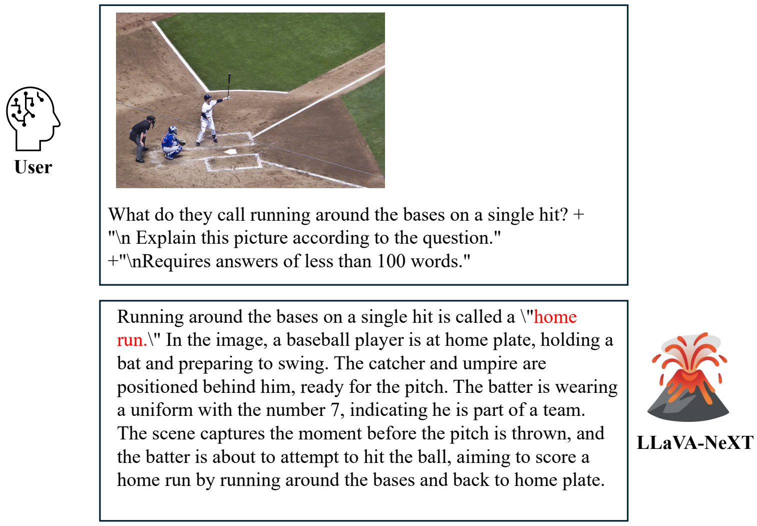
A. MLLM to generate relevant knowledge



What do they call running around the bases on a single hit? +
"\n Explain this picture according to the question."
+"\nRequires answers of less than 100 words."

Running around the bases on a single hit is called a \"home run.\" In the image, a baseball player is at home plate, holding a bat and preparing to swing. The catcher and umpire are positioned behind him, ready for the pitch. The batter is wearing a uniform with the number 7, indicating he is part of a team. The scene captures the moment before the pitch is thrown, and the batter is about to attempt to hit the ball, aiming to score a home run by running around the bases and back to home plate.

Figure A1. MLLM's internal implicit knowledge

We conducted an experiment to evaluate MLLM's ability to generate relevant implicit knowledge from images and questions. Specifically, we provided an image-question pair to MLLM and prompted it to interpret the image in the context of the question to extract relevant knowledge. As demonstrated in the Figure A1, MLLM successfully retrieved knowledge related to "*Home run*". However, when the same image and question were directly input into MLLM for reasoning, it incorrectly answered with "*Stealing*".

This result indicates that while MLLM possesses implicit knowledge about home runs, it fails to utilize it effectively when reasoning, instead selecting an incorrect reasoning path based on knowledge of "*Stealing*". This misalignment occurs due to MLLM's reliance on noisy or misleading information during inference.

Our method addresses this issue by filtering noise from explicit knowledge and guiding MLLM to retrieve the correct implicit knowledge based on both explicit knowledge and image content. This structured approach ensures that MLLM follows a more accurate reasoning path, leading to improved answer quality.

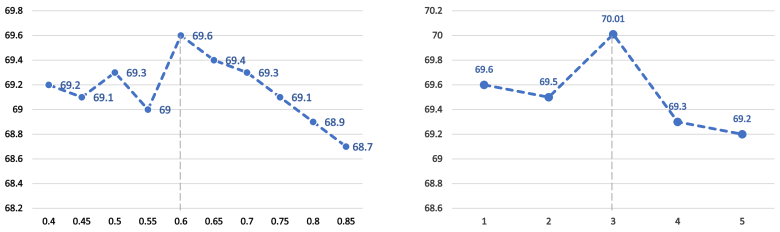
B. Parameter analysis

B.1. Choice of the number of λ

To effectively extract key image features that assist in answering questions while minimizing irrelevant information, we conducted a parametric analysis of the parameter λ. The results, as shown in Figure B1(a), indicate that the optimal performance of 69.6% is achieved when $\lambda = 0.6$. As λ increases beyond 0.6, performance likely drops because the model focuses on too few relevant features, introducing noise or negative effects. Conversely, when λ is less than 0.6, although the model initially benefits from more features, the excessive redundant information prevents it from reaching optimal results.

B.2. Choice of the number of candidate outputs

To mitigate variability in MLLM outputs, where the same input may yield different answers across runs, we store multiple generated answers as candidate outputs. These candidates are then leveraged to guide the model toward more consistent and reliable responses. The selection of the optimal number of iterations co is shown in Figure B1(b). We observe that when $co = 3$, MLLM tends to produce better results, as it effectively takes advantage of various answers without introducing noise. However, as co increases, the performance gain diminishes and may even result in negative optimization, where additional answers disrupt the model's reasoning process. This highlights the importance of carefully selecting model parameters to optimize performance while maintaining accuracy.



(a) Number of λ

(b) Number of co

Figure B1. Results of parameter analysis

B.3. Choice of the number of patch sizes

We conducted experiments to analyze the impact of patch size on model performance, keeping other parameters constant. As shown in Table B1, different patch sizes yielded varying results. These findings suggest that patch size significantly influences performance. Smaller patches capture fine-grained details but may introduce noise, while larger patches provide more global context but risk overlooking local details. Our results highlight the importance of selecting an optimal patch size to balance fine-grained and global feature extraction.

p	14	16	18	20
Score	68.1	69.8	69.1	68.6

Table B1. Results of different number of patch size p

B.4. Choice of the number of k

The parameter k controls the amount of retrieved knowledge from the external knowledge base, significantly impacting model performance. To explore this effect, we conducted experiments with varying k values while keeping other conditions unchanged, As shown in Table B2. As k increases from 1 to 5, the model score rises from 66.2 to 69.8, indicating that additional knowledge improves understanding and reasoning. However, at k=7, the score drops to 68.6, suggesting that excessive knowledge introduces redundancy and distracts the model. These results highlight that an optimal k balances knowledge richness and relevance, with k=5 yielding the best performance.

k	1	3	5	7
Score	66.2	68.1	69.8	68.6

Table B2. Results of different number of k

C. The relationship between identified regions and knowledge notes

To address the reviewer's suggestion of adding more quantitative comparisons, we conducted an extended experiment to rigorously measure the relationship between identified image regions and knowledge note tokens using correlation scores. Specifically, we employed $H_{i,j}$, derived from Eqs.7 and 8, to quantify the interaction between image patches and knowledge tokens.

Applying our method to a diverse set of image-question pairs, we observed clear patterns in the correlation scores, as illustrated in Figure C1. Tokens from the knowledge notes and semantically related image patches consistently exhibited higher correlation scores. For instance, in an image containing traffic-related elements with knowledge notes on traffic rules, patches corresponding to a traffic light and tokens referring to "green light" showed significantly high correlation, as evident in the heatmap.

Our method accurately identifies and emphasizes strong associations between relevant image regions and knowledge note tokens. These results provide both quantitative validation of our method's effectiveness in linking visual content with

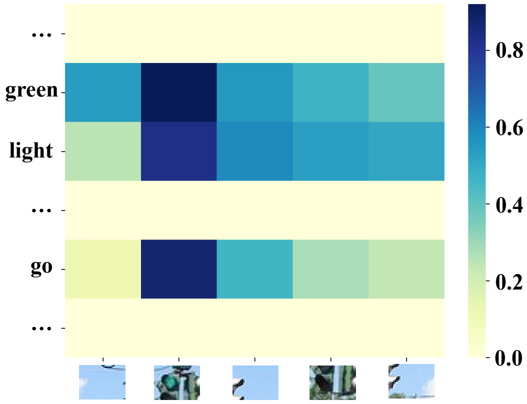


Figure C1. The attention scores between image patches and knowledge notes tokens.

knowledge-based information. This demonstrates our method's ability to better leverage the interplay between images and knowledge notes.

D. Details of prompt

c_k in Equation (5):

Visual context: V \n
Summary information about the retrieved knowledge linked to this image. If the retrieved knowledge is not relevant to the image, return the image captions. \n
Retrieved knowledge: P \n
Return answers that are as concise as possible and do not exceed 100 words.

Figure D1. Details of prompt c_k in Equation 5

c in Equation (10):

Your task is to perform visual question answering (VQA). You will receive three parts of information: \n
1. The original image for full context, providing a complete scene. \n
2. The Region of Interest (ROI) image called visual notes, where colored areas indicate focus and white areas are irrelevant. Mark the important areas related to the problem. Note that the visual notes may contain noise. \n
3. Additional knowledge, providing background information relevant to the issue. \n
Please answer the questions in combination with the three parts, mainly based on the information of the original image, using the visual notes as an aid, and referring to additional knowledge to improve accuracy. \n
Analyzing both the visual notes and the original image. Focus on the areas in the original image that correspond to the colored areas of the visual notes. \n
Ignore irrelevant noise areas and generate answers that match the question. \n
Original Image: V \n
Visual Notes: N_{ig} \n
Knowledge Notes: N_{kl} \n
question: q \n
Answer the question using a single word or phrase. \n
Answer:

Figure D2. Details of prompt c in Equation 10

$\bar{c}$ in Equation (11):

Your task is to perform visual question answering (VQA). You will receive three parts of information: \n
1. The original image for full context, providing a complete scene. \n
2. The Region of Interest (ROI) image called visual notes, where colored areas indicate focus and white areas are irrelevant. Mark the important areas related to the problem. Note that the visual notes may contain noise. \n
3. The candidate outputs for you to use as a reference, note that the candidate outputs may contain noise. \n
4. Additional knowledge, providing background information relevant to the issue. \n
Please answer the questions in combination with the three parts, mainly based on the information of the original image, using the visual notes as an aid, and referring to additional knowledge and candidate outputs to improve accuracy. \n
Analyzing both the visual notes and the original image. Focus on the areas in the original image that correspond to the colored areas of the visual notes. \n
Ignore irrelevant noise areas and generate answers that match the question. \n
original image: V \n
Visual Notes: N_{ig} \n
Candidate Outputs: Y \n
Knowledge Notes: N_{kl} \n
question: q \n
Answer the question using a single word or phrase. \n
Answer:

Figure D3. Details of prompt $\bar{c}$ in Equation 11

E. Specific elements of the knowledge

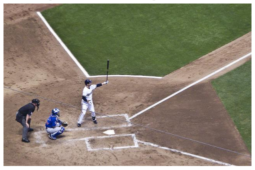
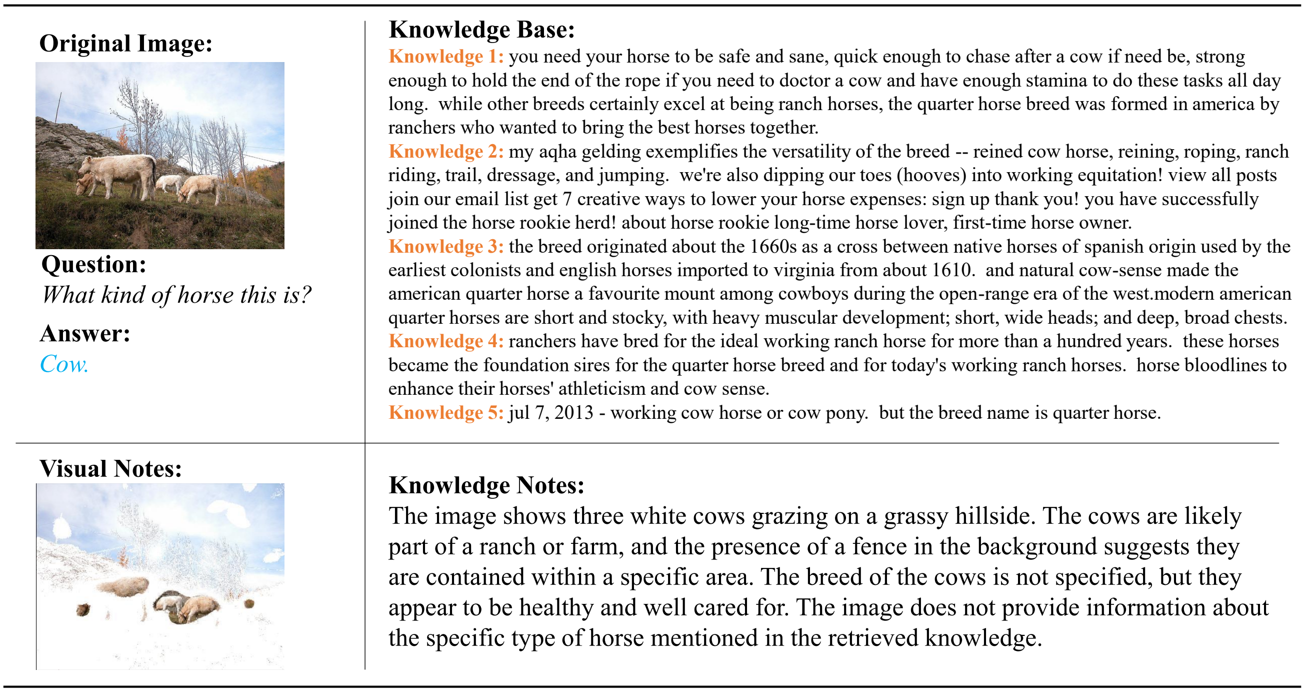
Original Image:  **Question:** *What do they call running around the bases on a single hit?* **Answer:** *Home run.*	**Knowledge Base:** **Knowledge 1:** home runs are often characterized by the number of runners on base at the time. a home run hit with the bases empty is seldom called a \"one-run homer\", but rather a solo home run, solo homer, or \"solo shot\". **Knowledge 2:** home runs: everybody (well, almost everybody) loves them. and, everybody's got a different name for them. **Knowledge 3:** a home run occurs when a batter hits a fair ball and scores on the play without being put out or without the benefit of an error. home runs can be a great stat for evaluating a hitter's power. they're also good for measuring the success of pitchers, who strive to limit home runs. now, however, most home runs feature players trotting around the bases after hitting the ball over the fence. **Knowledge 4:** if you hit one out of the park, you've hit a homer. is what a batter gets after hitting the ball and running all the way around the bases to home plate without being tagged out. if there are players on all the bases when a batter gets a homer, it's called a grand slam and it gives the team four points. you can use this word as a verb too, meaning \"hit a home run.\" there are also nearly twenty towns in the us called homer, but these are named for the ancient greek epic poet, homer. **Knowledge 5:** to accelerate through first base (french kissing), onto second base (\"heavy petting\“) to third base (oral sex) and finally coming around to home plate (sexual intercourse).wow, hit a home run with sheila tonight, gimme a high five barry!by cayle funck october 10, 20034243959 flagget a home run mug for your barber sarah.jan 12 word of the daypseudocoupcan be pronounced like sudoku if wordplay demands it.
Visual Notes: 	**Knowledge Notes:** A home run is a baseball hit that allows the batter to run around all four bases and score. It is a powerful hit that is often sought after by batters and can be a great stat for evaluating a hitter's power. Home runs can also be a good measure of a pitcher's success, as they strive to limit them. The term \"home run\" is often used colloquially to refer to a successful or impressive achievement.

Figure E1. Specific elements of the knowledge in case 1



Original Image:

Question:

What kind of horse this is?

Answer:

Cow.

Knowledge Base:

Knowledge 1: you need your horse to be safe and sane, quick enough to chase after a cow if need be, strong enough to hold the end of the rope if you need to doctor a cow and have enough stamina to do these tasks all day long. while other breeds certainly excel at being ranch horses, the quarter horse breed was formed in america by ranchers who wanted to bring the best horses together.

Knowledge 2: my aqha gelding exemplifies the versatility of the breed -- reined cow horse, reining, roping, ranch riding, trail, dressage, and jumping. we're also dipping our toes (hooves) into working equitation! view all posts join our email list get 7 creative ways to lower your horse expenses: sign up thank you! you have successfully joined the horse rookie herd! about horse rookie long-time horse lover, first-time horse owner.

Knowledge 3: the breed originated about the 1660s as a cross between native horses of spanish origin used by the earliest colonists and english horses imported to virginia from about 1610. and natural cow-sense made the american quarter horse a favourite mount among cowboys during the open-range era of the west.modern american quarter horses are short and stocky, with heavy muscular development; short, wide heads; and deep, broad chests.

Knowledge 4: ranchers have bred for the ideal working ranch horse for more than a hundred years. these horses became the foundation sires for the quarter horse breed and for today's working ranch horses. horse bloodlines to enhance their horses' athleticism and cow sense.

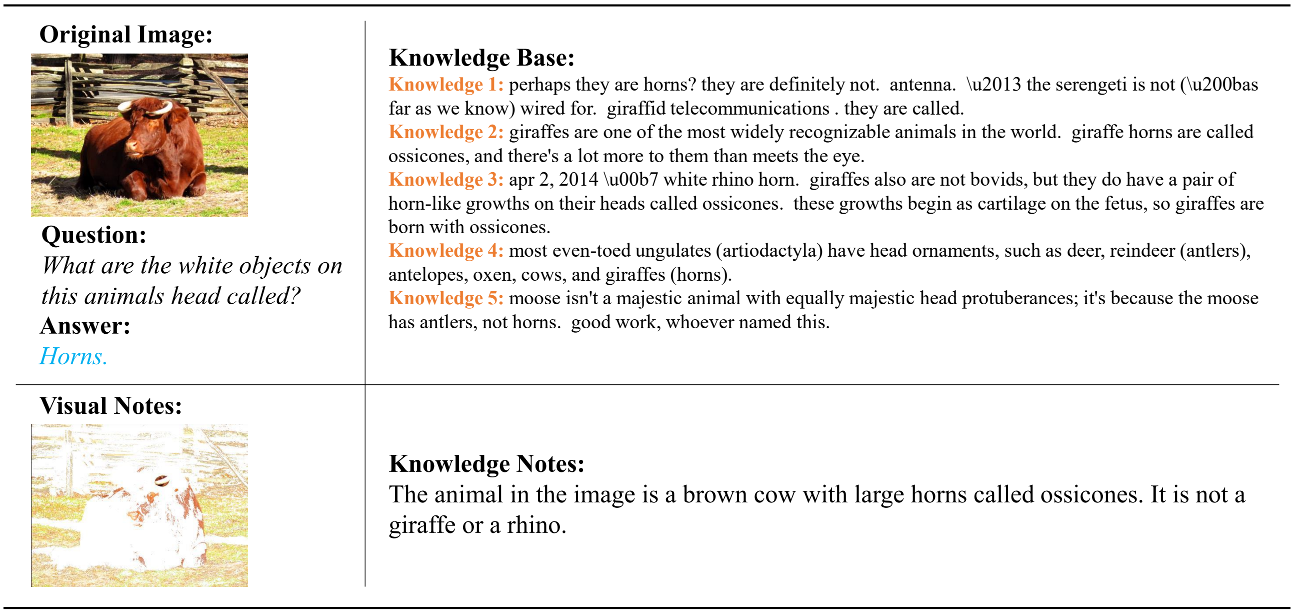
Knowledge 5: jul 7, 2013 - working cow horse or cow pony. but the breed name is quarter horse.

Visual Notes:

Knowledge Notes:

The image shows three white cows grazing on a grassy hillside. The cows are likely part of a ranch or farm, and the presence of a fence in the background suggests they are contained within a specific area. The breed of the cows is not specified, but they appear to be healthy and well cared for. The image does not provide information about the specific type of horse mentioned in the retrieved knowledge.

Figure E2. Specific elements of the knowledge in case 2



Original Image:

Question:

What are the white objects on this animals head called?

Answer:

Horns.

Knowledge Base:

Knowledge 1: perhaps they are horns? they are definitely not. antenna. \u2013 the serengeti is not (\u200bas far as we know) wired for. giraffid telecommunications . they are called.

Knowledge 2: giraffes are one of the most widely recognizable animals in the world. giraffe horns are called ossicones, and there's a lot more to them than meets the eye.

Knowledge 3: apr 2, 2014 \u00b7 white rhino horn. giraffes also are not bovids, but they do have a pair of horn-like growths on their heads called ossicones. these growths begin as cartilage on the fetus, so giraffes are born with ossicones.

Knowledge 4: most even-toed ungulates (artiodactyla) have head ornaments, such as deer, reindeer (antlers), antelopes, oxen, cows, and giraffes (horns).

Knowledge 5: moose isn't a majestic animal with equally majestic head protuberances; it's because the moose has antlers, not horns. good work, whoever named this.

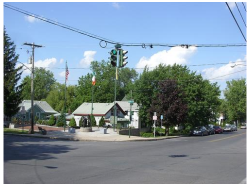
Visual Notes:

Knowledge Notes:

The animal in the image is a brown cow with large horns called ossicones. It is not a giraffe or a rhino.

Figure E3. Specific elements of the knowledge in case 3

Original Image:



Question:

What does the traffic signal indicate drivers should do?

Answer:

Go.

Knowledge Base:

Knowledge 1: if a traffic signal is not functioning at all at an intersection, all drivers must treat the intersection as if a stop sign is posted for all directions. a red light means you must make a complete stop before entering the crosswalk or intersection and wait until the light turns to green before proceeding. in this case you may go straight ahead only. a green light means you may proceed if it is safe to do so after stopping for pedestrians and yielding to vehicles within the intersection. appropriate travel lanes on a roadway utilizing a reversible lane system are indicated as follows: no travel allowed in this lane in the direction you are going.

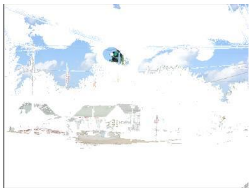
Knowledge 2: traffic signs are devices placed along, beside, or above a highway, roadway, pathway, or other routes to guide, warn, and regulate the flow of traffic, including motor vehicles, bicycles, pedestrians, equestrians, and other travelers. this sign means that you must make a complete stop before entering a crosswalk, passing the limit line, or entering the intersection. when safe, back out or turn around and go back to the road you were originally on. this warning sign indicates that there may be pedestrians crossing the roadway ahead.

Knowledge 3: how do you make a left turn with the signal is solid green and not a green arrow?if you are in the intersection making a left turn when the yellow light appears, proceed as soon as traffic allows and it is safe.how do you make a left turn with the signal is solid yellow?the solid yellow indication with an arrow means the signal for the turn is changing. drivers must yield to oncoming traffic and pedestrians in crosswalks, before turning. also, even if you want to go straight and an officer indicates that you must turn, you are required to turn.how do you handle a situation where the lights at an intersection are working but there is a police officer directing traffic?traffic signals are lights that regulate the flow oftraffic mainly through intersections.

Knowledge 4: red\u2013amber signal, and indicates that drivers may pass if no pedestrians are on the crossing. in some countries traffic signals will go into a flashing mode if the conflict monitor detects a problem, such as a fault that tries to display green lights to conflicting traffic.

Knowledge 5: as indicated with pennsylvania code title 67, chapter 212; how does penndot decide whether a traffic signal should be installed on a state highway? 5. how do i go about getting a traffic signal installed at an intersection? 7. most intersections would not necessarily be improved or made safer by the installation of a traffic signal. they can increase traffic on the side streets as drivers seek alternative routes through neighborhoods

Visual Notes:



Knowledge Notes:

The image shows a traffic light at an intersection with a green light. Drivers must stop if the light is red, proceed with caution if it is yellow, and go straight ahead if it is green. Traffic signals are installed at intersections to control the movement of vehicles and pedestrians.

Figure E4. Specific elements of the knowledge in case 4