DiffOSeg: Omni Medical Image Segmentation via Multi-Expert Collaboration Diffusion Model

Han Zhang[1], Xiangde Luo[1], Yong Chen[1], and Kang Li[1,2(✉)]

[1] West China Biomedical Big Data Center, West China Hospital, Sichuan University
[2] Shanghai Artificial Intelligence Laboratory, Shanghai, China
likang@wchscu.cn

Abstract. Annotation variability remains a substantial challenge in medical image segmentation, stemming from ambiguous imaging boundaries and diverse clinical expertise. Traditional deep learning methods producing single deterministic segmentation predictions often fail to capture these annotator biases. Although recent studies have explored multi-rater segmentation, existing methods typically focus on a single perspective—either generating a probabilistic "gold standard" consensus or preserving expert-specific preferences—thus struggling to provide a more omni view. In this study, we propose DiffOSeg, a two-stage diffusion-based framework, which aims to simultaneously achieve both consensus-driven (combining all experts' opinions) and preference-driven (reflecting experts' individual assessments) segmentation. Stage I establishes population consensus through a probabilistic consensus strategy, while Stage II captures expert-specific preference via adaptive prompts. Demonstrated on two public datasets (LIDC-IDRI and NPC-170), our model outperforms existing state-of-the-art methods across all evaluated metrics. Source code is available at https://github.com/string-ellipses/DiffOSeg.

Keywords: Annotation Variability · Diffusion Model · Medical Image Segmentation

1 Introduction

In medical image segmentation, ambiguous target boundaries or irregular shapes often lead to inter-expert annotation variability [3, 17, 23]. Essentially, this variability arises from two main factors: inherent image ambiguities and expert-specific preferences. For example, tumor boundaries are often ill-defined due to grayscale gradients or imaging artifacts, creating ambiguous annotation criteria. While subjectivity persists, multi-expert consensus can help define "relatively reasonable" boundaries. Besides, experts' annotation styles exhibit distinct variations–radiation oncologists may prioritize broader margins for tumor coverage [23], while surgeons prefer precise boundaries for tissue protection [5]. Therefore, it is essential to simultaneously address the dual demands for capturing consensus and preserving expertise.

Traditional segmentation networks [19, 21] typically generate single deterministic masks, often inadequate for clinical needs. Recent studies [2, 11, 17, 25, 29, 31]

have explored multi-rater segmentation, which can be generally categorized into three paradigms: meta-segmentation, diversified segmentation, and personalized segmentation. Meta-segmentation [25] reconstructs a pseudo gold standard through annotation fusion, though the existence of such a gold standard still remains debated in many medical contexts [3, 17]. Diversified segmentation [2, 11, 29] learns the distribution of multi-expert consensus and then samples multiple plausible annotations from it. Personalized segmentation [22, 31] captures expert-specific styles, producing annotations tailored to individual diagnostic preferences. While both meta- and diversified segmentation prioritize capturing consensus-driven patterns, they fail to preserve expert individuality. Conversely, preference-driven approaches (personalized segmentation) focus on individual experts' styles while ignoring shared clinical insights. Recent studies [14, 16] have attempted to unify consensus and personalized segmentation, but their deterministic point estimation overlooks expert annotation diversity. Although D-Persona [26] attempts to address this limitation, its performance is constrained by a limited expressive latent space and redundant expert-specific projection heads, which overlook inter-expert correlations.

Inspired by the success of Diffusion Probabilistic Models (DPMs) [7,8] in capturing complex data distribution and excelling in conditional generation tasks, we propose DiffOSeg, a diffusion-based multi expert collaboration framework that unifies population-level consensus learning and expert-specific preference adaptation Built on a categorical diffusion model [9, 12, 29], our approach operates in two learning stages: Stage I applys a probabilistic consensus strategy to integrate multi-expert consensus, enhancing segmentation diversity and generalization. Stage II employs learnable preference prompts that encode expert-specific styles through a plug-in prompt block, steering denoising toward expert-specific segmentation patterns. The framework achieves parameter efficiency with 13.1 M parameters, representing a 54% reduction compared to D-Persona's 28.46 M. Extensive experiments on the LIDC-IDRI and NPC-170 datasets demonstrate DiffOSeg's effectiveness in both consensus-driven and preference-driven segmentation tasks, achieving state-of-the-art performance across all evaluated metrics.

Our contributions includes: (1) First diffusion-based framework unifying consensus and preference-driven segmentation; (2) Stage I: Probabilistic consensus strategy integrating multi-rater agreements while preserving distributional diversity; (3) Stage II: Adaptive preference prompting mechanism dynamically encoding expert-specific annotation patterns; (4) Superior performance demonstrated on LIDC-IDRI and NPC-170 datasets for both segmentation paradigms.

2 Method

2.1 Consensus-Driven Segmentation

When annotating the same image, different experts exhibit distinct annotation patterns, reflecting both shared insights and individual preferences. Here raises a key question: How can we effectively capture inter-expert consensus while preserving diversity? To address this, we propose a probabilistic consensus strategy

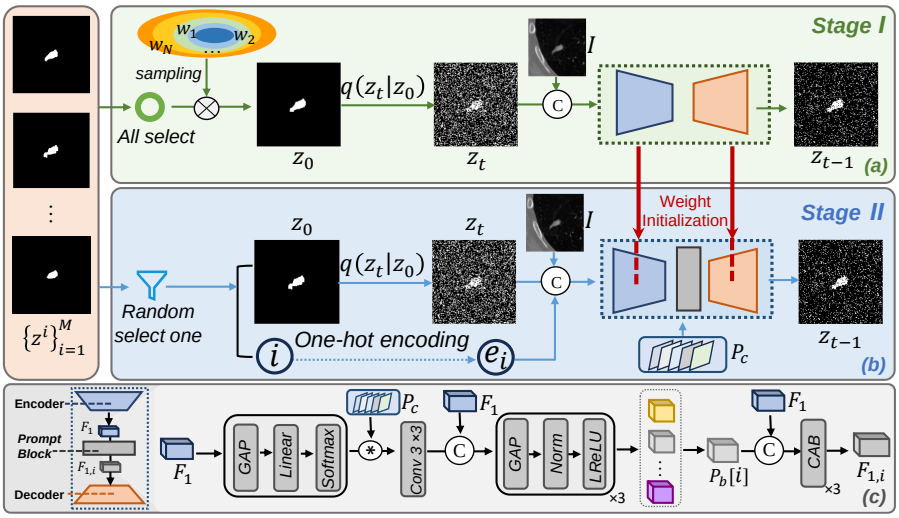

Fig. 1: **DiffOSeg pipeline.** (a) Stage I (Training): Probabilistic consensus stragety dynamiclly integrates multi-expert annotations (Sec 2.1); (b) Stage II (Training): Adaptive prompts model expert preferences (Sec 2.2); (c) Plug-in prompt block (Sec 2.2).

that dynamically integrates multi-expert annotations, enabling the model to reconcile both aspects by considering different expert subsets.

Given an image $I \in \mathbb{R}^{C\times H\times W}$ with M expert annotations $\{\mathbf{z}^i\}_{i=1}^M$, each annotation $\mathbf{z}^i \in \{e_1,\dots,e_L\}^K$ is structured as a $K \times L$ matrix. Here, each row corresponds to a one-hot encoded label, K denotes the number of pixels, and $\mathcal{L} = \{1,\dots,L\}$ represents the set of discrete labels. For example, if the k-th pixel is labeled l, then

$$\mathbf{z}^i[k,:] = e_l \in \{0,1\}^L, \tag{1}$$

Here e_l with 1 in position l and 0 elsewhere. To balance the contributions of multiple experts, we define a probabilistic weight distribution W, which generates a random vector $\mathbf{w} \in [0,1]^M$ that determines the importance of the M experts. The support set $\mathcal{S}$ of W includes all possible configurations of $\mathbf{w}$, representing the power set of the annotation space:

$$\mathcal{S} = \{\mathbf{w} \mid \mathbf{w} \in [0,1]^M, \|\mathbf{w}\|_1 \le M\}. \tag{2}$$

Here, $|\mathbf{w}|_1$ represents the number of active experts contributing to the consensus for a given sample. By sampling $\mathbf{w}$ from S, the model dynamically adjusts the participation of experts, enabling three distinct scenarios: 1) **Single Vote**: Only one expert contributes to the consensus, represented by $\|\mathbf{w}\|_1 = 1$; 2) **Subgroup Consensus**: A subset of experts contributes, where $\|\mathbf{w}\|_1 = k$ and $1 < k < M$; 3) **Full Vote**: All experts participate equally, represented by $\|\mathbf{w}\|_1 = M$.

The consensus label $\mathbf{z}^c$ is computed as a weighted linear combination of expert annotations $\left[\mathbf{z}^1, \mathbf{z}^2, \dots, \mathbf{z}^M\right]$, scaled by the probabilistic weight vector $\mathbf{w}$,

as formulated in Eq. (3) (left). We treated all latent variables $\mathbf{z}_{0:T}^c$ as categorical and modeled both forward and reverse transition distributions using categorical distribution. The initial state $\mathbf{z}_0^c$ is set to the consensus label $\mathbf{z}^c$. The reverse (generative) process is expressed as Eq. (3) (right).

$$\mathbf{z}^c = \mathbf{w} \times \left[\mathbf{z}^1, \mathbf{z}^2, \ldots, \mathbf{z}^M\right]^T, \quad p_\theta\left(\mathbf{z}_{0:T}^c \mid I\right) = p\left(\mathbf{z}_T^c\right) \prod_{t=1}^{T} p_\theta\left(\mathbf{z}_{t-1}^c \mid \mathbf{z}_t^c, I\right), \tag{3}$$

where the reverse transition probability is defined as (for clarity, we omit the superscript c in $\mathbf{z}_{0:T}^c$):

$$\begin{aligned} p_\theta\left(\mathbf{z}_{t-1} \mid \mathbf{z}_t, I\right) &= \begin{cases} \mathcal{C}(\mathbf{z}_0; \hat{\mathbf{p}}_0) & \text{if } t = 1, \\ \sum_{\mathbf{z}_0} \mathcal{C}(\mathbf{z}_{t-1}; \pi(\mathbf{z}_t, \mathbf{z}_0)) \cdot \mathcal{C}(\mathbf{z}_0; \hat{\mathbf{p}}_0) & \text{otherwise.} \end{cases} \\ \pi(\mathbf{z}_t, \mathbf{z}_0) &= \frac{1}{\tilde{\pi}} \left(\frac{\beta_t}{L} \mathbf{1} + \alpha_t \mathbf{z}_t \right) \odot \left(\frac{1 - \bar{\alpha}_{t-1}}{L} \mathbf{1} + \bar{\alpha}_{t-1} \mathbf{z}_0 \right), \end{aligned} \tag{4}$$

where $\tilde{\pi} = \frac{1-\bar{\alpha}_t}{L} + \bar{\alpha}_t \cdot \delta_{z_t}^{z_0}$, δ is the Kronecker delta and $\odot$ denotes element-wise multiplication. For detailed derivations about Eq. (4), please refer to [9, 29].

To approximate $p_\theta\left(\mathbf{z}_{t-1} \mid \mathbf{z}_t, I\right)$, we use a neural network to predict $\hat{\mathbf{p}}_0$. And the training objective is to minimize the Kullback-Leibler (KL) divergence between the forward process posterior $q(\mathbf{z}_{t-1} \mid \mathbf{z}_t, \mathbf{z}_0)$ and the predicted reverse distribution $p_\theta(\mathbf{z}_{t-1} \mid \mathbf{z}_t, I)$, ensuring the model effectively approximates the true reverse process:

$$\begin{aligned} \hat{\mathbf{p}}_0 &= f_\theta(\mathbf{z}_t, t, I) \in [0,1]^{D \times L}, \\ KL(q \parallel p) &= \sum_{d=1}^{D} KL\left(q(\mathbf{z}_{t-1} \mid \mathbf{z}_t, \mathbf{z}_0) \parallel p_\theta(\mathbf{z}_{t-1} \mid \mathbf{z}_t, I)\right). \end{aligned} \tag{5}$$

During inference, the denoising process begins with randomly sampled Gaussian noise and iteratively refines over T steps, guided by the corresponding image condition through the backbone denoising UNet ("backbone" is used to distinct from Stage II's one).

2.2 Preference-Driven Segmentation

Building upon Stage I's consensus features, Stage II specializes in capturing expert-specific deviations that reflect their unique annotation styles. We employ prompting-based techniques [10, 20], where tunable prompts encode individual preference of each expert. These prompts act as dynamic adapters, allowing the model to modulate its behavior based on the target expert's identity, without requiring separate networks for each expert.

As shown in Fig. 1(b), during each training step, we randomly sample an expert annotation z_i per image and encode its identity i as a one-hot vector, as described in [13, 30]. This vector and the corresponding image are initially concatenated with z_t. Then, to enhance feature specificity, our plug-in prompt

block (Fig. 1(c)) injects expert-specific prompts into features from the encoder, enabling dynamic adaptation to individual annotation styles. The prompt integration mechanism operates through two sequential phases:

Prompt-Weight Generation. To generate context-aware prompt-weights, we first extract global context from input features F_1 via spatial average pooling (GAP). This pooled feature is projected to a lower-dimensional space through a linear layer, followed by $softmax$ normalization to produce fusion weights $\lambda \in \mathbb{R}^N$. These weights dynamically adjust the prompt components P_c through a 3×3 convolution. The process is summarized as:

$$\lambda = \text{Softmax}\left(\text{Linear}\left(\text{GAP}\left(F_1\right)\right)\right), \quad P_m = \text{Conv}_{3\times3}(\lambda P_c) \tag{6}$$

Feature Modulation. The modulated prompts P_m are concatenated with F_1 and passed through three GAP-Norm-LReLU blocks to generate dynamic expert-specific weights P_b. For the i-th expert, the corresponding prompt $P_b[i]$ is selected and integrated with backbone features F_1 via cascaded Channel Attention Blocks (CABs) [28], which results in the final prompted feature $F_{1,i}$:

$$P_b = \text{Enhance}(P_m, F_1), \quad F_{1,i} = \text{Modulate}(P_b[i], F_1). \tag{7}$$

The training objective remains the same as in Sec. 2.1. During inference, the process starts from randomly sampled noise, with the image and one-hot encoded identity vector serving as the condition to guide the prompt-driven denoising UNet. After T denoising steps, the model outputs a complete segmentation mask that reflects the target expert's annotation style.

3 Experiments

3.1 Experimental Setup

Dataset. We evaluate our method on two benchmark datasets: LIDC-IDRI and NPC-170. The **LIDC-IDRI** dataset [1] consists of 1,609 2D thoracic CT scans from 214 subjects with manual annotations from four domain experts. To simulate the expert preferences, we manually rank the four annotated regions following the experimental protocols established in [26,31]. Experiments are conducted using the 4-fold cross-validation at the patient level, as described in [4,26]. The **NPC-170** dataset [18,27] includes 170 Nasopharyngeal Carcinoma (NPC) MRI subjects with GTVp annotations from 4 senior radiologists. It is split into training (100 subjects), validation (20 subjects), and testing (50 subjects) sets, following the setup in [26,27].

Implementation Details. All experiments are conducted on a single NVIDIA GeForce RTX 4090 GPU, with total 13.1 M parameters —significantly 54% fewer than D-Persona's 28.46 M [26]. Stage II is initialized using Stage I's pretrained

Table 1: Consensus-driven segmentation performance comparison ($Dice_{soft}$ are denoted as D_n^s and shown as percentages). Results for DiffOSeg and PhiSeg are obtained from our implementation, and others are cited from [26]. Best in **bold**.

Method	LIDC-IDRI						NPC-170	
	$GED_{10}(\downarrow)$	$GED_{30}(\downarrow)$	$GED_{50}(\downarrow)$	$D^s_{10}(\uparrow)$	$D^s_{30}(\uparrow)$	$D^s_{50}(\uparrow)$	$GED_{30}(\downarrow)$	$D^s_{30}(\uparrow)$
Prob. UNet [11]	0.2181	0.2169	0.2168	88.79	88.79	88.80	0.4466	75.34
PhiSeg [2]	0.1934	0.1921	0.1917	89.31	89.73	89.92	-	-
D-Persona (Stage I) [26]	0.1461	0.1375	0.1358	90.24	90.42	90.45	0.2385	**80.40**
DiffOSeg (Stage I)	**0.09154**	**0.0773**	**0.0745**	**91.21**	**92.20**	**92.37**	**0.1822**	79.83

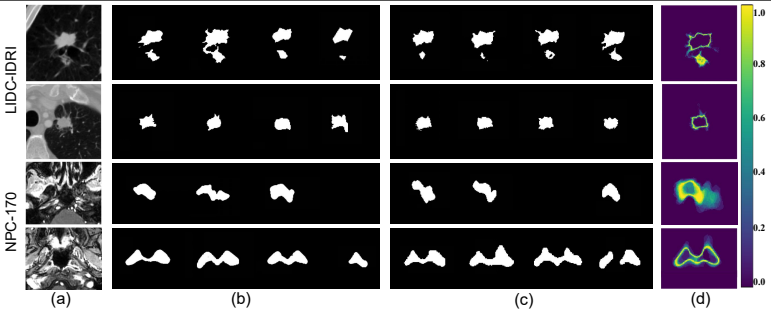

Fig. 2: Consensus-driven segmentation visualization on both datasets. (a) Original images, (b) expert annotations, (c) sampled DiffOSeg predictions for multi-expert consensus, (d) corresponding uncertainty maps (brighter = higher unncertainty).

weights and fine-tuned end-to-end with all parameters trainable. Both datasets are resized to 128×128 resolution with [0,1] intensity normalization. The diffusion process employs $T = 250$ steps under cosine noise scheduling. In the training phase, we use the Adam optimizer (lr=10^{-4}) with batch size 12. To enhance generalization, data augmentation techniques such as random rotations, flips, and intensity scaling are applied. For LIDC-IDRI, the model is trained for 120,000 iterations in Stage I and 9,0000 iterations in Stage II. For NPC-170, both Stage I and Stage II are trained for 400,000 iterations.

3.2 Consensus-Driven Segmentation

In Stage I, we compare DiffOSeg with some representative probabilistic segmentation methods, including Prob. UNet [11], PhiSeg [2] and D-Persona [26]. We employ two complementary metrics: the Generalized Energy Distance (GED) to assess segmentation diversity, and Threshold-Aware Dice ($Dice_{soft}$) [26] to evaluate segmentation fidelity. For $Dice_{soft}$, the thresholds are set as $\{0.1, 0.3, 0.5, 0.7, 0.9\}$. We denote the metrics computed with n samples using a subscript, i.e., GED_n and D_n^s, and n is set to common values found in the literature [6, 26, 29].

As shown in Table 1, on LIDC-IDRI (4-fold cross-validation average), DiffOSeg achieves superior improvements in GED metrics across all sample sizes (10/30/50), outperforming the second-best method by 37.4% for GED_{10} (0.0915 vs. 0.1461), 43.8% for GED_{30} (0.0773 vs. 0.1375), and 45.2% for GED_{50} (0.0745

Table 2: Preference-driven segmentation performance comparison (all metrics shown as percentages). Results for DiffOSeg and PhiSeg are obtained from our implementation, and other scores are cited from [26]. Best in **bold**.

Method	LIDC-IDRI					NPC-170				
	$D_{A_1}(\uparrow)$	$D_{A_2}(\uparrow)$	$D_{A_3}(\uparrow)$	$D_{A_4}(\uparrow)$	$D_{mean}(\uparrow)$	$D_{A_1}(\uparrow)$	$D_{A_2}(\uparrow)$	$D_{A_3}(\uparrow)$	$D_{A_4}(\uparrow)$	$D_{mean}(\uparrow)$
CM-Global [24]	86.13	88.76	88.99	86.18	87.51	77.92	74.65	71.97	75.47	75.00
CM-Pixel [31]	85.99	88.81	89.31	86.77	87.72	78.97	73.93	71.65	75.12	74.92
TAB [15]	85.00	86.35	86.77	85.77	85.97	77.33	73.45	74.83	71.67	74.32
Pionono [22]	87.94	89.11	89.55	88.76	88.84	78.87	74.11	71.97	75.41	75.09
D-Persona (Stage II) [26]	88.54	89.50	90.03	88.60	89.17	79.78	**74.60**	75.22	**75.17**	76.19
DiffOSeg (Stage II)	**91.85**	**90.15**	**91.3**	**90.68**	**90.99**	**81.16**	74.03	**76.47**	75.07	**76.88**

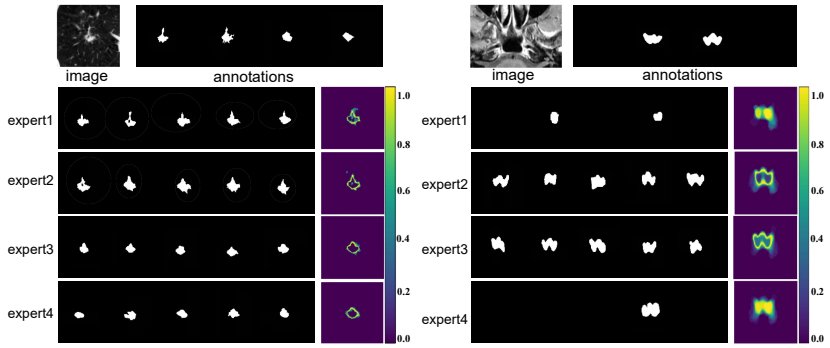

Fig. 3: Preference-driven segmentation visualization on both datasets. Row 1: original image & expert annotations (left to right: Experts 1–4). Rows 2-5: Sampled Expert-specific predictions & uncertainty map (computed by 100 samples).

vs. 0.1358). For $Dice_{soft}$, it also exhibits great performance across all sample sizes. On NPC-170, DiffOSeg reduces GED_{30} by 23.6% compared to D-Persona (0.1822 vs. 0.2385), though it slightly underperforms D-Persona on D^s_{30}. Visualization results shown in Fig. 2 demonstrate our model's ability to capture expert consensus while appropriately preserving diversity.

3.3 Preference-Driven Segmentation

In Stage II, we compare DiffOSeg with CM-Global [24], CM-Pixels [31], TAB [15], Pionono [22], and D-Persona. To evaluate performance in individualized preference modeling, we quantify segmentation fidelity for each expert using the Dice coefficient ($Dice$, denoted as D). This produces $N+1$ metrics: N expert-specific scores $(D_{A_1}, D_{A_2}, \ldots, D_{A_N})$ and the overall mean D_{mean} across all experts.

Table 2 shows our method achieves consistent improvements. On LIDC-IDRI, DiffOSeg surpasses all baselines for every expert $(A_1 - A_4)$, demonstrating robust superiority in multi-expert scenarios. On NPC-170, it achieves the best performance for experts (A_1, A_3) while marginally underperforming the strongest competitor (D-Persona) on others (A_2, A_4). Despite this, DiffOSeg achieves higher

Table 3: Consensus-driven segmentation performance of different annotation selection strategies on LIDC-IDRI. R: Random; A: Average; P: Our probabilistic consensus strategy. Best in **bold**.

Method	$GED_{10}(\downarrow)$	$GED_{30}(\downarrow)$	$GED_{50}(\downarrow)$	$D^s_{10}(\uparrow)$	$D^s_{30}(\uparrow)$	$D^s_{50}(\uparrow)$
R	0.0933	0.0774	**0.0742**	90.56	91.73	91.95
A	0.1194	0.1029	0.0990	89.57	90.43	90.69
P	**0.0915**	**0.0773**	0.0745	**91.21**	**92.20**	**92.37**

Table 4: Preference-driven segmentation performance of different expert identity encoding strategies on LIDC-IDRI. Best in **bold**.

Method	$D_{A_1}(\uparrow)$	$D_{A_2}(\uparrow)$	$D_{A_3}(\uparrow)$	$D_{A_4}(\uparrow)$	$D_{mean}(\uparrow)$
blind	90.35	88.50	89.73	89.10	89.42
concat.	90.96	89.16	90.19	89.64	89.99
ours	**91.41**	**89.46**	**90.78**	**90.09**	**90.44**

D_{mean} than all baselines. Fig. 3 qualitatively demonstrates our model's ability to capture expert-specific styles, showing five randomly sampled predictions and their corresponding uncertainty maps computed from 100 samples.

3.4 Ablation Study

Effect of Probabilistic Consensus Strategy. We compare our probabilistic consensus strategy (**P**) with two alternative approaches: random annotation selection (**R**) and average annotation aggregation (**A**). As shown in Table 3, our strategy demonstrates superior performance across both GED and $Dice_{soft}$ metrics on LIDC-IDRI. While achieving comparable GED_{30} and GED_{50} to R, our strategy shows clear advantages in segmentation fidelity ($Dice_{soft}$). By simulating a wide range of scenarios, the probabilistic consensus strategy enables robust integration of shared patterns while preserving annotation diversity.

Effect of Preference Prompts. We systematically evaluate three expert identity encoding strategies: 1) **blind**: randomly selects annotations from one expert per training step, without incorporating any expert identity information; 2) **concat.**: incorporates one-hot expert ID vectors via channel concatenation, excluding preference prompts subsequently; 3) **ours**: our proposed strategy, as described in Sec 2.2. As evidenced in Table 4, our prompt-based approach consistently outperforms baselines across all experts. The learned prompts demonstrate particular effectiveness in capturing nuanced annotation styles, without compromising overall segmentation fidelity.

4 Conclusion

We propose DiffOSeg, a novel diffusion-based framework that unifies consensus modeling and personalized preference adaptation for multi-rater medical image segmentation. The two-stage architecture first establishes population-level consensus via probabilistic consensus strategy, then adapts preference prompts encoding expert's individual styles. Extensive experiments on LIDC-IDRI and NPC-170 demonstrate that DiffOSeg achieves state-of-the-art performance in both tasks with parameter efficiency (13.1 M total vs. D-Persona's 28.46 M).

Future directions include few-shot adaptation for new unseen experts, and systematic modeling of annotation style variations to enhance clinical diagnostic applications.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Armato III, S.G., McLennan, G., Bidaut, L., McNitt-Gray, M.F., Meyer, C.R., Reeves, A.P., Zhao, B., Aberle, D.R., Henschke, C.I., Hoffman, E.A., et al.: The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. Medical physics **38**(2), 915–931 (2011)
2. Baumgartner, C.F., Tezcan, K.C., Chaitanya, K., Hötker, A.M., Muehlematter, U.J., Schawkat, K., Becker, A.S., Donati, O., Konukoglu, E.: Phiseg: Capturing uncertainty in medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22. pp. 119–127. Springer (2019)
3. Carass, A., Roy, S., Jog, A., Cuzzocreo, J.L., Magrath, E., Gherman, A., Button, J., Nguyen, J., Prados, F., Sudre, C.H., et al.: Longitudinal multiple sclerosis lesion segmentation: resource and challenge. NeuroImage **148**, 77–102 (2017)
4. Chen, S., Sun, P., Song, Y., Luo, P.: Diffusiondet: Diffusion model for object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 19830–19843 (2023)
5. Endo, M., Lin, P.P.: Surgical margins in the management of extremity soft tissue sarcoma. Chinese clinical oncology **7**(4), 37–37 (2018)
6. Gao, Z., Chen, Y., Zhang, C., He, X.: Modeling multimodal aleatoric uncertainty in segmentation with mixture of stochastic experts. arXiv preprint arXiv:2212.07328 (2022)
7. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
8. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
9. Hoogeboom, E., Nielsen, D., Jaini, P., Forré, P., Welling, M.: Argmax flows and multinomial diffusion: Learning categorical distributions. Advances in neural information processing systems **34**, 12454–12465 (2021)
10. Jia, M., Tang, L., Chen, B.C., Cardie, C., Belongie, S., Hariharan, B., Lim, S.N.: Visual prompt tuning. In: European conference on computer vision. pp. 709–727. Springer (2022)
11. Kohl, S., Romera-Paredes, B., Meyer, C., De Fauw, J., Ledsam, J.R., Maier-Hein, K., Eslami, S., Jimenez Rezende, D., Ronneberger, O.: A probabilistic u-net for segmentation of ambiguous images. Advances in neural information processing systems **31** (2018)
12. Kotelnikov, A., Baranchuk, D., Rubachev, I., Babenko, A.: Tabddpm: Modelling tabular data with diffusion models. In: International Conference on Machine Learning. pp. 17564–17579. PMLR (2023)

13. Li, X.L., Liang, P.: Prefix-tuning: Optimizing continuous prompts for generation. arXiv preprint arXiv:2101.00190 (2021)
14. Liao, Z., Hu, S., Xie, Y., Xia, Y.: Transformer-based annotation bias-aware medical image segmentation. In: International conference on medical image computing and computer-assisted intervention. pp. 24–34. Springer (2023)
15. Liao, Z., Hu, S., Xie, Y., Xia, Y.: Transformer-based annotation bias-aware medical image segmentation. In: International conference on medical image computing and computer-assisted intervention. pp. 24–34. Springer (2023)
16. Liao, Z., Hu, S., Xie, Y., Xia, Y.: Modeling annotator preference and stochastic annotation error for medical image segmentation. Medical Image Analysis **92**, 103028 (2024)
17. Luijnenburg, S.E., Robbers-Visser, D., Moelker, A., Vliegen, H.W., Mulder, B.J., Helbing, W.A.: Intra-observer and interobserver variability of biventricular function, volumes and mass in patients with congenital heart disease measured by cmr imaging. The international journal of cardiovascular imaging **26**, 57–64 (2010)
18. Luo, X., Wang, H., Xu, J., Li, L., Zhao, Y., He, Y., Huang, H., Xiao, J., Song, T., Zhang, S., et al.: Generalizable magnetic resonance imaging-based nasopharyngeal carcinoma delineation: Bridging gaps across multiple centers and raters with active learning. International Journal of Radiation Oncology* Biology* Physics (2024)
19. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**(1), 654 (2024)
20. Potlapalli, V., Zamir, S.W., Khan, S.H., Shahbaz Khan, F.: Promptir: Prompting for all-in-one image restoration. Advances in Neural Information Processing Systems **36**, 71275–71293 (2023)
21. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. pp. 234–241. Springer (2015)
22. Schmidt, A., Morales-Alvarez, P., Molina, R.: Probabilistic modeling of inter-and intra-observer variability in medical image segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 21097–21106 (2023)
23. Steenbakkers, R.J., Duppen, J.C., Fitton, I., Deurloo, K.E., Zijp, L., Uitterhoeve, A.L., Rodrigus, P.T., Kramer, G.W., Bussink, J., De Jaeger, K., et al.: Observer variation in target volume delineation of lung cancer related to radiation oncologist–computer interaction: a 'big brother'evaluation. Radiotherapy and Oncology **77**(2), 182–190 (2005)
24. Tanno, R., Saeedi, A., Sankaranarayanan, S., Alexander, D.C., Silberman, N.: Learning from noisy labels by regularized estimation of annotator confusion. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11244–11253 (2019)
25. Warfield, S.K., Zou, K.H., Wells, W.M.: Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation. IEEE transactions on medical imaging **23**(7), 903–921 (2004)
26. Wu, Y., Luo, X., Xu, Z., Guo, X., Ju, L., Ge, Z., Liao, W., Cai, J.: Diversified and personalized multi-rater medical image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11470–11479 (2024)
27. Wu, Y., Xie, Y., Luo, X., Wu, Q., Cai, J.: Dataset, challenge, and evaluation for tumor segmentation variability. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 11302–11303 (2024)

28. Xin, B., Ye, M., Axel, L., Metaxas, D.N.: Fill the k-space and refine the image: Prompting for dynamic and multi-contrast mri reconstruction. In: International Workshop on Statistical Atlases and Computational Models of the Heart. pp. 261–273. Springer (2023)
29. Zbinden, L., Doorenbos, L., Pissas, T., Huber, A.T., Sznitman, R., Márquez-Neila, P.: Stochastic segmentation with conditional categorical diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1119–1129 (2023)
30. Zhang, J., Xie, Y., Xia, Y., Shen, C.: 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)
31. Zhang, L., Tanno, R., Xu, M.C., Jin, C., Jacob, J., Cicarrelli, O., Barkhof, F., Alexander, D.: Disentangling human error from ground truth in segmentation of medical images. Advances in Neural Information Processing Systems **33**, 15750–15762 (2020)