
Dual-Stage Value-Guided Inference with Margin-Based Reward Adjustment for Fast and Faithful VLM Captioning Supplementary

Anonymous Author(s)

Affiliation

Address

email

1 Extended Qualitative Results

2 This section provides additional qualitative examples to supplement those presented in main pa-
3 per and appendix of the main paper. We include a broader set of visualization cases comparing
4 ViMaR-guided decoding with baseline methods across varied scenes and prompt types. Figures 1
5 through 14 showcase a series of challenging and representative examples that require fine-grained
6 visual understanding and precise language generation. These extended examples further highlight
7 ViMaR’s ability to produce more visually grounded, detailed, and hallucination-resistant captions.
8 Each figure illustrates both the model outputs and the corresponding attention maps to facilitate
9 deeper analysis of decoding behavior.

What's happening in the scene?



Attention Map after whole sentence generated



ViMaR Guided

In the scene, there is a clock tower with two clocks visible on its facade. The sky is overcast with dark clouds, suggesting it might be a cloudy or stormy day. In the foreground, there are trees and some buildings, indicating that the clock tower is part of a town or city center. There are no people clearly visible in the immediate vicinity of the tower, but there are a few cars parked or moving in the background, indicating some level of activity in the area. The setting appears to be a public space, possibly a park or a town square, given the presence of the clock tower and the benches visible in the foreground. The overall mood of the image is somewhat subdued due to the overcast sky and the absence of people in the immediate vicinity of the tower. The image captures a quiet moment in what might be a bustling area during other times of the day. The presence of the flag on top of the tower adds a touch of national pride or identity to the scene. There are power lines running across the image. The overall composition of the image, with the clock tower as the focal point, draws the viewer's eye upward, emphasizing the height and prominence of the tower in the landscape. The image is a snapshot of a moment in time, capturing the stillness of the scene despite the potential for activity in the background.

VisVM Guided

In the scene, there is a clock tower standing in the foreground with its face clearly visible. The sky appears to be cloudy, suggesting either an overcast day or a storm approaching. In the background, there are buildings, including a prominent structure with a green roof, and a street with cars and a truck parked along the side. The presence of a flag on top of the tower indicates that this might be a public or municipal building, possibly a town hall or a similar institution. The overall atmosphere of the image is somewhat somber due to the cloudy sky and the absence of people in the immediate vicinity of the tower.

CLIP-PRM Guided

The image shows a clock tower with a blue flag at the top, indicating it might be a public or government building. The sky is overcast, suggesting it might be a cloudy day. Below the tower, there's a street scene with shops and a car, indicating this is likely a town or city center. The flag flying at half-mast could signify a period of mourning or a special event that requires the flag to be lowered.

Greedy Decoding

In the scene, there is a clock tower with two clocks visible on its sides. The tower appears to be a prominent landmark, possibly in a town or city center. The sky is overcast, suggesting it might be a cloudy day. There are power lines running across the image, indicating that the tower is located in an area with electrical infrastructure. In the foreground, there are some trees and what looks like a public space with benches, which could be a park or a similar recreational area. There are also some buildings in the background, which might be part of a commercial or residential area. The overall atmosphere of the image is calm and quiet, with no visible activity that would indicate a specific event or time of day.

Figure 1: Case study 1: ViMaR and baseline captions.

What's happening in the scene?



Attention Map after whole sentence generated



ViMaR Guided

In the scene, there is a clock tower with two clocks visible on its facade. The sky is overcast with dark clouds, suggesting it might be a cloudy or stormy day. In the foreground, there are trees and some buildings, indicating that the clock tower is part of a town or city center. There are no people clearly visible in the immediate vicinity of the tower, but there are a few cars parked or moving in the background, indicating some level of activity in the area. The setting appears to be a public space, possibly a park or a town square, given the presence of the clock tower and the benches visible in the foreground. The overall mood of the image is somewhat subdued due to the overcast sky and the absence of people in the immediate vicinity of the tower. The image captures a quiet moment in what might be a bustling area during other times of the day. The presence of the flag on top of the tower adds a touch of national pride or identity to the scene. There are power lines running across the image. The overall composition of the image, with the clock tower as the focal point, draws the viewer's eye upward, emphasizing the height and prominence of the tower in the landscape. The image is a snapshot of a moment in time, capturing the stillness of the scene despite the potential for activity in the background.

VisVM Guided

In the scene, there is a clock tower standing in the foreground with its face clearly visible. The sky appears to be cloudy, suggesting either an overcast day or a storm approaching. In the background, there are buildings, including a prominent structure with a green roof, and a street with cars and a truck parked along the side. The presence of a flag on top of the tower indicates that this might be a public or municipal building, possibly a town hall or a similar institution. The overall atmosphere of the image is somewhat somber due to the cloudy sky and the absence of people in the immediate vicinity of the tower.

CLIP-PRM Guided

The image shows a clock tower with a blue flag at the top, indicating it might be a public or government building. The sky is overcast, suggesting it might be a cloudy day. Below the tower, there's a street scene with shops and a car, indicating this is likely a town or city center. The flag flying at half-mast could signify a period of mourning or a special event that requires the flag to be lowered.

Greedy Decoding

In the scene, there is a clock tower with two clocks visible on its sides. The tower appears to be a prominent landmark, possibly in a town or city center. The sky is overcast, suggesting it might be a cloudy day. There are power lines running across the image, indicating that the tower is located in an area with electrical infrastructure. In the foreground, there are some trees and what looks like a public space with benches, which could be a park or a similar recreational area. There are also some buildings in the background, which might be part of a commercial or residential area. The overall atmosphere of the image is calm and quiet, with no visible activity that would indicate a specific event or time of day.

Figure 2: Case study 2: ViMaR and baseline captions.

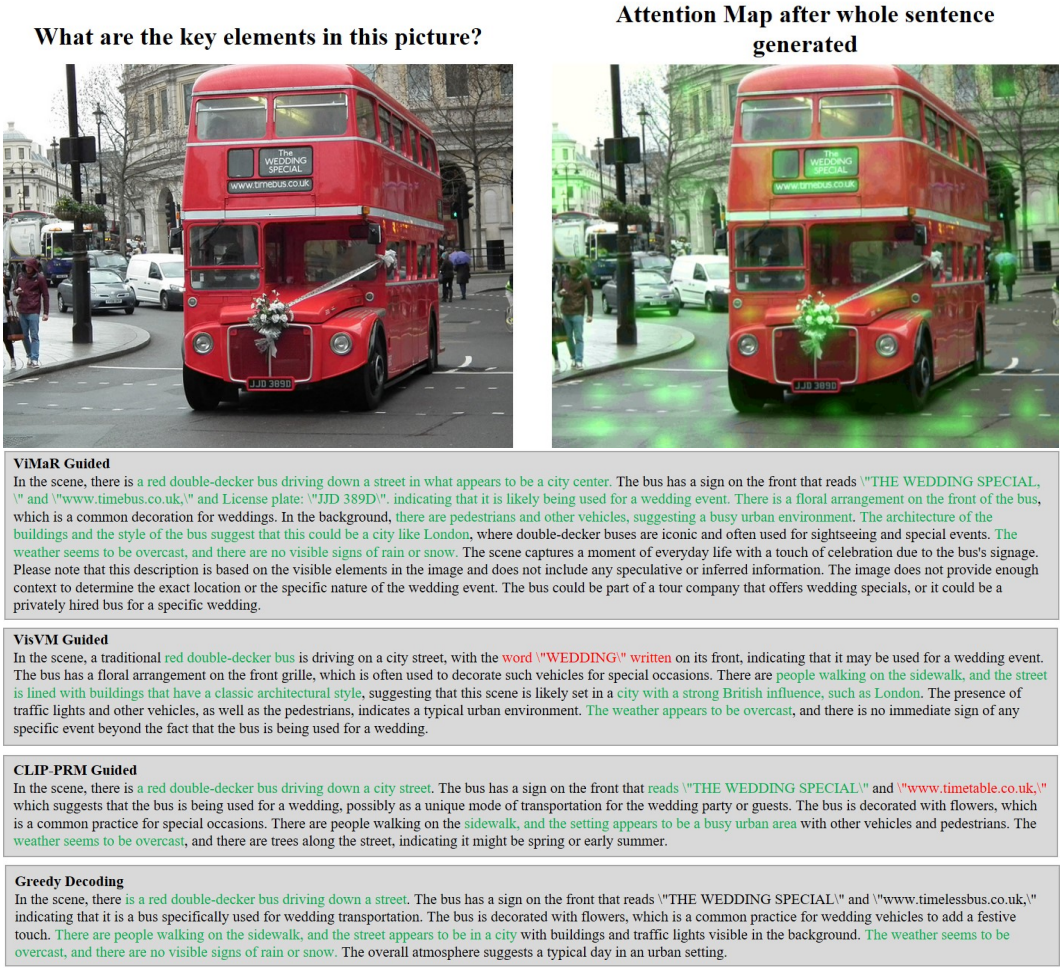


Figure 3: Case study 3: ViMaR and baseline captions.

Describe the following image.



Attention Map after whole sentence generated



ViMaR Guided

The image depicts a serene scene of a group of people on a boat in a body of water, which appears to be a river or a lake. The boat is a long, narrow vessel, with a pointed stern, and it's being poled through the water by an individual standing in the rear. There are several people seated in the boat, all facing the same direction, which is towards the left side of the image, suggesting they are moving upstream or towards the left bank. The boat is positioned in the foreground, with the people onboard appearing to be engaged in a leisurely activity, possibly sightseeing or enjoying a tour. They are dressed casually, and one person at the back of the boat is standing and wearing a cap, suggesting it might be a sunny day. In the background, there is a line of traditional huts with thatched roofs, which are partially obscured by the lush greenery of trees and foliage. The huts are constructed on stilts, indicating that the location may be prone to flooding or in a region where such construction is common for protection against rising water levels. The water has gentle ripples, reflecting the greenery and the huts, and the overall color palette of the image is dominated by shades of green, with the boat and the people providing a contrast with their earth tones. The image has a watermark or a filter applied to it, giving it a vintage or aged appearance, with a slight sepia tone that enhances the natural colors and adds a nostalgic feel to the scene. There is no visible text in the image, and the style of the image suggests it could be a photograph taken by a tourist or a traveler, capturing a moment during their visit to a cultural or natural site. The composition of the image, with the boat in the foreground and the huts in the background, leads the viewer's eye from the boat and its occupants to the tranquil and picturesque setting beyond them.

VisVM Guided

The image depicts a tranquil scene on a body of water, which appears to be a lake or a slow-moving river. In the foreground, there is a wooden boat with several people on board. They seem to be enjoying a leisurely ride, with one person standing at the front of the boat, possibly steering or navigating. The boat is painted in a reddish-brown hue and has a traditional design, suggesting it might be used for transportation or tourism in a region where such boats are common. In the background, there are several thatched-roof huts or bungalows situated along the water's edge. These structures have a rustic appearance and are likely part of a local community or a tourist facility. The surrounding vegetation is lush and green, indicating a tropical or subtropical climate. The water is calm, reflecting the greenery and the huts, which adds to the serene atmosphere of the scene. There is no visible text or branding in the image, and the style of the photograph is naturalistic, capturing the scene as it is without any apparent filters or artistic manipulations. The overall mood of the image is peaceful and idyllic, evoking a sense of relaxation and escape from the hustle and bustle of everyday life.

CLIP-PRM Guided

In the image, a small wooden boat is gently floating on a calm, greenish-brown river. The boat, with its pointed front and slightly curved rear, is filled with people who are enjoying a day out on the water. The passengers are dressed in casual clothing, some donning hats for shade. They are sitting close to each other, suggesting a sense of camaraderie and shared enjoyment. The river itself is nestled amidst lush greenery, with several huts or hut-like structures visible on the shore. These structures, likely made of bamboo, have thatched roofs, blending harmoniously with the natural surroundings. The vegetation appears dense, creating a serene and secluded environment. The boat is moving away from the camera, giving us a view of its rear as it ventures further down the river. The perspective from the shore looking towards the boat creates a sense of depth and distance in the image. The overall scene is one of leisure and tranquility, capturing a moment of relaxation and enjoyment in a natural setting.

Greedy Decoding

The image depicts a serene scene of a group of people on a boat in a body of water. The boat is a traditional design, possibly a canoe or a small motorboat, and it is occupied by several individuals who appear to be enjoying a leisurely outing. The boat is moving away from the shore, which is lined with thatched-roof huts, suggesting a tropical or subtropical location. The water is calm, reflecting the greenery of the surrounding vegetation, which includes trees and shrubs. The sky is overcast, casting a soft light over the scene. The overall atmosphere of the image is peaceful and idyllic, capturing a moment of relaxation and enjoyment in a natural setting.

Figure 4: Case study 4: ViMaR and baseline captions.

What is this photo about?

Attention Map after whole sentence generated



ViMaR Guided

The photo shows two young sheep, also known as lambs, in a grassy field. The lamb on the left appears to be black, while the lamb on the right is predominantly white with some black markings. In the background, there are other sheep, ducks, and big trees, suggesting this is a pastoral setting, possibly a farm or a rural area where sheep are raised. The lambs seem to be grazing or exploring the field, which is typical behavior for young sheep as they learn to navigate their environment. The presence of other sheep in the background indicates that this is part of a larger flock. The photo captures a peaceful and natural moment in the lives of these young animals. The weather appears to be overcast, and the grass is lush and green, indicating it might be spring or early summer when trees are not fully leafed. The overall scene conveys a sense of tranquility and the simple beauty of rural landscapes.

VisVM Guided

The photo depicts a pastoral scene with two young sheep, one black and one white with a light-colored stripe, walking in a grassy field. There are other animals in the background, including what appears to be ducks, grazing animals, and trees with bare branches, which suggests it might be spring or early summer when trees are not fully leafed. The presence of both sheep and ducks indicates a mixed-species environment, possibly on a farm or in a rural setting where different types of livestock can be found together. The scene is peaceful and represents a typical day in a rural area where animals are free to roam and graze.

CLIP-PRM Guided

This photo depicts a pastoral scene with two sheep, one black and one white, walking together in a grassy field. There are also other sheep in the background, grazing or resting. The presence of trees and a natural setting suggests that this is a rural or agricultural area. The sheep appear to be calm and are likely part of a flock that is not fully visible in the frame. The photo captures a peaceful moment in the lives of these animals, showcasing their natural behavior in a pasture.

Greedy Decoding

The photo shows two young sheep, also known as lambs, in a grassy field. The lamb on the left appears to be a black-faced sheep, possibly a breed like a Jacob or a Black Welsh Mountain, while the lamb on the right has a white face and a black body, which could be a breed like a Black Welsh Mountain or a similar black-faced sheep with a white face. They are walking together, and in the background, there are other sheep and trees, suggesting a rural or pastoral setting. The presence of a few birds, possibly ducks, adds to the natural, pastoral scene.

Figure 5: Case study 5: ViMaR and baseline captions.

Can you describe the main features of this image for me?



Attention Map after whole sentence generated



ViMaR Guided

In the image, there are two children standing on a sidewalk next to a white fire hydrant. The child on the left is wearing a red sweater, blue jeans, and white shoes, while the child on the right is dressed in a dark blue sweatshirt with a hood, blue-white pants, and white shoes. Both children are smiling at the camera and appear to be posing for the photo. In the background, there are trees, parked cars and a building with a red facade. The setting suggests an urban environment, possibly a residential area given the presence of the fire hydrant and the style of the buildings. The sky is clear, and there are no visible texts or distinctive brands in the image. The overall mood of the photo is cheerful and casual. The focus is on the children and the fire hydrant, with the background serving as context for the location of the photo. The fire hydrant has the number "8" visible on it, which is typically used to identify the hydrant for firefighting purposes. The image is well-lit and seems to have been taken during the day under natural light. The composition of the photo is balanced, with the children centrally positioned and the fire hydrant to their side.

VisVM Guided

In the image, there are two children standing on a sidewalk next to a white fire hydrant. The child on the left is wearing a red sweater and blue jeans, and the child on the right is dressed in a dark blue hoodie with white drawstrings, blue sweatpants, and white sneakers. They both appear to be standing in a residential area, as there are buildings in the background and a tree nearby. On the sidewalk, there are also some potted plants, and a man is standing a bit further down the sidewalk. On the right side of the image, there's a portion of a black car, and in the distance, there are more vehicles and buildings. The children seem to be posing for the photo, with the one on the right looking a bit more solemn than the other. The overall scene is casual and everyday, capturing a moment of childhood in an urban setting.

CLIP-PRM Guided

In this image, there are two young children standing on a sidewalk. The child on the left is wearing a red sweater and a pair of white shoes. They have short hair and are smiling at the camera. The child on the right is wearing a dark blue sweatshirt with a red stripe on the side, paired with dark pants and white sneakers. They have short, curly hair and also appear to be smiling. Between the two children, there is a white fire hydrant. The hydrant has a chain attached to it, and there are numbers and letters visible on its sides, although the specific details are not clear. The background shows a typical urban street scene with buildings, parked cars, and a tree lining the sidewalk. The lighting suggests it might be a sunny day, and there are no people other than the two children in the immediate vicinity.

Greedy Decoding

In the image, there are two children standing on a sidewalk next to a white fire hydrant. The child on the left is wearing a red sweater and blue jeans, while the child on the right is dressed in a blue jacket, blue pants, and white sneakers. Both children are smiling and appear to be posing for the photo. In the background, there are parked cars and a building with a red facade. The setting suggests an urban environment, possibly a residential area. The fire hydrant is a prominent feature in the foreground, indicating that the photo was taken in a location where fire safety is a concern.

Figure 6: Case study 6: ViMaR and baseline captions.

What do you see happening in this image?



Attention Map after whole sentence generated



ViMaR Guided

In the image, there is a male tennis player in the middle of a serve on a clay court. He is wearing a light blue shirt, dark blue shorts, and white shoes. The player is holding a red and white tennis racket and is in the process of striking a yellow tennis ball. He appears to be focused on the ball, which is in mid-air. In the background, there is a crowd of spectators watching the match, and there is a green advertisement board with the letters "BNP PARIBAS" on it, suggesting that the event is sponsored by that company. The setting suggests that this is a professional or high-level amateur tennis match. The player's posture and the position of the racket indicate that he is executing a powerful serve, which is a crucial part of the game in tennis. The clay court surface is characteristic of the French Open, one of the four Grand Slam tennis tournaments. The spectators are seated on chairs, and there are potted plants and flowers adding to the ambiance of the event. The overall scene captures the intensity and excitement of a live tennis match.

VisVM Guided

In the image, there is a tennis player in the middle of a serve. The player is wearing a blue shirt and dark shorts, with white shoes. He is holding a tennis racket with a red and white frame and a yellow ball in his other hand, preparing to toss the ball into the air and hit it. The player is looking upwards, likely tracking the ball's trajectory. In the background, there is a green advertisement board with the text "BNP PARIBAS" and a star logo, which suggests that the event is being sponsored by BNP Paribas, a French multinational banking corporation. There are also spectators in the stands watching the match. The setting appears to be a professional tennis court.

CLIP-PRM Guided

In the image, there is a male tennis player in action on a clay court. He is in the middle of a backhand shot, with his racket positioned behind him and his body turned towards the net. He is wearing a blue shirt, dark shorts, and white shoes. The player appears to be focused on the ball, which is in the air near him. In the background, there are spectators watching the match, and there is a sign that reads "BNP PARIBAS," which is likely a sponsor of the event. The setting suggests this is a professional tennis match.

Greedy Decoding

In the image, there is a male tennis player in the middle of a swing, hitting a tennis ball with his racket. He appears to be in the middle of a match, as indicated by the presence of a crowd in the background, and there are advertisements on the wall behind him. The player is wearing a blue shirt, dark shorts, and white shoes, which is typical attire for a tennis player. The setting suggests this is a professional tennis match.

Figure 7: Case study 7: ViMaR and baseline captions.

Can you elaborate on the elements of the picture provided?



Attention Map after whole sentence generated



ViMaR Guided

The image shows a beach scene with several elements: 1. Surfboards: There are four surfboards leaning against a railing on the left side of the image. They are of various colors and designs, suggesting they belong to different individuals. 2. Beachgoers: There are people on the beach, some of whom are sitting on the sand, while others are standing or walking. They appear to be enjoying the beach environment. 3. Ocean: The vast expanse of the ocean is visible in the background, with waves breaking on the shore. 4. Cliffs: In the background, you can see a large cliff or headland jutting out into the ocean, adding a rugged and natural feature to the beach landscape. 5. Railing: A black metal railing runs along the edge of the beach, providing a boundary between the sand and the road. 6. Signage: There is a sign on the right side of the image that reads "SALTBURN SURF SHOP OPEN," indicating that the surf shop is open for business. 7. Sky: The sky is clear and blue, suggesting a sunny day, which is ideal for beach activities. 8. Road: The road runs parallel to the beach, and it appears to be a public road with a sidewalk. The overall atmosphere of the image is relaxed and leisurely, typical of a day at the beach.

VisVM Guided

The image depicts a lively beach scene on a clear day with a blue sky and no visible clouds. The beach is sandy and appears to be a popular spot, as there are several people scattered across the area, some of whom are sitting on the sand, while others are standing or walking. In the foreground, there is a black metal railing that separates the beach from a paved area, which could be a boardwalk or a path for pedestrians. Leaning against the railing are surfboards of various colors, including yellow, blue, and white, suggesting that this is a place where surfing is a popular activity. On the paved area, there is a sign that reads "SALTBURN SURF SHOP OPEN," indicating that there is a surf shop nearby where visitors can rent or purchase surfing equipment. The presence of the surfboards and the surf shop sign suggests that this beach is known for surfing and possibly other water sports. The beach itself is lined with grassy areas, which might be used for picnicking or just relaxing, and the water appears to be calm with gentle waves. The overall atmosphere of the image is one of a fun, recreational environment where people come to enjoy the beach and the water activities it offers.

CLIP-PRM Guided

The image depicts a scenic beach view. In the foreground, there's a metal railing along the edge of the beach, likely to prevent people from accidentally falling into the ocean or to delineate the boundary of the beach area. To the right of the railing, there are several surfboards leaning against a sign that says "SALTBURN SURF SHOP OPEN," indicating that the shop is currently open for business. The surfboards are of different colors and designs, suggesting they are for rent or sale to visitors. Beyond the railing, the beach appears to be quite busy with several people enjoying the day. The sand is speckled with footprints, and there are a few small items scattered around, such as buckets, which are often used to collect shells or other small treasures from the beach. The ocean is visible in the background, with small waves indicating a calm sea. The sky is clear and blue, suggesting good weather. In the far background, there's a large rock formation that extends into the ocean. This cliff-like structure adds a dramatic backdrop to the scene and is a characteristic feature of the coastline. Overall, the image captures a typical day at a beachside surf shop, with people enjoying the beach and the shop offering surfing equipment for hire or sale.

Greedy Decoding

In the image, you can see a beach scene with the following elements: 1. Surfboards: There are several surfboards leaning against a railing on the left side of the image. The surfboards vary in color and design, suggesting they belong to different individuals. The presence of the surfboards indicates that surfing is a popular activity at this beach. 2. People and beach activity: There are several people on the sandy beach. Some are closer to the water's edge, while others are further back, possibly sunbathing or socializing. There are also beach umbrellas set up, providing shade for beachgoers. The presence of people and umbrellas suggests that it's a sunny day and the beach is a popular spot for relaxation and leisure activities. 3. Ocean: The vast body of water is visible, with waves crashing against the shore, which is a common feature of a coastal beach. The ocean's presence adds a dynamic element to the scene, with the constant movement of the waves. 4. Cliff: In the background, you can see a large cliff or headland jutting out into the ocean, adding a rugged and natural feature to the beach landscape.

Figure 8: Case study 8: ViMaR and baseline captions.

What do you see happening in this image?



Attention Map after whole sentence generated



ViMaR Guided

In the image, there is a **person riding a bucking bronco** during what appears to be a rodeo event. The rider is wearing a cowboy hat and a numbered on the back, which is typical attire for a rodeo competitor. The horse is in mid-buck, with its hind legs raised off the ground, and the rider is holding onto the horse's reins, trying to stay on. In the background, there is a crowd of spectators watching the event, and there is a banner with the words "LIGHT RODEO FANS" on it, suggesting that this event is part of a rodeo series. The setting looks like a dirt arena, which is common for rodeo competitions. The rider's posture and the position of the horse indicate a high level of skill and athleticism required for this sport. The spectators are focused on the action, indicating the excitement and intensity of the event. The image captures a moment of high adrenaline and athletic prowess. Please note that rodeo events can be controversial due to concerns about the welfare of the animals involved, particularly the horses and bulls used for bucking. The image does not provide enough context to determine the specific event or location, but it is clear that this is a professional rodeo event. The rider's attire and the presence of a numbered vest suggest that this is a competitive event, and the rider is likely a professional cowboy.

VisVM Guided

The image shows a person riding a horse, likely participating in a rodeo event, given the attire and the setting that includes spectators and arena seating. The rider is wearing a cowboy hat and appears to be engaged in an activity such as a bull-dogging competition, where the rider's goal is to throw a rope around the legs of a steer. The horse is in mid-gallop, and the rider is leaning forward, which is typical for riders attempting to catch a steer. The spectators are watching the event, and there is a banner in the background with the text "LIGHT RODEO FANS," suggesting that this is a public event where fans of rodeo sports are present. The setting looks like a typical rodeo arena, with metal barriers for safety.

CLIP-PRM Guided

In the image, there is a rodeo event taking place. A cowboy is riding a bucking bronco, which is a horse trained to buck and throw the rider off its back. The cowboy is holding onto a rigging, which is a strap that attaches to the horse's halter and the rider's belt, allowing the rider to stay on the horse. The horse is in mid-buck, and the cowboy is leaning forward, trying to maintain his balance. The audience is watching the event, and there is a sign in the background that reads "LIGHT RODEO FANS," indicating that this is a light rodeo event. The setting appears to be an outdoor arena, and the atmosphere is that of a competitive rodeo event.

Greedy Decoding

In the image, there is a rodeo event taking place. A cowboy is riding a bucking horse, which is a common event in rodeo competitions. The cowboy is holding onto a rigging, which is a strap that helps him stay on the horse. The horse is in mid-buck, with its hind legs raised off the ground, which is typical behavior for a bucking horse during a rodeo event. The cowboy is wearing a cowboy hat and a vest, which are typical attire for a rodeo cowboy. The audience is watching the event, and there is a banner in the background with the words "LIGHT RODEO FANS" on it, indicating that this is a light rodeo event. The setting appears to be an outdoor arena with seating for spectators.

Figure 9: Case study 9: ViMaR and baseline captions.

What are the key elements in this picture?

Attention Map after whole sentence generated



ViMaR Guided

The key elements in this picture include: 1. A person in mid-air, performing a trick with a frisbee. 2. The person is wearing a red shirt, black shorts, white sports shoes, and a black-and-white bandana. 3. The frisbee is neon yellow-green and appears to be in motion. 4. The setting is a beach with sand, and there are beach umbrellas in the background. 5. The background includes white parasols and people relaxing, indicating a recreational atmosphere. 6. The sky is clear and blue, suggesting good weather conditions. 7. A blue and yellow mat is laid on the sand. 8. There is a boat and some beach gear in the background, reinforcing the casual, coastal setting. 9. A few individuals are sitting or lounging on the sand. 10. The overall composition of the image conveys a sense of energy and fun, capturing a moment of athleticism and enjoyment.

VisVM Guided

The key elements in this picture are: 1. A person in mid-air performing a trick with a frisbee. 2. The frisbee is a yellow color. 3. The person is wearing a red shirt and black shorts. 4. They are wearing white shoes with black soles. 5. The person has a bandana around their head. 6. The background shows a beach setting with umbrellas, beach chairs, and other beachgoers. 7. The sky is clear, suggesting good weather conditions. 8. There is a body of water visible in the distance.

CLIP-PRM Guided

In the image, there are several key elements: 1. A person in mid-air, performing a frisbee trick. They are wearing a red shirt and black shorts, and they have a black bandana on their head. 2. The frisbee is in the air, close to the person's feet, indicating that they are either about to catch it or have just thrown it. 3. The person is on a beach, as evidenced by the sandy ground and the presence of beach umbrellas and chairs in the background. 4. There are other people on the beach, some of whom are seated under umbrellas, suggesting a relaxed, leisurely atmosphere. 5. The sky is clear and blue, indicating good weather conditions. 6. The overall scene suggests a recreational activity, possibly a casual game of frisbee, taking place on a sunny day at the beach.

Greedy Decoding

In this picture, there are several key elements: 1. A person in mid-air, performing a frisbee trick. The person is wearing a red shirt and black shorts, and they are holding a yellow frisbee. 2. A sandy beach setting with a clear blue sky. 3. A group of people sitting under umbrellas, which suggests a leisurely beach environment. 4. A boat visible in the background, indicating that the beach might be near a body of water. 5. A blue and yellow tarp on the sand, which could be part of the setup for the frisbee trick or simply a piece of beach equipment. 6. The person's dynamic pose and the frisbee in motion, which captures the action and skill involved in the trick. 7. The clear sky and the bright sunlight, which contribute to the overall vibrant and energetic atmosphere of the scene.

Figure 10: Case study 10: ViMaR and baseline captions.

What do you think is going on in this snapshot?



Attention Map after whole sentence generated



ViMaR Guided

In the image, we see a person standing on a paved area in front of a building with classical architecture. The person is holding a tennis racket and appears to be in the middle of a game or practice session, as indicated by the yellow-green tennis ball in motion nearby. The person is uniquely dressed in what appears to be a white robe or toga-like outfit with red crosses and blue accents, possibly a costume or themed attire rather than typical sportswear. In the background, the setting appears to be outside a classical-style building with tall columns and stone walls, resembling a historical or governmental structure, likely a cultural landmark. Several people are gathered behind a black railing on the raised platform of the building, observing the scene below. It's a sunny day, and the shadows on the ground suggest that the sun is low in the sky, which could indicate either early morning or late afternoon. The overall atmosphere is relaxed and leisurely, with people enjoying outdoor activities on a pleasant day. The presence of a backpack near the person with the racket might suggest that they have brought personal belongings or additional sports equipment with them.

VisVM Guided

In the image, we see a person standing on a paved area in front of a building with classical architecture. The individual is holding a tennis racket and appears to be in the middle of a tennis match or practice session. A sports ball, likely a tennis ball, is in mid-air nearby, suggesting that the person might have just hit the ball or is about to hit it. The person is dressed in sports attire, which includes a white and red outfit that is commonly associated with tennis. There are several other people in the background, some of whom are standing on a balcony above the tennis player, possibly watching the game or practice. The setting appears to be an urban environment, possibly a public park or a recreational area within a city. The presence of the tennis player and the sports equipment suggests that this is an outdoor recreational space where people can engage in sports activities.

CLIP-PRM Guided

In the image, we see a person standing on a paved area, holding a tennis racket and wearing a white outfit with red and blue accents. There is a tennis ball in the air, suggesting that the person is either about to serve or has just hit the ball. The person appears to be in the middle of a game or practice session, as indicated by the presence of the tennis ball and racket. In the background, there is a building with columns, which could be a public building or a sports facility. There are also other individuals in the distance, some of whom are standing on a balcony or a raised platform, possibly watching the person play tennis. The setting seems to be outdoors, and the presence of the tennis ball and racket suggests that the person is engaged in a sporting activity.

Greedy Decoding

In the image, there is a person standing on a paved area in front of a building with columns, which suggests it might be a public or institutional building. The person is holding a tennis racket and appears to be looking at a tennis ball on the ground. It seems like the person might be taking a break from playing tennis or is about to start a game. There are other people in the background, some of whom are standing on a balcony, which indicates that this might be a popular spot for people to gather and socialize. The presence of a backpack and a handbag suggests that the person might be a student or a visitor to the area. The overall atmosphere is casual and relaxed.

Figure 11: Case study 11: ViMaR and baseline captions.

Can you elaborate on the elements of the picture provided?



ViMaR Guided

The image shows a modern living room with a view of a city skyline through a large window. The room is well-lit, with natural light coming in from the window. Here are the key elements in the picture: Furniture: The living room is furnished with a variety of seating options, including a sectional sofa, two armchairs, a cylindrical side table and a round ottoman. The furniture appears to be of a contemporary style, with neutral tones that complement the room's decor. Flooring: The flooring is a light-colored hardwood, which adds warmth to the space. Window Treatments: The window is dressed with floor-to-ceiling curtains that are partially drawn, allowing for a balance between privacy and light. Decor: The room is decorated with a potted plant, which add a touch of greenery and life to the space. There are also decorative items such as books, a bowl, and a vase, which contribute to the room's aesthetic. Lighting: There is a ceiling light fixture that provides ambient lighting to the room. View: The most striking feature of the room is the view of the city through the large window. The window offers a panoramic view of the cityscape, with buildings of various heights and a river or body of water visible in the distance. Color Scheme: The color scheme of the room is neutral, with beiges, browns, and greys dominating the space. This creates a calm and inviting atmosphere. Ceiling: The ceiling has a modern design with exposed beams and a light fixture, which adds architectural interest to the room. Curtains: There are dark brown curtains that match the color of the drapes on the windows, providing a cohesive look. Overall, the living room is designed to be a comfortable and stylish space, with a focus on natural light and a connection to the urban environment outside.

VisVM Guided

In the image, you see a spacious living room that offers a view of a city skyline through a large floor-to-ceiling window. The room is well-lit, with natural light coming in from the window, and there is also a ceiling light fixture that provides additional illumination. The room is furnished with a variety of seating options, including a plush sofa and two armchairs, all arranged to create a comfortable seating area. The sofa and armchairs are upholstered in a neutral color palette, which complements the warm tones of the wooden flooring. In front of the seating area, there's a round ottoman, which can be used as a coffee table or additional seating. On the floor, there are a couple of decorative items, including a small plant in a pot and a decorative piece that could be a vase or a sculpture. The window not only provides a view but also allows for fresh air and natural light, which contributes to the room's inviting atmosphere. The curtains are partially drawn, which suggests that the room's occupants may want to control the amount of light and privacy coming in from outside. Overall, the room has been thoughtfully designed to create a welcoming space that combines comfort and style, while also offering a stunning view of the cityscape.

CLIP-PRM Guided

The image shows a spacious and well-lit living room with a view of a cityscape through floor-to-ceiling windows. The room features a variety of furniture, including: 1. A sectional sofa with beige cushions, which is the largest piece of furniture in the room. 2. A circular ottoman in the center of the room, which matches the sofa's color scheme. 3. Two armchairs, one of which is positioned closer to the window, providing a view of the city. 4. A small coffee table in front of the sofa, which appears to have a tray with books and a decorative item on it. 5. A floor lamp with a modern design, which is turned on, adding to the room's ambient lighting. 6. A large area rug under the coffee table, which has a geometric pattern and enhances the room's coziness. 7. A decorative item on the floor, possibly a sculpture or a piece of art, near the center of the room. 8. A large potted plant on the left side of the room, adding a touch of greenery and life to the space. 9. A window with a sheer curtain and a blackout curtain, allowing for adjustable light control. 10. A sliding glass door leading to a balcony, which is not visible in the image. The room's color palette is neutral with a mix of beige, brown, and black tones, creating a warm and inviting atmosphere. The cityscape in the background adds an urban touch to the space.

Greedy Decoding

The image shows a modern living room with a view of a city skyline through large windows. The room is well-lit, with natural light coming in from the windows, and there's a warm ambiance created by the lighting fixtures. Here are some key elements in the picture: Furniture: The room is furnished with a variety of seating options, including a sectional sofa, a couple of armchairs, and a round ottoman. The furniture appears to be of a contemporary style, with clean lines and neutral colors. Coffee Table: There's a coffee table in the center of the room, which seems to be made of wood and has a round shape. On top of it, there are a few decorative items, including a small plant and a couple of books. Rug: There's a large area rug on the floor, which adds texture and warmth to the space. Decor: The room is decorated with a few potted plants, which add a touch of greenery and life to the space. There are also some decorative items on the coffee table and on the windowsill, such as a small sculpture and a vase with flowers. Windows: The windows are large and offer a panoramic view of the city skyline. The windows are framed with white, which contrasts nicely with the darker tones of the room's interior. Curtains: The windows are dressed with dark curtains that can be drawn for privacy or to control the amount of light entering the room. Color Scheme: The room has a neutral color scheme, with beiges, browns, and creams dominating the space. This creates a calm and inviting atmosphere. Lighting: The room is well-lit, with a combination of natural light from the windows and artificial light from the ceiling fixtures. Balcony: There's a balcony visible through the windows, which adds an interesting architectural element to the view. Overall, the room is designed to be comfortable and stylish, with a focus on functionality and aesthetics. The view of the city adds a dynamic element to the space, making it a pleasant place to relax and enjoy the urban surroundings.

Attention Map after whole sentence generated



Figure 12: Case study 12: ViMaR and baseline captions.

Can you describe the main features of this image for me?



VIMaR Guided

The image shows a scene from a soccer (football) match. There are three players in the frame: one wearing a blue jersey and the others wearing a red jersey. The player in the red jersey, likely from the Arsenal team, is in possession of the yellow match ball and preparing to kick it, while the player in the blue jersey is attempting to intercept or challenge for the ball. The player in red is wearing white socks and shorts, while the players in blue are wearing blue socks and shorts. In the background, there is a crowd of spectators, indicating that this is likely a professional or semi-professional match. The stadium has multiple tiers of seating, suggesting it's a large venue designed to accommodate a significant number of spectators. The jerseys of the players have sponsor logos, which is common in professional sports. The player in the red jersey has 'Fly Emirates' visible on his jersey, and the number '15' is visible on his shorts. The grass of the field is well-maintained, which is typical for high-level soccer matches. Overall, the image captures a moment of competitive play in a soccer match, with the focus on the interaction between the two players in the center of the action. The atmosphere suggests a high-stakes game, given the intensity of the players and the presence of a large audience. The image does not provide enough context to determine the specific teams or the score of the match.

Attention Map after whole sentence generated



VisVM Guided

The image depicts a soccer match in progress. The main focus is on two players in the center of the image. The player in the foreground is wearing a blue and white kit, which is typical of the Arsenal football club, and he appears to be in a defensive stance, possibly preparing to intercept the ball. The player in the background is wearing a red and white kit, which is associated with the Arsenal football club, and he is in a forward position, possibly preparing to receive the ball or make a run. The soccer ball is visible on the ground, indicating that the game is in play. The player in the red and white kit is wearing yellow shoes, which is not a common color for soccer cleats and might be a sponsor's color or a personal preference. In the background, there are other players on the field, and a crowd of spectators can be seen in the stands, suggesting that this is a professional match being played in front of an audience. The grass of the field is well-maintained, and the stadium appears to be outdoors, based on the natural light and shadows. The atmosphere suggests a competitive sports event.

CLIP-PRM Guided

The image captures an intense moment in a soccer match. There are three players in the frame, each from different teams. The player in the foreground, wearing a red jersey, is in action, dribbling the ball with his left foot, while looking over his shoulder. His team's jerseys have white accents, and the jersey has the sponsor logo of Emirates. The player in the middle, wearing a dark blue jersey, is in a defensive stance, bending down and trying to intercept the ball. His team's jerseys have white and yellow stripes on the sides. The player in the background, wearing a red jersey similar to the first player's, is in motion, possibly running towards the ball or supporting his teammate. The crowd in the background is a blur of colors, indicating a large number of spectators in the stadium. The grass of the field is a vibrant green, typical for soccer pitches. The image conveys the competitive spirit of the game and the high level of athleticism on display.

Greedy Decoding

The image shows a soccer match in progress. There are three players in the foreground, with one player in a red jersey, likely from the Arsenal team, in possession of the ball and preparing to kick it. The other two players are in blue jerseys, which suggests they are from the opposing team, possibly the Gunners. The player in red is wearing white shorts, while the players in blue are wearing white socks and blue shorts. The background is filled with spectators, indicating that this is a professional match with an audience. The stadium has a large crowd, suggesting that the match is of significant importance or is a popular event. The players are on a well-maintained soccer field, and the weather appears to be clear and sunny, which is typical for outdoor sports. The image captures a moment of action, with the player in red looking down at the ball, possibly planning his next move, while the players in blue are positioned to defend or intercept. The focus is on the players and the ball, with the crowd serving as a backdrop to the action on the field.

Figure 13: Case study 13: ViMaR and baseline captions.

Describe the following image.

Attention Map after whole sentence generated



ViMaR Guided

The image depicts a street scene with a focus on a road intersection that has been cordoned off with yellow caution tape. The tape is attached to several orange traffic cones, which are placed on the pavement to block off the area. There are also a few white signs with black text, which appear to be related to the road work or closure. The signs include "Caution", "Electric", and "Days", suggesting that there may be ongoing electrical work or a temporary closure for a specific duration. The road itself is not visible due to the blockade, and there are no vehicles or pedestrians in the immediate vicinity of the cordoned-off area. The sidewalk on the right side of the image shows a few people walking, and there are a few cars and a truck parked or moving on the adjacent lanes. On the left side of the image, there is a building with a sign that reads "T&G Nails", indicating that it is a nail salon. The facade of the building is partially visible, and there is a small outdoor seating area with a potted plant. The sky is clear, and the lighting in the image is natural, with no artificial lights that are bright enough to overpower the ambient light. The overall style of the image is a candid, unposed photograph, capturing a typical urban street scene with a focus on the road closure. There are no visible people engaging with the road closure, and the photo seems to have been taken during the daytime.

VisVM Guided

The image depicts an urban street scene with a focus on a section of the road that has been cordoned off with yellow police tape. The tape is stretched across the street, indicating that the area is restricted or under investigation. Behind the tape, there are several orange traffic cones placed on the pavement, further emphasizing the restricted area. In the background, there is a building with a sign that reads "T&G NAILS." The street is lined with trees and street lamps, and there are parked cars along the side of the road. The sky is overcast, and the lighting suggests it might be late afternoon or early evening. There are no people visible in the immediate vicinity of the cordoned-off area, and the overall atmosphere of the image is quiet and still. The presence of the police tape and cones suggests that there may have been an incident that required the area to be cordoned off, but the exact nature of the incident is not visible in the image. The style of the image is a standard, non-stylized photograph that captures a moment in an urban environment. There are no visible texts other than the "T&G NAILS" sign.

CLIP-PRM Guided

The image depicts an urban street scene with a focus on traffic control measures. The street is cordoned off with yellow caution tape and orange traffic cones, indicating that there may be ongoing construction or maintenance work. There is a visible manhole cover with a sign that reads "Caution: Electrical Work." The sign suggests that there is work being done on the city's electrical infrastructure. On the right side of the image, there are parked cars, and on the left, there is a commercial building with signage. The sky is overcast, and the lighting suggests it is either late afternoon or early evening. The overall atmosphere is one of an active urban environment with ongoing city maintenance or construction activities.

Greedy Decoding

The image depicts a street scene with a focus on a section of road that has been cordoned off with yellow police tape. The tape is attached to several orange traffic cones, which are placed in a way that blocks access to the cordoned-off area. The tape and cones are clearly visible and appear to be in place to restrict access to the area. In the background, there is a commercial building with a sign that reads "T&G NAILS." The building has a green awning and a storefront with a display of plants. There are also other businesses visible in the distance, including a restaurant with outdoor seating. The street is lined with trees and streetlights, and there are parked cars along the side of the road. The sky is overcast, suggesting it might be a cloudy day. The overall style of the image is a standard, non-stylized photograph that captures a real-life urban scene.

Figure 14: Case study 14: ViMaR and baseline captions.