

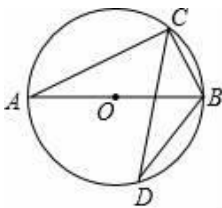
OneThinker: All-in-one Reasoning Model for Image and Video

Supplementary Material

A. Reasoning Examples

Image QA

Question: As shown in the figure, AB is the diameter of $\odot O$, and points C and D are on $\odot O$. If $\angle ABD = 50^\circ$, then the degree of $\angle BCD$ is () Choices: (A) 30° (B) 35° (C) 40° (D) 45°



OneThinker-8B: <think>Got it, let's try to solve this problem step by step. First, let's recall some circle theorems. \n\nWe know that AB is the diameter of circle O. So, any angle inscribed in a semicircle is a right angle. That means $\angle ACB$ should be 90° because triangle ACB is inscribed in the semicircle with AB as the diameter. Now, let's look at $\angle ABD$. The problem says $\angle ABD$ is 50° . Let's think about quadrilateral or triangles here. Maybe we can use the inscribed angle theorem, which states that an inscribed angle is half the measure of its intercepted arc. First, let's consider arc AD or arc AB. Since AB is the diameter, the arc AB is 180° . $\angle ABD$ is an inscribed angle that intercepts arc AD. So $\angle ABD$ intercepts arc AD, so the measure of arc AD would be $2 \times \angle ABD = 2 \times 50^\circ = 100^\circ$? Wait, maybe not. Let's check another angle. Alternatively, $\angle BAD$ and $\angle BCD$ might be related. Let's see. $\angle BAD$ is an inscribed angle intercepting arc BD. $\angle BCD$ is also an inscribed angle intercepting arc BD. So they might be equal or supplementary?

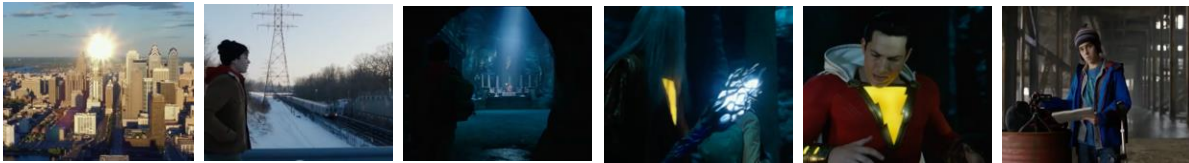
Wait, AB is the diameter, so $\angle ADB$ is also a right angle (since angle inscribed in a semicircle). So triangle ADB is a right-angled triangle at D. In triangle ADB, $\angle ABD$ is 50° , so $\angle BAD = 90^\circ - 50^\circ = 40^\circ$. \n\nNow, $\angle BCD$ and $\angle BAD$ are inscribed angles that intercept the same arc BD. Because $\angle BAD$ intercepts arc BD, and $\angle BCD$ also intercepts arc BD. So they should be equal. Therefore, $\angle BCD = \angle BAD = 40^\circ$. Let's verify again.\n\nAB is diameter, so $\angle ACB = 90^\circ$, $\angle ADB = 90^\circ$. In $\triangle ADB$, $\angle BAD = 90^\circ - 50^\circ = 40^\circ$. $\angle BCD$ and $\angle BAD$ both subtend arc BD, so they are equal. Hence, $\angle BCD$ is 40° , which is option C.</think>

<answer>C</answer>

Figure 5. Reasoning example of image question answering task.

Video QA

Question: In the context of the scenes described, what is the most likely narrative purpose of the glowing key (0:00:20-0:00:30) and its connection to the “SUPER STRENGTH TEST 1” sign (0:00:50-0:01:00), given the recurring appearances of superheroes and promotional content for a film platform? (A) The key unlocks a hidden villain base (B) The key is a mystical artifact that activates Shazam’s powers, and the test sign marks a public demonstration of his abilities to gain public trust (C) The key grants access to a magical realm where the superhero team must prove their worth via strength tests (D) The key is a McGuffin distracting from the “Don’t scream!” scene, while the test sign is unrelated promotional material for Filmisnow trailers



OneThinker-8B: <think>So, let's try to figure out this question. The options are about the glowing key and the “SUPER STRENGTH TEST 1” sign in the context of a superhero film, likely Shazam! since there's a mention of Shazam and the red suit superhero. First, let's recall the elements: the glowing key, the test sign, and the superhero (Shazam) context. Shazam's story involves a magical artifact (the Shazam stone) that grants powers, and tests of strength.....Key unlocks villain base, test is training. But the context is promotional for a film, so maybe not villain base..... C. Key grants access to magical realm for strength tests. The snowy industrial area with the test sign—maybe a test location. But the key's connection? D. McGuffin distracting, test sign unrelated. Unlikely, since it's promotional.

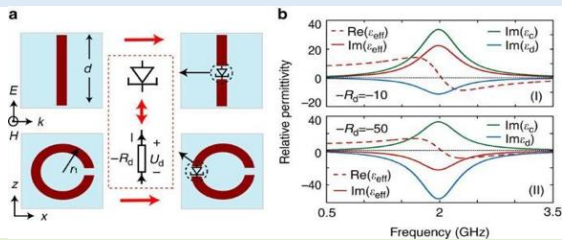
Wait, in Shazam, the Shazam stone is the key (glowing, magical). The “SUPER STRENGTH TEST” could be a test of his powers, maybe a public demonstration. But option C says “magical realm where team proves worth”. Alternatively, the key is the Shazam stone, and the test is a public test. But let's think again.\n\nThe video is a trailer for Shazam, so the glowing key might be the Shazam stone, and the test sign is a test of his strength. But option C: “key grants access to magical realm where superhero team proves worth via strength tests”. Maybe the key is a portal or artifact, and the test is part of that. Alternatively, the key is the stone, and the test is a public demo.\n\nWait, the correct answer is likely C..... The key (maybe the Shazam stone) is a magical artifact, and the test is a strength test in that realm.\n\nAlternatively, B: key activates Shazam’s powers, test sign public demo. But “public trust” might not be the main narrative purpose.\n\nAnother angle: Shazam's powers are activated by the stone, and tests of strength are part of his journey. The key (stone) is the source, and the test is a demonstration. But option C's “magical realm” might be the setting for the test.\n\nYes, C seems plausible. The key (maybe the stone) is a magical artifact, and the test is a strength test in a magical realm.</think>

<answer>C</answer>

Figure 6. Reasoning example of video question answering task.

Image Caption

Question: Give a description of this image.



OneThinker-8B: <think>So, let's try to describe this image. First, it's a scientific figure with two main parts, labeled (a) and (b)..... There are multiple curves: solid and dashed lines for real (Re) and imaginary (Im) parts of effective permittivity (ϵ_{eff}) and maybe another permittivity (ϵ_c , ϵ_d).....showing structural diagrams (a) and their electromagnetic properties (relative permittivity as a function of frequency) in (b), with parameters like R_d affecting the results.</think>

<answer>This image appears to be a scientific illustration, likely from a research paper in the field of physics or engineering, focusing on the study of metamaterials or resonators, particularly in the context of electromagnetic properties. The image is divided into two main sections, labeled (a) and (b).\n\n**Section (a):**\n- The left side shows two diagrams, one rectangular and one circular, both with a red-colored material, possibly representing a dielectric or resonant structure. The rectangular diagram has arrows indicating a dimension d and a wave vector k with an electric field E and magnetic field H direction, suggesting an electromagnetic wave interaction.\n- The right side of section (a) shows a schematic representation of an equivalent circuit model for these structures, with symbols indicating resistors and voltage sources, possibly representing the electrical behavior of the materials under study. There are dashed arrows indicating a transformation or correspondence between the physical structures and their circuit models.\n\n**Section (b):**\n- This section contains two graphs, labeled (I) and (II), plotting the relative permittivity against frequency in GHz.\n- Both graphs display curves for the real (Re) and imaginary (Im) parts of the effective permittivity (ϵ_{eff}) and another permittivity (ϵ_c), with different colors representing different values of a parameter, possibly a resistance R_d (with values -10 and -50 shown).\n- Graph (I) shows a higher variation in permittivity with frequency compared to graph (II), where the permittivity values are closer to zero, indicating a different response or material property at these specific resistance values.\n- The x-axis in both graphs is labeled "Frequency (GHz)" and the y-axis is "Relative permittivity," providing a quantitative analysis of how the permittivity of these materials changes with frequency under different conditions.\n\nOverall, the image seems to illustrate both the physical and circuit models of resonant structures and their corresponding electromagnetic properties, specifically focusing on how permittivity varies with frequency under different parameters, which is crucial for understanding and designing metamaterials or resonators for various applications in telecommunications, optics, or other fields requiring precise control over electromagnetic properties.</answer>

Figure 7. Reasoning example of image caption task.

Video Caption

Question: Give a description of this video.



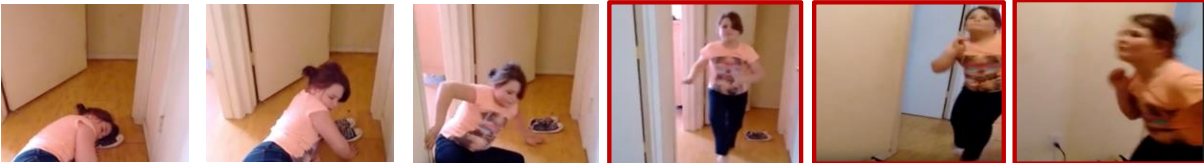
OneThinker-8B: <think>In a cozy, well-lit living room filled with bookshelves, a lamp, and a colorful couch, two individuals engage in an animated discussion. One person, seated on the couch, is focused on writing or taking notes in a notebook, wearing a vibrant, floral-patterned cardigan over a pink top. The other person.....Towards the end, the seated person appears to take notes or reflect, while the standing individual continues to speak, maintaining the educational or explanatory tone of the scene. The overall setting and interaction convey a friendly, intellectual exchange in a comfortable, lived-in space.</think>

<answer>The video takes place in a cozy, well-lit living room that exudes a warm, intellectual ambiance, characterized by a richly decorated space filled with bookshelves, a colorful striped couch, and a large whiteboard covered in scientific equations and diagrams. The room is bathed in a homely glow, enhanced by a lamp and a festive, beaded tree in the background, creating a setting that feels both scholarly and inviting. Two individuals are the focal points of the scene. One person, seated comfortably on the couch, is engrossed in writing or taking notes in a notebook, wearing a vibrant, floral-patterned cardigan over a pink top. Their focused demeanor suggests they are deeply engaged in the discussion or study session. The other person, standing near the whiteboard, is dressed in a blue shirt with colorful vertical stripes and yellow sleeves, and appears to be explaining or teaching, using a marker to point at and write on the whiteboard. Their animated gestures and active engagement indicate they are leading the discussion, possibly explaining complex scientific concepts. As the video progresses, the standing individual continues to elaborate on the whiteboard, gesturing towards various equations and diagrams with increasing emphasis, indicating a detailed and possibly intricate explanation. The seated person remains attentive, occasionally glancing up from their notebook with expressions of surprise or interest, reflecting their active participation and engagement in the session. The interaction is dynamic and educational, with the standing person frequently pointing and writing on the whiteboard, while the seated individual responds with focused listening and note-taking. The camera alternates between close-up shots of the individuals, capturing their expressions and actions in detail, and wider shots that provide a comprehensive view of the room and their interaction. Towards the end of the video, the seated person appears to be taking notes or reflecting, with a thoughtful expression, while the standing individual continues to explain, occasionally pausing and looking down, possibly to gather their thoughts. The overall atmosphere remains one of a lively, intellectual exchange, with both individuals deeply involved in a collaborative and educational session in a warm, inviting living room setting.</answer>

Figure 8. Reasoning example of video caption task.

Temporal Grounding

Question: Give the query: 'person runs back.', when does the described content occur in the video?



OneThinker-8B: <think>So, let's analyze the video. The query is 'person runs back.' Let's check each frame. Looking at the timestamps: The person was lying down, then gets up. Then, towards the end, the person runs towards the camera and then maybe back? Wait, the last part: the person runs towards the camera (maybe back direction). Let's see the movement. From the frames, around 15-22 seconds, the person gets up, walks, then runs. The action of running back—when does that happen? Let's check: The person starts lying on floor, then moves. Then, when they run towards the camera and then maybe back? Wait, the video shows the person running towards the camera (maybe towards the end). Let's check the time span.

Looking at the sequence: The person wakes up, moves, then runs. The running action towards the camera and then maybe back? The last part where the person runs towards the camera and then exits. The time when the person runs back—probably around 15-22 seconds. Let's confirm: The video shows the person lying down, then getting up, moving, then running. The action of running back would be when they run towards the camera and then maybe back. The correct time span for 'person runs back'—let's see the movement.

Yes, the person runs towards the camera (maybe back direction) in the later part. So the time span is around 15 to 22 seconds. Let's check:\n\n15-22 seconds: The person is moving quickly, running. So the time when the person runs back is in that range.</think>

<answer>{"time": [15.0, 22.0]}</answer>

Figure 9. Reasoning example of temporal grounding task.

Spatial Grounding

Question: Please provide the bounding box coordinate of the region this sentence describes: a bush of plant behind middle woman.



OneThinker-8B: <think>To find the bush behind the middle woman, first identify the middle woman: she is the one in the center, wearing a black shirt. Behind her is a large green bush. This bush sits in the background, positioned behind her as she is seated.</think>

<answer>{"boxes": [442, 56, 664, 357]}</answer>

Figure 10. Reasoning example of spatial grounding task.

Spatial-Temporal Grounding

Question: At which time interval in the video can we see an adult wearing blue pants plays the guitar occurring? When and where does the described content occur in the video? Please firstly give the start and end time, spatial bounding box corresponding to each integer second.



OneThinker-8B: <think>To find when an adult in blue pants plays the guitar, observe the video: the adult in blue pants playing guitar appears during a segment where the camera shifts focus. This occurs from around 5 seconds to 19 seconds. During this interval, the adult is seated on the right side of the frame, engaged in playing the guitar.</think>

```
<answer>{"time": [5.015, 19.511], "boxes": {"6": [670, 277, 998, 998], "7": [647, 281, 998, 998], "8": [616, 310, 998, 998], "9": [609, 327, 998, 998], "10": [609, 333, 998, 998], "11": [606, 333, 909, 998], "12": [600, 333, 889, 998], "13": [589, 333, 883, 998], "14": [583, 333, 877, 998], "15": [580, 333, 877, 998], "16": [577, 333, 877, 998], "17": [577, 333, 877, 998], "18": [577, 333, 877, 998], "19": [577, 333, 877, 998]}}
```

Figure 11. Reasoning example of spatial-temporal grounding task.

Tracking

Question: Given the bounding box [537,403,768,703] of the target object in the first frame, track this object in each frame and output its bounding box once per second.



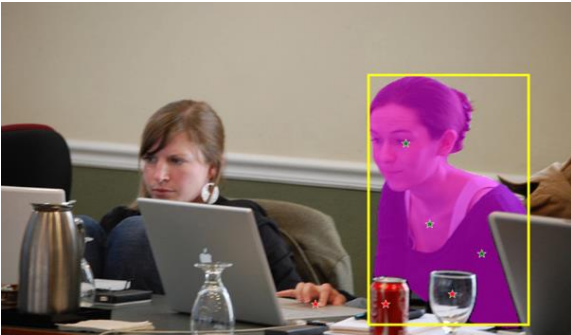
OneThinker-8B: <think>To track the white Ford Mustang convertible across each second up to 32 seconds, observe its movement pattern: the car travels forward along the road, so its position in the frame shifts subtly over time. Initially, the car occupies a central area of the frame. As seconds pass, the car moves closer to the camera, causing its bounding box to adjust incrementally. Over each second, the car's position shifts slightly toward the lower portion of the frame (as it approaches the camera) and its bounding box expands a bit to account for the closer perspective. Throughout the sequence, the car remains the central focus, with smooth, gradual changes in its placement and the size of the bounding box reflecting its forward motion and proximity to the camera.</think>

<answer>{"boxes": {"1": [537, 403, 768, 703], "2": [537, 353, 797, 690], "3": [532, 323, 827, 687], "4": [519, 282, 850, 690], "5": [515, 260, 877, 694], "6": [510, 246, 887, 694], "7": [509, 246, 897, 703], "8": [509, 232, 947, 710], "9": [509, 210, 972, 732], "10": [482, 200, 989, 787], "11": [469, 196, 995, 856], "12": [459, 203, 995, 929], "13": [422, 221, 995, 987], "14": [397, 244, 995, 994], "15": [389, 250, 995, 994], "16": [382, 250, 995, 994], "17": [370, 250, 995, 994], "18": [364, 250, 995, 994], "19": [362, 250, 995, 994], "20": [354, 244, 995, 994], "21": [350, 239, 995, 994], "22": [347, 239, 995, 994], "23": [347, 244, 965, 979], "24": [350, 244, 945, 944], "25": [350, 244, 904, 910], "26": [350, 244, 845, 860], "27": [347, 244, 797, 779], "28": [347, 244, 772, 750], "29": [350, 250, 764, 742], "30": [350, 250, 757, 739], "31": [350, 250, 750, 739], "32": [350, 250, 750, 739]}}

Figure 12. Reasoning example of tracking task.

Image Segmentation

Question: Given the referring expression "a woman touching a laptop wearing a white tank top under a black shirt", provide one bounding box, along with positive points located inside the object and negative points located outside, as segmentation hints.



OneThinker-8B: <think>Got it, let's tackle this task. First, identify the target object: \"a woman touching a laptop wearing a white tank top under a black shirt\". Let's look at the two women. The right woman has a black shirt with a white tank top underneath, and she's touching a laptop. So that's the target. \n\nNow the bounding box. Let's estimate coordinates. Let's see, the woman on the right. Let's set x1, y1, x2, y2. Let's check the image. She's on the right, so box around her. Let's approximate: x1 around 640, y1 around 278, x2 around 914, y2 around 957. Need to adjust.....The box around her would include her body, but maybe the table? Wait, she's sitting, so inside the box (her bounding box) might have parts like the table? No, the box is around her. Alternatively, if inside the box, non-object areas. But she's a person, so maybe the laptop she's touching? No, the laptop is part of the scene but she's the target. Alternatively, inside the box but not on her. Let's see. \n\nSo boxes: Let's define the box for the right woman. Let's say [642, 276, 917, 956] (adjusting). Then positive points: three points on her body. Like (700,450) [face], (754,650) [chest], (820,750) [arm]. Negative points: inside the box but not on her. Maybe on the laptop she's using? But the laptop is in front. Alternatively, the table items inside the box? Maybe the can or glass?Alternatively, adjust. \n\nLet's finalize: \n\nBox: Let's set x1=640, y1=275, x2=915, y2=955. \n\nPositive points: three points on her body. Like (704,460) (face), (745,680) (chest area), (834,760) (arm). \n\nNegative points: inside the box but not on her. Maybe on the laptop she's using? But the laptop is in front. Alternatively, the table items inside the box. For example, the red can, glass, etc. But need to be off the woman. So points on the table inside the box but not on her. Let's pick (670,900) (red can), (784,870) (glass), (550,900) (maybe a mouse? But need to check). Alternatively, inside the box but on background. Let's confirm.</think>

```
<answer>{"boxes": [640, 275, 915, 955], "positive_points": [[704, 460], [745, 680], [834, 760]], "negative_points": [[670, 900], [784, 870], [550, 900]]}</answer>
```

Figure 13. Reasoning example for an image segmentation task. The resulting answer will be forwarded to SAM2 to produce the mask.

Video Segmentation

Question: Given the referring expression “Which creature in the footage poses the greatest threat to an unwanted visitor encroaching on its domain?”, Provide the key frame time along with one bounding box, as well as positive points inside the object and negative points outside, to serve as segmentation hints.



OneThinker-8B: <think>So, let's tackle this task. The question is about identifying the creature that poses the greatest threat to an unwanted visitor in its domain. In the video, elephants are the main subjects, and the largest elephant would be the most dominant, so that's the target. Let's pick a time frame. Let's choose 1.0 second as a representative time. First, the bounding box for the large elephant. Let's look at the frame around 1.0s. The big elephant is the main one in the foreground. Let's estimate coordinates. Let's say the box is around the large elephant. Let's set x_1 , y_1 , x_2 , y_2 . Let's approximate: x_1 around 180, y_1 around 15, x_2 around 826, y_2 around 940 (adjusting for the video's aspect).

Positive points: inside the elephant. So pick three points on the body, like torso, head, leg. Negative points: need to be outside the object. If inside the box, maybe background. But the elephant is solid, so maybe inside the box but on the ground? Wait, the elephant's body is the object. So negatives could be on the ground inside the box but not on the elephant. Or outside.

Let's confirm: Time chosen: 1.0 seconds. Box around the large elephant. Let's set boxes as [175, 10, 828, 945] (adjusting). Positive points: three points inside the elephant. For example, (350, 400), (500, 300), (650, 500) — these are on the body. Negative points: three points outside. Maybe on the ground inside the box but not on the elephant? Or outside. Let's pick inside the box but on the ground. Like (200, 850) (on the ground near the elephant's leg), (750, 850) (near the leg), and (400, 900) (ground). These are outside the elephant's body.

Wait, need to ensure they are strictly off the object. So yes, on the ground inside the box but not on the elephant. That works.</think>

```
<answer>{"time": 1.0, "boxes": [175, 10, 828, 945], "positive_points": [[350, 400], [500, 300], [650, 500]], "negative_points": [[200, 850], [750, 850], [400, 900]]}</answer>
```

Figure 14. Reasoning example for an video segmentation task. The resulting answer will be forwarded to SAM2 to produce the mask.

B. Prompt Template

System Prompt

Please answer this question based on the visual content. Provide your thinking process between the `<think>` and `</think>` tags, and then give your final answer between the `<answer>` and `</answer>` tags. At the end, you must output the final answer in the format: `<answer><your_answer_here></answer>`

Figure 15. System prompt for all tasks.

Prompt for QA

"multiple choice": (
"Please provide only the single option letter (e.g., A, B, C, D, etc.) "
"within the `<answer>...</answer>` tags.\n"
"Example:\n`<answer>A</answer>`"
)

"numerical": (
"Please provide only the numerical value within the `<answer>...</answer>` tags.\n"
"Example:\n`<answer>3.14</answer>`"
)

"OCR": (
"Please provide only the transcribed text within the `<answer>...</answer>` tags.\n"
"Example:\n`<answer>Hello World</answer>`"
)

"open-ended": (
"Please provide only your text answer within the `<answer>...</answer>` tags.\n"
"Example:\n`<answer>The capital of France is Paris.</answer>`"
)

"regression": (
"Please provide only the numerical value within the `<answer>...</answer>` tags.\n"
"Example:\n`<answer>42.7</answer>`"
)

"math": (
"Please provide only the final result (a number or LaTeX formula) within the `<answer>...</answer>` tags.\n"
"Example:\n`<answer>$$-\dfrac{3}{2}$$</answer>`"
)

Figure 16. Prompt for QA tasks.

Prompt for Grounding and Tracking

```
"temporal grounding": (  
"Please provide only the time span in seconds as JSON within the <answer>...</answer>  
<tags> tags. Example:\n<answer>{\n\"time\": [12.3, 25.7]}</answer>  
)  
  
"spatial grounding": (  
"Please provide only the bounding box as JSON with key 'boxes' within the  
<answer>...</answer> tags.\n"  
"Example:\n<answer>{\n\"boxes\": [35, 227, 437, 932]}</answer>  
)  
  
"spatial-temporal grounding": (  
"Please provide only the time span in seconds and bounding boxes as JSON within the  
<answer>...</answer> tags. You MUST output one bounding box for every integer  
second within the given time span (inclusive).\n"  
"Example: <answer>{\n\"time\": [8.125, 13.483], \n\"boxes\": {\n\"9\": [317, 422, 582, 997], "  
"10\": [332, 175, 442, 369], \n\"11\": [340, 180, 450, 370]}}</answer>\n"  
"Note: Each key in 'boxes' must be an integer second within the span, and its value must  
be a 4-number bounding box [x1, y1, x2, y2]."  
)  
  
"tracking": (  
"Please track the target object throughout the video and provide one bounding box per  
second, within the <answer>...</answer> tags.\n"  
"Example: <answer>{\n\"boxes\": {\n\"1\": [405, 230, 654, 463], \n\"2\": [435, 223, 678,  
446], ..., \n\"32\": [415, 203, 691, 487]}}</answer>\n"  
)
```

Figure 17. Prompt for grounding and tracking tasks.

Prompt for Segmentation

```
"image_segmentation": (  
"This task prepares inputs for image object segmentation with a specialized model (e.g.,  
SAM2).\n"  
"Please provide ONE bounding box, 3 positive points (clearly INSIDE the object), and 3  
negative points (clearly OUTSIDE the object) within the <answer>...</answer> tags.\n"  
"Choose informative points that help distinguish object vs. background. Prefer negatives  
on clear non-object pixels INSIDE the box when safe; otherwise place them just outside  
on obvious background. Negatives must NEVER be on the object or on its boundary.\n"  
"Example:\n<answer>{\n\"boxes\": [x1, y1, x2, y2], \n\"positive_points\": [[x,y],[x,y],[x,y]],  
"negative_points\": [[x,y],[x,y],[x,y]]}</answer>  
)  
  
"video_segmentation": (  
"This task prepares inputs for video object segmentation with a specialized model (e.g.,  
SAM2).\n"  
"Please select ONE representative time (in seconds), and provide ONE bounding box, "  
"3 positive points (clearly INSIDE the object), and 3 negative points (clearly OUTSIDE  
the object) within the <answer>...</answer> tags.\n"  
"Choose informative points that help distinguish object vs. background. Prefer negatives  
on clear non-object pixels INSIDE the box when safe; otherwise place them just outside  
on obvious background. Negatives must NEVER be on the object or on its boundary.\n"  
"Example:\n<answer>{\n\"time\": <time_in_seconds>, \n\"boxes\": [x1, y1, x2, y2], "  
"positive_points\": [[x,y],[x,y],[x,y]], \n\"negative_points\": [[x,y],[x,y],[x,y]]}</answer>  
)
```

Figure 18. Prompt for segmentation tasks.