Cross-Modal Feature Alignment for Automated Diagnosis of Temporomandibular Joint Disorders Using Medical Imaging

Amie L. Mulvany
School of Computing, Clemson University, Clemson, SC, USA.
amiemail@clemson.edu

Abstract

Temporomandibular joint disorders (TMDs) comprise heterogeneous disc, bone, and inflammatory conditions whose imaging appearance varies across magnetic resonance imaging, computed tomography or cone-beam computed tomography, and three-dimensional intraoral scans. Automated diagnosis is difficult because no single modality captures both soft-tissue and osseous pathology, images are acquired with different geometry and contrast, and labelled cohorts are small and unevenly annotated. This manuscript develops a cross-modal feature alignment framework for TMD diagnosis that treats images from different modalities as partial observations of a shared latent disease state. The contribution is a methodological architecture: modality-specific encoders map each input into a latent representation; a shared projection space is trained with contrastive and registration-consistency objectives; anatomically homologous structures, including the condyle, glenoid fossa, articular eminence, disc, and dental occlusion, provide anchors; and an attention-weighted fusion layer produces multi-label diagnostic probabilities with uncertainty estimates. Diagnostic inference is formulated as latent-variable evidence integration, and alignment strength, registration accuracy, missing-modality behavior, and calibration are analyzed as coupled design choices. The analysis shows why latent alignment is preferable to early pixel-level fusion and late score averaging for heterogeneous TMJ imaging, identifies failure modes such as over-alignment, registration drift, and shortcut learning, and specifies patient-level evaluation and calibration criteria needed for clinical interpretation. The framework provides a reproducible design for combining complementary imaging evidence without treating any modality as a substitute for another.

Keywords

temporomandibular joint disorders; cross-modal alignment; medical image analysis; contrastive learning; uncertainty calibration; multimodal diagnosis

1. Introduction

Temporomandibular joint disorders (TMDs) are clinically diagnosed through history, examination, and imaging, with disc displacement, degenerative joint disease, effusion, and pain-related dysfunction representing distinct but frequently coexisting targets. The Diagnostic Criteria for Temporomandibular Disorders formalized a dual-axis structure that separates pain-related and intra-articular diagnoses and linked clinical assessment to specified imaging and examination findings [1]. Earlier Research Diagnostic Criteria established a common vocabulary for disc displacement and osteoarthrosis, but variability in image acquisition and interpretation limited direct conversion of imaging observations into automated labels [2]. Magnetic resonance imaging remains central for disc position,

morphology, and effusion, yet its diagnostic accuracy depends on slice orientation, reader experience, and the chosen reference standard [3]. Computed tomography and cone-beam computed tomography depict cortical bone, osteophytes, erosions, and joint-space change more directly, while three-dimensional intraoral scans capture dental occlusion and mandibular surface geometry. Clinical guidelines for orofacial pain consequently treat imaging as one component of a multimodal diagnostic process rather than as a stand-alone classifier [4].

The resulting computational problem is not simply image classification. Each modality measures a different physical property: MRI reflects proton density and relaxation behavior, CT reflects X-ray attenuation, and intraoral scanning reflects visible surface geometry. Their spatial resolutions, fields of view, and artifact profiles differ, and direct concatenation of voxel arrays or image-level scores discards the anatomical correspondences that make the modalities complementary. Multimodal machine learning addresses such settings by learning representations that are jointly predictive while preserving modality-specific information [5]. For TMJ diagnosis, however, the shared signal is concentrated in a small set of structures that move, deform, and are not always imaged in the same pose. The central gap is therefore a representation-level alignment problem: how to place heterogeneous TMJ observations into a common latent space in which homologous structures support diagnostic reasoning without forcing modalities to become indistinguishable.

The proposed framework makes one coherent contribution: it specifies a geometry-aware cross-modal alignment architecture and analytical evaluation logic for automated TMD diagnosis. Modality-specific encoders produce latent representations; a shared projection space is trained with contrastive alignment and registration consistency; anatomical anchors constrain the correspondence; and an attention-weighted fusion layer yields multi-label probabilities with entropy-based and calibration-based uncertainty reporting. The framework is developed as a methodological and analytical contribution, not as a report of measured clinical performance. Its scientific value lies in making explicit which alignments are identifiable, which failures are likely, and which evidence would be required before automated output could be interpreted alongside established diagnostic criteria.

2. Clinical and Imaging Foundations

2.1 Diagnostic targets and modality complementarity

A useful automated system must distinguish several partially independent targets: anterior disc displacement with and without reduction, degenerative change of the condyle and fossa, joint effusion, and broader masticatory dysfunction. These targets are not interchangeable. Disc position is primarily a soft-tissue judgment best supported by MRI, whereas cortical erosion, osteophytosis, and subchondral change are osseous judgments supported by CT or cone-beam CT. Intraoral scans do not visualize the joint space, but they record occlusal relations and surface landmarks that influence joint loading and mandibular position. The diagnostic label is therefore best represented as a multi-label vector rather than a single mutually exclusive class. This representation respects the coexistence of degenerative and disc-related findings and avoids forcing a patient with both effusion and osseous change into one category.

The imaging evidence also has unequal reliability. MRI can show disc displacement and effusion, but image contrast varies with sequence and the presence of motion or dental material. CT and cone-beam CT provide geometrically stable bone contours, but they do not

resolve the disc and expose the patient to ionizing radiation. Intraoral scans provide detailed dental geometry but cannot assess internal joint structures. These differences create a clinical rationale for cross-modal alignment: the modalities should compensate for one another rather than compete for diagnostic authority.

2.2 Why alignment is required

Early fusion concatenates image channels or voxel patches before classification. It assumes that corresponding voxels represent the same anatomy at the same scale, an assumption violated by different acquisition geometries and by the absence of a common intensity space across MRI and CT. Late fusion averages or stacks modality-specific scores. It is robust to missing modalities but cannot recover complementary evidence that is only visible through joint representation. Latent alignment occupies an intermediate position: each encoder preserves modality-specific structure, while a shared projection is trained to make representations of the same patient and anatomical region close and representations of unrelated regions or patients distant. This design follows the general logic of multimodal representation learning [5], but it must be constrained by TMJ anatomy because the shared signal is spatially localized and mechanically asymmetric.

3. Methodological Framework

3.1 Problem formulation

Let x = {x_MRI, x_CT, x_IOS} denote the available imaging inputs for a patient, where x_MRI is a magnetic resonance volume or slice stack, x_CT is a computed tomography or cone-beam computed tomography volume, and x_IOS is a three-dimensional intraoral surface scan. The diagnostic target y is a binary vector over K findings. Because modalities may be missing, the model must define a distribution over y conditional on any nonempty subset of x. The diagnostic posterior is written as an integration over a latent disease state z that is shared across modalities:

$$p(y \mid x) = \int p(y \mid z)\, p(z \mid x)\, dz, x = \{x_MRI, x_CT, x_IOS\}$$

Here p(z | x) is the modality-fusion encoder distribution and p(y | z) is the diagnostic head. The equation makes the central assumption explicit: modalities provide conditionally independent evidence about z once their modality-specific appearance is encoded. This assumption is analytically useful but not literally true, because image quality and acquisition parameters induce correlated errors. The framework therefore does not treat the integral as a claim of causal identification; it serves as a design principle for factorizing evidence while retaining modality-specific uncertainty.

3.2 Modality-specific encoders

Each modality is mapped to a modality-specific latent representation by an encoder f_θm:

$$z_m = f_\theta m(x_m), m \in \{MRI, CT, IOS\}$$

The symbol θm denotes the parameters of the encoder for modality m, and z_m is the resulting representation before alignment. For MRI and CT, a three-dimensional convolutional backbone with residual connections is appropriate because residual learning permits deep feature extraction without degrading optimization [6]. A U-Net-style auxiliary branch can produce segmentations of the condyle, fossa, disc, and joint space, providing spatially localized supervision that guides the latent representation [7]. For intraoral scans, point-based or voxelized encoders are preferable to slice-based two-dimensional networks

because surface geometry and occlusal relations are inherently three-dimensional. Transformer attention can model long-range dependencies between the condyle, dental arch, and occlusal plane, but its use should be constrained by anatomical priors and limited training data [8].

The encoders should not be trained in isolation indefinitely. Early frozen encoders reduce modality collapse but may prevent adaptation to TMJ-specific contrast. A staged schedule is preferable: first train each encoder on segmentation and modality-appropriate self-supervision, then jointly optimize the shared alignment objective with a small learning rate for the encoders. This balances feature preservation against cross-modal transfer.

3.3 Cross-modal alignment objectives

The shared space is defined by projection functions g_m that map each modality representation into a common dimension. The primary alignment loss minimizes the distance between projections of paired observations while preserving separation from unpaired observations:

$$L_align = \sum_\{m < n\} E[||g_m(z_m) - g_n(z_n)||_2\text{^}2]$$

In this expression, the expectation is over paired samples from the same patient and imaging session, and the sum is over modality pairs. The loss is intentionally simple, but it is insufficient by itself: a constant projection would satisfy it while destroying diagnostic information. The framework therefore couples alignment with a contrastive term that uses positive and negative pairs. For an anchor representation q_i with positive key k_i and negative keys k_j, the normalized temperature-scaled objective is:

$$L_cl = -\sum_i log(exp(sim(q_i, k_i)/\tau) / \sum_j exp(sim(q_i, k_j)/\tau))$$

Here sim(·,·) denotes cosine similarity, τ is a positive temperature controlling the sharpness of the distribution, and the index j ranges over negatives within a batch. Positives can be defined by same-patient correspondence or by anatomically matched regions after registration. This objective follows the logic of contrastive representation learning [11] and of large-scale cross-modal alignment in which paired views are brought together without requiring identical input formats [12]. For TMJ imaging, the negative set must be constructed carefully: contralateral joints, adjacent slices, and same-patient regions can be semantically similar, so indiscriminate negatives can teach the model to separate anatomy that should remain associated.

Registration consistency provides a geometric constraint that contrastive learning alone cannot supply. For CT-to-MRI alignment, a deformation field u maps coordinates in the CT volume to the MRI frame. A regularized registration loss can be written as:

$$L_reg = \sum_r ||x_CT(r + u(r)) - x_MRI(r)||_1 + \lambda \sum_r ||\nabla u(r)||_2\text{^}2$$

The variable r indexes spatial locations, u(r) is the displacement at r, and λ controls smoothness. The first term penalizes intensity mismatch after warping, while the second discourages anatomically implausible folding. Learned registration frameworks can estimate u efficiently [9], and classical symmetric registration provides a mature reference for regularization and evaluation [10]. The registered correspondences are not treated as ground truth; they are used to define positive pairs and to weight the alignment loss by local registration confidence.

3.4 Anatomical anchors and geometric constraints

TMJ alignment is underdetermined if the only requirement is that latent representations match. The condyle and fossa are small structures, the disc can be displaced, and the mandible can be imaged at different openings. Anatomical anchors reduce this ambiguity. Segmentation masks for the condylar head, glenoid fossa, articular eminence, disc, and dental occlusal surface define homologous regions. Landmark detection on intraoral scans can identify dental and skeletal reference points, and segmentation-guided semi-supervised landmark detection illustrates how sparse annotation can be used to regularize geometric correspondence [13]. Such landmarks are not used as a substitute for joint imaging; they anchor the dental and mandibular frame in which joint observations are compared.

The alignment objective should therefore include a masked or region-weighted term. Correspondences inside segmented condyle and fossa masks receive higher weight than background regions, and correspondences that cross anatomical boundaries are down-weighted. This weighting is important because intensity-based registration can align high-contrast dental material while misaligning the low-contrast disc. Region weighting also allows the model to distinguish alignment of osseous contours from alignment of soft-tissue spaces, which have different reliability across modalities.

3.5 Diagnostic fusion and uncertainty

After alignment, the shared representation must be fused for multi-label diagnosis. A simple average treats modalities as equally informative, which is inappropriate when MRI is unavailable or when CT is degraded by metal artifact. An attention-weighted fusion layer assigns modality weights from the latent representations:

$$p(y|x) = softmax(W\sum_m \alpha_m z_m + b), \alpha_m = softmax(a\text{^}T\, tanh(Vz_m + c))$$

In this expression, W and b are the classification parameters, V, a, and c parameterize the attention mechanism, and α_m is the normalized weight for modality m. The softmax is applied independently to each diagnostic label in the multi-label setting. The attention weights should be inspected rather than treated as explanatory by default. Gradient-based localization can identify which image regions contribute to a predicted finding [14], but such maps reveal model sensitivity, not clinical causation.

Uncertainty is reported at the level of the diagnostic output. Predictive entropy summarizes total uncertainty across labels:

$$H(y|x) = -\sum_{c} p(y = c|x)\log p(y = c|x)$$

Here c indexes diagnostic classes, and H is high when the fused prediction is diffuse. Entropy does not separate aleatoric image ambiguity from epistemic model uncertainty. Approximate Bayesian dropout can represent epistemic uncertainty by sampling multiple stochastic forward passes [15], while deep ensembles provide a complementary and often stronger estimate through disagreement across independently trained models [16]. These estimates should be calibrated before being presented as risk. Expected calibration error provides a simple summary of the gap between confidence and observed accuracy across confidence bins:

$$ECE = \sum_\{b = 1\}\text{^}\{B\}\ (|B_b|/N)\ |acc(B_b) - conf(B_b)|$$

Here B is the number of bins, B_b is the set of predictions in bin b, N is the total number of predictions, acc(B_b) is the observed accuracy in the bin, and conf(B_b) is the mean predicted confidence. The formula is a reporting tool, not evidence that a model is calibrated.

4. Analytical Assessment and Design Trade-offs

4.1 Failure modes and boundary conditions

The most important failure mode is over-alignment. If the contrastive weight and temperature force all modalities into an almost identical representation, the model can discard modality-specific evidence such as bone erosion visible only on CT or disc displacement visible only on MRI. The alignment loss should therefore be balanced against reconstruction, segmentation, and diagnostic losses. A second failure is registration drift: intensity-based registration may align dental restorations or air spaces while displacing the condyle. Smoothness regularization reduces folding but cannot guarantee anatomical correctness. Region weighting and registration-confidence estimates are necessary, and alignment should be rejected when the optimized deformation exceeds plausible bounds for the imaged field of view.

A third failure is shortcut learning. A model may associate scanner type, slice thickness, or the presence of dental hardware with a diagnostic label. Because these variables correlate with clinical setting, a model can appear accurate within one site while failing under external acquisition conditions. Adversarial or gradient-reversal objectives can reduce site-specific dependence, but they can also remove clinically relevant signal if the site and disease distributions are confounded. Generative augmentation with adversarial [17] or variational [18] models can increase apparent modality coverage, but synthetic anatomy may distort the very boundaries that define degenerative change. Synthetic data should therefore be evaluated as an augmentation hypothesis, not as a substitute for external validation.

4.2 Evaluation logic

Evaluation must be patient-level, not image-level, because multiple slices or surfaces from one patient are not independent observations. Splitting should be performed by patient and, when possible, by acquisition site. The reference standard should combine clinical diagnosis and independent imaging review, with disagreement reported rather than collapsed. Discrimination metrics such as area under the receiver operating characteristic curve are useful but insufficient: multi-label calibration, sensitivity at clinically meaningful specificity, and decision-curve analysis are more informative for a diagnostic adjunct. Missing-modality performance must be reported separately, because a model that requires all three modalities has limited use in routine care. The ECE defined above should be accompanied by reliability plots and subgroup analysis across scanner type, sex, age, and disease severity.

4.3 Interpretability and deployment

Interpretability should answer a narrow clinical question: does the model rely on the same structures that a clinician would inspect? Attention weights and saliency maps can support error analysis but are unstable under image perturbation and should not be presented as explanations of disease mechanism. A more defensible approach is to report region-specific evidence, such as predicted probabilities derived from condyle, fossa, disc, and occlusal regions, together with the alignment confidence for each region. This allows a clinician to see whether a diagnosis is being driven by a plausible anatomical source or by an artifact outside the joint.

5. Technical Discussion

5.1 Relationship to alternative fusion strategies

Late fusion is attractive because it tolerates missing modalities and preserves modality-specific classifiers. Its weakness is that each classifier must learn from a smaller effective sample and cannot exploit cross-modal correspondences during training. Early voxel fusion is geometrically demanding and vulnerable to misregistration, especially when MRI and CT are acquired at different times or jaw positions. Latent alignment avoids both extremes: it permits modality-specific encoders, supports missing inputs, and creates a shared space in which complementary evidence can be compared. The approach is consistent with multimodal representation learning [5] and with feature recalibration mechanisms that adapt channel importance to context [19]. Predictive coding offers a related objective for learning representations that anticipate related observations [20], but it does not replace anatomical anchors because the TMJ is too small and mechanically variable for unconstrained predictive alignment.

5.2 Implications for TMJ diagnosis

The framework shifts attention from a single best modality to the conditions under which modalities can be jointly interpreted. MRI contributes soft-tissue discrimination, CT and cone-beam CT contribute osseous geometry, and intraoral scans contribute occlusal context. Cross-modal alignment does not make these modalities interchangeable; it makes their disagreement visible. A prediction with high entropy, low registration confidence, or unstable attention across modalities should be treated differently from a confident prediction supported by aligned evidence. This distinction matters because TMD diagnosis often involves uncertain or conflicting findings, and an automated system that suppresses disagreement would misrepresent the clinical problem.

6. Conclusion

Cross-modal feature alignment offers a principled route to automated TMD diagnosis using heterogeneous medical imaging. The proposed framework separates modality-specific encoding from shared representation, constrains alignment with anatomical anchors and registration consistency, fuses evidence through attention-weighted multi-label inference, and reports uncertainty and calibration rather than a bare classification score. Its main analytical claims are that unconstrained alignment is underdetermined for the TMJ, that over-alignment can destroy diagnostically relevant modality-specific information, and that clinical interpretation requires patient-level external evaluation and explicit calibration. These claims define a testable design space for future studies without substituting architectural detail for clinical evidence.

References

1. Schiffman, E., Ohrbach, R., Truelove, E., Look, J., Anderson, G., Goulet, J. P., List, T., & International RDC/TMD Consortium Network. (2014). Diagnostic criteria for temporomandibular disorders (DC/TMD) for clinical and research applications: Recommendations of the International RDC/TMD Consortium Network and Orofacial Pain Special Interest Group. Journal of Oral & Facial Pain and Headache, 28(1), 6–27.

2. Dworkin, S. F., & LeResche, L. (1992). Research diagnostic criteria for temporomandibular disorders: Review, criteria, examinations and specifications, critique. Journal of Craniomandibular Disorders, 6(4), 301–355.

3. Tasaki, M. M., & Westesson, P. L. (1993). Temporomandibular joint: Diagnostic accuracy with sagittal and coronal MR imaging. Radiology, 186(3), 723–729.

4. de Leeuw, R., & Klasser, G. D. (Eds.). (2018). Orofacial pain: Guidelines for assessment, diagnosis, and management (6th ed.). Quintessence Publishing.

5. Baltrušaitis, T., Ahuja, C., & Morency, L.-P. (2019). Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(2), 423–443.

6. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 770–778).

7. Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015 (pp. 234–241). Springer.

8. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (Vol. 30, pp. 5998–6008).

9. Balakrishnan, G., Zhao, A., Sabuncu, M. R., Guttag, J., & Dalca, A. V. (2019). VoxelMorph: A learning framework for deformable medical image registration. IEEE Transactions on Medical Imaging, 38(8), 1788–1800.

10. Avants, B. B., Epstein, C. L., Grossman, M., & Gee, J. C. (2008). Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. Medical Image Analysis, 12(1), 26–41.

11. Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In Proceedings of the 37th International Conference on Machine Learning (pp. 1597–1607).

12. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In Proceedings of the 38th International Conference on Machine Learning (pp. 8748–8763).

13. Liu, S., Jin, J., Li, W., & Yan, X. (2026). Segmentation-Guided Semi-Supervised Landmark Detection on Intraoral 3D Scans. Preprints. https://doi.org/10.20944/preprints202609.1722.v1

14. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision (pp. 618–626).

15. Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In Proceedings of the 33rd International Conference on Machine Learning (pp. 1050–1059).

16. Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. In Advances in Neural Information Processing Systems (Vol. 30, pp. 6402–6413).

17. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In Advances in Neural Information Processing Systems (Vol. 27, pp. 2672–2680).

18. Kingma, D. P., & Welling, M. (2014). Auto-encoding variational Bayes. In Proceedings of the International Conference on Learning Representations.

19. Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 7132–7141).

20. Oord, A. van den, Li, Y., & Vinyals, O. (2018). Representation learning with contrastive predictive coding. arXiv. https://arxiv.org/abs/1807.03748