Supplementary Materials for LLaVA-SP: Enhancing Visual Representation with Visual Spatial Tokens for MLLMs

1. Implementation details

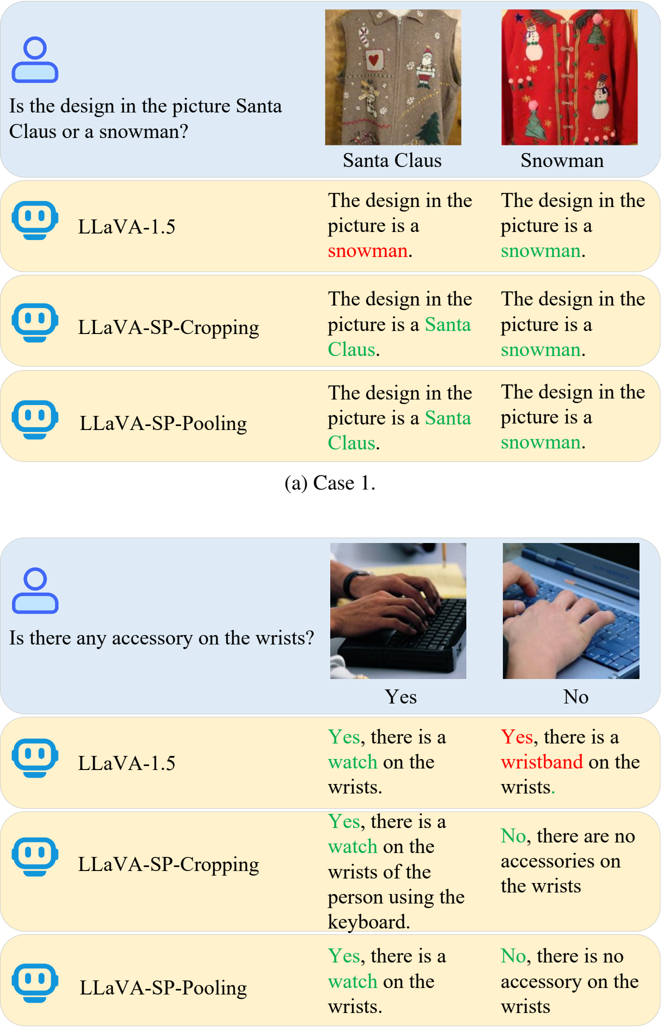
Hyperparameters. The experimental setup follows LLaVA-1.5 [3]. The training strategy consists of pre-training and instruction tuning. In the pre-training stage, the LLM is frozen, and the projector is trained to align vision and language representation. In the instruction tuning stage, both the LLM and the projector are trained to enhance the model's ability to follow human instructions. Specific training hyperparameters are detailed in Tab. 1.

Hyperparameter	Pre-training	Instruction Tuning
Global batch size	256	128
Projector LR	1×10^{-3}	2×10^{-5}
LLM LR	-	2×10^{-4}
LR schedule	Cosine decay	
Warmup ratio	0.03	
LoRA rank	-	128
Optimizer	AdamW	
Epoch	1	
Weight decay	0	
Deepspeed stage	2	

Table 1. **Training hyperparameters.** LR indicates learning rate.

Details on Spatial Feature Extractor. We use convolutional kernels to extract six visual spatial tokens. The convolution input channels are 1024, which match the dimension of vision encoder outputs, and the output channels are 512. The total parameters of the projector are 1536 MiB. We also experimented with transformer blocks, a 4 layers encoder-decoder structure, where the input and output channels are both 1024. The total parameters of the projector are 836 MiB. As mentioned in the main paper, the convolutional kernels outperforms the transformer blocks.

Details on Detail Feature Integrator. We use a simple linear layer with layer normalization to implement the Q and KV matrices for the cross-attention mechanism, where the input and output channel dimension are both 512.



Is the design in the picture Santa Claus or a snowman?

	Santa Claus	Snowman
LLaVA-1.5	The design in the picture is a snowman.	The design in the picture is a snowman.
LLaVA-SP-Cropping	The design in the picture is a Santa Claus.	The design in the picture is a snowman.
LLaVA-SP-Pooling	The design in the picture is a Santa Claus.	The design in the picture is a snowman.

(a) Case 1.

Is there any accessory on the wrists?

	Yes	No
LLaVA-1.5	Yes, there is a watch on the wrists.	Yes, there is a wristband on the wrists.
LLaVA-SP-Cropping	Yes, there is a watch on the wrists of the person using the keyboard.	No, there are no accessories on the wrists
LLaVA-SP-Pooling	Yes, there is a watch on the wrists.	No, there is no accessory on the wrists

(b) Case 2.

Figure 1. Examples of LLaVA-SP-Cropping and LLaVA-SP-Pooling on MMVP dataset. Correct and incorrect answers are marked in green and red respectively.

2. Qualitative Results

2.1. Case Study on MMVP Benchmark

We provide a case study of LLaVA-SP (comprising LLaVA-SP-Cropping and LLaVA-SP-Pooling) on the MMVP [7], to investigate their enhanced capabilities compared to the base LLaVA-1.5. From the output answers, we observe that: 1) In Fig. 1a, LLaVA-1.5 incorrectly identified both designs as a snowman. For example, when presented with a sweater design featuring Santa Claus-like elements such as a red color scheme, white trim, and Santa-like patterns, LLaVA-1.5 still answered that it was a snowman. In contrast, both LLaVA-SP-Cropping and LLaVA-SP-Pooling correctly identified the design as Santa Claus. The SFE in LLaVA-SP-Cropping likely focused on detailed regional features. For instance, it could have zeroed in on the specific Santa-like patterns and the red-white color combination that is characteristic of Santa Claus designs. Similarly, LLaVA-SP-Pooling, with its pooling operation in SFE, captured the overall visual context effectively. It could recognize the combination of elements that are typical of Santa Claus designs, rather than misinterpreting them as those of a snowman. 2) In Fig. 1b, when determining whether there are accessories on wrists, LLaVA-1.5 made mistakes. For example, in an image where a person was using a keyboard and had a watch on their wrist, LLaVA-1.5 incorrectly stated that there were no accessories on the wrist. Both LLaVA-SP-Cropping and LLaVA-SP-Pooling demonstrated superiority. LLaVA-SP-Cropping was able to extract relevant visual cues through its cropping-based SFE. It could focus on the wrist area and accurately identify the presence of the watch. LLaVA-SP-Pooling also performed well. Its pooling-based SFE captured the overall visual context, and DFI helped in integrating relevant features. This enabled it to correctly identify the presence of accessories on the wrist.

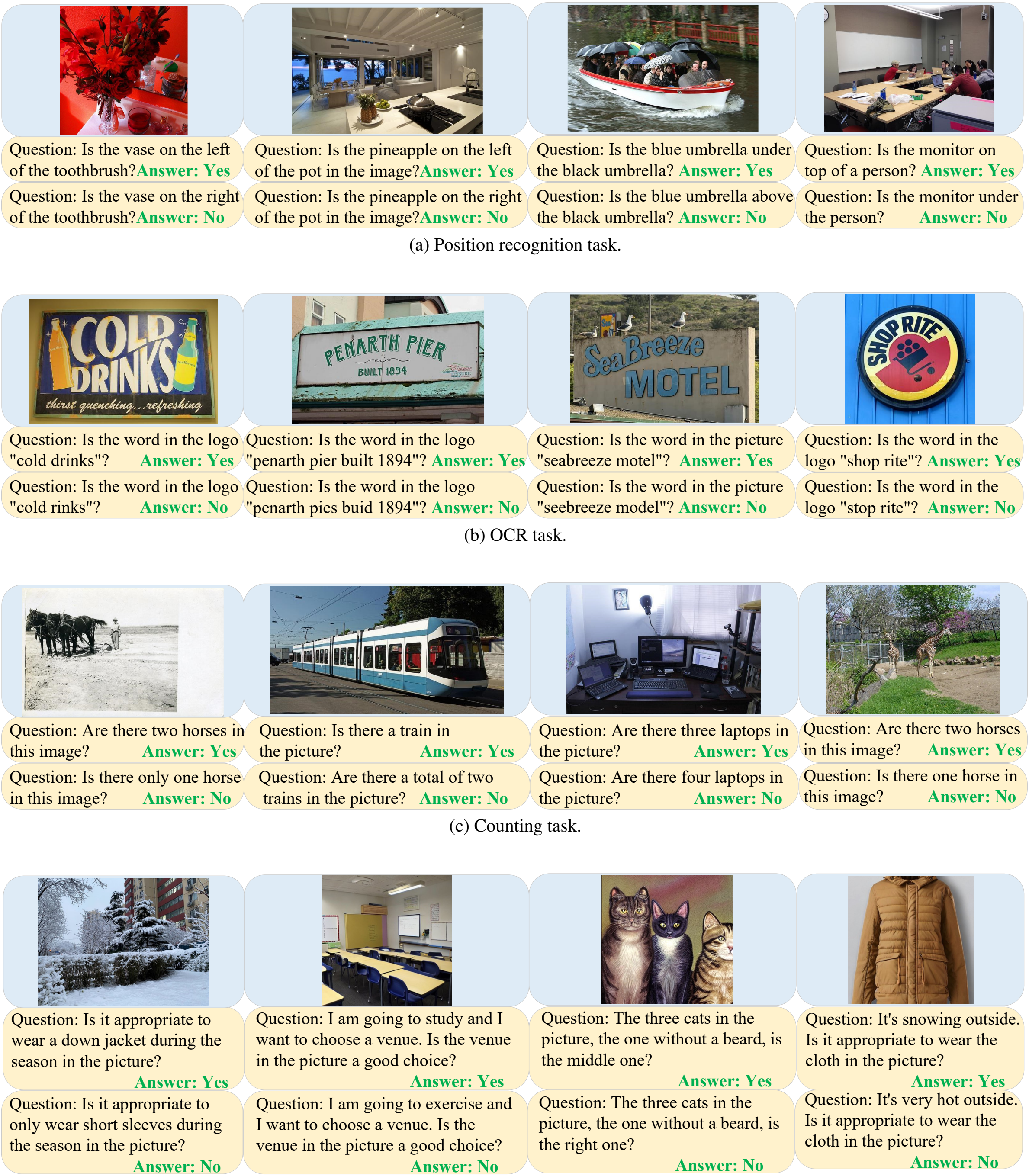
Overall, through the innovative designs of SFE and DFI, LLaVA-SP can distinguish differences between “CLIP-bind pairs” images that CLIP perceives as similar despite their clear visual differences.

2.2. Case Study on MME Benchmark

We provide a case study of LLaVA-SP (comprising LLaVA-SP-Cropping and LLaVA-SP-Pooling) on the MME [1]. From the output answers, we observe that: 1) In the position recognition task, LLaVA-SP demonstrated excellent performance. It was able to accurately determine the spatial relationships between objects. For example, in Fig. 2a, faced the question “Is the pineapple on the left of the pot in the image?”, LLaVA-SP can accurately analyze the visual layout. When asked another question about the same image, “Is the pineapple on the right of the pot in the image”, LLaVA-SP again precisely assesses the position. The SFE dissects the visual scene, focusing on the relative positions of the pineapple and the pot. This ability to handle multiple position-related questions about a single image accurately and consistently is a testament to the superiority of LLaVA-SP. 2) In the OCR task, LLaVA-SP effectively recognized text in images, accurately identifying words in logos. For example, in Fig. 2b, when presented with an image of a drinks shop logo that has the text “COLD DRINKS” in a unique cursive font, LLaVA-SP’s SFE first pinpoints the text region by detecting the contrast between the lighter text and the darker background. It then zooms in on each letter. For the letter “C”, it carefully analyzes the curved shape, the smooth transition of the stroke, and the way it connects to the following letter “O”. The SFE is able to handle the complexity of the cursive style and the decorative elements. The DFI further refines the recognition. It picks up on the minute variations in the thickness of the strokes and the small loops in the letters. This enables LLaVA-SP to accurately recognize “COLD DRINKS”. 3) In counting tasks, LLaVA-SP provided accurate counts. For example, in Fig. 2c(c) when presented with the image and the question “Are there three laptops in the picture?” LLaVA-SP’s SFE scanned the image, identifying the laptops based on their characteristic shapes and visual patterns. It differentiated laptops from other objects. The DFI then enhanced the analysis by focusing on details like the screen bezels and keyboard markings. This allowed LLaVA-SP to precisely count the laptops, answering both “Are there three laptops in the picture?” and “Are there four laptops in the picture?” accurately. It could handle occlusions and variations in laptop appearances, outperforming models that might miscount or miss some laptops. 4) In commonsense reasoning tasks, LLaVA-SP exhibited strong reasoning capabilities, offering correct answers to situational questions. For example, in Fig. 2d, when presented with an image and asked “It’s snowing outside. Is it appropriate to wear the cloth in the picture?” and “It’s very hot outside. Is it appropriate to wear the cloth in the picture?”, LLaVA-SP’s SFE analyzes the visual details of the cloth. It focuses on characteristics such as thickness and material texture. In snowy conditions, the SFE recognizes that the cloth appears warm and suitable. The DFI then further refines these features by integrating fine-grained details like the style and any associated accessories that suggest cold-weather wear. For the hot weather question, LLaVA-SP realizes the cloth is too heavy and inappropriate.

3. Qualitative Analysis of LLaVA-SP

To evaluate the effectiveness of LLaVA-SP in visual understanding, we qualitatively analyze its performance in comparison with LLaVA. The analysis highlights the strengths of SFE’s design, including cropping-based strategy, which enhances the model’s ability to capture fine-grained re-


Question: Is the vase on the left of the toothbrush?Answer: Yes
Question: Is the vase on the right of the toothbrush?Answer: No
Question: Is the pineapple on the left of the pot in the image?Answer: Yes
Question: Is the pineapple on the right of the pot in the image?Answer: No
Question: Is the blue umbrella under the black umbrella? Answer: Yes
Question: Is the blue umbrella above the black umbrella? Answer: No
Question: Is the monitor on top of a person? Answer: Yes
Question: Is the monitor under the person? Answer: No
Question: Is the word in the logo "cold drinks"? Answer: Yes
Question: Is the word in the logo "cold rinks"? Answer: No
Question: Is the word in the logo "penarth pier built 1894"? Answer: Yes
Question: Is the word in the logo "penarth pies buid 1894"? Answer: No
Question: Is the word in the picture "seabreeze motel"? Answer: Yes
Question: Is the word in the picture "seebreeze model"? Answer: No
Question: Is the word in the logo "shop rite"? Answer: Yes
Question: Is the word in the logo "stop rite"? Answer: No
Question: Are there two horses in this image? Answer: Yes
Question: Is there only one horse in this image? Answer: No
Question: Is there a train in the picture? Answer: Yes
Question: Are there a total of two trains in the picture? Answer: No
Question: Are there three laptops in the picture? Answer: Yes
Question: Are there four laptops in the picture? Answer: No
Question: Are there two horses in this image? Answer: Yes
Question: Is there one horse in this image? Answer: No
Question: Is it appropriate to wear a down jacket during the season in the picture? Answer: Yes
Question: Is it appropriate to only wear short sleeves during the season in the picture? Answer: No
Question: I am going to study and I want to choose a venue. Is the venue in the picture a good choice? Answer: Yes
Question: I am going to exercise and I want to choose a venue. Is the venue in the picture a good choice? Answer: No
Question: The three cats in the picture, the one without a beard, is the middle one? Answer: Yes
Question: The three cats in the picture, the one without a beard, is the right one? Answer: No
Question: It's snowing outside. Is it appropriate to wear the cloth in the picture? Answer: Yes
Question: It's very hot outside. Is it appropriate to wear the cloth in the picture? Answer: No

(a) Position recognition task.

(b) OCR task.

(c) Counting task.

(d) Commonsense reasoning task.

Figure 2. Examples of LLaVA-SP-Cropping and LLaVA-SP-Pooling on MME dataset.

gional details, and pooling-based strategy, which enables adaptive global reasoning. Two example images (Figure 3a and Figure 3b) are used for this analysis.

In Figure 3a, a desk scene with objects such as a laptop, books, and a chair is presented. LLaVA produces a general description, but it makes errors in object recognition, such as misidentifying the number of potted plants and failing to recognize the specific spatial arrangement of items. In contrast, LLaVA-SP generates more accurate and detailed descriptions. Specifically, using SFE with a cropping-based strategy, LLaVA-SP correctly identifies the presence and positions of key objects like the laptop and books. Meanwhile, with the pooling-based strategy, it focuses on summarizing the scene's overall calm and organized atmosphere, emphasizing its ability to balance detail with context.

Similarly, in Figure 3b, a beach scene is depicted, featuring people, boats, and trees. LLaVA fails to detect small objects like the boats and provides an inaccurate count of individuals in the scene, estimating at least 11 people when there are fewer. On the other hand, LLaVA-SP demonstrates significant improvements. The cropping-based strategy identifies the small boats and describes interactions between people and nearby objects with high precision. In contrast, the pooling-based strategy captures the overarching aesthetic of the beach scene, such as the presence of boats in the water and the relaxing coastal environment.

The superior performance of LLaVA-SP can be attributed to the combined effects of SFE and DFI. The cropping-based approach leverages SFE to focus on localized regions, making it effective for tasks that require object-level recognition. In contrast, the pooling-based approach benefits from DFI to aggregate information globally, providing a more abstract understanding of the scene. These two strategies offer complementary strengths, allowing LLaVA-SP to excel in both fine-grained and holistic reasoning.

In summary, LLaVA-SP significantly outperforms LLaVA in capturing both regional details and global context. The cropping-based approach is particularly suitable for precise object-level analysis, while the pooling-based approach excels in summarizing scene-level information. Together, they demonstrate the versatility and robustness of LLaVA-SP in handling diverse visual reasoning tasks.

4. Deep Analysis between LLaVA-SP-Cropping and LLaVA-SP-Pooling

The SFE enhances the vision encoder by introducing six visual spatial tokens, which can be obtained through two distinct methods: cropping and pooling. While both approaches aim to enrich the visual representation, their focus and mechanisms differ significantly, leading to distinct performance advantages in different scenarios. In this section, we conduct a deep analysis of LLaVA-SP-Cropping and LLaVA-SP-Pooling, comparing their outputs across the examples shown in Figs. 3 and 4.

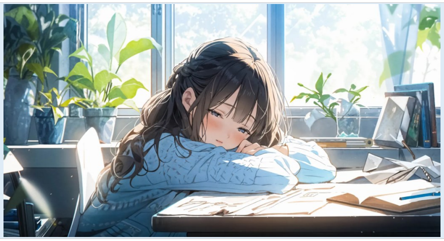
Cropping Method: The cropping approach extracts regional features by progressively narrowing the focus of the ViT patch features, starting from the global context and cropping inward toward the central region. This process generates multi-scale features arranged in the order of 'central region to global,' ensuring that detailed information from small but crucial regions is prioritized. For example, in Fig. 3a, LLaVA-SP-Cropping accurately captures specific details such as the posture of the individual and the texture of the surrounding elements, which are critical for precise understanding. Similarly, in Fig. 3b, cropping captures fine-grained features like the intricate design of objects and their interactions with the environment, showcasing its effectiveness in reasoning at a detailed level. Lastly, in Fig. 4, cropping excels at identifying the specific Mercedes logo on the clothing, a region-specific detail that might otherwise be overlooked in a global representation.

Pooling Method: In contrast, the pooling approach adopts a hierarchical strategy, using adaptive pooling layers to generate multi-scale features that range from abstract to specific. These features are arranged sequentially, first capturing the global structure and then transitioning to finer details, mimicking the way humans perceive visual scenes [6]. Pooling is particularly effective in scenarios requiring a holistic understanding. For instance, in Fig. 3a, pooling captures the overall context of the scene, emphasizing the arrangement and spatial composition of the surroundings. In Fig. 3b, pooling highlights the broader interactions between objects, such as the relationship between the primary elements and the background environment. Finally, in Fig. 4, pooling focuses on the subject's confident posture and overall compositional balance, offering a cohesive interpretation of the entire image.

Comparative Analysis: The differences between cropping and pooling stem from their respective focuses. Cropping excels in tasks requiring fine-grained image understanding, as it isolates and highlights specific regions with high precision. This is evident in examples such as the distinct feature in Fig. 4 and the detailed object interactions in Fig. 3b. Pooling, on the other hand, is better suited for tasks demanding a comprehensive understanding of the scene, as it captures the global structure and integrates contextual information. This is particularly beneficial in scenarios like Fig. 3a, where pooling effectively conveys the overall spatial arrangement and mood of the scene.

Both methods leverage the same foundational SFE mechanism, reshaping ViT patch features into their original 2D structures before processing them into multi-scale features. However, their strategies for organizing these features ("from central region to global" for cropping and "from abstract to specific" for pooling) lead to distinct strengths. Cropping prioritizes regional detail, making it more effective in capturing small but critical features. Pooling, by focusing on abstract-to-specific hierarchical information, provides a more holistic understanding of the image. Together, these methods complement each other, offering a flexible framework for addressing both fine-grained and global visual reasoning tasks.

Please describe this image.



LLaVA-1.5

The image depicts a woman sitting at a desk, surrounded by various books and potted plants. She is leaning on her elbows, possibly resting or studying. The desk is filled with numerous books, some of which are placed closer to her and others further away.
In addition to the books, there are two potted plants in the scene, one on the left side and another on the right side of the desk. A chair is also present in the scene, positioned near the desk. The woman appears to be the main focus of the image, as she is the only person visible in the scene.

LLaVA-SP-Cropping

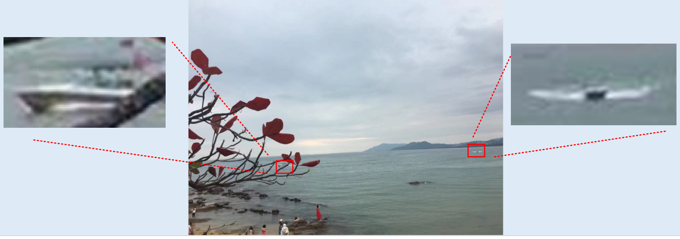
The image features a woman sitting at a desk in a room with a window. She is resting her head on her hand, possibly taking a break or daydreaming. The room is filled with various potted plants, with some placed on the desk and others scattered around the space.
There are multiple books on the desk, with some placed in front of the woman and others spread out across the desk. A laptop is also present on the desk, indicating that the woman might be working or studying. A chair is positioned next to the desk, providing a comfortable seating area for the woman.

LLaVA-SP-Pooling

The image features a woman sitting at a desk, with her head resting on her hand. She appears to be sleepy or tired, possibly from studying or working. The desk is surrounded by various books, with some placed on the desk and others scattered around the room.
There are several potted plants in the room, adding a touch of greenery to the space. One is located near the woman, while the others are placed around the room. A chair can be seen in the foreground, positioned close to the desk. The overall atmosphere of the scene is calm and peaceful.

(a) LLaVA-1.5 mistakenly pointed out that there are only two plants in the image. LLaVA-SP-Cropping identified details in the image, such as the books and laptop. LLaVA-SP-Pooling captured the overall atmosphere of this image.

Please describe this image.



LLaVA-1.5

The image features a beautiful beach scene with a large body of water and a rocky shore. A tree with red flowers is situated near the water, adding a touch of color to the landscape. Several people are enjoying the beach, with some standing close to the water and others further away. There are at least 11 people in the scene, scattered across the beach and near the water. Some of them are closer to the tree with red flowers, while others are positioned further away. The overall atmosphere of the scene is peaceful and serene, with people taking in the beauty of the beach and the ocean.

LLaVA-SP-Cropping

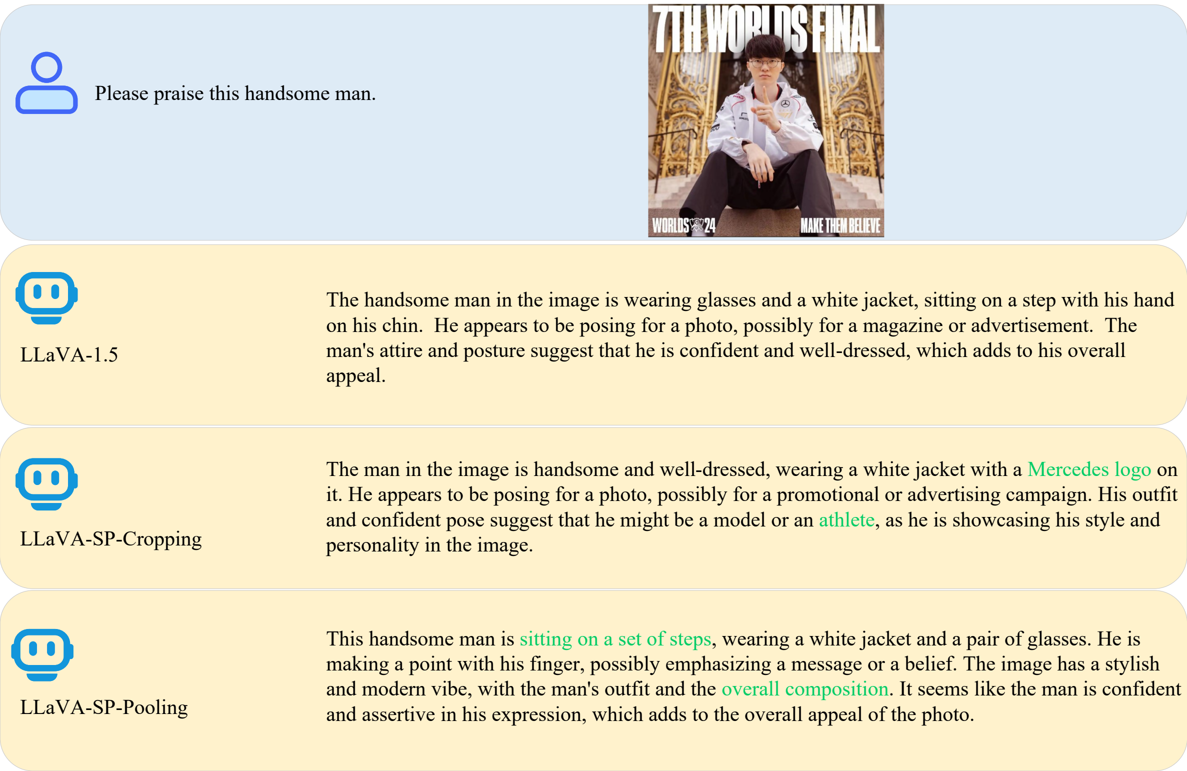
The image features a beautiful beach scene with a large tree near the water's edge. The tree is adorned with numerous red flowers, creating a vibrant and lively atmosphere. Several people are gathered around the tree, enjoying the view and the flowers. In addition to the people, there are a few boats visible in the water, adding to the coastal ambiance. The combination of the tree, the flowers, and the people creates a picturesque and inviting beach setting.

LLaVA-SP-Pooling

The image features a beach scene with a large tree near the water. The tree is adorned with red flowers, creating a beautiful contrast against the blue ocean. Several people are gathered around the tree, enjoying the view and the serene atmosphere.In the background, there are two boats visible on the water, one closer to the left side of the image and the other further to the right. The people are spread out across the beach, with some standing closer to the water and others further back. The overall scene is a picturesque representation of a beach day with friends and family.

(b) LLaVA-SP-Cropping and LLaVA-SP-Pooling all detected the small boat in the image, whereas LLaVA-1.5 failed to describe the small boat and incorrectly stated the number of people.

Figure 3. Deep Analysis between LLaVA-SP-Cropping, LLaVA-SP-Pooling and LLaVA-1.5. Correct and incorrect answers are marked in green and red respectively.



Please praise this handsome man.

LLaVA-1.5

The handsome man in the image is wearing glasses and a white jacket, sitting on a step with his hand on his chin. He appears to be posing for a photo, possibly for a magazine or advertisement. The man's attire and posture suggest that he is confident and well-dressed, which adds to his overall appeal.

LLaVA-SP-Cropping

The man in the image is handsome and well-dressed, wearing a white jacket with a Mercedes logo on it. He appears to be posing for a photo, possibly for a promotional or advertising campaign. His outfit and confident pose suggest that he might be a model or an athlete, as he is showcasing his style and personality in the image.

LLaVA-SP-Pooling

This handsome man is sitting on a set of steps, wearing a white jacket and a pair of glasses. He is making a point with his finger, possibly emphasizing a message or a belief. The image has a stylish and modern vibe, with the man's outfit and the overall composition. It seems like the man is confident and assertive in his expression, which adds to the overall appeal of the photo.

Figure 4. Compared to LLaVA-1.5, LLaVA-SP-Cropping captured the Mercedes logo on Faker's clothing, while LLaVA-SP-Pooling focused on the overall composition.

5. Limitation and Future Work

1) Did not utilize larger-scale LLMs: The experiments were conducted only on LLMs with 7B parameters, and the effectiveness of the method has not been validated on larger-scale LLMs. Future work will involve experiments on various LLMs such as Qwen2.5 [5], Mistral [2], and LLaMA3 [4].
2) Large Parameters: The SFE employs convolutional kernels to extract spatial information from visual features. The input and output channels of the convolutional kernels are 1024 (equal to the visual feature dimension output by ViT) and 512, respectively. Additionally, large kernels with a size of 16 are used. These will lead to a large number of parameters. According to the convolutional parameter calculation formula, the parameters for each convolutional kernel is:

$$parameters = C_{in} \times C_{out} \times Width \times Height, \quad (1)$$

where the C_{in} and C_{out} denotes the input channels and output channels respectively, $Width$ and $Height$ denotes the size of convolutional kernel.

In the future, we will adopt more efficient model design approaches to reduce parameters, improve training and inference speed, and achieve a trade-off between model performance and efficiency.

References

[1] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023. 2

[2] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023. 6

[3] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv:2310.03744*, 2023. 1

[4] Meta. Llama3, 2024. 6

[5] Qwen Team. Qwen2.5: A party of foundation models, 2024. 6

[6] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *arXiv preprint arXiv:2404.02905*, 2024. 4

[7] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *CVPR*, pages 9568–9578, 2024. 2