VGD: Value-Guided Diffusion Toward High-Utility Medical Image Segmentation

Hongyu Zhang[1,2], **Haipeng Chen**[1,2], **Chengxin Yang**[1,2] **Yingda Lyu**[3]

[1]College of Computer Science and Technology, Jilin University, China
[2] Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, China
[3]Center for Public Education Research, Jilin University, China
{zhanghongyu22,yangcx24}@mails.jlu.edu.cn, {chenhp, ydlv}@jlu.edu.cn

Abstract

Progress in medical image segmentation is fundamentally constrained by the scarcity of annotated data. While diffusion models offer a promising solution by generating high-fidelity image–mask pairs, their utility for downstream tasks remains underexplored. A key bottleneck lies in the misalignment between generation outputs and task-specific needs—samples are produced independently of their utility for downstream training. To this end, we propose **Value-Guided Diffusion (VGD)**, a lightweight sampling framework that integrates downstream model feedback into the generative inference process. VGD estimates a value score for each sample based on its utility to downstream training, and leverages this signal to iteratively guide the denoising trajectory toward high-reward regions of the data manifold. Crucially, VGD can be seamlessly integrated into existing medical diffusion models without any additional training or architectural modifications. Extensive experiments across multiple diffusion backbones and segmentation benchmarks demonstrate that VGD significantly boosts downstream segmentation performance while maintaining visual fidelity. Our findings highlight a task-aware sampling principle with potential to underpin future synthetic segmentation pipelines.

Supp. materials & code — https://github.com/JackCD99

Introduction

Diffusion Models (DMs) (Ho, Jain, and Abbeel 2020; Song et al. 2021), empowered by strong generative modeling capabilities, have emerged as a powerful tool for medical image analysis (Qiu et al. 2025; Li et al. 2025a). Beyond merely augmenting datasets, DM-based synthetic training has shown promise in addressing long-standing challenges in real-world medical imaging, including image corruption, missing annotations (Chen et al. 2025), privacy constraints (Giuffrè and Shung 2023), and class imbalance in long-tailed distributions (Nie et al. 2018).

Motivated by these advantages, recent efforts have increasingly explored the use of DMs to synthesize image–mask pairs for augmenting data-scarce medical segmentation tasks (Qiu et al. 2025; Li et al. 2025a; Zhang et al. 2024b). Paradoxically, despite the narrowing fidelity gap between synthetic and real images, existing medical DMs often yield only limited improvements in downstream performance (*c.f.* Tab. 2). This disconnect raises a central question: *Why doesn't the strong generative capacity of DMs reliably translate into greater downstream utility?*

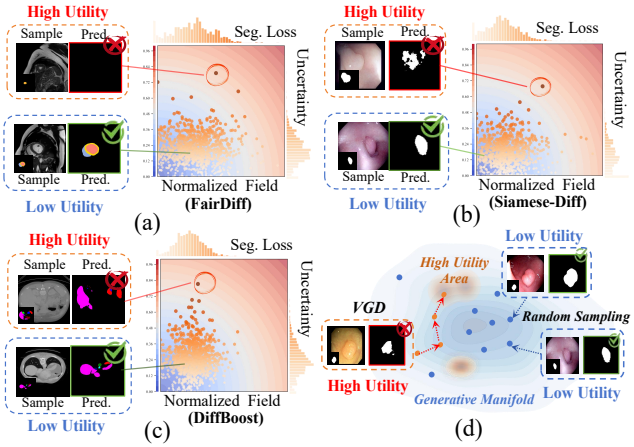

Figure 1: **(a–c)**: Loss fields of synthetic samples produced by SOTA medical DMs, evaluated via the downstream segmenter (nnU-Net). Most samples lie in low-temperature areas, providing limited training signal. **(d)**: Illustration of **VGD**, which leverages value-based guidance to actively navigate the generative manifold toward high-utility regions.

To investigate this discrepancy, we construct a fixed-budget synthetic dataset using three SOTA medical DMs—FairDiff (Li et al. 2024), SiameseDiff (Qiu et al. 2025), and DiffBoost (Zhang et al. 2024b)—and evaluate each sample based on its resulting segmentation loss and predictive uncertainty under a downstream model. As shown in Fig. 1, most samples exhibit low loss and uncertainty, indicating they are overly simple or semantically redundant from the model's perspective. We attribute this to representational overlap: Since the generator is trained or fine-tuned on the same dataset as the downstream model, its generative manifold tends to concentrate on patterns already familiar to and well-modeled by that model. In contrast, informative samples—those that challenge the model or expose failure

modes—tend to occupy sparse, underexplored regions rarely reached by naive sampling. This underscores a key limitation: *Despite high fidelity, DM-generated samples often fail to meet the utility demands of downstream tasks.*

To address this limitation, prior efforts can be broadly grouped into three categories: **(i)** scaling DM training on task-specific datasets to expand manifold coverage (Jimenez-Perez et al. 2025); **(ii)** incorporating reinforcement learning (Miao et al. 2024) or auxiliary objectives (Fan et al. 2024) to enhance generative diversity; and **(iii)** establishing feedback loops between the DM and the downstream model to enable joint optimization (Zhang et al. 2025; Frisch et al. 2025). However, retraining DMs or recollecting large-scale datasets purely for downstream gains contradicts the original goal of mitigating data scarcity under constrained computational and annotation budgets.

This work takes a different stance: the bottleneck lies not in the generator itself, but in the absence of utility-aware guidance. Modern DMs have the capacity to generate informative samples (Nolan et al. 2025), as evidenced by occasional high-loss outliers in Fig. 1. Yet, without explicit utility-aware signals, sampling traverses the manifold indiscriminately, leaving these valuable instances latent and rarely surfaced. This motivates a shift in perspective: *Instead of redesigning or retraining DMs, can we unlock their latent potential by rethinking the sampling process?*

Building on this insight, and as illustrated in Fig. 1(d), we introduce **Value-Guided Diffusion (VGD)**—a plug-and-play, training-free sampling strategy that transforms diffusion from *passive* random sampling into *active*, utility-driven generation. During denoising, VGD incorporates feedback from the downstream model to estimate a value score at each intermediate state, serving as a proxy for the sample's expected training utility. This feedback signal iteratively modulates the sampling trajectory, guiding generation toward high-utility regions of the generative manifold. Crucially, unlike adversarial perturbations (Chen et al. 2023) or diffusion-based augmentations (Hu and Shi 2025) that pursue utility at the cost of drifting off the generative manifold, VGD imposes a trust region constraint to reconcile fidelity and utility. This allows it to uncover naturally occurring high-value samples—such as rare anatomical variations, subtle foregrounds, and complex textures—that are typically missed by random sampling ($cf.$ Fig. 4).

In summary, our contributions are threefold:

- We identify a fundamental mismatch between generative supply and task-specific demand in diffusion-based data generation, wherein random sampling often fails to surface informative or task-relevant examples, limiting downstream segmentation performance.
- We introduce **VGD**, a plug-and-play sampling strategy that dynamically guides the diffusion process toward high-utility regions by leveraging feedback from downstream models.
- Extensive experiments on three medical benchmarks and diverse diffusion backbones demonstrate that VGD consistently improves downstream segmentation performance, validating its generality and practical utility.

Related Work

Generative Models. Recent advances in generative modeling include autoregressive transformers (Xie et al. 2024), flow-matching networks (Labs et al. 2025), and Diffusion Models (DMs) (Ho, Jain, and Abbeel 2020; Song et al. 2021). Among these, DMs—particularly Stable Diffusion (SD) (Rombach et al. 2022)—have emerged as the leading paradigm for medical imaging tasks, owing to their stability, flexibility, and open accessibility. SD variants offer strong visual priors, enabling effective transfer across tasks such as classification (Luo et al. 2024), reconstruction (Stringer and Pachitariu 2025), segmentation (Qiu et al. 2025), detection (Li et al. 2025b), and image-report alignment (Liu et al. 2024). Our work builds on this paradigm and specifically targets its limitations in downstream utility.

DM-Based Synthetic Data for Medical Segmentation. Recent efforts have centered on building high-performance DMs that synthesize high-fidelity, anatomically consistent image-mask pairs. For instance, (Chen et al. 2024) integrates anatomical priors, (Qiu et al. 2025) promote morphological diversity via real-image references, (Li et al. 2024) leverage point-cloud guidance to improve image–mask alignment; and (Zhang et al. 2024a) introduce texture control for fine-grained lesion generation. In contrast, our work shifts the focus from the generative supply side to task-driven demand, asking: *How can we extract high-utility samples from a fixed DM to maximize downstream performance?* Related motivations appear in GAUDA (Frisch et al. 2025), which heuristically resamples hard examples during training, and GenSeg (Zhang et al. 2025), which jointly trains the generator and segmenter via feedback loops. By contrast, VGD is training-free, model-agnostic, and readily deployable across diverse diffusion backbones, providing a simple yet effective approach to enhancing downstream segmentation.

Task-Guided Sampling in DMs aims to steer diffusion toward task-relevant outputs by leveraging the iterative and differentiable structure of the denoising process (Ye et al. 2024). A canonical approach is classifier guidance (Dhariwal and Nichol 2021), which injects gradients from an auxiliary noise classifier to condition sampling on target classes. This idea has been extended to a variety of downstream tasks, including out-of-distribution synthesis (Dhariwal and Nichol 2021), counterfactual generation (Wang et al. 2024), adversarial sample crafting (Chen et al. 2023; Xue et al. 2023), and data augmentation (Hu and Shi 2025; Shama Sastry, Dumpala, and Oore 2024). *Compared to these guidance techniques, VGD departs in three key ways:* (1) It eliminates the need for auxiliary networks by repurposing the downstream model itself as a value estimator, enabling a training-free, plug-and-play solution. (2) It explicitly targets downstream utility, as opposed to proxy goals such as attack success or class alignment. (3) It preserves visual fidelity while enhancing informativeness, avoiding the unnatural artifacts often introduced by conventional guidance methods.

Preliminaries

We begin by formalizing the task of generating high-utility samples. Let $\mathbf{x}_0 \in \mathcal{X}$ denote a medical image, and $\mathbf{m} =$

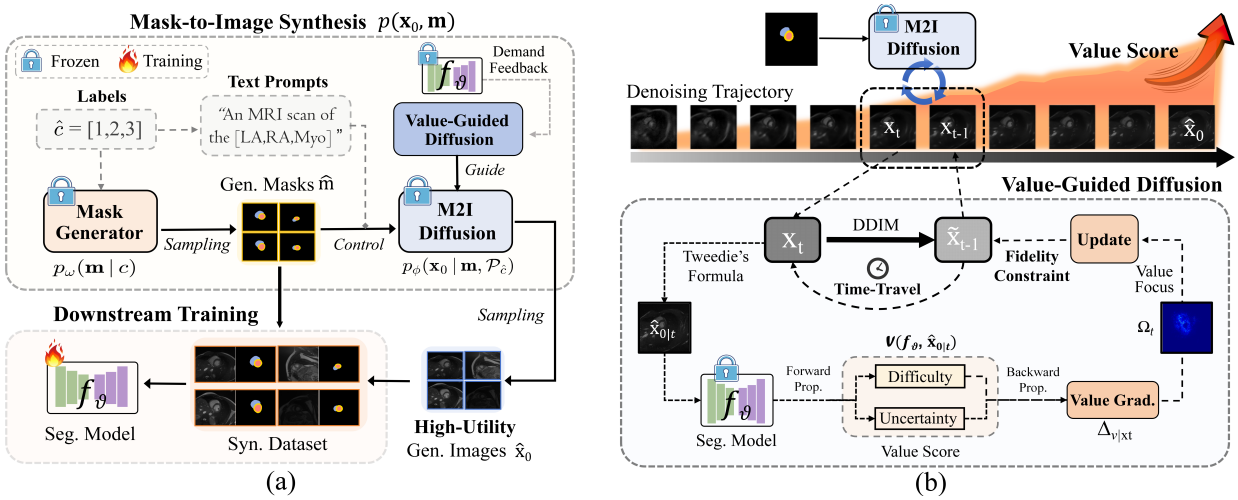

Figure 2: **(a) Image–mask generation** via a mask generator $p_\omega(\mathbf{m}|c)$ and an M2I DM $p_\phi(\mathbf{x}_0|\mathbf{m}, \mathcal{P}_{\hat{c}})$. **(b) VGD workflow:** intermediate states are iteratively guided toward higher-value regions along the denoising trajectory, thereby accumulating greater terminal value. Both p_ϕ and f_ϑ are used for inference and gradient computation, while remaining *frozen*.

$(\mathbf{s}, c) \in \mathcal{S} \times \mathcal{C}$ its semantic mask, where $\mathbf{s}$ is a spatial mask and c specifies the class label. Consider a pre-trained image-mask generative pipeline $p(\mathbf{x}_0, \mathbf{m})$ and a target segmentation model $f_\vartheta : \mathcal{X} \to \mathcal{S} \times \mathcal{C}$ trained on real data and awaiting synthetic augmentation. The goal of VGD is to sample from $p(\mathbf{x}_0, \mathbf{m})$ to maximize the downstream utility of f_ϑ.

Mask-to-Image Synthesis. In the medical domain, most diffusion-based generative pipelines for segmentation follow the Mask-to-Image (M2I) paradigm (Konz et al. 2024), which can be decomposed into a two-stage process:

$$(\hat{\mathrm{x}}_0, \hat{\mathrm{m}}) \sim p(\mathbf{x}_0, \mathbf{m}) \iff \begin{cases} \hat{\mathrm{m}} \sim p_\omega(\mathbf{m}|\hat{c}), \\ \hat{\mathrm{x}}_0 \sim p_\phi(\mathbf{x}_0|\hat{\mathrm{m}}, \mathcal{P}_{\hat{c}}). \end{cases} \tag{1}$$

As shown in Fig. 2(a), masks $\hat{\mathrm{m}}$ are first sampled from $p_\omega(\mathbf{m}|c)$ conditioned on labels $\hat{c} \sim \mathcal{C}$, using sources such as ground-truth, augmentations, or lightweight priors like VAEs (Sohn, Lee, and Yan 2015). Next, the M2I DM $p_\phi(\mathbf{x}_0|\mathbf{m}, \mathcal{P}_{\hat{c}})$ synthesizes a corresponding image $\hat{\mathrm{x}}_0$ conditioned on $\hat{\mathrm{m}}$ and an optional prompt $\mathcal{P}_{\hat{c}}$ (*e.g.*, "`A [Modality] scan of the [CLS]`"). The resulting $(\hat{\mathrm{x}}_0, \hat{\mathrm{m}})$ serve as synthetic training data for f_ϑ.

Diffusion Backbone. While VGD supports arbitrary iterative DMs, we instantiate p_ϕ with Stable Diffusion (v1.5) (Rombach et al. 2022), one of the most widely used diffusion backbones. SD operates in the latent space of a VQ-VAE (Van Den Oord, Vinyals et al. 2017), where a noise predictor $\mathcal{E}_\phi$ performs denoising over image latents. The text prompt $\mathcal{P}_{\hat{c}}$ is embedded by a frozen CLIP encoder (Radford et al. 2021). To support spatial conditioning, we incorporate ControlNet (Zhang, Rao, and Agrawala 2023) as a mask-guided branch and fine-tune the model following (Qiu et al. 2025). At inference, a random noise $\mathrm{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is iteratively denoised by $\mathcal{E}_\phi(\cdot, \mathcal{P}_{\hat{c}}, \hat{\mathrm{m}}, t)$ over timesteps $t : T \to 0$, producing a terminal latent x_0. This is subsequently decoded via the VQ-VAE decoder to obtain the final image $\hat{\mathrm{x}}_0 = \mathcal{D}_{\text{VAE}}(\mathrm{x}_0)$. See Supp. for more details.

High-Utility Sampling via VGD

Value Score as a Proxy for Sample Utility

In active learning, data utility is typically quantified via acquisition functions such as mutual information (Huang et al. 2024) or expected model change (Song, Zhang, and King 2023). However, these require access to real data and online model updates, making them unsuitable for generative scenarios where samples are synthesized on-the-fly and unavailable *a priori*. This motivates the need for a *training-free proxy* evaluable during sampling.

Value Criteria. As shown in Fig. 1, random sampling frequently yields redundant or trivial examples that contribute little to downstream learning. Drawing inspiration from active learning (Ribeiro Marnet et al. 2024), we argue that high-utility samples are those that reveal model weaknesses and offer greater optimization leverage. Specifically, such samples exhibit: **(1)** *segmentation difficulty*, indicating high empirical risk and promoting updates near the decision boundary; and **(2)** *predictive uncertainty*, capturing epistemic ambiguity or atypical structures with high informational value. Our goal is to extract such samples directly from p_ϕ based on these criteria, transforming sampling from a passive to a targeted, demand-driven process.

Value Score. To operationalize the above criteria, we define a differentiable *value score* $\mathcal{V}(f_\vartheta, \hat{\mathrm{x}}_0)$ that quantifies the utility of a generated image $\hat{\mathrm{x}}_0 \sim p_\phi$ *w.r.t.* the downstream model f_ϑ. Higher values indicate greater training utility and guide the sampling process:

$$\mathcal{V}(f_\vartheta, \hat{\mathrm{x}}_0) := \zeta\left(\mathcal{L}_{\text{CE}}(f_\vartheta(\hat{\mathrm{x}}_0), \hat{\mathrm{m}})\right) + \lambda \cdot \mathcal{M}_s(f_\vartheta(\hat{\mathrm{x}}_0)), \tag{2}$$

where $\hat{\mathrm{m}} \sim p_\omega$ is the target mask, $\mathcal{L}_{\text{CE}}$ is the normalized softmax cross-entropy loss (empirical risk), and $\zeta(\cdot)$ is a log-cosh transformation (Barron 2019) used to stabilize gradients. The uncertainty term $\mathcal{M}_s$ is computed via

$$\mathcal{M}_s(f_\vartheta(\hat{\mathrm{x}}_0)) = \frac{1}{|\chi|} \sum_{i \in \chi} 1 - \left(p_\vartheta(i)_{[1]} - p_\vartheta(i)_{[2]}\right), \tag{3}$$

where $p_\vartheta(i)_{[1]}$ and $p_\vartheta(i)_{[2]}$ denote the top-1 and top-2 predicted class probabilities at pixel i, and χ is the set of all pixels. Unlike entropy-based measures, this margin-based uncertainty isolates ambiguity between the most probable classes, offering sharper and more localized guidance during sampling (Ribeiro Marnet et al. 2024).

Maximizing Value under Fidelity Constraints

While $\mathcal{V}(f_\vartheta, \hat{\mathrm{x}}_0)$ acts as a utility proxy, its reliability hinges on the fidelity of $\hat{\mathrm{x}}_0$. Without constraints, optimization may cause $\hat{\mathrm{x}}_0$ to degenerate into adversarial noise that achieves high $\mathcal{V}$ yet provides limited supervisory value. To prevent degenerate solutions, we impose a *fidelity constraint* that restricts $\hat{\mathrm{x}}_0$ to the support of the conditional distribution $p(\mathbf{x}_0|\hat{\mathrm{m}}, \cdot)$ defined by the M2I model p_ϕ:

$$\hat{\mathrm{x}}_0 = \arg\max_{\hat{\mathrm{x}}_0} \mathcal{V}(f_\vartheta, \hat{\mathrm{x}}_0), \text{ s.t. } \hat{\mathrm{x}}_0 \in \mathrm{Supp}\left(p_\phi(\mathbf{x}_0|\hat{\mathrm{m}}, \cdot)\right). \tag{4}$$

In DMs, $\hat{\mathrm{x}}_0 \sim p_\phi$ is generated via a reverse denoising chain $\mathbf{x}_T \to \cdots \to \mathbf{x}_0$. Exploiting the Markov property, we reformulate the global constraint in Eq. (4) as a sequence of stepwise local constraints (Guo et al. 2024):

$$\hat{\mathrm{x}}_0 = \arg\max_{\hat{\mathrm{x}}_0} \mathcal{V}(f_\vartheta, \hat{\mathrm{x}}_0), \tag{5}$$

$$\hat{\mathrm{x}}_0 \sim \int \cdots \int p(\mathbf{x}_T) \prod_{t=1}^{T} p_\phi(\mathbf{x}_{t-1}|\mathbf{x}_t, \hat{\mathrm{m}}, \cdot)\, \mathrm{d}\mathbf{x}_{1:T}, \tag{6}$$

$$\text{s.t.} \quad \mathrm{x}_{t-1} \in \mathrm{Supp}\left(p_\phi(\mathbf{x}_{t-1}|\mathbf{x}_t, \hat{\mathrm{m}}, \cdot)\right), \quad \forall t \in [0, T], \tag{7}$$

where $p_\phi(\mathbf{x}_{t-1}|\mathbf{x}_t, \hat{\mathrm{m}}, \cdot)$ is the step-t transition kernel. This formulation enables value maximization within the local support of each denoising step, ensuring that the final sample $\hat{\mathbf{x}}_0$ remains aligned with the generative manifold.

Value Guidance in the Denoising Process

Solving Eqs. (5)–(7) is intractable, as $\hat{\mathrm{x}}_0$ only emerges at the final denoising step. To address this, we adopt a tractable approximation following (Ye et al. 2024). At each timestep t, we estimate a clean sample $\hat{\mathrm{x}}_{0|t}$ from latent x_t, evaluate $\mathcal{V}(f_\vartheta, \hat{\mathrm{x}}_{0|t})$, and update x_t via value-guided ascent. The updated latent propagates forward, progressively increasing the terminal score along the denoising trajectory.

Value Accumulation. As illustrated in Fig. 2(b), at each t, we first estimate a proxy $\hat{\mathrm{x}}_{0|t}$ of the clean image $\hat{\mathrm{x}}_0$ from the current latent x_t via a first-order approximation to Tweedie's formula (Efron 2011):

$$\hat{\mathrm{x}}_{0|t} = \mathcal{D}_{\mathrm{VAE}}\left(\frac{1}{\sqrt{\bar{\alpha}_t}}\left(\mathrm{x}_t - \sqrt{1-\bar{\alpha}_t}\cdot\mathcal{E}_\phi(\mathrm{x}_t, \mathcal{P}_{\hat{c}}, \hat{\mathrm{m}}, t)\right)\right), \tag{8}$$

where $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$, $\alpha_t = 1 - \beta_t$ and β_t defines the variance schedule. The estimated $\hat{\mathrm{x}}_{0|t}$ is evaluated by the frozen model f_ϑ to compute the intermediate score $\mathcal{V}(f_\vartheta, \hat{\mathrm{x}}_{0|t})$. To guide generation, we backpropagate through $\mathcal{V}$ $w.r.t.$ x_t to obtain the value gradient $\Delta_{v|\mathrm{x}_t}$:

$$\Delta_{v|\mathrm{x}_t} = \tau \cdot \mathrm{sign}\left(\frac{\partial \mathcal{V}(f_\vartheta, \hat{\mathrm{x}}_{0|t})}{\partial \hat{\mathrm{x}}_{0|t}} \cdot \frac{\partial \hat{\mathrm{x}}_{0|t}}{\partial \mathrm{x}_t}\right), \tag{9}$$

where $\tau \in \mathbb{R}^+$ is the step size, and $\mathrm{sign}(\cdot)$ denotes the sign function, resulting in a fixed-magnitude $\Delta_{v|\mathrm{x}_t}$. The latent update $\mathrm{x}_t \to \mathrm{x}_{t-1}$ combines DDIM sampling (Song, Meng, and Ermon 2021) with value-guided refinement:

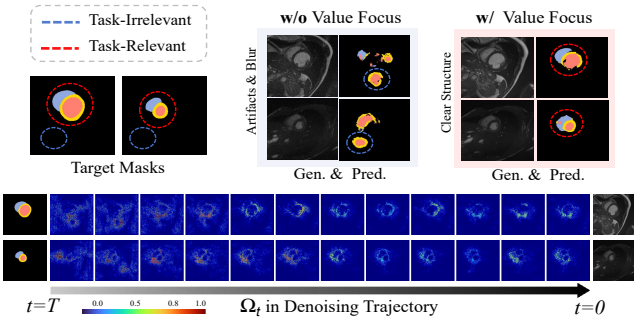

Figure 3: **Top:** Target mask $\hat{\mathrm{m}}$, generated image $\hat{\mathrm{x}}_0$, and predicted segmentation $f_\vartheta(\hat{\mathrm{x}}_0)$. **Bottom:** Ω_t amplify gradients in task-relevant structures, preventing shortcut solutions that exploit background artifacts to inflate value scores.

$$\tilde{\mathrm{x}}_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\cdot\hat{\mathrm{x}}_{0|t} + \sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2}\cdot\mathcal{E}_\phi(\mathrm{x}_t, \cdot) + \sigma_t\cdot\epsilon, \tag{10}$$

$$\mathrm{x}_{t-1} \leftarrow \tilde{\mathrm{x}}_{t-1} + \Delta_{v|\mathrm{x}_t}, \tag{11}$$

where $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is standard Gaussian noise, $\mathcal{E}_\phi(\mathrm{x}_t, \cdot)$ denotes the predicted noise as in Eq. (8), and $\sigma_t \in [0, 1]$ controls the stochasticity of sampling.

Fidelity Constraint. To ensure the refined latent x_{t-1} remains within the support of the local transition kernel, *i.e.*, $\mathrm{x}_{t-1} \in \mathrm{Supp}(p_\phi(\mathbf{x}_{t-1}|\mathbf{x}_t, \hat{\mathrm{m}}, \cdot))$, we impose a time-dependent constraint by projecting it onto an ℓ_∞ ball centered at the nominal DDIM prediction $\tilde{\mathrm{x}}_{t-1}$, with radius $\varepsilon_t = k \cdot \sqrt{\beta_t}$ for a small trust coefficient $k \ll 1$, *i.e.*,

$$\text{s.t.} \quad \|\mathrm{x}_{t-1} - \tilde{\mathrm{x}}_{t-1}\|_\infty \leqslant \varepsilon_t \approx \mathcal{O}(\sqrt{\beta_t}). \tag{12}$$

The noise scale $\sqrt{\beta_t}$ provides a conservative upper bound on the standard deviation of the transition kernel, thereby defining an $\mathcal{O}(\sqrt{\beta_t})$ trust region. This constraint reprojects updates back into the high-probability region of the generative path, preventing off-manifold drift due to over-optimization. As the projection normalizes the update magnitude, the step size η is fixed to 1 without loss of generality.

Time-Travel. While Eq. (12) enforces updates to remain on the manifold, its conservative nature may lead to undersaturated value scores. A naive remedy is to relax k, but this requires careful tuning. We adopt a more principled strategy, termed the *Time-Travel* (Lugmayr et al. 2022), wherein x_{t-1} is stochastically perturbed to revisit an earlier state x_t via

$$\mathrm{x}_t \leftarrow \sqrt{\alpha_t}\cdot\mathrm{x}_{t-1} + \sqrt{1-\alpha_t}\cdot\epsilon', \quad \epsilon' \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{13}$$

after which Eqs. (8)–(12) are re-executed. This stochastic rewind enables additional ascent steps while preserving fidelity, facilitating stable value accumulation.

Rebalancing Value Gradients via Value Focus

As shown in Fig. 3, although the procedure yields structurally plausible $\hat{\mathrm{x}}_0$, it frequently induces background artifacts that misalign with the target mask $\hat{\mathrm{m}}$. This issue arises

from uniformly applying value gradients $\Delta_{v|\mathrm{x}_t}$ across all pixels in Eq. (11). In medical images, background pixels dominate the spatial domain and dilute the gradient signal, suppressing updates to task-relevant regions like foregrounds and boundaries. Consequently, optimization becomes biased toward shortcut solutions that exploit background cues to spuriously boost value scores.

To address this, we propose a *value focus* mechanism that amplifies gradients in task-relevant regions ($e.g.$, segmentation targets and adjacent anatomy) while suppressing updates to irrelevant background. These regions are identified by measuring the local sensitivity of the value score to input perturbations. The core intuition is that task-critical pixels—typically located near decision boundaries—exert strong influence on model predictions, where minor perturbations can cause significant output changes (Xu et al. 2023). Accordingly, we approximate the sensitivity map $\mathcal{S}_t$ using the squared ℓ_2 norm of the per-pixel gradient:

$$\mathcal{S}_t(i) \approx \left\| \nabla_{\mathrm{x}_{t,i}} \mathcal{V}(f_\vartheta, \hat{\mathrm{x}}_{0|t}) \right\|_2^2 . \tag{14}$$

To stabilize sensitivity estimates over time, we maintain an Exponential Moving Average (EMA), $i.e.$,

$$\overline{\mathcal{S}}_t = \begin{cases} \mathcal{S}_T, & t = T, \\ (1-\gamma)\cdot \mathcal{S}_t + \gamma \cdot \overline{\mathcal{S}}_{t+1}, & t \in [0, T), \end{cases} \tag{15}$$

where $\gamma \in [0, 1]$ is the EMA decay rate. We apply min-max normalization to $\overline{\mathcal{S}}_t$ to obtain a focus mask $\Omega_t \in [0, 1]^{H \times W}$, which modulates the value gradient in Eq. (11):

$$\mathrm{x}_{t-1} \leftarrow \tilde{\mathrm{x}}_{t-1} + \Omega_t \odot \Delta_{v|\mathrm{x}_t}, \tag{16}$$

where $\odot$ denotes element-wise multiplication. As shown in Fig. 3, Ω_t focuses updates on foregrounds while suppressing spurious gradients in the background. This reduces artifacts in $\hat{\mathrm{x}}_0$ and yields more interpretable downstream errors.

Experiments

Experimental Setup

Datasets. We evaluate VGD on three medical benchmarks spanning distinct imaging modalities: *ACDC* (Bernard et al. 2018): cine-MRI scans with annotations for 3 cardiac structures; *Synapse Multi-Organ Segmentation* (Landman and Warfield 2015): clinical CT scans with 9 annotated abdominal organs; *Polyps- CVC-ClinicDB & Kvasir* (Bernal et al. 2015; Jha et al. 2019): RGB endoscopy images for colonic polyp segmentation with binary lesion masks. All datasets are resized to 256×256 and split into training (for generative backbone and f_ϑ pretraining) and test in a 7:3 ratio.

Implementation Details. Our *default generation backbone* uses a pre-trained SD v1.5 with ControlNet for M2I synthesis, and a conditional VAE for mask generation. All experiments are conducted on dual RTX 4090 GPUs. For each downstream task, we synthesize a dataset equal in size to the training set using DDIM (100 steps, η=0.2), with guidance scales set to 9.0 for Polyps and 7.5 for others. Unless otherwise noted, VGD adopts default hyperparameters: λ=1.0, τ=1.0, k=0.04, and γ=0.90. We adopt nnU-Net (Isensee et al. 2021) and SwinUNet (Cao et al. 2022) as segmentation backbones, using their official pretrained configurations. These models serve both as value estimators f_ϑ and as consistent anchors for evaluating downstream utility.

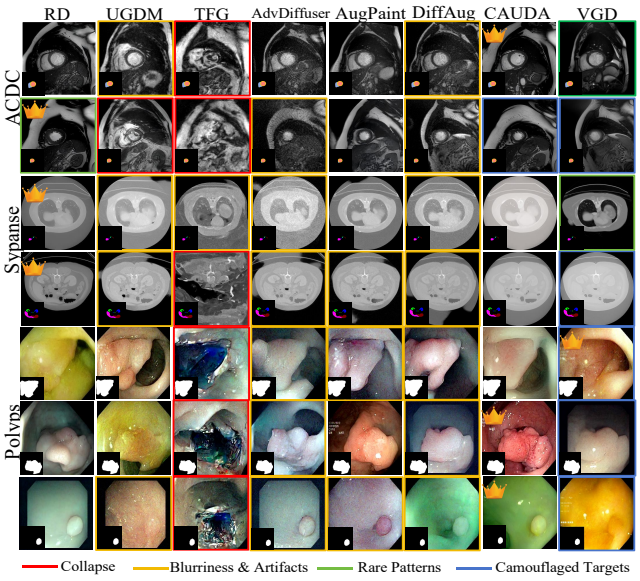

Figure 4: Qualitative comparison of generated samples. The "*Crown*" icon indicates the expert-annotated best-quality sample. Augmentation- and guidance-based methods often introduce exhibit instability, while VGD maintains visual fidelity and yields samples with higher downstream utility.

Metrics. We evaluate generative fidelity via Fréchet Inception Distance (FID↓) and CLIP-Image (CLIP-I↑). Segmentation performance is measured by Dice Similarity Coefficient (DSC%↑) and Average Surface Distance (ASD↓).

Comparative Evaluation

Compared Methods. We benchmark against seven methods selected based on the following criteria: *(i) explicitly designed to improve downstream performance*; *(ii) based on DMs*; and *(iii) require no retraining and are agnostic to the diffusion backbone.* These include: **Guided Sampling:** RD (random), uncertainty-guided UGDM (Luo et al. 2024), and general-purpose TFG (Ye et al. 2024) (with $\mathcal{V}(f_\vartheta, \hat{\mathrm{x}}_0)$ as guidance function); **Diffusion-based Augmentation:** adversarial attack (AdvDiffuser (Chen et al. 2023)), inpainting (AugPaint (Hu and Shi 2025)), and random perturbation (DiffAug (Shama Sastry, Dumpala, and Oore 2024)); **Post-hoc Scoring:** GAUDA (Frisch et al. 2025), re-implemented as a pre-generation selector using real-data-derived metrics. All methods share the same generation (*default*) and downstream model for fairness; see Supp. for details.

Results. As shown in Tab. 1, VGD consistently surpasses competing methods in segmentation performance by prioritizing high-utility samples within a fixed sampling budget. The slight reduction in low-level fidelity compared to unconstrained methods (RD and GAUDA) reflects a deliberate trade-off, which we further examine in the next subsection. Among Guided methods, TFG and UGDM integrate

Methods	Gen. Fidelity		ACDC				Synapse				Polyps			
			nnU-Net		SwinUNet		nnU-Net		SwinUNet		nnU-Net		SwinUNet	
	aFID↓	aCLIP-I↑	DSC↑	ASD↓	DSC↑	ASD↓	DSC↑	ASD↓	DSC↑	ASD↓	DSC↑	ASD↓	DSC↑	ASD↓
Baseline	-	-	87.3	1.95	88.7	1.54	73.2	19.8	76.6	11.4	78.9	7.32	81.8	5.77
RD	119.8	**0.772**	89.0±0.6	1.5±0.3	89.8±0.5	1.2±0.4	74.5±1.2	16.6±0.9	78.1±1.3	10.9±0.7	80.4±0.8	5.9±1.0	82.7±0.4	5.1±0.3
UGDM	145.1	0.680	87.2±1.0	1.8±0.4	88.6±0.9	1.4±0.3	73.2±0.8	21.0±0.9	75.9±0.9	16.2±0.6	78.2±0.6	6.5±0.5	80.5±0.6	6.7±0.5
TFG	189.5	0.433	78.5±0.6	4.7±0.6	82.7±0.8	3.9±0.5	68.9±1.7	24.6±2.7	69.3±1.2	23.9±1.5	72.5±1.5	9.7±0.9	77.3±1.2	8.4±0.7
AdvDiffuser	141.2	0.731	80.1±0.9	4.4±0.5	83.3±0.8	3.6±0.5	74.8±1.0	18.2±0.5	77.9±1.0	13.4±0.7	79.1±0.9	6.7±0.3	82.3±0.7	4.1±0.4
AugPaint	134.7	0.725	89.4±0.5	1.7±0.1	90.3±0.5	1.1±0.4	75.4±1.1	13.9±1.3	79.2±0.9	10.0±0.6	80.8±0.7	5.6±0.2	83.0±0.5	5.2±0.6
DiffAug	129.9	0.739	85.9±0.8	2.2±0.5	87.5±0.7	1.7±0.3	75.6±0.6	14.3±0.5	79.0±0.6	9.3±0.8	80.5±0.8	4.6±0.7	82.9±0.4	5.0±0.3
GAUDA	**116.5**	0.755	89.1±0.4	1.4±0.5	90.0±0.3	1.2±0.1	74.7±0.8	16.2±1.6	78.5±0.8	14.1±0.3	81.0±0.6	4.7±0.3	83.2±0.3	4.2±0.2
VGD	122.3	0.758	**90.5**±0.3	**1.1**±0.1	**91.2**±0.4	**0.9**±0.2	**76.6**±0.2	**12.4**±0.3	**80.3**±0.4	**8.9**±0.2	**82.1**±0.5	**3.8**±0.1	**83.8**±0.2	**3.5**±0.1

Table 1: Generative fidelity (averaged across all datasets) and downstream utility. Synthetic data are generated using $n_{\text{seed}} = 3$ random seeds. Downstream performance is reported as mean±std over the three runs. *Baseline:* real-data-only training.

Methods	Gen. Pipelines	ACDC				Synapse				Polyps			
		nnU-Net		SwinUNet		nnU-Net		SwinUNet		nnU-Net		SwinUNet	
	(Image & Mask)	DSC↑	ASD↓	DSC↑	ASD↓	DSC↑	ASD↓	DSC↑	ASD↓	DSC↑	ASD↓	DSC↑	ASD↓
Baseline	Real & GT	87.3	1.95	88.7	1.54	73.2	19.8	76.6	11.4	78.9	7.32	81.8	5.77
SegDiff	DPM & GT	87.6±0.5	1.9±0.6	89.1±0.4	1.4±0.4	73.5±0.8	19.6±1.1	77.9±1.2	12.3±0.7	78.6±0.7	7.1±0.7	81.3±0.9	6.6±0.5
+VGD		89.0±0.5	0.6±0.4	90.7±0.3	1.3±0.2	74.9±0.8	18.4±0.6	78.2±1.1	12.6±1.3	78.8±1.1	6.5±0.5	82.5±1.3	4.9±0.8
FairDiff	LDM & *Cloud.*	88.5±0.7	1.3±0.3	89.9 ±1.2	1.1±0.3	74.4±1.3	14.8±0.9	78.0±1.5	10.3±0.6	80.9±1.3	6.7±0.6	82.1±0.6	4.7±0.4
+VGD		89.2±0.6	0.9±0.3	91.0±0.8	0.9±0.6	75.7±0.8	13.9±0.3	79.1±0.3	9.0±0.5	81.5±0.7	4.6±0.3	83.0±0.8	4.1±0.2
SiameseDiff	SDv1.5 & GT†	88.9±0.8	1.3±0.1	90.1±1.0	1.3±0.6	75.2±1.1	15.4±1.9	78.3±0.8	12.7±0.5	81.3±0.8	5.2±0.5	82.8±0.5	4.9±0.3
+VGD		91.4±0.7	1.0±0.2	92.0±0.9	0.8±0.3	77.1±0.7	13.6±0.8	79.8±0.9	12.5±0.6	82.7±0.4	3.7±0.3	83.9±0.3	3.2±0.2
DiffBoost	SDv1.5 & GT	89.1±0.5	1.7±0.2	89.2±0.9	1.7±0.8	75.0±0.6	18.7±1.0	77.1±1.2	11.9±0.4	80.1±0.8	5.8±0.4	82.0±0.7	5.2±0.3
+VGD		90.3±0.4	1.5±0.5	91.5±0.3	1.2±0.5	76.2±0.2	12.9±0.8	79.5±0.4	10.3±0.5	81.6±0.6	4.5±0.3	83.6±0.4	3.8±0.1

Table 2: Integration of VGD into diverse M2I backbones ($n_{\text{seed}} = 3$). *Baseline:* real-data-only training. **Mask sources:** "GT" denotes ground-truth; "†" indicates shape augmentations derived from GT; "*Cloud.*" refers to point cloud–based masks.

task-aware signals but lack fidelity regularization, often resulting in over-optimized yet degraded samples (see Fig. 4). AugPaint and DiffAug promote sample diversity through perturbation-based augmentation, but rely on fixed priors and cannot adaptively align with task- or model-specific requirements, yielding limited downstream gains. GAUDA, while emphasizing hard examples, tends to oversample from familiar modes, leading to redundancy.

Generation Visualizations. As shown in Fig. 4, guidance- and augmentation-based methods often introduce artifacts or structural distortions. In contrast, VGD uncovers naturally occurring high-value instances—such as subtle foregrounds and complex textures—that are typically underrepresented in randomly sampled data.

Plug-and-Play Compatibility

We integrate VGD as a *drop-in* sampling module across four representative SOTA medical M2I backbones with diverse architectures and training paradigms: SegDiff (Konz et al. 2024), FairDiff (Li et al. 2024), SiameseDiff (Qiu et al. 2025), and DiffBoost (Zhang et al. 2024b). As official checkpoints are unavailable, all models are retrained following their respective configurations, with VGD replacing the default sampling process during synthetic data generation. As shown in Tab. 2, VGD consistently improves performance across all backbones, validating its effectiveness as a

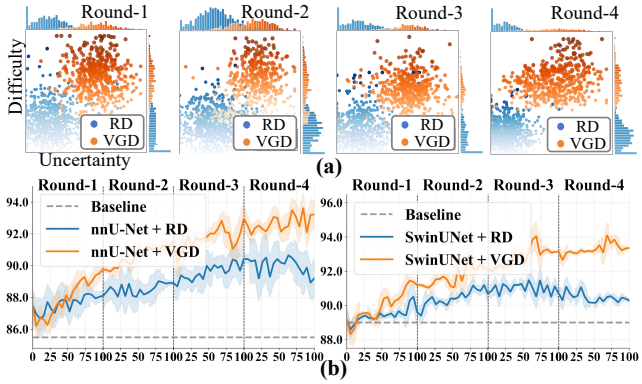

Figure 5: **(a)** Sample-wise normalized value scores of synthesized data across four synthesize-train rounds. Higher-scoring samples cluster in the upper-right region. **(b)** Performance (DSC) over training epochs using the corresponding data from (a). *Baseline*: training with real data only.

model-agnostic, plug-and-play sampler for boosting downstream utility without retraining. **Insight.** VGD's gains scale with the capacity of the diffusion backbone, *e.g.*, stronger on SiameseDiff (SD) than SegDiff (vanilla DPM). This is because VGD relies on accurate estimates of $\hat{x}_0$ in Eq. (8); low-capacity models introduce bias, which VGD may amplify, degrading guidance. To maximize utility gains, we recommend pairing VGD with high-capacity diffusion backbones.

ID	Core Components	Gen. Fidelity		nnU-Net		SwinUNet	
		FID↓	CLIP-I↑	DSC↑	ASD↓	DSC↑	ASD↓
–	RD (Baseline)	**156.2**	**0.837**	89.0	1.52	89.8	1.23
1	w/o $\mathcal{L}_{\text{CE}}$	161.9	0.810	89.8	1.59	89.2	1.33
2	w/o $\mathcal{M}_s$	159.4	0.826	89.4	1.38	90.3	1.19
3	w/o Ω_t	186.5	0.804	88.6	1.72	88.9	1.63
4	w/o Time Travel (T.t.)	157.8	0.825	89.9	1.43	90.5	1.06
5	w/o Fidelity Const. $\xi(\hat{\mathbf{x}}_0)$	-	-	Invalid due to sampling collapse			
6	**VGD (Ours)**	163.2	0.819	**90.5**	**1.14**	**91.2**	**0.93**

Table 3: **Ablation study** of VGD components on ACDC.

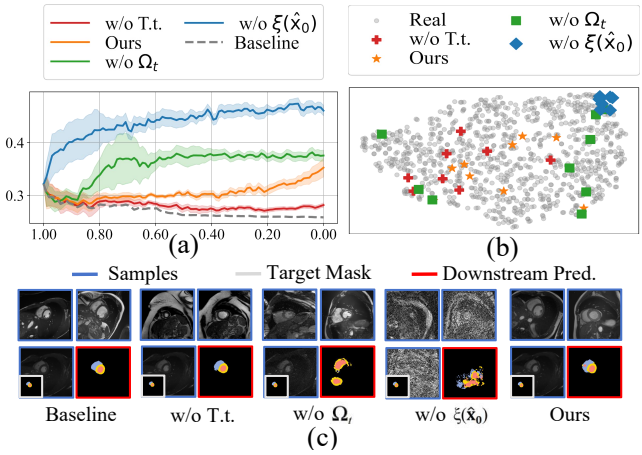

Figure 6: **(a)** Evolution of average value scores during denoising for a batch of samples. **(b)** t-SNE comparison between the sampled data from (a) and the real data distribution. **(c)** Qualitative comparison of sample quality under different settings, with corresponding downstream predictions.

Discussions

We perform diagnostic analyses to answer four key questions regarding VGD: **(Q1)** How well does the value score correlate with downstream utility? **(Q2)** What is the impact of each VGD component on overall performance? **(Q3)** How does the value score govern the fidelity–utility trade-off? **(Q4)** What is the sampling overhead introduced by VGD?

(Q1) Value Score as Utility Proxy. We evaluate whether the value score reliably reflects the training utility of samples by tracking its distribution and the corresponding downstream performance across synthesize–train rounds on ACDC, using RD as a reference. As shown in Fig. 5, VGD consistently selects higher-$\mathcal{V}$ samples, yielding sustained improvements across synthesize–train rounds. In contrast, RD suffers from rapid value degradation as models improve, making high-utility samples increasingly unlikely to be selected. This leads to noticeable stagnation—or even regression—in downstream performance by Round-3 for both models. The observed correlation confirms that the value score is a reliable and transferable proxy for downstream utility.

(Q2) Component Ablation. As shown in Tab. 3, removing either $\mathcal{L}_{\text{CE}}$ or $\mathcal{M}_s$ from the value score consistently impairs utility. These terms capture complementary aspects of sample utility—empirical risk and informativeness. Excluding either term weakens the expressiveness of the value score. Settings 3–6 isolate the contributions of guidance components, with their effects illustrated in Fig. 6. (1) Removing $\xi(\hat{\mathbf{x}}_0)$ leads to sampling collapse, as unconstrained value maximization causes drift off the data manifold—underscoring *fidelity as a prerequisite for meaningful optimization*. (2) Disabling Ω_t inflates the value of task-irrelevant regions, introducing background artifacts (see Fig. 6(c)). (3) Omitting T.t. hinders value saturation by preventing the accumulation of utility signals across timesteps. All components are indispensable to VGD's efficacy.

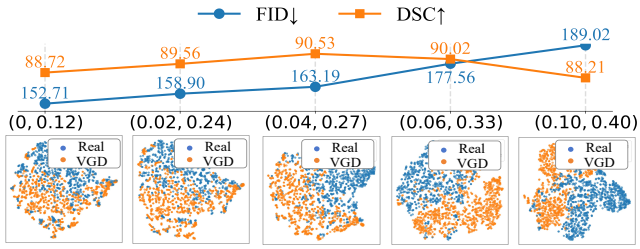

Figure 7: **Top:** Relationship between value level and generative fidelity as well as downstream utility (nnU-Net on ACDC). The x-axis $(\cdot, \cdot)$ denotes $(k, \overline{\mathcal{V}})$. **Bottom:** t-SNE of real data and VGD samples stratified by value level.

(Q3) Value–Fidelity–Utility. We vary the trust coefficient k in Eq. (12) to control the average value score $\overline{\mathcal{V}}$, and assess its effects on fidelity and downstream utility. Fig. 7 reveals two key trends: (1) *Higher value reduces fidelity*, reflecting a trade-off between task specificity and generative realism (Meng et al. 2022), as value-guided sampling shifts distributional support toward task-relevant but marginal regions. (2) *Moderate value increases improve utility*, while excessive value pursuit ($k \geqslant 0.06$) violates fidelity constraints, causing manifold drift ($cf.$ t-SNE) and degraded performance. **Takeaway.** Conservative thresholds ($e.g.$, k=0.04) achieve a stable fidelity–utility balance.

(Q4) Sampling Overhead. Under the same default settings, generating a 256×256 image takes 3.02s for the baseline (DDIM) and 4.85s with VGD, *introducing a ∼1.61× overhead.* As synthesis is typically performed offline prior to downstream training, this overhead remains acceptable.

Additional Results. Supp. provides additional analyses on sampling steps, EMA decay rate γ, value score variants, failure cases, and extended visualizations.

Conclusion

This work shifts the perspective from supply-driven to demand-driven generation, presenting VGD—a training-free, utility-aware sampling framework that leverages downstream feedback to guide diffusion models toward high-value samples. VGD is plug-and-play compatible with diverse medical M2I pipelines and consistently delivers substantial segmentation improvements without retraining.

Acknowledgments

This work is supported in part by the National Key R&D Program of China (2024YFB3311605) and the National Natural Science Foundation of China (62276112).

References

Barron, J. T. 2019. A general and adaptive robust loss function. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4331–4339.

Bernal, J.; Sánchez, F. J.; Fernández-Esparrach, G.; Gil, D.; Rodríguez, C.; and Vilariño, F. 2015. WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized medical imaging and graphics*, 43: 99–111.

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

Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; and Wang, M. 2022. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*, 205–218. Springer.

Chen, Q.; Chen, X.; Song, H.; Xiong, Z.; Yuille, A.; Wei, C.; and Zhou, Z. 2024. Towards generalizable tumor synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11147–11158.

Chen, X.; Gao, X.; Zhao, J.; Ye, K.; and Xu, C.-Z. 2023. Advdiffuser: Natural adversarial example synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4562–4572.

Chen, Y.; Liu, Y.; Lu, M.; Fu, L.; and Yang, F. 2025. Multi-consistency for semi-supervised medical image segmentation via diffusion models. *Pattern Recognition*, 161: 111216.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34: 8780–8794.

Efron, B. 2011. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496): 1602–1614.

Fan, C.; Zhu, M.; Chen, H.; Liu, Y.; Wu, W.; Zhang, H.; and Shen, C. 2024. Divergen: Improving instance segmentation by learning wider data distribution with more diverse generative data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3986–3995.

Frisch, Y.; Bornberg, C.; Fuchs, M.; and Mukhopadhyay, A. 2025. GAUDA: Generative Adaptive Uncertainty-guided Diffusion-based Augmentation for Surgical Segmentation. *arXiv preprint arXiv:2501.10819*.

Giuffrè, M.; and Shung, D. L. 2023. Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. *NPJ digital medicine*, 6(1): 186.

Guo, Y.; Yuan, H.; Yang, Y.; Chen, M.; and Wang, M. 2024. Gradient guidance for diffusion models: An optimization perspective. *Advances in Neural Information Processing Systems*, 37: 90736–90770.

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

Hu, X.; and Shi, Y. 2025. Inpainting is All You Need: A Diffusion-based Augmentation Method for Semi-supervised Medical Image Segmentation. *arXiv preprint arXiv:2506.23038*.

Huang, Y.; Pi, Y.; Shi, Y.; Guo, W.; and Wang, S. 2024. Adaptive graph active learning with mutual information via policy learning. *Expert Systems with Applications*, 255: 124773.

Isensee, F.; Jaeger, P. F.; Kohl, S. A.; Petersen, J.; and Maier-Hein, K. H. 2021. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2): 203–211.

Jha, D.; Smedsrud, P. H.; Riegler, M. A.; Halvorsen, P.; De Lange, T.; Johansen, D.; and Johansen, H. D. 2019. Kvasir-seg: A segmented polyp dataset. In *International conference on multimedia modeling*, 451–462. Springer.

Jimenez-Perez, G.; Osório, P.; Cersovsky, J.; Montalt-Tordera, J.; Hooge, J.; Vogler, S.; and Mohammadi, S. 2025. DiNO-Diffusion: Scaling Medical Diffusion Models via Self-Supervised Pre-Training. In *Annual Conference on Medical Image Understanding and Analysis*, 257–274. Springer.

Konz, N.; Chen, Y.; Dong, H.; and Mazurowski, M. A. 2024. Anatomically-controllable medical image generation with segmentation-guided diffusion models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 88–98. Springer.

Labs, B. F.; Batifol, S.; Blattmann, A.; Boesel, F.; Consul, S.; Diagne, C.; Dockhorn, T.; English, J.; English, Z.; Esser, P.; et al. 2025. FLUX. 1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space. *arXiv preprint arXiv:2506.15742*.

Landman, B. A.; and Warfield, S. K. 2015. Multi-Atlas Labeling Beyond the Cranial Vault - Workshop and Challenge. https://www.synapse.org/\#!Synapse:syn3193805/wiki/217789. MICCAI 2015 Workshop: Multi-Atlas Labeling Beyond the Cranial Vault.

Li, C.; Liu, X.; Li, W.; Wang, C.; Liu, H.; Liu, Y.; Chen, Z.; and Yuan, Y. 2025a. U-kan makes strong backbone for medical image segmentation and generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 4652–4660.

Li, W.; Xu, H.; Zhang, G.; Gao, H.-a.; Gao, M.; Wang, M.; and Zhao, H. 2024. Fairdiff: Fair segmentation with point-image diffusion. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 617–628. Springer.

Li, X.; Tan, X.; Chen, Z.; Zhang, Z.; Zhang, R.; Guo, R.; Jiang, G.; Chen, Y.; Qu, Y.; Ma, L.; et al. 2025b. One-for-more: Continual diffusion model for anomaly detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 4766–4775.

Liu, C.; Wan, Z.; Wang, H.; Chen, Y.; Qaiser, T.; Jin, C.; Yousefi, F.; Burlutskiy, N.; and Arcucci, R. 2024. Can Medical Vision-Language Pre-training Succeed with Purely Synthetic Data? *arXiv preprint arXiv:2410.13523*.

Lugmayr, A.; Danelljan, M.; Romero, A.; Yu, F.; Timofte, R.; and Van Gool, L. 2022. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11461–11471.

Luo, Y.; Yang, Q.; Fan, Y.; Qi, H.; and Xia, M. 2024. Measurement guidance in diffusion models: Insight from medical image synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Meng, C.; He, Y.; Song, Y.; Song, J.; Wu, J.; Zhu, J.-Y.; and Ermon, S. 2022. SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations. In *International Conference on Learning Representations (ICLR)*.

Miao, Z.; Wang, J.; Wang, Z.; Yang, Z.; Wang, L.; Qiu, Q.; and Liu, Z. 2024. Training diffusion models towards diverse image generation with reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10844–10853.

Nie, D.; Trullo, R.; Lian, J.; Wang, L.; Petitjean, C.; Ruan, S.; Wang, Q.; and Shen, D. 2018. Medical image synthesis with deep convolutional adversarial networks. *IEEE Transactions on Biomedical Engineering*, 65(12): 2720–2730.

Nolan, O.; Stevens, T. S.; van Nierop, W. L.; and van Sloun, R. J. 2025. Active Diffusion Subsampling. *Transactions on Machine Learning Research*, 2025.

Qiu, K.; Gao, Z.; Zhou, Z.; Sun, M.; and Guo, Y. 2025. Noise-Consistent Siamese-Diffusion for Medical Image Synthesis and Segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 15672–15681.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Ribeiro Marnet, L.; Brodskiy, Y.; Grasshof, S.; and Wasowski, A. 2024. Uncertainty driven active learning for image segmentation in underwater inspection. In *International Conference on Robotics, Computer Vision and Intelligent Systems*, 66–81. Springer.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Shama Sastry, C.; Dumpala, S. H.; and Oore, S. 2024. DiffAug: A Diffuse-and-Denoise Augmentation for Training Robust Classifiers. *Advances in Neural Information Processing Systems*, 37: 20745–20785.

Sohn, K.; Lee, H.; and Yan, X. 2015. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28.

Song, J.; Meng, C.; and Ermon, S. 2021. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations*.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2021. Score-Based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations*.

Song, Z.; Zhang, Y.; and King, I. 2023. No change, no gain: Empowering graph neural networks with expected model change maximization for active learning. *Advances in neural information processing systems*, 36: 47511–47526.

Stringer, C.; and Pachitariu, M. 2025. Cellpose3: one-click image restoration for improved cellular segmentation. *Nature Methods*, 1–8.

Van Den Oord, A.; Vinyals, O.; et al. 2017. Neural discrete representation learning. *Advances in neural information processing systems*, 30.

Wang, K.; Chen, Z.; Zhu, M.; Li, Z.; Weng, J.; and Gu, T. 2024. Score-based counterfactual generation for interpretable medical image classification and lesion localization. *IEEE transactions on medical imaging*.

Xie, J.; Mao, W.; Bai, Z.; Zhang, D. J.; Wang, W.; Lin, K. Q.; Gu, Y.; Chen, Z.; Yang, Z.; and Shou, M. Z. 2024. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*.

Xu, Y.; Sun, Y.; Goldblum, M.; Goldstein, T.; and Huang, F. 2023. Exploring and Exploiting Decision Boundary Dynamics for Adversarial Robustness. The Eleventh International Conference on Learning Representations.

Xue, H.; Araujo, A.; Hu, B.; and Chen, Y. 2023. Diffusion-based adversarial sample generation for improved stealthiness and controllability. *Advances in Neural Information Processing Systems*, 36: 2894–2921.

Ye, H.; Lin, H.; Han, J.; Xu, M.; Liu, S.; Liang, Y.; Ma, J.; Zou, J. Y.; and Ermon, S. 2024. Tfg: Unified training-free guidance for diffusion models. *Advances in Neural Information Processing Systems*, 37: 22370–22417.

Zhang, H.; Liu, Y.; Yang, J.; Wan, S.; Wang, X.; Peng, W.; and Fua, P. 2024a. LeFusion: Controllable Pathology Synthesis via Lesion-Focused Diffusion Models. *arXiv preprint arXiv:2403.14066*.

Zhang, L.; Jindal, B.; Alaa, A.; Weinreb, R.; Wilson, D.; Segal, E.; Zou, J.; and Xie, P. 2025. Generative AI enables medical image segmentation in ultra low-data regimes. *Nature Communications*, 16(1): 6486.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3836–3847.

Zhang, Z.; Yao, L.; Wang, B.; Jha, D.; Durak, G.; Keles, E.; Medetalibeyoglu, A.; and Bagci, U. 2024b. Diffboost: Enhancing medical image segmentation via text-guided diffusion model. *IEEE Transactions on Medical Imaging*.