MedFlowSeg: Flow Matching for Medical Image Segmentation with Frequency-Aware Attention

Zhi Chen[1*] **Runze Hu**[2*] **Le Zhang**[1†]
[1]School of Engineering, University of Birmingham
[2]School of Public Health, Peking University
{zxc561@student.bham.ac.uk, 2110306221@stu.pku.edu.cn, l.zhang.16@bham.ac.uk}

Abstract

Flow matching has recently emerged as a principled framework for learning continuous-time transport maps, enabling efficient ODE-based sampling without relying on stochastic diffusion processes. While generative modeling has shown promise for medical image segmentation, particularly in capturing uncertainty and complex anatomical variability, existing approaches are predominantly based on diffusion models, which require iterative sampling and incur substantial computational overhead. In this work, we propose **MedFlowSeg**, a conditional flow matching framework that formulates medical image segmentation as learning a time-dependent vector field that transports a simple prior distribution to the target segmentation distribution. Compared to diffusion-based methods, our formulation enables more efficient inference through solving an ordinary differential equation, while preserving the flexibility of generative modeling. To effectively incorporate conditional information, we introduce a dual-conditioning mechanism. Specifically, we propose a *Dual-Branch Spatial Attention* (DB-SA) module to inject multi-frequency structural priors, and a *Frequency-Aware Attention* (FA-Attention) module to model interactions between spatial and spectral representations via discrepancy-aware fusion and time-dependent modulation. These components improve the alignment between noisy intermediate states and clean semantic features, leading to better structural consistency and boundary delineation. We conduct extensive experiments across multiple medical imaging modalities, where MedFlowSeg consistently outperforms prior state-of-the-art (SOTA) baselines, including diffusion-based and flow-based methods. *Code is available at* `https://github.com/zrs03/MedFlowSeg`.

1 Introduction

Medical image segmentation aims to partition medical images into meaningful anatomical regions, which plays a crucial role in many clinical applications such as diagnosis, treatment planning, and image-guided interventions. In recent years, deep learning-based approaches have achieved remarkable success in this field, including classical convolutional neural networks (CNNs) such as U-Net and its variants [37, 23], as well as transformer-based architectures that enhance global contextual modeling [7, 32, 15, 35]. More recently, generative models, particularly diffusion probabilistic models (DPMs) [13], have shown strong potential for medical image segmentation by modeling structural uncertainty and producing diverse predictions [33, 34, 36].

Despite their success, diffusion-based segmentation methods suffer from several limitations. First, they rely on iterative denoising processes with a fixed number of sampling steps, resulting in limited

*Equal contribution.
†Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

flexibility during inference. Second, these methods often struggle to effectively integrate conditional features with the denoising process, leading to suboptimal alignment between noisy intermediate representations and clean semantic priors. As a result, their performance is typically sensitive to specific task settings, achieving strong results on certain benchmarks while exhibiting limited generalization across different modalities and anatomical structures. Although recent works attempt to improve conditional integration, they still lack an effective mechanism to address this issue.

Motivated by these challenges, we explore an alternative generative paradigm for medical image segmentation based on flow matching. Compared with diffusion models, flow matching directly learns a continuous deterministic velocity field [19, 20], enabling more efficient generation. More importantly, the sampling process is inherently flexible and can be adaptively adjusted according to task requirements, allowing a better trade-off between efficiency and accuracy. Recent studies [4] have shown that adopting UNet-style backbones within the flow matching framework can already achieve competitive performance in medical image segmentation. However, existing flow-based methods still struggle to effectively leverage conditional information, particularly when aligning noisy intermediate representations with clean anatomical priors, and are often tailored to specific tasks [4, 22], limiting their generalization across diverse modalities and segmentation scenarios.

To address these issues, we propose **MedFlowSeg**, a novel flow matching framework for medical image segmentation. The core idea is to explicitly model the discrepancy between noisy intermediate flow states and clean conditional priors. To this end, we introduce a dual-conditioning architecture that decomposes conditional guidance into structural alignment and semantic alignment. For structural alignment, MedFlowSeg injects multi-frequency anatomical priors into shallow flow representations, enabling stable guidance for both coarse structures and fine boundaries. For semantic alignment, we model cross-domain interactions in both spatial and spectral spaces, allowing discrepancy-aware feature fusion and time-dependent modulation between noisy flow states and clean semantic features.

Our main contributions can be summarized as: (1) We propose a unified flow matching based framework for general medical image segmentation, enabling consistent modeling across diverse tasks and imaging modalities. **(2)** We are the first to integrate attention mechanisms into a flow-based generative model for general medical image segmentation, improving the effectiveness of conditional information modeling. **(3)** We design a dual-conditioning architecture, including a Dual-Branch Spatial Attention (DB-SA) module for multi-frequency structural guidance and a Frequency-Aware Attention (FA-Attention) mechanism for discrepancy-aware semantic alignment across spatial and spectral domains. **(4)** Extensive experiments on multiple benchmarks across different modalities demonstrate that MedFlowSeg consistently achieves state-of-the-art performance over CNN-based, transformer-based, diffusion-based, and flow-based methods.

2 Related Work

Deterministic Models for Medical Segmentation U-Net and its variants have long been the dominant architectures for medical image segmentation, owing to their effective encoder–decoder design and strong ability to preserve spatial details [23]. With the rise of vision transformers, recent works further enhance global context modeling by incorporating self-attention mechanisms, such as TransUNet, Swin-UNet, Swin-UNETR, and DS-TransUNet [7, 5, 11, 18]. However, medical imaging data often exhibit substantial variations in acquisition devices, noise levels, and pathological characteristics, resulting in a complex data distribution. Consequently, a single deterministic prediction may fail to capture the full range of plausible interpretations. Although these models achieve strong performance on specific benchmarks, their generalization across datasets and modalities remains limited. Moreover, the lack of explicit uncertainty modeling makes them less effective in handling irregular anatomical structures and ambiguous boundaries, leading to suboptimal results in challenging cases.

Generative Models for Medical Segmentation Recently, generative models have shown promising potential for medical image segmentation by reformulating it as a conditional generation problem. Diffusion-based methods, such as EnsemDiff [30], MedSegDiff [33], and MedSegDiff-V2 [34], leverage stochastic sampling to model uncertainty and improve predictions on ambiguous boundaries, but rely on fixed iterative denoising, limiting inference flexibility. Flow matching provides a more efficient alternative by learning a continuous velocity field for efficient ODE-based generation with deterministic trajectories given an initial condition, with recent works such as FlowSDF [4] and LatentFM [22] exploring this direction. However, existing methods still adopt simple conditioning

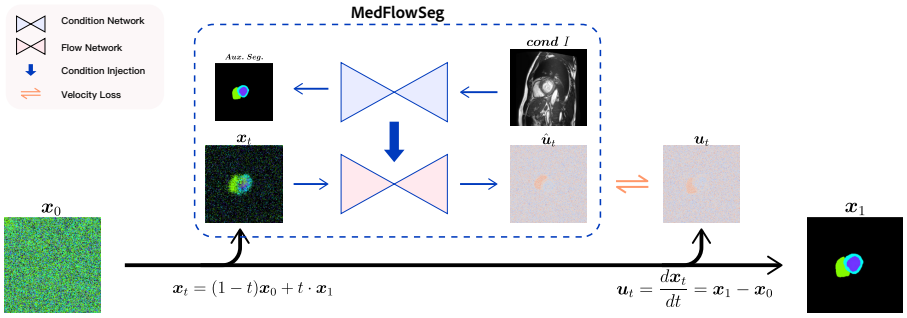

Figure 1: Overview of the proposed MedFlowSeg pipeline. The generative process is defined as a continuous trajectory from a random noise prior x_0 to the final segmentation label x_1. At a given time step t, the intermediate state x_t is generated. The raw medical image I is processed by a Condition Network to provide structural and semantic guidance to the Flow Network. The model takes x_t, I, and t as inputs to predict the image-guided velocity $\hat{u}_t$. During training, the model is optimized by minimizing the velocity loss between $\hat{u}_t$ and the target velocity $u_t = x_1 - x_0$, together with an auxiliary segmentation loss applied on the condition branch.

strategies and fail to effectively bridge the gap between noisy intermediate states and clean semantic priors, limiting their generalization. Therefore, a unified flow matching framework for general medical image segmentation is needed to enable consistent modeling across diverse tasks and modalities.

3 Methodology

In this section, we present the details of the proposed MedFlowSeg framework. We first briefly review the formulation of conditional flow matching. Then, we provide an overall pipeline of the architecture, highlighting how conditional features are injected into the flow network. Finally, we elaborate on our two core conditioning mechanisms: the Dual-Branch Spatial Attention (DB-SA) for structural condition, and the Frequency-Aware Attention (FA-Attention) for semantic condition.

3.1 Preliminaries

We have designed our model based on the conditional flow matching framework [19, 20]. Unlike conventional diffusion models that reverse a multi-step noise corruption process, flow matching generative models directly learn a continuous deterministic velocity field. As illustrated by the overall pipeline in Figure 1, MedFlowSeg establishes an end-to-end generation procedure. It defines a continuous trajectory starting from a random noise sample $x_0 \sim \mathcal{N}(0, 1)$ and progressively evolving toward the final segmentation label x_1, structurally guided by the raw image condition I. This probability path and its target velocity u_t at time step $t \in [0, 1]$ can be formulated as follows:

$$\begin{aligned} x_t &= (1-t)x_0 + t \cdot x_1, \\ u_t &= \frac{dx_t}{dt} = x_1 - x_0, \end{aligned} \tag{1}$$

where x_t represents the intermediate state along the path, and u_t denotes the constant target velocity pointing directly from the noise prior x_0 to the clean segmentation x_1. Within this pipeline, we utilize a neural network parameterized by θ to estimate the velocity. To achieve precise segmentation, conditional features extracted from the raw medical image I are injected into the flow network:

$$\hat{u}_t = F_\theta(x_t, I, t), \tag{2}$$

where the network F_θ takes the intermediate state x_t, the condition I, and the time step t as inputs to predict the image-guided velocity $\hat{u}_t$. During training, the predicted velocity $\hat{u}_t$ is constrained to match the target velocity u_t. Notably, instead of the conventional mean squared error, our pipeline specifically minimizes the expected L_1 error. This design choice is employed to robustly penalize velocity estimation errors and better preserve sharp structural boundaries in the segmentation:

$$\mathcal{L}_{\text{vel}} = \mathbb{E}_{x_0, x_1, t}\left[\|\hat{u}_t - u_t\|_1\right]. \tag{3}$$

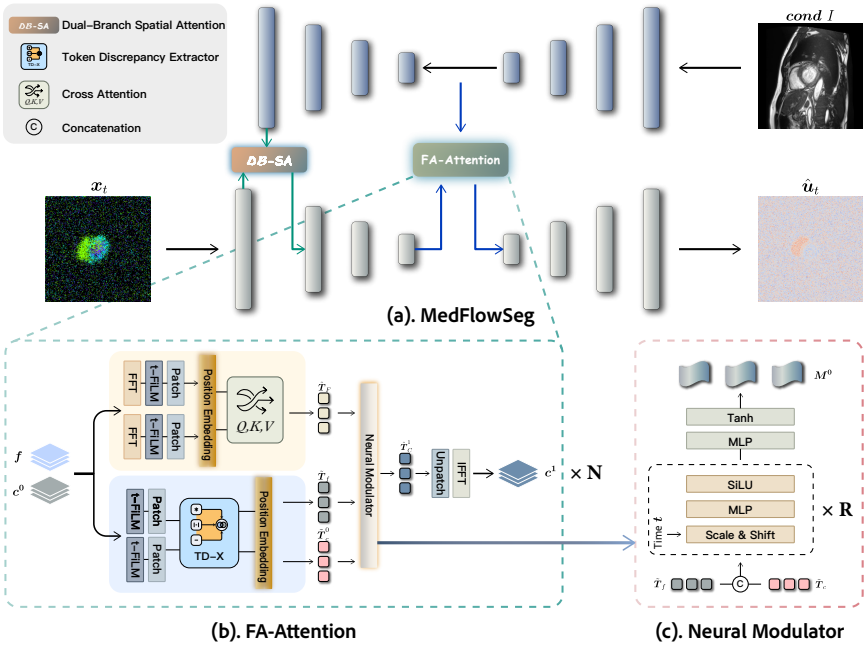

Figure 2: An illustration of MedFlowSeg, which starts from (a) an overview of the two-stream architecture, and continues with zoomed-in diagrams of individual modules, including (b) FA-Attention, and (c) Neural Modulator.

In addition to the velocity supervision, the conditional network is equipped with an auxiliary segmentation head to provide direct structural guidance. This auxiliary objective is denoted as $\mathcal{L}_{\text{aux}}$, which will be detailed in Section 3.2. At inference time, starting from a random initialization $x_0 \sim \mathcal{N}(0, 1)$, the pipeline recovers the clear mask step by step by solving an Ordinary Differential Equation (ODE) along the predicted velocity field, ultimately yielding the final segmentation result x_1.

3.2 Overall Architecture

The overall architecture of MedFlowSeg is shown in Figure 2(a). At each time step t of the flow matching trajectory, the intermediate state x_t is fed into a UNet-style Flow Network, while the raw medical image I is processed by a parallel standard UNet, termed the Condition Network.

To bridge the two streams, we introduce two complementary conditioning mechanisms, namely a structural condition and a semantic condition. The structural condition is imposed at the shallow stage of the Flow Network through Dual-Branch Spatial Attention (DB-SA). DB-SA takes the final high-resolution decoder feature of the Condition Network as input, which is transformed and injected into the first encoder stage of the Flow Network. By decomposing this structural prior into low-frequency and high-frequency components, DB-SA provides stable local guidance, including coarse anatomical layouts and fine boundary details.

The semantic condition is introduced at the bottleneck through Frequency-Aware Attention (FA-Attention). At this stage, deep condition features from the Condition Network are combined with the flow features to facilitate interaction between the clean semantic representations and the noisy intermediate state. By modeling their relationship in a shared representation space, FA-Attention enables effective information exchange between the two streams, thereby providing high-level semantic guidance for the generation process.

MedFlowSeg is trained in an end-to-end manner. The primary objective, $\mathcal{L}_{\text{vel}}$ in Equation (3), measures the L_1 distance between the predicted velocity $\hat{u}_t$ and the constant target velocity u_t. To further enhance the structural representation ability of the Condition Network, we attach an auxiliary segmentation head to its final decoder feature. Specifically, the auxiliary loss $\mathcal{L}_{\text{aux}}$ is defined as a

combination of a multi-class Dice loss and a Cross-Entropy loss:

$$\mathcal{L}_{\text{aux}} = \mathcal{L}_{\text{dice}} + \alpha \mathcal{L}_{\text{ce}}. \tag{4}$$

The overall training objective is then formulated by jointly optimizing the flow matching loss and the auxiliary segmentation loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{vel}} + \lambda \mathcal{L}_{\text{aux}}, \tag{5}$$

where λ and α are hyper-parameters that balance the auxiliary supervision and the cross-entropy term, respectively. In our implementation, we set $\lambda = 0.1$ and $\alpha = 10$.

3.3 Structural Condition Guidance

To provide stable structural guidance at the early stage of the flow branch, we introduce a Dual-Branch Spatial Attention (DB-SA) mechanism. Unlike directly using shallow encoder features from the condition branch, DB-SA takes the final high-resolution decoder feature of the Condition Network f_{cond}^{-1} as its input. We first extract a structural condition feature directly from this representation:

$$f_{str} = f_{cond}^{-1} * k_{Conv_{3\times3}}, \tag{6}$$

where $k_{Conv_{3\times3}}$ is a learnable convolution kernel used to capture local spatial context. Next, the structural condition feature is decomposed into low-frequency and high-frequency components:

$$\begin{aligned} f_{low} &= \left(f_{str} * k_{Gauss}^{l}\right) * k_{Conv_{3\times3}}^{l}, \\ f_{high} &= \left(f_{str} - \left(f_{str} * k_{Gauss}^{h}\right)\right) * k_{Conv_{3\times3}}^{h}, \end{aligned} \tag{7}$$

where k_{Gauss}^{l} and k_{Gauss}^{h} denote learnable Gaussian filters for the low- and high-frequency branches, respectively. To capture complementary structural priors at different spatial scales, the corresponding kernel sizes are set to 3 and 5. The low-frequency branch captures coarse anatomical structures, while the high-frequency branch emphasizes local boundary details. After obtaining the complementary features, they are concatenated and converted into an adaptive attention map:

$$g = \text{Sigmoid}\left(\text{Concat}(f_{high}, f_{low}) * k_{Conv_{1\times1}}^{g}\right), \tag{8}$$

where $\text{Concat}(\cdot)$ denotes channel-wise concatenation and $k_{Conv_{1\times1}}^{g}$ is a learnable 1×1 convolution kernel used to aggregate the fused features. Finally, the generated attention map is applied to the first encoder feature of the Flow Network:

$$f_{flow}^{\prime 0} = f_{flow}^{0} + g \cdot f_{flow}^{0}, \tag{9}$$

where $\cdot$ denotes element-wise multiplication. Through this process, DB-SA injects multi-frequency structural cues derived from the decoder output of the Condition Network into the shallow flow representation, thereby providing more stable anatomical guidance and sharper boundary awareness for subsequent generation.

3.4 Semantic Condition Alignment

The generative trajectory in flow matching evolves from noisy intermediate states to structured segmentation targets, while the Condition Network provides clean and static semantic priors extracted from the input image. This creates an intrinsic discrepancy between the two streams, particularly at the bottleneck stage where features mainly encode compressed global anatomical semantics. As the flow features are strongly timestep-dependent and often contaminated by unstable local responses during generation, directly coupling them with deterministic condition features in a shared representation space may lead to inconsistent semantic alignment and unreliable feature interaction.

To address this issue, we introduce Frequency-Aware Attention (FA-Attention), which refines the bottleneck condition feature c^0 using the current flow feature f as a dynamic reference. As illustrated in Figure 2(b), FA-Attention processes f and c^0 through two complementary pathways: a spectral branch for global semantic alignment and a spatial-token branch for discrepancy modeling.

In the spectral branch, both bottleneck features are first transformed by Fast Fourier Transform (FFT). The resulting frequency-domain features are then processed by FiLM modulation, patch embedding,

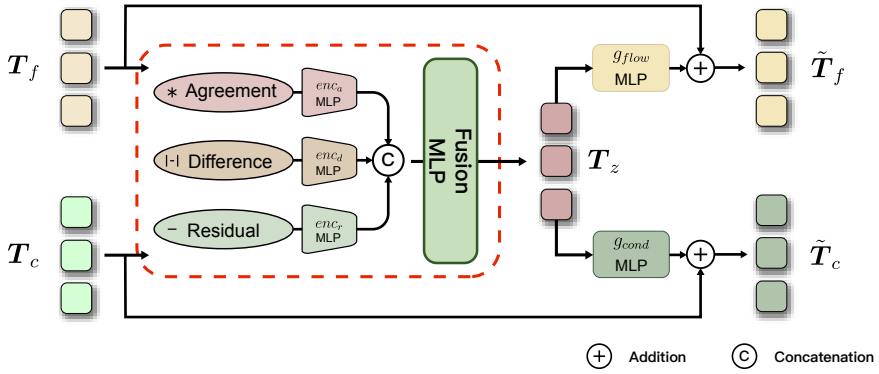

Figure 3: Architecture of the TD-X module. Given the patchified flow token T_f and condition token T_c, TD-X constructs three complementary discrepancy cues, namely the agreement token, the difference token, and the residual token. These cues are independently encoded and fused into a unified evidence token T_z, which is further used to recalibrate the original tokens, yielding $\hat{T}_f$ and $\hat{T}_c$.

and positional encoding to form token sequences $\hat{T}_f$ and $\hat{T}_c$. We perform cross-attention with $\hat{T}_c$ as the query and $\hat{T}_f$ as the key-value pairs:

$$\hat{T}_F = \mathrm{CrossAttn}(\hat{T}_c, \hat{T}_f). \tag{10}$$

Since FFT represents features from a global spectral view, this branch helps the condition tokens capture long-range dependencies and align coarse anatomical structures with current noisy flow state.

In parallel, the spatial-token branch operates on the original real-valued bottleneck features, preserving the relative spatial organization of feature tokens. After FiLM modulation and patch embedding, we obtain token representations T_f and T_c. To explicitly characterize their local mismatch, we introduce the Token Discrepancy Extractor (TD-X), as shown in Figure 3. TD-X constructs three complementary cues:

$$T_a = T_f \odot T_c, \quad T_d = |T_f - T_c|, \quad T_r = T_f - T_c, \tag{11}$$

where $\odot$ denotes element-wise multiplication. The agreement cue T_a captures co-activated responses, the difference cue T_d measures discrepancy magnitude, and the residual cue T_r preserves directional variation. These cues are independently encoded and fused into an evidence token T_z, which recalibrates the original tokens and produces discrepancy-aware spatial tokens $\tilde{T}_f$ and $\tilde{T}_c$.

Finally, a Neural Modulator integrates the spectral and spatial-token evidence. It takes $\hat{T}_F$, $\tilde{T}_f$, and $\tilde{T}_c$ as input, and applies $R = 2$ time-conditioned transformation blocks to generate a modulation mask M^0. The mask is applied to the frequency-aware condition tokens:

$$\hat{T}_C^1 = \hat{T}_F \odot (1 + M^0). \tag{12}$$

The modulated tokens are then transformed back through inverse patch embedding and Inverse FFT (IFFT), yielding the refined bottleneck condition feature c^1. In our implementation, FA-Attention is stacked $N = 4$ times. By combining global spectral alignment with spatial-token discrepancy modeling, FA-Attention provides a more stable semantic prior for subsequent vector-field estimation.

4 Experiments

4.1 Datasets

To evaluate our approach, we conduct experiments across five distinct medical image segmentation datasets. The public ACDC dataset [3], annotated with three cardiac structures, is utilized to verify multi-class segmentation performance on cine-MRI. Additionally, to demonstrate the model's robustness across diverse imaging modalities, we employ four other public datasets: the BraTS-2021

Table 1: Quantitative results on the ACDC cardiac segmentation dataset. Dice and IoU scores are reported for individual cardiac structures (RV, Myo, and LV), alongside the overall average metrics including Dice, IoU, and HD95. ↑/↓ indicates that higher/lower values represent better performance.

Category	Methods	Average			RV		Myo		LV	
		Dice ↑	IoU ↑	HD95 ↓	Dice ↑	IoU ↑	Dice ↑	IoU ↑	Dice ↑	IoU ↑
Deterministic Segmentation Methods	UNet	87.55	78.37	9.58	87.10	77.20	80.63	67.58	94.92	90.33
	nnUNet	90.29	82.35	7.49	88.92	80.05	89.13	80.38	92.84	86.63
	TransUNet	89.71	81.56	8.50	86.67	76.47	87.27	77.41	95.18	90.80
	SwinUNet	88.07	78.94	9.03	85.77	75.09	84.42	73.04	94.03	88.73
	MT-UNet	90.43	82.97	6.69	86.64	76.44	89.04	81.70	95.62	90.77
Generative Segmentation Methods	EnsemDiff	92.76	86.61	6.25	94.52	89.61	89.03	80.23	94.73	89.99
	SegDiff	91.80	84.90	6.84	91.68	84.64	89.34	80.73	94.37	89.34
	MedSegDiff	91.89	85.12	7.16	92.55	86.13	87.92	78.44	93.19	87.25
	MedSegDiff-V2	92.68	86.39	6.03	94.07	88.80	90.18	82.11	93.79	88.30
	FlowSDF	91.43	84.25	6.38	91.28	83.95	88.71	79.71	94.30	89.21
Proposed	Aux. Objective	90.61	82.43	7.55	89.23	80.66	88.25	78.13	94.34	88.51
	MedFlowSeg	**93.57**	**87.73**	**5.27**	**94.68**	**89.90**	**90.32**	**82.35**	**95.72**	**90.97**

dataset [2] for brain tumor segmentation in MRI, REFUGE-2 [10] for optic disc and cup segmentation in fundus images, the GlaS dataset [24] for gland segmentation in histology, and the CAMUS dataset [17] for echocardiographic segmentation in ultrasound images. For 3D datasets, all metrics are computed per patient after reconstructing volumes from slice-wise predictions.

4.2 Implementation Details

All models were developed in PyTorch Lightning and trained and evaluated on four NVIDIA A100 GPUs, each with 40 GB of VRAM. All images were uniformly resized to a resolution of 256×256 pixels. For grayscale medical images such as MRI and ultrasound, we used a single-channel input. For color images such as fundus and histology images, the RGB format was preserved as three-channel input. The backbone of both the Flow Network and the Condition Network follows a four-level UNet-style encoder–decoder design, with channel dimensions set to $\{64, 128, 256, 256\}$. Each level contains two residual blocks, and the time embedding dimension is set to 256 for the time-conditioned flow branch. We optimized the networks end-to-end using AdamW [21] with a batch size of 48 and an initial learning rate of 1×10^{-4}. During training, we maintained an exponential moving average (EMA) of model parameters with a decay rate of 0.9999. At inference time, we used 50 sampling steps by default, which is much fewer than the 100 sampling steps in MedSegDiff-V2 [34] and FlowSDF [4]. For each case, the model was run 10 times to generate an ensemble of segmentation samples, which were subsequently fused using the STAPLE algorithm [29]. Segmentation performance was assessed using Dice, IoU, and HD95 metrics. Higher Dice and IoU values, together with lower HD95 values, indicate better agreement with the ground truth.

4.3 Comparison with SOTA Methods

Evaluation on multi-class medical image segmentation To verify the segmentation performance on cardiac structures, we evaluate MedFlowSeg on the ACDC dataset [3] and compare it with a wide range of representative methods, including CNN-based models (U-Net [23], nnUNet [14]), transformer-based models (TransUNet [7], Swin-Unet [5], MT-UNet [26]), as well as generative approaches (EnsemDiff [31], SegDiff [1], MedSegDiff [33], MedSegDiff-V2 [34]) and the flow-based method FlowSDF [4]. All baselines were retrained using the same preprocessing pipeline, input resolution, data augmentation strategy, and train/validation/test splits as our method to ensure a fair comparison. The quantitative results are shown in Table 1.

From Table 1, we observe that segmentation performance is not solely determined by architectural complexity or the generative paradigm. Transformer-based models (e.g., Swin-Unet) do not consistently outperform strong CNN baselines like nnUNet, and early diffusion methods (e.g., EnsemDiff, SegDiff) also fail to reliably surpass deterministic approaches. Notably, the auxiliary objective, despite being based on U-Net, achieves competitive segmentation performance under joint training,

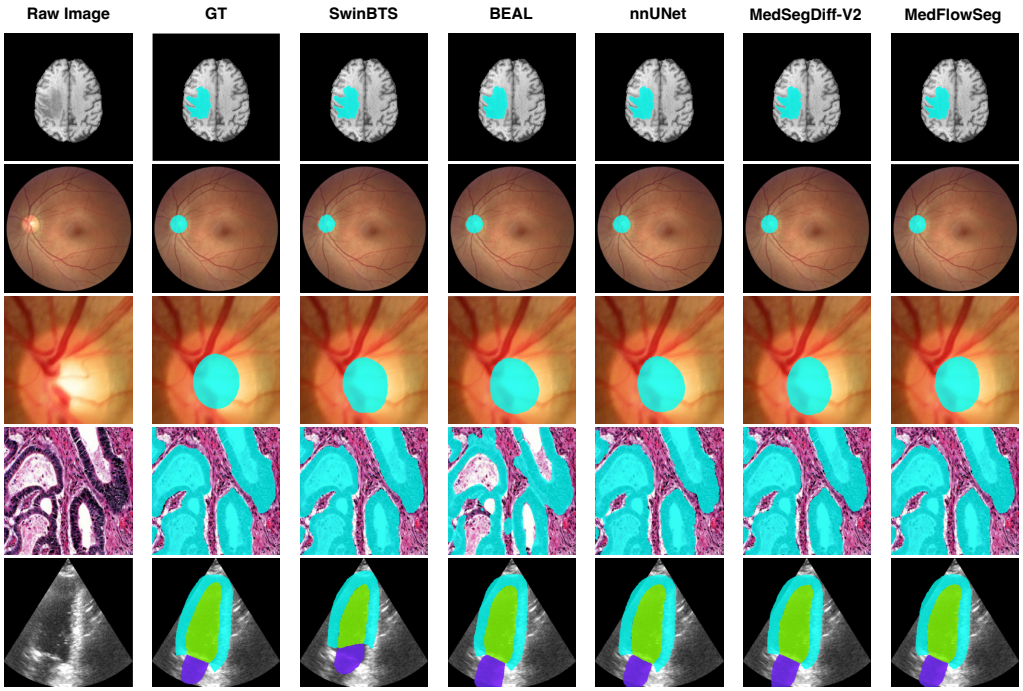

Figure 4: Visual comparison of our method against the representative baselines presented in Table 2. The rows, from top to bottom, display brain tumor segmentation, optic disc segmentation, optic cup segmentation, gland segmentation and echocardiographic segmentation, respectively.

demonstrating the effectiveness of auxiliary supervision. Although more advanced designs such as MedSegDiff-V2 improve performance, results remain sensitive to how conditional information is integrated. In contrast, MedFlowSeg achieves consistent improvements across all cardiac structures, performing well on both large and thin regions. This indicates that effective conditional modeling, rather than architecture alone, is key to achieving accurate and robust segmentation.

Evaluation on multi-modality medical image segmentation We further evaluate MedFlowSeg on multiple segmentation tasks with diverse imaging modalities. The results are reported in Table 2, where we compare with task-specific models such as TransBTS [28], SwinBTS [16], ResUNet [9], BEAL [27], DCAN [6], TA-Net [25], UltraUNet [17], and ECHO-SegNet [8], as well as general-purpose segmentation models like UNetr [12] and Swin-UNetr [11]. All baselines were retrained using the same preprocessing pipeline, input resolution, data augmentation strategy, and train/validation/test splits as our method to ensure a fair comparison.

These results show that MedFlowSeg generalizes well across diverse modalities (MRI, fundus, histology, ultrasound) despite large variations in appearance, noise, and anatomy. Task-specific methods perform strongly on their target datasets but lack cross-domain robustness, while generative models are sensitive to how conditional information is integrated, leading to inconsistent gains. In contrast, MedFlowSeg achieves stable improvements across all modalities. This robustness stems from its dual-conditioning design: DB-SA introduces multi-frequency structural priors early on, preserving both global layouts and fine boundaries, while FA-Attention aligns noisy intermediate states with clean features in spatial and frequency domains. Together, these components improve region consistency and boundary accuracy. As shown in Figure 4, MedFlowSeg produces more coherent segmentations, particularly in challenging regions.

4.4 Analysis and Discussion

Ablation Study We conduct an ablation study to verify the effectiveness of each component in the dual-conditioning framework, as summarized in Table 3. Starting from the backbone without the proposed conditioning modules, adding standard Spatial Attention (SA) brings only marginal improvement, indicating its limited ability to handle noisy flow features. In contrast, the proposed DB-

Table 2: Comparison of MedFlowSeg against SOTA segmentation models across multiple image modalities. Methods explicitly proposed for a given task are highlighted with a grey background.

Task	Methods	BraTs-2021			REFUGE2-Disc		REFUGE2-Cup		GlaS		CAMUS	
		Dice	IoU	HD95	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU
Brain Tumor	TransBTS	87.6	78.4	12.44	94.1	87.2	85.4	75.7	86.6	76.4	88.2	79.4
	SwinBTS	88.7	81.2	10.03	95.2	87.7	85.7	75.9	84.3	72.9	86.9	79.0
Optic Disc/Cup	ResUNet	78.4	71.3	18.71	92.9	85.5	80.1	72.3	82.1	69.6	87.9	82.7
	BEAL	78.8	71.7	18.53	93.7	86.1	83.5	74.1	84.4	73.0	89.1	82.8
Histology Gland	DCAN	81.0	68.1	13.86	92.3	85.7	82.1	69.6	91.2	83.8	85.9	80.3
	TA-Net	79.3	65.7	15.62	93.1	87.1	82.9	70.8	90.2	82.2	86.5	81.5
Cardiac Ultrasound	UltraUNet	84.7	76.9	13.75	91.2	82.3	83.6	74.2	86.7	76.5	90.8	83.1
	ECHO-SegNet	83.2	71.2	14.79	92.9	86.7	83.5	71.7	85.2	74.2	**91.6**	84.5
Det. Seg.	nnUNet	88.5	80.6	11.20	94.7	87.3	84.9	75.1	89.1	80.3	89.8	82.4
	TransUNet	86.6	79.0	13.74	95.0	87.7	85.6	75.9	84.6	73.3	86.7	79.2
	UNetr	87.3	80.6	12.81	94.9	87.5	83.2	73.3	89.0	80.1	86.2	80.1
	Swin-UNetr	88.4	81.8	11.36	95.3	87.9	84.3	74.5	87.8	78.3	89.9	80.3
Gen. Seg.	EnsemDiff	88.7	80.9	10.85	94.3	87.8	84.2	74.4	87.4	77.6	87.1	80.4
	SegDiff	85.7	77.0	14.31	92.6	85.2	82.5	71.9	87.2	77.4	88.0	80.7
	MedSegDiff	88.9	81.2	10.41	95.1	87.6	85.9	76.2	89.4	80.8	87.1	82.5
	MedSegDiff-V2	90.8	83.4	7.53	**96.7**	88.9	87.9	**80.3**	90.7	82.9	87.6	83.9
	FlowSDF	86.1	75.6	12.39	92.7	86.4	85.7	75.0	89.9	81.7	87.5	82.8
Proposed	**MedFlowSeg**	**91.2**	**84.5**	**6.76**	95.8	**90.3**	**88.6**	80.1	**92.8**	**86.8**	91.0	**84.7**

Table 3: Ablation study on the effectiveness of each component in the MedFlowSeg pipeline.

Structural Condition		Semantic Condition		Datasets (Dice %)		
SA	DB-SA	FA-Attention (w/o Modulator)	Neural Modulator	Brain Tumor	Optic Cup	Histology Gland
–	–	–	–	87.6	84.2	87.1
✓	–	–	–	88.5	85.0	86.9
–	✓	–	–	89.0	86.4	90.3
–	✓	✓	–	90.7	87.9	91.5
–	✓	✓	✓	**91.2**	**88.6**	**92.8**

SA consistently improves performance across all datasets by introducing multi-frequency structural priors into shallow flow features. For semantic conditioning, FA-Attention further enhances the results by reducing the discrepancy between noisy flow states and clean condition features. Finally, incorporating the Neural Modulator achieves the best performance.

Analysis of Uncertainty. In Table 4, we assess performance stability using the per-image standard deviation (SD) of Dice scores across 10 stochastic predictions against a fixed reference, averaged over the test set, and distribution alignment using Generalized Energy Distance (GED). MedFlowSeg achieves the lowest SD and highest Dice, indicating more stable segmentation performance and higher accuracy than the compared methods. It also achieves lower GED than all compared diffusion-based methods and FlowSDF. The individual-module variants exhibit different trade-offs: DB-SA alone provides greater performance stability, whereas FA-Attention alone achieves better distribution alignment. Combining both modules further improves stability and accuracy while maintaining competitive distribution alignment, suggesting complementary benefits from DB-SA and FA-Attention. Although its GED is slightly higher than that of FA-Attention alone, the combined model offers a favorable overall balance across the three metrics. These results suggest that improved performance stability can be achieved alongside high accuracy and close alignment with the reference distribution.

Table 4: Comparison of model complexity and generated sample uncertainty. The number of parameters is reported in millions (M), and Dice scores are expressed as percentages (%).

Methods	Params (M) ↓	GFLOPs ↓	SD ↓	GED ↓	Dice (%) ↑
EnsemDiff	53	2203	6.5	28.9	84.2
SegDiff	53	2399	10.5	26.4	82.5
MedSegDiff	85	1770	8.9	27.9	85.9
MedSegDiff-V2	137	983	6.3	23.5	87.9
FlowSDF	**42**	1997	7.7	36.5	85.7
bone + DB-SA	–	–	6.2	22.4	85.1
bone + FA-A.	–	–	7.1	**21.4**	87.4
MedFlowSeg	111	**789**	**5.6**	21.9	**88.6**

Table 5: Robustness analysis under different input resolutions on the ACDC dataset. Images are first downsampled and then resized back to 256×256 for inference.

Scale	Resolution	Dice (%) ↑	IoU (%) ↑	HD95 ↓
$1.0\times$	256×256	93.57	87.73	5.27
$0.75\times$	192×192	91.89	85.17	7.43
$0.5\times$	128×128	90.10	83.26	7.96
$0.25\times$	64×64	86.72	81.59	10.37

Model Efficiency and Complexity In Table 4, we further compare model complexity and computational cost. GFLOPs measure the cost of processing a single 256×256 image until convergence, defined as when the variance over the last ten samples is below 0.1%. MedFlowSeg achieves the lowest GFLOPs while also delivering the best segmentation performance. In contrast, diffusion-based methods require many sampling steps, leading to higher cost. Parameter count is also not indicative of efficiency; for example, MedSegDiff-V2 has higher GFLOPs due to iterative sampling. This underscores that MedFlowSeg's efficiency stems not only from its ODE-based formulation without stochastic denoising, but also from more effective utilization of conditional information, enabling faster convergence with fewer sampling steps and a smaller ensemble of segmentation samples.

Analysis of Resolution Robustness We further evaluate the robustness of MedFlowSeg under different input resolutions on the ACDC dataset. Specifically, the original MR images are first downsampled to lower resolutions and then resized back to 256×256 for inference. As summarized in Table 5, MedFlowSeg maintains relatively stable segmentation performance despite the progressive loss of fine anatomical details. Although performance gradually decreases as the resolution becomes lower, the degradation remains moderate even under severe downsampling. This suggests that the proposed dual-conditioning mechanism effectively preserves both global anatomical structure and local semantic consistency under resolution degradation. In particular, DB-SA provides stable structural priors across scales, while FA-Attention helps align degraded image features with intermediate flow representations, leading to robust segmentation performance under varying resolutions.

5 Conclusion

In this paper, we propose MedFlowSeg, a novel generative framework that formulates medical image segmentation as a conditional flow matching problem.To bridge the gap between noisy flow states and clean priors, we adopt a two-stream conditioning architecture with DB-SA for spatial guidance and FA-Attention for spectral alignment. Comprehensive experiments across five diverse medical imaging datasets demonstrate that MedFlowSeg significantly outperforms existing state-of-the-art models. By shifting the generative paradigm from multi-step diffusion to highly efficient, flexible flow matching, we believe MedFlowSeg provides a highly accurate and scalable solution that will serve as a strong foundation for future generative medical image analysis.

Acknowledgments and Disclosure of Funding

This study was supported by the High-Performance Computing Platform of Peking University.

The computations described in this research were performed using the Baskerville Tier 2 HPC service (https://www.baskerville.ac.uk/). Baskerville was funded by the EPSRC and UKRI through the World Class Labs scheme (EP/T022221/1) and the Digital Research Infrastructure programme (EP/W032244/1) and is operated by Advanced Research Computing at the University of Birmingham.

References

[1] T. Amit, E. Nachmani, T. Shaharbany, and L. Wolf. Segdiff: Image segmentation with diffusion probabilistic models. *arXiv preprint arXiv:2112.00390*, 2021.

[2] U. Baid, S. Ghodasara, S. Mohan, M. Bilello, E. Calabrese, E. Colak, K. Farahani, J. Kalpathy-Cramer, F. C. Kitamura, S. Pati, et al. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv preprint arXiv:2107.02314*, 2021.

[3] O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P.-A. Heng, I. Cetin, K. Lekadir, O. Camara, M. A. G. Ballester, et al. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE transactions on medical imaging*, 37(11):2514–2525, 2018.

[4] L. Bogensperger, D. Narnhofer, A. Falk, K. Schindler, and T. Pock. Flowsdf: Flow matching for medical image segmentation using distance transforms. *International Journal of Computer Vision*, 2025.

[5] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. *arXiv preprint arXiv:2105.05537*, 2021.

[6] H. Chen, X. Qi, L. Yu, and P.-A. Heng. Dcan: deep contour-aware networks for accurate gland segmentation. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2487–2496, 2016.

[7] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

[8] P. Dhivya, M. Shobana, N. Kumar, et al. Echo-segnet framework for accurate 2d echocardiographic image segmentation using the camus dataset. In *2025 International Conference on Next Generation Computing Systems (ICNGCS)*, pages 1–8. IEEE, 2025.

[9] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu. Resunet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 162:94–114, 2020.

[10] H. Fang, F. Li, J. Wu, H. Fu, X. Sun, J. Son, S. Yu, M. Zhang, C. Yuan, C. Bian, et al. Refuge2 challenge: A treasure trove for multi-dimension analysis and evaluation in glaucoma screening. *arXiv preprint arXiv:2202.08994*, 2022.

[11] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. Roth, and D. Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. *arXiv preprint arXiv:2201.01266*, 2022.

[12] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 574–584, 2022.

[13] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[14] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021.

[15] W. Ji, S. Yu, J. Wu, K. Ma, C. Bian, Q. Bi, J. Li, H. Liu, L. Cheng, and Y. Zheng. Learning calibrated medical image segmentation via multi-rater agreement modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12341–12351, 2021.

[16] Y. Jiang, Y. Zhang, X. Lin, J. Dong, T. Cheng, and J. Liang. Swinbts: A method for 3d multimodal brain tumor segmentation using swin transformer. *Brain sciences*, 12(6):797, 2022.

[17] S. Leclerc, E. Smistad, J. Pedrosa, A. Østvik, F. Cervenansky, F. Espinosa, T. Espeland, E. A. R. Berg, P.-M. Jodoin, T. Grenier, et al. Deep learning for segmentation using an open large-scale dataset in 2d echocardiography. *IEEE transactions on medical imaging*, 38(9):2198–2210, 2019.

[18] A. Lin, B. Chen, J. Xu, Z. Zhang, G. Lu, and D. Zhang. Ds-transunet: Dual swin transformer u-net for medical image segmentation. *IEEE Transactions on Instrumentation and Measurement*, 71:1–15, 2022.

[19] Y. Lipman, R. T. Chen, H. Ben-Hamu, M. Nickel, and M. Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

[20] X. Liu, C. Gong, and Q. Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

[21] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[22] H. T. Ngoc, H. A. N. Kim, T. N. Hai, and L. T. Quoc. Latentfm: A latent flow matching approach for generative medical image segmentation. *arXiv preprint arXiv:2512.04821*, 2025.

[23] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015*, pages 234–241. Springer, 2015.

[24] K. Sirinukunwattana, J. P. Pluim, H. Chen, X. Qi, P.-A. Heng, Y. B. Guo, L. Y. Wang, B. J. Matuszewski, E. Bruni, U. Sanchez, et al. Gland segmentation in colon histology images: The glas challenge contest. *Medical image analysis*, 35:489–502, 2017.

[25] H. Wang, M. Xian, and A. Vakanski. Ta-net: Topology-aware network for gland segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1556–1564, 2022.

[26] H. Wang, S. Xie, L. Lin, Y. Iwamoto, X.-H. Han, Y.-W. Chen, and R. Tong. Mixed transformer u-net for medical image segmentation. In *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 2390–2394. IEEE, 2022.

[27] S. Wang, L. Yu, K. Li, X. Yang, C.-W. Fu, and P.-A. Heng. Boundary and entropy-driven adversarial learning for fundus image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 102–110. Springer, 2019.

[28] W. Wang, C. Chen, M. Ding, H. Yu, S. Zha, and J. Li. Transbts: Multimodal brain tumor segmentation using transformer. In *International conference on medical image computing and computer-assisted intervention*, pages 109–119. Springer, 2021.

[29] S. K. Warfield, K. H. Zou, and W. M. Wells. 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.

[30] J. Wolleb, R. Sandkühler, F. Bieder, P. Valmaggia, and P. C. Cattin. Diffusion models for implicit image segmentation ensembles. *arXiv preprint arXiv:2112.03145*, 2021. doi: 10.48550/arXiv.2112.03145.

[31] J. Wolleb, R. Sandkühler, F. Bieder, P. Valmaggia, and P. C. Cattin. Diffusion models for implicit image segmentation ensembles. In *International conference on medical imaging with deep learning*, pages 1336–1348. PMLR, 2022.

[32] J. Wu, H. Fang, F. Shang, D. Yang, Z. Wang, J. Gao, Y. Yang, and Y. Xu. Seatrans: learning segmentation-assisted diagnosis model via transformer. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 677–687. Springer, 2022.

[33] J. Wu, R. Fu, H. Fang, Y. Zhang, Y. Yang, H. Xiong, H. Liu, and Y. Xu. Medsegdiff: Medical image segmentation with diffusion probabilistic model. In *Medical Imaging with Deep Learning*, volume 227 of *Proceedings of Machine Learning Research*, pages 1623–1639, 2024.

[34] J. Wu, W. Ji, H. Fu, M. Xu, Y. Jin, and Y. Xu. Medsegdiff-v2: Diffusion-based medical image segmentation with transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 6030–6038, 2024.

[35] L. Zhang, M. Pereañez, C. Bowles, S. Piechnik, S. Neubauer, S. Petersen, and A. Frangi. Missing slice imputation in population cmr imaging via conditional generative adversarial nets. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 651–659. Springer, 2019.

[36] L. Zhang, F. Wu, K. Bronik, and B. W. Papiez. Diffuseg: domain-driven diffusion for medical image segmentation. *IEEE Journal of Biomedical and Health Informatics*, 29(5):3619–3631, 2025.

[37] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang. Unet++: A nested u-net architecture for medical image segmentation. In *International workshop on deep learning in medical image analysis*, pages 3–11. Springer, 2018.

Appendix

Overview. In this supplementary, we present more results and details:

§A: More details on our method.

§B: Evaluation on more datasets.

§C: More empirical studies.

§D: Discussion of the strengths, limitations and broader impacts.

A More details on our method

Dataset Augmentation To improve data diversity and enhance generalisation, we apply an online data augmentation pipeline during training. All geometric transformations are jointly applied to the input images and corresponding masks to preserve spatial consistency. The augmentation strategy includes: **(1)** geometric transformations, including random horizontal/vertical flips, scaling within $\pm 10\%$, translations up to 8 pixels, and rotations within $\pm 30^{\circ}$; **(2)** intensity perturbations, including random brightness (± 0.08), contrast ($\pm 15\%$), gamma adjustment ($[1/1.2,\ 1.2]$), Gaussian blur ($p = 0.2$, $\sigma \leq 0.8$), and additive Gaussian noise ($\sigma = 0.02$). All augmentations are applied on-the-fly during training and are disabled during validation and testing.

Visual Comparison on ACDC Figure 5 presents qualitative results on the ACDC dataset with 50 sampling steps. Compared with representative baselines, our method achieves more accurate and robust segmentation results, particularly in challenging regions with ambiguous boundaries and low contrast. As shown in the figure, MedFlowSeg produces more precise and anatomically consistent delineation of cardiac structures, with clearer boundaries and reduced noise.

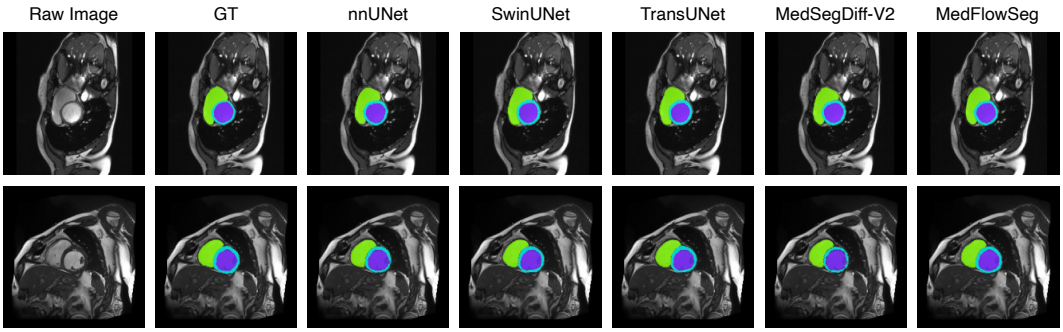

Figure 5: Visual comparison on the ACDC dataset. From left to right: raw image, ground truth, and segmentation results of different methods.

Figure 6 further analyzes the effect of the number of sampling steps. Even with a single step (N=1), our method already yields competitive segmentation results, demonstrating the effectiveness of the learned flow field. As the number of steps increases, the segmentation quality is progressively refined, leading to sharper boundaries and better alignment with the ground truth. Notably, MedFlowSeg quickly reaches near-optimal performance within a small number of sampling steps, indicating fast convergence of the generation process. This highlights that MedFlowSeg not only supports efficient one-step inference, but also benefits from iterative refinement when additional steps are applied.

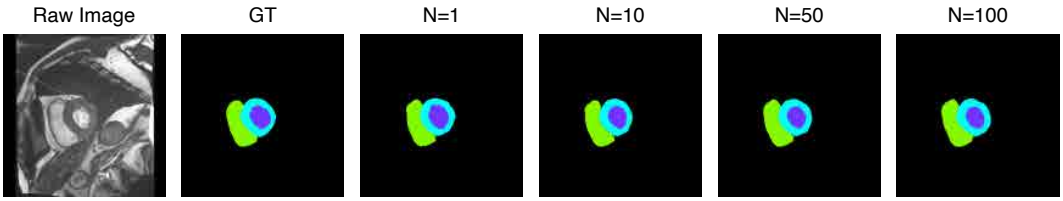

Figure 6: Effect of sampling steps on segmentation quality. From left to right: raw image, ground truth, and results with different numbers of sampling steps (N=1, 10, 50, 100), where N denotes the number of sampling steps.

B Evaluation on more datasets

Qualitative Evaluation on COVID-19 CT We further evaluate the generalization ability of MedFlowSeg on a COVID-19 CT dataset. As shown in Figure 7, MedFlowSeg is able to accurately capture infection regions with diverse shapes and appearances. Compared with the ground truth, the predicted masks exhibit better spatial coherence and more complete coverage of lesion areas, particularly in challenging cases with scattered or low-contrast infections. Notably, MedFlowSeg produces smoother and more anatomically consistent boundaries, while reducing fragmented predictions and missing regions. This demonstrates that the proposed dual-conditioning mechanism can effectively adapt to unseen data distributions and maintain robust segmentation performance beyond the training domain.

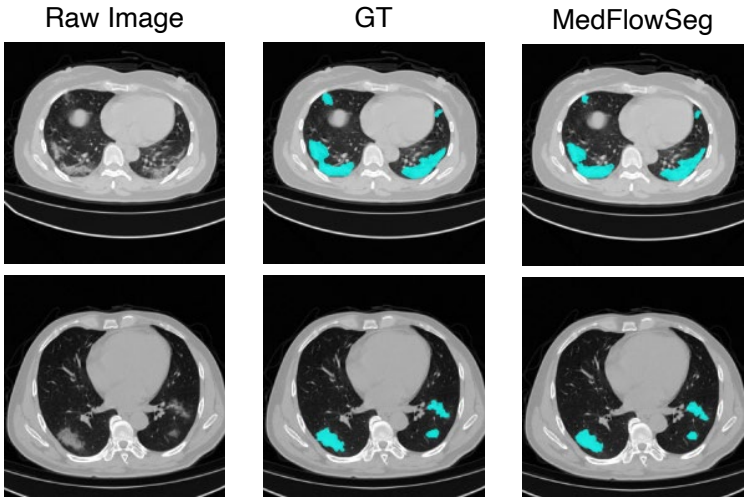

Figure 7: Qualitative comparison on a COVID-19 CT dataset. From left to right: raw image, ground truth, and the prediction of MedFlowSeg.

Qualitative Evaluation on MoNuSeg We further evaluate MedFlowSeg on the MoNuSeg dataset, which focuses on challenging nuclei segmentation in histopathology images. As shown in Figure 8, the dataset contains densely packed and highly overlapping nuclei with significant variations in shape, size, and staining appearance. MedFlowSeg demonstrates strong capability in capturing fine-grained structures and separating closely adjacent nuclei. Compared with the ground truth, the predicted masks exhibit improved instance separation and reduced merging errors, particularly in crowded regions. In addition, the model produces more regular and well-defined boundaries, effectively preserving the morphology of individual nuclei. These results highlight the effectiveness of MedFlowSeg in handling high-frequency structural details and complex spatial layouts, indicating its robustness across different medical imaging modalities.

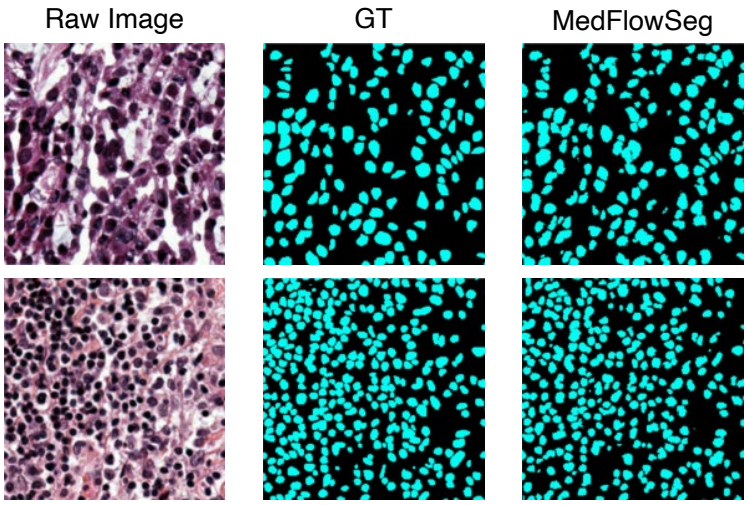

Figure 8: Qualitative comparison on MoNuSeg dataset. From left to right: raw image, ground truth, and the prediction of MedFlowSeg.

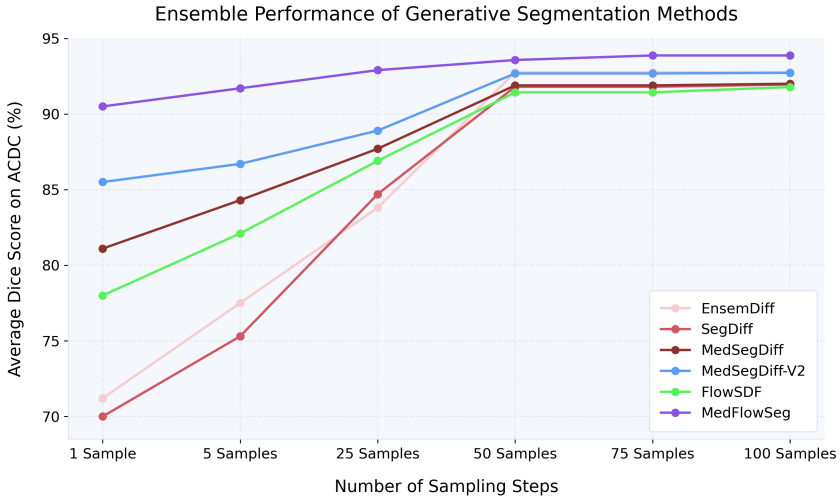

Figure 9: Comparison of the ensemble performance of generative segmentation methods. The average Dice score on the ACDC dataset is reported as the number of sampling steps increases.

C More empirical studies

Implicit Ensemble Effect We evaluate the implicit ensemble behavior of generative segmentation models on the ACDC dataset, as illustrated in Figure 9. The performance is measured by the average Dice score with increasing numbers of sampling runs. As shown in the figure, all methods benefit from ensemble sampling, with performance improving rapidly in the early stages and gradually saturating. MedFlowSeg consistently achieves the best performance across all sampling budgets. Notably, it starts from a significantly stronger baseline (over 90% Dice with a single sample), outperforming other methods by a clear margin. Moreover, MedFlowSeg converges faster than competing approaches. While other methods require a larger number of samples to reach stable performance, MedFlowSeg approaches its peak performance within fewer sampling steps and maintains a consistent advantage throughout the entire range. This indicates that the generated samples are both high-quality and stable, reducing the reliance on large ensemble sizes. These results suggest that MedFlowSeg provides a better initial estimation of the target distribution and produces more reliable predictions during sampling. As a result, it achieves a higher performance ceiling with improved efficiency, demonstrating the effectiveness of its design in stabilizing the generation process.

D Discussion

Strengths and Key Findings MedFlowSeg consistently improves segmentation accuracy, computational efficiency, and uncertainty modeling across datasets. Compared with diffusion-based methods, it achieves higher Dice scores with significantly fewer sampling steps, indicating that flow matching provides a more efficient approximation of the target distribution. Moreover, MedFlowSeg achieves a better balance between diversity and reliability, as evidenced by lower GED and higher CI, suggesting that the generated samples are both diverse and concentrated within plausible regions, unlike diffusion models that often trade stability for diversity.

Limitations Despite these advantages, MedFlowSeg still has limitations. The model contains a relatively large number of parameters due to its dual-branch and transformer-based design, which may limit deployment in resource-constrained settings. It also operates on 2D slices without modeling 3D spatial consistency, and its performance may degrade under severe domain shifts, suggesting the need for lightweight design, 3D extensions, and stronger domain generalization.

Broader Impact and Future Work MedFlowSeg provides an efficient framework for uncertainty-aware generative segmentation by combining flow modeling with structured feature learning. Its ability to produce reliable and diverse predictions is valuable for clinical applications. Future work will focus on improving efficiency via lightweight architectures and adaptive sampling, as well as extending to 3D modeling and enhancing robustness across domains.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of the work in Appendix Section D.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: The paper does not present formal theoretical theorems or proofs, but instead focuses on methodological design and empirical validation.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper provides detailed descriptions of datasets, architecture, training settings, loss functions, and evaluation protocols sufficient to reproduce the experiments (Section 3 and Section 4)

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: All datasets used in this work are publicly available. The code is publicly available via an anonymized repository.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: All experimental settings and procedures have been described in detail in Section 4 and Appendix Section A.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The paper evaluates stability via uncertainty metrics such as GED and CI in Section 4.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper provides details of compute resources, including GPU type, training setup, and inference sampling steps in Section 4.4.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We conform, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the broader impacts of our work in the Appendix Section D.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: Our work does not have the risk for misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All the assets used in our paper are publicly available for academic research.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We provide well-documented code with instructions for reproducibility.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: Our work does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: Our work does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method development of our work does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.