This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Masked-Diffusion Autoencoders for 3D Medical Vision Representation Learning

Jiachen Tu[*†1] Guanghui Qin[*2] Theodore Zhengde Zhao[2] Jeya Maria Jose Valanarasu[2]
Sheng Zhang[2] Tristan Naumann[2] Fan Lam[1] Sheng Wang[3] Hoifung Poon[2]
[1]University of Illinois Urbana-Champaign [2]Microsoft [3]University of Washington

Abstract

Effective medical image analysis requires representations that capture both global anatomical structure and fine-grained tissue texture. Current self-supervised approaches exhibit limited capacity to address both requirements simultaneously. Invariance-based methods learn through augmentation consistency but face challenges in medical imaging where common augmentations may discard diagnostically relevant intensity patterns. Masked image modeling approaches employ high masking ratios to enforce holistic reasoning, yet inherently limit exposure to fine-grained texture. Recent work in general-domain vision demonstrates that generative and semantic objectives can mutually benefit each other, yet this paradigm remains unexplored for 3D medical imaging. We introduce Masked-Diffusion Autoencoders (MDAE), a self-supervised framework that imposes concurrent spatial masking and diffusion corruption, encouraging the model to learn complementary objectives: masked region reconstruction for structural coherence and visible region denoising for textural characteristics. This dual corruption enables the network to learn structure-texture representations within a unified time-conditioned objective. Evaluated on brain MRI across tumor classification, molecular marker detection, and dense segmentation benchmarks, MDAE consistently outperforms state-of-the-art baselines, with improvements most pronounced in cross-modal generalization tasks.

1. Introduction

Medical image annotation requires specialized domain expertise and is considerably more resource-intensive than natural image labeling [27, 37, 49, 75]. Self-supervised learning (SSL) has emerged as a promising solution and achieved remarkable success in natural vision [6, 7, 9, 18, 23, 24, 44, 74]. However, medical imaging presents unique challenges as diagnostic information is often distributed continuously across spatial scales—from organ-level geometry to voxel-level texture, unlike natural images that contain discrete objects against backgrounds [16, 17, 36, 56, 70]. Effective representation learning shall therefore encode holistic anatomical structure and fine-grained tissue characteristics simultaneously.

This dual requirement connects to a broader paradigm shift in visual representation learning. Historically, semantic encoders and generative models have operated along separate paths, stemming from assumptions that encoders capturing semantics are unsuited for faithful reconstruction [72]. Recent work challenges this dichotomy: Representation Autoencoders demonstrate that semantic encoders enable high-fidelity reconstruction while maintaining structured latent spaces [73, 82]. Furthermore, evidence demonstrates that generative pretraining can also enhance discriminative representations [12, 38, 64, 67], showing that the two objectives can mutually benefit each other. Yet, these advances remain confined to 2D natural images. How to effectively incorporate generative pretraining objectives to enhance 3D medical vision SSL representation remains an open question as medical imaging presents distinct challenges: data scarcity [4, 62], non-object-centric nature, partial volume effects that blur tissue boundaries [8, 41, 56], and the need for simultaneous structure-texture learning across continuous anatomical fields.

Current SSL approaches offer partial solutions. Invariance-based methods [6, 7, 9, 18, 23, 44, 74] learn representations by aligning augmented views, but may face difficulties when applied to non-object-centric medical data [11, 13, 32]. A fundamental limitation lies in augmentation design: common augmentations suitable for natural images often fail in medical imaging—color jittering disrupts diagnostic intensity relationships, Gaussian blurring may suppress pathological features, and heavy random cropping might risks discarding critical diagnostic regions. Masked autoencoders (MAE) and their medical adaptations [19, 20, 24, 62, 86] reconstruct masked regions using high masking ratios [65], favoring holistic structure at the cost of visible fine texture and relying on the assumption that objects form spatially continuous structures. However,

[*] Equal contribution.
[†] Work done during an internship at Microsoft.
Project page: https://jiachentu.github.io/MDAE/

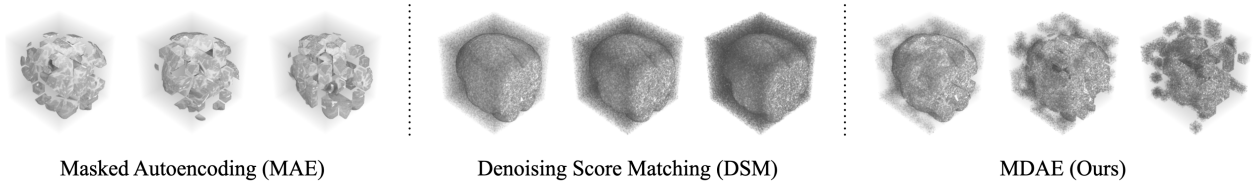


Figure 1. Comparison of corruption schemes. **Left:** MAE applies fixed masking corruption. **Middle:** DSM applies diffusion noise. **Right:** MDAE combines spatial masking with diffusion noising, creating dual objectives for holistic reconstruction and fine-grained denoising.

in medical imaging, crucial regions of interest often occupy small portions and differ markedly from neighboring anatomy. Diffusion and score-based models excel at capturing fine details through iterative denoising [26, 50], yet in medical imaging they have been applied primarily to synthesis [48, 71], reconstruction [35, 76, 77], and denoising [58, 59] rather than discriminative representation learning for 3D volumes. These limitations motivate the need for a discriminative framework that simultaneously addresses holistic structural reasoning and fine-grained texture preservation for 3D medical volumes.

We introduce Masked-Diffusion Autoencoders (MDAE), a unified discriminative SSL framework that adapts the generative-semantic paradigm to 3D medical imaging through global dual corruption. As illustrated in Figure 1, MDAE performs spatial masking and diffusion noising together, creating two complementary learning objectives: masked region reconstruction enforces holistic structural reasoning, while visible region denoising preserves fine-grained textural characteristics. The dual corruption approach enables effective learning at lower masking ratios—having both denoising and reconstruction objectives ensures that reconstruction remains challenging even when more context is visible, as the added noise prevents trivial pattern copying. This dual-objective design is unified through a time-conditioned framework that adapts reconstruction based on corruption severity, enabling the network to learn representations that capture both anatomical structure and tissue texture beneficial for medical diagnosis. We evaluate MDAE across 16 clinical benchmarks spanning tumor classification, molecular marker detection, and dense segmentation. MDAE achieves 73.6% average AUROC on in-distribution tasks and 78.6% on out-of-distribution tasks. Ablation studies reveal synergistic benefits when combining both corruptions, with variable masking further improving over fixed-ratio alternatives.

2. Related Work

Three-dimensional medical vision representation learning faces distinct challenges including data scarcity [4, 62], annotation complexity [37, 49, 75], and non-object-centric continuous anatomical fields [8, 41]. Drawing inspiration from general domain SSL, medical vision SSL approaches can be broadly categorized into invariance-based / contrastive methods and masked image modeling. Invariance-based / contrastive methods [6, 7, 9, 18, 23, 44, 74] have achieved strong performance in natural vision by learning representations through aligning augmented views, yet encounter challenges in medical imaging. Augmentation strategies effective for natural images can prove problematic for medical data—color transformations may alter diagnostically meaningful intensity patterns, blur operations risk obscuring subtle pathological features, and aggressive spatial cropping may remove small but critical anatomical structures [10, 13, 32]. Masked image modeling [1, 3, 24, 68, 85] has shown success through reconstruction-based learning, employing high masking ratios (75% for natural images [24], 60-95% across domains [19, 45, 57]) to prevent trivial solutions by enforcing holistic understanding, yet consequently reduces the model's exposure to fine-grained texture information. Medical imaging presents additional challenges, as diagnostically important regions may be spatially localized and exhibit appearance patterns distinct from surrounding tissue [79]. Medical-specific adaptations have emerged through rotation prediction [53], multi-scale restoration [20, 25, 83, 84, 86], position-aware learning [66], and large-scale foundation models [15, 30, 31, 42, 47, 52, 62, 69, 78, 80], yet the fundamental trade-off between holistic structure and textural detail remains unresolved.

Recent work motivates rethinking of the relationship between semantic and generative objectives. Representation Autoencoders (RAE) [82] show that semantic encoders enable high-fidelity reconstruction, REPA [73] aligns semantic representations with diffusion generation, and STELLAR [81] resolves the semantic-generative tension through spatial-semantic factorization of visual features. In parallel, growing evidence in 2D natural vision demonstrates that the two objectives can be mutually beneficial [12, 29, 38, 39, 55, 64, 67]. However, these benefits remain unexplored in 3D medical imaging, where few works have incorporated generative pretraining [21, 64] and existing approaches fo-

cus primarily on generation quality rather than discriminative learning.

3. Preliminaries

Data and notation. Let $\mathcal{D} = \{X^{(i)}\}_{i=1}^{N}$ be a dataset of N 3D medical imaging volumes, each $X^{(i)} \in \mathbb{R}^{D\times H\times W}$ with z-score normalization. For notational simplicity, we present the single-modality case; the framework naturally extends to multi-channel inputs $X \in \mathbb{R}^{C\times D\times H\times W}$ by applying corruption and reconstruction channel-wise. We use X_0 to denote a **clean ground-truth volume** (without noise or masking corruption), and (D, H, W) denote spatial dimensions. Let $M \in \{0,1\}^{D\times H\times W}$ denote a binary mask where $M_j = 1$ indicates masked voxels and $M_j = 0$ indicates visible voxels at spatial location $j = (d, h, w)$. We denote $M_v = (\mathbf{1} - M)$ as the visible mask. We define $\Omega_M = \{j \mid M_j = 1\}$ as the set of masked voxels with cardinality $|\Omega_M| = \sum_j M_j$, and $\Omega_V = \{j \mid M_j = 0\}$ as the set of visible voxels with cardinality $|\Omega_V| = \sum_j M_{v,j}$.

Masked auto-encoder (MAE). Given a patch-based mask M where each non-overlapping patch is independently masked with probability p_{mask}, the masked input $\tilde{X}^M = M_v \odot X_0$ zeros out masked regions. MAE [24] reconstructs these regions by minimizing:

$$\mathcal{L}_{\text{MAE}}(\theta) = \mathbb{E}_{X_0,M}\left[\frac{1}{|\Omega_M|}\cdot\left\|M\odot\left(g_\theta\left(\tilde{X}^M\right) - X_0\right)\right\|_2^2\right] \tag{1}$$

where the loss is computed only over masked voxels Ω_M. High masking ratios (typically 75%) are often used to prevent trivial reconstruction from neighboring patches.

Denoising score matching (DSM). Denoising Score Matching (DSM) [60] is a framework for learning the score function (i.e., the gradient of the log-density) of the data distribution by training a model to reverse the corruption process of adding Gaussian noise. In DSM, given a clean volume $X_0 \in \mathbb{R}^{D\times H\times W}$ and a noise level $\sigma_t > 0$, a noisy observation $\tilde{X}_t$ is generated by:

$$\tilde{X}_t = X_0 + \sigma_t Z, \quad Z \sim \mathcal{N}(0, I) \tag{2}$$

where I denotes the identity matrix of appropriate dimension. The objective is to train a denoising function $g_\theta : \mathbb{R}^{D\times H\times W} \times \mathbb{R} \to \mathbb{R}^{D\times H\times W}$, parameterized by θ, that estimates the clean volume X_0 from its noisy counterpart $\tilde{X}_t$. This is achieved by minimizing the expected mean squared error (MSE) loss:

$$\mathcal{L}_{\text{DSM}}(\theta) = \mathbb{E}_{X_0,t,Z}\left\|g_\theta\left(\tilde{X}_t, t\right) - X_0\right\|_2^2 \tag{3}$$

where $t \sim \mathcal{U}(0,1)$ parametrizes the noise level and the expectation is taken over the data distribution, noise levels, and random noise. Vincent [60] demonstrated that optimizing the denoising function g_θ using the MSE loss is equivalent to learning the score function of the noisy data distribution $p_t(\tilde{X}_t)$ up to a scaling factor. Specifically, using Tweedie's formula, $\nabla_{\tilde{X}_t}\log p_t(\tilde{X}_t) = \frac{1}{\sigma_t^2}\left(g_\theta(\tilde{X}_t, t) - \tilde{X}_t\right)$. By learning g_θ, we implicitly learn $\nabla_{\tilde{X}_t}\log p_t(\tilde{X}_t)$, which forms the foundation of score-based generative models [26, 50].

4. MDAE Framework

Figure 2 illustrates the MDAE framework. We apply global dual corruption: volumes are corrupted with diffusion noise before spatial masking, enabling simultaneous learning of structural and fine-grained features.

4.1. Dual corruption process

Given a clean volume $X_0 \in \mathbb{R}^{D\times H\times W}$, MDAE applies simultaneous dual corruptions combining *spatial masking* and *diffusion noise* (Figure 2).

Spatial masking We use a masking ratio p_{mask} to generate a binary mask $M \in \{0,1\}^{D\times H\times W}$ using *blocky patch-based masking*: the volume is partitioned into non-overlapping 16^3 voxel patches, each independently masked ($M_j = 1$ for all voxels in the patch) with probability p_{mask}, or kept visible ($M_j = 0$) otherwise. We follow MAE's patch-based masking strategy [24] for compatibility with transformer architectures; for our CNN-based implementation, masked regions are replaced with zeros. This blocky strategy ensures spatially contiguous masked regions, forcing the network to reason about volumetric structures rather than interpolating from immediately adjacent voxels.

Standard MAE [24] requires fixed high ratios (75%) because lower ratios enable trivial reconstruction from neighboring patches without diffusion noise. MDAE adopts a variable masking distribution $p_{\text{mask}} \sim \mathcal{U}(p_{\min}, p_{\max})$, [1] enabling the model to learn both low-level textural details (at lower ratios where more context is visible) and holistic anatomical structure (at higher ratios requiring global reasoning) within a unified pretraining objective. This is feasible because MDAE's dual corruption strategy—combining masking with diffusion noise on visible regions—maintains task difficulty even at low masking ratios.

Diffusion noise corruption Simultaneously, we sample a diffusion timestep $t \sim \mathcal{U}(0,1)$ and Gaussian noise $Z \sim \mathcal{N}(0, I)$ to create the noise-corrupted volume via variance-

[1] We empirically set $p_{\min} = 0.01$ and $p_{\max} = 0.99$, providing comprehensive coverage of the masking spectrum while avoiding extreme values that could destabilize training.

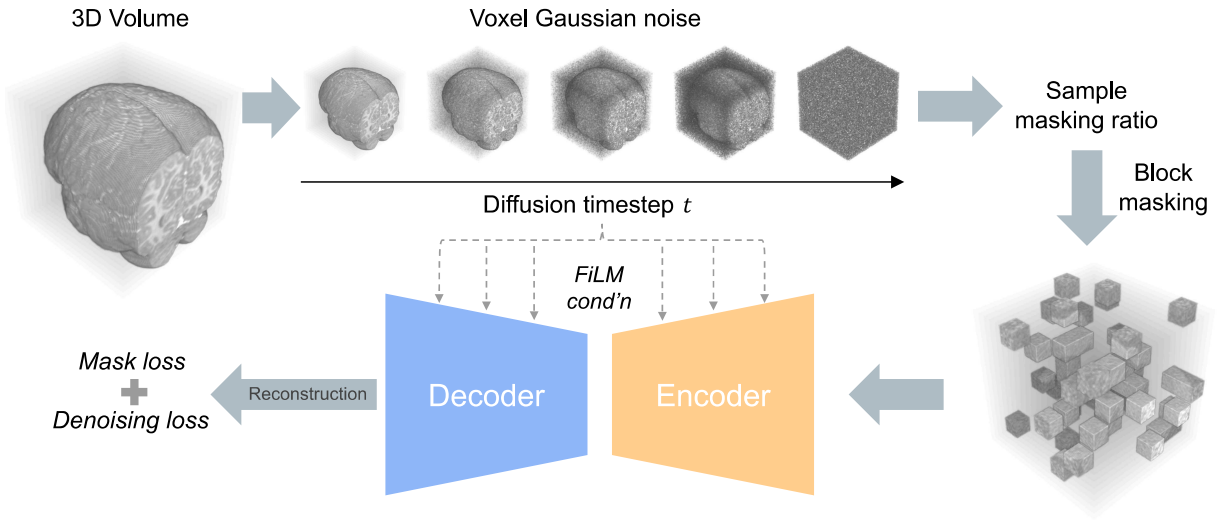


Figure 2. **MDAE Framework.** A 3D volume is jointly corrupted by patch-based masking (blocky 16^3 voxel patches) and diffusion noise (σ_t). The doubly-corrupted input is processed by a time-conditioned ResNet U-Net $g_\theta(\tilde{X}_t^M, t)$ with skip connections (orange: encoder, blue: decoder), where the diffusion timestep t modulates each block through FiLM conditioning. The network learns to predict clean anatomy for both visible and masked regions, supervised by two reconstruction targets: denoising on visible voxels (M_v, $\mathcal{L}_{\text{visible}}$ with $w(\sigma_t)$ weighting) and inpainting on masked voxels (M, $\mathcal{L}_{\text{masked}}$), both compared against the clean target X_0. Combined loss $\mathcal{L}_{\text{MDAE}} = \lambda_{\text{masked}} \cdot \mathcal{L}_{\text{masked}} + \lambda_{\text{visible}} \cdot \mathcal{L}_{\text{visible}}$ couples spatial structure and texture learning.

exploding (VE) corruption:

$$\tilde{X}_t = X_0 + \sigma_t Z \tag{4}$$

where $\sigma_t = \sigma(t)$ is the noise schedule with $\sigma(t) = t \cdot \sigma_{\max}$, and $\sigma_{\max}$ is a hyperparameter controlling maximum corruption severity. We use VE corruption for its notation simplicity with more analysis provided in Section 5.2.

Doubly-corrupted input The final input combines both corruptions: $\tilde{X}_t^M = M_v \odot \tilde{X}_t$, which zeros out masked regions (superscript M) while preserving noise-corrupted visible regions (subscript t denoting the time/noise level). This unified notation $\tilde{X}_t^M$ clarifies the simultaneous application of both masking and diffusion corruptions. This joint corruption strategy creates complementary learning signals. See Algorithm 1 (Appendix B) for detailed implementation.

4.2. Time-conditioned architecture

MDAE employs an encoder-decoder network $g_\theta : \mathbb{R}^{D\times H\times W} \times \mathbb{R} \to \mathbb{R}^{D\times H\times W}$ with parameters θ that predicts the clean volume X_0 from doubly-corrupted inputs (Figure 2). Unlike standard MAE which processes corrupted inputs as $\hat{X} = g_\theta(\tilde{X}^M)$, MDAE uses a **time-conditioned** architecture: $\hat{X} = g_\theta(\tilde{X}_t^M, t)$. The time parameter t informs the network about the noise level, enabling adaptive reconstruction strategies. Specifically, t is mapped to a 256-dimensional embedding via sinusoidal positional encoding followed by an MLP, then injected at each encoder and decoder stage through Feature-wise Linear Modulation (FiLM) [46]: $h_{\text{out}} = h_{\text{in}} \odot (\gamma(t_{\text{emb}}) + 1) + \beta(t_{\text{emb}})$, where γ and β are learned projections mapping the time embedding to scale and shift parameters. Time conditioning allows the network to dynamically balance between spatial inpainting and intensity denoising, learning representations robust across the full corruption spectrum.

4.3. Learning

The MDAE training objective combines two complementary components that operate on different spatial regions:

$$\mathcal{L}_{\text{MDAE}}(\theta) = \lambda_{\text{masked}} \cdot \mathcal{L}_{\text{masked}}(\theta) + \lambda_{\text{visible}} \cdot \mathcal{L}_{\text{visible}}(\theta) \tag{5}$$

where λ_{masked} and λ_{visible} are weighting hyperparameters. [2] The masked region loss reconstructs hidden content from corrupted visible context:

$$\mathcal{L}_{\text{masked}}(\theta) = \mathbb{E}\left[\frac{1}{|\Omega_M|} \cdot \left\| M \odot \left(g_\theta\left(\tilde{X}_t^M, t\right) - X_0\right)\right\|_2^2\right] \tag{6}$$

where the expectation is taken w.r.t. the patch-based mask M with masking probability p_{mask}, $t \sim \mathcal{U}(0,1)$, $Z \sim \mathcal{N}(0, I)$, and Ω_M denotes the set of masked voxels. This objective learns global anatomical structure by forcing the

[2] We empirically set $\lambda_{\text{masked}} = \lambda_{\text{visible}} = 1.0$. They could be adapted based on the masking ratio p_{mask} for task-specific optimization.

network to infer masked regions from corrupted visible context—analogous to the MAE objective but with additional noise robustness. The visible region loss denoises corrupted content back to the clean target:

$$\mathcal{L}_{\text{visible}}(\theta) = \mathbb{E}\left[\frac{w(\sigma_t)}{|\Omega_V|} \cdot \left\| M_v \odot \left(g_\theta\left(\tilde{X}_t^M, t\right) - X_0\right)\right\|_2^2\right] \tag{7}$$

where the expectation is taken w.r.t. the same distributions, Ω_V denotes the set of visible voxels, and $w(\sigma_t)$ is the noise-level weighting function. Following practices from prior works [33, 34, 50], this weighting ensures that loss contributions remain approximately constant across different noise levels, preventing optimization from being dominated by high-noise samples. This formulation connects to noise-conditional denoising score matching: optimizing the weighted MSE loss is equivalent to learning the score function via Tweedie's formula [60]. This objective learns fine-grained tissue texture by removing diffusion noise—analogous to denoising score matching but applied to non-masked regions within a masked autoencoding framework. Both objectives target the clean volume X_0 as the reconstruction target (not the noise-corrupted $\tilde{X}_t$), requiring the network to jointly learn spatial inpainting and intensity denoising. See Algorithm 1 (Appendix B) for training details.

Limiting behavior Eq. 5 interpolates MAE and DSM (up to per-voxel normalization and noise-level weighting):

$$\mathcal{L}_{\text{MDAE}} \to \begin{cases} \mathcal{L}_{\text{MAE}} & \text{as } \sigma_{\max} \to 0 \\ \mathcal{L}_{\text{DSM}} & \text{as } p_{\text{mask}} \to 0 \end{cases}$$

In the first limit, the visible loss vanishes as noise-corrupted visible regions become clean ($\tilde{X}_t \to X_0$, thus $\tilde{X}_t^M \to \tilde{X}^M$), leaving only masked reconstruction. In the second limit, $\mathbb{P}(|\Omega_M| > 0) \to 0$ so the expected masked loss vanishes as no voxels are masked ($|\Omega_M| \to 0$, thus $\tilde{X}_t^M \to \tilde{X}_t$), leaving only denoising (derivations provided in Appendix D).

5. Experiments and Results

Pretraining data. We pretrain MDAE on OpenMind [62], comprising 114,570 3D brain MRI volumes from 34,191 subjects. All volumes are preprocessed to 160^3 voxels via center cropping or zero-padding with stratified splits for reproducible evaluation.

Evaluation protocol. We evaluate across three scenarios: (1) *In-distribution performance* on T1/T2 modalities present during pretraining, covering 3 tasks—tumor type classification (BraTS23: glioma vs metastasis), MGMT methylation status, and IDH1 mutation detection—each evaluated on T1 and T2; (2) *Cross-modal generalization* to 6 out-of-distribution tasks using rare sequences in pretraining for molecular marker detection, survival prediction, demographic classification, and resection status classification; (3) *Multi-modality integration* for classification on BraTS18 (4 modalities) and UCSF-PDGM [5] (6 modalities), and dense prediction for tumor segmentation on UCSF-PDGM and BraTS18 benchmarks. For multi-modality tasks, each modality is processed independently through the pretrained encoder with mean-pooled embeddings. Segmentation uses the nnUNet framework [28] with encoders initialized from pretrained weights. Model finetuning details are provided in Algorithm 2, 3 in Appendix C.

Baseline methods. We compare against representative SSL methods spanning three paradigms: (1) *Invariance-based contrastive learning*: SimCLR [9] (general domain) and VoCo [66] (medical-specific volume contrast); (2) *Masked image modeling*: MAE [24] (high masking ratio reconstruction); (3) *Medical-specific multi-objective methods*: Models Genesis [86] (multi-scale restoration), Volume Fusion [63] (pseudo-segmentation), and SwinUNETR [54] (combined inpainting, rotation prediction, and contrastive learning). We additionally compare against large-scale foundation models BrainMVP [47], BrainIAC [52], and MRI-Core [15]. All SSL baselines are pretrained on identical OpenMind splits with matched compute budget; foundation models use official pretrained weights. Detailed experimental configurations are provided in Appendix A.

5.1. Main Results

In-distribution performance. We evaluate MDAE on T1 and T2 sequences abundant in pretraining across 6 task-modality combinations spanning tumor classification and molecular marker detection (Table 1). MDAE achieves 73.6% mean AUROC and 71.6% mean AP, outperforming SSL baselines by 4.1% over MAE (69.5%) and foundation models by 6.2% over BrainIAC (67.4%). For tumor type discrimination (BraTS23: glioma vs metastasis), MDAE achieves 96.3% AUROC on T1 and 96.6% on T2. On molecular marker detection, IDH1 mutation detection reaches 69.6% (T1) and 60.4% (T2) AUROC, while MGMT methylation achieves 58.9% (T1) and 59.6% (T2). On individual metrics, BrainMVP achieves competitive MGMT AP (59.6% vs. 57.6%) and SwinUNETR matches on T2 IDH1 AP (60.0% vs. 56.4%), though MDAE leads decisively on aggregate.

Cross-modal generalization. To assess generalization capability, we evaluate MDAE on 6 out-of-distribution tasks using sequences rare or absent in OpenMind pretraining: FLAIR, T1-gadolinium, ASL, and SWI (Table 2). MDAE

Table 1. Performance comparison on T1 and T2 modalities. All methods are evaluated on three brain tumor classification tasks: BraTS23 glioma vs metastasis (Tumor Type), RSNA-MICCAI MGMT methylation status (MGMT Methylation), and UPenn-GBM IDH1 mutation detection (IDH1 Status). Values represent AUROC and average precision (AP) in %. Best results per row in **bold**, second-best underlined.

	Tumor Type				**MGMT Methylation**				**IDH1 Status**				**Average**	
	T1		T2		T1		T2		T1		T2		AUROC	AP
Method	AUROC	AP	AUROC	AP	AUROC	AP	AUROC	AP	AUROC	AP	AUROC	AP		
No pretrain	85.4	86.1	86.7	78.3	38.0	43.2	54.1	55.7	51.3	53.3	48.6	53.0	60.7	61.6
SSL Baselines														
MAE	86.3	78.7	96.2	85.0	52.5	57.1	58.6	57.7	63.4	57.6	59.8	58.6	69.5	65.8
SimCLR	81.4	70.4	93.0	83.4	58.3	58.5	54.7	53.8	46.9	51.0	53.9	51.4	64.7	61.4
VoCo	91.4	82.9	96.1	90.2	53.8	56.5	53.0	53.4	50.3	50.7	57.3	56.4	67.0	65.0
MG	71.8	92.6	80.1	72.7	56.6	57.7	53.4	51.9	53.2	51.0	53.4	52.8	61.4	63.1
VF	83.8	71.4	77.8	70.6	52.1	52.7	54.8	56.0	44.5	49.0	44.0	49.3	59.5	58.2
S3D	91.0	93.7	77.5	69.0	54.5	55.2	52.9	53.6	59.6	55.7	53.2	53.8	64.8	63.5
SwinUNETR	80.4	66.3	87.3	78.3	54.3	57.6	50.4	51.7	49.7	51.1	59.1	**60.0**	63.5	60.8
Foundation Models														
BrainIAC	86.4	90.9	92.8	85.0	51.4	53.1	58.8	58.5	59.0	55.3	55.9	54.0	67.4	66.1
BrainMVP	92.6	87.7	77.5	69.5	58.7	**59.6**	52.1	52.5	64.3	59.3	56.2	56.0	66.9	64.1
MRI-Core	58.3	53.7	80.6	73.1	56.3	57.4	52.4	50.8	52.8	51.0	48.3	51.3	58.1	56.2
General Vision Models														
DinoV2	72.5	67.4	76.8	61.0	54.2	54.4	51.6	51.0	49.8	50.7	55.7	57.8	60.1	57.1
ResNet-50	75.4	69.1	87.0	74.8	52.3	51.8	45.7	49.4	42.4	47.0	48.9	51.1	58.6	57.2
MDAE	**96.3**	**95.0**	**96.6**	**95.4**	**58.9**	57.6	**59.6**	**61.0**	**69.6**	**64.3**	**60.4**	56.4	**73.6**	**71.6**

achieves 78.6% mean AUROC and 75.4% mean AP, improving 8.6% over MAE (70.0%) and 10.7% over BrainIAC (67.9%). MDAE achieves 91.8% AUROC for age stratification on FLAIR and 84.7%/81.9% AUROC for IDH classification on ASL/SWI respectively, providing 6.2% and 6.0% improvements over MAE on the latter two tasks. Notably, DinoV2—a general-domain vision model without medical pretraining—achieves 72.1% mean AUROC, competitive with domain-specific SSL methods; MDAE's 6.5% improvement over DinoV2 demonstrates the additional value of medical-specific dual corruption pretraining.

Multi-modality integration. Clinical brain tumor diagnosis relies on integrating complementary information from multiple MRI modalities. We evaluate MDAE on both classification and segmentation using BraTS18 (4 modalities: T1, T1-Gd, T2, FLAIR) and UCSF-PDGM (6 modalities: T1, T1c, T2, FLAIR, SWI, ASL). Each modality is processed independently through the pretrained encoder with mean pooling aggregation (Table 3). Due to computational constraints of full finetuning foundation models on multi-modality data, we focus on SSL methods with shared encoder architecture. For classification, MDAE achieves 92.1% AUROC on BraTS18 tumor grading and 93.0% on UCSF-PDGM IDH classification, representing 2.0% and 0.7% improvements over the strongest baselines respectively. Complete classification metrics (AP, F1, balanced accuracy) are provided in Appendix F. For dense prediction, MDAE achieves 85.2% Dice and 88.1% NSD on UCSF-PDGM segmentation and 81.4% Dice and 75.3% NSD on BraTS18 test set, outperforming all baselines. Detailed per-region results are provided in Appendix E. Qualitative results in Appendix E (Figure 4) further show improved boundary delineation with fewer false positives at tumor margins.

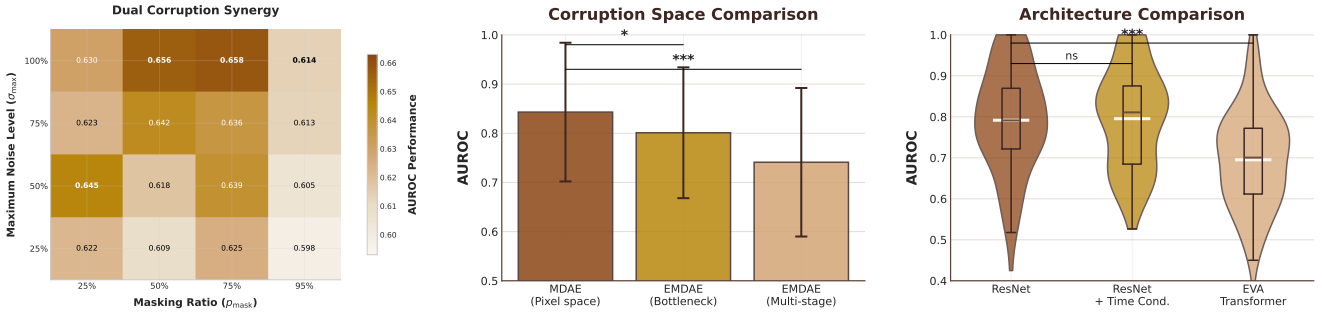
5.2. Ablation Studies

We ablate five design axes of MDAE: (i) the synergy between masking and diffusion corruption, (ii) corruption space (pixel vs. embedding), (iii) masking strategy and ratio, (iv) noise schedule, and (v) encoder architecture. Unless stated otherwise, all ablation models are pretrained on a 10% subset of OpenMind for 100 epochs and evaluated by AUROC on BraTS18 LGG-vs-HGG classification, isolating each design choice under matched data and compute. Key findings—particularly the superiority of variable masking (Table 4)—are corroborated by full-scale results, where the final model trained with variable masking on the complete dataset achieves the best downstream performance (Tables 1–3).

Dual corruption synergy: Figure 3a sweeps the corruption parameter space over four masking ratios

Table 2. Generalization performance on diverse MRI modalities. Tasks include UCSF-PDGM IDH classification using arterial spin labeling (ASL) and susceptibility-weighted imaging (SWI), UPenn-GBM age group classification using FLAIR, UPenn-GBM gross total resection (GTR) status using T1-gadolinium (T1GD), and TCGA-GBM 1-year disease-specific survival (DSS) and progression-free interval (PFI) using mixed-contrast protocols. Values represent AUROC and AP in %. Best results per row in **bold**, second-best underlined. Some foundation models did not provide predictions for TCGA tasks.

	IDH Classification				**Age Group**		**GTR Status**		**TCGA Survival**				**Average**	
	ASL		SWI		FLAIR		T1GD		DSS 1-year		PFI 1-year		AUROC	AP
Method	AUROC	AP	AUROC	AP	AUROC	AP	AUROC	AP	AUROC	AP	AUROC	AP		
No pretrain	57.8	58.5	63.1	61.3	83.2	82.5	50.5	50.8	68.2	64.2	40.6	47.5	60.6	60.8
SSL Baselines														
MAE	78.5	74.3	75.9	65.5	84.5	82.9	55.1	53.4	57.8	60.3	68.4	68.6	70.0	67.5
SimCLR	76.8	73.0	72.6	65.3	86.1	84.4	51.7	51.6	70.1	72.9	64.7	67.8	70.3	69.2
VoCo	78.8	76.0	75.9	71.4	86.2	85.4	53.6	56.1	62.3	63.3	69.2	64.9	71.0	69.5
MG	75.9	71.7	71.0	62.1	85.5	85.8	56.9	58.4	40.3	46.2	43.6	48.6	62.2	62.2
VF	79.7	70.4	75.8	70.9	85.0	84.8	47.4	49.3	52.6	59.5	60.9	59.1	66.9	65.7
S3D	60.7	58.6	62.4	59.0	85.1	83.9	50.9	51.4	65.6	70.0	54.1	57.4	63.2	63.4
SwinUNETR	78.9	72.2	69.0	61.5	86.9	86.8	58.3	57.7	35.1	45.8	54.1	55.9	63.7	63.3
Foundation Models														
BrainIAC	79.2	75.1	74.1	70.8	76.2	74.1	47.1	49.4	70.8	73.7	60.2	61.1	67.9	67.4
BrainMVP	65.6	59.4	69.1	62.7	55.5	56.1	53.2	54.2	46.8	51.5	53.4	56.8	57.3	56.8
MRI-Core	69.3	60.2	52.5	52.1	56.9	54.3	45.4	48.1	58.3	61.0	69.9	**71.5**	58.7	57.8
General Vision Models														
DinoV2	78.8	74.0	77.0	69.5	87.1	86.2	60.3	**60.6**	57.1	61.0	72.2	70.2	72.1	70.2
ResNet-50	66.1	61.1	63.8	59.8	58.1	59.8	44.3	48.1	70.1	70.0	68.4	63.1	61.8	60.3
MDAE	**84.7**	**81.8**	**81.9**	**73.6**	**91.8**	**92.2**	**63.0**	60.1	**76.0**	**73.9**	**74.4**	70.6	**78.6**	**75.4**



(a) Dual corruption parameter landscape (b) Corruption space comparison (c) Architecture comparison

Figure 3. **Ablation study overview.** (a) AUROC landscape over masking ratio and noise level, showing synergistic interaction between masking and noise (see §5.2 for analysis). (b) Pixel-space vs. embedding-space (EMDAE) corruption across encoder depths. (c) Convolutional (ResNet), ResNet with time conditioning, and transformer (EVA) encoder architectures.

($p_{\text{mask}} \in \{25, 50, 75, 95\}\%$) and four noise levels ($\sigma_{\max} \in \{25, 50, 75, 100\}\%$). The global optimum of 0.658 AUROC occurs at 75% masking with $\sigma_{\max}$=100%, substantially above the weakest configuration (0.598 at 95% masking, $\sigma_{\max}$=25%). Notably, the optimal noise level shifts with masking ratio: low masking (25%) peaks at moderate noise ($\sigma_{\max}$=50%, 0.645), whereas high masking (75%) peaks at full noise ($\sigma_{\max}$=100%, 0.658). This coupling indicates that masking and diffusion noise are not independent augmentations but jointly shape the quality of learned representations.

Corruption space: Pixel-space corruption (MDAE) outperforms EMDAE (Embedding-space MDAE), which instead applies masking and noise in the encoder's latent space, at both bottleneck and multi-stage injection points by 4.2–10.2% AUROC (Figure 3b), indicating that corrupting raw inputs yields stronger representations than corrupting learned features.

Table 3. **Multi-modality brain tumor evaluation** on BraTS18 (4 modalities) and UCSF-PDGM (6 modalities). Classification reports AUROC; segmentation reports mean Dice and NSD. Best results in **bold**, second-best underlined.

	BraTS18 (4 modalities)			UCSF-PDGM (6 modalities)		
	Cls.	Segmentation		Cls.	Segmentation	
Method	AUROC	Dice	NSD	AUROC	Dice	NSD
MAE	90.1	80.6	75.0	87.8	84.8	87.7
SimCLR	78.7	80.1	74.1	80.3	84.1	87.1
VoCo	84.8	80.9	74.7	91.6	84.1	87.2
MG	84.9	80.7	75.1	92.3	84.3	87.3
VF	83.7	80.7	74.7	90.2	83.6	86.5
S3D	89.4	80.2	74.3	81.9	84.2	87.1
SwinUNETR	82.7	80.7	75.1	86.5	83.7	86.5
MDAE	**92.1**	**81.4**	**75.3**	**93.0**	**85.2**	**88.1**

Table 4. Masking ratio ablation with $\sigma_{\max} = 1.0$. Variable masking outperforms all fixed-ratio variants.

Masking Ratio	75%	80%	85%	90%	95%	Var. $\mathcal{U}(.01, .99)$
Mean AUROC	0.625	0.610	0.623	0.595	0.568	**0.662**

Masking strategy and ratio: Block-wise masking with 16^3 voxel patches outperforms random voxel masking (0.623 vs. 0.600 AUROC), showing that spatially contiguous masking is more effective for encouraging volumetric structural reasoning. Fixed-ratio analysis (Table 4) shows that performance peaks at 75% masking (0.625 AUROC) and remains competitive in the 75–85% range, while degrading at extreme masking (95%: 0.568 AUROC). Notably, all fixed masking ratios underperform variable masking, indicating that sampling p_{mask} during pretraining is more effective than optimizing for a single operating point.

Noise schedule: Variance-Preserving (VP) [26], Variance-Exploding (VE) [51], and Flow Matching [40] schedules yield nearly identical performance (VE: 0.615, Flow Matching: 0.614, VP: 0.613 AUROC), indicating robustness to the choice of diffusion corruption scheme.

Encoder architecture: ResNet outperforms the Vision Transformer [61] by 12.1–13.8% AUROC (Figure 3c), suggesting that convolutional inductive biases are well suited to medical imaging's continuous anatomical structures. Adding time conditioning via FiLM layers [46, 59] to ResNet provides marginal gains, concentrated on texture-sensitive tasks (see Appendix C for implementation details).

Loss weight balance: We ablate the loss weights λ_{masked} and λ_{visible} on UCSF-PDGM (6 modalities/contrasts), which provides both classification and segmentation benchmarks (Table 5). Equal weighting ($\lambda_{\text{masked}} = \lambda_{\text{visible}} = 1.0$) achieves the best overall balance across tasks, yielding the highest mean segmentation Dice (92.6) and the highest mean classification score (90.7). Single-objective variants show a clear tradeoff: visible-only optimization (λ_{masked}=0, λ_{visible}=1) remains competitive on classification (90.4 mean) but underperforms on segmentation (91.8 mean), whereas masked-only optimization (λ_{masked}=1, λ_{visible}=0) improves segmentation relative to visible-only (92.3 mean) but reduces classification performance (86.4 mean). Intermediate asymmetric settings remain competitive but do not surpass equal weighting overall. These results support the use of balanced loss weights in MDAE and further suggest that the masked reconstruction and visible-region denoising terms provide complementary supervision.

Table 5. **Loss weight ablation on UCSF-PDGM.** Validation performance for segmentation (Dice %) and classification (AUROC/AP %) with varying loss weights.

		Segmentation				Classification		
λ_m	λ_v	NET	ED	ET	Mean	AUROC	AP	Mean
0.0	1.0	92.1	90.3	**93.0**	91.8	91.5	**89.3**	90.4
1.0	0.0	93.3	**91.1**	92.4	92.3	89.7	83.1	86.4
0.5	1.0	93.7	90.7	92.1	92.2	89.5	83.9	86.7
1.0	0.5	94.3	90.5	92.4	92.4	91.7	89.2	90.5
1.0	1.0	**94.5**	90.7	92.7	**92.6**	**92.2**	89.1	**90.7**

6. Conclusion

We introduced MDAE, a self-supervised framework that imposes concurrent spatial masking and diffusion corruption to learn medical image representations. This dual corruption creates complementary learning objectives—masked reconstruction for anatomical structure and visible denoising for tissue texture—within a unified time-conditioned framework. Across 16 brain MRI benchmarks spanning tumor classification, molecular marker detection, and dense segmentation, MDAE achieves consistent performance improvements, with particularly strong cross-modal generalization. Ablation studies confirm that dual corruption provides complementary supervision, with variable masking and balanced loss weighting each contributing measurably. While our evaluation focuses on brain MRI, the approach is extensible to other medical imaging modalities and demonstrates that combining structural and textural objectives provides effective self-supervised pretraining for medical imaging.

References

[1] Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Michael Rabbat, and Nicolas Ballas. Masked siamese networks for label-efficient learning. In *European conference on computer vision*, pages 456–473. Springer, 2022. 2

[2] Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel

Bilello, Martin Rozycki, Justin S Kirby, John B Freymann, Keyvan Farahani, and Christos Davatzikos. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific data*, 4(1): 1–13, 2017. 13

[3] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 2

[4] Valay Bundele, Karahan Sarıtaş, Bora Kargi, Oğuz Ata Çal, Kıvanç Tezören, Zohreh Ghaderi, and Hendrik Lensch. Evaluating self-supervised learning in medical imaging: A benchmark for robustness, generalizability, and multi-domain impact. *arXiv preprint arXiv:2412.19124*, 2024. 1, 2

[5] Evan Calabrese, Javier E Villanueva-Meyer, Jeffrey D Rudie, Andreas M Rauschecker, Ujjwal Baid, Spyridon Bakas, Soonmee Cha, John T Mongan, and Christopher P Hess. The university of california san francisco preoperative diffuse glioma mri dataset. *Radiology: Artificial Intelligence*, 4 (6):e220058, 2022. 5, 13

[6] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020. 1, 2

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 1, 2

[8] Gabriel Chartrand, Phillip M Cheng, Eugene Vorontsov, Michal Drozdzal, Simon Turcotte, Christopher J Pal, Samuel Kadoury, and An Tang. Deep learning: a primer for radiologists. *Radiographics*, 37(7):2113–2131, 2017. 1, 2

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020. 1, 2, 5, 13

[10] Ting Chen, Calvin Luo, and Lala Li. Intriguing properties of contrastive losses. *Advances in Neural Information Processing Systems*, 34:11834–11845, 2021. 2

[11] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758, 2021. 1

[12] Xiangxiang Chu, Renda Li, and Yong Wang. Usp: Unified self-supervised pretraining for image generation and understanding. *arXiv preprint arXiv:2503.06132*, 2025. 1, 2

[13] Ching-Yao Chuang, R Devon Hjelm, Xin Wang, Vibhav Vineet, Neel Joshi, Antonio Torralba, Stefanie Jegelka, and Yale Song. Robust contrastive learning against noisy views. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16670–16681, 2022. 1, 2

[14] Maria Correia de Verdier, Rachit Saluja, Louis Gagnon, Dominic LaBella, Ujjwall Baid, Nourel Hoda Tahon, Martha Foltyn-Dumitru, Jikai Zhang, Maram Alafif, Saif Baig, et al. The 2024 brain tumor segmentation (brats) challenge: Glioma segmentation on post-treatment mri. *arXiv preprint arXiv:2405.18368*, 2024. 13

[15] Haoyu Dong, Yuwen Chen, Hanxue Gu, Nicholas Konz, Yaqian Chen, Qihang Li, and Maciej A Mazurowski. Mricore: A foundation model for magnetic resonance imaging. *arXiv preprint arXiv:2506.12186*, 2025. 2, 5, 13

[16] Tim Donovan and David J Manning. The radiology task: Bayesian theory and perception. *The British journal of radiology*, 80(954):389–391, 2007. 1

[17] Harold W Fischer. Principles of chest roentgenology. *Archives of Internal Medicine*, 117(2):308–308, 1966. 1

[18] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020. 1, 2

[19] Anubhav Gupta, Islam Osman, Mohamed S Shehata, W John Braun, and Rebecca E Feldman. Medmae: A self-supervised backbone for medical imaging tasks. *Computation*, 13(4):88, 2025. 1, 2

[20] Fatemeh Haghighi, Mohammad Reza Hosseinzadeh Taher, Zongwei Zhou, Michael B Gotway, and Jianming Liang. Transferable visual words: Exploiting the semantics of anatomical patterns for self-supervised learning. *IEEE transactions on medical imaging*, 40(10):2857–2868, 2021. 1, 2

[21] Philippe Hansen-Estruch, Sriram Vishwanath, Amy Zhang, and Manan Tomar. Unified auto-encoding with masked diffusion. *arXiv preprint arXiv:2406.17688*, 2024. 2

[22] Ali Hatamizadeh, Jiaming Song, Guilin Liu, Jan Kautz, and Arash Vahdat. Diffit: Diffusion vision transformers for image generation. In *European Conference on Computer Vision*, pages 37–55. Springer, 2024. 14

[23] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 1, 2

[24] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16000–16009, 2022. 1, 2, 3, 5, 13

[25] Yuting He, Guanyu Yang, Rongjun Ge, Yang Chen, Jean-Louis Coatrieux, Boyu Wang, and Shuo Li. Geometric visual similarity learning in 3d medical image self-supervised pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9538–9547, 2023. 2

[26] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 3, 8

[27] Marie Humbert-Droz, Pritam Mukherjee, and Olivier Gevaert. Strategies to address the lack of labeled data for supervised machine learning training with electronic health records: Case study for the extraction of symptoms from

clinical notes. *JMIR Medical Informatics*, 10(3):e32903, 2022. 1

[28] 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, 15

[29] Sarah Jabbour, Gregory Kondas, Ella Kazerooni, Michael Sjoding, David Fouhey, and Jenna Wiens. Depict: Diffusion-enabled permutation importance for image classification tasks. In *European Conference on Computer Vision*, pages 35–51. Springer, 2024. 2

[30] Daniel P Jeong, Suhana Bedi, Cliff Wong, Sheng Zhang, Jeya Maria Jose Valanarasu, Qianchu Liu, Reuben Tan, Zelalem Gero, Jaspreet Bagga, Juan Manuel Zambrano Chaves, et al. Pretraining patient foundation models on multimodal patient journeys. In *NeurIPS 2025 Workshop on Learning from Time Series for Health*, 2025. 2

[31] Chongcong Jiang, Tianxingjian Ding, Chuhan Song, Jiachen Tu, Ziyang Yan, Yihua Shao, Zhenyi Wang, Yuzhang Shang, Tianyu Han, and Yu Tian. Medical sam3: A foundation model for universal prompt-driven medical image segmentation. *arXiv preprint arXiv:2601.10880*, 2026. 2

[32] Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. *arXiv preprint arXiv:2110.09348*, 2021. 1, 2

[33] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022. 5, 18

[34] Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *Advances in neural information processing systems*, 34:21696–21707, 2021. 5, 18

[35] Yilmaz Korkmaz, Tolga Cukur, and Vishal M Patel. Self-supervised mri reconstruction with unrolled diffusion models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 491–501. Springer, 2023. 2

[36] HL Kundel and CF Nodine. A visual concept shapes image perception. *Radiology*, 146(2):363–368, 1983. 1

[37] Sonja Kunzmann, Mathias Öttl, Prathmesh Madhu, Felix Denzinger, and Andreas Maier. An unobtrusive quality supervision approach for medical image annotation. *arXiv preprint arXiv:2211.06146*, 2022. 1, 2

[38] Alexander C Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2206–2217, 2023. 1, 2

[39] Daiqing Li, Huan Ling, Amlan Kar, David Acuna, Seung Wook Kim, Karsten Kreis, Antonio Torralba, and Sanja Fidler. DreamTeacher: Pretraining image backbones with deep generative models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16698–16708, 2023. 2

[40] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 8

[41] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017. 1, 2

[42] Zixuan Liu, Hanwen Xu, Addie Woicik, Linda G. Shapiro, Marian Blazes, Yue Wu, Verena Steffen, Catherine Cukras, Cecilia S. Lee, Miao Zhang, Aaron Y. Lee, and Sheng Wang. OCTCube-M: A 3d multimodal optical coherence tomography foundation model for retinal and systemic diseases with cross-cohort and cross-device validation. *arXiv preprint arXiv:2408.11227*, 2024. 2

[43] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993–2024, 2014. 13

[44] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 1, 2, 13

[45] Yatian Pang, Wenxiao Wang, Francis E.H. Tay, Wei Liu, Yonghong Tian, and Li Yuan. Masked autoencoders for point cloud self-supervised learning. In *European Conference on Computer Vision (ECCV)*, pages 604–621, 2022. 2

[46] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 4, 8, 14

[47] Shaohao Rui, Lingzhi Chen, Zhenyu Tang, Lilong Wang, Mianxin Liu, Shaoting Zhang, and Xiaosong Wang. Multimodal vision pre-training for medical image analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5164–5174, 2025. 2, 5, 13

[48] Sushmita Sarker, Prithul Sarker, George Bebis, and Alireza Tavakkoli. Can score-based generative modeling effectively handle medical image classification? In *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2025. 2

[49] Marc-Andre Schulz, Bertrand Thirion, Alexandre Gramfort, Gaël Varoquaux, and Danilo Bzdok. Label scarcity in biomedicine: Data-rich latent factor discovery enhances phenotype prediction. *arXiv preprint arXiv:2110.06135*, 2021. 1, 2

[50] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 2, 3, 5, 15

[51] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 8

[52] Divyanshu Tak, Biniam A Garomsa, Tafadzwa L Chaunzwa, Anna Zapaishchykova, Juan Carlos Climent Pardo,

Zezhong Ye, John Zielke, Yashwanth Ravipati, Sri Vajapeyam, Maryam Mahootiha, et al. A foundation model for generalized brain mri analysis. *medRxiv*, 2024. 2, 5, 13

[53] Aiham Taleb, Winfried Loetzsch, Noel Danz, Julius Severin, Thomas Gaertner, Benjamin Bergner, and Christoph Lippert. 3d self-supervised methods for medical imaging. *Advances in neural information processing systems*, 33:18158–18172, 2020. 2

[54] Yucheng Tang, Dong Yang, Wenqi Li, Holger R Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, and Ali Hatamizadeh. Self-supervised pre-training of swin transformers for 3d medical image analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20730–20740, 2022. 5, 13

[55] Changyao Tian, Chenxin Tao, Jifeng Dai, Hao Li, Ziheng Li, Lewei Lu, Xiaogang Wang, Hongsheng Li, Gao Huang, and Xizhou Zhu. ADDP: Learning general representations for image recognition and generation with alternating denoising diffusion process. In *International Conference on Learning Representations (ICLR)*, 2024. 2

[56] Jussi Tohka. Partial volume effect modeling for segmentation and tissue classification of brain magnetic resonance images: A review. *World journal of radiology*, 6(11):855, 2014. 1

[57] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022. 2

[58] Jiachen Tu, Yaokun Shi, and Fan Lam. An ambient denoising score matching based self-supervised denoising approach for multicontrast low-field MRI. In *Proceedings of the International Society for Magnetic Resonance in Medicine (ISMRM)*, 2025. 2

[59] Jiachen Tu, Yaokun Shi, and Fan Lam. Score-based self-supervised mri denoising. In *International Conference on Learning Representations*, 2025. 2, 8

[60] Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011. 3, 5

[61] Tassilo Wald, Saikat Roy, Fabian Isensee, Constantin Ulrich, Sebastian Ziegler, Dasha Trofimova, Raphael Stock, Michael Baumgartner, Gregor Köhler, and Klaus Maier-Hein. Primus: Enforcing attention usage for 3d medical image segmentation. *arXiv preprint arXiv:2503.01835*, 2025. 8

[62] Tassilo Wald, Constantin Ulrich, Jonathan Suprijadi, Sebastian Ziegler, Michal Nohel, Robin Peretzke, Gregor Köhler, and Klaus H. Maier-Hein. An openmind for 3d medical vision self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23839–23879, 2025. 1, 2, 5, 13

[63] Guotai Wang, Jia Fu, Jianghao Wu, Xiangde Luo, Yubo Zhou, Xinglong Liu, Kang Li, Jingsheng Lin, Baiyong Shen, and Shaoting Zhang. Volume fusion-based self-supervised pretraining for 3d medical image segmentation. *IEEE Transactions on Image Processing*, 2025. 5, 13

[64] Chen Wei, Karttikeya Mangalam, Po-Yao Huang, Yanghao Li, Haoqi Fan, Hu Xu, Huiyu Wang, Cihang Xie, Alan Yuille, and Christoph Feichtenhofer. Diffusion models as masked autoencoders. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16284–16294, 2023. 1, 2

[65] Daniel Wolf, Tristan Payer, Catharina Silvia Lisson, Christoph Gerhard Lisson, Meinrad Beer, Michael Götz, and Timo Ropinski. Self-supervised pre-training with contrastive and masked autoencoder methods for dealing with small datasets in deep learning for medical imaging. *Scientific Reports*, 13(1):20260, 2023. 1

[66] Linshan Wu, Jiaxin Zhuang, and Hao Chen. Voco: A simple-yet-effective volume contrastive learning framework for 3d medical image analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22873–22882, 2024. 2, 5, 13

[67] Weilai Xiang, Hongyu Yang, Di Huang, and Yunhong Wang. Denoising diffusion autoencoders are unified self-supervised learners. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15802–15812, 2023. 1, 2

[68] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9653–9663, 2022. 2

[69] Hanwen Xu, Naoto Usuyama, Jaspreet Bagga, Sheng Zhang, Rajesh Rao, Tristan Naumann, Cliff Wong, Zelalem Gero, Javier González, Yu Gu, et al. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 630(8015):181–188, 2024. 2

[70] Yan Xu, Rixiang Quan, Weiting Xu, Yi Huang, Xiaolong Chen, and Fengyuan Liu. Advances in medical image segmentation: A comprehensive review of traditional, deep learning and hybrid approaches. *Bioengineering*, 11(10): 1034, 2024. 1

[71] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. arxiv 2022. *arXiv preprint arXiv:2209.00796*, 2022. 2

[72] Qihang Yu, Mark Weber, Xueqing Deng, Xiaohui Shen, Daniel Cremers, and Liang-Chieh Chen. An image is worth 32 tokens for reconstruction and generation. *Advances in Neural Information Processing Systems*, 37:128940–128966, 2024. 1

[73] Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *ICLR*, 2025. 1, 2

[74] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International conference on machine learning*, pages 12310–12320. PMLR, 2021. 1, 2

[75] Yu Zhang, Jing Chen, Xiangxun Ma, Gang Wang, Uzair Aslam Bhatti, and Mengxing Huang. Interactive medical image annotation using improved attention u-net with compound geodesic distance. *Expert systems with applications*, 237:121282, 2024. 1, 2

[76] Ruiyang Zhao and Fan Lam. An adaptive, disentangled representation for multidimensional mri reconstruction. *arXiv preprint arXiv:2512.24674*, 2025. 2

[77] Ruiyang Zhao, Jiachen Tu, and Fan Lam. Latent diffusion based adaptive image representation for multidimensional mri. In *2025 IEEE 10th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 246–250. IEEE, 2025. 2

[78] Theodore Zhengde Zhao, Yu Gu, Jianwei Yang, Naoto Usuyama, Ho Hin Lee, Tristan Naumann, Jianfeng Gao, Angela Crabtree, Jacob Abel, Christine Moung-Wen, et al. Biomedparse: a biomedical foundation model for image parsing of everything everywhere all at once. *arXiv preprint arXiv:2405.12971*, 2024. 2

[79] Theodore Zhengde Zhao, Sid Kiblawi, Naoto Usuyama, Ho Hin Lee, Sam Preston, Hoifung Poon, and Mu Wei. Boltzmann attention sampling for image analysis with small objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25950–25959, 2025. 2

[80] Theodore Zhengde Zhao, Ho Hin Lee, Alberto Santamaria-Pang, Noel C Codella, Sid Kiblawi, Yu Gu, Yu Fang, Wenxuan Teng, Naiteek Sangani, Ivan Tarapov, et al. Biomedparse-v: Scaling foundation model for universal text-guided volumetric biomedical image segmentation. In *CVPR 2025: Foundation Models for 3D Biomedical Image Segmentation*, 2025. 2

[81] Theodore Zhengde Zhao, Sid Kiblawi, Jianwei Yang, Naoto Usuyama, Reuben Tan, Noel C Codella, Tristan Naumann, Hoifung Poon, and Mu Wei. Learning sparse visual representations via spatial-semantic factorization. *arXiv preprint arXiv:2602.01905*, 2026. 2

[82] Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. *arXiv preprint arXiv:2510.11690*, 2025. 1, 2

[83] Hong-Yu Zhou, Chixiang Lu, Sibei Yang, Xiaoguang Han, and Yizhou Yu. Preservational learning improves self-supervised medical image models by reconstructing diverse contexts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3499–3509, 2021. 2

[84] Hong-Yu Zhou, Chixiang Lu, Chaoqi Chen, Sibei Yang, and Yizhou Yu. A unified visual information preservation framework for self-supervised pre-training in medical image analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):8020–8035, 2023. 2

[85] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021. 2

[86] Zongwei Zhou, Vatsal Sodha, Md Mahfuzur Rahman Siddiquee, Ruibin Feng, Nima Tajbakhsh, Michael B Gotway, and Jianming Liang. Models genesis: Generic autodidactic models for 3d medical image analysis. In *International conference on medical image computing and computer-assisted intervention*, pages 384–393. Springer, 2019. 1, 2, 5, 13