Imagen Diagnóstica
https://imagendiagnostica.org/

Imadi. 2024, **15**(1), 1–9.

Una evaluación sistemática de la competencia multimodal de GPT-4V en el análisis de radiografías de tórax

Minerva Montero Díaz[1,*], Roberto Rodríguez Morales[2], Luis Antonio Rodríguez Sánchez[3]

[1]Departamento de Matemática, Instituto de Cibernética, Matemática y Física, La Habana, Cuba
[2] Departamento de Matemática Interdisciplinaria, Instituto de Cibernética, Matemática y Física, La Habana, Cuba
[3] Departamento de Neurología, Hospital General Docente Roberto Rodríguez , Ciego de Ávila, Cuba

(Received: 12 June 2024. Accepted: 23 September 2024. Published online: 30 December 2024.)

Resumen

Este estudio presenta una evaluación sistemática del desempeño multimodal de GPT-4V para el análisis de radiografías de tórax en tres tareas clínicamente relevantes: generación de informes radiológicos, respuesta a preguntas visuales médicas y localización visual médica. Para cada tarea, diseñamos conjuntos de prompts específicos orientados a elicitar las competencias propias de cada una necesarias para obtener resultados clínicamente significativos (p. ej., elaboración de informes estructurados, razonamiento condicionado por preguntas y localización a nivel regional). Evaluamos GPT-4V mediante tres enfoques complementarios —puntuación cuantitativa automática, evaluación humana experta y estudios de casos cualitativos— con el fin de aportar tanto amplitud como profundidad al análisis. Los resultados muestran que GPT-4V demuestra una sólida comprensión global de las radiografías de tórax, generando informes de alta calidad y respondiendo con precisión a muchas consultas clínicas fundamentadas en la imagen. Sin embargo, su capacidad de localización visual sigue siendo comparativamente débil, lo que limita una localización fiable y el razonamiento específico por regiones. Además, observamos una discrepancia consistente entre las conclusiones sugeridas por las métricas automáticas estándar y aquellas derivadas del juicio experto, lo que subraya la necesidad de protocolos de evaluación clínicamente alineados para los modelos multimodales de lenguaje de gran escala en radiología.

Palabras clave: radiografía de tórax; modelo multimodal de lenguaje de gran escala; GPT-4V; generación de informes radiológicos; respuesta a preguntas visuales médicas; localización visual; evaluación humana; métricas clínicamente alineadas.

1. Introducción

Los modelos de lenguaje de gran tamaño (LLM, por sus siglas en inglés) han avanzado rápidamente en su capacidad para generar texto coherente y sensible al contexto, así como para resolver diversas tareas centradas en el lenguaje [1–3]. Motivado por la necesidad de extender estas capacidades más allá del texto, trabajos recientes han impulsado el desarrollo de grandes modelos multimodales (LMM) que razonan conjuntamente sobre imágenes y lenguaje [4–6]. Entre ellos, GPT-4V integra comprensión visual con generación guiada por instrucciones, lo que le permite interpretar imágenes y producir salidas ricas en lenguaje natural. Estudios iniciales sobre razonamiento visual genérico y comprensión de imágenes en entornos abiertos informan un rendimiento sólido, pero también destacan limitaciones en tareas que requieren localización precisa, restricciones estrictas de salida o atribución fina a nivel de región [7, 8].

En medicina y atención sanitaria, los sistemas de razonamiento multimodal podrían reducir la carga de documentación, mejorar el acceso a la experiencia especializada y apoyar la toma de decisiones clínicas [9, 10]. La radiología es un banco de pruebas especialmente convincente porque la interpretación clínica depende de un acoplamiento estrecho entre la evidencia visual (p. ej., hallazgos de imagen) y un lenguaje preciso (p. ej., impresiones estructuradas, consideraciones diferenciales y recomendaciones). Las radiografías de tórax se encuentran entre los estudios de imagen más solicitados y son centrales para el cribado y el triaje en flujos de trabajo de urgencias, hospitalización y consulta externa. Sin embargo, el uso clínico requiere más que texto fluido: los modelos deben (i) generar informes clínicamente fieles, (ii) responder de forma fiable a preguntas fundamentadas en la imagen y (iii) apoyar el razonamiento específico por regiones cuando la localización es necesaria para la transparencia, la auditabilidad o el uso posterior. Las investigaciones médicas existentes

*Corresponding author (minerva@icimaf.cu)

sobre GPT-4V han comenzado a explorar estas capacidades, pero muchas siguen siendo limitadas en escala, se centran en una sola tarea o se basan principalmente en estudios de caso ilustrativos [11]. En consecuencia, siguen abiertas cuestiones importantes sobre el rendimiento multitarea, la sensibilidad a los *prompts*, los modos de fallo clínicamente relevantes y, de manera crítica, la validez de métricas automáticas comunes de evaluación en radiología.

Para abordar estas lagunas, presentamos una evaluación sistemática de la competencia multimodal de GPT-4V en el análisis de radiografías de tórax a través de tres tareas clínicas centrales: (i) generación de informes radiológicos, (ii) pregunta–respuesta visual médica (VQA) y (iii) *visual grounding* médico. Estas tareas exploran competencias complementarias: la generación de informes evalúa la comprensión global de la imagen y la síntesis con estilo clínico; VQA evalúa el razonamiento condicionado por la pregunta; y el *grounding* visual pone a prueba si el modelo puede vincular el lenguaje con evidencia espacialmente localizada necesaria para la interpretación específica por región. Para cada tarea, diseñamos conjuntos de *prompts* orientados a inducir los comportamientos relevantes tanto en escenarios *zero-shot* como *few-shot*, permitiendo un análisis fundamentado del rendimiento y de la sensibilidad al *prompting*.

Un desafío central en este ámbito es la evaluación: las métricas automáticas convencionales pueden divergir del juicio clínico cuando múltiples formulaciones son aceptables, cuando errores clínicamente relevantes no se capturan mediante solapamiento superficial de texto, o cuando los *benchmarks* imponen formatos rígidos de respuesta. En consecuencia, adoptamos un marco de evaluación de tres componentes que combina (a) puntuación automática, (b) valoración humana por expertos y (c) estudios de caso cualitativos. Este diseño permite la comparabilidad con trabajos previos a la vez que preserva la interpretabilidad clínica y facilita el análisis de errores. En conjunto, observamos que GPT-4V presenta un buen desempeño en la comprensión global de radiografías de tórax (informes y muchas preguntas VQA), pero sigue siendo relativamente débil en *visual grounding*, donde se requiere localización fiable. Además, documentamos una discordancia sistemática entre métricas y juicio experto, lo que motiva protocolos de evaluación clínicamente alineados que reflejen mejor el criterio especialista y la utilidad en el mundo real.

Nuestras contribuciones principales son:

- **Evaluación clínica multitarea.** Evaluamos GPT-4V en radiografías de tórax a través de generación de informes, VQA médica y *visual grounding* médico dentro de un diseño experimental unificado, lo que permite una interpretación coherente entre tareas.
- **Análisis de sensibilidad al *prompting*.** Introducimos conjuntos de *prompts* orientados a cada tarea (variantes *zero-shot* y *few-shot*) y analizamos cómo la composición de ejemplos y las restricciones de salida afectan el rendimiento y los modos de fallo.
- **Evaluación con fundamento clínico.** Triangulamos métricas automáticas con evaluación humana experta y estudios de caso cualitativos, y cuantificamos explícitamente el acuerdo entre métricas y expertos para revelar puntos ciegos de evaluación.
- **La localización como cuello de botella.** Mostramos que el *visual grounding* sigue siendo una debilidad clave frente a tareas centradas en texto, resaltando una barrera práctica para el razonamiento por regiones y la explicabilidad en radiología.

2. Metodología y Experimentos

La Figura 1 resume nuestro flujo de evaluación: (i) definir criterios específicos por tarea y requisitos de salida, (ii) diseñar *prompts* para inducir esas competencias, (iii) generar salidas del modelo, (iv) evaluar mediante métricas automáticas y valoración experta, y (v) sintetizar tendencias cuantitativas con análisis de casos cualitativos. Describimos las tareas de evaluación (§2.1), el proceso de evaluación (§2.2), los conjuntos de datos (§2.3) y la organización de resultados (§2.4).

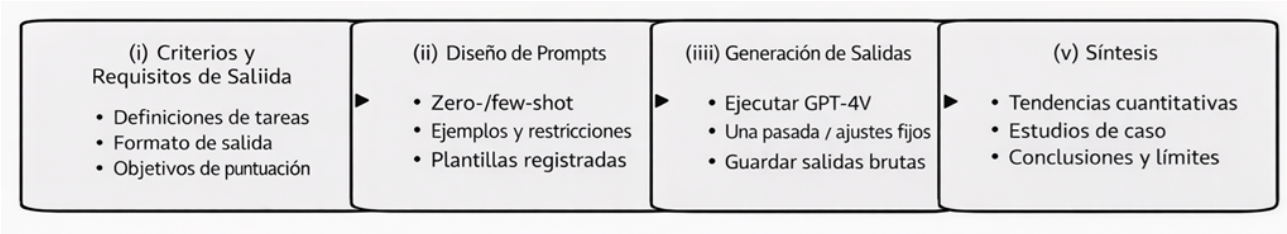

Figure 1: Visión general del flujo de evaluación: (i) definir criterios específicos por tarea y requisitos de salida, (ii) diseñar *prompts* para inducir competencias objetivo, (iii) generar salidas del modelo, (iv) evaluar mediante métricas automáticas y valoración experta, y (v) sintetizar tendencias cuantitativas con análisis de casos cualitativos.

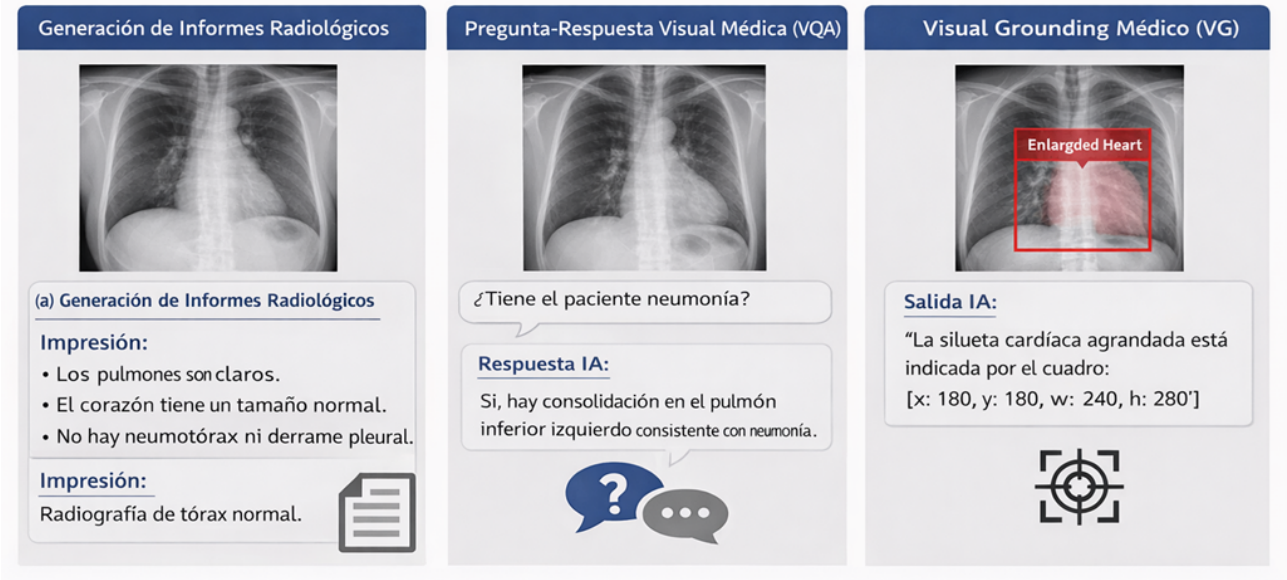

Figure 2: Ejemplos de las tres tareas de evaluación para el análisis de radiografías de tórax: (a) generación de informes radiológicos (R2Gen), (b) pregunta–respuesta visual médica (VQA) y (c) *visual grounding* (VG). En conjunto, estas tareas cubren la interpretación global, el razonamiento condicionado por la pregunta y la localización a nivel de región.

2.1 Tareas de Evaluación

Evaluamos GPT-4V en tres tareas multimodales representativas del análisis de radiografías de tórax (Figura 2): generación de informes radiológicos (R2Gen), pregunta–respuesta visual médica (VQA) y *visual grounding* médico (VG). En conjunto, estas tareas cubren la interpretación global, el razonamiento condicionado por la pregunta y la localización a nivel de región.

La generación de informes radiológicos (R2Gen) está estrechamente relacionada con el *image captioning*, pero impone requisitos sustancialmente mayores en cuanto a corrección médica, completitud y estilo narrativo estructurado [12–14]. R2Gen es desafiante debido a (i) la redacción de texto largo, (ii) anormalidades sutiles y de grano fino, y (iii) el sesgo hacia estudios normales (el problema de *data bias*), que puede inducir salidas excesivamente normales o genéricas.

Trabajos previos suelen seguir dos direcciones: mejoras arquitectónicas para una mejor extracción de características visuales y generación de informes (p. ej., diseños jerárquicos y basados en *Transformers*) [15–17], y mitigación del sesgo de datos mediante conocimiento externo o señales clínicas auxiliares [18–21]. Más recientemente, se han explorado enfoques basados en LLM para la generación de informes radiológicos de formato largo [9, 19].

La pregunta–respuesta visual (VQA) asigna a un par imagen–pregunta una respuesta fundamentada en el contenido de la imagen [22, 23]. Los sistemas tradicionales de VQA médica suelen formularse como problemas de clasificación [24–26], mientras que los métodos de generación reflejan mejor el cuestionamiento clínico de tipo

abierto [27, 28]. Los LLM también han habilitado la construcción de conjuntos de datos [29] y han mejorado el razonamiento en VQA [30], lo que convierte a VQA en una dimensión crítica para la evaluación de GPT-4V.

El *visual grounding* médico (VG) requiere predecir un cuadro delimitador (*bounding box*) correspondiente a una frase que describe un hallazgo o una región anatómica [31]. El progreso se ha visto limitado por la escasez de datos médicos de *grounding*, pero MS-CXR ha permitido un *benchmarking* sistemático y el desarrollo de nuevos métodos [32–35]. A diferencia de la generación de informes y VQA, las salidas de VG son coordenadas estructuradas en lugar de texto libre; trabajos recientes indican que modelos de estilo LLM pueden ser guiados mediante *prompts* para producir coordenadas directamente, lo que motiva nuestra evaluación del comportamiento de *grounding* de GPT-4V.

2.2 Método/Proceso de Evaluación

Nuestro diseño de evaluación busca ser (i) clínicamente interpretable, (ii) comparable con *benchmarks* previos y (iii) reproducible. Adoptamos un protocolo centrado en *prompts* con restricciones explícitas de salida, y triangulamos métricas automáticas, evaluación humana experta y estudios de caso cualitativos.

Para reducir variabilidad evitable, estandarizamos la configuración de inferencia entre tareas y registramos explícitamente: plantillas de *prompts*, reglas de selección de ejemplos, restricciones de salida y reglas de posprocesamiento. Salvo indicación contraria, cada instancia imagen–*prompt* se consulta una sola vez y se evalúa sobre la salida producida, reflejando un escenario de despliegue realista. Cuando los parámetros de generación son configurables, recomendamos reportarlos explícitamente (p. ej., temperatura, longitud máxima de salida y cualquier instrucción a nivel de sistema) y mantenerlos fijos entre tareas para garantizar comparabilidad.

Evaluamos:

- ***Zero-shot.*** GPT-4V recibe la radiografía de tórax y una instrucción para producir un informe con *Hallazgos* e *Impresión*.
- ***Few-shot* (aprendizaje en contexto).** Proporcionamos pares ejemplo imagen–informe para enseñar la estructura y calibrar el lenguaje normal/anormal sin actualizar parámetros. Probamos tres composiciones: solo normales, solo anormales y mixtas (un normal + un anormal). En nuestros análisis cualitativos, los ejemplos mixtos mejoran de forma consistente la calibración y reducen el sesgo inducido por el *prompt*; por ello, adoptamos *few-shot* mixto para experimentos de informes a escala de *benchmark* en MIMIC-CXR.

Calculamos BLEU [36], ROUGE [37], METEOR [38] y CIDEr [39]. Estas métricas capturan propiedades complementarias de solapamiento léxico, pero pueden penalizar paráfrasis clínicamente válidas o enunciados visualmente correctos que no aparecen en el informe de referencia. Para garantizar equidad, aplicamos una normalización de texto consistente (p. ej., normalización de espacios en blanco y conversión a minúsculas) en todos los sistemas evaluados.

Dado que el solapamiento léxico puede no reflejar la corrección clínica, adicionalmente calculamos puntuaciones clínicas basadas en etiquetas extrayendo un conjunto fijo de observaciones clínicas tanto de los informes generados como de los informes de referencia mediante un etiquetador automático estándar de informes (p. ej., CheXbert) y luego reportamos precisión, exhaustividad (recall) y F1 sobre las etiquetas correspondientes. (Eliminamos aquí la cita conflictiva para evitar referenciar erróneamente a GPT-4V como etiquetador.) Esta vista complementaria refleja mejor si se preservan hallazgos clínicamente significativos, incluso cuando la redacción difiere.

Un radiólogo entrenado evalúa una muestra estratificada de pares de informes (referencia vs. GPT-4V) en términos de relevancia, corrección factual, consistencia interna entre *Hallazgos* e *Impresión*, completitud y legibilidad. Analizamos el acuerdo entre las valoraciones humanas y las métricas automáticas para cuantificar la discordancia entre métricas y expertos.

Los *benchmarks* de VQA médica a menudo restringen el formato de respuesta. Para reducir la discrepancia de formato en salidas generativas, adoptamos un *prompt few-shot* que (i) enseña a GPT-4V a clasificar las preguntas como cerradas vs. abiertas y (ii) restringe las respuestas cerradas a una forma mínima (p. ej., *sí/no*), permitiendo prosa informativa para preguntas abiertas (Figura 3). Este diseño busca mejorar tanto la corrección como la evaluabilidad.

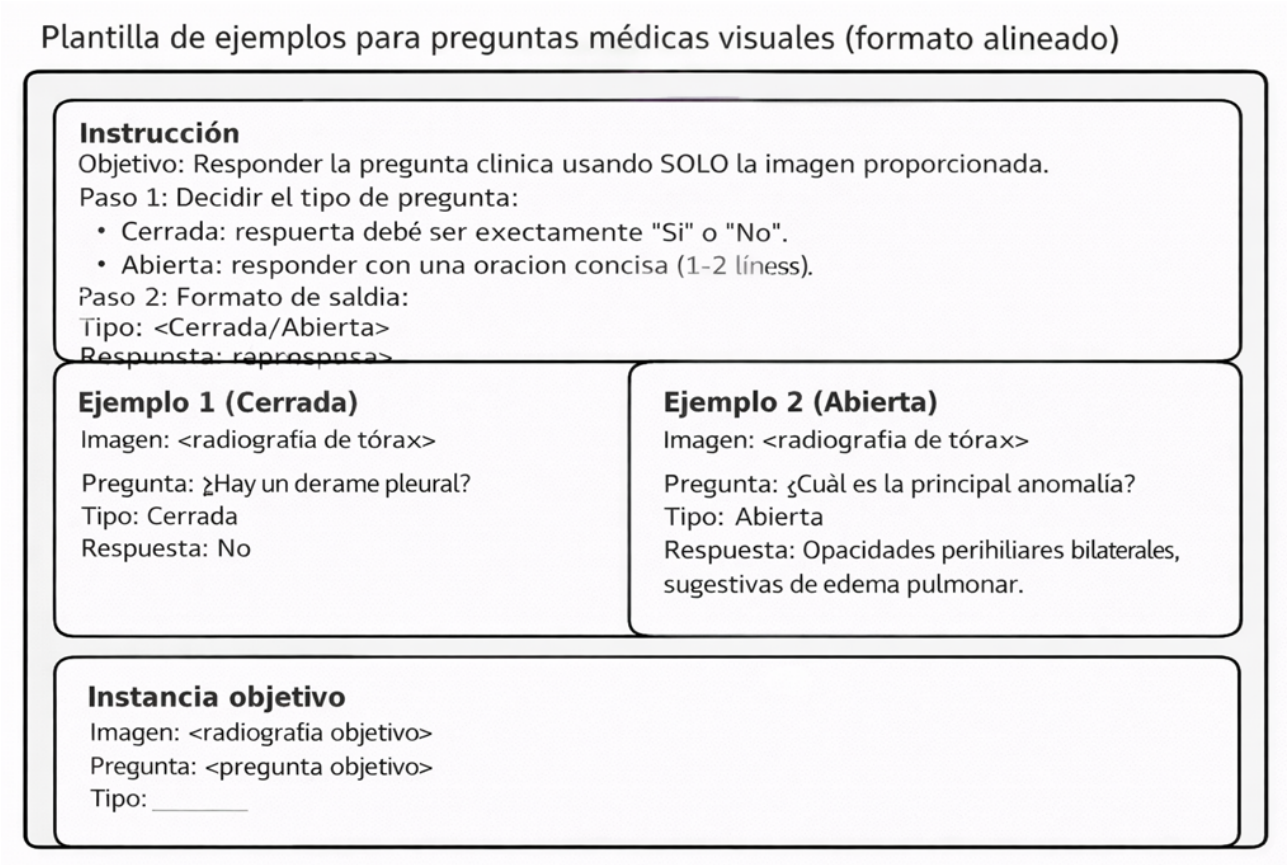

Figure 3: Plantilla de *prompt few-shot* para VQA médica utilizada para reducir la discrepancia de formato en salidas generativas. El *prompt* enseña a GPT-4V a clasificar las preguntas como cerradas frente a abiertas y restringe las respuestas cerradas a una forma mínima (p. ej., "Sí/No"), mientras permite prosa concisa para preguntas abiertas.

Reportamos exactitud para preguntas cerradas y BLEU-4 [36] para preguntas abiertas (en línea con reportes previos). Para preguntas cerradas, aplicamos una regla de normalización transparente que mapea variantes comunes (p. ej., "Sí.", "No, no lo es.") a etiquetas canónicas antes de puntuar; reportamos la regla explícitamente para asegurar reproducibilidad. Para preguntas abiertas, calculamos BLEU-4 tras la normalización estándar del texto, señalando que las puntuaciones basadas en solapamiento pueden infravalorar respuestas semánticamente correctas pero léxicamente diversas.

Para abordar limitaciones conocidas de la puntuación automática rígida en VQA generativa, además realizamos una calificación humana experta de un subconjunto muestreado de preguntas abiertas, reportando la proporción juzgada clínicamente correcta y analizando modos de fallo representativos (p. ej., respuestas basadas en conocimiento que descuidan la imagen).

Para VG, empleamos un *prompt* con formato restringido que instruye a GPT-4V a producir un único cuadro delimitador para la frase objetivo (Figura 4). Evitamos ejemplos en contexto para prevenir sobreajuste a formulaciones específicas de los ejemplos y para evaluar mejor la generalización.

Requerimos cuadros delimitadores en un orden fijo (p. ej., $(x_{\min}, y_{\min}, x_{\max}, y_{\max})$). Las predicciones que violan restricciones básicas (p. ej., $x_{\min} \geq x_{\max}$) se tratan como inválidas y se puntúan con IoU cero para evitar inflar el rendimiento mediante correcciones *post-hoc*.

Calculamos el *mean Intersection-over-Union* (mIoU) a través de todas las muestras de prueba y comparamos frente a líneas base publicadas en MS-CXR [40–45]. También incluimos ejemplos cualitativos de *grounding* para ilustrar errores característicos de localización.

Para cuantificar la validez de las métricas, reportamos (i) la correlación de rangos entre métricas automáticas y valoraciones expertas usando Kendall τ, y (ii) las pruebas de significación asociadas bajo una hipótesis nula de ausencia de asociación. Cuando corresponde, recomendamos reportar incertidumbre para métricas agregadas (p.

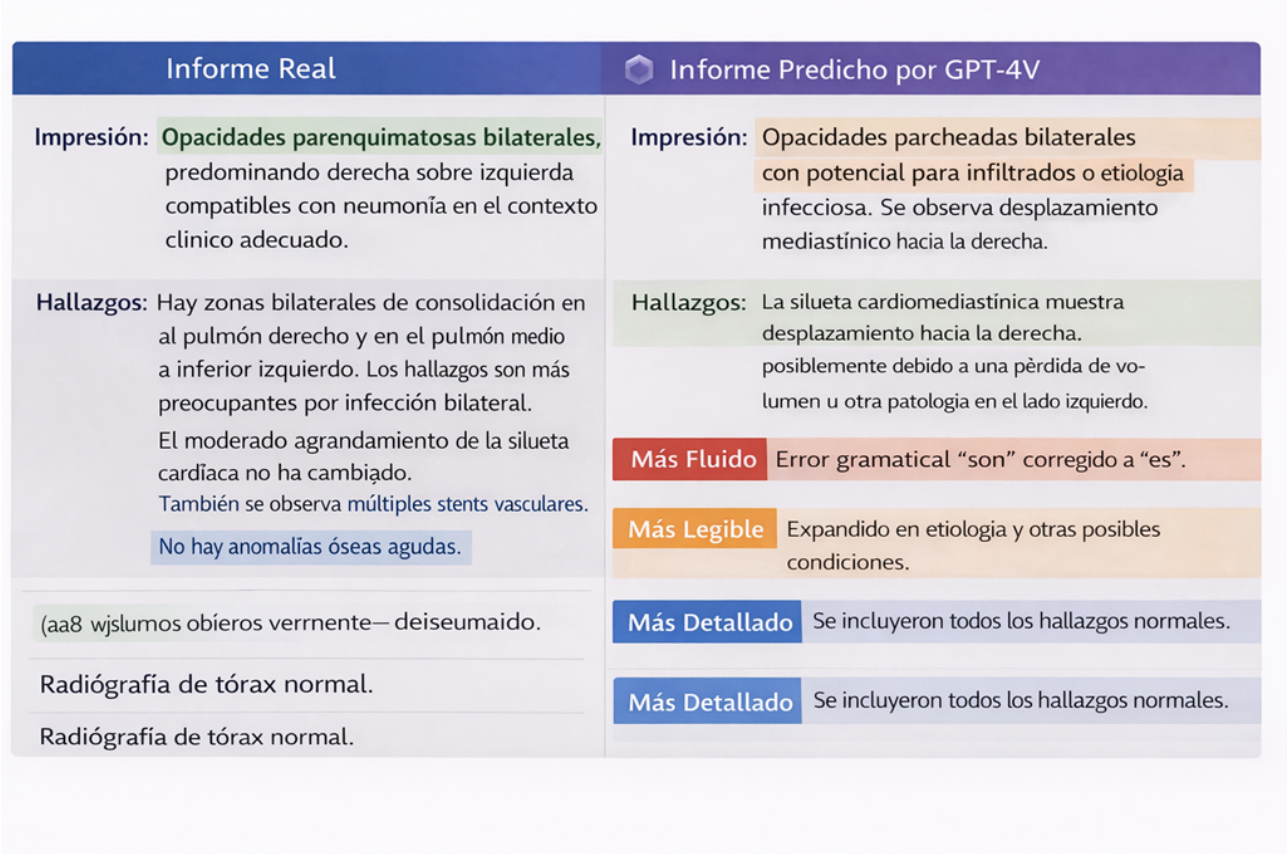

Figure 4: Plantilla de *prompt* con formato restringido para *visual grounding* (VG) médico. El *prompt* instruye a GPT-4V a producir un único cuadro delimitador para la frase objetivo en un formato fijo de coordenadas. Se evitan deliberadamente ejemplos en contexto para reducir el sobreajuste a formulaciones específicas de los ejemplos y evaluar mejor la generalización.

ej., intervalos de confianza bootstrap al 95% sobre las muestras de prueba) para distinguir diferencias sustantivas del ruido muestral.

2.3 Conjuntos de Datos de Evaluación

Usamos MIMIC-CXR, un conjunto de datos a gran escala de radiografías de tórax emparejadas con informes de texto libre. Seguimos la partición oficial y evaluamos en el conjunto de prueba (3,858 estudios). Para revisión experta, muestreamos 100 pares de informes (referencia vs. GPT-4V) del conjunto de prueba.

Usamos VQA-RAD, que contiene 315 imágenes radiológicas y 3,515 pares pregunta–respuesta. Seguimos la partición oficial y evaluamos en el subconjunto de prueba (451 pares QA). Para revisión experta, muestreamos 100 pares QA de preguntas abiertas para calificación por radiólogo.

Usamos MS-CXR, que proporciona pares imagen–oración con cuadros delimitadores a nivel de frase anotados por radiólogos certificados. Evaluamos el rendimiento de *grounding* usando mIoU y comparamos con líneas base publicadas en este *benchmark*.

2.4 Resultados de la Evaluación

Presentamos resultados en tres vistas complementarias: (i) resultados cuantitativos de *benchmark* (§2.4.1), (ii) evaluación humana experta (§2.4.2) y (iii) estudios de caso cualitativos (§2.4.3).

2.4.1 Resultados cuantitativos

Comparamos GPT-4V frente a modelos estándar de *captioning* [12,13,16] y métodos especializados de generación de informes [17] en MIMIC-CXR. Reportamos BLEU/ROUGE/METEOR/CIDEr [36–39] y precisión/exhaustividad/F1

basados en etiquetas clínicas, y analizamos cómo el *prompting few-shot* mixto afecta el rendimiento en relación con el *prompting zero-shot*.

En VQA-RAD, comparamos GPT-4V con sistemas representativos de VQA médica [24–26, 28], reportando exactitud en preguntas cerradas y BLEU-4 [36] en preguntas abiertas. Adicionalmente, analizamos modos de error inducidos por rigidez de formato y verbosidad en respuestas cerradas.

En MS-CXR [35], evaluamos GPT-4V frente a líneas base de *grounding* médico [35, 40–45] usando mIoU. Reportamos ejemplos cualitativos para contextualizar fallos típicos de localización.

2.4.2 Resultados de evaluación humana

Realizamos evaluación humana experta para generación de informes y VQA, donde la corrección semántica no se captura de forma fiable mediante métricas basadas en solapamiento. Para VG, utilizamos mIoU y visualización directa.

Un radiólogo entrenado califica 100 pares de informes en niveles ordinales de calidad según corrección clínica, relevancia, consistencia interna, completitud y legibilidad. Cuantificamos el acuerdo entre métricas y expertos mediante Kendall's τ y analizamos discrepancias para identificar modos de fallo de métricas convencionales.

Un radiólogo entrenado juzga la corrección en un subconjunto muestreado de preguntas abiertas. Contrastamos la corrección según el experto con la puntuación automática estricta y proporcionamos ejemplos representativos donde respuestas semánticamente correctas son penalizadas por formatos rígidos del *benchmark*.

2.4.3 Estudios de caso

Proporcionamos ejemplos cualitativos a través de tareas para revelar fortalezas y patrones de fallo clínicamente significativos que pueden quedar ocultos por puntuaciones agregadas, incluyendo sesgo normal/anormal inducido por *prompting* en informes, errores de *image-neglect* en VQA y cuadros delimitadores burdos o anatómicamente desalineados en *grounding*.

3. Conclusiones y Limitaciones

3.1 Conclusiones

Presentamos una evaluación sistemática de la competencia multimodal de GPT-4V en el análisis de radiografías de tórax a través de tres tareas clínicas—generación de informes radiológicos, VQA médica y *visual grounding* médico—utilizando un marco de evaluación de tres componentes que combina puntuación automática, evaluación humana experta y estudios de caso cualitativos. Este diseño permite tanto comparabilidad con *benchmarks* como un análisis clínicamente interpretable de los modos de fallo.

En conjunto, GPT-4V muestra un desempeño sólido en comprensión global de radiografías de tórax: puede generar informes coherentes con estilo clínico y responder con precisión a muchas preguntas clínicas fundamentadas en la imagen. Sin embargo, el rendimiento depende de forma significativa del diseño del *prompt*. En particular, ejemplos *few-shot* mixtos (normal/anormal) mejoran la calibración y reducen el sesgo inducido por el *prompt* respecto a ejemplos solo normales o solo anormales. También observamos que métricas basadas en solapamiento (p. ej., BLEU [36], CIDEr [39]) pueden infravalorar salidas clínicamente plausibles cuando GPT-4V utiliza paráfrasis válidas o menciona hallazgos visualmente evidentes ausentes del informe de referencia, contribuyendo a la discordancia entre métricas y expertos.

Para VQA, la flexibilidad generativa de GPT-4V produce respuestas clínicamente informativas, pero es penalizada por formatos rígidos del *benchmark* y convenciones estrictas de emparejamiento diseñadas originalmente para sistemas de tipo clasificación [22–24, 27, 28]. La revisión experta indica de forma consistente una corrección mayor de la que sugieren las puntuaciones automáticas convencionales, lo que subraya la necesidad de métodos de evaluación que midan la corrección semántica y contextual, más allá del emparejamiento léxico exacto.

En *visual grounding*, GPT-4V sigue siendo comparativamente débil: aunque puede generar coordenadas de cuadros delimitadores, la localización precisa en radiografías de tórax es poco fiable, lo que limita el razonamiento específico por región y la explicabilidad. Enfoques recientes como *Set-of-Mark prompting* [46] podrían ofrecer

una vía práctica para mejorar el *grounding* al hacer más explícitas las referencias regionales; establecer su efectividad para la localización fina específica de radiología sigue siendo una dirección importante para trabajos futuros.

En conjunto, nuestros resultados destacan dos implicaciones clave: (i) GPT-4V ya es capaz en varias tareas radiológicas centradas en texto, pero (ii) la localización fiable y una evaluación clínicamente alineada siguen siendo cuellos de botella para un despliegue seguro e interpretable.

3.2 Limitaciones

Este estudio tiene limitaciones que motivan trabajos futuros. Primero, la evaluación humana experta está necesariamente limitada en escala. Aunque nuestra revisión muestreada está alineada con estudios médicos multimodales previos, no puede capturar plenamente la diversidad de hallazgos y estilos de informe en conjuntos grandes como MIMIC-CXR. Futuras investigaciones deberían considerar muestreo estratificado o selección de subconjuntos basada en *clustering* para cubrir mejor hallazgos raros y casos límite difíciles.

Segundo, nuestra evaluación experta utiliza un esquema ordinal de calificación relativamente grueso. Un rúbrica más granular—por ejemplo, corrección por hallazgo, omisiones/comisiones clínicamente significativas y accionabilidad—podría ofrecer una visión diagnóstica más fina y mejorar la fiabilidad de la medición.

Tercero, nuestro enfoque principal es la inferencia basada en *prompting*. Canalizaciones aumentadas con herramientas o estructuras (p. ej., propuestas explícitas de regiones, andamiajes de segmentación o marcado regional estilo SoM [46]) pueden afectar sustancialmente el *grounding* y deberían evaluarse de forma sistemática. Finalmente, los *benchmarks* públicos actuales imponen restricciones (p. ej., formatos fijos de respuesta en VQA-RAD y definiciones frase–cuadro en MS-CXR [35]) que pueden confundir la evaluación de modelos generativos. Es probable que el progreso requiera nuevos conjuntos de datos y métricas clínicamente alineadas que recompensen la corrección semántica, penalicen errores clínicamente relevantes y acomoden la variabilidad aceptable en la redacción experta.

Referencias

[1] Touvron H, Lavril T, Izacard G, et al. Llama: Open and Efficient Foundation Language Models. arXiv preprint arXiv:2302.13971. 2023.

[2] OpenAI. Gpt-4 technical report. ArXiv abs/2303.08774; 2023. https://api.semantics cholar.org/CorpusID:257532815.

[3] Anil R, Dai AM, Firat O, et al. Palm 2 Technical Report. arXiv preprint arXiv: 2305.10403. 2023.

[4] Ye Q, Xu H, Xu G, et al. Mplug-Owl: Modularization Empowers Large Language Models with Multimodality. arXiv preprint arXiv:2304.14178. 2023.

[5] Li J, Li D, Savarese S, Hoi S. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597. 2023.

[6] Awadalla A, Gao I, Gardner J, et al. Openflamingo: An Open-Source Framework for Training Large Autoregressive Vision-Language Models. arXiv preprint arXiv: 2308.01390. 2023.

[7] Wu Y, Wang S, Yang H, et al. An Early Evaluation of Gpt-4v (Ision). arXiv preprint arXiv:2310.16534. 2023.

[8] Yang Z, Li L, Lin K, et al. The Dawn of Lmms: Preliminary Explorations with Gpt-4v (Ision). arXiv preprint arXiv:2309.17421. 2023.

[9] Wang Z, Liu L, Wang L, Zhou L. R2gengpt: radiology report generation with frozen llms. arXiv preprint arXiv:2309.09812. 2023.

[10] Singhal K, Tu T, Gottweis J, et al. Towards Expert-Level Medical Question Answering with Large Language Models. arXiv preprint arXiv:2305.09617. 2023.

[11] Wu C, Lei J, Zheng Q, et al. Can gpt-4v (ision) serve medical applications? case studies on gpt-4v for multimodal medical diagnosis. arXiv preprint arXiv:2310.09909. 2023.

[12] Vinyals O, Toshev A, Bengio S, Erhan D. Show and tell: a neural image caption generator. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015. IEEE Computer Society; 2015:3156–3164.

[13] Xu K, Ba J, Kiros R, et al. Show, attend and tell: neural image caption generation with visual attention. In: Bach FR, Blei DM, eds. Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015. JMLR.org; 2015:2048–2057. URL: http://proceedings.mlr.press/v37/xuc15.html.

[14] Pan Y, Yao T, Li Y, Mei T. X-linear attention networks for image captioning. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020. Computer Vision Foundation/IEEE; 2020:10968–10977.

[15] Li Y, Liang X, Hu Z, Xing EP. Hybrid retrieval-generation reinforced agent for medical image report generation. In: Bengio S, Wallach HM, Larochelle H, Grauman K, Cesa- Bianchi N, Garnett R, eds. Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018. Canada: Montreal; 2018:1537–1547. URL: https://proceedings.neurip s.cc/paper/2018/hash/e07413354875be01a996dc560274708e-Abstract.html.

[16] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: Guyon I, von Luxburg U, Bengio S, et al., eds. Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA; 2017:5998–6008. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

[17] Chen Z, Song Y, Chang T, Wan X. Generating radiology reports via memory-driven transformer. In: Webber B, Cohn T, He Y, Liu Y, eds. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020. Association for Computational Linguistics; 2020:1439–1449. https://doi.org/10.18653/v1/2020.emnlp-main.112.
[18] Zhang Y, Wang X, Xu Z, Yu Q, Yuille A, Xu D. When radiology report generation meets knowledge graph. Proceedings of the AAAI Conference on Artificial Intelligence. 2020.
[19] Liu F, Wu X, Ge S, Fan W, Zou Y. Exploring and distilling posterior and prior knowledge for radiology report generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021:13753–13762.
[20] Li M, Lin B, Chen Z, Lin H, Liang X, Chang X. Dynamic graph enhanced contrastive learning for chest x-ray report generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023:3334–3343.
[21] Huang Z, Zhang X, Zhang S. Kiut: knowledge-injected u-transformer for radiology report generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023:19809–19818.
[22] Jiang H, Misra I, Rohrbach M, Learned-Miller E, Chen X. In defense of grid features for visual question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020:10267–10276.
[23] Wu C, Liu J, Wang X, Li R. Differential networks for visual question answering. In: Proceedings of the AAAI Conference on Artificial Intelligence. 2019:8997–9004.
[24] Nguyen BD, Do TT, Nguyen BX, Do T, Tjiputra E, Tran QD. Overcoming data limitation in medical visual question answering. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2019: 522–530.
[25] Finn C, Abbeel P, Levine S. Model-agnostic meta-learning for fast adaptation of deep networks. In: International Conference on Machine Learning. PMLR; 2017: 1126–1135.
[26] Eslami S, de Melo G, Meinel C. Does Clip Benefit Visual Question Answering in the Medical Domain as Much as it Does in the General Domain? arXiv Preprint arXiv: 2112. 2021:13906.
[27] Ambati R, Dudyala CR. A sequence-to-sequence model approach for imageclef 2018 medical domain visual question answering. In: 2018 15th IEEE India Council International Conference (INDICON). IEEE; 2018:1–6.
[28] Khare Y, Bagal V, Mathew M, Devi A, Priyakumar UD, Jawahar C. Mmbert: multimodal bert pretraining for improved medical vqa. In: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI). IEEE; 2021:1033–1036.
[29] Pellegrini C, Keicher M, €Ozsoy E, Navab N. Rad-restruct: a novel vqa benchmark and method for structured radiology reporting. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2023: 409–419.
[30] Li C, Wong C, Zhang S, et al. Llava-med: training a large language- and-vision assistant for biomedicine in one day. arXiv preprint arXiv:2306.00890. 2023.
[31] Kamath A, Singh M, LeCun Y, Synnaeve G, Misra I, Carion N. Mdetr-modulated detection for end-to-end multi-modal understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021:1780–1790.
[32] Huang W, Zhou H, Li C, Yang H, Liu J, Wang S. Enhancing Representation in Radiography-Reports Foundation Model: A Granular Alignment Algorithm Using Masked Contrastive Learning. arXiv preprint arXiv:2309.05904. 2023.
[33] Sun J, Wei D, Xu Z, et al. You've got two teachers: Co-evolutionary image and report distillation for semi-supervised anatomical abnormality detection in chest x-ray. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2023:363–373.
[34] Sun Z, Lin M, Zhu Q, et al. A scoping review on multimodal deep learning in biomedical images and texts. J Biomed Inf. 2023:104482.
[35] Boecking B, Usuyama N, Bannur S, et al. Making the most of text semantics to improve biomedical vision–language processing. In: European Conference on Computer Vision. Springer; 2022:1–21.
[36] Papineni K, Roukos S, Ward T, Zhu W. Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002. Philadelphia, PA, USA: ACL; 2002: 311–318. URL: https://aclanthology.org/P02-1040/.
[37] Lin CY. ROUGE: a package for automatic evaluation of summaries. In: Text Summarization Branches Out. Barcelona, Spain: Association for Computational Linguistics; 2004:74–81. URL: https://aclanthology.org/W04-1013.
[38] Banerjee S, Lavie A. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In: Goldstein J, Lavie A, Lin C, Voss CR, eds. Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation And/or Summarization@ACL 2005. Ann Arbor, Michigan, USA: Association for Computational Linguistics; 2005:65–72. https://aclanthology.org/ W05-0909/.
[39] Vedantam R, Zitnick CL, Parikh D. Cider: consensus-based image description evaluation. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015. IEEE Computer Society; 2015:4566–4575.
[40] Bannur S, Hyland S, Liu Q, et al. Learning to exploit temporal structure for biomedical vision-language processing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023:15016–15027.
[41] Li M, Sigal L. Referring transformer: a one-step approach to multi-task visual grounding. Adv Neural Inf Process Syst. 2021;34:19652–19664.
[42] Du Y, Fu Z, Liu Q, Wang Y. Visual grounding with transformers. In: 2022 IEEE International Conference on Multimedia and Expo (ICME). IEEE; 2022:1–6.
[43] Zhu C, Zhou Y, Shen Y, et al. Seqtr: a simple yet universal network for visual grounding. In: European Conference on Computer Vision. Springer; 2022:598–615.
[44] Deng J, Yang Z, Chen T, Zhou W, Li H. Transvg: end-to-end visual grounding with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021:1769–1779.
[45] Chen Z, Zhou Y, Tran A, et al. Medical phrase grounding with region-phrase context contrastive alignment. In: International Conference on Medical Image Computing and Computer- Assisted Intervention. Springer; 2023:371–381.
[46] Yang J, Zhang H, Li F, Zou X, Li C, Gao J. Set-of-mark Prompting Unleashes Extraordinary Visual Grounding in Gpt-4v. arXiv preprint arXiv:2310.11441. 2023.