This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

# PoCoLT: Position-aware Cross-modal Learning with LLM-decoupled Hierarchical Reports for TMD Diagnosis

Xinyi Zeng[1], Hao Li[1], Pinxian Zeng[2], Xuwei Luo[1], Jize Han[3], Shuai Tan[4], Yan Wang[1,✉]

[1] School of Computer Science, Sichuan University, China
wangyanscu@hotmail.com
[2] Faculty of Science and Technology, University of Macau, Macao SAR
[3] China Mobile (Suzhou) Software Technology Company Limited, China
[4] Computer Science and Engineering, Shanghai Jiao Tong University, China

**Abstract.** Temporomandibular Disorders (TMD), affecting the temporomandibular joint and associated musculature, often cause significant functional impairments. Recent advances in deep learning have boosted diagnostic efficiency, yet TMD diagnosis remains challenging due to the requirement of evaluating dual-position (open/closed-mouth) images. Existing models either fail to effectively capture inter-position interactions or rely heavily on expensive, pixel-level annotations as extra guidance. Moreover, they generally overlook the valuable insights offered by raw diagnostic reports, which offer holistic and mixed descriptions of both positions. In this paper, we propose **PoCoLT**, a novel framework that synergizes **Po**sition-aware **C**ross-m**o**dal **L**earning with L**L**M-guided hierarchical reports for enhanced **T**MD diagnosis. Notably, PoCoLT establishes a novel TMD-tailored Position-Modal paradigm that encompasses both representation construction and alignment learning strategies. For Position-aware Intra-modal Construction (PIC), the visual branch employs a cross-position visual fusion (CVF) module to merge position-specific representations into a cross-position one, assigning higher weights to channels with richer global understanding. For the textual branch, holistic raw reports undergo an LLM-guided Textual Decomposition (LTD) process to generate three hierarchical sub-reports, deriving two position-specific and a cross-position textual embedding. Founded on PIC-derived hierarchical representations, PoCoLT devises a Position-aware Cross-modal Alignment (PCA) strategy at corresponding levels. Operated in a supervised manner, PCA progressively shapes a well-clustered textual space and performs precise semantic alignment in the cross-modal joint space. Experiments show that PoCoLT outperforms representative diagnostic and vision-language baselines in practical report-free TMD pre-evaluation.

**Keywords:** Temporomandibular Disorders (TMD) Diagnosis, Dual-Position MRI, Representation Learning, Large Language Model (LLM).

The first two authors (Xinyi Zeng and Hao Li) contribute equally to this work.

## 1 Introduction

Temporomandibular Disorders (TMD) refer to a group of conditions that affect the temporomandibular joint (TMJ) and associated muscles, often causing pain and functional impairment [1]. TMD diagnosis presents great challenges due to its multifaceted nature, requiring the evaluation of two mouth positions (open and closed) from both the left and right sides of the TMJ for each patient [2]. Traditional diagnosis primarily relies on subjective interpretations of magnetic resonance imaging (MRI) scans, which demand extensive expertise and resources [3]. Therefore, adopting an automated approach is crucial to enhancing the accuracy and efficiency of the diagnostic process.

Thanks to deep learning (DL), the development of medical diagnostic models has seen significant progress. Convolutional Neural Networks (CNNs) [4-6] have achieved impressive performance by learning spatial feature hierarchies through convolutional layers, while Vision Transformer (ViT) and its variants [7-9] have gained attention for their ability to model long-range dependencies via self-attention mechanisms. Notably, for dual-position TMD classification, existing DL-based models generally follow two approaches. The first approach [10] involves separate single-position classification of open- or closed-mouth images, a straightforward method that, however, has limited clinical applicability due to its deviation from standard diagnostic procedures. The second approach [11] employs an indirect segmentation-then-classification pipeline to focus on annotated joint structures, achieving strong performance but constrained by the need for costly, extensive 3D pixel-level annotations for training.

The challenges of using dual-position (open/closed-mouth) images for direct TMD classification can be attributed to two main factors. First, existing DL-based classification models, typically designed for single-image inputs, are ill-equipped to effectively capture the interactions between features from different mouth positions, hindering the extraction and fusion of discriminative diagnostic information. For instance, the widely adopted input-level channel concatenation technique for spatially aligned multi-modal MRI fusion [12] often fails to generalize to the TMD settings, as the two MRI scans exhibit significant spatial misalignment due to position shifts. Second, current TMD classification models [10, 11] primarily rely on image-based classification and overlook the valuable prior information encapsulated in paired radiological reports, which provide critical insights to inform model training [13]. However, since these raw reports typically offer holistic descriptions that lack clear delineation of position-specific details, directly leveraging these reports to train with dual-position images via conventional unsupervised contrastive learning frameworks (e.g., ConVIRT [14], CLIP [15]), which operate exclusively at the sample level, may lead to misalignment of position-specific semantics, potentially degrading performance in the report-free inference.

In this paper, we propose **PoCoLT**, a novel task-tailored framework that synergizes **Po**sition-aware **C**ross-m**o**dal learning with L**L**M-decoupled hierarchical reports for **T**MD diagnosis. Tailored to the characteristics of dual-position imaging and training-only reports in TMD diagnostic workflows, PoCoLT is underpinned by a TMD-tailored Position-Modal paradigm, which comprises two core strategies: Position-aware Intra-Modal Construction (PIC) and Position-aware Cross-Modal Alignment (PCA). For PIC, we construct hierarchical textual and visual representations: (1) For the visual

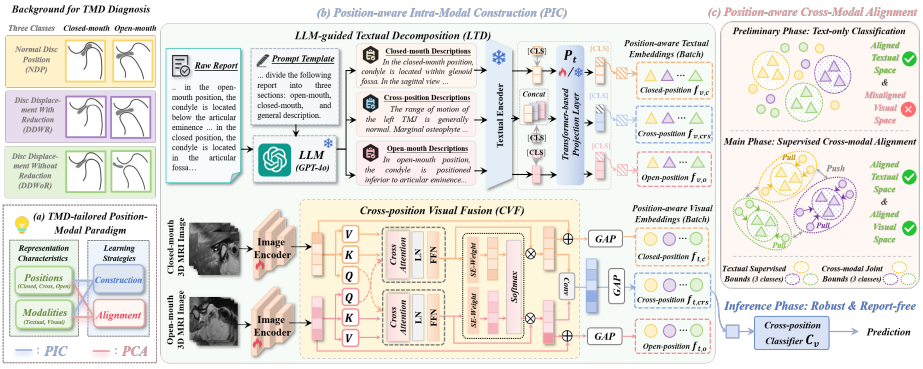


**Fig. 1.** Overview of our PoCoLT. (a) The Position-Modal paradigm serves as the core design philosophy of PoCoLT, outlining two key TMD-tailored components: PIC and PCA. (b) PIC constructs hierarchical multi-modal representations by adopting an LTD process to decouple diagnostic reports and a CVF module to fuse dual-position image features. (c) PCA achieves precise supervised cross-modal alignment at hierarchical levels, conducted in two sequential phases.

branch, position-specific features from dual-position MRI scans are fused into a cross-position visual representation via the Cross-position Visual Fusion (CVF) module, which prioritizes globally informative channels to mitigate spatial misalignment. (2) For the textual branch, inspired by the widespread use of large language models (LLMs) in medical vision [16-18], we introduce the LLM-guided Textual Decomposition (LTD) process to decouple unstructured holistic raw reports into three hierarchical sub-reports, which are encoded into two position-specific and one cross-position textual embeddings. Founded on the hierarchical multi-modal representations by PIC, PCA conducts supervised classification to first refine a tightly clustered textual feature space, then perform fine-grained cross-modal alignment at corresponding levels within the joint space. This design enables PoCoLT to leverage report-derived guidance during training while supporting robust report-free inference. Our contributions are fourfold:

(1) We present PoCoLT, a task-tailored framework synergizing position-aware cross-modal learning with LLM-decoupled hierarchical reports for TMD diagnosis. Notably, its design philosophy lies in the TMD-tailored Position-Modal paradigm, which comprises two core strategies for representation construction and alignment.

(2) To construct hierarchical multi-modal representations, our PIC strategy adopts a CVF module to fuse dual-position visual features, prioritizing globally informative channels to mitigate spatial misalignment. PIC also uses an LTD process to decouple holistic raw reports, deriving position-specific and cross-position textual embeddings.

(3) For hierarchical cross-modal alignment, our PCA strategy employs two supervised training phases: the first exclusively refines a tightly clustered textual space, and the second achieves fine-grained cross-modal alignment at corresponding hierarchical levels, facilitating effective knowledge transfer from textual to visual representations.

(4) Extensive experiments on a private TMD dataset show that PoCoLT outperforms current state-of-the-art methods by a large margin, highlighting its superiority in practical pre-evaluation scenarios where complete diagnostic reports are unavailable.

## 2 Methodology

**Task Definition.** For TMD diagnosis, each patient has bilateral temporomandibular joints (TMJs). Given that the left and right TMJs are anatomically independent structures requiring separate assessments, each joint is treated as an individual sample. Each sample is denoted as $(I_c, I_o, T, y)$, where $I_c$ and $I_o$ represent the closed-mouth and open-mouth MRI images of a single TMJ. $T$ denotes the corresponding raw diagnostic report, which contains descriptions of both position-specific and shared characteristics of the closed- and open-mouth imaging findings. Notably, since TMD reports are typically generated during or after diagnosis, incorporating them for full-modal inference offers limited utility in real-world diagnostic workflows. Therefore, $T$ is exclusively utilized during the training phase in our defined setting. $y$ denotes the TMD diagnostic label, which can be one of the three categories: normal disc position (NDP), disc displacement with reduction (DDWR), or disc displacement without reduction (DDWoR).

### 2.1 TMD-tailored Position-Modal paradigm

Automated TMD diagnosis demands analysis on dual-position MRIs (open/closed-mouth) and can potentially leverage holistic radiological reports to enhance diagnostic performance, yet existing methods fail to integrate these resources effectively: they either overlook position-specific feature interactions in images or misalign cross-modal semantics, thereby limiting representation quality and consistency.

To guide the conceptualization of PoCoLT, we first propose a TMD-tailored Position-Modal Learning Paradigm, which comprehensively accounts for the data characteristics and learning strategies involved in TMD diagnosis. As shown in Fig. 1(a), the left panel defines two dimensions of representation characteristics: Position (covering three levels: open, closed, and cross-position) and Modalities (encompassing textual raw reports and visual MRI scans). The right delineates two representation learning strategies: Position-aware Intra-Modal Construction (PIC), which constructs hierarchical multi-modal representations across positions and modalities, and Position-aware Cross-Modal Alignment (PCA), which aligns semantic content between text and image at corresponding position levels. Notably, we adopt cross-modal alignment instead of fusion in PoCoLT, which minimizes reliance on diagnostic reports during inference.

### 2.2 Position-aware Intra-Modal Construction

Guided by the above paradigm, our PIC strategy builds matched hierarchical position-specific (open/closed) and cross-position representations across visual and textual modalities, encompassing two task-tailored components: the Cross-position Visual Fusion (CVF) module and the LLM-guided Textual Decomposition (LTD) process.

**Cross-position Visual Fusion.** Existing multi-modal models typically rely on channel concatenation for input-level fusion, which often fails to capture interactions between open- and closed-mouth features due to spatial misalignment in the dual-position setting. To address this, we propose the Cross-position Visual Fusion (CVF) module to

effectively integrate position-specific image features, ensuring that the fused cross-position embeddings fully exploit complementary information from both mouth states.

The CVF module begins with bilateral cross-attention to model the correlations between position-specific features. Specifically, we project the closed-mouth feature $F_{v,c}$ and open-mouth feature $F_{v,o}$ via $Proj_c$ and $Proj_o$ to generate query, key, and value matrices. The cross-attention interaction is then realized by enabling the query from one position to interact with the key and value from the other position:

$$[Q_c, K_c, V_c] = Proj_c\big(F_{v,c}\big), [Q_o, K_o, V_o] = Proj_o(F_{v,o}), \quad (1)$$

$$F'_{v,c} = Attn(Q_o, K_c, V_c), F'_{v,o} = Attn(Q_c, K_o, V_o). \quad (2)$$

We then employ the well-designed weighted channel attention to fuse dual-position features. This is accomplished through: (1) implicitly modeling the dependencies between channels within positions using the weighting module $W_{SE}$ in SE-Net [19], and (2) recalibrating via a Softmax function to balance the weights between positions:

$$attn_p = Softmax(W_{SE}(F'_{v,p})), \text{ where } p = c/o. \quad (3)$$

The resulting position-specific attention scores $attn_p$ are multiplied with the features of the corresponding position, with greater weights assigned to channels with greater cross-position understanding. The two feature maps are then concatenated ($Cat$) to form a unified representation $F_{v,crs}$ with enhanced global semantic relevance:

$$f_{v,crs} = GAP(Conv(Cat[F'_{v,c} \odot attn_c, F'_{v,o} \odot attn_o])), \quad (4)$$

$$f_{v,c} = GAP\big(F_{v,c} + F'_{v,c} \odot attn_c\big), f_{v,o} = GAP(F_{v,o} + F'_{v,o} \odot attn_o), \quad (5)$$

where "$\odot$" represents element-wise multiplication, and $Conv(\cdot)$ halves the channel dimension, and $GAP(\cdot)$ stands for global average pooling. The CVF yields three hierarchical embeddings ($f_{v,c}$, $f_{v,o}$, $f_{v,crs}$). Residual connections are incorporated to preserve original position-specific information in $f_{v,c}$, $f_{v,o}$, and the cross-position embedding $f_v^{crs}$ effectively integrates complementary information from both mouth positions.

**LLM-guided Textual Decomposition.** As shown in Fig. 1(b), raw TMD reports provided by clinicians are typically unstructured, holistic descriptions. They lack clear delineations of position-specific details and cross-position information, hindering precise alignment between position-specific visual features and their textual counterparts.

Inspired by the text parsing capability of large language models (LLMs), we propose the LLM-guided Textual Decomposition (LTD) process, which prompts a general LLM (GPT-4o [20]) to decompose raw reports into three semantically distinct components via a task-specific prompt template. Formally, given a prompt instruction template $\mathcal{P}$, the process generates three LTD-decoupled Sub-Reports (LTD-SR):

$$\{T_c, T_o, T_{crs}\} = LLM(\mathcal{P}; T_{raw}), \quad (6)$$

where $T_c$ and $T_o$ denote closed-mouth and open-mouth descriptions, while $T_{crs}$ encapsulates shared/position-agnostic details. These three LTD-SRs are passed through a textual encoder to produce hierarchical embeddings. Specifically, position-specific tokens are concatenated with a learnable [CLS] token and fed into a Transformer-based Projection layer $P_t(\cdot)$ to obtain a refined [CLS] token, serving as $f_{t,o}$ and $f_{t,c}$. For the cross-position embedding $f_{t,crs}$, tokens from all three LTD-SRs are concatenated with the [CLS] token to structurally aggregate holistic information across positions. Notably, diagnostic reports are used only in the offline LTD process during training, while inference requires only dual-position MRI scans.

### 2.3 Position-aware Cross-Modal Alignment

Traditional contrastive learning [15, 21, 22] typically assumes that captions within a batch are entirely distinct, treating only diagonal pairs as positive and all other pairs as negative. However, in TMD diagnosis, due to the limited number of categories and small-scale datasets, intra-batch semantic similarity is elevated (i.e., samples within a batch are more likely to share the same class and similar texts), which violates the unsupervised assumption of sample distinctiveness. Founded on hierarchical visual ($f_{v,c}$, $f_{v,o}$, $f_{v,crs}$) and textual ($f_{t,c}$, $f_{t,o}$, $f_{t,crs}$) embeddings by PIC, we propose the Position-aware Cross-Modal Alignment (PCA) strategy, operating in two learning phases.

**Preliminary Phase - Text-only Classification.** To balance semantic quality and computational cost, we lock the LLM and the textual encoder in the LTD pipeline, with only the textual projection layer $P_t(\cdot)$ set as learnable. Our goal is to optimize this lightweight layer to derive category-discriminative embeddings. Specifically, we introduce a classification head $C_t(\cdot)$ (mapping embeddings into three *textual supervised bounds*) and compute cross-entropy loss over hierarchical outputs ($f_{t,c}$, $f_{t,o}$, $f_{t,crs}$):

$$\mathcal{L}_{pre} = \mathcal{L}_{ce}(C_t(f_{t,c}), y) + \mathcal{L}_{ce}(C_t(f_{t,o}), y) + \mathcal{L}_{ce}(C_t(f_{t,crs}), y), \tag{7}$$

where $y$ is the TMD category label for supervision. By pre-refining with $\mathcal{L}_{pre}$, we drive the projection layer to map initial encodings into a compact, intra-class clustered textual space, serving as a reliable semantic benchmark for subsequent alignment.

**Main Phase - Supervised Cross-modal Alignment.** Unlike unsupervised contrastive learning (which only treats diagonal sample pairs as positives), this phase leverages TMD labels to define off-diagonal positive pairs (same class, regardless of sample identity) to align with the ***textual supervised bounds***. To achieve this, we fix $P_t$ and adopt a supervised alignment loss $\mathcal{L}_{sup}$ to refine multi-modal representations into ***cross-modal joint bounds*** in this phase, which pulls visual features of the same category within a batch closer and pushes features of different categories apart, mitigating potential intra-class separation inherent in unsupervised contrastive settings:

$$\mathcal{L}_{sup}(f_v, f_t) = \sum_{i=1}^{B}\sum_{j=1}^{B} \mathcal{L}_{ce}\left(sim\left(f_v^i, f_t^j\right), \mathbf{R}_{i,j}\right), \text{ where } \mathbf{R}_{i,j} = \begin{cases} 1, & if\ y_i = y_j, \\ 0, & if\ y_i \neq y_j. \end{cases} \tag{8}$$

where $sim(\cdot)$ denotes the inner product and $\mathbf{R}_{i,j}$ is a defined category indicator. To align with the hierarchical representations from PIC, the whole alignment loss $\mathcal{L}_{SCA}$ covers three levels (closed-position, open-position, cross-position), formulated as:

$$\mathcal{L}_{SCA} = \mathcal{L}_{sup}(f_{v,c}, f_{t,c}) + \mathcal{L}_{sup}(f_{v,o}, f_{t,o}) + \mathcal{L}_{sup}(f_{v,crs}, f_{t,crs}). \tag{9}$$

Through progressive alignment in the preliminary (textual space) and main phase (cross-modal space), PoCoLT ensures precise semantic alignment and enhances the robustness of visual representations. Notably, during the main phase, we adopt joint training to perform cross-modal alignment and visual classification simultaneously, which allows mutual reinforcement between the two tasks to accelerate convergence:

$$\mathcal{L}_{main} = \mathcal{L}_{vis-cls}(C_v(f_{v,crs}), y) + \alpha \cdot \mathcal{L}_{SCA}, \tag{10}$$

where $C_v(\cdot)$ denotes the visual classification head, and $\alpha$ is a hyperparameter that balances the two objectives. During the inference phase, we follow the report-free setting, where only the visual branch of PoCoLT classifies the dual-position images.

**Table 1.** Quantitative comparison in terms of five metrics.

| Type | Method | ACC (%) | F1 (%) | AUC (%) | Precision (%) | Recall (%) |
|---|---|---|---|---|---|---|
| Vision-only Classification | Med3D [24] | 60.81±1.73 | 54.58±4.74 | 77.63±5.74 | 58.59±9.47 | 60.56±1.75 |
| | M3T [25] | 62.90±0.88 | 57.67±3.72 | 82.85±0.73 | 69.20±5.24 | 62.81±0.85 |
| | SFusion [26] | 73.23±0.79 | 72.87±0.88 | 87.46±0.96 | 73.89±1.43 | 73.26±0.80 |
| | VoCo [27] | 71.29±0.82 | 70.83±0.96 | 87.95±0.61 | 72.10±1.65 | 71.32±0.82 |
| Vision-Language Classification | ConVIRT [14] | 65.32±0.88 | 62.77±2.64 | 83.45±0.93 | 68.01±4.82 | 65.27±1.02 |
| | MGCA [28] | 70.16±0.88 | 70.00±0.90 | 87.39±0.35 | 69.98±0.97 | 70.21±0.90 |
| | IMITATE [29] | 73.06±0.82 | 71.88±1.50 | 88.23±1.04 | 74.98±2.26 | 73.04±0.87 |
| | T3D [30] | 75.16±0.94 | 74.98±1.00 | 89.27±1.41 | 75.56±0.96 | 75.23±0.98 |
| | **PoCoLT (Ours)** | **80.97**±0.40 | **81.11**±0.44 | **90.73**±0.69 | **81.65**±0.60 | **81.02**±0.39 |

## 3 Experiments

**Datasets.** Considering the scarcity of public datasets, we constructed an in-house TMD dataset to evaluate our PoCoLT framework. The dataset comprises 248 patients, yielding 496 samples (one per left/right TMJ, as each is independently diagnosed), with English reports curated by three specialists, corresponding to 3D MRIs of both mouth states. The dataset was split into training and test sets at a 3:1 ratio.

**Implementation Details.** We utilize 3D ResNet-18 [5] as the learnable visual encoder and ClinicalBERT [23] as the frozen textual encoder. Before training, raw reports were decomposed into three LTD-SRs via GPT-4o [20] via the LTD process, splitting open-mouth, closed-mouth, and cross-position descriptors. In the preliminary training phase, only $P_t$ and $C_t$ are trained. In the main phase, the entire textual branch is kept frozen while the full visual branch and $C_v$ are trained. Following [15], both visual and textual modalities are projected into a 512-dimensional space to perform alignment. Adam optimizer is used for both training phases, with a weight decay of $1 \times 10^{-8}$ and initial learning rate of $2 \times 10^{-5}$. The preliminary and main phases are trained for 30 and 50 epochs, respectively. The balancing coefficient $\alpha$ was selected from preliminary validation trials and fixed to 0.4 for all reported experiments. All experiments were conducted on an RTX 5090 with a batch size of 16. Each experiment was repeated five times, and average results and standard deviation are reported.

**Comparison Analysis.** We conducted comparative experiments involving four vision-only 3D classification methods (Med3D, M3T, SFusion, VoCo) and four vision-language learning methods (ConVIRT, MGCA, IMITATE, T3D). For VL methods, given the misalignment risks of channel-dimension stacking, we concatenated dual-position images along the depth dimension to build single-volume inputs. To ensure fairness, 2D methods [14, 28, 29] were adapted by replacing their 2D encoders with 3D ResNet-18 and adopting the same joint training protocol. As quantified in Table 1, VL approaches generally outperform image-only ones due to the inclusion of complementary textual information. Notably, PoCoLT achieves the highest performance across all five metrics, surpassing the top multi-modal method T3D by 5.81% in ACC. For qualitative validation, CAM visualizations of open/closed-mouth images in Fig. 2 show that baselines like M3T and VoCo either miss key TMJ anatomical regions or only attend to one mouth position, whereas PoCoLT precisely highlights critical areas of both positions.

**Table 2.** Quantitative comparison of ablation models in terms of ACC, F1, AUC.

| Model | Description | ACC (%) | F1 (%) | AUC (%) |
|---|---|---|---|---|
| A | Baseline | 67.58±0.60 | 66.31±1.62 | 85.99±0.64 |
| B | A + PIC (CVF) | 74.68±0.82 | 74.02±1.79 | 89.07±1.13 |
| C | B + PCA (Main (Cross, Unsup)) | 74.19±0.51 | 73.36±1.45 | 88.32±1.38 |
| D | B + PCA (Main (Cross, Sup)) | 75.65±0.60 | 75.70±0.69 | 89.39±0.95 |
| E | D + PCA (Preliminary) | 76.94±1.09 | 76.65±0.94 | 89.88±0.66 |
| F | E + PCA (Main (Specific, Sup)) + PIC (Reg-SR) | 78.39±0.60 | 78.10±0.91 | 90.17±0.69 |
| G (Ours) | E + PCA (Main (Specific, Sup)) + PIC (LTD-SR) | 80.97±0.40 | 81.11±0.44 | 90.73±0.69 |

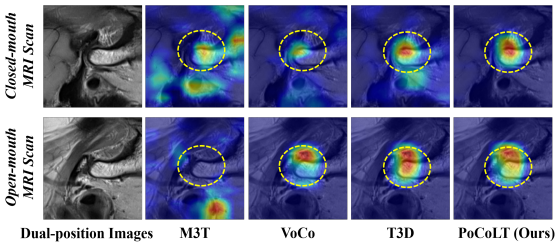


**Fig. 2.** Visualizations of class activation maps for two mouth positions (open/closed) with SOTA methods.

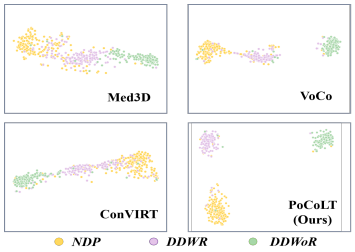


**Fig. 3.** t-SNE visualization of visual embeddings with SOTA methods.

Furthermore, t-SNE visualizations presented in Fig. 3 reveal significant inter-class overlap in Med3D, while ConVIRT only achieves partial separation due to its coarse global text-visual alignment strategy. By contrast, enabled by PCA's precise supervised alignment at hierarchical levels, PoCoLT exhibits robust feature discriminability with distinct clusters and minimal overlap. Collectively, both quantitative and qualitative results validate its superior performance in TMD pre-evaluation.

**Ablation studies.** To validate the effectiveness of each module, we progressively created ablation models. The baseline (Model-A) uses simple convolutional layers with average pooling for visual fusion. Model-B extends Model-A with the CVF module in PIC. Model-C adds the main phase of PCA but adopts only unsupervised cross-modal alignment at the cross-position level (Cross, Unsup), while Model-D adopts the supervised counterpart (Sup). Model-E integrates the preliminary phase (text-only classification) of PCA into Model-D. Model-F and our proposed Model-G extend Model-E with position-specific alignment in the main phase of PCA (Specific, Sup), where subreports are constructed via general regular expression-based decomposition for Model-F (Reg-SR), and via the LTD process for Model-G (LTD-SR). As shown in Table 2, Model-A exhibits limited performance due to its simplistic visual fusion. Integrating the CVF module in Model B boosts ACC by 7.10%, confirming its ability to enhance cross-position visual interaction. Model C adds cross-modal alignment but leads to a slight drop, as its unsupervised formulation induces undesirable intra-class separation. Conversely, Model D implements supervised alignment, which raises ACC to 75.65%, indicating that supervised methods are more effectively suited to the limited-category nature of TMD diagnosis. Model-E integrates the preliminary phase of PCA into Model D, pushing ACC to 76.94%. Extending Model-E with position-specific alignment, Model F (using Reg-SR) reaches 78.39% ACC. Instead, our PoCoLT adopts LTD-SR, raising ACC to 80.97%. This verifies the LTD's superiority over rule-based parsing, as it allows more accurate textual semantics to construct hierarchical representations.

## 4 Conclusions

In this paper, we propose PoCoLT, a novel vision-language framework that addresses key limitations in TMD diagnosis. Targeting dual-position MRI images, PoCoLT establishes a TMD-tailored Position-Modal paradigm encompassing representation construction and alignment. Specifically, Position-aware Intra-Modal Construction (PIC) integrates a cross-position visual fusion (CVF) module and LLM-guided textual decomposition (LTD) process to build hierarchical multi-modal representations. Position-aware Cross-Modal Alignment refines a well-clustered textual space and enables precise supervised alignment at hierarchical levels. Experiments show that PoCoLT outperforms representative diagnostic and vision-language baselines in practical report-free TMD pre-evaluation. PoCoLT currently functions as a TMD pre-evaluation aid and multi-center validation will be pursued in future work to expand its applicability.

**Acknowledgments.** This work is supported by National Natural Science Foundation of China (NSFC 62371325), Sichuan Science and Technology Program 2025NSFJQ0050, Sichuan Science and Technology Program 2024ZDZX0018.

**Disclosure of Interests.** The authors have no competing interests to declare.

## References

1. Gauer, R.L., Semidey, M. J.: Diagnosis and treatment of temporomandibular disorders. American Family Physician 91(6), 378–386 (2015)
2. Ahmad, M., Schiffman, E. L.: Temporomandibular joint disorders and oro-facial pain. Dental Clinics of North America 60(1), 105 (2015)
3. Johnson, M., et al.: Actual applications of magnetic resonance imaging in dentomaxillofacial region. Oral Radiology 38(1), 17–28 (2022)
4. Krizhevsky, A., Sutskever, I., Hinton, G. E.: ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems 25 (2012)
5. Hara, K., Kataoka, H., Satoh, Y.: Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet?. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6546–6555 (2018)
6. Milletari, F., Navab, N., Ahmadi, S. A.: V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV), pp. 565–571. IEEE (2016)
7. Dosovitskiy, A., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
8. Liu, Z., et al.: Swin Transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021)
9. Yang, Y., et al.: BTMuda: a bi-level multisource unsupervised domain adaptation framework for breast cancer diagnosis. IEEE Transactions on Radiation and Plasma Medical Sciences 9(3), 313-324 (2024)
10. Ozsari, S., et al.: Interpretation of magnetic resonance images of temporomandibular joint disorders by using deep learning. IEEE Access 11, 49102–49113 (2023)

11. Kao, Z. K., et al.: Classifying temporomandibular disorder with artificial intelligent architecture using magnetic resonance imaging. Annals of Biomedical Engineering 51(3), 517–526 (2023)
12. Zhou, T., et al.: Deep learning methods for medical image fusion: A review. Computers in Biology and Medicine 160, 106959 (2023)
13. Chauhan, G., et al.: Joint modeling of chest radiographs and radiology reports for pulmonary edema assessment. In: Martel, A.L., et al. (eds.) MICCAI 2020, pp. 529–539. Springer, Cham (2020)
14. Zhang, Y., et al.: Contrastive learning of medical visual representations from paired images and text. In: Machine Learning for Healthcare Conference, pp. 2–25. PMLR (2022)
15. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763. PMLR (2021)
16. Li, C., et al.: LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. In: Advances in Neural Information Processing Systems 36, pp. 28541–28564 (2023)
17. Wang, S., et al.: ChatCAD: Interactive computer-aided diagnosis on medical image using large language models. *arXiv preprint* arXiv:2302.07257 (2023)
18. Kim, K., et al.: LLM-Guided Multi-modal Multiple Instance Learning for 5-Year Overall Survival Prediction of Lung Cancer. In: Linguraru, M.G., et al. (eds.) MICCAI 2024, pp. 239–249. Springer, Cham (2024)
19. Hu, J., Shen, L., Sun, G.: Squeeze-and-Excitation Networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7132–7141. IEEE (2018)
20. Achiam, J., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
21. Jia, C., et al.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International Conference on Machine Learning, pp. 4904–4916. PMLR (2021)
22. Xu, J., De Mello, S., Liu, S., et al.: GroupViT: Semantic segmentation emerges from text supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18134–18144 (2022)
23. Alsentzer, E., et al.: Publicly available clinical BERT embeddings. arXiv preprint arXiv:1904.03323 (2019)
24. Chen, S., Ma, K., Zheng, Y.: Med3D: Transfer learning for 3D medical image analysis. *arXiv preprint* arXiv:1904.00625 (2019)
25. Jang, J., Hwang, D.: M3T: Three-dimensional medical image classifier using Multi-plane and Multi-slice Transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 20718–20729 (2022)
26. Liu, Z., Wei, J., Li, R., Zhou, J.: SFusion: Self-attention based n-to-one multimodal fusion block. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 159–169. Springer (2023)
27. Wu, L., Zhuang, J., Chen, H.: VoCo: A Simple-yet-Effective Volume Contrastive Learning Framework for 3D Medical Image Analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 22873–22882 (2024)
28. Wang, F., et al.: Multi-granularity cross-modal alignment for generalized medical visual representation learning. Advances in Neural Information Processing Systems 35, 33536–33549 (2022)
29. Liu, C., Cheng, S., Shi, M., Shah, A., Bai, W., Arcucci, R.: Imitate: Clinical prior guided hierarchical vision-language pre-training. *IEEE Transactions on Medical Imaging* (2024)

30. Liu, C., et al.: T3D: Towards 3D medical image understanding through vision-language pre-training. arXiv preprint arXiv:2312.01529 (2023)