CHANNEL-SPATIAL GATED AND CONTOUR ALIGNED LATENT DIFFUSION FOR MEDICAL IMAGE SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion models have recently been employed for medical image segmentation due to their strong generative priors and ability to model rational anatomical structure consistent with real distribution. However, existing latent diffusion segmentation methods are largely adapted from natural image synthesis and remain constrained by parameter-heavy U-Net architectures, static skip connections, and insufficient contour-aware supervision. To align the latent diffusion paradigm with medical image segmentation tasks, we propose **GCLDS**, a channel-spatial **G**ated and **C**ontour Aligned **L**atent **D**iffusion **S**egmentation framework, with three bottom-up improvements at the architecture, feature, and loss levels. Our method redesigns the diffusion backbone with a compressed bottleneck to reduce parameter count and computational overhead without compromising segmentation accuracy. We design a channel-spatial disentangled gated skip connection to enable adaptive, context- and timestep-aware feature fusion during the denoising process. In addition, a contour alignment optimization loss is derived to enhance boundary localization and improve anatomical plausibility. Extensive experiments on multiple benchmarks demonstrate that GCLDS achieves superior segmentation accuracy and boundary delineation while maintaining efficient computation consumption.

1 INTRODUCTION

Medical image segmentation serves as a cornerstone of computer-aided diagnosis by accurately delineating anatomical structures and pathological regions, thereby supporting downstream clinical analysis. A wide range of deep learning models, including CNNs (Lei et al., 2025; Rahman & Marculescu, 2025; Ronneberger et al., 2015), Transformers (Cao et al., 2022; Chen et al., 2024a; Wang et al., 2025), and recently proposed architectures such as Mamba (Ruan et al., 2025; Zhu et al., 2024) and KAN (Li et al., 2025; Zhang et al., 2025b), have achieved remarkable success in this field, and remain the mainstream paradigm for medical image segmentation to date. These methods inherently formulate segmentation as a pixel-wise classification task. While this paradigm excels at optimizing local pixel-wise prediction accuracy, it focuses predominantly on individual pixel classification without explicit constraints on the structural rationality of the entire output segmentation mask. This inherent limitation may lead to unsatisfactory predictions for clinical applications, such as discontinuous target regions, blurred object boundaries, and outputs that do not align with the structural expectations of clinical diagnosis.

Diffusion models (DMs), as powerful generative frameworks originally designed for high-fidelity image synthesis (Ho et al., 2020), have recently been deployed for medical image segmentation to address the above issues (Ahn et al., 2025; Wu et al., 2024a; Xing et al., 2025). Benefiting from their strong generative priors, DMs can better model the inherent spatial distribution of anatomical structures and produce structurally coherent and clinically plausible segmentation masks. To overcome the inefficiency of iterative pixel-space diffusion, recent advances adopt latent diffusion models (LDMs) (Rombach et al., 2022) within an efficient single-step inference paradigm (Lin et al., 2024; Yang et al., 2026), achieving competitive performance and making them a promising solution for medical image segmentation.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Despite the success of efficient LDM segmentation pipelines, they inherit the paradigm designed for high-fidelity natural image synthesis, which may be misaligned with the requirements of medical image segmentation. This paradigm mismatch manifests in three limitations. First, the standard LDM U-Net backbone is originally designed for modeling intricate textures, complex semantics, and diverse visual distributions in image synthesis tasks. When adapted to medical image segmentation, it incurs substantial parameter overhead while providing limited performance benefits, as modeling structured anatomical mask priors requires less high-dimensional texture modeling than image synthesis. Second, vanilla skip connections perform feature fusion between encoder and decoder in a static, content-agnostic manner, disregarding the time-varying denoising context across different timesteps during training. This rigid design forces the decoder to passively aggregate encoder features without adaptively considering the noise level, encoder-provided information, and decoder needs, thereby introducing irrelevant noisy features and reducing the efficiency of structural information propagation. Finally, the training objectives of LDMs are native designs for natural image synthesis, acting as global pixel-wise reconstruction signals with uniform spatial weighting that don't prioritize boundary regions. The lack of explicit boundary supervision would lead to insufficient contour localization accuracy, which is significant for medical image segmentation tasks.

To mitigate this paradigm mismatch, we introduce Channel-Spatial **G**ated and **C**ontour Aligned **L**atent **D**iffusion **S**egmentation (**GCLDS**) for medical image segmentation. It unifies three complementary designs spanning architecture-level, feature-level, and loss-level optimizations, to systematically align the latent diffusion paradigm with medical image segmentation requirements. First, as architecture-level optimization, we tailor a lightweight bottleneck through channel compression, which effectively reduces computational overhead while preserving the intact encoder-decoder structure and their valuable pre-trained generative priors. Second, as feature-level optimization, we equip the U-Net skip connections with a time-aware, encoder-decoder guided disentangled channel-spatial gating mechanism for adaptive feature fusion. This mechanism dynamically modulates information propagation from encoder to decoder, and adaptively selects features across both channel and spatial dimensions tailored to each timestep of the denoising tasks during training. Finally, as loss-level optimization, we devise a latent-space contour alignment loss to enhance contour accuracy of segmentation. It enforces consistency between Sobel gradient responses of predicted and ground-truth latent features to deliver explicit supervision for fine-grained boundary refinement.

In summary, main contributions of this work are four-fold. (i) We propose a lightweight diffusion backbone redesign with targeted bottleneck compression at architecture level to improve parameter efficiency while maximizing the retention of pre-trained generative priors and maintaining segmentation performance. (ii) We design a time-aware, channel-spatial disentangled gating mechanism for skip connections at feature level, where channel modulation is governed by timestep embeddings and spatial weighting is jointly guided by encoder-decoder features, enabling adaptive multi-scale feature fusion across dynamic denoising contexts. (iii) We propose a targeted latent-space contour alignment loss to explicitly optimize contour accuracy at loss level, compensating for the limitations of native global reconstruction objectives adopted in LDMs. (iv) Extensive experiments on multiple medical segmentation benchmarks demonstrate that GCLDS achieves state-of-the-art segmentation performance and outperforms existing diffusion-based methods w.r.t. computational consumption.

2 Related Work

Diffusion and Latent Diffusion Models for Medical Image Segmentation. Diffusion models (DMs) (Ho et al., 2020; Song et al., 2021a;b) have emerged as dominant generative models in medical image analysis, supporting various downstream tasks including image-to-image translation (Kim & Park, 2024), medical image synthesis (Chen et al., 2024b; Li et al., 2026), and segmentation (Kazerouni et al., 2023; Rahman et al., 2023; Zhang et al., 2024). Different from conventional discriminative segmentation methods, DMs excel at modeling complex spatial distributions of segmentation masks. Several pixel-space diffusion segmentation methods have achieved substantial progress in boundary refinement, long-range dependency modeling, and uncertainty estimation, covering foundational frameworks such as MedSegDiff (Wu et al., 2024a) and its transformer-enhanced variant MedSegDiff-V2 (Wu et al., 2024b), as well as subsequent optimized models including MoDiff (Ahn et al., 2025), Diff-UNet (Xing et al., 2025), DiffTCL (Li et al., 2024), and DiffOSeg (Zhang et al., 2025a). Recent studies have explored improving diffusion models for 3D medical imaging (Du et al., 2026; Wu et al., 2026). However, such pixel-level diffusion paradigms suffers from

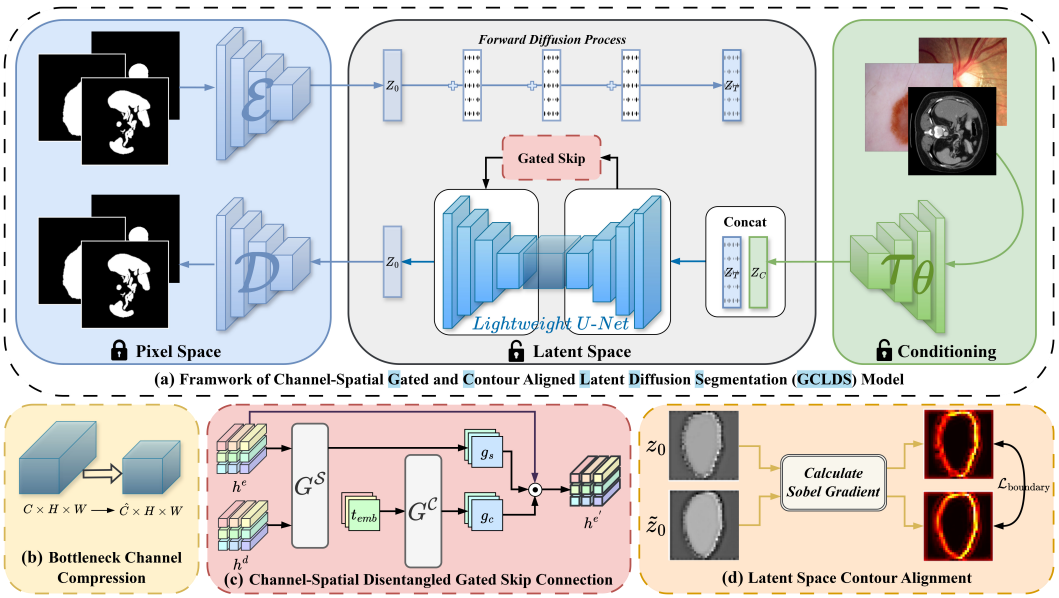

Figure 1: Overview of Channel-Spatial Gated and Contour Aligned Latent Diffusion Segmentation.

heavy computational costs due to iterative full-resolution denoising, limiting clinical deployment. To address this efficiency bottleneck, latent diffusion models (LDMs) (Rombach et al., 2022) perform denoising in compact VAE-encoded latent spaces (Kingma & Welling, 2013; Van Den Oord et al., 2017), effectively reducing computational consumption while preserving essential structural information. This efficient paradigm has been widely applied to medical segmentation, where pioneering work SDSeg (Lin et al., 2024) validates the effectiveness of latent-space segmentation, and follow-up methods including HiDiff (Chen et al., 2024c), MedSegLatDiff (Ngoc et al., 2025), and TSLDSeg (Yang et al., 2026) further improve structural preservation and semantic adaptation for medical data. Despite improved efficiency, existing LDM-based segmentation methods inherit architectures and training paradigms from natural image synthesis. These designs incur high computational costs, use static skip connections, and rely on uniform objectives without targeted contour supervision. These limitations motivate our proposed GCLDS framework.

3 Method

3.1 Preliminaries: Diffusion Models for Segmentation

In this work, we perform medical image segmentation following the latent diffusion model paradigm. Formally, given a segmentation mask x_0 and its corresponding image condition c (i.e., the medical image to be segmented), we first encode them into the latent space z using a VAE encoder $\mathcal{E}$ and an image encoder τ_θ. The forward diffusion process progressively adds Gaussian noise to the clean latent z_0 over T time steps:

$$z_t = \sqrt{\bar{\alpha}_t}\, z_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \tag{1}$$

where $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$ denotes the cumulative noise schedule. A U-Net backbone ϵ_θ is trained as the denoiser to predict the noise ϵ conditioned on the noisy latent z_t, timestep t, and image condition z_c.

Existing work (Lin et al., 2024) adopts a dual-loss training objective that supervises both noise prediction and latent reconstruction:

$$\mathcal{L}_{\text{base}} = \mathcal{L}_{\text{noise}} + \mathcal{L}_{\text{latent}}. \tag{2}$$

Both losses are implemented with mean absolute error:

$$\begin{aligned} \mathcal{L}_{\text{noise}} &= \mathbb{E}_{z_0,t,\epsilon} \left\| \epsilon - \epsilon_\theta\left(z_t, t, z_c\right) \right\|_1, \\ \mathcal{L}_{\text{latent}} &= \mathbb{E}_{z_0,t,\epsilon} \left\| z_0 - \hat{z}_0\left(z_t, t, z_c\right) \right\|_1, \end{aligned} \tag{3}$$

where $\|\cdot\|_1$ denotes the mean absolute error, and the clean latent $\hat{z}_0$ is recovered in closed form from the predicted noise:

$$\hat{z}_0 = \frac{z_t - \sqrt{1-\bar{\alpha}_t}\,\epsilon_\theta\left(z_t, t, z_c\right)}{\sqrt{\bar{\alpha}_t}}. \tag{4}$$

During inference, we follow the single-step sampling paradigm Lin et al. (2024) by default. The pipeline feeds a pure Gaussian noise latent $z_T \sim \mathcal{N}(0, I)$, the maximum time step T, and the latent image condition z_c into the U-Net to obtain the predicted noise $\hat{\epsilon}$. The clean latent $\hat{z}_0$ is reconstructed via Equation equation 4, then decoded to pixel-space segmentation mask $\hat{x}_0$ using the VAE decoder $\mathcal{D}$. Other mainstream diffusion samplers (e.g., DDIM Song et al. (2021a)) are also compatible with our framework.

To establish an effective medical image segmentation framework that aligns the native latent diffusion paradigm with medical image segmentation requirements, we propose the Channel-Spatial **G**ated and **C**ontour Aligned **L**atent **D**iffusion **S**egmentation (**GCLDS**) in this paper. As illustrated in Figure 1, our GCLDS framework consists of three bottom-up complementary components, covering architecture-, feature-, and loss-level optimizations. We first introduce an architecture-level lightweight bottleneck that improves parameter efficiency by reducing trainable parameters in the diffusion backbone while preserving the core encoder–decoder structure and segmentation capability. For the feature-level improvement, we equip skip connections with a channel-spatial disentangled gating mechanism. It dynamically adapts to the time-varying denoising context to improve the efficiency of multi-scale encoder-decoder feature fusion. Finally, to strengthen the clinically critical boundary supervision at the loss level, we add a latent-space loss that explicitly optimizes contour alignment. The three technical components are elaborated in the subsequent sections.

3.2 Bottleneck Channel Compression

As the architecture-level improvement of our GCLDS framework, we first propose a parameter-efficient bottleneck compression strategy. Standard diffusion U-Nets are originally designed for high-fidelity natural image synthesis and are typically equipped with sufficient capacity to model complex semantics, fine-grained textures, and diverse visual distributions. Such capacity inherently exceed the requirements of medical image segmentation, which focuses on predicting structured anatomical masks rather than generating detailed visual content. This difference in task requirements motivates us to explore moderate bottleneck compression to reduce computational overhead while preserving segmentation accuracy. Accordingly, we introduce a lightweight bottleneck to improve computational efficiency while maintaining segmentation accuracy.

A core design principle of our modification is to maximize retention of pre-trained natrual image generative priors, which enables effective knowledge transfer to medical data. To this end, we keep the vast majority of the encoder and decoder architectures intact. Since the encoder and decoder are tightly coupled via layer-wise skip connections, modifying any internal layer in either module would break the dimensional alignment of cross-scale features, disrupt the native skip connection pipeline, and invalidate the pre-trained weights of the entire backbone. We therefore target the bottleneck for compression. Modifications on it only require minimal dimensional adjustments to its two immediately adjacent layers of the encoder and decoder, with no changes to all other encoder and decoder layers. This brings negligible impact to the encoder-decoder architecture and pre-trained weight compatibility, while delivering maximum compression gains as the bottleneck consists of the deepest layers of the U-Net with the largest channel dimension. We redesign the bottleneck by reducing its channel count from C to $\hat{C}$ ($\hat{C} < C$) to cut parameters while maximally preserving pre-trained generative priors. Further experimental analysis of bottleneck channel compression is provided in the the Appendix / Supplementary Material..

3.3 Channel-Spatial Disentangled Gated Skip Connection

As the feature-level improvement of our framework, we propose a channel-spatial disentangled gated skip connection. The core of this design is to implement adaptive feature selection via learnable weight modulation, which simultaneously adapts to the dynamic denoising process of diffusion models and the structural preservation requirement of medical image segmentation. Standard U-Net skip connections directly concatenate encoder features into the decoder in a static, content-agnostic

manner. While effective for deterministic segmentation, this design is suboptimal in diffusion-based frameworks for two core reasons. First, diffusion denoising is a time-dependent dynamic process. Encoder features passed through skip connections carry varying levels of noise and semantic information across timesteps, with drastically changing utility for decoder denoising. Static skip connections lack a mechanism to adaptively filter these encoder features based on the current noise level, leading to inefficient feature propagation. Second, encoder and decoder features are semantically misaligned in the diffusion paradigm. Encoder features preserve more fine-grained spatial details and structural priors, as they lie closer to the input segmentation mask and image conditioning. In contrast, decoder features contain more abstract semantic information, as they are optimized for noise prediction in the diffusion denoising process. Blindly fusing these misaligned features without targeted selection introduces irrelevant background noise and degrades structural mask recovery.

To address the above two limitations, we propose a channel-spatial disentangled gating mechanism. The core design principle is to assign the signal that inherently determines the feature property to its corresponding modulation branch. As illustrated in Figure 1 (c), channel gating is solely conditioned on timestep embedding. This is because the noise intensity $\bar{\alpha}_t$ in Equation equation 1 is a global signal that directly governs which feature channels are informative for the denoising objective at the current training timestep. At high-noise timesteps, channels encoding global semantic priors are prioritized, while at low-noise timesteps, channels preserving fine-grained structural details are more critical. The timestep-based channel gating thus adaptively enhances channels matched to the current noise intensity and training denoising demand, while suppressing redundant ones. Meanwhile, spatial gating is exclusively conditioned on both encoder and decoder features because the informativeness of a local spatial region is content-dependent. To judge whether a local feature is valid for spatial-level optimization, we need to consider both the structural content provided by the encoder feature, and the denoising and mask prediction demand of the decoder feature. Only by jointly modeling the encoder's input and decoder's demand can we reliably highlight discriminative target regions and suppress noisy background responses.

Formally, let encoder features $\mathbf{h}^e \in \mathbb{R}^{C_h \times H \times W}$ and decoder features $\mathbf{h}^d \in \mathbb{R}^{C_h \times H \times W}$ be the paired feature maps from the corresponding layer of the skip connection, with identical channel size C_h and spatial resolution $H \times W$. Given the timestep embedding $\mathbf{t}_{emb} \in \mathbb{R}^{C_t}$, we compute two complementary gating masks with independent design paradigms:

$$\mathbf{g}_c = G^{\mathcal{C}}(\mathbf{t}_{emb}) = \sigma\big(f_c(\mathbf{t}_{emb})\big), \quad \mathbf{g}_s = G^{\mathcal{S}}(\mathbf{h}^e, \mathbf{h}^d) = \sigma\big(f_s(\mathbf{h}^e \oplus \mathbf{h}^d)\big), \tag{5}$$

where f_c and f_s denote lightweight learnable modules implemented with tiny MLPs, $\sigma(\cdot)$ means Sigmoid activation function, $\oplus$ means feature concatenation. Specifically, f_c maps the global timestep embedding to a channel-wise gating mask $\mathbf{g}_c \in \mathbb{R}^{C_h \times 1 \times 1}$, which performs global channel reweighting to adapt feature activation patterns across different denoising timesteps. Meanwhile, f_s learns pixel-level spatial affinity via cross-feature matching between encoder and decoder features, outputting a spatial-wise gating mask $\mathbf{g}_s \in \mathbb{R}^{1 \times H \times W}$ to highlight semantically relevant regions and suppress noisy background responses across all channels. The final modulation mask is obtained by unifying global channel scaling and local spatial weighting in a single tensor $\mathbf{g}_{total} \in \mathbb{R}^{C_h \times H \times W}$:

$$\mathbf{g}_{total} = \mathbf{g}_c \odot \mathbf{g}_s. \tag{6}$$

where $\odot$ denotes the Hadamard product, with dimension broadcasting aligned with the input feature layout. The gated encoder features $\mathbf{h}^{e\prime}$ are computed by element-wise modulating:

$$\mathbf{h}^{e\prime} = \mathbf{h}^e \odot \mathbf{g}_{total}, \tag{7}$$

which are then concatenated with the decoder features $\mathbf{h}^d$ for subsequent denoising.

3.4 Latent-space Contour Alignment

We finally introduce the latent-space contour alignment as the loss-level improvement of our framework. Accurate boundary and contour localization is a core requirement for medical image segmentation. However, dominant training objectives in existing diffusion-based segmentation methods, such as $\mathcal{L}_{noise}$ and $\mathcal{L}_{latent}$ in Equation equation 3, are formulated as global pixel-wise regression tasks. These objectives apply uniform weights to all spatial regions of the feature map, while boundary regions only account for a tiny proportion of the total pixels. This leads to extremely low contribution of boundary optimization to the total loss, leaving the model lacking explicit, dedicated supervision for fine-grained boundary regions.

Table 1: Comparison with state-of-the-art methods on BTCV, ISIC2017 and ISIC2018 datasets. (Best and second-best metrics are shown in **bold** and underline, respectively. ↑ indicates the higher the better and ↓ indicates the lower the better. DMs denote methods based on diffusion models.)

Methods	DMs	Metrics on BTCV				Metrics on ISIC2017				Metrics on ISIC2018			
		Dice ↑	IoU ↑	HD95 ↓	ASD ↓	Dice ↑	IoU ↑	HD95 ↓	ASD ↓	Dice ↑	IoU ↑	HD95 ↓	ASD ↓
U-Net Ronneberger et al. (2015)	✗	84.52	74.36	19.10	4.90	89.50	82.81	13.73	5.32	86.59	78.41	21.78	8.70
U-Net(w/ R50) He et al. (2016)	✗	90.97	83.52	9.55	2.72	90.63	84.15	13.44	5.37	87.58	79.38	18.66	8.00
nnU-Net Isensee et al. (2021)	✗	90.55	82.92	9.61	2.71	90.59	84.29	11.36	4.52	87.42	79.33	18.57	8.19
SwinU-Net Cao et al. (2022)	✗	85.19	74.41	11.72	3.16	88.82	81.52	12.59	5.30	87.10	79.12	19.40	8.42
TransU-Net Chen et al. (2024a)	✗	91.88	85.04	9.80	2.73	90.54	83.93	10.49	4.27	87.93	79.91	17.89	7.57
U-KAN Li et al. (2025)	✗	89.13	80.56	6.87	1.58	90.50	84.14	12.61	4.78	86.56	78.49	21.53	8.40
VM-UNet Ruan et al. (2025)	✗	82.36	70.17	15.68	4.35	90.33	84.12	11.14	4.53	87.52	80.20	19.82	7.73
KM-UNet Zhang et al. (2025b)	✗	90.67	83.05	8.42	2.24	89.13	82.47	13.28	5.22	88.18	81.08	17.73	6.93
LoMix Rahman & Marculescu (2025)	✗	91.82	84.98	6.79	2.04	91.06	84.88	10.36	3.89	88.43	80.82	17.68	7.13
MedSegDiff Wu et al. (2024a)	✓	85.72	75.55	10.19	2.87	84.14	76.30	21.64	8.95	84.72	76.40	24.60	10.41
MedSegDiff-V2 Wu et al. (2024b)	✓	87.65	77.39	9.64	2.74	84.95	76.70	22.29	8.99	86.65	78.69	22.14	8.89
SDSeg Lin et al. (2024)	✓	92.15	85.51	8.17	2.00	90.77	84.38	10.99	4.17	88.15	80.34	18.25	7.40
TSLDSeg Yang et al. (2026)	✓	91.46	84.32	7.63	2.12	90.06	83.90	12.08	4.80	88.46	81.18	17.39	6.62
GCLDS (ours)	✓	**93.07**	**87.13**	**5.20**	**1.35**	**92.11**	**86.36**	**9.21**	**3.25**	**89.83**	**83.16**	**16.32**	**5.94**

A common remedy for boundary enhancement is to impose contour constraints in the pixel space. However, this approach introduces non-negligible extra computation, as it requires decoding latent features back to high-resolution pixel space in every training iteration. In contrast, we argue that contour alignment can be directly performed in the latent space. As validated by prior work (Lin et al., 2024), the VAE-encoded latent z_0 faithfully preserves spatial structures and contour information of segmentation masks, which we further verify via dedicated ablation study in Section 4.2.3. This enables us to directly model contour consistency at the latent space, with negligible computational overhead and full compatibility with the native diffusion framework.

Motivated by this, we introduce a latent contour alignment loss, as illustrated in Figure 1 (d), to align the edge gradient maps of the predicted latent $\hat{z}_0$ and the ground-truth latent z_0. We adopt Sobel operators, a classic differential kernel that extracts boundary-aware gradient maps in a fully differentiable and computationally efficient manner. Let S_x and S_y denote 3×3 horizontal and vertical Sobel kernels, which we apply depthwise (independently per channel) to the multi-channel latent features. We compute the gradient maps for z_0 and $\hat{z}_0$ as:

$$\begin{aligned} \nabla_x z_0, \nabla_y z_0 &= \text{Conv}(z_0, S_x), \text{Conv}(z_0, S_y), \\ \nabla_x \hat{z}_0, \nabla_y \hat{z}_0 &= \text{Conv}(\hat{z}_0, S_x), \text{Conv}(\hat{z}_0, S_y). \end{aligned} \tag{8}$$

We then minimize the mean absolute error between paired gradient maps to enforce explicit contour alignment:

$$\mathcal{L}_{\text{contour}} = \mathbb{E}_{z_0,t,\epsilon}\left[\|\nabla_x z_0 - \nabla_x \hat{z}_0\|_1 + \|\nabla_y z_0 - \nabla_y \hat{z}_0\|_1\right]. \tag{9}$$

To preserve the convergence stability and ensure fair comparisons with baseline methods, we fully retain the original diffusion losses. Our latent contour alignment loss acts as a complementary objective that focuses exclusively on clinically critical contour regions. Finally, the total training objective is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{noise}} + \mathcal{L}_{\text{latent}} + \mathcal{L}_{\text{contour}}. \tag{10}$$

4 EXPERIMENTS

This section presents experimental results and analysis of GCLDS for medical image segmentation under a fully supervised setting. We compare GCLDS with state-of-the-art approaches across multiple datasets, including BTCV, ISIC2017, ISIC2018, and REFUGE2, with the REFUGE2 results reported in the Appendix, and conduct ablation studies to assess the contributions of individual components and provide corresponding analysis and discussion. Datasets, evaluation metrics, implementation details, additional ablation analyses, and supplementary results with visualizations are provided in the Appendix / Supplementary Material.

Table 2: Ablation study on the effectiveness of individual components.

Components			Metrics on ISIC2017			Metrics on ISIC2018		
Bottleneck Compression	Channel-Spatial Disentangled Gate	Latent Contour Alignment Loss	Dice ↑	IoU ↑	HD95 ↓	Dice ↑	IoU ↑	HD95 ↓
✗	✗	✗	90.77	84.38	10.99	88.15	80.34	18.25
✔	✗	✗	90.79	84.48	10.81	88.39	81.01	17.83
✗	✔	✗	91.57	85.55	10.52	89.16	82.25	17.22
✗	✗	✔	<u>91.78</u>	<u>85.89</u>	<u>10.08</u>	89.14	82.19	<u>17.16</u>
✔	✔	✗	91.66	85.78	10.15	<u>89.29</u>	<u>82.26</u>	17.17
✔	✔	✔	**92.11**	**86.36**	**9.21**	**89.83**	**83.16**	**16.32**

4.1 Comparison with State-of-the-Art Methods

To evaluate the effectiveness of GCLDS, we select representative approaches in the field for comparison. Specifically, U-Net (Ronneberger et al., 2015), U-Net (w/ R50) [1], nnU-Net (Isensee et al., 2021) and LoMix (Rahman & Marculescu, 2025) are based on CNN. TransU-Net (Chen et al., 2024a) and SwinU-Net (Cao et al., 2022) are based on Transformer. U-KAN (Li et al., 2025), VM-UNet (Ruan et al., 2025) and KM-UNet (Zhang et al., 2025b) are based on emerging architectures such as KAN and Mamba. MedSegDiff (Wu et al., 2024a), MedSegDiff-V2 (Wu et al., 2024b), TSLDSeg (Yang et al., 2026) and SDSeg (Lin et al., 2024) are Diffusion-based generative methods.

Table 1 compares the methods on the BTCV abdominal organ segmentation dataset. Due to large anatomical variability and low soft-tissue contrast in CT images, this task requires strong structural modeling and accurate boundary localization. As shown in the table, GCLDS achieves the best performance on all metrics, with Dice and IoU reaching 93.07% and 87.13%, respectively. In particular, GCLDS reduces HD95 and ASD by 2.97 and 0.65 over SDSeg, demonstrating superior boundary delineation and better capture of anatomical structures. Compared with conventional discriminative models, GCLDS shows clear advantages in all metrics, highlighting the effectiveness of diffusion-based generative priors for medical image segmentation.

Table 1 reports quantitative results on the ISIC2017 and ISIC2018 datasets. GCLDS achieves the best results across all metrics, with Dice/IoU of 92.11%/86.36% and 89.83%/83.16%, respectively, improving SDSeg by 1.34, 1.98, 1.68, and 2.82 percentage points. It also substantially reduces HD95 and ASD, demonstrating the superiority of GCLDS for accurate and efficient lesion segmentation.

4.2 Ablation Studies

4.2.1 Ablation Study on the Effectiveness of Individual Components

Table 2 reports ablation results to verify the independent effectiveness and complementary synergy of each proposed component in GCLDS on ISIC2017 and ISIC2018 datasets, with the first row (no modules added) corresponding to the baseline method (Lin et al., 2024).

The second to fourth rows show that introducing each of our three components individually delivers consistent performance improvements over the baseline. Introducing the lightweight bottleneck achieves lossless accuracy retention with better boundary localization, as HD95 are reduced by 0.18 and 0.42 respectively. Adding the channel-spatial disentangled gated skip connection brings significant gains in segmentation accuracy indicated by the improvement of Dice and IoU in both datasets. Utilizing the latent contour alignment loss also delivers a substantial reduction in boundary-sensitive HD95. These results validate the independent effectiveness of our architecture-level, feature-level, and loss-level targeted improvements, respectively.

The last two rows further demonstrate that integrating our complementary components yields continuous performance gains. The full GCLDS achieves optimal overall performance with Dice improving by 1.34% and 1.68%, respectively, and HD95 decreasing by 1.78 and 1.93. This confirms strong synergy of our three designs, which jointly address the limitations of native LDM segmentation method, and achieve simultaneous improvements in segmentation accuracy and boundary precision.

[1] Replace the encoder of U-Net to a pre-trained ResNet50 (He et al., 2016).

Table 3: Ablation study on the Channel-Spatial disentangled gated skip connection.

Components		Metrics on ISIC2017			Metrics on ISIC2018		
Channel Gate	Spatial Gate	Dice ↑	IoU ↑	HD95 ↓	Dice ↑	IoU ↑	HD95 ↓
✗	✗	90.77	84.38	10.99	88.15	80.34	18.25
✓	✗	91.18	85.16	<u>10.56</u>	88.54	81.02	<u>17.43</u>
✗	✓	<u>91.23</u>	<u>85.19</u>	10.68	<u>88.62</u>	<u>81.27</u>	17.66
✓	✓	**91.57**	**85.55**	**10.52**	**89.16**	**82.25**	**17.22**

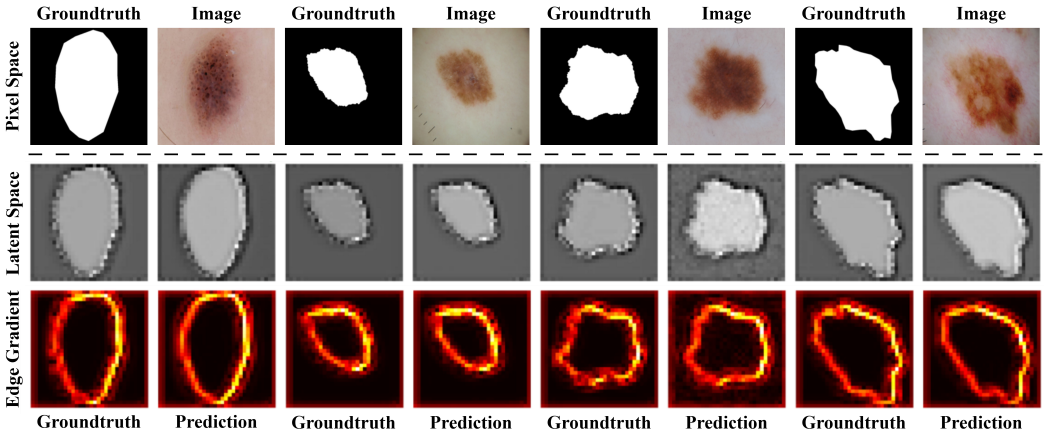

Figure 2: Qualitative visualization of latent space edge structures.

4.2.2 Ablation Study on the Channel-Spatial Disentangled Gated Skip Connection

To validate the independent effectiveness of the channel gate and spatial gate in our disentangled gating mechanism, we conduct this ablation study on the vanilla baseline model without other proposed improvements (i.e., lightweight bottleneck and latent contour alignment loss). We construct four controlled variants of the gating component: (i) *No Gate*: baseline with vanilla skip connections; (ii) *Channel-only*: only timestep-aware channel gating is applied; (iii) *Spatial-only*: only encoder-decoder guided spatial gating is applied; (iv) *Channel-Spatial*: our full disentangled gating mechanism. Results in Table 3 show that both channel and spatial gating bring consistent performance gains across both datasets, verifying their complementary effectiveness. Specifically, channel gating helps adapting to time-varying noise levels across denoising steps, while spatial gating enhances structural patterns alignment via encoder-decoder feature matching. Combining both disentangled gates yields the best overall performance, demonstrating the superiority of our channel-spatial gating for both global temporal modulation and local spatial selection.

4.2.3 Analysis of the Latent Contour Alignment Loss

To verify the core rationality of our proposed latent contour alignment loss, we conduct qualitative analysis on the converged GCLDS model, with results shown in Figure 2. We visualize the input images c, pixel-space ground-truth masks x_0, their corresponding VAE-encoded latent representations z_0, model-predicted latents $\hat{z}_0$, and Sobel edge gradient maps derived from these latent features. We observe that, despite the dimensionality compression by the VAE encoder, the core contour and key fine-grained boundary structures of the segmentation masks are well preserved. This confirms that z_0 retains sufficient structural information to support reliable boundary modeling. Furthermore, the high consistency between Sobel gradient maps of $\hat{z}_0$ and z_0 validates that our proposed contour loss can effectively enforce latent space contour alignment. These visualizations demonstrate that our latent contour alignment loss reliably and effectively improves the segmentation boundary accuracy.

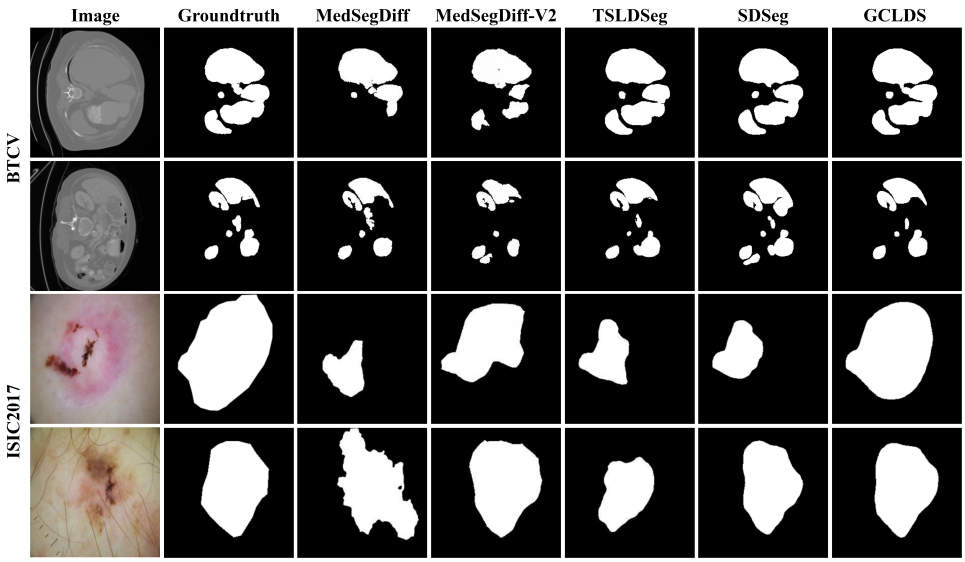

Figure 3: Visualization comparison on the BTCV and ISIC2017 datasets.

Table 4: Comparison of computational efficiency on the BTCV dataset.

Methods	Trainable Parameters	Training Memory	Inference Time	Dice ↑
MedSegDiff Wu et al. (2024a)	130 M	≈ 14.2 GB	≈ 16.55 h	85.72
MedSegDiff-V2 Wu et al. (2024b)	129 M	≈ 14.3 GB	≈ 16.55 h	87.65
TSLDSeg Yang et al. (2026)	339 M	≈ 13.5 GB	≈ 0.21 h	91.46
SDSeg Lin et al. (2024)	329 M	≈ 12.4 GB	≈ 0.09 h	<u>92.15</u>
GCLDS (ours)	225 M	≈ 10.0 GB	≈ 0.07 h	**93.07**

4.2.4 Visualization Results

Figure 3 presents a qualitative comparison between GCLDS and other state-of-the-art diffusion-based approaches on the BTCV and ISIC2017 datasets. The competing methods exhibit different types of errors across the examples. In contrast, GCLDS avoids these issues and achieves consistently accurate segmentation results across all displayed examples.

4.2.5 Computational Efficiency Comparison

Table 4 compares the computational cost of GCLDS with other diffusion-based methods. The inference time indicates the processing time for the entirety of BTCV test set. Compared with pixel-space iterative methods (MedSegDiff and MedSegDiff-V2), GCLDS achieves orders-of-magnitude faster inference speed, eliminating the limitation of traditional multi-step diffusion pipelines. For latent-space baselines (TSLDSeg and SDSeg), our lightweight bottleneck design reduces trainable parameters by up to 31.6% and training memory consumption by 19.4% versus SDSeg, while further shortening inference time and improving the Dice score from 92.15% to 93.07%. Overall, GCLDS achieves a favorable accuracy-efficiency trade-off with lower computational overhead.

5 Conclusion

In this work, we present Channel-Spatial **G**ated and **C**ontour Aligned **L**atent **D**iffusion **S**egmentation (**GCLDS**), an efficient latent diffusion framework for medical image segmentation. By combining a lightweight bottleneck, disentangled channel-spatial gating, and latent-space contour alignment, GCLDS improves feature fusion and boundary accuracy with reduced computational cost. Experiments on multiple medical benchmarks demonstrate strong segmentation performance with fewer parameters, lower memory consumption, and shorter inference latency than existing diffusion-based methods. **Limitation and Future Work:** GCLDS currently supports binary segmentation, and extending it to multi-class segmentation remains future work.

AI USE STATEMENT

In this work, we used generative AI tools for manuscript language refinement and code debugging. We have not used generative AI tools for generating core model designs, experimental results, or technical contributions. All AI-assisted content has been fully reviewed and verified by the authors. We take full responsibility for the final content of this work.

ETHICS STATEMENT

To the best of our knowledge, the research in this paper conforms to the ICLR Code of Ethics. The study involves standard medical image segmentation experiments on publicly available datasets and does not raise issues related to human subjects, privacy, safety, misuse, or other ethical concerns addressed by the guidelines.

REPRODUCIBILITY STATEMENT

Due to page limitations in the main paper, we provide the datasets, evaluation metrics, and implementation details used in our experiments in *Section B.1 Experimental Setup*, including *Section B.1.1 Datasets and Evaluation Metrics* and *Section B.1.2 Implementation Details.* We also present the algorithms of the channel-spatial disentangled gating mechanism and the latent contour alignment loss in *Section A Training Algorithm.*

REFERENCES

Jung Su Ahn, Ki Hoon Kwak, Jung Woo Seo, and Young-Rae Cho. MoDiff: A morphology-emphasized diffusion model for ambiguous medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 386–395. Springer, 2025.

Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-Unet: Unet-like pure transformer for medical image segmentation. In *Proceedings of the European Conference on Computer Vision*, pp. 205–218. Springer, 2022.

Jieneng Chen, Jieru Mei, Xianhang Li, Yongyi Lu, Qihang Yu, Qingyue Wei, Xiangde Luo, Yutong Xie, Ehsan Adeli, Yan Wang, Matthew P. Lungren, Shaoting Zhang, Lei Xing, Le Lu, Alan Yuille, and Yuyin Zhou. TransUNet: Rethinking the u-net architecture design for medical image segmentation through the lens of transformers. *Medical Image Analysis*, 97:103280, 2024a.

Qi Chen, Xiaoxi Chen, Haorui Song, Zhiwei Xiong, Alan Yuille, Chen Wei, and Zongwei Zhou. Towards generalizable tumor synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11147–11158, 2024b.

Tao Chen, Chenhui Wang, Zhihao Chen, Yiming Lei, and Hongming Shan. HiDiff: Hybrid diffusion framework for medical image segmentation. *IEEE Transactions on Medical Imaging*, 43(10): 3570–3583, 2024c.

Noel Codella, Veronica Rotemberg, Philipp Tschandl, M. Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, Harald Kittler, and Allan Halpern. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC). *arXiv preprint arXiv:1902.03368*, 2019.

Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC). In *IEEE International Symposium on Biomedical Imaging*, pp. 168–172, 2018.

Chenhe Du, Qing Wu, Xuanyu Tian, Jingyi Yu, Hongjiang Wei, and Yuyao Zhang. Improving 2d diffusion models for 3d medical imaging with inter-slice consistent stochasticity. In *International Conference on Learning Representations*, pp. 39622–39642, 2026.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

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

Amirhossein Kazerouni, Ehsan Khodapanah Aghdam, Moein Heidari, Reza Azad, Mohsen Fayyaz, Ilker Hacihaliloglu, and Dorit Merhof. Diffusion models in medical imaging: A comprehensive survey. *Medical image analysis*, 88:102846, 2023.

Jonghun Kim and Hyunjin Park. Adaptive latent diffusion model for 3d medical image to image translation: Multi-modal magnetic resonance imaging study. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 7604–7613, 2024.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Bennett Landman, Zhoubing Xu, Juan Igelsias, Martin Styner, Thomas Langerak, and Arno Klein. MICCAI multi-atlas labeling beyond the cranial vault–workshop and challenge. In *Proc. MICCAI multi-atlas labeling beyond cranial vault—workshop challenge*, volume 5, pp. 12. Munich, Germany, 2015.

Mengqi Lei, Haochen Wu, Xinhua Lv, and Xin Wang. ConDSeg: A general medical image segmentation framework via contrast-driven feature enhancement. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 4571–4579, 2025.

Chenxin Li, Xinyu Liu, Wuyang Li, Cheng Wang, Hengyu Liu, Yifan Liu, Zhen Chen, and Yixuan Yuan. U-KAN makes strong backbone for medical image segmentation and generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 2025.

Fei Li, Diping Song, Han Chen, Jian Xiong, Xingyi Li, Hua Zhong, Guangxian Tang, Sujie Fan, Dennis SC Lam, Weihua Pan, et al. Development and clinical deployment of a smartphone-based visual field deep learning system for glaucoma detection. *NPJ digital medicine*, 3(1):123, 2020.

Xiang Li, Huihui Fang, Mingsi Liu, Yanwu Xu, and Lixin Duan. Diffusion-enhanced transformation consistency learning for retinal image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 221–231. Springer, 2024.

Xinran Li, Yi Shuai, Chen Liu, Qi Chen, Tianyu Lin, Pengfei Guo, Dong Yang, Can Zhao, Pedro RAS Bassi, Daguang Xu, et al. Text-driven tumor synthesis. *IEEE Transactions on Medical Imaging*, 2026.

Tianyu Lin, Zhiguang Chen, Zhonghao Yan, Weijiang Yu, and Fudan Zheng. Stable diffusion segmentation for biomedical images with single-step reverse process. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 656–666. Springer, 2024.

Huynh Trinh Ngoc, Toan Nguyen Hai, Ba Luong Son, and Long Tran Quoc. Diffusion model in latent space for medical image segmentation task. *arXiv preprint arXiv:2512.01292*, 2025.

José Ignacio Orlando, Huazhu Fu, João Barbosa Breda, Karel Van Keer, Deepti R Bathula, Andrés Diaz-Pinto, Ruogu Fang, Pheng-Ann Heng, Jeyoung Kim, JoonHo Lee, et al. REFUGE Challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. *Medical Image Analysis*, 59:101570, 2020.

Aimon Rahman, Jeya Maria Jose Valanarasu, Ilker Hacihaliloglu, and Vishal M Patel. Ambiguous medical image segmentation using diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11536–11546, 2023.

Md Mostafijur Rahman and Radu Marculescu. LoMix: Learnable weighted multi-scale logits mixing for medical image segmentation. *Advances in Neural Information Processing Systems*, 2025.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241. Springer, 2015.

Jiacheng Ruan, Jincheng Li, and Suncheng Xiang. VM-UNet: Vision mamba unet for medical image segmentation. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2025.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021a.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021b.

Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The HAM10000 dataset: A large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5: 180161, 2018. doi: 10.1038/sdata.2018.161.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in Neural Information Processing Systems*, 30, 2017.

Zhonghua Wang, Junyan Lyu, and Xiaoying Tang. SET: Superpixel embedded transformer for skin lesion segmentation. *Medical Image Analysis*, pp. 103738, 2025.

Junde Wu, Rao Fu, Huihui Fang, Yu Zhang, Yehui Yang, Haoyi Xiong, Huiying Liu, and Yanwu Xu. MedSegDiff: Medical image segmentation with diffusion probabilistic model. In *Medical Imaging with Deep Learning*, pp. 1623–1639. PMLR, 2024a.

Junde Wu, Wei Ji, Huazhu Fu, Min Xu, Yueming Jin, and Yanwu Xu. MedSegDiff-V2: Diffusion-based medical image segmentation with transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 6030–6038, 2024b.

Shangqian Wu, Siyuan Shen, Yahan Li, Zhijian Huang, Ziyu Fan, Yuanpeng Zhang, YI WANG, and Lei Deng. Cross-timestep: 3d diffusion model with trans-temporal memory lstm and adaptive priori decoding strategy for medical segmentation. In *International Conference on Learning Representations*, pp. 75674–75693, 2026.

Zhaohu Xing, Liang Wan, Huazhu Fu, Guang Yang, Yijun Yang, Lequan Yu, Baiying Lei, and Lei Zhu. Diff-UNet: A diffusion embedded network for robust 3d medical image segmentation. *Medical Image Analysis*, 105:103654, 2025.

Zongjian Yang, Chunquan Li, and Jiquan Ma. TSLDSeg: A texture-aware and semantic-enhanced latent diffusion model for medical image segmentation. *Pattern Recognition*, pp. 112795, 2026.

Han Zhang, Xiangde Luo, Yong Chen, and Kang Li. DiffOSeg: Omni medical image segmentation via multi-expert collaboration diffusion model. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 128–138. Springer, 2025a.

Hongsheng Zhang, Yuting Duan, Ting Liu, Weifeng Zhang, and Hongzhong Tang. KMUNet: A novel medical image segmentation model based on KAN and Mamba. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 298–307. Springer, 2025b.

Zheyuan Zhang, Lanhong Yao, Bin Wang, Debesh Jha, Gorkem Durak, Elif Keles, Alpay Medetalibeyoglu, and Ulas Bagci. DiffBoost: Enhancing medical image segmentation via text-guided diffusion model. *IEEE Transactions on Medical Imaging*, 44(9):3670–3682, 2024.

Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision Mamba: Efficient visual representation learning with bidirectional state space model. In *International Conference on Machine Learning*, volume 235, pp. 62429–62442, 2024.

Channel-Spatial Gated and Contour Aligned Latent Diffusion for Medical Image Segmentation

Appendix / Supplementary Material

This document provides the detailed training algorithm, experimental setup, complementary analyses, and limitations of the proposed method supporting the main text. In Section A, we present the algorithms of the channel-spatial disentangled gating mechanism and the latent contour alignment loss. In Section B.1, we describe the datasets, evaluation metrics, and implementation details used in our experiments. In Section B.2, we provide additional experimental results and analysis, including results on the REFUGE2 dataset (Section B.2.1), analysis of the lightweight bottleneck (Section B.2.2), analysis of the channel-spatial disentangled gating mechanism (Section B.2.3), supplemental visualization results (Section B.2.4), and statistical significance analysis (Section B.2.5). In Section C, we discuss the limitations of GCLDS and future directions for improvement.

A Training Algorithm

A.1 Algorithm of Channel-Spatial Disentangled Gating Mechanism

Algorithm 1 outlines the proposed channel-spatial disentangled gating process. Given the decoder feature map $\mathbf{h}^d$, encoder feature map $\mathbf{h}^e$, and timestep embedding $\mathbf{t}_{emb}$, the algorithm first computes the channel gate $\mathbf{g}_c = \sigma(f_c(\mathbf{t}_{emb}))$ from the timestep embedding, then derives the spatial gate $\mathbf{g}_s = \sigma\left(f_s([\mathbf{h}^e \oplus \mathbf{h}^d])\right)$ by projecting $\mathbf{h}^e$ and $\mathbf{h}^d$ into a shared hidden space via small MLPs. The two gates are fused as $\mathbf{g}_{total} = \mathbf{g}_c \odot \mathbf{g}_s$, and the final gated encoder feature map is obtained by $\mathbf{h}^{e\prime} = \mathbf{h}^e \odot \mathbf{g}_{total}$. This disentangled design enables adaptive skip feature modulation by jointly considering the diffusion timestep and the encoder-decoder feature content.

A.2 Algorithm of Latent Contour Alignment Loss

Algorithm 2 outlines the proposed latent contour alignment loss. Given the ground-truth latent variable z_0 and the predicted latent variable $\hat{z}_0$, we first apply fixed 3×3 Sobel filters to extract horizontal and vertical gradient maps in the latent space. The loss then minimizes the ℓ_1 distance (i.e., mean absolute error) between the corresponding gradients of z_0 and $\hat{z}_0$, thereby encouraging the predicted latent representation to preserve contour-consistent structural details aligned with the ground truth.

B Experiments

B.1 Experimental Setup

B.1.1 Datasets and Evaluation Metrics

To comprehensively evaluate the effectiveness of the proposed Channel-Spatial **G**ated and **C**ontour Aligned **L**atent **D**iffusion **S**egmentation (**GCLDS**) framework, we conduct experiments on four representative benchmark datasets, including BTCV, ISIC2017, ISIC2018, and REFUGE2. These datasets cover both CT and RGB imaging modalities, as well as organ segmentation and lesion segmentation tasks, enabling a thorough assessment of the proposed method across diverse anatomical structures, imaging characteristics, and segmentation challenges. For all experiments, we apply an identical task formulation and preprocessing pipeline to our method and all compared baselines (including both diffusion and non-diffusion approaches), ensuring fair comparisons across all evaluations.

BTCV dataset is a multi-class abdominal organ segmentation dataset originating from the 2015 MICCAI workshop, *Multi-Atlas Labeling Beyond the Cranial Vault* (Landman et al., 2015). It con-

Algorithm 1 Channel-Spatial Disentangled Gated Skip Connection

Require: Decoder feature map $\mathbf{h}^d$, encoder skip feature map $\mathbf{h}^e$, timestep embedding $\mathbf{t}_{emb}$
Ensure: Gated encoder feature map $\mathbf{h}^{e\prime}$

$\mathbf{g}_c \leftarrow \sigma\big(f_c(\mathbf{t}_{emb})\big)$ (Compute global channel-wise weights)
$\mathbf{g}_s \leftarrow \sigma\big(f_s([\mathbf{h}^e \oplus \mathbf{h}^d])\big)$ (Compute local spatial-wise weights)
$\mathbf{g}_{total} \leftarrow \mathbf{g}_c \odot \mathbf{g}_s$ (Fuse channel and spatial gates)
$\mathbf{h}^{e\prime} \leftarrow \mathbf{h}^e \odot \mathbf{g}_{total}$ (Apply gating to encoder features)
return $\mathbf{h}^{e\prime}$

Algorithm 2 Latent Contour Alignment Loss

Require: Predicted latent variable $\hat{z}_0$, ground-truth latent variable z_0
Ensure: Contour loss $\mathcal{L}_{\text{contour}}$

$S_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}, \quad S_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix}$ (Define 3×3 Sobel kernels)
$\nabla_x z_0 \leftarrow \text{Conv2d}\left(z_0, S_x, \text{groups} = C\right)$ (Compute ground-truth horizontal gradient)
$\nabla_y z_0 \leftarrow \text{Conv2d}\left(z_0, S_y, \text{groups} = C\right)$ (Compute ground-truth vertical gradient)
$\nabla_x \hat{z}_0 \leftarrow \text{Conv2d}\left(\hat{z}_0, S_x, \text{groups} = C\right)$ (Compute predicted horizontal gradient)
$\nabla_y \hat{z}_0 \leftarrow \text{Conv2d}\left(\hat{z}_0, S_y, \text{groups} = C\right)$ (Compute predicted vertical gradient)
$\mathcal{L}_{\text{contour}} \leftarrow \mathbb{E}_{z_0,t,\epsilon}\left[\|\nabla_x z_0 - \nabla_x \hat{z}_0\|_1 + \|\nabla_y z_0 - \nabla_y \hat{z}_0\|_1\right]$ (Aggregate contour loss)
return $\mathcal{L}_{\text{contour}}$

tains 30 abdominal CT scans annotated by expert physicians and provided by Vanderbilt University Medical Center, with 14 segmentation labels in total, including the background and 13 abdominal organs. In our experimental setup, the 30 three-dimensional CT volumes are split into 18 training cases and 12 test cases, then cropped into two-dimensional slices. The 13 abdominal organs are merged into a single foreground class, thereby reformulating the task as binary segmentation. This preprocessing strategy is kept consistent with that of the baseline method (Lin et al., 2024), ensuring a fair comparison.

ISIC2017 dataset is a publicly available dermoscopic image dataset for melanoma segmentation, collected and annotated by the International Skin Imaging Collaboration (ISIC) (Codella et al., 2018). The dataset contains 2,000 RGB dermoscopic images with pixel-level ground-truth annotations, and the task is formulated as binary segmentation, i.e., distinguishing lesion regions from the background. In this study, the dataset is split into training and test sets in an 80%:20% ratio, containing 1,600 and 400 images, respectively, for fully supervised medical image segmentation experiments.

ISIC2018 is a publicly available dermoscopic image dataset for skin lesion segmentation, released by the International Skin Imaging Collaboration (ISIC) (Codella et al., 2019; Tschandl et al., 2018). It contains 3,594 RGB dermoscopic images with pixel-level ground-truth annotations, and the task is formulated as binary segmentation, i.e., separating lesion regions from the background. Compared with ISIC 2017, ISIC 2018 is constructed from a more diverse set of sources, with the official training data aggregated from HAM10000 and MSK. For the fully supervised experiments, we follow the official split and use 2,594 images for training and 1,000 images for testing.

REFUGE2 dataset is a publicly available color fundus photography (CFP) dataset for glaucoma analysis, providing segmentation annotations for the optic disc (OD) and optic cup (OC). It contains 1,200 CFP images and was released by the Zhongshan Ophthalmic Center (ZOC) for MICCAI (Li et al., 2020; Orlando et al., 2020). The dataset is divided into 800 training and 400 test images for all methods, with the task formulated as binary segmentation of the optic cup (OC) only. To better balance the foreground and background classes, we use preprocessed cropped images of size 500×500 pixels.

We employ several standard metrics for medical image segmentation, including the Dice Similarity Coefficient (Dice) and Intersection over Union (IoU) to measure region overlap. In addition, we utilize the 95% Hausdorff Distance (HD95) and the Average Surface Distance (ASD) to assess the accuracy of segmentation boundaries.

Table 5: Comparison with state-of-the-art methods on the REFUGE2 dataset for optic cup (OC) segmentation. (The best and second-best metrics are shown in **bold** and underline, respectively. ↑ indicates the higher the better.)

Methods	Publication	Model Paradigm	Dice ↑	IoU ↑
U-Net Ronneberger et al. (2015)	MICCAI 2015	CNN	82.34	70.94
U-Net(w/ R50) He et al. (2016)	CVPR 2016	CNN	84.28	73.58
nnU-Net Isensee et al. (2021)	Nature Methods 2021	CNN	85.15	74.83
LoMix Rahman & Marculescu (2025)	NIPS 2025	CNN	84.04	73.33
SwinU-Net Cao et al. (2022)	ECCV 2022	Transformer	84.21	73.50
TransU-Net Chen et al. (2024a)	MedIA 2024	Transformer	85.81	75.98
U-KAN Li et al. (2025)	AAAI 2025	KAN	83.97	73.27
VM-UNet Ruan et al. (2025)	TOMM 2025	Mamba	85.52	75.39
KM-UNet Zhang et al. (2025b)	MICCAI 2025	KAN+Mamba	83.60	72.86
MedSegDiff Wu et al. (2024a)	MIDL 2024	Diffusion	82.18	70.85
SDSeg Lin et al. (2024)	MICCAI 2024	Diffusion	<u>86.68</u>	<u>77.13</u>
TSLDSeg Yang et al. (2026)	Pattern Recognition 2026	Diffusion	85.64	75.49
GCLDS (ours)	This Paper	Diffusion	**87.86**	**79.01**

B.1.2 Implementation Details

GCLDS is trained on a single RTX 4090 D GPU with 24 GB RAM. It is optimized using the AdamW optimizer with a learning rate of 1×10^{-5}, a total of 100,000 training iterations, and a default batch size of 4. Following the prior work (Lin et al., 2024), GCLDS adopts a VAE and a U-Net backbone for latent diffusion modeling with downsampling factor of $r = H/h = W/w = 8$. The input to GCLDS is standardized as an RGB image of spatial resolution $H = W = 256$ pixels. For single-channel medical images such as CT scans, the input is replicated across three channels to match the input format expected by the pretrained latent diffusion model. As a result, the latent representation has a spatial size of $h = w = 32$ and $c = 4$ channels, which preserves compatibility with the standard latent diffusion configuration while accommodating the characteristics of medical images.

For weight initialization, our GCLDS architecture is customized accordingly. All base components, including the VAE and the U-Net backbone, are initialized with the pretrained weights released in Rombach et al. (2022) to benefit from transfer learning from natural image synthesis. Notably, since the bottleneck of U-Net in GCLDS has been structurally compressed ($C = 768$ and $\hat{C} = 576$), pretrained weights are not loaded for layers whose channel dimensions do not match; these layers are initialized randomly. During inference, GCLDS adopts an efficient single-step sampling strategy following existing work (Lin et al., 2024) by default, with compatibility for other mainstream samplers (e.g., DDIM (Song et al., 2021a)).

B.2 Complementary Experiments

B.2.1 Results on REFUGE2 Dataset

Table 5 reports the comparative results on the REFUGE2 optic cup segmentation dataset. For this task, the boundary between the optic cup and optic disc is often ambiguous with low tissue contrast in retinal fundus images. This places extremely high demands on accurate localization of the segmentation results. As shown in the table, GCLDS achieves the best Dice and IoU scores of 87.86% and 79.01%, respectively, outperforming all competing methods, indicating stronger overall segmentation capability and better region overlap. These results demonstrate that GCLDS is particularly effective in capturing the optic cup region, while maintaining competitive performance on challenging retinal segmentation tasks.

B.2.2 Analysis of the Lightweight Bottleneck

To justify our lightweight bottleneck design, we conduct the subsequent experiments. As shown in Figure 4, we train a baseline model with vanilla LDM U-Net on the BTCV dataset. We define the mean absolute activation of each bottleneck's channel as the importance score, and calculate their cumulative contribution. It can be found that 75% of channels contribute over 90% of the total

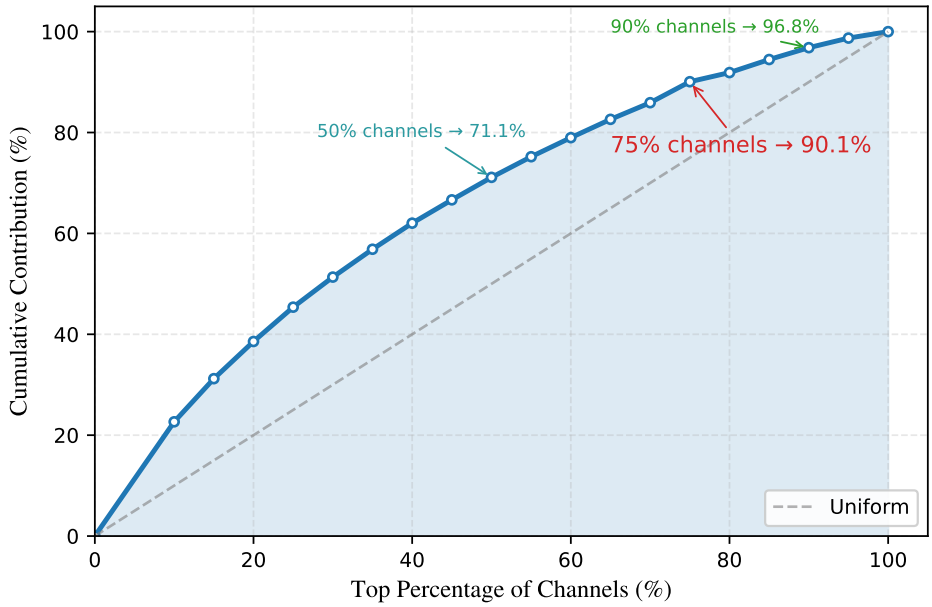

Figure 4: Cumulative contribution of channel importance in bottleneck of the diffusion model.

Table 6: Ablation study of the bottleneck with different channels.

Channel Count	Dice ↑	IoU ↑	HD95 ↓	ASD ↓
$C = 768$ (100%)	**92.15**	85.51	8.17	2.00
$C = 576$ (75%)	92.13	**85.53**	**6.75**	**1.81**
$C = 384$ (50%)	91.64	84.83	9.03	2.95

Table 7: Ablation study on different designs of gating mechanism.

Gating Mechanism	Dice ↑	IoU ↑	HD95 ↓	ASD ↓
Static encoder-decoder gating	91.02	84.87	11.32	4.11
Mixed signals channel gating	91.29	85.30	10.86	3.89
Coupled unified gating	91.48	85.39	10.62	3.82
Channel-Spatial disentangled gating	**91.57**	**85.55**	**10.52**	**3.70**

feature information. This observation indicates that the bottleneck has considerable capacity for channel compression and motivates our channel reduction strategy.

Moreover, we evaluate U-Net backbones with bottlenecks of varying channel dimensions on the BTCV dataset, with quantitative results summarized in Table 6. The setting $C = 768$ corresponds to the vanilla LDM U-Net, while $C = 576$ and $C = 384$ denote configurations where the bottleneck channel count is reduced to 75% and 50% of the original dimension, respectively. Our results demonstrate that retaining 75% of the original channels incurs no performance drop in region-level segmentation accuracy (i.e., Dice and IoU) and even improves boundary localization, as reflected by lower HD95 and ASD values. However, reducing the channel dimension to 50% leads to consistent degradation across all evaluation metrics. This trend is consistent with the observed concentration of channel contributions, suggesting that moderate compression can preserve the effective feature representation, whereas excessive compression may impair the network's capacity to model the dynamic diffusion denoising process. Meanwhile, retaining all channels incurs a higher parameter cost with limited performance benefits for medical image segmentation. Retaining 75% of the original channels therefore provides a favorable balance between model efficiency and segmentation performance. Accordingly, we set $C = 576$ as the default bottleneck channel dimension in GCLDS.

In addition, Figure 5 illustrates the feature importance curves. The baseline U-Net with the original bottleneck (i.e., $C = 768$) and the proposed lightweight U-Net with the redesigned bottleneck (i.e., $C = 576$) were trained on the BTCV dataset. The channel indices were normalized to the range of $[0, 1]$ and sorted by channel importance score in descending order for comparison. As shown in the figure, the lightweight bottleneck exhibits substantially higher average activation intensity across channels than the original structure. This observation suggests that the compressed bottleneck maintains stronger average feature responses with fewer channels, supporting its favorable efficiency–accuracy trade-off.

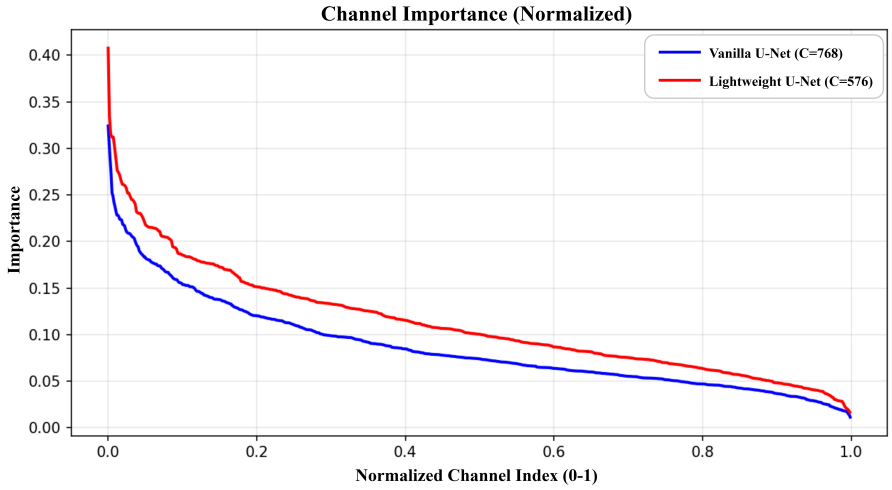

Figure 5: Normalized channel importance scores sorted in descending order.

B.2.3 Analysis of the Channel-Spatial Disentangled Gating Mechanism

To further validate the rationality of our proposed channel-spatial disentangled gating mechanism, we design several intuitive alternative gating variants for comparative ablation experiments:

(i) **Static encoder-decoder gating.** No timestep embedding $\mathbf{t}_{emb}$ is introduced, both the channel and spatial gates are computed exclusively from encoder features $\mathbf{h}^e$ and decoder features $\mathbf{h}^d$:

$$\mathbf{g}_c = G^{\mathcal{C}}(\mathbf{h}^e, \mathbf{h}^d) = \sigma\big(f_c(\mathbf{h}^e \oplus \mathbf{h}^d)\big), \mathbf{g}_s = G^{\mathcal{S}}(\mathbf{h}^e, \mathbf{h}^d) = \sigma\big(f_s(\mathbf{h}^e \oplus \mathbf{h}^d)\big). \tag{11}$$

(ii) **Mixed signals channel gating.** The spatial gating branch remains identical to our original design, while the channel gating is computed from the timestep embedding $\mathbf{t}_{emb}$ with additional encoder features $\mathbf{h}^e$ and decoder features $\mathbf{h}^d$ incorporated:

$$\mathbf{g}_c = G^{\mathcal{C}}(\mathbf{t}_{emb}, \mathbf{h}^e, \mathbf{h}^d) = \sigma\big(f_c(\mathbf{t}_{emb} \oplus \mathbf{h}^e \oplus \mathbf{h}^d)\big), \mathbf{g}_s = G^{\mathcal{S}}(\mathbf{h}^e, \mathbf{h}^d) = \sigma\big(f_s(\mathbf{h}^e \oplus \mathbf{h}^d)\big). \tag{12}$$

(iii) **Coupled unified gating.** A single full gating map with the size of $C_h \times H \times W$ is directly generated from the joint input of timestep embedding $\mathbf{t}_{emb}$, encoder features $\mathbf{h}^e$, and decoder features $\mathbf{h}^d$, with no disentanglement between channel and spatial modulation branches:

$$\mathbf{g}_{total} = G^{total}(\mathbf{t}_{emb}, \mathbf{h}^e, \mathbf{h}^d) = \sigma\big(f_{total}(\mathbf{t}_{emb} \oplus \mathbf{h}^e \oplus \mathbf{h}^d)\big). \tag{13}$$

As summarized in Table 7, our proposed channel-spatial disentangled gating outperforms other three alternative variants across all evaluation metrics. Compared with the static encoder-decoder gating, our timestep-aware channel gating enables adaptive selection of encoder channels that match the denoising demand of the current timestep, which validates the necessity of dynamic temporal modulation for diffusion-based segmentation. Compared with the mixed signals channel gating, our design strictly follows the core principle of assigning the matched signal to its corresponding modulation branch, so that each gating branch performs its dedicated function without cross-interference. Using mixed global and local signals to jointly modulate the channel gate will inevitably introduce feature confounding between the global temporal state and local spatial semantics, degrading the consistency of channel-wise modulation. Compared with the coupled unified gating, our disentangled design effectively eliminates the possible information conflict in coupled gating. Specifically, the timestep embedding is a one-dimensional global representation with no spatial location information, which provides no valid information gain for spatial-aware gating generation, and will only introduce irrelevant interference to spatial feature selection. These ablation results consistently verify that our proposed gating mechanism has a more theoretically sound design, and achieves superior segmentation accuracy and boundary localization performance.

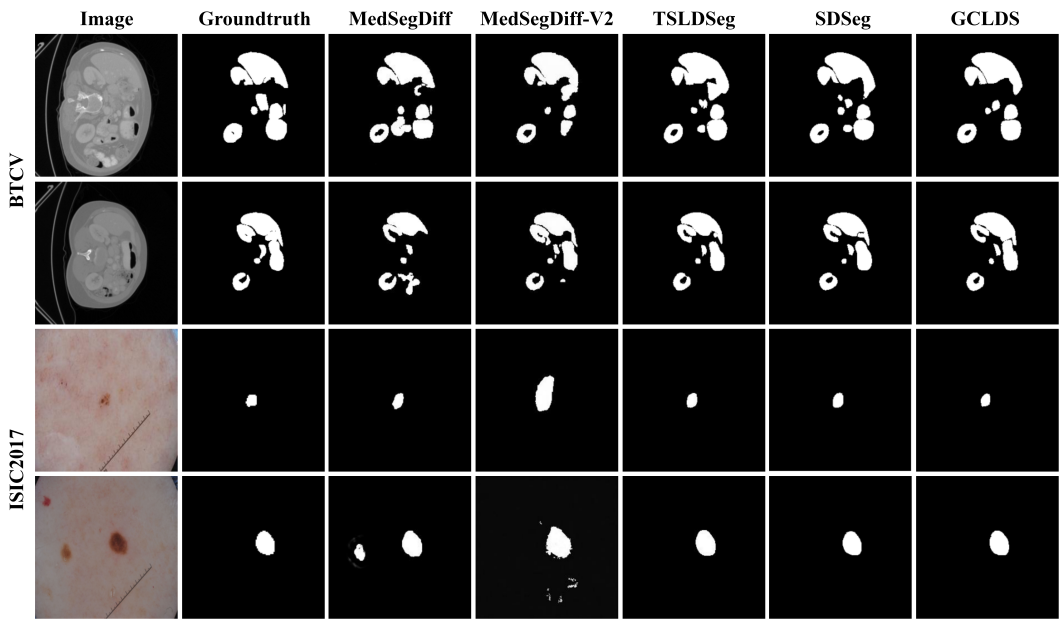

Figure 6: Supplemental visualization comparison on the BTCV and ISIC2017 datasets.

B.2.4 Supplemental Visualization Results

Supplemental segmentation results of GCLDS and other competing methods on the BTCV and ISIC2017 datasets are shown in Figure 6. Across both datasets, GCLDS generates segmentation masks that are more anatomically complete and spatially coherent than competing methods. In challenging abdominal slices on BTCV dataset with low contrast and ambiguous organ boundaries, several state-of-the-art methods exhibit fragmented predictions, missing regions, or obvious false positives, especially for small or adjacent structures. Similarly, on ISIC2017, some methods introduce spurious lesion regions, as observed in the fourth sample of MedSegDiff, while others produce noisy predictions, as in the fourth sample of MedSegDiff-V2, or fail to accurately capture lesion shape and boundary, as seen in the third sample of MedSegDiff-V2. In contrast, GCLDS consistently preserves the global structure of target regions while yielding sharper boundaries and fewer discontinuities. These qualitative results are consistent with the quantitative improvements on Dice, IoU, HD95, and ASD, further demonstrating that GCLDS better captures anatomical priors and boundary details in both organ segmentation and lesion segmentation.

B.2.5 Statistical Significance Analysis

We conducted paired Wilcoxon signed-rank tests between SDSeg and GCLDS on the ISIC2017 and ISIC2018 test sets. As shown in Table 8, GCLDS achieves statistically significant improvements over SDSeg across all evaluation metrics on both datasets. On ISIC2017, Dice and IoU improve by 1.02 and 1.35 percentage points, respectively, while HD95 and ASD decrease by 1.40 and 0.58. On ISIC2018, the corresponding improvements are 2.25 and 3.37 percentage points for Dice and IoU, together with reductions of 2.48 and 1.45 in HD95 and ASD.

All p-values are below 10^{-4} on ISIC2017 and below 10^{-40} on ISIC2018, indicating statistically significant differences between GCLDS and SDSeg under the paired Wilcoxon signed-rank test. Importantly, the test operates on paired predictions from individual test samples rather than on dataset-level averages, providing evidence that the observed performance differences are consistently reflected across the test samples.

The statistical significance is observed on both ISIC2017 ($N = 400$) and the larger ISIC2018 test set ($N = 1000$). The consistent results across the two independent benchmarks further support the robustness of the performance gains. Moreover, improvements are observed simultaneously in

Table 8: Paired statistical significance analysis between SDSeg and GCLDS on the ISIC2017 and ISIC2018 test sets using the Wilcoxon signed-rank test. Δ denotes the change from SDSeg to GCLDS. For Dice and IoU, positive values indicate improvement, whereas negative values indicate improvement for HD95 and ASD.

Dataset	Metric	SDSeg	GCLDS	Δ	p-value
ISIC2017 ($N = 400$)	Dice ↑	0.9086	**0.9188**	+0.0102	3.52×10^{-5}
	IoU ↑	0.8469	**0.8604**	+0.0135	1.90×10^{-5}
	HD95 ↓	11.1936	**9.7914**	-1.4022	1.31×10^{-5}
	ASD ↓	4.2053	**3.6281**	-0.5772	1.70×10^{-4}
ISIC2018 ($N = 1000$)	Dice ↑	0.8743	**0.8967**	+0.0225	1.15×10^{-57}
	IoU ↑	0.7950	**0.8287**	+0.0337	6.20×10^{-59}
	HD95 ↓	19.0570	**16.5820**	-2.4750	3.20×10^{-42}
	ASD ↓	7.7002	**6.2490**	-1.4512	2.23×10^{-58}

both region-based metrics (Dice and IoU) and boundary-based metrics (HD95 and ASD), which is consistent with the design motivation of GCLDS and its channel-spatial disentangled gating mechanism.

C LIMITATIONS

Despite the effectiveness and efficiency of our GCLDS framework, it still has a limitation that it is currently only validated for binary segmentation tasks. Unlike discriminative models that can directly predict multi-class probability maps, our latent diffusion segmentation framework relies on a pre-trained natural image VAE with a fixed 3-channel input/output dimension. It cannot natively support K-class segmentation, which requires K independent channels to model the distribution of each category. We emphasize that this constraint is tied to our choice of the pre-trained VAE, rather than an inherent limitation of the diffusion-based segmentation paradigm itself.

Two straightforward workarounds exist for multi-class extension, but both have notable practical drawbacks. The first is sequential class-conditional inference for each category followed by result aggregation, which linearly increases inference latency with the number of categories, making it not suitable enough for real clinical scenarios. The second is training a custom K-channel VAE for the target segmentation task, but this requires massive amounts of data and high computational costs to train a high-quality VAE model, and needs separate retraining for tasks with different category counts, resulting in poor flexibility and generalizability. In future work, we plan to explore lightweight and native multi-class latent diffusion designs that avoid the above limitations.