Scaling Inference-Time Search with Vision Value Model for Improved Visual Comprehension

Supplementary Material

GPT-4o and human evaluation prompt

Evaluate the following image captions generated by two vision-language models (VLMs) in response to a given image. Please decide which caption is "better" based on the following criteria:

Richness of Content: The caption should provide a comprehensive description of the visual elements in the image, detailing objects, actions, colors, or settings that enhance the viewer's understanding of the scene.
Accuracy: The caption must accurately reflect what is in the image without adding incorrect information or making erroneous assumptions about unseen aspects.
Harmlessness and Appropriateness: The caption should be free from any potentially harmful, offensive, or inappropriate content. Ensure that the caption does not make unnecessary assumptions about people's identities, backgrounds, or intentions.
Creativity and Elaboration: If the caption can remain accurate and harmless, it should include suitable and imaginative elaborations that enrich the scene's description (e.g., suggesting possible actions, moods, or a gentle interpretation of the scene's context).
Clarity and Coherence: The caption should be clear, concise, and well-structured, helping any viewer quickly understand the image's content without ambiguity.

After considering these criteria, please directly output your choice, the format of your choice should be 'Response1 is better', 'Response2 is better' or 'Tie'.
Response1: {}
Response2: {}

Table 6. Prompt used for image caption quality evaluation with GPT-4o and human.

A. Evaluation prompt for GPT and human evaluation

In this section, we provide the detailed prompt for GPT-4o and human evaluation in Section 4.1. We evaluate the caption quality from 5 aspects: Richness of Content, Accuracy, Harmlessness and Appropriateness, Creativity and Elaboration, Clarity and Coherence.

B. Details of VisVM training

Our training dataset consists of 378k <current sentence, current sentence clip score, next sentence, image>. Based on this training set, we train VisVM for 3 epochs with a learning rate of 5e-5 and a batch size of 1024. The latent dimension of the value head in VisVM is 2560. γ used for TD learning is 0.9. The entire training process is conducted on 8×80G A100 GPUs for 50 hours.

C. More experiments

In this section, we provide comparison with additional various finetuning (HADPO [73], POVID [74], CSR [75]) and guided decoding (VCD [24], CLIP-Guided) methods. Since all previous methods use LLaVA-1.5-7B as the base model, we also experiment on LLaVA-1.5-7B for generating description data and training VisVM with CLIP-PRM. The experimental results in Table 7 show that VisVM consistently outperforms all other methods across all metrics. Furthermore, VisVM focuses on inference-time search, making the improvement orthogonal to model finetuning, and capable of further enhancing model performance during inference.

D. A case study for VisVM Analysis

To better understand how VisVM influences VLM's response generation, this section examines how VisVM and CLIP-PRM select responses when presented with the same set of candidates. As illustrated in the upper part of Figure 5, given an image and prompt, LLAVA-Next generated three different response candidates. When using CLIP to directly score the responses, the second candidate received the highest clip score which is 0.2617 and is thus chosen as the current step response. However, VisVM considers potential hallucinations in subsequent responses induced by the sentence, resulting in a higher value 2.2695 for the third sentence among the given candidates. Therefore, the third sentence is selected. We search and obtain the final complete response using both CLIP-PRM and VisVM following their choices, as shown in the lower part of Figure 5. We observe that the response generated through VisVM search

Table 7. Comparison of VisVM-guided search with various finetuning and decoding methods. VisVM significantly outperforms all other methods across all metrics.

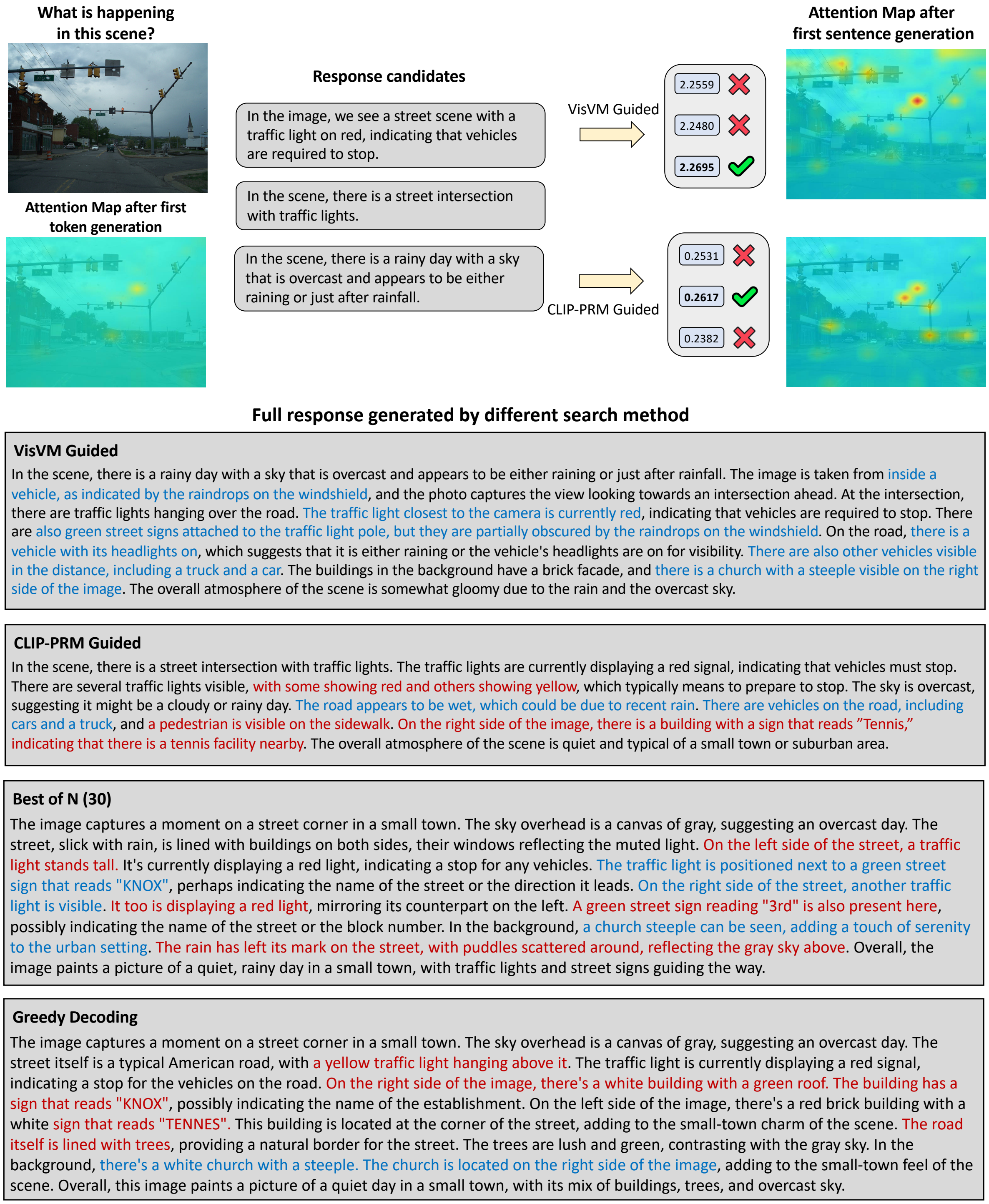
Base	Finetuning/Searching Method	CHAIRs ↓	CHAIRi ↓	MMHal ↑	AMBER Cov ↑
LLaVA-1.5-7B	Greedy (Default)	46.2	10.1	2.13	49.7
	HADPO (Finetuning)	43.6	9.7	2.27	48.9
	POVID (Finetuning)	39.7	9.4	2.45	46.8
	CSR (Finetuning)	35.2	8.6	2.69	44.3
	VCD (Searching)	33.4	8.9	2.45	51.9
	CLIP-Guided (Searching)	32.2	8.4	2.72	52.4
	VisVM-Guided	30.4	8.2	2.75	56.1

indeed contains more details and less hallucinations.

Additionally, Figure 5 also presents the changes in the LLAVA-Next's image-text cross-attention map following the selection of different candidates. We note significant differences in the attention maps after choosing different candidates. VisVM's attention map more comprehensively covers the entire image, enabling it to catch visual details such as "water droplets obscuring the road sign." In contrast, the CLIP-PRM's attention map over-emphasizes the area around the traffic light, leading to inaccurate description of the light's color and missing other visual details.

E. More case studies

In this section, we give more case studies from Table 8 to Table 16 to compare VisVM guided decoding results and CLIP-PRM guided decoding results.


What is happening in this scene?
Attention Map after first token generation
Response candidates
VisVM Guided
2.2559
2.2480
2.2695
CLIP-PRM Guided
0.2531
0.2617
0.2382
Attention Map after first sentence generation
Full response generated by different search method

In the image, we see a street scene with a traffic light on red, indicating that vehicles are required to stop.

In the scene, there is a street intersection with traffic lights.

In the scene, there is a rainy day with a sky that is overcast and appears to be either raining or just after rainfall.

VisVM Guided

In the scene, there is a rainy day with a sky that is overcast and appears to be either raining or just after rainfall. The image is taken from inside a vehicle, as indicated by the raindrops on the windshield, and the photo captures the view looking towards an intersection ahead. At the intersection, there are traffic lights hanging over the road. The traffic light closest to the camera is currently red, indicating that vehicles are required to stop. There are also green street signs attached to the traffic light pole, but they are partially obscured by the raindrops on the windshield. On the road, there is a vehicle with its headlights on, which suggests that it is either raining or the vehicle's headlights are on for visibility. There are also other vehicles visible in the distance, including a truck and a car. The buildings in the background have a brick facade, and there is a church with a steeple visible on the right side of the image. The overall atmosphere of the scene is somewhat gloomy due to the rain and the overcast sky.

CLIP-PRM Guided

In the scene, there is a street intersection with traffic lights. The traffic lights are currently displaying a red signal, indicating that vehicles must stop. There are several traffic lights visible, with some showing red and others showing yellow, which typically means to prepare to stop. The sky is overcast, suggesting it might be a cloudy or rainy day. The road appears to be wet, which could be due to recent rain. There are vehicles on the road, including cars and a truck, and a pedestrian is visible on the sidewalk. On the right side of the image, there is a building with a sign that reads ”Tennis,” indicating that there is a tennis facility nearby. The overall atmosphere of the scene is quiet and typical of a small town or suburban area.

Best of N (30)

The image captures a moment on a street corner in a small town. The sky overhead is a canvas of gray, suggesting an overcast day. The street, slick with rain, is lined with buildings on both sides, their windows reflecting the muted light. On the left side of the street, a traffic light stands tall. It's currently displaying a red light, indicating a stop for any vehicles. The traffic light is positioned next to a green street sign that reads "KNOX", perhaps indicating the name of the street or the direction it leads. On the right side of the street, another traffic light is visible. It too is displaying a red light, mirroring its counterpart on the left. A green street sign reading "3rd" is also present here, possibly indicating the name of the street or the block number. In the background, a church steeple can be seen, adding a touch of serenity to the urban setting. The rain has left its mark on the street, with puddles scattered around, reflecting the gray sky above. Overall, the image paints a picture of a quiet, rainy day in a small town, with traffic lights and street signs guiding the way.

Greedy Decoding

The image captures a moment on a street corner in a small town. The sky overhead is a canvas of gray, suggesting an overcast day. The street itself is a typical American road, with a yellow traffic light hanging above it. The traffic light is currently displaying a red signal, indicating a stop for the vehicles on the road. On the right side of the image, there's a white building with a green roof. The building has a sign that reads "KNOX", possibly indicating the name of the establishment. On the left side of the image, there's a red brick building with a white sign that reads "TENNES". This building is located at the corner of the street, adding to the small-town charm of the scene. The road itself is lined with trees, providing a natural border for the street. The trees are lush and green, contrasting with the gray sky. In the background, there's a white church with a steeple. The church is located on the right side of the image, adding to the small-town feel of the scene. Overall, this image paints a picture of a quiet day in a small town, with its mix of buildings, trees, and overcast sky.

Figure 5. A case study on VisVM-guided search. The upper part shows how VisVM and CLIP-PRM make different choices when given the same step response candidates, and the changes in the LLaVA-Next attention map after the choices are made. We can observe significant differences in the attention maps. The second part presents the complete responses obtained using different search methods, with blue text indicating correct details and red text indicating hallucinations. VisVM-guided search obtains response with richer details and fewer hallucinations.

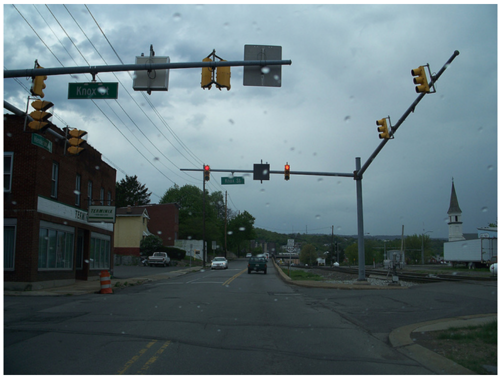
Example 1	
	
Question	What is happening in this scene?
VisVM Guided	In the scene, there is a rainy day with a sky that is overcast and appears to be either raining or just after rainfall. The image is taken from inside a vehicle, as indicated by the raindrops on the windshield, and the photo captures the view looking towards an intersection ahead. At the intersection, there are traffic lights hanging over the road. The traffic light closest to the camera is currently red, indicating that vehicles are required to stop. There are also green street signs attached to the traffic light pole, but they are partially obscured by the raindrops on the windshield. On the road, there is a vehicle with its headlights on, which suggests that it is either raining or the vehicle's headlights are on for visibility. There are also other vehicles visible in the distance, including a truck and a car. The buildings in the background have a brick facade, and there is a church with a steeple visible on the right side of the image. The overall atmosphere of the scene is somewhat gloomy due to the rain and the overcast sky.
CLIP-PRM Guided	In the scene, there is a street intersection with traffic lights. The traffic lights are currently displaying a red signal, indicating that vehicles must stop. There are several traffic lights visible, with some showing red and others showing yellow, which typically means to prepare to stop. The sky is overcast, suggesting it might be a cloudy or rainy day. The road appears to be wet, which could be due to recent rain. There are vehicles on the road, including cars and a truck, and a pedestrian is visible on the sidewalk. On the right side of the image, there is a building with a sign that reads "Tennis," indicating that there is a tennis facility nearby. The overall atmosphere of the scene is quiet and typical of a small town or suburban area.

Table 8

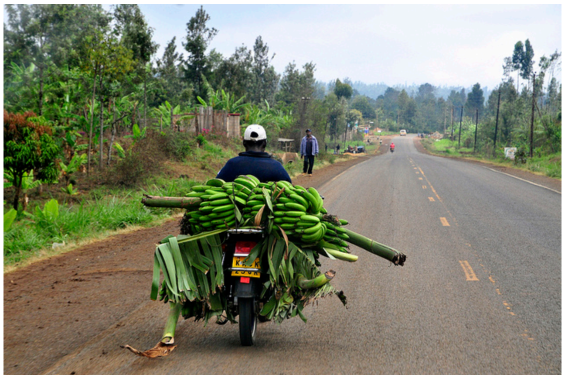
Example 2	
	
Question	What is happening in this scene?
VisVM Guided	In the image, there is a person riding a motorcycle or moped with a large bundle of bananas tied to the back. The person is wearing a helmet and appears to be on a rural road, with lush greenery and trees on both sides, suggesting a tropical or subtropical region. There are other people in the background, some of whom are walking along the road, and there's a small building or structure visible in the distance. The road seems to be a secondary road, with yellow markings indicating it's a two-way road, and there's a sign or symbol on the right side of the road that is partially obscured. The overall scene suggests a day-to-day activity in a rural or semi-rural area, possibly related to transportation of goods or agricultural produce.
CLIP-PRM Guided	In the image, there is a person riding a motorcycle loaded with a large quantity of green bananas. The bananas are tied to the back of the motorcycle, and the rider appears to be transporting them down a road. The setting is rural, with trees and greenery surrounding the road, suggesting that the bananas might have been harvested from nearby plantations. The person is wearing a helmet, which is a good safety practice when riding a motorcycle. The scene captures a moment of daily life in a region where bananas are a significant crop.

Table 9

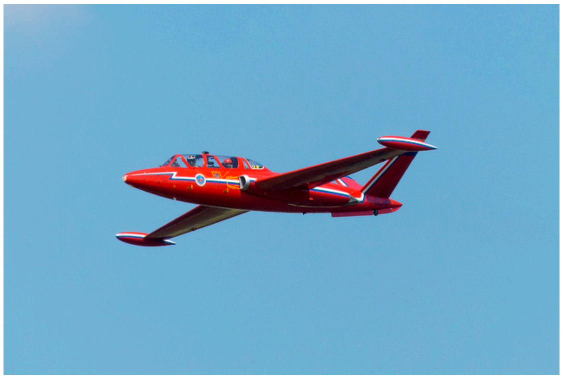
Example 3	
	
Question	Can you describe the main features of this image for me?
VisVM Guided	The image shows a single-engine, propeller-driven aircraft flying against a clear blue sky. The aircraft has a bright red fuselage with white and blue stripes running along its side, and a white and blue striped tail. There are no visible registration numbers or other identifying marks on the aircraft. The aircraft appears to be a small personal or recreational plane, given its size and the absence of any visible branding or commercial markings. The wings are swept back slightly, and there are no visible landing gear, suggesting that the aircraft is either in flight or has just taken off. The image captures the aircraft from a perspective below and to the side, providing a clear view of its underside and tail section.
CLIP-PRM Guided	The image shows a red airplane flying against a clear blue sky. The aircraft has a single propeller and a tail fin with a white stripe. It appears to be a small, single-engine plane, possibly a vintage or classic model, given its design and color scheme. The plane is captured in mid-flight, with its nose pointed slightly upward. The wings are swept back, which is a common feature in many small aircraft for improved aerodynamics. There are no visible windows on the fuselage, suggesting it may be a one-seater or a light utility aircraft. The overall impression is that of a leisurely flight or a demonstration of the aircraft's capabilities.

Table 10

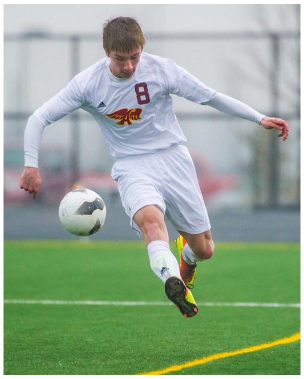
Example 4	
	
Question	Can you elaborate on the elements of the picture provided?
VisVM Guided	In the image, there is a soccer player in action on a field. The player is wearing a white soccer kit with a red number ”8” on the front, which is a common way to identify a player’s position or role within a team. The jersey also has an emblem that appears to be a bird, possibly a falcon, which could be the team’s logo. The player is wearing white socks with black cleats, which are typical for soccer players. The player is in the process of kicking a soccer ball, which is white with black pentagons, known as a pentagon ball, which is the standard shape for soccer balls. The playerś body is in a dynamic pose, with one leg extended forward and the other bent at the knee, indicating a powerful kick. The field is marked with white lines, which are standard for soccer fields, and there is a yellow line on the side, which could be the touchline or the goal line, depending on the specific context of the game. In the background, there is a chain-link fence, which is common for soccer fields to keep the ball from going out of play. There is also a vehicle, possibly a car or van, partially visible behind the fence, suggesting that the field might be located in a park or recreational area near a road. The sky is overcast, and the lighting suggests it might be late afternoon or early evening, as the shadows are long and the light is diffused. The overall atmosphere of the image is one of action and competition, capturing a moment of athletic endeavor.
CLIP-PRM Guided	In the image, we see a soccer player in action on a field. The player is wearing a white uniform with a red number ”8” on the front, which is likely his teamś jersey number. The uniform is the standard soccer attire, with shorts, a shirt, and possibly socks and shin guards that are not fully visible. The player is wearing yellow and orange cleats, which are common for soccer players to provide traction on the field. The player is in the process of kicking the soccer ball, which is in mid-air, indicating a dynamic moment in the game. The playerś right foot is extended towards the ball, suggesting a powerful strike. The player’s body is positioned with one leg bent and the other extended, which is a common stance for a player about to kick the ball. The background of the image shows a soccer field with a goal post visible, indicating that the player is near the goal area. The field appears to be well-maintained with clear boundary lines, and there are no visible spectators or other players, which puts the focus on the player and his action. Overall, the image captures a snapshot of an intense moment during a soccer game. The athlete’s concentration and the positioning of the ball and the playerś leg suggest a high level of skill and athleticism.

Table 11

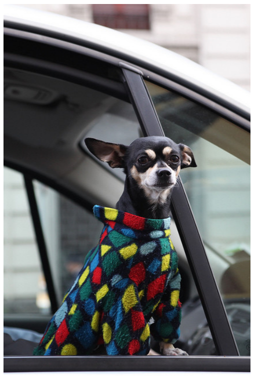
Example 5	
	
Question	Write a detailed description of the given image.
VisVM Guided	In the heart of a bustling city, a small Chihuahua with a vibrant coat of black and brown fur is the star of this scene. The dog, standing on its hind legs, has its front paws resting on the window sill of a white car. The car is parked on the side of a street, with a building visible in the background. The dog is dressed in a sweater that is a riot of colors - red, blue, green, and yellow. The sweater, made of soft fabric, adds a touch of whimsy to the otherwise mundane urban setting. The dog's head is tilted to the side, as if it's curiously observing its surroundings. Its eyes are wide open, taking in the sights of the city. Despite being inside a car, the dog seems to be enjoying the view outside. The image captures a moment of joy and curiosity in an everyday urban scene. It's a reminder that even in the hustle and bustle of city life, there's always room for a little bit of fun and playfulness.
CLIP-PRM Guided	In the heart of a bustling city, a small Chihuahua has found its perch on the passenger side of a car. The dog, dressed in a vibrant sweater, is the star of this urban tableau. The sweater, a riot of colors, is adorned with a playful polka dot pattern that stands out against the dog's dark fur. The dog is seated comfortably on the car's door sill, its front paws resting on the window ledge. Its head is tilted slightly to the left, as if it's curiously observing the world outside. The window itself is rolled down, allowing the dog to enjoy the sights and sounds of the city. The car itself is a stark white, contrasting with the dog's colorful attire. It's parked on a street lined with buildings, their details blurred by the distance. The overall scene paints a picture of a day in the life of this charming Chihuahua.

Table 12

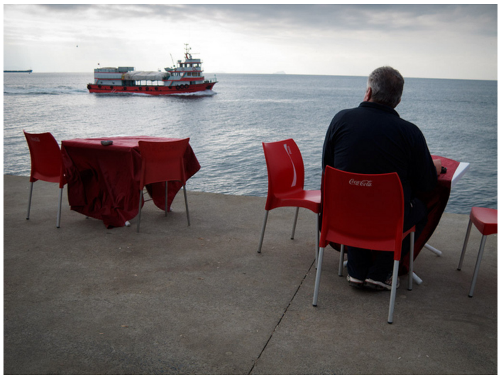
Example 6	
	
Question	Can you describe the main features of this image for me?
VisVM Guided	The image depicts a serene outdoor setting with a view of the ocean. In the foreground, there is a man seated at a table with a red chair. He appears to be looking out at the water, possibly enjoying the view or perhaps deep in thought. The table is covered with a red cloth, and there is a Coca-Cola logo on the chair, suggesting the location might be a restaurant or café with branded furniture. In the background, there is a large red and white boat on the water, which seems to be in motion, as suggested by the wake it is leaving behind. The sky is overcast, with a soft light that suggests either dawn or dusk. The water is calm, reflecting the colors of the sky and the boat. The overall scene is one of tranquility and solitude, with the man at the table as the main subject, framed by the expansive view of the ocean and the boat. The red color of the tablecloth and chair adds a vibrant contrast to the cooler tones of the water and sky.
CLIP-PRM Guided	The image shows a person sitting at a table on what appears to be a pier or waterfront. The person is facing away from the camera, looking out towards the water. The table is set with a red tablecloth and there are two red chairs, one of which the person is occupying. In the background, there is a large body of water, and on the water, there is a red and white boat that seems to be moving away from the pier. The sky is overcast, suggesting either early morning or late afternoon light conditions. The overall atmosphere of the image is calm and serene, with a sense of solitude and contemplation.

Table 13

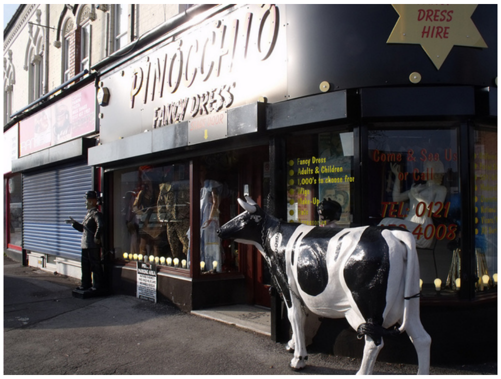
Example 7	
	
Question	What’s happening in the scene?
VisVM Guided	In the scene, there’s a store named ”Pinocchio Fancy Dress” which suggests it sells costumes or special occasion outfits. There’s a display in the window with various mannequins dressed in different styles of clothing, which might be used to showcase the types of outfits available inside the store. In front of the store, there’s a statue of a large cow with a spotted pattern, possibly representing a character from a story or just a decorative element. The cow is positioned in such a way that it appears to be walking towards the store, which adds a whimsical or playful touch to the storefront. On the right side of the image, there’s a sign that reads ”Dress Hire” indicating that the store offers rental services for costumes or dresses. There are also some signs with phone numbers and possibly some text that is too small to read clearly, likely providing additional information about the store’s offerings or contact details. The overall impression is that this is a storefront with a creative and engaging display designed to attract customers and convey the type of business it operates.
CLIP-PRM Guided	In the scene, there is a life-size statue of a cow with a black and white color pattern, standing in front of a store named ”Pino’s Fancy Dress.” The storefront has a sign that reads ”Dress Hire” and there are other signs indicating the services offered, such as ”Fancy Dress,” ”Adult & Children’s Fancy Dress,” and ”100’s Fancy Dress.” The cow statue appears to be a decorative element to attract attention to the store. To the left of the cow statue, there is a mannequin dressed in what appears to be a traditional or historical costume, which could be related to the dress hire services offered by the store. The overall setting suggests this is a shop that rents costumes for various occasions, and the cow statue could be a fun and quirky way to draw customers in.

Table 14

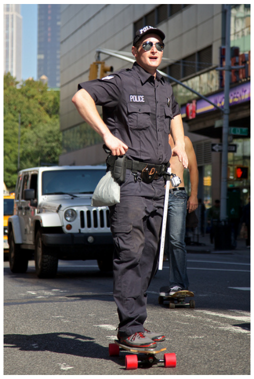
Example 8	
	
Question	Can you elaborate on the elements of the picture provided?
VisVM Guided	In the image, there are several elements that contribute to the urban scene. The most prominent figure is a police officer standing in the middle of the street. The officer is wearing a dark-colored uniform with the word ”POLICE” printed on the chest, indicating their role. They are also wearing a cap, sunglasses, and a holster, which are standard items of police attire. The officer is standing on a skateboard, which is a common mode of transportation in urban areas. This action could suggest that the officer is taking a break from duty or perhaps patrolling an area where skateboarding is allowed or encouraged. In the background, there is a busy street with vehicles, including a white SUV and a yellow taxi cab, which are typical of city traffic. There are also other pedestrians visible, contributing to the bustling atmosphere of the city. The architecture of the buildings suggests a modern cityscape, with high-rise buildings that are common in many urban centers. The presence of traffic lights and street signs further emphasizes the urban setting. The image captures a moment of everyday life in a city, with the police officer on a skateboard adding a unique and dynamic element to the scene.
CLIP-PRM Guided	The image shows a police officer riding a skateboard in a city street. The officer is wearing a uniform that includes a dark-colored shirt with the word ”POLICE” on the front, and heś also wearing a cap with a badge. He has a walkie-talkie on his belt and is holding a bag, possibly containing personal items or equipment. The officer is on a skateboard, which is a common mode of transportation for urban areas. The skateboard is red and black, and the officer is wearing casual shoes and sunglasses. In the background, there are other people and vehicles, indicating that this is a busy urban environment. There’s a white SUV to the left of the officer, and a traffic light is visible in the distance. The architecture suggests a modern cityscape, and the presence of a building with a sign that reads ”BROOKLYN” indicates that this scene might be taking place in a city that has a borough named Brooklyn. Overall, the image captures a moment of urban law enforcement using a skateboard as a means of navigating through the city. Itś a snapshot of a unique approach to policing in a contemporary urban setting.

Table 15

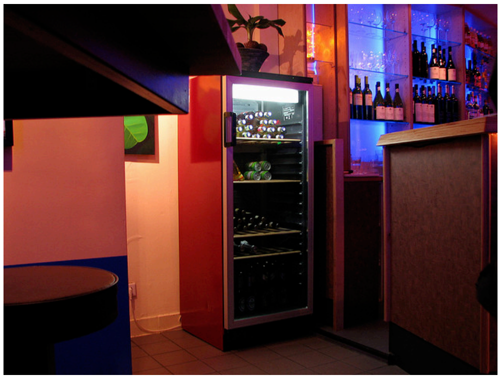
Example 9	
	
Question	Can you elaborate on the elements of the picture provided?
VisVM Guided	In this snapshot, it appears to be a scene from a bar or a restaurant with a focus on the beverage selection. The refrigerator, which is likely a beer cooler, is prominently stocked with various bottles of what seems to be beer. The lighting is dim, with a warm glow coming from the refrigerator, which is typical for establishments that want to highlight their beverage offerings. The presence of a potted plant and a shelf with more bottles in the background suggests a well-decorated and inviting space. It's possible that the image was taken during a quiet moment or just before or after opening or closing hours when the establishment is not busy.
CLIP-PRM Guided	In this snapshot, it appears to be an interior view of a bar or a restaurant with a focus on a refrigerator. The refrigerator is stocked with various bottles, likely containing beverages such as beer, wine, or cocktails. The lighting in the room is dim, with the refrigerator illuminated, which suggests that it might be evening or the place is intentionally dimly lit. The presence of wine glasses on a shelf above the refrigerator indicates that the establishment serves alcoholic beverages. The ambiance seems to be casual and relaxed, possibly catering to a social or dining experience.

Table 16