Downscaling Intelligence: Exploring Perception and Reasoning Bottlenecks in Small Multimodal Models

Supplementary Material

A1. Additional LLM Downscaling Results and Details

Per-tasks results. We present plots showing the performance dropoff from LLM downscaling across all evaluated tasks in Figure A1. As described in the main text, most tasks exhibit minimal performance decline when downscaling the LLM, except for a handful of vision-centric tasks that exhibit substantially larger drops (e.g., Grounding, NIGHTS, PieAPP).

Full decoupled results. We plot the performance dropoff from LLM downscaling of the perception and reasoning modules in Figure A2. We find that LLM downscaling of either module leads to performance degradation across a wide range of tasks. Notably, downscaling the perception module has a large effect on both tasks assessing perception (e.g., OCR-VQA, Fine-grained Perception) and reasoning (e.g., Logical Reasoning). One exception is Math, where LLM downscaling of the perception module has little impact. We expect this is because mathematical ability is limited primarily by the downstream process of operating on visual information (reasoning) rather than by the foundational perception ability.

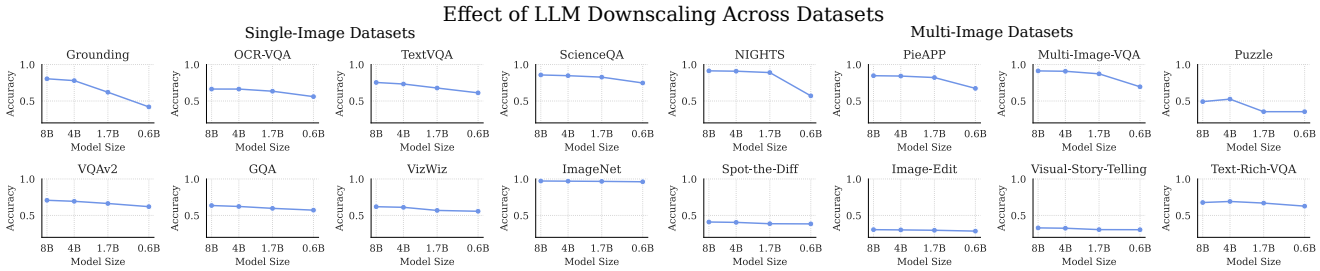


Figure A1. Performance dropoff from downscaling LLM across all datasets.

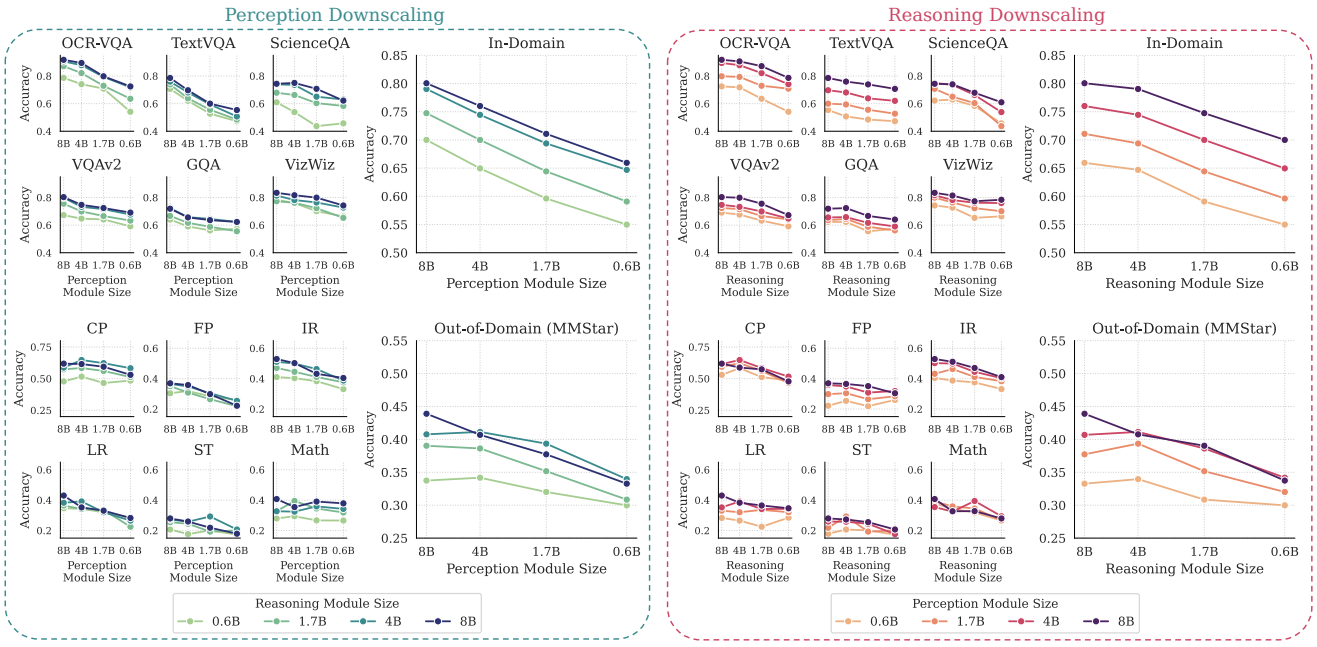


Figure A2. Full decoupled results. CP=Coarse Perception, FP=Fine-grained Perception, IR=Instance Reasoning, LR=Logical Reasoning, ST=Science & Technology.

LLaVA-OneVision as the perception module. While our main analysis used models trained from scratch for a controlled study, here we also experiment with using LLaVA-OneVision ($\in$ 0.5B, 7B) as the perception module in the decoupled framework. We first present decoupled results using the same reasoning module as in our experiments (Qwen3). As shown in Figure A3, this configuration produces results that are largely consistent with those obtained using our controlled model as the perception module, where LLM downscaling of either module hinders performance. We do observe, however, that downscaling the perception module has a smaller effect than in our controlled study for in-domain data. This is likely because LLaVA-OneVision includes extensive training on captioning, which we demonstrate alleviates the perception bottleneck.

Additionally, we experiment with using Qwen2 as the reasoning model in this setup (the LLM used in LLaVA-OneVision). As shown in Figure A4, relative to the earlier results using Qwen3 as the reasoning module, we see a larger impact from downscaling the reasoning module on the in-domain tasks, and overall performance is weaker than when using Qwen3. This outcome is not surprising, as Qwen3 has demonstrated stronger performance than Qwen2 on textual tasks, particularly for smaller model variants.

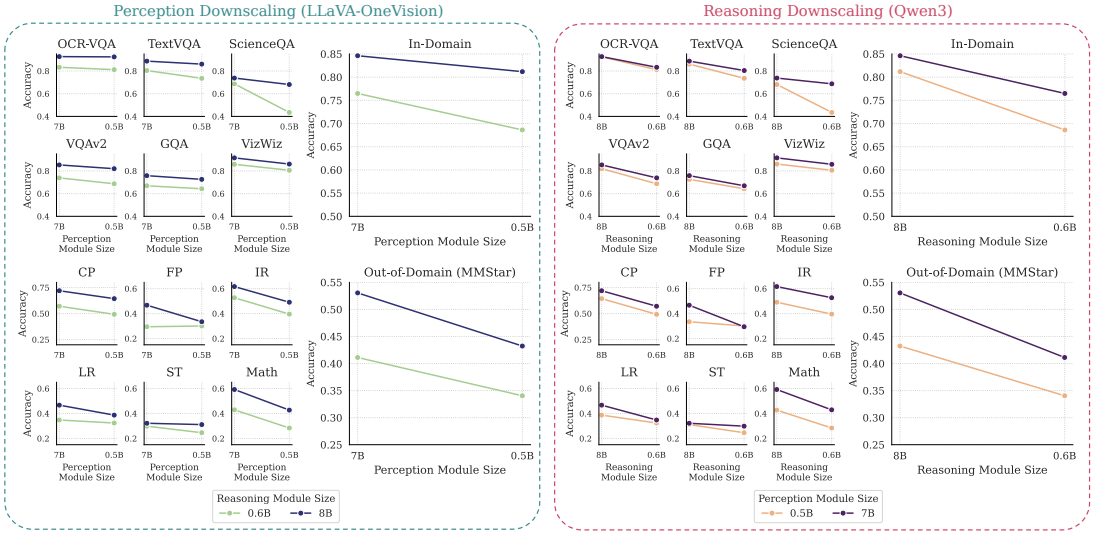


Figure A3. Decoupled analysis using LLaVA-OneVision as the perception module and Qwen3 as the reasoning module. CP=Coarse Perception, FP=Fine-grained Perception, IR=Instance Reasoning, LR=Logical Reasoning, ST=Science & Technology.

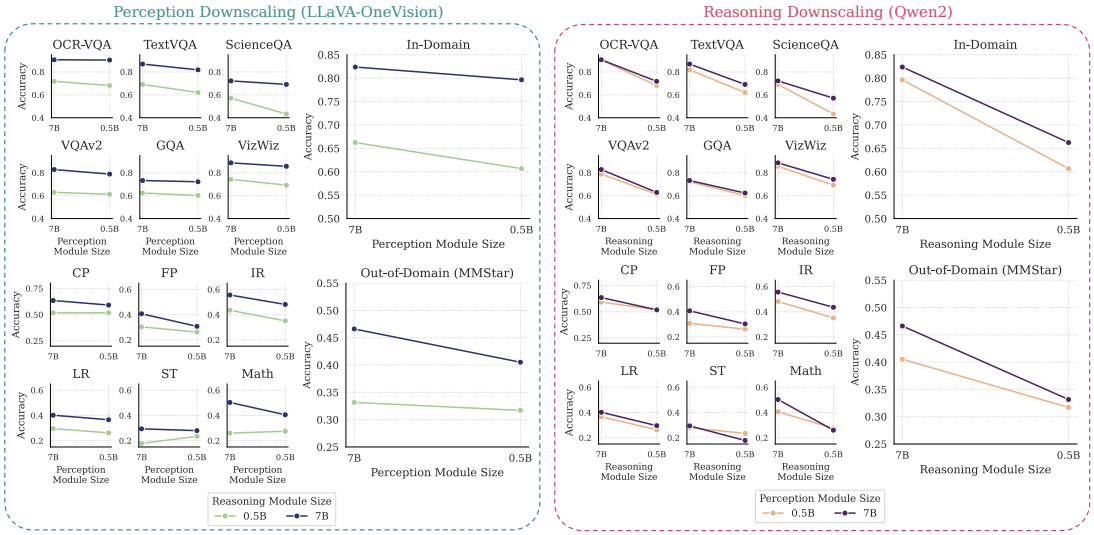


Figure A4. Decoupled analysis using LLaVA-OneVision as the perception module and Qwen2 as the reasoning module. CP=Coarse Perception, FP=Fine-grained Perception, IR=Instance Reasoning, LR=Logical Reasoning, ST=Science & Technology.

Prompts for decoupled analysis. We list prompt templates for our decoupled perception and reasoning analysis in Figure A5. These prompts follow the Prism framework, except that the initial instruction for obtaining question-specific information is run offline using the same model throughout, ensuring that deriving question-specific instructions does not influence our analysis of perception and reasoning downscaling. Thus, the question-specific information inserted into the perception module prompt is consistent across all model setups.

Perception bottleneck across model families. We additionally analyze Gemma3 (12B $\rightarrow$ 4B) and InternVL2.5 (8B $\rightarrow$ 2B), which offer multiple LLM sizes with a fixed vision encoder, enabling a controlled analysis. As shown in Table A1, evaluating on MMStar using the same procedure as §3.3, we observe consistent trends of decreasing performance when downscaling the perception module's LLM.

P Module	R Module	P LLM Size	Avg. Acc. Drop
Ours (§3)	Qwen3	8B $\rightarrow$ 1.7B	3.30
InternVL2.5	InternLM2	8B $\rightarrow$ 2B	4.33
Gemma3	Gemma3	12B $\rightarrow$ 4B	5.26

Table A1. Effect of downscaling the perception module on MMStar accuracy. For each model family, we vary the perception module's LM size while holding the reasoning module fixed, and report the accuracy drop averaged across both reasoning module sizes. P=Perception, R=Reasoning.

Question-specific Instruction Prompt

Your task is to give a concise instruction about what basic elements are needed to be described based on the given question. Ensure that your instructions do not cover the raw question, options or thought process of answering the question.
Examples:
Question: In which period the number of full time employees is the maximum?
Contents to observe: the number of full time employees
Question: What is the value of the smallest bar?
Contents to observe: the heights of all bars and their values
Question: What is the main subject of the image?
Contents to observe: the central theme or object
Question: What is the position of the catcher relative to the home plate?
Contents to observe: the spatial arrangement of the objects
Question: What is the expected ratio of offspring with white spots to offspring with solid coloring? Choose the most likely ratio.
Contents to observe: the genetic information
Now, perform the task, and format your answer as "Contents to observe:"
Question: <question>

Perception Module Prompt

Describe the fine-grained content of the image, including scenes, objects, relationships, instance location, and any text present.
Especially, pay attention to <question-specific info>

Reasoning Module Prompt (w/o thinking)

You are an excellent text-based reasoning expert. You are required to answer the question based on the detailed description of the image.
Description: <description>
Question: <question>
Answer directly with the optiońs letter in the format of "Answer:". Do not add anything other than the letter answer after "Answer:".

Reasoning Module Prompt (w/ thinking)

You are an excellent text-based reasoning expert. You are required to answer the question based on the detailed description of the image.
Description: <description>
Question: <question>
Please reason step by step, and give the final answer on the last line by itself in the format of "Answer:". Do not add anything other than the letter answer after "Answer:".

Figure A5. Prompt templates for the decoupled perception / reasoning analysis.

A2. Additional Visual Extraction Tuning Details

Visual extraction tuning data generation pipeline. We present the prompt templates for our pipeline generating visual extraction tuning data in Figure A6. In the first stage, each question–answer pair in a visual instruction example is converted into a declarative statement. These statement(s) are then inserted into a prompt that instructs the model to describe fine-grained visual details, with explicit emphasis on the information contained in the visual instruction example.

Visual extraction tuning data examples. We provide examples produced by our visual extraction tuning data generation pipeline in Figure A7. We see that the generated examples preserve the original information required to answer the questions from the visual instruction while also offering a more detailed and comprehensive depiction of the specific image content.

Question-answer Pair to Declarative Statement Prompt

Your task is to convert each question–answer pair about an image into a concise, fully self-contained declarative statement. The resulting statements should be understandable on their own, without requiring the reader to refer to the original question.

```
% for each QA pair in conv include:
```
Question: <question>
Answer: <answer>

```
% if len(conv) > 1 include:
```
As there are <len(conv)> questions, you should respond with <len(conv)> statements. Include each statement on its own line
Declarative Statement(s):

Visual Extraction Response Generation Prompt

Your task is to describe the fine-grained content of the image, including scenes, objects, relationships, instance location, and any text present.
As part of your description, you should incorporate the following information about the image.
<declarative statements>
Description:

Figure A6. Prompts used for visual extraction tuning data generation pipeline.

Image from GQA

Visual Instruction Example

Human: Is the cheese to the left of the food on the plate?
Provide your response using only one word or a short phrase.
Assistant: Yes
Human: What is the color of this plate
Assistant: Black
Human: Do the plate and the table have the same material
Assistant: No
Human: What material is the table made of?
Assistant: Wood
Human: How is the fried meat called
Assistant: Bacon
Human: Is the plate both round and white?
Assistant: No
Human: Is the cheese on the left of the photo?
Assistant: Yes
Human: Are there both forks and plates?
Assistant: No
Human: Which kind of food is not melted?
Assistant: Bacon
Human: Is the round plate made of plastic?
Assistant: Yes
Human: What is the plate made of?
Assistant: Plastic
Human: What is the meat to the left of the food on the plate?
Assistant: Bacon
Human: Is there bacon to the left of the food on the plate?
Assistant: Yes

Converted Visual Extraction Example

Human: Describe the fine-grained content of the image, including scenes, objects, relationships, instance location, and any text present. Especially, pay attention to positions, colors, materials, shapes, types of food, and objects on the plate.

Assistant: This is a close-up, overhead shot of a plated meal on a black, round plate made of plastic. The plate rests on a wooden table, which is visible in the background and has a warm, grainy texture, indicating it is made of wood — a material different from the plate.
The meal is composed of several distinct components arranged on the plate:
- On the left side of the photo, there is a cheeseburger patty topped with a generous amount of melted orange cheese. Draped over the cheese and burger are several strips of crispy, cooked bacon, which are not melted into the cheese but are laid on top. The bacon is visibly curled and has a reddish-brown color with crisp, slightly charred edges.
- To the right of the burger, there is a serving of bright green, steamed broccoli florets, which appear tender and slightly glistening.
- Above the broccoli and to the right of the burger, there is a pile of roasted Brussels sprouts. They are halved and show signs of roasting, with some charred edges and a glossy sheen from oil or seasoning.
- In the upper right portion of the plate, there are thick-cut, roasted potato wedges. They are golden brown with crispy-looking edges and appear seasoned with spices.

The food items are arranged in a somewhat clustered fashion, with the burger as the central focus on the left, and the vegetables and potatoes filling the rest of the plate. The lighting highlights the textures of the food, from the crisp bacon to the soft cheese and the roasted vegetables. There is no visible text or utensils in the image.

Image from VQAv2

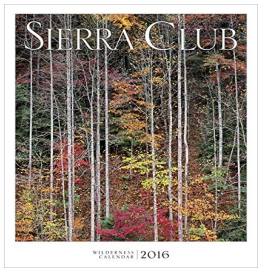
Visual Instruction Example

Human: What color is the snow?
Answer the question with just one word or a short phrase.
Assistant: white
Human: What is the person doing?
Assistant: skiing
Human: What color is the persons headwear?
Assistant: red

Converted Visual Extraction Example

Human: Describe the fine-grained content of the image, including scenes, objects, relationships, instance location, and any text present. Especially, pay attention to the color of the snow, the action of the person, the color of the headwear
Assistant: This is a dynamic, high-angle action shot of a skier descending a steep, snow-covered slope. The skier is captured mid-turn, leaning into the curve with their body angled sharply downwards, indicating speed and control. They are wearing a bright red helmet, which stands out against the white snow and their dark clothing. Their attire consists of a dark, possibly black, jacket with a lighter gray or beige panel on the chest, black ski pants, and gloves. A black backpack is strapped to their back. The skier is holding ski poles, with one pole planted in the snow to aid in balance and steering. The snow is uniformly white and appears powdery, with visible tracks from the skis and shadows cast by the surrounding trees. The background is dominated by the snow-covered hillside, with the dark, dense foliage of evergreen trees visible at the top of the frame, suggesting a forested mountain environment. The image conveys a sense of motion and adventure.

Image from OCR-VQA



Visual Instruction Example

Human: Who wrote this book?
Select the correct answer from the provided options and reply with the letter associated with it.
Assistant: Sierra Club
Human: What is the title of this book?
Assistant: Sierra Club Wilderness Calendar 2016
Human: What type of book is this?
Assistant: Calendars
Human: Is this a historical book?
Assistant: No
Human: Which year's calendar is this?
Assistant: 2016

Converted Visual Extraction Example

Human: Describe the fine-grained content of the image, including scenes, objects, relationships, instance location, and any text present. Especially, pay attention to the author's name, book title, book type, historical context, and calendar year.
Assistant: This is the cover of the "Sierra Club Wilderness Calendar 2016," a calendar book published by the Sierra Club. The cover features a vibrant, full-bleed photograph of a forest in autumn. The scene is dominated by tall, slender trees with pale, smooth bark, likely birches, standing vertically and creating strong, repeating lines that draw the eye upward. Their bare branches contrast with the colorful foliage on the forest floor and in the background, which displays a rich tapestry of fall colors: fiery reds, deep oranges, golden yellows, and patches of lingering green. The composition is dense and immersive, evoking a sense of wilderness and seasonal change.
Text is overlaid on the image in a clean, white serif font. At the top, the words "SIERRA CLUB" are prominently displayed in large, capitalized letters. At the bottom, centered, the text reads "WILDERNESS CALENDAR 2016," indicating the book's purpose and year. The overall design is minimalist and elegant, using the natural beauty of the photograph as the primary focus to convey the organization's connection to nature. The white border surrounding the image and text gives the cover a clean, finished look."

Figure A7. Visual extraction tuning data examples.

A3. Additional Step-by-step Reasoning Details and Results

NoWait Setup. To reduce overthinking of Qwen3 with thinking mode enabled, we use a logits processor that suppresses self-reflection tokens. Namely, we mask the logits of any token that contains one of the following keywords: {wait, alternatively, hmm, but, however, alternative, another, check, double-check, oh, maybe, verify, other, again, now, ah, anyway, anyhow}, while manually excluding words that only contain a keyword as a substring but are not reflexive (e.g., waiter).

Full results. We present results from performing step-by-step visual reasoning across all tasks in Figure A8. Expectedly, we find that Math heavily benefits from CoT reasoning (consistent with findings in text-only Math tasks).

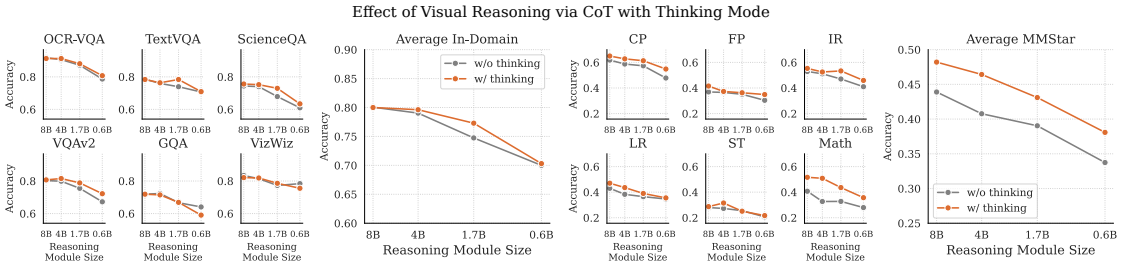


Figure A8. Full results showing impact of step-by-step reasoning on in-domain and out-of-domain (MMStar) performance.

A4. Additional EXTRACT+THINK Analyses

Inference latency analysis. In Figure A9, we characterize the tradeoff between our approach's parameter/data efficiency and its inference latency. We plot latency across end-to-end baselines, PrismCaptioner (1.8/70B), and our EXTRACT+THINK method (1.7/4B w/ and w/o CoT). While end-to-end approaches incur lower latency from generating far fewer tokens, our method achieves 27.4% improvement on MMStar and 8.8% on VMCBench tasks. Additionally, vLLM greatly increases generation throughput for longer output sequences, narrowing the gap, particularly at a higher batch size.

Comparing against PrismCaptioner underscores the benefit of visual extraction tuning: even w/o CoT, our method improves performance by 9.80% on MMStar and 10.9% on VMCBench while using a smaller reasoning module and reducing latency. This indicates that increasing the token budget via a two-stage design alone does not drive performance; rather, it is the visual extraction tuning that drives gains within this framework. Applying test-time scaling through CoT further improves performance, particularly on MMStar.

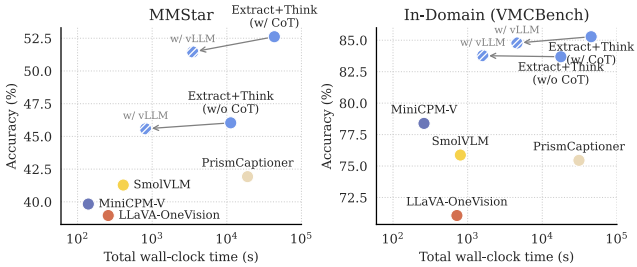


Figure A9. Inference latency on 1 L40s using Transformers (BS=1) or vLLM (BS=32), evaluated on MMStar and In-Domain (VMCBench).

Information bottleneck in two-stage design. While the two-stage approach offers a modular framework for independently studying and improving perception and reasoning and has been shown to enable strong performance, a drawback is that its design requires all relevant visual information for answering the question to be captured and adequately described in the initial stage. This stage acts as an intermediate representation: the reasoning module bases its final response on this description rather than the image itself. In Figure A10, we show examples where this leads to loss of critical visual information, ultimately leading the reasoning module astray. For instance, a musician may be described as playing an instrument without specifying which one, or players may be described without indicating their exact, relative spatial relationship to the camera, rendering certain questions unanswerable. Nonetheless, the proposed visual extraction tuning approach is specifically designed to mitigate such issues by targeting the extraction of question-relevant visual details and has been shown to effectively improve performance across evaluated tasks.

Question: How many people are playing guitar in the image?
Options: A: One, B: Two, C: Three, D: Four
Answer: C (Three)

Extraction output: "In the foreground, a guitarist is prominently featured, holding a black electric guitar and playing it with a focused expression... Behind him, another guitarist is visible, also playing an electric guitar, though slightly out of focus. **A third musician**, partially obscured, is visible in the background, also **playing an instrument**..."

Final prediction: B (Two)

Question: What color is the softball?
Options: A: Blue, B: Red, C: Yellow, D: Green
Answer: D (Green)

Extraction output: "The player on the right, wearing a white pinstripe jersey with red sleeves and blue pants, is in a defensive stance, reaching out with her right hand to block or intercept the ball. Her jersey has the word "Raiders" in red script across the chest, and the number "4" is visible below it. She is wearing a brown leather catcher's mitt on her left hand, which is holding a **bright yellow softball**..."

Final prediction: C (Yellow)

Question: What is the primary color of the shoes worn by the player closest to the camera?
Options: A: Yellow, B: Black, C: Gray, D: Orange
Answer: C (Gray)

Extraction output: "The central focus is on a player in a white jersey with orange trim and the word "TENNESSEE" across the chest, who is dribbling a basketball with both hands. She is wearing bright orange sneakers and is in motion, leaning forward with her body low to the ground, suggesting she is maneuvering around defenders... Surrounding her are three opposing players in maroon jerseys... **The players are positioned close to the camera**, creating a sense of immediacy and intensity in the moment."

Final prediction: D (Orange)

Figure A10. Examples showing failure modes of two-stage framework for visual question answering (using EXTRACT+THINK w/ 1.7B perception module).