

## Issue1: Not visual search-centric.

**Case1:** Question may be solved by exploiting cross-referencing among textual entities in the question.

Question:

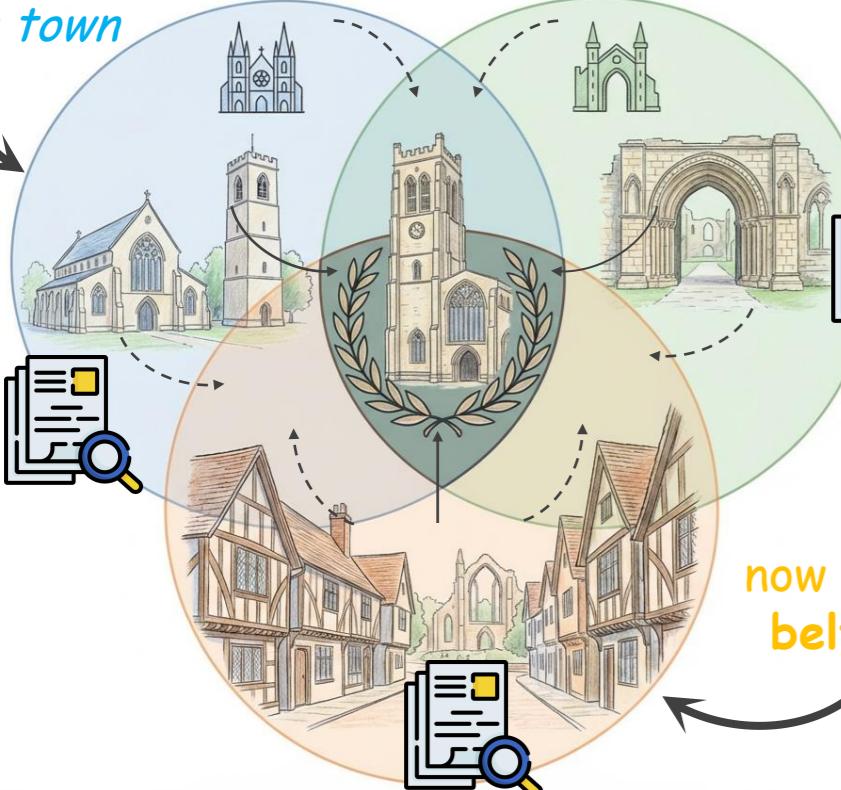
In the image, in a historic English town known for its rich medieval heritage, there stands a structure. This gate, dating back to the early 12th century, now functions as the belfry for a local church that has since become a cathedral. What was the original purpose of this structure when it was first built?

Visual Entity: Norman Tower in Bury St Edmunds.

Ground truth:

The gateway for the Abbey Church.

English town



12th century

Image search is

The phrase 'early 12th century gate' points to a rare Norman-period gateway. Its later reuse as the belfry of a church that became a cathedral, points specifically to the Norman Tower in Bury St Edmunds.



MLLM

now functions as the belfry for a local church



**Case2:** Question may be solved by MLLM prior world knowledge.

Question:

Based on the provided image, what is the location where this match is being played? Options: A. Camp Nou B. Parc des Princes C. Signal Iduna Park D. Allianz Arena E. Wembley Stadium

Ground truth:

C. Signal Iduna Park



MLLM

The player on the left is wearing Borussia Dortmund's signature yellow home kit, which indicates that the match is taking place at their home stadium in Germany. Borussia Dortmund's home stadium is Signal Iduna Park, therefore the correct answer is C. Signal Iduna Park.



## Issue2: overly idealized whole-image retrieval

**Case1:** Fail to reflect the noisy, ambiguous, and iterative nature of real-world visual search.

**One-shot perfect retrieval becomes nearly impossible.**

Question:

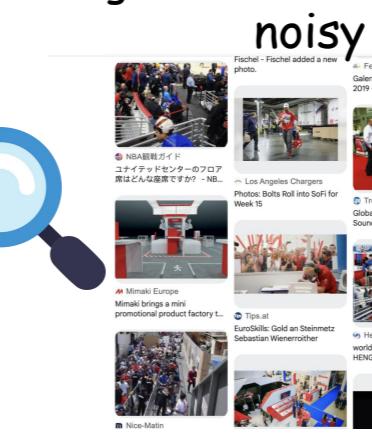
In the image, a crowd scene is shown at what appears to be a public event, where a person is holding a red umbrella featuring a recognizable logo. What brand does this umbrella belong to?

Ground truth:  
Ferrari.

Whole image Search



Crop Image Search



noisy



MLLM

The dominant visual of crowds and metal barriers suggests a generic 'public gathering' context. Combined with the blue path and distant stands, the scene implies a 'race track entrance,' leading me to identify the setting as an outdoor sporting venue. **Unable to find umbrella information**



MLLM

The distinct yellow shield points to a racing team's identity. The black prancing horse emblem points specifically to the **Ferrari logo**.

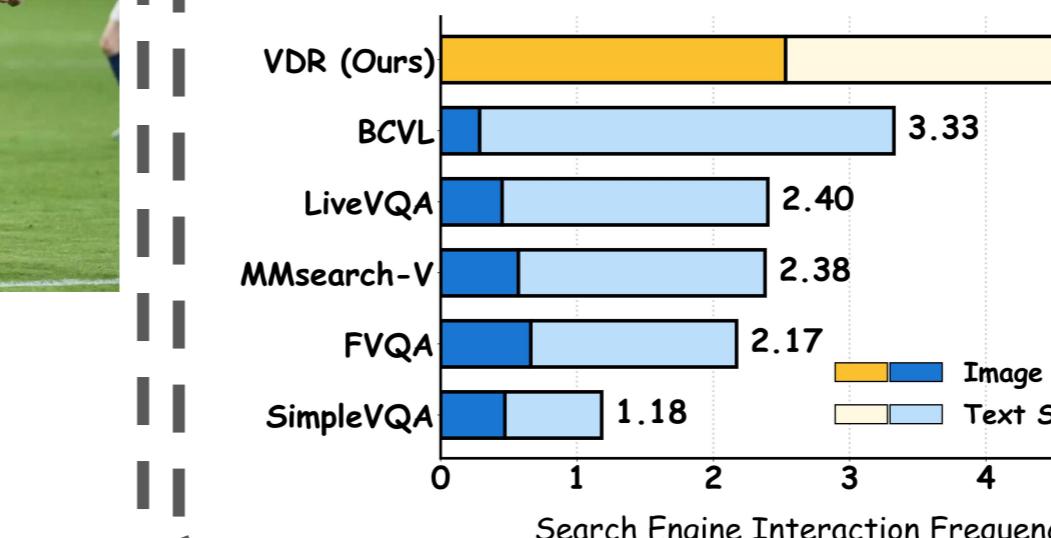
**Case2:** Fail to reflect the complexity of multi-hop reasoning and evidence aggregation.

Question:

Who is the developer of the game shown in the image?

Answer:

Rockstar Games.



Questions are overly shallow, underestimating multi-hop reasoning challenges.

