Enhancing Interpretability in CLIP with Optimal Transport-based Submodular Optimization for Ophthalmic Imaging

Sihong Lu[1], Jian Yang[1], and Lei Luo[1](✉)

[1]Nanjing University of Science and Technology, Nanjing 210094, China
Project Page: https://github.com/SihongLu/OTSMDL
lusihong@njust.edu.cn, luoleipitt@gmail.com

Abstract. Although Contrastive Language-Image Pretraining (CLIP) has shown promising transfer ability for ophthalmic image analysis, its opaque decision-making process hinders interpretability, especially in ophthalmic settings where probability-only outputs may be difficult for clinicians and patients to trust. Providing visual evidence for predictions can therefore make models more inspectable and interpretable. However, existing explainability methods often struggle to produce spatially consistent and semantically aligned heatmaps that highlight prediction-relevant anatomical or pathological evidence. To bridge this gap, we propose OTSMDL, a post-hoc framework for interpretable region selection that couples submodular-inspired scoring with unbalanced optimal transport. We formulate the identification of salient evidence regions as a subset selection problem. Specifically, submodular-inspired utilities first construct an initial candidate subset and a score-guided source distribution, which are then used to initialize the Optimal Transport (OT) process. Subsequently, OT performs global transport-based re-ranking over all candidates, constructing a unified cost matrix to encourage cross-region competition and align candidate superpixels with text prompts. We evaluate OTSMDL on seven ophthalmic datasets spanning color fundus photography, OCT, and glaucoma assessment, with additional ImageNet experiments for out-of-domain analysis. Under standard post-hoc Deletion/Insertion protocols and ground-truth mask localization metrics, OTSMDL generally outperforms baselines in region-level evidence grounding, achieving top-1/2 performance in most evaluation settings.

Keywords: Interpretability · CLIP · Optimal Transport · Submodular Optimization · Ophthalmic Imaging

1 Introduction

The Contrastive Language-Image Pretraining (CLIP) model [29] has shown exceptional potential in ophthalmology by aligning medical images with diagnostic text. Recent ophthalmic vision-language models have further extended this paradigm through joint pretraining for multimodal retinal analysis [32], ranking-aware prompting for ordinal diabetic retinopathy grading [41], and expert-informed

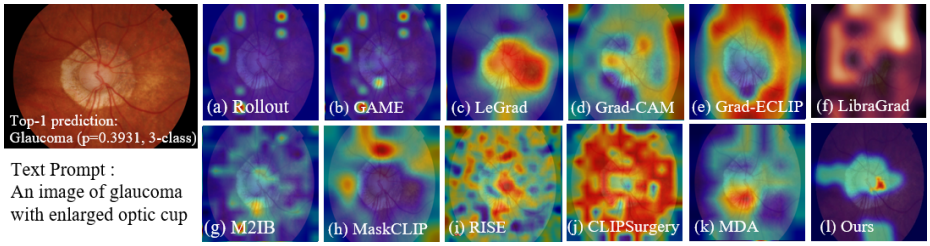

Fig. 1: Explanation heatmaps for a CLIP glaucoma prediction. (a) Rollout [1], (b) GAME [6], (c) LeGrad [3], (d) Grad-CAM [31], (e) Grad-ECLIP [42], (f) LibraGrad [24], (g) M2IB [37], (h) MaskCLIP [10], (i) RISE [27], (j) CLIPSurgery [22], (k) MDA [36], (l) OTSMDL (Ours).

text supervision for fine-grained retinal disease representation [33]. Despite these advances, CLIP-style models remain largely black boxes: prediction probabilities alone cannot reveal whether a model prediction is supported by disease-relevant anatomical or pathological evidence. Therefore, we argue that complementing probability scores with visual evidence can enhance the interpretability of ophthalmic models by making the evidential basis of each prediction inspectable.

Existing explainability methods can struggle in complex ophthalmic scenarios, typically producing heatmaps that are overly sparse, noisy, or semantically inconsistent. Figure 1 illustrates this issue with representative baseline visualizations using a CLIP model trained on ophthalmic datasets. For glaucoma, the optic disc/cup complex and cup-to-disc morphology provide key diagnostic cues [43]. However, attention/relevance-propagation methods such as Rollout [1] and GAME [6] tend to produce sparse or fragmented activations, failing to capture cohesive anatomical structures. For the other displayed baselines, although some methods partially activate regions near the optic disc, their heatmaps generally contain substantial background responses and show weak semantic alignment with the text prompt in this case. Among the displayed baselines, LeGrad is relatively more focused on the optic disc region, whereas most others fail to consistently concentrate on the optic disc/cup complex. In contrast, OTSMDL produces a more coherent and background-suppressed activation around the optic disc/cup complex. We attribute these limitations to the shift from object-centric natural images to pathology- and anatomy-centric medical images, where anatomical biomarkers, pathological structures, and complex backgrounds exacerbate the visual-textual semantic gap, which may explain why these baselines transfer less effectively to ophthalmic scenarios than to natural-image settings.

To mitigate the above limitations, we propose OTSMDL, an optimal transport-based submodular method for selecting explainable regions to improve interpretability in ophthalmic prediction tasks. We first partition the input image into multiple sub-regions. Assuming a concise subset of key structures or lesions suffices to explain a prediction, we design a three-dimensional submodular scoring function and couple it with Optimal Transport (OT) for global region selection. This joint formulation quantifies the interpretability task to promote the

selection of fewer regions that yield high information gain, reduce redundancy, and better align with the target category. Crucially, distinct from both Chen et al.'s *Less-is-More* approach and traditional greedy submodular optimization algorithms, our method assigns a dual role to the submodular-inspired scores: initializing the top-scoring candidate pool and formulating the source mass distribution for Optimal Transport, rather than treating the greedy subset as the final explanation. This design enables subsequent transport-based re-ranking, thereby helping to mitigate the potential local biases and semantic fragmentation that may arise from purely greedy strategies. In summary, the contributions of this paper are as follows:

1. We propose a novel submodular-inspired function that scores explanation regions from three dimensions: importance, contrast, and uniqueness. These scores initialize both the candidate pool and the source mass distribution for OT, and ablation studies verify the contribution of each metric.
2. Distinct from purely greedy selection, we propose an optimal transport-based global filtering mechanism for explainable regions. By combining visual and textual cost matrices, this mechanism encourages the selection of coherent and structurally consistent regions that align with target semantics. Ablation studies on the cost matrices demonstrate that removing cost components generally weakens the overall explanation performance.
3. Evaluated on seven ophthalmic datasets and ImageNet, OTSMDL achieves favorable visual evidence grounding, with further analyses confirming prompt robustness and providing practical guidance on hyperparameter selection.

2 Related Work

Explainability Methods in Computer Vision. Early gradient/CAM-based methods, such as Grad-CAM [31], Layer-CAM [16], and Grad-CAM++ [5], localize discriminative regions through class-specific gradients, but they were originally designed for convolutional architectures and can struggle with transformer-based models [35]. To adapt attribution to CLIP and ViT models, Grad-ECLIP [42], LeGrad [3], and LibraGrad [24] further exploit image-text similarity gradients, attention-map gradients, or balanced gradient flow to improve localization. Another line of work explains transformer predictions through attention or relevance propagation, including Rollout [1] and GAME [6], but these methods may produce sparse or fragmented maps. CLIP-specific dense localization methods, such as MaskCLIP [10] and CLIP-Surgery [22], leverage or intervene in CLIP feature representations to obtain dense activation maps, while their localization may still be imprecise in fine-grained medical images. Other attribution methods, including randomized masking-based RISE [27], information-bottleneck-based M2IB [37], metric-driven MDA [36], and foveation-based FovEx [26], estimate explanations through sampling, optimization, or human-inspired foveation. Although these approaches have shown promising results in natural images, they may still generate noisy, incomplete, or semantically misaligned explanations in ophthalmic scenarios. These limitations motivate the design of OTSMDL.

Submodular Optimization in Explanatory Tasks. Submodular optimization [12] involves functions that exhibit the characteristic of "diminishing returns," where the incremental gain decreases as more elements are selected. The goal in submodular optimization is typically to select a subset of elements that maximizes utility while minimizing redundancy, thereby improving efficiency. Several studies have applied this theory to model interpretability. For instance, SP-LIME [30], introduced in the LIME framework, uses submodular pick to select representative and non-redundant instances for explaining model behavior. Furthermore, Chen et al. (*Less-is-More*) [7] integrated the submodular function with confidence, effectiveness, consistency, and collaboration scores. Compared to existing attribution methods, this approach demonstrates superior performance on datasets such as Celeb-A [23], VGG-Face2 [4], and CUB-200-2011 [38]. However, to solve the optimization problem, this method relies heavily on a sequential greedy search algorithm. While computationally efficient, purely greedy heuristics are inherently short-sighted, making selections based solely on local marginal gains at each step. Therefore, to avoid these limitations of current interpretation methods, we introduce a novel framework that integrates submodular optimization and optimal transport. Both the submodular scores and visual-textual transport costs jointly contribute to selecting explanation regions.

Optimal Transport (OT). Optimal Transport (OT) is a fundamental mathematical framework for measuring the distance between probability distributions. It was initially proposed by Gaspard Monge [25] and later refined by Leonid Kantorovich [17]. While traditional OT provides meaningful structural correspondences, its classic formulation incurs prohibitive computational costs. This bottleneck is largely alleviated by Sinkhorn's regularization [9], which enables efficient approximations. Furthermore, classical balanced OT also assumes strict preservation of marginal distributions and total mass. However, this balanced matching imposes limitations in certain applications. Empirically, a small number of critical anatomical or pathological cues can often support clinical diagnosis. Forcing a balanced transport could inadvertently disperse the matching mass across extensive irrelevant backgrounds. Therefore, in this study, we adopt Unbalanced Optimal Transport (UOT) [8], which relaxes the strict mass preservation constraint to perform soft matching.

3 Methodology

In this section, we propose an Optimal Transport-based Submodular Optimization algorithm for selecting explainable regions in the CLIP model. In Section 3.1, we introduce the partitioning of sub-regions; in Section 3.2, we present the design of the submodular functions; and in Section 3.3, we describe how the Optimal Transport algorithm is utilized to select the most explainable regions. The overall workflow of OTSMDL is shown in Figure 2.

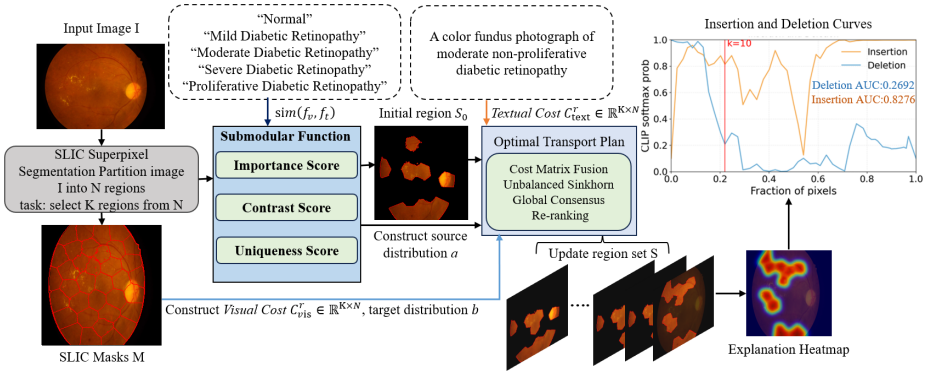

Fig. 2: Overview of the workflow of OTSMDL. We partition the input fundus image into N superpixels using the Simple Linear Iterative Clustering (SLIC) superpixel algorithm. A submodular function evaluates candidate regions based on importance, contrast, and uniqueness, initializing the target subset S_0 and the source distribution a. Then, an Optimal Transport (OT) module aligns the regions by combining visual and textual cost matrices. The OT weights w are used to select K regions, generating the explanation heatmap, evaluated by Insertion and Deletion curves.

3.1 Superpixel Region Segmentation

The Simple Linear Iterative Clustering (SLIC) superpixel algorithm is a k-means clustering-based method [2] with two key features: First, it reduces the computational complexity to linear by limiting the search region for distance calculations, significantly enhancing the speed compared to traditional k-means methods. Second, it employs a weighted distance metric that combines color and spatial proximity, allowing for control over the size and compactness of the superpixels. Recent works have also utilized SLIC superpixels for processing medical images [21], primarily due to SLIC's superior boundary adherence, which enables it to follow image boundaries well and accurately capture the structural boundaries of the image. In our work, we use SLIC superpixels to process retinal images, dividing them into multiple subregions for interpretability tasks. Compared to patch-level segmentation, SLIC better adheres to the boundary information of feature objects, especially for structures with highly irregular edges, such as exudates and blood vessels. Specifically, given an input image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$, we first perform SLIC-based over-segmentation to partition it into a set of N disjoint, boundary-aware superpixels, denoted as $\mathcal{R} = \{R_i\}_{i=1}^{N}$. For each region R_i, we define a spatial binary mask $m_i \in \{0,1\}^{H\times W}$ and extract the corresponding candidate sub-image $\mathbf{I}_i$ via:

$$m_i(u,v) = \begin{cases} 1, & \text{if } (u,v) \in R_i \\ 0, & \text{otherwise} \end{cases}, \quad \mathbf{I}_i = \mathbf{I} \odot m_i, \tag{1}$$

where $\odot$ denotes element-wise multiplication with channel-wise broadcasting. Consequently, we obtain the candidate pool of visual elements $M = \{\mathbf{I}_i\}_{i=1}^{N}$. As

shown in Figure 2, the SLIC Masks M represent the set of candidate regions, which will be used to select the explanatory regions in subsequent steps.

3.2 Submodular Function Design

In this section, we detail the design of the submodular functions used to score the regions selected for explanation. Firstly, we define a scoring function: the relevance of regions to the target class (importance), their contrast with other classes (contrast), and their diversity within the image (uniqueness). These criteria are integrated to guide the initial region selection. After obtaining the candidate pool M, let $f_i \in \mathbb{R}^d$ denote the visual feature of region m_i extracted by the image encoder $F(\cdot)$. Given a target class index y and its text embedding t_y, we sequentially define the three utility components of region i as follows.

Contrast Score. To enhance the distinction between the target class and other competing classes, and to suppress class-ambiguous or overlapping regions, we define the Contrast Score by measuring the difference in similarity between the region's visual feature and the target class's text embedding, compared to the most similar competing class. This metric primarily targets complex scenes, isolating the target class from multiple objects. Additionally, for ordinal tasks like Diabetic Retinopathy (DR) with ambiguous adjacent grades, this mechanism encourages focusing on target-specific pathological evidence. Specifically, this score is calculated as:

$$s_i^{\text{con}} = \Big[\cos(f_i, t_y) - \max_{c \neq y} \cos(f_i, t_c) \Big]_+^2, \tag{2}$$

Here, $[\cdot]_+$ denotes the positive part; this term is high only when region i aligns more with the target class y than any competing class c, promoting discriminative regions while suppressing ambiguous ones.

Importance Score. We quantify target relevance by combining direct region-text similarity and a deletion-based contribution gap. The direct region-text similarity measures how well a superpixel region matches the target class by calculating the cosine similarity between the visual feature of the region f_i and the target class text embedding t_y. The deletion-based contribution gap further evaluates the importance of the region by measuring the change in image similarity to the target class when the region m_i is removed from the image. Specifically, the gap g_i quantifies how much the alignment with the target class decreases when the region is masked out, where the greater the reduction in similarity, the more important the region is. This is computed as:

$$g_i = \max\Big(0, \cos(F(I), t_y) - \cos(F(I \setminus m_i), t_y)\Big), \tag{3}$$

where $I \setminus m_i$ denotes the image with region m_i masked out. By combining the direct similarity and the contribution gap, the overall Importance Score is computed as:

$$s_i^{\text{imp}} = \cos(f_i, t_y) + g_i. \tag{4}$$

Uniqueness Score. Besides target relevance, we also consider the visual distinctiveness of each candidate region. We operate under the assumption that in retinal images, areas containing potential pathological abnormalities (such as hemorrhages, hard exudates, or neovascularization) constitute a minority, and their visual distributions typically deviate from the background norm. Since the CLIP image encoder naturally maps visually homogeneous background patches into a dense cluster in the latent space, rare pathological lesions act as spatial outliers. To mathematically measure this feature deviation and highlight these salient regions, the uniqueness score is computed as:

$$s_i^{\text{uni}} = \|f_i - \bar{f}\|_2^2, \quad \bar{f} = \frac{1}{N}\sum\nolimits_{j=1}^{N} f_j, \tag{5}$$

where $\bar{f}$ represents the mean visual embedding across all N candidate regions. By calculating the squared L_2 distance between a region's feature f_i and the global average, this formulation explicitly assigns higher scores to spatial outliers. After computing the individual criteria for all candidate regions, we integrate them into a unified base utility score s_i:

$$s_i = \lambda_{\text{imp}} s_i^{\text{imp}} + \lambda_{\text{con}} s_i^{\text{con}} + \lambda_{\text{uni}} s_i^{\text{uni}}. \tag{6}$$

where $\lambda_{\text{imp}}, \lambda_{\text{con}}, \lambda_{\text{uni}} \geq 0$ are the weighting coefficients. Prior to the subsequent selection process, the utility score for each candidate superpixel is pre-computed and cached, effectively avoiding any redundant computation. Finally, we apply standard min-max normalization to yield the normalized base utility $\hat{s}_i \in [0, 1]$.

Diminishing Marginal Returns Construction. A fundamental property of submodular optimization is the effect of diminishing marginal returns, which inherently models objectives requiring coverage and diversity. Leveraging this property to explicitly suppress redundant regions, we first define a pairwise redundancy kernel to quantify the visual similarity between any two regions:

$$K_{ij} = \frac{1 + \cos(f_i, f_j)}{2}, \quad K_{ii} = 0. \tag{7}$$

This mapping bounds the redundancy penalty K_{ij} within $[0, 1]$, and setting $K_{ii} = 0$ explicitly avoids self-penalization. During the subset selection process, the marginal gain of adding a candidate region i to the currently selected set S is calculated by subtracting its redundancy penalty from the normalized base utility $\hat{s}_i$:

$$\Delta(i \mid S) = \hat{s}_i - \lambda_{\text{dr}} \sum_{j \in S} K_{ij}, \tag{8}$$

where λ_{dr} controls the penalty strength. Through this objective, a candidate's score dynamically decreases if it shares high visual similarity with regions already present in S. Finally, we iteratively select the region $i^* = \arg\max_{i \notin S} \Delta(i \mid S)$ until K regions are collected. This yields the subset S_0, which may only be a local optimum and will serve merely as an initialization set for global alignment in the subsequent optimal transport module.

3.3 Explainable Region Selection

The construction of diminishing marginal returns in Eq. 8 generates an initial subset S_0 shown in Fig. 2, with a budget of K regions. However, S_0 is only a score-guided initialization and still lacks a global matching perspective. Therefore, we use S_0 to initialize the Optimal Transport (OT) process and construct its source distribution, allowing all candidate regions to compete under a unified visual-textual cost criterion for more consistent region-level interpretations.

Starting from $S^{(0)} = S_0$, we iteratively update the selected set until it stabilizes or the maximum number of iterations $R_{\max}$ is reached. For each iteration r, let $S^{(r-1)} = u_1, \ldots, u_K$ be the current selected set. We formulate the OT problem by transporting mass from a source distribution $a^{(r)}$ defined over $S^{(r-1)}$ to a target distribution b defined over all N candidates. Specifically, $b = \frac{1}{N}\mathbf{1} \in \mathbb{R}^N$ acts as a uniform prior, assigning equal weight to each candidate. For the source distribution, the mass of each selected region is proportional to its base utility, normalized within the current subset, such that $a_i^{(r)} = \hat{s}u_i / \sum j = 1^K \hat{s}_{u_j}$ for $i = 1, \ldots, K$. To guide this global alignment, we construct an OT cost matrix that combines two complementary components: a visual term measuring feature discrepancy in the CLIP image space, and a text-guided term measuring semantic mismatch with respect to the target class embedding t_y.

Visual Cost. For a source index $u \in S^{(r-1)}$ and a candidate index $j \in \{1, \ldots, N\}$, we define the visual cost as:

$$C_{\text{vis}}^{(r)}(u, j) = \|f_u - f_j\|_2 + \|f_{u \cup j} - f_{S^{(r-1)}}\|_2, \tag{9}$$

where $f_{u \cup j}$ is the feature of the merged region pair (u, j), and $f_{S^{(r-1)}}$ is the feature of the image composed by the current selected set. This visual cost balances local similarity and global structural cohesion: the first term aligns visually similar regions, while the second penalizes deviations from the selected subset's consensus.

Text Cost. We define the target-aware similarities for an individual region i and a merged pair (u, j) as $\text{sim}_i = \cos(f_i, t_y)$ and $\text{sim}_{u \cup j} = \cos(f_{u \cup j}, t_y)$, respectively. Using these cosine similarities, the text-guided pairwise cost is formulated to penalize both low target alignment and semantic inconsistency between the source and candidate regions:

$$C_{\text{text}}^{(r)}(u, j) = (1 - \text{sim}_u) + (1 - \text{sim}_j) + (1 - \text{sim}_{u \cup j}) + |\text{sim}_u - \text{sim}_j|. \tag{10}$$

Notably, the extraction of visual and text features is uniformly performed via a single batched forward pass for all N candidate regions, and the results are globally cached. To alleviate the computational complexity, for the merged region pairs (u, j), we precompute their representations for the initial subset S_0 prior to the OT loop. During the iterative optimization process, we adopt an incremental update strategy. If the selected set $S^{(r-1)}$ updates, the computation is limited to the newly introduced regions. Experimental validation regarding computational resource consumption is presented in Section 4.2.

Algorithm 1: OT-based Explainable Region Selection

Input: Candidate pool $M = \{I_i\}_{i=1}^{N}$, cached region features $\mathcal{F}$, target text feature t_y, normalized base scores $\hat{\mathbf{s}}$, region budget K, and hyperparameters $\Theta_{\text{sub}}, \Theta_{\text{OT}}$.

Output: Selected set $S^\star \subseteq M$, $|S^\star| = K$, and transport weights $\mathbf{w}^\star$.

$S^{(0)} \leftarrow \text{Submodular}(M, \hat{\mathbf{s}}, K; \Theta_{\text{sub}})$; `// Initialize via Eq. 8`

Prepare cached unary and merged-pair features for $S^{(0)}$ with incremental updates;

stable $\leftarrow 0$;

for $r = 1$ **to** $R_{\max}$ **do**

  Build marginals $a^{(r)}$ and b from $S^{(r-1)}$, M, and $\hat{\mathbf{s}}$; `// Score-guided source, uniform target`

  Build OT cost matrix $C^{(r)}$ using $S^{(r-1)}$, M, $\mathcal{F}$, and t_y ; `// Eqs. 9-11`

  $\Pi^{(r)} \leftarrow \text{UnbalancedSinkhorn}(a^{(r)}, b, C^{(r)}; \Theta_{\text{OT}})$;

  $\mathbf{w}^{(r)} \leftarrow \text{Normalize}(\text{ColSum}(\Pi^{(r)}))$; `// Eq. 12`

  $S^{(r)} \leftarrow \text{TopK}(\mathbf{w}^{(r)}, K)$; `// Eq. 13`

  if $S^{(r)} == S^{(r-1)}$ **then**

    stable $\leftarrow$ stable $+ 1$;

  else

    stable $\leftarrow 0$;

  if stable $\geq p$ **then**

    break ; `// Terminate if unchanged for` $p = 5$ `iterations`

$S^\star \leftarrow S^{(r)}, \quad \mathbf{w}^\star \leftarrow \mathbf{w}^{(r)}$;

return $S^\star, \mathbf{w}^\star$;

Optimal Transport Alignment. Having defined the individual cost components, we fuse them into a rectangular OT cost matrix $C^{(r)} \in \mathbb{R}^{K \times N}$, whose rows correspond to the K regions in $S^{(r-1)}$ and columns to all N candidate regions:

$$C^{(r)} = \alpha C_{\text{vis}}^{(r)} + \beta C_{\text{text}}^{(r)}, \tag{11}$$

where α and β are balancing coefficients. As defined previously, the source marginal $a^{(r)}$ is constructed using the normalized base utilities $\hat{s}_i$. This design gives high-scoring regions greater transport priority, so the OT process is guided by the importance prior rather than a uniform prior. We then solve an unbalanced Sinkhorn problem to obtain the transport plan $\Pi^{(r)}$. This design is motivated by the observation that effective diagnostic evidence in medical imaging is typically scarce. If forced to balance, transported mass may be diluted across a large number of background areas. Unbalanced OT relaxes this strict marginal constraint, encouraging soft matching of local features to global features. Subsequently, the column-wise accumulated transported mass is computed as a global importance signal over candidates:

$$w_i^{(r)} = \frac{\sum_u \Pi_{u,i}^{(r)}}{\sum_j \sum_u \Pi_{u,j}^{(r)}}, \quad i = 1, \ldots, N. \tag{12}$$

Intuitively, the column-sum weight $w_i^{(r)}$ quantifies the degree to which a candidate region i receives global consensus from the source regions. This formulation essentially serves as a globally consistent re-ranking mechanism over all candidates. The selected set is then updated by extracting the regions with the aggregated transport weights:

$$S^{(r)} = \arg \max_{S \subseteq M, |S|=K} \sum_{i \in S} w_i^{(r)}. \tag{13}$$

This subset is extracted directly from the globally stabilized transport distribution. The final explainable region set upon convergence is denoted as $S^\star$. Section 4.1 reports the implementation details and practical hyperparameter choices. To empirically validate our formulation choices, Section 4.2 includes detailed ablation studies, specifically evaluating the score-guided source distribution against a uniform prior, and unbalanced OT (solved via Sinkhorn) against its balanced counterpart.

Overall, the OT-based region selection process is summarized in Algorithm 1. Starting from the submodularly initialized subset, OTSMDL iteratively formulates a score-guided transport problem between the current selected regions and the full candidate pool. The column-wise transported mass is then converted into global region weights, from which the top-K regions are updated until convergence or the maximum iteration budget is reached. Therefore, the final explanation $S^\star$ is obtained from the OT-refined consensus over all candidates, rather than from the greedy initialization alone.

4 Experiments

4.1 Data and Settings

Datasets. We evaluate our method on seven ophthalmic datasets spanning color fundus photography and OCT imaging across DR, macular disease, and glaucoma settings. For color fundus photography, we use three diabetic retinopathy (DR) datasets: APTOS 2019 [34] (3662 images), EyePACS [11] (training set: 35,126 images, testing set: 53,576 images), and IDRiD [28] (516 images). All three datasets are annotated with DR severity levels from 0 to 4, covering normal, mild, moderate, severe, and proliferative retinopathy. For OCT imaging, we use OCTID [13] (518 images), OCTDL [19] (2064 images), and OIMHS [40] (3859 images). OCTID and OCTDL are multi-disease OCT datasets with four and seven categories, respectively, covering representative pathologies such as macular hole, central serous retinopathy, diabetic retinopathy, age-related macular degeneration, diabetic macular edema, epiretinal membrane, retinal artery/vein occlusion, and vitreomacular interface disease. OIMHS is an OCT macular-hole dataset with manual segmentation annotations and is used for mask-based consistency evaluation. For glaucoma analysis, we use PAPILA [18] (488 images). To obtain the ophthalmic CLIP backbone, we train CLIP [29] on DeepEyeNet [15], MM-Retinal-v1 [39], and FFA-IR [20]. ImageNet [14] results are reported in the supplementary material for out-of-domain faithfulness evaluation.

Baselines. We adopt the following methods as baselines for the comparative experiments: raw attention, Grad-ECLIP [42], M2IB [37], MaskCLIP [10], CLIP-Surgery [22], Rollout [1], Grad-CAM [31], GAME [6], Less-is-More [7], RISE [27], LeGrad [3], LibraGrad [24], and MDA [36].

Evaluation Metric. To assess the faithfulness of the explanations, we calculate the Area Under the Curve (AUC) for the standard Insertion and Deletion protocols. Unless otherwise specified, all AUC scores reported in the main text are computed using the ground-truth class as the target. Furthermore, to explicitly quantify spatial localization accuracy against clinical ground-truth annotations, we employ the Mean Dice coefficient and Mean Intersection over Union (IoU). In most tables, the best and second-best results are highlighted in **bold** and underlined, respectively.

Implementation Details. We use CLIP (ViT-B/32) as the backbone model trained on ophthalmic data in our experiments. All images are resized to 224×224, and all experiments are conducted on a single NVIDIA GeForce RTX 5080 GPU. The SLIC superpixel size is set to 30, partitioning each image into 49 subregions. For AUC computation, we use a uniform black image as the constant masking baseline, and the pixel ratio replaced at each step corresponds to the proportion of pixels within a superpixel region. For the unbalanced Sinkhorn algorithm, the numerical convergence tolerance is set to 10^{-6}, the entropy regularization strength to 0.05, and the mass-relaxation coefficient to 1.0. For the OT module, unless specified otherwise, the maximum number of OT iterations $R_{\max}$ is set to 5. The cost weights are set to $\alpha = 0.1$ and $\beta = 0.6$ in Eq. 11; grid search shows that they operate on different numerical scales, with favorable ranges of $\alpha \in [0.06, 0.1]$ and $\beta \in [0.5, 0.8]$ on ophthalmic data. The weights $(\lambda_{\text{imp}}, \lambda_{\text{con}}, \lambda_{\text{uni}})$ of the submodular scoring function in Eq. 6 are all set to 0.8. Since the final utility score is min-max normalized before selection, the relative ratios of these weights are more important than their absolute scales. Further ablations on key hyperparameters are presented in Section 4.2.

4.2 Experiment Results

Comparison Results on Ophthalmic Datasets. As reported in Table 1, our method achieves the best Insertion AUC and the second-best Deletion AUC across all three diabetic retinopathy (DR) datasets. Similarly, in Table 2, our method achieves the best Insertion AUC on OCTDL and PAPILA, while ranking third on OCTID. For Deletion AUC, OTSMDL ranks first on OCTDL and OCTID and second on PAPILA. Notably, except on OCTID, OTSMDL significantly outperforms all baselines in Insertion AUC across the remaining ophthalmic datasets ($p < 0.05$). For Deletion AUC, OTSMDL remains competitive with *Less-is-More*. Overall, OTSMDL shows stable performance across different datasets and imaging modalities. In particular, its stronger Insertion AUC in most settings indicates that the selected regions provide more faithful evidence for progressively recovering the target prediction. These results support the effectiveness of combining submodular-inspired scoring with transport-based global

Table 1: Comparison of explanation methods on three DR datasets, using ground-truth labels as the target class, evaluated by Deletion and Insertion AUC scores.

Methods	IDRiD [28]		APTOS2019 [34]		EYEPACS [11]	
	Deletion (↓)	Insertion (↑)	Deletion (↓)	Insertion (↑)	Deletion (↓)	Insertion (↑)
Grad-ECLIP [42]	0.397	0.427	0.361	0.399	0.159	0.237
GAME [6]	0.357	0.305	0.363	0.226	0.151	0.123
Grad-CAM [31]	0.493	0.300	0.471	0.287	0.243	0.128
M2IB [37]	0.346	0.326	0.321	0.292	0.159	0.137
MaskCLIP [10]	0.465	0.323	0.479	0.291	0.223	0.139
RISE [27]	0.244	<u>0.583</u>	0.306	<u>0.444</u>	0.192	0.229
Rollout [1]	0.356	0.286	0.347	0.236	0.169	0.113
raw attention	0.533	0.305	0.509	0.272	0.237	0.103
Less-is-More [7]	**0.224**	0.450	**0.209**	0.380	**0.113**	0.179
CLIPSurgery [22]	0.329	0.413	0.337	0.281	0.138	0.182
LeGrad [3]	0.350	0.393	0.402	0.365	0.139	<u>0.245</u>
LibraGrad [24]	0.311	0.400	0.351	0.411	0.148	0.216
MDA [36]	0.256	0.452	0.305	0.428	0.128	0.239
Ours	<u>0.226</u>	**0.624**	<u>0.232</u>	**0.484**	<u>0.122</u>	**0.253**

Table 2: Comparison of ophthalmic OCT and glaucoma datasets, using ground-truth labels as the target class, evaluated by Deletion and Insertion AUC scores.

Methods	OCTDL [19]		OCTID [13]		PAPILA [18]	
	Deletion (↓)	Insertion (↑)	Deletion (↓)	Insertion (↑)	Deletion (↓)	Insertion (↑)
Grad-ECLIP [42]	0.307	0.208	0.433	**0.511**	0.230	0.702
GAME [6]	0.235	0.208	0.400	0.265	0.280	0.577
Grad-CAM [31]	0.339	0.192	0.449	0.480	0.256	0.573
M2IB [37]	0.259	0.364	0.425	0.317	0.208	0.622
MaskCLIP [10]	0.362	0.220	0.417	0.432	0.353	0.599
RISE [27]	<u>0.184</u>	<u>0.500</u>	<u>0.339</u>	0.236	0.158	0.573
Rollout [1]	0.348	0.222	0.343	0.257	0.365	0.553
raw attention	0.237	0.257	0.457	0.376	0.229	0.563
Less-is-More [7]	0.221	0.337	0.379	<u>0.506</u>	**0.089**	0.678
CLIPSurgery [22]	0.283	0.241	0.438	0.372	0.295	0.628
LeGrad [3]	0.245	0.365	0.402	0.433	0.190	<u>0.721</u>
LibraGrad [24]	0.291	0.263	0.378	0.440	0.250	0.714
MDA [36]	0.186	0.365	0.406	0.476	0.149	0.690
Ours	**0.164**	**0.530**	**0.334**	0.482	<u>0.148</u>	**0.725**

refinement for faithful region attribution. Experiments using CLIP's zero-shot top-1 prediction as the target label are reported in the supplementary material.

Computational Cost. The theoretical complexity of OTSMDL is $\mathcal{O}(NC_{\text{enc}} + R_{\max}(KN + TKN))$; C_{enc} denotes the cost of a CLIP encoding pass and T denotes the number of Sinkhorn iterations. We benchmark the runtime and hardware footprint of OTSMDL, with detailed results reported in the supplementary material. With batched forward passes, dense GPU feature caching, and feature reuse, OTSMDL requires 5.28 s/img on average. Its runtime is comparable to CLIPSurgery and substantially faster than Less-is-More, while maintaining lightweight resource usage, with ≈6.4 GB peak GPU memory and lower average GPU utilization than CLIPSurgery (≈50.3% vs. ≈76%).

Table 3: Comparison of spatial consistency and localization accuracy across different pixel insertion ratios on IDRiD [28].

Methods	Ratio = 0.25		Ratio = 0.50		Ratio = 0.75		Ratio = 1.00	
	mDice (↑)	mIoU (↑)	mDice (↑)	mIoU (↑)	mDice (↑)	mIoU (↑)	mDice (↑)	mIoU (↑)
Grad-ECLIP [42]	<u>0.137</u>	<u>0.077</u>	0.209	0.122	0.279	<u>0.171</u>	0.299	0.186
GAME [6]	0.091	0.049	0.169	0.095	0.208	0.119	0.219	0.127
Grad-CAM [31]	0.104	0.058	0.191	0.113	0.222	0.131	0.261	0.157
M2IB [37]	0.094	0.051	0.180	0.103	0.215	0.124	0.240	0.141
MaskCLIP [10]	0.129	0.074	0.205	0.118	0.242	0.144	0.282	0.173
RISE [27]	0.095	0.053	0.096	0.054	0.093	0.052	0.100	0.057
Rollout [1]	0.109	0.061	0.130	0.073	0.157	0.089	0.155	0.088
raw attention	0.044	0.023	0.059	0.032	0.062	0.034	0.074	0.040
Less-is-More [7]	**0.165**	**0.093**	**0.258**	**0.151**	<u>0.281</u>	0.167	0.307	0.185
CLIPSurgery [22]	0.083	0.046	0.135	0.076	0.189	0.111	0.225	0.134
LeGrad [3]	0.072	0.067	0.221	0.124	0.264	0.166	0.274	0.173
LibraGrad [24]	0.109	0.055	0.219	0.139	0.273	0.161	<u>0.314</u>	<u>0.198</u>
MDA [36]	0.130	0.059	0.208	0.133	0.246	0.150	0.246	0.149
Ours	0.131	0.073	<u>0.235</u>	<u>0.140</u>	**0.295**	**0.181**	**0.323**	**0.200**

Key Structural Consistency Assessment. To assess spatial alignment with ground-truth annotations, we use IDRiD and OIMHS for mask-based consistency evaluation. For IDRiD, lesion masks are formed by merging annotations of microaneurysms, hemorrhages, hard exudates, and soft exudates. For OIMHS, we use the union of non-background manual OCT annotations as the ground-truth foreground structural mask. As shown in Tables 3 and 4, OTSMDL achieves favorable localization consistency on both datasets. On IDRiD, although *Less-is-More*'s greedy subset selection better captures the most salient early focus, OTSMDL improves with increasing insertion ratios, obtaining four best and two second-best results among eight metrics. This indicates that OT-based refinement improves mask alignment beyond the initially salient regions. On OIMHS, OTSMDL achieves the best or tied-best results in six of eight metrics and ranks second in the remaining two, showing that the same region-selection strategy also applies to OCT structural-mask evaluation.

Ablation Study. We conduct ablation experiments on the IDRiD and OCTDL datasets to evaluate the key components of OTSMDL. Table 5 shows that removing both visual and textual transport costs reduces the framework to a purely submodular selection scheme, leading to the weakest overall Deletion/Insertion performance. Combining both costs yields the best overall results, with the textual transport cost playing a particularly important role. Table 6 further analyzes the submodular scores and OT variants. The scoring ablation shows that Importance (s^{imp}) mainly improves Deletion AUC, while Uniqueness (s^{uni}) contributes more to Insertion AUC. Although using all three scores does not always produce the lowest Deletion AUC, it provides a balanced trade-off and achieves the strongest Insertion performance. Even when all three score terms are removed, thereby disabling effective score-driven submodular guidance, the results remain competitive, suggesting that the OT process itself plays a key role in region explanation. Crucially, replacing unbalanced OT with balanced OT causes

Table 4: Comparison of spatial consistency and localization accuracy across different pixel insertion ratios on OIMHS [40].

Methods	Ratio = 0.25		Ratio = 0.50		Ratio = 0.75		Ratio = 1.00	
	mDice (↑)	mIoU (↑)	mDice (↑)	mIoU (↑)	mDice (↑)	mIoU (↑)	mDice (↑)	mIoU (↑)
Grad-ECLIP [42]	0.085	0.048	0.146	0.084	0.161	0.093	0.161	0.093
GAME [6]	0.323	0.214	0.423	0.277	0.424	0.278	0.425	0.278
Grad-CAM [31]	0.255	0.152	0.389	0.251	0.434	0.290	0.439	0.294
M2IB [37]	0.330	0.210	0.433	0.300	<u>0.525</u>	**0.367**	**0.599**	**0.444**
MaskCLIP [10]	0.065	0.036	0.118	0.067	0.134	0.076	0.135	0.077
RISE [27]	0.114	0.063	0.122	0.068	0.122	0.068	0.123	0.069
Rollout [1]	0.338	0.206	0.428	0.278	0.431	0.281	0.431	0.281
raw attention	0.201	0.116	0.295	0.181	0.319	0.199	0.320	0.199
Less-is-More [7]	0.081	0.043	0.079	0.042	0.099	0.055	0.184	0.110
CLIPSurgery [22]	<u>0.340</u>	<u>0.219</u>	<u>0.451</u>	<u>0.315</u>	0.517	0.358	0.550	0.365
LeGrad [3]	0.147	0.083	0.265	0.160	0.304	0.187	0.306	0.189
LibraGrad [24]	0.089	0.049	0.130	0.074	0.143	0.081	0.144	0.081
MDA [36]	0.148	0.086	0.250	0.154	0.301	0.189	0.307	0.193
Ours	**0.347**	**0.222**	**0.465**	**0.323**	**0.526**	**0.367**	<u>0.557</u>	<u>0.385</u>

Table 5: Ablation study of visual and text transport cost terms, evaluated by Deletion (Del.) and Insertion (Ins.) AUC scores.

Cost Terms		IDRiD [28]		OCTDL [19]	
Visual Transport Cost (α)	Text Transport Cost (β)	Del. (↓)	Ins. (↑)	Del. (↓)	Ins. (↑)
×	✓	**0.226**	<u>0.601</u>	<u>0.168</u>	<u>0.528</u>
✓	×	<u>0.274</u>	0.487	0.203	0.359
×	×	0.289	0.406	0.248	0.336
✓	✓	**0.226**	**0.624**	**0.164**	**0.530**

Table 6: Ablation study of the submodular function scores and OT variants.

Score Terms			IDRiD [28]		OCTDL [19]	
Importance (s^{imp})	Contrast (s^{con})	Uniqueness (s^{uni})	Del. (↓)	Ins. (↑)	Del. (↓)	Ins. (↑)
×	×	×	0.233	0.566	0.173	0.506
✓	×	×	<u>0.221</u>	0.524	0.165	0.510
×	✓	×	0.255	0.519	0.193	0.517
×	×	✓	0.231	<u>0.612</u>	0.173	0.521
✓	✓	×	0.227	0.524	**0.161**	**0.530**
✓	×	✓	**0.220**	0.603	0.165	0.528
×	✓	✓	0.237	0.609	0.175	0.523
Balanced OT (full scores)			0.313	0.426	0.293	0.346
a = uniform (full scores)			0.225	0.603	0.167	<u>0.529</u>
Default Settings (full scores)			0.226	**0.624**	<u>0.164</u>	**0.530**

clear performance degradation, while replacing the score-guided source distribution with a uniform prior weakens the overall results. These findings validate the necessity of both the visual-textual transport costs and the score-guided unbalanced OT design.

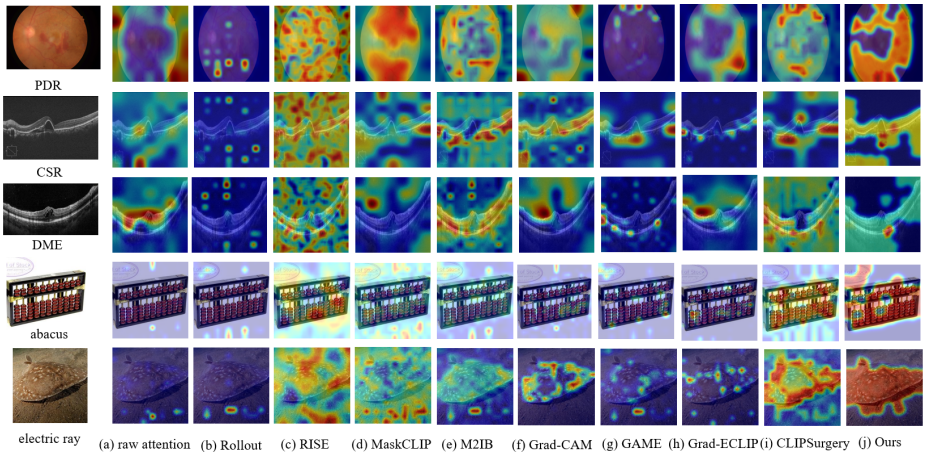

Fig. 3: Comparison of heat maps from: (a) raw attention; (b) Rollout [1]; (c) RISE [27]; (d) MaskCLIP [10]; (e) M2IB [37]; (f) Grad-CAM [31]; (g) GAME [6]; (h) Grad-ECLIP [42]; (i) CLIPSurgery [22]; (j) Ours (OTSMDL). The medical target classes correspond to Proliferative Diabetic Retinopathy (PDR), Central Serous Retinopathy (CSR), and Diabetic Macular Edema (DME).

Visualization Results. We visualize the top-K transport weights (w) in Eq. 12, with heatmaps from several baselines shown for reference in Fig. 3. Qualitative results indicate that OTSMDL is able to consistently localize target-related regions while suppressing background activation. Additional visualizations in the supplementary material further analyze how the region budget K affects the resulting heatmaps and discuss representative limitations of OTSMDL.

5 Conclusion

In this work, we proposed OTSMDL, a post-hoc region selection framework for interpretable ophthalmic CLIP. By combining submodular-inspired scoring with unbalanced optimal transport, OTSMDL globally re-ranks candidate superpixels to produce coherent and target-aligned visual evidence. Experiments on seven ophthalmic datasets demonstrate favorable region-level evidence grounding, achieving top-1 or top-2 performance against 13 baselines in most settings, while additional ImageNet results show competitive out-of-domain performance. Ablation studies further validate the effectiveness of the proposed scoring components and visual-textual transport costs. Future work will explore finer-grained localization and adaptive region budgeting.

Acknowledgements

This work was supported in part by the National Natural Science Fund of China (No. 62276135).

References

1. Abnar, S., Zuidema, W.: Quantifying attention flow in transformers. In: Annual Meeting of the Association for Computational Linguistics (2020), https://api.semanticscholar.org/CorpusID:218487351
2. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S.: Slic superpixels compared to state-of-the-art superpixel methods. IEEE transactions on pattern analysis and machine intelligence **34**(11), 2274–2282 (2012)
3. Bousselham, W., Boggust, A., Chaybouti, S., Strobelt, H., Kuehne, H.: Legrad: An explainability method for vision transformers via feature formation sensitivity. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 20336–20345 (2025)
4. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: Vggface2: A dataset for recognising faces across pose and age. In: 2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018). pp. 67–74. IEEE (2018)
5. Chattopadhay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N.: Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 839–847 (2018). https://doi.org/10.1109/WACV.2018.00097
6. Chefer, H., Gur, S., Wolf, L.: Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 387–396 (2021). https://doi.org/10.1109/ICCV48922.2021.00045
7. Chen, R., Zhang, H., Liang, S., Li, J., Cao, X.: Less is more: Fewer interpretable region via submodular subset selection. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=jKTUlxo5zy
8. Chizat, L., Peyré, G., Schmitzer, B., Vialard, F.X.: Unbalanced optimal transport: Dynamic and kantorovich formulations. Journal of Functional Analysis **274**(11), 3090–3123 (2018)
9. Cuturi, M.: Sinkhorn distances: Lightspeed computation of optimal transport. Advances in neural information processing systems **26** (2013)
10. Dong, X., Bao, J., Zheng, Y., Zhang, T., Chen, D., Yang, H., Zeng, M., Zhang, W., Yuan, L., Chen, D., Wen, F., Yu, N.: Maskclip: Masked self-distillation advances contrastive language-image pretraining. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10995–11005 (2023). https://doi.org/10.1109/CVPR52729.2023.01058
11. (EyePACS), E.P.A.C.S.: Eyepacs diabetic retinopathy dataset. Kaggle Dataset: https://www.kaggle.com/datasets/ascanipek/eyepacs-aptos-messidor-diabetic-retinopathy (2015), approximately 88,700 fundus images labeled with DR severity (0–4), Accessed: 2026-02-16
12. Fujishige, S.: Submodular functions and optimization, vol. 58. Elsevier (2005)
13. Gholami, P., Roy, P., Parthasarathy, M.K., Lakshminarayanan, V.: Octid: Optical coherence tomography image database. Computers & Electrical Engineering **81**, 106532 (2020)
14. Howard, A., Park, E., Kan, W.: Imagenet object localization challenge. https://kaggle.com/competitions/imagenet-object-localization-challenge (2018), kaggle
15. Huang, J.H., Yang, C.H.H., Liu, F., Tian, M., Liu, Y.C., Wu, T.W., Lin, I., Wang, K., Morikawa, H., Chang, H., et al.: Deepopht: medical report generation for retinal

images via deep models and visual explanation. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 2442–2452 (2021)
16. Jiang, P.T., Zhang, C.B., Hou, Q., Cheng, M.M., Wei, Y.: Layercam: Exploring hierarchical class activation maps for localization. IEEE Transactions on Image Processing **30**, 5875–5888 (2021). https://doi.org/10.1109/TIP.2021.3089943
17. Kantorovich, L.V.: On a problem of monge. Journal of Mathematical Sciences **4**(133), 1383–1383 (2006)
18. Kovalyk, O., Morales-Sánchez, J., Verdú-Monedero, R., Sellés-Navarro, I., Palazón-Cabanes, A., Sancho-Gómez, J.L.: Papila: Dataset with fundus images and clinical data of both eyes of the same patient for glaucoma assessment. Scientific Data **9**(1), 291 (2022)
19. Kulyabin, M., Zhdanov, A., Nikiforova, A., Stepichev, A., Kuznetsova, A., Ronkin, M., Borisov, V., Bogachev, A., Korotkich, S., Constable, P.A., et al.: Octdl: Optical coherence tomography dataset for image-based deep learning methods. Scientific data **11**(1), 365 (2024)
20. Li, M., Cai, W., Liu, R., Weng, Y., Zhao, X., Wang, C., Chen, X., Liu, Z., Pan, C., Li, M., et al.: Ffa-ir: Towards an explainable and reliable medical report generation benchmark. In: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2) (2021)
21. Li, S., Zhao, J., Liu, S., Dai, X., Zhang, C., Song, Z.: Sp3: Superpixel-propagated pseudo-label learning for weakly semi-supervised medical image segmentation. arXiv preprint arXiv:2411.11636 (2024)
22. Li, Y., Wang, H., Duan, Y., Li, X.: A closer look at the explainability of contrastive language-image pre-training. Pattern Recognit. **162**, 111409 (2023), https://api.semanticscholar.org/CorpusID:258079047
23. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of the IEEE international conference on computer vision. pp. 3730–3738 (2015)
24. Mehri, F., Baghshah, M.S., Pilehvar, M.T.: Libragrad: Balancing gradient flow for universally better vision transformer attributions. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 67–78 (2025)
25. Monge, G.: Mémoire sur la théorie des déblais et des remblais. Mem. Math. Phys. Acad. Royale Sci. pp. 666–704 (1781)
26. Panda, M.P., Tiezzi, M., Vilas, M., Roig, G., Eskofier, B.M., Zanca, D.: Fovex: Human-inspired explanations for vision transformers and convolutional neural networks. International Journal of Computer Vision **133**(10), 7437–7459 (2025)
27. Petsiuk, V., Das, A., Saenko, K.: Rise: Randomized input sampling for explanation of black-box models. ArXiv **abs/1806.07421** (2018), https://api.semanticscholar.org/CorpusID:49324724
28. Porwal, P., Pachade, S., Kamble, R., Kokare, M., Deshmukh, G., Sahasrabuddhe, V., Meriaudeau, F.: Indian diabetic retinopathy image dataset (idrid): a database for diabetic retinopathy screening research. Data **3**(3), 25 (2018)
29. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
30. Ribeiro, M.T., Singh, S., Guestrin, C.: " why should i trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. pp. 1135–1144 (2016)

31. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 618–626 (2017). https://doi.org/10.1109/ICCV.2017.74
32. Shi, D., Zhang, W., Yang, J., Huang, S., Chen, X., Xu, P., Jin, K., Lin, S., Wei, J., Yusufu, M., et al.: A multimodal visual–language foundation model for computational ophthalmology. npj digital medicine, 8, 381 (2025)
33. Silva-Rodriguez, J., Chakor, H., Kobbi, R., Dolz, J., Ayed, I.B.: A foundation language-image model of the retina (flair): Encoding expert knowledge in text supervision. Medical Image Analysis **99**, 103357 (2025)
34. Society, A.P.T.: Aptos 2019 blindness detection dataset. Kaggle Competition: https://www.kaggle.com/c/aptos2019-blindness-detection (2019), 3662 fundus images with DR severity labels (0–4), Accessed: 2026-02-16
35. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Neural Information Processing Systems (2017), https://api.semanticscholar.org/CorpusID:13756489
36. Walker, C., Jha, S., Ewetz, R.: Metric-driven attributions for vision transformers. In: International Conference on Learning Representations. vol. 2025, pp. 31446–31482 (2025)
37. Wang, Y., Rudner, T.G.J., Wilson, A.G.: Visual explanations of image-text representations via multi-modal information bottleneck attribution. ArXiv **abs/2312.17174** (2023), https://api.semanticscholar.org/CorpusID:266573768
38. Welinder, P., Branson, S., Mita, T., Wah, C., Schroff, F., Belongie, S., Perona, P.: Caltech-ucsd birds 200 (2010)
39. Wu, R., Zhang, C., Zhang, J., Zhou, Y., Zhou, T., Fu, H.: Mm-retinal: Knowledge-enhanced foundational pretraining with fundus image-text expertise. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 722–732. Springer (2024)
40. Ye, X., He, S., Zhong, X., Yu, J., Yang, S., Shen, Y., Chen, Y., Wang, Y., Huang, X., Shen, L.: Oimhs: An optical coherence tomography image dataset based on macular hole manual segmentation. Scientific Data **10**(1), 769 (2023)
41. Yu, Q., Xie, J., Nguyen, A., Zhao, H., Zhang, J., Fu, H., Zhao, Y., Zheng, Y., Meng, Y.: Clip-dr: Textual knowledge-guided diabetic retinopathy grading with ranking-aware prompting. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 667–677. Springer (2024)
42. Zhao, C., Wang, K., Zeng, X., Zhao, R., Chan, A.B.: Gradient-based visual explanation for transformer-based clip. In: International Conference on Machine Learning (2024), https://api.semanticscholar.org/CorpusID:272330536
43. Zhao, R., Chen, X., Liu, X., Chen, Z., Guo, F., Li, S.: Direct cup-to-disc ratio estimation for glaucoma screening via semi-supervised learning. IEEE Journal of Biomedical and Health Informatics **24**(4), 1104–1113 (2020). https://doi.org/10.1109/JBHI.2019.2934477