Supplementary Material for "Diffusion-Based Native Adversarial Synthesis for Enhanced Medical Segmentation Generalization"

Hongyu Zhang[1,2] **Haipeng Chen**[1,2] **Zhimin Xu**[3] **Chengxin Yang**[1,2] **Yingda Lyu**[4]✉

[1]College of Computer Science and Technology, Jilin University

[2]Key Laboratory of Symbolic Computation and Knowledge Engineering of the Ministry of Education, Jilin University

[3]Hospital of Stomatology, Jilin University [4]Public Computer Education and Research Center, Jilin University

{zhanghongyu22, yangcx24}@mails.jlu.edu.cn {chenhp, xuzhimin, ydlv✉}@jlu.edu.cn

Contents

1. Theoretical Analysis

1.1. Multi-step extension of the generalization gain

We extend the one-step analysis in Eq. (4) of the main text to K gradient steps on $\mathcal{U}_{\text{syn}}$ and show that the same projection structure persists.

Let $\ell_{\text{seg}} : \mathbb{R}^D \to \mathbb{R}_{\geq 0}$ denote a *continuously differentiable* empirical segmentation loss, and let $\vartheta_0 \in \mathbb{R}^d$ be the initial parameters of the downstream segmentation model f_ϑ. We perform $K > 1$ steps of gradient descent on a synthetic dataset $\mathcal{U}_{\text{syn}}$, using a fixed learning rate $0 < \gamma \ll 1$:

$$\vartheta_{k+1} = \vartheta_k - \gamma \nabla_\vartheta \ell_{\text{seg}}(\mathcal{U}_{\text{syn}}; \vartheta_k), \quad k = 0, \ldots, K-1. \tag{11}$$

Define the synthetic gradient at step $k \in [0, K)$ and the real-data gradient at initialization be defined as

$$g_k := \nabla_\vartheta \ell_{\text{seg}}(\mathcal{U}_{\text{syn}}; \vartheta_k) \in \mathbb{R}^d, \;\; g_{\text{real}} := \nabla_\vartheta \ell_{\text{seg}}(\mathcal{U}_{\text{real}}; \vartheta_0) \in \mathbb{R}^d. \tag{12}$$

The K-step generalization gain is defined as

$$\mathcal{G}_\vartheta^{(K)}(\mathcal{U}_{\text{syn}}) \in \mathbb{R} := \ell_{\text{seg}}(\mathcal{U}_{\text{real}}; \vartheta_0) - \ell_{\text{seg}}(\mathcal{U}_{\text{real}}; \vartheta_K). \tag{13}$$

Unrolling the update recursion yields the total parameter displacement after K steps as

$$\begin{aligned} \Delta\vartheta^{(K)} &:= \vartheta_K - \vartheta_0 \\ &= \sum_{k=0}^{K-1} (\vartheta_{k+1} - \vartheta_k) = -\gamma \sum_{k=0}^{K-1} g_k. \end{aligned} \tag{14}$$

We view $\Delta\vartheta^{(K)}$ as a function of γ. Expanding $\ell_{\text{seg}}(\mathcal{U}_{\text{real}}; \cdot)$ via its Taylor series in the direction $\Delta\vartheta^{(K)}$, we obtain:

$$\begin{aligned} &\ell_{\text{seg}}(\mathcal{U}_{\text{real}}; \vartheta_0 + \Delta\vartheta^{(K)}) \\ &= \sum_{m=0}^{\infty} \frac{1}{m!} D^m \ell_{\text{seg}}(\mathcal{U}_{\text{real}}; \vartheta_0)\big[\Delta\vartheta^{(K)}, \ldots, \Delta\vartheta^{(K)}\big] \\ &= \ell_{\text{seg}}(\mathcal{U}_{\text{real}}; \vartheta_0) + D\ell_{\text{seg}}(\mathcal{U}_{\text{real}}; \vartheta_0)\big[\Delta\vartheta^{(K)}\big] \\ &\quad + \sum_{m=2}^{\infty} \frac{1}{m!} D^m \ell_{\text{seg}}(\mathcal{U}_{\text{real}}; \vartheta_0)\big[\Delta\vartheta^{(K)}, \ldots, \Delta\vartheta^{(K)}\big], \end{aligned} \tag{15}$$

where $D^m \ell(\vartheta_0)$ denotes the m-th order derivative tensor evaluated at ϑ_0. Since $D\ell(\vartheta_0)[\cdot] = \nabla_\vartheta \ell(\vartheta_0)^\top(\cdot)$, the first-order term reduces to

$$D\ell_{\text{seg}}(\mathcal{U}_{\text{real}}; \vartheta_0)\big[\Delta\vartheta^{(K)}\big] = g_{\text{real}}^\top \Delta\vartheta^{(K)}, \tag{16}$$

✉ Corresponding author.

linking the descent path to the gradient at initialization. Substituting Eq. (15) into the definition of $\mathcal{G}_{\vartheta}^{(K)}(\mathcal{U}_{\text{syn}})$ yields

$$\begin{aligned}\mathcal{G}_{\vartheta}^{(K)}(\mathcal{U}_{\text{syn}}) &= \ell_{\text{seg}}(\mathcal{U}_{\text{real}};\vartheta_0) - \ell_{\text{seg}}(\mathcal{U}_{\text{real}};\vartheta_0+\Delta\vartheta^{(K)}) \\ &= \ell_{\text{seg}}(\mathcal{U}_{\text{real}};\vartheta_0) - \Bigg[\ell_{\text{seg}}(\mathcal{U}_{\text{real}};\vartheta_0) + g_{\text{real}}^{\top}\Delta\vartheta^{(K)} \\ &\qquad + \sum_{m=2}^{\infty}\frac{1}{m!}D^m\ell_{\text{seg}}(\mathcal{U}_{\text{real}};\vartheta_0)\big[\Delta\vartheta^{(K)},\dots,\Delta\vartheta^{(K)}\big]\Bigg] \\ &= -g_{\text{real}}^{\top}\Delta\vartheta^{(K)} - \sum_{m=2}^{\infty}\frac{1}{m!}D^m\ell_{\text{seg}}(\mathcal{U}_{\text{real}};\vartheta_0)\big[\Delta\vartheta^{(K)},\dots,\Delta\vartheta^{(K)}\big],\end{aligned} \tag{17}$$

where the first term is linear in $\Delta\vartheta^{(K)}$, and the second collects higher-order effects. Substituting Eq. (14), the linear term becomes

$$-g_{\text{real}}^{\top}\Delta\vartheta^{(K)} = -g_{\text{real}}^{\top}\big(-\gamma\sum_{k=0}^{K-1}g_k\big) = \gamma\,\bar{g}_{\text{syn}}^{(K)\top}g_{\text{real}},$$

where $\bar{g}_{\text{syn}}^{(K)} := \sum_{k=0}^{K-1} g_k \in \mathbb{R}^d$. is the aggregated synthetic gradient. Thus, the generalization gain rewrites as

$$\begin{aligned}&\mathcal{G}_{\vartheta}^{(K)}(\mathcal{U}_{\text{syn}}) = \\ &\gamma\,\bar{g}_{\text{syn}}^{(K)\top}g_{\text{real}} - \sum_{m=2}^{\infty}\frac{1}{m!}D^m\ell_{\text{seg}}(\mathcal{U}_{\text{real}};\vartheta_0)\big[\Delta\vartheta^{(K)},\dots,\Delta\vartheta^{(K)}\big].\end{aligned} \tag{18}$$

Eq. (18) takes the form of a power series in γ, where the leading term is linear and depends on the inner product between $\bar{g}_{\text{syn}}^{(K)\top}$ and g_{real}, while higher-order terms are at least quadratic. Applying the Cauchy–Schwarz inequality [3], we express the first-order term as

$$\bar{g}_{\text{syn}}^{(K)\top}g_{\text{real}} = \big\|\bar{g}_{\text{syn}}^{(K)}\big\|_2\,\big\|g_{\text{real}}\big\|_2\cos\zeta^{(K)}, \tag{19}$$

where $\zeta^{(K)}$ is the angle between the two vectors. Hence, the leading term admits a geometric projection interpretation:

$$\gamma\,\bar{g}_{\text{syn}}^{(K)\top}g_{\text{real}} = \gamma\big\|\bar{g}_{\text{syn}}^{(K)}\big\|_2\big\|g_{\text{real}}\big\|_2\cos\zeta^{(K)}. \tag{20}$$

To bound the residual, we assume $\ell_{\text{seg}}(\mathcal{U}_{\text{real}};\cdot)$ is β-smooth [24], *i.e.*,

$$\big\|\nabla_{\vartheta}\ell_{\text{seg}}(\mathcal{U}_{\text{real}};\vartheta) - \nabla_{\vartheta}\ell_{\text{seg}}(\mathcal{U}_{\text{real}};\vartheta')\big\|_2 \le \beta\|\vartheta-\vartheta'\|_2. \tag{21}$$

Under this assumption, the Taylor remainder is upper-bounded as

$$\left|\sum_{m=2}^{\infty}\frac{1}{m!}D^m\ell_{\text{seg}}(\mathcal{U}_{\text{real}};\vartheta_0)\big[\Delta\vartheta^{(K)},\dots,\Delta\vartheta^{(K)}\big]\right| \le \frac{\beta}{2}\big\|\Delta\vartheta^{(K)}\big\|_2^2. \tag{22}$$

Using Eq. (14) and the triangle inequality, and assuming $\|g_k\|_2 \le \mathcal{G}_{\max}$, We obtain

$$\big\|\Delta\vartheta^{(K)}\big\|_2^2 \le \gamma^2\Big(\sum_{k=0}^{K-1}\|g_k\|_2\Big)^2 \le \gamma^2(K\mathcal{G}_{\max})^2 = \gamma^2K^2\mathcal{G}_{\max}^2, \tag{23}$$

leading to the bound:

$$\left|\sum_{m=2}^{\infty}\frac{1}{m!}D^m\ell_{\text{seg}}(\mathcal{U}_{\text{real}};\vartheta_0)\big[\Delta\vartheta^{(K)},\dots,\Delta\vartheta^{(K)}\big]\right| \le \frac{\beta}{2}\gamma^2K^2\mathcal{G}_{\max}^2. \tag{24}$$

Combining with Eq. (18), we recover the K-step analogue of Eq. (4):

$$\mathcal{G}_{\vartheta}^{(K)}(\mathcal{U}_{\text{syn}}) = \gamma\big\|\bar{g}_{\text{syn}}^{(K)}\big\|_2\big\|g_{\text{real}}\big\|_2\cos\zeta^{(K)} + \mathcal{O}\big(\beta\gamma^2K^2\mathcal{G}_{\max}^2\big), \tag{25}$$

For small γ and moderate K , the second-order term becomes negligible and the projection term dominates. If the synthetic gradient direction remains nearly constant across steps (*i.e.*, $g_k \approx g_0, \forall k$), then

$$\bar{g}_{\text{syn}}^{(K)} = \sum_{k=0}^{K-1}g_k \approx \sum_{k=0}^{K-1}g_0 = Kg_0, \tag{26}$$

and Eq. (25) simplifies to

$$\begin{aligned}\mathcal{G}_{\vartheta}^{(K)}(\mathcal{U}_{\text{syn}}) &\approx \gamma\big\|Kg_0\big\|_2\big\|g_{\text{real}}\big\|_2\cos\zeta \\ &= K\gamma\,\big\|\nabla_{\vartheta}\ell_{\text{seg}}(\mathcal{U}_{\text{syn}};\vartheta_0)\big\|_2\,\big\|\nabla_{\vartheta}\ell_{\text{seg}}(\mathcal{U}_{\text{real}};\vartheta_0)\big\|_2\cos\zeta.\end{aligned} \tag{27}$$

demonstrating that multi-step generalization gain scales linearly in K, preserving the alignment-based interpretation from the one-step case.

1.2. Training-Free Dual Modeling for Skewed Distributions

By taking the logarithm and computing the gradient $w.r.t.$ $\mathbf{x}$ in Eq. (6), we obtain:

$$\begin{aligned}&\nabla_{\mathbf{x}}\log q_{\phi}^{\text{adv}}(\mathbf{x}\mid\hat{\mathbf{y}}_s,\cdot) \\ &= \nabla_{\mathbf{x}}\Big[\log q_{\phi}(\mathbf{x}\mid\hat{\mathbf{y}}_s,\cdot) + \lambda\,\ell_{\text{seg}}(f_{\vartheta}(\mathbf{x}),\hat{\mathbf{y}}_s) - \log\mathcal{Z}(\hat{\mathbf{y}}_s)\Big] \\ &= \nabla_{\mathbf{x}}\log q_{\phi}(\mathbf{x}\mid\hat{\mathbf{y}}_s,\cdot) + \lambda\,\nabla_{\mathbf{x}}\ell_{\text{seg}}(f_{\vartheta}(\mathbf{x}),\hat{\mathbf{y}}_s).\end{aligned} \tag{28}$$

Following the diffusion trajectory, we set $\mathbf{x}=\mathbf{x}_t$ and define the skewed score as:

$$\begin{aligned}s_{\phi}^{\text{adv}}(\mathbf{x}_t,t\mid\hat{\mathbf{y}}_s,\cdot) &:= \nabla_{\mathbf{x}_t}\log q_{\phi}^{\text{adv}}(\mathbf{x}_t\mid\hat{\mathbf{y}}_s,\cdot) \\ &= \nabla_{\mathbf{x}_t}\log q_{\phi}(\mathbf{x}_t\mid\hat{\mathbf{y}}_s,\cdot) + \lambda\,\nabla_{\mathbf{x}_t}\ell_{\text{seg}}\big(f_{\vartheta}(\mathbf{x}_t),\hat{\mathbf{y}}_s\big) \\ &\approx s_{\phi}(\mathbf{x}_t,t\mid\hat{\mathbf{y}}_s,\cdot) + \lambda\,\nabla_{\mathbf{x}_t}\ell_{\text{seg}}\big(f_{\vartheta}(\hat{\mathbf{x}}_{0|t}),\hat{\mathbf{y}}_s\big),\end{aligned} \tag{29}$$

where $s_{\phi}(\mathbf{x}_t,t\mid\hat{\mathbf{y}}_s,\cdot) \approx \nabla_{\mathbf{x}_t}\log q_{\phi}(\mathbf{x}_t\mid\hat{\mathbf{y}}_s,\cdot)$ is the base score function, and $\hat{\mathbf{x}}_{0|t}$ is evaluated on a Tweedie-denoised estimate:

$$\hat{\mathbf{x}}_{0|t} = \frac{\mathbf{x}_t + (1-\bar{\alpha}_t)\,s_{\phi}(\mathbf{x}_t,t\mid\hat{\mathbf{y}}_s,\cdot)}{\sqrt{\bar{\alpha}_t}}, \qquad \bar{\alpha}_t = \prod_{i=1}^{t}\alpha_i. \tag{30}$$

Together, Eqs. (29) and (30) define a training-free guidance rule at inference time, where the skewed score simply augments the base diffusion score with the gradient of a downstream loss evaluated on the denoised reconstruction. This yields a dual formulation to the energy-based preference modeling in Eq. (6), enabling adversarial modulation without additional model training or architectural modification.

1.3. Solution Space of Eq. (10) Lies Within the Subspace of Eq. (6)

We show that the optimal solutions induced by the Native Adversariality Mining (NAM) objective in Eq. (10) are contained in the energy-tilted family defined by Eq. (6).

Fix a mask $\hat{\mathrm{y}}_s \sim q_\omega$, and denote the base diffusion distribution by $q(\mathbf{x}) := q_\phi(\mathbf{x} \mid \hat{\mathrm{y}}_s, \cdot)$. Let $P_\xi(\mathbf{x})$ be the distribution of $\hat{\mathbf{x}}_s^r$ generated via DDIM sampling from noise $\hat{\mathbf{x}}_T^r \sim \mathcal{N}^r(\Delta_{\boldsymbol{\mu}}, \mathbf{I}+\Delta_{\boldsymbol{\Sigma}})$ ($cf.$ Sec. 3.3). Since DDIM defines a deterministic and shared mapping from noise to image space, the KL divergence between the shifted distribution and the denoising prior is preserved under pushforward:

$$\ell_{\mathrm{KL}}\big(\mathcal{N}^r(\Delta_{\boldsymbol{\mu}}, \mathbf{I}+\Delta_{\boldsymbol{\Sigma}}) \,\|\, \mathcal{N}(\mathbf{0}, \mathbf{I})\big) = \mathrm{KL}\big(P_\xi(\mathbf{x}) \,\|\, q(\mathbf{x})\big). \quad (31)$$

Letting $u(\mathbf{x}, \hat{\mathrm{y}}_s) := \min\big(\kappa_{\mathrm{up}}, \ell_{\mathrm{seg}}(f_\theta(\mathbf{x}), \hat{\mathrm{y}}_s)\big)$, the conditional NAM objective becomes:

$$J(P_\xi) := \mathbb{E}_{\mathbf{x}\sim P_\xi}\big[u(\mathbf{x}, \hat{\mathrm{y}}_s)\big] - \beta\, \mathrm{KL}\big(P_\xi(\mathbf{x}) \,\|\, q(\mathbf{x})\big). \quad (32)$$

with $P_\xi \in \mathcal{P}_\xi := \{P_\xi : \xi \in \Xi\}$. In contrast, Eq. (6) optimizes over all $P \ll q$, leading to the unconstrained variational problem:

$$\max_{P \ll q} \mathbb{E}_{\mathbf{x}\sim P}\big[u(\mathbf{x}, \hat{\mathrm{y}}_s)\big] - \beta\, \mathrm{KL}\big(P(\mathbf{x}) \,\|\, q(\mathbf{x})\big), \quad (33)$$

where the maximization is over all densities P such that $P \ll q$. Introducing a Lagrange multiplier α for the normalization constraint $\int P(\mathbf{x})d\mathbf{x} = 1$, we define the functional

$$\mathcal{L}[P] := \int P(\mathbf{x})u(\mathbf{x}, \hat{\mathrm{y}}_s)\, d\mathbf{x} - \beta \int P(\mathbf{x}) \log \frac{P(\mathbf{x})}{q(\mathbf{x})}\, d\mathbf{x} + \alpha\left(\int P(\mathbf{x})\, d\mathbf{x} - 1\right). \quad (34)$$

Taking the first variation and solving $\delta\mathcal{L}/\delta P = 0$ yields the unique maximizer:

$$\frac{\delta\mathcal{L}}{\delta P(\mathbf{x})} = u(\mathbf{x}, \hat{\mathrm{y}}_s) - \beta\Big(\log \frac{P(\mathbf{x})}{q(\mathbf{x})} + 1\Big) + \alpha = 0. \quad (35)$$

Solving for $P(\mathbf{x})$ gives

$$\log P^\star(\mathbf{x}) = \log q(\mathbf{x}) + \frac{1}{\beta} u(\mathbf{x}, \hat{\mathrm{y}}_s) + C,$$
$$P^\star(\mathbf{x}) \propto q(\mathbf{x})\, \exp\Big(\frac{1}{\beta} u(\mathbf{x}, \hat{\mathrm{y}}_s)\Big), \quad (36)$$

where C absorbs the constants. Thus the unique maximizer of Eq. (33) over all $P \ll q$ is the energy-tilted distribution

$$P^\star(\mathbf{x} \mid \hat{\mathrm{y}}_s) = \frac{1}{\mathcal{Z}(\hat{\mathrm{y}}_s)}\, q(\mathbf{x})\, \exp\Big(\lambda\, u(\mathbf{x}, \hat{\mathrm{y}}_s)\Big), \quad \lambda := \frac{1}{\beta}, \quad (37)$$

which is exactly of the form of Eq. (6), with the energy given by the (clipped) adversariality u.

Meanwhile, Eq. (10) optimizes $J(P_\xi)$ in Eq. (32) over the restricted family $\mathcal{P}_\xi$ induced by the miner $\mathcal{M}_\xi$. Denote the (possibly non-unique) maximizers by $\mathcal{P}_\xi^\star := \arg\max_{P_\xi \in \mathcal{P}_\xi} J(P_\xi)$. By construction,

$$\mathcal{P}_\xi^\star \subseteq \arg\max_{P \ll q} \Big\{\mathbb{E}_{\mathbf{x}\sim P}\big[u(\mathbf{x}, \hat{\mathrm{y}}_s)\big] - \beta\, \mathrm{KL}\big(P(\mathbf{x}) \,\|\, q(\mathbf{x})\big)\Big\}, \quad (38)$$

Hence, the solution space of Eq. (10) lies within the exponential tilting family in Eq. (6), up to the representational capacity of $\mathcal{M}_\xi$ and the clipping in u. In other words, NAM implements a restricted energy tilt of the base diffusion model via reweighted noise sampling, without modifying either DM or sampling procedure.

1.4. Gradient Approximation

Recall the NAM objective from Eq. (10):

$$\boldsymbol{\Omega}(\xi) := \arg\max_{\xi} \mathbb{E}_{\hat{\mathbf{y}}_s \sim q_\omega,\ \hat{\mathbf{x}}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \Big[\ell_{\mathrm{adv}}\big(f_\theta(\hat{\mathbf{x}}_s^r), \hat{\mathbf{y}}_s\big) - \beta\, \ell_{\mathrm{KL}}\big(\mathcal{N}^r(\Delta_{\boldsymbol{\mu}}, \mathbf{I}+\Delta_{\boldsymbol{\Sigma}}) \,\|\, \mathcal{N}(\mathbf{0}, \mathbf{I})\big)\Big], \quad (39)$$

where $\hat{\mathbf{x}}_s^r$ is obtained via deterministic DDIM sampling from reparameterized noise $\hat{\mathbf{x}}_T^r$. The full gradient $w.r.t.\ \xi$ follows the chain rule:

$$\frac{\partial \boldsymbol{\Omega}(\xi)}{\partial \xi} \propto \frac{\partial \boldsymbol{\Omega}(\xi)}{\partial \hat{\mathbf{x}}_s^r} \frac{\partial \hat{\mathbf{x}}_s^r}{\partial \hat{\mathbf{x}}_T^r} \frac{\partial \hat{\mathbf{x}}_T^r}{\partial \xi} - \beta\, \frac{\partial \ell_{\mathrm{KL}}}{\partial \xi}, \quad (40)$$

where the computational bottleneck arises from the Jacobian $\partial \hat{\mathbf{x}}_s^r / \partial \hat{\mathbf{x}}_T^r$, which accumulates across all timesteps $t : T \to 0$ in the sampling trajectory. For clarity, we drop superscript r and denote the DDIM trajectory as $\mathbf{x}_T \to \mathbf{x}_{T-1} \to \cdots \to \mathbf{x}_0 = \mathbf{x}_s$. Under the variance-preserving DDIM sampler [30], the update at each step is:

$$\mathbf{x}_{t-1} = a_t\, \mathbf{x}_t + b_t\, s_\phi(\mathbf{x}_t, t \,|\, \hat{\mathbf{y}}_s, \cdot), \quad (41)$$

with scalars a_t, b_t determined by the noise schedule. Differentiating Eq. (41) yields:

$$\frac{\partial \mathbf{x}_{t-1}}{\partial \mathbf{x}_t} = a_t\, \mathbf{I} + b_t\, \frac{\partial s_\phi(\mathbf{x}_t, t \,|\, \hat{\mathbf{y}}_s, \cdot)}{\partial \mathbf{x}_t}. \quad (42)$$

We adopt the *temporal stop-gradient heuristic* [22], treating the score function as fixed during backward pass:

$$\forall t, \quad \frac{\partial s_\phi(\mathbf{x}_t, t \,|\, \hat{\mathbf{y}}_s, \cdot)}{\partial \mathbf{x}_t} \equiv \mathbf{0}, \quad (43)$$

so the Jacobian simplifies to:

$$\frac{\partial \mathbf{x}_{t-1}}{\partial \mathbf{x}_t} \approx a_t\, \mathbf{I}. \quad (44)$$

The full Jacobian across all steps collapses to a scalar multiple:

$$\frac{\partial \mathbf{x}_s}{\partial \mathbf{x}_T} = \prod_{t=1}^{T} \frac{\partial \mathbf{x}_{t-1}}{\partial \mathbf{x}_t} \approx \left(\prod_{t=1}^{T} a_t\right) \mathbf{I} =: c_T\, \mathbf{I}. \quad (45)$$

where for VP-DDIM, $a_t = \sqrt{\bar{\alpha}_{t-1}/\bar{\alpha}_t}$, giving:

$$c_T = \prod_{t=1}^{T} \sqrt{\frac{\bar{\alpha}_{t-1}}{\bar{\alpha}_t}} = \sqrt{\frac{\bar{\alpha}_0}{\bar{\alpha}_T}} = \sqrt{\frac{1}{\bar{\alpha}_T}} \equiv \sqrt{\frac{1}{\alpha_T}}. \quad (46)$$

Thus, the gradient of the final sample $w.r.t.$ initial noise becomes:

$$\frac{\partial \hat{\mathbf{x}}_s^r}{\partial \hat{\mathbf{x}}_T^r} \approx \sqrt{\frac{1}{\alpha_T}}\, \mathbf{I}. \quad (47)$$

Substituting Eq. (47) into the chain rule in Eq. (40), and using $\partial \mathbf{\Omega}(\xi)/\partial \hat{\mathbf{x}}_s^r = \nabla_{\hat{\mathbf{x}}_s^r} \ell_{\text{adv}}(f_\theta(\hat{\mathbf{x}}_s^r), \hat{\mathbf{y}}_s)$, we arrive at the final approximation:

$$\begin{aligned}\frac{\partial \mathbf{\Omega}(\xi)}{\partial \xi} &\approx \mathbb{E}_{\hat{\mathbf{x}}_T, \hat{\mathbf{y}}_s}\left[\nabla_{\hat{\mathbf{x}}_s^r} \ell_{\text{adv}}^\top \frac{\partial \hat{\mathbf{x}}_s^r}{\partial \hat{\mathbf{x}}_T^r} \frac{\partial \hat{\mathbf{x}}_T^r}{\partial \xi}\right] - \beta\, \mathbb{E}_{\hat{\mathbf{x}}_T, \hat{\mathbf{y}}_s}\left[\frac{\partial \ell_{\text{KL}}}{\partial \xi}\right] \\ &\approx \sqrt{\frac{1}{\alpha_T}}\, \mathbb{E}_{\hat{\mathbf{x}}_T, \hat{\mathbf{y}}_s}\left[\nabla_{\hat{\mathbf{x}}_s^r} \ell_{\text{adv}}^\top \frac{\partial \hat{\mathbf{x}}_T^r}{\partial \xi}\right] - \beta\, \mathbb{E}_{\hat{\mathbf{x}}_T, \hat{\mathbf{y}}_s}\left[\frac{\partial \ell_{\text{KL}}}{\partial \xi}\right].\end{aligned} \tag{48}$$

In summary, under a DDIM sampler and the stop-gradient assumption, the dependence of $\hat{\mathbf{x}}_s^r$ on the noise $\hat{\mathbf{x}}_T^r$ collapses to a scalar factor $\sqrt{1/\alpha_T}$, effectively avoiding gradient explosion and enabling efficient, memory-light optimization.

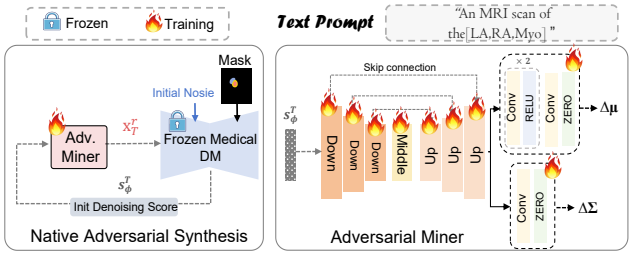

Figure 11. **Architecture of the Adversariality Miner**, where "Down" and "Up" refer to ResNet blocks with down- and up-sampling, respectively. "Middle" denotes the self-attention layer. "ZERO" refers to the zero-initialized convolution. The prompt encoder and timestep embedding are omitted for clarity.

Algorithm 2 Training Procedure for $\mathcal{M}_\xi$

Require: Score function s_ϕ, downstream model f_ϑ, adversariality miner $\mathcal{M}_{\xi^{(0)}}$, mask source q_ω, DDIM sampler, upper bound $\kappa_{\text{up}} = 0.5$, learning rate $lr = 1 \times 10^{-4}$, total iterations $N_{\text{iter}} = 3\text{K}$, truncated DDIM stride $\Delta_t = 10$.

for $i = 0, \ldots, N_{\text{iter}}$ **do**
  Sample $\hat{\mathbf{y}}_s \sim q_\omega$, $\hat{\mathbf{x}}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  // Compute initial score prediction
  $\mathcal{S}_\phi^{\text{Init}} \leftarrow \text{sg}\left(s_\phi\left(\hat{\mathbf{x}}_T, T | \hat{\mathbf{y}}_s\right)\right)$
  // Compute shifted distribution statistics
  $(\Delta_\mu, \Delta_{\mathbf{\Sigma}}) \leftarrow \mathcal{M}_{\xi^{(i)}}\left(\mathcal{S}_\phi^{\text{Init}}\right)$
  // Reselecting noise
  $\hat{\mathbf{x}}_T^r \leftarrow \Delta_\mu + (\mathbf{I} + \Delta_{\mathbf{\Sigma}})^{\frac{1}{2}} \odot \epsilon',\ \epsilon' \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  // Perform truncated DDIM sampling
  $\hat{\mathbf{x}}_s^r \leftarrow \hat{\mathbf{x}}_T^r$
  for $t = 1$ **to** Δ_t **do**
    $\hat{\mathbf{x}}_s^r \leftarrow \text{DDIM-Step}_{s_\phi}\left(\hat{\mathbf{x}}_s^r; \hat{\mathbf{y}}_s, t\right)$ **// No Grad**
  end for
  // Compute adversarial preference
  $\ell_{\text{adv}} \leftarrow \text{clamp}\left(f_\vartheta(\hat{\mathbf{x}}_s^r, \hat{\mathbf{y}}_s), \max = \kappa_{\text{up}}\right)$
  // Compute regularization term
  $\ell_{\text{KL}} \leftarrow \ell_{\text{KL}}\left(\mathcal{N}^r(\Delta_\mu, \mathbf{I} + \Delta_{\mathbf{\Sigma}}) \| \mathcal{N}(\mathbf{0}, \mathbf{I})\right)$
  // Compute the optimization objective
  $\mathbf{\Omega}^{(i)} \leftarrow \ell_{\text{adv}} - \beta \cdot \ell_{\text{KL}}$
  // Backpropagation step for optimization
  $\xi^{(i+1)} \leftarrow \xi^{(i)} + lr \cdot \frac{\partial \mathbf{\Omega}^{(i)}}{\partial \xi^{(i)}}$
end for
return $\mathcal{M}_{\xi^{(N_{\text{iter}})}}$

2. Details of the Adversariality Miner

2.1. Architecture Design

As shown in Fig. 11, the adversariality miner $\mathcal{M}_\xi$ follows the Res-Unet [26] architecture, comprising three successive ($2\times$) down-sampling and up-sampling ResNet blocks with skip connections for feature aggregation. The "Middle" layer integrates self-attention to enhance the model's representational capacity. The output is decoupled into two heads, estimating the statistics $\Delta_{\mathbf{\Sigma}}$ and Δ_μ, respectively. The final layer includes a "ZERO" convolution initialized with zeros, ensuring that $(\Delta_{\boldsymbol{\mu}}, \Delta_{\mathbf{\Sigma}}) \approx \mathbf{0}$ initially, minimizing early-stage corrections to stabilize optimization. Only $\mathcal{M}_\xi$ is *updated*, while the DM and f_θ remain *frozen*.

2.2. Training Procedure

The complete training procedure is summarized in Alg. 2. For clarity, we omit auxiliary components such as the VAE decoder, text/prompt encoder, timestep embedding, and classifier-free guidance (CFG). During DDIM sampling, no gradients are accumulated (*cf.* Eq. (47)); only $\mathcal{M}_\xi$ is updated, while all other modules remain *frozen*.

Thanks to the zero-initialized output heads and the regularizing effect of ℓ_{KL}, the reselected noise remains close to the denoising prior throughout training (as shown in Fig. 13, where $(\Delta_{\boldsymbol{\mu}}, \Delta_{\mathbf{\Sigma}}) \approx \mathbf{0}$), enabling the generative process to consistently produce high-quality samples across all training stages. Interestingly, we observe that $\mathcal{M}_\xi$ does not gradually improve but instead abruptly acquires the ability to mine highly adversarial samples, typically within fewer than 1.5K optimization steps (see Fig. 12); we refer to this behavior as the *sudden convergence phenomenon*.

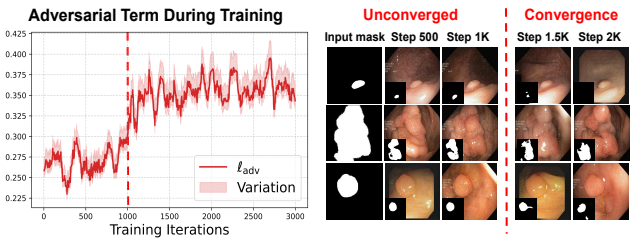

Figure 12. **The sudden convergence phenomenon**. Due to zero initialization and ℓ_{KL}, the training process consistently generates high-quality images. At a certain step (*e.g.*, 1.5K steps, marked by the dashed line), $\mathcal{M}_\xi$ abruptly gains the ability to mine high-adversarial samples.

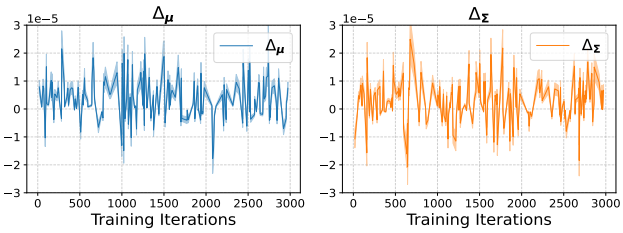

Figure 13. **Training dynamics** of the mean values of Δ_μ and Δ_Σ, which both remain close to zero throughout training.

Dataset	Modality	Classes	Train (70%)	Val (10%)	Test (20%)
ACDC [2]	MRI	4	4010	572	1146
Synapse [15]	CT	10	2645	378	756
Polyps [1, 13]	RGB	2	1128	161	323
EndoScene† [33]	RGB	2	—	—	60
ColonDB† [31]	RGB	2	—	—	380
ETIS† [29]	RGB	2	—	—	196
MMWHS* [42]	CT	8	3714	531	1060
MMWHS* [42]	MRI	8	2029	290	579

Table 7. **Dataset statistics.** Number of classes and split sizes. †Used for evaluation only. * Multi-modal dataset.

3. More Implementation Details

3.1. Datasets

We evaluate our method on seven public medical segmentation benchmarks spanning *MRI*, *CT*, and *RGB endoscopy*. All datasets are first split at the patient level into training, validation, and test sets with a ratio of 7:1:2, after which 3-D volumes are converted into axial 2-D slices and all images are resized to 256×256 (see Tab. 7 for exact counts).

- **ACDC** [2] (*cine-MRI, 3-D*). We follow the official split and segment **three cardiac structures**: left-ventricular cavity (`LV`), right-ventricular cavity (`RV`), and myocardium (`Myo`). Short-axis volumes are converted to 2-D slices for training and evaluation.
- **Synapse Multi-Organ** [15] (*contrast-enhanced CT, 3-D*). The dataset contains 30 abdominal CT scans with **nine organ** annotations: spleen, right kidney, left kidney, gallbladder, esophagus, liver, stomach, aorta, and pancreas. We retain axial slices containing at least one labeled organ, resulting in 3,779 2-D slices.
- **Polyps** [1, 13] (*RGB colonoscopy, 2-D*). We combine CVC-ClinicDB (612 frames) and Kvasir-SEG (1k frames) into a single colonic polyp segmentation benchmark. Each image is treated as a **binary** mask: foreground polyp (`POL`) vs. background (`BG`). Following [25], we train on the mixed training split and report (i) in-domain performance on their held-out test sets and (ii) cross-domain generalization on three unseen datasets: EndoScene [33] (60 samples), CVC-ColonDB [31] (380 samples), and ETIS [29] (196 samples).
- **MMWHS** [42] (*MRI & CT, 3-D*). Multi-Modality Whole Heart dataset with **seven** foreground labels: Left Ventricle (`LV`), Left Atrium (`LA`), Right Ventricle (`RV`), Right Atrium (`RA`), Myocardium (`Myo`), Ascending Aorta (`AA`), and Pulmonary Artery (`PA`). We use 5,305 CT slices and 2,898 MRI slices, training on one modality and evaluating on the held-out test split of the other modality to assess cross-modality generalization.

3.2. Generative Pipeline

We instantiate four M2I medical diffusion backbones: FairDiff [16], SiameseDiff [25], and DiffBoost [40], all built on LDM [27], and SegDiff [14] based on a vanilla DPM [10]. Since their original data splits are either undisclosed or incompatible with our setup, we re-train all models from their official implementations on our unified splits to avoid leakage and broadly evaluation.

Training. For SD-based backbones, we initialize from publicly released SD v1.5 checkpoints pretrained on large-scale text-to-image corpora [27, 39] and adopt the default optimization settings from the respective papers, unless otherwise stated. When textual conditioning is supported, we use a fixed prompt template "`a [Modality] imaging of [CLS]`". We compute FID [8] on the validation set every 500 iterations and select the checkpoint with the lowest FID.

Inference. We replace each method's original sampler with a deterministic DDIM sampler ($\eta = 0$) with 50 steps. The classifier-free guidance (CFG) scale [9] is fixed to 7.5 for SD-based models and 1.2 for SegDiff. Conditional masks $\hat{y}_s \sim q_\omega$, obtained by applying random horizontal flips and isotropic scaling to the ground-truth segmentations.

3.3. Downstream Segmentation Models

We employ nnU-Net [12] and SwinUNet [4] as downstream segmentation backbones. For each dataset, the backbone is first trained on the real training split, and the resulting checkpoint is treated as the *anchor* model, following the default training configurations in the original works. For synthetic-data training, we initialize from the *anchor* and adopt a dual-stream sampling strategy, with each mini-batch comprising an equal ratio (1:1) of real and synthetic samples. Within each batch, real–synthetic pairs are randomly formed, and CutMix [37] is applied to each pair with a probability of 0.5. All other training settings remained consistent with the initial pretraining stage.

3.4. Evaluation Protocol

We compare our method against a set of representative diffusion-based augmentation strategies:

- **Base**: A naive baseline using DDIM sampling with default experimental settings and no additional design.

- **UGDM** [18]: We omit the first stage of DDIM inversion, as it degrades generation quality, and directly apply uncertainty sampling with margin-based guidance ($\gamma = 3$) to generate $\hat{\mathrm{x}}_s$.
- **AdvDiffuser** [5]: A canonical adversarial guidance (AG) method combined with PGD [19]. We use $T = 50$ denoising steps, a step size of $\eta = 0.1$, and $I = 1$ adversarial update per sample.
- **Diff-PGD** [35]: Applies adversarial editing via SDEdit [21] and PGD. A Base synthetic image $\hat{\mathrm{x}}_s$ is transformed into an adversarial variant $\hat{\mathrm{x}}_s^{\text{adv}}$ using default parameters.
- **P2P** [20]: An AG variant similar to Textual Inversion [6], which optimizes text embeddings to induce adversarial perturbations. We use $\epsilon = 0.05$, AdamW optimizer (lr = $1\mathrm{e}{-5}$), and 500 training iterations.
- **AugPaint** [11]: An inpainting method where foreground regions in $\hat{\mathrm{x}}_s$ are masked and reconstructed via task-aware sampling to yield the augmented image $\hat{\mathrm{x}}_s^*$.
- **DiffAug** [28]: Generates structurally perturbed samples by partially denoising a Base synthetic image $\hat{\mathrm{x}}_s$ from a random intermediate timestep $t \sim \mathrm{Beta}(2, 4) \cdot T$.
- **CIG** [34]: We use only the Circle Interpolation strategy in "interp" mode. For input noise $\hat{\mathrm{x}}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, we randomly sample a real image $\mathrm{x}^{\text{ref}} \sim \mathcal{U}_{\text{real}}$, obtain its noisy latent $\tilde{\mathrm{x}}_T$ by forward noising to timestep T (omitting the original DDIM inversion), and perform circular interpolation between $\hat{\mathrm{x}}_T$ and $\tilde{\mathrm{x}}_T$ to obtain $\hat{\mathrm{x}}'_T$, which is then fed into the DM for generation.
- **Da-fusion** [32]: For each $\mathrm{x}^{\text{ref}} \sim \mathcal{U}_{\text{real}}$, we first generate an augmented variant $\hat{\mathrm{x}}_s^*$ using SDEdit [21] with a random starting noise level $t_0 \in \{1/4, 1/2, 3/4, 1\}T$, then apply Mixup [38] between x^{ref} and $\hat{\mathrm{x}}_s^*$ with $\alpha = 0.5$.
- **GAL** [41]: Offline filtering is applied with a validation score threshold $\tau = -0.05$. Sampling is repeated until the synthetic dataset reaches the target budget N_s.

Evaluation Protocol. All methods are evaluated under a unified setup: **(1)** a shared diffusion backbone is used across all methods; **(2)** each method generates a synthetic dataset matching the size of the real training set, using random seed 42; and **(3)** downstream models are retrained from the same anchor checkpoint with identical optimization settings. Unless otherwise specified, all hyperparameters follow official defaults. In cases where the default configurations yield degenerate results (e.g., excessive noise or failure), minimal tuning is performed to ensure valid sample generation.

4. Additional Experiment Results

4.1. Cross-model Validity

We assess whether adversarially enhanced synthetic datasets, generated using a *source* model f_A, can transfer generalization benefits to a distinct *target* model f_B. Concretely, we instantiate f_A as nnU-Net [12] and f_B as SwinUNet [4]. The default setting $f_A = f_B$ corresponds to within-model training, while $f_A \neq f_B$ probes cross-model generalization. As shown in Tab. 8, performance gains persist across architectures, albeit with mild attenuation, indicating that our method surfaces adversarial signals reflective of shared inductive blind spots, rather than model-specific failure modes. This suggests that native adversariality captures transferable hard cases, reinforcing its potential as a model-agnostic augmentation strategy.

Models		ACDC		Synapse		Polyps	
Source	Target	DSC↑	ASD↑	DSC↑	ASD↑	DSC↑	ASD↑
f_A†	–	86.51	2.57	73.95	20.10	78.83	6.62
f_B†	–	87.36	1.96	75.45	14.39	81.38	5.76
f_A	f_A	6.93±0.8	1.03±0.2	9.12±0.6	7.99±1.3	10.25±1.3	3.66±0.5
f_B	f_A	5.17±0.5	0.82±0.1	8.44±0.5	6.91±0.3	8.98±1.1	2.43±0.3
f_B	f_B	5.40±0.7	0.99±0.2	8.75±0.4	5.81±0.8	9.09±1.4	3.02±0.2
f_A	f_B	5.33±0.4	0.60±0.3	7.04±0.3	4.72±0.5	8.15±1.9	2.10±0.1

Table 8. **Cross-Model Validity**. The synthesis pipeline and experimental setup follow Tab. 1. f_A : nnU-Net [12] ; f_B : SwinUNet [4]; †: The Baseline is trained solely on $\mathcal{U}_{\text{train}}$ and reports absolute performance; all other entries report Δ over this baseline.

ℓ_{seg}	Synapse (DiffBoost)				Polyps (SiameseDiff)			
	nnU-Net [12]		SwinUNet [4]		nnU-Net [12]		SwinUNet [4]	
	Δ_{DSC}↑	Δ_{ASD}↑	Δ_{DSC}↑	Δ_{ASD}↑	Δ_{DSC}↑	Δ_{ASD}↑	Δ_{DSC}↑	Δ_{ASD}↑
—	4.09	2.97	3.17	2.19	5.05	2.12	4.38	1.18
ℓ_{CE}	9.12	**7.99**	8.75	**5.81**	**10.25**	3.66	**9.09**	**3.20**
ℓ_{Dice}	**10.53**	6.12	9.33	4.30	9.95	2.49	8.36	2.98
$\ell_{\text{CE}} + \ell_{\text{Dice}}$	9.77	7.26	**9.64**	4.22	10.01	**3.68**	8.90	3.58
ℓ_{Focal}	8.53	6.40	8.92	5.37	9.33	2.75	8.16	3.09

Table 9. **Comparison of loss variants** ℓ_{seg} on Synapse and Polyps. All results are reported as performance gains over the Baseline from Tab. 1. "—" denotes w/o adversarial enhancement.

4.2. Loss Variants

We perform an ablation study on several empirical segmentation loss variants used as adversarial metrics ℓ_{seg}. The losses include pixel-wise Cross-Entropy ℓ_{CE} (default), Dice loss [23] ℓ_{Dice}, a compound loss $\ell_{\text{CE}} + \ell_{\text{Dice}}$ commonly adopted in medical segmentation, and Focal loss [17] ℓ_{Focal}. As reported in Tab. 9, all variants yield comparable performance, suggesting that our method is robust to the choice of supervision signal. For simplicity and compatibility across datasets, we adopt ℓ_{CE} as the default loss in all experiments.

4.3. Failure Cases

Fig. 15 highlights representative failure cases observed in adversarially synthesized samples across both anatomical

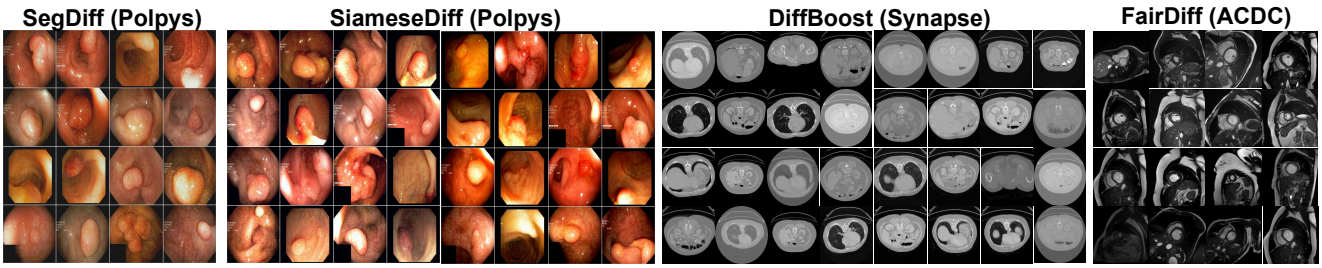

Figure 14. **Visualization of synthesized images.** Our method achieves enhanced adversarial effectiveness while preserving high fidelity.

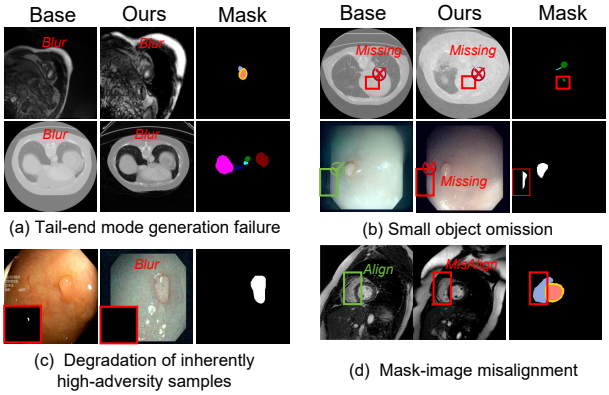

Figure 15. Examples of **failure cases** in synthetic datasets generated by the Base sampling and by our method.

and endoscopic domains. We categorize these into four distinct patterns:

- **Tail-end mode generation failure**: In rare scenarios involving uncommon anatomical configurations or extreme variations, both the base model and adversarial synthesis may hallucinate. These failures stem from insufficient representation of such outliers in the training data, limiting the model's capacity to form reliable generative priors.
- **Small object omission**: The method occasionally fails to preserve fine-grained structures such as nodules or vessel-like features. This occurs because small-scale elements contribute minimally to the adversarial reward (ℓ_{KL}), making them less likely to be preserved during sample optimization.
- **Degradation of inherently high-adversity samples**: In cases where the initial input noise $\hat{\mathbf{x}}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ already leads to high adversariality, our method may introduce artifacts or noise in its attempts to further optimize adversarial reward, potentially degrading the quality of the generated samples.
- **Mask-image misalignment**: Occasional spatial mismatches between synthesized images and their corresponding conditional masks suggest that $\mathcal{M}_\xi$ may exploit trivial adversarial shortcuts. While infrequent, such cases indicate the model's tendency to prioritize reward over spatial consistency.

We note that these failure cases do not exclusively arise from our method; the Base synthetic samples often exhibit comparable or even more severe artifacts. This suggests that many of these issues are rooted in the underlying diffusion backbone. Nevertheless, since our approach operates by amplifying adversarial preference signals, it may inadvertently intensify the generator's inherent limitations.

Mitigation strategies include enhancing the training distribution, refining conditional prompts, enforcing structural consistency [7], or applying filtering techniques such as rejection sampling [36].

4.4. Synthesis Visualizations

Fig. 14 presents qualitative results of our method across different DMs. The synthesized samples exhibit high visual fidelity and structural diversity across all datasets, demonstrating that native adversariality can be effectively amplified without compromising generative quality.

References

[1] Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized medical imaging and graphics*, 43: 99–111, 2015. 5

[2] Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al. 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, 2018. 5

[3] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. 2

[4] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*, pages 205–218. Springer, 2022. 5, 6

[5] Xinquan Chen, Xitong Gao, Juanjuan Zhao, Kejiang Ye, and Cheng-Zhong Xu. Advdiffuser: Natural adversarial example synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4562–4572, 2023. 6

[6] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *The Eleventh International Conference on Learning Representations*. 6

[7] Xiefan Guo, Jinlin Liu, Miaomiao Cui, Jiankai Li, Hongyu Yang, and Di Huang. Initno: Boosting text-to-image diffusion models via initial noise optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9380–9389, 2024. 7

[8] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 5

[9] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*. 5

[10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 5

[11] Xinrong Hu and Yiyu Shi. Inpainting is all you need: A diffusion-based augmentation method for semi-supervised medical image segmentation. *arXiv preprint arXiv:2506.23038*, 2025. 6

[12] 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. 5, 6

[13] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas De Lange, Dag Johansen, and Håvard D Johansen. Kvasir-seg: A segmented polyp dataset. In *International conference on multimedia modeling*, pages 451–462. Springer, 2019. 5

[14] Nicholas Konz, Yuwen Chen, Haoyu Dong, and Maciej A Mazurowski. Anatomically-controllable medical image generation with segmentation-guided diffusion models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 88–98. Springer, 2024. 5

[15] Bennett A. Landman and Simon K. Warfield. Multi-atlas labeling beyond the cranial vault - workshop and challenge. https://www.synapse.org/#!Synapse:syn3193805/wiki/217789, 2015. MICCAI 2015 Workshop: Multi-Atlas Labeling Beyond the Cranial Vault. 5

[16] Wenyi Li, Haoran Xu, Guiyu Zhang, Huan-ang Gao, Mingju Gao, Mengyu Wang, and Hao Zhao. Fairdiff: Fair segmentation with point-image diffusion. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 617–628. Springer, 2024. 5

[17] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 6

[18] Yimin Luo, Qinyu Yang, Yuheng Fan, Haikun Qi, and Menghan Xia. Measurement guidance in diffusion models: Insight from medical image synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 6

[19] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 6

[20] Yasamin Medghalchi, Moein Heidari, Clayton Allard, Leonid Sigal, and Ilker Hacihaliloglu. Prompt2perturb (p2p): Text-guided diffusion-based adversarial attack on breast ultrasound images. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 28564–28574, 2025. 6

[21] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2022. 6

[22] Boming Miao, Chunxiao Li, Xiaoxiao Wang, Andi Zhang, Rui Sun, Zizhe Wang, and Yao Zhu. Noise diffusion for enhancing semantic faithfulness in text-to-image synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23575–23584, 2025. 3

[23] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *Proceedings of the Fourth International Conference on 3D Vision (3DV)*, pages 565–571. IEEE, 2016. 6

[24] Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer Science & Business Media, 2003. 2

[25] Kunpeng Qiu, Zhiqiang Gao, Zhiying Zhou, Mingjie Sun, and Yongxin Guo. Noise-consistent siamese-diffusion for medical image synthesis and segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15672–15681, 2025. 5

[26] Hameedur Rahman, Tanvir Fatima Naik Bukht, Azhar Imran, Junaid Tariq, Shanshan Tu, and Abdulkareeem Alzahrani. A deep learning approach for liver and tumor segmentation in ct images using resunet. *Bioengineering*, 9(8):368, 2022. 4

[27] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 5

[28] Chandramouli Shama Sastry, Sri Harsha Dumpala, and Sageev Oore. Diffaug: A diffuse-and-denoise augmentation for training robust classifiers. *Advances in Neural Information Processing Systems*, 37:20745–20785, 2024. 6

[29] Juan Silva, Aymeric Histace, Olivier Romain, Xavier Dray, and Bertrand Granado. Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. *International journal of computer assisted radiology and surgery*, 9(2):283–293, 2014. 5

[30] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*. 3

[31] Nima Tajbakhsh, Suryakanth R Gurudu, and Jianming Liang. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE transactions on medical imaging*, 35(2):630–644, 2015. 5

[32] Brandon Trabucco, Kyle Doherty, Max A Gurinas, and Ruslan Salakhutdinov. Effective data augmentation with diffusion models. In *The Twelfth International Conference on Learning Representations*. 6

[33] David Vázquez, Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Antonio M López, Adriana Romero, Michal Drozdzal, and Aaron Courville. A benchmark for endoluminal scene segmentation of colonoscopy images. *Journal of healthcare engineering*, 2017(1):4037190, 2017. 5

[34] Yanghao Wang and Long Chen. Inversion circle interpolation: Diffusion-based image augmentation for data-scarce classification. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25560–25569, 2025. 6

[35] Haotian Xue, Alexandre Araujo, Bin Hu, and Yongxin Chen. Diffusion-based adversarial sample generation for improved stealthiness and controllability. *Advances in Neural Information Processing Systems*, 36:2894–2921, 2023. 6

[36] Chen Yu and Shuyang Gao. Improving compositional generation with diffusion models using lift scores. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*. PMLR, 2025. 7

[37] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*, pages 6023–6032, 2019. 5

[38] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 6

[39] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023. 5

[40] Zheyuan Zhang, Lanhong Yao, Bin Wang, Debesh Jha, Gorkem Durak, Elif Keles, Alpay Medetalibeyoglu, and Ulas Bagci. Diffboost: Enhancing medical image segmentation via text-guided diffusion model. *IEEE Transactions on Medical Imaging*, 2024. 5

[41] Muzhi Zhu, Chengxiang Fan, Hao Chen, Yang Liu, Weian Mao, Xiaogang Xu, and Chunhua Shen. Generative active learning for long-tailed instance segmentation. In *International Conference on Machine Learning*, pages 62349–62368. PMLR, 2024. 6

[42] Xiahai Zhuang and Juan Shen. Multi-scale patch and multi-modality atlases for whole heart segmentation of mri. *Medical Image Analysis (MIA)*, 31:77–87, 2016. 5