

Discriminative Perception via Anchored Description for Reasoning Segmentation

Supplementary Material

Query: When taking a dog for a walk on the beach, what part of the dog's body should be wet if it enjoys playing in the water?



Mask IoU=0.19 Baseline: 146 Tokens



Think: The image shows a dog on a beach with its paws in the water. The dog appears to be enjoying itself. The question asks what part of the dog's body should be wet if it enjoys playing in the water. Given the context, the paws of the dog are the part that would be most likely to be wet, as they are in direct contact with the water. The dog's paws are the part of the body that would be wet if it enjoys playing in the water. The paws are the only part of the dog's body that are in the water, and they are the part that would get wet from playing in the water. The answer should be the paws of the dog, as they are the part of the body that would be wet if the dog enjoys playing in the water.



Mask IoU=0.61 DPAD(Ours): 56 Tokens



Think: The thinking process involves identifying the part of the dog's body that would be wet if it enjoys playing in the water. The dog in the image is standing in shallow water, and the part of the dog's body that would be wet is the lower legs and paws.



Caption: The lower legs and paws of the dog.

Query: When a cat is interested in something and wants to observe it carefully, what part of the cat's body in the picture is most likely to be the focus of attention?



Mask IoU=0.25 Baseline: 160 Tokens



Think: The image shows a cat lying on a wooden floor. The cat's body is relaxed, and it is looking directly at the camera with a focused expression. The cat's ears are perked up, which is a common sign of interest. The eyes are wide open, and the pupils are dilated, indicating that the cat is paying close attention to something. In the context of the question, when a cat is interested in something and wants to observe it carefully, the most likely part of the cat's body to be the focus of attention would be the eyes. The eyes are the most expressive part of the cat's face and are the first thing that catches attention when a cat is observing something. In the image, the cat's eyes are clearly visible and appear to be focused on something in front of it. Therefore, the eyes are the most likely part of the cat's body to be the focus of attention.



Mask IoU=0.78 DPAD(Ours): 85 Tokens



Think: The thinking process involves identifying the part of the cat's body that is most likely to be the focus of attention when the cat is interested in something. Typically, when a cat is interested, it will look directly at the object of interest. In the picture, the cat is looking directly at the camera, which suggests it is focused on something in front of it. The part of the body that is most likely to be the focus of attention is the eyes. Caption: The cat's eyes.



Query: The spare tire that is not used currently.



Mask IoU=0.00 Baseline: 135 Tokens



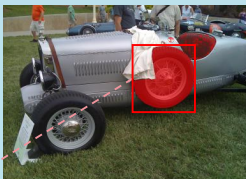
Think: The image shows a vintage race car on display with a spare tire mounted on the side. The car is parked on a grassy area, and there are people in the background observing it. The spare tire is located on the side of the car, and it is clearly visible in the image. The question asks to find the spare tire that is not used currently and compare the difference between objects. The spare tire is the object in question, and it is not currently in use. The difference between the spare tire and the car is that the spare tire is not part of the car's main structure and is not in contact with the ground, while the car is a functional vehicle with a complete structure and is in contact with the ground.



Mask IoU=0.90 DPAD(Ours): 57 Tokens



Think: The thinking process involves identifying the spare tire that is not currently in use. The spare tire is located at the rear of the vehicle, and in this image, it is visible on the right side of the car. The tire is not inflated and is covered with a protective cover, indicating it is not in use.



Caption: The spare tire that is not used currently.

Query: Girl on left no sleeves.



Mask IoU=0.00 Baseline: 133 Tokens



Think: The image shows a group of people sitting on the grass, with some wearing distinctive clothing and accessories. The person in the foreground, wearing a black tank top with text on the back, appears to be the "girl on left no sleeves". The tank top has a visible text that reads "WIR KÖNNEN UNS NICHT EINRICHTEN DEN WIR SIND IM TEIL VON SUCH!". The task is to find the difference between objects and compare the most closely matched one. The objects in the image are the people, and the most closely matched object in this context is the person wearing the black tank top with the text. The answer should be in JSON format, with the bbox of the person and the points of the two largest inscribed circles inside the interested object.



Mask IoU=0.78 DPAD(Ours): 53 Tokens

Think: The thinking process involves identifying the object of interest, which is 'girl on left no sleeves'. The object is the person on the left side of the image, wearing a sleeveless denim vest and a necklace. The focus is on the person's upper body and the sleeveless vest.
Caption: The girl on left no sleeves.



Query: In a baseball game, the pitcher throws the ball to the batter, who tries to hit it. what equipment does the batter use to hit the ball?



Mask IoU=0.00 Baseline: 83 Tokens

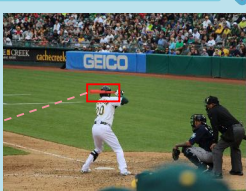


Think: The image shows a baseball game with a batter, a catcher, and an umpire. The batter is holding a baseball bat, which is the equipment used to hit the ball. The thinking process involves identifying the object that the batter uses to hit the ball, which is the baseball bat. The objects in the image are the batter, the baseball bat, the catcher, the umpire, and the baseball field. The baseball bat is the object that the batter uses to hit the ball.



Mask IoU=0.90 DPAD(Ours): 41 Tokens

Think: The thinking process involves identifying the object the batter uses to hit the ball. The object in question is the bat, which the batter holds and swings to hit the ball. The bat is the primary equipment used by the batter.
Caption: The bat used by the batter to hit the ball.



Query: Girl kicking.



Mask IoU=0.00 Baseline: 128 Tokens



Think: The image shows a group of children playing soccer on a grassy field. The focus is on a girl in a blue jersey who appears to be kicking a soccer ball. The other children are running, and there are adults in the background. The girl in the blue jersey is the one kicking the ball, and the other children are running towards the ball. The objects in the image include the soccer ball, the children's soccer uniforms, the grassy field, and the adults in the background. The girl in the blue jersey is the one kicking the ball, and the other children are running towards the ball. The most closely matched object is the girl in the blue jersey, as she is the one actively kicking the ball.



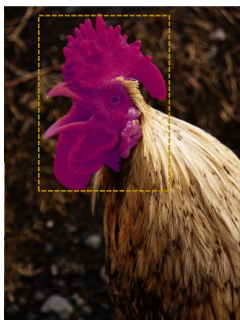
Mask IoU=0.96 DPAD(Ours): 55 Tokens

Think: The thinking process involves identifying the object of interest, which is the girl kicking the soccer ball. I look for the person who is actively kicking the ball, which is the girl in the white jersey. The other players and the coach are not the focus.
Caption: The girl in the white jersey kicking the soccer ball.

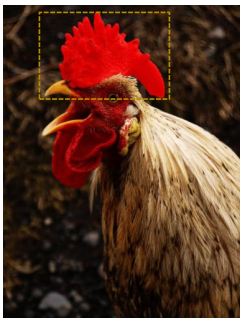


Figure 6. Additional qualitative comparison. Masks: GT (green), DPAD (red), and baseline (purple).

Query: In the picture, there is a type of bird with a distinctive feature on the top of its head, which usually indicates its gender, what part in the picture might have this characteristic?



<think>The image shows a close-up of a rooster, which is a type of bird. The distinctive feature on the top of its head is the comb, which is a red, fleshy structure. This comb is a characteristic feature of roosters and is used to indicate their gender. The rooster in the image has a red comb. The comb is located on the top of the rooster's head, above the beak and eyes. The rooster's comb is the part of the bird that has the distinctive feature, which is the red fleshy structure on the top of its head. </think><answer>{"bbox": [204, 21, 786, 258], "points_1": [289, 159], "points_2": [642, 225]}</answer><caption>the rooster's red head </caption> IoU=0.40
Think Format=1 Seg Format=1 IoU=0 Bbox=0 Points=1 DPAD=0 Final=3



<think> The thinking process involves identifying the distinctive feature on the top of the bird's head, which is typically a comb and wattle in roosters. The comb is the red, fleshy structure on top of the bird's head, and the wattle is the hanging flap of skin below the beak. In the picture, the red, fleshy structure on top of the bird's head is the comb. </think><answer>{"bbox": [202, 19, 572, 250], "points_1": [330, 130], "points_2": [420, 200]}</answer><caption>the red comb on the top of the rooster's head</caption> IoU=0.96
Think Format=1 Seg Format=1 IoU=1 Bbox=1 Points=1 DPAD=1 Final=6

Figure 7. Qualitative demonstration of spatial-semantic alignment evolution via discriminative perception.

6. Additional Qualitative Analysis

It can be observed that models lacking the “Discriminative Perception” reward often generate verbose and divergent reasoning chains, as detailed in Figure 6. Specifically, DPAD significantly reduces reasoning redundancy across the following three core dimensions:

Pruning Divergent Global Scenes. In the baseline method, the model frequently allocates excessive attention to global background descriptions irrelevant to the query. For instance, in the “spare tire” and “girl kicking” examples, the baseline model consumes a large number of tokens describing the “vintage race car... parked on a grassy area,” “people in the background,” or “adults in the background”. In contrast, DPAD, through the contrastive reward mechanism of anchored descriptions, effectively guides the model to disregard these divergent global details and anchor its reasoning directly on the target object (e.g., clearly locating the uninflated spare tire on the side of the car). This achieves accurate segmentation using significantly fewer tokens (e.g., reducing the count from 135 to 57).

Filtering Irrelevant Objects and Over-identification. When multiple objects or complex semantics are present in a scene, the baseline model is highly prone to enumerative behavior or deviating from the core task. In the “baseball game” example, the baseline not only describes the batter

but also redundantly lists the catcher, the umpire, and the baseball field. This is particularly evident in the “girl on left no sleeves” example, where the baseline diverges from the visual grounding objective to recognize and transcribe a large block of German text (“WIR KÖNNEN UNS NICHT EINRICHTEN...”) on the person’s clothing. DPAD effectively filters out these distractors and precisely converges on the “bat” and the “sleeveless denim vest,” which not only reduces the token count by more than half but also substantially improves the mask IoU from 0.00 to high-precision levels (0.90 and 0.78, respectively).

Reducing Over-reasoning and Self-confirmations. In the examples of the “dog’s paws” and “cat’s eyes,” the baseline exhibits a pronounced tendency toward “self-confirmation.” The model repeatedly restates the problem context and incorporates superfluous trivial details (such as repeatedly emphasizing that the cat’s “ears are perked up” and “pupils are dilated,” or that the dog “appears to be enjoying itself”). DPAD, on the other hand, presents a highly “Short & Focused” reasoning paradigm, addressing the core attributes of the target directly and concisely outputting the “lower legs and paws” or the “cat’s eyes.”

These additional visualization results provide compelling evidence that, by incentivizing target-relevant descriptions, DPAD structurally prunes divergent scenes, irrelevant objects, and uninformative self-confirmations. This

not only substantially compresses the length of the reasoning chains (mitigating the redundancy issues that degrade reasoning efficiency) but, more importantly, provides highly precise localization cues for downstream mask generation by eliminating contextual noise.

7. Evolution of DPAD Alignment

In the evolution analysis of the convergence process, we observe how the DPAD reward mechanism calibrates spatial outputs through differentiated feedback on intermediate states (as shown in Figure 7). During the under-converged stages of training, the model frequently exhibits a disconnect between semantics and geometry. At this stage, if the model relies solely on visual priors to yield a coarse bounding box covering most of the rooster’s head, but generates an over-generalized or inaccurate description (e.g., “the rooster’s red head”), our mechanism assigns a reward of $R = 0$ to suppress such blind localization lacking rigorous logical support. When the model demonstrates correct anatomical discrimination during its reasoning process (e.g., explicitly distinguishing between the “comb on top of the head” and the “wattle under the beak”) and generates the precise description “the red comb on the top of the rooster’s head”, the mechanism provides positive feedback of $R = 1$. Continuously driven by this mechanism, the model is able to correct the initially ambiguous boundaries, precisely contracting and locking the bounding box onto the comb region, raising the IoU to 0.96.