Explaining BiomedCLIP with Weighted Banzhaf Interactions Supported by Tree-Gram Parsing

Jakub Rymarski[1,2], Adam Rempała[1], Bartłomiej Sobieski[1,2] and Przemysław Biecek[1,2]

[1]University of Warsaw, Poland
[2]Centre for Credible AI, Warsaw University of Technology, Poland

Abstract

Vision-Language Models (VLMs) are demonstrating significant capabilities in medical tasks like radiology analysis, yet providing faithful and interpretable explanations remains a key consideration for their responsible deployment in clinical settings. However, existing explanation methods, such as the widely used FIxLIP framework, often struggle with the fine-grained nature of modern tokenizers. The tokenization problem fragments clinical concepts—splitting terms like "saddle embolus" into scattered, meaningless subwords—which leads to noisy, semantically incoherent cross-modal attributions. Such fragmentation also results in a combinatorial explosion of interaction possibilities, obscuring the model's true reasoning. To address this, we introduce ParseFIxLIP, an extension that incorporates the Tree-Gram Parsing into the Banzhaf interaction game used by FIxLIP. This semantically informed strategy utilizes dependency parsing trees to define explanation players by grouping related text tokens into semantically coherent units. Our `smart_depth` grouping strategy, merging tokens according to spaCy token dependency tree, successfully mitigates concept fragmentation, yielding substantially more interpretable cross-modal interactions by unifying complex medical concepts. Quantitatively, while baselines struggled with the high dimensionality of long captions, our parsing approach maintained statistical robustness and semantic parsimony. Qualitative analysis on BiomedCLIP, validated on medical imagery (ROCOv2) and general examples, confirms that the approach accurately captures the synergistic influence of grouped words on model predictions. In conclusion, our work offers intuitive and clinically relevant insights into VLM decision-making, fulfilling the critical need for coherent explanations in the medical domain.

Keywords

Vision-Language Models, Explainable AI, Weighted Banzhaf Interactions, Tree-Gram Parsing, BiomedCLIP, Medical Imaging

CEUR-WS.org/Vol-4339/paper_7.pdf

1. Introduction

Contrastive language–image pre-training has revolutionized the development of powerful Vision–Language Models (VLMs), enabling them to perform exceptionally well on a wide range of downstream tasks [1]. Examples include setting new benchmarks in zero-shot classification [2] and complex visual question answering [3]. In the medical domain, models such as BiomedCLIP [4] demonstrate strong zero-shot capabilities in processing complex radiological image–text pairs. However, these encoders remain highly opaque. Deploying them in high-stakes clinical decision-making requires not only high accuracy but also trustworthy and interpretable explanations of how predictions arise from specific visual and textual features [5, 6].

One way to validate these similarity predictions is through game-theoretic explainability frameworks that build upon established research on interaction indices and generalized additive models [7, 8, 9], such as the recently proposed FIxLIP [10]. By modeling input tokens as players in a cooperative game, FIxLIP faithfully decomposes the model's output into main effects and pairwise cross-modal interactions using Weighted Banzhaf values [11]. Yet, a fundamental limitation of applying such methods directly to modern VLMs lies in their reliance on subword tokenizers. In a clinical context, processing raw tokens leads to severe concept fragmentation—splitting precise medical terminology like "saddle embolus" into scattered, arbitrary fragments (e.g., "sad", "dle", "embol", "us"). This granularity creates two critical

EXPLIMED 2026 - Third Workshop on Explainable Artificial Intelligence for the medical domain - 15-17 August 2026, Bremen, Germany

jr472844@students.mimuw.edu.pl (J. Rymarski); ar406309@students.mimuw.edu.pl (A. Rempała); sobieski.bartlomiej.jan@gmail.com (B. Sobieski); przemyslaw.biecek@gmail.com (P. Biecek)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-11-26

problems: it generates semantically incoherent explanations that are unreadable for clinicians, and it triggers a combinatorial explosion of meaningless cross-modal interactions. Consequently, the explanation process suffers from multicollinearity and the curse of dimensionality, heavily penalizing goodness-of-fit metrics [12, 13], especially when analyzing long, complex radiology reports.

To overcome these challenges, we introduce **ParseFIxLIP**, a model-agnostic approach that incorporates linguistic structure into interaction-based explanations. Building upon the concept of Tree-Gram Parsing [14], we utilize Natural Language Processing (NLP) dependency parsing (e.g., via spaCy [15]) to formally project the original Banzhaf game into a lower-dimensional quotient game. By linking syntactically dependent modifiers to their roots, we redefine the players of the game as semantically coherent textual units. This enforces semantic parsimony, structurally eliminating the noise caused by subword fragmentation before the game-theoretic interactions are even computed.

Contributions. Our work advances the state of VLM interpretability in the following ways: **(1) A linguistically-informed game-theoretic explanation.** We introduce ParseFIxLIP, adapting the Banzhaf interaction framework by using dependency parsing trees. To our knowledge, this is the first work to address the token fragmentation problem in game-theoretic XAI by grouping tokens into coherent clinical concepts. **(2) Theoretical mitigation of the dimensionality curse.** We demonstrate that our quotient game formulation inherently solves the multicollinearity and high-dimensionality issues present in standard token-level interaction methods. By reducing the number of players, we stabilize the Weighted Banzhaf estimation under a fixed sampling budget. **(3) Empirical validation on medical data.** We evaluate multiple semantic grouping strategies on the ROCOv2 dataset [16]. Quantitative metrics show that while standard baselines fail (yielding negative Adjusted R^2) on long radiology captions, our semantic grouping maintains statistical robustness.

2. Related Work

Vision-Language Models in the Medical Domain. The success of contrastive language-image pre-training [1] has led to its rapid adoption in specialized domains, notably medicine [5, 17]. Models such as BiomedCLIP [4] leverage massive datasets of scientific image-text pairs to achieve strong zero-shot performance in tasks like radiology report analysis. However, the black-box nature of these multimodal architectures poses a significant barrier to clinical adoption. While traditional saliency maps have been applied to medical imaging to validate model predictions, they often fail to capture the nuanced, cross-modal dependencies critical to clinical reasoning, driving the need for more sophisticated, multimodal explainability frameworks.

Game-Theoretic Explainability for VLMs. To provide faithful feature attribution, game-theoretic methods rooted in Shapley and Banzhaf values have become increasingly prominent in machine learning [18]. In the multimodal space, existing gradient and attention based methods [19, 20] generally approximate only first-order attributions, acting as simple unimodal saliency maps. To capture synergistic cross-modal effects, the FIxLIP framework [10] introduced a computationally efficient approximation of Weighted Banzhaf Interactions for vision–language encoders. While FIxLIP successfully uncovers pairwise interactions between image patches and text, its direct application to medical models is severely limited by its reliance on raw subword tokens, which struggle to represent complex clinical terminology.

Linguistic Structure and Concept-Based Interpretability. The reliance of modern VLMs on subword tokenizers (e.g., BPE, WordPiece) creates a well-documented disconnect between mathematical model inputs and human-understandable concepts [21, 22]. In NLP, techniques like Tree-Gram Parsing [14] and dependency parsing have long been utilized to group words into modular semantic units. However, their integration into multimodal explainability remains largely unexplored. Previous attempts at concept-level explanations typically require auxiliary models, human-annotated concepts, or extensive linear probing [23, 24]. Our approach, ParseFIxLIP, bridges this gap by directly embedding NLP dependency trees into the game-theoretic formulation.

3. Methodology

3.1. Background: Vision-Language Similarity and the FIxLIP Game

A vision–language encoder f consists of a vision encoder $f_I : \mathbb{R}^{n_I} \to \mathbb{R}^d$ and a language encoder $f_T : \mathbb{R}^{n_T} \to \mathbb{R}^d$. For a given input pair of n_I image patches $\mathbf{x}_I \in \mathbb{R}^{n_I}$ and n_T text tokens $\mathbf{x}_T \in \mathbb{R}^{n_T}$, the model computes the cosine similarity $f(\mathbf{x}_I, \mathbf{x}_T)$ of their d-dimensional embeddings. Let $N = N_I \cup N_T$ denote the combined index set of raw input features, where $N_I = \{1, \ldots, n_I\}$ represents the image patches and $N_T = \{n_I + 1, \ldots, n_I + n_T\}$ represents the text subword tokens produced by a standard tokenizer.

The FIxLIP framework explains this similarity prediction by formulating it as a cooperative game. For a given input pair $(\mathbf{x}_I, \mathbf{x}_T)$ and a baseline $(\mathbf{b}_I, \mathbf{b}_T)$, the masking operator $\oplus_S$ retains the features in subset S and replaces the hidden rest with their baseline values. For any subset of visible tokens (a mask) $M \subseteq N$, the model yields a similarity score:

$$\nu(M) = f(\mathbf{x}_I \oplus_{M \cap N_I} \mathbf{b}_I, \mathbf{x}_T \oplus_{M \cap N_T} \mathbf{b}_T) \tag{1}$$

FIxLIP decomposes this score into main effects and pairwise interactions by fitting a 2-additive game approximation $\hat{\nu}_{\mathbf{e}}(M)$:

$$\hat{\nu}_{\mathbf{e}}(M) = e_0 + \sum_{i \in M} e_i + \sum_{\{i,j\} \subseteq M : i \neq j} e_{\{i,j\}} \tag{2}$$

where e_0 is a constant baseline, e_i is the individual attribution of token i, and $e_{\{i,j\}}$ represents the pairwise interaction between tokens i and j. The explanation parameters are found by minimizing a p-weighted least squares error (approximating Weighted Banzhaf Interactions) over uniformly sampled masks. Explicitly, given a computational budget of K masks sampled uniformly at random, $M_1, \ldots, M_K \sim \mathcal{U}(2^N)$, we minimize the empirical loss:

$$\hat{\mathcal{L}}(\mathbf{e}) = \frac{1}{K} \sum_{k=1}^{K} p(M_k) \left(\nu(M_k) - \hat{\nu}_{\mathbf{e}}(M_k)\right)^2 \tag{3}$$

where the weighting term $p(M_k) = p^{|M_k|}(1-p)^{|N|-|M_k|}$ reflects the chosen probability $p \in (0, 1)$ of an individual token remaining visible.

3.2. The Fragmentation Problem and Multicollinearity

While FIxLIP provides a rigorous explanation basis, applying it directly to VLMs at the subword level introduces critical statistical and semantic flaws. Modern tokenizers split complex clinical terms into arbitrary fragments (e.g., “saddle” into “sad” and “dle”).

In the random masking process, evaluating a state where “sad” is visible but “dle” is hidden creates an out-of-distribution and meaningless text input. Furthermore, from a theoretical perspective, artificially splitting a single underlying concept into multiple sub-tokens creates highly correlated variables in the interaction space. This multicollinearity destabilizes the Banzhaf regression estimation. As the length of the medical report increases, the dimensionality of the raw token set $|N|$ grows. Driven by the curse of dimensionality, the combinatorial space of possible masks explodes. Under a fixed computational budget of sampled masks, the regression fails to isolate meaningful signals from the noise, resulting in poor goodness-of-fit.

3.3. ParseFIxLIP: Tree-Gram Parsing and Semantic Grouping

To solve both the semantic fragmentation and the mathematical instability, we introduce ParseFIxLIP. Instead of computing interactions on the raw token set N, we project the explanation task into a reduced dimensional space. We define a surjective mapping function $\pi : N \to N'$ that groups raw text

tokens into a smaller set of semantically coherent macro-players N', based on the Tree-Gram Parsing paradigm [14].

This transformation defines a *reduced explanation game* $\mathcal{G}' = (N', \nu')$. For any subset of macro-players $M' \subseteq N'$, the corresponding similarity score is evaluated by mapping the macro-mask back to the raw tokens:

$$\nu'(M') = \nu(\{i \in N \mid \pi(i) \in M'\}) \tag{4}$$

When a macro-player in N' is masked, all its constituent sub-tokens from N are masked simultaneously. To construct the mapping π, we utilize the spaCy dependency parser to implement four grouping strategies: **(1)** `raw`**:** The baseline tokenizer splitting (identity mapping). **(2)** `words`**:** Subwords are strictly merged into full dictionary words. **(3)** `noun_chunks`**:** Multi-word noun phrases are grouped into single entities, leaving remaining tokens as standalone players. **(4)** `smart_depth`**:** Our novel syntactic strategy, where heads of dependency trees are recursively grouped with their modifiers.

Theoretical justification. Formulating the explanation on a reduced game space provides a strong theoretical guarantee for long reports. By enforcing semantic grouping, we ensure that $|N'| \ll |N|$. Drastically reducing the number of players under a fixed sampling budget (e.g., $K = 50,000$ perturbations) directly reduces the variance of weighted least squares estimator. This effectively neutralizes the multicollinearity problem and maintains semantic parsimony, structurally guaranteeing a higher and more stable Adjusted R^2 fit compared to raw token baselines.

4. Quantitative Experiments

We evaluate the proposed parsing strategies on a subset of the ROCOv2 dataset following the evaluation protocol of [10]. Since explanation quality strongly depends on report length, we stratify the analysis by caption length.

The faithfulness is measured by Spearman's Rank Correlation and Adjusted R^2 to check if Tree-Gram Parsing increases it. Spearman's correlation assesses the general capability of the explanation method to recover the ranking of similarity scores for uniformly masked inputs compared to the original model predictions. Adjusted R^2 is used to evaluate the precise goodness-of-fit. It accounts for the varying number of players resulting from different parsing strategies.

In terms of Spearman correlation on short captions, `noun_chunks` and `smart_depth` demonstrate the highest stability with a median score close to 0.9. This disparity is more visible in the Adjusted R^2 metric, where both `smart_depth` and `noun_chunks` achieve median values above 0.7 (Figure 1).

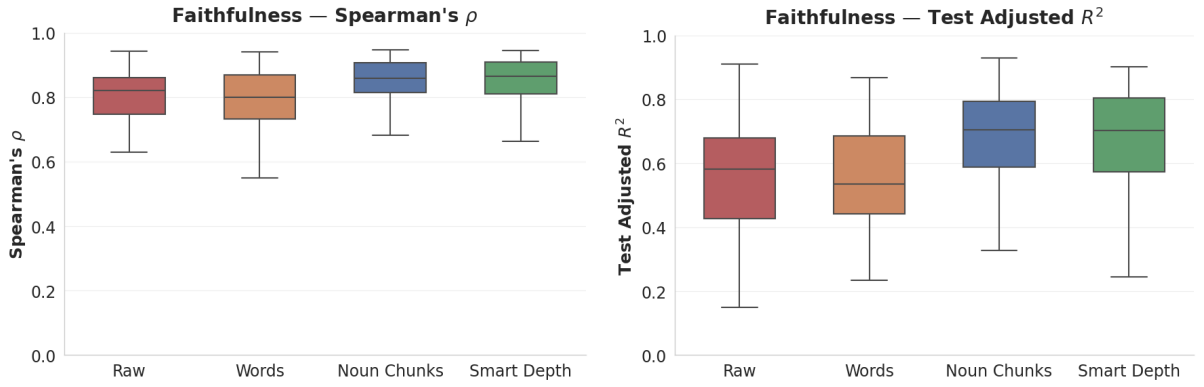


Figure 1: Faithfulness evaluation on 128 short captions samples. **Left:** Spearman Rank Correlation. **Right:** Adjusted R^2. `noun_chunks` and `smart_depth` **consistently outperform the token-level baselines**.

We extend our quantitative evaluation to the challenging subset of long captions (> 30 words). This setting results in a high-dimensional feature space populated largely by functional words and background noise. The interaction space also expands drastically. We employed a sampling budget of

$N = 50,000$ perturbations. The following results suggest that this budget might be insufficient for the token-level baseline to stabilize the signal-to-noise ratio in such a sparse feature space.

First, we analyze the model fit in Figure 2. Evaluated using the Friedman test and subsequent pairwise Wilcoxon signed-rank tests at a significance level of $\alpha = 0.05$, the Standard R^2 (Left) shows statistically significant differences across strategies ($p < 0.001$). However, a critical divergence appears in the Adjusted R^2 (Right), where semantic grouping drastically outperforms the raw token baseline ($p < 0.0001$). The `raw` strategy drops to negative values for two samples. This phenomenon likely reflects the curse of dimensionality, as the `raw` method attempts to fit a linear explanation using over 70 variables (tokens + patches). The penalty for this high complexity, combined with the multicollinearity of noisy linguistic tokens, outweighs the marginal gains in fit. In contrast, `smart_depth` effectively performs dimensionality reduction, condensing the input into < 15 semantic concepts. This allows the method to recover a statistically robust signal.

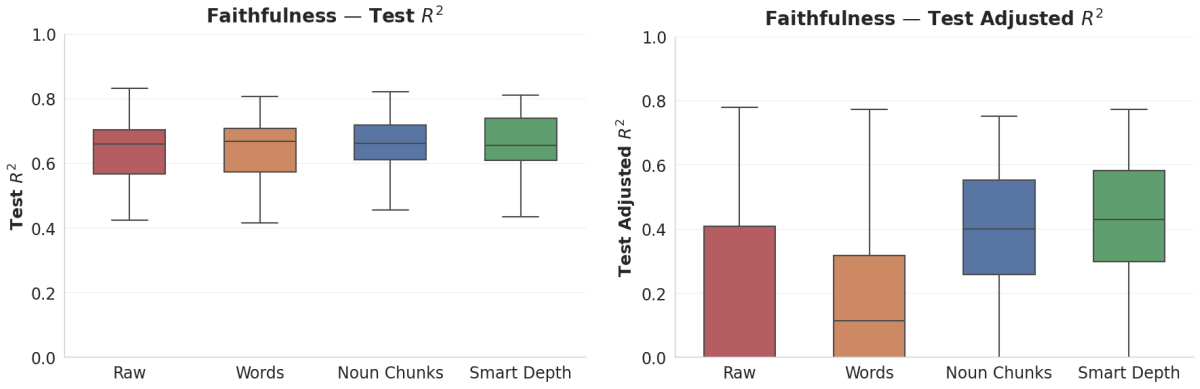


Figure 2: Goodness-of-fit metrics on long captions. **Left:** Standard R^2 shows comparable explanatory potential. **Right:** Adjusted R^2 reveals the inefficiency of the `raw` strategy, which is penalized for its high feature count, whereas `smart_depth` remains robust due to semantic parsimony.

We further assess the quality of the explanations, to check if Tree-Gram Parsing improves them, using Insertion and Deletion curves, quantified by the Area Between Curves (AID). These metrics measure how rapidly the model's confidence changes as important features are added to or removed from the input.

In the Figure 3 (Left), semantic grouping strategies achieve consistently higher AID scores compared to token-level baselines on short captions. `smart_depth` yields the highest AID (0.79 ± 0.41), outperforming the `raw` strategy (0.73 ± 0.34). Interestingly, on long captions (Figure 3, Right), the `raw` strategy achieves slightly higher AID scores, as it may track the model's scattered attention to background noise and adverbs, which semantic grouping intentionally abstracts away.

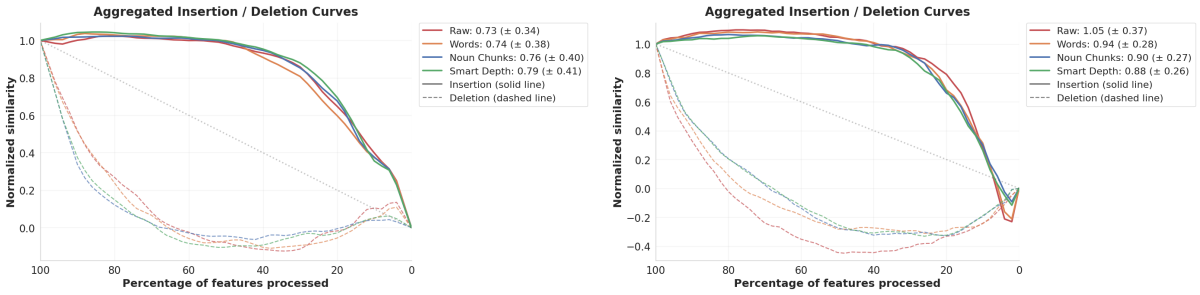


Figure 3: Insertion and Deletion curves. **Left:** Short Captions (< 10 words). Semantic grouping outperforms raw tokens on short texts. **Right:** Long Captions (> 30 words). On long texts, the raw baseline tracks scattered attention to background noise, slightly inflating its AID score compared to semantic grouping.

5. Qualitative Experiments

5.1. Parsing Enables the Understanding of Complex Medical Concepts

In this section, we investigate whether Tree-Gram Parsing mitigates the key limitation of standard VLM explanations, namely the fragmentation of medical concepts caused by tokenization. Figure 4 illustrates the progression of interpretability across different parsing strategies for the caption "CT angiogram with saddle embolus". While our primary qualitative analysis focuses on cross-modal interactions, ParseFIxLIP naturally yields clearer visualizations of main effects (standalone feature attribution). The visualizations show both cross-modal interaction strengths and the main effects of single-modality players.

In the `raw` strategy, the medical term "saddle embolus" is split into arbitrary tokens (e.g., "sad", "dle", "embol", "us"). The explanation is noisy, with importance distributed across meaningless fragments and significant weight given to the general modality "angiogram" rather than the specific pathology "saddle embolus". The `words` strategy merges subwords into recognizable terms. However, it splits similarity score treating "saddle" and "embolus" as separate players, still highlighting general term "angiogram". In contrast, our proposed semantic parsing strategy, `smart_depth` successfully groups these tokens into a coherent clinical concept. The parser identifies "saddle embolus" as a single entity. Consequently, the explanation concentrates mostly on this unified concept and results in the best quantitative metrics.

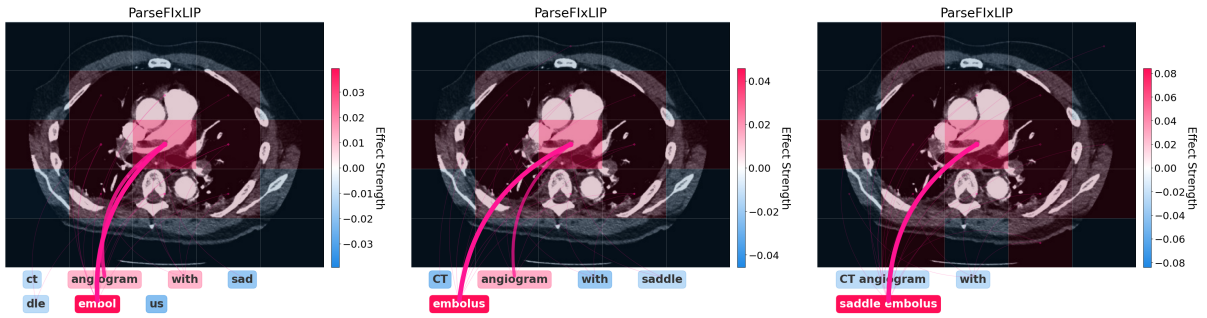


Figure 4: Qualitative comparison of explanations. **Left:** `raw` – Concept fragmented, noisy explanation. **Middle:** `words` – Subwords merged into words. **Right:** `smart_depth` – "saddle embolus" as one concept, a clean explanation.

This capability extends effectively to longer and more complex phrases. Figure 5 showcases the interaction analysis for the long caption containing the phrase "the right kidney". The `words` strategy fragments the concept into single words, leading to the inability to distinguish between the left and right kidney. Conversely, `noun_chunks` processes the entire phrase as a single unit, assigning interaction importance directly to the observed kidney and its anatomical view.

We further illustrate the capability of our `smart_depth` strategy to capture complex, multi-word medical concepts compared to the fragmented baseline (Figures 6, 7 and 8).

Radiological images frequently contain explicit annotations (arrows, rulers, text, like in the Figure 9) which pose a unique challenge. Our framework connects these straightforward, not interesting from a medical perspective, interactions. For instance, the word "arrow" is heavily connected to the red graphical marker pointing towards the spinal pathology. Some other seemingly important relations can be omitted.

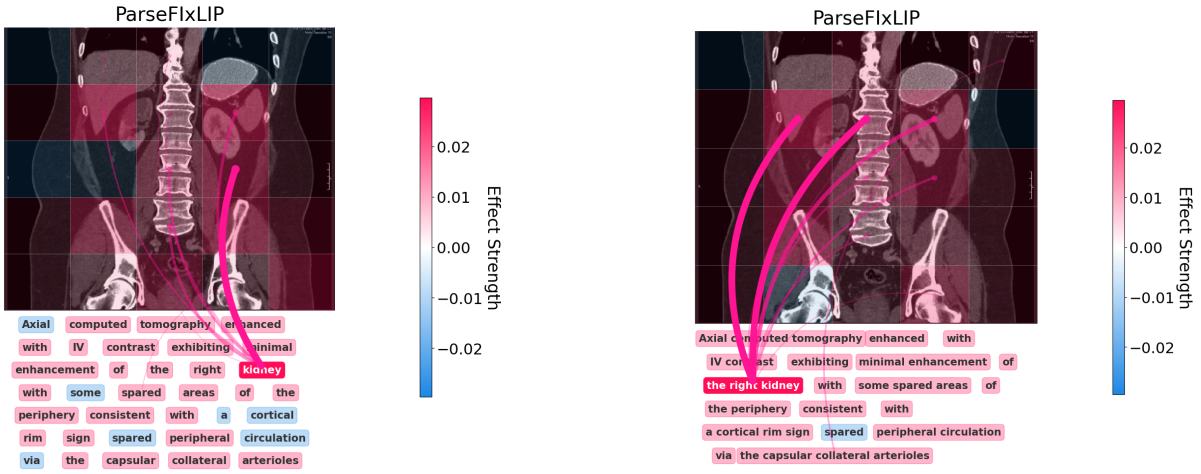


Figure 5: Visual explanation comparison for the long caption containing "the right kidney". **Left:** `words` strategy – Fragmentation leads to the inability to distinguish between the kidneys. **Right:** `noun_chunks` – Semantic grouping unifies these tokens, assigning importance directly to the right kidney.

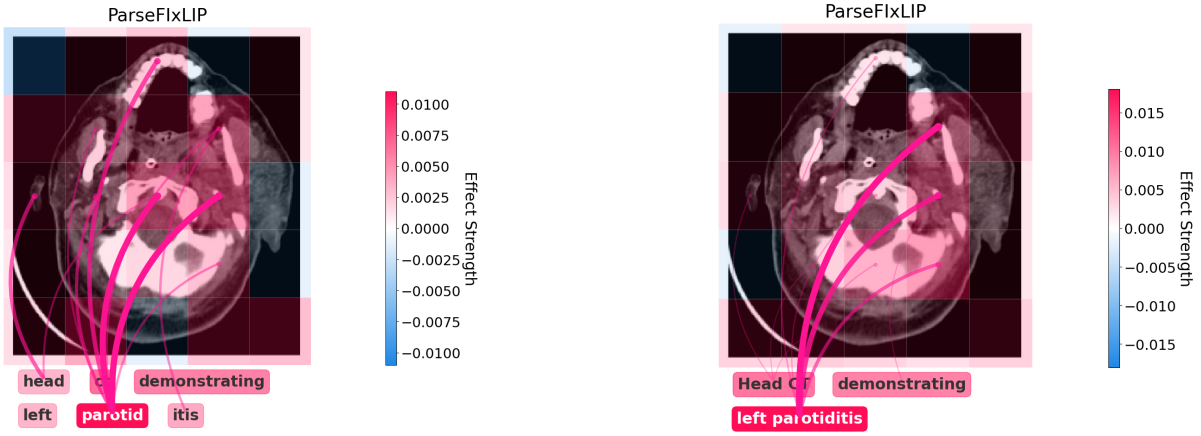


Figure 6: Visual explanation comparison for the caption "Head CT demonstrating left parotiditis". **Left:** `raw` – Fragmenting the concept into "left", "parotid", and "itis" results in scattered attention. **Right:** `smart_depth` – Processing "left parotiditis" as a single semantic unit yields a precise explanation. Semantic grouping eliminates the noise seen with fragmented tokens.

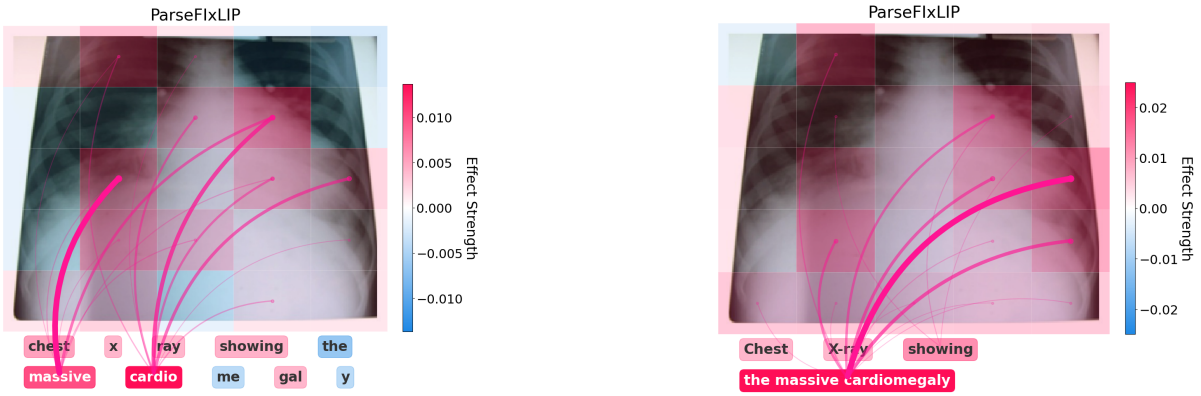


Figure 7: Visual explanation comparison for "massive cardiomegaly". **Left:** `raw` – Fragmentation of the medical term generates significant visual noise outside the heart region. **Right:** `smart_depth` — Processing the entire phrase as a single unit eliminates noise.

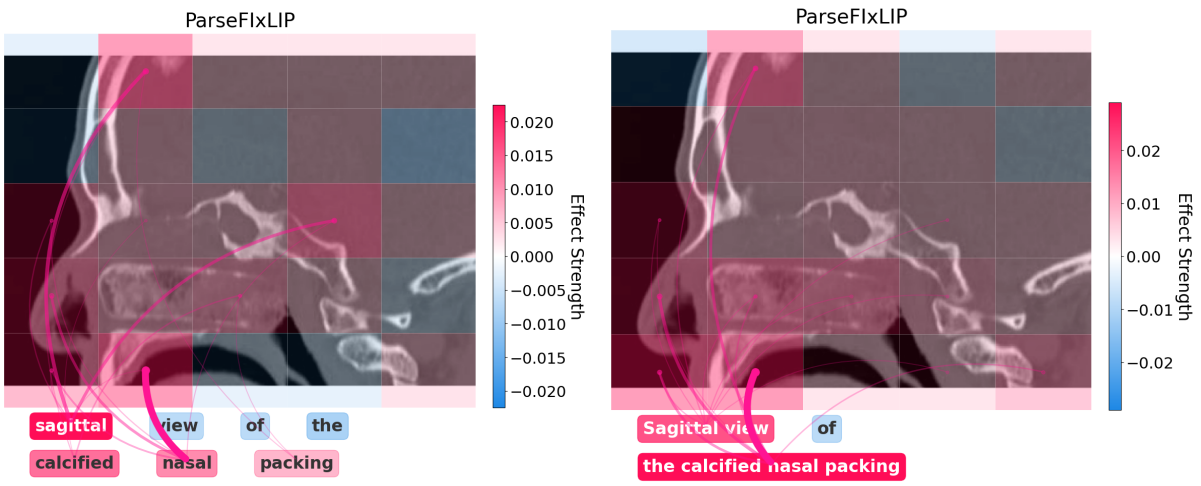


Figure 8: Visual explanation comparison for "Sagittal view of the calcified nasal packing". **Left:** `raw` – Noisy distribution of importance. **Right:** `smart_depth` – Clear semantic grounding where relevant image parts receive positive attribution.

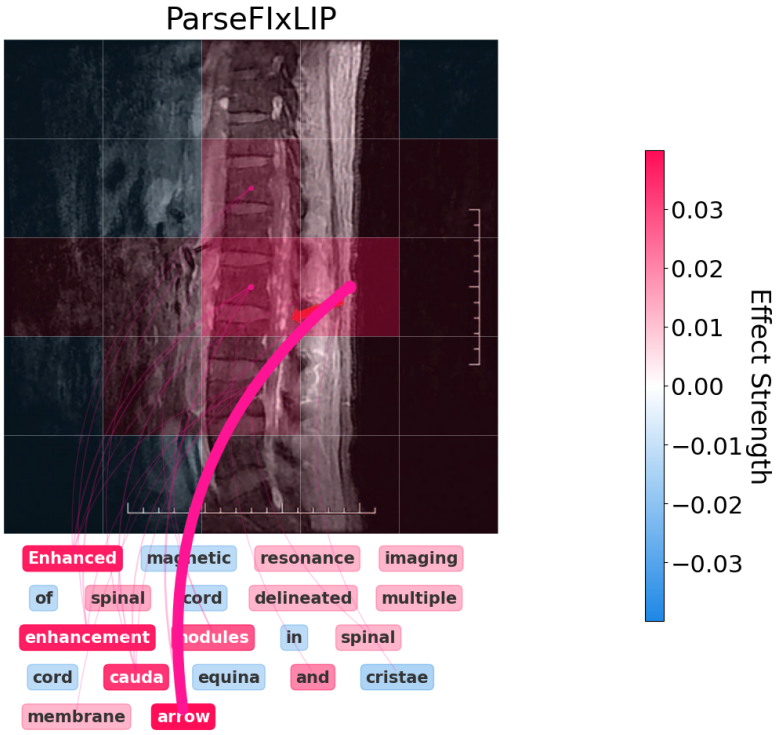


Figure 9: Visual grounding of explicit annotations. The word "arrow" is connected to the red graphical marker pointing towards the spinal pathology.

5.2. Image transformation test

We investigate the robustness of BiomedCLIP and our parsing strategy to spatial transformations such as upside down rotation. The baseline normal image yields a model output similarity of 0.4587, whereas the transformed version drops to 0.3061. This performance degradation is likely because the model was trained on medical images always taken from a standardized position, thus the model does not recognize the concept of the right lobe in the rotated image and the main contributor are the numbers.

Grouping tokens strengthens "the right lobe" score for normal view. The model demonstrates robustness to transformations regarding printed numbers and symbols, but fails to maintain orientation-awareness for anatomical structures.

The model exhibits unexpected behavior when processing inverted images. Deducing from the Figures 10 and 11, it seems that the model is learning "multiloculated hepatic abscess" from areas other than the liver. This may be related to "hepatic" domination seen in the `raw`, which word regularly appears in the vision context (and can be linked with liver-neighboring organs).

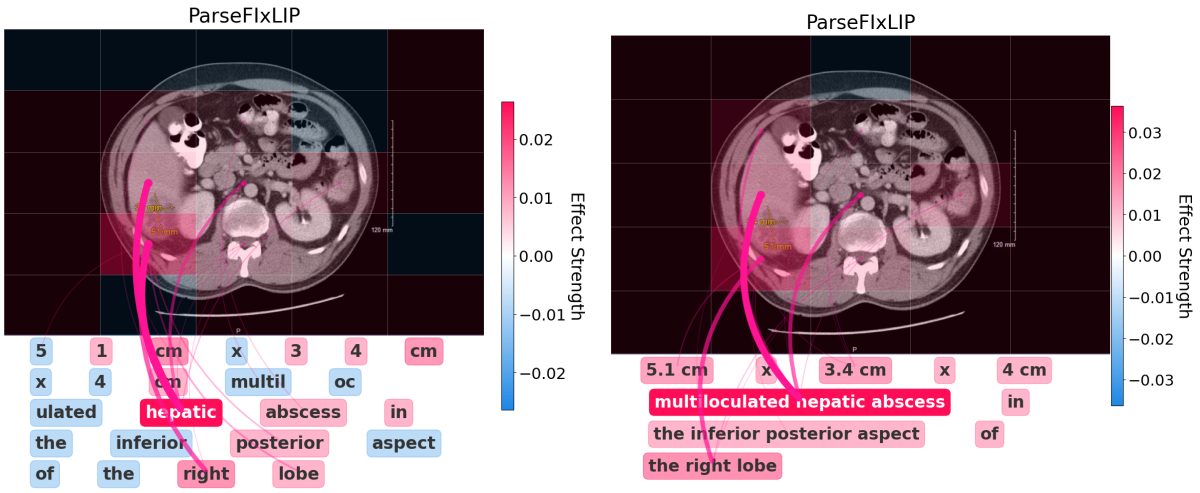


Figure 10: Image transformation comparison. **Left:** Normal `raw`. **Right:** Normal `smart_depth`.

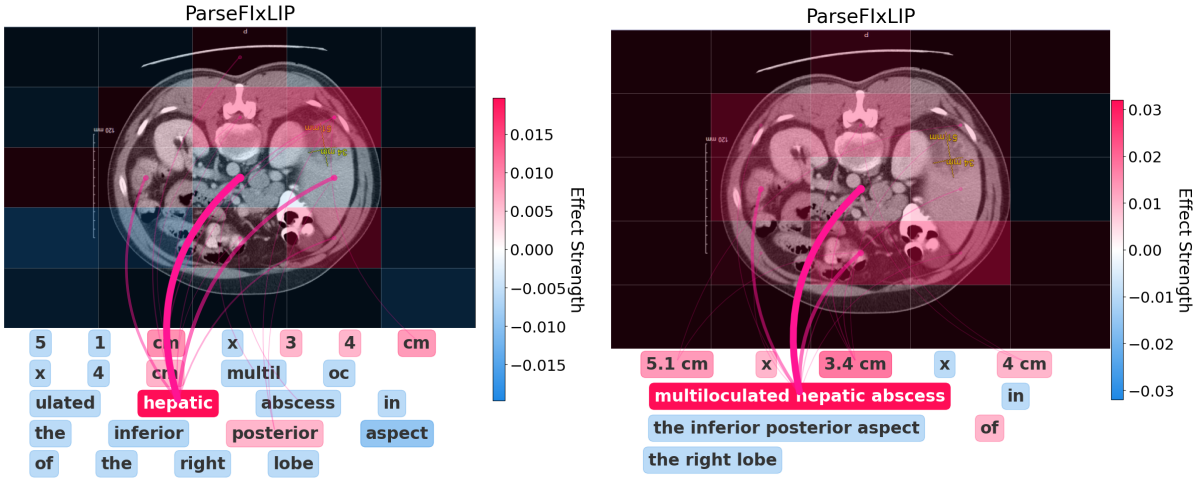


Figure 11: Image transformation comparison. **Left:** Upside down `raw`. **Right:** Upside down `smart_depth`.

5.3. False label test

Can the model detect left/right change in the text? The model can distinguish between the left and the right breast (Figures 12, 13), probably considering its placement in the image. The breast position on the image is meaningful for selecting the correct breast, not the arrow. Connecting the tokens together changes the part of the image indicated by breast — breast appearing always with the correct left/right indicator is linked with the image part correctly, not as the right.

The model fails in recognizing the correct breast in the transformed image (not surprising, as breasts are quite similar and model is not transformation-robust). A notable limitation observed in this experiment is that the Tree-Gram Parsing algorithm separated the "breast" token from the "nodule"

token (probably due to being separate nouns). It would be interesting to see if the model is more likely to classify such grouped chunk to the left breast (false label) or the nodule (detected correctly on the right breast). Interestingly, both the nodule and the arrow pointing to it are visible in the image. They are on separate grids, but all the experiments show that the arrow is more meaningful for nodule detection in the image than the nodule itself.

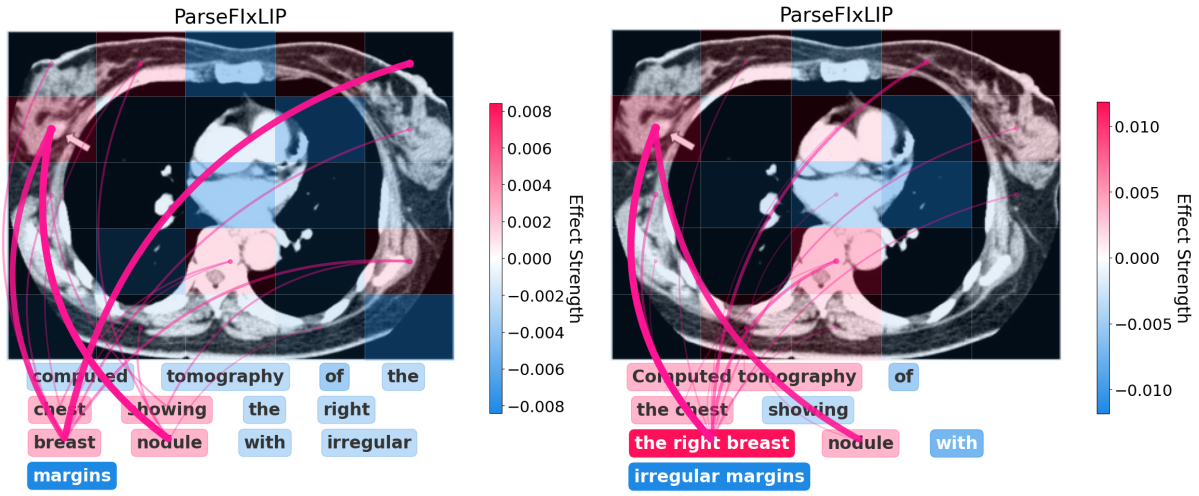


Figure 12: True label. **Left:** `raw`. **Right:** `smart_depth`.

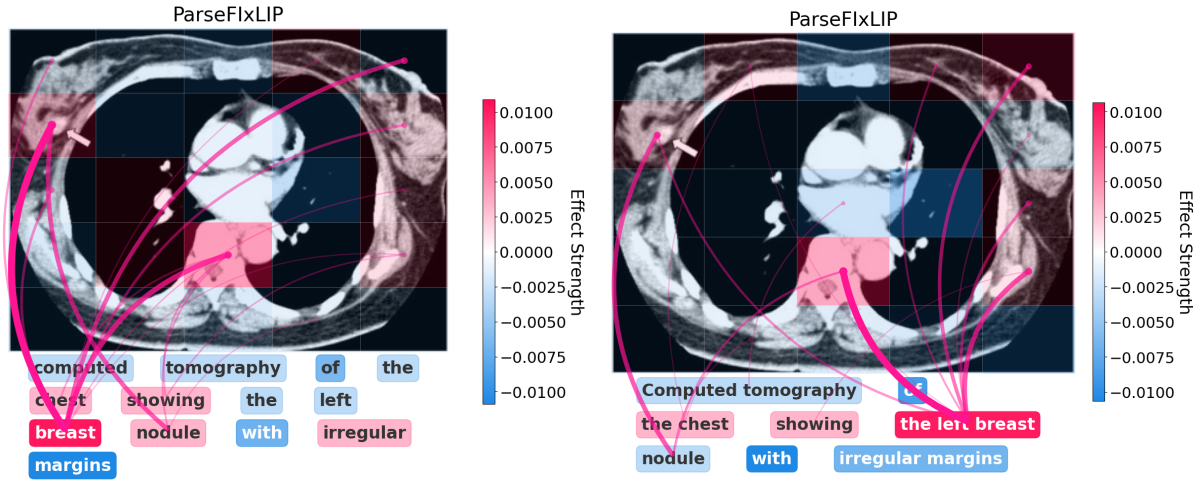


Figure 13: False label. **Left:** `raw`. **Right:** `smart_depth`.

6. Conclusions

One of the goals of our work is to faithfully explain the inner workings of BiomedCLIP. To demonstrate this capability and evaluate our method, we conducted several qualitative stress tests. Through these tests, we observe that BiomedCLIP itself is not robust to certain conditions, which our explanations successfully capture. **Transformation Fragility.** Because medical datasets contain standardized images, BiomedCLIP severely lacks robustness to spatial transformations. When images are rotated (Figure 11), the model's similarity score drops drastically. ParseFIxLIP faithfully reflects this confusion: the interaction weight shifts away from the misplaced anatomy and heavily focuses on transformation-robust artifacts, such as printed numbers or text symbols. **Shortcut Learning via Annotations.** Radiological images frequently contain explicit markers like arrows or text. Our explanations reveal that

BiomedCLIP frequently assigns disproportionately high interaction weights to these explicit pointers rather than the actual pathology. This highlights a critical warning: the model often learns to "cheat" by looking at human annotations, meaning interpretations of BiomedCLIP outputs require extreme caution.
Parser Limitations. While Tree-Gram Parsing effectively unifies clinical concepts, it is fundamentally constrained by the underlying NLP parser (spaCy). As observed in our false label tests (Section 5.3), the parser occasionally separates related concepts (e.g., decoupling "breast" from "nodule").

To conclude, adding Tree-Gram Parsing to the FIxLIP algorithm allows us to understand more about how tokens influence the model results and find out the synergy of the grouped tokens. Quantitative evaluation shows that our parsing-based approach achieves performance higher and more stable than the raw token-based FIxLIP, efficiently neutralizing the curse of dimensionality on complex reports. Qualitative analysis demonstrates that our dependency-based grouping significantly mitigates the fragmentation issues inherent in the baseline method, offering more intuitive and clinically relevant insights into the model's decision-making process. Crucially, ParseFIxLIP exposes BiomedCLIP's severe limitations, such as spatial transformation fragility and shortcut learning via visual annotations (arrows, numbers).

Declaration on Generative AI

During the preparation of this work, the authors used *Gemini* in order to: Grammar and spelling check. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, PMLR, 2021, pp. 8748–8763.

[2] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, T. Duerig, Scaling up visual and vision-language representation learning with noisy text supervision, in: International Conference on Machine Learning, PMLR, 2021, pp. 4904–4916.

[3] J. Li, D. Li, C. Xiong, S. Hoi, Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, in: International Conference on Machine Learning, PMLR, 2022, pp. 12888–12900.

[4] S. Zhang, Y. Xu, N. Usuyama, H. Xu, J. Bagga, R. Tinn, S. Preston, R. Rao, M. Wei, N. Valluri, et al., A multimodal biomedical foundation model trained from fifteen million image–text pairs, NEJM AI 2 (2024). doi:10.1056/AIoa2400640.

[5] J. N. Acosta, G. J. Falcone, P. Rajpurkar, E. J. Topol, Multimodal biomedical ai, Nature Medicine 28 (2022) 1773–1784.

[6] J. Amann, A. Blasimme, E. Vayena, D. Frey, V. I. Madai, Precise4Q Consortium, Explainability for artificial intelligence in healthcare: a multidisciplinary perspective, BMC medical informatics and decision making 20 (2020) 310.

[7] S. Bordt, U. von Luxburg, From shapley values to generalized additive models and back, in: International Conference on Artificial Intelligence and Statistics, PMLR, 2023, pp. 709–745.

[8] F. Fumagalli, M. Muschalik, P. Kolpaczki, E. Hüllermeier, B. Hammer, SHAP-IQ: Unified approximation of any-order shapley interactions, in: Advances in Neural Information Processing Systems, 2023.

[9] G. D. Pelegrina, L. T. Duarte, M. Grabisch, A k-additive choquet integral-based approach to approximate the shap values for local interpretability in machine learning, Artificial Intelligence 325 (2023) 104014.

[10] H. Baniecki, M. Muschalik, F. Fumagalli, B. Hammer, E. Hüllermeier, P. Biecek, Explaining

similarity in vision-language encoders with weighted banzhaf interactions, in: Advances in Neural Information Processing Systems, 2025.
[11] J.-L. Marichal, P. Mathonet, Weighted banzhaf power and interaction indexes through weighted approximations of games, European Journal of Operational Research 211 (2011) 352–358.
[12] I. Covert, S.-I. Lee, Improving kernelshap: Practical shapley value estimation using linear regression, in: International Conference on Artificial Intelligence and Statistics, PMLR, 2021, pp. 3457–3465.
[13] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer Science & Business Media, 2009.
[14] K. Sima'an, Tree-gram parsing: Lexical dependencies and structural relations, in: Proceedings of the 38th annual meeting of the association for computational linguistics, 2000, pp. 37–44.
[15] M. Honnibal, I. Montani, S. Van Landeghem, A. Boyd, spacy: Industrial-strength natural language processing in python, 2020. URL: https://spacy.io. doi:10.5281/zenodo.1212303.
[16] J. Rückert, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, C. S. Schmidt, S. Koitka, O. Pelka, A. B. Abacha, A. G. Seco de Herrera, et al., Rocov2: Radiology objects in context version 2, an updated multimodal image dataset, Scientific Data 11 (2024) 688.
[17] M. Moor, O. Banerjee, Z. S. H. Abad, H. M. Krumholz, J. Leskovec, E. J. Topol, P. Rajpurkar, Foundation models for generalist medical artificial intelligence, Nature 616 (2023) 259–265.
[18] I. Covert, S. Lundberg, S.-I. Lee, Explaining by removing: A unified framework for model explanation, Journal of Machine Learning Research 22 (2021) 1–90.
[19] H. Chefer, S. Gur, L. Wolf, Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers, in: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 397–406.
[20] C. Zhao, K. Wang, X. Zeng, R. Zhao, A. B. Chan, Gradient-based visual explanation for transformer-based CLIP, in: International Conference on Machine Learning, PMLR, 2024, pp. 61072–61091.
[21] S. J. Mielke, Z. Alyafeai, E. Salesky, C. Raffel, M. Dey, M. Gallé, A. Raja, C. Si, W. Y. Lee, B. Sagot, et al., Between words and characters: A brief history of open-vocabulary modeling and tokenization in nlp, arXiv preprint arXiv:2112.10508 (2021).
[22] K. Bostrom, G. Durrett, Byte pair encoding is suboptimal for language model pretraining, in: Findings of the Association for Computational Linguistics: EMNLP 2020, 2020, pp. 4617–4624.
[23] M. Yuksekgonul, M. Wang, J. Zou, Post-hoc concept bottleneck models, in: International Conference on Learning Representations, 2023.
[24] T. Oikarinen, T.-W. Weng, Clip-dissect: Automatic description of neuron representations in deep vision networks, in: International Conference on Learning Representations, 2023.