Medición del balance multimodal en modelos CLIP médicos usando MM-SHAP

Andrés Alberto Góngora-Ramos, Pablo Pancardo García,
Luis Enrique Ramon-Pedrero

Universidad Juárez Autónoma de Tabasco,
México

alberthg.ramos@gmail.com., pablo.pancardo@ujat.mx,
242H21005@alumno.ujat.mx

Resumen. Los modelos médicos de visión–lenguaje basados en CLIP han mostrado buen desempeño en tareas que combinan imágenes y texto; sin embargo, las métricas globales no permiten determinar si ambas modalidades contribuyen de forma equilibrada. En este trabajo se propone un marco de análisis para medir el balance multimodal mediante la aplicación experimental de MM-SHAP. Cada muestra se representa como tokens textuales y parches visuales, cuyas contribuciones se agregan mediante *TScore* e *IScore*. El análisis se evaluó en *Image–Sentence Alignment* (ISA) sobre ROCO y en *Visual Question Answering* (VQA) sobre VQA-Med 2019, considerando cuatro modelos en ISA y dos en VQA. Los resultados muestran que PubMedCLIP presenta el comportamiento más equilibrado, mientras que BioMedCLIP, RCLIP y WhyXRayCLIP exhiben distintos grados de sesgo hacia la modalidad visual. Estos hallazgos evidencian que métricas como similitud imagen–texto o exactitud no reflejan por sí solas la integración multimodal, y resaltan la utilidad de incorporar métricas de explicabilidad en la evaluación de modelos médicos.

Palabras clave: Explicabilidad multimodal, MM-SHAP, modelos de visión–lenguaje, CLIP médico, balance multimodal.

Multimodal Balance Measurement in Medical CLIP Models Using MM-SHAP

Abstract. CLIP-based medical vision-language models have shown good performance in tasks combining images and text; however, global metrics do not allow us to determine whether both modalities contribute in a balanced way. This paper proposes an analytical framework to measure multimodal balance through the experimental application of MM-SHAP. Each sample is represented as textual tokens and visual patches, whose contributions are aggregated using TScore and IScore. The analysis was evaluated in Image-Sentence Alignment (ISA) on ROCO and in Visual

Question Answering (VQA) on VQA-Med 2019, considering four models in ISA and two in VQA. The results show that PubMedCLIP exhibits the most balanced behavior, while BioMedCLIP, RCLIP, and WhyXRay-CLIP show varying degrees of bias toward the visual modality. These findings show that metrics such as image-text similarity or accuracy do not by themselves reflect multimodal integration, and highlight the usefulness of incorporating explainability metrics in the evaluation of medical models.

Keywords: Multimodal explainability, MM-SHAP, vision-language models, medical CLIP, multimodal balance.

1. Introducción

Una red neuronal multimodal, del inglés *Multimodal Neural Network* (MNN), procesa simultáneamente distintos tipos de datos, como texto e imágenes, integrando información proveniente de diferentes modalidades [15,1]. Dentro de este tipo de modelos, *Contrastive Language–Image Pre-training* (CLIP) [13] ha cobrado relevancia en el análisis de imágenes y reportes médicos, debido a su capacidad para relacionar información visual con descripciones textuales [19]. Sus versiones preentrenadas se han utilizado en aplicaciones clínicas como diagnóstico de enfermedades torácicas y segmentación de órganos [14,12,9].

En medicina, la transparencia de los modelos de inteligencia artificial (IA) es fundamental, ya que la falta de explicaciones puede limitar su adopción en escenarios clínicos [8]. En este contexto, la inteligencia artificial explicable, del inglés *Explainable Artificial Intelligence* (XAI), permite analizar por qué un modelo toma determinadas decisiones. Para sistemas multimodales, MM-SHAP extiende la técnica SHAP (*SHapley Additive exPlanations*) con el fin de estimar la contribución relativa de cada modalidad en una predicción [10].

El balance multimodal se refiere al grado en que distintas modalidades contribuyen de manera equilibrada al resultado final. Cuando un modelo depende excesivamente del texto o de la imagen, puede presentar un sesgo modal que no siempre se refleja en métricas tradicionales de desempeño. Dado lo anterior, en este trabajo se propone un marco de análisis para medir el balance multimodal en modelos médicos basados en CLIP mediante la aplicación experimental de MM-SHAP.

El resto del artículo se organiza de la siguiente manera. La Sección 2 presenta los trabajos relacionados; la Sección 3 describe la metodología; la Sección 4 detalla los datos y el diseño experimental; la Sección 5 presenta los resultados y la discusión; finalmente, la Sección 6 expone las conclusiones.

2. Trabajo relacionado

MM-SHAP extiende SHAP a modelos multimodales al cuantificar la contribución de cada modalidad, como texto e imagen, a nivel de muestra y de conjunto de datos [10]. Su aplicación ha permitido identificar dependencias modales

en modelos de visión–lenguaje, como una mayor dependencia textual en BLIP y un comportamiento más equilibrado en BLIP2 y FLAVA [3].

En medicina, SHAP se ha usado para explicar modelos multimodales, por ejemplo, en la predicción de comorbilidades en epilepsia [7]. Además, SHAP-CAT utiliza valores de Shapley para integrar modalidades histopatológicas y mejorar la clasificación de cáncer [16]. En VQA, se ha mostrado que los modelos pueden depender excesivamente del lenguaje e ignorar la imagen, lo que motiva analizar explícitamente el balance multimodal [5].

3. Metodología

En este trabajo se propone un marco de análisis para evaluar el balance multimodal en modelos médicos de visión–lenguaje mediante la aplicación experimental de MM-SHAP. La formulación matemática original de MM-SHAP no se modifica; en su lugar, se adapta su uso a modelos tipo CLIP entrenados o ajustados al dominio médico. Esta adaptación consiste en: i) representar cada entrada como una combinación de unidades textuales y visuales explicables, ii) definir una salida escalar compatible con SHAP para cada tarea evaluada, iii) implementar un esquema de enmascaramiento para tokens y parches visuales, y iv) agregar las atribuciones por modalidad para calcular métricas de contribución y balance multimodal.

3.1. Descripción general del marco de análisis propuesto

El marco propuesto estima la contribución relativa de texto e imagen en la salida de un modelo multimodal a partir de una entrada (x^{txt}, x^{img}). Para ello, cada muestra se transforma en una representación conjunta compuesta por unidades explicables: tokens en la modalidad textual y parches o regiones en la modalidad visual. Sobre esta representación se aplica MM-SHAP para obtener atribuciones a nivel de característica. Posteriormente, los valores SHAP se agregan por modalidad con el fin de cuantificar la contribución total de texto e imagen en cada predicción.

3.2. Adaptación experimental de MM-SHAP a modelos médicos de visión–lenguaje

MM-SHAP fue propuesto como un marco de explicabilidad para modelos multimodales basado en valores de Shapley. En este trabajo no se modifica su formulación matemática original; la adaptación realizada corresponde a su aplicación experimental en modelos médicos de visión–lenguaje tipo CLIP. Para ello, cada entrada se representa como $\mathbf{z} = [z_1, z_2, \ldots, z_M]$, donde cada z_i es una unidad explicable: un token textual o un parche visual. Las imágenes y los textos se preprocesan con el procesador de cada modelo, respetando su resolución de entrada, normalización visual, tokenización y longitud máxima de secuencia.

La función de predicción $f(\mathbf{z})$ se define como una salida escalar asociada a la decisión analizada. En *Image–Sentence Alignment* (ISA), corresponde a la puntuación de similitud entre la imagen médica y su descripción textual. En *Visual Question Answering* (VQA), corresponde a la puntuación asignada a una respuesta candidata, condicionada por la imagen y la pregunta. En este último caso, la entrada textual combina la pregunta y la respuesta candidata.

Bajo esta formulación, MM-SHAP estima la contribución marginal de cada token textual y parche visual respecto a $f(\mathbf{z})$ mediante coaliciones de características presentes y ausentes. Las atribuciones obtenidas se agregan por modalidad para calcular las métricas de contribución textual y visual. Por tanto, la adaptación propuesta no introduce nuevas ecuaciones, sino que especifica cómo representar las entradas médicas y cómo transformar las salidas de los modelos en una puntuación escalar comparable entre tareas y arquitecturas.

3.3. Enmascarador personalizado y envoltura de predicción

Para aplicar MM-SHAP a los modelos evaluados, se implementaron un enmascarador personalizado y una envoltura de predicción. El enmascarador genera muestras parcialmente observadas a partir de una coalición binaria, preservando la estructura requerida por los codificadores textual y visual. En texto, las características ausentes se sustituyen por un token de relleno compatible con el tokenizador, manteniendo los tokens especiales. En imagen, los parches ausentes se reemplazan por una representación base sobre el tensor de entrada, sin alterar la geometría esperada por el codificador visual.

El enmascaramiento se utiliza solo como una perturbación controlada para estimar contribuciones, no como parte de la predicción final. En imágenes médicas, se aplica a nivel de parches y no de estructuras anatómicas completas; por ello, las explicaciones deben interpretarse como atribuciones visuales aproximadas, no como segmentaciones clínicas ni localizaciones diagnósticas.

La envoltura de predicción transforma la salida del modelo en una función escalar adecuada para SHAP. Para una entrada enmascarada $\tilde{\mathbf{z}}$, el predictor devuelve una puntuación $f(\tilde{\mathbf{z}})$ asociada a la decisión analizada, como similitud imagen–texto en ISA o afinidad con una respuesta candidata en VQA.

3.4. Métricas de contribución y balance multimodal

Las métricas de contribución se calcularon siguiendo la formulación original de MM-SHAP, sin modificar sus ecuaciones. Una vez obtenidos los valores SHAP, las atribuciones individuales se agregan por modalidad:

$$S_{txt} = \sum_{i \in \mathcal{T}} |\phi_i|, \qquad S_{img} = \sum_{i \in \mathcal{I}} |\phi_i|,$$

donde $\mathcal{T}$ y $\mathcal{I}$ representan los conjuntos de características textuales y visuales, respectivamente, y ϕ_i es la atribución SHAP de la característica i.

A partir de estas cantidades se calculan puntajes modales normalizados:

$$TScore = \frac{S_{txt}}{S_{txt} + S_{img}}, \qquad IScore = \frac{S_{img}}{S_{txt} + S_{img}}.$$

Estas métricas permiten medir la contribución relativa de cada modalidad en una muestra. Valores cercanos entre *TScore* e *IScore* indican mayor balance multimodal, mientras que diferencias grandes sugieren sesgo hacia una modalidad dominante. A nivel global, estas puntuaciones se agregan sobre el conjunto de datos para identificar tendencias de dependencia modal y comparar el comportamiento de distintos modelos y tareas.

4. Datos y diseño experimental

4.1. Conjuntos de datos

Se utilizaron dos conjuntos de datos médicos de visión–lenguaje. Para *Image–Sentence Alignment* (ISA) se empleó **ROCO**[1] (*Radiology Objects in Context*) [11], compuesto por pares de imágenes radiológicas y descripciones textuales. Para *Visual Question Answering* (VQA) se utilizó **VQA-Med 2019**[2] [2], compuesto por imágenes médicas, preguntas clínicas y respuestas de referencia. Estos conjuntos permiten evaluar dos escenarios complementarios: alineación imagen–texto y razonamiento visual condicionado por lenguaje.

4.2. Tareas y formulación experimental

Se consideraron dos tareas. La primera fue ISA, formulada como un problema de correspondencia entre una imagen médica y un texto, donde el modelo debe asignar mayor afinidad a pares semánticamente consistentes. La segunda fue VQA, formulada como una tarea de *answer selection*, en la que el modelo puntúa un conjunto de respuestas candidatas condicionado por la imagen y la pregunta, seleccionando aquella con mayor afinidad. Esta formulación permitió analizar el balance multimodal bajo dos regímenes distintos: uno centrado en la alineación entre modalidades y otro en la integración visual–textual para la selección de respuestas.

4.3. Modelos evaluados

Se evaluaron cuatro modelos médicos de visión–lenguaje tipo CLIP: **PubMedCLIP** [4], **BioMedCLIP** [18], **RCLIP** [6] y **WhyXRayCLIP** [17]. Estos modelos fueron seleccionados por su disponibilidad pública, su especialización biomédica o radiológica y su capacidad para estimar afinidad imagen–texto, lo que permite comparar distintos patrones de dependencia textual y visual.

[1] `https://www.kaggle.com/datasets/virajbagal/roco-dataset`
[2] `https://github.com/abachaa/VQA-Med-2019`

Todos los modelos fueron integrados bajo el mismo *pipeline* explicable. En ISA se evaluaron los cuatro modelos. En VQA solo se consideraron **PubMedCLIP** y **BioMedCLIP**, ya que son compatibles con la formulación de *similarity-based scoring*. **RCLIP** y **WhyXRayCLIP**, orientados principalmente a alineamiento imagen–texto, no se aplicaron directamente a esta tarea.

4.4. Implementación y configuración

Los experimentos se realizaron en *Google Colab Pro* utilizando una GPU NVIDIA T4 de 15 GB. El entorno experimental se implementó en Python 3.10 con `PyTorch`, `Transformers`, `OpenCLIP` y `SHAP` como librerías principales. Para el análisis y visualización de resultados se utilizaron `NumPy`, `Pandas` y `Matplotlib`. Adicionalmente, se empleó una infraestructura modular desarrollada para este estudio, orientada a la carga de datos, ejecución de inferencia y cálculo de explicaciones multimodales.

La configuración experimental mantuvo parámetros consistentes entre modelos para asegurar comparabilidad. Las explicaciones se calcularon con un presupuesto de entre 20 y 50 evaluaciones SHAP por instancia, dependiendo de la complejidad del modelo. La representación visual se definió en términos de 49, 196 o 256 parches, de acuerdo con el codificador visual correspondiente, mientras que la entrada textual se limitó a un máximo de 77 tokens siguiendo la configuración estándar de CLIP.

4.5. Protocolo de evaluación

El análisis se realizó a nivel de muestra y a nivel de conjunto de datos. Para cada instancia se obtuvieron valores SHAP por característica, que posteriormente se agregaron por modalidad para calcular las métricas de contribución textual y visual descritas en la sección anterior. Estas métricas se resumieron globalmente para comparar el balance multimodal entre modelos y tareas.

En ROCO, el *split* de validación se utilizó para análisis preliminar, mientras que el *split test* se reservó para el reporte final en ISA. En VQA-Med 2019, el *split* de entrenamiento se empleó para análisis preliminar y el *split test* para la evaluación final. Ambos conjuntos de datos se procesaron mediante cargadores dedicados que filtran muestras inválidas y estandarizan el formato multimodal de entrada.

5. Resultados y discusión

5.1. Resultados en ISA

La Tabla 1 resume los resultados de balance multimodal para la tarea de *Image–Sentence Alignment* (ISA) en el *split test* de ROCO.

En general, **PubMedCLIP** es el único modelo con comportamiento cercano al equilibrio, con $TScore = 0.544$, $IScore = 0.456$ y $\Delta = 0.089$, además del mayor porcentaje de muestras balanceadas (35.2 %).

En contraste, **BioMedCLIP**, **RCLIP** y **WhyXRayCLIP** presentan sesgo hacia la modalidad visual, reflejado en valores negativos de Δ y menores porcentajes de muestras balanceadas.

Este sesgo es moderado en **BioMedCLIP** ($\Delta = -0.219$) y **RCLIP** ($\Delta = -0.264$), y más pronunciado en **WhyXRayCLIP** ($\Delta = -0.384$), que además registra solo 0.4 % de muestras balanceadas.

En términos de similitud imagen–texto, **RCLIP** obtiene el mayor valor medio de ℓ_{norm} (0.849), mientras que **WhyXRayCLIP** presenta el menor (0.357). Estos resultados indican que un mayor alineamiento imagen–texto no implica necesariamente un mejor balance multimodal, ya que el modelo con mayor similitud media no es el más equilibrado.

Table 1. Resultados finales de balance multimodal y similitud imagen–texto en ISA (*split test*). Los mejores resultados se resaltan en negritas. Para Δ, se resalta el valor más cercano a cero.

Modelo	**TScore** ($\mu \pm \sigma$)	**IScore** ($\mu \pm \sigma$)	Δ (μ)	**% bal.** ($\lvert\Delta\rvert < 0.1$)	ℓ_{norm} ($\mu \pm \sigma$)
PubMedCLIP	$\mathbf{0.544 \pm 0.090}$	$\mathbf{0.456 \pm 0.090}$	**0.089**	**35.2**	0.515 ± 0.131
BioMedCLIP	0.390 ± 0.084	0.610 ± 0.084	-0.219	21.8	0.503 ± 0.141
RCLIP	0.368 ± 0.101	0.632 ± 0.101	-0.264	17.6	$\mathbf{0.849 \pm 0.131}$
WhyXRayCLIP	0.308 ± 0.061	0.692 ± 0.061	-0.384	0.4	0.357 ± 0.146

La Fig. 1 resume el comportamiento modal de los modelos evaluados en la tarea de ISA sobre el *split test*. **PubMedCLIP** es el único modelo con contribuciones relativamente equilibradas entre texto e imagen, mientras que **BioMedCLIP**, **RCLIP** y **WhyXRayCLIP** muestran un predominio de la modalidad visual. Este sesgo es más pronunciado en **WhyXRayCLIP**, que presenta la mayor diferencia entre *IScore* y *TScore*. En conjunto, la figura confirma que el balance multimodal varía significativamente entre arquitecturas, aun dentro de modelos especializados en el dominio biomédico.

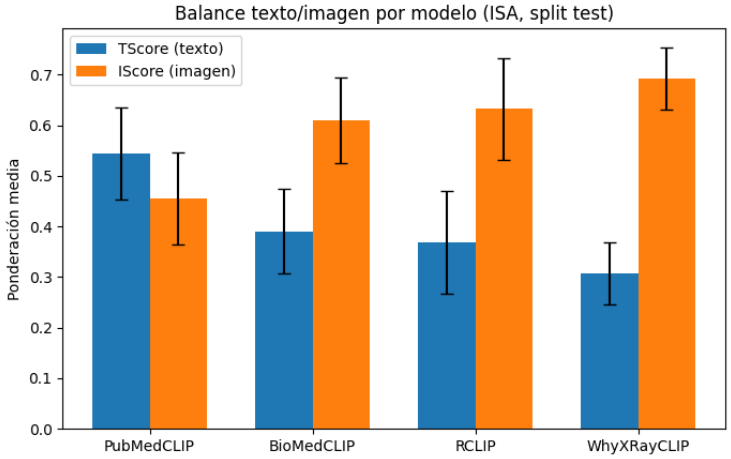


Fig. 1. Comparación de TScore e IScore por modelo en ISA (*split test*).

La Fig. 2 presenta un ejemplo cualitativo de ISA en ROCO. **PubMedCLIP** muestra la distribución más equilibrada entre tokens y parches visuales (*TScore* = 46.0%, *IScore* = 53.9%), mientras que **BioMedCLIP**, **RCLIP** y **WhyXRayCLIP** asignan mayor peso a la modalidad visual, con *IScore* superiores al 60%. Esto sugiere que, para esta muestra, la afinidad imagen–texto se apoya principalmente en regiones visuales en estos modelos, lo cual coincide con las tendencias cuantitativas del conjunto de prueba.

5.2. Resultados en VQA

La Tabla 2 resume los resultados de balance multimodal en la tarea de *Visual Question Answering* (VQA) sobre el *split test* de VQA-Med 2019. Ambos modelos presentan sesgo hacia la modalidad visual, aunque con distinta intensidad. **PubMedCLIP** muestra un sesgo moderado ($\Delta = -0.133$), con 29.2 % de muestras balanceadas, mientras que **BioMedCLIP** exhibe un sesgo visual mucho más pronunciado ($\Delta = -0.517$) y ausencia total de muestras balanceadas.

A pesar de esta diferencia en balance multimodal, ambos modelos alcanzan una exactitud similar (24.2 % en **PubMedCLIP** y 24.0 % en **BioMedCLIP**), lo que indica que el desempeño final no refleja por sí solo el grado de dependencia modal. En términos de similitud normalizada, ambos mantienen valores comparables, con ligera ventaja para **PubMedCLIP** (0.588 frente a 0.554).

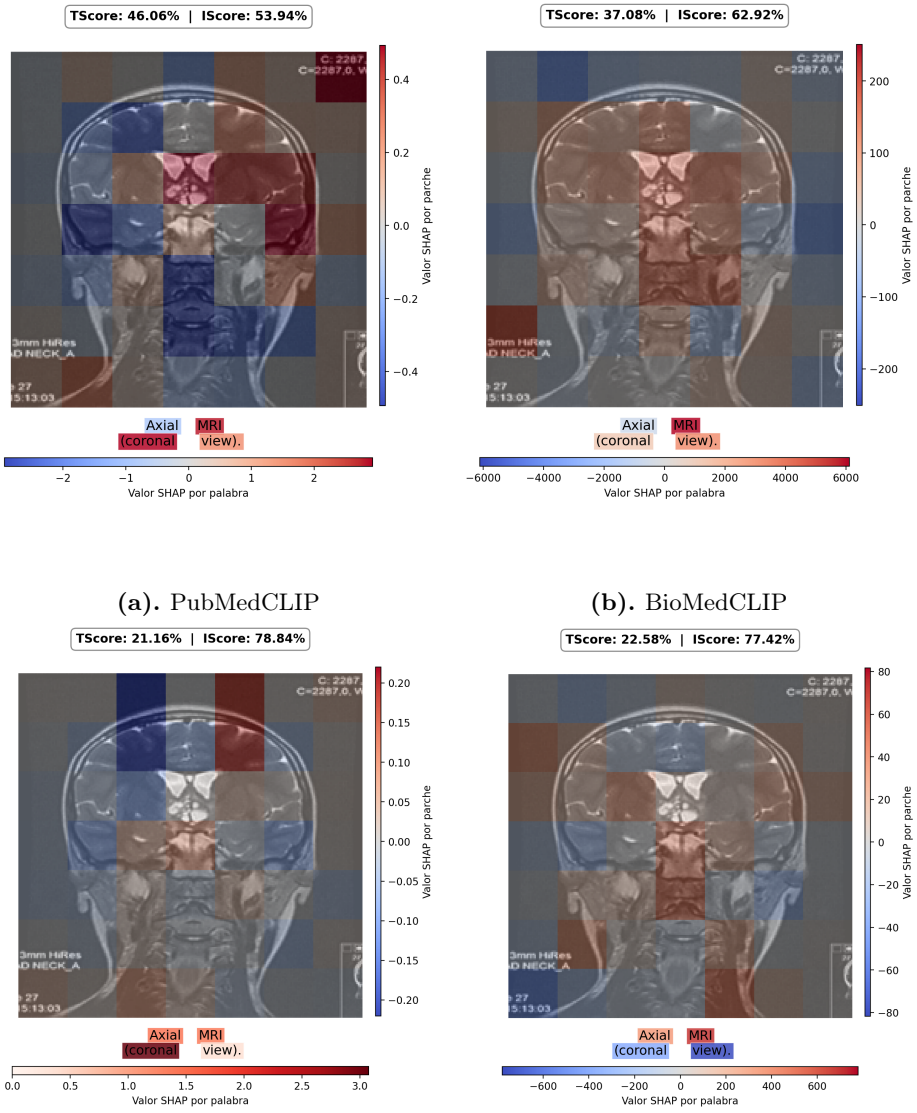
Table 2. Resultados finales de balance multimodal en VQA (*split test*). Los mejores resultados se resaltan en negritas. Para Δ, se resalta el valor más cercano a cero.

Modelo	**TScore** ($\mu \pm \sigma$)	**IScore** ($\mu \pm \sigma$)	Δ (μ)	**% bal.** ($\lvert\Delta\rvert < 0.1$)	**Accuracy (%)**	ℓ_{norm} ($\mu \pm \sigma$)
PubMedCLIP	$\mathbf{0.434 \pm 0.100}$	$\mathbf{0.566 \pm 0.100}$	$\mathbf{-0.133}$	**29.2**	**24.2**	$\mathbf{0.588 \pm 0.196}$
BioMedCLIP	0.242 ± 0.057	0.758 ± 0.057	-0.517	0.0	24.0	0.554 ± 0.172

La Fig. 3 resume el balance multimodal de los modelos evaluados en la tarea de VQA sobre el *split test*. En ambos casos se observa un predominio de la modalidad visual, aunque con distinta intensidad. **PubMedCLIP** presenta un sesgo moderado hacia la imagen, con valores de *TScore* e *IScore* relativamente cercanos, mientras que **BioMedCLIP** muestra una separación mucho mayor entre ambas métricas, evidenciando una dependencia visual más pronunciada. En conjunto, la figura confirma que dos modelos con exactitud similar pueden diferir sustancialmente en su balance multimodal. La Fig. 4 muestra un ejemplo cualitativo de MM-SHAP en VQA. En ambos modelos predomina la modalidad visual, aunque **BioMedCLIP** conserva una contribución textual relativamente mayor que **PubMedCLIP** en esta muestra. Este patrón coincide con la tendencia global hacia la modalidad visual observada en VQA.

5.3. Discusión

Los resultados muestran que el balance multimodal varía entre modelos y tareas. En ISA, **PubMedCLIP** fue el modelo más equilibrado, mientras que



(a). PubMedCLIP **(b).** BioMedCLIP

(c). RCLIP **(d).** WhyXRayCLIP

Fig. 2. Heatmaps ISA — Muestra 0.

BioMedCLIP, **RCLIP** y **WhyXRayCLIP** presentaron sesgo hacia la modalidad visual. En VQA, ambos modelos evaluados también mostraron predominio de la imagen, aunque este fue más marcado en **BioMedCLIP**.

Estas diferencias pueden explicarse por la interacción entre el preentrenamiento de cada modelo y la formulación de los conjuntos de datos. Mientras ISA evalúa

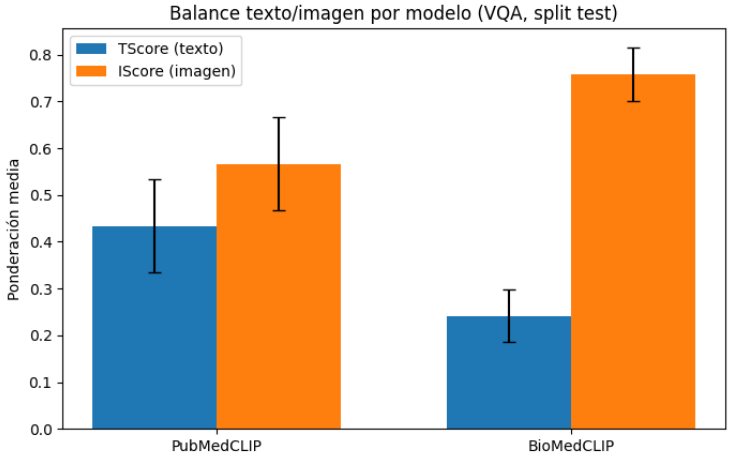


Fig. 3. Comparación de TScore e IScore por modelo en VQA (*split test*).

la correspondencia global entre imagen y texto, VQA requiere integrar imagen, pregunta y respuesta candidata. Por ello, el sesgo modal observado no debe atribuirse solo a la arquitectura, sino también a los datos de entrenamiento y a la tarea evaluada.

En conjunto, los resultados indican que métricas tradicionales como similitud imagen–texto o exactitud no reflejan por sí solas el grado de integración multimodal. Así, MM-SHAP y las métricas *TScore*, *IScore* y Δ aportan una caracterización complementaria del comportamiento de los modelos médicos de visión–lenguaje.

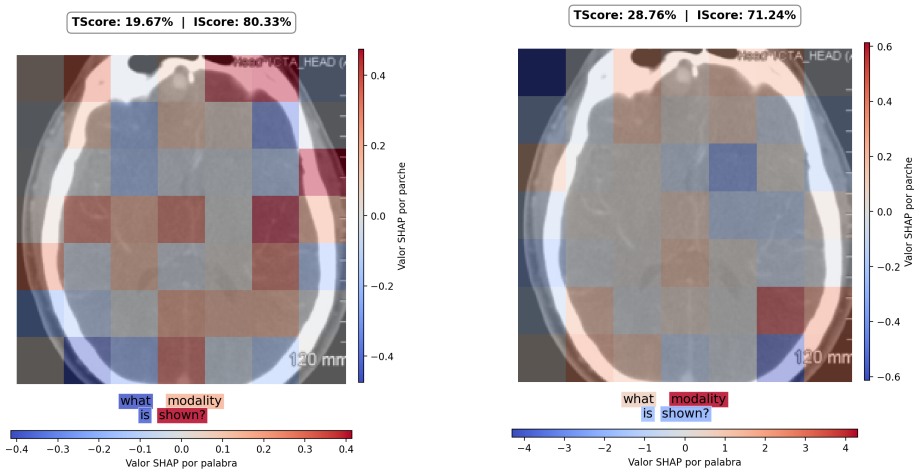
Desde una perspectiva práctica, este tipo de análisis puede apoyar la selección y auditoría de modelos médicos multimodales, al identificar casos en los que un desempeño alto oculta una dependencia excesiva de una sola modalidad.

6. Conclusiones

En este trabajo se presentó un marco de análisis para estudiar el balance multimodal en modelos médicos de visión–lenguaje mediante la aplicación experimental de MM-SHAP. El enfoque permitió cuantificar la contribución relativa de texto e imagen a nivel de muestra y de conjunto de datos, proporcionando una caracterización complementaria a las métricas tradicionales de desempeño.

Los resultados en ISA y VQA mostraron diferencias importantes entre modelos. **PubMedCLIP** presentó el balance más estable, mientras que **BioMedCLIP**, **RCLIP** y **WhyXRayCLIP** exhibieron distintos grados de sesgo hacia la modalidad visual. Asimismo, se observó que métricas como la similitud imagen–texto o la exactitud no reflejan por sí solas el grado de integración entre modalidades.

Como limitación, el análisis se restringe a dos tareas y a un conjunto acotado de modelos médicos tipo CLIP. Por ello, se plantea extender la evaluación a más



(a). PubMedCLIP **(b).** BioMedCLIP

Fig. 4. Heatmaps VQA — Muestra 0.

arquitecturas, conjuntos de datos y métodos de explicabilidad multimodal, así como estudiar la estabilidad de las explicaciones bajo distintas configuraciones.

Estos hallazgos evidencian la utilidad de incorporar métricas de explicabilidad multimodal en la evaluación de modelos médicos, especialmente en escenarios donde la interpretabilidad y la confianza son factores críticos. Además, este tipo de análisis puede apoyar la selección y auditoría de modelos al identificar dependencias excesivas hacia una sola modalidad.

References

1. Baltrusaitis, T., Ahuja, C., Morency, L.P.: Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence 41(2), 423–443 (Feb 2019)
2. Ben Abacha, A., Hasan, S.A., Datla, V.V., Liu, J., Demner-Fushman, D., Müller, H.: VQA-Med: Overview of the Medical Visual Question Answering Task at ImageCLEF 2019. In: Working Notes of CLEF 2019. CEUR Workshop Proceedings, vol. 2380 (2019), `https://ceur-ws.org/Vol-2380/paper_272.pdf`
3. Beňová, I., Gregor, M., Gatt, A.: Cv-probes: Studying the interplay of lexical and world knowledge in visually grounded verb understanding. arXiv preprint arXiv:2409.01389 (2024)
4. Eslami, S., de Melo, G., Meinel, C.: PubMedCLIP: How Much Does CLIP Benefit Visual Question Answering in the Medical Domain? In: Findings of the Association for Computational Linguistics: EACL 2023. pp. 1181–1193. Association for Computational Linguistics (2023), `https://aclanthology.org/2023.findings-eacl.88`
5. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In:

Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6904–6913 (2017)
6. Kaveh: RCLIP: CLIP Model Fine-Tuned on Radiology Images and Captions. https://huggingface.co/kaveh/rclip (2024), hugging Face model card. Consultado: 2026-05-07
7. Linden, T., De Jong, J., Lu, C., Kiri, V., Haeffs, K., Fröhlich, H.: An explainable multimodal neural network architecture for predicting epilepsy comorbidities based on administrative claims data. Frontiers in Artificial Intelligence 4, 610197 (2021), https://www.frontiersin.org/journals/artificial-intelligence
8. Liu, C., Jin, Y., Guan, Z., et al.: Visual–language foundation models in medicine. The Visual Computer (2024)
9. Liu, J., Zhang, Y., Chen, J., Xiao, J., Lu, Y., Landman, B., Yuan, Y., Yuille, A., Tang, Y., Zhou, Z.: Clip-driven universal model for organ segmentation and tumor detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21152–21164 (2023)
10. Parcalabescu, L., Frank, A.: Mm-shap: A performance-agnostic metric for measuring multimodal contributions in vision and language models and tasks. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics (2023)
11. Pelka, O., Koitka, S., Rückert, J., Nensa, F., Friedrich, C.M.: Radiology Objects in COntext (ROCO): A Multimodal Image Dataset. In: Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis. Lecture Notes in Computer Science, vol. 11043, pp. 180–189. Springer (2018)
12. Pellegrini, C., Keicher, M., Özsoy, E., Jiraskova, P., Braren, R., Navab, N.: Xplainer: From x-ray observations to explainable zero-shot diagnosis. arXiv preprint arXiv:2303.13391 (2023)
13. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763. PMLR (18–24 Jul 2021), https://proceedings.mlr.press/v139/radford21a.html
14. Tiu, E., Talius, E., Patel, P., Langlotz, C.P., Ng, A.Y., Rajpurkar, P.: Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. Nature Biomedical Engineering 6, 1399–1406 (2022)
15. Truhn, D., Eckardt, J.N., Ferber, D., Kather, J.N.: Large language models and multimodal foundation models for precision oncology. NPJ Precision Oncology 8(1), 72 (2024)
16. Wang, J., Mao, Y., Guan, N., Xue, C.J.: Shap-cat: A interpretable multi-modal framework enhancing wsi classification via virtual staining and shapley-value-based multimodal fusion. arXiv preprint arXiv:2410.01408 (2024)
17. YYUPenn: WhyXrayCLIP. https://huggingface.co/yyupenn/whyxrayclip (2024), hugging Face model card. Accessed: 2026-05-07
18. Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., Wong, C.C., Tupini, A., Wang, Y., Mazzola, M., Shukla, S., Liden, L., Gao, J., Crabtree, A., Piening, B., Bifulco, C., Lungren, M.P., Naumann, T., Poon, H.: Biomedclip: A multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv preprint arXiv:2303.00915 (2023)

19. Zhao, Z., Liu, Y., Wu, H., Li, Y., Wang, S., Teng, L., Liu, D., Cui, Z., Wang, Q., Shen, D.: Clip in medical imaging: A comprehensive survey. arXiv preprint arXiv:2312.07353 (2023)