The Fortieth AAAI Conference on Artificial Intelligence (AAAI-26)

MedSAMix: A Training-Free Model Merging Approach for Medical Image Segmentation

Yanwu Yang[1,2*], Guinan Su[3*], Jiesu Hu[4,5*], Francesco Sammarco[1],
Jonas Geiping[3,6,7†], Thomas Wolfers[1,2,8 †]

[1]Department of Psychiatry and Psychotherapy, University Hospital Tübingen, Tübingen, Germany
[2]German Center for Mental Health (DZPG), partner site Halle/Jena/Magdeburg, Germany
[3]Max Planck Institute for Intelligent Systems
[4]Harbin Institute of Technology at Shenzhen, Shenzhen, China
[5]Peng Cheng Laboratory, Shenzhen, China
[6]ELLIS Institute Tübingen
[7]Tübingen AI Center
[8]Department of Psychology, Friedrich Schiller University of Jena, Germany
yangyanwu1111@gmail.com

Abstract

Universal medical image segmentation models have emerged as a promising paradigm due to their strong generalizability across diverse tasks, showing great potential for a wide range of clinical applications. This potential has been partly driven by the success of general-purpose vision models such as the Segment Anything Model (SAM), which has inspired the development of various fine-tuned variants for medical segmentation tasks. However, fine-tuned variants like MedSAM are trained on comparatively limited medical imaging data that often suffers from heterogeneity, scarce annotations, and distributional shifts. These challenges limit their ability to generalize across a wide range of medical segmentation tasks. In this regard, we propose MedSAMix, a training-free model merging method that integrates the strengths of both generalist models (e.g., SAM) and specialist models (e.g., MedSAM) for medical image segmentation. In contrast to traditional model merging approaches that rely on manual configuration and often result in suboptimal outcomes, we propose a zero-order optimization method to automatically discover optimal layer-wise merging solutions. Furthermore, for clinical applications, we develop two regimes to meet the demand of domain-specificity and generalizability in different scenarios by single-task optimization and multi-objective optimization respectively. Extensive evaluations on 25 medical segmentation tasks demonstrate that MedSAMix effectively mitigates model bias and consistently improves performance in both domain-specific accuracy and generalization, achieving improvements of 6.67% on specialized tasks and 4.37% on multi-task evaluations.

Code — https://github.com/podismine/MedSAMix
Model — https://huggingface.co/guinansu/MedSAMix

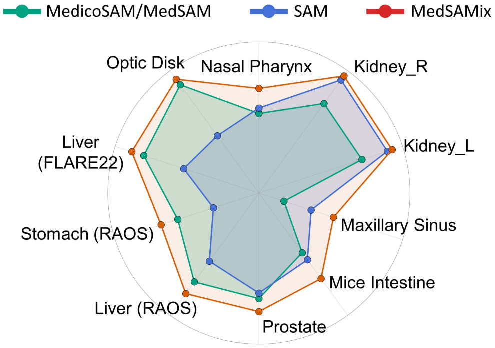


Figure 1: Segmentation performance of different SAM-based models across diverse medical tasks, evaluated using the Dice coefficient. Blue lines indicate the superior results between MedSAM and MedicoSAM, while green lines denote the performance of the original SAM. Although MedSAM and MedicoSAM are domain-specific, they still underperform on certain tasks. In contrast, our MedSAMix (red) consistently achieves higher performance across all tasks.

Introduction

Universal medical image segmentation has become a widely adopted solution for diverse medical imaging tasks, enabling broad applicability without the need for extensive annotations or clinical expertise (Butoi et al. 2023; Czolbe and Dalca 2023). These generalizable models facilitate adaptability to domain shifts and even generalize to previously unseen tasks (Hu et al. 2025b). Fine-tuned models derived from general-purpose foundation models for segmentation, such as the Segment Anything Model (SAM) (Kirillov et al. 2023), represent a significant step toward universal medical image segmentation, e.g., MedSAM (Ma et al. 2024a) and MedicoSAM (Archit, Freckmann, and Pape 2025).

Despite the promising results of these fine-tuned variants

---

[*]These authors contributed equally.
[†]Senior authors

in universal medical image segmentation, we find their performance unbalanced across different tasks. As shown in Fig. 1, MedSAM/MedicoSAM achieve strong results on certain familiar tasks, such as optic disk and liver segmentation. However, on others such as kidney and maxillary sinus segmentation they even underperform the original SAM (see additional results in the Results section). This highlights the stronger generalization capability of SAM, despite the absence of domain-specific optimization, and reveals the limited adaptability of its fine-tuned variants.

This can be mainly attributed to the inherent complexities of medical imaging data. Due to heterogeneity, domain shifts, and class imbalance, medical imaging datasets introduce complex optimization landscapes during fine-tuning, often leading models to converge to suboptimal local minima (Sanjeev et al. 2024; Li et al. 2020). These complexities make fine-tuned models such as MedSAM more susceptible to suboptimal generalization. In contrast, SAM, which is trained on large-scale natural image datasets with smoother optimization landscapes may retain stronger global generalization and thus outperform MedSAM on certain medical tasks. Furthermore, fine-tuned variants are often affected by catastrophic forgetting during adaptation, especially in the absence of strategies to preserve their original generalization capabilities. As a result, these models may lose part of their broader segmentation ability when adapted to the medical imaging domain (Kemker et al. 2018; Aleixo et al. 2023).

This raises a critical question: **How can we enhance domain-specific capabilities while mitigating the compromise of generalization?** Noting that fine-tuned models initialized from the same pre-trained weights often converge to similar loss basins (Neyshabur, Sedghi, and Zhang 2020), model merging has emerged as an effective strategy to unify diverse solution modes into a single model without additional training (Su and Geiping 2025; Su et al. 2025; Yadav et al. 2024). By integrating parameters or representations from multiple models, model merging provides a promising approach to improve performance and mitigate single-model biases, resulting in more stable, diverse, and generalizable predictions (Almakky et al. 2024). As medical models are often trained separately on data from different clinical centers due to privacy constraints, model merging offers a promising solution for effectively integrating these models without requiring data sharing, and its potential in the medical domain remains largely underexplored. Moreover, most existing model merging methods in other domains either rely on manually crafted configurations, which often result in suboptimal performance (Maron et al. 2022; Yadav et al. 2024), or require computationally intensive merging during the training process (Sanjeev et al. 2024; Qazi et al. 2024), making them particularly costly for large foundation models such as SAM and MedSAM. In addition, these methods typically lack support for multi-objective optimization, making it difficult to ensure the generalization ability of the merged models.

In this study, we propose MedSAMix, an efficient training-free model merging framework to balance generalization and domain-specific capabilities for SAM-based medical image segmentation. Specifically, we explore the potential of MedSAMix from two perspectives: (1) **Expert capability:** MedSAMix merges model variants through single-task optimization using only a few calibration samples, tailored to task-specific distributions and improved performance without retraining. (2) **General capability:** We introduce a multi-objective optimization to capture diverse aspects of model performance across tasks for improving universal medical image segmentation.

MedSAMix employs a zero-order optimization approach that selects merging configurations based on their empirical performance during the search, enabling efficient exploration of the solution space given only a few samples. This allows MedSAMix to adaptively balance task-specificity and generalization by combining model variants, while mitigating suboptimal generalization and fine-tuning issues. In summary, our main contributions are as follows:

- We propose MedSAMix, a training-free model merging method leveraging the strengths of general and expert knowledge for medical image segmentation.
- We introduce flexible ways of merging for MedSAMix to accommodate different scenarios: single-task merging focuses on expert capabilities for specific domain tasks, while multi-task merging promotes generalization across diverse tasks by jointly optimizing multiple objectives.
- Extensive experiments on 25 medical image segmentation tasks show that MedSAMix facilitates the enhancement of expert-level performance and generalization without retraining, achieving improvements of 6.67% on specialized tasks and 4.37% on multi-task evaluations.

Related Work

1. Universal Medical image segmentation

Recently, universal models for medical image segmentation have become powerful tools for medical image segmentation without retraining (Ma et al. 2024a; Butoi et al. 2023; Zhao et al. 2025; Hu et al. 2025a). These models typically leverage image-mask pairs or symbolic prompts such as bounding boxes and points, offering strong zero-shot generalization capabilities. Among them, in-context learning (ICL) methods (Hu et al. 2024; Gao et al. 2025) relate rich spatial alignment information by utilizing image-mask pairs as prompts, including UniverSeg (Butoi et al. 2023), SegGPT (Wang et al. 2023), Neuralizer (Czolbe and Dalca 2023), Tyche (Rakic et al. 2024) and Neuroverse3D (Hu et al. 2025b). These models have achieved accurate segmentation performance on unseen data or tasks. However, they are inherently constrained by their small parameter scale and the computational cost of in-context learning, both of which limit their generalizability.

In addition, building on the Segment Anything Model (SAM), variants including MedSAM (Ma et al. 2024a), SAM-Med (Ye et al. 2023), MedicoSAM (Archit, Freckmann, and Pape 2025) have been proposed by fine-tuning on large-scale medical imaging datasets with more parameters, showing promising performance in various medical tasks. Compared to ICL-based models, SAM-based approaches benefit from larger model capacity and broader

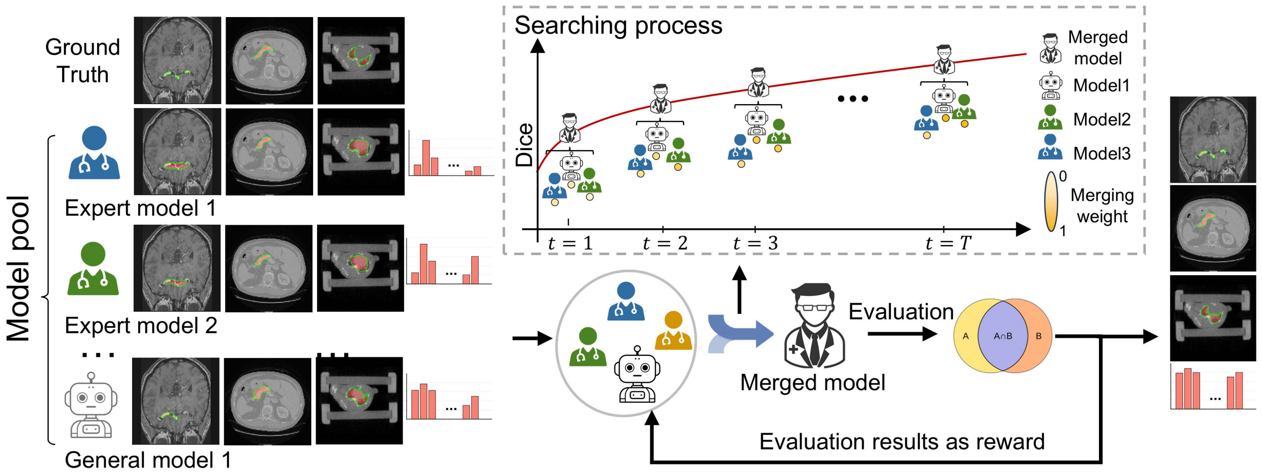


Figure 2: Overview of our model merging framework. Given a pool of models, MedSAMix searches for optimal merging configurations by using single-task or multi-task performance as rewards. While individual models may exhibit varying behaviors across tasks, MedSAMix adaptively combines them to optimize performance for the target task.

training coverage, which contribute to their superior generalization. However, the success of such foundation models often comes at the cost of large data requirements and substantial training overhead. In contrast, our proposed MedSAMix is an efficient and training-free approach that operates at the model level, reducing dependence on large-scale datasets while preserving strong generalization across diverse segmentation tasks.

2. Model Merging

Early model merging relied on direct weight averaging (Utans 1996; White 2016), which lacked fine-grained control over model behaviors. Subsequent methods introduced parameter-space transformations, either through adjustment matrices (Matena and Raffel 2022; Jin et al. 2022) or task vectors defined by fine-tuning deltas (Ilharco et al. 2022), enabling algebraic composition of capabilities across tasks. To address parameter conflicts, sparsity-driven strategies like TIES-Merging (Yadav et al. 2024) and DARE (Yu et al. 2024) selectively retain and rescale parameters based on magnitude. More recent approaches explored parameter-level hyperparameter control (Yang et al. 2023; Du et al. 2024), though still limited to task-vector formulations, while evolutionary merging (Akiba et al. 2024) suffers from high search complexity. In this study, our framework integrates multiple merging approaches with fine-grained hyperparameter control tailored for image segmentation. Unlike prior medical imaging approaches (Qazi et al. 2024; Sanjeev et al. 2024), which merge networks during training to produce a robust model, our method adopts a fundamentally different paradigm: MedSAMix enables efficient post-hoc merging across different SAM fine-tuned models.

Methodology

In this section, we present a comprehensive overview of our approach. We start by formalizing the problem and introducing an optimization framework built upon three core components: (1) Search Space: We design a layer-wise search space specifically for SAM-series models based on Vision Transformers (ViT). This space allows for fine-grained control over merging strategies by supporting multiple merging methods at varying layer granularities across key modules, including the image encoder, prompt encoder, and mask decoder. (2) Optimization Objectives: The framework accommodates both task-specific objectives, enabling targeted enhancements for specialized domains, and multi-objective formulations that aim to identify Pareto-optimal configurations across diverse segmentation tasks. (3) Search Algorithm: We adopt SMAC optimization algorithm (Lindauer et al. 2022), which leverages the defined objectives as rewards to steer the search toward effective merge configurations. The following subsections elaborate on each component and explain how they collectively contribute to robust model merging for segmentation applications.

Problem Formulation

Given a pre-trained segmentation base model M_{base} and a set of candidate segmentation models $\mathcal{M} = \{M_1, M_2, ..., M_K\}$ fine-tuned from the same base architecture, our goal is to construct an optimal merged model that maximizes performance across a single dataset or multiple datasets.

These combinations are determined by a set of hyperparameters $\omega \in \Omega$, where Ω represents the search space of all possible merging configurations. Each configuration ω defines a specific strategy to combine components from candidate models to form a merged model M_ω. The performance of the merged model is evaluated using an objective function $f(M_\omega)$ that measures effectiveness across segmentation tasks. This leads to our optimization problem:

$$\omega^* = \arg\min_{\omega \in \Omega} f(M_\omega) \tag{1}$$

where ω^* represents the optimal hyperparameter configuration that yields the best-performing merged model according to the chosen objective.

Search Space

Our candidate models are all based on the Segment Anything Model (SAM) architecture (Kirillov et al. 2023). The network is built on the transformer architecture, specifically incorporating a Vision Transformer-based (Dosovitskiy et al. 2020) image encoder with l transformer layers responsible for extracting image features, a prompt encoder with k convolutional downsampling layers for integrating user interactions, and a lightweight mask decoder with z transformer layers and p transposed convolutional layers that generate segmentation results and confidence scores. The search space Ω encompasses all possible configurations for constructing our merged model. For different components (image encoder, prompt encoder, and mask decoder), we employ varying layer granularity to determine how layers are grouped for merging, where each group shares the same merge hyperparameters. This granularity-based approach allows us to balance the size of the search space with fine-grained control over the merging process. Specifically, we define granularity g_{enc}, g_{prompt}, and g_{dec} for the three components respectively, dividing the layers into $G = \lceil l/g_{\text{enc}} \rceil + \lceil k/g_{\text{prompt}} \rceil + \lceil (z+p)/g_{\text{dec}} \rceil$ groups in total. For each layer group i, we specify a merge method $d_i \in \{1, 2, \ldots, D\}$ selected from D available merging techniques, and associated hyperparameters $\mathbf{h}_i = [h_{i,1}, h_{i,2}, \ldots, h_{i,P_i}]$, where P_i is the number of hyperparameters for merge method d_i. Therefore, a complete configuration $\omega \in \Omega$ is represented as:

$$\omega = \{\{(d_1, \mathbf{h}_1), (d_2, \mathbf{h}_2), \ldots, (d_G, \mathbf{h}_G)\}\} \tag{2}$$

This formulation provides a structured framework for exploring the space of possible model merging configurations while maintaining computational tractability through hierarchical grouping.

Optimization Objective

To evaluate the quality of the merged model, we define both single-objective and multi-objective optimization that measures the model's effectiveness across tasks. Specifically, we measure performance on calibration datasets $\mathcal{D}$, quantifying metrics such as Dice coefficient for segmentation.

For single-task optimization, we focus on maximizing performance on a specific calibration target task T:

$$f_{\text{single}}(M_\omega) = \mathcal{L}(M_\omega, \mathcal{D}_T) \tag{3}$$

where $\mathcal{L}$ represents the segmentation loss function (e.g., Dice loss, cross-entropy) evaluated on the target dataset $\mathcal{D}_T$.

For multi-task optimization across tasks $\mathcal{T} = \{T_1, T_2, \ldots, T_m\}$, we employ Pareto Efficient Global Optimization (ParEGO) (Knowles 2006) to identify Pareto-optimal solutions:

$$\begin{aligned} f_{\text{multi}}(M_\omega, \lambda) = \max_{i=1,\ldots,m} \{\lambda_i \cdot f_{\text{single},i}(M_\omega)\} \\ + \alpha \sum_{i=1}^{m} \lambda_i \cdot f_{\text{single},i}(M_\omega) \end{aligned} \tag{4}$$

where $f_{\text{single},i}(M_\omega)$ is the i-th objective function, λ_i is the corresponding weight satisfying $\sum_{i=1}^{m} \lambda_i = 1$ and $\lambda_i \geq 0$, and α is a small positive constant (typically 0.05). Each task-specific objective is defined as:

$$f_{\text{single},i}(M_\omega) = \mathcal{L}_i(M_\omega, \mathcal{D}_{T_i}) \tag{5}$$

The optimizer outputs a Pareto front of merging configurations representing different trade-offs between tasks. For evaluation, we selected the configurations yielding the best Pareto front on the calibration set.

Search Algorithm

To efficiently navigate the large search space Ω and find optimal merging configurations, we employ Bayesian optimization based on SMAC (Lindauer et al. 2022) with Random Forest (Breiman 2001) as the surrogate model. Given evaluated configurations $\mathcal{H}_t = \{(\omega_1, f(M_{\omega_1})), \ldots, (\omega_t, f(M_{\omega_t}))\}$ at iteration t, the next configuration is selected by:

$$\omega_{t+1} = \arg\max_{\omega \in \Omega} \text{EI}(\omega) \tag{6}$$

where $\text{EI}(\omega) = \mathbb{E}[\max(f^* - \hat{f}(\omega), 0)]$ is the Expected Improvement acquisition function, f^* is the best observed value, and $\hat{f}(\omega)$ is the Random Forest prediction. The process iteratively selects promising configurations until convergence. The whole process is described in Alg. 1

Algorithm 1: Optimization Process of MedSAMix

```
Require: Base model $M_{\text{base}}$, candidate models $\mathcal{M}$, calibration datasets $\mathcal{D}$, maximum iterations $T_{\max}$
Ensure: Optimal merging configuration $\omega^*$
Initialize search space with $g_{\text{enc}}, g_{\text{prompt}}, g_{\text{dec}}$
Initialize evaluation history $\mathcal{H}_0 = \emptyset$
Random initial configurations and evaluate to get $\mathcal{H}_1$
for $t = 1$ to $T_{\max}$ do
    Train Random Forest surrogate model on $\mathcal{H}_t$
    Compute Expected Improvement: $\text{EI}(\omega) = \mathbb{E}[\max(f^* - \hat{f}(\omega), 0)]$
    Select next configuration: $\omega_{t+1} = \arg\max_{\omega \in \Omega} \text{EI}(\omega)$
    // Model Merging based on Configuration $\omega_{t+1}$
    for each layer group $i = 1$ to $G$ do
        Apply merge method $d_i$ with hyperparameters $\mathbf{h}_i$
        Merge layers from candidate models
    end for
    // Evaluation
    if single-task optimization then
        Compute $f_{\text{single}}(M_{\omega_{t+1}}) = \mathcal{L}(M_{\omega_{t+1}}, \mathcal{D}_T)$
    else
        Compute $f_{\text{multi}}(M_{\omega_{t+1}}, \lambda)$ using ParEGO on tasks
    end if
    Update history: $\mathcal{H}_{t+1} = \mathcal{H}_t \cup \{(\omega_{t+1}, f(M_{\omega_{t+1}}))\}$
    Update best configuration: $f^* = \min_{\omega \in \mathcal{H}_{t+1}} f(M_\omega)$
end for
return $\omega^* = \arg\min_{\omega \in \mathcal{H}_{T_{\max}}} f(M_\omega)$
```

idx	Task Name	Supervised Upper-bound	In-context-learning models			SAM-based					
		nnU-Net	SegGPT	UniverSeg	Neuro3D	SAM-Med	Medico SAM	MedSAM	SAM	Ours-S	Ours-M
1	Brain Tumor	75.83	17.24	19.20	68.05	67.26	73.17	70.56	70.04	**78.36**	75.18
2	Vascular	85.87	28.33	68.21	**80.27**	36.85	31.35	28.80	50.86	60.70	62.90
3	Cerebral Cortex	87.72	47.08	69.51	**87.08**	31.77	56.13	48.83	55.00	58.00	55.03
4	Hippocampus	82.30	25.52	71.33	**75.92**	36.17	52.17	43.36	53.23	62.14	57.51
5	Thalamus	83.40	38.70	74.57	**78.09**	56.49	67.60	56.99	44.08	70.98	67.60
6	Lateral Ventricle	84.76	45.60	75.23	**82.44**	53.83	55.74	27.52	74.68	75.71	77.60
7	Putamen	83.03	21.92	73.38	**78.41**	25.74	36.93	26.26	32.49	45.33	33.61
8	Amygdala	73.43	7.77	59.37	57.85	30.33	48.62	39.58	29.18	**61.92**	52.03
9	FLARE22 Liver	96.59	68.72	79.33	90.73	92.03	91.83	85.89	84.60	**94.05**	92.88
10	FLARE22 Kidney_R	94.66	63.24	84.21	76.34	93.95	90.04	84.95	95.05	**95.86**	95.12
11	FLARE22 Kidney_L	95.39	66.75	83.33	79.25	92.92	89.64	83.87	94.25	**95.06**	94.16
12	Maxillary Sinus	91.23	51.30	80.57	60.29	60.90	75.54	56.40	80.45	**84.47**	79.61
13	Nasal Cavity	88.37	39.83	**69.76**	60.36	54.23	60.15	40.42	58.70	62.51	65.29
14	Nasal Pharynx	91.40	46.35	84.87	83.61	87.44	84.68	70.52	85.62	**89.01**	88.83
15	Prostate	86.78	50.26	76.36	33.57	88.40	89.16	79.01	88.21	**91.36**	90.64
16	Mice-Lung	89.28	62.17	75.67	**83.25**	61.08	65.93	49.04	71.96	80.62	75.43
17	Mice-Pancreas	83.27	53.52	65.21	66.03	85.35	83.68	74.17	85.20	**89.20**	88.44
18	Cardiac	85.00	43.23	68.96	64.21	77.06	74.31	66.18	74.87	**84.41**	83.03
19	FLARE22 Spleen	95.14	64.57	70.59	82.61	93.46	92.93	83.86	92.60	**95.06**	94.76
20	FLARE22 Pancreas	79.92	2.36	36.38	16.89	71.66	76.59	58.14	70.26	77.66	**78.36**
21	RAOS Liver	93.74	70.81	81.49	81.83	88.19	89.90	84.64	85.50	**92.38**	89.82
22	RAOS Kidney	92.39	63.14	**80.14**	39.74	62.34	72.44	66.01	69.13	74.66	72.72
23	RAOS Stomach	87.73	33.03	57.92	33.50	79.85	85.67	77.64	79.23	**88.74**	86.00
24	Optic Cup	85.68	74.32	80.56	-	71.61	86.42	86.41	61.93	**87.70**	82.11
25	Optic Disk	96.67	**96.27**	95.61	-	85.53	93.70	94.01	83.17	95.17	94.73
	Avg.	87.58	47.28	71.27	67.84	67.38	72.97	63.32	70.81	79.64 (25)	77.34 (18)

Table 1: Segmentation performance across 25 tasks in terms of Dice coefficient score (%). The values in parentheses in the last row indicate the number of tasks in which each method outperforms the baseline MedicoSAM/MedSAM/SAM model.

**Ours-S:* MedSAMix-S. The model is merged under a specific single-task setting and evaluated on the testing set.
**Ours-M:* MedSMix-M. The model is merged under a multi-task setting and evaluated on the test data across all 25 tasks.

Experiments

Datasets. In this study, we incorporate 25 publicly available medical image segmentation datasets across a wide range of organs (e.g., liver, kidney, spleen) and imaging modalities (e.g., CT, MRI, MRA). These datasets cover diverse tasks, including brain tumor segmentation (BraTS) (Menze et al. 2014), vascular segmentation (Topcow) (Yang et al. 2024), optic disc and cup segmentation from retinal fundus images (Fundus) (Staal et al. 2004), and abdominal organ segmentation for the kidneys, liver, spleen, pancreas (FLARE22) (Ma et al. 2024b), liver, kidney, and stomach (RAOS) (Luo et al. 2024). We also include maxillary sinus, nasal cavity segmentation, and nasal pharynx (Nasal) (Zhang et al. 2024), cardiac structure segmentation (MSD) (Antonelli et al. 2022), and prostate segmentation (PROMISE) (Litjens et al. 2014). For a broad and fair comparison as well as to enhance the diversity, we include additional commonly studied segmentation tasks, such as body organ segmentation in rodents (e.g., lungs and intestines) using the Mice dataset (Rosenhain et al. 2018), as well as various neuroanatomical structure segmentation from the ADNI dataset (Jack Jr et al. 2008), including the cerebral cortex, hippocampus, thalamus, amygdala, lateral ventricles, and putamen. More details of the data size are reported in the Supplementary.

Benchmarks. We compare our method with ICL-based baselines, including UniverSeg (Butoi et al. 2023), SegGPT (Wang et al. 2023), and Neuroverse3D (Neuro3D) (Hu et al. 2025b), as well as SAM-based models such as SAM-Med2D (Ye et al. 2023), SAM (Kirillov et al. 2023), MedSAM (Ma et al. 2024a), and MedicoSAM (Archit, Freckmann, and Pape 2025). In addition, we use the fully-supervised nnU-Net (Isensee et al. 2021) as an upper bound for reference. Additional comparisons are supplied in the Supplementary.

Evaluations. For each dataset, 80% is reserved for testing to ensure robust evaluation. The remaining 20% serves multiple purposes: as a calibration set for MedSAMix during the merging search, as context prompts for ICL-based models, and as training data for nnU-Net. SAM-based models are evaluated in a zero-shot setting without access to training data but are provided with inferred bounding boxes for guidance. All tasks are evaluated using the Dice coefficient.

Model Selection. We use SAM (Kirillov et al. 2023) as the base and merge it with MedSAM (Ma et al. 2024a) and MedicoSAM (Archit, Freckmann, and Pape 2025) as representative fine-tuned variants. Although other variants exist, such as the Medical SAM Adapter (Wu et al. 2025), they differ architecturally from SAM, making them incompatible with our model merging framework.

Objective and Optimizer. For expert capability evaluation, we focus on single-task search. For each task, the model is optimized using the calibration set to obtain the optimal merging configuration, which is then evaluated on the corresponding test set. This procedure is repeated for all 25 tasks. For general capability evaluation, we perform multi-objective optimization across a representative subset of tasks to reflect overall performance. To ensure both diversity and representativeness, we select eight tasks for the search process: optic disk, tumor, vascular, lateral ventricle, mice-lung, cardiac, nasal pharynx, and prostate. The merged models obtained from single-task and multi-task settings are referred to as **MedSAMix-S** and **MedSAMix-M**, respectively.

Implementation details. In this study, we leverage the commonly used base architectures of SAM and its variants, MedSAM, and MedicoSAM for model merging, with $l = 12$ Transformer layers, 768 hidden features, 12 heads, $k = 4$ prompt encoder layers, and a lightweight mask decoder with $z = 2$ Transformer layers and $p = 3$ transposed convolutional layers. Our model merging framework adaptively searches for optimal layer combinations, with candidate merging techniques including TIES, task arithmetic, linear combination, and SLERP. For single-task, we perform 120 trials using two GPUs, while for multi-task optimization, we conduct 200 trials on four GPUs. The layer granularity is searched within the range [1, 4].

Results

1. Evaluations on domain-specific single tasks

We first evaluated the performance in domain-specific single-task applications. The results are shown in Table 1, where our merged model is denoted as MedSAMix-S. From the results, we can see that MedSAMix-S achieves significant improvements over the original SAM and other fine-tuned models across all the tasks. For instance, on the brain tumor and Nasal Pharynx parcellation tasks, MedSAMix-S outperforms the MedicoSAM by 5.19% and 4.33% in terms of Dice coefficient, respectively. Moreover, for tasks where MedSAM and MedicoSAM already perform well, such as optic disk segmentation, MedSAMix-S still yields further improvements of around 1.2%. This demonstrates the advantage of model merging in leveraging the complementary strengths of different models, even when those models are already well-adapted to the task. In addition, we report the fully supervised results from nnU-Net (Isensee et al. 2021) as an upper bound for reference. Compared to nnU-Net, although there remains a performance gap of around 8%, our model achieves comparable or even better results on tasks such as brain tumor and prostate segmentation. This underscores the strong potential of foundation models, highlighting their promising few-shot capabilities.

2. Evaluations on universal segmentation tasks

In addition to the single-task evaluations, we also compare merged MedSAMix-M with other ICL-based and SAM-based models across multiple tasks to assess its generalization capability in medical image segmentation. The results are reported as MedSAMix-M in Table 1, where it outperforms baselines on 18 out of 25 tasks. Among all universal segmentation models, MedSAMix-M achieves the best overall performance, with an average improvement of 4.37% over the second-best model. Furthermore, most ICL-based models perform even worse than the original SAM. As shown in Fig. 3, visual comparisons further illustrate this trend. For example, while SegGPT performs well on certain tasks, its average performance across all tasks remains relatively low, suggesting that these domain-specific foundation models still exhibit limited generalization. These findings underscore the critical importance of generalization capability in medical image segmentation.

Type	Models	Avg.	Improv.
Baseline	MedSAM	63.32	-
	SAM	70.81	-
	MedicoSAM	72.97	-
Merged	TIES	66.31	-6.66 %
	TA	72.19	-0.78 %
	Linear	74.00	1.03 %
	SLERP	75.28	2.31 %
	MedSAMix-S	79.64	6.67 %
	MedSAMix-M	77.34	4.37 %

Table 2: Comparisons with other model merging methods on medical image segmentation in terms of the averaged Dice coefficient (%) score across 25 tasks.

3. Comparisons with other model merging methods

Given the effectiveness of model merging in medical image segmentation, we compare our MedSAMix with several model merging baselines, including Task-Arithmetic (TA) (Ilharco et al. 2022), TIES (Yadav et al. 2024), SLERP (White 2016), and linear weighted combination. which we used in our search space. The results are summarized in Table 2, which reports the average Dice coefficient across all 25 tasks, along with the corresponding improvements over the MedicoSAM baseline. Detailed results are provided in the Supplementary Material. The results show that most model merging baselines offer limited improvement over MedicoSAM. In contrast, MedSAMix(-S/-M) consistently outperforms them. These findings highlight the effectiveness and adaptability of MedSAMix in identifying optimal merging configurations for medical image segmentation.

4. Search space sensitivity analysis

Layer granularity. We conducted a sensitivity analysis on layer granularity for both single-task and multi-task optimization, varying the granularity from 1 to 4. The results are presented in Fig. 4 (1), where the performance under single-task and multi-task settings is shown in orange and blue, respectively. The results reveal that single-task performance is relatively insensitive to changes in layer granularity, whereas multi-task performance is notably affected. This suggests that for single-task scenarios, there remains redundancy in the network parameters, allowing for merging even

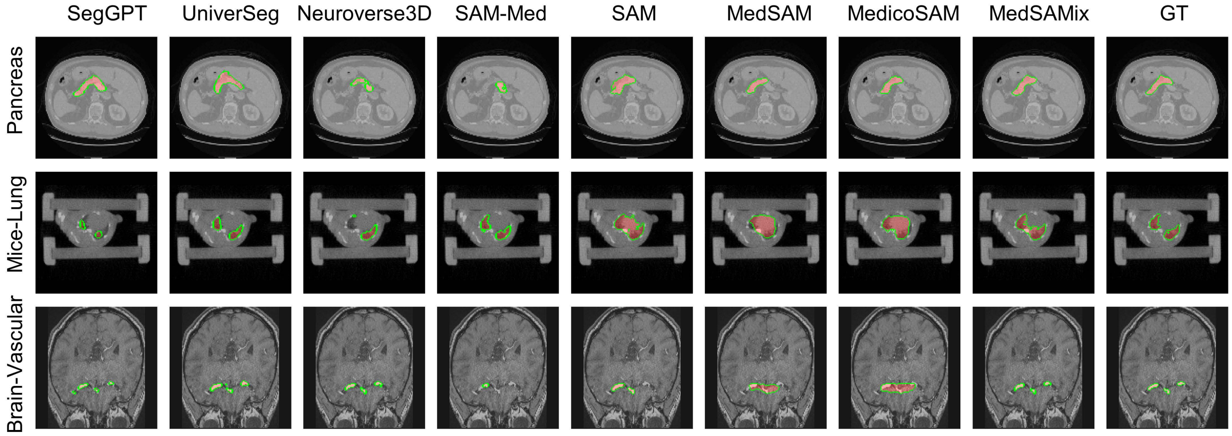
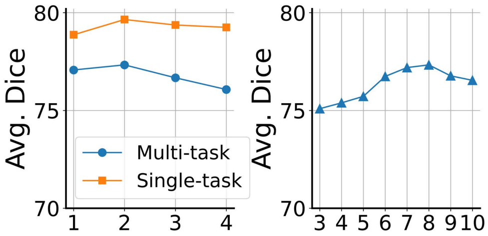
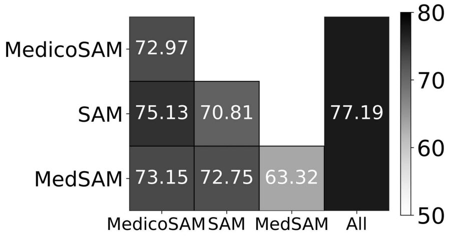


Figure 3: Visual examples of segmentation results of implemented models on pancreas, mouse lung, and brain vascular datasets.

with coarse granularity. In contrast, multi-task performance is more dependent on fine-grained control over layer merging. Notably, the best performance in the multi-task setting is achieved when the granularity is set to 2.



(1) Layer granularity (2) Task number (multi-task)



(3) Different combinations

Figure 4: Sensitivity analysis of different merging settings in terms of (1) layer granularity, (2) task for multi-task optimization, and (3) different model combinations.

Number of searched tasks. In addition, we investigated the impact of the number of tasks used for multi-task optimization. As shown in Fig. 4 (2), we find a notable improvement when the number of search tasks exceeds six. The best performance is achieved when eight tasks are used. However, further increasing the number of tasks leads to a decline in performance. This may be attributed to the increasing difficulty in maintaining Pareto-optimal solutions and effective model selection as the task number grows.

Combinations of models. Finally, we compare model merging on different combinations of models, as shown in Fig. 4 (3). We find that merging MedicoSAM and SAM yields the best pairwise performance, consistent with their strong individual results. Although MedSAM achieves the worst performance among the three models, combining all three models can lead to further improvements, demonstrating the benefit of leveraging complementary knowledge from multiple sources. This demonstrates the great potential of our framework. Incorporating more diverse models is expected to further enhance performance.

Discussion and Conclusion

In this study, we propose MedSAMix, a training-free model merging method for medical image segmentation. Our work represents a pioneering exploration of model merging in medical image analysis. While domain-specific foundation models achieve strong performance on specialized tasks, they struggle with generalizability and, in some cases, underperform the original SAM model. Our MedSAMix surpasses all baselines across 25 tasks, showing both stronger domain-specific performance and better generalization.

This study highlights the inherent trade-off between specialization and generalization in medical image segmentation. While expert models like MedicoSAM perform well across tasks, they still underperform the base model in certain cases. In contrast, the original SAM demonstrates strong generalizability, underscoring the value of foundation models. Based on this insight, our MedSAMix effectively balances this trade-off and enhances both multi-task generalization and single-task performance without the need for additional data or retraining. This highlights the power of model merging in mitigating suboptimal generalization and points to the potential of hybrid strategies that achieve performance gains purely through model-level optimization. Further details are provided in the extended version (Yang et al. 2025).

Limitation. We acknowledge that, despite covering 25 diverse tasks, our study cannot fully represent the entire landscape of medical image segmentation. Future work will expand task diversity and develop systematic methods to assess generalization and robustness.

Acknowledgments

Y.Y., F.S., and T.W. acknowledge support from the German Research Foundation (DFG) Emmy Noether Program (513851350, T.W.) and the BMBF/DLR project FEDORA (01EQ2403G, T.W.). They also acknowledge the computational resources provided by the de.NBI Cloud, part of the German Network for Bioinformatics Infrastructure (de.NBI). J.H. acknowledges support from the Major Key Project of PCL (PCL2025A09). Y.Y., G.S., and J.G. acknowledge their affiliation with the International Max Planck Research School for Intelligent Systems (IMPRS-IS). G.S. and J.G. further acknowledge support from the Hector II Foundation.

References

Akiba, T.; Shing, M.; Tang, Y.; Sun, Q.; and Ha, D. 2024. Evolutionary optimization of model merging recipes. *arXiv preprint arXiv:2403.13187*.

Aleixo, E. L.; Colonna, J. G.; Cristo, M.; and Fernandes, E. 2023. Catastrophic forgetting in deep learning: A comprehensive taxonomy. *arXiv preprint arXiv:2312.10549*.

Almakky, I.; Sanjeev, S.; Hashmi, A. U. R.; Qazi, M. A.; Wang, H.; and Yaqub, M. 2024. MedMerge: merging models for effective transfer learning to medical imaging tasks. *arXiv preprint arXiv:2403.11646*.

Antonelli, M.; Reinke, A.; Bakas, S.; Farahani, K.; Kopp-Schneider, A.; Landman, B. A.; Litjens, G.; Menze, B.; Ronneberger, O.; Summers, R. M.; et al. 2022. The medical segmentation decathlon. *Nature communications*, 13(1): 4128.

Archit, A.; Freckmann, L.; and Pape, C. 2025. MedicoSAM: Towards foundation models for medical image segmentation. *arXiv preprint arXiv:2501.11734*.

Breiman, L. 2001. Random forests. *Machine learning*, 45: 5–32.

Butoi, V. I.; Ortiz, J. J. G.; Ma, T.; Sabuncu, M. R.; Guttag, J.; and Dalca, A. V. 2023. Universeg: Universal medical image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 21438–21451.

Czolbe, S.; and Dalca, A. V. 2023. Neuralizer: General neuroimage analysis without re-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6217–6230.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Du, G.; Lee, J.; Li, J.; Jiang, R.; Guo, Y.; Yu, S.; Liu, H.; Goh, S. K.; Tang, H.-K.; He, D.; et al. 2024. Parameter competition balancing for model merging. *Advances in Neural Information Processing Systems*, 37: 84746–84776.

Gao, Y.; Liu, D.; Li, Z.; Li, Y.; Chen, D.; Zhou, M.; and Metaxas, D. N. 2025. Show and Segment: Universal Medical Image Segmentation via In-Context Learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 20830–20840.

Hu, J.; Cao, J.; Yang, Y.; Ye, C.; Zhang, Y.; Peng, H.; and Ma, T. 2025a. Medverse: A universal model for full-resolution 3d medical image segmentation, transformation and enhancement. *arXiv preprint arXiv:2509.09232*.

Hu, J.; Peng, H.; Yang, Y.; Guo, X.; Shang, Y.; Shi, P.; Ye, C.; and Ma, T. 2025b. Building 3D In-Context Learning Universal Model in Neuroimaging. *arXiv preprint arXiv:2503.02410*.

Hu, J.; Shang, Y.; Yang, Y.; Guo, X.; Peng, H.; and Ma, T. 2024. Icl-sam: Synergizing in-context learning model and sam in medical image segmentation. *Medical Imaging with Deep Learning*, 641–656.

Ilharco, G.; Ribeiro, M. T.; Wortsman, M.; Gururangan, S.; Schmidt, L.; Hajishirzi, H.; and Farhadi, A. 2022. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089*.

Isensee, F.; Jaeger, P. F.; Kohl, S. A.; Petersen, J.; and Maier-Hein, K. H. 2021. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2): 203–211.

Jack Jr, C. R.; Bernstein, M. A.; Fox, N. C.; Thompson, P.; Alexander, G.; Harvey, D.; Borowski, B.; Britson, P. J.; L. Whitwell, J.; Ward, C.; et al. 2008. The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 27(4): 685–691.

Jin, X.; Ren, X.; Preotiuc-Pietro, D.; and Cheng, P. 2022. Dataless knowledge fusion by merging weights of language models. *arXiv preprint arXiv:2212.09849*.

Kemker, R.; McClure, M.; Abitino, A.; Hayes, T.; and Kanan, C. 2018. Measuring catastrophic forgetting in neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4015–4026.

Knowles, J. 2006. ParEGO: A hybrid algorithm with online landscape approximation for expensive multiobjective optimization problems. *IEEE transactions on evolutionary computation*, 10(1): 50–66.

Li, H.; Wang, Y.; Wan, R.; Wang, S.; Li, T.-Q.; and Kot, A. 2020. Domain generalization for medical imaging classification with linear-dependency regularization. *Advances in neural information processing systems*, 33: 3118–3129.

Lindauer, M.; Eggensperger, K.; Feurer, M.; Biedenkapp, A.; Deng, D.; Benjamins, C.; Ruhkopf, T.; Sass, R.; and Hutter, F. 2022. SMAC3: A versatile Bayesian optimization package for hyperparameter optimization. *Journal of Machine Learning Research*, 23(54): 1–9.

Litjens, G.; Toth, R.; Van De Ven, W.; Hoeks, C.; Kerkstra, S.; Van Ginneken, B.; Vincent, G.; Guillard, G.; Birbeck, N.; Zhang, J.; et al. 2014. Evaluation of prostate segmentation algorithms for MRI: the PROMISE12 challenge. *Medical image analysis*, 18(2): 359–373.

Luo, X.; Li, Z.; Zhang, S.; Liao, W.; and Wang, G. 2024. Rethinking Abdominal Organ Segmentation (RAOS) in the clinical scenario: A robustness evaluation benchmark with challenging cases. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 531–541. Springer.

Ma, J.; He, Y.; Li, F.; Han, L.; You, C.; and Wang, B. 2024a. Segment anything in medical images. *Nature Communications*, 15(1): 654.

Ma, J.; Zhang, Y.; Gu, S.; Ge, C.; Mae, S.; Young, A.; Zhu, C.; Yang, X.; Meng, K.; Huang, Z.; et al. 2024b. Unleashing the strengths of unlabelled data in deep learning-assisted pan-cancer abdominal organ quantification: the FLARE22 challenge. *The Lancet Digital Health*, 6(11): e815–e826.

Maron, R. C.; Hekler, A.; Haggenmüller, S.; von Kalle, C.; Utikal, J. S.; Müller, V.; Gaiser, M.; Meier, F.; Hobelsberger, S.; Gellrich, F. F.; et al. 2022. Model soups improve performance of dermoscopic skin cancer classifiers. *European Journal of Cancer*, 173: 307–316.

Matena, M. S.; and Raffel, C. A. 2022. Merging models with fisher-weighted averaging. *Advances in Neural Information Processing Systems*, 35: 17703–17716.

Menze, B. H.; Jakab, A.; Bauer, S.; Kalpathy-Cramer, J.; Farahani, K.; Kirby, J.; Burren, Y.; Porz, N.; Slotboom, J.; Wiest, R.; et al. 2014. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE transactions on medical imaging*, 34(10): 1993–2024.

Neyshabur, B.; Sedghi, H.; and Zhang, C. 2020. What is being transferred in transfer learning? *Advances in neural information processing systems*, 33: 512–523.

Qazi, M. A.; Almakky, I.; Hashmi, A. U. R.; Sanjeev, S.; and Yaqub, M. 2024. Dynammo: Dynamic model merging for efficient class incremental learning for medical images. In *Annual Conference on Medical Image Understanding and Analysis*, 245–257. Springer.

Rakic, M.; Wong, H. E.; Ortiz, J. J. G.; Cimini, B. A.; Guttag, J. V.; and Dalca, A. V. 2024. Tyche: Stochastic in-context learning for medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11159–11173.

Rosenhain, S.; Magnuska, Z. A.; Yamoah, G. G.; Rawashdeh, W.; Kiessling, F.; Gremse, F.; et al. 2018. A preclinical micro-computed tomography database including 3D whole body organ segmentations. *Scientific data*, 5(1): 1–9.

Sanjeev, S.; Zhaksylyk, N.; Almakky, I.; Hashmi, A. U. R.; Qazi, M. A.; and Yaqub, M. 2024. FissionFusion: fast geometric generation and hierarchical souping for medical image analysis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 131–141. Springer.

Staal, J.; Abràmoff, M. D.; Niemeijer, M.; Viergever, M. A.; and Van Ginneken, B. 2004. Ridge-based vessel segmentation in color images of the retina. *IEEE transactions on medical imaging*, 23(4): 501–509.

Su, G.; and Geiping, J. 2025. Fine, I'll Merge It Myself: A Multi-Fidelity Framework for Automated Model Merging. *arXiv preprint arXiv:2502.04030*.

Su, G.; Shen, L.; Yin, L.; Liu, S.; Yang, Y.; and Geiping, J. 2025. GPTailor: Large Language Model Pruning Through Layer Cutting and Stitching. *arXiv preprint arXiv:2506.20480*.

Utans, J. 1996. Weight averaging for neural networks and local resampling schemes. In *Proc. AAAI-96 Workshop on Integrating Multiple Learned Models. AAAI Press*, 133–138. Citeseer.

Wang, X.; Zhang, X.; Cao, Y.; Wang, W.; Shen, C.; and Huang, T. 2023. Seggpt: Towards segmenting everything in context. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1130–1140.

White, T. 2016. Sampling generative networks. *arXiv preprint arXiv:1609.04468*.

Wu, J.; Wang, Z.; Hong, M.; Ji, W.; Fu, H.; Xu, Y.; Xu, M.; and Jin, Y. 2025. Medical sam adapter: Adapting segment anything model for medical image segmentation. *Medical image analysis*, 102: 103547.

Yadav, P.; Tam, D.; Choshen, L.; Raffel, C. A.; and Bansal, M. 2024. Ties-merging: Resolving interference when merging models. *Advances in Neural Information Processing Systems*, 36.

Yang, E.; Wang, Z.; Shen, L.; Liu, S.; Guo, G.; Wang, X.; and Tao, D. 2023. Adamerging: Adaptive model merging for multi-task learning. *arXiv preprint arXiv:2310.02575*.

Yang, K.; Musio, F.; Ma, Y.; Juchler, N.; Paetzold, J. C.; Al-Maskari, R.; Höher, L.; Li, H. B.; Hamamci, I. E.; Sekuboyina, A.; et al. 2024. Benchmarking the cow with the topcow challenge: Topology-aware anatomical segmentation of the circle of willis for cta and mra. *ArXiv*, arXiv–2312.

Yang, Y.; Su, G.; Hu, J.; Sammarco, F.; Geiping, J.; and Wolfers, T. 2025. Medsamix: A training-free model merging approach for medical image segmentation. *arXiv preprint arXiv:2508.11032*.

Ye, J.; Cheng, J.; Chen, J.; Deng, Z.; Li, T.; Wang, H.; Su, Y.; Huang, Z.; Chen, J.; Jiang, L.; et al. 2023. Sa-med2d-20m dataset: Segment anything in 2d medical imaging with 20 million masks. *arXiv preprint arXiv:2311.11969*.

Yu, L.; Yu, B.; Yu, H.; Huang, F.; and Li, Y. 2024. Language models are super mario: Absorbing abilities from homologous models as a free lunch. In *Forty-first International Conference on Machine Learning*.

Zhang, Y.; Wang, J.; Pan, T.; Jiang, Q.; Ge, J.; Guo, X.; Jiang, C.; Lu, J.; Zhang, J.; Liu, X.; et al. 2024. Nasalseg: A dataset for automatic segmentation of nasal cavity and paranasal sinuses from 3d ct images. *Scientific Data*, 11(1): 1329.

Zhao, J.; Yang, F.; Li, X.; Jiao, Z.; Zhai, Q.; Li, X.; Wu, D.; Fu, H.; and Cheng, H. 2025. SegMIC: A Universal Model for Medical Image Segmentation Through In-Context Learning. *Pattern Recognition*, 112179.