

# Creación de un Modelo de Descripciones de Imágenes Especializado en Arqueología Griega

## *Creating an Image Description Model Specialized in Greek Archaeology*

Enrique Garcia-Arias, Ana Garcia-Serrano

Universidad Nacional de Educación a Distancia (UNED)

Madrid, España

e.garcia@scc.uned.es, agarcia@lsi.uned.es

**Resumen:** La generación automatizada de descripciones de imágenes (IM, Image Captioning) ha experimentado un considerable impulso en los últimos años con la incorporación de los LLM (Large Language Models). En contextos generalistas, los resultados son bastante ajustados, sin embargo, los desafíos son significativos en dominios especializados, como es el caso del proyecto Arqueogriegos. El corpus multimodal de este trabajo está formado por fotos, planos y textos en un contexto arqueológico y se refieren a yacimientos, artefactos y su entorno histórico, un ámbito complejo para interpretar estas imágenes descontextualizadas y carentes de un texto descriptivo (caption) adecuado. El objetivo principal de este estudio es generar descripciones automáticas optimizadas que superen esta desconexión entre imágenes y textos, abordando las limitaciones de las imágenes arqueológicas aisladas. Para ello, en lugar de recurrir a soluciones directas o vía API, que han resultado insuficientes para la complejidad del problema, se ha diseñado una metodología innovadora que divide los componentes clave en fases, evaluando e implementando en cada una la solución más efectiva, constituyendo así la principal contribución del trabajo al superar las deficiencias de los modelos de IM y LLM multimodal existentes.

**Palabras clave:** Descripciones de imágenes, Corpus multimodal, Arqueología, Modelos del Lenguaje.

**Abstract:** The automated generation of image descriptions (IM, Image Captioning) has seen significant progress in recent years with the integration of LLMs (Large Language Models). In generalist contexts, the results are quite accurate; however, challenges remain substantial in specialized domains, as exemplified by the Arqueogriegos project. The multimodal corpus of this study comprises photos, plans, and texts within an archaeological context, encompassing sites, artifacts, and their historical environment—a particularly complex domain due to the difficulty of interpreting these decontextualized images, lacking an adequate descriptive text (caption). The primary objective of this study is to generate optimized automatic descriptions that address the disconnect between images and texts, tackling the limitations of isolated archaeological images. To achieve this, rather than relying on direct solutions or APIs, which have proven insufficient for the problem's complexity, an innovative methodology was designed, breaking down key components into phases and evaluating and implementing the most effective solution at each stage. This approach constitutes the main contribution of the work, overcoming the shortcomings of existing IM and multimodal LLM models.

**Keywords:** Image Caption, Multimodal Corpus, Archeology, Large Language Models.

## 1 Introducción

En los últimos años emergen los modelos de difusión con unas capacidades generativas (Rombach et al., 2022) que permiten tareas tanto de ima-

gen a texto como de texto a imagen y ofrecen soluciones más versátiles (Song et al., 2023) que las arquitecturas basadas en CNN (Convolutional Neural Network) y Transformers, más simples de

implementar pero que tienden a generar descripciones de imágenes con menor diversidad y son propensas al sobreajuste. La reciente emergencia de los modelos de lenguaje (LLM) (Alayrac et al., 2022), (Chowdhery et al., 2023), (Touvron et al., 2023) marca un cambio de paradigma en estas tecnologías, obteniendo mejores resultados en tareas multimodales en escenarios sin entrenamiento (Fei et al., 2023), (Liu et al., 2023), (Su et al., 2022) o con poco entrenamiento (Huang et al., 2023), (Tsimpoukelli et al., 2021) con la generación de descripciones de imágenes.

La incorporación de contexto específico a la generación de descripciones en las aplicaciones prácticas, como por ejemplo en *Wikipedian Captioning* (wikiCap) (Nguyen et al., 2023), permite facilitar a los autores de Wikipedia la asociación de descripciones de imágenes basadas en sus páginas. Posteriormente, el concepto *Contextualized Image Captioning* (CIC) (Burns et al., 2023) amplía este enfoque en un modelo más formal y generalista.

Los Modelos de Visión-Lenguaje (VLM) son capaces de procesar entradas tanto textuales como visuales (Alayrac et al., 2022), (Liu et al., 2023). La integración de Modelos de Lenguaje de Gran Escala (LLM) en los VLM ha incrementado notablemente su capacidad para interpretar preguntas complejas (Zou et al., 2023). Entre los desarrollos más avanzados destacan modelos como *Flamingo* (Alayrac et al., 2022), *FROMAGE* (Koh, Salakhutdinov, y Fried, 2023) e *InstructBLIP* (Liu et al., 2023) que incorporan mecanismos de atención con capacidades de instrucciones centradas en el contenido visual. De forma similar, *MiniGPT4* (Zhu et al., 2023) y *LlaVA* (Liu et al., 2023) han sido diseñados para gestionar interacciones conversacionales complejas con entradas multimodales.

Sin embargo, estos modelos enfrentan desafíos significativos en dominios especializados como la arqueología, donde las imágenes de yacimientos y artefactos requieren un conocimiento contextual específico para su interpretación. Las aplicaciones en este campo siguen siendo escasas, en gran parte debido a la complejidad técnica de los sistemas de inteligencia artificial empleados, que suelen requerir la intervención de especialistas informáticos. (Caspari y Crespo, 2019).

Para superar las limitaciones de soluciones genéricas o vía API's externas, en este trabajo se presenta una metodología innovadora que optimiza el proceso mediante un enfoque por fases.

Este artículo se estructura como sigue. En el segundo apartado se presentan trabajos relaciona-

dos en el dominio de la arqueología en Humanidades Digitales. En el apartado 3 se propone, en detalle, la metodología definida para desarrollar un modelo híbrido de creación de descripciones de imágenes asociadas a un contexto textual. En el apartado cuarto se incluyen los resultados de las evaluaciones realizadas y se finaliza con el apartado de conclusiones.

## 2 Trabajos relacionados

La aplicación de la inteligencia artificial en varias áreas de la arqueología incluye, entre otros: (1) uso de herramientas software para crear estímulos en la organización de exhibiciones, uso de robots humanoides y presentaciones holográficas como guías de interacción e implicación de los visitantes de museos; (2) análisis de los métodos de clasificación de los fragmentos en las excavaciones arqueológicas y la reconstrucción de cerámicas así como la recomposición de partes de textos faltantes en los documentos históricos y epigrafías; (3) catálogo y estudio de restos humanos para la comprensión del contexto histórico y social; (4) detección de hallazgos terrestres de dificultad particular; (5) diseño y estudio para la exploración de lugares arqueológicos marinos situados a profundidades no alcanzables por el hombre, mediante la construcción de visualizaciones 3D (Mantovan y Nanni, 2020).

Estas líneas generales se concretan en una serie de casos reales que ilustran el potencial de la IA aplicada a contextos arqueológicos específicos. Dentro de las aplicaciones para inferir posibles localizaciones de yacimientos, uno de los primeros trabajos en la utilización de CNN para la detección automática de imágenes históricas ha sido la identificación del eje qanat en el Kurdistán utilizando imágenes del satélite CORONA de la guerra fría (Soroush et al., 2020). La disponibilidad de conjuntos de datos (datasets) de imágenes satelitales, la introducción de drones y la reducción de costes de computación han impulsado estas prácticas. Otra línea de crecimiento es la denominada *sensibilidad remota* (remote sensing), definida como la práctica de extraer información del territorio utilizando imágenes y la captura de la radiación electromagnética reflejada o emitida por la superficie de la Tierra (Soroush et al., 2020), (Berganzo-Besga et al., 2021).

El proyecto ArchAIDE (Gualandi et al., 2016), (Gualandi, Gattiglia, y Anichini, 2021) es un proyecto europeo de 3 años de duración (junio 2016–mayo 2019) para el reconocimiento automatizado de cerámicas arqueológicas con técnicas innovadoras. La clasificación de cerámicas

es de importancia fundamental para la comprensión y datación de los contextos arqueológicos, las dinámicas de producción y los flujos de intercambio, para lo que es necesario disponer de conjuntos de entrenamiento relativamente grandes, como ya ocurría con los primeros trabajos para procesar imágenes con el soporte de información textual (García-Serrano et al., 2009), (Martínez-Fernández et al., 2006).

Resaltamos los trabajos en la necrópolis medieval de Santa Ana (Guzmán, Burgos) donde se utilizan técnicas no invasivas basadas en la prospección superficial pedestre intensiva y el empleo de los Sistemas de Información Geográfica (SIG) para facilitar el estudio del área sin excavación arqueológica y mejorar su comprensión, superando los obstáculos derivados de la erosión debida a las labores agrícolas realizadas en el área (Tobalina Pulido, Polo Romero, y Suárez López, 2025).

Respecto a los Modelos del Lenguaje, se destaca la evolución hacia los modelos multimodales como LLaVA (Liu et al., 2023), Flamingo (Alayrac et al., 2022) visual-GPT (Chen et al., 2022) que integran el tratamiento de visión y lenguaje, y que están mostrando avances muy positivos tanto en la precisión como en la relevancia de las descripciones. Sin embargo, cuando hemos aplicado estas tecnologías al dominio de las Humanidades Digitales (García Serrano y Menta Garuz, 2022) y concretamente en arqueología, nos hemos enfrentado a serios retos debido a la especial característica de las imágenes de yacimientos y las dificultades con las interpretaciones iconográficas o culturalmente específicas al no contar con datos de entrenamiento extensos (Mantovan y Nanni, 2020), (Caspari y Crespo, 2019).

En este escenario de limitaciones y oportunidades, la metodología de este trabajo propone un modelo híbrido utilizando OpenCLIP para extraer características visuales, Kmeans para clasificar las imágenes y un enfoque de contextualización (spaCy, RAG) con el corpus multimodal del proyecto Arqueogriegos (García-Serrano et al., 2025), con un compromiso de equilibrio entre innovación tecnológica y pertinencia cultural.

### 3 Metodología de desarrollo

Como fuente de información se parte del proyecto Arqueogriegos que incluye un corpus multimodal de miles de fotografías tomadas por un estudio filohelena de la Antigua Grecia (D. César Fernández), así como planos y textos explicativos, en los que no se referencian las fotografías. Actualmente en el portal orientado al estudio y difusión del trabajo, se incorporan 8 regiones de las

20 de la Antigua Grecia, visualizable en el nodo CLARIAH-ES (<https://arqueogriegos.iatext.ulpgc.es/>) y disponible en abierto. Las imágenes y los textos utilizados por los procesos desarrollados en esta investigación se limitan a tres regiones de la antigua Grecia: Ática, El Peloponeso y Tracia (Lo que supone información de 13 yacimientos y 7 museos en 70 ficheros Word y 486 imágenes de Ática; de 45 yacimientos y 20 museos en 216 ficheros Word y 1543 imágenes de El Peloponeso; de 3 yacimientos y 2 museos en 14 ficheros Word y 107 imágenes de Ática).

Las fases del proceso (Fase I a Fase III) se muestran en la figura 1 y se explican en lo que sigue. El código del proyecto está accesible en <https://github.com/enriquegarciaarias/Descriplmg>.

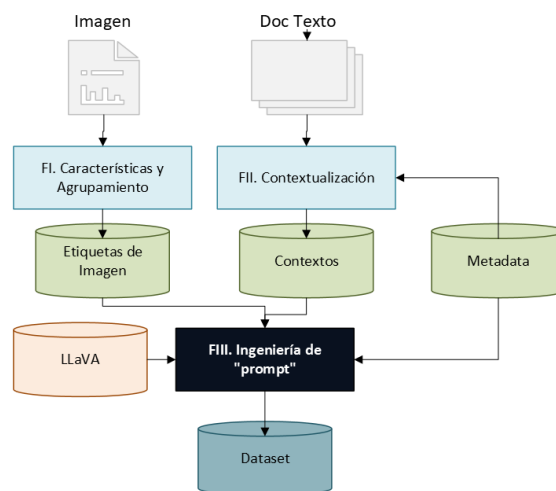


Figura 1: Estructura del proceso.

#### Fase I. Características y Agrupamiento

En esta fase se entrenan los modelos que serán utilizados posteriormente para la agrupación de nuevas imágenes. Se realizan 4 procesos: Extracción de características, Agrupación (Cluster), Etiquetado y Almacenamiento de los modelos, como se describe a continuación.

En primer lugar, se extraen las características de la imagen mediante OpenCLIP, que es la implementación de CLIP (Contrastive Language Image Pretraining) desarrollado por OpenAI, utilizando la variante ViT-L/14, que es un modelo de visión basado en Vision Transformer (ViT) y entrenado en el conjunto de datos laion2b\_s32b\_b82k que contiene 2 billones de pares de imagen y texto. OpenCLIP ViT-L/14, convierte cada imagen  $I$  en un vector de embeddings  $e_I \in \mathbb{R}^d$ , donde  $d = 768$  es la dimensionalidad del espacio de características.

La elección de CLIP se fundamenta en su capacidad de realizar tareas de clasificación de imágenes.

nes sin necesidad de entrenamiento específico (zero-shot learning). La elección de la variante ViT-L/14 se realiza en función del número de parámetros, dado que tiene una influencia fundamental en el rendimiento; los modelos con un mayor número de parámetros pueden identificar un mayor número de características.

Con el conjunto de características de las imágenes de entrenamiento, se procede a agrupar imágenes con características similares (Clustering). En concreto, una vez que las imágenes han sido convertidas en embeddings por OpenCLIP, son agrupadas por el algoritmo de agrupación (cluster) Kmeans minimizando la función objetivo:

$$J = \sum_{i=1}^n \min_{\mu_j} \|e_i - \mu_j\|^2,$$

donde  $\mu_j$  son los centroides de los  $k = 4$  clústeres. La elección de Kmeans (Arthur y Vassilvitskii, 2006) frente a otras alternativas como DBSCAN, Mean Shift o GMM, se fundamenta en su facilidad de utilización, dado que requiere sólo 2 parámetros: el número de clústeres (4 en nuestro caso) y la inicialización de los centroides. Es una buena opción cuando se precisa un equilibrio entre simplicidad y rendimiento.



**Figura 2:** Asignación de grupos “clusters” de imágenes.

En este proceso de agrupación interviene la supervisión manual para decidir el número óptimo de categorías, teniendo en cuenta que las características de las imágenes son específicas del contexto de la arqueología y se identifican las etiquetas de dichos grupos. Se seleccionan 4 categorías:

**Panorámica:** Fotografía panorámica, mostrando elementos estructurales, muchas veces arte-

factos dispersos, pero con alguna estructura, organización, disposición y estado. Es una visión general en la que se presenta el yacimiento integrado en el paisaje.

**Dibujos:** Dibujo que representa con detalle un elemento del yacimiento, como una vivienda o un plano.

**Detalles:** Fotografía de detalle, muestra el elemento principal que protagoniza la escena. Es una fotografía de exteriores, pero, a diferencia de la Panorámica, centra su atención en un elemento singular.

**Exposición:** Fotografía de exposición en interiores, mostrando un elemento principal con fondo plano.

Es preciso comprobar la estabilidad de la asignación de las imágenes a los grupos adecuados, para lo que se realizan pruebas sobre conjuntos de imágenes y se realiza una comprobación visual de la coherencia de las asignaciones a las 4 categorías. A continuación, se procede a estabilizar el modelo y se almacenan los siguientes resultados para los procesos posteriores:

**Características (Features):** Constituyen los embeddings de características generadas a partir de las imágenes con el modelo OpenCLIP

**Centroids:** Representan los puntos centrales de cada cluster en el espacio de características. Resultan del “promedio” de todos los puntos en las imágenes. Sirven para asignar nuevas características por su cercanía al punto central.

**Etiquetas de Clúster:** Son los valores numéricos que identifican a cada clúster.

**Algoritmo PCA:** que reduce la dimensionalidad de los embeddings de las características de las imágenes para mejorar la eficiencia y calidad de las agrupaciones. Deben conservarse para que sean aplicados de modo homogéneo en todos los casos.

Una vez se dispone de un modelo de agrupación generado en la fase anterior, se aplica sobre cada nuevo conjunto de imágenes. Estos son tratados por el componente de características (Features) como paso previo a poder aplicar las reglas de agrupamiento entrenadas en la fase anterior.

El uso de estos modelos entrenados garantiza que el criterio de aplicación de los grupos sea homogéneo. Tras la agrupación con las nuevas imágenes, los resultados se utilizan para retroalimentar el modelo inicial haciéndolo más sólido y eficiente. La supervisión manual puede, adicionalmente, refinar los resultados y aumentar la eficiencia.

## Fase II. Contextualización

Los modelos de generación de descripciones

a partir de las imágenes son buenos cuando las imágenes presentan objetos reconocibles de tipo generalista y que pueden asociarse correctamente con los conjuntos de datos (datasets) disponibles, es el caso de edificios, animales, personas, etc. Sin embargo, las imágenes correspondientes a yacimientos arqueológicos como las mostradas en la imagen 3 pueden no presentar elementos claramente identificables. Las pruebas realizadas generan descripciones muy vagas del tipo “La imagen muestra un yacimiento de rocas y árboles en un bosque. Las rocas están dispuestas en un montón y se encuentran rodeadas de árboles.” Este tipo de descripción es insuficiente y no puede utilizarse.



El Tétaro

El teatro

**Figura 3:** Imágenes de difícil interpretación automatizada.

Para resolver esta carencia se desarrolla un componente de enriquecimiento contextual que aporta elementos relevantes que puedan asociarse a las características identificadas de forma automatizada y, de este modo, pueda mejorar el posterior proceso de generación de indicaciones (prompt) correspondiente a la siguiente fase. El componente de enriquecimiento contextual cuenta con las siguientes fuentes de información:

*El título de la imagen:* Describe la imagen.

*La zona y el yacimiento:* Identifican la localización; el conjunto de localizaciones incluidas incluye los yacimientos de las zonas de Ática, Peloponeso y Tracia.

*El documento de texto extenso:* Un documento en formato Word, escrito por el filoheleno en el que se describe el contexto de un conjunto de imágenes que forman parte de una determinada zona de yacimientos, aunque no se referencia en el texto. La asociación contextual puede realizarse cuando en el contenido de este documento se mencionan características que puedan asociarse a la imagen.

El nombre de la imagen aparece con el formato “Diapo 13.36 Templete del recinto de Diogiton.jpg” o “Diapo 13.11 La Casa Oeste.JPG” o incluso “Diapo 13.30.JPG”. El proceso debe extraer el texto relevante para poder utilizarlo en la

búsqueda de contexto disponible en el texto del documento extenso. Para ello, realizamos primero una identificación de entidades, es decir, “Diogiton”, y el sujeto de la frase, esto es, “Casa Oeste” para componer las palabras clave que van a ser utilizadas en la búsqueda de contexto. Esta búsqueda de contexto se realiza sobre el texto, para lo que se procede a su segmentación utilizando un modelo de Procesamiento de Lenguaje Natural con la librería spaCy. A continuación, se procede a generar los embeddings de las palabras clave y los párrafos del documento para poder calcular la similitud del coseno entre el embedding medio de las palabras clave y los párrafos de texto (Lastra-Díaz et al., 2021). Como resultado de este proceso, se seleccionan los 4 párrafos con mayor relevancia para componer el contexto con el que enriquecer las descripciones de la imagen.

La presencia de palabras clave y de su significado contextual mejora la precisión al seleccionar los párrafos relevantes de contexto para la imagen. A continuación, se muestra un ejemplo.

**Imagen:** Diapo 13.21 El templo de Némesis.

**Palabras Clave:** El templo de Némesis, Némesis, Ática, Ramnous.

**Contexto:** *Némesis, además, y debido a sus particularidades ctónicas, estaba relacionada en el Ática con los muertos, y en su celebración anual, las Nemeas, tenían lugar combates de antorchas entre los adolescentes que hacían el servicio militar en la fortaleza. Más adelante nos encontramos con el santuario de Némesis (9). La Némesis que era adorada en Ramnous era una diosa de la agricultura. La estatua de Némesis de mármol de Paros, obra de Agorákrito de Paros, un discípulo de Fidias, y, frente a ella, la mesa de ofrendas. Némesis sacudió sus plumas resignadamente y fue a Esparta, donde Leda, esposa del rey Tindáreo, encontró poco después en un pantano un huevo de color de jacinto que llevó a su casa y ocultó en un cofre; de ese huevo salió Helena de Troya.*

En aquellos casos en que no se identifica el nombre de la imagen, como ocurre en “Diapo 13.30.JPG”, o que el nombre no está referenciado en el texto extenso como en el caso de “Recinto funerario de Diofantos”, se ha optado por no incluir contexto. En la tabla 1 se resumen los algoritmos y parámetros utilizados en esta fase.

### Fase III. Ingeniería de prompts

Se completa el proceso de producción de las descripciones de las imágenes con la utilización de las capacidades de los Modelos de Visión-Lenguaje (VLM) y concretamente el modelo LLaVA “liuhaotian/llava-v1.5-7b”. La elección de este

Algoritmo	Función	Parámetros
OpenCLIP ViT-L/14	Embeddings	$d = 768$ , laion-2b
KMeans	Agrupamiento	$k = 4$ , k-means++
spaCy	Segmentación de texto	es_core_news_lg

**Tabla 1:** Resumen de algoritmos y funciones.

modelo se basa en las limitaciones computacionales de la configuración del equipo de desarrollo con NVIDIA GeForce RTX 4060 con 8 GB de VRAM; se trata de un modelo de alta capacidad y que, con una adecuada optimización, ha podido ejecutarse en el hardware especificado mediante técnicas como la cuantización (4-bit quantization) y offloading en RAM. Por lo tanto, se ha optado por un compromiso entre precisión y recursos, en comparación con modelos más grandes, como GPT-4V, LLaVA-7B ofrece un equilibrio adecuado entre rendimiento y prestaciones, evitando la necesidad de infraestructura especializada.

En nuestro caso, se necesita partir de la identificación de características relevantes de la imagen en el Prompt 1 para después, en la segunda parte de la cadena con el Prompt 2, retroalimentar al modelo con la respuesta inicial forzándole a realizar un razonamiento estructurado, evitando la generación de respuestas de forma libre y perfeccionando la coherencia en cada paso.

En la contribución del conocimiento externo que denominamos contexto y categorías de agrupación de imágenes, aplicamos el modelo Recuperación Generación Aumentada (RAG) (Lewis et al., 2020). El contexto de las categorías se aplica al Prompt 1, obtenido del texto extenso (documento Word) que acompaña al conjunto de imágenes. La incorporación del contexto textual se produce en el Prompt 2, enriqueciendo la respuesta del modelo al Prompt 1. El objetivo es limitar las descripciones a los contextos establecidos, mejorando la precisión y evitando contenido “inventado”.

Con la técnica **RAILS** (Reinforced Artificial Intelligence Language Structures) se guía y estructuran las respuestas con reglas o plantillas, de modo que se cumplan ciertos criterios de estandarización o patrones de respuesta, generando descripciones con mayor coherencia. Esta implementación se realiza en dos fases (Topic y Fact-Checking RAILS):

**Topic RAILS:** Estructurando guías temáticas para asegurar el foco en temas específicos, las respuestas del LLM deberán incluir de forma ho-

mogénea los elementos de descripción del elemento principal que protagoniza la escena, la descripción de la disposición y organización espacial de la imagen, el entorno o lugar donde se encuentra el yacimiento y el contexto arqueológico. Esta normalización se aplica según los criterios de agrupación realizados: Panorámica, Dibujos, Detalles y Exposición, como se indica a continuación.

### Panorámica

- (C1) Para esta fotografía panorámica.
- (C2)\*\*Ubicación y entorno\*\*:  
Describe el paisaje y el tipo de terreno.
- (C3)\*\*Elemento principal yacimiento arqueológico\*\*:  
Explica la estructura y su disposición.
- (C4) Ten en cuenta que es la vista general y amplia de un yacimiento arqueológico no son simples piedras.

### Dibujos

- (C1) Para este dibujo que muestra con detalle un elemento del yacimiento arqueológico.
- (C2)\*\*Composición y contorno\*\*:  
Describe su estructura, composición, apariencia.
- (C3)\*\*Elemento principal\*\*:  
Explica qué representa, su función.
- (C4) Ten en cuenta que es la representación de una estructura arqueológica singular y que puede estar incompleta.

### Detalles

- (C1) Para esta fotografía de detalle.
- (C2)\*\*Ubicación y entorno\*\*:  
Describe el entorno y cómo se ubica el elemento principal.
- (C3)\*\*Elemento principal que protagoniza la imagen\*\*:  
Explica la estructura y qué representa.
- (C4) Ten en cuenta que es un elemento principal de un yacimiento arqueológico no son simples piedras.

### Exposición

- (C1) Para esta fotografía de exposición.
- (C2)\*\*Composición\*\*:  
Describe su composición y contorno.
- (C3)\*\*Elemento principal\*\*:  
Explica sus características y qué representa.
- (C4) Ten en cuenta que es un objeto arqueológico de valor singular.

**Fact-Checking RAILS:** Para reducir inexactitudes guiando al LLM hacia respuestas basadas en evidencia y penalizando afirmaciones especulativas, se exige no usar lenguaje subjetivo como

“parece ser”, “probablemente” y evitar al máximo las inferencias sin respaldo verificable. Un elemento adicional que no se utiliza en esta implementación, es contrastar la respuesta con una fuente de información externa verificada.

El primer Prompt de la secuencia es:

“{C1} representando a '{titulo}', ten en cuenta sin mencionar explícitamente que {C4} y realiza una descripción en castellano, cññete a estos dos items: Item 1 {C2}. Item 2 {C3}. Máximo 20 palabras, no mencionar deteriorado y/o antiguo.” Donde {Cn} corresponde a la personalización resultado de la categorización de la imagen realizada en el proceso de clúster descrito anteriormente.

El segundo Prompt es:

Partiendo de la descripción inicial: {descripInicial}, adopta un perfil de arqueólogo y mejórala evitando subjetividades y suposiciones, no menciones evidencias para un arqueólogo como que es antiguo o deteriorado.

Donde “descripInicial” es la respuesta generada por el LLM al primer prompt y “contexto” es el contexto extraído del documento de texto largo en formato Word facilitado por el arqueólogo, en función de las menciones a la imagen vía las palabras clave extraídas en la Fase de Contextualización. A este texto se añade, en función de la disposición o no de contexto:

**Con contexto:** “y enriquece con este contexto si puedes relacionarlo con la imagen: {contexto}. Utiliza un máximo de 50 palabras.”

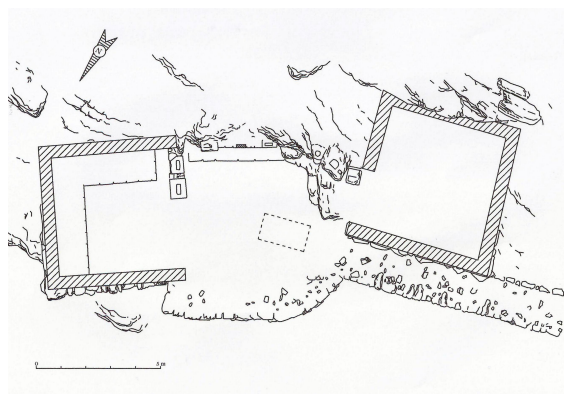
**Sin contexto:** “Utiliza un máximo de 20 palabras”

Con este proceso se elaboran los dos prompts mencionados. A continuación se reproducen los prompts y su resultado en 3 imágenes ejemplo.

#### La Planta del Amfiareio. Fig.4:

**Prompt1:** Para este dibujo que muestra con detalle un elemento o una estructura representando a “Planta del Amfiareio”, Ten en cuenta sin mencionar explícitamente que es la representación de una estructura arqueológica singular y que puede estar incompleta y realiza una descripción en castellano cññete a estos dos items: Item 1 **\*\*Composición y contorno\*\***: Describe su estructura, composición, apariencia. Item 2 **\*\*Elemento principal\*\***: Explica qué representa, su función. Máximo 20 palabras, no mencionar deteriorado y/o antiguo.

**Prompt2:** Partiendo de la descripción inicial: “La planta del amfiareio es una estructura compuesta por bloques de piedra unidos con mortero. Tiene un contorno irregular y se asemeja a un castillo. La planta del amfiareio representa una estructura arqueológica singular que podría estar



**Figura 4:** Planta del Amfiareio, yacimiento de RAMNOUS en zona de ÁTICA. El Amfiareio es una estructura arqueológica compuesta por bloques de piedra unidos con mortero, que podría estar incompleta. Se asemeja a un castillo y podría haber sido utilizado como hospital para los heridos de la fortaleza. En un principio, era un santuario medicinal y ctónico del héroe médico Aristómaco, como se menciona en las epigrafías y fuentes filológicas. El conjunto escultórico, uno de los más bellos del Ática, se encuentra en el Museo Arqueológico Nacional de Atenas.

incompleta. No se menciona que está deteriorada o antigua.”, adopta un perfil de arqueólogo y mejórala evitando subjetividades y suposiciones no menciones evidencias para un arqueólogo como que es antiguo o deteriorado y enriquece con este contexto si puedes relacionarlo con la imagen: “Casas parcelas construcciones agrícolas pozos vallas y almacenes daban otro aspecto al lugar Quizás el Amfiareio se utilizase como hospital para los heridos de la fortaleza El Anfiareio era en un principio un santuario medicinal y ctónico del héroe médico Aristómaco como sabemos por las epigrafías y las fuentes filológicas con la cabeza de Apolo saliendo de su útero Este conjunto escultórico uno de los más bellos del Ática se encuentra en el Museo Arqueológico Nacional de Atenas.”. Utiliza un máximo de 50 palabras y construye un texto enlazado.

#### Templete del recinto de Diogiton. Fig.5

**Prompt1:** Para esta fotografía de exposición representando a “Templete del recinto de Diogiton”, Ten en cuenta sin mencionar explícitamente que es un objeto arqueológico de valor singular y realiza una descripción en castellano cññete a estos dos items: Item 1 **\*\*Composición\*\***: Describe su composición y contorno. Item 2 **\*\*Elemento principal\*\***: Explica sus características y qué representa. Máximo 20 palabras, no mencionar deteriorado y/o antiguo.

**Prompt2:** (SIN CONTEXTO) Partiendo de la descripción inicial: “La escultura está hecha de



**Figura 5:** Templete del recinto de Diogiton, yacimiento de RAMNOUS en zona de ÁTICA. La escultura representa a un hombre y una niña, con la niña sosteniendo un libro. La escultura está hecha de piedra y tiene un contorno elegante.

pedra y tiene un contorno elegante. La escultura representa a un hombre y una niña, con la niña sosteniendo un libro.”, adopta un perfil de arqueólogo y mejórala evitando subjetividades y suposiciones no mencionadas evidencias para un arqueólogo como que es antiguo o deteriorado. Utiliza un máximo de 20 palabras y construye un texto enlazado.

#### **El Anfiareio. Fig.6**

*Prompt1:* Para esta fotografía que enfoca un detalle representando a 'El Anfiareio', Ten en cuenta sin mencionar explícitamente que es un elemento principal de un yacimiento arqueológico no son sólo piedras y realiza una descripción en castellano cñete a estos dos items: Item 1 **\*\*Ubicación y entorno\*\***: Describe el entorno y cómo se ubica el elemento principal. Item 2 **\*\*Elemento principal que protagoniza la imagen\*\***: Explica la estructura y qué representa. Máximo 20 palabras, no mencionar deteriorado y/o antiguo.

*Prompt2:* Partiendo de la descripción inicial: “El elemento principal se encuentra en medio de un bosque, rodeado de árboles y rocas. El elemento principal es una estructura subterránea, que podría ser parte de un yacimiento arqueológico.”, adopta un perfil de arqueólogo y mejórala evitan-



**Figura 6:** El Anfiareio, yacimiento de RAMNOUS en zona de ÁTICA. El yacimiento arqueológico se encuentra en un terreno rocoso, rodeado de piedras y vegetación. El elemento principal representa a 'El Anfiareio', un personaje de la mitología griega, que era en un principio un santuario medicinal y ctónico del héroe médico Aristómaco, como se menciona en las epigrafías y fuentes filológicas. El pequeño santuario de Anfiarao se encuentra al SO de la puerta principal de entrada a la fortaleza encima de las rocas de la colina.

do subjetividades y suposiciones no mencionadas evidencias para un arqueólogo como que es antiguo o deteriorado y enriquece con este contexto si puedes relacionarlo con la imagen: “El Anfiareio era en un principio un santuario medicinal y ctónico del héroe médico Aristómaco como sabemos por las epigrafías y las fuentes filológicas El pequeño santuario de Anfiarao se encuentra al SO de la puerta principal de entrada a la fortaleza encima de las rocas de la colina.”. Utiliza un máximo de 50 palabras y construye un texto enlazado.

Tras el proceso de construcción de las descripciones, se genera un dataset que incluye: la ruta a la ubicación física de la imagen; la etiqueta de clúster asignada; el título de la imagen; metadatos (yacimiento, región); palabras clave; contexto asociado a la imagen; Prompt elaborado [1 y 2] y la descripción generada.

## **4 Resultados y validación**

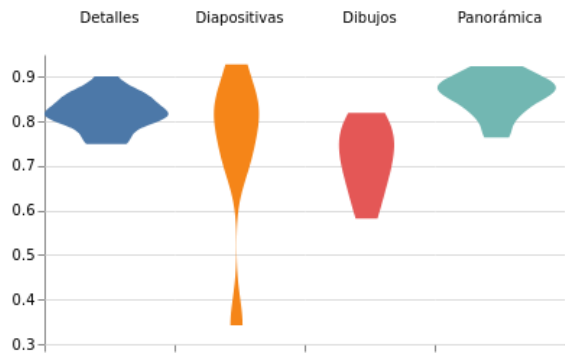
Se analiza la consistencia y relevancia del contexto utilizado, la precisión en la asignación de los “cluster”, la comparación con otros modelos generativos, y la eficiencia en términos de tiempo de respuesta.

### **4.1 Evaluación del contexto**

Se mide la calidad de las descripciones generadas al incorporar el contexto, con tres métricas: relevancia semántica, diversidad léxica y relevancia

léxica, comparando descripciones generadas con y sin contexto.

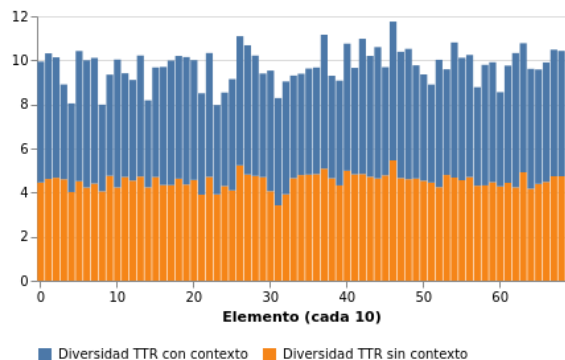
Para la relevancia semántica, utilizamos modelos de embeddings (Sentence-BERT) con los resultados mostrados en la Figura 7.



**Figura 7:** Relevancia semántica de las descripciones.

Este enfoque es clave ya que evalúa el significado de las descripciones en lugar de basarse solo en la coincidencia léxica. Constituye un indicador que mide en qué grado las descripciones reflejan el significado del contexto. Se evalúa la mejora que se produce al incluir el contexto en las indicaciones “prompts”.

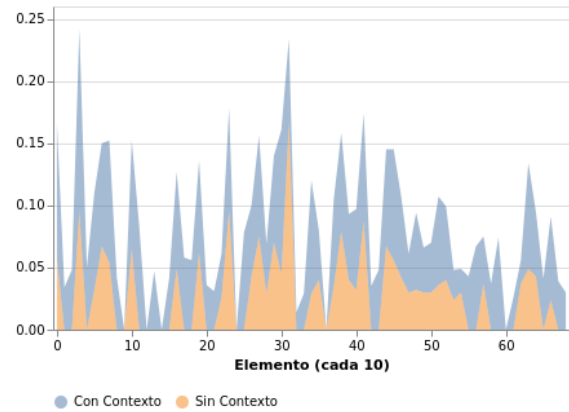
La diversidad léxica Figura 8, mide la variedad de palabras únicas (formas léxicas diferentes) presentes en las descripciones utilizando el índice TTR (Type-Token Ratio) que divide el número de palabras únicas entre el total de palabras de la descripción generada. Al incorporar el contexto, el modelo no sólo tiene acceso a la información visual de la imagen, sino que lo enriquecemos con términos y conceptos específicos de nuestra área de especialización, enriqueciendo, por tanto, el vocabulario.



**Figura 8:** Diversidad léxica.

La relevancia léxica, Figura 9 indica el grado de alineamiento de las descripciones con el contexto de referencia, no sólo en el sentido de número de palabras sino de su pertinencia. Este

valor implica la coincidencia de términos comparando cuántas palabras o sinónimos del contexto están presentes en las descripciones, así como la pertinencia semántica verificando si los términos usados reflejan el significado del contexto.



**Figura 9:** Relevancia léxica.

## 4.2 Precisión del clúster

Se evalúa la precisión del agrupamiento de imágenes en distintas categorías (Panorámica, Detalles, Dibujos, Exposición) comparándolos con una clasificación manual. Los niveles muestran un buen rendimiento que puede incrementarse con la retroalimentación del modelo al incorporar el proceso de nuevas imágenes y optimizar los “centroides” con mayor variedad de imágenes.

Las mayores convergencias se identifican en la agrupación de Panorámica y Detalles, dado que ambas se refieren a imágenes de exteriores donde Panorámica identifica un paisaje amplio con un yacimiento extenso o disperso, y Detalles muestra con claridad un elemento principal protagonista del yacimiento.

## 4.3 Comparación de las descripciones generadas

Se analizan las descripciones generadas por LLaVA en comparación con GPT-4 y Fuyu-8B respecto a la relevancia semántica y riqueza léxica, utilizando métricas como Sentence-BERT y TTR.

Para utilizar la comparación con GPT-4, facilitamos la imagen, su título y el documento de texto extenso de contexto. Construimos las siguientes indicaciones similares a las desarrolladas en nuestra investigación:

Indicaciones (prompt) GPT-4: “*Para esta imagen de título <título de la imagen > vamos a proceder por partes, primero extrae las características y define la imagen, luego con esta descripción inicial realizas mejoras asociando elementos del contexto (documento Word) y finalmente realizas*

*mejoras dando linealidad a la respuesta y evitando una respuesta especulativa. Recuerda que solo puedes utilizar la información facilitada, nada de fuentes externas. Máximo 50 palabras”.*

Los resultados del modelo desarrollado son mejores que los de GPT-4 tanto en Relevancia Semántica como en Riqueza Léxica cuando disponemos de contexto porque encontramos referencias a la imagen en el texto largo.

#### 4.4 Discusión de resultados

El modelo propuesto ofrece mayor calidad de las descripciones, lo cual es especialmente notable en los clústeres de elevada ambigüedad visual (Panorámica, Detalles). Con la incorporación del contexto no sólo se mejora la relevancia semántica y léxica, sino que también se enriquece el vocabulario con lo que las descripciones resultan de mayor utilidad para investigadores y el público en general. La precisión en los clústeres Dibujos y Exposición tiene margen de mejora ampliando los conjuntos de entrenamiento con una mayor diversidad de imágenes. Estos resultados destacan la relevancia del enriquecimiento contextual en las aplicaciones de humanidades digitales y abren nuevas posibilidades en el área de la generación automatizada de descripciones de imágenes en el contexto de la arqueología.

### 5 Conclusiones y trabajos futuros

Se ha mostrado que las técnicas de generación de descripciones son eficientes en entornos generalistas, pero insuficientes en un contexto específico como el de las imágenes de yacimientos arqueológicos a las que nos enfrentamos en este estudio. Tras la categorización de grupos “clúster” de imágenes se ha evidenciado que aquellas que corresponden a “Panorámica” y “Detalles”, que en su mayoría muestran artefactos dispersos del yacimiento o restos individuales, reciben una descripción muy vaga y ambigua. En el caso de “Dibujos” y, sobre todo, en imágenes “Exposición” en las que se exhibe un artefacto muy detallado, los métodos generalistas generan descripciones más relevantes, pero sin alcanzar los objetivos de este estudio. Por lo tanto, debía realizarse un tratamiento más especializado.

En la investigación se ha incorporado la información de características de la imagen, información de metadato (nombre del yacimiento, zona del yacimiento, título de la imagen y entidades extraídas del título) y el contexto arqueológico extraído del documento de texto largo provisto por el estudioso filoheleno para que, con la ingeniería de indicaciones (prompt) puedan generarse

respuestas más relevantes y mejor adaptadas a nuestro contexto especializado que los modelos generales.

Para la evaluación de este resultado hemos aplicado las métricas descritas, evidenciando que con el incremento en especialización arqueológica de nuestro modelo de Prompts, los resultados resultan en una mejora tanto en riqueza léxica como en relevancia semántica. Se ha resaltado que tanto la utilización de las técnicas de visión por computador con la extracción de las características (features) de las imágenes, la agrupación (clúster) y la asociación de metadatos, así como la adecuada asociación de contexto han permitido generar descripciones más relevantes.

Un aspecto que ha sido crucial es el diseño de un enfoque metodológico global. Ha sido clave probar las diferentes soluciones y poder realizar las adaptaciones que, debido a la especialización requerida tanto en el dominio como en la capacidad de computación, han sido abordables con la arquitectura presentada. Los resultados obtenidos sugieren que este enfoque puede ser útil en otros ámbitos de las Humanidades Digitales donde la descripción automática de imágenes sea relevante.

El rendimiento del modelo es un factor clave, por lo tanto, es preciso avanzar en el conocimiento no solo del ajuste fino de parámetros sino también de los procesos internos del modelo y avanzar para disponer de mayores prestaciones con menores recursos. Queda pendiente realizar una evaluación con expertos ante la dificultad de evaluación sistemática de las descripciones.

#### Agradecimientos

Los autores desean agradecer especialmente la obra del estudioso filoheleno D. César Fernández y mostrar su admiración por su “Catálogo Universal de la Antigua Grecia” disponible en el repositorio e-CienciaDatos del Consorcio Madroño (Fernández y García-Serrano, 2025). Además, este trabajo se ha financiado parcialmente por el el Ministerio de Ciencia e Innovación/AEI en el marco del proyecto GRESEL-UNED (PID2023-151280OB-C22).

#### Bibliografía

- Alayrac, J.-B., J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, ..., y K. Simonyan. 2022. Flamingo: a Visual Language Model for Few-Shot Learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, Diciembre.
- Arthur, D. y S. Vassilvitskii. 2006. k-

- means++: The advantages of careful seeding. *ilpubs.stanford.edu*.
- Berganzo-Besga, I., H. A. Orengo, F. Lumbresas, M. Carrero-Pazos, J. Fonte, y B. Vilas-Estévez. 2021. Hybrid MSRM-based deep learning and multitemporal sentinel 2-based machine learning algorithm detects near 10k archaeological tumuli in north-western iberia. *Remote Sensing*, 13(20).
- Burns, A., K. Srinivasan, J. Ainslie, G. Brown, B. A. Plummer, K. Saenko, J. Ni, y M. Guo. 2023. A Suite of Generative Tasks for Multi-Level Multimodal Webpage Understanding, Octubre. arXiv:2305.03668 [cs].
- Caspari, G. y P. Crespo. 2019. Convolutional neural networks for archaeological site detection – Finding “princely” tombs. *Journal of Archaeological Science*, 110:104998, Octubre.
- Chen, J., H. Guo, K. Yi, B. Li, y M. Elhoseiny. 2022. Visualgpt: Data-efficient adaptation of pretrained language models for image captioning. En *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, páginas 18030–18040, June.
- Chowdhery, A., S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, ..., y N. Fiedel. 2023. PaLM: Scaling Language Modeling with Pathways. *Journal of Machine Learning Research*, 24(240):1–113.
- Fei, J., T. Wang, J. Zhang, Z. He, C. Wang, y F. Zheng. 2023. Transferable Decoding with Visual Entities for Zero-Shot Image Captioning. En *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, páginas 3136–3146.
- Fernández, C. y A. Garcia-Serrano. 2025. Arqueogriegos - catálogo universal de museos y yacimientos arqueológicos de la antigua grecia. Informe técnico, repositorio en abierto; e-cienciaDatos del Consorcio Madroño.
- García-Serrano, A., X. Benavent, R. Granados, y J. M. Goñi-Menoyo. 2009. Some results using different approaches to merge visual and text-based features in CLEF’08 photo collection. En C. Peters T. Deselaers N. Ferro J. Gonzalo G. J. F. Jones M. Kurimo T. Mandl A. Peñas, y V. Petras, editores, *Evaluating Systems for Multilingual and Multimodal Information Access*, páginas 568–571, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Garcia Serrano, A. y A. Menta Garuz. 2022. La inteligencia artificial en las humanidades digitales: dos experiencias con corpus digitales. *Revista de Humanidades Digitales*, 7:19–39, jun.
- García-Serrano, A., F. Chuquimarca, F. Paños Merino, y C. Fernández, 2025. *Diseño del acceso a un catálogo sobre la Antigua Grecia*, capítulo 3. En prensa, ISBN (versión en papel): 978-84-9749-895-1.
- Gualandi, M. L., G. Gattiglia, y F. Anichini. 2021. An Open System for Collection and Automatic Recognition of Pottery through Neural Network Algorithms. *Heritage*, 4(1):140–159, Enero.
- Gualandi, M. L., R. Scopigno, L. Wolf, J. Richards, J. B. I. Garrigos, M. Heinzelmann, M. A. Hervas, L. Vila, y M. Zallocco. 2016. ArchAIDE - Archaeological Automatic Interpretation and Documentation of cEramics. Artwork Size: 4 pages ISBN: 9783038680116 ISSN: 2312-6124 Pages: 4 pages Publication Title: Eurographics Workshop on Graphics and Cultural Heritage.
- Huang, S., L. Dong, W. Wang, Y. Hao, S. Singhal, S. Ma, T. Lv, L. Cui, O. K. Mohammed, B. Patra, Q. Liu, K. Aggarwal, Z. Chi, N. Bjorck, V. Chaudhary, S. Som, X. Song, y F. Wei. 2023. Language Is Not All You Need: Aligning Perception with Language Models. *Advances in Neural Information Processing Systems*, 36:72096–72109, Diciembre.
- Koh, J. Y., R. Salakhutdinov, y D. Fried. 2023. Grounding Language Models to Images for Multimodal Inputs and Outputs. En *Proceedings of the 40th International Conference on Machine Learning*, páginas 17283–17300. PMLR, Julio. ISSN: 2640-3498.
- Lastra-Díaz, J. J., J. Goikoetxea, M. A. Hadj Taieb, A. Garcia-Serrano, M. Ben Aouicha, E. Agirre, y D. Sánchez. 2021. A large reproducible benchmark of ontology-based methods and word embeddings for word similarity. *Information Systems*, 96:101636.
- Lewis, P., E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. T. Yih, T. Rocktäschel, S. Riedel, y D. Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. En *Advances in Neural Information Processing Systems*, volumen 33, páginas 9459–9474. Curran Associates, Inc.

- Liu, H., C. Li, Q. Wu, y Y. J. Lee. 2023. Visual Instruction Tuning. *Advances in Neural Information Processing Systems*, 36:34892–34916, Diciembre.
- Mantovan, L. y L. Nanni. 2020. The Computerization of Archaeology: Survey on Artificial Intelligence Techniques. *SN Computer Science*, 1(5):267, Septiembre.
- Martínez-Fernández, J. L., J. V. Román, A. M. García-Serrano, y J. C. González-Cristóbal. 2006. Combining textual and visual features for image retrieval. En C. Peters F. C. Gey J. Gonzalo H. Müller G. J. F. Jones M. Kluck B. Magnini, y M. de Rijke, editores, *Accessing Multilingual Information Repositories*, páginas 680–691, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Nguyen, K., A. F. Biten, A. Mafla, L. Gomez, y D. Karatzas. 2023. Show, Interpret and Tell: Entity-Aware Contextualised Image Captioning in Wikipedia. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(2):1940–1948, Junio. Number: 2.
- Rombach, R., A. Blattmann, D. Lorenz, P. Esser, y B. Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. En *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, páginas 10674–10685, New Orleans, LA, USA, Junio. IEEE.
- Song, Y., Z. Zhang, Z. Lin, S. Cohen, B. Price, J. Zhang, S. Y. Kim, y D. Aliaga. 2023. ObjectStitch: Object Compositing with Diffusion Model. En *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, páginas 18310–18319, Vancouver, BC, Canada, Junio. IEEE.
- Soroush, M., A. Mehrtash, E. Khazraee, y J. A. Ur. 2020. Deep Learning in Archaeological Remote Sensing: Automated Qanat Detection in the Kurdistan Region of Iraq. *Remote Sensing*, 12(3):500, Febrero.
- Su, Y., T. Lan, Y. Liu, F. Liu, D. Yogatama, Y. Wang, L. Kong, y N. Collier. 2022. Language Models Can See: Plugging Visual Controls in Text Generation, Mayo. arXiv:2205.02655 [cs].
- Tobalina Pulido, L., L. A. Polo Romero, y P. A. Suárez López. 2025. La necrópolis de Santa Ana (Guzmán, Burgos). Localización y caracterización mediante prospección superficial. *Arqueología y Territorio Medieval*, 32:e9051, Febrero.
- Touvron, H., L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, ..., y T. Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models, Julio. arXiv:2307.09288 [cs].
- Tsimpoukelli, M., J. L. Menick, S. Cabi, S. M. A. Eslami, O. Vinyals, y F. Hill. 2021. Multimodal Few-Shot Learning with Frozen Language Models. En *Advances in Neural Information Processing Systems*, volumen 34, páginas 200–212. Curran Associates, Inc.
- Zhu, D., J. Chen, X. Shen, X. Li, y M. Elhoseiny. 2023. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models, Octubre. arXiv:2304.10592 [cs].
- Zou, A., Z. Wang, N. Carlini, M. Nasr, J. Z. Kolter, y M. Fredrikson. 2023. Universal and Transferable Adversarial Attacks on Aligned Language Models. Version Number: 2.