set to $r = 16$ for the text encoder and alignment modules, and $r = 32$ for the vision encoder. Corresponding scaling factors were $\alpha = 32$ (text/alignment) and $\alpha = 64$ (vision). A dropout rate of 0.1 was applied, with no bias terms trained.

Training setup: Models were trained with AdamW (learning rate 1×10^{-4}, weight decay 0.01) for 10 epochs, using cosine learning rate scheduling with 100 warmup steps. Batch size was 32 with gradient accumulation of 1. Mixed-precision training (AMP) was enabled. Early stopping was applied with patience of 3 epochs and $\Delta = 0.001$.

Data: Training used the μ-Bench datasets across cardiovascular, neuropathology, hematopathology, and gastrointestinal domains, with splits of 0.1, 0.25, 0.5, and 0.75 for ablation. Each experiment tracked coarse-, fine-, and combined-question VQA tasks. A random seed of 42 was used

Evaluation: Performance was assessed with accuracy and confidence metrics, both overall and per question type. Predictions were saved for downstream analysis, using a confidence threshold of 0.5.

Hardware and Compute: All experiments were run on NVIDIA GPU RTX 6000 Ada with CUDA support on Runpod as the cloud provider. A network volume of 125 GB was used for memory.Fine-tuning a single organ-specific LoRA adapter took approximately 5-20 min depending on selected datasets size. Adapter soup merging required roughly 1-5 minutes per soup. Across all experiments, the total compute is estimated at 20 GPU-hours.

F Detailed Results from Cross Domain Analysis of LoRA Fine-Tuning

Cross-domain evaluation highlights the trade-offs between specialization and generalization in organ-specific fine-tuning. Each model fine-tuned on fine-grained tasks achieves its highest accuracy within its own domain, but performance degrades when transferred to other domains, sometimes matching or underperforming the Base BioMedCLIP model, which maintains more consistent cross-domain accuracy (Table 1 in Appendix F). This drop illustrates the limited robustness of organ-specialized models. At the coarse- and combined-task levels (Tables 2 and 3 in Appendix F), accuracies are higher overall, particularly within-domain. Combined-task fine-tuning yields a more balanced trade-off, demonstrating substantially higher robustness across domains and mitigating over-specialization,

offering a promising strategy for more transferable medical foundation models. Based on these results from Figure 5, the LoRA-tuned model shows clear gains across all pathology subdomains, achieving substantially higher fine-grained, coarse-grained, and combined accuracies compared to the base BioMedCLIP.

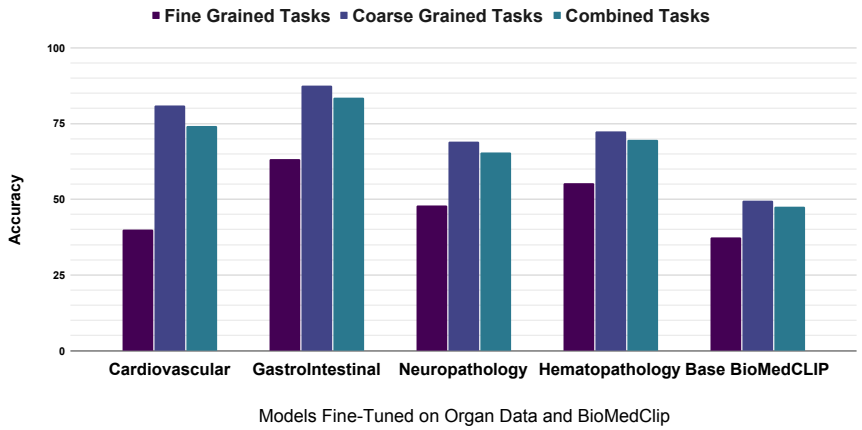

Figure 5: Cross-domain evaluation of organ-specific fine-tuned models compared to the BioMedCLIP baseline. Models were fine-tuned across an organs dataset and evaluated across the complete pathology dataset and the results are shown above. Accuracy is reported across fine-grained tasks, coarse-grained tasks, and combined tasks (fine- and coarse- grained tasks).

Tables 4–6 report fine-grained, coarse-grained, and combined-task accuracies (±95% CI) for organ-specific models and the baseline BioMedCLIP. Each table corresponds to the task granularity used during fine-tuning (fine, coarse, or combined). Rows indicate the organ on which a model was fine-tuned, while columns show evaluation performance across all organ-specific datasets. The final row in each table reports the performance of the unmodified BioMedCLIP model, serving as a baseline for comparison.

As expected, models achieve their highest accuracy within their training domain, with substantial improvements over BioMedCLIP. However, performance drops markedly when transferred across domains, in some cases approaching or falling below baseline. This highlights the trade-off between specialization and generalization: while organ-specific fine-tuning enhances in-domain sensitivity, it limits robustness across tasks.

Table 1: Fine-grained accuracy ± CI for each organ-specific task and complete pathology across models fine tuned on combined tasks.

Model	Cardiovascular	GastroIntestinal	Neuropathology	Hematopathology	Complete Pathology
Cardiovascular	76.67% (60.05–81.04)	44.17% (40.03–47.84)	40.72% (33.26–46.96)	29.23% (24.65–33.64)	39.97% (37.18–42.58)
GastroIntestinal	46.67% (31.81–56.29)	95.31% (92.75–96.13)	39.69% (32.30–45.94)	25.38% (21.07–29.68)	63.20% (60.31–65.63)
Neuropathology	50.00% (34.73–59.26)	51.13% (46.88–54.74)	74.74% (66.75–78.89)	22.05% (18.00–26.21)	47.98% (45.08–50.59)
Hematopathology	43.33% (28.94–53.26)	35.76% (31.86–39.40)	34.02% (27.06–40.27)	96.15% (92.82–96.72)	55.35% (52.43–57.91)
Base BioMedCLIP	38.33% (24.74–48.63)	52.27% (48.00–55.85)	35.05% (28.00–41.31)	23.08% (18.94–27.28)	37.51% (34.84–40.18)

Table 2: Coarse-grained accuracy ± CI for each organ-specific task and complete pathology across models fine tuned on combined tasks.

Model	Cardiovascular	GastroIntestinal	Neuropathology	Hematopathology	Complete Pathology
Cardiovascular	100.00% (98.74–100.00)	74.73% (73.08–76.13)	93.29% (91.17–94.33)	81.23% (79.28–82.74)	80.91% (79.87–81.81)
GastroIntestinal	96.00% (91.93–96.49)	100.00% (99.88–100.00)	82.25% (79.39–84.20)	70.56% (68.36–72.41)	87.55% (86.66–88.29)
Neuropathology	63.33% (56.95–67.79)	61.41% (59.62–63.03)	100.00% (99.61–100.00)	64.97% (62.70–66.93)	68.96% (67.77–70.05)
Hematopathology	74.00% (67.82–77.71)	60.64% (58.84–62.27)	53.46% (50.10–56.37)	100.00% (99.80–100.00)	72.42% (71.27–73.47)
Base BioMedCLIP	66.33% (59.97–70.61)	47.01% (45.20–48.71)	70.38% (67.15–72.89)	40.05% (37.82–42.16)	49.48% (48.25–50.71)

Table 3: Combined accuracy ± CI for each organ-specific task and complete pathology across models fine tuned on combined tasks.

Model	Cardiovascular	GastroIntestinal	Neuropathology	Hematopathology	Complete Pathology
Cardiovascular	96.11% (92.57–96.66)	69.66% (68.09–71.05)	84.52% (82.05–86.21)	72.56% (70.60–74.21)	74.11% (73.07–75.04)
GastroIntestinal	87.78% (83.07–89.85)	99.22% (98.78–99.36)	75.15% (72.34–77.30)	63.03% (60.96–64.86)	83.50% (82.61–84.28)
Neuropathology	50.00% (34.73–59.26)	59.71% (58.07–61.21)	95.79% (94.16–96.48)	57.82% (55.71–59.71)	65.48% (64.37–66.50)
Hematopathology	68.89% (63.20–72.73)	56.52% (54.86–58.04)	50.21% (47.18–52.92)	99.36% (98.78–99.45)	69.59% (68.51–70.58)
Base BioMedCLIP	61.67% (55.90–65.89)	47.88% (46.23–49.44)	64.49% (61.48–66.97)	37.22% (35.23–39.14)	47.49% (46.34–48.59)

Table 4: Average (Avg) and harmonic mean (HM) accuracies for organ-specific models - models fine-tuned on an organs complete dataset - evaluated across different task types of all organs. The average provides a straightforward measure of overall performance, while the harmonic mean emphasizes consistent performance across all tasks

Model	Avg Fine (%)	HM Fine (%)	Avg Coarse (%)	HM Coarse (%)	Avg Combined (%)	HM Combined (%)
Cardiovascular	47.20	44.73	87.31	86.02	80.21	78.11
GastroIntestinal	51.76	42.80	87.20	85.97	81.30	79.24
Neuropathology	49.48	44.09	72.43	71.12	65.33	63.73
Hematopathology	52.82	46.26	72.53	71.54	68.75	65.94
Base BioMedCLIP	37.68	37.16	55.94	54.73	52.82	50.69

G Detailed Model Souping Results

Table 5: Average (Avg) and harmonic mean (HM) accuracies across fine-, coarse-, and combined-pathology tasks for organ-specific BioMedCLIP (BMC) hybrids (Base + Organ), the all-organ average soup, and the baseline BioMedCLIP model.

Model	Avg Fine (%)	HM Fine (%)	Avg Coarse (%)	HM Coarse (%)	Avg Combined (%)	HM Combined (%)
Cardiovascular + BMC	44.83	40.74	85.52	84.47	78.74	77.60
GastroIntestinal + BMC	53.34	44.35	87.89	86.72	82.14	80.43
Neuropathology + BMC	48.19	42.23	71.95	69.72	67.99	65.85
Hematopathology + BMC	52.91	46.55	72.48	69.56	69.22	65.93
All Organ Average Soup	41.50	38.28	72.46	71.66	67.30	66.59
Base BioMedCLIP	37.68	37.16	55.94	54.73	52.82	50.69

Table 6: Average (Avg) and harmonic mean (HM) accuracies across fine-, coarse-, and combined-pathology tasks for different weight interpolation techniques as described in Section 3.3

Model	Avg Fine (%)	HM Fine (%)	Avg Coarse (%)	HM Coarse (%)	Avg Combined (%)	HM Combined (%)
Linear Average Soup	41.50	38.28	72.46	71.66	67.30	66.59
SLERP Soup	53.95	49.46	91.20	90.92	84.99	84.63
Task Arithmetic Soup	37.55	34.99	59.66	57.89	55.97	54.49
TIES Top 25	33.16	30.84	32.27	30.95	32.60	31.15
WiSE-FT Avg	40.51	37.39	71.28	70.49	66.15	65.47
WiSE-FT SLERP	51.75	47.24	90.50	90.25	84.04	83.75

H Results of Training with Varied Percentages of the Organ - Specific Dataset on the Remaining Organ Domain-Specific Models

We created multiple training subsets (10%, 25%, 50%, 70% of available data) to enable ablation studies on data efficiency and model scalability. Data from each organ category was processed independently to support domain-specific analysis while maintaining consistent evaluation protocols.

Table 7: Fine-grained, coarse-grained & combined dataset-tuned hematopathology model perception accuracy for the hematopathology dataset, measured under varying training dataset proportions from 10% to 70%.

Model & Data	Dataset Percentage (%)	Accuracy ± CI (%)
Coarse Grained Tuned Model	0.10	100.00% (99.80, 100.00)
x Coarse-grained data	0.25	100.00% (99.80, 100.00)
	0.50	100.00% (99.80, 100.00)
	0.70	100.00% (99.80, 100.00)
Fine Grained Tuned Model	0.10	71.28% (65.91, 74.85)
x Fine-grained data	0.25	86.67% (82.09, 88.84)
	0.50	93.08% (89.21, 94.29)
	0.70	94.62% (90.99, 95.53)
Combined Dataset Tuned Model	0.10	97.31% (96.41, 97.73)
x Combined data	0.25	98.85% (98.16, 99.04)
	0.50	99.53% (98.99, 99.57)
	0.70	99.36% (98.78, 99.45)

Table 8: Fine-grained, coarse-grained & combined dataset-tuned neuropathology model perception accuracy for the neuropathology dataset, measured under varying training dataset proportions from 10% to 70%.

Model & Data	Dataset Percentage (%)	Accuracy ± CI (%)
Coarse Grained Tuned Model	0.10	100.00% (99.61, 100.00)
x Coarse-grained data	0.25	100.00% (99.61, 100.00)
	0.50	100.00% (99.61, 100.00)
	0.70	100.00% (99.61, 100.00)
Fine Grained Tuned Model	0.10	60.82% (52.63, 66.24)
x Fine-grained data	0.25	67.53% (59.35, 72.42)
	0.50	70.10% (61.97, 74.75)
	0.70	74.74% (66.75, 78.89)
Combined Dataset Tuned Model	0.10	93.55% (91.68, 94.52)
x Combined data	0.25	94.75% (93.01, 95.58)
	0.50	95.79% (94.16, 96.48)
	0.70	95.79% (94.16, 96.48)

Table 9: Fine-grained, coarse-grained & combined dataset-tuned model perception accuracy for the Cardiovascular Dataset, measured under varying training data proportions from 10% to 70%

Model & Data	Dataset Percentage (%)	Accuracy ± CI (%)
Coarse Grained Tuned Model	0.10	100.00% (98.74, 100.00)
x Coarse-grained data	0.25	100.00% (98.74, 100.00)
	0.50	100.00% (98.74, 100.00)
	0.70	100.00% (98.74, 100.00)
Fine Grained Tuned Model	0.10	58.33% (42.25, 66.46)
x Fine-grained data	0.25	68.33% (51.72, 74.65)
	0.50	75.00% (58.35, 79.80)
	0.70	66.67% (50.11, 73.32)
Combined Dataset Tuned Model	0.10	91.67% (87.39, 93.14)
Combined data	0.25	95.28% (91.57, 96.03)
	0.50	94.44% (90.58, 95.38)
	0.70	96.11% (92.57, 96.66)

This study shows that training with additional data yields only marginal gains with significant improvements achieved with only 10%of the available data. This suggests that our models exhibit strong data efficiency, maintaining high performance with limited data which is a key advantage in biomedical domains where labeled samples are scarce and costly.

Table 10: Fine-grained, coarse-grained & combined dataset-tuned gastrointestinal model perception accuracy for the gastrointestinal dataset, measured under varying training dataset proportions from 10% to 70%.

Model & Data	Dataset Percentage (%)	Accuracy ± CI (%)
Coarse Grained Tuned Model	0.10	99.97% (99.88, 100.00)
x Coarse-grained data	0.25	100.00% (99.88, 100.00)
	0.50	100.00% (99.88, 100.00)
	0.70	100.00% (99.88, 100.00)
Fine Grained Tuned Model	0.10	88.19% (84.86, 89.96)
x Fine-grained data	0.25	92.07% (89.10, 93.38)
	0.50	93.69% (90.91, 94.77)
	0.70	95.15% (92.57, 95.99)
Combined Dataset Tuned Model	0.10	98.47% (97.92, 98.72)
x Combined data	0.25	98.98% (98.50, 99.15)
	0.50	99.06% (98.60, 99.22)
	0.70	99.22% (98.78, 99.36)

Individual Performance of the Models on the Different Datasets: Overall, both coarse-grained dataset tuned and combined dataset tuned models performed best with coarse-grained data, seconded by combined data. They both performed the least on the fine grained data.
The fine-grained dataset tuned model performed the best on the fine-grained dataset, followed by the coarse grained data and the least on the combined data.

Model Performance Across Data Ratios: Across all combined, fine-grained and coarse-grained data sets, the coarse-grained data set tuned and the models tuned to the combined dataset, the highest precision was achieved while training on 70% of the entire data set, while the fine-grained dataset tuned model achieved its highest accuracy at 50% of the entire dataset.

Highest and Lowest Model Accuracy: The combined dataset tuned model had the highest accuracy across all the kinds of dataset with the fine-grained dataset tuned model recorded the least accuracy.

I Detailed Results from Model Souping

To provide a more granular view of performance, Appendix E reports detailed results from both organ-specific hybrids (Base + Organ) and multi-organ soups across cardiovascular, gastrointestinal, neuropathology, hematopathology, and complete pathology settings.

I.1 Organ-Specific Hybrids (Base + Organ) and Baseline

For organ-specific hybrids, BioMedCLIP was interpolated with a single organ adapter and compared against the baseline BioMedCLIP and an all-organ average soup. The tables report fine-, coarse-, and combined-pathology accuracies with 95% confidence intervals, highlighting organ-specific gains and the variability introduced by dataset imbalance.

Table 11: Cardiovascular results - fine, coarse, and combined-pathology accuracy (± CI) for BioMedCLIP hybrids with cardiovascular tuning, compared against all-organ average soup and the base BioMedCLIP model.

Model	Fine Grained ± CI	Coarse Grained ± CI	Combined ± CI
Cardiovascular + BioMedCLIP	73.33% (60.99%, 82.87%)	100.00% (98.74%, 100.00%)	95.56% (92.90%, 97.25%)
Neuropathology + BioMedCLIP	48.33% (36.18%, 60.69%)	63.33% (57.74%, 68.59%)	60.83% (55.70%, 65.74%)
GastroIntestinal + BioMedCLIP	45.00% (33.09%, 57.51%)	95.67% (92.73%, 97.45%)	87.22% (83.38%, 90.41%)
Hematopathology + BioMedCLIP	40.00% (28.57%, 52.63%)	73.33% (68.02%, 78.02%)	67.78% (62.78%, 72.40%)
All Organ Avg Soup	50.00% (37.34%, 62.65%)	84.33% (80.28%, 88.45%)	78.61% (73.26%, 82.81%)
Base BioMedCLIP	38.33% (26.03%, 50.64%)	66.33% (60.99%, 71.68%)	61.67% (55.90%, 65.89%)

Table 12: Gastrointestinal results - fine-, coarse-, and combined-pathology accuracy (± CI) for BioMedCLIP hybrids with gastrointestinal tuning, compared against all-organ average soup and the base BioMedCLIP model.

Model	Fine Grained ± CI	Coarse Grained ± CI	Combined ± CI
Cardiovascular + BioMedCLIP	44.17% (40.03%, 47.84%)	73.92% (72.35%, 75.44%)	68.99% (67.49%, 70.46%)
Neuropathology + BioMedCLIP	51.31% (48.65%, 56.50%)	60.00% (58.48%, 61.82%)	59.82% (58.43%, 61.62%)
GastroIntestinal + BioMedCLIP	93.85% (91.67%, 95.49%)	100.00% (99.88%, 100.00%)	98.98% (98.60%, 99.26%)
Hematopathology + BioMedCLIP	41.42% (37.60%, 45.35%)	61.25% (59.53%, 62.95%)	57.97% (56.37%, 59.54%)
All Organ Avg Soup	55.99% (51.70%, 59.51%)	65.05% (63.27%, 66.62%)	63.55% (61.92%, 65.01%)
Base BioMedCLIP	52.27% (48.23%, 56.20%)	47.01% (45.23%, 48.76%)	49.44% (46.23%, 52.50%)

Table 13: Neuropathology results - fine-, coarse-, and combined-pathology accuracy (± CI) for BioMedCLIP hybrids with neuropathology tuning, compared against all-organ average soup and the base BioMedCLIP model.

Model	Fine Grained ± CI	Coarse Grained ± CI	Combined ± CI
Cardiovascular + BioMedCLIP	39.18% (32.58%, 46.19%)	93.29% (91.54%, 94.70%)	84.26% (82.06%, 86.24%)
Neuropathology + BioMedCLIP	69.07% (62.25%, 75.15%)	100.00% (99.61%, 100.00%)	94.84% (93.42%, 95.97%)
GastroIntestinal + BioMedCLIP	39.69% (33.07%, 46.71%)	84.00% (81.56%, 86.18%)	76.61% (74.09%, 78.95%)
Hematopathology + BioMedCLIP	33.51% (27.24%, 40.41%)	55.01% (51.86%, 58.11%)	51.42% (48.55%, 54.28%)
All Organ Avg Soup	36.60% (29.82%, 43.38%)	79.05% (76.49%, 81.61%)	71.97% (69.08%, 74.24%)
Base BioMedCLIP	35.05% (28.34%, 41.77%)	70.38% (67.51%, 73.26%)	64.49% (61.48%, 66.97%)

Table 14: Hematopathology results - fine-, coarse-, and combined-pathology accuracy (± CI) for BioMedCLIP hybrids with hematopathology tuning, compared against all-organ average soup and the base BioMedCLIP model.

Model	Fine Grained ± CI	Coarse Grained ± CI	Combined ± CI
Cardiovascular + BioMedCLIP	28.21% (23.97%, 32.87%)	80.46% (78.64%, 82.16%)	71.75% (69.89%, 73.54%)
Neuropathology + BioMedCLIP	62.25% (56.71%, 67.92%)	85.22% (83.71%, 86.65%)	80.70% (78.86%, 82.44%)
GastroIntestinal + BioMedCLIP	25.90% (21.80%, 30.47%)	71.69% (69.65%, 73.65%)	64.06% (62.09%, 65.98%)
Hematopathology + BioMedCLIP	93.08% (90.11%, 95.20%)	100.00% (99.80%, 100.00%)	98.85% (98.33%, 99.21%)
All Organ Avg Soup	24.62% (20.36%, 28.88%)	65.69% (63.43%, 67.64%)	58.85% (56.74%, 60.73%)
Base BioMedCLIP	23.08% (18.90%, 27.26%)	40.05% (37.87%, 42.26%)	37.22% (35.23%, 39.14%)

Table 15: Complete pathology results - fine-, coarse-, and combined-pathology accuracy (± CI) across all four organ domains, comparing organ-specific BioMedCLIP hybrids, the all-organ average soup, and the baseline BioMedCLIP model.

Model	Fine Grained ± CI	Coarse Grained ± CI	Combined ± CI
Cardiovascular + BioMedCLIP	39.25% (36.60%, 41.98%)	79.91% (78.91%, 80.88%)	73.16% (72.15%, 74.14%)
Neuropathology + BioMedCLIP	47.90% (45.15%, 50.66%)	69.02% (67.86%, 70.14%)	65.50% (64.43%, 66.56%)
GastroIntestinal + BioMedCLIP	62.25% (59.54%, 64.89%)	88.10% (87.28%, 88.88%)	83.81% (82.96%, 84.62%)
Hematopathology + BioMedCLIP	56.54% (53.79%, 59.25%)	72.79% (71.68%, 73.87%)	70.09% (69.05%, 71.11%)
All Organ Avg Soup	40.29% (37.49%, 42.90%)	68.16% (67.01%, 69.30%)	63.53% (62.40%, 64.57%)
Base BioMedCLIP	37.51% (34.84%, 40.18%)	49.48% (48.25%, 50.71%)	47.49% (46.34%, 48.59%)

I.2 Multi-Organ Soup Accuracy

For multi-organ soups, four organ adapters (Cardiovascular, Gastrointestinal, Neuropathology, Hematopathology) were merged using strategies such as Linear Average, SLERP, Task Arithmetic, TIES, and WiSE-FT. The tables summarize fine-, coarse-, and combined-pathology accuracies with 95% confidence intervals, providing a detailed view of how merging strategies influence cross-organ generalization.

Table 16: Fine-, coarse-, and combined-pathology accuracy (± CI) when evaluated on the cardiovascular dataset across all multi-organ soup strategies: Linear Average, SLERP, Task Arithmetic, TIES, WiSE-FT (Avg), and WiSE-FT (SLERP).

Model Soup	Fine Grained ± CI	Coarse Grained ± CI	Combined ± CI
Linear Average Soup	50.00% (37.34%, 62.65%)	87.31% (80.22%, 88.45%)	78.61% (73.26%, 81.71%)
SLERP Soup	61.67% (45.45%, 86.42%)	99.00% (98.84%, 99.24%)	92.78% (88.65%, 94.05%)
Task Arithmetic Soup β=0.5	38.33% (24.74%, 48.63%)	66.67% (57.49%, 68.63%)	64.44% (58.18%, 68.53%)
TIES Top 25	50.00% (34.73%, 59.28%)	42.33% (37.49%, 47.45%)	42.11% (38.63%, 43.21%)
WiSE-FT Avg	73.33% (57.89%, 82.63%)	100.00% (99.61%, 100.00%)	92.22% (88.09%, 93.71%)
WiSE-FT SLERP	58.33% (42.25%, 66.48%)	98.67% (98.24%, 99.00%)	91.49% (87.70%, 93.37%)

Table 17: Fine-, coarse-, and combined-pathology accuracy (± CI) when evaluated on the gastrointestinal dataset across all multi-organ soup strategies: Linear Average, SLERP, Task Arithmetic, TIES, WiSE-FT (Avg), and WiSE-FT (SLERP).

Model Soup	Fine Grained ± CI	Coarse Grained ± CI	Combined ± CI
Linear Average Soup	55.99% (51.70%, 59.51%)	65.05% (63.27%, 66.62%)	63.55% (61.92%, 65.01%)
SLERP Soup	65.05% (59.51%, 68.93%)	90.80% (89.63%, 91.67%)	88.44% (87.28%, 89.33%)
Task Arithmetic Soup β=0.5	52.67% (51.05%, 54.35%)	52.71% (50.45%, 54.59%)	50.35% (48.12%, 52.41%)
TIES Top 25	28.24% (20.48%, 29.26%)	32.67% (29.25%, 31.43%)	32.21% (30.59%, 32.59%)
WiSE-FT Avg	65.59% (61.28%, 68.91%)	82.05% (80.94%, 83.16%)	80.84% (80.28%, 81.23%)
WiSE-FT SLERP	74.44% (70.30%, 77.28%)	88.40% (85.17%, 89.73%)	87.47% (83.74%, 89.33%)

Table 18: Fine-, coarse-, and combined-pathology accuracy (± CI) when evaluated on the neuropathology dataset across all multi-organ soup strategies: Linear Average, SLERP, Task Arithmetic, TIES, WiSE-FT (Avg), and WiSE-FT (SLERP).

Model Soup	Fine Grained ± CI	Coarse Grained ± CI	Combined ± CI
Linear Average Soup	36.60% (29.82%, 43.38%)	79.05% (76.49%, 81.61%)	71.97% (69.08%, 74.24%)
SLERP Soup	49.54% (43.24%, 54.24%)	83.29% (80.47%, 85.17%)	82.28% (79.47%, 83.46%)
Task Arithmetic Soup β=0.5	36.55% (30.58%, 39.75%)	38.47% (35.39%, 39.47%)	37.71% (36.23%, 38.97%)
TIES Top 25	23.29% (20.58%, 25.67%)	26.73% (25.67%, 28.57%)	24.62% (23.67%, 25.04%)
WiSE-FT Avg	52.03% (48.11%, 54.99%)	81.31% (79.03%, 83.26%)	80.05% (78.34%, 81.44%)
WiSE-FT SLERP	74.44% (70.39%, 77.03%)	85.94% (81.83%, 89.34%)	84.33% (81.67%, 85.95%)

Table 19: Fine-, coarse-, and combined-pathology accuracy (± CI) when evaluated on the hematopathology dataset across all multi-organ soup strategies: Linear Average, SLERP, Task Arithmetic, TIES, WiSE-FT (Avg), and WiSE-FT (SLERP).

Model Soup	Fine Grained ± CI	Coarse Grained ± CI	Combined ± CI
Linear Average Soup	24.62% (20.36%, 28.88%)	65.69% (63.43%, 67.64%)	58.85% (56.74%, 60.73%)
SLERP Soup	50.36% (42.27%, 55.24%)	92.59% (88.27%, 93.57%)	82.65% (80.39%, 83.94%)
Task Arithmetic Soup β=0.5	23.08% (18.93%, 25.67%)	46.82% (43.95%, 48.95%)	42.86% (40.91%, 44.81%)
TIES Top 25	21.54% (17.55%, 25.67%)	25.06% (23.57%, 26.45%)	24.01% (23.14%, 25.34%)
WiSE-FT Avg	40.62% (36.55%, 44.43%)	63.77% (62.47%, 65.02%)	59.92% (58.23%, 61.34%)
WiSE-FT SLERP	52.44% (48.95%, 55.38%)	81.18% (80.13%, 82.35%)	83.25% (81.09%, 84.32%)

Table 20: Fine-, coarse-, and combined-pathology accuracy (± CI) when evaluated on the complete pathology dataset across all multi-organ soup strategies: Linear Average, SLERP, Task Arithmetic, TIES, WiSE-FT (Avg), and WiSE-FT (SLERP).

Model Soup	Fine Grained ± CI	Coarse Grained ± CI	Combined ± CI
Linear Average Soup	40.29% (37.49%, 42.90%)	68.16% (67.01%, 69.30%)	63.53% (62.40%, 64.57%)
SLERP Soup	50.70% (45.15%, 53.79%)	90.34% (89.54%, 90.95%)	84.75% (84.20%, 85.50%)
Task Arithmetic Soup β=0.5	38.33% (33.44%, 40.99%)	54.85% (52.73%, 56.79%)	52.19% (50.91%, 53.29%)
TIES Top 25	25.67% (21.36%, 28.99%)	30.61% (28.96%, 31.85%)	30.75% (29.89%, 31.63%)
WiSE-FT Avg	44.23% (40.65%, 46.97%)	65.75% (64.43%, 67.01%)	61.44% (60.12%, 62.24%)
WiSE-FT SLERP	58.58% (51.08%, 61.72%)	89.53% (87.89%, 90.44%)	83.33% (81.52%, 84.43%)

J Representational Geometry of Adapter Weights

To better understand why some interpolation methods, like SLERP and WiSE-FT Slerp, outperform others, we analyzed how similar the adapters are to each other by measuring the cosine similarity between their weight representations (Figure 6). The analysis shows that Linear Average, SLERP, Task Arithmetic, WiSE-FT Slerp, and WiSE-FT Average are almost identical to each other (cosine similarity >= 0.99). This means that the organ-specific adapters learn very similar patterns rather than distinct or complementary ones. Consequently, approaches designed to leverage highly divergent adapters, such as TIES Top 25, provide limited benefit, as the adapters exhibit minimal variation to exploit.

SLERP and WiSE-FT Slerp perform slightly better because they merge adapters in a smoother, more balanced way that respects their shared structure, avoiding the kind of interference seen in simpler averaging. On the other hand, TIES-Weighted has much lower similarity (around 0.72–0.73), showing that it heavily distorts the learned features. This aligns with its poor empirical performance, as its aggressive pruning removes too much shared information. Overall, these results suggest that biomedical adapters are closely related, and performance differences mainly depend on how carefully each method combines these overlapping representations rather not on discovering new or divergent features.

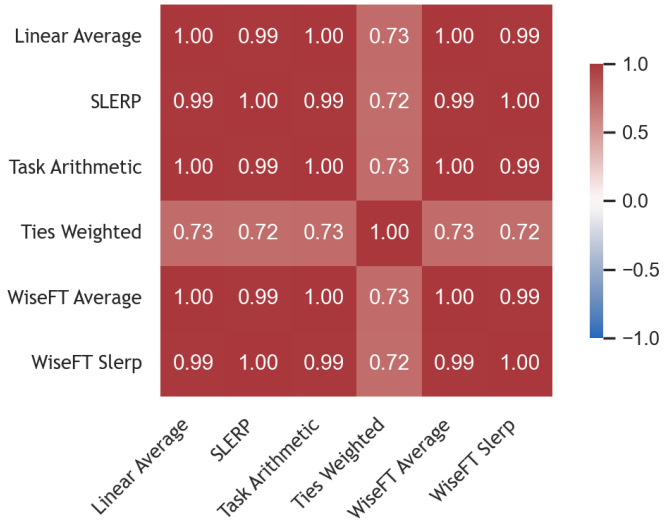

Figure 6: Pairwise cosine similarity between adapter fusion methods, illustrating the representational proximity among Linear Average, SLERP, Task Arithmetic, and WiSE-FT variants, and their divergence from TIES-Weighted. These similarities highlight that most interpolation methods operate within near-identical geometric subspaces, while TIES introduces a distinct representation shift.

References

[1] S. Azizi, L. Culp, J. Freyberg, B. Mustafa, S. Baur, S. Kornblith, T. Chen, P. MacWilliams, S. S. Mahdavi, E. Wulczyn, B. Babenko, M. Wilson, A. Loh, P.-H. C. Chen, Y. Liu, P. Bavishi, S. M. McKinney, J. Winkens, A. Guha Roy, Z. Beaver, F. Ryan, J. Krogue, M. Etemadi, U. Telang, Y. Liu, L. Peng, G. S. Corrado, D. R. Webster, D. Fleet, G. Hinton, N. Houlsby, A. Karthikesalingam, M. Norouzi, and V. Natarajan, "Robust and efficient medical imaging with self-supervision," *arXiv preprint arXiv:2205.09723*, 2022.

[2] C. Perone, P. Ballester, R. Barros, and J. Cohen-Adad, "Unsupervised domain adaptation for medical imaging segmentation with self-ensembling," in *Proc. Int. Conf. Medical Imaging with Deep Learning (MIDL)*, 2018.

[3] D. C. Castro, I. Walker, and B. Glocker, "Causality matters in medical imaging," *arXiv preprint arXiv:1912.08142*, 2019.

[4] Y. Zhai, S. Tong, X. Li, M. Cai, Q. Qu, Y. J. Lee, and Y. Ma, "Investigating the catastrophic forgetting in multimodal large language models," *arXiv preprint arXiv:2309.10313*, 2023.

[5] A. Lozano, J. Nirschl, J. Burgess, S. R. Gupte, Y. Zhang, A. Unell, and S. Yeung-Levy, "mu-bench: A vision-language benchmark for microscopy understanding," *arXiv preprint arXiv:2407.01791*, 2024.

[6] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen *et al.*, "Lora: Low-rank adaptation of large language models," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2022.

[7] A. Prabhakar, Y. Li, K. Narasimhan, S. Kakade, E. Malach, and S. Jelassi, "Lora soups: Merging loras for practical skill composition tasks," *arXiv preprint arXiv:2410.13025*, 2024.

[8] E. Kesim and S. S. Helli, "Multi lora meets vision: Merging multiple adapters to create a multi-task model," *arXiv preprint arXiv:2411.14064*, 2024.

[9] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn. (ICML)*. PMLR, 2021, pp. 8748–8763.

[10] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *Proc. Int. Conf. Mach. Learn. (ICML)*. PMLR, 2022, pp. 12 888–12 900.

[11] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds *et al.*, "Flamingo: A visual language model for few-shot learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 23 716–23 736, 2022.

[12] S. Zhang, Y. Xu, N. Usuyama, H. Xu, J. Bagga, R. Tinn, S. Preston, R. Rao, M. Wei, N. Valluri *et al.*, "Biomedclip: A multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs," *arXiv preprint arXiv:2303.00915*, 2023.

[13] M. Wortsman, G. Ilharco, S. Y. Gadre, R. Roelofs, R. Gontijo-Lopes, A. S. Morcos, H. Namkoong, A. Farhadi, Y. Carmon, S. Kornblith *et al.*, "Model soups: Averaging weights of multiple fine-tuned models improves accuracy without increasing inference time," in *Proc. Int. Conf. Mach. Learn. (ICML)*. PMLR, 2022, pp. 23 965–23 998.

[14] E. Yang, L. Shen, G. Guo, X. Wang, X. Cao, J. Zhang, and D. Tao, "Model merging in llms, mllms, and beyond: Methods, theories, applications and opportunities," *arXiv preprint arXiv:2408.07666*, 2024.

[15] A. Kleiman, G. K. Dziugaite, J. Frankle, S. Kakade, and M. Paul, "Soup to go: Mitigating forgetting during continual learning with model averaging," *arXiv preprint arXiv:2501.05559*, 2025.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction frame a focused goal—improving biomedical VLM performance in organ-centered microscopy by a concrete, novel combination of parameter-efficient LoRA adaptation and adapter-level weight interpolation (Linear Average, SLERP, Task Arithmetic, TIES, WiSE-FT). They clearly bound the scope to BioMedCLIP and μ-Bench organs, state contributions (a systematic benchmark of five merging families and empirical analyses of dataset imbalance/domain similarity), and preview calibrated accuracy gains that are later substantiated in the Results section. The text avoids over-generalization, explicitly backs claims with evaluation data, and thus matches the methodology and evidence presented in the paper.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper clearly discusses its limitations in the dedicated Limitations and Future Work section. It identifies strong assumptions (e.g., high structural similarity across organ datasets) and shows how these constrain divergence-based methods like Task Arithmetic and TIES. It also reflects on dataset imbalance, where large domains dominate smaller ones, and acknowledges the scope of claims by noting that experiments were only run on a single foundation model (BioMedCLIP). Furthermore, it addresses robustness concerns by proposing evaluation on noisier real-world datasets such as TCGA, CAMELYON, and PANDA, and by extending testing to other biomedical and general-purpose VLMs. These acknowledgments demonstrate transparency about the work's boundaries, align with the checklist guideline that empirical results depend on implicit assumptions, and provide a roadmap for how future studies could address these weaknesses.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The paper does not present formal theorems or proofs. However, the mathematical formulations of the interpolation and souping strategies (e.g., Linear Averaging, SLERP, Task Arithmetic, TIES, and WiSE-FT) are clearly defined in the methods section, with equations and details provided in the appendix to ensure reproducibility and correctness.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper fully discloses experimental details needed for reproducibility. All datasets are publicly available (μ-Bench), and the fine-tuning process with LoRA adapters is described with clear hyperparameters. Model souping strategies (Linear Average, SLERP, Task Arithmetic, TIES, and WiSE-FT) are defined mathematically in the methods section and appendix. Tables and figures report average and harmonic mean accuracies across fine-, coarse-, and combined-pathology tasks, with explicit breakdowns by organ system. Together, these details provide sufficient information for independent researchers to replicate the key experiments and validate the claims.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce

the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The experiments use publicly available datasets (μ-Bench) and the BioMedCLIP model. Scripts to reproduce the main experimental results, including training and evaluation of LoRA adapters and adapter soups, will be made publicly available on GitHub upon publication. Instructions, dependencies, and environment details will be included to ensure reproducibility.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies the essential experimental details to understand the results. Training and evaluation are conducted on μ-Bench, with clearly defined organ-specific domains (Cardiovascular, Gastrointestinal, Hematopathology, Neuropathology). LoRA fine-tuning hyperparameters (e.g., rank, learning rate, epochs) and data split strategies are described in the methods and appendix. All interpolation strategies (Linear Average, SLERP, Task Arithmetic, TIES, WiSE-FT) are defined mathematically, with parameter sensitivities (e.g., β, trim thresholds, τ) reported in controlled ablation studies. Together, these details allow readers to interpret the results and replicate the experimental setup, while additional specifications (e.g., optimizer choice and implementation details) are provided in the appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The paper reports confidence intervals for key accuracy metrics, including overall performance and organ-specific evaluations. Variability is captured through accuracy distributions across test splits, and confidence intervals are explicitly computed (95% CI) rather than omitted. While the paper does not include extensive hypothesis testing, the provided confidence bounds are sufficient to establish statistical reliability for the main claims.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper specifies the GPU class used (NVIDIA RTX 6000 Ada) and core training settings (optimizer, epochs, batch size, schedule), memory required, per-run and total estimated compute time.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: The work conforms to the NeurIPS Code of Ethics. It uses publicly available microscopy benchmark datasets (μ-Bench) for evaluation, introduces no new sensitive data, and does not involve human or animal subjects. The research focuses on methodological contributions to biomedical vision–language modeling and avoids privacy, fairness, or misuse concerns.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The paper discusses the broader impact of improving biomedical VLMs, including potential benefits for pathology workflows and medical research by enabling more accurate, resource-efficient analysis of microscopy images. It also acknowledges risks such as possible misuse or over-reliance on automated systems, which could lead to misdiagnoses if models are deployed prematurely. These points are framed as considerations for responsible application of the research.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper does not release new datasets or pretrained models, and the methods are evaluated only on μ-Bench, which is already publicly available. Since no new resources with dual-use risks are introduced, explicit safeguards were not required.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All datasets used in this work, including μ-Bench, are publicly available and were cited appropriately. Their licenses and terms of use (e.g., CC BY 4.0) were respected in all experiments.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We release the trained LoRA adapter soup models described in the paper. The GitHub repository will include instructions for loading and reproducing results, licensing information (CC BY 4.0), and notes on intended use and limitations. The assets will be fully documented upon publication.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.

- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The study did not involve any crowdsourcing or research with human subjects; all experiments were conducted on existing publicly available biomedical microscopy datasets.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The study used only pre-existing, publicly available biomedical microscopy datasets from repositories such as Zenodo, Dataverse, Dryad, and BBBC. All data were shared under permissive licenses (e.g., CC BY 4.0) that allow derivatives and redistribution, and no new data were collected from human participants; therefore, IRB approval was not required.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The research did not employ LLMs as an important, original, or non-standard component of the proposed methods; any LLM use was limited to routine editing of the paper or code-editing assistance, which does not require declaration under the NeurIPS LLM policy.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.