A. More Cases

Fig. 10, Fig. 11, and Fig. 12 showcase descriptions of a given ordinary image generated by various MLLMs from single-object, multi-object, and global perspectives. Specifically, LLaVA-1.6-7B enhances the static scene with a vivid sense of imagery using phrases such as "magical night ride" and "serene celebration", which add an appealing layer to the description. Descriptions from Qwen-VL-7B-Chat and Qwen-VL-7B+CLoT are similarly vivid and imaginative, incorporating metaphor and personification with phrases like "the tram was like a mobile diner, rolling down the street with its windows wide open" and "the car mirror seemed to be staring back at me with a mysterious smile", thereby heightening the appeal. VisualGLM-6B's descriptions focus on evoking a peaceful and gentle atmosphere. For instance, in describing the tram, it uses expressions like "quietly glides" and "gentle, calming atmosphere", conveying a serene urban landscape and providing a soothing visual experience. XMeCap also enhances the appeal by using dynamic and emotionally charged language, such as "glided down the street like a classic diner on wheels" and "twinkling lights on the palm tree dance alongside it", which create a more vivid and dramatic scene. In contrast, descriptions generated by CogVLM-7B are comparatively plain. While they accurately describe the elements in the image, they lack vividness. For example, the description includes "the headlights resemble a pair of curious eyes", which is somewhat figurative but fails to capture the scene's dynamism fully. Descriptions from MiniGPT-4-7B and mPLUG-Owl-I are also straightforward and concise, offering statements like "The tram moves on the street" and "The tram is next to some palm trees and lights". Though accurate, these descriptions lack rich detail and emotional depth. Descriptions from InstructBLIP-7B, while detailed, tend to be relatively lengthy, with the abundance of information potentially diluting the impact of the description. Upon comparison, CharmNet clearly excels in generating more appealing descriptions, characterized by a rich vocabulary and expressive language that enhance the image's appeal.

B. Error Analysis and Exploratory Analysis

We examine some less successful descriptions. For example, in an image showing a man leading a boy carrying a teddy bear, the generated description reads "a boy walking with a man in a crowd", which lacks vivid detail and emotional depth, rendering it less appealing.

Moreover, bias may exist due to the limited coverage of specific cultural or domain-specific images in the selected datasets. Therefore, we collect 100 images of iconic landmarks, such as the Eiffel Tower and the Egyptian pyramids, and apply CharmNet to generate culturally relevant captions. Our evaluation indicates that these captions are highly appealing, with an average human rating of 4.6 in 5. For instance, a generated appealing description like "In Times Square, the billboards shine brighter than my dreams, and buses move faster than my 5G!" effectively captures the vibrancy of the location, potentially attracting tourists and promoting sightseeing, demonstrating real-world applicability. Future work will explore more datasets to make further validation.

C. Other Details

Figure 7 presents additional samples of descriptions generated from various perspectives, with corresponding prompts detailed in Table 12. Figure 8 showcases further examples of different tasks, with prompts for each task provided in Table 13. Detailed evaluation metrics for these tasks are reported in Table 11. Ground-truth labels for all tasks were subjected to human evaluation, following the guidelines illustrated in Figure 9. We calculated Krippendorff's Alpha (IRA) and retained only annotations with scores exceeding 0.7. Additionally, for each task, we required a Pearson correlation coefficient greater than 0.9 between GPT-4o's predictions and human judgments. The influence of different dataset sizes in heuristic active learning and the effect of different difficulty levels across tasks are presented in Fig. 5 and Fig. 6, respectively, with results for other perspectives showing similar trends.

Task	Predicted Answer	Ground Truth	Score	Criteria
Discrimination	A	A	1	Same as the ground truth
	B	A	0	Different from the ground truth
Selection	A	A	1	Same as the ground truth
	B; C	A	0	Different from the ground truth
Rank	ABCD	ABCD	1	Only one combination with all 4 letters matching the relative positions in the ground truth
	ABDC; BACD;	ABCD	0.86	Only two combinations with 3 letters matching the relative positions in the ground truth
	ACBD; ACDB; ADBC; BCDA; CABD; DABC	ABCD	0.71	Only one combination with 3 letters matching the relative positions in the ground truth
	BCAD	ABCD	0.57	Four combinations with 2 letters matching the relative positions in the ground truth
	BADC; BDAC; CADB; CBAD	ABCD	0.43	Three combinations with 2 letters matching the relative positions in the ground truth
	ADCB; BDCA; CBDA; CDAB; DACB; DBAC	ABCD	0.29	Two combinations with 2 letters matching the relative positions in the ground truth
	CDBA; DBCA; DCAB	ABCD	0.14	One combination with 2 letters matching the relative positions in the ground truth
	DCBA	ABCD	0	No combinations matching the relative positions in the ground truth
Generation	[TEXT]	[TEXT]	1, 2, 3, 4, 5	appealing level of predicted answer

Table 11. Evaluation metrics for the four tasks based on AppealImage.

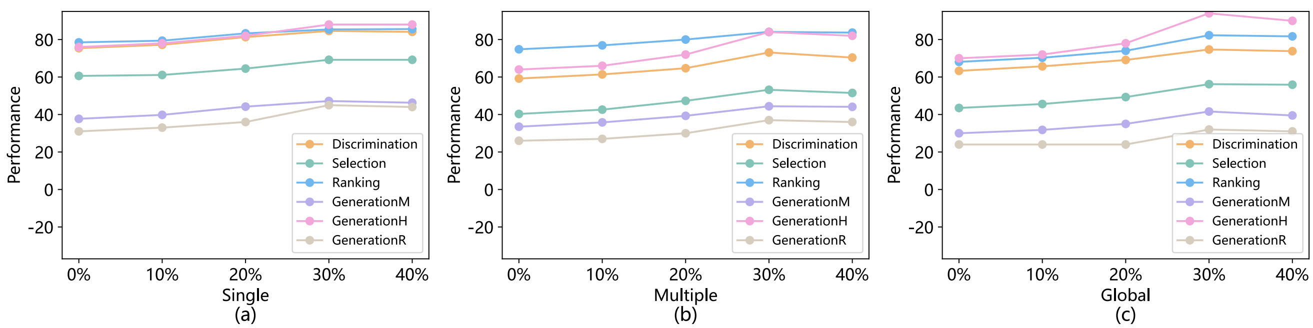

Figure 5. Effect of difficult data proportion in heuristic active learning across single-object, multi-object, and global perspectives. GenerationM, GenerationR, and GenerationH represent machine scores (e.g., BLEU, ROUGE, CIDEr, METEOR), referee scores, and human scores, respectively.

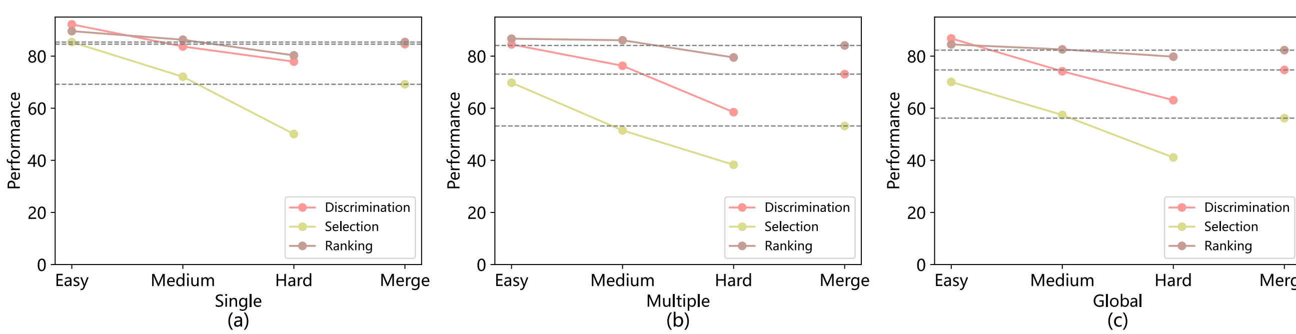

Figure 6. Performance on tasks across varying difficulty levels from single-object, multi-object, and global perspectives. GenerationM, GenerationR, and GenerationH represent machine scores (e.g., BLEU, ROUGE, CIDEr, METEOR), referee scores, and human scores, respectively.

Image	Description
	**Single** [Person] When in doubt, pedestrian about. Wordplay [Car] This car is tired of the street life; it just wants to retire to a nice, quiet driveway. Reversal [Traffic light] The traffic light's favorite color? Clearly, it's green with envy at the cars passing by. Wordplay [Building] This building never gets into office politics; it prefers to stay out of the concrete jungle drama. Reversal [Trees] Local trees branch out to meet new buds; unfortunately, leafing their current spot is out of the question. Phonetic similarity [Sidewalk] Sidewalks are the fashion runways of the urban jungle, but they always feel walked all over. Wordplay [Crosswalk] The crosswalk is where zebras in the city feel right at home — stripe for stripe. Wordplay [Street] This street always wanted to be a novel, but sadly it ended up as a thoroughfare-tale instead. Phonetic similarity [Posters] Posters are the gossips of the urban world, always sticking their faces where they shouldn't. Reversal **Multiple** [Car, Traffic light] The car was so red, it looked like it was trying to stop traffic with its dazzling charm! Reversal [Traffic light, Building] The traffic light felt overshadowed by the building, saying, "I may be bright, but you really know how to tower over me!" Reversal [Building, Trees] The building told the trees, "I may have more floors, but you definitely have more leaves!" Reversal [Trees, Sidewalk] The trees whispered to the sidewalk, "You really know how to pave the way for a good time!" Wordplay [Sidewalk, Crosswalk] The sidewalk said to the crosswalk, "You may get all the attention, but I'm the one who really supports the community!" Reversal [Crosswalk, Posters] The crosswalk joked, "I'm just here to help people cross over, while those posters are busy crossing out my fame!" Wordplay [Posters, Person] The posters were jealous of the person, saying, "You get to walk around while we just hang around!" Wordplay [Street, Car] The street sighed, "I'm just here for the ride, but that car is really taking me for a spin!" Wordplay **Global** He took the "walk" signal so seriously, he decided to take it for a walk. Wordplay

Figure 7. A sample from AppealImage with complete object descriptions.

Image


(a) Discrimination

Requirement: Identify which description is appealing.

Option A

Single Description: *This train is on the right track to light up your day!*
Multiple Description: *This train is so lit, it could give the lights a run for their money while chugging through the trees and mountains!*
Global Description: *This train is clearly on the right track to becoming the next big star in the locomotive light show!*
Highly appealing

Option B

Single Description: *This is train number 71, a historic steam locomotive.*
Multiple Description: *This is train number 71, a historic steam locomotive decorated with lights.*
Global Description: *This is train number 71, a historic steam locomotive decorated with lights, photographed in a mountainous area during the evening.*
Normal

(d) Generation

Requirement: Generate a highly appealing description.

Worst Best

(b) Selection

Requirement: Choose the most appealing description.

Option A

Single Description: *This train is on the right track to light up your day!*
Multiple Description: *This train is so lit, it could give the lights a run for their money while chugging through the trees and mountains!*
Global Description: *This train is clearly on the right track to becoming the next big star in the locomotive light show!*
Highly appealing

Option B

Single Description: *Train number 71 is an old steam locomotive.*
Multiple Description: *Train number 71 is a historic steam locomotive adorned with lights.*
Global Description: *Train number 71 is an old steam locomotive, decorated with lights, standing against a backdrop of mountains in the evening.*
Low appealing

Option C

Single Description: *The train is traveling down the tracks as evening approaches.*
Multiple Description: *The train, with its lights on, moves steadily through the landscape.*
Global Description: *This train, illuminated by its lights, continues its journey through the evening.*
Low appealing

(c) Rank

Requirement: Rank descriptions from most to least appealing.

Option A

Single Description: *This train is on the right track to light up your day!*
Multiple Description: *This train is so lit, it could give the lights a run for their money while chugging through the trees and mountains!*
Global Description: *This train is clearly on the right track to becoming the next big star in the locomotive light show!*
Highly appealing

Option B

Single Description: *The train travels down the tracks, standing out against the evening sky.*
Multiple Description: *The train, with its lights shining, moves smoothly through the evening landscape.*
Global Description: *This train, brightly lit, makes its way through the serene evening scenery.*
Second highly appealing

Option C

Single Description: *This train is moving along the tracks through the evening landscape.*
Multiple Description: *The lights on this train shine brightly as it moves through the mountains and trees.*
Global Description: *This train, adorned with lights, is traveling on its route in the evening.*
Moderately appealing

Option D

Single Description: *The train moves along the tracks.*
Multiple Description: *The train, with lights on, travels through the area.*
Global Description: *This train, lit up, continues on its route in the evening.*
Low appealing

Figure 8. Four tasks applied to a sample from AppealImage across single-object, multi-object, and global perspectives.

Single object	Generate humorous and creative sentences for the given image using one of the following techniques: 1. Wordplay - Utilizing puns or idiomatic expressions to create a clever and amusing sentence, e.g. “I’m on a seafood diet. I see food and I eat it!” 2. Phonetic similarity - Employing words that sound alike to generate a humorous twist, e.g. “I’m feeling grape!” 3. Reversal - Using literary techniques such as unexpected contrasts, irony, personification, or hyperbole to enhance the humor, e.g. “I’m not a complete idiot. Some parts are missing. Here is a list of objects (including human) in the given image. List of objects: [object list] Provide an interesting sentence for each object by strictly follow the template. Each object should be mentioned once, so there should be [number] sentences in total. Be careful and only use the object name in the list. The sentence should only focus on the object and shouldn’t mention other objects. You should only consider object from the list. - object: sentence
Multiple object	Generate humorous and creative sentences for the given image using one of the following techniques: 1. Wordplay - Utilizing puns or idiomatic expressions to create a clever and amusing sentence, e.g. “I’m on a seafood diet. I see food and I eat it!” 2. Phonetic similarity - Employing words that sound alike to generate a humorous twist, e.g. “I’m feeling grape!” 3. Reversal - Using literary techniques such as unexpected contrasts, irony, personification, or hyperbole to enhance the humor, e.g. “I’m not a complete idiot. Some parts are missing.” Here is a list of objects (including human) in the given image. List of objects: [object list] Provide an interesting sentence for each object by strictly follow the template and don’t use any other notations. Each object should be mentioned as the main object at least once, so there should be [number] sentences in total. In the sentence of each object, also mention another object in the list to increase the humor. Here is a template example for two sentences: - main object - other objects - sentence - main object - other objects - sentence
Global object	Generate one humorous and creative sentence for the given image using one of the following techniques: 1. Wordplay - Utilizing puns or idiomatic expressions to create a clever and amusing sentence, e.g. “I’m on a seafood diet. I see food and I eat it!” 2. Phonetic similarity - Employing words that sound alike to generate a humorous twist, e.g. “I’m feeling grape!” 3. Reversal - Using literary techniques such as unexpected contrasts, irony, personification, or hyperbole to enhance the humor, e.g. “I’m not a complete idiot. Some parts are missing.”
Extract object	Extract the objects from the image by strictly follow the template. The name of object should less than three words. - object1 - object2

Table 12. Instructions for generating appealing descriptions and extracting objects from images.

Discrimination	Identify which of the two descriptions, one highly appealing and one normal, is more appealing from single-object, multiple-object, and global perspectives, respectively. “Single-object” refers to a description focused on a specific object in the image, “multiple-object” refers to a description involving the interaction between the specific object and other objects in the image, and “global” refers to a description of the entire image as a whole.
Selection	Choose the most appealing description from three options: one highly appealing and two low appealing, from single-object, multiple-object, and global perspectives, respectively. “Single-object” refers to a description focused on a specific object in the image, “multiple-object” refers to a description involving the interaction between the specific object and other objects in the image, and “global” refers to a description of the entire image as a whole.
Rank	Rank four descriptions from most to least appealing based on their level of appeal from single-object, multiple-object, and global perspectives, respectively. “Single-object” refers to a description focused on a specific object in the image, “multiple-object” refers to a description involving the interaction between the specific object and other objects in the image, and “global” refers to a description of the entire image as a whole.
Generation	Generate a highly appealing description from single-object, multiple-object, and global perspectives, respectively. “Single-object” refers to a description focused on a specific object in the image, “multiple-object” refers to a description involving the interaction between the specific object and other objects in the image, and “global” refers to a description of the entire image as a whole.

Table 13. Task design instructions for the four tasks based on AppealImage.

User Questionaire:

Discrimination Task

Users are asked to choose the appealing description between two options. The descriptions will be provided in three contexts: "Single" refers to a description focused on a specific object in the image, "Multiple" refers to a description involving the interaction between the specific object and other objects in the image, and "Global" refers to a description of the entire image as a whole.

Image: [Insert Image]

Description A: [Insert Description A]

Description B: [Insert Description B]

Question: Which description is appealing?

Options: A / B

Selection Task

Users choose the most appealing description from three options. The descriptions will be provided in three contexts: "Single" refers to a description focused on a specific object in the image, "Multiple" refers to a description involving the interaction between the specific object and other objects in the image, and "Global" refers to a description of the entire image as a whole.

Image: [Insert Image]

Option A: [Insert Description A]

Option B: [Insert Description B]

Option C: [Insert Description C]

Question: Which description is the most appealing?

Options: A / B / C

Ranking Task

Users rank four descriptions by their engagement level. The descriptions will be provided in three contexts: "Single" refers to a description focused on a specific object in the image, "Multiple" refers to a description involving the interaction between the specific object and other objects in the image, and "Global" refers to a description of the entire image as a whole.

Image: [Insert Image]

Description A: [Insert Description A]

Description B: [Insert Description B]

Description C: [Insert Description C]

Description D: [Insert Description D]

Question: Please rank the following four descriptions by engagement level (from most appealing to least appealing).

Ranking: ABCD/.../ACBD

Generation Task

Users rate the generated descriptions on a scale from 1 to 5, where 1 means the worst and 5 means the best. We anonymize the model name. The descriptions will be provided in three contexts: "Single" refers to a description focused on a specific object in the image, "Multiple" refers to a description involving the interaction between the specific object and other objects in the image, and "Global" refers to a description of the entire image as a whole.

Image: [Insert Image]

Description: [Insert Generated Description]

Question: Please rate the following description, with 1 being the worst and 5 being the best.

Scoring Criteria: 1 / 2 / 3 / 4 / 5

Figure 9. User evaluation questionnaire for assessing options in discrimination, selection, and ranking tasks, as well as predictions in the generation task.

Single Description

CogVLM-7B:
[Tram] The tram moved steadily down the street, its windows open, revealing the passengers inside.
[Lights] The lights were bright enough to illuminate the area, making things a little clearer.
[Car mirror] The car mirror reflected the same scene, unchanged for a while.

InstructBLIP-7B
[Tram] Tram.
[Lights] Lights are visible.
[Car mirror] It's a train.

Llava-1.6-7B:
[Tram] "This tram is so long, it's **like a mobile city** with its own flag and streetlights!"
[Lights] The lights on the trolley are so bright, they're making the passengers **feel like they're on a magical night ride through a fairy tale.**
[Car mirror] The car mirror is so big, it's **like having a personal paparazzi** on the road.

MiniGPT-4-7B:
[Tram] The tram moves on the street.
[Lights] The lights are on.
[Car mirror] The car mirror shows a reflection.

mPLUG-Owl-I:
[Tram] The tram moves along the tracks.
[Lights] The lights are on.
[Car mirror] The car mirror reflects things.

Otter-I:
[Tram] The tram is a vehicle, it moves.
[Lights] Some lights are here.
[Car mirror] The mirror shows stuff.

Qwen-VL-7B-Chat:
[Tram] The tram was **like a mobile diner**, rolling down the street with its windows wide open.
[Lights] Those lights were so captivating, I nearly forgot about the tram itself!
[Car mirror] That car mirror? It's been showing me the same view for what feels like miles!

Qwen-VL-7B+CLoT:
[Tram] The tram **looked like it was going to jump** onto the tracks, but it was just the reflection in the car mirror.
[Lights] The string of lights above the train **looks like a colorful chandelier**, illuminating the streets below.
[Car mirror] The car mirror seemed to be **staring back at me with a mysterious smile**, as if it **had a secret or two hidden** behind its reflective surface.

VisualGLM-6B:
[Tram] The tram quietly glides down the tracks, offering a simple and smooth ride through the city.
[Lights] The lights cast a gentle glow, softly illuminating the surrounding area without being overpowering.
[Car mirror] The car mirror reflects the passing scenery, adding a quiet depth to the view.

XMeCap:
[Tram] The tram glided down the street **like a classic diner** on wheels, its windows inviting the world to join in.
[Lights] Those lights? They were so mesmerizing, I nearly missed the tram gliding past!
[Car mirror] That car mirror? It's been offering the same view for what **feels like an eternity**!

CharmNet:
[Tram] This tram **zips** through the streets **with a mission**, **leaving an exciting trail in its wake**!
[Lights] These lights **sparkle like stars come down to party**, casting a magical glow wherever they go!
[Car mirror] The car mirror captures more than just reflections; it **freezes the coolest moments** in the city, **each glance a snapshot of adventure**!

Ground Truth:
[Tram] This tram is so stylish, it must **have a degree in track-tical fashion**!
[Lights] These lights are **stringing everyone along with their bright ideas**!
[Car mirror] The car mirror is always **reflecting on the past**, but it's still **moving forward**.

Figure 10. Descriptions of a given ordinary image from other MLLMs in the single-object perspective.

Multiple Description

CogVLM-7B:
[Tram, Palm tree, Lights] The tram slowed as it passed the palm tree, with the lights casting a soft glow nearby.
[Car mirror, Building] The rearview mirror showed the building's outline, blending the old and new in a simple way.
[Lights, Flag] The building was large, making the lights and flag hanging from it seem much smaller by comparison.

InstructBLIP-7B
[Tram, Palm tree, Lights] The tram is a unique mode of transportation that is both historic and modern, as it combines the old-fashioned trolley system with the convenience of a modern tram. The tram is traveling down the street, passing by a palm tree, and is accompanied by a building, lights, and a car mirror. The tram is also carrying a flag, which adds to the festive atmosphere of the street.
[Car mirror, Building] There's a reflection.
[Lights, Flags] Building has lights and a flag on it.

Llava-1.6-7B:
[Tram, Palm tree, Lights] The tram glides by, illuminated by the soft glow of lights, as the palm tree stands tall, adding a touch of nature to the urban scene.
[Car mirror, Building] Reflected in the car mirror, the building stands tall, its lights twinkling faintly, merging the modern with the reflective.
[Lights, Flags] The building's lights and flags create a quiet display, marking a serene celebration within the cityscape.

MiniGPT-4-7B:
[Tram, Palm tree, Lights] The tram goes by a palm tree and some lights.
[Car mirror, Building] The lights reflect in the car mirror and on the building.
[Lights, Flags] The building has some lights and a flag.

mPLUG-Owl-I:
[Tram, Palm tree, Lights] The tram is next to some palm trees and lights.
[Car mirror, Building] The lights are shining. The mirror and building are there too.
[Lights, Flags] The building has some lights and flags on it.

Otter-I:
[Tram, Palm tree, Lights] There is a tram, and there is a palm tree, and also some lights.
[Car mirror, Building] Lights are reflecting, and there's a building.
[Lights, Flags] Flags, lights, and a building are all there.

Qwen-VL-7B-Chat:
[Tram, Palm tree, Lights] The tram had to slow down as it rolled by, catching glimpses of the palm tree twinkling with lights beside it.
[Car mirror, Building] The rearview mirror reflects the building's outline, together creating a unique blend of the old and the modern.
[Lights, Flag] The building was so massive that the lights and flag hanging from it seemed tiny by comparison.

Qwen-VL-7B+CLoT:
[Tram, Palm tree, Lights] As the tram passed by, the palm trees gently swayed to the rhythm of the city's vibrant lights, creating a serene scene.
[Car mirror, Building] The car mirror reflected the towering building, creating a fascinating blend of old and new architecture.
[Lights, Flags] The building adorned with bright lights and fluttering flags stood out as a beacon of celebration, casting a festive glow across the street.

VisualGLM-6B:
[Tram, Palm tree, Lights] The tram moves alongside palm trees and lights, creating a gentle, calming atmosphere as it goes.
[Car mirror, Building] The lights shine quietly, reflecting off the mirror and building, adding a touch of brightness to the scene.
[Lights, Flags] The building with lights and flags adds a gentle touch of festivity, making the street just a bit more lively.

XMeCap:
[Tram, Palm tree, Lights] The tram gently slowed down, allowing the twinkling lights on the palm tree to **dance alongside it in a fleeting waltz**.
[Car mirror, Building] The rearview mirror captured the building's silhouette, merging the past and present into a single, timeless frame.
[Lights, Flag] The building loomed so large that the lights and flag draped across it looked almost miniature in comparison.

CharmNet:
[Tram, Palm tree, Lights] The tram races by, **turning the city into a vibrant carnival** with palm trees swaying and lights flashing in celebration!
[Car mirror, Building] The lights **shine** so brightly that even the buildings and mirrors **can't help but join in**, reflecting the night's energy and excitement!
[Lights, Flags] **Dressed in** festive lights and fluttering flags, this building is ready to **steal the spotlight**, **shining like a star** in the city's celebration!

Ground Truth:
[Tram, Palm tree, Lights] This tram is so cool, it even has palm trees and lights to **keep it company**!
[Car mirror, Building] These lights are so bright, even the car mirror and building are **jealous**!
[Lights, Flags] The building is feeling festive with all the lights and flags, **it's ready for a party**!

Figure 11. Descriptions of a given ordinary image from other MLLMs in the multi-object perspective.

Global Description

CogVLM-7B:
The tram's front lights **resemble a pair of curious eyes**, as if it's always watching and waiting to see where the next adventure takes it.

InstructBLIP-7B
There's something there.

Llava-1.6-7B:
Why did the trolley car go to the beach? Because it heard the tracks were leading to a sandy destination!

MiniGPT-4-7B:
There is a train on the tracks

mPLUG-Owl-I:
The tram is moving.

Otter-I:
Tram on street.

Qwen-VL-7B-Chat:
I saw a bus that looked like a train and thought, "Oh, this is a trolley train."

Qwen-VL-7B+CLoT:
Why did the bus stop for the red light? Because it **didn't want to get a ticket**!

VisualGLM-6B:
The tram makes its way through the city, adding a soft, understated rhythm to the urban landscape.

XMeCap:
I saw **a bus masquerading as a train** and thought, "Oh, a trolley **with a secret**!"

CharmNet:
The tram is **leading a parade**, bringing life and color to the city as it **glides by with effortless charm**!

Ground Truth:
This tram must be **on a seafood diet** because it's definitely **a shell of its former self**!

Figure 12. Descriptions of a given ordinary image from other MLLMs in the global perspective.