Patch Matters: Training-free Fine-grained Image Caption Enhancement via Local Perception

Supplementary Material

A. Experiments Details

A.1. Metrics Details

BLEU. The BLEU [1] score was introduced for machine translation. This metric measures the similarity between model-generated text and reference text. BLEU score typically calculates consistent n-grams between the two texts. A higher score indicates greater similarity between the texts.

ROUGE. The ROUGE [2] score was initially designed to evaluate summarization. ROUGE measures the overlap of n-grams. Its variant, ROUGE-L, calculates the similarity between generated text and reference text based on the longest common subsequence.

METEOR. METEOR [3] is another machine translation metric. It is based on the harmonic mean of unigram precision and recall between generated sentences and reference sentences, with improved comparison in aspects such as synonyms, and stem variations.

CIDEr. CIDEr [4] evaluates image descriptions using TF-IDF (Term Frequency-Inverse Document Frequency) weighted n-grams, calculating cosine similarity between candidate captions and reference captions, incorporating both precision and recall.

SPICE. SPICE [5] is a semantic-based metric for image description evaluation. It converts image descriptions into scene graphs, extracting objects, attributes, and their relationships to measure semantic similarity between candidate and reference descriptions.

WMD. WMD [6] is a semantic text similarity metric that uses word vectors to represent words and computes a distance measure to determine the similarity between two text sequences.

CAPTURE. CAPTURE [7] evaluates image descriptions by identifying key visual elements. It uses a scene graph parser to extract objects, attributes, and relationships from both candidate and reference texts. Abstract nouns are filtered using a stop-word list. F1 scores are calculated based on exact matches, synonym matches, and soft matches, aligning with human evaluation standards. The final CAPTURE score is a weighted combination of these F1 scores, defined as:

$$\text{CAPTURE} = \frac{\alpha \cdot F1_{obj} + \beta \cdot F1_{attr} + \gamma \cdot F1_{rel}}{\alpha + \beta + \gamma}, \quad (1)$$

where $\alpha = 5,\ \beta = 5,\ \gamma = 2$.

Polos. Polos[8] is a supervised automatic evaluation metric, which computes scores from multimodal inputs using a parallel feature extraction mechanism that leverages embeddings trained through large-scale contrastive learning. It is designed to better align with human judgments and handle diverse images and texts.

CHAIR. CHAIR[9] is designed for evaluating object hallucination in image captions. It calculates the proportion of objects that appear in a caption but not in an image. Its two variants, CHAIRi and CHAIRs, evaluate the hallucination at the object instance level and the sentence level, respectively. They are calculated as follows:

$$\text{CHAIRi} = \frac{\{\text{hallucinated objects}\}}{\{\text{all objects mentioned}\}}, \quad (2)$$

$$\text{CHAIRs} = \frac{\{\text{sentences with hallucinated objects}\}}{\{\text{all sentences}\}}. \quad (3)$$

LIN-Bench. LIN-Bench [10] is an evaluation framework introduced in the article to assess the readability and linguistic complexity of generated image descriptions. It uses metrics such as ARI, FK, and SMOG. ARI focuses on the number of words in a sentence and the average number of characters per word, FK is based on sentence length and syllable count, and SMOG measures the use of polysyllabic words. Higher scores on these metrics typically indicate that the text contains more information and detail.

CLIP-score and DINO-score. CLIP-score utilizes the CLIP [11] model to extract image embeddings and calculate cosine similarity between the generated image and a candidate image. DINO-score uses the DINOv2 [12] model to extract features from both images and compute cosine similarity. CLIP is trained on image-text datasets and captures semantic features of images, allowing CLIP-score to reflect high-level semantic similarity. DINOv2 is a self-supervised vision model that effectively captures fine-grained visual features, making DINO-score well-suited for assessing detailed visual similarity.

VQA. The VQA task [13] is designed to assess caption quality in conveying image content. We conducted this experiment using 625 images from the VQA-V2 validation set [14] (5,000 questions total). A text-only LLM [15] answered questions based on captions from various methods.

POPE. POPE [16] is a mainstream evaluation metric for multimodal models, primarily focusing on object-level hallucinations. It employs three polling strategies: sampling objects randomly, selecting from popular objects, and choosing among frequently co-occurring objects that do not exist in the image, which is referred to as adversarial sampling. We conduct our evaluation on the MSCOCO [17] validation dataset, which consists of 500 images, each accompanied by 6 questions. The evaluation metrics include Accuracy, Precision, Recall, and F1 score.

A.2. Prompt for Semantic Filtering

In Fig. 2, we demonstrate the detailed prompt used to guide the LLM in analyzing the semantic content of candidate descriptions and categorizing them into three groups. First, we specify the task goal for the LLM. Then, we emphasize key points to remember during the extraction process: 1) Identify sentences that describe the same object across different descriptions and consolidate them into a single sentence; 2) Identify contradictory sentences from different descriptions; 3) Identify sentences that appear only in one description but describe important objects. We also emphasize that each sentence should belong to only one category. Finally, we provide the LLM with manually labeled contextual examples to enhance its ability to follow instructions.

A.3. Prompt for Aggregation

In Figs. 3 to 7, we demonstrate LLM prompts used for aggregation. Fig. 3 shows intra-patch aggregation, where descriptions of the same patch are merged into a single description. In this example, we input three candidate descriptions along with a high-confidence description obtained through semantic filtering and instruct the LLM to merge them based on the high-confidence description, ensuring accuracy and avoiding redundancy. Fig. 4 show prompts for merging when the IoU between the semantic patch and the global image exceeds a certain threshold. The LLM combines descriptions of the key semantic regions based on the high-confidence descriptions obtained through semantic filtering, filling in the missing parts of the global image and correcting any errors within them. Fig. 5 and Fig. 6 show prompts for merging when the IoU between spatial patches exceeds a certain threshold. The LLM combines the descriptions of the two regions based on the high-confidence descriptions obtained through semantic filtering, resulting in a unified description for the merged region. Finally, Fig. 7 provides an example of how to merge descriptions from different spatial patches into a global description. Here, we assume that the IoU between the four spatial patches is below the threshold, so we have four patch descriptions and one global description as input. We prompt the LLM to use the high-confidence patch descriptions to supplement and correct potential hallucinations in the global description.

IoU	Blip2score	CIDEr	METEOR	ROUGE	SPICE	WMD
0.3	0.2	4.03	19.16	20.76	20.89	44.40
0.3	0.3	3.51	19.36	20.77	21.09	44.31
0.3	0.4	4.18	18.54	20.52	20.97	44.21
0.4	0.2	3.35	19.85	21.00	21.16	44.49
0.4	0.3	4.55	19.69	21.04	21.39	44.64
0.4	0.4	4.12	19.21	20.83	21.10	44.46

Table 1. Ablation study of Blip2Score and IoU on DID-Bench.

Method	CIDEr	METEOR	ROUGE	SPICE	WMD
LLaVA-1.6					
Global	0.74	14.18	19.86	19.79	42.24
w/. 2 spatial patches	1.97	15.47	18.98	19.74	42.80
w/. object level	3.11	20.67	20.88	21.40	45.04
ours	4.55	19.69	21.04	21.39	44.64

Table 2. Ablation study of patch numbers and objec level description on DID-Bench.

Method	CIDEr	METEOR	ROUGE	SPICE	WMD
Cambrian					
Global	0.00	6.77	12.62	10.31	35.06
Ours+Cambrian	3.31	15.47	18.41	16.27	39.95
CogVLM					
Global	0.00	7.89	13.78	15.78	38.65
Ours+CogVLM	1.25	14.49	18.00	18.90	41.88

Table 3. Visually Enhanced MLLMs on DID-Bench.

B. Additional Results and Experiments

B.1. Ablation Study in DID-Bench

To evaluate our method regarding the selection of Blip2Score and IoU thresholds, we employ grid search for the experiments, as shown in the Tab. 1. The small variation in the metrics indicates that the nearby thresholds are not sensitive to the experimental results and all surpass the vanilla MLLM. In addition, we have carried out experiments on two aspects: one is solely splitting the image into two patches, and the other is adding object-level information. The results presented in Tab. 2 demonstrate that dividing the image into two patches hinders the model from acquiring perception, while adding object-level information extracted by GRiT[18] is beneficial.

B.2. Experiment on Visually Enhanced MLLMs

To assess the effectiveness of our method on MLLMs with stronger visual capabilities, we conducted experiments on the CogVLM[19] and Cambrian[20]. The results are presented in Tab. 3. As indicated by the data in the table, our method continues to significantly enhance performance on these models.

B.3. CAPTURE Score in DID-Bench

We also evaluated the CAPTURE score on DID-Bench [10], as shown in Tab. 4 and Tab. 5. We present the CAPTURE scores for both open-source and closed-source large models, and it is clear that our method consistently improves the performance of these models on this metric. An interesting observation is that LLaVA and Mini-Gemini outperform many closed-source large models in the CAPTURE score, which could be attributed to their use of the GPT-4V-annotated ShareGPT4V [21] dataset during training. This further highlights the critical role of high-quality image-text descriptions in improving model performance.

B.4. Experiment on LIN-Bench

We evaluated our method on LIN-Bench [10] using images from DID-Bench. LIN-Bench focuses on readability and descriptive detail to assess the quality and complexity of generated text. We also conducted a statistical analysis of the description lengths generated by different methods on the DID-Bench dataset, as shown in the Tab. 7. Since this benchmark is only suitable for descriptions longer than 100 words, we did not use PoCa [13] as a baseline. As shown in Tab. 6, our method achieves higher scores across LIN-Bench metrics (ARI, FK, SMOG), demonstrating that it produces detailed descriptions.

Description	CAPTURE	$F1_{obj}$	$F1_{attr}$	$F1_{rel}$
LLaVA1.5	49.81	59.17	38.26	55.25
LLaVA1.5+IT [10]	53.86	61.96	44.74	56.67
LLaVA1.5+PoCa [13]	44.84	56.21	31.27	50.37
LLaVA1.5+Syn [7]	58.51	65.74	52.24	56.12
LLaVA1.5+Ours	**59.61**	**66.78**	**53.11**	**57.95**
LLaVA1.6	59.58	65.66	54.12	58.01
LLaVA1.6+IT [10]	61.60	67.56	56.68	59.00
LLaVA1.6+PoCa [13]	54.53	60.54	49.11	53.02
LLaVA1.6+Syn [7]	62.67	67.93	59.17	58.31
LLaVA1.6+Ours	**64.48**	**69.61**	**61.27**	**59.71**
Mini-Gemini	62.28	67.46	59.00	57.58
Mini-Gemini+IT [10]	63.24	68.13	60.36	58.16
Mini-Gemini+PoCa [13]	55.05	60.54	50.65	52.49
Mini-Gemini+Syn [7]	64.22	68.75	61.84	58.85
Mini-Gemini+Ours	**65.55**	**69.69**	**63.75**	**59.86**

Table 4. CAPTURE scores of open-source models on DID- Bench are weighted combinations of various F1 scores, where higher F1 scores for objects, attributes, and relations indicate better performance in capturing these aspects.

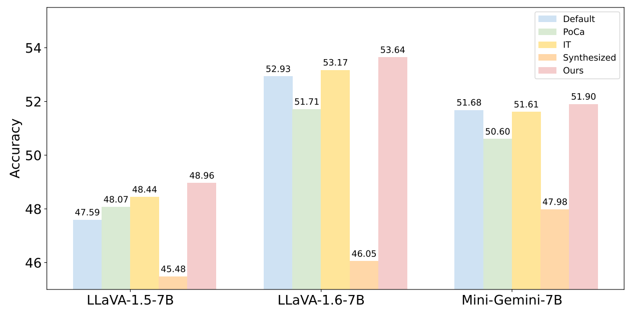


Figure 1. The performance of different methods and models on the VQA task shows that our method achieves the best results.

Description	CAPTURE	$F1_{obj}$	$F1_{attr}$	$F1_{rel}$
GLM-4V-Plus	56.04	64.12	47.84	56.30
GLM-4V-Plus+IT [10]	59.37	66.75	52.50	58.10
GLM-4V-Plus+PoCa [13]	53.25	59.04	47.60	52.88
GLM-4V-Plus+Syn [7]	62.91	68.57	59.17	58.09
GLM-4V-Plus+Ours	**64.72**	**70.55**	**60.83**	**59.89**
GPT-4o	57.40	64.39	50.46	57.29
GPT-4o+IT [10]	60.53	67.26	54.48	58.83
GPT-4o+PoCa [13]	50.40	56.87	43.81	50.70
GPT-4o+Syn [7]	62.93	68.98	58.89	57.93
GPT-4o+Ours	**63.28**	**68.97**	**59.07**	**59.57**
Claude-3.5	51.76	52.14	53.43	46.65
Claude-3.5+IT [10]	64.23	68.68	62.29	57.98
Claude-3.5+PoCa [13]	55.83	59.92	53.32	51.86
Claude-3.5+Syn [7]	63.66	67.92	61.94	57.30
Claude-3.5+Ours	**64.72**	**69.21**	**62.77**	**58.37**

Table 5. CAPTURE scores of close-source models on DID- Bench are weighted combinations of various F1 scores, where higher F1 scores for objects, attributes, and relations indicate better performance in capturing these aspects.

Description	ARI	FK	SMOG	Avg
LLaVA1.5	8.69	8.18	10.80	9.22
LLaVA1.5+IT [10]	8.85	8.35	10.86	9.35
LLaVA1.5+Syn [7]	9.74	8.93	10.97	9.88
LLaVA1.5+Ours	**11.50**	**10.42**	**12.24**	**11.38**
LLaVA1.6	9.71	9.19	11.41	10.10
LLaVA1.6+IT [10]	10.05	9.48	11.58	10.37
LLaVA1.6+Syn [7]	10.75	9.85	11.79	10.80
LLaVA1.6+Ours	**12.49**	**11.37**	**12.98**	**12.28**
Mini-Gemini	9.31	8.55	10.87	9.57
Mini-Gemini+IT [10]	9.52	8.75	10.97	9.74
Mini-Gemini+Syn [7]	10.69	9.57	11.35	10.53
Mini-Gemini+Ours	**12.11**	**10.77**	**12.31**	**11.73**

Table 6. LIN-Bench Results. Our outputs contain a higher number of syllables and characters.

Description	Chars	Sentences	Words
Ground_truth	1211.49	12.29	245.82
LLaVA1.6	587.21	7.27	128.08
LLaVA1.6+IT [10]	709.77	8.05	155.48
LLaVA1.6+Syn [7]	723.62	7.73	156.12
LLaVA1.6+PoCa [13]	275.22	2.53	59.73
LLaVA1.6+Ours	1044.40	9.53	221.70

Table 7. Statistical comparison of image description lengths generated by different methods on the DID-bench dataset, measured in average characters, words, and sentences per description.

B.5. VQA Task

As shown in Fig. 1, our method consistently enhances accuracy across models, with improvements from 0.22% to

Tuning Data	Adversarial				Random				Popular			
	Acc	Precision	Recall	F1	Acc	Precision	Recall	F1	Acc	Precision	Recall	F1
/	82.70	85.62	78.60	81.96	87.67	95.35	79.20	86.53	85.73	91.10	79.20	84.74
{LLaVA}	83.77	88.00	78.20	82.81	87.83	96.86	78.20	86.54	87.07	**95.06**	78.20	85.81
Ours-{LLaVA}	**84.37**	**88.10**	**79.47**	**83.56**	**88.60**	**97.23**	**79.47**	**87.45**	**87.33**	94.30	**79.47**	**86.25**

Table 8. LLaVA-1.5-7B performance with and without fine-tuning on synthesized detailed caption data on the POPE benchmark. "{LLaVA}" refers to detailed captions generated directly by LLaVA-1.5-7B, while "Ours-{LLaVA}" refers to the data constructed using our method. "Acc" denotes accuracy, and "F1" denotes the F1 score.

GroundTruth	Tuning Data	BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDEr	METEOR	ROUGE	SPICE	WMD
GT-{LLaVA}	/	9.16	5.70	3.37	2.07	0.47	10.62	19.54	17.65	41.37
	{LLaVA}	10.02	6.25	3.79	2.40	0.00	11.05	20.19	18.26	41.81
	Ours-{LLaVA}	**34.09**	**19.33**	**10.91**	**6.47**	**5.63**	**17.34**	**23.35**	**20.35**	**43.11**
GT-{GPT4-V}	/	7.50	4.05	1.94	0.98	0.00	9.11	15.67	13.29	37.74
	{LLaVA}	9.05	4.78	2.32	1.20	0.00	9.71	16.31	14.12	37.83
	Ours-{LLaVA}	**27.68**	**13.84**	**6.68**	**3.50**	**3.11**	**15.13**	**19.75**	**16.30**	**39.26**

Table 9. LLaVA-1.5-7B performance with and without fine-tuning on synthesized detailed caption data on the DID-Bench, with GT-{LLaVA} and GT-{GPT4-V} as ground truth captions generated by LLaVA and GPT-4Vision, respectively. "{LLaVA}" refers to detailed captions generated directly by LLaVA-1.5-7B, while "Ours-{LLaVA}" refers to the data constructed using our method.

1.37%, reflecting higher informational richness and enabling more comprehensive responses. In contrast, some methods introduce errors or omit critical information, leading to incorrect responses.

B.6. Fintune Result in POPE and DID-Bench

Experiment Settings. We fine-tuned the LLaVA-1.5-7B model using LoRA [22] with the default pipeline parameters. The learning rate was set to $2e^{-4}$, the LoRA rank was 128, and the scaling factor was 256. For fine-tuning, we used 10k image-text pairs sourced from COCO [17], VG [23], and SAM [24], which were annotated using our method based on LLaVA-1.5-7B. We then compared the performance of our annotations with the directly annotated data from LLaVA.

POPE and DID-Bench Results. We evaluated the impact of our annotation method on model performance by conducting experiments on two benchmarks: POPE [16] and DID-Bench [10]. As shown in the results of POPE Tab. 8, even with only 10k annotated pairs for fine-tuning, our approach significantly mitigates hallucinations compared to the baseline model, which uses direct annotations. While direct annotations provide some improvement, our method, which uses higher-quality annotations, yields more substantial gains. In Tab. 9, we show the results on the DID-Bench benchmark, where fine-tuning with captions generated by our method substantially improves the model's ability to generate high-quality captions, outperforming the baseline.

C. More Case Studies

We provide more qualitative comparisions between MLLM-generated and our-generated image descriptions in Fig. 8 and Fig. 9.

References

[1] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 1

[2] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 1

[3] Alon Lavie and Abhaya Agarwal. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, 2007. 1

[4] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 1

[5] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 382–398. Springer, 2016. 1

[6] Mert Kilickaya, Aykut Erdem, Nazli Ikizler-Cinbis, and Erkut Erdem. Re-evaluating automatic metrics for image captioning. In Mirella Lapata, Phil Blunsom, and Alexander Koller, editors, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 1: Long Papers*, pages 199–209. Association for Computational Linguistics, 2017. 1

[7] Hongyuan Dong, Jiawen Li, Bohong Wu, Jiacong Wang, Yuan Zhang, and Haoyuan Guo. Benchmarking and improving detail image caption. *arXiv preprint arXiv:2405.19092*, 2024. 1, 3

[8] Yuiga Wada, Kanta Kaneda, Daichi Saito, and Komei Sugiura. Polos: Multimodal metric learning from human feedback for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13559–13568, 2024. 1

[9] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045, 2018. 1

[10] Renjie Pi, Jianshu Zhang, Jipeng Zhang, Rui Pan, Zhekai Chen, and Tong Zhang. Image textualization: An automatic framework for creating accurate and detailed image descriptions. *arXiv preprint arXiv:2406.07502*, 2024. 1, 2, 3, 4

[11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1

[12] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 1

[13] Delong Chen, Samuel Cahyawijaya, Etsuko Ishii, Ho Shu Chan, Yejin Bang, and Pascale Fung. What makes for good image captions? *arXiv preprint arXiv:2405.00485*, 2024. 1, 3

[14] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 1

[15] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 1

[16] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305, 2023. 1, 4

[17] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 2, 4

[18] Jialian Wu, Jianfeng Wang, Zhengyuan Yang, Zhe Gan, Zicheng Liu, Junsong Yuan, and Lijuan Wang. Grit: A generative region-to-text transformer for object understanding. *arXiv preprint arXiv:2212.00280*, 2022. 2

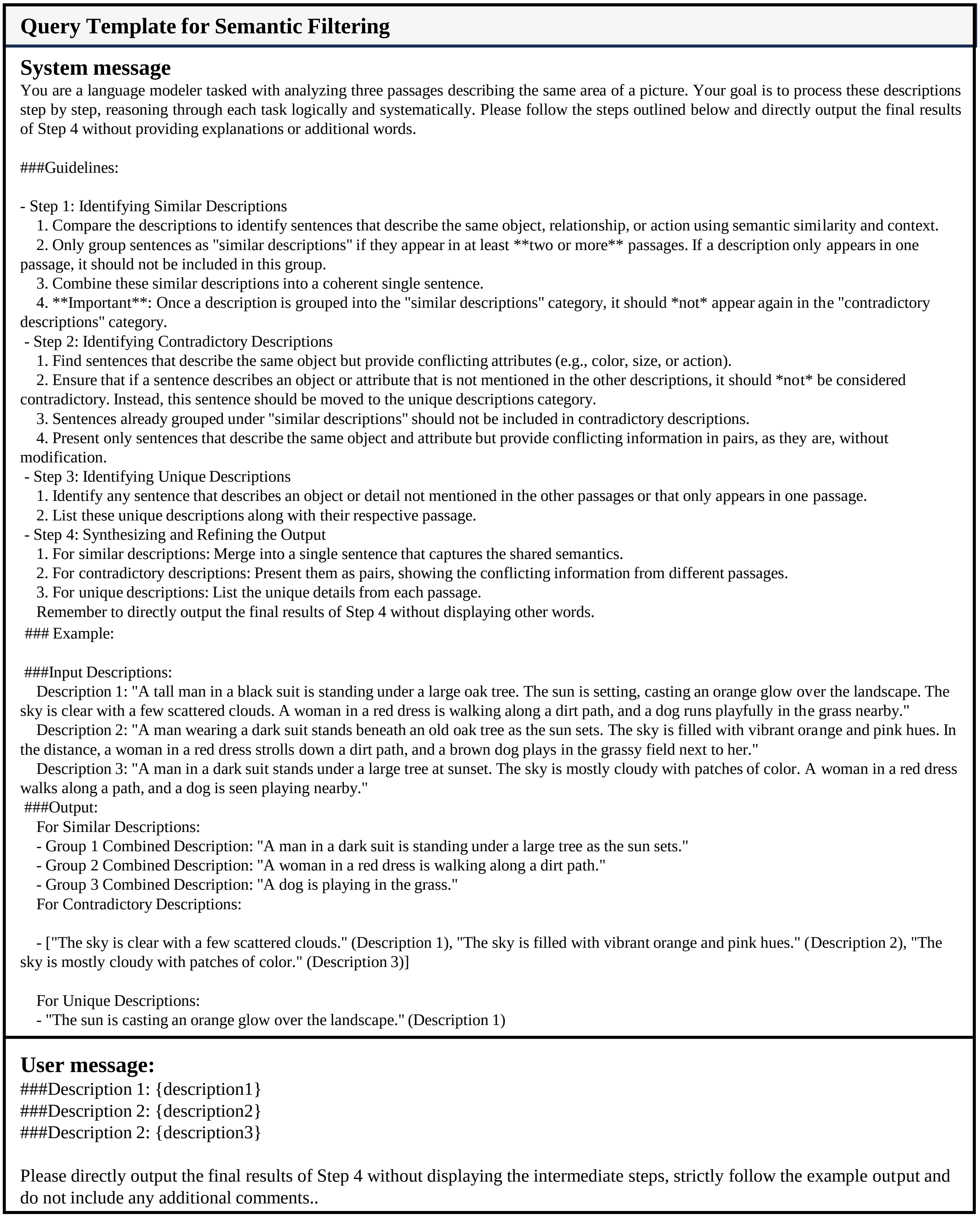
[19] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Song XiXuan, et al. Cogvlm: Visual expert for pretrained language models. *Advances in Neural Information Processing Systems*, 37:121475–121499, 2025. 2

[20] Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356, 2025. 2

[21] 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. *arXiv preprint arXiv:2311.12793*, 2023. 3

[22] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. 4

[23] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense



Query Template for Semantic Filtering

System message

You are a language modeler tasked with analyzing three passages describing the same area of a picture. Your goal is to process these descriptions step by step, reasoning through each task logically and systematically. Please follow the steps outlined below and directly output the final results of Step 4 without providing explanations or additional words.

###Guidelines:

- Step 1: Identifying Similar Descriptions
    1. Compare the descriptions to identify sentences that describe the same object, relationship, or action using semantic similarity and context.
    2. Only group sentences as "similar descriptions" if they appear in at least **two or more** passages. If a description only appears in one passage, it should not be included in this group.
    3. Combine these similar descriptions into a coherent single sentence.
    4. **Important**: Once a description is grouped into the "similar descriptions" category, it should *not* appear again in the "contradictory descriptions" category.
- Step 2: Identifying Contradictory Descriptions
    1. Find sentences that describe the same object but provide conflicting attributes (e.g., color, size, or action).
    2. Ensure that if a sentence describes an object or attribute that is not mentioned in the other descriptions, it should *not* be considered contradictory. Instead, this sentence should be moved to the unique descriptions category.
    3. Sentences already grouped under "similar descriptions" should not be included in contradictory descriptions.
    4. Present only sentences that describe the same object and attribute but provide conflicting information in pairs, as they are, without modification.
- Step 3: Identifying Unique Descriptions
    1. Identify any sentence that describes an object or detail not mentioned in the other passages or that only appears in one passage.
    2. List these unique descriptions along with their respective passage.
- Step 4: Synthesizing and Refining the Output
    1. For similar descriptions: Merge into a single sentence that captures the shared semantics.
    2. For contradictory descriptions: Present them as pairs, showing the conflicting information from different passages.
    3. For unique descriptions: List the unique details from each passage.
    Remember to directly output the final results of Step 4 without displaying other words.

### Example:

###Input Descriptions:

Description 1: "A tall man in a black suit is standing under a large oak tree. The sun is setting, casting an orange glow over the landscape. The sky is clear with a few scattered clouds. A woman in a red dress is walking along a dirt path, and a dog runs playfully in the grass nearby."

Description 2: "A man wearing a dark suit stands beneath an old oak tree as the sun sets. The sky is filled with vibrant orange and pink hues. In the distance, a woman in a red dress strolls down a dirt path, and a brown dog plays in the grassy field next to her."

Description 3: "A man in a dark suit stands under a large tree at sunset. The sky is mostly cloudy with patches of color. A woman in a red dress walks along a path, and a dog is seen playing nearby."

###Output:

For Similar Descriptions:
- Group 1 Combined Description: "A man in a dark suit is standing under a large tree as the sun sets."
- Group 2 Combined Description: "A woman in a red dress is walking along a dirt path."
- Group 3 Combined Description: "A dog is playing in the grass."

For Contradictory Descriptions:

- ["The sky is clear with a few scattered clouds." (Description 1), "The sky is filled with vibrant orange and pink hues." (Description 2), "The sky is mostly cloudy with patches of color." (Description 3)]

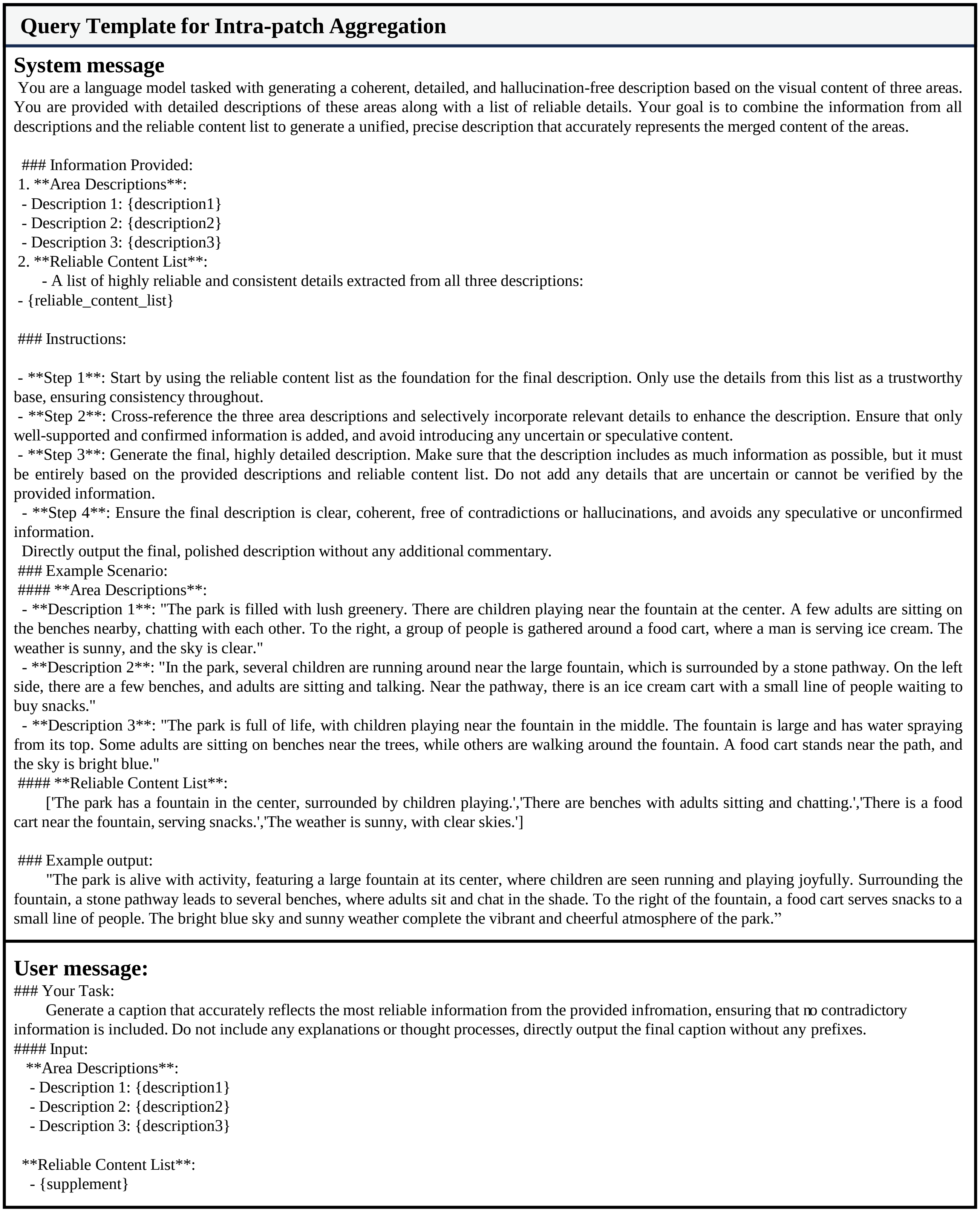
For Unique Descriptions:
- "The sun is casting an orange glow over the landscape." (Description 1)

User message:

###Description 1: {description1}
###Description 2: {description2}
###Description 2: {description3}

Please directly output the final results of Step 4 without displaying the intermediate steps, strictly follow the example output and do not include any additional comments..

Figure 2. The system and user prompts used for semantic filtering query.



Query Template for Intra-patch Aggregation

System message

You are a language model tasked with generating a coherent, detailed, and hallucination-free description based on the visual content of three areas. You are provided with detailed descriptions of these areas along with a list of reliable details. Your goal is to combine the information from all descriptions and the reliable content list to generate a unified, precise description that accurately represents the merged content of the areas.

### Information Provided:
1. **Area Descriptions**:
- Description 1: {description1}
- Description 2: {description2}
- Description 3: {description3}
2. **Reliable Content List**:
- A list of highly reliable and consistent details extracted from all three descriptions:
- {reliable_content_list}

### Instructions:

- **Step 1**: Start by using the reliable content list as the foundation for the final description. Only use the details from this list as a trustworthy base, ensuring consistency throughout.
- **Step 2**: Cross-reference the three area descriptions and selectively incorporate relevant details to enhance the description. Ensure that only well-supported and confirmed information is added, and avoid introducing any uncertain or speculative content.
- **Step 3**: Generate the final, highly detailed description. Make sure that the description includes as much information as possible, but it must be entirely based on the provided descriptions and reliable content list. Do not add any details that are uncertain or cannot be verified by the provided information.
- **Step 4**: Ensure the final description is clear, coherent, free of contradictions or hallucinations, and avoids any speculative or unconfirmed information.

Directly output the final, polished description without any additional commentary.

### Example Scenario:
#### **Area Descriptions**:
- **Description 1**: "The park is filled with lush greenery. There are children playing near the fountain at the center. A few adults are sitting on the benches nearby, chatting with each other. To the right, a group of people is gathered around a food cart, where a man is serving ice cream. The weather is sunny, and the sky is clear."
- **Description 2**: "In the park, several children are running around near the large fountain, which is surrounded by a stone pathway. On the left side, there are a few benches, and adults are sitting and talking. Near the pathway, there is an ice cream cart with a small line of people waiting to buy snacks."
- **Description 3**: "The park is full of life, with children playing near the fountain in the middle. The fountain is large and has water spraying from its top. Some adults are sitting on benches near the trees, while others are walking around the fountain. A food cart stands near the path, and the sky is bright blue."
#### **Reliable Content List**:

['The park has a fountain in the center, surrounded by children playing.','There are benches with adults sitting and chatting.','There is a food cart near the fountain, serving snacks.','The weather is sunny, with clear skies.']

### Example output:

"The park is alive with activity, featuring a large fountain at its center, where children are seen running and playing joyfully. Surrounding the fountain, a stone pathway leads to several benches, where adults sit and chat in the shade. To the right of the fountain, a food cart serves snacks to a small line of people. The bright blue sky and sunny weather complete the vibrant and cheerful atmosphere of the park."

User message:

### Your Task:

Generate a caption that accurately reflects the most reliable information from the provided infromation, ensuring that no contradictory information is included. Do not include any explanations or thought processes, directly output the final caption without any prefixes.

#### Input:
Area Descriptions:
- Description 1: {description1}
- Description 2: {description2}
- Description 3: {description3}

Reliable Content List:
- {supplement}

Figure 3. The system and user prompts used for intra-patch aggregation query.

Query Template for Inter-patch Aggregation (if semantic patch's IoU > threshold)

System message

### Input ###

- You will receive a global description that provides an overall view of the image.
- Additionally, you will be provided with a detailed region description, which focuses on a specific area within the image.

### Task Objective ###

- Modify and enhance the global description by integrating accurate details from the region description.
- Do not introduce any elements or details that are not present in the given region description or the original global description.
- The output should be an enriched global description with relevant details seamlessly integrated from the region description.

### Input INFORMATION EXPLANATION ###

1. Global Description: This is the initial, broader description of the image, covering the main elements and objects but potentially lacking specific details.
2. Region Description: It offers detailed information about a specific section of the image, often containing more precise or additional details about objects and actions.

### Guidelines ###

- Integrate specific details from the region description into the global description where relevant.
- Ensure the updated global description remains coherent, natural, and more detailed than the original.
- Do not add any new elements or make assumptions beyond the given descriptions.

### Example ###

Global Description:
A busy marketplace with various stalls can be seen, with people walking around and shopping. There are different goods on display, such as fruits, vegetables, and clothes. In the background, the sky is partly cloudy, and a few birds are flying.

Region Description:
A fruit stand in the center of the market is displaying piles of bright oranges, green apples, and ripe bananas. A vendor wearing a green apron is helping a customer select some oranges. The customer is holding a wicker basket.

Your Modified Global Description:
A busy marketplace with various stalls can be seen, with people walking around and shopping. There are different goods on display, such as fruits, vegetables, and clothes. In the center of the market, a fruit stand showcases bright oranges, green apples, and ripe bananas, while a vendor wearing a green apron assists a customer in selecting some oranges. The customer holds a wicker basket. In the background, the sky is partly cloudy, and a few birds are flying.

User message:

### Your Task:
Please provide the modified description directly.
Global Description:
{description1}
Region Description:
{description2}

Figure 4. The system and user prompts used for inter-patch aggregation (if semantic patch's IoU >threshold) query.

image annotations. *International journal of computer vision*, 123:32–73, 2017. 4

[24] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 4

Query Template for Inter-patch Aggregation (if spatial patch’s IoU > threshold)

System message

You are a language model tasked with generating a coherent and hallucination-free caption based on the visual content of two image regions. You are provided with detailed descriptions of these regions along with a list of reliable details extracted from the visual content. Your goal is to combine the information from both descriptions and the reliable content list to generate a unified caption that accurately represents the merged visual content of both regions.

### Information Provided:
1. **Image Details**:
- Image Size: {width, height}
2. **Region 1**:
- Location: {region1_location}
- Description: {region1_description}
3. **Region 2**:
- Location: {region2_location}
- Description: {region2_description}
4. **Reliable Content List**:
- A list of highly reliable and consistent details extracted from both Region 1 and Region 2:
- {reliable_content_list}

### Instructions:
- **Step 1**: Use the reliable content list to establish the foundation of the final caption, ensuring that only trustworthy information is included.
- **Step 2**: Cross-reference the descriptions of Region 1 and Region 2 to enhance the caption, ensuring that the final description is coherent and accurately merges the visual content of both regions.
- **Step 3**: Generate a final, hallucination-free caption that avoids any contradictions or conflicting information, while ensuring the description remains clear and cohesive.

### Example Scenario:
#### Image Details:
- Image Size: [1024, 768]
#### Region 1:
- Location: [150, 250, 550, 650]
- Description: "A man in a green jacket is standing near a large tree, with a park bench nearby. He is holding a small book, and there are flowers around the base of the tree. The scene suggests a calm, outdoor setting."
#### Region 2:
- Location: [600, 250, 1000, 650]
- Description: "A man in a green jacket is sitting on a park bench next to a tree, holding a book. The bench is surrounded by flowers, and there is a small bird perched on the back of the bench. The atmosphere feels peaceful, and the weather appears clear."
#### Reliable Content List:
-['A man in a green jacket is near a tree.’ ,'The man is holding a book.’,'There are flowers around the tree.’,'The man is near or sitting on a park bench.’,'A bird is perched on the back of the bench.’,'The atmosphere is peaceful and calm.’]
#### Example Output:
"A man in a green jacket is holding a small book. Flowers surround the base of the tree, and a bird is perched on the back of the bench. The scene suggests a peaceful, calm outdoor environment, with the man seemingly enjoying a moment of quiet reflection.“

User message:

### Your Task:
Generate a caption that accurately reflects the most reliable information from the provided triples and image regions, ensuring that no contradictory information is included. Do not include any explanations or thought processes, directly output the final caption without any prefixes.
#### Input:
Image Size: {image_size}
Region 1:
- Location: {region1_location}
- Description: {region1_description}
Region 2:
- Location: {region2_location}
- Description: {region2_description}
Reliable Content List:
- {supplement}

Figure 5. The system and user prompts used for inter-patch aggregation (if spatial patch’s IoU >threshold) query.

Query Template for Inter-patch Aggregation (combine all patches description into a global caption)

System message

###Input###

• You will receive a global caption that provides an initial description of an image, capturing the main elements and semantics. Note that the global caption may contain errors, including hallucinated objects or vague descriptions.

• Along with the global caption, you will also receive region-specific captions, which focus on particular parts of the image and are more reliable due to a hallucination filter. Each region caption is accompanied by location coordinates that define a rectangle within the image using a normalized coordinate system (x and y range from 0 to 1).

###Task Objective###

• Your goal is to modify and enhance the global caption by integrating accurate details from the region captions and their location.

• The global caption should be enriched with specific, accurate details from the regions and corrected where necessary.

• Focus on using the region captions to correct any inaccuracies or hallucinations in the global caption.

• The updated global caption must contain more detail than the original global caption by including relevant information from the region captions.

• You only give the updated global caption as output, without any additional information.

• Do NOT give any explanation or notes on how you generate this caption.

###Input INFORMATION EXPLANATION###

1. Global Description: It provides the initial global image description, which captures the primary semantic information of the image. However, some of the described objects are hallucinated, and certain details are either missing or insufficiently described, requiring additional information for correction and enhancement.

2. Region Description: It provides descriptions of different regions, focusing on specific parts of the image. These include more detailed object features and finer details. Additionally, this section has undergone hallucination filtering, making the descriptions more reliable compared to the global description.

3. Region Location: It uses a normalized coordinate system where both x (horizontal) and y (vertical) axes range from 0 to 1. The x-coordinate starts at 0 on the image's left edge and increases to 1 towards the right edge. Similarly, the y-coordinate starts at 0 at the top edge and increases to 1 towards the bottom. This system uses four coordinates to define the corners of a rectangle within the image: [x1, y1, x2, y2], representing the top-left and bottom-right corners of the rectangle, respectively.

###Guidelines###

• Through the extra information of different regions, some objects may represent the same thing. When adding objects to the original description, it is important to avoid duplication.

• Combine Information: Extract and integrate key details from both the global and local (region) captions, giving priority to the region captions for more specific or accurate details.

• Modify and Enhance: Add relevant details from the region captions to enrich the global description. Correct any hallucinations or inaccuracies in the global caption using the region captions.

• Consider Location: Ensure that spatial information from the region captions is incorporated to provide a more coherent and accurate description of the image.

• Filter Noise: Remove any conflicting or irrelevant information from the global caption, especially if it contradicts verified details in the region captions.

• Enhance Detail: Ensure that the final global caption contains more detailed and refined visual information than the original, using the region captions to add specificity.

###In-Context Examples###

[Chain of thought is placed within a pair of "@@@" (remember only in the Examples will you be provided with a chain of thoughts to help you understand; in the actual task, these will not be given to you.)]

###Example:###

Global Description:

Three friends are sitting on a bench in the park, chatting and laughing. The sun is shining brightly, and people are scattered around the park, enjoying the weather. A man is jogging along the path, and there's a pond with ducks swimming nearby.

Region 1:

- Location: [0.10, 0.25, 0.40, 0.60]
- Description: Two women are sitting on a bench under a tree. One is wearing a blue T-shirt and shorts, while the other is dressed in a white sundress. They are chatting and laughing, with one of the women holding a cup of coffee. There's a picnic blanket on the ground near the bench with some snacks on it.

Region 2:

- Location: [0.50, 0.10, 0.80, 0.50]
- Description: A man in a green T-shirt and jeans is standing next to the bench, holding a water bottle in one hand while looking at the two women. He seems to be engaged in their conversation, smiling and occasionally glancing at his phone.

Figure 6. The system prompts used for inter-patch aggregation (combine all patches description into a global caption) query (first half).

Query Template for Iner-patch Aggregation (combine all patches description into a global caption)

System message (continued)

Region 3:
- Location: [0.60, 0.70, 0.80, 0.90]
- Description: A man is jogging along a path, wearing headphones and a blue tank top. He is passing by a group of trees and a flower bed filled with brightly colored flowers.

Region 4:
- Location: [0.10, 0.60, 0.30, 0.90]
- Description: A small pond with ducks swimming near the shore. A child is throwing breadcrumbs into the water, and the ducks are gathering around. A couple is sitting on a bench nearby, watching the scene.

Chain of Thought:
1. The global description mentions "three friends," but the region descriptions confirm there are only two women sitting on the bench and one man standing nearby. This discrepancy should be corrected.
2. The global description correctly refers to a man jogging and a pond with ducks, which are supported by Region 3 and Region 4. These details should remain but be expanded with more specific information from the region descriptions.
3. Additional details in Region 1, such as the picnic blanket and coffee cup, should be included to enrich the description.
4. The spatial arrangement of the park (e.g., the bench under a tree, jogging path, and pond) should be better represented in the global description, incorporating all region-specific details.

Modified Description:
In a sunny park, two friends are sitting on a bench under a tree, chatting and laughing. One woman is wearing a blue T-shirt and shorts, while the other is dressed in a white sundress, holding a cup of coffee. A picnic blanket is spread on the ground next to the bench, with snacks scattered across it. Standing next to the bench is a man in a green T-shirt and jeans, holding a water bottle in one hand while engaging in their conversation, occasionally glancing at his phone and smiling. Further along the path, a man in a blue tank top and headphones is jogging, passing by a group of trees and a flower bed filled with vibrant flowers. Near the edge of the park, a small pond is home to several ducks swimming near the shore. A child is throwing breadcrumbs into the water, attracting the ducks, while a couple sits on a bench nearby, watching the scene with a sense of calm. The sun continues to shine brightly, casting a warm light over the park as people enjoy the beautiful weather.

User message:

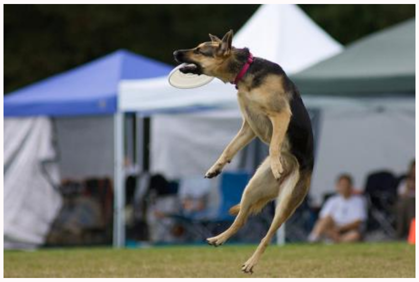
###TASK###
Please provide the modified description directly.
Global Description:
{global_description}
Region1
- Location: {region1_location}
- Description: {region1_description}

Region2
- Location: {region2_location}
- Description: {region2_description}

Region3
- Location: {region3_location}
- Description: {region3_description}

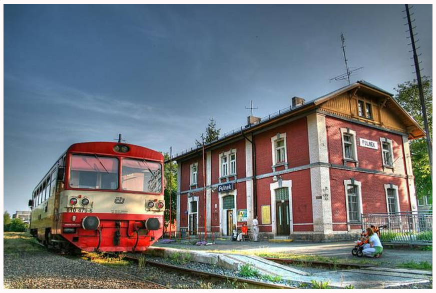
Region4
- Location: {region4_location}
- Description: {region4_description}

Figure 7. The system and user prompts used for inter-patch aggregation (combine all patches description into a global caption) query (second half).



GLM-4V-Plus: The image showcases a German Shepherd dog in mid-air, leaping to catch a white frisbee. The dog is wearing a pink collar and is captured at the peak of its jump, with all four legs off the ground. The background features a grassy area with a few people sitting on chairs, and there are tents set up in the distance, suggesting an outdoor event or gathering.

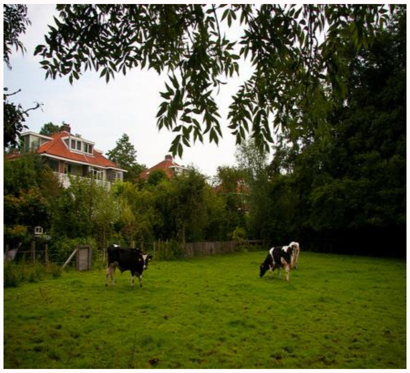
GLM-4V-Plus+Ous: A German Shepherd dog is leaping to catch a white frisbee in mid-air, its pink collar visible as it soars through the air with all four paws lifted off the ground. The scene is set in an outdoor area with several people seated on chairs, possibly observing the dog's performance or waiting for their turn to play. Several tents, including a large white tent, a smaller blue tent and a green tent, are visible in the background. The dog's mouth is wide open, displaying its teeth, and it is holding the frisbee firmly between its teeth, showcasing its agility and focus. The background features several tents, suggesting an outdoor setting, with greenery adding a touch of nature to the scene. The image captures a moment of action and playfulness, showcasing the dog's agility and enthusiasm for the game, likely in a park or a field where people and their pets gather to enjoy recreational activities.



GPT-4o: The image shows a train station with a train on the tracks. The station building is made of red brick and has a sign that reads \"Fulnek.\" There are a few people near the building, including someone with a stroller. The scene is set in a calm, sunny environment.

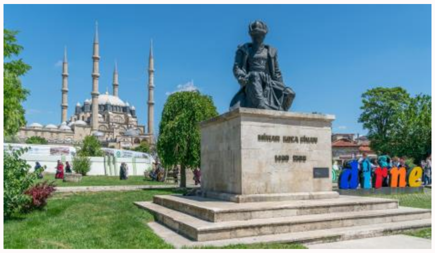
GPT-4o+Ours: The train station with a red and cream-colored train is situated in a calm, sunny environment surrounded by greenery. The train, stationary on the tracks, offers a clear view of its front, including windows and lights. The station building, made of red brick, bears a sign that reads \"Fulnek.\" A few people are nearby, including someone with a stroller,a person in red sitting, and another individual standing near the entrance of the station building. The scene is set against a clear blue sky with a few clouds, and trees are visible on both sides of the image. The station building features a brick exterior with white accents around the windows and a sloped roof with a window beneath the gable. The surrounding area is peaceful with no apparent activity happening around the train.

Figure 8. Visualization of the original description and the modified description. Shared information and newly added details are highlighted in different colors for clarity.



GLM-4V-Plus: The image showcases a peaceful rural scene with a group of cows grazing in a lush green field. The cows are scattered across the field, with some standing closer to the foreground and others further in the background. The field is surrounded by a wooden fence, providing a boundary for the grazing area. In the background, there are houses with red roofs, indicating a residential area close to the field. The sky is overcast, suggesting a cloudy day. The overall atmosphere of the image is serene and tranquil, capturing the essence of rural life.

GLM-4V-Plus+Ours: A peaceful rural scene with a group of cows grazing in a lush green field, surrounded by a wooden fence that encloses the grazing area. The field, with its vibrant and well-maintained grass, indicates a healthy pasture. The field is surrounded by a residential area with several houses, each with a red roof and white walls, one featuring a balcony on the second floor, creating a sense of tranquility. Three black and white cows are scattered across the field, with one on the left side, another in the center, and the third on the right side, all appearing to be in a relaxed state, enjoying their time in the open field. The trees with green leaves, located in the residential area, has vibrant and healthy leaves, indicating thriving trees, and their branches create a natural archway above. The sky is overcast, suggesting a cloudy day, contributing to the serene and tranquil atmosphere that captures the essence of rural life.



LLaVA-1.5: The image features a statue of a man sitting on a pedestal in a park. The statue is located near a tree, and there is a bench nearby for visitors to sit and enjoy the surroundings. The park is situated in front of a large building, possibly a mosque, which can be seen in the background. There are several people in the park, some of them standing and others sitting on the bench. A few individuals are also carrying handbags, indicating that they might be visitors or park-goers. The overall atmosphere of the scene is peaceful and inviting.

LLaVA-1.5+Ours: The image features a majestic mosque with a multi-domed structure, set against a clear blue sky. The mosque is adorned with four towers and a domes, giving it a sense of grandeur and majesty. The tall and majestic minarets stand in front of the mosque, creating a picturesque scene. The overall atmosphere is serene, with the combination of the mosque's grandeur and the serene sky creating a captivating atmosphere for viewers. The mosque is located in a city, surrounded by greenery, adding a touch of nature to the scene. In front of the mosque, a statue of a man sits on a pedestal, surrounded by trees in a peaceful and serene atmosphere, with a blue sky above. The park is situated in front of a large stone wall with a predominantly light color and a slightly greyish tone, giving it a rustic and historical appearance. Several people are in the park, some standing and others sitting on benches, enjoying the surroundings. A few individuals are also carrying handbags, indicating that they might be visitors or park-goers. The overall atmosphere of the scene is peaceful and inviting, with a social gathering or event taking place outdoors, as people are scattered throughout the park, enjoying their time.

Figure 9. Visualization of the original description and the modified description. Shared information and newly added details are highlighted in different colors for clarity.