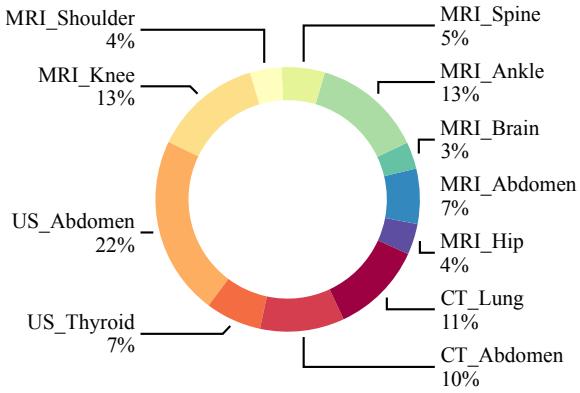
Low-Rank Knowledge Decomposition for Medical Foundation Models

A. Three Pre-training Datasets

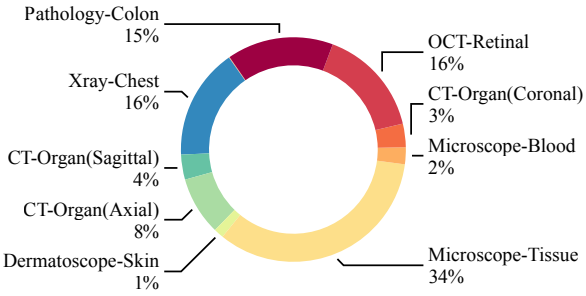
- The dataset RadImageNet [12] consists of 1.35 million images, covering 11 tasks and 3 common modalities. The distribution of diseases is shown in Figure 1. In our experiments, we decompose the pre-trained models on RadImageNet [12] (which have been publicly released by the authors) into 11 lightweight expert models corresponding to task IDs.
- The dataset MedMnist is selected from MedMnistV2 [21], consisting of 705,689 images, covering 10 tasks and 7 different modalities. The distribution of diseases is shown in Figure 2. In our experiments, we decompose the fully pre-trained models on MedMnist into 10 lightweight expert models corresponding to task IDs.
- The dataset Med-ML is a multi-task dataset we constructed, consisting of 119,655 images, covering 8 tasks and 5 different modalities, including APTOS [3], ISIC [4], BUSI [7], Kvasir [15], Shenzhen X-ray [6], Shoulder X-ray [5], VinDr [14] and Bone [8]. The distribution of diseases is shown in Figure 3. In our experiments, we decompose the fully pre-trained models on Med-ML into 8 lightweight expert models corresponding to task IDs.



Radimagenet [12]

Task ID	Name	Modality	Region	Labels	Number
1	Lung	CT	Chest	6	152528
2	Abdomen	CT	Abdomen	28	139825
3	Thyroid	Ultrasound	Neck	2	92599
4	Abdomen	Ultrasound	Abdomen	13	297286
5	Knee	MRI	Knee	18	179555
6	Shoulder	MRI	Shoulder	14	52407
7	Spine	MRI	Spine	9	71674
8	Ankle	MRI	Foot	25	181603
9	Abdomen	MRI	Abdomen	26	91348
10	Brain	MRI	Head	10	44671
11	Hip	MRI	Hip	14	51417

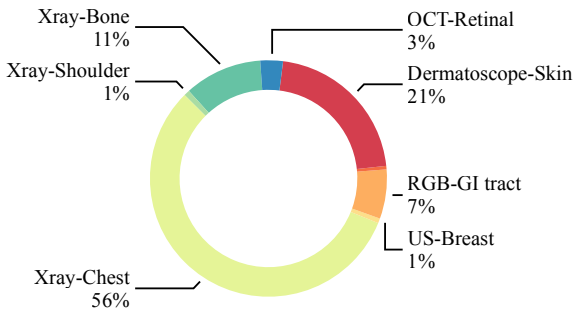
Figure 1. Disease distribution in Radimagenet.



MedMnist

Task ID	Name	Modality	Region	Labels	Number
1	Colon	Pathology	Colon	9	107180
2	Retinal	OCT	Eye	4	109309
3	OrganC	CT	Abdomen	11	23660
4	Cell	Microscope	Blood	8	17092
5	Breast	Ultrasound	Breast	2	780
6	Tissue	Microscope	Kidney cortex	8	236386
7	Skin	Dermatoscope	Skin	7	10015
8	OrganA	CT	Abdomen	11	58850
9	OrganS	CT	Abdomen	11	25221
10	Chest	Xray	Chest	2	112120

Figure 2. Disease distribution in MedMnist.



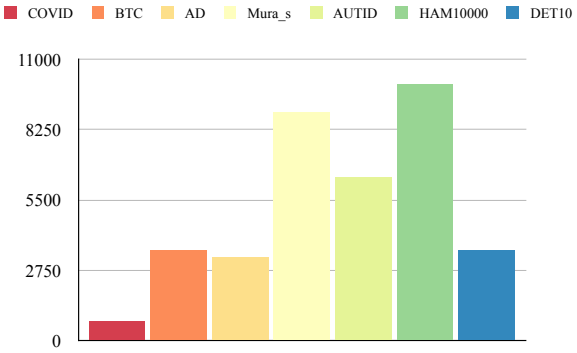
Med-MT

Task ID	Name	Modality	Region	Labels	Number
1	Retinal [3]	OCT	Eye	5	3662
2	Skin [4]	Dermatoscope	Skin	3	25331
3	Breast [7]	Ultrasound	Breast	8	780
4	GI tract [15]	RGB	Gastrointestinal	8	8000
5	Lung [6]	Xray	Chest	2	566
6	Shoulder [5]	Xray	Shoulder	4	945
7	Lung [14]	Xray	Chest	15	67914
8	Bone [8]	Xray	Bone	12	12611

Figure 3. Disease distribution in Med-MT.

B. Seven Downstream Datasets

In the experiments, the downstream datasets we used include COVID [20], BTC [17], AD [1], Mura [16], AUITD [2], HAM10000 [18], and DET10 [11]. These datasets cover five common modalities and are used to thoroughly validate the effectiveness and generalization of our method. The description of these datasets is shown in Figure 4.



Seven Downstream Datasets

Task ID	Name	Modality	Region	Labels	Number
1	COVID [20]	CT	Chest	2	746
2	BTC [17]	MRI	Head	4	3538
3	AD [1]	MRI	Head	4	3264
4	Mura_shoulder [16]	MRI	Shoulder	2	8942
5	AUTID [2]	Ultrasound	Neck	3	6400
6	HAM10000 [18]	Dermatoscope	Skin	7	10015
7	DET10 [11]	Xray	Chest	10	3543

Figure 4. Seven downstream datasets used in our paper.

C. Correspondence between experts and datasets

For STL-based methods, because they train models independently for each task in the pre-training dataset, the trained models can be considered as "expert models" lacking general knowledge. In this case, fine-tuning is only performed when the downstream task matches the model. In the main text, the "-" symbol is used to indicate whether there is a match.

For MTL-based methods, we fine-tune their shared encoders on all downstream datasets since MTL-based methods do not generate task-specific experts.

For KF and our method LoRKD, we fine-tune the corresponding expert models on each downstream dataset. The correspondence between expert models and downstream datasets can be seen in Table 1. The symbol † indicates the absence of a corresponding expert model (due to task or modality mismatch). Following the work of [22], in such cases, we fine-tune a shared backbone that incorporates general knowledge learned from multiple tasks.

Table 1. Correspondence between expert models and downstream datasets.

Pre-trained data	COVID [20]	BTC [17]	AD [1]	Mura_s [16]	AUTID [2]	HAM10000 [18]	DET10 [11]
Radimagenet	Expert_1	Expert_10	Expert_10	Expert_6	Expert_3	†	Expert_1
MedMnist	Expert_10	†	†	†	Expert_5	Expert_7	Expert_10
Med-MT	†	†	†	†	†	Expert_2	†

D. Efficiency Analysis

The role of our EKS conv is to construct personalized low-rank adapters for each sample in the mini-batch, and there is not a unique way to achieve this goal. Thus, to demonstrate the efficiency advantage of EKS conv, we follow FLoRA [19] to compare the computational costs of different methods [9, 19] from a theoretical perspective (as shown in Table. 2). Following [19], we denote b and l as the batch size and the maximum sequence length in the input batch, and $W_0 \in \mathbb{R}^{d\times k}$, $B_i \in \mathbb{R}^{d\times r}$, $A_i \in \mathbb{R}^{r\times k}$. In addition, c_1 and c_2 represent the computational coefficients of batched matmuls (bmm, "φ") and matrix multiplication ("$\circ$") respectively. We also omit the cost of "$\circ$" and set $d = k$ as [19], and T is the number of tasks.

Table 2. Efficiency comparison of different methods for constructing personalized low-rank experts for each sample in a mini-batch.

Method	Improved Operation	Computational Cost
LoRA [9, 19]	$\mathbf{Y} = \mathbf{X}W_0 + \varphi(\varphi(\mathbf{X}, \mathbf{B}), \mathbf{A})$	$2c_1(dblr) + c_2(bld^2)$
FLoRA [19]	$\mathbf{Y} = \mathbf{A} \circ ((\mathbf{B} \circ \mathbf{X})\mathbf{W_0})$	$c_2(rbld^2)$
EKS conv (ours)	$\mathbf{Y} = \mathbf{X}(\mathbf{W_0} + \sum_{i=1}^{T}(\widetilde{\mathbf{BA}} \odot \mathbf{M})_i)$	$Tc_2(rd^2) + c_2(bld^2)$

Since both our EKS conv and FLoRA aim to construct personalized low-rank adapters for each sample in a mini-batch, but their principles for improving efficiency are different. Specifically, FLoRA replaces expensive batched matmuls (bmm) with cheap element-wise multiplications, while we perform parameter fusion before the forward pass of DNNs. If we say the efficiency of EKS conv is better than that of FLoRA, the condition in the following must be satisfied:

$$\frac{rbld^2}{Td^2r + bd^2l} \geq 1 \Longrightarrow \frac{Tr}{bl} + 1 \leq r$$

Note that this inequality holds true in most real-world cases, as $bl > Tr$ and $r > 2$ are common training settings. In addition, using broadcasting to improve efficiency as [19] cannot be widely generalized to convolution operations, while our method is not subject to this limitation.

E. Results on Larger foundation models

Considering that larger foundation models may be encountered in the real world, here we add LVM-Med [13] and BioMedCLIP [23] to further validate the effectiveness of our LoRKD on dataset Med-MT.

Table 3. Comparison of larger foundation models with the best baseline on the Med-MT dataset in terms of decomposition performance.

Table I	Method	Retinal	Skin	Breast	GI tract	Lung	Shoulder	Lung	Bone	Avg
LVM-Med	best baseline	78.14	78.57	77.85	87.94	69.91	79.81	64.37	49.41	73.25
	LoRKD	79.64	82.42	78.76	88.25	75.45	82.69	64.87	53.94	75.75
BioMedCLIP	best baseline	78.14	78.57	77.85	87.94	69.91	79.81	64.37	49.41	73.25
	LoRKD	79.64	80.52	76.89	89.19	77.88	85.58	65.12	52.53	75.92

Table 4. Comparison of larger foundation models with the best baseline on the Med-MT dataset in terms of transferability.

Table II	Method	COVID	BTC	AD	Mura_s	AUITD	HAM10000	DET10	Avg
LVM-Med	best baseline	82.76	76.65	77.48	77.09	97.49	74.92	87.15	81.93
	LoRKD	84.24	79.70	77.23	74.96	97.77	77.28	87.23	82.63
BioMedCLIP	best baseline	82.76	76.65	77.48	77.09	97.49	74.92	87.15	81.93
	LoRKD	84.24	78.68	77.98	76.91	97.49	77.28	87.34	82.96

We can summarize two following points: 1) Compared with results in submission, decomposing the larger foundation models achieves the better decomposition and transferring performance. 2) Compared with results in these two tables, the superiority of our method over best baselines still holds, which confirms the advantage of our method.

F. Discussion of Comparison methods

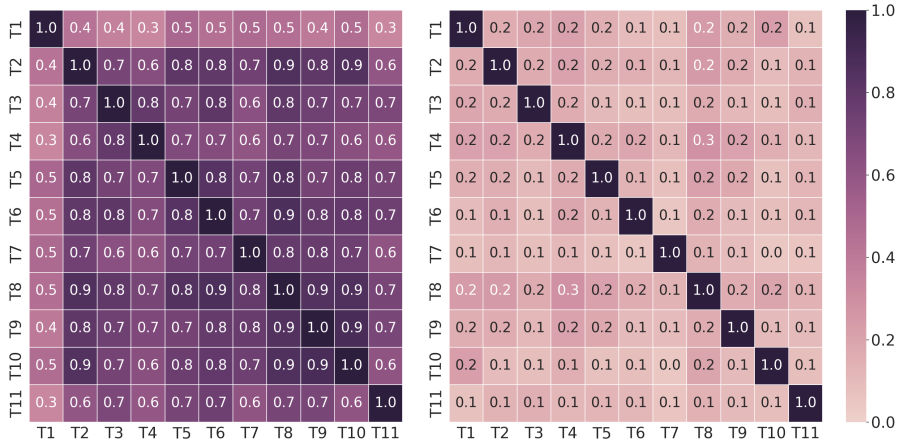
Knowledge decomposition of foundation models to save cost during serving is relatively a new topic, especially in the medical area. The only directly correlated and available baseline is KF [22], which is proposed and verified in natural domains. Thus, we try to verify the effectiveness of LoRKD as much as possible in the following perspectives for comprehensive comparison.

Table 5. Different types of comparison methods

Type	Comparison Methods
Pre-training baseline	Foundation models
Single-task direct training baseline	STL
SOTA Multi-task training baselines	MTL, MoCo-MTL, Aligned-MTL
Knowledge Distillation (**KD**)	STL-KD, MTL-KD
Knowledge Decomposition (**KDe**)	KF

Generally, we would like to form three following points by comparison: 1) By means of both pre-training models and small models decomposed from pre-training models, the downstream task performance should be better than directly training or multi-task collaborative training on the narrow downstream data. 2) Small models decomposed from pre-training models will maintain and even outperform the performance of pre-training models in specific tasks, due to the merits of distillation on pre-training model. 3) Naive distillation from pre-training models to a specific model is not better than the distillation from pre-training models to our mixture of Low-rank Expert modules as we consider the heterogeneity harmony and the task collaboration benefits in design.

G. Knowledge Disentanglement



(a) MTL (b) Ours

Figure 5. The CKA feature similarity matrices of MTL and Ours.

Figure 5 shows the Centered Kernel Alignment (CKA) feature similarity matrices [10] of our method and MTL on Radimagenet dataset. It is evident that our method exhibits significantly lower CKA feature similarity between different tasks compared to MTL, which confirms the knowledge disentanglement ability of our method. This phenomenon can be attributed to our low-rank expert modules being embedded at the convolutional level, which facilitates the simultaneous decomposition of shallow knowledge and deep knowledge. Meanwhile, our proposed efficient knowledge decomposition convolution ensures that this knowledge decomposition pattern can be achieved at a low cost.

H. Notation table

We add a notation table here to ease reading which is summarized as below.

Notation	Description	Shape
W_0	Shared weight in backbone	$\mathbb{R}^{C^{\text{out}} \times C^{\text{in}} \times k \times k}$
B_t	Low rank factors	$\mathbb{R}^{C^{\text{out}} k \times rk}$
A_t	Low rank factors	$\mathbb{R}^{rk \times C^{\text{in}} k}$
h_t	Input features	$\mathbb{R}^{B \times C^{\text{in}} \times H \times W}$
g_t	Output features	$\mathbb{R}^{B \times C^{\text{out}} \times H \times W}$
o_{ij}	Output feature unit	$\mathbb{R}^{B \times C^{\text{out}}}$
$h_{(i)(j)}$	Input feature unit	$\mathbb{R}^{B \times C^{\text{in}}}$
ω	Convolution weight unit	$\mathbb{R}^{C^{\text{in}} \times C^{\text{out}}}$
M	Task label	$\mathbb{R}^{B \times T}$
W'	Aggregated weight	$\mathbb{R}^{B \times C^{\text{out}} \times C^{\text{in}} \times k \times k}$

References

[1] Alzheimer's dataset, Kaggle dataset. https://www.kaggle.com/datasets/tourist55/alzheimers-dataset-4-class-of-images. 2

[2] Algerian ultrasound images thyroid dataset: Auitd,Kaggle dataset. https://www.kaggle.com/datasets/azouzmaroua/algeria-ultrasound-images-thyroid-dataset-auitd. 2

[3] Aptos 2019 blindness detection,Kaggle dataset. https://www.kaggle.com/c/aptos2019-blindness-detection/data. 1, 2

[4] Skin lesion images for melanoma classification,Kaggle dataset. https://www.kaggle.com/datasets/andrewmvd/isic-2019. 1, 2

[5] Shoulder x-ray classification,Kaggle dataset. https://www.kaggle.com/datasets/dryari5/shoulder-xray-classification. 1, 2

[6] Lung masks for shenzhen hospital chest x-ray set,Kaggle dataset. https://www.kaggle.com/datasets/yoctoman/shcxr-lung-mask. 1, 2

[7] Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in brief*, 28: 104863, 2020. 1, 2

[8] Safwan S Halabi, Luciano M Prevedello, Jayashree Kalpathy-Cramer, Artem B Mamonov, Alexander Bilbily, Mark Cicero, Ian Pan, Lucas Araújo Pereira, Rafael Teixeira Sousa, Nitamar Abdala, et al. The rsna pediatric bone age machine learning challenge. *Radiology*, 290(2):498–503, 2019. 1, 2

[9] 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. 3

[10] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pages 3519–3529. PMLR, 2019. 4

[11] Jingyu Liu, Jie Lian, and Yizhou Yu. Chestx-det10: Chest x-ray dataset on detection of thoracic abnormalities, 2020. 2

[12] Xueyan Mei, Zelong Liu, Philip M Robson, Brett Marinelli, Mingqian Huang, Amish Doshi, Adam Jacobi, Chendi Cao, Katherine E Link, Thomas Yang, et al. Radimagenet: an open radiologic deep learning research dataset for effective transfer learning. *Radiology: Artificial Intelligence*, 4(5):e210315, 2022. 1

[13] Duy MH Nguyen, Hoang Nguyen, Nghiem Diep, Tan Ngoc Pham, Tri Cao, Binh Nguyen, Paul Swoboda, Nhat Ho, Shadi Albarqouni, Pengtao Xie, et al. Lvm-med: Learning large-scale self-supervised vision models for medical imaging via second-order graph matching. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[14] Ha Q Nguyen, Khanh Lam, Linh T Le, Hieu H Pham, Dat Q Tran, Dung B Nguyen, Dung D Le, Chi M Pham, Hang TT Tong, Diep H Dinh, et al. Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations. *Scientific Data*, 9(1):429, 2022. 1, 2

[15] Konstantin Pogorelov, Kristin Ranheim Randel, Carsten Griwodz, Sigrun Losada Eskeland, Thomas de Lange, Dag Johansen, Concetto Spampinato, Duc-Tien Dang-Nguyen, Mathias Lux, Peter Thelin Schmidt, Michael Riegler, and Pål Halvorsen. Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection. In *Proceedings of the 8th ACM on Multimedia Systems Conference (MMSYS)*, pages 164–169, 2017. 1, 2

[16] Pranav Rajpurkar, Jeremy Irvin, Aarti Bagul, Daisy Ding, Tony Duan, Hershel Mehta, Brandon Yang, Kaylie Zhu, Dillon Laird, Robyn L Ball, et al. Mura: Large dataset for abnormality detection in musculoskeletal radiographs. *arXiv preprint arXiv:1712.06957*, 2017. 2

[17] Ahmad Saleh, Rozana Sukaik, and Samy S. Abu-Naser. Brain tumor classification using deep learning. In *2020 International Conference on Assistive and Rehabilitation Technologies (iCareTech)*, pages 131–136, 2020. 2

[18] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9, 2018. 2

[19] Yeming Wen and Swarat Chaudhuri. Batched low-rank adaptation of foundation models. *arXiv preprint arXiv:2312.05677*, 2023. 3

[20] Yang Xingyi, He Xuehai, Zhao Jinyu, Zhang Yichen, Zhang Shanghang, and Xie Pengtao. Covid-ct-dataset: a ct image dataset about covid-19. *arXiv preprint arXiv:2003.13865*, 2020. 2

[21] Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023. 1

[22] Xingyi Yang, Jingwen Ye, and Xinchao Wang. Factorizing knowledge in neural networks. In *European Conference on Computer Vision*, pages 73–91. Springer, 2022. 2, 4

[23] 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 preprint arXiv:2303.00915*, 2023. 3