Text as a Compass: Semantic-Navigated 3D Bone Collapse Prediction via Orthogonal Micro-Volume Primitives

Qingyuan Zeng[1†], Zixin Guan[2†], Yusen Chen[3], Zifeng Wu[3], Hao Li[3], Qian Ma[2], Zixiao Lu[2], Wei He[3], Leilei Chen[3], Wu Zhou[2✉], and Jintai Chen[1✉]

[1] The Hong Kong University of Science and Technology (Guangzhou), China
jintaichen@hkust-gz.edu.cn
[2] School of Medical Information Engineering, Guangzhou University of Chinese Medicine, China
zhouwu@gzucm.edu.cn
[3] The Third Clinical Medical College of Guangzhou University of Chinese Medicine, China

Abstract. Accurate collapse prediction is pivotal for joint-preserving treatment in osteonecrosis of the femoral head (ONFH). However, mainstream 3D-CNN-based methods struggle to capture decisive fine-grained pathological signs due to the intrinsic global receptive field dilution effect, where volumetric isotropic aggregation diffuses sparse pathological singularities. Conversely, clinical reports explicitly document key regional features indicative of collapse, providing critical semantic cues to guide visual focus and enhance prediction. Driven by this clinical insight and the critical need to mitigate the volumetric signal dilution effect, we propose COMPASS (**C**linical **O**rthogonal **M**icro-volume **P**rimitive **A**ligned **S**emantic **S**earch), a novel text-guided fine-grained evidence identification framework for ONFH collapse prediction. First, we introduce the triplanar orthogonal micro-volume primitive (OMVP), a topological manifold representation that concentrates singularity energy onto orthogonal bases to counteract volumetric smoothing. Second, we devise a dual-stage optimization strategy. It begins with semantic space alignment to proactively navigate from macro-ROIs to micro-lesions using textual priors. Furthermore, it incorporates counterfactual causal verification to eliminate spurious correlations via hybrid supervision. Independent evaluations across diverse clinical scenarios confirm COMPASS's state-of-the-art accuracy, yielding 97.01% AUC on a cohort with fine-grained lesion masks and 84.91% on a public dataset relying solely on coarse regional annotations, alongside causally verified localization.

Keywords: Osteonecrosis of the Femoral Head · Text-Driven Prediction · Micro-Volume Primitives · Counterfactual Verification.

† Equal contribution.
✉ Corresponding authors.

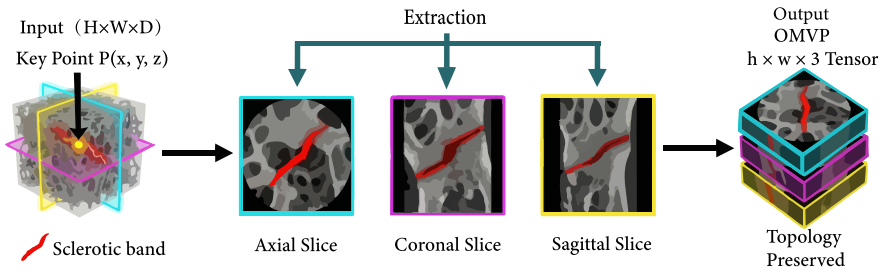

Fig. 1. Illustration of the OMVP extraction process, projecting localized 3D neighborhoods onto tri-planar orthogonal manifolds.

1 Introduction

In osteonecrosis of the femoral head (ONFH), early collapse prediction is crucial for enabling joint-preserving surgeries over inevitable total hip arthroplasty [12]. However, collapse pathogenesis depends less on overall necrotic volume and more on fine-grained micro-structural signs, such as the morphological variations and structural integrity of the sclerotic rim [26]. Consequently, effective predictive algorithms must move beyond macroscopic necrosis localization to precisely capture the subtle textural cues that forecast impending structural failure [8].

Current ONFH collapse prediction predominantly relies on 3D-CNNs, including recent state-of-the-art methods that integrate structured clinical labels [8, 15]. However, these methods struggle with a perceptual bottleneck known as the Global Receptive Field Dilution Effect. As 3D kernels perform isotropic aggregation, the signal energy of sparse pathological singularities, such as sharp sclerotic rim contours, is inevitably diluted within the volumetric support [25]. Consequently, overwhelming background spatial redundancy dominates the gradient, leading networks to prioritize macroscopic patterns while discarding high-frequency evidence as noise.

Addressing this perceptual bottleneck, we shift our focus to radiological reports. In clinical practice, radiologists explicitly document key structural features (e.g., clear sclerotic band) as diagnostic indicators. While these expert descriptions are traditionally used solely for diagnosis, we hypothesize that they can effectively guide neural networks to target the specific imaging regions dictating collapse severity. Regrettably, existing models neglect these semantic priors, failing to retrieve and verify low-level visual evidence [16]. Recognizing this clinical value, we argue that deeply mining these information-rich texts to navigate visual feature extraction is essential for improving collapse prediction [7, 4].

In this work, we reformulate the collapse prediction task as a text-guided fine-grained evidence identification problem and propose a novel framework named COMPASS. While medical vision-language (VL) learning has shown immense promise in medical image analysis [27, 16, 4], existing paradigms typically rely on standard isotropic 3D feature aggregation. Consequently, they remain trapped by the global receptive field dilution effect, smoothing out the very micro-lesions

the text is trying to locate. To resolve this, we design a novel representation unit: the Tri-Planar Orthogonal Micro-Volume Primitive (OMVP). As illustrated in Fig. 1, instead of treating the 3D ROI as an indivisible volume, OMVP topologically transforms local neighborhoods to explicitly preserve the energy of sparse pathological singularities (e.g., sclerotic rims) against volumetric smoothing [23, 10], providing a high-fidelity structural foundation for semantic retrieval.

Furthermore, our technical design fundamentally departs from the associative alignment mechanisms used in conventional VL frameworks [16, 4], which often suffer from spurious correlations and text-induced hallucinations [14]. We devise a dual-stage optimization strategy tailored for fine-grained clinical evidence identification. Stage I performs Semantic Space Alignment, utilizing the clinical report as a semantic prior to proactively navigate from the macroscopic ROI toward potential micro-lesions [7, 4]. Crucially, Stage II introduces Counterfactual Causal Verification. By employing hybrid supervision with counterfactual texts, we explicitly enforce the model to verify the causal presence of the lesion, eliminating spurious cross-modal alignments. This ensures that our text-driven predictions are not merely correlation-based, but firmly grounded in causally verified, logically consistent anatomical evidence.

Our main contributions are: (1) We propose the first text-driven framework COMPASS for ONFH collapse prediction, utilizing information rich clinical reports to transform the task from passive voxel classification into active, semantically guided pathological evidence retrieval. (2) We construct a semantic navigation framework based on OMVP. This approach resolves the smoothing defects of 3D networks via OMVP representation, uniquely synergizing implicit semantic navigation to locate potential lesions with explicit counterfactual reasoning to verify their causality. (3) Independent multi-dataset evaluations demonstrate that COMPASS achieves state-of-the-art accuracy. Notably, our architecture demonstrates strong generalizability across varying levels of visual supervision, seamlessly adapting from ONFH cohorts with fine-grained lesion masks to related bone deterioration scenarios relying solely on coarse regional annotations.

2 Methodology

2.1 Problem Formulation and Overview

Clinically, collapse prediction hinges on locating subtle micro-lesions rather than assessing macroscopic volume. To explicitly target these signs, we reformulate the task from passive volume classification into text-guided evidence identification. Formally, given a multimodal dataset $\mathcal{D} = \{(X_i, T_i, y_i)\}_{i=1}^{N}$, where X_i is the 3D volume, T_i is the clinical report, and $y_i \in \{0, 1\}$ is the collapse label, a topological operator $\mathcal{T} : X_i \to \mathcal{V}_i$ projects the volume into a bag of orthogonal micro-volume primitives (OMVPs), $\mathcal{V}_i$ (Sec. 2.2). As illustrated in Fig. 2, COMPASS maps $\mathcal{F}_\theta(\mathcal{V}_i, T_i) \to \hat{y}_i$ via a dual-stage constrained optimization framework:

$$\min_{\theta} \sum_{i} \mathcal{L}_{\text{task}}(\hat{y}_i, y_i) \quad \text{s.t.} \quad \Phi_{\text{align}}(\mathcal{V}_i, T_i) \geq \tau,\ \Psi_{\text{causal}}(\mathcal{V}_i, T_i, T_{cf}) \to 0 \tag{1}$$

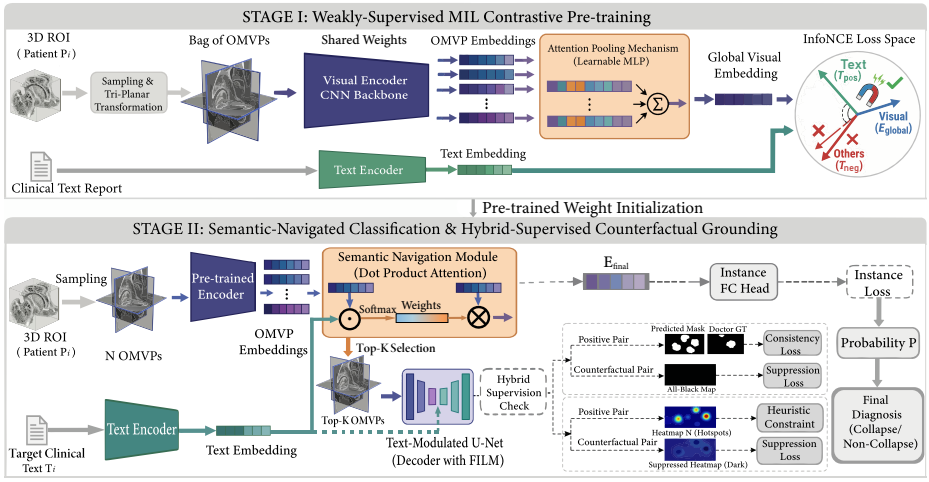

Fig. 2. Overview of the COMPASS framework. 3D volumes are first projected into OMVP manifolds. The optimization then proceeds via Stage I (Semantic Space Alignment) to establish visual-textual priors, followed by Stage II (Counterfactual Causal Verification) for text-navigated lesion identification.

Here, the constraints explicitly dictate the training progression: Φ_{align} denotes semantic space alignment (Stage I), and Ψ_{causal} enforces counterfactual causal verification using counterfactual text T_{cf} (Stage II).

2.2 Orthogonal Micro-Volume Primitives

Clinically, early collapse signs like sclerotic rims are delicate structures easily drowned out by the surrounding bone mass in volumetric assessments. To maximize the pathological signal-to-noise ratio and mimic how radiologists isolate lesions via orthogonal planes, we propose the OMVP (Fig. 1). Formally, we discretize the volumetric domain Ω via a uniform 3D grid to sample a set of anchor coordinates $\mathcal{P} = \{p_k\}_{k=1}^{K}$. For each p_k, we extract a local neighborhood $\mathcal{N}(p_k)$ with fixed spatial dimensions $h \times w \times d$ to define the local receptive field. We then apply a manifold projection operator $\mathcal{T}$ that maps $\mathcal{N}(p_k)$ onto orthogonal basis subspaces:

$$v_{p_k} = \mathcal{T}(p_k, \mathcal{N}(p_k)) = [\Pi_{xy}(p_k), \Pi_{yz}(p_k), \Pi_{zx}(p_k)], \tag{2}$$

where $\Pi_{\{\cdot\}}$ denotes the center-aligned sampling functions along the cardinal planes. This strategy yields a bag of K instances per patient, where each OMVP $v_{p_k} \in \mathbb{R}^{3\times h\times w}$ captures localized structural semantics.

Theoretically, fine-grained pathological signs (e.g., the boundaries of the sclerotic rim) constitute low-rank geometric singularities within the volumetric space. While standard 3D convolution performs a volumetric integral that inevitably dilutes these sparse signals (where signal density decays with kernel

volume), our orthogonal projection concentrates the singularity's energy onto aligned planar manifolds. By stacking these projections channel-wise, OMVP explicitly preserves pathological discontinuities as sharp edges rather than diffuse voxel gradients. Essentially, any anisotropic singularity in 3D space retains its maximal structural footprint on at least one orthogonal projection plane, ensuring critical edge features survive the aggregation process.

2.3 Stage I: Semantic Space Alignment via MIL

Mimicking the clinical workflow where physicians utilize radiological text as a diagnostic prior to actively locate specific pathological regions, we formulate Stage I as a weakly-supervised multiple instance learning (MIL) task (see the top of Fig. 2) [11, 1]. Let f_v and f_t denote visual and textual encoders. For patient i, we extract local visual embeddings $\mathcal{H}_i = \{h_{i,j} | h_{i,j} = f_v(v_{i,j})\}$ from the OMVP bag and a global text embedding $t_i = f_t(T_i)$. To synthesize a holistic representation z_i, we employ a learnable attention aggregation to dynamically weight primitives based on semantic relevance:

$$\alpha_{i,j} = \frac{\exp(w^\top \tanh(V h_{i,j}^\top))}{\sum_k \exp(w^\top \tanh(V h_{i,k}^\top))}, \quad z_i = \sum_j \alpha_{i,j} h_{i,j}, \tag{3}$$

where w, V are learnable parameters. This mechanism implicitly suppresses non-informative background noise while highlighting pathological regions. Finally, we enforce alignment Φ_{align} by maximizing the mutual information of matched pairs (z_i, t_i) via a symmetric InfoNCE loss [19, 27, 24]:

$$\mathcal{L}_{align} = -\frac{1}{2N} \sum_{i=1}^{N} \left(\log \frac{e^{\text{sim}(z_i,t_i)/\tau}}{\sum_{k=1}^{N} e^{\text{sim}(z_i,t_k)/\tau}} + \log \frac{e^{\text{sim}(t_i,z_i)/\tau}}{\sum_{k=1}^{N} e^{\text{sim}(t_i,z_k)/\tau}} \right). \tag{4}$$

This stage instills the encoders with clinical semantic awareness, effectively building a coarse-grained index that maps textual concepts to their potential visual manifestations, establishing a semantic prior for precise navigation in Stage II.

2.4 Stage II: Counterfactual Causal Verification

Reliable clinical diagnosis requires ruling out spurious visual correlations by verifying genuine anatomical evidence. To endow the model with this confirmatory capability, this stage transforms the implicit feature alignment into explicit evidence identification via a dual-path mechanism (Fig. 2, bottom). First, semantic navigation mechanism acts as a noise filter. By computing the similarity distribution $w_{i,j}$ between the clinical text t_i and OMVP bag $\mathcal{V}_i$, we perform a top-k selection to discard non-informative background primitives, yielding a critical subset $\hat{\mathcal{V}}_i$ [17]. The final prediction relies solely on the text-weighted aggregation of critical evidence, supervised by the standard cross-entropy loss $\mathcal{L}_{task}(\hat{y}_i, y_i)$:

$$\hat{y}_i = \mathcal{C}\left(\sum_{v_{i,j} \in \hat{\mathcal{V}}_i} w_{i,j} \cdot f_v(v_{i,j}) \right), \quad w_{i,j} = \text{softmax}(\text{sim}(f_v(v_{i,j}), t_i)/\tau), \tag{5}$$

Table 1. Quantitative comparison and ablation study (%). The first, second, and third best results are highlighted in slate blue, light blue, and light gray, respectively. The superscripts in the last row indicate the absolute improvements over the 3D-CNN baseline. 'Sem. Align.' and 'Causal Verif.' refer to Stage I and Stage II of our framework.

	Private Dataset (Ours)					Public Dataset (OAI)				
Method	**AUC**	**Acc**	**Sen**	**Spe**	**F1**	**AUC**	**Acc**	**Sen**	**Spe**	**F1**
Baselines										
Text-Only	59.33	60.15	53.34	65.03	53.34	63.47	63.79	33.33	77.50	36.36
3D-CNN (ResNet-38)	73.08	69.81	76.92	62.96	73.68	74.19	63.79	71.41	61.36	52.38
Vision-Language (3D+Text)	73.36	83.02	80.77	66.67	84.75	69.16	72.41	64.29	75.01	57.14
State-of-the-Art (MICCAI'25)										
Label Tokenization	87.46	86.79	96.05	77.77	87.72	75.16	77.59	50.07	86.36	51.85
Ablation Study (Ours)										
w/o OMVP (Mono-planar)	92.45	88.67	96.13	81.48	91.22	72.56	77.58	64.28	81.81	62.50
w/o Sem. Align. (Stage I)	85.76	81.13	88.46	74.07	84.21	77.76	74.14	71.43	75.02	61.11
w/o Causal Verif. (Stage II)	94.58	84.90	96.15	74.05	88.13	79.71	72.41	78.57	70.45	61.54
w/ ViT Backbone	75.92	66.03	84.61	48.14	73.01	73.86	70.69	71.46	70.42	57.89
Proposed Framework										
COMPASS (ours)	97.01$^{\uparrow 23.9}$	90.57$^{\uparrow 20.8}$	92.31$^{\uparrow 15.4}$	88.89$^{\uparrow 25.9}$	92.59$^{\uparrow 18.9}$	84.91$^{\uparrow 10.7}$	81.03$^{\uparrow 17.2}$	74.74$^{\uparrow 3.3}$	78.18$^{\uparrow 16.8}$	65.53$^{\uparrow 13.2}$

where $\mathcal{C}$ denotes the classification head. To verify that this focus stems from genuine pathology rather than spurious correlations [2], we introduce hybrid-supervised counterfactual grounding. A decoder $\mathcal{G}$ fuses visual and textual embeddings via cross-attention [16] to project features back to spatial maps under varying semantic conditions. For positive pairs $(v_{i,j}, t_i)$, we enforce anatomical consistency via available priors [18] (clinical masks Y_{mask} or heuristic sparsity).To enforce sparsity where masks are unavailable, we apply a sparse loss $\mathcal{L}_{\text{sparse}}$ that minimizes the average activation of heatmaps outside those masks:

$$\begin{aligned}\mathcal{L}_{pos} = \mathbb{I}_{mask} \cdot \mathcal{L}_{dice}\big(\mathcal{G}(f_v(v_{i,j}), t_i), Y_{mask}\big) \\ + (1 - \mathbb{I}_{mask}) \cdot \mathcal{L}_{sparse}\big(\mathcal{G}(f_v(v_{i,j}), t_i)\big).\end{aligned} \tag{6}$$

Conversely, we construct counterfactual pairs $(v_{i,j}, t_{cf})$ by introducing semantic contradictions (e.g., pairing a "clear sclerotic band" image with "healthy trabecular bone" text), where $t_{cf} = f_t(T_{cf})$. Causally, valid grounding requires the concurrency of visual presence and textual query [22]. Thus, under such conflicting conditions, the model must suppress activation to avoid hallucinations:

$$\mathcal{L}_{cf} = \|\mathcal{G}(f_v(v_{i,j}), t_{cf})\|_2^2 \to 0. \tag{7}$$

The joint objective $\mathcal{L}_{total} = \mathcal{L}_{task} + \lambda_1 \mathcal{L}_{pos} + \lambda_2 \mathcal{L}_{cf}$ optimizes the Stage II. In practice, Stage I initializes the encoders via $\mathcal{L}_{align}$ to fulfill the semantic constraint $\Phi_{\text{align}} \geq \tau$, while Stage II utilizes the auxiliary losses to enforce the causal constraint $\Psi_{\text{causal}} \to 0$, comprehensively solving the formulation in Eq. 1.

3 Experiments

Datasets and Clinical Cohorts. We validate COMPASS via a multi-dataset setup: (1) Private ONFH Dataset: 175 ARCO-II hips (97 collapsed, 78 stable)

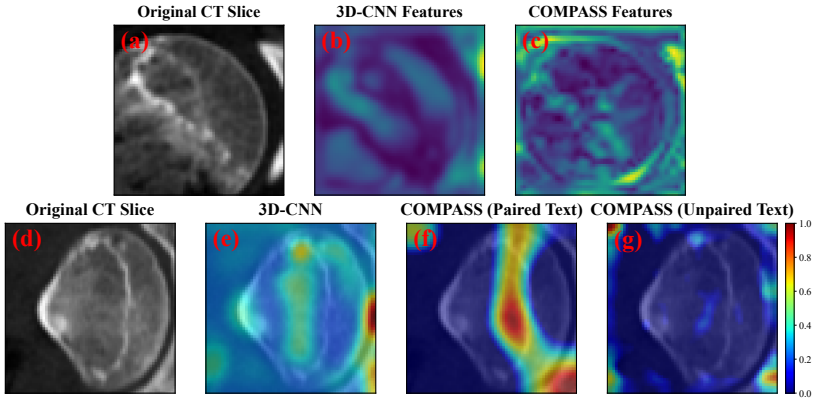

Fig. 3. Sclerotic rim visualization. (a)(d) Original CT. (b)(e) 3D-CNN features/heatmaps. (c) COMPASS features. COMPASS heatmaps guided by (f) paired (*"clear sclerotic band"*) and (g) counterfactual (*"healthy trabecular bone"*) texts.

from 101 patients. Clinical texts and fine-grained masks were acquired at admission; ground-truth collapse was determined via longitudinal follow-up; retrospective, anonymized study approved by the Ethics Committee (PJ-XS-20250725-002), registered at ChiCTR2500107150, in accordance with the Declaration of Helsinki. (2) Public OAI Dataset [21]: We use 191 samples (46 progressive, 145 stable) to assess adaptability to related structural deterioration under a coarse-annotation regime. To establish semantic priors, patient diagnostic variables were formulated into standardized narratives via predefined clinical templates. The progression label (`tkr_incident_108`) denotes total joint arthroplasty within 108 months, mirroring the ONFH collapse endpoint.

Data Preprocessing. ROIs are extracted via minimal 3D bounding boxes from available masks (fine-grained necrotic regions for private; coarse femoral heads for OAI). Boxes are padded by 2 voxels, trilinearly resampled to $64 \times 64 \times 64$, and min-max normalized. We extract a dense bag of $3 \times 32 \times 32$ OMVPs via grid-based uniform sampling across the ROI. During training, a subset of 256 OMVPs is randomly sampled per epoch to enhance model robustness, whereas a fixed subset of the Top-30 OMVPs is deterministically selected during testing.

Implementation Details. COMPASS is implemented in PyTorch [20] and trained on an NVIDIA 2080Ti via 5-fold cross-validation. To mitigate non-standardized terminology in clinical reports, BioBERT [13] extracts robust textual embeddings, while a 2D ResNet [9] processes tri-planar OMVPs as 3-channel images. Stage II auxiliary loss dynamically applies $\mathcal{L}_{dice}$ for the annotated private cohort and $\mathcal{L}_{sparse}$ for the unannotated OAI dataset. Optimization utilizes Adam (batch size 64, lr 5×10^{-4}, $\lambda_1 = \lambda_2 = 1.0$). We report 5-fold averages of standard metrics (AUC, Acc, Sen, Spe, F1).

Comparison with State-of-the-Art. Table 1 summarizes quantitative performance. The Text-Only baseline performs worst, as texts alone lack direct mor-

phological evidence. The standard 3D-CNN [3, 8] yields a 73.08% AUC on the private dataset, confirming the global receptive field dilution effect where isotropic aggregation smooths sparse pathological singularities. Conventional VL frameworks [16, 4] improve results but remain trapped by this volumetric smoothing and lack explicit causal reasoning, risking spurious correlations. For fair comparison with the recent SOTA (Label Tokenization, MICCAI'25) [15], we provided it with all structured metadata (e.g., ARCO staging, demographics). While effective for macroscopic risk stratification, it misses fine-grained regional cues documented in clinical reports. Consequently, for precise micro-structural assessment, COMPASS significantly outperforms it (97.01% vs. 87.46% AUC). This validates our core premise: actively navigating OMVP-preserved visual singularities via clinical semantic priors and counterfactual verification is a superior paradigm. On the public OAI dataset, performance expectedly shifts (AUC 84.91%) due to the transition from precise clinician-drawn masks to coarse annotations. Nevertheless, COMPASS substantially outperforms all competitors, highlighting its task-agnostic versatility even without fine-grained regional masks.

Ablation Study. We ablate our proposed modules in Table 1. First, replacing OMVP with mono-planar projections substantially outperforms 3D convolutions [5, 23, 10] by escaping volumetric smoothing, yet it inherently lacks spatial completeness. Our tri-planar OMVP circumvents this by preserving high-frequency planar features while maintaining spatial geometric integrity. Second, bypassing either Semantic Alignment (Stage I) or Causal Verification (Stage II) leads to notable performance drops. Specifically, omitting Stage I deprives the model of the semantic priors necessary to actively navigate toward potential micro-lesions, whereas removing Stage II leaves it vulnerable to spurious correlations. This combined degradation highlights the necessity of coupling initial semantic guidance with explicit counterfactual reasoning for reliable evidence identification. Finally, replacing the ResNet backbone with a Vision Transformer (ViT) [6] causes a severe performance drop. We attribute this to transformers' data-hungry nature, causing them to overfit on limited medical imaging cohorts [28].

Qualitative Analysis. Fig. 3 visualizes intermediate features and attention heatmaps, highlighting a distinct sclerotic rim (white band) in the original CT. 3D-CNNs severely blur this critical fine-grained structure due to volumetric smoothing (Fig. 3b). In contrast, our OMVP explicitly preserves the sharp contour of the sclerotic rim, yielding a highly informative representation (Fig. 3c). Furthermore, text-guided navigation enables COMPASS to precisely concentrate its attention heatmap on this specific sclerotic rim when given paired clinical reports (Fig. 3f), overcoming the diffuse localization of 3D-CNNs (Fig. 3e). Crucially, applying unpaired counterfactual text successfully suppresses hallucinated activations (Fig. 3g), confirming our model's robust causal grounding.

4 Conclusion

We propose COMPASS, a text-driven ONFH collapse prediction framework that utilizes OMVPs to preserve high-frequency pathological singularities against 3D

isotropic smoothing. By integrating clinical reports via a dual-stage strategy of semantic alignment and counterfactual verification, COMPASS transforms passive classification into active, causally-grounded evidence retrieval. Extensive evaluations demonstrate significant superiority over existing voxel-classification paradigms, validating its architectural robustness across both expert-annotated and weakly-supervised cohorts. Consequently, COMPASS provides a highly interpretable clinical tool to support critical joint-preserving treatment decisions.

Acknowledgments. This work was supported by the Guangdong Basic and Applied Basic Research Foundation (2026A1515011793), the Youth S&T Talent Support Programme of Guangdong Provincial Association for Science and Technology (SKXRC2025467), the Platform Project of Big Data Research Center for Traditional Chinese Medicine at Guangzhou University of Chinese Medicine, the National Natural Science Foundation of China (82374478), the Guangdong Administration of Traditional Chinese Medicine (20231118), and the Science and Technology Department of Guangdong Province (2023A1515010551).

Disclosure of Interests. The authors have no competing interests.

References

1. Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., Busam, K.J., Brogi, E., Reuter, V.E., Klimstra, D.S., Fuchs, T.J.: Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nature Medicine **25**(8), 1301–1309 (2019)
2. Castro, D.C., Walker, I., Glocker, B.: Causality matters in medical imaging. Nature Communications **11**(1), 3673 (2020)
3. Chen, S., Ma, K., Zheng, Y.: Med3D: Transfer learning for 3D medical image analysis. arXiv preprint arXiv:1904.00625 (2019)
4. Chen, Y., Liu, C., Huang, W., Liu, X., Shi, H., Cheng, S., Arcucci, R., Xiong, Z.: GTGM: Generative text-guided 3D vision-language pretraining for medical image segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops. pp. 6715–6724 (2025)
5. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: Proceedings of the Medical Image Computing and Computer Assisted Intervention (MICCAI). pp. 424–432. Springer (2016)
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: Proceedings of the International Conference on Learning Representations (ICLR) (2021)
7. Feng, C.M.: Enhancing label-efficient medical image segmentation with text-guided diffusion models. In: Proceedings of the Medical Image Computing and Computer Assisted Intervention (MICCAI). pp. 253–262. Springer Nature Switzerland (2024)

8. Gao, S., Zhu, H., Wen, M., He, W., Wu, Y., Li, Z., Peng, J.: Prediction of femoral head collapse in osteonecrosis using deep learning segmentation and radiomics texture analysis of MRI. BMC Medical Informatics and Decision Making **24**(1), 320 (2024)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016)
10. He, Y., Guo, P., Tang, Y., Myronenko, A., Nath, V., Xu, Z., Yang, D., Zhao, C., Simon, B., Belue, M., Harmon, S.A., Turkbey, B., Xu, D., Li, W.: VISTA3D: A unified segmentation foundation model for 3D medical imaging. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 20863–20873 (2025)
11. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: Proceedings of the International Conference on Machine Learning (ICML). pp. 2127–2136. PMLR (2018)
12. Karampinas, P., Galanis, A., Papagrigorakis, E., Vavourakis, M., Vlachos, C., Zachariou, D., Pneumaticos, S., Vlamis, J.: Osteonecrosis of the femoral head. optimizing the early-stage joint-preserving surgical treatment? Maedica **17**(4), 948–954 (2022)
13. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics **36**(4), 1234–1240 (2020)
14. Leng, S., Zhang, H., Chen, G., Li, X., Lu, S., Miao, C., Bing, L.: Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13872–13882 (2024)
15. Li, G., Otake, Y., Kameda, Y., Uemura, K., Takashima, K., Mae, H., Kono, S., Hamada, H., Okada, S., Sugano, N., Sato, Y.: Predicting femoral head collapse risk in osteonecrosis using label tokenization: A multi-modality survival analysis approach. In: Proceedings of the Medical Image Computing and Computer Assisted Intervention (MICCAI). pp. 512–522. Springer Nature Switzerland (2025)
16. Li, Z., Li, Y., Li, Q., Zhang, Y., Wang, P., Guo, D., Lu, L., Jin, D., Hong, Q.: LViT: Language meets vision transformer in medical image segmentation. IEEE Transactions on Medical Imaging **43**(1), 96–107 (2023)
17. Lu, M.Y., Williamson, D.F.K., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nature Biomedical Engineering **5**(6), 555–570 (2021). https://doi.org/10.1038/s41551-020-00682-w
18. Milletari, F., Navab, N., Ahmadi, S.A.: V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In: Proceedings of the International Conference on 3D Vision (3DV). pp. 565–571. IEEE (2016)
19. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
20. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 32 (2019)
21. Peterfy, C.G., Schneider, E., Nevitt, M.: The osteoarthritis initiative: report on the design rationale for the magnetic resonance imaging protocol for the knee. Osteoarthritis and Cartilage **16**(12), 1433–1441 (2008)

22. Schölkopf, B., Locatello, F., Bauer, S., Ke, N.R., Kalchbrenner, N., Goyal, A., Bengio, Y.: Toward causal representation learning. Proceedings of the IEEE **109**(5), 612–634 (2021)
23. Shen, Y., Li, J., Shao, X., Romillo, B.I., Jindal, A., Dreizin, D., Unberath, M.: FastSAM3D: An efficient segment anything model for 3D volumetric medical images. In: Proceedings of the Medical Image Computing and Computer Assisted Intervention (MICCAI). pp. 542–552. Springer Nature Switzerland (2024)
24. Tiu, E., Talius, E., Patel, P., Langlotz, C.P., Ng, A.Y., Rajpurkar, P.: Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning. Nature Biomedical Engineering **6**(12), 1399–1406 (2022)
25. Wang, S., Cao, G., Wang, Y.L., Liao, S., Wang, Q., Shi, J., Li, C., Shen, D.: Review and prospect: Artificial intelligence in advanced medical imaging. Frontiers in radiology **1** (2021), https://api.semanticscholar.org/CorpusID:245118639
26. Yang, T.j., Chen, T.x., Zhang, Y., Luo, Y., Wen, P.p.: Senescence-driven osteonecrosis of the femoral head in the elderly—a distinct pathophysiological entity. Frontiers in Medicine **12**, 1653037 (2025)
27. Zhang, Y., Jiang, H., Miura, Y., Manning, C.D., Langlotz, C.P.: Contrastive learning of medical visual representations from paired images and text. In: Proceedings of the Machine Learning for Healthcare Conference. pp. 2–25. PMLR (2022)
28. Zhou, Y., Li, L., Lu, L., Xu, M.: nnWNet: Rethinking the use of transformers in biomedical image segmentation and calling for a unified evaluation benchmark. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 20852–20862 (2025)