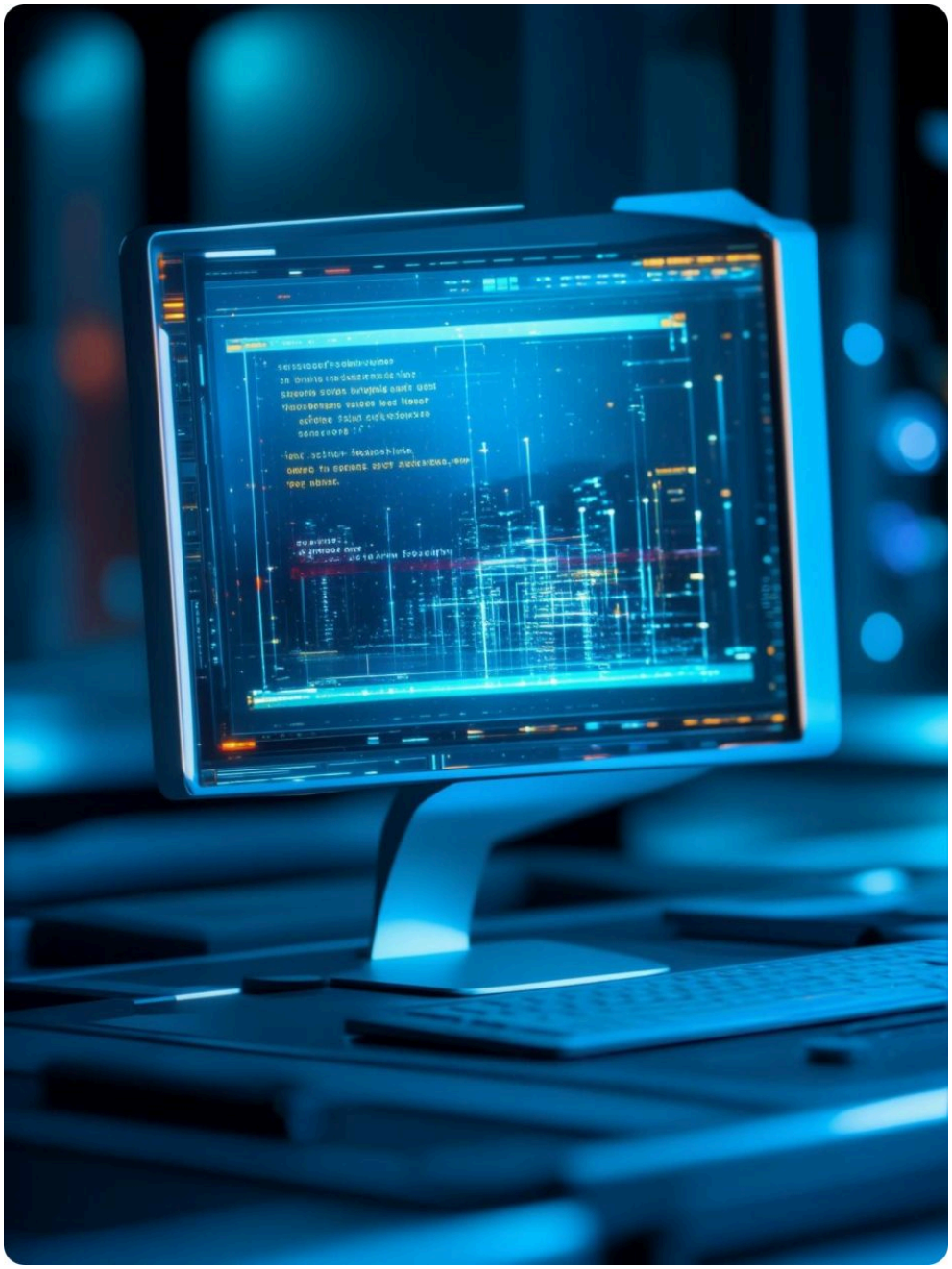
Unir visión y lenguaje

Los modelos multimodales relacionan píxeles, palabras e instrucciones.



El tipo de salida distingue comparar, describir, localizar y razonar.

El tema recorre CLIP, BLIP, Grounding DINO y un VLM como Qwen2.5-VL. Los cuatro aceptan información visual y lingüística, pero resuelven tareas distintas.

Una representación reutilizable

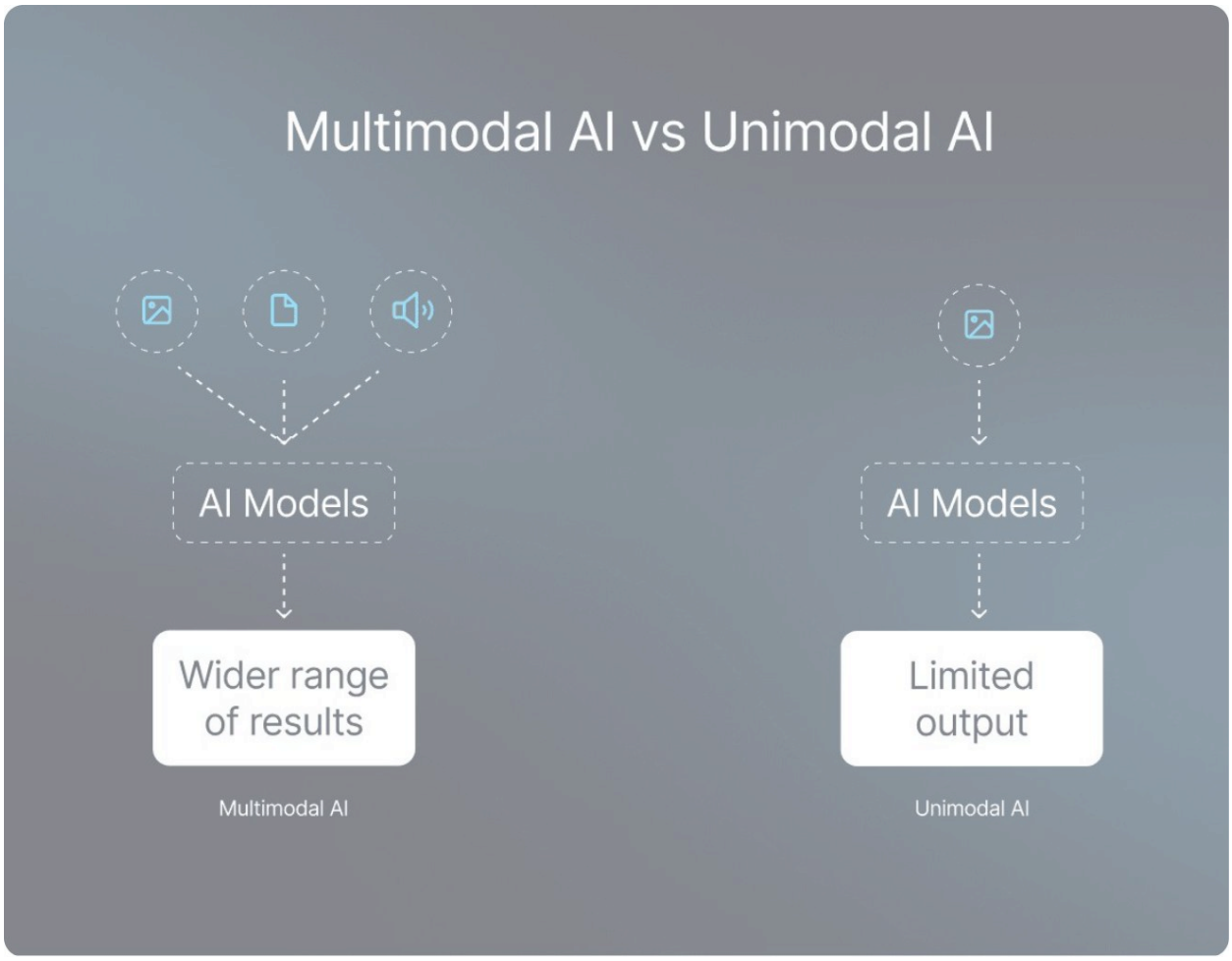
El preentrenamiento masivo produce características transferibles a tareas nuevas.

General	Transferible	Evaluable
Aprende patrones que sirven en varios problemas.	Se adapta con prompts, ejemplos o ajuste.	Su utilidad depende del dominio y la tarea.

CLIP aprende con pares imagen-texto; DINO con imágenes sin etiquetas; SAM con un gran motor de máscaras. La escala ayuda a generalizar, pero no garantiza rendimiento en una aplicación concreta.

Dos dominios, un espacio común

Imagen y texto se convierten en representaciones que el modelo puede relacionar.

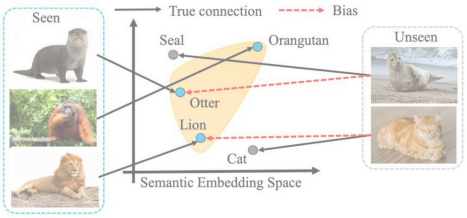


La unión puede servir para medir similitud o alimentar un generador de lenguaje.

CLIP compara representaciones; BLIP genera texto condicionado por imagen; Grounding DINO alinea expresiones con regiones; un VLM inserta tokens visuales en el contexto de un modelo de lenguaje.

Nuevas categorías escritas

En zero-shot, las clases llegan como texto y no como una cabeza reentrenada.



Quieres separar fotos de interior y exterior sin entrenar un clasificador nuevo.

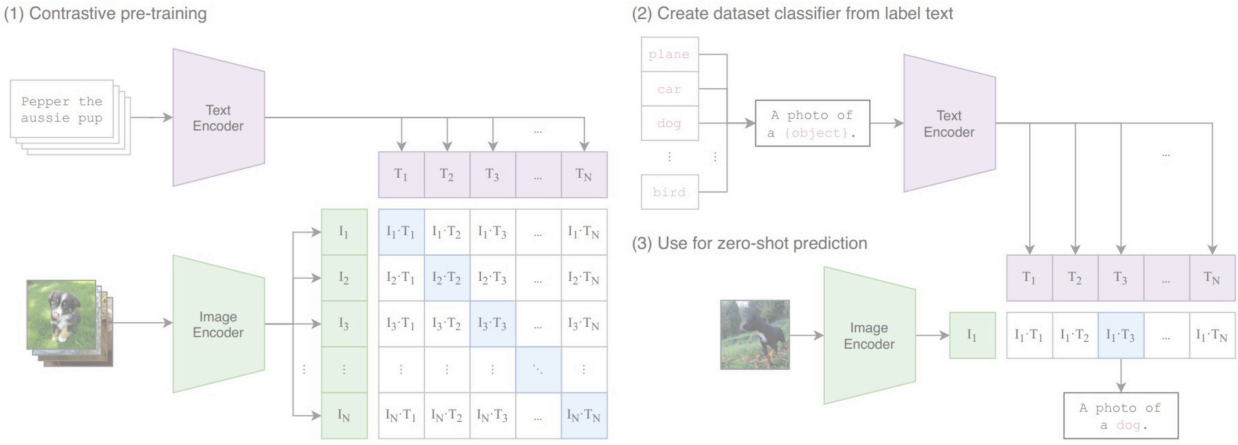
- A Comparar la imagen con ambas descripciones de texto.
- B Cambiar la resolución de la foto.
- C Usar una caja fija.

Respuesta A. Exacto: CLIP ordena categorías propuestas como texto.

La composicionalidad permite combinar atributos como “rojo”, “perro” y “corriendo”. Esto acelera la experimentación, aunque las categorías propuestas y la formulación de los textos pueden cambiar el resultado.

Comparar imagen y texto

CLIP acerca pares compatibles en un espacio de embeddings compartido.



Dos codificadores producen vectores comparables mediante similitud coseno.

Usos habituales: clasificación zero-shot, búsqueda de imágenes mediante texto y filtrado inicial de catálogos. CLIP recupera o clasifica entre textos candidatos; no genera frases ni cajas.

De una foto a una opción

La predicción es el texto cuyo vector queda más cerca de la imagen.

Imagen

El encoder produce un vector.

Textos

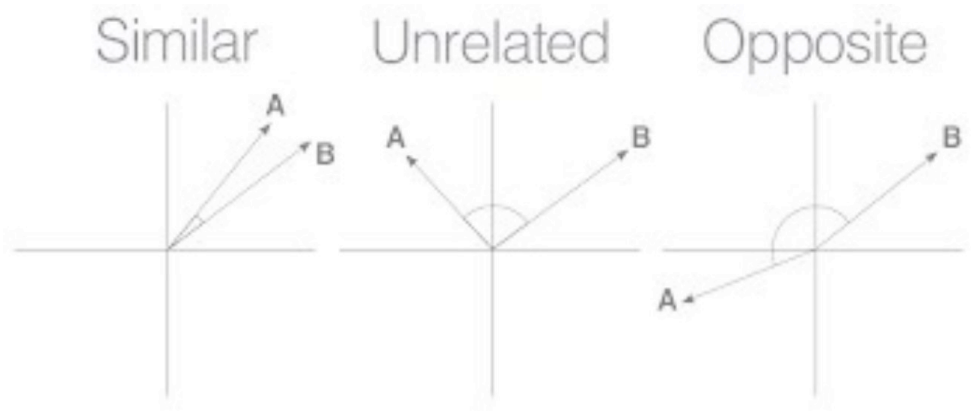
Cada opción produce otro vector.

Similitud

Se comparan todos los pares.

Orden

La puntuación prioriza opciones.



Aprender por asociación

En cada lote, CLIP acerca las parejas correctas y separa las combinaciones incorrectas.

La diagonal representa los pares imagen-texto que sí corresponden.

Un encoder visual y otro textual trabajan en paralelo. La pérdida contrastiva aumenta la similitud de cada imagen con su texto asociado y la reduce frente a los demás textos del lote.

CLIP no resuelve todo

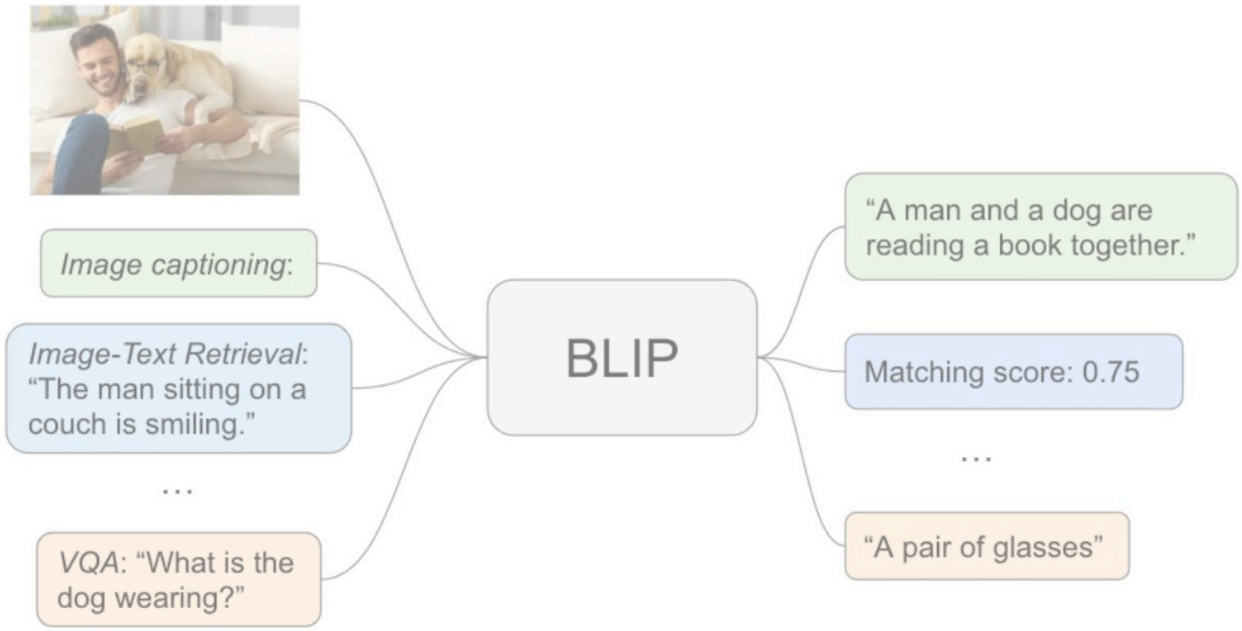
Un embedding global no equivale a una explicación detallada de la escena.

Sí aporta	No aporta solo
Similitud, recuperación y clasificación entre textos candidatos.	Frases, cajas, máscaras, OCR fiable o conteo preciso.

También puede heredar sesgos del entrenamiento, confundir atributos finos y ser sensible a la plantilla textual. Las negaciones suelen ser especialmente difíciles; conviene comparar formulaciones afirmativas y validar con datos propios.

De la imagen a una frase

BLIP genera texto condicionado por características visuales.



La atención cruzada permite al decodificador consultar la representación visual.

BLIP combina un encoder de imagen con componentes de lenguaje. Se puede usar para captioning y preguntas visuales sencillas. Una descripción plausible puede omitir elementos o inventar detalles, por lo que debe contrastarse cuando afecte a una decisión.

Texto que señala una región

Grounding DINO devuelve cajas para los objetos que coinciden con el prompt.

Prompt que nombra el concepto.

Cajas asociadas a fragmentos del texto.

Umbral que se valida en tu dominio.

Quieres localizar un defecto industrial nombrándolo en texto.

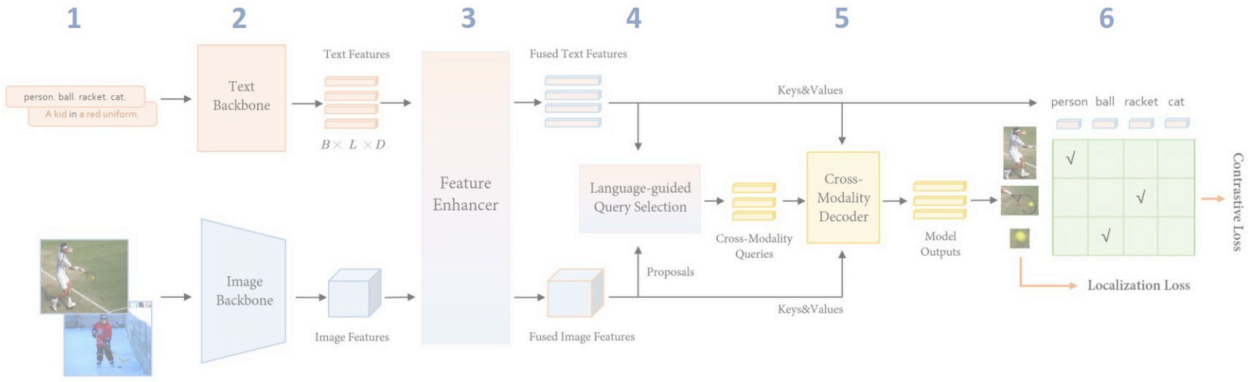
- A Usar Grounding DINO y evaluar con imágenes etiquetadas.
- B Suponer que el prompt elimina falsos positivos.
- C Medir solo el tamaño del checkpoint.

Respuesta A. Exacto: el prompt guía la búsqueda, pero la evaluación valida el uso.

El vocabulario abierto permite cambiar el concepto sin reentrenar una cabeza de clases fijas. Es útil para autoetiquetado y búsquedas ad hoc, pero no elimina falsos positivos ni falsos negativos.

El lenguaje guía dónde mirar

Las características visuales y textuales se fusionan antes de predecir las cajas.



La atención entre modalidades alinea palabras con regiones de la imagen.

El texto se codifica con un modelo lingüístico y la imagen con un backbone visual. Un bloque de fusión y un decoder basado en Transformer usan queries condicionadas por lenguaje para producir cajas y puntuaciones alineadas con el prompt.

Detección abierta y dirigida

La consulta define qué buscar y puede incluir atributos o referencias espaciales.

Sin clases fijas

El vocabulario llega en el prompt.

Sin anchors

Las queries predicen cajas directamente.

Fusión

Texto e imagen interactúan.

Autoetiquetado

Las cajas pueden iniciar un dataset.

Las expresiones compositivas y referenciales permiten pedir “coche rojo” o “persona de la izquierda”. La calidad depende de la formulación y de cuánto se parezca el dominio al preentrenamiento.

El prompt también puede fallar

Ambigüedad, objetos pequeños y escenas densas exigen validación específica.

Funciona bien como punto de partida

Inspección, retail, seguridad, agricultura y anotación asistida.

Requiere atención

Jerga, solapes, conceptos vagos y detalles diminutos.

En escenas críticas hay que medir precisión y exhaustividad por clase, tamaño y condición visual. Una caja generada automáticamente puede revisarse o combinarse con SAM para obtener una máscara.

Cómo ve un modelo de lenguaje

Un adaptador convierte las características visuales en tokens que el LLM puede recibir.

Visión	Adaptador	LLM
El encoder transforma píxeles en características.	Alinea dimensión y formato con el lenguaje.	Procesa tokens visuales e instrucción.

La imagen suele convertirse en una secuencia de tokens visuales que se combina con tokens de sistema y de usuario. El LLM genera la respuesta condicionado por ese contexto multimodal.

Conversar y estructurar

Un VLM puede describir, leer documentos y devolver información con un formato pedido.

Comprender	Razonar	Estructurar
OCR contextual, gráficos y diagramas.	Comparaciones e instrucciones paso a paso.	JSON o campos definidos por el usuario.

Qwen2.5-VL admite distintas relaciones de aspecto y puede producir coordenadas o datos estructurados. Un VLM suele requerir más memoria y tiempo que CLIP, BLIP o Grounding DINO; conviene reservarlo para tareas donde el razonamiento o la generación aportan valor.

Cada modelo pide otro lenguaje

El prompt debe describir la salida que el modelo realmente puede producir.

CLIP	Grounding	VLM
Plantillas afirmativas y varias formulaciones.	Conceptos breves separados y atributos visibles.	Instrucción, criterio y formato de salida.

Ejemplos: CLIP “una foto de {clase}”; Grounding DINO “persona con casco . persona sin casco”; VLM “lista los defectos visibles y devuelve JSON con tipo y evidencia”. Evitar preguntas vagas como “¿está bien?”.

La salida elige el modelo

Usa la herramienta más sencilla que produzca la respuesta necesaria.

CLIP recupera o clasifica.

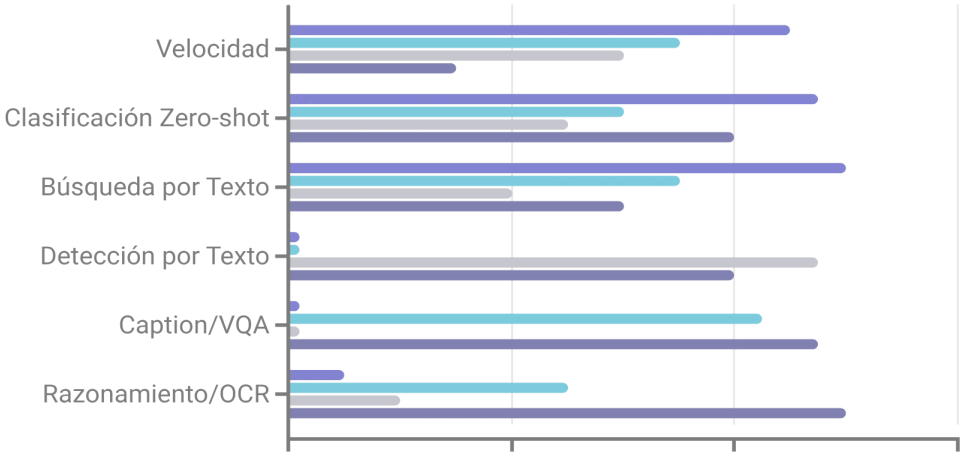
BLIP describe o responde algo sencillo.

Grounding DINO localiza; un VLM razona.

Necesitas una frase que describa una imagen.

- A BLIP.
- B CLIP.
- C Grounding DINO.

Respuesta A. Exacto: BLIP genera texto condicionado por una imagen.



Una respuesta fluida puede ser falsa

Evalúa alucinaciones, sesgos y errores con ejemplos representativos.

Datos

Casos reales y difíciles.

Métricas

Según clasificación, cajas o texto.

Umbrales

Ajustados al coste del error.

Revisión

Humana cuando la decisión es crítica.

Los benchmarks públicos no garantizan rendimiento en tu población o entorno. Conviene registrar prompts, versiones y condiciones; analizar falsos positivos y negativos; y exigir evidencia visual o un formato verificable cuando el modelo genere texto.

Cuatro formas de mirar

Compara CLIP, BLIP, Grounding DINO y Qwen con la misma imagen.

1. Guarda una copia en Drive.
2. Cambia imagen y prompts.
3. Contrasta salidas y errores.

Abrir en Colab ↗

Usa la misma imagen para separar capacidades: CLIP ordena textos, BLIP genera una descripción, Grounding DINO localiza conceptos y Qwen responde instrucciones. Registra el prompt, el modelo y el error observado; una salida más larga o fluida no implica una respuesta más correcta.