# GRADPATH: TASK-SENSITIVE PATHWAY TUNING FOR MEDICAL IMAGE ANALYSIS

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

## ABSTRACT

Parameter-efficient fine-tuning (PEFT) commonly adapts pretrained models either by inserting newly initialized modules at manually specified locations or by updating independently selected pretrained weights. The former introduces additional computation and placement sensitivity, while the latter overlooks the structural coupling within pretrained networks. We propose GradPath, a structured PEFT method that treats each hidden pathway as a paired adaptation unit consisting of an input-projection row and its corresponding output-projection column. GradPath estimates pathway importance using an activation–gradient sensitivity score and jointly updates the paired projections of the selected pathways. The selected pathways form a bottleneck-shaped subnetwork within the pretrained transformation, without introducing new projection matrices. We further develop GradPath-S, which enhances selected spatial pathways using gated multi-kernel depthwise convolutions. Extensive experiments on medical image classification, segmentation, and few-shot learning demonstrate strong performance across diverse modalities, architectures, and supervision levels. GradPath further reduces training memory, computation, and latency. These results establish structured pretrained pathways as an effective and efficient unit for parameter-efficient medical image adaptation.

## 1 INTRODUCTION

Large pretrained models have substantially advanced natural language processing Chowdhary (2020) and computer vision Szeliski et al. (2011), but full fine-tuning incurs considerable computational and memory costs Thompson et al. (2020); Mahendran (2021) and complicates downstream adaptation Jiang et al. (2025), particularly in specialized domains such as medical imaging Strubell et al. (2019); Ji et al. (2021); Wang et al. (2021). Parameter-efficient fine-tuning (PEFT) alleviates these costs by freezing backbone parameters and updating only a small set of task-specific parameters. For example, Visual Prompt Tuning (VPT) Jia et al. (2022) introduces learnable prompts into the input space, Adapter Houlsby et al. (2019) inserts lightweight modules into the backbone, and SSF Lian et al. (2022) adapts intermediate features using learnable scaling and shifting parameters.

Adapter-based PEFT methods introduce additional parameters and computational branches into pretrained models, and their performance is sensitive to module design, insertion position, and parameter initialization. Moreover, fewer trainable parameters do not necessarily translate into shorter training time because GPUs typically execute pretrained and Adapter layers sequentially, reducing hardware utilization and introducing additional latency. Since architectures differ in their attention, FFN, and activation components, suitable insertion positions may not transfer across tasks.

As shown in Fig. 1, we apply Adapter, MoNA, and DKA at three common insertion positions and evaluate them across multiple medical image classification datasets. The preferred position varies across modules and datasets, with no single placement consistently performing best. These results suggest that fixed external insertion positions may not adapt equally well to different downstream tasks, motivating us to discover task-relevant adaptation structures within the pretrained model.

Unlike Adapter-based methods, GPS Zhang et al. (2024) uses gradients to select a small number of scalar connections from the pretrained model for updating, without introducing new learnable parameters. However, GPS records the selected parameters using an element-wise mask with the same shape as the original weight matrix, incurring additional storage and memory-access overhead. Moreover, connections are selected independently across weight matrices without explicit structural



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

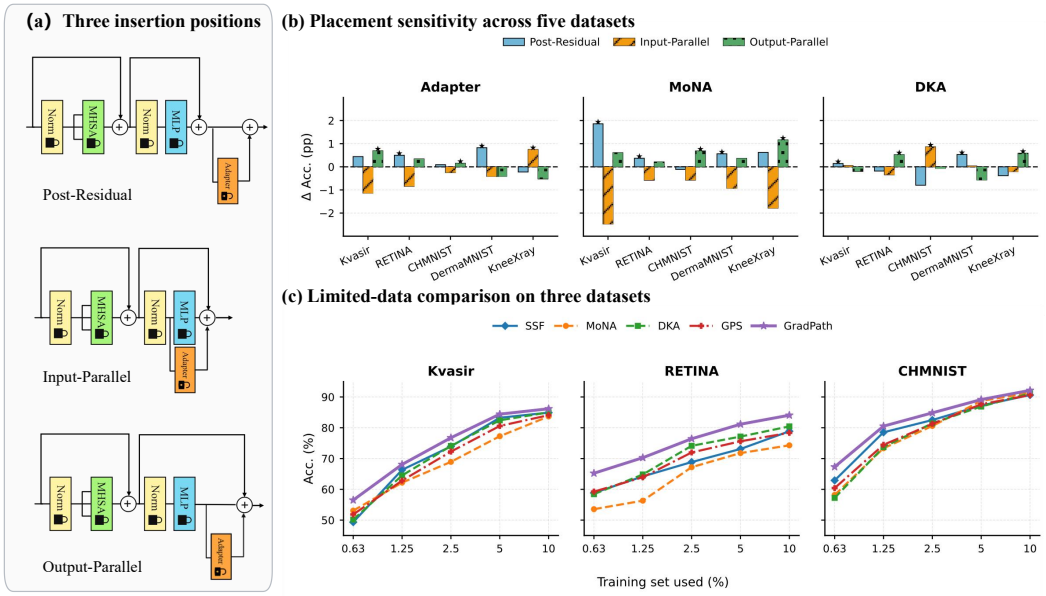



Figure 1: Analysis of Adapter placement and limited-data adaptation. (a) Three insertion positions. (b) Placement sensitivity across medical datasets. (c) PEFT performance under progressively reduced training-data ratios.

constraints, making their reliability more dependent on gradient estimates obtained from limited samples.

This issue is particularly important in low-data medical imaging. Privacy regulations and institutional data-sharing policies leave many medical downstream tasks with only limited labeled data. To simulate this setting, we progressively reduce the training-data ratio from $10\%$ to $0.63\%$ on three medical image classification datasets and compare MoNA, DKA, GPS, and other PEFT methods. Across the 15 dataset–training-ratio combinations, either MoNA or DKA outperforms GPS in 12 cases. Although the performance of different Adapters still varies across tasks, structured adaptation is generally more stable under limited supervision. These results reveal the complementary strengths of the two approaches. Adapters perform structured adaptation through complete down- and up-projections but require newly initialized parameters, whereas GPS directly reuses pretrained weights but treats scalar connections independently. This leads to the following critical question:

*Can we discover a structured parameter subset directly within a pretrained model while retaining pretrained-weight reuse?*

To this end, we use the complete hidden pathway as the basic adaptation unit. Hidden unit $i$ couples input-projection row $\boldsymbol{W}_{1,i,:}$ with output-projection column $\boldsymbol{W}_{2,:,i}$. GradPath ranks each pathway using the product of its activation and loss gradient, and then jointly updates both projections. The selected pathways form a bottleneck-shaped subnetwork within the pretrained FFN, without introducing new projection matrices. GradPath achieves the best performance across all corresponding low-data settings, demonstrating the effectiveness of structured pretrained pathway adaptation.

Medical images additionally require local detail and broader spatial context. Our optional GradPath-Spatial variant applies multi-kernel depthwise convolutions only to selected hidden channels and injects their response through a learnable residual gate. Our contributions are:

- We discover complete FFN pathways and jointly tune their paired input and output projections. The selected pathways form a bottleneck-shaped subnetwork within the pretrained FFN, reusing pretrained weights without introducing new projection parameters.
- We demonstrate the effectiveness of GradPath across medical image classification and segmentation, as well as few-shot adaptation of a biomedical vision–language model, covering diverse architectures and supervision levels.

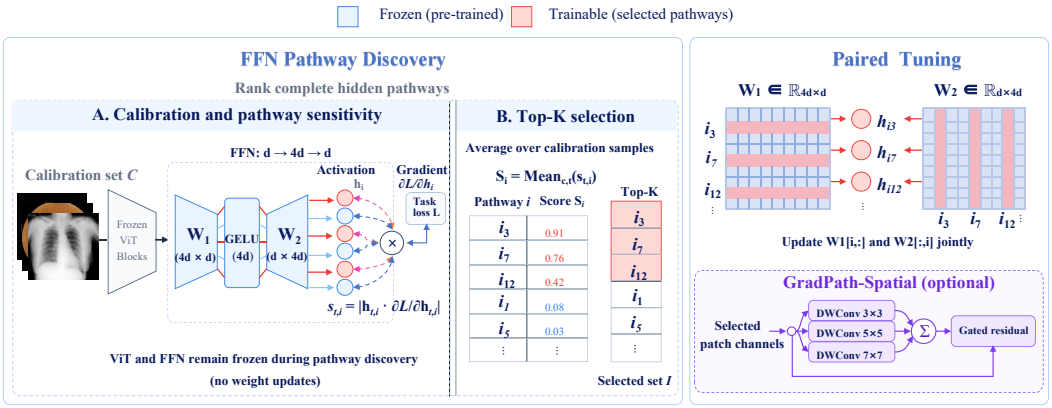



Figure 2: GradPath overview. Activation–gradient scores computed over the training set rank hidden FFN pathways. GradPath jointly tunes the paired input row and output column of each selected pathway; GradPath-Spatial additionally processes selected patch channels with gated multi-kernel depthwise convolutions.

## 2 GradPath: Structured Pathway Tuning

### 2.1 Structured Pathways in Pretrained FFNs

As illustrated in Fig. 2, instead of inserting an external adaptation module, we identify structured adaptation units already present in the pretrained FFN. Consider an FFN:

$$\boldsymbol{h} = \phi(\boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}_1), \qquad \boldsymbol{y} = \boldsymbol{W}_2\boldsymbol{h} + \boldsymbol{b}_2, \tag{1}$$

where $\boldsymbol{W}_1 \in \mathbb{R}^{m\times d}$ and $\boldsymbol{W}_2 \in \mathbb{R}^{d\times m}$. Its output can be expanded as:

$$\boldsymbol{y} = \sum_{i=1}^{m} \boldsymbol{W}_{2,:,i}\, \phi(\boldsymbol{W}_{1,i,:}\boldsymbol{x} + b_{1,i}) + \boldsymbol{b}_2. \tag{2}$$

Each hidden unit $i$ therefore defines a complete pathway: row $\boldsymbol{W}_{1,i,:}$ determines which input features it extracts, while column $\boldsymbol{W}_{2,:,i}$ determines how its activation is written back to the output. For selected indices $\mathcal{I}$ and their complement $\bar{\mathcal{I}}$, Eq. equation 2 can be written exactly as:

$$\boldsymbol{y} = \underbrace{\boldsymbol{W}_{2,:,\bar{\mathcal{I}}}\phi(\boldsymbol{W}_{1,\bar{\mathcal{I}},:}\boldsymbol{x} + \boldsymbol{b}_{1,\bar{\mathcal{I}}})}_{\text{frozen FFN pathways}} + \underbrace{\boldsymbol{W}_{2,:,\mathcal{I}}\phi(\boldsymbol{W}_{1,\mathcal{I},:}\boldsymbol{x} + \boldsymbol{b}_{1,\mathcal{I}})}_{\text{selected trainable pathways}} + \boldsymbol{b}_2. \tag{3}$$

For $K = |\mathcal{I}|$, the second term maps $d \to K \to d$, matching the projection–nonlinearity–projection structure of a bottleneck Adapter. Unlike a conventional Adapter, its projections are inherited from pretrained $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$, and only the corresponding rows and columns are updated; no newly initialized branch is required.

### 2.2 Pathway Discovery

GradPath measures the sensitivity of each hidden pathway using one pass over the full training set $\mathcal{D}_{\text{train}}$. Consider a small relative perturbation to pathway $i$ in block $l$:

$$\Delta h_{t,i}^{(l)} = \rho_i h_{t,i}^{(l)}. \tag{4}$$

A first-order Taylor expansion gives:

$$\Delta\mathcal{L} \approx \sum_i \rho_i h_{t,i}^{(l)} \frac{\partial\mathcal{L}}{\partial h_{t,i}^{(l)}}. \tag{5}$$

Thus, $\left|h_{t,i}^{(l)}\partial\mathcal{L}/\partial h_{t,i}^{(l)}\right|$ measures the empirical first-order sensitivity of pathway $i$. We aggregate this quantity over all training examples and task-relevant tokens $\mathcal{T}$:

$$S_i^{(l)} = \frac{1}{|\mathcal{D}_{\text{train}}||\mathcal{T}|} \sum_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{D}_{\text{train}}} \sum_{t\in\mathcal{T}} \left| h_{t,i}^{(l)} \frac{\partial\mathcal{L}}{\partial h_{t,i}^{(l)}} \right|. \tag{6}$$

GradPath selects and updates the $K_l$ pathways with the largest scores in block $l$:

$$\mathcal{I}^{(l)} = \mathrm{TopK}\left(\{S_i^{(l)}\}_{i=1}^{m}, K_l\right). \tag{7}$$

The first-order justification of this selection rule is provided in Appendix B.

### 2.3 Comparison with External Adapters

Consider a bottleneck Adapter applied after a frozen transformation $\boldsymbol{u} = F(\boldsymbol{x})$:

$$\boldsymbol{h}_A = \phi(\boldsymbol{A}_1\boldsymbol{u} + \boldsymbol{b}_1^A), \qquad \boldsymbol{y} = \boldsymbol{u} + \boldsymbol{A}_2\boldsymbol{h}_A + \boldsymbol{b}_2^A, \tag{8}$$

where $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ are newly introduced down- and up-projection matrices. Let $\boldsymbol{g} = \nabla_{\boldsymbol{y}}\mathcal{L}$ and

$$\boldsymbol{\delta}_A = (\boldsymbol{A}_2^\top \boldsymbol{g}) \odot \phi'(\boldsymbol{A}_1\boldsymbol{u} + \boldsymbol{b}_1^A). \tag{9}$$

Omitting bias gradients, backward propagation gives:

$$\begin{aligned} \nabla_{\boldsymbol{u}}\mathcal{L} &= \boldsymbol{g} + \boldsymbol{A}_1^\top \boldsymbol{\delta}_A, \\ \nabla_{\boldsymbol{A}_2}\mathcal{L} &= \boldsymbol{g}\boldsymbol{h}_A^\top, \\ \nabla_{\boldsymbol{A}_1}\mathcal{L} &= \boldsymbol{\delta}_A\boldsymbol{u}^\top. \end{aligned} \tag{10}$$

Because the Adapter takes the output of $F$ as its input, the two components are executed sequentially. It also introduces additional projections and requires the Adapter-specific activation $\boldsymbol{h}_A$ to be retained for backpropagation.

GradPath instead updates pathways within the pretrained FFN in Eq. equation 1. Define

$$\boldsymbol{\delta} = (\boldsymbol{W}_2^\top \boldsymbol{g}) \odot \phi'(\boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}_1). \tag{11}$$

Its input gradient follows the original FFN computation graph:

$$\nabla_{\boldsymbol{x}}\mathcal{L} = \boldsymbol{W}_1^\top \boldsymbol{\delta}. \tag{12}$$

For the selected pathway set $\mathcal{I}$, only the corresponding gradient slices are applied:

$$\begin{aligned} \nabla_{\boldsymbol{W}_{2,:,\mathcal{I}}}\mathcal{L} &= \boldsymbol{g}\boldsymbol{h}_{\mathcal{I}}^\top, \\ \nabla_{\boldsymbol{W}_{1,\mathcal{I},:}}\mathcal{L} &= \boldsymbol{\delta}_{\mathcal{I}}\boldsymbol{x}^\top. \end{aligned} \tag{13}$$

Thus, GradPath reuses the native FFN activations and pretrained projections, without introducing a serial external branch or an additional input-gradient path. The selected weights are updated in place, preserving the original inference graph.

**Functional interpretation of paired tuning.** Around the selected pre-activation $\boldsymbol{z}_{\mathcal{I}} = \boldsymbol{P}_1\boldsymbol{x}$, let $\boldsymbol{D}_{\mathcal{I}} = \mathrm{diag}(\phi'(\boldsymbol{z}_{\mathcal{I}}))$. The locally linearized transformation of the selected pathways is $\boldsymbol{M}_{\mathcal{I}} = \boldsymbol{P}_2\boldsymbol{D}_{\mathcal{I}}\boldsymbol{P}_1$. Jointly updating both projections induces:

$$\begin{aligned} \Delta\boldsymbol{M}_{\mathcal{I}} = \boldsymbol{P}_2\boldsymbol{D}_{\mathcal{I}}\Delta\boldsymbol{P}_1 + \Delta\boldsymbol{P}_2\boldsymbol{D}_{\mathcal{I}}\boldsymbol{P}_1 \\ + \Delta\boldsymbol{P}_2\boldsymbol{D}_{\mathcal{I}}\Delta\boldsymbol{P}_1. \end{aligned} \tag{14}$$

Although only $K$ pathways are updated, the resulting change can affect the complete input–output feature space, while $\mathrm{rank}(\Delta\boldsymbol{M}_{\mathcal{I}}) \le 2K$. Paired tuning therefore jointly adapts what each pathway extracts and how its activation is written back. The complete derivation is provided in Appendix C.

### 2.4 Selected-Pathway Spatial Enhancement

MoNA Yin et al. (2025) demonstrates that introducing depthwise convolutions into Adapters can substantially improve visual adaptation by incorporating local spatial priors. Such convolutional inductive biases are particularly useful for medical image analysis, where accurate recognition often depends on fine boundaries, small lesions, and subtle texture variations. However, applying spatial operators to all FFN hidden channels introduces unnecessary computation. Inspired by partial convolution Chen et al. (2023), which applies spatial convolution to only a subset of input channels

while leaving the remaining channels unchanged, GradPath-Spatial enhances only the task-relevant pathways selected by GradPath.

Let the selected patch activations be $\boldsymbol{H}_{\mathcal{I}} \in \mathbb{R}^{B\times N\times K}$, where $N = H \times W$. After reshaping them into $\boldsymbol{Z} \in \mathbb{R}^{B\times K\times H\times W}$, we apply multi-kernel depthwise convolutions:

$$\boldsymbol{Z}_{\mathrm{ms}} = \tfrac{1}{3}\left(\mathrm{DWConv}_{3\times3}(\boldsymbol{Z}) + \mathrm{DWConv}_{5\times5}(\boldsymbol{Z}) + \mathrm{DWConv}_{7\times7}(\boldsymbol{Z})\right). \tag{15}$$

The depthwise branches capture spatial context at different scales while preserving the identity of each selected pathway. A learnable gate $\alpha = \sigma(\gamma)$ then combines the original and spatially enhanced features:

$$\boldsymbol{Z}' = (1-\alpha)\boldsymbol{Z} + \alpha\boldsymbol{Z}_{\mathrm{ms}}. \tag{16}$$

We initialize $\gamma < 0$ to limit the influence of the spatial enhancement at the beginning of training. The enhanced features are written back only to the selected patch channels, while unselected channels and non-spatial tokens remain unchanged.

# 3 Experiments

We evaluate GradPath on three medical imaging tasks: image classification, image segmentation, and few-shot learning, to assess its effectiveness across diverse tasks.

## 3.1 Datasets and Metrics

**Medical image classification.** For standard classification, we use seven datasets covering diverse imaging modalities and disease types: BTMRI Nickparvar (2021), COVID Tahir et al. (2021), DermaMNIST Codella et al. (2019), Kvasir Pogorelov et al. (2017), CHMNIST Kather et al. (2016), KneeXray Chen (2018), and RETINA Köhler et al. (2013). We report the classification accuracy (%) on each dataset and the average across all seven datasets.

**Medical image segmentation.** We evaluate on five datasets: BTMRI Cheng (2017), GLaS Sirinukunwattana et al. (2017), Kvasir-SEG Jha et al. (2019), TN3K Gong et al. (2023), and ISIC2017 Codella et al. (2018), covering brain tumor, gland, gastrointestinal polyp, thyroid nodule, and skin lesion segmentation. IoU score (%) is used as the evaluation metric.

**Few-shot learning.** We evaluate the 4-, 8-, and 16-shot settings on seven datasets. Four datasets, BTMRI, COVID, Kvasir, and CHMNIST, overlap with the standard classification benchmark, while CTKIDNEY Islam et al. (2022), LC25000 Borkowski et al. (2019), and OCTMNIST Kermany et al. (2018) are additionally included to cover more organs and imaging modalities.

## 3.2 Baselines

We compare GradPath with full fine-tuning, linear probing, and representative PEFT methods, including Adapter Houlsby et al. (2019), LoRA Hu et al. (2021), VPT Jia et al. (2022), SSF Lian et al. (2022), MoNA Yin et al. (2025), GPS Zhang et al. (2024), and DKA Zhu et al. (2026). For few-shot learning, we additionally include prompt-learning methods such as CoOp Zhou et al. (2022b), CoCoOp Zhou et al. (2022a), and BiomedCoOp Koleilat et al. (2025), as well as representative adaptation methods such as CLIP-Adapter Gao et al. (2024) and Evi-Steer Koleilat et al. (2026).

Detailed dataset descriptions, loss functions, and schematic illustrations of the three task pipelines are provided in Appendix E.

## 3.3 Implementation Details

For fully supervised classification and segmentation, we freeze the pretrained backbone and update only the FFN pathways selected by GradPath and the task head. GradPath-S additionally optimizes its spatial enhancement modules. In few-shot classification, predictions are obtained directly from the cosine similarity between image and class-text features without a task-specific head. Therefore, only the selected FFN pathways in the visual branch are updated, while the text branch remains frozen.

Table 1: Comparison of classification accuracy (%) on seven medical image classification datasets using ViT-B as the backbone. **Bold** and underlined values indicate the best and second-best results, respectively. GradPath denotes our structured FFN pathway tuning method, while GradPath-S denotes its spatially enhanced variant. Average accuracy is computed over all seven datasets.

| Methods | BTMRI | COVID | DermaMNIST | Kvasir | CHMNIST | KneeXray | RETINA | Average |
|---|---|---|---|---|---|---|---|---|
| FULL | 97.44 | 94.43 | 88.68 | 91.92 | 97.21 | 66.55 | 89.35 | 89.37 |
| Linear probing | 91.61 | 89.20 | 84.49 | 90.83 | 92.95 | 54.35 | 86.75 | 84.31 |
| Adapter | 98.37 | 94.25 | 90.21 | 92.42 | 96.88 | 68.90 | 93.85 | 90.70 |
| LoRA | 89.40 | 54.94 | 80.20 | 90.00 | 89.36 | 50.12 | 76.50 | 75.79 |
| VPT | 87.13 | 87.78 | 80.65 | 82.08 | 91.16 | 40.52 | 84.31 | 79.09 |
| SSF | 97.32 | 93.76 | 88.68 | 91.17 | 86.81 | 65.76 | 90.54 | 87.72 |
| MONA | 96.97 | 95.26 | 87.58 | 86.58 | 95.15 | **69.99** | 89.91 | 88.78 |
| GPS | 97.67 | 95.24 | 91.42 | 93.06 | 97.07 | 68.36 | 92.59 | 90.77 |
| DKA | 97.61 | 95.34 | 92.32 | 91.92 | 96.08 | 67.51 | 92.51 | 90.47 |
| GradPath | 97.85 | 95.24 | 92.37 | 92.83 | 97.47 | 67.69 | 93.30 | 90.96 |
| GradPath-S | **98.72** | **95.43** | **92.67** | **93.42** | **97.61** | 68.60 | **94.40** | **91.55** |

For image classification, we adopt ViT-B Dosovitskiy (2020) as the backbone and resize input images to $224 \times 224$. Random horizontal flipping is used for data augmentation. The bottleneck dimension of Adapter-based methods is set to 64, and GradPath selects and updates 64 pathways per FFN block for a fair comparison. Models are trained for 100 epochs using AdamW with a batch size of 32 and a learning rate of $1 \times 10^{-4}$.

For image segmentation, we adopt DINOv3-S+ Siméoni et al. (2025) and extract intermediate features from the 3rd, 6th, 9th, and 12th layers. These features are processed by a dense prediction head based on DPT Ranftl et al. (2021). Input images are resized to $256 \times 256$ and augmented using random rotation, horizontal flipping, and vertical flipping, each applied with a probability of 0.5. The bottleneck dimension of Adapter-based methods is set to 96, and GradPath selects and updates 96 pathways per FFN block, resulting in comparable numbers of trainable parameters. Models are trained for 300 epochs using AdamW with a batch size of 12 and a learning rate of $1 \times 10^{-4}$.

For few-shot learning, we adopt BiomedCLIP Zhang et al. (2023) with a ViT-B/16 visual backbone and use "a photo of a [CLASS]" as the text prompt. Training images are augmented with random resized cropping to a resolution of $224 \times 224$. GradPath is applied only to the visual branch, where 64 pathways are selected and updated in each FFN block. We evaluate the 4-, 8-, and 16-shot settings. Models are trained for 100 epochs using SGD with a batch size of 32 and a learning rate of $2 \times 10^{-3}$. Few-shot results are averaged over three independent runs. All experiments are conducted on an NVIDIA GeForce RTX 4090 GPU.

### 3.4 Main Results

**Medical image classification.** Table 1 reports the results on seven classification datasets. GradPath-S achieves the best average accuracy of 91.55%, improving full fine-tuning by 2.18% and the strongest competing PEFT method, GPS, by 0.78%. GradPath-S achieves the best performance on six of the seven datasets, while GradPath also remains competitive across different medical domains. Together, the two variants demonstrate the effectiveness of structured pathway adaptation for medical image classification. We further evaluate the architecture generalization of GradPath using Swin-T Liu et al. (2021) and ConvNeXt-T Liu et al. (2022). As shown in Table 2, GradPath consistently performs well across different backbone designs. On

Table 2: Classification accuracy (%) across architectures.

| Architecture | Method | Derma | BTMRI | CHMNIST | Kvasir | RETINA |
|---|---|---|---|---|---|---|
| Swin-T | Full-tuning | 92.02 | 98.78 | 97.67 | 92.92 | 93.93 |
| | DKA | 91.32 | 97.79 | 96.54 | 92.00 | 91.64 |
| | GPS | 88.28 | 96.68 | 95.88 | 92.50 | 91.72 |
| | GradPath | **92.67** | **98.06** | **96.81** | **92.92** | **92.98** |
| ConvNeXt-T | Full-tuning | 91.92 | 98.08 | 97.27 | 91.50 | 92.43 |
| | DKA | 86.08 | 97.15 | 95.74 | 90.75 | 87.70 |
| | GPS | 87.88 | 96.56 | 95.94 | 90.17 | 89.67 |
| | GradPath | **89.53** | **97.55** | **96.67** | **90.83** | **90.69** |

Table 3: Comparison of segmentation performance (%) on five medical image segmentation datasets. "Trained Params" denotes the number of trainable parameters in the backbone.

| Methods | BTMRI | GLaS | Kvasir | TN3K | ISIC2017 | Trained Params | Extra Structure |
|---|---|---|---|---|---|---|---|
| FULL | 84.91 | **86.35** | 82.70 | 76.30 | 78.37 | 28.70M | ✗ |
| Adapter | 84.39 | 83.05 | 82.14 | 76.31 | 78.72 | 2.08M | ✓ |
| LoRA | 83.30 | 80.95 | 80.56 | 75.45 | 78.94 | 1.73M | ✓ |
| SSF | 84.53 | 83.42 | 82.24 | 75.92 | 78.60 | 0.10M | ✓ |
| MONA | 84.17 | 84.11 | 82.65 | <u>76.68</u> | 78.76 | 2.24M | ✓ |
| GPS | 84.91 | 83.60 | 82.76 | 76.22 | 78.73 | 0.60M | ✗ |
| DKA | 84.57 | 84.12 | 81.79 | 76.55 | 79.08 | 5.22M | ✓ |
| GradPath | **85.60** | <u>84.45</u> | **84.07** | 76.00 | <u>79.19</u> | 1.28M | ✗ |
| GradPath-S | <u>85.18</u> | 84.43 | <u>83.53</u> | **76.73** | **79.41** | 1.43M | ✓ |

ConvNeXt-T, it achieves the best PEFT result on the five datasets, demonstrating that the proposed structured pathway tuning is not restricted to the standard ViT architecture. The results on Swin-T further evaluate its applicability to hierarchical vision Transformers. We further evaluate the generalization of GradPath on VTAB-1KZhai et al. (2019), with the results reported in Appendix F.

**Medical image segmentation.** Table 3 shows that GradPath and GradPath-S are complementary across dense prediction tasks. GradPath performs best on BTMRI and Kvasir and second best on GLaS and ISIC2017, whereas GradPath-S ranks first on TN3K and ISIC2017 and second on BTMRI and Kvasir. Thus, our variants obtain the best result on four of the five datasets. GradPath uses only 1.28M trainable backbone parameters, compared with 28.70M for full fine-tuning, and introduces no additional structure. GradPath-S increases the trainable count only to 1.43M, showing that spatial enhancement can improve dense prediction with a small parameter overhead.

**Few-shot learning.** Table 4 summarizes the few-shot classification results with BiomedCLIP, with zero-shot performance included as a reference. GradPath achieves the second-best average accuracy in the 4-shot setting and the best average accuracy in both the 8- and 16-shot settings. With 8 and 16 shots, GradPath obtains average accuracies of 81.92% and 85.70%, outperforming Evi-Steer by 1.77% and 1.57%, respectively. Compared with GPS, the corresponding improvements are 7.53% and 5.00%. GradPath further achieves the best performance on four datasets under the 8-shot setting and five datasets under the 16-shot setting. These results demonstrate that jointly adapting complete pretrained pathways provides effective and data-efficient transfer under limited supervision.

**Summary.** Across the three tasks, GradPath consistently improves parameter-efficient transfer without adding task-specific modules. GradPath-S further benefits spatially demanding tasks with only a small increase in trainable parameters, supporting the effectiveness of structured pathway tuning for medical image adaptation. To further examine robustness under extremely limited supervision, we evaluate GradPath under progressively reduced training-data ratios and one- and two-shot settings, with detailed results provided in Appendix G.

**Paired pathway tuning.** Table 5 validates the importance of jointly selecting and updating the paired FFN projections. For hidden unit $i$, row $\boldsymbol{W}_{1,i,:}$ determines which input features are activated, while column $\boldsymbol{W}_{2,:,i}$ controls how the activation is written back to the output. Updating only one projection limits the adaptation capability, whereas independently selecting the two projections may break their hidden-unit correspondence. Gradient-only pairing preserves this correspondence but does not account for pathway activation. By combining activation and gradient information, GradPath identifies complete task-relevant pathways and consistently provides the strongest performance.

**Choosing important hyperparameters.** Table 6 examines two key hyperparameters of GradPath-S. *Selecting the number of updated pathways per layer.* Performance generally improves as $K$ increases from 16 to 64, whereas updating more pathways provides no further benefit. We therefore set $K = 64$ for each layer. *Selecting the initial spatial scale.* A moderately negative $\gamma$ performs best by limiting the spatial enhancement at the beginning of training while allowing it to be gradually learned. Accordingly, we use $\gamma = -3$ by default.

Table 4: Accuracy (%) comparison under different few-shot settings.

| Shots | Method | BTMRI | CHMNIST | COVID | CTKIDNEY | Kvasir | LC25000 | OCTMNIST | Average |
|---|---|---|---|---|---|---|---|---|---|
| 0 | BiomedCLIP | 56.79 | 30.65 | 43.80 | 42.43 | 54.58 | 50.03 | 30.00 | 44.04 |
| 4 | CoOp | 78.12 | 68.92 | 68.64 | 68.12 | 70.89 | 84.30 | 52.77 | 70.25 |
| | CoCoOp | 67.83 | 58.58 | 63.78 | 61.06 | 68.94 | 77.42 | 48.57 | 63.74 |
| | CLIP-Adapter | 58.44 | 34.46 | 49.24 | 47.65 | 56.28 | 53.74 | 52.57 | 50.34 |
| | LP | **78.96** | 71.72 | 69.99 | 67.22 | 69.47 | 86.08 | 44.00 | 69.63 |
| | BiomedCoOp | 77.46 | 72.05 | **74.17** | 69.27 | 72.22 | 86.43 | 52.63 | 72.03 |
| | Evi-Steer | 78.28 | 75.29 | 69.93 | 71.98 | 74.22 | 86.12 | **72.33** | **75.45** |
| | GPS | 69.56 | 70.23 | 60.29 | 68.62 | 68.47 | 81.46 | 59.87 | 68.36 |
| | GradPath (Ours) | 78.28 | **75.93** | 66.73 | **72.15** | **75.25** | **86.65** | 64.80 | 74.26 |
| 8 | CoOp | 77.81 | 74.98 | 74.10 | 77.40 | 77.72 | 87.23 | 60.50 | 75.68 |
| | CoCoOp | 71.71 | 66.58 | 69.36 | 73.93 | 72.94 | 85.56 | 55.40 | 70.78 |
| | CLIP-Adapter | 59.70 | 37.41 | 49.05 | 46.86 | 57.33 | 56.71 | 52.50 | 51.37 |
| | LP | 79.98 | 77.15 | 69.04 | 80.77 | 75.03 | 89.64 | 68.30 | 77.13 |
| | BiomedCoOp | 78.51 | 75.00 | **76.05** | 80.03 | 77.69 | 89.93 | 61.20 | 76.92 |
| | Evi-Steer | **83.42** | **82.78** | 74.69 | 77.71 | 79.64 | 90.39 | 72.40 | 80.15 |
| | GPS | 71.04 | 77.04 | 68.42 | 77.95 | 74.11 | 84.92 | 67.27 | 74.39 |
| | GradPath (Ours) | 82.00 | 81.58 | 75.21 | **83.26** | **80.08** | **91.04** | **80.30** | **81.92** |
| 16 | CoOp | 82.90 | 79.23 | 75.79 | 83.52 | 79.19 | 92.06 | 66.30 | 79.86 |
| | CoCoOp | 78.45 | 72.19 | 74.52 | 77.70 | 75.19 | 87.37 | 60.67 | 75.16 |
| | CLIP-Adapter | 60.69 | 44.22 | 52.04 | 48.63 | 57.28 | 62.18 | 52.77 | 53.97 |
| | LP | 81.98 | 78.13 | 76.17 | 73.37 | 79.36 | 92.64 | 63.40 | 77.86 |
| | BiomedCoOp | 82.53 | 79.19 | 77.85 | 81.59 | 78.30 | 92.40 | 68.13 | 80.00 |
| | Evi-Steer | **86.39** | 84.22 | 73.44 | **90.11** | 82.33 | 92.05 | 80.37 | 84.13 |
| | GPS | 81.54 | 84.57 | 74.53 | 83.22 | 78.47 | 89.72 | 72.87 | 80.70 |
| | GradPath (Ours) | 85.89 | **86.72** | **78.46** | 88.96 | **84.00** | **95.02** | **80.87** | **85.70** |

Table 5: Ablation on parameter-selection strategies.

| Variant | BTMRI | COVID | RETINA |
|---|---|---|---|
| $\boldsymbol{W}_1$ only | 97.61 | 95.13 | 92.90 |
| Gradient Pair | 97.25 | 95.18 | 92.69 |
| $\boldsymbol{W}_2$ only | 97.37 | 94.87 | 92.77 |
| Independent | **97.85** | 94.71 | 92.61 |
| GradPath | **97.85** | **95.24** | **93.30** |

Table 6: Hyperparameter sensitivity of GradPath: (a) the number of selected FFN pathways $K$; (b) the spatial scale $\gamma$ of GradPath-S.

| $K$ | BTMRI | COVID | RETINA |
|---|---|---|---|
| 16 | 97.26 | 94.95 | 92.40 |
| 32 | 97.44 | **95.40** | 92.98 |
| 64 | **97.85** | 95.24 | **93.30** |
| 128 | 97.61 | 94.90 | 93.26 |

| $\gamma$ | BTMRI | COVID | RETINA |
|---|---|---|---|
| $-5$ | 98.37 | 95.12 | 94.16 |
| $-3$ | **98.72** | **95.43** | **94.40** |
| $-1$ | 98.31 | 95.15 | 94.01 |
| 0 | 98.08 | 95.26 | 92.59 |

### 3.5 Effect of path selection across blocks

To investigate the effect of selected FFN paths at different layers, we evaluate the accuracy degradation when excluding selected paths from a specific FFN block or from the top-$k$ FFN blocks. As shown in Fig. 3(a), excluding paths from different blocks causes different levels of performance drop, suggesting that the importance of selected paths varies across layers. Moreover, Fig. 3(b) shows that the accuracy drop increases as more top-ranked blocks are excluded. This observation indicates that informative FFN paths are broadly distributed across the network.

### 3.6 Task sensitivity of selected paths

To evaluate task sensitivity, we compare target-specific FFN paths with transferred and random paths. For each target dataset, the transferred result is the average accuracy obtained by applying the task-specific paths selected from the other three datasets. As shown in Fig. 4(a), target-specific paths consistently outperform both alternatives, while the low cross-task Jaccard overlap in Fig. 4(b) indicates that different tasks tend to select distinct FFN pathways. Together, these results support the task-sensitive nature of GradPath.

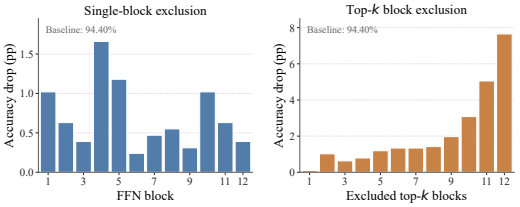



Figure 3: Impacts of different FFN path selection locations. (a) Accuracy drop caused by excluding selected paths from the $k$-th FFN block. (b) Accuracy drop caused by progressively excluding selected paths from the top-$k$ FFN blocks.

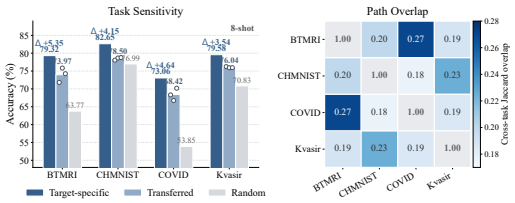
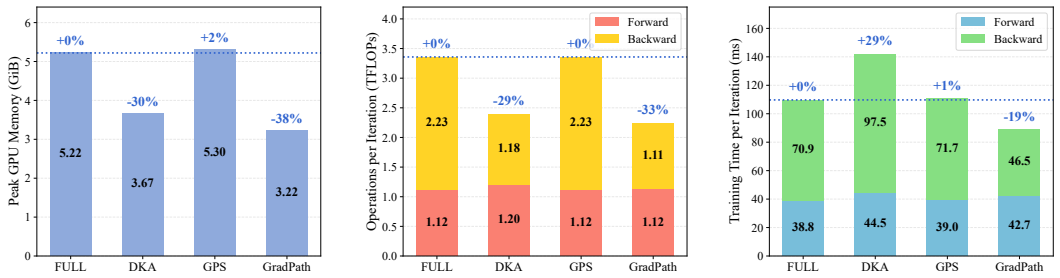



Figure 4: Task sensitivity of selected FFN paths. (a) Performance comparison among target-specific, transferred, and random FFN paths under the 8-shot setting. (b) Jaccard overlap of selected FFN paths across different tasks.




(a) Peak GPU memory. (b) Operations per iteration. (c) Training time per iteration.

Figure 5: Training efficiency of ViT-B on the DermaMNIST classification task with a batch size of 32. We compare (a) peak GPU memory, (b) operations per iteration, and (c) training time per iteration. Dotted lines denote full fine-tuning, and percentages show relative changes in total cost.

### 3.7 All About Efficiency

**Training Memory.** Fig. 5(a) compares peak GPU memory under the same training setting. GradPath requires only 3.22 GiB, reducing memory usage by 38% over full fine-tuning, 39% over GPS, and 12% over DKA.

**Training Computation.** As shown in Fig. 5(b), GradPath reduces the total computation to 2.23 TFLOPs per iteration, 33% lower than full fine-tuning and GPS. This saving mainly comes from avoiding dense weight-gradient computation for frozen FFN parameters.

**Training Time.** Fig. 5(c) shows that GradPath requires 89.2 ms per iteration, achieving a 19% reduction over both full fine-tuning and GPS and a 37% reduction over DKA. These results demonstrate that structured pathway tuning provides practical efficiency gains without introducing additional sequential Adapter branches.

Additional ablations on spatial kernel configurations and Grad-CAM visualizations for the Kvasir-SEG segmentation task are provided in Appendices H and I, respectively.

## 4 Conclusion

We introduced GradPath, a structured PEFT method that discovers and adapts task-sensitive pathways within pretrained transformations. GradPath treats the paired input and output projections of each hidden unit as a complete adaptation unit, preserving their feature extraction and write-back correspondence while reusing pretrained weights. GradPath-S further incorporates multi-scale spatial context into selected pathways with limited overhead. Experiments across classification, segmentation, few-shot learning, different backbones, and limited-supervision settings demonstrate the effectiveness and efficiency of structured pathway adaptation as an alternative to external modules and unstructured parameter selection.

### AI USE STATEMENT

Large language models (LLMs) were used to assist with writing, including grammar correction and improving sentence clarity. They were also used to assist with literature search and to check mathematical derivations. All references suggested by LLMs were manually verified against their original sources, and all AI-assisted mathematical content was independently checked by the authors. The authors reviewed all AI-assisted content and take full responsibility for the final text, claims, and results presented in this work.

### ETHICS STATEMENT

We confirm that this study adheres to the ICLR Code of Ethics. All datasets used in this work are publicly available and have been widely adopted in prior research. This study does not involve the collection of private patient information or new experiments with human subjects.

### REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the proposed method, experimental settings, dataset splits, and evaluation protocols in the main text and Appendix E. Complete theoretical derivations and proofs are provided in Appendices B, C, and D. Code is available at `https://anonymous.4open.science/r/GradPath-2F97`.

## REFERENCES

Andrew A Borkowski, Marilyn M Bui, L Brannon Thomas, Catherine P Wilson, Lauren A DeLand, and Stephen M Mastorides. Lung and colon cancer histopathological image dataset (lc25000). *arXiv preprint arXiv:1912.12142*, 2019.

Jierun Chen, Shiu-hong Kao, Hao He, Weipeng Zhuo, Song Wen, Chul-Ho Lee, and S-H Gary Chan. Run, don't walk: chasing higher flops for faster neural networks. In *2023 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 12021–12031. IEEE, 2023.

Pingjun Chen. Bknee osteoarthritis severity grading dataset, 2018.

Jun Cheng. Brain tumor dataset, 2017.

KR1442 Chowdhary. Natural language processing. *Fundamentals of artificial intelligence*, pp. 603–649, 2020.

Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*, 2019.

Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pp. 168–172. IEEE, 2018.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. *arXiv preprint arXiv:1803.03635*, 2018.

Trevor Gale, Erich Elsen, and Sara Hooker. The state of sparsity in deep neural networks.(2019). *arXiv preprint cs.LG/1902.09574*, 2019.

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International journal of computer vision*, 132(2):581–595, 2024.

Haifan Gong, Jiaxin Chen, Guanqi Chen, Haofeng Li, Guanbin Li, and Fei Chen. Thyroid region prior guided attention for ultrasound segmentation of thyroid nodules. *Computers in biology and medicine*, 155:106389, 2023.

Yunhui Guo, Honghui Shi, Abhishek Kumar, Kristen Grauman, Tajana Rosing, and Rogerio Feris. Spottune: transfer learning through adaptive fine-tuning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4800–4809. IEEE, 2019.

Song Han, Huizi Mao, and William J Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*, 2015.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pp. 2790–2799. PMLR, 2019.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Md Nazmul Islam, Mehedi Hasan, Md Kabir Hossain, Md Golam Rabiul Alam, Md Zia Uddin, and Ahmet Soylu. Vision transformer and explainable transfer learning models for auto detection of kidney cyst, stone and tumor from ct-radiography. *Scientific Reports*, 12(1):11440, 2022.

D Jha, PH Smedsrud, MA Riegler, P Halvorsen, T De Lange, D Johansen, and HD Johansen. Kvasir-seg: A segmented polyp dataset. multimedia modeling. *MMM 2020. Lecture Notes in Computer Science*, 11962, 2019.

Cheng Ji, Fan Wu, Zongwei Zhu, Li-Pin Chang, Huanghe Liu, and Wenjie Zhai. Memory-efficient deep learning inference with incremental weight loading and data layout reorganization on edge systems. *Journal of Systems Architecture*, 118:102183, 2021.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European conference on computer vision*, pp. 709–727. Springer, 2022.

Wenyu Jiang, Zhenlong Liu, Zejian Xie, Songxin Zhang, Bingyi Jing, and Hongxin Wei. Exploring learning complexity for efficient downstream dataset pruning. In *International Conference on Learning Representations*, volume 2025, pp. 67259–67279, 2025.

Jakob Nikolas Kather, Cleo-Aron Weis, Francesco Bianconi, Susanne M Melchers, Lothar R Schad, Timo Gaiser, Alexander Marx, and Frank Gerrit Zöllner. Multi-class texture analysis in colorectal cancer histology. *Scientific reports*, 6(1):27988, 2016.

Daniel S Kermany, Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *cell*, 172(5):1122–1131, 2018.

Thomas Köhler, Attila Budai, Martin F Kraus, Jan Odstrčilik, Georg Michelson, and Joachim Hornegger. Automatic no-reference quality assessment for retinal fundus images using vessel segmentation. In *Proceedings of the 26th IEEE international symposium on computer-based medical systems*, pp. 95–100. IEEE, 2013.

Taha Koleilat, Hojat Asgariandehkordi, Hassan Rivaz, and Yiming Xiao. Biomedcoop: Learning to prompt for biomedical vision-language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14766–14776. IEEE, 2025.

Taha Koleilat, Hassan Rivaz, and Yiming Xiao. Evi-steer: Learning to steer biomedical vision-language models through efficient and generalizable evidential tuning. *arXiv preprint arXiv:2605.26292*, 2026.

Dongze Lian, Daquan Zhou, Jiashi Feng, and Xinchao Wang. Scaling & shifting your features: A new baseline for efficient model tuning. *Advances in Neural Information Processing Systems*, 35: 109–123, 2022.

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 *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 9992–10002. Ieee, 2021.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 11966–11976. IEEE, 2022.

N Mahendran. Analysis of memory consumption by neural networks based on hyperparameters. *arXiv: 2110.11424*, 2021.

Masoud Nickparvar. Brain tumor mri dataset, 2021.

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, et al. 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, 2017.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12159–12168. IEEE, 2021.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Korsuk Sirinukunwattana, Josien PW Pluim, Hao Chen, Xiaojuan Qi, Pheng-Ann Heng, Yun Bo Guo, Li Yang Wang, Bogdan J Matuszewski, Elia Bruni, Urko Sanchez, et al. Gland segmentation in colon histology images: The glas challenge contest. *Medical image analysis*, 35:489–502, 2017.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in nlp. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 3645–3650, 2019.

Richard Szeliski et al. *Computer vision: algorithms and applications*, volume 1. Springer London, 2011.

Anas M Tahir, Muhammad EH Chowdhury, Amith Khandakar, Tawsifur Rahman, Yazan Qiblawey, Uzair Khurshid, Serkan Kiranyaz, Nabil Ibtehaz, M Sohel Rahman, Somaya Al-Maadeed, et al. Covid-19 infection localization and severity grading from chest x-ray images. *Computers in biology and medicine*, 139:105002, 2021.

Neil C Thompson, Kristjan Greenewald, Keeheon Lee, Gabriel F Manso, et al. The computational limits of deep learning. *arXiv preprint arXiv:2007.05558*, 10(2), 2020.

Jian Wang, Hengde Zhu, Shui-Hua Wang, and Yu-Dong Zhang. A review of deep learning on medical image analysis. *Mobile Networks and Applications*, 26(1):351–380, 2021.

Runxin Xu, Fuli Luo, Zhiyuan Zhang, Chuanqi Tan, Baobao Chang, Songfang Huang, and Fei Huang. Raise a child in large language model: Towards effective and generalizable fine-tuning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 9514–9528, 2021.

Dongshuo Yin, Leiyi Hu, Bin Li, Youqun Zhang, and Xue Yang. 5%¿ 100%: Breaking performance shackles of full fine-tuning on visual recognition tasks. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20071–20081. IEEE, 2025.

Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, et al. The visual task adaptation benchmark. 2019.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. Biomedclip: A multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arxiv 2023. *arXiv preprint arXiv:2303.00915*, 70, 2023.

Zhi Zhang, Qizhe Zhang, Zijun Gao, Renrui Zhang, Ekaterina Shutova, Shiji Zhou, and Shanghang Zhang. Gradient-based parameter selection for efficient fine-tuning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28566–28577. IEEE, 2024.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 16795–16804. IEEE, 2022a.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International journal of computer vision*, 130(9):2337–2348, 2022b.

Ziquan Zhu, Hanruo Zhu, Si-Yuan Lu, Xiang Li, Yanda Meng, Yunxiao Zhang, Gaojie Jin, Lu Yin, Lijie Hu, Di Wang, et al. Dual-kernel adapter: Expanding spatial horizons for data-constrained medical image analysis. In *International Conference on Learning Representations*, volume 2026, pp. 116057–116088, 2026.

# Appendix

## CONTENTS

**A Related Work** ..... **14**

**B First-Order Justification of Pathway Selection** ..... **15**

**C Functional Analysis of Paired Pathway Tuning** ..... **15**

**D Descent Property in the Selected Pathway Subspace** ..... **16**

**E Additional Experimental Details** ..... **16**

**F Generalization on the Visual Task Adaptation Benchmark (VTAB-1K)** ..... **17**

**G Performance under Limited Training Data** ..... **20**

**H Selecting Spatial Kernels** ..... **20**

**I Complementary Grad-CAM Visualization** ..... **20**

## A RELATED WORK

### A.1 PARAMETER-EFFICIENT FINE-TUNING.

Parameter-efficient fine-tuning (PEFT) adapts pretrained models by optimizing a small set of task-specific parameters while keeping most backbone weights frozen. Prompt-based methods learn task cues in the input or embedding space: VPT Jia et al. (2022) inserts learnable visual tokens into Vision Transformers, while CoOp Zhou et al. (2022b), CoCoOp Zhou et al. (2022a), and Biomed-CoOp Koleilat et al. (2025) optimize textual prompts for vision-language models. Another line introduces lightweight modules into pretrained blocks. Classical Adapters employ bottleneck projections, whereas recent visual variants incorporate image-specific inductive biases. MoNA Yin et al. (2025) combines multiple visual filters with scaled normalization, while DKA Zhu et al. (2026) integrates small- and large-kernel operators to preserve local details and enlarge spatial context. Despite updating only a small number of parameters, these methods still introduce additional prompt tokens or adaptation modules: the former increases the input sequence length, while the latter adds extra computational branches. Consequently, parameter efficiency does not necessarily translate into computational efficiency. Moreover, the newly introduced parameters must be learned from limited downstream supervision, which may lead to unstable optimization under extreme data scarcity.

### A.2 SUB-NETWORK TRAINING

Sub-network training Pruning technique Gale et al. (2019)Han et al. (2015)uncovers the importance of subnetworks. The lot tery ticket hypothesis Frankle & Carbin (2018) articulates that subnetworks can reach the accuracy of the original model. Fine-tuning sub networks are widely studied. SpotTune Guo et al. (2019) designs a pol icy network to make routing decisions for subset networks. Child-tuning Xu et al. (2021) iteratively updates a subset of parame ters by masking out some gradients during the backward process. However, these methods primarily focus on dynamic routing or training regularization rather than parameter-efficient adaptation by updating a fixed subset of pretrained parameters. More recently, GPS Zhang et al. (2024) ranks scalar connections in the pretrained model according to their gradient magnitudes and updates only the selected weights

without introducing new trainable parameters. However, its standard dense implementation requires an element-wise mask of the same shape as each weight matrix to control parameter updates, resulting in additional storage and irregular memory-access overhead. Moreover, connections across different weight matrices are selected independently without explicit structural constraints.

## B First-Order Justification of Pathway Selection

For clarity, we omit the layer superscript. Let $\mathcal{Q} = \mathcal{D}_{\text{train}} \times \mathcal{T}$ contain all training example–token pairs used to compute the pathway scores. For $q \in \mathcal{Q}$, define

$$a_i^{(q)} = h_i^{(q)} \frac{\partial \mathcal{L}^{(q)}}{\partial h_i^{(q)}}. \tag{17}$$

The score used by GradPath can be written as

$$S_i = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} |a_i^{(q)}|. \tag{18}$$

Suppose that $K$ pathways are selected for update. For a candidate set $\mathcal{I}$ with $|\mathcal{I}| = K$, consider the relative activation perturbation $\Delta h_i^{(q)} = \rho_i^{(q)} h_i^{(q)}$, where

$$\operatorname{supp}(\boldsymbol{\rho}^{(q)}) \subseteq \mathcal{I}, \qquad \|\boldsymbol{\rho}^{(q)}\|_\infty \le \epsilon. \tag{19}$$

**Proposition 1 (First-order optimality of Top-$K$ selection).** The pathway set maximizing the average worst-case first-order loss variation is given by the $K$ largest activation–gradient scores:

$$\begin{aligned} \mathcal{I}^* &= \arg\max_{|\mathcal{I}|=K} \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \max_{\substack{\operatorname{supp}(\boldsymbol{\rho}^{(q)}) \subseteq \mathcal{I} \\ \|\boldsymbol{\rho}^{(q)}\|_\infty \le \epsilon}} \left| (\mathbf{a}^{(q)})^\top \boldsymbol{\rho}^{(q)} \right| \\ &= \operatorname{TopK}\left( \{S_i\}_{i=1}^m, K \right). \end{aligned} \tag{20}$$

**Proof.** For any $q$ and $\mathcal{I}$,

$$\begin{aligned} \left| (\mathbf{a}^{(q)})^\top \boldsymbol{\rho}^{(q)} \right| &\le \sum_{i \in \mathcal{I}} |a_i^{(q)}| |\rho_i^{(q)}| \\ &\le \epsilon \sum_{i \in \mathcal{I}} |a_i^{(q)}|. \end{aligned} \tag{21}$$

The upper bound is attained by choosing $\rho_i^{(q)} = \epsilon \operatorname{sign}(a_i^{(q)})$ for $i \in \mathcal{I}$. Therefore,

$$\begin{aligned} \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \max_{\boldsymbol{\rho}^{(q)}} \left| (\mathbf{a}^{(q)})^\top \boldsymbol{\rho}^{(q)} \right| &= \frac{\epsilon}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \sum_{i \in \mathcal{I}} |a_i^{(q)}| \\ &= \epsilon \sum_{i \in \mathcal{I}} S_i. \end{aligned} \tag{22}$$

This expression is maximized by selecting the $K$ largest values of $S_i$, which proves Eq. equation 20.

## C Functional Analysis of Paired Pathway Tuning

Let

$$\boldsymbol{P}_1 = \boldsymbol{W}_{1,\mathcal{I},:} \in \mathbb{R}^{K \times d}, \qquad \boldsymbol{P}_2 = \boldsymbol{W}_{2,:,\mathcal{I}} \in \mathbb{R}^{d \times K} \tag{23}$$

denote the paired projections of the selected pathways. Around a given activation state $\boldsymbol{z}_\mathcal{I} = \boldsymbol{P}_1 \boldsymbol{x}$, define

$$\boldsymbol{D}_\mathcal{I} = \operatorname{diag}\left( \phi'(\boldsymbol{z}_\mathcal{I}) \right). \tag{24}$$

The locally linearized transformation is

$$\boldsymbol{M}_\mathcal{I} = \boldsymbol{P}_2 \boldsymbol{D}_\mathcal{I} \boldsymbol{P}_1 \in \mathbb{R}^{d \times d}. \tag{25}$$

After updating both projections, $\boldsymbol{P}_1' = \boldsymbol{P}_1 + \Delta\boldsymbol{P}_1$ and $\boldsymbol{P}_2' = \boldsymbol{P}_2 + \Delta\boldsymbol{P}_2$. Holding the local activation pattern fixed, the induced change is

$$\begin{aligned}\Delta\boldsymbol{M}_{\mathcal{I}} = \boldsymbol{P}_2\boldsymbol{D}_{\mathcal{I}}\Delta\boldsymbol{P}_1 + \Delta\boldsymbol{P}_2\boldsymbol{D}_{\mathcal{I}}\boldsymbol{P}_1 \\ + \Delta\boldsymbol{P}_2\boldsymbol{D}_{\mathcal{I}}\Delta\boldsymbol{P}_1.\end{aligned} \tag{26}$$

Each component belongs to $\mathbb{R}^{d\times d}$ and may contain nonzero entries across the complete input–output feature space. Meanwhile,

$$\operatorname{rank}\left(\boldsymbol{P}_2\boldsymbol{D}_{\mathcal{I}}\Delta\boldsymbol{P}_1\right) \le K, \qquad \operatorname{rank}\left(\Delta\boldsymbol{P}_2\boldsymbol{D}_{\mathcal{I}}\boldsymbol{P}_1\right) \le K. \tag{27}$$

Moreover, the first and third terms can be grouped as $(\boldsymbol{P}_2 + \Delta\boldsymbol{P}_2)\boldsymbol{D}_{\mathcal{I}}\Delta\boldsymbol{P}_1$. Consequently,

$$\operatorname{rank}(\Delta\boldsymbol{M}_{\mathcal{I}}) \le 2K. \tag{28}$$

Therefore, paired pathway tuning induces a potentially dense but rank-constrained functional perturbation. The input projection controls what each pathway extracts, whereas the output projection controls how its activation is written back. Updating them jointly preserves this read–write correspondence without introducing an external adaptation branch.

## D Descent Property in the Selected Pathway Subspace

After pathway discovery, the selected set $\mathcal{I}$ is fixed. Let $\Pi_{\mathcal{I}}$ denote the coordinate projection that retains gradients only on the selected pathway parameters. The update can be written as

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t - \eta\Pi_{\mathcal{I}}\nabla\mathcal{L}(\boldsymbol{\theta}^t). \tag{29}$$

By the $\beta$-smoothness of $\mathcal{L}$, for any update $\Delta\boldsymbol{\theta}^t$,

$$\mathcal{L}(\boldsymbol{\theta}^{t+1}) \le \mathcal{L}(\boldsymbol{\theta}^t) + \left\langle\nabla\mathcal{L}(\boldsymbol{\theta}^t), \Delta\boldsymbol{\theta}^t\right\rangle + \frac{\beta}{2}\left\|\Delta\boldsymbol{\theta}^t\right\|_2^2. \tag{30}$$

Substituting $\Delta\boldsymbol{\theta}^t = -\eta\Pi_{\mathcal{I}}\nabla\mathcal{L}(\boldsymbol{\theta}^t)$ and using the fact that $\Pi_{\mathcal{I}}$ is an orthogonal coordinate projection gives

$$\mathcal{L}(\boldsymbol{\theta}^{t+1}) \le \mathcal{L}(\boldsymbol{\theta}^t) - \eta\left\langle\nabla\mathcal{L}(\boldsymbol{\theta}^t), \Pi_{\mathcal{I}}\nabla\mathcal{L}(\boldsymbol{\theta}^t)\right\rangle + \frac{\beta\eta^2}{2}\left\|\Pi_{\mathcal{I}}\nabla\mathcal{L}(\boldsymbol{\theta}^t)\right\|_2^2 \tag{31}$$

$$= \mathcal{L}(\boldsymbol{\theta}^t) - \eta\left(1 - \frac{\beta\eta}{2}\right)\left\|\Pi_{\mathcal{I}}\nabla\mathcal{L}(\boldsymbol{\theta}^t)\right\|_2^2. \tag{32}$$

For $0 < \eta < 2/\beta$, the coefficient of the projected gradient norm is positive. If $\mathcal{L}$ is lower bounded, summing the inequality over iterations yields

$$\sum_{t=0}^{\infty}\left\|\Pi_{\mathcal{I}}\nabla\mathcal{L}(\boldsymbol{\theta}^t)\right\|_2^2 < \infty, \tag{33}$$

and hence

$$\lim_{t\to\infty}\left\|\Pi_{\mathcal{I}}\nabla\mathcal{L}(\boldsymbol{\theta}^t)\right\|_2 = 0. \tag{34}$$

Therefore, GradPath converges to a first-order stationary point in the selected pathway subspace.

## E Additional Experimental Details

Tables 7 and 8 provide detailed information on the classification and segmentation datasets, respectively. We use the cross-entropy loss for classification and combine cross-entropy with Dice loss for segmentation. To prevent background regions from dominating pathway selection in dense prediction, we compute the mean gradient response over foreground and background regions separately and use their sum to determine pathway importance. The complete pipelines for standard classification, segmentation, and few-shot learning are illustrated in Fig. 6 and 7.

Table 7: Overview of the datasets used for standard classification and few-shot learning. C and F denote standard classification and few-shot classification, respectively. Dataset sizes are reported as training/validation/test splits.

| Modality | Organ | Dataset | Classes | Split | Task |
|---|---|---|---|---|---|
| Computed Tomography | Kidney | CTKIDNEY | Kidney cyst, kidney stone, kidney tumor, and normal kidney | 6221/2487/3738 | F |
| Dermatoscopy | Skin | DermaMNIST | Actinic keratosis, basal cell carcinoma, benign keratosis, dermatofibroma, melanocytic nevus, melanoma, and vascular lesion | 7007/1003/2005 | C |
| Endoscopy | Gastrointestinal tract | Kvasir | Dyed lifted polyps, normal cecum, esophagitis, dyed resection margins, normal pylorus, normal Z-line, polyps, and ulcerative colitis | 2000/800/1200 | C/F |
| Fundus Photography | Retina | RETINA | Cataract, diabetic retinopathy, glaucoma, and normal retina | 2108/841/1268 | C |
| Histopathology | Lung/Colon | LC25000 | Colon adenocarcinoma, colon benign tissue, lung adenocarcinoma, lung benign tissue, and lung squamous cell carcinoma | 12500/5000/7500 | F |
| | Colorectal | CHMNIST | Adipose tissue, complex stroma, debris, empty background, immune cells, normal mucosal glands, simple stroma, and tumor epithelium | 2496/1000/1504 | C/F |
| Magnetic Resonance Imaging | Brain | BTMRI | Glioma, meningioma, pituitary tumor, and normal brain | 2854/1141/1717 | C/F |
| Optical Coherence Tomography | Retina | OCTMNIST | Choroidal neovascularization, drusen, diabetic macular edema, and normal retina | 97477/10832/1000 | F |
| X-Ray | Chest | COVID | COVID-19, lung opacity, normal lungs, and viral pneumonia | 10582/4232/6351 | C/F |
| | Knee | KneeXray | No, doubtful, minimal, moderate, and severe osteoarthritis | 5778/826/1656 | C |

Table 8: Details of the medical image segmentation datasets. Dataset sizes are reported as training/validation/test splits. A dash indicates that no separate split is provided.

| Dataset | Modality | Anatomical Region | Segmentation Target | Train/Val/Test |
|---|---|---|---|---|
| BTMRI | MRI | Brain | Brain tumor | 2738/1321/1005 |
| GlaS | Histopathology | Colorectum | Gland | 85/80/– |
| Kvasir-SEG | Endoscopy | Gastrointestinal tract | Polyp | 880/120/– |
| TN3K | Ultrasound | Thyroid | Thyroid nodule | 2303/576/614 |
| ISIC-2017 | Dermoscopy | Skin | Skin lesion | 2000/150/600 |

## F GENERALIZATION TO VTAB-1K

To evaluate whether GradPath generalizes beyond medical imaging, we conduct experiments on the Visual Task Adaptation Benchmark (VTAB-1K) (Zhai et al., 2019), which comprises 19 visual classification tasks grouped into Natural, Specialized, and Structured categories. Following previous work (Jia et al., 2022), all images are resized to $224 \times 224$ and normalized using the ImageNet mean and standard deviation. As shown in Table 9, GradPath achieves the best average performance of 76.93%, outperforming the strongest baseline, GPS, by 1.78%. Although no method performs best on every task due to the diversity of VTAB-1K, GradPath obtains the best result on 11 of the 19 tasks. Fig. 8 further shows that GradPath performs best across all three task groups, with particularly clear gains on the Structured group, demonstrating its effectiveness across diverse visual domains.

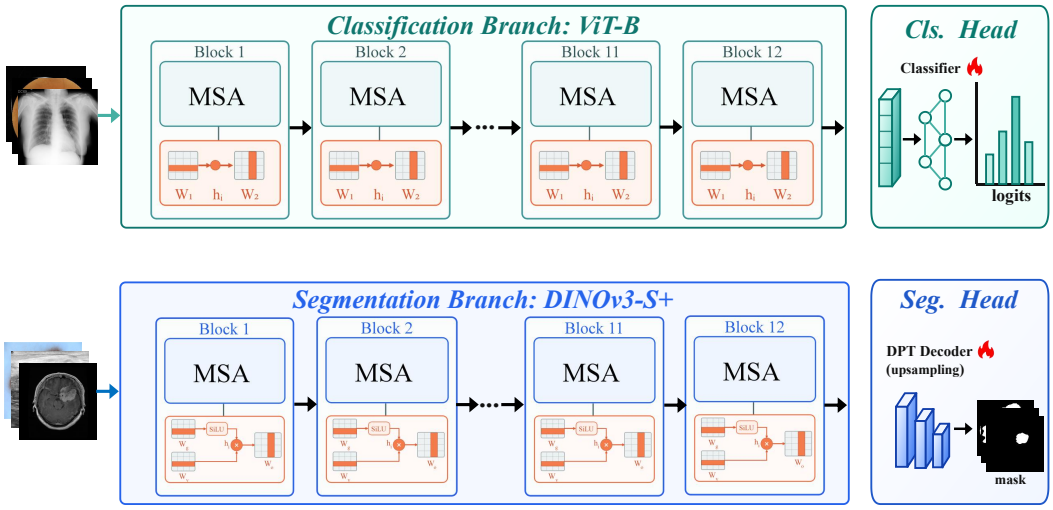



Figure 6: Pipelines for standard medical image classification and segmentation. For classification, GradPath adapts selected pathways in the ViT-B backbone, followed by a trainable classification head. For segmentation, GradPath-S adapts the DINOv3-S+ backbone, and multi-level features are decoded by a DPT-based segmentation head.

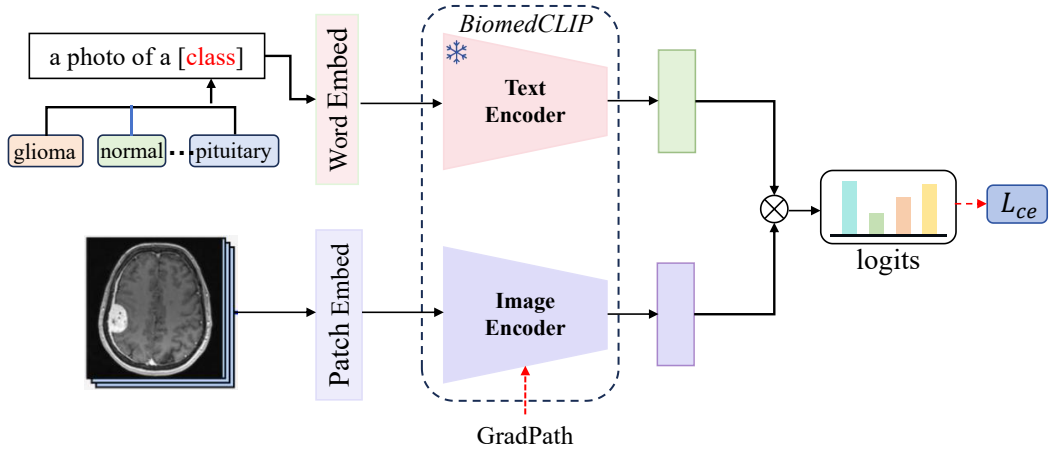



Figure 7: Pipeline for few-shot medical image classification. Class names are inserted into the prompt template and encoded by the frozen text encoder of BiomedCLIP. GradPath adapts only selected pathways in the image encoder, and predictions are obtained from image–text similarities optimized with cross-entropy loss.

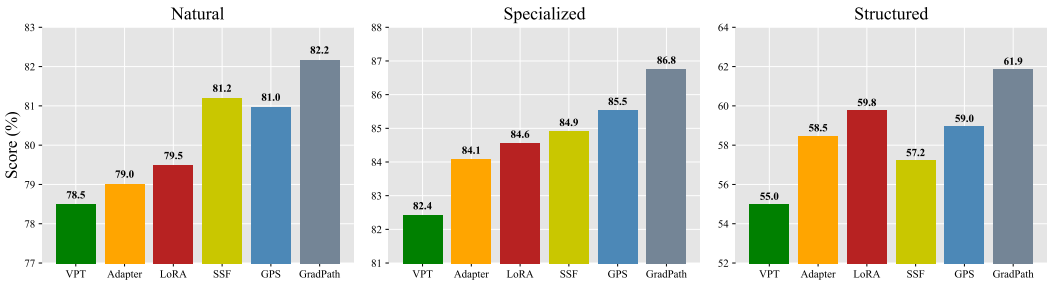



Figure 8: Group-wise average results on VTAB-1K. GradPath achieves the best performance across the Natural, Specialized, and Structured groups, demonstrating consistent transfer across diverse visual domains.

Table 9: Full results on the VTAB-1K benchmark. "Average" denotes the mean of the three group-wise averages shown in Fig. 8. **Bold** and underlined values indicate the best and second-best results among PEFT methods, respectively. GradPath achieves the best result on 11 of the 19 tasks.

| Method | Natural | | | | | | | Specialized | | | | Structured | | | | | | | | Average |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | CIFAR100 | Caltech101 | DTD | Flower102 | Pets | SVHN | Sun397 | Camelyon | EuroSAT | Resisc45 | Retinopathy | Clevr-Count | Clevr-Dist | DMLab | KITTI-Dist | dSpr-Loc | dSpr-Ori | sNORB-Azim | sNORB-Ele | |
| *Traditional Finetuning* | | | | | | | | | | | | | | | | | | | | |
| Full | 68.9 | 87.7 | 64.3 | 97.2 | 86.9 | 87.4 | 38.8 | 79.7 | 95.7 | 84.2 | 73.9 | 56.3 | 58.6 | 41.7 | 65.5 | 57.5 | 46.7 | 25.7 | 29.1 | 68.97 |
| Linear | 64.4 | 85.0 | 63.2 | 97.0 | 86.3 | 36.6 | 51.0 | 78.5 | 87.5 | 68.5 | 74.0 | 34.3 | 30.6 | 33.2 | 55.4 | 12.5 | 20.0 | 9.6 | 19.2 | 57.68 |
| *PETL methods* | | | | | | | | | | | | | | | | | | | | |
| VPT | **78.8** | 90.8 | 65.8 | 98.0 | 88.3 | 78.1 | 49.6 | 81.8 | 96.1 | 83.4 | 68.4 | 68.5 | 60.0 | 46.5 | 72.8 | 73.6 | 47.9 | **32.9** | 37.8 | 71.97 |
| Adapter | 69.2 | 90.1 | 68.0 | 98.8 | 89.9 | 82.8 | 54.3 | 84.0 | 94.9 | 81.9 | 75.5 | 80.9 | 65.3 | 48.6 | 78.3 | 74.8 | 48.5 | 29.9 | 41.6 | 73.86 |
| AdaptFormer | 70.8 | 91.2 | 70.5 | 99.1 | 90.9 | 86.6 | 54.8 | 83.0 | 95.8 | 84.4 | **76.3** | 81.9 | 64.3 | 49.3 | 80.3 | 76.3 | 45.7 | 31.7 | 41.1 | 74.75 |
| LoRA | 67.1 | 91.4 | 69.4 | 98.8 | 90.4 | 85.3 | 54.0 | 84.9 | 95.3 | 84.4 | 73.6 | **82.9** | **69.2** | 49.8 | 78.5 | 75.7 | 47.1 | 31.0 | **44.0** | 74.60 |
| SSF | 74.5 | 89.6 | 72.8 | 99.2 | **91.5** | 84.5 | 56.3 | 84.2 | 95.3 | 84.9 | 75.2 | 79.1 | 66.7 | 47.0 | 77.9 | 73.5 | 50.8 | 26.1 | 36.7 | 74.44 |
| GPS | 71.8 | 90.1 | 70.4 | 99.2 | 91.2 | 88.0 | 56.1 | 85.2 | **96.3** | 86.0 | 74.6 | 79.9 | 65.3 | 50.9 | 81.2 | 71.8 | 53.0 | 31.5 | 38.1 | 75.15 |
| GradPath | 73.7 | **91.8** | **73.56** | **99.5** | 91.3 | **88.9** | **56.44** | **88.3** | 96.2 | **87.1** | 75.4 | 81.1 | 68.9 | **51.7** | **82.1** | **83.7** | **54.50** | 32.1 | 40.9 | **76.93** |

Table 10: Classification accuracy (%) under different training-data ratios. **Bold** and underlined values denote the best and second-best results, respectively.

| Dataset | Method | 0.63% | 1.25% | 2.5% | 5% | 10% |
|---|---|---|---|---|---|---|
| Kvasir | SSF | 49.33 | 66.33 | 73.92 | 83.25 | 84.92 |
| | MoNA | 53.08 | 62.17 | 68.92 | 77.25 | 83.67 |
| | DKA | 50.17 | 64.50 | 74.08 | 82.42 | 85.00 |
| | GPS | 51.83 | 62.83 | 72.25 | 80.58 | 84.08 |
| | GradPath | **56.50** | **68.08** | **76.75** | **84.42** | **86.17** |
| RETINA | SSF | 58.83 | 64.20 | 68.85 | 73.11 | 78.86 |
| | MoNA | 53.55 | 56.31 | 67.19 | 71.77 | 74.29 |
| | DKA | 58.43 | 64.82 | 74.13 | 77.13 | 80.44 |
| | GPS | 59.25 | 63.92 | 71.97 | 75.60 | 78.36 |
| | GradPath | **65.22** | **70.27** | **76.42** | **81.15** | **84.07** |
| CHMNIST | SSF | 62.90 | 78.52 | 82.51 | 87.23 | 90.63 |
| | MoNA | 58.18 | 73.34 | 80.59 | 88.70 | 91.36 |
| | DKA | 57.22 | 73.72 | 81.45 | 86.90 | 91.16 |
| | GPS | 60.40 | 74.46 | 81.25 | 87.63 | 90.62 |
| | GradPath | **67.29** | **80.52** | **84.84** | **89.16** | **92.15** |

## G PERFORMANCE UNDER LIMITED TRAINING DATA

Table 10 compares different PEFT methods under progressively reduced training-data ratios. GradPath achieves the best result in all 15 dataset–ratio combinations, with particularly clear advantages under the lowest-data settings. These results demonstrate that jointly selecting and tuning pretrained pathways provides stable adaptation when supervision is severely limited.

Table 11 further evaluates the more challenging one- and two-shot settings using BiomedCLIP, with zero-shot performance included as a reference. GradPath achieves the best average accuracy in both settings and ranks first in seven of the eight dataset–shot combinations. It outperforms the strongest competing method by 2.15% and 3.02% on average under the one- and two-shot settings, respectively, further demonstrating its effectiveness under extremely limited supervision.

## H SELECTING SPATIAL KERNELS

Table 12 compares different kernel configurations for the selected-pathway spatial enhancement. Multi-kernel configurations generally outperform single-kernel variants, demonstrating the benefit of combining complementary receptive fields. The $[3, 5, 7]$ configuration achieves the best performance across all three datasets and is therefore adopted as the default setting.

## I COMPLEMENTARY GRAD-CAM VISUALIZATION

To qualitatively examine the spatial responses learned by different PEFT methods, we apply Grad-CAM to the last encoder layer on Kvasir-SEG. As shown in Figure 9, GPS produces fragmented responses within the polyp, suggesting that independently selected sparse connections may not preserve spatially coherent representations. DKA introduces large-kernel convolutions into its Adapter to enlarge the receptive field, but its activation also extends substantially into the surrounding background. In contrast, GradPath produces a more compact and continuous response that better aligns with the foreground region. This observation is consistent with its structured pathway tuning, which preserves paired feature extraction and write-back while emphasizing task-relevant regions.

Table 11: Classification accuracy (%) under zero-, one-, and two-shot settings using BiomedCLIP as the backbone. **Bold** and underlined values denote the best and second-best results within each shot setting, respectively. The average is calculated over the four datasets.

| Shots | Method | BTMRI | COVID | Kvasir | CHMNIST | Average |
|---|---|---|---|---|---|---|
| 0 | BiomedCLIP | 56.79 | 43.80 | 54.58 | 30.65 | 46.46 |
| 1 | CLIP-Adapter | 60.26 | 50.48 | 56.25 | 33.62 | 50.15 |
| | CoCoOp | 59.47 | 69.77 | 59.45 | 49.09 | 59.45 |
| | GPS | 60.61 | 59.25 | 62.08 | 53.39 | 58.83 |
| | BiomedCoOp | 64.07 | 70.26 | **62.42** | 58.42 | 63.79 |
| | GradPath | **66.65** | **72.62** | 62.40 | **62.10** | **65.94** |
| 2 | CLIP-Adapter | 57.66 | 45.14 | 55.81 | 32.98 | 47.90 |
| | CoCoOp | 64.12 | 68.80 | 65.50 | 50.84 | 62.32 |
| | GPS | 63.74 | 60.16 | 65.86 | 60.15 | 62.48 |
| | BiomedCoOp | 70.12 | 73.80 | 65.86 | 61.81 | 67.90 |
| | GradPath | **73.40** | **77.29** | **67.69** | **65.29** | **70.92** |

Table 12: Ablation on kernel configurations in GradPath-S. Each list denotes the kernel sizes of the depthwise convolution branches. **Bold** and underlined values indicate the best and second-best results, respectively. The shaded row indicates the default configuration.

| Kernel Settings | BTMRI | COVID | RETINA | Average |
|---|---|---|---|---|
| $[3]$ | 98.48 | 95.37 | 94.01 | 95.95 |
| $[5]$ | 98.31 | 95.15 | 93.53 | 95.66 |
| $[7]$ | 98.37 | 95.27 | 93.75 | 95.80 |
| $[3, 5]$ | 98.60 | 95.40 | 94.16 | 96.05 |
| $[5, 7]$ | 98.43 | 95.38 | 94.16 | 95.99 |
| $[3, 5, 7]$ | **98.72** | **95.43** | **94.40** | **96.18** |

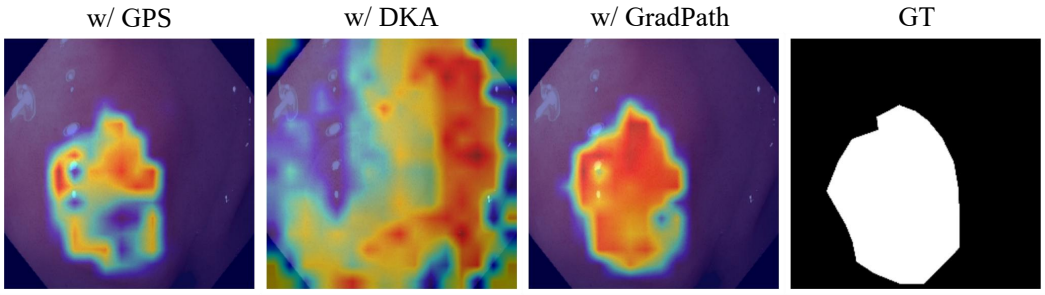



Figure 9: Grad-CAM visualizations on Kvasir-SEG, extracted from the last encoder layer. Warmer colors indicate stronger responses. GPS produces fragmented activations within the polyp, while DKA captures a broader region but also responds strongly to the background. GradPath yields a more continuous and foreground-focused response that better aligns with the ground-truth mask.