Enhancing Descriptive Captions with Visual Attributes for Multimodal Perception

Yanpeng Sun[1,2], Jing Hao[4], Ke Zhu[5], Jiang-Jiang Liu[3], Xiaofan Li[3]
Na Zhao[2], Zechao Li[1,*], Jingdong Wang[3*]
[1]NJUST, [2]SUTD, [3]Baidu Inc., [4]HKU, [5]NJU

A. More Details of Cap-Workflow pipeline

A.1. The Specific Visual Expert Model

In Cap-Workflow, we employ different visual expert models to obtain object-level and relation-level attribute labels. The main paper provides an overview of the process for obtaining different attributes and briefly introduces the types of Visual Specialists used. In Table 1, we will further elaborate on the off-the-shelf visual specialists employed in the Cap-Workflow pipeline. Based on our experience, existing open-source OCR models exhibit suboptimal performance in accurately recognizing text in images. To enhance OCR accuracy, we leverage Baidu's API [1], which offers superior text recognition capabilities. Additionally, for attributes such as logo, celebrity, and landmark recognition, the limited availability of labeled data poses a challenge. To mitigate this, we also utilize Baidu's API [2] to better capture these attributes, ensuring more precise and comprehensive annotations. Meanwhile, as a simple and versatile framework, Cap-Workflow allows users to flexibly select the most suitable visual specialist models based on their available resources.

A.2. The Detail of Relation Attributes

Relation Attributes play a crucial role in image captioning. Among them, HOI captures human-object interactions, while 2D Absolute Location describes an object's position within the image. Since such information is relatively sparse, we retain all relations in the captions. In contrast, 2D/3D Relative Location encodes spatial relationships between objects, which can become abundant and redundant as the number of objects increases. To address this, we retain only one dominant relationship per image for 2D/3D Relative Location, prioritizing the one with the greatest relative distance to ensure accuracy.

*Corresponding author.
[1]https://ai.baidu.com/tech/ocr/general
[2]https://ai.baidu.com/tech/imagerecognition

A.3. The Prompt Templates of Cap-Workflow

To effectively transform extracted visual attributes into natural language, Cap-Workflow designs two structured prompt templates that guide the LLM to convert discrete visual signals into coherent and readable descriptions. The first template focuses on object-level attributes, integrating multiple properties of each individual object. The second focuses on relation-level attributes, summarizing the relationships among different objects. The entire process is designed to mimic human perception: *first describing what each object is, and then explaining how these objects relate to each other within the scene.*

For the object-level prompt template (as shown in Figure 1), the LLM uses structured attributes extracted by visual specialists—such as category, color, texture, emotion, fine-grained classification (*e.g.*, replacing "dog" with "golden retriever"), OCR content, and depth—to check whether the reference caption already includes these details. If a specific attribute is missing, it will be added; if it conflicts with the extracted value, the information from the visual specialist takes priority. This process allows the model to merge all attributes into a richer region-level description without redundancy or hallucinated content. For example, the prompt explicitly defines conditional rules such as *"If a fine-grained animal category is available, replace the coarse label with the specific species; otherwise, keep the original name."* Likewise, it guides the LLM to naturally embed OCR text, detected emotions, and visual context into the caption, producing region-level descriptions that are both accurate and fluent.

For the relation-level prompt template (as shown in Figure 2), the LLM incorporates the relationships among different objects into the overall image caption. Among these, the P2O relations are directly obtained from the HOI model, while other types of relations are derived from detection and depth models. Specifically, the detection model provides the count of objects and, based on their bounding boxes, determines both 2D absolute locations and 2D relative locations to describe spatial relationships on the image plane.

The depth model, on the other hand, extracts a depth value for each object from its bounding box and the depth map. By comparing the depth values between objects, the model can infer their 3D spatial relationships, such as whether object A is in front of or behind object B. Moreover, incorporating the object-level captions further enriches the image caption with detailed visual information. This structured prompting process ensures that the model maintains spatial grounding while merging scattered regional details into one coherent and comprehensive image-level description.

In practice, both templates are implemented as single-turn structured prompts rather than multi-turn reasoning. Each prompt directly guides the LLM to generate captions based on the provided visual attributes and contextual cues, without requiring iterative dialogue. The object-level template focuses on enriching region descriptions with fine-grained visual details, while the relation-level template integrates spatial and interaction information to form a complete image-level caption. Together, these two prompts enable Cap-Workflow to seamlessly connect perception-level attributes with natural language expression, producing captions that are more detailed, precise, and contextually grounded than those from conventional LMMs.

B. Analysis of Cap-Workflow datasets

The Cap-Workflow dataset consists of two parts: Cap-Workflow-1M, comprising 1 million diverse image-text pairs sampled from the Laion dataset [12], and Cap-Workflow-118K, comprising 118,000 real image-text pairs from the COCO dataset [7]. Next, we will analyze the captions in Cap-Workflow-118K and compare them with captions annotated by humans [3] and generic LMM models [4, 6].

B.1. The Caption Length

In general, the longer caption could convey more detailed visual content. We compared the caption length of Cap-Workflow-118k with human annotations as well as captions generated by advanced MLLM models, InternVL2-26B [4] and LLaVA-Next-34B [6]. The results are summarized in Table 2. It was observed that human-generated captions were the shortest, as they typically focus on only the most salient objects. The InternVL2-26B could generate more longer captions, with approximately 106 tokens. The captions generated by LLaVA-Next-34B were the longest, averaging around 228 tokens, while Cap-Workflow-118k produced captions with an average of 218 tokens, approximately 10 tokens fewer than those of LLaVA-Next-34B.

B.2. The Lexical Composition

We conducted a detailed analysis of the lexical composition of the captions to evaluate how effectively each model described the visual content. This analysis examined the variety, frequency, and distribution of different word categories, including nouns, verbs, adjectives, adverbs, numerals, and more. As shown in Fig. 3, Cap-Workflow-118k contained the highest average number of lexical elements per sentence, demonstrating a more diverse and complex linguistic structure compared to other captions generated by other datasets. This richer composition indicates that Cap-Workflow-118k was better equipped to deliver nuanced and detailed descriptions of visual scenes, using a wider range of grammatical constructs to convey more comprehensive information.

B.3. The Word Clouds

In Fig. 4, we present word clouds for the captions generated by InternVL2-26B, LLaVA-Next-34B, and our Cap-Workflow pipeline. These visualizations highlight the most frequently used words across the different captioning methods, providing an intuitive comparison of the lexical patterns and focus areas of each caption. By examining the word clouds, we observed that the captions generated by our Cap-Workflow exhibited a notably diverse vocabulary. In particular, there was a significantly higher frequency of words describing the relative spatial relationships of objects, both in 2D and 3D space like *'left side', 'right region', 'front'*, and *'behind'*, compared to captions produced by other methods. This indicates that Cap-Workflow not only captured a wider range of visual details but also excelled in conveying the spatial context of objects within the scenes, offering more comprehensive descriptions of the visual content.

C. Training Details

We elaborate the training details and hyper-parameters used in our experiments for evaluating the effectiveness of Cap-Workflow-1M generated by our Cap-Workflow pipeline. The whole training step consists of three stages, as shown in Table. 3. During the pre-alignment stage, we exclusively train the projector, resulting in a more stable and consistent vision-language connection. In the pre-training phase, similar with ShareGPT4V [1], we unfreeze the Vision Encoder (VE) for the last 12 layers, the Language Model (LM), and the projector. Regarding the instruction tuning stage, we use the open-source LLaVA-mix-665K [8] and LLaVA-NeXT-data to fine-tune both the projector and language model of the LLaVA-v1.5 [8] and LLaVA-NeXT [6] models, respectively.

D. Detalied Evaluation Results

To highlight the significant improvements Cap-Workflow-generated image descriptions bring to model performance, we present a detailed evaluation of the MMbench results in Table 4. These results demonstrate how high-quality im-

```
messages = ["role" : "system", "content": f """ You are an AI visual assistant tasked with generating a detailed region caption by combining multiple visual attributes. Given a brief
reference caption of the region and the object attributes provided by various visual experts, create a single, cohesive description that includes all relevant details.

Ensure that the final caption:
1. Integrates the reference caption with the attributes to produce a richer, more comprehensive description.
2. Retains all region-level attribute information, such as colors, textures, object types, and spatial relationships.

It is important to preserve region-level attributes information. Remember you could not return any digital coordinates."""}]

brief _region_queries = "The brief description of this region is {reference caption}. "

if region_attributes["object"] is not None:
    cat_name = region_attributes["object"]
    det_query = "The detection model found that this is {cat_name}. "

if region_attributes["emotion"] is not None:
    emotion = region_attributes[" emotion "]
    emotion_query = "The emotion model found that the person with {emotion} in the caption."

if region_attributes["OCR"] is not None:
    OCR_str = region_attributes["OCR"]["str"]
    OCR_bbox = region_attributes["OCR"]["bbox"]
    ocr_query = "The OCR model found that the OCR information in {OCR_bbox} and add the information '{OCR_str}'."

if region_attributes["fine_grained"]["aircraft"] is not None:
    aircraft_name = region_attributes["fine_grained"]["aircraft"]
    aircraft_query = "If the airplane exits in the region, use the {aircraft_name} with airplane in the caption; otherwise, do not mention {aircraft_name}."

if region_attributes["fine_grained"]["animal"] is not None:
    animal_name = region_attributes["fine_grained"]["animal"]
    animal_query = "{cat_name} exits in the region and {animal_name} is the {cat_name}'s subclass, use {animal_name} in the caption;  otherwise, do not mention {animal_name}. "

if region_attributes["fine_grained"]["plants"] is not None:
    plants_name = region_attributes["fine_grained"]["animal"]
    plants_query = "{cat_name} exits in the region and {plants_name} is the {cat_name}'s subclass, use {plants_name} in the caption; otherwise, do not mention {plants_name}. "
…

if region_attributes["fine_grained"]["logo"] is not None:
    logo_name = box["logo"]
    logo_query += "If {logo_name} exits in the region, add the {logo_name} in the caption; otherwise, do not mention {logo_name}."

query = " ".join([brief _region_queries, det_query, emotion_query, ocr_query, aircraft_query, animal_query, plants_query , … logo_query ])
messages.append({"role":"user", "content":'\n'.join(query)}
```

Figure 1. The prompt for using LLM to generate an region caption by considering object attributes and reference captions.

Table 1. The Specific Visual Expert Model of Cap-Workflow.

Detection Model		**Depth Model**	**OCR Model**	**HoI Model**	**Emotion Model**	
In-domain	Open world					
Group Detr [2]	LaMI-DETR [5]	Depth Anything V2 [14]	API[1]	RLIPv2 [15]	[11]	
Fine-Grained Model						
Animal	Plant	Aircrafts	Logo	Landmark	Food	Celebrity
BioClip [13]	BioClip [13]	MMALNet [16]	API[2]	API[2]	PreNet [10]	API[2]

Table 2. Comparison of the different caption datasets. The *"ATL"* abbreviates the *"Average Token Length"*. The token length is counted by the tokenlizer of Vicuna-v1.5

Cpationed by	**Image Source**	**Samples**	**ATL of Caption**
Human	COCO	118k	14.67
InternVL2-26B			105.80
LLaVA-NeXT-34B			227.68
Cap-Workflow			217.71

age descriptions enhance the model's capabilities, particularly in logical reasoning, attribute reasoning, and relational reasoning. This improvement is driven by Cap-Workflow's ability to accurately capture object relationships and detailed attributes within images, enabling more effective reasoning and a deeper understanding of object interactions and characteristics.

Moreover, Cap-Workflow-generated descriptions achieve competitive performance in both fine-grained and coarse perception tasks, showcasing the effectiveness of integrating various visual experts to emulate manual annotation. This approach enriches visual information, resulting in higher accuracy and robustness in tasks like visual question answering and image comprehension.

```
messages = [{"role": "system", "content": f"""You are an AI visual assistant tasked with creating a more complete image description by merging the following information. You are provided with a brief description of the entire image and some descriptions of specific image regions.

The region descriptions consist of two parts: 1. The location of the region on the image. 2. A detailed description of the region. The location is represented as a bounding box in the format (x1, y1, x2, y2), with floating-point values ranging from 0 to 1. These values correspond to the coordinates of the top-left corner (x1, y1) and the bottom-right corner (x2, y2).

Please identify the correspondence between the objects mentioned in the brief description and those in the region descriptions. The region descriptions might be related to objects mentioned in the overall image description or in other region descriptions. Avoid repeating the description of the same object.

Note that the person providing the region descriptions can only see parts of the image, so the focus of these descriptions may differ. Your final output should be a complete image description that integrates all the relevant information. You do not need to address any contradictions between the brief description and the region descriptions, simply retain the useful information.

It is important to preserve OCR information, relative location information within the image, and the spatial relationships between objects as much as possible. Remember you could not return any digital coordinates."""}]

brief _image_queries = "The brief description of this image is {reference caption}. "
hoi_queries = " In the image, "
for person_box, relation in imaga_attributes["hoi"].items():
    hoi_queries += "the person in{person_box} is{relation}, "

count_queries = "In the image, there is "
for category, count in imaga_attributes["count"].items():
    count_queries += "{count} {category}, "

region_queries = ""
for instance_attribute in instance_attributes:
    pos = instance_attribute ["bbox"]
    category = instance_attribute ["object"]
    region_caption = instance_attribute ["detail_caption"]
    2d_location = imaga_attributes["2D_Relative_Location"][pos]
    region_query = "In {pos}, there is a {category} in {2d_location} and the brief description of this region is: {region description} ".
    region_queries  = " ".join([region_queries, region_query])

3d_location_queries = " "
for region, 3d_relation in imaga_attributes["3D_Relative_Location"].items():
    category_0, bbox_0 = region[0][cls_name],  region[0] [bbox]
    category_1, bbox_1 = region[1][cls_name],  region[1] [bbox]
    3d_location_query =  "Relative to the camera, the {category_0} in {bbox_0} of the image is {3d_relation} {category_1} in {bbox_1}  of the image"
    3d_location_queries = " ".join([3d_location_queries , 3d_location_query])

query = " ".join([brief_image_queries, hoi_queries, count_queries, region_queries, 3d_location_queries ])
messages.append({"role":"user", "content":'\n'.join(query)}
```

Figure 2. The prompt for LLM to generate an image caption by considering relation attributes, region location information and captions.

Table 3. Training details and hyper-parameters used in our experiments. 'VE' means the vision encoder of CLIP for the last 12 layers, and 'LM' refers to the language model.

Hyper-parameter	Pre-aligning	Pre-training	Instruction Tuning
Batch Size	256	256	128
Learning Rate	2e-5	2e-5	2e-5
LR Schedule		cosine decay	
LR Warmup Ratio	0.01	0.01	0.01
Weight Decay	0	0	0
Trainable Module	Projector	**LLaVA-v1.5**: Projector, VE, LM **LLaVA-NeXT**: Full Model	**LLaVA-v1.5**: Projector, LM **LLaVA-NeXT**: Full Model
Epoch	1	1	1
Optimizer		AdamW	
DeepSpeed stage	3	3	3
Dataset	Cap-Workflow-1M	Cap-Workflow-1M	**LLaVA-v1.5**: LLaVA-mix-665K **LLaVA-NeXT**: LLaVA-NeXT-data

E. Visualizations on Cap-Workflow

To visually demonstrate the quality of captions annotated by Cap-Workflow, we compared them with captions generated by generic LMMs, such as InternVL2-26B and LLaVA-NeXT-34B. The visualization highlights the differences in caption quality, providing a clear comparison of Cap-Workflow's detailed and accurate descriptions against

Table 4. CircularEval multi-choice accuracy results on MMBench [9] dev set. We adopt the following abbreviations: LR for Logical Reasoning; AR for Attribute Reasoning; RR for Relation Reasoning; FP-C for Fine-grained Perception (Cross Instance); FP-S for Finegrained Perception (Single Instance); CP for Coarse Perception.

Annotation Method	MMBench-CN							MMBench						
	Overall	LR	AR	RR	FP-S	FP-C	CP	Overall	LR	AR	RR	FP-S	FP-C	CP
InternVL2-26B [4]	56.9	28.8	58.8	**59.1**	54.9	44.1	74.0	64.8	35.6	68.3	**57.4**	**70.3**	52.4	77.4
LLaVA-NeXT-34B [6]	56.5	28.8	60.3	56.5	54.3	42.0	74.3	64.9	31.4	68.3	57.4	69.3	**54.5**	**79.7**
Cap-Workflow	**58.2**	**29.7**	**62.3**	57.4	**56.7**	**45.5**	**74.7**	**65.8**	**37.3**	**71.4**	57.4	70.0	53.8	78.4
InternVL2-26B [4]	58.8	31.4	60.3	51.3	56.0	51.0	78.4	66.7	36.4	71.4	59.1	66.9	**62.9**	80.1
LLaVA-NeXT-34B [6]	59.8	31.3	60.8	54.8	56.7	49.7	**80.7**	67.2	38.1	**69.8**	57.4	**69.3**	60.1	82.1
Cap-Workflow	**60.1**	**31.4**	**62.3**	**55.7**	**57.0**	**51.0**	79.1	**68.5**	**39.0**	**72.4**	**67.8**	67.9	59.4	**82.8**

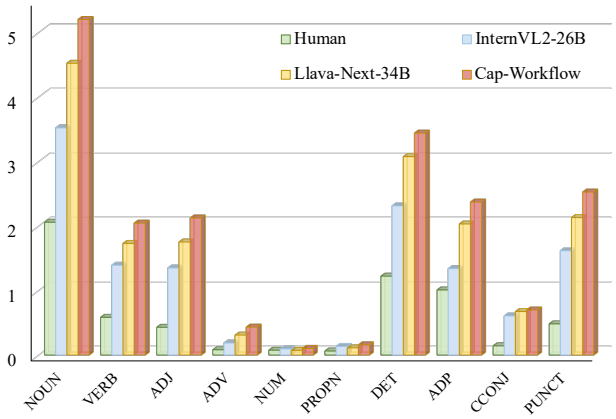

Figure 3. Comparison of lexical composition of the captions generated by different methods. The y-axis represents the average frequency of each class of lexical per sentence in the datasets.

those produced by the LMMs. As shown in Figure 5, Cap-Workflow captions consistently capture more nuanced object attributes, relationships, and contextual details, showcasing its superior annotation capabilities. This comparison underscores Cap-Workflow's effectiveness in generating high-quality captions that enhance downstream visual-language tasks.

References

[1] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, pages 370–387, 2024. 2

[2] Qiang Chen, Xiaokang Chen, Jian Wang, Shan Zhang, Kun Yao, Haocheng Feng, Junyu Han, Errui Ding, Gang Zeng, and Jingdong Wang. Group detr: Fast detr training with group-wise one-to-many assignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6633–6642, 2023. 3

[3] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 2

[4] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024. 2, 5

[5] Penghui Du, Yu Wang, Yifan Sun, Luting Wang, Yue Liao, Gang Zhang, Errui Ding, Yan Wang, Jingdong Wang, and Si Liu. Lami-detr: Open-vocabulary detection with language model instruction. In *European Conference on Computer Vision*, pages 312–328, 2025. 3

[6] Bo Li, Kaichen Zhang, Hao Zhang, Dong Guo, Renrui Zhang, Feng Li, Yuanhan Zhang, Ziwei Liu, and Chunyuan Li. Llava-next: Stronger llms supercharge multimodal capabilities in the wild, 2024. 2, 5

[7] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755, 2014. 2

[8] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306, 2024. 2

[9] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2024. 5

[10] Weiqing Min, Zhiling Wang, Yuxin Liu, Mengjiang Luo, Liping Kang, Xiaoming Wei, Xiaolin Wei, and Shuqiang Jiang. Large scale visual food recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8): 9932–9949, 2023. 3

[11] Andrey Savchenko. Facial expression recognition with adaptive frame rate based on multiple testing correction. In *International Conference on Machine Learning*, pages 30119–30129, 2023. 3

[12] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo

InternVL2 Llava-NeXT Cap-Workflow

Figure 4. Word Cloud of captions generated by InternVL2, LLaVA-Next and Cap-Workflow.

Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems*, pages 25278–25294, 2022. 2

[13] Samuel Stevens, Jiaman Wu, Matthew J Thompson, Elizabeth G Campolongo, Chan Hee Song, David Edward Carlyn, Li Dong, Wasila M Dahdul, Charles Stewart, Tanya Berger-Wolf, et al. Bioclip: A vision foundation model for the tree of life. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 19412–19424, 2024. 3

[14] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. In *Advances in Neural Information Processing Systems*, pages 21875–21911, 2024. 3

[15] Hangjie Yuan, Shiwei Zhang, Xiang Wang, Samuel Albanie, Yining Pan, Tao Feng, Jianwen Jiang, Dong Ni, Yingya Zhang, and Deli Zhao. Rlipv2: Fast scaling of relational language-image pre-training. In *Proceedings of the IEEE International Conference on Computer Vision*, 2023. 3

[16] Fan Zhang, Meng Li, Guisheng Zhai, and Yizhao Liu. Multi-branch and multi-scale attention learning for fine-grained visual categorization. In *International Conference on MultiMedia Modeling*, pages 136–147, 2021. 3

(1)

InternVL2: The image shows a group of six people posing for a photo with two elephants. The elephants are standing on either side of the group, with their trunks extended towards the people. The people are smiling and appear to be enjoying the experience. The setting looks like a natural environment, possibly a wildlife reserve or safari park. The ground is dirt, and there are trees and bushes in the background. One person is holding a stick or pole, possibly for guiding or interacting with the elephants.

LLava-Next: In the heart of a verdant landscape, a group of six individuals and two majestic elephants share a moment of tranquility. The elephants, their gray skin a testament to their age and wisdom, stand on a dirt path that cuts through the lush greenery. Their long trunks and tusks are clearly visible, adding to their imposing presence. The people, clad in jackets and hats, stand in front of the elephants, their faces lit up with smiles. They are positioned in such a way that they form a semi-circle around the elephants, creating a sense of unity and connection. Their casual attire suggests a day of exploration and adventure. The background is a tapestry of trees and bushes, their leaves a vibrant green that contrasts beautifully with the gray of the elephants. The sky above is overcast, casting a soft light over the scene and enhancing the overall serene atmosphere. This image captures not just a moment, but a story - a story of humans and animals co-existing in harmony, a story of adventure and exploration, and a story of nature's beauty.

Cap-Workflow:
The image depicts a group of six people enjoying a unique photo opportunity with three African Bush Elephants (Loxodonta africana) in what appears to be a natural environment, possibly a wildlife reserve or safari park. The elephants, characterized by their large ears and long tusks, stand on either side of the group, with their trunks extended towards the people. The elephants are positioned behind a wooden fence, which is visible in the left region of the image. The people are smiling and seem to be thoroughly enjoying the experience. One individual is holding a stick or pole, potentially for guiding or interacting with the elephants. The setting is outdoors, with a dirt ground and a backdrop of trees and bushes that suggest a savanna or forested area. The people are dressed casually and warmly, indicating a possibly cooler climate. The elephants are spatially arranged with the one on the left behind the one in the center relative to the camera.

(2)

InternVL2: The image shows a stone balustrade with two large, ornate urns placed on it. Each urn is filled with pink and green flowers. In the background, there is a statue of a man standing on a pedestal. The statue is surrounded by trees and greenery. The setting appears to be a park or a public garden.

LLava-Next: The image depicts a serene park setting. In the foreground, there's a stone wall with two large, black vases on top. The vases are filled with pink flowers, adding a splash of color to the scene. The wall appears to be part of a larger structure, possibly a monument or a historical building, as suggested by the architectural details. In the background, there's a statue of a man standing on a pedestal. The statue is positioned in the center of the image, drawing the viewer's attention. The man is dressed in a suit and tie, suggesting a formal or historical figure. The park is lush with trees and greenery, providing a natural backdrop to the man-made structures. The sky is overcast, casting a soft light over the scene. The overall atmosphere of the image is peaceful and contemplative.

Cap-Workflow:
The image depicts a serene outdoor setting, likely a park or public garden, featuring a stone balustrade with two large, ornate urns placed on it. The urns, which are black with elegant handles, are filled with pink flowers and green foliage, adding a vibrant touch of color to the scene. One urn is positioned in the bottom-right region of the image, slightly in front of the other urn located in the center. Both urns are placed against a backdrop of well-maintained greenery and a stone structure, contributing to the lush garden atmosphere. In the background, there is a statue of a man standing on a pedestal with his hands clasped behind his back, situated among the trees and greenery. The statue and the surrounding environment suggest a carefully landscaped area, possibly a historical or cultural site within the park or garden. The overall composition of the image highlights the ornate urns, the colorful flowers, and the statue, creating a visually appealing and tranquil scene.

(3)

InternVL2: The image shows a woman kneeling on the floor in a store, holding a leash attached to a white dog with a fluffy coat. The dog appears to be panting with its tongue out. The woman is wearing glasses and a dark-colored sweatshirt with a Christmas tree design. She is also holding a black purse. The store shelves around them contain various items, including what appears to be a microwave oven and pet supplies. There is a shopping cart in the background.

LLava-Next: In the image, a woman and her white dog are the main subjects. The woman, dressed in a black sweater and blue jeans, is kneeling on the floor. She is holding a leash attached to her dog, which is sitting on the floor with its tongue hanging out, giving the impression of excitement or anticipation.\n\nThe setting appears to be a store, as indicated by the presence of shelves stocked with various items. The shelves are filled with a variety of products, including pet supplies and household items. The store seems well-stocked and organized, suggesting a wide range of products available for customers.\n\nThe woman and her dog are positioned in the foreground of the image, drawing the viewer's attention immediately. Their actions and expressions suggest they might be shopping for something, possibly related to their pet.\n\nOverall, the image captures a moment of interaction between a pet owner and her dog in a retail setting. The woman's attire and the dog's behavior add a touch of everyday life to the scene.

Cap-Workflow :
In the image, a woman is kneeling on a light-colored floor in a store, situated behind a large white dog with a fluffy coat. The dog is sitting on the floor, panting with its tongue out, indicating it might be a bit tired or excited. The woman is holding a leash that is attached to the dog's blue collar, ensuring the pet stays close by. She is wearing glasses, a dark-colored sweatshirt with a Christmas tree design, and jeans. The woman is also holding a black purse and appears to be examining a product on a nearby shelf. The store shelves around them are filled with various items, including what seems to be a microwave oven and pet supplies, suggesting they are in the pet section of the store. In the background, a shopping cart is visible, likely belonging to the woman as she shops with her dog. The scene captures a moment of everyday life in a store, where a woman is shopping with her companion animal, a large white dog, by her side.

(4)

InternVL2: The image shows two young girls outdoors. The girl in the foreground is sitting on the grass, wearing a light blue dress and a white hat with a floral pattern. The girl in the background is standing on a white bench, also wearing a light blue dress and a similar white hat with a floral pattern. Both girls appear to be enjoying themselves in a casual, outdoor setting.

LLava-Next: In the image, there are two young girls engaged in a playful moment. The girl standing is wearing a light blue dress with a white patterned hat. She is standing on a white bench, her arms outstretched as if she's balancing or perhaps pretending to fly. Her position on the bench and her outstretched arms give a sense of motion and joy. Below her, on the ground, is another girl. She is also dressed in a light blue dress, and she's wearing a white hat with a floral pattern. She is lying on her stomach, her hands reaching up towards the standing girl. It appears as if she is trying to touch the standing girl's feet, adding to the playful interaction between the two. The background of the image is a simple white wall, which provides a stark contrast to the colorful scene of the girls' play. The wall's simplicity allows the focus to remain on the girls and their interaction. Overall, the image captures a delightful moment of childhood play, with the girls' matching outfits and hats adding a sense of harmony and unity to the scene.

Cap-Workflow:
The image depicts two young girls enjoying a sunny day outdoors. In the foreground, a girl is sitting on the grass, dressed in a light blue dress and a white hat adorned with a floral pattern. She is positioned in front of a white wooden bench. The background of this scene showcases a grassy area with some greenery, adding to the casual and natural setting. In the center of the image, another girl stands on the white bench, also wearing a light blue dress and a similar white hat with a floral pattern. Her arms are outstretched, and she wears colorful shoes, capturing the joy and freedom of the moment. The plain, light-colored wall behind her provides a simple backdrop that allows the girls and their activities to stand out. Both girls are engaged in their own activities, yet they share a connection through their matching attire and the outdoor setting. The scene is filled with a sense of innocence and playfulness, as the girls seem to be enjoying the simple pleasures of a day spent outside.

Figure 5. A comparison of image captions generated by InternVL2-26B, LLaVA-Next-34B, and Cap-Workflow. We highlight different types of information, including Object Attributes, OCR, HOI, 2D spatial relations and 3D spatial relations.