HIERARCHICAL GATING MIXTURE-OF-EXPERTS FOR MIXED-DOMAIN MEDICAL IMAGE SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning a unified model that effectively captures domain-specific segmentation patterns remains challenging in mixed-domain medical image segmentation (MD-MIS), where images collected from diverse domains exhibit substantial heterogeneity despite sharing consistent anatomical structures. While Mixture-of-Experts (MoE) holds strong potential for mixed-domain medical image segmentation through expert specialization, existing MoEs suffer from limited feature expressiveness and unstable expert allocation, leading to inadequate domain-discriminative characterization of implicit task heterogeneity and intensified expert competition or collapse. To address the aforementioned challenges, we propose Hierarchical Gating Mixture-of-Experts (HG-MoE), introducing a hierarchical gating framework comprising two complementary mechanisms for effective expert specialization. Firstly, Micro-level Feature Gating is proposed to enhance feature expressiveness through an input-adaptive non-linear gating modulation, thereby enhancing domain-discriminative characterization of implicit task heterogeneity. Secondly, Macro-level Expert Gating is introduced to promote stable task-specific expert allocation through Top-K routing guided by domain-aware expert load distributions, thereby mitigating expert competition and collapse. Extensive experiments on three public datasets across diverse domain scenarios demonstrate the consistent superiority and versatility of HG-MoE. HG-MoE improves Dice by up to 7.30% under the mixed-domain setting, while achieving improvements of 3.04% in Dice and 30.81% in region-level F1 under the single-domain setting, respectively. Source code is publicly available at `https://***/HG-MoE`.

1 INTRODUCTION

Mixed-domain medical image segmentation (MD-MIS) has become increasingly prevalent in real-world clinical applications (Ulrich et al., 2023; Ma et al., 2025). Medical images are often aggregated from multiple institutions, imaging devices and acquisition protocols, naturally forming a mixed-domain segmentation scenario (Li et al., 2026a;b). However, substantial heterogeneity across domains leads to distinct feature distributions and domain-specific segmentation patterns, endowing MD-MIS with inherent multi-task characteristics (Wang et al., 2020; Bian et al., 2020; Ouyang et al., 2022). Consequently, developing a unified segmentation framework capable of effectively modeling heterogeneous mixed-domain data remains an important yet challenging problem in medical image analysis.

However, existing methods struggle to capture domain-specific segmentation patterns without explicit domain annotations in heterogeneous segmentation scenarios (Guan & Liu, 2021; Yoon et al., 2024). Existing approaches for MD-MIS generally follow two paradigms: shared-representation learning and domain-specific representation learning. Shared-representation methods learn a unified feature space across all samples, eliminating the need for domain annotations but potentially overlooking domain-specific segmentation patterns and suffering from interference among heterogeneous domains (Isensee et al., 2021; Ma et al., 2024a). In contrast, domain-specific representation learning relies on explicit domain annotations to separate heterogeneous domains and learn specialized representations, limiting their applicability to realistic mixed-domain scenarios where sample-level domain identities are unavailable (Wang et al., 2020; Ouyang et al., 2022). Consequently,

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.

broader applications of existing methods remains limited by the lack of a unified mechanism for characterizing and exploiting domain-specific segmentation patterns across heterogeneous inputs.

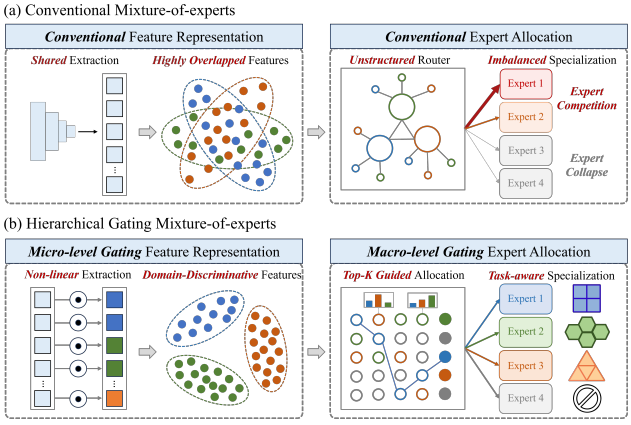


Figure 1: (a) Existing MoEs suffer from insufficiently discriminative feature representations and unstructured expert allocation. (b) HG-MoE addresses these challenges via domain-discriminative feature representations and Top-K selection guided expert allocation.

Mixture-of-Experts (MoE) has demonstrated the potential for mixed-domain medical image segmentation. By enabling individual experts to specialize in distinct segmentation patterns and feature characteristics, MoE allows individual experts to focus on different segmentation patterns and feature characteristics, thereby enhancing modeling flexibility for heterogeneous mixed-domain data. Moreover, its adaptive expert selection mechanism allows different samples to dynamically exploit appropriate expert knowledge, facilitating more effective utilization of heterogeneous information (Jiang & Shen, 2024; Rezvani et al., 2025). Owing to its flexibility, MoE has achieved remarkable success in a variety of heterogeneous learning scenarios, including multi-task learning (Ma et al., 2018), multimodal learning (Mustafa et al., 2022), domain generalization (Li et al., 2022) and uncertainty modeling (Wang et al., 2025).

However, directly applying MoE to capture domain-specific segmentation patterns in MD-MIS remains challenging, as conventional MoE architectures lack effective mechanisms to identify heterogeneous characteristics and establish reliable expert specialization (as Fig. 1(a)). The key difficulty lies not only in how to allocate samples to experts, but also in whether the representations used for expert allocation can reliably reveal implicit domain-dependent segmentation patterns. Specifically, the two fundamental challenges are: **(1) Existing MoE frameworks lack sufficiently domain-discriminative representations to identify domain-specific segmentation patterns in MD-MIS.** Although MD-MIS exhibits inherent multi-task characteristics, these task differences are implicitly encoded within heterogeneous domain distributions and often manifest as highly overlapping feature representations. Consequently, conventional feature extraction mechanisms within existing MoE frameworks struggle to learn task-aware representations that can effectively characterize domain-specific segmentation patterns. **(2) Existing MoE frameworks suffer from expert competition and expert collapse when learning heterogeneous mixed-domain segmentation patterns in MD-MIS.** To establish expert specialization, existing MoE frameworks typically rely on unstructured routing mechanisms to distribute input representations across experts. However, such routing strategies encourage intensive competition among experts for routing assignments, causing a small subset of experts to dominate the routing process while others remain underutilized. Consequently, expert collapse frequently occurs, preventing stable expert specialization and limiting the capability of MoE to fully exploit heterogeneous data.

To address the aforementioned challenges, we propose Hierarchical Gating Mixture-of-Experts (HG-MoE) for learning specialized representations across heterogeneous domains (as Fig. 1(b)). Globally, HG-MoE formulates expert specialization as a hierarchical gating process, which progressively enhances task-aware representations and performs stable expert allocation through two complementary gating mechanisms. Specifically, **(1) The Micro-level Feature Gating enhances task-aware representations through non-linear feature modulation, emphasizing subtle domain-dependent characteristics embedded in heterogeneous mixed-domain data.** The proposed Micro-level Feature Gating constructs a domain-discriminative representation feature space by adaptively modulating non-linear feature transformations. By amplifying task-relevant characteristics while suppressing domain-irrelevant variations, the resulting representations better capture the implicit task differences embedded within heterogeneous domains, thereby providing a reliable foun-

dation for subsequent stable expert allocation for MD-MIS. (**2) The Macro-level Expert Gating performs balanced and stable expert allocation, promoting effective and efficient expert specialization while alleviating expert competition and expert collapse.** The proposed Macro-level Expert Gating adopts a Top-K guided lightweight expert allocation paradigm that establishes a clearer correspondence between samples and experts rather than unstructured routing mechanisms. By encouraging more balanced expert utilization, the proposed Macro-level Expert Gating mitigates expert competition and expert collapse, thereby achieving more stable expert specialization and more effective exploitation of heterogeneous mixed-domain knowledge for MD-MIS.

In summary, our main contributions are listed as follows:

• For the first time, we investigate mixed-domain medical image segmentation from the perspective of expert specialization, exploiting the implicit multi-task characteristics of heterogeneous domains to achieve effective domain-specific segmentation patterns modeling.

• We propose a Micro-level Feature Gating mechanism that enhances task-aware representations through non-linear feature modulation, enabling domain-discriminative characterization of implicit task differences in heterogeneous mixed-domain medical images.

• We propose a Macro-level Expert Gating mechanism that establishes stable expert allocation and specialization via Top-K routing, promoting effective expert utilization while mitigating expert competition and collapse.

• Extensive experiments on three public datasets demonstrate the effectiveness and versatility of proposed HG-MoE in both mixed-domain and single-domain medical image segmentation scenarios.

2 Related Works

2.1 Heterogeneous Medical Image Segmentation

Existing methods for heterogeneous medical image segmentation can be divided into two categories. The first category learns unified segmentation models with shared representations, such as U-Net (Ronneberger et al., 2015), V-Net (Milletari et al., 2016), nnU-Net (Isensee et al., 2021), and recent medical segmentation foundation models (Ma et al., 2024a). These methods use shared encoder-decoder architectures or unified promptable segmentation frameworks, achieving strong performance across different organs, modalities, and segmentation tasks (Antonelli et al., 2022; Campello et al., 2021). The second category explicitly considers domain discrepancy, mainly including domain adaptation and domain generalization methods. Domain adaptation methods usually reduce distribution gaps between domains through explicit alignment or domain-aware representation learning. Uncertainty-aware domain alignment estimates prediction uncertainty to guide anatomical structure alignment across domains (Bian et al., 2020), while DoFE embeds domain-oriented information into feature representations to improve fundus image segmentation on unseen datasets (Wang et al., 2020). Domain generalization methods further aim to learn domain-invariant or domain-robust representations for unseen domains. FedDG introduces episodic learning in continuous frequency space to improve federated domain generalization for medical image segmentation (Liu et al., 2021), while causality-inspired domain generalization uses causal augmentation to encourage models to learn more robust features from a single source domain (Ouyang et al., 2022).

Although these methods have achieved promising progress, they mainly aim to learn unified and aligned representations across heterogeneous data which fails to achieve optimal performance for each individual domain due to the weakened domain-specific segmentation patterns. To this end, the proposed HG-MoE enhances domain-discriminative representations via non-linear feature representation and performs domain-aware stable expert allocation, thereby achieving superior performance in MD-MIS applications.

2.2 MoEs for Medical Image Segmentation

Mixture-of-Experts (MoE) has been increasingly adopted in medical image segmentation thanks to its ability to learn diverse and complementary representations through specialized expert branches (Jacobs et al., 1991; Shazeer et al., 2017). The expert specialization in MoE allows the model to better capture anatomical structures with varying appearances, scales, shapes, and boundaries,

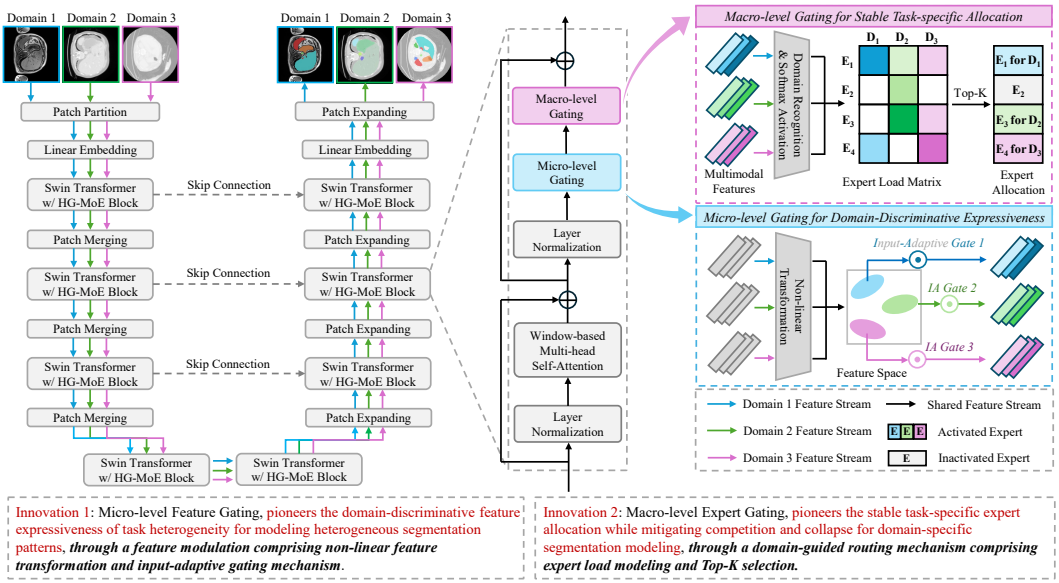


Figure 2: The proposed HG-MoE addresses the challenge of stable expert specialization by enhancing discriminative feature expressiveness *(Sec. 3.1)* and facilitating stable expert allocation *(Sec. 3.2)*, respectively, thereby demonstrating superior effectiveness in MD-MIS.

providing stronger feature expressiveness and adaptability in multiple application scenarios (Fedus et al., 2022; Riquelme et al., 2021). In multimodal medical image segmentation, M^{4}oE formulates different modalities as expert selection problems and improves cross-modal generalization through modality-specific experts and gated fusion (Jiang & Shen, 2024). MoME incorporates vision-language information into expert routing, where text embeddings guide multi-scale visual feature selection for semantically aware medical image segmentation (Rezvani et al., 2025). For medical foundation model adaptation, SegMoTE introduces token-level expert selection to alleviate negative transfer during fine-tuning on large-scale heterogeneous medical data (Lu et al., 2026). HoME combines MoE with state space models and uses hierarchical soft experts to capture long-range dependencies and structural heterogeneity in 3D medical images (Płotka et al., 2026). Patch-MoE Mamba further models different patch sequences with directional experts to enhance local structure preservation and spatial dependency modeling (Adame et al., 2026).

Although existing MoE-based methods can relatively learn representations for heterogeneity, they still struggle in MD-MIS because domain-dependent differences are implicit and difficult to separate from task-relevant anatomical patterns before expert allocation. Moreover, the inherent expert imbalance in existing MoE-based approaches can hinder stable expert specialization and limit the ability of MoE to exploit heterogeneous mixed-domain data. To address these limitations, our HG-MoE introduces a hierarchical gating strategy to learn non-linear discriminative representations and establish stable and balanced expert allocation.

3 METHODS

We propose Hierarchical Gating Mixture-of-Experts (HG-MoE), which provides a reliable solution for heterogeneous segmentation modeling, improving the robustness and reliability of unified segmentation under mixed-domain scenarios (as Fig. 2). HG-MoE provides a unified framework for modeling heterogeneous segmentation patterns in MD-MIS. Specifically, the Micro-level Feature Gating enhances domain-discriminative representations through non-linear feature modulation, enabling the model to characterize subtle task-dependent characteristics embedded in mixed-domain data (*Sec. 3.1*). Secondly, the Macro-level Expert Gating performs task-specific expert allocation to establish stable correspondence between heterogeneous tasks and experts, facilitating effective expert specialization while alleviating expert competition and collapse (*Sec. 3.2*).

3.1 Micro-level Feature Gating Enhances Feature Expressiveness for Heterogeneous Characterization

Feature expressiveness is essential for modeling task heterogeneity in MD-MIS, which provides stronger domain-discriminative capability to discriminate heterogeneous segmentation patterns across medical images from mixed-domains. The proposed Micro-level Feature Gating adaptively emphasizes latent task-relevant characteristics via non-linear feature modulation while suppressing irrelevant variations, thereby producing more domain-discriminative representations for heterogeneous task characterization.

To achieve this goal, the proposed Micro-level Feature Gating introduces an input-adaptive non-linear modulation mechanism to modulate feature representations. The proposed Micro-level Feature Gating learns a non-linear transformation to reshape the feature space and improve the separability of heterogeneous segmentation characteristics by exploiting the implicit task-dependent patterns embedded in the extracted representations. Therefore, the Micro-level Feature Gating mechanism dynamically modulates the transformed representations to reveal and enhance latent task-dependent characteristics, rather than relying on representations with limited capability to distinguish heterogeneous task patterns.

Specifically, given input medical image X, a transformer encoder is first employed to extract the initial feature representation H, which is further transformed into an enhanced representation Z through a non-linear transformation:

$$H = F_\theta(X) \tag{1}$$

$$Z = T(H) = W_2\delta(W_1H + b_1) + b_2 \tag{2}$$

where $F_\theta(\cdot)$ denotes the backbone transformer network and $T(\cdot)$ represents the non-linear transformation function.The non-linear transformation maps the original representation into a more expressive feature space by modeling non-linear dependencies among feature dimensions, allowing heterogeneous task-dependent characteristics to be better captured.

A gating function is further introduced to the transformed representation for input-adaptive modulation coefficients:

$$M = g(H) = \sigma(W_gH + b_g), \tag{3}$$

where M represents the dimension-wise modulation vector and $\sigma(\cdot)$ denotes the sigmoid function, which converts the learned responses into adaptive gating coefficients for element-wise feature modulation. Different from conventional feature weighting strategies that focus on selecting high-response features, the proposed gating mechanism estimates the input-dependent contributions of transformed feature dimensions, adaptively retaining task-relevant responses while attenuating task-irrelevant variations.

The final representation $\hat{H}$ is obtained by applying the learned modulation coefficients:

$$\hat{H} = M \odot Z, \tag{4}$$

where $\odot$ denotes element-wise multiplication. Through this non-linear feature modulation, the proposed modulation mechanism dynamically enhances task-relevant responses in the transformed feature space while suppressing irrelevant variations, increasing the separability of heterogeneous task-dependent segmentation patterns with domain-discriminative representations and providing a more reliable basis for the subsequent task-specific expert allocation.

Summarized Advantages: Micro-level Feature Gating, pioneers the domain-discriminative feature expressiveness of heterogeneity for modeling heterogeneous segmentation patterns, through a feature modulation comprising non-linear feature transformation and input-adaptive gating mechanism.

3.2 Macro-level Expert Gating Facilitates Stable Expert Allocation for Task-specific Assignment

Stable expert allocation is essential for maintaining balanced expert load distribution in MD-MIS, as implicitly task-dependent characteristics from heterogeneous domains are only exhibited in the feature space, making it difficult for conventional routing strategies to reliably distinguish heterogeneous segmentation patterns. The proposed Macro-level Expert Gating performs task-specific expert allocation by routing heterogeneous segmentation patterns to corresponding experts according to the

inferred task characteristics, promoting stable utilization of expert knowledge for domain-specific segmentation modeling while mitigating expert collapse and expert competition in MD-MIS.

Specifically, the proposed Macro-level Expert Gating introduces a domain-guided strategy that performs stable expert allocation based on the task-discriminative representations enhanced by the Micro-level Feature Gating. Since the heterogeneous segmentation characteristics in MD-MIS are exhibited in domain distribution shifts, we employ a domain recognition branch to learn domain-guided priors from the enhanced representations, thereby providing a reliable basis for expert allocation with the feature representation H as folows:

$$\mathbf{\Pi}^{e|d} = \mathrm{Softmax}_e\left(W_t \cdot \mathrm{GAP}\left(f_\phi(H)\right) + b_t\right), \tag{5}$$

The domain recognition network first extracts domain-discriminative responses, which are then globally aggregated and projected into the expert allocation space. $f_\phi(\cdot)$, $\mathrm{GAP}(\cdot)$ denotes the domain recognition network and global average pooling. $\mathbf{\Pi}^{e|d} \in \mathbb{R}^{N_d \times N_e}$ represents the domain-aware expert load distribution. N_d and N_e denote the numbers of domains and experts, respectively. $\mathrm{Softmax}_e(\cdot)$ indicates the normalization performed along the expert dimension. Each row of $\mathbf{\Pi}^{e|d}$ describes the expert load distribution associated with a specific domain.

Domain recognition network is supervised by domain labels during training procedure, where d_i denotes the one-hot domain label of the input sample:

$$\mathcal{L}_{dom} = -\sum_{i=1}^{N_d} d_i \log p_i^d, \tag{6}$$

During inference, the domain label is not externally provided. Instead, the domain recognition branch predicts the domain category of the input sample, and the corresponding row of $\mathbf{\Pi}^{e|d}$ is used for expert allocation.

Based on the selected domain-conditioned expert load distribution, the Top-K experts with the largest load scores are activated:

$$\mathcal{K} = \mathrm{TopK}\left(\boldsymbol{\pi}^{e|d}\right), \tag{7}$$

where $\mathcal{K}$ denotes the index set of the selected experts. The outputs of these selected experts are then aggregated according to their normalized expert-load weights:

$$Y = \sum_{k \in \mathcal{K}} \alpha_k E_k(\hat{H}), \qquad \alpha_k = \frac{\pi_k^{e|d}}{\sum_{j \in \mathcal{K}} \pi_j^{e|d}}, \tag{8}$$

where $E_k(\cdot)$ denotes the k-th expert and α_k represents the normalized selected expert load weight. In this way, the proposed Macro-level Expert Gating first estimates the domain-aware expert load distribution and then selects the Top-K experts according to the load pattern associated with the corresponding domain, which explicitly models the allocation relationship between domains and experts, reduces unstructured expert allocation caused by overlapped feature distributions, and maintains a more stable expert load distribution for domain-specific segmentation modeling.

Summarized Advantages: Macro-level Expert Gating, pioneers the stable task-specific expert allocation while mitigating expert competition and collapse for heterogeneous segmentation modeling, through domain-guided Top-K routing mechanism based on expert load modeling.

4 Experiments

4.1 Experimental Setups

4.1.1 Datasets.

Three public datasets, including FLARE22 (Ma et al., 2024b), AMOS22 (Ji et al., 2022), and MSWAL (Wu et al., 2025), are utilized to establish complementary evaluation settings covering both multi and single-domain scenarios, enabling a comprehensive assessment across different sources of data heterogeneity. All datasets are preprocessed using the same pipeline as (Vu et al., 2020).

• FLARE22 and AMOS22 are jointly employed to construct a heterogeneous multi-domain segmentation setting. FLARE22 contains 50 labeled contrast-enhanced abdominal CT scans with annotations of 14 organs, whereas AMOS22 includes 600 labeled abdominal scans, comprising 500 CT and 100 MRI cases with annotations of 16 anatomical structures. Their combination introduces both cross-dataset and cross-modality heterogeneity, enabling a comprehensive evaluation of domain-dependent variations modeling.

• MSWAL is employed to construct a single-domain segmentation setting, which further evaluates the robustness and reliability of HG-MoE in a single-domain scenario. It contains CT scans from 694 patients with 191,417 slices and annotations of seven common abdominal lesion types.

4.1.2 Implementation Details.

All training and inference procedure is conducted on a single NVIDIA RTX 4090 GPU. The models are trained for 150 epochs with a batch size of 24. The MoE backbone architecture follows Swin U-Net (Cao et al., 2022). To ensure fair comparison, the encoder-decoder structure, input resolution, and patch size were kept identical to (Jiang & Shen, 2024). The encoder was initialized using MAE-pretrained weights to stabilize representation learning. Models are optimized using an Adam optimizer (Kingma & Ba, 2014) with a learning rate of 10^{-4}.

4.1.3 Evaluation Metrics.

Four widely used metrics are adopted for comprehensive evaluation of both segmentation accuracy and reliability, including Dice coefficient (Taha & Hanbury, 2015), Recall (Powers, 2011), mean Intersection over Union (mIoU) (Everingham et al., 2010), and region-level F1 (Powers, 2011). Dice evaluates overall spatial overlap while Recall reflects sensitivity to target regions. Furthermore, mIoU provides a stricter class-wise assessment of segmentation accuracy and region-level F1 balances false-positive and false-negative errors.

4.2 Experimental Results

Table 1: Comparison of mean Dice (%) on FLARE22 and AMOS22 datasets demonstrate the superiority of the proposed HG-MoE.

Method	FLARE22	AMOS22
DoDNet (Zhang et al., 2021)	87.62	80.44
Med3D (Chen et al., 2019)	86.33	73.97
Liu's (Liu et al., 2022)	74.13	–
SAM-Med2D (Cheng et al., 2023)	73.67	71.34
M^{4}OE (Jiang & Shen, 2024)	87.68	83.99
HG-MoE (Ours)	**94.08**	**88.84**

Table 2: Extended evaluation metrics (%) on FLARE22 and AMOS22 datasets further validate the superiority of the proposed HG-MoE.

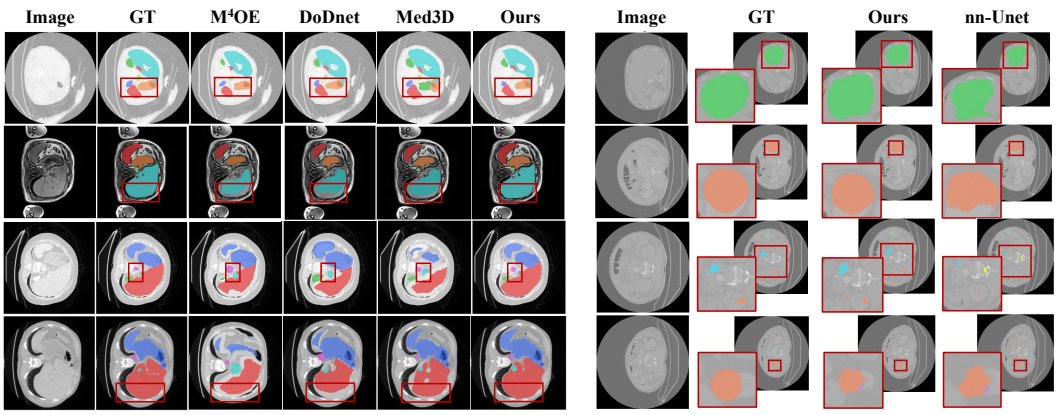
Metrics	FLARE22	AMOS22
FG Dice	94.08	88.84
FG Recall	94.00	89.54
Macro Dice	69.34	61.08
mIoU	61.13	50.83

Table 3: Per-organ Dice comparison on FLARE22 demonstrates the superior performance of HGM. For clarity and conciseness, Dice scores are presented for seven representative organs, whereas the reported average is computed across all 14 annotated organs. Red and blue indicate the best and second-best results in each column, respectively.

Method	Liver	Left Kidney	Right Kidney	Esophagus	Aorta	Spleen	Stomach	Overall Average
DoDNet (Zhang et al., 2021)	89.15	89.66	88.99	68.28	85.37	84.78	75.76	87.62
Med3D (Chen et al., 2019)	88.34	87.23	88.47	71.60	84.61	91.98	71.60	86.33
Liu's (Liu et al., 2022)	89.30	76.32	78.40	57.19	75.15	83.07	67.82	74.13
SAM-Med2D (Cheng et al., 2023)	80.37	77.96	78.19	62.63	88.19	82.47	75.01	73.67
M^{4}OE (Jiang & Shen, 2024)	96.19	92.88	91.92	66.74	88.67	95.74	88.92	87.68
HG-MoE (Ours)	96.75	93.25	93.33	71.33	92.58	96.38	86.17	94.08

4.2.1 Evaluation on Mixed-Domain Segmentation.

Table 1 demonstrates the quantitative evaluation among the proposed HG-MoE and advanced methods on FLARE22 and AMOS22 datasets. The proposed HG-MoE achieves the highest mean Dice on both FLARE22 and AMOS22, reaching 94.08% and 88.84%, respectively, improving the mean



(a) Qualitative comparison results on FLARE22 and AMOS22

(b) Qualitative comparison results on MSWAL

Figure 3: Comparative qualitative results on three public datasets among ground truths, advanced methods and the proposed HG-MoE demonstrate both better alignment with ground truths and higher per-class accuracy.

Table 4: Per-organ Dice comparison on AMOS22 demonstrates the superior performance of HGM. For clarity and conciseness, Dice scores are presented for seven representative organs, whereas the reported average is computed across all 16 annotated organs. Red and blue indicate the best and second-best results in each column, respectively.

Method	Liver	Left Kidney	Right Kidney	Esophagus	Aorta	Spleen	Stomach	Overall Average
DoDNet (Zhang et al., 2021)	50.33	76.16	82.74	65.05	76.55	75.47	79.99	80.44
Med3D (Chen et al., 2019)	59.42	74.31	81.45	45.68	77.12	79.34	64.65	73.97
SAM-Med2D (Cheng et al., 2023)	86.39	79.68	67.89	51.57	85.00	86.61	77.87	71.34
M^{4}OE (Jiang & Shen, 2024)	87.74	81.12	83.55	55.56	84.33	82.59	73.28	83.99
HG-MoE (Ours)	92.78	82.19	84.33	57.00	85.12	89.90	69.99	88.84

Dice by 7.30% on FLARE22 and 5.77% on AMOS22 compared with the corresponding runner-up methods. The consistent improvements indicate the capacity of effectively capturing the inherent heterogeneity across different data domains of the proposed HG-MoE. Table 2 further provides a complementary evaluation across several extended metrics, consistently demonstrating the superior ability of HG-MoE to achieve reliable segmentation in mixed-domain scenarios.

Furthermore, the per-organ results on FLARE22 and AMOS22, illustrated in Tables 3 and 4, demonstrate the effectiveness of HG-MoE. Across both datasets, HG-MoE achieves the highest Dice scores on five of the seven representative organs including liver, left kidney, right kidney, aorta, and spleen. Notably, on FLARE22 and AMOS22, HG-MoE achieves maximum organ-level relative Dice gains of 4.41% and 5.74%, respectively. Consistent gains at organ-specific level across the two datasets highlight the effectiveness of HG-MoE for modeling heterogeneity across domains.

Fig. 3(a) provides qualitative comparisons among the proposed HG-MoE and advanced methods on both FLARE22 and AMOS22. HG-MoE generates segmentations that more closely align with the ground-truth annotations, particularly in preserving anatomical boundaries. Visual improvements across heterogeneous datasets further demonstrate the effectiveness and robustness of HG-MoE for MD-MIS in heterogeneous segmentation scenarios.

4.2.2 Evaluation on Single-Domain Segmentation.

Fig. 4 illustrates the quantitative evaluation among HG-MoE and six advanced methods including (Isensee et al., 2021; 2024; Hatamizadeh et al., 2021; Roy et al., 2023; Zhou et al., 2023; McConnell et al., 2023) on MSWAL dataset, where HG-MoE achieves the best overall performance. Compared with runner-up methods, HG-MoE improves Dice by 3.04% and region-level F1 by 30.81%, demonstrating the segmentation effectiveness and robustness in single-domain scenarios, confirming broader applicability across both mixed-domain and single-domain medical image segmentation scenarios of the proposed HG-MoE.

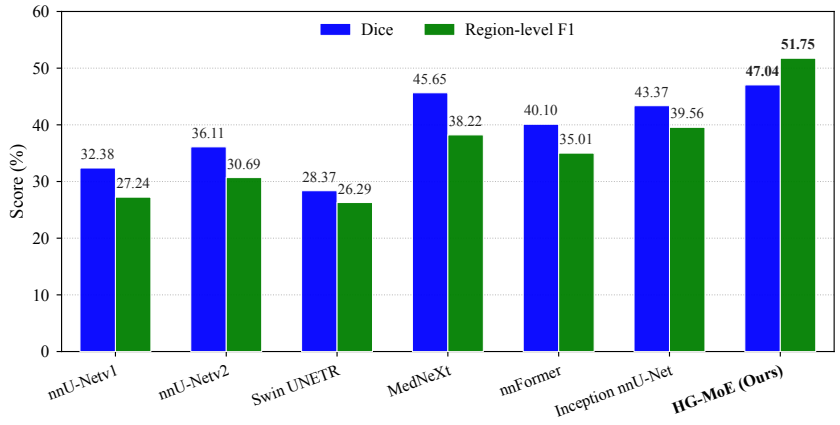


Figure 4: Quantitative evaluation demonstrate the effectiveness of the proposed HG-MoE in scenarios of single-domain medical image segmentation, enhancing the robustness and generalizability of the proposed MG-MoE.

Table 5: Ablation study of the two gating components on FLARE22 and AMOS22 (mean Dice, %) validates the effectiveness of model innovations.

Baseline	Micro	Macro	FLARE22	AMOS22
✓	×	×	86.39	73.65
✓	×	✓	88.64	77.91
✓	✓	×	91.67	76.81
✓	✓	✓	**94.08**	**88.84**

Fig. 3(b) provides qualitative comparison among HG-MoE and existing methods on MSWAL dataset. HG-MoE generates predictions that more closely match the ground-truth annotations, particularly in delineating irregular lesion boundaries, further confirming the effectiveness and versatility of HG-MoE in a single-domain scenario.

4.2.3 Ablation Study.

Table 5 reports the conducted ablation study evaluating four model variants. Both Macro- and Micro-Gating independently improve segmentation performance, achieving relative gains of up to 5.78% and 6.11%, respectively, confirming the effectiveness of both innovative components. When jointly incorporated, HG-MoE achieves the best performance on both datasets, with improvements of 8.9% and 20.62% on FLARE22 and AMOS22 over the baseline, underscoring the superiority and effectiveness of their complementary integration in modeling domain-specific segmentation patterns for the hierarchical gating mechanism in MD-MIS.

5 Conclusion

In this work, we proposed Hierarchical Gating Mixture-of-Experts (HG-MoE), addressing the challenge of learning domain-specific segmentation patterns from mixed-domain medical images. Specifically, we proposed Micro-level Feature Gating, which enhances feature expressiveness through a feature modulation comprising non-linear feature transformation and input-adaptive gating mechanism, thereby enhancing domain-discriminative characterization of implicit task heterogeneity. Secondly, we designed Macro-level Expert Gating, which performs Top-K routing according to domain-aware expert load distributions to promote stable task-specific expert allocation while mitigating expert competition and collapse. Experimental results on three public datasets demonstrated HG-MoE consistently outperforms existing methods across different domain settings ans achieves SOTA performance across multiple evaluation metrics. Through multifaceted analysis and comprehensive evaluation, HG-MoE offers a versatile and reliable solution for MD-MIS.

AI USE STATEMENT

In this work, we used generative AI tools to assist with English language editing and polishing of this manuscript, including improvements to grammar, clarity, and readability. The authors carefully reviewed and revised all AI-assisted wording to ensure that it accurately conveys the intended meaning. The authors take full responsibility for the final content of the paper.

ETHICS STATEMENT

All experiments used publicly available medical imaging datasets in accordance with their respective terms of use. We did not collect new patient data, access identifiable patient information, or release additional patient-level data. Therefore, this work raises no additional ethical concerns.

REPRODUCIBILITY STATEMENT

To support reproducibility, we provide code for the key components of HG-MoE as supplementary material. The main paper and appendix describe the experimental procedures, computational resources, parameter settings, dataset preprocessing, and definitions and implementation details of the primary evaluation metrics. We will release the complete codebase in a public repository.

REFERENCES

Diego Adame, Fabian Vazquez, Jose A Nunez, Huimin Li, Jinghao Yang, Erik Enriquez, DongChul Kim, Haoteng Tang, Bin Fu, and Pengfei Gu. Patch-moe mamba: A patch-ordered mixture-of-experts state space architecture for medical image segmentation. *arXiv preprint arXiv:2605.17719*, 2026.

Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. The medical segmentation decathlon. *Nature communications*, 13(1):4128, 2022.

Cheng Bian, Chenglang Yuan, Jiexiang Wang, Meng Li, Xin Yang, Shuang Yu, Kai Ma, Jin Yuan, and Yefeng Zheng. Uncertainty-aware domain alignment for anatomical structure segmentation. *Medical Image Analysis*, 64:101732, 2020.

Victor M Campello, Polyxeni Gkontra, Cristian Izquierdo, Carlos Martin-Isla, Alireza Sojoudi, Peter M Full, Klaus Maier-Hein, Yao Zhang, Zhiqiang He, Jun Ma, et al. Multi-centre, multi-vendor and multi-disease cardiac segmentation: the m&ms challenge. *IEEE Transactions on Medical Imaging*, 40(12):3543–3554, 2021.

Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*, pp. 205–218. Springer, 2022.

Sihong Chen, Kai Ma, and Yefeng Zheng. Med3d: Transfer learning for 3d medical image analysis. *arXiv preprint arXiv:1904.00625*, 2019.

Junlong Cheng, Jin Ye, Zhongying Deng, Jianpin Chen, Tianbin Li, Haoyu Wang, Yanzhou Su, Ziyan Huang, Jilong Chen, Lei Jiang, et al. Sam-med2d. *arXiv preprint arXiv:2308.16184*, 2023.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2): 303–338, 2010.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Hao Guan and Mingxia Liu. Domain adaptation for medical image analysis: a survey. *IEEE Transactions on Biomedical Engineering*, 69(3):1173–1185, 2021.

Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI brainlesion workshop*, pp. 272–284. Springer, 2021.

Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.

Fabian Isensee, Tassilo Wald, Constantin Ulrich, Michael Baumgartner, Saikat Roy, Klaus Maier-Hein, and Paul F Jaeger. nnu-net revisited: A call for rigorous validation in 3d medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 488–498. Springer, 2024.

Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

Yuanfeng Ji, Haotian Bai, Chongjian Ge, Jie Yang, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhanng, Wanling Ma, Xiang Wan, et al. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. *Advances in neural information processing systems*, 35:36722–36732, 2022.

Yufeng Jiang and Yiqing Shen. M4oe: A foundation model for medical multimodal image segmentation with mixture of experts. In *international conference on medical image computing and computer-assisted intervention*, pp. 621–631. Springer, 2024.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Bo Li, Yifei Shen, Jingkang Yang, Yezhen Wang, Jiawei Ren, Tong Che, Jun Zhang, and Ziwei Liu. Sparse mixture-of-experts are domain generalizable learners. *arXiv preprint arXiv:2206.04046*, 2022.

Fanding Li, Xiangyu Li, Xianghe Su, Xingyu Qiu, Suyu Dong, Wei Wang, Kuanquan Wang, Gongning Luo, and Shuo Li. Ambiguity-aware truncated flow matching for ambiguous medical image segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 6082–6090, 2026a.

Xiangyu Li, Fanding Li, Yongfeng Yuan, Suyu Dong, Kuanquan Wang, Yi Shen, Guohua Wang, Gongning Luo, and Shuo Li. Plato: Probabilistic hierarchical multi-head model for plug-and-play ambiguous medical image segmentation. *Knowledge-Based Systems*, pp. 115746, 2026b.

Quande Liu, Cheng Chen, Jing Qin, Qi Dou, and Pheng-Ann Heng. Feddg: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1013–1023, 2021.

Wentao Liu, Weijin Xu, Songlin Yan, Lemeng Wang, Haoyuan Li, and Huihua Yang. Combining self-training and hybrid architecture for semi-supervised abdominal organ segmentation. In *MICCAI challenge on fast and low-resource semi-supervised abdominal organ segmentation*, pp. 281–292. Springer, 2022.

Yujie Lu, Jingwen Li, Sibo Ju, Yanzhou Su, He Yao, Yisong Liu, Min Zhu, and Junlong Cheng. Segmote: Token-level mixture of experts for medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 36332–36342, 2026.

Jiaqi Ma, Zhe Zhao, Xinyang Yi, Jilin Chen, Lichan Hong, and Ed H Chi. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 1930–1939, 2018.

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature communications*, 15(1):654, 2024a.

Jun Ma, Yao Zhang, Song Gu, Cheng Ge, Shihao Mae, Adamo Young, Cheng Zhu, Xin Yang, Kangkang Meng, Ziyan Huang, et al. 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, 2024b.

Qinghe Ma, Jian Zhang, Lei Qi, Qian Yu, Yinghuan Shi, and Yang Gao. Unleashing the power of intermediate domains for mixed domain semi-supervised medical image segmentation. *IEEE Transactions on Medical Imaging*, 2025.

Niccolò McConnell, Nchongmaje Ndipenoch, Yu Cao, Alina Miron, and Yongmin Li. Exploring advanced architectural variations of nnunet. *Neurocomputing*, 560:126837, 2023.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pp. 565–571. Ieee, 2016.

Basil Mustafa, Carlos Riquelme, Joan Puigcerver, Rodolphe Jenatton, and Neil Houlsby. Multimodal contrastive learning with limoe: the language-image mixture of experts. *Advances in Neural Information Processing Systems*, 35:9564–9576, 2022.

Cheng Ouyang, Chen Chen, Surui Li, Zeju Li, Chen Qin, Wenjia Bai, and Daniel Rueckert. Causality-inspired single-source domain generalization for medical image segmentation. *IEEE Transactions on Medical Imaging*, 42(4):1095–1106, 2022.

Szymon Płotka, Gizem Mert, Maciej Chrabaszcz, Ewa Szczurek, and Arkadiusz Sitek. Mamba goes home: Hierarchical soft mixture-of-experts for 3d medical image segmentation. *Advances in Neural Information Processing Systems*, 38:97871–97909, 2026.

David Powers. Evaluation: From precision, recall and f-measure to roc, informedness, markedness & correlation. *Journal of Machine Learning Technologies*, 2(1):37–63, 2011.

Arghavan Rezvani, Xiangyi Yan, Anthony T Wu, Kun Han, Pooya Khosravi, and Xiaohui Xie. Mome: Mixture of visual language medical experts for medical imaging segmentation. *arXiv preprint arXiv:2510.26996*, 2025.

Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. *Advances in Neural Information Processing Systems*, 34:8583–8595, 2021.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Saikat Roy, Gregor Koehler, Constantin Ulrich, Michael Baumgartner, Jens Petersen, Fabian Isensee, Paul F Jaeger, and Klaus H Maier-Hein. Mednext: transformer-driven scaling of convnets for medical image segmentation. In *International conference on medical image computing and computer-assisted intervention*, pp. 405–415. Springer, 2023.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017.

Abdel Aziz Taha and Allan Hanbury. Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool. *BMC medical imaging*, 15(1):29, 2015.

Constantin Ulrich, Fabian Isensee, Tassilo Wald, Maximilian Zenk, Michael Baumgartner, and Klaus H Maier-Hein. Multitalent: A multi-dataset approach to medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 648–658. Springer, 2023.

Minh H Vu, Guus Grimbergen, Tufve Nyholm, and Tommy Löfstedt. Evaluation of multislice inputs to convolutional neural networks for medical image segmentation. *Medical Physics*, 47 (12):6216–6231, 2020.

Jiangnan Wang, Caixia Zhou, and Yaping Huang. Contour-aware multi-expert model for ambiguous medical image segmentation. *IEEE Transactions on Medical Imaging*, 2025.

Shujun Wang, Lequan Yu, Kang Li, Xin Yang, Chi-Wing Fu, and Pheng-Ann Heng. Dofe: Domain-oriented feature embedding for generalizable fundus image segmentation on unseen datasets. *IEEE Transactions on Medical Imaging*, 39(12):4237–4248, 2020.

Zhaodong Wu, Qiaochu Zhao, Ming Hu, Yulong Li, Haochen Xue, Zhengyong Jiang, Angelos Stefanidis, Qiufeng Wang, Imran Razzak, Zongyuan Ge, et al. Mswal: 3d multi-class segmentation of whole abdominal lesions dataset. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 378–388. Springer, 2025.

Jee Seok Yoon, Kwanseok Oh, Yooseung Shin, Maciej A Mazurowski, and Heung-Il Suk. Domain generalization for medical image analysis: A review. *Proceedings of the IEEE*, 112(10):1583–1609, 2024.

Jianpeng Zhang, Yutong Xie, Yong Xia, and Chunhua Shen. Dodnet: Learning to segment multi-organ and tumors from multiple partially labeled datasets. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1195–1204, 2021.

Hong-Yu Zhou, Jiansen Guo, Yinghao Zhang, Xiaoguang Han, Lequan Yu, Liansheng Wang, and Yizhou Yu. nnformer: Volumetric medical image segmentation via a 3d transformer. *IEEE transactions on image processing*, 32:4036–4045, 2023.

A Appendix

A.1 Appendix 1: Detailed Problem Definition of Mixed-Domain Medical Image Segmentation

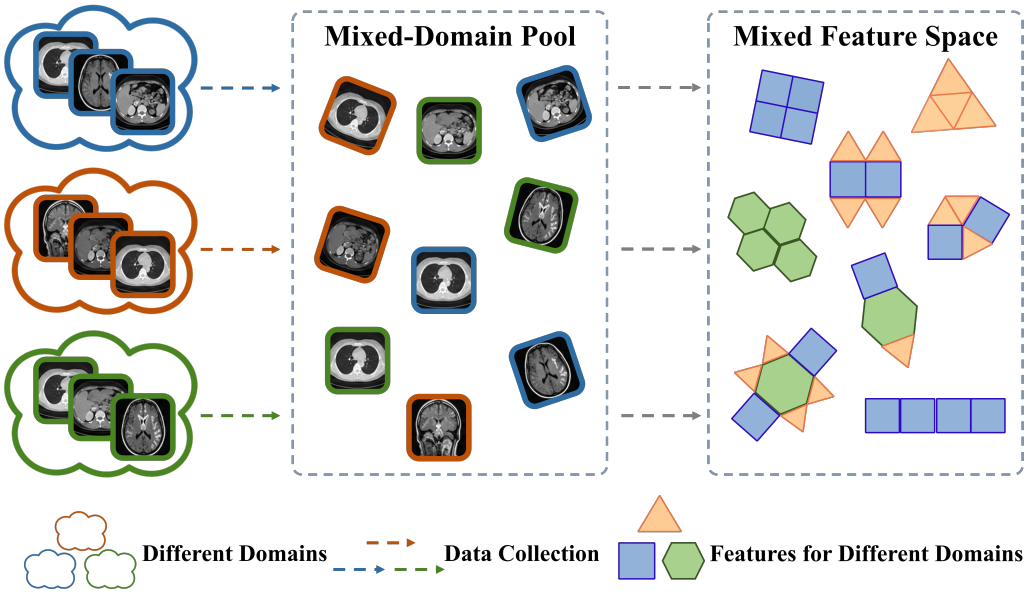


Figure 5: Problem definition for mixed-domain medical image segmentation (MD-MIS). Medical images acquired from different institutions, scanners, and acquisition protocols are typically aggregated into a unified training set without directly provided sample-level domain annotations. MD-MIS exhibits inherent multi-task characteristics, requiring a segmentation framework to simultaneously learn shared anatomical knowledge and domain-dependent representations from heterogeneous mixed-domain data.

As Fig. 5, mixed-domain medical image segmentation (MD-MIS) aims to train a unified segmentation model on medical images aggregated from multiple heterogeneous domains, including different

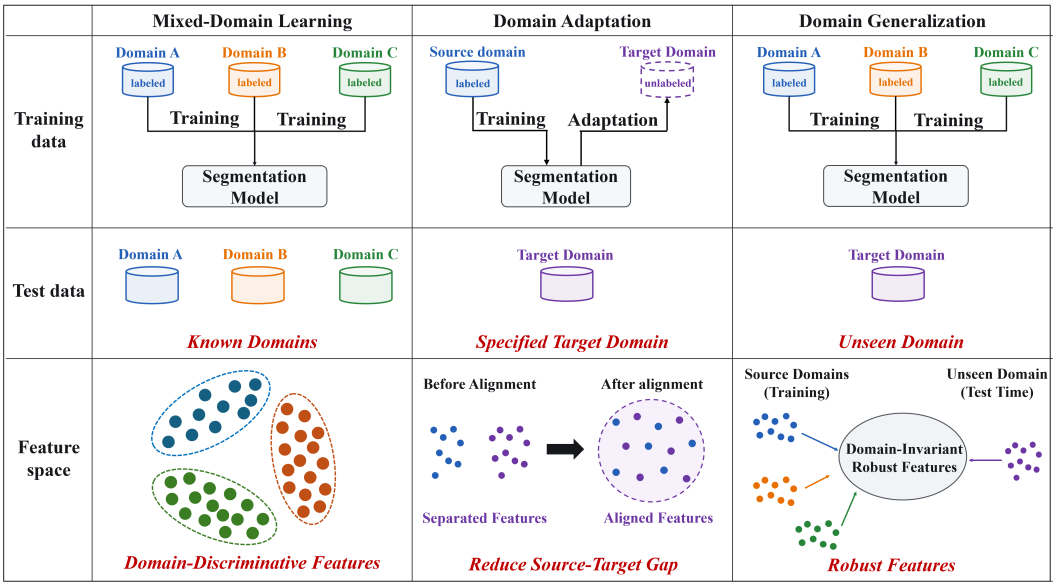


Figure 6: Comparison of mixed-domain learning, domain adaptation, and domain generalization in terms of training data, test domains, feature-space organization, and learning objectives. Mixed-domain learning jointly trains a unified model on multiple labeled and known domains, aiming to capture shared semantics while preserving domain-specific characteristics for coordinated prediction across all observed domains. Domain adaptation follows an asymmetric source-to-target setting, where labeled source-domain data and accessible target-domain data are used to reduce the source–target distribution gap and improve performance on a specified target domain. Domain generalization, in contrast, learns from multiple labeled source domains without accessing the target domain during training, seeking domain-robust representations that can generalize to previously unseen domains.

institutions, scanners, modalities, and acquisition protocols. Despite sharing common anatomical structures and segmentation semantics, these domains exhibit distinct appearance distributions and domain-dependent segmentation patterns. Consequently, the model must learn shared anatomical knowledge from the mixed-domain data while dynamically adapting its representation and prediction to the domain characteristics of each input image.

Furthermore, as illustrated in Fig. 6, MD-MIS is related to but fundamentally different from domain adaptation (DA) and domain generalization (DG). DA typically assumes an asymmetric source-to-target transfer setting, in which labeled source-domain data and available target-domain data are jointly used to reduce the distribution gap and improve performance on a specified target domain. DG, by contrast, learns from multiple source domains without accessing the target domain during training, aiming to extract domain-robust representations that can generalize to unseen domains. Similar to DG, MD-MIS jointly learns from multiple heterogeneous domains; however, its objective is not to generalize beyond the observed domains, but to achieve coordinated and specialized prediction across multiple known domains. Accordingly, rather than completely aligning different domains into a unified domain-invariant feature space, MD-MIS seeks to preserve shared anatomical semantics while retaining domain-dependent representations that support adaptive segmentation for each input domain.

A.1.1 Further Discussion of Underlying Mechanism

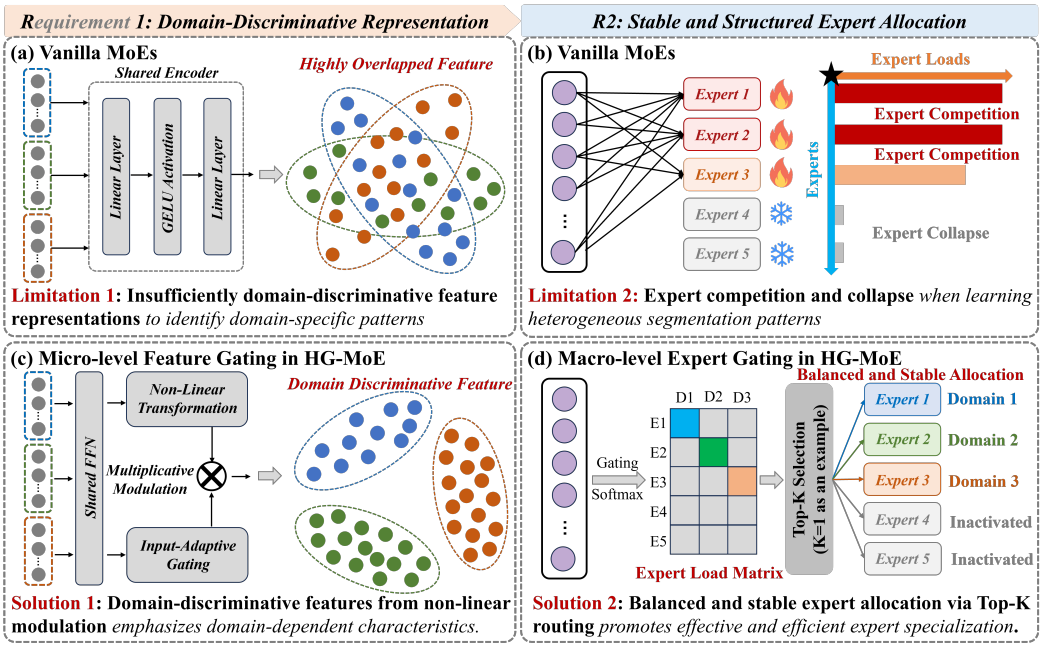


Figure 7: Limitations of vanilla MoEs and the solutions offered by the proposed HG-MoE. From R1 to R2, they are the necessary requirements of performing reliable MD-MIS. From (a) to (b), they are the limitations of vanilla MoEs, and from (c) to (d), they are the solutions and advantages of HG-MoE, respectively. (a) Vanilla MoEs lack domain-discriminative feature representations due to insufficient feature expressiveness, resulting in highly overlapping features across heterogeneous domains. (b) Vanilla MoEs suffer from expert competition and collapse due to unstructured routing mechanisms and imbalanced expert loads. (c) Micro-level Feature Gating in the proposed HG-MoE achieves domain-discriminative feature representations through nonlinear transformation and input-adaptive multiplicative modulation. (d) Macro-level Expert Gating in the proposed HG-MoE achieves balanced and stable expert allocation through domain-guided Top-K routing based on expert-load modeling.

A.1.2 Potential Underlying Mechanism of Mixture-of-Experts

Mixture-of-Experts frameworks (MoEs) demonstrate strong potential for mixed-domain medical image segmentation (MD-MIS) through expert specialization, a mechanism that distributes heterogeneous knowledge among multiple experts within a unified framework. Specifically, MoE preserves shared representation learning while dynamically assigning input-dependent characteristics to different experts, allowing each expert to capture distinct segmentation patterns. This design reveals that a unified model does not need to encode all heterogeneous knowledge within a single undifferentiated parameter space; instead, domain-dependent characteristics can be separately modeled through specialized experts. Although MoE was originally developed to improve model capacity and conditional computation, its architecture implicitly provides a flexible mechanism for heterogeneous data modeling. This property is particularly valuable for MD-MIS, where images from different domains exhibit subtle latent distribution shifts despite sharing the same imaging modality, anatomical region, and segmentation semantics. By interpreting these domain-dependent variations as implicit task heterogeneity, MoE enables the preservation of shared anatomical knowledge while supporting domain-specific segmentation modeling. Thus, MoE not only serves as an efficient conditional modeling framework, but also inspires a reconsideration of how expert specialization can be exploited to learn heterogeneous segmentation patterns in MD-MIS.

A.1.3 Superior Mechanism of the proposed Hierarchical Gating Mixture-of-Experts for MD-MIS Inspired by MoEs

The proposed Hierarchical Gating Mixture-of-Experts (HG-MoE) further realizes the potential of MoE for MD-MIS through a hierarchical reformulation of expert specialization, which coordinates heterogeneous characterization and expert allocation at the feature and routing levels. Specifically, HG-MoE retains the multi-expert architecture of conventional MoE while replacing direct single-stage routing with two complementary gating processes. Micro-level Feature Gating first enhances feature expressiveness and exposes the implicit task heterogeneity underlying subtle domain shifts, while Macro-level Expert Gating subsequently transforms these domain-dependent characteristics into domain-aware expert load distributions for stable Top-K routing. This hierarchical design reveals that effective expert specialization depends not only on selecting experts, but also on constructing sufficiently discriminative representations before expert assignment. Although conventional MoE provides flexible model capacity through conditional expert activation, its direct application to MD-MIS remains limited by indistinguishable heterogeneous patterns and unstable expert allocation. HG-MoE addresses these limitations by progressively establishing a reliable correspondence between heterogeneous segmentation characteristics and task-relevant experts. Thus, HG-MoE not only inherits the conditional modeling flexibility of MoE, but also reformulates generic expert routing into a hierarchical specialization mechanism tailored to MD-MIS.

A.2 Appendix 2: Further Discussion on the Innovations of HG-MoE

As illustrated in Fig. 7, effective expert specialization in MD-MIS involves two complementary requirements: domain-discriminative representation learning (**R1**) and stable and structured expert allocation (**R2**). However, vanilla MoEs fail to adequately satisfy these requirements because their shared feature transformations produce highly overlapping representations across heterogeneous domains (Fig. 7(a)), while their insufficiently guided routing mechanisms may result in imbalanced expert loads, expert competition, and expert collapse (Fig. 7(b)). Accordingly, HG-MoE introduces Micro-level Feature Gating and Macro-level Expert Gating as two hierarchical solutions to representation enhancement and expert allocation, respectively.

Innovation 1: Micro-level Feature Gating for Domain-Discriminative Representation. The proposed Micro-level Feature Gating addresses the challenge of implicit task heterogeneity in MD-MIS by reshaping the original, highly overlapping feature space into a more domain-discriminative representation space, as illustrated in Fig. 7(c). Specifically, it couples nonlinear feature transformation with input-adaptive multiplicative modulation. The nonlinear transformation enhances representation expressiveness by modeling nonlinear dependencies among feature dimensions, while the input-adaptive gating mechanism dynamically adjusts feature responses according to latent task-dependent characteristics. As a result, the modulated representations exhibit improved separability among heterogeneous segmentation patterns, enabling more effective utilization of heterogeneous

knowledge and providing a more reliable representation foundation for subsequent task-specific expert allocation.

Innovation 2: Macro-level Expert Gating for Stable and Structured Expert Allocation. The proposed Macro-level Expert Gating addresses expert competition and collapse in MD-MIS by establishing a stable and structured allocation relationship between heterogeneous domains and specialized experts, as illustrated in Fig. 7(d). Specifically, the domain recognition branch captures domain-discriminative cues from the enhanced feature representations and projects them into a domain-aware expert-load matrix, where each domain is associated with a corresponding expert-load distribution. Based on this distribution, Top-K selection routes each domain-related input to the most relevant subset of experts. Consequently, the proposed mechanism performs explicit expert assignment based on inferred domain-dependent characteristics, reducing ambiguous routing decisions caused by entangled feature distributions, balancing expert utilization, and mitigating expert competition and collapse. This structured allocation provides a stable foundation for effective expert specialization and domain-specific segmentation pattern modeling.

Together, these two innovations constitute a hierarchical gating process. Micro-level Feature Gating first enhances the separability of heterogeneous segmentation representations, determining what information is provided for expert allocation. Macro-level Expert Gating subsequently exploits these enhanced representations to determine how specialized expert knowledge is allocated. Their complementary integration enables HG-MoE to jointly improve domain-discriminative representation learning and expert-allocation stability for MD-MIS.

A.3 Appendix 3: Detailed Network Structures

A.3.1 Structure of the HG-MoE Backbone.

The HG-MoE backbone adopts a four-level Swin-Unet encoder–decoder architecture with symmetric skip connections. Given an input image, a convolutional patch-embedding layer partitions it into non-overlapping 4×4 patches and projects them into 96-dimensional tokens. The four encoder stages operate with feature dimensions of 96, 192, 384, and 768 and employ 3, 6, 12, and 24 attention heads, respectively. Each stage contains alternating window-based multi-head self-attention and shifted-window attention blocks with a default window size of 7. Patch-merging layers halve the spatial resolution and double the feature dimension between consecutive stages. Within each Swin block, the conventional feed-forward network is replaced by a task-conditioned MoE-FFN containing five experts. Each expert consists of two linear layers with a $C \to 4C \to C$ structure and a GELU activation. A softmax gate estimates token-wise expert weights and aggregates the outputs of all experts. The decoder progressively restores the spatial resolution through patch-expansion layers and fuses the corresponding encoder features through concatenation and linear projection. At the bottleneck, a task gate applies global token pooling, layer normalization, and a two-layer MLP to estimate the task identity. Finally, a fourfold patch-expansion layer reconstructs the full-resolution feature map, after which the selected task-specific 1×1 convolutional head generates the segmentation logits.

A.3.2 Structure of the Swin U-Net.

The Swin U-Net processes 224×224 three-channel images using a four-level encoder–decoder architecture with symmetric skip connections. A 4×4 convolution with stride 4 converts each image into a 56×56 token grid with an embedding dimension of 96. The four encoder stages contain $[2, 2, 2, 2]$ Swin Transformer blocks and operate at resolutions of 56^2, 28^2, 14^2, and 7^2. Their corresponding feature dimensions are $[96, 192, 384, 768]$, with $[3, 6, 12, 24]$ attention heads. Each block alternates window-based and shifted-window attention using a window size of 7 and an MLP expansion ratio of 4. The model uses zero dropout and attention dropout, while the stochastic-depth rate increases linearly from 0 to 0.1. Absolute positional embedding is disabled, whereas layer normalization after patch embedding is enabled. The decoder begins with a twofold patch-expansion layer and subsequently applies three decoding stages containing $[2, 2, 2]$ Swin blocks. Decoder features with dimensions 384, 192, and 96 are concatenated with the corresponding encoder features and projected using linear layers. A final fourfold patch-expansion layer restores the feature map to 224×224. For the three-task setting, the task gate globally averages the 768-dimensional bottleneck tokens and applies a $768 \to 384 \to 3$ MLP. The resulting task prediction selects one of

three task-specific 1×1 convolutional segmentation heads. Compatible pretrained parameters are loaded by matching both parameter names and tensor dimensions, while the decoder, MoE modules, task gate, and output heads are initialized for task-specific optimization.

A.4 Appendix 4: Details of Evaluation Metrics

We adopted four widely used metrics for comprehensive evaluation of the segmentation performance, including Dice coefficient, Recall, mean Intersection over Union (mIoU), and region-level F1-score. Let TP, FP, and FN denote true positives, false positives, and false negatives, respectively.

- The Dice coefficient (DSC) measures the spatial overlap between the predicted segmentation and the ground-truth annotation, and is widely used in medical image segmentation tasks. A higher Dice value indicates better agreement between the predicted region and the reference annotation. The Dice coefficient is calculated as follows:

$$\mathrm{Dice} = \frac{2TP}{2TP + FP + FN} \tag{9}$$

- The recall score evaluates the proportion of ground-truth foreground regions that are correctly detected by the model. In medical image segmentation, a higher Recall indicates fewer missed target regions. Recall is computed as

$$\mathrm{Recall} = \frac{TP}{TP + FN} \tag{10}$$

- The mean Intersection over Union (mIoU) is commonly used to evaluate multi-class segmentation performance, which imposes a stricter penalty on false positives and false negatives. For each class, IoU is calculated as

$$\mathrm{IoU} = \frac{TP}{TP + FP + FN} \tag{11}$$

  The mIoU is then obtained by averaging the IoU values over all classes:

$$\mathrm{mIoU} = \frac{1}{C} \sum_{c=1}^{C} \mathrm{IoU}_c \tag{12}$$

  where C denotes the number of classes.

- The region-level F1-score evaluates the ability to correctly identify individual lesion regions by balancing region-level false positives and false negatives. A predicted region is regarded as a true positive when its IoU with a ground-truth region is at least 0.5. Specifically, it is calculated as follows:

$$\mathrm{F1}_{\mathrm{reg}} = \frac{2 \times \mathrm{Precision}_{\mathrm{reg}} \times \mathrm{Recall}_{\mathrm{reg}}}{\mathrm{Precision}_{\mathrm{reg}} + \mathrm{Recall}_{\mathrm{reg}}} \tag{13}$$

  where

$$\mathrm{Precision}_{\mathrm{reg}} = \frac{TP_{\mathrm{reg}}}{TP_{\mathrm{reg}} + FP_{\mathrm{reg}}}. \tag{14}$$

A.5 Appendix 5: More Qualitative Results

We have further provided more qualitative visualizations of segmentation results (Fig. 8 for FLARE22 dataset, Fig. 9 for AMOS22 dataset and Fig. 10 for MSWAL dataset), demonstrating that HG-MoE produces accurate and consistent segmentation predictions, further validating its effectiveness across different cases. Moreover, we reconstructed 3D anatomical structures from the slice-level segmentation masks of both ground-truths and predictions with 3D Slicer (as illustrated in Fig. 11). The resulting 3D visualizations demonstrate that HG-MoE preserves inter-slice continuity, structural integrity, and anatomical consistency, providing further evidence of the effectiveness beyond individual two-dimensional slices.

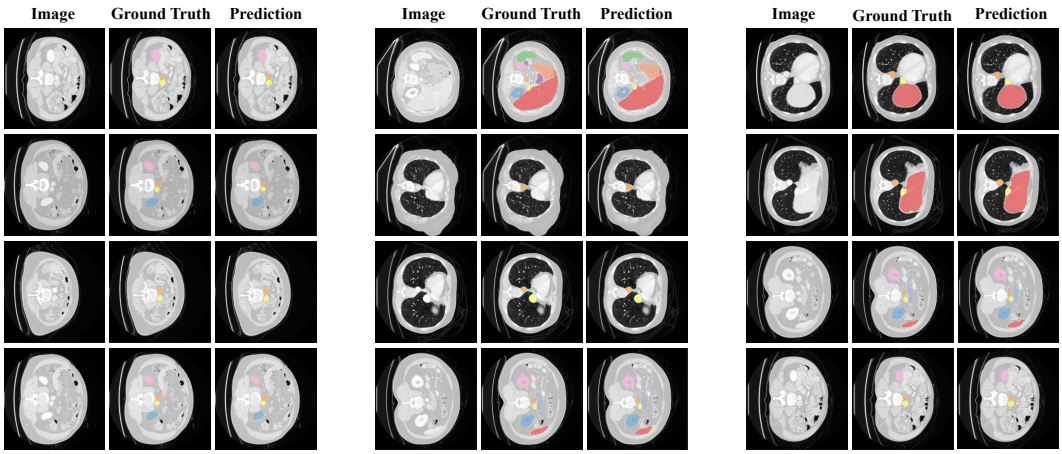


Figure 8: More visualization results of predictions on the FLARE22 dataset demonstrate effectiveness of proposed HG-MoE.

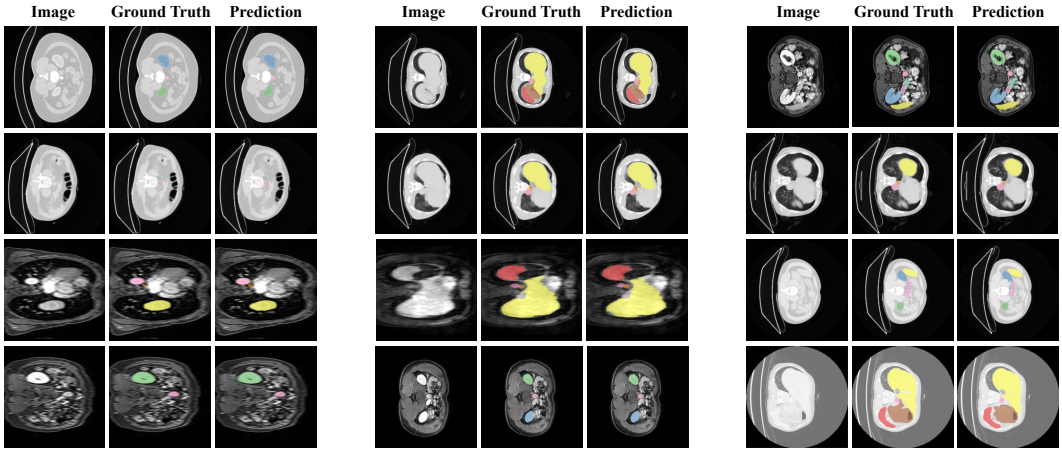


Figure 9: More visualization results of predictions on the AMOS22 dataset demonstrate effectiveness of proposed HG-MoE.

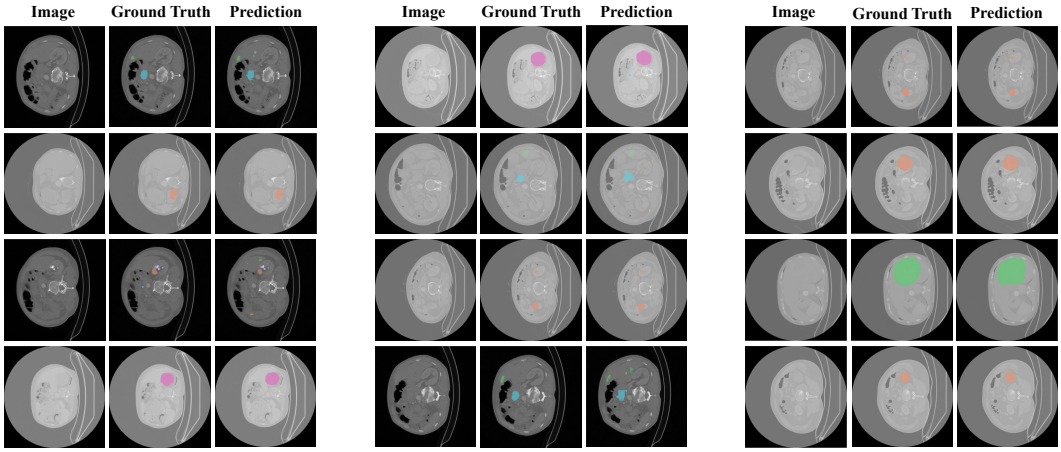


Figure 10: More visualization results of predictions on the MSWAL dataset demonstrate effectiveness of proposed HG-MoE.

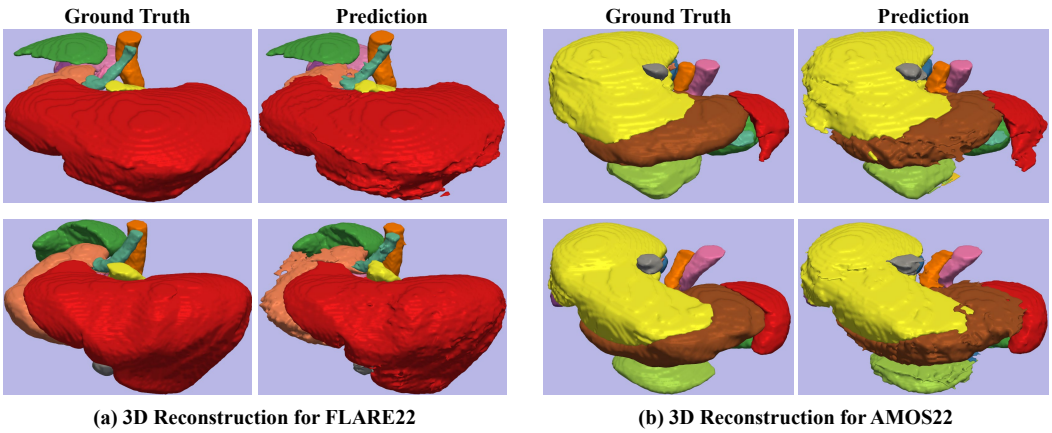


Figure 11: 3D reconstruction visualizations based on slice-level segmentation predictions on the FLARE22 and AMOS22 datasets further demonstrate the effectiveness of the proposed HG-MoE.