VOLT: 3D MEDICAL FOUNDATION MODEL FOR VOLUMETRIC LATENT TOKENIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Medical foundation model (FM) development faces a critical trade-off between semantically aligned multimodal models constrained by paired data and scalable scan-only models with weaker clinical grounding, compounded by volumetric anisotropy and the scarcity of paired scan–text data. To bridge this gap, we introduce **VOLT**, a scan-only 3D medical FM for high-fidelity volumetric tokenization trained through a two-stage strategy. Stage-1 learns robust through-plane priors through Depth Shuffle reconstruction and discretizes latents into compact codewords using the proposed Additive Binary Spherical Quantization (ABSQ) module. Stage-2 aligns these representations with clinical semantics through patch-consistency matching and distillation from a frozen vision-language teacher, without requiring paired reports for the training scans. **VOLT** yields semantically structured representations that transfer effectively across segmentation, classification, visual question answering (VQA), report generation, and synthetic data generation. Compared with the strongest reported baseline for each metric, **VOLT** achieves relative improvements of 13.57% in PSNR and 6.22% in SSIM for volumetric reconstruction, 0.35% in Dice and a 10.98% reduction in HD95 for segmentation, and 1.84% in mean VQA accuracy. For report generation, it improves BLEU, METEOR, and BERTScore by 4.47%, 2.11%, and 0.52%, respectively. Code and dataset will be open-sourced.

1 INTRODUCTION

Foundation models (FMs) have shifted computer vision from task-specific architectures toward general-purpose pretrained backbones transferable across diverse tasks with minimal supervision. This progress spans ImageNet scale supervised pretraining (Deng et al., 2009; Krizhevsky et al., 2012), transformer-based reconstruction (Dosovitskiy, 2020; He et al., 2022), self-supervised contrastive and distillation learning (Chen et al., 2020; He et al., 2020; Caron et al., 2021), multimodal image-text alignment (Radford et al., 2021; Kirillov et al., 2023)

In medical imaging, however, this paradigm breaks down. Unlike natural images, medical data are shaped by modality-specific acquisition physics (signal generation, encoding, reconstruction, etc.) and heterogeneous protocols. Key diagnostic cues are often subtle, sparse, and distributed across anisotropic 3D volumes (*e.g.*, CT/MRI) rather than isolated 2D slices (Litjens et al., 2017; Jia et al., 2015). Moreover, large-scale 3D medical corpora are scarce, and paired scan-text datasets are even rarer and heavily domain-restricted. This creates a central tension for medical FMs: scan-text pretraining is clinically aligned but depends on scarce paired data, whereas scalable scan-only pretraining risks weak clinical grounding.

Recent medical FM research follows two main directions: vision-language pretraining, which enables transferable multimodal embeddings and generalist assistants through contrastive alignment and instruction tuning (Zhang et al., 2022; Li et al., 2023); and volumetric visual pretraining, where self-supervised learning, universal 3D architectures, and modality-agnostic tokenization aim to build scalable scan-native representations (Zhou et al., 2019; Chen et al., 2019; Luo et al., 2023; Varma et al., 2025; Ma et al., 2025). However, neither yields general-purpose representations that fully capture 3D clinical structure. Existing FMs either (i) operate purely in 2D or treat 3D volumes as independent slices, losing volumetric context (Liu et al., 2025a); (ii) rely on paired image-text supervision, limiting scalability to domains with abundant radiology reports (Ryu et al.,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

2025; Zhang et al., 2023; Wang et al., 2022); or (iii) use task-specific supervision rather than general-purpose pretraining (Liu et al., 2025b; Chen et al., 2024b). Moreover, existing medical tokenizers either lack autoregressive modeling (Varma et al., 2025) or operate only in 2D (Ma et al., 2025), limiting their use as universal volumetric FMs.

To enable semantically structured 3D representation learning while retaining scan-only scalability, we propose **VOLT**, which combines a novel *codebook-free discrete* volumetric latent space with a two-stage pretraining scheme that separates *3D anatomical/volumetric priors* from *semantic organization*. This injects clinical semantics into a scan-only backbone via cross-model distillation, allowing the model to inherit the semantic geometry of vision-language models without large-scale radiology reports. In **Stage 1**, our *Depth Shuffle* pretext task encourages volumetric consistency, while our ABSQ regularizes the encoder to produce stable discrete volumetric tokens without an explicit codebook, yielding anatomically coherent, through-plane consistent 3D representations across heterogeneous scans. In **Stage 2**, clinical semantics are introduced without paired scan-text supervision by aligning encoder representations with a frozen 3D vision teacher using patch-consistency and feature-distillation objectives. We pretrain **VOLT** on 73,648 3D volumes spanning multiple modalities and anatomical regions, demonstrating its general-purpose utility across segmentation, disease classification, VQA, report generation, and synthetic data generation. Our contributions are:

- We introduce **VOLT**, a 3D medical FM that learns discrete volumetric representations from diverse, scan-only CT/MRI datasets, bypassing scarce scan-text pairs.
- We propose *Depth Shuffle*, a corruption strategy that preserves per-slice structure while perturbing inter-slice order, encouraging the encoder to capture through-plane organization, long-range depth dependencies, and global anatomical progression beyond standard patch-masking.
- We introduce *Additive Binary Spherical Quantization (ABSQ)*, a stable high-capacity discrete bottleneck with additive residual refinement, and present a scan-only semantic injection strategy via patch-level consistency matching and distillation from a frozen scan-text teacher, transferring clinical semantic priors without requiring paired volumes.
- We curate 73,648 high-quality 3D volumes spanning CT and MRI across diverse anatomical regions, and show that a single frozen **VOLT** transfers across synthesis, segmentation, VQA, report generation, and classification, achieving state-of-the-art performance over prior tokenizers.

2 Related Work

Medical Foundation Models: Medical FMs encompass 2D encoders and volumetric 3D backbones. In 2D radiology, contrastive scan-text learning (Zhang et al., 2022) builds transferable embeddings, while GLoRIA (Huang et al., 2021) and BioViL (Boecking et al., 2022) add local image-text alignment for fine-grained pathology. MedCLIP (Wang et al., 2022) uses unpaired data, and BiomedCLIP (Zhang et al., 2023) scales to biomedical figure-caption pairs. LLaVA-Med (Li et al., 2023) and CheXagent (Chen et al., 2024b) combine visual encoders with LLMs for VQA and reasoning. These approaches strengthen semantic alignment but largely operate on 2D images, limiting their ability to capture volumetric context. For volumetric learning, Med3D (Chen et al., 2019) pools 3D datasets, Models Genesis (Zhou et al., 2019) uses restoration-based self-supervision, and UNETR (Hatamizadeh et al., 2022) captures long-range context for segmentation. Spatially Adaptive Networks (Luo et al., 2023) handle modality heterogeneity, while E3D-GPT (Lai et al., 2024) supports multimodal reasoning. **VOLT** learns native 3D representations from scan-only data via semantic distillation for visual and vision-language tasks.

Visual Tokenizers for Representation Learning and Generation: Visual tokenizers compress high-dimensional inputs into compact latent representations for efficient modeling. VQ-VAE (Van Den Oord et al., 2017) introduced discrete codebooks for autoregressive modeling, and VQ-GAN (Esser et al., 2021) improved fidelity through perceptual and adversarial objectives. LDM (Rombach et al., 2022) and MedDiffusion (Wang et al., 2025) use continuous latents for diffusion-based synthesis. VideoGPT (Yan et al., 2021) and Wan (Wan et al., 2025) adopt spatiotemporal tokenization, while MotionAura (Susladkar et al., 2024) combines VAE compression, vector quantization, and masked token modeling. Beyond synthesis, **VOLT** learns discrete volumetric latents that preserve clinical morphology while supporting multimodal integration.

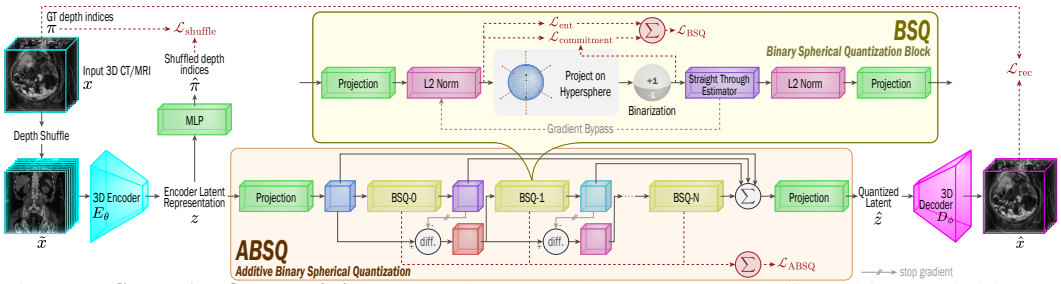


Figure 1: **Stage-1 VOLT training.** 3D volume x undergoes Depth Shuffle and is encoded by 3D encoder E_θ. The latent z is discretized via the proposed ABSQ module Q_Δ, composed of residual BSQ blocks, and decoded by D_ϕ to reconstruct $\hat{x}$. Training jointly optimizes $\mathcal{L}_{\text{rec}}$ for anatomical reconstruction, $\mathcal{L}_{\text{shuffle}}$ for depth-order reasoning, and $\mathcal{L}_{\text{ABSQ}}$ for stable discrete quantization.

Visual Tokenizers in Medical Imaging: Latent compression enables efficient volumetric synthesis and image-to-image translation (Pinaya et al., 2022a). This is particularly useful for 3D scans, where direct voxel-space modeling is computationally demanding. MedITok (Ma et al., 2025) and MedVAE (Varma et al., 2025) connect medical image synthesis and interpretation, while LDAE (Lozupone et al., 2025) and VQ-VAE-based anomaly detection (Pinaya et al., 2022b) demonstrate the value of reconstruction-based pretraining. MedITok's 2D design limits volumetric context, while spatial resizing in PUMIT (Luo et al., 2023) and MedVAE can lose fine-grained clinical detail. **VOLT** combines codebook-free discrete tokenization with native-resolution volumetric processing for both dense understanding and generation.

3 Method

Design Motivation. **VOLT** targets three complementary failure modes of existing 3D medical tokenizers. **(L1)** *Slice-continuity shortcuts:* high inter-slice correlation lets reconstruction-based models exploit local texture rather than global anatomical depth progression. **(L2)** *Codebook fragility:* single-stage codebook or binary quantizers suffer collapse, assignment noise, or residual loss. **(L3)** *Patch-level inconsistency:* patch-wise training yields locally accurate but globally incoherent latents lacking scan-level semantics. Each **VOLT** component targets one failure mode: Depth Shuffle addresses **L1** by forcing through-plane reasoning from voxel content rather than positional continuity (§3.1); ABSQ addresses **L2** via codebook-free additive binary residuals for coarse-to-fine discrete compression (§3.1); and Stage-2 patch-consistency matching with semantic distillation addresses **L3** by aligning within-volume patches and distilling clinical organization from a frozen teacher (§3.2).

Overview. Given a 3D medical volume $x \in \mathbb{R}^{D\times H\times W}$, **VOLT** learns a compact, discrete representation $\hat{z}$ resolving **L1**–**L3** without paired scan-text supervision. It comprises a 3D encoder (E_θ), a codebook-free discrete bottleneck (ABSQ; Q_Δ), and a 3D decoder (D_ϕ). The encoder extracts a continuous latent $z = E_\theta(x)$; ABSQ discretizes it as $\hat{z} = Q_\Delta(z)$; and the decoder reconstructs $\hat{x} = D_\phi(\hat{z})$. Training proceeds in two stages: **Stage-1** learns depth-aware volumetric representations with discrete tokenization; **Stage-2** refines the latent space via patch-consistent semantic distillation.

3.1 Stage 1: Through-Plane Shuffle Pretraining

The objective of this stage is to equip **VOLT** with depth-aware 3D representations. As shown in Figure (1), Stage-1 trains **VOLT** on 3D CT/MRI to jointly learn volumetric reconstruction and a depth-aware pretext task for through-plane semantics. An input volume x undergoes Depth Shuffle corruption to produce $\tilde{x}$, which is encoded by 3D encoder E_θ. The resulting latent z supports two branches: reconstruction and shuffled depth-index prediction. Before decoding, z is discretized by the proposed ABSQ module, producing a tokenizer-like discrete latent space for later stages.

Depth Shuffle pretraining objective: Masking objectives (*e.g.*, MIM) (Chen et al., 2023; Prakash et al., 2026) reconstruct missing patches from visible context, but in 3D CT/MRI often (i) favor local interpolation over global anatomy and (ii) provide no supervision for through-plane order. Our Depth Shuffle instead preserves voxel content while permuting slices along depth, producing locally plausible yet globally inconsistent volumes. Recovering slice placement and anatomical progression requires inter-slice volumetric reasoning beyond local texture matching. We supervise

this with an index prediction head on E_θ using cross-entropy: $\mathcal{L}_{\text{shuffle}} = \text{CE}(\hat{\pi}, \pi)$, where $\hat{\pi}$ and π denote predicted and ground-truth shuffled indices.

3D encoder-decoder backbone: To learn rich, stable volumetric representations, we insert the ABSQ module between E_θ and D_ϕ to discretize continuous latents (backbone and initialization details in Appendix).

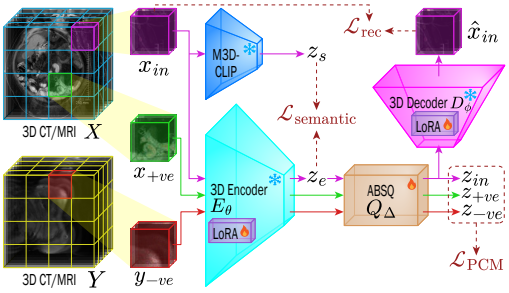


Figure 2: **Stage-2 training.** Initializing from Stage-1, anchor, positive, and negative patches are quantized and optimized via $\mathcal{L}_{\text{rec}}$ (anatomical fidelity), $\mathcal{L}_{\text{PCM}}$ (volume consistency), and $\mathcal{L}_{\text{semantic}}$ (frozen M3D-CLIP distillation) to yield scan-consistent semantic representations.

Additive Binary Spherical Quantization (ABSQ): E_θ outputs continuous latent z, which ABSQ (Figure 1) maps to a discrete manifold via N progressive quantization stages, each capturing residual information missed by prior stages. Concretely, we initialize the remainder $r_0 = z$. At stage i, a Binary Spherical Quantization (BSQ) block produces quantized component q_i, and the remainder is updated by subtracting q_i with a stop-gradient, so each stage learns residuals only: $q_i := \text{BSQ}(r_i), \quad r_{i+1} = r_i - \text{sg}(q_i), \quad \hat{z} = \sum_{i=0}^{N-1} q_i$, where $sg(.)$ denotes stop-gradient. This yields a high-capacity discrete bottleneck without a large learned codebook. ABSQ is trained by summing losses across all BSQ blocks: $\mathcal{L}_{\text{ABSQ}} = \sum_{i=0}^{N-1} \mathcal{L}_{\text{BSQ}}^{(i)}$. This encourages quantization-consistent, well-utilized codes across the batch, preventing codebook collapse.

Binary Spherical Quantization (BSQ): Following (Zhao et al., 2024), each BSQ block (Figure 1) applies a learnable projection, normalizes features onto a unit hypersphere, and binarizes them to $\{+1, -1\}^d$. A straight-through estimator preserves gradient flow through the non-differentiable sign operation, producing compact discrete codes, while a final projection restores decoder latent dimensionality. Each BSQ block is trained with a commitment loss that stabilizes code assignments via MSE between the pre-quantized spherical vector and its binarized target, and an entropy regularization term that prevents collapse by encouraging balanced bit usage across the batch: for $q \in \{+1, -1\}^d$, define $p_k = \mathbb{P}(q_k = +1)$ and $H_k = -p_k \log p_k - (1 - p_k) \log(1 - p_k)$, then maximize $H = \frac{1}{d} \sum_{k=1}^{d} H_k$ (driving $p_k \approx 0.5$) to increase effective code capacity. The BSQ loss is $\mathcal{L}_{\text{BSQ}} = \mathcal{L}_{\text{ent}} + \mathcal{L}_{\text{commit}}$.

Stage-1 objective: We decode the quantized latent $\hat{z}$ to reconstruct x using voxel-wise, structural, and perceptual losses: $\mathcal{L}_{\text{rec}} = \|x - \hat{x}\|_1 + \big(1 - \text{SSIM}(x, \hat{x})\big) + \text{LPIPS}(x, \hat{x})$. The full objective is

$$\mathcal{L}_{\text{stage-1}} = \mathcal{L}_{\text{rec}} + \alpha \mathcal{L}_{\text{shuffle}} + \beta \mathcal{L}_{\text{ABSQ}}. \tag{1}$$

Together, Depth Shuffle encourages E_θ to capture through-plane semantics, while ABSQ discretizes them into a compact latent space for later stages. We set $\alpha = 2.2$ and $\beta = 1.5$ empirically.

3.2 Stage 2: Cross-Patch Semantic Refinement

Patch-wise volumetric autoencoders can reconstruct patches accurately yet still produce visible block/seam artifacts when stitching full volumes (Pedersen et al., 2022; Tsoli et al., 2019), since reconstruction alone does not enforce shared global style factors across same-scan patches, such as MR bias-field shading, scanner-dependent contrast, or noise statistics, allowing patch-level drift. Moreover, scan-only pretraining, even with strong structural objectives, may learn features less aligned with clinical semantics than scan-text-supervised representations. Stage-2 tackles both issues by reshaping the Stage-1 latent space to be (i) consistent across patches from the same volume and (ii) semantically aligned with a scan-text embedding space, without paired training reports.

As shown in Figure 2, Stage-2 starts from the Stage-1 checkpoint and retains the same pipeline. To preserve the volumetric geometry learned in Stage-1 while enabling efficient adaptation, we freeze the encoder-decoder and insert LoRA adapters into both modules, training only the LoRA parameters and ABSQ module. In parallel, we use a pretrained frozen M3D-CLIP (Bai et al., 2024) (Vision Encoder) as a semantic teacher which is trained on CT, MRI and PET modalities. This setup allows us to inject semantics through distillation while avoiding catastrophic drift of encoder.

Training uses local 3D patch triplets. From the same volume X, we sample an anchor patch x_{in} and a positive patch x_{+ve} at different locations. From another randomly selected volume Y, we sample a negative patch y_{-ve}. All three patches are passed through the same LoRA-adapted frozen encoder followed by ABSQ, producing $z_{\text{in}} = Q_\Delta\big(E_\theta(x_{\text{in}})\big), \quad z_{+ve} = Q_\Delta\big(E_\theta(x_{+ve})\big), \quad z_{-ve} = Q_\Delta\big(E_\theta(y_{-ve})\big)$. This construction forces patches from the same scan to map to nearby representations, discouraging the model from assigning inconsistent global/style explanations to neighboring patches. This stabilizes scan-wide latent codes and smooths stitched reconstructions.

To impose semantic organization, we feed the anchor patch x_{in} into the frozen M3D-CLIP vision teacher to obtain a dense semantic feature map $\mathbf{z}_s$, and extract the pre-quantized encoder feature $\mathbf{z}_e$ from E_θ. A lightweight $1\times1\times1$ projector π_ϕ maps $\mathbf{z}_e$ to the teacher's channel dimension and trilinear resampling aligns it to the teacher's spatial grid, yielding a one-to-one token correspondence $u \in \Omega$. At each token, we form categorical distributions via a softmax along the channel axis with temperature τ=2.0, $p_s^{(u)} = \mathrm{softmax}(\mathbf{z}_s^{(u)}/\tau)$ and $p_e^{(u)} = \mathrm{softmax}(\tilde{\mathbf{z}}_e^{(u)}/\tau)$, and apply token-wise KL distillation under a stop-gradient on the teacher: $\mathcal{L}_{\text{semantic}} = \tau^2\, \mathbb{E}_{x_{in}}\big[\frac{1}{|\Omega|}\sum_{u\in\Omega} \mathrm{KL}(p_s^{(u)} \,\|\, p_e^{(u)})\big]$. Intuitively, this distills the semantic geometry learned through scan-text pretraining into our scan-only encoder, encouraging clinically similar representations to remain close in latent space.

Alongside semantic alignment, we apply a Patch Consistency Matching (PCM) loss, $\mathcal{L}_{\text{PCM}} = \ell_{\text{InfoNCE}}(z_{\text{in}}, z_{\text{+ve}}, z_{\text{-ve}})$, on the triplet embeddings to encourage same-volume similarity and cross-volume separability. Formulated as a patch-to-patch InfoNCE loss, this objective acts as a cross-patch regularizer, forcing the latent space to consistently capture global scan-level factors across local regions, a crucial constraint absent in reconstruction-only patch training. To preserve spatial detail during semantic alignment, we also reconstruct the anchor patch: after quantization and decoding, $\hat{x}_{\text{in}}$ is supervised with the Stage-1 reconstruction loss $\mathcal{L}_{\text{rec}}$. The Stage-2 objective, We set $\alpha = 1.8$, $\beta = 0.6$, and $\gamma = 1.5$ empirically.

$$\mathcal{L}_{\text{Stage-2}} = \mathcal{L}_{\text{rec}} + \alpha\mathcal{L}_{\text{semantic}} + \beta\mathcal{L}_{\text{ABSQ}} + \gamma\mathcal{L}_{\text{PCM}}. \tag{2}$$

4 Experimental Results

Details of the pretraining corpora, downstream tasks and datasets, and implementation details are provided in Appendix A, Appendix B, and Appendix C respectively.

4.1 Quantitative Results

3D Reconstruction: Table 1 compares **VOLT** against baselines. VQ-VAE (Razavi et al., 2019), designed for natural images, performs worst, highlighting the need for medical pretraining. PUMIT-3D (Luo et al., 2023) improves with medical data but struggles with global coherence. MedVAE-3D (Varma et al., 2025) achieves the highest PSNR (30.88), while MedITok-2D (Ma et al., 2025) leads perceptually (SSIM 89.02, LPIPS 11.56) but suffers inter-slice inconsistencies. **VOLT** achieves the best overall mean performance, while also exhibiting the lowest variability across runs, indicating that its reconstruction advantage is consistent rather than driven by a favorable single estimate. Stage-1 already surpasses all baselines, while Stage-2 improves PSNR by $+4.19$ over MedVAE-3D and SSIM/LPIPS by $+5.54/-3.75$ over MedITok-2D. We therefore select MedVAE and MedITok, the top baselines from Table 1, for downstream evaluation against **VOLT** to assess representation generalization across diverse tasks. Inference time in Table 1 is measured for $512 \times 512 \times 512$ patches using a single A100 GPU under identical settings.

Table 1: **Reconstruction results.** Mean $\pm$ standard deviation.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	Time ↓ (ms)
VQ-VAE	25.47±1.10	82.28±1.10	15.88±0.98	**700**
PUMIT-3D	28.82±0.92	85.37±0.92	13.07±0.78	736
MedITok-2D	29.92±1.33	89.02±1.63	11.56±1.23	907
MedVAE-3D	30.88±1.70	87.67±1.50	12.33±1.17	875
VOLT	**35.07±0.30**	**94.56±0.55**	**7.81±0.47**	891

3D Synthetic Data Generation: We evaluate learned codes as tokenizers for semantic-aware 3D CT synthesis on BTCV using diffusion models (TRACE (Shao et al., 2025), BTB3D (Hamamci et al., 2025)), an autoregressive model (MAGViT-V2 (Luo et al., 2024)),

and a flow-matching model (VicTr (Susladkar et al., 2025)), while fixing architectures, training schedules, and sampling. We vary only the tokenizer (MedITok-2D, MedVAE-3D, **VOLT**). As shown in Table 2, **VOLT** consistently performs best across all priors: mFVD drops of 400→312, 322→277, 302→222, and 278→219 for TRACE, BTB3D, MAGViT-V2, and VicTr, respectively. Compared with MedITok-2D and MedVAE-3D, the latter benefits from true 3D tokenization over slice-stacked designs but still trails **VOLT**.

Table 2: **Synthetic data generation** (mFVD ↓).

Tokenizer	**TRACE** Diffusion	**BTB3D** Diffusion	**MAGViT-V2** Autoregressive	**VicTr** Flow Matching
MedITok	400	322	302	278
MedVAE	340	295	–	232
VOLT	**312**	**277**	**222**	**219**

3D Segmentation: In Table 3, **VOLT** achieves the best BTCV transfer, outperforming both fully supervised specialists and prior frozen FMs. With decoder-only training, VOLT achieves 84.87 Dice/20.02 HD95, surpassing the strongest baseline (MedVAE) and the fully supervised nnFormer. Task-specific baselines (TransUNet, nnUNet, nnFormer) lag behind because they train only on BTCV and suffer architectural limitations, including nnUNet's local CNN bias and nnFormer's attention-heavy skip connections that blur fine boundaries. Among frozen FMs, MedVAE-3D performs strongest due to its continuous 3D latent preserving intensity fidelity, while MedITok remains competitive but can miss subtle edges from discrete tokenization despite semantic supervision. UNITER captures global context but lacks spatial inductive bias for precise boundaries. VOLT preserves patch consistency and fine boundary cues, critical for subtle, overlapping BTCV organs. Its codebook-free discrete space avoids lossy spatial resizing, retaining fine-grained clinical detail. All VAEs share an identical training setup.

Table 3: **Segmentation results on BTCV**.

Methods	**Dice (↑)**	**HD95 (↓)**
Transunet (Chen et al., 2021)	71.21	28.77
nnUNet (Isensee et al., 2018)	83.32	26.63
nnFormer (Zhou et al., 2023)	84.57	25.32
MedITok (Ma et al., 2025)	76.32	23.17
UNITER (Shaker et al., 2024)	76.77	23.44
MedVAE-3D (Varma et al., 2025)	78.57	22.49
VOLT	**84.87**	**20.02**

Table 4: **Performance on 3D medical vision-language tasks.** Left: report generation on M3D-Cap. Right: close-ended VQA on M3D-VQA. Higher is better for all metrics.

Method	**M3D-Cap report generation**				**M3D-VQA closed-ended VQA**					
	BLEU	*ROUGE*	*METEOR*	*BERT Score*	*Plane*	*Phase*	*Organ*	*Abnormality*	*Location*	*Mean*
RadFM (Wu et al., 2025)	12.23	16.49	11.57	87.93	19.65	28.70	16.80	18.92	14.88	19.79
M3D-LaMed (Bai et al., 2024)	15.15	19.55	14.38	88.46	98.80	79.75	74.75	66.65	58.94	75.78
Med3DVLM (Xin et al., 2025)	36.88	40.25	36.42	88.11	**99.15**	87.50	77.45	70.17	64.49	79.75
MedITok (Ma et al., 2025)	35.42	41.47	37.48	87.92	98.23	86.23	75.92	66.54	65.75	77.77
MedVAE (Varma et al., 2025)	36.89	43.03	39.29	88.63	97.33	88.29	77.89	69.45	66.89	79.92
VOLT	**38.54**	**43.17**	**40.12**	**89.09**	98.45	**90.15**	**79.41**	**71.44**	**67.55**	**81.39**

3D Close-ended Visual Question Answering: Table 4 reports close-ended M3D-VQA (Bai et al., 2024) VQA across five categories (plane, phase, organ, abnormality, location). **VOLT**+MedGemma1.5 (Sellergren et al., 2025) achieves the best mean accuracy (81.39%), outperforming Med3DVLM (79.75%) and M3D-LaMed (75.78%) by +1.64% and +5.61%. Under the same VLM, **VOLT** improves over MedITok and MedVAE by +3.62% and 1.47%, indicating more informative volumetric features. Category-wise, **VOLT** leads on phase, organ, abnormality, and location while remaining competitive on plane. Compared to MedVAE+MedGemma1.5, largest gains are on organ, abnormality, and phase, suggesting improved semantic separability for fine-grained anatomical and pathological concepts.

Radiology Report Generation: Table 4 reports report generation under a fixed language backbone (MedGemma1.5), enabling controlled comparison of visual representations. **VOLT** performs best across all metrics. Relative to the strongest prior baseline (Med3DVLM), gains are +1.66 BLEU, +2.92 ROUGE, +3.70 METEOR, and +0.98 BERTScore indicating improved lexical precision, content coverage, and semantic fidelity. Margins over MedITok are larger (+3.12 BLEU, +1.70 ROUGE, +2.64 METEOR, +1.17 BERTScore). Scan-text and generic FM baselines (M3D-LaMed, RadFM) trail substantially, highlighting the importance of robust volumetric modeling and semantically structured 3D features.

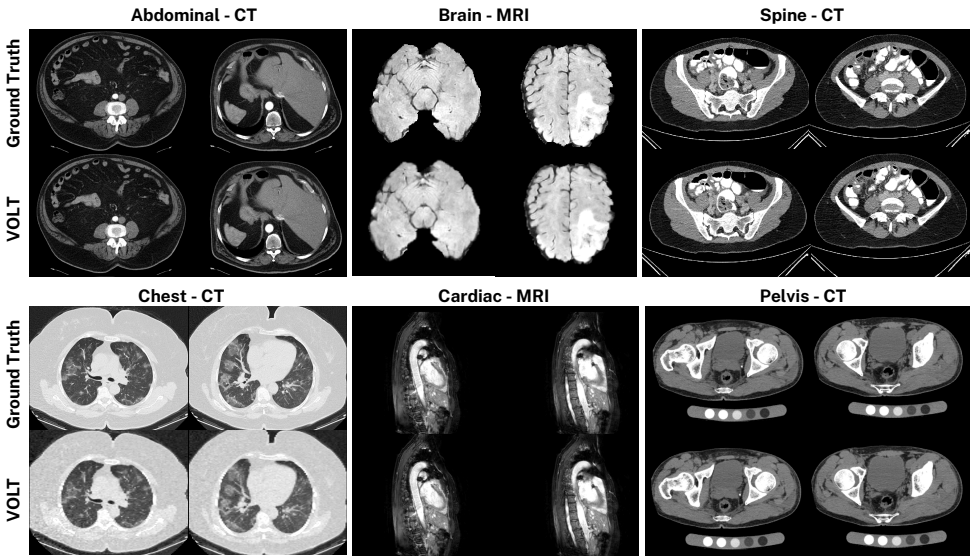


Figure 3: **3D reconstruction across anatomies and modalities. VOLT** preserves global anatomy, organ boundaries, and modality-specific contrast, demonstrating consistent volumetric fidelity.

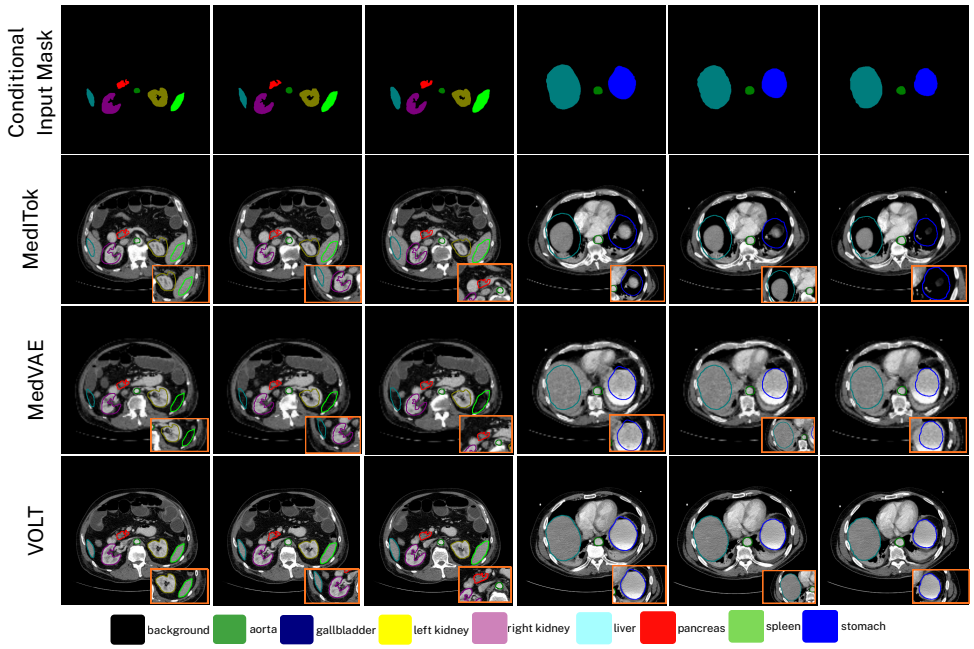


Figure 4: **Controllable 3D synthetic data generation.** Given input masks, **VOLT** better preserves mask boundaries, organ interfaces, and background consistency than prior methods.

4.2 Qualitative Results

3D Reconstruction: Qualitative results (Fig. 3) show that **VOLT** preserves global anatomical structure and fine-grained intensity detail across diverse anatomies (Abdominal-CT, Brain-MRI, Spine-CT, Chest-CT, Heart-MRI, Pelvis-CT). Reconstructions closely match ground-truth in organ shape, tissue contrast, and boundary sharpness, with MRI cases retaining coherent contrast without texture distortion or intensity drift. Consistent performance across anatomically and acquisitionally diverse scans suggests that **VOLT** learns reusable volumetric anatomical priors rather than overfitting to a single organ or modality, a key requirement for a general-purpose 3D medical FM. Additional comparisons across baselines and anatomical planes (axial/coronal/sagittal) are provided in Appendix G, and clinical validation of reconstructed volumes by two radiologists is reported in Appendix E.

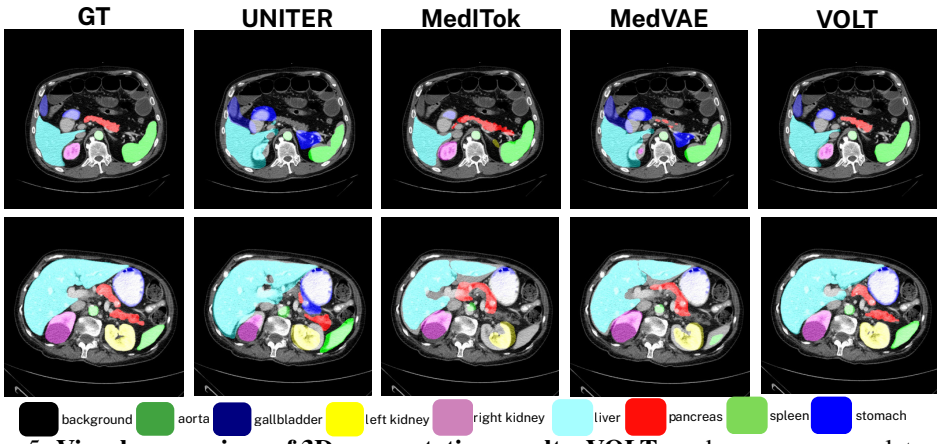


Figure 5: **Visual comparison of 3D segmentation results. VOLT** produces more complete regions and anatomically coherent boundaries with fewer spurious artifacts.

3D Synthetic Data Generation: Fig. 4 compares mask-conditioned 3D CT synthesis, with the top row showing input masks and subsequent rows showing generated volumes from MedITok, MedVAE, and **VOLT** under identical conditioning. MedITok leaks generation outside masked regions, causing surrounding texture and intensity drift. MedVAE better preserves background but produces blurred, under-conditioned content inside masks with less crisp boundaries. **VOLT** best matches intended behavior: generated anatomy adheres closely to mask locations while unmasked context remains stable. This manifests as (i) tighter alignment to conditional shapes, (ii) cleaner organ interfaces with fewer boundary drifts, and (iii) fewer global style shifts, indicating more controllable and volumetrically consistent conditioning for downstream priors.

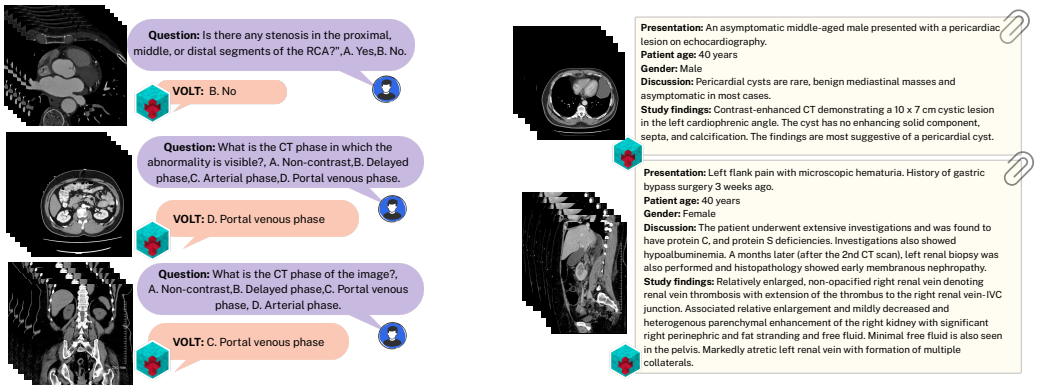
3D Segmentation: As shown in Fig. 5, on BTCV, **VOLT** predictions match GT more closely in region completeness and boundary sharpness. UNITER exhibits noticeable label noise and fragmentation around central vessels and soft tissue, with occasional spillover into surrounding tissue. MedITok produces cleaner regions but retains small boundary offsets at organ interfaces. MedVAE captures major organs well yet shows localized over/under-segmentation near the liver and vessel regions. In contrast, **VOLT** yields more coherent masks with smoother, anatomically plausible contours and fewer spurious islands; particularly around closely packed structures.

3D Visual Question Answering (Closed-ended): Figure 6(a) presents qualitative closed-ended VQA examples from M3DVQA, where the model must answer MCQs that require (i) recognizing pathology presence/absence and (ii) identifying acquisition/contrast phase from volumetric CT context. Across the shown cases, our model's predictions exactly match the ground-truth answers, correctly identifying no RCA stenosis in the cardiac CTA example, and correctly predicting the portal venous phase in two separate phase-recognition questions (one asking the phase where the abnormality is visible, and one asking the scan phase directly). Qualitatively, these examples highlight that the learned 3D representation retains both fine diagnostic cues and global contextual cues.

Table 5: **Ablations based on architectural choices and vector quantization methods.**

Experiments	PSNR (↑)	SSIM (↑)	LPIPS (↓)
Main Baseline (Ours with Two Stage Pre-trainng and ABSQ Module)	35.07	94.56	7.81
Depth Shuffle Pre-training Objective			
Ours with Augmentations and Only Stage-1 Pre-training	30.09	89.1	11.03
Ours with Masking and Only Stage-1 Pre-training	32.57	90.75	10.15
Ours with Depth Shuffle and Only Stage-1 Pre-training	32.98	91.27	9.72
Stage-2 Pre-training			
Ours with Masking and Two Stage Pre-training	34.68	93.98	8.02
Vector Quantizer			
VQ(Van Den Oord et al., 2017)	30.31	91.08	10.43
BSQ(Zhao et al., 2024)	32.47	91.78	9.19
LFQ(Yu et al., 2023)	31.97	90.11	9.87
RVQ(Chen et al., 2010)	33.23	93.17	9.43
Additive Remainder			
LFQ with Proposed Additive Remainder Idea	34.11	94.04	8.82

3D Report Generation: Figure 6(b) shows qualitative report generation examples, where **VOLT** produces clinically structured text consistent across presentation, discussion, and findings. In the first case, the report correctly localizes the lesion to the left cardiophrenic angle with a concrete size estimate and discriminative negative descriptors. In the second, it captures a complex vascular finding with correct laterality and extent. Overall, **VOLT** generates coherent long-form reports that capture key findings and clinically relevant negatives.



(a) Closed-Ended VQA (b) Report Generation

Figure 6: **3D VQA and Report Generation. VOLT** supports anatomically grounded VQA and clinically consistent reports.

4.3 Ablation Studies

Table 5 validates the key design choices in **VOLT**, grouped by (i) the Stage-1 corruption objective, (ii) the contribution of Stage-2, (iii) the vector quantizer, and (iv) the additive remainder mechanism.

Importance of Stage-1 Depth Shuffle: When Stage-1 is trained with generic augmentations instead, performance drops sharply. Augmentations alter appearance but do not create an explicit supervisory challenge tied to through-plane anatomy ordering, so the encoder learns weaker volumetric organization. Additionally, replacing our Depth Shuffle with MIM reduces performance. This proves masking does not explicitly pressure the encoder to encode through-plane ordering and long-range volumetric structure.

Importance of Stage-2 pretraining: To isolate Stage-2's effect, we applied the Stage-2 pipeline to an MIM-trained Stage-1 model. Even without Depth Shuffle, Stage-2 improves reconstruction, confirming its effectiveness. The full model (Depth Shuffle + Stage-2) remains best, showing complementary gains: Depth Shuffle strengthens volumetric priors while Stage-2 refines the representation.

Vector quantizer ablation: We replace ABSQ with alternative quantizers under identical training. Classic VQ performs worst, as hard codebook assignments introduce quantization noise. BSQ improves stability through codebook-free binarization but remains capacity-limited for high-fidelity 3D volumes. RVQ achieves the best PSNR/SSIM among alternatives but worse LPIPS than BSQ, suggesting that multi-stage codebook quantization preserves average intensity while causing perceptual artifacts (*e.g.*, texture/edge distortions). More broadly, codebook-based methods require parametric lookup tables and are susceptible to collapse under distributional shift, whereas codebook-free methods (BSQ, ABSQ) avoid these failure modes and scale better to 3D medical data.

Importance of additive remainder: Additive residual refinement significantly improves lookup-free discretization by reducing quantization distortion relative to a single-shot bottleneck. This also explains BSQ's weaker standalone performance. Within ABSQ, additive residual capacity substantially improves BSQ by preserving fine anatomical boundaries and perceptual quality.

5 Conclusion

We introduce **VOLT**, a 3D foundation model that learns reusable, clinically meaningful discrete volumetric representations from scan-only medical data. VOLT's latent space preserves through-plane anatomical structure, maintains global patch coherence, and aligns volumes with clinical concepts without paired scan-text supervision. These frozen features support accurate boundary-aware segmentation, coherent token sequences for controllable 3D synthesis, and semantics that language models can reliably decode for volumetric VQA. Overall, our results show that semantically structured 3D medical representations can be learned from diverse scan-only data, providing a practical foundation for volumetric generation and vision-language applications.

AI USE STATEMENT

We used large language models (LLMs) solely to polish the writing of this paper, including improving sentence clarity and correcting grammar. We also used an AI coding agent to help optimize our implementation and accelerate the execution of experiments. All research ideas, methods, experimental designs, and analyses are our own, and we carefully reviewed all AI-assisted text and code. The authors take full responsibility for the content of this paper.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide complete implementation details, including all training and evaluation hyperparameters, in Appendix C. Together, these materials are sufficient to replicate our results. In line with our commitment to open research, we will publicly release our code, data and model weights upon publication.

ETHICS STATEMENT

This study used publicly available, de-identified medical imaging data set comprising MRI and CT volumes. We compiled with the applicable data set license and data use requirements and did not attempt to identify any individuals. No new participant recruitment or data collection was conducted.

REFERENCES

Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. The medical segmentation decathlon. *Nature communications*, 13(1):4128, 2022.

Fan Bai, Yuxin Du, Tiejun Huang, Max Q-H Meng, and Bo Zhao. M3d: Advancing 3d medical image analysis with multi-modal large language models. *arXiv preprint arXiv:2404.00578*, 2024.

Patrick Bilic, Patrick Christ, Hongwei Bran Li, Eugene Vorontsov, Avi Ben-Cohen, Georgios Kaissis, Adi Szeskin, Colin Jacobs, Gabriel Efrain Humpire Mamani, Gabriel Chartrand, et al. The liver tumor segmentation benchmark (lits). *Medical image analysis*, 84:102680, 2023.

Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, et al. Making the most of text semantics to improve biomedical vision–language processing. In *European conference on computer vision*, pp. 1–21. Springer, 2022.

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*, pp. 9650–9660, 2021.

Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

Sihong Chen, Kai Ma, and Yefeng Zheng. Med3d: Transfer learning for 3d medical image analysis. *arXiv preprint arXiv:1904.00625*, 2019.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020.

Yinda Chen, Che Liu, Xiaoyu Liu, Rossella Arcucci, and Zhiwei Xiong. Bimcv-r: A landmark dataset for 3d ct text-image retrieval. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 124–134. Springer, 2024a.

Yongjian Chen, Tao Guan, and Cheng Wang. Approximate nearest neighbor search by residual vector quantization. *Sensors*, 10(12):11259–11273, 2010.

Zekai Chen, Devansh Agarwal, Kshitij Aggarwal, Wiem Safta, Mariann Micsinai Balan, and Kevin Brown. Masked image modeling advances 3d medical image analysis. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1970–1980, 2023.

Zhihong Chen, Maya Varma, Jean-Benoit Delbrouck, Magdalini Paschali, Louis Blankemeier, Dave Van Veen, Jeya Maria Jose Valanarasu, Alaa Youssef, Joseph Paul Cohen, Eduardo Pontes Reis, et al. Chexagent: Towards a foundation model for chest x-ray interpretation. In *AAAI 2024 Spring Symposium on Clinical Foundation Models*, 2024b.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12873–12883, 2021.

Ibrahim Ethem Hamamci, Sezgin Er, Furkan Almas, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Irem Dogan, Muhammed Furkan Dasdelen, Omer Faruk Durugol, Bastian Wittmann, Tamaz Amiranashvili, Enis Simsar, Mehmet Simsar, Emine Bensu Erdemir, Abdullah Alanbay, Anjany Sekuboyina, Berkan Lafci, Christian Bluethgen, Mehmet Kemal Ozdemir, and Bjoern Menze. Developing generalist foundation models from a multimodal dataset for 3d computed tomography, 2024.

Ibrahim Ethem Hamamci, Sezgin Er, Suprosanna Shit, Hadrien Reynaud, Dong Yang, Pengfei Guo, Marc Edgar, Daguang Xu, Bernhard Kainz, and Bjoern Menze. Better tokens for better 3d: Advancing vision-language modeling in 3d medical imaging. *arXiv preprint arXiv:2510.20639*, 2025.

Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 574–584, 2022.

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*, pp. 9729–9738, 2020.

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*, pp. 16000–16009, 2022.

Nicholas Heller, Fabian Isensee, Dasha Trofimova, Resha Tejpaul, Zhongchen Zhao, Huai Chen, Lisheng Wang, Alex Golts, Daniel Khapun, Daniel Shats, et al. The kits21 challenge: Automatic segmentation of kidneys, renal tumors, and renal cysts in corticomedullary-phase ct. *arXiv preprint arXiv:2307.01984*, 2023.

Shih-Cheng Huang, Liyue Shen, Matthew P Lungren, and Serena Yeung. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3942–3951, 2021.

Fabian Isensee, Jens Petersen, Andre Klein, David Zimmerer, Paul F Jaeger, Simon Kohl, Jakob Wasserthal, Gregor Koehler, Tobias Norajitra, Sebastian Wirkert, et al. nnu-net: Self-adapting framework for u-net-based medical image segmentation. *arXiv preprint arXiv:1809.10486*, 2018.

Debesh Jha, Onkar Kishor Susladkar, Vandan Gorade, Elif Keles, Matthew Antalek, Deniz Seyithanoglu, Timurhan Cebeci, Halil Ertugrul Aktas, Gulbiz Dagoglu Kartal, Sabahattin Kaymakoglu, et al. Large scale mri collection and segmentation of cirrhotic liver. *Scientific Data*, 12 (1):896, 2025.

Yuanfeng Ji, Haotian Bai, Chongjian Ge, Jie Yang, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhanng, Wanling Ma, Xiang Wan, et al. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. *Advances in neural information processing systems*, 35:36722–36732, 2022.

Yuanyuan Jia, Zhongshi He, Ali Gholipour, and Simon K Warfield. Single anisotropic 3-d mr image upsampling via overcomplete dictionary trained from in-plane high resolution slices. *IEEE journal of biomedical and health informatics*, 20(6):1552–1561, 2015.

Anahita Fathi Kazerooni, Nastaran Khalili, Xinyang Liu, Debanjan Haldar, Zhifan Jiang, Anna Zapaishchykova, Julija Pavaine, Lubdha M Shah, Blaise V Jones, Nakul Sheth, et al. Brats-peds: Results of the multi-consortium international pediatric brain tumor segmentation challenge 2023. *arXiv preprint arXiv:2407.08855*, 2024.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

Haoran Lai, Zihang Jiang, Qingsong Yao, Rongsheng Wang, Zhiyang He, Xiaodong Tao, Wei Wei, Weifu Lv, and S Kevin Zhou. E3d-gpt: enhanced 3d visual foundation for medical vision-language model. *arXiv preprint arXiv:2410.14200*, 2024.

Bennett Landman, Zhoubing Xu, Juan Igelsias, Martin Styner, Thomas Langerak, and Arno Klein. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In *Proc. MICCAI multi-atlas labeling beyond cranial vault—workshop challenge*, 2015.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36: 28541–28564, 2023.

Wenxuan Li, Chongyu Qu, Xiaoxi Chen, Pedro RAS Bassi, Yijia Shi, Yuxiang Lai, Qian Yu, Huimin Xue, Yixiong Chen, Xiaorui Lin, et al. Abdomenatlas: A large-scale, detailed-annotated, & multi-center dataset for efficient transfer learning and open algorithmic benchmarking. *Medical Image Analysis*, 97:103285, 2024.

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.

Han Liu, Bogdan Georgescu, Yanbo Zhang, Youngjin Yoo, Michael Baumgartner, Riqiang Gao, Jianing Wang, Gengyan Zhao, Eli Gibson, Dorin Comaniciu, et al. Revisiting 2d foundation models for scalable 3d medical image classification. *arXiv preprint arXiv:2512.12887*, 2025a.

Peirong Liu, Oula Puonti, Xiaoling Hu, Karthik Gopinath, Annabel Sorby-Adams, Daniel C Alexander, W Taylor Kimberly, and Juan E Iglesias. A modality-agnostic multi-task foundation model for human brain imaging. *arXiv preprint arXiv:2509.00549*, 2025b.

Pengbo Liu, Hu Han, Yuanqi Du, Heqin Zhu, Yinhao Li, Feng Gu, Honghu Xiao, Jun Li, Chunpeng Zhao, Li Xiao, et al. Deep learning to segment pelvic bones: large-scale ct datasets and baseline models. *International Journal of Computer Assisted Radiology and Surgery*, 16(5):749–756, 2021.

Gabriele Lozupone, Alessandro Bria, Francesco Fontanella, Frederick JA Meijer, Claudio De Stefano, and Henkjan Huisman. Latent diffusion autoencoders: Toward efficient and meaningful unsupervised representation learning in medical imaging. *arXiv preprint arXiv:2504.08635*, 2025.

Lingxiao Luo, Xuanzhong Chen, Bingda Tang, Xinsheng Chen, Rong Han, Chengpeng Hu, Yujiang Li, and Ting Chen. Building universal foundation models for medical image analysis with spatially adaptive networks. *arXiv preprint arXiv:2312.07630*, 2023.

Zhuoyan Luo, Fengyuan Shi, Yixiao Ge, Yujiu Yang, Limin Wang, and Ying Shan. Open-magvit2: An open-source project toward democratizing auto-regressive visual generation. *arXiv preprint arXiv:2409.04410*, 2024.

Chenglong Ma, Yuanfeng Ji, Jin Ye, Zilong Li, Chenhui Wang, Junzhi Ning, Wei Li, Lihao Liu, Qiushan Guo, Tianbin Li, et al. Meditok: A unified tokenizer for medical image synthesis and interpretation. *arXiv preprint arXiv:2505.19225*, 2025.

Jun Ma, Yao Zhang, Song Gu, Cheng Zhu, Cheng Ge, Yichi Zhang, Xingle An, Congcong Wang, Qiyuan Wang, Xin Liu, et al. Abdomenct-1k: Is abdominal organ segmentation a solved problem? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6695–6714, 2021.

Raghav Mehta, Angelos Filos, Ujjwal Baid, Chiharu Sako, Richard McKinley, Michael Rebsamen, Katrin Dätwyler, Raphael Meier, Piotr Radojewski, Gowtham Krishnan Murugesan, et al. Qu-brats: Miccai brats 2020 challenge on quantifying uncertainty in brain tumor segmentation-analysis of ranking scores and benchmarking results. *The journal of machine learning for biomedical imaging*, 2022:https–www, 2022.

Adriënne M Mendrik, Koen L Vincken, Hugo J Kuijf, Marcel Breeuwer, Willem H Bouvy, Jeroen De Bresser, Amir Alansary, Marleen De Bruijne, Aaron Carass, Ayman El-Baz, et al. Mrbrains challenge: online evaluation framework for brain image segmentation in 3t mri scans. *Computational intelligence and neuroscience*, 2015(1):813696, 2015.

André Pedersen, Erik Smistad, Tor V Rise, Vibeke G Dale, Henrik S Pettersen, Tor-Arne S Nordmo, David Bouget, Ingerid Reinertsen, and Marit Valla. H2g-net: A multi-resolution refinement approach for segmentation of breast cancer region in gigapixel histopathological images. *Frontiers in medicine*, 9:971873, 2022.

Walter HL Pinaya, Petru-Daniel Tudosiu, Jessica Dafflon, Pedro F Da Costa, Virginia Fernandez, Parashkev Nachev, Sebastien Ourselin, and M Jorge Cardoso. Brain imaging generation with latent diffusion models. In *MICCAI workshop on deep generative models*, pp. 117–126. Springer, 2022a.

Walter HL Pinaya, Petru-Daniel Tudosiu, Robert Gray, Geraint Rees, Parashkev Nachev, Sebastien Ourselin, and M Jorge Cardoso. Unsupervised brain imaging 3d anomaly detection and segmentation with transformers. *Medical Image Analysis*, 79:102475, 2022b.

Tushar Prakash, Onkar Susladkar, Sparsh Mittal, and Inderjit S Dhillon. Pyramidal spectrum: Frequency-based hierarchically vector quantized vae for videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2073–2082, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019.

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*, pp. 10684–10695, 2022.

Ji Seung Ryu, Hyunyoung Kang, Yuseong Chu, and Sejung Yang. Vision-language foundation models for medical imaging: a review of current practices and innovations. *Biomedical Engineering Letters*, 15(5):809–830, 2025.

Anjany Sekuboyina, Malek E Husseini, Amirhossein Bayat, Maximilian Löffler, Hans Liebl, Hongwei Li, Giles Tetteh, Jan Kukačka, Christian Payer, Darko Štern, et al. Verse: a vertebrae labelling and segmentation benchmark for multi-detector ct images. *Medical image analysis*, 73:102166, 2021.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

Abdelrahman Shaker, Muhammad Maaz, Hanoona Rasheed, Salman Khan, Ming-Hsuan Yang, and Fahad Khan. Unetr plus plus: Delving into efficient and accurate 3d medical image segmentation. *IEEE Transactions on Medical Imaging*, 2024.

Minye Shao, Xingyu Miao, Haoran Duan, Zeyu Wang, Jingkun Chen, Yawen Huang, Xian Wu, Jingjing Deng, Yang Long, and Yefeng Zheng. Trace: Temporally reliable anatomically-conditioned 3d ct generation with enhanced efficiency. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 627–637. Springer, 2025.

Yanzhao Shi, Xiaodan Zhang, Junzhong Ji, Haoning Jiang, Chengxin Zheng, Yinong Wang, and Liangqiong Qu. Hsenet: Hybrid spatial encoding network for 3d medical vision-language understanding. *arXiv preprint arXiv:2506.09634*, 2025.

Onkar Susladkar, Jishu Sen Gupta, Chirag Sehgal, Sparsh Mittal, and Rekha Singhal. Motionaura: Generating high-quality and motion consistent videos using discrete diffusion. *arXiv preprint arXiv:2410.07659*, 2024.

Onkar Susladkar, Gayatri Deshmukh, Yalcin Tur, Gorkem Durak, and Ulas Bagci. Victr: Vital consistency transfer for pathology aware image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22772–22782, 2025.

Aggeliki Tsoli, Antonis Argyros, et al. Patch-based reconstruction of a textureless deformable 3d surface from a single rgb image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 0–0, 2019.

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

Maya Varma, Ashwin Kumar, Rogier Van der Sluijs, Sophie Ostmeier, Louis Blankemeier, Pierre Chambon, Christian Bluethgen, Jip Prince, Curtis Langlotz, and Akshay Chaudhari. Medvae: Efficient automated interpretation of medical images with large-scale generalizable autoencoders. *arXiv preprint arXiv:2502.14753*, 2025.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Haoshen Wang, Zhentao Liu, Kaicong Sun, Xiaodong Wang, Dinggang Shen, and Zhiming Cui. 3d meddiffusion: A 3d medical latent diffusion model for controllable and high-quality medical image generation. *IEEE Transactions on Medical Imaging*, 2025.

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2022, pp. 3876, 2022.

Jakob Wasserthal, Hanns-Christian Breit, Manfred T Meyer, Maurice Pradella, Daniel Hinck, Alexander W Sauter, Tobias Heye, Daniel T Boll, Joshy Cyriac, Shan Yang, et al. Totalsegmentator: robust segmentation of 104 anatomic structures in ct images. *Radiology: Artificial Intelligence*, 5(5):e230024, 2023.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Hui Hui, Yanfeng Wang, and Weidi Xie. Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. *Nature Communications*, 2025.

Yu Xin, Gorkem Can Ates, Kuang Gong, and Wei Shao. Med3dvlm: An efficient vision-language model for 3d medical image analysis. *IEEE Journal of Biomedical and Health Informatics*, 2025.

Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021.

Lijun Yu, José Lezama, Nitesh B Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Vighnesh Birodkar, Agrim Gupta, Xiuye Gu, et al. Language model beats diffusion–tokenizer is key to visual generation. *arXiv preprint arXiv:2310.05737*, 2023.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Machine learning for healthcare conference*, pp. 2–25. PMLR, 2022.

Yue Zhao, Yuanjun Xiong, and Philipp Krähenbühl. Image and video tokenization with binary spherical quantization. *arXiv preprint arXiv:2406.07548*, 2024.

Hong-Yu Zhou, Jiansen Guo, Yinghao Zhang, Xiaoguang Han, Lequan Yu, Liansheng Wang, and Yizhou Yu. nnformer: Volumetric medical image segmentation via a 3d transformer. *IEEE transactions on image processing*, 32:4036–4045, 2023.

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*, pp. 384–393. Springer, 2019.

Xiahai Zhuang, Lei Li, Christian Payer, Darko Štern, Martin Urschler, Mattias P Heinrich, Julien Oster, Chunliang Wang, Örjan Smedby, Cheng Bian, et al. Evaluation of algorithms for multi-modality whole heart segmentation: an open-access grand challenge. *Medical image analysis*, 58:101537, 2019.

Appendix

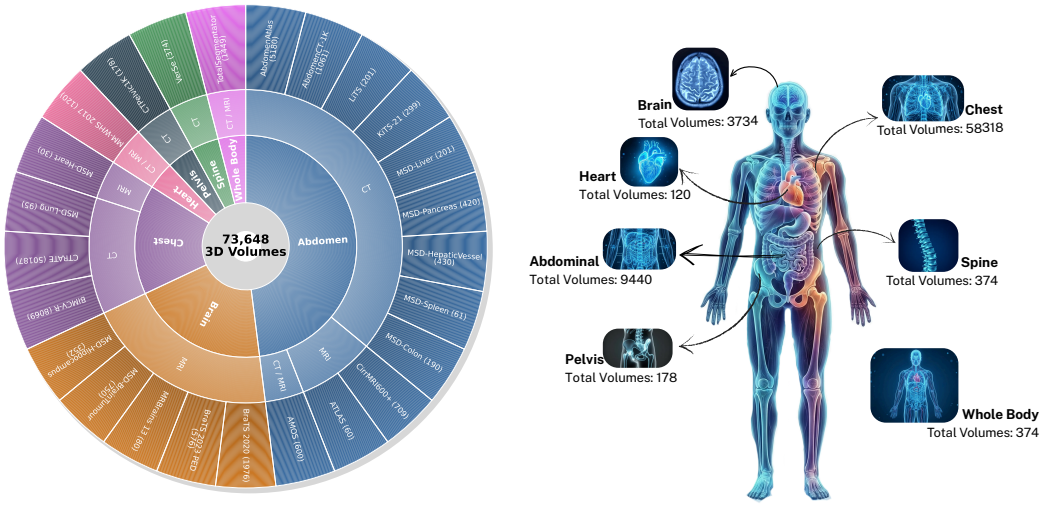


Figure 7: **VOLT pretraining corpus.** Distribution of 73,648 3D volumes, spanning seven anatomical regions: brain, chest, abdominal, spine, heart, pelvis, and whole-body curated from 25 publicly available CT and MRI datasets.

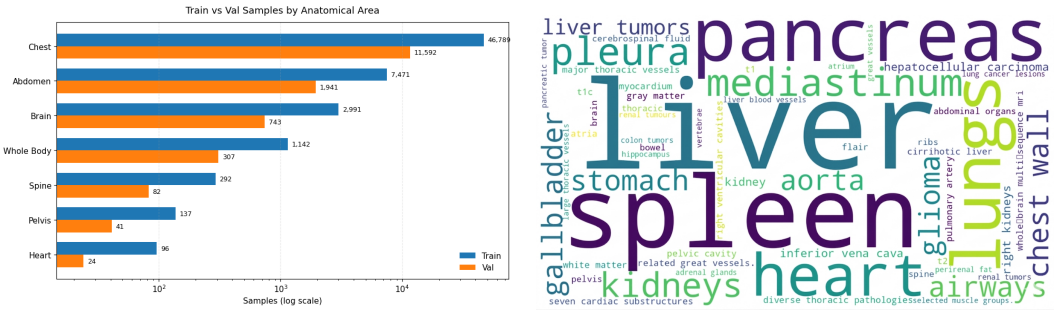
A Pretraining Corpus

To pre-train **VOLT**, we construct a large-scale curated corpus by aggregating publicly available 3D CT and MRI datasets spanning multiple anatomies, pathologies, and acquisition protocols. The goal is to expose the model to diverse real-world heterogeneity, including scanner/protocol differences, reconstruction settings, contrast usage, field-of-view, anisotropic spacing, and dataset-specific conventions, so that learned representations transfer robustly across downstream institutions and tasks. The curated corpus contains 73,648 3D volumes, split into 80% training and 20% validation volumes. All train, validation, and test splits were constructed with strict patient-level separation, ensuring that volumes from the same patient never appeared in more than one split and preventing data leakage across all reported experiments.

A.1 Diversity and Anatomical Coverage

Rather than focusing on a single anatomy, our pre-training corpus spans a broad range of body regions, including the *abdomen*, *thorax/chest*, *whole body*, *brain*, *pelvis*, *heart*, and *spine*. This anatomical diversity exposes the model to substantial inter-organ and inter-protocol variability (*e.g.* differences in field-of-view, contrast usage, reconstruction characteristics, and spatial resolution), which is critical for learning transferable 3D representations.

Fig. 7 provides an at-a-glance summary of the curated pre-training corpus, showing the hierarchical composition by *anatomical area* $\rightarrow$ *modality* $\rightarrow$ *dataset source*, along with the number of 3D volumes contributed by each dataset (totaling 73,648 volumes). The corpus is assembled from 25 publicly released datasets (Mendrik et al., 2015; Li et al., 2024; Antonelli et al., 2022; Ma et al., 2021; Sekuboyina et al., 2021; Chen et al., 2024a; Ji et al., 2022; Mehta et al., 2022; Heller et al., 2023; Kazerooni et al., 2024; Liu et al., 2021; Hamamci et al., 2024; Jha et al., 2025; Bilic et al., 2023; Zhuang et al., 2019; Wasserthal et al., 2023). In addition to highlighting the breadth across CT and MRI, the sunburst visualization makes the long-tail nature of medical data explicit: large-scale sources (e.g., chest CT collections) coexist with smaller, anatomy-specific datasets, together yielding a clinically heterogeneous training distribution. Fig. 8a (a) further reports the distribution of samples across anatomical areas separately for the training and validation splits. Finally, Fig. 8b (b) provides a qualitative view of the anatomical breadth by summarizing frequently occurring organ and region-level terms (*e.g.*liver, spleen, pancreas, mediastinum, etc.), reflecting the multi-organ and multi-region nature of the curated pre-training data.



(a) Train vs. validation sample distribution across anatomical regions (log scale).

(b) Word cloud of anatomical terms showing organ coverage.

Figure 8: **Dataset statistics and anatomical coverage**

A.2 Dataset Heterogeneity

As shown in Fig. 9, the curated pre-training corpus exhibits substantial variability in voxel spacing, *i.e.* the physical distance represented by each voxel along the three spatial axes of a 3D scan. This quantity determines the effective spatial resolution of the volume and directly affects how anatomical structures are sampled and represented. The figure shows that the in-plane spacings (dx, dy) are relatively concentrated, whereas the through-plane spacing (dz) spans a much broader range on a logarithmic scale. This indicates that the co rpus contains both near-isotropic high-resolution volumes and thick-slice acquisitions, reflecting the diversity of scanners, reconstruction settings, anatomical coverage, and imaging protocols encountered in real clinical data. Such heterogeneity is particularly important for pre-training a 3D medical foundation model: exposure to varying spatial resolutions and slice thicknesses encourages the model to learn representations that are less tied to a single acquisition regime and more robust across downstream datasets and tasks.

B Downstream Tasks & Datasets

We evaluate reconstruction quality on the validation split. For downstream evaluation, we benchmark **VOLT** on BTCV (Landman et al., 2015) for volumetric segmentation and synthetic data generation, and use M3D-VQA (Bai et al., 2024) for 3D visual question answering (VQA) and report generation. For classification, we use CirrMRI600+ (Jha et al., 2025) liver cirrhosis dataset, comprising MRI volumes with binary labels (healthy vs. cirrhotic/damaged liver).

C Implementation Details

Framework and distributed training: in PyTorch and trained with for synchronous multi-GPU optimization. Unless otherwise specified, we use mixed-precision training and standard gradient clipping when needed for stability. Stage-1 and Stage-2 pretraining are conducted on 4×8 NVIDIA A100 GPUs (80 GB each; 32 GPUs total). Downstream segmentation uses $8\times$A100 GPUs, while VQA, report generation, and synthesis are trained on 3×8 A100 GPUs (24 GPUs total).

Backbone initialization and volumetric adaptation: We initialize the encoder and decoder backbones from WAN2.2 VAE (Wan et al., 2025), which employs an overall compression of $\times16$ in-plane and $\times8$ along the temporal/depth axis. To better accommodate 3D medical volumes with higher through-plane redundancy, we extend this

All experiments are implemented `DistributedDataParallel` (DDP)

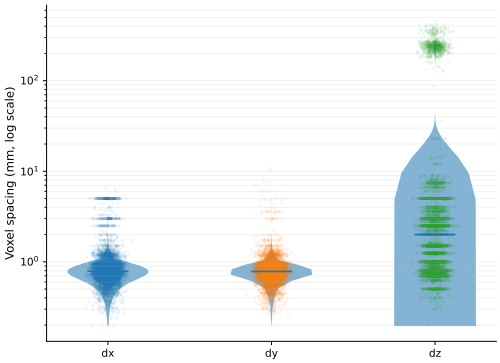


Figure 9: **Voxel spacing distribution of the pre-training corpus.**

backbone by adding one additional downsampling
and upsampling convolutional block in the depth pathway, resulting in a uniform compression factor of $\times 16$ along both spatial (H$\times$W) and depth (D) dimensions.

Volumetric patch-based training: To enable efficient training on high-resolution scans under memory constraints, we adopt patch-based volumetric training. For any volume with in-plane resolution exceeding 300×300, we randomly crop a sub-volume of size $256 \times 256 \times 32$ (H$\times$W$\times$D) and apply the corresponding objectives on the cropped region. For smaller volumes, we use center cropping or padding to match the training resolution. All methods are trained with the same patching strategy whenever applicable to ensure fair comparison.

Stage-1 optimization: Stage-1 is trained with per-GPU batch size 3 and gradient accumulation 4, yielding an effective batch size of 384. We optimize with AdamW using an initial learning rate of 1×10^{-4} and a cosine annealing schedule. Stage I is trained for 400 optimization steps.

Stage-2 optimization: Stage-2 initializes from the Stage I checkpoint and inserts LoRA modules into the VAE encoder and decoder with rank r=64. We train with per-GPU batch size 2 and gradient accumulation 2 (effective batch size 128). Optimization uses AdamW with initial learning rate 1×10^{-5} and cosine annealing scheduling. Unless stated otherwise, only LoRA parameters are updated during Stage II while the remaining backbone weights remain fixed.

Downstream tasks and fair baseline protocols: To ensure a controlled comparison across representations, we match the downstream training protocol across all tokenizers/backbones. In particular, we keep the pretrained VAE encoder *frozen* for **VOLT** as well as for the MedVAE and MedITok baselines and take encoder latent representation, and we keep optimizer type, learning-rate schedules, batch sizes, and the number of optimization steps identical whenever applicable.

Segmentation: For multi-class segmentation, we freeze the pretrained encoder and train a task-specific segmentation decoder/head. Training uses $8\times$A100 GPUs for 100 steps with per-GPU batch size 4, AdamW, initial learning rate 1×10^{-4}, and cosine annealing scheduling. The same protocol is applied to **VOLT**, MedITok, and MedVAE (with their respective frozen encoders) to isolate the effect of the learned representation.

VQA and report generation (VLM fine-tuning): We use MedGemma-1.5 as the medical VLM and fine-tune it with PEFT using LoRA (rank r=32) for MedGemma. Across all visual encoders (our **VOLT**, MedITok, and MedVAE), we keep the VLM architecture, LoRA configuration, and optimization settings fixed to attribute performance differences solely to the visual representation. We train on 3×8 A100 GPUs (24 GPUs total) with per-GPU batch size 4 and gradient accumulation 4 (effective batch size 384), using Adam with learning rate 1×10^{-5} and a linear learning-rate schedule.

Medical image synthesis: For medical image synthesis, we use the same compute and batching configuration as VQA/report generation. We optimize with Adam with learning rate 1×10^{-5} with a linear schedule, keeping all other hyperparameters consistent across tokenizers to ensure fair comparison. We use **VOLT** and other baseline as tokenized and dtokenized our 3D-Volumes for prior generative models like VicTr (Susladkar et al., 2025) , BTB3D (Hamamci et al., 2025) and MAGVITV2 (Yu et al., 2023)

Classification: To isolate the contribution of the tokenizer latent space, we adopt a linear-probing protocol for classification. Specifically, we freeze the entire pretrained encoder for all tokenizers and train only a lightweight MLP classification head on top of the fixed 16-dim per-volume embedding. This ensures that performance differences are attributable to the quality of the learned representations rather than differences in end-to-end fine-tuning capacity.

Inference-time volume processing: VOLT is fully volumetric and is not restricted to a fixed input volume size at inference. During inference, larger volumes are processed using a sliding-window strategy over the full volume, and window outputs are aggregated to reconstruct the full scan.

Table 6: **Open-Ended VQA Results.**

Method	Metric	Plane	Phase	Organ	Abnormality	Location	Mean
RadFM (Wu et al., 2025)	BLEU	14.24	14.25	14.24	15.64	23.58	16.39
	ROUGE	25.40	25.41	25.38	25.38	29.09	26.13
	METEOR	20.62	20.63	20.61	20.60	24.19	21.33
	BERTScore	92.68	92.04	86.79	85.84	86.26	88.72
M3D-LaMed (Bai et al., 2024)	BLEU	98.37	74.41	34.20	15.91	24.00	49.38
	ROUGE	98.42	78.63	37.87	19.27	27.74	52.39
	METEOR	49.20	63.58	23.78	12.83	18.50	33.58
	BERTScore	99.47	95.55	88.97	86.08	87.60	91.53
Med3DVLM (Xin et al., 2025)	BLEU	**98.85**	78.17	40.22	18.99	25.66	52.38
	ROUGE	**98.89**	84.20	45.22	23.27	29.99	56.31
	METEOR	**49.43**	68.50	29.32	16.21	20.32	36.76
	BERTScore	**99.83**	96.47	90.47	86.27	87.88	92.18
MediTok (Ma et al., 2025)	BLEU	97.32	79.34	39.82	19.25	24.31	52.01
	ROUGE	98.01	83.27	45.17	22.12	27.89	55.29
	METEOR	42.19	68.45	25.49	15.48	19.42	34.2
	BERTScore	97.79	95.32	89.45	85.39	86.58	90.91
MedVAE (Varma et al., 2025)	BLEU	98.22	80.11	41.44	19.95	26.19	53.18
	ROUGE	98.11	84.52	47.01	24.54	31.32	57.11
	METEOR	44.01	69.67	31.21	18.09	22.01	36.99
	BERTScore	97.56	96.92	91.23	87.88	88.45	92.41
VOLT	BLEU	98.45	**81.09**	**43.43**	**21.56**	**28.11**	**54.52**
	ROUGE	98.79	**85.78**	**48.35**	**27.52**	**33.33**	**58.75**
	METEOR	48.54	**71.45**	**31.39**	**20.23**	**24.38**	**39.19**
	BERTScore	98.34	**97.83**	**93.01**	**89.35**	**89.34**	**93.57**

D Extended Quantitative Comparison

D.1 Open-ended VQA Results

Table 6 reports open-ended M3DVQA performance broken down by question type, revealing a consistent split between orientation style queries (Plane) and clinically grounded queries (Phase / Organ / Abnormality / Location). RadFM is substantially lower across metrics, reflecting weaker specialization for fine-grained 3D VQA. Among all, Med3DVLM attains the strongest Plane scores (BLEU 98.85, BERTScore 99.83), which is expected because plane identification has a small answer space and strong global cues, so an end-to-end model can tightly fit the dataset's canonical phrasing and nearly saturate token-overlap metrics. In contrast, **VOLT** achieves the best overall mean (BLEU 54.52, ROUGE 58.75, METEOR 39.19, and BERTScore 93.57) and leads on clinically demanding categories, particularly Abnormality (BLEU 21.56 vs. MedVAE 19.95) and Location (BLEU 28.11 vs. MedVAE 26.19), where answers depend on semantically meaningful evidence and spatial grounding rather than global orientation alone. In such cases, **VOLT** consistently improves over MediTok and MedVAE, indicating that its discrete representation exposes more reliable cues for both contrast-phase reasoning and localized pathology/anatomy description. Notably, **VOLT** matches the "fully trained" systems while using a frozen visual backbone, suggesting the gains come from the quality and organization of the learned 3D representation rather than task-specific training.

D.2 Classification:

We evaluate the representation quality on the dataset CirrMRI600+ (Jha et al., 2025) by performing binary classification of 3D liver volumes into healthy and cirrhotic categories. For a fair comparison, we freeze the backbone for all methods and train only a linear classification head on top of the extracted features. As shown in Table 7, **VOLT** achieves the highest classification accuracy among the evaluated methods, suggesting that it learns more discriminative representations.

Table 7: **Classification results on CirrMRI600+.**

Model	Accuracy
MedVAE	72.88
MedITok	68.99
M3D-CLIP	73.20
VOLT	**76.09**

E Clinical Validation

To assess clinical fidelity of **VOLT** reconstructions, two radiologists independently evaluated 25 reconstructed 3D volumes spanning abdominal CT, abdominal MRI, chest CT, and brain scans by comparing **VOLT** outputs against ground truth scans. Each reconstruction was scored as either acceptable or unacceptable based on its adherence to the anatomical and pathological coherence of the original image. The experts found 22 of the 25 cases (88%) to be clinically acceptable, noting that the model produced anatomically coherent results with high-fidelity preservation of three-

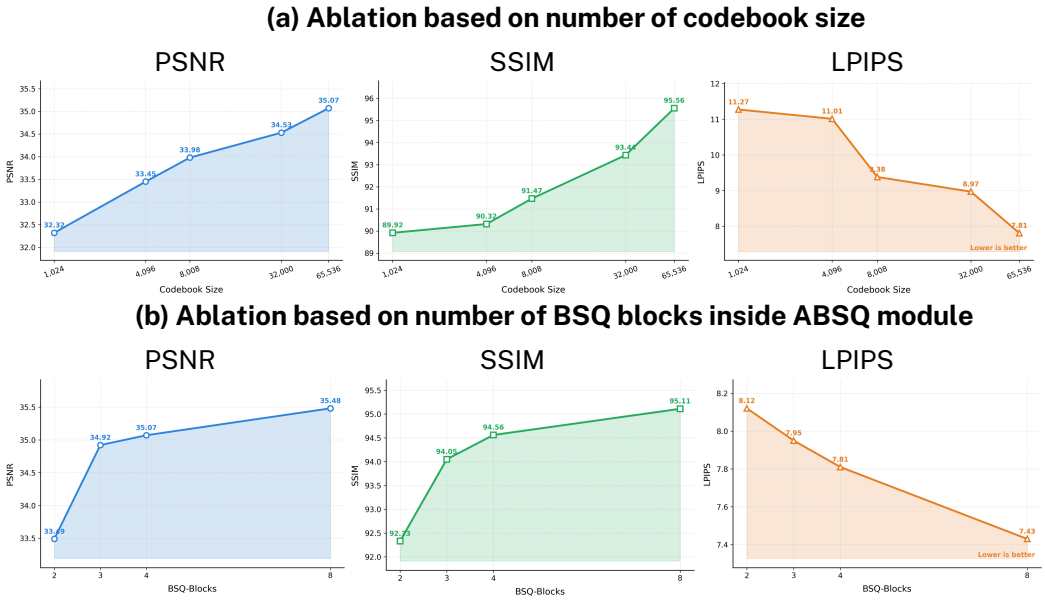


Figure 10: **Ablation based on number of codebook and BSQ blocks**

dimensional structures. Specifically, the model maintained strong through-plane consistency and realistic anatomical continuity across diverse regions, addressing a common limitation of existing 3D image generation approaches

F Ablation Study

Table 8 and Fig 10 examine two factors that directly control the quality of **VOLT** discrete volumetric representation: (i) which loss terms are necessary, and (ii) how much discrete capacity (codebook size) is required to reconstruct diverse 3D CT/MRI volumes faithfully. **Ablation to show importance of loss functions:** As shown in Table 8, removing L_{ABSQ} causes the largest degradation, indicating that quantizer regularization is essential to keep token assignments stable and well-utilized; without it, the discrete bottleneck becomes noisy/collapsed and reconstruction quality drops. Removing both L_{PCM} and $Lsemantic$ also yields a clear decline, showing that Stage-2 improvements are not coming from reconstruction alone. These terms actively organize the latent space to be consistent across patches and semantically grounded. When removed individually, dropping $LPCM$ hurts more than dropping $Lsemantic$, which is expected because PCM directly addresses the patch-wise drift problem that heavily affects perceptual consistency, while semantic alignment mainly refines the concept-level structure of features. Finally, removing the shuffle loss reduces performance as well, supporting the claim that Depth Shuffle provides supervision specifically tied to through-plane structure that pure reconstruction does not enforce.

Table 8: **Ablation based on loss and codebook size.**

Experiments	PSNR (↑)	SSIM (↑)	LPIPs (↓)
Main Baseline (**VOLT**with Two Stage Pre-trainng and ABSQ Module)	35.07	94.56	7.81
Ablation to show importance of Losses			
w/o $\mathcal{L}_{ABSQ}$	32.18	91.27	9.45
w/o $\mathcal{L}_{PCM}$ & $\mathcal{L}_{semantic}$	33.08	92.78	9.11
w/o $\mathcal{L}_{PCM}$	33.87	93.48	8.75
w/o $\mathcal{L}_{shuffle}$	34.12	93.47	8.81
w/o $\mathcal{L}_{semantic}$	34.22	94.05	8.92

Ablation on codebook size: Fig. 10 (a) shows the effect of varying the effective codebook size of ABSQ. Reconstruction quality improves consistently with larger codebook capacity, with PSNR rising from 32.13 to 35.07, SSIM from 89.92 to 95.16, and LPIPS decreasing from 11.27 to 7.61, confirming that larger binary code spaces provide sufficient capacity to capture fine-grained anatomical detail.

Ablation on number of BSQ blocks inside ABSQ: Fig. 10 (b) shows the effect of varying the number of BSQ blocks within ABSQ. Performance improves consistently as blocks increase from 2 to 8, with PSNR improving from 33.51 to 35.49, SSIM from 92.51 to 95.11, and LPIPS decreasing from 8.22 to 7.61. This validates the additive residual design; each additional BSQ block captures infor-

Table 10: **Analysis of dependence on the pretrained semantic teacher.** (a) Reconstruction and generation performance when frozen teacher visual features are directly used as latent representations and the same reconstruction interface is trained. (b) Effect of replacing the frozen semantic teacher used during Stage 2. All other architecture and training settings are held fixed.

(a) Frozen teacher features

Representation	PSNR ↑	SSIM ↑	LPIPS ↓	r-FID ↓
HSENet-3D ()	28.27	80.18	13.37	**16.67**
CT-CLIP-3D ()	25.25	79.37	14.02	17.89
M3D-CLIP ()	**29.37**	**82.22**	**11.28**	17.01

(b) Alternative Stage-2 teachers

Stage-2 teacher	PSNR ↑	SSIM ↑	LPIPS ↓	r-FID ↓
VOLT + HSENet-3D (Shi et al., 2025)	**35.17**	92.25	**7.45**	9.27
VOLT + CT-CLIP-3D (**?**)	33.29	91.56	8.39	10.29
VOLT + M3D-CLIP (Bai et al., 2024)	35.07	**94.56**	7.81	**8.88**

mation missed by prior stages, progressively reducing quantization error and improving perceptual quality.

Ablation to validate semantic grounding from Stage-2 using report generation: To verify that Stage-2 actually improves the semantic quality of the learned representation, we evaluate frozen Stage-1 and Stage-2 encoders on report generation, a task that directly probes clinical language grounding in the latent representation quality. As shown in Table 9, Stage-2 consistently outperforms Stage-1 across all metrics (BLEU 38.54 vs. 36.89, ROUGE 43.17 vs. 42.02, METEOR 40.12 vs. 38.75, BERTScore 89.09 vs. 87.89), confirming that Stage-2's combined distillation and consistency objectives do indeed enrich the latent space with clinically meaningful semantics beyond what volumetric pretraining alone provides. To isolate the contribution of semantic distillation, we additionally evaluate a Stage-2 variant trained without $\mathcal{L}_{semantic}$. Removing this objective degrades all metrics (BLEU 37.11, ROUGE 41.43, BERTScore 86.92), falling below even Stage-1 on BERTScore, which indicates that patch-consistency matching alone reshapes the latent geometry but does not inject the clinical grounding needed for report generation; the teacher distillation loss is essential for transferring scan-text semantic structure into the scan-only encoder.

Table 9: **Ablation of Stage-2 semantic distillation on report generation.**

	BLEU	ROUGE	METEOR	BERTScore
VOLT-stage-II	38.54	43.17	40.12	89.09
VOLT-stage-I	36.89	42.02	38.75	87.89
VOLT-stage-II (w/o $\mathcal{L}_{semantic}$)	37.11	41.43	38.75	86.92

Ablation to evaluate VOLT's Dependence on Teacher Features: Although **VOLT** is trained on a scan-only corpus, its Stage-2 semantic objective transfers clinical priors from a frozen teacher previously pretrained using paired scan-text data. Thus, **VOLT** does not require reports paired with its 73,648 training volumes, but it does not eliminate scan-text supervision from the complete training lineage. To determine whether **VOLT** merely inherits the teacher representation, we directly use the frozen visual features of HSENet-3D, CT-CLIP-3D, and M3D-CLIP as latent representations and train the same reconstruction interface. As shown in Table 10 (a), these representations produce substantially lower reconstruction and generation quality than **VOLT**. Although they contain clinically aligned semantics, the teacher features are optimized primarily for cross-modal discrimination rather than invertible volumetric compression. Moreover, M3D-CLIP produces a 768-dimensional visual representation, whereas **VOLT** uses a compact 16-dimensional latent representation. These results show that **VOLT** does not simply reuse the frozen teacher features; it incorporates transferred semantic structure into a compact latent space while preserving substantially more reconstructive information.

Ablation on sensitivity of VOLT to the choice of semantic teacher: To evaluate whether the Stage-2 improvement depends specifically on M3D-CLIP, we replace it with HSENet-3D and CT-CLIP-3D while keeping the **VOLT** architecture, training corpus, objectives, reconstruction interface, and optimization protocol fixed. As shown in Table 10 (b), the benefit of semantic refinement generalizes across different teachers. HSENet-3D achieves the best PSNR and LPIPS, whereas M3D-CLIP obtains the best SSIM and r-FID. CT-CLIP-3D transfers less effectively, consistent with the mismatch between its CT-specific pretraining and our mixed CT/MRI training distribution. These results show that **VOLT** is not tied to a particular semantic teacher, although teacher modality coverage influences the transferred representation.

Downstream Component Ablations: The reconstruction experiments evaluate how the proposed components affect volumetric fidelity, but do not directly determine whether these components improve the transferability of the learned representation. We therefore conduct component-wise downstream ablations on BTCV segmentation, CirrMRI600+ classification, and closed-ended M3D-VQA. For each downstream task, the pretrained visual encoder is frozen. All variants use the same

Table 11: **Component-wise downstream ablation.**

Pretrained Variant	BTCV Dice ↑	BTCV HD95 ↓	CirrMRI600+ Accuracy ↑	M3D-VQA Overall Accuracy ↑
VOLT (with Two Stage Pre-training and ABSQ Module)	84.87	20.02	76.09	81.39
VOLT (w/o Stage-2)	80.24	25.11	71.03	77.01
VOLT (Stage-2 w/o semantic distillation)	83.18	23.44	73.28	78.72
VOLT (Stage-2 w/o Patch Consistency Matching (PCM))	83.27	24.08	72.28	78.17
VOLT (w/o Depth Shuffle approach in Stage-1)	82.08	25.12	72.45	77.96
VOLT (with BSQ instead of proposed ABSQ module)	81.27	25.36	72.37	77.16

patient-level data split, preprocessing, task-specific downstream architecture and hyperparameters such as learning rate, batch size, etc.

As shown in Table 11, the complete **VOLT** performs best on every downstream metric. Relative to Stage-1 alone, Stage-2 improves BTCV Dice by 4.63 points, reduces HD95 by 5.09, and improves classification and VQA accuracy by 5.06 and 4.38 points, respectively. Stage-2 without semantic distillation recovers part of this improvement, but remains below the complete model. Removing PCM, or Depth Shuffle degrades every downstream task, showing that cross-patch consistency, and through-plane pretraining make complementary contributions. Replacing ABSQ with a single-stage BSQ bottleneck also consistently reduces transfer performance. These results demonstrate that the proposed components improve the transferability of the learned volumetric representation in addition to its reconstruction fidelity.

G Additional Qualitative Analysis

Latent Space Analysis. To evaluate the robustness and anatomical organization of the learned latent space, we conduct an unsupervised t-SNE analysis on the validation set using encoder representations. For each 3D volume, we extract a 16-dimensional embedding and project it to 2D with identical t-SNE settings across all methods. As shown in Fig. 11, where each color denotes a different anatomical region, **VOLT** forms more compact and better-separated clusters with clearer inter-region margins compared to MedVAE and MedITok, which exhibit more dispersed and overlapping manifolds. This confirms **VOLT** captures stronger region-specific structure and reduces anatomical ambiguity in its latent space.

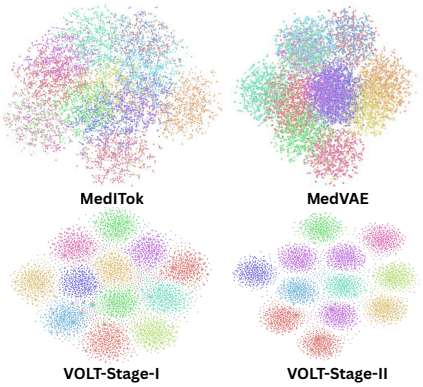


Figure 11: **t-SNE visualization of latent representations.**

Latent Space Interpretability. To better assess whether **VOLT** learns semantically meaningful representations, we perform a text-driven localization analysis using the **frozen** M3D-CLIP text encoder. Concretely, we encode organ prompts (e.g., *"gallbladder"*, *"right kidney"*, *"liver"*, *"stomach"*) into text embeddings, and extract the corresponding volume-level (and/or token-level) latent representations from **VOLT**. We then compute a semantic matching score between the text embedding and the visual latent features, and backpropagate the score to the input volume to obtain a 3D Grad-CAM style relevance map (*i.e.* gradients are aggregated over channels to produce a spatial importance volume, which is then visualized on representative slices).

As shown in Fig. 13, the resulting attention maps consistently highlight the correct anatomical regions across multiple organs, producing sharper and more spatially coherent activations than diffuse responses. Notably, although **VOLT** is trained in a scan-only regime (without paired text supervision), the text-conditioned gradients yield meaningful organ-localization maps, suggesting that **VOLT** learns a latent geometry that is well aligned with the semantic manifold induced by M3D-CLIP. This qualitative evidence supports that our latent space preserves clinically relevant semantics and enables zero-shot text-to-region interpretability through simple similarity-based probing.

3D VQA (Open-ended). Fig. 12 presents examples that illustrate **VOLT**'s ability to generate clinically grounded, reasoning-driven responses from frozen 3D representations. In the first case, **VOLT** correctly identifies collateral ventilation as the core mechanism underlying left upper lobe hyperin-

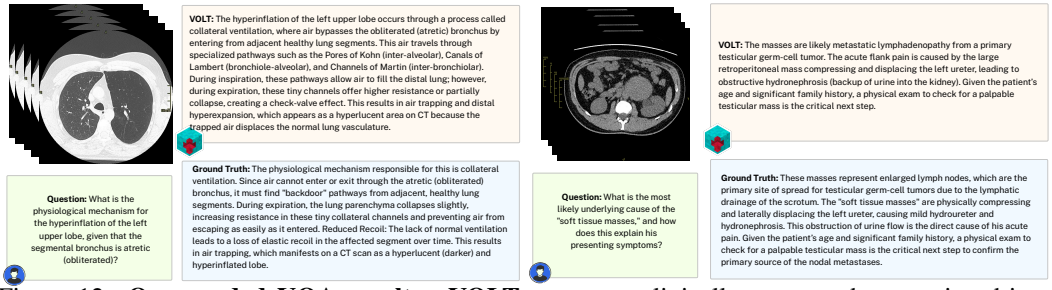


Figure 12: **Open-ended VQA results.** **VOLT** generates clinically structured, reasoning-driven answers that match diagnostic conclusions and reflect deeper anatomical reasoning.

flation, and further enriches the explanation by naming specific collateral pathways (Pores of Kohn, Canals of Lambert, Channels of Martin) and describing the check-valve effect during expiration, demonstrating depth beyond the ground truth. In the second case, **VOLT** accurately diagnoses metastatic lymphadenopathy from a primary testicular germ-cell tumor, correctly explains the causal chain from retroperitoneal mass compression to ureteral displacement and obstructive hydronephrosis, and appropriately identifies the critical next clinical step. Across both examples, **VOLT** produces internally consistent, anatomically grounded responses that align closely with ground truth in clinical reasoning and diagnostic conclusions. This shows that the learned 3D representations encode not just visual structure but semantically meaningful clinical context that supports complex medical question answering.

We further observe a consistent refinement from **VOLT**-StageI to **VOLT**-StageII. While **VOLT**-StageI produces relatively cluttered clusters with increased mixing between neighboring anatomical regions, **VOLT**-StageII yields tighter intra-region groupings and sharper boundaries. This improvement indicates that our patch-level contrastive training encourages boundary-aware, region-consistent features by enforcing local semantic consistency, and that the semantic teacher–student distillation further regularizes the embedding geometry toward more discriminative and semantically aligned representations. In contrast, baseline tokenizers lack these constraints, leading to weaker anatomical separability and increased overlap in the projected space.

Reconstruction Results. Figs. 14-18 present qualitative reconstruction comparisons across five anatomies and modalities; evaluated across all three anatomical planes (axial, coronal, sagittal). Across all cases, **VOLT** reconstructions most closely match the ground truth in terms of tissue contrast, organ boundary sharpness, and structural coherence. In Abdominal-CT (Fig. 14), **VOLT** preserves fine vascular structures and organ interfaces in all three planes, while MedVAE shows mild intensity smoothing and MedITok introduces subtle boundary artifacts, particularly visible in coronal and sagittal views. In Brain-MRI (Fig. 15), **VOLT** faithfully retains cortical folding patterns and white/grey matter contrast across all planes; MedITok notably loses fine sulcal detail in sagittal views. In Chest-MRI (Fig. 16), **VOLT** maintains vessel sharpness and boundaries, while MedVAE and MedITok produce blurred reconstructions. In Pelvis-CT (Fig. 17 and Spine-CT (Fig. 18), **VOLT** accurately reconstructs bony structures, joint spaces, and soft tissue boundaries, whereas others tend to over-smooth bone edges and show inconsistencies across planes. Overall, these multi-plane comparisons demonstrate that **VOLT** produces volumetrically consistent reconstructions, not just in the primary axial plane but across all viewing orientations. This confirms that the learned discrete latent captures true 3D anatomical structure rather than slice-wise appearance.

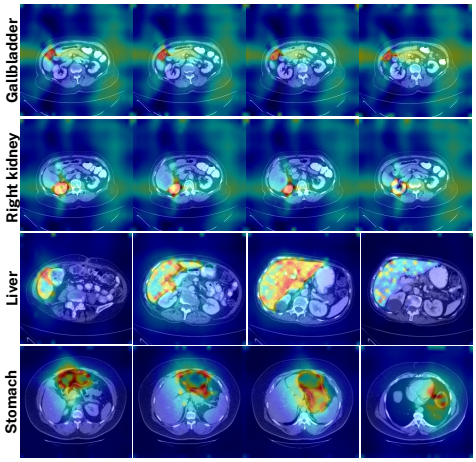


Figure 13: **Attention Maps.**

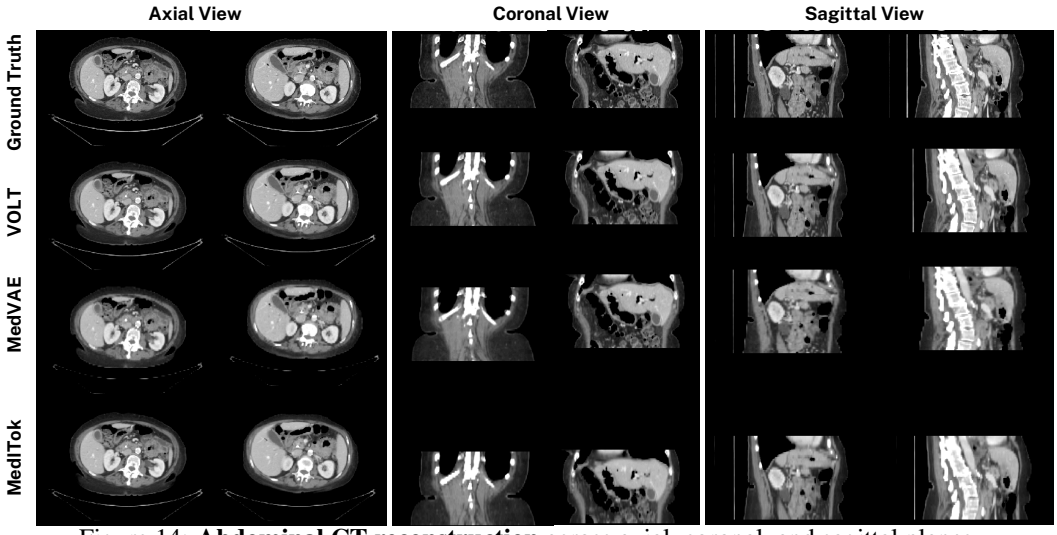


Figure 14: **Abdominal CT reconstruction** across axial, coronal, and sagittal planes

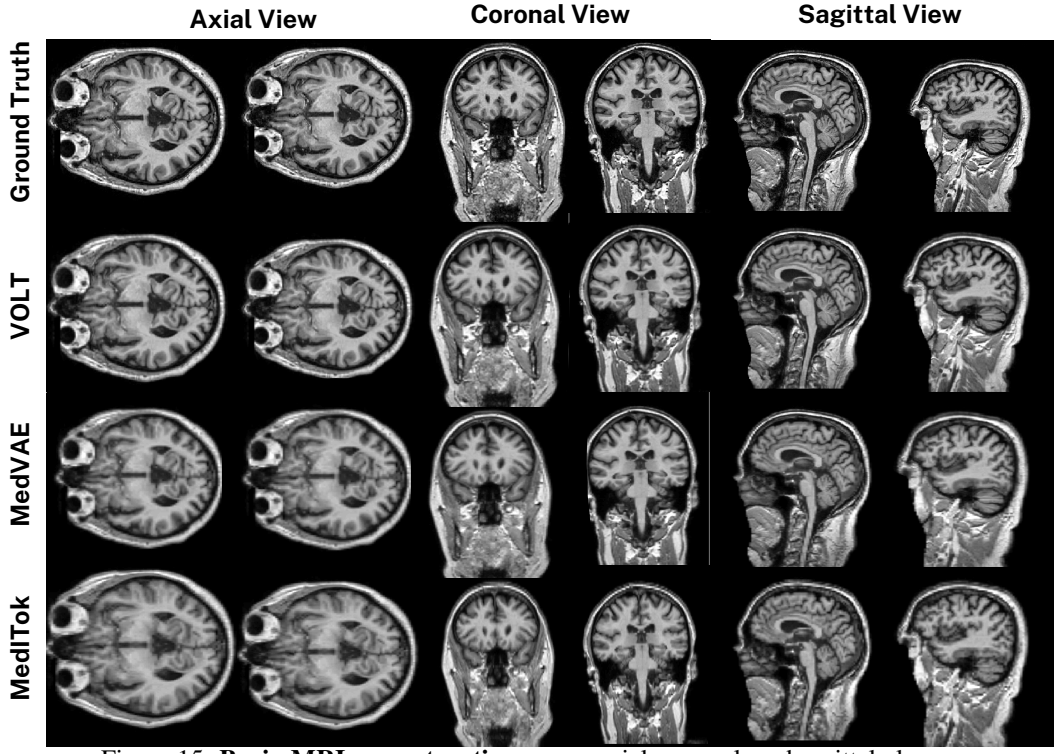


Figure 15: **Brain MRI reconstruction** across axial, coronal, and sagittal planes

H Limitations and Future Work

While **VOLT** achieves consistent improvements across diverse downstream tasks, we identify several promising directions for further development. Our pretraining corpus spans 73,648 CT and MRI volumes with broad anatomical coverage; extending this to additional 3D modalities such as PET or 3D ultrasound is a natural next step that could further validate the generality of the proposed tokenization framework. Additionally, stage-2 semantic distillation leverages a frozen M3D-CLIP teacher to inject clinical semantics without requiring paired reports but in future if more stronger 3D vision-language model emerge, **VOLT**'s distillation framework can readily incorporate them for richer semantic grounding.

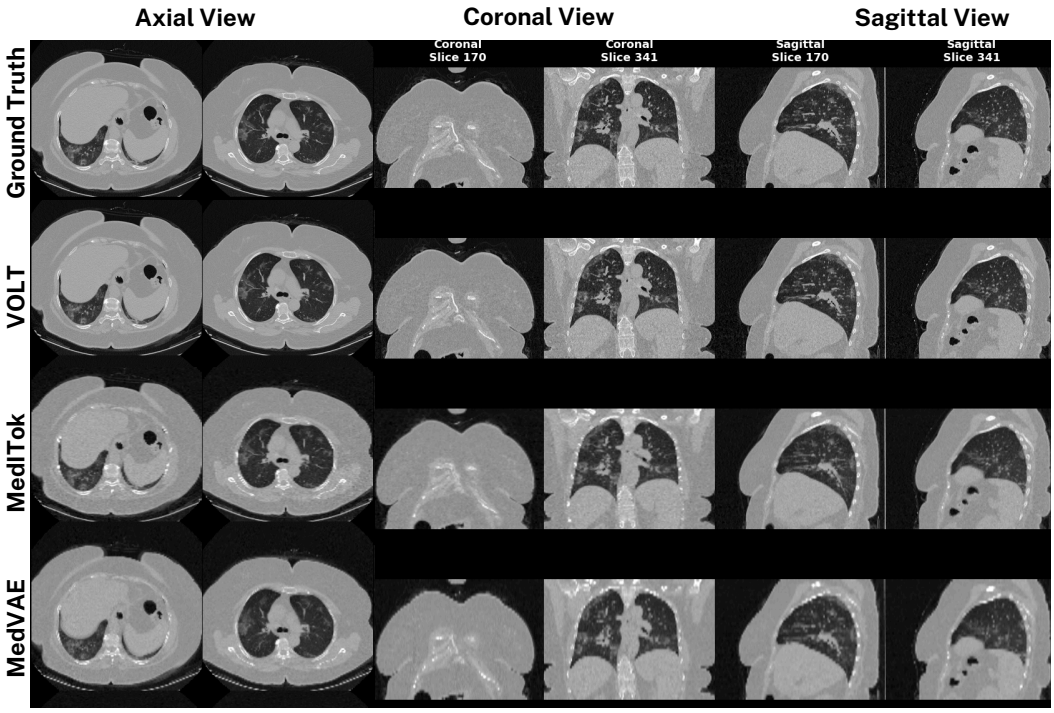


Figure 16: **Chest CT reconstruction** across axial, coronal, and sagittal planes

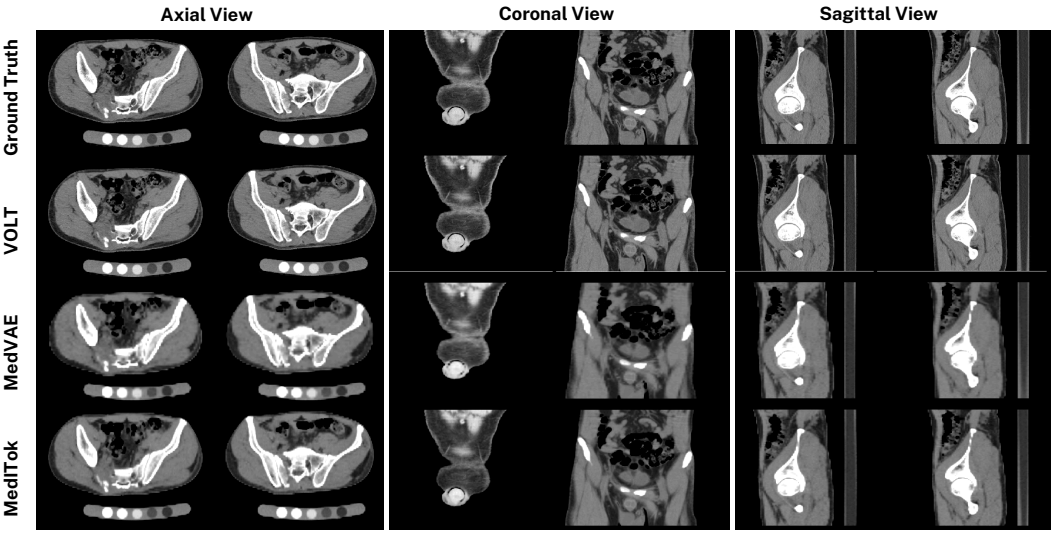


Figure 17: **Pelvis CT reconstruction** across axial, coronal, and sagittal planes

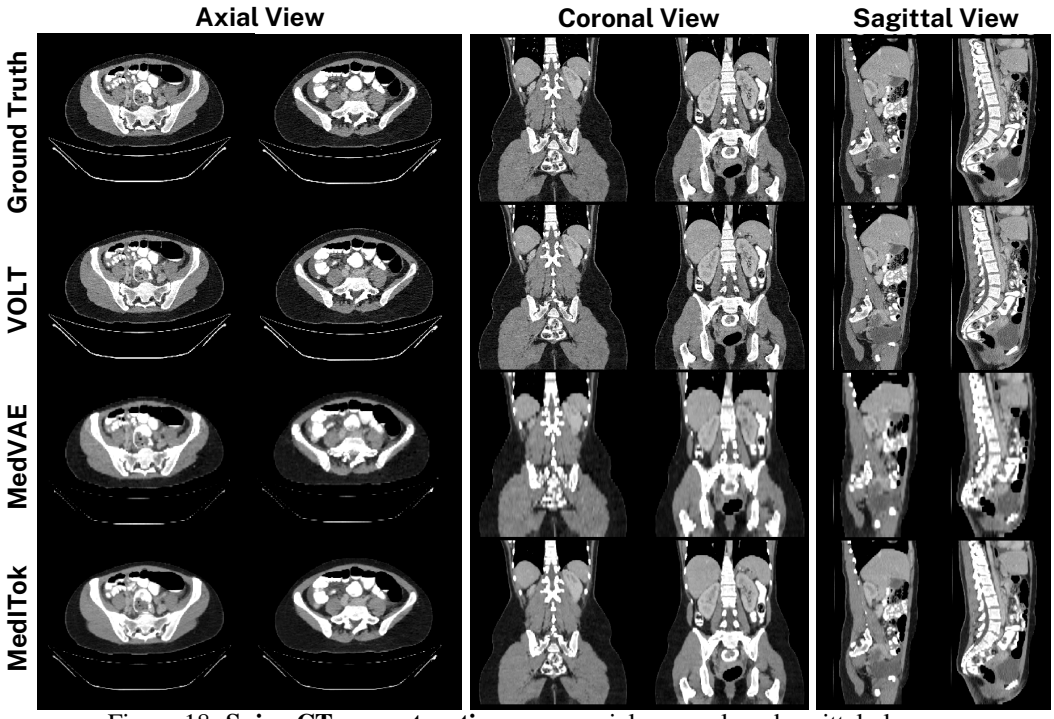


Figure 18: **Spine CT reconstruction** across axial, coronal, and sagittal planes