# OTCHA: Optimal Transport-driven Confidence-aware Latent Hub Alignment for Multi-View Medical Image Classification

Jiwoong Yang[1 [0009−0000−2412−2959]], Haejun Chung[1† [0000−0001−8959−237X]], and Ikbeom Jang[2† [0000−0002−6901−983X]]

[1] Hanyang University, Seoul, Republic of Korea
[2] Hankuk University of Foreign Studies, Yongin, Republic of Korea
{jiwoong, haejun}@hanyang.ac.kr, ijang@hufs.ac.kr

**Abstract.** Multi-view imaging, such as mammography and chest radiography, is a standard component of clinical practice. However, medical images are often unregistered and contain view-specific artifacts or irrelevant background cues that can obscure diagnostically relevant findings. Many existing methods directly fuse per-view representations, allowing such irrelevant content to contaminate the fused embedding and reducing robustness under varying view configurations. We propose OTCHA, a confidence-aware latent hub token alignment module based on optimal transport (OT) that refines patch tokens before fusion for multi-view classification. OTCHA introduces a set of learnable latent hub tokens shared across views. For each view, we compute an OT plan between patch tokens and hub tokens that jointly considers feature similarity and geometry, and augment the OT formulation with token-conditional dustbins to enable partial matching and discard irrelevant tokens. The resulting transport plan provides token-wise matching confidence, which gates hub-mediated message passing and weights a novel optimal-transport-based representation alignment loss to stabilize refinement. Experiments on three multi-view medical image datasets demonstrate consistent improvements over competing baselines across diverse anatomies and view configurations. Our code is available at https://github.com/labhai/OTCHA.

**Keywords:** Multi-view Medical Image Fusion · Confidence-Aware Fusion · Latent Hub · Token-Level Refinement · Cross-view Learning

## 1 Introduction

While 3D medical imaging is increasingly used for complex diagnosis, 2D imaging, such as X-ray, 2D ultrasound, and mammography, remains dominant in most countries. However, a single 2D projection inevitably superimposes overlapping anatomical structures, limiting the visibility of clinically relevant findings. Standard clinical practice therefore involves acquiring multiple views to

† Corresponding authors.

cross-reference complementary information from distinct projections. With recent advances in deep learning, multi-view learning methods have been developed to jointly leverage these multiple projections for improved prediction [14,25,16]. However, robust multi-view fusion remains challenging in real-world clinical scenarios. Clinical images are often unregistered and vary in coverage and acquisition, so many patch tokens correspond to background or view-specific artifacts. When dense tokens are fused without explicitly modeling token reliability, irrelevant content can contaminate the fused embedding and obscure localized discriminative cues.

Prior methods in multi-view medical image analysis often directly fuse per-view representations. Early approaches such as MVC-Net [29] employ late fusion by concatenating or pooling view-wise features. Subsequent approaches strengthen interaction via cross-view attention and the exchange of token features [24,23,20]. Recent methods enforce consistency between single-view and fused predictions via mutual distillation [1] and explore state-space sequence models for efficient cross-view exchange [27,28]. To better address unregistered views, several works have also begun to explicitly learn local cross-view correspondences to guide attention [7,4]. However, these methods offer limited control over token reliability and partial matching, allowing irrelevant background or view-specific artifacts to dilute clinically informative cues in the fused representation. Several methods are designed for domain-specific use, limiting scalability to datasets with varying numbers and compositions of views. In this context, optimal transport (OT) offers a principled framework for learning regularized set-to-set correspondences, where the resulting transport plan can naturally derive confidence measures. OT has been widely used for patch- and feature-level correspondence [18,15] and robust matching with partial or unbalanced formulations [26,5], often incorporating rejection mechanisms such as dustbins [21,12]. In medical imaging, OT-based approaches are also gaining traction, including cross-view representation learning [9] and multimodal alignment [22]. However, using OT to provide token-level partial matching or token-wise confidence remains underexplored in multi-view medical image classification.

In this paper, we propose Optimal Transport Confidence-aware Latent Hub Alignment (OTCHA) for multi-view medical image classification. Unlike conventional fusion designs, OTCHA introduces a set of learnable latent hub tokens as a shared meeting point across views, enabling token refinement prior to fusion. For each view, OTCHA computes an entropic transport plan that aligns patch tokens to hub tokens using a fused cost combining feature similarity with a geometry-aware prior, augmented with token-conditional dustbins for partial matching. The resulting transport plan provides token-level matching confidence, which gates hub-mediated message passing and weights a novel Optimal Transport Representation Alignment (OTRA) loss to stabilize refinement. The main contributions are as follows: (1) We introduce OTCHA, a module that refines per-view patch tokens via feature–geometry fused OT with token-conditional dustbins to a learnable latent hub before fusion. (2) We propose OTRA, a confidence-weighted alignment loss with stop-gradient targets that stabilizes hub-mediated refine-

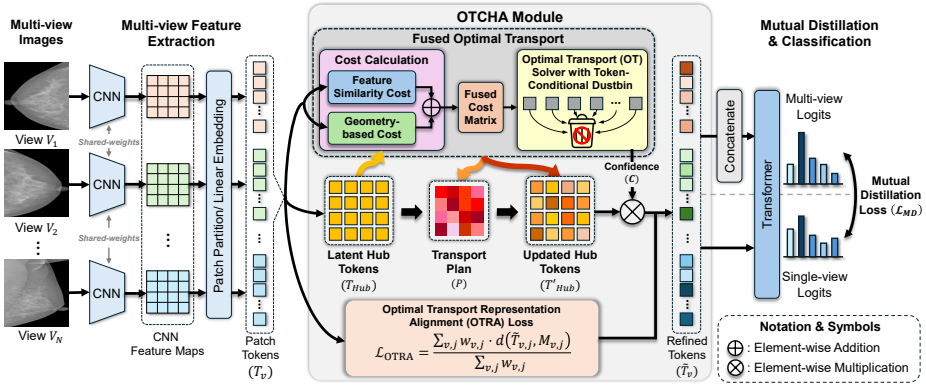


**Fig. 1.** Overview of the proposed framework. OTCHA is inserted between patch embedding and multi-view fusion. It refines per-view patch tokens before fusion using feature- and geometry-based optimal transport to a latent hub, where token-conditional dustbins reject unmatched or irrelevant tokens. OT-based matching confidence guides hub-mediated refinement and weights the OTRA alignment loss, and the refined tokens are fused for classification with mutual distillation.

ment. (3) OTCHA is motivated by radiological cross-referencing practice, and its hub assignment map offers interpretable cross-view correspondences. (4) Experiments on three multi-view medical image datasets across distinct anatomies show consistent improvements over state-of-the-art methods.

## 2 Method

Given $N$ multi-view images $\{\mathbf{I}_v\}_{v=1}^{N}$, we predict the clinical condition by leveraging complementary information across views. OTCHA refines per-view tokens before multi-view fusion as illustrated in Fig. 1. It performs (i) confidence-aware OT matching to a learnable latent hub with token-conditional dustbins and (ii) hub-mediated message passing. We further regularize the refinement process using the proposed confidence-weighted OTRA loss.

### 2.1 Preliminaries

**Base Framework.** We adopt the shared CNN–Transformer encoder and mutual distillation framework of MV-HFMD [1] as the backbone for fair comparison. Each view image $\mathbf{I}_v$ is encoded by a shared CNN encoder and patch-embedding layer into $\mathbf{X}_v \in \mathbb{R}^{S\times D}$, augmented with learnable positional and view-identity encodings. Multi-view prediction is obtained by concatenating tokens from all views and passing them through the Transformer, while single-view logits are computed in parallel for mutual distillation as in [1].

**Optimal Transport.** Optimal transport (OT) finds a minimum-cost assignment between two sets of elements. Given a source set of $S$ elements and a target

set of $M$ elements, OT solves for a transport plan $\mathbf{P} \in \mathbb{R}_{\geq 0}^{S \times M}$ that minimizes the total cost $\langle \mathbf{C}, \mathbf{P} \rangle$ subject to marginal constraints $\mathbf{P}\mathbf{1}_M = \boldsymbol{\mu}$ and $\mathbf{P}^\top \mathbf{1}_S = \boldsymbol{\nu}$, where $\mathbf{C} \in \mathbb{R}^{S \times M}$ is a pairwise cost matrix and $\boldsymbol{\mu} \in \mathbb{R}^S$, $\boldsymbol{\nu} \in \mathbb{R}^M$ are the source and target marginals. We use entropic OT solved by the Sinkhorn algorithm [3] with regularization $\epsilon$; when using an affinity matrix $\mathbf{A}$, we set $\mathbf{C} = -\mathbf{A}$.

### 2.2 Proposed Method

Our proposed method **OTCHA** introduces learnable latent hub tokens $\mathbf{H} \in \mathbb{R}^{M \times D}$ as a shared intermediary, where $M$ is set to match the number of spatial positions in the CNN feature map. For each view, an optimal transport plan aggregates patch-token information into the hub, and the hub then broadcasts the aggregated information back to the view, yielding refined tokens $\tilde{\mathbf{X}}_v$.

**Fused Optimal Transport.** For each view $v$, we compute a transport plan between raw patch tokens $\mathbf{X}_v = \{\mathbf{x}_{v,j}\}_{j=1}^S$ and latent hub tokens $\mathbf{H} = \{\mathbf{h}_k\}_{k=1}^M$ using a fused score matrix. We obtain normalized query and key representations $\mathbf{Q}_v \in \mathbb{R}^{S \times D}$ and $\mathbf{K}_h \in \mathbb{R}^{M \times D}$ via layer normalization and learned linear projections, and compute the fused affinity:

$$\mathbf{A} = \mathbf{Q}_v \mathbf{K}_h^\top + \lambda_{\text{geo}} \cdot \mathbf{A}_{\text{geom}}, \tag{1}$$

where $\mathbf{A}_{\text{geom}}(j,k) = -\|\mathbf{p}_{v,j} - \tanh(\mathbf{p}_{h,k})\|^2 / 2\sigma_{\text{geo}}^2$ is a Gaussian kernel over learnable hub positions $\mathbf{p}_{h,k}$. Here, $\mathbf{p}_{v,j}$ are normalized 2D patch coordinates on the $S$–token grid, and $\sigma_{\text{geo}}$ is a fixed bandwidth.

**Token-Conditional Dustbin.** Among all view tokens, some may contain irrelevant information such as background regions or imaging artifacts, which can introduce noise into the hub representation if forced to participate in transport. To enable partial matching that excludes such tokens, we augment the fused score matrix $\mathbf{A} \in \mathbb{R}^{S \times M}$ with a dustbin row and column:

$$\bar{\mathbf{A}} = \begin{bmatrix} \mathbf{A} & \mathbf{b}_u \\ \mathbf{b}_h^\top & b_0 \end{bmatrix} \in \mathbb{R}^{(S+1)\times(M+1)}, \tag{2}$$

where $b_{u,j} = b_0 + \text{MLP}_{\text{bin}}(\text{LN}(\mathbf{x}_{v,j}))$ and $b_{h,k} = b_0 + \text{MLP}_{\text{bin}}(\text{LN}(\mathbf{h}_k))$, with $b_0$ being a learnable base score. Unlike a global dustbin [21], this token-conditional design lets each token independently control its unmatched mass based on its content. We solve for $\mathbf{P}^* \in \mathbb{R}^{(S+1)\times(M+1)}$ via the Sinkhorn algorithm [3] with entropic regularization $\epsilon$ and rectangular marginals $\boldsymbol{\mu} = [\mathbf{1}_S; M]$, $\boldsymbol{\nu} = [\mathbf{1}_M; S]$. $\mathbf{1}_S$ and $\mathbf{1}_M$ denote all-ones vectors.

**Matching Confidence.** From $\mathbf{P}^*$, we derive confidence scores that modulate the subsequent message passing and representation alignment. Let $\mathbf{P} = \mathbf{P}^*_{1:S,\,1:M}$ denote the transport plan excluding dustbin entries. The view-token confidence $c_{v,j} = 1 - \mathbf{P}^*_{j,\,M+1}$ measures how much mass token $j$ retains for matching rather than being discarded. The hub-token confidence $c_{h,k} = \sum_j \mathbf{P}_{j,k}$ reflects the total mass hub token $k$ receives from the view. The view-level reliability $\beta_v = \frac{1}{S}\sum_j c_{v,j}$ summarizes the overall matching quality of view $v$.

**Hub-Mediated Message Passing.** The hub collects information from all views and broadcasts the fused representation back to each view.
**(i) View → Hub:** For each view $v$, we compute a weighted message from view tokens to each hub token $k$ using the transport plan $\mathbf{P}$:

$$\mathbf{m}_{h,k} = \frac{\sum_j \mathbf{P}_{j,k}\, \mathbf{W}_V \mathbf{x}_{v,j}}{c_{h,k} + \delta}, \tag{3}$$

where $\mathbf{W}_V$ is a learned value projection. Here and in the following, we add a small constant $\delta$ to denominators for numerical stability. A learned sigmoid gate $g_{h,k} = \sigma(\mathrm{MLP}_{\mathrm{gate}}([\mathbf{h}_k;\, c_{h,k};\, \beta_v]))$ modulates the message based on the hub token and matching confidence, yielding the per-view update $\Delta\mathbf{h}_k = g_{h,k} \cdot c_{h,k} \cdot \mathbf{m}_{h,k}$. The hub is updated by averaging across all $N$ views with a residual connection: $\mathbf{H} \leftarrow \mathbf{H} + \mathbf{W}_O\big(\frac{1}{N}\sum_v \Delta\mathbf{H}^{(v)}\big)$, where $\mathbf{W}_O$ is a learned output projection and $\Delta\mathbf{H}^{(v)}$ denotes the per-view hub update assembled from $\{\Delta\mathbf{h}_k\}_{k=1}^{M}$.
**(ii) Hub → View:** Since the hub now carries aggregated multi-view information, the correspondences differ from the initial step. Using the updated hub, we recompute the fused optimal transport (Eq. (1)–(2)) to obtain a new transport plan $\tilde{\mathbf{P}}$, confidence scores $\tilde{c}_{v,j}$, and gate $\tilde{g}_{v,j}$. Each view token is then refined as:

$$\tilde{\mathbf{x}}_{v,j} = \mathbf{x}_{v,j} + \mathbf{W}_O\Big(\tilde{g}_{v,j} \cdot \tilde{c}_{v,j} \cdot \frac{\sum_k \tilde{\mathbf{P}}_{j,k}\, \mathbf{W}_V \mathbf{h}_k}{\tilde{c}_{v,j} + \delta}\Big). \tag{4}$$

All projection weights are shared across both message-passing directions for parameter efficiency.

### 2.3 Loss Function

We train OTCHA with the same classification and mutual distillation loss used in [1], and add our OTRA loss to stabilize hub-mediated refinement. The overall training objective is:

$$\mathcal{L} = \mathcal{L}_{\mathrm{s}} + \mathcal{L}_{\mathrm{m}} + \alpha\, \mathcal{L}_{\mathrm{MD}} + \gamma\, \mathcal{L}_{\mathrm{OTRA}}, \tag{5}$$

where $\mathcal{L}_{\mathrm{s}}$ and $\mathcal{L}_{\mathrm{m}}$ denote the single-view and multi-view classification losses, respectively, and $\mathcal{L}_{\mathrm{MD}}$ is the mutual distillation loss defined in base framework [1]. We keep $\alpha$ identical to the setting in [1], and $\gamma$ weights the OTRA term.
**OTRA Loss.** OTRA provides an explicit token-level learning signal for the hub-mediated message passing and stabilizes training. For each view token $\tilde{\mathbf{x}}_{v,j}$, we define its hub-to-view message $\tilde{\mathbf{m}}_{v,j} = \sum_k \tilde{\mathbf{P}}_{j,k}\, \mathbf{W}_V \mathbf{h}_k \,/\, (\tilde{c}_{v,j} + \delta)$ in Eq. (4), and measure a cosine distance after $\ell_2$-normalization ($\hat{\mathbf{x}} = \mathrm{Norm}_{\ell_2}(\mathrm{LN}(\mathbf{x}))$), weighted by the learned gate and OT-derived view reliability:

$$\mathcal{L}_{\mathrm{OTRA}} = \frac{\sum_{v,j} \overbrace{\tilde{g}_{v,j}\, \beta_v}^{w_{v,j}} \Big(1 - \hat{\tilde{\mathbf{x}}}_{v,j}^{\top}\, \hat{\tilde{\mathbf{m}}}_{v,j}\Big)}{\sum_{v,j} w_{v,j} + \delta}, \tag{6}$$

where $w_{v,j}$ combines the gate $\tilde{g}_{v,j}$ and the OT-derived view reliability $\beta_v$ (Sec. 2.2). We stop gradients through the hub message and the confidence weights to prevent degenerate solutions and stabilize optimization.

## 3 Experiments and Results

### 3.1 Experimental Details

**Datasets and Evaluation.** We evaluate OTCHA on three public multi-view datasets. We split each dataset at the patient level into train/validation/test sets to prevent data leakage. The VinDr-Mammo [17] is a full-field digital mammography dataset labeled with BI-RADS categories. We formulate a binary task by mapping BI-RADS 1–2 to normal and 3–5 to abnormal and following [8] for ROI cropping and background suppression. Each study contains four standard views (LCC, LMLO, RCC, RMLO). We evaluate (i) a breast-level setting using two views per breast and (ii) a study-level setting using all four views, where a study is labeled abnormal if at least one breast is abnormal. The breast-level splits contain 7,199/800/2,000 samples and the study-level splits contain 3,599/400/1,000 studies. The MURA [19] is a musculoskeletal abnormality detection dataset formulated as a binary classification. We select studies with at least two radiographic views and randomly sample two per study, yielding 11,056/1,270/1,098 studies. The CheXpert [11] contains chest radiographs annotated for 13 distinct observations. We use VisualCheXbert labels [13] and include only studies with both frontal and lateral views, and split them into 23,577/3,923/3,913 studies. For all datasets, images are resized to 224×224. We report AUROC for all datasets, using the average over 13 labels for CheXpert. All results are reported as mean and standard deviation over five training runs.

**Implementation Details.** OTCHA is implemented in PyTorch and trained on an NVIDIA RTX 6000 Ada GPU. The backbone is a shared hybrid CNN–Transformer (ResNet26+ViT small pre-trained on ImageNet [6]). We set $\alpha = 0.1$ following [1]. Key hyperparameters were selected via grid search, yielding $\lambda_{\text{geo}} = 0.2$, entropic regularization $\epsilon = 0.2$ with Sinkhorn for 20 iterations. We employ $\gamma = 0.3$ in Eq. (5). We use a batch size of 32, a learning rate of $1e^{-3}$ with SGD, weight decay of $5e^{-4}$, and a cosine annealing scheduler (min LR $1e^{-6}$). The model is trained for up to 100 epochs with early stopping.

**Table 1.** OTCHA achieves the best performance on all dataset configurations. '$*$': statistically significant (Mann-Whitney U, FDR-corrected $p < 0.05$). '–': not supported.

| Method | Backbone | #P(M) | VinDr-Mammo | | MURA | CheXpert |
|---|---|---|---|---|---|---|
| | | | breast-level | study-level | | |
| MVC-Net [29] | ResNet26 | 71.10 | 0.783±.007* | – | 0.859±.003* | 0.900±.001* |
| CVT [24] | ResNet26 | 63.55 | 0.786±.009* | – | 0.849±.005* | 0.898±.002* |
| MVT [2] | DeiT-s | 21.67 | 0.782±.013* | 0.748±.012* | 0.852±.006* | 0.896±.002* |
| ETMC [10] | ResNet26 | 22.35 | 0.797±.007* | 0.745±.008* | 0.869±.005* | 0.903±.001* |
| MV-Swin-T [20] | Swin-T | 58.32 | 0.615±.008* | – | 0.704±.006* | 0.878±.006* |
| MV-HFMD [1] | ViT-s | 36.05 | 0.813±.008 | 0.759±.003* | 0.886±.003* | 0.904±.002* |
| XFMamba [28] | VMamba-s | 58.73 | 0.816±.004* | – | 0.889±.002* | 0.906±.001* |
| OTCHA | ViT-s | 36.73 | **0.826±.006** | **0.782±.005** | **0.895±.003** | **0.908±.002** |

**Table 2.** Computational cost on VinDr-Mammo (per-sample latency and peak memory). The gap over the base MV-HFMD reflects the OTCHA module overhead.

| Method | Stage | 2-view (breast-level) | | 4-view (study-level) | |
|---|---|---|---|---|---|
| | | Lat (ms) | Mem (GB) | Lat (ms) | Mem (GB) |
| MV-HFMD (base) | Inference | 1.08 | 0.96 | 2.30 | 1.63 |
| | Training | 3.47 | 6.64 | 7.19 | 12.93 |
| OTCHA | Inference | 1.65 | 0.97 | 3.42 | 1.65 |
| | Training | 4.75 | 6.85 | 9.22 | 13.34 |

### 3.2 Results and Analysis

**Comparison with State-of-the-Arts.** We compare OTCHA against SOTA multi-view methods across three medical imaging datasets, as summarized in Table 1. For fair comparison, all baselines were retrained under a unified protocol, using official implementations where available. In all datasets and view configurations, OTCHA achieves the best performance with only marginal parameter overhead and statistically significant gains over most baselines. Table 2 further shows that it adds only marginal latency and memory overhead over its base framework MV-HFMD. This demonstrates that refinement-before-fusion can substantially improve multi-view recognition while preserving practical efficiency, supporting OTCHA's effectiveness across diverse anatomies and view configurations. Notably, the performance gain is most pronounced in the study-level (4-view) setting of VinDr-Mammo, suggesting that it offers substantial benefits when integrating complementary views with richer information.

**Ablation studies (Table 3). (I):** Baseline without OTCHA and OTRA. **(II):** Add the OTCHA module with fused OT and token-conditional dustbins, but without OTRA loss. **(III):** Incorporate OTRA loss with uniform weighting. **(IV):** Remove dustbins entirely, reducing to balanced OT. **(V):** Replace token-conditional dustbins with a global dustbin. **(VI):** Remove the geometry cost from fused OT. **(VII):** Full model with all components and confidence-weighted OTRA. (II)–(III) and (VII) show that each component contributes incremen-

**Table 3.** Ablation studies on VinDr-Mammo (study-level) demonstrate the significant improvement of the baseline.

| Num. | Variant | **OTCHA module** | | | **OTRA** | | AUROC ↑ | $\Delta$ (%) |
|---|---|---|---|---|---|---|---|---|
| | | Geo | Dust. | T-cond. | Loss | Conf-wt. | | |
| (I) | Baseline | | | | | | 0.7589 | - |
| (II) | + OTCHA | ✓ | ✓ | ✓ | | | 0.7628 | +0.39 |
| (III) | + OTRA loss | ✓ | ✓ | ✓ | ✓ | | 0.7759 | +1.70 |
| (IV) | w/o dustbin | ✓ | | | ✓ | | 0.7731 | +1.42 |
| (V) | global dustbin | ✓ | ✓ | | ✓ | ✓ | 0.7747 | +1.58 |
| (VI) | w/o geometry | | ✓ | ✓ | ✓ | ✓ | 0.7763 | +1.74 |
| (VII) | **Ours** | ✓ | ✓ | ✓ | ✓ | ✓ | **0.7816** | **+2.27** |

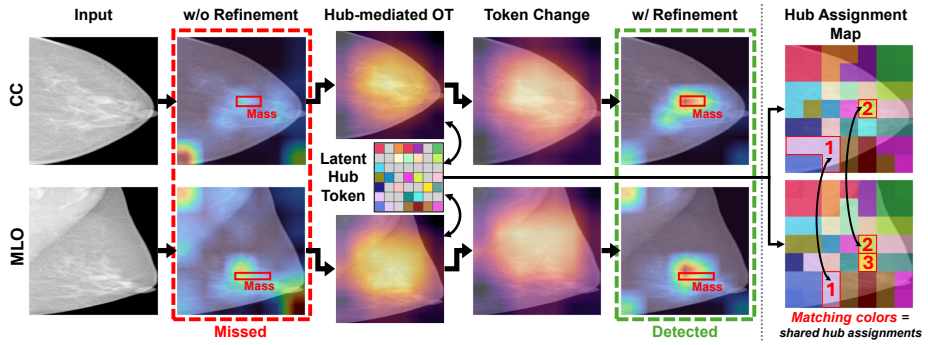


**Fig. 2.** Qualitative visualization of OTCHA refinement. The hub assignment map colors each patch by its assigned hub; matching colors across views indicate cross-view correspondence, while view-specific colors represent regions without counterparts.

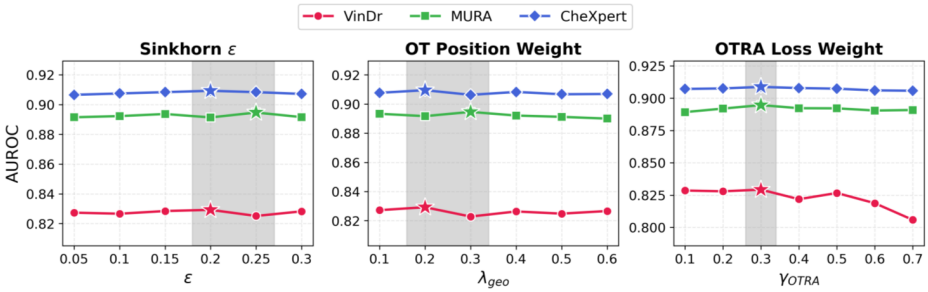


**Fig. 3.** Sensitivity analysis of key hyperparameters. (★ = best result)

tally, while (IV)–(VI) confirm that geometry cost, partial matching, and token-level dustbin control each provide complementary gains.

**Qualitative evaluation (Fig. 2).** We visualize the OTCHA process on a representative mammography case. Without refinement, the model attends broadly to background regions, producing weak activations around the mass. After hub-mediated OT, background activations are suppressed while the mass region receives substantially stronger responses, resulting in localized and concentrated attention. The hub assignment map confirms semantically meaningful token-level correspondences between CC and MLO views. By suppressing view-specific artifacts and reinforcing shared findings across views, OTCHA reflects how radiologists cross-reference views to confirm suspicious regions.

**Hyperparameter sensitivity (Fig. 3).** We analyze the sensitivity of Sinkhorn regularization ($\epsilon$), geometry cost weight ($\lambda_{\text{geo}}$), and OTRA loss weight ($\gamma$) across all datasets. Performance remains stable over a wide range of $\epsilon$ and $\lambda_{\text{geo}}$, demonstrating robustness to these hyperparameters. When $\gamma = 0.3$, optimal performance is achieved, while larger values degrade performance as the alignment term dominates the classification objective.

## 4 Conclusion

We propose OTCHA, a confidence-aware refinement-before-fusion module that leverages optimal transport with token-conditional dustbins. It refines per-view patch tokens via hub-mediated message passing. A confidence-weighted OTRA loss further stabilizes refinement by aligning refined tokens with their hub-broadcast messages. By suppressing clinically irrelevant tokens, OTCHA aligns with how radiologists selectively attend to and cross-reference clinically meaningful regions across views. Moreover, the hub assignment map provides interpretable token-level correspondences across views, potentially facilitating the localization and verification of suspicious findings. Experiments across three datasets/anatomies show consistent gains with marginal parameter overhead, suggesting a promising direction for diverse multi-view medical applications. Although OTCHA supports variable numbers of views, as reflected in the 2- and 4-view settings in Table 1, our evaluation uses all available views; future work will explore missing-view settings, additional modalities, and downstream tasks.

**Acknowledgments.** This work was supported by NRF (RS-2024-00338048, RS-2024-00455720, RS-2026-25487796), National Institute of Health (2026-ER0901-00, 2026-ER0904-00), Hankuk University of Foreign Studies Research Fund of 2025, KOCCA R&D Program (RS-2024-00332210), IITP AI Graduate School Program (RS-2020-II201373, Hanyang University), IITP AI Semiconductor Support Program (RS-2023-00253914), and AI Seoul Tech Research Support Program of Seoul Future Foundation.

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

## References

1. Black, S., Souvenir, R.: Multi-view classification using hybrid fusion and mutual distillation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 270–280 (2024)
2. Chen, S., Yu, T., Li, P.: Mvt: Multi-view vision transformer for 3d object recognition. arXiv preprint arXiv:2110.13083 (2021)
3. Cuturi, M.: Sinkhorn distances: Lightspeed computation of optimal transport. Advances in neural information processing systems **26** (2013)
4. Dabboussi, M., Huard, M., Gousseau, Y., Gori, P.: Self-supervised multiview xray matching. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 578–588. Springer (2025)
5. De Plaen, H., De Plaen, P.F., Suykens, J.A., Proesmans, M., Tuytelaars, T., Van Gool, L.: Unbalanced optimal transport: A unified framework for object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3198–3207 (2023)
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
7. Du, Y., Chen, L., Dvornek, N.C.: Geometry-guided local alignment for multi-view visual language pre-training in mammography. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 299–310. Springer (2025)

8. Ghosh, S., Poynton, C.B., Visweswaran, S., Batmanghelich, K.: Mammo-clip: A vision language foundation model to enhance data efficiency and robustness in mammography. In: International conference on medical image computing and computer-assisted intervention. pp. 632–642. Springer (2024)
9. Gorade, V., Sing, A., Mishra, D.: Otcxr: Rethinking self-supervised alignment using optimal transport for chest x-ray analysis. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 7143–7152. IEEE (2025)
10. Han, Z., Zhang, C., Fu, H., Zhou, J.T.: Trusted multi-view classification with dynamic evidential fusion. IEEE transactions on pattern analysis and machine intelligence **45**(2), 2551–2566 (2022)
11. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al.: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 590–597 (2019)
12. Izquierdo, S., Civera, J.: Optimal transport aggregation for visual place recognition. In: Proceedings of the ieee/cvf conference on computer vision and pattern recognition. pp. 17658–17668 (2024)
13. Jain, S., Smit, A., Truong, S.Q., Nguyen, C.D., Huynh, M.T., Jain, M., Young, V.A., Ng, A.Y., Lungren, M.P., Rajpurkar, P.: Visualchexbert: addressing the discrepancy between radiology report labels and image labels. In: Proceedings of the Conference on Health, Inference, and Learning. pp. 105–115 (2021)
14. Ji, C., Du, C., Zhang, Q., Wang, S., Ma, C., Xie, J., Zhou, Y., He, H., Shen, D.: Mammo-net: Integrating gaze supervision and interactive information in multi-view mammogram classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 68–78. Springer (2023)
15. Le, T., Nguyen, K., Sun, S., Ho, N., Xie, X.: Integrating efficient optimal transport and functional maps for unsupervised shape correspondence learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23188–23198 (2024)
16. Manigrasso, F., Milazzo, R., Russo, A.S., Lamberti, F., Strand, F., Pagnani, A., Morra, L.: Mammography classification with multi-view deep learning techniques: Investigating graph and transformer-based architectures. Medical Image Analysis **99**, 103320 (2025)
17. Nguyen, H.T., Nguyen, H.Q., Pham, H.H., Lam, K., Le, L.T., Dao, M., Vu, V.: Vindr-mammo: A large-scale benchmark dataset for computer-aided diagnosis in full-field digital mammography. Scientific Data **10**(1), 277 (2023)
18. Ni, J., Li, Y., Huang, Z., Li, H., Bao, H., Cui, Z., Zhang, G.: Pats: Patch area transportation with subdivision for local feature matching. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 17776–17786 (2023)
19. Rajpurkar, P., Irvin, J., Bagul, A., Ding, D., Duan, T., Mehta, H., Yang, B., Zhu, K., Laird, D., Ball, R.L., et al.: Mura: Large dataset for abnormality detection in musculoskeletal radiographs. arXiv preprint arXiv:1712.06957 (2017)
20. Sarker, S., Sarker, P., Bebis, G., Tavakkoli, A.: Mv-swin-t: Mammogram classification with multi-view swin transformer. In: 2024 IEEE International Symposium on Biomedical Imaging (ISBI). pp. 1–5. IEEE (2024)
21. Sarlin, P.E., DeTone, D., Malisiewicz, T., Rabinovich, A.: Superglue: Learning feature matching with graph neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4938–4947 (2020)

22. Shaaban, M.A., Saleem, T.J., Papineni, V.R.K., Yaqub, M.: Motor: Multimodal optimal transport via grounded retrieval in medical visual question answering. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 459–469. Springer (2025)
23. Sun, Z., Jiang, H., Ma, L., Yu, Z., Xu, H.: Transformer based multi-view network for mammographic image classification. In: International conference on medical image computing and computer-assisted intervention. pp. 46–54. Springer (2022)
24. Van Tulder, G., Tong, Y., Marchiori, E.: Multi-view analysis of unregistered medical images using cross-view transformers. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 104–113. Springer (2021)
25. Wan, P., Zhang, S., Shao, W., Zhao, J., Yang, Y., Kong, W., Xue, H., Zhang, D.: Correlation-adaptive multi-view ceus fusion for liver cancer diagnosis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 188–197. Springer (2024)
26. Xu, M., Gould, S.: Temporally consistent unbalanced optimal transport for unsupervised action segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14618–14627 (2024)
27. Yang, Z., Zhang, J., Wang, G., Kalra, M.K., Yan, P.: Cardiovascular disease detection from multi-view chest x-rays with bi-mamba. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 134–144. Springer (2024)
28. Zheng, X., Chen, X., Gong, S., Griffin, X., Slabaugh, G.: Xfmamba: Cross-fusion mamba for multi-view medical image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 672–682. Springer (2025)
29. Zhu, X., Feng, Q.: Mvc-net: Multi-view chest radiograph classification network with deep fusion. In: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI). pp. 554–558. IEEE (2021)