BOX2LESION: TRAINING-FREE SPATIAL CONTROL BY INVERTING FROZEN DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion models have shown strong capability for controllable generation of medical images with lesions. However, in many lesion-synthesis settings, the generator produces an image from spatial guidance without explicitly providing a precise mask of the realized lesion. We study whether a *frozen* conditional diffusion model can provide a denoising-based objective for controllable lesion generation, without training an additional segmentation or mask-generation network. To address this question, we introduce **Box2Lesion**, a training-free framework that reuses a frozen lesion synthesis model to recover a refined binary lesion mask from a healthy source image and a coarse bounding box. The box is first used as a coarse mask condition to generate a provisional lesion image. We then optimize only a compact 8×8 residual representation using the conditional denoising behavior of the same frozen model. A source-aware objective encourages the candidate mask to match the box-conditioned denoising response on the provisional lesion image while remaining close to the unconditioned response on the corresponding healthy source. Stochastic mask optimization together with a topology-constrained projection enables refinement of the lesion mask while maintaining spatial coherence without imposing a predefined lesion shape. The recovered mask is then fed back into the same frozen generator to produce paired synthetic lesion images and masks without additional training.

We evaluate Box2Lesion on LIDC-IDRI lung CT and PI-CAI prostate MRI using nnU-Net and SwinUNETR. Across both datasets and backbones, Box2Lesion augmentation improves all reported metrics over the real-data baseline and achieves the best performance among the compared augmentation strategies across all dataset–backbone–metric combinations. These results show that spatial information already encoded in a pretrained lesion generator can be recovered at test time and reused for controllable synthesis and downstream supervision. Code will be made available upon acceptance.

1 INTRODUCTION

Diffusion models have become powerful generative priors for medical image synthesis and editing (Guo et al., 2025; Zhang et al., 2025; Lei et al., 2025; Shin et al., 2025; Yeganeh et al., 2025), enabling localized changes through spatial conditions such as masks or bounding boxes. This capability is particularly appealing in medical imaging (Nazir et al., 2025), where synthetic abnormalities may help reduce the scarcity of rare pathological examples. We ask whether a lesion generator, once trained, can also be reused for spatial reasoning without additional training or supervision.

We study this question in the context of controllable lesion generation. Given a pretrained lesion generator, we would like to specify only a coarse candidate region and recover a more precise lesion mask without training an additional mask generator or segmentation model. This is particularly useful when bounding boxes are available but pixel-level lesion masks are expensive or unavailable.

The challenge is that successful synthesis does not reveal which pixels are responsible for the generated pathology. We therefore ask: *can the conditional denoising behavior of a frozen diffusion model be used to refine coarse spatial guidance into a precise mask of the generated lesion?*

Rather than extracting attention maps or training an additional segmentation or mask-generation network, Box2Lesion directly optimizes the spatial conditioning of a frozen lesion generator. Given

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.

a healthy source image I_H and a coarse bounding box B, we first use B as a coarse mask condition to obtain a provisional lesion image I_0. The unknown lesion mask is represented by a compact 8×8 spatial bottleneck, where only an additive residual is optimized at test time.

Candidate masks are evaluated using a source-aware diffusion distortion that encourages consistency with the box-conditioned response on I_0 while remaining close to the unconditioned response on I_H. Stochastic mask exploration and topology-constrained projection further refine the mask while maintaining spatial coherence without imposing a predefined lesion shape. The recovered mask is then reused by the same frozen generator to synthesize a paired lesion image and mask from coarse bounding-box guidance.

We evaluate Box2Lesion on LIDC-IDRI lung CT and PI-CAI prostate MRI using two downstream segmentation architectures, and further examine its synthetic-only zero-shot setting and key design components. Our main contributions are:

- We introduce **Box2Lesion**, a training-free approach that recovers a refined lesion mask from a healthy image and coarse bounding box by reusing the conditioning behavior of a frozen lesion generator, without training an additional mask or segmentation model.
- We formulate mask recovery as a compact test-time optimization with only 64 learnable residual variables, combining source-aware denoising responses, stochastic Binary Concrete exploration, and topology-constrained binary projection.
- The recovered masks can be fed directly back into the same generator to produce paired synthetic images and masks. Across two modalities and two downstream architectures, these pairs consistently improve segmentation over real-only training, while the zero-shot results further demonstrate that non-trivial lesion-related spatial information can be extracted from the pretrained generator itself.

2 Related Work

2.1 Diffusion Models for Medical Lesion Synthesis

Diffusion models have become increasingly popular for medical image synthesis due to their ability to model complex image distributions and support conditional generation. For pathology synthesis, a particularly useful setting is to generate lesions on top of existing healthy anatomy, such that the synthesized images can be used to augment rare pathological cases while preserving realistic background structures. Recent methods have therefore incorporated spatial conditions, lesion masks, or lesion-specific objectives into diffusion models to improve the controllability of pathological image synthesis (Dorjsembe et al., 2024a;b; Sharma et al., 2024).

LeFusion (Zhang et al., 2025), for example, introduces a lesion-focused diffusion framework that explicitly concentrates the diffusion learning objective on pathological regions while preserving real background content during generation. LesionDiffusion (Lei et al., 2025) similarly targets controllable lesion synthesis using a dedicated lesion-mask generation component followed by mask-conditioned lesion editing. These approaches demonstrate that explicitly modeling lesion geometry can substantially improve the controllability and usefulness of synthetic medical data.

Our goal is complementary but fundamentally different. We consider the setting in which a generative model has *already been trained* for medical image synthesis and ask whether its learned knowledge can be reused for spatial reasoning without training another model. Rather than learning a new mask generator or segmentation network, we freeze the generative model and recover a lesion support directly by optimizing its spatial conditioning variable.

2.2 Spatial Knowledge in Pretrained Diffusion Models

A growing body of work suggests that pretrained diffusion models encode rich spatial and semantic information that can be reused beyond image generation. DiffuMask (Wu et al., 2023) exploits text–image cross-attention in a pretrained text-to-image diffusion model to derive pixel-level annotations for synthetic images. DiffSeg (Tian et al., 2024) uses attention representations extracted from Stable Diffusion to obtain zero-shot segmentation masks, while EmerDiff (Namekata et al.,

2024) demonstrates that fine-grained pixel-level semantic structure can emerge from intermediate diffusion representations without additional segmentation training.

These methods primarily obtain segmentation by extracting spatial information from internal diffusion features or attention maps. In contrast, we do not assume that any particular intermediate representation directly corresponds to a lesion mask. Instead, we use the conditional denoising error of the frozen model as the optimization objective and search for the spatial condition that best explains the generated change. The spatial support is therefore obtained through inverse optimization over the conditioning interface, rather than through feature extraction or attention aggregation. This distinction also makes the formulation applicable to conditional diffusion models whose spatial knowledge is not explicitly represented by text-image attention.

2.3 Training-Free Spatial Control and Test-Time Diffusion Optimization

Another related direction explores training-free control of pretrained diffusion models. RePaint (Lugmayr et al., 2022) performs image inpainting with a pretrained diffusion model by enforcing consistency with known image regions during the reverse process. BoxDiff (Xie et al., 2023) introduces spatial constraints into the denoising process to control object generation from user-provided boxes or scribbles without additional training.

Our problem can be viewed as the inverse of this conventional control setting. Rather than asking how a given spatial condition can force a pretrained model to generate an object at a desired location, we ask which spatial condition is supported by a change that the model has already generated. Starting from a coarse bounding box, we optimize a finer latent mask so that its conditional diffusion response remains consistent with the synthesized lesion but not with the corresponding healthy source. The resulting framework therefore turns training-free diffusion control into an inverse spatial inference problem.

3 Methodology

3.1 Problem Formulation

We consider a frozen mask-conditioned diffusion model that has already been trained for lesion synthesis, where the spatial mask condition is injected through ControlNet (Zhang et al., 2023) to control the lesion location. At inference time, however, we assume that a precise lesion mask is unavailable and only a candidate bounding box $B \in \{0,1\}^{h\times w}$ is given together with a healthy source image I_H. We first use B as a provisional spatial support in the generation pipeline to obtain a lesion-containing image I_0. The frozen generator itself remains mask-conditioned; B is used as a coarse binary mask condition through the mask-conditioned ControlNet.

Our goal is to recover a binary lesion support, *i.e.*, a segmentation mask $M^\star \in 0, 1^{h\times w}$ corresponding to the generated lesion and contained within B, using only the capabilities of the frozen diffusion model. The recovered support can subsequently be reused as a more precise mask condition for controllable lesion generation.

At a high level, we formulate lesion-mask recovery as

$$M^\star = \underset{M\in\mathcal{M}(B)}{\arg\min}\ \mathcal{D}(M) + \beta\mathcal{R}(M), \tag{1}$$

where $\mathcal{D}(M)$ is the source-aware diffusion distortion, $\mathcal{R}(M)$ regularizes the refined spatial estimate, and $\mathcal{M}(B)$ denotes the set of admissible masks within B induced by the topology-constrained projection Π. In practice, $\mathcal{R}$ is applied to the compact latent residual (r) introduced below; the sensitivity to β is analyzed in Appendix A.4.

Since M is discrete, we optimize a low-resolution continuous representation and evaluate binary feasible masks using a straight-through estimator. The overall framework is illustrated in Fig. 1.

3.2 Compact Spatial Bottleneck

We parameterize the lesion mask using a compact 8×8 latent logit grid inside the candidate box. The grid is initialized from the discrepancy between the provisional lesion image I_0 and the healthy

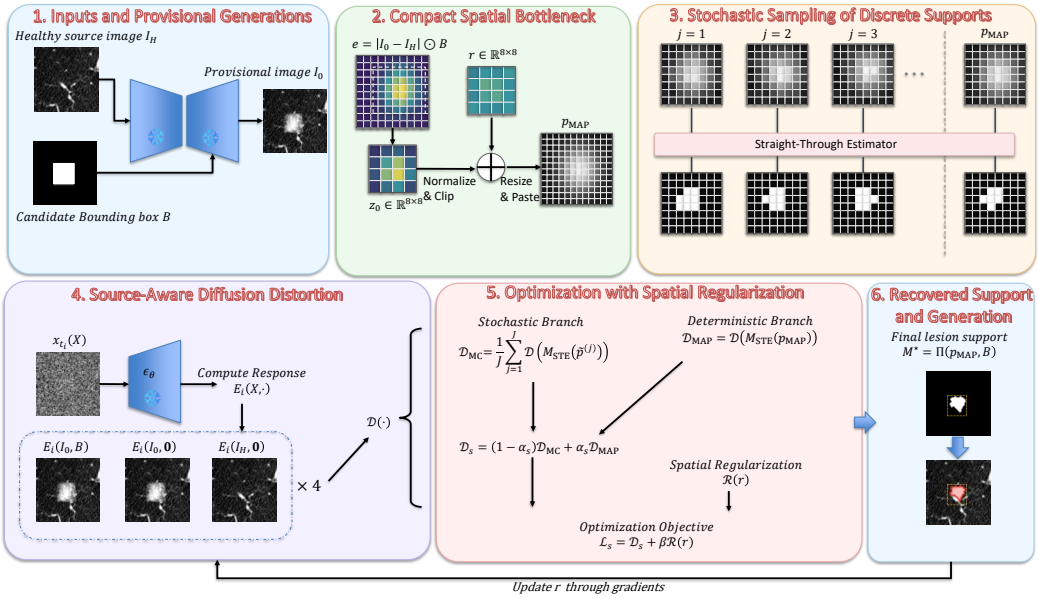

Figure 1: **Box2Lesion:** Overview of the proposed training-free lesion mask recovery framework. The figure illustrates optimization at stage s, with Binary Concrete temperature T_s. $\mathcal{D}_{\mathrm{MC}}$ denotes the Monte Carlo distortion averaged over J stochastic mask samples, while $\mathcal{D}_{\mathrm{MAP}}$ evaluates the deterministic branch. The residual r is the only learnable parameter and represents the correction applied to the fixed discrepancy-derived initialization z_0.

source I_H:

$$e = \mathrm{Resize}_{8\times8}[\mathrm{Crop}_B(|I_0 - I_H| \odot B)], \tag{2}$$

$$z_0 = \frac{e - \mu_e}{\max(\sigma_e, \varepsilon)}, \tag{3}$$

where μ_e and σ_e are the mean and standard deviation of the resized discrepancy map. In practice, z_0 is bounded for numerical stability. The implementation details are provided in Appendix A.1.

The discrepancy-derived grid $z_0 \in \mathbb{R}^{8\times8}$ provides a fixed initialization of the spatial logits. We optimize only an additive residual $r \in \mathbb{R}^{8\times8}$, initialized to zero, so that $z_0 + r$ forms the refined estimate. The residual therefore represents the correction to the source-derived initialization:

$$p_{\mathrm{MAP}} = \mathrm{Sigmoid}\left(\mathrm{Paste}_B\left[\mathrm{Resize}_B(z_0 + r)\right]\right) \odot B. \tag{4}$$

Here, Resize_B resizes the latent field to the spatial extent of B, and Paste_B places it back into the full image resolution. Thus, only 64 latent variables are optimized instead of a full-resolution mask.

3.3 Stochastic Optimization of Discrete Supports

Since the desired lesion mask is binary, we use the Binary Concrete relaxation (Maddison et al., 2016) to enable stochastic exploration of discrete masks. For the j-th sample, we draw

$$u_{a,b}^{(j)} \sim \mathcal{U}(0,1), \qquad g_{a,b}^{(j)} = \log u_{a,b}^{(j)} - \log(1 - u_{a,b}^{(j)}), \tag{5}$$

where $\mathcal{U}(0,1)$ denotes the uniform distribution and (a,b) indexes the 8×8 latent grid. Then we construct

$$\tilde{p}^{(j)} = \mathrm{Sigmoid}\left(\mathrm{Paste}_B\left[\mathrm{Resize}_B\left(\frac{z_0 + r + g^{(j)}}{T_s}\right)\right]\right) \odot B. \tag{6}$$

Following the idea of standard Concrete relaxation (Maddison et al., 2016), the optimization is divided into five stages, with 16 steps per stage. At each step, we draw $J = 16$ stochastic Binary Concrete samples. The temperature T_s is fixed within each stage and annealed between stages. A

higher temperature produces smoother relaxed samples, while decreasing T_s progressively sharpens them toward binary masks. The stochastic fields $\widetilde{p}^{(j)}$ are used to explore alternative discrete supports, whereas p_{MAP} in Eq. 4 represents the current deterministic estimate.

Let p denote either p_{MAP} or a stochastic field $\widetilde{p}^{(j)}$, we evaluate the corresponding binary support using a straight-through estimator (Bengio et al., 2013),

$$\begin{aligned} M &= \Pi(p, B), \\ M_{\mathrm{STE}}(p) &= p + (M - p)\,.\,\mathrm{detach}(), \ \text{for } p \in \{p_{\mathrm{MAP}}, \widetilde{p}^{(j)}\}, \end{aligned} \tag{7}$$

where the projection operator Π is defined in Appendix A.1. Thus, the forward pass uses a discrete feasible mask M while gradients are propagated through the continuous probability field p.

The stochastic samples are used only during optimization.

3.4 Source-Aware Diffusion Distortion

We evaluate a candidate mask through the conditional denoising error of the frozen diffusion model. For probe i, we fix a timestep t_i and Gaussian noise $\epsilon_i \sim \mathcal{N}(0, I)$, and construct

$$x_{t_i}(X) = \sqrt{\bar{\alpha}_{t_i}} X + \sqrt{1 - \bar{\alpha}_{t_i}} \epsilon_i$$

using the standard forward diffusion process. The conditional denoising error is

$$E_i(X, M) = \left\| \epsilon_\theta\big(x_{t_i}(X), t_i, M\big) - \epsilon_i \right\|_{1,w}, \tag{8}$$

where ϵ_θ is the frozen conditional denoiser and $\|\cdot\|_{1,w}$ denotes a spatially weighted L_1 norm with weight 100 inside B and 1 outside B. Thus, $E_i(X, M)$ measures the noise-prediction error of image X under mask condition M. The same (t_i, ϵ_i) pair is used for all conditions within each probe.

Using the box-conditioned provisional lesion image I_0 and its corresponding healthy source I_H, we define

$$\mathcal{D}(M) = \frac{1}{P} \sum_{i=1}^{P} \frac{|E_i(I_0, M) - E_i(I_0, B)| + |E_i(I_H, M) - E_i(I_H, \mathbf{0})|}{|E_i(I_0, \mathbf{0}) - E_i(I_0, B)| + \varepsilon}, \tag{9}$$

where $\mathbf{0}$ denotes an all-zero mask and P is the number of diffusion probes.

The first term in the numerator encourages M to preserve the box-conditioned denoising behavior of the provisional lesion image, while the second encourages the same mask to remain close to the unconditioned response on the healthy source. The denominator normalizes each probe by the change induced by the original box condition.

3.5 Spatially Regularized Support Optimization

The residual r refines the source-derived initialization z_0. To prevent the compact latent representation from encoding unnecessary spatial detail, we regularize both the magnitude and spatial variation of the residual:

$$\mathcal{R}(r) = \frac{1}{2} \left[\mathrm{mean}(r^2) + \frac{1}{2} \mathrm{mean}(\Delta_x r)^2 + \frac{1}{2} \mathrm{mean}(\Delta_y r)^2 \right]. \tag{10}$$

The magnitude term limits excessive deviation from the initialization, while the finite-difference terms encourage spatially smooth corrections. Further details on the regularization weights are provided in Appendix A.1.

At each optimization step, we evaluate two branches. The *stochastic branch* uses J Binary Concrete samples $\widetilde{p}^{(j)}$ to explore alternative discrete masks, whereas the *deterministic branch* evaluates the current MAP estimate p_{MAP}. Their distortions are

$$\mathcal{D}_{\mathrm{MC}} = \frac{1}{J} \sum_{j=1}^{J} \mathcal{D}\left(M_{\mathrm{STE}}(\widetilde{p}^{(j)}) \right), \tag{11}$$

and

$$\mathcal{D}_{\mathrm{MAP}} = \mathcal{D}\left(M_{\mathrm{STE}}(p_{\mathrm{MAP}})\right), \tag{12}$$

where $\mathcal{D}_{\mathrm{MC}}$ is the Monte Carlo estimate obtained by averaging the stochastic branch.

Here, $s \in \{1, \ldots, 5\}$ indexes the optimization stage. Within each stage, the Binary Concrete temperature T_s and mixing weight α_s are fixed. Across stages, T_s is decreased while α_s is increased, progressively shifting the optimization from stochastic exploration toward deterministic refinement:

$$\mathcal{D}_s = (1 - \alpha_s)\mathcal{D}_{\mathrm{MC}} + \alpha_s \mathcal{D}_{\mathrm{MAP}}. \tag{13}$$

The resulting stage-wise objective is

$$\mathcal{L}_s = \mathcal{D}_s + \beta \mathcal{R}(r). \tag{14}$$

The exact schedules and optimization hyperparameters are provided in Appendix A.1.

After the final stage, stochastic sampling is discarded and the recovered binary lesion mask is

$$M^\star = \Pi(p_{\mathrm{MAP}}, B). \tag{15}$$

The mask $M^\star$ is then reused as the spatial condition of the same frozen diffusion model, together with the healthy source image I_H, to synthesize the final lesion-containing image using RePaint-based sampling (Lugmayr et al., 2022).

4 Experiments

4.1 Experimental Protocol

Datasets. We evaluate our method on two medical lesion datasets with different imaging modalities. The detailed dataset splits are provided in the Reproducibility Statement.

LIDC-IDRI: Lung Nodule CT. We use LIDC-IDRI (Armato III et al., 2011), which provides thoracic CT scans with expert pulmonary nodule annotations. For augmentation, we use 297 normal candidate regions with coarse bounding-box guidance to generate 297 synthetic lesion image–mask pairs.

PI-CAI: Prostate Lesion MRI. We use PI-CAI (Saha et al., 2024), a prostate MRI benchmark with images standardized to 0.5×0.5 mm in-plane resolution. Using the same augmentation protocol as LIDC, we generate 297 synthetic lesion image–mask pairs from normal candidate regions and add them to the real training set.

Compared Augmentation Strategies. We compare four training settings. The first uses only real annotated data and serves as the **Baseline**. The remaining settings augment the same real training set with synthetic lesion image–mask pairs.

TumorCP (Yang et al., 2021) is a copy-paste-based augmentation strategy that transfers lesion regions from annotated pathological images to new backgrounds. It provides a representative non-generative lesion augmentation baseline.

RePaint w/ Real Lesion Masks (Lugmayr et al., 2022) uses diffusion-based inpainting to synthesize lesion appearance within real lesion masks. This comparison is particularly informative because it is supplied with ground-truth lesion geometry during synthesis, whereas our method does not require real lesion masks to determine the support of newly generated lesions.

Ours (Box2Lesion) starts from coarse bounding-box guidance and uses the frozen conditional diffusion model itself to recover a refined lesion support through test-time optimization. No additional segmentation network, mask generator, or trainable adapter is introduced during support recovery. The resulting support (synthetic mask) is subsequently used to generate the final lesion image, producing a paired synthetic image and mask for downstream training.

Zero-Shot Setting. To further assess the zero-shot capability of Box2Lesion, we additionally consider a setting in which the downstream segmentation model is trained using only the synthetic

Table 1: **Downstream segmentation performance on LIDC-IDRI.** HD95 is reported in pixels. ↑ indicates higher is better and ↓ indicates lower is better.

	Method	Dice ↑	IoU ↑	Recall ↑	Precision ↑	HD95 (px) ↓
2D nnU-Net	Baseline	0.8143	0.7114	0.8455	0.8155	2.9964
	TumorCP	0.8097	0.7085	0.8520	0.8145	2.7462
	RePaint w/ Real Lesion Masks	0.8056	0.7040	0.8413	0.8225	2.6127
	Ours (Zero-shot)	0.6833	0.5550	**0.9556**	0.5798	3.6210
	Ours	**0.8211**	**0.7233**	0.8653	**0.8260**	**2.4355**
2D SwinUNETR	Baseline	0.7820	0.6797	0.8200	0.8132	3.3817
	TumorCP	0.7886	0.6827	0.8206	0.7765	3.4225
	RePaint w/ Real Lesion Masks	0.7878	0.6833	0.8501	0.7874	3.1114
	Ours (Zero-shot)	0.6548	0.5226	**0.9101**	0.5634	4.1598
	Ours	**0.7920**	**0.6915**	0.8630	**0.8237**	**2.9776**

image–mask pairs recovered from the frozen diffusion model, without real annotated lesion samples. This experiment is not intended as a competing augmentation strategy; instead, it evaluates how much task-relevant spatial information can be extracted from a pretrained lesion generator alone.

For all augmentation methods, we keep the amount of real training data fixed and add the same number of synthetic samples. This isolates the effect of synthetic sample quality rather than differences in training-set size.

Downstream Segmentation Models. Following the common evaluation protocol for lesion synthesis methods, we assess synthetic data by measuring its utility for downstream segmentation. We train two segmentation architectures with substantially different inductive biases: 2D nnU-Net (Isensee et al., 2021) and 2D SwinUNETR (Hatamizadeh et al., 2021). For each combination of dataset, augmentation strategy, and segmentation architecture, the model checkpoint is selected using the validation set and evaluated once on the fixed test set. The same downstream training protocol is used across all augmentation strategies. The further details are provided in the Reproducibility Statement.

Evaluation Metrics. We report Dice coefficient, intersection-over-union (IoU), Recall, Precision, and the 95th-percentile Hausdorff distance (HD95). Higher Dice, IoU, Recall, and Precision indicate better segmentation performance, whereas lower HD95 indicates more accurate boundary localization. For LIDC, physical spacing is unavailable in the image representation used by our downstream benchmark, and HD95 is therefore reported in pixels. For PI-CAI, images are standardized to 0.5×0.5 mm in-plane spacing, and HD95 is reported in millimeters. The absolute HD95 values should therefore not be compared directly across the two datasets. All metrics are first computed independently for each test slice and then macro-averaged over the test set.

4.2 Effectiveness of Synthetic Data for Segmentation

The principal purpose of our generation framework is not merely to produce visually plausible lesions, but to obtain synthetic image–mask pairs that provide useful supervision for downstream learning. We therefore compare segmentation models trained on real data alone with models trained using the same real data augmented by different synthesis strategies.

Results on LIDC. Table 1 summarizes the lung nodule segmentation results. Across both nnU-Net and SwinUNETR, Box2Lesion improves all five metrics over the real-data baseline and achieves the best Dice, IoU, Precision, and HD95 among the compared settings. The zero-shot setting attains higher Recall, but with substantially lower overlap and Precision and worse HD95, indicating less localized predictions. TumorCP and RePaint improve selected metrics but do not provide the same consistent gains across both backbones. Overall, the results show that our synthetic image–mask pairs provide useful supervision across both convolutional and Transformer-based segmentors.

Results on PI-CAI. Table 2 reports the prostate lesion segmentation results. Across both nnU-Net and SwinUNETR, Box2Lesion achieves the best performance on all five reported metrics. Tu-

Table 2: **Downstream segmentation performance on PI-CAI.** HD95 is reported in millimeters. ↑ indicates higher is better and ↓ indicates lower is better.

	Method	Dice ↑	IoU ↑	Recall ↑	Precision ↑	HD95 (mm) ↓
2D nnU-Net	Baseline	0.7055	0.5714	0.7398	0.7432	3.4220
	TumorCP	0.7210	0.5862	0.7823	0.7545	3.4420
	RePaint w/ Real Lesion Masks	0.7235	0.5890	0.7720	0.7605	3.5423
	Ours (Zero-shot)	0.5332	0.3870	0.5109	0.7733	5.7469
	Ours	**0.7280**	**0.5949**	**0.8075**	**0.7801**	**3.3608**
2D SwinUNETR	Baseline	0.7083	0.5697	0.8083	0.7111	3.7590
	TumorCP	0.7057	0.5665	0.7772	0.7369	3.7063
	RePaint w/ Real Lesion Masks	0.6997	0.5604	0.7774	0.7029	3.8257
	Ours (Zero-shot)	0.5720	0.4230	0.5911	0.7343	5.3717
	Ours	**0.7115**	**0.5723**	**0.8103**	**0.7400**	**3.6745**

morCP and RePaint improve selected metrics with nnU-Net, but their gains are less consistent and do not transfer as reliably to SwinUNETR. In contrast, Box2Lesion maintains improvements across both backbones, indicating that the benefit of the generated image–mask pairs extends beyond a single segmentation architecture. Together with the LIDC results, this suggests that the proposed recovery-and-generation framework generalizes across substantially different imaging modalities and anatomical settings.

Cross-Dataset and Cross-Backbone Consistency. Across two datasets, two downstream architectures, and five evaluation metrics, Box2Lesion achieves the best performance among the compared augmentation strategies across all reported dataset–backbone–metric combinations. The consistent gains across datasets and architectures indicate that the recovered masks provide useful supervision beyond a particular downstream segmentor. The downstream results therefore provide complementary evidence that the recovered supports are not only plausible spatial explanations of the generated lesions, but are sufficiently aligned with lesion appearance to serve as effective segmentation supervision.

4.3 Visual Quality Assessment

Figure 2 qualitatively compares lesion synthesis from the same healthy source image and target region. TumorCP and real-mask RePaint rely on fixed donor lesion masks whose geometry is independent of the current source image. In contrast, our method first produces a provisional lesion synthesis and then optimizes a source-adaptive binary mask $M^\star$ for the current image. The resulting lesions appear more naturally integrated with the surrounding anatomy and show better visual compatibility with the local image background. This behavior is observed in both LIDC CT and PI-CAI MRI examples, suggesting that the optimized mask can adapt its shape to different imaging modalities and anatomical contexts.

Figure 3 further illustrates the interaction between mask optimization and lesion generation. Starting from the coarse box B, the mask progressively evolves into the final source-adaptive binary mask $M^\star$. The same optimized mask is then used directly as the spatial condition during the subsequent RePaint denoising process. Therefore, $M^\star$ is not obtained by segmenting or post-processing the final synthetic image; it is optimized before final synthesis and serves as the spatial condition for lesion generation. This coupling between mask recovery and generation yields an image–mask pair through the same generative process, helping reduce spatial mismatch between the synthesized lesion and its associated mask.

Additional Experiments. Additional experiments on controlled lesion support recovery and component ablations are provided in Appendix A.2 and Appendix A.3, respectively.

5 Conclusion

We introduced Box2Lesion, a training-free framework that reuses a frozen mask-conditioned diffusion model to recover refined lesion masks from coarse bounding-box guidance. Rather than train-

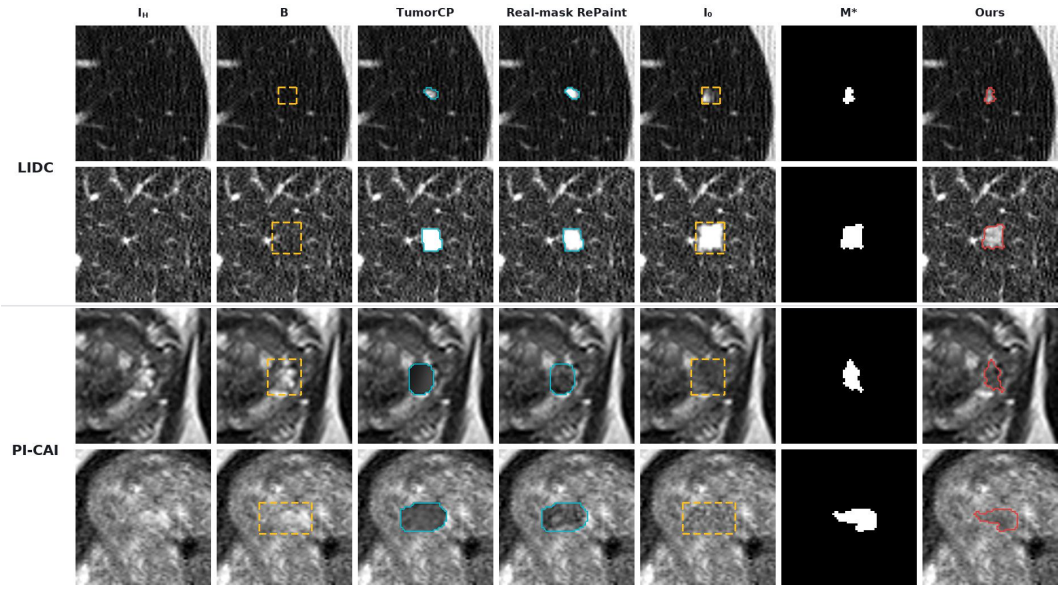

Figure 2: **Qualitative comparison of lesion synthesis.** Starting from the same healthy source image I_H and target region, we compare TumorCP, RePaint using real lesion masks, and our method on LIDC CT and PI-CAI MRI. Our method derives a source-adaptive non-rectangular mask $M^\star$ from the provisional synthesis I_0, rather than reusing a fixed donor mask. The resulting lesions show better visual compatibility with the surrounding anatomy and local image background.

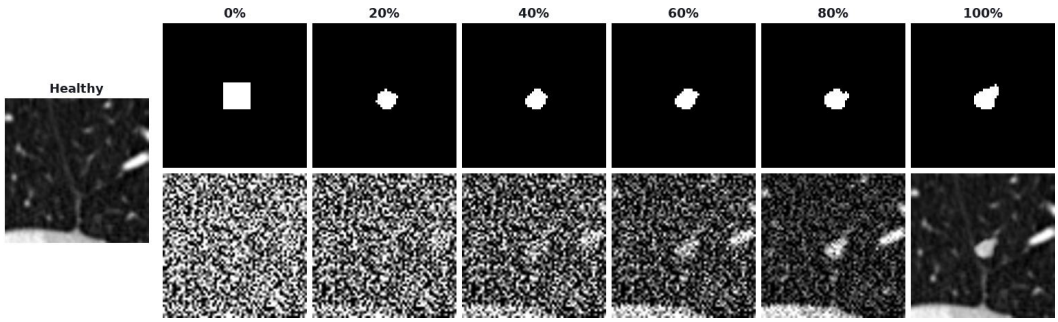

Figure 3: **Mask optimization and subsequent lesion generation.** Top: the coarse box B is progressively refined into the final binary lesion mask $M^\star$ during optimization. Bottom: the same $M^\star$ is used as the spatial condition during the subsequent RePaint denoising process.

ing an additional segmentation or mask-generation network, Box2Lesion optimizes only a compact spatial residual and feeds the recovered mask back into the same generator for controllable lesion synthesis. Experiments on LIDC-IDRI CT and PI-CAI MRI show that the resulting synthetic image–mask pairs provide useful supervision across different downstream segmentation architectures, while the zero-shot results further indicate that pretrained lesion generators contain reusable spatial information beyond image synthesis.

The current study is limited to 2D lesion generation with a predefined candidate box and a single connected lesion support, and the test-time optimization introduces additional per-sample inference cost. Future work will investigate extension to 3D and multi-lesion settings, broader conditional generators and imaging modalities, and more efficient optimization for lesion-mask recovery.

AI USE STATEMENT

Generative AI tools were used to assist with writing and language refinement, literature search and analysis, methodology presentation (the idea is from human) and feedback. All AI-assisted outputs were critically reviewed, verified, and revised by the authors as necessary. We take full responsibility for the final content of this work, including all text, claims, analyses, and other artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The primary datasets used in this study were obtained from publicly available online repositories and were subsequently processed and curated by the authors. We used the data solely for research purposes and in accordance with the applicable licenses, terms of use, and data-access policies of the respective datasets. The study did not involve direct interaction with human participants or the collection of new personal data. No personally identifiable information was intentionally collected, retained, or disclosed during data processing or analysis.

REPRODUCIBILITY STATEMENT

Diffusion models. LIDC-IDRI and PI-CAI use independently trained diffusion models with identical configurations. All diffusion models operate directly in the 64×64 2D pixel space on single-channel images normalized to $[-1, 1]$, without a VAE or latent representation. Training is performed on a single NVIDIA RTX 6000 Ada Generation 48 GB GPU.

The generator contains two stages. Stage 1 is an unconditional DDPM (Ho et al., 2020) with a U-Net using 64 base channels, channel multipliers $(1, 2, 4, 8)$, and eight attention heads. We use 300 diffusion steps with a cosine noise schedule and train with an L_1 noise-prediction loss for 300 epochs using batch size 64. Optimization uses AdamW with learning rate 10^{-4}, zero weight decay, cosine learning-rate decay, BF16 mixed precision, gradient-norm clipping at 1.0, and EMA with decay 0.995. The best checkpoint is selected using the validation EMA ϵ-L_1 loss.

Stage 2 loads and freezes the best Stage-1 EMA U-Net and trains only a mask-conditioned ControlNet. The condition is a single-channel binary mask with condition-embedding width 16. Training runs for 100 epochs with batch size 64 using AdamW with learning rate 10^{-4} and weight decay 0.01. The noise-prediction loss uses spatial weights 100 inside the lesion mask and 1 outside; the remaining precision and learning-rate settings follow Stage 1. The best checkpoint is selected by validation weighted ϵ-L_1. Only horizontal flipping with probability 0.5 is used for diffusion training. The diffusion-model train/validation splits contain 12,248/1,336 slices for LIDC-IDRI and 1,768/165 slices for PI-CAI.

nnU-Net. We use nnU-Net v2.8.1 (Isensee et al., 2021) with the default 2D `nnUNetTrainer`, `nnUNetPlans`, and fold 0. Inputs are single-channel 64×64 images with per-image Z-score normalization. The automatically configured five-level PlainConvUNet (Isensee et al., 2021) uses 32/64/128/256/512 feature channels, InstanceNorm, LeakyReLU, and deep supervision. Models are trained from scratch for 1000 epochs with 250 updates per epoch (250,000 updates in total). We use equally weighted Dice and cross-entropy losses and SGD with Nesterov momentum 0.99, initial learning rate 10^{-2}, weight decay 3×10^{-5}, and polynomial decay

$$10^{-2}(1 - e/1000)^{0.9}.$$

Training uses AMP, gradient-norm clipping at 12, and the default nnU-Net 2D spatial and intensity augmentations. Batch sizes are 35 for LIDC and 97 for PI-CAI. The best checkpoint is selected using the exponential moving average of foreground pseudo-Dice on the real validation set.

SwinUNETR. SwinUNETR is implemented with MONAI 1.5.1 using single-channel 64×64 inputs with per-slice Z-score normalization. We use `feature_size=24`, `patch_size=2`, `depths=(2,2,2,2)`, `num_heads=(3,6,12,24)`, `window_size=7`, and a drop-path rate of 0.1, resulting in approximately 6.3M parameters. All models are trained for 1000 epochs with 29 updates per epoch (29,000 updates in total) and batch size 32. The loss is an equal combination of foreground Dice and cross-entropy. We use AdamW with peak learning rate 3×10^{-4}, weight

decay 10^{-5}, linear warm-up followed by cosine decay, AMP, and gradient-norm clipping at 1.0. Augmentation includes horizontal and vertical flipping, random affine transforms, mild intensity perturbations, and Gaussian noise. The best checkpoint is selected by the highest slice-macro Dice on the real validation set.

Data splits and comparison protocol. The fixed downstream train/validation/test splits are 630/70/175 images for LIDC and 630/165/454 images for PI-CAI. Each augmentation method adds 297 synthetic image–mask pairs to the same 630 real training images, while validation and test sets contain only real images. All train/validation/test partitions are patient-disjoint, and no downstream test subject is used to train the diffusion models. All methods use identical data splits, model configurations, and optimization budgets. Downstream segmentation experiments are run on a single RTX 4090, while a RTX 6000 Ada is used specifically for diffusion-model training.

REFERENCES

Samuel G Armato III, Geoffrey McLennan, Luc Bidaut, Michael F McNitt-Gray, Charles R Meyer, Anthony P Reeves, Binsheng Zhao, Denise R Aberle, Claudia I Henschke, Eric A Hoffman, et al. The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. *Medical physics*, 38(2):915–931, 2011.

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

Zolnamar Dorjsembe, Hsing-Kuo Pao, Sodtavilan Odonchimed, and Furen Xiao. Conditional diffusion models for semantic 3d brain mri synthesis. *IEEE Journal of Biomedical and Health Informatics*, 28(7):4084–4093, 2024a.

Zolnamar Dorjsembe, Hsing-Kuo Pao, and Furen Xiao. Polyp-ddpm: Diffusion-based semantic polyp synthesis for enhanced segmentation. In *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 1–7. IEEE, 2024b.

Pengfei Guo, Can Zhao, Dong Yang, Ziyue Xu, Vishwesh Nath, Yucheng Tang, Benjamin Simon, Mason Belue, Stephanie Harmon, Baris Turkbey, et al. Maisi: Medical ai for synthetic imaging. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 4430–4441. IEEE, 2025.

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

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

Wenhui Lei, Hengrui Tian, Linrui Dai, Hanyu Chen, and Xiaofan Zhang. Lesiondiffusion: Towards text-controlled general lesion synthesis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 327–336. Springer, 2025.

Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 11451–11461. IEEE, 2022.

Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*, 2016.

Koichi Namekata, Amirmojtaba Sabour, Sanja Fidler, and Seung Wook Kim. Emerdiff: Emerging pixel-level semantic knowledge in diffusion models. In *International Conference on Learning Representations*, volume 2024, pp. 11348–11370, 2024.

Maham Nazir, Muhammad Aqeel, and Francesco Setti. Diffusion-based data augmentation for medical image segmentation. In *2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 1341–1350. IEEE, 2025.

Anindo Saha, Joeran S Bosma, Jasper J Twilt, Bram Van Ginneken, Anders Bjartell, Anwar R Padhani, David Bonekamp, Geert Villeirs, Georg Salomon, Gianluca Giannarini, et al. Artificial intelligence and radiologists in prostate cancer detection on mri (pi-cai): an international, paired, non-inferiority, confirmatory study. *The Lancet Oncology*, 25(7):879–887, 2024.

Vanshali Sharma, Abhishek Kumar, Debesh Jha, Manas Kamal Bhuyan, Pradip K Das, and Ulas Bagci. Controlpolypnet: Towards controlled colon polyp synthesis for improved polyp segmentation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2325–2334. IEEE, 2024.

Yejee Shin, Yeeun Lee, Hanbyol Jang, Geonhui Son, Hyeongyu Kim, and Dosik Hwang. Anatomical consistency and adaptive prior-informed transformation for multi-contrast mr image synthesis via diffusion model. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 30918–30927. IEEE, 2025.

Junjiao Tian, Lavisha Aggarwal, Andrea Colaco, Zsolt Kira, and Mar Gonzalez-Franco. Diffuse, attend, and segment: Unsupervised zero-shot segmentation using stable diffusion. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3554–3563. IEEE, 2024.

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 *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1206–1217. IEEE, 2023.

Jinheng Xie, Yuexiang Li, Yawen Huang, Haozhe Liu, Wentian Zhang, Yefeng Zheng, and Mike Zheng Shou. Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7418–7427. IEEE, 2023.

Jiawei Yang, Yao Zhang, Yuan Liang, Yang Zhang, Lei He, and Zhiqiang He. Tumorcp: A simple but effective object-level data augmentation for tumor segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 579–588. Springer, 2021.

Yousef Yeganeh, Azade Farshad, Ioannis Charisiadis, Marta Hasny, Martin Hartenberger, Björn Ommer, Nassir Navab, and Ehsan Adeli. Latent drifting in diffusion models for counterfactual medical image synthesis. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7685–7695. IEEE, 2025.

Hantao Zhang, Yuhe Liu, Jiancheng Yang, Shouhong Wan, Xinyuan Wang, Wei Peng, and Pascal Fua. Lefusion: Controllable pathology synthesis via lesion-focused diffusion models. In *International Conference on Learning Representations*, volume 2025, pp. 13232–13253, 2025.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3813–3824. IEEE, 2023.

A Appendix

A.1 Implementation Details

Initialization and Resizing Details. Crop_B extracts the rectangular region specified by the candidate bounding box B. All spatial resizing operations use bilinear interpolation. Specifically, $\mathrm{Resize}_{8\times 8}$ maps the cropped discrepancy map to the fixed 8×8 latent grid, while Resize_B upsamples an 8×8 logit field to the height and width of B. Paste_B then places the resized field back into the corresponding location of the full-resolution image canvas, with values outside B set to zero.

After normalization, the initialization z_0 is clipped element-wise to $[-3, 3]$:

$$z_0 \leftarrow \mathrm{clip}(z_0, -3, 3). \tag{16}$$

Since z_0 is standardized by its mean and standard deviation, this clipping limits extreme initialization logits to three standard deviations and prevents a small number of large discrepancy values from dominating the initial probability field. The same bound is used for all experiments.

Topology-Constrained Projection. Conceptually, the procedure $\Pi(p, B)$ contains the following sequence of operations applied to p:

$$p \;\rightarrow\; \text{binary mask} \;\rightarrow\; \text{remove fragments/holes} \;\rightarrow\; \text{check thin neck} \;\rightarrow\; \text{optional separation}.$$

The role of the projection is to convert the continuous probability field produced by the latent optimization into a spatially coherent binary lesion mask that can be passed to the frozen diffusion model. At a high level, the projection first removes disconnected fragments and internal holes, and then checks whether the remaining connected mask contains multiple substantial regions joined only by a narrow bridge. Only when such a thin-neck configuration is detected do we further separate the mask and retain the region most strongly supported by the optimized probability field.

Base projection. We first threshold the probability field at 0.5 inside B:

$$S = \mathbf{1}[p \geq 0.5] \odot B. \tag{17}$$

Among the 4-connected foreground components of S, we retain the dominant component using the optimized probability field p. In our implementation, this operation is denoted by $\mathrm{KeepLargest4CC}(S, p, B)$:

$$M_0 = \mathrm{KeepLargest4CC}(S, p, B). \tag{18}$$

We next remove enclosed background cavities. $\mathrm{FillBG8}$ fills background regions that are not connected to the exterior under 8-connectivity, while $\mathrm{FillBG4}$ performs the analogous operation using 4-connectivity. Applying both operations removes small cavity configurations that depend on the chosen digital-connectivity convention.

We then apply $\mathrm{OpenCross3}$, a binary opening consisting of erosion followed by dilation with a fixed 3×3 cross-shaped structuring element. This removes small spur-like or very thin local structures without imposing a global shape on the lesion:

$$M_1 = \mathrm{OpenCross3}\left(\mathrm{FillBG4}\left[\mathrm{FillBG8}(M_0)\right]\right). \tag{19}$$

Because opening may separate a weakly connected structure, we again retain a single 4-connected component and fill any remaining internal holes:

$$M = \Pi_{\mathrm{base}}(p, B) = \mathrm{FillBG4}\left[\mathrm{KeepLargest4CC}(M_1, p, B)\right]. \tag{20}$$

Conditional thin-neck detection. A mask may satisfy 4-connectivity while still containing two substantial regions connected only by a narrow bridge. To detect this case, we apply a 3×3 square opening as a diagnostic probe:

$$Q = \mathrm{OpenSquare3}(M), \qquad \{C_1, \ldots, C_K\} = \mathrm{CC}_4(Q), \tag{21}$$

where CC_4 returns the 4-connected foreground components. The square opening is more restrictive than the cross-shaped opening and therefore removes narrow connections that cannot support the full 3×3 structuring element. Importantly, Q is used only to detect stable cores and is never returned as the projected mask.

If $K \leq 1$, the probe does not reveal multiple stable cores and no additional separation is performed:

$$\Pi(p, B) = M. \tag{22}$$

Conditional basin separation. When $K \geq 2$, the components $\{C_i\}$ indicate regions that remain substantial after the thin connection has been removed by the diagnostic opening. We then compute the Euclidean distance transform inside the original mask:

$$d(x) = \mathrm{EDT}(M)(x), \tag{23}$$

where $d(x)$ is the Euclidean distance from foreground pixel x to the nearest background pixel. Large values of d therefore occur near the interior of thicker regions, while narrow connections have smaller values.

The detected cores are used as markers for a 4-connected marker-controlled watershed on $-d$, restricted to the original mask M:

$$\{R_1, \ldots, R_K\} = \mathrm{Watershed}_4\left(-d; \mathrm{markers} = \{C_i\}, \mathrm{domain} = M\right). \tag{24}$$

Using $-d$ turns the interior distance maxima into watershed minima, allowing the markers to partition the original connected mask into basins associated with the detected cores. Because the watershed operates on M rather than on the opened probe Q, the original irregular lesion boundary is retained.

We keep the basin with the largest optimized probability mass,

$$i^\star = \arg\max_i \sum_{x \in R_i} p(x), \tag{25}$$

rather than selecting a basin solely by geometric area. A final connectivity and cavity cleanup gives

$$\Pi(p, B) = \mathrm{FillBG4}\left[\mathrm{FillBG8}\left[\mathrm{KeepLargest4CC}\left(R_{i^\star}, p, B\right)\right]\right]. \tag{26}$$

Thus, the additional watershed step is activated only when the diagnostic opening reveals multiple stable cores. It removes narrow-bridge degeneracies while avoiding a predefined global lesion geometry.

Straight-Through Evaluation. The same projection Π is used for both the deterministic field p_{MAP} and each stochastic Binary Concrete field $\widetilde{p}^{(j)}$. For either case,

$$M_{\mathrm{STE}}(p) = p + \left(\Pi(p, B) - p\right).\mathrm{detach}(). \tag{27}$$

The forward value is therefore exactly the projected binary mask $\Pi(p, B)$, whereas the backward pass propagates gradients through the continuous probability field p. Consequently, the frozen diffusion model is always evaluated with a binary feasible mask while the latent logits remain optimizable.

Regularization Weighting. The residual regularizer used in Sec. 3.5 is

$$\mathcal{R}(r) = \frac{1}{2}\,\mathrm{mean}(r^2) + \frac{1}{4}\,\mathrm{mean}(\Delta_x r)^2 + \frac{1}{4}\,\mathrm{mean}(\Delta_y r)^2, \tag{28}$$

where $\Delta_x r$ and $\Delta_y r$ denote first-order finite differences along the two axes of the 8×8 latent grid. The magnitude term therefore receives total weight $1/2$, while the remaining half is split equally between horizontal and vertical variation. This prevents the smoothness contribution from doubling simply because the latent grid has two spatial directions.

Optimization Schedule. We use a five-stage optimization schedule that gradually shifts from stochastic exploration to deterministic refinement. The MAP mixing weights and Binary Concrete temperatures are fixed as

$$(\alpha_1, \ldots, \alpha_5) = (0.10,\ 0.25,\ 0.50,\ 0.75,\ 1.00), \tag{29}$$

and

$$(T_1, \ldots, T_5) = (1.00,\ 0.75,\ 0.50,\ 0.35,\ 0.25). \tag{30}$$

Each stage contains 16 optimization steps, giving 80 steps in total. At each step, we use $P = 4$ diffusion probes and $J = 16$ Binary Concrete samples. As T_s decreases, the stochastic samples become progressively sharper, while the increasing α_s shifts the distortion from the stochastic Monte Carlo branch toward the deterministic MAP branch.

The residual r is optimized using Adam with learning rate 0.05, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, and zero weight decay. Gradients are clipped at 5, and the residual values are constrained to $r \in [-6, 6]$ throughout optimization. All optimization settings are kept fixed across experiments.

Table 3: Controlled lesion support recovery against the hidden editing support G. The recovery procedure observes only (I_H, B, I_0); G is used exclusively for post-hoc evaluation. Higher Dice, IoU, Recall, and Precision are better, while lower HD95 is better.

Dataset	Method	Dice ↑	IoU ↑	Recall ↑	Precision ↑	HD95 ↓
LIDC-IDRI	BBox	0.2342	0.1445	1.0000	0.1445	12.5778 px
	z_0-only	0.5122	0.4011	**0.8516**	0.4235	**5.9577 px**
	Ours	**0.5261**	**0.4171**	0.8179	**0.4494**	6.2671 px
PI-CAI	BBox	0.3390	**0.2270**	1.0000	0.2270	5.9446 mm
	z_0-only	0.3369	0.2241	0.4356	0.3163	4.6287 mm
	Ours	**0.3578**	0.2250	0.4403	**0.3173**	**4.5920 mm**

A.2 Controlled Lesion Support Recovery

The downstream segmentation experiments evaluate whether the recovered supports produce useful synthetic image–mask pairs, but they do not directly test whether the proposed inverse optimization can recover the spatial support responsible for a provisional generation. We therefore construct a controlled support-recovery experiment in which the editing support used to produce the provisional lesion image is known but hidden from the recovery procedure.

For each sample, we start from a healthy source image I_H and a candidate bounding box B. A binary editing support $G \subset B$ is obtained from a real lesion mask in the corresponding training split and deterministically transformed and placed inside B. The frozen lesion generator then uses G to produce a provisional lesion-containing image I_0. During recovery, however, the method is given only (I_H, B, I_0); the hidden support G is never exposed to the optimization and is used only after inference for evaluation. This yields the controlled setting

$$(I_H, B, I_0) \xrightarrow{\text{support recovery}} M^\star, \qquad M^\star \text{ evaluated against hidden } G. \tag{31}$$

This experiment should therefore be interpreted as a mechanism-level validation of lesion support recovery rather than as a downstream segmentation benchmark. In particular, G is a transformed editing support derived from training-side lesion annotations and is not a native expert annotation of the healthy target image.

Compared variants. We compare three support estimates. **BBox** directly uses the candidate bounding box, $M = B$. **z_0-only** uses the source-derived discrepancy initialization defined in Eq. 3 and applies the same topology-constrained projection as the full method, but performs no test-time optimization. **Full** uses the complete proposed procedure, including the compact spatial bottleneck, Binary Concrete sampling, source-aware diffusion distortion, spatial regularization, and topology-constrained projection.

We evaluate all 297 samples in each dataset without filtering by recovery quality or optimization outcome. Dice, IoU, Recall, Precision, and HD95 are computed independently for each slice and macro-averaged. HD95 is reported in pixels for LIDC-IDRI and in millimeters for PI-CAI.

Results. Table 3 shows that the recovered supports capture substantially more localized lesion structure than the coarse bounding box, while the magnitude of refinement differs across datasets. The bounding-box baseline attains Recall $= 1$ by construction because the hidden editing support satisfies $G \subset B$; this reflects complete containment rather than precise support localization.

On LIDC-IDRI, the full method achieves the best Dice, IoU, and Precision, showing that test-time optimization further refines the source-derived initialization. On PI-CAI, the full method also achieves the best Dice and Precision, together with the lowest HD95.

Overall, the full method improves Dice over z_0-only on both datasets, providing controlled evidence that the frozen diffusion response supplies an effective optimization signal beyond the source-derived discrepancy prior. These results validate the feasibility of the proposed support-recovery mechanism rather than representing an independent lesion-segmentation benchmark.

Table 4: Downstream ablation of the support-recovery components on LIDC-IDRI using 2D nnU-Net. All variants augment the same real training set with 297 synthetic image–mask pairs. Higher Dice, IoU, Recall, and Precision are better, while lower HD95 is better.

Variant	Dice ↑	IoU ↑	Recall ↑	Precision ↑	HD95 ↓
Ours	**0.8211**	**0.7233**	0.8653	0.8260	2.4355
w/o source	0.8182	0.7165	0.8528	**0.8280**	2.6992
MAP-only	0.8194	0.7188	**0.8823**	0.8035	2.9487
Threshold-only	0.8080	0.7051	0.8745	0.7882	2.9287
Basic topology	0.8186	0.7188	0.8792	0.8077	**2.1137**

A.3 Ablation of Support-Recovery Components

We ablate the main components of the proposed support-recovery procedure to examine how each design choice affects the downstream utility of the generated image–mask pairs.

w/o source removes the healthy-source term from the source-aware diffusion distortion. In the full formulation, a candidate support is required not only to preserve the conditional response associated with the provisional lesion image I_0, but also to remain unnecessary for explaining the corresponding healthy source I_H. This ablation tests the contribution of this source-aware contrast.

MAP-only removes the stochastic Binary Concrete branch and optimizes only the deterministic p_{MAP} branch. The full method uses stochastic supports to explore alternative discrete configurations before progressively shifting toward deterministic refinement. This variant tests the role of stochastic support exploration.

Threshold-only replaces the topology-constrained projection Π with direct thresholding of the probability 0.5. It therefore removes the spatial feasibility constraints imposed on the discrete support, including connectivity and hole filling.

Basic topology retains the base topology projection but removes the conditional thin-neck detection and separation used in the full projection. This variant tests whether the additional handling of narrow-bridge degeneracies is useful beyond basic spatial coherence.

The full method provides the strongest overall downstream performance and is therefore used in the main experiments. The largest degradation is observed when topology-constrained projection is replaced by direct thresholding, indicating that enforcing a coherent discrete support is an important part of the recovery procedure. Removing the healthy-source term or stochastic exploration leads to smaller changes, suggesting that these components provide complementary refinement. Basic topology remains competitive, while the full projection additionally handles thin-neck degeneracies that are not addressed by the simplified variant.

Overall, we keep the full formulation because it combines all three design principles of the proposed method: source-aware spatial evidence, stochastic exploration followed by deterministic refinement, and topology-constrained support projection.

A.4 Sensitivity to the Spatial Regularization Weight

We study the sensitivity to the regularization weight β in Eq. 14,

$$\mathcal{L}_s = \mathcal{D}_s + \beta \mathcal{R}(r),$$

which controls the strength of the spatial regularization applied to the learnable residual r. A larger β places greater emphasis on limiting the magnitude and spatial variation of the latent correction, whereas $\beta = 0$ removes this regularization entirely.

The results show that a moderate regularization weight is preferable to strong regularization. The default setting $\beta = 0.10$ gives the strongest overall performance on LIDC-IDRI and remains competitive on PI-CAI, whereas larger values tend to degrade downstream performance. Very small or zero regularization can also perform competitively, indicating that the diffusion distortion remains the primary optimization signal. We therefore use $\beta = 0.10$ as a balanced default across datasets.

Table 5: Sensitivity to the spatial regularization weight β. All other components of the support-recovery and downstream training pipelines (2D nnU-Net) are kept fixed. Higher Dice, IoU, Recall, and Precision are better, while lower HD95 is better.

β	Dice ↑	IoU ↑	Recall ↑	Precision ↑	HD95 ↓
LIDC-IDRI					
0.00	0.8136	0.7124	**0.8833**	0.7959	2.5180
0.03	0.8036	0.7030	0.8242	**0.8290**	2.5804
0.10	**0.8211**	**0.7233**	0.8653	0.8260	**2.4355**
0.30	0.8124	0.7109	0.8476	0.8258	2.5458
1.00	0.8121	0.7105	0.8690	0.8062	2.5520
PI-CAI					
0.00	0.7003	0.5703	0.7184	0.7674	3.4703
0.03	0.7057	0.5693	0.7459	0.7621	3.4985
0.10	**0.7280**	**0.5949**	**0.8075**	**0.7801**	**3.3608**
0.30	0.6843	0.5485	0.7239	0.7574	3.6763
1.00	0.6970	0.5601	0.7260	0.7651	3.5784