LAW & ORDER: Adaptive Spatial Weighting for Medical Diffusion and Segmentation

Anonymous authors
Paper under double-blind review

Abstract

Medical image analysis relies on accurate segmentation, and benefits from controllable synthesis (of new training images). Yet both tasks of the cyclical pipeline face spatial imbalance: lesions occupy small regions against vast backgrounds. In particular, diffusion models have been shown to drift from prescribed lesion layouts, while efficient segmenters struggle on spatially uncertain regions. Adaptive spatial weighting addresses this by learning where to allocate computational resources. This paper introduces a pair of network adapters: 1) Learnable Adaptive Weighter (LAW) which predicts per-pixel loss modulation from features and masks for diffusion training, stabilized via a mix of normalization, clamping, and regularization to prevent degenerate solutions; and 2) Optimal Region Detection with Efficient Resolution (ORDER) which applies selective bidirectional skip attention at late decoder stages for efficient segmentation. Experiments on polyp and kidney tumor datasets demonstrate that LAW achieves 20% FID generative improvement over a uniform baseline (52.28 vs. 65.60), with synthetic data then improving downstream segmentation by 4.9% Dice coefficient (83.2% vs. 78.3%). ORDER reaches 6.0% Dice improvement on MK-UNet (81.3% vs. 75.3%) with 0.56 GFLOPs and just 42K parameters, remaining 730× smaller than the standard nnUNet.

1 Introduction

Medical image segmentation and synthesis are essential for diagnosis, treatment planning, and data augmentation. Accurate segmentation localizes lesions precisely (Isensee et al., 2021; Chen et al., 2024; Wei et al., 2021), while controllable synthesis generates diverse training data to address annotation scarcity (Ho et al., 2020; Song et al., 2021b; Rombach et al., 2022; Shorten & Khoshgoftaar, 2019). The rise of large-scale, multi-modal biomedical datasets further accelerates this progress (Lozano et al., 2025). Recent work makes significant progress: ControlNet-style conditioning improves spatial control in diffusion models (Zhang et al., 2023; Mou et al., 2024), and lightweight U-Nets achieve efficiency gains for segmentation (Valanarasu & Patel, 2022; Tang et al., 2024; Ruan et al., 2023).

Unfortunately, both approaches face a common challenge: **spatial imbalance** (Fig. 1). In medical images, lesions occupy small, irregular regions against vast backgrounds. This imbalance causes diffusion models to drift from prescribed masks during synthesis (Wu et al., 2023; Du et al., 2023), and prevents efficient segmenters from allocating capacity to spatially uncertain regions (Fan et al., 2020; Dong et al., 2023).

Prior work addresses these issues separately. Adaptive diffusion methods reweight losses to emphasize lesions (Du et al., 2023), but use fixed, ratio-based priors that cannot adapt to local feature complexity. Lightweight segmenters reduce parameters by 100× (Valanarasu & Patel, 2022; Ruan et al., 2023; Liu et al., 2024), yet struggle on ambiguous boundaries where heavier models excel (Isensee et al., 2021; Chen et al., 2024). These approaches fail because they rely on static heuristics or sacrifice capacity: fixed priors cannot respond to varying denoising difficulty across image regions, while parameter reduction limits the model's ability to handle spatially ambiguous cases. What is missing is a **common principle** that addresses spatial imbalance across both paradigms.

Lesion: ∼7% of pixels
Background dominates loss

(a) Synthesis: mask drift

Uniform skip fusion
wastes capacity on easy regions

(b) Segmentation: boundary uncertainty

Figure 1: Spatial imbalance in medical imaging. (a) Lesions occupy a small fraction of pixels, causing uniform loss weighting to prioritize background and drift from target masks. (b) Uniform skip connections fail to emphasize uncertain boundary regions where errors concentrate.

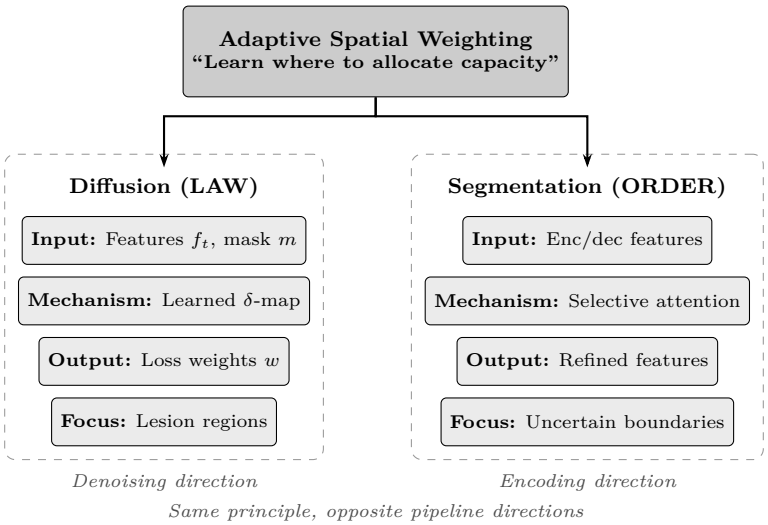


Figure 2: Adaptive spatial weighting as a unifying principle. Both LAW (for diffusion) and ORDER (for segmentation) learn to emphasize spatially sparse, difficult regions, though they operate at opposite ends of the imaging pipeline.

Our Key Insight. We observe that spatial imbalance, a challenge plaguing both generative and discriminative medical imaging, can be addressed through a unified principle: **adaptive spatial weighting**. Rather than treating synthesis and segmentation as separate problems requiring separate solutions, we recognize that both benefit from learning where to allocate computational resources. In diffusion, the model focuses denoising effort on lesion regions. In segmentation, the network emphasizes spatially uncertain areas during feature fusion. Despite operating in opposite directions (denoising versus encoding), both instantiate the same underlying idea: **learn where to learn** (Fig. 2).

Our solution. LAW (Learnable Adaptive Weighter) for mask-conditioned diffusion predicts per-pixel loss modulation from intermediate features and the conditioning mask, building on ratio-based priors but adding learned adjustments stabilized via normalization, clamping, and a Dice regularizer. This yields region-aware denoising without extra supervision (Dhariwal & Nichol, 2021; Ho & Salimans, 2021). ORDER (Optimal Region Detection with Efficient Resolution) applies the same idea to segmentation by keeping MK-UNet's multi-kernel encoder (Rahman & Marculescu, 2025) and inserting lightweight bidirectional attention with learned confidence gates on the final two skips. Decoder–encoder similarity drives the bidirectional attention maps, and a learned confidence gate scales both residual updates before fusion, so ORDER focuses capacity where semantics are richest while staying within a 42K-parameter budget.

Supervision \ *Weight source*	Fixed	Learned
Uniform	ControlNet	—
Ratio	ArSDM	**LAW (ours)**

(a) Diffusion weights

Scope \ *Direction*	Unidir.	Bidir.
All	Attn U-Net	Full BiAttn
Selective	Late-stage	**ORDER (ours)**

(b) Segmentation attention

Figure 3: Design space for adaptive spatial mechanisms. (a) LAW combines learned feature-dependent weights with ratio-based priors. (b) ORDER applies bidirectional attention selectively at late decoder stages. Shaded cells indicate our contributions.

1.1 Contributions

The contributions of this work are as follows:

- LAW for diffusion (synthesis) extends ratio-based priors with learned, feature-dependent adjustments, stabilized via normalization, clamping, and Dice regularization. This achieves a 20% FID generative improvement over a uniform baseline ControlNet (52.28 vs. 65.60) and then a 6% Dice improvement (83.2 vs. 78.3) when generated images are used *downstream* to train the standard nnUNet.
- ORDER for efficient segmentation couples selective bidirectional skip attention with a 42K-parameter backbone, achieving 81.3% Dice and 81.7% IoU on Polyps-a +8% Dice gain (81.3 vs. 75.3) over MK-UNet while remaining 730× smaller than nnUNet.

See Fig. 3 for a visual placement of our scope. The methods build on latent diffusion (Rombach et al., 2022), ControlNet conditioning (Zhang et al., 2023), and efficient U-Net designs (Valanarasu & Patel, 2022), but the adaptive weighting mechanisms are orthogonal and broadly applicable. Comprehensive experiments on Polyps (Ali et al., 2023) and KiTS19 (Heller et al., 2021) datasets show that adaptive spatial weighting benefits both paradigms.

2 Related Work

Three research threads inform this work: spatial control in diffusion models, efficient segmentation architectures, and adaptive mechanisms for visual learning. Each addresses spatial imbalance differently, yet none connect these insights across generative and discriminative paradigms.

2.1 Spatial Control in Diffusion Models

Denoising diffusion probabilistic models (Ho et al., 2020; Song et al., 2021b) provide powerful generative frameworks; latent diffusion (Rombach et al., 2022) and accelerated samplers such as DDIM (Song et al., 2021a; Naman et al., 2026) enable high-resolution, tractable synthesis. In medical imaging they mitigate data scarcity by creating supervisory data while supporting anomaly detection, temporal modeling, and MRI reconstruction (Wu et al., 2023; Nguyen et al., 2023; Qiu et al., 2025a; Rahman et al., 2023; Kim & Ye, 2022; Wolleb et al., 2022; Pinaya et al., 2022; Chung & Ye, 2022).

Spatial control began with ControlNet (Zhang et al., 2023), which copies frozen encoders with zero-initialized connections for mask conditioning. Lighter adapters such as T2I (Mou et al., 2024) and ControlNet++ (Li et al., 2024) further improve efficiency. Yet medical conditioning data is limited and lesions occupy irregular, tiny regions, so background pixels dominate the loss and models drift from the prescribed masks.

Adaptive spatial weighting tackles this drift. ArSDM (Du et al., 2023) reweights the loss with a mask-derived ratio prior, but the fixed heuristic cannot react to local denoising difficulty and can suppress backgrounds

excessively. Qiu et al. (Qiu et al., 2025b) train a teacher-student pair instead, raising training cost and tying the weights to the teacher's biases. LAW continues this line by adapting the prior with intermediate features, clamps to avoid degeneration, and removes the need for auxiliary pretrained segmenters.

2.2 Efficient Segmentation Architectures

Medical segmentation has progressively shifted from heavy encoder-decoders to compact designs. U-Net (Ronneberger et al., 2015) popularized skip-connected decoders, with successors such as U-Net++ (Zhou et al., 2018), V-Net (Milletari et al., 2016), Attention U-Net (Oktay et al., 2020), and transformer hybrids (Chen et al., 2024; Cao et al., 2022; Azad et al., 2022; Heidari et al., 2023) trading accuracy for substantial computation.

Efficiency-focused architectures draw on mobile design principles (Howard et al., 2017; Sandler et al., 2018). UNeXt (Valanarasu & Patel, 2022) mixes convolutions and MLPs, CMUNeXt (Tang et al., 2024) uses large kernels, EGE-UNet (Ruan et al., 2023) leverages group convolutions, and MK-UNet (Rahman & Marculescu, 2025) attains 27K parameters via multi-kernel inverted residual blocks.

Despite these innovations, lightweight models still fail on ambiguous boundaries even when they slash parameter counts by 100×. Uniform skip fusions treat every region identically, so scarce capacity is wasted on easy background pixels instead of lesion edges. ORDER keeps MK-UNet's compact encoder but adds targeted bidirectional skip attention, computing one similarity matrix reused in both directions (Kitaev et al., 2020) and activating it only in the final decoder stages where semantics peak.

2.3 Adaptive Mechanisms in Visual Models

Attention and reweighting mechanisms are longstanding tools in vision. Channel/spatial attention (Woo et al., 2018; Oktay et al., 2020; Hu et al., 2018) teaches discriminative models to emphasize salient features, while classifier or classifier-free guidance (Dhariwal & Nichol, 2021; Ho & Salimans, 2021) tune diffusion globally rather than spatially.

Medical synthesis studies extend these ideas with mask-aware loss weighting (Du et al., 2023; Qiu et al., 2025b), yet they confine the adaptivity to the diffusion loop. The same spatial imbalance afflicts segmentation, implying a unified treatment would be more effective. LAW and ORDER therefore deploy adaptive weighting across both paradigms, demonstrating that learning *where* to allocate capacity simultaneously boosts generative fidelity and discriminative accuracy.

3 Method

Two instantiations of adaptive spatial weighting are presented: LAW (Learnable Adaptive Weighter) for mask-conditioned diffusion (Section 3.1) and ORDER (Optimal Region Detection with Efficient Resolution) for efficient segmentation (Section 3.2). Both share the principle of learning where to allocate computational resources, but operate in complementary settings.

3.1 LAW: Learnable Adaptive Weighting for Diffusion

LAW augments ControlNet-based diffusion with feature-dependent spatial weighting. We first review mask-conditioned diffusion for completeness, then describe the teacher-student setup that supplies supervision before detailing the adaptive weighting module. See Fig. 4 for the final pipeline.

3.1.1 Preliminaries: Mask-Conditioned Diffusion

LAW builds on latent diffusion models (Rombach et al., 2022) that denoise in a compressed latent space. Given an image $\mathbf{x}_0$, a pretrained encoder $\mathcal{E}$ maps it to latent $\mathbf{z}_0 = \mathcal{E}(\mathbf{x}_0)$. The forward diffusion process adds Gaussian noise over T steps:

$$\mathbf{z}_t = \sqrt{\bar{\alpha}_t}\mathbf{z}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{1}$$

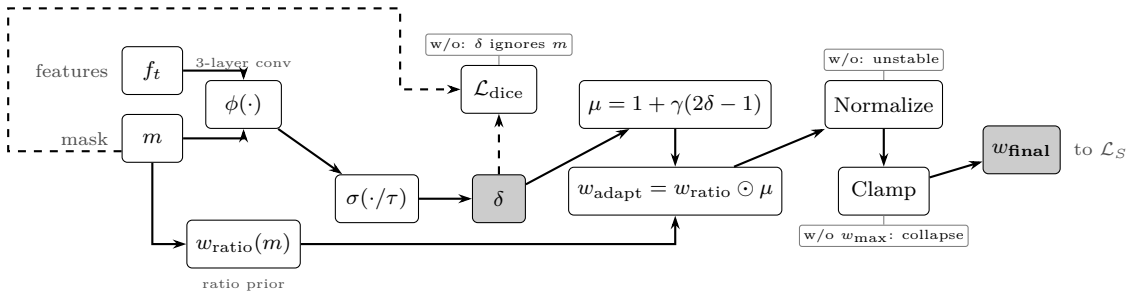


Figure 4: LAW weight computation with stabilization mechanisms. The learned delta map δ modulates a ratio-based prior, with normalization and clamping ensuring training stability. Annotations indicate failure modes when each component is ablated (cf. Table 4).

where $\bar{\alpha}_t$ controls the noise schedule. A denoising network ϵ_θ predicts the noise given $\mathbf{z}_t$, timestep t, and conditioning. For mask-conditioned synthesis, ControlNet (Zhang et al., 2023) based models inject spatial control via a segmentation mask $\mathbf{m} \in [0,1]^{H \times W}$ by adding trainable copies of the encoder with zero-initialized connections. The standard training objective minimizes:

$$\mathcal{L}_{\text{base}} = \mathbb{E}_{t,\mathbf{z}_0,\boldsymbol{\epsilon}} \left[\|\epsilon_\theta(\mathbf{z}_t, t, \mathbf{m}) - \boldsymbol{\epsilon}\|_2^2 \right]. \tag{2}$$

In medical images, lesions occupy small regions (often <10% of pixels) against vast backgrounds. Uniform loss weighting causes models to prioritize background reconstruction, leading to poor lesion-mask alignment. Prior work (Du et al., 2023; Qiu et al., 2025b) addresses this via ratio-based spatial weighting:

$$w_{\text{ratio}}(\mathbf{m}) = \mathbf{m} \cdot (1 - r) + \bar{\mathbf{m}} \cdot r, \tag{3}$$

where $\bar{\mathbf{m}} = 1 - \mathbf{m}$ and $r = \frac{|\mathbf{m}|}{HW}$ is the lesion-area ratio. This linear prior balances lesion and background contributions based on their relative coverage, but remains oblivious to local feature complexity or denoising difficulty.

3.1.2 Adaptive Weight Learning

LAW learns spatially adaptive weights from the student features and the conditioning mask. Let $\mathbf{f}_t = \hat{\boldsymbol{\epsilon}}_t^S$ denote the student prediction. LAW predicts a delta map $\boldsymbol{\delta} \in [0,1]^{h \times w}$ that modulates the ratio-based prior:

$$\boldsymbol{\delta} = \sigma\left(\frac{1}{\tau} \cdot \phi(\mathbf{f}_t, \mathbf{m})\right), \tag{4}$$

where ϕ is a lightweight convolutional network (3 layers, 32 hidden channels), σ denotes the sigmoid function, and $\tau = 3.0$ is a temperature parameter. The delta map adjusts the base weights via:

$$\boldsymbol{\mu} = 1.0 + \gamma \cdot (2\boldsymbol{\delta} - 1.0), \tag{5}$$

where $\gamma = 0.2$ in our implementation, producing multipliers confined to $[0.8, 1.2]$ after clamping. The final adaptive weights combine the ratio prior with learned adjustments:

$$w_{\text{adapt}} = w_{\text{ratio}}(\mathbf{m}) \odot \boldsymbol{\mu}, \tag{6}$$

where $\odot$ denotes element-wise multiplication. To ensure stability and prevent model collapse, LAW normalizes to preserve mean magnitude and applies both minimum and maximum weight bounds:

$$w_{\text{final}} = \text{clamp}\left(\frac{w_{\text{adapt}}}{\text{mean}(w_{\text{adapt}})}, w_{\min}, w_{\max}\right), \tag{7}$$

with $w_{\min} = 10^{-3}$ and $w_{\max} = 2.0$. The maximum bound is critical to prevent extreme weights that cause the model to ignore background regions entirely, leading to degenerate solutions. The spatially weighted student loss becomes:

$$\mathcal{L}_S = \mathbb{E}_{t,\mathbf{z}_0,\boldsymbol{\epsilon}} \left[w_{\text{final}} \odot \|\hat{\boldsymbol{\epsilon}}_t^S - \boldsymbol{\epsilon}\|_2^2 \right]. \tag{8}$$

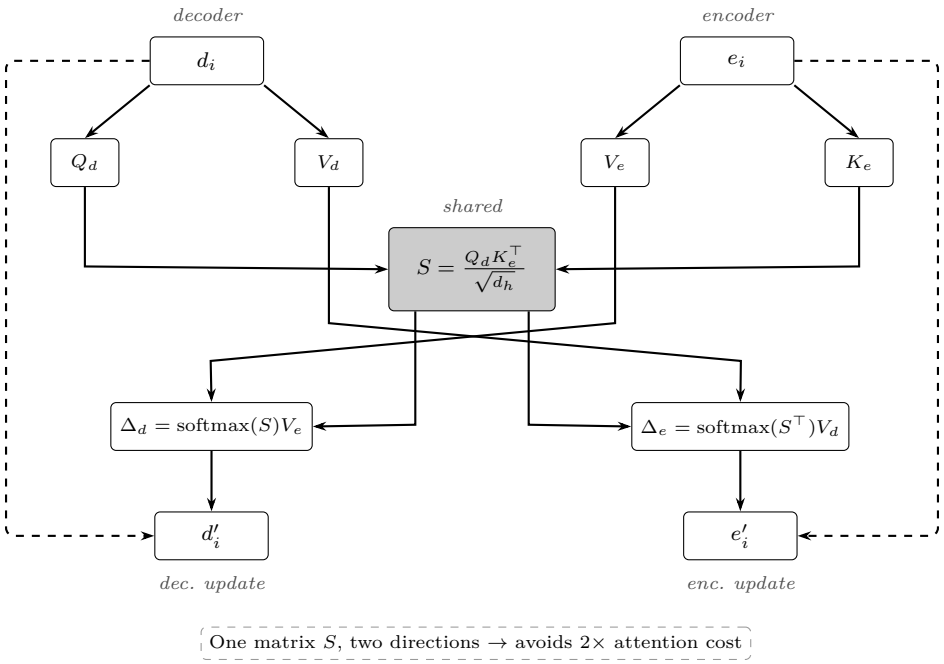


Figure 5: Bidirectional skip attention in ORDER. A single similarity matrix S is computed from decoder queries and encoder keys, then reused for both dec→enc and enc→dec attention via S and $S^\top$, halving the attention cost compared to separate computations.

To prevent degenerate solutions where $\boldsymbol{\delta}$ ignores the mask, LAW adds a Dice-style regularizer that encourages spatial alignment:

$$\mathcal{L}_{\text{dice}} = 1 - \frac{2\sum_{i,j} \delta_{ij} m_{ij} + \epsilon_s}{\sum_{i,j} \delta_{ij} + \sum_{i,j} m_{ij} + \epsilon_s}, \tag{9}$$

where $\epsilon_s = 10^{-6}$ prevents division by zero. In addition to $\mathcal{L}_S$ we can supervise the teacher prediction and optionally distill it back to the student:

$$\mathcal{L}_T = \mathbb{E}_{t,\mathbf{z}_0,\boldsymbol{\epsilon}} \left[\|\hat{\boldsymbol{\epsilon}}_t^T - \boldsymbol{\epsilon}\|_2^2 \right], \tag{10}$$

$$\mathcal{L}_{S \leftarrow T} = \mathbb{E}_{t,\mathbf{z}_0} \left[w_{\text{final}} \odot \|\hat{\boldsymbol{\epsilon}}_t^S - \hat{\boldsymbol{\epsilon}}_t^T\|_2^2 \right]. \tag{11}$$

The overall training objective follows Qiu et al.'s design for ratio based weighing with LAW based losses as

$$\mathcal{L}_{\text{total}} = \mathcal{L}_S + \beta_T \mathcal{L}_T + \beta_D \mathcal{L}_{S \leftarrow T} + \lambda_{\text{dice}} \mathcal{L}_{\text{dice}}, \tag{12}$$

where $\lambda_{\text{dice}} = 1.0$ with teacher and distillation weights fixed as $\beta_T = \beta_D = 0.05$.

3.2 ORDER: Adaptive Efficient Segmentation

ORDER extends lightweight MK-UNet baselines with selective bidirectional attention and confidence-aware fusion. We summarize the backbone before detailing how the bidirectional skip attention operates (further illustrated in Fig. 5).

3.2.1 Preliminaries: Multi-Kernel Lightweight Architecture

ORDER builds on the ultra-lightweight MK-UNet architecture (Rahman & Marculescu, 2025), which achieves 27K parameters through multi-kernel inverted residual (MKIR) blocks. The encoder consists of five stages with multi-kernel depthwise convolutions that capture features at multiple scales. Each stage i produces feature maps $\mathbf{e}_i \in \mathbb{R}^{C_i \times H_i \times W_i}$ with progressively reduced spatial resolution and increased channels.

Following (Rahman & Marculescu, 2025), each encoder stage uses MKIR blocks that expand channels, apply parallel depthwise convolutions with multiple kernel sizes $\mathcal{K} = \{1, 3, 5\}$, and project back:

$$\text{MKIR}(\mathbf{x}) = \mathbf{x} + \text{Proj}\left(\sum_{k \in \mathcal{K}} \text{DWConv}_k(\text{Expand}(\mathbf{x}))\right), \tag{13}$$

where Expand uses 1×1 convolution with expansion factor 2, DWConv_k applies depthwise convolution with kernel k, and Proj projects back to the original channels. This design achieves efficiency through depthwise separable convolutions while capturing multi-scale context. The decoder mirrors the encoder structure with simple additive skip connections.

While MK-UNet achieves remarkable parameter efficiency, it struggles on spatially ambiguous regions where lesion boundaries are unclear. This limitation stems from uniform skip connections that treat all spatial regions equally, failing to emphasize uncertain areas that require more representational capacity.

3.2.2 Selective Bidirectional Skip Attention

ORDER adds bidirectional cross attention to the final decoder stages where semantic information is most critical. Inspired by Reformer (Kitaev et al., 2020), the module computes one similarity matrix that is reused in both directions, avoiding the quadratic cost of separate attentions. Let $\mathbf{d}_i$ denote decoder features and $\mathbf{e}_i$ the corresponding encoder skip. In the implementation, projections are

$$\begin{aligned} &\mathbf{Q}_d = \text{Conv}_Q(\mathbf{d}_i), \qquad \mathbf{K}_e = \text{Conv}_K(\mathbf{e}_i), \qquad \mathbf{V}_d = \text{Conv}_V(\mathbf{d}_i), \\ &\mathbf{V}_e = \text{Conv}'_V(\mathbf{e}_i). \end{aligned} \tag{14}$$

We first form a shared similarity matrix using the decoder queries and encoder keys:

$$\mathbf{S} = \frac{\mathbf{Q}_d \mathbf{K}_e^\top}{\sqrt{d_h}}, \tag{15}$$

where d_h is the per-head dimension. Decoder-side updates attend to encoder values, while encoder-side updates re-use the transposed weights to attend back to decoder values:

$$\begin{aligned} \Delta_d &= \text{softmax}(\mathbf{S})\mathbf{V}_e, \\ \Delta_e &= \text{softmax}(\mathbf{S}^\top)\mathbf{V}_d. \end{aligned} \tag{16}$$

Both feature streams are enhanced residually:

$$\begin{aligned} \mathbf{d}'_i &= \mathbf{d}_i + c_i \Delta_d, \\ \mathbf{e}'_i &= \mathbf{e}_i + c_i \Delta_e. \end{aligned} \tag{17}$$

where $c_i \in [0, 1]$ is a confidence gate predicted from globally pooled decoder and encoder features. The decoder then uses the fused skip $\mathbf{d}'_i + \mathbf{e}'_i$. For other decoder stages, simple additive skip connections suffice. The final segmentation output is produced via 1×1 convolution followed by sigmoid activation. Training optimizes binary cross-entropy combined with Dice loss:

$$\mathcal{L}_{\text{seg}} = \mathcal{L}_{\text{BCE}}(\hat{\mathbf{y}}, \mathbf{y}) + \mathcal{L}_{\text{Dice}}(\hat{\mathbf{y}}, \mathbf{y}), \tag{18}$$

where $\hat{\mathbf{y}}$ denotes the predicted mask and $\mathbf{y}$ denotes the ground truth. This combination handles class imbalance while encouraging spatial overlap.

4 Experiments

Experiments evaluate LAW and ORDER on two medical imaging tasks: mask-conditioned synthesis and segmentation. Results validate that adaptive spatial weighting improves both generative and discriminative models.

Table 1: Synthesis quality comparison. LAW achieves superior quality over all baselines, with 20% FID improvement over uniform baseline and 22% improvement over adaptive distillation.

Method	FID↓	CLIP-I↑	FID↓	CLIP-I↑
	Polyps		KiTS19	
ControlNet (Uniform) (Zhang et al., 2023)	65.60	0.884	69.24	0.833
ArSDM (Ratio-based) (Du et al., 2023)	98.08	0.845	104.12	**0.852**
Adaptive Distillation (Qiu et al., 2025b)	66.60	**0.901**	70.78	0.839
LAW (ours)	**52.28**	0.893	**63.12**	0.845

4.1 Experimental Setup

Datasets. Experiments use two publicly available medical datasets: (1) Polyps, combining Kvasir, CVC-ClinicDB, CVC-ColonDB, CVC-300, and ETIS-LaribPolypDB and (2) KiTS19 (Heller et al., 2021). For LAW training, we follow the same recipe as Qiu et al. (2025b). For ORDER training, we follow the recipe used in Fan et al. (2020). Both datasets are widely used benchmarks in medical image analysis (Antonelli et al., 2022). Both datasets exhibit severe spatial imbalance with lesions occupying <10% of pixels. Mean Dice (mDice) and mean IoU (mIoU) are reported, computed as the average of per-image metrics across all test samples.

Implementation details. For diffusion, LAW builds on Stable Diffusion v1.5 with ControlNet conditioning. LAW uses a 3-layer convolutional network with 32 hidden channels. Training runs for 10 epochs (approximately 500 steps) with batch size 4, learning rate 10^{-5}, $\lambda_{\text{dice}} = 1.0$, $w_{\min} = 10^{-3}$, and $w_{\max} = 2.0$. The maximum weight bound is critical for training stability. For segmentation, ORDER adopts channels $[4, 8, 16, 24, 32]$ with bidirectional attention applied to the final two decoder stages and a global confidence gate inside skip fusion; optimization uses BCE+Dice. Training uses 100 epochs on 512×512 crops with batch size 16, AdEMAMix (Pagliardini et al., 2025) optimizer (learning rate 10^{-4}), cosine annealing, and random flips/rotations.

Baselines. For diffusion, comparisons include: (1) ControlNet with uniform loss (Zhang et al., 2023), (2) ArSDM with ratio-based weighting (Du et al., 2023), and (3) adaptive distillation (Qiu et al., 2025b). These represent the progression from no spatial weighting (uniform) to fixed spatial weighting (ratio-based) to learned weighting with teacher models (adaptive distillation), allowing us to isolate the contribution of feature-dependent weight learning. For segmentation, comparisons include lightweight models: MK-UNet (Rahman & Marculescu, 2025), UNeXt (Valanarasu & Patel, 2022), EGE-UNet (Ruan et al., 2023), and the heavy baseline nnUNet (Isensee et al., 2021). These span the efficiency-accuracy spectrum, from ultra-lightweight (27K-1.3M parameters) to heavy (31M parameters), enabling evaluation of ORDER's accuracy-efficiency tradeoff.

4.2 Diffusion Results

Image quality. Table 1 shows synthesis quality metrics. LAW achieves the best performance across both datasets, with FID of 52.28 on Polyps (20% improvement over uniform baseline's 65.60) and 63.12 on KiTS19 (9% improvement over uniform's 69.24). Compared to adaptive distillation, LAW achieves 22% better FID (52.28 vs 66.60) on Polyps while maintaining comparable CLIP-I scores (0.893 vs 0.901). LAW's learned, feature-dependent weighting produces higher-quality synthesis than both fixed ratio-based priors and teacher-student distillation approaches. This reveals that learned adaptation from intermediate features is more effective than fixed heuristics or teacher-guided learning, suggesting that local feature complexity varies significantly across denoising steps and spatial regions.

Downstream segmentation. To evaluate lesion-mask alignment, nnUNet is trained on synthetic data generated by different methods. Table 2 shows that images generated with LAW achieve 83.2% Dice on Polyps and 98.0% Dice on KiTS19. Compared to training on real data only, LAW-generated synthetic data improves Dice by 4.9% on Polyps and 1.5% on KiTS19. Compared to uniform weighting, LAW improves by

Table 2: Downstream segmentation when training nnUNet on synthetic data. LAW-generated data achieves the best downstream performance, improving Dice 4.9% on Polyps and 1.5% on KiTS19 over real data alone.

Synthetic Data	mDice↑	mIoU↑	mDice↑	mIoU↑
	Polyps		KiTS19	
Real Only	78.3	70.9	96.5	93.4
+ ControlNet (Uniform)	75.9	68.8	96.1	93.2
+ Adaptive Distillation	79.5	72.7	97.9	**96.0**
+ LAW (ours)	**83.2**	**84.5**	**98.0**	95.5

7.3% on Polyps and 1.9% on KiTS19, while outperforming adaptive distillation by 3.7% and 0.1% respectively. This demonstrates that learned adaptive weighting produces better-aligned synthetic training data that enhances downstream segmentation performance. Importantly, the lack of correlation between FID scores and downstream segmentation performance (ratio-based has worst FID but better segmentation than uniform) reveals that lesion-mask alignment, not overall image quality, is the critical factor for effective data augmentation in medical imaging.

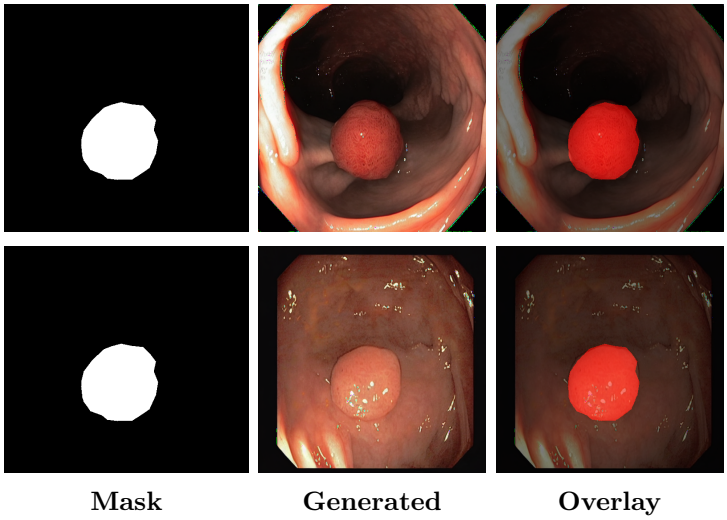


Figure 6: Qualitative comparison of mask-conditioned synthesis. **Top row**: Adaptive Distillation baseline. **Bottom row**: LAW (ours). LAW produces more photorealistic images with improved anatomical fidelity and texture detail. The overlay column demonstrates superior spatial adherence to the conditioning mask, with generated lesions exhibiting natural tissue appearance and realistic lighting characteristics that closely match the prescribed layout.

Qualitative analysis. Fig. 6 shows visual comparisons of synthesis quality. Uniform weighting produces images with poor mask adherence and unrealistic texture, reflected in its higher FID (65.60). Ratio-based weighting improves alignment but generates artifacts and unnatural appearance, resulting in the worst FID (98.08). Adaptive distillation achieves reasonable quality (FID 66.60), incorporating teacher model training. LAW produces the most photorealistic results with natural tissue appearance and accurate spatial correspondence (FID 52.28). Learned feature-dependent weighting generates clinically plausible images with superior visual fidelity. The visual comparison reveals that fixed ratio-based priors can over-emphasize lesions at the expense of background coherence, while LAW balances both through adaptive, context-aware weighting.

Table 3: Segmentation performance comparison. ORDER achieves the best accuracy-efficiency tradeoff.

Category	Method	Params	FLOPs	mDice↑	mIoU↑
Standard	nnUNet	31M	50G	**85.7**	**83.4**
Edge	UNeXt	1.3M	2G	74.9	70.1
	EGE-UNet	0.5M	1G	74.1	69.2
	MK-UNet	27K	0.15G	75.3	71.2
	ORDER (ours)	42K	0.56G	**81.3**	**81.7**

4.3 Segmentation Results

Accuracy vs efficiency. Table 3 compares segmentation performance. ORDER delivers 6.0% higher mean test Dice than MK-UNet (27K params) while lifting mIoU to 81.7%. Compared to UNeXt (1.3M params), ORDER improves Dice by 6.4% with 30× fewer parameters. Against the heavy baseline nnUNet (31M params), ORDER reaches a 4.4% lower Dice (81.3% vs. 85.7%) but uses 730× fewer parameters. Selective bidirectional attention balances accuracy and efficiency even in this ultra-compact setting. This demonstrates that the bottleneck in lightweight segmentation is not parameter count per se, but rather how those parameters are allocated: selective attention at semantically critical stages provides more value than uniform capacity distribution across all layers.

Computational cost. ORDER requires 0.56 GFLOPs at 256^2 resolution versus MK-UNet's 0.15 GFLOPs, yet still adds 6.0% Dice. Selective bidirectional attention and confidence-aware losses therefore improve accuracy with a modest compute budget.

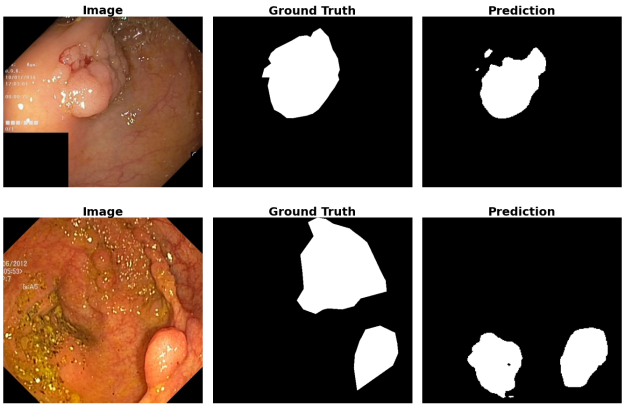


Figure 7: Qualitative segmentation results from ORDER on challenging polyp cases. Each row shows: input image, ground truth mask, and predicted segmentation. ORDER produces accurate boundaries even on difficult cases with varying polyp sizes and shapes, demonstrating the effectiveness of selective bidirectional attention.

Qualitative analysis. Fig. 7 shows segmentation results from ORDER on challenging polyp cases from the Kvasir test set. ORDER accurately segments polyps of varying sizes and shapes, producing clean boundaries even in cases with complex morphology. The first example shows successful detection of a large polyp with irregular boundaries, while the second demonstrates accurate segmentation of a smaller lesion with subtle color variations.

4.4 Ablation Studies

LAW components. Table 4 shows how each component of LAW builds upon the adaptive distillation baseline of Qiu et al. Injecting a learned delta map without safeguards destabilizes training, indicating

Table 4: Ablation of LAW components. Each component contributes to stability and performance on Polyps dataset.

Configuration	FID↓	CLIP-I↑	mDice↑	Stability
Adaptive Distillation (Qiu et al., 2025b)	66.60	**0.901**	95.2	Stable
+ Delta map (no norm)	70.34	0.854	93.3	Unstable
+ Normalization	70.11	0.876	93.9	Stable
+ Min-weight clamp	66.51	0.881	94.3	Stable
+ Dice regularizer (LAW)	**52.28**	0.893	**96.1**	Stable

Table 5: Ablation of ORDER components on Polyps dataset. Selective attention provides the best tradeoff.

Configuration	Params	mDice↑	mIoU↑
MK-UNet (base)	27K	75.3	71.2
+ BiAttn (all stages, ORDER)	50K	77.3	73.3
+ BiAttn (selective, ORDER)	42K	**81.3**	**81.7**

that feature-dependent weights alone are insufficient. Normalization and the min-weight clamp progressively rein in the weights and recover the teacher baseline. The Dice regularizer then enforces alignment between δ and the mask, delivering the full LAW result. We also tried applying Dice directly to the ratio prior (adaptive distillation, no delta map); it improved mDice by only 0.2% while FID stayed above 95, so it is omitted. These trends show that LAW's gains arise from the combined effect of learned weights together with normalization, clamping, and Dice alignment, not from any single component.

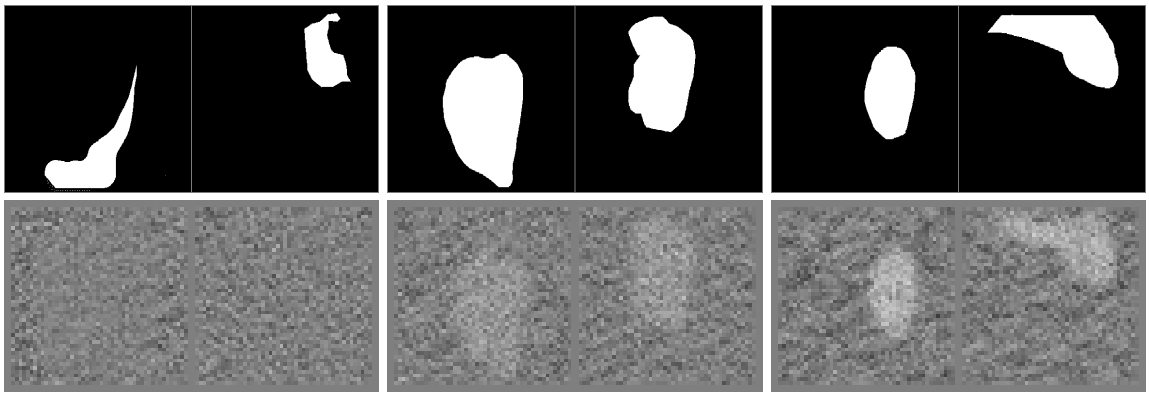
ORDER components. Table 5 ablates ORDER components. Starting from MK-UNet, adding bidirectional attention to all stages improves Dice by 2.0% but increases parameters to 50K. Applying attention selectively to the final two stages yields the best tradeoff, adding 6.0% Dice and 10.5% IoU with only 42K parameters. Late-stage attention is most effective while maintaining parameter efficiency. This reveals a key insight: early encoder features capture low-level textures that benefit less from adaptive attention, while late decoder features contain semantic information where selective emphasis on uncertain regions provides maximum value.

Table 6: Ablation of attention stage selection. Late decoder stages provide the best tradeoff.

Stages with BiAttn	Params	mDice↑	mIoU↑
None (MK-UNet)	27K	75.3	71.2
Stages 3-4 (early)	35K	78.1	74.8
Stages 2-3 (mid)	38K	78.3	75.1
Stages 0-1 (ours)	42K	**81.3**	**81.7**
All stages	50K	79.2	75.9

Attention stage selection. Table 6 varies which decoder stages use bidirectional attention. Early stages (stages 3-4; high number is due to mirroring the stages of the encoder) contribute modest gains (+2.8% Dice) as features are still low-level, and mid stages add +3.0% Dice. Late stages (stages 0-1) provide the largest jump (+6.0% Dice, +10.5% IoU) with only a 55% parameter increase over MK-UNet, whereas enabling attention at all stages raises parameters by 85% for smaller gains. This confirms that concentrating attention on semantic decoder stages yields the best accuracy-efficiency balance. Interestingly, applying attention to all stages actually degrades performance compared to selective late-stage attention, suggesting that adaptive mechanisms can interfere with low-level feature extraction when applied indiscriminately.

Attention visualization. Fig. 8 visualizes the evolution of learned attention maps from LAW during training. Initially (step 0), attention is uniformly distributed. As training progresses, the delta maps



Step 0 **Step 250** **Step 500**

Figure 8: Visualization of learned attention maps from LAW during training. **Top row**: Conditioning masks. **Bottom row**: Learned attention maps (delta maps). At step 0, attention is uniform. By step 250, the model begins focusing on lesion regions. At step 500, attention strongly emphasizes lesion boundaries and small lesions, validating adaptive spatial weighting.

increasingly attend to lesion boundaries and small lesions, validating that LAW learns to emphasize difficult regions. Background regions receive lower attention, confirming that denoising effort focuses where needed.

5 Conclusion

Adaptive spatial weighting helps both generative and discriminative medical-imaging models by directing attention to the regions that are hardest to denoise or segment. LAW shows that even very lightweight, feature-aware weighting can outperform fixed heuristics, while ORDER demonstrates that selectively applying attention only where it matters can yield strong accuracy–efficiency benefits. Together, they highlight that adaptive mechanisms need not be complex to be effective and that learned strategies generally surpass hand-designed ones. These insights suggest broader applicability to other vision tasks, with promising future directions including multi-scale adaptive designs and extensions beyond medical imaging.

References

Sharib Ali, Debesh Jha, Noha Ghatwary, Stefano Realdon, Renato Cannizzaro, Osama E Salem, Dominique Lamarque, Christian Daul, Michael A Riegler, Kim V Anonsen, Andreas Petlund, Pål Halvorsen, Jens Rittscher, Thomas de Lange, and James E East. A multi-centre polyp detection and segmentation dataset for generalisability assessment. *Scientific Data*, 10(1):75, 2023.

Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A. Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M. Summers, Bram van Ginneken, Michel Bilello, Patrick Bilic, Patrick F. Christ, Richard K. G. Do, Marc J. Gollub, Stephan H. Heckers, Henkjan Huisman, William R. Jarnagin, Maureen K. McHugo, Sandy Napel, Jennifer S. Golia Pernicka, Kawal Rhode, Catalina Tobon-Gomez, Eugene Vorontsov, James A. Meakin, Sebastien Ourselin, Manuel Wiesenfarth, Pablo Arbeláez, Byeonguk Bae, Sihong Chen, Laura Daza, Jianjiang Feng, Baochun He, Fabian Isensee, Yuanfeng Ji, Fucang Jia, Ildoo Kim, Klaus Maier-Hein, Dorit Merhof, Akshay Pai, Beomhee Park, Mathias Perslev, Ramin Rezaiifar, Oliver Rippel, Ignacio Sarasua, Wei Shen, Jaemin Son, Christian Wachinger, Liansheng Wang, Yan Wang, Yingda Xia, Daguang Xu, Zhanwei Xu, Yefeng Zheng, Amber L. Simpson, Lena Maier-Hein, and M. Jorge Cardoso. The medical segmentation decathlon. *Nature Communications*, 13(1):4128, 2022.

Reza Azad, Moein Heidari, Moein Shariatnia, Ehsan Khodapanah Aghdam, Sanaz Karimijafarbigloo, Ehsan Adeli, and Dorit Merhof. TransDeepLab: Convolution-free transformer-based DeepLab v3+ for medical

image segmentation. In *Predictive Intelligence in Medicine (PRIME)*, volume 13564 of *Lecture Notes in Computer Science*, pp. 91–102. Springer, 2022.

Huazhu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-Unet: Unet-like pure transformer for medical image segmentation. In *European Conference on Computer Vision (ECCV) Workshops*, 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 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, 2024.

Hyungjin Chung and Jong Chul Ye. Score-based diffusion models for accelerated MRI. *Medical Image Analysis*, 80:102479, 2022.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 8780–8794, 2021.

Bo Dong, Wenhai Wang, Deng-Ping Fan, Jinpeng Li, Huazhu Fu, and Ling Shao. Polyp-PVT: Polyp segmentation with pyramid vision transformers. *CAAI Artificial Intelligence Research*, 2(2):81–88, 2023.

Yuhao Du, Yuncheng Jiang, Shuangyi Tan, Xusheng Wu, Qi Dou, Zhen Li, Guanbin Li, and Xiang Wan. ArSDM: Colonoscopy images synthesis with adaptive refinement semantic diffusion models. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 339–349. Springer, 2023.

Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. PraNet: Parallel reverse attention network for polyp segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 263–273. Springer, 2020.

Moein Heidari, Amirhossein Kazerouni, Milad Soltany, Reza Azad, Ehsan Khodapanah Aghdam, Julien Cohen-Adad, and Dorit Merhof. HiFormer: Hierarchical multi-scale representations using transformers for medical image segmentation. In *Proceedings of IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 6202–6212. IEEE, 2023.

Nicholas Heller, Fabian Isensee, Klaus H. Maier-Hein, Xiaoshuai Hou, Chunmei Xie, Fengyi Li, Yang Nan, Guangrui Mu, Zhiyong Lin, Miofei Han, Guang Yao, Yaozong Gao, Yao Zhang, Yixin Wang, Feng Hou, Jiawei Yang, Guangwei Xiong, Jiang Tian, Cheng Zhong, Jun Ma, Jack Rickman, Joshua Dean, Bethany Stai, Resha Tejpaul, Makinna Oestreich, Paul Blake, Heather Kaluzniak, Shaneabbas Raza, Joel Rosenberg, Keenan Moore, Edward Walczak, Zachary Rengel, Zach Edgerton, Ranveer Vasdev, Matthew Peterson, Sean McSweeney, Sarah Peterson, Arveen Kalapara, Niranjan Sathianathen, Nikolaos Papanikolopoulos, and Christopher Weight. The state of the art in kidney and kidney tumor segmentation in contrast-enhanced ct imaging: Results of the kits19 challenge. *Medical Image Analysis*, 67, 2021.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. MobileNets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. URL `https://arxiv.org/abs/1704.04861`.

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7132–7141. IEEE, 2018.

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.

Boah Kim and Jong Chul Ye. Diffusion deformable model for 4D temporal medical image generation. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 539–548. Springer, 2022.

Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. In *International Conference on Learning Representations (ICLR)*, 2020. URL `https://openreview.net/forum?id=rkgNKkHtvB`.

Ming Li, Taojiannan Yang, Huafeng Kuang, Jie Wu, Zhaoning Wang, Xuefeng Xiao, and Chen Chen. ControlNet++: Improving conditional controls with efficient consistency feedback. In *European Conference on Computer Vision (ECCV)*, pp. 129–147, 2024.

Yutong Liu, Haijiang Zhu, Mengting Liu, Huaiyuan Yu, Zihan Chen, and Jie Gao. Rolling-UNet: Revitalizing MLP's ability to efficiently extract long-distance dependencies for medical image segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 38, pp. 3819–3827, 2024.

Alejandro Lozano, Min Woo Sun, James Burgess, Liangyu Chen, Jeffrey J Nirschl, Jeffrey Gu, Ivan Lopez, Josiah Aklilu, Austin Wolfgang Katzer, Collin Chiu, Anita Rau, Xiaohan Wang, Yuhui Zhang, Alfred Seunghoon Song, Robert Tibshirani, and Serena Yeung-Levy. BIOMEDICA: An open biomedical image-caption archive, dataset, and vision-language models derived from scientific literature. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19724–19734, 2025.

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

Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2I-Adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 38, pp. 4296–4304, 2024.

Anugunj Naman, Gaibo Zhang, Ayushman Singh, and Yaguang Zhang. Disk: Dynamic inference skipping for world models, 2026. URL `https://arxiv.org/abs/2602.00440`.

Quang Nguyen, Truong Vu, Anh Tran, and Khoi Nguyen. Dataset diffusion: Diffusion-based synthetic dataset generation for pixel-level semantic segmentation. In *Proceedings of the 37th International Conference on Neural Information Processing Systems (NeurIPS)*, pp. 76872–76892, 2023.

Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Hindtoft, Daniel Cain, Kevin Zhou, Guo-Jun Luo, and Daniel Rueckert. Attention U-Net: Learning where to look for the pancreas. *Medical Image Analysis*, 61:101664, 2020.

Matteo Pagliardini, Pierre Ablin, and David Grangier. The adEMAMix optimizer: Better, faster, older. In *International Conference on Learning Representations (ICLR)*, 2025. URL `https://openreview.net/forum?id=jj7b3p5kLY`.

Walter H. L. Pinaya, Mark S. Graham, Robert Gray, Pedro F. Da Costa, Petru-Daniel Tudosiu, Paul Wright, Yee H. Mah, Andrew D. MacKinnon, James T. Teo, Rolf Jager, David Werring, Geraint Rees, Parashkev Nachev, Sebastien Ourselin, and M. Jorge Cardoso. Fast unsupervised brain anomaly detection and segmentation with diffusion models. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 705–714. Springer, 2022.

Kunpeng Qiu, Zhiqiang Gao, Zhiying Zhou, Mingjie Sun, and Yongxin Guo. Noise-consistent siamese-diffusion for medical image synthesis and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15672–15681, 2025a.

Kunpeng Qiu, Zhiying Zhou, and Yongxin Guo. Adaptively distilled ControlNet: Accelerated training and superior sampling for medical image synthesis. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, volume 15969 of *Lecture Notes in Computer Science*, pp. 55–65, Cham, 2025b. Springer.

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 (CVPR)*, pp. 11536–11546, 2023.

Md Mostafijur Rahman and Radu Marculescu. MK-UNet: Multi-kernel lightweight CNN for medical image segmentation. In *Proceedings of IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*. IEEE/CVF, 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 (CVPR)*, pp. 10684–10695, 2022.

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

Jiacheng Ruan, Mingye Xie, Jingsheng Gao, Ting Liu, and Yuzhuo Fu. Ege-unet: An efficient group enhanced unet for skin lesion segmentation. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 481–490. Springer, 2023.

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. MobileNetV2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4510–4520. IEEE, 2018.

Conner Shorten and Taghi Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):1–48, 2019.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*, 2021a. URL `https://openreview.net/forum?id=St1giarCHLP`.

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 (ICLR)*, 2021b.

Fenghe Tang, Jianrui Ding, Quan Quan, Lingtao Wang, Chunping Ning, and S. Kevin Zhou. CMUNeXt: An efficient medical image segmentation network based on large kernel and skip fusion. In *IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 1–5. IEEE, 2024.

Jeya Maria Jose Valanarasu and Vishal M. Patel. UNeXt: MLP-based rapid medical image segmentation network. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 23–33. Springer, 2022.

Jun Wei, Yiwen Hu, Ruimao Zhang, Zhen Li, S. Kevin Zhou, and Shuguang Cui. Shallow attention network for polyp segmentation. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 699–709. Springer, 2021.

Julia Wolleb, Florentin Bieder, Robin Sandkühler, and Philippe C. Cattin. Diffusion models for medical anomaly detection. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, volume 13438, pp. 35–45. Springer, 2022.

Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: Convolutional block attention module. In *European Conference on Computer Vision (ECCV)*, pp. 3–19. Springer, 2018.

Weijia Wu, Yuzhong Zhao, Mike Zheng Shou, Hong Zhou, and Chunhua Shen. DiffuMask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1206–1216, 2023.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3836–3847, 2023.

Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. UNet++: A nested U-Net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pp. 3–11. Springer, 2018.